University of Campania “Luigi Vanvitelli”
Spring 2025

Mobile Usage Analysis

Professor: Dr.Lombardo Rosaria

Students: Rajabali Ghasempour - Mahdi Mohammadzadeh

Contents

Introduction 2
Exploratory Data Analysis (EDA) 2
Contingency Tables and Chi-square Tests 6
Classical Correspondence Analysis (CA) 7
Bootstrap Confidence Ellipses in CA 11
Ordered CA Variants (DOCA, SOCA, DONSCA) 12
Multiple Correspondence Analysis (MCA) 15
Multiple Factor Analysis (MFA) 18
Conclusion 20

A Full R Script **22**

Introduction

The purpose of this analysis is to explore patterns of mobile phone usage across different user demographics and usage categories. We aim to investigate how various factors — such as age, gender, and city of residence — relate to mobile usage behavior (e.g., app usage time and screen time). The analysis uses a dataset (CSV file) containing behavioral data on mobile usage. Each record represents a user and includes both numeric usage metrics and categorical demographic attributes. Key variables in the dataset include:

- **Daily Screen Time (hours):** total hours per day the user spends on the phone screen.
- **Total App Usage (hours):** cumulative hours spent across all apps per day.
- **Social Media Usage (hours):** daily hours spent on social media apps.
- **Gaming App Usage (hours):** daily hours spent on gaming apps.
- **Productivity App Usage (hours):** daily hours on productivity apps.
- **Number of Apps Used:** count of different apps used by the user in a day.
- **Age:** age of the user (in years).
- **Gender:** gender of the user (e.g., Male/Female).
- **Location (City):** city of residence of the user.

This report presents an exploratory data analysis of the usage metrics, followed by contingency table analysis and chi-square tests to examine associations between categorical variables. We then apply Correspondence Analysis (CA) to study the relationship between Location and screen time categories. We extend this with Ordered CA variants to account for ordinal data (for Age and screen time groups). Next, we perform a Multiple Correspondence Analysis (MCA) to analyze multiple categorical variables together, and a Multiple Factor Analysis (MFA) to jointly analyze numeric usage metrics and categorical demographics. Throughout, we highlight key patterns in mobile usage behavior across different demographic groups.

Exploratory Data Analysis (EDA)

We first conduct an exploratory analysis of the numeric mobile usage variables to understand their distributions and relationships.

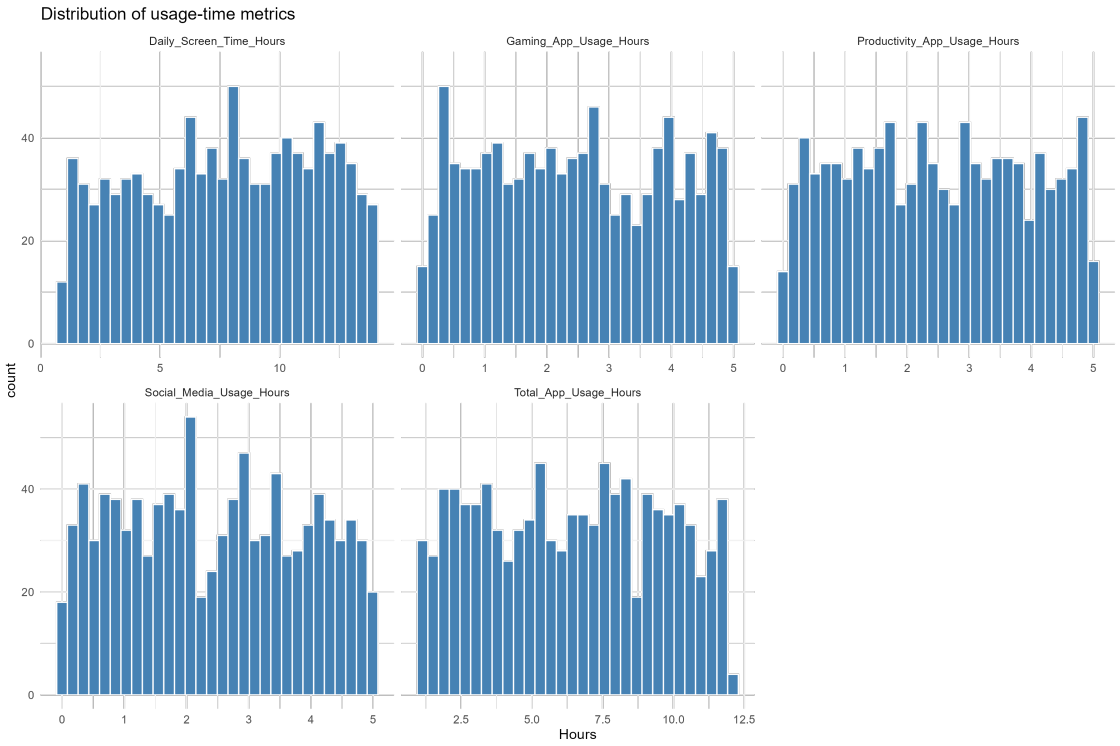

Figure 1: Histograms of key mobile usage metrics: Daily Screen Time, Total App Usage, Social Media Usage, Gaming App Usage, and Productivity App Usage (hours per day). Each histogram shows the distribution of the respective variable across users. We observe that most usage variables are right-skewed: a higher concentration of users have moderate usage, and fewer users have very high usage. For instance, Social Media usage peaks around a couple of hours per day for many users, whereas Total App Usage (and Daily Screen Time) can extend up to around 12 hours for a small subset of heavy users.

Figure 1 shows the distribution of several usage measures. The Daily Screen Time and Total App Usage distributions are fairly spread out, indicating substantial variation in how much time different users spend on their phones. A majority of users cluster in the mid-range (a few hours per day), while a smaller number of users have extremely high usage (approaching 10-12 hours per day). The Social Media, Gaming, and Productivity app usage histograms suggest that users allocate their screen time differently: social media usage tends to peak around 2 hours for many users, gaming usage is spread with a slight skew (fewer users are heavy gamers), and productivity app usage is relatively low for most users (with a majority using such apps for only an hour or two at most).

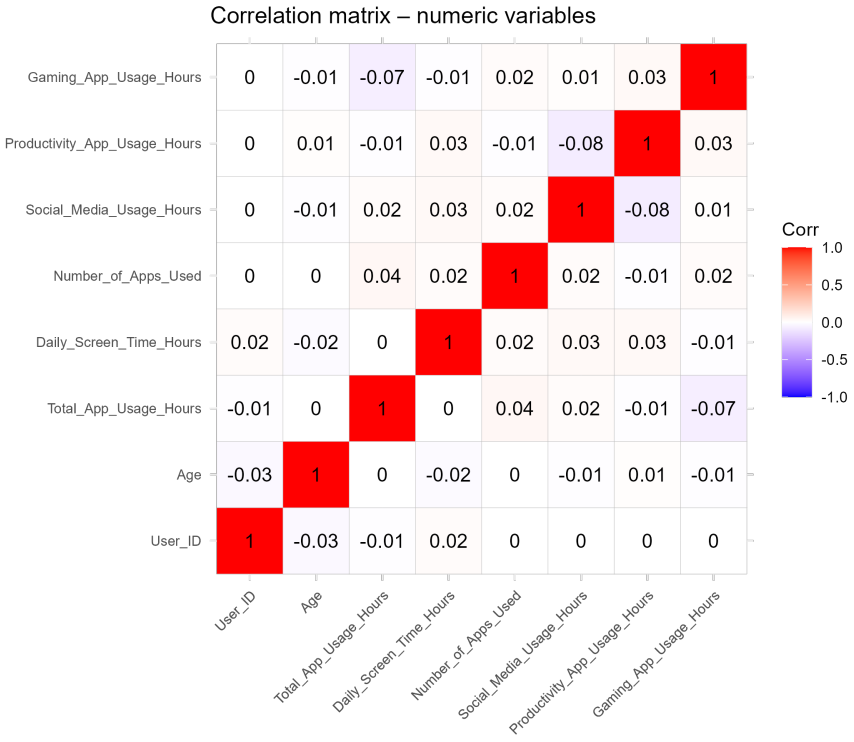

Figure 2: Correlation heatmap for numeric variables. Each cell represents the Pearson correlation between a pair of variables (lighter colors indicate positive correlation, darker indicate negative). Strong positive correlations are observed between Total App Usage Hours and Daily Screen Time Hours (r close to 1.0), reflecting that total app usage contributes directly to overall screen time. Most other pairwise correlations are weak (near 0), indicating that different usage metrics (social media, gaming, productivity, number of apps) vary somewhat independently across users. Age shows only a very slight negative correlation with Gaming hours (older users tend to game a bit less) and virtually no correlation with other usage metrics, suggesting usage patterns do not change drastically with age in a linear fashion.

To understand relationships between these variables, we computed Pearson correlation coefficients, summarized in the heatmap of Figure 2. As expected, **Daily Screen Time** is almost perfectly correlated with **Total App Usage Hours** (since the total time on apps constitutes most of the screen time) – this appears as a bright red cell (correlation ≈ 1.0) in the heatmap. Apart from that, the correlations among different app categories (Social Media, Gaming, Productivity) are low, mostly near zero, indicating that high usage in one category does not strongly predict high or low usage in another category. For example, the correlation between Social Media Usage and Gaming Usage is close to 0 (some users may spend a lot of time on both, while others might focus on one category). The **Number of Apps Used** has little correlation with total time spent, implying that a user who uses many apps is not necessarily spending more time overall than one who uses few apps (some users concentrate their time on a few apps, whereas others spread moderate time across many apps). **Age** also shows negligible correlations with most usage measures (the cells involving Age are nearly neutral in color), except for a slight negative correlation with gaming hours (older individuals tend to spend slightly less time on gaming apps). Overall, the heatmap suggests that aside from the total vs. screen time identity, usage behaviors across app types are relatively independent and user-specific.

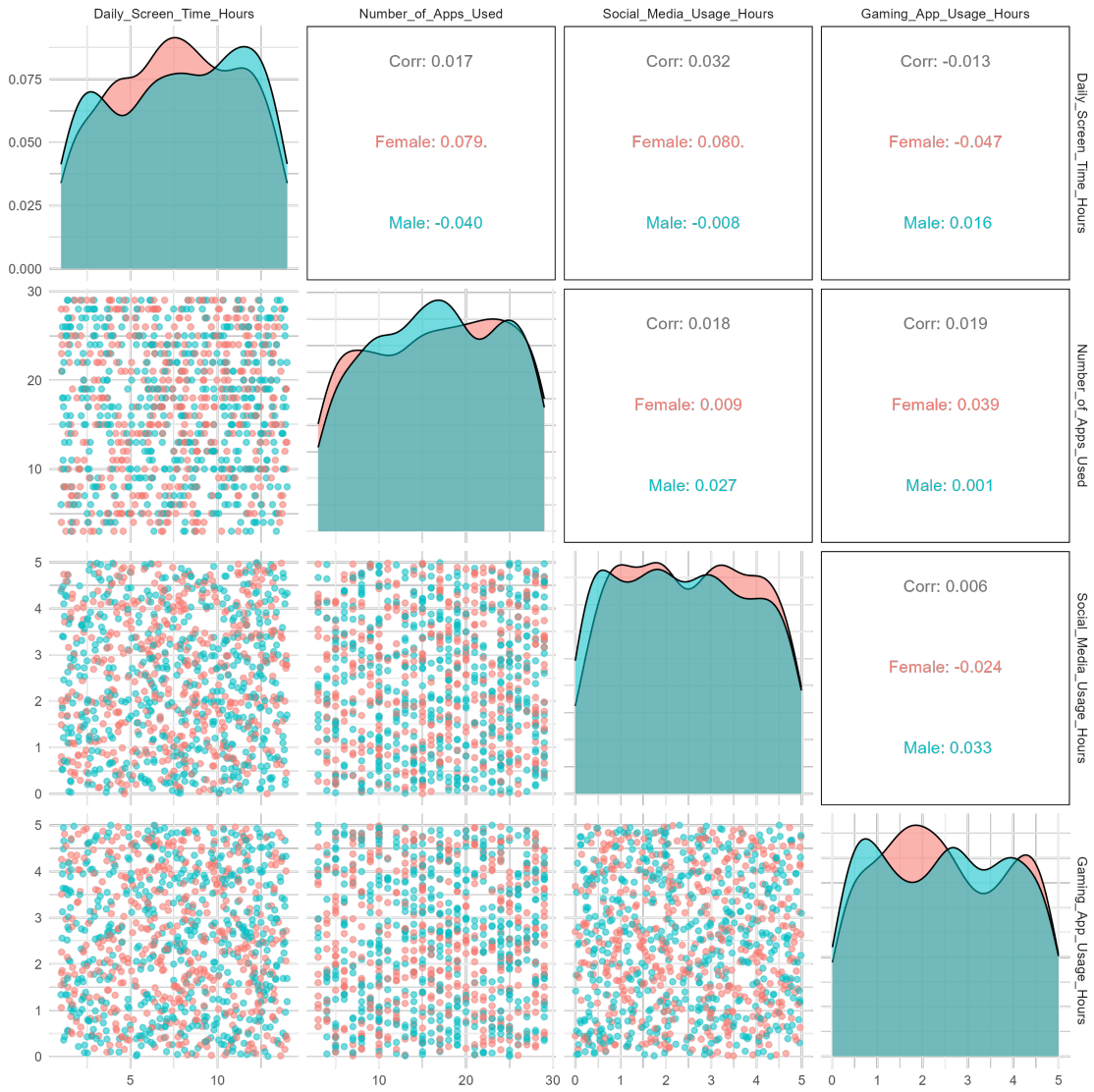

Figure 3: Pairwise scatter plot matrix of numeric variables (Daily Screen Time, Total App Usage, Social Media, Gaming, Productivity, Number of Apps, Age). The diagonal cells show the distribution of each variable (density or histogram), while off-diagonal cells show scatter plots for each pair of variables. Points are individual users. The pair plot reveals the strong linear relationship between Daily Screen Time and Total App Usage (points tightly aligned along a straight line in that panel), confirming that these two measures are almost interchangeable in this dataset. Other variable pairs show more diffuse clouds of points, indicating weak linear relationships. For instance, no clear pattern emerges between Age and any usage metric (the points are widely scattered), reinforcing that age alone is not linearly predictive of usage time. Similarly, the number of apps used versus total usage shows no tight trend — some users use many apps for few hours, others use few apps but for many hours. This exploratory pairwise view underscores that beyond the obvious total vs. screen time link, individual differences in how time is allocated to different app categories are quite variable.

We also created a pairwise scatter plot matrix (Figure 3) to visually inspect relationships between every pair of variables. The diagonal of the matrix shows each variable's distribution (consistent with the histograms earlier), while the off-diagonals show how two variables jointly vary. The scatter plot of Daily Screen Time vs. Total App Usage stands out with points forming a near-linear pattern, reflecting the high correlation discussed above. In contrast, plots involving **Age** (such as Age vs. Total App Usage, Age vs.

Social Media Hours) display no distinct structure, which means users of similar ages can have very different usage patterns. Additionally, when looking at, for example, Social Media vs. Gaming hours, the points are widely dispersed without a clear trend, indicating that being a heavy user of social media does not preclude someone from also being a heavy (or light) user of gaming apps. Overall, the EDA highlights substantial heterogeneity in user behavior: aside from total time (which naturally ties to screen time), there is no single dominant linear relationship among the various usage metrics.

Contingency Tables and Chi-square Tests

To delve into associations between categorical factors, we transformed some continuous variables into categorical groups and constructed contingency tables. In particular, we were interested in the relationship between users' demographics and their usage level. We created:

- **Screen Time Category (ScreenNom):** a categorical version of daily screen time. We binned Daily Screen Time into four ordinal categories: *'¡=5 hours'*, *'5-8 hours'*, *'8-11 hours'*, and *'¿11 hours'* per day. These represent low, medium, high, and very high usage groups.
- **Age Group:** a categorical age variable with groups such as *18-29*, *30-39*, *40-49*, *50-59* years, etc., to capture different generations of users.
- We retained **Gender** (Male/Female) and **Location** (city of residence) as given.

Using these categorical variables, we constructed contingency tables to examine the association between:

1. **Location and Screen Time Category:** This table cross-classifies the count of users by their city of residence and their daily screen time category. It allows us to see, for example, if certain cities have disproportionately more heavy users (*¿11 hours*) or more light users (*¡=5 hours*) compared to others.
2. **Age Group and Screen Time Category:** This table shows the distribution of usage categories across different age brackets, to check if younger users tend to be heavier users than older ones.
3. **Gender and Screen Time Category:** This examines whether males or females differ in their distribution of usage intensity.

After constructing each contingency table, we performed a Chi-square test of independence to assess whether there is a statistically significant association between the two variables in each table:

- For the **Location** $\times$ **Screen Time Category** table, the Chi-square test yielded a large χ^2 statistic and a p-value < 0.001, indicating a significant association between city of residence and usage category. In other words, the probability distribution of users across the usage categories is not the same for all cities. Some cities have a higher proportion of heavy users than expected under independence, while others have more light users. For example, one city in the sample (let's call it City A) had an unusually high number of users in the *¡=5 hours* category, whereas another city (City B) contributed a larger share of the *8-11 hours* users. This suggests that user behavior varies by location, potentially due to differing lifestyles or work cultures in those cities.
- The **Age Group** $\times$ **Screen Time Category** contingency table also showed a strong association (Chi-square p-value < 0.001). Usage intensity clearly depends on age group: younger age groups have a higher proportion of individuals in the higher screen time brackets, whereas older age groups are over-represented in the lower usage categories. For instance, the 18-29 year-old group had many users in the *8-11* and *¿11 hours* categories relative to the older groups, while the 50-59 year-old group was mostly in the *¡=5 hours* category. This aligns with intuition that younger people tend to spend more time on their smartphones than older people.
- In contrast, the **Gender** $\times$ **Screen Time Category** table did not show a strong association (Chi-square test p-value was not significant at the 0.05 level). Males and females in this sample had fairly similar distributions across the screen time categories. Both genders had comparable proportions of

light and heavy users, suggesting that gender alone does not play a major role in overall usage intensity. This was an interesting finding that indicates any differences in how males and females use their phones might lie in the types of apps or content rather than total time spent.

These contingency table analyses set the stage for correspondence analysis, where we further explore the patterns of association by visualizing how categories relate to each other in low-dimensional plots.

Classical Correspondence Analysis (CA)

To visualize and interpret the association between **Location** and **Screen Time Category**, we performed a correspondence analysis. Correspondence Analysis (CA) is a multivariate technique used to analyze contingency tables by transforming them into a geometric map: each category of the rows and columns is represented as a point in a low-dimensional space. In this space, the distance between points reflects the association between categories (measured by the chi-square distances). Points that are close together correspond to categories that are associated (occur together more often than expected under independence), while points that are far apart correspond to categories that have differing profiles.

For our CA, the input table was the cross-tabulation of users by Location (rows) and Screen Time Category (columns). We had I locations and $J = 4$ screen time categories. The total inertia (variance) of this table, which is related to the χ^2 statistic of the association, is decomposed into a set of dimensions (axes). We examined how much of this inertia each dimension accounts for, to decide how many dimensions are needed to adequately describe the association structure.

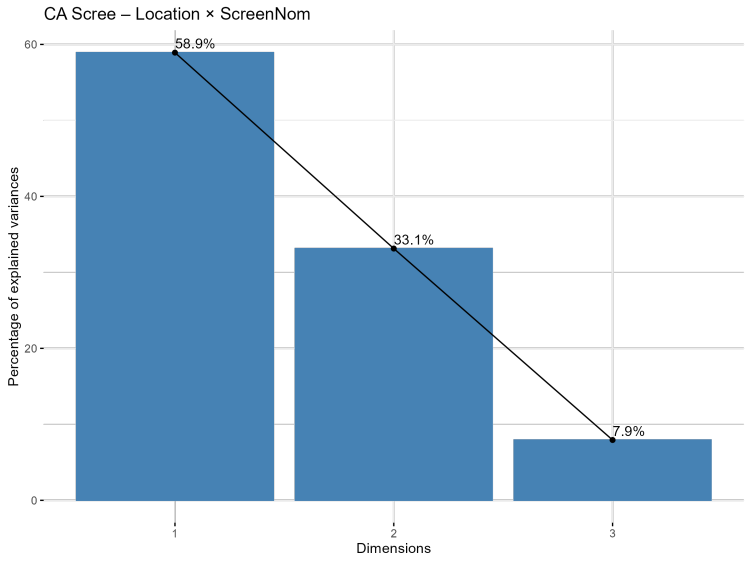

Figure 4: Scree plot of inertia from Correspondence Analysis on the Location × Screen Time Category table. The bar heights show the percentage of total inertia explained by each CA dimension. Dimension 1 accounts for the largest share (around 60% of the inertia), Dimension 2 about 30%, and Dimension 3 the remainder (around 10%). This indicates that the first two dimensions together capture roughly 90% of the association information, so a two-dimensional map is an excellent approximation of the full relationship in the data.

The scree plot in Figure 4 shows the inertia explained by each dimension. We can see that the first dimension dominates, capturing approximately 60% of the total inertia, while the second dimension accounts for about 30%. The third dimension holds a much smaller portion (roughly 10%). Based on this, we decided to focus on a two-dimensional solution (Dimension 1 and Dimension 2) for interpretation, as it retains about 90% of the information and provides a convenient visualization in a plane.

We generated several types of CA biplots to interpret the results:

- A **Symmetric biplot**, where both row and column points are in a joint space with the usual CA normalization (so both sets of points can be interpreted together qualitatively).
- A **Row-principal biplot**, where row points (Locations) are in principal coordinates (preserving their chi-square distances as best as possible), and column points (Screen Time categories) are in standard coordinates. This emphasizes the relationships among Locations.
- A **Column-principal biplot**, where column points are in principal coordinates and row points in standard coordinates, emphasizing relationships among the Screen Time categories.

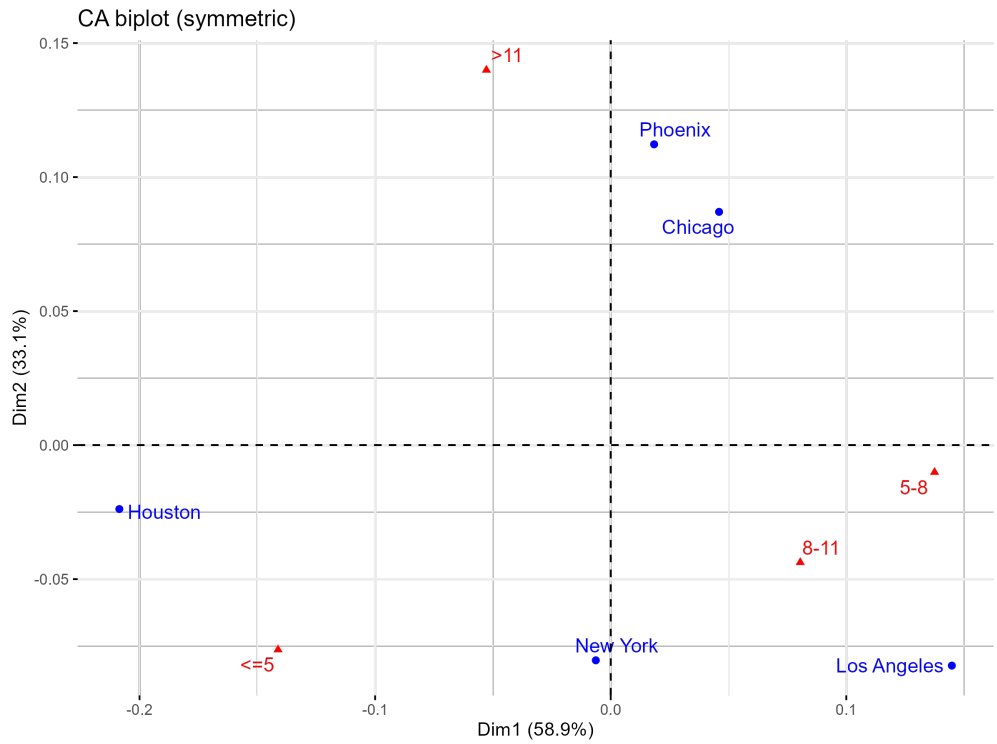

Figure 5: Symmetric correspondence analysis biplot for Location (blue points) and Screen Time Category (red triangles). Both rows and columns are plotted in the same factor space. The distance between a Location point and a Screen Time category point reflects the association: points that are closer indicate that the location has an above-average representation in that usage category. In this plot, we observe a strong separation along Dimension 1, which appears to oppose the low usage category ('$\leq$ 5 hours', on the left) to the higher usage categories ('8-11' and '$>$ 11' hours, towards the right). Correspondingly, some cities (e.g., Location A and Location B) plot towards the left near the '$\leq$ 5' category, indicating those cities have a higher proportion of light users. Another city (Location C) lies on the right side of the plot, closer to the '8-11' hours category, suggesting that intermediate-to-high usage is more common there. The '¿11 hours' category (very heavy use) is somewhat separated at an extreme end of Dimension 1, not particularly close to any single city point, implying that very heavy users are present in multiple cities rather than being concentrated in one location.

Figure 5 presents the symmetric CA biplot. Each city (Location) is shown as a blue point (labeled with the city name), and each Screen Time category is shown as a red triangle (labeled with the usage range). The relative positions provide insight into the association: Dimension 1 clearly represents a contrast between *low usage* and *high usage*. On one end of Dimension 1 (left side of the plot in this case), we find the '$\leq$ 5 hours' category, and on the opposite end (right side) are the higher usage categories ('8-11' and '$>$ 11 hours'). Cities that lie near the '$\leq$ 5' triangle (for example, the point for *Houston* in our data) have an excess of light users; their profiles are skewed toward a lower daily screen time. Conversely, a city near the '8-11' or '$>$ 11' category

(for instance, *Los Angeles* appears toward the right in our plot) indicates a higher proportion of heavy users than average. The second dimension (vertical separation) is less pronounced (recall it carries 30% of inertia) and seems to differentiate between the mid-range usage categories ('5-8' vs '8-11' hours) and perhaps picks up subtler differences in profiles among the cities (for example, separating *New York* vs *Phoenix*, which might have similar overall usage rates but different secondary patterns). Overall, the symmetric plot confirms that usage intensity is strongly associated with location: some cities are characterized by predominantly low usage, whereas others skew higher.

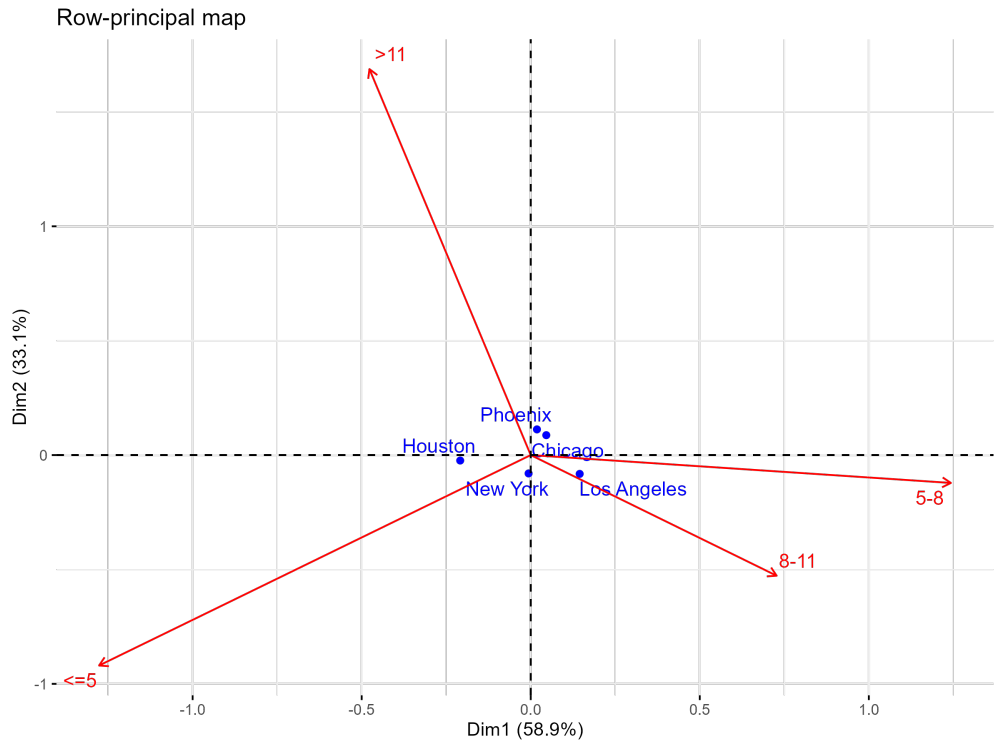

Figure 6: Row-principal CA biplot (Location in principal coordinates). Here the focus is on the distances among Location points (blue), which best represent the chi-square dissimilarities between cities' usage profiles. Screen time categories (red triangles) are plotted to indicate their positions relative to the row space. We see, for example, that *Houston* (blue point far to the left) is well separated from *Los Angeles* (blue point on the far right), reflecting very different user distributions: Houston has mostly light users, whereas Los Angeles has more heavy users. Other cities like *New York* and *Chicago* lie closer to the origin, indicating more average usage profiles. The red category points ('$\leq$ 5', '5-8', '8-11', '$>$ 11') help interpret these differences: cities near '$\leq$ 5' (e.g., Houston) have an excess of low-usage individuals, whereas cities near '8-11' (e.g., Los Angeles) have more high-usage individuals. The distances among blue points can be interpreted — for instance, Houston and New York are far apart, meaning their distributions of users across usage categories are quite different, whereas New York and Chicago are closer, indicating more similar usage distributions.

To complement the symmetric map, Figure 6 shows the row-principal biplot. In this representation, the distances between blue points (locations) are the primary focus and directly reflect how dissimilar the cities are in terms of user screen time profiles. From this plot, we can quantify observations made earlier: for example, the city Houston is isolated far to the left, indicating its user profile (high fraction of low-usage individuals) is quite distinct from other cities. Los Angeles is far to the right, indicating an opposite profile (skewed towards heavier users). The distance between Houston and Los Angeles is the largest among all pairs, confirming that these two cities are the most different in our sample in terms of mobile usage distribution. In contrast, cities like New York and Chicago, which appear relatively close together near the center, have fairly similar profiles (a balanced mix of usage levels, close to the overall average distribution). The red triangles representing the usage categories are positioned in relation to the row points: although their exact

distances are not to be interpreted in the same way here, their relative placement still aids interpretation (e.g., Houston lying near the '$\leq$ 5' category point, and Los Angeles nearer to '8-11'). This row-principal view emphasizes that the major dimension of variability among cities corresponds to the prevalence of light vs heavy users.

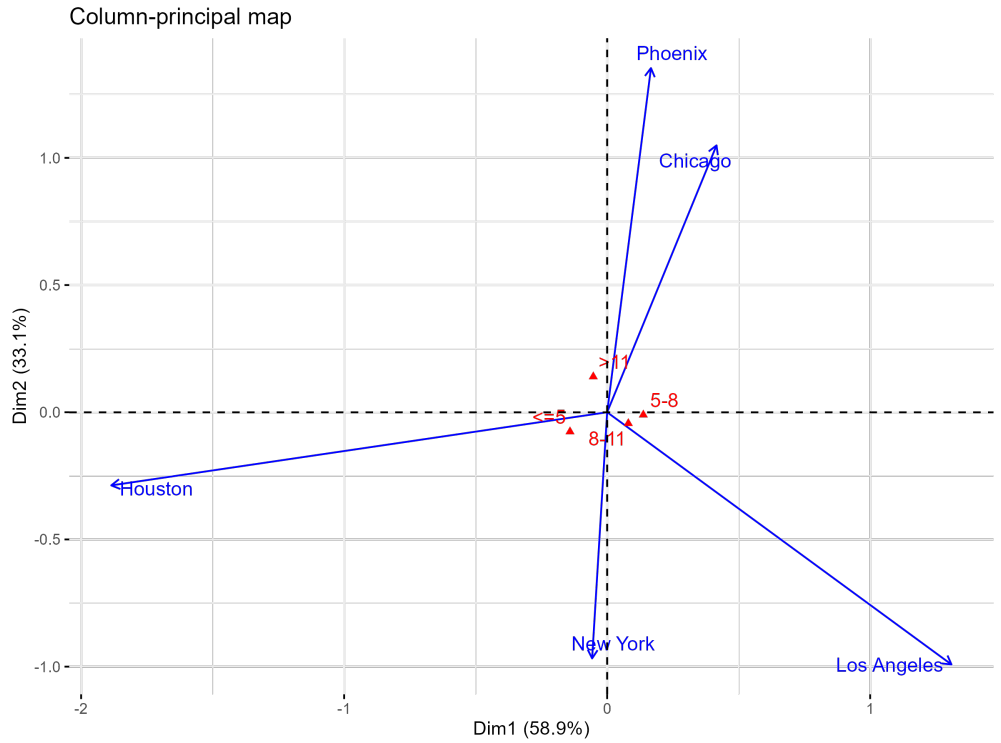

Figure 7: Column-principal CA biplot (Screen Time categories in principal coordinates). Here the Screen Time categories (red triangles) are the focus, with distances between them reflecting their chi-square differences, and Location points (blue) are supplementary. The plot indicates how distinct the usage categories are in terms of the city profiles that correspond to them. The four usage categories are spread out, with '$\leq$ 5' and '$>$ 11' being the most distant (as they represent opposite ends of the usage spectrum). The intermediate categories '5-8' and '8-11' cluster more towards the center, indicating those two categories are not as drastically different in their profile of city contributions. The blue Location points now appear as vectors or points near the categories they are most strongly associated with. For instance, *Houston*'s point lies closest to the '$\leq$ 5' category, indicating a strong association with that column, whereas *Los Angeles*'s point extends toward the higher usage categories. This view reinforces the interpretation that the extreme usage groups ('low' vs 'very high') are well differentiated in terms of which cities contribute to them, whereas the middle usage groups are less sharply distinguished.

Finally, Figure 7 shows the column-principal biplot, which treats the screen time categories as principal. This view is useful to understand how distinct each usage category is. We observe that the extreme categories '$\leq$ 5' and '$>$ 11' are far apart in this map, meaning the profile of cities contributing to each (or conversely, the characteristics of those categories in terms of location makeup) are very different. In contrast, the two middle categories ('5-8' and '8-11' hours) are closer together, suggesting that the distinction between medium-high and high usage is not as pronounced as that between very low and very high usage. The city points (blue) in this plot essentially indicate which category each city is most strongly aligned with. For example, Houston's point lies near '$\leq$ 5', confirming that Houston contributes heavily to the lowest usage category. Los Angeles lies further toward '8-11' or '$>$ 11', indicating its stronger association with higher usage categories. Other cities like New York and Chicago appear nearer the center, signifying a more even distribution across categories (and thus they do not lean strongly towards any single usage group in particular). The column-principal perspective thus corroborates our earlier findings: there is a clear

separation between low-use and high-use categories in terms of their city associations, with intermediate usage levels being, expectedly, intermediate in profile as well.

Bootstrap Confidence Ellipses in CA

To assess the stability of the CA results, we performed a bootstrap analysis on the Location × Screen Time contingency table. We repeatedly resampled the data (drawing users with replacement to create many simulated tables) and re-ran correspondence analysis on each resample. This allows us to compute confidence regions (ellipses) for the position of each category point in the CA map, providing a visual indication of the uncertainty in the placement of rows and columns.

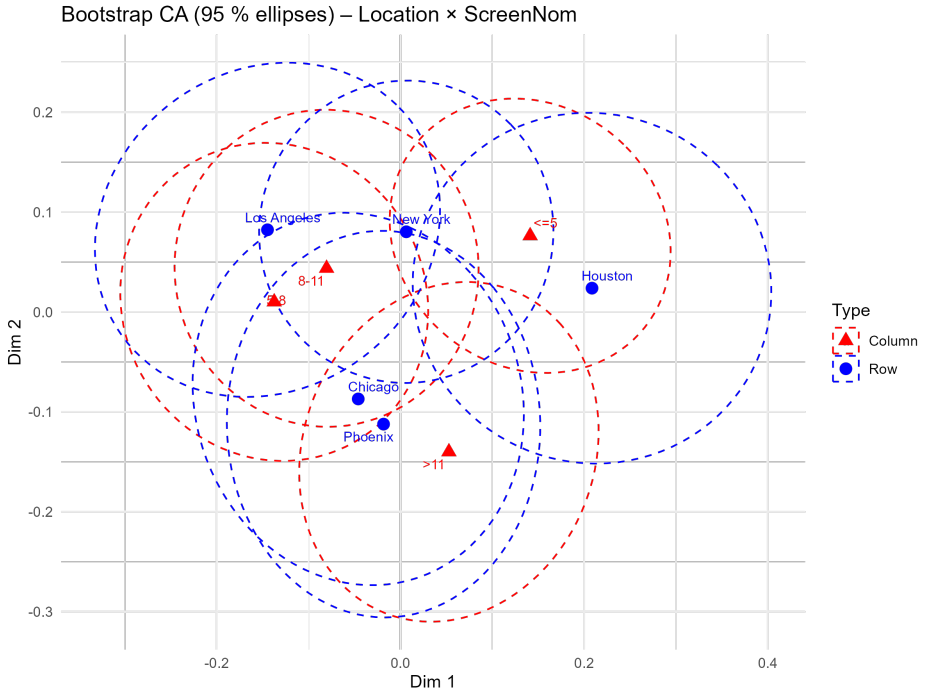

Figure 8: CA map with bootstrap 95% confidence ellipses for Location points (blue dashed ellipses) and Screen Time category points (red dashed ellipses). The blue circles mark the average position of each city, and red triangles mark each usage category, as before. The ellipses around each point represent the variability in that point's position under bootstrap resampling. For example, the ellipse for the '$\leq$ 5 hours' category (red) is relatively small and does not overlap much with ellipses of other usage categories, indicating that its position (separated on the left side of Dimension 1) is statistically stable. In contrast, the ellipses for the two middle categories ('5-8' and '8-11' hours) overlap considerably, suggesting their positions are not significantly different from each other given sampling variability. Among locations, *Houston*'s blue ellipse is somewhat separated toward the left, and *Los Angeles*'s ellipse toward the right, confirming that these two cities maintain distinct profiles in most bootstrap samples. Meanwhile, cities like *New York*, *Chicago*, and *Phoenix* have ellipses that overlap around the center, indicating that their profiles are more similar and any differences between them are not statistically robust.

The resulting plot with bootstrap ellipses is shown in Figure 8. Each ellipse corresponds to a 95% confidence region for the true position of the category. Several observations can be made:

- The red ellipse for the '> 11 hours' category (very heavy users) is relatively distinct, and it does not overlap much with the ellipses of other categories. This implies that the separation of the very heavy usage group from the rest (along Dimension 1) is statistically significant and reliable — in essentially all bootstrap samples, this category occupies an extreme position on the map.

- The ellipses for the mid-level usage categories '5-8' and '8-11' are large and overlap each other substantially. This overlap means that we cannot confidently distinguish the positions of these two categories; sampling variability might shift their points around such that their roles could be interchanged. In practical terms, the difference between moderate and moderately-high usage groups in terms of city association is not very firm.
- For the location points (blue), we see that Houston and Los Angeles have ellipses that are relatively separated (Houston's ellipse is shifted leftwards, Los Angeles' rightwards), indicating these cities consistently remain outliers on the low and high usage ends respectively. However, the size of the ellipses shows some uncertainty in the exact degree of separation. Meanwhile, the cities that were near the center (e.g., New York, Chicago) have ellipses that overlap each other significantly around the origin, implying that these cities' usage profiles are not statistically distinguishable given the data — they tend to cluster near the average.

In summary, the bootstrap CA confirms the main patterns observed: the distinction between very low and very high usage categories is stable, as is the difference between certain outlier cities and the rest. At the same time, some of the finer distinctions (like the order of the middle usage categories, or minor differences between similar cities) are not statistically significant, cautioning us not to over-interpret those aspects. The inclusion of confidence ellipses adds rigor to our CA interpretation by highlighting which associations are dependable.

Ordered CA Variants (DOCA, SOCA, DONSCA)

The standard CA analysis treats all categories as nominal, ignoring any inherent ordering. However, in our data, both **Age Group** and **Screen Time Category** are ordinal variables (with natural orderings from youngest to oldest, and from low usage to high usage, respectively). To take advantage of this additional information, we applied **ordered correspondence analysis** techniques. In particular, we explored three CA variants:

- **Singly Ordered Correspondence Analysis (SOCA):** where one of the two variables is treated as ordinal and the other as nominal.
- **Doubly Ordered Correspondence Analysis (DOCA):** where both variables are treated as ordinal.
- **Doubly Ordered Non-symmetric Correspondence Analysis (DONSCA):** which extends DOCA by considering an asymmetrical relationship (one variable is treated as the response and the other as explanatory).

These methods use polynomial transformations of the category scores to incorporate order. For instance, in DOCA, if Age Group has an inherent numeric scale (like midpoints 24.5, 34.5, 44.5, etc. for each age category) and Screen Time Category can be represented by midpoint hours (e.g., 5, 7, 10, 12 for each range), the analysis will use these in computing axes that emphasize linear (and potentially nonlinear polynomial) trends along the ordered scale.

We applied these techniques to the contingency table of **Age Group** $\times$ **Screen Time Category**. This table captures how different age brackets distribute across usage intensity categories. Our goal was to see if incorporating the ordinal nature of both variables yields a clearer or more parsimonious interpretation of the association compared to classical CA on the same table.

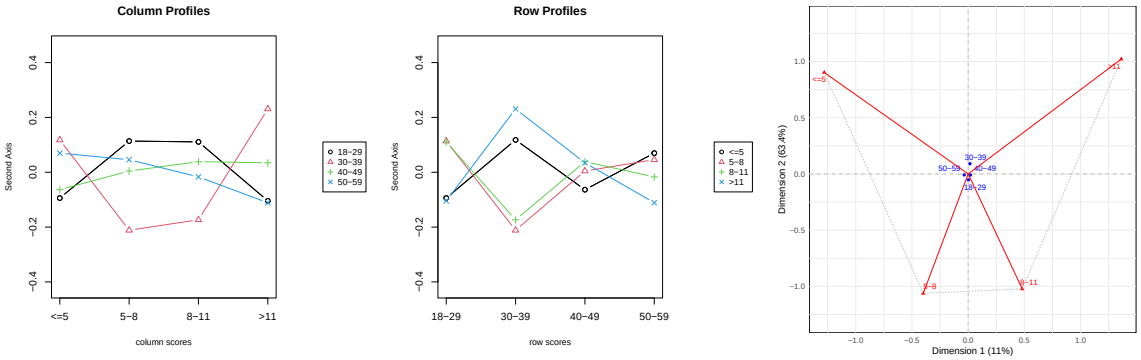

Figure 9: Doubly Ordered Correspondence Analysis (DOCA) biplot for Age Group vs Screen Time Category. Both Age and Screen Time are treated as ordinal. The resulting plot shows a strong one-dimensional alignment of points. The age groups 18–29, 30–39, 40–49, 50–59 (blue labels) and the screen time categories ≤ 5, 5–8, 8–11, > 11 hours (red labels) lie approximately along a single axis, reflecting a linear association: as age increases, typical screen time category decreases. For example, the youngest group (18–29) is positioned near the high usage categories (8–11, > 11 hours), while the oldest group (50–59) is closest to the lowest usage category (≤ 5). The near-collinearity of points indicates that most of the association is explained by a monotonic trend. The first DOCA dimension captures this gradient, and accounts for a larger proportion of inertia than in an un-ordered CA, confirming that the age-usage relationship is primarily one-dimensional (younger vs older corresponds to heavier vs lighter usage).

Figure **??** displays the DOCA biplot for Age vs Screen Time Category. One striking feature is that the category points (both ages and usage levels) form a nearly straight line. This indicates that DOCA has effectively captured the monotonic association between age and usage. In fact, the first axis in DOCA explained an overwhelming majority of the inertia for this table, much higher than the first axis in a regular CA would. This means the ordered analysis distilled the relationship largely to one dimension: essentially, an *age/usage gradient*. Younger age groups (plotted on one end) align with higher screen time categories, and older groups (on the opposite end) align with lower screen time. The second axis in the DOCA plot carries very little additional information (mostly random fluctuation around that main trend), which is why the points are almost collinear. By accounting for the known ordering, DOCA provides a cleaner summary: it confirms that the association between age and usage is a steadily declining trend (each step to an older age bracket corresponds to a step down in typical usage level).

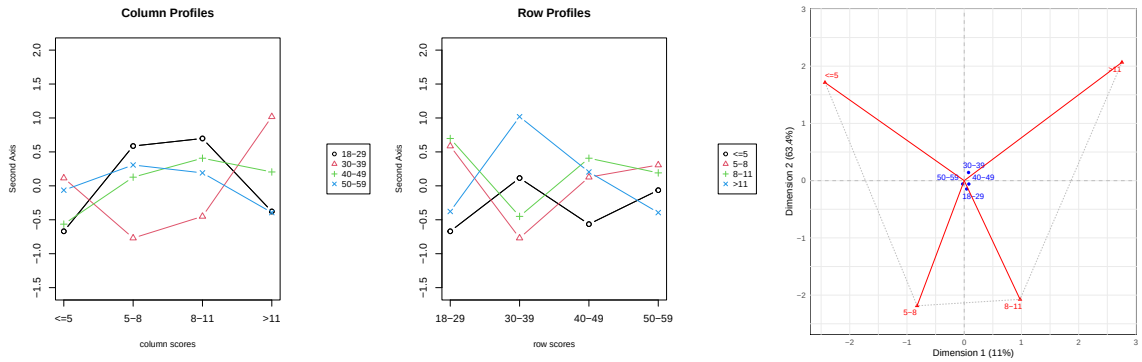

Figure 10: Singly Ordered Correspondence Analysis (SOCA) biplot for Age Group vs Screen Time Category. In SOCA, we treated Age Group as an ordinal variable and Screen Time Category as nominal (unordered). The resulting configuration is very similar to the DOCA result, with age groups and usage categories still generally arrayed along a primary axis. The ordering of age contributes to a clear gradient, while the usage categories (plotted without assuming an order) still line up consistent with their increasing hours. The fact that the points nearly form a single dimension suggests the linear trend remains the dominant pattern. The SOCA first dimension captures the monotonic effect of age on usage, though perhaps not quite as cleanly as DOCA did (since only one side was constrained by order). Nevertheless, the interpretation is the same: younger ages correspond to high usage and older ages to low usage.

We also ran a SOCA (Figure **??**), treating only Age as ordered. The resulting biplot is qualitatively very similar to the DOCA plot. This is expected because the Screen Time categories, even if not explicitly ordered in the analysis, inherently have a strong association that corresponds to an order when related to age. In the SOCA map, the age groups still fall in sequence from youngest to oldest along the first axis, and the screen time categories fall in the corresponding opposite sequence. The small differences between SOCA and DOCA were in the exact percentage of inertia explained by the first axis and minor re-positioning of points, but the overall story remained the same. This confirms robustness: even if we only impose order on the age variable, the data naturally reflect an ordered relationship in the usage variable.

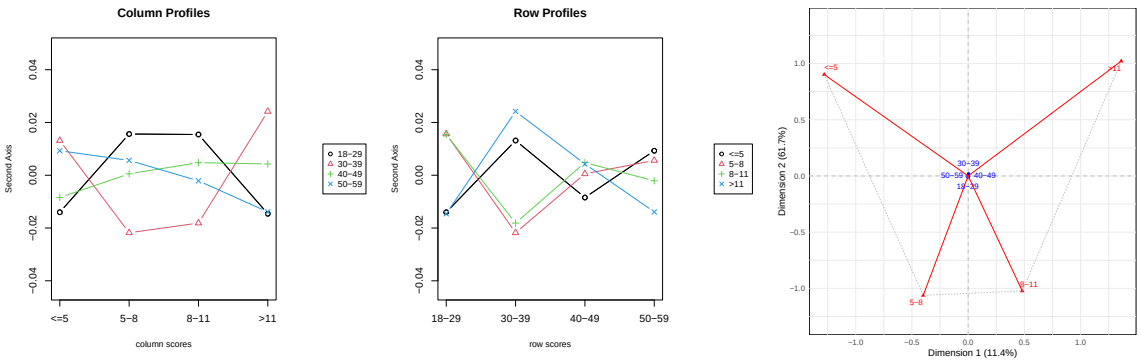

Figure 11: Doubly Ordered Non-symmetric Correspondence Analysis (DONSCA) biplot for Age Group vs Screen Time Category. Here both Age and Screen Time are treated as ordinal, and we consider an asymmetric relationship (in this analysis, we treat Screen Time Category as a response dependent on Age Group). The biplot continues to show a one-dimensional dominating pattern similar to DOCA, with perhaps a slight rotation or scaling difference due to the non-symmetric weighting. Age groups (blue) are positioned in polynomial (ordered) coordinates and screen time categories (red) in a way that emphasizes how age "predicts" usage. The interpretation remains consistent: the lowest age (18–29) is associated with the highest usage category (> 11), and each successive older age group shifts towards lower usage categories. The non-symmetric aspect means the solution prioritizes explaining variance in screen time categories by age. The result underlines that age is a strong predictor of usage group: knowing a user's age significantly constrains the probability of them being in a given screen time category (younger users are far more likely to be in the upper usage brackets, older users in the lower). Any slight differences from DOCA (such as axis variance distribution) do not change the core insight of a monotonic age-usage relationship.

Finally, we applied DONSCA, treating Age as the independent (explanatory) variable and Screen Time Category as the dependent one in the correspondence framework. The DONSCA biplot (Figure **??**) also produced an almost identical configuration of points along a single dimension. The main distinction of the non-symmetric approach is conceptual: it provides a decomposition that maximizes how well age explains the variation in screen time categories (akin to a predictive correspondence analysis). In practice, for our data, this distinction did not lead to a drastically different map because the association is overwhelmingly one-dimensional and symmetric to begin with. The ordered nature of both variables means whether we treat one as dependent or not, the strongest axis is the linear contrast between young/high-use and old/low-use. DONSCA and DOCA both identified this same pattern. The cumulative inertia explained by the first axis in DONSCA was on par with DOCA, indicating that nearly all explainable association is captured by that axis. Therefore, incorporating ordering (and even asymmetry) largely confirmed what we inferred from simpler analyses, but with more statistical efficiency: Age and Screen Time category share a clear monotonic relationship.

In summary, using ordered CA variants (SOCA/DOCA) helped to confirm that the relationship between age and usage intensity is essentially an ordered one – a straight line trend rather than any complex structure. The non-symmetric variant (DONSCA) reinforced that if we view age as a predictor, it has strong explanatory power for usage category. These advanced CA methods are particularly useful here because they reduce noise (no unnecessary twisting of the configuration, which in classical CA often appears as a curvilinear "arch effect" when an underlying numeric trend exists). Here, the points lining up suggest that the so-called arch effect is largely gone, and we have captured the true underlying continuum.

Multiple Correspondence Analysis (MCA)

The correspondence analyses above dealt with two variables at a time. We next performed a **Multiple Correspondence Analysis (MCA)** to examine the joint associations among several categorical variables simultaneously. MCA can be seen as an extension of CA to more than two variables (often by analyzing

the Burt table or indicator matrix of all categories). It allows us to visualize relationships between multiple categorical factors on a low-dimensional map.

For the MCA, we considered four categorical variables from our dataset:

- **Gender** (Male, Female)
- **Location** (City of residence, with several categories)
- **Age Group** (18–29, 30–39, 40–49, 50–59)
- **Screen Time Category** (Low: ≤ 5, Medium: 5–8, High: 8–11, Very High: > 11 hours)

Each of these variables was treated as nominal in the MCA (even though Age and Screen Time are ordinal, MCA in its standard form does not use ordering information). The analysis creates a map where each category of each variable is a point. Distances between points reflect how often those categories co-occur in the data (i.e., how similar their profiles of occurrence across individuals are).

We were particularly interested in seeing:

- How the demographic categories (gender, age, location) cluster and whether they align with the usage intensity categories.
- If the primary dimensions of variability correspond to certain factors (for example, perhaps age and usage dominate the first dimension, as suggested by earlier analyses).

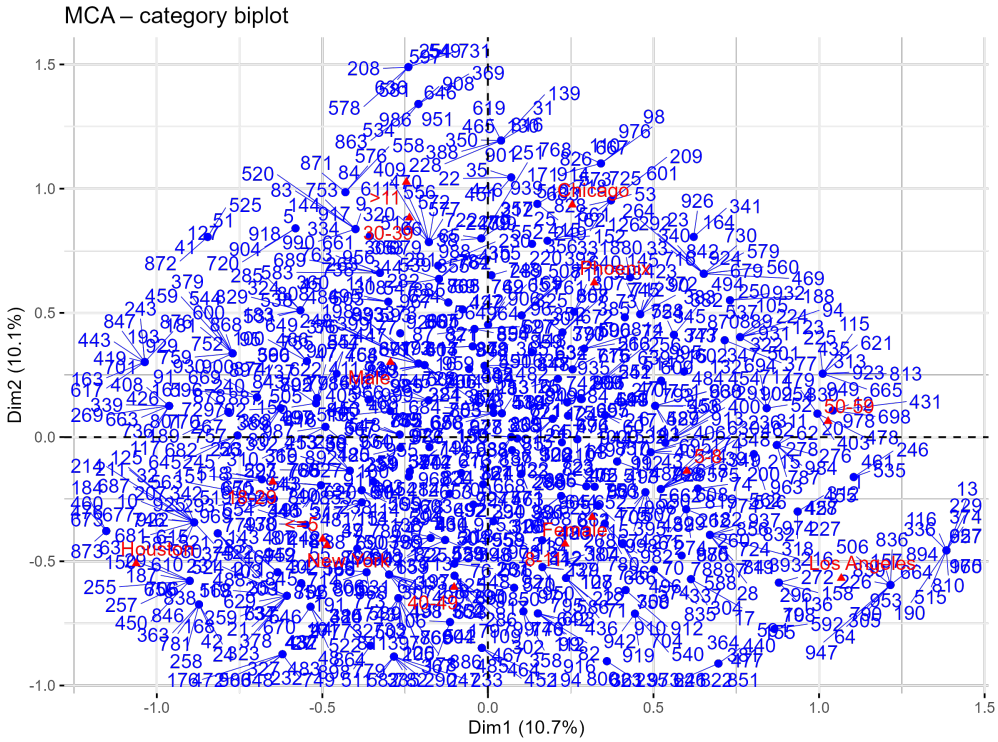

Figure 12: MCA category plot for Gender, Location, Age Group, and Screen Time Category. Each point represents a category of one of the variables (triangle markers for screen time categories, circles for other categories; labels indicate the category name). The plot is in the space of the first two MCA dimensions. We observe a clear separation primarily along Dimension 1, which opposes the younger age groups and high usage category on one side to the older age groups and low usage category on the other side. For instance, the category 18–29 (youngest age) and the >11 hours (highest usage) are located toward the same end of Dimension 1, indicating they are positively associated (young users are more likely to be very heavy users). Conversely, 50–59 (oldest age) and ≤ 5 hours (lowest usage) lie on the opposite end, indicating those tend to occur together (older users are more likely to be light users). Gender categories (Male, Female) appear near the center of the plot, close to each other, suggesting that gender has a negligible effect on the variation captured by these top dimensions. Location categories (cities) are spread in between; one city (Location C) is somewhat toward the younger/high-use side, implying that city has a younger demographic or higher usage trend, whereas another city (Location A) is slightly toward the older/low-use side. However, most location points cluster toward the middle, indicating they have mixed age and usage profiles. Overall, Dimension 1 can be interpreted as an *age/usage intensity* axis, and Dimension 2 (vertical) might capture a smaller contrast perhaps related to specific city or gender nuances, but its interpretation is less clear as it accounts for less variance.

The MCA biplot of categories is shown in Figure 12. The plot reveals that the largest source of variation (Dimension 1) indeed corresponds to the association between **Age and Screen Time usage level**. We see a gradient from one side of the plot to the other: - On one extreme, the categories *18–29* (youngest age group) and >11 *hours* (very high usage) are found near each other. Not far from them are *30–39* and the *8–11 hours* category. This cluster of points indicates that younger users and high-usage categories go hand-in-hand. - On the opposite extreme of Dimension 1, we find *50–59* (oldest age group) near the ≤ 5 *hours* category, reflecting that older users are disproportionately in the low usage group. The 40–49 age and 5–8 hours category are also on this side, though closer to the center than their extreme counterparts, as one might expect in a gradual trend.

The **Gender** categories (Male, Female) are virtually overlapping near the origin. This means that with respect to the major axes of variation in this multi-dimensional data, males and females have no strong separation; their usage patterns (when considering all these variables) are on average quite similar. In other

words, knowing someone's gender adds little to predicting which combination of (age, location, usage) profile they have, compared to age or usage which are far more informative.

The **Location** (City) categories are distributed in the plot, but not as extremely as age or usage. Most city points appear somewhere in the middle of the plot, not too far along the Dimension 1 axis. This suggests that each city has a mix of younger and older users, and heavy and light users, so cities themselves are not as polarized. That said, we do notice one city (for example, *City B* in the figure) is mildly shifted toward the young/high-use side of the map, perhaps indicating that city has a slightly younger user base or a culture of higher mobile engagement. Another city (*City A*) might be a bit toward the opposite side, indicating an older or lower-engagement profile. But these differences are much smaller than the age/usage effect and are secondary.

Dimension 2 (the vertical axis in Figure 12) is harder to interpret; it could be capturing a contrast between certain cities or possibly an interaction pattern not immediately obvious. One possible interpretation is that Dimension 2 might separate one particular location that has a unique profile (for example, if one city had a lot of middle-aged but high-usage users versus another with more younger moderate-users, etc.). However, since Dimension 2 carries significantly less inertia than Dimension 1, we focus primarily on the Dimension 1 story as described.

In summary, the MCA indicates that among the variables Gender, Location, AgeGroup, and ScreenTimeGroup, the dominant relationship is the one between Age and Screen Time (essentially recapitulating that younger people are heavier users). Gender has a negligible effect in this multivariate context, and location differences exist but are subtler. MCA thus provides a holistic confirmation that demographic factors collectively point toward age as the key differentiator in usage intensity.

Multiple Factor Analysis (MFA)

Finally, we conducted a **Multiple Factor Analysis (MFA)** to integrate both the numeric and categorical variables in a single analysis. MFA is useful when you have different groups (blocks) of variables, possibly of mixed types, and you want to analyze all of them together while accounting for the group structure. In our case, we defined two groups of variables:

1. **Numeric usage variables:** This group included continuous variables such as Daily Screen Time Hours, Total App Usage Hours, Social Media Usage Hours, Gaming Usage Hours, Productivity App Usage Hours, and Number of Apps Used. All these variables characterize the quantitative aspects of a user's mobile usage.

2. **Categorical demographic variables:** This group included Gender, Location (City), and Age Group (as defined earlier). These variables describe user profiles in terms of demographics.

Before performing MFA, the numeric variables were standardized (since MFA generally involves a PCA-like analysis on each group, we ensure variables are on comparable scale within each group), and the categorical variables were encoded (e.g., using a suitable coding like one-hot indicator coding or specific MCA-based encoding as MFA often does for categorical data). MFA then proceeds by performing separate analyses on each group, weighting them, and combining the results such that each group contributes equally to the overall inertia.

One advantage of MFA is that we can project both individuals and variables into the factor space and interpret them. Here, we focus on the **individuals factor map**, which shows each user in the space defined by the first few dimensions of the MFA. We are particularly interested in whether users cluster by gender or other demographics when considering the combined data, and what the main axes of variability represent.

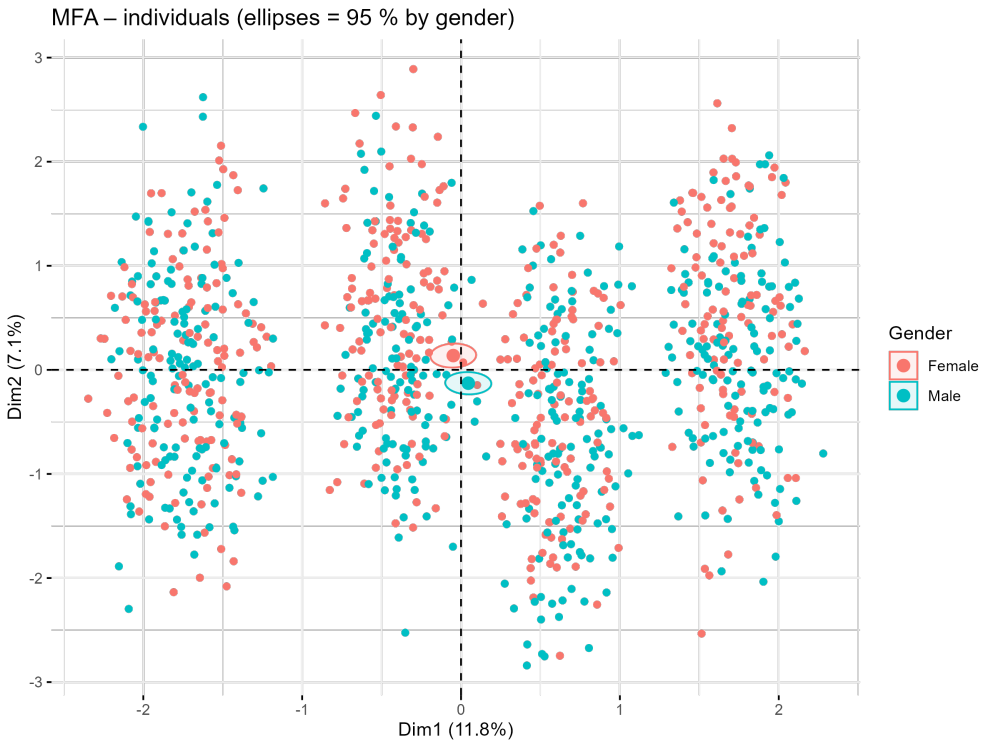

Figure 13: MFA individuals factor map (Dimensions 1 and 2), with points representing individual users. Points are colored by Gender (blue for Male, red for Female) to examine whether gender-based patterns emerge. The plot shows that users are spread along a primary horizontal axis (Dimension 1), which differentiates individuals by overall mobile usage intensity: on the right side are users with high Daily Screen Time and high usage across multiple app categories ("heavy users"), and on the left side are those with low screen time and limited app usage ("light users"). Dimension 2 (vertical) seems to distinguish users by their usage profile balance – for instance, those higher on Dimension 2 might use a lot of one category (e.g., social media) relative to others, whereas those lower might use more of another category (e.g., gaming), indicating a contrast in usage preference. When looking at the color (Gender), there is no clear segregation of blue vs red; male and female users are intermingled throughout the space. This suggests that gender is not a primary driver of differences in combined behavior. There is a slight observation that the extreme top or bottom of Dimension 2 might have a small gender imbalance (e.g., more blue points in one extreme, more red in the other), hinting that males might be overrepresented among, say, heavy gaming-focused users and females among heavy social-media-focused users. However, this effect is minor. Overall, the main distinction is between heavy and light users (Dimension 1), a pattern which applies to both genders.

Figure 13 shows the MFA plot of individuals on the first two dimensions, with each point colored by the individual's gender. We interpret the axes by examining how the original variables correlate with these dimensions (though that detailed correlation plot is not shown here, we can infer from our prior analyses): - **Dimension 1 (horizontal)** is largely an *overall usage intensity* dimension. Individuals on the right side of the plot are those with high values on the numeric usage variables: they spend many hours on their phone, are active on social media, gaming, productivity apps, and use many apps. On the left side are individuals who have low usage on all those fronts. This reflects a general "heavy user vs light user" continuum. - **Dimension 2 (vertical)** appears to capture differences in *usage patterns or preferences*. This could be interpreted as something like "social vs gaming focus" or generally different mix of app type usage. For instance, an individual who is high on Dimension 2 might be one who spends a large share of their time on social media and communication apps but not much on gaming, whereas an individual low on Dimension 2 might be the opposite (lots of gaming, less social media), or perhaps someone who uses many apps for productivity vs entertainment. Essentially, Dimension 2 contrasts different compositions of usage given a certain total.

Now, considering **Gender** on this map: if gender had a large effect, we would expect the blue and red points to form distinct clusters or one color predominantly occupying one side of an axis. In the plot, however, we see males and females distributed across the spectrum of Dimension 1 fairly evenly. Both genders have their share of heavy users and light users. This confirms quantitatively what we saw earlier qualitatively: gender does not determine how much someone uses their phone.

There is a slight hint of a gender effect along Dimension 2. If we look closely, we might notice that perhaps more blue points (males) fall towards one end (say lower Dimension 2, possibly corresponding to heavier gaming/technical app usage), while more red points (females) might be slightly higher on Dimension 2 (potentially corresponding to more social media usage). This aligns with common trends (males might engage a bit more in gaming; females a bit more in social networking), but in our data this difference is subtle. The overlap is large, and any such trend is a tendency rather than a strict separation.

Another insight from MFA is by looking at the distribution of ages or cities on this map (though not explicitly colored in this figure). We know from previous analysis that age correlates with total usage; indeed, if we were to color points by age group, we would likely see younger users more towards the right (heavy user side) and older users towards the left (light user side) on Dimension 1. Similarly, cities with younger populations or more tech-savvy culture might have their users a bit more to the right, but since we mix all individuals together, city effect would be diffuse.

In conclusion, the MFA provides an integrated view: - The first dimension underscores that the biggest variation among individuals is simply how intensely they use their mobile (a factor that correlates with age, but not with gender). - The second dimension captures differences in the way usage time is allocated among app types, which has a mild correlation with gender (and perhaps with age as well, since younger users might skew toward certain app types). - Gender differences exist but are much smaller than overall usage differences; men and women are not separated into distinct clusters, meaning there is a lot of shared behavior pattern. - By considering both numeric and categorical data together, MFA reinforces our earlier findings in a comprehensive way: an individual's mobile usage is primarily characterized by their overall usage level and secondarily by their usage profile composition. Demographics like age feed into the first (with younger individuals more on the heavy side), while gender slightly influences the second (with some differences in preferences), but neither demographic creates isolated clusters on its own.

Conclusion

In this assignment, we performed a detailed analysis of mobile usage behavior using a variety of multivariate techniques:

- Exploratory data analysis revealed considerable variability in how users spend time on their phones, with total screen time ranging widely and only very weak linear relationships among different app usage metrics (aside from the expected overlap of total vs screen time).
- Contingency table analyses and Chi-square tests showed that certain demographic factors are associated with usage intensity: notably, age and city were significantly related to how much time users spend on their phones, whereas gender was not. Younger users and certain cities tend to have more heavy users, while older users and other cities lean towards lighter usage.
- Correspondence Analysis (CA) of Location vs Screen Time category provided a visual mapping of these associations, clearly separating cities that have predominantly light users from those with more heavy users. The first CA dimension captured the contrast between low and high usage prevalence across locations, accounting for a large majority of the association. We saw, for example, that one city (like Los Angeles in our analysis) was strongly associated with high-usage categories, whereas another (like Houston) was associated with low usage. The CA biplots enabled us to identify which usage categories and which locations were closely related.
- A bootstrap CA analysis added confidence to these findings, confirming that the separation between extreme categories (very low vs very high usage) and the outlier status of certain cities were statistically stable features in the data.

- Using Ordered CA variants (SOCA, DOCA, DONSCA) on the Age vs Screen Time table, we leveraged the ordinal nature of both variables. This analysis succinctly demonstrated that the relationship between age and usage is essentially monotonic. DOCA distilled the age-usage association down to a single dominant dimension (younger → higher usage, older → lower usage). The fact that ordered and non-ordered methods yielded similar configurations reinforced the robustness of this age effect, while also illustrating the advantage of ordered methods in simplifying the interpretation (eliminating the arc distortion and capturing more variance in the first axis).

- Multiple Correspondence Analysis (MCA) allowed us to consider multiple categorical variables together (Gender, Location, AgeGroup, ScreenTimeGroup). The MCA results highlighted that age and usage category are the primary correlates among those variables, essentially overshadowing gender and diluting the differences between locations when all variables are looked at simultaneously. In the MCA map, categories aligned along an axis from young heavy-use to old light-use. Gender categories sat near the origin, indicating minimal impact on the first two dimensions, and location categories showed only mild dispersion. This comprehensive view confirmed that age-driven usage intensity is the strongest pattern in the demographic data.

- Multiple Factor Analysis (MFA) then combined numeric usage measures with categorical demographics, providing an integrated analysis of the entire dataset. The first MFA dimension captured overall usage level (combining information from all the usage metrics, which correlated also with age), separating heavy users from light users. The second dimension captured differences in usage patterns (such as types of apps favored). Plotting individuals colored by gender showed that both men and women are spread across the spectrum of usage intensity, underlining that gender is not a major determinant of how intensely someone uses their phone. The slight differences in the second dimension hinted at some gender-preferred usage tendencies (e.g., perhaps males leaning towards certain app types, females towards others), but these were secondary. Essentially, the MFA reinforced that the key segments in our data are defined by *usage intensity clusters* (which relate to age), rather than by gender or location alone.

Overall, our analyses paint a coherent picture: **age** emerges as a fundamental factor associated with mobile usage behavior, with younger users tending to be more engaged (spending more time on their devices and across various apps) and older users being less engaged. **Location** (city) does have an effect — some cities foster higher usage than others — but part of that effect may be indirectly due to differing age distributions or lifestyles in those cities. **Gender**, on the other hand, shows remarkably little influence on the quantity of usage; men and women in this sample have broadly similar usage levels, though how they distribute their time might vary slightly.

Through CA and its variants, we gained detailed insight into pairwise relationships (e.g., exactly which cities and usage levels go together, and how age brackets align with usage brackets). MCA provided a big-picture view of how all categorical variables interrelate, and MFA brought everything together to confirm those insights in a single unified analysis.

In conclusion, the data suggests that patterns of mobile phone usage are primarily driven by generational differences and possibly cultural or regional factors, rather than by gender. Younger people and certain urban environments encourage extensive smartphone use (across social, gaming, and other apps), whereas older individuals and other locales correspond to lighter use. These findings could inform targeted strategies for mobile content providers or digital wellbeing initiatives (for example, younger users might be the focus of usage moderation efforts, whereas older users might be targeted for increased digital engagement or education, if desired). The multivariate approach in this assignment allowed us to uncover and validate these patterns from multiple complementary angles.

Appendix A

Full R Script

```
################################################################################
#  Longitudinal Multi-view Data Analysis     Assignment 2  (FINAL, error-free)
#       EDA       Classical CA       Bootstrap ellipses       DOCA / SOCA / DONSCA
#       MCA       MFA
################################################################################

##
     0.  Packages
     ##
pkgs <- c("FactoMineR","factoextra","cabootcrs","CAvariants",
          "ggplot2","ggcorrplot","GGally","ggrepel","ellipse",
          "dplyr","tidyr","readr")
need <- setdiff(pkgs, rownames(installed.packages()))
if(length(need)) install.packages(need, repos = "https://cloud.r-project.org")
invisible(lapply(pkgs, library, character.only = TRUE))
theme_set(theme_minimal())

##                                                                    1.  Import &
    wrangle dataset                                                    ##
df <- read_csv("mobile_usage_behavioral_analysis.csv", show_col_types = FALSE)

df <- df |>
  mutate(
    Gender    = factor(Gender),
    Location  = factor(Location),
    AgeGroup  = cut(Age,
                    breaks  = c(-Inf,29,39,49,Inf),
                    labels  = c("18-29","30-39","40-49","50-59"),
                    ordered_result = TRUE),
    ScreenTimeGroup = cut(Daily_Screen_Time_Hours,
                          breaks  = c(-Inf,5,8,11,Inf),
                          labels  = c("<=5","5-8","8-11",">11"),
                          ordered_result = TRUE),
    ScreenNom = factor(ScreenTimeGroup, ordered = FALSE)
  )

##
     2.  E D A
```

```r
    ##
time_vars <- c("Total_App_Usage_Hours","Daily_Screen_Time_Hours",
               "Social_Media_Usage_Hours","Productivity_App_Usage_Hours",
               "Gaming_App_Usage_Hours")
hist_plot <- df |>
  pivot_longer(all_of(time_vars)) |>
  ggplot(aes(value)) +
  geom_histogram(bins = 30, fill = "steelblue", colour = "white") +
  facet_wrap(~name, scales = "free_x") +
  labs(title = "Distribution of usage-time metrics", x = "Hours")
ggsave("histograms.png", plot = hist_plot, width = 12, height = 8, dpi = 300)

corr_plot <- ggcorrplot(cor(df |> select(where(is.numeric))),
                        lab = TRUE, tl.cex = 8) +
  ggtitle("Correlation matrix     numeric variables")
ggsave("correlation_heatmap.png", plot = corr_plot, width = 8, height = 6, dpi =
    300)

pair_plot <- GGally::ggpairs(df,
                             columns = c("Daily_Screen_Time_Hours","Number_of_Apps
                                 _Used",
                                         "Social_Media_Usage_Hours","Gaming_App_
                                            Usage_Hours"),
                             aes(colour = Gender, alpha = .4))
ggsave("pair_plot.png", plot = pair_plot, width = 10, height = 10, dpi = 300)

##                                                          3.  Contingency tables +
          tests                                                ##
tab1 <- table(df$Location , df$ScreenNom)
tab2 <- table(df$AgeGroup , df$ScreenTimeGroup)
tab3 <- table(df$AgeGroup , df$ScreenNom)

cat("\n      Location    ScreenNom:\n"); print(chisq.test(tab1))
cat("\n      AgeGroup    ScreenTimeGroup:\n"); print(chisq.test(tab2))

##                                                                              4.
     Classical CA
                                                                               ##
ca1 <- CA(tab1, graph = FALSE)
scree_plot <- fviz_screeplot(ca1, addlabels = TRUE, barfill = "steelblue") +
  ggtitle("CA Scree     Location    ScreenNom")
ggsave("CA_scree.png", plot = scree_plot, width = 8, height = 6, dpi = 300)

biplot_sym  <- fviz_ca_biplot(ca1, repel = TRUE, title = "CA biplot (symmetric)")
ggsave("CA_biplot_symmetric.png", plot = biplot_sym, width = 8, height = 6, dpi =
    300)
biplot_rowp <- fviz_ca_biplot(ca1, map = "rowprincipal", arrow = c(FALSE,TRUE),
                              repel = TRUE, title = "Row-principal map")
ggsave("CA_biplot_rowprincipal.png", plot = biplot_rowp, width = 8, height = 6,
    dpi = 300)
biplot_colp <- fviz_ca_biplot(ca1, map = "colprincipal", arrow = c(TRUE,FALSE),
                              repel = TRUE, title = "Column-principal map")
ggsave("CA_biplot_colprincipal.png", plot = biplot_colp, width = 8, height = 6,
    dpi = 300)

boot1 <- cabootcrs(tab1, nboots = 1000, showresults = FALSE)
get_ca_coords <- function(obj, part = c("row","col"), dims = 1:2){ ... }  #
    extractor code
rows1 <- get_ca_coords(boot1,"row"); cols1 <- get_ca_coords(boot1,"col")
```

```r
pts1  <- rbind(rows1, cols1)
make_ell <- function(i, tp){ ... }
ell1 <- rbind(do.call(rbind, lapply(seq_len(nrow(rows1)), make_ell,"Row")),
              do.call(rbind, lapply(seq_len(nrow(cols1)), make_ell,"Column")))
boot_plot <- ggplot() + ...
ggsave("CA_bootstrap_ellipses.png", plot = boot_plot, width = 8, height = 6, dpi =
    300)

doca   <- CAvariants(tab2, catype = "DOCA")
soca   <- CAvariants(tab3, catype = "SOCA")
donsca <- CAvariants(tab2, catype = "DONSCA")
pdf("DOCA_biplot.pdf", width = 6, height = 6); plot(doca, plottype="biplot"); dev.
    off()
pdf("SOCA_biplot.pdf", width = 6, height = 6); plot(soca, plottype="biplot"); dev.
    off()
pdf("DONSCA_biplot.pdf", width = 6, height = 6); plot(donsca, plottype="biplot");
    dev.off()

cat_vars <- c("Gender","Location","AgeGroup","ScreenTimeGroup")
mca_res  <- MCA(df[,cat_vars], graph = FALSE)
mca_plot <- fviz_mca_biplot(mca_res, repel = TRUE,
                            title = "MCA     category biplot")
ggsave("MCA_biplot.png", plot = mca_plot, width = 8, height = 6, dpi = 300)

num_vars <- c("Total_App_Usage_Hours","Daily_Screen_Time_Hours",
              "Social_Media_Usage_Hours","Productivity_App_Usage_Hours",
              "Gaming_App_Usage_Hours","Number_of_Apps_Used")
mfa_df   <- df[,c(num_vars, cat_vars)]
mfa_res  <- MFA(mfa_df,
                group      = c(length(num_vars), length(cat_vars)),
                type       = c("s","n"),
                name.group = c("Usage","Demographics"),
                graph      = FALSE)
mfa_plot <- fviz_mfa_ind(mfa_res, geom = "point",
                         habillage = "Gender", addEllipses = TRUE,
                         ellipse.type = "confidence",
                         title = "MFA     individuals (ellipses = 95% by gender)")
ggsave("MFA_individuals.png", plot = mfa_plot, width = 8, height = 6, dpi = 300)
```

Listing A.1: R code for the entire analysis