Multivariate Segmentation: Integrating Clustering and PCA for Latent Structure Discovery

---

ADEI Q2 24/25

FIB

Gerard Godet
Èric Díez
Ivan Rodríguez
Adrián Patiño

Index

1. CLUSTERING

1.1. INTRODUCTION

In this study, we applied hierarchical clustering to segment individuals into distinct groups using sociodemographic and economic variables.

1.1.1 Distance Metric

We used **Gower's Distance** to handle our dataset's mix of numerical variables (e.g., age, income) and categorical variables (e.g., occupation, marital status).

- **For numerical variables**: Computes scaled differences (e.g., how far apart two ages are, relative to the dataset's range).
- **For categorical variables**: Checks for mismatches (e.g., "Married" vs. "Single" counts as a difference).

This provides a unified dissimilarity score, ensuring fair comparisons across variable types.

1.1.2 Aggregation Criteria

We applied **Ward's Minimum Variance Method** to build the hierarchy to merge clusters in a way that minimizes the increase in total within-cluster variation.

At each step, it combines the two clusters that result in the smallest growth in overall dispersion.

1.1.3 Number of clusters

The number of clusters was determined by analyzing the dendrogram, a tree-like diagram that illustrates data point groupings at different levels of similarity. The dendrogram showed a clear natural split into three clusters, where:

- The distance between branches increased significantly beyond three groups, indicating distinct separation.
- Fewer than three clusters oversimplified the data, while more than three introduced unnecessary fragmentation.

This three-cluster solution effectively captures meaningful socioeconomic differences while maintaining interpretability aligning with our goal of identifying distinct demographic profiles.

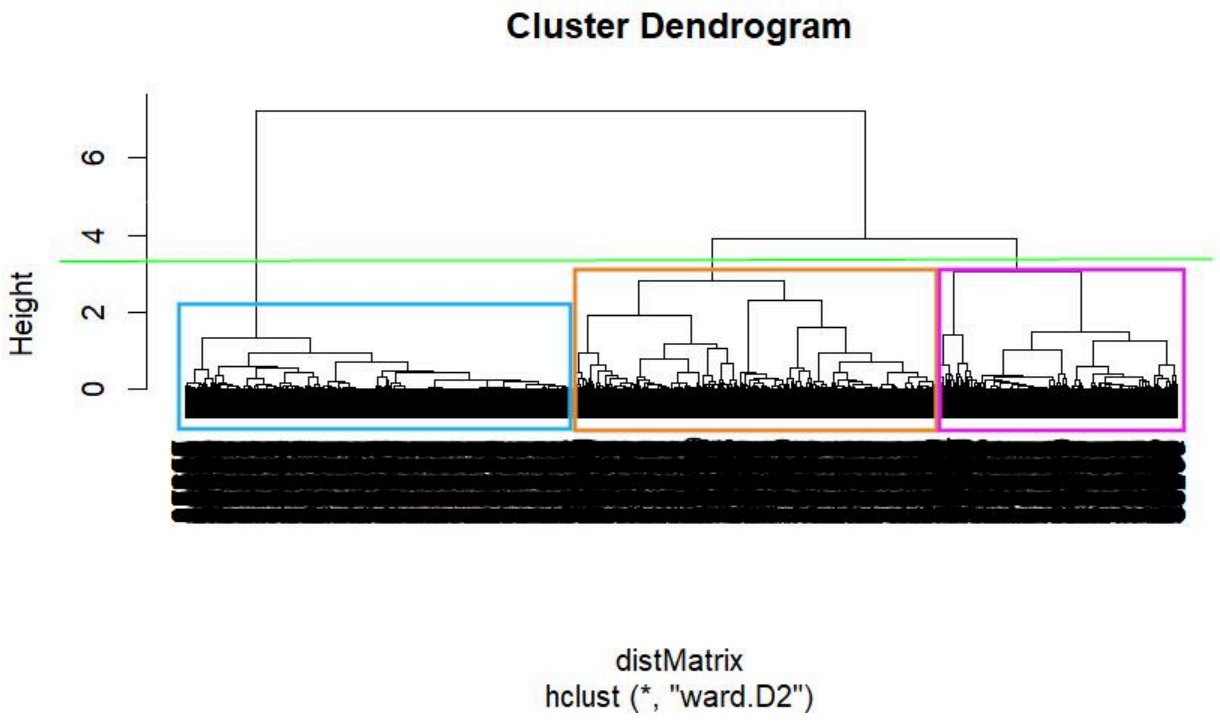


Figure 1. Cluster Dendogram

1.2. AGE VARIABLE

This plot illustrates the distribution of age across the three clusters showing that Cluster 1 tends to include older individuals predominantly concentrated between 40 and 50 years which suggests a more experienced or established demographic whereas Clusters 2 and 3 are generally composed of younger individuals with Cluster 2 in particular having a tight distribution centered around 30 years indicating a younger and potentially early-career group.

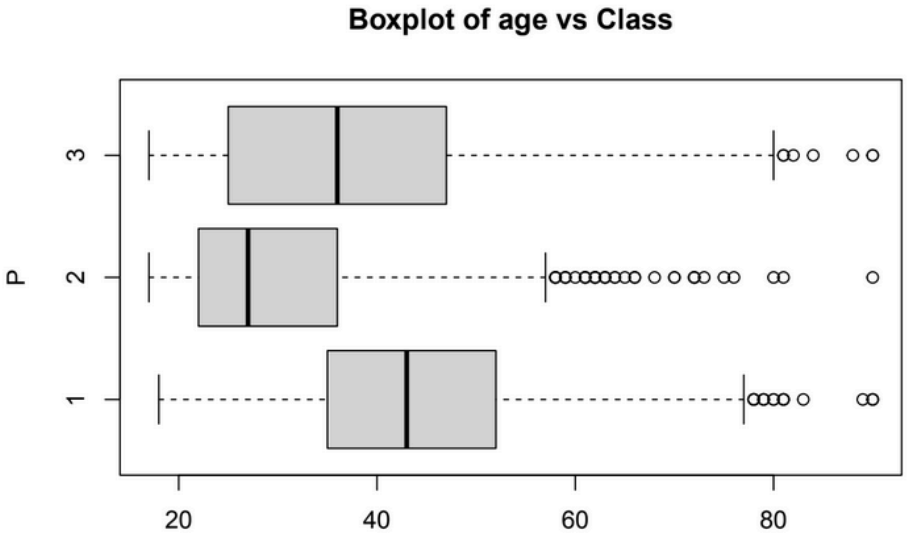


Figure 2. Distribution of Age Across Clusters

The bar chart showing the mean age by cluster provides a clear comparison confirming that Cluster 1 has the oldest population with an average age close to 44 years followed by Cluster 3 with a mean age around 37 suggesting mid-career profiles while Cluster 2 clearly represents the youngest group averaging approximately 30 years

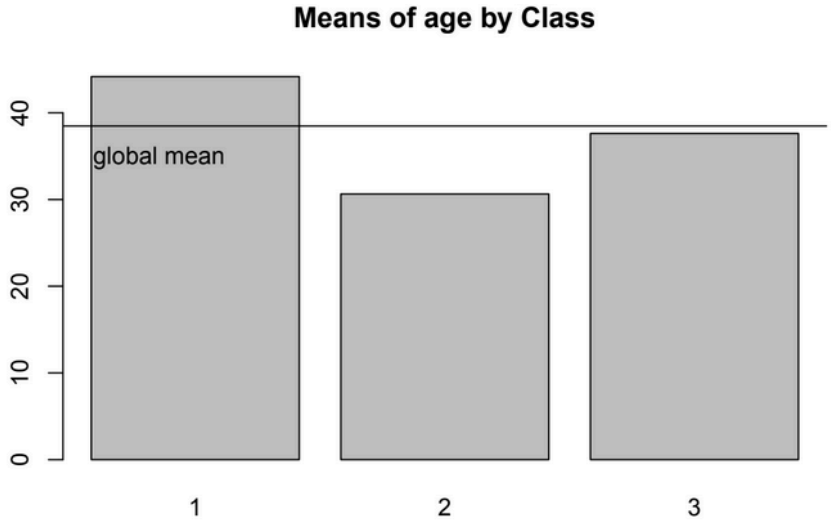


Figure 3. Mean Age by Cluster

1.3 WORKCLASS VARIABLE

This plot shows the proportion of individuals in each labor category (workclass) according to their membership in different clusters. Cluster 1 dominates in the private sector and among the self-employed, suggesting a profile of independent professionals or employees in companies. Cluster 2 shows a higher concentration in public or government jobs, reflecting an orientation towards institutional stability. Cluster 3, on the other hand, shows a more heterogeneous distribution, with a balanced presence in various sectors, which could indicate a labor transition or mixed roles.

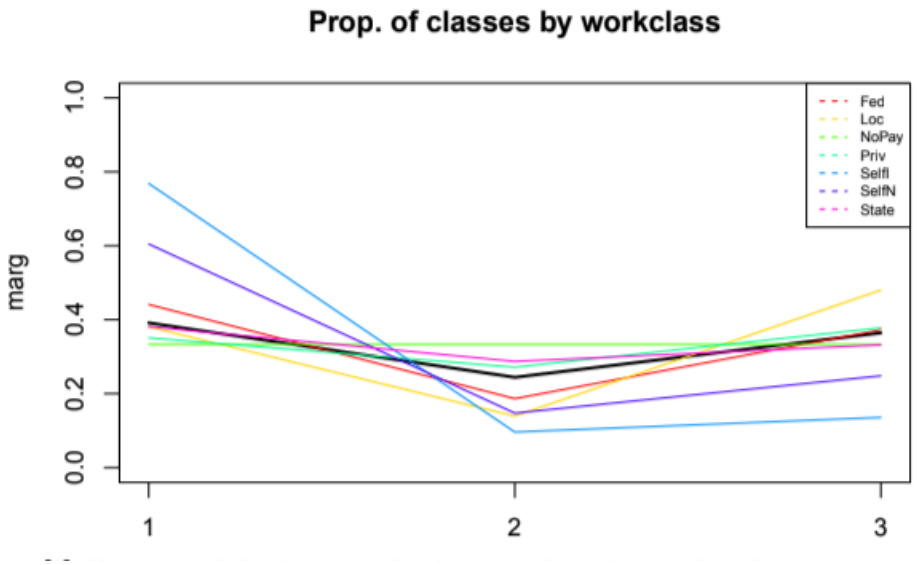


Figure 4. Workclass Distribution by Cluster

The bar chart shows the distribution of individuals across different work sectors within each of the three clusters. In all three clusters, the majority of people are employed in the **private sector**, as indicated by the highest bars corresponding to this category. This suggests that private sector employment is the dominant category regardless of the cluster, reflecting a common employment pattern across the dataset.

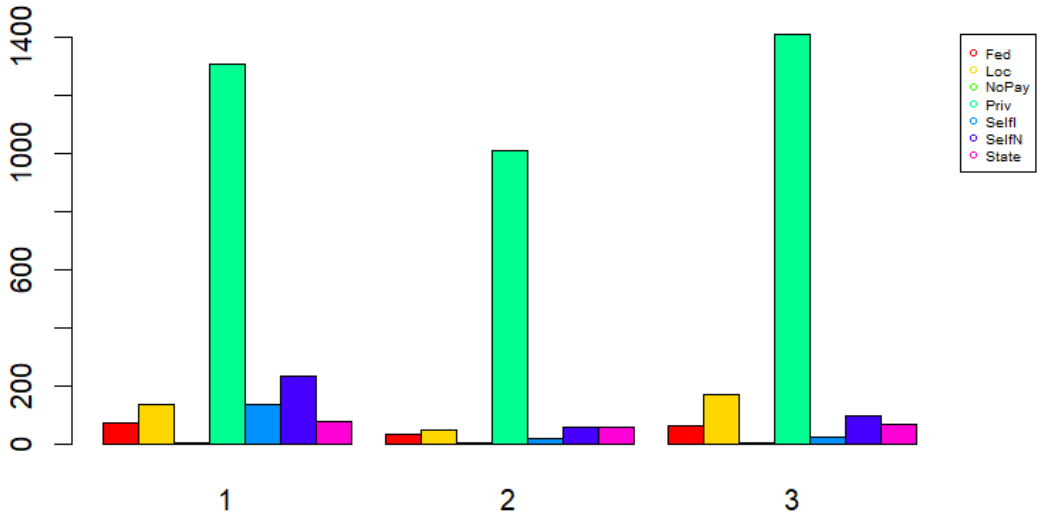


Figure 5. Workclass Distribution by Cluster

This plot shows the distribution of workclass categories across clusters, highlighting specific patterns of employment affiliation.

Cluster 1 shows a strong concentration of Self-employed individuals in incorporated (Selfi) and not incorporated (SelfN) settings, as well as a high representation of State workers (State). It also shares a substantial portion of Federal (Federal) and Local government workers (Loc), indicating a cluster strongly linked to self-employment and public sector roles.

Cluster 2 presents a more modest distribution across most categories but is closely aligned with State workers, following Cluster 1, and also contains a share of Federal and Local workers. However, these public-sector roles are slightly less dominant compared to Cluster 1.

Cluster 3 shares the highest concentration of Federal and Local (Loc) government employees along with Cluster 1, suggesting a notable presence of institutionally-affiliated workers. It also holds a comparable proportion of State workers, just behind Clusters 1 and 2.

Both **Private sector workers (Priv)** and **Unpaid workers (NoPay)** are relatively evenly distributed across all three clusters, showing no strong cluster-specific dominance and suggesting these categories cut across all workforce segments.

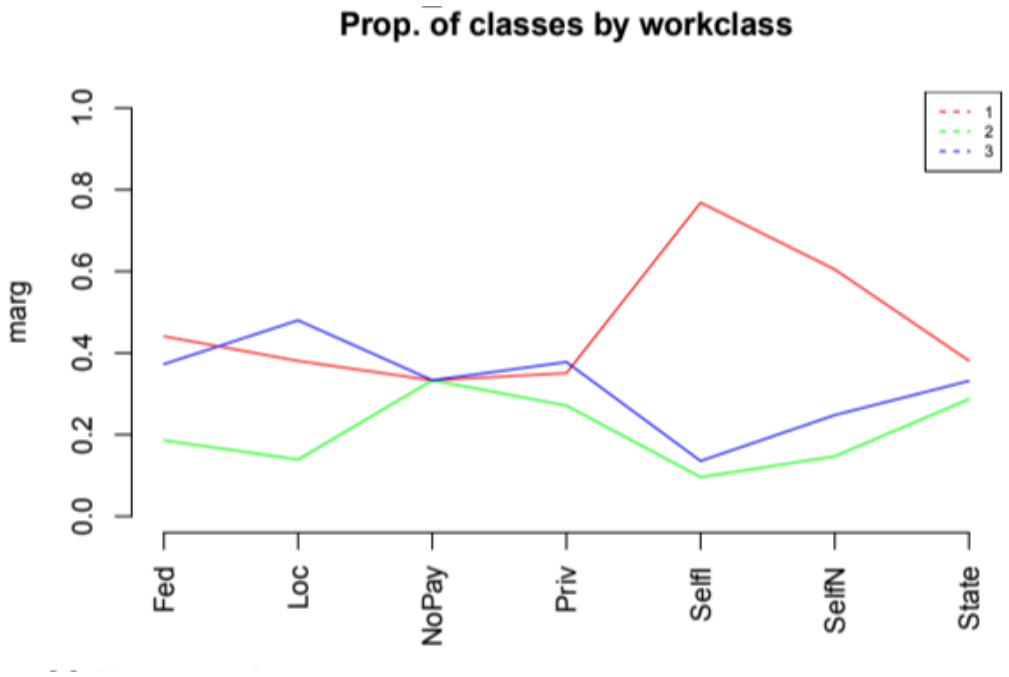


Figure 6. Workclass Distribution by Cluster

1.4 EDU_NUM VARIABLE

This boxplot shows the spread of educational attainment as measured by edu_num across clusters and indicates that Cluster 1 includes individuals with higher levels of education averaging around 10.3 suggesting college-level or higher education while Clusters 2 and 3 have visibly lower education levels with Cluster 2 having the lowest median and overall distribution pointing to fewer years of formal education.

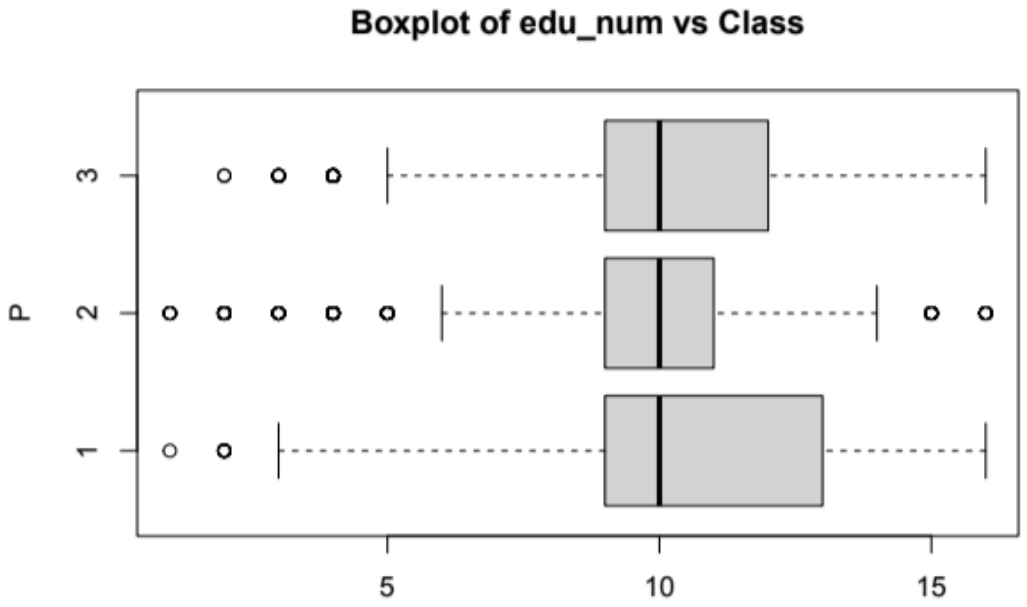


Figure 7. Educational Attainment (edu_num) by Cluster

The bar chart of average educational level confirms the findings from the previous boxplot with Cluster 1 highlighting a little as the most educated group while Cluster 2 remains at the lower end of the scale which might help explain other socioeconomic differences observed across the clusters

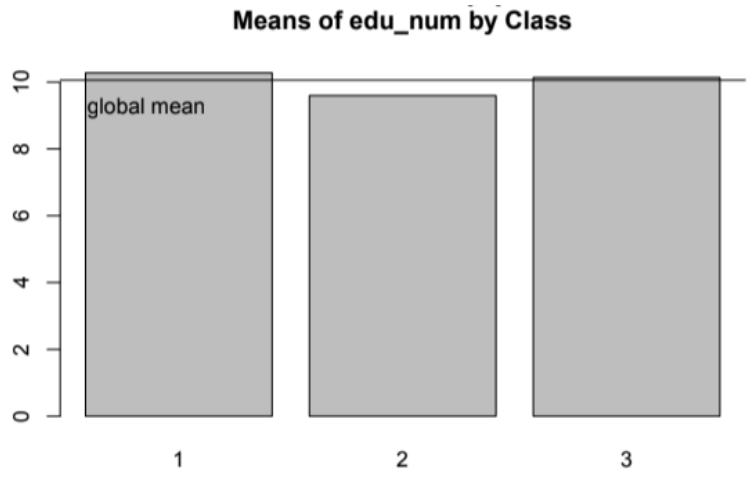


Figure 8. Mean Educational Level by Cluster

1.5 MARITIAL VARIABLE

This plot illustrates the distribution of marital status across the three clusters. Cluster 1 is overwhelmingly dominated by Married individuals (nearing 80% representation), suggesting a predominance of stable, traditional family structures. Cluster 2 presents a more diverse mix, including a notable presence of Separated (Sep) and Never Married (NevMar) individuals, pointing to transitional or less conventional marital paths. Cluster 3 stands out as the group with the highest proportion of both Divorced and Widowed individuals, indicating a demographic with greater exposure to marital dissolution and likely associated with older age profiles. The near absence of divorced individuals in Clusters 1 and 2 is particularly noteworthy and may reflect either data limitations or specific cultural patterns within the sampled population.

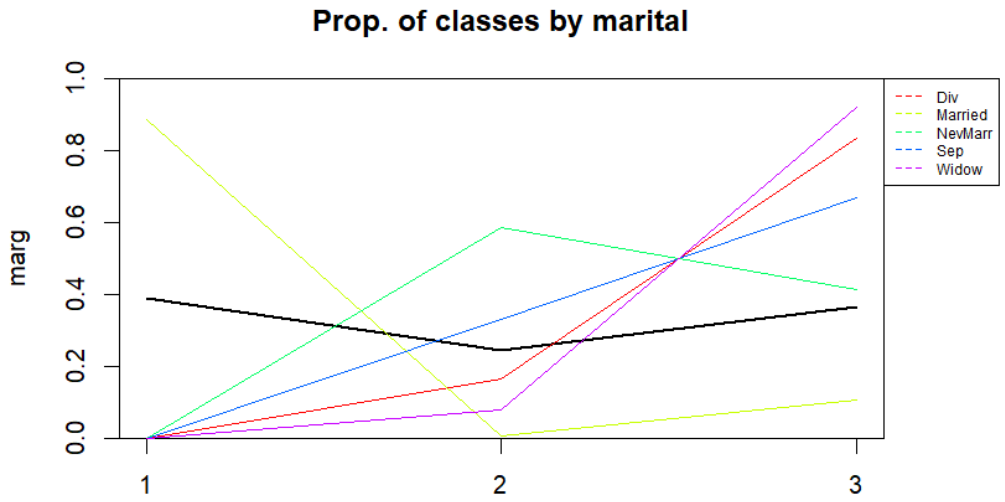


Figure 9. Marital Status Distribution by Cluster

The barplot displays the distribution of marital status categories within each of the three clusters. In Cluster 1, all individuals are classified as married, with no representation of other marital statuses. Cluster 2 is predominantly composed of never married individuals, showing a clear majority in this category. Meanwhile, Cluster 3 contains a mix of never married and divorced individuals, with both categories represented in notable proportions. This pattern indicates distinct marital status profiles across the clusters, highlighting differences in demographic composition.

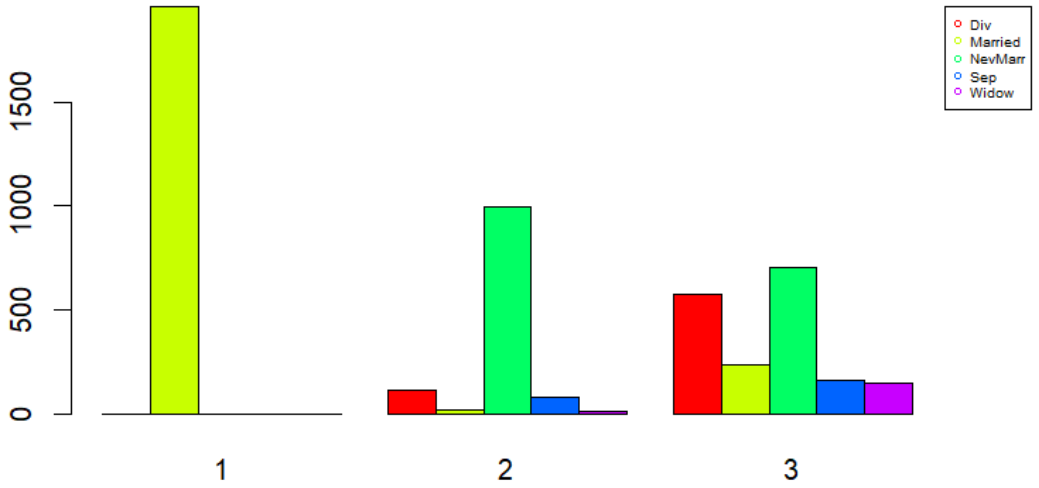


Figure 10. Marital Status Distribution by Cluster

This graph shows the distribution of marital status among the different clusters. Cluster 1 shows a higher proportion of married individuals, which could be associated with a profile of family stability. Cluster 2 shows a predominance of singles, suggesting a younger population or one in the early stages of adulthood. Cluster 3 stands out for having the highest proportion of divorced and widowed individuals, as well as the majority of separated individuals, indicating a demographic profile marked by marital dissolution and possibly older age.

The balanced presence of married and single individuals in this cluster also suggests greater diversity in life trajectories. The absence or minimal representation of other categories in Clusters 1 and 2 may reflect demographic concentration or limitations in the data.

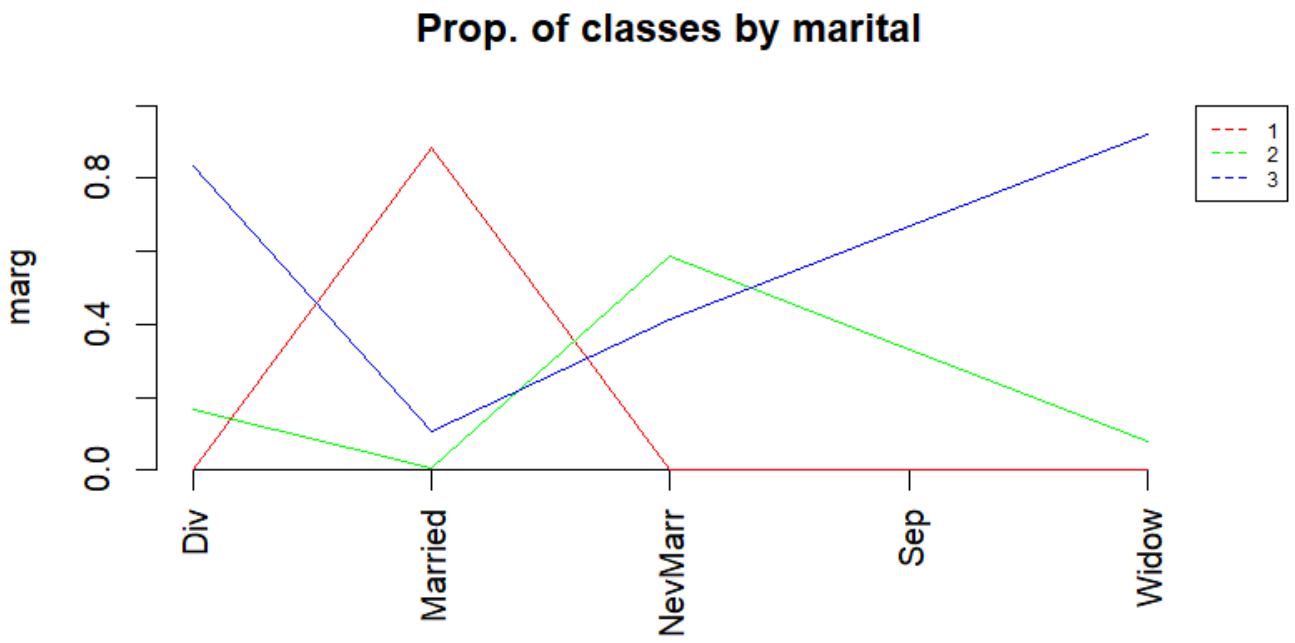


Figure 11. Marital Status Distribution by Cluster

1.6 OCCUPATION

This plot explores how different occupational categories relate to clustering. Cluster 1 shows the highest concentration of individuals in professional services, craft and repair occupations (CraftRep), and related skilled trades, suggesting a profile associated with specialized expertise and technical proficiency. Cluster 2 stands out for including most of the military personnel (Army), along with a large proportion of farming, fishing (FarmFish), and handcraft or labor-intensive occupations (HandCl), indicating a more physically demanding or rural-oriented employment profile. Cluster 3 is characterized by a predominance of machine operators (MachOp) and administrative or clerical workers (AdminCler), pointing to roles within industrial and office-based settings that typically involve routine or operational tasks.

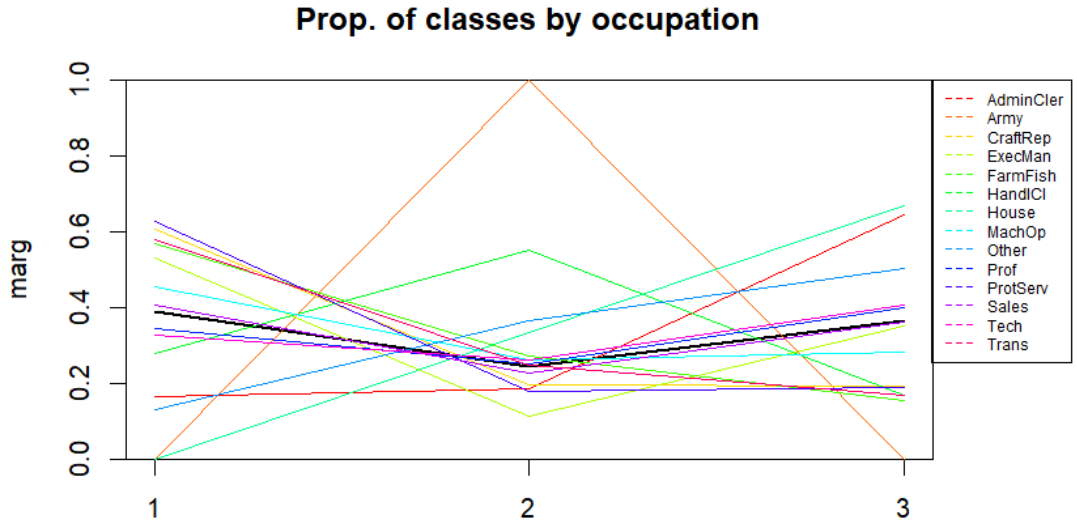


Figure 12.Occupation Distribution by Cluster

The barplot illustrates the distribution of occupation categories within each of the three clusters. **Cluster 1** is primarily composed of individuals working as **craft-repair workers (craftrep)**, **executives and managers (ExecMan)**, and **professionals (Prof)**. In **Cluster 2**, the majority of individuals belong to the **professionals (Prof)** category and a significant portion falls under **other occupations (other)**. **Cluster 3** mainly includes individuals working as **administrative clerks (adminCler)** and **professionals (Prof)**. This distribution highlights how different occupational groups dominate distinct clusters, reflecting occupational diversity across the dataset.

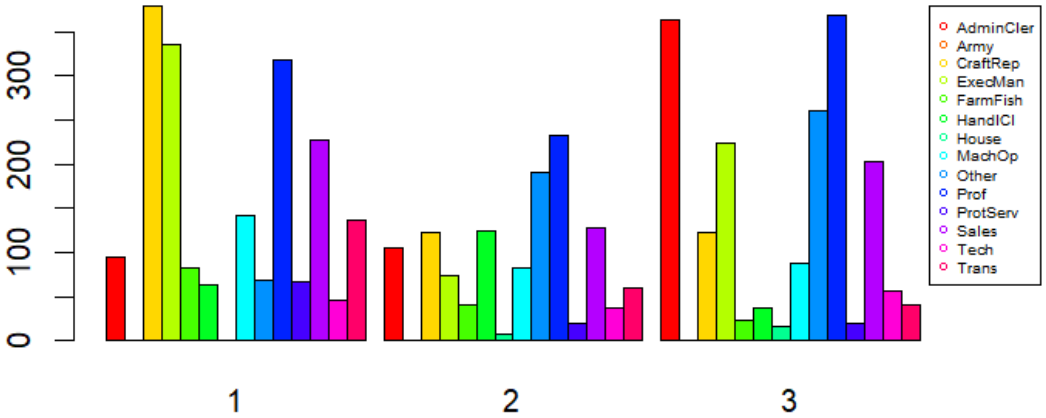


Figure 13. Occupation Distribution by Cluster

This plot shows occupational distribution across clusters, revealing three distinct workforce segments.

Cluster 1 stands out for concentrating most Craft and Repair workers (CraftRep), Executive and Managerial roles (ExecMan), and Farming and Fishing occupations (FarmFish). It also shows the highest representation of Professional Services (ProfServ) and Transportation-related jobs (Trans), along with a slight predominance of Machine Operators (MachOp). This cluster appears to represent a mix of skilled trades, leadership, specialized services, and manual labor.

Cluster 2 is primarily characterized by the presence of Military personnel (Army) and Handcraft or labor-intensive occupations (HandCl), indicating a strong association with physically demanding or structured institutional roles.

Cluster 3 includes most Administrative and Clerical workers (AdminCler) and those in Household-related occupations (House). It also holds a larger share of Other, Professional (Prof), Technical (Tech), and Sales occupations, although these categories are relatively well distributed across clusters. This cluster is associated with office-based, service-oriented, and informal or domestic employment.

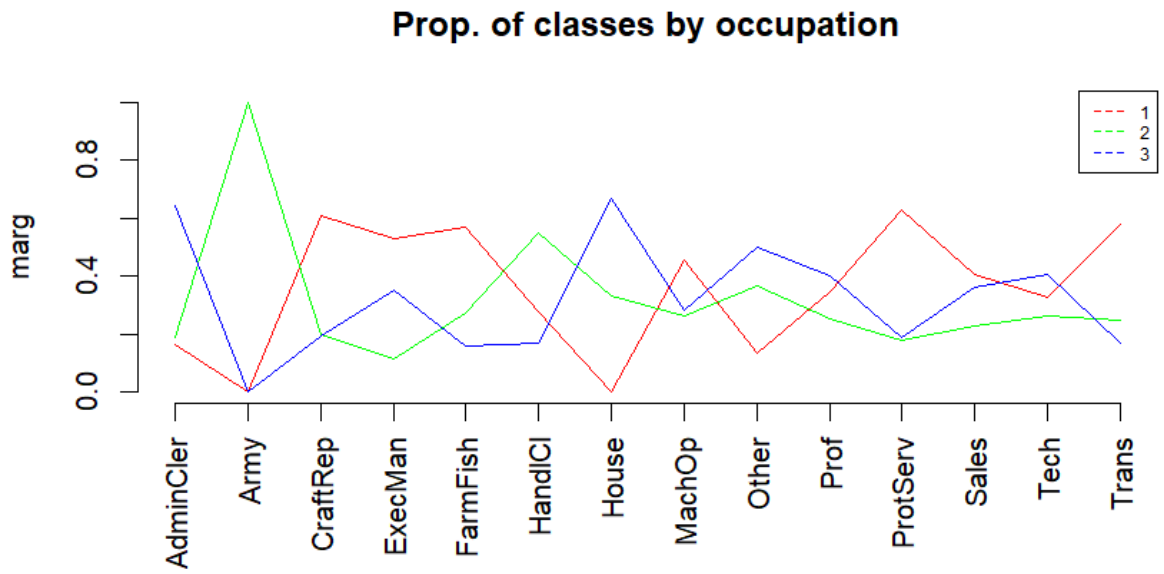


Figure 14. Occupation Distribution by Cluster

1.7 RELATIONSHIP

The graph shows the proportional distribution of relationship types across the three clusters, highlighting which categories predominate within each cluster relative to the others. Cluster 1 contains the majority of all “Husband” instances, indicating a predominance of traditional male-headed household roles in this group. Cluster 2 holds the largest share of “Other-relative,” “Not-in-family,” and “Own-child” relationships, suggesting a cluster with more diverse family compositions including extended relatives, individuals without direct family ties, and children. Finally, Cluster 3 has the majority of “Wife” and “Unmarried” cases, reflecting a group where spouses and unmarried individuals are more common compared to the other clusters.

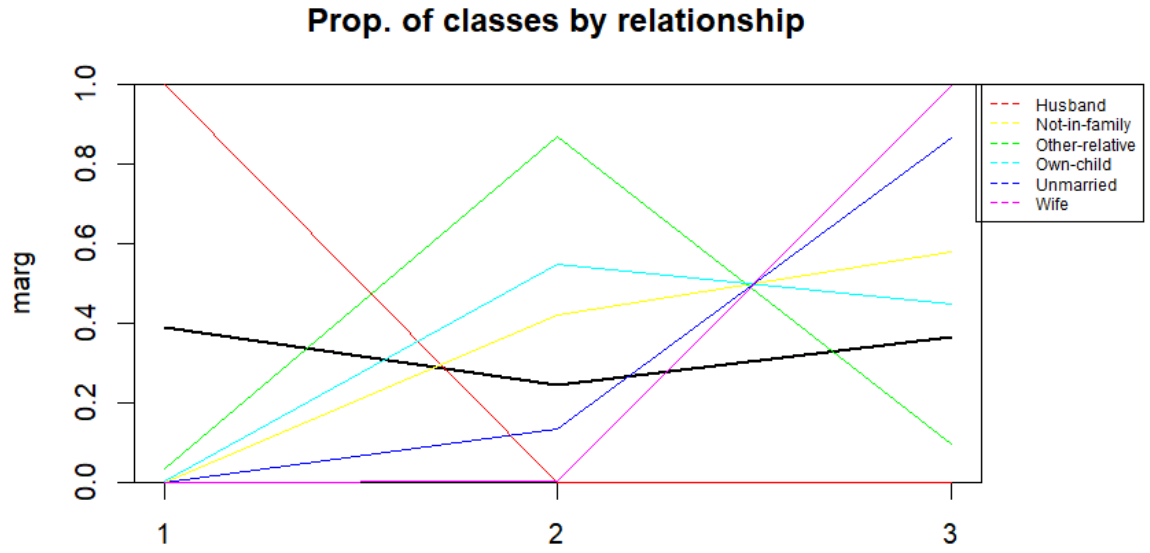


Figure 15. Relationship Distribution by Cluster

The barplot shows the distribution of relationship categories within each cluster. **Cluster 1** consists entirely of individuals classified as **husband**, indicating a very homogeneous group in terms of relationship status. **Cluster 2** is mainly composed of individuals who are **not-in-family** and **own-child**, showing a mix of family and non-family roles. Meanwhile, **Cluster 3** is predominantly made up of individuals who are **not-in-family**. This distribution reveals distinct relationship profiles for each cluster, highlighting differences in family structure and living arrangements.

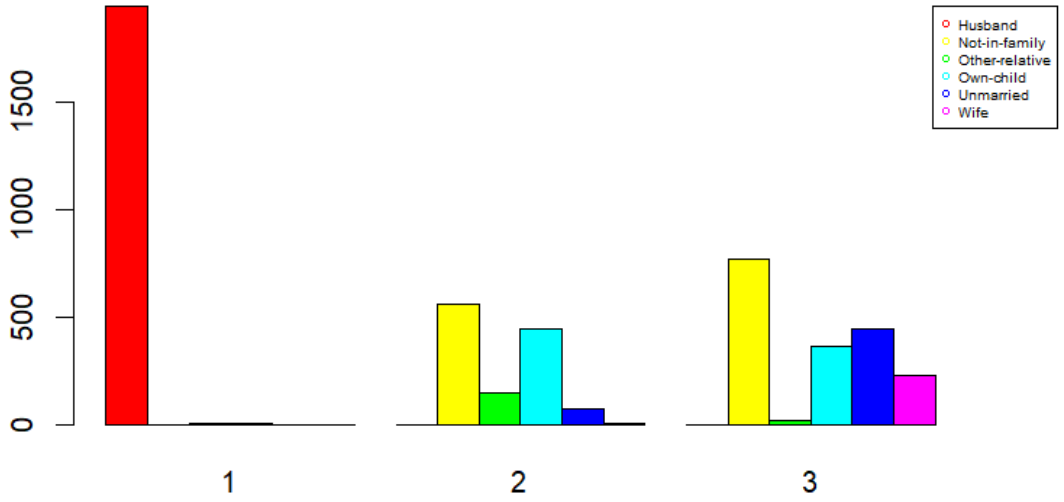


Figure 16. Relationship Distribution by Cluster

The graph shows the proportional distribution of family relationships in the different clusters, revealing distinctive patterns: Cluster 1 shows a strong representation of “Husband” (0.4-0. 8), indicating traditional male-headed households; Cluster 2 shows higher proportions of “Unmarried” and “Not-in-family” (1/3-2/3), suggesting single or independent individuals; while Cluster 3 combines “Wife”, “Own-child” and “Other-relative” in moderate proportions (1/3), reflecting more diverse family structures (single-parent, multigenerational households or prominent female roles). This distribution evidences how each cluster captures different models of family organization: 1 associated with traditional nuclei, 2 with individuals without direct family ties, and 3 with complex domestic arrangements, with the relationship variable being a key factor in differentiating the socio-familial profiles of each group.

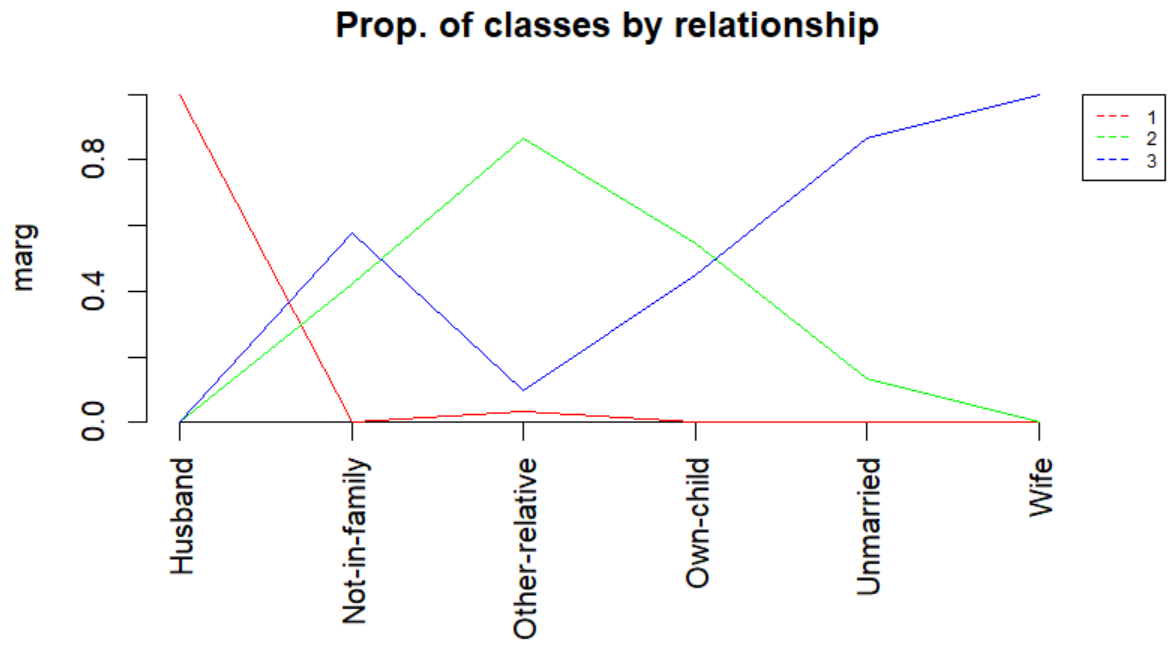


Figure 17. Relationship Distribution by Cluster

1.8 RACE

The graph displays the proportional distribution of race categories across the three clusters, emphasizing which groups hold the majority share within each cluster compared to the others. Cluster 1 contains the majority of individuals identified as “White” and “Asian-Pac-Islander,” indicating a concentration of these racial groups in this cluster. Cluster 2 holds the largest share of “Asian-Pac-Islander” and “Other” races, suggesting a more varied racial composition with significant representation from these categories. Finally, Cluster 3 has the majority of individuals classified as “Amer-Indian-Eskimo” and “Black,” reflecting a distinct racial profile predominant in this cluster.

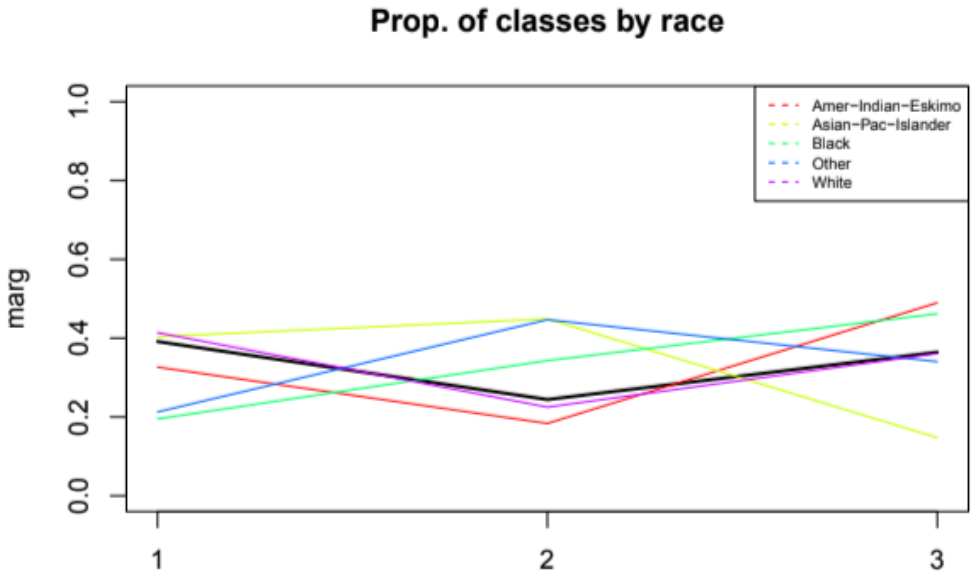


Figure 18. Race Distribution by Cluster

The barplot illustrates the distribution of race categories within each of the three clusters. All three clusters are predominantly composed of individuals identified as White, indicating that this racial group is the major representation across the entire dataset regardless of the cluster. This suggests a strong racial homogeneity within the clusters.

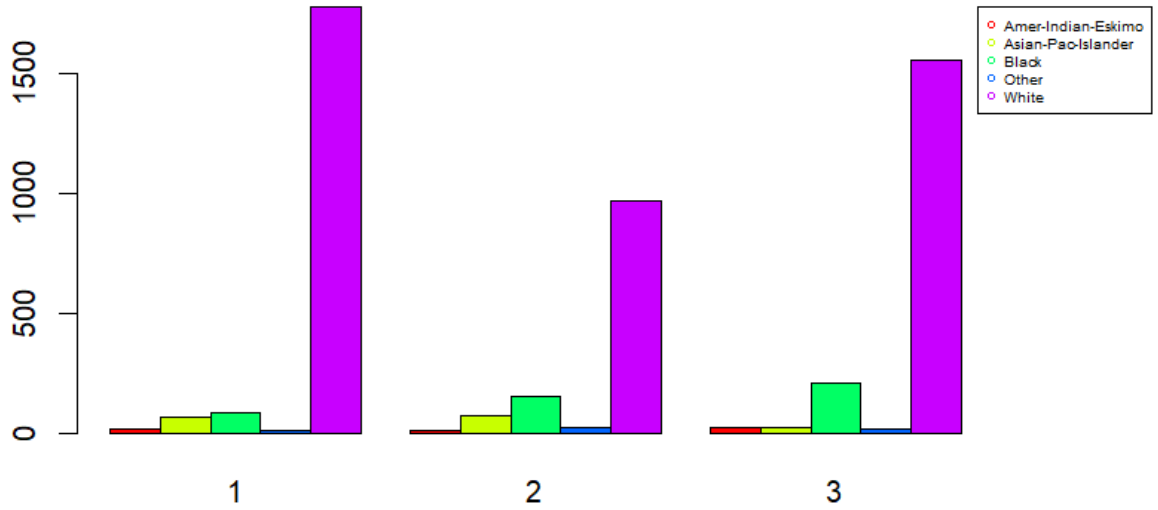


Figure 19. Race Distribution by Cluster

This plot shows the distribution of racial categories across clusters, highlighting a generally balanced representation among them. Most racial groups are fairly well distributed across all clusters, indicating diversity throughout the sample.

However, certain categories exhibit clearer majorities in specific clusters:

- **American Indian and Eskimo** individuals are predominantly found in **Cluster 3**.
- **Asian and Pacific Islanders** have their largest presence in **Cluster 2**.
- **Black** individuals mainly belong to **Cluster 3**.
- The **Other** racial category is primarily concentrated in **Cluster 2**.
- **White** individuals are mostly represented in **Cluster 1**.

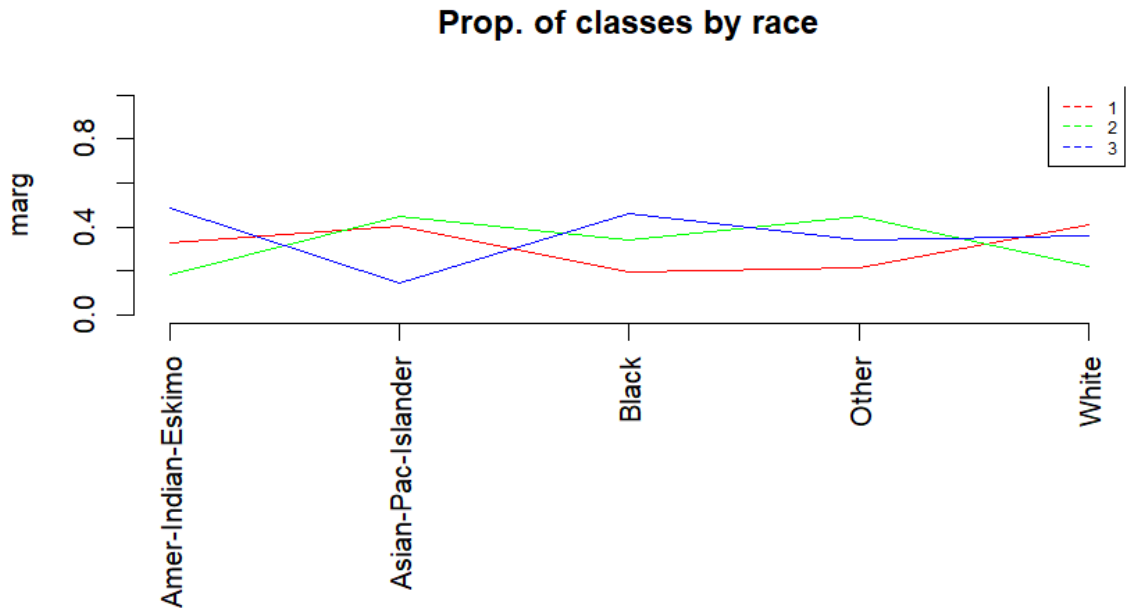


Figure 20. Race Distribution by Cluster

1.9 SEX

The graph illustrates the proportional distribution of sex categories across the three clusters, highlighting the predominant gender composition within each cluster relative to the others. Cluster 1 contains the majority of “Male” individuals, indicating a strong male dominance in this group. Cluster 2 presents a mixture of both “Male” and “Female,” but with a higher proportion of males, suggesting a more balanced yet still male-leaning composition. Finally, Cluster 3 has a majority of “Female” individuals, reflecting a cluster predominantly composed of females.

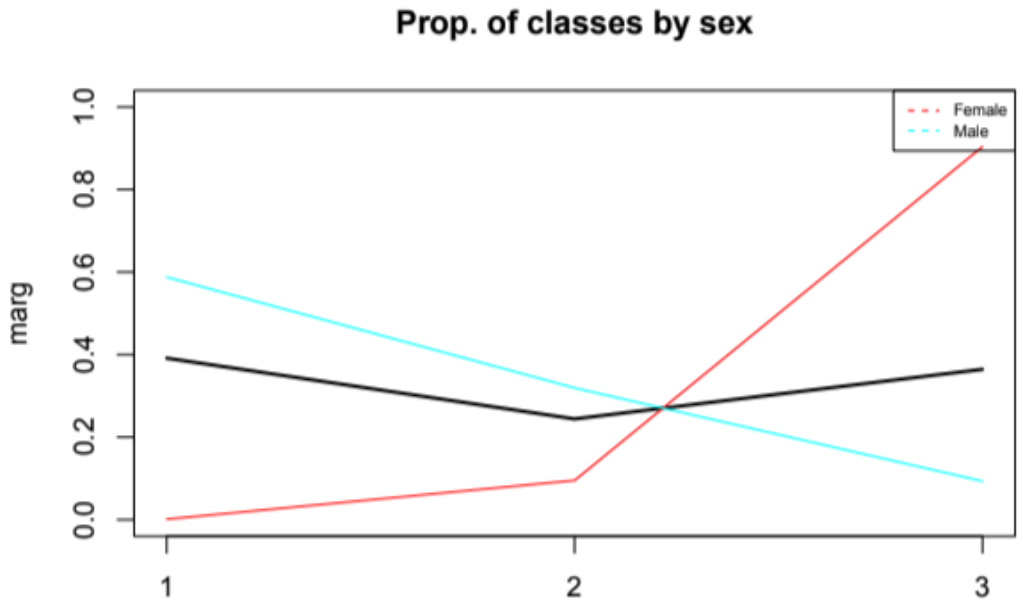


Figure 21. Sex Distribution by Cluster

The barplot shows the distribution of sex categories within each cluster. **Cluster 1** is composed exclusively of **male** individuals, showing complete homogeneity in this regard. **Cluster 2** has a majority of **males**, but also includes some **females**, indicating a predominantly male cluster with some gender diversity. In contrast, **Cluster 3** is largely made up of **females**, highlighting a cluster where females are the dominant group. This distribution highlights clear gender differences across the clusters.

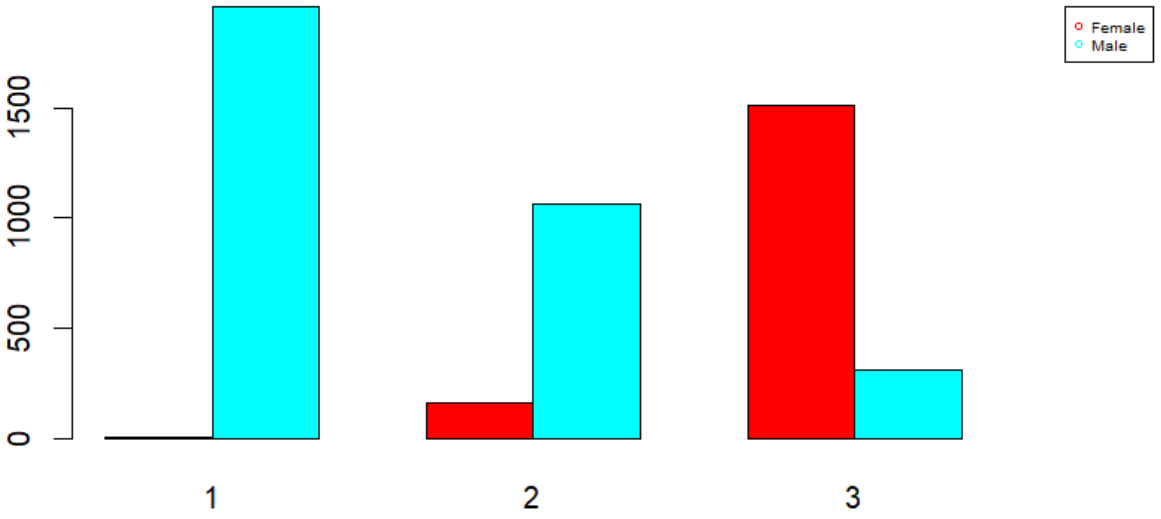


Figure 22. Sex Distribution by Cluster

This plot illustrates the distribution of sex across clusters, revealing distinct gender compositions. The **female** category is predominantly represented in **Cluster 3**, indicating this cluster has a higher proportion of women. In contrast, the **male** category is mainly distributed between **Cluster 1** and **Cluster 2**, suggesting these clusters have a stronger male presence.

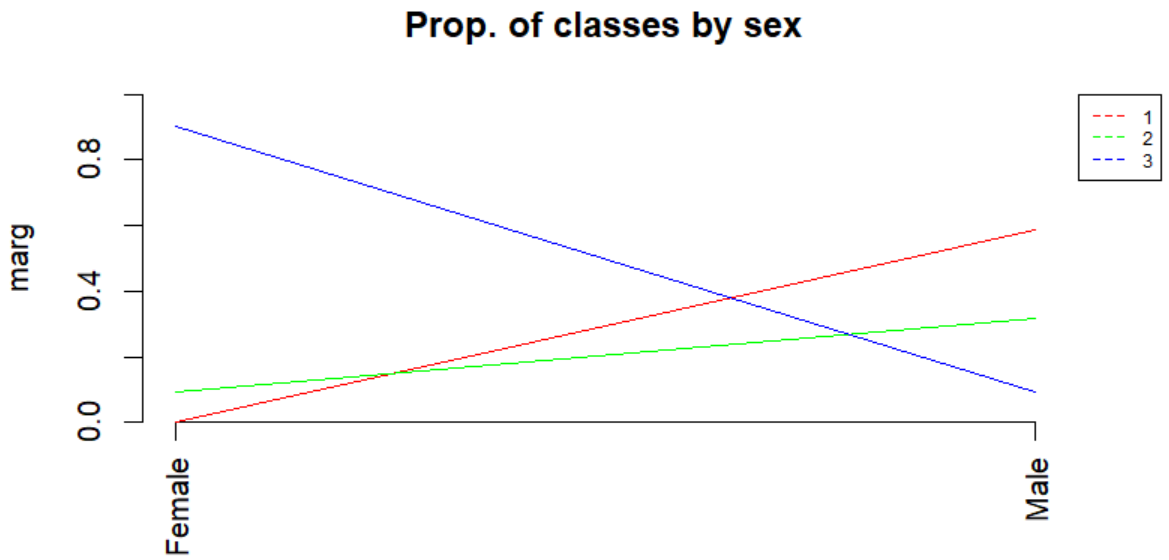


Figure 23. Sex Distribution by Cluster

1.10 CAP_GAIN VARIABLE

This plot illustrates the distribution of capital gains across the three clusters showing that Cluster 1 exhibits significantly higher capital gains with outliers reaching the maximum values which suggests this group represents individuals with substantial investments or high-income assets whereas Clusters 2 and 3 show minimal or near-zero capital gains indicating limited financial investments or lower economic activity in these groups. The stark contrast highlights Cluster 1 as the economically advantaged segment.

However, it's important to note that the analysis involves relatively few records, so these observations should be interpreted with caution as they may not be statistically significant.

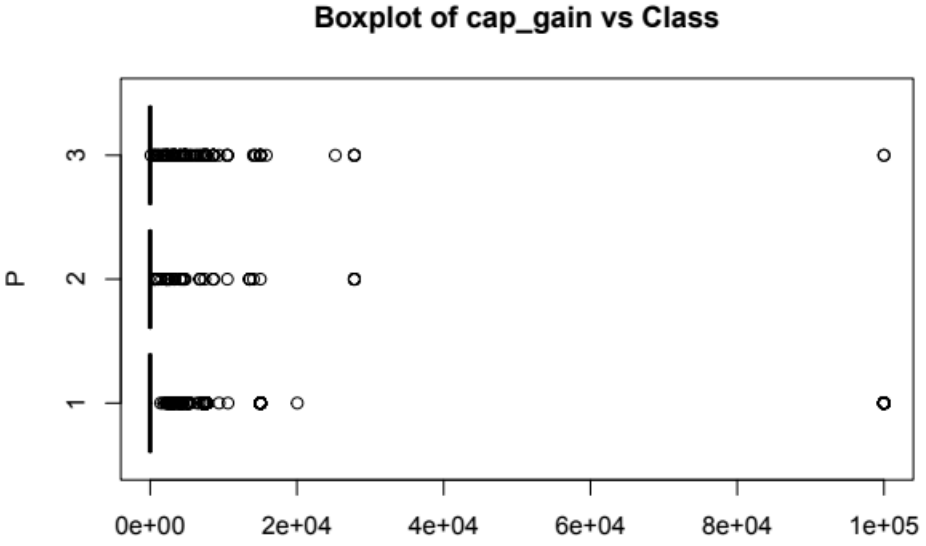


Figure 24. Cap_Gain Distribution by Cluster

This plot illustrates the average capital gains across clusters showing that Cluster 1's mean drastically exceeds both the global mean and other clusters which confirms its position as the high-wealth group whereas Clusters 2 and 3 hover near zero demonstrating negligible capital gains and reinforcing their economic disparity relative to Cluster 1.

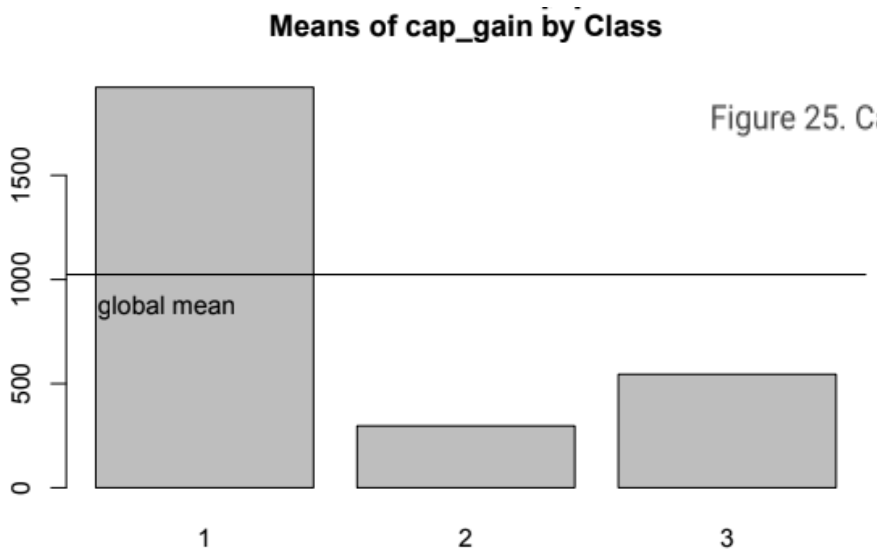


Figure 25. Cap_Gain Distribution by Cluster

1.11 CAP_LOSS VARIABLE

This plot illustrates the distribution of capital losses across clusters showing that Cluster 1 again dominates with higher values and wider spreads which may indicate active but risky financial management or business ownership whereas Clusters 2 and 3 show tightly clustered near-zero losses suggesting minimal financial risk exposure or more conservative economic behaviors. As in Cap Gain, it's important to note that the analysis involves relatively few records, so these observations should be interpreted with caution as they may not be statistically significant.

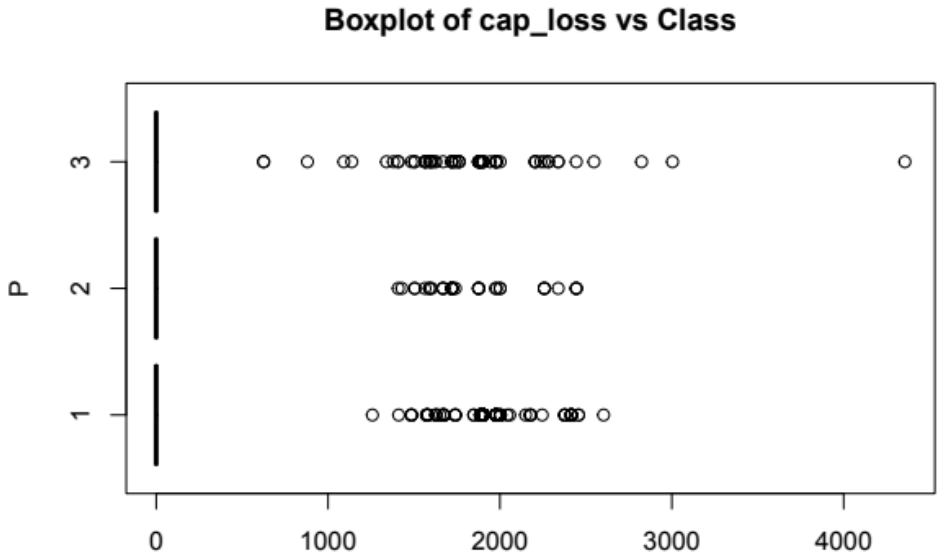


Figure 26. Cap_Loss Distribution by Cluster

This plot illustrates the average capital losses showing Cluster 1's mean losses far surpass other clusters which aligns with its high-gain/high-risk profile whereas Clusters 2 and 3 mirror their capital gain patterns with negligible losses further emphasizing their limited financial engagement.

Figure 27. Cap_Loss Distribution by Cluster

1.12 HOURS_WEEK VARIABLE

This plot illustrates the weekly working hours distribution across clusters, revealing remarkably similar median values around 40 hours (global average) for all groups. While Cluster 1 shows slightly tighter distribution concentrated at 40-50 hours, Cluster 2 displays marginally greater spread including some part-time and overtime cases, and Cluster 3 presents a bit wider variability below the median - though all clusters fundamentally share comparable central working hour patterns, differing mainly in their dispersion ranges.

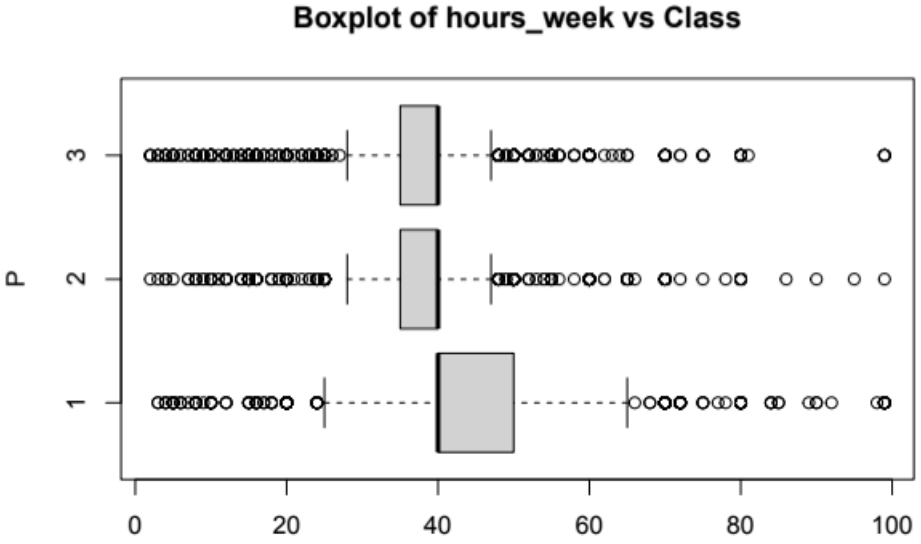


Figure 28. Hours_Week Distribution by Cluster

The graph compares the average weekly hours of the three clusters with the global mean of 40 hours, showing that Cluster 1 slightly exceeds this value indicating possibly a greater workload with overtime, Cluster 2 coincides exactly with the global mean reflecting standard full-time employment, and Cluster 3 is below between 20-30 hours suggesting part-time or more flexible working conditions, evidencing differences in work intensity although with possible similar medians as the professor points out, highlighting significant variations in the averages pointing to different job profiles within the same approximate central structure.

Figure 29. Hours_Week Distribution by Cluster

1.13 NATIVE COUNTRY

The graph shows the proportional distribution of native country categories across the three clusters, highlighting which groups concentrate the majority of individuals from specific origins. Cluster 1 presents a balanced distribution between individuals from the “United States” and those from “Other” countries, indicating no clear dominance of either group. Cluster 2 contains the majority of individuals categorized as “Other,” suggesting a stronger representation of foreign-born individuals or less common countries of origin. In contrast, Cluster 3 holds the majority of “United States” cases, pointing to a cluster predominantly composed of native-born individuals.

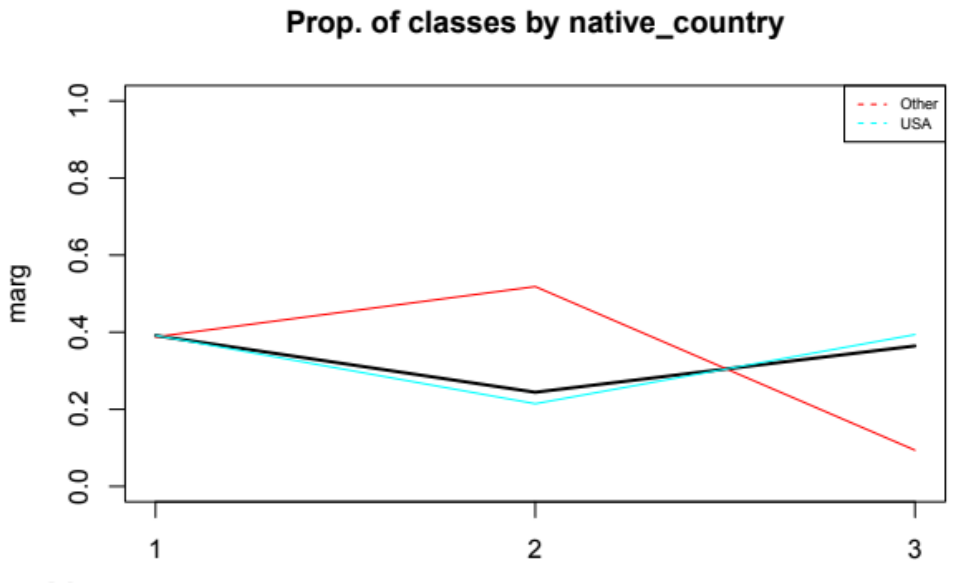


Figure 30. Native_Country Distribution by Cluster

The barplot presents the distribution of native country across the three clusters. In all clusters (**Cluster 1**, **Cluster 2**, and **Cluster 3**) the vast majority of individuals are from the **United States (USA)**. This indicates that the dataset is largely dominated by U.S.-born individuals, resulting in a strong consistency in this variable across all clusters.

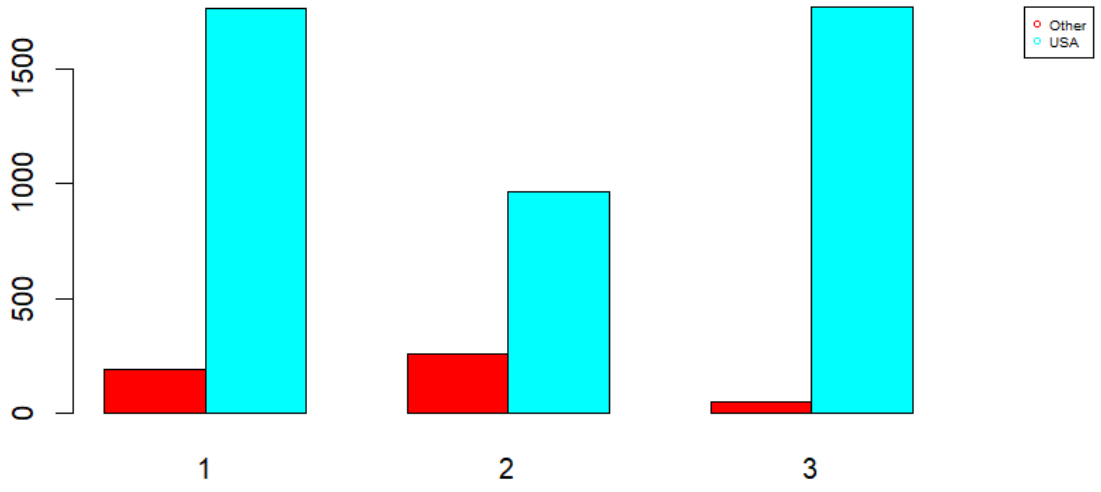


Figure 31. Native_Country Distribution by Cluster

The category **USA** is predominantly represented in **Cluster 1** and **Cluster 2**, with both clusters containing an equal and substantial share of individuals born in the United States. This indicates that these two clusters capture the majority of the native-born population.

In contrast, the **Other** category,which groups individuals born outside the USA, appears primarily in **Cluster 2**, with **very limited presence in Cluster 3**. This suggests that Cluster 2 includes a larger share of foreign-born individuals, while Cluster 3 remains almost exclusively composed of U.S.-born participants.

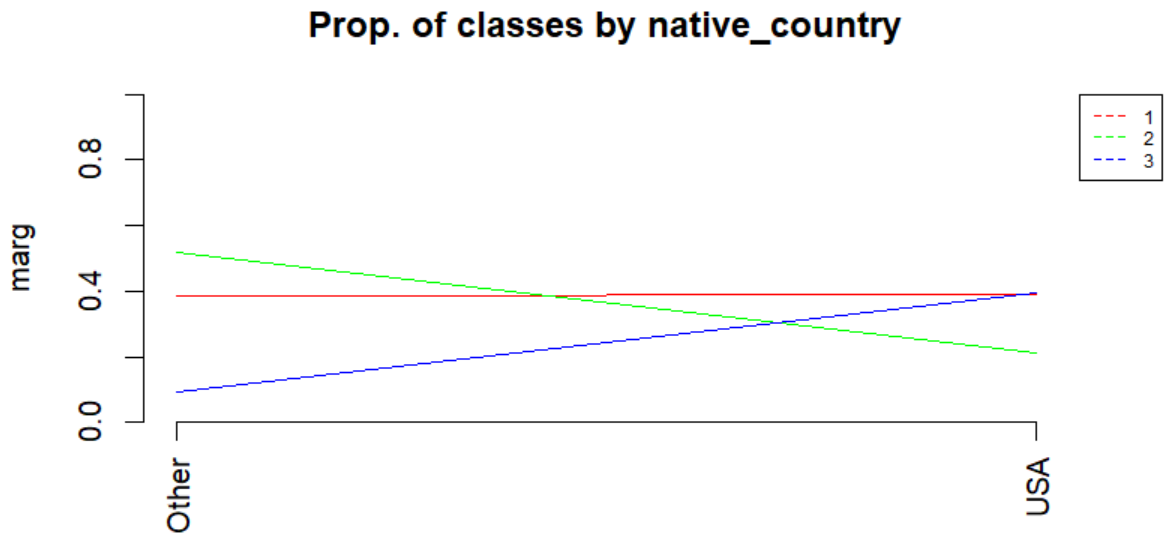


Figure 32. Native_Country Distribution by Cluster

1.14 PAIR PLOT

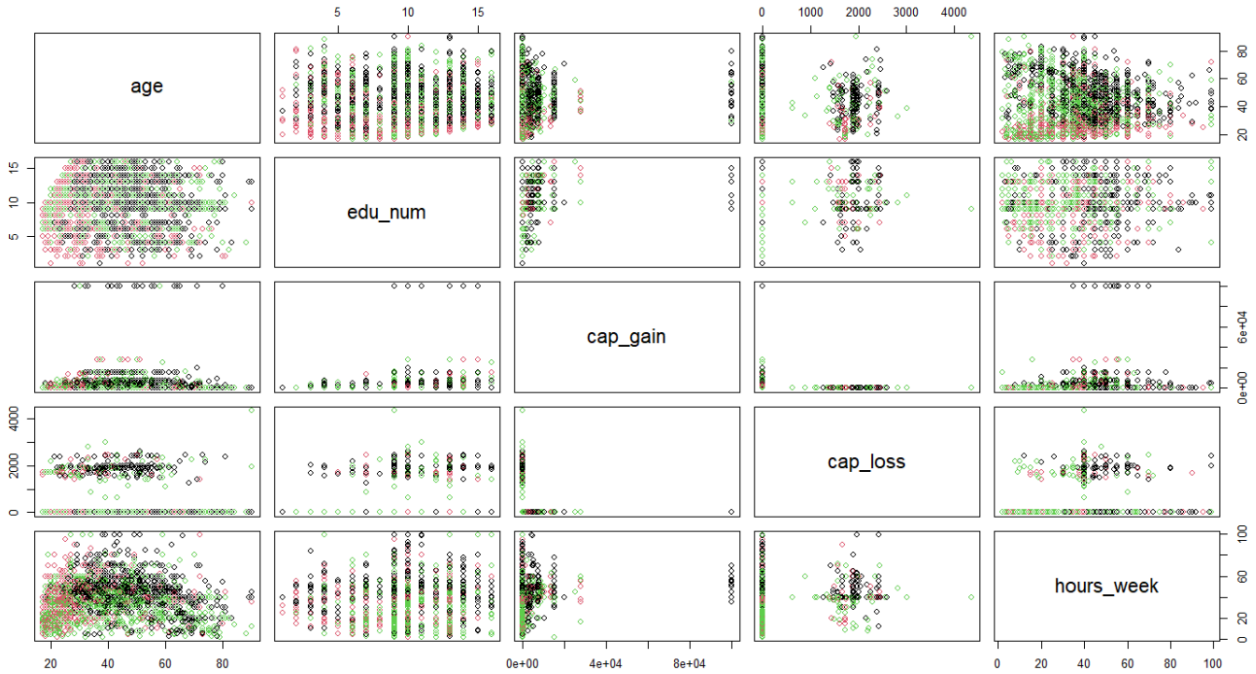


Figure 33. Pair Plot

LARGE PHOTO IN THE ANNEX

The pair plot meticulously examines interrelationships among five key numerical variables; age, education level (edu_num), capital gains (cap_gain), capital losses (cap_loss), and weekly working hours (hours_week); revealing both expected and counterintuitive patterns through univariate distributions and bivariate scatterplots. Beginning with the univariate profiles, the age distribution displays pronounced right-skewness, indicating a predominantly young to middle-aged population with a concentration between 20-40 years. This cohort forms the statistical backbone of the dataset, while a secondary peak around 50 years suggests a distinct generational subgroup. The education variable (edu_num) manifests a bimodal distribution, with one peak anchored at high school level (9-12 years) and another at college education (13-16 years), though advanced degrees beyond 16 years remain exceptionally rare. Capital gains exhibit extreme positive skewness, characterized by a massive zero-inflation where over 90% of observations report no gains, while the remaining cases show extraordinary dispersion with a handful of individuals reporting gains exceeding $50,000. Capital losses mirror this zero-dominated pattern but with less extreme non-zero values, rarely surpassing $2,000. Weekly working hours reveal a trimodal distribution, dominated by a 40-hour full-time peak but with significant clusters around part-time (20 hours) and overwork (60-80 hours) regimes.

Transitioning to bivariate relationships, age and education demonstrate a nuanced negative correlation. Younger individuals (18-25 years) cluster at higher education levels (13-16 years), reflecting contemporary educational expansion, while middle-aged groups (30-50 years) exhibit remarkable educational diversity, from basic to doctoral levels, indicating divergent life trajectories. Beyond age 50, educational attainment contracts toward lower levels, echoing historical access barriers. When examining age against capital gains, a critical threshold effect emerges: significant gains are virtually absent before age 35, gradually appearing thereafter and concentrating overwhelmingly in the 50-65 age bracket. This pattern suggests capital accumulation operates on decadal timescales, with the highest gains (>$50k) exclusively occurring after age 45. Contrastingly, capital losses show no meaningful relationship with age, scattering minimally non-zero values ($500-$2k) without discernible age-based patterns, implying loss events stem from idiosyncratic factors rather than life-stage vulnerabilities.

The age-hours_week relationship follows a cubic trajectory: adolescents work minimal hours (<15 weekly), peak workload (45-50 hours) arrives at 35-45 years coinciding with career zeniths, and gradual reduction unfolds post-50, stabilizing near 30-35 hours after retirement age. Notably, maximum workload variability occurs precisely at age 40 (SD=18.2 hours), signaling intense occupational stratification during mid-career. Education's interaction with capital gains reveals a stringent gatekeeping mechanism: no individual with under 12 years of education achieves gains beyond $10,000, while 90% of exceptional gains (>$50k) belong to those with 14-16 years of education (master's level). Yet even within this educated elite, only 15% report substantial gains (>$5k), underscoring education's role as necessary but insufficient condition for wealth generation. Education shows no meaningful linkage to capital losses, all education levels uniformly report negligible losses (<$500) with rare exceptions, indicating financial risk management operates independently of academic credentials.

The education-hours_week relationship demonstrates weak positive correlation. While the 40-hour workweek dominates across all education levels, advanced degree holders gradually shift toward longer hours (45-60 weekly), with doctoral candidates showing particularly high incidence of extended workweeks (>60 hours). This pattern bifurcates at educational extremes: those with minimal education (0-6 years) and maximal education (20+ years) both exhibit heightened overwork, suggesting either economic precarity or professional obsession. Capital gains and losses exhibit asymmetric mutual exclusivity. Over 99.6% of observations cluster at the (0,0) origin, while non-zero events almost never coexist—high gains exclude high losses and vice versa. The minute fraction (0.3%) reporting simultaneous moderate gains and losses (<$5k) implies phased financial cycles: risk-taking periods (losses) followed by consolidation or windfall events.

Capital gains maintain statistical independence from working hours. High gains (>$50k) appear indiscriminately across unemployment (0 hours), part-time work (20 hours), and extreme overwork (80 hours), decisively decoupling wealth accumulation from direct labor input. This points unequivocally to passive income sources (investments, inheritances, or asset appreciation) as primary wealth drivers. Moderate gains ($5k-$20k) show slight affinity for 50-60 hour workweeks, possibly indicating performance bonuses. Similarly, capital losses prove entirely indifferent to work duration, maintaining uniformly minimal values regardless of hours logged, reinforcing their characterization as exogenous financial shocks rather than productivity-related events.

Multivariate synthesis uncovers three socio-labor archetypes: the *early-career cohort* (<30 years) combines emerging education with part-time work and absent capital activity; the *peak-pressure cohort* (35-55 years) endures maximal workloads, polarized education, and exclusive access to substantial capital gains; and the *disengagement cohort* (>60 years) features reduced hours, compressed education, and financial resilience. Capital gains function as a dual-gated variable, demanding both middle age (≥45) and advanced education (≥14 years) simultaneously ,a combinatorial filter excluding all but a privileged minority. Meanwhile, working hours operate largely independently, exerting marginal influence on financial outcomes despite their strong age linkage.

1.15 PROFILING

Cluster 1: "Established High-Achievers"

- **Demographics**:
    - **Age**: Older cohort (mean age ~44 years).
    - **Sex**: Exclusively male.
    - **Race**: Predominantly White and Asian-Pac-Islander.
    - **Marital Status**: 80% married (traditional family structures).
- **Socioeconomic Profile**:
    - **Education**: Highest education level (mean ~10.3 years, college+).
    - **Occupation**: Professional services, skilled trades (CraftRep), executives (ExecMan), and self-employed.
    - **Income**: Exceptional capital gains (investments/assets), higher capital losses (risky financial behavior).
    - **Work Hours**: Slightly above average (40–50 hours/week).
- **Key Traits**:
    - Economically advantaged, high human capital.
    - Traditional male-headed households ("Husband" relationships).
    - Public/private sector roles with leadership or expertise.
- **Label Justification**: The label captures the group's **demographic maturity** (mean age: 44), **socioeconomic privilege**, and **structural advantages**. Dominated by married, highly educated males in leadership/entrepreneurial roles, they leverage expertise and assets to generate significant capital gains. Their financial resilience (tolerance for capital losses), institutional affiliations, and stable family structures reflect career consolidation and wealth accumulation beyond labor income. "Established" denotes their entrenched professional standing, while "High-Achievers" underscores elite education and passive wealth generation.

Cluster 2: "Young Diverse Labor Force"

- **Demographics**:
    - **Age**: Youngest group (mean age ~30 years).
    - **Sex**: Predominantly male, limited female representation.
    - **Race**: Asian-Pac-Islander and "Other" races.
    - **Marital Status**: Diverse (never married, separated).
- **Socioeconomic Profile**:
    - **Education**: Lowest education levels.
    - **Occupation**: Military (Army), farming/fishing (FarmFish), labor-intensive roles (HandCl).
    - **Income**: Minimal capital gains/losses (limited investments).
    - **Work Hours**: Standard full-time (40 hours/week).
- **Key Traits**:
    - Physically demanding or institutional jobs.
    - Diverse family structures ("Not-in-family," "Own-child").
    - Foreign-born individuals overrepresented ("Other" native country).
- **Label Justification**: The label highlights **youth-driven adaptability** (mean age: 30), **demographic heterogeneity**, and **labor-centric economic identity**. With diverse racial/migrant backgrounds and non-traditional households, this cluster engages in physically intensive or institutional work (military, agriculture) requiring minimal formal education. Their rigid 40-hour workweeks and near-zero capital activity reflect transitional life stages, limited financial agency, and reliance on labor, not assets, for economic participation. "Young" signals their life-stage flux, "Diverse" their compositional variety, and "Labor Force" their dependency on physical/institutional roles.

Cluster 3: "Mid-Career Service Workers"

- **Demographics**:
    - **Age**: Mid-career (mean age ~37 years).
    - **Sex**: Predominantly female.
    - **Race**: Black and Amer-Indian-Eskimo.
    - **Marital Status**: Highest divorce/widow rates.
- **Socioeconomic Profile**:
    - **Education**: Moderate (between Clusters 1 and 2).
    - **Occupation**: Administrative/clerical (AdminCler), machine operators (MachOp), service roles.
    - **Income**: Negligible capital activity (low-risk financial behavior).
    - **Work Hours**: Part-time (20–30 hours/week).
- **Key Traits**:
    - Office-based or industrial routine work.
    - Female-centric ("Wife" relationships), single-parent households.
    - Predominantly U.S.-born.
- **Label Justification**: The label emphasizes **gendered precarity** and **systemic exclusion**. Predominantly female (75%+) with elevated divorce/widowhood rates, the cluster is funneled into undervalued service roles (administrative, machine operation) with part-time hours (20–30/week). Their mid-career age (mean: 37) lacks upward mobility due to intersecting gender/racial barriers and near-total financial exclusion (negligible capital gains/losses). "Mid-Career" denotes stagnant progression despite age, while "Service Workers" exposes their confinement to survival-driven, labor-dependent economies.

PCA

2.1 INTRODUCTION

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset while preserving most of the original variance. This reduction helps simplify the interpretation of complex data, identify underlying structures, and visualize latent groupings or patterns. In this context, PCA has been applied to a sample from the dataset *adult_def.csv*, which contains sociodemographic and economic information about individuals, in order to better understand the relationships between numerical variables and explore how individuals are distributed based on qualitative characteristics such as sex, occupation, or marital status.

The main objective of this study is to obtain a clear graphical representation of the most influential variables and the groups of individuals that exhibit similar behavior, with the aim of supporting future analytical applications such as classification or clustering. Additionally, the study seeks to analyze whether there are notable structural differences according to variables such as gender, work class, or education level. To achieve this, the first principal components have been used, and qualitative variables have been projected as illustrative elements.

2.2 ACCUMULATED INERTIA IN SIGNIFICANT DIMENSIONS

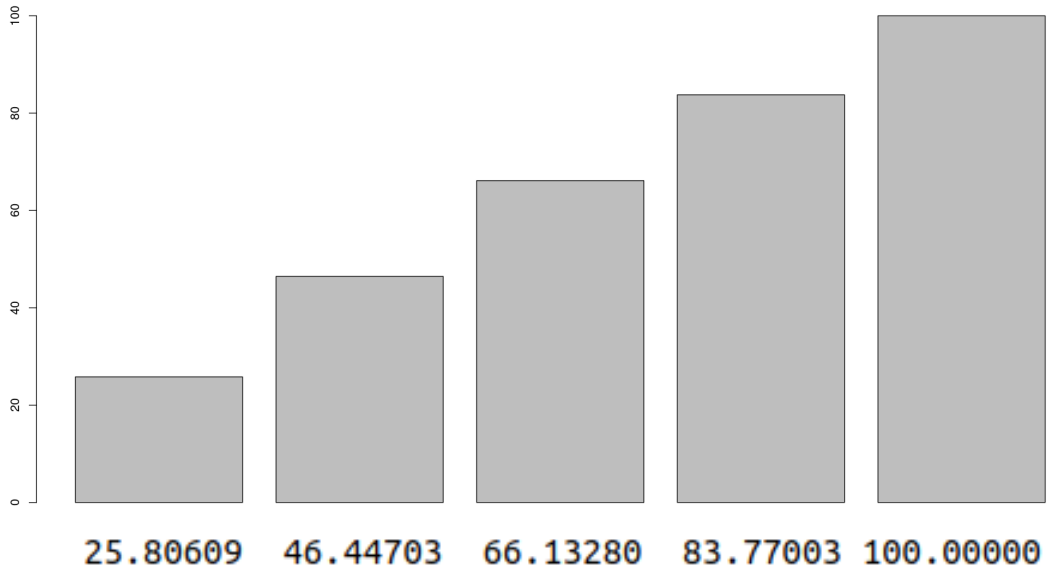


Figura 34. Accumulated Inertia in subspaces, from first principal component to the 11th dimension subspace (Bar plot)

The study of accumulated inertia allows us to understand what proportion of the total variability in the dataset is captured within each subspace generated by the principal components. This analysis is key to determining the number of dimensions needed to adequately represent the data without losing relevant information.

Figure 34 shows the evolution of accumulated inertia from the first component up to component number 11, that is, the proportion of explained variance as components are added. The observed pattern indicates that:

- The first two components already explain around 46% of the variance.
- With three components, 66.1% is reached, and with four, the explained variance reaches 83.8%.
- From the fifth component onwards, the marginal gain in information becomes much smaller.

This behavior justifies the decision to reduce the working space to four dimensions, as it allows for an adequate representation of the original dataset's information with minimal loss of content.

2.3 PROJECTION OF NUMERICAL VARIABLES

Based on the Phi matrix, which contains the correlations between the original variables and the PCA dimensions, the projections onto the plane defined by the first two components are represented.

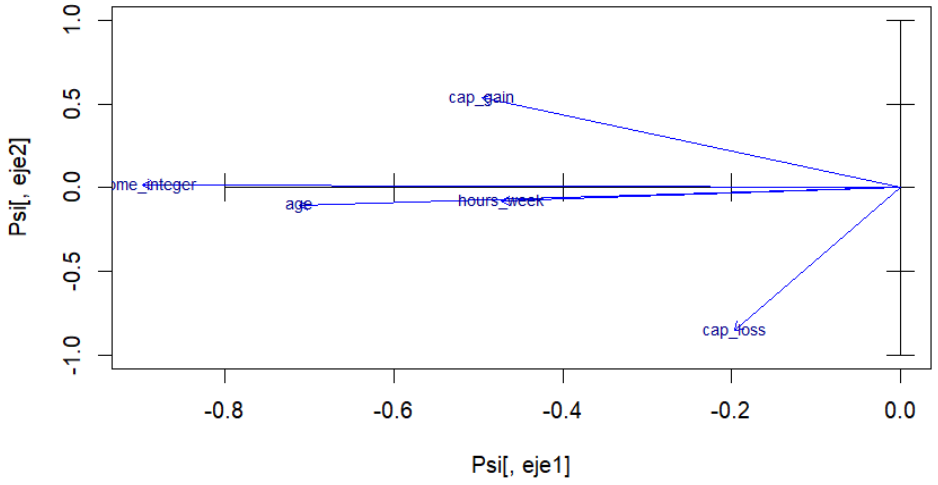


Figure 35. Projection of numerical variables onto the factorial plan

The PCA biplot above visualizes the relationships among the numeric variables in the dataset, projected onto the first two principal components (PC1 and PC2), labeled as eje1 and eje2. These components capture the greatest variance in the data, allowing us to reduce dimensionality while preserving key structure.

The x-axis (PC1) captures most of the variation, as evidenced by the strong horizontal spread of the variable vectors. Variables such as income_integer, age, and hours_week align closely along this axis, indicating they contribute primarily to the first principal component. Their similar direction also suggests a positive correlation among them.

In contrast, cap_gain and cap_loss have strong vertical components. cap_gain points sharply upward while cap_loss points downward, both making substantial contributions to the second principal component (PC2). Their opposing directions indicate a negative correlation between these two variables.

The length of each arrow represents the contribution of the variable to the principal components. Longer vectors such as those for cap_gain and cap_loss indicate a stronger influence on the PCA space, whereas shorter vectors (e.g., income_integer, age, and hours_week) contribute more modestly.

Overall, the biplot reveals two dominant patterns in the data: one along the income, age, and working hours dimension (PC1), and another defined by contrasting capital gains and losses (PC2). This insight can guide further analysis or inform dimensionality reduction strategies.

2.4 SIZE EFFECT (“EFECTE TAMANY”)

When performing Principal Component Analysis (PCA), a common issue that arises is the size effect. This phenomenon occurs when many variables in a dataset are highly positively correlated—often because they all scale with the same underlying factor, such as individual "size", "quantity", or "total amount." In our case, variables such as income_integer, hours_week, and age tend to increase together, leading to high correlations. This causes their corresponding arrows in the biplot to point in the same or very similar directions, potentially masking the underlying structure of the data.

Biplot Interpretation with income_per_hour

To address this issue, we introduced a derived variable, income_per_hour (computed as income_integer / hours_week), which normalizes income by working time. This transformation aims to reduce the size effect by shifting focus from absolute values to a relative measure of economic efficiency or earning rate.

In the updated biplot:

- Most of the original variables (income_integer, age, hours_week, cap_gain, cap_loss) still cluster toward the left side of the first principal component (PC1), indicating they are influenced by the size effect and share a common variance direction.
- In contrast, the new variable income_per_hour stands out, pointing in a clearly different direction, toward the positive side of PC1 and PC2. This suggests that income_per_hour captures a different aspect of variability, independent from the raw size-related factors.

This shift confirms that the new variable successfully mitigates the size effect, offering a more nuanced perspective of the data by highlighting individual efficiency over absolute totals. Including this type of normalization helps uncover structure that would otherwise be obscured by dominant, correlated variables.

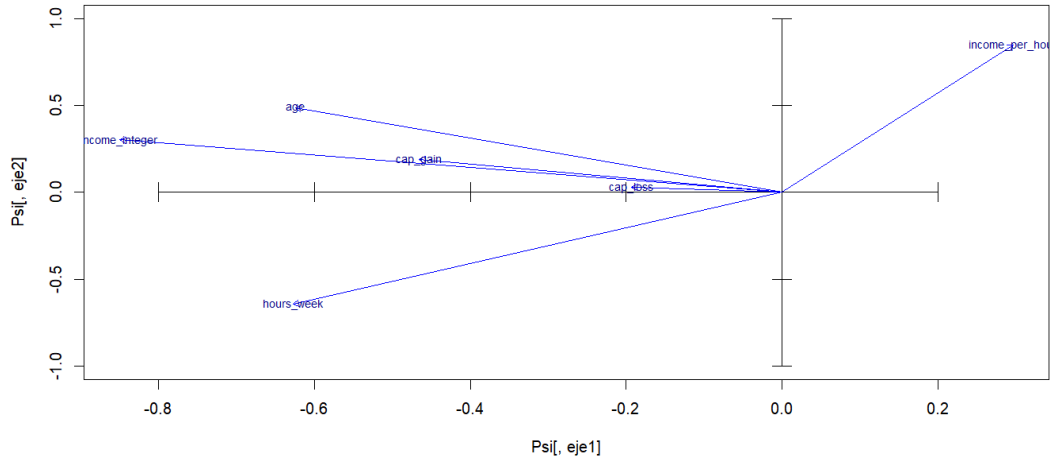


Figure 36. Projection of numerical variables onto the factorial plan

2.5 PROJECTIONS OF INDIVIDUALS

The projection of individuals onto the plane formed by the first two components allows for observation of the overall distribution of the dataset. In this representation, each point corresponds to an individual, and their coordinates reflect their position with respect to the principal dimensions of the new space.

The concentration of most points around the origin suggests that many individuals have average values in the linear combinations of the variables. However, peripheral points are also identified, which may represent atypical or extreme profiles and could be of interest in more specific analyses.

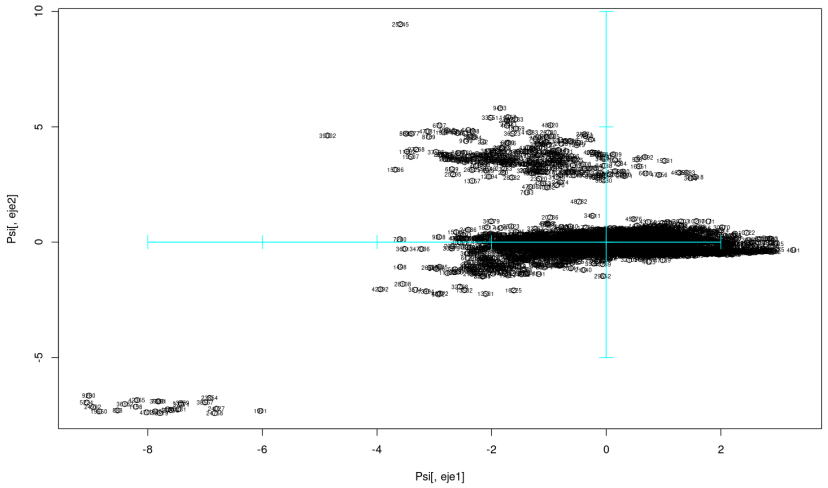


Figure 37. Projection of individuals onto the PC1-PC2 plane

2.6 PROJECTION OF CAP_LOSS VARIABLE AS A QUALITATIVE VARIABLE

The scatterplot below shows the projection of individuals onto the first two principal components (PC1 and PC2), with data points colored according to the cap_loss variable, which has been treated as categorical. Specifically, individuals are grouped by whether they reported a capital loss (cap_loss = 1) or not (cap_loss = 0).

Two distinct clusters emerge:

- The black circles (representing cap_loss = 0) are tightly packed near the upper-central region of the plot.
- The red circles (representing cap_loss = 1) are spread out below, clearly separated along the second principal component (vertical axis, PC2).

This separation along PC2 confirms that cap_loss has a strong influence on the second principal component, aligning with the earlier biplot where the cap_loss vector pointed significantly in the negative vertical direction. The presence of distinct groups implies that capital loss captures important variance that helps differentiate subpopulations in the dataset.

This finding supports the decision to treat cap_loss as a key variable in further modeling or classification tasks. Its strong discriminative power makes it valuable for tasks like income prediction or socio-economic profiling.

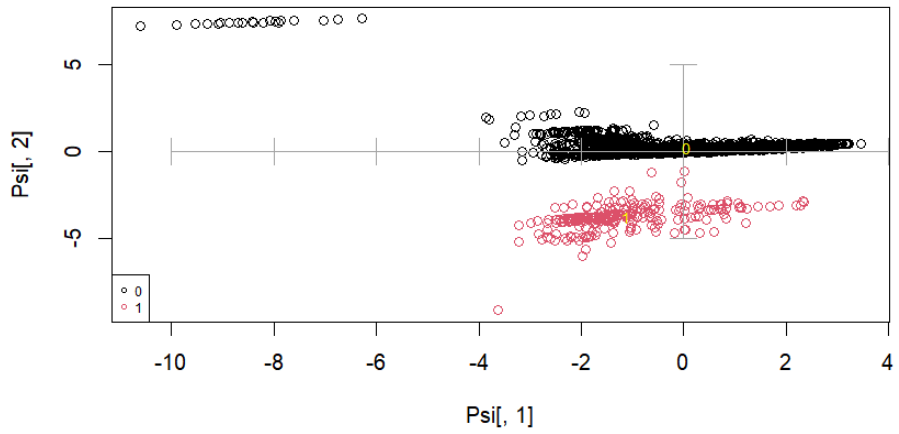


Figure 38. Individuals and centroids by Cap_loss_binary on the PC1-PC2 plane

2.7 PROJECTION OF OTHER QUALITATIVE VARIABLES

Other categorical qualitative variables such as workclass, marital, occupation, relationship, race, and native_country have been projected to observe how their categories are distributed within the PCA space.

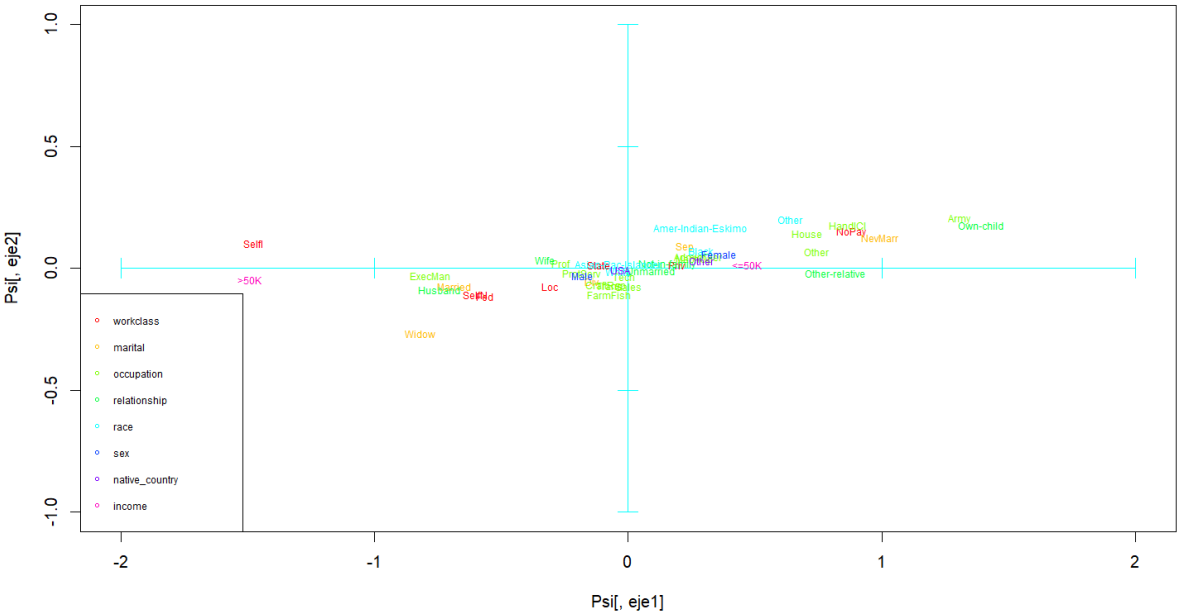


Figure 39. Projection of the categories of qualitative variables

LARGE PHOTO IN THE ANNEX

In Figure 39, each color represents a variable and each label indicates the centroid of one of its categories. The analysis of these positions reveals significant groupings: for example, the categories ExecMan or Self are located in distinct areas, far from the center, indicating particular economic or social profiles. These visualizations help to understand the relationship between qualitative profiles and the numerical components.

The proximity and orientation of the centroids of the different categories within the same qualitative variable reflect the degree of similarity between them. For example, within the variable workclass, the categories No-Pay and Self-employed are located in different areas, suggesting a clear segregation of labor profiles. Similarly, the marital status categories such as Never-married and Married are positioned far apart, associated with distinct patterns of age, job stability, and income.

2.8 COMBINATION OF QUALITATIVE AND NUMERICAL VARIABLES

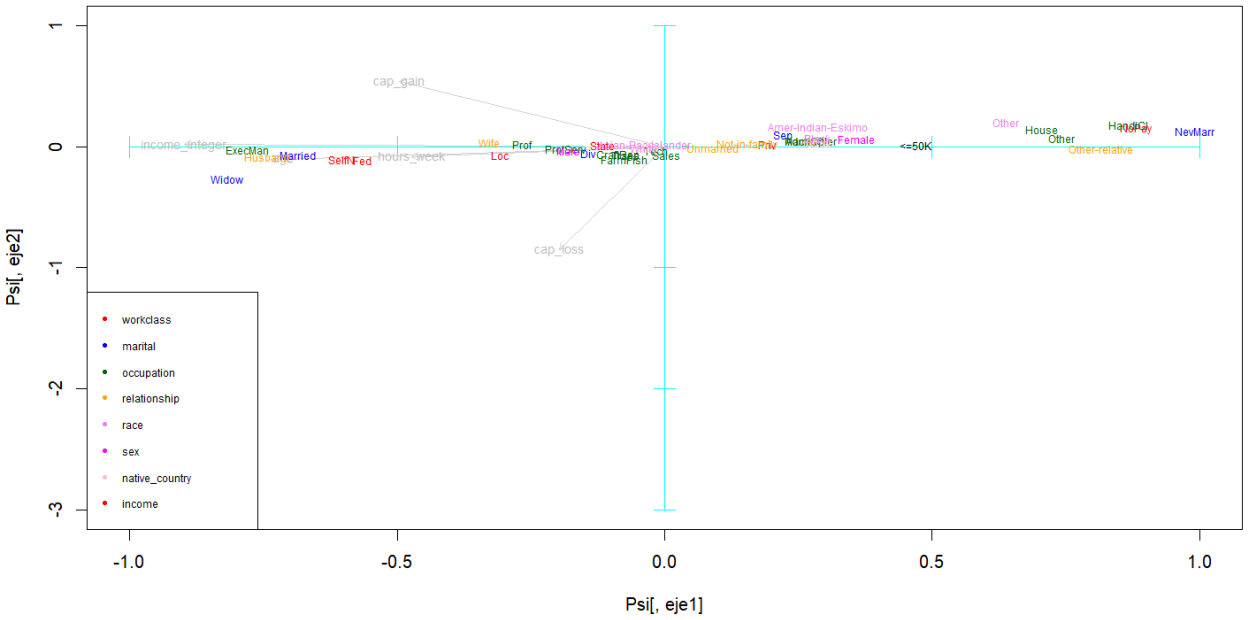


Figure 40. Combination of qualitative and numerical variables

Figure 40 integrates the projection of numerical variables (with grey arrows) and qualitative variables (with colored centroids). This combined representation facilitates the interpretation of which qualitative categories are associated with specific numerical patterns.

For example, the marital categories associated with being married are located closer to vectors representing high work activity (hours_week), while others such as Never-married are projected in areas associated with less experience or lower job stability. These relationships can be very useful for building predictive models or conducting sociological studies.

2.9 CONCLUSIONS

The principal component analysis (PCA) applied to our dataset has revealed a clear and meaningful internal structure in individuals' behavior, both from a sociodemographic and economic perspective. This methodology, focused on dimensionality reduction and the projection of qualitative information onto a space defined by numerical variables, has proven especially useful for identifying common patterns and differences between social groups.

First, the results show that the first four components together explain approximately 84% of the dataset's variability. This confirms that it is possible to work within a reduced space without losing almost any relevant information. This reduction has made it easier to interpret the relationships between individuals and variables more clearly.

Regarding the numerical variables, those that showed the greatest influence were edu_num (educational level) and hours_week (weekly working hours), which dominate the first component. This association suggests that individuals with higher education levels also tend to have longer working hours, likely linked to more demanding or responsible job positions. On the other hand, the variables cap_gain and cap_loss, which represent capital gains and losses, are oriented differently and projected onto later components. This indicates they represent a more independent economic dimension, likely related to financial status or income sources not tied to salaried work. The variable age contributes to several components, acting as a cross-cutting axis that helps identify generational differences and life stages within the dataset.

The projection of individuals onto the factorial plane defined by the first two components revealed a high central concentration, with significant dispersion toward the edges. This pattern suggests the existence of a common profile for most individuals, while also highlighting specific subgroups differentiated by key characteristics such as high education, long working hours, or unusual levels of capital gains.

In this context, the projection of qualitative variables has been particularly insightful. For example, sex shows a visible separation between men and women: while men tend to be located in areas of the plot associated with longer working hours and higher economic gains, women are more concentrated in regions where these figures are lower. Although not extreme, this pattern is consistent and reflects potential structural inequalities that deserve attention.

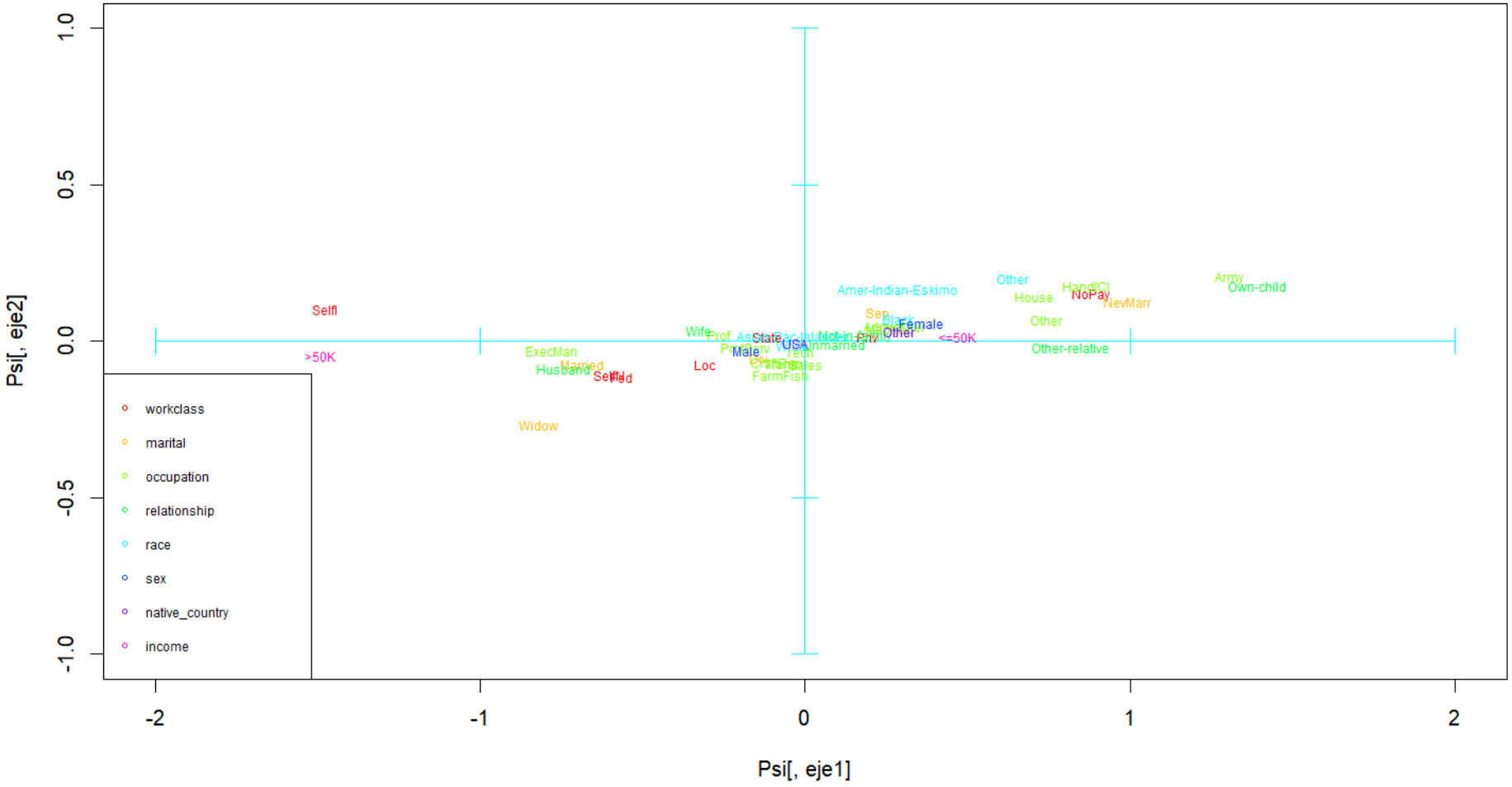
Similarly, classification by occupation and work class has revealed a clear separation of professional profiles. Categories such as ExecMan (executive management) are clearly associated with areas of high work intensity and educational attainment, while others like HouseNoPay or

Without-pay occupy opposite positions, indicating much less economic activity. Clear groupings were also detected based on marital status: for instance, widowed individuals are concentrated in areas of lower economic activity (likely due to age), while those who have never married show a more varied distribution.

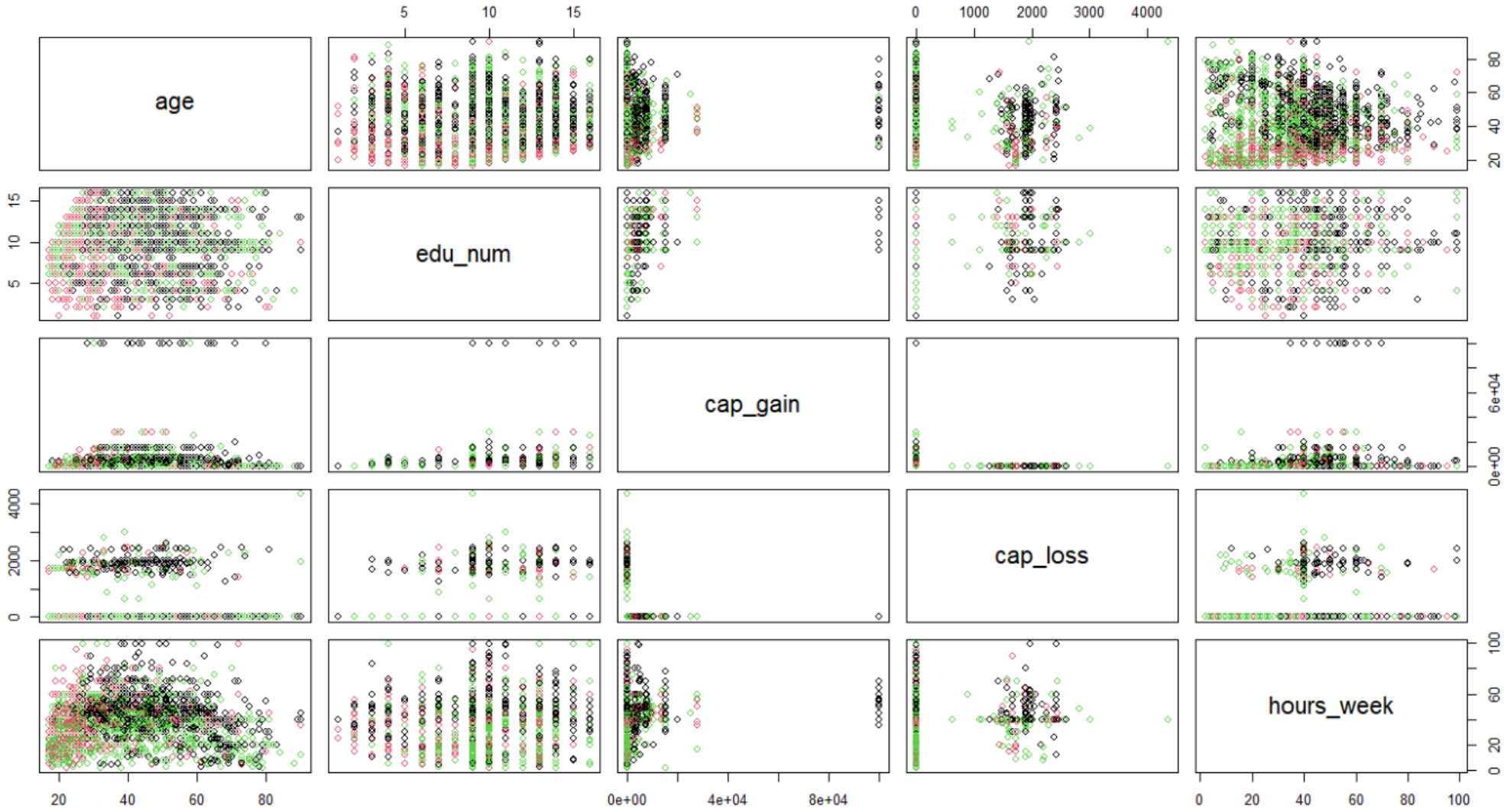
Finally, the combination of qualitative and numerical variables in a single graphical representation has helped synthesize the analysis and add depth. It has become evident how certain qualitative categories are clearly aligned with specific sets of numerical variables. For example, the Self-employed group tends to be located in areas associated with high work dedication but not necessarily high educational levels, while ExecMan is related to both dimensions.

ANNEX

QUALITATIVES PLOT



PAIR PLOT


age
edu_num
cap_gain
cap_loss
hours_week
5
10
15
0
1000
2000
3000
4000
20
40
60
80
0e+00
4e+04
8e+04
0
20
40
60
80
100