Introduction

This project tackles a raw, unlabeled dataset (mystery_dataset_complex.csv) designed to simulate real-world data chaos — with missing values, mixed types, embedded JSON, and no column names.

Our goal is to decode its structure, clean and normalize the data, uncover meaningful patterns through EDA, apply clustering, and build a simple interactive dashboard to present key insights.

1. Column Interpretation and Naming Justification

After taking a bird's eye view of the dataset, meaningful names were assigned to each column:

a. **Name:** Self-explanatory. Retained initially for reference but not used in analysis due to lack of analytical value.

b. **Age:** Numeric value representing each person's age.

c. **Gender:** Self-identified gender, typically categorized as male, female, or other.

d. **Nationality:** Country of origin. Very high-cardinality column.

e. **Continent:** Derived from the nationality column using a country-to-continent mapping to allow continent-level grouping (e.g., Asia, Europe).

f. **City (made-up):** Fictional or anonymized city names. Very high-cardinality column.

g. **Days Since Applying:** Created by cleaning and converting the original date column into datetime format and calculating the duration since application.

h. **Monthly Income:** Possibly family income, as values exist even for unemployed individuals. Cleaned from a mix of JSON-encoded strings and raw numeric values.

i. **Current Employment Status:** Individual's current work status (e.g., employed, unemployed, freelancer).

j. **Most Recent Occupation:** The most recent job title or functional role held. Very high-cardinality column.

k. **Experience (in Years):** Professional experience expressed as numeric years. Takes discrete values in 0-10.

l. **DBMS:** Skills related to database management systems, including categories like SQL, Excel, Python.

m. **Visualization Tools:** Knowledge of tools such as Tableau, Power BI, or R.

n. **Marital Status:** Relationship status (single, married, divorced, etc.).

o. **Phone:** Retained as-is for potential referencing but not useful for modeling.

---

2. Removed Columns

a. **Email:** Dropped due to privacy concerns and high cardinality, making it unsuitable for analysis.

b. **Firm Name:** Removed due to a high rate of missing values and lack of modeling utility. Very high-cardinality column.

c. **Date:** Invalid entries were set to NA and converted to datetime. A new feature, **Days Since Applying**, was derived to provide more analytical value, then the original date column was dropped.

---

3. Binary and Categorical Indicators — Ambiguous Fields

(Note: *ns* stands for "not sure," *sb* stands for "some boolean.")

a. **ns1:** Categorical field with values like A, B, and C; standardized to lowercase.

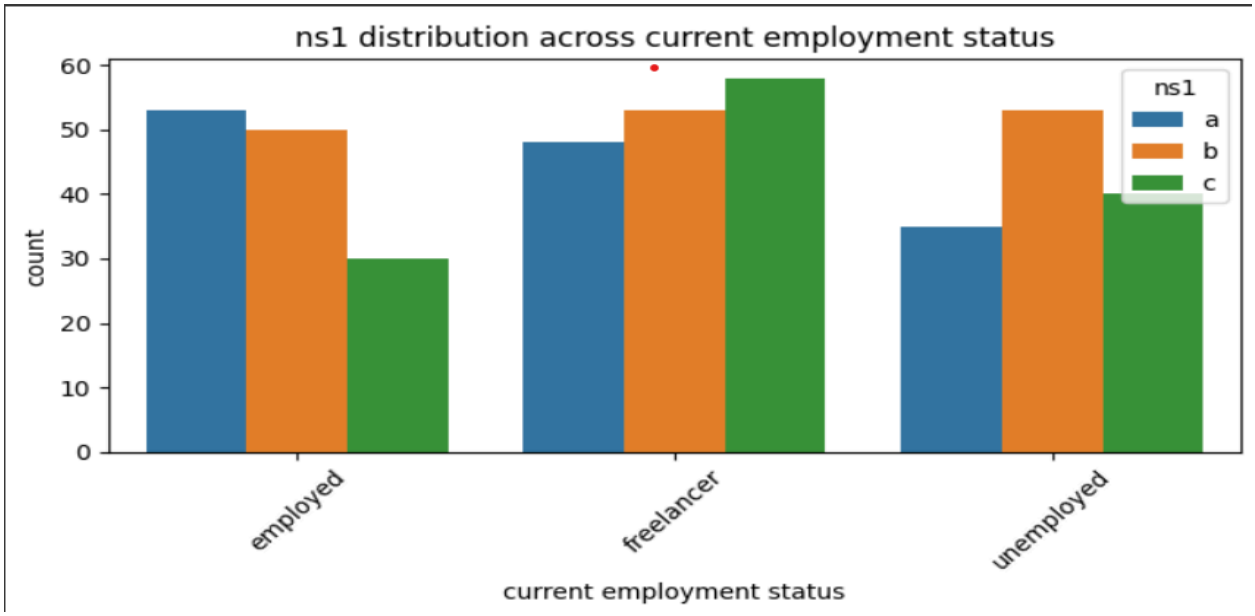
b. **ns2:** Originally encoded as "Y" or "N"; cleaned and converted to binary numeric format (1 or 0), suggesting a yes/no flag.

c. **ns3:** Categories include 001, 002, 003, and abc. The meaning is unclear; retained in original form for further exploration.

d. **sb:** Contains True/False values. Meaning ambiguous; preserved for potential future investigation.

2. EDA : Key Insights

1. Interpreting ns1: Possiblly Academic Stream or Education Level



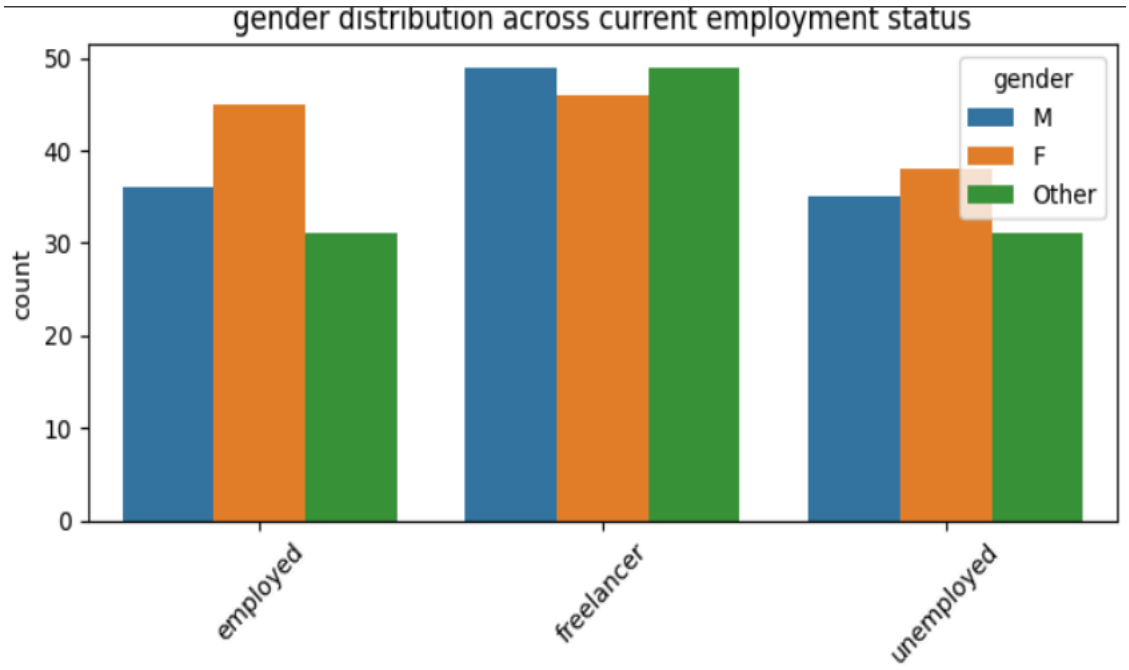
Category **'a'** is most represented among the employed, **'c'** is favored by freelancers, while unemployed individuals are mostly in category **'b'**.
This pattern suggests that ns1 likely represents **academic stream**, with:

- **a** = Science
- **b** = Commerce
- **c** = Arts

Alternatively, ns1 could also correspond to **education level**, where:

- **a** = Undergraduate (UG)
- **b** = Postgraduate (PG)
- **c** = PhD

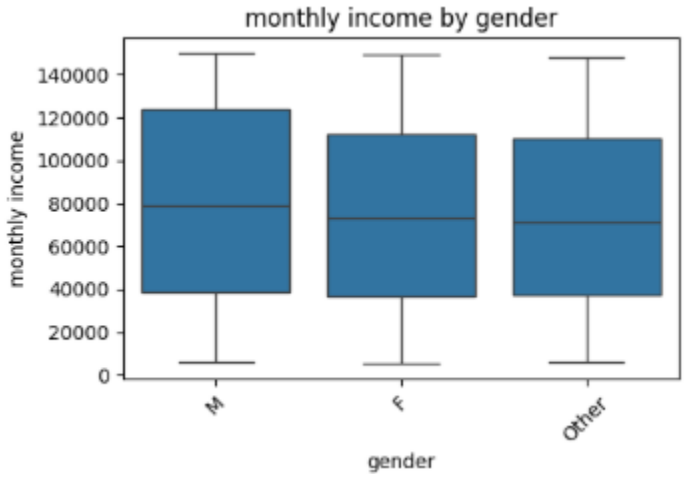
2. Gender Disparities in Employment Status: "Other" Group Leans Toward Freelancing



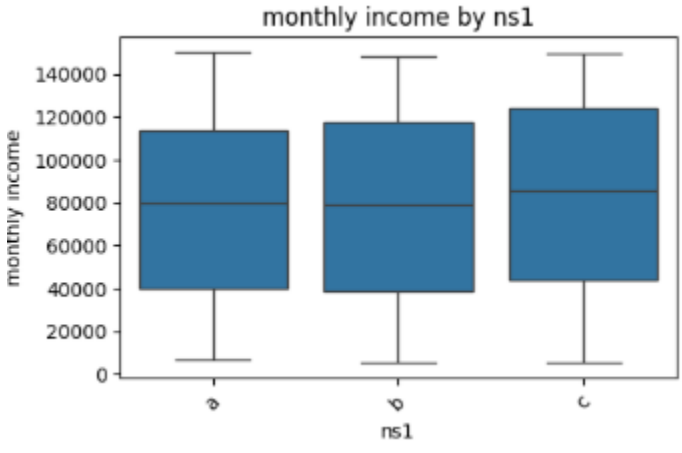
The "Other" Gender Group Has Lowest Representation in Both Employed and Unemployed Categories. This might hint at barriers in formal employment channels and a shift toward freelance work as an alternative.

Monthly Income Insights

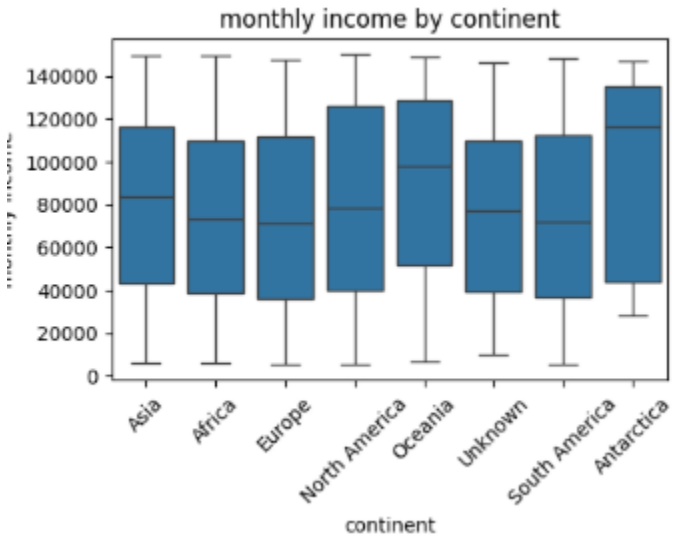
- Median income for **Males** appears slightly higher than **Females** and **Other**.



- Group **'c'** of ns1 appears to have **higher income**. This supports our assumption that category **'c'** might represent individuals from an Arts background, as people pursuing Arts are often from families with higher income levels. Similarly, people who have attained higher education levels like PhD. are also more likely to have higher income.

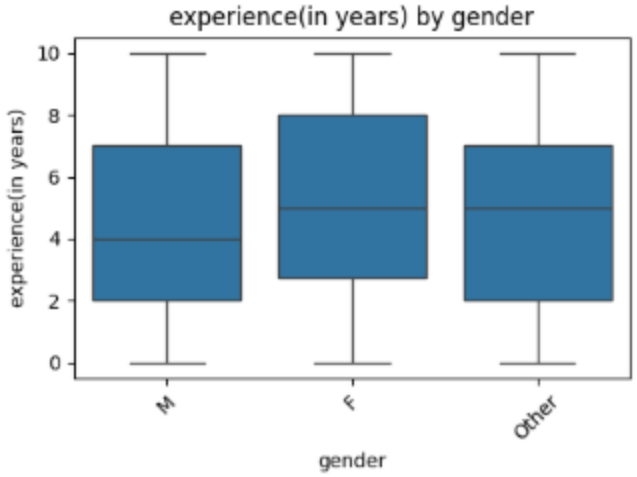


- Oceania and North America exhibit the highest median incomes, while Africa and South America tend to have lower ones. This suggests a strong correlation between continent and earning potential, likely influenced by broader economic conditions and regional disparities.

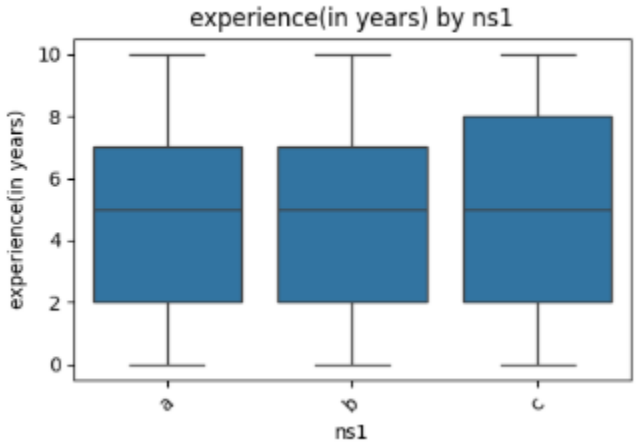


Experience Insights

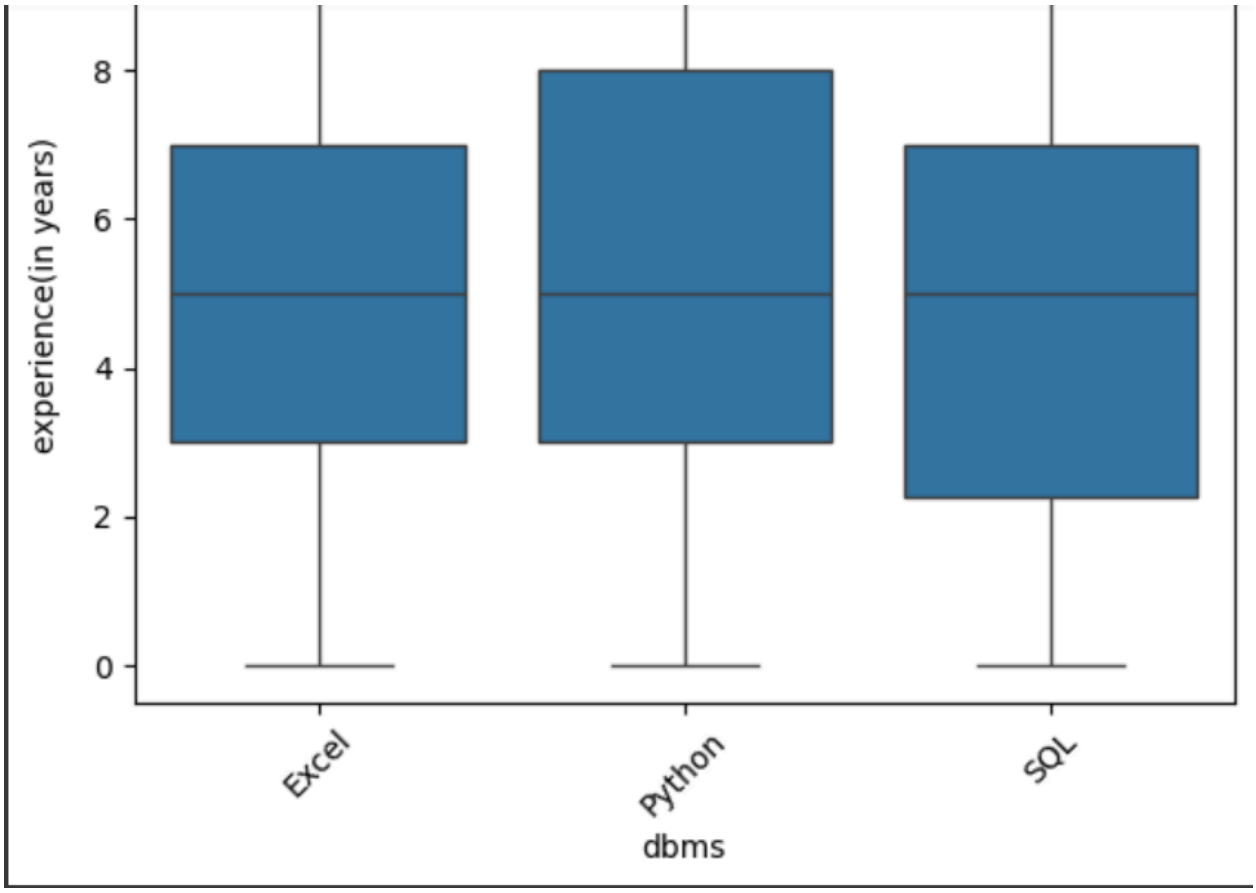
- Females have slightly higher median experience than males. Could indicate longer tenure or retention among females in the dataset.



- Group 'c' again shows higher experience levels. Suggests group 'c' may represent senior roles, certifications, or education level.

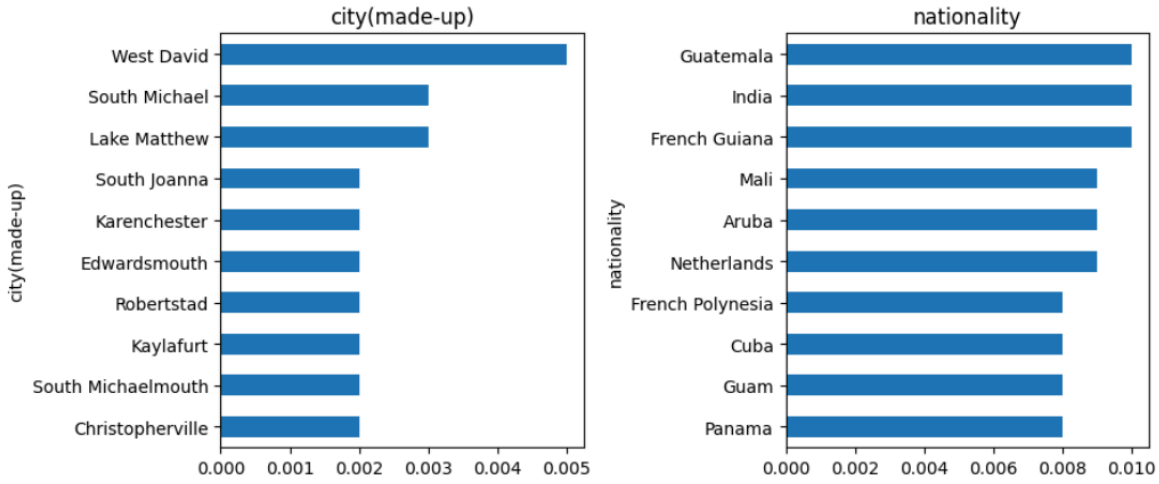


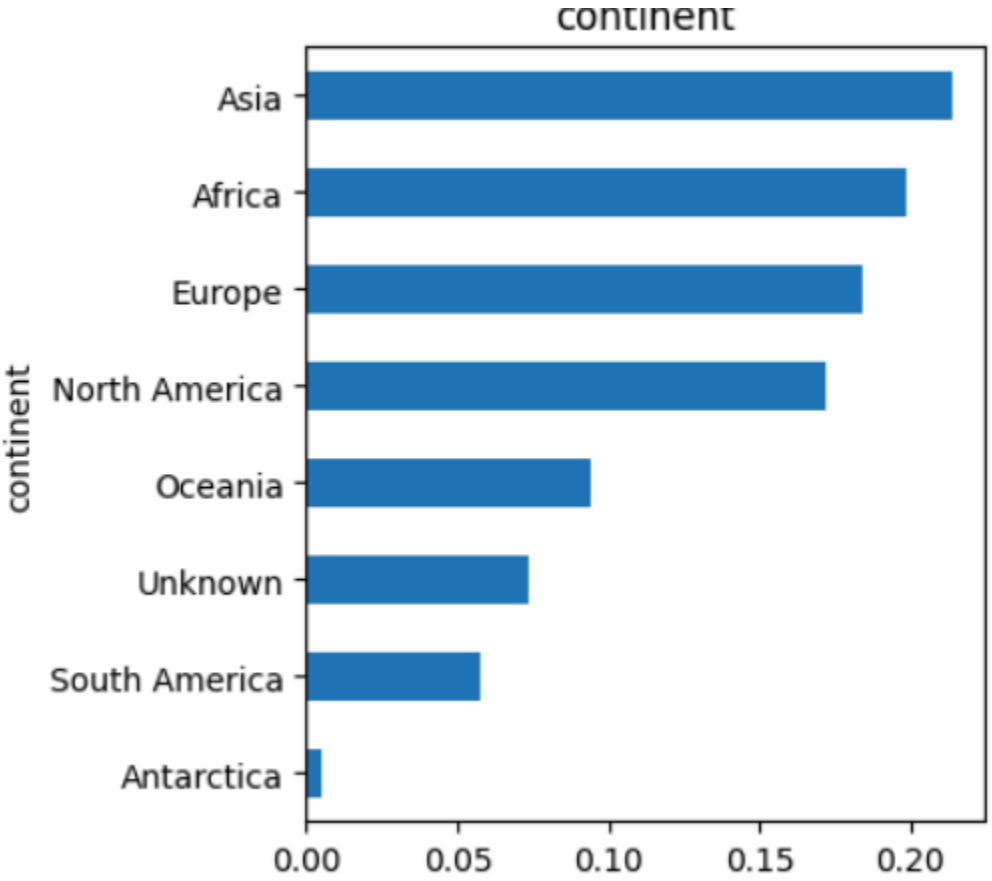
6. Which tool do experienced people use?



Python users have a slightly higher upper quartile, suggesting more experienced individuals tend to use Python.

7. Which locations dominate in the data?


city(made-up)
West David
South Michael
Lake Matthew
South Joanna
Karenchester
Edwardsmouth
Robertstad
Kaylafurt
South Michaelmouth
Christopherville
0.000
0.001
0.002
0.003
0.004
0.005
nationality
Guatemala
India
French Guiana
Mali
Aruba
Netherlands
French Polynesia
Cuba
Guam
Panama
0.000
0.002
0.004
0.006
0.008
0.010


continent
Asia
Africa
Europe
North America
Oceania
Unknown
South America
Antarctica
0.00
0.05
0.10
0.15
0.20

3. Imputation of Missing Values

- Numerical columns like **age, application_year, and days_since_applying** were imputed using KNN imputation to estimate values based on similar records.

- Categorical columns such as **ns1, ns3, gender, marital status, current employment status, dbms, and visualization tools** were filled with a "missing" label to retain information without introducing bias.

4. Preparation of dataset for clustering.

Step 1: Removed high-cardinality columns such as name, phone, city, and nationality.

Step 2: The *most recent occupation* column was transformed into a low-cardinality categorical feature by:

- Generating sentence embeddings using the all-MiniLM-L6-v2 model,

- Applying KMeans clustering (k=8) to group similar occupations,

- Interpreting each cluster with TF-IDF to identify dominant terms,

- Mapping clusters to interpretable occupation categories.
- The original column was then dropped.

Step 3: Dropping Low-Impact Features

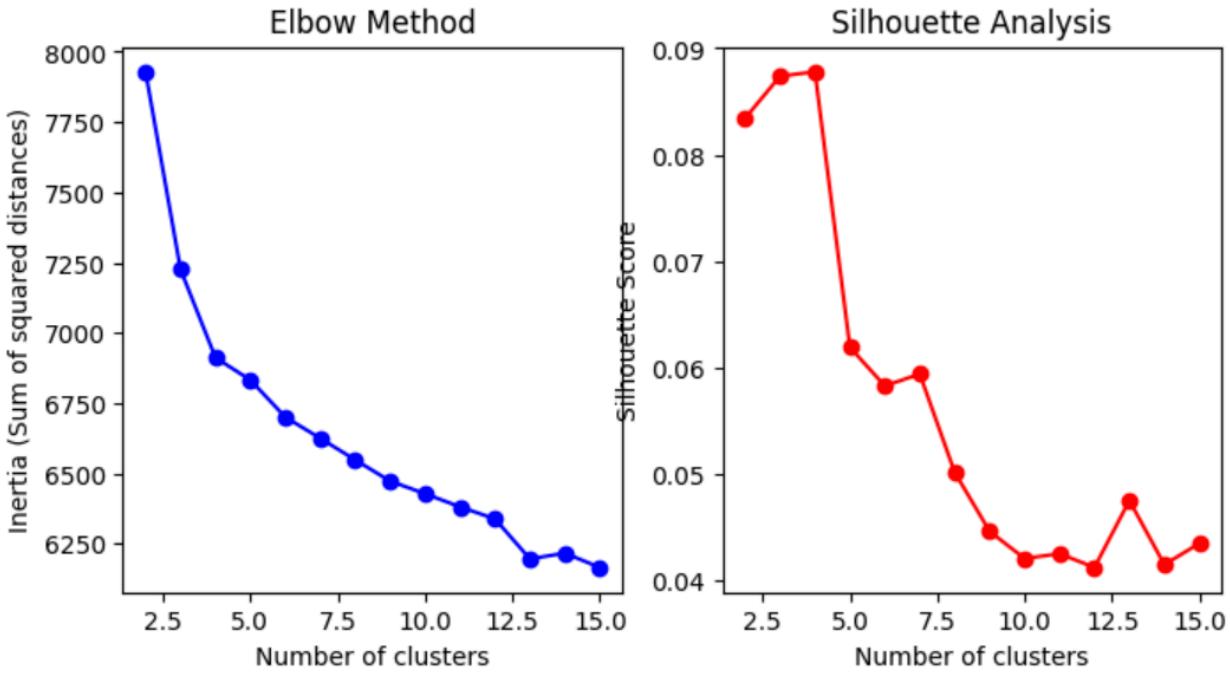
Based on insights from exploratory analysis, the following columns were dropped due to low variance or poor clustering value:

- **age** and **days_since_applying**: had high missingness and contributed minimally to clustering.

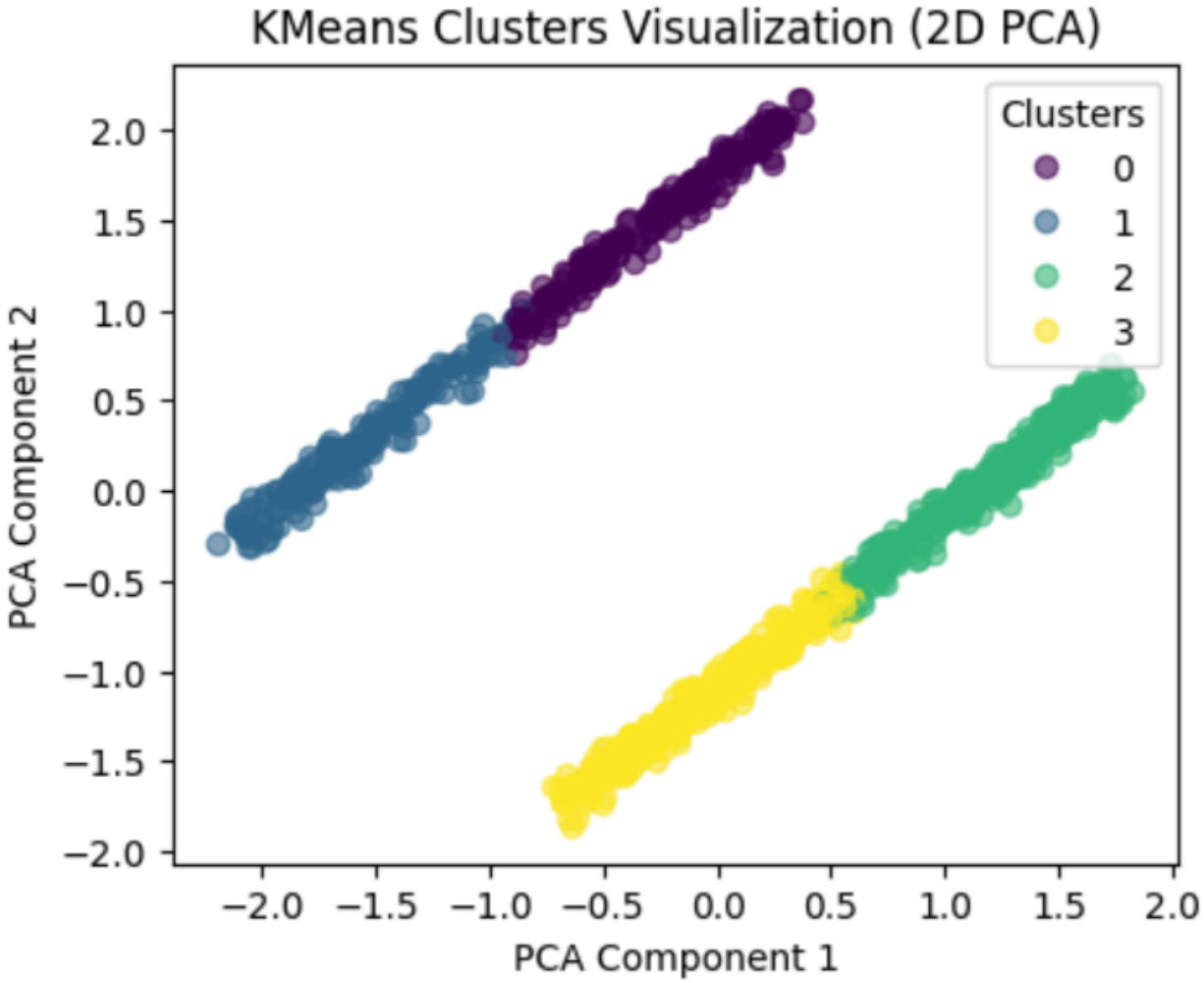
- **sb**: lacked interpretability and meaningful variation across profiles.

5. Clustering and Visualization

- To determine the optimal number of clusters, the Elbow Method and Silhouette Analysis were applied across values of k from 2 to 15. Both plots indicated that 4 clusters provide a balance between compactness and separation.



- A complete pipeline combining preprocessing and KMeans clustering with k = 4 was then constructed, and cluster labels were assigned to each profile.
- For visualization, PCA was used to reduce the dataset to 2 dimensions, and the clusters were plotted. The resulting plot revealed reasonably distinct groupings, supporting the quality of the clusters.



6. Interactive Dashboard

To make the clustered data actionable and accessible, we developed an interactive dashboard using **Streamlit** and deployed it via **GitHub Codespaces**.

Click to run the app: Cluster Dasboard

Click for dashboard pre-view: Preview

7. Assumptions made:

1. The dataset represents applicant data either within an organization, a program, or a general survey.
2. The monthly income feature could also be family income and it's in dollars across all cities and nationalities.
3. The cities/locations are hypothetical.