BUSINESS INTELLIGENCE 381 GROUP D

MILESTONE 2: DATA PREPARATION

CONTENTS

DATA DESCRIPTION

INTRODUCTION

The dataset used in this project contains various attributes related to personal and financial information. These attributes include details about individuals' demographics, marital status, education, occupation, and financial metrics such as salary and gross pay. This report outlines the data preparation and cleaning steps undertaken, including column selection, data splitting, handling missing values, imputation, and outlier detection.

DATA SELECTION

The dataset originally contained a combined column with multiple fields, including information such as year of birth, marital status, street address, postal code, and more. To make the dataset easier to analyze, we split this combined column into separate fields. After splitting, the following key columns were retained for further analysis:

1. **Year of Birth** (year_of_birth): Numeric values indicating the year in which an individual was born.
2. **Marital Status** (marital_status): Categorical data showing the current marital status of individuals.
3. **Education** (Education): Categorical data indicating the highest education level attained.
4. **Occupation** (Occupation): Categorical data describing the individual's current occupation.
5. **Annual Salary** (Annual.Salary): Numeric data representing the annual salary of individuals.
6. **Gross Pay Last Paycheck** (Gross_Pay_Last_Paycheck): Numeric data showing the gross pay from the last paycheck.
7. **Gross Year to Date** (Gross_Year_To_Date): Numeric data showing the gross earnings for the year to date.
8. **Gross Year to Date FRS Contribution** (Gross_FRS_Contribution): Numeric data showing the gross year-to-date contributions to the Florida Retirement System (FRS).
9. **Household Size** (household_size): Numeric data representing the size of the household.
10. **Years of Residence** (yrs_residence): Numeric data indicating how many years an individual has been living at their current residence.

DATA CLEANING PROCESS

1. SPLITTING COMBINED COLUMNS

The first step in the data preparation process involved splitting a combined column that contained multiple fields. After splitting, we retained only the relevant fields for analysis (as listed above). The original combined column was subsequently removed from the dataset.

2. HANDLING MISSING VALUES AND IMPUTATION

In several columns, there were missing values that needed to be addressed. Here is how they were handled:

MARITAL STATUS

We cleaned the marital_status column by standardizing various terms for clarity:

- Replaced variations like "Divorc", "separ.", and "Divorc(ed)" with Divorced.
- Replaced variations like "mar-af", "mabsent", and "married" with Married.
- Replaced "widow" with Widow.
- Combined "NeverM" and "single" into a unified category Single.
- Missing values were replaced with the most common value (mode), that was Unknown.

EDUCATION

We cleaned the Education column by removing any invalid data (e.g., email addresses) and replaced missing values with the mode that represented the most common education level.

OCCUPATION

Values that should not be present in the Occupation column (such as "Bach", "HS-grad", or "Masters") were removed, and missing values were imputed with the mode (most common occupation).

HOUSEHOLD SIZE

Missing values in the household_size column were imputed using the median value to prevent the impact of extreme values on the dataset.

3. NUMERIC CONVERSION AND DATA TRANSFORMATION

Several columns were converted from character to numeric format to ensure they could be used in further analysis:

- Annual.Salary, Gross_Pay_Last_Paycheck, Gross_Year_To_Date, Gross_FRS_Contribution, yrs_residence, and household_size were all converted to numeric after removing any commas.
- A new column called Age was created by subtracting the year_of_birth from the current year. An age_group column was also added, classifying individuals into three categories based on their age:

✓ **Youth** (16-24 years old)

✓ **Adult** (25-64 years old)

✓ **Senior** (65+ years old)

OUTLIER DETECTION AND HANDLING

To ensure the quality of the analysis, outliers were detected and handled across multiple numeric columns. Outliers were identified using the Interquartile Range (IQR) method, which calculates the first quartile (Q1), third quartile (Q3), and the IQR (Q3 - Q1). Data points outside the range of [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] were considered outliers.

Outliers were removed or capped in each of these columns to prevent them from skewing the analysis.

CORRELATION ANALYSIS

As part of our data exploration, we generated a correlation matrix to better understand the relationships between key numerical variables in the dataset. The color scale ranges from dark blue (strong positive correlation) to white (no correlation) to light blue (negative correlation). The matrix provided the following insights:

STRONG POSITIVE CORRELATIONS (DARK BLUE)

- **Gross_Year_To_Date with Annual Salary and Gross_FRS_Contribution**: There is a strong positive correlation between these three financial variables. This relationship indicates that as **Annual Salary** increases, so do **Gross_Year_To_Date** earnings and **Gross_FRS_Contribution**, which is logical given that these variables reflect overall financial compensation.

- **Annual Salary with Gross_FRS_Contribution**: Like the above, **Annual Salary** is strongly positively correlated with **Gross_FRS_Contribution**, suggesting a clear link between salary and FRS contributions.

MODERATE POSITIVE CORRELATIONS

- **Age with yrs_residence**: A moderate positive correlation between **Age** and **yrs_residence** indicates that as individuals age, they are more likely to have lived longer in a particular location, which could be due to a natural tendency to settle as people grow older.

WEAK TO MODERATE NEGATIVE CORRELATIONS

- **household_size with Age and yrs_residence**: There is a weak to moderate negative correlation between **household_size** and **Age**, as well as **yrs_residence**. This suggests that older individuals or those who have lived in one place for a longer period may have smaller household sizes, due to grown children leaving home.
- **Age.1 with household_size**: There is a negative correlation between **Age.1** and **household_size**. The presence of two age variables (Age and Age.1) may indicate an error or duplication in the data, which should be investigated further.

NO APPARENT CORRELATION (WHITE CELLS)

- **Gross_Year_To_Date, Annual Salary, Gross_FRS_Contribution with Age, household_size, or yrs_residence**: There is no significant correlation between the financial variables and demographic

data such as **Age**, **household_size**, or **yrs_residence**. This suggests that earnings do not strongly depend on age, the size of the household, or how long a person has lived in one place.

PERFECT CORRELATION (DIAGONAL)

The diagonal of the matrix shows perfect correlation (dark blue), indicating that each variable correlates perfectly with itself, as expected.