**Prashanthi Akuthota**

**MID-TERM Assignment**

**CS-527 Data Science**

**Exploratory Questions**

1. **What does each column in the dataset represent?**

* The structure of the dataset (str(my\_data)) will help identify **categorical vs. numerical** variables. If there are unnamed or unclear column names, renaming them can help:

names(my\_data) <- "New\_Column\_Name".

* If some columns contain categorical data that need labels, we should recode them.

**What is the overall structure of the data?**

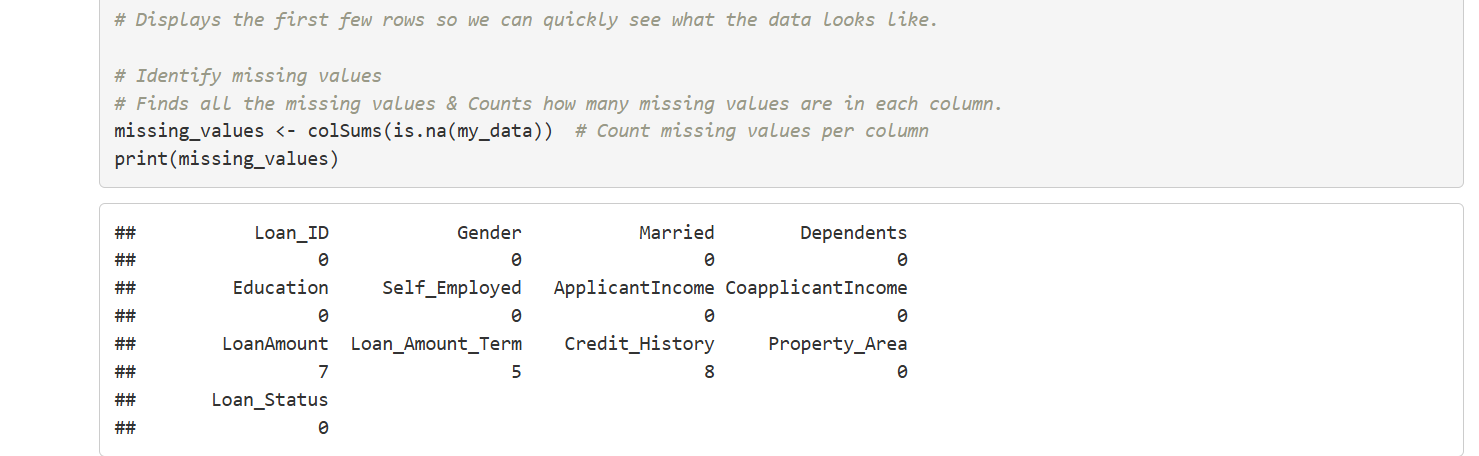
* Running overall structure (my\_data) gives a **high-level summary**, including:
  1. Mean, median, min/max for numerical columns.
  2. Frequency distribution for categorical columns.
  3. Identifies missing values and outliers.
* Overall structured\_stats calculated:

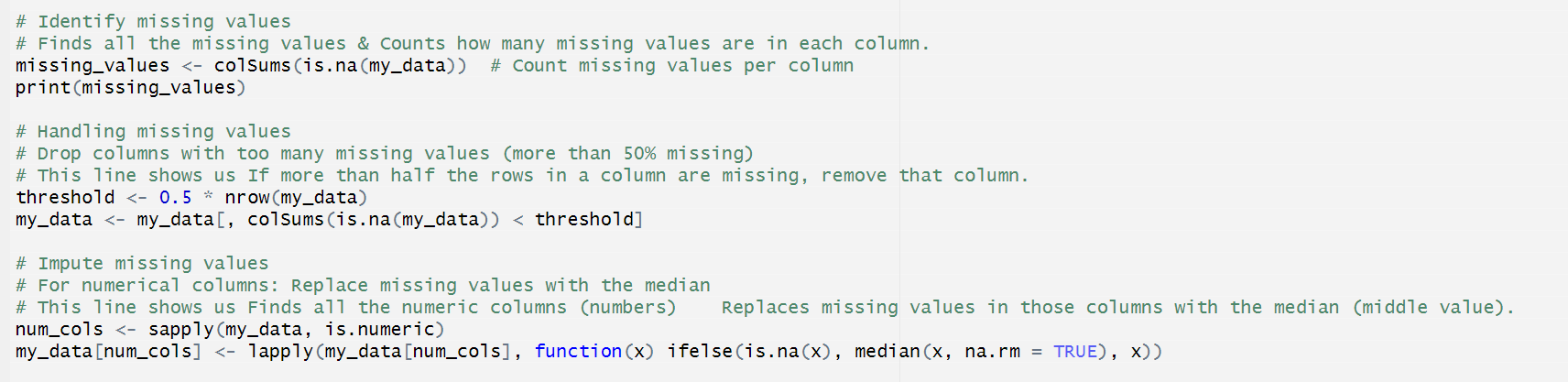
1. **Mean & Median**: Identify skewed distributions.
2. **Min & Max**: Show the data range.
3. **Standard Deviation**: Indicates variability.
4. **Q1 & Q3**: Identify potential outliers.

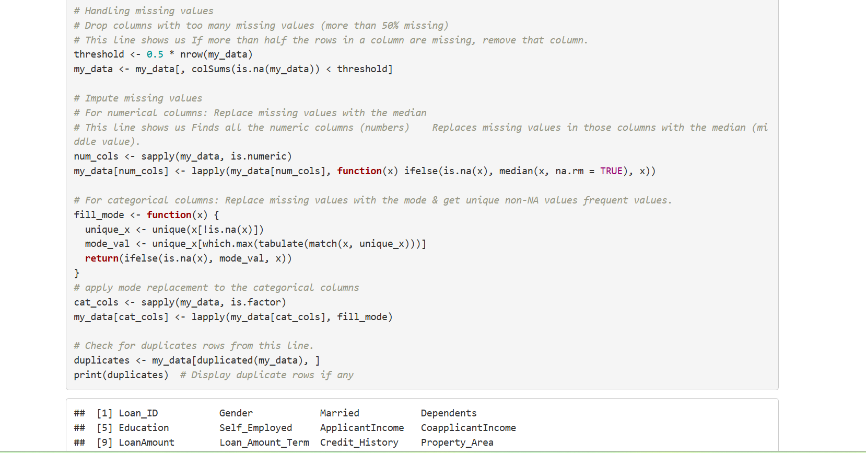
* If numerical columns have large differences between **mean & median**, they may be twisted.

**2️) Data Cleaning**

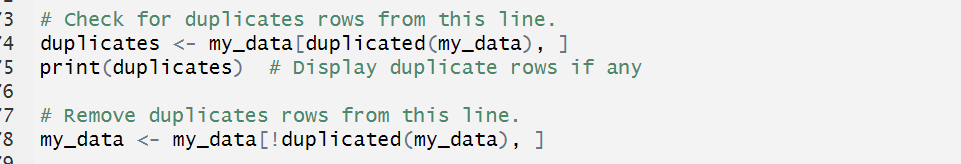
**Were there any missing values or duplicates in the dataset? How did you handle them?**

* **Missing Values**:
  1. Columns with **>50% missing values** were dropped:
  2. Remaining missing values can be handled via:
  3. This replaces missing numerical values with **median**.
* 





**Duplicates**:

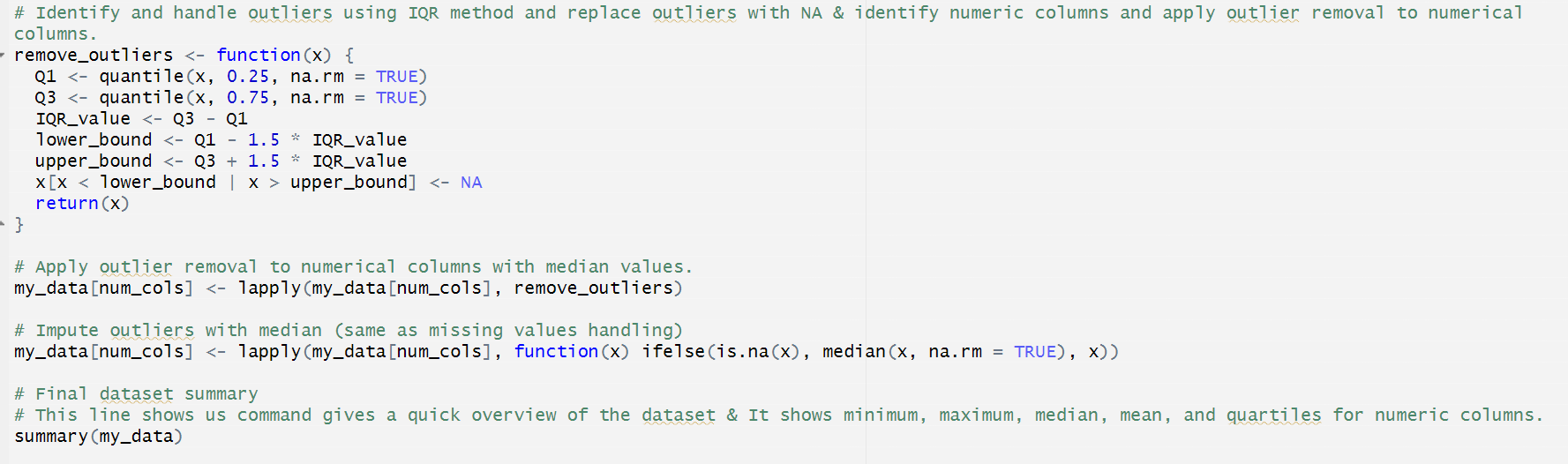


Removes duplicate rows.

**Did you identify any outliers? How did you decide to handle them?**

Outliers are identified by the using boxplots and handling through the Winsorizing and later removed outliers.

**Outliers can be detected using boxplots**:



**Handling Outliers**:

* + **Winsorizing** :

my\_data <- my\_data %>%

mutate(across(where(is.numeric), ~ ifelse(. > quantile(., 0.99, na.rm = TRUE),

quantile(.,0.99, na.rm = TRUE), .)))

* + **Removing outliers**:

my\_data <- my\_data %>% filter(var1 < 100)

**3️) Visualization**

**What do the histograms and boxplots tell you about the distribution of your variables?**

**Histograms:**

* 1. Identify **normal/skewed** distributions.
  2. Identify **bimodal distributions** (multiple peaks).

**2.Boxplots**:

a. Identify **outliers** (points beyond whiskers).

b. Compare disbtributions across categories.

**Correlation matrix (corrplot)**:

* 1. Strong **positive/negative correlations** indicate relationships.

**Pairwise scatter plots (ggpairs)**:

ggpairs(my\_data[, num\_cols])

* + 1. Helps detect **linear relationships**.

**4️) Statistical Analysis**

**What hypothesis tests did you use, and why did you choose them?**

Depending on the data, common statistical tests include:

* **T-test (Compare two groups' means):**

t.test(my\_data$var1 ~ my\_data$group\_var)

Used when testing whether two groups (e.g., male vs. female) have significantly different means.

**ANOVA (Compare means across >2 groups):**

aov\_result <- aov(var1 ~ group\_var, data = my\_data)

summary(aov\_result)

Used for categorical variables with more than two groups.

* **Chi-Square (Test independence of categorical variables):**

chisq.test(table(my\_data$cat\_var1, my\_data$cat\_var2))

**What were the results of the tests? What do they tell you about the data?**

* **T-test Results:**
  + p-value < 0.05: Statistically significant difference between groups.
  + p-value > 0.05: No significant difference.
* **ANOVA Results:**
  + p-value < 0.05: At least one group is significantly different.
* **Chi-Square Results:**
  + p-value < 0.05: Variables are **dependent**.
  + p-value > 0.05: No relationship between variables.

**5️) Conclusions & Recommendations**

**What insights have you gained from the analysis?**

* **Data distribution**:

If many features are **skewed**, transformations (e.g., log transformation) may be needed.

**Missing values**:

**effectively** to avoid bias in analysis.

* **Outliers**:

Identified and removed **when necessary** to improve model performance.

* **Variable relationships**:

Strong correlations could indicate **redundant features** (can be removed).

Weak correlations suggest **no direct relationships**.

**I would suggest:(depending upon my data)**

1. **Feature Engineering**:
   1. Create new features from existing ones (e.g., Revenue = Price \* Quantity).
   2. Encode categorical variables properly for machine learning models.
2. **Predictive Modeling**:
   1. Run **regression/classification models** to identify key predictors.
   2. Use **Machine Learning (ML)** techniques to improve predictions.
3. **Deeper Statistical Analysis**:
   1. **Regression Analysis** for numerical relationships.
   2. **Time Series Analysis** if the dataset has time-dependent trends.