****

**Module 2 Final Project**

**Milestone 1: EDA Code Walthrough**

**ALY6040: Data Mining Applications**

Date: April 17, 2024

Group 5

Team Members: Nidhi Jeetesh Patel

Yukang Lin

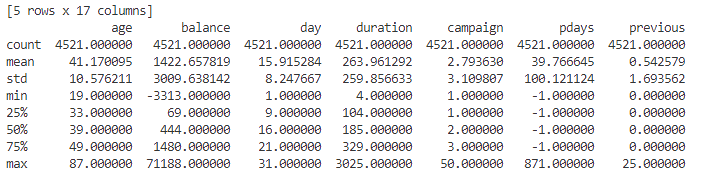
Ramish Fatima

Instructor: Hema Seshadri Ph.D

**INTRODUCTION**

**DESCRIPTIVE STATISTICS**

The numerical columns provide a statistical summary including count, mean, standard deviation, min, and max values, as well as quartiles.

**age** appears to be normally distributed with a mean of approximately 41 and a standard deviation of around 10.6.

**balance** has a wide range and is right-skewed, with a mean significantly higher than the median (444), and the presence of negative values could be considered for further investigation.

**day**, **duration**, **campaign**, **pdays**, and **previous** also show right-skewness.

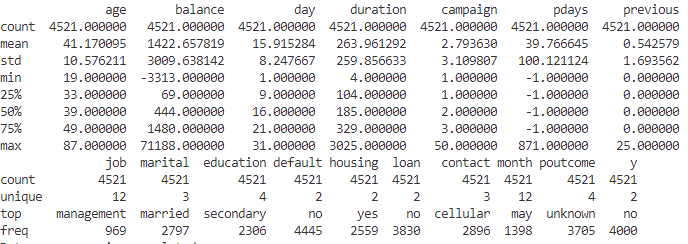
**pdays** has a large number of **-1** values, which may indicate missing or not applicable data in the context of the dataset.

**Categorical Variables**

1. Categorical variables include job types, marital status, education level, and others.
2. management is the most common job type, married is the most common marital status, and secondary is the most frequent education level.
3. default, housing, loan, contact, month, outcome, and y have binary or limited unique values, they could be encoded as binary or ordinal features for modeling.

**Potential Issues:**

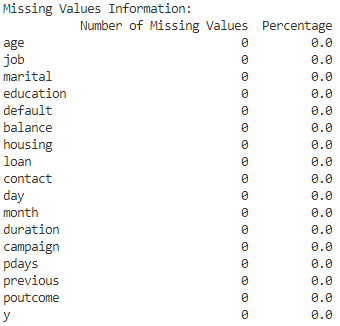
1. Negative balances might need to be investigated.
2. The -1 values in pdays could be placeholders for 'not previously contacted'; we might need to handle this.
3. High standard deviation relative to the mean in columns like balance, duration, and pdays suggests the presence of outliers or wide variation.



**DATA CLEANING**

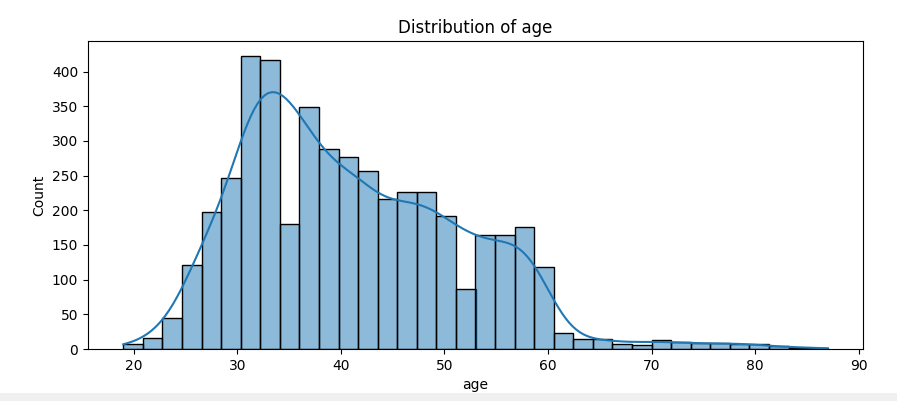
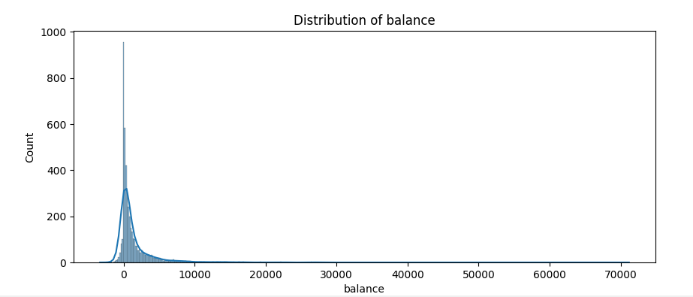
**Section#1 : Handling Missing Values**

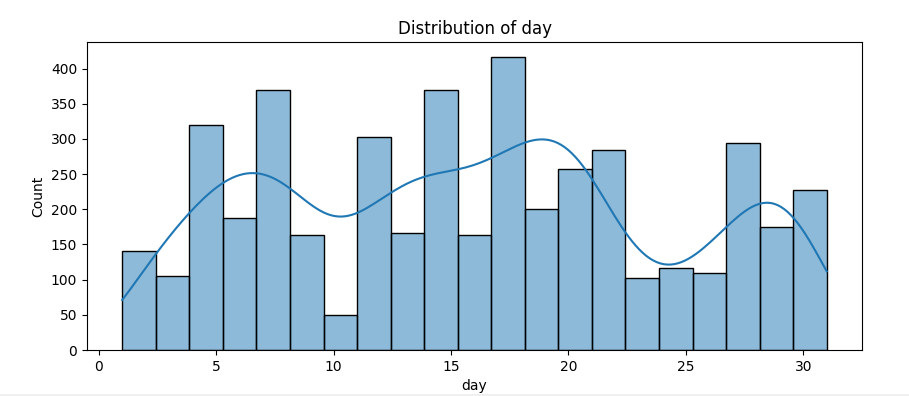
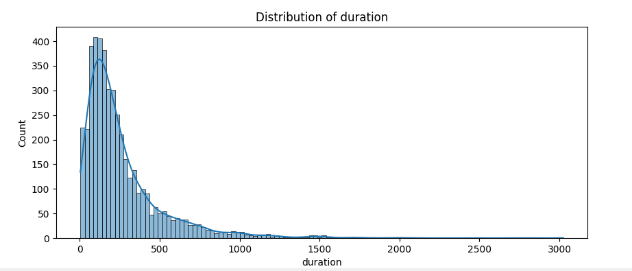
**Output:** There are no missing values in the dataset as the count for each column matches the number of entries and the missing values percentage is 0%.

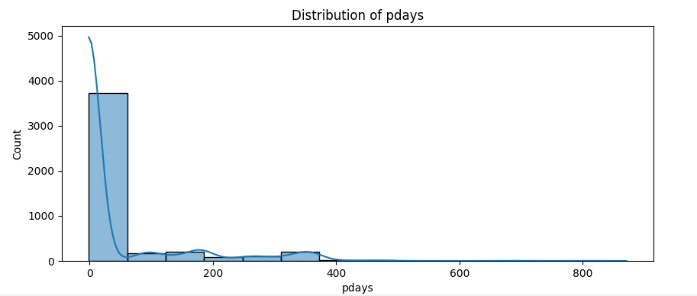
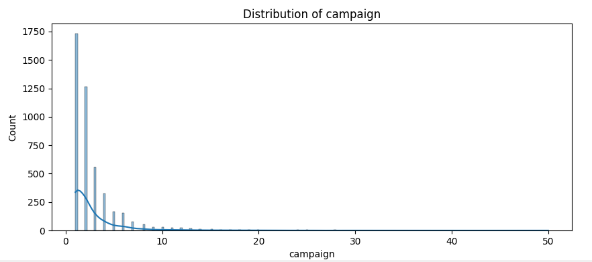


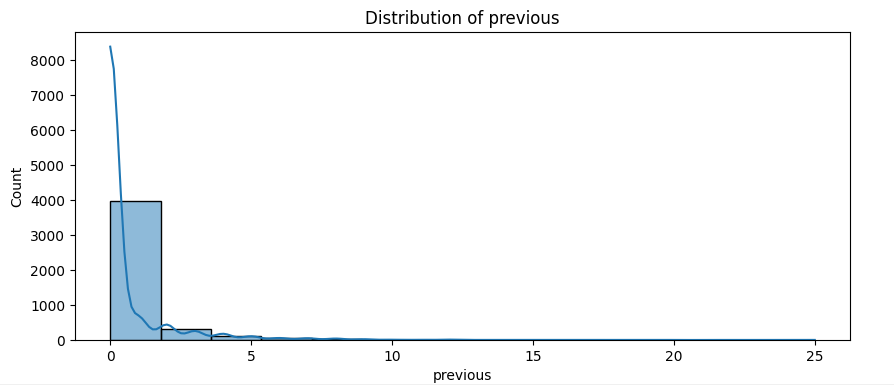
**Section#2 : Outlier Detection**

**Method 1:** Visualization of distributions of numerical features









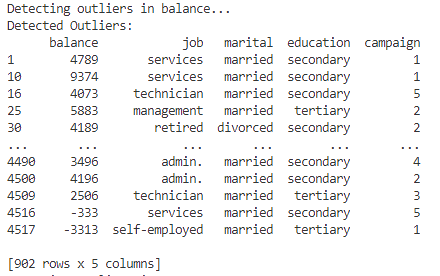
**ANALYSIS:**

|  |  |
| --- | --- |
| **Age** | The distribution of age is somewhat normally distributed, with a slight right-skew. Most clients are in their 30s to 50s. There are no obvious outliers from this histogram alone. |
| **Balance** | The balance is heavily right-skewed, with most clients having a balance in the lower range, while a few have very high balances. This right-skew indicates the presence of outliers towards the higher end. |
| **Day** | This appears to be the day of the month. The distribution is somewhat uniform with slight peaks, which may correspond to specific days when marketing campaigns are more active. |
| **Duration** | Duration of contact is also right-skewed, indicating that most calls are short, but there are a few very long calls, which could be potential outliers. |
| **Campaign** | The number of contacts performed during this campaign for a client is heavily right-skewed, with most clients being contacted a few times, and a small number being contacted many times, which may be outliers. |

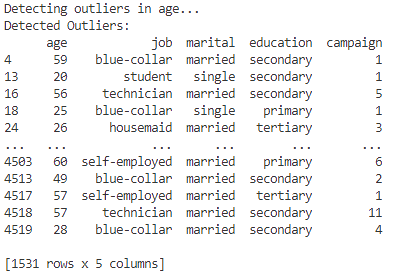
For 'balance', 'duration', and 'campaign', since these are heavily skewed, a small percentage of very high values are common and these can be considered outliers.   
For such skewed distributions, rather than using a standard IQR method, we will use other methods that are less sensitive to extreme values, such as:

**Isolation Forest:**

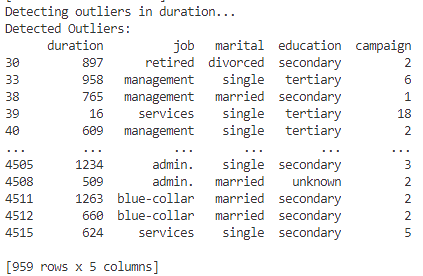
1. **balance**:
   * **902 outliers detected**: These include accounts with unusually high or low balances compared to the norm, which indicate legitimate extremes (e.g., very wealthy or indebted clients).



1. **age**:
   * **1531 outliers detected**: Age outliers include very young or old customers relative to the typical banking clientele.



1. **duration**:
   * **959 outliers detected**: Extremely long or short call durations could reflect ineffective calls or highly successful interactions, respectively.



1. **campaign**:
   * **965 outliers detected**: A high number of contacts in a campaign could either be a strategy for hard-to-convince clients or could indicate over-targeting, potentially leading to customer dissatisfaction.

