Answers 5.4
Ashleigh Byers
8/7/2024

Data Shape:

Raw data: 991 rows, 14 columns

Cleaned data: 979 rows, 13 columns

Current_Clients: 776 rows, 13 columns

Former_Clients: 203 rows, 13 columns

Data Cleaning:

Missing Values:

- Last_Name – 1 missing value
  - Valid methods would be to either remove the entire row (which in this case contains other valuable information) or fill in the missing value with 'Unknown'.
  - Chosen method: filling in the missing value with 'Unknown'.

- Credit Score – 3 missing values
  - The mean and median are very close which indicates there likely aren't any major outliers.
  - Chosen method: Imputation with the mean of credit scores.

- Gender – 1 missing value
  - There are 528 males
  - There are 462 females
  - There are several ways of handling this missing data
    - Impute the missing gender record with the mode (most common gender) -- This could introduce bias
    - Remove the record altogether -- This would remove data that may have been beneficial from the other columns.
    - Fill in the missing value as 'unknown' -- When performing aggregations or charts based on the gender variable, having an unknown gender would not benefit the analysis.
  - Chosen method: Remove the record. Due to there only being one missing value, I chose to remove this record as there will be very minimal impact to the analysis. Had there been many missing values, another method would have been chosen.

- Age – 1 missing value
  - There are several ways of handling this missing data
    - Impute the missing age record with the mean age
    - Remove the record altogether -- This would remove data that may have been beneficial from the other columns.
    - Fill in the missing value as 'unknown' -- When performing aggregations or charts based on the age variable, having an unknown age would not benefit the analysis.
  - Chosen method: Impute the record with the mean age.

- Estimated Salary – 2 missing values

- The mean and median are very close which indicates there likely aren't any major outliers.
- Chosen method: Imputation with the mean estimated salary.

Mixed-Data Types:
There were no mixed-data types found, however the following data types were changed.

Column Name	Original Data Type	Updated Data Type
Credit Score	float64	int64
Age	float64	int64
Balance	object	float64
HasCrCard?	int64	boolean
IsActiveMember	int64	boolean
Estimated Salary	object	float64
ExitedFromBank?	int64	boolean

Duplicates:
There were no duplicates found.

Data Wrangling:

Dropped Columns:
'Row_Number' was dropped as it was just an index column and not relevant to the analysis.

Data Consistency:

Country and Gender columns had abbreviated and spelled out values. Corrections shown in the table below.

Country Code Abbreviations: https://www.yourdictionary.com/articles/country-abbreviations

Column Name	Value to Change	Changed to
Country	FR	France
Country	ES	Spain
Country	DE	Germany
Gender	M	Male
Gender	F	Female

There were 11 records with Age less than 18. Confirmed with stakeholders that these should be removed. Dropping these records only eliminates ~ 1% of the dataset and will not have any major impact on the analysis.

Answers 5.4
Ashleigh Byers
8/7/2024

Top Risk Factors:

1. Active Member: Non active members are higher risk
2. Number of Products: Only having 1 product is a higher risk
3. Gender: Females have a higher risk
4. Age: 36 and 55 years old has a higher risk