BUS211A – Foundations of Data Analytics - Group 1

Report 1: Data Preparation

North Star Consultants

North Star Consultants is a leading consultancy dedicated to designing and implementing cutting-edge Database Management Systems (DBMS) for businesses across various industries. With a team of highly skilled database experts and a passion for data optimization, we specialize in helping financial institutions harness the full potential of their data resources.

Mission Statement:

"Our mission is to offer comprehensive DBMS solutions that enable financial institutions to unlock the true value of their data assets and ensure that our clients thrive in a rapidly evolving financial landscape."

Vision Statement:

"To be the trusted partner of choice for financial institutions seeking excellence in optimizing data storage, retrieval, and analysis, ensuring our clients stay ahead in today's data-driven world.

Our Services:

1. Data Cleansing and Quality Assurance

2. Data Transformation and Structuring

3. Insights Generation to Support Strategic Planning and Forecast

4. Visualization and Reporting

Why Invest in us?

We have a history of over 100+ successful database implementations and satisfied clients in our 2 years of existence. Our solutions are high in demand and our expertise and commitment to excellence ensure investor confidence in our ability to execute and deliver results.

Customer Background

Horizon Bank is a prominent bank with over 50 branches across the United States. The bank wants to launch a new loan product for its loyal customers (more than 2 years with the bank). In addition, the bank is grappling with soaring marketing expenditures and an alarming drain on resources due to inefficient marketing campaigns. Horizon Bank is in dire need of a database with insights for optimizing their marketing strategies targeting the right customers, reducing expenses, and maximizing the effectiveness of their campaigns to maintain a competitive edge in the financial sector.

Value Proposition

1. To Build a relational database for managing the data set of the bank and leverage advanced data analytics and predictive modeling

2. To identify customers who qualify for the loan loyalty product for targeted marketing

3. To provide customer insights to enable the bank to develop a marketing strategy that minimizes expenses while delivering superior results.

Why did we choose the dataset?

Our data source was Kaggle, which is an accredited data source. The dataset had clear explanations to the field labels and a usability rating of 7.5. This dataset, in particular, tests our knowledge of data wrangling.

Our choice of this dataset is driven by its relevance, credibility, and potential to unlock actionable insights that can revolutionize marketing approaches within the banking industry, aligning perfectly with our consultancy's objectives and commitment to data-driven excellence. The dataset encompasses direct phone call marketing campaigns conducted by a prominent US banking institution over an extensive period, spanning from May 2008 to November 2010. This offers a rich source of historical data that enables us to delve deep into the dynamics of marketing strategies and customer responses.

Additionally, though the dataset was obtained from Kaggle, it is readily accessible through the UCI Machine Learning Repository, a reputable and widely recognized resource for high-quality datasets.

Data Preparation

a) Document the initial data quality and how credible you find the data source where you found the data.

Comprehension: The dataset had labels (column names) that were difficult to comprehend though it answers the questions being asked.

Clean/Completeness: Though there were no missing values, the dataset had quite a number of unknown values.

Chosen: Generally, the dataset had less irrelevant/confusing data. There was however an irrelevant field called ‘contact’ which specifies the communication type used to reach a customer. This is irrelevant to the goal of the project and has many ‘unknown’ entries.

Generally, the dataset is good for the type of analysis we intend to do. Though the dataset was obtained from Kaggle, it is readily accessible through the UCI Machine Learning Repository, a reputable and widely recognized resource for high-quality datasets. The repository was created and is maintained by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. Being part of this repository underscores the dataset's credibility, completeness, and suitability for advanced analytics and machine learning applications.

b) Document your data preparation process, including your data cleaning flow (steps / substeps), and any insights you have.

Renaming of columns

We began our data cleaning steps by renaming several columns in our dataset to enhance clarity and facilitate subsequent analyses. These column name changes include: Renaming [loan] to [personal_loan], Renaming [duration] to [duration_seconds], Renaming [month] to [last_contact_month], Renaming [campaign] to [campaign_times], Renaming [pdays] to [past_days], Renaming [previous] to [previous_contact], Renaming [poutcome] to [previous_outcome], Renaming [day] to [day_of_month] and renaming [y] to [subscribe_result].

These new column names will make it easier for us to analyze the dataset effectively in subsequent steps.

Exclusion of irrelevant values

We proceeded by filtering out values that held no relevance for our data analysis. Specifically, we addressed the issue of **'unknown'** values within columns [previous_outcome], [default], and [housing_loan]. These columns are binary variables, encompassing only two distinct values, 'yes' and 'no.' Given their binary nature, attempting mathematical imputation methods, such as mean or mode, to resolve 'unknown' values would not be feasible. The presence of these 'unknown' values could potentially compromise the integrity of the data crucial for our project's accuracy. Consequently, we made the decision to systematically remove these 'unknown' values from consideration

Replacing of values

Subsequently, **we replaced ‘unknown’ with 'other' within the [education] column**. This approach was taken to accommodate cases where individuals might have been reluctant to divulge their personal information during a standard telephone campaign. By categorizing such cases as 'other,' we aimed to ensure a more comprehensive and sensible representation of the data.

Filtering out values

Ultimately, we made the decision to **filter out all values below '10' in the [duration_second] column**. Based on our collective experience, when a person answers a telephone campaign or cold call, it typically takes around 5 seconds for sellers to provide a brief product introduction. Subsequently, it may take an additional 5 seconds for individuals to respond with a polite rejection or simply disconnect the call.

c) Document the final data quality after you clean it.

Now, our dataset is clean, making it easy to read, with labels that are easy to search. The entire dataset is now well-prepared for the database preparation and data analysis, resulting in outcomes that are easy to understand and interpret.

To be more precise, our dataset now exhibits key features that prove its quality:

1. Accuracy: All null and unknown values have been either removed or modified.

2. Relevance: There are no irrelevant or meaningless variables present in this dataset.

3. Validity: All variables have been converted into

the proper format for the subsequent analysis procedures.

4. Completeness: The dataset contains all the necessary information without any crucial elements missing.

5. Consistency: Data elements and their relationships remain uniform and coherent throughout the dataset.

6. Completeness: No essential variables are missing following the cleaning process.

d) What are your plans for possibly grouping/segmenting the rows of your data? Are there some numeric variables that might be good to divide into bins or categories, e.g., low, medium, high. Are there any data types that might need changing, e.g., converting a string

into a date?

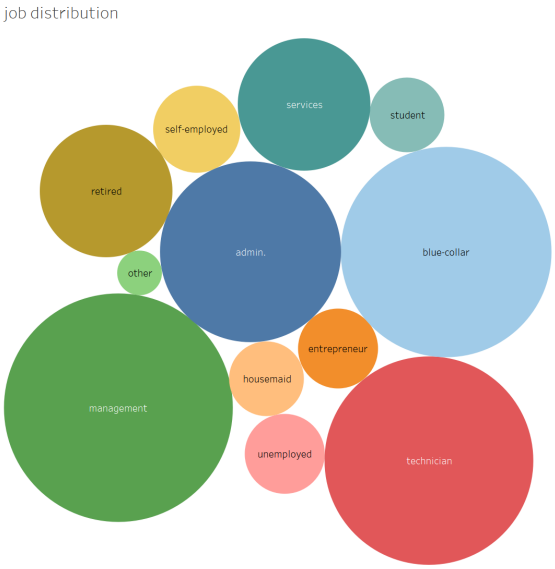
In order to enhance the visual representation of our dataset, we have implemented specific grouping strategies. Firstly, we have categorized the 'age' variable into intervals of 10 years, creating segments such as '18-30,' '30-40,' and so forth, up to 'over 80.' This categorization allows for a more structured analysis and visualization of age-related trends. Additionally, we have considered applying a similar approach to the 'balance' variable, where we intend to create intervals, possibly divided by increments of $2500.

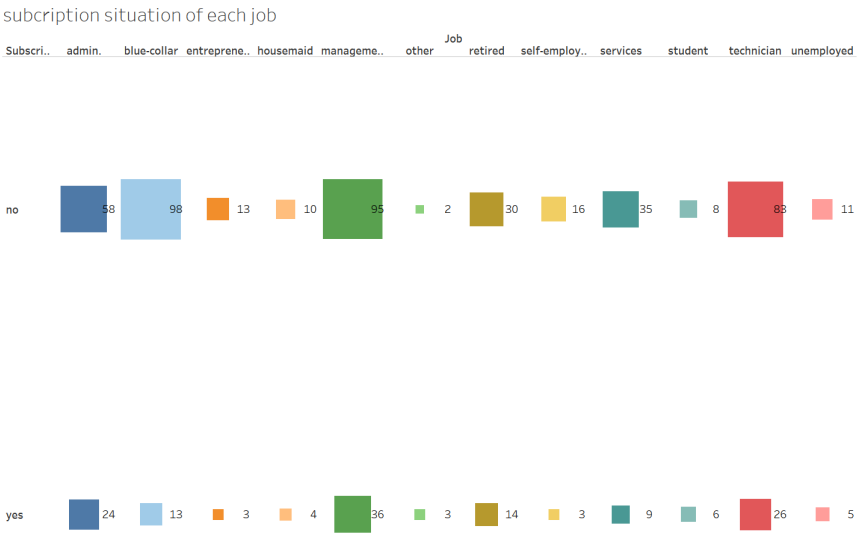
Furthermore, we have conducted extensive data type conversions during the data preparation phase using Tableau Prep. However, there remains one critical adjustment in our plan. We aim to amalgamate the 'last_contact_month' and 'day_of_month' variables to form a new column named 'last_contact_date.' Subsequently, we will convert this newly created column from its current string format to a date format. This conversion serves a dual purpose: it enables us to perform date-based calculations and allows for the precise calculation of the time elapsed (in days) since the last contact with customers. This time-based analysis can offer valuable insights into customer interactions and behaviors.

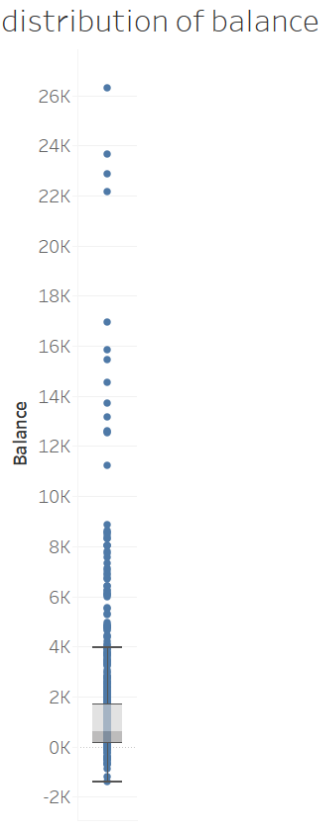
For a better visualization effect, we think [age] can set an interval by 10, like ‘18-30’, ‘30-40’, …, ‘70-80’, ‘over 80’. We also consider that balance can set an interval by 2500.

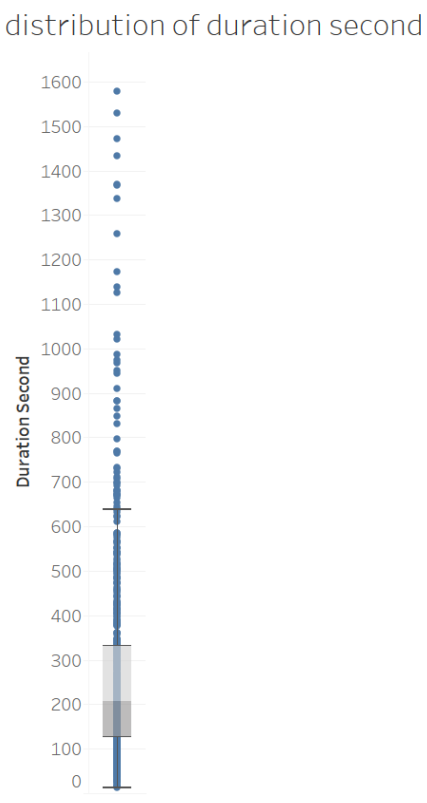
Also, we have already done most of the data type converting work in advance through tableau prep, but there is still one adjustment need to be done: now we have [last_contact_month] and [day_of_month], maybe we will combine these two together, create a new column called [last_contact_date] and convert it form a string to a date. With this column, we can even calculate the exact period(days) from the last contact up to now.

e) Show descriptive or summary statistics for a few key variables to show the data quality and data distributions. 10%









We harnessed the capabilities of data preparation tools, notably Tableau Prep, Microsoft Excel, and Tableau Desktop, to enhance and streamline our data quality assurance procedures.

a) In our Data Quality Assurance Process:

Tableau Prep's Contribution:

Handling unknown values: Tableau Prep enabled us to swiftly eliminate unknown values, by removing rows that were not critical but will prevent us from being able build an effective database for subsequent analytics.

Grouping/Segmentation and Binning: We effectively utilized Tableau Prep to segment and bin data, making attributes like age and account balance more amenable for analysis.

Filtering out values

Tableau Prep assumes a critical role in data refinement by enabling the filtration of non-relevant values. As an illustrative example, we harnessed Tableau Prep's capabilities to eliminate values falling below '10' in the [duration_second] column.

Replacing values

Tableau prep serves as a powerful tool for replacing one value with another value. We leveraged on Tableau Prep to replace ‘unknown’ values with ‘others’.

b) In Combination with Other Database / Software Tools:

Our project exemplified the seamless synergy between Tableau Prep, Microsoft Excel, and Tableau Desktop. While we initially received our data in CSV format, it turned out to be a text file. We efficiently imported it into Tableau Prep, which automatically organized the dataset for cleaning

purposes. In the concluding phrases, we exported the refined dataset to Excel and employed Tableau Desktop for data visualization. Furthermore, the resulting output is also compatible with Power BI for visualization purposes.

c) In a Collaborative Workflow Within or Across Organizational Boundaries:

Tableau Prep facilitated collaborative efforts among team members by enabling them to concurrently address various data preparation tasks within a shared workflow. Additionally, its version control systems played a crucial role in overseeing and monitoring alterations made by team members.

Simultaneously, Tableau Desktop provided a platform for team members to collaboratively craft visualizations using the same dataset, fostering seamless cooperation and creativity. Moreover, beyond organizational confines, Tableau Online offered the capability to effortlessly distribute and publish these visualizations, enhancing accessibility and sharing across diverse boundaries.