

12/8/2023

FOUNDATIONS OF DATA ANALYTICS

Final Report



Susan Arzapalo
Project Manager



Yian Chen
Data Analyst



Sean Kagugube
Business Analyst



Ella Acheampong
Data Engineer

About Us: North Star Consultants

Founded in 2000 in Boston, 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.

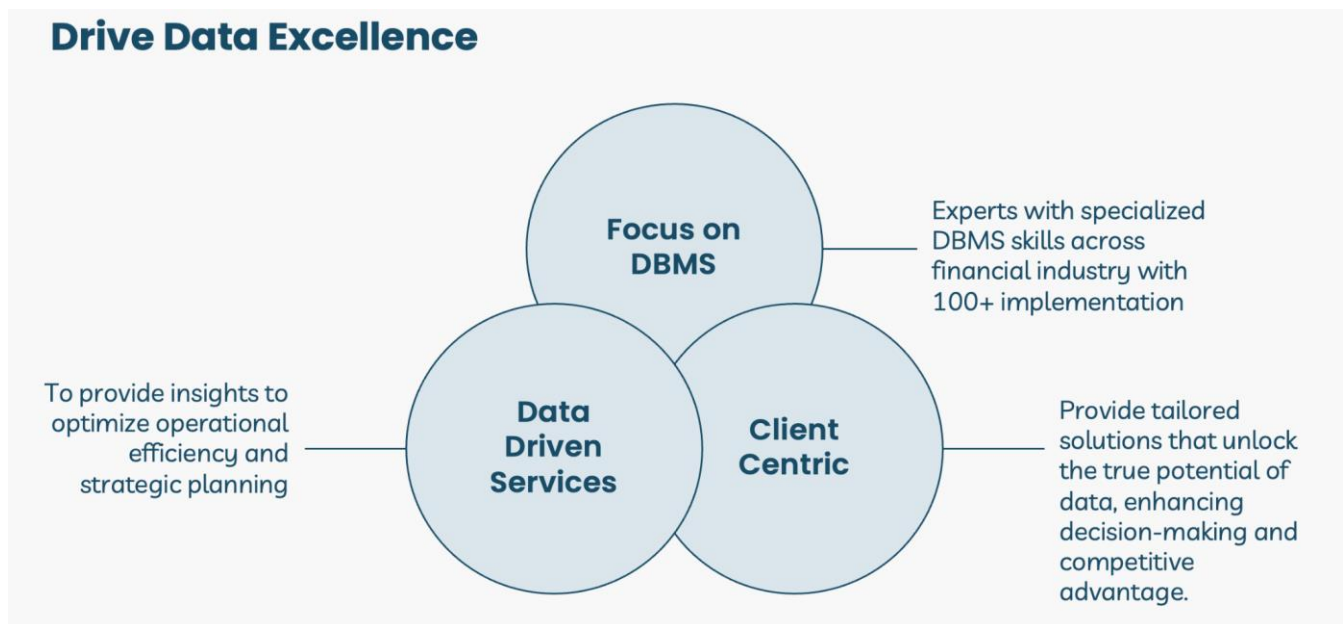
Mission Statement:

“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 rapid financial landscape through customized data driven services.”

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.

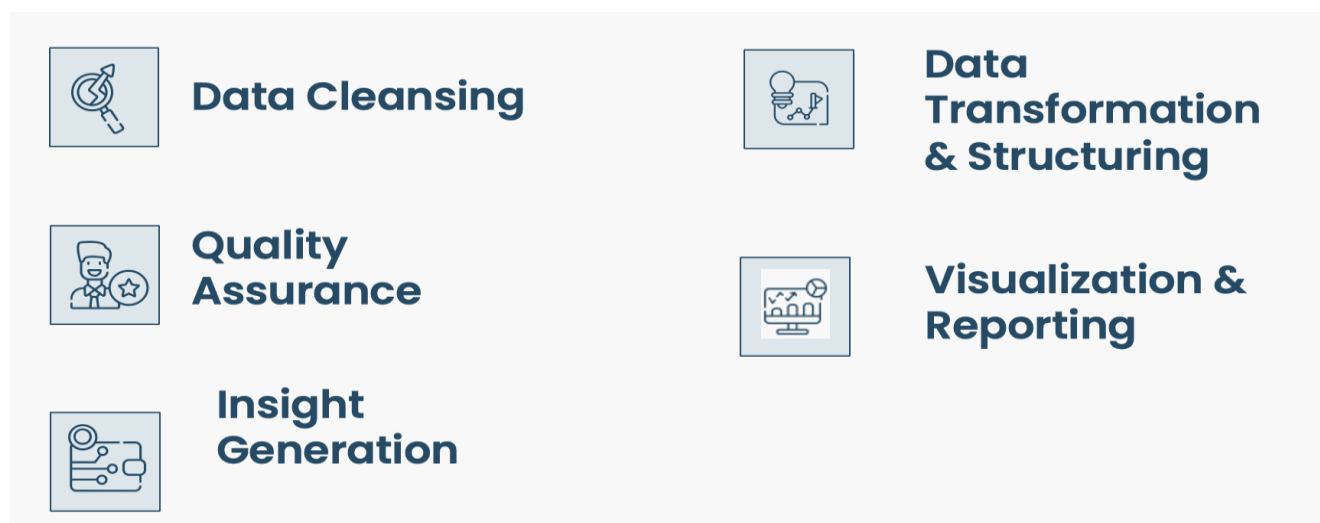
Value Proposition:



Why Invest in us?

We have a history of over 5000+ successful database implementations and satisfied clients in our 23 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.

Our Services:



Customer Background: Horizon Bank

With \$1.5B Assets, \$250M Equity, and 100M Customers, Horizon Bank (the bank) is a prominent bank with over 50 branches across the United States Northeast region. Horizon Bank services medium income communities consisting of mid-level managers, blue collar, management, and technicians.

Challenges of Horizon Bank:

- The bank doesn't have a centralized database, leading to silo data coming from CRM, SAP, Oracle, and other applications from their 50 branches without consolidation.
- No Data leadership, the bank doesn't have a data management strategy.
- Soaring marketing expenditures and an alarming drain on resources due to inefficient marketing campaigns.

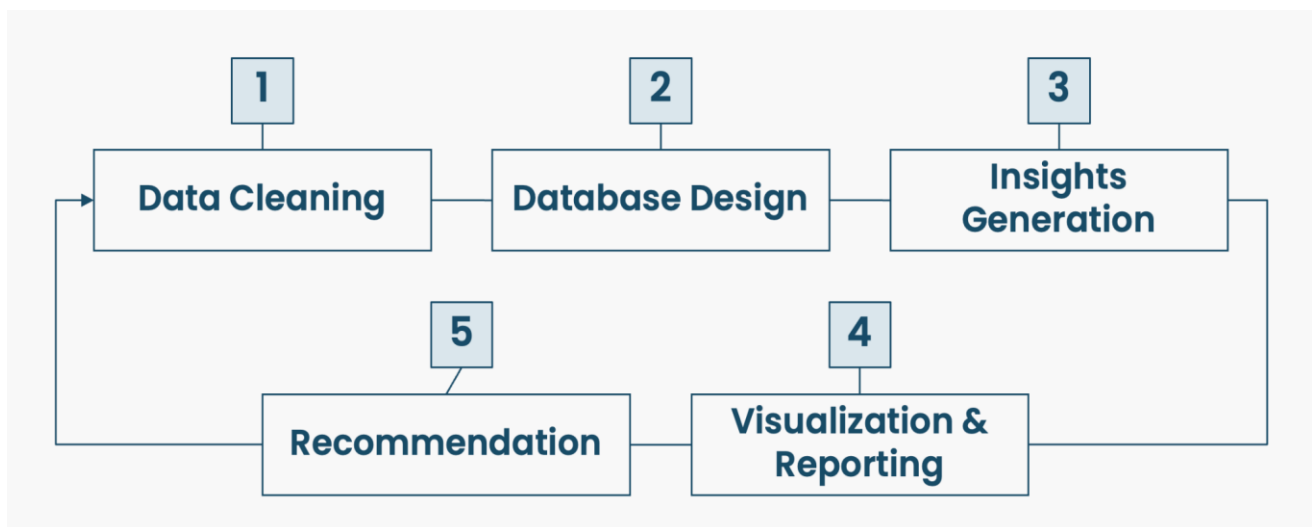
Goals: Project Scope

North Star Consultants aim to:

- Provide a centralized database to give a 360 degree view of their client's preferences, activities, transactions, and behavior.
- Help the bank gain insights to transform product development to outperform the competition
- Optimize marketing strategy and improve effectiveness, targeting the right customers, reducing expenses, and maximizing the effectiveness of their campaigns to maintain a competitive edge in the financial sector.

Proposed Project Phases:

Below is the diagram of the approved proposal of the steps North Star Consultants will follow to deliver the goals. The main focus will be to consolidate the bank's data in a central database and extract insights to support the marketing strategy, optimize its results, and reduce expenditure.



Data Quality

Why did we choose the dataset?

Our data originates from Kaggle, a well-respected source, providing clear field labels and boasting a usability rating of 7.5. This dataset serves as an excellent platform to test our data wrangling expertise.

The selection of this dataset was driven by its relevance, credibility, and its potential to unlock pivotal insights capable of revolutionizing marketing strategies within the banking industry. This aligns seamlessly with our consultancy's dedication to data-driven excellence.

This dataset encapsulates direct phone call marketing campaigns executed by a prominent US banking institution, spanning an extensive period from May 2008 to November 2010. Its wealth of historical data empowers us to deeply analyze marketing dynamics and customer responses.

Furthermore, despite its origin on Kaggle, the dataset is easily accessible through the esteemed UCI Machine Learning Repository. This repository is widely recognized as a reputable source for top-quality datasets.

Data Preparation

a) Evaluating the initial data quality revealed some noteworthy aspects.



<https://www.truinsights.ai/>

In terms of comprehension, while the dataset provided answers to the required queries, the column names were challenging to interpret. Regarding cleanliness and completeness, although no missing values were detected, a considerable number of entries were marked as 'unknown.'

The dataset showed a general lack of irrelevant or confusing data, save for one field labeled 'contact,' detailing the communication method with customers. This field, while containing numerous 'unknown' entries, doesn't align with the project's objectives and is therefore considered irrelevant.

Overall, the dataset remains suitable for our intended analysis despite these identified issues. Notably, although sourced from Kaggle, it's accessible via the UCI Machine Learning Repository, administered by the esteemed Center for Machine Learning and Intelligent Systems at the University of California, Irvine. This association underscores its credibility, completeness, and suitability for advanced analytics and machine learning applications.

b) The data preparation process involved several crucial steps to refine and structure the dataset.

Renaming of Columns

Initiating our data cleaning process, we implemented a comprehensive renaming strategy for several columns within our dataset. These adjustments aimed to refine column names for enhanced clarity and easier data analysis. We undertook multiple changes, including renaming [loan] to [personal_loan], [duration] to [duration_seconds], [month] to [last_contact_month], [campaign] to [campaign_times], [pdays] to [past_days], [previous] to [previous_contact], [poutcome] to [previous_outcome], [day] to [day_of_month], and [y] to [subscribe_result]. These new column labels were chosen to facilitate a more straightforward and effective dataset analysis in subsequent steps.

Exclusion of Irrelevant Values

Addressing the issue of 'unknown' values in certain columns—[previous_outcome], [default], and [housing_loan], all of which are binary variables representing 'yes' and 'no'—we found that employing mathematical imputation methods for 'unknown' values was impractical due to their binary nature. Consequently, we opted to systematically remove these 'unknown' values, considering their potential to impact the integrity of the data necessary for the accuracy of our project's outcomes.

Replacing Values

Further refinements involved replacing 'unknown' entries in the [education] column with 'other.' This substitution aimed to accommodate scenarios where respondents might have been hesitant to disclose personal information during standard telephone campaigns. Labeling such cases as 'other' sought to ensure a more inclusive and sensible representation of the dataset.

Filtering Out Values

In a final refinement step, we filtered out all entries below '10' in the [duration_second] column. This decision was informed by our collective experience, suggesting that during typical telephone campaigns, it takes approximately 5 seconds for sellers to introduce a product and an additional 5 seconds for individuals to respond or decline politely, prompting the exclusion of values below this threshold.

These cleaning processes aimed to refine and optimize our dataset, ensuring a more accurate and coherent foundation for our analytical endeavors.



c) Final Data Quality Documentation

Our dataset has undergone comprehensive cleaning procedures resulting in enhanced readability and searchability. It stands well-prepared for database preparation and subsequent data analysis, ensuring outcomes that are easily comprehensible and interpretable.

Outlined below are the key quality features of our refined dataset:

1. **Accuracy:** All null and unknown values have been meticulously addressed by either removal or modification, ensuring the dataset's accuracy.
2. **Relevance:** We have eliminated any irrelevant or meaningless variables, enhancing the dataset's relevance to our analysis objectives.
3. **Validity:** Through rigorous formatting adjustments, all variables now conform to appropriate formats essential for subsequent analysis procedures, ensuring the dataset's validity.
4. **Completeness:** The dataset now encompasses all necessary information without any crucial elements missing, ensuring its completeness for analytical pursuits.
5. **Consistency:** Data elements and their relationships maintain uniformity and coherence throughout the entire dataset, ensuring consistent and reliable analyses.

6. Data Integrity: Following the cleaning process, no essential variables crucial to our analysis are absent, confirming the dataset's completeness and integrity.



This meticulous cleansing has fortified our dataset, instilling confidence in its quality and reliability for driving insightful and accurate data-driven analyses.

d) Grouping and Segmenting Data Strategy

To optimize the visual representation and analysis of our dataset, we've initiated specific grouping strategies. Primarily, we've categorized the 'age' variable into distinct intervals of 10 years, creating segments such as '18-30,' '30-40,' and continuing in similar increments up to 'over 80.' This stratification allows for a structured and insightful analysis, especially when examining age-related trends. Additionally, we're considering a comparable approach for the 'balance' variable, intending to segment it into intervals, potentially divided by \$2500 increments.

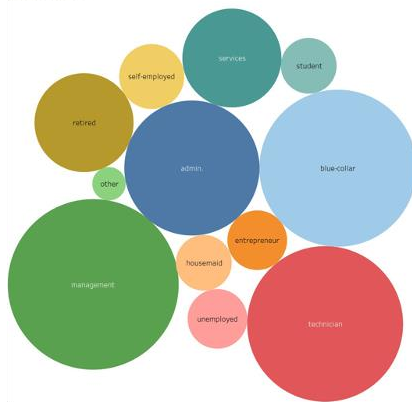
Moreover, our data preparation, primarily carried out through Tableau Prep, included extensive data type conversions. However, one crucial adjustment remains. Our plan involves amalgamating the 'last_contact_month' and 'day_of_month' variables to craft a new column named 'last_contact_date.' Subsequently, we'll convert this newly created column from its existing string format into a date format. This conversion holds significance as it facilitates date-based calculations and precise determination of the time elapsed (in days) since the last customer contact. This time-based analysis is pivotal in discerning customer interaction patterns and behaviors.

In summary, we're strategizing to group the [age] variable into 10-year intervals and potentially segment the [balance] variable in \$2500 increments. Additionally, the merging of [last_contact_month] and [day_of_month] into [last_contact_date] and its conversion to a date format will enable nuanced date-based analyses crucial for understanding customer engagement patterns.

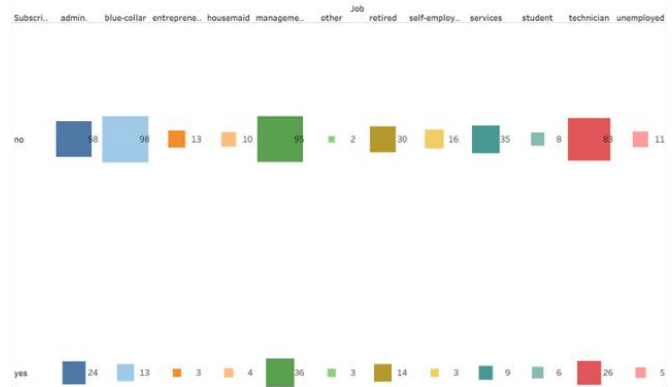
e) Descriptive Statistics and Data Quality Assurance

We leveraged an array of data preparation tools, primarily relying on Tableau Prep, Microsoft Excel, and Tableau Desktop, to bolster our data quality assurance procedures, ensuring meticulous attention to data distributions and overall quality.

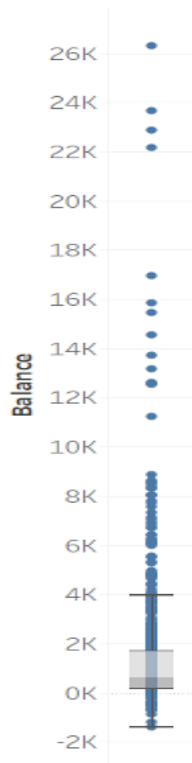
job distribution



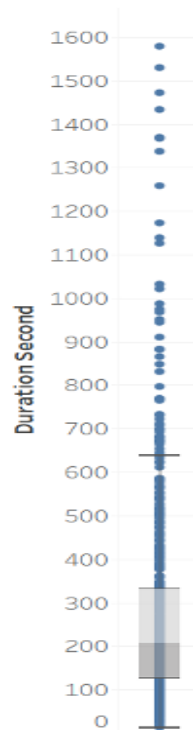
subscription situation of each job



distribution of balance



distribution of duration second



i) Data Quality Assurance Process:

Tableau Prep's Contributions:

Handling Unknown Values: Tableau Prep facilitated the swift removal of unknown values, enabling the elimination of non-critical rows that might impede the creation of an effective database for subsequent analytics.

Grouping, Segmentation, and Binning: Effectively utilizing Tableau Prep, we segmented and binned data, particularly refining attributes like age and account balance, rendering them more conducive for in-depth analysis.

Filtering Out Values: Tableau Prep played a pivotal role in refining data by filtering out non-relevant values. For instance, we utilized its capabilities to exclude values below '10' in the

[duration_second] column, ensuring data integrity and relevance for our analysis.

Value Replacement: Serving as a potent tool for value replacement, Tableau Prep was instrumental in substituting 'unknown' values with 'others,' enhancing the clarity and completeness of our dataset.

In parallel, to exhibit the data quality and distributions, descriptive or summary statistics will be showcased for key variables, highlighting the integrity and distribution patterns within our refined dataset.

ii) In Combination with Other Database/Software Tools:

Our project embodies a harmonious integration of Tableau Prep, Microsoft Excel, and Tableau Desktop. Initially presented in CSV format, our data unexpectedly arrived as a text file. Swiftly imported into Tableau Prep, this software automatically structured the dataset, streamlining the data cleaning process. Upon refinement, we seamlessly transitioned the dataset to Excel for final touches before harnessing the visualization prowess of Tableau Desktop. Additionally, the resulting output maintains compatibility with Power BI for comprehensive visualization capabilities.

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

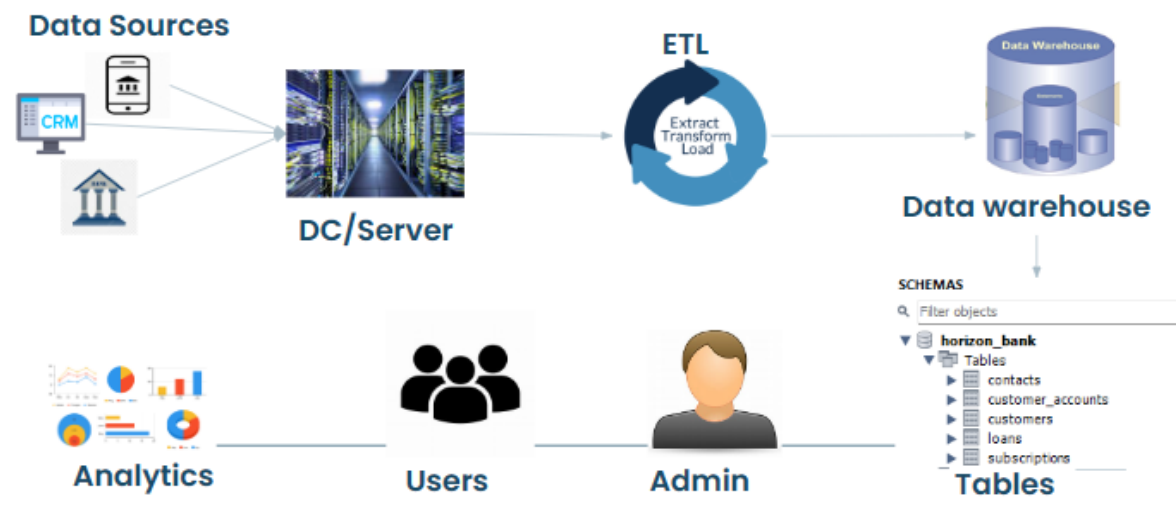
Tableau Prep served as a collaborative hub for team members, enabling concurrent engagement in diverse data preparation tasks within a shared workflow. The system's version control mechanisms ensured comprehensive oversight of alterations made by team members, fostering a coherent and well-monitored collaborative environment.



Simultaneously, Tableau Desktop provided a unified platform for team members to collaboratively construct visualizations using the refined dataset. This facilitated seamless cooperation, enabling the collective exploration of data insights and creative visualization strategies. Moreover, transcending organizational borders, Tableau Online facilitated effortless distribution and publication of these visualizations, ensuring widespread accessibility and sharing across diverse boundaries.

Database Management System Implementation for Horizon Bank

Data Management Architecture



As part of our deliverables, was to develop a data management architecture for Horizon bank. The data management architecture above was designed and developed by the engineering team of North Star Consultants. It shows the flow of data from data sources such as the mobile app, the 50 branches, contact center and the customer relationship management tool of the bank. The data is then stored on the bank's server located in its Data Center in California with backup in New York for data recovery in case of any disaster. The Technology team has been trained in how to do an Extraction, Transformation and Loading process each day to extract the unstructured data in the server and transform it into relational data suitable to fit in the bank's tables. This is then loaded into the data warehouse for access to users for their reports and analytics. The principal database is called Horizon Bank Database. Access to the tables in the database is restricted and users will require access from the Database Administrators of the bank. Now let's delve into the Horizon Bank Database.

Database Schema: The database developed by North Star Consultants is a robust solution designed to meet Horizon Bank's critical needs for data management and marketing optimization. It serves as a central repository for storing and managing essential customer-related information, enabling targeted marketing strategies and in-depth campaign analysis.

SCHEMAS

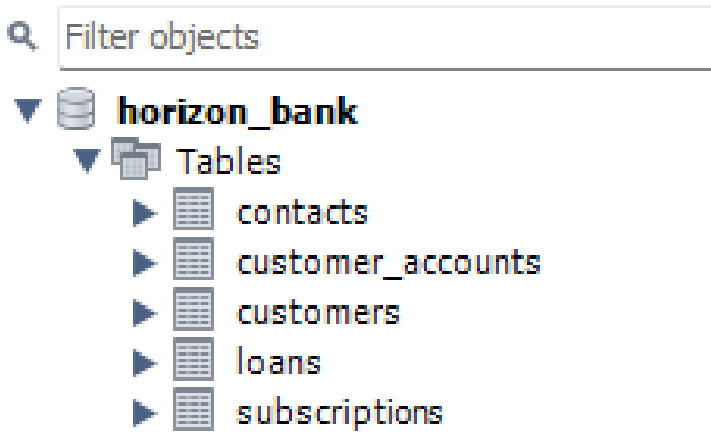


Table Descriptions

The database comprises five principal tables, each with a well-defined purpose:

- **Customers:** This table holds comprehensive customer data, encompassing demographics and personal details.
- **Customer Accounts:** It manages customer account-related information, including account balance.
- **Loans:** This table records customer loan details.
- **Contacts:** It tracks customer interactions, campaign data, and contact history.
- **Subscriptions:** This table stores information pertaining to customer subscriptions.

Some Use cases:

1. **Marketing Department:** They can utilize the customer demographics (e.g., age, marital status) for targeted marketing campaigns.

2. **Customer Relationship Management (CRM) Department:** This will enable them to maintain and update customer profiles for personalized services.

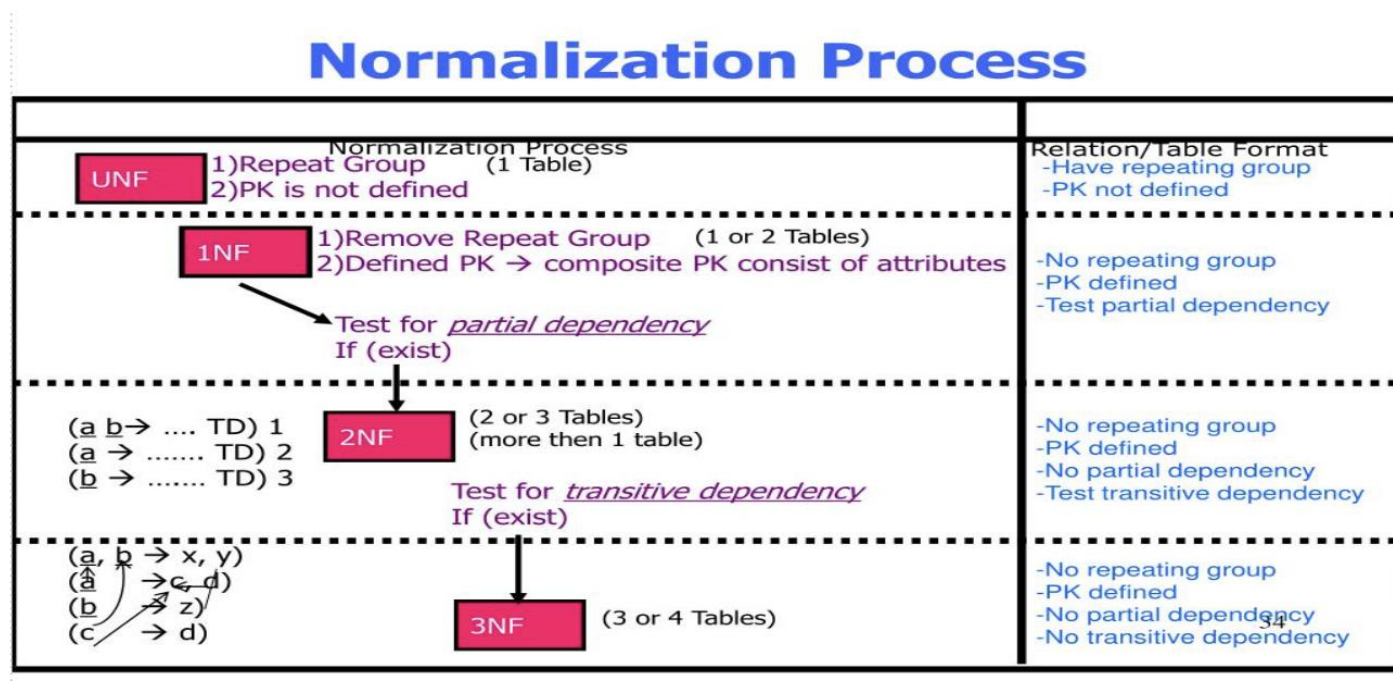
Risk Management Department: In order to monitor account balances and financial transactions to assess and mitigate risks.

3. **Finance and Accounting Department:** This will enable the team to access account balance and transaction data for financial reporting and auditing.

Normalization

The normalization process began with one large table which we broke down into 5 tables by the end of the normalization process. Fortunately, we didn't have repeating groups and each group record was uniquely identified by a unique primary key. Our data already conforms to all the 3 normalization forms. However, in order to ensure access to data is restricted to the specific need of the user, we separated the data into five tables with each serving a different purpose as shown above.

Below is the sequence of the normalization process by stages:



First Normal Form (1NF)

All tables discussed are in compliance with the First Normal Form (1NF). They ensure atomicity, meaning each column contains unique, indivisible values. Specific examples include the Customer table with distinct data like First_Name and Last_Name, Customer_Account holding singular data points like Balance, the Loan table with independent values in Housing_Loan and Personal_Loan, Contact table featuring singular entries like Contact details, and the Subscription table with discrete values such as Subscribe_Result. Each table successfully maintains the criteria of 1NF by avoiding duplicate columns and ensuring each column has a unique identifier.

Second Normal Form (2NF)

All tables adhere to the Second Normal Form (2NF). They are already in the First Normal Form (1NF) and exhibit no partial dependencies. Each table has a primary key (Customer_ID) with all non-key attributes (like MaritalStatus, Balance, Housing_Loan, Contact details, and Subscribe_Result) fully dependent on the primary key, ensuring the absence of partial dependencies. This complete reliance of attributes on the primary key across all tables, including Customer, Customer_Account, Loans, Contacts, and

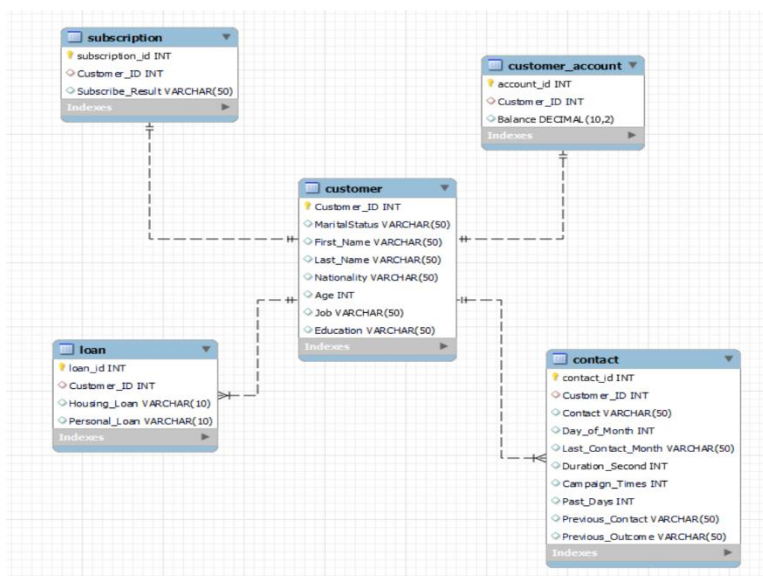
Subscriptions, confirms their compliance with 2NF.

Third Normal Form (3NF)

All tables in the dataset, including Customer, Customer_Account, Loan, Contact, and Subscription, meet the requirements of the Third Normal Form (3NF). In each table, non-key attributes solely depend on the primary key (Customer_ID) and exhibit no transitive dependencies. This consistent dependency structure across all tables ensures their compliance with 3NF, signifying a normalized database design to avoid data redundancy and maintain data integrity.

EER Diagram

The Entity-Relationship Diagram (ERD) provides a visual representation of these relationships



Explanation of Entities and Relationships

In Our Entity-Relationship (ER) diagram, the 5 tables work together to form a comprehensive database for managing customer data and interactions within Horizon Bank. The **Customers** table serves as the central hub, representing bank customers. It is linked to other tables,

including **Customer_Accounts** for tracking financial accounts, **Loans** for managing loans, **Contacts** for recording customer interactions, and **Subscriptions** for monitoring service. Together, these tables create a holistic view of customer relationships, financial activities, and interactions, enabling Horizon Bank to tailor its services, make informed decisions, and optimize customer engagement.

1. Customers Table (One-to-One and One-to-Many Relationships):

The customers table has a one-to-many relationship with all the tables except the loans table and contacts table. The one-to-many relationship indicates that one customer can have more than one loan and has been contacted more than once by the bank. The primary key of the customers table is the Customer_ID which serves as a foreign key in all the other tables.

2. Customer_Accounts Table (One-to-One): Each customer can have one account, and each account is associated with one customer. This one-to-one relationship links customers to their account details, including balances and transactions. The primary key on this table is the Account_ID which uniquely identifies each customer account.

3. Loans Table (One-to-Many): The loan table maintains a one-to-many relationship with customers. Each customer may have either a Personal Loan or Loan or both. The primary key in this table is the Loan_ID which uniquely identifies each loan.

4. Contacts Table: This table has a one-to-many relationship with Customer Table: Many contact records can be associated with one customer. This allows for tracking multiple interactions a customer may have with the bank. The primary key here is the Contact_ID.

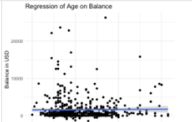
5. Subscription Table: This table has a one-to-one relationship with Customer Table. A customer may have a subscription option of yes or no. The primary key here is the subscription_ID.

Visualization

1. visualization logic introduction & concise frame

Our client sometimes has some incorrect ‘public perceptions’. Therefore, we used these perceptions and basic economic knowledge in the banking industry to bring forward some hypotheses, then we use visualization to prove or reject these hypotheses in order to give some useful and recommendations to our clients.

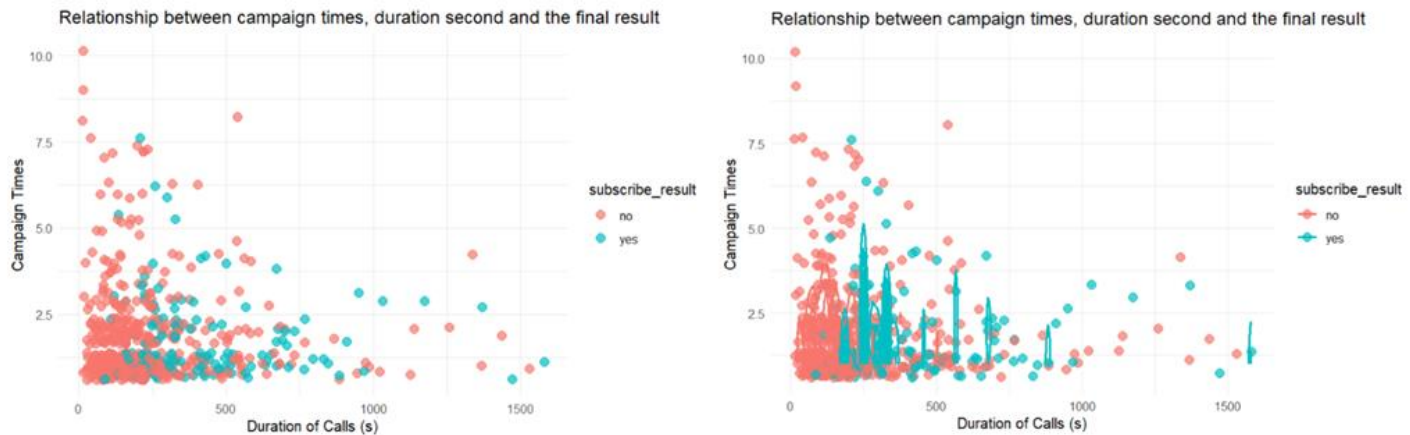
Briefly, this frame includes our insights with visualization.

| | public perception- tier1 | tier2 | 4 hypothesis | visualization(prove) | result (proved/reject) | interpretation |
|---|--|---|--|--|---------------------------|---|
| 1 | campaign result may relate to campaign times and campaign duration | | more campaign times and higher duration will increase success rate |  | partly proved | those people with few campaign times and long campaign duration will be more likely to accept this term deposit campaign. |
| success result = few campaign times + long campaign duration (following hypothesis are based on this conclusion) | | | | | | |
| 2 | campaign result may relate to customers' balance and age | balance relates to age | higher age, more balance |  | rejected | a person's balance does not relate to his/her age |
| | | subscription result relates to age, because people in specific age period are | people in 30-50 bracket will be more interested in this product |  | rejected | a person's acception possibility have no relationship with his/her age |
| | | subscription result relates to balance | people with higher balance are more likely to accept this campaign |  | proved | Those people with higher balance may have more idle money, and may be more willing to buy a term desposit product from horizon bank to earn some interests. |

2. Detailed visualization and Interpretation

- **Perception 1:** Campaign result (say yes/no to Horizon Bank’s telephone campaign) may be related to campaign times and campaign duration.
- **Hypothesis 1:** More campaign times and longer campaign duration may lead to a successful campaign result.

★ **Plot 1:** Relationship between campaign times, duration, and final result, grouped by campaign.



★ **Result & Interpretation:**

Partly proved !

The right density plot shows the concentration of the points in the scatter plot, which prepares us for the following interpretation of our scatter plot.

From the left scatter plot, we can see that people who reject the campaign may have a shorter campaign duration and more campaign times. We speculate that when the campaign times exceed 5, customers are more likely to become weary and reject the campaign.

Additionally, an increase in call duration could be associated with a higher likelihood of a customer subscribing to a service. We infer that those who are interested in our product may tend to gather more information, leading to a longer call duration with telemarketers.

- **Perception 2 tier 1:** Campaign result is personalized. The result may be related to customers' balance and age.
- **Perception 2 Tier 2-1:** Balance related to customers' age.

Hypothesis 2: Older customers will have a higher balance.

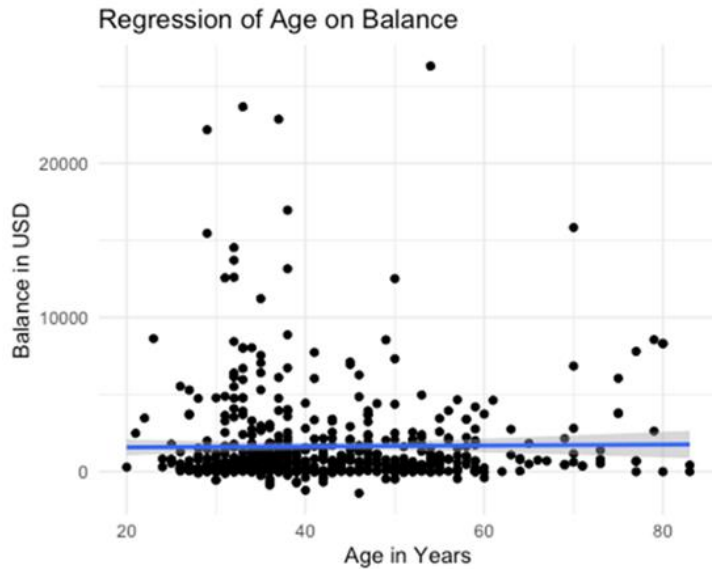
★ **Plot 2:** Regression of age and balance

★ **Result & Interpretation:**

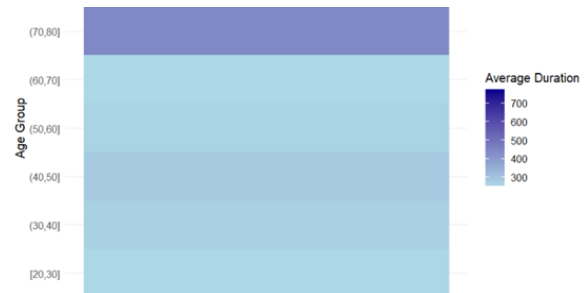
Rejected!

The graph shows that there isn't a strong correlation between customer age and account balance. When the age increases, the balance does not change too much.

From this graph we can now determine that age shouldn't be a useful factor when improving our client's telephone campaign efficiency.



- **Perception 2 Tier 2-2:** subscription result is related to age, because people in specific age periods are more likely to choose this product.
- **Hypothesis 3:** People in the 30-50 bracket will be more interested in this product (term deposit).
- ★ **Plot 3:** average contact times & average campaign duration, grouped by ages (age brackets)



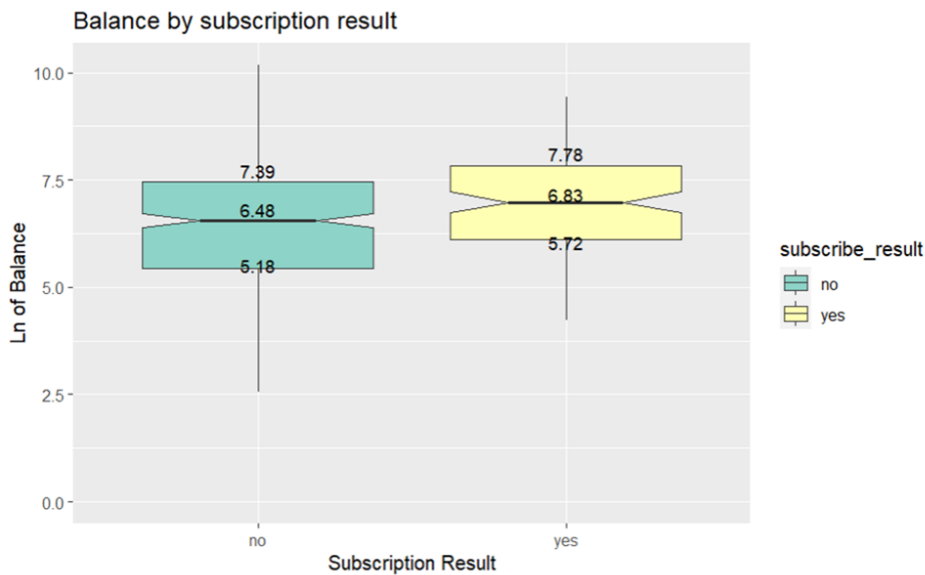
★ **Result and Interpretation:**

Rejected!

From hypothesis 1 and plot 1, we have a brief conclusion that: Successful Campaign Result(yes) = More Campaign Times + Longer Average Duration', we used those two elements (contacts times and duration).

For the term deposit product, a person's acceptance possibility has no relationship with his/her age, which aligns with plot 2, indicating the same result: age is not an effective indicator.

- **Perception 2 Tier 2-3:** • Subscription results are closely related to people's balance.
- **Hypothesis 4: People with higher balance will be more likely to subscribe for the term deposit through telephone campaign.**
- ★ **Plot 4: Balance by subscription result.**



- ★ **Result and Interpretation:**

Proved!

From the boxplot with notches, we can see those people who subscribe for the term deposit, their balance is significantly higher than those people who say no.

We conjecture that 'yes' group with higher balance may have more idle money. Therefore, they will be more willing to buy a term deposit product from horizon bank to earn some interest.

Recommendations

- Horizon Bank should consider a strategic shift in their campaign approach. We recommend a more concentrated and personalized engagement strategy tailored towards understanding the unique needs of each potential customer. This shift from numerous, rushed contacts to focused interactions could significantly enhance the effectiveness of their marketing efforts.
- It's our recommendation that the bank refocuses its marketing initiatives towards individuals with higher account balances. By prioritizing this segment, Horizon Bank can allocate resources more effectively, crafting targeted offerings that resonate with these high-value customers and potentially improving campaign success rates.
- Moreover, we suggest that Horizon Bank actively engages with customers having lower balances. Understanding their specific financial circumstances and needs can be pivotal in nurturing their accounts towards qualification for lucrative products, thereby broadening the bank's customer base.
- Additionally, it's our recommendation that Horizon Bank diversifies its subscription options. Introducing a range of plans tailored to different preferences and financial capacities can attract a more diverse customer base, offering inclusivity and accommodating various customer needs effectively.

Reflection

- The bank's policy

Reflecting on the bank's current policy that requires a customer to have existing loans to qualify for the new term deposit product seems to exclude a significant portion of potential customers. This approach may be biased against those without loans but with deposit needs. In response, the bank should consider revising this requirement or offering alternatives to cater to these customers' financial needs. For those without loans but with deposit requirements, the bank could provide attractive deposit rates, making it more inclusive.

- Dataset Insights:

The dataset provided valuable insights into Horizon Bank's customer demographics, financial behaviors, and interaction patterns. Analyzing this data unveiled critical details about account balances, customer preferences, and the effectiveness of marketing campaigns. However, certain limitations in the dataset, such as incomplete or missing data points, challenged the depth of our analysis.

- Team Collaboration:

Working within a team environment proved instrumental in leveraging diverse perspectives and expertise. Each team member brought unique skills to the table, contributing to different facets of the project. Collaboration and open communication were key strengths that facilitated the pooling of ideas, enhancing the quality of our analyses and recommendations.

- Challenges Faced:

Navigating through complex data structures and handling discrepancies in data formats posed challenges. Additionally, ensuring data accuracy and integrity demanded meticulous attention to detail. Overcoming these challenges required a coordinated effort, rigorous validation processes, and strategic problem-solving skills within the team.

- Learnings and Growth:

This project provided invaluable hands-on experience in data analysis and visualization. It emphasized the significance of data quality, effective teamwork, and the iterative nature of problem-solving.