BIN381 - PROJECT

Milestone 1

Members

Jo-Anne van der Wath (577394)
Henry Roux (577440)
Armandre Erasmus (577311)
Chaleigh Storm (577716)

Contents

Table of Figures

Introduction

The primary factor for customer eligibility for service contracts with satellite internet service provider LangaSat is annual salary, with a barrier of R50,000. Nonetheless, the business thinks that additional elements, like client demographics and lifestyle choices, can offer insightful information on qualifying and credit risk prediction. The goal of this project is to create a classification model that can forecast client eligibility based on extra variables, enhancing decision-making accuracy. This project will be guided through several phases, including business understanding, data understanding, data preparation, modelling, evaluation, and deployment, using the CRISP-DM framework.

Background

The only criteria for eligibility for a customer service contract that LangaSat currently uses is annual salary, which ignores other possible signs of a customer's capacity to fulfil the contract. The business has seen that eligibility and credit risk may be impacted by variables such as years of residency, marital status, occupation, and degree of education. By identifying and incorporating these factors into a recommendation system, this project aims to help LangaSat improve the accuracy and efficiency of its client assessment procedure.

Business Understanding

Business Problem

LangaSat currently only considers the annual salary of an applicant for determining eligibility, which may not account for all variables affecting credit risk. Currently, only prospective customers with an annual salary of R50 000 or more qualify for the services offered.

According to the observations made by LangaSat, prospective customers who have otherwise good credit risk scores may be turned away if wage is the only qualifying factor. By considering a variety of parameters, the organisation hopes to increase the accuracy of its eligibility evaluations (e.g., job title, years of residence, household size, etc.).

The objective is to create a recommender model that, in addition to salary, can more accurately determine whether clients are eligible for their service based on a variety of other factors. This will not only allow for a larger customer selection, but this will also reduce the overall credit risk to the company.

Business Objectives

1. Primary Objective:

Create a classification model that considers more attributes than just salary to accurately determine the eligibility of a customer for the satellite internet services provided by LangaSat.

2. Secondary Objectives:
   - Increase the precision of decisions made about customer eligibility.
   - Reduce the quantity of false positives and false negatives (i.e., make sure that eligible and unqualified consumers aren't mistakenly rejected and admitted).

- Examine many elements (such as occupation, length of residency, and educational attainment) to determine important components that affect customer eligibility.
- To help decision-makers better understand the model's output, present the results using visualisations (such as Power BI dashboards).
- Increasing the number of customers who are eligible for the services.

The model's improvements in accuracy, precision, and recall should allow it to exceed the baseline, which is now based solely on income.

Stakeholder Requirements and Success Criteria

There are multiple stakeholders in the project and each stakeholder has different requirements for the project.

1. LangaSat

LangaSat, as the primary stakeholder in the project, have the following requirements:

- An intelligent recommender model that should predict whether a customer is eligible for satellite internet service.
- The classification model should determine eligibility by accurately identifying the credit risk of the customer.
- The classification model should be able to determine eligibility based on more factors than the customer salary alone. This is to include a wider range of customers to be considered.
- The classification model should allow LangaSat to acquire more customers while minimizing the credit risk of the customers.

2. Project Supervisor

The project supervisor, Mr Gift T. Mudare, will have the following requirements from the team:

- The Cross-Industry Process for Data Mining Methodology (CRISP-DM) should be followed during the planning of the project. Thereafter, the project will be implemented using the planning framework derived from the CRISP-DM plan.
- The project supervisor must be made aware of the project goals and must agree with the goals before the continuation of the project to the following milestones.

- The project team will be required to follow the guidance of the project supervisor throughout the project.

3. Development Team

The project team will have the following requirements:

- The exploratory data analysis (EDA) of the raw data (CustData2.csv) to better understand the dataset.
- The preparation of the raw data to ensure high quality input.
- The development of an accurate classification model that will determine customer eligibility based on credit risk as mentioned under the LangaSat requirements.
- Model performance will need to be evaluated using performance metrics, such as confusion matrix, F1 scores or the accuracy of the model (Tavish, 2024).

- The documentation of the project, included but not limited to the methodologies used, preprocessing of data, exploration of data and model training.
- The model should be free from bias and should not be discriminatory. This is an ethical requirement.

Success Criteria

Success Criteria can be defined as precise standards that a project must satisfy in order to be considered successful (awork, 2024).

The requirements will be considered successfully met if:

- The model accurately classifies customers as eligible, or ineligible based on their credit risk.
- The model accurately determines the credit risk of the customer.
- The model performance should be above acceptable performance metrics.
- There is an increase in the number of customers that are eligible for services.
- There is a reduced financial risk by excluding customers with high credit risk from being eligible.
- Stakeholders are able to identify which variables contribute to credit risk and therefor customer eligibility the most.
- The model meets all the requirements set by the stakeholders.
- The project adheres to the CRISP-DM framework throughout all stages of the project.
- Each milestone of the project is completed before the due date.
- The work for the milestone is completed fully according to the project plan and CRISP-DM within the given time frame without compromising on the quality of the work.
- The model does not contain any biases or result in discriminatory classifications of customer eligibility.

Inventory of Resources

Data Resources

- Dataset: The offered dataset (CustData2.csv) includes variables including department names, salary, job titles, and household sizes.
- Data Source: Supplied with consumer demographic data for the LangaSat scenario.

Software & Tools:

1. Data Analysis Tools:
    - R: For constructing models, cleaning date, the preprocessing phase and data exploration. (Python could also be used in data analysis, but in this case, R will be used).
    - RStudio: Integrating environments for development with coding. (Jupyter Notebook could also be used)

2. Visualization Tools:
    - Power BI will be used to create dashboards and visualisations which will be used to show correlations, patterns, and model results.

3. Collaboration Tools:
   - For exchanging documents, code, and collaborative projects, use GitHub, Google Drive or Microsoft Teams.

4. Machine *Learning Libraries:*
   - Relevant R libraries could be used for constructing models for classification.

5. Human Resources:
   - All the fellow students working on the project, each with a specific responsibility (e.g., data cleaning, model building, reporting), make up the group.
   - Project Supervisor: Instructor offering direction and input on the project's development. (In our case this would be our lector giving feedback)

6. Hardware Resources:
   - Computers/Laptops: Individual machines will be used for modelling, analysis, and visualisation.
   - Cloud systems: For computational power and cooperation, Google Colab or other comparable cloud-based systems will be used if needed.

7. Reference Materials:
   - Documentation of the CRISP-DM Methodology to direct the project stages.
   - Textbooks and Online Resources: Any scholarly or research articles that aid in the comprehension of data mining, business intelligence, and customer classification models.

CRISP-DM

The CRISP-DM framework ensures that data mining projects follow a structured, repeatable process. It allows flexibility, so earlier stages can be revisited as new insights are gained during the project (SMARTVISION, 2024). The CRISP-DM framework consists of the following phases:

1. Business Understanding:
   - This phase focuses on identifying what the business wants to achieve. For your project, the primary business goal is to build a model that can recommend customer eligibility for a satellite internet service based on more than just salary.
   - Key Questions:
     - What factors beyond salary could impact customer eligibility?
     - How will the model's success be measured (accuracy, precision, etc.)?
   - The business is looking for a model that can reduce credit risk by finding more accurate indicators of eligibility.
2. Data Understanding:
   - Once the business goals are clear, the dataset (CustData2.csv) will need to be analysed. This involves initial data exploration to understand the attributes and types of data in the dataset.
   - This stage is critical for uncovering initial insights and forming a basis for the model.
   - Key Actions will include:
     - Doing descriptive statistics to summarize data.
     - Identifying missing values, inconsistencies, and outliers.

  - Developing an initial hypothesis about which features might be important for predicting eligibility.

3. Data Preparation:
   - The data preparation phase involves transforming the raw dataset into a format suitable for modelling.
   - Properly prepared data will ensure that the modelling phase proceeds smoothly.
   - Key Tasks:
     - **Cleaning**: Handling missing values, duplicates and noisy data. At this stage imputation may be done, missing values will need to be filled with means or removed. Noisy data will need to be cleaned.
     - **Feature Engineering**: Creating new features or modifying existing ones. Values such as age will be created using the year of birth and such.
     - **Encoding Categorical Variables**: Since many variables like job title, department, and city of residence are categorical, they will need to be encoded into numerical values using one hot encoding or other similar methods.
     - **Scaling**: Normalize or standardize numerical data such as salary if required by the chosen model.
4. Modelling:
   - In this phase the model will be created. The model can be created using machine learning algorithms. If the model accuracy is sub optimal, different algorithms may be implemented and the modelling phase will be reiterated until satisfactory performance occurs.
   - Key Considerations:
     - **Algorithm Choice**: Common options include logistic regression, decision trees, or random forests for classification problems like this.
     - **Training**: The dataset will be split into a training set and a test set to train the model and evaluate its performance on unseen data.
     - **Tuning**: Fine-tuning of hyperparameters (e.g., max depth for decision trees) may occur to optimize the model's performance.
5. Evaluation:
   - Once a model is built, it must be evaluated to ensure it meets the business objectives set out in the first phase.
   - The model performance will need to be compared to the current baseline.
   - Key Metrics:
     - **Accuracy**: This is an indication of how often the model accurately predicts eligibility. The accuracy of model should be at least 85%.
     - **Precision**: Of the customers predicted to be eligible, how many are actually eligible?
     - **Recall**: Of all customers who are actually eligible, how many does the model predict correctly?
     - **F1-Score**: A balance between precision and recall.
6. Deployment:
   - Finally, the model will be deployed in a real-world environment where it can be used by stakeholders.
   - Key Steps:
     - A user-friendly interface will need to be created for non-technical stakeholders.
     - The model will need to be maintained, monitored, and updated over time.

Risks, Assumptions and Constraints

Risks

Risks cannot be avoided for sure, however, the most likely risks involving the project can be determined and planned for. Proper risk mitigation strategies can be implemented to minimise the impact of the risks on the project outcome.

Risks can be classified based on the impact on the project. They are classified as low, medium or high risk. A risk matrix can be used to determine the impact of the risk using the likelihood and severity of the risk should it occur.

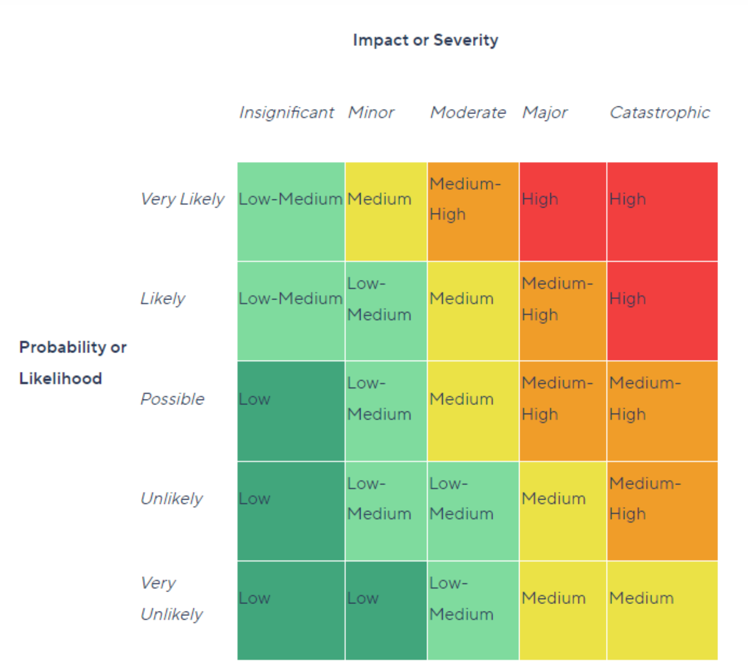
The likelihood of a risk occurring can be as follows:

- Very unlikely
- Unlikely
- Possible
- Likely
- Very likely

The severity of the risk on the project should they occur can be as follows:

- Insignificant
- Minor
- Moderate
- Major
- Catastrophic

Boogaard (2024), provides the following risk matrix:



Impact or Severity

		Insignificant	*Minor*	*Moderate*	*Major*	*Catastrophic*
Probability or Likelihood	*Very Likely*	Low-Medium	Medium	Medium-High	High	High
	Likely	Low-Medium	Low-Medium	Medium	Medium-High	High
	Possible	Low	Low-Medium	Medium	Medium-High	Medium-High
	Unlikely	Low	Low-Medium	Low-Medium	Medium	Medium-High
	Very Unlikely	Low	Low	Low-Medium	Medium	Medium

Figure 1 Risk Matrix

The risks of the project will mainly revolve around risks associated with the data analytics aspects of the project but is not limited to the data. The risks are as follows:

1. **Data Quality**

Data quality is determined by the accuracy, completeness, validity, consistency, uniqueness, timeliness and relevance of the data (IBM, 2024).

- Risk: Low data quality as a result of missing values, duplicates, outliers or noisy data will have a direct impact on the accuracy of the model. If the model is inaccurate, customer eligibility cannot be determined accurately. This directly impacts the decision-making of LangaSat.
- Severity: Major
- Likelihood: Possible
- Impact: High-Medium Risk
- Mitigation: The data will need to be cleaned before it can be used. Missing values will need to be handled. Outliers will need to be dealt with. Noise will need to be removed to prevent inaccurate classifications.

2. **Biased data**

Human biases can be present in the dataset (Javapoint, 2024).

- Risk: If there are biases present in the dataset, this will lead to these biases carrying over to the model. This ultimately leads to biased or inaccurate classifications from the model.
- Severity: Major
- Likelihood: Likely
- Impact: Medium-High Risk
- Mitigation: The integrity of the data will need to be verified. Furthermore, the data will need to be cleaned to remove any biased or dirty data that cannot be properly classified (Roshaan, 2024).

3. **Overfitting**

Overfitting of a model occurs when the model is trained too perfectly according to the training data that it cannot account for any variability when implemented and given other data (Javapoint, 2024).

- Risk: Overfitting means that the model will not be able to make an accurate prediction should the input vary from the data the model was trained on.
- Severity: Moderate
- Likelihood: Possible
- Impact: Medium Risk
- Mitigation: The dataset should contain a wide variety of data. Furthermore, the data should be free of noise that could be misinterpreted as a trend or pattern that could be classified (Roshaan, 2024).

4. **Data Privacy**

The model will be trained on the data of customers. This data will be sensitive.

- Risk: According to Roshaan (2024), the training data of a model can be reproduced. This means that attackers can recreate and steal the training data and use this data with malicious intent.
- Severity: Catastrophic

- Likelihood: Unlikely
- Impact: Medium Risk
- Mitigation: The Probability Approximately Correct (PAC) Privacy framework can be implemented to protect the data. This will allow for the minimum needed amount of noise to be added to the model that will prevent the recreation of the training data.

5. **Missing Deadlines**

Each milestone of the project will need to be completed within a specific timeframe.

- Risk: With any project, there is a risk of delays occurring which may lead to missed deadlines. If the deadlines are missed it may cause the project to not be completed in time.
- Severity: Moderate
- Likelihood: Possible
- Impact: Medium Risk
- Mitigation: Proper time management must occur. Dashboards can be used to monitor progress, and timelines can be created to ensure every aspect of the project is completed in a timely manner.

Assumptions

In a project, assumptions are the components of a project that are believed to be true, even if there is no proof (Malsam, 2022).

To complete the project, the following assumptions will be made:

1. The provided dataset “CustData2.cvs” is a complete, accurate representation of the customer population for LangaSat.
2. The dataset will contain all the necessary data to create a model to accurately determine the customer eligibility based on credit risk.
3. Data preprocessing, such as data cleaning in which missing values will be handled and noise will be removed from data, will be able to be completed the quality of the data being significantly compromised.
4. The data in the dataset is relevant to determining the credit risk of customers.
5. The dataset is free of biases that will influence the model to make biased classifications.
6. The development team will use R and Power BI for data analysis and modelling.
7. Currently, customers with an average yearly salary of R50 000 or more are eligible for the services provided by LangaSat.
8. LangaSat has the proper infrastructure that will be able to allow for the easy integration of the model.
9. The supervisor will be available for guidance through all phases of the project.
10. The project will adhere to the CRISP-DM framework.

Constraints

The constraints in the project are the limitations that are set that the team will have to work within (monday.com, 2024).

The project will have the following constraints:

1. The CRISP-DM framework must be followed.
2. The supervisor must be consulted on all data mining goals before the team will be able to continue with subsequent milestones.
3. R and Power BI are the tools that should be used to prepare the data and create the models. This could limit the data analysis in terms of the capabilities of the tools.
4. The accuracy of the model may be limited if the dataset contains many missing values, or the quality of the data is low.
5. If there are only a limited number of records in the dataset will mean that we are working with limited data on which the data may be trained. This could lead to less accurate results from the model.
6. The time to complete each milestone is limited and not completing each milestone within the set timeframe may compromise the quality of the project.
7. The development team is limited to the four members of the team and the knowledge of these members.
8. The model will need to comply with ethical guidelines.
9. Privacy limitations may limit the way the data can be stored and used in analysis.

Data Understanding

Data Mining Goals and Success Criteria

Data mining goals help to create clear, measurable objectives that will be used to measure the success of the project. The project will have the following goals:

1. Primary Goal:

The main objective is to create a classification model that can recommend whether a customer is eligible for the satellite internet service. The company currently uses salary as the only criterion, but the model will incorporate other factors, such as job title, education level, and household size, to improve accuracy.

- Specific Objectives are:
  - **Improve Customer Selection**: To reduce the number of false positives (customers incorrectly predicted as eligible) and false negatives (eligible customers not recommended).
  - **Feature Importance**: To determine which variables (besides salary) are the most important for predicting eligibility.
  - **Baseline Comparison**: The aim is to outperform the model that currently only uses salary as the predictor.
- Evaluation Metrics that should be met, include:
  - **Accuracy**: The percentage of correct predictions. The accuracy of the model should be at be at least 85%, to be deemed acceptable. The higher the accuracy, the better.

- **Precision and Recall**: Important metrics if the business cares more about minimizing false positives (bad credit risks) or false negatives (missed opportunities).
- **F1-Score**: A harmonic mean of precision and recall, often used in classification problems where balancing false positives and false negatives is important.

2. Secondary Goals (Optional):

a. **Clustering**: Group customers into different segments based on shared characteristics (e.g., education level, city of residence).

b. **Trend Analysis**: Look for trends over time (e.g., do customers from certain age groups or professions tend to qualify more frequently?).

Dataset Overview

The CustData2.csv dataset that will be used has the following attributes:

- **Customer_ID**: A unique identifier for each customer. It does not have predictive value but is used to track individual records.
- **Job_Title**: Job titles can be a strong predictor of financial stability and service eligibility. This can be used to determine if there are groups of customers with certain jobs that are more eligible for services than others.
- **Department**: Similar to job titles, certain departments (e.g., IT, Finance) might be more indicative of eligibility than others. This variable can help create broader segments.
- **Annual_Salary**: This is the current criterion for eligibility. Although it is a strong predictor for eligibility, it should not be the sole predictor. The new model can be tested against the current model to test if the new model will be more accurate.
- **Year_of_Birth**: Year of Birth can be used to calculate the age of the customers. This will allow insights to be gathered on whether certain age groups have more financial responsibility and resources, which may impact other factors contributing to eligibility.
- **Marital_Status**: Marital status might indicate different financial responsibilities, with married customers potentially having higher household expenses.
- **City_of_Residence**: Different cities may have varying economic conditions. Examining the area from which customers come, may allow for insight to be gained on the city’s financial situation.
- **Years_of_Residence**: Length of residence might indicate stability or financial responsibility, with longer residence possibly correlating with lower credit risk.
- **Level_of_Education**: Higher education levels often correlate with better-paying jobs and financial stability, which could impact eligibility.
- **Occupation**: This provides further context for the Job_Title and Department attributes. Certain occupations might have lower financial risks. Additionally, customers in varying occupational field may have different incomes for the same job titles.
- **Household_Size**: A larger household could indicate higher expenses and thus a greater financial burden, which might affect eligibility.

Dataset Considerations:

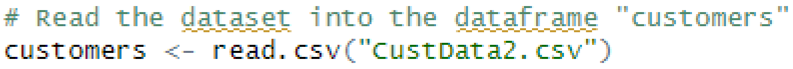
- **Missing Data**: The dataset will need to be checked for missing or incomplete values. Missing values will be handled appropriately to avoid bias in the model.
- **Outliers**: The dataset will most probably contain outliers. It is important to either remove or properly handle outliers, as they could skew the model.

- **Feature Engineering**: New features can be created such as **age** (from year of birth) or combining similar job titles into broader categories.

Data Quality Assessment

The quality of the data plays a very important role in creating an accurate intelligent recommender model. The GIGO (garbage in, garbage out) concept supports this by stating that in any system, the quality of the input determines the quality of the output (Awati, 2023). Therefore, the quality of the dataset must be assessed for quality problems, missing values, duplicates, outliers and any other relevant metrics. Any identified quality concerns must be mitigated in the third CRISP-DM phase (Data Preparation).

RStudio will be used to analyse the dataset. To import the dataset into RStudio the following code needs to be executed:



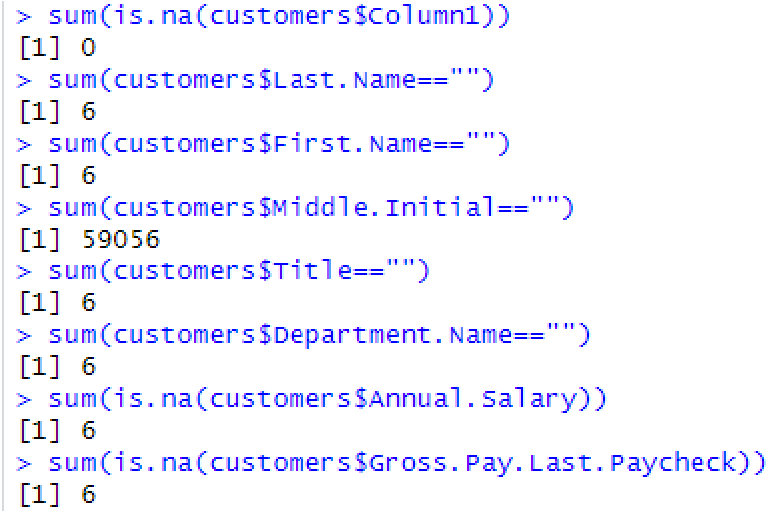
```
# Read the dataset into the dataframe "customers"
customers <- read.csv("CustData2.csv")
```

Figure 2: Read the dataset into a data frame

Handling Missing Values

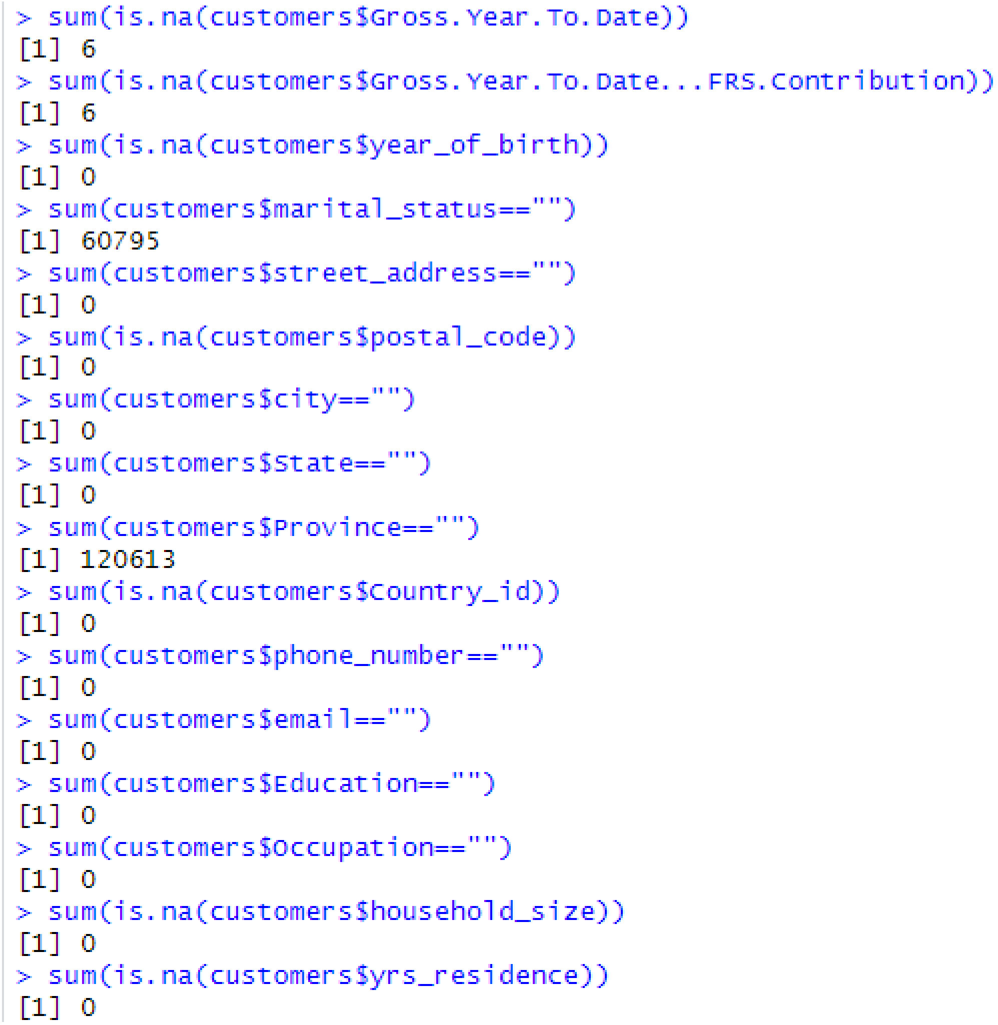
This is done by determining whether entries in the dataset is empty/missing or N/A, as these could have an impact on the analysis' findings. Depending on the amount of missing data and its significance for the analysis, missing values are either imputed (using the mean, median, or mode) or deleted, this will only be done in the next phase of the CRISP-DM lifecycle as stated above.

Figures 3 and 4 below contains code for every attribute that counts the number of empty cells and their respective output.



```
> sum(is.na(customers$Column1))
[1] 0
> sum(customers$Last.Name=="")
[1] 6
> sum(customers$First.Name=="")
[1] 6
> sum(customers$Middle.Initial=="")
[1] 59056
> sum(customers$Title=="")
[1] 6
> sum(customers$Department.Name=="")
[1] 6
> sum(is.na(customers$Annual.Salary))
[1] 6
> sum(is.na(customers$Gross.Pay.Last.Paycheck))
[1] 6
```

Figure 3: Missing values for each attribute - Part 1



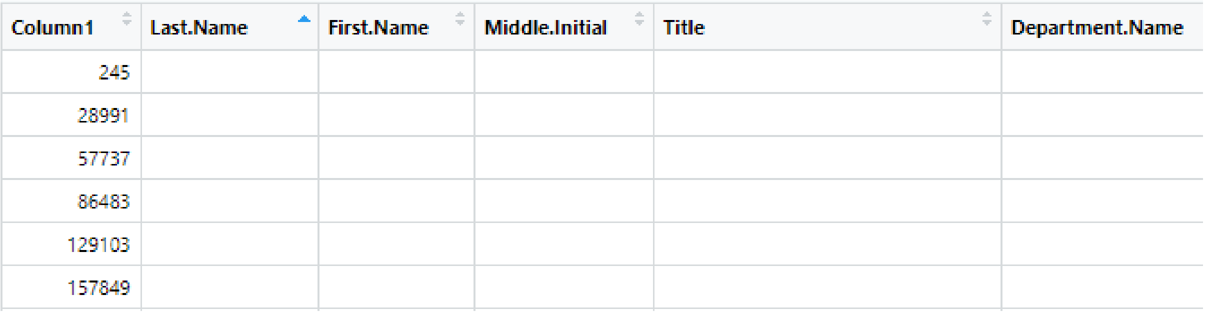
```
> sum(is.na(customers$Gross.Year.To.Date))
[1] 6
> sum(is.na(customers$Gross.Year.To.Date...FRS.Contribution))
[1] 6
> sum(is.na(customers$year_of_birth))
[1] 0
> sum(customers$marital_status=="")
[1] 60795
> sum(customers$street_address=="")
[1] 0
> sum(is.na(customers$postal_code))
[1] 0
> sum(customers$city=="")
[1] 0
> sum(customers$State=="")
[1] 0
> sum(customers$Province=="")
[1] 120613
> sum(is.na(customers$Country_id))
[1] 0
> sum(customers$phone_number=="")
[1] 0
> sum(customers$email=="")
[1] 0
> sum(customers$Education=="")
[1] 0
> sum(customers$Occupation=="")
[1] 0
> sum(is.na(customers$household_size))
[1] 0
> sum(is.na(customers$yrs_residence))
[1] 0
```

Figure 4: Missing values for each attribute - Part 2

The following was observed:

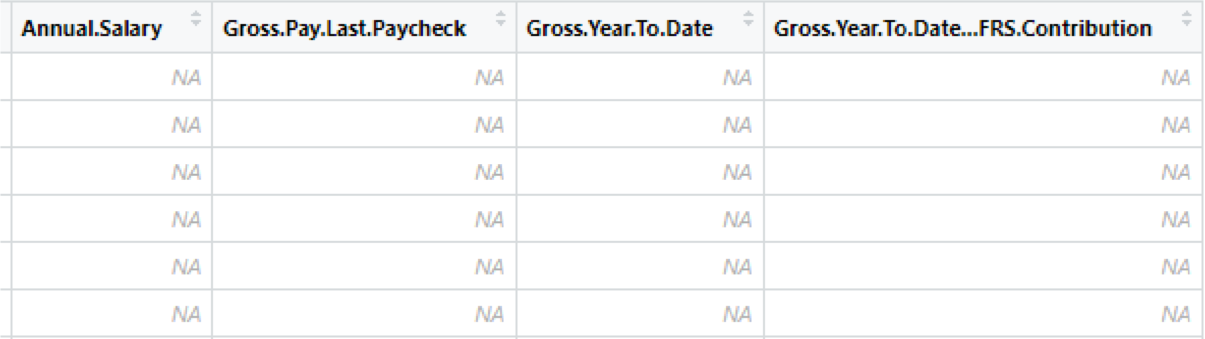
- There are three attributes that are missing a large number of entries:
  - Middle.Initial: It is not abnormal for this attribute to be empty, seeing as not every person has a middle name.
  - Marital_status: All the cells for this attribute must contain an entry, such as “Married” or “Single”. This shows weak data quality, seeing as this attribute can be used to train the model.
  - Province: All the cells for this attribute must contain an entry. This also shows weak data quality, seeing as this attribute can be used to train the model.
- There are numerous attributes that are only missing six entries. This is not a dangerous number of entries, but it still needs to be investigated. The following was discovered:

The six missing entries for each attribute, is missing for the same 6 records. This could have been caused by corruption during data extraction or data entry issues.



Column1	Last.Name	First.Name	Middle.Initial	Title	Department.Name
245					
28991					
57737					
86483					
129103					
157849					

Figure 5: Six records missing the same attributes - Part 1

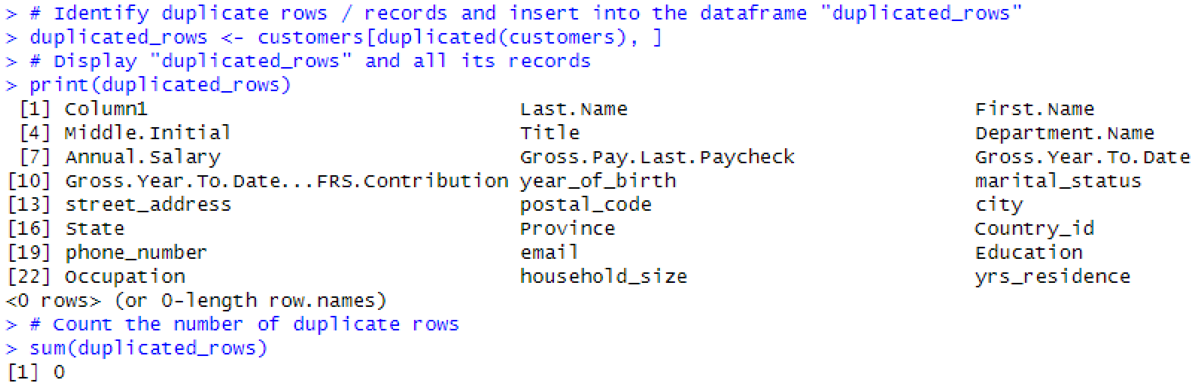


Annual.Salary	Gross.Pay.Last.Paycheck	Gross.Year.To.Date	Gross.Year.To.Date...FRS.Contribution
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA

Figure 6: Six records missing the same attributes - Part 2

Duplicate Records

Duplicate records exist when a record/row appears more than once in a dataset. This is redundant and unnecessarily uses more storage. Any duplicate records are identified using the following R code:



```
> # Identify duplicate rows / records and insert into the dataframe "duplicated_rows"
> duplicated_rows <- customers[duplicated(customers), ]
> # Display "duplicated_rows" and all its records
> print(duplicated_rows)
 [1] Column1                               Last.Name                             First.Name
 [4] Middle.Initial                        Title                                 Department.Name
 [7] Annual.Salary                         Gross.Pay.Last.Paycheck               Gross.Year.To.Date
[10] Gross.Year.To.Date...FRS.Contribution year_of_birth                         marital_status
[13] street_address                        postal_code                           city
[16] State                                 Province                              Country_id
[19] phone_number                          email                                 Education
[22] Occupation                            household_size                        yrs_residence
<0 rows> (or 0-length row.names)
> # Count the number of duplicate rows
> sum(duplicated_rows)
[1] 0
```

Figure 7: Output of the duplication check

Figure 7 above shows the output that is generated when the code is run. As can be seen, there are no duplicate records in this dataset. Thus, no further steps are required regarding duplicate records.

Outliers

Box plots and other visualizations are used to identify outliers in the dataset. Figure 8 below shows a boxplot for the column/attribute “Annual.Salary”. This boxplot was created using the following R code:

```
# Outliers shown in boxplot of the "Annual.Salary"
boxplot(customers$Annual.Salary)
```

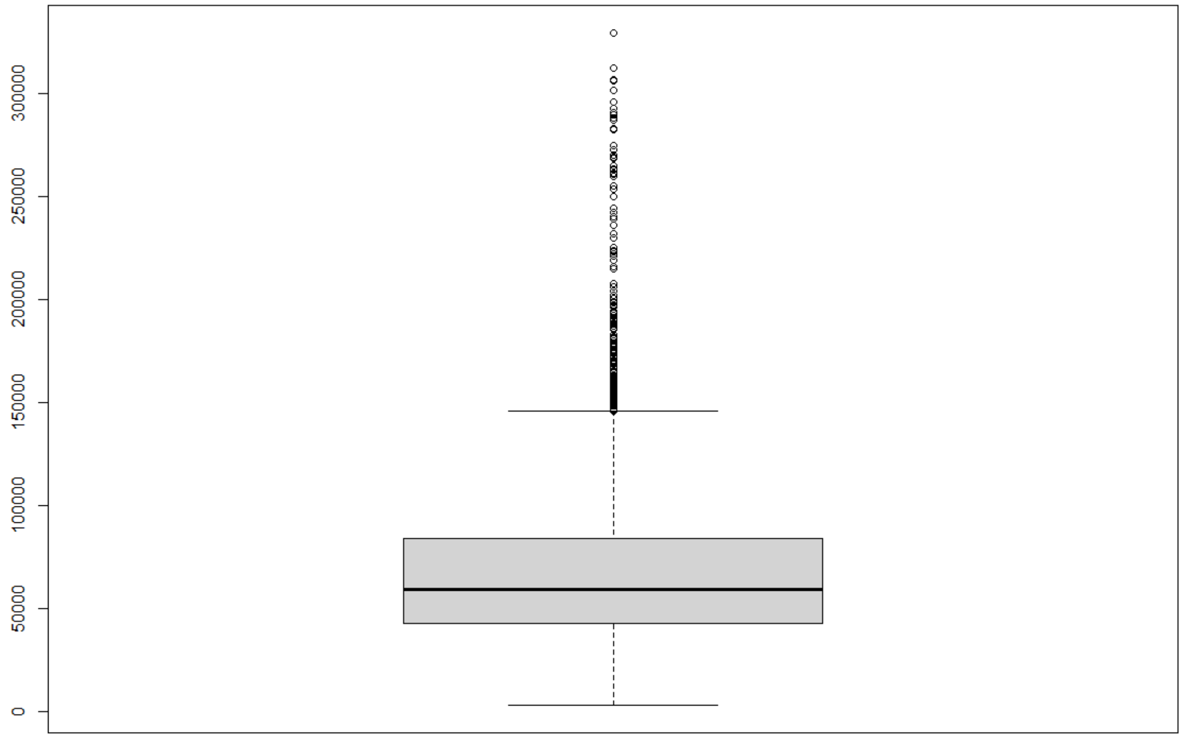


Figure 8: Boxplot of Annual.Salary

It can be stated with confidence that there are several outliers within the “Annual.Salary” column of this dataset (shown by the dots outside the boxplot).

After outliers are located, the subsequent actions need to be performed:

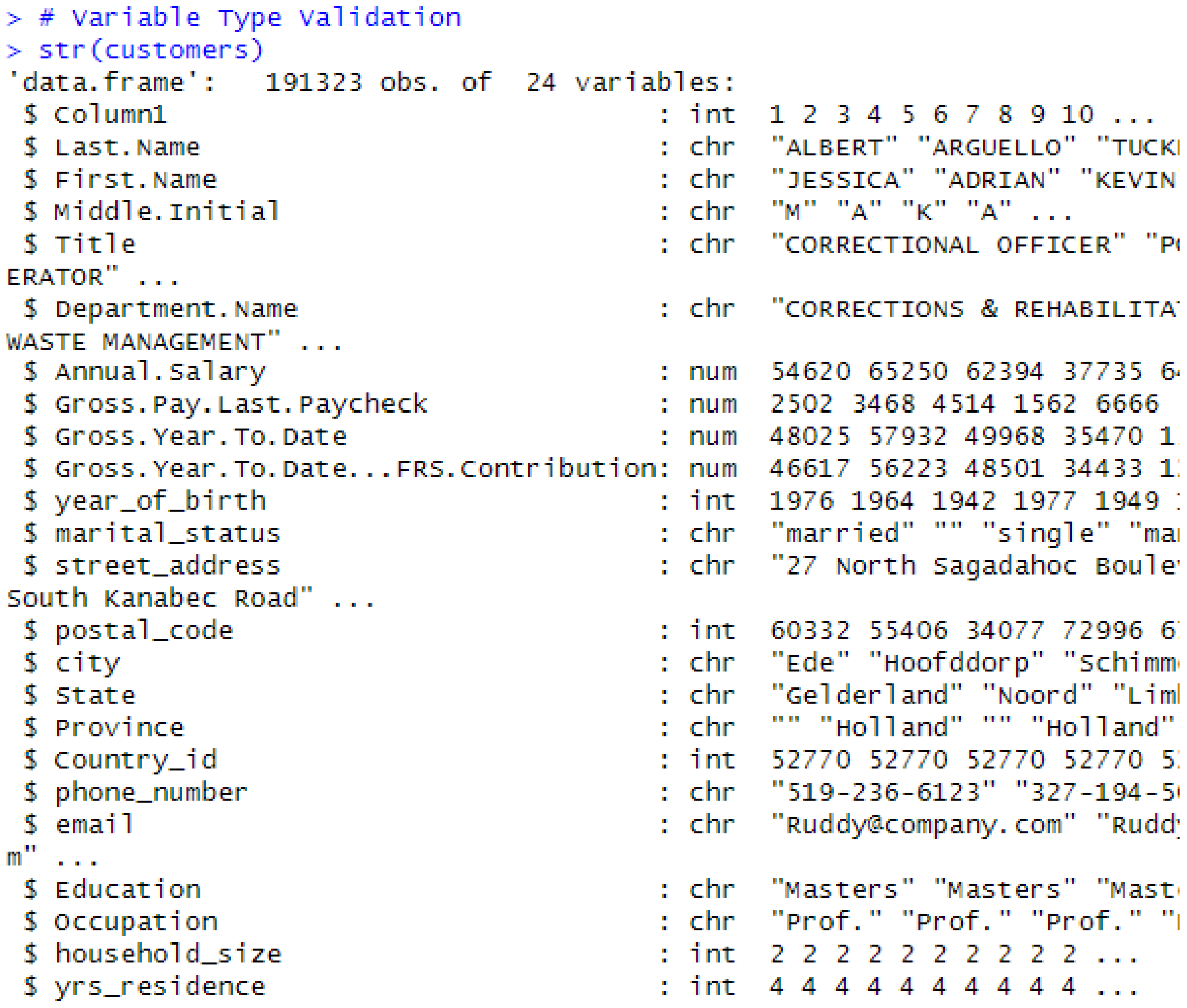
- The dataset is cleaned up of outliers that were obviously the product of incorrect data entry.
- Valid extreme outliers were handled differently, recognizing their possible significance for the analysis.

Variable Type Validation

Verify the proper formatting of continuous and categorical variables. To do this, the following command was used:

```
# Variable Type Validation
str(customers)
```

Figure 9 below shows the output generated by the command:



```
> # Variable Type Validation
> str(customers)
'data.frame':	191323 obs. of  24 variables:
 $ Column1                              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Last.Name                            : chr  "ALBERT" "ARGUELLO" "TUCK
 $ First.Name                           : chr  "JESSICA" "ADRIAN" "KEVIN
 $ Middle.Initial                       : chr  "M" "A" "K" "A" ...
 $ Title                                : chr  "CORRECTIONAL OFFICER" "P
ERATOR" ...
 $ Department.Name                      : chr  "CORRECTIONS & REHABILITA
WASTE MANAGEMENT" ...
 $ Annual.Salary                        : num  54620 65250 62394 37735 6
 $ Gross.Pay.Last.Paycheck              : num  2502 3468 4514 1562 6666
 $ Gross.Year.To.Date                   : num  48025 57932 49968 35470 1
 $ Gross.Year.To.Date...FRS.Contribution: num  46617 56223 48501 34433 1
 $ year_of_birth                        : int  1976 1964 1942 1977 1949
 $ marital_status                       : chr  "married" "" "single" "ma
 $ street_address                       : chr  "27 North Sagadahoc Boule
South Kanabec Road" ...
 $ postal_code                          : int  60332 55406 34077 72996 6
 $ city                                 : chr  "Ede" "Hoofddorp" "Schimm
 $ State                                : chr  "Gelderland" "Noord" "Lim
 $ Province                             : chr  "" "Holland" "" "Holland"
 $ Country_id                           : int  52770 52770 52770 52770 5
 $ phone_number                         : chr  "519-236-6123" "327-194-5
 $ email                                : chr  "Ruddy@company.com" "Rudd
m" ...
 $ Education                            : chr  "Masters" "Masters" "Mast
 $ Occupation                           : chr  "Prof." "Prof." "Prof." "
 $ household_size                       : int  2 2 2 2 2 2 2 2 2 2 ...
 $ yrs_residence                        : int  4 4 4 4 4 4 4 4 4 4 ...
```

Figure 9: Output confirming variable type validation

Analysis of the output given in Figure 4 shows that every attribute of the dataset is read/stored as the correct/appropriate data type. This states that no changes to attribute types need to be made.

Data Distribution Analysis

To find any skewness or non-normal distributions, the distribution of numerical variables (in this case, "Annual.Salary") was evaluated. If required, transformations (such as the logarithmic transformation) must be used to standardize the data. The following R code was used to plot the graph in Figure 10 below:

```
# Data Distribution
hist(customers$Annual.Salary)
```

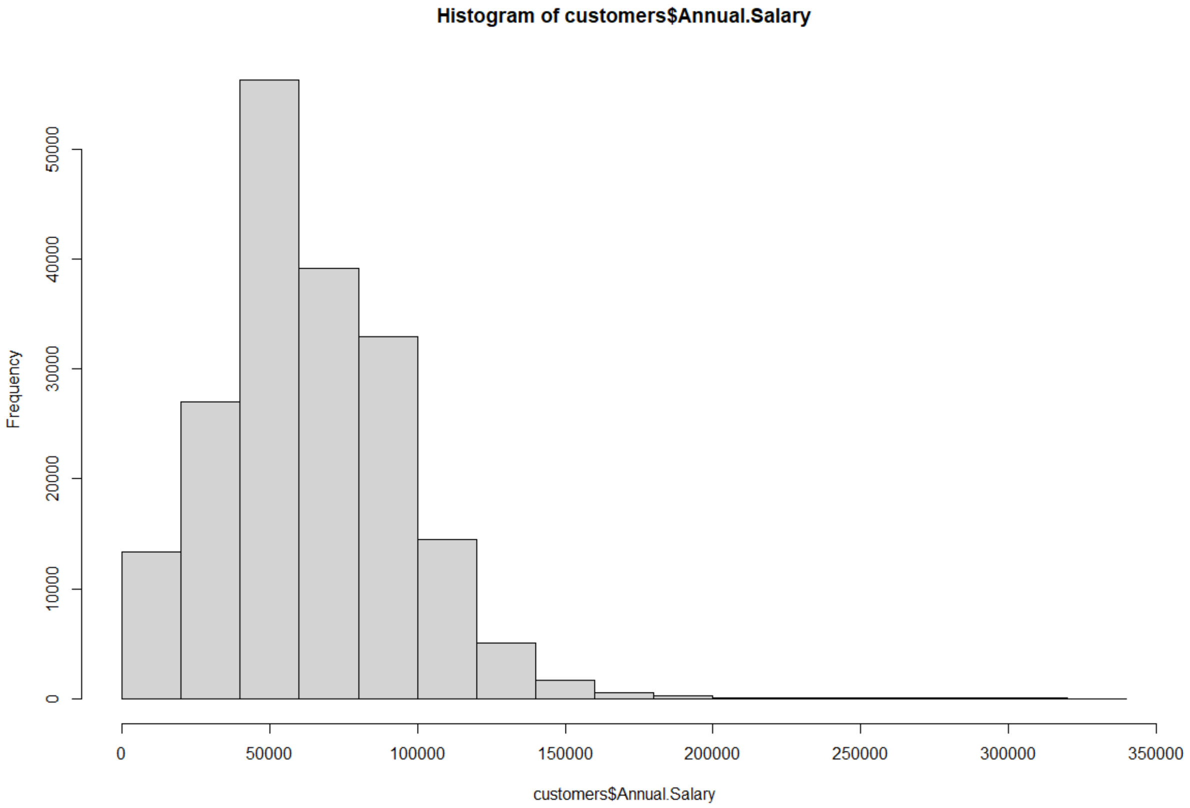
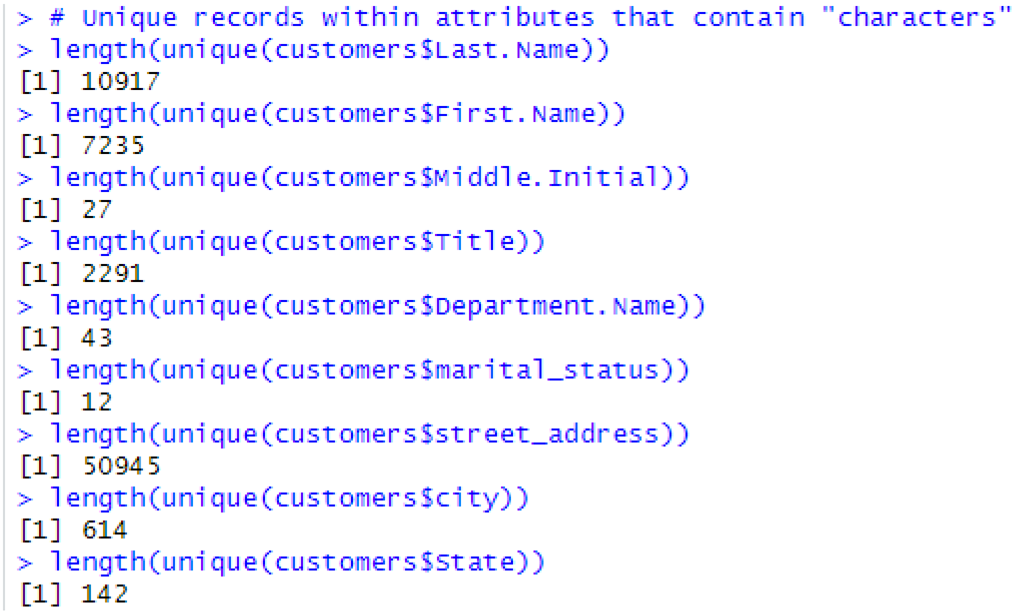


Figure 10: Histogram displaying the distribution of salary

Figure 10 shows a normal distribution of the salary data.

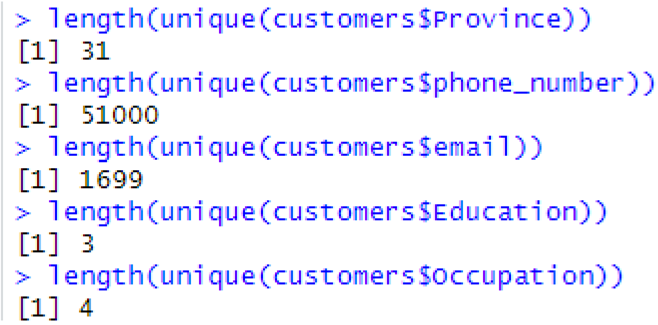
Unique values per attribute/column

The quality of the data is also determined by assessing the values contained in attributes that store/contain values of type "character" or "string". Figures 11 and 12 contains the code used to count the number of unique values within each attribute, along with the output:



```
> # Unique records within attributes that contain "characters"
> length(unique(customers$Last.Name))
[1] 10917
> length(unique(customers$First.Name))
[1] 7235
> length(unique(customers$Middle.Initial))
[1] 27
> length(unique(customers$Title))
[1] 2291
> length(unique(customers$Department.Name))
[1] 43
> length(unique(customers$marital_status))
[1] 12
> length(unique(customers$street_address))
[1] 50945
> length(unique(customers$city))
[1] 614
> length(unique(customers$State))
[1] 142
```

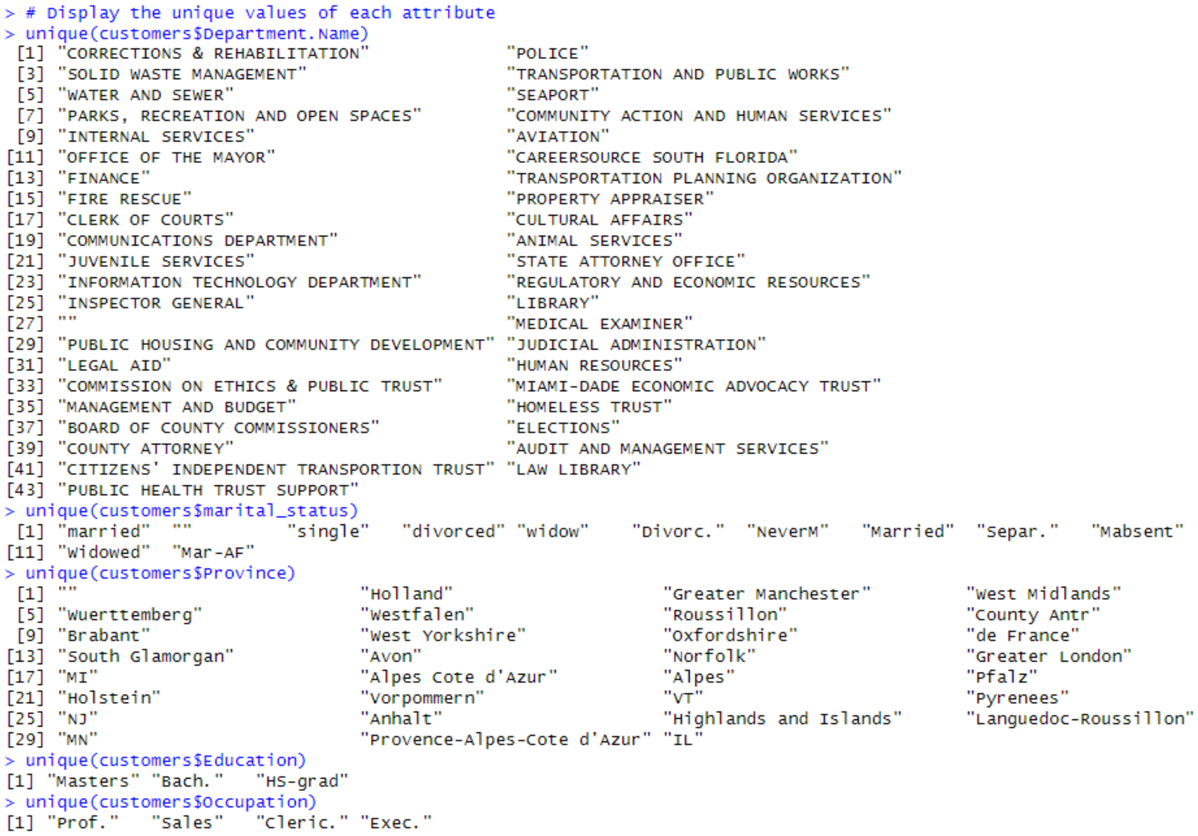
Figure 11: Unique values of "character" attributes – Part 1



```
> length(unique(customers$Province))
[1] 31
> length(unique(customers$phone_number))
[1] 51000
> length(unique(customers$email))
[1] 1699
> length(unique(customers$Education))
[1] 3
> length(unique(customers$Occupation))
[1] 4
```

Figure 12: Unique values of "character" attributes – Part 2

All the attributes that contain a significant number of unique values (more than 100) will be disregarded, seeing as this shows a high cardinality. It can be concluded that these attributes will not play a significant role in the training/creation of the model. Attributes such as name and surname should have high cardinality seeing as people have different names. “Middle.Initial” will also be discarded, it contains all the letters contained in the alphabet (That is why there are 27 unique values). Further analysis is conducted on the attributes with low cardinality. The following was discovered:



```
> # Display the unique values of each attribute
> unique(customers$Department.Name)
 [1] "CORRECTIONS & REHABILITATION"             "POLICE"
 [3] "SOLID WASTE MANAGEMENT"                   "TRANSPORTATION AND PUBLIC WORKS"
 [5] "WATER AND SEWER"                          "SEAPORT"
 [7] "PARKS, RECREATION AND OPEN SPACES"        "COMMUNITY ACTION AND HUMAN SERVICES"
 [9] "INTERNAL SERVICES"                        "AVIATION"
[11] "OFFICE OF THE MAYOR"                      "CAREERSOURCE SOUTH FLORIDA"
[13] "FINANCE"                                  "TRANSPORTATION PLANNING ORGANIZATION"
[15] "FIRE RESCUE"                              "PROPERTY APPRAISER"
[17] "CLERK OF COURTS"                          "CULTURAL AFFAIRS"
[19] "COMMUNICATIONS DEPARTMENT"                "ANIMAL SERVICES"
[21] "JUVENILE SERVICES"                        "STATE ATTORNEY OFFICE"
[23] "INFORMATION TECHNOLOGY DEPARTMENT"        "REGULATORY AND ECONOMIC RESOURCES"
[25] "INSPECTOR GENERAL"                        "LIBRARY"
[27] ""                                         "MEDICAL EXAMINER"
[29] "PUBLIC HOUSING AND COMMUNITY DEVELOPMENT" "JUDICIAL ADMINISTRATION"
[31] "LEGAL AID"                                "HUMAN RESOURCES"
[33] "COMMISSION ON ETHICS & PUBLIC TRUST"      "MIAMI-DADE ECONOMIC ADVOCACY TRUST"
[35] "MANAGEMENT AND BUDGET"                    "HOMELESS TRUST"
[37] "BOARD OF COUNTY COMMISSIONERS"            "ELECTIONS"
[39] "COUNTY ATTORNEY"                          "AUDIT AND MANAGEMENT SERVICES"
[41] "CITIZENS' INDEPENDENT TRANSPORTION TRUST" "LAW LIBRARY"
[43] "PUBLIC HEALTH TRUST SUPPORT"
> unique(customers$marital_status)
 [1] "married"  ""         "single"   "divorced" "widow"    "Divorc."  "NeverM"   "Married"  "Separ."   "Mabsent"
[11] "Widowed"  "Mar-AF"
> unique(customers$Province)
 [1] ""                           "Holland"                    "Greater Manchester"         "West Midlands"
 [5] "Wuerttemberg"               "Westfalen"                  "Roussillon"                 "County Antr"
 [9] "Brabant"                    "West Yorkshire"             "Oxfordshire"                "de France"
[13] "South Glamorgan"            "Avon"                       "Norfolk"                    "Greater London"
[17] "MI"                         "Alpes Cote d'Azur"          "Alpes"                      "Pfalz"
[21] "Holstein"                   "Vorpommern"                 "VT"                         "Pyrenees"
[25] "NJ"                         "Anhalt"                     "Highlands and Islands"      "Languedoc-Roussillon"
[29] "MN"                         "Provence-Alpes-Cote d'Azur" "IL"
> unique(customers$Education)
[1] "Masters" "Bach."   "HS-grad"
> unique(customers$Occupation)
[1] "Prof."   "Sales"   "Cleric." "Exec."
```

Figure 13: Unique records for low cardinality attributes

“Department.Name”, “Province”, “Education” and “Occupation” contain different values that are not related in any way; therefore, these attributes do not show any issues. “Marital_status” contains many different phrases that mean the same things. These phrases need to be transformed/normalized so that only one phrase is used for people that fall under the following categories: “Married”, “Single”, “Widowed” and “Divorced”.

Data Quality Conclusion

After completion of the data quality assessments, it can be concluded that there are many quality problems within this dataset. Data preparation steps must be taken, to create an accurate intelligent recommender model.

Preliminary Data Visualization

All of the following visualisations have been done using RStudio.

1. Correlation Matrix:

The first, and arguably most important data visualization that can be done on the dataset is to create a correlation heatmap. This heatmap will assist in finding strong correlations between attributes.

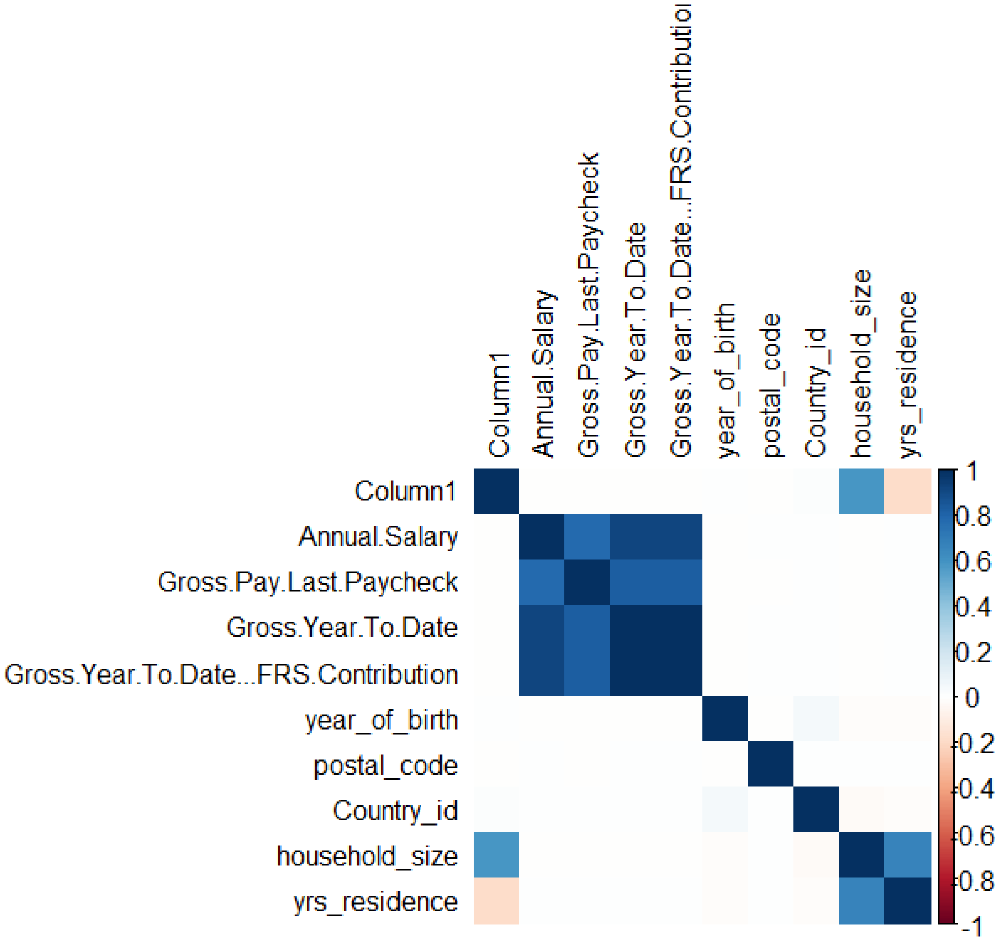


Figure 14 Correlation Matrix

From Figure 11, it can be observed that there is a very strong correlation between the Groos.Year.To.Date...FRS.Contribution attribute and the Gross.Year.To.Date attribute. Furthermore, there is also a strong correlation between the Annual.Salary, Gross.Pay.Last.Paycheck, and Gross.Year.To.Date.

There is also a correlation between the household_size and yrs_residence attributes.

2. Pair Plots

The correlations between attributes can also be visualised using pair plots. The pair plots will visualise the relationships between pairs of attributes. The following pair plots have been created.

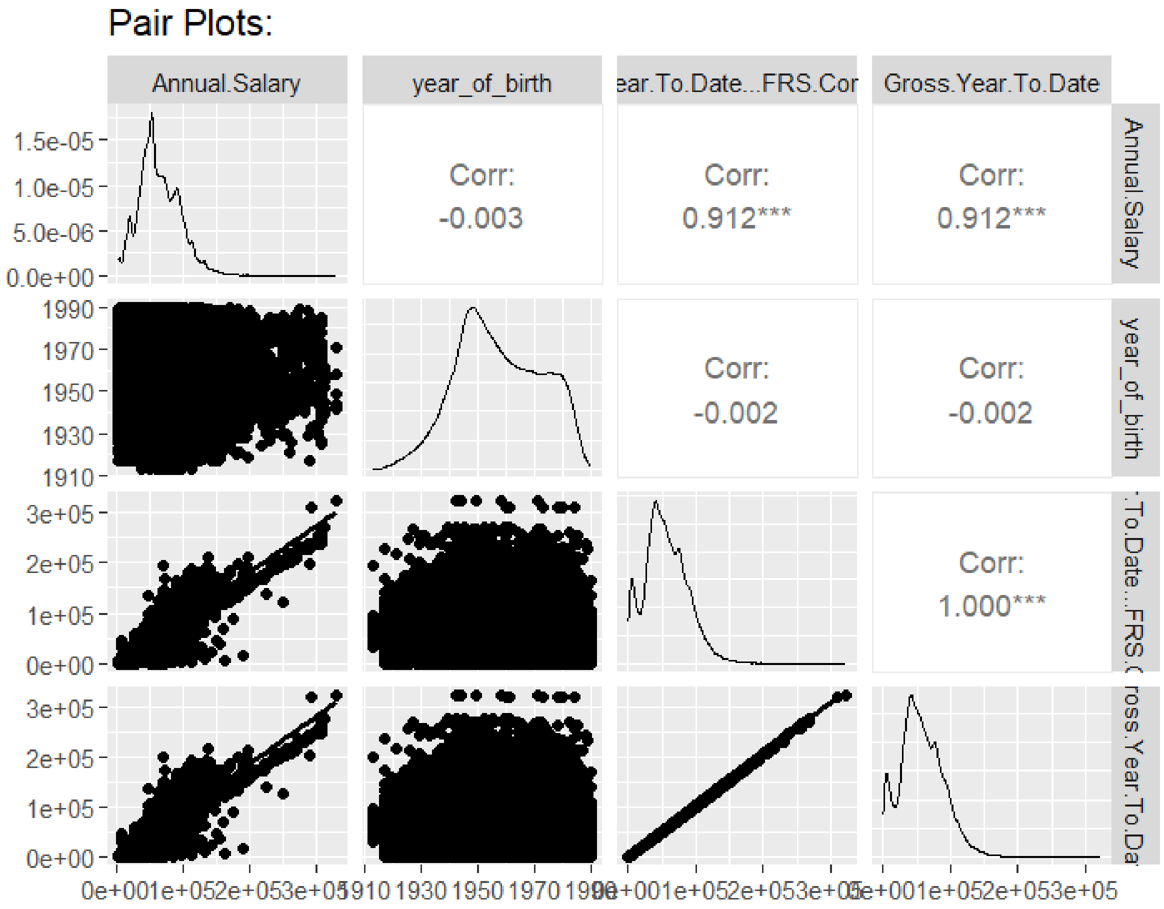


Figure 15 Pair Plots between attributes

The first relationship to observe is that between Annual Salary and Year of Birth. Although there is a correlation between this pair it is very weak. The correlation is also slightly negative, which suggests that as the Year of Birth increases, the salary also decreases. However, as mentioned there is not a strong relationship between these attributes.

The next relationship is that between Annual Salary and Gross Year To Data (FRS Constribution). There is a very strong positive correlation between these attributes. This means that as the Annual Salary increases, so does the Gross Year To Data (FRS Constribution).

The relationship between Annual Salary and Gross Year To Date is observed next. There is a strong correlation between these attributes as seen in the upward slope of the scatter plot. This means that as the Annual Salary increases, so does the Gross Year To Date.

Next, the correlation between the Year of Birth and the Gross Year To Date (FRS Contibution) is examined. There is a weak correlation between the attributes. Furthermore, the correlation is slightly negative. This suggests that as the Year of Birth increases, the Gross Year to Date (FRS Contribution) decreases very slightly.

The correlation between the Year of Birth and Gross Year to Date has a similar correlation as the previous pair. There is a weak negative correlation, and this suggests that as the Year of Birth increases, the Gross Year to Date decreases slightly.

Finally, the correlation between the Gross Year to Date and the Gross Year to Date (FRS Contribution) is examined. The correlation between these attributes is perfect. This is seen in the almost perfect diagonal line of the scatter plot as well as the correlation coefficient of 1. This means that either attribute can be 100% accurately detected using the other.

3. Boxplot for Annual Salary by Department

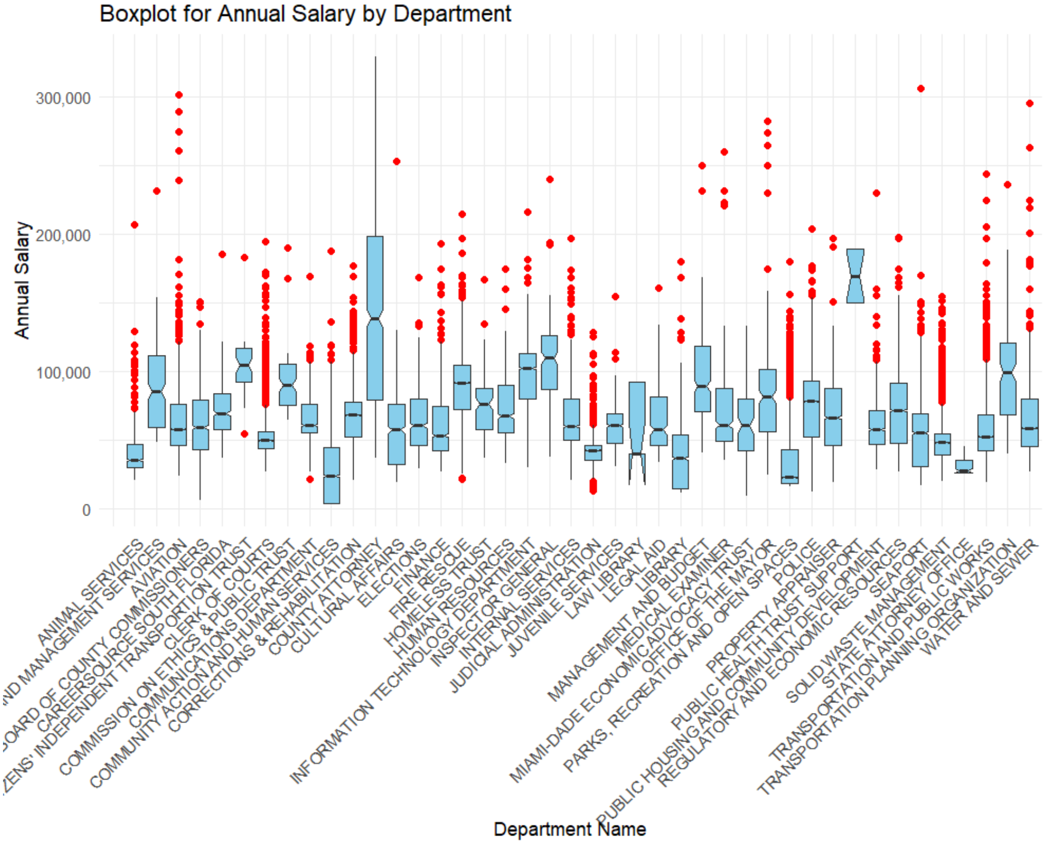


Figure 16 Boxplot for Annual Salary by Department

This graph shows the distribution of the annual salaries across different departments. As cab be seen in Figure 13, most customers have lower salaries with their being a few outliers with higher salaries. The median salaries belonging to different departments also vary, with some having much higher median salaries then others.

4. Scatter Plot for Relationship between Annual Salary and Gross Pay Last Paycheck

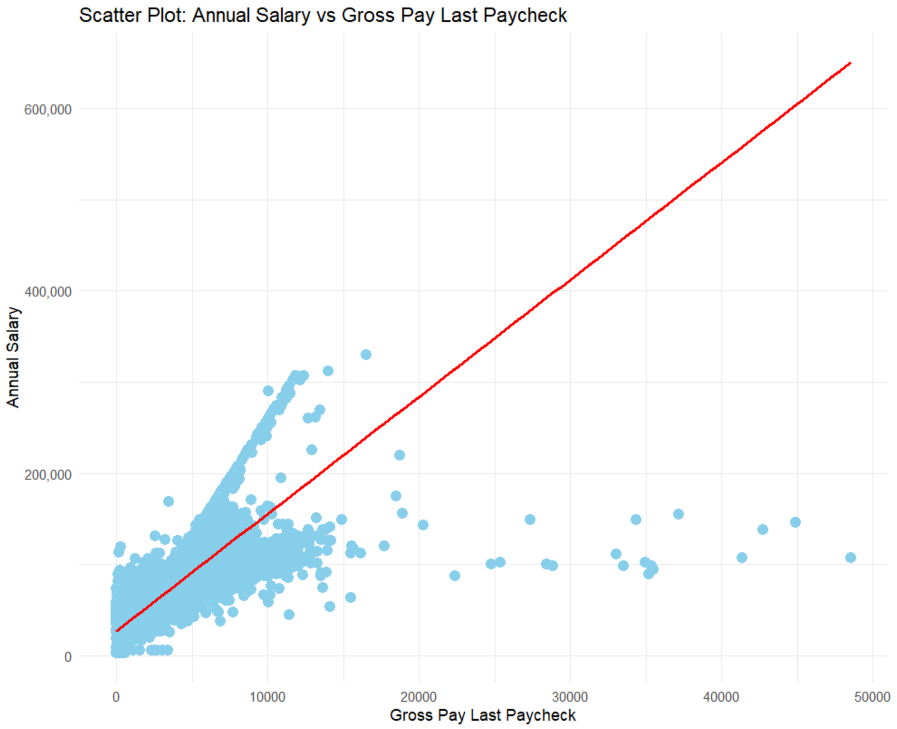


Figure 17 Scatter Plot for Annual Salary vs Gross Pay Last Paycheck

The scatter plot in Figure 14 shows the correlation between the Annual Salary and Gross Pay Last Paycheck. The upward slope of the scatter plot suggests a string positive correlation between the attributes. This means that as the Annual Salary increases, so does the Gross Pay Last Paycheck.

Although the scatter plot is mostly linear, there are outliers that can be observed. This will give un insight into individuals who have unusual salary patterns.

5. Histogram for Distribution of Year of Birth

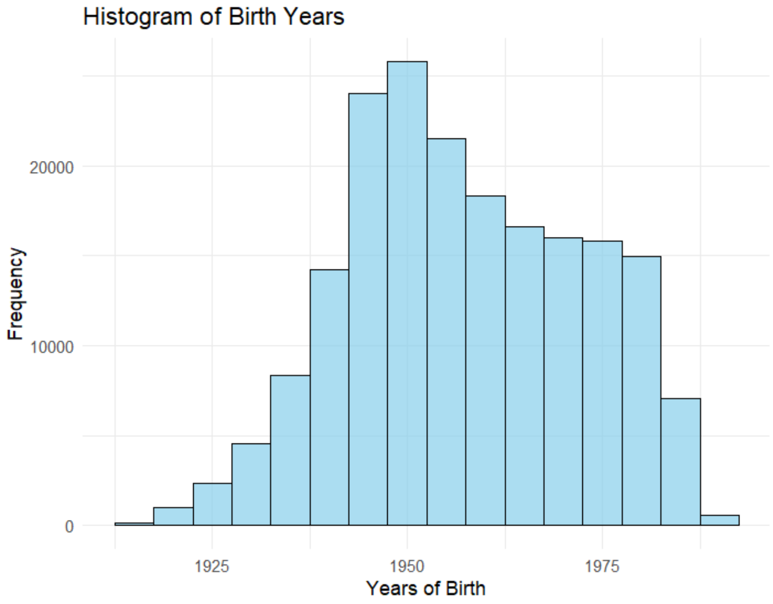


Figure 18 Histogram for Distribution of Year of Birth

The Histogram in Figure 15 shows the distribution of customers across the Year of Birth of the customers.

The first observation that can be made, is that the histogram is skewed to the right. This means that there are more customers born in recent years.

Furthermore, it can be seen that most of the current customers were born around 1950. Although the number of customers born after 1950 have declined, the decline is not major and there is still a large number of the customers that have been born from the 1950s to the 1980s.

Preliminary Dashboard Visualization

Key performance indicators (KPIs) from several marketing channels are tracked by a data visualization dashboard, which then creates an eye-catching report from them. You can use it to help you understand your complex facts, much like an infographic. The use of a data visualization dashboard tool can be advantageous to clients as well as marketers. Figure 19 contains the preliminary dashboard, which will be discussed in further detail.

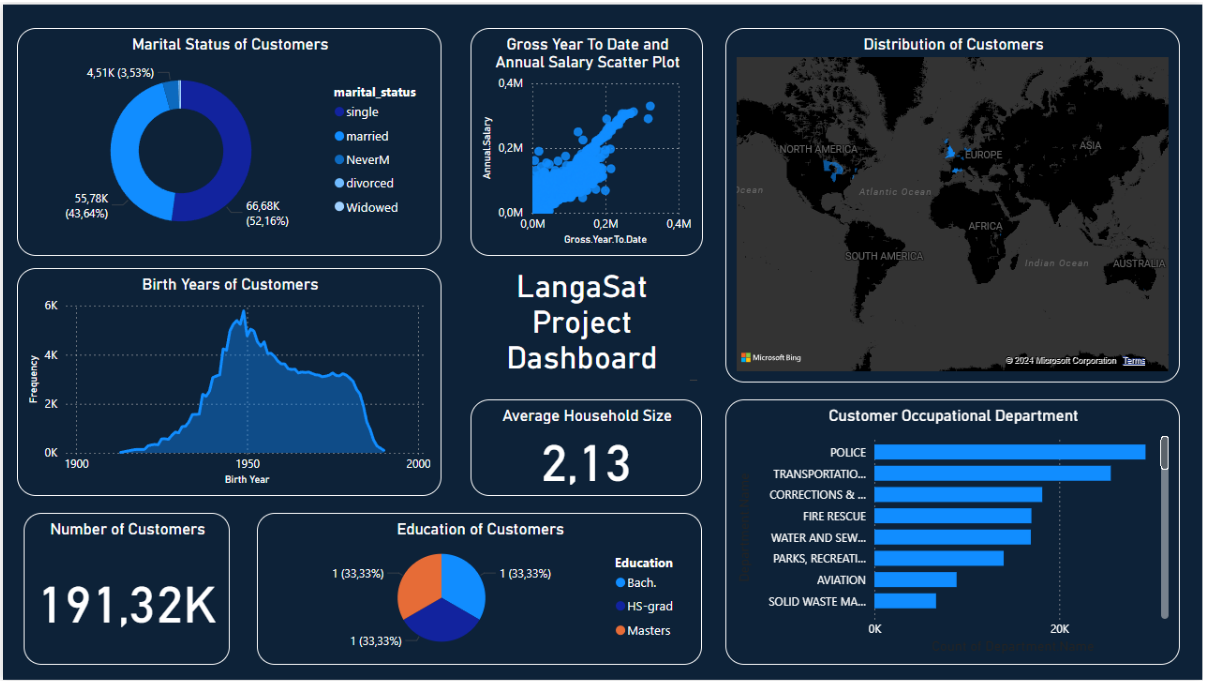


Figure 19: Preliminary Dashboard in Power BI

Each graph/visual will be focused on and explained below:

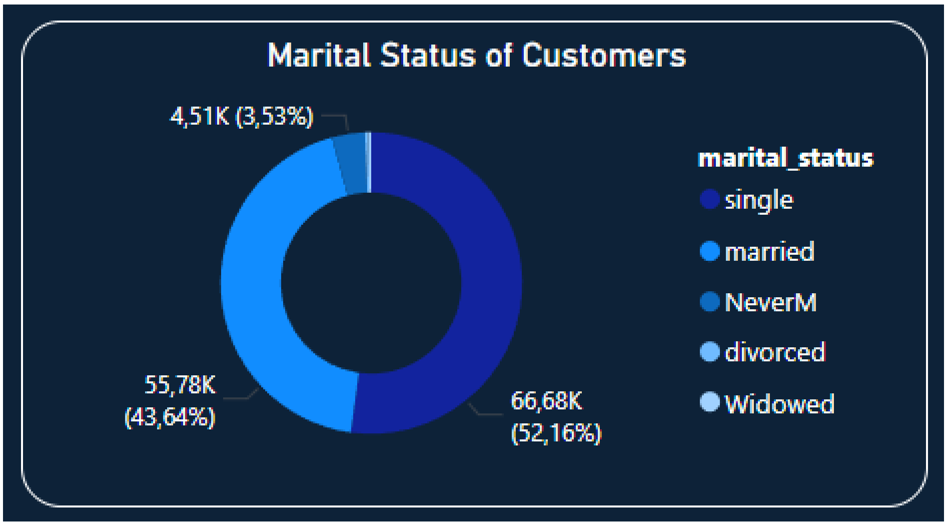


Figure 20: Marital Status of Customers

This visual contains a pie chart (or in this case a "doughnut" chart) for the marital status of the customers. It contains the number for each category captured under marital status along with the percentage. As can be seen majority of the customers are "single", standing at 52.16%.

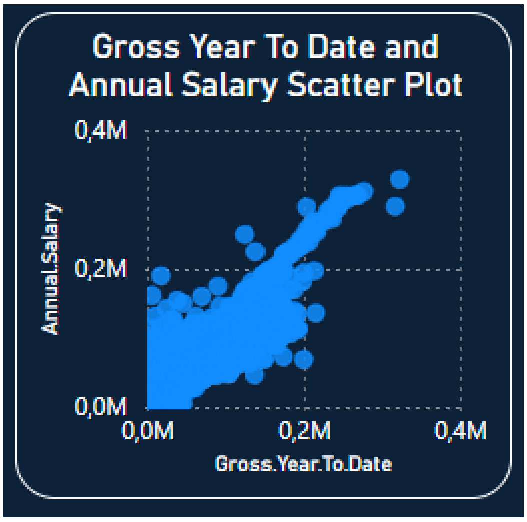


Figure 21: Gross Year to Date and Annual Salary Scatter Plot

This figure displays a scatter plot that contains the "Gross Year to Date" and "Annual Salary" attributes. This shows a correlation between these two attributes, meaning that the annual salary increased directly to the gross year to date.

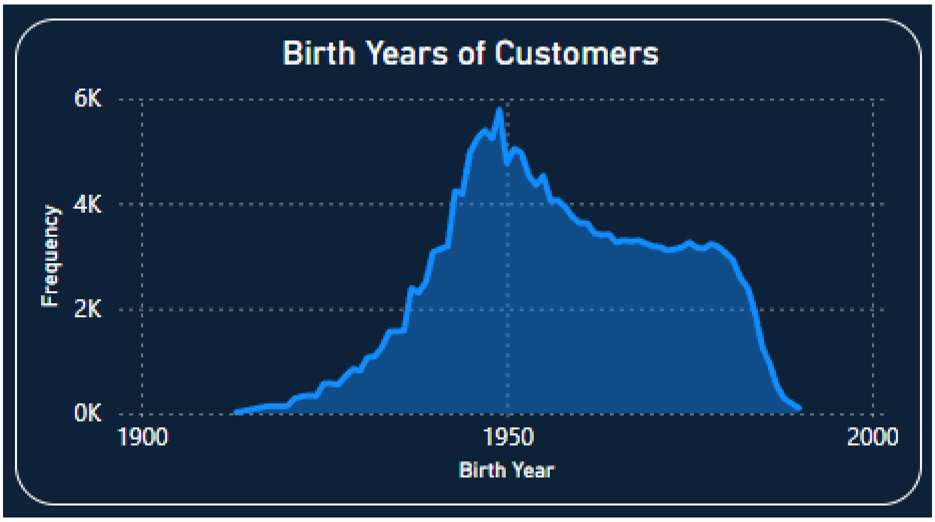


Figure 22: Birth Years of Customers

This figure displays a line graph that shows the frequency of people's birth years. It can be seen that the year the greatest number of customers were born was around 1950.

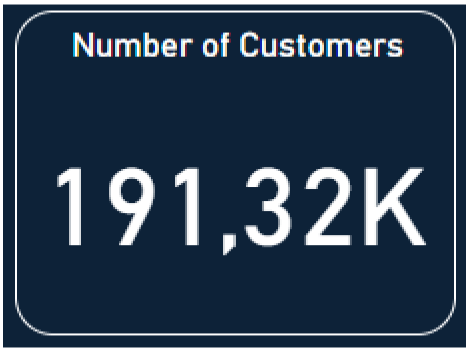


Figure 23: Number of Customers

This figure contains a simple "card" that displays the number of customers the company currently has.

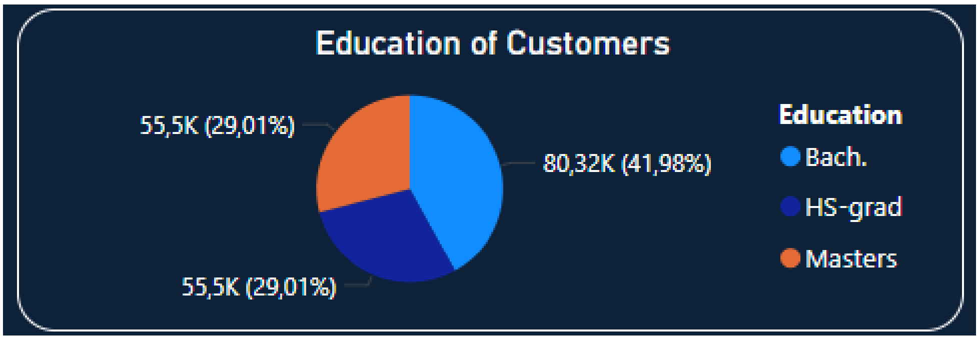
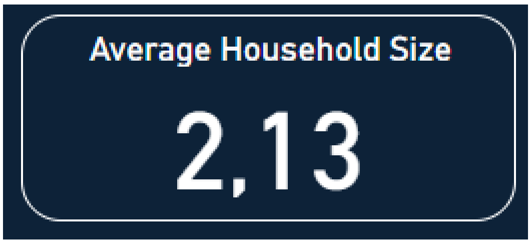


Figure 24: Education of Customers

This simple pie graph displays the highest educational level of the customers. It can be concluded that most of the customers have a bachelor's degree.

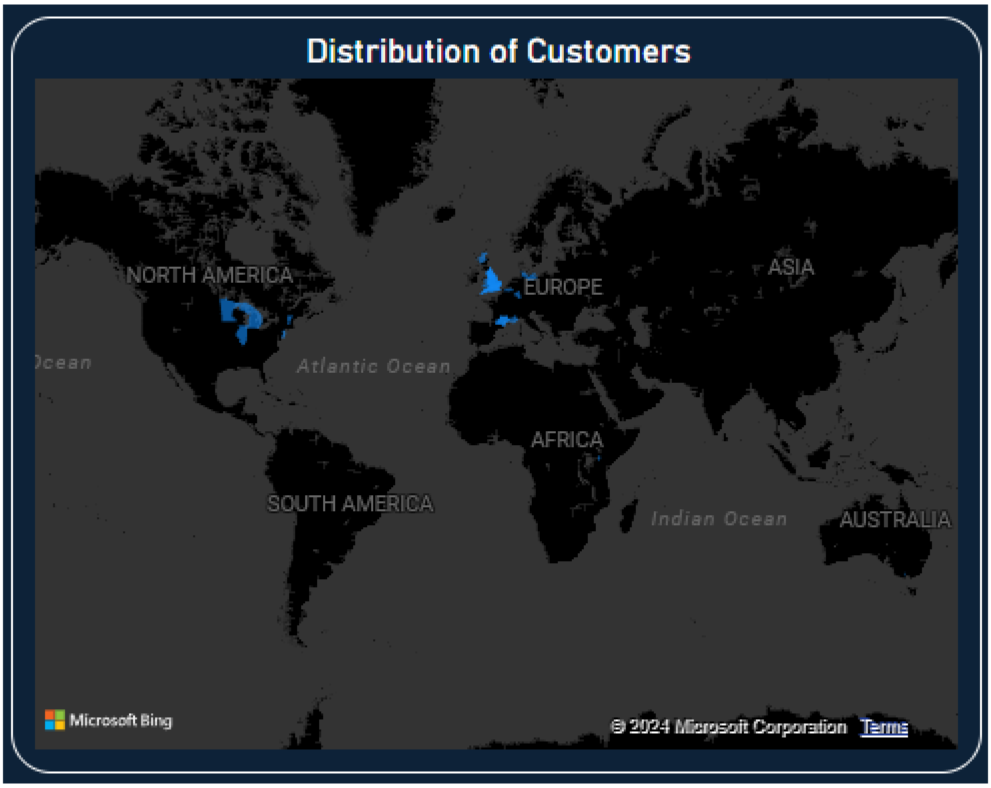


Average Household Size

2,13

Figure 25: Average Household Size

This is another simple “card” visual that displays the average amount of people per household.



Distribution of Customers

Figure 26: Distribution of Customers

Figure 26 above contains a map of the earth, and within the map there are blue parts which shows were the customers are based.

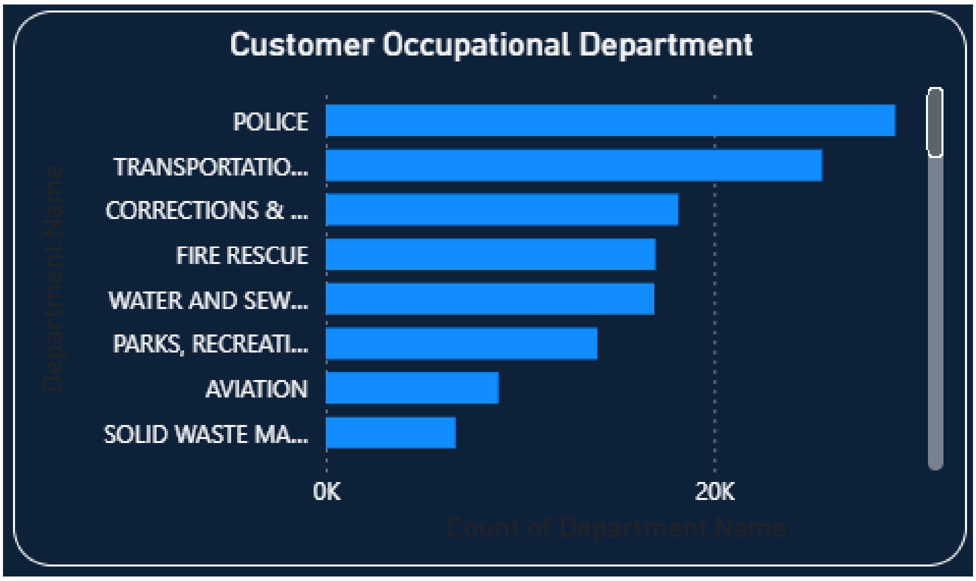


Figure 27: Customer Occupational Department

The last visual contains a bar graph that displays the number of customers within each occupational department. The greatest number of customers are in the "Police" department.

Conclusion

The data was investigated in this first project milestone to fully comprehend its quality, structure, and applicability for developing a LangaSat customer eligibility classification algorithm. Missing values, outliers, and inconsistent data formatting were among the problems found during the data quality evaluation. These issues must be fixed during the data preparation stage. Intriguing trends within consumer demographics like education and marital status were also found, as were relationships across variables, most notably the substantial connections between wage and other financial indicators.

A helpful summary of the main linkages and patterns in the dataset was given by the preliminary visuals. These realizations will direct the project as it continuous on to the CRISP-DM process's subsequent stages, which include feature engineering, data cleansing, and model construction. Through the resolution of data quality concerns and the utilization of important variables, there is the strong potential to develop a solid model that enhances LangaSat's client eligibility assessments.

This significant achievement lays the groundwork for the next phases, which will guarantee clean data and well-informed insights as the project works to improve LangaSat decision-making accuracy.

Bibliography

Awati, R., 2023. *garbage in, garbage out (GIGO)*. [Online]
Available at: https://www.techtarget.com/searchsoftwarequality/definition/garbage-in-garbage-out
[Accessed 4 October 2024].

awork, 2024. *Success Criteria*. [Online]
Available at: https://www.awork.com/glossary/success-criteria#:~:text=Success%20criteria%20are%20measurable%20and,order%20to%20be%20considered%20successful
[Accessed 1 October 2024].

Boogaard, K., 2024. *What is a risk matrix?*. [Online]
Available at: https://www.wrike.com/blog/what-is-risk-matrix/
[Accessed 1 October 2024].

IBM, 2024. *What is data quality?*. [Online]
Available at: https://www.ibm.com/topics/data-quality#:~:text=Data%20quality%20measures%20how%20well,governance%20initiatives%20within%20an%20organization
[Accessed 1 October 2024].

Javapoint, 2024. *Risks of Machine Learning*. [Online]
Available at: https://www.javatpoint.com/risks-of-machine-learning
[Accessed 1 October 2024].

Malsam, W., 2022. *Project Assumptions: A Quick Guide*. [Online]
Available at: https://www.projectmanager.com/blog/project-assumptions
[Accessed 1 October 2024].

monday.com, 2024. *Project Assumptions: What They Are and Why You Should Care*. [Online]
Available at: https://monday.com/blog/project-management/project-assumptions/
[Accessed 1 October 2024].

Roshaan, E., 2024. *Top tips: Watch out for these 4 machine learning risks*. [Online]
Available at: https://blogs.manageengine.com/corporate/general/2024/04/04/top-tips-watch-out-for-these-4-machine-learning-risks.html
[Accessed 1 October 2024].

SMARTVISION, 2024. *What is the CRISP-DM methodology?*. [Online]
Available at: https://www.sv-europe.com/crisp-dm-methodology/
[Accessed 2 October 2024].

Tavish, 2024. *12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2024)*. [Online]
Available at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
[Accessed 1 October 2024].