

Insure ABC

Database Documentation and Insights

**Joseph Deery**

**40078793**

Contents

[1.0 Introduction 2](#_Toc57153470)

[2.0 Database Documentation 2](#_Toc57153471)

[2.1 Database development in Microsoft Access 2](#_Toc57153472)

[2.2 Data Quality Report 6](#_Toc57153473)

[3.0 Insights Report 9](#_Toc57153474)

[Appendix 1: SQL Code 13](#_Toc57153475)

[Appendix 2: R Code 15](#_Toc57153476)

[Appendix 3: Screenshots 17](#_Toc57153477)

# 1.0 Introduction

This project is designed around the business problem posed by Insure ABC, who are seeking to set up an analytics function to enable a deeper analysis of the characteristics of customers across different insurance policies, as well as finding out what the preferred marketing communication channels are to allow for better strategic planning and data-driven decision making. The purpose of this initial project is to create a small database from the data provided by Insure ABC, for both ease of access and security, so that further analytics tasks can be carried out by Data Analysts within the function. A data quality analysis will be completed followed by an early descriptive summary on the new database to demonstrate its functionality as well as providing some preliminary analysis and recommendations. The project is split into 2 main sections. The Database Documentation in Section 2.0 outlines the steps taken to develop and created the Database and Analytics Base Table, as well as details on the structure and content of each table and the ABT, and how they were joined in Microsoft Access. Section 2.2 also contains a Data Quality Report detailing the quality issues both identified and addressed in RStudio. The Insights Report in Section 3.0 contains a discussion on the key insights from the initial descriptive analysis carried out using SQL, including recommendations to the business as well as some limitations.

# 2.0 Database Documentation

## 2.1 Database development in Microsoft Access

Insure ABC provided 4 Microsoft Excel tables showing details of Customer Characteristics and Insurance policies taken out between 1/1/2018 and 31/12/2018. The tables are labelled “Customer”, “Health Policies”, “Motor Policies” and “Travel Policies.” These tables were imported and combined using Microsoft Access into and a 5th table labelled “Insurance ABT” via their primary key columns “HealthID”, “MotorID” and “TravelID” to the respective foreign keys of the same label in the “Customer” Table. The left-join function was used to combine each table, to retain every row of data in the ABT for analytics purposes. The content of each table in the database is shown below:

**Table**: Customer

**Description**: Contains Customer ID’s across the 3 insurance policies, as well customer characteristics such as location, credit-card type, gender and preferred marketing channels. The foreign keys in this table were used to join the tables into a combined ABT.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| CustomerID | Number | Primary Key |
| Title | Short Text | Mr/Miss/Mrs etc |
| GivenName | Short Text | self explanatory |
| Surname | Short Text | self explanatory |
| CreditCardType | Short Text | Visa or Amex |
| Occupation | Short Text | Job Type |
| Gender | Short Text | Male/Female |
| Age | Number | self explanatory |
| Location | Short Text | Rural or Urban |
| PrefChannel | Short Text | Preferred Marketing Communication (Phone/SMS/Email) |
| MotorID | Number | Foreign Key (Motor\_Policies) |
| HealthID | Number | Foreign Key (Health\_Policies) |
| TravelID | Number | Foreign Key (Travel\_Policies) |

**Table** : Motor\_Policies

**Description**: Contains Primary motorID Key, as well as customer characteristics across the motor Insurance Policy such as start/end date, vehicle Value, vehicle type, number of claims and the type of Insurance Policy held by the customer.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Details** |
| motorID | Number | Primary Key |
| PolicyStart | Date/Time | Policy Start Date |
| PolicyEnd | Date/Time | Policy End Date |
| exposure | Number | Exposure level to accidents (level 0-1) |
| clm | Short Text | occurrence of claim(yes/no) |
| numclaims | Number | number of claims |
| veh\_body | Short Text | type of vehicle(BUS/CONVT/COUPE/HBACK/HDTOP/MCARA/MIBUS/PANVN/RDSTR/SEDAN/STNWG/TRUCK) |
| veh\_age | Number | 1 to 4(1 being youngest age level) |
| LastClaimDate | Date/Time | Date of the most recent insurance claim |

**Table :** Health\_Policies

**Description:** Contains Primary healthID Key as well as health insurance policy dates, types and dependents.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Details** |
| healthID | Number | Primary Key |
| HealthType | Short Text | Type of Health Insurance Held (Level1, Level2 or Level3) categories |
| PolicyStart | Date/Time | Policy Start Date |
| PolicyEnd | Date/Time | Policy End Date |
| HealthDependentsAdults | Number | Number of dependent adults on customers' health policy |
| DependentsKids | Number | Number of dependent children on customers' health policy |

**Table :** Travel\_Policies

**Description:** Contains Primary travelID Key as well as type of travel insurance held by each customer and the start/end dates of each insurance policy

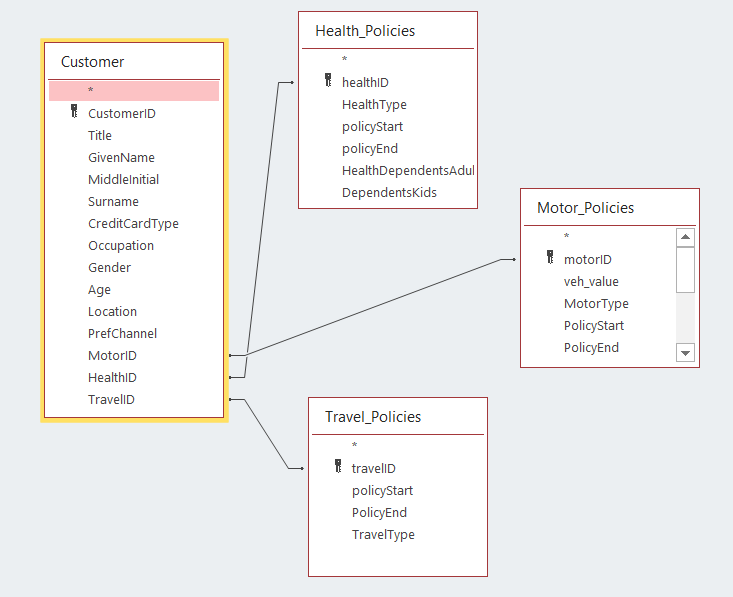
|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Details** |
| travelID | Number | Primary Key |
| PolicyStart | Date/Time | Policy Start Date |
| PolicyEnd | Date/Time | Policy End Date |
| TravelType | Short Text | Travel Insurance Type held by customer (Backpacker,Senior,Business,Premium or Standard) |

**ABT:** Insurance\_ABT

**Description:** This is an ABT created by combining each of the above policy tables via the 3 foreign keys in the customer table. The LEFT JOIN function was used in SQL, as shown in Appendix 1. This maintained the total number of records in the database of 4085, whereas an INNER JOIN would reduce that number significantly to 975 records, as each customer included should have all 3 insurance policies. This was determined not conducive for this project and a potential hinderance to any future analytics tasks carried out on this database. The table below shows the columns included in this ABT. Figure 1 also shows the concept for joining the primary and foreign keys using SQL in MS Access.

|  |  |  |
| --- | --- | --- |
| **Column** |  |  |
| CustomerID | T1motorID | Motor\_Policies\_PolicyEnd |
| Title | Customer\_healthID | exposure |
| GivenName | Health\_Policies\_HealthID | clm |
| MiddleInitial | T1\_PolicyStart | numclaims |
| Surname | T1\_PolicyEnd | clmst0 |
| CreditCardType | HealthDependentsAdults | veh\_body |
| Occupation | DependentsKids | veh\_age |
| Gender | Motor\_Policies\_motorID | LastClaimDate |
| Age | veh\_value | Travel\_Policies\_travelID |
| Location | MotorType | PolicyStart/PolicyEnd |
| PrefChannel | Motor\_Policies\_PolicyStart | TravelType |

**Figure 1**: MS Access Design View showing concept of joining 4 tables via Primary and Foreign Keys



Microsoft Access was a good solution for combining tables and creating this small database for access and use by analysts in the company, however it does have a few limitations and there are options for overcoming these. A well-known limitation of MS Access is that it is not “groupware”, meaning it cannot be accessed over the internet for analytics purposes, nor can multiple people work on the database over a global network. It can be accessed by up by a maximum of 255 users over a local network, however as the function and company grow other options should be considered. Microsoft Access is probably adequate for a few users or administrators, but as the company expands this function further it could look to incorporate something like MongoDB using nodeJS as an alternative, which will allow real-time workflow incorporation, increasing efficiency considerably. Access is also integrated into Windows systems exclusively, and with Mac being a popular OS amongst analysts and data engineers, limiting the function to Windows by solely using MS Access would not only be unpopular but could lead to a sluggish analytics department overall. Access also has limitations on the security side. Data encryption is difficult and cumbersome. With such personal data being held on customers, including names, types of car, location etc, particularly in an era where GDPR and data security are becoming ever more important, this is not ideal. If security is a real concern, SQL server offers both Column Level and Transparent Data encryption integration, and more recently, the DbDefense Database Encryptor which can be installed and connected directly to any SQL Server Database, overcoming this issue Overall, MS Access, like most RDBMS’s, has its upsides and limitations. It is widely available, relatively simple to use and great for small database management and analytics, however its usefulness ultimately depends on its purpose. If the company wish to expand the function beyond a few users, then it is perhaps necessary to look at some of the recommendations made. Company growth in terms of data and customer-base will likely dictate this, along with an increasing need for greater data security.

## 2.2 Data Quality Report

The Data Quality Analysis was carried out using R/RStudio. The tables were imported and combined using the “dplyr” package with left joins. The new Insurance ABT was then explored for any data quality issues such as extreme outliers, incorrect data, or missing data. The tables below show a data quality report for both the continuous and categorical features of the ABT prior to any fixes being made:

*Continuous Features*:

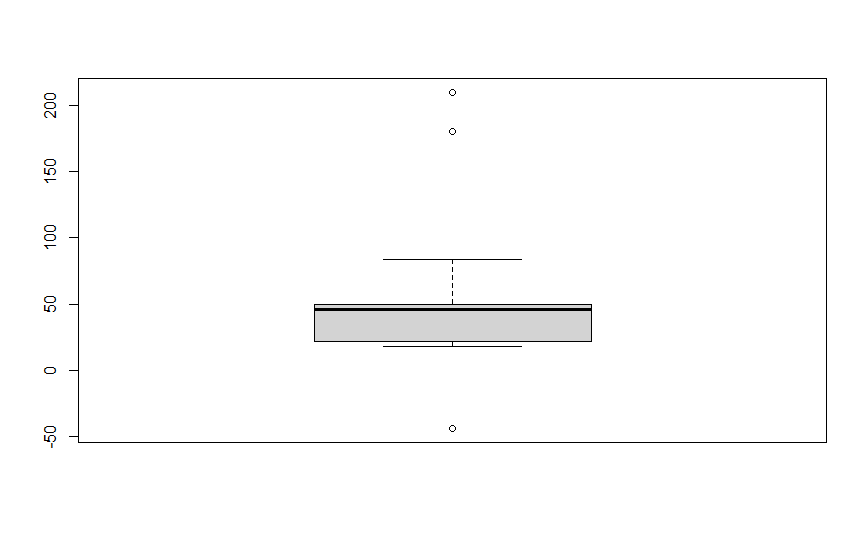


*Categorical Features*

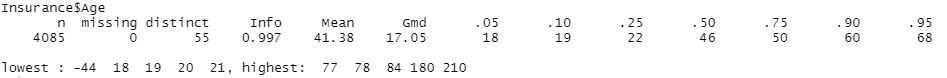


The first step with RStudio is to ensure variables were in the correct format. RStudio will automatically save categorical variables as character vectors, so several variables had to be converted to factors to be of use in any preliminary/further analysis. Some basic explorations could then be carried out to identify outliers in the data. Figure 2 for example shows the age distribution prior to any data quality fixes, with clear invalid outliers between the -44 and 200 mark. The describe function in the “Hmisc” package shows the exact spread of the data (shown in figure 3). The new age range was set between 16 and 110 to remove invalid outliers, ensuring a more normal distribution of the data. It was also important to be conscious of maintaining the integrity of the original data, so, as with all removal of outliers, some critical judgement was used here. This approach was applied to each variable where outliers were identified. The number of child dependents column for example showed an extreme outlier of over 40, heavily right skewing the data. The new limit was set as a maximum of 10, while there were also some extreme vehicle values that were removed. It is worth noting that, unlike the age variable, these could be valid outliers, however they were deemed so extreme that they would be counterproductive for any analysis carried out on the dataset, skewing any possible visualisations or the inferential/predictive accuracy of any models produced from this dataset.

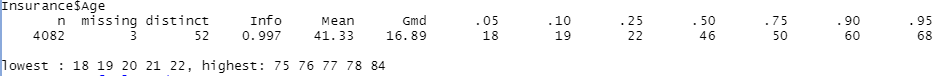
**Figure 2**: Boxplot showing age distribution of Insurance Data Frame (pre quality fixes)



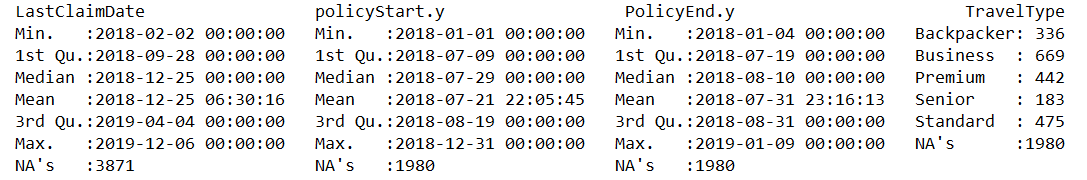
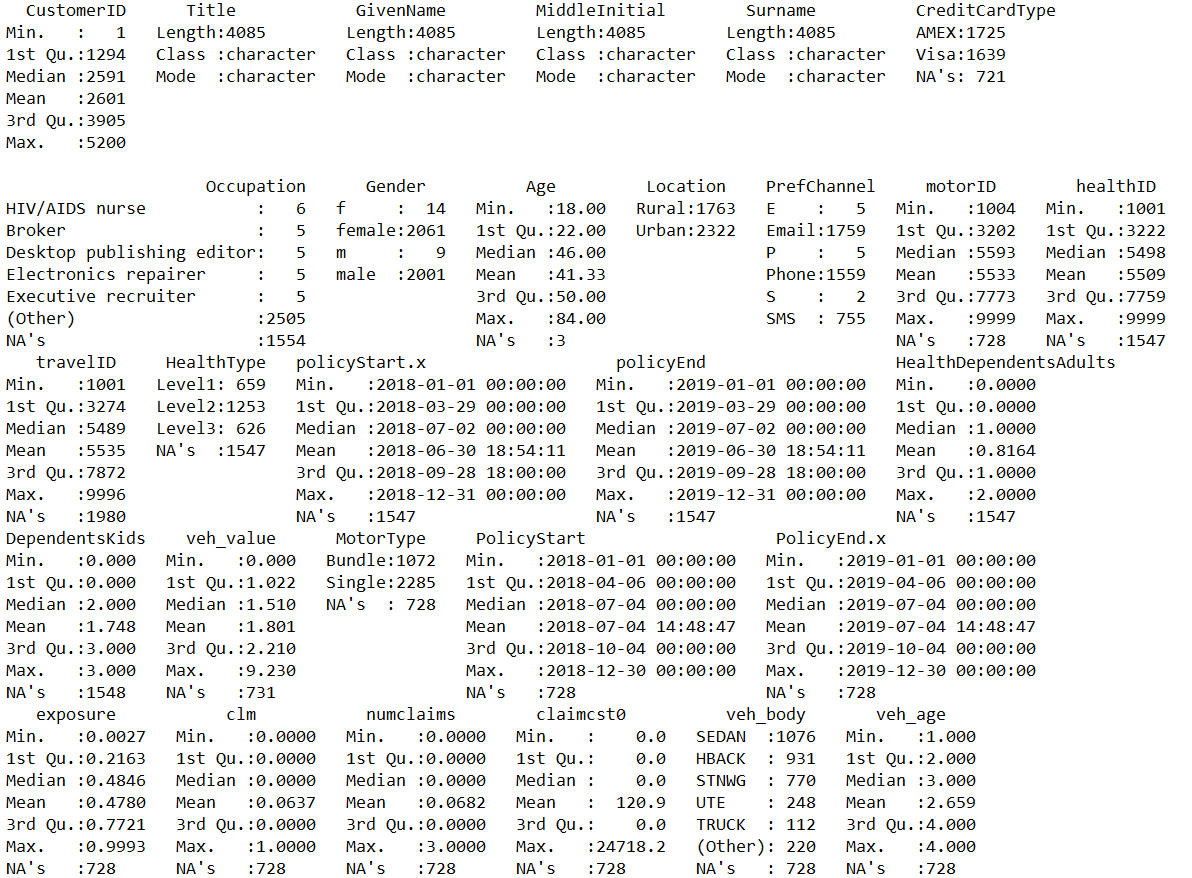
**Figure 3**: Output showing distribution of age data (pre quality fixes)



**Figure 4:** Output showing post quality fix age distribution statistics



**Figure 5:** Summary output for clean Insurance Data

****

There were also several missing values in this dataset. These were determined to be a characteristic of the dataset itself, rather than errors, therefore the decision was taken to leave them untouched. For example, not every customer in the dataset had all 3 Insurance Policies. Missing Data may also be of use in further analytics tasks and can provide use in predictive modelling. Ultimately, analysts can decide to remove data on an individual case-by-case basis. Figure 4 shows a clean distribution output for age, with the new range now sitting between 18 and 84, with a mean of 41, while Figure 5 shows the summary for the entire clean dataset. Further outputs from this section can be found in Appendix 3. A few steps can be taken to prevent/limit any quality issues in future datasets. Cleaning the data at source level is the best place to start, thus limiting time spent cleaning data at the database level. This can be done at admin level, or the software/website used to collect data could be programmed to normalise data at the data entry level. This process can be automated, saving time. Another option is to integrate a full ETL phase into the business. Use of ETL technologies will not only allow for better organisation of data, but it will improve data quality, reducing time at the analytics level and improving ROI. Its also highly scalable as more data is available for use within the business.

3.0 Insights Report

Gaining insights through data analysis has become a huge part of the insurance industry in the last decade. Understanding the preferences, tastes, habits, and needs of a company’s customer base is key driving-force for strategic decision making. The purpose of the preliminary analysis carried out in this project is to provide Insure ABC with information that will be useful in understanding the characteristics of their consumer base across different policies, while also informing potential marketing strategies through a deeper understanding of marketing channel preferences. This early descriptive analytics process should also provide the foundations for deeper future analysis, but it is limited in its depth and scope. These limitations will be highlighted, while recommendations of how the business could improve decision making further in this area will be produced. Figure 6 shows the outputs of some preliminary descriptive analysis carried out in MS Access on the combined Insurance ABT. It shows a relatively equal split of gender on the whole dataset, with slightly more females than males in the customer database. This could simply reflect the demographics in the area. Average age is also around the 41 and similarly, this could be explained by area demographics. Not much useability can be gained from these insights alone, and further analysis has been carried out below to show the relationships between age and preferred communication channels for example, however it is useful to know these figures for decision making purposes on the marketing side.

**Figure 6**: Basic Descriptive SQL outputs for ABT

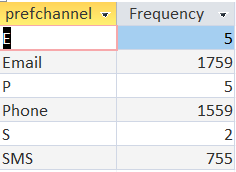
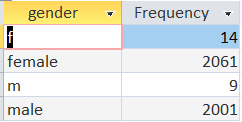


Figure 6 also shows preferred marketing channels over the entire population, with SMS clearly being the least popular, and email and phone being almost equally as popular with email occurring 200 more times. This information can be used to dictate marketing strategy however as further analysis shows, it would be better to target specific age groups with different marketing strategies. There are 1587 distinct occupations in this dataset of around 4000 total customers, meaning there is little use for analysis from this column. A suggestion would be to split these occupations into different sectors or categories for analysis. For example, public/private sector, consultancy, healthcare, manual labour etc. Having distinct categories for occupations can allow for deeper analysis of customer behaviour to determine if there are any patterns in customer product or marketing preferences.

**Figure 7**: Query Output of Age by Marketing Channel Preference (Slit by Age Group, Young, Middle Aged and Senior).

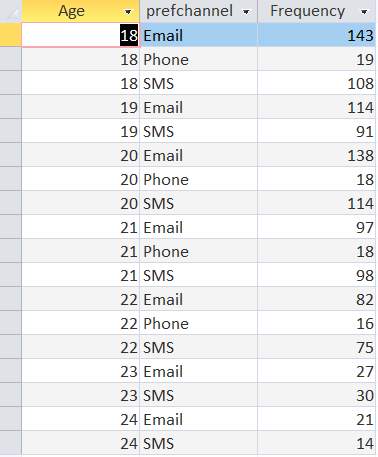
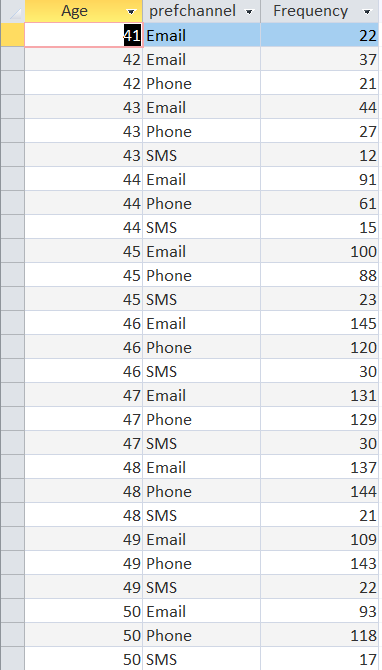
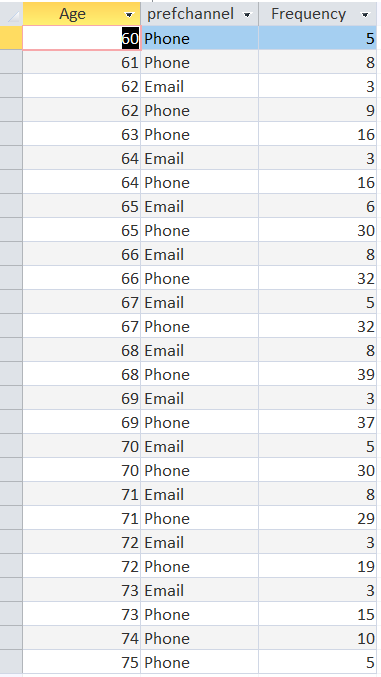


Figure 7 shows the marketing preferences for specific age groups. For the 18 to 30 range emails seem the most popular option. SMS is also popular while phone communications seem to be unpopular. SMS however is the least popular amongst the middle-aged and senior age groups. This is likely explained by generational differences, with younger customers growing up in a highly technological era and therefore more comfortable with less direct marketing channels. On the other hand, as the data shows, as customers are of an older age, telephone marketing takes increasing preference. This can certainly assist Insure ABC in determining different strategies for different age groups, opting to target younger audiences with SMS/Email and older audiences with direct telephone marketing. It is worth remembering that these preferences will likely change with time, as in theory, the senior age groups in 10/15 years should be more comfortable with technology. The company must therefore ensure surveys like these are completed regularly to fully capitalise on the information provided.

**Figure 8**: Query Output for Average Age Range and Insurance Categories

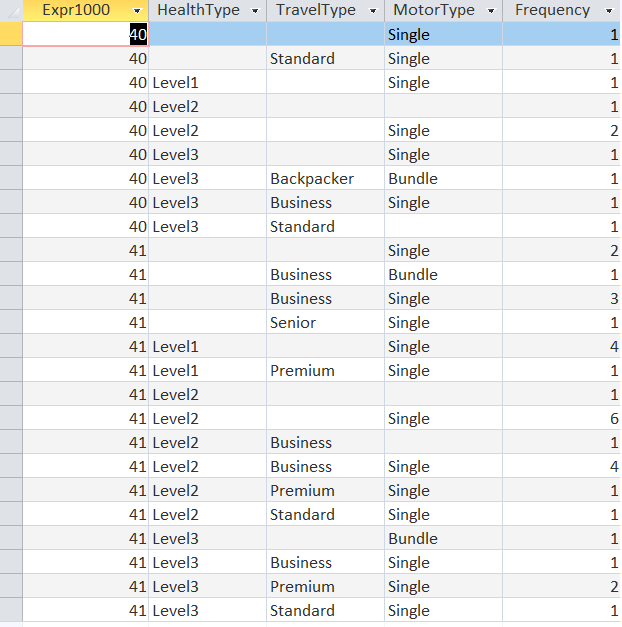
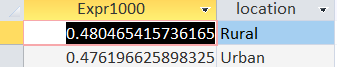


Figure 8 shows the Insurance Types around the average age mark of 41, used to compare the most popular policy for this age. Evidently, from this output, the Single Motor Policy is more popular than Bundle. Level 2 and Level 3 Health Insurance policies are also more popular, while Backpacker seems to be least popular amongst the travel policies. This is likely due to age itself, for example, younger customers are more likely to travel and therefore need this type of insurance. This information can certainly help Insure ABC target customers with more appropriate products based on their age. However, the data here from this output is limited, as seen from the low frequency count. More data on this across more age ranges should help, and a deeper analysis through logistic regression modelling for example would provide a more solid foundation for data-driven decision making on targeted marketing through preferred marketing communications.

**Figure 9:** Output for Exposure Level (to accidents) by location and Travel Insurance Type popularity by location.



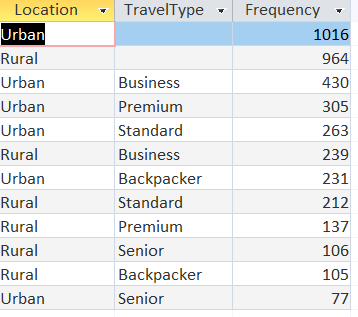


Figure 9 shows popular Travel Insurance types by location (Rural/Urban). Interestingly, business overall is the most popular Travel Insurance Type amongst customers in both Urban and Rural areas, however both Urban Premium and Urban Standard are more popular than Business in Rural Areas. Senior and Backpacker overall are the least common amongst customers. The low number for senior is likely due to the demographics within the customer base, with the average age being 41 and a higher population of younger customers compared to those above 60. Backpacker Travel insurance simply seems to be a more niche product popular almost exclusively amongst younger customers. This information can be used in decision making, targeting customers of specific areas with the right product. For example, a heavy focus can be placed on targeting Urban customers with Business, Premium and Standard travel insurance. The figure also shows exposure levels by area, with Rural customers being marginally more exposed to accidents. This may be due to higher speed zones and poorer quality roads in these areas. There is also likely to be fewer speed cameras. This information may be used to determine pricing strategies by location; however, further analysis is recommended as the difference is negligible. Furthermore, a recommendation would be to categorise locations further. Splitting Rural/Urban by population, traffic levels, settlement type or district, could lead to deeper analysis in determining exposure levels as well as the popularity of travel insurance types by area, leading to better marketing and pricing strategies. Overall, the analysis in this report produced some interesting insights. Age for example, is a characteristic that clearly effects consumer product and marketing communications preferences. However, this analysis is preliminary and merely descriptive. Insure ABC should take on- board some of the recommendations provided, such as developing on the location and occupation categories, as the insights that could be gained from these columns based on the current data is limited. Naturally, as more data is collected, the level and accuracy of insights will increase, however analysts within the function should seek to further visualise this data so it is better communicated to decision-makers within the business. Additionally, further analysis could be done to provide inferential data that can produce predictive models to determine a customer’s preference based on their characteristics. Following these steps should ultimately lead to better data-driven decision making within the business.

# ***Appendix 1: SQL Code***

**CREATING ABT**

**Query 1:** Left Join Health\_Policies to Customer Table

SELECT \* INTO T1

FROM Customer LEFT JOIN Health\_Policies ON Customer.healthID = Health\_Policies.healthID;

**Query 2:** Left Join Motor\_Policies

SELECT \* INTO T2

FROM T1 LEFT JOIN Motor\_Policies ON T1.motorID = Motor\_Policies.motorID;

**Query 3 :** Left Join Travel Policies into Insurance\_ABT

SELECT \* INTO Insurance\_ABT

FROM T2 LEFT JOIN travel\_Policies ON T2.travelID = travel\_Policies.travelID;

**DESCRIPTIVE ANALYSIS/SUMMARIES**

**Average Age:**

SELECT avg(age) AS AverageAge

FROM Insurance\_ABT

WHERE Age BETWEEN 16 AND 110

**Gender Count:**

SELECT gender, COUNT(Gender) AS Frequency

FROM Insurance\_ABT

GROUP BY gender

**Preferred Channel Count:**

SELECT prefchannel, COUNT(prefchannel) AS Frequency

FROM Insurance\_ABT

Group By prefchannel

**Location by Traveltype Count:**

SELECT Location, TravelType, count(\*) AS Frequency

FROM Insurance\_ABT

GROUP BY location, traveltype

ORDER BY count (\*) DESC

**Age Ranges by Preferred Channel:**

(18 to 30)

SELECT DISTINCT Age , prefchannel, count(\*) AS Frequency

FROM Insurance\_ABT

WHERE AGE BETWEEN 18 AND 30

GROUP BY age, prefchannel

HAVING COUNT (\*) >10

(30 to 50)

SELECT DISTINCT Age , prefchannel, count(\*) AS Frequency

FROM Insurance\_ABT

WHERE AGE BETWEEN 30 AND 50

GROUP BY age, prefchannel

HAVING COUNT (\*) >10

(Senior)

SELECT DISTINCT Age , prefchannel, count(\*) AS Frequency

FROM Insurance\_ABT

WHERE AGE BETWEEN 60 AND 80

GROUP BY age, prefchannel

HAVING COUNT (\*) >2

**Average Exposure by Location:**

SELECT avg (exposure), location

FROM Insurance\_ABT

GROUP BY location

**Average Age by Insurance Types**

SELECT avg(Age) , HealthType, TravelType, MotorType, count(\*) AS Frequency

FROM Insurance\_ABT

WHERE Age BETWEEN 40 AND 41

GROUP BY age, HealthType, TravelType, MotorType

**Distinct Occupations**

SELECT Count(OCCUPATION) AS DistinctOccupations

FROM (SELECT DISTINCT occupation FROM Insurance\_ABT) AS [%$##@\_Alias]

# Appendix 2: R Code

#check working directory

getwd()

#import datasets

#install dplyr

install.packages("dplyr")

library(dplyr)

#rename variables to match

customer <- customer %>%

rename(motorID=MotorID,

healthID=HealthID,

travelID=TravelID)

#combine datasets

Insurance <- left\_join (customer, health\_policies, by="healthID")

Insurance <- left\_join (Insurance, motor\_policies, by="motorID")

Insurance <- left\_join (Insurance, travel\_policies, by="travelID")

#summarise new ABT

summary(Insurance)

#quick exploration to identify quality issues

boxplot(Insurance$Age)

hist(Insurance$DependentsKids)

#data wrangling

##converting relevant variables to factors

Insurance$CreditCardType <- as.factor(Insurance$CreditCardType)

Insurance$Occupation <- as.factor(Insurance$Occupation)

Insurance$Gender <- as.factor(Insurance$Gender)

Insurance$Location <- as.factor(Insurance$Location)

Insurance$PrefChannel <- as.factor(Insurance$PrefChannel)

Insurance$HealthType <- as.factor(Insurance$HealthType)

Insurance$MotorType <- as.factor(Insurance$MotorType)

Insurance$veh\_body <- as.factor(Insurance$veh\_body)

Insurance$TravelType <- as.factor(Insurance$TravelType)

##deal with outliers

Insurance$Age[Insurance$Age>110]<- NA

Insurance$Age[Insurance$Age<16] <-NA

Insurance$DependentsKids[Insurance$DependentsKids>10]<- NA

Insurance$veh\_value[Insurance$veh\_value>10]<- NA

#summary of clean data

summary(Insurance)

#structure overview

str(Insurance)

#basic descriptives using hmisc

install.packages("Hmisc")

library(Hmisc)

describe(Insurance$Age)

describe(Insurance$PrefChannel)

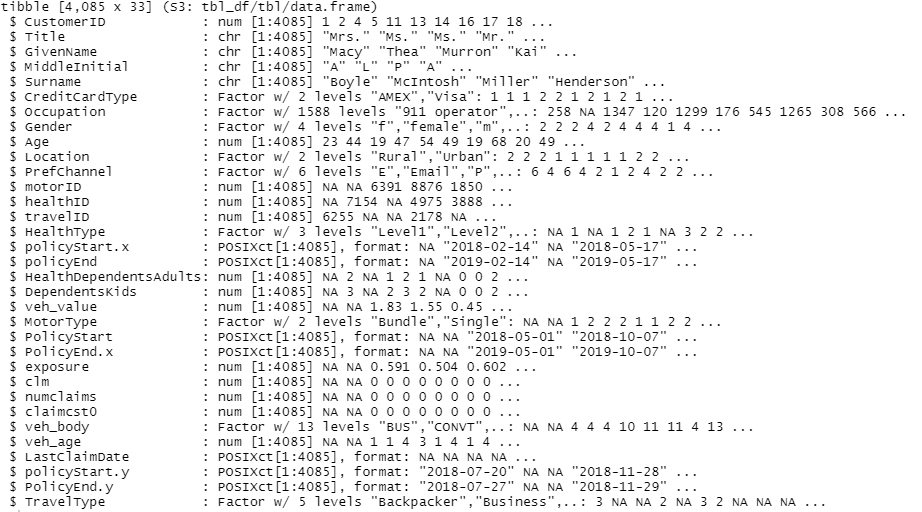
describe(Insurance$Gender)

describe(Insurance$Location)

describe(Insurance$TravelType

# Appendix 3: Screenshots

**Screenshot 1:** Insurance ABT Structure Output (RStudio)



**Screenshots 2**: Hmisc Descriptive Ouputs in RStudio

