Queen's University Management School.

MGT7178: Data Management: Assignment 1

Semester 1, 2022/2023

Student: BHARAT GURUNATHAN HARI DOSS
Student ID: 40387258

WORD COUNT:2158

TABLE OF CONTENTS

INTRODUCTION -

This study explores consumer features across numerous insurance products and preferred marketing channels in an insurance company, and how these insights may assist future sales, data management, and modelling. Identifying Potential Markets, Gaining a 360-degree View of Customers, and Understanding Needs Leads to Increased Profitability. Increasing operating costs and regulatory constraints harm insurance firms, as most customers want fast and individualized service.

This article addresses life insurance techniques for different consumer segments. The study's purpose is to analyse client traits and factors that lead to effective tactics.

The company sells travel, health, and auto insurance. The company has given all the data for the three above-mentioned policies and their clients' distribution channels

This paper will show how to utilise SQL and R to improve data quality and analysis, as the company's analysts are educated in SQL but want to move to R soon.

1. Database Documentation

1.1 Database development in Microsoft Access

This database contains details of customer insurance policies. There are four tables in the database. Each table details are summarised below shows the fields, data quality concerns within the table. Tables can be joined using the primary and foreign keys.

Customer Data – This document contains data on 4085 observations, including Customer personal information and policies ID numbers. I discovered and fixed data quality issues since the data was raw in Customer Title, Card information, Customer Age, Communication Channel, and Gender, among other areas. Primary key:CustomerID

Foreign keys: MotorID, HealthID, and TravelD.

Motor Policy: The overall records in this are 3361 that includes variables like Motor IDs, a vehicle worth of $10,000, Claim values are either Yes or No, and v body (vehicle body) is indicated as the type of vehicle body, such as BUS, COUPE HATCHBACK, HARD TOP, SEDAN, and so on. 1 2 3 4 is the code for v age (age of the vehicle) which is ranked from 1 being the youngest to 4 being the oldest. The total number of claims filed, as well as the amount of each claim and the most current claim date.

Primary key: MotorID.

Health Policy: This file contains information on 2543 records on HealthID, Policy starts and ends dates, HealthType (which specifies the type of health insurance held by the customer,

Level1, Level2, or Level3), and HealthType (which specifies the type of health insurance held by the customer, Level1, Level2, or Level3) fields. Adults and children are counted as dependents.
Primary key: HealthID.

Travel Policy: The total data is 2108. Attribute of this file includes TravelID, Travel type coded as Backpacker, Business, Premium, Senior Standard. Policy starts and end date. Primary key : TravelID.

1.2 Analytics Base Table (ABT):

I'll join all the tables into a single table and naming it as ABT which contains all the customer's information as well as the data of all the Insurance products that they purchased. ABT will also be used to gain a better understanding of consumer insurance purchasing behaviour to make dta decision for formulating.
The processes for creating the database in MS Access are illustrated below. The primary keys used in each table are listed below and how to join the tables. The LEFT Join method is used here.

1.3 Procedure to import all the files

Open MS Access and import all the data provided by the organisation.
Select External Data -> New Data Source -> Choose from File -> Click on Excel. (As the data provided is in xlsx format)
Specify the source of the files and import all four data files (Customer, Health polices, Motor polices, and Travel Polices) into MS Access.

1.4 Addressing Data Quality Issues and Joining Table

Before merging the tables, the aim is to fix the data quality problems. Below are the steps which I used for data quality analysis in Customer Table.
Click on Create tab and open Query Design to create Query.
Evaluate which fields in the Customer data require quality assurance. Now we are going to create a new table Clean_Customer_Table where all the data quality issues in Customer table will be resolved in both SQL and R.

Column Name	Data quality issues
Age Column	Three records in the Age Column include incorrect values such as -44, 180, and 210. This will be corrected and migrated to a new table called Clean Customer Table.
Title Column	The Title column has six fields in which the title 'Mr' is used instead of 'Mr.'
Gender Column	It has 23 fields in which F is used instead of Female and M is used instead of Male. The same has been corrected and updated.

CardType Column	column contains 721 instances in which the Card type is specified as 0. I altered it to "No Card."
ComChannel Column	There are few records in the ComChannel column where S is referenced instead of SMS, P is used instead of phone, and E is used instead of email.

Verify that the common value in both tables is written in the same format (Number or Text). Changing the fields MotorID, HealthID, and HealthID in Number format for doing analysis.

1.4.1 Joining the Clean_Customer_Table and Health_polices data and generating a new table called "Customer_Health Table"

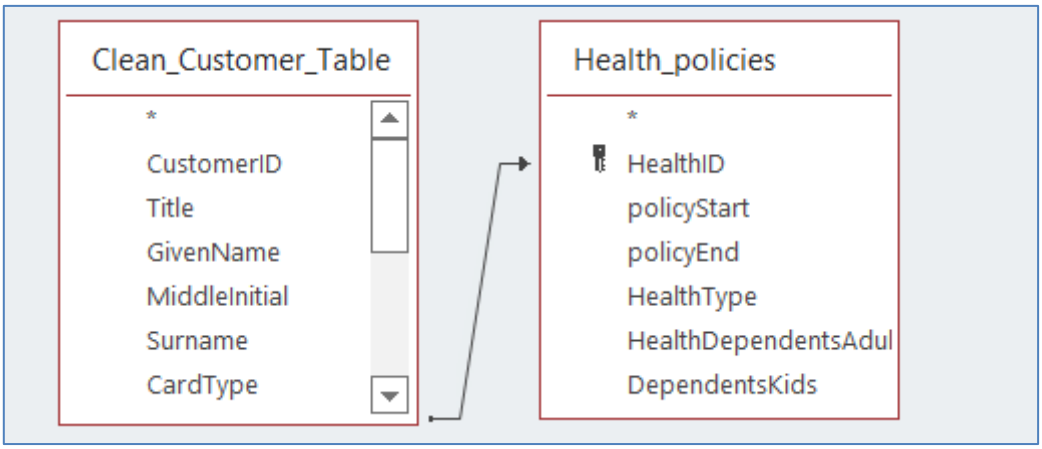


Figure 1: Design View: Clean_Customer_Table + Health_policies Table -> Customer_Health Table

1.4.2 Resolving Data Quality issues in Motor_policies and Joining the Customer_Health and Motor_policies

The occurrence of claim column clm will be modified as 0 = "No" & 1 = "Yes". Then joining the Customer_Health table and Health_polices table which is names as Customer_Health_Motor

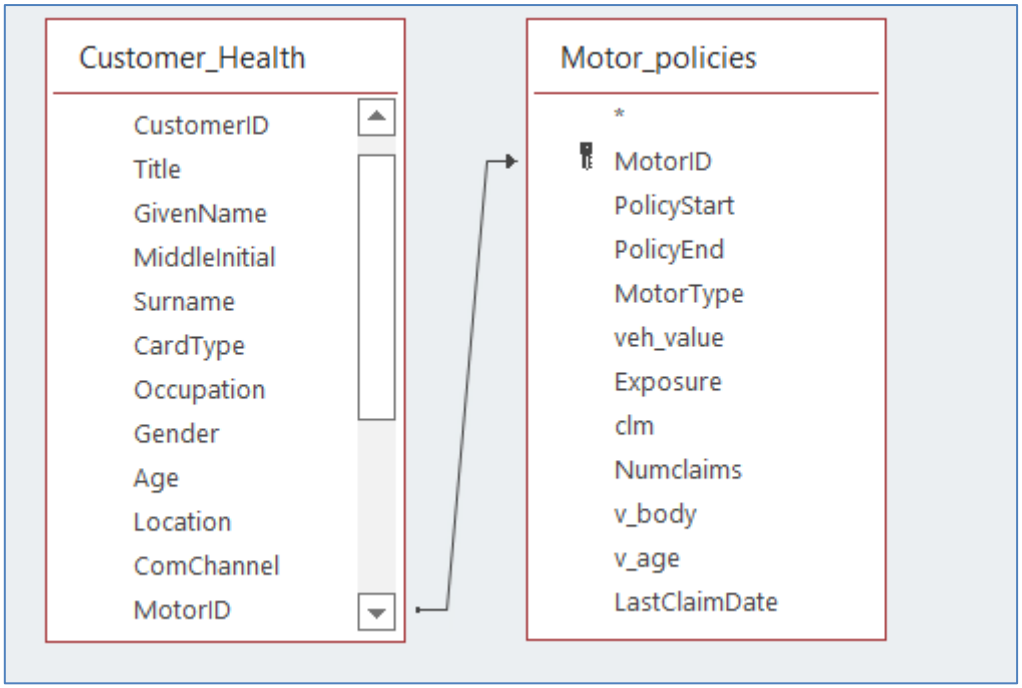


Figure 2: Design View: Customer_Health and Health_polices -> Customer_Health_Motor

1.4.3 Joining the Customer_Health_Motor and Travel_policies INTO *Customer_Health_Motor_Travel*

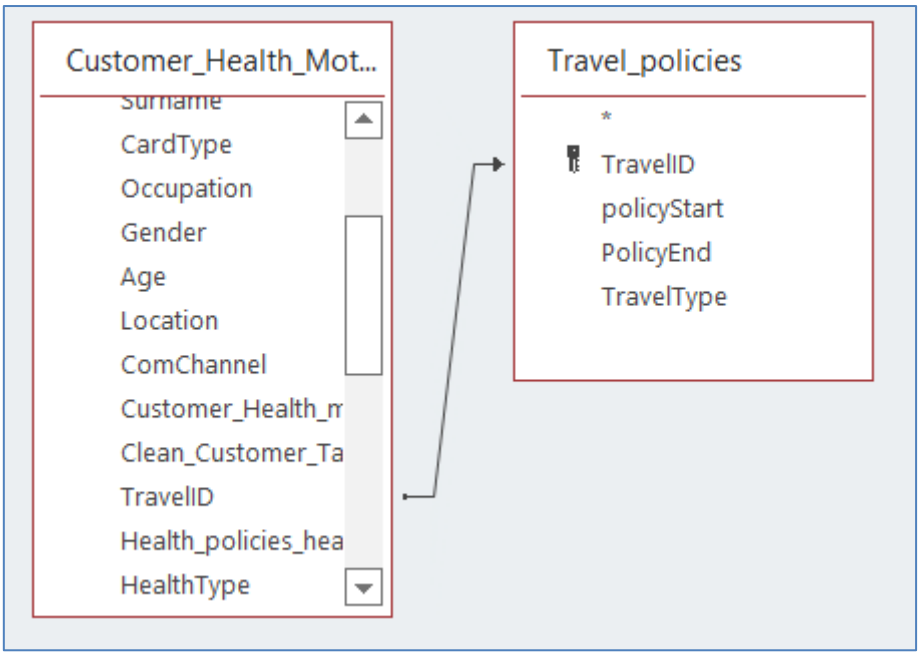


Figure 3: Design View : Customer_Health_Motor and Travel_policies -> Customer_Health_Motor_Travel

2 SQL Limitations and overcoming with R programming:

When dealing with database creation and resolving data quality issues., SQL has been discovered as a particularly useful technique in joining tables and getting insights from data by running queries. Loading data, transforming, manipulating, aggregating, graphing, and sharing your analyses are all very simple with R, and the workflow is much more fluid than in SQL.
R can perform complex data transformations with simple syntax using the dplyr, Tidyverse group of packages, without the limitations or difficulties that SQL provides. R's capabilities for advanced analytics such as predictions, modelling is industry-leading and used by Business Analysts all over the world. And R's charting and plotting capabilities (through packages like as ggplot2) are powerful and customizable.

I utilised R to join the tables, addressing the data quality issues, visualization, and for further descriptive data analysis in this document.

3 Data Quality Report -

The data quality issues are explained below. However, other data quality issues are addressed in the appendix part written in R code. The scenario below shows how I utilised R to improve data quality analysis. The combined data of all files is named ABT.

Age column in Customer Data: It has been determined that three records have incorrect values such as -44, 180, and 210. (Considered as Outliers). This is determined by running plots in R for Age Data and obtaining a summary of the Age Data.
Below is a summary of age data (With Outliers)

```
> summary(ABT$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-44.00   22.00   46.00   41.38   50.00  210.00
```

STEP1: To identify the outliers in Age data use the Box Plot function and ggplot. Below the images, we can see the Outliers -44,180,220.

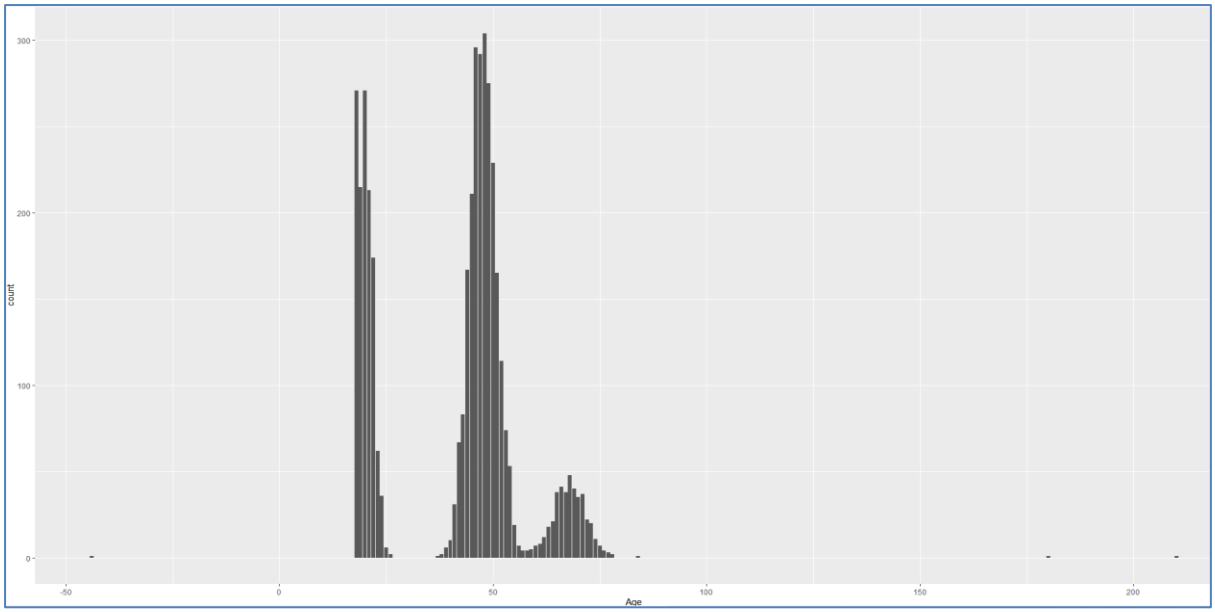


Figure 4: ggplot of Age data

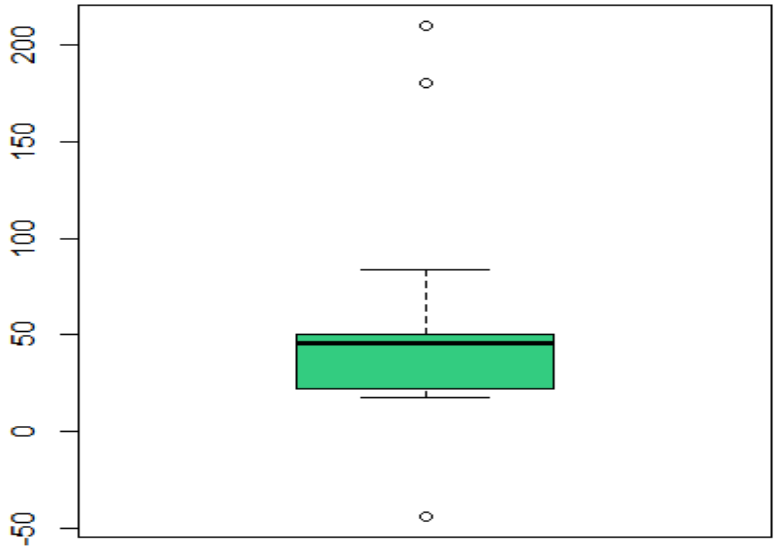


Figure 5: Boxplot of Age data

STEP2: Using the preceding procedure, we may identify numbers that are greater than 100 and less than one as misfits.

STEP3: Age Data where age is more than 100 is assigned as NA and Less than 1 is assigned as NA. When all of the outliers have been designated as NA. We can examine the data summary.

```
ABT$Age[ABT$Age>100]<-NA
ABT$Age[ABT$Age<1]<-NA
summary(ABT$Age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
18.00   22.00   46.00   41.33   50.00   84.00       3
```

Figure 6: Data Summary of Age after rectifying data issues

Title Column: The Title column has six fields in which the title 'Mr' is used instead of 'Mr.' which is resolved using commands below.

```
> summary(ABT$Title)
   Length     Class      Mode
     4085 character character
> table(ABT$Title)

 Dr.    Mr  Mr. Mrs.   Ms.
 127     6 1938  948 1066
> ABT$Title[ABT$Title == 'Mr'] <- "Mr."
> # AFTER RECTIFYING
> table(ABT$Title)

 Dr.   Mr. Mrs.   Ms.
 127 1944  948 1066
```

Figure 7: Rectified data issues of Title Column

Gender Column: It has 23 fields in which F is used instead of Female and M is used instead of Male. The same has been corrected and updated.

```
> summary(ABT$Gender)
   Length     Class      Mode
     4085 character character
> table(ABT$Gender)

     f female       m   male
    14   2061       9   2001
> ABT$Gender[ABT$Gender == "m"] <- 'male'
> ABT$Gender[ABT$Gender == "f"] <- 'female'
> # AFTER RECTIFYING
> table(ABT$Gender)

female   male
  2075   2010
```

Figure 8: Rectified data issues of Gender Column

CardType Column: column contains 721 instances in which the Card type is specified as 0. I altered it to "No Card."

```
> summary(ABT$CardType)
   Length     Class      Mode
     4085 character character
> table(ABT$CardType)

         0 Mastercard       Visa
       721       1725       1639
> ABT$CardType[ABT$CardType == "0"] <- "No Card"
> # AFTER RECTIFYING
> table(ABT$CardType)

Mastercard     No Card       Visa
      1725         721       1639
```

Figure 9: Rectified data issues of CardType Column

ComChannel Column: There are few records in the ComChannel column where S is referenced instead of SMS, P is used instead of phone, and E is used instead of email.

```
> table(ABT$ComChannel)

    E Email      P Phone      S    SMS
    5  1759      5  1559      2    755
> ABT$ComChannel[ABT$ComChannel == "E"] <- 'Email'
> ABT$ComChannel[ABT$ComChannel == "P"] <- 'Phone'
> ABT$ComChannel[ABT$ComChannel == "S"] <- 'SMS'
> # AFTER RECTIFYING
> table(ABT$ComChannel)

Email Phone   SMS
 1764  1564   757
```

Figure 10: Rectified data issues of ComChannel Column

clm (Claim Occurrence) column in Motor_Policies Data:
If the data is in the form of 0 and 1, I will clean it up by converting it to the correct format, where 0 represents no claim('NO') taken and 1 represents a claim taken('YES'). Below is the summary and cleaning the data.

```
> summary(ABT$clm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0000  0.0000  0.0000  0.0637  0.0000  1.0000     728
> table(ABT$clm)

   0    1
3143  214
> ABT$clm[ABT$clm == "0"] <- "No"
> ABT$clm[ABT$clm == "1"] <- "Yes"
> # AFTER RECTIFYING
> table(ABT$clm)

  No  Yes
3143  214
```

Figure 11: Rectified data issues of clm Column

Change the data set into factor and generating the summary: 3143 cases where the claim is not made and 214 case where a claim is taken by the customer.

```
> ABT$clm <- as.factor(ABT$clm)
> summary(ABT$clm)
  No  Yes NA's
3143  214  728
```

Figure 12 :NA value summary

Health Policy: Health Type

Change the data set into factor and generating summary of the variable where Data quality issues are identified and resolved.

```
> summary(ABT$HealthType)
   Length     Class      Mode
     4085 character character
> table(ABT$HealthType)

Level1 Level2 Level3
   659   1253    626
> ABT$HealthType <- as.factor(ABT$HealthType)
> summary(ABT$HealthType)
Level1 Level2 Level3   NA's
   659   1253    626   1547
```

Figure 13 : NA value summary

Travel Policy: TravelType

Summary of TravelType column where Data quality issues like NA values are identified and resolved.

```
> table(ABT$TravelType)

Backpacker   Business    Premium     Senior   Standard
       336        669        442        183        475
> ABT$TravelType <- as.factor(ABT$TravelType)
> summary(ABT$TravelType)
Backpacker   Business    Premium     Senior   Standard       NA's
       336        669        442        183        475       1980
```

Figure 14 :summary of TravelType

4. Insights Report *:*

Insight 1: Evaluating the Communication Modes of Travel Insurance Customers Over the Age of 70

Using the data, I collected information such as card details, travel policy type, and preferred method of communication for all older persons over the age of 70. According to the search, 70 consumers satisfied the specified criteria. Only 13 customers do not have a credit card, while the remaining 57 do. All these customers had previously acquired travel insurance, with 42 opting for the Senior Policy. 53 clients prefer to be reached by phone, while 16 prefer to be reached by email. Only 1 client prefers to be reached by SMS

With the information provided above, firms can market Senior insurance products and explain their benefits via their preferred communication channel.

Here is the output table of the SQL query.

GivenName	Age	CardType	TravelType	ComChannel
Alice	73	Mastercard	Backpacker	Email
Autumn	84	No Card	Business	Phone
Milly	71	Visa	Business	Phone
Lachlan	77	Mastercard	Business	Email
Jack	72	No Card	Business	Email
Bobby	71	Mastercard	Business	Phone
Ronnie	72	Visa	Business	Phone
Leon	71	Visa	Business	Phone
Bradley	71	Visa	Business	Phone
Harvey	71	Mastercard	Business	Phone
Anthony	74	Mastercard	Premium	Phone
Teegan	72	Mastercard	Premium	Phone
Ethan	71	Mastercard	Senior	Phone
Nicholas	75	Mastercard	Senior	Phone
Millie	73	No Card	Senior	Phone

Insight 2: Preferred Communication Channel based on Location for all Customers

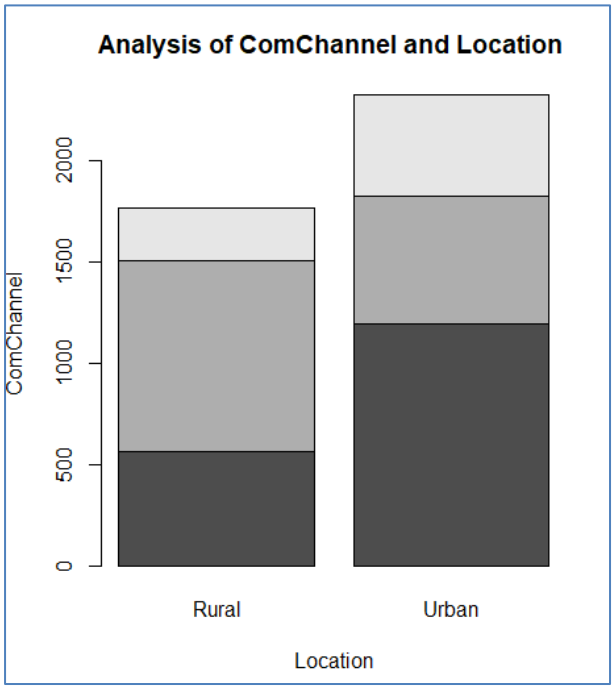
The table below summarises the preferred method of contact for all 4082 consumers across all goods. The figure shows that email is the most popular strategy among customers which is 1763 records. This is followed by Phone (1564 users) and SMS (755 users).
When the data is further separated by location, it becomes evident that the majority of customers in urban areas choose Email, while the majority of the customers in rural areas prefer Phone.
Given the preceding information, it is suggested that the company implement a marketing plan in which new insurance products and offers are primarily communicated via email to consumers in urban areas and phone calls are made to clients in rural areas in order to attract more customers. Below is the output of SQL queries of data analysis and the Bar plot of the data done in R.

Output of Query a

Channel	Count_Customer
Email	1763
Phone	1564
SMS	755

Output of Query b

Location	Channel	Count_Customer
Rural	Email	567
Rural	Phone	939
Rural	SMS	256
Urban	Email	1196
Urban	Phone	625
Urban	SMS	499



Insight 3: Analysis of Motor Policy with Policy Age and Communication Channel

This analysis provides information on the customer's motor insurance coverage (Single or Bundle). This analysis clearly illustrates that single policy is selected by clients over bundled policies.

When the nature of the policy purchased is compared to the age of the vehicle (1-Youngest, 2, 3, 4-Oldest), it is discovered that the customer purchases a maximum of 1270+ single policy types when the car is old (at stage 3 and 4).

Moreover, when the policy type is compared to the communication channel, it is determined that the preferred communication channels for clients seeking Single type insurance are Phone and Email. Businesses should focus their marketing awareness approach on a certain sort of policy, with phone and email as their key communication channels. Likewise, if necessary, we can dive deep into additional variables such as gender and location.

Output of Query a

MotorType	Count_Motc
Bundle	1069
Single	2285

Output of Query b

MotorType	Vehcile_Age	Count_Motor_policy
Bundle	1	201
Bundle	2	250
Bundle	3	324
Bundle	4	294
Single	1	435
Single	2	573
Single	3	620
Single	4	657

Output of Query c

ComChanne	MotorType	Count_Customer
Email	Bundle	522
Phone	Bundle	184
SMS	Bundle	363
Email	Single	976
Phone	Single	1106
SMS	Single	203

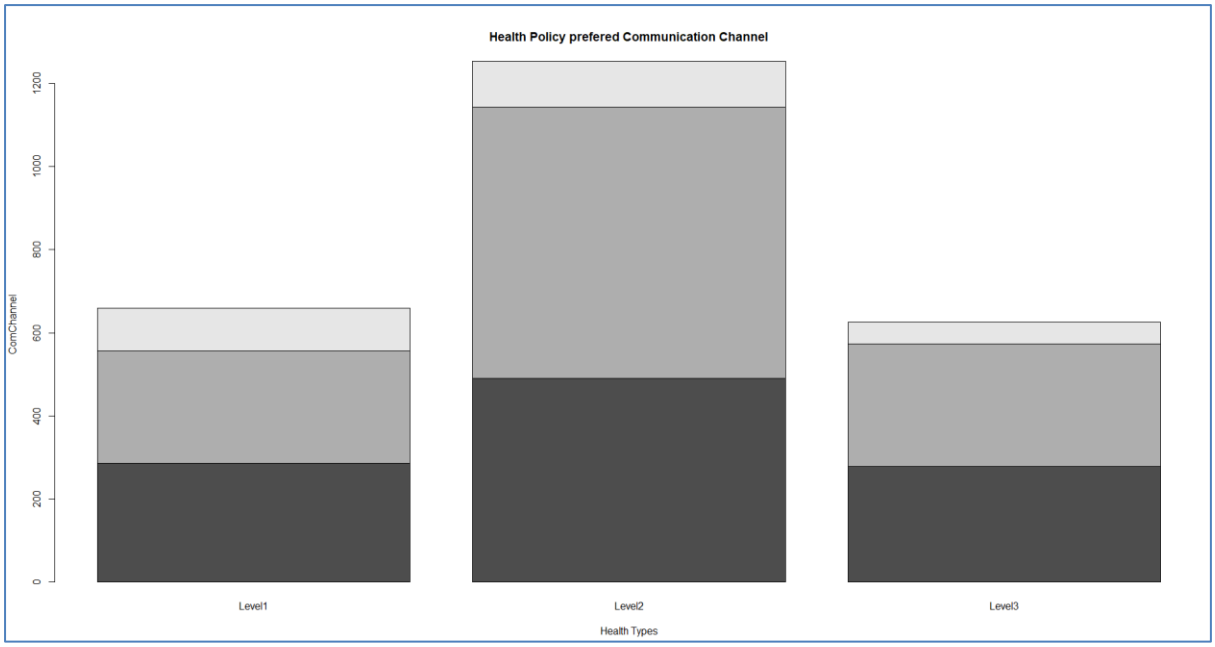
Insight 4: Analysis of Health Policy with HeathType of preferred Communication Channel of Customers over age of 50

By analyzing the data, I gathered information such as health type, customer age, and favourite mode of communication for all those above the age of 50. According to the results of the search, 696 consumers met the stated criteria. All of these clients had previously purchased health insurance, with 123 choosing Level 1, 314 choosing Level 2, and 256 choosing Level 3. While 505 clients prefer to be contacted by phone, 165 prefer to be contacted via email. Only 26 clients like to be contacted by SMS.

Firms can promote Senior insurance products with Health Type levels and explain their benefits using the information supplied above using their choice communication channel.

The SQL query's output table is shown below and the bar plot of the data analysed for better understanding

GivenName	Age	HealthType	ComChannel
Billy	51	Level1	Email
George	52	Level3	Email
Charles	54	Level1	Email
William	51	Level2	Email
Samantha	52	Level1	Email
Christian	52	Level3	Email
Alex	53	Level2	Email
Sophie	51	Level3	Email
Alexander	52	Level2	Email
Jack	72	Level3	Email
Sian	67	Level3	Email
Scott	51	Level3	Email
Jamie	74	Level3	Email
Sienna	71	Level3	Email



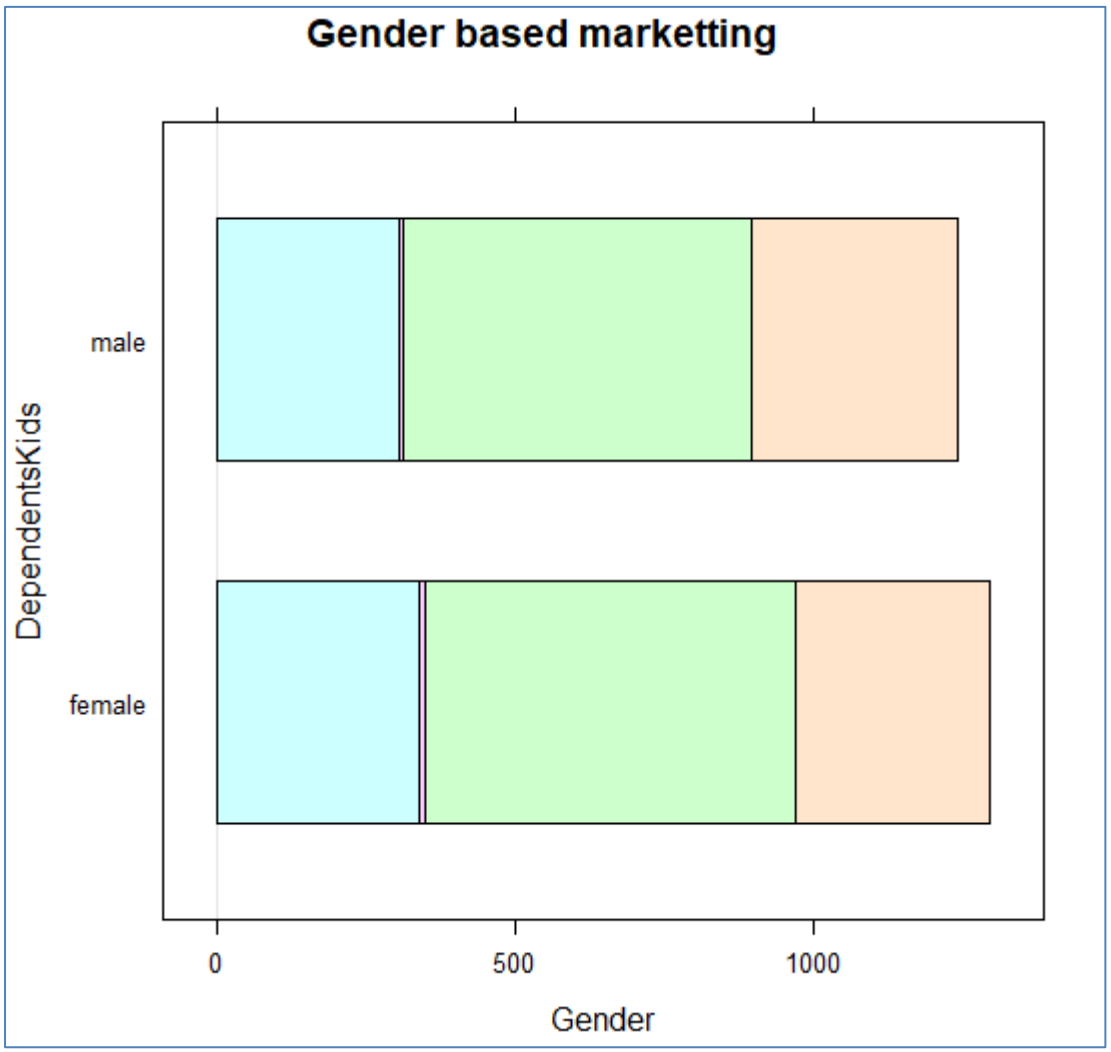
Insight 5: Evaluating the Health Insurance based on Dependents Kids of both Genders of Customer:

Using R programming, it was discovered that females who chose the health policy had significantly more dependent children than males who chose the health policy. Where there is 0 Dependents Kids, the female count is 340, For 1 Dependent Kids, the female count is 620. For 2 Dependents Kids, the female count is 620. For 3 Dependents Kids, the female count is

326. All the values of the female gender is more than male. So, if we provide marketing strategy for based on the analysis, it will improve the firm.

```
> table(ABTv1$DependentsKids,ABTv1$Gender)

   female male
0     340  306
1      10    8
2     620  581
3     326  346
```



Appendix 1: SQL Code

SQL code is used to identify and resolve data quality concerns.

Customer Table- DQ analysis:

- **Age Column** where 3 Records are having wrong values like -44, 180,210. The same will be rectified and transferred to the new table name Clean_Customer_Table

Query:

```
SELECT Customer.*
INTO Clean_Customer_Table
FROM Customer
WHERE Customer.Age >0 AND Age<100;
```

- **The gender column** has 23 fields where F is mentioned instead of Female and M is mentioned instead of Male. The same needs to update

Query:

```
UPDATE Clean_Customer_Table
SET Gender = "female"
WHERE Gender="f";
UPDATE Clean_Customer_Table
SET Gender = "male"
WHERE.Gender="m";
```

- **CardType column** has 721 records where Card type is mentioned as 0. Which we will change to "No Card".

Query:

```
UPDATE Clean_Customer_Table
SET CardType = "No Card"
WHERE CardType="0";
```

- **ComChannel column** has a few records where S is mentioned Instead of SMS, P is mentioned instead of Phone, and E is mentioned instead of Email.

Query:

```
UPDATE Clean_Customer_Table
SET ComChannel = "Email"
WHERE ComChannel="E";
UPDATE Customer SET ComChannel = "Phone"
WHERE ComChannel="P";
UPDATE Customer SET ComChannel = "SMS"
WHERE ComChannel="S";
```

- **Motor_policies -DQ Analysis:** Column clm - occurrence of claim to be updated as 0 = no & 1 = yes

Query:

```
UPDATE Motor_policies SET clm = "No"
WHERE clm ="0" ;
UPDATE Motor_policies SET clm = "Yes"
WHERE clm ="1" ;
```

- **Joining Table:**
  - **Clean_Customer_Table and Health_polices.**

Ensure the Common value in both the tables should be in the same format (Number or Text). Changing the field MotorID HealthID and HealthID in Number format. Joining the

Clean_Customer_Table data and Health_polices Data and creating a new Table named ***Customer_Health***

Query:

SELECT Clean_Customer_Table.,*
Health_policies.HealthID, Health_policies.HealthType,
Health_policies.policyStart AS HealthStartDate,
Health_policies.policyEnd AS HealthEndDate,
Health_policies.HealthDependentsAdults, Health_policies.DependentsKids

INTO ***Customer_Health***

FROM Clean_Customer_Table ***LEFT JOIN*** *Health_policies ON Clean_Customer_Table.HealthID = Health_policies.HealthID;*

- **Joining Table: Customer_Health and Health_polices -> Customer_Health_Motor**

Query:

SELECT Customer_Health.,*
*Motor_policies.MotorID, Motor_policies.veh_value*10000 AS Vehicle_Value,*
Motor_policies.MotorType, Motor_policies.PolicyStart AS MotorPolicyStart,
Motor_policies.PolicyEnd AS MotorPolicyEnd, Motor_policies.clm AS ClaimOccurrence,
Motor_policies.numclaims, Motor_policies.claimcst0 AS MotorClaimCost,
Motor_policies.v_body, Motor_policies.v_age

INTO Customer_Health_Motor

FROM Customer_Health ***LEFT JOIN*** *Motor_policies ON Customer_Health.MotorID = Motor_policies.MotorID;*

- **Joining the Customer_Health_Motor and Travel_policies**

Query**:**

SELECT Customer_Health_Motor ., Travel_policies.**
INTO ***Customer_Health_Motor_Travel*** *FROM Customer_Health_Motor* ***LEFT JOIN*** *Travel_policies ON Customer_Health_Motor .TravelID = Travel_policies.TravelID;*

SQL Query of Insight 1: Evaluating the Communication Modes of Travel Insurance Customers Over the Age of 70

Query:

SELECT GivenName, Age, CardType,TravelType,ComChannel
FROM Customer_Health_Motor_Travel
WHERE (((Age)>70) AND ((TravelType)<>''))
ORDER BY ComChannel;

SQL Query of Insight 2 : Preferred Communication Channel based on Location for all Customers

Query:

a . SELECT ComChannel AS Channel, Count(CustomerID) AS Count_Customer
FROM Customer_Health_Motor_Travel
GROUP BY ComChannel;

b. SELECT Location, ComChannel AS Channel, Count(CustomerID) AS Count_Customer
FROM Customer_Health_Motor_Travel
GROUP BY Location, ComChannel;

SQL Query of Insight 3: Analysis of Motor Policy with Policy Age and Communication Channel

Query:

 # Type motor policy purchased by customers

```
SELECT MotorType, Count(Motor_policies_motorID) AS Count_Motor_policies
FROM Customer_Health_Motor_Travel
WHERE MotorType <> ''
GROUP BY MotorType;
```

#Type of Motor Policy purchased relation with the age of the Vehicle

```
SELECT MotorType, v_age, Count(Motor_policies_motorID) AS Count_Motor_policies
FROM Customer_Health_Motor_Travel
WHERE MotorType <> ''
GROUP BY MotorType, v_age;
```

#CommunicationChannel suitable for the type of motor policy (Single Or Bundle)

```
SELECT ComChannel, MotorType,Count(CustomerID) AS Count_Customer
FROM Customer_Health_Motor_Travel
WHERE MotorType <> ''
GROUP BY MotorType, ComChannel;
```

SQL Query of Insight 4: Analysis of Health Policy of preferred Communication Channel of Customers over age of 50

Query:

```
SELECT GivenName, Age, HealthType ,ComChannel
FROM Customer_Health_Motor_Travel
WHERE (((Age)>50) AND ((HealthType)<>''))
ORDER BY ComChannel;
```

R code for Insight 5:

Code:

```
summary(ABT$DependentsKids)
#for resolving data quality
ABT%>%
mutate(DependentsKids=replace(DependentsKids, DependentsKids<0, NA),
DependentsKids=replace(DependentsKids, DependentsKids>3,NA))->ABTv1

table(ABTv1$DependentsKids,ABTv1$Gender)
#barchart for better visualization
barchart(table(ABTv1$Gender,ABTv1$DependentsKids),main="Gender  based  marketting",
xlab = "Gender",ylab = "DependentsKids")
```

Appendix 2: R Code

```
#Seting up the working directory
setwd("C:/Users/Bharatgh/Desktop/Data Management/Assignment 1")
getwd()

#To see the files inside the directory
dir()

# Install Packages and Load libraries
install.packages("dplyr ")
install.packages("tidyverse ")
install.packages("readxl ")
install.packages("Hmisc ")
library("dplyr")
library(readxl)
library(tidyverse)
install.packages("Hmisc")


#Import Data into R
Customer <- read_xlsx("Data 1_Customer.xlsx")
MotorPolicies <- read_xlsx("Data 2_Motor Policies.xlsx")
HealthPolicies <- read_xlsx("Data 3_Health Policies.xlsx")
TravelPolicies <- read_xlsx("Data 4_Travel Policies.xlsx")

#View Files
View(Customer )
View(MotorPolicies)
View(HealthPolicies)
View(TravelPolicies)

#Joining Data - Customer  + HealthPolicies -> CustHealth
CustHealth <-dplyr::left_join(Cust,Health,by = c("HealthID" = "HealthID"))

#Joining Data - CustHealth + MotorPolicies -> CustHealthMotor
```

```
CustHealthMotor <- dplyr::left_join(CustHealth,Motor,by = c("MotorID"="MotorID"))

#Joining Data - CustHealthMotor + TravelPolicies -> ABT
ABT<- dplyr::left_join(CustHealthMotor,Travel,by = c("TravelID"="TravelID"))

#Data Quality issues

#Age column
summary(ABT$Age)
boxplot(ABT$Age , breaks=40 , col=rgb(0.2,0.8,0.5,1) , main="" , xlab="Boxplot of Age data") ABT$Age[ABT$Age>100]<-NA #Age value more than 100 is assigned as NA
ABT$Age[ABT$Age<1]<-NA # Age Value less than 1 is assigned as NA
ABT<- subset(ABT, !is.na(ABT$Age)) #Removing NA values
summary(ABT$Age)

#Title Column

ABT$Title[ABT$Title == 'Mr'] <- "Mr."
table(ABT$Title) # After Rectifying checking the again
ggplot(ABTv3_clean) + geom_bar(aes(x=Age)) # plot for visualization



#Gender Column
summary(ABT$Gender)
#Notice that M and F should be converted to Male and Female
ABT$Gender[ABT$Gender == "f"] <- "female"
ABT$Gender[ABT$Gender == "m"] <- "male"

#Convert this to a factor
ABT$Gender <- as.factor(ABT$Gender)
summary(ABT$Gender)
table(ABT$Gender)


#Cardtype Column
table(ABT$CardType)
summary(ABT$CardType)

#Notice that data entered as "0" should be converted to No Card
ABT$CardType[ABT$CardType == "0"] <- "No Card"

#Convert this to a factor
ABT$CardType <- as.factor(ABT$CardType)
summary(ABT$CardType)

#ComChannel Column
summary(ABT$ComChannel)
table(ABT$ComChannel)
```

```
#Notice that data entered as "S" "E" and "P" should be converted to SMS, Email and Phone
ABT$ComChannel[ABT$ComChannel == "S"] <- "SMS"
ABT$ComChannel[ABT$ComChannel == "E"] <- "Email"
ABT$ComChannel[ABT$ComChannel == "P"] <- "Phone"
ABT$ComChannel <- as.factor(ABT$ComChannel)
summary(ABT$ComChannel)
table(ABT$ComChannel)


#HealthType Column
summary(ABT$HealthType)
ABT$HealthType <- as.factor(ABT$HealthType)
summary(ABT$HealthType)
table(ABT$ HealthType)

#Location Column
summary(ABT$Location)
ABT$Location <- as.factor(ABT$Location)
summary(ABT$Location)
table(ABT$ Location)


#clm (Claim Occurance) Column
summary(ABT$clm)
ABT$clm[ABT$clm == "0"] <- "no"
ABT$clm[ABT$clm == "1"] <- "yes"
ABT$clm <- as.factor(ABT$clm)
summary(ABT$clm)
table(ABT$ clm)


#TravelType column
summary(ABT$TravelType)
ABT$TravelType <- as.factor(ABT$TravelType)
summary(ABT$TravelType)
table(ABT$ TravelType)


# statistical information
describe(ABT)
```

Appendix 3:

R-Code

 # Motor Policy with Policy Age and Communication Channel

plot(table(ABTv_age,ABTComChannel,ABT$MotorType),main="Motor Policy with Policy Age and Communication Channel")

Output

#Age column after resolving data quality

summary(ABT$Age)

hist(ABT$Age, xlab="Age")

Output

```
summary(ABT$Age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
18.00   22.00   46.00   41.33   50.00   84.00       3
```

Analysis of Health Policy of preferred Communication Channel of Customers over age of 50

barplot(table(ABT$ComChannel,ABT$HealthType),main="Health Policy prefered Communication Channel", xlab = "Health Types", ylab = "ComChannel")

Output