Assignment 1

Data Management

Pratik Prakash Brahmapurkar

40331504

TOTAL WORD COUNT: 2,111

Table of Contents

1.0 Introduction

An insurance firm maintains its sensitive data in .csv file format, but they intend to roll out an environment in which only R language is used in the future. The customer dataset examines the characteristics of customers who purchased insurance contracts between 2019 and 2020. Customer information is analyzed, and it contains three foreign keys that will help to connect the other 3 tables. The dataset is evaluated by focusing on key attributes.

The dataset is initially read into a variable before being analysed for outliers and missing values. Analytical Base table is implemented by using all the 4 tables (Customer, Health, Motor and Travel). Before running the SQL queries, the data is cleaned. Insights are offered with the help of suitable logical tables generated from the queries.

2.0 Database Documentation (958 words)

2.1 Database development in Microsoft Access

All the 4 text files customer.csv, health_policies.csv, motor_policies.csv, and travel_policies.csv were imported in Microsoft Access. If the file is in .csv format then text format is taken into consideration. While importing the files default Date format was also verified, if the format is MMDDYY then it needs to be changed to DDMMYY. If the format is not changed then the date values above 12 would be seen as blank values, as it would consider 12 as a month. While uploading the text file in MS Access it is made sure that the data types of the foreign keys are similar. In this case, the data type was Short Text for the columns MotorID, HealthID, and TravelID which were then converted to Long Integer in the Customer table.

Finally, when the required data is imported, we are good to join the tables. Customer Table is the master table that consists of all the important details about the customer who has taken at least one policy. Motor_Policies, Health_Policies, and Travel_Policies tables are used to identify motor insurance, health insurance, and travel insurance of customers respectively. The Primary Key is a column that identifies unique rows in the table and the foreign key is a column that helps to connect 2 tables, each table has its primary key. Here CustomerID is a primary key and there are 3 foreign keys which are MotorID, HealthID, and Travel ID which are primary keys respectively in Tables Motor_policies, Health_policies, and travel_policies table. Before the creation of the Analytical Base Table, it is noted that there are no duplicates present in the table by using a distinct count function and comparing the count in main tables.

An analytical Base table (ABT) named 'Insurance' is created using left join connecting all the 4 tables and customer table connected at the start. A left join is used for ABT because all the records from the left table which is the customer table need to be extracted. There are few integer and text fields in the database. Names of a few attributes were also changed and the veh_value attribute was multiplied by 10,000. 4085 rows were pasted into the Insurance table. There are a few attributes that are neglected as they might not provide many insights to the problem. The important attributes which are considered in Analytical Base Table are shown in Fig 1. Below. In this scenario, Age is considered as the target feature because it helps to analyse the marketing strategy, the mode of communication and the premium to be charged by the company.

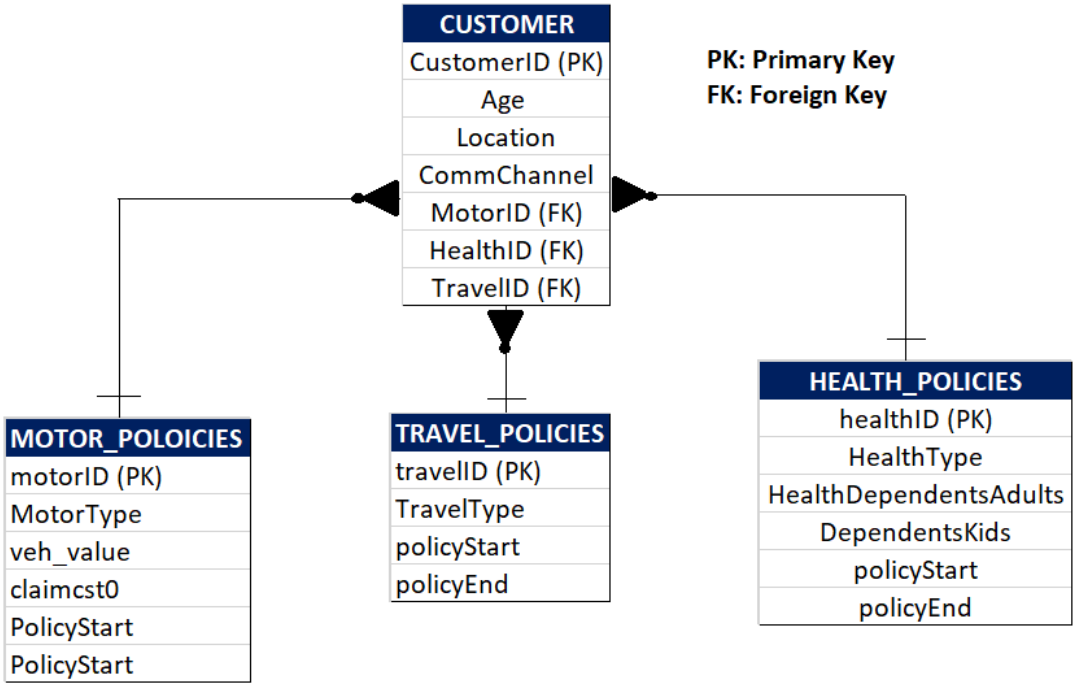

Fig 1. Shows Customer Table connecting other 3 Tables

2.2 Data Quality Report

The Insurance Table (Analytical Base Table) has a few outliers (extreme values), incorrect data, and missing data. The data is also not implemented correctly to analyse. So, data cleaning is performed to proceed. A few of the data quality concerns occurred as a result of the Gender name being incomplete. For 'Male,' it was just 'm' or 'male,' and for 'Female,' it was simply 'f' or 'female.' The same was true for the ComChannel; there were a few data mismatch issues, such as Email being simply 'E,' SMS being 'S,' and Phone being 'P.' Changing of the names in SQL was done by using the UPDATE query and the name change in R was done replacing the values. The code is given in the appendix section. And all these attributes are converted into factors in R.

There were few anomalies in Age, DepentKids and Vehicle Values. As seen in Table 1 below, the minimum age is -44 and the maximum age is 210. The age can never be negative nor in today's era it can exceed above 100. Considering this scenario the age was changed to 41, which is the mean value. In Table 2 below one of the customer has 40 dependent kids. This also is an invalid data. Instead of entering 4 the customer might have entered an extra zero. So, the DepentKids column with 40 is then changed to 4. In the below fig 2, outlier are also identified in the Vehicle price. As observing the 2 dots at extreme left hand side, it is noticed that there are 2 prices above 130000 USD which are the outliers. By observing table 3, we can see that some values are 0 as well. In all circumstances, the median value of 15100 USD may be used to replace these figures. All of the above attributes are modified in SQL by using the UPDATE query. All of the attributes that needed to be updated were changed, and the DepentKids attribute was afterwards transformed into a factor.

Age			
Minimum	**Median**	**Mean**	**Maximum**
-44	46	41.38	210

Table 1. Shows Summary Age of Insurance Table

DependentsKids	0	1	2	3	40
Total Count	648	18	1202	674	1

Table 2. Shows dependent kids with respect to customer

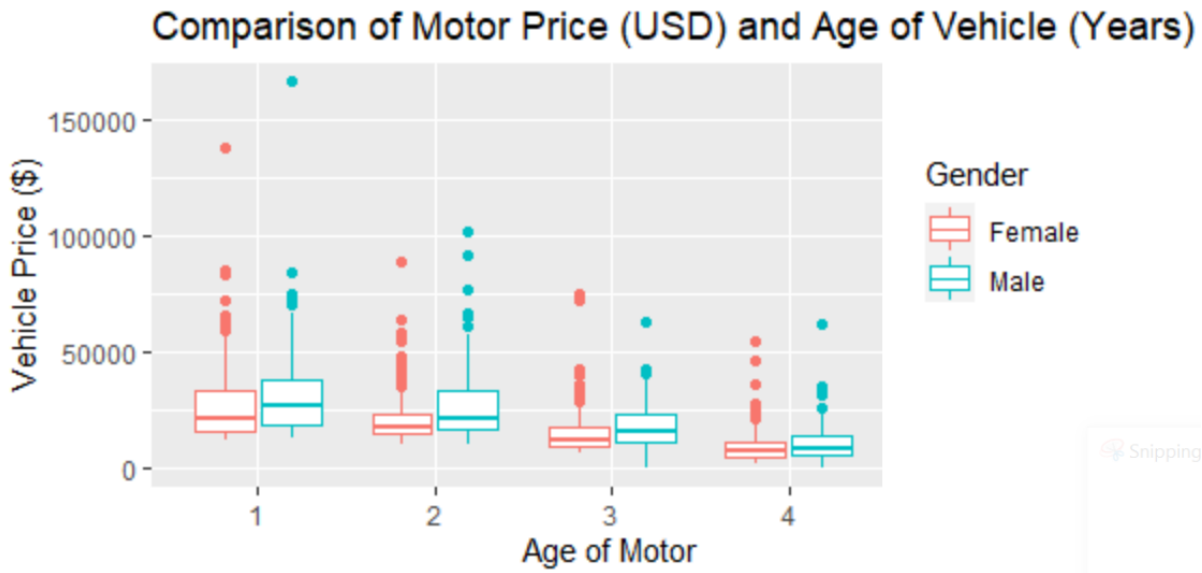

Fig 2. Shows Vehicle Price and Age of Motors where we can see outliers in Vehicle price

Vehicle Price (USD)			
Minimum	Median	Mean	Maximum
0	15100	18116	166900

Table 3. Shows Summary of Vehicle price

Other attributes like HealthDependentsAdults, HealthType, MotorType, v_age and TravelType were converted into factor in R. And if CardType is 0 then it was renamed to 'Other Mode of Transaction'.

While creating Analytical Base table in R, it had to made sure that foreign key names are similar to join the tables. As the column names were not similar customer table had foreign keys start with capital, they were then changed by colnames() function in R. Also there were few similar columns in different tables like policyStart and policyEnd, these column names also had to be changed. All of the tables were connected using three left joins before the creation of ABT. The unwanted columns were eliminated when all four tables were joined. The Mutate function was also used to alter the values of veh_value.

These data quality concerns can be avoided if certain conditions are implemented in the system. For example, if a negative age or an age greater than 99 is input into the system, the system

must display a pop-up error message indicating that the age is incorrect. If the customer gives the correct data, the chances of encountering a data quality issue lowers.

3.0 Insights Report (1025 Words)

Out of 4085 customers only 975 customers have subscribed to Motor insurance policy, Health insurance policy, and Travel insurance policy which is 23.87% of total customers. We can also observe in Table 4 that the majority of the subscribed insurance is Motor Insurance policy and the least preferred is the Travel Insurance Policy. Some tourists never leave home without travel insurance. Most individuals don't bother with travel insurance since it costs extra money and they find it unnecessary (Stevens, 2013)Travel Insurance and health insurance depends on an individual whether that person is willing to buy one. As this data belongs to the USA, motor insurance is compulsory in almost every state of the US except Hawaii, Michigan, Montana, and New Hampshire (Insurance Information Institute, 2021). 728 out of 3357 (total customers with motor insurance) customers who don't have motor insurance probably must be staying in US states where it's not compulsory to have motor insurance. The company can focus and target individual customers who stay in the US where having motor insurance is compulsory. If the customer has taken motor insurance from some other company, they can also attract them back with amazing schemes.

Location	Motor Insurance	Health Insurance	Travel Insurance	Total
Rural	1449	1199	799	1763
Urban	1908	1339	1306	2322

Table 4. Shows Total Count of Insurance with respect to Location

It is also observed in table 4 that maximum customers from urban areas prefer to subscribe the insurance as compared to rural areas. It can also be observed that there is a high ratio of the rural area not preferring travel insurance.

As shown in Table 5, the most preferred communication channel by Motor insurance and Travel Insurance customers is Email and the least one is SMS. Emails can be very detailed and they can also send documents to customers. Wherever the customer is that person can access the policy document or any other document via cell phone/tab/computer. SMS is the least preferred communication channel in all because it has got outdated and it is not that interactive like email. SMS is generally preferred by the young public as shown in the below table 5. The phone is preferred by the customers who are around their 50s, it might be expected that they are not up to date with the technologies like email. Here we can observe that if the company is planning to advertise or sell their product, Email and Phone marketing must be preferred. Phone must be preferred to customers who are in their middle age or above.

Communication Channel	Motor Insurance	Health Insurance	Travel Insurance	Average Age
Email	1499	1054	1007	38
Phone	1290	1219	639	51
SMS	568	265	459	28

Table 5. Shows Table Communication Channel preferred by Customers and Average Age

As shown in Table 6, the older is the vehicle the average vehicle value decreases drastically. One thing is noticed that for a vehicle of Age 1 the average claim amount is 130.52 USD, but for Vehicle age 2 the amount increases to 156.49 USD. And for Vehicle of age 3, the average claim amount is 137.55 which is still greater than the average claim amount of vehicle age 1. This scenario only takes

place if there is less amount of data. It might also be considered that if the vehicle is new there were fewer accidents as compared to the older vehicles, and if even there were accidents there was less expense involved to repair or replace a covered vehicle. All the motor insurance policies were for 1 year.

Vehicle Age	Average Claim Amount (USD)	Average Vehicle Value(USD)	Average Age	Total Motors	Policy Days
1	130.52	27861.83	41	636	365.8
2	156.49	22497.29	42	825	365.9
3	137.55	16113.54	41	945	365.8
4	67.18	9545.22	43	951	365.9

Table 6. Shows Vehicle Age, Claim amount, Vehicle Value and Average Age, motors and duration of policy

As shown in Table 7 it is observed that the Level 2 health insurance type consists of the maximum count of dependent adults. The average age for level 2 health insurance type is 48 years. Mostly Level 3 health insurance type is considered by senior customers who are above 50. And level 1 health insurance type customers are the youngest amongst all having average dependence per customer as 2. The policy in this insurance company for all the customers is for 1 year as well. The average dependent for level 2 health insurance type customers is 3. It might be a possibility that the dependent might be 1 spouse and 2 kids. There are few customers also that didn't even have single dependent insurance. The customer who has a spouse and kids can be targeted. The customer can be advised to subscribe to health policy for their dependents. Which will increase the sales of the insurance company.

Health Type	Count of Dependent Adults	Average Dependant Per customer (Adult + Kids)	Average Age	Average Policy Days
Level 1	659	2.1	40	365.8
Level 2	1253	2.9	48	365.8
Level 3	626	2.4	53	365.8

Table 7. Show Health insurance Type, dependents data and average age of customer

As shown in table 8, it is observed that Travel insurance held by Backpackers are the youngest customers with an average age of 25 who are planning a long vacation where the average days of insurance active are 21. Premium travel insurance, Standard, and Senior Insurance have almost the same average days of insurance active which is 10 to 11 days. Customers from senior travel insurance have an average age of 66, they are the senior-most travelers. Business insurance Traveller has the most count of customer amongst all where the average age is around 39, the average active policy days is 4.

Mostly for business travel insurance subscribers mostly the trips are covered by corporate giants or companies they are working in as it is business or work-related travel. So, the customer who is working in the corporate world and whoever is likely to opt for Business insurance can be considered the best candidate to target. As the chances of customers are very high that they will subscribe to travel insurance for business trips as compared to the other travel type, because anyhow they might be getting sponsorship.

Travel Type	Count of Customer	Average Age	Average Days of Policy Active
Backpacker	336	25	21
Premium	442	32	10
Business	669	39	4
Standard	475	44	11
Senior	183	66	10

Table 8. Shows Travel Type Insurance, Count of Customer, Average Age and Active policy duration

CardType	Health Insurance	Travel Insurance	Motor
MasterCard	1064	899	1422
Visa	1005	851	1356
Other mode	469	355	579

Table 9. Shows Card Types

Referring to Table 9 it can be analyzed that maximum customers prefer to pay by MasterCard card type. The second most preferred card type is Visa.

The insurance company needs to focus more on the duration of the health and motor policies. By observing table 6 and table 7 the policy lasts only for a year. The company needs to implement term insurance policies which can be five, 10, 15, or 20 years policies for health insurance and up to 3 years insurance for motor insurance. So, that the company doesn't need to attract the same customers every year. Communication Channel helps us to understand better insights that which customers can be targeted with a mode of communication for a better marketing strategy. In this company, the communication channel is limited to SMS, email, and phone. Communication should also be possible by the customers on a social media platform as well like Twitter, Facebook, and Instagram.

Appendix 1: SQL Code

Creation of Analytical Base Table:

```
SELECT C.CustomerID, C.CardType, C.Age, C.Location, C.ComChannel, M.motorID,
M.veh_value*10000 AS 'Vehicle Value(USD)', M.v_age AS 'Vehicle Age', M.claimcst0 AS 'Claim
Amount (USD)', H.healthID, H.HealthType, H.HealthDependentsAdults, H.DependentsKids,
T.travelID, T.TravelType, M.PolicyStart AS MotorPolicyStartDate, M.PolicyEnd AS
MotorPolicyEndDate, H.policyStart AS HealthPolicyStartDate, H.policyEnd AS HealthPolicyEndDate,
T.policyStart AS TravelPolicyStartDate, T.policyEnd AS TravelPolicyEndDate
INTO Insurance FROM
((Customer AS C LEFT JOIN Health_policies AS H ON C.HealthID = H.healthID) LEFT JOIN
Motor_policies AS M ON C.MotorID = M.motorID) LEFT JOIN Travel_policies AS T ON C.TravelID =
T.travelID
Order By C.CustomerID;
```

Data Quality Issues:

Data Quality Issue in SQL:

```
UPDATE Customer SET CardType = ‘Other mode of transcation’ WHERE CardType = '0';
UPDATE Customer SET Gender = 'Male' WHERE Gender = 'male' or Gender = 'm';
UPDATE Customer SET Gender = 'Female' WHERE Gender = 'female' or Gender = 'f';
UPDATE Customer SET ComChannel = 'Email' WHERE ComChannel = 'E';
UPDATE Customer SET ComChannel = 'SMS' WHERE ComChannel = 'S';
UPDATE Customer SET ComChannel = 'Phone' WHERE ComChannel = 'P';
UPDATE Customer SET Age = 40 WHERE Age > 85 or Age < 0;
UPDATE Health_policies SET DependentsKids = 4 WHERE DependentsKids= 40;
UPDATE Insurance SET ['Vehicle Value(USD)'] = 15100 WHERE ((((['Vehicle Value(USD)'])=0 Or
(Insurance.['Vehicle Value(USD)'])>130000));
```

SQL Queries for Insights:

1.
```
SELECT  Insurance.Location AS 'Location', Count(Insurance.motorID) AS ['Motor Insurance'],
Count(Insurance.healthID) AS ['Health Insurance'], Count(Insurance.travelID) AS ['Travel
Insurance'], Count(Insurance.CustomerID) AS ['Total']
FROM Insurance
GROUP BY Insurance.Location;
```

2.
```
SELECT Insurance.ComChannel AS 'Communication Channel', COUNT(Insurance.motorID) AS
'Motor Insurance', COUNT(Insurance.healthID) AS 'Health Insurance',
COUNT(Insurance.travelID) AS 'Travel Insurance', Round(Avg(Insurance.Age)) AS 'Average
Age'
FROM Insurance GROUP BY Insurance.ComChannel
ORDER BY COUNT(Insurance.motorID) DESC, COUNT(Insurance.healthID) DESC;
```

3.
```
SELECT Insurance.['Vehicle Age'], ROUND(Avg(Insurance.['Claim Amount (USD)']),2) AS
['Average Claim Amount (USD)'], ROUND(Avg(Insurance.['Vehicle Value(USD)']),2) AS ['Avg
Vehicle Value(USD)'], ROUND(Avg(Insurance.Age)) AS 'Average Age'
    FROM Insurance
    GROUP BY Insurance.['Vehicle Age'] HAVING Insurance.['Vehicle Age'] <> 0
ORDER BY Insurance.['Vehicle Age'];
```

4.
```
SELECT Insurance.HealthType, Count(Insurance.HealthDependentsAdults) AS ['Total Count
having Dependent Adults'],
Round(Avg(Insurance.HealthDependentsAdults+Insurance.DependentsKids),1) AS ['Average
Dependant per Customer'], Round(Avg(Insurance.Age)) AS ['Average age'],
Round(Avg(Insurance.HealthPolicyEndDate-Insurance.HealthPolicyStartDate),1) AS ['Average
Policy days']
FROM Insurance
GROUP BY Insurance.HealthType
HAVING (((Insurance.HealthType)<>''))
ORDER BY Insurance.HealthType ASC;
```

5.
```
SELECT Insurance.TravelType, Count(*) AS ['Total Count of Travel Type'], Round(Avg(Age)) AS
['Average Age'], Round(Avg(TravelPolicyEndDate-TravelPolicyStartDate)) AS ['Average days
of policy active']
FROM Insurance
```

```
GROUP BY Insurance.TravelType
HAVING (((Insurance.[TravelType])<>''))
ORDER BY Round(Avg(Age));
```

6. SELECT Insurance.CardType, Count(Insurance.healthID) AS 'Health Insurance', Count(Insurance.travelID) AS 'Travel Insurance', Count(Insurance.motorID) AS 'Motor Insurance'
   FROM Insurance
   GROUP BY Insurance.CardType
   ORDER BY Count(Insurance.healthID) DESC , Count(Insurance.travelID) DESC;

   Customers present in all the 3 tables query:
   SELECT Customer.*, Health_policies.*, Motor_policies.*, Travel_policies.*
   FROM ((Customer INNER JOIN Health_policies ON Customer.HealthID = Health_policies.healthID) INNER JOIN Motor_policies ON Customer.MotorID = Motor_policies.motorID) INNER JOIN Travel_policies ON Customer.TravelID = Travel_policies.travelID;

Appendix 2: R Code

```
#Install the required packages
#Read the Packages
library(dplyr)
library(tidyverse)

#Set Working Directory
setwd('D:/Business Analytics/Data Management')

#Read the excel sheet into variable
cust <- read.csv('D:/Business Analytics/Data Management/Assignment 1/customer.csv')
motor <- read.csv('D:/Business Analytics/Data Management/Assignment 1/motor_policies.csv')
health <- read.csv('D:/Business Analytics/Data Management/Assignment 1/health_policies.csv')
travel <- read.csv('D:/Business Analytics/Data Management/Assignment 1/travel_policies.csv')

#:: Creation of Analytical Base Table ::

#Change Column Names. Make it same as the other tables for motorID, heatlthID,travelID
colnames(cust)[12] <- 'motorID'
colnames(cust)[13] <- 'healthID'
colnames(cust)[14] <- 'travelID'

#Change the names of policyStart and policyEnd for all the tables
colnames(health)[3] <- 'HealthPolicyStart'
colnames(health)[4] <- 'HealthPolicyEnd'
colnames(motor)[4] <- 'MotorPolicyStart'
colnames(motor)[5] <- 'MotorPolicyEnd'
```

```
colnames(travel)[2] <- 'TravelPolicyStart'
colnames(travel)[3] <- 'TravelPolicyEnd'

#Join 2 tables and then join all the 4 tables at last
cust_health <- left_join(cust,health,"healthID")
cust_health_motor <- left_join(cust_health,motor,"motorID" )
cust_health_motor_travel <- left_join(cust_health_motor,travel,"travelID")
#cust_health_motor_travel Table is created from all the 4 tables

#Remove all the unwanted Columns from cust_health_motor_travel table, and change the
veh_value by multiplying by 10000

Insurance <- cust_health_motor_travel %>%
  distinct() %>%
  select(-`Title`,-`GivenName`,-`MiddleInitial`,-`Surname`,`Occupation`,-`MotorType`,-`clm`,-
`numclaims`,-`v_body`,-`LastClaimDate`) %>%
  mutate(veh_value = veh_value*10000)

#Analytical Base Table named 'Insurance' is created

#::::: Data Quality Issues and Action ::::::

##Re-summarize CardType 0 as 'Other Mode of Transaction'
Insurance$CardType[Insurance$CardType == 0] <- "Other Mode of Transaction"
#Convert CardType in factor
Insurance$CardType <- as.factor(Insurance$CardType)


#Re-summarize Gender as 'Male' WHERE Gender = 'male' or 'm'
Insurance$Gender[(Insurance$Gender == 'male')] <- 'Male'
Insurance$Gender[(Insurance$Gender == 'm')] <- 'Male'
#Re-summarize Gender as 'Female' WHERE Gender = 'female' or 'f'
Insurance$Gender[(Insurance$Gender == 'female')] <- 'Female'
Insurance$Gender[(Insurance$Gender == 'f')] <- 'Female'
#Convert Gender into Factor
Insurance$Gender <- as.factor(Insurance$Gender)

#Re-summarize ComChannel as 'Email' where ComChannel = 'E'
Insurance$ComChannel[Insurance$ComChannel == 'E'] <- 'Email'
#Re-summarize ComChannel as 'SMS' where ComChannel = 'S'
Insurance$ComChannel[Insurance$ComChannel == 'S'] <- 'SMS'
#Re-summarize ComChannel as 'Phone' where ComChannel = 'P'
Insurance$ComChannel[Insurance$ComChannel == 'P'] <- 'Phone'
#Convert ComChannel into factor
Insurance$ComChannel <- as.factor(cust$ComChannel)
```

```
#Re-summarize Age to 41 (average age 41.38) for the age above 100 or negative
Insurance$Age[Insurance$Age > 100] <- 41
Insurance$Age[Insurance$Age < 0] <- 41

#Re-summarize DependentsKids in table Health_policies to 4 from 40
Insurance$DependentsKids[Insurance$DependentsKids == 40] <- 4
#Convert DependentsKids into Factor
Insurance$DependentsKids <- as.factor(health$DependentsKids)

#Re-summarize veh_value 0 and above 130000 to median value 15100\
Insurance$veh_value[Insurance$veh_value > 130000] <- 15100
Insurance$veh_value[Insurance$veh_value == 0] <- 15100

#Convert HealthDependentsAdults into Factor
Insurance$HealthDependentsAdults <- as.factor(Insurance$HealthDependentsAdults)

#Convert HealthType into factor
Insurance$HealthType <- as.factor(health$HealthType)

#Convert MotorType into factor
Insurance$MotorType <- as.factor(Insurance$MotorType)

#Convert Vehicle Age into factor
Insurance$v_age <- as.factor(Insurance$v_age)

#Convert Travel type into factor
Insurance$TravelType <- as.factor(Insurance$TravelType)

#GGPLOT GRAPH For Representing Vehicle Price Outliers
Insurance %>% ggplot(aes(x=Insurance$v_age,y=Insurance$veh_value, color = Gender, na.rm =
TRUE))+
  geom_boxplot()+
  labs(title = "Comparison of Motor Price (USD) and Age of Vehicle (Years)", x="Age of Motor", y=
"Vehicle Price ($)", fill = "Gender")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
scale_fill_brewer(palette = "Dark2")
```

References

Insurance Information Institute, 2021. Background on: Compulsory Auto/Uninsured Motorists. *Insurance Information Institute,* March.

Stevens, S. B. a. P., 2013. Do You Really Need Travel Insurance?. *HeinOnline,* pp. 26-29.