



MGT7178: Data Management

Title : Empowering Insights: Using Data
Analysis to Create Targeted Insurance
Products and Marketing Strategies

Name: Dhanush Mathighatta Shobhan Babu

Student ID: 40412492

Word Count:2122

Table Of Content

Sl. No	Content	Page No
1.	1. Introduction and Background 1.1 Background and objectives of the problem	1
2.	2.Database development 2.1 Microsoft Access 2.2 Steps taken to create database 2.3 Description of given table and creation of Analytical Base Table 2.4 Database Structure 2.5 Limitation of the approach and technologies used 2.6 Overcoming limitations for the business	1-10
3.	3. Data Quality Report (R) 3.1 Steps in ABT table in R 3.2 Identification of Data Quality Issues 3.3 Implication of Data Quality Issues 3.4 Approaches Used to Address Data Quality 3.5 Business Strategies for Preventing Data Quality Issues	11-15
4.	4. Insights	16-19
5.	5. Appendix	20-26

Table Of Figures

Sl. No	Content	Page. No
1.	Figure 2.4	6
2.	Figure 3.4.1	12
3.	Figure 3.4.2	13
4.	Figure 3.4.3	13
5.	Figure 3.4.4	13
6.	Figure 3.4.5	14
7.	Figure 3.4.6	14
8.	Figure 4.1	15
9.	Figure 4.2	16
10.	Figure 4.3	17
11.	Figure 4.4	17
12.	Figure 4.5	18
13.	Figure 4.6	18

1. Introduction

1.1 Background and Objectives of the Problem

In the highly competitive insurance market, managing all data is difficult. Our consultant must consolidate data types from numerous files in the organization's repository. The firm struggles to find significant data due to its widespread distribution. The primary objective of our study is to employ SQL and R programming languages to analyse a dataset and identify significant associations among consumer characteristics, various insurance types, and the effectiveness of diverse marketing techniques.

2. Database development

2.1 Microsoft Access

The data analysis of the provided dataset was conducted using SQL, with Microsoft Access being employed as the software tool. Microsoft Access is chosen due to its user-friendly interface, which facilitates the efficient creation of databases. Additionally, it facilitates seamless integration of data management across multiple platforms and offers a cost-effective solution. Therefore, it is a suitable selection for various projects.

2.2 Steps taken to create database

We created a new database using the create function on Microsoft access and then we imported all the 4 excel sheets as tables to our database using the external data import function.

2.3 Description of Given Data and Creation of Analytical Base Table (ABT)

The information is given in tabular form, as explained below:
Using SQL queries, this descriptive analysis was obtained.

1. File name : Data 1_Customer.xlsx

Sl.No	Variable Name	Type of Data	Description	No.Of Distinct Values
1.	Customer ID	Number	Distinct Value (Primary key)	N/A
2.	Title	Short Text	The name of the client (Dr., Ms., Mr., Mr., Mrs.)	5
3.	Given Name	Short Text	The names of each individual client	N/A

4.	Middle Initial	Short Text	The customer's middle initials	N/A
5.	Surname	Short Text	The last name of the client	N/A
6.	Card Type	Short Text	The customer's credit card type (0, Mastercard, Visa)	3
7.	Occupation	Short Text	distinct occupation of the customer	
8.	Gender	Short Text	Gender classification of the customer (f, female, m, male)	4
9.	Age	Number	Age of the client	N/A
10.	Location	Short Text	Displays the client's location (Urban, Rural).	2
11.	ComChannel	Short Text	The chosen method of communication for the consumer (E, Email, P, Phone, S, SMS)	6
12.	Motor ID	Number	Unique value - taken from motor policy data (Foreign key)	N/A
13.	Health ID	Number	Distinct Value - taken from health policy data (Foreign key)	N/A
14.	Travel ID	Number	Distinct Value - taken from travel policy data (Foreign key)	N/A

Table 2.3.1: Description Customer Table data

2. File name : Data 2_Motor Policies.xlsx

Sl.No	Variable Name	Type of Data	Description	No.Of Distinct Values
1.	Motor ID	Number	It only includes distinct values (Primary Key)	N/A
2.	Policy Start	Date/Time	Date of policy start	N/A
3.	Policy End	Date/Time	Date of policy termination	N/A
4.	Motor Type	Short Text	What type of auto insurance (bundle or single) the client possesses	2
5.	veh_value	Number	Car Worth	N/A
6.	Exposure	Number	From 0 to 1.	N/A
7.	Clm	Number	Insurance claim status: 0/1 (zero = not claimed, one = claimed)	2

8.	Numclaims	Number	No of claims: 0, 1, 2, 4, (0 indicates there are no claims).	4
9.	v_body	Short Text	Vehicle body type, with codes for HBACK, HDTOP, COUPE, BUS, and CONVT RDSTR, SEDAN, STNWG, TRUCK, UTE, MCARA, MIBUS, and PANVN.	13
10.	v_age	Number	Age of vehicle: 1, 2, 3, 4, with 1 being the youngest.	4
11.	Last Claim Date	Date and Time	previous auto insurance claim.	N/A

Table 2.3.2: Description Motor Table data

3. File name: Data 3_Health Policies.xlsx

Sl.No	Variable Name	Type of Data	Description	No.Of Distinct Values
1.	Health ID	Number	Distinct Value (Primary key)	N/A
2.	Policy Start	Date/Time	Date of policy start	N/A
3.	Policy End	Date/Time	Date of policy termination	N/A
4.	Health Type	Short Text	The customer's type of health insurance is (Level 1, Level 2, or Level 3).	3
5.	Health Dependents Adults	Number	The number of adults who are reliant on that individual's health coverage (0,1,2)	3
6.	Dependents Kids	Number	The number of children who are insured by the person (0,1,2,3,40)	5

Table 2.3.3: Description Health Table data

4. File name: Data 4_Travel Policies.xlsx

Sl.No	Variable Name	Type of Data	Description	No. Of Distinct Values
1.	Travel ID	Number	It only includes distinct values. (Primary Key)	N/A
2.	Policy Start	Date/Time	Date of policy start	N/A
3.	Policy End	Date/Time	Date of policy termination	N/A
4.	Travel Type	Short Text	Which type of travel insurance does the customer have. (Business, Premium, Senior, Standard, Backpacker)	5

Table 2.3.4: Description Travel Table data

5. ABT table :

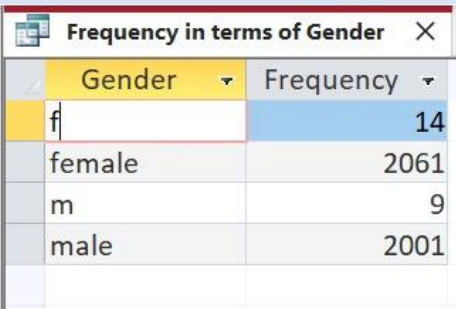
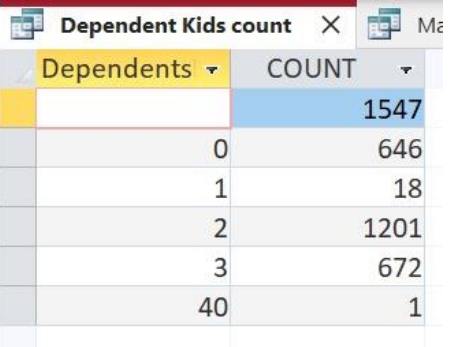

Sl.No	Data Variable Name	Data Type	Description	No.Of Distinct Values
1.	Customer ID	Number	Distinct Value (Primary key)	N/A
2.	Title	Short Text	The name of the client (Dr., Ms., Mr., Mr., Mrs.)	5
3.	GivenName	Short Text	The names of each individual client	N/A
4.	MiddleInitial	Short Text	The customer's middle initials	N/A
5.	Surname	Short Text	The last name of the client	N/A
6.	CardType	Short Text	The customer's credit card type (0, Mastercard, Visa)	3
7.	Occupation	Short Text	distinct occupation of the customer	N/A
8.	Gender	Short Text	Gender classification of the customer (f, female, m, male)	4
9.	Age	Number	Age of the client	N/A
10.	Location	Short Text	Displays the client's location (Urban, Rural).	2
11.	ComChannel	Short Text	The chosen method of communication for the	6

			consumer (E, Email, P, Phone, S, SMS)	
12.	MotorID	Number	Unique value - taken from motor policy data (Foreign key)	N/A
13.	HealthID	Number	Distinct Value - taken from health policy data (Foreign key)	N/A
13.	TravelID	Number	Distinct Value - taken from travel policy data (Foreign key)	N/A
14.	PolicyStart	Number	Date of policy start	N/A
15.	PolicyEndx	Date/Time	Date of policy termination	N/A
16.	MotorType	Date/Time	What type of auto insurance (bundle or single) the client possesses	N/A
17.	veh_value	Short Text	Car Worth	2
18.	Exposure	Number	From 0 to 1.	N/A
19.	Clm	Number	Insurance claim status: 0/1 (zero = not claimed, one = claimed)	N/A
20.	Numclaims	Number	No of claims: 0, 1, 2, 4, (0 indicates there are no claims).	2
21.	v_body	Short Text	Vehicle body type, with codes for HBACK, HDTOP, COUPE, BUS, and CONVT RDSTR, SEDAN, STNWG, TRUCK, UTE, MCARA, MIBUS, and PANVN.	4
22.	v_age	Age	Age of vehicle: 1, 2, 3, 4, with 1 being the youngest.	13
23.	LastClaimDate	Number	previous auto insurance claim.	4
24.	PolicyStartx	Date/Time	Date of policy start	N/A
25.	PolicyEnd	Date/Time	Date of policy termination	N/A
26.	HealthType	Short Text	The customer's type of health insurance is (Level 1, Level 2, or Level 3).	3
27.	Health Dependent Adults	Number	The number of adults who are reliant on that individual's health coverage (0,1,2)	3
28.	DependentsKids	Number	The number of children who are insured by the person (0,1,2,3,40)	5

29.	Polycystarty	Date/Time	Date of policy start	N/A
30.	PolicyEndy	Date/Time	Date of policy termination	N/A
31.	TravelType	Short Text	Which type of travel insurance does the customer have. (Business, Premium, Senior, Standard, Backpacker)	5

Table 2.3.5: Description ABT Table data

Furthermore We conducted descriptive analysis using a few simple SQL queries, and the results are shown in the table below:

Descriptive Analysis Picture	Sql Code														
 <table border="1"> <thead> <tr> <th>Gender</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>f</td> <td>14</td> </tr> <tr> <td>female</td> <td>2061</td> </tr> <tr> <td>m</td> <td>9</td> </tr> <tr> <td>male</td> <td>2001</td> </tr> </tbody> </table>	Gender	Frequency	f	14	female	2061	m	9	male	2001	<pre>SELECT Gender, COUNT(*) AS Frequency FROM ABT GROUP BY Gender;</pre>				
Gender	Frequency														
f	14														
female	2061														
m	9														
male	2001														
 <table border="1"> <thead> <tr> <th>Dependents</th> <th>COUNT</th> </tr> </thead> <tbody> <tr> <td></td> <td>1547</td> </tr> <tr> <td>0</td> <td>646</td> </tr> <tr> <td>1</td> <td>18</td> </tr> <tr> <td>2</td> <td>1201</td> </tr> <tr> <td>3</td> <td>672</td> </tr> <tr> <td>40</td> <td>1</td> </tr> </tbody> </table>	Dependents	COUNT		1547	0	646	1	18	2	1201	3	672	40	1	<pre>SELECT ABT.DependentsKids, Count(*) AS [COUNT] FROM ABT GROUP BY ABT.DependentsKids;</pre>
Dependents	COUNT														
	1547														
0	646														
1	18														
2	1201														
3	672														
40	1														
 <table border="1"> <thead> <tr> <th>MaxAge</th> <th>MinAge</th> </tr> </thead> <tbody> <tr> <td>210</td> <td>-44</td> </tr> </tbody> </table>	MaxAge	MinAge	210	-44	<pre>SELECT MAX(Age) AS MaxAge, MIN(Age) AS MinAge FROM ABT;</pre>										
MaxAge	MinAge														
210	-44														

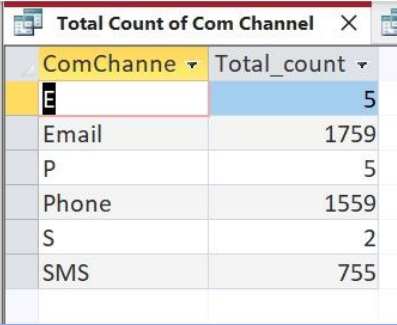
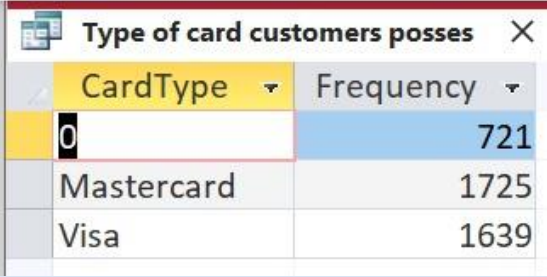
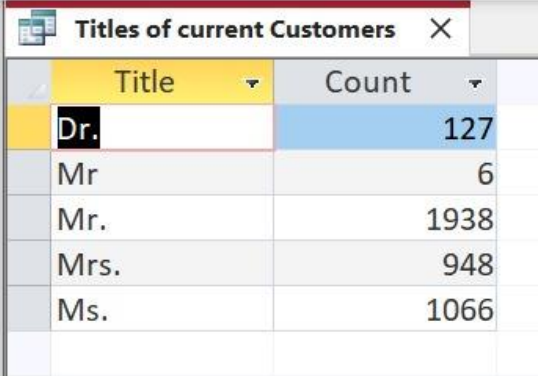
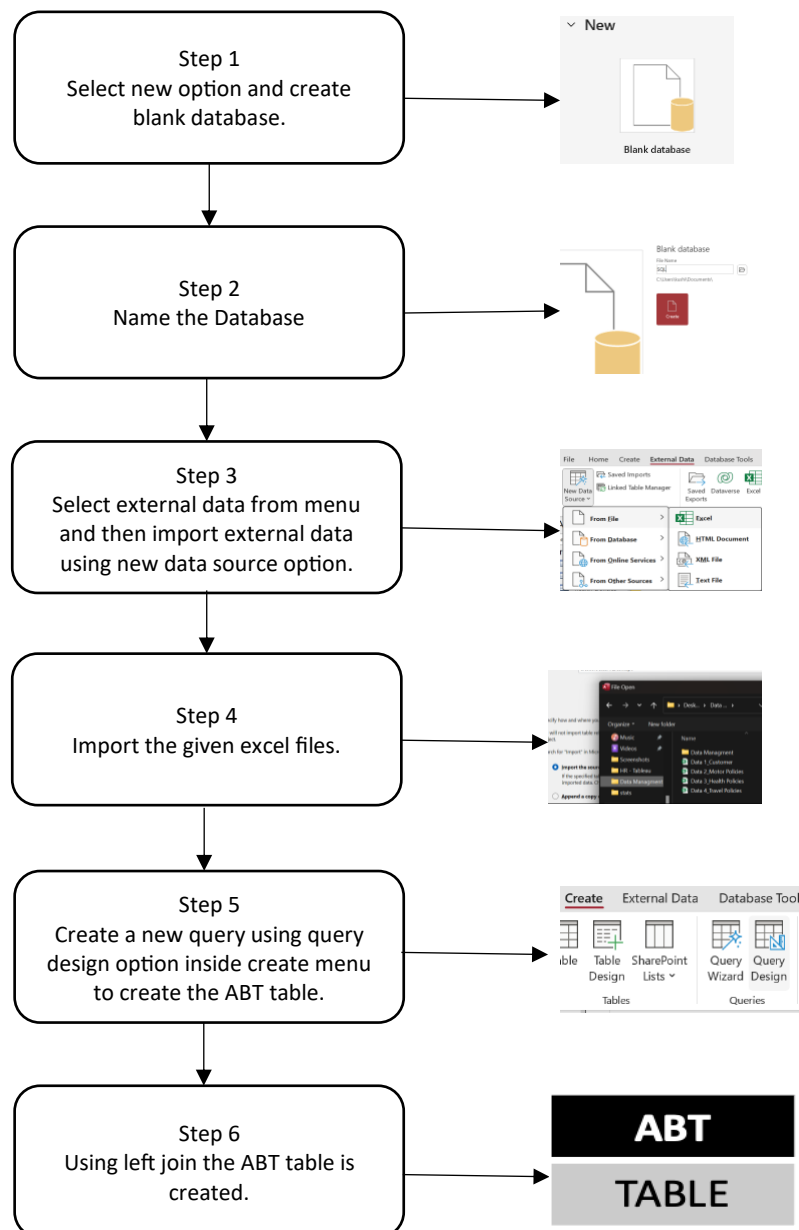
 <table border="1"> <thead> <tr> <th>ComChannel</th> <th>Total_count</th> </tr> </thead> <tbody> <tr> <td>E</td> <td>5</td> </tr> <tr> <td>Email</td> <td>1759</td> </tr> <tr> <td>P</td> <td>5</td> </tr> <tr> <td>Phone</td> <td>1559</td> </tr> <tr> <td>S</td> <td>2</td> </tr> <tr> <td>SMS</td> <td>755</td> </tr> </tbody> </table>	ComChannel	Total_count	E	5	Email	1759	P	5	Phone	1559	S	2	SMS	755	<pre>SELECT ComChannel, COUNT(*) AS Total_count FROM ABT GROUP BY ComChannel;</pre>
ComChannel	Total_count														
E	5														
Email	1759														
P	5														
Phone	1559														
S	2														
SMS	755														
 <table border="1"> <thead> <tr> <th>CardType</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>721</td> </tr> <tr> <td>Mastercard</td> <td>1725</td> </tr> <tr> <td>Visa</td> <td>1639</td> </tr> </tbody> </table>	CardType	Frequency	0	721	Mastercard	1725	Visa	1639	<pre>SELECT CardType, COUNT(*) AS Frequency FROM ABT GROUP BY CardType;</pre>						
CardType	Frequency														
0	721														
Mastercard	1725														
Visa	1639														
 <table border="1"> <thead> <tr> <th>Title</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>Dr.</td> <td>127</td> </tr> <tr> <td>Mr</td> <td>6</td> </tr> <tr> <td>Mr.</td> <td>1938</td> </tr> <tr> <td>Mrs.</td> <td>948</td> </tr> <tr> <td>Ms.</td> <td>1066</td> </tr> </tbody> </table>	Title	Count	Dr.	127	Mr	6	Mr.	1938	Mrs.	948	Ms.	1066	<pre>SELECT Title, COUNT(*) AS [Count] FROM ABT GROUP BY Title;</pre>		
Title	Count														
Dr.	127														
Mr	6														
Mr.	1938														
Mrs.	948														
Ms.	1066														

Table 2.3.6: Descriptive Analysis using SQL

Using the query design option under the create menu we created a query and joined all the four quires using left join. The selection of a left join allows for the inclusion of all records from the primary database, while also combining matching data from linked tables. This methodology guarantees the maintenance of data integrity and offers significant insights into the connections among the data. This functionality ensures the retention of all primary table records, regardless of any matches in the linked table, hence providing a comprehensive view of data connections.



Flow chart 2.3: Steps in access to create ABT

2.4 Database Structure

As mentioned previously, left join was used to create a single analytical table. Our Main table is the customer table, which comprises foreign keys as well as a primary key.

The primary key of the customer table is "*CustomerID*," while the foreign keys are "*HealthID*," "*MotorID*," and "*TravelID*." Foreign keys specified in the customer table serve as primary keys in the following tables:

- "*HealthID*" is the primary key for Health policies table
- "*MotorID*" is the primary key for Motor policies table
- "*TravelID*" is the primary key for Travel policies table

We have joined and created the ABT table based The above keys mentioned.

The following figure constitutes the database's structure:

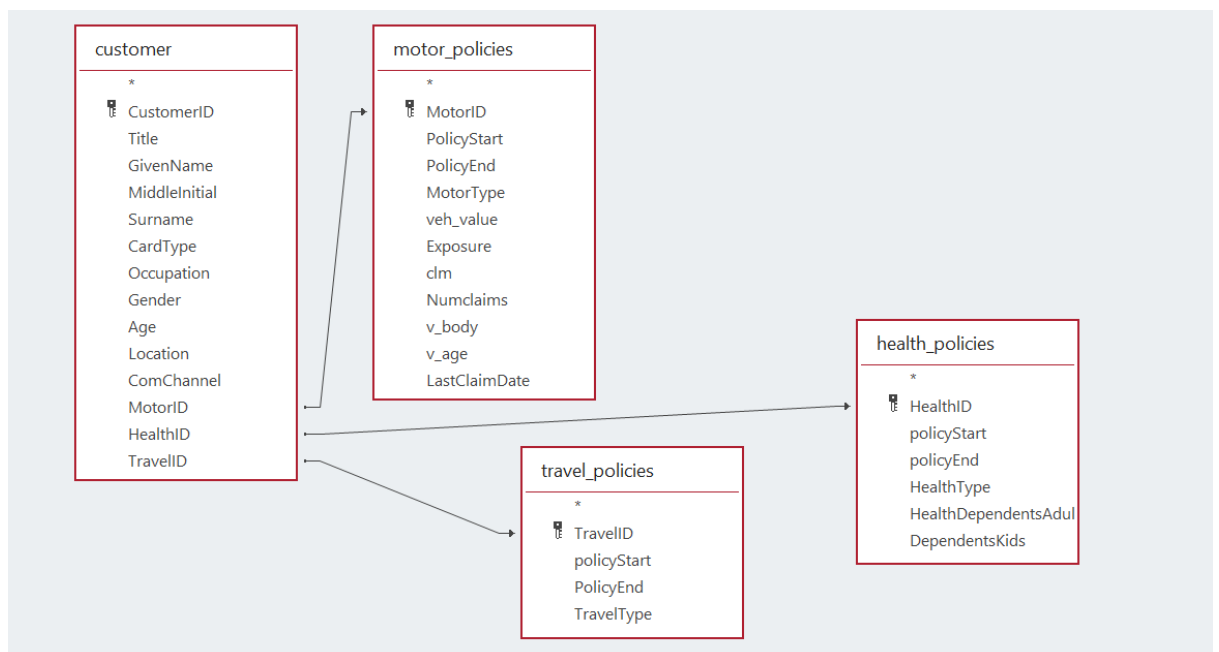


Fig 2.4: Database Structure

2.5 Limitations of the Approach and Technologies used

The limits of SQL encompass various aspects, such as scalability concerns when dealing with extensive datasets, the intricacy of handling unstructured data, the learning curve associated with executing sophisticated queries, and the possible difficulties in achieving optimal speed in concurrent contexts.

2.6 Overcoming Limitations for the Business

Scalability is crucial for huge datasets. NoSQL databases or SQL partitioning could be beneficial. Effective data management requires SQL and NoSQL database integration and SQL's JSON/XML capabilities for managing different data formats. The inclusion of R into data analytic workflows improves statistical analysis, data visualisation, and machine learning, enabling SQL to deliver complete data insights

3.Data Quality Report (R)

3.1 Steps to create ABT table in R

The necessary libraries in the R programming language, such as "tidyverse", "readxl", and "dplyr", are invoked. The initial step involves importing the provided Raw data into the R programming environment using the import function. In order to facilitate coding, we proceed to construct variables named data1, data2, data3, and data4, and assign the respective imported data to each variable. The table "ABT" is constructed by the utilisation of the left_join function from the "dplyr" package. This left_join operation is conducted by matching the primary and foreign keys specified earlier. Additionally, the pipe function is employed to facilitate the execution of the left_join operation.

3.2 Identification of Data Quality Issues

After creating the ABT we must find out the data quality issues, we find that by using various functions in r such as count (), summary(), str(), is.na() and plotting various box plots. The use of these functions are as follows:

count(): Calculates the number of occurrences within a dataset.

summary(): Generates descriptive statistics summarizing dataset properties.

str(): Displays the structure and attributes of an object.

is.na(): Identifies and flags missing values within a dataset.

Plotting Box Plots: Visualizes data distribution, outliers, and range across different categories or variables.

The identified data quality issues are shown in the below table:

Data	Data type	Outliers and quality issues	Issue
Title	Varchar [Categorical]	'Mr'	There exist two separate categories for the title "Mr.," namely "Mr" and "Mr." The presence of these two categories is

			considered incorrect. There are six entries of the title "Mr" that need to be appended to the abbreviation "Mr."
Cardtype	Varchar [Categorical]	'0'	There exists an entry labelled '0' for the CardType, which requires modification.
Gender	Varchar [Categorical]	'f', 'm'	There are a total of four categories within the gender classification, namely "male," "female," "f," and "m." The usage of the abbreviations 'f' and 'm' is deemed incorrect and should be replaced with the terms 'male' and 'female'.
Age	Integer [Numerical]	-44,180,210	The illogical ages -44, 180, and 210 must be filtered or substituted with the mean.
ComChannel	Varchar [Categorical]	'E', 'P', 'S'	There are only three distinct categories, namely 'Email', 'Phone', and 'Email'. The letters 'E', 'P', and 'S' necessitate renaming.
DepedentsKids	Integer [Numerical]	40	The entry labelled '40' lacks logical coherence and should be subjected to filtration or replaced with the mean value.

Table 3.2: Data quality analysis

3.3 Implication of Data Quality Issues

Inconsistent Categorical Values: Inconsistency is introduced, and categorical analysis may be distorted by the existence of disparate representations for titles, genders, and communication channels. Correcting these guarantees a consistent and precise comprehension of demographic distributions.

Inaccurate Numerical Entries: Statistical computations and summary statistics are distorted by illogical or outlier numerical values, such as negative ages or abnormally high values. For precise numerical studies and average computations, these abnormalities must be corrected.

Data Accuracy and Integrity: Inaccurate entries that cause problems for trend analysis or demographic profiling are those that start with '0' for CardType or '40' for

Effect on Analysis and Decision-Making: Inaccurate data may distort analysis, resulting in incorrect conclusions. To improve the validity and applicability of any findings or plans created from the data, these problems must be fixed.

3.4 Approaches Used to Address Data Quality

In order to effectively resolve the aforementioned data quality concerns, we utilised the mutate () and filter () functions from the "dplyr" library, as well as the replace () function from the Base R package. Pipes function also have been incorporated.

These functions in Rare used for a variety of data manipulation tasks.

1. replace (): With the help of this function, you can add new values to certain locations within a vector, matrix, or data frame. It enables precise substitutions, changing certain components according to predetermined standards.

2. mutate (): This dplyr package function adds new variables to a data frame or alters ones that already present. To create derived variables, it is frequently used to create new columns by applying operations to pre-existing columns.

3. filter (): Filter (), another function in the dplyr package, is used to extract rows from a data frame according to predetermined standards or criteria. By choosing rows that satisfy predetermined logical criteria, users can subset data.

The subsequent images depict the solved data issues prior to and after their resolution:

We have created a new variable ABT_clean to reflect the changes made:

1. Title: By mutating Mr to Mr., the corresponding entries have been appended to Mr.

```
> count(ABT,Title)
```

```
# A tibble: 5 × 2
```

```
  Title      n  
  <chr> <int>
```

```
1 Dr.      127  
2 Mr.       6  
3 Mr.     1938  
4 Mrs.     948  
5 Ms.     1066
```

After



cleaning

```
> count(ABT_clean,Title)
```

```
# A tibble: 4 × 2
```

```
  Title      n  
  <chr> <int>
```

```
1 Dr.      127  
2 Mr.     1943  
3 Mrs.     947  
4 Ms.     1065  
> |
```

Fig 3.4.1: Data analysis of Title

2. CardType : We have replaced the entry "0" with the string "Not_given" using mutate function

```
> count(ABT, CardType)
# A tibble: 3 x 2
  CardType      n
  <chr>    <int>
1 0          721
2 Mastercard 1725
3 Visa      1639
> |
```

After
cleaning

```
> count(ABT_clean, CardType)
# A tibble: 3 x 2
  CardType      n
  <chr>    <int>
1 Mastercard 1724
2 Not_Given   720
3 Visa      1638
>
```

Fig 3.4.2: Data analysis of CardType

3. Gender : The entries "f" and "m" have been substituted with "male" and "female" respectively during the process of mutation.

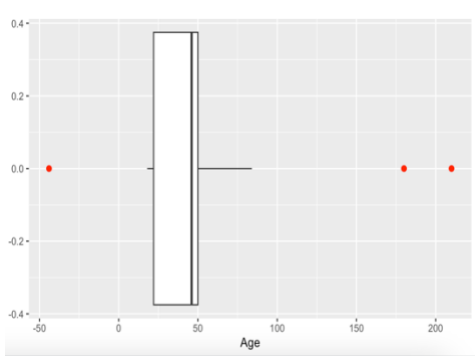
```
> count(ABT, Gender)
# A tibble: 4 x 2
  Gender      n
  <chr>    <int>
1 f          14
2 female 2061
3 m           9
4 male    2001
```

After
cleaning

```
> count(ABT_clean, Gender)
# A tibble: 2 x 2
  Gender      n
  <chr>    <int>
1 female 2073
2 male   2009
>
```

Fig 3.4.3: Data analysis of Gender

4. Age : The filter() function is utilized to exclude irrational ages -44, 180, and 210. The ggplot package was employed to generate visual representations of the outliers.



After
cleaning

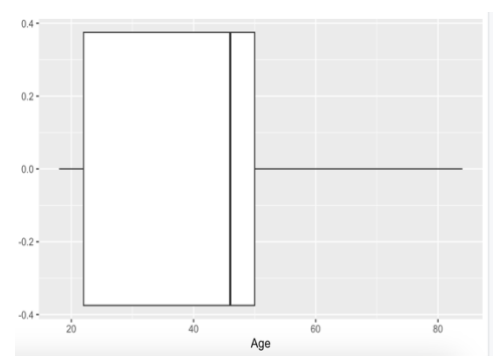


Fig 3.4.4: Data analysis of Age

5. ComChannel: The letters "E," "P," and "S" have been changed into "Email," "Phone," and "SMS," respectively using mutate function:

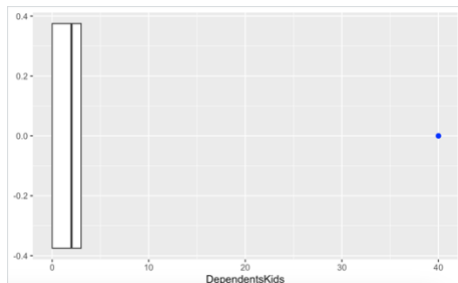
```
> count(ABT, ComChannel)
# A tibble: 6 x 2
  ComChannel      n
  <chr>      <int>
1 E             5
2 Email       1759
3 P             5
4 Phone       1559
5 S             2
6 SMS         755
> |
```

After
cleaning

```
> count(ABT_clean, ComChannel)
# A tibble: 3 x 2
  ComChannel      n
  <chr>      <int>
1 Email       1763
2 Phone       1564
3 SMS          755
>
```

Fig 3.4.5: Data analysis of ComChannel

6. DependentsKids: The outlier value "40" has been substituted with the mean value of the DependentsKids variable, which is "2".



After
cleaning

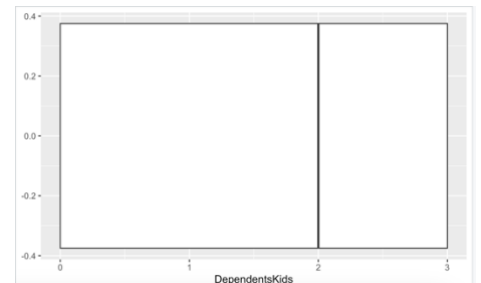


Fig 3.4.6: Data analysis of DependentsKids

3.5 Business Strategies for Preventing Data Quality Issues

Standardised gathering techniques, regular audits, and complete employee training ensure accurate data entry and reduce future data quality issues. Automated validation tools, explicit governance rules, and frequent data maintenance are crucial. Modern technologies and a data-focused culture support data accuracy and reliability efforts.

4. Insights

4.1 SQL Functions used for insights

SUM(): Determines the total number of policies for various insurance types and counts the number of renewed policies by adding up the provided values.

IIF(): Conditional function used to count a renewed policy if the difference between PolicyStart and PolicyEnd is greater than or equal to 365 days; if not, it counts other conditions based on the logic of the query.

DATEDIFF(): Determines the difference between two dates; it is used to calculate the time interval between PolicyStart and PolicyEnd in order to determine the status of policy renewals.

COUNT(): Determines the overall count of policies for various insurance kinds by counting the instances of a given value or row in a dataset.

GROUP BY: Used in a variety of queries to aggregate data for analysis based on particular categories like policy type, geography, age group, or gender, this function groups the return set according to designated columns.

4.2 insights Obtained

The ABT_clean data exported from R is utilised to derive insights. The next step involves importing the ABT_clean dataset into Microsoft Access in order to utilise SQL for deriving insights.

1. **Insights on Retention rate**

The study provides evidence of the variability in client retention rates for insurance products. Health insurance plans demonstrate a higher degree of client retention compared to automobile insurance plans, while travel insurance programmes exhibit the lowest level of customer retention. In order to foster customer retention, it is imperative to develop strategies including all three policies, with a particular focus on travel and vehicle plans owing to their comparatively lower costs.

PolicyType	RenewedPolic	TotalPolicies	RetentionRate
Motor	1089	3354	32.4686940966011
Health	1089	2538	42.9078014184397
Travel	483	2102	22.9781160799239

Fig 4.1 Query Output For retention rate of Each policy

2. Insights on Gender, Age and different policies

A uniform pattern can be seen in all age groups and both genders when looking at the provided data table, suggesting that there aren't many transactions in the travel insurance area. This pattern emphasises how critical it is to have strategic strategies in place to boost sales in the travel insurance industry. It is advisable to implement a comprehensive plan that combines targeted promotional campaigns, targeted marketing initiatives, and significant customer education efforts in order to effectively address this scenario.

Gender	age_range	total_custon	health_polic	motor_polic	travel_polic
female	18-25	645	161	497	448
female	26-40	8	5	5	4
female	41-55	1195	942	1034	475
female	56-60	13	9	12	6
female	Above 60	212	179	147	134
male	18-25	603	141	457	424
male	26-40	13	11	12	6
male	41-55	1185	911	1029	467
male	56-60	14	14	13	9
male	Above 60	194	165	148	129

Fig 4.2 Query Output for Gender Age and different policies.

3. Insights on Location, Gender and Count of all policies

The data indicates that urban residents, regardless of gender, exhibit a higher propensity to embrace policy. The identification of this pattern presents an opportunity to improve rural marketing strategies. To enhance sales and hence augment revenue in rural regions, it is advisable to employ promotional tactics that emphasise the advantages of the policy. The aforementioned advantages must to be articulated through customised approaches, emphasising the reduction of risks and the consideration of long-term consequences. Enhancing policy sales in rural areas can be achieved by customising advertising messaging to address the distinct needs and concerns of individuals residing in these regions.

Query3 X Location VS Gender VS Policy Count		
Location	Gender	PolicyCount
Rural	female	894
Rural	male	868
Urban	female	1179
Urban	male	1141

Fig 4.3 Query Output for Location, Gender and Count of all policies

4. Insights on ComChannel , Location and Customer Count

Urban clientele, due to their familiarity with digital media, tend to prioritise the use of email and SMS. These modes are well-suited for the urban lifestyle characterised by a strong preoccupation with technology. There is a preference for rural locations in terms of call response times, as they tend to be faster for clients located in these regions. Phone interactions have the potential to enhance the sense of connection and reliability among rural clients due to their direct and personalised nature. The communication strategy should be revised in geographically isolated areas to prioritise methods that rely on telephonic means. Engaging in direct telephone communication with clients facilitates more personalised discussions and effectively caters to their distinct requirements and challenges.

ComChannel	Location	CustomerCo
Email	Rural	567
Email	Urban	1196
Phone	Rural	939
Phone	Urban	625
SMS	Rural	256
SMS	Urban	499

Fig 4.4 Query Output for ComChannel , Location and Customer Count

5. Insights on Age group and total claims

Age cohorts vary in insurance purchase propensity. The 41–60 group has the highest customer engagement and insurance coverage. Customers between 26 and 40 have fewer insurance coverage due to lower engagement. To reduce the age gap in insurance policy uptake, customised marketing campaigns should stress the importance and benefits of insurance for 26-to-40-year-olds. Reduce insurance premiums for 26–40-year-olds. Additional advantages boost insurance enrolment. Digital platforms and social media can be used for focused outreach. Use online ads to target specific age groups' tastes and behaviours.

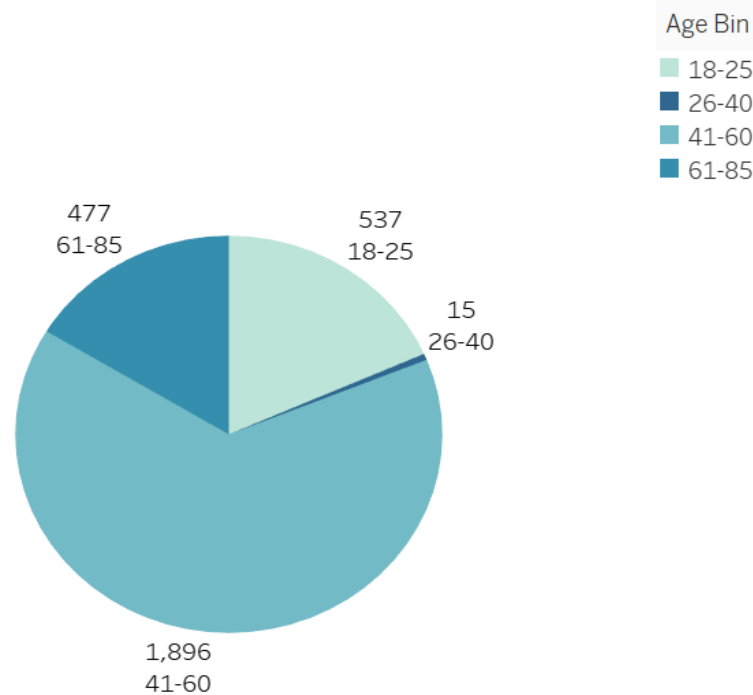


Fig 4.5 Tableau Output for Age group and total claims

6. Insights on AgeGroup , Count of Customer and AvgClaimPercent

As people near retirement age, insurance claims peak in the 56-60 age bracket. Health issues often accompany this phenomena. Due to rising health issues, insurance claims in this age group are rising. Age Group 60-85 Retirement Category and Lowest Claims: The lowest claim frequency is among pensioners aged 60–85. Retirees may buy less auto and travel insurance, reducing claims. Due to their caution and risk aversion, 41-55-year-olds may make fewer insurance claims. They may have fewer accidents or health issues than younger or older people. Between 18 and 25, people have better health and do less insurance-risky activities. These individuals may have fewer chronic health conditions, reducing insurance claims.

AgeGroup	CustomerCount	AvgClaimspe
18-25 Entry Level / Early Professional	1248	3.98
26-40 Mid-Career / Intermediate Professional	21	5.88
41-55 Senior/Experienced Professional	2380	3.2
56-60 Pre-Retirement/Late Career	27	8
60-85 Retired	406	2.03

Fig 4.6 Query Output for AgeGroup , Count of Customer and AvgClaimPercent

The aforementioned insights and marketing suggestions should be applied as marketing strategies to enhance policy sales.

5. Appendix

SQL code :

. Code for left join in SQL

```
SELECT customer.CustomerID, customer.Title, customer.GivenName,
customer.MiddleInitial, customer.Surname, customer.CardType, customer.Gender,
customer.Occupation, customer.Age, customer.Location, customer.ComChannel,
customer.MotorID, motor_policies.PolicyStart, motor_policies.PolicyEnd,
motor_policies.MotorType, motor_policies.veh_value, motor_policies.Exposure,
motor_policies.clm, motor_policies.Numclaims, motor_policies.v_body,
motor_policies.v_age, motor_policies.LastClaimDate, health_policies.policyStart,
health_policies.policyEnd, health_policies.HealthType,
health_policies.HealthDependentsAdults, health_policies.DependentsKids,
travel_policies.policyStart, travel_policies.PolicyEnd, travel_policies.TravelType
INTO ABT
FROM ((customer LEFT JOIN motor_policies ON customer.[MotorID] =
motor_policies.[MotorID]) LEFT JOIN health_policies ON customer.[HealthID] =
health_policies.[HealthID]) LEFT JOIN travel_policies ON customer.[TravelID] =
travel_policies.[TravelID];
```

2. Descriptive Analysis using SQL

```
SELECT Gender, COUNT(*) AS Frequency
FROM ABT
GROUP BY Gender;
```

```
SELECT ABT.DependentsKids, Count(*) AS [COUNT]
FROM ABT
GROUP BY ABT.DependentsKids;
```

```
SELECT MAX(Age) AS MaxAge, MIN(Age) AS MinAge
FROM ABT;
```

```
SELECT ComChannel, COUNT(*) AS Total_count
FROM ABT
GROUP BY ComChannel;
```

```
SELECT CardType, COUNT(*) AS Frequency
FROM ABT
GROUP BY CardType;
```

```
SELECT Title, COUNT(*) AS [Count]
FROM ABT
GROUP BY Title;
```

3. Insights Queries using SQL

Retention rate

```
SELECT 'Motor' AS PolicyType,
      SUM(IIF(DATEDIFF('d', PolicyStart, PolicyEnd) >= 365, 1, 0)) AS RenewedPolicies,
      COUNT(*) AS TotalPolicies,
      (SUM(IIF(DATEDIFF('d', PolicyStart, PolicyEnd) >= 365, 1, 0)) / COUNT(*)) * 100 AS
RetentionRate
FROM ABT_clean
WHERE MotorID IS NOT NULL
UNION ALL
SELECT 'Health' AS PolicyType,
      SUM(IIF(DATEDIFF('d', policyStart, policyEnd) >= 365, 1, 0)) AS RenewedPolicies,
      COUNT(*) AS TotalPolicies,
      (SUM(IIF(DATEDIFF('d', policyStart, policyEnd) >= 365, 1, 0)) / COUNT(*)) * 100 AS
RetentionRate
FROM ABT_clean
WHERE HealthID IS NOT NULL
UNION ALL
SELECT 'Travel' AS PolicyType,
      SUM(IIF(DATEDIFF('d', policyStart, PolicyEnd) >= 365, 1, 0)) AS RenewedPolicies,
      COUNT(*) AS TotalPolicies,
      (SUM(IIF(DATEDIFF('d', policyStart, PolicyEnd) >= 365, 1, 0)) / COUNT(*)) * 100 AS
RetentionRate
FROM ABT_clean
WHERE TravelID IS NOT NULL;
```

Communication VS Location

```
SELECT ComChannel, Location, COUNT(*) AS CustomerCount
FROM ABT_clean
GROUP BY ComChannel, Location;
```

Location VS Gender VS Policy Count

```
SELECT Location, Gender, Count(*) AS PolicyCount
FROM ABT_clean
GROUP BY Location, Gender;
```

Age-Group VS Total_Claims

```
SELECT
  Gender,
  IIF(Age BETWEEN 18 AND 25, '18-25',
    IIF(Age BETWEEN 26 AND 40, '26-40',
```



```

        IIF(Age BETWEEN 41 AND 55, '41-55',
            IIF(Age BETWEEN 56 AND 60, '56-60',
                IIF(Age > 60, 'Above 60', 'Other')
            )
        )
    ) AS age_range,
    COUNT(CustomerID) AS total_customers,
    COUNT(IIF(HealthID IS NOT NULL, customerID, NULL)) AS health_policies,
    COUNT(IIF(MotorID IS NOT NULL, customerID, NULL)) AS motor_policies,
    COUNT(IIF(TravelID IS NOT NULL, customerID, NULL)) AS travel_policies
FROM
    ABT_clean
GROUP BY
    Gender,
    IIF(age BETWEEN 18 AND 25, '18-25',
        IIF(age BETWEEN 26 AND 40, '26-40',
            IIF(age BETWEEN 41 AND 55, '41-55',
                IIF(age BETWEEN 56 AND 60, '56-60',
                    IIF(age > 60, 'Above 60', 'Other')
                )
            )
        )
    );

```

Age_group VS Claim %

```

SELECT
    IIF(ABT_clean.Age BETWEEN 18 AND 25, '18-25 Entry Level / Early Professional',
        IIF(ABT_clean.Age BETWEEN 26 AND 40, '26-40 Mid-Career / Intermediate
Professional',
            IIF(ABT_clean.Age BETWEEN 41 AND 55, '41-55 Senior/Experienced
Professional',
                IIF(ABT_clean.Age BETWEEN 56 AND 60, '56-60 Pre-Retirement/Late
Career', 'Other')
            )
        )
    ) AS AgeGroup,
    COUNT(*) AS CustomerCount,
    ROUND(( AVG(ABT_clean.Numclaims)/MAX(ABT_clean.Numclaims))*100,2) AS
AvgClaimspercentage
FROM ABT_clean
GROUP BY
    IIF(ABT_clean.Age BETWEEN 18 AND 25, '18-25 Entry Level / Early Professional',
        IIF(ABT_clean.Age BETWEEN 26 AND 40, '26-40 Mid-Career / Intermediate
Professional',

```

```

        IIF(ABT_clean.Age BETWEEN 41 AND 55, '41-55 Senior/Experienced
Professional',
        IIF(ABT_clean.Age BETWEEN 56 AND 60, '56-60 Pre-Retirement/Late
Career', 'Other')
    )
)
);

```

Gender VS Diff Policies under diff Age_Range

```

SELECT
    Gender,
    IIF(Age BETWEEN 18 AND 25, '18-25',
        IIF(Age BETWEEN 26 AND 40, '26-40',
            IIF(Age BETWEEN 41 AND 55, '41-55',
                IIF(Age BETWEEN 56 AND 60, '56-60',
                    IIF(Age > 60, 'Above 60', 'Other')
                )
            )
        )
    ) AS age_range,
    COUNT(CustomerID) AS total_customers,
    COUNT(IIF(HealthID IS NOT NULL, customerID, NULL)) AS health_policies,
    COUNT(IIF(MotorID IS NOT NULL, customerID, NULL)) AS motor_policies,
    COUNT(IIF(TravelID IS NOT NULL, customerID, NULL)) AS travel_policies
FROM
    ABT_clean
GROUP BY
    Gender,
    IIF(age BETWEEN 18 AND 25, '18-25',
        IIF(age BETWEEN 26 AND 40, '26-40',
            IIF(age BETWEEN 41 AND 55, '41-55',
                IIF(age BETWEEN 56 AND 60, '56-60',
                    IIF(age > 60, 'Above 60', 'Other')
                )
            )
        )
    )
);

```

R Code:

```
#Import required libraries
library(tidyverse)
library(dplyr)

#get the working directory
getwd()

#set the working directory
setwd("/Users/dhanush/Desktop/Bussiness analytics /Data management ")

#Give alias to tables

data1<- Data_1_Customer
data2<- Data_2_Motor_Policies
data3<- Data_3_Health_Policies
data4<- Data_4_Travel_Policies

#Join tables using dplyr
ABT <- data1 %>%
  left_join(data2,by = "MotorID") %>%
  left_join(data3,by ="HealthID") %>%
  left_join(data4,by="TravelID")

#data quality analysis
#summarising the data
summary(ABT)
str(ABT)

#Fiding out data quality issues and outlier
count(ABT, CustomerID)
count(ABT, Title) # Mr and Mr. are one and the same and Mr must be renaned to Mr.
count(ABT,GivenName)
count(ABT,MiddleInitial)
count(ABT, Surname)
count(ABT, CardType)# There is a outlier in cardtype which says 0
count(ABT, Occupation)
count(ABT, Gender)#There is an outiler with f , m which must be changed to male and
female
count(ABT, Age)

#ggplot age
ggplot(ABT)+
  geom_boxplot(aes(x=Age),outlier.color = "red",outlier.size=2)#we can find that there are 3
outlier in data
```

```
summary(ABT$Age)#summarising age
```

```
count(ABT, Location)
count(ABT, MotorID)
count(ABT, HealthID)
count(ABT, TravelID)
count(ABT, MotorType)
count(ABT, veh_value)
count(ABT, Exposure)
count(ABT, clm)
count(ABT, Numclaims)
count(ABT, v_body)
count(ABT, v_ae)
count(ABT, HealthType)
count(ABT, HealthDependentsAdults)
count(ABT, DependentsKids)#there is an outlier 40
```

```
#ggplot of DependentKids
ggplot(ABT)+
  geom_boxplot(aes(x=DependentsKids), outlier.color = "blue", outlier.size=2)#we can find
that there are 3 outliers in data
```

```
summary(ABT$DependentsKids)
```

```
count(ABT, TravelType)
count(ABT, ComChannel)#there are outliers for email, phone, sms like e, p and s which must
be recoded as email, phone and sms
```

```
is.na(ABT)#finding out any null values
```

```
count(ABT, ComChannel)
```

```
#Addressing data quality issues mentioned above
```

```
ABT_clean <- ABT %>%
  mutate(Title = ifelse(Title == "Mr", "Mr.", Title)) %>%
  mutate(CardType = ifelse(CardType == 0, "Not_Given", CardType)) %>%
  mutate(Gender = ifelse(Gender == "f", "female", Gender)) %>%
  mutate(Gender = ifelse(Gender == "m", "male", Gender)) %>%
  mutate(ComChannel = ifelse(ComChannel == "E", "Email", ComChannel)) %>%
  mutate(ComChannel = ifelse(ComChannel == "P", "Phone", ComChannel)) %>%
  mutate(ComChannel = ifelse(ComChannel == "S", "SMS", ComChannel)) %>%
  filter(!(Age < 18 | Age > 85))
```

```
#replace value "40" DependentKids with mean 2
```

```
ABT_clean$DependentsKids <- replace(ABT_clean$DependentsKids,  
ABT_clean$DependentsKids>4,2)
```

```
#ggplot age after cleaning
```

```
ggplot(ABT_clean)+  
  geom_boxplot(aes(x=Age),outlier.color = "red",outlier.size=2)
```

```
#ggplot Dependentskids after cleaning
```

```
ggplot(ABT_clean)+  
  geom_boxplot(aes(x=DependentsKids),outlier.color = "blue",outlier.size=2)
```

```
#Exporting clean ABT data
```

```
install.packages("openxlsx")
```

```
library(openxlsx)
```

```
write.xlsx(ABT_clean,"ABT_clean.xlsx")
```