Description and explanation of the POLICY data and the CLAIMS data’s columns and variables

The **policy** data it is a data that contains 227469 observations and 21 variables:

1. CUST_ID
2. EXECUTIVE
3. BODY
4. MAKE
5. MODEL
6. USE_OF_VEHICLE
7. MODEL_YEAR
8. CHASSIS_NO
9. REGN
10. POLICY_NO
11. POL_EFF_DATE
12. POL_EXPIRY_DATE
13. SUM INSURED
14. POL_ISSUE_DATE
15. PREMIUM2
16. DRV_DOB
17. DRV_DLI
18. VEH_SEATS
19. PRODUCT
20. POLICYTYPE
21. NATIONALITY

P.S: we have 5% of the rows in the policy data that have 50% and more of missing values

The **claims** data it is a data that contains 62521 observations and 24 variables:

1. Account Code
2. "DATE OF INTIMATION"
3. "DATE OF ACCIDENT"
4. "PLACE OF LOSS"
5. CLAIM NO
6. AGE
7. TYPE
8. "DRIVING LICENSE ISSUE"
9. BODY TYPE
10. MAKE
11. MODEL
12. YEAR
13. CHASIS NO
14. REG
15. SUM INSURED
16. POLICY NO
17. POLICY START
18. POLICY END
19. "INTIMATEDAMOUNT"
20. "INTIMATEDSF"
21. EXECUTIVE
22. PRODUCT
23. POLICYTYPE
24. NATIONALITY

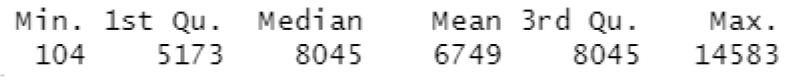
P.S: we have 0 rows from the claims data that have 50% and more of missing values

POLICY DATA

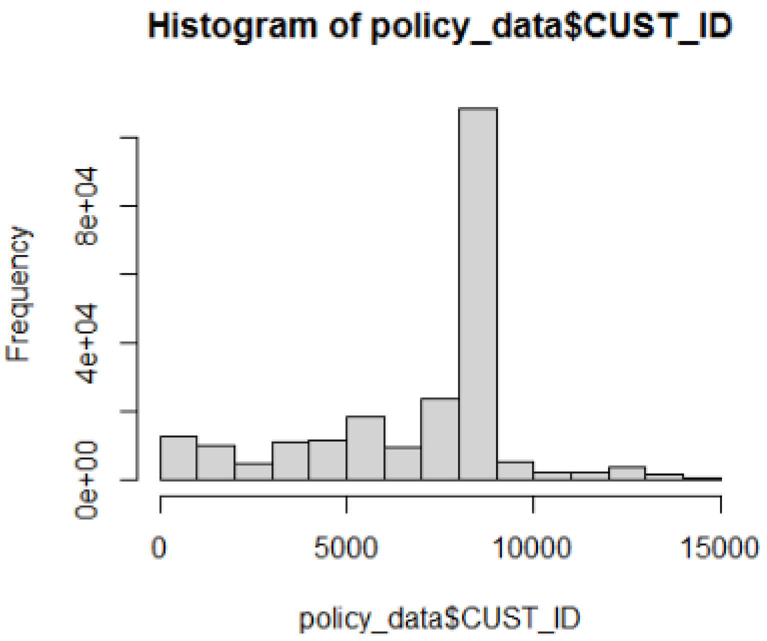
POLICY DATA VARIABLES:

1. CUST_ID:

- It is a quantitative variable that represents the unique value to identify a customer who bought car insurance.
- It is ranging from 104 to 14583



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    104    5173    8045    6749    8045   14583
```



Histogram of policy_data$CUST_ID

2. EXECUTIVE:

- It is a qualitative variable that represents the agent's name who sold the policy to the customer
- It contains 110 agent's name
- The agent BR has the highest sales of insurance product and JOHN and MANOHAR also in our policy data

3. BODY:

- It is a qualitative variable that represents the type of the body of the vehicle (if it is a Bus or Jeep or Truck or...) to be insured
- It contains 90 types of the body of the vehicle
- It contains 10274 blanks and 7 "-" values and 8 "." Values
- The SALOON and 4WD and SUV types of the body of the vehicles have the highest frequency in our policy data

4. MAKE:

- It is a qualitative variable that represents the name of the Maker or Manufacturing company of the insured vehicle
- It contains 240 maker companies of the vehicle
- It contains 10306 blanks
- The HONDA and HYUNDAI and FORD and CHEVROLET and BMW maker companies have the highest frequency in our policy data

5. MODEL:

- It is a qualitative variable that represents the Vehicle Model or Brand or Name by which the manufacturer sells everywhere
- It contains 2207 vehicle models
- It contains 10517 blanks and 11 "-" values and 679 ". " Values and 7 “1” value and 51 “09/03/2022” value, or I think these values are a typo
- And also it contains 2008 and 2015 and 2000 and 2002 values… which are typo values
- The 180 CC and 150 CC and 220 CC vehicle name have the highest frequency in our policy data

6. USE_OF_VEHICLE:

- It is a qualitative variable that indicates either the vehicle is used for commercial or private purpose
- It contains 12 possibilities (PRIVATE, COMMERCIAL, SCHOOL BUS, PRIVATE/COMMERCIAL, TOUR OPERATION, SPORTS CAR, 4 WD STATION, PICK UP DOUBLE CABIN, RENTACAR, CONSTRUCTION EQUIP, FORK LIFT, DELIVERY VEHICLES) of usage of the vehicle
- It contains 10303 blanks
- The PRIVATE and the COMMERCIAL usage of vehicle have the highest frequency in our policy data

7. MODEL_YEAR:

- It is a quantitative variable that indicates the Year of manufacturing that tells the age of the vehicle
- It is ranging from 2 to 9999 years, which is illogic that's why I think it should be ranging from 1944 year till 2021 year, and we can do this because we can see that all the manufacturing year smaller than 1944 or greater 2021 have a low frequency in our policy data
- The 2015, 2008, 2016, 2014, 2007, 2013, 2009, 2006, 2012 years of manufacturing of the vehicle have the highest frequency in our policy data
- We can also note that we have 1296 number of vehicles that the manufacture year of them is 1996
- We can also note that we have 404 number of vehicles that the manufacture year of them is 1995
- We can also note that we have 341 number of vehicle that the manufacture year of them is 1994

8. CHASSIS_NO:

- It is a qualitative variable that represents the unique number pertaining to a particular vehicle
- It contains 176166 different numbers
- It contains 10274 blanks and 2 "…." values and 2 "." Values and 4 zeros values and 1 "#NAME" value and 1 "#REF" value and 1 "10-Dec" value

9. REGN:

- It is a qualitative variable that represents the place of registration of the policy
- It contains 53 regions
- It contains 10277 blanks and 1 "40298" value
- DUBAI, SHJ(Sharjah is the third largest emirate in the UAE), AD, AJMAN, UAQ, RAK regions are the highest place of registration in our policy data
- We can see that we have a region ”A D” i think this space is a typo and this can be corrected to ”AD”

10. POLICY_NO:

- It is a quantitative variable that represents the unique number related to the insurance policy that customer bought for her car/vehicle
- It contains 227461 different number, or this should be 227469 different number because this represent the unique number related to the insurance policy

11. POL_EFF_DATE:

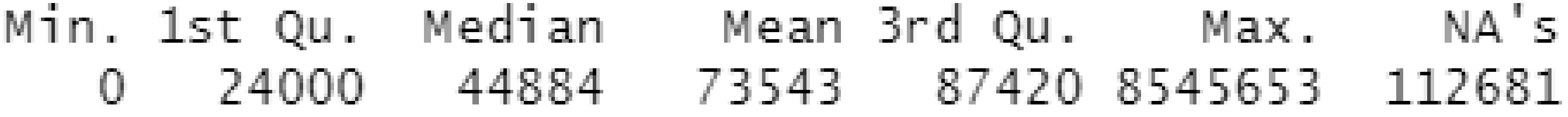
- This is the date from where insurance policy start for the vehicle of the customer, in other words it is the date the policy is activated.
- It is a date variable that is going from date 18/6/2002 till 26/3/2021
- It contains 2347 different effective dates

12. POL_EXPIRY_DATE:

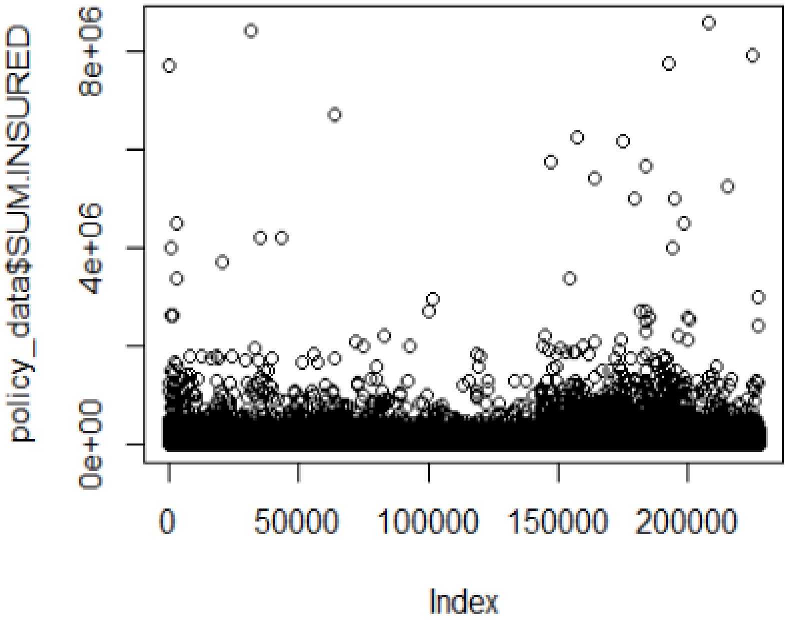
- This is the date till then insurance policy ends/remain in force for the vehicle of the customer
- It is a date variable that is going from date 19/3/2015 till 25/4/2022
- It contains 2335 different expiry dates

13. SUM INSURED:

- It is a quantitative variable that represents the total sum assured under the policy for a particular vehicle
- It is ranging from 0 to 8545653
- It contains 112681 blanks and 7797 zero values
- We can see that 25% of the policies have a sum insured <= 24 000
- We can see that 50% of the policies have a sum insured <= 44 884
- We can see that 75% of the policies have a sum insured <= 87 420
- The 30 000 sum insured and 25 000 sum insured and 20 000 sum insured have the highest frequency in our policy data



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0   24000   44884   73543   87420 8545653  112681
```



14. POL_ISSUE_DATE:

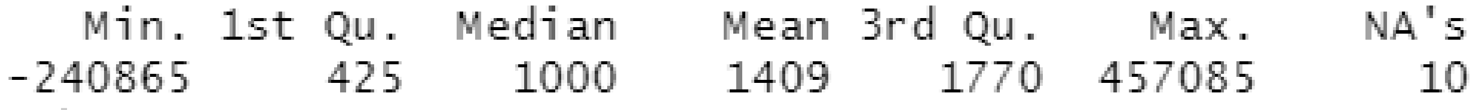
- This variable represents the date when a particular policy was issue to the customer on her vehicle
- Date of issue simply refers to the date the insurer created the contract (the insurance policy), which isn't necessarily when the coverage starts.
- The issue date can be on or before the effective date, but never after because the effective date is the day that the coverage actually begins.
- It is a date variable that is going from date 4/1/2015 till 31/12/2020
- We should pay attention that this variable contain the precise hour of the day that the policy was issued at it

P.S: we have 5.68% of policies that have an issue date after the effective date.

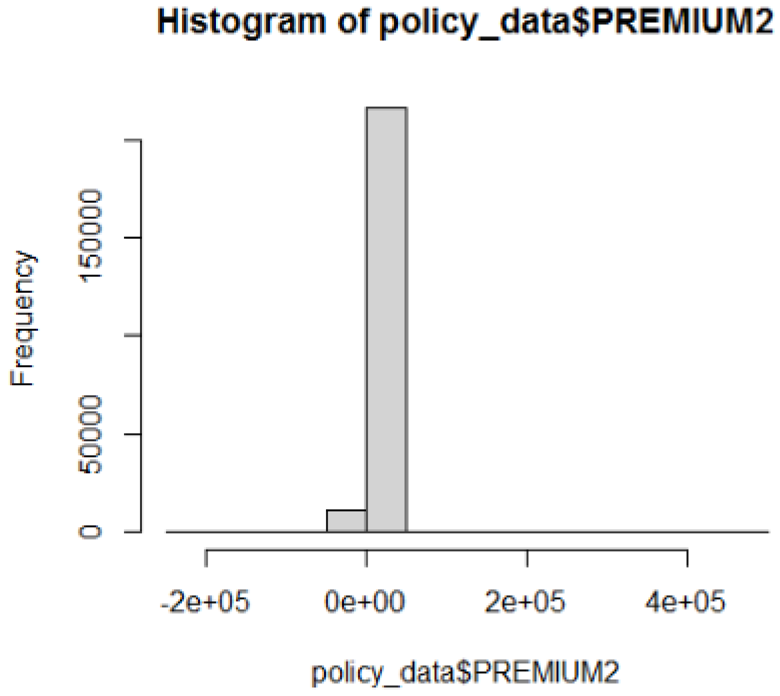
P.S: All the expiry dates of all the policies are after the effective date, which is good.

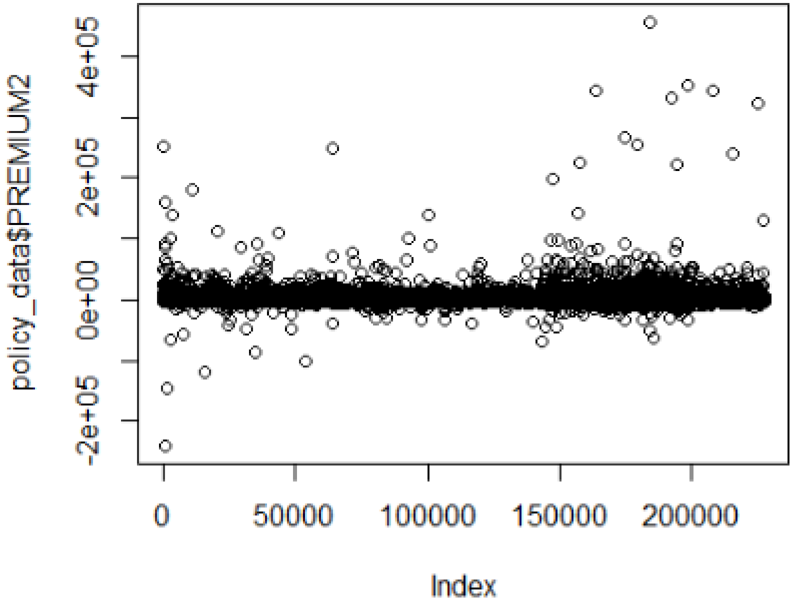
15. PREMIUM2:

- It is a quantitative variable that represents the premium paid by the customer for the Insurance policy on her vehicle
- It is ranging from -240865 to 457085
- We can note that we have a lot of negative values for the premium amount which is illogic!!!
- It contains 10 blanks and 741 zero values
- We can see that 25% of the premiums amount of our policies are <= 425
- We can see that 50% of the premiums amount of our policies are <= 1000
- We can see that 75% of the premiums amount of our policies are <= 1770



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
-240865     425    1000    1409    1770  457085      10 
```





16. DRV_DOB:

- This is a variable that represents the date of the birth of the driver (if the vehicle is driven by a driver)
- It is a date variable that is going from year 1/1/1900 till 30/1/9999
- It contains 14266 different dates of birth
- We can see that we have something illogic because the driver date of birth can't be in 1900 year (because he should be died specially that this data was uploaded 1 year ago) and moreover we have a driver date of birth in 9999 and 2021 and 2020 and 2024 ... which is illogic because a child can't be driving and because 9999 year doesn't exist yet

17. DRV_DLI:

- This is a variable that represents the date of the driving license of the driver (if the vehicle is driven by a driver)
- It is a date variable that is going from year 2/4/1900 till 31/1/9997
- It contains 17463 different dates
- We can see that we have something illogic because the license of the driver can't be in 1900 year and moreover we have a date in 9997 and 2024 and 6300 ... which is illogic

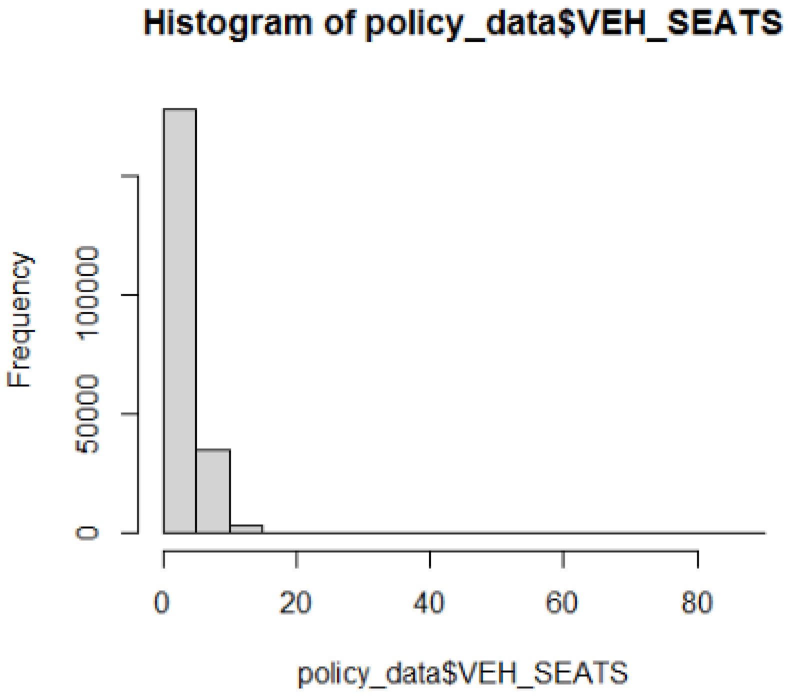
P.S: There is something illogic in the description of the policy data of Mr. Sumit:

- Because how the date of birth of the driver (DRV_DOB) can be the same date as the driving license (DRV_DLI)? (such as observation number 101429 in the policy data)
- And also because how can the date of birth of someone (DRV_DOB) greater than the date of the driving license (DRV_DLI)? (such as observations number: 10148, 10145, 101411, 101554, 180107 in the policy data)

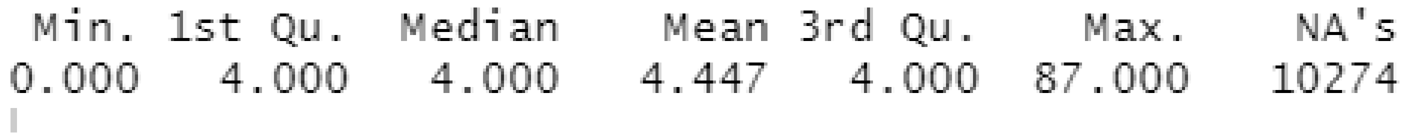
So in my opinion I think that these 2 columns should be split by their meaning.

If DRV_DLI is the date of the driving license of the driver and DRV_DOB is the date of birth of the driver, we have 86.33% policies in our data that have a date of birth after the date of the driving license which is illogic!!!
But if we consider DRV_DOB is the date of the driving license of the driver and DRV_DLI is the date of birth of the driver, we have only 3.76% policies in our data that have a date of birth after the date of the driving license!!!

18. VEH_SEATS:



- It is a quantitative variable that represents the Number of seats in the vehicle
- It is ranging from 0 to 87
- It contains 65 different numbers of seats
- It contains 10274 blanks and 2071 zero values
- We can see that 25% of the vehicles of our policies have a seats number <= 4
- We can see that 50% of the vehicles of our policies have a seats number <= 4
- We can see that 75% of the vehicles of our policies have a seats number <= 4
- The vehicle with 4 seats has the highest frequency in our policy data than after it we have the vehicle with 6 seats
- We have 1257 vehicles with 14 seats in our policy data



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
0.000   4.000   4.000   4.447   4.000  87.000   10274 
```

19. PRODUCT:

- It is a qualitative variable that contains 10 different types of products (STANDARD, NOT CLASSIFIED, M 2.5, M 2019, M 2.25, RENT A CAR, LUXURY, PASSENGER, TOURISM, TP)
- The TP and STANDARD and NOT CLASSIFIED product type have the highest frequency in our policy data

<u>20. POLICYTYPE:</u>

- It is a qualitative variable that represent the type of guaranty (or coverage) in the insurance policy (OD: Own Damage, TP: Third Party, COMP : Comprehensive etc.)
- It contains 2 different types of policies: COMP with a 49.58742% and TP with a 50.41258%

<u>21. NATIONALITY:</u>

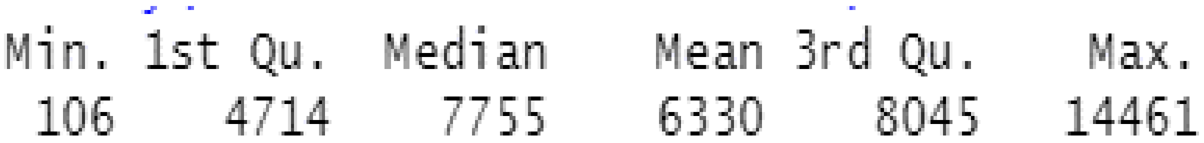
- It is a qualitative variable that represent the nationality of the agent or executive who sold the vehicle insurance policy
- It contains 165 different agents nationality
- It contains 169349 blanks
- We can see that the most agent nationality are from the company itself and from INDIAN and from EMIRATE

CLAIMS DATA

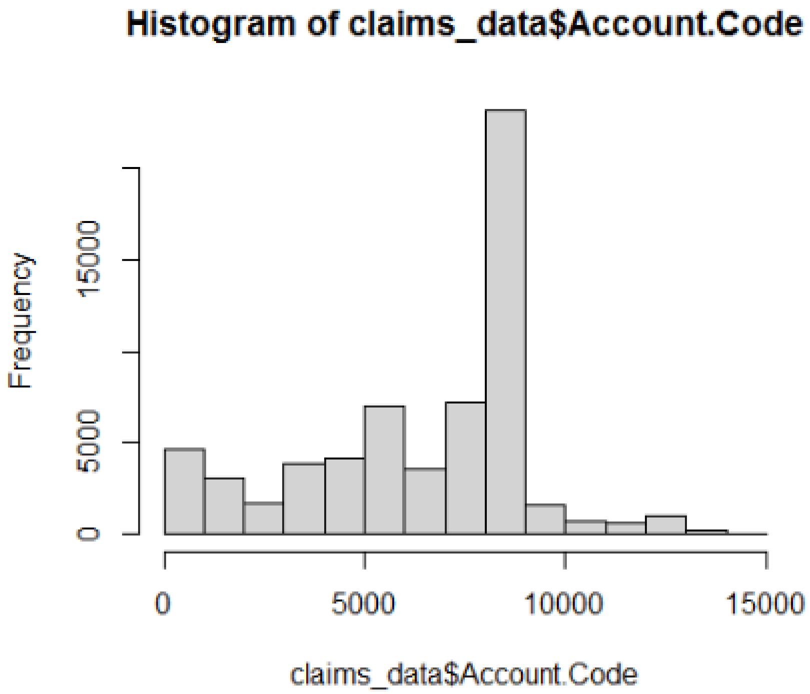
CLAIMS DATA VARIABLES:

1. Account Code:

- it is a quantitative variable that represents the unique value to identify a customer who bought car insurance.
- I think this is the same variable as CUST_ID in the policy data but I’m not sure because we have 72 Account Code that are not existing in the CUST_ID variable in the policy data
- It is ranging from 106 till 14461



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  106    4714    7755    6330    8045   14461
```



2. DATE OF INTIMATION:

- It is the date on which the insured inform the insurer that an accident is occurs
- It is a date variable that is going from date 4/1/2015 till 31/12/2020
- It contains 60223 different date of intimation
- We should pay attention that this variable contain the precise hour of the day that the insurer was inform at it

P.S: we can note that we have 0.06% rows of our claims that have a date of accident after the date of the policy end and these rows contains an intimated amount, which is illogic!!

3. DATE OF ACCIDENT:

- It is the date of the occurrence of the accident
- It is a date variable that is going from date 9/8/2010 till 30/12/2020
- It contains 2493 different date of accident

4. PLACE OF LOSS:

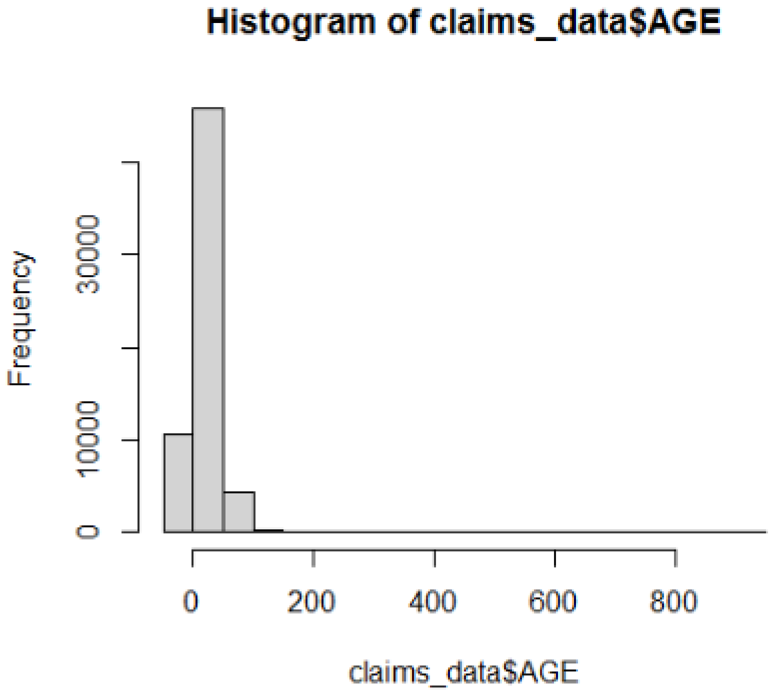
- It is a qualitative variable that represents the place where the accident was occurred
- It contains 15 regions
- DUBAI,SHARJAH, Dubai, ABU DHABI, AJMAN, Sharjah regions are the highest place of loss in our claims data
- I think that DUBAI and Dubai is the same region and also i think that SHARJAH and Sharjah is the same region (same fo Ajman and AJMAN and for Abu Dhabi and ABU DHABI) so i think we should consider them same place of losses

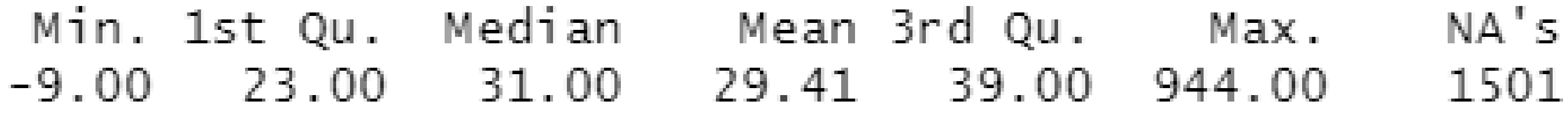
5. CLAIM NO:

- It is a qualitative variable that represents the unique number pertaining to the claim
- It contains 61972 different numbers
- It contains 42 blanks

6. AGE:

- This is a quantitative variable that represents the age of the driver that have a claim
- It is going from -9 till 944
- It contains 94 different ages
- It contains 10505 of zero values
- It contains 1501 Blanks
- We can see that we have something illogic because the age can't be 944 or 220 year (because he should be died) and moreover we have a driver age of 220...
- We can see that 25% of the drivers that are doing claims have an age <= 23
- We can see that 50% of the drivers that are doing claims have an age <= 31
- We can see that 75% of the drivers that are doing claims have an age <= 39
- We can also see that we have ages of: -9, -8, -6, -5, -2, -1, 0, 1, 3, 5 which is illogic
- We can also see that the age of 30 have the highest frequency that’s mean that the most of the claims of our data are caused by driver with an age of 30 years





```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 -9.00   23.00   31.00   29.41   39.00  944.00    1501 
```

7. TYPE:

- It is a qualitative variable that represents the type of the claim that occurs for ex if it is Own Damage claims or Third Party claims or…
- It contains 17 different types of claims
- It contains 36 Blanks
- We can also remark that we have 2 types of claims: “Loss Of Revenue SHJ taxi” and “Loss of Revenue(SHJ Taxi)”, that I think we should consider it the same type of claims. And same remarks we can talk about it for the 2 types of claims: “TP CLAIM” and “TP Claim”.
- We can also remark that the TP Claim and RECOVERY CLAIM and OD Claim represent the major types of claims that has occurs in our claims data (because they have the higher frequency)

8. DRIVING LICENSE ISSUE:

- It is a date variable that represent the issue date of the driving license of the driver that has occur a claim
- It is going from 1/1/1900 till 31/1/9970
- It contains 744 Blanks
- We can remarks that the License who has an issue date in the year 2021 till year 9970 have a low frequency and I think we should remove them because this data was uploaded 1 year ago so it can not contains a date issue for driving license after 2020
- We can also remark that the date issue 1/1/1900 of the driving license has a high frequency of 49.11%

P.S: There is something illogic in the DRIVING LICENSE ISSUE date and AGE of the driver, we have 49.56% (most of these values are for driving license issue year 1900) of the claim data that have a DRIVING LICENSE ISSUE before they were born because for ex: for the observation number 57 we have an age of 30 for the driver and the date of the driving license issue is in year 1900 which is illogic because in year 1900 the person was not yet born!!! Same for the observations numbers 64,65,66…

P.S: **There is something confusing in the DRIVING LICENSE ISSUE, the DRIVING LICENSE ISSUE date should be the same as the DRV_DLI in the policy data (because DRV_DLI means Driving License Issue). But we have 52.65% of the components of the DRIVING LICENSE ISSUE variable that is not present in the DRV_DLI component's variable.**
But as we said before in slide 10 if we split the meaning of DRV_DOB and DRV_DLI and then we compare the components of the DRIVING LICENSE ISSUE in the claims data and the components of the DRV_DOB in the policy data we can see that 4.57% of the components of the DRIVING LICENSE ISSUE variable that is not present in the DRV_DOB component's variable.
So I think DRV_DOB and DRV_DLI meaning should be split.

9. BODY TYPE:

- It is a qualitative variable that represents the type of the body vehicle, for ex: if it is a truck or bus or motorcycle or…
- I think that this variable is same as the BODY variable in the policy data
- It contains 65 different types of vehicle
- It contains 247 Blanks
- The SALOON and 4WD and SUV types of the body of the vehicles have the highest frequencies in our claims data

10. MAKE:

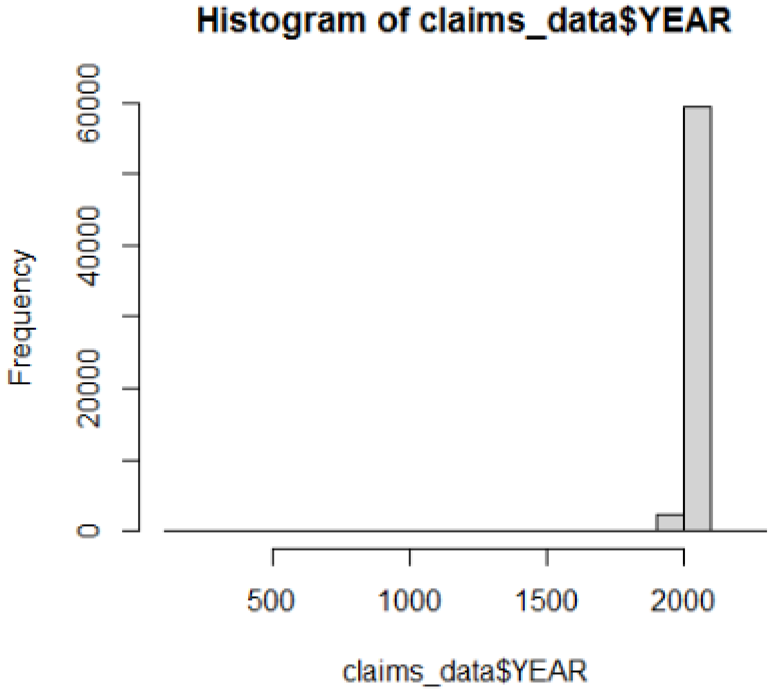
- It is a qualitative variable that represents the Maker or Manufacturing company of the insured vehicle
- I think that this variable is same as the MAKE variable in the policy data
- It contains 144 different manufacturing companies
- It contains 314 "." values
- We can remark that the TOYOTA, NISSAN, MITSUBISHI, HONDA are the manufacturing companies of the vehicles that have the highest frequencies in our claims data, so we can conclude that most of the claims that occurred in our data are caused for the TOYOTA's company vehicle

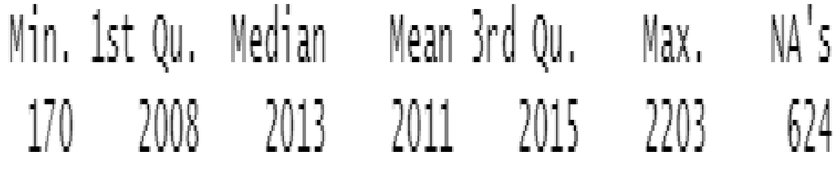
11. MODEL:

- It is a qualitative variable represent the Vehicle Model or Brand or Name by which the manufacturer sells everywhere
- I think that this variable is the same as the MODEL variable in the policy data
- It contains 633 Blanks
- It contains 225 "." values
- It contains 15 "09/03/2022" values which should be removed for sure because they don't represent a brand of vehicle, that I think that are a typo
- It contains 1 "1" value that I think that is a typo
- And we can also remarks that it contains numbers values such as 2000, 2008, 1840, 1831, 159, 156, 1626… that I think that are a typo
- We can also remarks that the COROLLA, CAMRY, YARIS, SUNNY vehicle brands that that have the highest frequencies in our claims data, so we can conclude that most of the claims that occurred in our data are caused for the COROLLA's vehicle

12. YEAR:

- It is a quantitative variable that indicates the Year of manufacturing that tells the age of the vehicle
- I think it is the same variable as the Model_YEAR variable in the policy data, but we remark that in the MODEL_YEAR variable we don't have 1900 year but in the YEAR variable of the claims data we have a year of 1900 which is illogic if we are saying that these 2 columns are the same
- It is ranging from 170 to 2203 years, which is illogic that's why I think it should be ranging from 1900 year till 2021 year, and we can do this because we can see that all the years: 170,2095, 2203 have a small frequency in our claims data and I think that this is a typo
- It contains 624 Blanks
- We can see that 25% of the policies that realize a claim have a manufacturing year <= 2008
- We can see that 50% of the policies that realize a claim have a manufacturing year <= 2013
- We can see that 75% of the policies that realize a claim have a manufacturing year <= 2015
- The 2015, 2016, 2014, 2013, 2008, 2009, 2012, 2017, 2007 years of manufacturing of the vehicle have the highest frequency in our claims data
- The 1991, 1990, 1900, 2203, 170, 1980, 1988, 1989, 2095 years of manufacturing of the vehicle have the smallest frequency in our claims data





```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 170    2008    2013    2011    2015    2203     624
```

13. CHASIS NO:

- It is a qualitative variable that represents the unique number pertaining to a particular vehicle
- I think this is the same variable as CHASSIS_NO in the policy data, but I’m not sure because we have 717 CHASIS NO in the claims data that are not existing in the CHASISS_NO column in the policy data

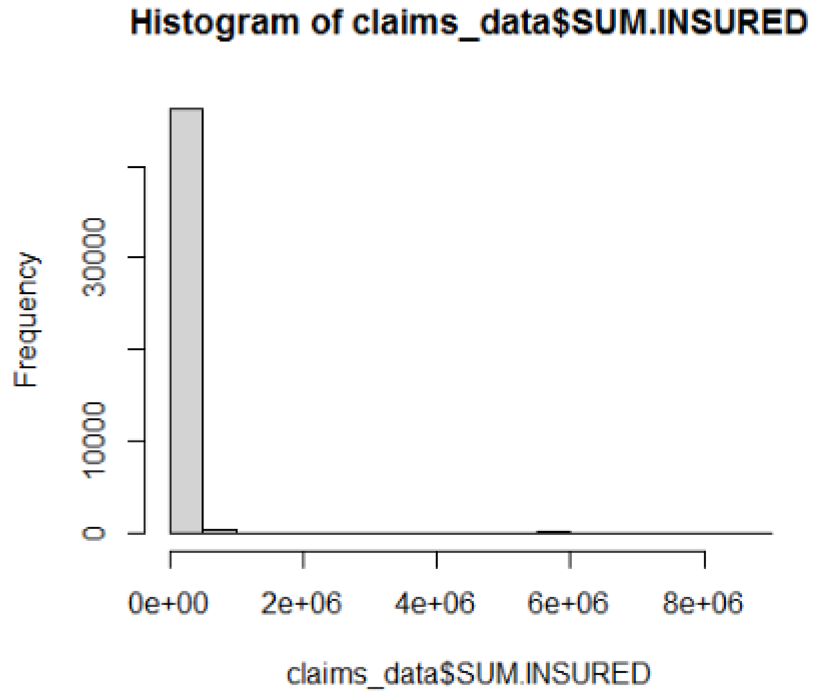
14. REG:

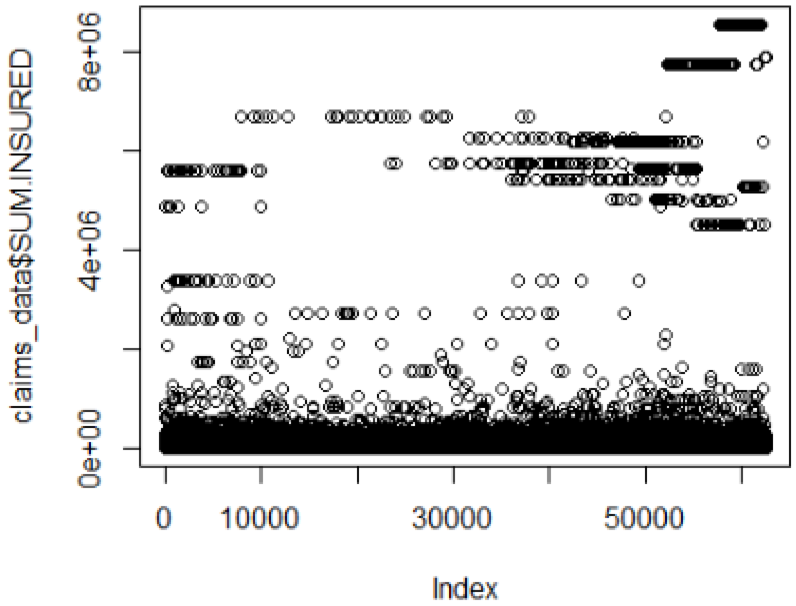
- It is a qualitative variable that represents the place of registration of the policy
- I think that this variable is the same variable as REGN in the policy data
- It contains 34 different regions
- It contains 5 values of “29713”, I think this is a typo
- It contains 1 Blank
- I think that the “SHJ I” it is a typo
- I think that the “DUBAI D” it is also a typo, and I think that this is same as “DUBAI – D”
- The DUBAI, SHJ, AD, AJMAN are the place registration of the policy that have the highest frequencies in our claims data

15. SUM NSURED:

- It is a quantitative variable that represents the total sum assured under the policy for a particular vehicle
- It is ranging from 0 to 8 545 653
- It contains 15055 blanks
- It contains 874 zero values
- I think that the value “1” which have a frequency of 5 it is a typo
- We can see that 25% of the policies that realize a claim have a sum insured <= 29 000
- We can see that 50% of the policies that realize a claim have a sum insured <= 50 500
- We can see that 75% of the policies that realize a claim have a sum insured <= 100 000
- The 30 000 sum insured and 20 000 sum insured have the highest frequency in our claims data, so we note that the most claims are caused for the 30 000 sum insured and 20 000

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
      0   29000   50500  170171  100000 8545653   15055 
```





16. POLICY NO:

- It is a quantitative variable that represents the unique number related to the insurance policy that customer bought for her car/vehicle
- I think this is the same variable as the POLICY_NO variable in the policy data but I'm not sure because We have 41953 POLICY NO that exist in the claims data that are not existing in the POLICY_NO column in the policy data

17. POLICY START:

- I think this is the date from where insurance policy start for the vehicle of the customer
- I think that should be the same as the variable POL_EFF_DATE, or in reality they are not the same because the POL_EFF_DATE is a variable going from year 2014 till 2021 or the POLICY START contains more dates such as year 2009, 2010, 2011, 2012, 2013.
- It is a date variable that is going from date 23/12/2009 till 9/12/2020
- It contains 2561 different policy start dates

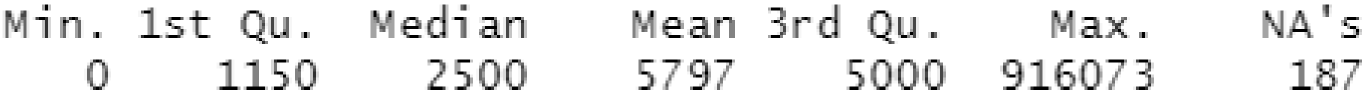
18. POLICY END:

- I think this is the date from where insurance policy end for the vehicle of the customer
- I think that should be the same as the variable POL_EXPIRY_DATE, or in reality they are not the same because we have 389 dates that are existing in the POLICY END column in the claims data but not existing in the POL_EXPIRY_DATE in the policy date, furthermore, the POLICY END contains dates such as year 2010, 2011, 2012, 2013, 2014 that are not present in the POL_EXPIRY_DATE.
- It is a date variable that is going from date 22/12/2010 till 8/1/2022
- It contains 2548 different policy end dates

P.S: all the rows in the claims data have a POLICY START date before the POLICY END date, which is good sign!

19. INTIMATED AMOUNT:

- It is a quantitative variable that represents the amount of the claims that is occur
- It is ranging from 0 to 916 073
- It contains 187 blanks
- It contains 2569 zero values
- We can see that 25% of the claims amounts in our data are <= 1 150
- We can see that 50% of the claims amounts in our data are <= 2 500
- We can see that 75% of the claims amounts in our data are <= 5 000
- The 30 000 sum insured and 20 000 sum insured have the highest frequency in our claims data, so we note that the most claims are caused for the 30 000 sum insured and 20 000

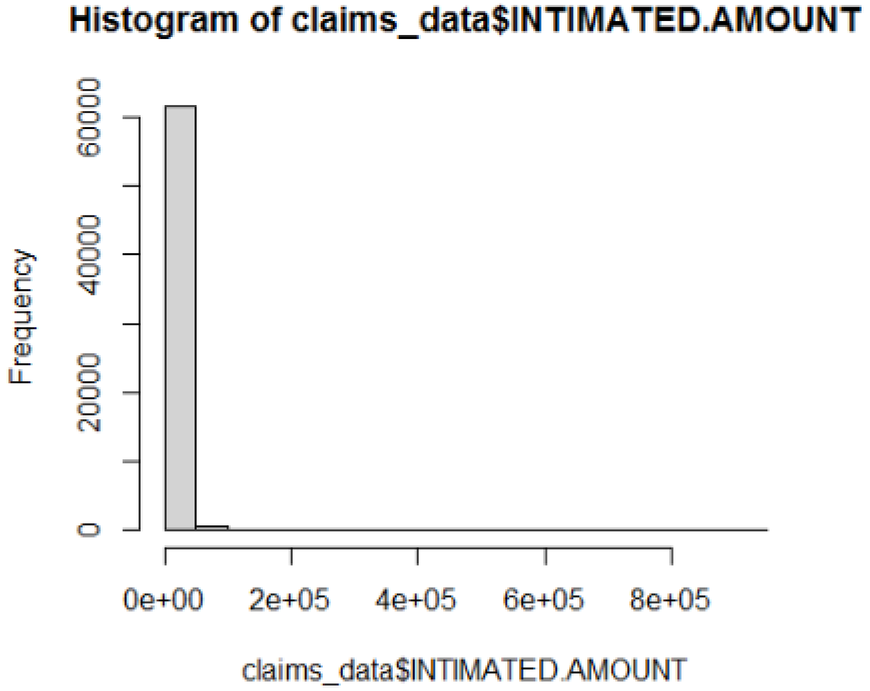


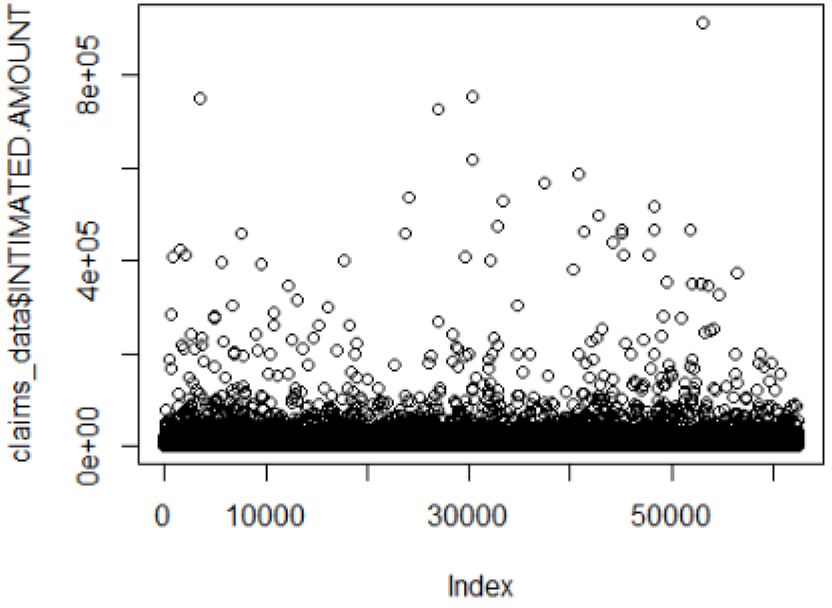
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0    1150    2500    5797    5000  916073     187
```

P.S: for each claim in the claims data the sum insured is <= to the intimated amount of this claim.

But we have 1.34% of our claims data that have a sum insured = 0 or an intimated amount = 0.

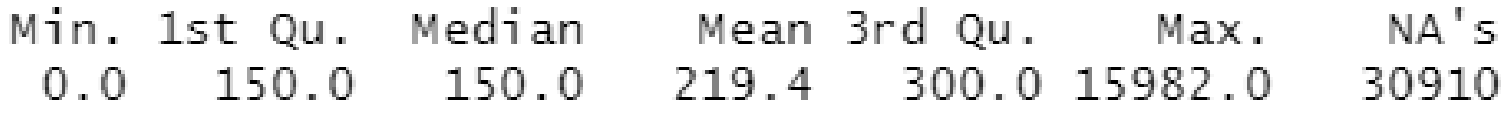
And we also have 0.31% of our claims data that have an intimated amount > sum insured, for ex: the observation number 2110, 3766, 4143, 7695…



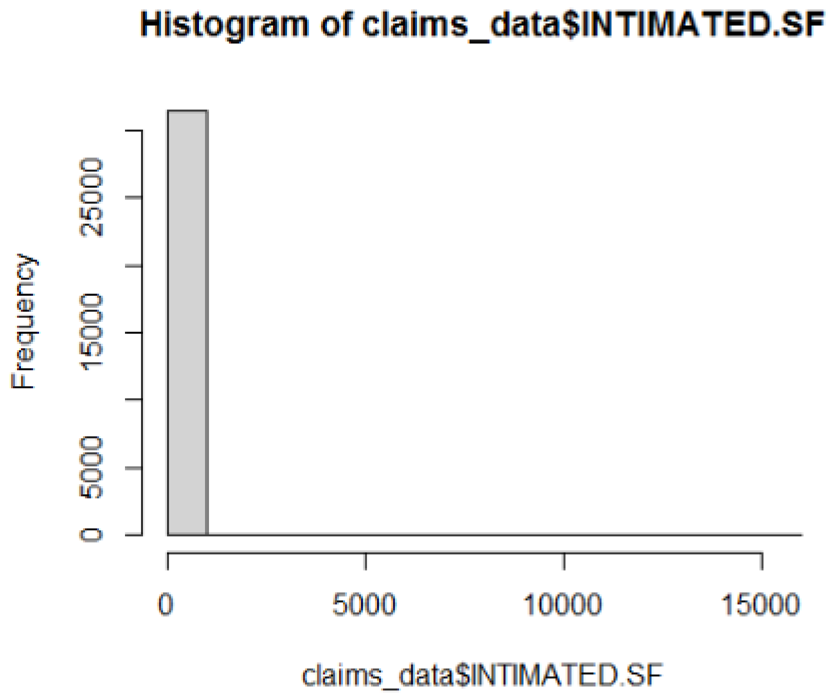


20. INTIMATED SF:

- It is a quantitative variable
- It is ranging from 0 to 15982
- It contains 30910 blanks
- It contains 2901 zero values
- We can see that 25% of the intimated sf in our data are <= 150
- We can see that 50% of the intimated sf in our data are <= 150
- We can see that 75% of the intimated sf in our data are <= 300

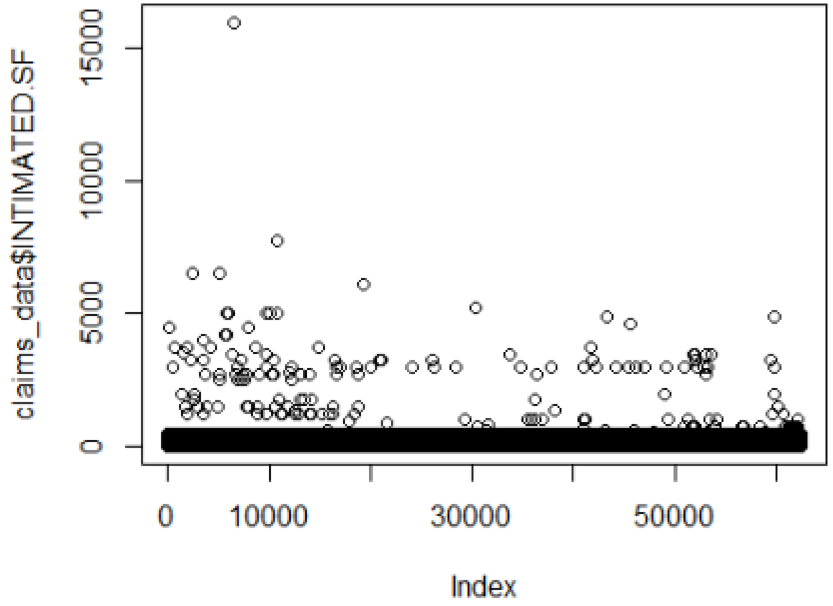


```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.0   150.0   150.0   219.4   300.0 15982.0   30910 
```



P.S: I tried to understand the meaning of this variable but I couldn’t

The INTIMATED SF it is the frequency of the claims?



21. EXECUTIVE:

- It is a qualitative variable that represents the agent's name who sold the policy to the customer who realized a claim
- I think this variable is the same variable EXECUTIVE that exist in our policy data
- It contains 85 agent's name
- The agent BR has the highest sales of insurance product that realize claims and MANOHAR and JOHN also in our claims data

22. PRODUCT:

- It is a qualitative variable that contains 10 different types of products (STANDARD, NOT CLASSIFIED, M 2.5, M 2019, M 2.25, RENT A CAR, LUXURY, PASSENGER, TOURISM, TP)
- I think this variable is the same variable PRODUCT that exist in our policy data
- The STANDARD and TP and M2.5 are the product types that have the highest frequency in our claims data

23. POLICYTYPE:

- It is a qualitative variable that represent the type of guaranty (or coverage) in the insurance policy that occur a claim (OD: Own Damage, TP: Third Party, COMP : Comprehensive etc.)
- I think this is the same variable as the variable of POLICY TYPE in the policy data
- It contains 2 different types of policies and here are the % of each possible type of policy that have realize a claim: COMP with a 74.48217% and TP with a 25.51783%
- So we can remarks that 74.48% of our claims are occurred for the COMP policy type and 25.51% of our claims are occurred for the TP policy type. **Thus, we can conclude that the Comprehensive type policy realize more claims.**

24. NATIONALITY:

- It is a qualitative variable that represent the nationality of the agent or executive who sold the vehicle insurance policy and that this vehicle insurance policy has realized a claim
- It contains 151 different agents nationality
- It contains 24362 blanks
- We can see that the most agent nationality that sold a vehicle insurance policy who realize a claim are from the INDIAN and then we have the PAKISTANI agent's nationality and then EMIRATE and then from the COMMPANY

The new slides that were added

- we have 5% of the rows in the policy data that have 50% and more of missing values
- we have 0 rows from the claims data that have 50% and more of missing values
- To see what are the missing criteria in each row and if they are important or no and if I can continue using the row without them? We can see a statistic table for the blanks in the policy data in the sheet named “stat_table_policy_blanks” in the excel file named “data_explanation”. And We can see a statistic table for the blanks in the claims data in the sheet named “stat_table_claims_blanks” in the excel file named “data_explanation”.

P.S: “x” in the tables means that the value of this column (or variable) is missed in the row

- The CUST_ID neither the POLICY_NO neither the CHASSIS_NO are unique value so how can I use them to link the policy and the claims datasets? They (CUST_ID, POLICY_NO, CHASSIS_NO) are duplicates in the policy data and each duplicates have some different entries so when I want to merge them how can I merge them if I have 2 same CUST_ID in the policy data? And if the total row is the same for all the CUST_ID, well I can remove duplicates but sometimes the rows that have the same CUST_ID have different entries for the other variables in my data so I cant remove it in order to merge data!!
- I was thinking if for ex the customer can be a company (not an individual person) and all the vehicles of this company are insured here under the same ID of CUST_ID with different vehicles and drivers… that’s why the CUST_ID isn’t unique.
- The CUST_ID (variable in policy data) and Account Code (variable in policy data) should be the same columns between the 2 datasets, so the CUST_ID should be linked to the Account Code in order to merge the 2 datasets, or in the variable Account Code we have 72 components of account code that are not a components in the policy data (specially in the column of the CUST_ID)
- So I think that to know the contract that appears in the claims data refers to which contract in the policy data? We should have for the contract (in the claims data) same Account Code (claim data) as the CUST_ID (policy data) and same CHASIS NO (claim data) as the CHASSIS_NO (policy data) and same POLICY NO (claim data) as the POLICY_NO (policy data) at the same time in order to know the qualify the unique contract.
- But I’m confused how can I merge these 2 data (claims and policy data) base on 3 variables (CUST_ID, POLICY_NO, CHASSIS_NO) at the same time? By Microsoft ACCESS?

- The weird entries are highlighted in the sheet named “policy components” for the weird policy data components and the sheet named “claims components” for the weird claims data components in the excel file named “data_explanation”.
- If DRV_DLI is the date of the driving license of the driver and DRV_DOB is the date of birth of the driver, we have 86.33% policies in our data that have a date of birth after the date of the driving license which is illogic!!! But if we consider DRV_DOB is the date of the driving license of the driver and DRV_DLI is the date of birth of the driver, we have only 3.76% policies in our data that have a date of birth after the date of the driving license!!! So I think that the DRV_DOB should be the date of the driving license issue and the DRV_DLI should be the date of birth of the driver.
- Is the AGE (in claims data) related to the DRV_DOB (in policy data)? I think NO (we can see this in the excel file named “data_explanation” specially in the sheet named “DRV_DOB vs AGE”), I think the AGE is more likely related to the DRV_DLI because as I said before the meaning of the DRV_DOB and DRV_DLI should be split. But still we can see from the sheet named “DRV_DLI vs AGE” in the excel file named “data_explanation” that the AGE is not related to the DRV_DLI also.

Is the DRIVING LICENSE ISSUE variable (claims data) have the same components as the variable DRV_DLI variable (policy data)? I don’t think that the DRIVING LICENSE ISSUE (claim data) is the linked to the DRV_DLI, I think it is more likely related to the DRV_DOB because as I said before the meaning of the DRV_DOB and DRV_DLI should be split. Because if the DRIVING LICENSE ISSUE is linked to the DRV_DLI we have 3755 DRIVING LICENSE ISSUE dates that are present in the claims data and not present in the DRV_DLI (in policy data). But if the DRIVING LICENSE ISSUE is linked to the DRV_DOB we only have 326 DRIVING LICENSE ISSUE dates that are present in the claims data and not present in the DRV_DOB (in policy data). That’s why I think that the DRIVING LICENSE ISSUE is related to the DRV_DOB and not to the DRV_DLI in the policy data.

Is the BODY TYPE variable (claims data) have the same components as the variable BODY variable (policy data)? YES

Is the MAKE variable (claims data) have the same components as the variable MAKE variable (policy data)? Kind of YES, because these 2 components (JINBEI CARGO VAN, JAC HFC) of the MAKE variable (in claims data) are not presents in the components of the MAKE variable (in policy data). The rows in claims data that contain a MODEL=SMART (ACURA) should be replaced by SMART CAR (ACURA).

Is the MODEL variable (claims data) have the same components as the variable MODEL variable (policy data)? Kind of YES, because these 3 components (1046K, 2235, SY6483A3) of the MODEL variable (in claims data) are not presents in the components of the MODEL variable (in policy data). The rows in claims data that have a MODEL=420 I CONVERTIBLE should be replaced by 420 I only, same the rows in claims data that contain a MODEL=E 500 (E 320) should be replaced by E500 (E 320).

Is the YEAR variable (claims data) have the same components as the variable MODEL_YEAR variable (policy data)? YES, but we have 1 component (1900) of the YEAR variable (in claims data) is not presents in the components of the MODEL_YEAR variable (in policy data). But I don't think it is a big deal because these component (1900) has a low frequency of 0.0048% in our claims data, and I think it should be replaced by 2013 because it is a weird entry, and because these contract that have YEAR=1900 is a contract in the policy data with a MODEL_YEAR=2013, that's why I think it should be replaced by 2013.

Is the CHASIS NO variable (claims data) have the same components as the variable CHASSIS_NO variable (policy data)? Kind of YES, because we have 716 CHASIS NO (in claims data) that are not presents in the CHASSIS_NO variable (in policy data).

Is the REG variable (claims data) have the same components as the variable REGN variable (policy data)? YES, but we have 2 weird components (REFER LIST, 29713) in the variable REG (in claims data) that should be replaced respectively by (DUBAI,TBR).

Is the POLICY NO variable (claims data) have the same components as the variable POLICY_NO variable (policy data)? Kind of YES, but we have 1622 POLICY NO (in claims data) that are not presents in the POLICY_NO variable (in policy data).

P.S: most of (not all of them) the rows that have a CHASIS_NO that is present in the claims data but not present in the policy data have also a POLICY_NO that is present in the claims but not present in the policy data

Is the POLICY START variable (claims data) have the same components as the variable POL_EFF_DATE variable (policy data)? Not absolutely YES, because we have 5 POLICY START years (in claims data) (such as year 2009, 2010, 2011, 2012, 2013) that are not presents in the POL_EFF_DATE variable (in policy data).

Is the POLICY END variable (claims data) have the same components as the variable POL_EXP_DATE variable (policy data)? Not absolutely YES, because we have 5 POLICY END years (in claims data) (such as year 2010, 2011, 2012, 2013, 2014) that are not presents in the POL_EXP_DATE variable (in policy data).

P.S: the rows in the claims data that have a policy start date one of these years (2009, 2010, 2011, 2012, 2013) (known as the weird entries) have for sure a policy end date one of these years (2010, 2011, 2012, 2013, 2014, 2015) (known as the weird entries) and all these rows have a POLICY NO that is not present in the column of POLIVY_NO in the policy data and some of them (not all these rows) also have a CHASIS NO that is not present in the column of the CHASSIS_NO in the policy data

Is the EXECUTIVE variable (claims data) have the same components as the variable EXECUTIVE variable (policy data)? YES

Is the PRODUCT variable (claims data) have the same components as the variable PRODUCT variable (policy data)? YES

Is the POLICY TYPE variable (claims data) have the same components as the variable POLICY_TYPE variable (policy data)?
YES

Is the NATIONALITY variable (claims data) have the same components as the variable NATIONALITY variable (policy data)?
Not absolutely YES, because we have 17 NATIONALITY (in claims data) that are not presents in the components of the NATIONALITY variable (in policy data). When I tried to see what are the rows in the claims data that have a Nationality that is not a component of the NATIONALITY variable in the policy data, I remarked that some rows contains the same entries as the policy data's row but the policy data's row have missed value for the nationality columns, so I think we can fill it by the nationality present in the claims data's row. And I also remarked that the other rows are same as the policy data's row but in the claim's row the nationality is for ex ANGOLAN (SWAZI, MALDIVIAN) or in the policy the nationality is TUNISIA (SWEDISH, PHILIPPINE) which is very weird!!!

- The sum insured in the policy data and claim data are related? The answer is in the excel file named "data_explanation" and specially in the sheet named "sum_ins policy vs claim". Furthermore, 2.06% of the components of the sum insured variable in the claims data are not components in the sum insured variable in the policy data.
- Could the intimated amount be a part of the sum insured ? For each claim in the claims data the sum insured is <= to the intimated amount of this claim. And we have 1.34% of our claims that have a sum insured = 0 or an intimated amount = 0. And we also have 0.31% of our claims data that have an intimated amount > sum insured, for ex: the observation number 2110, 3766, 4143, 7695…
- Is the INTIMATED SF variable (claims data) represent the frequency of the claims?
- I make sure that all the rows in the claims data have a POLICY START date before the POLICY END date, which is good sign!
- I make sure that all the expiry dates of all the policies are after the effective date, which is good.

- we have 5.68% of policies that have an issue date after the effective date!! Which is illogic because the issue date can be on or before the effective date, but never after because the effective date is the day that the coverage actually begins. When I tried to see the rows that have an issue date after the effective date, I remarked that most of these rows contains a lot of blanks columns (51.14% blanks in the row) so I think these rows should be removed totally. And for the other one I'm confused what should I do with them?
- We can note that we have 0.06% of the rows of our claims that have a date of accident after the date of the policy end and these rows contains an intimated amount, which is illogic!! How can the insurance cover an accident that occurs after the policy was end? For this idea I tried to see what are the rows that have a date of accident after the policy end date, and I noted that some (not all of them such as the observations number 9577, 12668 in the claims data) of these rows are present in the policy data with a same policy effective (or start) date but with a different policy expiry date between claims data and policy data, so for these rows I think we should put the policy expiry date of the policy data because it is after the date of accident.