Insurance company

The following text is the official description of the data:

The data for this case study comes from the former Swedish insurance company Wasa, and concerns partial casco insurance, for motorcycles this time. The reason for using this rather old data set is confidentiality; more recent data for ongoing business can not be disclosed.

Data Format

A data frame with 64548 observations on the following 9 variables.

-agarald: The owners age, between 0 and 99, a numeric vector

-kon: The owners age, between 0 and 99, a factor with levels K M

-zon: Geographic zone numbered from 1 to 7, in a standard classification of all Swedish parishes, a numeric vector

-mcklass: MC class, a classification by the so called EV ratio, defined as (Engine power in kW x 100) / (Vehicle weight in kg + 75), rounded to the nearest lower integer. The 75 kg represent the average driver weight. The EV ratios are divided into seven classes, a numeric vector

-fordald: Vehicle age, between 0 and 99, a numeric vector

-bonuskl: Bonus class, taking values from 1 to 7. A new driver starts with bonus class 1; for each claim-free year the bonus class is increased by 1. After the first claim the bonus is decreased by 2; the driver can not return to class 7 with less than 6 consecutive claim free years, a numeric vector

-duration: the number of policy years, a numeric vector

-antskad: the number of claims, a numeric vector

-skadkost: the claim cost, a numeric vector

My main objective for this case, is to identify different patterns in the data to explain which groups of clients have a higher risk of having at least one accident.

I will start by loading the data and looking at the different columns to see if they contain any NA.

```
options(warn = -1)
library(insuranceData)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data(dataOhlsson)
df=dataOhlsson
apply(df, 2, function(x) any(is.na(x)))
```

```
##  agarald      kon      zon  mcklass  fordald  bonuskl duration  antskad
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## skadkost
##    FALSE
```

Now that I know that all the registers are complete, I will take a quick look at the data. In this way, I will be able to understand it better and thus to analyze it in a more efficient way.

With this fist look, I have been able to detect 2 things. The first one, is that the variable "bonuskl" explains how good a driver is. According to the description, the drivers start with a bonus level of one and for every claim-free year, the level increases by one, until reaching the maximum level, which is seven.

However, the data frame, contains many registers of clients that, even if they have been using the policy for less than a year, their bonus level is higher than 1. I will keep this is mind so I can deeply analyze it latter.

The last thing that I have seen in the data frame, is that there are insured vehicles that, according to the data, are owned by underaged people. Since it is legal for a underaged person to own property I will analyze the data normally, nonetheless I will be careful when taking conclusions from these registers, since they may just be mistakes.

Now I will look at the format of the data.

```
str(df)
```

```
## 'data.frame':    64548 obs. of  9 variables:
##  $ agarald : int  0 4 5 5 6 9 9 9 10 10 ...
##  $ kon     : Factor w/ 2 levels "K","M": 2 2 1 1 1 1 1 2 2 2 ...
##  $ zon     : int  1 3 3 4 2 3 4 4 2 4 ...
##  $ mcklass : int  4 6 3 1 1 3 3 4 3 2 ...
##  $ fordald : int  12 9 18 25 26 8 6 20 16 17 ...
##  $ bonuskl : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ duration: num  0.175 0 0.455 0.173 0.181 ...
##  $ antskad : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ skadkost: int  0 0 0 0 0 0 0 0 0 0 ...
```

Most of the variables, have an "integer" format. Even if in most of the cases a "factor" format would be better, for now, I will leave it with the original format.

I will continue with the data as integers, because, based on the nature of this data, I presume that I will be regularly using logic operators to extract information from the discrete variables (greater than 1, lower than 7...).

The names of the different features are a little confusing for me, I prefer to change them.

```
df=df %>%
  rename(
    Owner_age = agarald ,
    Geo_zone = zon,
    power_by_weight = mcklass,
    vehicle_age=fordald,
    bonus_of_owner=bonuskl,
    policy_years=duration,
    number_of_claims=antskad,
    claim_cost=skadkost)
```

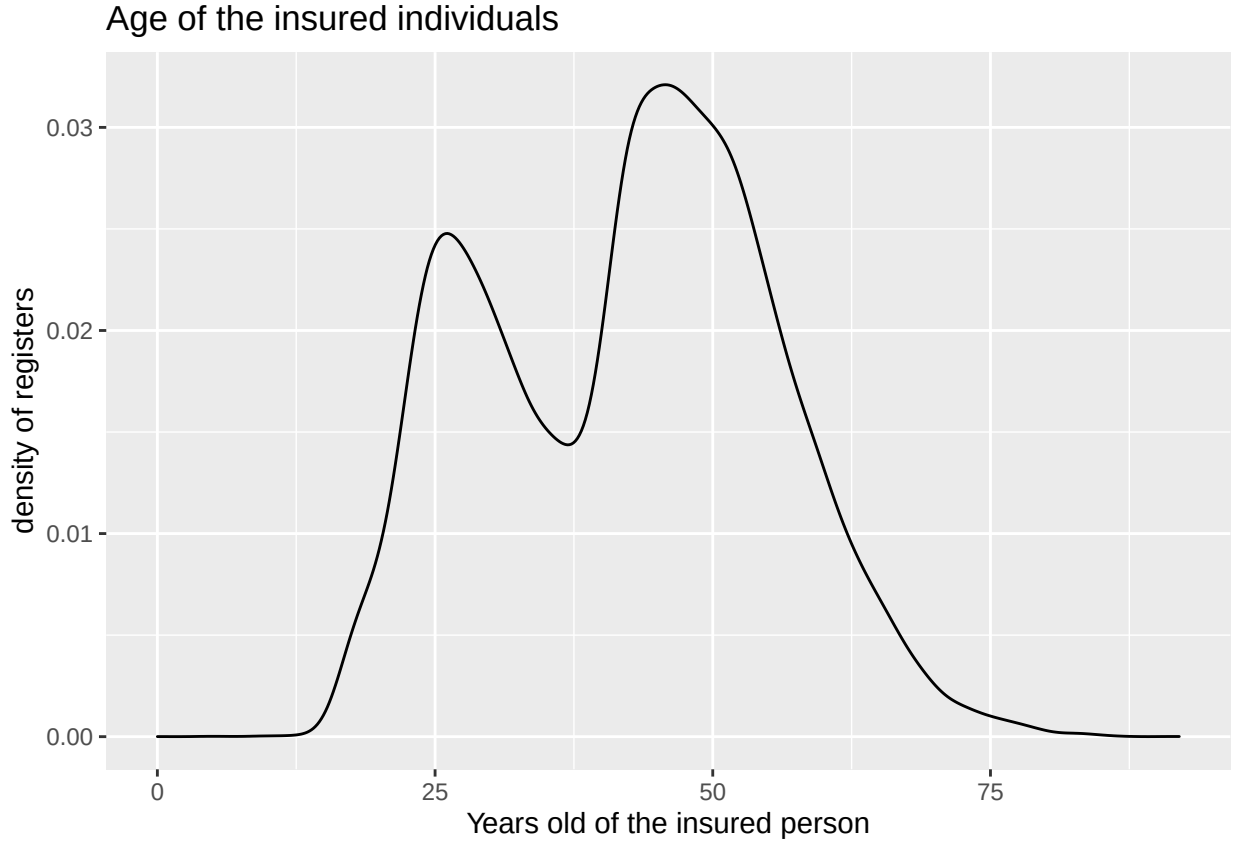
- START OF THE ANALYSIS -

TARGET

I will start the analysis understanding better who the target is of this policy.

The first characteristic that I will consider will be the age of the insured person.

```
ggplot(df)+
  geom_density(aes(Owner_age))+
  labs(x="Years old of the insured person",
       y="density of registers",
       title="Age of the insured individuals")
```



In this graph, I can see that the business has two main targets for this policy.

The first target is the "young drivers", from 22 to 27 years old. The second one, are the more experienced drivers, from 45 to 50 years old.

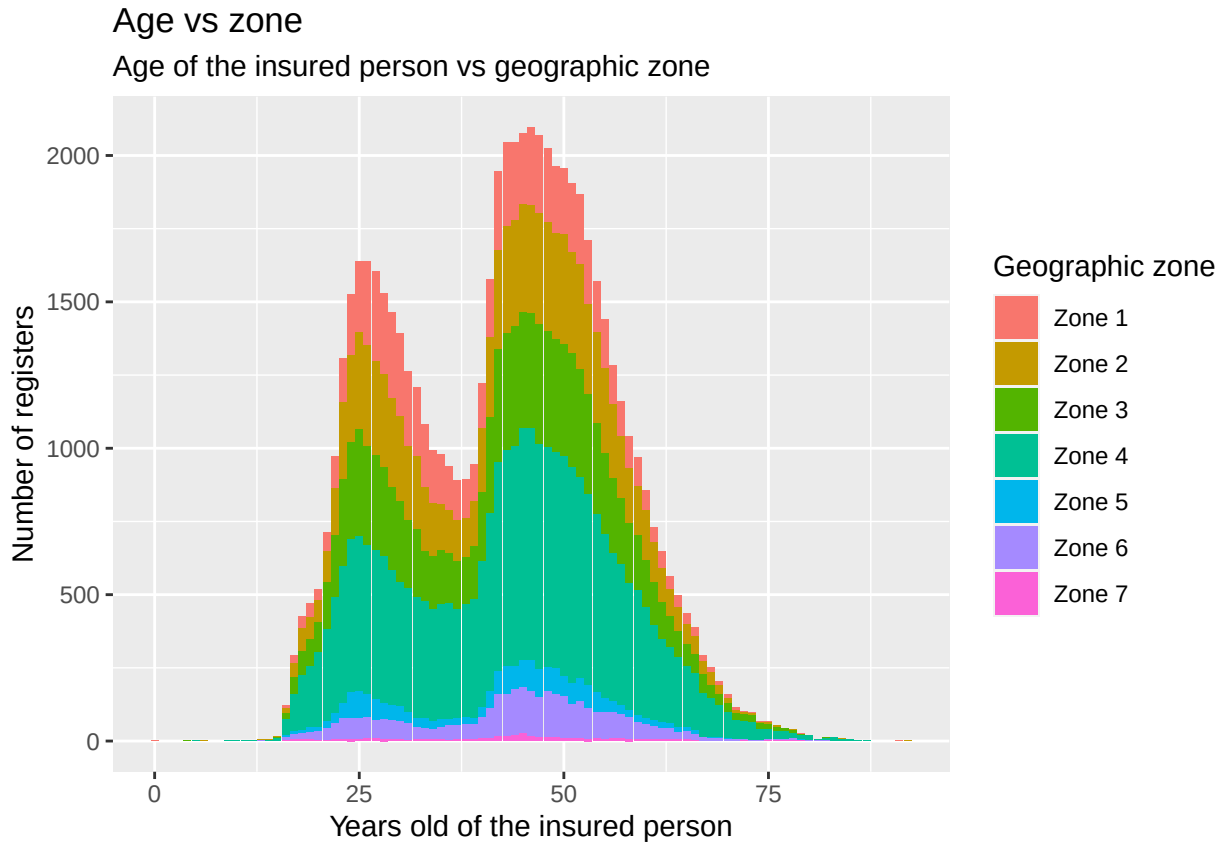
An other interesting information that it is possible to obtain from this plot, is that from these two targets, the business has a clear preference towards the more mature one.

The "young drivers" represent almost 25% of the total number of insured individuals, and the more experienced drivers, represent more than 30%. Normally, to define a target, a business also considers the geographic position of the consumer. That is why, I would like to know, if this is also the case for this policy.

I will create a bar plot and I will paint it considering the different geographic zones.

```
ggplot()+
  geom_bar(aes(df$Owner_age,fill=as.factor(df$Geo_zone)))+
  labs(x="Years old of the insured person",
```

```
      y="Number of registers",
      title="Age vs zone",
      subtitle = "Age of the insured person vs geographic zone",
      fill = "Geographic zone")+
      scale_fill_discrete(labels = c("Zone 1", "Zone 2","Zone 3","Zone 4","Zone 5","Zone 6","Zone 7"))
```

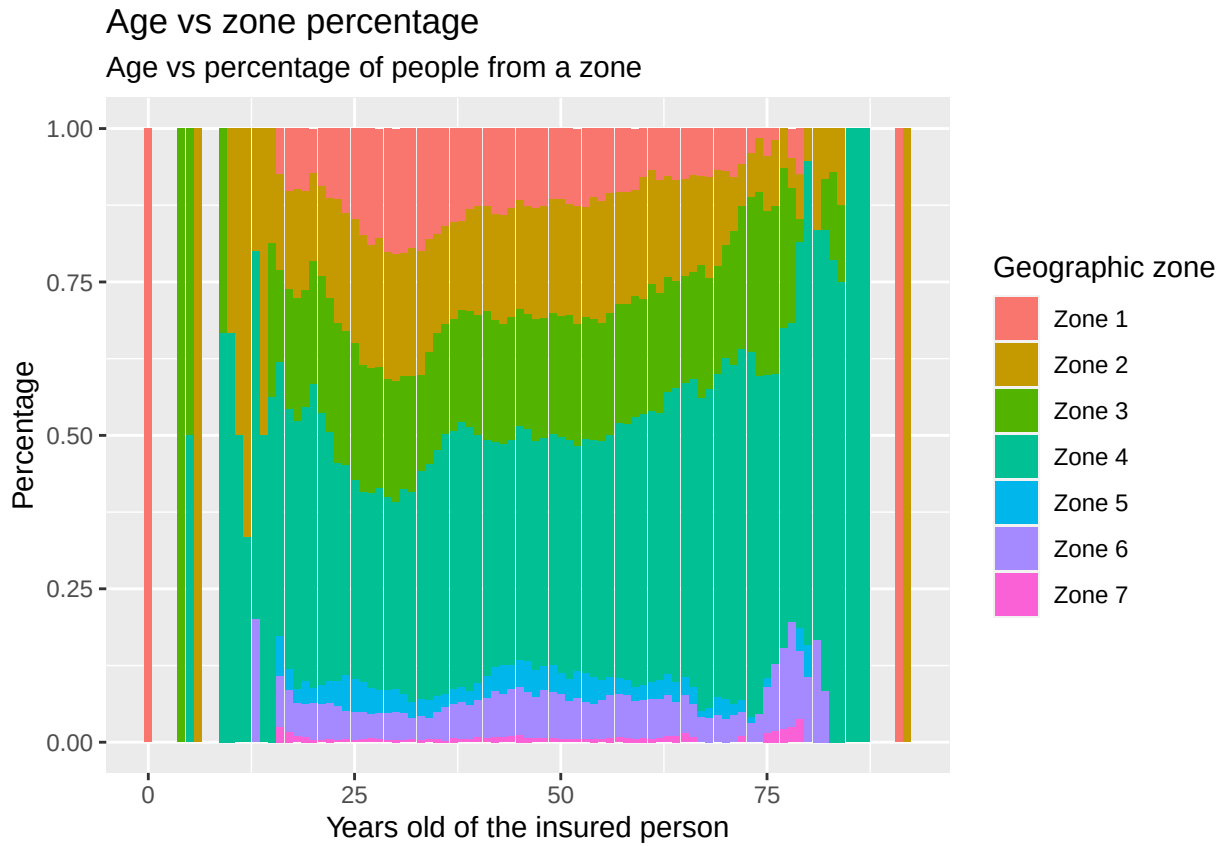


It is possible to see that the business is focused on specific geographic zones. Customers from the first, second, third and especially forth zones, seem to be the target of the business.

In this plot, we can also see, that the percentage of individuals from the different zones, change notoriously depending on the age of the insured. I would now like to focus on the percentage that every zone represents depending on the age of the customer.

I will use a bar plot, but this time with the command 'position="fill"'. This will force the graph to express the results in percentage, not in number of registers.

```
ggplot()+
  geom_bar(aes(df$Owner_age,fill=as.factor(df$Geo_zone)),position="fill")+
  labs(x="Years old of the insured person",
       y="Percentage",
       title="Age vs zone percentage",
       subtitle = "Age vs percentage of people from a zone",
       fill = "Geographic zone")+
       scale_fill_discrete(labels = c("Zone 1", "Zone 2","Zone 3","Zone 4","Zone 5","Zone 6","Zone 7"))
```



In this graph, we can see an interesting pattern. The insured individuals from the zones one, two and three together, agglomerate more than 50% of the "young drivers" target.

From that point on, the percentual presence of this areas, is heavily reduced. It seems that even if the business still is active on these zones, the company focus its attention on the fourth zone as the age of the customers grow.

HIGH RISK CLIENTS

Now that I know the target of the policy, I will start to focus on finding the characteristics that constitute a client with a high risk of having an accident/suing aclaim.

The most obvious one, is the age. Many studies show that "young drivers" are less aware of the danger. Due to the lack of experience and the false security that they feel, young people, normally have a driving style that is more likely to cause and accident than the driving style adopted by more experienced drivers.

I will now create a table that will group people by ages. The first group, the "young drivers" will be formed by individuals from 0 to 30 years, the second group, from 31 to 50, the third one from 51 to 70 and the last one from 70 to 100.

This table will express, from the total amount of individuals insures, which percentage represent every group of age.

I will also create two additional tables. The first one will show, from all the 1-accident claims that the clients have issued, which percentage corresponds to every group of age. The second one, will do the same, but this time, with the 2-accident claims.

```
prop.table(table(cut(df$Owner_age,breaks = c(-0.1,30,50,70,100))))
```

```
##
##  (-0.1,30]    (30,50]    (50,70]   (70,100]
```

```
## 0.24264114 0.46816323 0.27853690 0.01065873
```

```
prop.table(table(cut(df$Owner_age,breaks = c(-0.1,30,50,70,100)),df$number_of_claims==1),2)[,2]
```

```
## (-0.1,30]   (30,50]   (50,70]  (70,100]
## 0.5054432 0.3374806 0.1570762 0.0000000
```

```
prop.table(table(cut(df$Owner_age,breaks = c(-0.1,30,50,70,100)),df$number_of_claims==2),2)[,2]
```

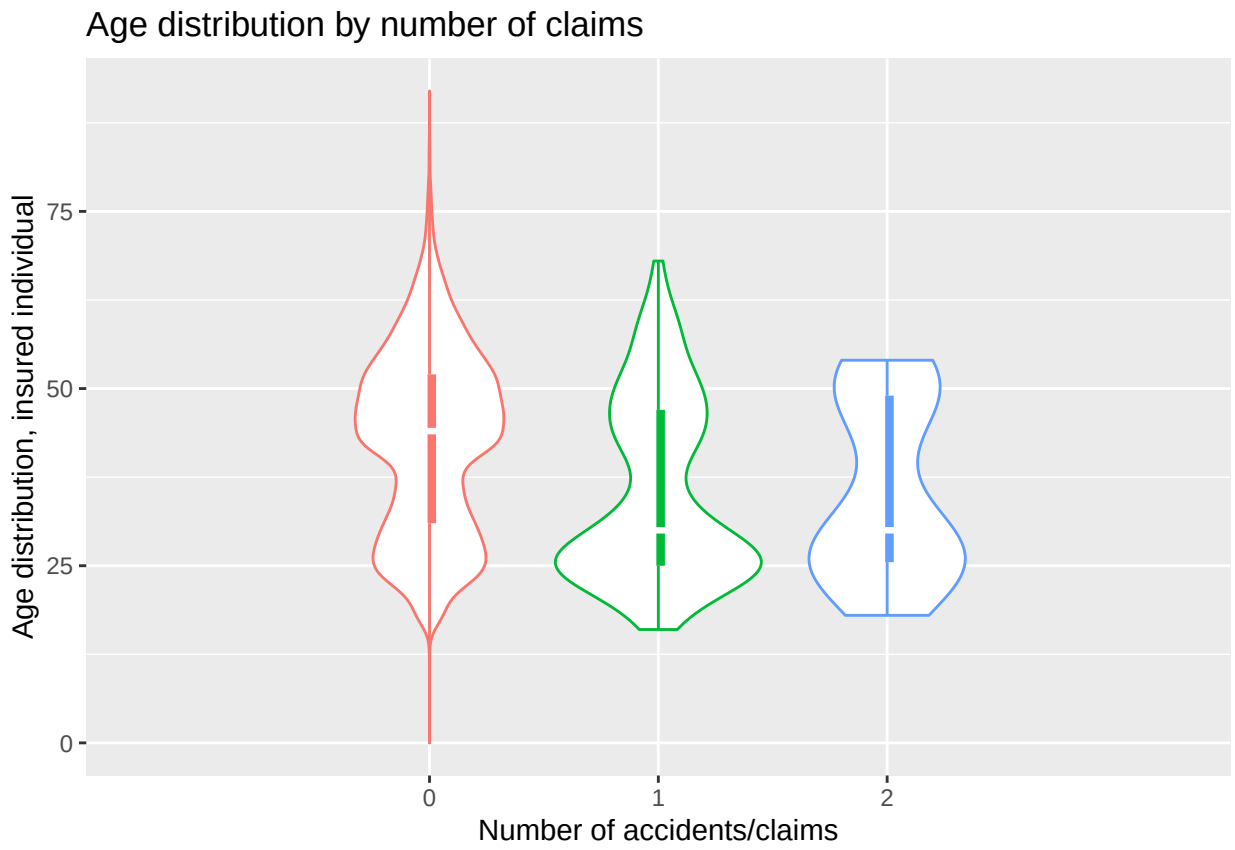
```
## (-0.1,30]   (30,50]   (50,70]  (70,100]
## 0.5185185 0.3333333 0.1481481 0.0000000
```

These tables, show that the customers who have less than 31 years old, represented just the 24.26% of the total individuals insured. Nonetheless, they issued the 50.54% of the 1-accident claims and the 51.85% of the 2-accident claims.

With the objective of seeing the distribution of ages based on the number of claims, I will divide the data by quantity of accidents and create a violin plot and a box plot.

```
library(ggthemes)

ggplot(df,aes(as.factor(number_of_claims), Owner_age, color=as.factor(number_of_claims)))+
  geom_violin()+
  geom_tufteboxplot(median.type = "line",width = 3)+
  theme(legend.position = "none")+
  labs(x="Number of accidents/claims",
       y="Age distribution, insured individual",
       title="Age distribution by number of claims")
```



According to this plot, among the drivers who had 0 registered claims, is possible to see two clear groups, the

biggest one that corresponds to the "experienced drivers" and a smaller one that corresponds to the "young drivers".

Regarding the 1-accident claims and 2-accident claims groups, a great majority of the distribution is concentrated in the lower part of the graph. That means that the clients who are younger than 35-37 years old have a great majority of the accidents.

Before proceeding with this analysis, I would like to clarify, that from now on, due to the reduced number of 2-accident claims, I will just differentiate those clients who issued 0 claims and those that issued at least 1.

I will now examine the geographic zone by creating two percentual tables.

The first one will allow us to see, from the total number of insured individuals, which percentage are from a specific zone. With the second one, we will see, from all the claims issued, which percentage of them is every zone responsible for.

```
perc_costum_per_zone=prop.table(table(df$Geo_zone))
accidents_costum_per_zone=prop.table(table(df[df$number_of_claims!=0,"Geo_zone"]))
perc_costum_per_zone
```

```
##
##           1           2           3           4           5           6
## 0.132955320 0.182716738 0.197093636 0.384458078 0.036825308 0.060172275
##           7
## 0.005778645
```

```
accidents_costum_per_zone
```

```
##
##           1           2           3           4           5           6
## 0.258208955 0.241791045 0.176119403 0.283582090 0.013432836 0.025373134
##           7
## 0.001492537
```

In these tables, we can see a remarkably interesting fact.

38.45% of the clients are from the fourth zone but this zone just causes a 28.36% of the accidents.

The area with more clients, have the biggest difference between the percentage of clients insured and the percentage of accidents caused by them.

```
perc_costum_per_zone-accidents_costum_per_zone
```

```
##
##            1            2            3            4            5            6
## -0.125253635 -0.059074307  0.020974233  0.100875988  0.023392472  0.034799141
##            7
##  0.004286108
```

This is a demonstration that the business has chosen the target for this policy wisely. Nevertheless, there is one thing that we need to consider. As we can see, the zones 3, 5, 6 and 7 also obtain positive results.

The fact that most of the territories are in "positive" numbers, means that the percentages of the zones 1 and 2 are, as we can see, greatly negative, meaning that they have a high number of accidents. In other words, even if the ratio ("number of claims)/(number of insured people)" from the fourth zone is positive for the insurance company. The percentage that the fourth zone has obtained, is so positive due to the high number of accidents that the zones 1 and 2 have had.

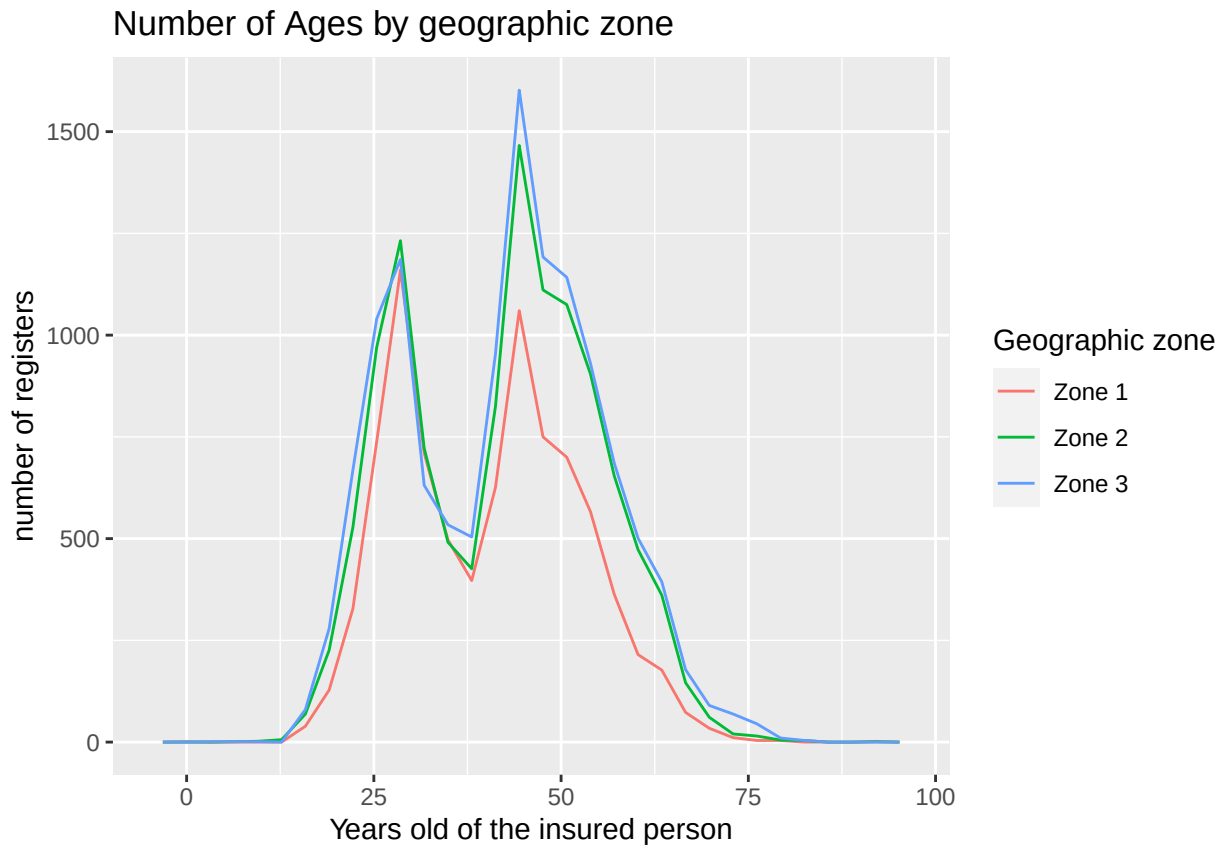
In fact, the third plot that we used in this analysis, showed how representative the zones 1, 2 and 3 were among the young drivers. We have also seen that unexperienced drivers have more accidents. That is why, I assume that this is the reason why the zones 1 and 2 obtain this bad result.

Notice that just the zone 1 and 2 have negative results, but the third territory, which as we saw in the third graph of this analysis, showed a similar behavior as the first two zones, obtains a positive result.

The results for the first 3 areas are the following ones: A 13.29% of the insured individuals are from the first zone, this area accumulates the 25.82% of the accidents, 13.29%-25.82% = -12.53% A 18.27% of the insured individuals are from the second zone, this area accumulates the 24.17% of the accidents, 18.27%-24.17% = -5.90% A 19.71% of the insured individuals are from the third zone, this area accumulates the 17.61% of the accidents, 19.71%-17.61% = 2.10%

Now I would like to understand why the results of these 3 zones are so different if, as we have seen, their evolution was so similar. To understand better these results, I will use a Frequency polygon plot, with 30 bins. I will just focus these 3 first zones.

```
ggplot(df[df$Geo_zone<=3,])+
  geom_freqpoly(aes(Owner_age,color=as.factor(Geo_zone)),bins = 30)+
  labs(x="Years old of the insured person",
       y="number of registers",
       title="Number of Ages by geographic zone",
       color = "Geographic zone")+
       scale_color_discrete(labels = c("Zone 1", "Zone 2","Zone 3"))
```



In this graph we can see that the first zone, has more "young drivers" than mature ones. With this fact in mind, makes sense for this territory, to present the highest percentage of accidents.

Regarding the second and third areas, they have a remarkably similar number of "young drivers". Nevertheless, the third zone has significantly more "mature drivers" in different points of the plot.

I presuppose that the difference on the results between the areas 2 and 3, are mostly explained by this fact. However, there might be other factors involved, like for example, the types of vehicles that are used in the

different zones.

Now I will focus on analyzing the automobile that the insured persons were using. According to different studies that I have found, the age of the vehicle is a particularly important factor. People tend to feel that a newer automobile is safer and more reliable, and thus, they may adopt a riskier way of driving. This normally leads to a higher number of accidents.

I will start by diving the different insured vehicles in groups of years. From 0 years to 5, from 6 to 10, from 11 to 15... until reaching 30. An automobile that has more than 30 years is normally considered incredibly old and not many people is willing to drive it.

That is why the last group of vehicles will be from 31 to 100 years.

I will show the results in a percentage table.

```
prop_age=prop.table(table(cut(df$vehicle_age,breaks = c(-0.1,5,10,15,20,25,30,100))))
prop_age
```

```
##
##   (-0.1,5]     (5,10]    (10,15]    (15,20]    (20,25]    (25,30]   (30,100]
## 0.25879965 0.17638037 0.25497304 0.18847989 0.04920369 0.02438495 0.04777840
```

As we can see, the most popular automobiles are the ones that have less than 5 years and the ones that have between 11 and 15 years. The vehicles that have between 6 and 10 years and those that have between 16 and 20 years are also popular, but not that much. An interesting conclusion that we can extract from this table, is that many people don't want a vehicle that has more than 20 years. The percentage of insure automobiles that have between 21 and 100 years is just 12.14% (0.04920369+0.02438495+0.04777840) of the total.

Now, considering the percentage of vehicles insured in every group of age, I will do another table to see, from the total number of vehicles that suffered at least one accident, which percentage corresponds to every group.

```
prop_age_accidents=prop.table(table(cut(df$vehicle_age,breaks = c(-0.1,5,10,15,20,25,30,100)),df$number_
prop_age_accidents
```

```
##    (-0.1,5]      (5,10]     (10,15]     (15,20]     (20,25]     (25,30]
## 0.438805970 0.232835821 0.210447761 0.085074627 0.016417910 0.007462687
##    (30,100]
## 0.008955224
```

With this last table, I can confirm that people tend to be more confident while driving a new vehicle, which leads to more accidents. The automobiles that have less than 5 years, suffered the 43.88% of the accidents, even if they just represent the 25.88% of the total insured ones.

As we can see, the percentage of cases suffered by every group of age, decreases as the vehicle gets older. In fact, those that have between 16 and 20 years, suffered just 8.50% of the accidents. Considering that these vehicles represent a 18.85% of the total number of the insured vehicles, their ratio of accident is very reduced.

Even if it is clear that the age of the vehicle affects the number of issued claims, there is one more factor that we need to consider. We need to know which impact the "young drivers" are having in these results.

To see which percentage of young people, use automobiles that have 15 years or less, I will create the following table:

```
df_vehicle_years=mutate(df,
       vehicle_years= cut(vehicle_age,breaks = c(-0.1,5,10,15,20,25,30,100)),
       client_years=cut(Owner_age,breaks = c(-0.1,30,50,70,100)))

prop.table(table(df_vehicle_years$client_years,df_vehicle_years$vehicle_age<=15),1)[1,]
```

```
##     FALSE      TRUE
## 0.2023369 0.7976631
```

As we can see, just 20.23% of young people have a vehicles older than 15 years.

We learned before, that more than half of the accidents are caused by "young drivers". It is sure to say that the percentage of accidents with automobiles that have less than 15 years is very influenced by the presence of "young drivers".

To learn more about it, I will create two plots. The first one, will show the percentage of accidents that the different vehicles, divided by groups of age, have suffered.

The second one, will contain the same information but this time, I will remove the "Young drivers". In this way it will be possible to see how the age of the automobile affects the individuals who are older than 30 years old.

```
prop.table(table(df_vehicle_years[df_vehicle_years$Owner_age>30,"vehicle_years"],df_vehicle_years[df_ve
```

```
##
##                   FALSE        TRUE
##   (-0.1,5] 0.988190914 0.011809086
##   (5,10]   0.992179215 0.007820785
##   (10,15]  0.993992044 0.006007956
##   (15,20]  0.995953814 0.004046186
##   (20,25]  0.997419830 0.002580170
##   (25,30]  0.997785978 0.002214022
##   (30,100] 0.998859316 0.001140684
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

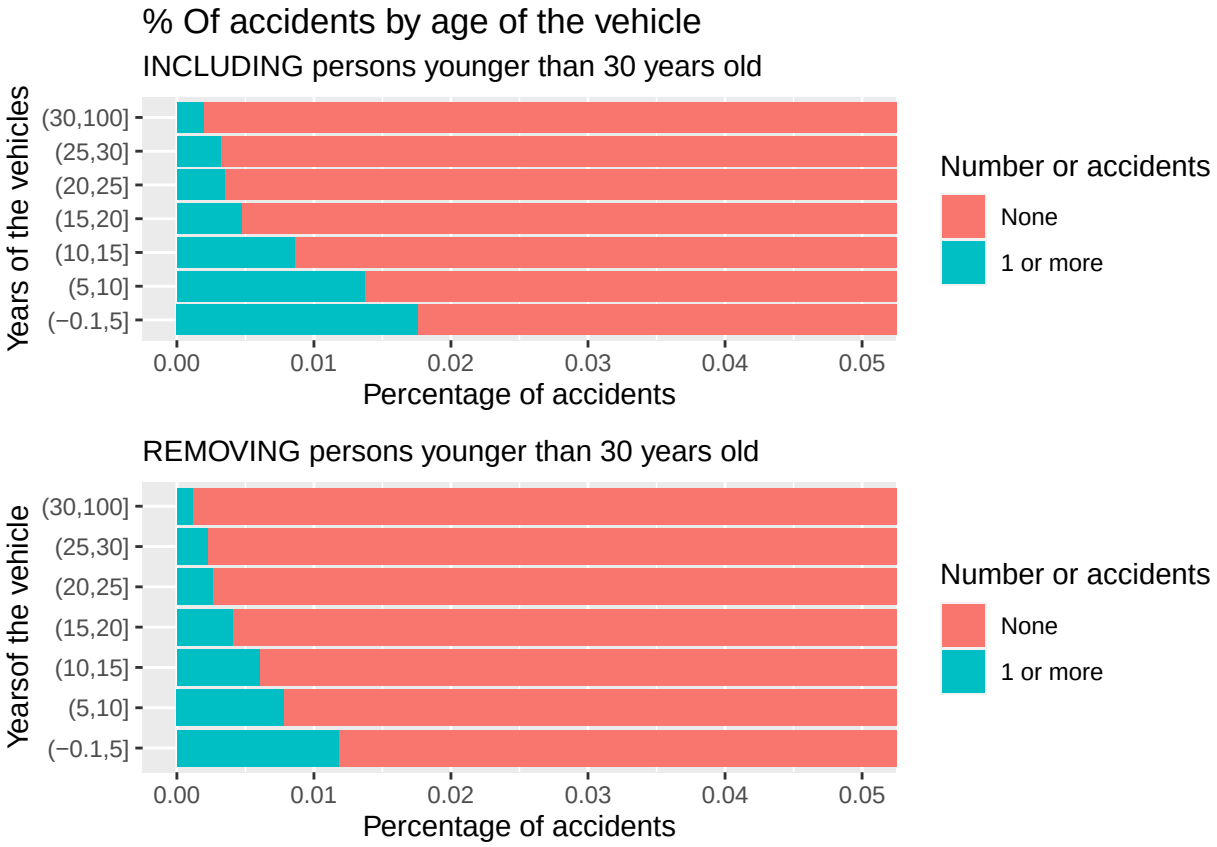
```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
plot1=ggplot(df_vehicle_years)+
  geom_bar(aes(vehicle_years,fill=number_of_claims!=0),position="fill")+
  labs(x="Years of the vehicles",
       y="Percentage of accidents",
       title = "% Of accidents by age of the vehicle",
       subtitle = "INCLUDING persons younger than 30 years old",
       fill = "Number or accidents")+
  scale_fill_discrete(labels = c("None", "1 or more"))+
  coord_flip()+
  scale_y_continuous(limits=c(0,0.05),oob = rescale_none)


plot2=ggplot(df_vehicle_years[df_vehicle_years$Owner_age>30,])+
```

```
  geom_bar(aes(vehicle_years,fill=number_of_claims!=0),position="fill")+
  labs(x="Yearsof the vehicle",
       y="Percentage of accidents",
       subtitle = "REMOVING persons younger than 30 years old",
       fill = "Number or accidents")+
  scale_fill_discrete(labels = c("None", "1 or more"))+
  coord_flip()+
  scale_y_continuous(limits=c(0,0.05),oob = rescale_none)


grid.arrange(plot1, plot2, nrow=2)
```



For the correct interpretation of these plots, please notice that I have used the command "scale_y_continuous(limits=c(0,0.05))" to limit the Y axis.

I have done such thing because more than 95% of the vehicles didn't suffer any accident. Since we are interested in the ones that did have an accident, I thought that the best way to observe the results was to reduce the Y axis.

With these two plots, we can confirm that the "young drivers" are greatly increasing the number of accidents for the different groups of age of the vehicles. Especially the vehicles with less than 16 years old.

Additionally, we can see that even removing the "younger individuals" a newer automobile, has more probabilities of suffering an accident.

Another thing that I would like to see is, if there is a relation between how powerful a vehicle is and the number of accidents.

I presume that if a vehicle can run extremely fast, the drivers may misuse this possibility. The faster a vehicle is circulating, the more difficult it is to control it and thus it is easier to have an accident.

As we can see in the description of the case, Power_by_weight is a variable that takes into consideration the power of the vehicle and divides it by the average weight it is expected to carry. The result is rounded to the nearest lower integer.

If the result is 1, it means that the vehicle has a low power compared with its weight and thus, it won't run much. If the result is 7, it means that the vehicle can run a lot.

As I have been doing until now, the first thing I will do to understand this new variable is to create two tables.

The first one, with the percentage that the vehicles of the different levels of power represent against the total number of automobiles insured.

The second one, with the percentage of accidents that the vehicles of the different levels of power have caused against the total number of accidents.

```
perc_vehicle_by_power=prop.table(table(df$power_by_weight))
perc_vehicle_by_power
```

```
##
##          1          2          3          4          5          6          7
## 0.10894218 0.08062217 0.29288282 0.19176427 0.18305757 0.13024416 0.01248683
```

In the first table, we can see that the vehicles with a power of 1 and 2 are not extremely popular among the insured people. The vehicles with a power of 3 are the most used ones. Finally, the vehicles with a power of 4, 5 and 6 seem to have a good level of acceptance.

However, the most relevant thing to notice from this table, is the extremely low level of vehicles that have a power of 7. I would like to investigate more about why there are so few automobiles of this kind in this policy.

```
accidents_vehicle_by_power=prop.table(table(df[df$number_of_claims!=0,"power_by_weight"]))

accidents_vehicle_by_power
```

```
##
##           1           2           3           4           5           6
## 0.067164179 0.085074627 0.235820896 0.138805970 0.214925373 0.250746269
##           7
## 0.007462687
```

Now I will subtract the first table from the second one. If, as I said before there is a positive correlation between the power of a vehicle, and the number of accidents, we will find the following results:

1-The percentage of the vehicles with a power of 1,2,3 and maybe 4, will be negative. This will happen, because, they will have a higher percentage of insured vehicles than accidents caused by these low-power automobiles.

2-The results of the vehicles with a power of 5, 6 and 7 will be positive, since they will have a higher percentage of accidents.

```
accidents_vehicle_by_power-perc_vehicle_by_power
```

```
##
##            1            2            3            4            5            6
## -0.041778003  0.004452454 -0.057061920 -0.052958298  0.031867804  0.120502109
##            7
## -0.005024145
```

This table seems to confirm that there is a correlation between the power of the vehicle and the number of claims.

The vehicles with a power lower than 4 tend to have less accidents and those with a higher power tend to have more of them.

However, it is important to highlight that this pattern seems not to apply to the vehicles with a power of 2 and 7.

The vehicles with a power of 2, have a higher number of accidents than expected and the vehicles with a power of 7, have a lower number of them.

I suppose, that most of the vehicles with a power of 2, will have less than 5 years and/or will be driven by young individuals and that most of the vehicles with a power of 7 will be rather old and used by more mature drivers.

Now, I will focus on the vehicles with power of 2 and 7. First, I will create the data frame "data_2_7" that will contain the information from these vehicles.

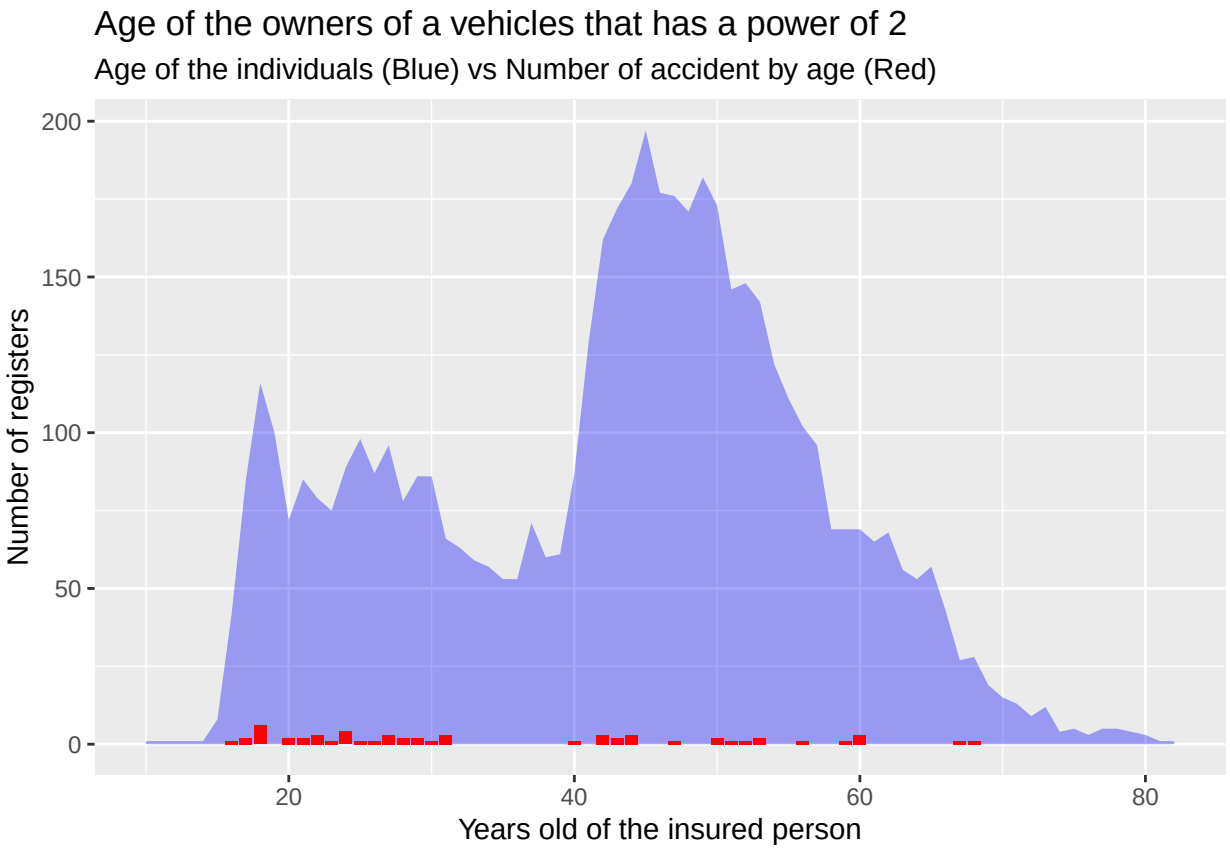
```
data_2_7=df[(df$power_by_weight==2)| (df$power_by_weight==7),]
```

I will first examine the distribution of the age of the drivers of the vehicles with power of 2 with an area plot.

Furthermore, with a bar plot, I will mark the ages of the drivers that had at least 1 accident.

```
tab=data.frame(table(data_2_7[data_2_7$power_by_weight==2,"Owner_age"]))
tab$Var1=as.numeric(as.character(tab$Var1))
tab$Freq=as.numeric(tab$Freq)
accidents=data_2_7[(data_2_7$power_by_weight==2) & (data_2_7$number_of_claims!=0),"Owner_age"]

ggplot()+
  geom_area(aes(tab$Var1,tab$Freq),fill="blue",alpha=0.35)+
  geom_bar(aes(accidents),fill="red")+
  labs(x="Years old of the insured person",
     y="Number of registers",
     title="Age of the owners of a vehicles that has a power of 2",
     subtitle = "Age of the individuals (Blue) vs Number of accident by age (Red)")
```

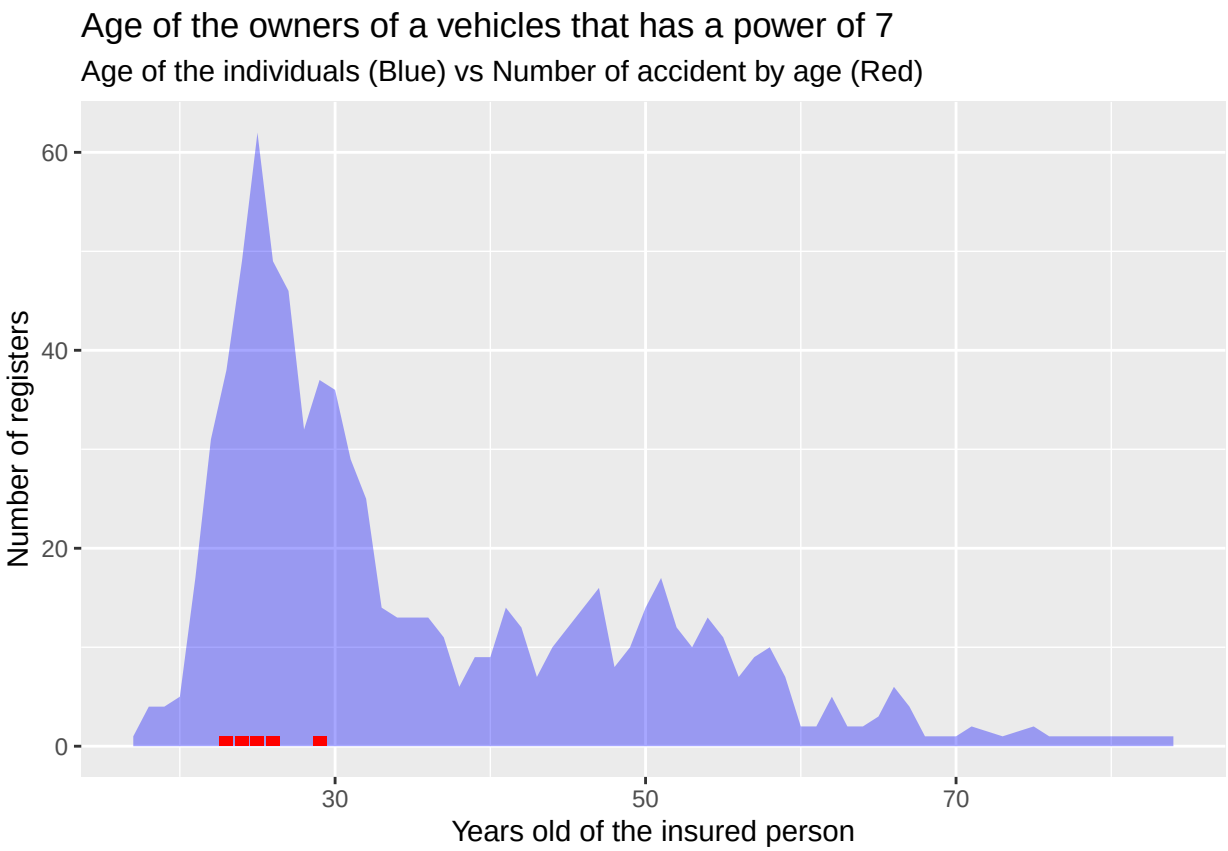


This plot shows that the drivers of the vehicles with a power of 2, tend to have between 40 and 57 years old.

Now I will apply the same procedure to the automobile with a power of 7.

```
tab=data.frame(table(data_2_7[data_2_7$power_by_weight==7,"Owner_age"]))
tab$Var1=as.numeric(as.character(tab$Var1))
tab$Freq=as.numeric(tab$Freq)
accidents=data_2_7[(data_2_7$power_by_weight==7) & (data_2_7$number_of_claims!=0),"Owner_age"]

ggplot()+
  geom_area(aes(tab$Var1,tab$Freq),fill="blue",alpha=0.35)+
  geom_bar(aes(accidents),fill="red")+
  labs(x="Years old of the insured person",
       y="Number of registers",
       title="Age of the owners of a vehicles that has a power of 7",
       subtitle = "Age of the individuals (Blue) vs Number of accident by age (Red)")
```

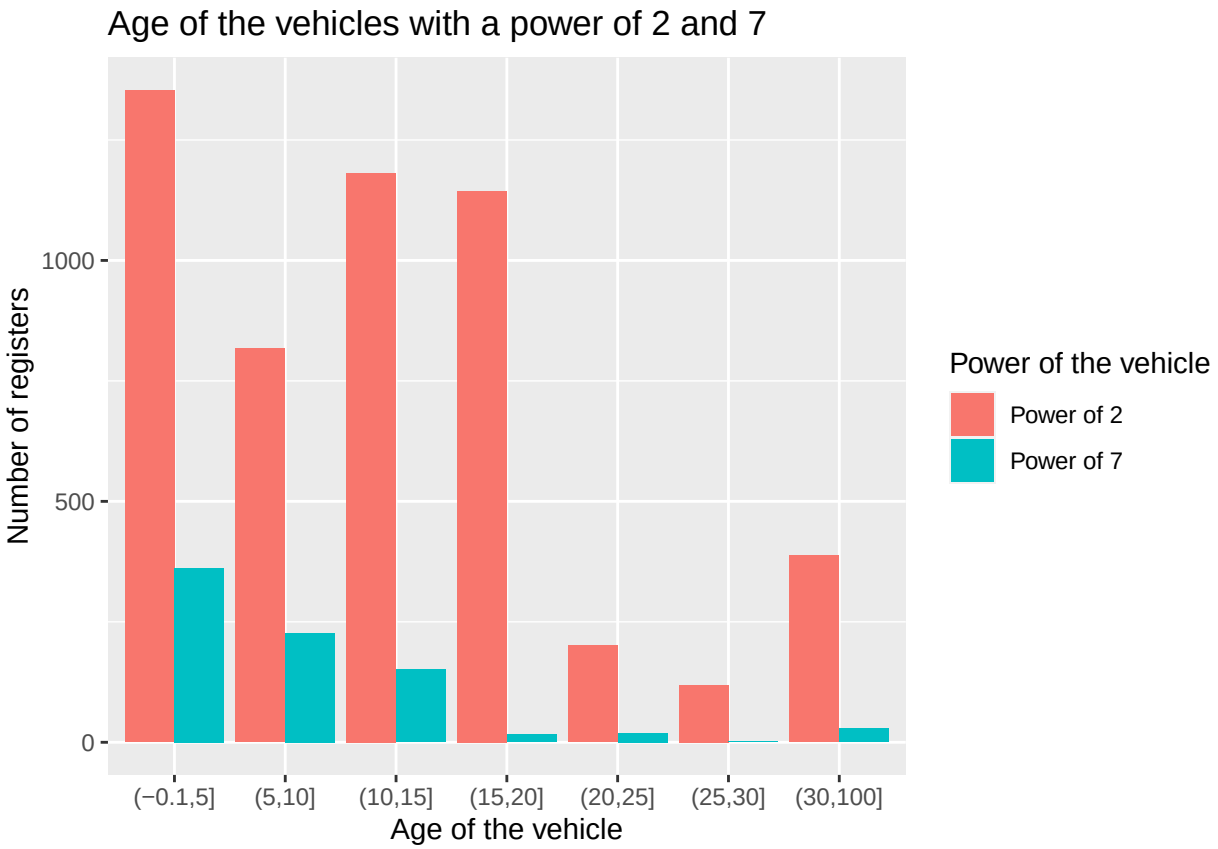


In this plot we can see that the vehicles with a power of 7 are used by different ages, but especially by drivers who have between 22 and 32 years old, and thus, are mostly considered "young drivers".

I can't conclude that the age of the drivers explains the exception that these two kinds of automobiles experience.

I will now proceed to create a bar plot that will allow us to see the age of the vehicles.

```
ggplot(data_2_7)+
  geom_bar(aes(cut(vehicle_age,breaks = c(-0.1,5,10,15,20,25,30,100)),fill=as.factor(power_by_weight)),p
  labs(x="Age of the vehicle",
       y="Number of registers",
       title="Age of the vehicles with a power of 2 and 7",
       fill = "Power of the vehicle")+
  scale_fill_discrete(labels = c("Power of 2", "Power of 7"))
```



From this graph we can extract 3 important ideas:

-A great majority of the vehicles with a power of 7, have less than 15 years. Furthermore, a good part of them have less than 5 years.

-The vehicles with a power of 2 show an irregular distribution. Even if most vehicles have less than 20 years, their distribution shows that these vehicles tend to be older than those with a power of 7.

-The number of vehicles with a power of 7 is very reduced compared with the quantity of automobiles with a power of 2. This might be caused by a high popularity of these last vehicles or by restrictions imposed on the automobiles with a power of 7.

We can conclude this brief analysis of these two kinds of vehicles, by saying that the age of the driver and the age of the vehicle, can't explain why they behave differently.

Now I will focus on a fact that I highlighted at the beginning of this analysis, the "bonus level" of the insured person. Maybe this variable will be able to explain the behavior of these two kinds of vehicles.

The feature " bonus_of_owner" explains "how good" a driver is. Every person starts with a level 1 of " bonus_of_owner". For every year without any claim, the driver goes up one level.

If this is the case, we could assume that a driver that has been using this policy for less than a year, would have a "bonus_of_owner" equal to 1. However, this is not always true.

Before proceeding with this analysis, I would like to clarify WHY I am interested in analyzing the different levels of bonus.

As we have seen, the percentage of accidents of the vehicles with a power of 7 is lower than expected.

At the same time, the number of vehicles of this kind that are insured, is very reduced compared with the other kind of vehicles.

I presume that the business is limiting the number of vehicles of this kind that are being insured with this policy. Maybe, the corporation is just allowing the "good drivers" to insure a automobile with a power of 7 with this policy. That may be happening, because the company knows that with such a powerful vehicle, the risk of an accident is high.

In the following table I will include the bonus level of the different clients. I will also divide them is two groups, the drivers who have been using the policy for less than 1 year, and those who have been using it from 1 year to 11 years.

```
table(data_2_7$bonus_of_owner,cut(data_2_7$policy_years,breaks = c(-0.1,0.999,11)))
```

```
##
##     (-0.1,0.999] (0.999,11]
##   1         1299        459
##   2          683        284
##   3          439        150
##   4          340         83
##   5          255         66
##   6          284         77
##   7          988        603
```

In this table, it is possible to notice, that most of the clients that have been subscribed to this policy for less than one year, have a bonus higher than 1. Some even have a bonus of 7, the maximum level.

I suppose that this is the case because the clients are allowed to change from a policy to another, while keeping their bonus level.

Now I will create a table that will show us, the number of drivers that have the maximum level of bonus, for both kinds of vehicles. Additionally, I will divide this same table, by the total number of insured automobiles of these two types.

The result will tell us, from the total number of vehicles insured, which percentage of them are driven by persons with a bonus level of 7. In other words, individuals that the busines considers "perfect drivers".

```
table(data_2_7[data_2_7$bonus_of_owner==7,"power_by_weight"])
```

```
##
##    2    7
## 1236  355
```

```
table(data_2_7[data_2_7$bonus_of_owner==7,"power_by_weight"])/table(data_2_7$power_by_weight)
```

```
##
##         2         7
## 0.2375096 0.4404467
```

As we can see in this table, 23,75% of the total insured vehicles with a power of 2 and 44.04% of the total insured automobiles with a power of 7,have a bonus level of 7 and thus are considered "perfect drivers". It can't be a coincidence, that almost half of the drivers with the most powerful automobiles have the best possible driving record. I presume that the busines is, in fact, limiting the individuals that are eligible to be a part of this policy, based on the power of their vehicle and their driving skills.

I will now divide all the registers of the clients in this data frame by how good they are driving. If they have a bonus level equal or less than 4, I will tag them as "bad-average drivers". If their level is higher than 4, the tag will be "good drivers". With this division in mind, I will create a new table to see the percentage of good drivers" that these two kinds of vehicles have.

```
table(data_2_7[data_2_7$bonus_of_owner>=5,"power_by_weight"])/table(data_2_7$power_by_weight)
```

```
##
```

```
##         2         7
## 0.3403151 0.6228288
```

According to the tables 62.28% of the drivers who have a vehicle with a power of 7 could be classified as "good drivers" It is now clear that the insurance company shows a preference for individuals with a low record of accidents for the automobiles with a power of 7.

Regarding the vehicles with a power of 2, we can see that 65.96% (0.6596849 = 1-0.3403151) are considered "average-bad drivers".

This level of "average-bad drivers" among the automobiles with power 2 and the high volume of "good drivers" among the vehicles with power 7 could explain why we are seeing more accidents with the first kind of automobile, and less with the second one.

If what I said until now, is true, if I create a plot that classifies all the insured vehicles by its power, and I divide that group by whether the bonus level of the drivers is LOWER than 5, we should be able to see:

1- That the percentage of "good drivers" among the vehicles with a power of 2 is the lowest one.

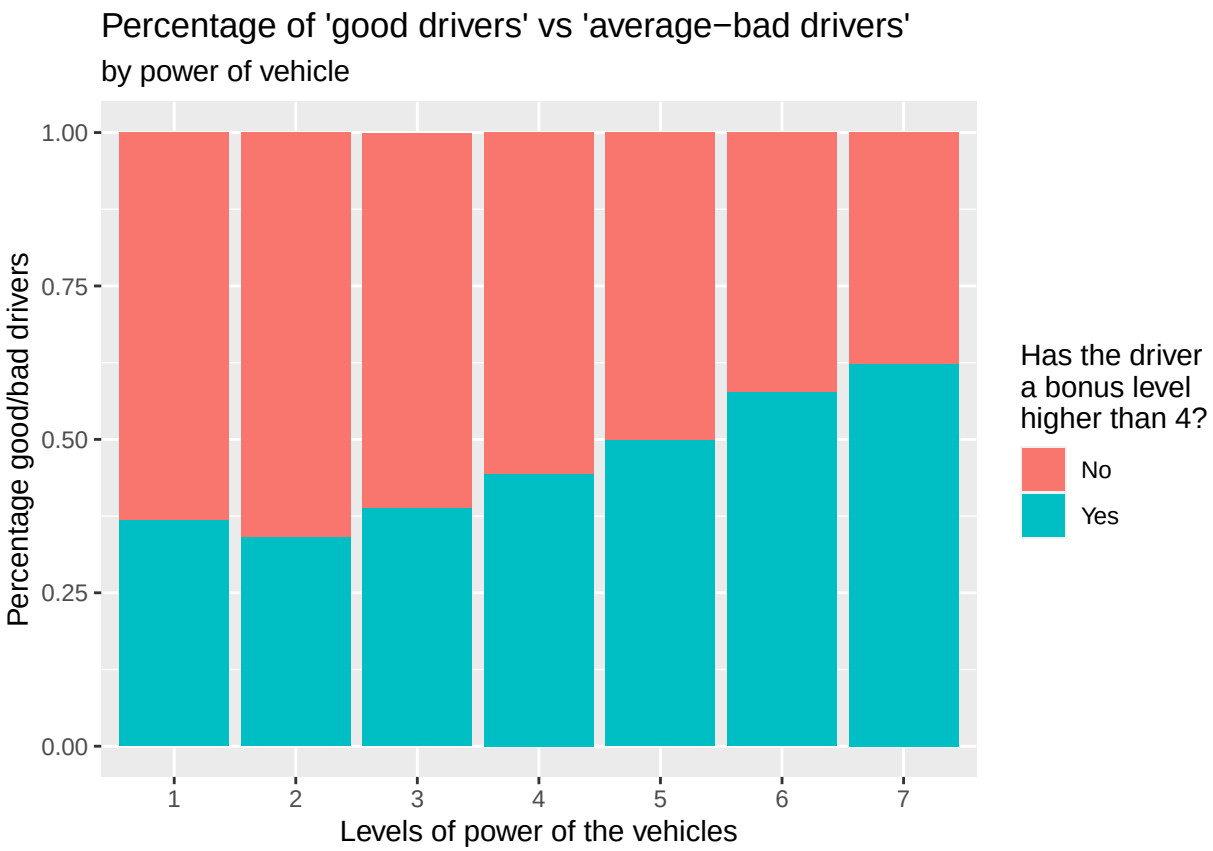
2- That the percentage of "good drivers" increases as the power of the automobile rises.

```
prop.table(table(df$power_by_weight,df$bonus_of_owner>4),1)
```

```
##
##         FALSE      TRUE
##   1 0.6318259 0.3681741
##   2 0.6596849 0.3403151
##   3 0.6125364 0.3874636
##   4 0.5567135 0.4432865
##   5 0.5022004 0.4977996
##   6 0.4234566 0.5765434
##   7 0.3771712 0.6228288
```

```
freq=data.frame(table(df$power_by_weight,df$bonus_of_owner>4))

ggplot()+
  geom_bar(aes(freq$Var1,freq$Freq,fill=freq$Var2),stat="identity",position = "fill")+
  labs(x="Levels of power of the vehicles",
       y="Percentage good/bad drivers",
       title="Percentage of 'good drivers' vs 'average-bad drivers'",
       subtitle ="by power of vehicle"  ,
       fill="Has the driver \na bonus level \nhigher than 4?")+
       scale_fill_discrete(labels = c("No", "Yes"))
```



As we can see, both assumptions are fulfilled and thus we can confirm that the business requires the driver to meet a defined standard to get insured by this policy with a powerful vehicle.

Finally, I will explore the costs of the claims that the drivers issued. One important detail that we need to know, is that the information of the original data frame, does not specify if the claim cost for those users who had 2 accidents, is the result of adding the price of both claims or if it is just the mean of the costs.

I think that to use the claim cost of the users who had 2 accidents, without understanding the meaning of these registers, will give us misleading conclusions.

```
table(df[df$number_of_claims>1,"number_of_claims"]) #there are just 27 registers
```

```
##
##  2
## 27
```

Since the number of these register is reduced, I prefer to omit them. I will start the analysis of the claim cost, by creating the data frame "df_1_accident". This data set will remove the customers that issued 2 claims and will include two new columns. One column will divide the age of the vehicle in different groups and the other one will do the dame with the age of the owners.

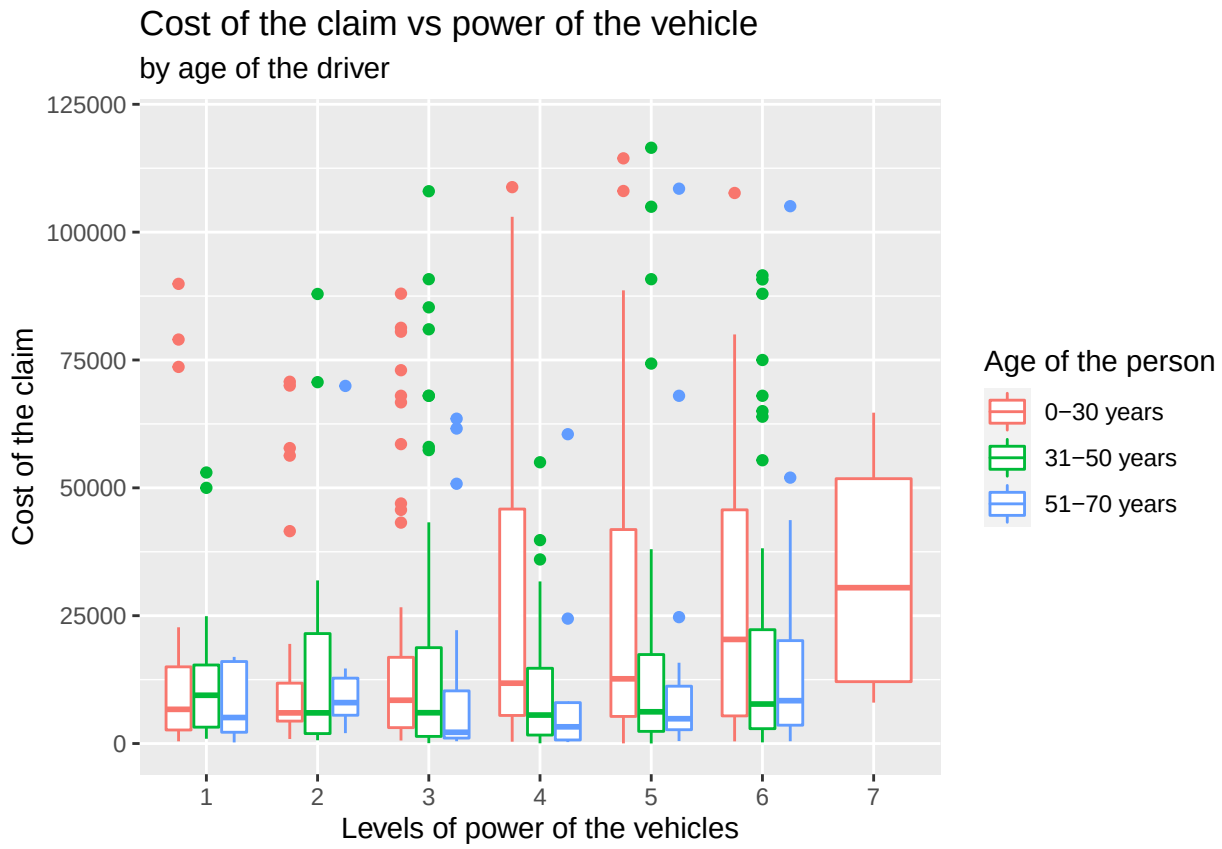
```
df_1_accident=df%>%
  filter(number_of_claims == 1)%>%
  mutate(vehicle_years= cut(vehicle_age,breaks = c(-0.1,5,10,15,20,25,30,100)),
         client_years = cut(Owner_age,breaks = c(-0.1,30,50,70,100)))
```

First, I will learn more about the relation between the claim cost, the power of the vehicle and the age of the insured.

I will use a boxplot that will show us the distribution of the cost of the claim, dividing it by the power or the

automobile that had the accident. I will also make a second division based on the different groups of age of the clients.

```
ggplot(df_1_accident)+
  geom_boxplot(aes(as.factor(power_by_weight),claim_cost,color=client_years))+
  ylim(c(0,120000))+
  labs(x="Levels of power of the vehicles",
     y="Cost of the claim",
     title="Cost of the claim vs power of the vehicle",
     subtitle ="by age of the driver"  ,
     color="Age of the person")+
     scale_color_discrete(labels = c("0-30 years", "31-50 years","51-70 years"))
```



In this next plot, we can see an interesting pattern, the cost of the accidents when the power of the vehicle is lower than 3, is approximately equal among the 3 groups of age.

However, the difference between the group of "young drivers" and the other two, tend to grow as the power of the vehicle increases.

I would require more data, to determine which is the cause of this difference. Nonetheless, with the data that I have, I would assume that it is related with the tendency of speeding that some "young drivers" show.

Now I will continue analyzing the relation between the claim cost, the vehicle years, and the client years.

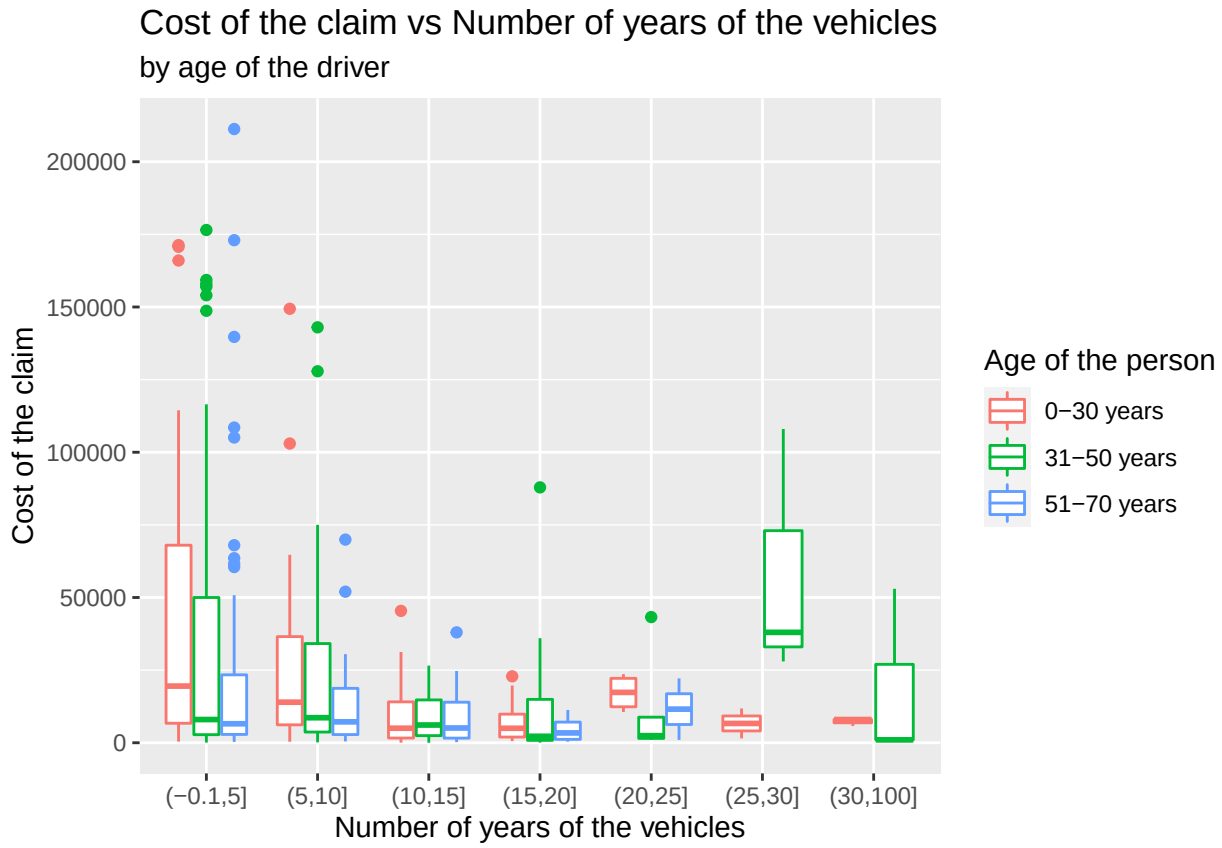
I will create a plot like the last one, but instead of dividing the claim cost by the power of the vehicle, I will use its age.

Before proceeding, I need to clarify that, as we have seen before, the number of vehicles with more than 21 years is reduced. Due to the reduced volume of this data, it may not be representative of the reality nor good enough to take conclusions out of it. That is why I will focus just on the automobiles with less than 21 years.

```
table(df_1_accident$vehicle_years)
```

```
##
## (-0.1,5]   (5,10]  (10,15]  (15,20]  (20,25]  (25,30] (30,100]
##      274      152      138       57       11        5        6
```

```
ggplot(df_1_accident)+
  geom_boxplot(aes(vehicle_years,claim_cost,color=client_years))+
  labs(x="Number of years of the vehicles",
     y="Cost of the claim",
     title="Cost of the claim vs Number of years of the vehicles",
     subtitle ="by age of the driver",
     color="Age of the person")+
     scale_color_discrete(labels = c("0-30 years", "31-50 years","51-70 years"))
```




```
options(warn = 0)
```

In this plot, we can see that the vehicles, show an interesting pattern. There is a great difference between the cost of the accidents of the "young drivers" and the other two groups of age within the vehicles with less than 11 years.

This difference on cost, is reduced as the vehicle's gets older. With a vehicle that has more than 10 years, the cost of the accidents caused by "young people" is similar to the other age groups.

I will finish this analysis explaining our 4 main findings:

1-Young drivers: From all the factors that we have seen during this analysis, the most important one, has

been the age of the insured person. The violin plot that we made, has showed us that a great majority of the accidents were caused by individuals who are less than 37 years old. In fact, 50.54% of the 1-accident claims and the 51.85% of the 2-accident claims were suffered by clients who are less than 30.

2-The age of the vehicle: 1.76% of the automobiles with less than 5 years have had at least one accident. This is relevant, because the data shows that the percentage of accidents is reduced as the vehicle gets older. The percentage of claims for the vehicles with more than 30 years is only 0.19%.

In this analysis, we have found, that 79.77% of the "young drivers" were using vehicles that had less than 15 years, and thus, they were increasing the number of accidents. However, when we removed them from the result, we have found that more experienced drivers, show the same pattern. The vehicles with less than 5 years suffered 1.18% of the accidents, and those with more than 30 years just represented a 0.11%. One last thing that I would like to highlight, is that the cost of the accidents that the "younger drivers" have, is notoriously higher for the cars that have less that 5 years. The cost of the claims decreases as the car gets older.

3-Power of the car: The EV ratio, expresses how powerful the engine of a vehicle is, considering the weight that it will normally carry. In this analysis, we have found that automobiles that have a reduced EV ratio, tend to have a lower level of accidents than those that have a higher power. The main problem with the cars that have an elevated EV ratio, is not that the percentage of accidents grow as the power increases, but that the cost of the claims also soars. This is especially notorious among "young drivers".

4-Good drivers: In this analysis, we have seen that the percentage that the "good drivers" represent from the total number of persons that insured a vehicle with a certain power, increases as the power of the automobile grows. For example, from the vehicles that have an EV ratio of 1, just 36.82% of the insured individuals are considered to be "good drivers" by the insurance company; the percentage of "good drivers" is 62.28% for the automobiles with a power of 7.