Exercice 1

1. The dimension of the data (num. of rows, num. of columns) is (***67856***, ***11***).

2. The first six lines of this data are

X	veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
1	1.06	0.3039014	0	0	0	HBACK	3	F	C	2
2	1.03	0.6488706	0	0	0	HBACK	2	F	A	4
3	3.26	0.5694730	0	0	0	UTE	2	F	E	2
4	4.14	0.3175907	0	0	0	STNWG	2	F	D	2
5	0.72	0.6488706	0	0	0	HBACK	4	F	C	2
6	2.01	0.8542094	0	0	0	HDTOP	3	M	C	4

3. Using the R function `str()`, we get the structure of the data

```
'data.frame':   67856 obs. of  11 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : chr  "HBACK" "HBACK" "UTE" "STNWG" ...
 $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : chr  "F" "F" "F" "F" ...
 $ area     : chr  "C" "A" "E" "D" ...
 $ agecat   : int  2 4 2 2 2 4 4 6 3 4 ...
```

4. We made use of the function `subset()` to delete the first column of `dataCar` and the function `transform()` to transform the variables `clm`, `numclaims`, `veh_body`, `veh_age`, `gender`, `area`, and `agecat` to a factor. The summary of the resulting data is

```
   veh_value      exposure       clm         numclaims    claimcst0
 Min.   : 0   Min.   :0.003   0:63232   0:63232   Min.   :    0
 1st Qu.: 1   1st Qu.:0.219   1: 4624   1: 4333   1st Qu.:    0
 Median : 2   Median :0.446             2:  271   Median :    0
 Mean   : 2   Mean   :0.469             3:   18   Mean   :  137
 3rd Qu.: 2   3rd Qu.:0.709             4:    2   3rd Qu.:    0
 Max.   :35   Max.   :0.999                       Max.   :55922

    veh_body      veh_age    gender     area       agecat
 SEDAN  :22233   1:12257   F:38603   A:16312   1: 5742
 HBACK  :18915   2:16587   M:29253   B:13341   2:12875
 STNWG  :16261   3:20064             C:20540   3:15767
 UTE    : 4586   4:18948             D: 8173   4:16189
```

```
 TRUCK  : 1750                          E: 5912   5:10736
 HDTOP  : 1579                          F: 3578   6: 6547
 (Other): 2532
```

5. Below is a Barplot of numclaims.

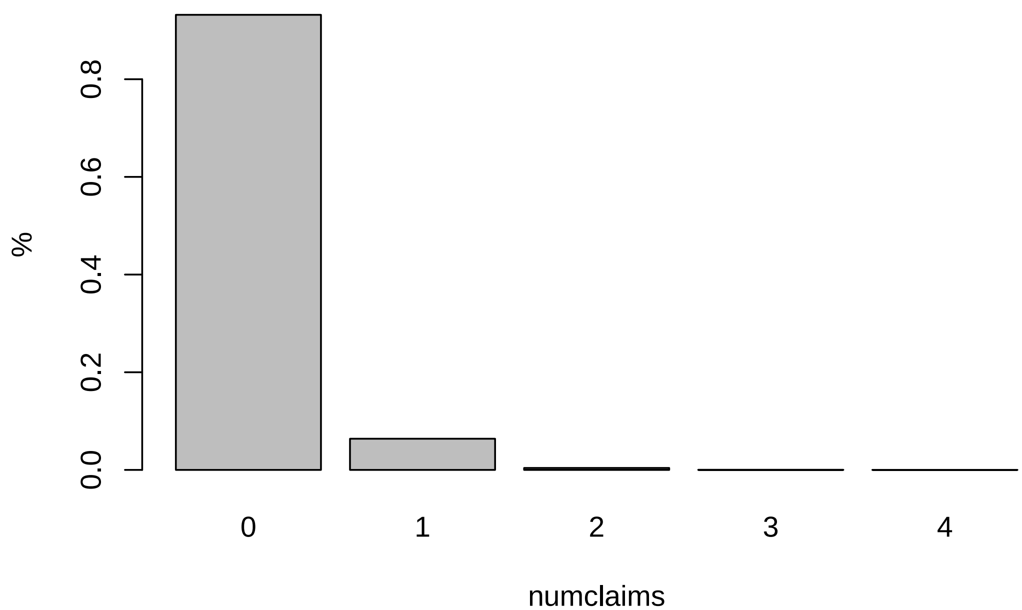

Figure 1: Barplot of ‘numclaims‘

6. We define `dataCar0` to be the subset data with *only variables* `claimcst0` and `veh_value` and *only subjects* with (`claimcst0` >0) and (`agecat` = 3 or 4). In the flowing we will work with this data. Its summary appears below.

```
   claimcst0       veh_value
 Min.   :  200   Min.   : 0.00
 1st Qu.:  354   1st Qu.: 1.07
 Median :  748   Median : 1.56
 Mean   : 1929   Mean   : 1.84
 3rd Qu.: 2035   3rd Qu.: 2.31
 Max.   :47297   Max.   :11.54
```

7. We fit a linear regression model with veh_value as independent variable and claimcst0 as dependent variable. We also fit another linear model but this time with log(claimcst0) as independent variable. The summary of each model is given below.

- `claimcst0` $\sim$ `veh_value`

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2053.3584  131.83266 15.575490 5.511928e-52
veh_value    -67.4188   60.78594 -1.109118 2.674995e-01
```

					confint	
	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
claimcst0 ~ veh_value						
(Intercept)	2053.358	131.833	15.575	0.000	1794.830	2311.887
veh_value	-67.419	60.786	-1.109	0.267	-186.622	51.785
log(claimcst0) ~ veh_value						
(Intercept)	6.834	0.047	145.356	0.000	6.741	6.926
veh_value	-0.023	0.022	-1.076	0.282	-0.066	0.019

- `log(claimcst0)` ~ `veh_value`

```
               Estimate Std. Error    t value  Pr(>|t|)
(Intercept)  6.83354029 0.04701249 145.355840 0.0000000
veh_value   -0.02333101 0.02167671  -1.076317 0.2819028
```

8. We compute, for each model, a 95% confidence intervals (confint) for the intercept and the slope parameters. We then use the `kableExtra` functions `kbl()`, `kable_styling()`, `pack_rows()` and `add_header_above()`, to construct the following table.

9. Figure 2 below show the scatterplots `claimcst0~veh_value` and `log(claimcst0)~veh_value` (side by side) with the corresponding least squares regression lines.

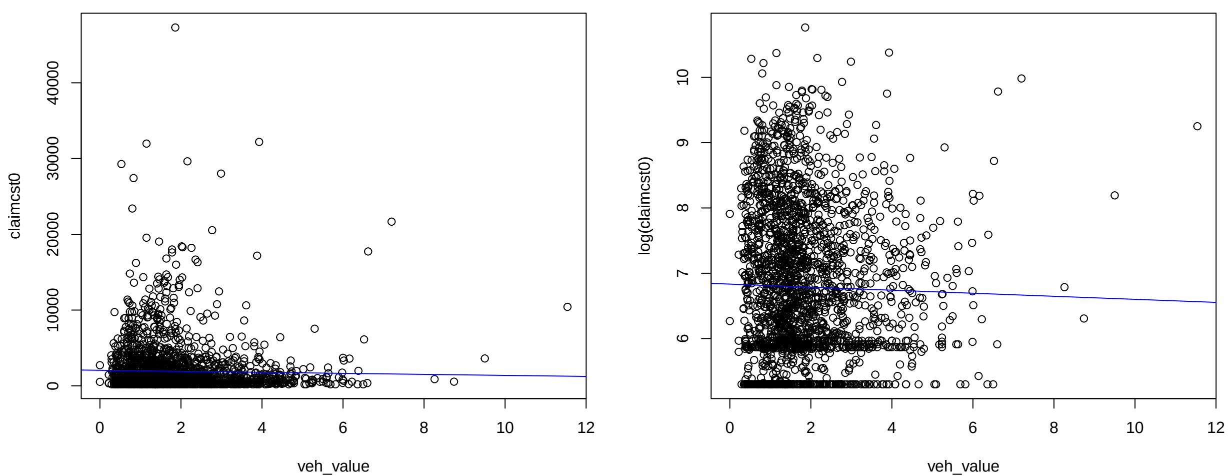

Figure 2: Least Squares Regression Lines; (a) Y = claimcst0 and (b) Y = log(claimcst0)

To lean more about linear regression, visit the website of Introduction to Modern Statistics.