The Bucketing way that I used for each variable:

For the variable: MODEL

I grouped all the MODEL by the alphabet that's mean that I group all the MODEL that starts by the letter A under the group "A", etc.. And for the values of MODEL that start with a numeric number I grouped them under the group "nombre"

For the variable: MAKE

I grouped all the MAKE by the alphabet that's mean that I group all the MAKE that starts by the letter A under the group A, etc..

For the variable: BODY

For example, I grouped PICK UP DOUBLE CABIN and PICK UP TRUCK and PICK UP WITH CRANE and PICKUP WITH BOX & TAIL LIFT and PICKUP WITH TIPPER under the group "PICK UP"

For the variable VEH_SEATS

I created 7 intervals: (0,4] (4,5] (5,6] (6,10] (10,30] (30,60] (60,Inf]

For the variable PREMIUM

I created 3 intervals: (868,1000] (1000,1770] (1770,Inf]

For the variable AGE

I created 6 intervals: (17,23] (23,26] (26,31] (31,44] (44,73] (73,Inf] (60,INF]

For the variable MODEL_YEAR

I created 5 intervals: (1996,2002] (2002,2006] (2006,2010] (2010,2015] (2015,Inf]

For the variable SUM.INSURED

I created 4 intervals: (0,24000] (24000,44884] (44884,87420] (87420,Inf]

For the variable LICENSE_AGE

I created 7 intervals: (0,5] (5,10] (10,20] (20,30] (30,50] (50,60] (60,INF]

Application of the GLM model on the INTIMATED.AMOUNT (in other word application of the GLM model on the severity of the claims)

```
glm_severity_2=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2,
family=Gamma(link ="log"),data=data)
```

AIC = 182443

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2

```
glm_severity_3=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+BODY,family=Gamma(link
="log"),data=data)
```

AIC = 182287

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- 5 components of the variable BODY
- The interval (1.77e+03,Inf] of the variable PREMIUM2

```
glm_severity_4=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+BODY+MAKE.x,family=Gamma(link ="log"),data=data)
```

AIC = 182077

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 4 components of the variable BODY
- 4 components of the variable MAKE

```
glm_severity_5=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+BODY+MAKE.x+MODEL.x,family=Gamma(link ="log"),data=data)
```

AIC = 181966

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 4 components of the variable BODY
- 6 components of the variable MAKE
- 16 components of the variable MODEL

```
glm_severity_6=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE,family=Gamma(link ="log"),data=data)
```

AIC = 182383

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The 2 components of the variable USE_OF_VEH: PRIVATE, PRIVATE/COMERCIAL

P.S: The interval (1.77e+03,Inf] of the variable PREMIUM2 here in this model is significant with a p-value < 0.1

```
glm_severity_7=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS,family=Gamma(link ="log"),data=data)
```

AIC = 182376

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The 2 components of the variable USE_OF_VEH: PRIVATE, PRIVATE/COMERCIAL

P.S: The interval (1.77e+03,Inf] of the variable PREMIUM2 here in this model is significant with a p-value < 0.1

P.S: The component SHARJAH of the variable PLACE.OF..LOSS here in this model is significant with a p-value < 0.01

```
glm_severity_8=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x,family=Gamma(link ="log"),data=data)
```

AIC = 182158

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 5 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- The 2 components of the variable USE_OF_VEH: PRIVATE, PRIVATE/COMERCIAL
- 4 components of the variable MAKE

P.S: The component SHARJAH of the variable PLACE.OF..LOSS here in this model is significant with a p-value < 0.01

```
glm_severity_9=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x,family=Gamma(link ="log"),data=data)
```

AIC = 182044

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 5 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- The 2 components of the variable USE_OF_VEH: PRIVATE, PRIVATE/COMERCIAL
- The component SHARJAH of the variable PLACE.OF..LOSS
- 6 components of the variable MAKE
- 9 components of the variable MODEL

```
glm_severity_10=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY,family=Gamma(link ="log"),data=data)
```

AIC = 181969

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 6 components of the variable MAKE
- 3 components of the variable BODY
- 17 components of the variable MODEL

P.S: The 2 components SHARJAH and DUBAI of the variable PLACE.OF..LOSS here in this model is significant with a p-value < 0.01

```
glm_severity_11=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE,family=Gamma(link ="log"),data=data)
```

AIC = 181455

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 3 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 6 components of the variable MAKE
- 3 components of the variable BODY
- 21 components of the variable MODEL
- The component TP DEATH CLAIM of the variable type

P.S: The 2 components SHARJAH and DUBAI of the variable PLACE.OF..LOSS here in this model is significant with a p-value < 0.01

```
glm_severity_12=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+POLICYTYPE.x,family=Gamma(link ="log"),data=data)
```

AIC = 181456

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 3 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 6 components of the variable MAKE
- 3 components of the variable BODY
- 21 components of the variable MODEL
- The component TP DEATH CLAIM of the variable type

P.S: The 2 components SHARJAH and DUBAI of the variable PLACE.OF..LOSS here in this model is significant with a p-value < 0.01

```
glm_severity_13=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+POLICYTYPE.x+PRODUCT.x,family=Gamma(link ="log"),data=data)
```

AIC = 181305

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 3 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 7 components of the variable MAKE
- 3 components of the variable BODY
- 20 components of the variable MODEL
- The component TP DEATH CLAIM of the variable type
- 7 components of the variable PRODUCT

```
glm_severity_14=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+POLICYTYPE.x+PRODUCT.x+REGN,family=Gamma(link ="log"),data=data)
```

AIC = 181311

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 3 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 7 components of the variable MAKE
- 3 components of the variable BODY
- 20 components of the variable MODEL
- The component TP DEATH CLAIM of the variable type
- 7 components of the variable PRODUCT

```
glm_severity_15=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+POLICYTYPE.x+PRODUCT.x+REGN+VEH_SEATS,family=Gamma(link ="log"),data=data)
```

AIC = 181046

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 3 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (1.77e+03,Inf] of the variable PREMIUM2
- 7 components of the variable MAKE
- 4 components of the variable BODY
- 18 components of the variable MODEL
- The component TP DEATH CLAIM of the variable type
- 7 components of the variable PRODUCT

```
glm_severity_16=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+PLACE.OF
..LOSS+MAKE.x+MODEL.x+BODY+REGN+VEH_SEATS,family=Gamma(link ="log"),data=data)
```

AIC = 181701

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- 6 components of the variable MAKE
- 3 components of the variable BODY
- 15 components of the variable MODEL

P.S: The interval (1.77e+03,Inf] of the variable PREMIUM2 here in this model is significant with a p-value < 0.1

P.S: The interval (6,10] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

```
glm_severity_17=glm(INTIMATED.AMOUNT~SUM.INSURED.x+AGE+MODEL_YEAR+PREMIUM2+USE_OF_VEHICLE+MAKE.x+
MODEL.x+BODY+REGN+VEH_SEATS,family=Gamma(link ="log"),data=data)
```

AIC = 181698

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- 7 components of the variable MAKE
- 3 components of the variable BODY
- 15 components of the variable MODEL

P.S: The interval (6,10] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

P.S: The 2 components FUJAIRAH and DUBAI of the variable REGN here in this model is significant with a p-value < 0.1

P.S: The interval (1.77e+03,Inf] of the variable PREMIUM2 here in this model is significant with a p-value < 0.1

```
glm_severity_21=glm(INTIMATED.AMOUNT~BODY+MAKE.x+MODEL.x+REGN+POL_EFF_DATE+POL_EXPIRY_DATE+MODEL_YEAR+SUM.INSURED.x+AGE+PLACE.OF..LOSS+DRV_DLI+DRV_DOB+VEH_SEATS+DATE.OF..ACCIDENT+LICENSE_AGE+POLICYTYPE.x+PRODUCT.x+TYPE+USE_OF_VEHICLE+PREMIUM2,family=Gamma(link ="log"),data=data)
```

AIC = 180575

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable SUM.INSURED: (44884,87420] (87420,Inf]
- The 2 intervals of the variable AGE: AGE(31,44], AGE(44,73]
- 7 components of the variable MAKE
- 4 components of the variable BODY
- 19 components of the variable MODEL
- 7 components of the variable PRODUCT
- The component TP DEATH CLAIM of the variable type
- The interval (1.77e+03,Inf] of the variable PREMIUM2

So as a conclusion I think the most important variable for the estimation of the severity of the claims are:

1. SUM.INSURED
2. AGE
3. MAKE
4. MODEL
5. BODY
6. PRODUCT
7. TYPE
8. PREMIUM2
9. USE_OF_VEH
10. PLACE.OS..LOSS
11. MODEL_YEAR

P.S: The model “glm_severity_21” has the smallest AIC

Application of the GLM model on the NB (in other word application of the GLM model on the frequency of the claims)

```
glm_freq_1=glm(NB~BODY+MAKE.x+MODEL.x+REGN+POL_EFF_DATE+POL_EXPIRY_DATE+MODEL_YEAR+SUM.INSURED.x+AGE+PLACE.OF..LOSS+DRV_DLI+DRV_DOB+VEH_SEATS+DATE.OF..ACCIDENT+LICENSE_AGE+POLICYTYPE.x+PRODUCT.x+TYPE+USE_OF_VEHICLE+PREMIUM2,family=poisson(link ="log"),data=data)
```

AIC = 26941

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable LICENSE_AGE: (10,20] (20,30]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- 2 components of the variable REGN
- The interval (6,10] of the variable VEH_SEATS

P.S: 3 components of the variable MAKE here in this model is significant with a p-value < 0.1

```
glm_freq_11=glm.nb(NB~BODY+MAKE.x+MODEL.x+REGN+POL_EFF_DATE+POL_EXPIRY_DATE+MODEL_YEAR+SUM.INSURED.x+AGE+PLACE.OF..LOSS+DRV_DLI+DRV_DOB+VEH_SEATS+DATE.OF..ACCIDENT+LICENSE_AGE+POLICYTYPE.x+PRODUCT.x+TYPE+USE_OF_VEHICLE+PREMIUM2,data=data)
```

AIC = 26943

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable LICENSE_AGE: (10,20] (20,30]
- The 4 intervals of the variable AGE: AGE(26,31], AGE(31,44], AGE(44,73], AGE(73,Inf]
- 2 components of the variable REGN
- The interval (6,10] of the variable VEH_SEATS
- 2 components of the variable BODY

P.S: 3 components of the variable MAKE here in this model is significant with a p-value < 0.1

```
glm_freq_2=glm(NB~BODY+MAKE.x+REGN+POL_EFF_DATE+POL_EXPIRY_DATE+MODEL_YEAR+SUM.INSURED
.x+AGE+PLACE.OF..LOSS+DRV_DLI+DRV_DOB+VEH_SEATS+DATE.OF..ACCIDENT+LICENSE_AGE+POLICYTYPE.x+
PRODUCT.x+TYPE+USE_OF_VEHICLE+PREMIUM2,family=poisson(link ="log"),data=data)
```

AIC = 26936

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 2 intervals of the variable LICENSE_AGE: (10,20] (20,30]
- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- 1 components of the variable REGN
- 4 components of the variable BODY
- 2 components of the variable MAKE

P.S: The interval (6,10] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

P.S: 3 components of the variable MAKE here in this model is significant with a p-value < 0.1

```
glm_freq_3=glm(NB~BODY+MAKE.x+REGN+MODEL_YEAR+SUM.INSURED.x+AGE+PLACE.OF..LOSS+VEH_SEATS
+LICENSE_AGE+POLICYTYPE.x+PRODUCT.x+TYPE+USE_OF_VEHICLE+PREMIUM2,family=poisson(link
="log"),data=data)
```

AIC = 27036

The significant variables (with a p-value<0.05) of this model are:

- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (20,30] of the LICENSE_AGE
- 2 components of the REGN
- 3 components of the variable BODY
- 3 components of the variable MAKE

P.S: The interval (10,30] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

```
glm_freq_4=glm(NB~BODY+MAKE.x+REGN+MODEL_YEAR+SUM.INSURED.x+AGE+PLACE.OF..LOSS+VEH_SEATS
+LICENSE_AGE+TYPE+USE_OF_VEHICLE,family=poisson(link ="log"),data=data)
```

AIC = 27123

The significant variables (with a p-value<0.05) of this model are:

- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- The interval (20,30] of the LICENSE_AGE
- 3 components of the variable BODY
- 2 components of the variable MAKE
- 2 components of the REGN

P.S: The interval (10,30] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

```
glm_freq_5=glm(NB~BODY+MAKE.x+REGN+MODEL_YEAR+AGE+PLACE.OF..LOSS+VEH_SEATS+LICENSE_AGE+
USE_OF_VEHICLE,family=poisson(link ="log"),data=data)
```

AIC = 27357

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- 3 components of the variable BODY
- 2 components of the variable MAKE
- 2 components of the variable REGN

P.S: The interval (10,30] of the variable VEH_SEATS here in this model is significant with a p-value < 0.1

```
glm_freq_6=glm(NB~AGE+MODEL_YEAR+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+
POLICYTYPE.x+PRODUCT.x+REGN,family=poisson(link ="log"),data=data)
```

AIC = 27111

The significant variables (with a p-value<0.05) of this model are:

- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- 4 components of the variable BODY
- 2 components of the variable MAKE
- 2 components of the variable REGN

```
glm_freq_6=glm(NB~AGE+MODEL_YEAR+USE_OF_VEHICLE+PLACE.OF..LOSS+MAKE.x+MODEL.x+BODY+TYPE+POLICYTYPE.x+PRODUCT.x+REGN,family=poisson(link ="log"),data=data)
```

AIC = 27155

The significant variables (with a p-value<0.05) of this model are:

- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- 4 components of the variable BODY
- 2 components of the variable MAKE
- 2 components of the variable REGN

P.S: The intercept here in this model is significant with a p-value < 0.1

```
glm_freq_8=glm(NB~AGE+LICENSE_AGE+MODEL_YEAR+MAKE.x+BODY+REGN+USE_OF_VEHICLE,family=poisson(link ="log"),data=data)
```

AIC = 27369

The significant variables (with a p-value<0.05) of this model are:

- The intercept
- The 4 intervals of the variable AGE: AGE(23,26], AGE(26,31], AGE(31,44], AGE(44,73]
- 4 components of the variable BODY
- 2 components of the variable MAKE
- 2 components of the variable REGN

CONCLUSION

So as a conclusion I think the most important variable for the estimation of the frequency of the claims are:

1. AGE
2. REGN
3. LICENSE_AGE (DRIVING..LICENSE.ISSSUE)
4. BODY
5. MAKE
6. VEH_SEATS

P.S: The model “glm_freq_2” has the smallest AIC