ST3002: Lab 8

Ross Finnegan, 15320532

Part 1.

Figure 1 outlines the Severity of the Claims in euro by Age (left side) and by Vehicle Use (right side). The age of drivers has been divided into eight different age brackets labelled A to H (A is 17-20 and the getting older). Each box is divided by a solid black line which marks the midpoint of the data for that division. One important takeaway from the Severity of Claims by Age graphic is that the median value does not vary dramatically between each age bracket. The interquartile range is contained within the coloured boxes. The box with the most significant spread is the boxplot for Age Group A. This age has a large deviation from the median to its upper quartile which indicates that the cost of claims for 50% of customers in comparison to the next 25% rises significantly.

The Severity of Claims by Vehicle Use graph shows the steady decline of value of the claim from Business vehicles to the other car types. The circles above and below the Business plot indicates outliers in this data. The upper outlier is the Business Car Types associated with customers in the A Age Group. There were only 5 claims in this category last year so this may skew results. The lower outlier is the Business Car Types associated with customers in the E Age Group. On this occasion there were 166 claims. Some investigation could be done to understand the reason for this dip in claim value.

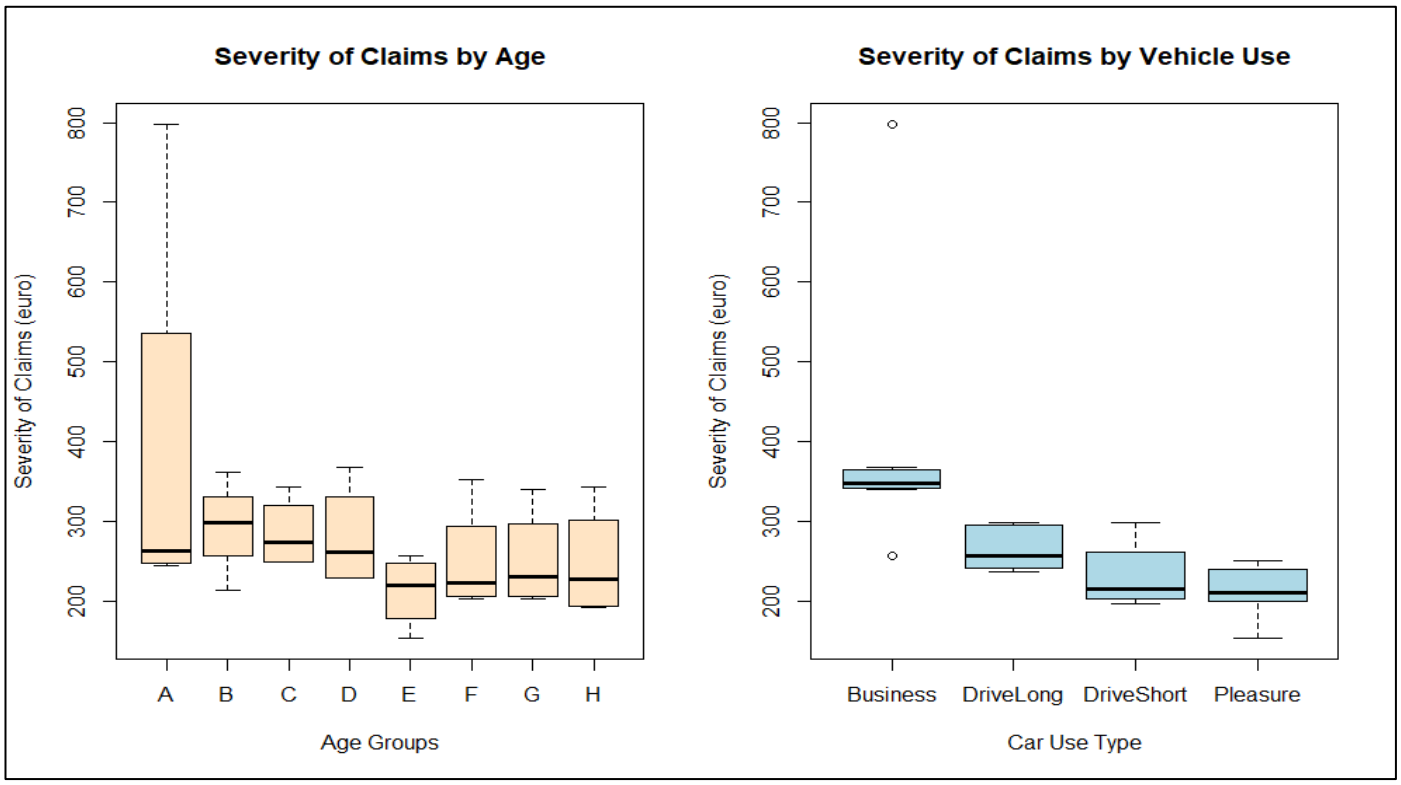

Figure 1: Severity Analysis

Part 2.

Figure 2 outlines the Number of Claims by Age (left side) and by Car Use Type (right side). It is very apparent that the number of claims varies dramatically depending on the Age Group. It is evident that the youngest customers (Group A) are very unlikely to claim. This allows Figure 1 to make sense as those in the Age Group A are only likely to claim when the value is quite high. I would anticipate that this is to avoid an increase in their already expensive premium as young drivers. It should also be noted that the number of claims has a far greater spread than the value of the claim for most age groups. This would lend me to believe that greater analysis should be made to investigate the factors affecting the number of claims. There is also a large disparity between the number of claims by vehicle use. One interesting takeaway is that those who drive their vehicles to work appear to be more likely to claim.

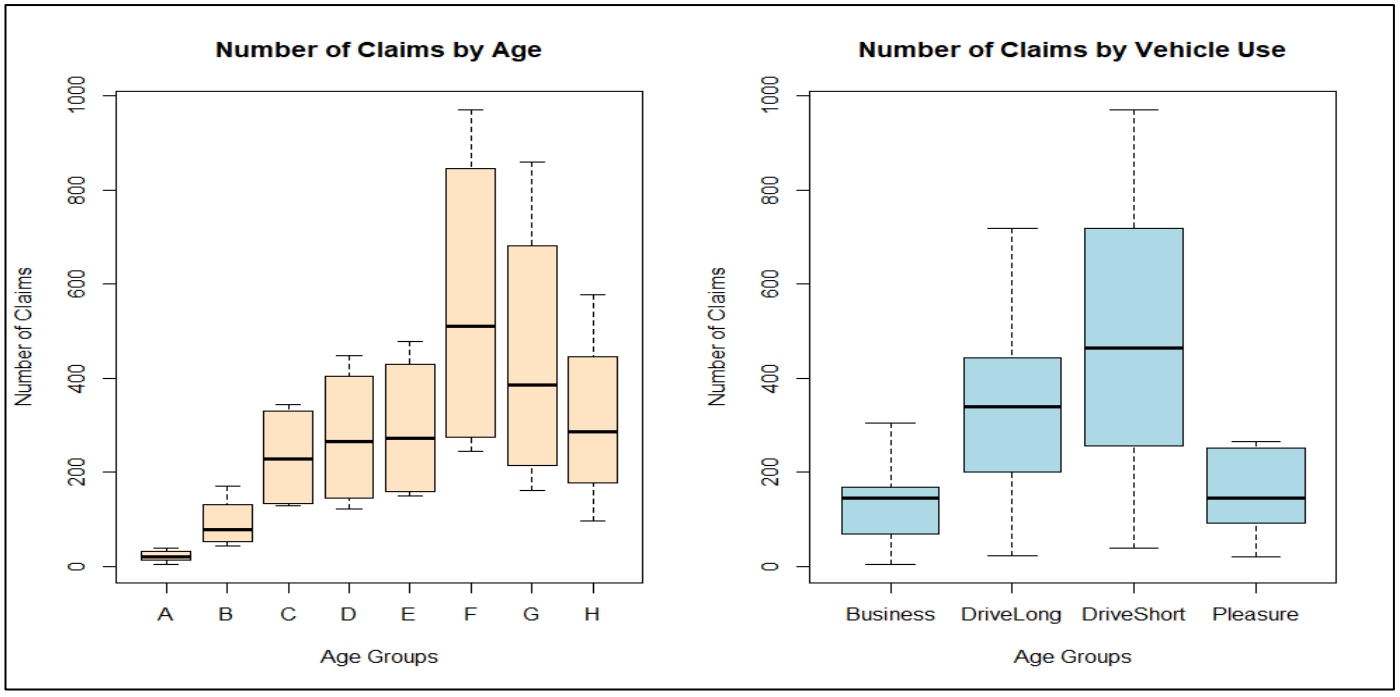

Figure 3: Number of Claims

Part 3.

A predictive model has been produced to help with indemnity planning for next year. A fitted generalised linear model was used to model predictions. These values are then transformed to allow for predictions to be on the correct scale (Figure 3). Figure 4 outlines the model used and can be recreated should a pool of test data be provided.

Age	Vehicle_Use	Claim_Count	Predicted_Count
A	Pleasure	21	12
A	DriveShort	40	38
A	DriveLong	23	26
A	Business	5	10
B	Pleasure	63	52
B	DriveShort	171	160
B	DriveLong	92	112
B	Business	44	44

Age	Vehicle_Use	Claim_Count	Predicted_Count
C	Pleasure	140	131
C	DriveShort	343	404
C	DriveLong	318	281
C	Business	129	111
D	Pleasure	123	156
D	DriveShort	448	478
D	DriveLong	361	333
D	Business	169	132

Age	Vehicle_Use	Claim_Count	Predicted_Count
E	Pleasure	151	167
E	DriveShort	479	511
E	DriveLong	381	356
E	Business	166	141
F	Pleasure	245	317
F	DriveShort	970	973
F	DriveLong	719	678
F	Business	304	269

Age	Vehicle_Use	Claim_Count	Predicted_Count
G	Pleasure	266	254
G	DriveShort	859	778
G	DriveLong	504	542
G	Business	162	215
H	Pleasure	260	176
H	DriveShort	578	541
H	DriveLong	312	377
H	Business	96	149

Figure 2: Predicted Claim Count

Part 4.

Number of things that could be done to improve the model in the next few months:

1. Data regarding the value and age of the vehicle claimed on may impact a policy holder's decision whether or not to claim.
2. Greater investigation into the what motivates customers to claim will allow for more accurate predictive models to be produced.
3. As always, a greater sized sample would improve the accuracy of the model.
4. Policy holders' income details may also drive new insights as the value of the damage outweighing their salary may increase the likelihood of a claim.

```
Call:
glm(formula = Claim_Count ~ Age + Vehicle_Use, family = poisson,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7079  -1.7508   0.4497   1.5586   5.8413

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.37020    0.10979  21.588  < 2e-16 ***
AgeB               1.42487    0.11806  12.069  < 2e-16 ***
AgeC               2.34655    0.11096  21.148  < 2e-16 ***
AgeD               2.51534    0.11020  22.825  < 2e-16 ***
AgeE               2.58209    0.10993  23.488  < 2e-16 ***
AgeF               3.22470    0.10809  29.834  < 2e-16 ***
AgeG               3.00189    0.10860  27.641  < 2e-16 ***
AgeH               2.63906    0.10972  24.053  < 2e-16 ***
Vehicle_UseDriveLong   0.92463    0.03604  25.652  < 2e-16 ***
Vehicle_UseDriveShort  1.28557    0.03446  37.307  < 2e-16 ***
Vehicle_UsePleasure    0.16591    0.04145   4.002 6.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6064.97  on 31  degrees of freedom
Residual deviance:  184.72  on 21  degrees of freedom
AIC: 430.81
```

Figure 4: Predictive Model