**Run an Exploratory data analysis (EDA) to get preliminary impressions about the associations (report associated plots to substantiate your answers). Which associations do you expect, and in which directions?**

The first graph analysis run was a histogram the frequency of both smoking and number of glasses of alcohol consumed per week. The histograms are both right skewed, the majority of the data points are in the lower range. Next a count of male and female participants was conducted. There is a visually higher number of male participants. A count of the data reveals there are 173 more male participants. This would improve the accuracy of predictions in males, but it may also introduce possible confounding factors associated with males specifically.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

The spread of the data was segregated by sex and analysed. In both variables drinking and smoking females had a lower mean, median and smaller standard deviations than males. This would be indicative that males on average consume more cigarettes and alcohol than females but there is more variation in the datapoints contributing to this result. If we look closer into the spread of the data the median and mean of female packYears is quite different indicating a non-normal distribution (4 and 11 respectively, rounded) but this isn’t present in the glass\_week variable (7 and 9). This could be due to the missing data points. Although the male results have a larger variation the data is relatively normal in its distribution when looking at the mean and median.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sex | **packYears**.mean | **packYears**.sd | **packYears**.median |
| 1 | female | 11.34645 | 15.70678 | 3.6 |
| 2 | male | 20.64162 | 20.02877 | 16.8 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sex | **glass\_week**.mean | **glass\_week**.sd | **glass\_week**.median |
| 1 | female | 8.94703 | 7.462796 | 7 |
| 2 | male | 14.96734 | 10.968173 | 12 |

There is also 108 missing data points in the packYears column that reduces our confidence in the results, 46 in females and 62 in males. Careful consideration with what to do with those data points is needed. They could be removed or imputed, for simplicity the values were removed with the understanding that this will impact the power of the results.

To observe any trends in the data a scatter plot was created and a smooth line(left) was added to capture any non-linear trends and a linear regression line(right) on a separate graph.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

We can see that in the smooth line curve on the left that the data is not quite monotonastic but isn’t overly concerning. As the quantity of packYears increases the line starts to curve downwards in a sort of arc. The model may be over fitting the data or outliers are pulling the line down, but this isn’t too dramatic. There is also an increase in the variance indicated by the grey around the line diverging meaning we may not be as confident when we reach higher levels of packYears. With this in mind a linear model is likely suitable. The linear model on the right gives us an idea of the linear trends in the data i.e there is a positive relationship between the variables. These graphs do not take the possible effect of gender into account. They were repeated to do such.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

When stratified for gender we can again see the downward arc in the smooth line curve however it is much more dramatic for females rather than males. Also at the beginning of the lines there is a bump in both male and females in opposite directions. The model is possibly overfitting the data in this region and outliers are pulling it down when packyears values increase. The variance seen in the model also increases dramatically when seggregated for sex, being much higher for females. The straight line model may not be as appropriate when segregated as the monotanicity isn’t held, however the linear model again confirms the positive relationship between the variables in both sexes.

If we look at a visual comparison of the means:

Chart, scatter chart, qr code

Description automatically generated Chart, scatter chart

Description automatically generated

Again we can see that in both cases of smoking and alcohol consumption men have a higher mean than females indicated in the male regression line being above the females. We can also see the relative concentration of the female points being in the lower limits with a few outliers in the upper limits. These could be the points responsible for the drag down of the smooth line fit.

**Smoking is a count variable. Can you add it to the model in its original scale, i.e. does linearity/monotonicity hold for both variables? Motivate your answer;- *pack years vs glass week by sex***

No it cannot be added in its original scale as it is a count variable. It needs to be modelled using Poisson. If we didn’t we would be assuming the linearity and monotonicity of the relationship between the dependent and independent variables. Further we would be assuming a change in the dependent variable with each unit increase of the independent variable is constant which may not be accurate in the case of count data. When we looked at the scatter plots the data appears to meet these assumptions but when we separate for sex it is clear that this is not the case.

**Explain why the use of the linear regression models may not be the adequate choice to analyse this dataset;**

Linear regression models have certain assumptions that need to be met in order to be considered reliable. The data violates these assumptions:

**Linearity:** There must be a linear relationship between the variables or interest. As we can see in the scatter plots with the smooth line this doesn’t look to be the case, monotonicity is not achieved so the relationship may not be parametric.

**Variance:** should be equal at all levels of the data which isn’t present in this data set. In both smooth line graphs with and without sex separations the variance increases as we reach the higher values of drinking vs smoking. This is also present in the linear models. This data could be considered heteroscedastic and fails this assumption.

**Distribution:** The data must also be normally distributed. This is the case in three groupings as seen in the summary stats but the female pack years does not appear to be normally distributed.

**Outliers:** Visually on the scatter plots there are outliers in pack years, we can all see this in the histograms with long tails. These outliers can impart a bias on the data if a linear model was used.

**Run simple Poisson regression models with all explanatory variables, one at a time. Are they significantly linked to the outcome? Interpret the associated effects;**

The models were built with glass\_week as the dependent variable. One model was built with sex as the independent variable and the second with packYears.

* Model 1: This model reported a statistically significant relationship between gender and the number of drinks consumed (P= <2e-16). The coefficient for the male gender is 0.51455. The confidence interval does not cross 0 (sexmale: 0.4828033-0.5464163) indicating a significant difference from 0, i.e. there is a significant relationship between the variables
* Ln(drinks/day) pred = 2.2 + 0.51\*sex.
* Antilog: 1.556 = 56%.
* This means that in the simple model males drink 56% more than females in glass per week.
* Model 2: This model reported a statistically significant relationship between packYears and the number of drinks consumed (p= <2e-16). The coefficient for packYears is 0.01. The confidence interval does not cross 0 (packYears: 0.009386967-0.01079321) indicating a significant difference from 0, i.e. there is a significant relationship between the variables.
* Ln(drinks/day) pred = 2.3 + 0.01\*packYears
* Antilog: 1.025 = 2.5%
* This means that in the simple model for every unit increase of packYears drinks increase by 2.5%.
* The model will estimate the value of the intercept and slope coefficient, which represents the expected log count of glass\_week for a one-unit increase in packYears.

The AIC values for both models are very similar indicating that both have relatively the same predictive power and fit of the data (model 1: 14980, model 2: 14209).

**Run a multiple model with all Xs (mutually adjusted for each other). Write down the estimated multiple model in its full extension (the equation);**

When we run the multiple model both packYears and the male gender are statistically significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 2.1101804 | 0.0147119 | 143.43 | <2e-16 \*\*\* |
| packYears | 0.0078011 | 0.0003752 | 20.79 | <2e-16 \*\*\* |
| sexmale | 0.4282911 | 0.0172332 | 24.85 | <2e-16 \*\*\* |

* Ln(drinks/day) pred = 2.1 + 0.01\*packYears + 0.43\*sex.

**Draw conclusions regarding observed associations, without forgetting to interpret them one more time. Are there differences between the crude (question 4) and adjusted (question 5) associations? If yes, what would be the possible explanations, i.e. what is the third variable problem in the data, confounding or interaction? (Please note that in the latter case, the adjusted model needs to be extended to consider an interaction term);**

The model was reconsidered with the possibility of either a confounding or interaction effect.

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 2.0582827 | 0.0175543 | 117.252 | < 2e-16 \*\*\* |
| packYears | 0.0114790 | 0.0007355 | 15.607 | < 2e-16 \*\*\* |
| sexmale | 0.5085661 | 0.0224343 | 22.669 | < 2e-16 \*\*\* |
| packYears:sexmale | -0.0048689 | 0.0008544 | -5.699 | 1.21e-08 \*\*\* |

* Ln(drinks/day) pred = 2.05 + 0.01\*packYears + 0.5\*sexmale – 0.005\*packYears:sexmale.

From this we can see that there is an interaction effect between the two predictor variables. The packyears:sexmale is statistically significant indicating that there is a relationship between being male and the number of cigarettes consumed, i.e. the effect of packYears on glass\_week is different for males and females. Since this value is negative this means the effect is weaker for males than females. The effect of packYears is dependent on sex thus indicating an interaction effect.

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| (Intercept) | 2.023726047 | 2.092538633 |
| packYears | 0.010025796 | 0.012908899 |
| sexmale | 0.464656199 | 0.552598914 |
| packYears:sexmale | -0.006535783 | -0.003186542 |

The confidence intervals of the model are also very narrow and do not cross zero. This also indicates that these variables are significant and that there is an interaction effect between being male and smoking habits.

**Once selecting the final model, check its goodness of fit (GOF). Did you encounter any over-dispersion problem? Motivate your answer;**

A chi squared test was used to determine whether a model with fewer predictors (the null model) fits the data as well as a more complex model with additional predictors which in this case were packYears and sex. The result of the chi square on the complex model including accounting for the interaction effect was 0 meaning the model was significant in its predicitions.

However, looking further into the models deviance is eight times larger (8216.3) than the 95% critical value (1392.2). Meaning that the model is not good at explaining the majority of variation in the data and it does not capture the relationship between the variables. This was run on the model that accounts for the interaction effect so there may be another factor not accounted for.

Further the deviance vs the degrees of freedom are imbalanced. 8216.3:1307, the variance is 6.9 times larger than the mean. In poisson we would expect this value to be 1. This high value of deviance would be indicative of over dispersion.

**If yes (question 7), re-run the model again, this time using the ‘*quasipoisson*’ family. Are the conclusions the same as the ones from “*Poisson*” model?**

The model was re-run with quasipoisson as the family which allows for the mean and the deviance to differ unlike the poisson family. Using this approach has changed the significance of the variables. While all still significant the interaction effect of being male and smoking habits becomes weaker and almost falls out of significance. The original model thus may not be able to account for the dispersion seen in the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 2.058283 | 0.046059 | 44.688 | < 2e-16 \*\*\* |
| packYears | 0.011479 | 0.001930 | 5.948 | 3.47e-09 \*\*\* |
| sexmale | 0.508566 | 0.058863 | 8.640 | < 2e-16 \*\*\* |
| packYears:sexmale | -0.004869 | 0.002242 | -2.172 | 0.03 \* |

The confidence intervals of the model also become broader using this model. This is understandable as the model accounts for more of the dispersion in the data.

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| (Intercept) | 1.966964744 | 2.1475305988 |
| packYears | 0.007616056 | 0.0151820462 |
| sexmale | 0.393603795 | 0.6243745926 |
| packYears:sexmale | -0.009210627 | -0.0004204311 |

The final model equation:

* Ln(drinks/day) pred = 2.06 + 0.01\*packYears + 0.5\*sexmale -0.005\*packYears\*sexmale.

Overall quasipoisson had very little effect on the values of the coefficients but reduced the significance and confidence in the values. This approach however did not help in accounting for the majority of the variance as the 95% critical value was 1392.2 which is marginally higher. The deviance of this model was 8216.3. This dispersion of this model is still very high at 6.9.

The intercept value of 2.06 is what the log count we would expect for non-smoker females. The packYears (0.01) is the expected change in log count of glass\_week for each one unit increase of packYears when all other variables are held constant. sexmale (0.5) is the log difference between sexes when all other variables are held to zero. Finally, the interaction account dictates the log count change in glass\_week related to packYears but this value is to be reduced (-0.009) when the sex is male.

**Write down a brief paragraph describing the objective of your investigation and main conclusions based on your analyses (no statistical jargon, please- just a summary of the results in plain and accessible language);**

The objective of this study was to investigate if there was a relationship between gender and smoking habits in the amount of alcohol consumed on a weekly basis.

From the analysis there was a significant association between the predictor variable smoking habits and the outcome variable of drinking consumption. This could be due to a cultural link between smoking and alcohol consumption with the notion of social smoking in locations of social gatherings like pubs and nightclubs. People consuming alcohol in these locations will often smoke in order to partake in conversations as the group moves outdoors to do so. This may affect the assumption of independence of the variables.

Sex was also a significant predictor of alcohol consumption but there was also an interaction effect, meaning that being male altered the effect of smoking habits on the predicted alcohol consumption. When sex and smoking habits are considered together the amount of alcohol predicted to be consumed on a weekly basis for a male needs to be reduced. Meaning that increase in alcohol consumption does not increase as steeply for males as it does for females as seen in the graph below. Males may consume more alcohol but the amount increases steadily compared to females as smoking increases. Looking at the scatter plot the smooth line for females climbs at a steeper rate than males towards the end of the graph. This could be due to outliers in the female related data points with a few points varying quite drastically from where the bulk of female data points are situated.

Chart, scatter chart

Description automatically generated

Overall the data reported was quite varied and contained missing datapoints which reduces our confidence in the results we have obtained so results must be considered with caution.

**Appendix**

**Q1.**

**Text

Description automatically generatedText

Description automatically generated**

**Text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Q4**

**Graphical user interface, text, email

Description automatically generated**

**Q5**

**Text

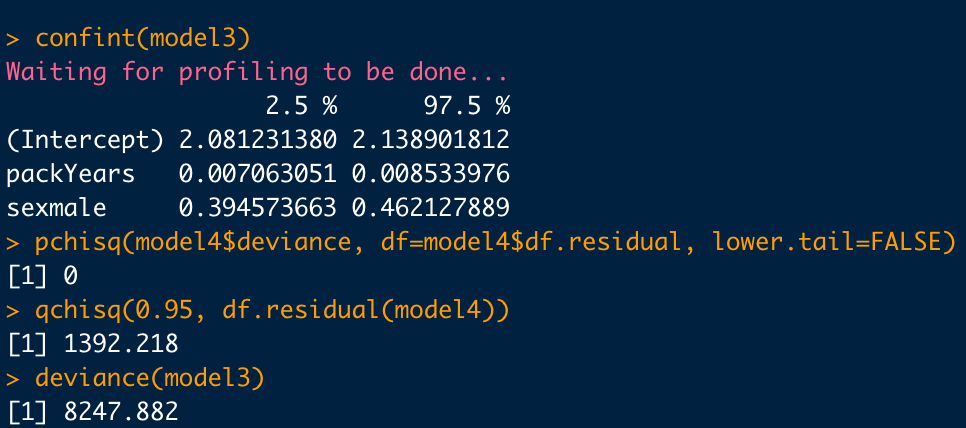
Description automatically generated**

**Q6**

**Text

Description automatically generated**

**Q7**

****

**Q8**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Q9**

**Graphical user interface, text

Description automatically generated**