**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 out 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 and a linear regression line(right) on a separate graph.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

On the left we can see that the monotonicity of the data holds. There is a continuous increase in the data that as alcohol consumption increases so does smoking. Towards the end of this line the spread in the variance deviated a little meaning we may not be as confident in the upper limits of the data. There is no drastic variation in the smooth line meaning the model isn’t overfitting the data either. We can see in the straight line model that the data trends upwards again.

This was repeated but segregated for sex.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

In the smooth line model the monotanicity only holds for males and not females. This graph suggests that as alcohol consumption increases for females the consumption of cigarettes decreases. This is contrary to the model above. This could be due to having a higher number of male participants confounding the data. The linear model tells us that overall on average the data does trend upwards. ???

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.

**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:** is one of the assumptions and as we saw from the scatter plot of glass\_week vs packYears by sex that does not give us a linear relationship and fails that assumption.

**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. The variance is tight initially but when we reach 20 glasses per week the variation begins to spread. This data could be considered heteroscedastic and fails this assumption.

**Distribution:** The data must also be normally distributed. This is the case in three contexts 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 combine