Q2: What is the relationship or association between first period exam and absences?

Answer:

- two variables: G1 and absence
- To see the relationship between the variables we perform a correlation test. Before that, we check for normality using Shapiro-Wilkins test.
- P-value for G1 < 2.2e-16 < 0.05. **G1 is not normal**
- P-value for absence = 2.454e-06 < 0.05. **absence is not normal.**
- Since both variables are not normal, we use **spearman correlation.**
- Using Spearman Correlation Test, we find that:
  - P-value = 0.9293 > 0.05, hence G1 and absence are not statistically significant. One doesn't affect the other. Also, rho (correlation coefficient) = 0.0044 indicating there is no significant association.
  - **Hence there is no relationship between the two variables**

```
#Q2: What is the relationship or association between first period exam (G1) and absebces (absences)

# Use correlation test, but check for normality before

shapiro.test(df$absences) #p-value < 2.2e-16 < 0.05. data not normal

shapiro.test(df$G1) #p-value = 2.454e-06 < 0.05. data not normal

# since data not normal use spearman correlation

cor.test(df$absences, df$G1, alternative = "two.sided", method = "spearman", conf.level = 0.95)
```

```
> shapiro.test(df$absences) #p-value < 2.2e-16 < 0.05. data not normal

        Shapiro-Wilk normality test

data:  df$absences
W = 0.66683, p-value < 2.2e-16

> shapiro.test(df$G1) #p-value = 2.454e-06 < 0.05. data not normal

        Shapiro-Wilk normality test

data:  df$G1
W = 0.97491, p-value = 2.454e-06

> cor.test(df$absences, df$G1, alternative = "two.sided", method = "spearman", conf.level = 0.95)

        Spearman's rank correlation rho

data:  df$absences and df$G1
S = 10225570, p-value = 0.9293
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.004479359
```

Q3: Visualize the correlations between second exam and family relations and absences.

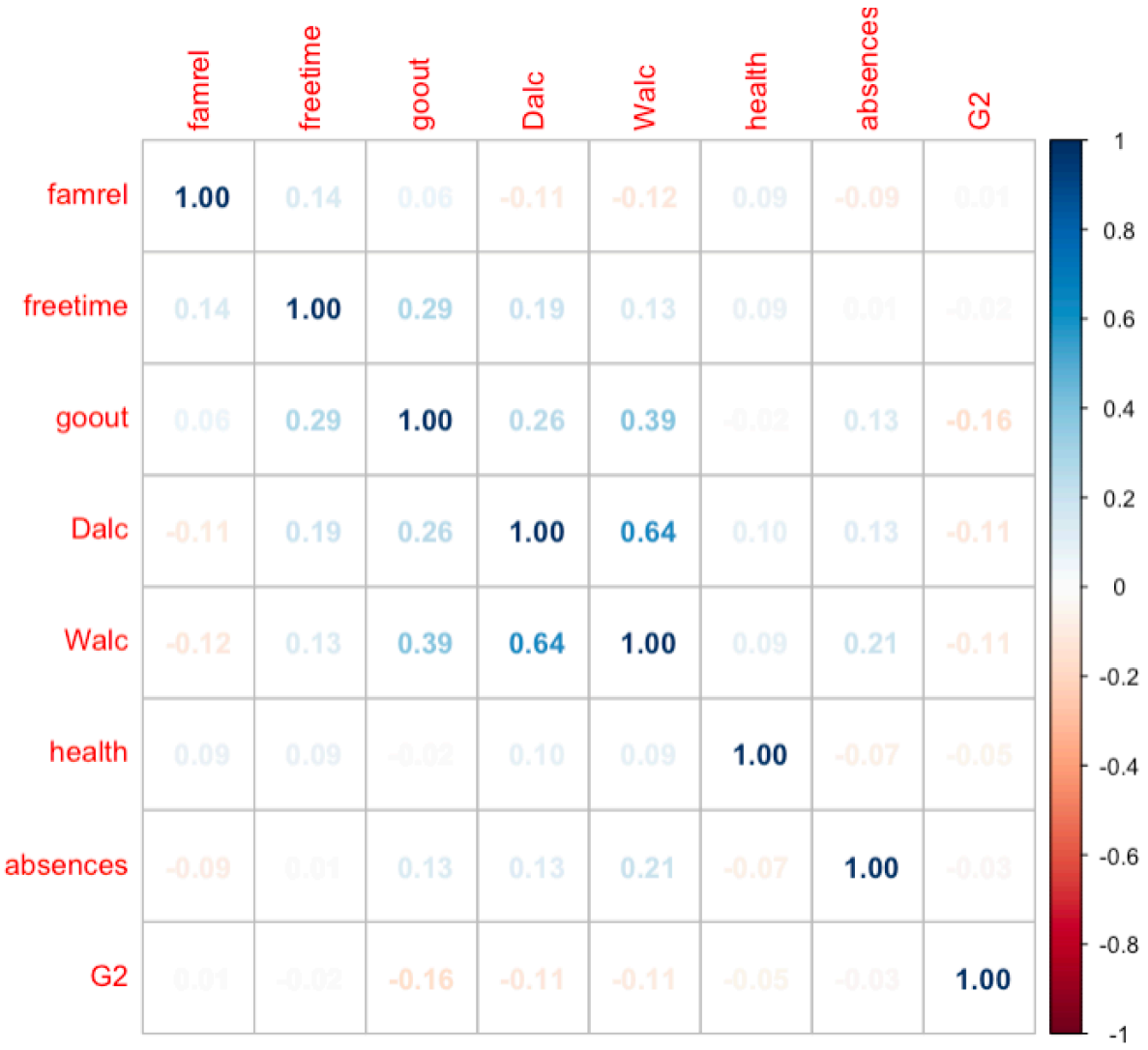
Answer:

- Variable sets: (G2, famrel, absences)
- The answer for if we plotted them separately (G2 and famrel) and (G2 and absences) would be similar, except here we get an extra correlation result between (famrel and absences).
- Before deciding which correlation test to choose, we check for normality.
  - P-value for absences < 2.2e-16 < 0.05. absences is not normal.
  - P-value for famrel < 2.2e-16 < 0.05. famrel is not normal
  - P-value for G2 = 2.084e-07 < 0.05. G2 is not normal.
- Since all three variables are not normal, we choose **Spearman correlation** test and visualize the correlation coefficients.

```
#Q3: Visualize the correlations between second exam (G2) and family relations to absences

library(corrplot)

names(df)

df_g2 <- df[, c(24:30, 32)]
corrplot(cor(df_g2, method = "spearman"), method="number")
```

```
> library(corrplot)
> names(df)
 [1] "school"     "sex"        "age"        "address"    "famsize"    "Pstatus"    "Medu"       "Fedu"       "Mjob"
[10] "Fjob"       "reason"     "guardian"   "traveltime" "studytime"  "failures"   "schoolsup"  "famsup"     "paid"
[19] "activities" "nursery"    "higher"     "internet"   "romantic"   "famrel"     "freetime"   "goout"      "Dalc"
[28] "Walc"       "health"     "absences"   "G1"         "G2"         "G3"
> df_g2 <- df[, c(24:30, 32)]
> corrplot(cor(df_g2, method = "spearman"), method="number")
```

Correlation Plot



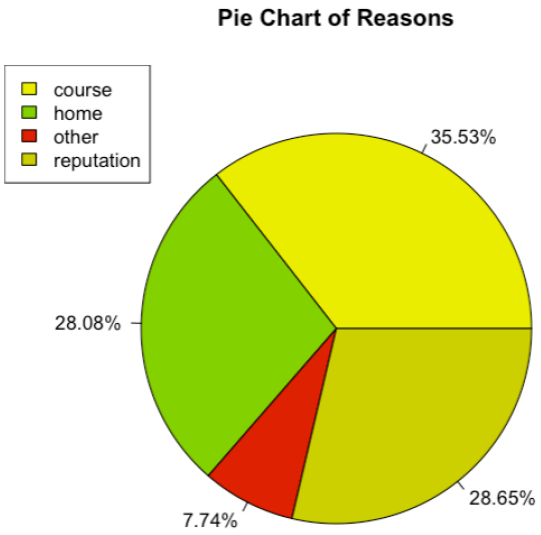
- From the plot above, we can see there is little to no association between the three variables.
- G2 and the following variables have no association-famrel(correlation coefficient = 0)
- G2 and no other variables have strong positive or negative correlations
- G2 and the following variables have **weak negative correlation**:
  - goout: -0.16 -> increase in going out with friends leads to decrease in G2
  - Dalc: -0.11 -> increase in workday alcohol consumption leads to decrease in G2
  - Walc: -0.11 -> increase in weekend alcohol consumption leads to decrease in G2
  - health: -0.05 -> increase in health status leads to decrease in G2
  - absences: -0.03 -> increase in absences leads to decrease in G2
  - Freetime : -0.02 -> increase in free time leads to decrease in G2'

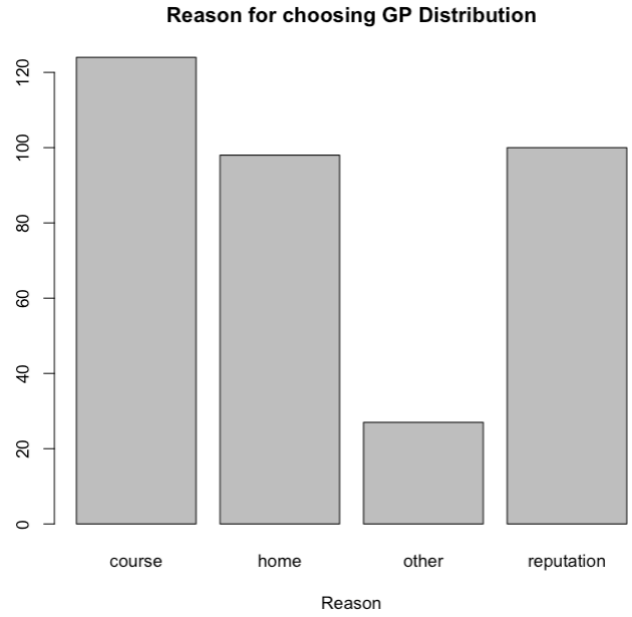
Q4: Create a visualization and answer the question below, which will provide an interesting story or insight within this data.

- Who is your audience?
- What is the application insight?
- What does this application insight mean for the audience? Why is it important for the audience to know?

Answer:

- The plot I've chosen to make is a bar plot and pie chart. showing the distribution of the reason for why students chose the school GP and the percentage of students under each reason.
- The target audience for my visualization is the authority running the GP school - this plot could potentially help in deciding their strong and weak points, such as if the courses or bad.
- From this visualization, we can see that a majority of the student's have mentioned the reason for choosing the school GP is because of it's course offerings (it has the tallest bar). The number of students who chose GP due to it's reputation is slightly higher than the students who chose it because it's close to home.
- The pie chart shows the percentage of students who chose each reason while the bar plot shows number of students (not percentage) who chose each reason.





Q5: There are other types of regression models outside of linear and logistic regression. Locate a journal article, which utilizes Elastic Net Regression. Write a summary of the article and how it utilizes the model.

Answer:

This paper talks about the prediction of a rapidly progressing disease called the Creutzfeldt-Jakob Disease (CJD) that affects a small population of the United States. It is a "prion disease" - meaning it is a type of protein that causes normal proteins in the brain to fold abnormally that can be caused due to infectious meat products. This type of disease leads to degeneration of different parts of the brain, resulting in decline in physiological aspects. This disease presents various symptoms such as depression and hallucinations and could even develop into worse conditions such as dementia or Alzheimer's. Due to the severity of this disease, the team have chosen various techniques such as elastic net regression, LSTMs and Random Forest Regressors to predict the existence of this disease which would potentially help in early detection of this disease in patients.

The model uses 8 different variables to predict CJD levels. This data was recorded between 1979 and 2015. In addition, they've made a finding that states there is a strong correlation between CJD levels and beef production in the US. The 8 variables were collected yearly and involved exposure to various chemicals, beer consumption, smoking etc. The data was collected from various sources and put together for the regressor. Correlation tests were performed between the variables, and it was found there is a high degree of correlation between the variables. They chose ENR instead of regular multivariate regression models because ENR performs well with multicollinearity. They use Pearson correlation for comparing CJD levels with the other independent variables and found that the variables they chose were all significant and could have an impact on predicting CJD levels accurately.

Out of the three models trained – ENR, LSTM and RF – the ENR model gave the most accurate results as its RMSE and MAE values were 0.179 and 0.136 and the other models had RMSE > 0.2 and MAE > 0.13. Through the model weights, it was found that beer consumption, obesity and tobacco usage can lead to a weakened immune system which can increase risk of catching CJD and variables such as beef production and nitrogen usage did not have a big impact on CJD contraction.

Link: https://arxiv.org/pdf/2108.04972.pdf

Citation:
Bhakta, A., & Byrne, C. (2021, August 11). Creutzfeldt-Jakob disease prediction using machine learning techniques. [2108.04972] Creutzfeldt-Jakob Disease Prediction Using Machine Learning Techniques. Retrieved October 6, 2021, from http://arxiv-export-lb.library.cornell.edu/abs/2108.04972.