Assignment 4 - Analyzing Student Alcohol Consumption

Stephen Quinn Turner - sqturner - 20576575

Rebecca Rayner - rsrayner - 20562488

Yusuf Khaled - yhbkhale - 20508841

Samantha Villaluz - skvillal - 20582656

Friday, April 05, 2019

1.0 Objective

For our final assignment for MSCI 718, we are asked to analyze a dataset of our choice and generate useful insights.

We chose to analyze student alcohol consumption in secondary schools. The data includes social, gender and study data from secondary school students. The data was provided by the UC Irving Machine Learning Repository and is available on Kaggle (https://www.kaggle.com/uciml/student-alcohol-consumption).

Our objective for this analysis was to analyze the correlation between alcohol consumption, and other characteristics among Portuguese post-secondary school students.

2.0 Hypothesis

From the set of variables collected in this dataset, our hypothesis is that there is correlation between alcohol consumption and each of the following variables: * health (negative correlation) * frequency of going out (positive) * past class failures (positive) * final average (negative)

Each of these hypothesis come from personal experience and/or curiosity. As students we know that when you go out you more you tend to drink more, which can cause you to sleep less, which can impact your grades and your overall health. For this reason we felt the variables chosen had a good chance of showing some correlation with alcohol consumption.

3.0 Background

This analysis will use two datasets, one set includes information on students taking a math course (matData) and the other includes information on students taking a Portugese course (porData).

3.1 Raw Variables

There are 35 variables in this dataset. Descriptions of a select group of variables in the dataset are shown below (provided by the original Kaggle source):

- sex - student’s sex (binary: ‘F’ - female or ‘M’ - male)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)

- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- G3 - final grade (numeric: from 0 to 20, output target)

3.2 Transformed Variables

Based on the description of the variables, we deemed that it was appropriate to make the following transformations to the imported dataset:

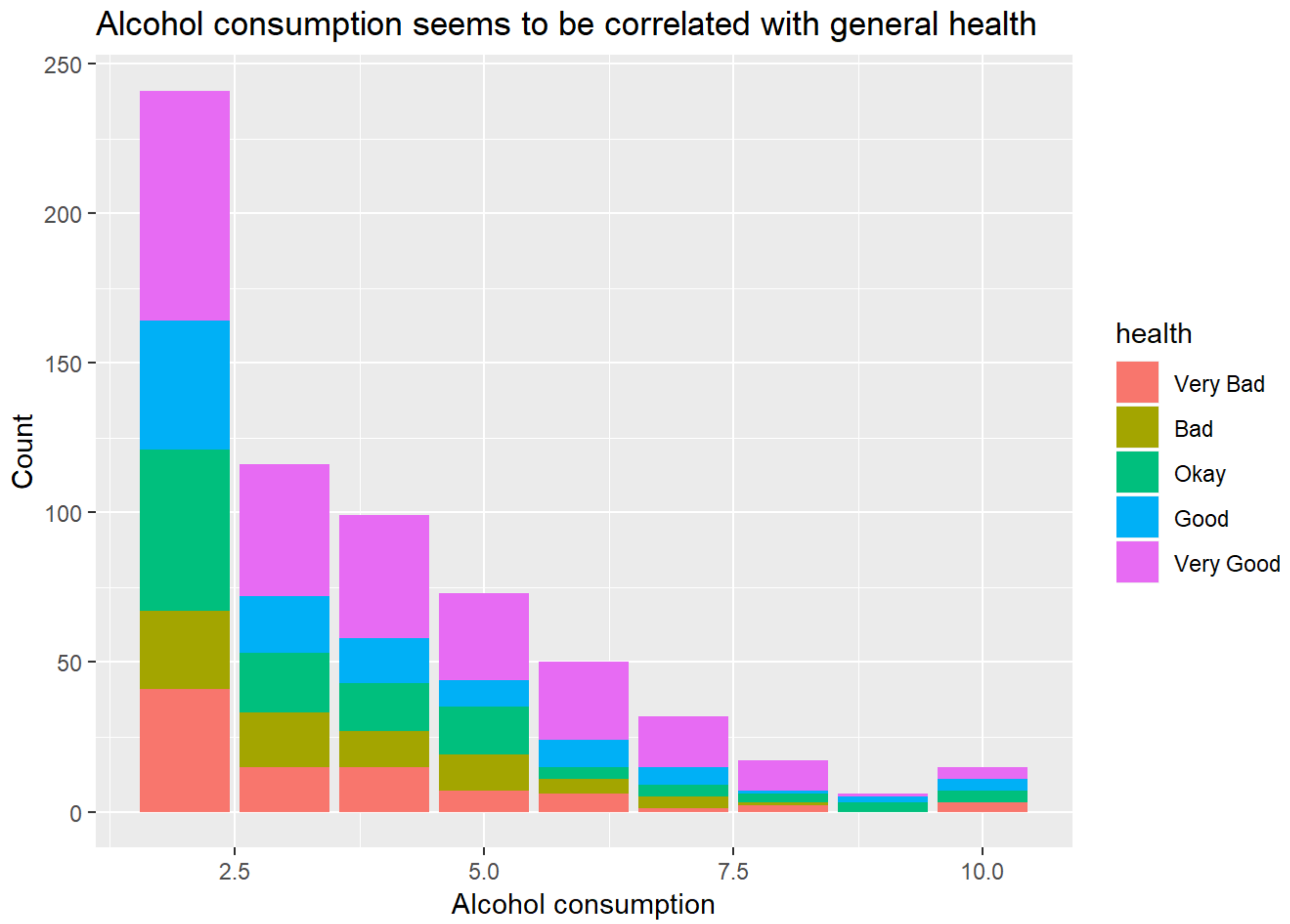
- Add variable `Talc` *<2 - very low to 10 - very high> (Sum of students' drinking habits)* –> factored so that 2 = No Alcohol, 3-6 = Low Alcohol, 7-10 = High Alcohol
- Add factor `famrel` *<1 - very bad to 5 - excellent>*
- Add factor `freetime` *<1 - very low to 5 - very high>*
- Add factor `goout` *<1 - very low to 5 - very high>*
- Add factor `health` *<1 - very bad to 5 - very good>*

Additional note: there are 382 students that belong to both datasets so there are only 13 students who took math did not take Portugese. It is not likely or appropriate that we can find statistically or practically significant differences between the students who took math and not Portugese or vice-versa. This is because we don't know whether the remaining Portugese students took math or not since they could have been in another class that was not included in the data collection process. Since most students who took math also took Portugese but not vice versa, we will use the Portugese data whenever analyzing traights unrelated to the course type.

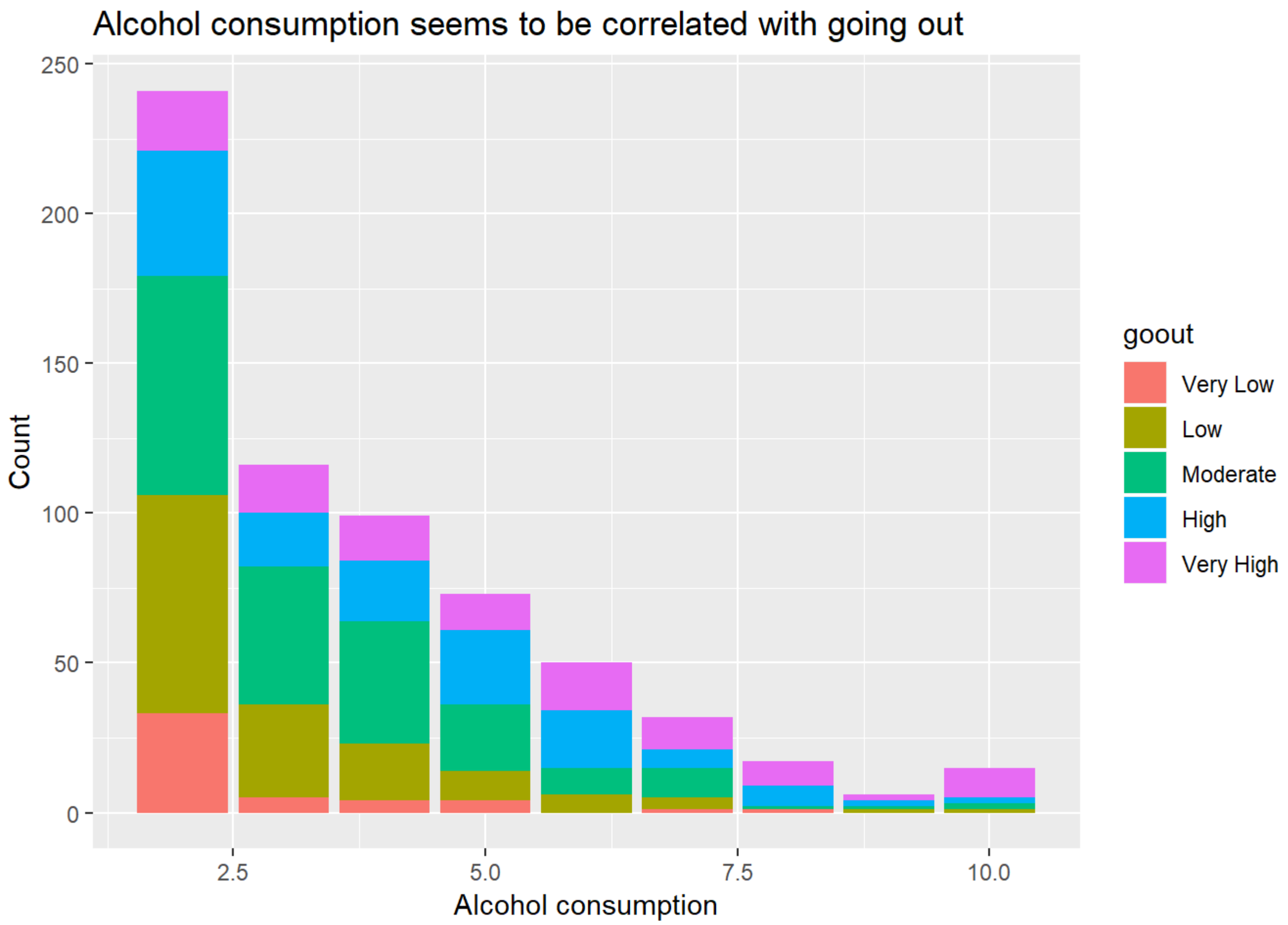
3.3 Discovery

Data visualizations involving the variables health and going out. Alcohol consumption is indicated by the different colours.

```
# Lets examine two of our feature variables and analyze their correlation to alcohol consumption
ggplot(porData, aes(Talc))+
  geom_bar(aes(fill = health), position = position_stack(reverse = TRUE))+
  xlab("Alcohol consumption")+
  ylab("Count")+
  ggtitle("Alcohol consumption seems to be correlated with general health")
```



```
ggplot(porData, aes(Talc))+
  geom_bar(aes(fill = goout), position = position_stack(reverse = TRUE))+
  xlab("Alcohol consumption")+
  ylab("Count")+
  ggtitle("Alcohol consumption seems to be correlated with going out")
```



From this we can see that overall women drink less than men, and there are especially more men who drink higher levels of alcohol. This may mean that we want to separate by sex when looking at alcohol effects. We also see that our initial assumptions about which feature variables to examine could be vindicated, based on the fact that both going out and general health seem to have some correlation with alcohol consumption - at least upon visual inspection.

Now that we've set up our data, and have some confidence going forward that our initially chosen hypothesis variables could show some correlation to alcohol consumption, let's move forward with our analysis.

4.0 Analysis: Contrasts, Effect Size and Anova

The four variables we deemed interesting to assess against alcohol consumption were 1. Health 2. Past Class Failures 3. Frequency of Going Out 4. Final Grade in Class

We have broken down the analysis in three parts:

1. Contrasts Explore how each of the four variables differ against heavy vs. moderate drinkers vs. non-drinkers. The data is segmented in two different ways to ensure that statistical significance is not only due to a high number of data points in the whole dataset.

- Gender (Male, Female)
- Alcohol Consumption (None, Low Alcohol, High Alcohol)

2. Effect Size Evaluate the effect size of the results found in 4.1 Contrasts.

3. Anova

4.1 Contrasts

We desired to see if there was a difference in significance when we segmented the data using Total, just males, and just females when assessing correlation between alcohol consumption and final average.

Below is a sample of the code used. We repeated this three more times with the other three variables.

Results: All genders, Male Datasets

- Stat sig difference in final average between drinkers and non-drinkers
- Stat sig difference in final average between heavy drinkers and non-drinkers

```
summary.lm(GradeEffect.Total)
```

```
## 
## Call:
## aov(formula = G3 ~ Talc2, data = porData)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.3693  -1.8757   0.1243   2.1243   7.1243 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 11.56739    0.15542  74.428  < 2e-16 ***
## Talc2Alcohol vs. No Alcohol -0.40095    0.09774  -4.102 4.61e-05 ***
## Talc2High vs. Low           -0.70930    0.20930  -3.389 0.000744 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.188 on 646 degrees of freedom
## Multiple R-squared:  0.02942,    Adjusted R-squared:  0.02642 
## F-statistic: 9.791 on 2 and 646 DF,  p-value: 6.471e-05
```

```
summary.lm(GradeEffect.Male)
```

```
##
## Call:
## aov(formula = G3 ~ Talc2, data = porData.Male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7971  -1.2695  -0.0357   1.7305   7.7305
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  11.3674     0.2114  53.783  < 2e-16 ***
## Talc2Alcohol vs. No Alcohol  -0.7148     0.1532  -4.665 4.91e-06 ***
## Talc2High vs. Low            -0.6169     0.2521  -2.447   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 263 degrees of freedom
## Multiple R-squared:  0.08258,    Adjusted R-squared:  0.0756
## F-statistic: 11.84 on 2 and 263 DF,  p-value: 1.196e-05
```

Results: Female Dataset

- Not stat sig difference in final average between drinkers and non-drinkers
- Not stat sig difference in final average between heavy drinkers and non-drinkers

```
summary.lm(GradeEffect.Female)
```

```
##
## Call:
## aov(formula = G3 ~ Talc2, data = porData.Female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3096  -2.1977  -0.1977   1.8023   6.8023
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12.216725   0.299517  40.788   <2e-16 ***
## Talc2Alcohol vs. No Alcohol  0.009525   0.164863   0.058    0.954
## Talc2High vs. Low           -0.083394   0.433120  -0.193    0.847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.132 on 380 degrees of freedom
## Multiple R-squared:  0.0003563,  Adjusted R-squared:  -0.004905
## F-statistic: 0.06772 on 2 and 380 DF,  p-value: 0.9345
```

4.2 Cohen’s d for Grades between Male Drinkers

We wish to obtain a comparison of two means of the results we obtained through the contrasts. It was gathered that there is a significant difference in male moderate drinkers vs. male heavy drinkers. Through finding the cohen's d value, we have obtained results for a SMALL effect size.

```
cohen.d(lowMale$G3, highMale$G3)
```

```
##
## Cohen's d
##
## d estimate: 0.3752798 (small)
## 95 percent confidence interval:
##      lower      upper
## 0.06153945 0.68902015
```

4.3 Ancova

When determining whether alcohol had an effect on a male student's grades, it is important to consider other factors are correlated and may influence the analysis.

In this case, we decided that `goout` would be an interesting covariate. The ANCOVA analysis revealed that the interaction is not significant, so the slope across groups is not different. However, the the category variable (`goout`) is significant, so the intercepts among groups are different.

```
GradeEffect2.Total1 <- aov(G3 ~ goout + Talc2 + goout:Talc2, data=porData.Male)
Anova(GradeEffect2.Total1, type="III")
```

	Sum Sq	Df	F value	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	656.38950	1	67.892436	9.619518e-15
goout	51.89586	4	1.341938	2.548737e-01
Talc2	48.76150	2	2.521778	8.235007e-02
goout:Talc2	93.17591	8	1.204685	2.965599e-01
Residuals	2426.68807	251	NA	NA

5 rows

5.0 Conclusion

For our variables of focus, there is a statically significant effect found when all data is used, but many of these effects disappear when data is separated by sex and degree of alcohol consumption during weekdays and weekends. The statistical significance could be due to the higher number of datapoints, but that doesn't necessarily point to practical significance.

Through contrasts analysis, it was found that the mean of final averages are statistically significant between 1) non drinkers vs. drinkers and 2) moderate drinkers vs. heavy drinkers for MALES ONLY. More granular analysis through effect size Cohen's d disputed this result by showing that it had a small effect size. Ancova dove deeper into the other three variables that could possibly be displaying an effect such as going out.

Overall, it was learned that a result can be validated in a number of ways. Health and failures can only be proven to be negatively correlated in males. All variables were not statistically significant when segmenting by females only. Female performance is not closely correlated to alcohol consumption.