Data Science Final Assignment Prepared by Michelle Nicome and Damico Nicome

Introduction

This project examines two datasets created from surveying students taking math and Portuguese language classes. Each dataset contains data on various aspects of students daily lives. An analyses of various variables is conducted to determine the factors that contribute to alcohol consumption and the resulting effect on the student's final grade. The datasets are imported into R as comma-separated values (csv) and R logic is used to consolidate them so that comprehensive analyses of multiple variables from both datasets can be performed. The consolidated file is transformed into an R dataframe for ease of manipulation.

The analyses conducted in this project hopes to shed light on key indicators educators can use to recognize students who may have an existing problem with alcohol, or who may be predisposed to the condition. Increased awareness of the latter may enable early intervention that helps students to change behavioral patterns and minimize or prevent damage to their academic career.

A description of the variables contained in both datasets follows:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9thgrade, 3 - secondary education or 4 - higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

Note on document organization: General information about the project is included in the body of the document (i.e. outside the R code chunk) and comments about the R logic used to perform the analysis is included within the appropriate R code chunk.

```
## Clear environment
rm(list=ls())
```

##Data Organization

There were 85 students, who took both Math and Portuguese, creating duplicate data in each dataset. However, previous studies on this dataset found that math and Portuguese grades were highly correlated. As such this study assumes that course subject has negligible implication on average grade and it is therefore safe to remove duplicate observations (Cortez & Silva, 2008). The code chunk below imports both datasets and saves them as R data frames for ease of manipulation and subsequent retrieval. After loading and saving the data, the datasets are merged in the next code chunk.

```
##Load both csv files, convert them to dataframes and reload them in .Rdata format.

student_mat <- read_csv("C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Da
ta Science/Final Project/student-mat.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   absences = col_double(),
##   G1 = col_double(),
##   G2 = col_double(),
##   G3 = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
##View(student_mat)##Used only to verify initial loading of data
math_data <- data.frame(student_mat)
#math_data
save(math_data, file = "C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Dat
a Science/Final Project/math_data.Rdata")
load("C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Data Science/Final Pr
oject/math_data.Rdata")
##View(math_data)##Used only to validate data was reloaded correctly in .Rdata format

student_por <- read_csv("C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Da
ta Science/Final Project/student-por.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   absences = col_double(),
##   G1 = col_double(),
##   G2 = col_double(),
##   G3 = col_double()
## )
## See spec(...) for full column specifications.
```

```
##View(student_por)##Used only to verify initial loading of data
por_data <- data.frame(student_por)
#por_data
save(por_data, file = "C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Data
Science/Final Project/por_data.Rdata")
load("C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Data Science/Final Pr
oject/por_data.Rdata")
##View(por_data)##Used only to validate data was reloaded correctly in .Rdata format
```

The datasets are combined and 85 duplicates are removed to adjust for the students, who participated in both courses. The resulting dataset contains 959 observations across 30 independent and 3 dependent variables.

```
##This logic combines both datasets appending the columns.
comb_math_por_data <- rbind(math_data, por_data)
##Remove duplicates from combined dataset
comb_math_por_data_no_dupes <- comb_math_por_data%>%distinct(school,sex,age,address,fams
ize,Pstatus,
                                                             Medu,Fedu,Mjob,Fjob,reason,
                                                             guardian,traveltime,studyti
me,failures,
                                                             schoolsup, famsup,activitie
s,nursery,higher,internet,
romantic,famrel,freetime,goout,Dalc,Walc,
                                                             health,absences,.keep_all =
TRUE)

#add a column with average grades (math or Portuguese, whichever is available)
comb_math_por_data_no_dupes <- comb_math_por_data_no_dupes%>%mutate(avggrades=rowMeans(c
bind(
  comb_math_por_data_no_dupes$G1,
  comb_math_por_data_no_dupes$G2,
  comb_math_por_data_no_dupes$G3)))
##Saving dataset as R data frame for later use.
save(comb_math_por_data_no_dupes, file = "C:/Users/dnico/OneDrive - B&N Enterprises/Desk
top/Vanderbilt/Data Science/Final Project/comb_math_por_data_no_dupes.Rdata")

##Verify that duplicate rows were removed from merged datasets
nrow(comb_math_por_data)##1044
```

```
## [1] 1044
```

```
nrow(comb_math_por_data_no_dupes)##959 (85 rows removed)
```

```
## [1] 959
```

```
##Display top and bottom of merged dataset
head(comb_math_por_data_no_dupes)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
## 4     GP   F  15       U     GT3       T    4    2   health services
## 5     GP   F  16       U     GT3       T    3    3    other    other
## 6     GP   M  16       U     LE3       T    4    3 services    other
##       reason guardian traveltime studytime failures schoolsup famsup paid
## 1     course   mother          2         2        0       yes     no   no
## 2     course   father          1         2        0        no    yes   no
## 3      other   mother          1         2        3       yes     no  yes
## 4       home   mother          1         3        0        no    yes  yes
## 5       home   father          1         2        0        no    yes  yes
## 6 reputation   mother          1         2        0        no    yes  yes
##   activities nursery higher internet romantic famrel freetime goout Dalc
## 1         no     yes    yes       no       no      4        3     4    1
## 2         no      no    yes      yes       no      5        3     3    1
## 3         no     yes    yes      yes       no      4        3     2    2
## 4        yes     yes    yes      yes      yes      3        2     2    1
## 5         no     yes    yes       no       no      4        3     2    1
## 6        yes     yes    yes      yes       no      5        4     2    1
##   Walc health absences G1 G2 G3 avggrades
## 1    1      3        6  5  6  6  5.666667
## 2    1      3        4  5  5  6  5.333333
## 3    3      3       10  7  8 10  8.333333
## 4    1      5        2 15 14 15 14.666667
## 5    2      5        4  6 10 10  8.666667
## 6    2      5       10 15 15 15 15.000000
```

```
tail(comb_math_por_data_no_dupes)
```

```
##     school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 954     MS   F  18       R     GT3       T    4    4  teacher  at_home
## 955     MS   F  19       R     GT3       T    2    3 services    other
## 956     MS   F  18       U     LE3       T    3    1  teacher services
## 957     MS   F  18       U     GT3       T    1    1    other    other
## 958     MS   M  17       U     LE3       T    3    1 services services
## 959     MS   M  18       R     LE3       T    3    2 services    other
##         reason guardian traveltime studytime failures schoolsup famsup
## 954 reputation   mother          3         1        0        no    yes
## 955     course   mother          1         3        1        no     no
## 956     course   mother          1         2        0        no    yes
## 957     course   mother          2         2        0        no     no
## 958     course   mother          2         1        0        no     no
## 959     course   mother          3         1        0        no     no
##     paid activities nursery higher internet romantic famrel freetime goout
## 954   no        yes     yes    yes      yes      yes      4        4     3
## 955   no        yes      no    yes      yes       no      5        4     2
## 956   no         no     yes    yes      yes       no      4        3     4
## 957   no        yes     yes    yes       no       no      1        1     1
## 958   no         no      no    yes      yes       no      2        4     5
## 959   no         no      no    yes      yes       no      4        4     1
##     Dalc Walc health absences G1 G2 G3 avggrades
## 954    2    2      5        4  7  9 10  8.666667
## 955    1    2      5        4 10 11 10 10.333333
## 956    1    1      1        4 15 15 16 15.333333
## 957    1    1      5        6 11 12  9 10.666667
## 958    3    4      2        6 10 10 10 10.000000
## 959    3    4      5        4 10 11 11 10.666667
```

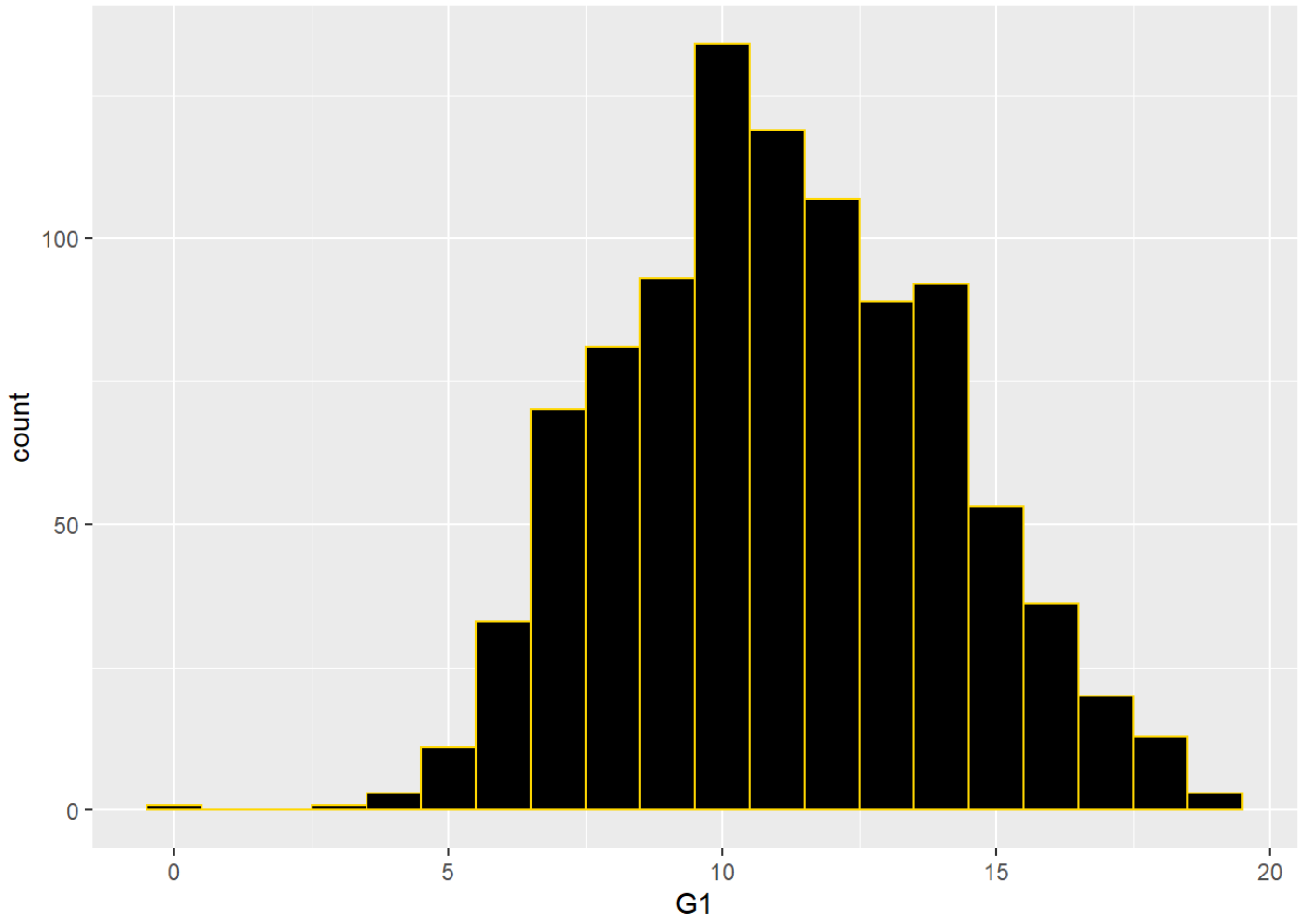
#Exploratory Data Analysis

The following chunks of code generate charts showing grade distribution by period. Though the dependent variable examined here is the average of the three grading periods, reviewing the distribution of individual grade periods may provide additional insights that help to explain variation in the overall average (e.g. changes in alcohol consumption as the school year progresses). This could be useful in designing future studies. One notable observatioin is that there is a marked increase in the number of “0” grades in the third period. Furthermore, it appears as though the number of “0” grades increase with each period. However, further analysis is necessary to determine whether this increase is due to alcohol consumption.

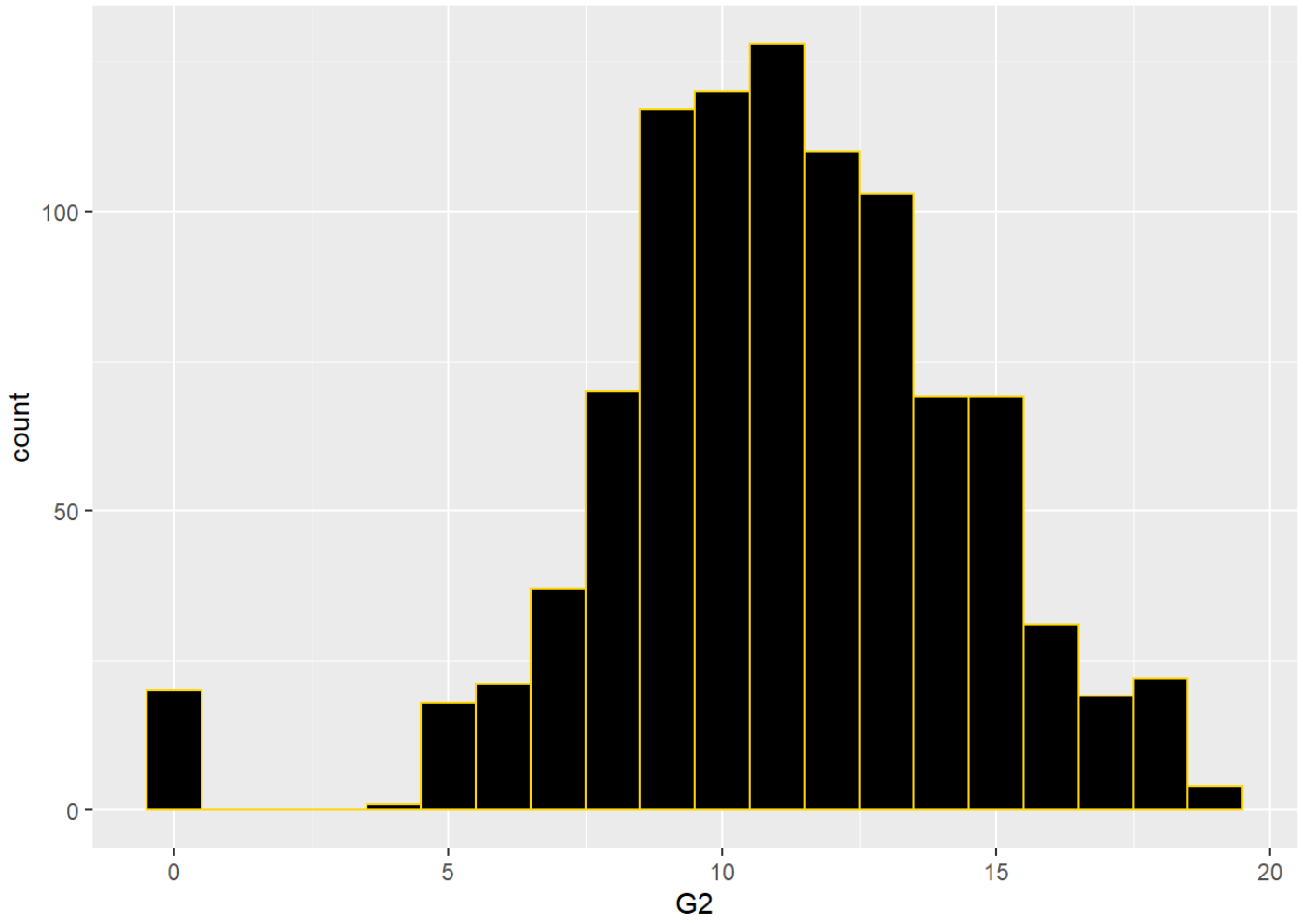
```
## Display top of data frame
head(comb_math_por_data)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
## 4     GP   F  15       U     GT3       T    4    2   health services
## 5     GP   F  16       U     GT3       T    3    3    other    other
## 6     GP   M  16       U     LE3       T    4    3 services    other
##       reason guardian traveltime studytime failures schoolsup famsup paid
## 1     course   mother          2         2        0       yes     no   no
## 2     course   father          1         2        0        no    yes   no
## 3      other   mother          1         2        3       yes     no  yes
## 4       home   mother          1         3        0        no    yes  yes
## 5       home   father          1         2        0        no    yes  yes
## 6 reputation   mother          1         2        0        no    yes  yes
##   activities nursery higher internet romantic famrel freetime goout Dalc
## 1         no     yes    yes       no       no      4        3     4    1
## 2         no      no    yes      yes       no      5        3     3    1
## 3         no     yes    yes      yes       no      4        3     2    2
## 4        yes     yes    yes      yes      yes      3        2     2    1
## 5         no     yes    yes       no       no      4        3     2    1
## 6        yes     yes    yes      yes       no      5        4     2    1
##   Walc health absences G1 G2 G3
## 1    1      3        6  5  6  6
## 2    1      3        4  5  5  6
## 3    3      3       10  7  8 10
## 4    1      5        2 15 14 15
## 5    2      5        4  6 10 10
## 6    2      5       10 15 15 15
```

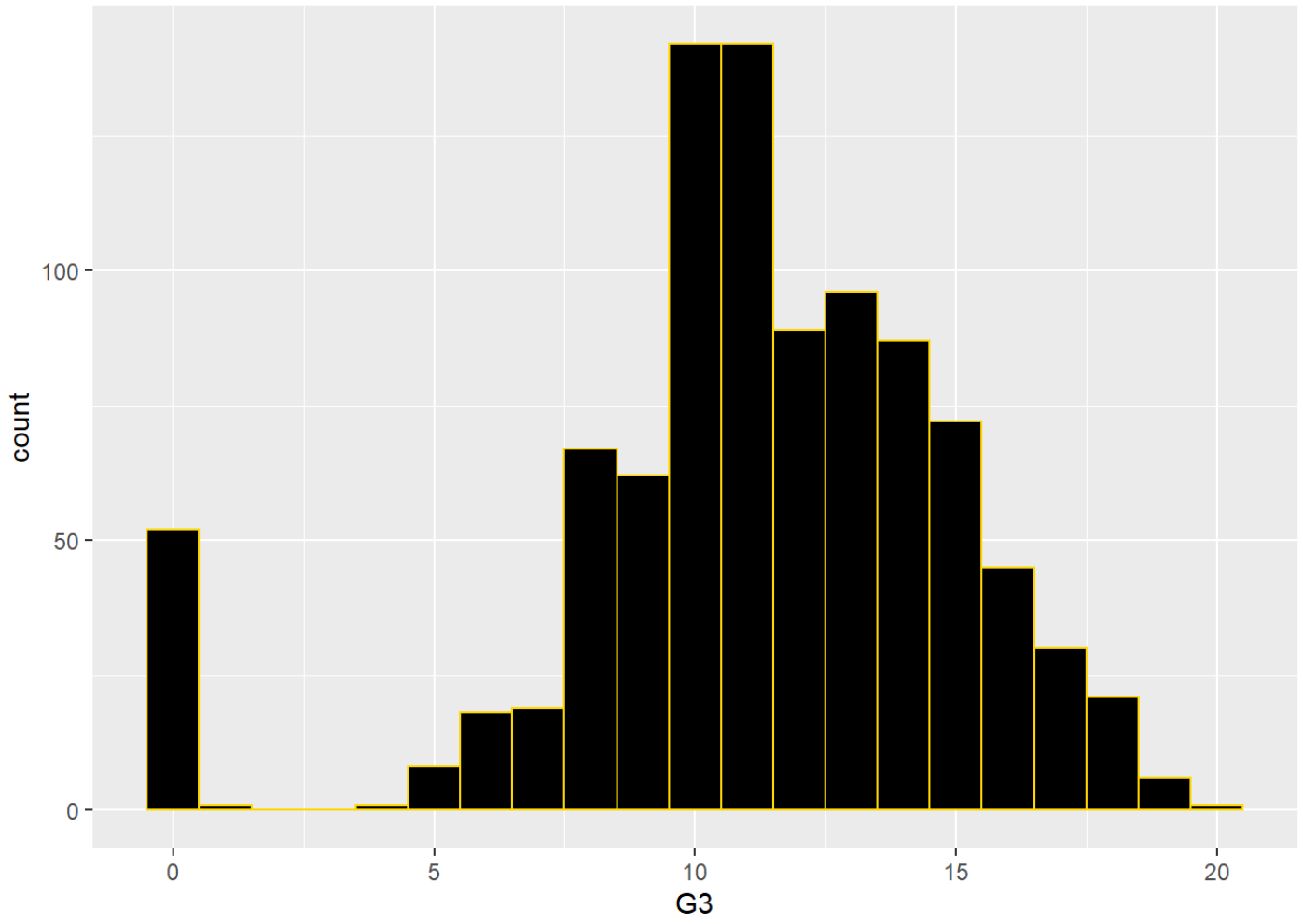
```
##Create a histogram of grades by period
first_per_grade_dist<-ggplot(comb_math_por_data_no_dupes,aes(x=G1)) ## First period grad
e distribution
first_per_grade_dist<-first_per_grade_dist+geom_histogram(fill = "black", color = "gold"
, binwidth = 1)
first_per_grade_dist
```




```
second_per_grade_dist<-ggplot(comb_math_por_data_no_dupes,aes(x=G2)) ## Second period gr
ade distribution
second_per_grade_dist<-second_per_grade_dist+geom_histogram(fill = "black", color = "gol
d", binwidth = 1)
second_per_grade_dist
```




```
third_per_grade_dist<-ggplot(comb_math_por_data_no_dupes,aes(x=G3)) ## Third period grad
e distribution
third_per_grade_dist<-third_per_grade_dist+geom_histogram(fill = "black", color = "gold"
, binwidth = 1)
third_per_grade_dist
```



The distribution for all three grading periods appear to be normally distributed.Further evidence of this is seen below after calculating measures of central tendency (i.e. mean,median, mode). All measures are within one point of each other. Worth noting, is the is that period three observations were bimodal.

```
##Calculate measures of central tendency for period 1 grades
mean(comb_math_por_data_no_dupes$G1, na.rm = TRUE)
```

```
## [1] 11.07716
```

```
median(comb_math_por_data_no_dupes$G1, na.rm = TRUE)
```

```
## [1] 11
```

```
mode_per_1 <- table(as.vector(comb_math_por_data_no_dupes$G1))##Calculating mode for fir
st period one gradds
mode_per_1 <- names (mode_per_1)[mode_per_1==max (mode_per_1)]
mode_per_1
```

```
## [1] "10"
```

```
mean(comb_math_por_data_no_dupes$G2, na.rm = TRUE)
```

```
## [1] 11.10636
```

```
median(comb_math_por_data_no_dupes$G2, na.rm = TRUE)
```

```
## [1] 11
```

```
mode_per_2 <- table(as.vector(comb_math_por_data_no_dupes$G2))##Calculating mode for sec
ond period one gradds
mode_per_2 <- names (mode_per_2)[mode_per_2==max (mode_per_2)]
mode_per_2
```

```
## [1] "11"
```

```
##Calcullating mode for second perioe gradds
mean(comb_math_por_data_no_dupes$G3, na.rm = TRUE)
```

```
## [1] 11.17623
```

```
median(comb_math_por_data_no_dupes$G3, na.rm = TRUE)
```

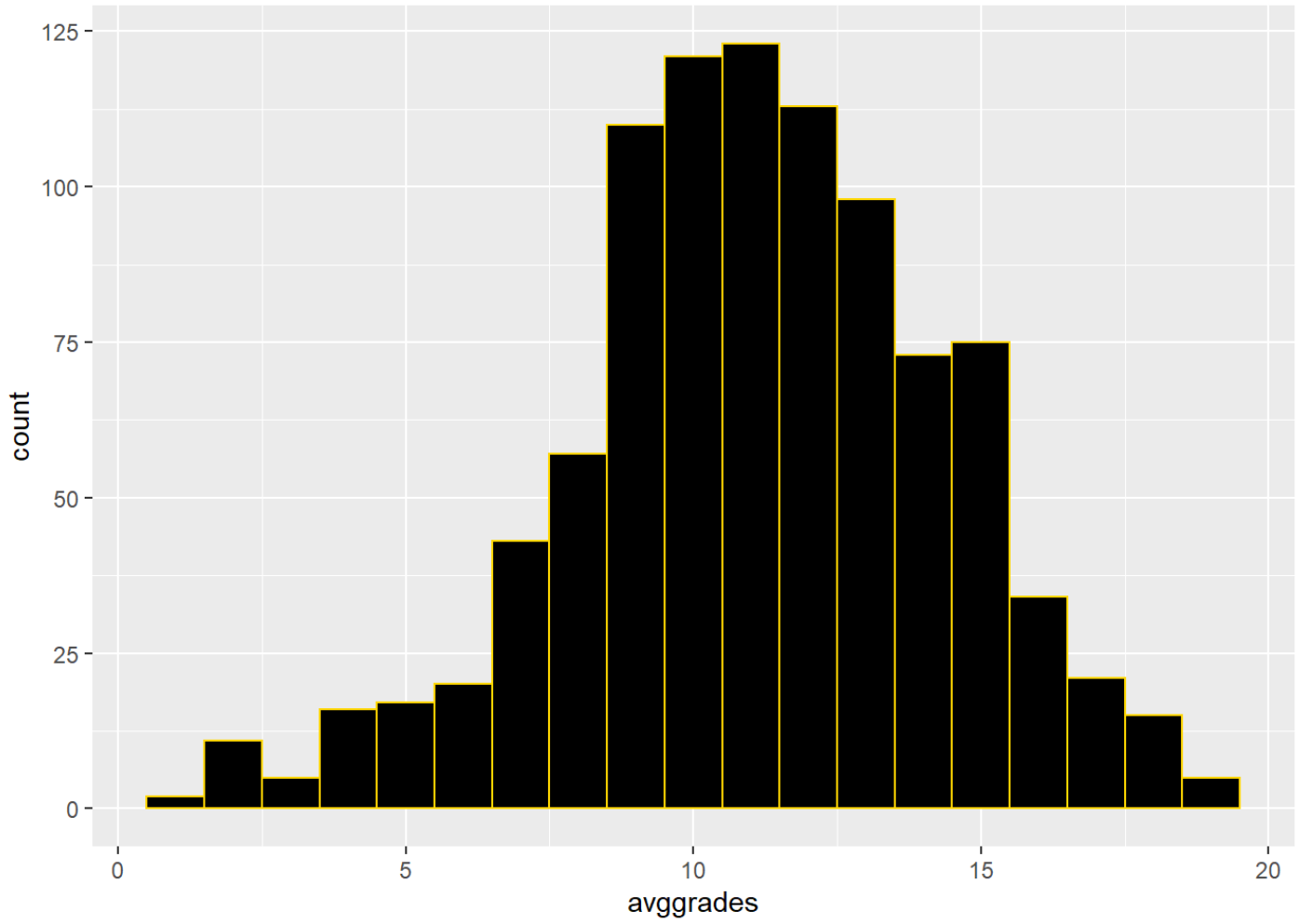
```
## [1] 11
```

```
mode_per_3 <- table(as.vector(comb_math_por_data_no_dupes$G3))##Calculating mode for fin
al period one gradds
mode_per_3 <- names (mode_per_3)[mode_per_3==max (mode_per_3)]
mode_per_3
```

```
## [1] "10" "11"
```

The code chunk below now examines the distribution of average final grade, calculated using the “avggrades” variable calculated above.

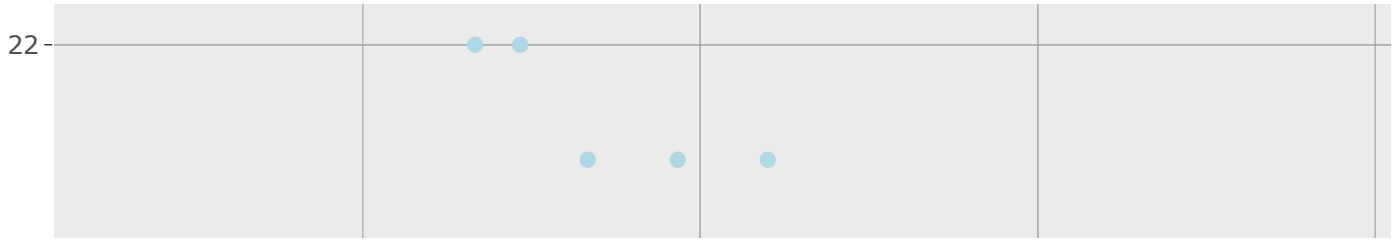
```
avg_final_grade_dist<-ggplot(comb_math_por_data_no_dupes,aes(x=avggrades,)) ## Third per
iod grade distribution
avg_final_grade_dist<-avg_final_grade_dist+geom_histogram(fill = "black", color = "gold"
, binwidth = 1)
avg_final_grade_dist
```

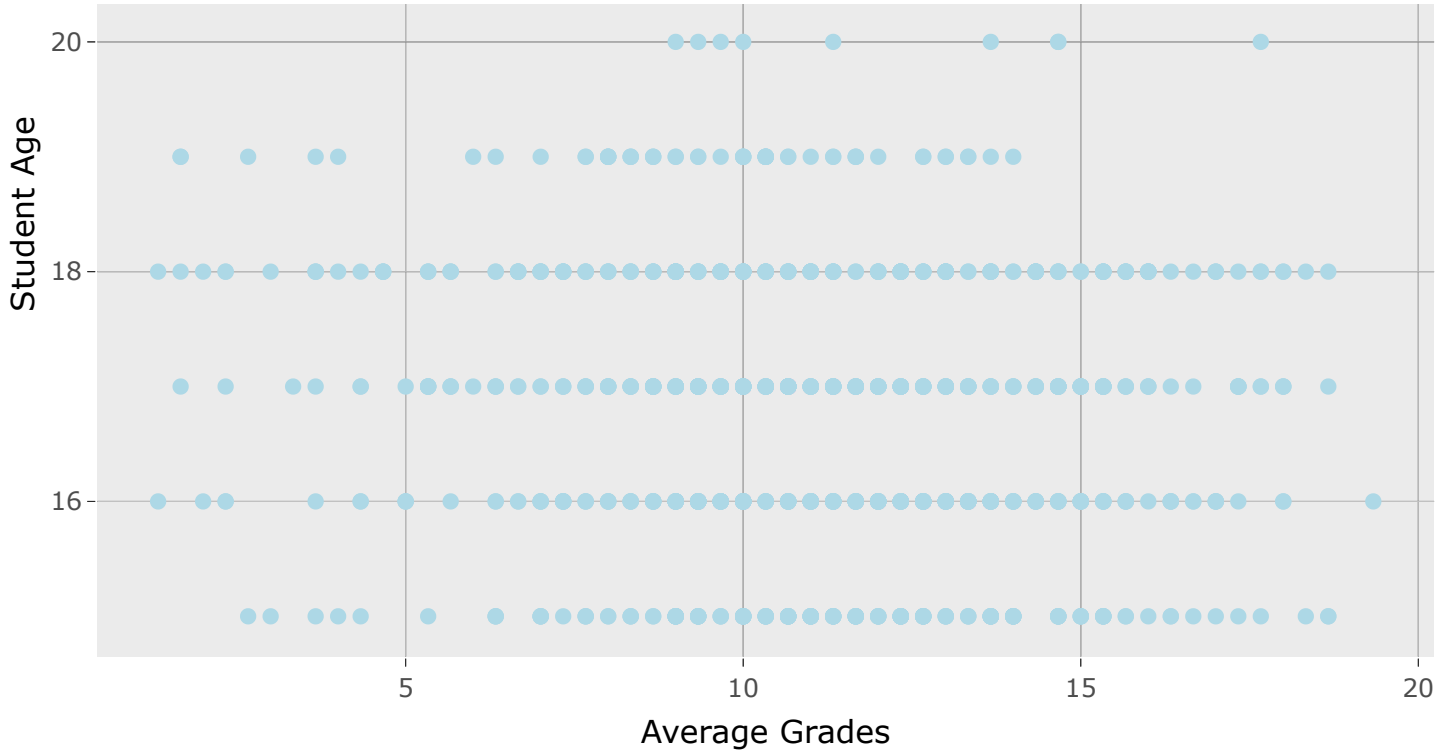


The next code chunk presents an interactive plot of average grades by student age. Hovering over each point shows their level of daily and weekly alcohol consumption.

```
gg<-ggplot(comb_math_por_data_no_dupes,
           aes(x=avggrades, y=age,
               text=paste0("Daily Alcohol Consumption: ",
                           Dalc,
                           "<br>",
                           "Weekly Alcohol Consumption: ",
                           Walc) ))
gg<-gg+geom_point(color="lightblue")
gg<-gg+xlab("Average Grades")+ylab("Student Age")
gg<-ggplotly(gg)

gg
```



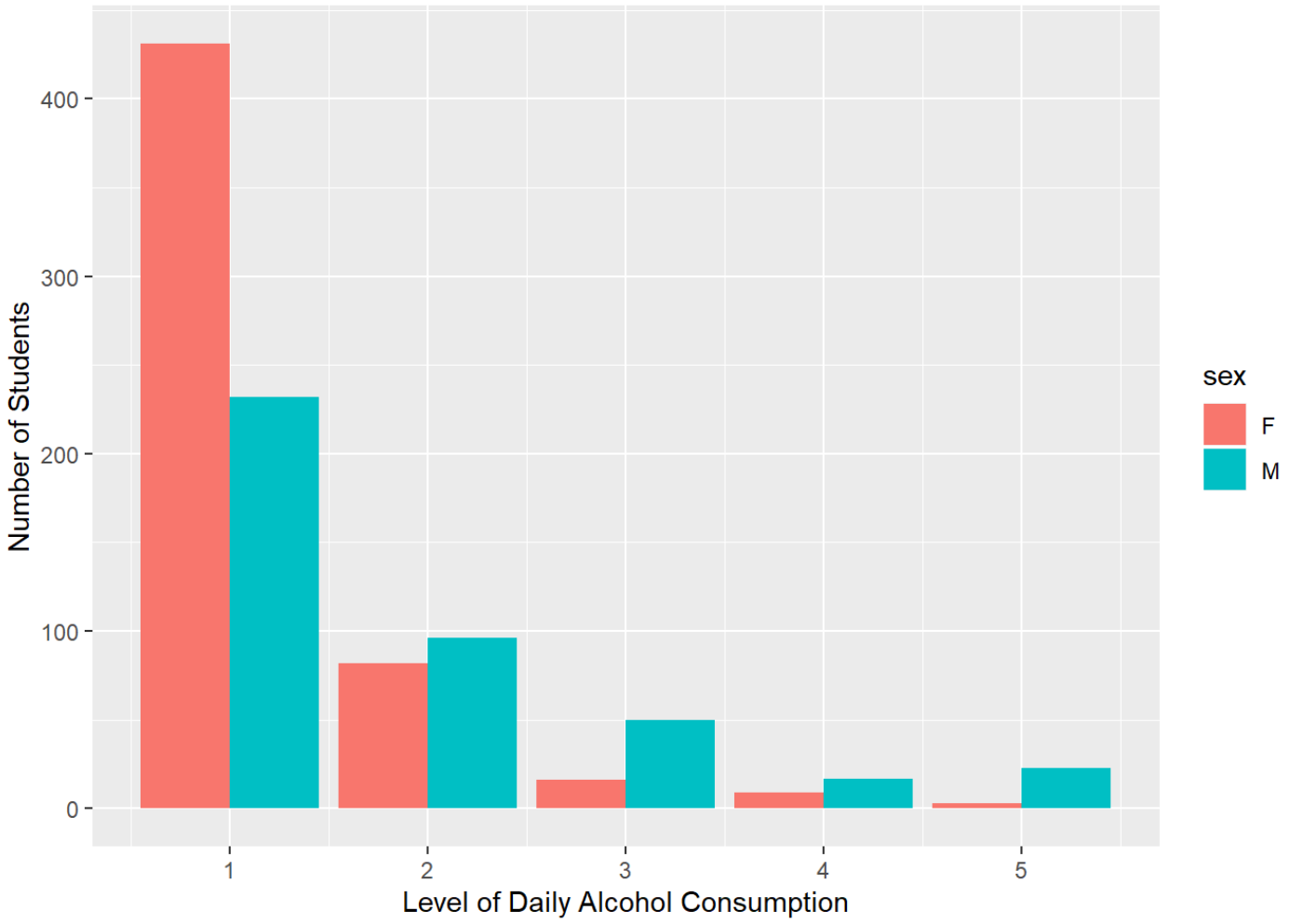


As with the distribution of each grading period, the average grade distribution also appears normally distribution. However, the distribution of grades below the median appear smoother than what was observed in the distribution of the individual grading periods.We will now examine the distribution of average final grades by level of alcohol consumption in the code chunk below.

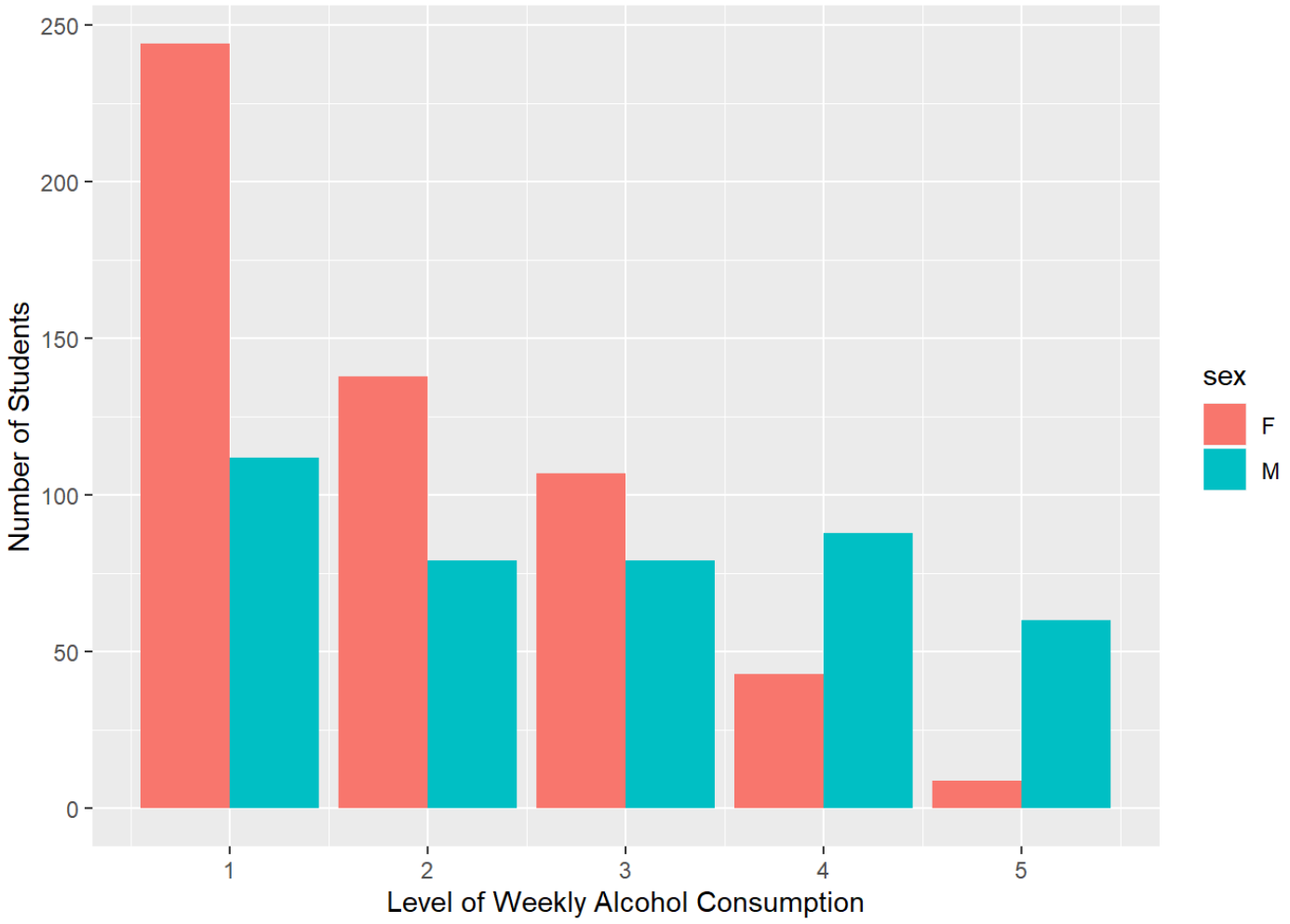
```
##Calculate the mean Average grade
comb_math_por_data_no_dupes%>%
  summarize(avg_grade=mean(avggrades))##11.11992
```

```
##   avg_grade
## 1  11.11992
```

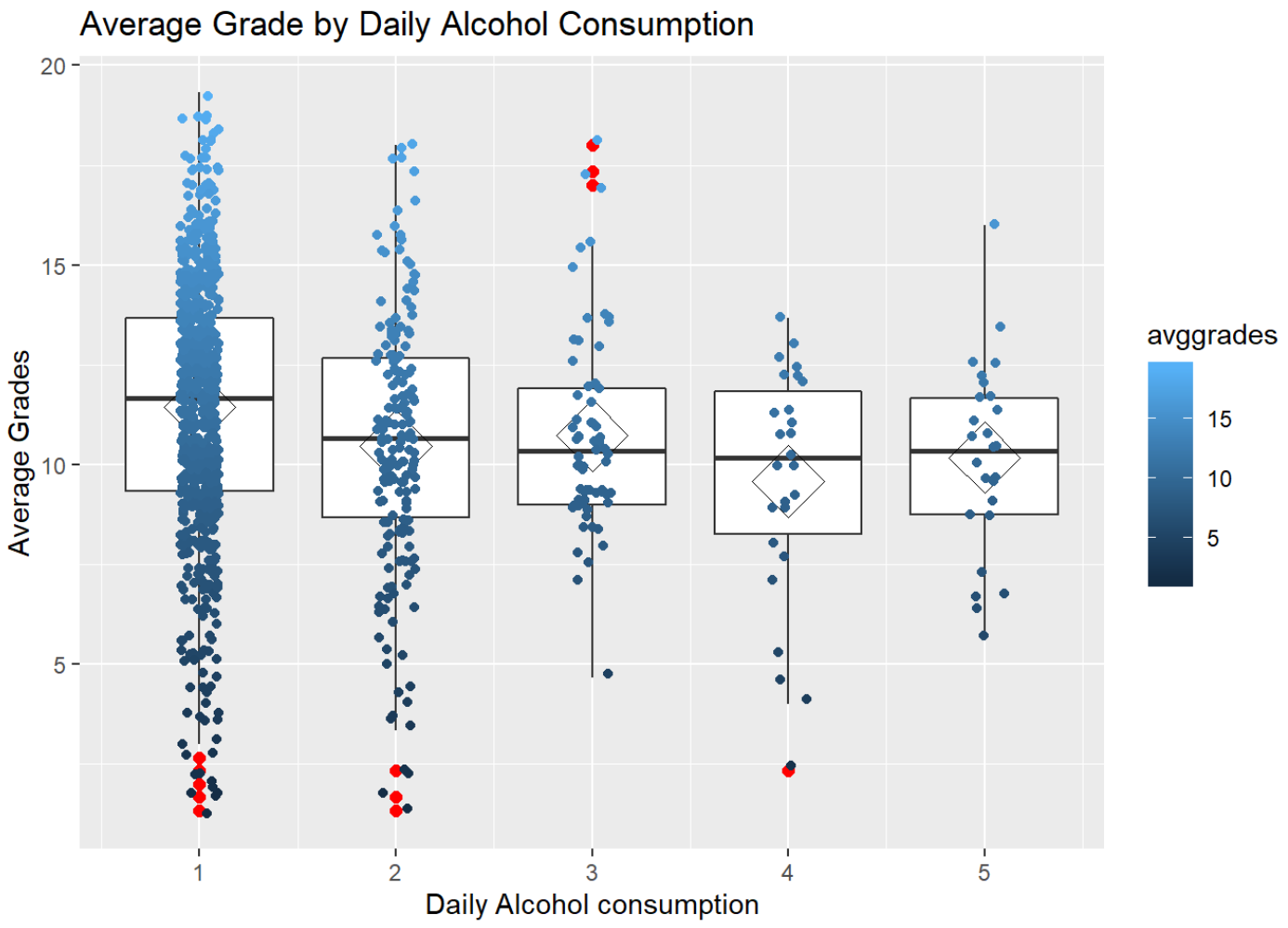
```
##Display a bar chart showing the number of students at each level of daily alcohol cons
umption
dalc_avg_grade_by_sex<-ggplot(comb_math_por_data_no_dupes,aes(x=Dalc, group=sex, fill=se
x))
dalc_avg_grade_by_sex<-dalc_avg_grade_by_sex+geom_bar(position = "dodge")
dalc_avg_grade_by_sex<-dalc_avg_grade_by_sex+ylab("Number of Students")+xlab("Level of D
aily Alcohol Consumption")
dalc_avg_grade_by_sex
```



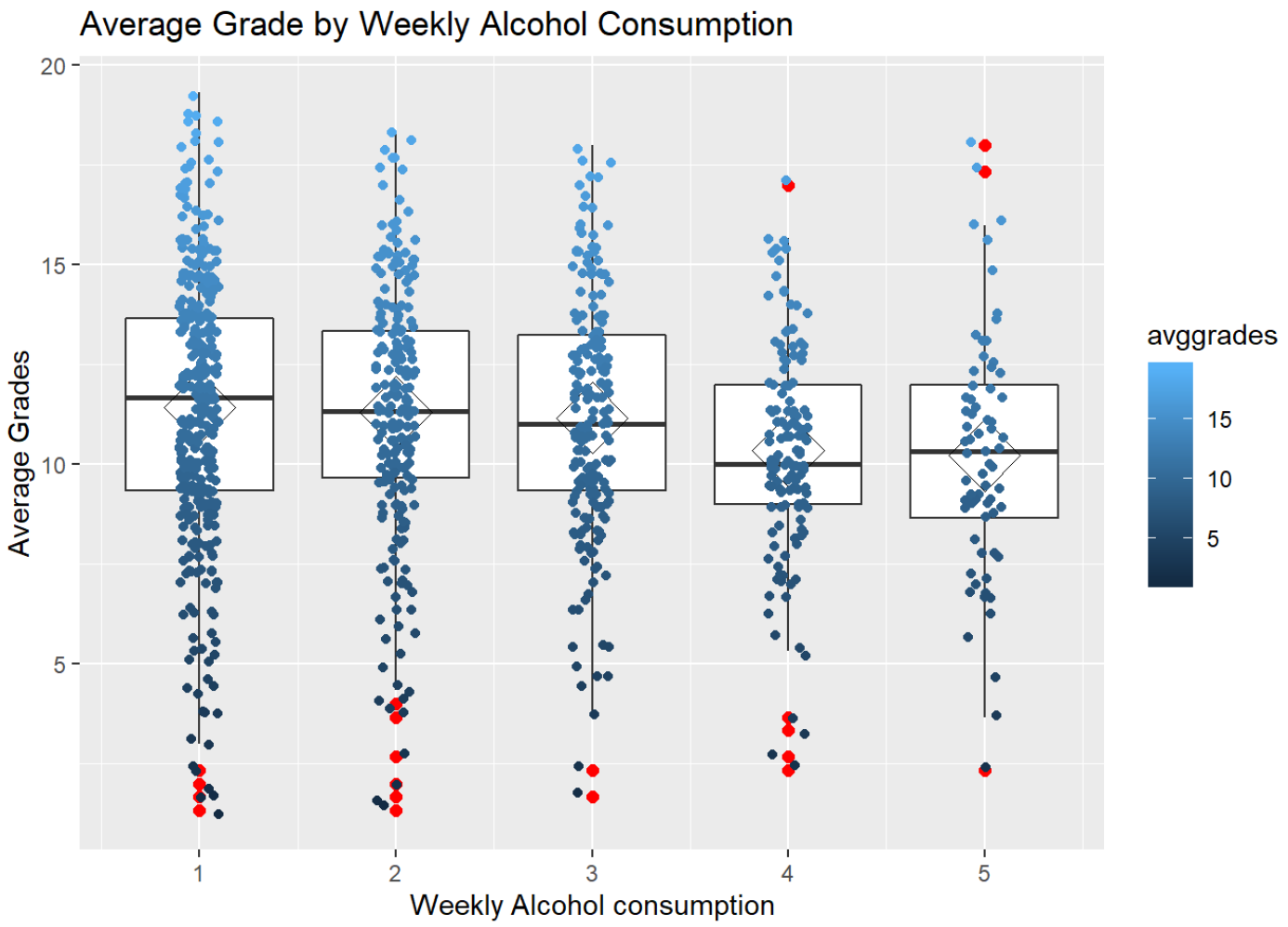

```
##Display a bar chart showing the number of students at each level of weekly alcohol con
sumption
walc_avg_grade_by_sex<-ggplot(comb_math_por_data_no_dupes,aes(x=Walc, group=sex, fill=se
x))
walc_avg_grade_by_sex<-walc_avg_grade_by_sex+geom_bar(position = "dodge")
walc_avg_grade_by_sex<-walc_avg_grade_by_sex+ylab("Number of Students")+xlab("Level of W
eekly Alcohol Consumption")
walc_avg_grade_by_sex
```




```
## Display a boxchart of average grades by daily alcohol consumption
ggplot(comb_math_por_data_no_dupes, aes(x=Dalc, y=avggrades, group=Dalc, color=avggrade
s))+
  geom_boxplot(outlier.colour="red", outlier.shape=16,
               outlier.size=2, notch=FALSE)+ #Highlights outlier observations in red
  stat_summary(fun.y=mean, geom="point", shape=23, size=10)+#Adds a diamond to represent
the mean of each consumption level
  geom_jitter(shape=16, position=position_jitter(0.1))+#Overlays the boxchart with a sca
tterplot showing number of observations
  theme(legend.position="right")+ #Adds legend to the right of the chart
  xlab("Daily Alcohol consumption")+#Labels the x axis
  ylab("Average Grades")+#Labels the y axis
  ggtitle("Average Grade by Daily Alcohol Consumption")#Displays the title of the chart
```




```
## Display a boxchart of average grades by daily alcohol consumption
ggplot(comb_math_por_data_no_dupes, aes(x=Walc, y=avggrades, group=Walc, color=avggrade
s))+
  geom_boxplot(outlier.colour="red", outlier.shape=16,
               outlier.size=2, notch=FALSE)+ #Highlights outlier observations in red
  stat_summary(fun.y=mean, geom="point", shape=23, size=10)+#Adds a diamond to represent
the mean of each consumption level
  geom_jitter(shape=16, position=position_jitter(0.1))+#Overlays the boxchart with a sca
tterplot showing number of observations
  theme(legend.position="right")+ #Adds legend to the right of the chart
  xlab("Weekly Alcohol consumption")+#Labels the x axis
  ylab("Average Grades")+#Labels the y axis
  ggtitle("Average Grade by Weekly Alcohol Consumption")#Displays the title of the chart
```



The bar charts and boxplots above suggest that there are a fewer number of students who consume large amounts of alcohol on a daily basis. The majority of the observations for daily consumption occur at levels 1 and 2, and then show a decline as alcohol consumption increases to level 5. In contrast, there are a greater number of students, who consume larger amounts of alcohol at least once a week, as shown by the more evenly distributed weekly observations. Both boxplots also show a higher average grade for students whose daily and weekly consumption is very low. However, this is insufficient evidence of causality. The following code chunk runs a multimodel regression analysis to identify coefficients of significance. This will help us to identify the best predictor variables for average grade.

There appears to be a difference in alcohol consumption between gender with a larger number of female students consuming smaller amounts of alcohol than their maile counterparts. Furthermore, this behavior is consistent for daily and weekly alcohol consumption. However, this does not imply that gender is significant in determining average final grades. The observations is given only for informational purposes.

#Models and Methods

The following code chunks creates several linear models to determine which of the 30 variables show significance on final grade and warrant further investigation.

```
##multiple regression model
multi_model_regression<-lm(comb_math_por_data_no_dupes$avggrades~., data=comb_math_por_d
ata_no_dupes[1:30])##runs a regression model for all independent variables, located in c
olumns 1 thru 30, against the depoendent variable avggrades.
summary(multi_model_regression)#Summarizes result in tabular format
```

```
##
## Call:
## lm(formula = comb_math_por_data_no_dupes$avggrades ~ ., data = comb_math_por_data_no_
dupes[1:30])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7373  -1.4725   0.1352   1.8551   7.9877
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.561842   1.732544   5.519 4.44e-08 ***
## schoolMS         -0.412755   0.245492  -1.681 0.093037 .
## sexM              0.001044   0.215393   0.005 0.996133
## age               0.022626   0.087662   0.258 0.796384
## addressU          0.254954   0.231408   1.102 0.270858
## famsizeLE3        0.373380   0.211974   1.761 0.078495 .
## PstatusT          0.031347   0.302073   0.104 0.917372
## Medu              0.147378   0.133179   1.107 0.268750
## Fedu              0.067936   0.118120   0.575 0.565334
## Mjobhealth        0.903440   0.464979   1.943 0.052325 .
## Mjobother         0.065328   0.274056   0.238 0.811643
## Mjobservices      0.559211   0.324083   1.726 0.084770 .
## Mjobteacher      -0.001245   0.435529  -0.003 0.997720
## Fjobhealth        0.024220   0.652747   0.037 0.970410
## Fjobother         0.045749   0.408154   0.112 0.910779
## Fjobservices     -0.152078   0.428622  -0.355 0.722817
## Fjobteacher       1.226590   0.576966   2.126 0.033775 *
## reasonhome        0.146675   0.243833   0.602 0.547629
## reasonother       0.145208   0.325850   0.446 0.655970
## reasonreputation  0.333324   0.253634   1.314 0.189109
## guardianmother   -0.178059   0.232585  -0.766 0.444133
## guardianother     0.370334   0.442821   0.836 0.403200
## traveltime       -0.106162   0.138211  -0.768 0.442618
## studytime         0.398494   0.121836   3.271 0.001113 **
## failures         -1.480093   0.153191  -9.662  < 2e-16 ***
## schoolsupyes     -1.365736   0.305068  -4.477 8.53e-06 ***
## famsupyes        -0.305398   0.199457  -1.531 0.126078
## paidyes          -0.664939   0.230106  -2.890 0.003947 **
## activitiesyes     0.103146   0.192791   0.535 0.592769
## nurseryyes       -0.007701   0.235800  -0.033 0.973953
## higheryes         1.355599   0.349932   3.874 0.000115 ***
## internetyes       0.337320   0.245160   1.376 0.169183
## romanticyes      -0.448438   0.199984  -2.242 0.025175 *
## famrel            0.092980   0.100933   0.921 0.357188
## freetime          0.034100   0.098468   0.346 0.729193
## goout            -0.212224   0.094889  -2.237 0.025555 *
## Dalc             -0.125062   0.132462  -0.944 0.345349
## Walc             -0.013376   0.104119  -0.128 0.897803
```

```
## health           -0.139644   0.067679  -2.063 0.039362 *
## absences         -0.013469   0.015514  -0.868 0.385499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.806 on 919 degrees of freedom
## Multiple R-squared:  0.2751, Adjusted R-squared:  0.2444
## F-statistic: 8.944 on 39 and 919 DF,  p-value: < 2.2e-16
```

The multi-linear regression analysis suggest that the top five predictors of average grades (in no particular order)are:

1. failures
2. schoolsupyes
3. higheryes
4. paidyes
5. studytime

Note that neither daily or weekly alcohol consumption appear to have any affect on average grades. The next code creates test and train datasets that will be used to develop and test incremental linear models that determine the effect that the top 5 predictors have on the dependent variable (final grade).

```
set.seed(1) # Set Seed so that same sample can be reproduced for future verification
# Now Selecting 50% of data as sample from total 'n' rows of the data to use as test/tra
in datasets
sample <- sample.int(n = nrow(comb_math_por_data_no_dupes),
                     size = floor(.50*nrow(comb_math_por_data_no_dupes)), replace = F)##
Sampling without rgeplacement
comb_math_por_data_no_dupes_train <- comb_math_por_data_no_dupes[sample, ]
##Save data in df format for later retrieval
save(comb_math_por_data_no_dupes_train, file = "C:/Users/dnico/OneDrive - B&N Enterprise
s/Desktop/Vanderbilt/Data Science/Final Project/comb_math_por_data_train.Rdata")
comb_math_por_data_no_dupes_test  <- comb_math_por_data_no_dupes[-sample, ]##Creates tes
t dataset from remaining rows
##Save data in df format for later retrieval
save(comb_math_por_data_no_dupes_test, file = "C:/Users/dnico/OneDrive - B&N Enterprise
s/Desktop/Vanderbilt/Data Science/Final Project/comb_math_por_data_test.Rdata")
##Checking that dataset was correctly divided
nrow(comb_math_por_data_no_dupes_train)##479 rows
```

```
## [1] 479
```

```
nrow(comb_math_por_data_no_dupes_test)##480 rows
```

```
## [1] 480
```

The code chunks below create incremental models for the top 5 predictors identified in our multi-linear regression model. The models will be run against the training dataset initially, and then tested using the testing dataset. The root mean square error (RMSE)is calculated for each increment to determine if there was improvement in the model. RMSE calculations for the train and test datasets will be compared to determine the suitability of the model.

```
mod_1<-lm(G3~failures,
          data=comb_math_por_data_no_dupes_train)
            train_mod_1_rmse<-rmse(mod_1,
                 data = comb_math_por_data_no_dupes_train);train_mod_1_rmse
```

```
## [1] 3.686785
```

```
mod_2<-lm(G3~schoolsup+failures,
          data=comb_math_por_data_no_dupes_train)
            train_mod_2_rmse<-rmse(mod_2,
                 data = comb_math_por_data_no_dupes_train);train_mod_2_rmse
```

```
## [1] 3.675372
```

```
mod_3<-lm(G3~as.factor(higher)+failures+schoolsup,
          data=comb_math_por_data_no_dupes_train)
            train_mod_3_rmse<-rmse(mod_3,
                 data = comb_math_por_data_no_dupes_train);train_mod_3_rmse
```

```
## [1] 3.617506
```

```
mod_4<-lm(G3~as.factor(higher)+failures+schoolsup+as.factor(paid),
          data=comb_math_por_data_no_dupes_train)
            train_mod_4_rmse<-rmse(mod_4,
                 data = comb_math_por_data_no_dupes_train);train_mod_4_rmse
```

```
## [1] 3.606779
```

```
mod_5<-lm(G3~as.factor(studytime)+as.factor(higher)+failures+schoolsup+as.factor(paid),
          data=comb_math_por_data_no_dupes_train)
            train_mod_5_rmse<-rmse(mod_5,
                 data = comb_math_por_data_no_dupes_train);train_mod_5_rmse
```

```
## [1] 3.565139
```

The code chunk below makes predictions about final grades using the top five variables identified in the multi-linear regression model above and the testing dataset.

```
##Loading test dataset
load(file = "C:/Users/dnico/OneDrive - B&N Enterprises/Desktop/Vanderbilt/Data Science/F
inal Project/comb_math_por_data_test.Rdata")

mod_1<-lm(G3~failures,
          data=comb_math_por_data_no_dupes_test)
            test_mod_1_rmse<-rmse(mod_1,
                 data = comb_math_por_data_no_dupes_test);test_mod_1_rmse
```

```
## [1] 3.513103
```

```
##Add model 1 predictions to testing dataset
comb_math_por_data_no_dupes_test<-comb_math_por_data_no_dupes_test%>%add_predictions(mod
_1,var = "pred1")

mod_2<-lm(G3~schoolsup+failures,
          data=comb_math_por_data_no_dupes_test)
            test_mod_2_rmse<-rmse(mod_2,
                 data = comb_math_por_data_no_dupes_test);test_mod_2_rmse
```

```
## [1] 3.503728
```

```
##Add model 2 predictions to testing dataset
comb_math_por_data_no_dupes_test<-comb_math_por_data_no_dupes_test%>%add_predictions(mod
_2,var = "pred2")

mod_3<-lm(G3~as.factor(higher)+failures+schoolsup,
          data=comb_math_por_data_no_dupes_test)
            test_mod_3_rmse<-rmse(mod_3,
                 data = comb_math_por_data_no_dupes_test);test_mod_3_rmse
```

```
## [1] 3.480498
```

```
##Add model 3 predictions to testing dataset
comb_math_por_data_no_dupes_test<-comb_math_por_data_no_dupes_test%>%add_predictions(mod
_3,var = "pred3")

mod_4<-lm(G3~as.factor(higher)+failures+schoolsup+as.factor(paid),
          data=comb_math_por_data_no_dupes_test)
            test_mod_4_rmse<-rmse(mod_4,
                 data = comb_math_por_data_no_dupes_test);test_mod_4_rmse
```

```
## [1] 3.472225
```

```
##Add model 4 predictions to testing dataset
comb_math_por_data_no_dupes_train<-comb_math_por_data_no_dupes_test%>%add_predictions(mo
d_4,var = "pred4")

mod_5<-lm(G3~as.factor(studytime)+as.factor(higher)+failures+schoolsup+as.factor(paid),
          data=comb_math_por_data_no_dupes_test)
            test_mod_5_rmse<-rmse(mod_5,
                 data = comb_math_por_data_no_dupes_test);test_mod_5_rmse
```

```
## [1] 3.452568
```

```
##Add model 5 predictions to training dataset
comb_math_por_data_no_dupes_test<-comb_math_por_data_no_dupes_train%>%add_predictions(mo
d_5,var = "pred5")
##Verifying that model predictions were added to the table
head(comb_math_por_data_no_dupes_test)
```

```
##    school sex age address famsize Pstatus Medu Fedu     Mjob    Fjob
## 2      GP   F  17       U     GT3       T    1    1  at_home   other
## 6      GP   M  16       U     LE3       T    4    3 services   other
## 7      GP   M  16       U     LE3       T    2    2    other   other
## 8      GP   F  17       U     GT3       A    4    4    other teacher
## 9      GP   M  15       U     LE3       A    3    2 services   other
## 10     GP   M  15       U     GT3       T    3    4    other   other
##        reason guardian traveltime studytime failures schoolsup famsup paid
## 2      course   father          1         2        0        no    yes   no
## 6  reputation   mother          1         2        0        no    yes  yes
## 7        home   mother          1         2        0        no     no   no
## 8        home   mother          2         2        0       yes    yes   no
## 9        home   mother          1         2        0        no    yes  yes
## 10       home   mother          1         2        0        no    yes  yes
##    activities nursery higher internet romantic famrel freetime goout Dalc
## 2          no      no    yes      yes       no      5        3     3    1
## 6         yes     yes    yes      yes       no      5        4     2    1
## 7          no     yes    yes      yes       no      4        4     4    1
## 8          no     yes    yes       no       no      4        1     4    1
## 9          no     yes    yes      yes       no      4        2     2    1
## 10        yes     yes    yes      yes       no      5        5     1    1
##    Walc health absences G1 G2 G3 avggrades    pred1    pred2    pred3
## 2     1      3        4  5  5  6  5.333333 11.61007 11.70790 11.80207
## 6     2      5       10 15 15 15 15.000000 11.61007 11.70790 11.80207
## 7     1      3        0 12 12 11 11.666667 11.61007 11.70790 11.80207
## 8     1      1        6  6  5  6  5.666667 11.61007 10.90813 10.90323
## 9     1      1        0 16 18 19 17.666667 11.61007 11.70790 11.80207
## 10    1      5        0 14 15 15 14.666667 11.61007 11.70790 11.80207
##       pred4    pred5
## 2  11.93588 12.12397
## 6  11.35878 11.41561
## 7  11.93588 12.12397
## 8  11.08934 11.19116
## 9  11.35878 11.41561
## 10 11.35878 11.41561
```

The RMSE calculation from the test dataset consistenly outperform RMSE scores across all models, the lowest score being 3.48 for test data and 3.56 for train data. This suggests that the models are suitable to apply to alternative data sources. The code chunks below perform additional cross validaton using the complete dataset.

```
comb_math_por_data_no_dupes_cv<-comb_math_por_data_no_dupes %>%
  crossv_mc(n=100,test=.2)
comb_math_por_data_no_dupes_cv
```

```
## # A tibble: 100 x 3
##    train      test       .id
##    <list>     <list>     <chr>
##  1 <resample> <resample> 001
##  2 <resample> <resample> 002
##  3 <resample> <resample> 003
##  4 <resample> <resample> 004
##  5 <resample> <resample> 005
##  6 <resample> <resample> 006
##  7 <resample> <resample> 007
##  8 <resample> <resample> 008
##  9 <resample> <resample> 009
## 10 <resample> <resample> 010
## # ... with 90 more rows
```

The next code chunk converts all of the individual training datasets to tibbles. Then model 5, the model returning the lowest RMSE above, is run on each training dataset. Preditction from the model are then generated for each testing dataset, and then calculates the rmse from each of the testing datasets.

```
mod_5_rmse_cv<-comb_math_por_data_no_dupes_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod_5, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse

mod_5_rmse_cv
```

```
## # A tibble: 100 x 2
##    .id    rmse
##    <chr> <dbl>
##  1 001    3.34
##  2 002    3.49
##  3 003    3.62
##  4 004    3.58
##  5 005    3.44
##  6 006    3.81
##  7 007    3.65
##  8 008    3.27
##  9 009    3.35
## 10 010    3.52
## # ... with 90 more rows
```

Creates a summary table and plots each rmse calucalation

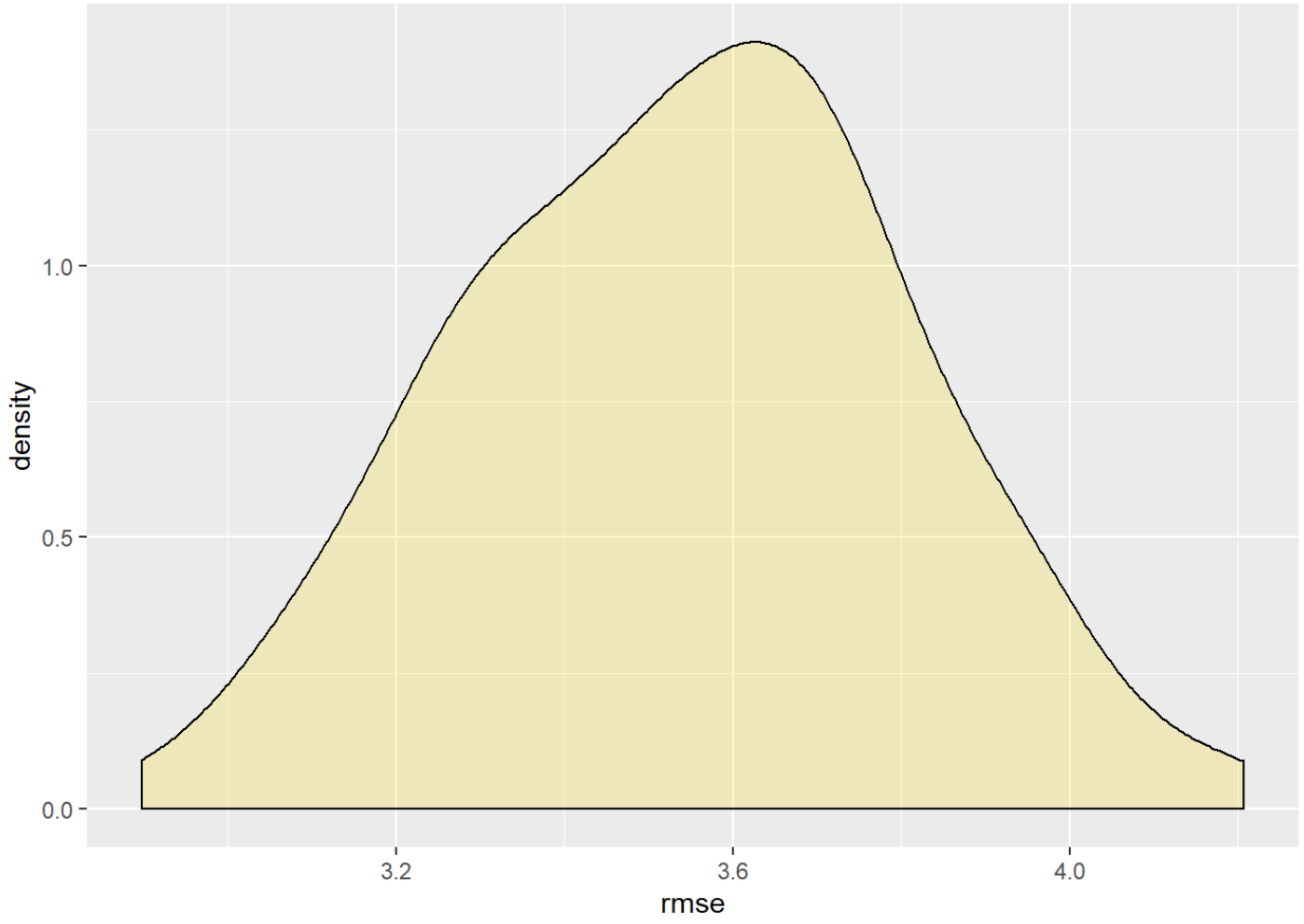
```
summary(mod_5_rmse_cv$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.898   3.345   3.558   3.546   3.715   4.206
```

```
gg<-ggplot(mod_5_rmse_cv,aes(rmse))
gg<-gg+geom_density(bins=50,fill="gold",alpha=.2)
```

```
## Warning: Ignoring unknown parameters: bins
```

```
gg
```



#Summary

This study attempted to find a causal relationship between final grade and alcohol consumption in college students. Data was colleced from college students in a Portuguese lanquage and mathematics class containing a total of 959 observations, after the removal of duplicate records. A multiple linear regression model did not show that alcohol consumption was a significant variable in predicting final grades. The top five predictors were:

1. failures
2. schoolsupyes

3. higheryes
4. paidyes
5. studytime

Cross validation of the model using 100 datasets showed a normally distributed rmse distribution with a mean of 3.568. This suggests that the model is valid as a predictor. However, it does not suggest that it is the best model to predict final grades as their may be other variables not included in this study that show greater levels of significance.

#References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.