Jack Dumanski, Ryan O'Neill

Dr. Perry

STAT 351 001

7 November 2025

Exploratory Data Analysis

Our data set (MathScores.csv) is surrounding students and their success rate in their math classes based on several factors on the personal and academic lives of the students. The target variable is categorical and called "G4" which has entries "F" – Fail, "NI" – Needs Improvement, "S" – Satisfactory, and "E" – Excellent. This variable is related to the student's success in their math class. This model will aim to explain the students' math class success based primarily on variables outside of school performance. To outline each factor, we made plots and calculations below for all types of variables:

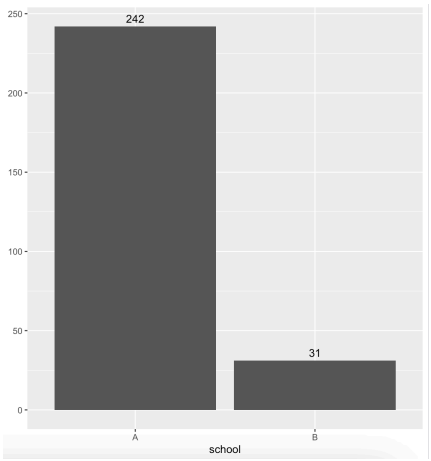
<u>Categorical Variables:</u>

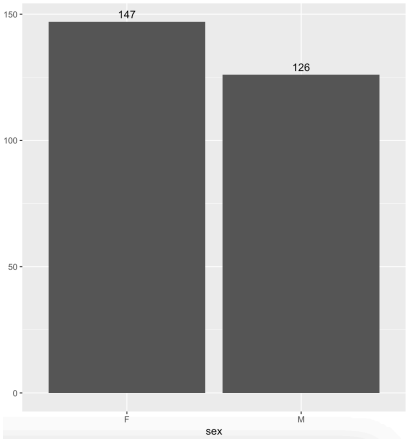
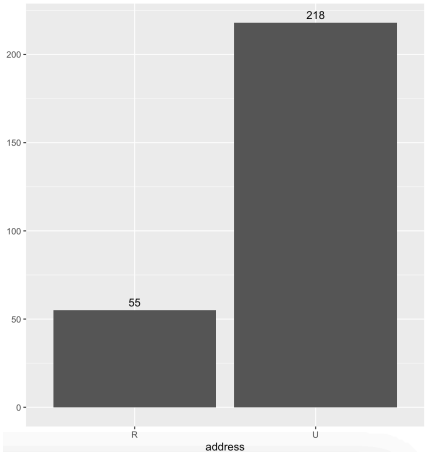
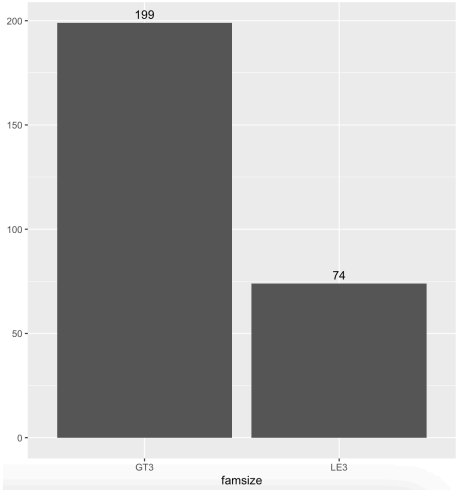
library(ggplot2)

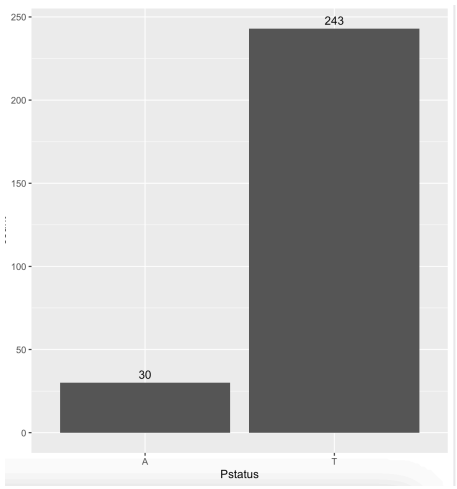
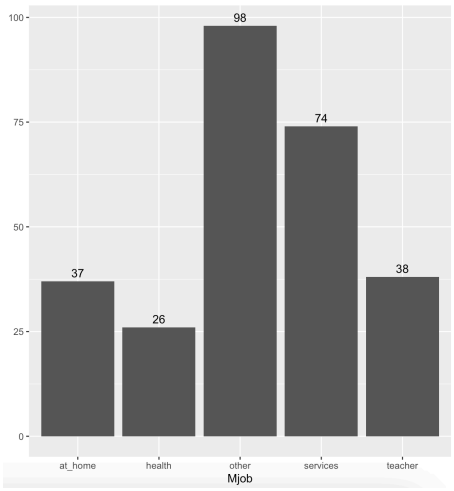
School:

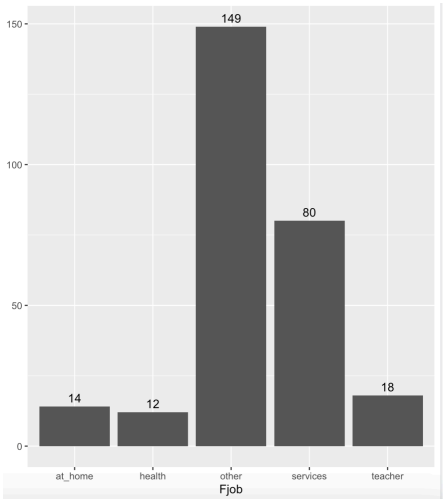
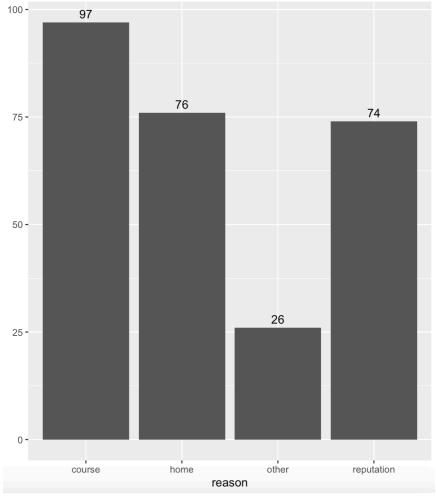
```
ggplot(MathScores, aes(x=school)) + geom_bar() +
geom_text(stat="count", aes(label=..count..), vjust=-.5)
```

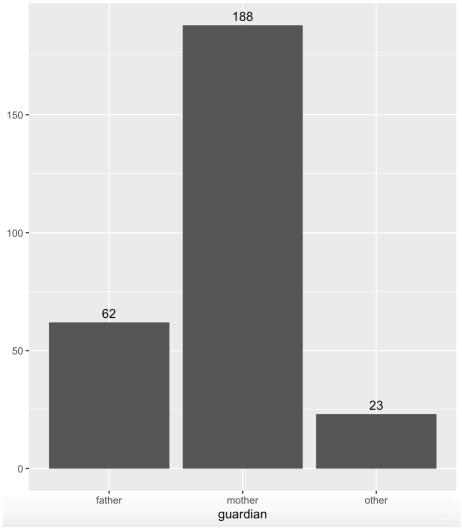
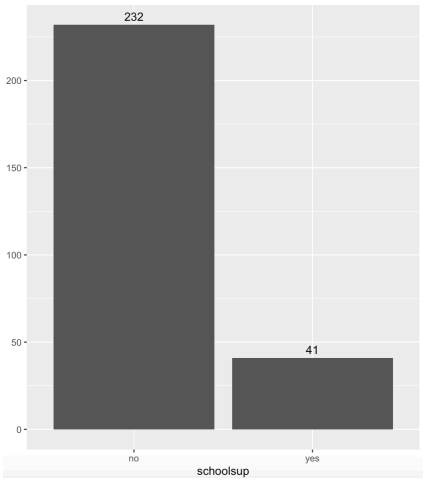
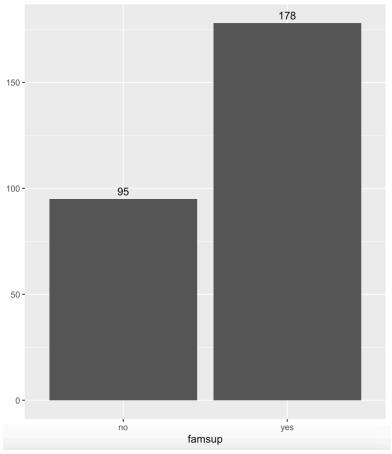
The vast majority of students in the dataset attended school A as opposed to school B.

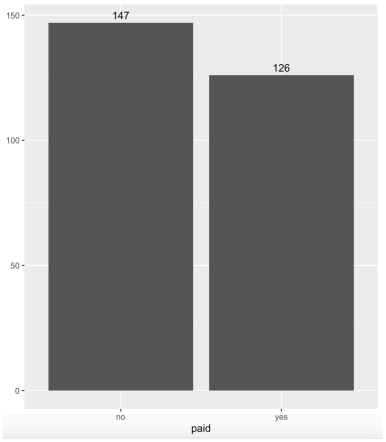
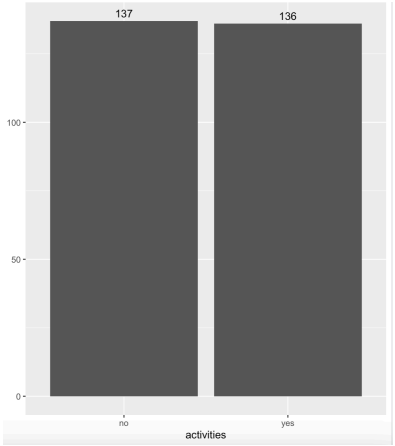


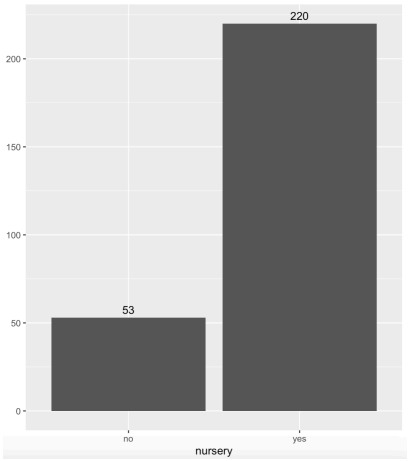
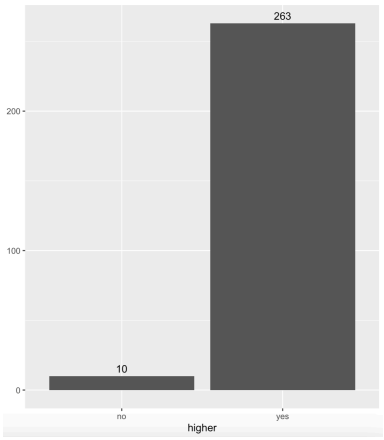
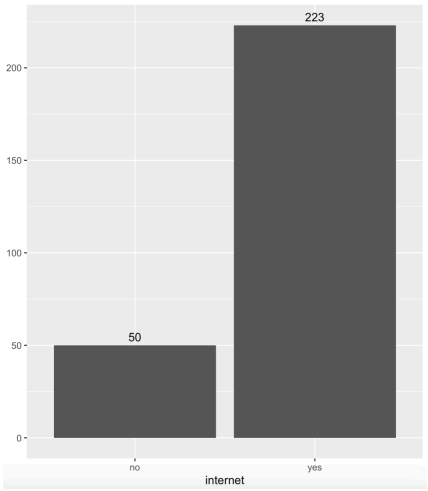
| | |
|---|---|
| Sex:<br>ggplot(MathScores, aes(x=sex)) + geom_bar() +<br>geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The distribution of Male and Female students is fairly even. There are slightly more female than male students. |  |
| Address:<br>ggplot(MathScores, aes(x=address)) + geom_bar() +<br>geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students in this study are from urban addresses rather than rural. |  |
| Family size: ggplot(MathScores, aes(x=famsize)) +<br>geom_bar() + geom_text(stat="count",<br>aes(label=..count..), vjust=-.5)<br><br>The family size for most students is greater than 3, which is around three times the amount of families with size less than 3. |  |

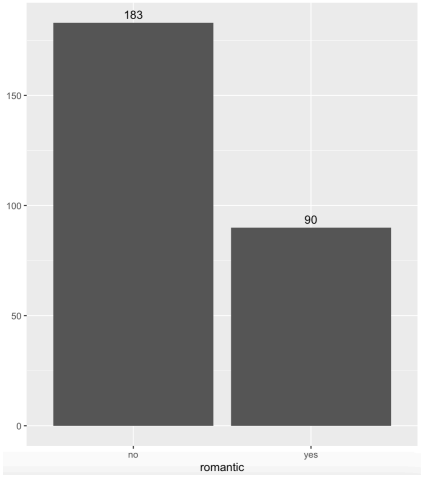
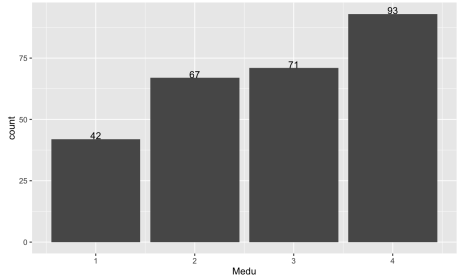
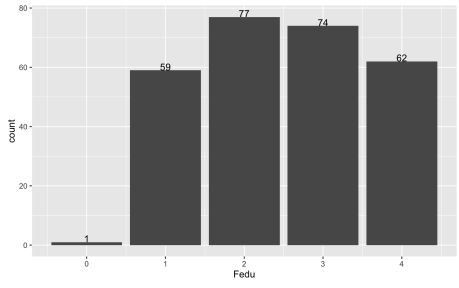
| | |
|---|---|
| Parent Status: ggplot(MathScores, aes(x=Pstatus)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>An overwhelming majority of students in this sample have their parents together rather than apart. |  |
| Mother's Job: ggplot(MathScores, aes(x=Mjob)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>Out of the common jobs listed (at-home work, healthcare, services, teaching) for the students' mothers, service jobs are the most popular. However, a large number of students' mothers do not have a common job. |  |

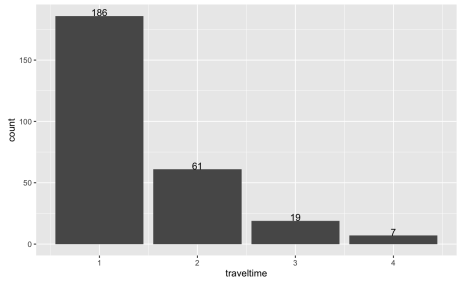
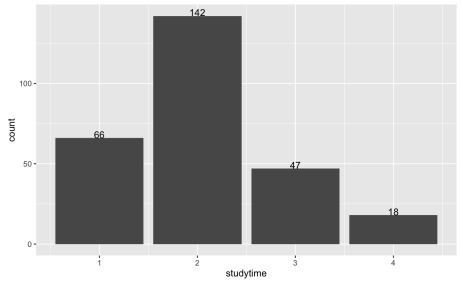
| | |
|---|---|
| Father's job: ggplot(MathScores, aes(x=Fjob)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>Students' fathers tend to have the most jobs in services rather than the other listed common jobs, but the majority of jobs fall in the "other" category. |  |
| Reason for choosing school: ggplot(MathScores, aes(x=reason)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The most common reason students chose their school is due to their course offerings, with proximity to home and the school's reputation not too far behind. |  |

| | |
|---|---|
| Guardian choice: ggplot(MathScores, aes(x=guardian)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students list their mother as their primary guardian, around 3 times as common as they list their father. |  |
| School support: ggplot(MathScores, aes(x=schoolsup)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students do not have additional school support, but 41 students do. |  |
| Family support: ggplot(MathScores, aes(x=famsup)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5) |  |

| | |
|---|---|
| The majority of students do have family educational support, and it is around twice as often that students do have family educational support than those who don’t. | |
| Extra math classes: ggplot(MathScores, aes(x=paid)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>There is a roughly even split between students who have extra paid Math classes and those who don’t. |  |
| Activities: ggplot(MathScores, aes(x=activities)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>There is a nearly even split of students who participate in extra-curricular activities and those who don’t. |  |

| | |
|---|---|
| Nursery school: ggplot(MathScores, aes(x=nursery)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students attended nursery school, but there is still a strong population of students who didn't. |  |
| Pursuing higher education: ggplot(MathScores, aes(x=higher)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>Nearly every student in this study has goals to pursue higher education. |  |
| Internet access: ggplot(MathScores, aes(x=internet)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>A large majority of students in this study have internet access at home, but there is a strong population of students without home internet access. |  |

| | |
|---|---|
| Romantic relationship: ggplot(MathScores, aes(x=romantic)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students in this study do not consider themselves in a romantic relationship, which is around twice as often as those who are in a romantic relationship |  |
| Mother's education: ggplot(MathScores, aes(x=Medu)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>There is a slight negative skew in the level of mother's education where the highest bar is higher education, then high school, then middle school, then some elementary school. |  |
| Father's education: ggplot(MathScores, aes(x=Fedu)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5) |  |

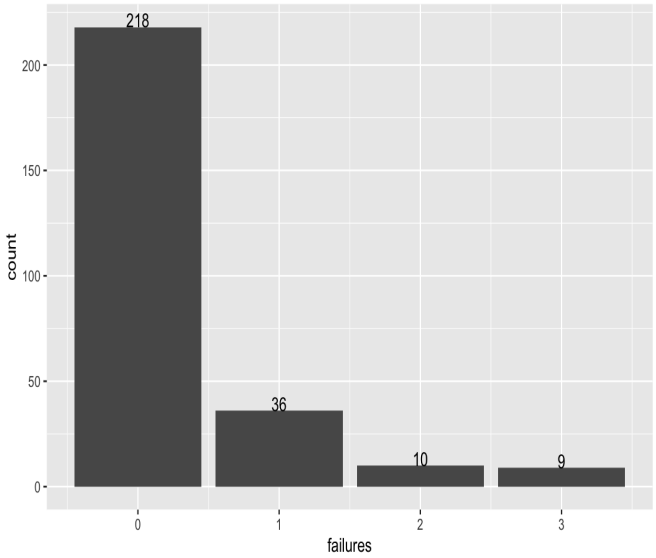
| | |
|---|---|
| The bars for father's education are all roughly equal with the same strata as the mother's education, but there is one student whose father had no education at all. | |
| Traveltime: ggplot(MathScores, aes(x=traveltime)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The majority of students are less than 15 minutes away from their school, and there is a positive skew in the data in terms of distance away per student. |  |
| Studytime: ggplot(MathScores, aes(x=studytime)) + geom_bar() + geom_text(stat="count", aes(label=..count..), vjust=-.5)<br><br>The most common amount of time spent studying per student is 2-5 hours per week, with 0-2 and 5-10 being the next highest bars, and the fewest students studying 10+ hours. |  |

Numerical Variables:

Failed classes: ggplot(aes(x = failures))

+

geom_bar()




The majority of students did not fail any classes, and this distribution has a strong positive skew, with very few students failing any classes, especially more than 1.

summary(scores$failures)

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 0.000 | 0.304 | 0.000 | 3.000 |

mean(scores$failures)

0.3040293

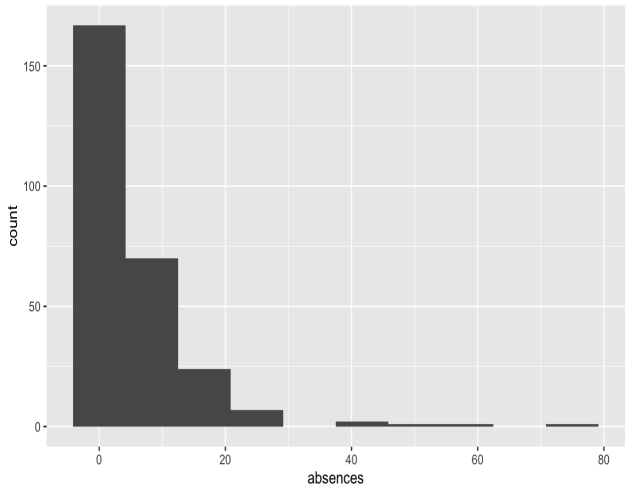
sd(scores$failures)

0.6960117

The average amount of classes failed by students is 0.3 with a standard deviation of 0.69.

Absenses: ggplot(aes(x = absences)) +

geom_histogram(bins = 10)




This distribution of absences per student has a strong positive skew, and students usually missed no more than 10 classes, and virtually no students missed more than 20.

summary(scores$absences)

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 4.000 | 6.011 | 8.000 | 75.000 |

mean(scores$absences)

6.010989

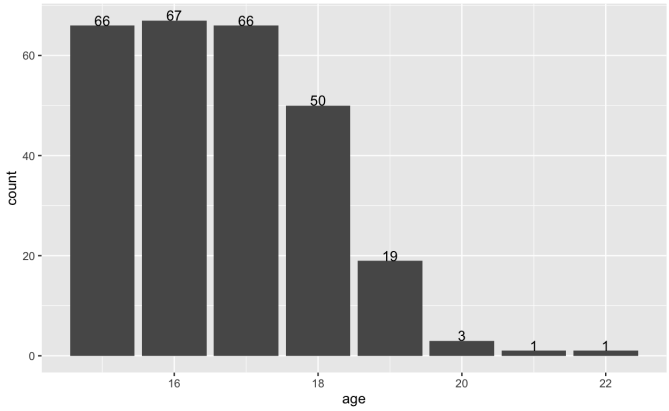
sd(scores$absences)

8.739483

The average absences per student is 6.01 with a standard deviation of 8.74, which is influenced by the large outliers.

Age: ggplot(aes(x = age)) + geom_bar() +

geom_text(stat = "count", aes(label = ..count..), vjust = 0 )




The majority of the students polled were from ages 15-18, which is traditional for high school students. A few outliers were older, up to 22 years old.

summary(scores$age)

Min. 1st Qu. Median Mean 3rd Qu. Max.

15.00 16.00 17.00 16.66 18.00 22.00

mean(scores$age)

16.65934

sd(scores$age)

1.346726

The average age of students polled is 16.65 with a standard deviation of 1.35, which means 68% of students are aged between 15.3 and 18.

Family Relationship: ggplot(aes(x = Famrel)) + geom_bar() +

geom_text(stat = "count", aes(label = ..count..), vjust = 0 )

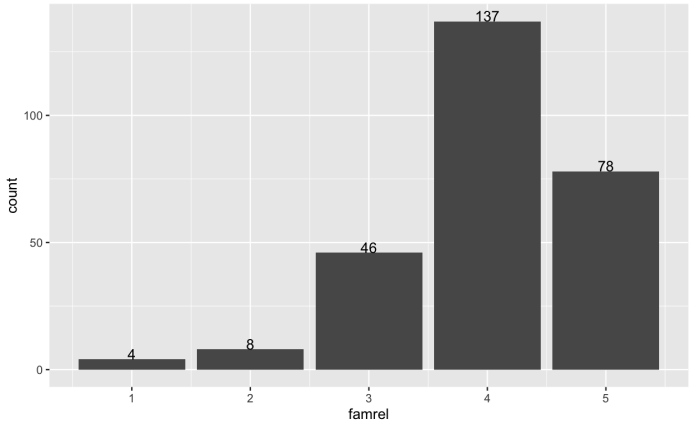
The majority of students had a 3+ relationship with their family on a 1-5 scale of quality of their family relationship. This distribution is negatively skewed.

summary(scores$famrel)

mean(scores$famrel)

sd(scores$famrel)

The average rating of students' family relationship is 4.01/5, with a 0.84 standard deviation.




Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 4.000 4.000 4.015 5.000 5.000

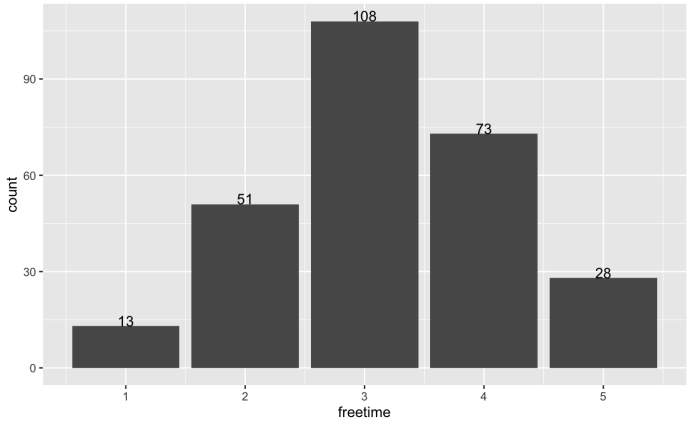
4.014652

0.8400398

Free time: ggplot(aes(x = freetime)) +

geom_bar() +

geom_text(stat = "count", aes(label = ..count..), vjust = 0 )




This graph is roughly normally distributed, with the majority of students centered around ¾ amount of freetime.

summary(scores$freetime)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 3.00 3.00 3.19 4.00 5.00

mean(scores$freetime)

3.190476

sd(scores$freetime)

1.011142

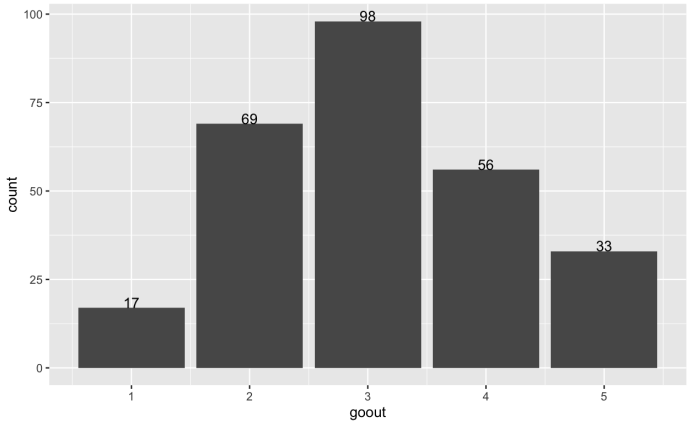
The average rating of students' free time from 1-5 is 3.2 with a standard deviation of 1.01.

Going out: ggplot(aes(x = goout)) +

geom_bar() +

geom_text(stat = "count", aes(label =

..count..), vjust = 0 )




The distribution for students' going out time rating is approximately normal centered around ¾.

summary(scores$goout)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 2.00 3.00 3.07 4.00 5.00

mean(scores$goout)

3.069597

sd(scores$goout)

1.090867

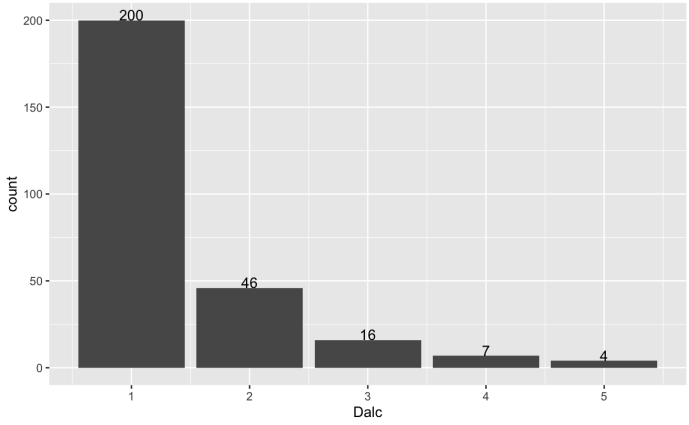
The average rating out of 5 for students' going out time is 3.07 with a standard deviation of 1.09.

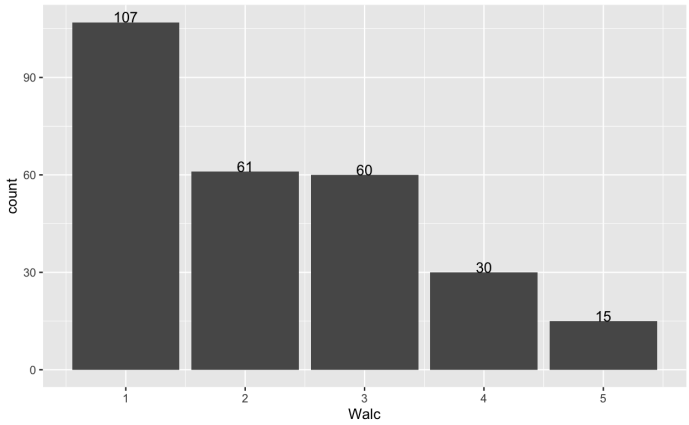
Work day alcohol consumption:

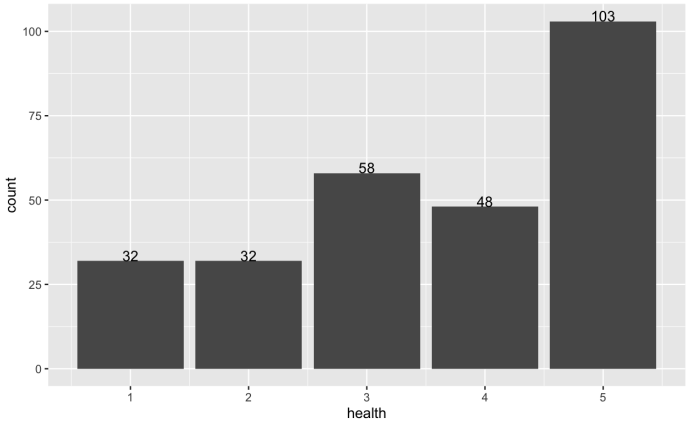
ggplot(aes(x = Dalc)) + geom_bar() +

geom_text(stat = "count", aes(label =

..count..), vjust = 0 )



| | |
|---|---|
| This distribution of rating of workday alcohol consumption is strongly positively skewed.<br><br>summary(scores$Dalc)<br>mean(scores$Dalc)<br>sd(scores$Dalc)<br>The average rating out of 5 of alcohol consumption on a workday is 1.42 with a standard deviation of 0.832. | Min. 1st Qu. Median Mean 3rd Qu. Max.<br>1.000 1.000 1.000 1.421 2.000 5.000<br>1.421245<br>0.8326004 |
| Weekend Alcohol consumption:<br>ggplot(aes(x = Walc)) + geom_bar() +<br>geom_text(stat = "count", aes(label = ..count..), vjust = 0 )<br><br>This distribution has a strong positive skew, with the majority of students rating ⅗ or less for their weekend alcohol consumption.<br><br>summary(scores$Walc)<br>mean(scores$Walc) |  <br>Min. 1st Qu. Median Mean 3rd Qu. Max.<br>1.000 1.000 2.000 2.212 3.000 5.000<br>2.212454<br>1.22726 |

| | |
|---|---|
| sd(scores$Walc)<br>The average weekend alcohol consumption rating out of 5 is 2.21 with a standard deviation of 1.227. | |
| Health rating: ggplot(aes(x = health)) + geom_bar() + geom_text(stat = "count", aes(label = ..count..), vjust = 0 )<br><br>This distribution is largely negatively skewed, with the majority of students rating their health to be "very high" or close to it.<br>summary(scores$health)<br>mean(scores$health)<br>sd(scores$health)<br>The average health rating for students is 3.58 with a standard deviation of 1.39, showing high average health. |  <br><br>Min. 1st Qu. Median Mean 3rd Qu. Max.<br>1.000 3.000 4.000 3.579 5.000 5.000<br>3.578755<br>1.393995 |

Conclusion:

Most of the variables in this dataset are categorical or based on perceived ratings on 1-5 scales, so a tree diagram may be the best method to form a model for this data. We may first form a tree with every variable, factored, in a tree model but prune the tree until its predictive ability is adequate enough, or we may begin with variables we think are most important and then test for significance in each additional variable. The most pertinent variables we noticed are Studytime, Schoolsup, Famsup, Paid, Internet, Famrel, Absences. To us, these represented direct connections to success in school or psychological effects that may prohibit students from succeeding. Again, since the data has very few numerical variables, the tree will do well to split the data based on categorical strata, for example health <= 3. This is nonparametric because we can't make assumptions about the underlying distributions of these variables and simply add them as categories into our tree.