How Do Family Factors Affect Student Performance in Portugese Class?

Name: Izzy Rankin, Nina Sun, Jeanny Zhang

March 16, 2020

Introduction

What contributes to students' failure in school? Does life outside of school have an impact on one's performance in school? Our project will explore these larger questions through the following specific research question: how does family background and family life affect secondary school student performance? Our data was collected from secondary schools in Portugal, and contains data on students' performance in school, their family makeup, and the quality of family life. Previous research by Paulo Cortez and Alice Silva, which uses the same data as we do, investigates what factors affect student achievement and whether it is possible to predict a students achievement based on certain factors. They found that they could predict a student's final grade relatively accurately if the first and second period school grades were given, and that previous performance in school strongly affects a student's current school achievements. Their study particularly focused on the factors of previous grades and performance in school in predicting the final grade (Cortez and Silva, 2008). A different study by M. Livaditis et al. investigates the relationship between secondary school performance and family and psychological factors, in which secondary school students in Greece were randomly sampled. The study found that male gender, low socioeconomic status, low parental education, and parental separation are all positively associated with school failure. This study focused not on the effect of previous school performance on school failure, but on the effect of familial and demographic factors on school failure (M. Livaditis et al., 2003). The second study suggests that there is a relationship between family and demographic factors and school performance, while the first study provides information on our data set and the context of our research.

Our research will blend elements of both of these studies, using the same data set and a focus on what factors affect final grades from the first study, but taking the approach of the second study, which looks at how family and demographic data affects a student's performance in school. We are interested in exploring the relationship between family and school performance, so we specifically focused our analysis on the relationship between a student's performance, based on Portugese final grade, and family demographic variables. We hypothesize that there is a relationship between family background and life and a student's achievement in school, and that students whose parents have a higher education and higher-performing jobs will have higher achievement in school, and those students whose parents have less education and a lower-performing job or no job will have lower achievement in school. We also hypothesize that students who have a stable and supportive family life will be higher achieving in school than those students who have a less stable and less supportive family life.

Materials and Methods

We found our raw data set on Kaggle, and the data that we use in our research is collected from two public schools located in the Alentejo region in Portugal from the 2005-2006 school year. The data itself was collected from school reports, which include the variables G3, absences, and failures, and a questionnaire given to the students which covers the remaining variables which relate to family life and background. The key variables that we will be focusing on are: "G3", the student's final grade in Portugese class which ranges from 0 to 20, with 0 to 3.4 defined as a poor grade, 3.5 to 9.4 defined as a poor grade, 9.5 to 13.4 defined as a sufficient grade, 13.5 to 15.4 defined as a good grade, 15.5 to 17.4 defined as a very good grade, and 17.5 to 20 defined as an excellent grade; "familysize", the size of the student's family which is a binary of either less than three or larger than 3; "Pstatus", the student's parent's cohabitation status which is a binary

of either living together or living apart; "famrel", the quality of family relationships which ranges from 1 to 5, with 1 defined as a very bad relationship and 5 defined as an excellent relationship; "famsup", family educational support which is a binary of either yes or no; "guardian", which is the student's guardian, defined as either mother, father, or other, which includes any guardian that is not the mother or father of the student; "Mjob" and "Fjob", which are the job of the mother and father, including options such as teacher, health care related, civil services (e.g. administrative or police), at home, or other; and "Medu" and "Fedu", which are the mother and father's education that range from 0 to 4, with 0 defined as none, 1 defined as primary education, 2 defined as 5th to 9th grade, 3 defined as secondary education, and 4 defined as higher education. We did not make any modifications to this data.

In our research, we used data visualization such as boxplots and tables to summarize statistics, and analysis of variance tests (ANOVA), t-tests, permutation tests, linear regression, and multivariable regression to quantify the association between predictor and outcome variables. We performed ANOVA tests for the predictor variables which have multiple subsets (mother and father's education and job and the quality of family relationships) to see if there is a difference between the mean final grades when these subsets are different. We also performed permutation tests for the predictor variables with a binary response (family size, parent's cohabitation status, and family support for education) to see whether or not the differences were significant and had an effect on final grade. We used multivariable regression to look at all of the variables together in order to determine which have the most significance and greatest effect out of all of the variables on the final grade. It shows us how significant each variable is and how much it influences the final grade by keeping all other variables the same. From this test, we determined which variables were most significant, and thus which ones we should perform more tests on.

Results

We base secondary school performance as a whole on secondary school student's final grades in Portugese class. We decided to use this final grade as a marker of school performance as a whole, with a poor grade indicative of lower success in school and a good grade indicative of high success in school. We use the student's guardian, quality of their family relationship, family size, parents' cohabitation status, and family support for their education as reflective of overall family life, as these variables are likely indicative of a stable, supportive, and healthy home environment, which potentially leads to a higher grade. We used mother and fathers' education and job as reflective of family background and we explored the relationship both of these have to a student's success in school by examining the association between each variable and the final grade.

Initially, we performed a multiple variable regression analysis on the student's final grade against all the other variables in the dataset as shown in Table 1. We discovered that the variables of mother's and father's education levels, some of mother's and father's jobs, and having "other" as the guardian all have significance and impact on a student's final grade. We specifically found that the mother having a healthcare related job has the most significant impact on a student's final grade. Thus, we decided to explore these significant variables in depth with more tests.

Table 1: Multivariable Regression of Final Grade and All Other Variables

	estimate	std.error	p.value	adj.r.squared
(Intercept)	10.186	0.300	< 0.001	0.056
Medu	0.684	0.109	< 0.001	
(Intercept)	10.471	0.288	< 0.001	0.043
Fedu	0.622	0.113	< 0.001	
(Intercept)	11.812	0.151	< 0.001	0.000
famsize LE3	0.318	0.278	0.252	
(Intercept)	11.044	0.273	< 0.001	0.038
Mjob health	2.018	0.533	< 0.001	
Mjob other	0.626	0.337	0.063	

	estimate	std.error	p.value	adj.r.squared
Mjob services	1.103	0.385	0.004	
Mjob teacher	2.094	0.462	< 0.001	
(Intercept)	11.429	0.495	< 0.001	0.014
Fjob health	1.137	0.832	0.172	
Fjob other	0.462	0.523	0.377	
Fjob services	0.201	0.549	0.714	
Fjob teacher	2.155	0.729	0.003	
(Intercept)	12.203	0.261	< 0.001	0.005
guardian mother	-0.306	0.301	0.310	
guardian other	-1.300	0.567	0.022	
(Intercept)	11.665	0.204	< 0.001	0.002
famsup yes	0.392	0.260	0.132	
(Intercept)	11.064	0.536	< 0.001	0.002
famrel	0.214	0.133	0.107	
(Intercept)	11.913	0.361	< 0.001	-0.002
Pstatus T	-0.007	0.386	0.985	

We first generated boxplots comparing student's final grade and mother's and father's education levels (Figure 1 and 2), which both show a general increase in the final grade as the education level increases. We then performed analysis of variance tests to determine whether students with different parents' education have the same mean final grade or not. This yielded a significant p-value of 5.75e-10 for mother's education level and 5.12e-08 for father's education level relative to a significance level of 0.05. Thus, there seems to be strong evidence that there is a difference between the mean final grades when mother and father's educational level is different. Later, we performed permutation tests that compared the subsets of the variables to one another to determine which levels produces a significant difference in mean final grades and to see if the trend we saw in the boxplot was correct. For the permutation test for mother's education, we found a significant p-value of 0.0156 between the primary education and 5th to 9th grade education level and a p-value of 0.002 between the secondary education and higher education level. For the permutation test for father's education, we only found one significant p-value of 0.0176 between the primary education and 5th to 9th grade education level. These results match the multiple regression analyses on the student's final grade against mother's and father's education levels where the mother's education has a much lower lower p-value (0.000414) than that of the father's education (0.0528). The p-values indicate that there is a extremely small chance (less than 5%) of yielding these results given there is no difference in mean final grade between different education levels. The reason that we conducted a separate multiple regression analysis other than the one shown in Table 1 is because we were trying to limit the effects of other variables by just focusing on mother's and father's education.

Figure 1. Final Grade Grouped by Mother's Education

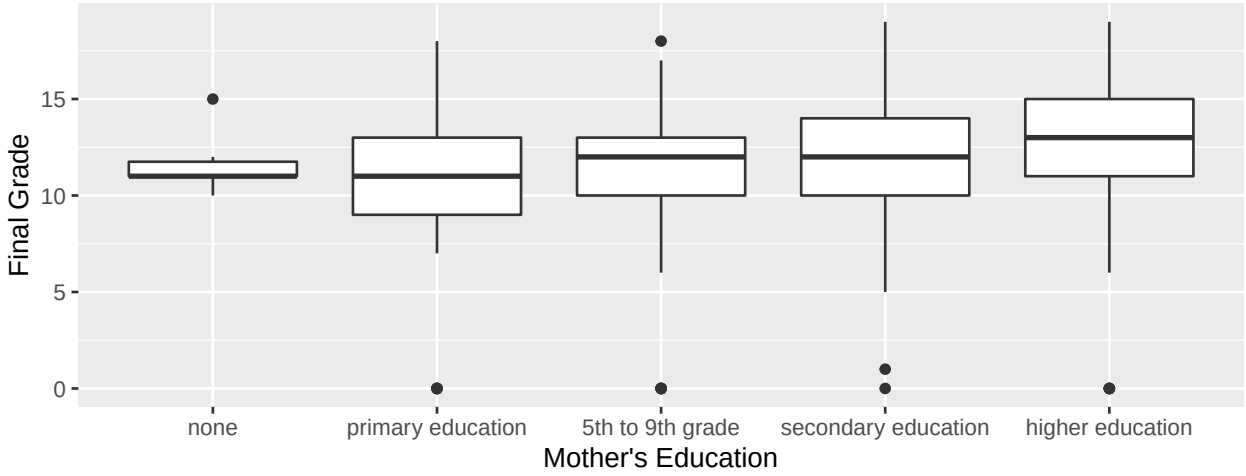
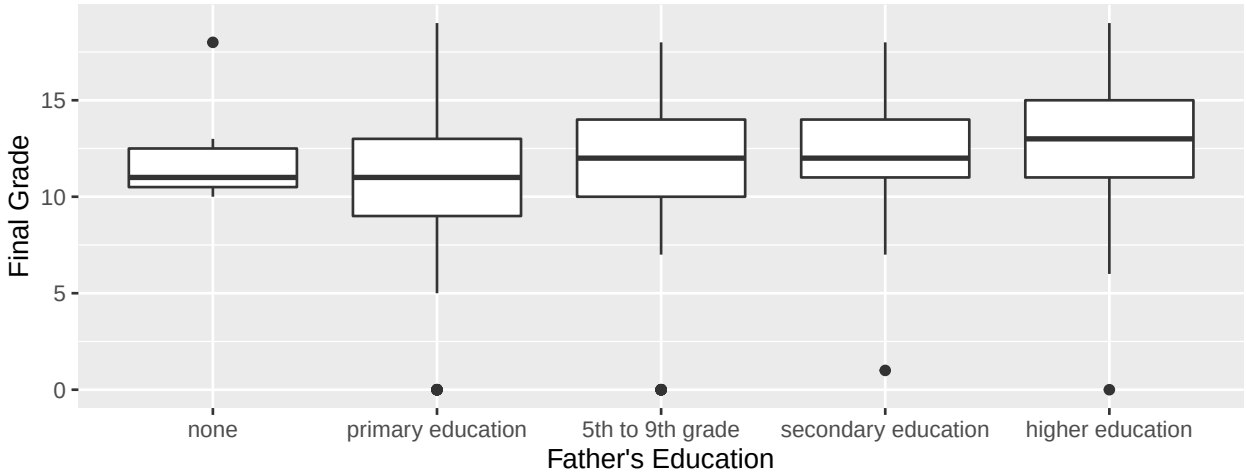


Figure 2. Final Grade Grouped by Father's Education



We also generated boxplots comparing student's final grade and mother's and father's jobs (Figure 3 and 4). We obsereved that both plots have the lowest mean final grade for "at home", which means they stay at home and don't work. We then performed analysis of variance tests to determine whether students with different parents' job have the same mean final grade or not. For the ANOVA for mother's job we found a p-value of 8.31e-06, which, relative to a significance level of 0.05, shows that this variable is significant. For the ANOVA for father's job we found a p-value of 0.0114, which also shows that this variable is significant. These p-values indicate that there is strong evidence that there is a difference between the mean final grades for different parents' jobs. Since "at home" has a obvious lower mean shown in the boxplot, we decided to perform permutation tests for the differences in final grade between each job and "at home". For mother's job, we found significance between "at home" and all jobs except for "other", with a p-value of 0.0002 between "at home" and health care related job, 0.0036 between "at home" and civil services, and 0.0002 between "at home" and teacher. There is strong evidence that there is a difference between the mean final grades for different jobs. The p-value between "at home" and "other" jobs is actually 0.07, which could be considered as weak evidence. For father's job, we only found significance between "at home" and teacher with a p-value of 0.0044. The p-values indicate that there is a extremely small chance (less than 5%) of yielding these results given there is no difference in mean final grade between different jobs. These results align perfectly with the multiple regression analyses on the student's final grade against mother's and father's jobs as shown in Table 2 and 3, which display statistical significance in the same subsets.

Figure 3. Final Grade Grouped by Mother's Job

Figure 4. Final Grade Grouped by Father's Job

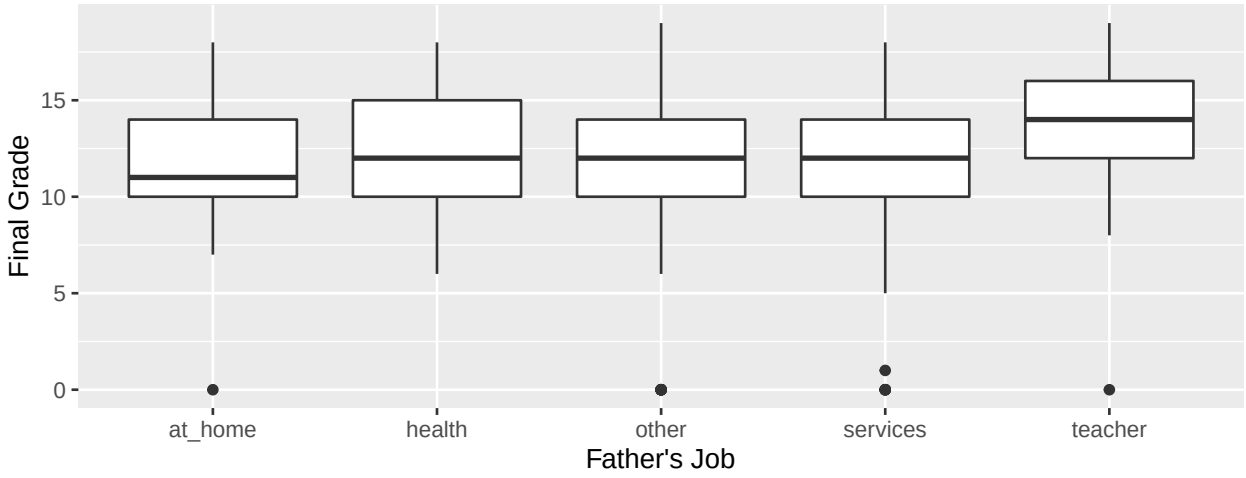


Table 2: Multiple Regression of Final Grade and Mother's Job

	estimate	std.error	p.value	adj.r.squared
(Intercept)	11.044	0.273	< 0.001	0.038
Mjob health	2.018	0.533	< 0.001	
Mjob other	0.626	0.337	0.063	
Mjob services	1.103	0.385	0.004	
Mjob teacher	2.094	0.462	< 0.001	

Table 3: Multiple Regression of Final Grade and Father's Job

	estimate	std.error	p.value	adj.r.squared
(Intercept)	11.429	0.495	< 0.001	0.014
Fjob health	1.137	0.832	0.172	
Fjob other	0.462	0.523	0.377	
Fjob services	0.201	0.549	0.714	
Fjob teacher	2.155	0.729	0.003	

After performing permutation tests for the variables family size, parents' cohabitation status, and family support for education, we found that none of these variables are significant compared to a 0.1 significance level. A higher significance level is used here to avoid Type II error. For the permutation test for family size, we found a p-value of 0.249. For the permutation test for parents' cohabitation status, we found a p-value of 1.0. For the permutation test for family support for education, we found a p-value of 0.135. For quality of family relationship we performed an analysis of variance test and found a p-value of 0.107. These p-values are not significant, which means that there is a high chance that there is no difference in mean final grades caused by the difference in these variables.

To find out if the student's guardian has an effect on their final grade, we performed a multivariable linear regression where the father as guardian was the reference because there are more than two categories within the variable. We found a p-value of 0.310 for mother as the guardian taking father as guardian as reference, which means that it is not significant and there is weak evidence for there being a difference in having a mother versus a father as guardian towards the final grade. We found a p-value of 0.0221 for someone who is not the mother or father as the guardian taking father as guardian as reference, which, relative to a significance level of 0.05, means that it is significant and there is strong evidence for there being a difference in having father versus some other guardian towards the final grade. Thus, taking father as a reference, there is a difference in final grade for those students who have a guardian that is not one of their parents.

Discussion

In our research, we found that the quality of family life and home environment does not have significant effect on a student's achievement in school, but parent's education level and job does have an effect on a student's achievement in school. Out of all of the variables we used to be indicative of family and home environment, guardian was the only variable to have any significance, and it only had an effect on a student's achievement in school when the guardian was someone who is not the mother or father of the student. We were surprised that none of the other family environment related variables had any significance, as we assumed that at least some of them would have some sort of relationship, whether it be positive or negative, with students' final grade. This indicates that the quality and stability of home life does not have a great enough effect on students' school life that it would affect their final grade. We did find that parent's education level and job have an effect on the student's final grade, but specifically within the variables of parents' job, only father's job as teacher was significant and had an effect on the final grade while all of mother's job except for "other" were significant and had an effect on the final grade. Generally, mother's job has more of an effect on final grade than father's job does. For the parent's education level, we also found that mother's education level had more significance than father's overall, and for mother's education level, higher education generally results in a higher final grade, while father's does not. These findings align with those in the study of M. Livaditis et al. in which students whose parents have a lower education level and lower-level job will likely have less success in school (M. Livaditis et al., 2003). This research has important implications for the schools in Portugal, as it shows that there may be outside factors, other than just a student's previous performance in school, that have an effect on a student's overall achievement in school.

Because we narrowed our research to the relationship between family life and background and a student's achievement in school based on final grade, we did not use a majority of the data from the raw data set. Much of this data included variables related to previous performance in school, social life, and other demographic data about the student. Many of these variables could be possible confounding variables, as they could have more of an effect on the final grade than family life or demographic does, but we did not take any of them into account in our research so it is impossible to know for certain. Our research was limited by how general some of the variables and subsets of variables were, specifically mother and father's job grouped many types of jobs into the category of "other". Without more knowledge about "other" and what is included in it, we cannot reasonably infer why it is less significant than the other types of jobs. Due to this, we are not able to make as many strong conclusions about the relationship between mother and father's job and final grade. Additionally, some of the variables are vague in what they mean, such as the variable of guardian, which is unclear about what the role of guardian is and how it is determined. The greatest weakness of our analysis is not having the ability to fully analyze and know which type of job or education level results in the highest

mean grade on the final tests. We are not able to rank the types of jobs and education levels based on mean final grades. Our analysis is also weak in that we did not look at relationships between education level and job, which might have more association with the final grade. On the other hand, a strength of our analysis is the use of multivariable regression which directly shows which variables are most significant and have the greatest effect out of all of the variables on the final grade.

Additionally, because our data is taken from Portugese secondary school students, it cannot be generalized to a larger population due to the cultural specificities of school, family life, social life, etc. in Portugal. It is impossible to know whether or not these findings are applicable in other countries or other situations, as it may be unique due to Portugese family life or the relationship between students, their families, and school in Portugal. We suggest performing similar studies on the relationship between family life and background and a student's success in school in other countries around the world to see if there is a relationship or not, and whether or not it is similar to the relationship found in this study.

Work Cited

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. http://www3.dsi.uminho.pt/pcortez/student.pdf

Livaditis, M., Zaphiriadis, K., Samakouri, M., Tellidou, C., Tzavaras, N., & Xenitidis, K. (2003). Gender Differences, Family and Psychological Factors Affecting School Performance in Greek Secondary School Students. Educational Psychology, 23(2), 223–231. https://doi.org/10.1080/01443410303226

Uci. (2016, October 19). Student Alcohol Consumption. Retrieved from https://www.kaggle.com/uciml/student-alcohol-consumption

Appendix

```
# Variable 1: Mother's Education
momedu<- table(spf$Medu)
momedu
prop.table(momedu)
# ANOVA of Mother's Education and Corresponding Mean Final Grade
spf %>% group_by(Medu) %>% summarise(mean(G3), sd(G3))
ggplot(spf, aes(sample = G3)) + geom_qq() + geom_qq_line() + facet_grid(. ~ Medu)
medu.aov <- aov(G3 ~ Medu, data = spf)
summary(medu.aov)
# Figure 1: boxplot of Final Grade Grouped by Mother's Education
spf$Medvar <- factor(spf$Medu)
levels(spf$Medvar)
levels(spf$Medvar) <- c("none", "primary education", "5th to 9th grade",
                        "secondary education", "higher education")
ggplot(spf, aes(x = Medvar, y = G3)) + geom_boxplot()
# Permutation Tests of the Differences in Final Grades between Subsets
Medu0 <- spf %>% filter(Medvar == "none") %>% select(G3, Medu)
Medu1 <- spf %>% filter(Medvar == "primary education") %>% select(G3, Medu)
Medu2 <- spf %>% filter(Medvar == "5th to 9th grade") %>% select(G3, Medu)
Medu3 <- spf %>% filter(Medvar == "secondary education") %>% select(G3, Medu)
Medu4 <- spf %>% filter(Medvar == "higher education") %>% select(G3, Medu)
Com1 <- merge(Medu0, Medu1, all = TRUE)
Com2 <- merge(Medu1, Medu2, all = TRUE)
```

```r
Com3 <- merge(Medu2, Medu3, all = TRUE)
Com4 <- merge(Medu3, Medu4, all = TRUE)
permTest(G3 ~ Medu, data = Com1)
permTest(G3 ~ Medu, data = Com2)
permTest(G3 ~ Medu, data = Com3)
permTest(G3 ~ Medu, data = Com4)

# Variable 2: Father's Education
papedu<- table(spf$Fedu)
papedu
prop.table(papedu)
# ANOVA of Father's Education and Corresponding Mean Final Grade
spf %>% group_by(Fedu) %>% summarise(mean(G3), sd(G3))
ggplot(spf, aes(sample = G3)) + geom_qq() + geom_qq_line() + facet_grid(. ~ Fedu)
fedu.aov <- aov(G3 ~ Fedu, data = spf)
summary(fedu.aov)
# Figure 2: boxplot of Final Grade Grouped by Father's Education
spf$Fedvar <- factor(spf$Fedu)
levels(spf$Fedvar)
levels(spf$Fedvar) <- c("none", "primary education", "5th to 9th grade",
                        "secondary education", "higher education")
ggplot(spf, aes(x = Fedvar, y = G3)) + geom_boxplot()
# Permutation Tests of the Differences in Final Grades between Subsets
Fedu0 <- spf %>% filter(Fedvar == "none") %>% select(G3, Fedu)
Fedu1 <- spf %>% filter(Fedvar == "primary education") %>% select(G3, Fedu)
Fedu2 <- spf %>% filter(Fedvar == "5th to 9th grade") %>% select(G3, Fedu)
Fedu3 <- spf %>% filter(Fedvar == "secondary education") %>% select(G3, Fedu)
Fedu4 <- spf %>% filter(Fedvar == "higher education") %>% select(G3, Fedu)
Com5 <- merge(Fedu0, Fedu1, all = TRUE)
Com6 <- merge(Fedu1, Fedu2, all = TRUE)
Com7 <- merge(Fedu2, Fedu3, all = TRUE)
Com8 <- merge(Fedu3, Fedu4, all = TRUE)
permTest(G3 ~ Fedu, data = Com5)
permTest(G3 ~ Fedu, data = Com6)
permTest(G3 ~ Fedu, data = Com7)
permTest(G3 ~ Fedu, data = Com8)

# Variable 3: Mother's job
mjob <- table(spf$Mjob)
mjob
prop.table(mjob)
# Figure 3: boxplot of Final Grade Grouped by Mother's job
ggplot(spf, aes(x = Mjob, y = G3)) + geom_boxplot()
# ANOVA of Mother's Job and Corresponding Mean Final Grade
spf %>% group_by(Mjob) %>% summarise(mean(G3), sd(G3))
ggplot(spf, aes(sample = G3)) + geom_qq() + geom_qq_line() + facet_grid(. ~ Mjob)
Mjob.aov <- aov(G3 ~ Mjob, data = spf)
summary(Mjob.aov)
# Multiple Regression of Mother's Job and Final Grade
edu2.lm <- modelsum(G3 ~ Mjob, data = spf)
summary(edu2.lm, title="Multiple Regression of Final Grade and Mother's Job")
# Permutation Tests of the Differences in Final Grades between Subsets
Com9 <- spf %>% filter(spf$Mjob == "at_home" | spf$Mjob == "health") %>% select(G3, Mjob)
```

```r
Com10 <- spf %>% filter(spf$Mjob == "at_home" | spf$Mjob == "other") %>% select(G3, Mjob)
Com11 <- spf %>% filter(spf$Mjob == "at_home" | spf$Mjob == "services") %>% select(G3, Mjob)
Com12 <- spf %>% filter(spf$Mjob == "at_home" | spf$Mjob == "teacher") %>% select(G3, Mjob)
permTest(G3 ~ Mjob, data = Com9)
permTest(G3 ~ Mjob, data = Com10)
permTest(G3 ~ Mjob, data = Com11)
permTest(G3 ~ Mjob, data = Com12)


# Variable 4: Father's Job
fjob <- table(spf$Fjob)
fjob
prop.table(fjob)
# Figure 4: boxplot of Final Grade Grouped by Father's job
ggplot(spf, aes(x = Fjob, y = G3)) + geom_boxplot()
# ANOVA of Father's Job and Corresponding Mean Final Grade
spf %>% group_by(Fjob) %>% summarise(mean(G3), sd(G3))
ggplot(spf, aes(sample = G3)) + geom_qq() + geom_qq_line() + facet_grid(. ~ Fjob)
Fjob.aov <- aov(G3 ~ Fjob, data = spf)
summary(Fjob.aov)
# Multiple Regression of Father's Job and Final Grade
edu3.lm <- modelsum(G3 ~ Fjob, data = spf)
summary(edu3.lm, title="Multiple Regression of Final Grade and Father's Job")
# Permutation Tests of the Differences in Final Grades between Subsets
Com13 <- spf %>% filter(spf$Fjob == "at_home" | spf$Fjob == "health") %>% select(G3, Fjob)
Com14 <- spf %>% filter(spf$Fjob == "at_home" | spf$Fjob == "other") %>% select(G3, Fjob)
Com15 <- spf %>% filter(spf$Fjob == "at_home" | spf$Fjob == "services") %>% select(G3, Fjob)
Com16 <- spf %>% filter(spf$Fjob == "at_home" | spf$Fjob == "teacher") %>% select(G3, Fjob)
permTest(G3 ~ Fjob, data = Com13)
permTest(G3 ~ Fjob, data = Com14)
permTest(G3 ~ Fjob, data = Com15)
permTest(G3 ~ Fjob, data = Com16)

# Variable 5: Family Size
familysize<- table(spf$familysize)
familysize
prop.table(familysize)
# Permutation Test of the Difference in Mean Final Grade
ggplot(spf, aes(x = famsize, y = G3)) + geom_boxplot()
permTest(G3 ~ famsize, data = spf)

# Variable 6: Parent's Cohabitation Status
parentlive<- table(spf$Pstatus)
parentlive
prop.table(parentlive)
ggplot(spf, aes(x = Pstatus, y = G3)) + geom_boxplot()
# Permutation Test of the Difference in Mean Final Grade
permTest(G3 ~ Pstatus, data = spf)

# Variable 7: Guardian
guard <- table(spf$guardian)
guard
prop.table(guard)
```

```r
# Boxplot of Final Grade Grouped by Guardian
ggplot(spf, aes(x = guardian, y = G3)) + geom_boxplot()
# Multiple Regression of Guardian and Mean Final Grade
summary(spf$guardian)
guardian1.lm <-lm(G3 ~ guardian, data=spf)
summary(guardian1.lm)

# Variable 8: Family Educational Support
fedusup <- table(spf$famsup)
fedusup
prop.table(fedusup)
# Permutation Test of the Difference in Mean Final Grade
ggplot(spf, aes(x = famsup, y = G3)) + geom_boxplot()
permTest(G3 ~ famsup, data = spf)

# Variable 9: Quality of Family Relationships
famrela <- table(spf$famrel)
famrela
prop.table(famrela)
# ANOVA for Quality of Family Relationships and Mean Final Grade
spf %>% group_by(famrel) %>% summarise(mean(G3), sd(G3))
ggplot(spf, aes(sample = G3)) + geom_qq() + geom_qq_line() + facet_grid(. ~ famrel)
famrel.aov <- aov(G3 ~ famrel, data = spf)
summary(famrel.aov)
# Linear Regression of Quality of Family Relationships and Mean Final Grade
famrel1.lm <-lm(G3 ~ famrel, data=spf)
summary(famrel1.lm)

# Multiple Regression of Final Grade Vs Mother's and Father's Education
edu.lm <- lm(G3 ~ Medu+Fedu, data = spf)
summary(edu.lm)

# Multiple Regression of Final Grade Vs All the Other Variables
library(arsenal)
overall <- modelsum(G3 ~ Medu+Fedu+famsize+Mjob+Fjob+guardian+famsup+famrel+Pstatus, data = spf)
summary(overall, title= 'Multivariable Regression of Final Grade and All Other Variables')
```