Post1: How Does GGPLOT2 Help Interpret the Results of UGBA 102A (Introduction to Financial Accounting) Midterm 1

Katie Li

10/29/2017

```
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Introduction

- The post combines my favorite topic from the course, ggplot2, with a class that I am currently taking, UGBA 102A (Introduction to Financial Accounting). As a business major, I often don't have the opportunity to conduct statistical analysis, let alone on the midterm results of a business class. In UGBA 102A, we recently received our midterm scores. While I know my score, the basic statistics of the class (mean, median, max, min, etc.), I also want to visualize the results in greater detail. GGPLOT2 is perfect for this purpose, given its visually appealing and user-frienly nature. In Stats 133, we learned the basic functions of ggplot2 and applied it to one or two psuedo scenarios (in that the situations are made up). Here, I will be applying ggplot2 to a real-life scenario, exploring and visualizing different students' performance (rank & score) in UGBA 102A based on variables such as their gender and the amount of time that they spent studying for the test. Note: While the scores and ranking are based on real data, the study time is based on my assumption.

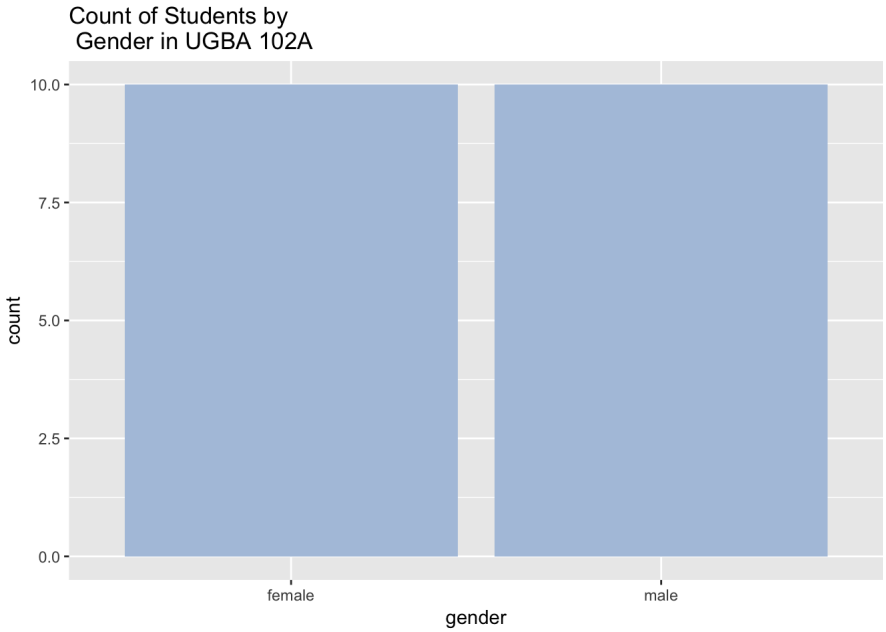
1.) Understanding the Basic Statistics of the Data Frame: Midterm1

```
rank <- c(seq(from = 1, to = 20, by = 1))
score <- c(100, 98, 96, 90, 88, 86, 83, 80, 76, 75, 74, 70, 69, 68, 66, 64, 62, 60, 58, 55)
gender <- c(rep('male', 10), rep('female', 10))
studytime <- c(460, 468, 455, 420, 402, 414, 396, 384, 362, 351, 365, 370, 320, 300, 290, 267, 250, 247, 230, 220)
midterm1 <- data.frame(rank, score, gender, studytime)
midterm1
```

```
##    rank score gender studytime
## 1     1   100   male       460
## 2     2    98   male       468
## 3     3    96   male       455
## 4     4    90   male       420
## 5     5    88   male       402
## 6     6    86   male       414
## 7     7    83   male       396
## 8     8    80   male       384
## 9     9    76   male       362
## 10   10    75   male       351
## 11   11    74 female       365
## 12   12    70 female       370
## 13   13    69 female       320
## 14   14    68 female       300
## 15   15    66 female       290
## 16   16    64 female       267
## 17   17    62 female       250
## 18   18    60 female       247
## 19   19    58 female       230
## 20   20    55 female       220
```

Count of Students

```
gender_count <- ggplot(midterm1, aes(x = gender)) + geom_bar(fill='lightsteelblue') + ggtitle("Count of Students b
y \n Gender in UGBA 102A")
gender_count
```

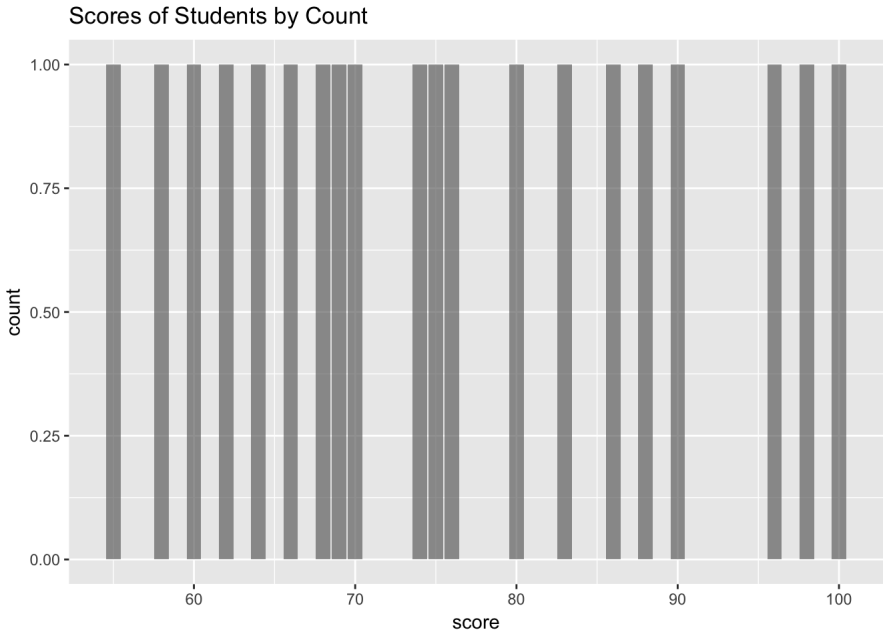


Explanation of Count of

Students by Gender in UGBA 102A * In understanding students' performance on the midterm, it is important to first understand the basic statistics. Here, I created a visualization on the number of students by gender. By using `geom_bar()`, which displays bar charts, it is clear that the number of male and female students equates (at 10 each). In addition to the basic ggplot2 functions we learned in class, I also learned a new trick on a website called Environmental Comuputing. By using `\n` between titles, a long title is able to separate into two lines, exemplified by "Count of Students by" and "Gender in UGBA 102A" listed above.

Counting Scores of Students

```
count_scores <- ggplot(midterm1) + aes(x=score) + geom_bar(alpha = 0.6) + ggtitle("Scores of Students by Count")
count_scores
```



Explanation of Scores of

Students by Count * The above graph might seem overly simplistic at first: it shows the number of students who received a particular score on UGBA 102A midterm1. Since there are no multiple students who share the same score, each vertical bar has the same height. Morever, since score is a discrete variable, there is an interval between each score on the x-axis. This simple visualization, however, could be manipulated if we utilized convert scores from numeric into factor. An application of such is shown in part 2 of the research, under a graph called "Score Range by Gender".

- By utilizing the original dataframe "midterm1", I created an additional variable using dplyr's mutate function called "scores_factor", which converts all scores from numeric into factors. Yet, beyond the basic factor conversions that we learned in class, I did additional research on `cut()` in a video called Plotting in R Tutorial: Gorgeous Graphs with GGPLOT2, which divides the range of the x variable (in our case, the students' scores) into intervals. I found `cut()` to be especially helpful for the purpose of this research because intervals are an efficient way of capturing students' performance in the class. For this problem, the first interval that I chose is 0-50, as no students scored below 50 points on the test. After 50, I selected an interval of 10 for the remaining scores, as there are multiple students whose scores lie in the range.

Turning Scores into Factors

```
scores_factor <- cut(score, breaks= c(0, 50, 60, 70, 80, 90, 100))
added_scores_factor <- mutate(midterm1, scores_factor)
added_scores_factor
```

```
##    rank score gender studytime scores_factor
## 1     1   100   male       460      (90,100]
## 2     2    98   male       468      (90,100]
## 3     3    96   male       455      (90,100]
## 4     4    90   male       420       (80,90]
## 5     5    88   male       402       (80,90]
## 6     6    86   male       414       (80,90]
## 7     7    83   male       396       (80,90]
## 8     8    80   male       384       (70,80]
## 9     9    76   male       362       (70,80]
## 10   10    75   male       351       (70,80]
## 11   11    74 female       365       (70,80]
## 12   12    70 female       370       (60,70]
## 13   13    69 female       320       (60,70]
## 14   14    68 female       300       (60,70]
## 15   15    66 female       290       (60,70]
## 16   16    64 female       267       (60,70]
## 17   17    62 female       250       (60,70]
## 18   18    60 female       247       (50,60]
## 19   19    58 female       230       (50,60]
## 20   20    55 female       220       (50,60]
```

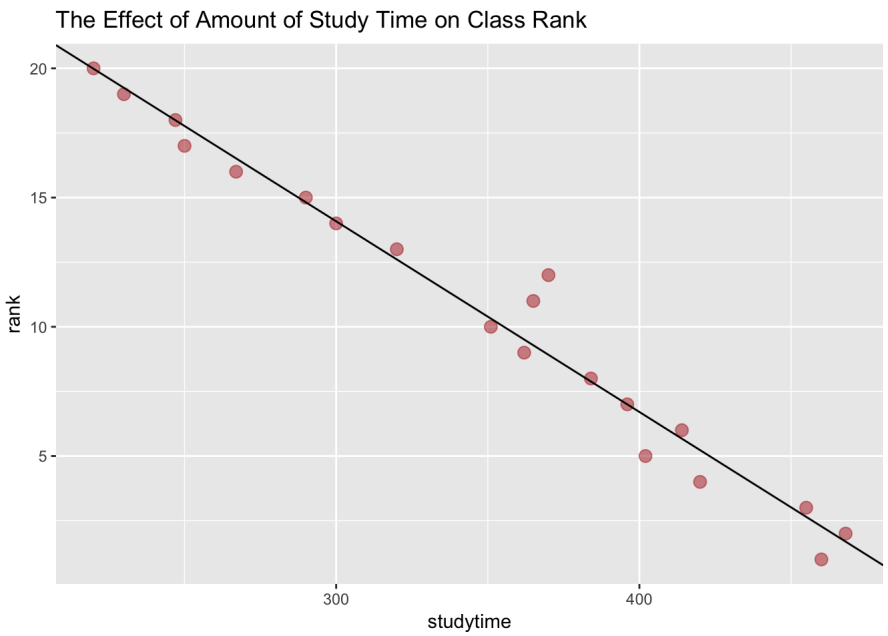
2.) Understanding the Variables that Led to Higher Performance

```
coef(lm(rank ~ studytime), data = midterm1)
```

```
## (Intercept)   studytime
## 36.21339056 -0.07377246
```

The Relationship between Study Time and Class Rank

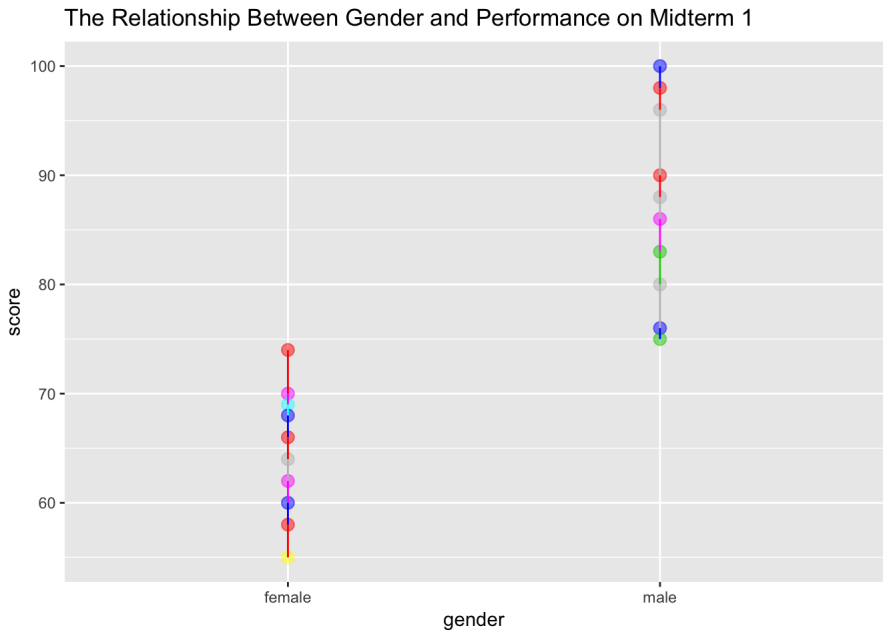
```
studytime_rank <- ggplot(midterm1, aes(studytime, rank)) + geom_point(color="firebrick", size = 3, alpha = 0.5) +
ggtitle("The Effect of Amount of Study Time on Class Rank") + geom_abline(intercept = 36.21339056, slope = -0.0737
7246)
studytime_rank
```



Explanation on The Effect of Amount of Study Time on Class Rank * While we learned to plot scatter plots in ggplot2, we did not learn explicitly on how to add a best fit line to the graph. After doing some research on Tidy Verse, I learned that it is necessary to deduce the intercept and slope using a separate line of code such as `coef(lm(rank ~ studytime), data = midterm1)`, if we wish to find the relationship between a student's study time and their ranking on the test. In addition, something that we didn't have too much opportunity to explore in class is ggplot2's versatile color selections. In the GGPLOT2 Cheatsheet, I found a color called "firebrick", which is used in the graph above. * The way that ggplot2 visualizes data frames fascinates me. When the midterm result was initially released, each student in my class received a ranking based on their test score. While it shows their relative standing in the class, it does not immediately show the relationship bewteen their study time and ranking. The scatter plot above demonstrates how on average, when students study more, their ranking is higher. There are some outliers, as shown through those points above the best-fit line, representing the students who perform better than their peers who studied around the same amount time. However, based on the best-fine line and the fact that most points fit closely to the best-fit line, we can conclude that when students study more, their ranking does increase in the class.

The Relationship Between Gender and Performance on Midterm 1

```
gender_score <- ggplot(midterm1, aes(gender, score)) + geom_point(color=score, size = 3, alpha=0.5) + ggtitle("The
Relationship Between Gender and Performance on Midterm 1") + geom_path(color=score)
gender_score
```

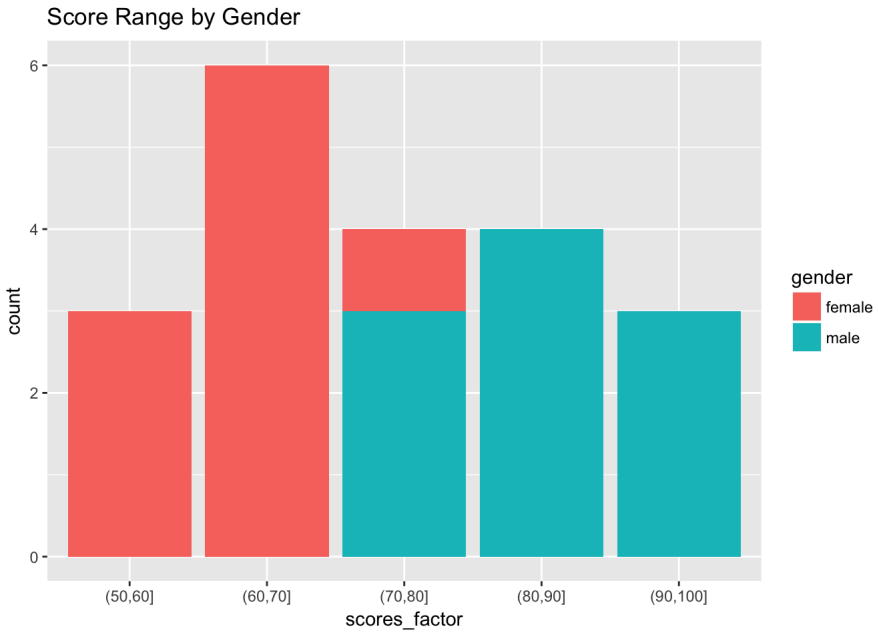


Explanation of the Graph

Above * In addition to discovering the relationship between the amount of time students spent studying and their respective ranking in the class, I also wanted to see the correlation between students' gender and their scores, as displayed by the ggplot above. Using `geom_step()` geom_step(), which shows exactly when changes occur (shown through the line that connects all the points), I am able to tell that changes all occured simultaneously, and that male students scored higher on average than female students. The two respective lines are completely vertical since we are dealing with a categorical variable (gender). The fact that all points for male students are higher than those for female students shows that all male students scored higher than the female students in the class. The observations described above could be deduced by reading the data frame, but it is much clearer and eaiser when the results are visualized on ggplot2.

Count of Male / Female Students Who Scored in Specific Range

```
ggplot(added_scores_factor) + aes(x = scores_factor, fill = gender) + geom_bar() + ggtitle("Score Range by Gender"
)
```



Explanation of Score Range by Gender

- The graph above captures two important factors. First, it shows the number of students whose test scores fall between sepcific score intervals. More importantly, it also displays the gender of the students whose scores fall under these ranges. In the previous graph "The Relationship Between Gender and Performance on Midterm 1", we can distinguish from the point patterns that all male students performed better than their female counterparts on the test. However, what if we want to dive deeper in the available data? How many female students fall into the bottom test range (50-60 points) and how many of them fall into the same score range as the male students? In the bar chart above, we are able to answer these questions. Here, there are 3 feamle students who received a score between 50-60, and there are 4 students who scored between 70 - 80 points, the only score range where there is a combination of both male and female students.The idea and applications are visualized on plot.ly, a site that enhanced my understanding of bar chart visualizations.

Conclusion

- The research idea for this project came from a simple phenomenon that I encounter in daily life: test scores. Something about UGBA 102A: Introduction to Financial Accounting that stood out to me is that the professor ranks students by their scores, and students' test scores vary considerably from one another. Using these variables, as well as the fact that the gender of the students are evenly divided, I created a data frame with some real statistics (ranking / scores) along with assumptions (the amount of time each student spent studying / gender of students for each test score). The beauty of ggplot 2, in the words of its creator Wickham - All Hail GGPLOT2, is that it has the power to

empower curious individuals who understake quirky projects. It is designed for people who would have struggled without the toolkit otherwise. As a business student, I often don't have the skill nor the capacity to explore the relationship between students' performance on midterms with variables such as their study time and gender. GGPLOT 2 helped me to clearly visualize these results and understand the attributes that are related to good performance on the midterm. The results derived from this research, such as the in the graph "Score Range by Gender", also have future implications. For example,since there are both male and female students scoring between 70 and 80 points, researchers could conduct additional research on these students and observe similarities among them. The research results could be applied to other score ranges, which could potentially bridge the score gap bewteen genders.

Future Steps

- The research that I have covered in this post is rather rudimentary. In the future, I am intersted to observe data visualizations on bigger data sets. The image below, which captures statistics on the 50 states in the country are quite interesting. This could be useful for election, healthcare implementation rates, and other national statistics. I would love to further explore if there is opportunity in the future!