Exams Project

Julton Wagner

2022-11-21

Introduction

The Exams data set is comprised of a mix of demo graphical information about participants who took assessments in various topics (Mathematics, Reading and Writing) and what factors could have possibly influenced those scores. In total, the data set has eight (8) variables, five (5) of which are qualitative variables and the remaining three (3) are quantitative variables. The five qualitative variables that describes features of the participants which include gender, ethnicity, parental level of education, subsidized/standard lunch, and whether or not they sat the test preparation course. For the quantitative variables, these are all scores earned by each participant for Mathematics, Reading and Writing. The output below depicts this more clearly.

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "male" "female" "male" "male" ...
##  $ race.ethnicity             : chr  "group A" "group D" "group E" "group
B" ...
##  $ parental.level.of.education: chr  "high school" "some high school"
"some college" "high school" ...
##  $ lunch                      : chr  "standard" "free/reduced"
"free/reduced" "standard" ...
##  $ test.preparation.course    : chr  "completed" "none" "none" "none" ...
##  $ math.score                 : int  67 40 59 77 78 63 62 93 63 47 ...
##  $ reading.score              : int  67 59 60 78 73 77 59 88 56 42 ...
##  $ writing.score              : int  63 55 50 68 68 76 63 84 65 45 ...
```

We see that the qualitative variables are picked up as characters in R while the quantitative variables are recognized as integer variables. In determining the goal of this project, one had to think outside of the box, because they were several research questions that could possibly be asked. So in determining possible questions, I engaged in further inspection and analysis and let the data guide my curiosity. Key steps performed during the project include, but are limited to

- Data retrieval & inspection
- Data cleaning
- Data manipulation
- Data partitioning
- Graphing & Sketching
- Functions

Analysis

This project began with the reading of the exams.csv file into R using the read.csv function. From there that output is saved into a variable called performance. A quick peek into performance yields the following

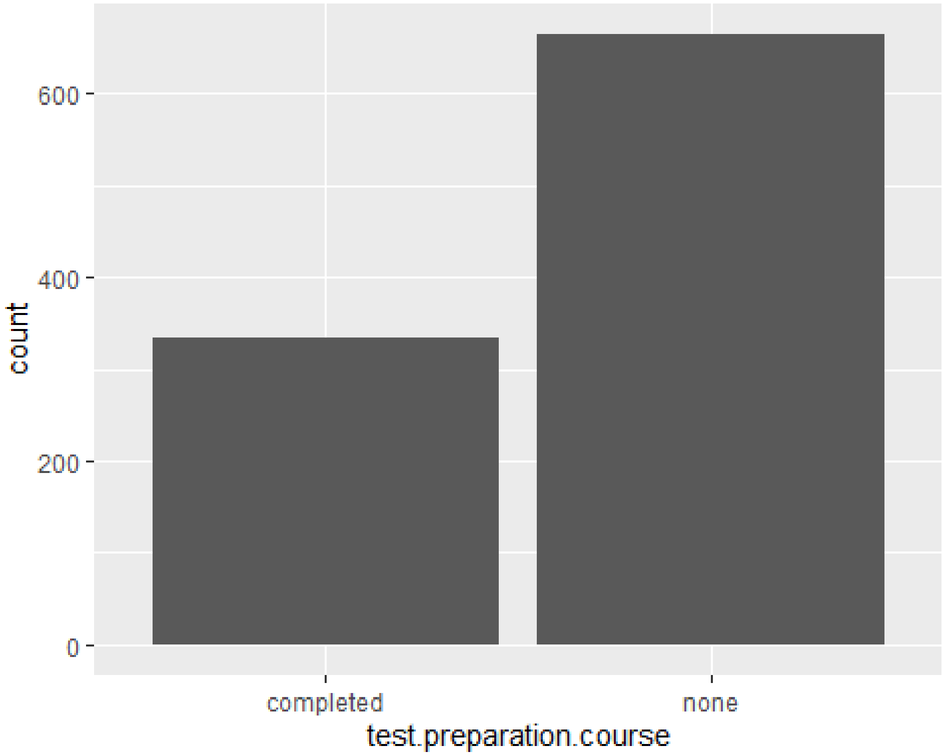
```
##   gender race.ethnicity parental.level.of.education        lunch
## 1   male        group A                 high school     standard
## 2 female        group D            some high school free/reduced
## 3   male        group E                some college free/reduced
## 4   male        group B                 high school     standard
## 5   male        group E          associate's degree     standard
## 6 female        group D                 high school     standard
##   test.preparation.course math.score reading.score writing.score
## 1               completed         67            67            63
## 2                    none         40            59            55
## 3                    none         59            60            50
## 4                    none         77            78            68
## 5               completed         78            73            68
## 6                    none         63            77            76
```

We note that we are provided with a mix of both quantitative and qualitative variables in this data set. This distinction provides us with varied approaches to consider when deciding on models as well shape possible narratives to investigate. In recycling some previously used code, an interesting observation can be made

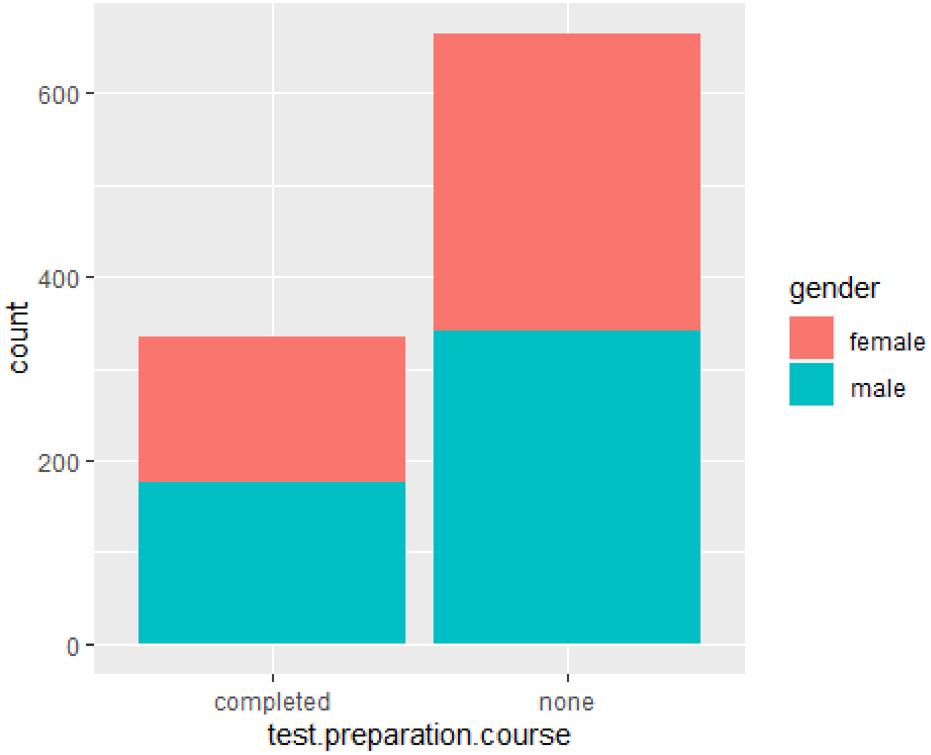
```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "male" "female" "male" "male" ...
##  $ race.ethnicity             : chr  "group A" "group D" "group E" "group
B" ...
##  $ parental.level.of.education: chr  "high school" "some high school"
"some college" "high school" ...
##  $ lunch                      : chr  "standard" "free/reduced"
"free/reduced" "standard" ...
##  $ test.preparation.course    : chr  "completed" "none" "none" "none" ...
##  $ math.score                 : int  67 40 59 77 78 63 62 93 63 47 ...
##  $ reading.score              : int  67 59 60 78 73 77 59 88 56 42 ...
##  $ writing.score              : int  63 55 50 68 68 76 63 84 65 45 ...
```

For all of the qualitative variables, they all returned as characters. It is always best practice to convert those characters to factors to facilitate visualizations using ggplot2 alongside models serving as good predictors with qualitative data (logistic, knn, svm, trees etc.). Having said that, all the character variables were converted to factors using as.factor and before a decision was made to begin creating visualizations, averages of all the integer variables alongside an average of those averages were taken and stored in their respective variables for later analysis.

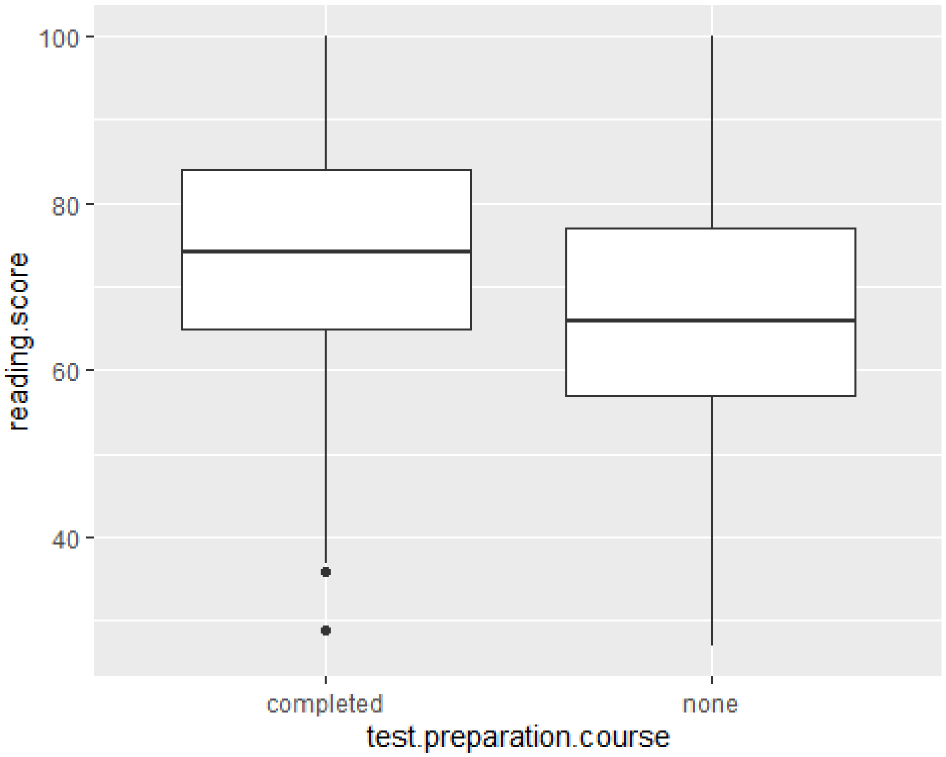
A key step in data analysis is the reproduction of visualizations as it lets the researcher discern various patterns in the data that could warrant further investigation. An example showing the breakdown of participants who took the test preparation course can be seen below.



From this example, we see that about 60% of the participants decided not to do the test preparation course. This result will be kept in mind as we continue to do exploratory data analysis. Following that thought process, we check to see if there is a massive gender difference in those who sat the test preparation course and those who did not. The chart below depicts this.



Based on the illustration, we do not see any massive difference in the gender breakdown of participants who did/did not the test preparation course. In determining a research question and goal for this project, I had to use conventional wisdom and test theories that should hold when presented with data. One such wisdom was that participants who did the test preparation should perform better than those who did not. Let us see if this holds visually.



The boxplot above confirms the conventional wisdom that those who did the prep course performed better than those who did not for reading. When this was repeated for the other two courses, it affirmed the outcome stated previously. From the visual representations done previously, a lingering question remains, does the test preparation course lead to better marks? From the evidence above, it certainly seems so, but we require more evidence. Given that the relationship between test prep participants and improved scores was a feature in all three subjects, one can use those grades to guess who did the course or not. Specifically, can we use the marks obtained by each participant to build prediction models that can accurately determine whether or not they did the course? To answer this question, we can try three different approaches which include

- Logistic Regression
- Decision Trees
- Random Forest

But before starting any of these, we have to set our random seed, so the results can be reproducible by others and also set up our train/test split. A snippet of the code can be seen below.

```
library(caret)

## Warning: package 'caret' was built under R version 4.2.1

## Loading required package: lattice
```

```r
set.seed(99)
series <- createDataPartition(y = performance$test.preparation.course, times
= 1, p = 0.3, list = FALSE)
train <- performance[-series,]
test <- performance[series,]
```

The code states that we want to create a data partition whereby it is partitioned once and follows a 70/30 Where it states “p=0.3”. This means that 70% of the observations will be used to train the models which will try and accurately predict the remaining 30% of outcomes in the test set. This split was chosen because it provides sufficient observations to train the model while avoiding over/under training. It also gives us adequate observations in the test set to predict against. We proceed by creating the logistic model by regressing the test preparation course on the scores obtained for math, reading and writing using the training data. We then get a summary of the model to check for variable significance and R-Squared. From this, we use the predict function to see how well the model performs against test data. The predicted data is then put through a decision rule which states that provided a probability that exceeds 0.5, it is likely that participant completed the test prep course. If the probability is equal to or below 0.5, we assume that candidate did not do the test prep. This process occurs for all predicted values. Lastly, to determine accuracy, we check to see what forecasted values were different than actual data and average those results. This provides us with our mean square error (mse), a measure of dispersion.

Finally, the process is repeated in building the decision tree and random forest model and these models are used to check for the mse to see which of the three is the most accurate. Those results can be seen in the next section.

Results

The error rate obtained by the logistic model can be seen below

```
## [1] 0.6976744
```

The miss rate of 0.698 or 69.8% is quite high and rather unexpected given that logistic models tend to work well with qualitative data. Let us see if using decision trees leads to any improvement in the mse.

```
## [1] 0.345515
```

Note that by using decision trees, the error rate more than halved to 0.346 or 34.6%. This is a notable improvement in model performance, but we can go further through random forest. The results using random forests can be seen below.

```
## [1] 0.3222591
```

Using random forests provides an error rate of 0.322 or 32.2%, which is the most accurate model we have seen out of the three. For model selection, the random forest model should be chosen as that has the lowest mse of the three.

```
##   misclasserror.logit misclasserror.tree misclasserror.rf
## 1           0.6976744           0.345515        0.3222591
```

Conclusion

To summarize, the exams project presented data on student performance in three subjects (mathematics, reading and writing) and possible categorical variables that may explain the scores obtained. From the graph sketching, the data illustrated that gender played no role in explaining the marks. However, the box plots suggested that sitting of the test preparation course could explained the varies performance across participants. We framed our research question by stating if the grades obtained by the participants could determine who sat the test prep course and who did not. To empirically answer this, we employed three models (logistic, decision trees and random forest) to see which would give the best results. It turns out that the random forest model was the best performer of the bunch.

Some limitations to this study include

- Small sample size
- Only three models considered
- Only using mse (perhaps include rmse as well)
- Possible noise in some of the estimates

Future work on this topic will facilitate the inclusion of these limitations and more diagnostic testing to ensure estimates are robust.