EOC Data Mining Final Research Recommendation Report

Executive Summary

The research project was conducted on 2014-15 Department of Elementary and Secondary Education End of Course Standardized Test Scores. The scores were extracted from the Missouri Comprehensive Data System's website ("MCDS", n.d.). The sample space consisted of 41 local St. Louis area school districts. The EOC's taken into consideration was Algebra 1 and Biology. Both of the EOC's are required to be completed by all accountable students prior to graduation. The individual 6 populations (Total, Asian/Pacific Islander, Black, Hispanic, Multiracial and White) were filtered, sorted by descending order per below-basic test scores and transposed to their own spreadsheet. The spreadsheets were partitioned into five columns: name of data set, below-basic percentage, basic percentage, proficient percentage and advanced percentage. Tables were constructed in R Studio composed of the individual data sets using the read.table() function. The tables were manipulated to stack the four different classifications of test scores on top of each other and then indexed and labeled with their corresponding school district. The reshape2 package was used to complete this process (Wickham, 2015). After the data had been formatted, the ggplot2 package was used to create aesthetic pleasing stacked bar graphs ("ggplot2", n.d.). The topo.colors() function was implemented to provide the color for the continuous stacked graphs to differentiate the four classifications of scores. The graphs were sorted in descending order per the below-basic scores. The Algebra 1 and Biology were compared within their own total and racial classifications. The topology of the graphs compared was similar in area for proficient and advanced test scores for all 6 populations. The difference in the topology was evident at the below-basic test scores. The graphs suggest inequalities within certain school districts and or certain racial classifications. Further research will be conducted on the feeder schools within the 41 school districts with the same method to search for any correlations or patterns.

Problem Description

The focal point of the research project was End of Course Standardized Test Scores accumulated during the 2014-15 school year by the Department of Elementary and Secondary Education. The EOC's taken into consideration were Algebra 1 and Biology. The sample spaces contained four different classifications of test scores from 41 St. Louis area school districts. The sample sample was disaggregated by 5 different racial classifications too. The main objective of the research was to look for any trends and or inequalities within the 41 school districts. Graphical analysis was used to compare the total population and the five different racial classifications.

Analysis Technique

The test scores were retrieved from the Missouri Comprehensive Data System ("MCDS", n.d.). The scores were compiled in one master Excel spreadsheet. The master spreadsheet was filtered down to the 41 school districts Algebra 1 and Biology test scores. The spreadsheet contained disaggregated scores per racial classification for each school district. Then, each total population and each racial classification was copied and pasted to new spreadsheets. The spreadsheets were partitioned to into five columns (population, below-basic percentage, basic percentage, proficient percentage and advanced percentage), sorted by descending below-basic scores and then transposed. The data was transferred into R Studio by using the read.table() function. The reshape2 package was used to allow the data to be manipulated (Wickham, 2015). The melt() function was used to convert the table into a stacked data frame. The individual school

district's test scores were stacked on top of each other. The rep() function was used to assign the unique values 1-4 to each row of the four classifications of the stacked scores. The cbind() function was then implemented to attached the row values to the corresponding data frame constructed by the melt() function. The ggplot2 package enabled the reshaped data to be graphed in a multilayered colored stacked graph ("ggplot2", n.d.). The individual graphs were constructed with the following code:

```
ggplot(df, aes(x=variable, y=value, fill=row)) +
scale_fill_gradientn(colours = topo.colors(10)) +
geom_bar( stat="identity") +
xlab("\nSample") +
ylab("Percentage\n") +
theme_bw()+
theme(axis.text.x=element_text(angle=20, size=12))+
guides(fill=FALSE)
```

The ase() function represented the x and y variables and declared the fill of the variables. The scale_fill_gradientn() function implemented the topology color palette. The geom_bar() function constructed the data into continuous stacked-bar graphs. The x and y axis was labeled with the xlab() and ylab() functions. The theme() function outlined the graphs and set the angle of the of the x variable's label ("ggplot2", n.d.). The ggplot2 code was executed to produce aesthetic pleasing and differentiable stacked bar graphs. The x variables were sorted by descending below-basic corresponding y values.

Assumptions

I assumed the Department of Elementary and Secondary Education's data was accurate.

Results

The Algebra 1 graphs were compared to the Biology graphs. There were 6 population comparisons: Total, Asian/Pacific Islander, Black, Hispanic, Multiracial and White. The conclusion was the topology of the graphs was similar for the total and the five different classifications of race. The proficient and advanced y-values covered similar areas in each subject when compared to each other. There was a difference in the topology at the below-basic y-values in all six comparisons. The graphs suggest inequalities within certain school districts and certain racial classifications. The data analysis was conducted on high school students. The next step in the research process is to analyze the feeder schools of the 41 school districts with the same method.

Issues

This was a great research project. It was a rigorous, but rewarding process. It took some time to understand the structure of ggplot2. The coding was an issue at first, but I am comfortable with the coding of the graphs now. The ggplot2 package is a very powerful tool for statistical analysis.

References

Wickham, H (2015, February, 20). *Flexibly Reshape Data: A Reboot of the Reshape Package.* retrieved November 2015, from CRAN R Project Web Site: https://cran.r-project.org/web/packages/reshape2/reshape2.pdf

. (n.d) retrieved November 2015, from ggplot2 Web Site: http://ggplot2.org/

. (n.d) retrieved November 2015, from Missouri Comprehensive Data System Web Site: http://mcds.dese.mo.gov/quickfacts/Pages/State-Assessment.aspx/