Data Science in Education: Mini-Project #1

In this assignment, you will use linear regression modeling to analyze selected variables in the 2022 South Carolina school report card data, focusing only on traditional high schools. You should use R markdown to create a write-up of your analysis and submit your knitted HTML file to Canvas when finished. Use the headings below in your report. Include all R code used to generate your output and appropriate commentary throughout, but hide any messages and warnings and adjust font sizes in visual displays as necessary to create a polished report.

Introduction

10 pts

- Open the **ReportCardData_AdditionalInfo2022.xlsx** file posted to this assignment in Canvas.
- Go to the sheet labeled KEY. Browse through the variables in the dataset.
- Select **one numeric response variable** for your analysis. Some possibilities include: dropout rate, average ACT score, average SAT score, AP pass rate, percent of seniors completing college applications, percent of seniors eligible for LIFE scholarships, chronic absenteeism rate, etc.
- Select **at least three potential explanatory variables** for your analysis.
- In the Introduction section of your report, identify your selected variables by stating the statistical research question your regression analysis will be investigating. Then write a few sentences to provide some insight into why the variables you selected are of interest to you. No references are needed, just some initial thoughts on the potential variable relationships based upon your expertise in the field of education.

Load Packages

5 pts

- In this section, you only need to provide the R code for installing and/or loading the packages you used to complete this assignment.
- At a minimum, this will include `readxl`, `tidyverse`, `dataedu`, `GGally`, and `sjPlot`.

Import Data

5 pts

- Use the `read_excel` function of the `readxl` package to import each sheet of the data file containing one of your selected variables.
- Note that with the `sheet` argument of the `read_excel` function, you can specify the Excel sheet to import either by position or by name.
- You will need to use the `na` argument of the `read_excel` function to specify the codes used for missing values. Note that these vary from sheet to sheet of the data file. Some sheets use more than one code and in this case be sure to use the combine function, `c()`, in the `na` argument.

Process Data

10 pts

- Use a `join` function to merge the different sheets of the data file you imported by both SCHOOLID and SCHOOLTYPECD.
- Use the filter function to reduce the observations to only high schools (H) based on the SCHOOLTYPECD variable.
- Use the `filter` function in combination with the `str_detect` function with the `negate = TRUE` option to reduce the observations to only traditional high schools. This involves removing any school name that includes the strings "Academy", "Charter", "College", "Magnet", "School Of", or "School For" as well as any district name that includes the strings "Charter", "Unified", or "Governor's".
- Use the `select` function to reduce your data frame down to include only the variables for district name, school name, school ID, and the variables you selected for your analysis.
- If any of your selected variables have dollar signs or commas in what should be a column of numeric data, R will think these are character variables. If you try to coerce them to numeric variables without removing these symbols, R will convert them to missing data. You will first need to remove the dollar signs using the `mutate` function in combination with the `str_sub` function with the `start =` option. To remove commas, use the `gsub` function, with `pattern = ","` and `replacement = ""`. At this point, you should be able to coerce the column into numeric data using the `as.numeric` function.

Analysis

Your data analysis should include three components: a univariate data analysis, a bivariate data analysis, and a linear regression analysis.

Univariate Data Analysis

20 pts

- Plot each of your variables using a style of your choosing in `ggplot()`. Optionally, you can try fitting multiple plots on the same page using the `grid.arrange` function of the `gridExtra` package.
- Calculate summary statistics for each of your selected variables. For numeric variables, calculate the mean, standard deviation, minimum, maximum, median, and IQR. For categorical variables, provide a frequency table.
- Describe the distribution of each variable in a sentence or two.

Bivariate Data Analysis

20 pts

- Use `ggpairs` to provide a scatterplot and correlation matrix that includes all numeric variables you selected. Comment on which of the numeric explanatory variables has the strongest relationship with the response variable. Also comment on any strong relationships between the numeric explanatory variables, as this could be a sign of multicollinearity in a regression model.
- If you have any categorical explanatory variables, use one of the visualization options provided in Lab 5 to display the relationship between the numeric response variable and the categorical explanatory variable. Then calculate the mean, standard deviation, minimum, maximum, median, and IQR of the response variable separately for each category of the categorical explanatory variable. Comment on the strength of the association.

Linear Regression Analysis

20 pts

- Use `lm()` to fit the multiple regression model that includes all of your selected explanatory variables. Use `tab_model` to display a summary of the fitted model.
- Follow a backward elimination model selection procedure by removing, one at a time, the least significant variable from the model until all remaining variables are statistically significant at the 5% level (or there is only one variable left in the model).
- Use `tab_model` to display a summary of the fitted model in each step of the backward elimination procedure and include commentary at each step explaining your rationale for the removal of any variables from the model.
- For the final model, create both a residual vs. fitted plot and a normal plot of the residuals. Comment on what these plots reveal about the model assumptions.
- Provide interpretations of each coefficient in your fitted model. Also, if multicollinearity is potentially present based on your bivariate analysis, briefly comment on how this might be impacting your results.

Conclusion

10 pts

In the conclusion, summarize your methods and findings in a few sentences. Discuss whether the relationships you expected to see were present and also discuss any surprising relationships you found.