Data 608 Final project proposal.

QUESTION TO ANSWER:

What is the future impact for children who play sports vs children that do not play sports.

WHY:

The societal impact of income is well known. Can sports shape a childs life who otherwise would not be exposed to the positive influences and challenges that sports can bring.

APPROACH:

I have not found any datasets that will give me sports activity per child. Short if that I will use a dataset that includes Schools that provide sports programs and those that do not. I will then compare the student grades, scores and percentage that attend college.

ARTICLES:

<https://www.theatlantic.com/education/archive/2017/09/whats-lost-when-only-rich-kids-play-sports/541317/>

https://www.freep.com/story/life/2016/08/01/child-dreams-financial-olympics/87913520/

<https://www.activekids.com/football/articles/youth-sports-participation-by-the-numbers>

<http://harvardsportsanalysis.org/2016/12/which-sports-league-has-the-most-parity/>

Data Sets:

<https://ope.ed.gov/athletics/#/>

<https://libguides.geneseo.edu/c.php?g=67454&p=434785#s-lg-box-1300425>

New Datasets:

<https://www.statista.com/statistics/197272/us-high-schools-with-athletic-programs-for-girls-2010/>

<https://researchguides.library.tufts.edu/c.php?g=249153&p=2519648>

<https://eddataexpress.ed.gov/>

<https://catalog.data.gov/dataset?tags=high-school&page=3>

Compare regular school stats to Maxprep lists of programs. Look at other school support programs.

<https://www.maxpreps.com/list/rankings_by_sport.aspx>

<https://data.cityofnewyork.us/browse/select_dataset?Dataset-Information_Agency=Department+of+Education+%28DOE%29&nofederate=true&suppressed_facets%5B%5D=domain&utf8=%E2%9C%93>

<https://data.cityofnewyork.us/Education/DOE-High-School-Directory-2013-2014/u553-m549/data>

**Presentation Text**

INTRO

Hi My name is Anthony Pagan and my project is on The impact of School Programs on Children.

The objective of this project is to determine if school programs have an impact on a child education and future. This small objective has a impact not only on a childs life but society itself because an educated child becomes an educated adults that contribute our society.

My approach was to use school data that included the success of the schools and their extra curricula programs, and determine if it contributed to a successful student, but how do I measure this. My assumption was if a school has a high graduation and college enrollment rate, then it is putting a high rate of children in position to succeed. Next I needed to prove that the additional programs made a difference so I decided to compare schools with programs vs school without programs to see if there was a difference in graduation and college enrollment rates.

The scope of all schools in the US was too large for this project so I narrowed the scope down to NYC schools and narrowed the years down to 2011-2012. The data was retrieved from

<https://data.cityofnewyork.us/Education/DOE-High-School-Directory-2013-2014/u553-m549/data>.

The first part of my presentation will be to go through a markup from Rpub that I created to analyze , model and visualize the data. The second part will be the shiny app that I created based on this analysis.

DATA EXPLORATION

I begin by importing the data and setting the missing values to NA. This would allow me to replace or ignore NA values if I needed to later on. There were 2 datasets, one was data with graduation information and the second dataset had school programs. They linked with a DBN Id. The first visualization is the missing data. There is quite a lot of NA strings, but I did not use all of the columns and the plots take account of the NA data, so NA data remained.

DATA PREPARATION

In the data preparation step I renamed the DBN column as it imported with additional stings in the column, select a subset of the columns of interest and renamed columns with shorter names.

DATA SUMMARY

In the data summary section I am displaying some summary stats. The boxplot and linear plot scompares graduation rate to college enrollment and classifies the sports code, 1 for schools with sports programs and 0 with no sports programs. I begin with sports because it was my original topic for this project, but the analysis pulled me a different direaction as you will see later on. The linear chart does show a positive correlation of grad vs college enrollment which does make sense, but the box plot does not show a lot a variation for sports vs no sports.

MODELING

The modeling begins with a GLM comparison of sports vs graduation and college enrollment. The only significant variable was college enroll 2012. We then switch graduation and college as the response and sports code as the predictor and sports becomes slightly significant

I then switch the a linear model setting college 2012 enrollment as the response vs all the codes including language, leader, onlineAP, onlineLanguage and sports code. This model shows leader code is significant and language code as slightly significant. The last model removes all insignificant predictors and only compares college enroll 2012 to leader and language code and both codes become significant.

VISUALIZATIONS

In the visualizations we set college enroll to response and graduation as predictor and split the data by sports, leader and language codes. The sports split in the xyplot show there is no difference in a school having sports or no sports although the Graduation and College enrollment is positively correlated. For the language and lead plots you can see the schools with no language and no lead codes disappear. The GGplots confirm this, as you see the 0 for no sports make no difference , but lead and language programs do make a difference.

In the density plots we see a shift in having more 1 than 0’s at around the 80% grad and college enroll points. So I narrow down schools that have >80% graduation and college enroll rates in the histogram charts for each program and you can see that even the sports programs have a higher number of sports for schools that go past the 80% graduation and college enroll rate and for language and lead programs the college with neither program disappears.

MAP

In the Map we narrowed down to schools that have a >80% graduation and college enrollment rate. You can drill down to map the click a school to get more information on location, programs and number of students in the school

CONCLUSION

My conclusion is that this basic modelling shows that language and lead programs may be significant in helping kids graduate and enroll in college and in turn help them become successful. However, additional analysis must be done.

As an additional part of this project I a shiny app to analyze more variables in the dataset.

<https://apag101.shinyapps.io/final_project/>

The shiny app does the following;

* Drop down to select predictors and response variables
* Display linear, density and histogram plots
* Display basic statistical data
* Give table of schools where 80% or more students graduated and enrolled in college
* Give map of all schools with details of AP, Language, Leadership and sports programs