Sample Markdown

Austin Team

3/19/2022

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

This is a R Markdown Document that will serve as an example dashboard for the purposes of deliverable one. The goal of this document is to present a clean, organized, and simple layout that includes information about the topic of interest that the Austin Team wishes to explore and convey to all potential readers, displaying graphs, images of interest that will aid in analysis, and pride a layout that is user friendly. Ultimately, a more robust and fitting dashboard will be chosen later as the project draws completion.

## Topic

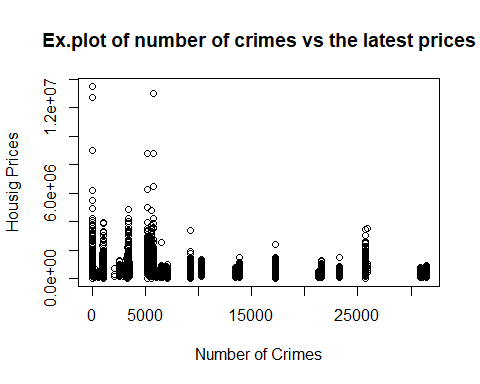
For Team Austin's project, we wish to preform analysis to determine if there is a connection to the increase price of houses in Austin as a relation to crimes in Austin. For this example, a merged data set that contains both housing prices and unique crimes committed are both sorted by zip codes and will be brought in.

To test how a machine learning model could be utilized, we can utilize functions in R studio to display a simple linear model test, whereas the housing prices will be the dependent and the number of crimes will be a single dependent variable:

## Bringing in Data

Firstly, the merged data sheet must be brought in and saved into R Studio to later be called upon:

Here is a plot that shows our two variables of interest:



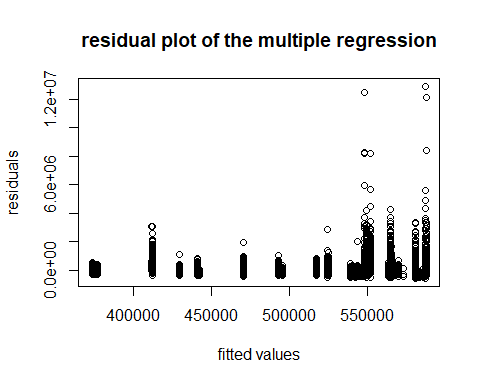
## Machine Learning Rundown

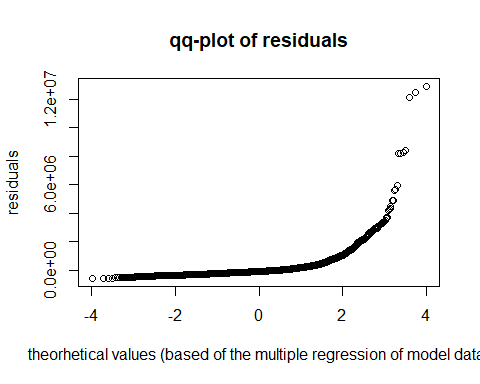
In Pandas, Machine learning was utilized to illustrate the results of a simple linear regression using the two datatypes. However, in a dashboard, we want little, if not any, lines of code for presentation; R Markdown allows us to hide code easily:

##   
## Call:  
## lm(formula = md$latestPrice ~ md$Count\_Crimes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -581872 -195807 -95275 49071 12912628   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.878e+05 5.554e+03 105.8 <2e-16 \*\*\*  
## md$Count\_Crimes -6.811e+00 3.804e-01 -17.9 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 448500 on 15169 degrees of freedom  
## Multiple R-squared: 0.02069, Adjusted R-squared: 0.02063   
## F-statistic: 320.5 on 1 and 15169 DF, p-value: < 2.2e-16

To quickly explain the results above, the numbers give us a good indication of what our linear model performances was. Without going into too many details, there are some indicators that tell us how appropriate our linear model was. An important value to consider is the R score; in the summary we see it is low, nearly at 0.0263 for r squared score. This essentially means that these two data types that we chose were not appropriate in the current state they were utilized in.

## Checking “Appropriateness” in our data

There are many ways to determine whether data utilzed for a model such as this was appriate or not. One way to check is to see what our data is trending towards. 

The residual plot tells us if there is any "clustering" of data present. In an ideal scenario, we want to see a flat cluster of data point across the board. Here, we see there is heavy clustering of points at the end of our data. We can explore further in the next plot. 

## Post Summary

Plots like a qq-plot can tell us what is happening to our data. In summary, we ideally want a linear regression to have all the data points as a straight line on this qq-plot. Our data portrays an exponential curve, this means higher values of our chosen variables are not matching up in a linear manner. This info can be useful in our post analysis where we talk further in the completed project. Going forward, we can try to improve our Machine learning model in a variety of ways, such as new merges, introducing new variables to consider in a Machine Learning model, etc. It is possible that after trial and error, the final model may still not be appropriate or that we couldn't make a stronger connection as anticipated.

More importantly, we can still talk about the model and the data itself. The main take away to note here is as data analytics, we are not trying to prove anything. A lot can be said about the data we are trying to use, even if our model isnt a good fit.