Project One: Data Analysis and Preprocessing

Allison Bolen
CIS 678
Wolffe
Winter 2018

Project One: Data Analysis and Preprocessing

Preprocessing

The data set provided contains some NaN values. This is a problem that needs addressing. After evaluating the data set it was observed that of 744 data entries only 8 of them were incomplete/invalid. Deleting these incomplete entries would not have an effect on the overall results of the data set, so removing them is sufficient for solving the invalid entries problem.

Visual

After deleting the incomplete entries in the data set, it was possible to generate a scatter plot of the remaining entries. The scatter plot is shown in figure 1.

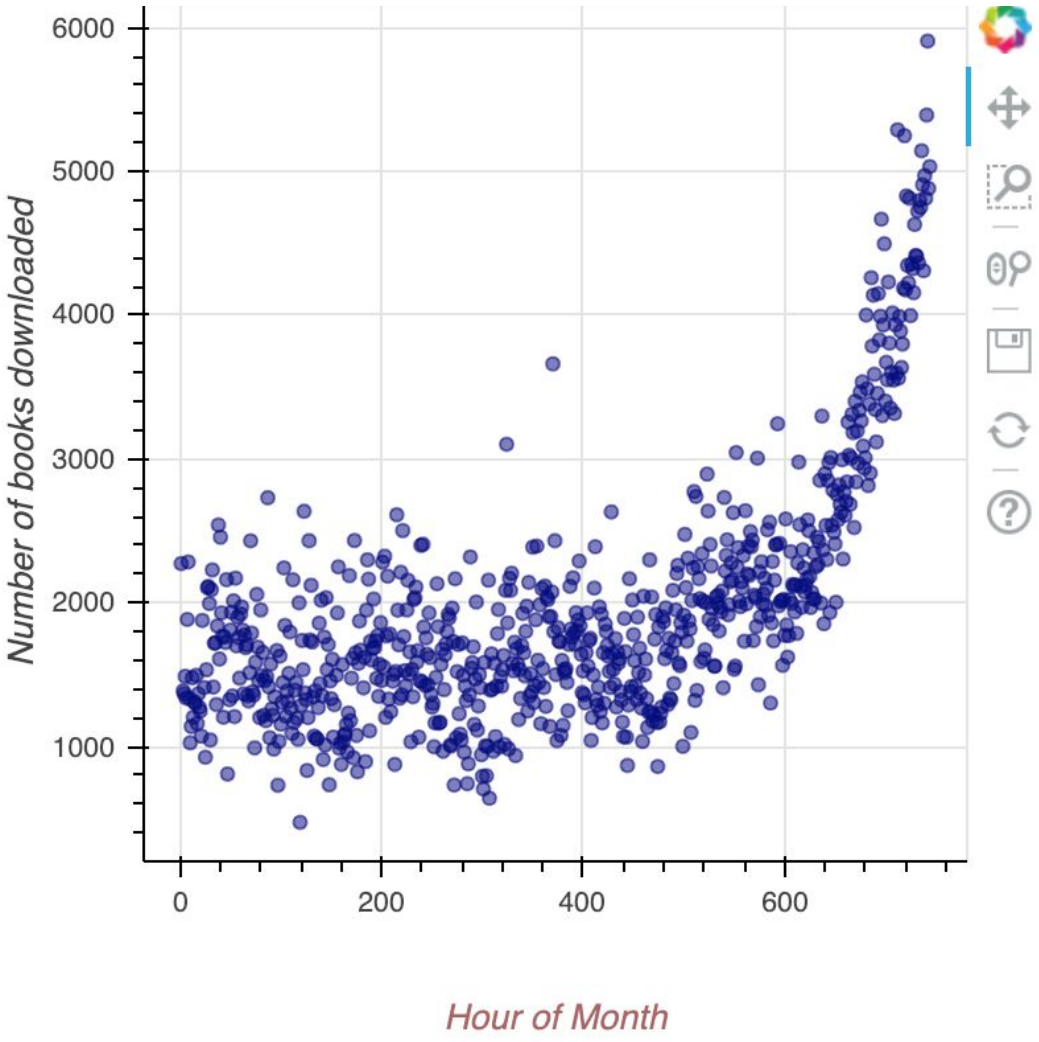

Figure 1: Scatter plot depicting number of books downloaded per hour over the course of one month.

This graph shows that as the month has gone on the book has gotten more popular and more copies have been downloaded. The growth trend represented by the data points roughly mimics an exponential function.

Project One: Data Analysis and Preprocessing

Analysis

Using linear regression we can determine a line of best fit to approximate values for data points in the future. Our linear regression line is mapped by the equation: $y = 0.6437135197845897x + 1724.7297164403078$. calculating the number of books downloaded at noon on the 5th day of the next month would require calculating the hour that noon on the 5th would be. This would be the result of 31 + 4 = 35, to get to the 4th day of the next month, then 35 * 24 = 840 the total hours occuring in 35 days so the end of the day on the 4th of next month. Then adding 12 to that value gets to the 5th day of next month at noon, so noon on the 5th of next month would be hour 852. Using the regression equation, you would expect to see 2273 downloads at noon on the fifth day of the next month.
A visual representation of the regression line is shown in figure 2.

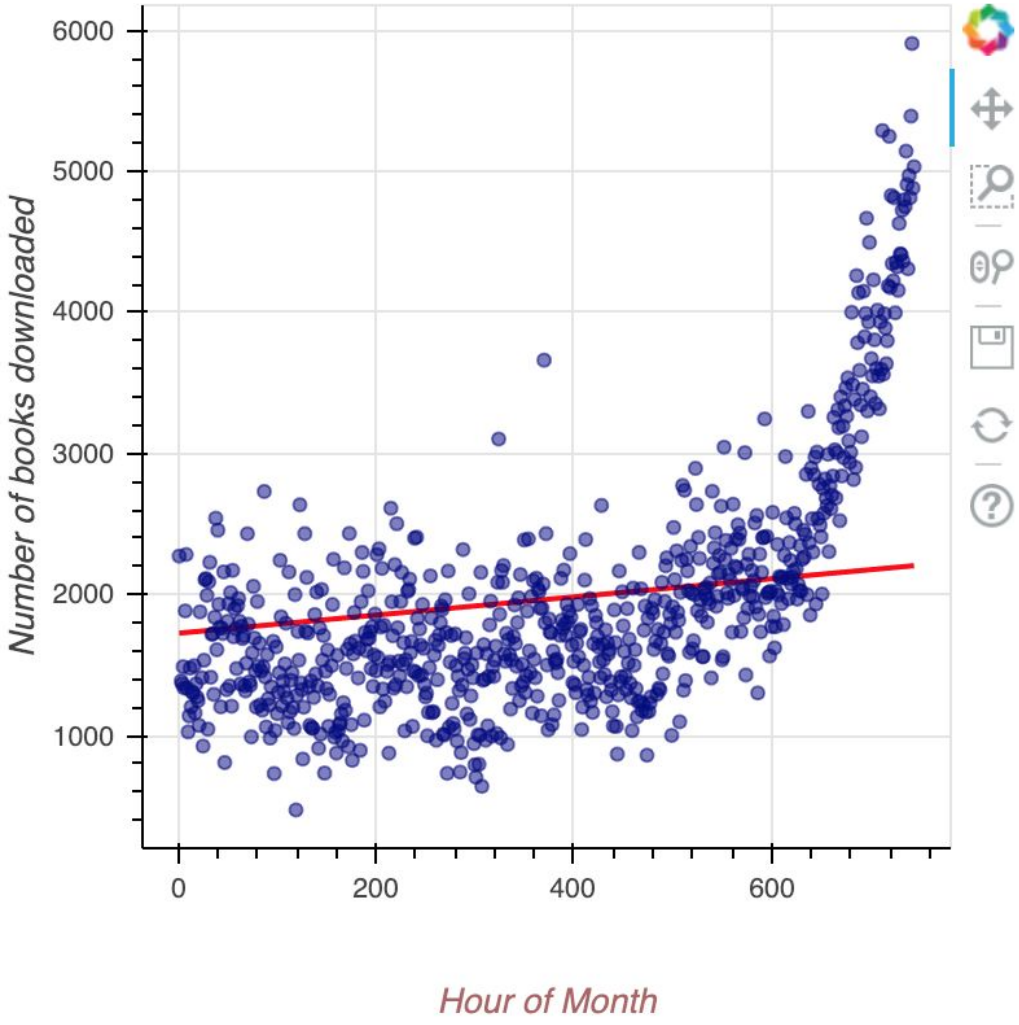

Figure 2: The linear regression line has is shown above.

Improvements

The linear regression analysis technique does not seem to account for the increase in downloads towards the end of the month due to a bloggers reviews on the book. To account for these changes it would be worth exploring exponential regression as an analysis technique. This would help account for the drastic increase towards the end of the month.

Alternatives

I used Weka as well to evaluate roughly what the scatter plot should look like and you can see the results in figure 3. The scatter plots seem to match up well.

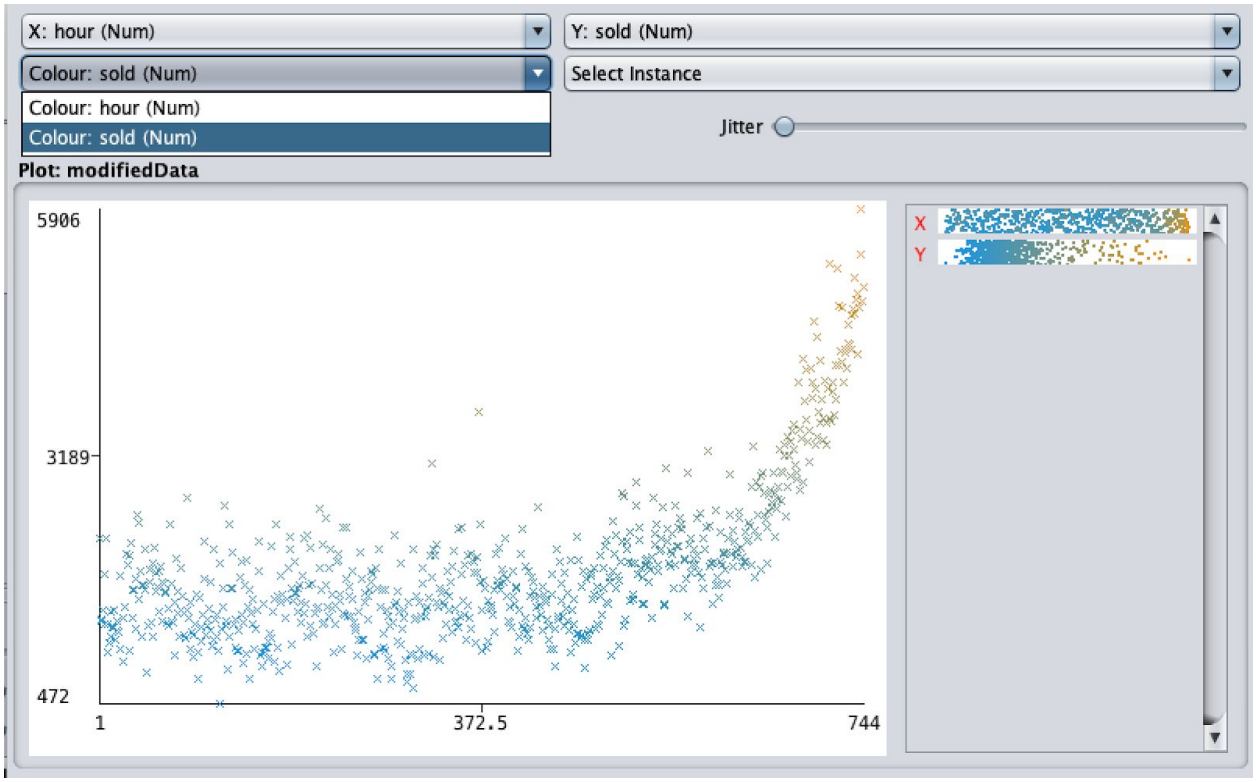

Figure 3: Scatter plot from Weka.

Weka also suggested that the regression equation would be $downloaded = 2.6193 * hour + 983.2275$. This equation would be visualized as shown in figure 4.

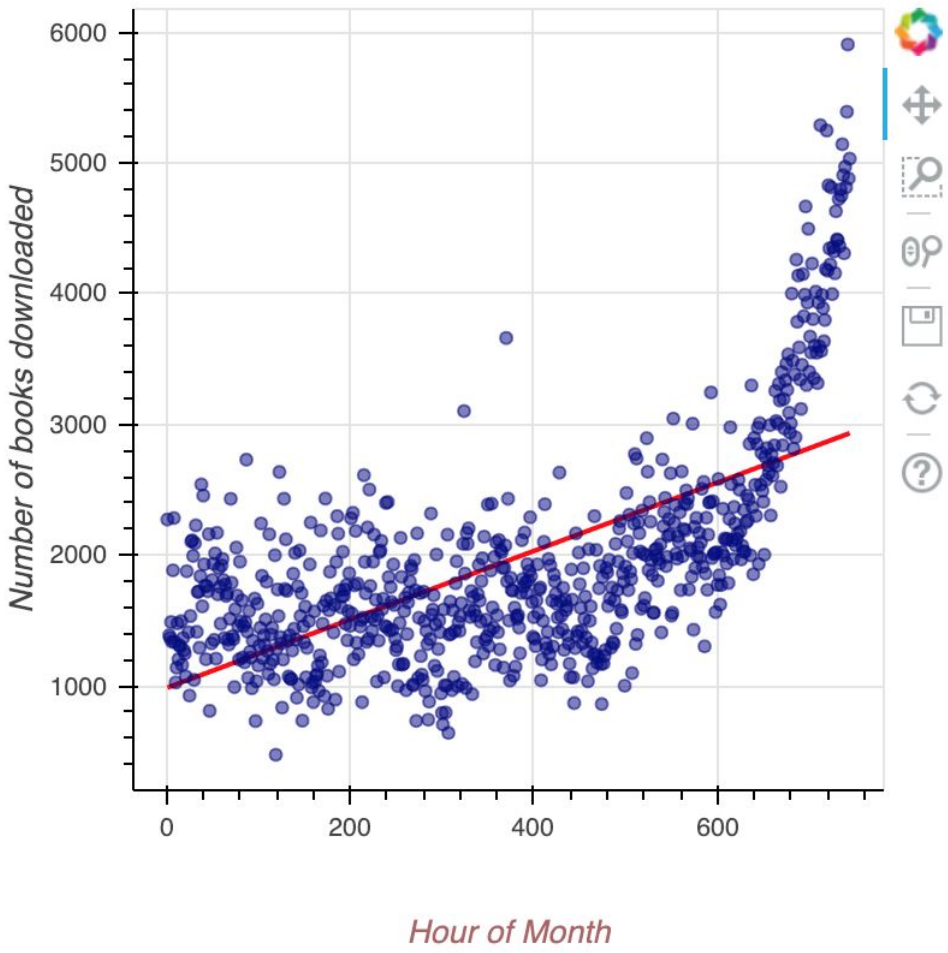

Figure 4: Weka's suggested equation compared to the scatter plot.

This differs from the linear equation I calculated in that is has a much steeper slope and a slightly lower intercept point. This seems to be better at accounting for the drastic increase in downloads towards the end of the month, so Weka must calculate linear regression in a slightly different manner.

Program

I chose to write my program using python3. I used the Pandas library for reading and navigating the data file and the Bokeh library for graphing the results. I would say that overall this project went smoothly. If I had a problem, I just googled it and usually found a working solution, most of my question revolved around Pandas syntax. It's something I look forward to learning how to use better.