Design & Discussion

**Getting Started**

I chose to use Python along with the matplotlib library for my data processing and graphing. I had only minimal experience with Python prior to this assignment, but I found it relatively intuitive with my Java background.

**Pre-processing**

I began by reading the provided data file’s contents into a two-dimensional list. I then replaced missing values with the mean of the next adjacent valid values on either side of it. If a missing value was not preceded or followed by a valid value, then it was left as “None” and excluded from the graph and calculations. I wrote then program this way to avoid issues if there were one or more missing values at the beginning or end of the data. I chose to fill in missing values in that were surrounded by valid values after observing that the data lacked seasonality but did have a trend. Using the mean of the adjacent values allowed me to keep the time series complete while using values that conformed to the local trend.

**Visualization**

As mentioned above, I used matplotlib to graph my processed data. The methods provided with pyplot made it very simple to create my scatter chart and later add my trendline. I also recreated the chart in Excel to confirm its accuracy.

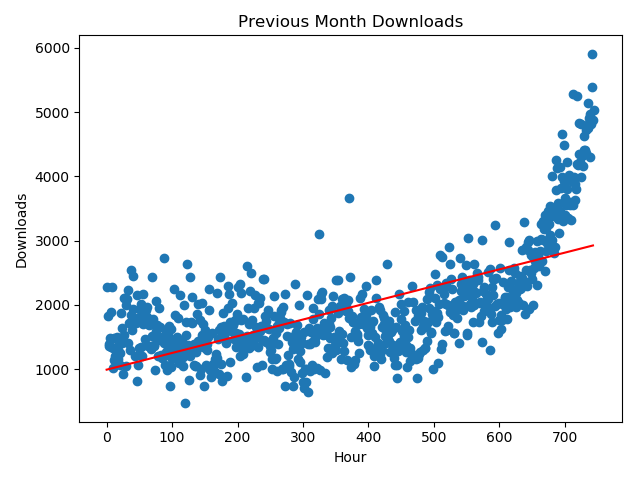
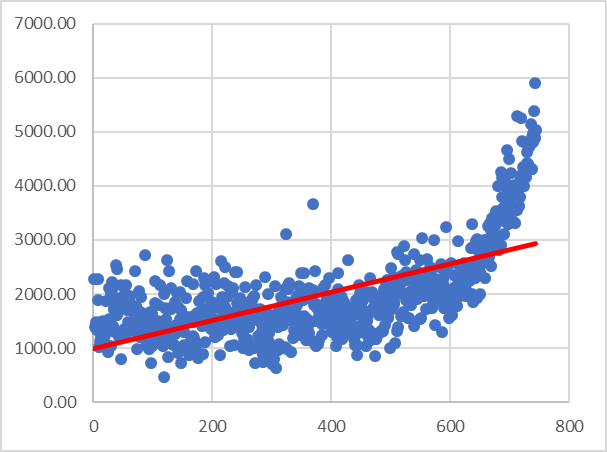


Figure : Excel

Figure : Python and matplotlib

The visualization allowed me to see that the book sales were had a wide range with a relatively steady mean for the first half of the month before taking off into a virtually exponential increase in the second half.

**Analysis**

I used the least-squares method provided in the assignment to define my linear regression equation and create my trendline. Using the equation, I was able to estimate that there would be 3209 downloads at noon on the fifth day of the next month (see sample output page).

Regarding the blogger who mentioned the book, I do not think this linear analysis fully captured the effect that it had. The linear trendline gives the appearance of a steady increase over the course of the month, but the reality is that the downloads were mostly flat until shooting up at the end. The blogger’s impact might be better visualized through a non-linear trendline that shows a rolling average of the past several hours or days.