Week 5: Finding Patterns within Library Data

Natalia DuBon

I. Abstract

For my project for this week, I would like to continue what I had started last week but dive deeper into pattern recognition and hopefully create a better machine learning model for prediction. Regression is a method of modeling a target value based on independent predictors. This method is mostly used for *forecasting* and finding out the *cause and effect relationship* between variables (association). Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables. The idea is to apply linear regression to multiple sets of data as I had started with last week. It's the mathematically best way of determining the trend over the hour for each ItemNumber and it will select out only the ones with a positive trend. For this week's student forum on patterns, I decided to explore the Seattle Public Library dataset to find statistical correlations between the progression of time and a subject's total corresponding checkouts. For this, I've decided to choose all items that relate to Data Science (as last week we found that Data Science had the greatest upward trend). I essentially want to discover if I can make a predictable statistical linear model that will be able to answer my question regarding such correlation. All queries are cited along the descriptions/analysis and can also be found in its own section further below. Note that I have chosen to use both SQL and R for this week's assignment due to some limitations I find SQL to have in comparison to R regarding running statistical methods.

II. Addressing Previous Issues From Last Week

Problem: Not enough data points to create an accurate model

A. Solution: Last week, I categorized the data by year, when I should have used smaller time increments such as progression of months/days/or even hours. After the individual project meetings, the final conclusion has been to focus on the year 2019 and use weekly increments. This should yield 52 points which is much larger than my original 17 points from last week.

Problem: Relationships do not always follow a linear pattern

B. Solution: Implement a non-linear regression model such as a cubic one if needed. Additionally, the previous model had included values during the Pandemic which are obvious outliers and as a consequence, diminished the quality of the machine learning regression model. The model becomes harder to train for this reason, and I've decided to use values from before the pandemic.

Problem: Frequency pattern mining is *not necessarily about total volume but more about performance over time*

C. Solution: By improving the above points and also training my machine learning model further, I should be able to make future predictions this time and test them against observed values which I was unable to do at all last week. I will be able to recognize patterns over the performance in time and predict future values using such pattern and have it be mathematically accurate.

III. Description and Analysis

Regression is a method of modeling a target value based on independent predictors. This method is mostly used for *forecasting* and finding out the *cause and effect relationship* between variables (association). Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.In order to make appropriate linear predictions, there must be some correlation between independent(explanatory) and dependent (response) variables. Though I still remain unsure of the correlation until I pull data, I'd like to start this prediction model by first exploring the popularity of computer science and data science books. This choice is largely based on the common idea that computer science careers and related fields are growing rapidly in comparison to other occupations. According to the US Bureau of Labor Statistics "overall employment in computer and information technology occupations is projected to grow 15 percent from 2021 to 2031, much faster than the average for all occupations"[1]. The natural question to ask here is if this growth in job availability also inspires public interest in the field (or vice versa). Most relevantly, does this inspiration translate to checkouts at the Seattle Library? The first query starts to simply explore the entire data set for items that have the words "computer science" or "data science" within their title. The purpose of this query is to essentially visually see how many data entries are received in order to assess whether we have enough to create an accurate prediction model [**Query A**]. Looking through the results, we have at least 1,000 data entries which is sufficient in building a prediction model. A quick overview of the results also shows that there are multiple instances of duplicates, which I have decided to keep moving forward considering that each copy is

vital in exploring the overall popularity of the results. There is no need for distinct titles in this search and I can move on to cleaning and organizing the data. We will only be using the Data Science portion of this data set!

[**Query A**] **Duration:** 20.173 sec / **Fetch time**: 42.295 sec

```
SELECT *
FROM spl_2016.outraw
WHERE TITLE LIKE '%computer science%' OR TITLE LIKE '%data science%';
```

CSV:

- Query A - Week_3_Query_A.pdf

Now that I've been able to explore the data, I now plan on organizing it into a table of independent (explanatory) and dependent (response) variables [**Query D**]. The independent variable here would be time, in this case either year or month values. For the purpose of organization and making the data easily readable, I first am using both measurements. The dependent variables are items checked out entailing computer science and data science. The purpose of this query is to count each checked out title by summing each one individually. The yielded results were quite interesting in that the numbers are never exceedingly large, which is unexpected; most checkouts stay within a one to two digit range, never exceeding 30. Data Science titles specifically didn't showcase any results until 2014, truly showcasing the emergence of the field. Despite having a younger history than Computer Science, we can see that in recent times, it seems to stay in the same ballpark of values as its counterpart, which is impressive considering that it is more

niche than the latter. Again, the Data Science portion is the only set that we will be proceeding with for this week considering the results of last week's statistical pattern analysis.

[Query D] **Duration:** 80.243 sec / **Fetch time**: 0.000 sec

```
SELECT YEAR(cout) AS Year, month(cout) AS Month,
SUM(CASE
WHEN title LIKE '%computer science%' Then 1
ELSE 0 END) AS 'Computer Science',
SUM(CASE
WHEN title LIKE '%data science%' Then 1
ELSE 0 END) AS 'Data Science'
FROM spl_2016.inraw
WHERE
YEAR(cout) >= '2006'
GROUP BY month(cout), YEAR(cout)
ORDER BY YEAR(cout) ,  month(cout);
```

CSV:

- Query D - Week_3_Query_D.pdf

It is in this step where I have to make proactive changes to my data set in order to create a more accurate prediction model. We can see that Data Science has a peak in 2019, so I'm going to choose the year prior (2018) to separate/categorize weekly (rather than yearly

from last week). This will yield more data points for our regression model later which in turn will make it more accurate. After I train the model and detect patterns, I should be able to predict the surge in Data Science item checkouts in 2019. (**Query E**)

[**Query E**] **Duration:** 80.243 sec / **Fetch time**: 0.000 sec

```
SELECT WEEK(cout) AS Week,
SUM(CASE
WHEN title LIKE '%data science%' Then 1
ELSE 0 END) AS 'Data Science'
FROM spl_2016.inraw
WHERE
YEAR(cout) = 2018
GROUP BY WEEK(cout)
ORDER BY WEEK(cout) ;
```

CSV:

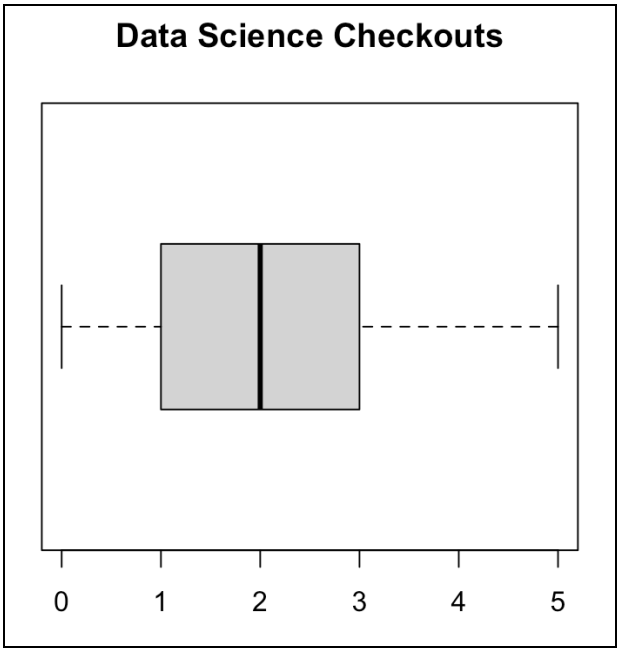
- **Query E (part 2) - WeeklyData2.pdf**

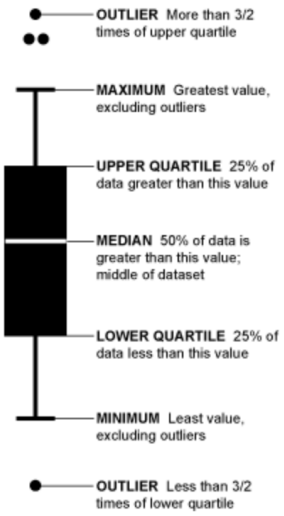
Since making a new dataframe from the Seattle Library is prohibited (I am not allowed access to do this from my knowledge), I exported the csv file and imported it into Google Sheets. I then transferred over to RStudio, a desktop version of R, to further run statistical analysis on the newly created dataset. In this dataset, 'week' is my independent predictor variable. After I was able to import the data into RStudio [**Query E**], I ran a summary report which allowed me to view the minimum, median, mean, and

maximum values of the independent variable and dependent variables (data science checkouts).

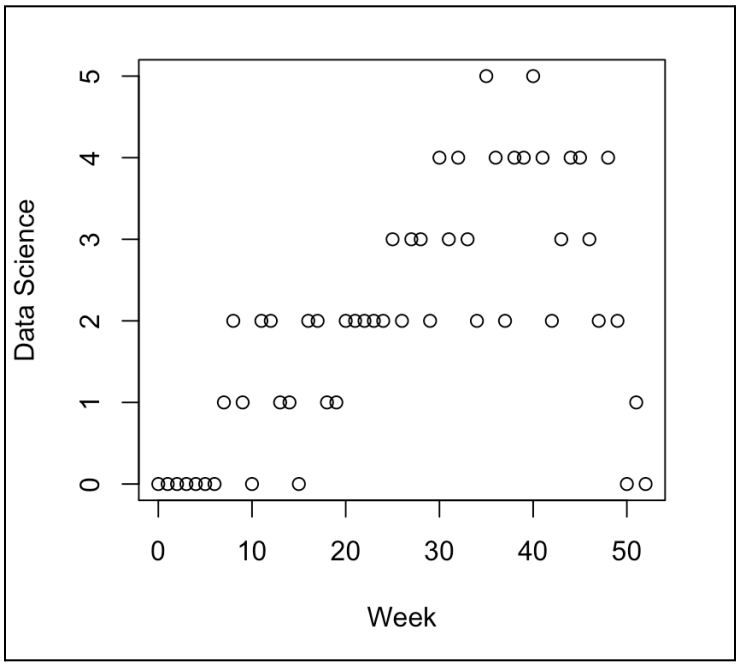
```
> summary(df1)
      Week       Data Science
 Min.   : 0   Min.   :0.000
 1st Qu.:13   1st Qu.:1.000
 Median :26   Median :2.000
 Mean   :26   Mean   :2.038
 3rd Qu.:39   3rd Qu.:3.000
 Max.   :52   Max.   :5.000
```

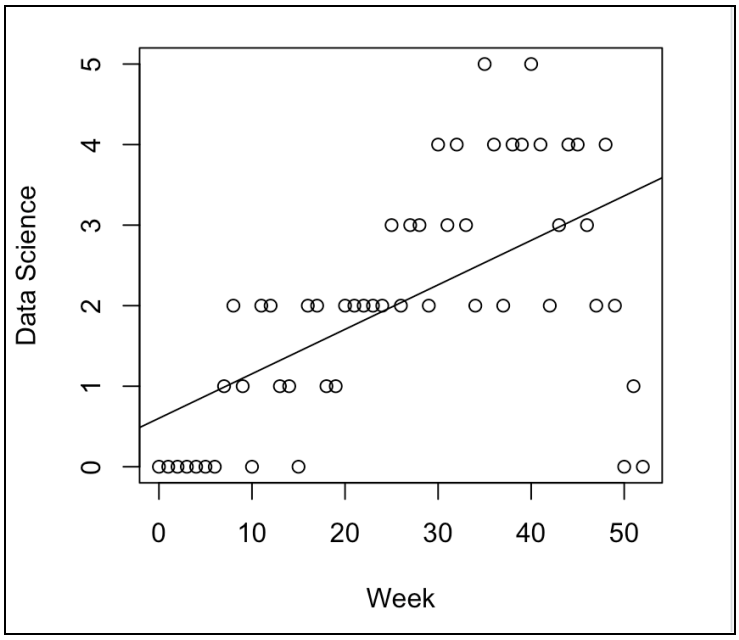
We can see overall that checkouts for Data Science items per week span anywhere from zero to five at most. There is not much that is too surprising or informing in this summary, so let's continue to visualize. Below is a boxplot primarily showcasing this statistical summary in visualization form. I've attached an explanatory diagram, as well [2]. The relationship between the independent and dependent variable must be linear in order to run statistical methods regarding linear regression. Therefore, my next step is to test this visually with a scatter plot to see if the distribution of data points could be described with a straight line [**Query F**]. We can see that we've fixed our previous problem from last week concerning our lack of data points; we've increased our dataset and can overall see an upward trend much more clearly than we did last week. There seems to be outliers near the end where we get zero checkouts, but overall we see an upward trend through the entirety of 2018, which will allow us to train the model more precisely.





Scatter Plot for Data Science:





[Query F] **Duration:** 80.243 sec / **Fetch time**: 0.000 sec

```
install.packages("ggplot2")

install.packages("dplyr")

install.packages("broom")

install.packages("ggpubr")

install.packages('googlesheets4')


#Load the required library

library(googlesheets4)

library(ggplot2)

library(dplyr)

library(broom)

library(ggpubr)
```

```
#Reads data into R
df1<-
read_sheet('https://docs.google.com/spreadsheets/d/1bKnxgh8qmrYusCGZzyPngs2Z4KF
UtGzvK-sdER60uZI/edit?usp=share_link')
#Prints the data
df1
summary(df1)
boxplot(df1$`Data Science`, horizontal=TRUE, main="Data Science Checkouts")
plot1<-plot(`Data Science` ~ Week, data = df1)
plot1
abline(lm(`Data Science` ~ Week, data = df1))
df.lm <- lm(`Data Science` ~ Week, data = df1)
summary(df.lm)

new <- data.frame(Week=c(77))
predict(df.lm, newdata=new)
```

Visually, we can see that there is a positive influence in our data set and can compute the most accurate regression line, but we should test a null and alternative hypothesis to run further statistical analysis. The null hypothesis states that there is no relationship between the variables being studied (one variable does not affect the other). It states the results are due to *chance* and are not significant in terms of supporting the

idea being investigated. The alternative hypothesis states that the independent variable did affect the dependent variable, and the results are significant in terms of supporting the theory being investigated (i.e. not due to chance). To test to see whether there is a significant positive relationship between weekly progression and data science checkout, I ran a statistical analysis procedure in which I turned the data into a linear model and then grabbed a statistical summary on it [**Query F**]. The final three lines are model diagnostics – the most important thing to note is the p-value (here it is 0.00000498, or almost zero), which will indicate whether the model fits the data well. From these results, ***we can say that there is a significant positive relationship between time progression and data science item checkouts (p-value < 0.001), with a 0.05523 -unit (+/- 0.01) increase in checkouts for every unit increase in time (year).*** It is in this step specifically where we can mathematically see the biggest improvement from last week's model. The p-value last week was 0.0001629, which was good, but now we were able to decrease it drastically which proves our machine learning model will be more accurate than the previous. Therefore we reject the null hypothesis and suggest the alternative.

```
Call:
lm(formula = `Data Science` ~ Week, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4738 -0.6569  0.0233  0.8518  2.4652

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.60168    0.32649   1.843   0.0712 .
Week         0.05523    0.01082   5.103 4.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.205 on 51 degrees of freedom
Multiple R-squared:  0.338,     Adjusted R-squared:  0.3251
F-statistic: 26.04 on 1 and 51 DF,  p-value: 4.989e-06
```

Outside of statistical analysis and finding correlation, a good model should be able to accurately predict future values. In order to fully understand the following, we should look back at our summary report and take note of The Residual Standard Error. The Residual Standard Error is the average amount that the response variable will deviate from the true regression line. To test this, I chose two random weeks in 2019 using the seed function in R (to make sure it's random and unbiased). Say the week is 38 in 2019 as in the first example, I would then add that to 52 (to follow the time series from 2018) to create a mathematical prediction for how many checkouts will be made for that week's total.

Predicting Future Checkouts from Patterns (Examples):

Week 38 in 2019 (Week 90 in total):

```
> new <- data.frame(Week=c(90))
> predict(df.lm, newdata=new)
       1
5.57265
```

From the actual dataset in SQL, we can see that the actual value for week 38 in 2019 is 5 checkouts, which shows an accurate reflection from our model. There is residual error between the predicted value and the observed value, but it's obvious in this case that the model predicted the value accordingly.

Week 25 in 2019 (Week 77 in total):

```
> new <- data.frame(Week=c(77))
> predict(df.lm, newdata=new)
       1
4.85462
```

Again, from the actual dataset in SQL, we can see that the actual value for week 25 in 2019 is 7 checkouts which, if you view the entirety of the dataset, is one of the highest values of that year and occurs in the middle of the year rather than at the end (in some ways like an outlier). Here, our model predicts a lower value than was actually observed. Even in considering residual standard error (1.205), the model can only predict upwards to 6.0596 which is still lower than the observed value but closer to its range. This could suggest that this particular day may be an outlier in its dataset.

IV. Final Conclusion

A statistically significant result cannot prove that a research hypothesis is correct (as this implies 100% certainty). Instead, we may state our results "provide support for" or "give evidence for" our research hypothesis (as there is still a slight probability that the results occurred by chance and the null hypothesis was correct – e.g. less than 1%). The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis (**there's no effect in**

the population). The smaller the p-value, the more likely you are to reject the null hypothesis. After running statistical analysis procedures on the data regarding checkouts made across the weeks of 2018 for Data Science, it seemed that there was a statistical significance that indicates an upward pattern/trend in the data. In fact, from my results, ***we can say that there is a significant positive relationship between time progression and data science item checkouts (p-value < 0.001), with a 0.05523 -unit (+/- 0.01) increase in checkouts for every unit increase in time (year).*** It is likely safe to say that outside influence on the popularity of Data Science (through media, word of mouth, company needs/interest) through the passage of time has had a statistically predictable effect on corresponding checkouts at the Seattle Public Library. As a result, we could finally complete a model that has recognized this pattern and can predict future ones.

V. Resources

[1]https://www.mat.ucsb.edu/~g.legrady/academic/courses/17w259/freqpatternMining.pdf

[2]https://learnsql.com/blog/high-performance-statistical-queries-sql-part-1-calculating-frequencies-histograms/

[3]https://www.cs.put.poznan.pl/mwojciechowski/papers/adbis99b.pdf

[4]https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a