Week 4: Finding Patterns within Library Data

Natalia DuBon

I. Abstract

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 Computer Science and/or Data Science. 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. Description and Analysis

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.

[**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

Another observation made is that nearly all of the titles have their own dewey class, meaning (as expected) most titles are non-fiction. However, this raises the question, which computer science books are fiction and why? After running another search, we see that the only fictional item is titled “Lauren Ipsum: A Story About Computer Science and Other Improbable Things”, which after a quick Google search, shows this item is a

children's book [**Query B**]. This brings up a vital personal question, do fictional items influence popularity in an otherwise non-fictional field? Because influence spans multiple generations (those younger and older audiences), I've made the personal decision to keep this item within my data set and continue with tabling my data. However, I want to verify this by creating a table that showcases how many dewey and non-dewey titles circulate per year in order to compare the two [**Query C**]. You can see that there is an obvious growth in the non-fictional side, and almost a hyperbolic growth for the fictional side. The most checkouts the fictional side ever gets within the span 16 years is 23 in the year 2015. It then proceeds to decrease in the following years which is interesting. For now, I plan to stick to my original plan to use both types of data as my logic remains that any interest in computer science material, whether through fictional storytelling or not, shows an influence on the overall public appeal towards the field.

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

```
SELECT *
FROM spl_2016.outraw
WHERE deweyClass = "" AND (TITLE LIKE '%computer science%' OR TITLE LIKE
'%data science%');
```

Week_3_Query_B - Week_3_Query_B.pdf

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

```
SELECT
 Year(cout) AS year,
 SUM(CASE
```

```
 WHEN deweyClass = "" THEN 1
 ELSE 0
 END) AS nonDewey,
 SUM(CASE
 WHEN deweyClass != "" THEN 1
 ELSE 0
 END) AS Dewey
FROM
 spl_2016.inraw
WHERE
 YEAR(cout) >= 2006
 AND TITLE LIKE '%computer science%' OR TITLE LIKE '%data science%'
GROUP BY year
ORDER BY year
```

CSV:

- Query C - Week_3_Query_C.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.

[**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:

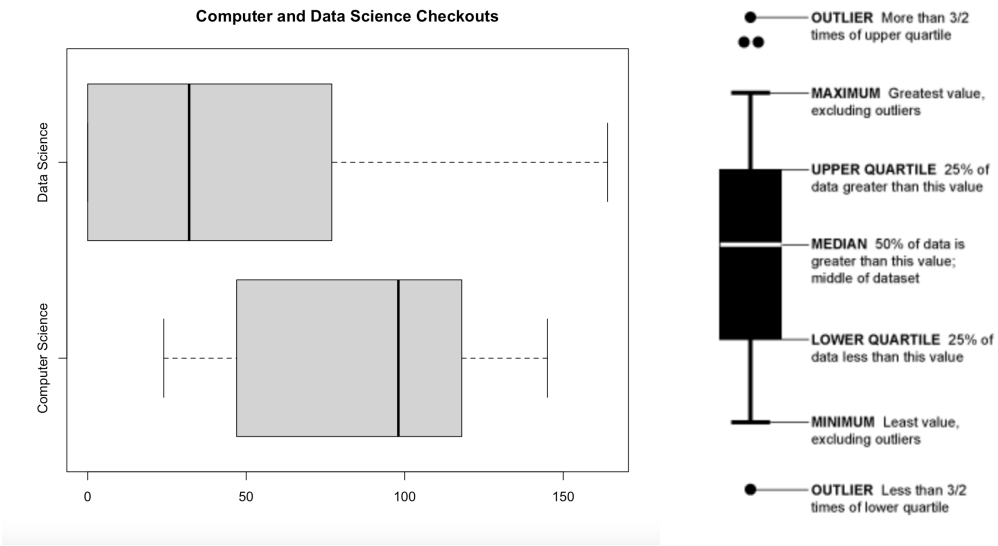
- Query D - Week_3_Query_D.pdf

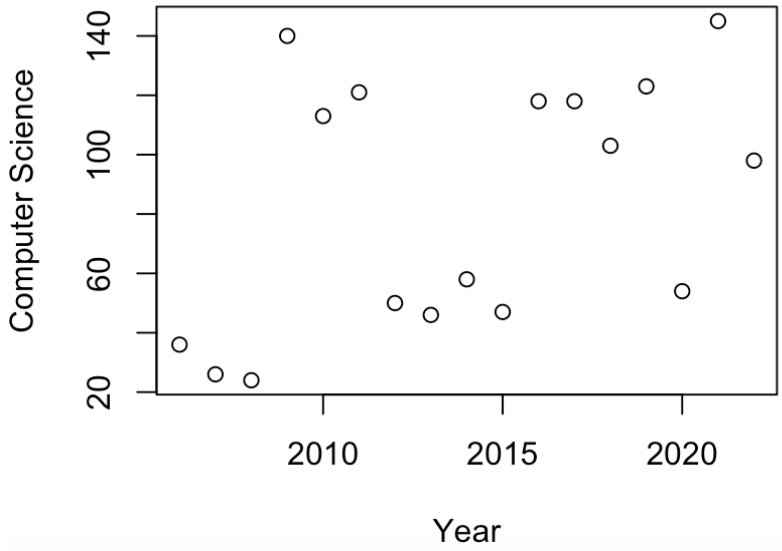
Since making a new dataframe from the Seattle Library is prohibited (I am not allowed access to do this from my knowledge), I exported the csv file and imported it into Google Sheets. I then transferred over to RStudio, a desktop version of R, to further run statistical analysis on the newly created dataset. It should be noted that I made the choice to drop the month column from my dataframe as if I had made the initial choice to count each month continuously, that would yield a timeline of 192 entries (192 months within 16 year span); this is longer than I would personally like in terms of later plotting this dataframe. So, instead I decided to use year as my independent 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 (computer science and data science checkouts).

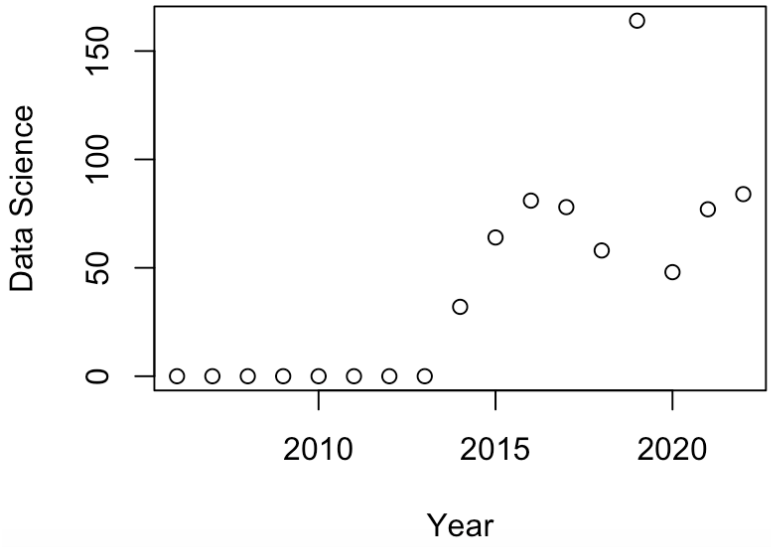
```
> summary(df)
      Year      Computer Science  Data Science
 Min.   :2006   Min.   : 24.00   Min.   :  0.00
 1st Qu.:2010   1st Qu.: 47.00   1st Qu.:  0.00
 Median :2014   Median : 98.00   Median : 32.00
 Mean   :2014   Mean   : 83.53   Mean   : 40.35
 3rd Qu.:2018   3rd Qu.:118.00   3rd Qu.: 77.00
 Max.   :2022   Max.   :145.00   Max.   :164.00
```

We can see overall that checkouts for Computer Science items are higher than that of Data Science by examining the mean of each (83.53 and 40.35 respectively). However, the maximum for Data Science is slightly higher which is also surprising given the lower mean! This was a discovery I didn't initially notice when I first ran the data in SQL. It is

in 2019 when Data Science reaches this peak, and Computer Science reaches 123 checkouts. 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 E**]. In general, there are not a lot of data points to work with and there are many outliers. Visually, Computer Science looks a lot more scattered than Data Science does. What's important to note though is that Data Science starts to show linearity past the year 2014 (when it was first checked out). Therefore, I went back to SQL to only extract data past 2014 for Data Science specifically [**Query F**].







[Query F] using SQL

SELECT YEAR(cout) AS Year,

SUM(CASE

```
WHEN title LIKE '%data science%' Then 1
ELSE 0 END) AS 'Data Science'
FROM spl_2016.inraw
WHERE
YEAR(cout) >= '2014'
GROUP BY YEAR(cout)
ORDER BY YEAR(cout) ;
```

- Week_3_Query_F - Week_3_Query_F.pdf

To test to see whether this is a significant positive relationship between year progression and computer 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 G**]. The final three lines are model diagnostics – the most important thing to note is the p-value (here it is 0.0001629, 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 7.38 -unit (+/- 0.01) increase in checkouts for every unit increase in time (year).*** However, when running the same analysis for Computer Science items instead, the P-value is not great enough to dictate a significant positive relationship between time progression and computer science item checkouts. The data is much more sporadic than Data Science and would require further advanced analysis or a much broader data set.

For Data Science:

```
Call:
lm(formula = `Data Science` ~ Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-36.632 -15.392  -8.353  15.507  86.748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14822.77    2981.37  -4.972 0.000167 ***
Year             7.38       1.48   4.985 0.000163 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.9 on 15 degrees of freedom
Multiple R-squared:  0.6236,    Adjusted R-squared:  0.5985
F-statistic: 24.85 on 1 and 15 DF,  p-value: 0.0001629
```

For Computer Science:

```
Call:
lm(formula = `Computer Science` ~ Year, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-51.77 -31.59 -15.18  27.06  75.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7380.118   3875.638  -1.904   0.0762 .
Year            3.706      1.924   1.926   0.0733 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.87 on 15 degrees of freedom
Multiple R-squared:  0.1982,    Adjusted R-squared:  0.1448
F-statistic: 3.709 on 1 and 15 DF,  p-value: 0.0733
```

[Query G] using RSTUDIO

df.lm <- lm(`Computer Science` ~ Year, data = df)

summary(df.lm)

```
df.lm2 <- lm(`Data Science` ~ Year, data = df)
summary(df.lm2)
```

III. Final Conclusion

After running statistical analysis procedures on the data regarding checkouts made across the years for Computer Science and Data Science, it seemed that only the latter somewhat shows signs of linear regression. 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 7.38 -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.

IV. Queries (Collection)

[**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

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

```
SELECT *
FROM spl_2016.outraw
WHERE deweyClass = "" AND (TITLE LIKE '%computer science%' OR TITLE LIKE
'%data science%');
```

- Week_3_Query_B - Week_3_Query_B.pdf

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

```
SELECT
 Year(cout) AS year,
 SUM(CASE
 WHEN deweyClass = "" THEN 1
 ELSE 0
 END) AS nonDewey,
 SUM(CASE
 WHEN deweyClass != "" THEN 1
 ELSE 0
 END) AS Dewey
FROM
 spl_2016.inraw
WHERE
 YEAR(cout) >= 2006
 AND TITLE LIKE '%computer science%' OR TITLE LIKE '%data science%'
GROUP BY year
ORDER BY year
```

CSV:

- Query C - Week_3_Query_C.pdf

[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

[Query E] **using RSTUDIO**

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages('googlesheets4')
#Load the required library
```

```
library(googlesheets4)

library(ggplot2)

library(dplyr)

#Reads data into R

df <-

read_sheet('https://docs.google.com/spreadsheets/d/1XUU0i03ASXCNML0QcWrx1sVX

7VYmIXTzRlj2_CUGrwE/edit?usp=sharing')

#Prints the data

df

summary(df)

# Boxplot:

boxplot(df[,-1],horizontal=TRUE, main="Computer and Data Science Checkouts")
```

[Query F] using SQL

```
SELECT YEAR(cout) AS Year,

SUM(CASE

WHEN title LIKE '%data science%' Then 1

ELSE 0 END) AS 'Data Science'

FROM spl_2016.inraw

WHERE

YEAR(cout) >= '2014'

GROUP BY YEAR(cout)

ORDER BY YEAR(cout) ;
```

[Query G] using RSTUDIO

```
df.lm <- lm(`Computer Science` ~ Year, data = df)
summary(df.lm)
df.lm2 <- lm(`Data Science` ~ Year, data = df)
summary(df.lm2)
```

V. References

[1]https://www.bls.gov/ooh/computer-and-information-technology/home.htm#:~:text=Overall%20employment%20in%20computer%20and,new%20jobs%20over%20the%20decade.

[2]https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

[3]https://flowingdata.com/2012/05/15/how-to-visualize-and-compare-distributions/