Week 8: Finding Outliers in Data

Natalia DuBon

I. Abstract

This week's assignment calls for us to explore outliers in the Seattle Library database. For this project, I decided to conduct a statistical experiment that would allow me to search for outliers within a database in addition to statistically proving whether or not that outlier has a negative influence on the overall scope of the data and regression model. My research involves heavily on more complicated statistical approaches beyond just calculating the standard deviations of the dataset, but are explained simplistically throughout this paper in order to provide an easier understanding of the analysis attached to these methods.

II. Query Explorations

To start off a topic for this project, I decided to focus on the categorical data of dewey numbers. This is simply done in order to focus on a subset of the data to find outliers. I ran a general query that would extract the total checkout of books per query number, which I chose to do just to choose the dewey number with the most overall checkouts to continue my analysis for the project.

[**Query A**] **Duration:** 172.243 sec

```
SELECT
 YEAR(cout) AS Years,
 SUM(CASE
 WHEN deweyClass > 000 AND deweyClass < 100 THEN 1
```

```
ELSE 0

END) AS Dewey000_099,

SUM(CASE

WHEN deweyClass > 100 AND deweyClass < 200 THEN 1

ELSE 0

END) AS Dewey100_199,

SUM(CASE

WHEN deweyClass > 200 AND deweyClass < 300 THEN 1

ELSE 0

END) AS Dewey200_299,

SUM(CASE

WHEN deweyClass > 300 AND deweyClass < 400 THEN 1

ELSE 0

END) AS Dewey300_399,

SUM(CASE

WHEN deweyClass > 400 AND deweyClass < 500 THEN 1

ELSE 0

END) AS Dewey400_499,

SUM(CASE

WHEN deweyClass > 500 AND deweyClass < 600 THEN 1

ELSE 0

END) AS Dewey500_599,

SUM(CASE
```

```
 WHEN deweyClass > 600 AND deweyClass < 700 THEN 1
 ELSE 0
 END) AS Dewey600_699,
 SUM(CASE
 WHEN deweyClass > 700 AND deweyClass < 800 THEN 1
 ELSE 0
 END) AS Dewey700_799,
 SUM(CASE
 WHEN deweyClass > 800 AND deweyClass < 900 THEN 1
 ELSE 0
 END) AS Dewey800_899,
 SUM(CASE
 WHEN deweyClass > 900 AND deweyClass < 1000 THEN 1
 ELSE 0
 END) AS Dewey900_999
FROM
 spl_2016.inraw
WHERE
 itemtype like '%bk'
 AND YEAR(cout) >= '2006'
 AND YEAR(cout) <= '2019'
GROUP BY YEAR(cout);
```

CSV RESULTS (*click* on file link below):

- Week 8 Query A - Week8_QueryA.pdf

Looking at the extracted data, we could see that overall the checkout for books for each dewey class is evenly distributed across each. However, it is clear that the 700 dewey class has the most checkouts almost every year, so I'm choosing to move forward with this specific classification because of the immense amount of data points (that will yield more accurate results).

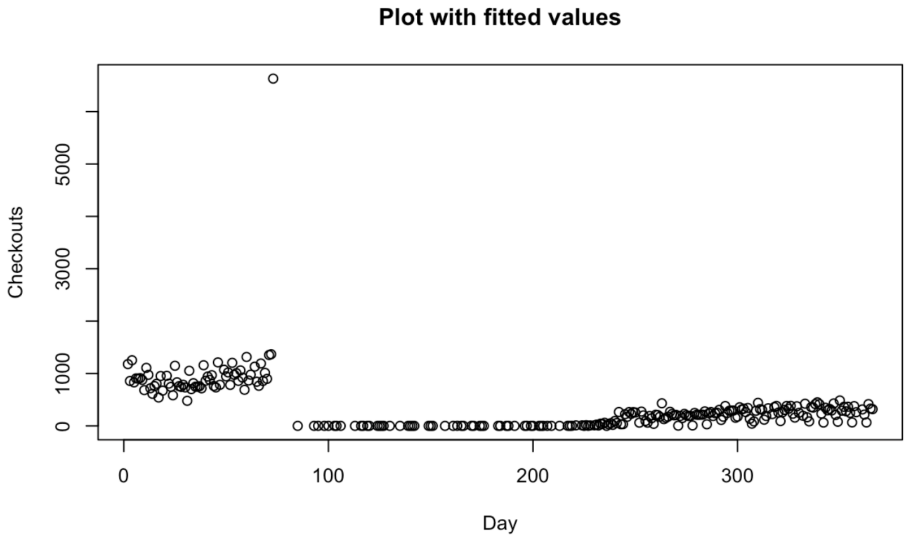
By Day:

```
SELECT DAYOFYEAR(cout) AS Day,
SUM(CASE
WHEN (deweyClass > 700 AND deweyClass < 800
AND  itemtype like '%bk') Then 1
ELSE 0 END) AS 'Checkouts'
FROM spl_2016.inraw
WHERE
YEAR(cout) = 2020
GROUP BY DAYOFYEAR(cout)
ORDER BY DAYOFYEAR(cout) ;
```

CSV RESULTS (*click* on file link below):

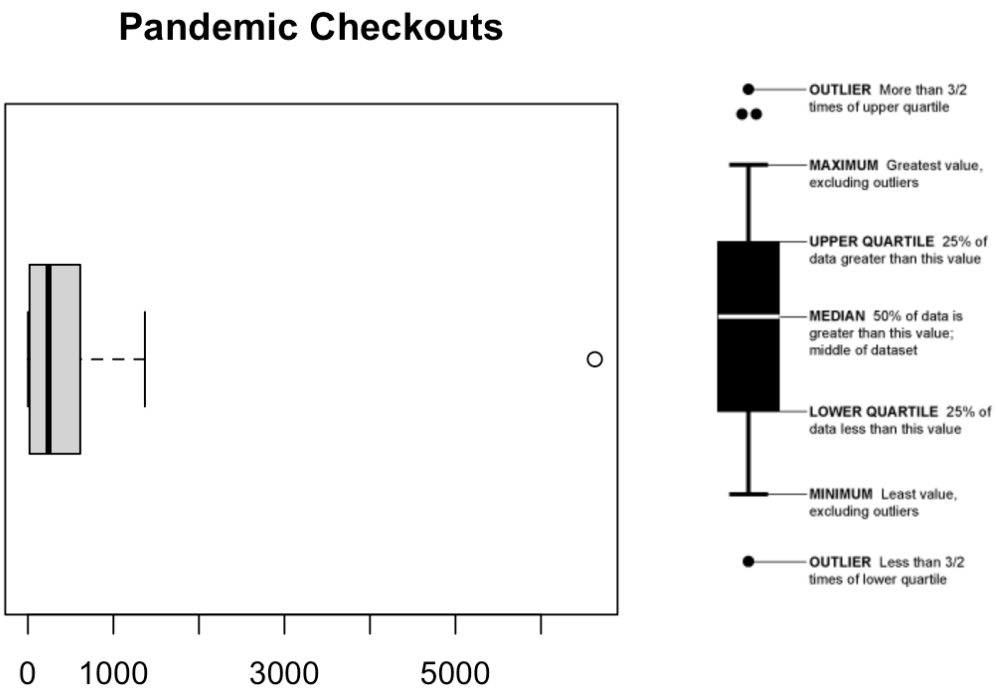
- Pandemic - Week8QueryC.pdf

Though initially I ran two queries that were categorized by week and day respectively, I ultimately decided to use the latter. I also made the decision to focus on the year 2020, the start of the pandemic in order to examine the outliers that arise from viewing such database and seeing if such outliers have an influence on a future machine learning regression model. Having created the csv file, I will now move forward using R because of its expansive statistical library. The goal will be to run some diagnostics to detect outliers and to see if such outliers actually have an influence in making future predictions.



The first procedure is to plot the values conventionally as shown above. We can see visually the cluster of checkouts on the left hand side before reaching an interesting peak below the 100th day threshold (this will be interesting to explore moving forward). Then, as expected, we have a noticeable decrease in values right after, where many days total zero checkouts. We expect to see this due to the pandemic restrictions, so these low values are no surprise. Then, towards the end

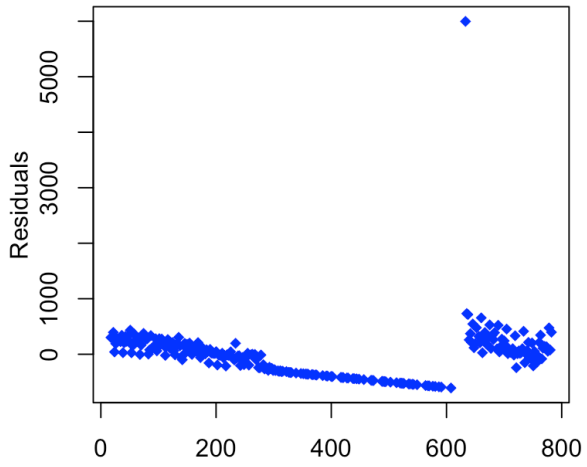
of the year we see a steady rise in checkouts with no sudden or abrupt peaks. My next approach is to create a boxplot to get a quick summary of the data.



We can visually see the minimum to maximum points of the dataset and an outlier which accounts for more than 3/2 times of upper quantile. We can extract the data point to see exactly when this occurs. While running more code in R, I am able to see that the point corresponds to observation number 69 and represents a total of 6630 checkouts in a single day which is over 6,000 above the mean. Even if we don't include the data points during the obvious months of the pandemic, this particular observation still intensely exceeds the mean of the data by thousands!

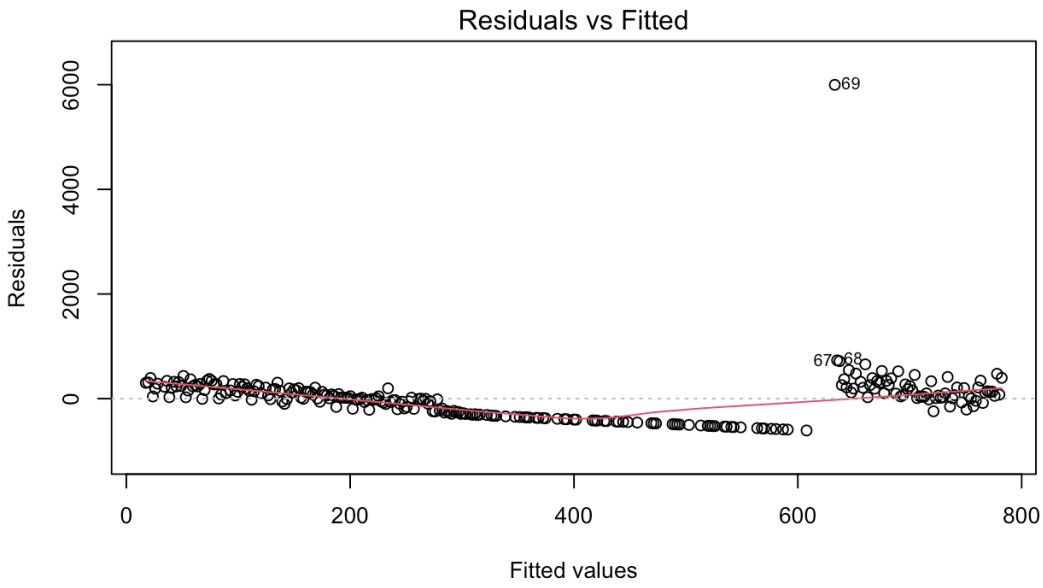
```
   Checkouts
Min.   :   0
1st Qu.:  19
Median : 241
Mean   : 371
3rd Qu.: 612
Max.   :6630
```

Now that we have been able to detect the outlier, we can run further statistical analysis to see if this outlier influences the greater data set to the point that if further models were made, we'd need to exclude such observation. I will first look into graphing the fitted points (the optimal points for a regression model) versus the residuals (how far the actual observed points deviate from the fitted model). A fitted value is a statistical model's prediction of the mean response value when you input the values of the predictors, factor levels, or components into the model.

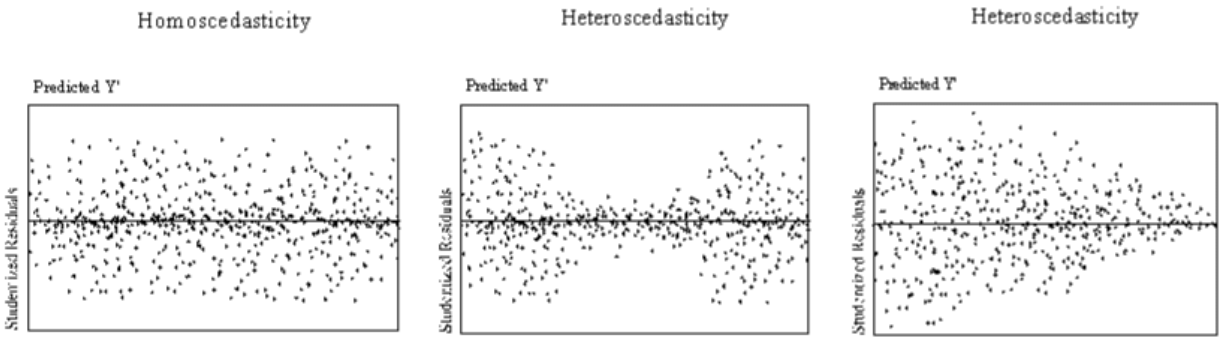


The following is a graphical diagnostic approach to check if the random errors have constant variance by checking homoscedasticity. Homoscedasticity, or homogeneity of variances, is an assumption of equal or similar variances in different groups being compared. This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results. Here is plotted the residuals (vertical axis) against the fitted values (horizontal axis). By looking at the plot we should observe constant symmetrical variation (Homoscedasticity) in the vertical

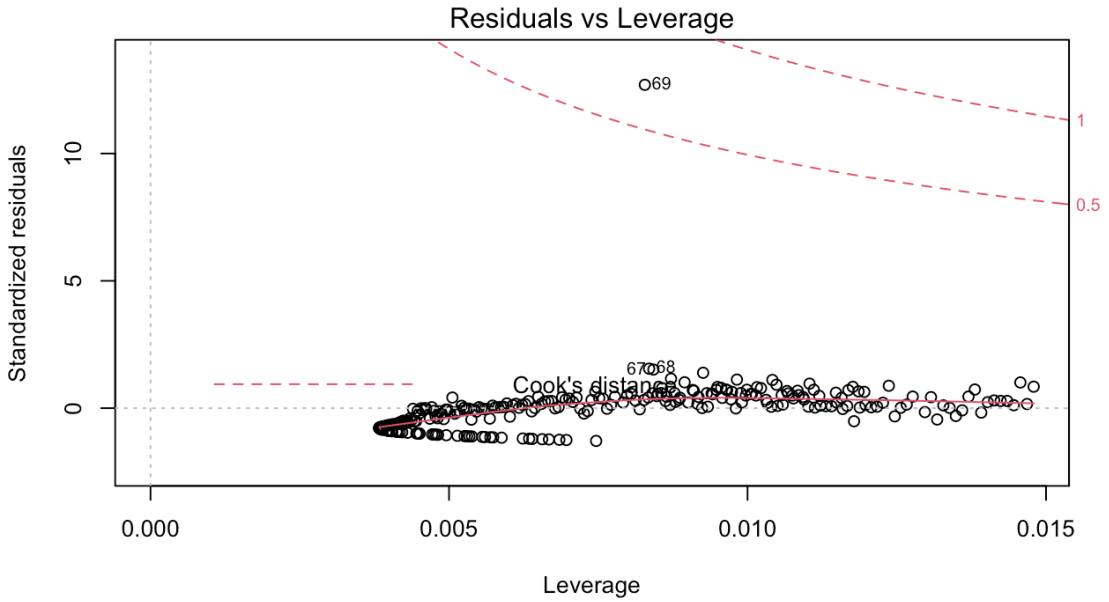
direction, but we don't necessarily in this case. The values don't seem to be evenly distributed across, especially once we get further across to the right and we notice a change in concentration of values (including the outlier). This plot is much more related to heteroscedasticity than homoscedasticity (therefore, there is not a constant variance).



Here is a demonstration to show you exactly what sort of distribution I'm looking for and how I am conducting my analysis:



Next, I will be conducting an analysis on “Cook’s distance”. Cook’s distance is used in Regression Analysis to find influential outliers in a set of predictor variables. In other words, it’s a way to identify points that negatively affect your regression model. The measurement is a combination of each observation’s leverage and residual values; the higher the leverage and residuals, the higher the Cook’s distance.

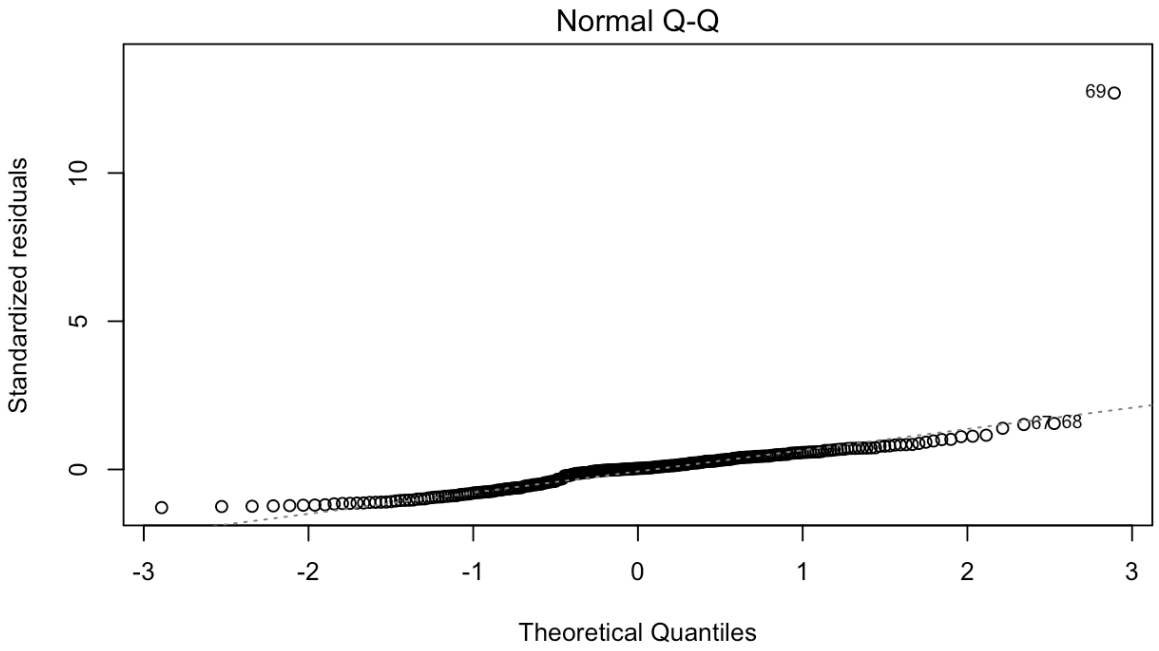


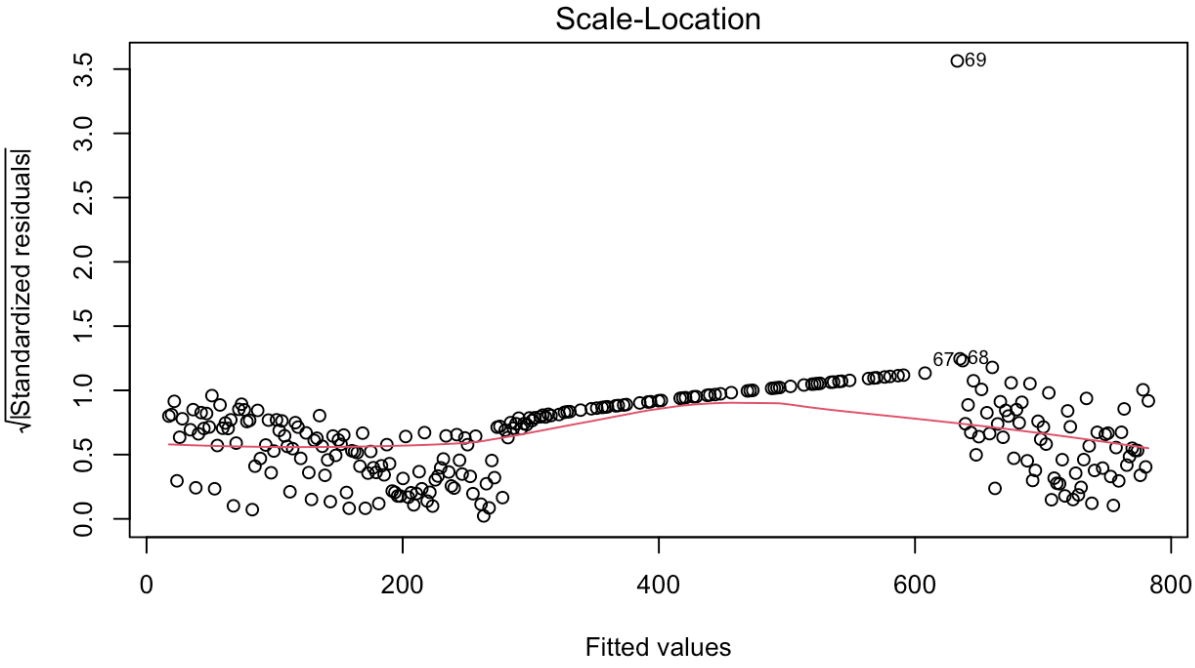
Visually, we can see that multiple observations lie below the border of Cook's distance (the bottom dashed line). This means that (visually) there exists an overly influential point(s) in the dataset. However, we can check this mathematically as well in order to extract the exact point that does so. This requires further coding in R until I finally receive the information that despite how it visually looks, there is only one single point that mathematically has an overly influential effect on the dataset. This would entail the point in between the top dotted lines, which refers

to observation 69. Therefore, not only is this observation an outlier, it negatively affects our regression model and the data set as a whole.

Next, my analysis will focus on the quantiles of the dataset to check for normality amongst our residuals, which again will verify for if the outlier will affect any further statistical analysis. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. From the Q-Q Plot, we can see that most of the residuals of the regression model actually follow the diagonal line very well! This was actually surprising to me! In previous weeks, I assumed data from the pandemic wouldn't be useful because of a lack of normality, but this diagnostic has proved me wrong. Overall, we can assume the residuals are normally distributed with the caveat of the 69th observation. The Scale-Location plot should have a roughly horizontal (zero slope) line/distribution, to which it comes close to but doesn't exactly execute.

We can also see that the red line for Scale-Location isn't exactly horizontal across the plot, but it doesn't deviate too wildly at any point. We would likely declare that the assumption of equal variance is not violated in this case (but can be improved).





After all this statistical analysis, we can definitively conclude that the outlier is observation 69, which refers to day 73 with a total of 6630 checkouts! Not only this, but we have also statistically proved that this outlier has a negative influence on the model/dataset. If we were to choose to continue to analyze this data set or use it for any other means, **it would ultimately be advisable to remove this outlier**.

I will finish off by returning to the MySQL Database to extract data on which books exactly contributed to the outlier found throughout all this analysis.
The 73rd day of 2020 corresponds to March 13th, 2020; this timeline matches that of the pandemic as on March 12th 2020, the city of Seattle announces that it will temporarily close all library locations.

For March 13th, 2020:

```
SELECT
 title, count(title) AS Count
FROM
 spl_2016.outraw
where DATE(cout) = '2020-03-13'
AND deweyClass > 700 AND deweyClass < 800
AND  itemtype like '%bk'
GROUP BY title
ORDER BY Count desc;
```

CSV RESULTS (*click* on file link below):

- Week8QueryD - Week8QueryD.pdf

For March 12th, 2020 (one day before outlier):

```
SELECT
 title, count(title) AS Count
FROM
 spl_2016.outraw
where DATE(cout) = '2020-03-12'
AND deweyClass > 700 AND deweyClass < 800
AND  itemtype like '%bk'
GROUP BY title
ORDER BY Count desc;
```

CSV RESULTS (*click* on file link below):

- Week8QueryE - Week8QueryE.pdf

Looking into the actual checkout per title on this specific day, there is not anything specifically strange or out of nature. The highest checkout accumulates to 10 and isn't out of the ordinary as the following checkouts are around the same range and decrease steadily. I've included the checkouts for the previous day to compare the actual differences between the two, which indeed showcases how though the individual checkouts within the scope of March 13th don't stand out

amongst each other, *together* they act as an anomaly when compared to the rest of the year.

City of Seattle to Temporarily Close All Library Locations, Community Centers to Public to Prevent Further Spread of COVID-19

release date: 03/12/2020

March 13th, 2020: (The outlier) VS March 12th, 2020 (Not an outlier)

title	Count
Walking Seattle 35 tours of the Jet Citys parks l…	10
PelÃƒÂ© the king of soccer	10
Urban trails Seattle Shoreline Renton Kent Vash…	9
My favorite thing is monsters Book one	8
Bone Vol 2 The great cow race	8
Read learn create The ocean craft book	7
Soccer	7
Fly Guy presents castles	7
Captain America Sam Wilson 1 Not my Captain…	7
Lightsaber battles	7
Star Wars the last Jedi incredible cross sections	7
Junior maker experiments to try crafts to create…	7
Out of the box	6
Bone Vol 8 Treasure hunters	6
Bossypants	6
What is a droid	6
unbeatable Squirrel Girl Vol 1 Squirrel power	6
Scott Pilgrim 3 Scott Pilgrim the infinite sadness	6
Bone Vol 3 Eyes of the storm	6
Bone Vol 6 Old mans cave	6
Anyas ghost	6
Hey kiddo	6
Spider man Velocity	6
Edgar an autobiography	6

title	Count
I spy mystery a book of picture riddles	6
Spill zone 1	4
Star Wars the rise of Skywalker the visual dictio…	4
Spill zone 2 The broken vow	3
Simpsons comics colossal compendium Volume…	3
Just kids	3
plain Janes	3
ultimate guide to paper airplanes 35 amazing st…	3
boy who became a dragon a Bruce Lee story	3
kids book of hand lettering 20 lessons and proje…	3
Through the woods	3
This one summer	3
Minecraft for beginners	3
Open book	3
Born a crime stories from a South African childh…	3
Ashs quest the essential guidebook	3
Spider man Velocity	3
Are you my mother a comic drama	3
Guinness world records Gamers edition 2019	3
Crafty gifts	3
My favorite thing is monsters Book one	2
Archie Volume two	2
Art sparks draw paint make and get creative wit…	2
I spy a book of picture riddles	2

III. Conclusion

After conducting a multitude of statistical analysis relating to Data Science, I was ultimately able to find an anomaly/outlier within the dataset pertaining to a specific dewey category. Throughout this experiment, I was able to test for

homoscedasticity, leverage, and overall variance as it is an important assumption of parametric statistical tests do so because models are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results. After conducting this analysis, I was able to find that the 69th observation, which pertains to the 73rd day of 2020 with a total of 6630 checkouts, was an **immense outlier** within the entirety of the dataset, even excluding the data from post-lockdown! Not only was it an outlier, I was able to statistically prove that it had a negative influential effect on the overall model of the data and if I were to conduct further statistical analysis or create anything from this data set, it would be advised to exclude such a point moving forward. After these tests, I returned to MySQL to extract the exact titles that contributed to the outlier in the original database, but came to find out there was not a single title that particularly stood out amongst its peers. Though the individual checkouts within the scope of March 13th don't stand out amongst each other, *together* they act as an anomaly when compared to the rest of the year. It is the drastic accumulation of these items that added up to the grand total of 6630, which was *thousands* of checkouts more than the mean or any other day within that year. Overall, this week's topic was incredibly interesting to me and though it required a lot of work in terms of statistical analysis, it has proven to be my favorite so far in this course.

R Code:

```
library(googlesheets4)
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)

#Reads data into R
df<-
read_sheet('https://docs.google.com/spreadsheets/d/1uoqr3K6F57uFSoXmsj8XFuV7bfqy-pC6yUHr7nkwkJU/edit?usp=sharing')
df

#Summary
summary(df)

#Boxplot
boxplot(df$`Checkouts`, horizontal=TRUE, main="Pandemic Checkouts")

#Extract outlier
which(df$`Checkouts`>2000)

#Extract observation outlier refers to
df[69,]

#Regression model
model <- lm(`Checkouts`~`Day` , data = df)

# fitted response values
fitted_values <- fitted(model)
head(fitted_values)

# residuals
residual_values <- residuals(model)
head(residual_values)

# plot residuals
plot(fitted(model), residuals(model), xlab="",
    ylab="", col="blue", pch=18);mtext(side=2, text="Residuals", line=2)

# find hatvalues of model
hatv<- hatvalues(model)
```

```
# cook's distance
cook1<- (r1^2/dim(X)[2])* hatv/(1-hatv) ## Using formula
cook2<- cooks.distance(model) ##Built-in function
cook2
n <- nrow(df)
h = 4/n
sum(cook2 > h)

# plot diagnosis of model
par(mfrow=c(2,2))
plot(model)

# plot of regression model
plot(`Checkouts` ~ `Day`, data = df,
    main = "Plot with fitted values")
abline(model, col = "Red")
```

IV. References

[1]https://www.spl.org/about-us/news-releases/city-of-seattle-to-temporarily-close-all-library-locations-community-centers-to-public-to-prevent-further-spread-of-covid-19

[2]https://data.library.virginia.edu/understanding-q-q-plots/

[3]https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html#:~:text=Cook's%20D%20Bar%20Plot,-Bar%20Plot%20of&text=Cook's%20distance%20was%20introduced%20by,y%20value%20of%20the%20observation.