Week 10: Future Predictions

Natalia DuBon

I. Abstract

For most of this course, we primarily focused on analyzing data from the past and keeping it there. For this project, I thought I could focus more on future predictions as another means of exploring a topic that hasn't been assigned yet. During week 5, I did a similar project focusing on trends, but this time I plan on solely focusing on prediction using week 8's data set (outliers) with some tweaking. The goal is to focus on the future versus the past.

II. Query Exploration

As stated previously, I will be using last week's query to focus on predictions. The following query uses the dewey class numbers between 700 and 800 that pertain to a book itemtype that have all been checked out during the year 2020, the height of the pandemic.

```
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

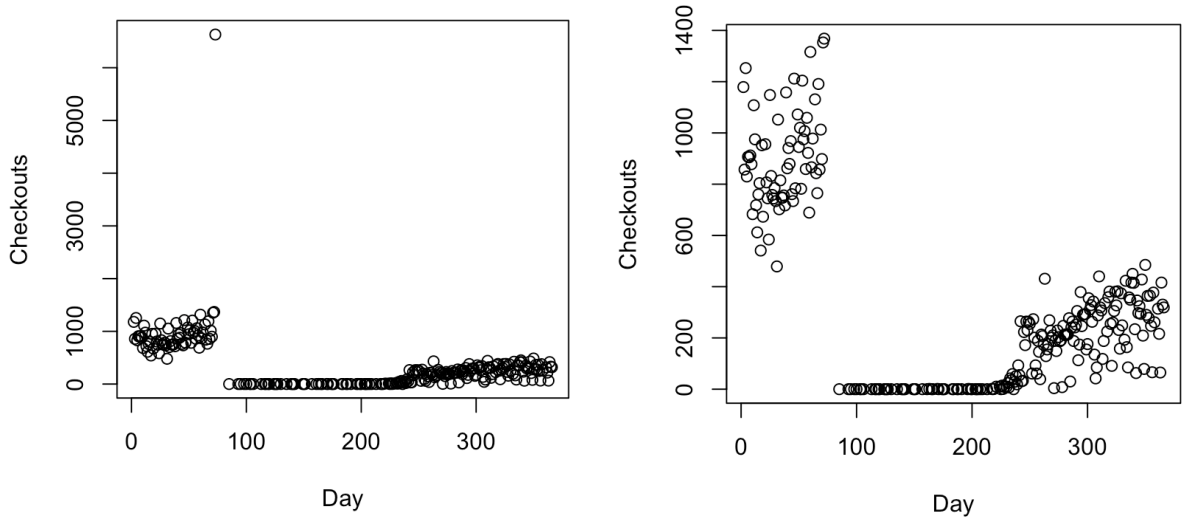
Last week, we were able to conduct a very thorough analysis on this data set to conclude that the 69th observation was an outlier and refers to day 73 with a total of 6630 checkouts. It is heavily advised to remove this outlier if we are to make a prediction model. Why? Because, as we proved consistently last week, this single outlier has a significant negative influence on the rest of the dataset. Therefore, future predictions would be heavily impacted by this single outlier and would likely produce less accurate results (hence, why the point is considered an "outlier" in the first place). Therefore, our next course of action is to remove this point from the data set and start a regression model.

Before (with outlier) VS **After (without outlier)**

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

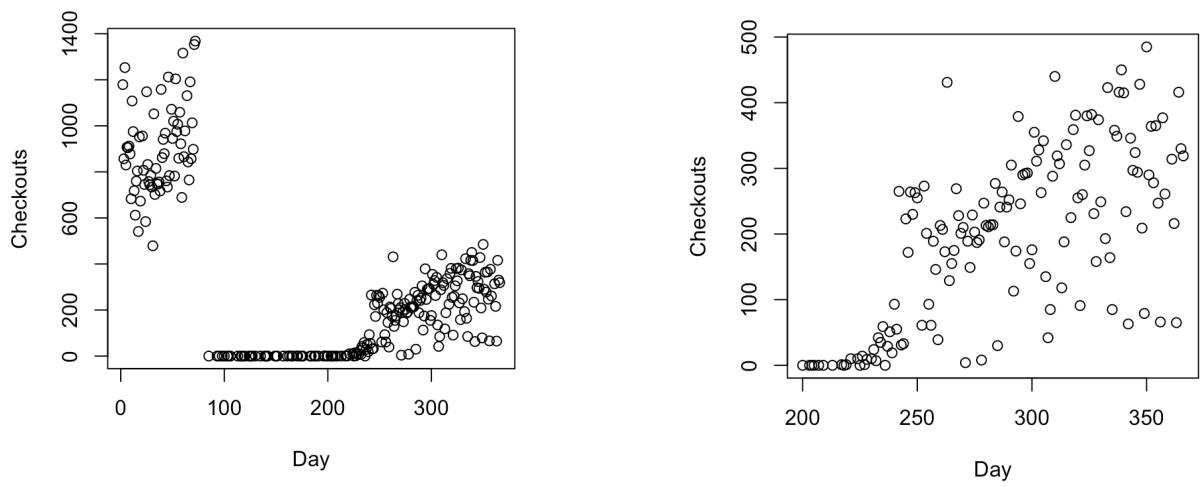
```
  Checkouts
Min.   :   0.0
1st Qu.:  17.8
Median : 237.5
Mean   : 347.2
3rd Qu.: 591.0
Max.   :1368.0
```

Before (with outlier) VS **After (without outlier)**



After plotting the previous and newly updated data set, we can see that the latter is much more easily visible. Since we have two cluster areas, I've also made the conscious decision to remove the left handed side. We can see a linear pattern on the right hand side, so that will make for a much more accurate model, post-pandemic. Therefore the next step is to make these adjustments and finally fit a linear model for the updated data set. For simplification purposes, I chose to make a subset of the data from day 200 forward (roughly leaving out some of the repeated zeros and the left handed cluster).

Before (including lockdown) VS **After (without lockdown)**



The right plot now appears as a zoomed in version of the original data set we had. Visually, we can see a linear pattern and can now create a prediction model. Normally, we would run statistical analysis on this final data set, but that is not the point of this week's assignment. The point is to predict the future outcomes. After creating the prediction model, I went back to MySQL in order to pull data from 2021 regarding the same criteria as the first query. This is the data I will use to compare my prediction results to test for accuracy.

```
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) = 2021
GROUP BY DAYOFYEAR(cout)
ORDER BY DAYOFYEAR(cout) ;
```

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

2021 Dataset - Week10_QueryA.pdf

Returning back to R, I ran a summary report of my regression model, just to view the residual standard error. The residual standard error (highlighted in blue below)

refers to the average amount that the response variable will deviate from the true regression line. In simpler words, it provides a leeway of error for our predictions, in this case 97.4 above or below the predicted value.

```
Call:
lm(formula = Checkouts ~ Day, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-279.86  -60.44   -4.45   69.17  283.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -371.685     51.907   -7.16  3.6e-11 ***
Day            1.974      0.178   11.09  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97.4 on 146 degrees of freedom
Multiple R-squared:  0.457,     Adjusted R-squared:  0.454
F-statistic:  123 on 1 and 146 DF,  p-value: <2e-16
```

Let’s see if we can predict the checkouts for the first day of 2021, just one day after our 2020 dataset ends. This point would refer to day 367 for a continuous timeline (where the first day of 2021 is simply added to the previous 366 days prior).

```
> new <- data.frame(Day=c(367))
> predict(model, newdata=new)
    1
352.8
```

Day	Checkouts
2	267

Subtracting and adding the residual standard error (97.4), there should be about 254.6 to 450.2 checkouts for the first day of 2021. Looking at the dataset for 2021, where we can see the actual observed values, indeed there are 267 checkouts that day and that falls perfectly between our boundary points. Let's test another random day using the seed function.

```
> new2 <- data.frame(Day=c(417))
> predict(model, newdata=new2)
    1
451.4
```

Day	Checkouts
52	465

For the 52nd day of 2021, our model predicted that there would be 451.4 checkouts with upper and lower boundaries of 548.8 and 354, respectively. We can see that our model was incredibly close to the actual observed checkout, which is great news!

```
> new3 <- data.frame(Day=c(569))
> predict(model, newdata=new3)
    1
751.5
```

Day	Checkouts
204	799

Choosing a random day towards the end of 2021, we can see that once again our prediction model proves its accuracy. We predicted that there would be about 751.5 books checked out (again with a 97.4 leeway), and indeed this was very close to actual observed value!

III. Conclusion

Overall, we can see that future predictions can be made after analyzing past results. Though I used data from 2020 to predict values in 2021 (in order to test accuracy) and just for demonstration for this assignment, this prediction model could be used for years ahead of our own right now. If we use data from 2022, we could predict the future values for 2023. This can be useful in tracking foot traffic and predicting overall popularity of several titles.

IV. References

[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

V. Code in R

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

```
#Reads data into R
df<-
read_sheet('https://docs.google.com/spreadsheets/d/1uoqr3K6F57uFSoXmsj8XFu
V7bfqy-pC6yUHr7nkwkJU/edit?usp=sharing')
df

#Remove 69th observation
df1 <- df[-c(69), ]
df1

#Summary
summary(df)
summary(df1)

#Plots
plot(df)
plot(df1)

#Create final dataset
df2 <- subset(df, `Day` >= 200)
df2

#Final Plot
plot(df2)
abline(lm(`Checkouts`~`Day`, data = df2))

#Regression Model
model <- lm(`Checkouts`~`Day` , data = df2)
summary(model)

#Prediction Model
new <- data.frame(Day=c(367))
predict(model, newdata=new)

new2 <- data.frame(Day=c(417))
predict(model, newdata=new2)

new3 <- data.frame(Day=c(569))
predict(model, newdata=new3)
```