Brianna Griffin
MAT 265

Week 8: Outlier Assignment

I. Introduction: For this assignment, I queried for outliers in two different groups of data from the SPL database: checkouts of rock CDs and horror DVDs. It is valuable to detect and find outliers within a data set because these observations differ significantly from the majority. They have a heavy impact on statistics like the average and the standard deviation which we commonly rely on to explain large sets of data. Also, detecting outliers can lead to finding anomalies or problems within the database which are important to catch. In my analysis, I found outliers by assuming normality of the data and looking for data that was outside of three standard deviations from the mean in both directions. This led to interesting results and conclusions.

II. (1) Queries & Analysis:

A. I began by looking at checkouts for rock style CDs at the Seattle Public Library. I wrote a query that searched for the number of checkouts per distinct rock CD. Then, I found the average and standard deviation for the number of checkouts for rock CD's at SPL. Once these two statistics were computed, I was able to distinguish outliers by querying for observations that were outside of three standard deviations from the mean.

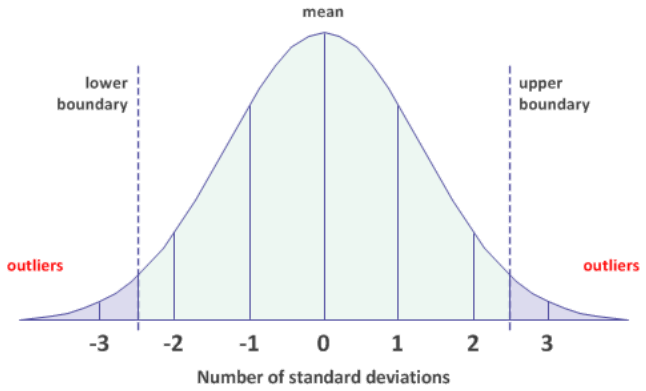
B. Query:

```
SELECT
  *
FROM
  (SELECT
    title,
      num_CO,
      CASE
        WHEN num_CO < (average - (3 * standarddev)) THEN 'outlier'
        WHEN num_CO > (average + (3 * standarddev)) THEN 'outlier'
        ELSE 'not outlier'
      END label
  FROM
    (SELECT DISTINCT
    (title), COUNT(title) AS num_CO
  FROM
    (SELECT
    A.title, A.bibNumber, B.subject
  FROM
    spl_2016.outraw AS A
```

```
INNER JOIN spl_2016.subject AS B ON A.bibNumber = B.bibNumber
WHERE
   A.itemtype LIKE '%cd'
      AND B.subject LIKE '%rock%') sub1
GROUP BY 1) sub2
CROSS JOIN (SELECT
   AVG(num_CO) AS average, STDDEV(num_CO) AS standarddev
FROM
   (SELECT DISTINCT
   (title), COUNT(title) AS num_CO
FROM
   (SELECT
   A.title, A.bibNumber, B.subject
FROM
   spl_2016.outraw AS A
INNER JOIN spl_2016.subject AS B ON A.bibNumber = B.bibNumber
WHERE
   A.itemtype LIKE '%cd'
      AND B.subject LIKE '%rock%') sub5
GROUP BY 1) sub6) sub4) sub7
WHERE
   label = 'outlier';
```

C. Output: CSV table

1. There are 109 outliers out of the 15,489 distinct rock CDs that have been checked out at the SPL over time. All 109 of the outliers were greater than the average value + 3*standard deviation. In other words, their checkouts were distinctively higher than the rest of the rock CDs. Visually, an interpretation of the outliers looks like this.



The outliers that I found are outliers because they differ significantly from the rest of the observations. Since they are each greater than the “upper

boundary", this means that they are significantly greater than the other observations. Put into the context of our dataset, they have a much higher number of checkouts than the rest of the rock CDs. They are and have been the most popular rock CDs to be checked out at the SPL.

- One of the outliers is 19 by Adele. I would not consider this rock, but interesting that it was flagged as this in the data set. Adele is considered as a soulful pop singer.

19

Adele, 1988-

(351 ratings) R

Music CD, 2008

- Here is another one of the dvd that is an outlier:

← Search results

Fleet Foxes

Fleet Foxes (Musical group)

(86 ratings) Rate this

Music CD, 2008 | Downloadable Music

In-library use only

This one is a rock album in fact.

III. (2) Queries & Analysis:

A. Now, I will use a similar query to look at checkouts of Horror movies at SPL. There have been a total of 1,423 of these DVDs at the SPL checked out over time. This number is high enough to assume a normal distribution. Therefore, we can find outliers as before by looking for observations that lie outside of 3 standard deviations away from the mean as before.

B. Query:

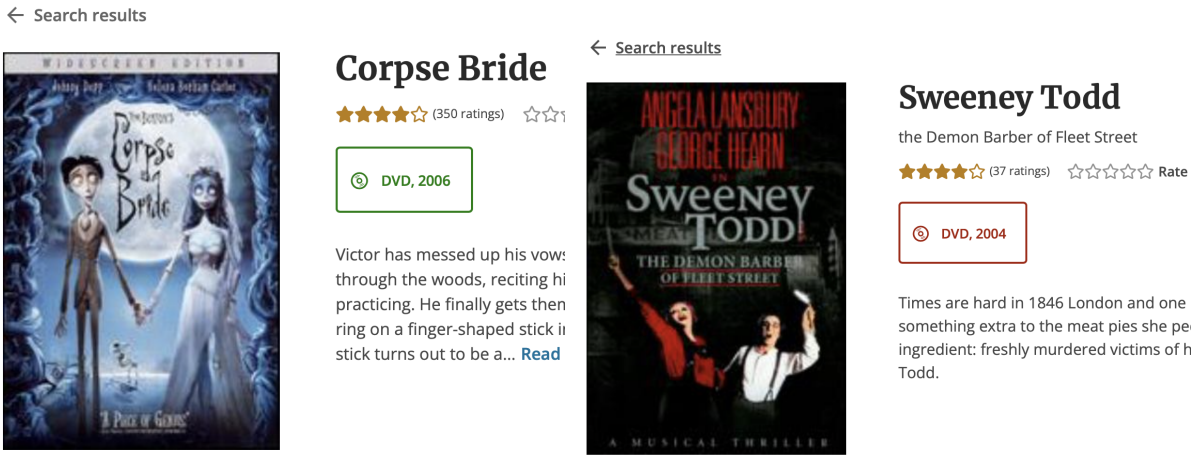
```
SELECT
  *
FROM
  (SELECT
    title,
```

```
        num_CO,
        CASE
            WHEN num_CO < (average - (3 * standarddev)) THEN 'outlier'
            WHEN num_CO > (average + (3 * standarddev)) THEN 'outlier'
            ELSE 'not outlier'
        END label
    FROM
      (SELECT DISTINCT
      (title), COUNT(title) AS num_CO
    FROM
      (SELECT
      A.title, A.bibNumber, B.subject
    FROM
      spl_2016.outraw AS A
    INNER JOIN spl_2016.subject AS B ON A.bibNumber = B.bibNumber
    WHERE
      A.itemtype LIKE '%dvd'
        AND B.subject LIKE '%horror%') sub1
    GROUP BY 1) sub2
    CROSS JOIN (SELECT
      AVG(num_CO) AS average, STDDEV(num_CO) AS standarddev
    FROM
      (SELECT DISTINCT
      (title), COUNT(title) AS num_CO
    FROM
      (SELECT
      A.title, A.bibNumber, B.subject
    FROM
      spl_2016.outraw AS A
    INNER JOIN spl_2016.subject AS B ON A.bibNumber = B.bibNumber
    WHERE
      A.itemtype LIKE '%dvd'
        AND B.subject LIKE '%horror%') sub5
    GROUP BY 1) sub6) sub4) sub7
  WHERE
    label = 'outlier';
```

C. Output: CSV output

It was found that there are 34 outliers from this category. This is smaller than in the previous query. Also, for this sample we have that the average number of checkouts per horror DVD is

710.5 and the standard deviation is **948.1**. Therefore, from the output CSV it is seen that the outliers are mixtures of upper and lower outliers. Putting this into the context of the dataset, some of the outliers had a lot more checkouts than the majority, while others were checked out way less frequently than the majority of horror DVDs.



The movies Corpse Bride and Sweeney Todd are 2 outliers that stand out. Corpse Bride is a famous stop motion Tim Burton movie featuring Johnny Depp and Emily Watson. The movie Sweeney Todd is also a Tim Burton movie featuring Johnny Depp as well. (It seems that Tim Burton movies are extraordinarily popular at SPL!) It won an Academy Award for Best Production Design explaining why it was an outlier in this data set for checkouts at the SPL.

IV. Conclusion: In conclusion, this assignment pushed me significantly in building MySQL queries. I used a lot of the functions that I learned in prior weeks to build a query that determined outliers from the SPL database. This included classifying statistics in the query such as the average and standard deviation followed by a lot of subqueries and joins. The end result was successful in determining outliers in the way that I wanted it to mathematically! Since I looked at the number of checkouts for both of my examples, I was able to verify some of the outliers in each. They both made sense as there were examples in the output table that were very substantially popular in society after doing some outside research. I found that it is really important to consider all of the values of the dataset when you are trying to find patterns and generalizations. This includes finding and maybe removing outliers.

V. References:

A. https://sql-snippets.count.co/t/identify-outliers-in-bigquery/169
B. https://seattle.bibliocommons.com/dashboard
C. Outlier — Why is it important? - Towards Data Sciencehttps://towardsdatascience.com › outlier-why-is-it-impo…