Brianna Griffin
MAT 265

Random Sampling Assignment

I. Introduction:

A. I will take a look at a random sampling of the check-ins and checkouts at the Seattle Public library. I will randomly select the month of 2% of the returns from the SPL filtering by 4 distinct item types. Following, I will compare the results of the queries and identify any possible patterns that arise.

II. Queries & Analysis

The Output CSV are all linked below the code for the SQL queries.

A. Check-ins

1. Books

```
SELECT
  month_return, COUNT(month_return) AS freq
FROM
  (SELECT
    MONTH(cin) AS month_return
  FROM
    spl_2016.inraw
  WHERE
    RAND() < .002 AND itemtype LIKE '%bk'
  ORDER BY RAND()) sub
GROUP BY 1
ORDER BY 1 ASC;
```

Output CSV - Book CIN

The query is showing the number of books with checkouts in each month. The overall data that it is pulling from is 2% of the overall check ins at the SPL. It is seen that the month with the maximum number of checkins is July. The month with the minimum number of returns is February. There are no significant outliers or differences in the frequencies throughout the months, though.

2. CDs

```
SELECT
  month_return, COUNT(month_return) AS freq
FROM
  (SELECT
    MONTH(cin) AS month_return
  FROM
    spl_2016.inraw
  WHERE
    RAND() < .002 AND itemtype LIKE '%cd'
```

```
   ORDER BY RAND()) sub
GROUP BY 1
ORDER BY 1 ASC;
```

Output CSV - Check In CDs

From the output of the above query it is seen that the month with the maximum number of returns is March. The month with the minimum number is September. Once again this table is produced from the same logic as before and random sampling.

3. DVDs

```
SELECT
   month_return, COUNT(month_return) AS freq
FROM
   (SELECT
      MONTH(cin) AS month_return
   FROM
      spl_2016.inraw
   WHERE
      RAND() < .002 AND itemtype LIKE '%dvd'
   ORDER BY RAND()) sub
GROUP BY 1
ORDER BY 1 ASC;
```

Output CSV - check in DVD

Once doing the random sampling, the month with the maximum number of checkouts is March. The month with the minimum number of checkouts is September.

4. Records

```
SELECT
   month_return, COUNT(month_return) AS freq
FROM
   (SELECT
      MONTH(cin) AS month_return
   FROM
      spl_2016.inraw
   WHERE
      RAND() < .002 AND itemtype LIKE '%rec
   ORDER BY RAND()) sub
GROUP BY 1
ORDER BY 1 ASC;
```

Output CSV - record check in

From the random sampling of 5% of checkouts at the SPL it is seen that November has the highest number of checkouts and February, March, June, and July all have the minimum number of checkouts. This data output has a very low frequency as records are not checked out that frequently. This is most likely due to the fact that there are not many at the SPL.

5. Video (VHS)

```
SELECT
   month_return, COUNT(month_return) AS freq
FROM
   (SELECT
      MONTH(cin) AS month_return
   FROM
      spl_2016.inraw
   WHERE
      RAND() < .002 AND itemtype LIKE '%vhs
   ORDER BY RAND()) sub
GROUP BY 1
ORDER BY 1 ASC;
```

[Output CSV - check out_VHS]

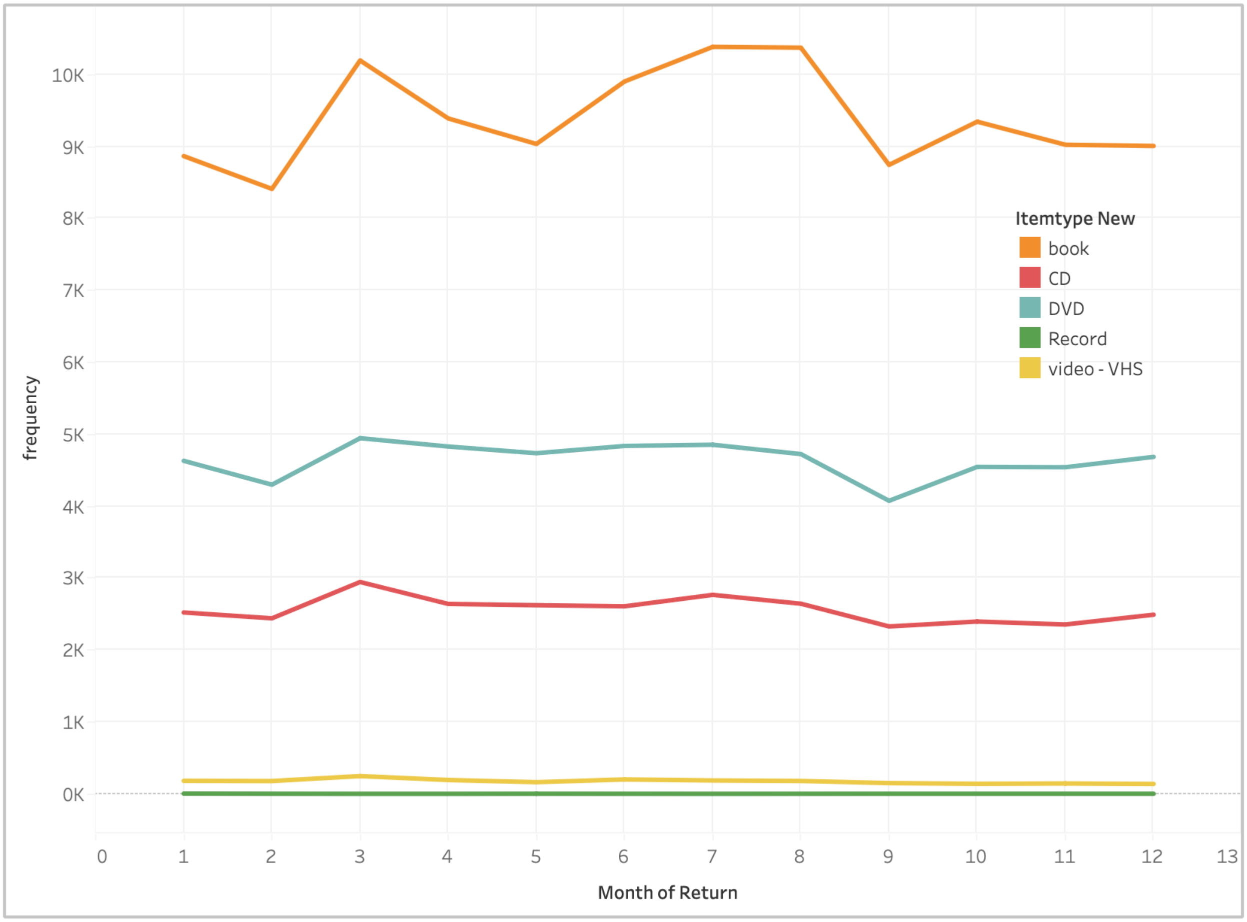
For the random sampling of VHS checkouts at the SPL we see that the highest number of checkouts are in March. The fewest number of checkouts are in October.

6. Check In - Comparison

```
SELECT
   CASE
      WHEN itemtype LIKE '%bk' THEN 'book'
      WHEN itemtype LIKE '%cd' THEN 'CD'
      WHEN itemtype LIKE '%dvd' THEN 'DVD'
      WHEN itemtype LIKE '%rec' THEN 'Record'
      WHEN itemtype LIKE '%vhs' THEN 'video - VHS'
      ELSE NULL
   END AS itemtype_new,
   month_return,
   COUNT(itemtype)
FROM
   (SELECT
      itemtype, MONTH(cin) AS month_return
   FROM
      spl_2016.inraw
   WHERE
      RAND() < .002
   ORDER BY RAND()) sub
GROUP BY 1 , 2
ORDER BY 2 ASC;
```

output css - comparison

For the last query, I decided to join the prior ones using a subquery and a Case When statement. Now, we can see the trends for each random sampling by the 5 distinct, different item types. I will graph it visually so that it is easier to interpret.



We can see that 3 out of the 5 chosen itemtypes hit their absolute maximum during March. 4 out 5 of the itemtypes have a relative maximum in March. Also, 3 groups have a minimum during September. It is interesting to see these trends arise while doing random sampling of the check ins. We randomly selected only 5% of the total returns. The random group that was selected for each itemtype ended up following a similar trend which is very fascinating. The only thing that can explain this is randomness.

III. Conclusion

A. In conclusion, this week's assignment allowed me to experiment and learn about SQL's rand() function. I found it to be very useful and interesting. In relation to my queries, I noticed 4 out of the 5 random samples that I chose tended to follow a similar pattern which I did not expect. Records at the SPL did not follow the trend. This was due to their low frequency of returns. They do not have much stock at the libraries I am assuming. I think it is very fascinating how the graph shows that books, CD's and DVD's even follow a very similar curve for the months of checkout. There were infinitely many options for the random samples that I could have queried, but this was the output and results that I randomly received.

IV. References

A. How to Select a Random Sample of Records in MySQL - PHPFog.com

B. https://www.mat.ucsb.edu/~g.legrady/academic/courses/08w259/itemTypes.pdf

C. https://www.w3schools.com/sql/func_sqlserver_rand.asp