Arnav Kumar's Project 1

Initial Concept:

The focus of this analysis is on the distribution of checkouts between electronic and physical books in the field of computer science, which is known for the fact that a wealth of valuable information could be gathered about the field online, especially for free, which is likely due to its groundings in the theme of technology and the Internet in general. Thus, it is of value to attempt to discover whether there is a higher frequency of checkouts for electronic books about computer science as compared to physical books, and even more interesting would be to see the differentials during the COVID-19 pandemic as compared between the two modalities. Due to limitations with the data, as well as the relative lack of popularity of specialized topics in computer science with the general public, this analysis opted to focus on introductory topics in computer science, which is the vast majority of the content of such material in any case.

Process:

Although initial research revealed that the MAT 259 SPL database does not contain records for E-books, further consulting showed that the Socrata data API for Seattle's open dataset on SPL contains such records, albeit the database has a different schema. Thus, queries were formed for both datasets to have outputs that matched in schema as closely as possible. The queries are as shown below:

```
SELECT
   CheckoutYear, CheckoutMonth,
   SUM(Checkouts) as Checkouts
WHERE
   CheckoutYear >= 2019
      AND MaterialType LIKE "%EBOOK"
      AND (Subjects LIKE '%Computer Science'
      OR Subjects LIKE '%Computing'
      OR Subjects LIKE '%Computer'
      OR Subjects LIKE '%Technology'
      OR Subjects LIKE '%Linux'
      OR Subjects LIKE '%Windows'
      OR Subjects LIKE '%Coding'
      OR Subjects LIKE '%Hacking'
      OR Subjects LIKE '%System Administration')
      AND Subjects != ''
GROUP BY CheckoutYear, CheckoutMonth
```

Listing 1: Socrata API Query

This query aggregates checkouts by month and year granularities, and the database query's select clause requires a sum aggregation on the checkouts column due to the database schema design. The where clause includes common computer science keyword filters, especially in an introductory context, on a column describing subject tags for each E-book. It also includes other filters such as %EBOOK to filter for the target items, and a not equals to clause to account for errors in database inputting. The group by clause allows for the actual aggregation by month and year granularity for each book checkout. Following this is the MySQL query executed directly against the MAT 259 database:

```
SELECT
   YEAR(cout) AS year, month(cout) as month, SUM(CASE when 1 = 1 then 1 else 0 end) as
Checkouts
FROM
   spl_2016.outraw
WHERE
   cout > '2019-01-01'
      AND itemtype LIKE '%bk'
      AND deweyClass >= 004
      AND deweyClass <= 006
      AND !( deweyClass = '')
GROUP BY year(cout) , month(cout)
LIMIT 0 , 1000;
```

Listing 2: MySQL Query

Similar to the Socrata API query, this query includes a modified sum aggregation in the select statement to account for the database schema and to produce a clean output as early as possible in the analysis process. The from clause includes the target checkout database, and the where clause includes a deweyClass filter which filters from 004 to 006. The numbers were chosen after careful analysis from preliminary exploration of the database to try to match the schema and theme of the Socrata API query as closely as possible. The group by clauses filter on the same variables, and the limit expression is a result of the limit setting set in MySQL workbench.

Cleaning:

The Socrata Data API query results were directly in memory due to calling the API from code, and the MySQL results were exported from csv through pandas and converted to a DataFrame like the Socrata API query results. After online research and code sample adaptation, they were then visualized in two separate graphs, as shown below:

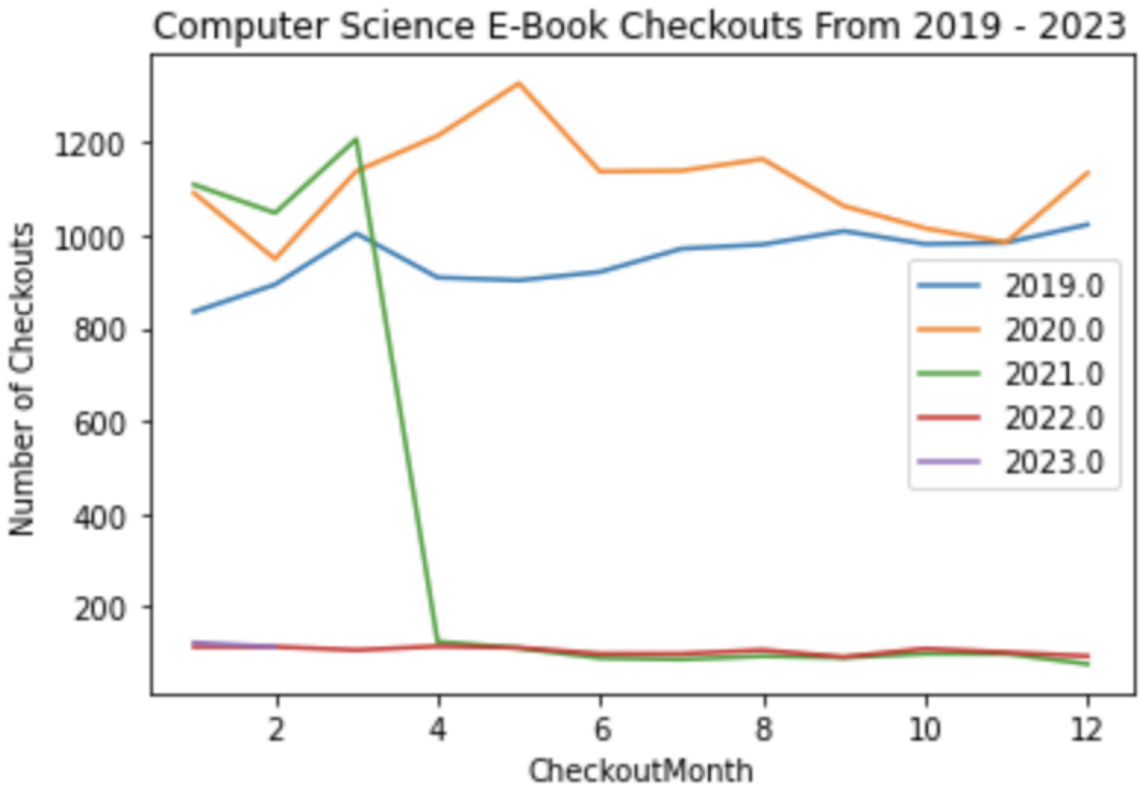

Figure 1: The number of electronic book checkouts against months per year starting from 2019.

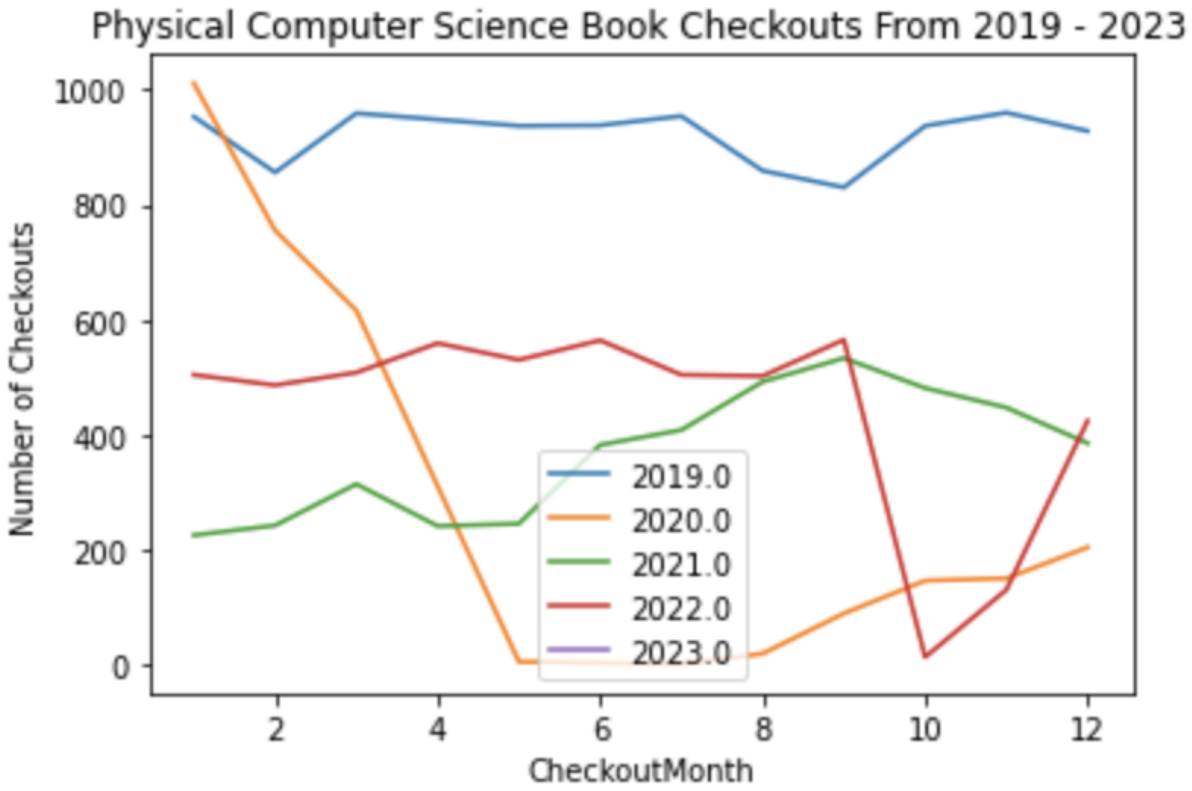

Figure 2: The number of physical book checkouts against months per year starting from 2019.

Although it is clear that electronic books likely remained popular during the start of the pandemic in 2020, with little difference in popularity throughout the year as opposed to the dip around May for physical books, it is more difficult to explain why physical books might have remained consistent and relatively popular as compared to electronic books, which saw a large collapse around April of 2021. Similarly, electronic books had little popularity in general while physical books continued to enjoy relative popularity until around October, after which they quickly rebounded. 2019 seems to be the only year where consistency in popularity, as well as the magnitude of popularity itself, which is reasonable since it is before the pandemic, of whose effects are still felt today. The difference in popularity, in general, is not as significant between the two modalities as was expected, and the unexplained differences as shown in this analysis warrant future queries analyzing these discrepancies.

Notes: The 2023 year was not considered in the analysis, although it appears in the Figure 2 legend, due to the lack of data for the year. All data cleaning after obtaining query results were in Python, using the pandas, matplotlib.pyplot, and csv packages. The sodapy package was used for API abstraction for querying in code for the Socrata Data API for SPL's public dataset.