**Impact of Best-Selling Books on Authors’ Other Works Using Seattle Checkouts Data**

**Jeffrey Cain, Brandon Fisher**

**Objectives**

The purpose of this project was to search for insights about how popular and best-selling books have impact beyond a single title. To examine that impact, an inspection of authors is a good place to start. Does a best-selling book by an author have cause that author to see a significant increase in the popularity of their other works? Building off that foundational question, this project developed more questions, examined data from Seattle checkouts, and visualized the results to better understand what happens when a book becomes incredibly popular.

**Background**

Throughout the project, a number of different tools and technologies were used to pursue the objectives. Notepad++ was a primary tool for writing and editing code, as well as viewing and formatting results data. Pyspark was the primary coding language and data science platform used. To run the Pyspark code, Google Cloud was utilized, specifically the Dataproc and Storage tools. Tableau was used for visualization of the results. Liquid Studio XML was used as a viewing tool for the large file.

The dataset used for this project came from Kaggle, an online source with thousands of free and public datasets. The Seattle checkouts information is found at: <https://www.kaggle.com/city-of-seattle/seattle-checkouts-by-title>. Once downloaded and unzipped, the dataset exists as a 6.48 GB csv file.

**Methodology**

The scope of the project existed in a few main stages. There was first a period of investigating, familiarizing and viewing the large dataset. Next came a phase of processing the data and preparation for the main analysis phase. Third was analysis and phase where results data were compiled. And finally, the results data were then put into a visualization.

**Data Familiarization**

The file presented a few challenges and the first was that it was too large to be opened using the first tools we tried. Microsoft Excel and Notepad++ both failed to open the file because it was so large. To open the file and examine some of the data, Liquid Studio XML, a tool specifically designed to quickly open very large files, was downloaded and used. From there, roughly 1000 rows were copied into a separate file for small scale testing purposes.

This led to the next challenge with the file. It was in a csv format, but some of the fields including commas within the values. A title, or list of subjects, for example could include commas. This is common in a csv format, and quotation marks were used to indicate a field that included a comma in the value. A tool such as Microsoft Excel can handle this formatting automatically and keep the in-value commas when reading a csv file. However, the issue came when attempting to use Pyspark. The split(“,”) function which was used to parallelize the data did not account for quotation marks indicating an in-value comma. This would lead to some lines have many more entries than others after applying the split function. To fix this, a short chunk of python code was used to read the large file in csv format, and then save it in a tsv format with tabs separating the values. This allowed the use of Pyspark with the split(“\t”) function.

**Processing and Preparation**

At this point, there was a large file that could be opened and processed using Pyspark within Google Cloud. To begin, a search was drawn up to find some of the most popular best-selling books within the records. An initial mapping phase was used that first split each line by tabs [split(“\t”)] and specifically assigned index corresponding to the Checkouts count as the key. Next a filtering stage was used to include only records for books that had been checked out. This avoided other data such as audiobooks, movies, CDs, laptops, or anything else that can be checked out. There was additionally different RDDs set up for each year. Finally, a sortByKey(False) was used to put the records in order with the greatest number of checkouts being at the top. And finally a take(10) returned the top ten from each RDD. The table structure meant each row was a book-month, or a book combined with one month of checkouts for that book. So the processing and preparation phase returned the top ten book-months for each year.

At that point, a visual inspection of the top ten from each year led to a list of potential authors to analyze. Not all prospective authors turned out to have enough data to do a meaningful analysis. And, as will be discussed in the summary, a number of other factors besides a single popular book are at play when it comes to an author’s overall popularity. Nonetheless, the authors that made it through final analysis are: Lee Child, Suzanne Collins, Anthony Doerr, Gillian Flynn, Jonathan Franzen, Stieg Larsson, Ann Patchett, and Maria Semple.

**Analysis**

The analysis phase involved separate RDDs for each author that was examined. The mapping began the same as before, with the checkouts as the key. Next, a much more specific filtering phase took place. The process was done twice for each author, once filtering to only the checkouts of the best-selling book or series, and once filtering to all other checkouts except the best-selling book or series. Stieg Larsson was the exception, since his only works were a part of the same popular series.

After the filtering phase had checked the existence, or lack of existence of a specific title, it also filtered to search for the author’s name in the appropriate field. Finally, it filtered to a specific set of years that was most relevant based on when the popular book appeared in the top-10 list accumulated during the processing and preparation phase.

Next, a critical mapping phase reduced the rows such that the key consisted of the year and the month, while the value consisted of the checkouts count. This allowed for the next stage, a reduceByKey(add), which consolidated all the rows in the data set with matching year and month, into a single row with a cumulative checkout total. A final sortByKey() was executed and the data was in a results form.

This process was repeated for all the authors chosen for analysis. Combining the data for an author’s popular book or series along with the data for the rest of their works finished the process of creating results data.

**Visualization**

In order to visualize data, we filtered out all of the extra characters created by pyspark. This turned the file into a comma separated value file, which can be read by Tableau.

Tableau allowed us to visualize the different pieces of data and compare how the popularity of a single book, or series, could affect the checkout rates of the other books by the same author. The data in Tableau was first joined on the key columns that were created in pyspark. These columns were the Year and Month columns. This allows us to create 2 graphs on top of eachother comparing the author’s popular book checkouts with their another book checkouts. the Orange graphs are the books that reached top 10 checked out book for the year. The Blue graphs are the other books by the same author as the orange graph. The following graphs show the information that we had gathered.

**Results**

Thanks to visualizing our data, we were able to find quite a few different situations that occur when an author creates a popular book. For better resolutions on the results data, see the associated slides. Figure 1 is the visualization of Suzanne Collins’ checkouts and the impact of the Hunger Games series. Figure 2 shows Anthony Doerr and the impact of All The Light We Cannot See. Figure 3 shows Ann Patchett and the impact of Commonwealth. Figure 4 shows Gillian Flynn and the impact of Gone Girl. Figure 5 shows Lee Child and the impact of The Midnight Line of the Jack Reacher series. Figure 6 shows Jonathan Franzen and the impact of Freedom. Figure 7 shows Maria Semple and the impact of Where’d You Go, Bernadette. Figure 8 shows Stieg Larsson and the impact of the Millenium series around the time that the movie The Girl With The Dragon Tattoo was released.

One of the more interesting graphs is figure 1. This graph shows how many people checkout out Suzanne Collins Hunger Games books and any of her other books. Her book series stay at a relatively constant checkout numbers, and then by the time her third book in the Hunger Games series is released, there is a massive increase in average book checkouts. This could be accounted for by how popular her series had become and that a movie series was on the way.

Similar to the first figure, figure 4 and figure 2 show the same increase that happened to Suzanne Collins. This increase is more exaggerated because these authors previous books didn’t have high checkout averages. Both authors have a massive increase in book checkout averages once they release their initial top 10 book of the year.

Not every author sees an obvious increase in book checkouts. Authors that already have had popular books, or series in the past have minimal if no obvious increase in book checkouts. Each author receives a minor bump in book checkouts, however since the average book checkouts doesn’t increase as obviously as with figure 2 or figure 4, it is not possible to conclusively say that those popular books had an affect on the author’s other books.

Ultimately, we were able to answer what we were looking for when we first started this project. Authors do see an increase in average book checkouts from the rest of their catalog when one of their books they release gain popularity.

**Summary**

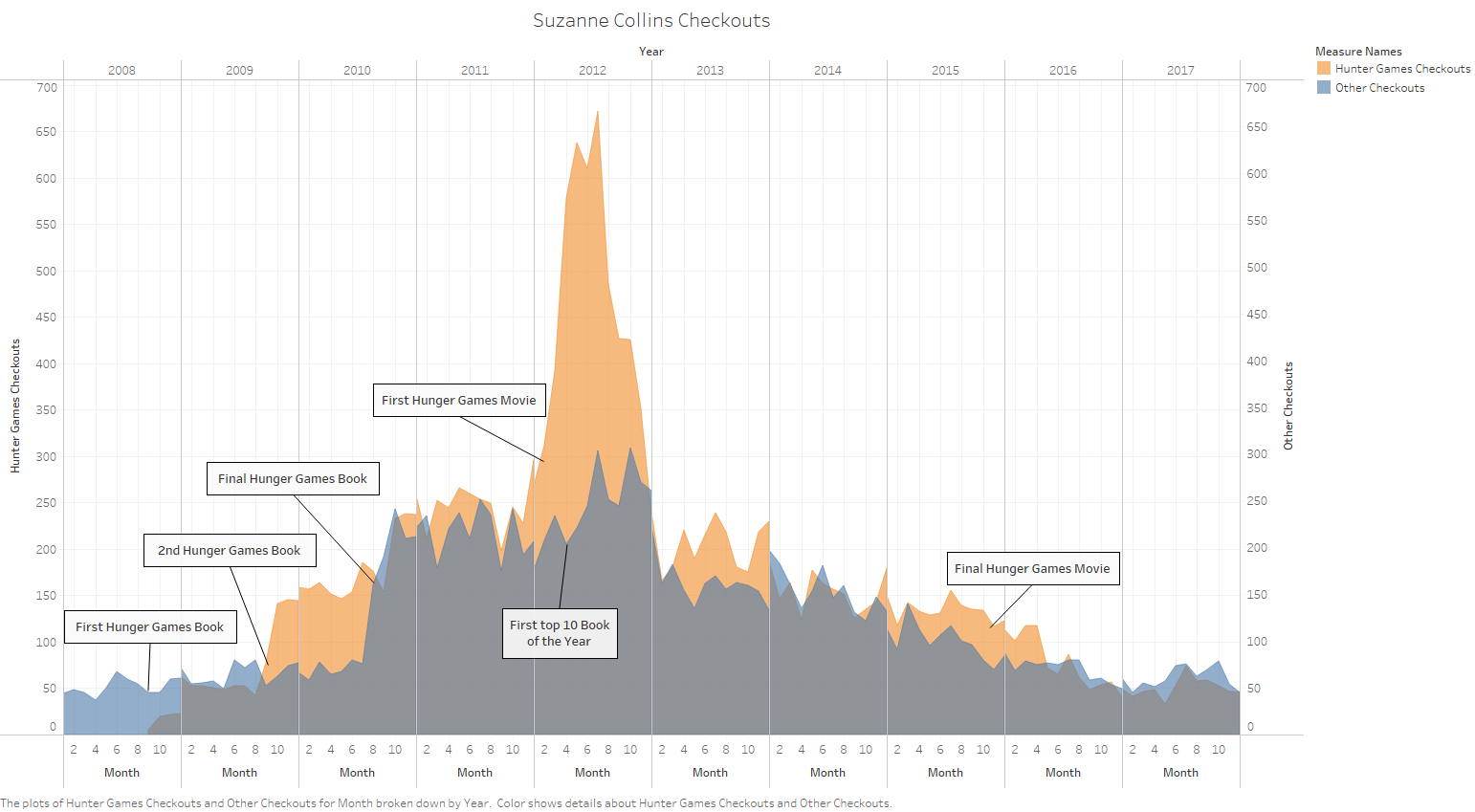
The project originally set out to look at a specific question: Does a best-selling book by an author have cause that author to see a significant increase in the popularity of their other works? In the end, a few interesting case studies were compiled and visualized to examine that impact.

Additionally, a number of other factors were at least discovered and worthy of note. For example, it is reasonable to expect some impact to be different depending on whether the best-selling book was a part of a series or a standalone. Also, how popular and experienced is the author? Are there even a number of other books by that author for readers to checkout? Did a movie come out that could impact the popularity of the story or series? Wherever possible, some of these things were noted in the resulting visualizations. However, this certainly provides a limitation on the ability to draw a conclusion about specifically how much a best-selling book can impact the popularity of an author’s other works.

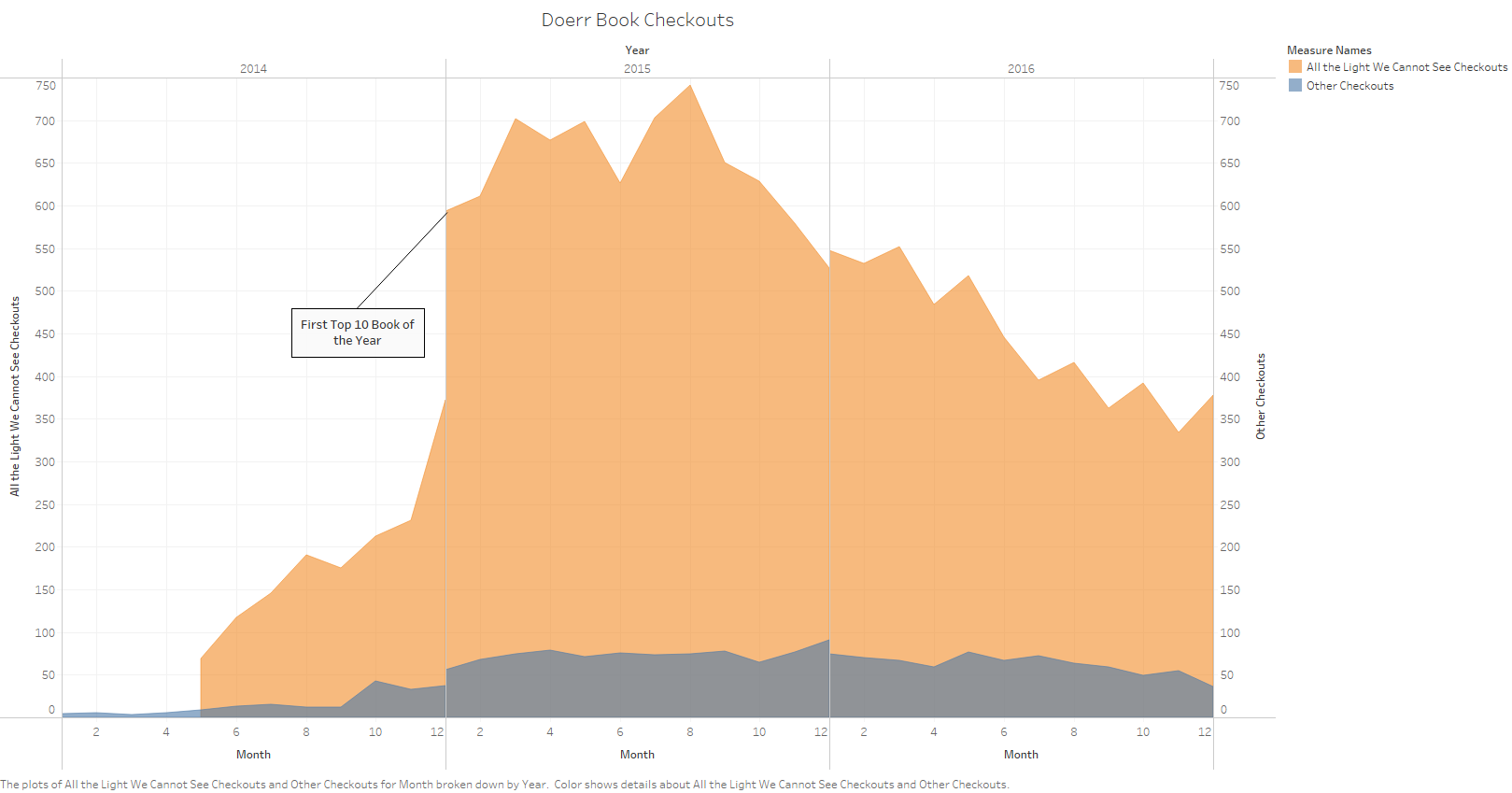
One of the biggest data science takeaways from this project is the importance of good visualizations. Even to experienced computer programmers, the numbers that come out as results are not especially meaningful. Once the data is visualized, it can truly be examined to find meaning or search for areas in need of further investigation. Additionally, the challenges of working with large datasets cannot be overlooked. This project involved dealing with files that are too large to view, in the wrong format, and that a lot of noise that is hard to detect. Each of these provides real challenge, even before the data analysis begins.

In regards to future work, a number of improvements could be made. As the project stands, authors need to be picked out by hand, and the code to analyze each case must also be written by hand. If this process were to be automated, results could be found for potentially hundreds or even thousands of best-selling authors, instead of just a few. Setting out to do this would include a few additional challenges. For example, the dataset regularly contained multiple entries for the same book with a slightly different capitalization scheme. There was also variations in comma and punctuation usage within the title strings. A simple string match would not be an adequate way to automate this process.

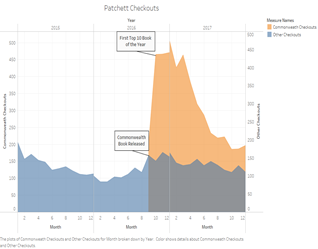
Handling a larger number of authors would present both a challenge and an opportunity regarding the number of other factors that contribute to checkout popularity. It would be challenging for an automated system capable of handling thousands of authors to identify which books are a part of a series, or which ones were adapted into popular movies. But with the ability to take on greater sets of authors, it could become possible to analyze specifically the impact of a movie release, compared to the impact of a new release in a popular series, compared to a best-selling standalone title etc.

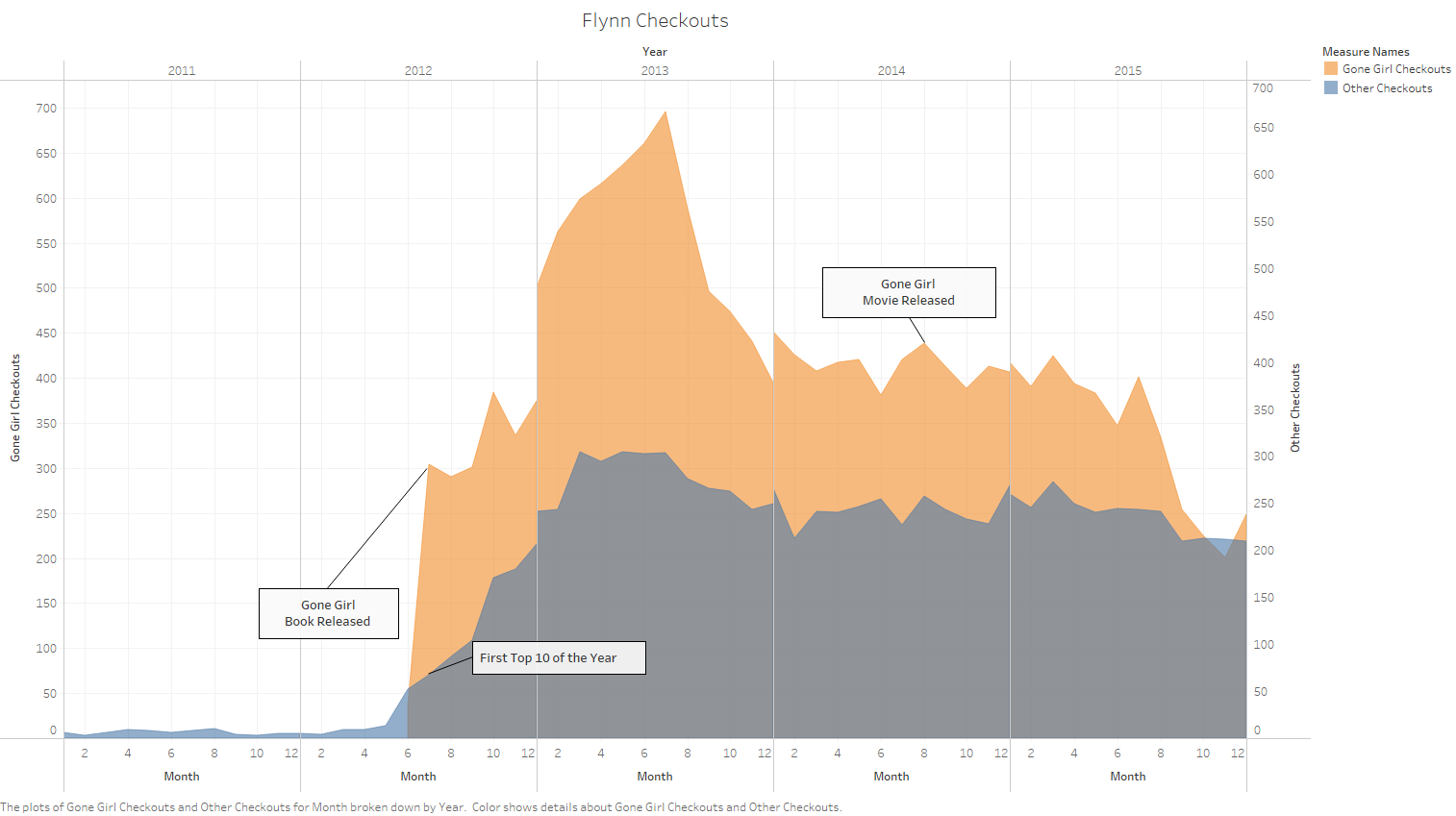
****

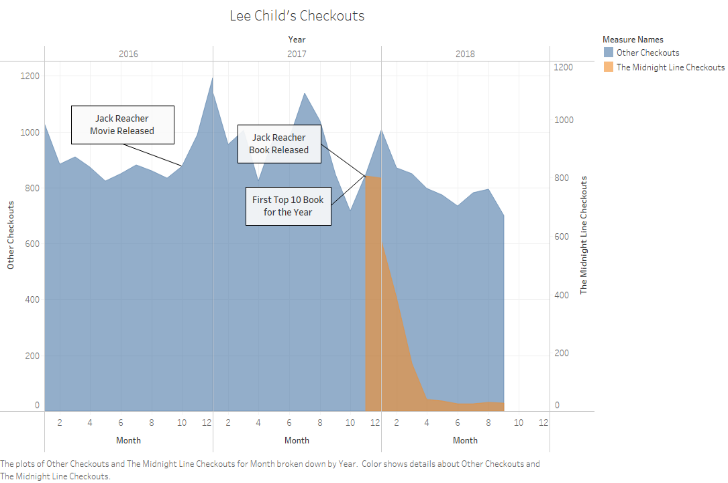
**Fig1:** Collins Checkouts



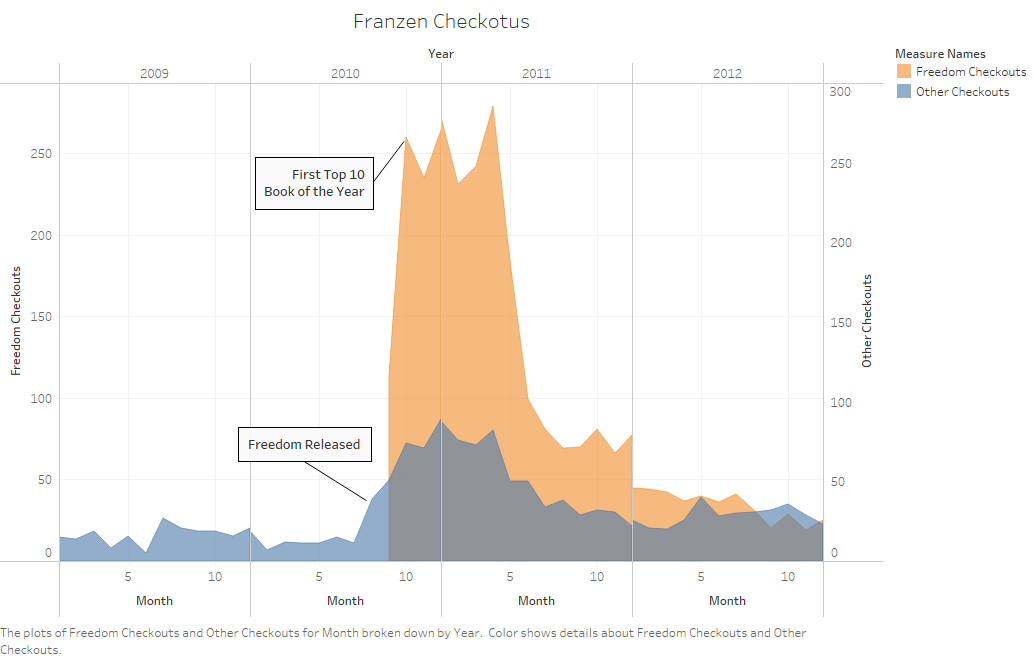
**Fig2**: Doerr Checkouts



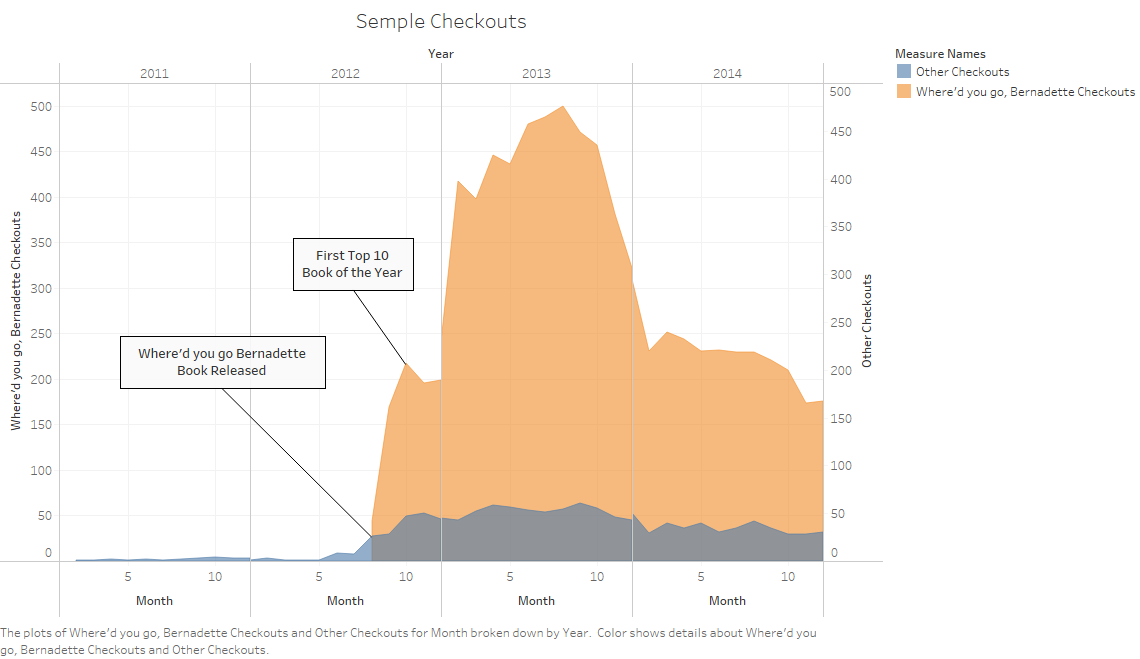
**Fig3:** Patchett Checkouts

**Fig4:** Flynn Checkouts

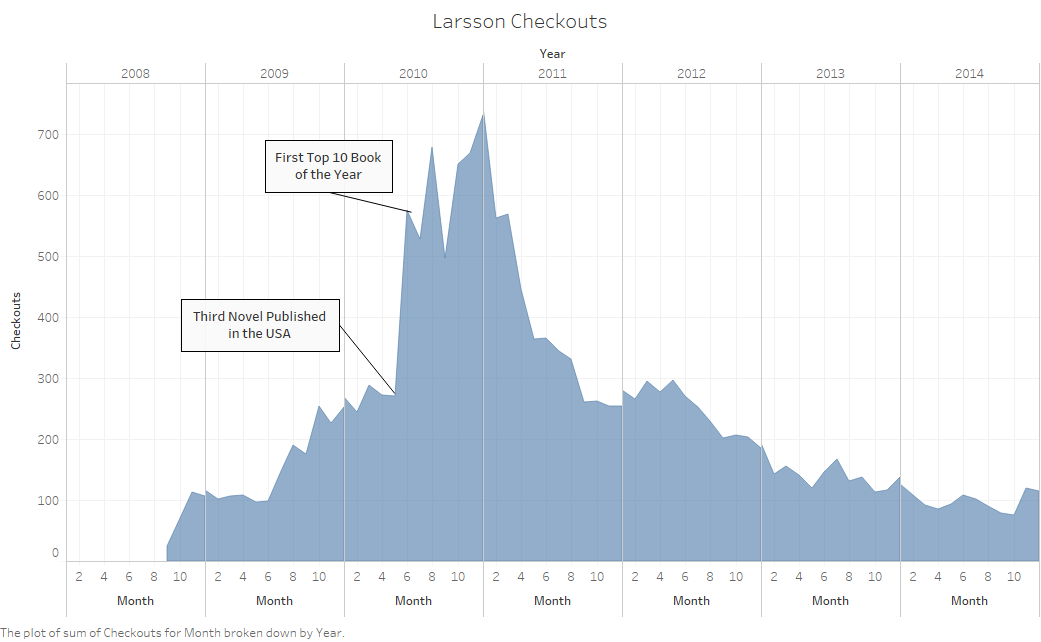
**Fig5:** Lee’s Checkouts



**Fig6**: Franzen Checkouts



**Fig7:** Semple Checkouts

**Fig8**: Larsson Checkouts