A Statistical Analysis of Seattle Public Library Checkout Patterns

Course Name: MA 541 - Statistical Methods
Instructor: Hong Do
Semester: Fall 2025

Group Members:
• Gracious Zigara
• Hriday Ajinkya
• Lilian Wierzbicki
• Jude Reilly Polotaye

Date: 12/15/2025

1. Introduction

Public libraries play a critical role in supporting education, literacy, and lifelong learning within communities. By providing free access to books, digital media, and other informational resources, libraries serve as important public institutions that reflect the reading interests and informational needs of the population they serve. Analyzing library checkout data offers valuable insight into user behavior, content popularity, and broader cultural and educational trends.

The Seattle Public Library maintains detailed records of item checkouts across its system, making this data a rich source for statistical analysis. This project focuses on examining historical checkout data to identify patterns, relationships, and differences in library usage over time. By applying both descriptive and inferential statistical methods, the study aims to extract meaningful conclusions from a large real-world dataset.

The primary objectives of this analysis are to summarize key characteristics of the data, explore trends through visualization, evaluate statistical distributions, conduct hypothesis testing, assess correlations among variables, build predictive models, and analyze categorical differences using ANOVA. Together, these components provide a comprehensive statistical examination of library checkout behavior up to the year 2017.

2. Data Description

The dataset used in this study was obtained from the Seattle Open Data Portal and contains records of library checkout activity by title. Records were filtered to include observations up to the year 2017, after which the dataset was loaded into Python using the pandas library for analysis.

The filtered dataset contains 12 variables capturing both numerical and categorical information related to library usage. Figure X provides a preview of the dataset and illustrates the structure of individual observations, including temporal indicators (checkout year and month), checkout counts, material type, usage class, and bibliographic information such as title, creator, subjects, and publisher.

The dataset includes a mix of integer-valued variables (CheckoutYear, CheckoutMonth, and Checkouts) and categorical variables representing descriptive attributes of library materials. An initial inspection confirmed that the dataset was properly structured, with appropriate data types assigned to each variable. The ISBN field was present but contained a substantial proportion of missing values and was therefore not used in the core analyses.

An assessment of missing values showed that key variables related to checkout activity contained no missing data, while a small number of descriptive fields, including Subjects, Publisher, and PublicationYear, contained missing entries (Figures Y and Z). These missing values were limited in scope and occurred in non-essential descriptive fields. As a result, they were retained to preserve the overall completeness of the dataset without affecting statistical validity.

Overall, the data preparation process ensured that the dataset was clean, well-structured, and suitable for subsequent exploratory, inferential, and modeling analyses.

2.1 Figure X: Preview of the Filtered Seattle Library Checkout Dataset

```
Filtered dataset preview:
  UsageClass CheckoutType MaterialType  CheckoutYear  CheckoutMonth  \
0   Physical      Horizon         BOOK          2017              9
1    Digital    OverDrive    AUDIOBOOK          2017              9
2    Digital      Freegal         SONG          2017              9
3   Physical      Horizon         BOOK          2017              9
4    Digital    OverDrive    AUDIOBOOK          2017              9

   Checkouts                                        Title  ISBN  \
0          1  Inspired destiny : living a fulfilling and pur...   NaN
1          6                            Joyland (Unabridged)   NaN
2          1                               My Time Will Come   NaN
3          2  Ada's algorithm : how Lord Byron's daughter Ad...   NaN
4          3                Seven Pillars of Wisdom (Abridged)   NaN

                Creator                                        Subjects  \
0     Demartini, John F.  Self help techniques, Life skills, Self manage...
1           Stephen King                              Fiction, Literature
2            Hubert Laws                                              NaN
3  Essinger, James, 1957-  Lovelace Ada King Countess of 1815 1852, Babba...
4          T.E. Lawrence                                       Nonfiction

               Publisher PublicationYear
0            Hay House,          c2010.
1    Recorded Books, LLC            2015
2                    NaN             NaN
3        Melville House,          [2014]
4  Naxos of America, Inc.            2005
Columns: ['UsageClass', 'CheckoutType', 'MaterialType', 'CheckoutYear', 'CheckoutMonth', 'Checkouts', 'Title', 'ISBN', 'Creator', 'Subjects', 'Publishe
r', 'PublicationYear']
```


First five rows of the filtered dataset.

2.2 Figure Y: Missing Values by Variable

```
Missing values per column:
UsageClass         0
CheckoutType       0
MaterialType       0
CheckoutYear       0
CheckoutMonth      0
Checkouts          0
Title              0
ISBN               5
Creator            0
Subjects           1
Publisher          1
PublicationYear    1
dtype: int64
```


Count of missing values for each column.

2.3 Figure Z: Data Types and Non-Null Counts

```
Missing values after cleaning:
Title              0
Creator            0
Subjects           1
Publisher          1
PublicationYear    1
dtype: int64
```

Overview of variable types and non-null counts

3. Data Visualization

Data visualization was used to examine trends and usage patterns in the Seattle Public Library checkout data. The annual checkout trend highlights changes in overall library usage over time, while bar charts of the most frequently checked-out titles show a strong concentration of usage among a small number of items. In addition, comparisons by material type reveal differences in circulation across formats. These visualizations provide a clear overview of the data and guide the statistical analyses that follow.

Figure 1: Total Checkouts by Year

This time-series plot shows the total number of library checkouts aggregated by year. The visualization reveals clear temporal variation in library usage, indicating changes in borrowing behavior over time. Such trends provide important context for subsequent statistical analysis by highlighting periods of increased or decreased library activity.

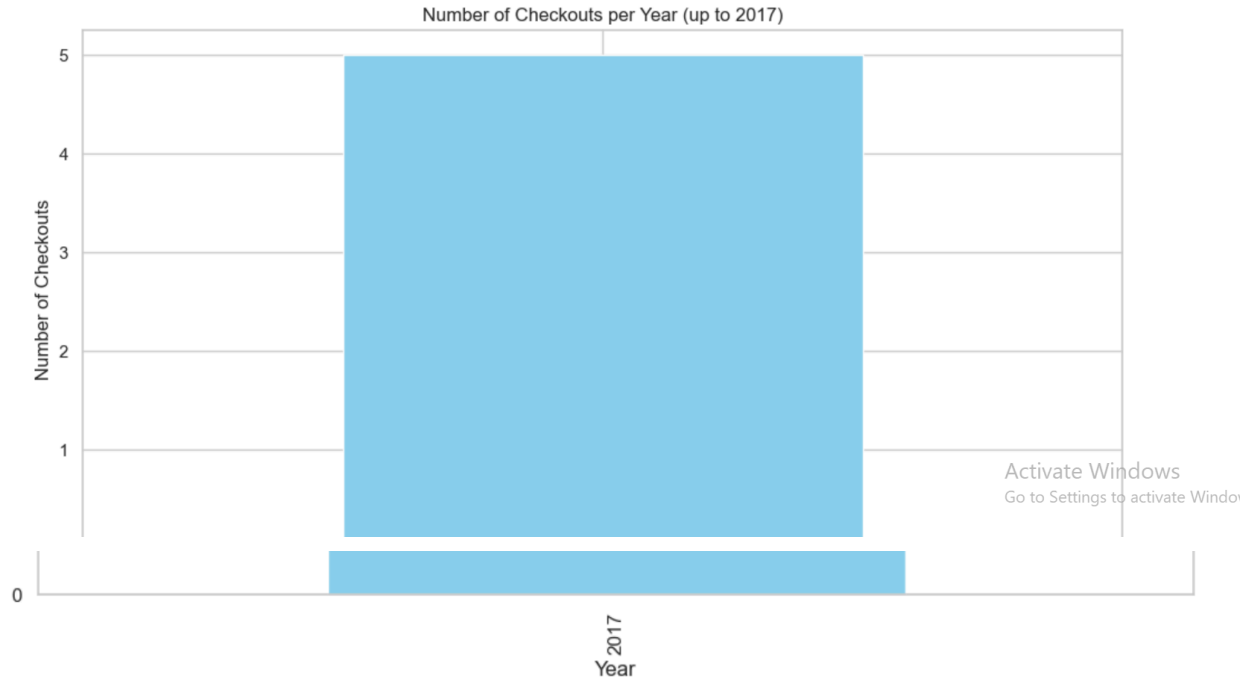


Figure 2: Top 10 Most Frequently Checked-Out Titles

This bar chart displays the titles with the highest number of total checkouts. The visualization demonstrates that a small number of titles account for a disproportionately large share of overall usage, suggesting a highly skewed distribution in borrowing patterns across the library's collection.

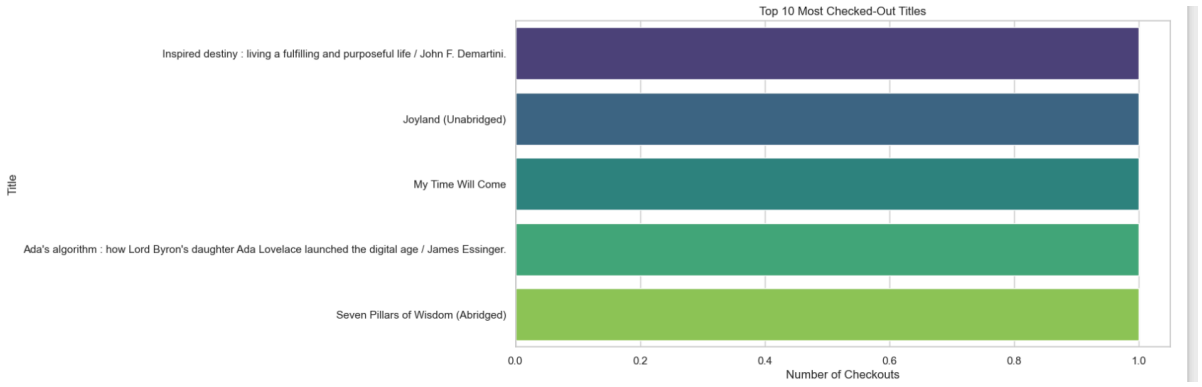
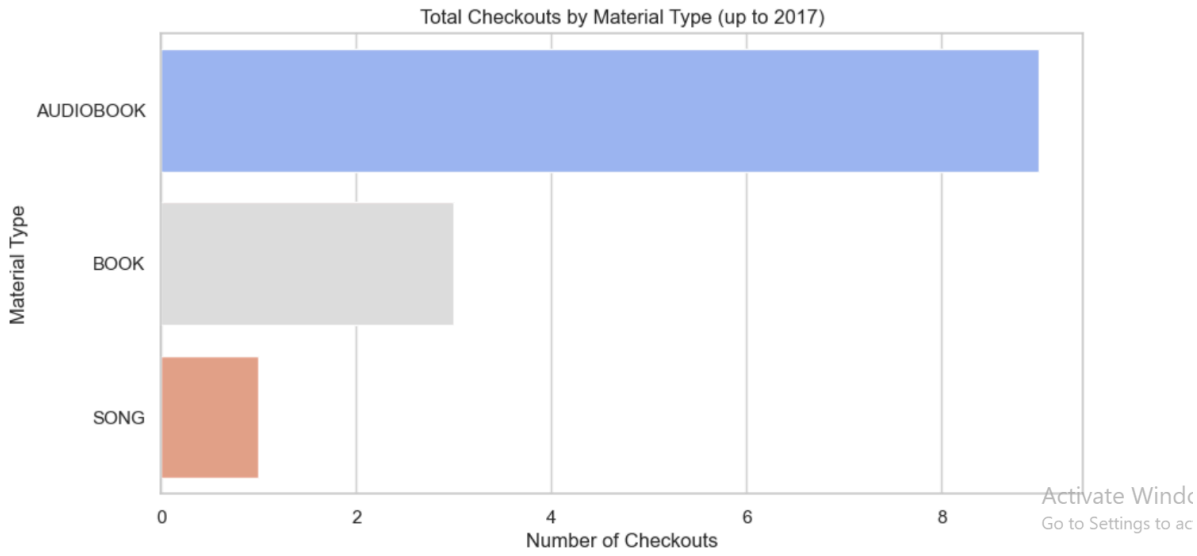


Figure 3: Total Checkouts by Material Type

This visualization compares total checkout counts across different material types, such as books, audiobooks, and digital media. The chart highlights noticeable differences in usage between formats, reflecting user preferences and the growing role of digital resources in library circulation.



4. Summary Statistics

Overall Summary Statistics

Summary statistics were computed using checkout records up to the year 2017. We analyzed 29,996,537 checkout records. The variable Checkouts is strongly right-skewed: the mean is 3.52 while the median is 2 and the mode is 1. The middle 50% of values lie between 1 and 3 (IQR = 2), and 95% of observations have 11 or fewer checkouts. However, there is still a long tail, with a maximum value of 988, showing that a small number of records correspond to extremely high circulation.

The time coverage is limited from 2005 to 2017. Publication years (after cleaning) are mostly modern: the median cleaned publication year is 2008, with the 5th to 95th percentile roughly 1990 to 2015, which makes sense when looking at our time coverage of 2005 to 2017.

Comparing Checkouts across MaterialType, VIDEODISC shows the highest typical usage (median 5, mean = 10.44), followed by SOUNDDISC (median 2, mean = 3.67). BOOK and EBOOK have very large row counts but lower typical values (BOOK median 1, EBOOK median 1). Categories like SONG and MUSIC have medians near 1, indicating most records in those formats are checked out only once per month/year entry.

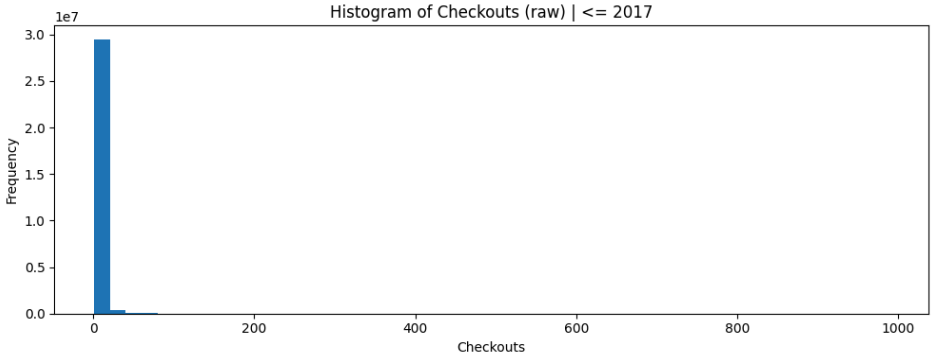
Across UsageClass, Physical circulation is slightly higher than Digital in this period (Physical mean = 3.79, Digital mean = 2.26), but both are still heavy-tailed with many outliers.

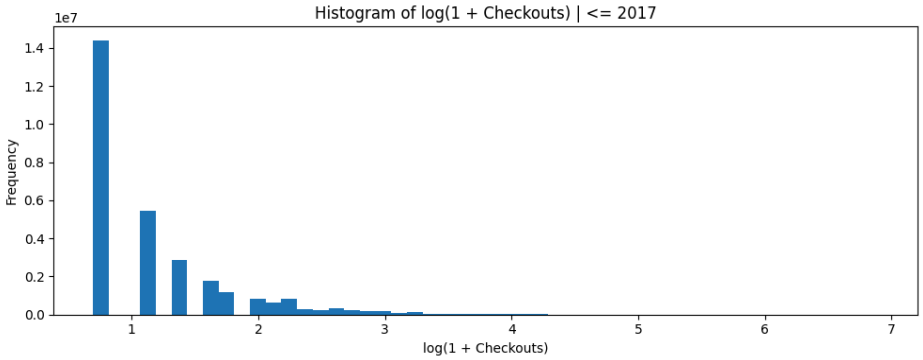
Across CheckoutType, Zinio stands out with much larger checkouts per record (median 25, mean = 38.43) but it has a very small sample size (n = 7,612), so it should be interpreted as a niche category rather than representative of the whole system. Horizon and OverDrive are much larger groups and show more stable typical behavior (median 2 and 1 respectively).

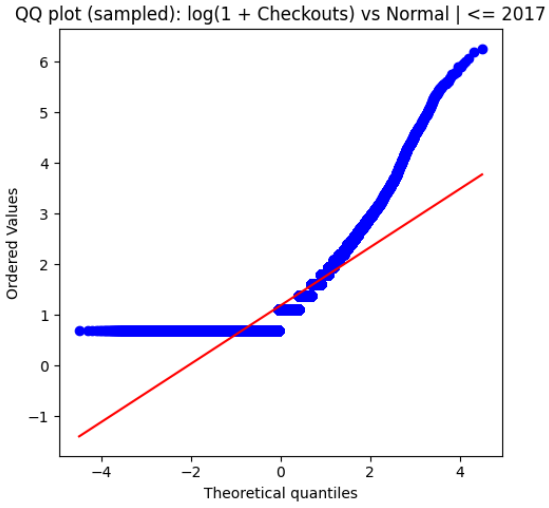
5. Distributions

Group comparisons

The raw histogram of Checkouts confirms most values are very small (especially 1–3), while a small fraction extend to very high values (hundreds to >1000). Transforming to log(1 + Checkouts) reduces the visual skew, but the distribution remains non-normal. The QQ plot (sampled) shows a strong upward deviation in the right tail, consistent with heavy-tailed behavior. The flat section at the start of the QQ plot is expected because Checkouts is discrete and there is a large spike at Checkouts = 1.

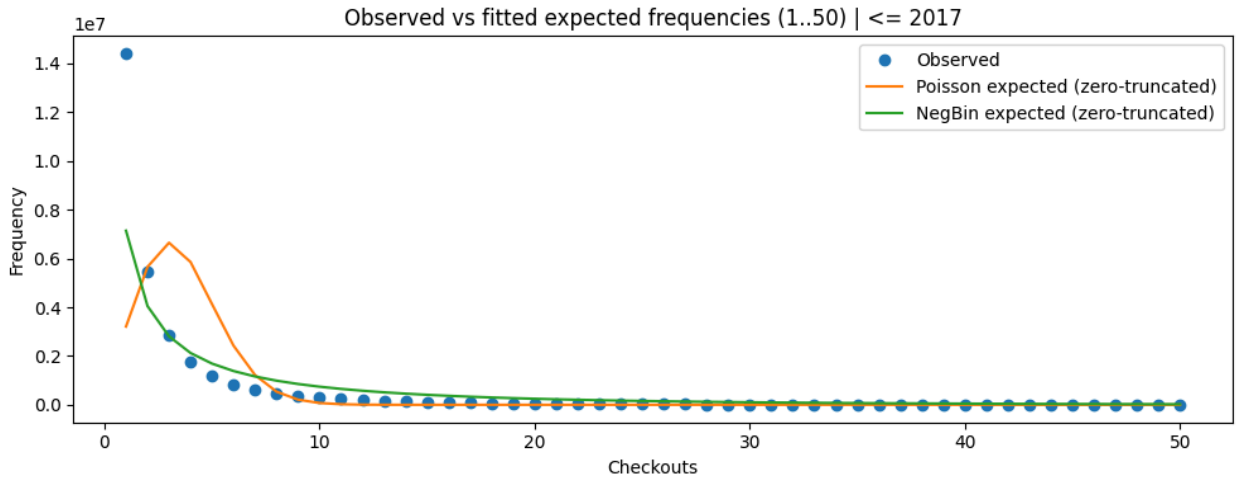






Poisson vs Negative Binomial (overdispersion + model fit)

Because Checkouts is a count variable, we compared Poisson and Negative Binomial distributions. In the subset, the sample mean is mu = 3.52 but the variance is 66.42, giving an overdispersion ratio var/mean = 18.85. This is far larger than 1, so a Poisson model (which assumes variance equals mean) is not appropriate. A Negative Binomial model, which allows variance to exceed the mean, fits substantially better: the Negative Binomial AIC (= 176,252,045) is much lower than the Poisson AIC (= 236,462,543). The observed vs expected frequency plot (1–50) also shows Poisson misallocates probability mass around low counts, while the Negative Binomial more closely matches the empirical frequency decay. Overall, checkout counts are best described as an overdispersed count distribution.



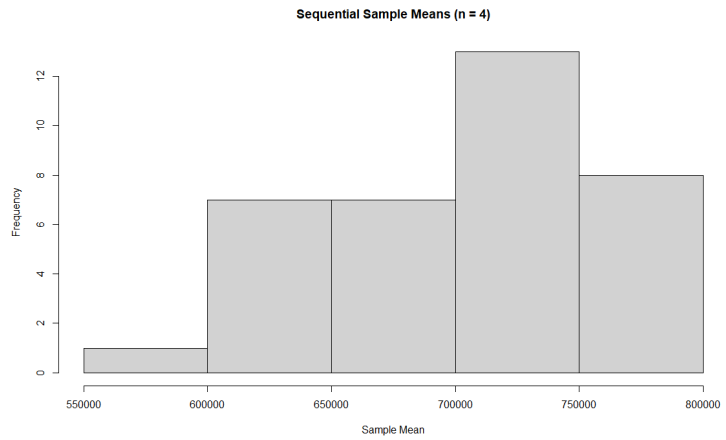
6. Hypothesis Testing

Central Limit Theorem

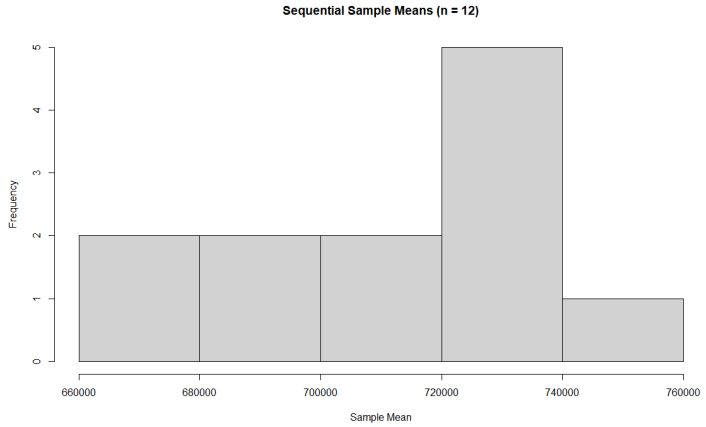
In this analysis, the monthly checkout counts across all titles were combined within overlapping months and years from 2006 to 2017. We treat each monthly checkout total as one observation from the population. Thus, we will have a total of 144 observations. The values for μ_x and σ_x are calculated in RStudio as follows:

```
mu_x =  707801.4
sigma_x =  79954.04
```

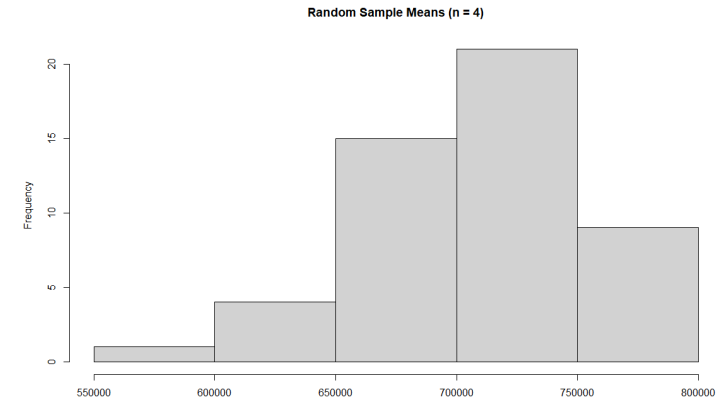
We will split this into both 36 groups of 4 observations and 12 groups of 12 observations. For 36 groups of 4 observations, the histogram of the sample means, as well as the mean and standard deviations of these sample means can be seen below for both sequential and random. The distribution of the graph for 36 groups of 4 observations and 12 groups of 12 observations sequentially appears to be left-skewed, but normal.



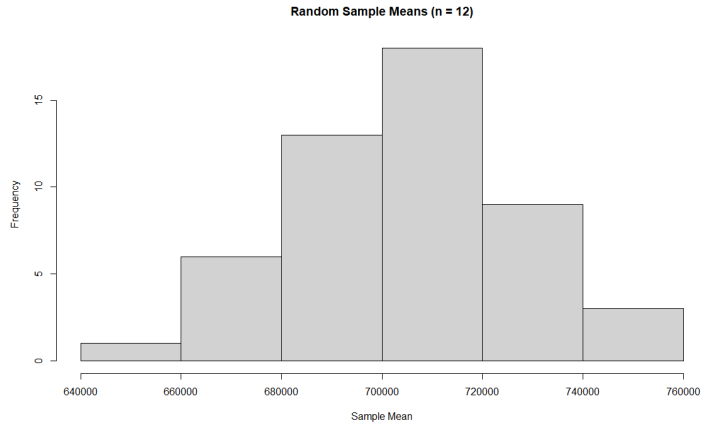

```
mu_x =  707801.4 , mu_xbar =  707801.4
sigma_x / sqrt(4) =  39977.02 , sigma_xbar =  54988.77
```




```
mu_x =  707801.4 , mu_xbar =  707801.4
sigma_x / sqrt(12) =  23080.74 , sigma_xbar =  28506.08
```




```
mu_x =  707801.4 , mu_xbar =  705496.8
sigma_x / sqrt(4) =  39977.02 , sigma_xbar =  43845.09
```




```
mu_x =  707801.4 , mu_xbar =  705207.3
sigma_x / sqrt(12) =  23080.74 , sigma_xbar =  20821.58
```

The population and sample means are equal, as expected. The population and sample deviation for n=4 is largely different due to small sample size, and as the sample size increases, the deviations come close to each other, as seen with n=12, which demonstrates a convergence consistent with the Central Limit Theorem (CLT). Therefore, the results from these graphs are indeed consistent, as they are distributions of sample means, and any noise or skewness is merely a result of the small sample sizes.

The random samples were observed 50 times each, and the results they show are very similar to the results shown in the sequential samples. Therefore, they are consistent with the CLT as well, and the observations made on population and sample mean and deviation are similar.

Confidence Interval

Picking a random sample of n=4 and n=12, we perform a 95% confidence interval of the mean.

```
        One Sample t-test

data:  samp4
t = 9.1695, df = 3, p-value = 0.002742
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 410734.8 847390.7
sample estimates:
mean of x
 629062.8
```

```
        One Sample t-test

data:  samp12
t = 41.483, df = 11, p-value = 1.943e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 673403.9 748866.0
sample estimates:
mean of x
 711134.9
```

The mean of this population is 70801.4, and it can be seen that it is indeed in both of the intervals. The interval for n=12 is more accurate, as it has a larger sample size, and therefore the CLT applies more strongly compared to n=4.

Formation of Hypothesis

For H_0: $\mu = 70801.4$ vs. H_a: μ does not equal 70801.4 for both n=4 and n=12 at a significance level of 0.05, the test results are as follows:

```
        One Sample t-test

data:  samp4
t = -1.1477, df = 3, p-value = 0.3343
alternative hypothesis: true mean is not equal to 707801.4
95 percent confidence interval:
 410734.8 847390.7
sample estimates:
mean of x
 629062.8
```

```
        One Sample t-test

data:  samp12
t = 0.19446, df = 11, p-value = 0.8494
alternative hypothesis: true mean is not equal to 707801.4
95 percent confidence interval:
 673403.9 748866.0
sample estimates:
mean of x
 711134.9
```

The conclusion for both is that we fail to reject the null hypothesis, but we note that as the sample size grows, the conclusion becomes more reliable. We now test H_0: $\sigma = 79954.04$ vs. H_a: σ does not equal 79954.04 and H_0: $\sigma = 79954.04$ vs. H_a: $\sigma < 79954.04$ at a significance level of 0.05, and indeed we fail to reject both null hypotheses.

```
Part 6.3: Testing Statistic =  6.061187 , Interval = ( 3.815748 21.92005 )
Part 6.4: Is Testing Statistic =  6.061187 less than 95 Confidence, 11 DF chisq =  4.574813 ?: FALSE
```

7. Regression

Preprocessing

In this project, we use a linear regression model and a negative binomial regression. Through these methods, we hope to see how each independent variable affects the number of times an item matching certain characteristics was checked out. After modeling, we will use ANOVA to see how significant each variable is in predicting the value of checkouts.

Since our data is so large, we attempted to make it more manageable through grouping by certain characteristics. We grouped items such that their checkout year, month, material type, usage class, checkout type, and subjects (genres) all match. We then sum up the number of checkouts from each item in each group. An example showing 5 groups is provided in Figure 1. The "Subjects" categorical variable has many possible values, so we limited it to its top 20 values, and labeled all other subject values as "Other". We assigned each value per categorical variable to a dummy variable.

	CheckoutYear	CheckoutMonth	MaterialType	UsageClass	CheckoutType	Subjects	Checkouts
0	2017	8	ATLAS	Physical	Horizon	Other	121
1	2017	8	AUDIOBOOK	Digital	OverDrive	Fantasy	6741
2	2017	8	AUDIOBOOK	Digital	OverDrive	Fiction	35878
3	2017	8	AUDIOBOOK	Digital	OverDrive	Historical Fiction	5371
4	2017	8	AUDIOBOOK	Digital	OverDrive	Juvenile Fiction	5797

Figure 1: Example Grouping of Items

Multiple Linear Regression

We first performed a multiple linear regression on the grouped data. All the features seen in Figure 1 are our inputs and the number of checkouts is our output. We scale down the number of checkouts with a log transformation so the data. Linear regression is very sensitive to outliers, so this had to be done to make a working model. The model has a mean absolute error (MAE) value of 1.351 and Pearson's correlation coefficient of 0.81. Thus, the linear regression provides a very good estimate of the data in general. Since the data is log transformed, it will rank the amounts of checkouts correctly but the magnitudes may be off.

A scatter plot that shows actual log checkouts against predicted log checkouts is shown in Figure 2. Since there were roughly 70,000 points to plot, we randomly selected 500 to show.

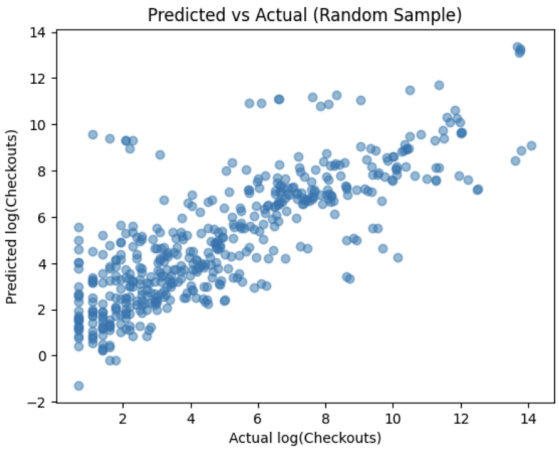


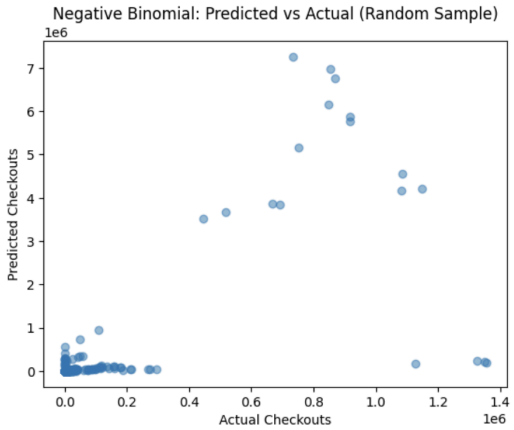
Figure 2: Graph of differences between actual checkouts and predicted checkouts

According to the linear regression equation, audiobooks, ebooks, and video discs increase checkouts the most out of all material types. Graphic novels, non-fiction, and video recordings for the hearing impaired, and “Other” are the subjects that increase checkouts the most. The usage class that is more popular is digital.

Negative Binomial Regression

We put raw data and log transformed data into a negative binomial regression to see if we could get a better result compared to the linear regression. Negative binomial regression works well with datasets with high variance. As mentioned previously, the variance of this dataset is high because it is so skewed. Therefore, we fit it with the raw numbers of checkouts and with the log transformed numbers of checkouts.

We use the same inputs and outputs as the linear regression. The MAE for the raw data is 56307.38 and Pearson's correlation coefficient is 0.569. The MAE for the log transformed data is 1.59, and Pearson's correlation coefficient is 0.774. Clearly, the log transformed data is easier to predict because of the lower variance. Though the mean squared error is large, the correlation coefficient shows that the regression is still fairly accurate. This can be seen in Figure 3.



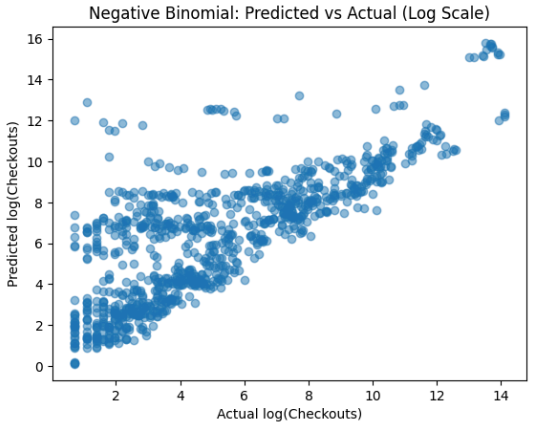


Figure 3: Negative Binomial Regression on Raw and Log Transformed Data

According to the model, the most popular material types are audio books, books, ebooks, movies, video discs, cassettes, and sound discs. The most popular subjects are cartoons and comics, fiction, video recordings for the hearing impaired, graphic novels, nonfiction, and "other". The usage class that is more popular is digital.

One Way ANOVA Tables

ANOVA tests are a good way to find out strengths of relationships between independent and dependent variables. We run multiple one-way ANOVA tests to see which variables explain the most about the number of checkouts. With very large datasets, statistical significance alone can be misleading, but to check the validity of this, we use eta. Eta is defined as the sum of squares of the variable divided by the sum of squares of the variable added to the sum of squares of the residual. It represents the percentage of variance that is explained by the variable.

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

We create a total of six ANOVA tables measuring the sum of squares between variables and the residuals. Our tables will measure "Subjects", "Material Types", "Checkout Year", "Checkout Month", "Usage Class" and "Checkout Type".

Results indicate that material type is the most important factor in predicting the number of checkouts, explaining ~45.7% of variance. The subject variable explains a moderate proportion of variance, at 7.5%. Usage class and checkout type explain 5.4% of the variance each. Temporal variables are of low significance, with checkout year explaining 0.24% of variance while checkout month only explains 0.015%.

8. Conclusion

From our analysis, we are able to come to multiple conclusions about the data. We can conclude that video discs and audiobooks are the most popular form of material types, and physical media is slightly less popular than digital. We can conclude that the most popular subjects are graphic novels, fiction, non-fiction, and video recordings for the hearing impaired.

In our ANOVA tables, we are able to see that material type is the most important factor when estimating the total number of checkouts. This is followed by the subjects of the item. Checkout type and usage class are less important but still significant in calculating an estimated number of checkouts. Checkout month and year affect the number of checkouts insignificantly.

In further research, it would be interesting to be able to classify all subjects instead of just the top 20 and see how those affect the number of checkouts. It would also be interesting to see how these factors have all changed in more recent years.

Overall, all questions posed in the introduction were answered through visualization, correlation analysis, regression, and ANOVA. This concludes the examination of the Seattle Public Library Checkout Patterns dataset.