DSCI 320 Project Proposal - MILESTONE I

Group Members:
Sadia Khan Durani
Heidi Lantz
Lillian Milroy

Title: **Analyzing Trends In Library Usage**

PART I : Initial Exploration

Dataset Link:
https://www.kaggle.com/datasets/datasf/sf-library-usage-data?select=Library_Usage.csv

Data Abstraction:

ATTRIBUTE TYPE	NAME	CARDINALITY	SEMANTICS
Quantitative	Total Checkouts	0 - 35907	The total number of items checked out from the library
	Total Renewals	0 - 8965	The total number of times checked-out items were renewed
	Supervisor District	1 - 11	Automated field given to those patrons with 'Outside of Country' set as true
	Circulation Active Year	2003 - 2016	The year the patron last checked out items
	Year Patron Registered	2003 - 2016	The year the patron registered with the library
Nominal	Patron Type Code	18	Code for type of patron
	Patron Type Definition	18	A description of the type of patron E.g., SENIOR, ADULT, VISITOR, TEACHER CARD
	Home Library Code	74	Code for the patron's original library branch registered
	Home Library Definition	35	A description of the patron's original library branch
	Notice Preference Definition	4	A description of the patron's preferred method of contact
	Provided Email Address	2	True - the patron provided their email.
	Outside of County	2	True - the patron's home address is not San Francisco
Ordinal	Age Range	10	A range of age groups E.g., 0 to 9 years, 20 to 24 years, etc.

	Notice Preference Code	4	Code for patron's preferred method of contact
Temporal	Circulation Active Month	12	The month the patron last checked out items

Exploratory Data Analysis:
The original data set consists of 423448 items and 15 attributes. After dropping items with null values, the dataset is left with 275127 items and 15 attributes. From the attributes listed above, we will use all 15 for EDA purposes to analyze distributions and learn the trends present in our dataset. Our visualizations below were created after removing null values and taking a random sample so that Altair can plot the data for the purpose of creating rough sketches.

LOADING THE DATA

```
data = pd.read_csv('archive-3/Library_Usage.csv', parse_dates = ['Circulation Active Month'])
```

```
data.columns
```

```
Index(['Patron Type Code', 'Patron Type Definition', 'Total Checkouts',
       'Total Renewals', 'Age Range', 'Home Library Code',
       'Home Library Definition', 'Circulation Active Month',
       'Circulation Active Year', 'Notice Preference Code',
       'Notice Preference Definition', 'Provided Email Address',
       'Year Patron Registered', 'Outside of County', 'Supervisor District'],
      dtype='object')
```

```
data.head(5)
```

	Patron Type Code	Patron Type Definition	Total Checkouts	Total Renewals	Age Range	Home Library Code	Home Library Definition	Circulation Active Month	Circulation Active Year	Notice Preference Code	Notice Preference Definition	Provided Email Address	Year Patron Registered	Outside of County	Supervisor District
141	3	SENIOR	469	282	65 to 74 years	C2	Chinatown	July	2016.0	z	email	True	2003	False	1.0
142	0	ADULT	256	102	45 to 54 years	P5	Portola	July	2016.0	z	email	True	2003	False	9.0
143	0	ADULT	552	105	55 to 59 years	R3	Richmond	July	2016.0	z	email	True	2003	False	1.0
144	0	ADULT	581	159	60 to 64 years	X	Main Library	August	2013.0	z	email	True	2003	False	10.0
145	0	ADULT	1245	1439	55 to 59 years	N6	North Beach	July	2016.0	z	email	True	2003	False	3.0

Frequency Tables

Quantitative Variables:

	Total Checkouts	Total Renewals	Supervisor District	Circulation Active Year	Year Patron Registered
min	0.000000	0.000000	1.0	2003.0	2003.0
max	35907.000000	8965.000000	11.0	2016.0	2016.0
mean	161.982097	59.657327	NaN	NaN	NaN

Nominal + Ordinal + Temporal:

	Patron Type Code	Patron Type Definition	Age Range	Home Library Code	Home Library Definition	Circulation Active Month	Notice Preference Code	Notice Preference Definition	Provided Email Address	Outside of County
unique	[3, 0, 16, 55, 5, 9, 4, 15, 100, 10, 12, 1, 10...	[SENIOR, ADULT, DIGITAL ACCESS CARD, RETIRED S...	[65 to 74 years, 55 to 59 years, 60 to 64 year...	[X, M8, P7, S7, M4, N4, E9, C2, R3, N6, P9, M6...	[Main Library, Mission Bay, Potrero, Sunset, M...	[November, October, January, February, July, D...	[z, p, a, -]	[email, phone, print, none]	[True, False]	[True, False]

Individual Frequency tables across the categorical variables were created to get a more in depth understanding of the relative frequencies of each level contained in the variables.

Patron Type:

Patron Type Code	Frequency	Relative Frequency
0	272251	0.642938
1	59208	0.139824
2	28816	0.068051
3	41619	0.098286
4	14931	0.035261
5	862	0.002036
8	40	0.000094
9	977	0.002307
10	415	0.000980
12	95	0.000224
15	1782	0.004208
16	1744	0.004119
55	157	0.000371
100	349	0.000824
101	47	0.000111
102	44	0.000104
103	66	0.000156
104	45	0.000106

Patron Type Definition	Frequency	Relative Frequency
ADULT	272251	0.642938
AT USER ADULT	349	0.000824
AT USER JUVENILE	47	0.000111
AT USER SENIOR	66	0.000156
AT USER TEEN	44	0.000104
AT USER WELCOME	45	0.000106
BOOKS BY MAIL	95	0.000224
DIGITAL ACCESS CARD	1744	0.004119
FRIENDS FOR LIFE	40	0.000094
JUVENILE	59208	0.139824
RETIRED STAFF	157	0.000371
SENIOR	41619	0.098286
SPECIAL	977	0.002307
STAFF	862	0.002036
TEACHER CARD	1782	0.004208
VISITOR	415	0.000980
WELCOME	14931	0.035261
YOUNG ADULT	28816	0.068051

Age Range:

Age Range	Frequency	Relative Frequency
0 to 9 years	38242	0.090357
10 to 19 years	58944	0.139271
20 to 24 years	29761	0.070318
25 to 34 years	91083	0.215208
35 to 44 years	67390	0.159227
45 to 54 years	52492	0.124026
55 to 59 years	21230	0.050161
60 to 64 years	19800	0.046783
65 to 74 years	30141	0.071216
75 years and over	14150	0.033433

Home Library:

Home Library Code	Frequency	Relative Frequency
A5	7183	0.016965
AQUIS	1	0.000002
B2	8417	0.019879
B2AAA	6	0.000014
B2AZZ	1	0.000002
...	...	...
YBJ	3	0.000007
YJJ	807	0.001906
YJJAA	2	0.000005
YLW	782	0.001847
YLWAA	5	0.000012

Circulation Active Month:

Circulation Active Month	Frequency	Relative Frequency
April	25172	0.070799
August	22081	0.062105
December	20787	0.058465
February	19936	0.056072
January	20623	0.058004
July	91566	0.257538
June	40316	0.113392
March	23968	0.067412
May	31381	0.088262
November	19430	0.054649
October	19716	0.055453
September	20568	0.057849

Notice Preference:

Notice Preference Code	Frequency	Relative Frequency
-	3	0.000007
a	31336	0.074002
p	68172	0.160993
z	323937	0.764998

Notice Preference Definition	Frequency	Relative Frequency
email	323937	0.764998
none	3	0.000007
phone	68172	0.160993
print	31336	0.074002

Email Address:

Outside of County:

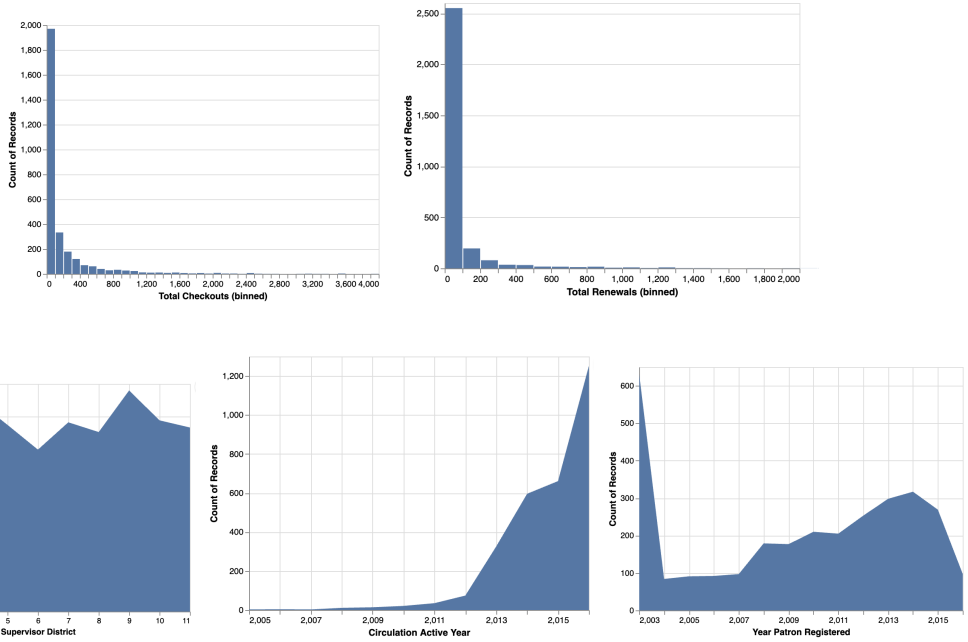
Provided Email Address	Frequency	Relative Frequency
False	87028	0.205522
True	336420	0.794478

Outside of County	Frequency	Relative Frequency
False	359628	0.849285
True	63820	0.150715

These separate frequency tables show the unique levels present within each variable. We can analyze the levels within a variable and see which ones are more frequent in this dataset. This will aid us in drawing relationships between the data. Below are univariate visuals of the variables to visually show the distributions of the data.

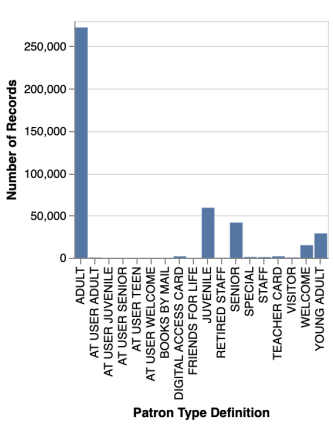
Univariate Visuals

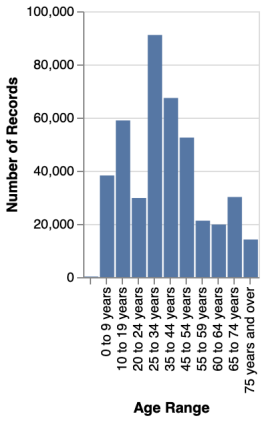
Quantitative Variables:

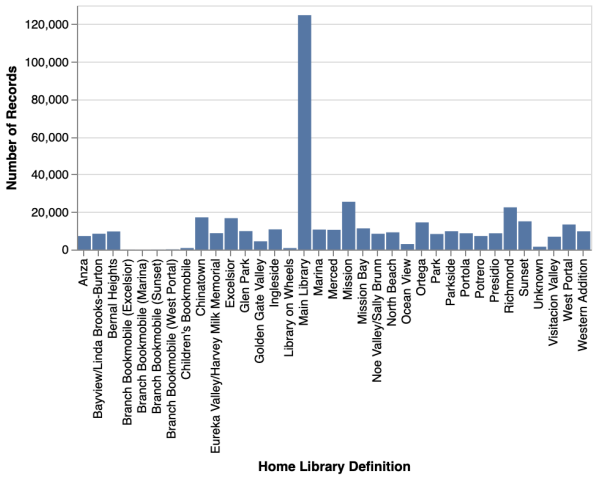


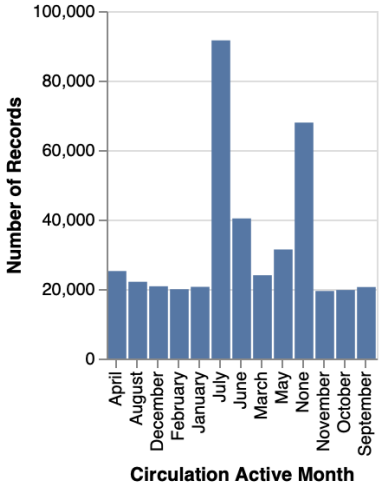
These plots show where the peak in data occurs across the variables. The first histogram shows that most patrons have total checkouts in the range from 0 to 200. The second histogram shows that over 2,500 patrons have done 0 to 300 renewals. Both of these plots show that the data is right-skewed for these 2 variables, hence the mean is not the best measure to state. The third visual indicates that the number of patrons are fairly distributed within the supervisor districts. The fourth visual shows that many patrons last checked out items from 2013 to 2016. Lastly, the fifth visual shows that many patrons registered in 2003, then a second increase from 2013-2015.

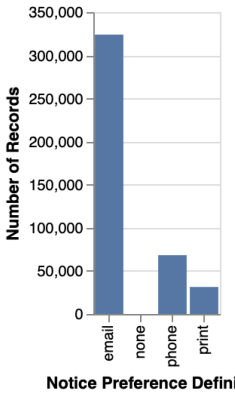
Categorical Variables:

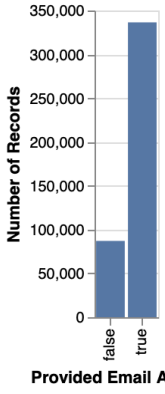


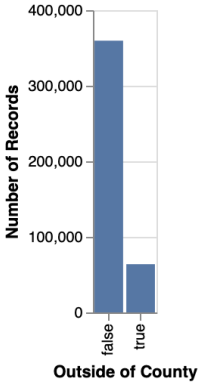






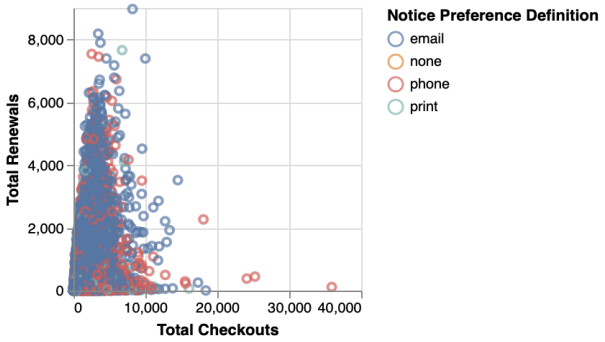


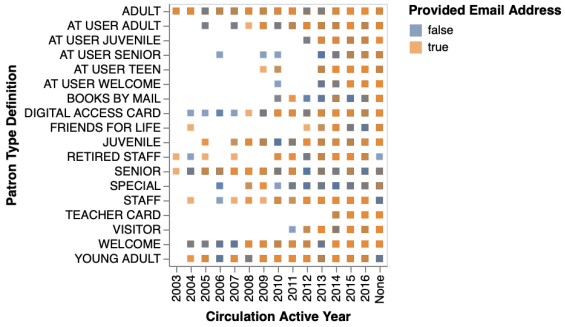


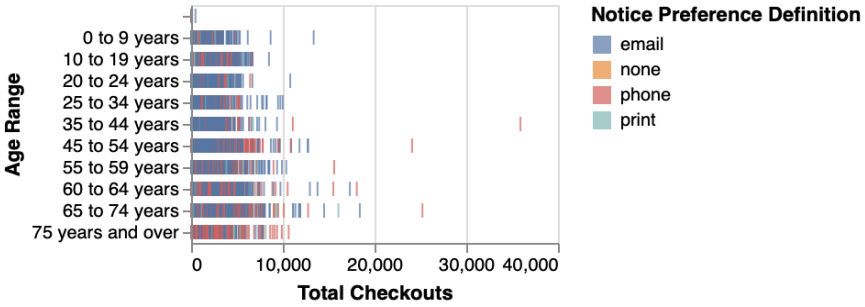


Of these bar charts created to visualize the categorical variables, we can tell that the data consist of patrons that are mostly in the ADULT section from ages 25 to 34. Many patrons were originally registered in their ‘Main Library.’ Most patrons opted for the email option to receive library notices. This suggests there may be an association between ‘Notice Preference Definition’ and ‘Age Range.’ More than half the patrons reside in the county.

Multivariate Visuals:







From these 3 multivariate visuals, a few examples of questions that we could answer are, what age group has the total number of checkouts? Or, did patrons frequently renew their checked-out items? Or, is there a year where most types of patrons last checked out? Or, what was the prevalent notice preference in the '75 years and over' age range? This lays out the base of our further analysis. In Part II, we dig deeper into further questions that could uncover the relationships and trends between the variables.

PART II : Project Scope

Introduction

Analyzing Trends in Library Usage

This project aims to analyze trends in library patron behavior using a dataset containing approximately 420,000 library patrons. We intend to identify trends in how various age demographics interact with media, and how these trends may affect other aspects of library usage. Our analysis will provide insight into how different kinds of patrons engage with the library system, and how to optimize their experience going forward.

Our intended audience for this project includes library administrators, researchers, and policymakers who are interested in understanding and improving library services.

The project seeks to provide valuable insights into library patron behavior to help library administrators make data-driven decisions. By understanding patron preferences, usage patterns, and demographics, libraries can tailor their services to better serve their communities and optimize resource allocation.

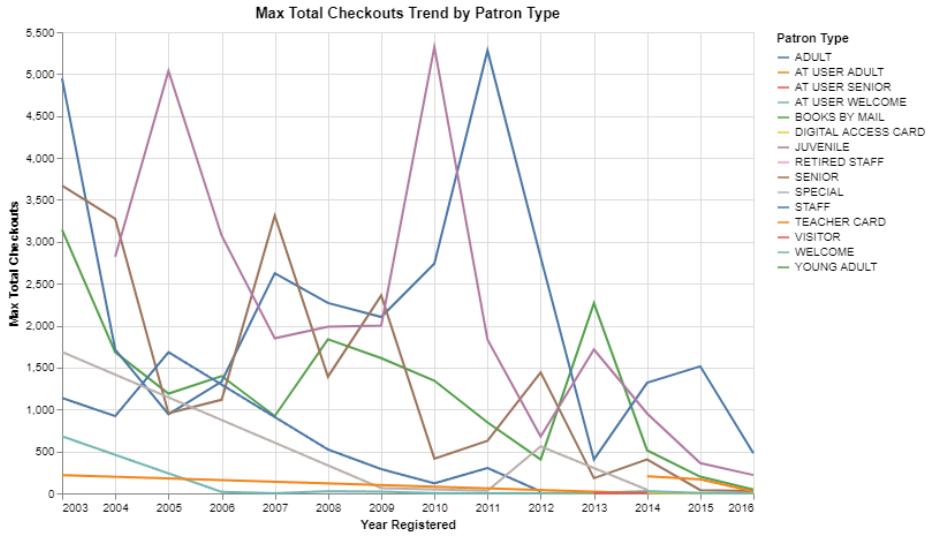
Task Analysis ⇒ Stasko's taxonomy for low-level tasks:

TASK TYPE	QUESTION
Retrieve Value	What is the highest total number of checkouts for adult patrons?
Filter	Which patron types have provided an email address for notifications?
Compute Derived Value	Calculate average total renewals for each age range
Characterize distribution	Look at distribution of year patron registered across age ranges
Correlation	Explore correlation between Circulation Active Year and Total Checkouts for each patron type

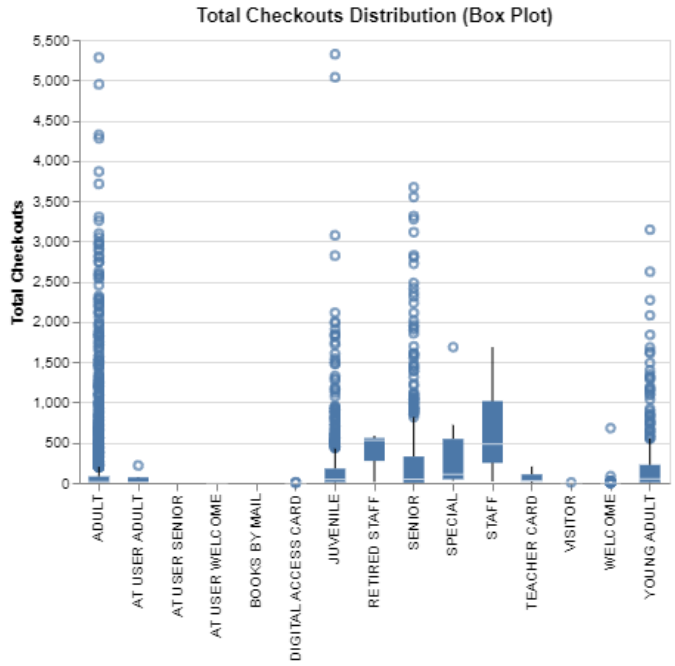
PART III : Visualization Ideas ⇒ **Preliminary Sketches**

Task 1: What is the highest total number of checkouts for adult patrons?

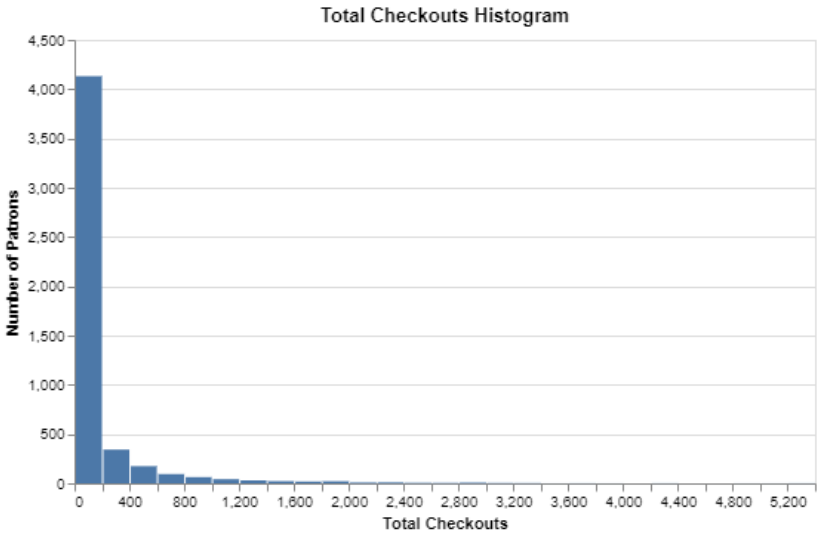
Sketch 1: Max Total Checkouts Line Chart



Sketch 2: Total Checkouts Box Plot



Sketch 3: Highest Total Checkouts Histogram



Critique:

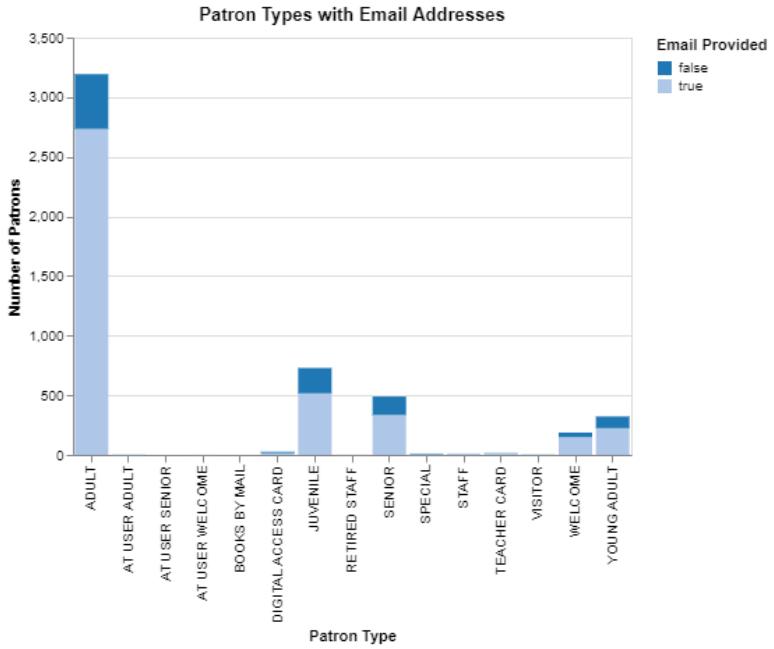
The line plot effectively displayed the maximum checkouts for each patron group in each year, offering valuable insights into library trends and group differences. However, the graph became cluttered due to numerous patron types and conflicting lines, making it challenging to distinguish between them. This tells us this graph has poor discriminability and affects its overall level of effectiveness. The overwhelming amount of patron types may affect the viewer's interpretation due to too many colors being in use. However, the graph is good at being expressive in terms of the max total checkouts.

The box plot provided the most informative view, offering a clear sense of the distribution of checkouts for each group. It does have low separability considering there is a lot of information and is hard to discern between groups. While it contained substantial information, it may not be suitable for some audiences as it contains a lot of information besides the maximum number of checkouts. If we are considering our audience, they likely would not have a statistical background, and therefore having knowledge about where the quantiles are and the general distribution isn't incredibly useful in this case.

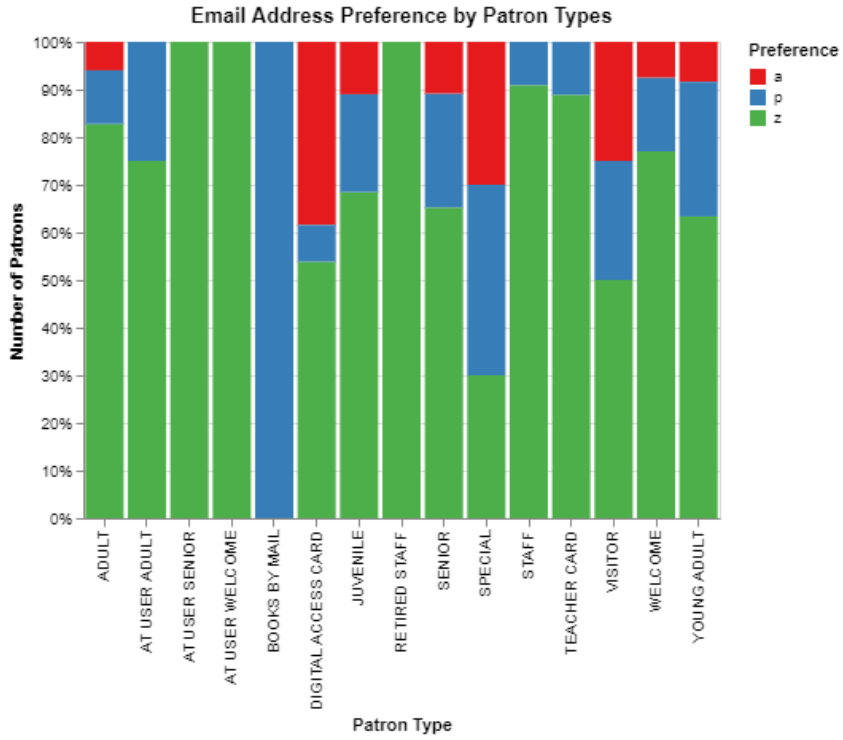
The histogram struggled to highlight or identify the patrons with the highest checkouts. Most patrons fell within the 0-200 checkout range, and the scaling made it challenging to spot those with over 5000 checkouts. This visualization, due to its scale, was less effective in showcasing the extreme values and identifying top patrons. There are significantly more patrons that fall within the 0-200 checkout range, and while that is useful information, it does not help us interpret the max total checkouts, regardless of bin size.

Task 2: Filter - Which patron types have provided an email address for notifications?

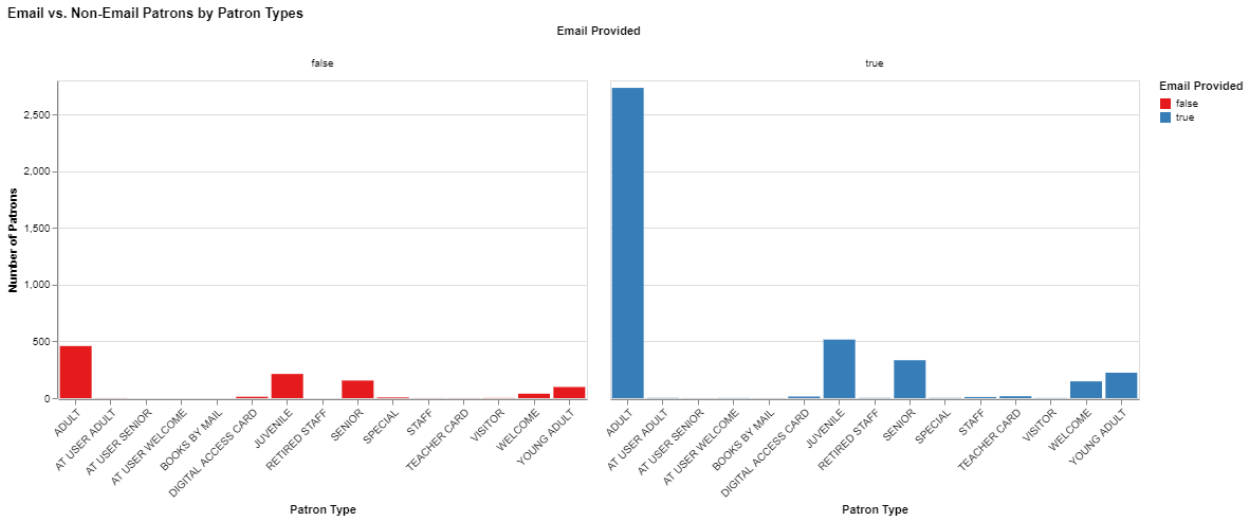
Sketch 1: Stacked Bar Chart



Sketch 2: Normalized Bar Chart



Sketch 3: Faceted Bar Chart



Critique:

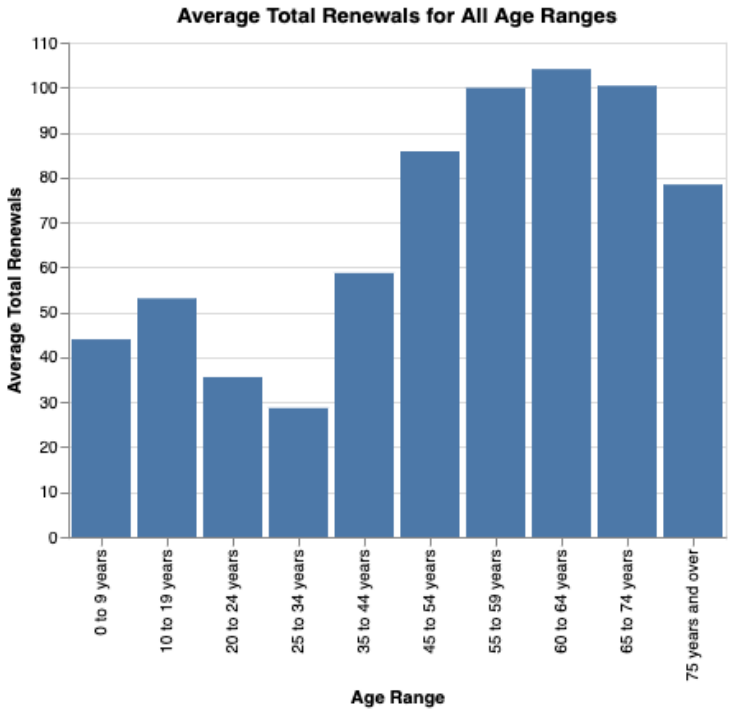
The stacked bar chart provides a clear visual representation of the number of patrons with and without email addresses, making it easy to compare email preferences across patron types. The use of colors helps differentiate between email provided and not, and the chart effectively conveys the quantity of patrons falling into each category. It makes this a suitable choice for this task, offering a balance of clear visual representation and numerical values, making it effective for understanding email address preferences among different patron types.

The normalized bar chart offers valuable insights into the proportion of patrons who provided email addresses, allowing for easy comparisons among patron types. It is extremely effective, as it is simplistic and lets us know which patron types are more likely to provide an email address. However, it lacks the direct numerical values, which can be a limitation when precise counts are preferred.

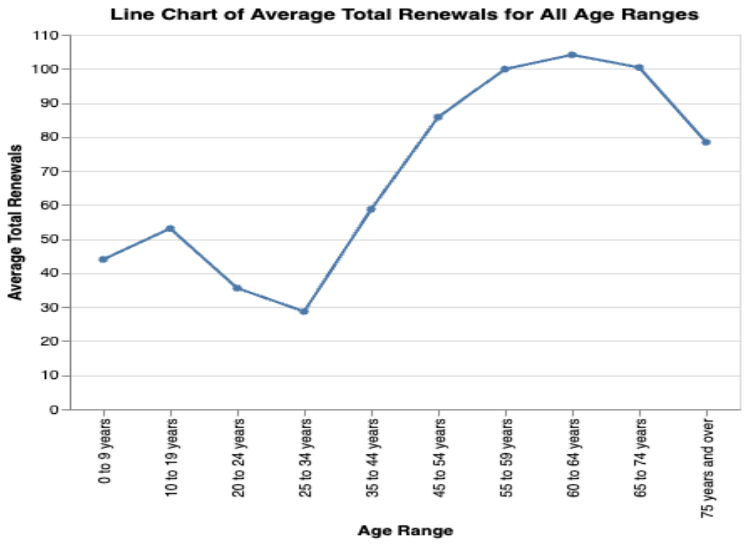
The faceted bar chart, while providing numerical values, makes it challenging to compare between email provided and not, as the facets separate the data. Even though it is on the same relative scale, the distance makes it hard for accurate comparison for the viewer's perspective. Therefore, although it is significantly expressive, it isn't very effective at communicating the main idea of differences in the amount of patrons providing their email.

Task 3: Calculate average total renewals for each age range

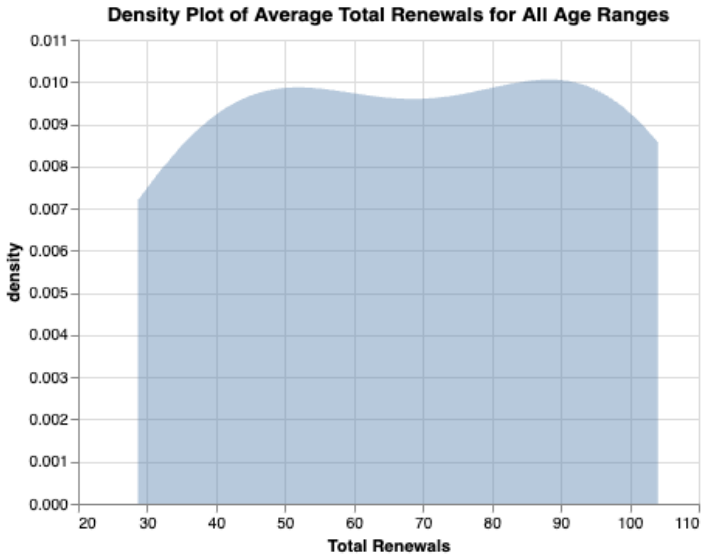
Sketch 1: Bar chart



Sketch 2: Line Chart



Sketch 3: Density Plot



Critique:

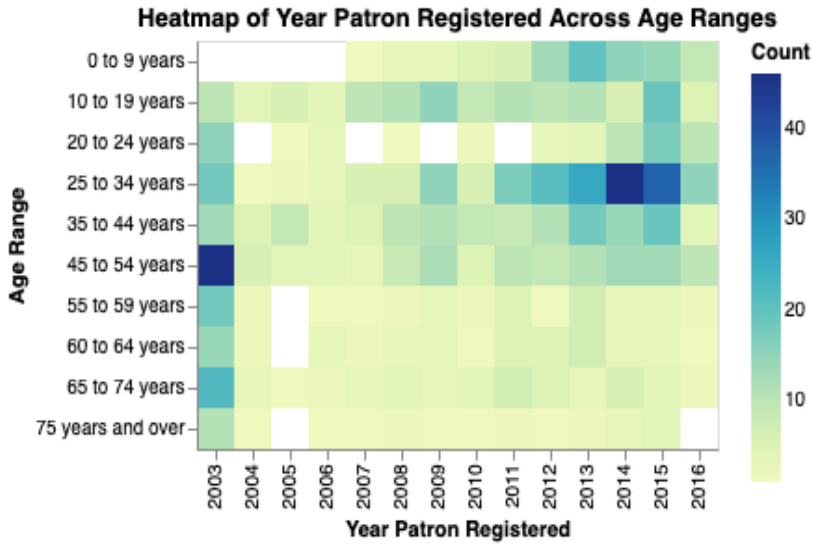
As is often the case, simpler is better for this objective. The density plot is not nearly as useful or intuitive as the bar chart or line chart.

Because a numeric quantity is being measured across groups, it is ideal that any visualization takes optimal advantage of the most important magnitude characteristic, position on a common scale. Both the line and bar plots do this, but the bar chart makes better use of the 2D space, adding another visual element allowing the viewer to quantify the quantities being represented.

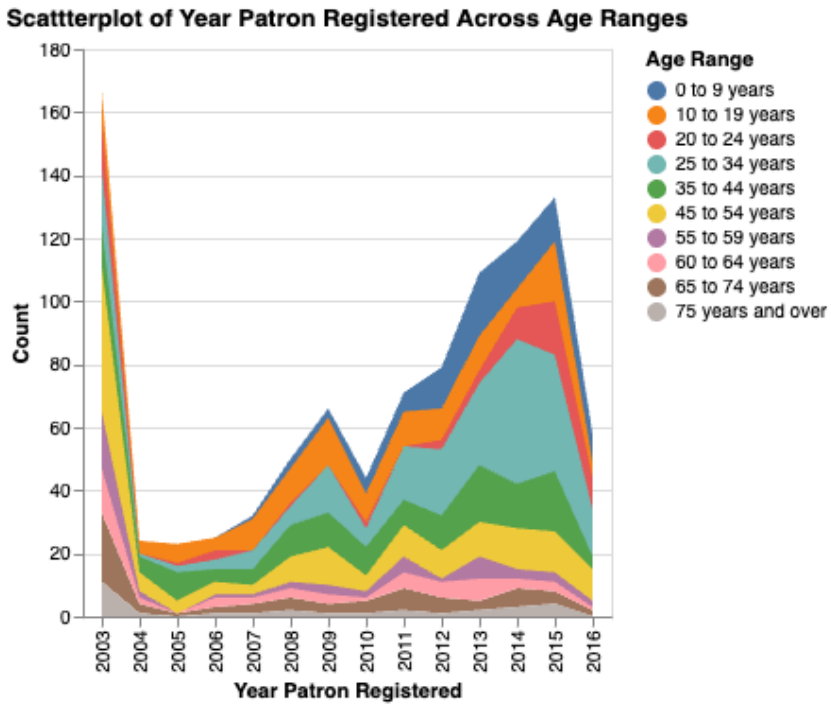
As such, the bar chart is the best candidate for this task, since it makes optimal use of the most effective relevant encoding channels.

Task 4: Look at distribution of year patron registered across age ranges

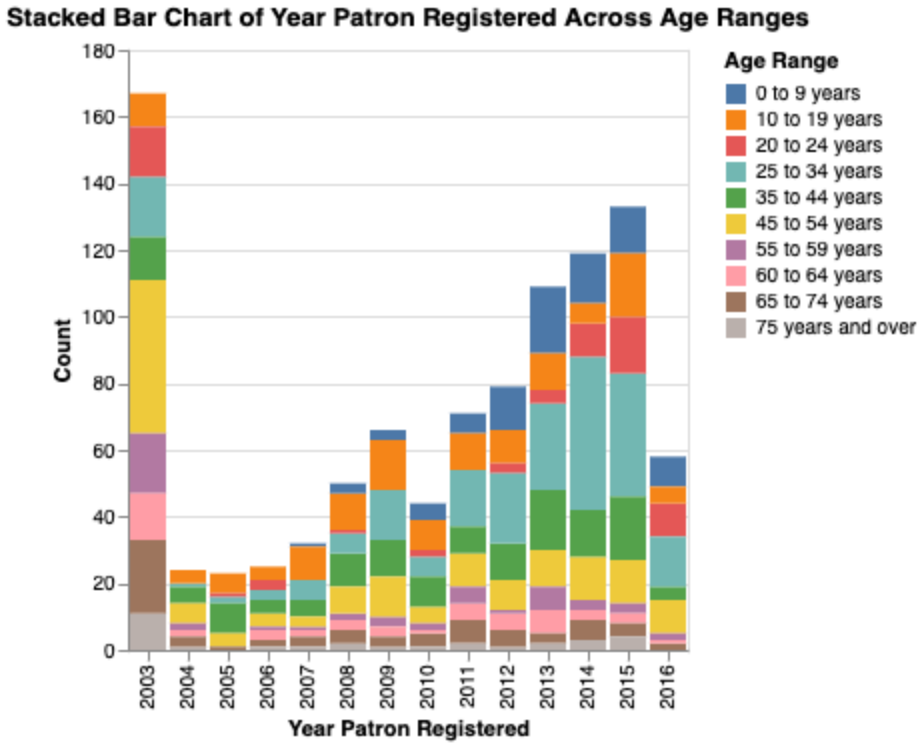
Sketch 1: Heatmap



Sketch 2: Area Chart



Sketch 3: Stacked Bar Chart



Critique:

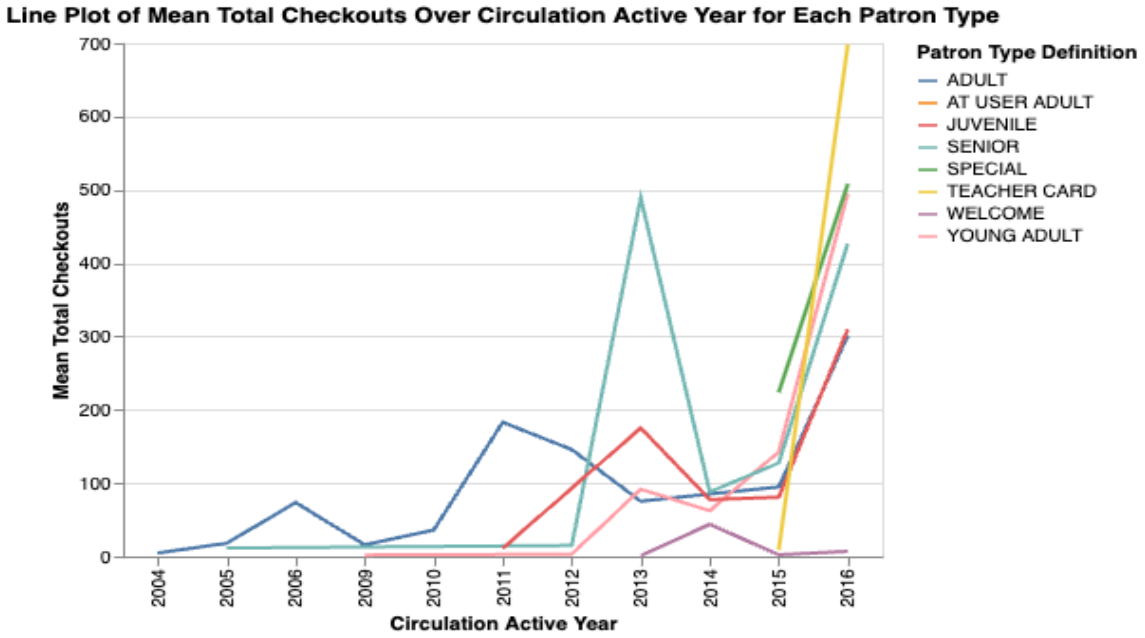
The heatmap is the least effective of the three visualizations - it isn't easy to identify trends in the data. Both the x-axis and y-axis are ordinal, and count is encoded by color, so one cannot immediately perceive the magnitude of the quantitative variable count for a given year/age group.

It is much easier to perceive the peaks and valleys that occur by age group in library membership registration looking at the area chart. However, its angularity might need to be adjusted to improve readability.

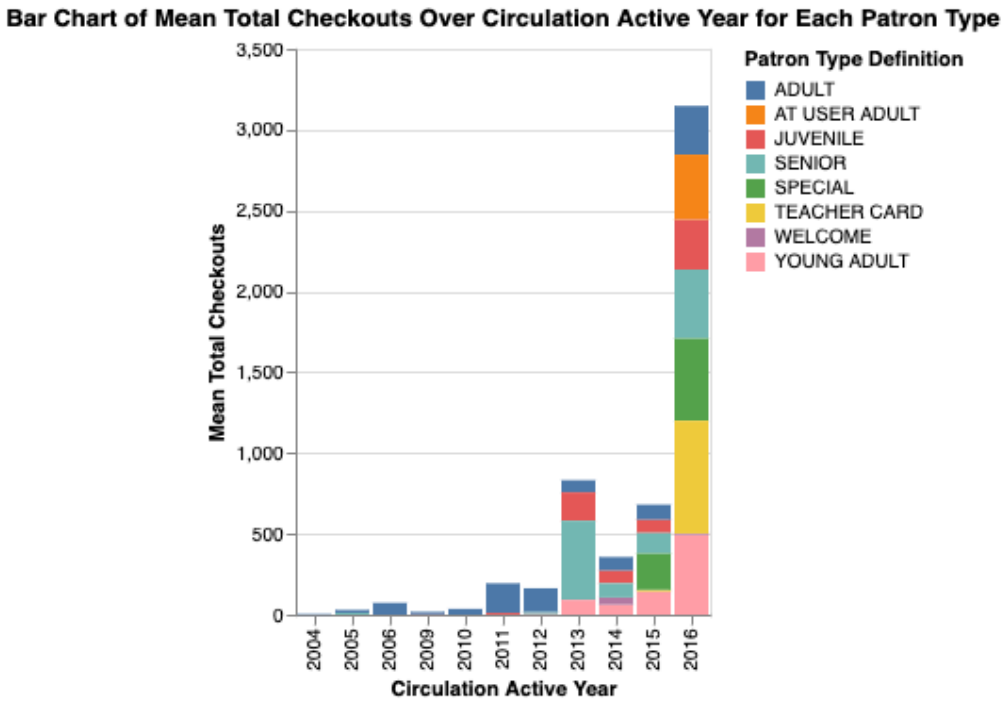
The stacked bar chart is probably the most effective visualization for the given task. Using colors to represent more than ~4 ordinal groups is not ideal (an issue also present with the area chart), but the stacked bar chart makes it most clear how many people registered per year, which age groups comprised the registrants, and in what proportions.

Task 5: Explore correlation between Circulation Active Year and Total Checkouts for each patron type

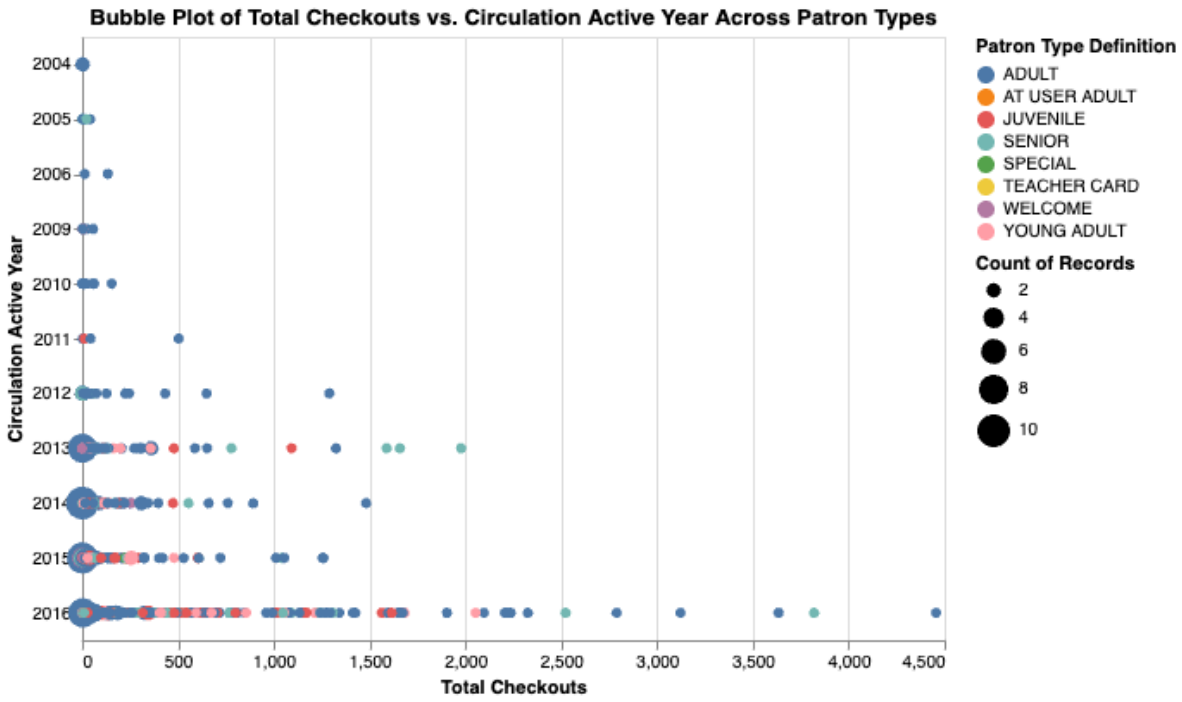
Sketch 1: Line Plot



Sketch 2:



Sketch 3: Bubble Chart



Critique:

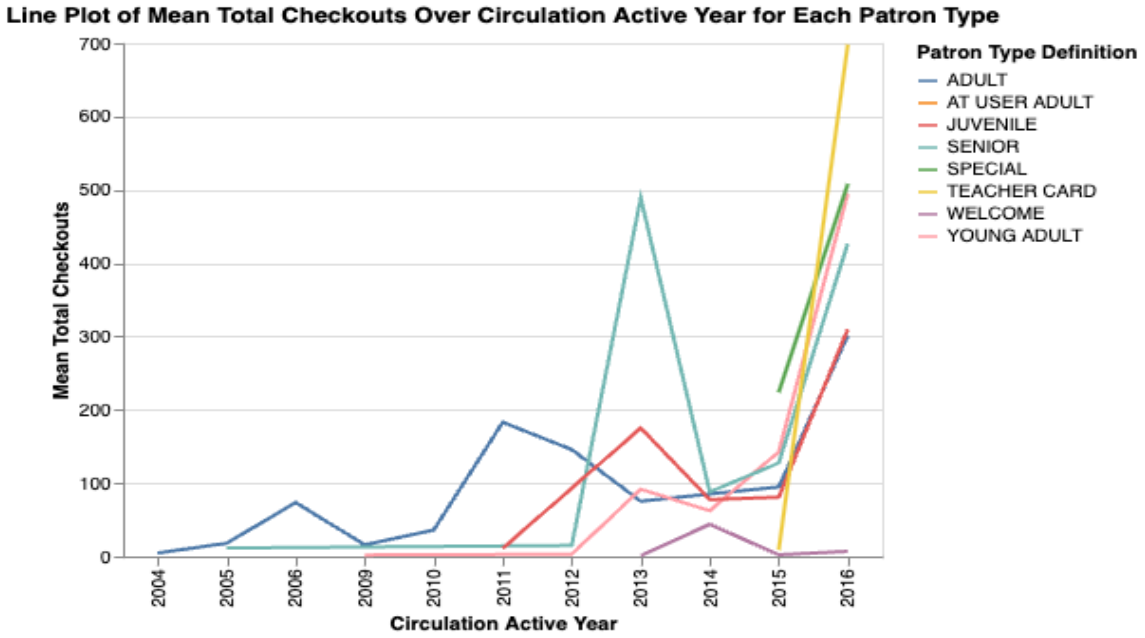
The bubble chart is the least effective of the visualizations. it has the most blank space, you have to look closely to determine the patterns in the data, and it does not do a great job of exploiting human visual tendencies.

The stacked bar chart more clearly illustrated trends in data; it is much easier to tell how many mean total checkouts there were in a given year, which patron groups comprised those checkouts, and in what proportions. It makes better use of the chart space than the bubble chart, but as with the previous task, it's not advisable to use color for too many variable levels.

The same comment holds for the line graph, but this option is the most elegant and efficient, because it conveys all the necessary information in the simplest, most user-friendly format. As such, the line chart is the most effective visualization for the task.

FINAL SKETCH:

The sketch we chose:



Critiques/Comments about this sketch:

The final line chart for Task 5, which explores the correlation between Circulation Active Year and Total Checkouts for each patron type, proved to be highly effective.

The graph we chose uses the line mark and has the channels of x as year, y as mean total checkouts, and color as patron type. The graph gives us insight into how often each group checks something out at the library and how that changes over time. We can notice a general upwards trend over the years for each group, and that certain groups, such as seniors, tend to have a higher mean value of checkouts than in comparison to the other groups.

When we talk about how effective the graph is, we are looking at whether the graph strikes an appropriate balance between simplicity and accuracy. We know that this is an effective graph because when we compared it to other visualizations for task 5, we saw this one communicated the information more efficiently but in a more simplistic manner. There is less “clutter” on the graph and allows the viewer to easily interpret what the graph is telling them. As well, we can see that the channels we used are optimally matched with the data types they represent.

When we talk about how expressive the graph is, we want to make sure that whatever channel is being used is representing only the data that it was meant to represent. We don’t have any instances of, for example, inferring ordering when there is none present. This tells us our line chart is meaningfully expressive.

Looking at the channel characteristics, we notice that this graph has good separability, integrability, and has no conflicting popouts that affect recognition. The discriminability on this graph could use some improvement, considering that it is recommended to have less than 5 color hues otherwise it is hard to tell the colors apart. Since we have 8 different patron types present in this graph, it makes the discriminability level something we should improve upon.

These principles, along with the chart's simplicity and user-friendliness, made it an elegant and efficient choice for visualizing the dataset and understanding trends over time. This line chart provides interesting insight into which groups are more active at the library and how that can change from year to year. It reminds us of the relevancy of our project, by highlighting the potential influence of these groups on library usage and operations.

PART IV : Next Steps

Outline

The next things that we plan to do as a group are:

- Set strict deadlines to ensure each member contributes equally for each section while considering their availability to ensure the completion of the project.
- We plan to follow the timeline below.
- **Week1: Nov 6th - Nov 12th**
    - Review project milestone II requirements and allocate tasks to each member.
    - Clear up any confusions with the teaching team.
    - Set up potential meeting days to meet to ensure each member is making progress within their part.
    - Begin implementing the visualizations in the notebook.
- **Week2: Nov 13th - Nov 19th**
    - Continue implementing the visualizations adhering to the requirements such as interactive based visualizations.
    - Set up a comprehensive report with our introduction, hypotheses and conclusions.
- **Week3: Nov 20th - Nov 26th**
    - Final Revisions of the implementation on Jupyter Notebook
    - Review and edit the report as a pdf
- **Nov 27th - Nov 28th**
    - Submit.