Seattle Public Library Checkouts

Capstone Project 2 by Samuel Ma

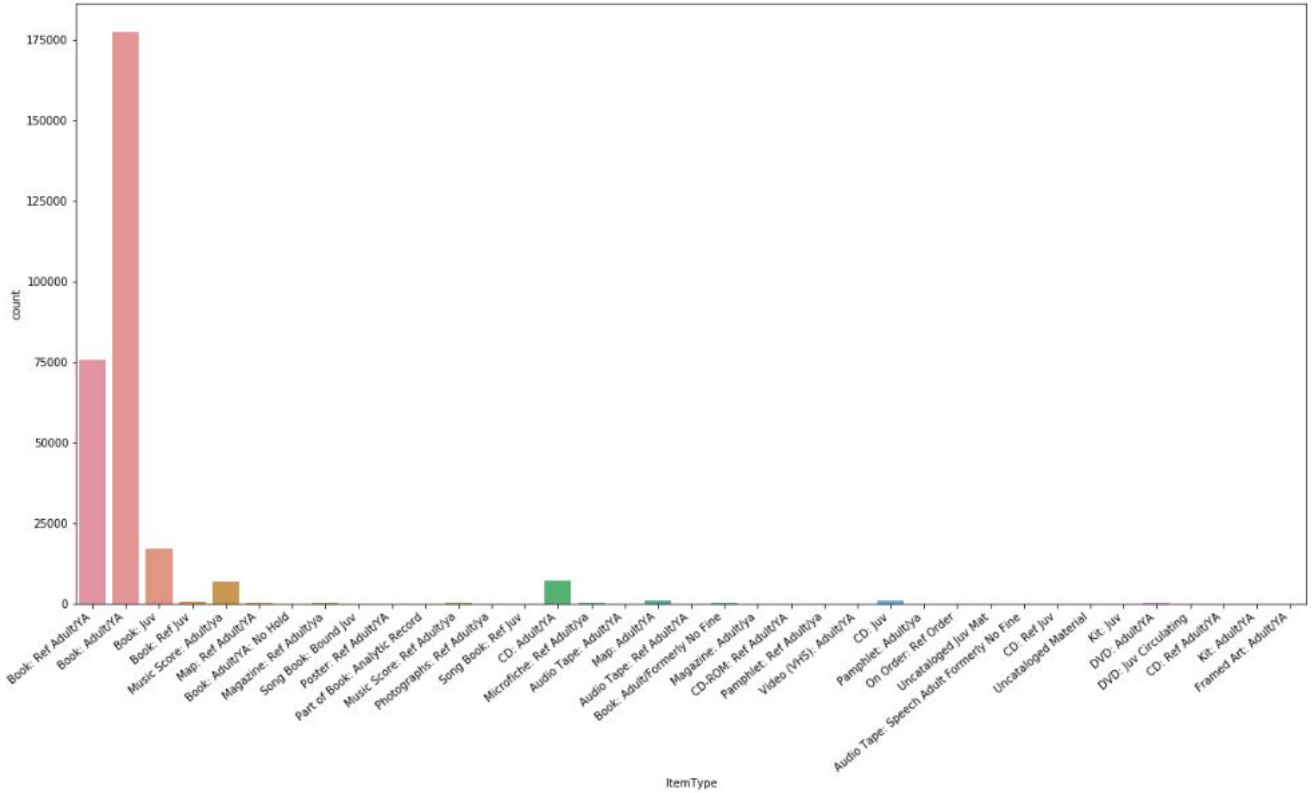
Data taken from Kaggle, Seattle Library inventory and Checkout records from 2017

Kaggle data:
https://www.kaggle.com/seattle-public-library/seattle-library-checkout-records

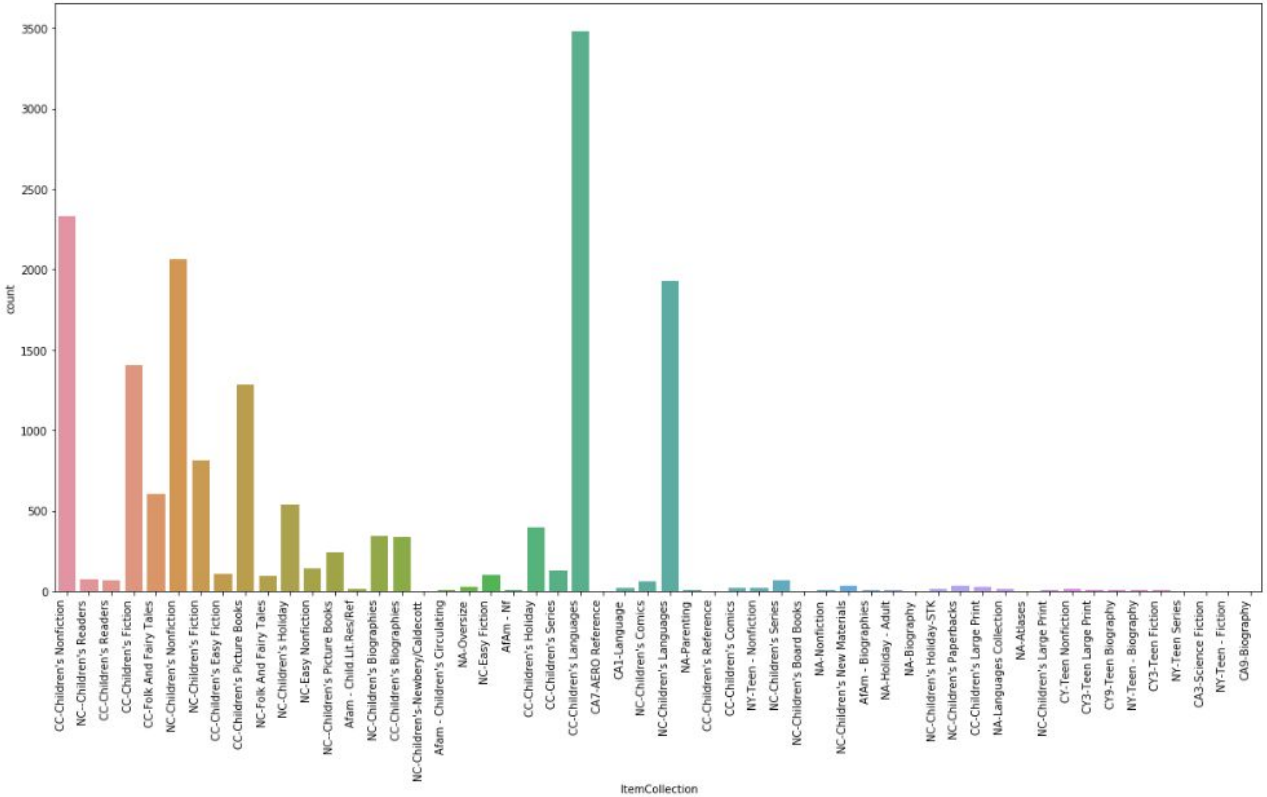
Original data can be taken from here:
https://data.seattle.gov/Community/Library-Collection-Inventory/6vkj-f5xf

https://data.seattle.gov/dataset/Checkouts-by-Title-Physical-Items-/3h5r-qv5w (note, as of 5/3/2020 it seems like the data was removed from government site but the data is still available on kaggle)

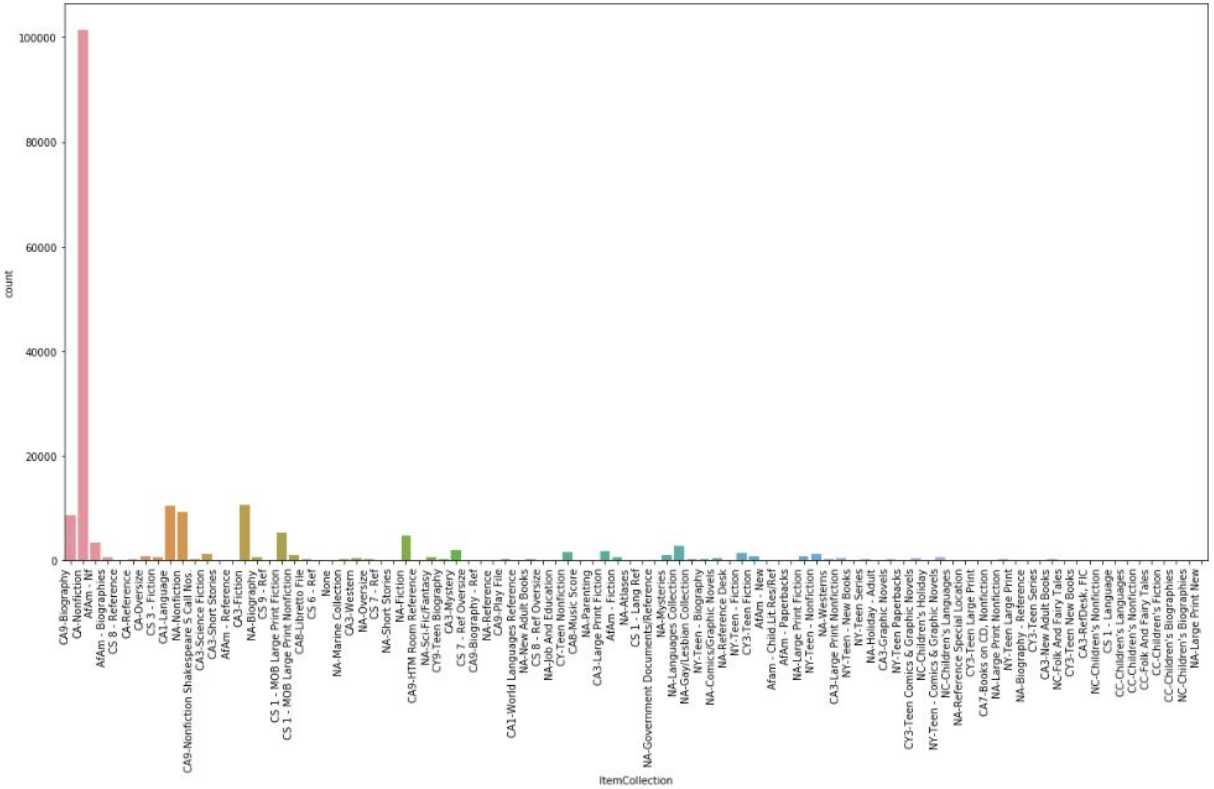
Initial Analysis - Counts of item types



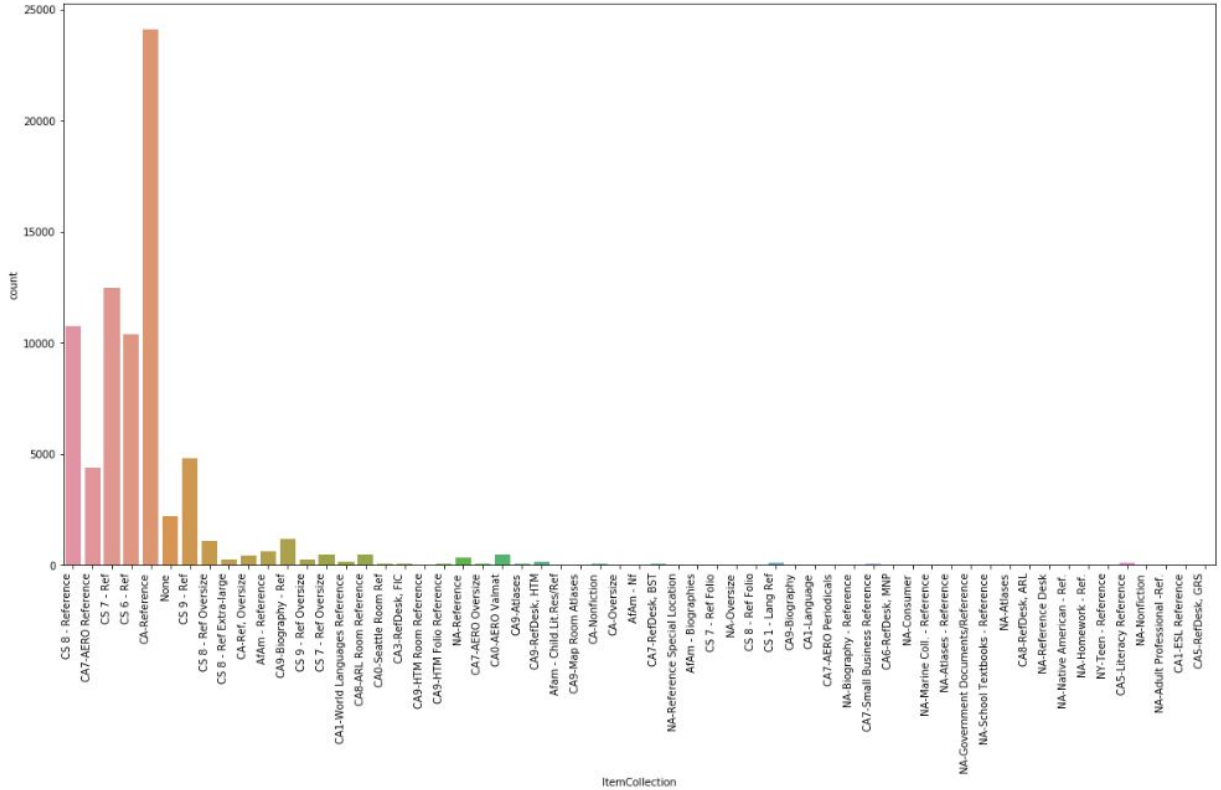
Juvenile Book Data by Collection



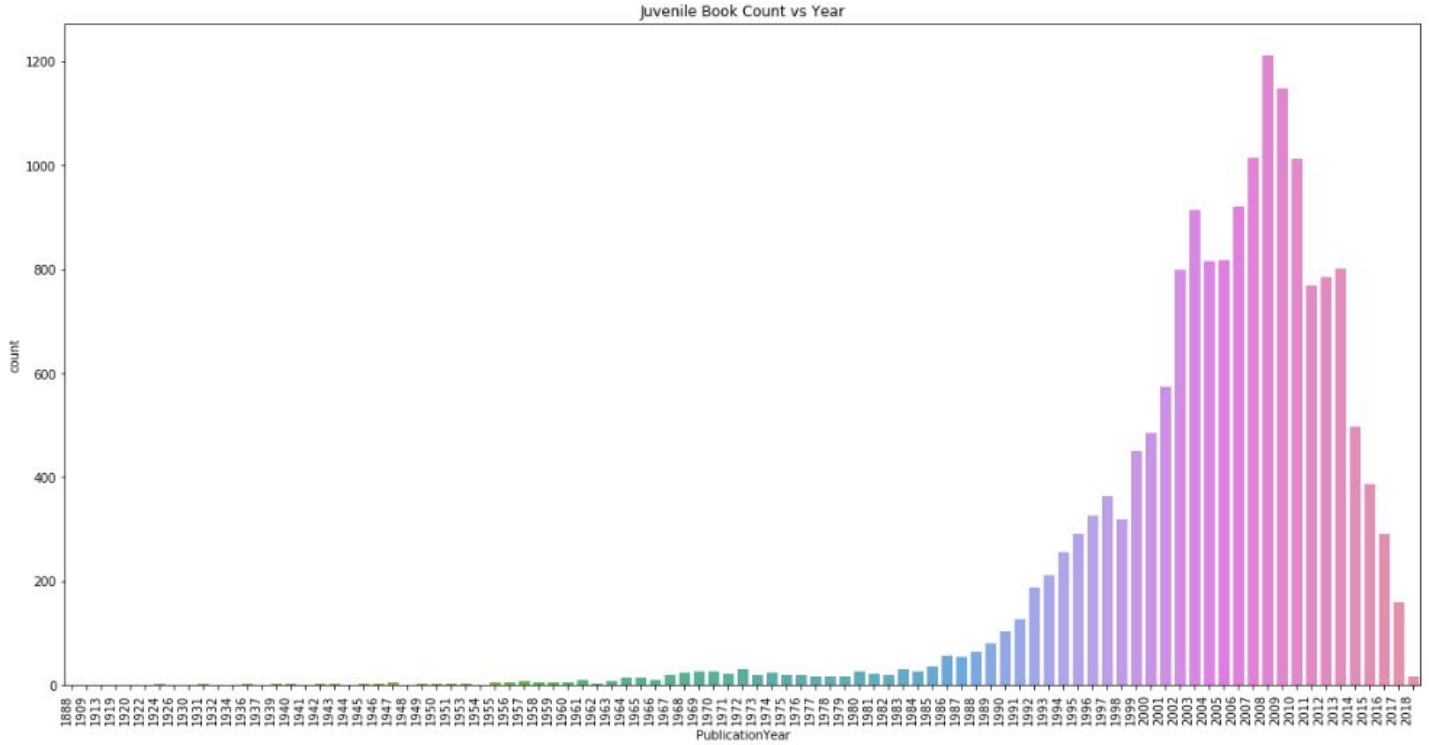
Adult Book Data by Collection



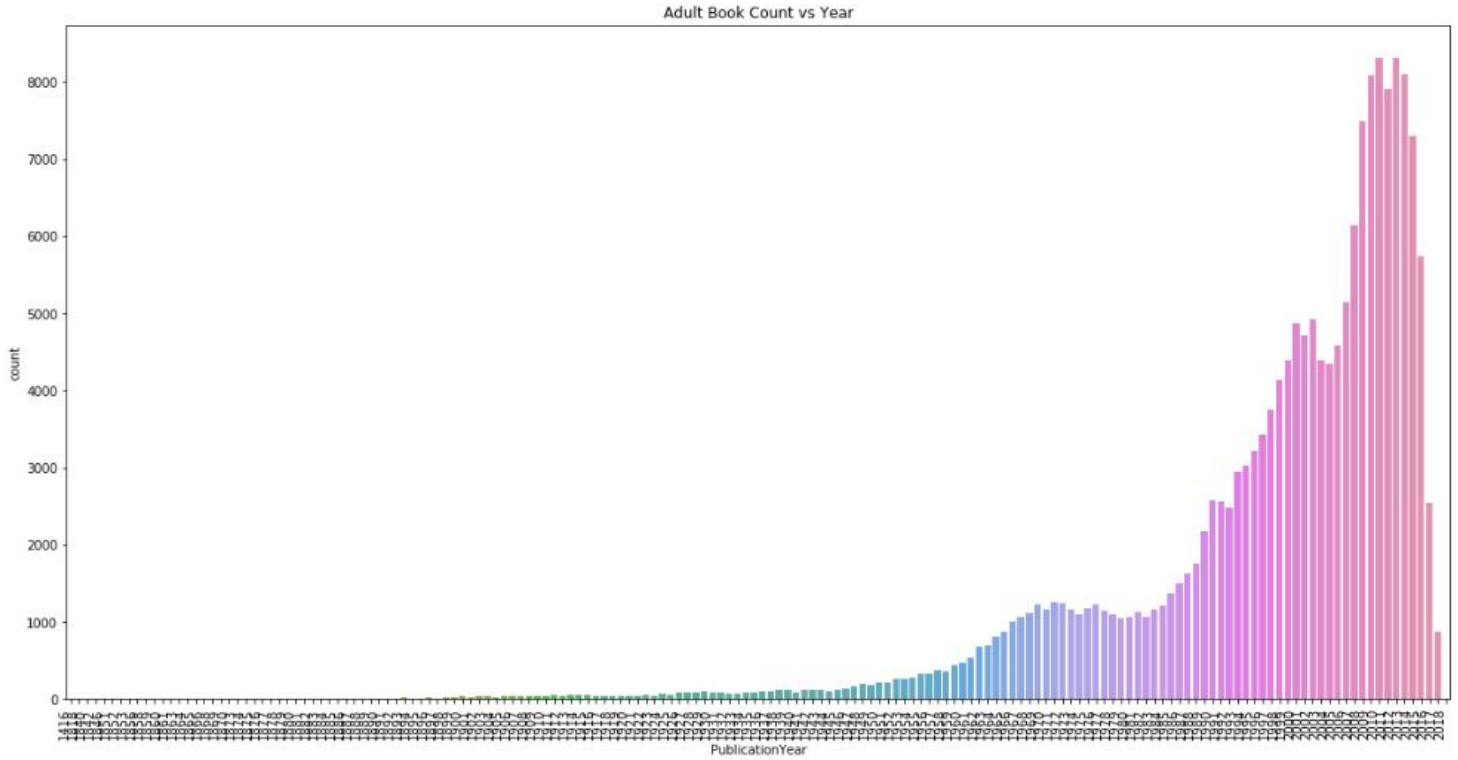
Reference Data by Collection



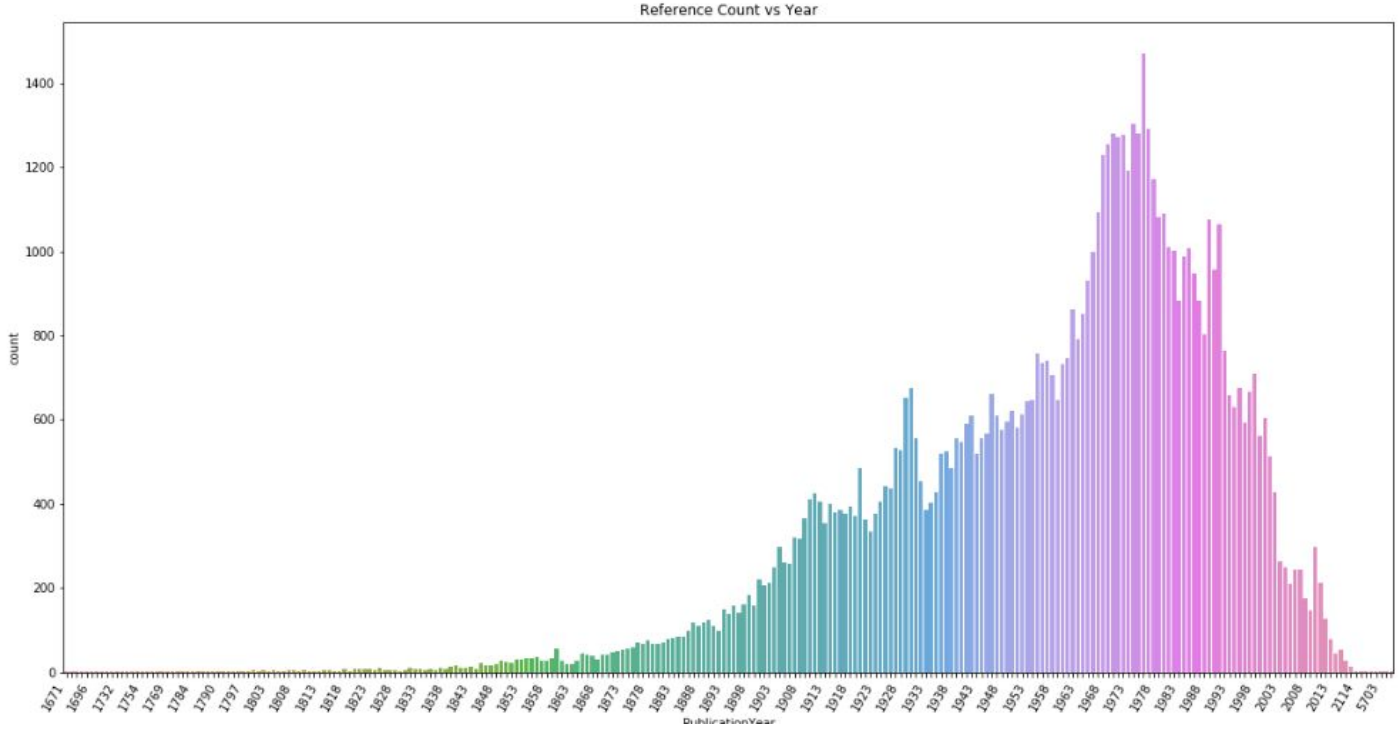
Juvenile Book Count vs Publication Year



Adult Book Data Count vs Publication Year



Reference Data Count vs Publication Year



Statistics

Chi-square test: compare year and type of item

1.
    a. Hypothesis Null Hypothesis (H0): Variables are independent.
    b. Alternative Hypothesis (H1): Variables are not independent
2. Use alpha value of 0.05
3. Data

PublicationYear	1954	1959	1966	1967	1968	1969	1970	1971	1972	1973	...	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Adult Books	152	183	495	363	398	430	582	699	544	633	...	38232	45562	57132	68438	88924	103118	123169	164146	317581	227828
Juvenile Books	578	551	354	1025	760	781	1686	1296	2028	1374	...	43012	50062	63989	74057	92163	116181	138488	184467	233445	116430
CDs	4	28	39	65	60	28	161	89	232	70	...	11810	14748	16015	19138	20557	26159	30471	37080	64329	38157

Statistics

chi-squared value: 119695.06464435393

p-value: 0

degree of freedom: 528

Interesting things to note:

Chi-squared value is very large due to the larger table size. With a p-value of 0 (most likely not exactly 0 but an extremely low number) it is pretty clear that we should reject the null hypothesis. Thus, we can conclude that the Publication Year and Item type are associated

Machine Learning Models

Random Forests

```
Mean absolute error:  0.43 degrees.
[0.38333333 0.925      0.         ... 0.         0.6        0.4        ]
```

```
Variable: PublicationYear      Importance: 0.37633
Variable: ItemType_Book: Ref Adult/YA Importance: 0.17513
Variable: ItemCount            Importance: 0.09484
Variable: ItemCollection_CA-Nonfiction Importance: 0.0104
Variable: ItemCollection_CA-Reference Importance: 0.00979
Variable: ItemCollection_NA-Nonfiction Importance: 0.00822
Variable: ItemCollection_CA1-Language Importance: 0.00811
Variable: ItemCollection_CA3-Fiction Importance: 0.00734
Variable: ItemCollection_AfAm - Nf Importance: 0.00717
Variable: ItemType_Book: Adult/YA Importance: 0.00693
Variable: ItemCollection_CA9-Biography Importance: 0.00686
Variable: ItemCollection_NA-Fiction Importance: 0.00667
Variable: ItemCollection_NA-Languages Collection Importance: 0.00643
Variable: ItemType_Book: Juv   Importance: 0.0063
```

Machine Learning Models

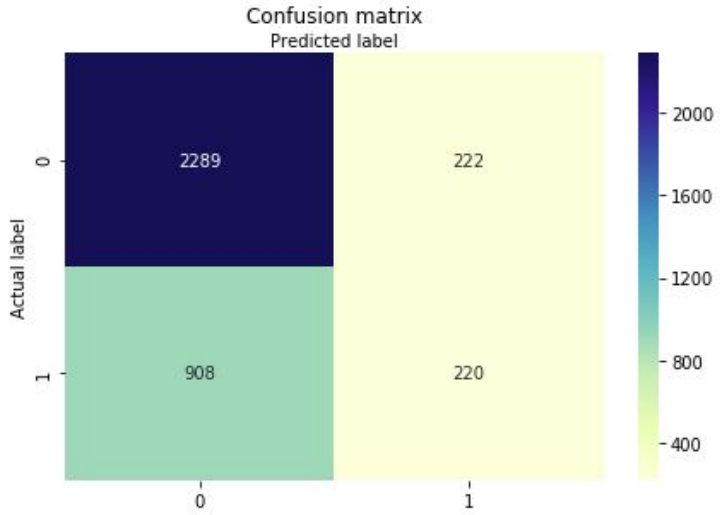
Decision Trees

Accuracy: 0.5221214619400935

*Refer to github code for implementation

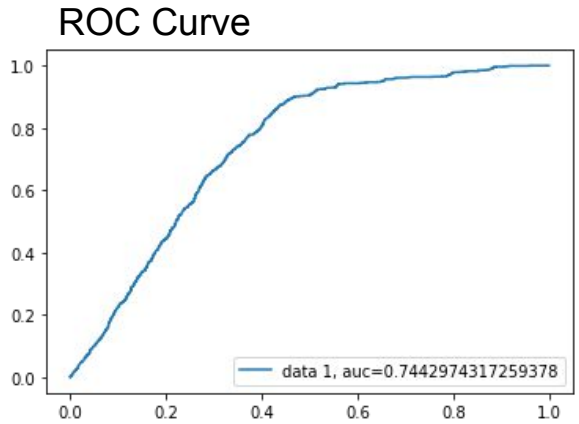
Machine Learning Models

Logistic Regression



Accuracy: 0.6894751305303655
Precision: 0.497737556561086
Recall: 0.1950354609929078

Accuracy higher than that of Decision Trees



Conclusions

- Publication Year of items play the biggest role in determining checkouts.
- Definite decrease in published material in the past few years.
- Books continue to be the most dominant form of checkout material even in the midst of cds/dvds/newer forms of library entertainment. This may be because latter items can be obtained from other sources.

Moving Forward

- Geographic locations of checkouts
- Inventory counts within each location
- Specific checkout times, trends during times of year?