Project 1

Evgeny Noi

January 15, 2020

1. Introduction

For my project with Seattle Public Library (SPL) I decided to visualize a journey of a book in space and time. Unfortunately, the provided dataset does not have any information on the readers or their places of residence, so I decided to employ a Gravity model prevalent in geography and spatial analysis as well as other probabilistic methods to assess the service areas of SPL branches and hypothesize about the trajectory of the book given random or pseudorandom processes.

2. Data Acquisition

2.1 SPL inventory

SPL publishes its data on City of Seattle Open Data Portal, where library inventory may be found and downloaded. The inventory file contains information on the resources owned by the library, including the bibNumber from the Library of Congress, as well as the ItemType and ItemCollection that could be used to geneate unique identifier to be joined to the spl_2016 database. Additionally, I found an ealier copy of the inventory on the Portal (2017) on kaggle.com. Logically, it was interesting to compare the prescription of books to branches from these two time frames: 2017 and 2020.

2.2 SPL branches

The information on branches is located on the official SPL website. Overall, there are 27 branches in Seattle. Their geographic coordinates can be downloaded from City of Seattle Open Data Portal.

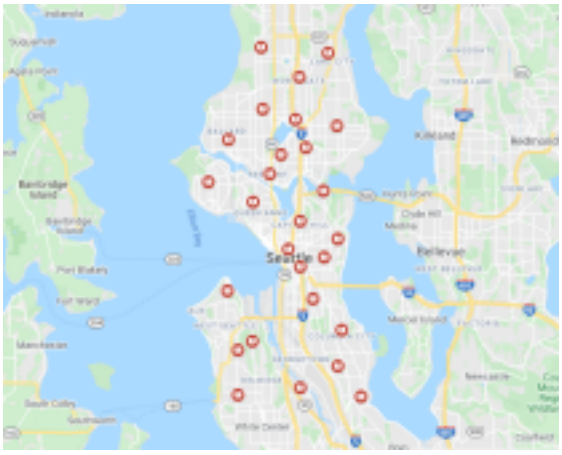

Figure 1: Branches of SPL library

2.3 SPL check-ins and check-outs

The SPL dataset is made available as part of the MAT259A course. MySQL Workbench was used to access data and run queries on the database.

3. Data Processing

There were five stages in data processing: 1. Downloading and pre-processing inventory data files 1. Inventory data from Kaggle (2017), denoted as $\mathcal{I}_{\infty}$, 2. Inventory data from Seattle Open Data Portal (2020), denoted as $\mathcal{I}_{\in}$, 2. Use 2017 inventory file as a baseline and keep only interesting records between the datasets. We can refer to this set as $\mathcal{I} \in \mathcal{I}_{\infty} \cap \mathcal{I}_{\in}$.

```
import pandas as pd

# df2 is inventory 2017, df is inventory 2020
df2['tycobib'] = df2.BibNum.astype('str') + df2.ItemCollection + df2.ItemType
df['tycobib'] = df.BibNum.astype('str') + df.ItemCollection + df.ItemType

# drop any records that have duplicates given unique identifier tycobib
dfu = df.drop_duplicates('tycobib', keep=False)
dfu2 = df2.drop_duplicates('tycobib', keep=False)

merger2 = pd.merge(left = dfu, right = dfu2[['tycobib', 'ItemLocation', 'FloatingItem']],
  on='tycobib', how='inner')
```

3. Creating unique identifier for merging inventory ($\mathcal{I}$) and check-ins/outs ($\mathcal{C}$)
    1. Because inventory files do not have a barcode or other unique identifiers for a join, I created a unique identifier for both inventory and check-ins/outs by concatenating *'bibNumber', 'collcode' / 'ItemCollection', and 'ItemType'*. See *Figure 2* for the formula. Overall, there are 10,900 items that could be merged based on the aforementioned criteria.
4. Running SQL query
    1. I copied the identifiers into a generated SQL query and ran it on the database to build the following query:

```
SELECT
  *
FROM spl_2016.inraw
WHERE cout>'2016-12-31' AND
  CONCAT(bibNumber,collcode,itemtype) in ('261cs9rarbk',
  -- < ... > here 10,900 identifiers were inserted
  '3304438canfacbk');
```

5. Connecting branches abbreviations to actual branch locations was done manually. Below, there are top 10 branches with highest number of items.

Branch	Number of items
CENTRAL	7161
SOUTHWEST	410
NORTHEAST	344
LAKE CITY	304
BALLARD	275
DOUGLASS-TRUTH	261
WEST SEATTLE	219
NORTHGATE	204
CAPITOL HILL	201
SOUTH PARK	190

Q2 fx Σ = =CONCAT("'", B2, J2, I2, "'", ",")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		BibNum	Title	Author	ISBN	Publicatio	Publisher	Subjects	ItemType	ItemColle	FloatingI	ItemLoca	ReportDa	ItemCou	locbib	tycobib	
2	0	261	American	Holt, Alfred Hubbar		1969	Gale Rese	Names G	arbk	cs9r		cen	###	1	cen2610	261cs9rarbk	'261cs9rarbk',
3	1	813	Central P	Johnston, Nancy		[1968]	Sierra Clu	Central P	acbk	canf		cen	###	1	cen8130	813canfacbk	'813canfacbk',
4	2	2524	The Virgi	Dowdey, Clifford, 1		[1969]	Little, Bro	Carter Ro	acbk	canf		cen	###	1	cen2524	2524canfacbk	'2524canfacbk',
5	3	3119	The light	MacDonald, George		[1969]	Farrar, St	Fairy tale	jcbk	ncfic		wts	###	1	wts3119	3119ncficjcbk	'3119ncficjcbk',
6	4	4439	An encyc	Franklyn,	8E+07	[1969]	Pergamo	Heraldry	arbk	cs9r		cen	###	1	cen4439	4439cs9rarbk	'4439cs9rarbk',
7	5	4764	Pearce pi	Pearce, Marvin J., 1		[1969]		Pierce far	arbk	cs9r		cen	###	1	cen4764	4764cs9rarbk	'4764cs9rarbk',
8	6	4974	History o	Morrison, Leonard		1971	Privately	Kimball f	arbk	cs9r		cen	###	1	cen4974	4974cs9rarbk	'4974cs9rarbk',
9	7	6212	Iggie's ho	Blume, Ju	0134508	[1970]	Bradbury	Race rela	jcbk	ncfic		cap	###	1	cap6212	6212ncficjcbk	'6212ncficjcbk',
10	8	6229	A financi	Myers, M	2.31E+08	1970	Columbia	Finance P	acbk	canf		cen	###	1	cen6229	6229canfacbk	'6229canfacbk',
11	9	6504	Quakeris	Carroll, Kenneth L. (		[1970]	Maryland	Society o	arbk	cs9r		cen	###	1	cen6504	6504cs9rarbk	'6504cs9rarbk',

Figure 2: Generating unique identifiers for joining inventory and check-ins/outs

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	itemNumber	bibNumber	cout	cin	collcode	itemtype	barcode	title	callNumber	deweyClass	subj
2	83770886	767844	6212	2017-02-09 18:12:00	2017-03-11 16:34:00	ncfic	jcbk	10013772867	Iggies house	J BLUME		NULL
3	84681966	767844	6212	2017-05-01 16:11:00	2017-05-21 16:10:00	ncfic	jcbk	10013772867	Iggies house	J BLUME		NULL
4	83254746	1693272	33200	2017-01-04 14:26:00	2017-01-22 16:32:00	canf	acbk	100489012	American slavery American freedom th	975.5 M821A	975.5	NULL
5	84686680	166431	46636	2017-05-01 15:31:00	2017-05-22 12:51:00	nchol	jcbk	10007090375	Humbug rabbit	E BALIAN		NULL
6	83408679	570489	70465	2017-01-04 19:29:00	2017-02-02 19:03:00	naaab	acbk	10012382163	Nigger an autobiography	B G86216		NULL
7	84267864	570489	70465	2017-03-28 12:35:00	2017-04-19 11:24:00	naaab	acbk	10012382163	Nigger an autobiography	B G86216		NULL
8	86132367	570489	70465	2017-08-24 16:00:00	2017-09-12 15:35:00	naaab	acbk	10012382163	Nigger an autobiography	B G86216		NULL
9	86165829	570489	70465	2017-09-13 14:50:00	2017-09-15 08:07:00	naaab	acbk	10012382163	Nigger an autobiography	B G86216		NULL
10	86169666	1130295	75901	2017-06-30 14:37:00	2017-09-15 13:53:00	canf	acbk	10004684923	Perspective for artists	742 L272P	742	NULL

Figure 3: The resulting table of inventory

4. Results and Exploratory Data Analysis

The resulting table is illustrated in *Figure 2* and has 18,005 records for check-ins/outs. This data will be further used to visualize the journey of a book in space-time using probabilistic models and p5.js.

L2 fx Σ = NULL

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	itemNumber	bibNumber	cout	cin	collcode	itemtype	barcode	title	callNumber	deweyClass	subj
2	82974281	3509394	2582482	2016-12-31 14:58:00	2016-12-31 16:18:00	ncenf	jcbk	10069913829	Pharaohs boat	J932.012 W439P 2009	932.012	NULL
3	82983278	3373447	2552116	2016-12-31 11:52:00	2017-01-02 13:45:00	nadvd	acdvd	10063533938	It happened on Fifth Avenue	DVD IT HAPP		NULL
4	82991434	3326994	2539876	2016-12-31 14:21:00	2017-01-02 16:41:00	nadvdnf	acdvd	10062636195	1000 places to see before you die Colle	DVD 910.2 A14 2008	910.2	NULL
5	82991697	4924988	2929834	2017-01-02 15:39:00	2017-01-02 16:44:00	cafic	acbk	10079457031	Prodigal summer a novel	FIC KINGSOL 2013		NULL
6	82994919	3878423	2669652	2016-12-31 13:02:00	2017-01-02 17:50:00	cacd	accd	10068206803	King of the beach	CD 782.42166 W36K	782.42166	NULL
7	82997417	1729394	1786481	2017-01-02 19:47:00	2017-01-02 19:52:00	nchol	jcbk	10033272575	Snowball	E CREWS		NULL
8	82997528	4866121	2939717	2016-12-31 12:59:00	2017-01-03 10:03:00	ccser	jcbk	10079430574	Wild born	J MULL		NULL
9	82998053	5146325	2990608	2017-01-02 17:02:00	2017-01-03 10:57:00	ncdvd	jcdvd	10082112730	Berenstain bears Carnival coasters	DVD J BERENST		NULL
10	82998741	5636305	2397260	2017-01-02 15:29:00	2017-01-03 11:34:00	nacd	accd	10059120732	Hotel California	CD 782.42166 Ea37H	782.42166	NULL
11	82999967	5773056	3166114	2016-12-31 11:36:00	2017-01-03 12:39:00	cadvd	acdvd	10088701056	Married with children Season ten	DVD MARRIED Season 10		NULL

Figure 4: The resulting table of check-ins/outs

Furthermore, preliminary visualizations of volume of books and other items is explored via a simple map depicted on Figure 4. We can see that most books are located in central branch.

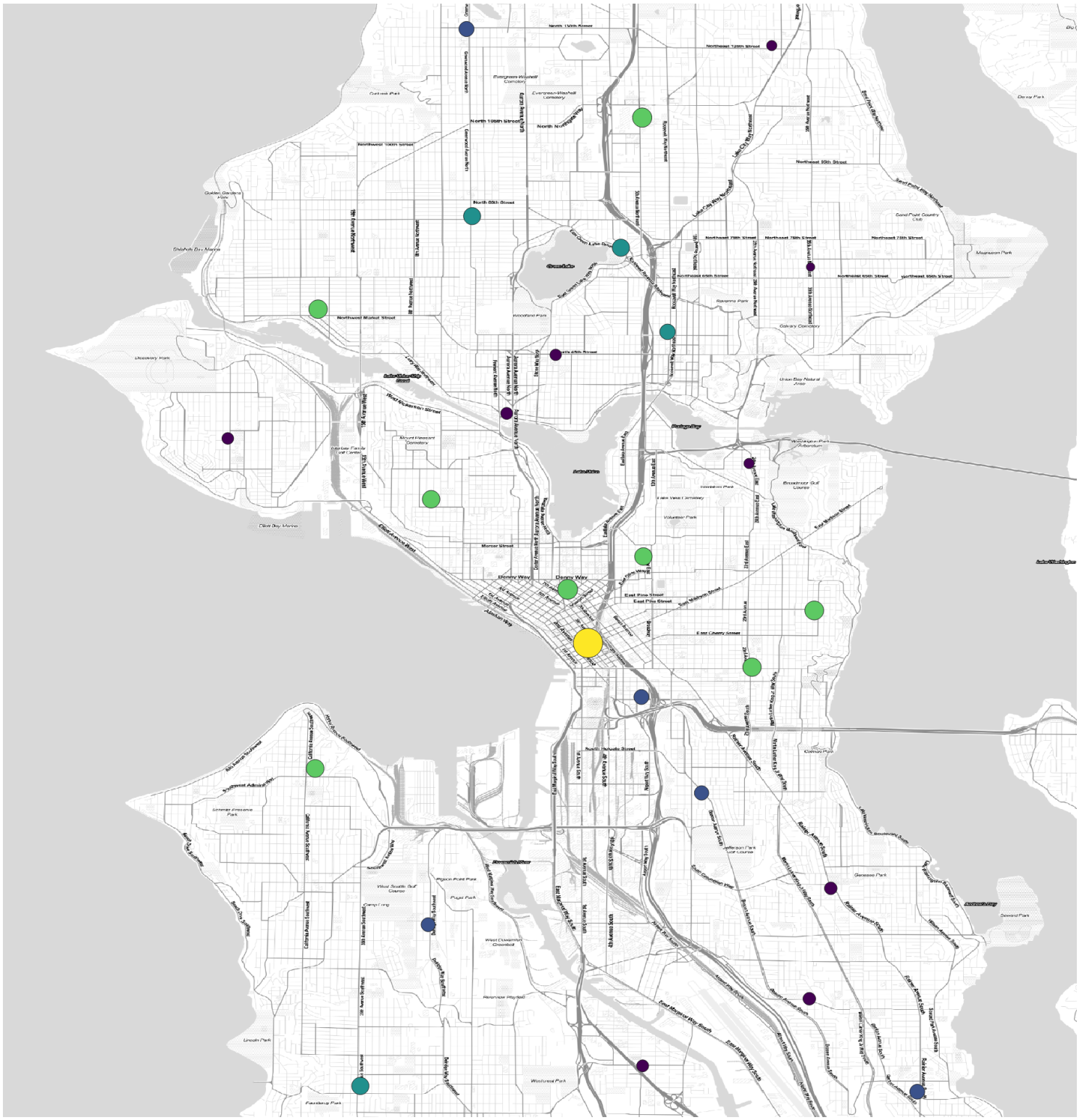

Figure 5: Number of items by branch location