MAT 259

Project 1 Knowledge Discovery

Meilin Shi

Introduction

I'm interested in the checkouts in Language within the Dewey Class 400. From the class, I picked **ten** specific languages: English (420), German (430), French (440), Italian (450), Spanish (460), Russian (491.7), Arabic (492.7), Chinese (495.1), Japanese(495.6) and Korean (495.7) that fall within my personal interest. I would like to see if there is a **relative popularity** in language learning at Seattle Public Library and the checkout patterns over time. I was also curious about **for how long people usually keep these items** and could a pattern be discerned?

Query (on Language Learning Interests)

```
SELECT
   YEAR(cout) AS years,
   COUNT(IF(deweyClass >= 420 AND deweyClass < 430, 1, NULL)) AS 'English',
   COUNT(IF(deweyClass >= 430 AND deweyClass < 440, 1, NULL)) AS 'German & related',
   COUNT(IF(deweyClass >= 440 AND deweyClass < 450, 1, NULL)) AS 'French & related',
   COUNT(IF(deweyClass >= 450 AND deweyClass < 460, 1, NULL)) AS 'Italian & related',
   COUNT(IF(deweyClass >= 460 AND deweyClass < 470, 1, NULL)) AS 'Spanish & related',
   COUNT(IF(deweyClass >= 491.7 AND deweyClass < 491.8, 1, NULL)) AS 'Russian',
   COUNT(IF(deweyClass >= 492.7 AND deweyClass < 492.8, 1, NULL)) AS 'Arabic',
   COUNT(IF(deweyClass >= 495.1 AND deweyClass < 495.2, 1, NULL)) AS 'Chinese',
   COUNT(IF(deweyClass >= 495.6 AND deweyClass < 495.7, 1, NULL)) AS 'Japanese',
   COUNT(IF(deweyClass >= 495.7 AND deweyClass < 495.8, 1, NULL)) AS 'Korean'
FROM
   spl_2016.outraw
WHERE
   deweyClass >= 420 AND deweyClass < 495.8
      AND YEAR(cout) BETWEEN 2006 AND 2018
```

```
GROUP BY YEAR(cout)
ORDER BY YEAR(cout);
```

Query (on average time people keep the items)

```
SELECT
   class, years, AVG(TIMESTAMPDIFF(DAY, cout, cin)) AS AVG_TIME
FROM
   (SELECT
      SUBSTRING(deweyClass, 1, 5) AS class,
      #SUBSTRING(deweyClass, 1, 2) AS class for deweyClass 420 to 470 (I manually changed these b/c I haven't found a way to fit all of them together)
         YEAR(cout) AS years,
         cin,
         cout,
         TIMESTAMPDIFF(DAY, cout, cin)
   FROM
      spl_2016.inraw
   WHERE
      YEAR(cout) BETWEEN 2006 AND 2018
         AND TIMESTAMPDIFF(DAY, cout, cin) > 0
         #AND deweyClass >= 420 AND deweyClass < 470
         AND (deweyClass >= 491.7 AND deweyClass < 491.8
         OR deweyClass >= 492.7 AND deweyClass < 492.8
         OR deweyClass >= 495.1 AND deweyClass < 495.2
         OR deweyClass >= 495.6 AND deweyClass < 495.7
         OR deweyClass >= 495.7 AND deweyClass < 495.8)
   GROUP BY years , class , cin , cout) AS aTable
GROUP BY class, years;
```

Result & Analysis

- Language learning interest from 2006 to 2018

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

#	years	English	German & related	French & related	Italian & related	Spanish & related	Russian	Arabic	Chinese	Japanese	Korean
1	2006	14486	1770	3890	1476	5820	748	328	2056	1555	222
2	2007	15113	1476	3399	1332	6221	695	334	2281	1245	246
3	2008	19565	1939	4233	1483	9025	861	457	3240	1617	349
4	2009	21434	1855	4447	1745	9704	813	578	3639	1722	437
5	2010	21444	1976	4828	1864	9393	1038	652	3879	1822	468
6	2011	19617	1773	4427	2062	8500	825	664	3633	1604	428
7	2012	16941	1665	3785	1583	7251	709	599	2987	1425	434
8	2013	16716	1573	3713	1495	7083	690	560	2621	1259	360
9	2014	13355	1312	3211	1536	5819	495	464	2229	1080	270
10	2015	12071	1367	3056	1291	4768	346	383	1745	1078	218
11	2016	10498	1153	2407	1105	4078	372	380	1460	896	219
12	2017	9453	956	2325	1061	3680	354	263	1336	851	268
13	2018	5772	714	1597	694	2457	252	170	742	628	203

- Average time people kept the items within English, German, French, Italian, Spanish

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

#	class	years	AVG_TIME
1	42	2006	58.7566
2	42	2007	55.6979
3	42	2008	49.6880
4	42	2009	46.8198
5	42	2010	40.3099
6	42	2011	29.2852
7	42	2012	26.0969
8	42	2013	25.7640
9	42	2014	27.6207
10	42	2015	26.7913
11	42	2016	26.9145
12	42	2017	25.9807
13	42	2018	27.7227
14	43	2006	56.1271
15	43	2007	51.3601
16	43	2008	49.8183
17	43	2009	53.4151

Result 3

Action Output

#	Time	Action	Message	Duration / Fetch
1	00:22:09	SELECT class, years, AVG(TIMESTAMPDIFF(DAY, cout, c...	65 row(s) returned	38.568 sec / 0.0000...

Running time for this query is relatively long, 38 seconds. I need to find a way to rearrange the table to make the class shown as rows under one descriptive name instead of numbers.

- Average time people kept the items within Russian, Arabic, Chinese, Japanese, Korean

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

#	class	years	AVG_TIME
1	491.7	2006	70.8342
2	491.7	2007	72.5580
3	491.7	2008	59.2756
4	491.7	2009	58.8150
5	491.7	2010	46.0711
6	491.7	2011	33.9357
7	491.7	2012	26.6723
8	491.7	2013	27.7061
9	491.7	2014	30.2255
10	491.7	2015	31.4726
11	491.7	2016	28.3440
12	491.7	2017	27.9447
13	491.7	2018	28.8245
14	492.7	2006	50.4817

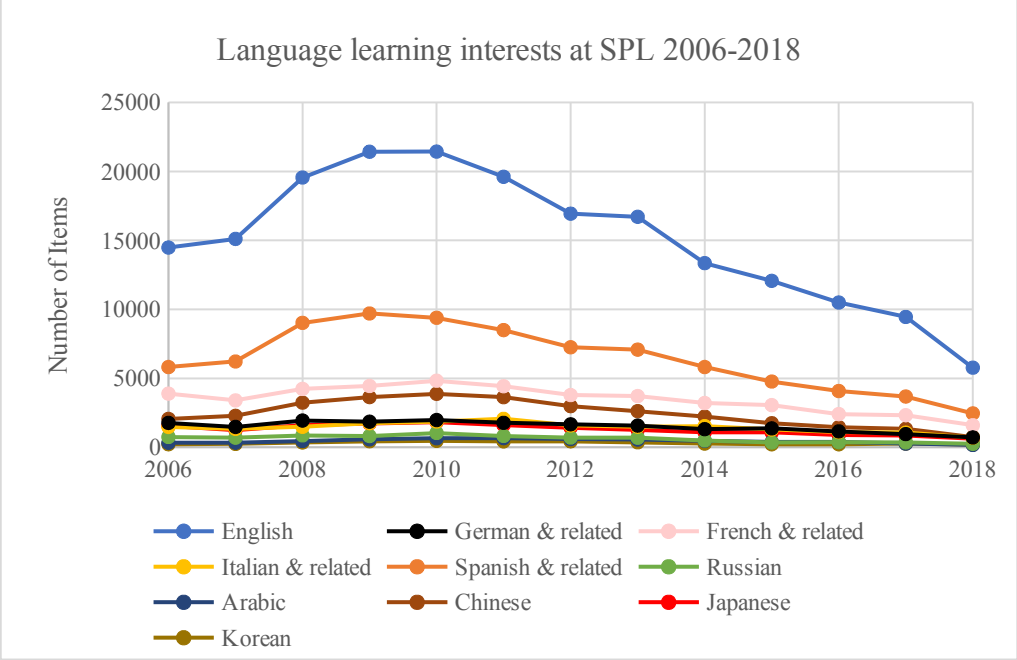
Result 1

Action Output

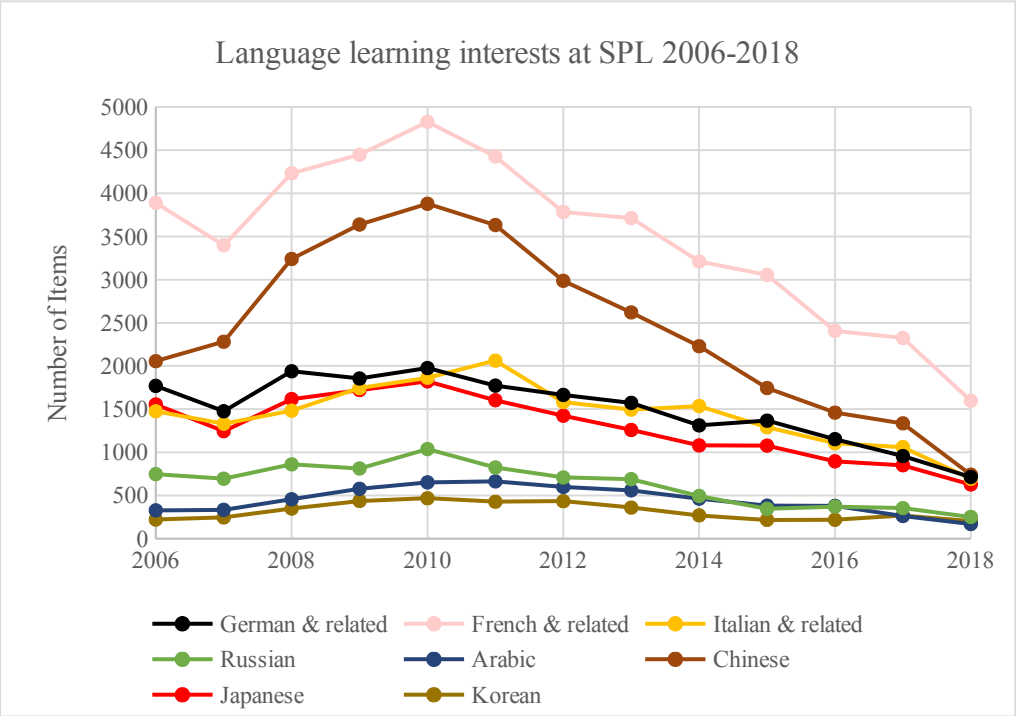
	#	Time	Action	Message	Duration / Fetch
✓	1	00:09:47	SELECT YEAR(cout) AS years, COUNT(IF(deweyClass ...	13 row(s) returned	25.792 sec / 0.0000...
✓	2	00:12:17	SELECT class, years, AVG(TIMESTAMPDIFF(DAY, cout, c...	65 row(s) returned	64.662 sec / 0.0000...

Same problem as above and this query takes twice as long. I need to find a way to express the non-continuous class more efficiently (maybe start sub-string from position 3).

- Rearrange and visualize the data in excel:



- Zoom-in version of languages other than English and Spanish related:

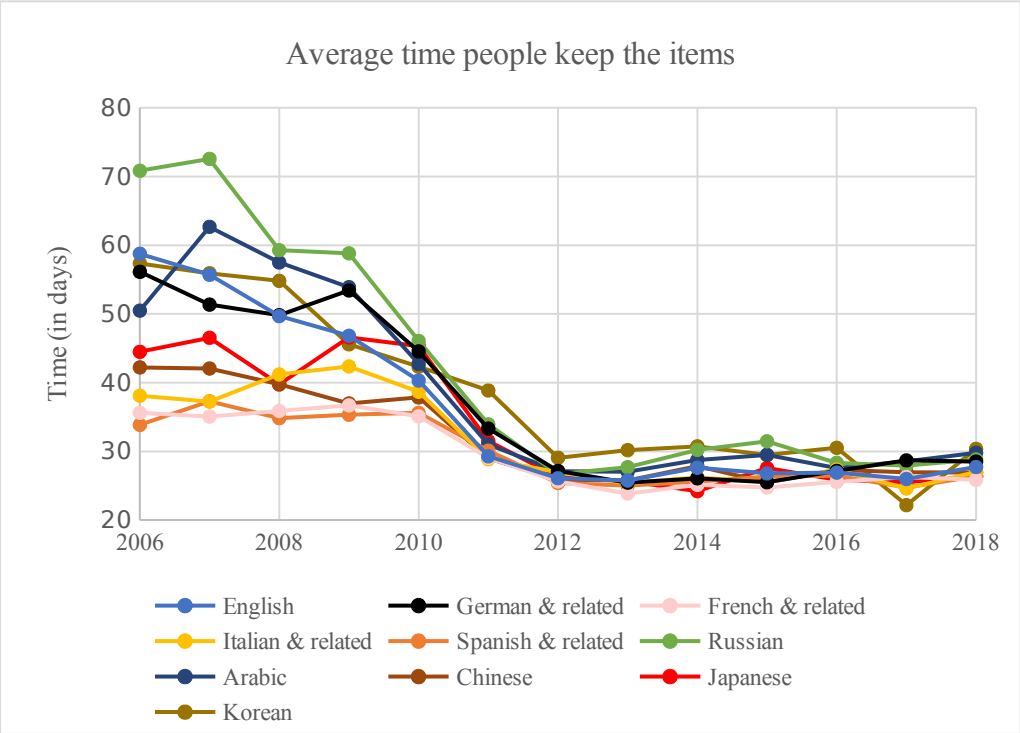


From this we can basically categorize the relative popularity as:

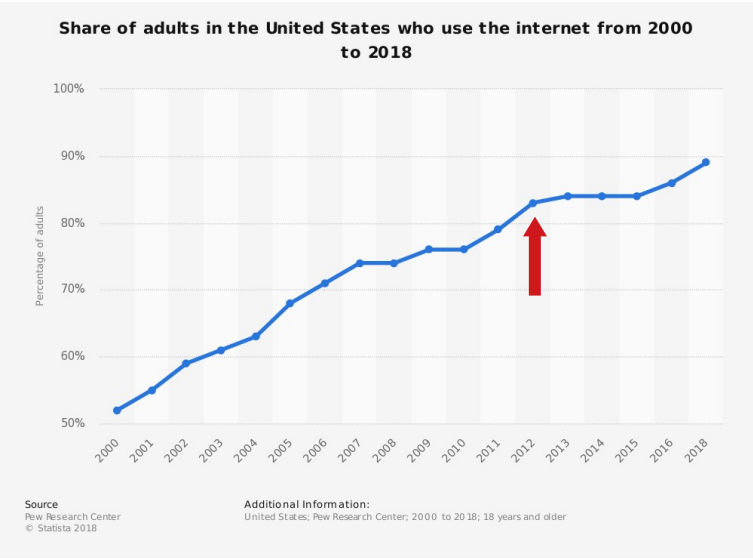
- English and Spanish (above 5000 checkouts)
- French and Chinese (2000-5000)
- German, Italian and Japanese (1000-2000)
- Russian, Arabic and Korean (0-1000)

Interesting pattern being discerned that overall checkouts are decreasing over time with a peak around year 2009/2010.

- Visualize the average time people kept the items



- Surprisingly, Russian has the highest average time among all retrieved data. But there is a possibility of outliers. Need to inspect the data specifically.
- Another interesting pattern spotted is that all of the ten language checkouts fall into the same average time for around 25-30 days from 2012 to 2018. People seem to have lost their interest in language books (CDs and other items) as they have more access to the internet for these items/information. According to the figure below, 2012 did seem as a turning point for internet user percentage – a huge jump from 2010 to 2012.



Retrieved from:

https://www.statista.com/statistics/185700/percentage-of-adult-internet-users-in-the-united-states-since-2000/