Seasonal trends in Seattle Public Library

Ilia Nikiforov

Seattle, being a city in the Northern part of the United States, usually experiences a good variation in weather conditions over the year. It is also known that Seattle experiences quite a lot of rain. For this report, I am interested in finding out whether activity in Seattle Public Library has some sort of seasonal component, and whether or not it is connected to the changes in weather conditions; or other factors, such as school breaks.

Seasonal trends

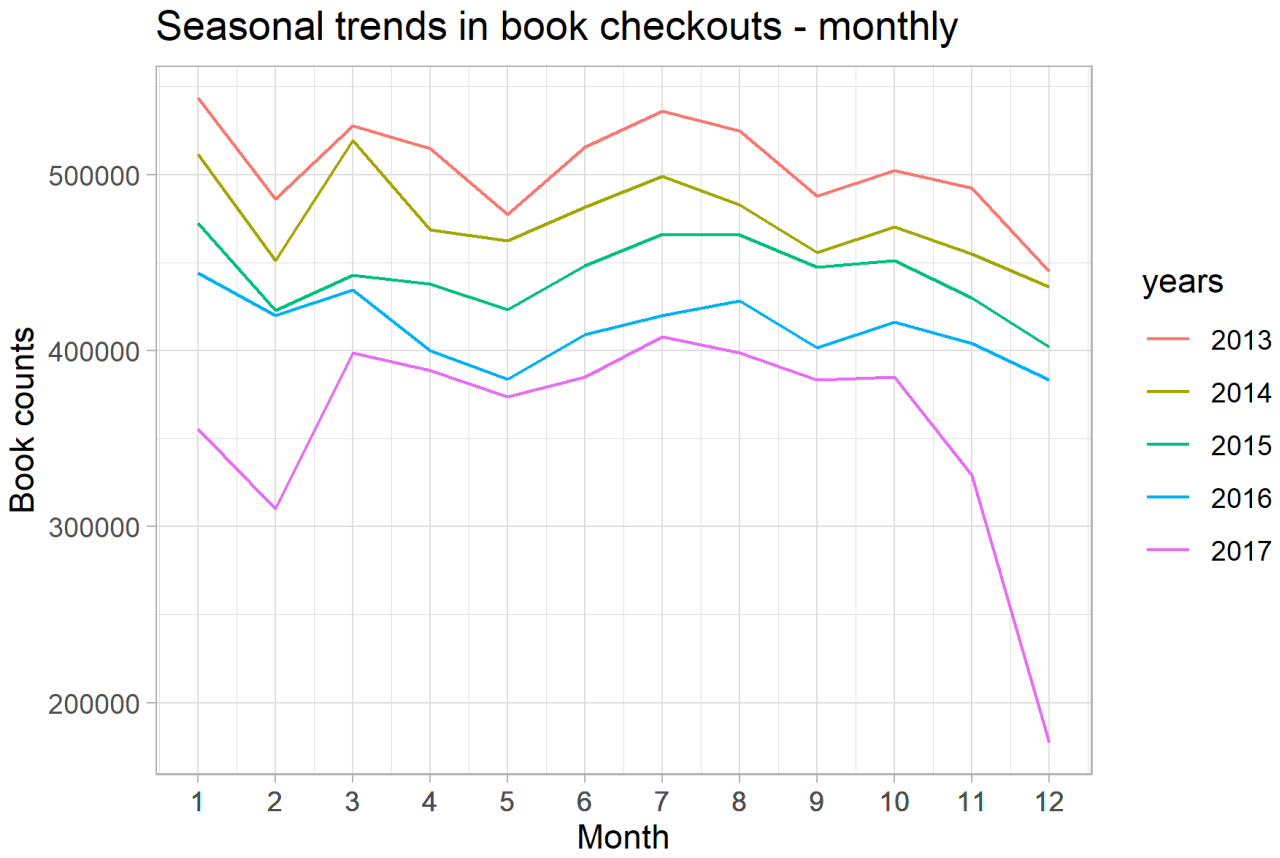
I begin exploring seasonal trends in book checkout in SPL by extracting summary data on checkouts between 2013 and 2017 included, aggregated by years and months separately. The following query makes use of the extract() command in order to create separate variables for a year and a month to make subsequent visualization easier using R.

Query – monthly book counts:

```
select

extract(year from cout) as years,

extract(month from cout) as months,

count(*) as counts

from spl_2016.inraw

where year(cout) >= 2013 and year(cout) < '2018'

group by 1,2

order by 1,2
```

Results:

The results are stored in the “seasons_split_monthly.csv” and visualized using ggplot2 in R Studio below. There appears to be some sort of monthly trend, although with different intensity for each year. Specifically, there is typically a decline in February, an increase in March, a decline by May, an increase by July, a decrease by September, followed by another decrease by the end of the year. The year of 2017 has the most dramatic seasonal changes.



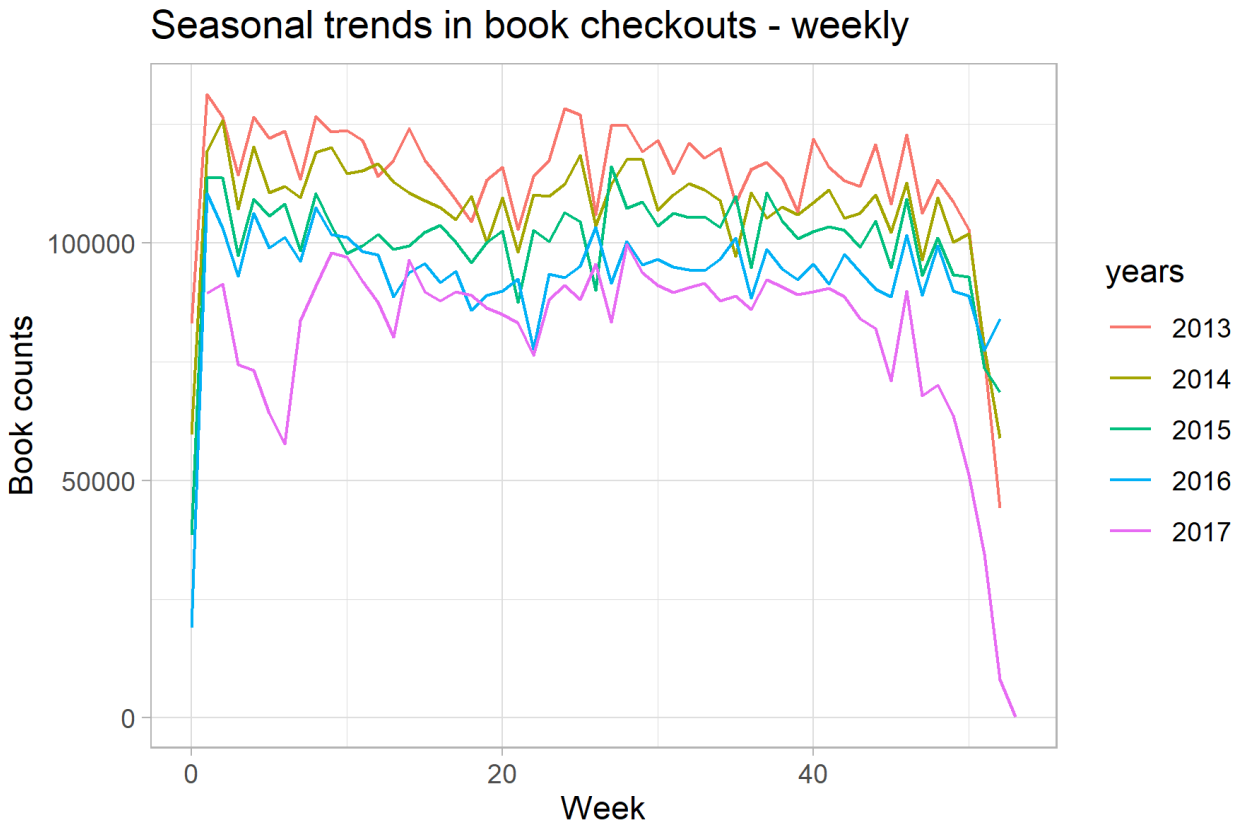
As the next step, I will look for a more granular result by aggregating checkout data by weeks rather than months. In the following query I will once again use extract() command to structure my data in the "long" format.

Query - weekly book counts:

```
select

extract(year from cout) as years,

extract(week from cout) as weeks,

count(*) as counts

from spl_2016.inraw

where year(cout) >= 2013 and year(cout) < '2018'

group by 1,2

order by 1,2
```

Result:

The result is stored in "seasons_split_weekly.csv" and visualized below. More granularity in this case resulted in more visual noise, but there are still more or less clear trends similar to those observed in the monthly data. That is, there are clear spikes approximately around weeks 25 and 50. Interestingly, years 2013-2016 have a spike in the beginning of the year (weeks 4-7), whereas year 2017 has a drop in book checkouts in the same timeframe.



Next, I will try to visually explore any correlational patterns between book checkouts and weather data. For weather data, I visited NOAA Online Weather Data service[1] and manually scraped the data on average monthly temperatures (in Fahrenheit) and average monthly precipitation (in inches). NOAA weather data is only available by month; therefore, I will take a step back in terms of granularity of book data and once again aggregated SPL data by months. For the next query, I will create separate columns with counts for each year as opposed to a long format of the data. I will do so to make it easier to assemble a data frame in R that would combine book data and weather data obtained from NOAA.

Query – monthly data (years as columns)

select

extract(month from cout) as months,

[1] https://www.weather.gov/wrh/climate?wfo=sew

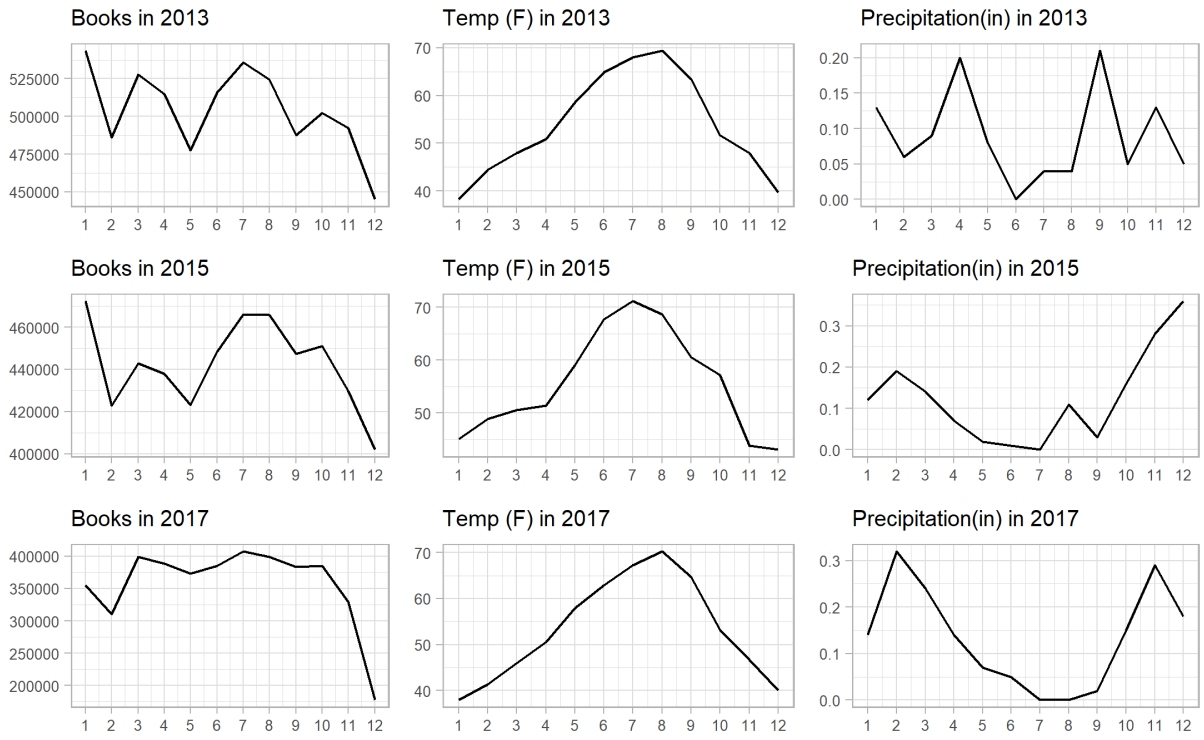
```
count(if(year(cout)='2013',1,NULL)) as 'counts_2013',

count(if(year(cout)='2014',1,NULL)) as 'counts_2014',

count(if(year(cout)='2015',1,NULL)) as 'counts_2015',

count(if(year(cout)='2016',1,NULL)) as 'counts_2016',

count(if(year(cout)='2017',1,NULL)) as 'counts_2017'

from spl_2016.inraw

where year(cout) >= 2013 and year(cout) < '2018'

group by 1

order by 1
```

Result:

The result is stored in “seasons_split_monthly_colyears.csv”. Once I obtained this result, I manually added the variables with temperature and precipitation data for years 2013, 2015, and 2017 in R studio. I then used ggplot2 to visualize them together. Since the philosophy of ggplot2 developers is against using 2 separate y axis on the same plot, I had to plot book counts, temperature, and precipitation on separate graphs as seen below.

The graphs demonstrate a few things:

1. Visitors of SPL tend to check out books more during hotter summer months. They also seems to check out less books towards the end of the year, as temperature goes down. On the other hand, in the beginning of the year (Jan, Feb, Mar), when temperature is still low, there seems to be no association with book checkouts – they fluctuate and sometimes rise regardless of the weather.
2. Average precipitation does seem to bother SPL visitors to a certain degree, but it is not clear. The drops in book checkout seem to correspond to increased precipitation (and possibly, rain or snowfall). For instance, February 2017 had increased precipitation and significant drop in books. On the other hand, a drop in February occurred in 2013 and 2015 when there was not much rainfall. Overall, it seems that extreme increases in precipitation do correspond to book checkout dynamics sometimes, but overall SPL readers seem to be fairly resilient to rain and snow.



My next question is whether certain item types have more pronounced trend dynamics. Specifically, I expect that content for kids (marked as itemtype beginning with "jc…" in the data) would display certain patterns. I expect, for instance, that seasonality in juvenile content can be connected to school break periods – specifically in the Summer, when school is not in session. This could go either way: kid literature might be in high demand since kids have more free time; or it can be in low demand since kids would rather spend their free time doing anything but reading.

For the next query, I will use case when() command to classify every itemtype beginning with "ac" as adult category, every item beginning with "jc" as juvenile category, and everything else as "other".

Query – age categories:

```
select

extract(year from cout) as years,

extract(month from cout) as months,

case when itemtype like 'ac%' then 'Adult'

        when itemtype like 'jc%' then 'Juvenile'

   else 'Others' end as itemtypes,

count(*) as counts
```

from spl_2016.inraw

where year(cout) >= 2013 and year(cout) < '2018'

group by 1,2,3

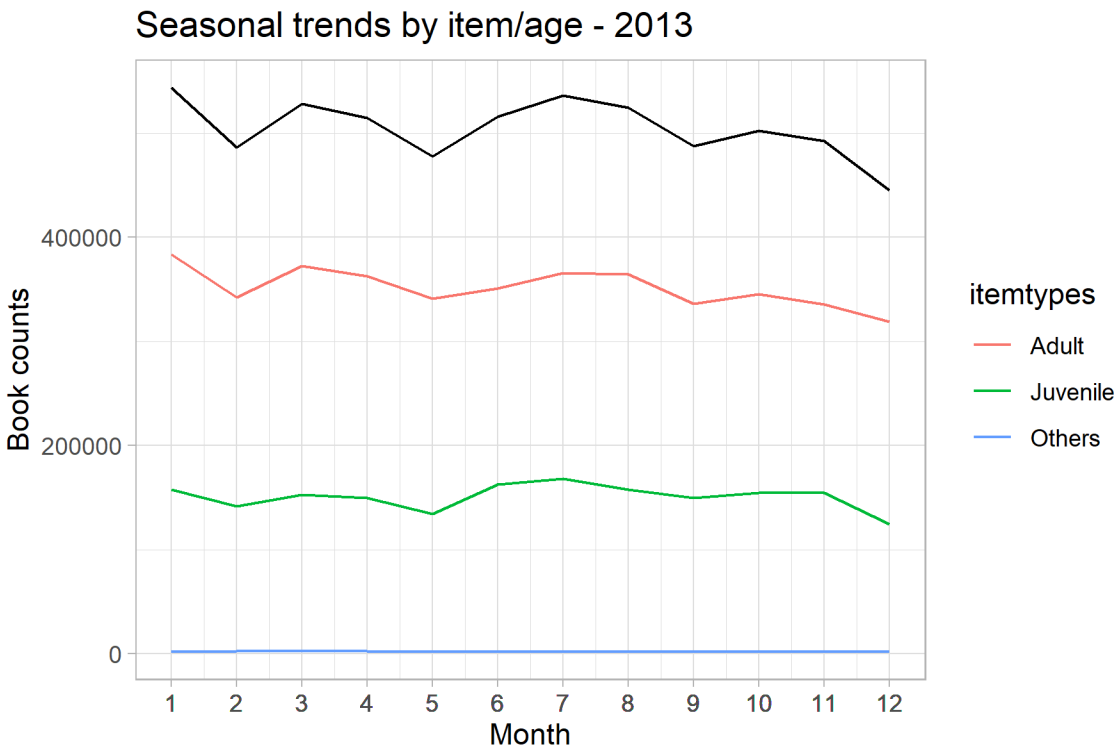
order by 1,2

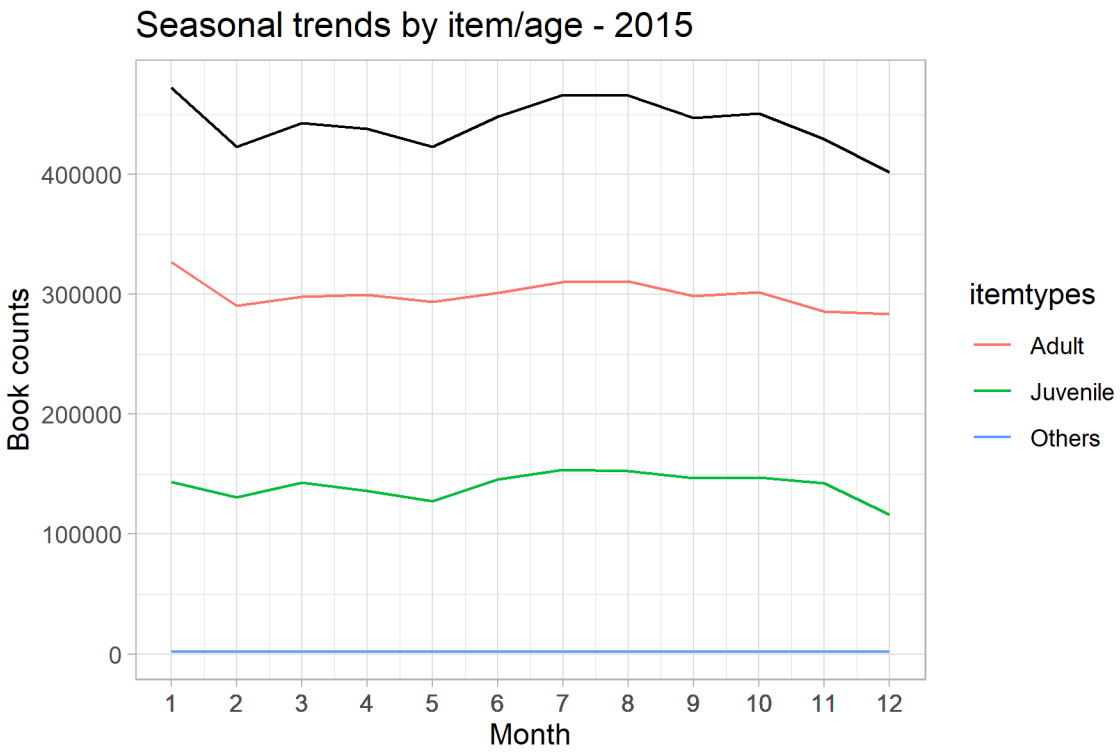
Result:

The result is stored in “seasons_split_plus_itemtypes_by_age.csv”. The graphs below illustrate the breakdown by age for selected years. The black line in each plot is the total counts for the year – the original seasonal line, for reference.

As we can see, there seems to be only minimal difference between seasonal trends for adults and kids. It is noticeable, however, that the increase in June has a slightly steeper curve for juvenile literature, as opposed to adult literature. Perhaps it is an evidence that juvenile item sections become relatively more active during summer break, compared to adults who work all year round anyway. The change, however, is minimal.

Interestingly, monthly trend becomes less pronounced once we break total counts down into different categories. It appears to be a form of Simpson’s paradox, where the effect present in high-level data starts to diminish, once we obtain more granular, detailed data.





Next, I will explore this pattern with another query, with a more granular classification, breaking down adult and juvenile items into specific most popular types (such as books, dvd, and cd). I will use a similar syntax.

Query – various itemtypes

```
select

extract(year from cout) as years,

extract(month from cout) as months,

case when itemtype='acbk' then 'Adult book'

        when itemtype='acdvd' then 'Adult DVD'

   when itemtype='accd' then 'Adult CD'

   when itemtype='jcbk' then 'Juvenile book'

   when itemtype='jcdvd' then 'Juvenile DVD'

   when itemtype='jccd' then 'Juvenile CD'

   else 'Others' end as itemtypes,
```

count(*) as counts

from spl_2016.inraw

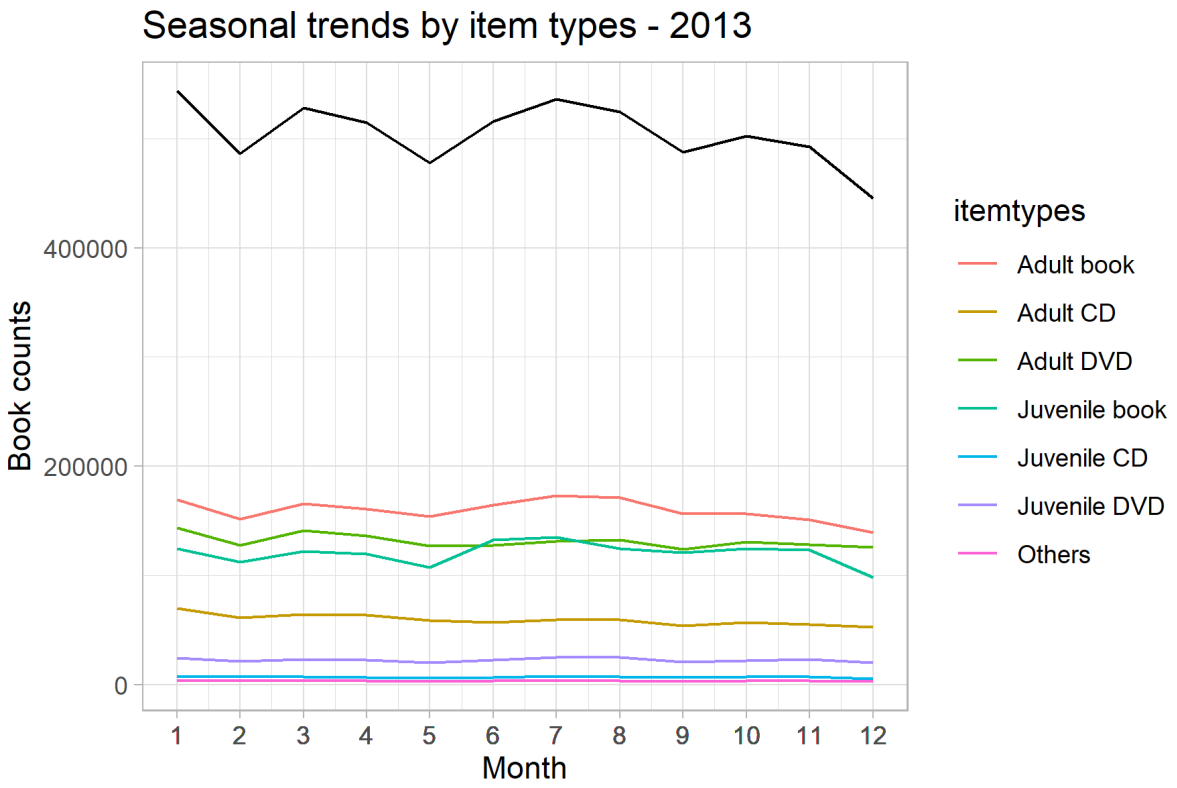
where year(cout) >= 2013 and year(cout) < '2018'

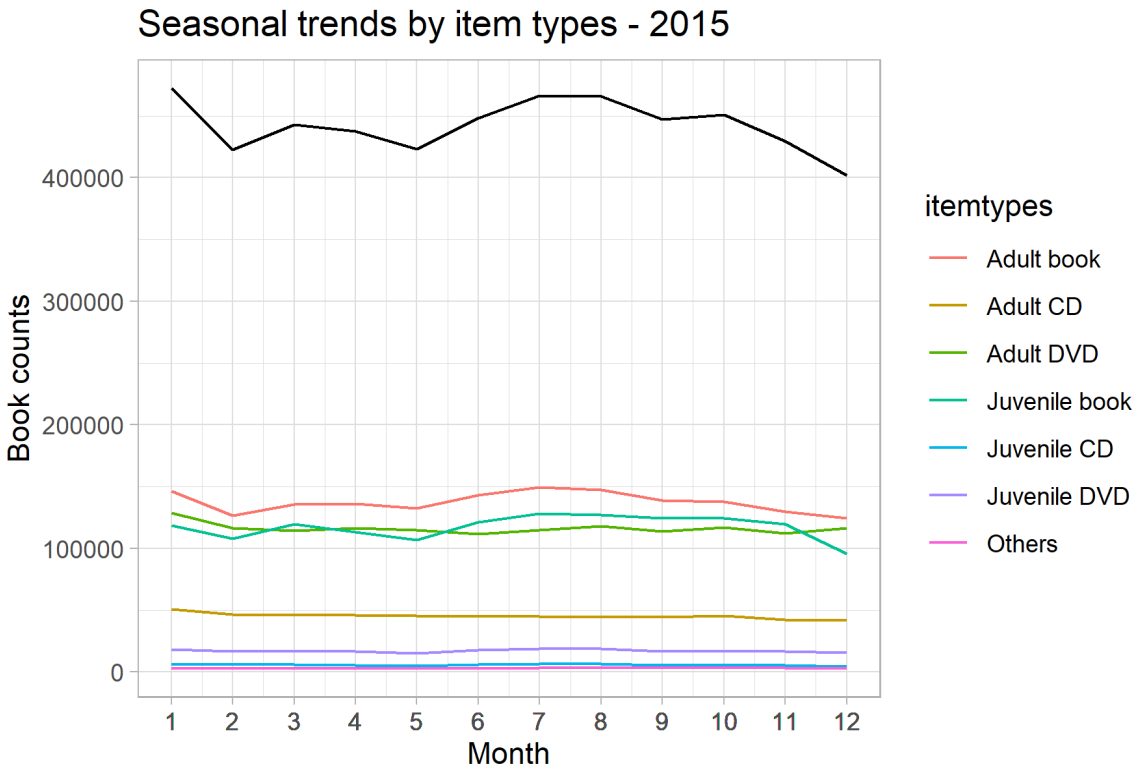
group by 1,2,3

order by 1,2

Result:

The result is stored in “seasons_split_plus_itemtypes.csv”. As expected, the more detailed view of the data we get, the less pronounced seasonal trends become. As evident in the following graphs, the trend line flattens significantly for most items. An obvious exception to that is the category of juvenile books, which is clearly seasonal, still.





Conclusion

Seattle Public Library book checkouts are not consistent throughout the year and fluctuate over time, with January, mid-Spring, and Summer being the most active months, whereas the Fall and December are usually the slowest. There is limited correlation with temperature, and very limited association with precipitation. Seasonal trends and monthly fluctuations become less and less apparent, once we break down high-level data into smaller categories, such as item types aggregated by age, or item types in general. The notable exception is kids' literature (which includes comic books) that retains general seasonal pattern for the given year. We can expect that juvenile literature section gets most active during the school break periods.

Appendix

1. NOAA Online Weather Data: https://www.weather.gov/wrh/climate?wfo=sew
2. R Studio script for analysis and plots: https://drive.google.com/file/d/1jkVbBcpUx28Kt8vIpMGs_w9_ePQ3DAmZ/view?usp=sharing