MAT 265
Assignment 4
Ilia Nikiforov

Daily seasonal trends in Seattle Public Library

Abstract

In this report, I dive deeper into the exploration of seasonal trends in the SPL, this time using daily Data from March 2015. I explore variables such as total checkouts, checkouts by type, diversity of titles, diversity of types, and the weight of a top performing item each day against temperature and precipitation. I use both visualization and a correlation matrix. The conclusion: there seems to be no seasonal trends based on daily weather in SPL, contrary to last week's findings.

Seasonal trends

This week I am adding more granularity and depth into my seasonal trend inquiry from the last week. I focus on March 2015 and aggregate the data by day. I start with total counts.

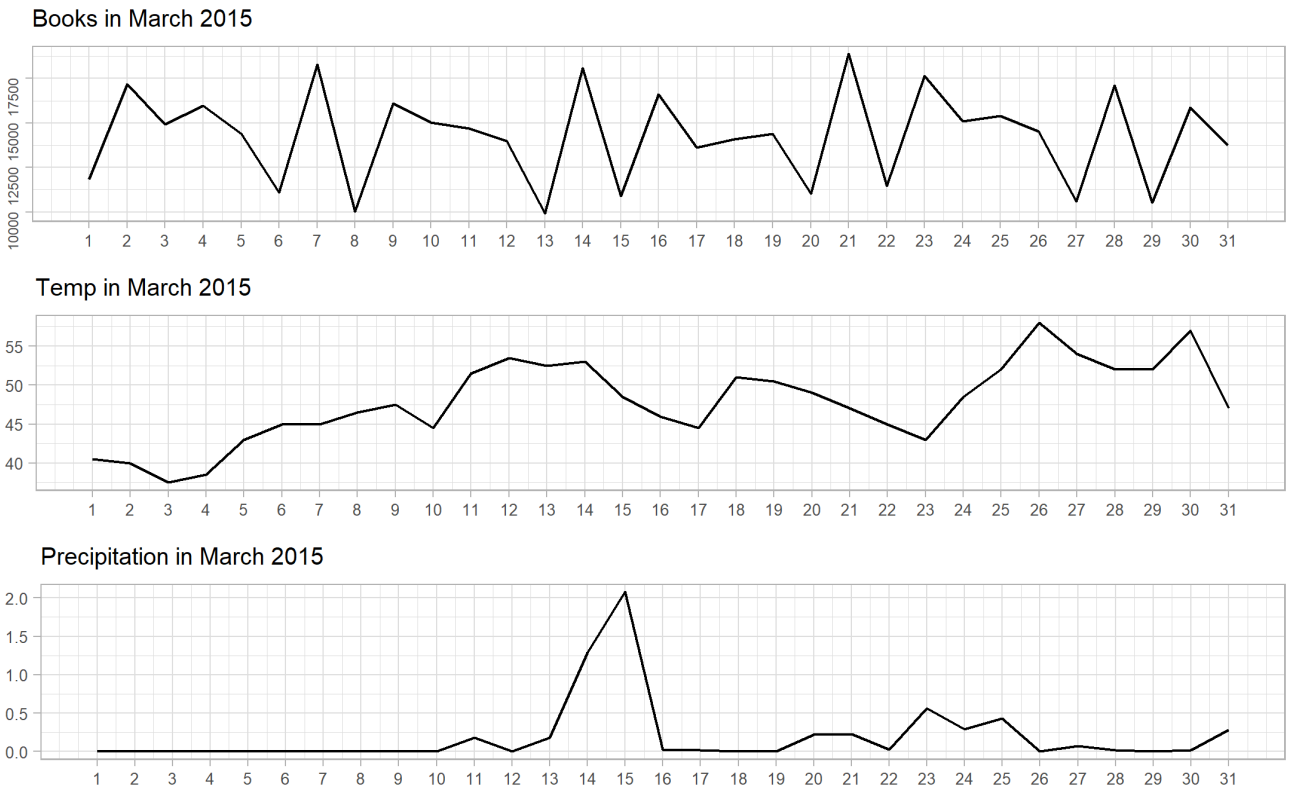
Query – daily book counts in March, 2015:

```
select

extract(day from cout) as days,

count(*) as counts

from spl_2016.inraw

where date_format(cout, '%Y-%m-%d') between '2015-03-01' and '2015-03-31'

group by 1

order by 1
```

Results:

The results are stored in the "cout_march_2015.csv", the daily weather data (temperature and precipitation) is once again taken from NOAA[1]. The plot below visualizes the results. As seen on the plot, there is seemingly no correlation between weather and book checkouts on the daily basis, apart from extreme scenarios. For instance, on March 15 there was heavy rainfall, and book checkouts dropped significantly.

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



As the next step, I test the granularity of the results following the last week's and break down the total counts by item types.

Query: daily item counts by major item types

```
select
extract(day from cout) as days,
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

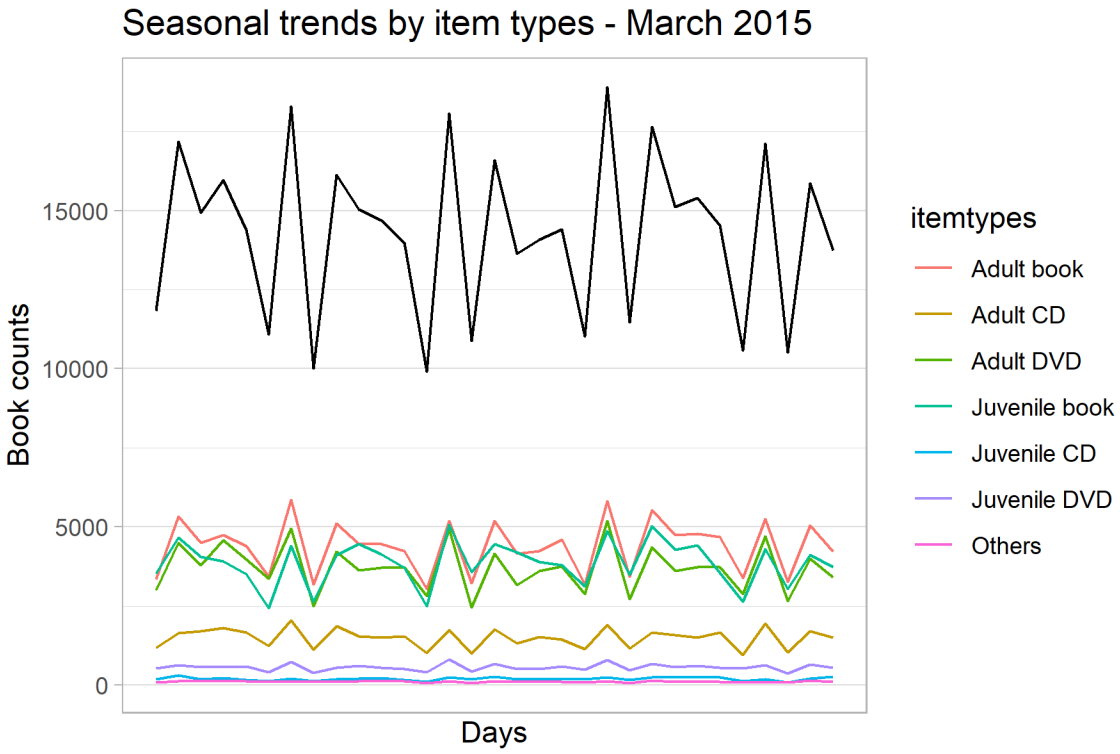
where date_format(cout, '%Y-%m-%d') between '2015-03-01' and '2015-03-31'

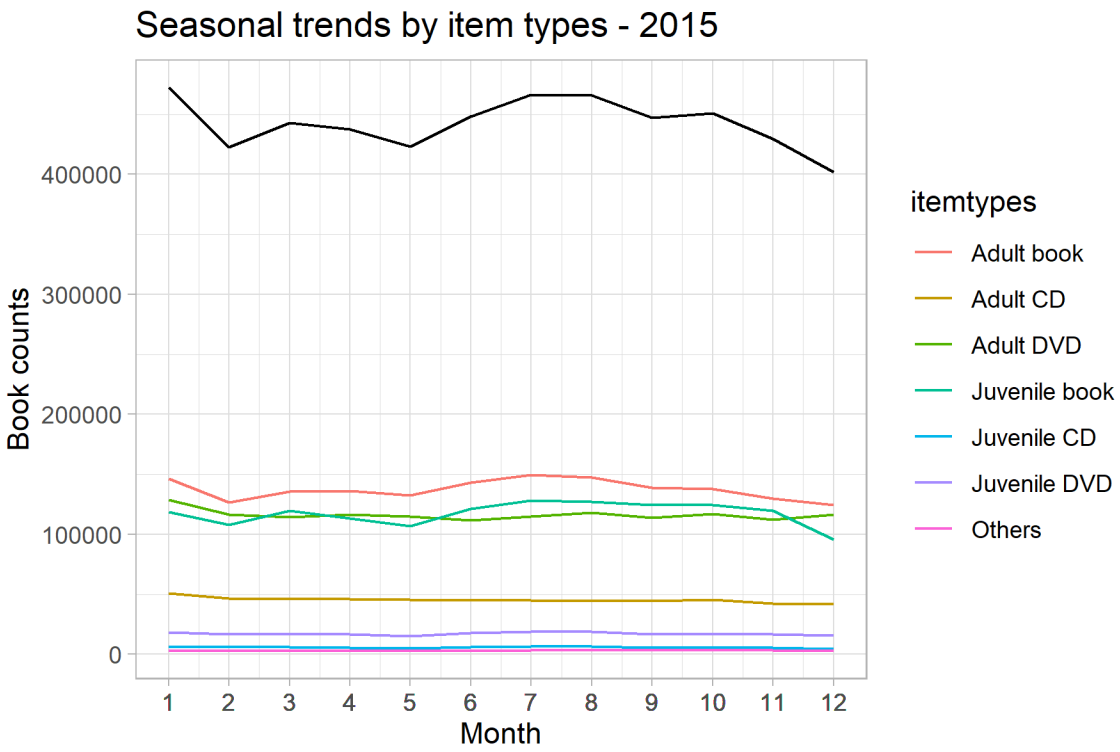
group by 1,2

order by 1

Result:

The result is stored in cout_march_2015_by_type.csv. and visualized below. I'm interested in comparing these results with less granular, monthly results from the last week. Last week I found out that when we break down check outs by item types, seasonality gets weaker except for children's literature. This time, looking at more detailed daily data, it is clear that seasonality, in fact, remains visible on a day-to-day basis. The first graph below shows daily data for March, the second one shows monthly data from last week.





For further inquiry, I was hoping to break down book checkouts by library branch. However, I was not able to match the collcode data from any of the tables to the latest data dictionary with location codes from the SPL website.[2] For now it seems that collcode does refer to specific item collections and not locations – I am not sure how to match these and would appreciate any insight into the coding scheme of collcode.

Without this option, I will instead try to detect any other seasonal patterns. Specifically, I will test the trends in distinct item counts, distinct types; and calculate the ratio of the counts for the top performing item each day to the total counts each day. That is, I am hoping to find out whether library patrons tend to focus on specific item types, or on specific book on particularly cold or rainy days.

Query – distinct items and types per day

```
select

extract(day from cout) as days,

count(distinct(title)) as title_counts,

count(distinct(itemtype)) as type_counts

from spl_2016.inraw
```

[2] https://data.seattle.gov/Community/Integrated-Library-System-ILS-Data-Dictionary/pbt3-ytbc/data

```
where date_format(cout, '%Y-%m-%d') between '2015-03-01' and '2015-03-31'
group by 1
order by 1
```

Query – counts of top performing items:

```
select
days,
max(counts) as top_counts
from
(select
extract(day from cout) as days,
title,
count(*) as counts
from spl_2016.inraw x
where date_format(cout, '%Y-%m-%d') between '2015-03-01' and '2015-03-31'
group by 1,2
order by 1,3 desc) y
group by 1
```

Result:

The results for these queries are stored in distinct_items_and_types_by_day.csv and top_items_per_day respectively. For analysis, I merge all of this information into a single dataframe in R studio and calculate the daily weight of a top performer by dividing the counts from the second query by distinct item counts from the first query. I then plot the results below.

Several things to note based on these results:

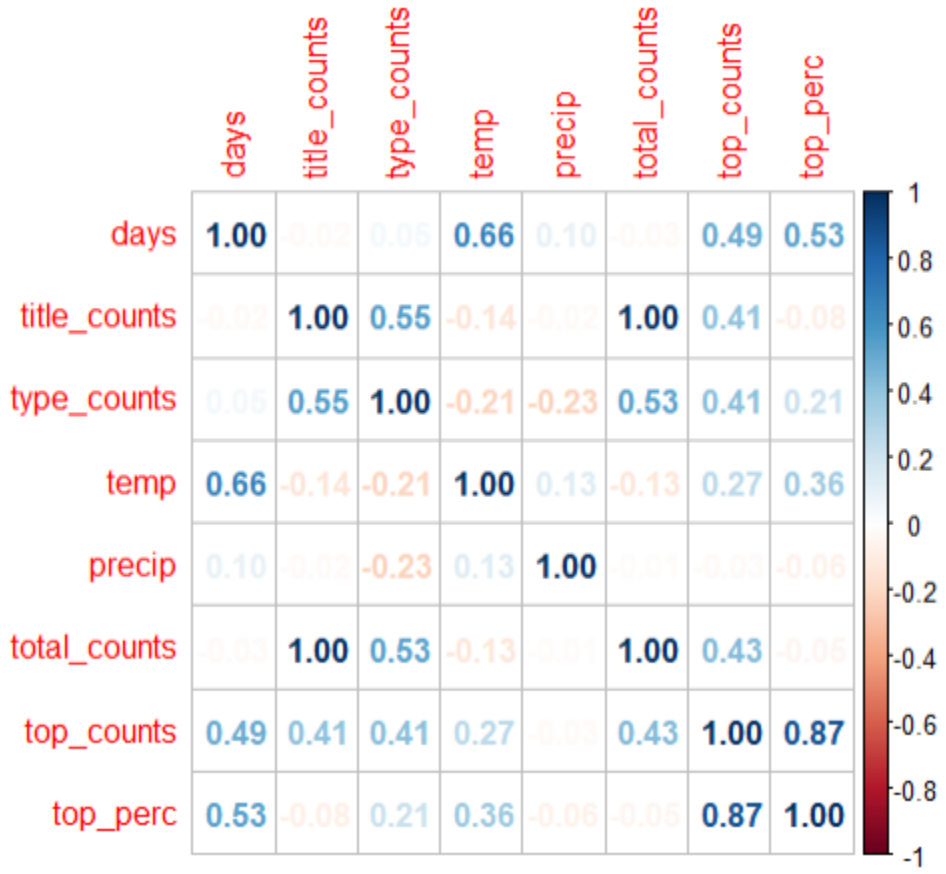
1. First, distinct titles closely follow the trends in total book checkouts, which is expected.
2. Second, in March 2015 there seems to be no particular day with a clearly dominant item. The counts of best performing items each day are in the range of 10-40, and as a result the account only for the 0.003% of total checkouts in any given day. It seems there were no viral book was released during that timeframe. Neither were the patrons willing to endure heavy rain to get a certain popular item – in fact, on March 15, the top performer that day had the lowest ratio to the total book checkouts.

Visually, graphs are a not very convenient form of analysis. I will follow up by calculating a proper correlation with my selected variables and building a correlation plot below.

The conclusions from the correlation plot are as follows:

1. Correlation of days with everything should be ignored as days are not a proper numeric variable per se. The only thing that makes sense is positive correlation between days and temperature (0.66, as expected).
2. Title counts (distinct titles) and total counts (all counts) are positively correlated with number of item types. The more items are being checked out, the more diverse they are.

3. Type counts have a weak inverse correlation with both temperature and precipitation – meaning that on hot and rainy days, people tend to focus on certain item types. This correlation, however, seems to be too low to be meaningful.
4. Both counts and percentage of top performing items are weakly positively correlated with temperature – on warmer days people seem to focus on a certain item. The proportion of top performers, however, is once again too low for this to be an indication of any meaningful trend.
5. Precipitation is not correlated with anything else. People in Seattle don't seem to change their reading habits because of rain.



	days	title_counts	type_counts	temp	precip	total_counts	top_counts	top_perc
days	1.00	-0.02	0.05	0.66	0.10	-0.03	0.49	0.53
title_counts	-0.02	1.00	0.55	-0.14	-0.02	1.00	0.41	-0.08
type_counts	0.05	0.55	1.00	-0.21	-0.23	0.53	0.41	0.21
temp	0.66	-0.14	-0.21	1.00	0.13	-0.13	0.27	0.36
precip	0.10	-0.02	-0.23	0.13	1.00	-0.01	-0.03	-0.06
total_counts	-0.03	1.00	0.53	-0.13	-0.01	1.00	0.43	-0.05
top_counts	0.49	0.41	0.41	0.27	-0.03	0.43	1.00	0.87
top_perc	0.53	-0.08	0.21	0.36	-0.06	-0.05	0.87	1.00

Conclusion:

Contrary to last week's findings, this time I found to strict seasonal correlation between temperature, precipitation, and parameters such as total book checkouts, daily breakdown by type, diversity of item types and titles, and the weight of a top performing item. It is clear only that extreme weather conditions (such as rain on March 15) result in less patrons attending, naturally. Other than that, the limited selection of variables this time is not affected by weather at all.

References:

1. NOAA Online Weather Data: https://www.weather.gov/wrh/climate?wfo=sew

2. SPL Integrated Library System (ILS) Data Dictionary: https://data.seattle.gov/Community/Integrated-Library-System-ILS-Data-Dictionary/pbt3-ytbc/data