MAT259 Project1

Guanyu Chen

January 16, 2020

Contents

Motivation

The primary interest of this project is forecasting monthly amounts of dewey books using historical data. The dataset used here, which contain 168-month chekc-out records for different sectors from January 2006 to December 2019, is provided by Seattle Public Library database and are alll time-correlated. It is possible for us to summary statistics and graphical representations of check-out records and conduct further predictions.

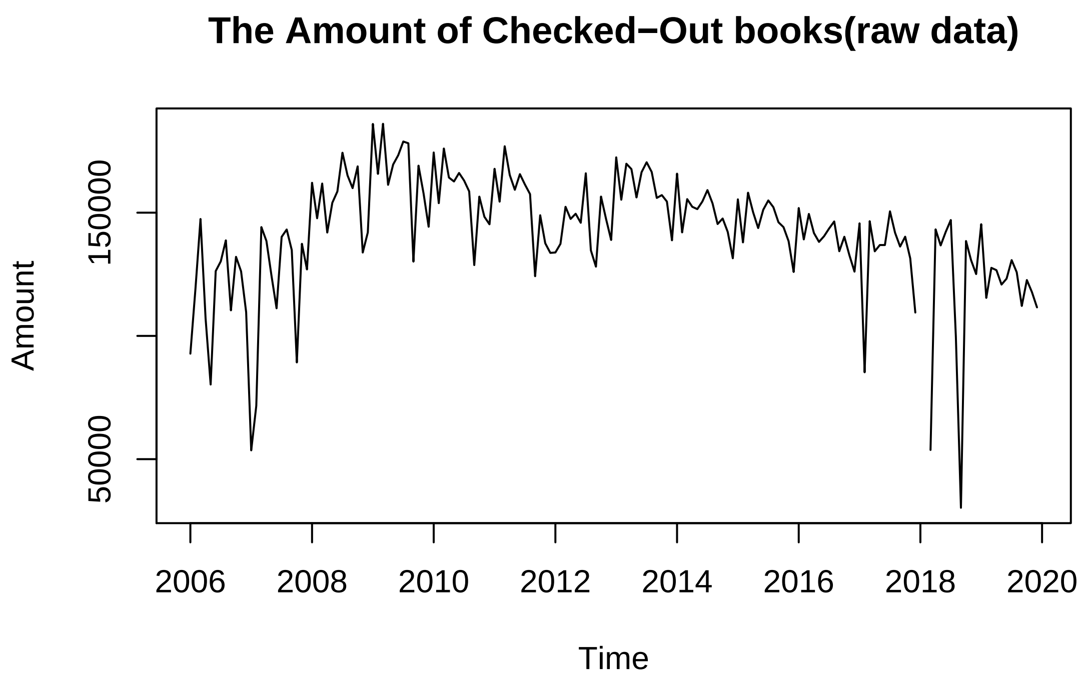
Difficulties

In reality, a collection of data is not always perfect for data analysis. Before starting data analysis, it is necessay to detect some errors, anomalies or hidden patterns.

- Missing Data: The dataset contain missing data in January 2018 and February 2018.
- Abnormal Data: In 2018, the amount of check-out books are extremly lower than previous year.

```
# data
# add NA values for 2018.1 & 2018.2
book = read.csv("Dewey_book.csv", header=TRUE)
book = as.data.frame(book)
book = insertRows(book, r=c(145,146),
                new=as.data.frame(rbind(c(2018, 1, NA), c(2018, 2, NA))))

# plot graph with missing data
book_ts = ts(book[,3],start=2006,f=12)
ts.plot(book_ts,main  = "The Amount of Checked-Out books(raw data)",
        ylab="Amount")
```

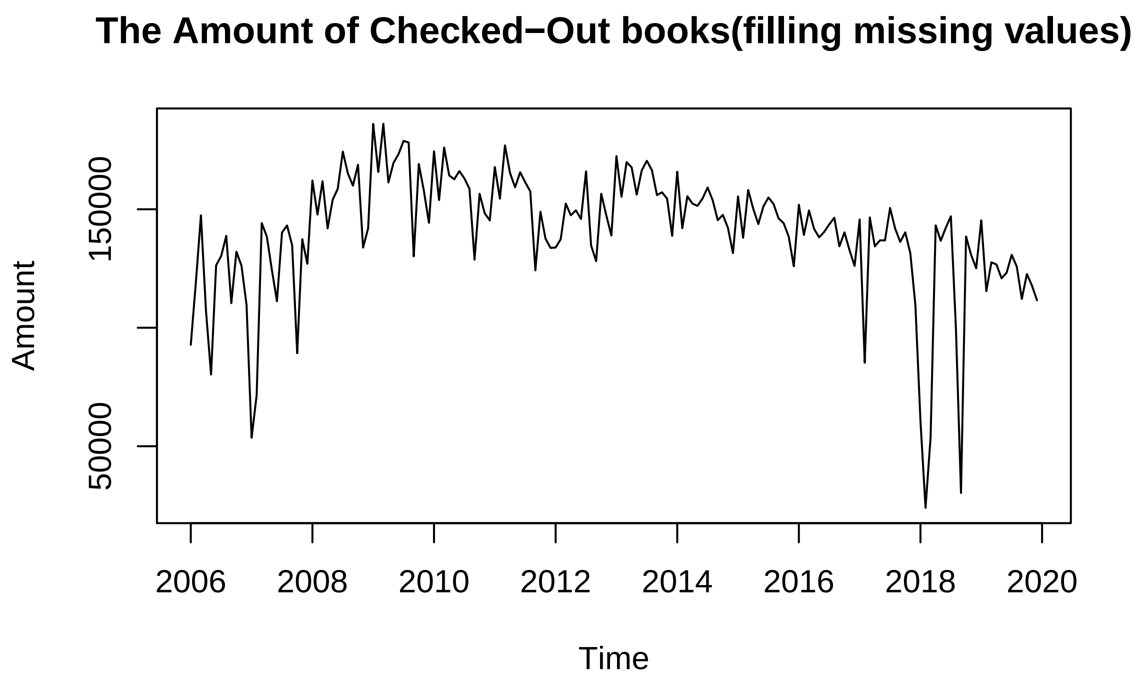


Methods

Due to missing data, in order to recognize long-term data pattern, I recollected all periodic data in system for analysis.

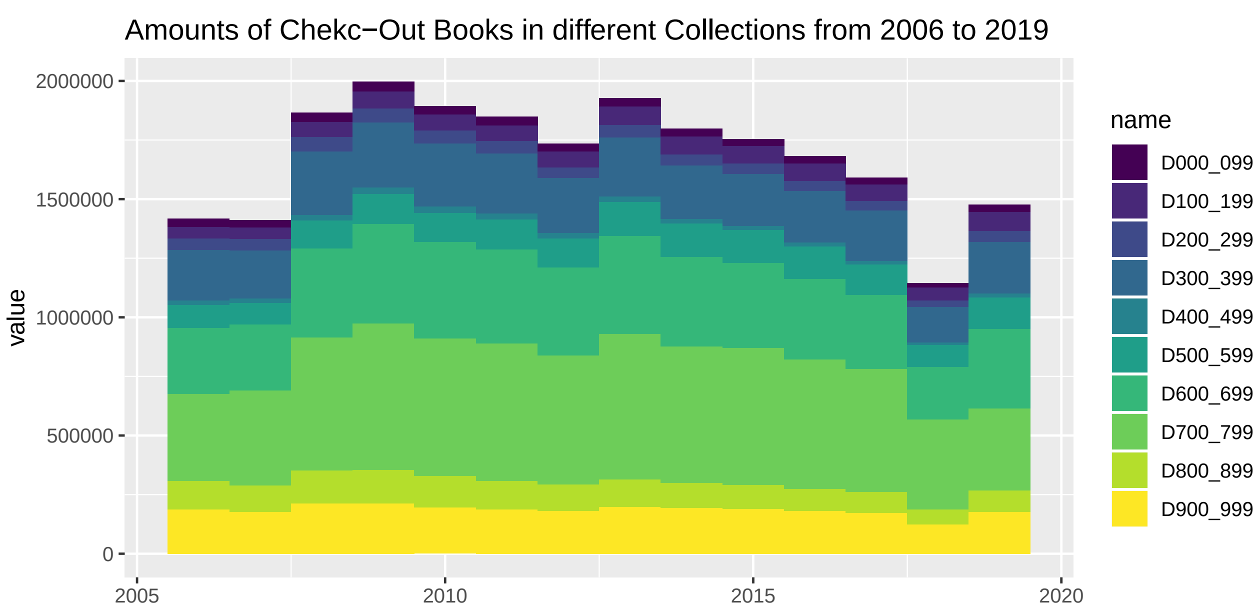
For missing values, applying interpolation method is one of way to solve the problem. Interpolation is used to extrapolate the missing data within the range of discrete set of known data points. The simplest way of interpolation is the linear interpolation that it can fillin a new value by the mean of twoadjacent known values.

```
# plot graph with filling missing data
book_fillna = na.interpolation(book, option ="spline")
book_fillna_ts = ts(book_fillna[,3],start=2006,f=12)
ts.plot(book_fillna_ts,
        main  = "The Amount of Checked-Out books(filling missing values)",
        ylab = "Amount")
```



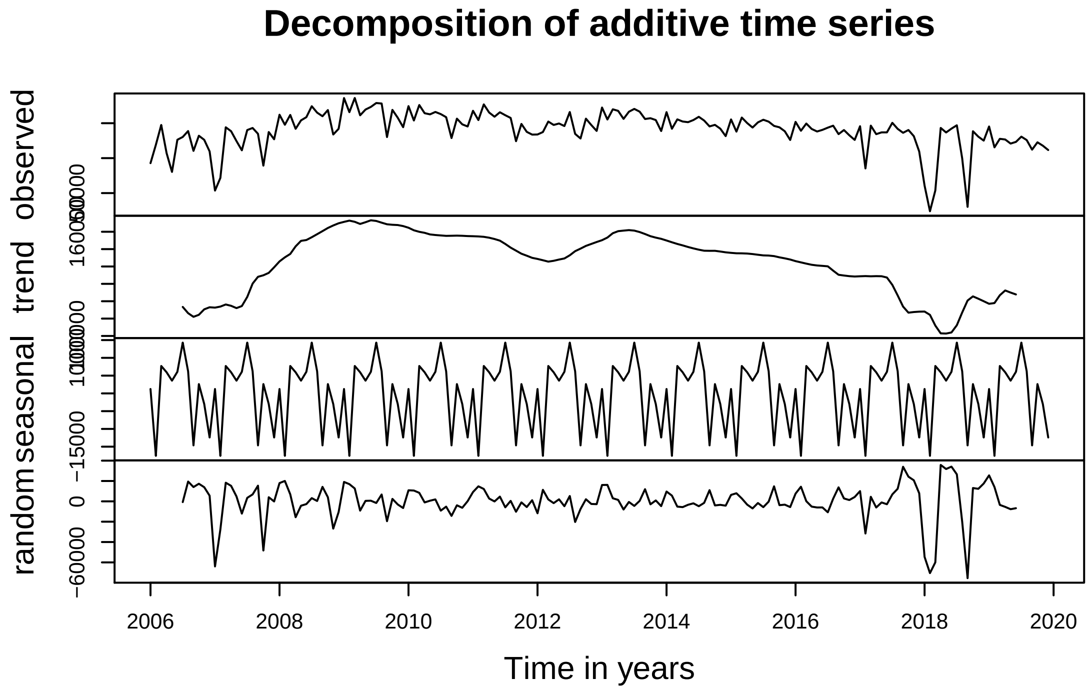
Also, we are able to dive into how did each sectors change over time to show any tendencies of checked-out books in library.

```
library(viridis)
ggplot(data2, aes(fill= name, y=value, x=year)) +
        geom_bar(position="stack", stat="identity", width = 1) +
        scale_fill_viridis(discrete = T) +
        ggtitle("Amounts of Chekc-Out Books in different Collections from 2006 to 2019") +
        xlab("")
```



In time series analysis, decomposing the original data into trend, seasonality and white noise can help us to understand changes of data and figure out proper models. For our data, the decomposition plot clearly illustrates the trend and seasonality inferred for the original data previously. Within a 12-month period of the seasonal component, there are obvious one checked-out books peak(July) and one bottom(December).

```r
# decompose
decom = decompose(book_fillna_ts)
plot(decom, xlab="Time in years")
```

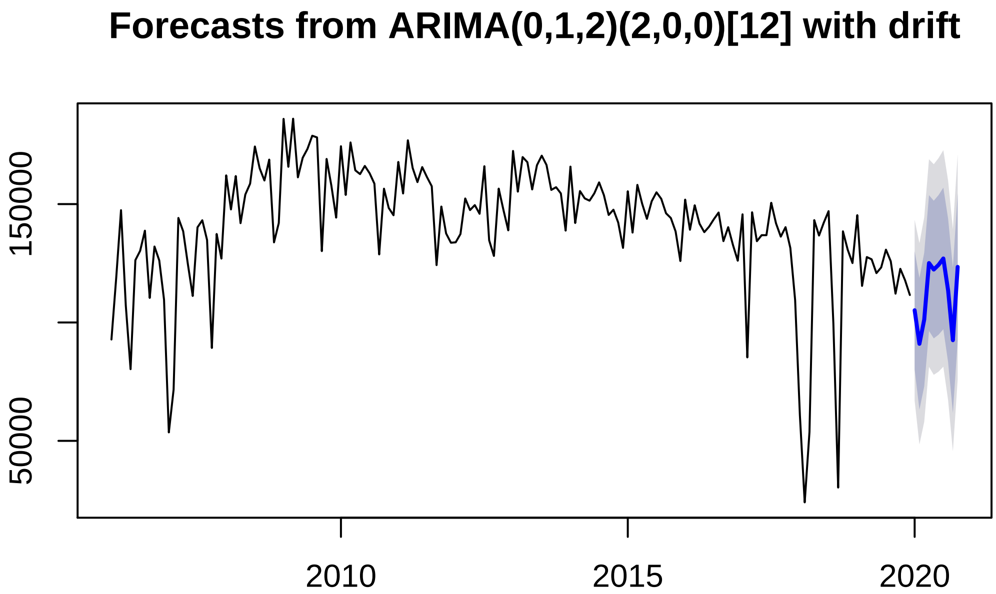


For statistical building model, ARIMA model is applied in data anlysis. The predicted amounts and 95% confidence region are plotted below. The predicted pattern looks close to historical patterns. It is noteworthy that the shape of confidence band widen with increasing horizon, which reflects longer term forecast has more uncertainty. This might also be a sign of need of a more stable model.

```r
# model & prediction
fit1 = auto.arima(book_fillna_ts)
kable(forecast(fit1,10))
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2020	105159	80173	130145	66946	143372
Feb 2020	90985	63181	118788	48463	133506
Mar 2020	101217	72983	129450	58037	144396
Apr 2020	125092	96434	153749	81264	168919
May 2020	122351	93276	151426	77885	166817
Jun 2020	124317	94830	153803	79221	169413
Jul 2020	127032	97139	156924	81315	172748
Aug 2020	113507	83214	143800	67178	159836
Sep 2020	92503	61815	123191	45569	139437
Oct 2020	123474	92396	154553	75944	171005

```r
plot(forecast(fit1,10))
```



Future Study

- Work on descriptive data analysis like most circulating books every year
- Add more details about building models (because ignoring model diagnostics and prediction accuracy)

Appendix

```
SELECT
    YEAR(cout) AS Years,
    MONTH(cout) AS Months,
    SUM(CASE
        WHEN deweyClass != '' THEN 1
        ELSE 0
    END) AS Dewey
FROM
    spl_2016.outraw
WHERE
    itemtype LIKE '%bk'
        AND YEAR(cout) >= '2006'
        AND YEAR(cout) <= '2019'
GROUP BY YEAR(cout) , MONTH(cout);
```

```
SELECT
    YEAR(cout) AS Years,
    MONTH(cout) AS Months,
    SUM(CASE
        WHEN deweyClass > 000 AND deweyClass < 100 THEN 1
        ELSE 0
    END) AS D000_099,
    SUM(CASE
        WHEN deweyClass > 100 AND deweyClass < 200 THEN 1
        ELSE 0
    END) AS D100_199,
    SUM(CASE
        WHEN deweyClass > 200 AND deweyClass < 300 THEN 1
        ELSE 0
    END) AS D200_299,
    SUM(CASE
        WHEN deweyClass > 300 AND deweyClass < 400 THEN 1
        ELSE 0
    END) AS D300_399,
    SUM(CASE
        WHEN deweyClass > 400 AND deweyClass < 500 THEN 1
        ELSE 0
    END) AS D400_499,
    SUM(CASE
        WHEN deweyClass > 500 AND deweyClass < 600 THEN 1
        ELSE 0
    END) AS D500_599,
    SUM(CASE
        WHEN deweyClass > 600 AND deweyClass < 700 THEN 1
        ELSE 0
    END) AS D600_699,
```

```sql
    SUM(CASE
        WHEN deweyClass > 700 AND deweyClass < 800 THEN 1
        ELSE 0
    END) AS D700_799,
    SUM(CASE
        WHEN deweyClass > 800 AND deweyClass < 900 THEN 1
        ELSE 0
    END) AS D800_899,
    SUM(CASE
        WHEN deweyClass > 900 AND deweyClass < 1000 THEN 1
        ELSE 0
    END) AS D900_999
FROM
    spl_2016.outraw
WHERE
    itemtype LIKE '%bk'
        AND YEAR(cout) >= '2006'
        AND YEAR(cout) <= '2019'
GROUP BY YEAR(cout) , MONTH(cout);
```