Homework on Exploratory Data Analysis
Data Science and Machine Learning, University of Tokyo

Name	: Hansen Hendra
Student ID	: 37205127
Department	: EEIS (Electrical Engineering and Information Systems)
Faculty	: Graduate School of Engineering

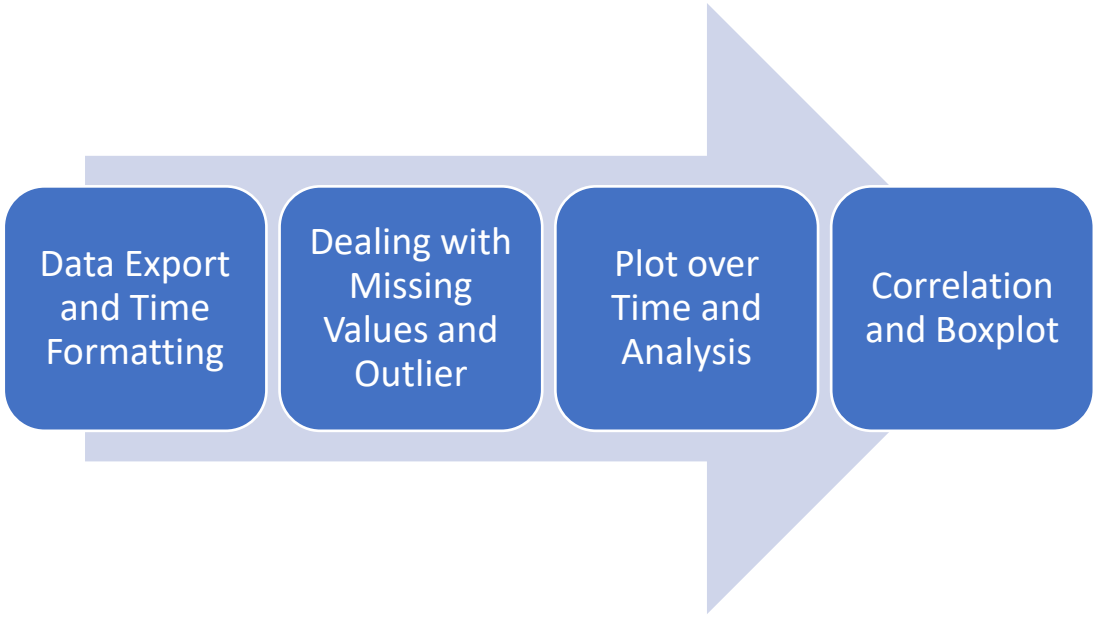
In this EDA, Python is used as the main program to do data preprocessing and visualization.
Pandas Dataframe is used for the data structure, **Seaborn and Matlplotlib** is used for visualization.

The HW1-Hansen Hendra.IPYNB is structured as below Table of Contents for both USA and Canada Data.
See the notebook and click on the hyperlink to see the detail of each part.

Table of Contents:

- USA DATA
  - Load Data and Time Formatting
  - Dealing with Missing Values
  - Finding Outlier Values
  - Pattern of Outlier Values
  - Plotting over time
    - Plotting Per City
    - Plotting Per Goods Sector
  - Correlation Between Goods
  - Boxplot
- CANADA DATA
  - Load Data and Time Formatting
  - Dealing with Missing Values
  - Finding Outlier Values
  - Plotting
    - Plotting Per City
    - Plotting Per Goods Sector
  - Correlation Between Goods
  - Boxplot

On this report, steps on the process below will be explained briefly.



1. Data Export and Time Formatting

For data export Pandas function read_excel() is simply used.

The time formatting is done by Pandas.to_datetime() function. This is preferable as Pandas can do time plotting easily with this standard format.

```
# convert the first columns into pandas datetime format
usa_df['TIME'] = pd.to_datetime(usa_df['TIME'], format='%Y:%m', errors='ignore')
# preview of %YYYY-%MM-%DD format
```

This snip of code will convert the string with format '%Y:%M' into datetime format.

	Unnamed: 0
0	1976:01
1	1976:02
2	1976:03
3	1976:04
4	1976:05

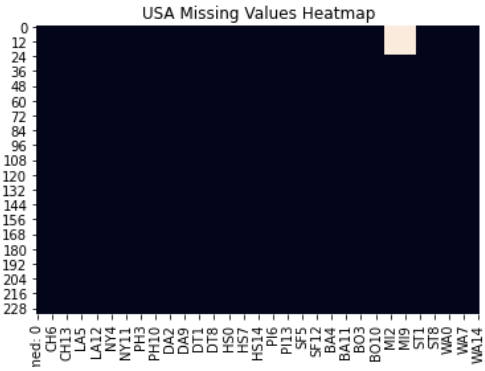
	TIME
0	1976-01-01
1	1976-02-01
2	1976-03-01
3	1976-04-01
4	1976-05-01

USA data time span range from January 1976 to May 1995.
Cana data time span range from January 1974 to May 1995.

2. Dealing with Missing Values and Outlier

A. America Data

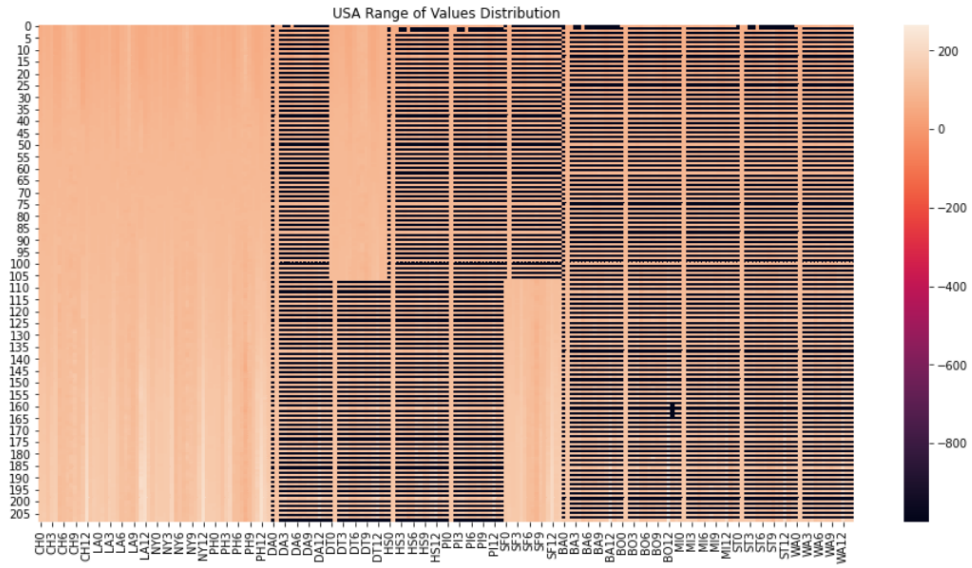
Heatmap is used to visualize how complete is the USA.xls data filled, whether there is empty value.



From the left figure, the black color represents rows with value. The light crème color represent empty value.

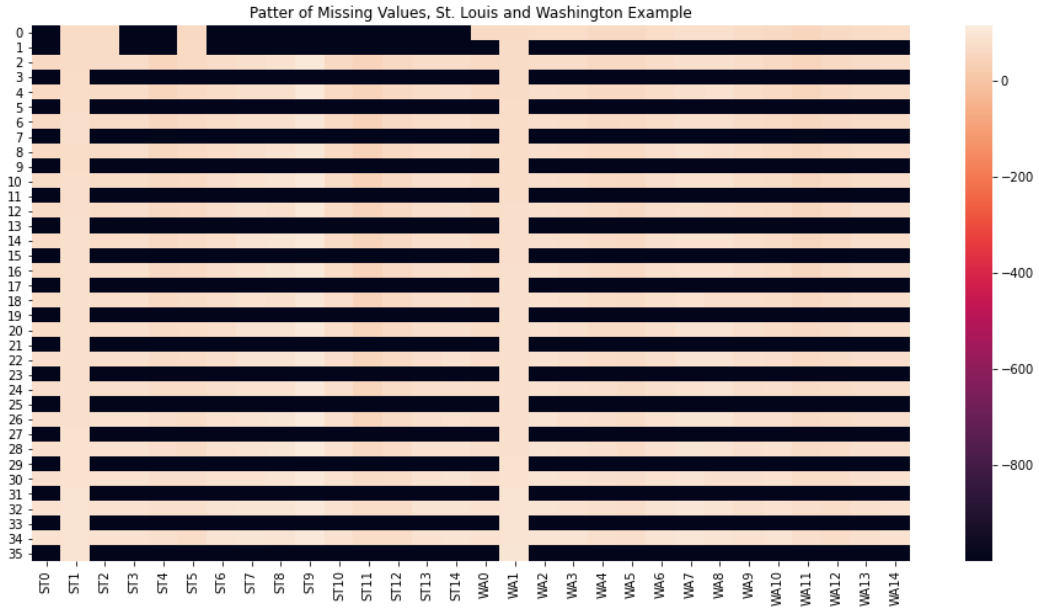
The first 24 months of Miami is empty.
I decided to remove the first 24 months for all cities so there will be no empty values in the process table.

After removing the first 24 months. Value range distribution is visualized with heatmap again as below figure:



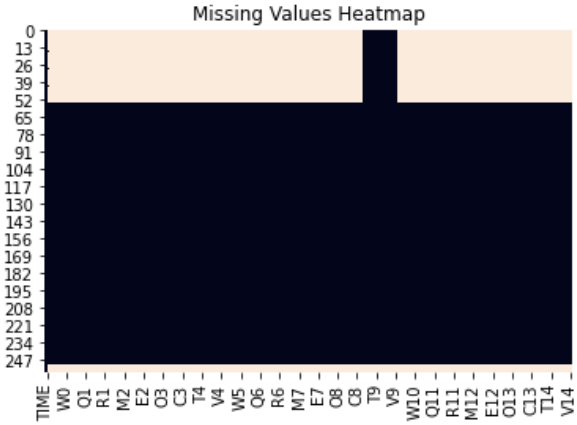
Black color represents very negative value of -999, I assume this is just empty cell that is filled with default value. Therefore we cannot use this black colored record.
Of course, some method for filling missing values like backfill, interpolation, or moving average can be used if it is needed for using all data. Since I do not have economy background, I will not use the data with record value.
<u>The first 4 cities of America, CH, LA, NY, and PH are used for the plotting later.</u>

Here is a further look on very negative value distribution on Washington and St. Louis. A missing value is found every 2 months.



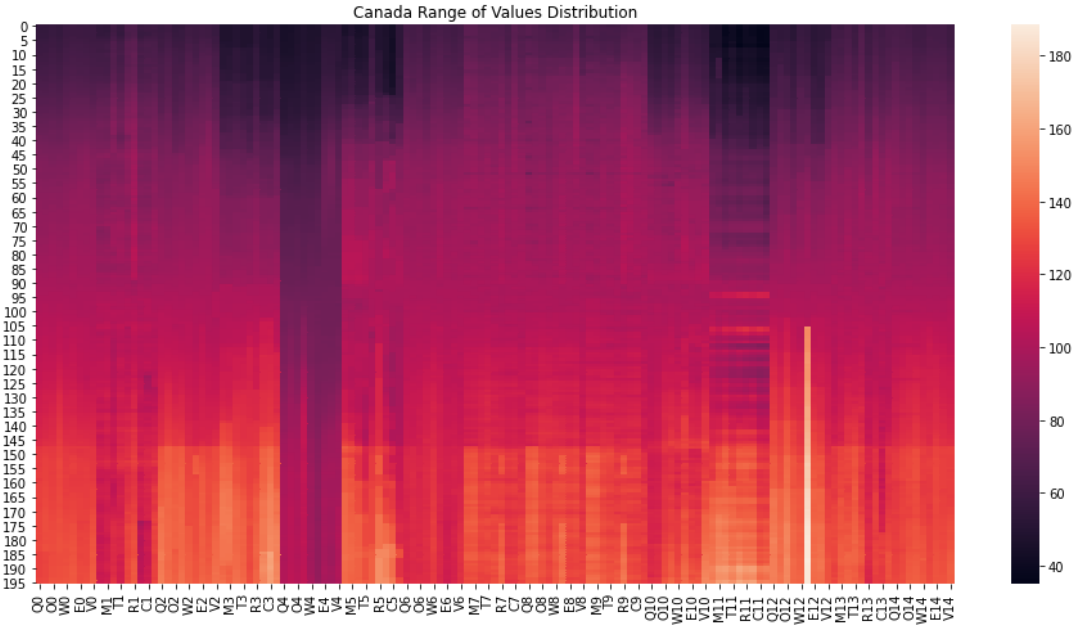
B. Canada Data

Heatmap is used to visualize how complete is the CAN.xls data filled as well.



Same countermeasure is used to clean the data. The first 60 months for all cities are removed so there will be no empty value anymore.

After removing the first 60 months. Value range distribution is visualized with heatmap again as below figure:

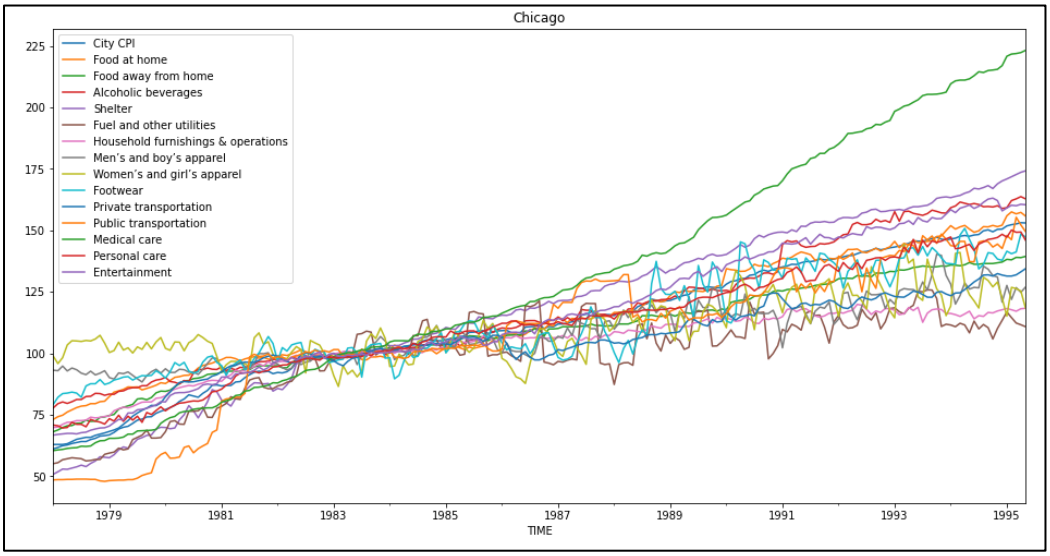


We can see the range for all value is inside -40 to 180. This is plausible value for the price index. **Therefore, all cities of Canada are used for the plotting later.**

3. Plot over Time and Analysis

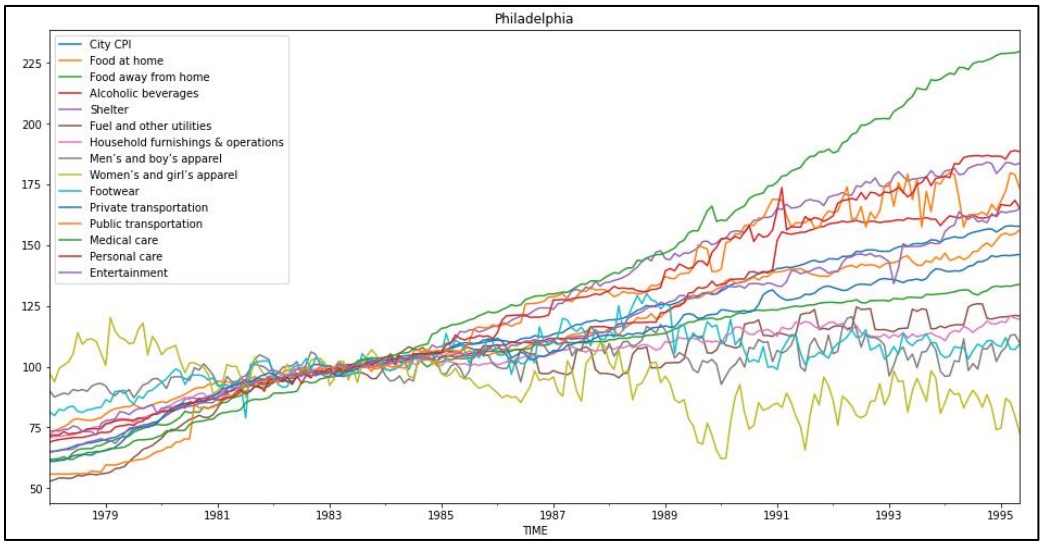
A. America Data

- Plot per city (Chicago and Philadelphia)



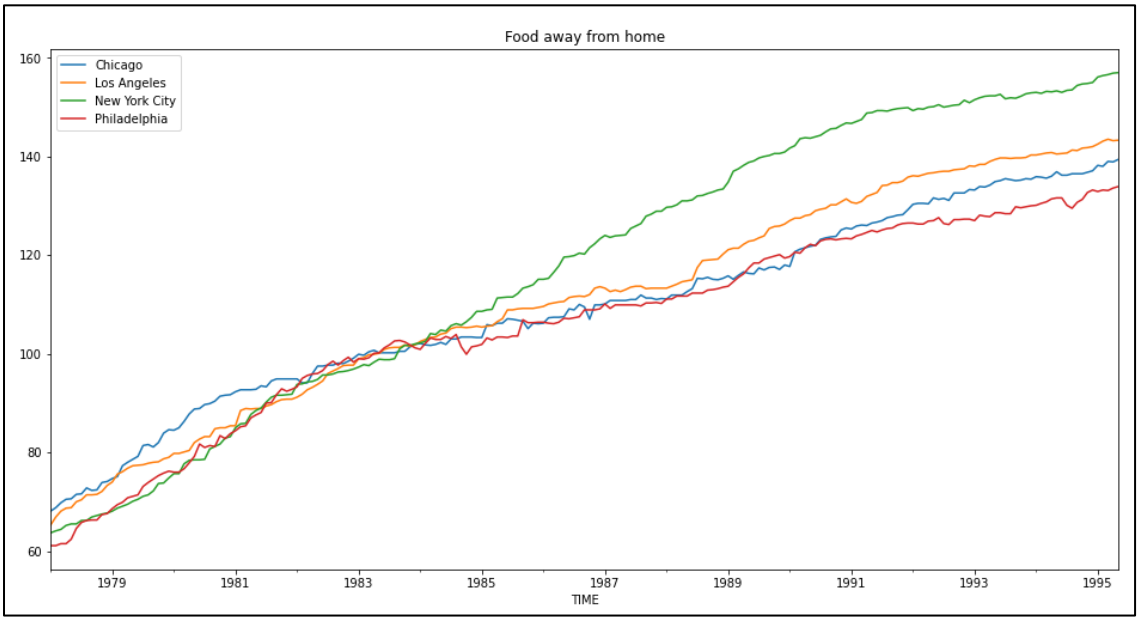
Overall price index in Chicago from 1977 to 1995 was rising in slow slope.

One sector is significantly has higher increase, this is the Food away from home. This sector started at only around 60 in 1977 and rising to around 220 in 1995. “Food away from home” price index rising is usually based on prices of both full- and limited service meals and snacks, vending machine , or even mobile vendors[1]. Another sector that can be noticed is public transportation that had the lowest price index <50 had risen to average CPI around 130 in 1995.

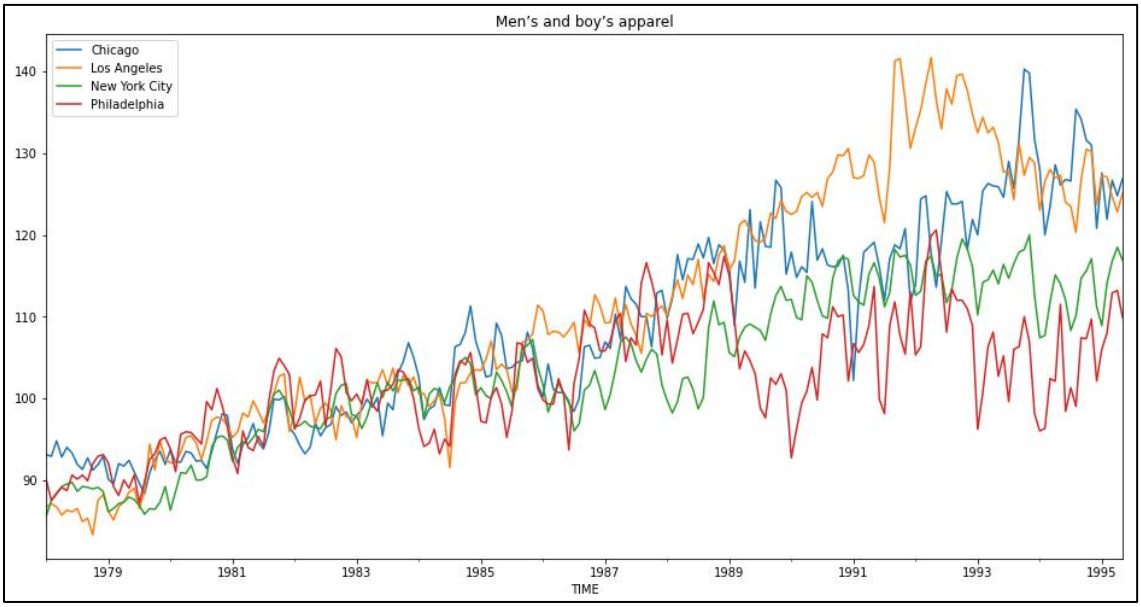


Overall price index in Philadelphia has more variation of trends compared to Chicago. The extreme sector is Women’s and girl’s apparel that led other sector at 1977 at around >100, plummeted to less almost 50 in 1995. On the other hand alcoholic beverage has significant rise from around 50 to about 170 at the end of record. Philadelphia until now is regarded as a state that always consumed large alcohol [2]. From this opinion, the high demand of alcohol most likely to increase the price as well.

- Plot per sector (all cities)



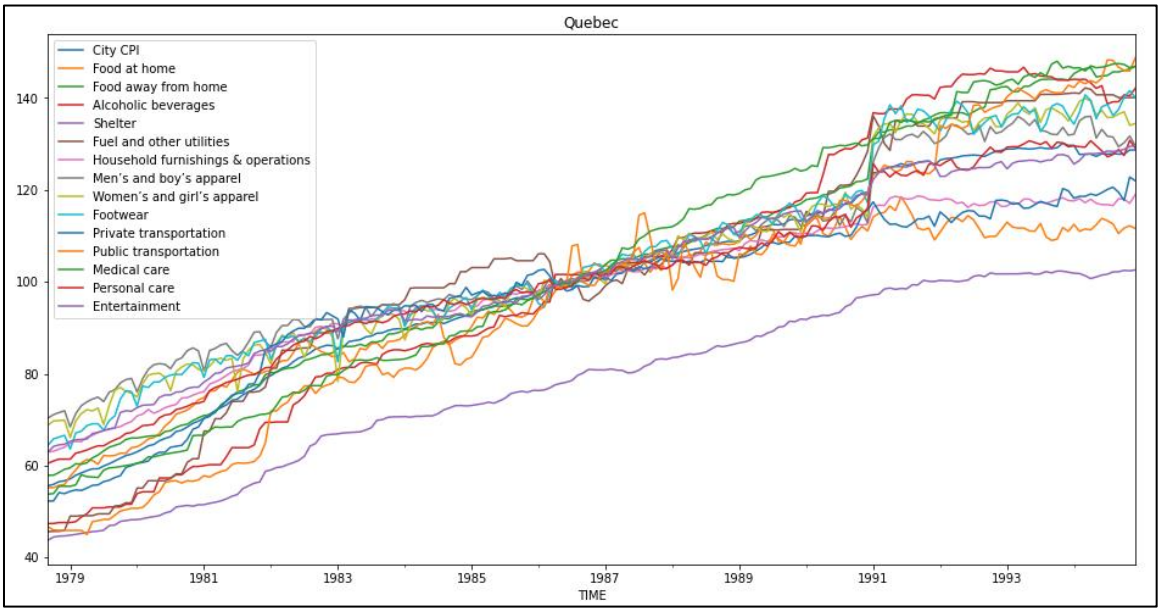
For comparison between the 4 cities, New York City has the highest growth of "Food away from home" price index. From one of article released in 1995 [3], NYC had more than 22000 additional restaurant employees in the city between 1994 and 1999. The growth rate was relatively high compared to another state. The number of restaurant employees surely highly related to the significant increase in "Food away from home" price index.



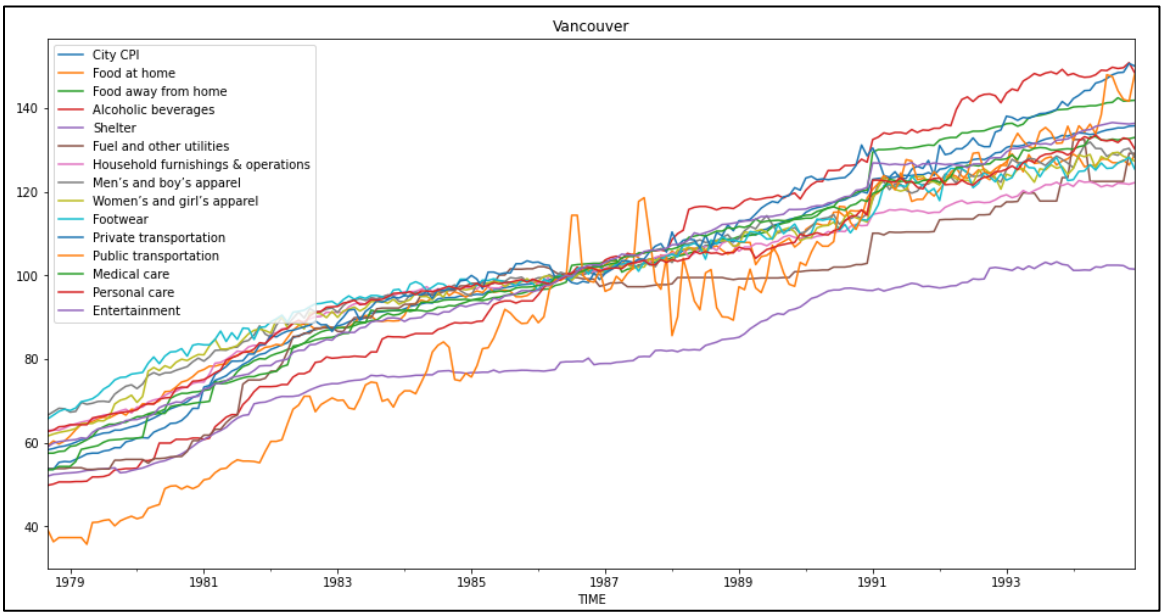
New York City again has the highest growth in "Men's and Boy's Apparel" throughout the period. NYC is surely known as one of the most updated fashion nowadays. But even since 1995, there were NYC Fashion Week with many popular trending celebrities showing off their fashion that takes the country attention. [4]

B. Canada Data

- Plot per city (Quebec and Vancouver)

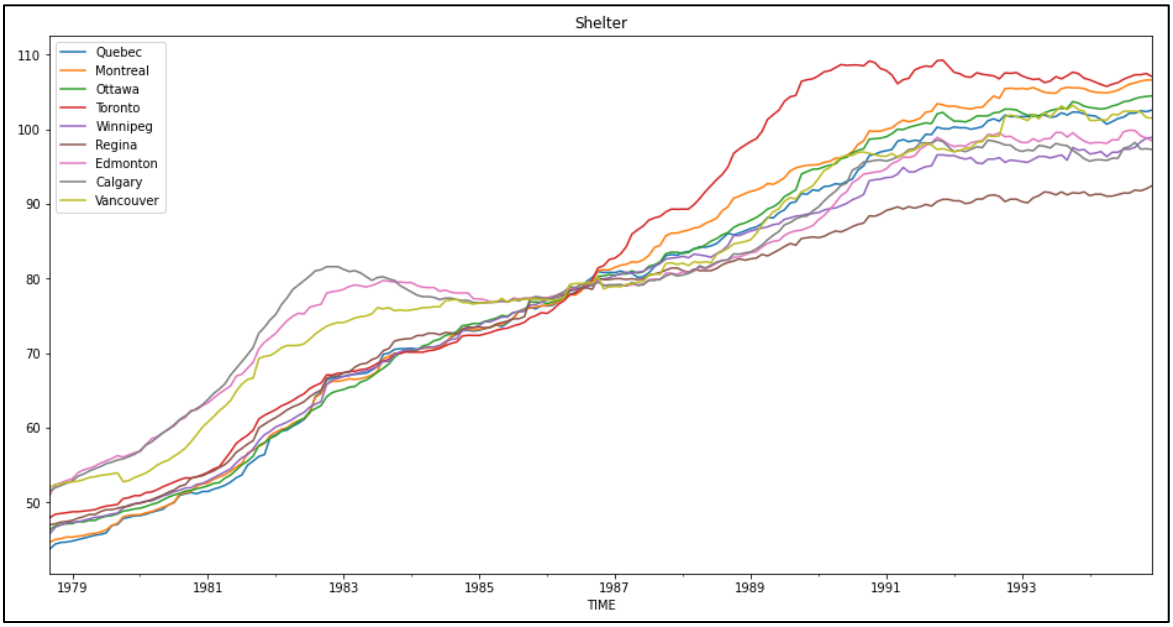


Quebec "entertainment" price sector has the lowest start and end of period.



Vancouver "public transportation" price index has massive growth throughout the period. It was said on "POLICY REPORT URBAN STRUCTURE/TRANSPORTATION" released on 2002[5], one of the key policies is "*Transport 2021"* (released in 1993) to provide a regional framework for managing the transportation system by integrating land use and transportation policies, applying transport demand management, adjusting transport service levels, and supplying transport capacity. The background of this policy is Downtown Vancouver was faced with the challenge of accommodating growth in population and employment and changes to land uses.

- Plot per sector (all cities)



It is clear from the chart above, Toronto had the highest growth of "Shelter" price index from 1978 to 1995. This fact is supported by an article about the sales and price of Toronto home started 1996 to 2017 [6]. Even this is the continuation from the chart above, it is strongly related of the consistency of high growth of shelter high price index.

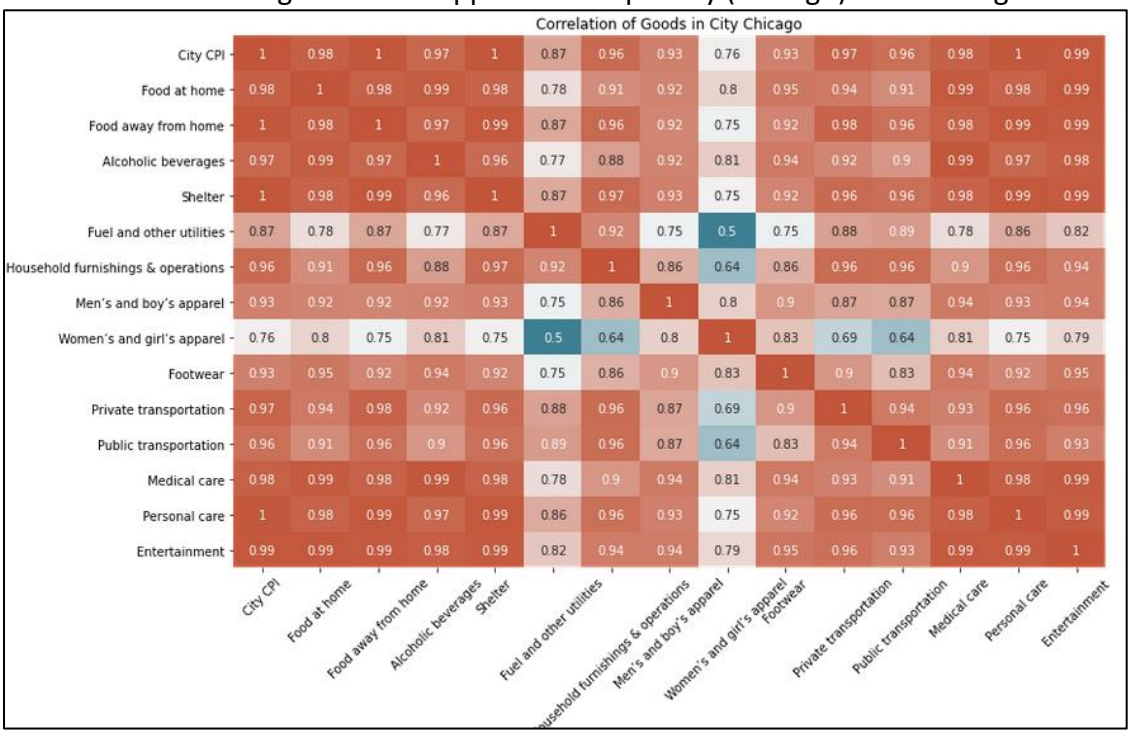
Below is the table showing sales and price of Toronto home:

Year	Sales	Average Sale Price
1995	39,273	$203,028
1996	55,779	$198,150
1997	58,014	$211,307
1998	55,344	$216,815
1999	58,957	$228,372
2000	58,343	$243,255
2001	67,612	$251,508
2002	74,759	$275,231
2003	78,898	$293,067
2004	83,501	$315,231
2005	84,145	$335,907
2006	83,084	$351,941
2007	93,193	$376,236
2008	74,552	$379,347
2009	87,308	$395,460
2010	85,545	$431,276
2011	89,096	$465,014
2012	85,496	$497,130
2013	87,049	$522,958
2014	92,782	$566,624
2015	101,213	$622,121
2016	113,040	$729,837
2017	92,263	$822,572

4. Correlation and Boxplot

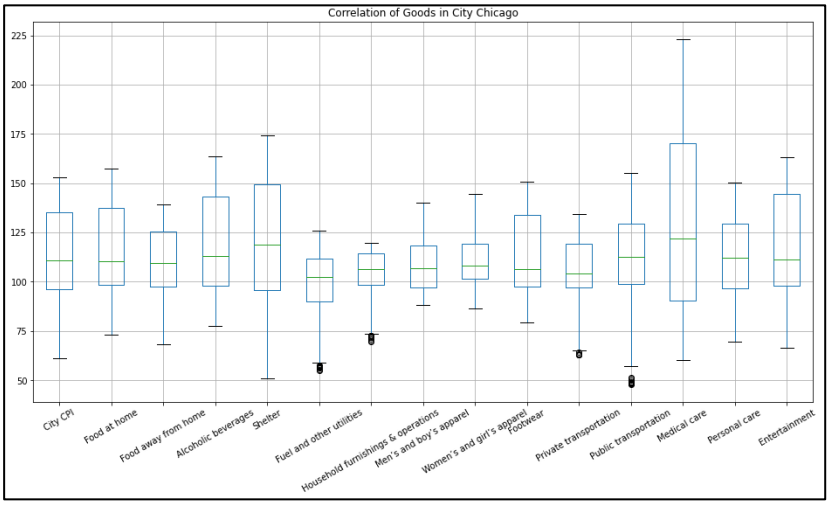
A. America Data

Correlation between goods are mapped for sample city (Chicago) as below figures.



One of the sector that's has relative low correlation with other sectors is Men's and boy's apparel. My hypothesis is fashion was not really affecting other sector, the extreme correlation is to "fuel and other utilities" and "public transportation. This apply also for the "Women's and girl's Apparel". On the other hand, an example of pair of sector that has strong relation is "Food at home" and "Alcoholic beverage".

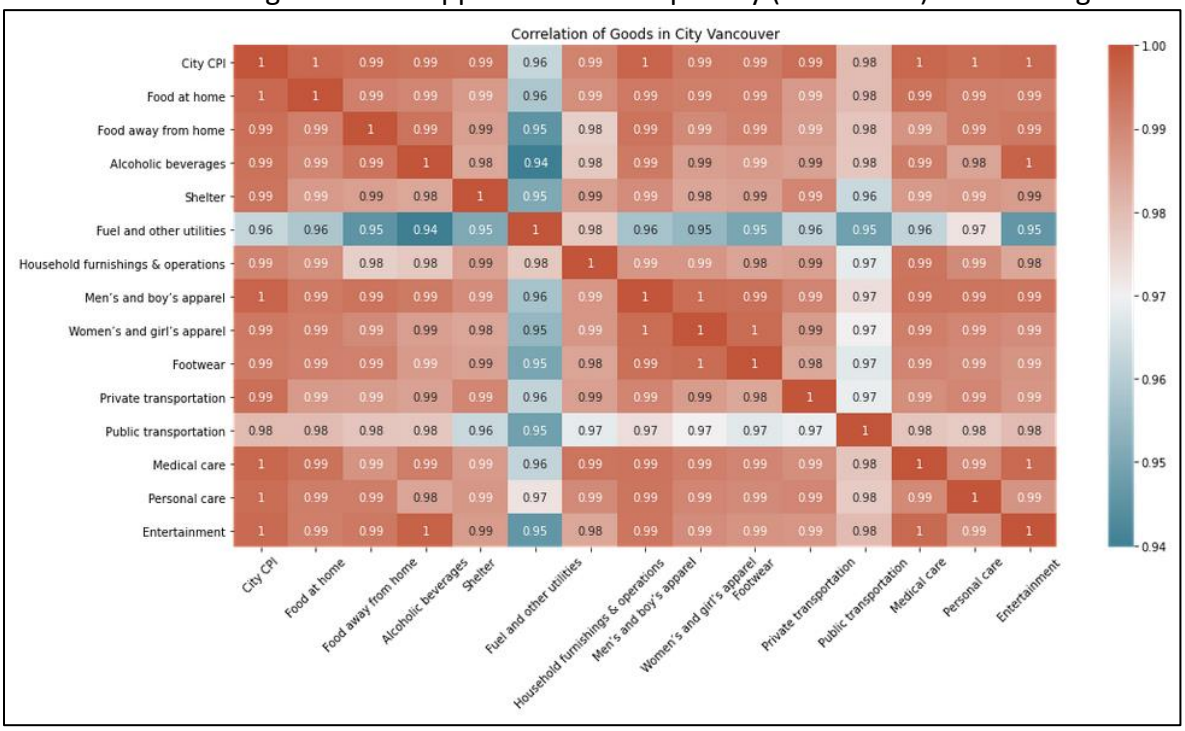
Another way to see the distribution of the price index is boxplot.



Boxplot of Chicago is visualized as below:
From the boxplot on the left. Most of the sectors seems to have normal distribution. Except for the 4 sectors:
1. "Fuel"
2. "Household"
3. "Public Transport"
4. "Private Transport"
has some outlier value.

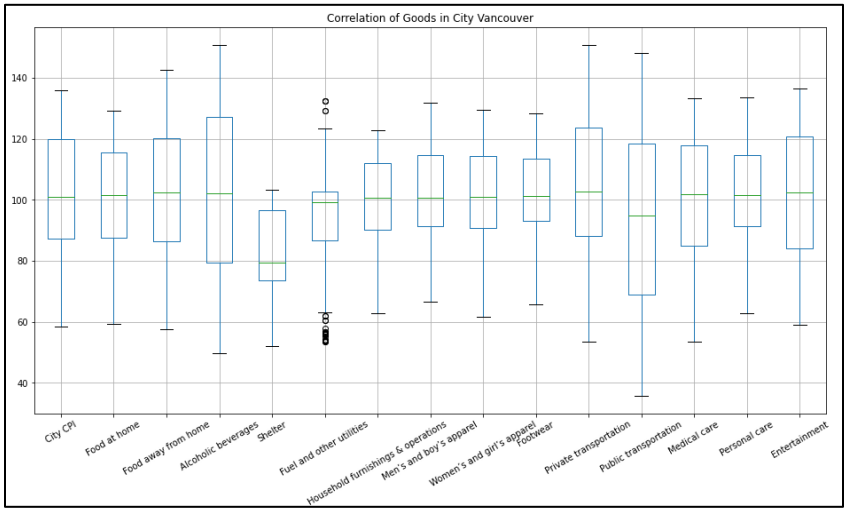
B. Canada Data

Correlation between goods are mapped for one sample city (Vancouver) as below figures:



For Vancouver, it seems that every sectors has strong correlation to any of other sectors. Even the lowest correlation of one sector "Fuel and other utilities" is still strongly correlated with other with minimum value of 0.95. This fact is supported by similar rising trend on the previous section for plot over time in Vancouver city.

Boxplot for Vancouver:



Almost all of the sectors has non skewed distribution. Only on secotr "Fuel and Other Utilities" has outliers.

Opinion on data quality:

USA	Canada
Has complete filled data except for the first 24 months in Miami city.	Has almost complete record for every time. Only the first 60 months (1974-1978) has empty value. Can use all cities data for the years after 1979.
Almost all cities (except the 4 [CH,PH,NY,LA]) has default values of -999 for every 2 months sampling. It seems that it is the method of default filling for empty values by the data provider.	Has perfectly good distribution for all time. The range is -40 to 180.
From the sampled boxplot distribution, most of the sector has normal distribution. Four sectors has outliers.	From the sampled boxplot distribution, almost all sector has unskewed normal distribution. Only one sector has outliers.

Both data generally has good sufficient and reasonable record values. I prefer to work on Canada dataset since it has complete records and unskewed distribution (most likely properly recorded data). Working on USA data is feasible if the data scientist has some background on the topic to make a proper assumption how to deal with missing values and default -999 values.

REFERENCE:

[1]About "Food away from Home" price index, https://www.bls.gov/opub/ted/2019/food-for-thought-changes-in-consumer-prices-for-food-at-home-and-away-from-home.htm?view_full

[2] "Philadelphians Sure Drink a Lot More Alcohol..", https://politicalcalculations.blogspot.com/2018/08/philadelphians-sure-drink-lot-more.html#.X42g8O0xVPY

[3]" Restaurant employment boom in New York", https://tobaccocontrol.bmj.com/content/10/2/199.1

[4] Article about New York Fashion Week 1995, https://www.wmagazine.com/gallery/new-york-fashion-week-celebrities-throwback/

[5] "Policy Report Urban Structure/Transportation by Vancouver Council about Downton Transportation Plan", https://council.vancouver.ca/020528/rr1.htm

[6] "1996 To 2017: A Toronto Real Estate Dynasty!", https://torontorealtyblog.com/blog/1996-to-2017-the-streak-is-over/