**Capstone Final Project Report (DATS6501)**

Yupeng Yang

Data Science, The George Washington University

December 03, 2018

**Motivations**

As a graduate student in our Data Science Program, I have been making significant progress since I came here. Based on my skills and knowledge acquired during the previous precious semesters, I decide to develop a Synthetic Project, including many subprojects and multiple techniques I learnt.

**Project Outline**

1. Creating self-designed Personal Webpage and performing an analysis focused on the Box Office Revenue;

2. Obtaining reliable data information from the most popular movie information website by Web Scraping technique (Data Source: https://www.boxofficemojo.com, which is the leading online box-office reporting service, the #1 movie website in the world);

3. Visualizing the data information by building various charts and distribution maps for several countries and companies in different years;

4. Making predictions and conclusions based on Time Series Analysis (such as using ARIMA Model), giving scientific suggestions and recommendations to film makers and customers.

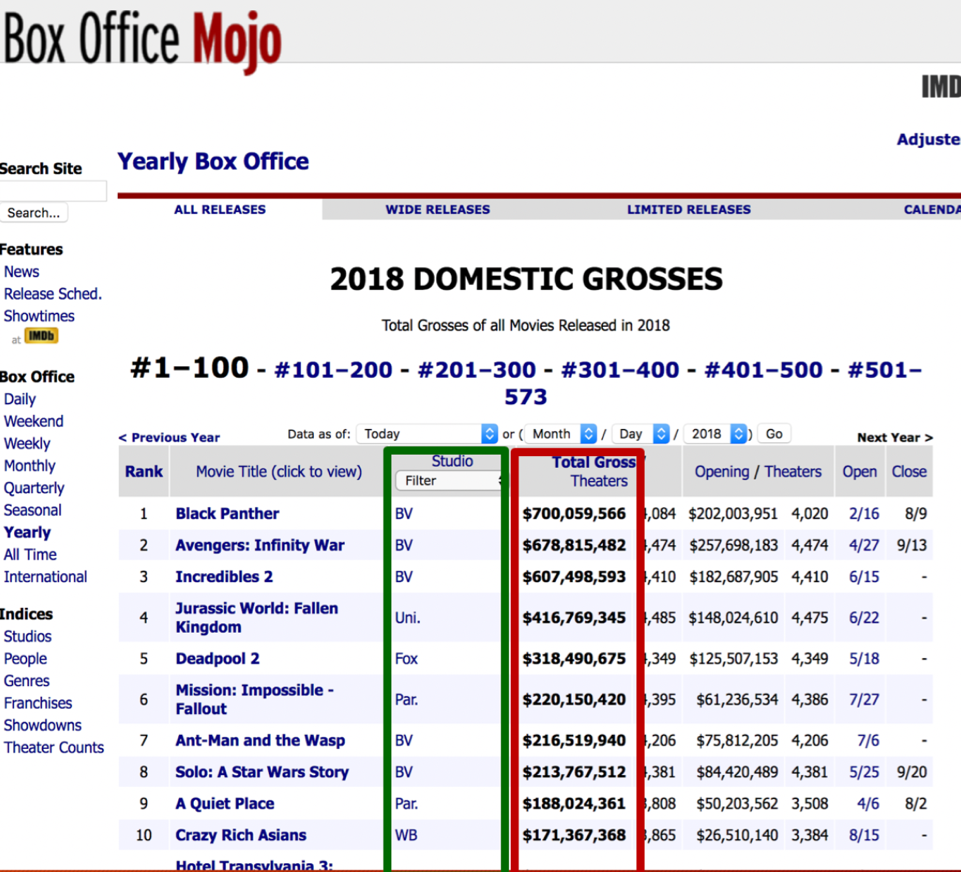
Programming Languages:Python, HTML, CSS, JavaScript

**Personal Webpage**

First, I designed header part in the beginning of my webpage, then I inserted my name, project name, university name and a background picture into this header. By adjusting the font and image format in my style.css file, the names and background picture can be put in the center of the header. Second, I inserted a menu bar under the header part. There are 4 portions in this menu bar, which are “Home”, “Intro”, “Data”, and “Analysis”. It shows different contents when clicking different bars. For example, under the “Home” bar, I wrote my current positions and stories in the content portion and added my contact information and some useful links to the sidebar, which is located in the right side of the page. I also inserted my recent interested videos below, which can be shown when clicking on them. Finally, I created a footer part in the end of the page.

**Data Description**

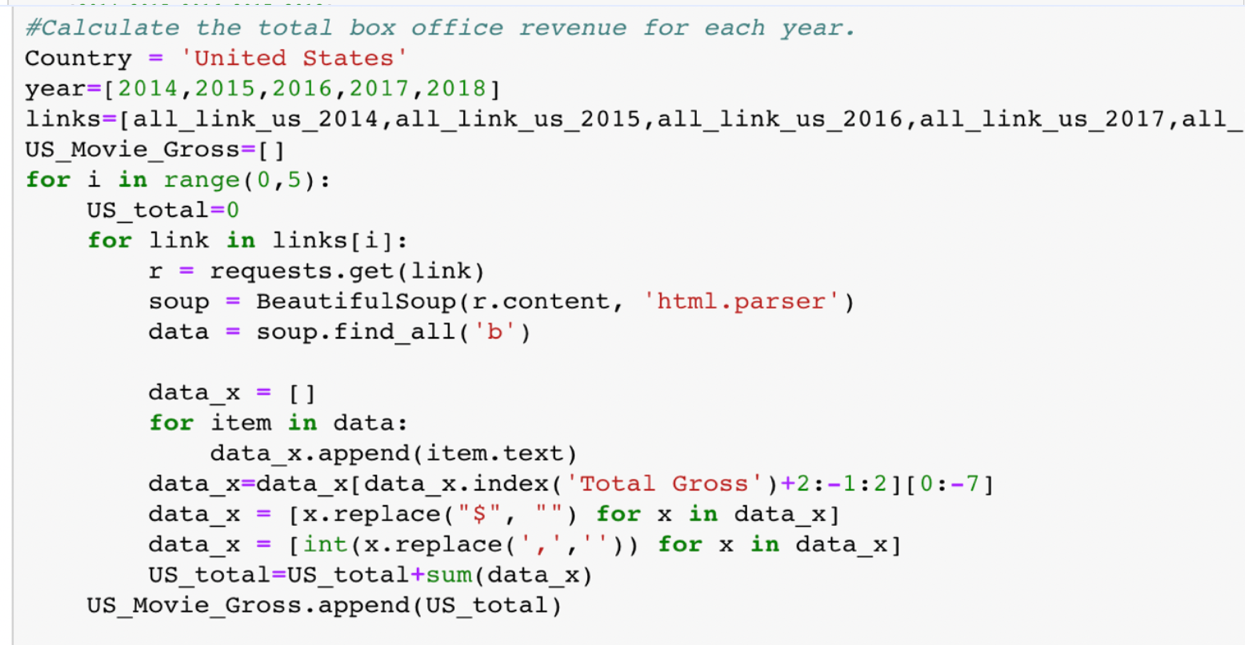
I found and scraped the box office revenue data information from the Box Office Mojo (http://www.boxofficemojo.com), which is the leading online box-office reporting service website. Box Office Mojo is owned and operated by IMDb (www.imdb.com), the #1 movie information website in the world.



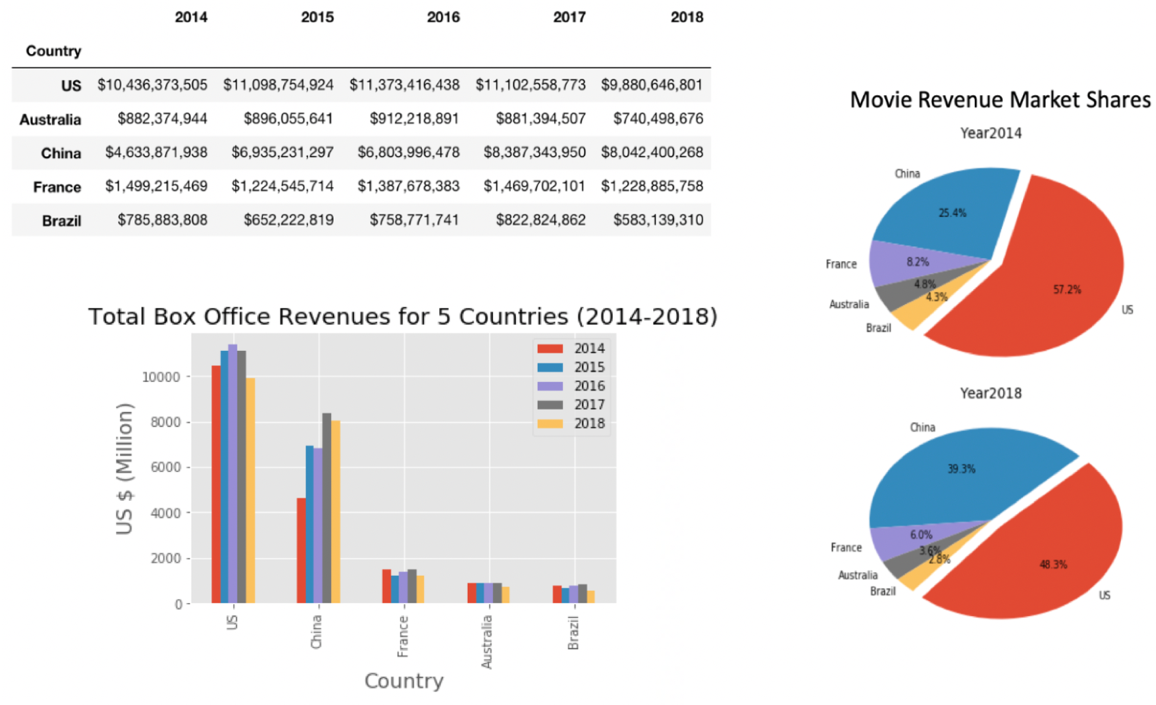
There is a list of the box office revenues of USA in 2018 (the picture above), we can easily find the production company and the exact revenue amount for each movie. We can also find other countries box office revenues in different years from this website. I obtained these data information by using web scraping techniques. For future analysis purpose, I selected 5 countries (Australia, Brazil, China, France and USA) and 3 major movie companies (Disney, Sony and Lionsgate) to scrape.

**Web Scraping**

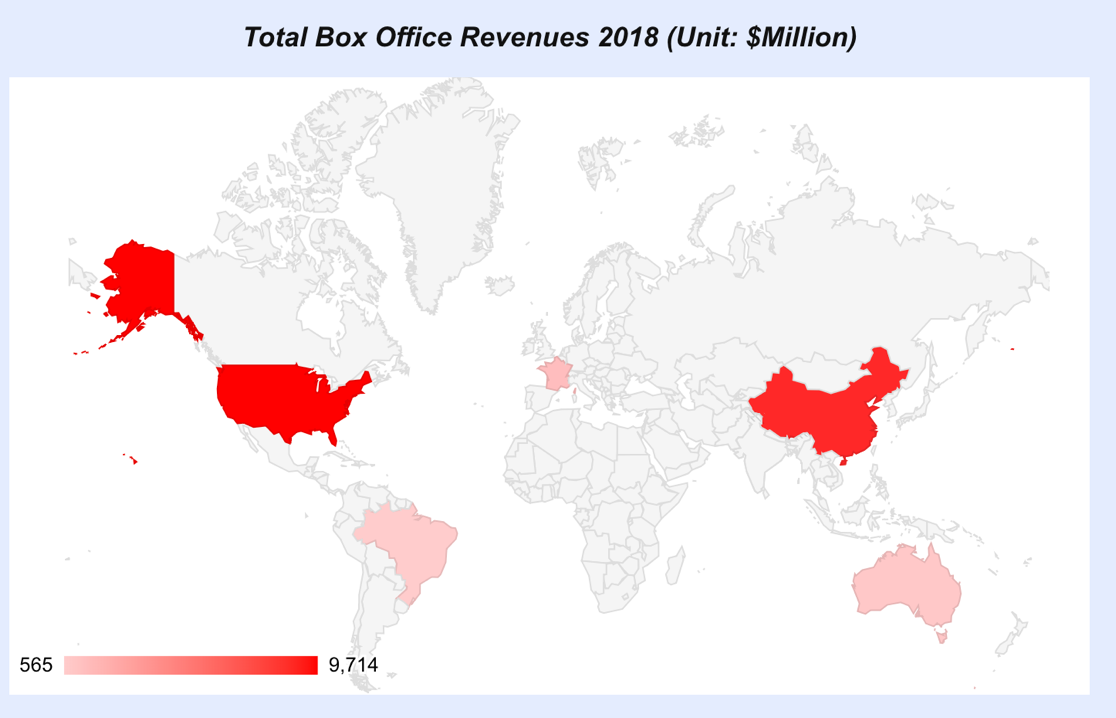
Web Scraping (also termed Screen Scraping) is a technique to extract large amounts of data from websites and save to a local file in your computer or database. For example, if I want to know the total box office revenues in USA in recent 5 years, I should select these URL links of USA, then open the page resource and find the html tabs before the exact revenue amount. There is a most popular and helpful python package named BeautifulSoup, which can be used to automatically scrape data by putting the selected tabs and patterns in that package. The scraping process details and codes can be found in the project codes folder.



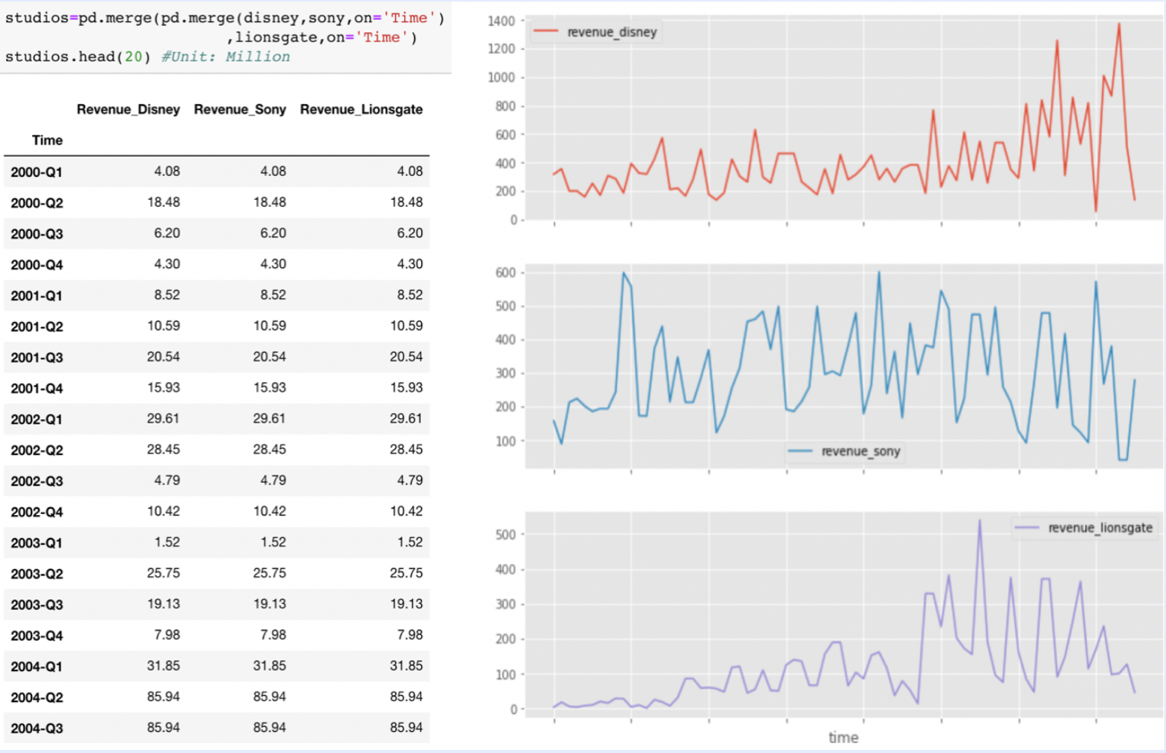
**EDA (Countries & Companies)**

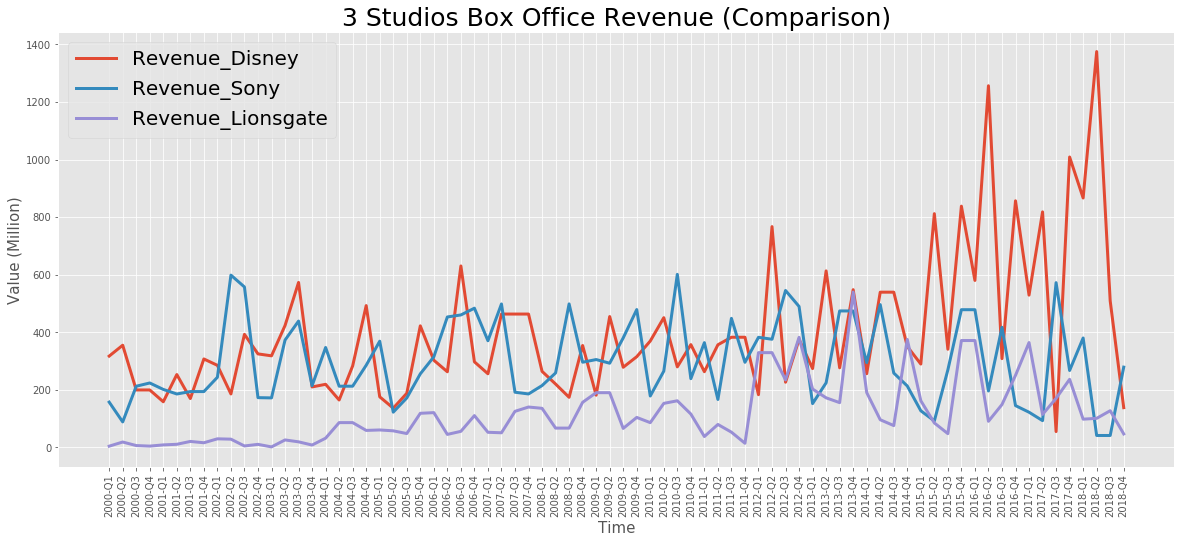


The above picture shows the scraped dataset and analytical charts of 5 countries box office revenues from 2014 to 2018. There are some interesting facts we can know from these charts, such as the pie charts can tell us that the total movie revenue market shares for USA was 57.2% in 2014, but now it becomes less than a half, which is 48.3%. Chinese movie revenue market share has a huge increasing in the recent years, from 25.4% in 2014 to 39.3% in 2018. We can also compare these countries box office revenues in different years, in the bar chart above.



The interactive distribution map above shows us these 5 countries box office revenues in 2018. The darker for the red color of the country, the more revenues in the country. When hovering the mouse on the country, we can see the exact value of revenue of the country. I built this interactive map by Google Map JavaScript API, which allows communication with Google Map Services and their JavaScript library.



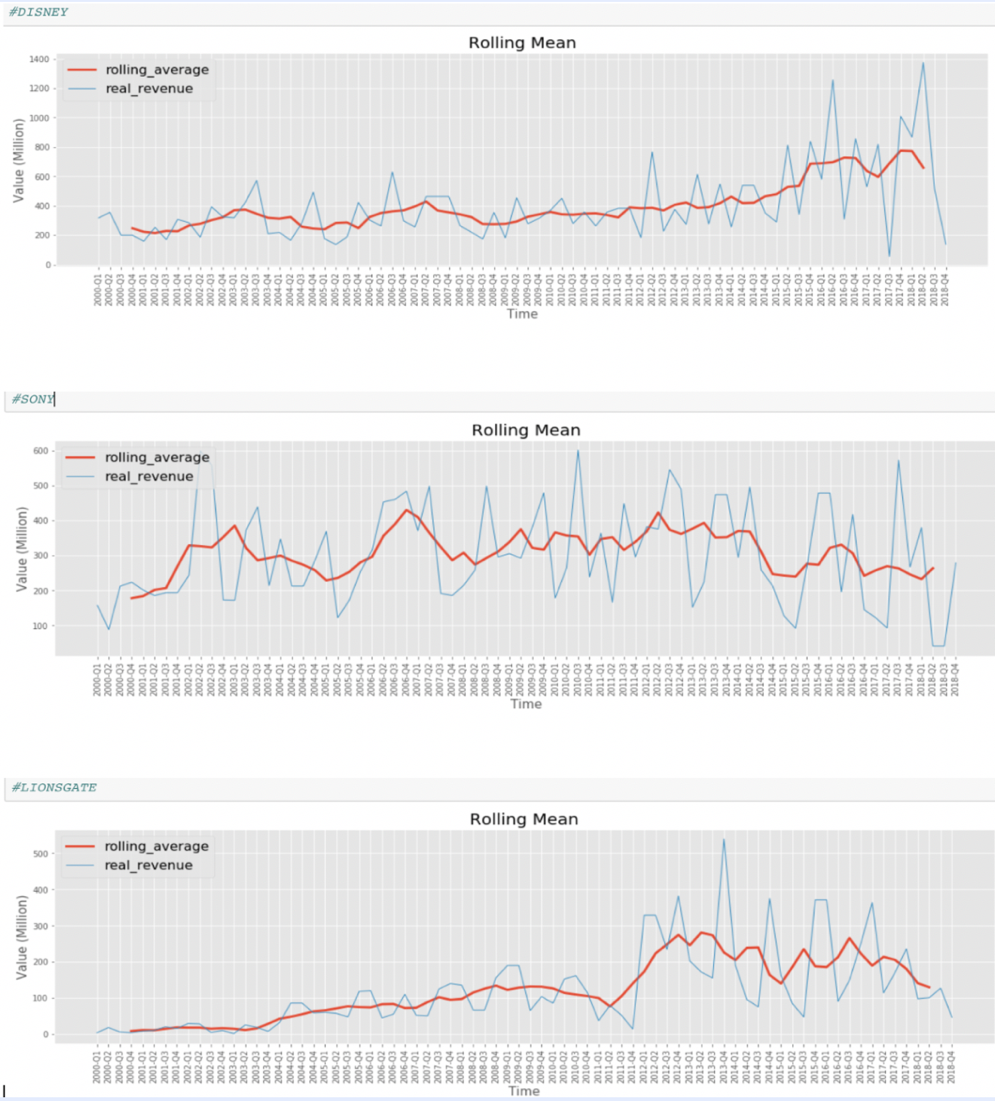


The pictures above show the scraped dataset and line charts of the 3 major studios total box office revenues from 2000 first quarter to 2018 fourth quarter. When I put their trends in one plot, we can see a clear comparison of each studio revenue situation. For instance, the revenue of Disney increases faster in the recent quarters; the revenues of Lionsgate were less than other two studios before 2012 but catching up others in the recent years.

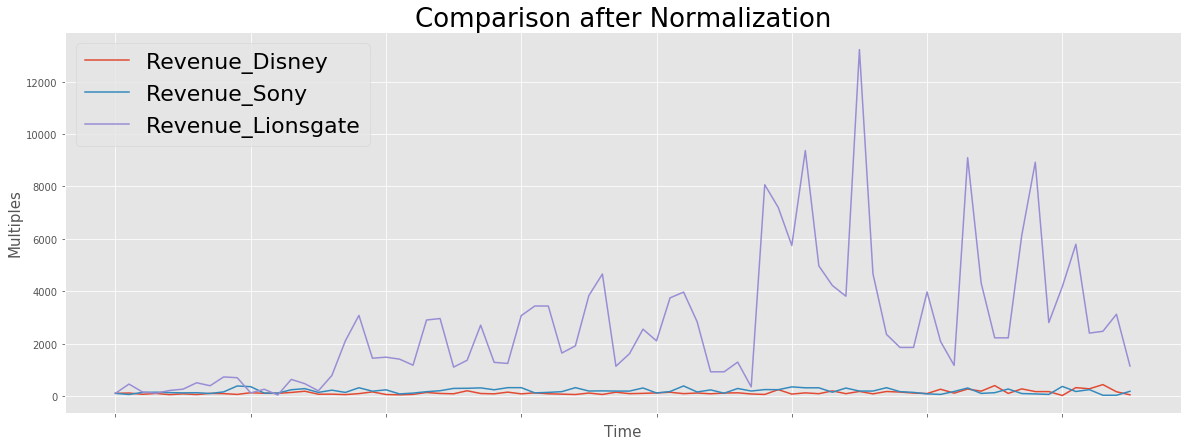
**Time Series Analysis**

I made an analysis for the 3 major studios by Time Series Analysis, which is a classic statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. There are some time series methods can be used to analyze whether the dataset has trend or not.

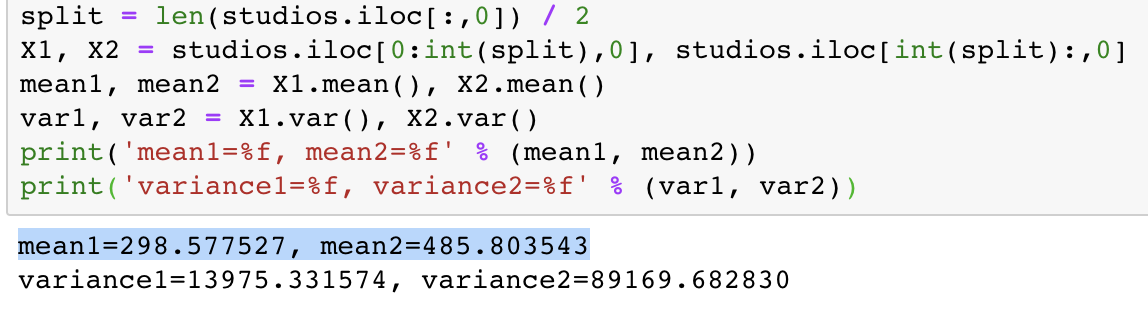
**Rolling Average** is a smoother version of the original trend plot. For example, by calculating the sum of previous 3 quarters revenue before 2010 first quarter and later 3 quarters revenue then divided by 6, we can get the mean value of 2010 first quarter. I calculated the mean value for every quarter, then plotted and lined them in the graph below (indicated by red line). We can see the tendency of revenues more clearly: From 2000 to 2018, Disney and Lionsgate are uptrend in general, but Sony is up and down.



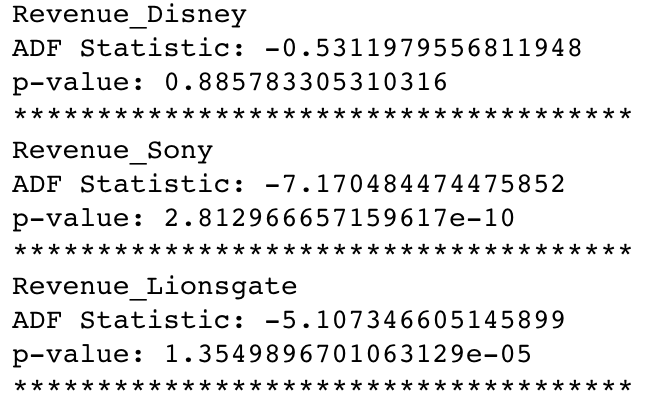
**Normalization** is another method to compare different studio revenue trends. The process is to divide each quarter's revenue from 2000 to 2018 by the first element, which is the first quarter in 2000, and then plotted and lined them in one graph (shown as picture below). This way can make these series all start from the same point and easily compare their growth rates. For example, there is even a quarter of the revenue of Lionsgate is more than 12000 times of the first quarter of 2000. We can conclude that Lionsgate outperforms other 2 companies’ revenue growths, because the normalized trend line of Lionsgate is higher than the other 2 studios over time.



In addition, testing the **stationarity** of the data is also useful in the time series analysis. The property of stationary time series is that observations in a stationary time series are independent to time and no trend or seasonal effects. The classical time series analysis and forecasting methods are concerned with making non-stationary time series data stationary. So, we need to check whether there is any evidence of a trend or seasonal effects and, if there is, we need to remove this effect first. For example, when testing the stationarity of the box office revenue trend of Disney, first I evenly split the time series into two contiguous sequences, then calculate mean and variance of each group. We can see the mean and variance of the two parts look very different (mean1 is not similar as mean2 and variance1 is not similar as variance2). So, the Disney box office revenue trend is a non-stationary time series.



The **Augmented Dickey-Fuller (ADF)** test is another method the check whether the time series is stationary or not. While the null hypothesis of the test is that the trend is not stationary (has some time-dependent structure); the alternate hypothesis is that the time series is stationary (the time series are independent to time).



Based on the p-values result above, we should reject the null hypothesis of Sony (p-value is much smaller than the significance level 5%) and conclude the time series is stationary for Sony, it does not have time-dependent structure; we fail to reject the null hypothesis of Disney and Lionsgate (p-values are higher than the significance level 5%), so the time series of these two companies are non-stationary, they have time-dependent structures.

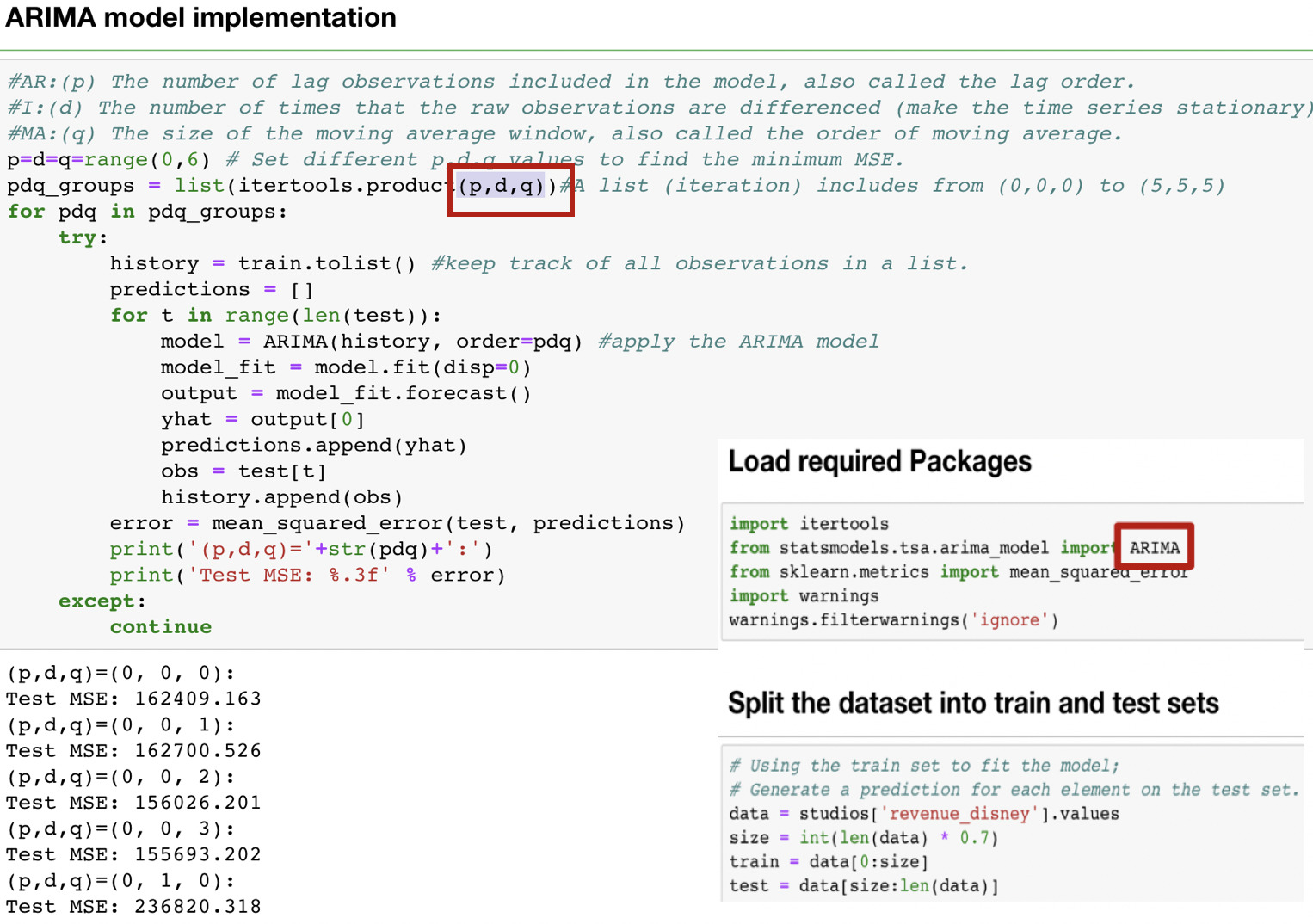
TheAutoregressive Integrated Moving Average Model (**ARIMA Model**)is one of the most popular and useful models in time series analysis, which can be used to **forecasting**.

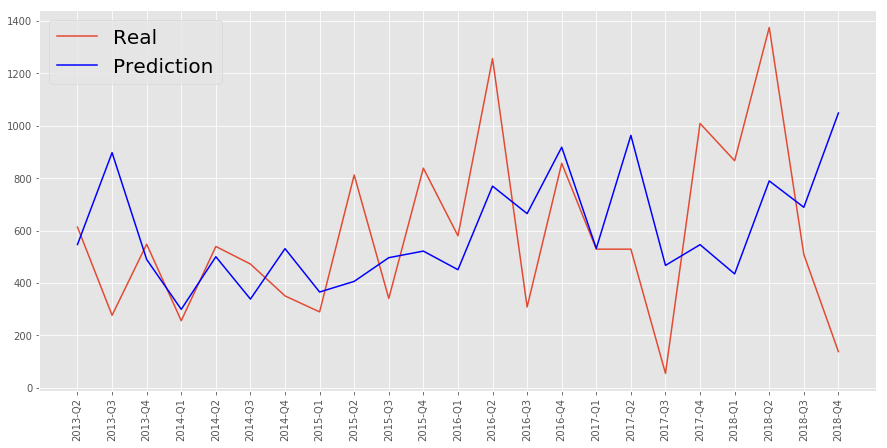
“AR” is stand for Autoregression. A model that uses the dependent relationship between a quarter revenue and some number of previous quarter revenues. For example, we use the 3 previous quarters before the 2011 fourth quarter box office revenue of Disney to predict the 2011 fourth quarter box office revenue. The autoregression model formula can be written as y\_2011Q4 = b1\* y\_2011Q3 + b2\* y\_2011Q2 +b3\* y\_2011Q1

“I” is stand for Integrated. The use of differencing of quarter revenues in order to make the time series stationary. (e.g. "I" = 1: subtracting the value of revenue in 2011 first quarter from 2011 second quarter only once.)

“MA” is stand for Moving Average. A model that uses the dependency between a quarter revenue and a residual error from a moving average model applied to previous quarter revenue. For example, if I set the moving average value to 3, I need to know the relationship of residual value of 2015 fourth quarter revenue of Disney and the previous 3 quarters’ residual values. The Moving Average model formula can be written as ε\_2015\_Q4 = b1\*ε\_2015\_Q3 + b2\*ε\_2015\_Q2+ b3\*ε\_2015\_Q1

There is a python package called statsmodels.tsa.arima\_model for the ARIMA model. The set of (p,d,q) are stand for number of lag observations included in the model for “AR”, the number of times that the raw observations are differenced for “I”, and current deviation from mean depends on the number of previous deviations for “MA” separately. By adjusting these 3 parameters from (0,0,0) to (5,5,5), we can find the model with minimum mean squared error (MSE). If we take a look at the plot of prediction box office revenue trend for Disney in the testing dataset, which from 2013 Q2 to 2018 Q3, by setting (3,1,0) as the value of (p,d,q), we can see the trend prediction (blue line) for Disney is generally similar with the real trend (red line).





This forecasting part by using ARIMA model is just a simple example. There are also many other better prediction models can be used in time series analysis, such as the LSTM model based Recurrent Neural Networks which I will research more in the machine learning class during this semester.

**References**

1. Justin Yek, “How to scrape websites with Python and BeautifulSoup”, Jun 10, 2017, <https://medium.freecodecamp.org/how-to-scrape-websites-with-python-and-beautifulsoup-5946935d93fe>

2. Ian Lurie, “How do I calculate a rolling average”, May 04, 2009, <https://www.portent.com/blog/analytics/rolling-averages-math-moron.htm>

3. Jason Brownlee on December 30, 2016, “How to Check if Time Series Data is Staionary with Python” <https://machinelearningmastery.com/time-series-data-stationary-python/>

4. Topta, “Is a random walk the same thing as a non stationary time series?”

<https://www.quora.com/Is-a-random-walk-the-same-thing-as-a-non-stationary-time-serie>

5. Siddharth Yadav, “Everything you can do with a time series”

<https://www.kaggle.com/thebrownviking20/everything-you-can-do-with-a-time-series>

6. Gurchetan Singh, “7 methods to perform Time Series forecasting”, Fevruary 8, 2018

<https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/>