Data Analysis for Shipping Dataset

1.Summary of Shipping Dataset

This dataset contains 2,697,549 entries and 60 features. It consists of shipping records from January to December in 2011. These features can be mainly described as the following aspects:

-Product Type (SI/BK)

-Timestamp

-Carrier

-Name and Address (Street, City, State, Country) for Requester, Shipper, Forwarder, Consignee, and Notify party

-Origin(Destination) City and Country

-Move Type

-Cargo Description

-Package Count and Cargo Weight

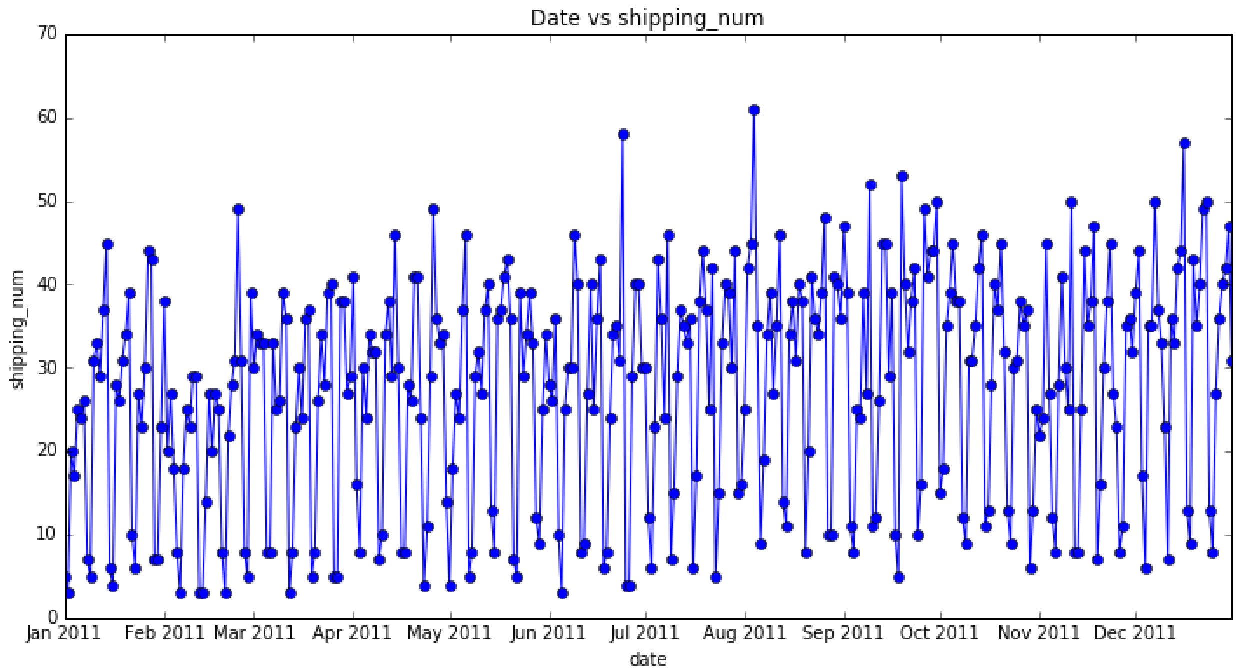
2.Detailed Analysis

2.1. Time Based Analysis

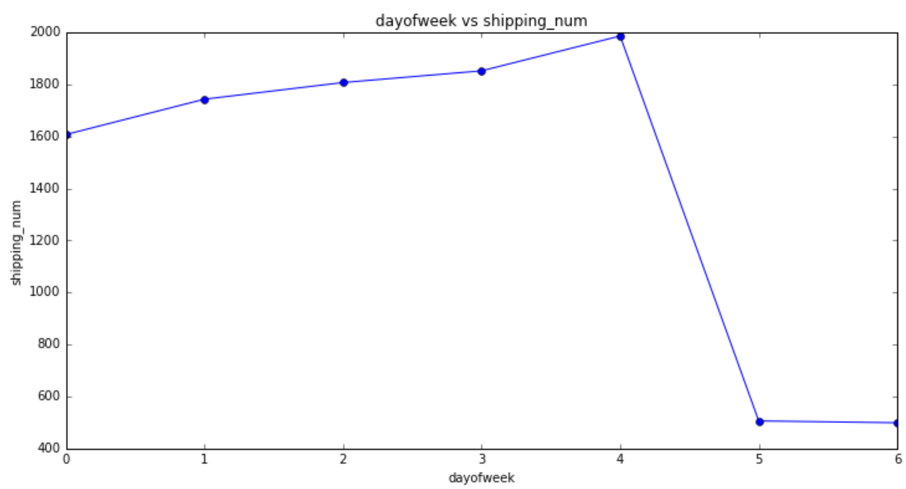
The following analysis is based on a sample with 10,000 rows which has been randomly selected from the original dataset.

The original format of timestamp in this data is like 'Year-Month-Day Hour: Minute: Second'. Then in order to explore some patterns related to month and day of week (weekend/ weekday), which is essential in the next modeling phase, 2 column features ('month', 'day of week') are added to the dataset.

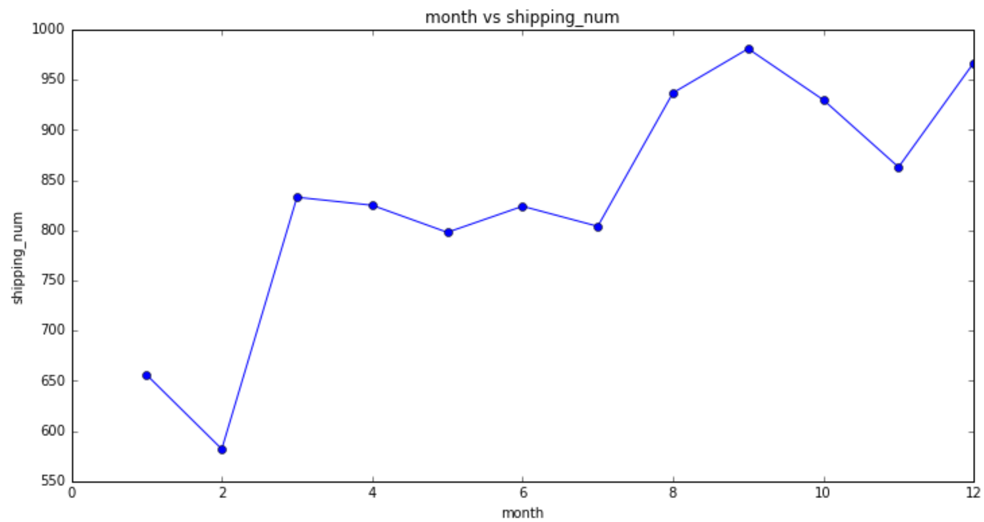
1) This plot shows the fluctuation of total shipping number for each day.



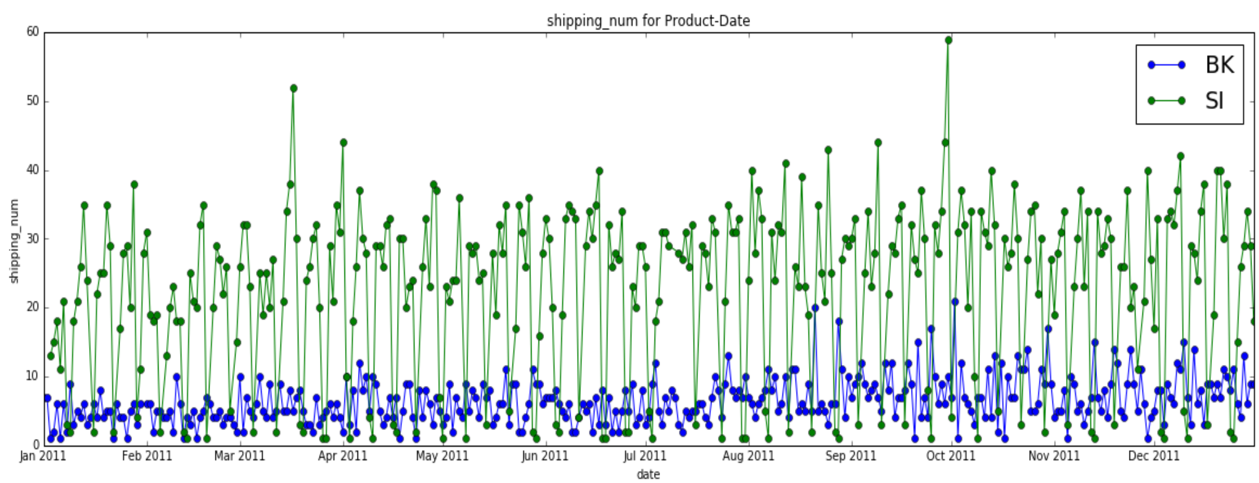
2) This plot of total shipping number for each day of week suggests that there's a significant decrease for weekends.

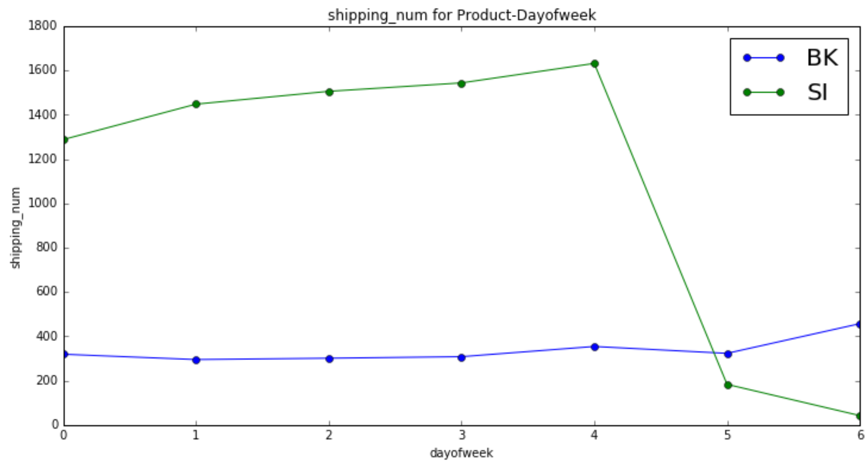


3) This plot of total shipping number for each month suggests that there's an overall increasing trend throughout the year.



4) There are two types of product in this shipping dataset, which are ‘SI’ and ‘BK’. This plot of total shipping number for each product type across the whole year shows on average the shipping number of product SI is higher than that of product BK. The plot of total shipping number for each product type on each day of week suggests that product SI is more sensitive to the effect of weekends, since the number of shipping for product SI drops dramatically on weekends. (0 corresponds t0 Monday, while 6 corresponds to Sunday)



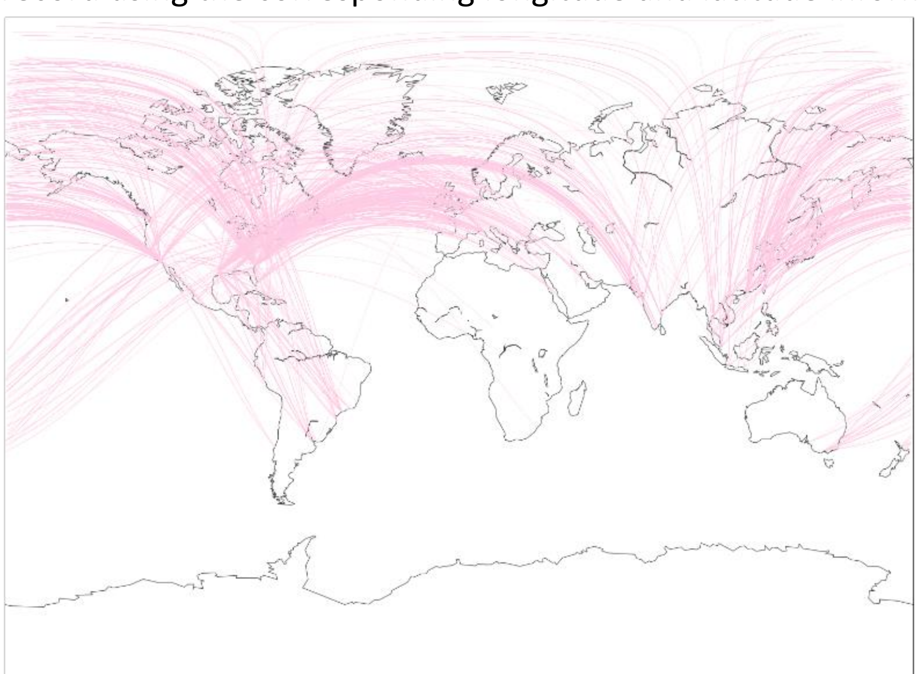


2.2. World Map with Shipping Plot

Use ‘pygeocoder’ package provided by Google’s geo-API to transform the physical address of each shipping record (‘origin city’ and ‘destination city’) to longitude and latitude information, then add 4 column features (‘origin longitude’, ‘origin latitude’, ‘destination longitude’, ‘destination latitude’) to the dataset.

From the randomly selected 10,000 sample, select those whose ‘Carrier’ name is ‘HAPAG-LLOYD’ in the sample dataset to generate a new dataset with about 2,673 rows.

Then use the ‘basemap’ package to make a plot of world map, which also contains the shipping route information from origin city to destination city for each shipping record using the corresponding longitude and latitude information.

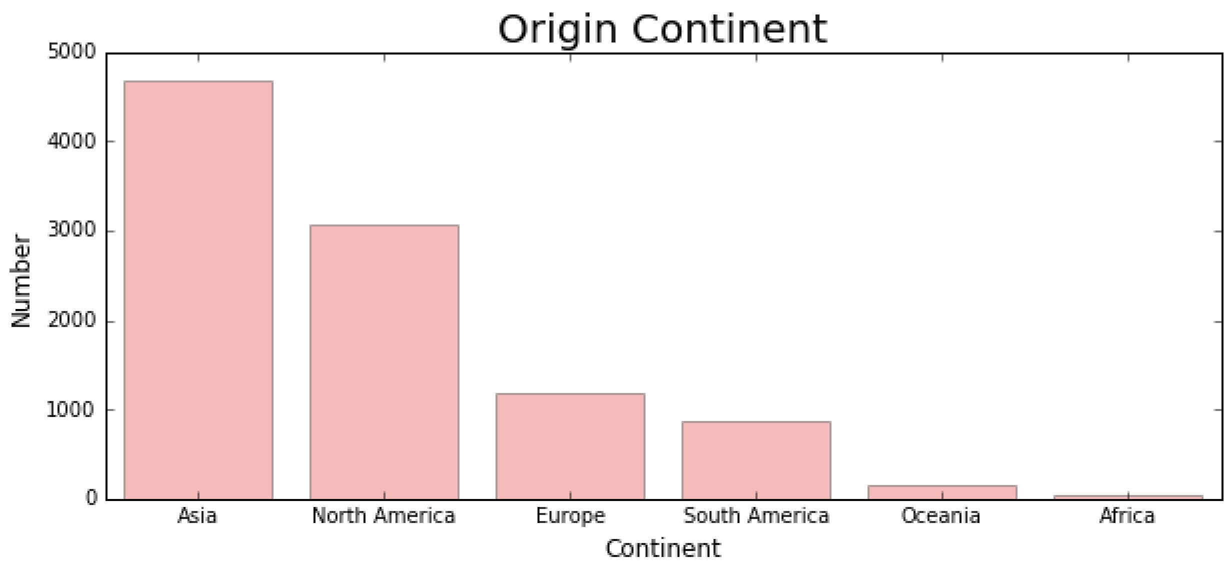


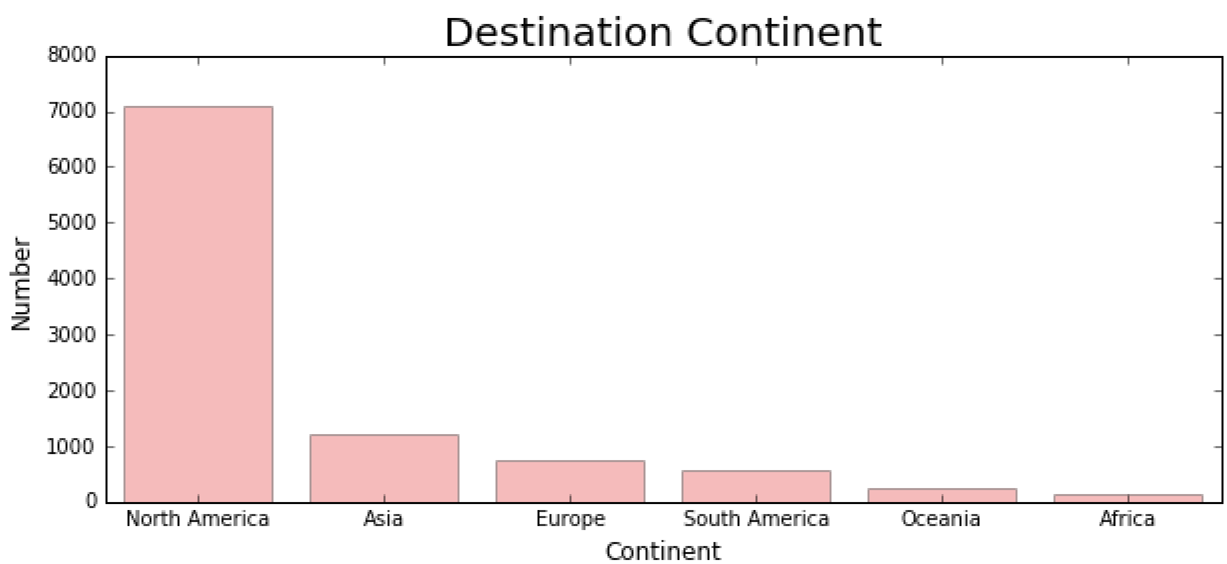
2.3. Continent Analysis

The following analysis is based on a sample with 10,000 rows which has been randomly selected from the original dataset.

Use ‘incf.countryutils’ package, which is a convenient API for transformations between country code and continent, to map the origin country and destination country for each shipping record to the corresponding continent, then add 2 column features (‘origin continent’, ‘destination continent’) to the dataset.

There are both six categories of continent for shipping origin and shipping destination. This plot of total shipping number for each continent suggests Asia is the largest export continent, while North America is the largest import continent.





2.4. NLP for Cargo Description

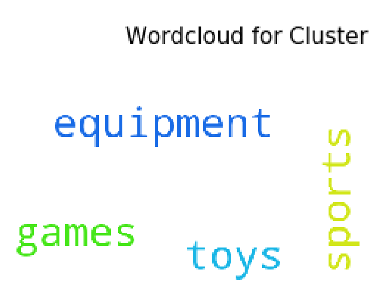
Since the cargo description column is in the text format of which the values are all very long and messy description, so we want to transform this kind of values to concise cargo categories.

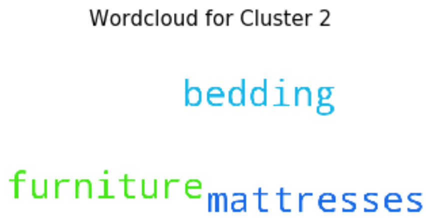
The following analysis is based on a sample with 152,589 rows of which the cargo description has the frequency with more than 1000 among the original dataset.

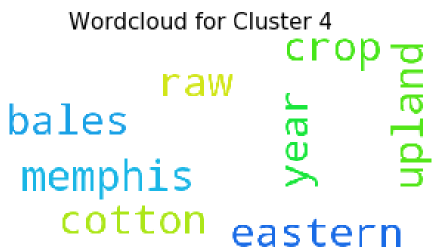
Detailed data manipulation:

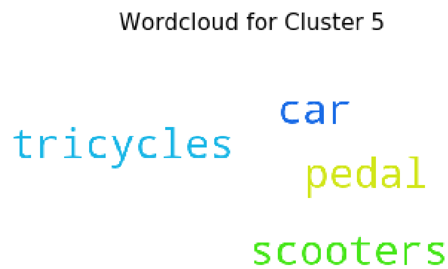
- Delete the punctuation, numbers and stop words for each of the cargo description.
- After the previous step of deletion, some of the cargo description may be null. The number of these null is about 29,925, so our data decreases to 122,664 rows.
- Do word stemming to transform words into their root form.
- Use "CountVectorizer" to convert text into a matrix of token counts to do word bagging. It includes two procedures. The first is to create a vocabulary of unique tokens (or words), and then construct a feature vector for each cargo description, which stores the count of words per cargo description. This feature vector has 122,664 rows and 100 features (unique words). For each cargo description, if it contains the corresponding word, then it will be marked as 1, and 0 otherwise.
- Then based on this feature vector, I have performed clustering(k_means) to divide them into 8 categories. Each cluster will contain several words. For example, cargo descriptions which belong to the 6th cluster have the following words: Scrap, Metal, Plastic, Mixed, Silicon.

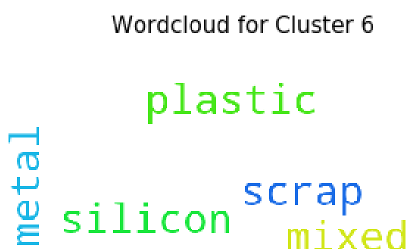
  Here’s the word cloud for each cluster:

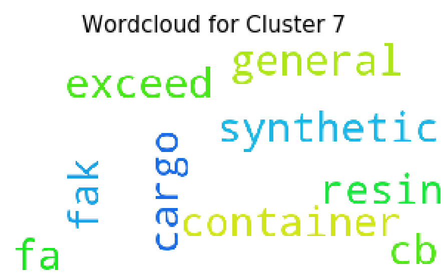




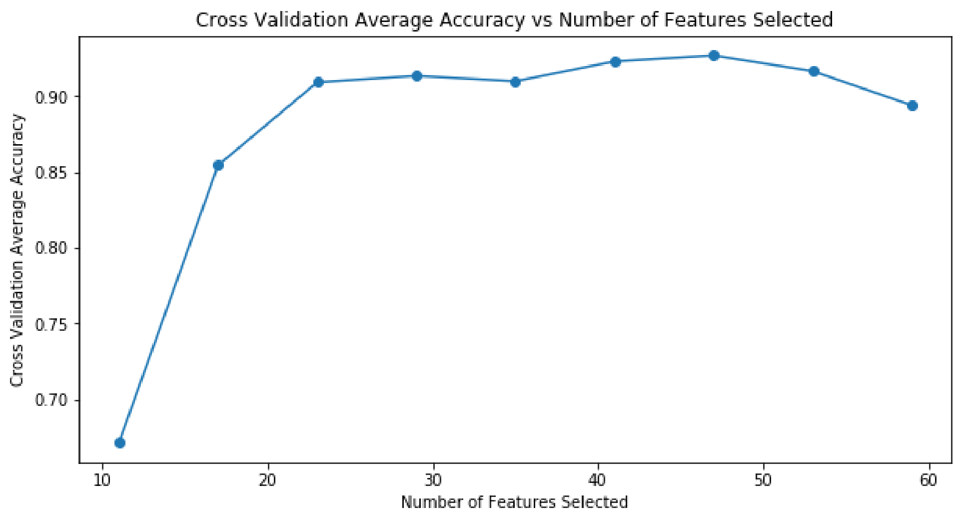








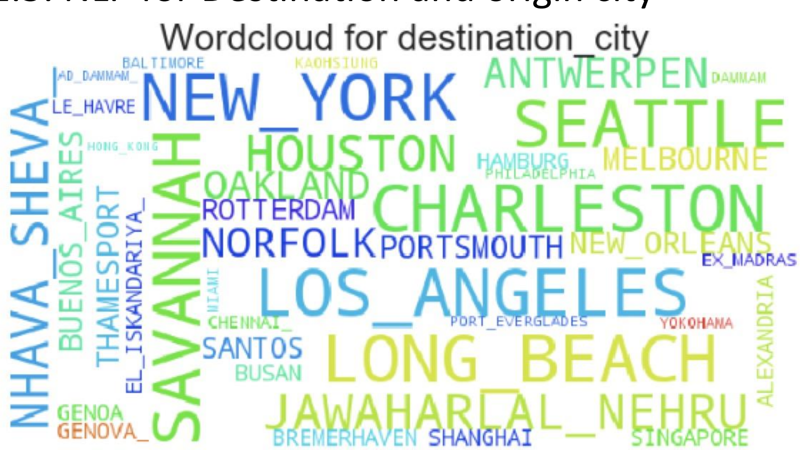
- Then I use these 8 categories as target labels to train a classification model which can be used to predict which cargo category is on the ship for each shipping record.
- Firstly, I perform the feature selection and find out the cross validation average accuracy is the highest when using about 40 to 50 features to build the model, which is about 92 percent. The size of full features is 59, and its cross validation average accuracy is about 89 percent.



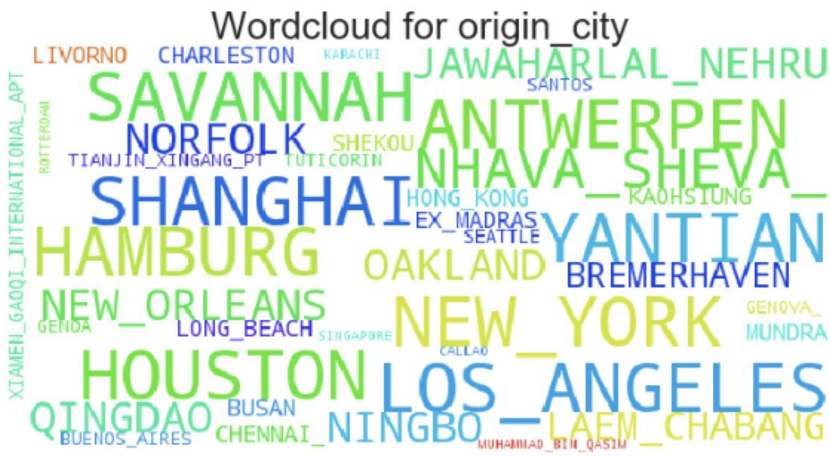
- Then using Decision Tree method to build this classification model, and the feature importance is as follows:

	feature	weight
0	notify_party_country	0.413381
1	consignee_state	0.186875
2	origin_country	0.097868
3	consignee_global_name	0.060882
4	forwarder_global_name	0.060611
5	notify_party_city	0.056478
6	transaction_id	0.042512
7	notify_party_street	0.027773
8	notify_party_postal_code	0.024694
9	notify_party	0.010654
10	forwarder_postal_code	0.008935
11	origin_voyage_number	0.003547
12	notify_party_state	0.002711
13	shipper_city	0.001798
14	requester_global_name	0.001282

2.5. NLP for Destination and origin city



Destination city counts reflects importing and consuming level of a city



Origin city counts reflects production and exporting level of a city

2.6. Visualization and analysis of clustering:

Algorithm: Truncated SVD, K-means, t-SNE

Clustering of “cargo_description” feature in the shipping data

A first look at the cargo description data:

```
0           17 X 20' CONTAINERS  SAID TO CONTAIN  297,500 ...
1           1 EMPTY RETURN ISOTANKS (SHIPPER OWNED  CONTAI...
2           BD 12 CUP DIGITAL COFFEEMAKER  BLACK  P.O.NO.:...
3           LUGGAGE SETS  PO NO:1852575497,  PO TYPE:0023 ...
4           FRESH APPLES  RECORDER NO. 18050101  VENTS 1/4...
5           S.T.C. HARDWARE STORE SUPPLIES    AES ITN: X20...
6           BUD1207 GDPK  SYNTHETIC RUBBER  NCM #: 4002.20...
7                                 PLASTIC HOUSEHOLD ARTICLES
8           BD 12 CUP DIGITAL COFFEEMAKER  BLACK  P.O.NO.:...
9           PO NO. 7952627119  - PO TYPE. 0023  - TOYS  - ...
10          ITEM DESCRIPTION:  GAS GRILL - BLACK  P.O.NO.:...
```

Noting that it contains numbers, stop words, punctuations which are irrelevant for clustering.

Data Selection:

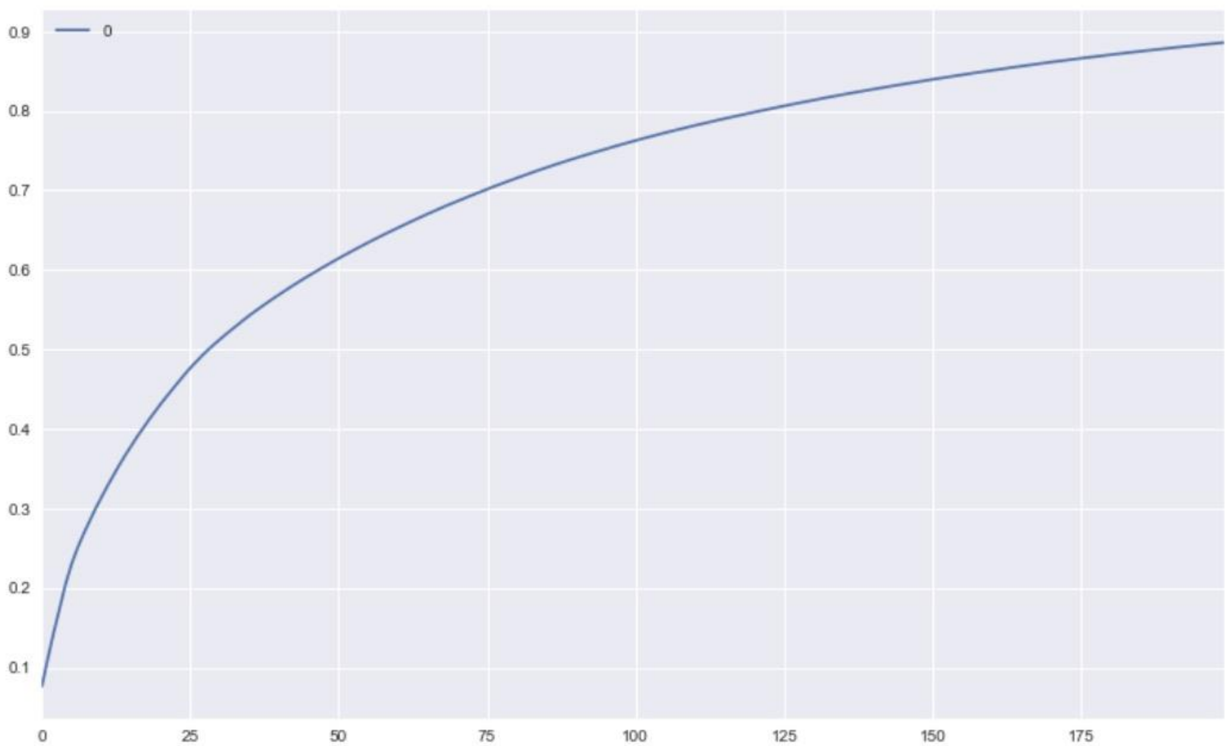
We only keep the records of cargo_description that appear more than 100 times, for the purpose of selecting representative data and running speed

After the data selection and cleaning:

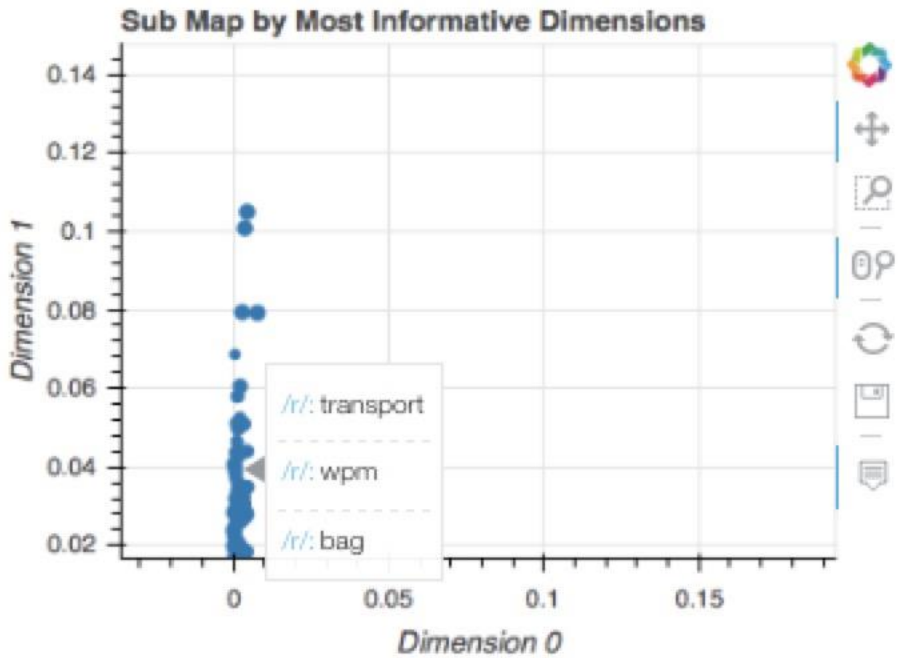
0	bud gdpk synthet rubber ncm metal box ncm x st...
1	NaN
2	empti wooden barrel
3	float glass
4	multi pli bag calcium casein kg net
5	contain exceed cb synthet resin
6	non haz chemic
7	cntr vehicl part invoic bs lr
8	NaN
9	guar gum powder
10	mix metal scrap
11	NaN
12	NaN
13	roll kraft paperboard

Noting that if there are only numbers, stop words and punctuations in the description, we ignore it, i.e NaN. Basically we follow the data manipulation process in the 2.4 NLP for cargo description part.

Use Truncated Singular Vector Decomposition for Dimensionality reduction:
We can capture around 88% of the original matrix (information) with the first N dimensions. Truncated because we only want part of the computation



Visualizing these dimensions:
Use Bokeh package to get hover-tooltips
Every dot in the plot is a cargo word scaled by size, pull mouse over to see it



We use KMeans from scikit-learn to cluster the cargo words into groups of buckets. Here, 8 groups are represented by 8 colors. We can manipulate the number of clusters for efficiency and quality. We use TSNE, t-distributed stochastic neighbor embedding, to nonlinearly reduce dimension. After embedding the data into a space of two or three dimensions, the scatter plot is visualized and human readable.

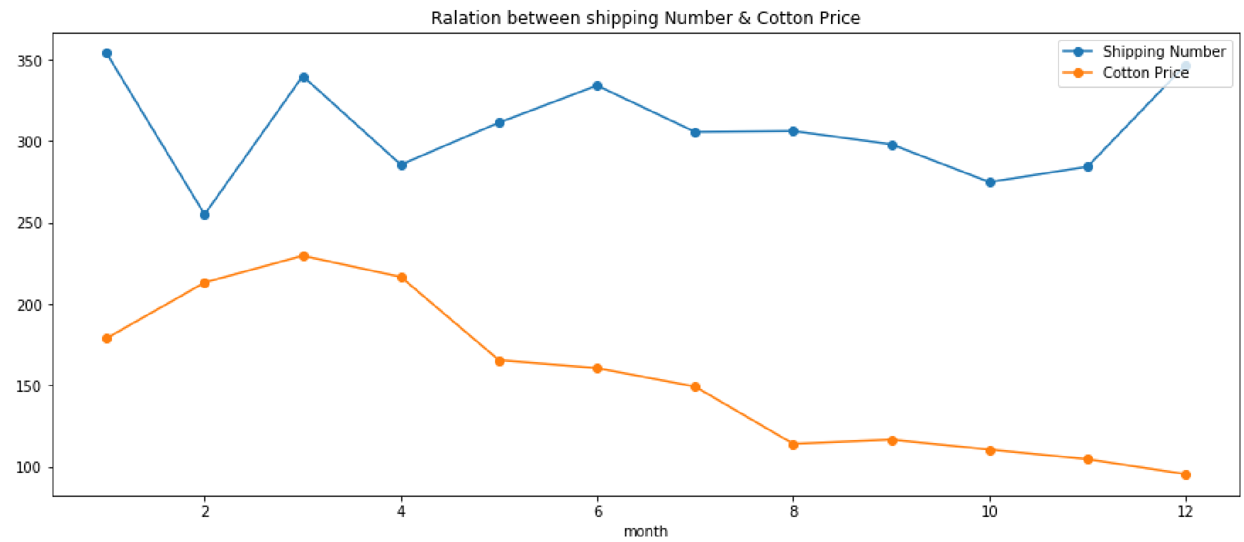
The points with same color are in the same cluster. From the example, concentr protein and milk are in the same cluster.

2.7. Cotton Price Analysis

From the initial dataset which contains about 2,697,548 number of shipping records, I choose the shipping records of which the cargo description contains the word 'cotton' to create a new dataset. Then the number of rows in this new dataset is 73,928. I calculate the number of shipping records for each month.

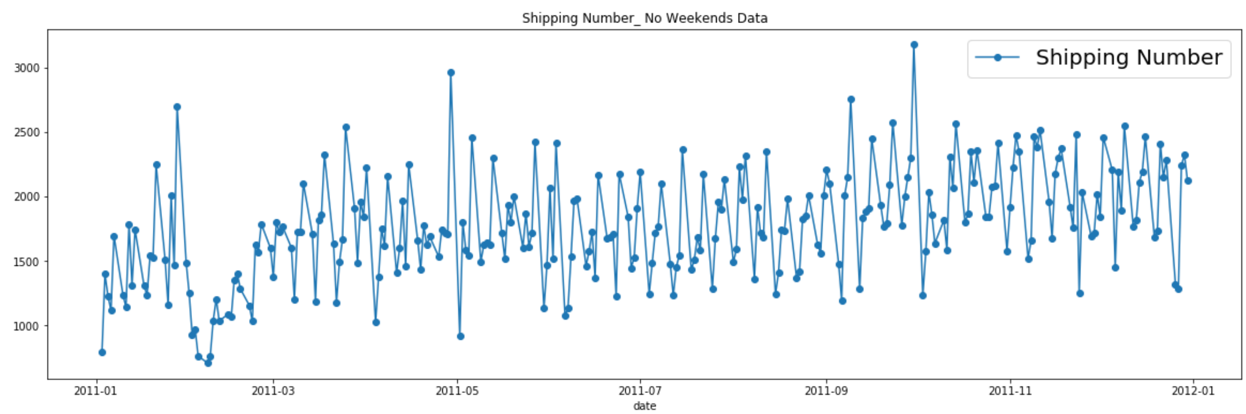
I have also got the global cotton price for each month in 2011, and the price unit is US cents per pound.

Then based on these two data, I make this plot of the relation between shipping number and cotton price for each month in 2011. The overall trend seems to be like the cotton price will decrease along with the increase of shipping number for cotton, which means a negative correlation.

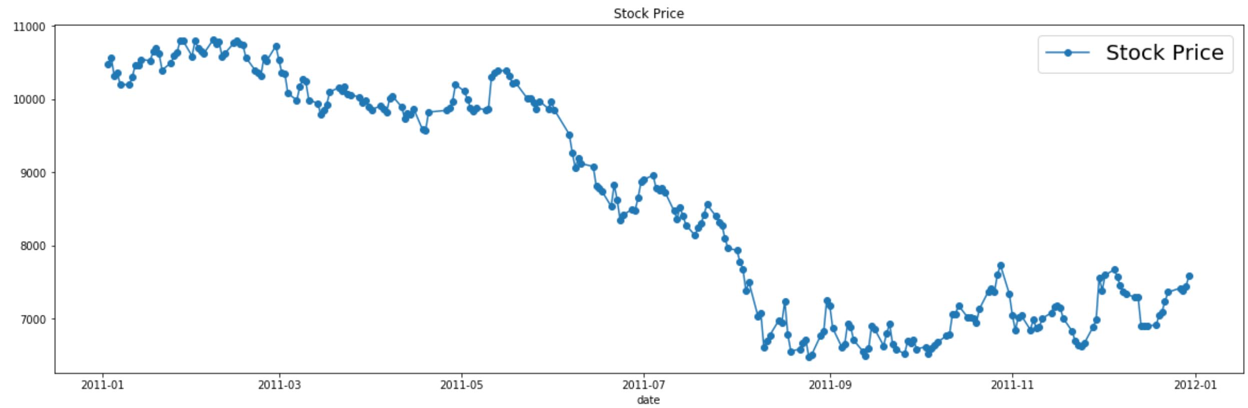


2.8. Carrier Stock Price Analysis

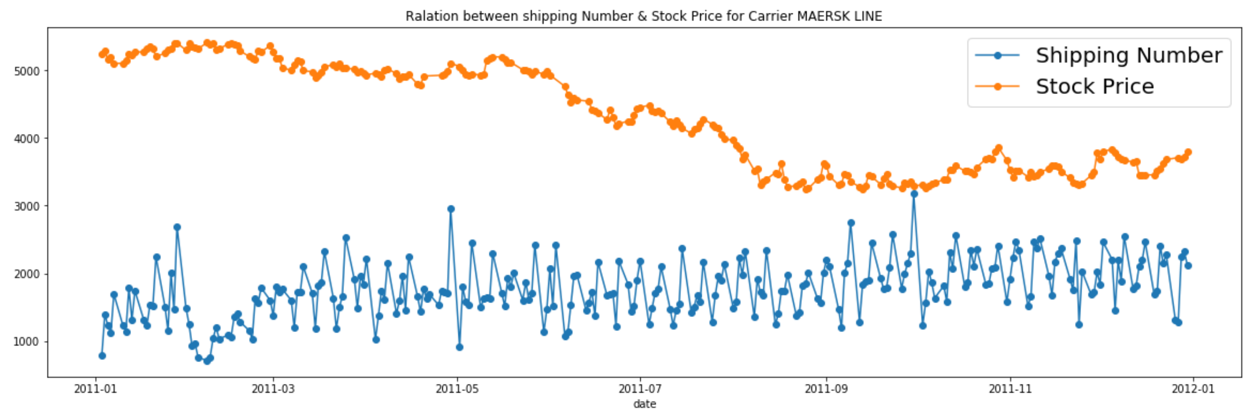
From the initial dataset which contains about 2,697,548 number of shipping records, I choose the shipping records of which the carrier name is 'MAERSK LINE' to create a new dataset. Then the number of rows in this new dataset is 475,981. I make this time-based plot of the number of shipping records for each weekday (eliminating the influence of weekends data).



I have also got the stock price in 2011 for this carrier- 'MAERSK LINE', which is one of the world's largest container shipping companies. Then I make this plot of the stock price for each weekday.



Then I combined these two plots together to see whether there's some relation between the stock price of this carrier and the number of shipping records this carrier has performed. The stock price has been divided by 2.



And here's a plot of monthly basis. The stock price has been divided by 5.