Data Wrangling Report

Rucha Joshi

August 6, 2018

Data Set:

The dataset was received from the ship chartering company.
The dataset includes about 185,000 observations which has voyages from 2009 across the world. The other datasets are Port details, Vessel details and Voyage details.

The data includes following important fields -
1. Voyage Id - Id assigned to each voyage
2. Vessel Id - Id assigned to each vessel
3. Commencing Date - Start date for the voyage
4. Completing Date - End date for the voyage
5. Vessel types - There are total 12 different types of vessels.
6. Vessel description - Is vessel a Tanker or a Bulk
7. Cargo types - There are 20 unique values of cargo types.
8. First load port - This is the port where first vessel got loaded.
9. Last discharge port - This is the last discharge port for the voyage. There can be multiple discharge ports for single voyage.
10. Cargo type - It gives information about what type of cargo vessel is carrying.
11. Cargo lift - It tells us how much cargo the vessel is carrying.
12. Dwt - Deadweight tonnage of the ship.
13. Estimated earnings and Actual Earnings
14. Trade Area - The area where the voyage is travelling.
15. Estimated voyage days and Actual voyage days
16. Ballast days - when vessel is empty
17. Laden days - when vessel is loaded
18. Base Currency - Currency of earnings
19. Operators details
20. Laden and Ballast Speed - this is finalized at the charter party.
21. Vessel properties such as built year, owner etc.

And there are few more fields for internal use of the shippers which are not relevant to our analysis.

Data Wrangling:

The dataset was comprehensive with few missing values. It required some cleanup and reformatting. The steps taken are described below. For detailed review of the code, R markdown document explaining each step with code is included.

Rows which were not needed for analysis were removed for example observations which had Id = -1 were removed from the dataset. In addition, column headings in the data set needed to be transformed from ambiguous names which gives clear description of each variable.
The dataset imported into R was stored in the data frame where I checked for NA's and spaces and tried to remove factor levels. These transformations were helpful to conduct preliminary exploration and data visualization.

In the ports data few observations had 0 longitude and 0 latitude, so we decided to remove these records.

Now data had lot of categorical variables like vessel type, cargo type; so, I tried to combine similar types into one category so that we can have limited categories to find the trends in the data.

Similarly, first loading port, last discharge port and trade area were replaced with region names so that we can find which are the regions gives more earnings and which trade area has more earnings.

For a single voyage there are multiple records with different data types as A - actual, E - Estimate and P-positional. We decided to use actual data for our analysis as estimate and positional type had lot of missing values.

If first loading port and/or last discharge port has blank value or actual earnings has zero value then those voyages got cancel due to some reason.

After considering all these factors from the data and cleaning up the data, now data is ready for further analysis.