# **ACME Data Analysis Report**

### **Anusha Palisetty**

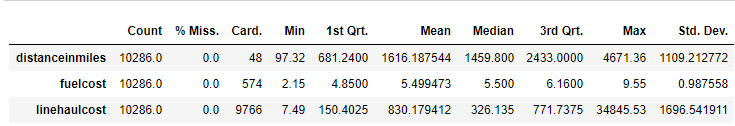
## **Introduction:**

For the Acme Inc manufacturing company, the shipping activity of the products in Acme inventory are captured. We perform deep data analysis with statistics modelling providing the descriptive statistics, relationship between features, check the correlation, identify the most expensive trucking companies, companies. Finally, we implement regression model and predict the “linehaulcost” for each shipping activity.

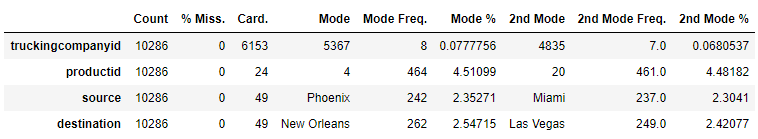
## **Summary Statistics:**

To start the analysis, we take the Acme data with 10,286 rows and 7 features. To get to know more about the data we create data quality reports, that describes the characteristics of each feature using statistical measures.

a) Continous Features



b) Categorical Features



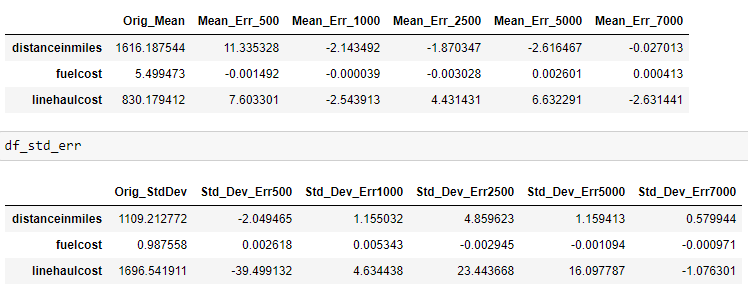
**Observations:**

1. From the report we observe that there are no missing values in the data.
2. For the “linehaulcost” data, the difference between 3rd quartile and Maximum (34,073.7925) is noticeably larger than Median and 3rd quartile (445.6025), this suggests the maximum value is unusual and like to be an outlier.
3. The cardinality of the “distanceinmiles“ is 48 which indicates some irregular cardinality. For 49 different source locations and 49 destination locations, we would have around 2,400 routes. There might be a chance that each distance may be same for few routes but here it looks like the 48 “distanceinmiles” values are repeated for all the 2,400 routes.

## **Sampling Methodology:**

The dataset contains total 10,286 shipping activities, if we want to sample data from this population, we need to make sure the we don’t loose the statistical significance of the population. We consider probability sampling here where each row has the same probability of being selected. We consider different sample sizes and compare the mean and standard deviation of population with that of sample data.

The above report shows the mean and std of population and the error of sample for different sizes (difference between average of population and average of sample).

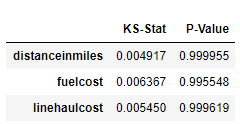


**Observations:**

We observe that when small sample size is considered we receive high error rate. But as the sample size increases the average of the sample data is getting closer to the population, as per Law of Large Numbers.

**Comparing Probability Distribution between Population and Sample:**

Kolmogorov-Smirnov test can be used to compare the distribution functions of continous variable between sample and population, here we consider 7,000 as sample size. The null hypothesis is that sample has same distribution as the population. We can reject the null hypothesis if p-value for any feature is < 0.05 confidence level.

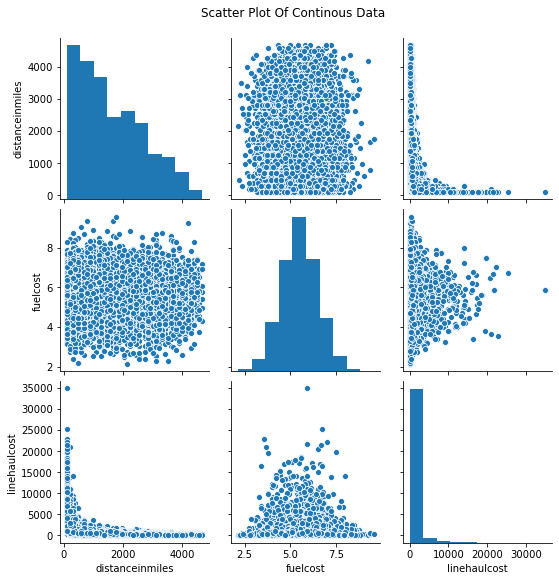


As the p-value for all the features is >0.05 we fail to reject the null hypothesis, and the KS test has been accepted. Thus, we can say our random sampling process can be used for data analysis with a sample data of 7,000 rows.

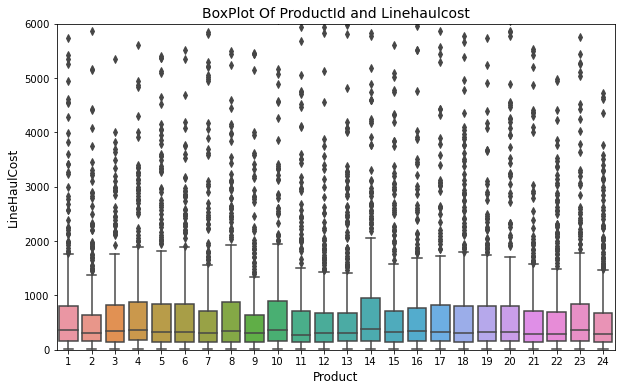
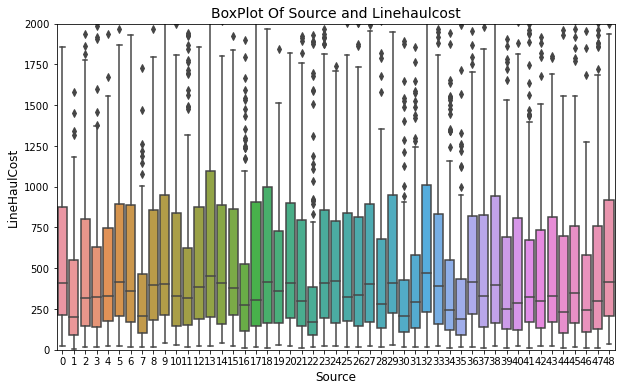
## **Relationship between features**

Visualizations are the best way to investigate the relationship between pairs of features. This helps us to know which feature might be useful to predict the target feature and help find the descriptive features that are closely related.

**Visualizing Pair of Continous Data:**



**Visualization of Categorical Data and Continous Data:**

**Observations:**

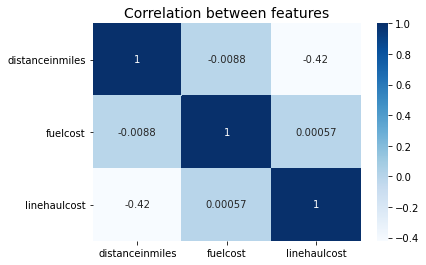
From the pair plot between “distanceinmiles”, ”fuelcost”, ”linehaulcost” the diagonal plots show the distribution of each feature and the lower matrix shows the relationship between them.

From the plot we observe that **no linear relationship exists “linehaulcost” and the remaining descriptive features.**

For the categorical and continous data, when a relationship exists between two features, the central tendency of the box plot should show some differing. So, from the box plots we observe that there is no differing in the central tendencies of “productid” and “source” features with respect to “linehaulcost”. Therefore, no linear relationship exist between “productid” , ”source” w.r.t “linehaulcost”.

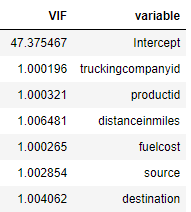
## **Collinearity:**

Correlation is a good measure of relationship between two continous features. Below is the correlation between “distanceinmiles”, ”fuelcost”, ”linehaulcost”.



**Variance Inflation Factor:**

This metric is used to detect the multi-collinearity in regression analysis i.e correlation between predictor variables in a model. The VIF between 3 to 5 are said to have high correlation and it can effect the regression model.



The VIF estimate is for all the variable is around 1, which indicate that collinearity doesn’t exist for the predictor variables.

## **Top 10 Most Expensive Trucking Companies**

To calculate the most expensive trucking companies, we calculate the linehaulcost per mile and fuelcost per mile, summation of these two variables gives the total cost for each “truckingcompanyid“ per “productid” for each shipping activity.

**Calculations:**

For each “truckingcompanyid“ we calculate average of the total\_cost across all the products.

**Step 1:** Truckingcompanyid, count(productid), sum (linehaulcostpermile, fuelcostpermile)

**Step 2:** Truckingcompanyid, sum (linehaulcostpermile, fuelcostpermile)/ count(productid) as final\_cost

From the step2 we extract the below top 10 trucking companies, sorted by “final\_cost”



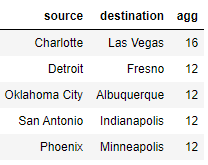
Assumption: Assuming the fuel cost at the time of pick up is the total fuel cost for the complete shipping activity.

## **Companies Services the most source-destination Routes:**

Most source-destination routes are the routes which have more frequency.

**Calculations:**

**Step1:** We calculate the source-destination frequency by grouping on source, destination and get the count of activities.



**Step2:** We see Charlotte-Las Vegas has more number of shipping activities and extract all the trucking companyid with this route.



Till now we have done the data exploration and learned that no linear relation exists between predictor variables and response, and no correlation between the predictive variables and the statistical significance of sample data is close to the population. After sampling we save the sample 7,000 rows to training\_data.csv and remaining rows to test\_data.csv.

## **Modelling:**

Now we perform machine learning modelling using the training and test data we saved earlier. The main focus is to predict the “linehaulcost” using the response variables. Here the response variable is continous data so we perform regression model and as we don’t find any linear relationship we implement non-linear regression models

## **Metrics to Evaluate the Model:**

In order to evaluate a regression model, or compare between models we use R2 and root mean square error for all the models and select the model with lowest root mean square value and high R2 score.

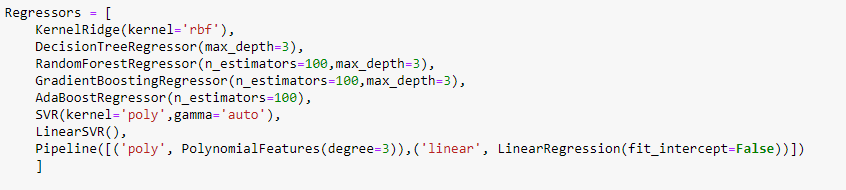
**RootMeanSquared Error:** The RMSE error allows us to rank the performance of multiple models in regression analysis.

**R2:** The R2 coefficient compares the performance of model on test set and predicts the average values. This coefficient values always falls between [0,1) and the higher the value indicate better model performance.

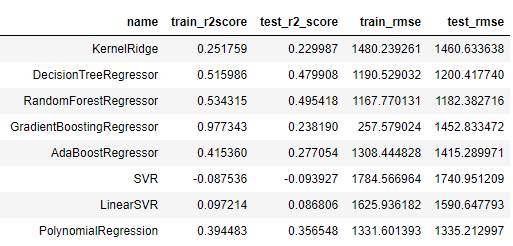
## **Models Used:**

Below are the models used for the training data. We calculate the training score and test score for all models.

The Regression models and their hyperparameters used for this data are:



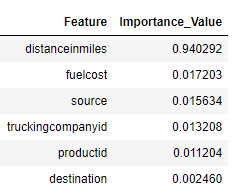
**Results for the Regression Models:**



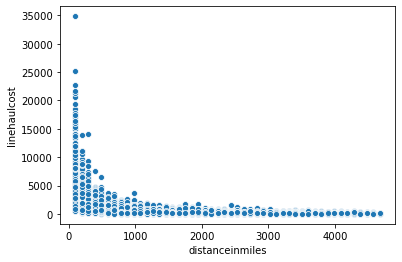
We observe that RandomForestRegressor has the lowest RMSE score and highest R2 on the test data. Based on this parameters we consider RandomForestRegressor as the winning model

## **Features Importance**

The random forest regressor gives the features importance. Below show the feature values used to train the model and their corresponding importance value.



From the feature importance value we observe that “distanceinmiles” is given the highest priority than the other features. When see the visualization of “distanceinmiles” and “linehaulcost” , it indicates an exponential relation between them.



Foe low distancein miles the frequency of linehaulcost is high and viceversa.

## **Residual Analysis:**

The residual analysis are plotted with the fitting data and residual. A residual plot with not having normal distribution is not a good fit data for Linear Model, since we are using non-linear regression we residual plot may not be a normal distribution.

