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

	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
distanceinmiles	10286.0	0.0	48	97.32	681.2400	1616.187544	1459.800	2433.0000	4671.36	1109.212772
fuelcost	10286.0	0.0	574	2.15	4.8500	5.499473	5.500	6.1600	9.55	0.987558
linehaulcost	10286.0	0.0	9766	7.49	150.4025	830.179412	326.135	771.7375	34845.53	1696.541911

b) Categorical Features

	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
truckingcompanyid	10286	0	6153	5367	8	0.0777756	4835	7.0	0.0680537
productid	10286	0	24	4	464	4.51099	20	461.0	4.48182
source	10286	0	49	Phoenix	242	2.35271	Miami	237.0	2.3041
destination	10286	0	49	New Orleans	262	2.54715	Las Vegas	249.0	2.42077

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).

	Orig_Mean	Mean_Err_500	Mean_Err_1000	Mean_Err_2500	Mean_Err_5000	Mean_Err_7000
distanceinmiles	1616.187544	11.335328	-2.143492	-1.870347	-2.616467	-0.027013
fuelcost	5.499473	-0.001492	-0.000039	-0.003028	0.002601	0.000413
linehaulcost	830.179412	7.603301	-2.543913	4.431431	6.632291	-2.631441

```
df_std_err
```

	Orig_StdDev	Std_Dev_Err500	Std_Dev_Err1000	Std_Dev_Err2500	Std_Dev_Err5000	Std_Dev_Err7000
distanceinmiles	1109.212772	-2.049465	1.155032	4.859623	1.159413	0.579944
fuelcost	0.987558	0.002618	0.005343	-0.002945	-0.001094	-0.000971
linehaulcost	1696.541911	-39.499132	4.634438	23.443668	16.097787	-1.076301

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.

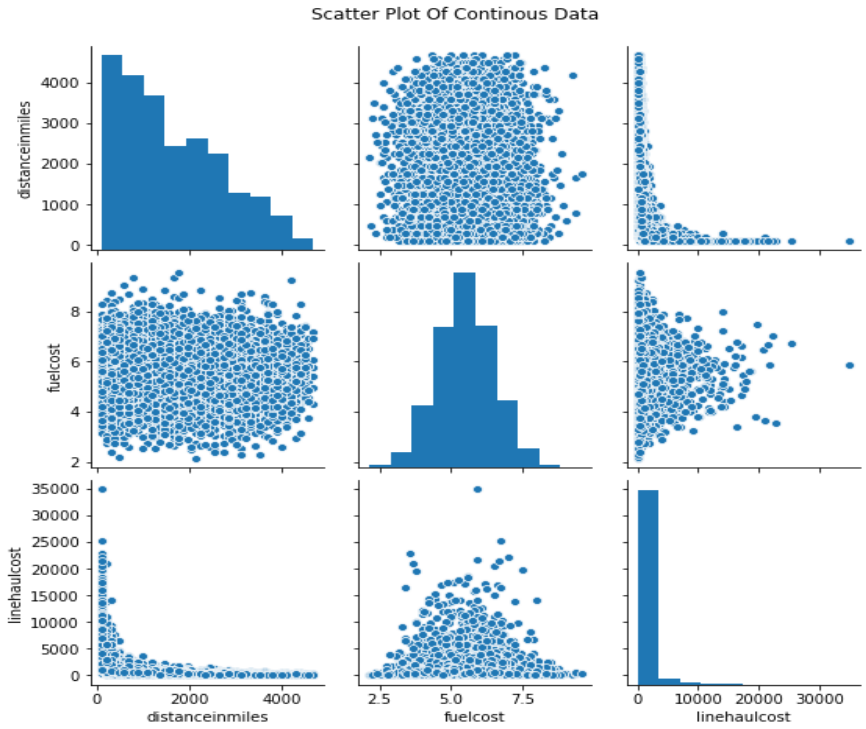
	KS-Stat	P-Value
distanceinmiles	0.004917	0.999955
fuelcost	0.006367	0.995548
linehaulcost	0.005450	0.999619

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.

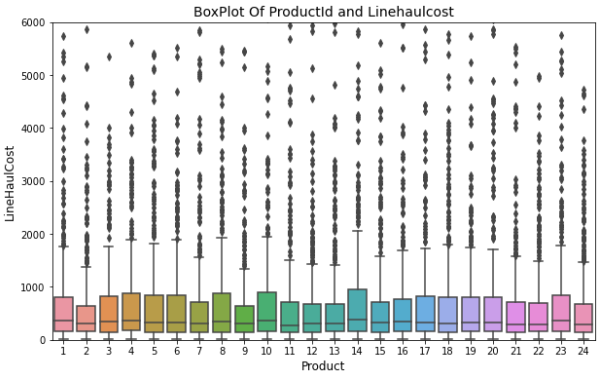
Relationship between features

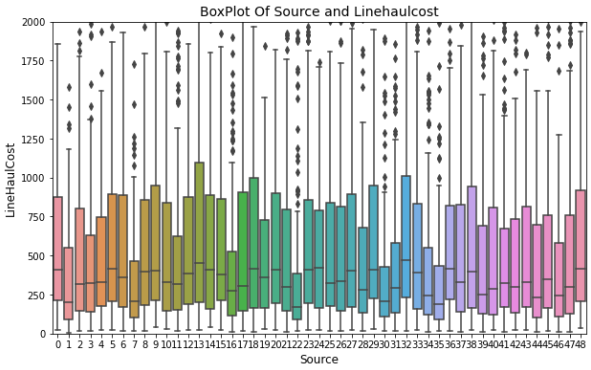
Visualizations are the best way to investigate the relationship between pairs of features. This helps us to know which feature might be useful to predict the target feature and help find the descriptive features that are closely related.

Visualizing Pair of Continous Data:



Visualization of Categorical Data and Continous Data:





Observations:

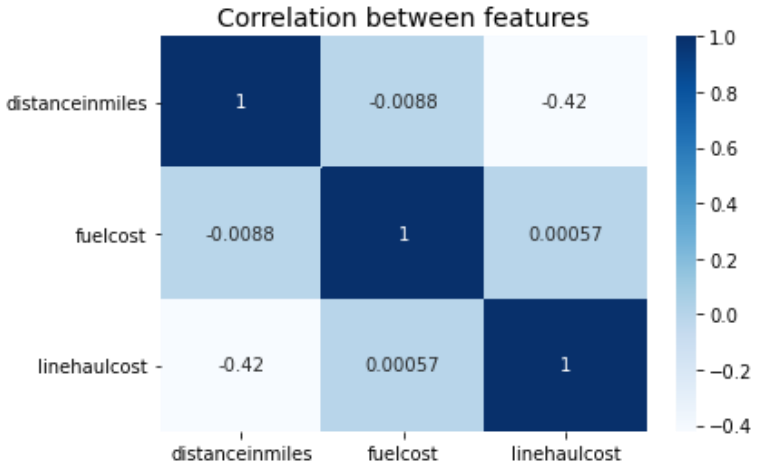
From the pair plot between “distanceinmiles”, ”fuelcost”, ”linehaulcost” the diagonal plots show the distribution of each feature and the lower matrix shows the relationship between them.

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.

VIF	variable
47.375467	Intercept
1.000196	truckingcompanyid
1.000321	productid
1.006481	distanceinmiles
1.000265	fuelcost
1.002854	source
1.004062	destination

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"

truckingcompanyid	productid	distanceinmiles	fuelcost	linehaulcost	source	destination	lhcpermile	fuelpermile	tot_cost
1699	2	97.32	5.88	34845.53	San Antonio	San Diego	358.051069	0.060419	358.111488
6665	20	97.32	7.02	22268.04	Chicago	Houston	228.812577	0.072133	228.884710
5287	8	97.32	6.49	20600.74	Albuquerque	Tucson	211.680436	0.066687	211.747123
8645	8	97.32	3.80	19459.44	Tampa	Tulsa	199.953144	0.039046	199.992191
9108	7	97.32	5.59	18611.15	Fort Worth	Columbus	191.236642	0.057439	191.294081
6832	12	97.32	4.89	17517.92	Albuquerque	Tucson	180.003288	0.050247	180.053535
6857	3	97.32	5.09	17357.32	Boston	El Paso	178.353062	0.052302	178.405364
2205	8	97.32	3.42	16400.68	Oklahoma City	Detroit	168.523222	0.035142	168.558364
3849	7	97.32	5.77	15684.42	Tulsa	Tampa	161.163379	0.059289	161.222667
1598	19	97.32	4.21	14285.38	Charlotte	Columbus	146.787711	0.043259	146.830970

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.

source	destination	agg
Charlotte	Las Vegas	16
Detroit	Fresno	12
Oklahoma City	Albuquerque	12
San Antonio	Indianapolis	12
Phoenix	Minneapolis	12

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

truckingcompanyid	productid	distanceinmiles	fuelcost	linehaulcost	source	destination	lhcpermile	fuelpermile	tot_cost
1001	21	1167.84	5.67	226.86	Charlotte	Las Vegas	0.194256	0.004855	0.199111
1328	19	1167.84	5.98	73.55	Charlotte	Las Vegas	0.062980	0.005121	0.068100
1755	14	1167.84	5.65	292.61	Charlotte	Las Vegas	0.250557	0.004838	0.255395
2160	15	1167.84	6.27	425.63	Charlotte	Las Vegas	0.364459	0.005369	0.369828
2726	12	1167.84	4.99	160.42	Charlotte	Las Vegas	0.137365	0.004273	0.141638
4016	4	1167.84	5.28	93.46	Charlotte	Las Vegas	0.080028	0.004521	0.084549
4417	20	1167.84	6.42	432.82	Charlotte	Las Vegas	0.370616	0.005497	0.376113
4551	11	1167.84	6.67	403.63	Charlotte	Las Vegas	0.345621	0.005711	0.351332
4928	9	1167.84	6.71	1042.41	Charlotte	Las Vegas	0.892597	0.005746	0.898342
5367	22	1167.84	6.68	330.03	Charlotte	Las Vegas	0.282599	0.005720	0.288319
5505	19	1167.84	6.55	619.88	Charlotte	Las Vegas	0.530792	0.005609	0.536401
6111	4	1167.84	6.71	137.91	Charlotte	Las Vegas	0.118090	0.005746	0.123835
7550	11	1167.84	5.28	315.28	Charlotte	Las Vegas	0.269968	0.004521	0.274490
7561	22	1167.84	5.97	743.56	Charlotte	Las Vegas	0.636697	0.005112	0.641809
9839	13	1167.84	4.18	322.02	Charlotte	Las Vegas	0.275740	0.003579	0.279319
9888	21	1167.84	6.44	873.48	Charlotte	Las Vegas	0.747945	0.005514	0.753459

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.

R^2: The R^2 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:

```
Regressors = [
    KernelRidge(kernel='rbf'),
    DecisionTreeRegressor(max_depth=3),
    RandomForestRegressor(n_estimators=100,max_depth=3),
    GradientBoostingRegressor(n_estimators=100,max_depth=3),
    AdaBoostRegressor(n_estimators=100),
    SVR(kernel='poly',gamma='auto'),
    LinearSVR(),
    Pipeline([('poly', PolynomialFeatures(degree=3)),('linear', LinearRegression(fit_intercept=False))])
    ]
```

Results for the Regression Models:

name	train_r2score	test_r2_score	train_rmse	test_rmse
KernelRidge	0.251759	0.229987	1480.239261	1460.633638
DecisionTreeRegressor	0.515986	0.479908	1190.529032	1200.417740
RandomForestRegressor	0.534315	0.495418	1167.770131	1182.382716
GradientBoostingRegressor	0.977343	0.238190	257.579024	1452.833472
AdaBoostRegressor	0.415360	0.277054	1308.444828	1415.289971
SVR	-0.087536	-0.093927	1784.566964	1740.951209
LinearSVR	0.097214	0.086806	1625.936182	1590.647793
PolynomialRegression	0.394483	0.356548	1331.601393	1335.212997

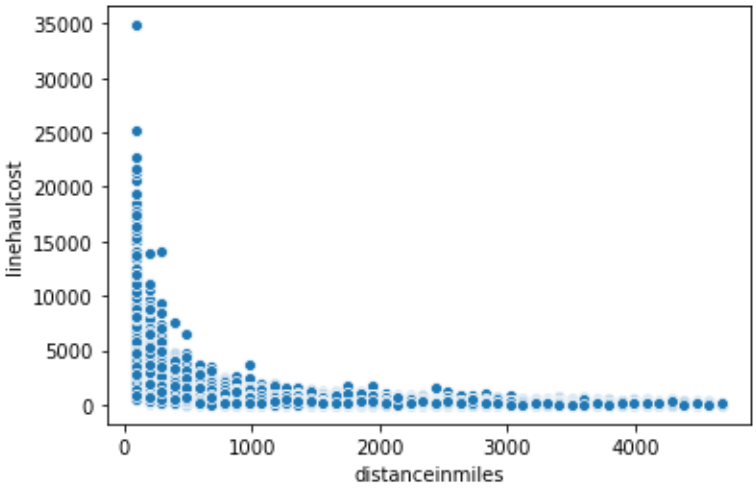
We observe that RandomForestRegressor has the lowest RMSE score and highest R^2 on the test data. Based on this parameters we consider RandomForestRegressor as the winning model

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.

Feature	Importance_Value
distanceinmiles	0.940292
fuelcost	0.017203
source	0.015634
truckingcompanyid	0.013208
productid	0.011204
destination	0.002460

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.