Shipping Cost Analysis

By :
VVS Kushwanth Reddy
Data Scientist (7 Years)
BTech Chemical Eng, Adv Mgmt in Business Analytics
kushwanth777@gmail.com
9619045522

Case Study

Problem Statement

- EASTMAN Company wants to understand the shipping cost that they are paying to shipping companies to transport their chemicals, fibers and Plastics from their warehouses to existing and new customers

Business Objective

- Obtain a model that will estimate shipping cost (column "Cost") for new selections of input variables. Assume we could change route class and priority if desired. The data provided includes all possible shipment origins but not all future shipment destinations as we regularly get new customers.
- Determine if there are any outliers in the historical data suggesting a higher-than-expected shipping cost that would warrant a follow-up with transportation suppliers regarding their rates.

Exploratory Data Analysis

- Entire data features can be broadly divided into below 3 subgroups.
  - Origin information.
  - Destination Information.
  - Navigation Information.
- Dataset has total of 306 rows and 8 columns
- There are No missing values in the dataset
- Y Variable(Dependent) = Cost
- X Variables(Independent) = 2 Numeric and 5 Categorical Variables

```
df.head()
```

	Origin City	Origin State	Destination City	Destination State	Route Class	Priority	Distance	Cost
0	ANNISTON	AL	BAYPORT	TX	Red	1	808	3105.12
1	ANNISTON	AL	HOUSTON	TX	Green	1	779	3357.28
2	MAGNESS	AR	KINGSPORT	TN	Green	2	657	2670.23
3	MAGNESS	AR	BAYPORT	TX	Red	1	610	2715.14
4	VERNON	CA	HARPERTOWN	CA	Red	1	235	2241.49

```
df.shape
```

```
(306, 8)
```

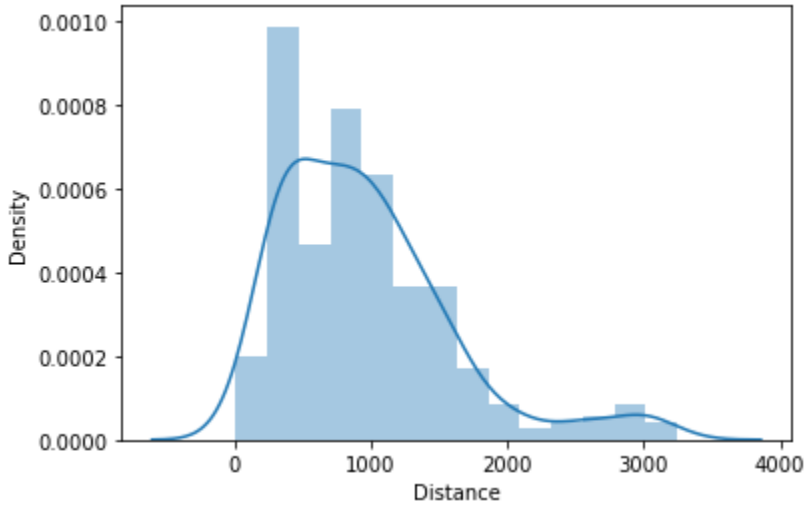
```
df.dtypes
```

```
Origin City           object
Origin State          object
Destination City      object
Destination State     object
Route Class           object
Priority               int64
Distance               int64
Cost                 float64
dtype: object
```

Numerical Feature Analysis

There are 2 Numerical Variables: 1.Distance and 2.Priority

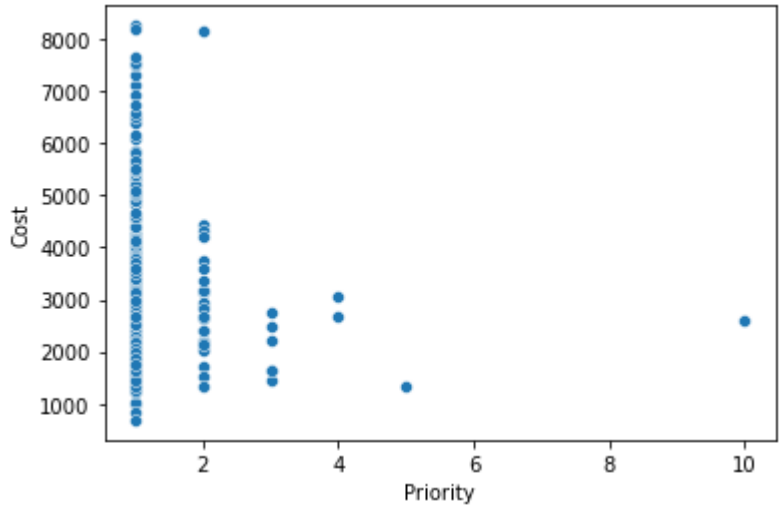
- **Distance** feature has Gaussian distribution
- No need of doing any transformations like (Log, Square root etc)



- **Priority** feature has Discrete Values(1,2,3..) and not continuous variable
- Let's Convert this from Integer data type to Categorical(Object) data type
- Most of the records have Priority-1.there is not much variance in the Variable

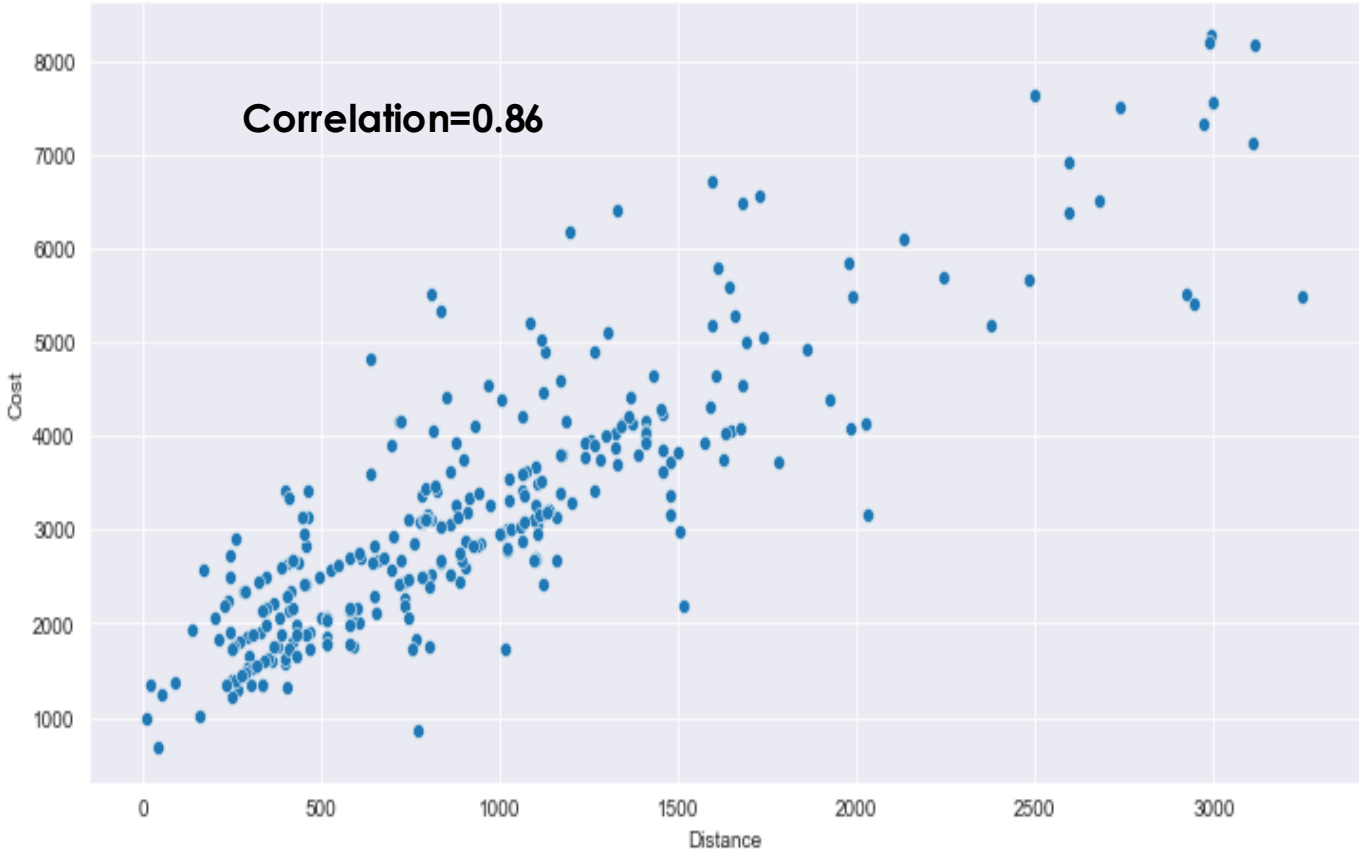
```
df['Priority'].value_counts()
```

```
1     267
2      29
3       6
4       2
10      1
5       1
Name: Priority, dtype: int64
```



Correlation

- Correlation is a statistical term describing the degree to which two variables has linear relationship.
- Distance feature is highly Positively correlated(0.86) with Cost Variable
- Distance vs Cost Scatter plot clearly shows Linear relationship
  - Outliers may exist in both Distance and Cost features
  - Will apply Inter Quartile Range (IQR) to identify Outliers



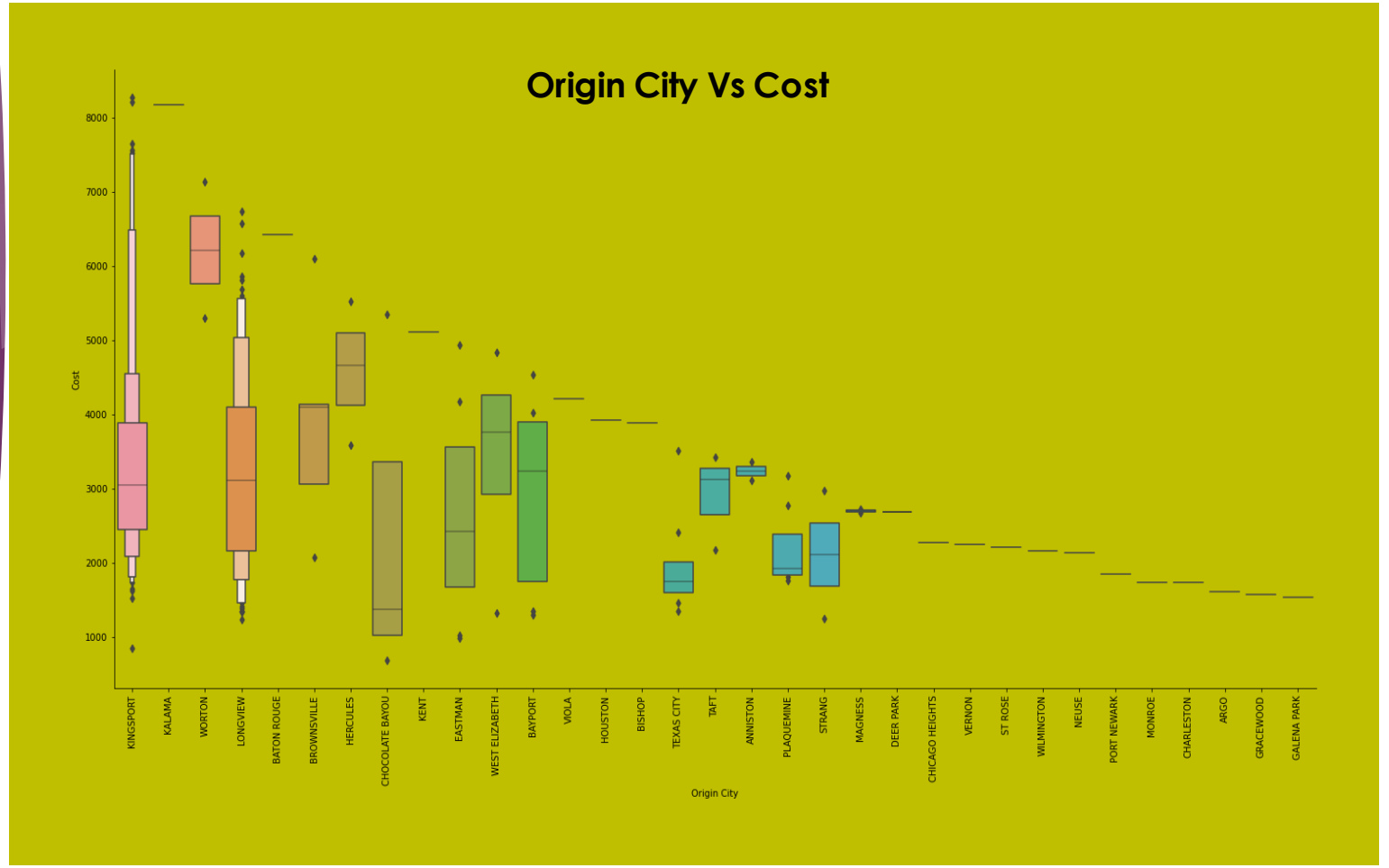
Categorical Features Analysis

- There are huge number of Levels in all categorical values especially in Destination City (due to multiple customers.
- **Priority** doesn't have much spread in the data
- **Origin Cities** could be EASTMAN company Warehouses

  Few Origin cities like

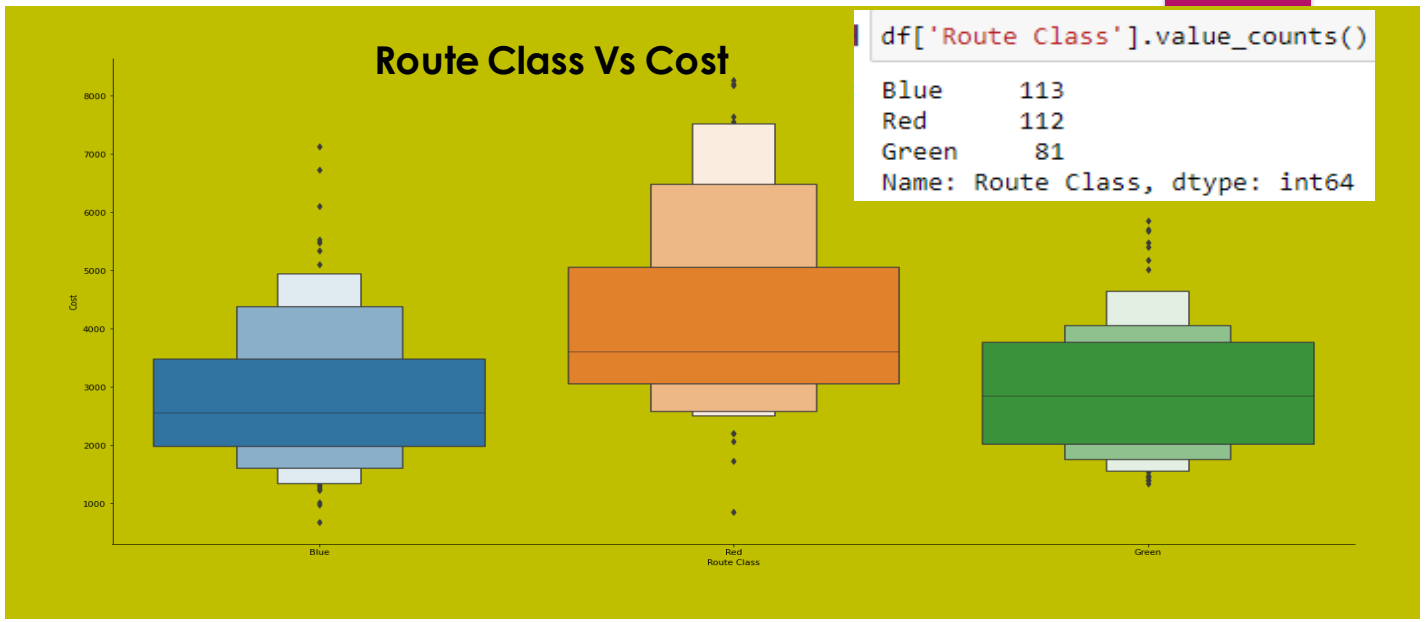
  - Kingsport
  - Longview have capability to send to multiple customers with Huge Cost Range

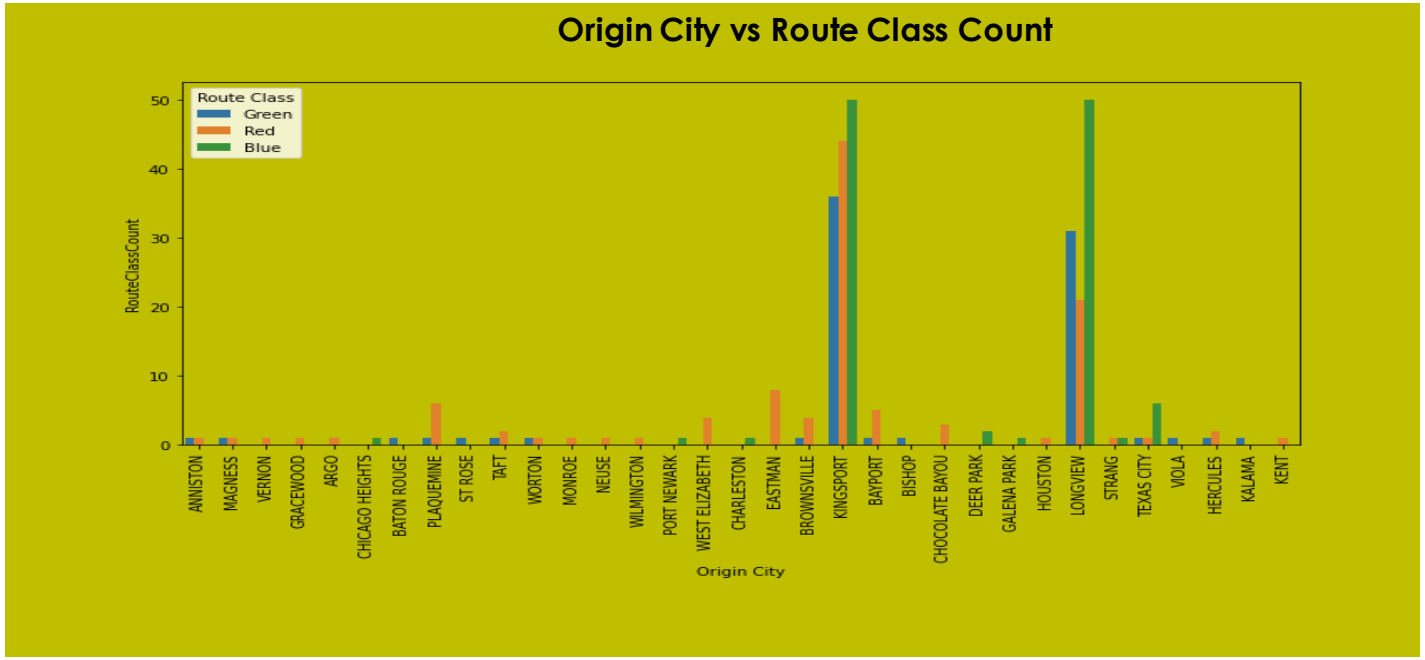
```
Origin City : 33 labels
Origin State : 15 labels
Destination City : 186 labels
Destination State : 39 labels
Route Class : 3 labels
Priority : 6 labels
```



Categorical Features Analysis

- **Route Class** has 3 levels where Blue, Red is almost equal and Green is relatively less.
- Is **Route Class** ordinal categorical variable?
  - Assuming it is not for this use case
- Origin City V S Route class count aggregation plot doesn't give much information
- All Categorical features are converted to One hot encoding features before fitting on any model





Outliers -Distance and Cost

- Applying Inter Quartile Range(IQR) on Distance and Cost Column to identify Records which is more than Upper and Lower range from dataset

```
Distance_outlierremoved_data_df=outlierremoved_data.loc[outlierremoved_data['Distance']>2484]
Distance_outlierremoved_data_df
```

	Origin City	Origin State	Destination City	Destination State	Route Class	Priority	Distance	Cost
21	WORTON	MD	VERNON	CA	Red	1	3112	7126.23
58	KINGSPORT	TN	CITY OF INDUSTRY	CA	Green	1	2973	7330.43
61	KINGSPORT	TN	VERNON	CA	Green	1	2594	6927.58
81	KINGSPORT	TN	SPARKS	NV	Green	1	2736	7518.08
86	KINGSPORT	TN	KENT	WA	Green	1	2922	5518.36
97	KINGSPORT	TN	ANAHEIM	CA	Green	1	2596	6388.32
108	KINGSPORT	TN	TOLENAS	CA	Green	1	2997	7557.31
117	KINGSPORT	TN	SANTA FE SPRINGS	CA	Green	1	2680	6510.20
129	KINGSPORT	TN	CARSON	CA	Green	1	2499	7641.99
155	KINGSPORT	TN	NORTH SAN JOSE	CA	Green	1	2990	8265.17
172	KINGSPORT	TN	SAN JOSE	CA	Green	1	2986	8207.39
243	LONGVIEW	TX	KALAMA	WA	Blue	1	2942	5404.06
275	LONGVIEW	TX	NEW WESTMINSTER	BC	Blue	1	3247	5481.73
304	KALAMA	WA	WORTON	MD	Green	2	3117	8172.43

```
cost_distance_outlierremoved_data_df=outlierremoved_data_cost.loc[outlierremoved_data_cost['Cost']>5158]
cost_distance_outlierremoved_data_df=cost_distance_outlierremoved_data_df.append(outlierremoved_data_cos
cost_distance_outlierremoved_data_df
```

	Origin City	Origin State	Destination City	Destination State	Route Class	Priority	Distance	Cost
8	BATON ROUGE	LA	WEST ELIZABETH	PA	Green	1	1328	6414.26
20	WORTON	MD	GARLAND	TX	Green	1	1658	5294.44
39	BROWNSVILLE	TN	VERNON	CA	Red	1	2131	6098.90
88	KINGSPORT	TN	LOS ANGELES	CA	Green	1	1681	6482.53
120	KINGSPORT	TN	EDMONTON	AB	Blue	1	2378	5173.27
144	KINGSPORT	TN	SALT LAKE CITY	UT	Blue	1	2243	5701.89
170	CHOCOLATE BAYOU	TX	ANNISTON	AL	Red	1	831	5339.84
191	LONGVIEW	TX	INDIAN ORCHARD	MA	Green	1	1726	6570.01
200	LONGVIEW	TX	WASHINGTON	WV	Green	1	1196	6171.22
205	LONGVIEW	TX	SEIPLE	PA	Green	1	1592	5195.46
210	LONGVIEW	TX	WORTON	MD	Green	1	1640	5593.37
223	LONGVIEW	TX	EAST BILLINGS	MT	Blue	1	1976	5853.20
225	LONGVIEW	TX	PINEVILLE	NC	Green	1	1083	5207.65
226	LONGVIEW	TX	SUCCASUNNA	NJ	Green	1	1610	5802.56
234	LONGVIEW	TX	CARSON	CA	Red	1	1596	6728.45
235	LONGVIEW	TX	CITY OF COMMERCE	CA	Red	1	1986	5477.66
253	LONGVIEW	TX	EDMONTON	AB	Blue	1	2483	5677.66
289	HERCULES	VA	ALERT	FL	Red	1	806	5525.39
150	KINGSPORT	TN	PLEASANT PRAIRIE	WI	Green	1	771	850.10
171	CHOCOLATE BAYOU	TX	HOUSTON	TX	Red	1	38	683.62

Objective 2 Findings

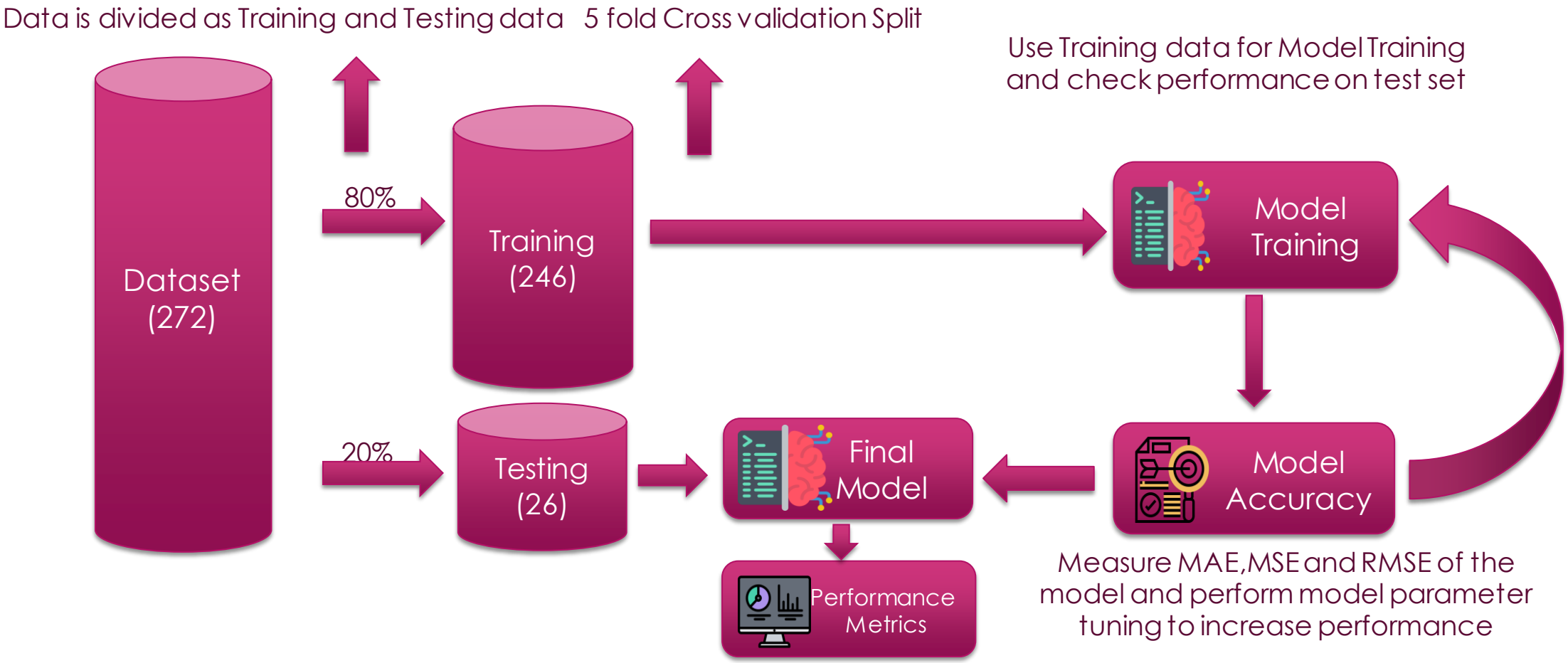
- Did Feature engineering by creating new feature 'cost_distance_ratio' by Cost/Distance
- Assuming if 'cost_distance_ratio' >7 is higher-than-expected shipping cost .
- These Routes and transactions need a follow-up with transportation suppliers regarding their rates.

```
In [452]: df.sort_values(by=['cost_distance_ratio']).tail(25)
```

Out[452]:

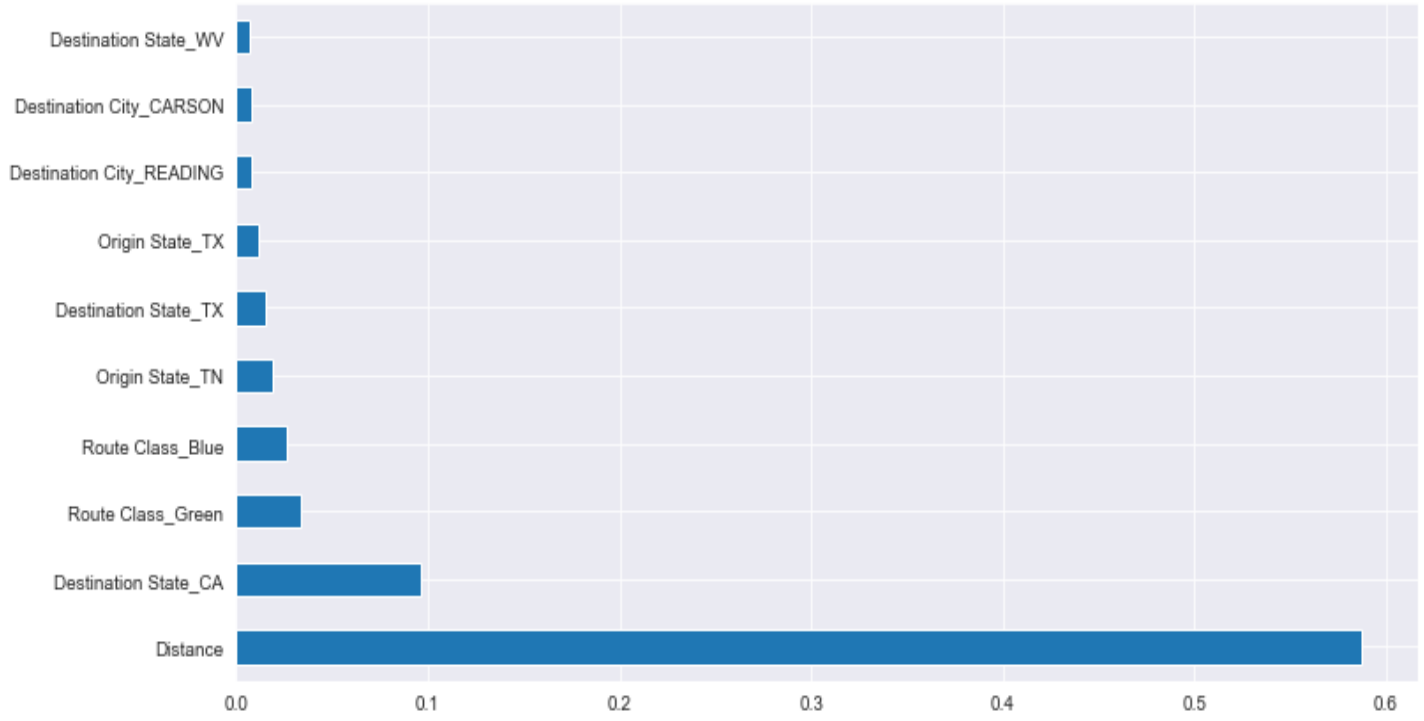
	Origin City	Origin State	Destination City	Destination State	Route Class	Priority	Distance	Cost	cost_distance_ratio
22	MONROE	NC	KINGSPORT	TN	Red	1	247	1737.27	7.033482
230	LONGVIEW	TX	BROWNSVILLE	TN	Green	1	443	3143.91	7.096862
46	KINGSPORT	TN	ASHEBORO	NC	Red	3	343	2497.14	7.280292
94	KINGSPORT	TN	RICHMOND	IN	Red	1	463	3418.12	7.382549
29	WEST ELIZABETH	PA	DANVILLE	VA	Red	1	635	4828.37	7.603732
148	KINGSPORT	TN	JAMESTOWN	NC	Red	1	320	2440.50	7.626563
74	KINGSPORT	TN	RICHMOND	KY	Red	1	244	1912.77	7.839221
103	KINGSPORT	TN	LELAND	NC	Red	1	409	3330.80	8.143765
126	KINGSPORT	TN	CHARLOTTE	NC	Red	1	285	2351.72	8.251649
104	KINGSPORT	TN	MOCKSVILLE	NC	Red	1	280	2338.76	8.352714
213	LONGVIEW	TX	MELENDY	TX	Red	1	212	1823.61	8.601934
75	KINGSPORT	TN	LAKE CITY	SC	Red	1	396	3415.71	8.625530
4	VERNON	CA	HARPERTOWN	CA	Red	1	235	2241.49	9.538255
31	EASTMAN	SC	BROWN SUMMIT	NC	Red	2	227	2196.19	9.674846
105	KINGSPORT	TN	DODDVILLE	SC	Red	1	200	2062.11	10.310550
76	KINGSPORT	TN	FOUNTAIN INN	SC	Red	1	241	2490.34	10.333361
66	KINGSPORT	TN	APPLE GROVE	WV	Red	1	259	2919.13	11.270772
240	LONGVIEW	TX	ODESSA	TX	Red	1	240	2721.40	11.339167
257	LONGVIEW	TX	LANCASTER	TX	Red	1	137	1947.72	14.216934
222	LONGVIEW	TX	TALLA BENA	LA	Green	1	166	2566.58	15.461325
183	CHOCOLATE BAYOU	TX	BAYPORT	TX	Red	1	86	1365.75	15.880814
182	CHOCOLATE BAYOU	TX	HOUSTON	TX	Red	1	38	683.62	17.990000
290	STRANG	TX	TEXAS CITY	TX	Red	1	50	1251.20	25.024000
294	TEXAS CITY	TX	BAYPORT	TX	Red	1	20	1349.62	67.481000
35	EASTMAN	SC	GASTON	SC	Red	1	9	982.18	109.131111

Objective 1- ML Model Design



Feature Importance

- Building a statistical model on 245 features doesn't give desired results and make any model complex. So Will have to find important features
- After applying Random Forest Regressor to find important features
- In top 10 important features , Distance variable clearly takes the major stake



Model-1&2 Linear Regression

- Applying standard scaler on 'Distance'
- Using all 278 independent variables to fit the model.

```
----------------------R-sqaure-----------------------------------------
R-sq for test data is 0.4902855509588536
R-sq for train data is 0.980576540806551
----------------------STANDARD ERROR/RMSE-------------------------------------
RMSE for test data is 773.6142589432704
RMSE for train data is 208.1660479567804
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Cost   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     15.12
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           4.14e-22
Time:                        22:21:10   Log-Likelihood:                -1648.8
No. Observations:                 244   AIC:                             3674.
Df Residuals:                      56   BIC:                             4331.
Df Model:                         187
Covariance Type:            nonrobust
```

- Using top 10 independent features instead of 278 features to fit the model.

```
----------------------R-sqaure-----------------------------------------
R-sq for test data is 0.4655939662834113
R-sq for train data is 0.8341608724593319
----------------------STANDARD ERROR/RMSE-------------------------------------
RMSE for test data is 705.5954755419804
RMSE for train data is 608.2611946539292
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Cost   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     169.6
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           2.52e-88
Time:                        22:24:46   Log-Likelihood:                -1910.4
No. Observations:                 244   AIC:                             3837.
Df Residuals:                     236   BIC:                             3865.
Df Model:                           7
Covariance Type:            nonrobust
```

Model-3&4 Linear Regression

- Using 6 independent variables to fit the model.

```
----------------------R-sqaure-----------------------------------------
R-sq for test data is 0.4403157709926716
R-sq for train data is 0.8314817096572966
----------------------STANDARD ERROR/RMSE-------------------------------------
RMSE for test data is 730.6686146478393
RMSE for train data is 613.1547974950796
--------
MAE: 474.2980164256009
MSE: 533876.6244313926
RMSE: 730.6686146478393
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Cost   R-squared:                       0.831
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     294.8
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           3.85e-91
Time:                        22:46:01   Log-Likelihood:                -1912.4
No. Observations:                 244   AIC:                             3835.
Df Residuals:                     239   BIC:                             3852.
Df Model:                           4
Covariance Type:            nonrobust
```

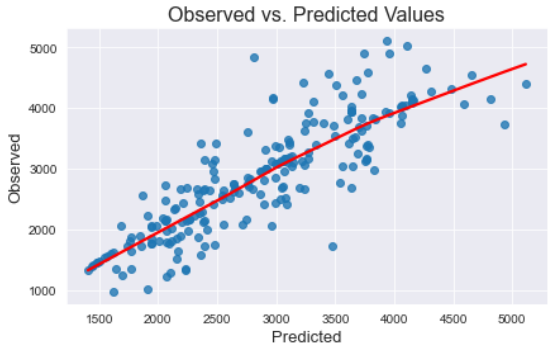
- Using top 9 independent features by dropping most important 'Distance' feature to fit the model.

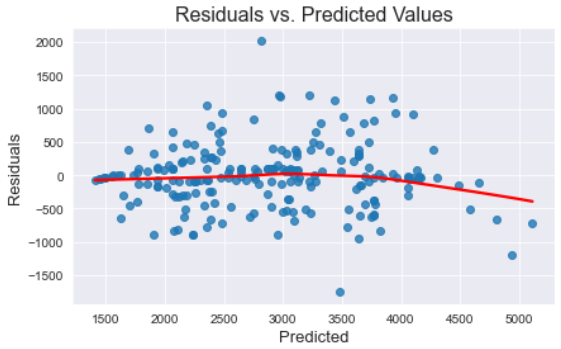
```
----------------------R-sqaure-----------------------------------------
R-sq for test data is -2.3555873823708953
R-sq for train data is 0.23206017927306954
----------------------STANDARD ERROR/RMSE-------------------------------------
RMSE for test data is 1167.6035716024203
RMSE for train data is 1308.9112560402239
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Cost   R-squared:                       0.232
Model:                            OLS   Adj. R-squared:                  0.213
Method:                 Least Squares   F-statistic:                     11.94
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           1.06e-11
Time:                        22:26:22   Log-Likelihood:                -2097.4
No. Observations:                 244   AIC:                             4209.
Df Residuals:                     237   BIC:                             4233.
Df Model:                           6
Covariance Type:            nonrobust
```

Linear Regression Model Evalution-with 6 imp Features

Observations on Multicollinearity== As VIF is <2..So no Multicollinearity in X

	feature	VIF
0	Distance	1.245033
1	Route Class_Green	1.231434
2	Route Class_Blue	1.893386
3	Destination State_TX	1.158457
4	Origin City_KINGSPORT	1.013222
5	Destination City_READING	1.022658
6	Destination City_DANVILLE	1.009695
7	Destination State_WV	1.014013
8	Route Class_Red	1.506096

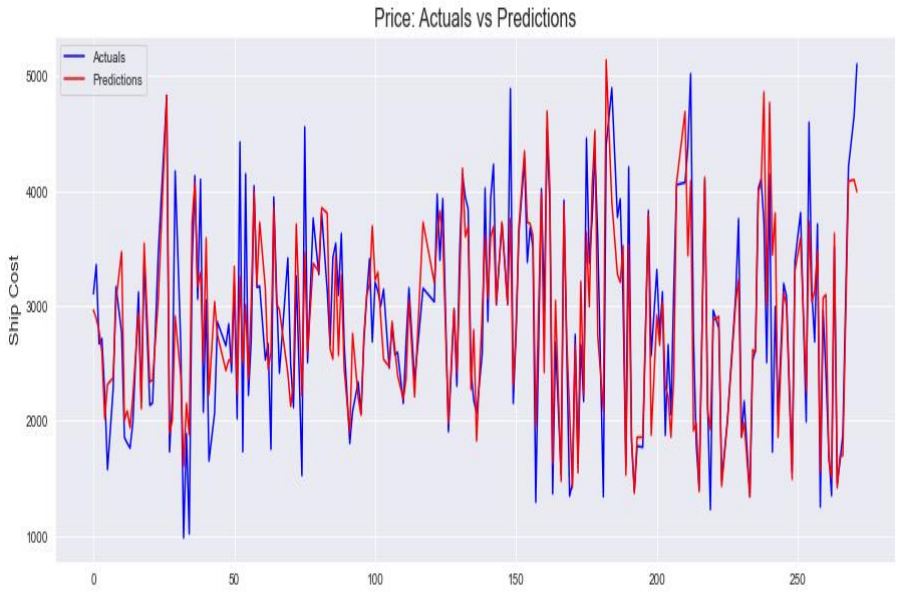




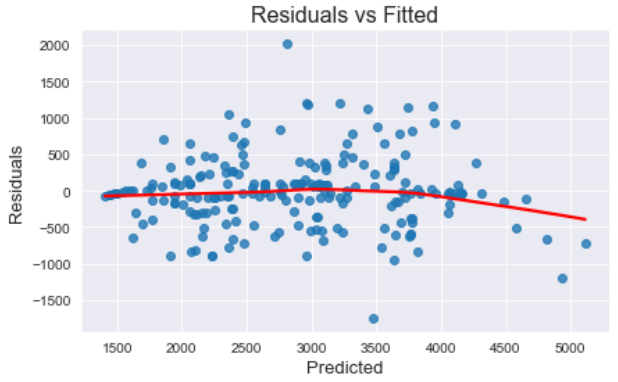
From above graphs :

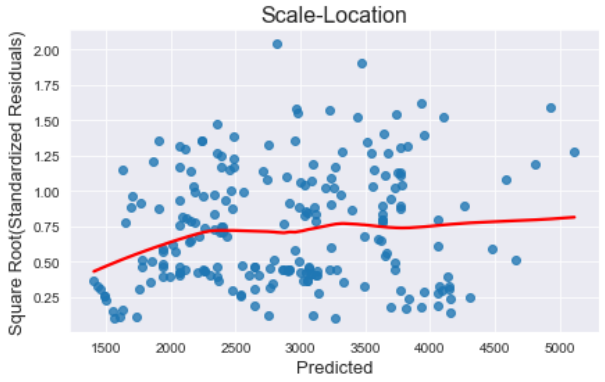
1. Obs vs predicted shows that most of the values are closer to the daigonal line, however some are not which is a problem

2. Resi vs pred graph does not give a conclusive evidence that residuals are evenly scattered around the zero line as Resi. values increase with increase in predicted values, SO ASSUMPTION OF LINEARITY CAN'T BE CONFIRMED.




```
homoscedasticity_test(lm_sm)
#both graphs show evenly spread residuals so homoscedasticity is present
```





Conclusion

Summary

- Linear Regression with top 4-5 important features is good to have for a simple model
- Random Forrest Algorithm with all independent features has done reasonably well
- Random Forrest Algorithm using all independent features with 10 Fold cross validation and Hyper parameter is better among the models built

Model	MAE	RMSE	R2
Random Forrest (CV, Hyper Parameter Tuning)	466	642	0.82
Random Forrest	464	700	0.61
Linear Regression	474	730	0.44

Follow up& Further Work

- Select top 5 frequent "Destination City" and group by rest of the cities as 1 to reduce no of classes from 186 to 6 and then apply Model to see if it helps
- If Possible, Collect more data and more numerical Features Periodically and retrain the model to identify new patterns and new customer segments.
- Dumped final model in pickle format which can be used to deploy
- Measure the Metrics periodically to monitor model performance.

Thank You!!!