Airline Cost prediction using Panel data Regression

Submitted by:

Soumya Paul (211391)

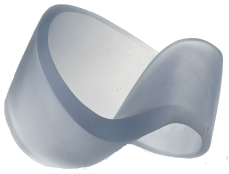
Tanu Denwal (201042)

Koustav Saha (200520)

Under the guidance of:

Sandipan Mitra and Rohan Kumar,

Mentors, Stamatics Project 2022



Stamatics, IIT Kanpur

Abstract

A person who purchases flight tickets most often would easily be able to predict and tell the right time to purchase a ticket to get the best deal. Most airlines change the prices of tickets for the management of their revenue. When the demand for flight tickets is expected to increase the airline may increase the price of flight tickets. It depends on many other factors too such as Fuel price, Load factor, and the average capacity utilization of the fleet. Here we will be using the Panel regression method for the prediction of the price of airlines in the future time also the variation of price in different places.

Acknowledgement

It is our pleasure to present a project on “Airline cost prediction busing Panel data Regression”. Every accomplishment has constant encouragement and advice from valuable and noble minds to guide us in putting our efforts in the right direction to bring out the project. We want to express our sincere gratitude to our mentors **Sandipan Mitra and Rohan Kumar** for their constant help and support throughout the completion of the project. Without their valuable guidance and motivation, it was nearly impossible to work on this project as a team and understand the practical aspect of the course “Regression Analysis”. Also we are thankful to all faculty members and seniors without whose support at various stages, this project would not have materialized. Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses.

Soumya Paul

Tanu Denwal

Koustav Saha

Contents

1 Introduction

Complex algorithms are used by Airline companies to calculate the prices of flights considering various situations present at that time. These methods take various social factors, and financial marketing into account to predict the prices of flights. As of now, the number of flights user has increased by a significant amount. It has become difficult for airlines to maintain constant prices because prices change dynamically due to different situations. This is the reason why we are trying to use machine learning to solve this kind of problem. This will help airlines in predicting what prices they can have. It will also be going to help customers to plan their journey accordingly and predict future flight prices. We are going to analyze the flight rate prediction using a Machine Learning dataset and necessary exploratory data analysis technique and predictions about the price of the flight based on factors such as what type of airline it is, what is fuel price, Load factor, the average capacity utilization of the fleet.

2 Objective

We want to predict the price of airlines at any time we want and for that prediction, we can face many problems such as variations in prices of fuel, demand for flight tickets, and load factors, passenger output. We will be using panel data analysis for the same. We can simply do the average of fuel price with respect to time or place for estimation but there is a problem with this, here the predicted variation of our data will not be reflected and also it will create a problem while we will test our model from the test dataset because we have removed all the variation of our dataset. With our panel regression model, we will predict the price of airlines in the future time also the variation of price in different places. We come up with different estimation methods like pooled OLS method of estimation, Fixed effect method and Random effect method.

3 Methodology

Time series data have data for different time periods and the difference between the time period is constant. In time-series data we have one particular stock over a period of time and the stock has different values at different times. Time-series data only observes one object recurrently over time. In Cross Section data we do not have a time component which means we have different data for a different stock for a single time period so in cross-section data we will have data for different cross-sections only for a single time period. cross-sectional data is described as one observation of multiple objects and corresponding variables at a specific point in time. Panel Data is a kind of combination data, when we combine time series respect and the cross-section aspect of data we get panel data. Panel data comprises characteristics of both into one model by collecting data from multiple, same objects over time.

4 Dataset

For building up the prediction model, we need experimental data,. We got the dataset from https://www.kaggle.com/datasets/s

For the prediction of the cost, we choose six firms and the data of the fifteen-year timelapse. Our predictors are 'Q' Output, in revenue passenger miles, index number, 'PF' = Fuel price, 'LF' = Load factor, and the average capacity utilization of the fleet. And we will be predicting 'C' the total cost

4.1 Viewing Dataset in Python

	I	T	C	Q	PF	LF
0	1	1	1140640	0.952757	106650	0.534487
1	1	2	1215690	0.986757	110307	0.532328
2	1	3	1309570	1.091980	110574	0.547736
3	1	4	1511530	1.175780	121974	0.540846
4	1	5	1676730	1.160170	196606	0.591167
...	...	...	...	...	...	...
85	6	11	381478	0.112640	874818	0.517766
86	6	12	506969	0.154154	1013170	0.580049
87	6	13	633388	0.186461	930477	0.556024
88	6	14	804388	0.246847	851676	0.537791
89	6	15	1009500	0.304013	819476	0.525775

90 rows 6 columns

5 Exploratory data analysis

EDA is an important step to build any machine learning project. Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

5.1 Dimension of Data

We have 90 on and 6 variables.

5.2 Basic Information about data

```
I        int64
T        int64
C        int64
Q      float64
PF       int64
LF     float64
dtype: float64
```

There are 4 variables which are of integer type and 2 variables which is of float type.

5.2.1 Checking if there is any missing value There is no missing value in the data

```
I     0
T     0
C     0
Q     0
```

```
PF     0
LF     0
dtype: int64
```

So, there is no missing values present in our dataset.

5.3 Reformatting our dataset

To give our dataset a compact view and creating dummy variable for I variable for future use, we have performed this reformatting.

		C	Q	PF	LF	I
I	T					
1	1	1140640	0.952757	106650	0.534487	1
	2	1215690	0.986757	110307	0.532328	1
	3	1309570	1.091980	110574	0.547736	1
	4	1511530	1.175780	121974	0.540846	1
	5	1676730	1.160170	196606	0.591167	1
...	...	...	... ...	...	...	
6	11	381478	0.112640	874818	0.517766	6
	12	506969	0.154154	1013170	0.580049	6
	13	633388	0.186461	930477	0.556024	6
	14	804388	0.246847	851676	0.537791	6
	15	1009500	0.304013	819476	0.525775	6

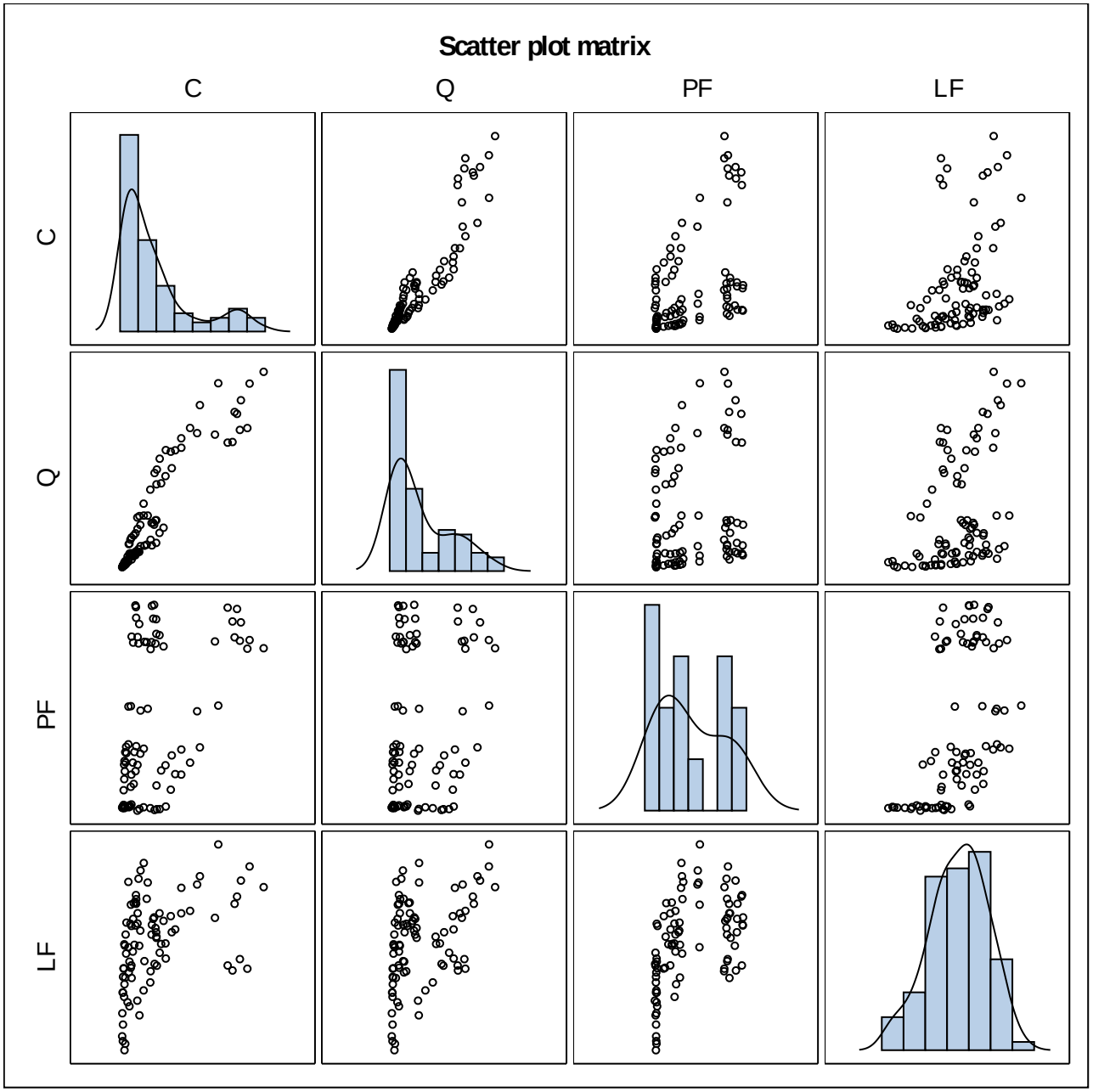
5.4 Data description

Here the number of observations, mean, standard deviation, minimum value, 1st quartile, 2nd quartile (median), 3rd quartile and maximum value is given as count, mean, std, min, '25%', '50%', '75%' and max respectively for each of the variables.

	C	Q	PF	LF
count	9.000000e+01	90.000000	9.000000e+01	90.000000
mean	1.122524e+06	0.544995	4.716830e+05	0.560460
std	1.192075e+06	0.533586	3.295029e+05	0.052793
min	6.897800e+04	0.037682	1.037950e+05	0.432066
25%	2.920460e+05	0.142128	1.298475e+05	0.528806
50%	6.370010e+05	0.305028	3.574335e+05	0.566085
75%	1.345968e+06	0.945278	8.498398e+05	0.594658
max	4.748320e+06	1.936460	1.015610e+06	0.676287

5.5 Scatter plot matrix

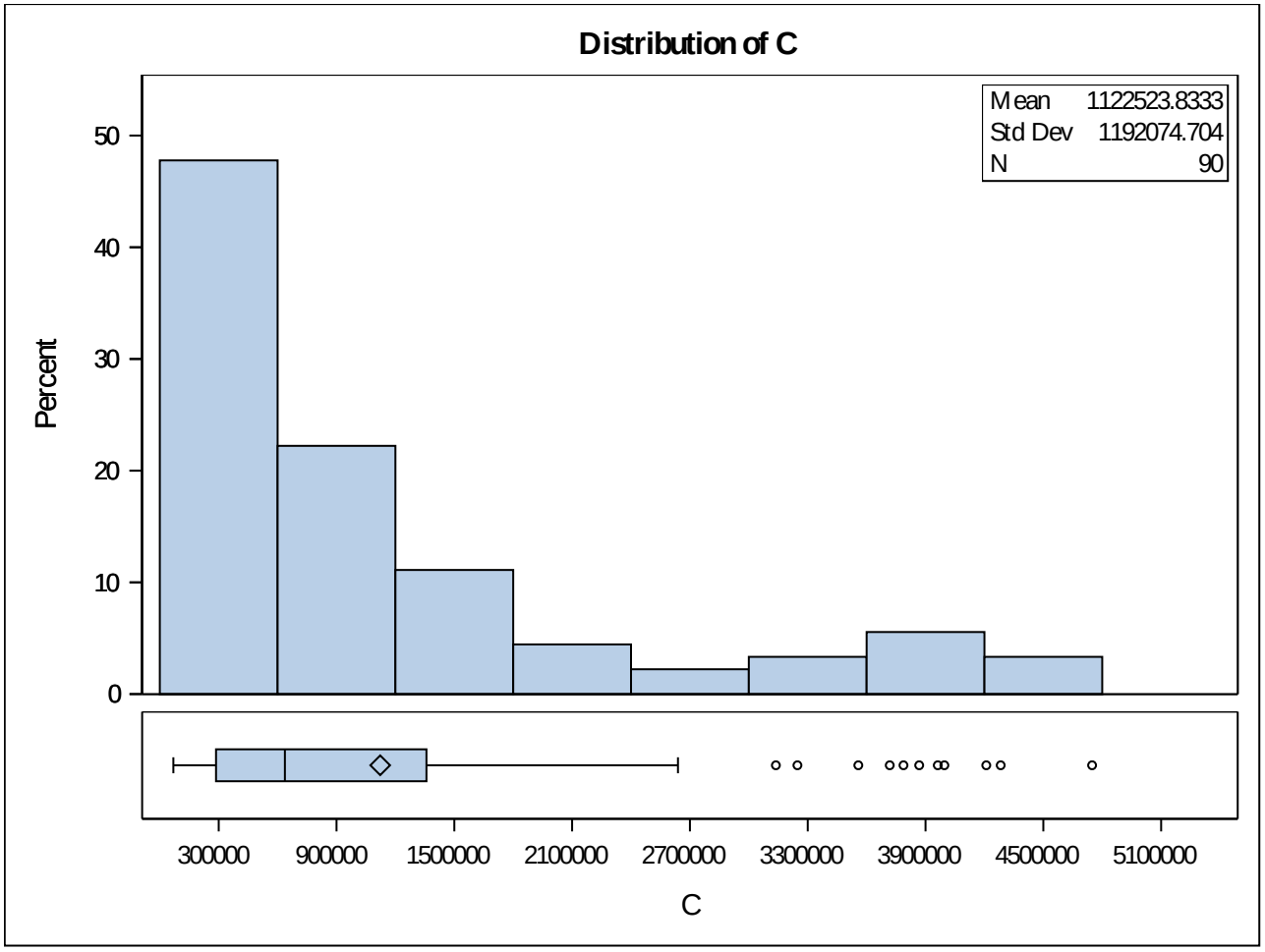
A grid of several scatter plots of up to five numeric variables is a scatter plot matrix. The matrix involves individual scatter plots for each and every combination of variables.



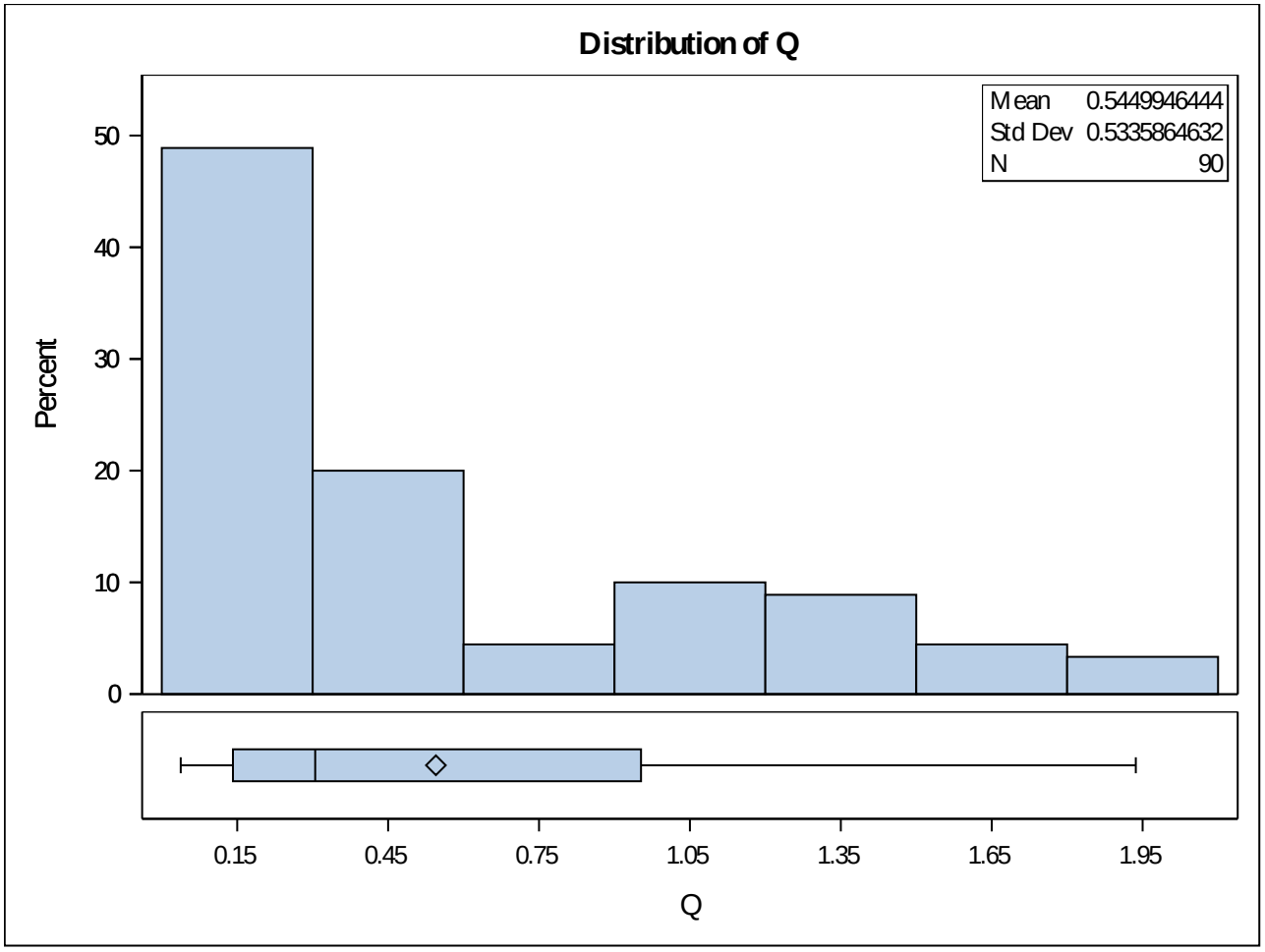
We can observe from our SAS output that there is a strong linear relationship between C and Q. Distribution of C and Q are almost positively skewed and distribution of LF looks like Gaussian.

5.6 Box Plot with Histogram

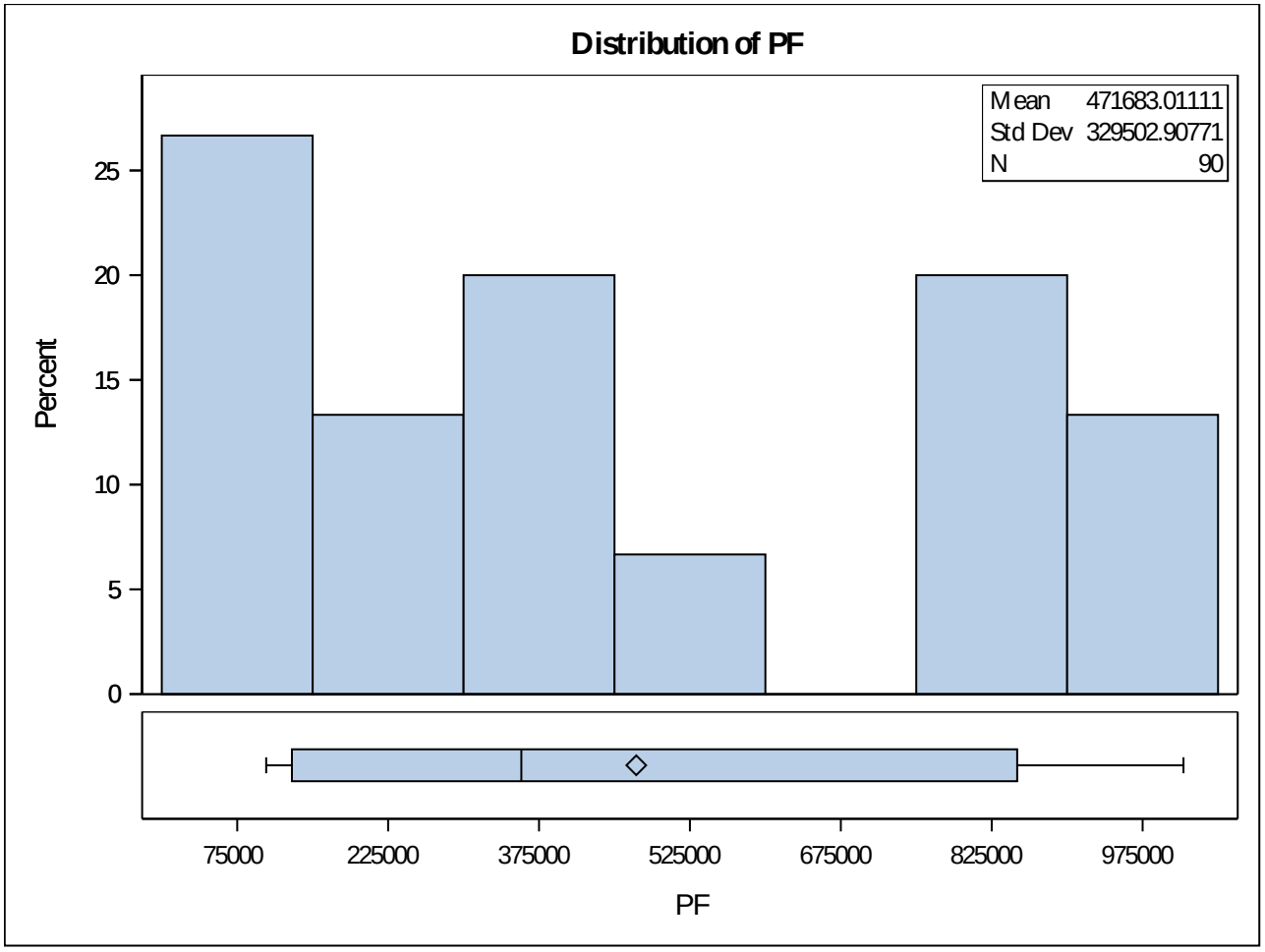
It is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. **Box plots** are also very beneficial when large numbers of data sets are involved or compared.



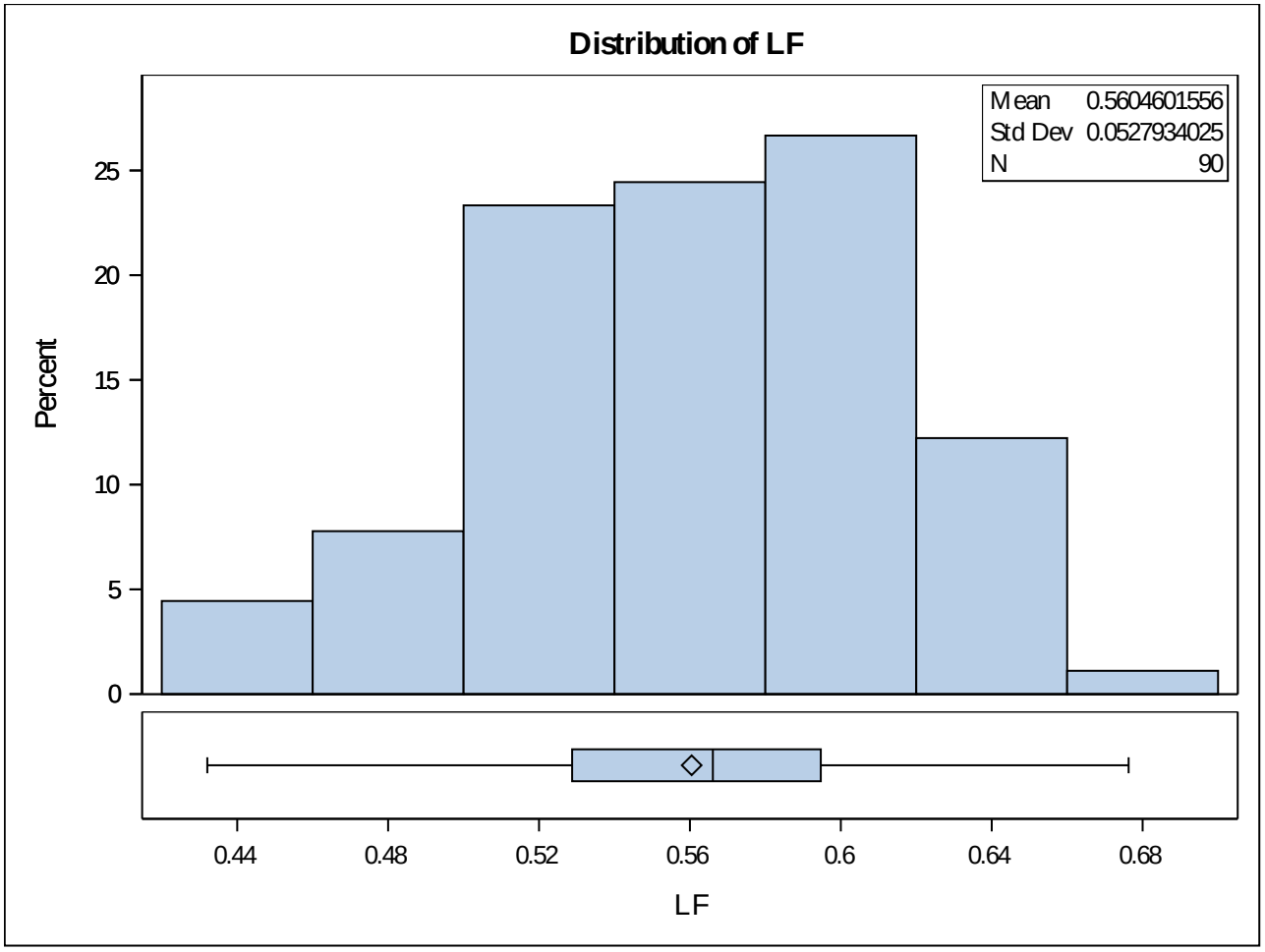
There are some outliers and distribution is positively skewed as median is nearer to lower hinge.



There is no outlier and distribution is positively skewed as median is nearer to lower hinge.



There is no outlier and distribution is slightly positively skewed as median is nearer to lower hinge.



There is no outlier and distribution is looks like gaussian.

6 Panel data regression analysis

Panel data which is also referred to as longitudinal data contains observations about different cross sections across time.It is a collection of quantities obtained across multiple individuals, that are assembled over even intervals in time and ordered chronologically.Like time series panel data contains observations collected at a regular interval.Examples of groups that may make up panel data series include countries, firms, individuals, or demographic groups.

Advantages of panel data regression:-

- Panel data contains more information, more variability, and more efficiency than pure time series data or cross-sectional data.
- Panel data can model both the common and individual behaviors of groups.
- Panel data can detect and measure statistical effects that pure time series or cross-sectional data can't.
- Panel data can minimize estimation biases that may arise from aggregating groups into a single time series.

Types of panel data:-

1. **Balanced panel**-Equal Number of observation for each panel member for entire period.
2. **Short panel -**Number of panels members are more than number of period.
3. **Long panel-**Number of period are more than number of panel members.
4. **Dynamic panel**-The dynamic panel data regression model is characterized by two sources of persistence over time: the presence of a lagged dependent variable as a regressor and cross section-specific unobserved heterogeneity.
5. **Unbalanced panel-**Number of observations are not same for all panel members.

6.1 Panel data regression model-

$$Y_{pt} = \beta_0 + \beta_1 X_{1,pt} + \beta_2 X_{2,pt} + \beta_3 X_{3,pt} + ... + \beta_n X_{n,pt} + \epsilon_{pt}$$

which can be rewritten as

$$Y_{pt} = \beta_0 + \beta_1 X_{1,pt} + \beta_2 X_{2,pt} + \beta_3 X_{3,pt} + ... + \beta_n X_{n,pt} + \mu_t + \omega_p + \eta_{pt}$$

(Here p =airline , t =year n = number of explanatory variables , μ_t=time dependent error term , ω_p=place dependent error term ,$X_{i,pt}$=predictor variable , Y_{pt}=Response variable, β_i=Coefficient of the ith explanatory variable and η_{pt} =idiosyncratic error)

6.2 Estimation methods used in our model-

There are several estimation method used to predict our model-

1. Pooled OLS method
2. Fixed effect method
3. Random effect method

6.3 Pooled OLS method of estimation-

A panel model is the most simple and accurate as it combines both time series and cross sectional data.In this model we do not consider time and individual dimension, so it is assumed that the behavior of corporate data is the same in various period.. This method can use the Ordinary Least Square (OLS) approach or the least squares technique to estimate the panel data model.

Assumption-

- Regression coefficients are same for all places.
- Regressors are non random.($Cov(X_{pt},\eta_{pt})$=0)

- η_{pt} follows $\mathcal{N}(0,,\sigma^2)$

PooledOLS Estimation Summary

Dep. Variable:	C	R-squared:	0.9461
Estimator:	PooledOLS	R-squared (Between):	0.9914
No. Observations:	90	R-squared (Within):	0.8786
Date:	Sun, Jun 12 2022	R-squared (Overall):	0.9461
Time:	22:29:52	Log-likelihood	-1255.0
Cov. Estimator:	Clustered		
		F-statistic:	503.12
Entities:	6	P-value	0.0000
Avg Obs:	15.000	Distribution:	F(3,86)
Min Obs:	15.000		
Max Obs:	15.000	F-statistic (robust):	2144.2
		P-value	0.0000
Time periods:	15	Distribution:	F(3,86)
Avg Obs:	6.0000		
Min Obs:	6.0000		
Max Obs:	6.0000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	1.159e+06	4.179e+05	2.7725	0.0068	3.279e+05	1.989e+06
Q	2.026e+06	8.134e+04	24.910	0.0000	1.864e+06	2.188e+06
PF	1.2253	0.3568	3.4345	0.0009	0.5161	1.9346
LF	-3.066e+06	1.04e+06	-2.9466	0.0041	-5.134e+06	-9.974e+05

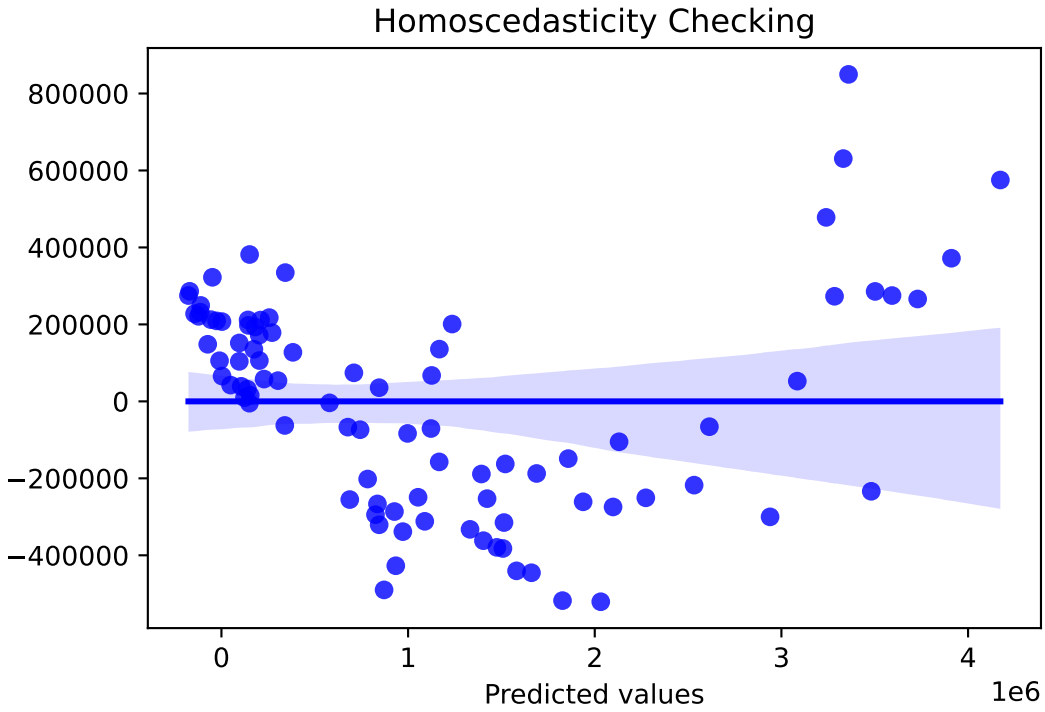
After performing this method of estimation we can see that R squared value is 0.9461.So we can say that our model can explain 94.16% of actual data accurately.

Now several problems arise when we perform this model. First problem is we have to determine whether the error is homoscedastic or heteroscedastic. Second problem is whether serial correlation is present in our data or not.

Homoscedasticity

Homoscedasticity, or homogeneity of variances, is an assumption of equal or similar variances in different groups being compared.Now in our model after performing the pooled OLS method of estimation , we get a plot between predicted value and residuals.

From the above plot we can see that with the increase in predicted value the variation of residual is increasing.So we can conclude our model is heteroscedastic.



Serial correlation check

```
# Durbin-Watson-Test
>>>from statsmodels.stats.stattools import durbin_watson
>>>pooled_OLS_dataset = pd.concat([dataset, pooledOLS_res.resids], axis=1)
>>>pooled_OLS_dataset = pooled_OLS_dataset.drop(['I'], axis = 1).fillna(0)
>>>durbin_watson_test_results = durbin_watson(pooled_OLS_dataset['residual'])
>>>print(durbin_watson_test_results)
0.4341624384518438
```

From the above python code we have got our DW statistics value is 0.4341624384518438.

Now there are three cases-

1. If DW statistics value is less than 2 and greater than 0 then positive correlation is present in our model.
2. If DW statistics value is equal to 2 then no correlation is present in our model.
3. If DW statistics value is greater than 2 then negative correlation is present in our model.

So we can conclude as our DW statistics value is in the range of 0 to 2 so positive serial correlation is present in our model.

To get rid of this problem we need to perform random effect method and fixed effect method.

6.4 Fixed effect method

6.4.1 LSDV approach

In statistics, a fixed effects model is a statistical model in which the model parameters are fixed or non-random quantities.So we will focus firstly on LSDV approach by introducing dummy variables and performing pooled OLS.

Equation-

$$Y_{pt} = \beta_0 + \beta_{1,pt} + ...\beta_{n,pt} + \omega_p + \eta_{pt}.....(1)$$

$$Y_{pt} = \beta_0 + \beta_{1,pt} + ...\beta_{n,pt} + \alpha_1 D_{1,p}..... + \alpha_{p-1} D_{P-1,p} + \eta_{pt}.....(2)$$

, where P is the total number of airlines and

$$D_{ip} = \begin{cases} 1 & \text{for } i\text{th airline} \\ 0 & \text{otherwise} \end{cases}$$

This is called one way FE model as intercept only vary across places but not across time.

PooledOLS Estimation Summary

Dep. Variable:	C	R-squared:	0.9716
Estimator:	PooledOLS	R-squared (Between):	1.0000
No. Observations:	90	R-squared (Within):	0.9294
Date:	Sat, Jun 11 2022	R-squared (Overall):	0.9716
Time:	01:00:54	Log-likelihood	-1226.1
Cov. Estimator:	Clustered		
		F-statistic:	346.92
Entities:	6	P-value	0.0000
Avg Obs:	15.000	Distribution:	F(8,81)
Min Obs:	15.000		
Max Obs:	15.000	F-statistic (robust):	-2.469e+16
		P-value	1.0000
Time periods:	15	Distribution:	F(8,81)
Avg Obs:	6.0000		
Min Obs:	6.0000		
Max Obs:	6.0000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-1.312e+05	8.545e+05	-0.1536	0.8783	-1.831e+06	1.569e+06
Q	3.319e+06	1.668e+05	19.898	0.0000	2.987e+06	3.651e+06
PF	0.7731	0.2797	2.7638	0.0071	0.2165	1.3296
LF	-3.797e+06	1.704e+06	-2.2291	0.0286	-7.187e+06	-4.079e+05
I.2	6.017e+05	8.93e+04	6.7382	0.0000	4.241e+05	7.794e+05
I.3	1.337e+06	1.624e+05	8.2352	0.0000	1.014e+06	1.66e+06
I.4	1.778e+06	1.908e+05	9.3147	0.0000	1.398e+06	2.157e+06
I.5	1.828e+06	2.071e+05	8.8285	0.0000	1.416e+06	2.24e+06
I.6	1.706e+06	2.155e+05	7.9180	0.0000	1.278e+06	2.135e+06

Here after performing this estimation model we can see that our R squared value is 0.9716.So we can conclude that our model can explain 97% of actual data accurately.

6.4.2 Group wise demeaning approach

Model

$$Y_{pt} = \beta_0 + \beta_1 X_{1,pt} + \beta_2 X_{2,pt} + \beta_3 X_{3,pt} + \ldots\ldots\ldots + \beta_n X_{n,pt} + \omega_p + \eta_{pt} \tag{1}$$

then

$$\overline{Y}_p = \beta_0 + \beta_1 \overline{X}_{1,p} + \beta_2 \overline{X}_{2,p} + \beta_3 \overline{X}_{3,p} + \ldots\ldots\ldots + \beta_n \overline{X}_{n,p} + \omega_p + \overline{\eta}_p \tag{2}$$

Performing (1)-(2) we gets

$$\ddot{Y}_{pt} = Y_{pt} - \bar{Y}_p = \beta_1 \ddot{X}_{1,pt} + \beta_2 \ddot{X}_{2,pt} + \beta_3 \ddot{X}_{3,pt} + \ldots\ldots\ldots + \beta_n \ddot{X}_{n,pt} + \ddot{\eta}_{pt} \tag{3}$$

which is equivalent to

$$Y_{pt} = \beta_{0p} + \beta_1 X_{1,pt} + \beta_2 X_{2,pt} + \beta_3 X_{3,pt} + \ldots\ldots\ldots + \beta_n X_{n,pt} + \ddot{\eta}_{pt}$$

where $\beta_{op} = \bar{Y}_p - \beta_1 \overline{X}_{1,p} + \beta_2 \overline{X}_{2,p} + \beta_3 \overline{X}_{3,pt} + \ldots\ldots\ldots + \beta_n \overline{X}_{n,p}$

PanelOLS Estimation Summary

Dep. Variable:	C	R-squared:	0.9294
Estimator:	PanelOLS	R-squared (Between):	0.4449
No. Observations:	90	R-squared (Within):	0.9294

Date:	Sun, Jun 12 2022	R-squared (Overall):	0.6394
Time:	23:01:52	Log-likelihood	-1226.1
Cov. Estimator:	Unadjusted		
		F-statistic:	355.25
Entities:	6	P-value	0.0000
Avg Obs:	15.000	Distribution:	F(3,81)
Min Obs:	15.000		
Max Obs:	15.000	F-statistic (robust):	355.25
		P-value	0.0000
Time periods:	15	Distribution:	F(3,81)
Avg Obs:	6.0000		
Min Obs:	6.0000		
Max Obs:	6.0000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	1.077e+06	3.108e+05	3.4662	0.0008	4.589e+05	1.696e+06
Q	3.319e+06	1.714e+05	19.369	0.0000	2.978e+06	3.66e+06
PF	0.7731	0.0973	7.9437	0.0000	0.5794	0.9667
LF	-3.797e+06	6.138e+05	-6.1869	0.0000	-5.019e+06	-2.576e+06

Here we can observe that our coefficients of the explanatory variables are same as LSDV approach but intercept will be the average of $\beta_{01}, \beta_{02}, ...\beta_{0p}$.

6.5 Random effect method

This method (DerSimonian 1986) incorporates heterogenerosity of all the places within the error term rather than specified as dummy variables while allowing for a common intercept(β_0)

Model-

$$Y_{pt} - \lambda\overline{Y_p} = \beta_0(1-\lambda) + \beta_1(X_{1,pt} - \lambda\overline{X_{1,p}}) + ... + \beta_n(X_{n,pt} - \lambda\overline{X_{n,p}}) + \eta_{pt} - \lambda\overline{\eta_p}$$

Where $\lambda = 1 - (\frac{\sigma_n^2}{\sigma_n^2 + T\sigma_\omega^2}) \in (0,1)$

(σ_n^2 →Variance of idiosyncratic error term)

(σ_ω^2 →Variance of place specific error term)

Note that

$\lambda = 0 \implies$ (Pooled OLS)

$\lambda = 1 \implies$ (fixed effect method)

Interpretation will be the same as Fixed effect (Group wise demeaning method)

RandomEffects Estimation Summary

Dep. Variable:	C	R-squared:	0.9113
Estimator:	RandomEffects	R-squared (Between):	0.9570
No. Observations:	90	R-squared (Within):	0.8974
Date:	Sun, Jun 12 2022	R-squared (Overall):	0.9331
Time:	22:29:52	Log-likelihood	-1248.3
Cov. Estimator:	Unadjusted		
		F-statistic:	294.50
Entities:	6	P-value	0.0000
Avg Obs:	15.000	Distribution:	F(3,86)
Min Obs:	15.000		
Max Obs:	15.000	F-statistic (robust):	294.50
		P-value	0.0000
Time periods:	15	Distribution:	F(3,86)
Avg Obs:	6.0000		
Min Obs:	6.0000		
Max Obs:	6.0000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	1.074e+06	3.775e+05	2.8461	0.0055	3.239e+05	1.825e+06
Q	2.289e+06	1.095e+05	20.902	0.0000	2.071e+06	2.506e+06
PF	1.1236	0.1034	10.862	0.0000	0.9180	1.3292
LF	-3.085e+06	7.257e+05	-4.2512	0.0001	-4.528e+06	-1.642e+06

Here from this python output we can see our R squared value is 0.9113 which is pretty high.So we can conclude that our model can explain 91% of actual data accurately.

6.6 Hausman test

Hausman test is helpful to select whether we have to go for FE or RE.

H_0 :The appropriate model is RE model and $Cov\left(X_{pt}, \omega_p\right) = 0$ vs H_1 :The appropriate model is FE model and $Cov\left(X_{pt}, \omega_p\right) \neq 0$

Test statistic

$$H = \left(\hat{\beta}^{RE} - \hat{\beta}^{FE}\right)' \left[Var\left(\hat{\beta}^{RE}\right) - Var\left(\hat{\beta}^{FE}\right)\right] \left(\hat{\beta}^{RE} - \hat{\beta}^{FE}\right) \sim \chi_n^2$$

If $P\left(H > observed\ H\right) < 0.05$. choose FE model. Otherwise, choose RE model.

```
chi-Squared: 60.869537603132564
degrees of freedom:4
p-Value:1.904399777228271e-12
```

As we can observe here our P value comes out to be much much less than 0.05, we can reject null hypothesis.

So we will consider fixed effect model as our final model which can explain around 97% or 93% of our actual data accurately depending on the method used.