37638 O-48953 2014/30/11 1010 37639 O-48960 2014-11-30 1031 37640 O-48972 2014-11-30 1031 37641 O-48977 2014-11-30 1000 37642 rows × 12 columns	408.3 Standard Class Consumer Office Supplies 314.22 Standard Class Consumer Technology 50.09 Standard Class Corporate Office Supplies 21.12 Same Day Consumer Office Supplies 33.57 First Class Consumer Office Supplies 20.72 Second Class Consumer Office Supplies 1.2 Standard Class Consumer Office Supplies Office Supplies Office Supplies Office Supplies	30 5% 45 3% 14 3% 14 5% 14 5% 30 3% 14 0%	No Yes No Yes Yes Yes Yes Yes	Yes 2011-02-17 Yes 2011-02-10 No 2014-12-31 No 2014-12-31 No 2014-12-16 No 2014-12-15 No 2014-12-30		
df['value'] = np.arange(len(df)) OrderID InvoiceDate CustomerID OrderID 0 O-100 2011-01-03 1024 1 O-101 2011-01-03 1024 2 O-102 2011-01-05 1006 3 O-104 2011-01-06 1023 4 O-105 2011-01-06 1009	ParVolume ShipMode Segment Category Payment 276.1 Same Day Corporate Office Supplies 35.88 Same Day Corporate Office Supplies 66.12 Second Class Consumer Office Supplies 408.3 Standard Class Consumer Office Supplies 314.22 Standard Class Consumer Technology	IntTerm Discount ExistingPoint 30 5% 30 5% 30 3% 45 3%	yes Yes No Yes	Yes 2011-02-18 0 No 2011-02-18 1 Yes 2011-02-06 2 Yes 2011-02-17 3 Yes 2011-02-10 4		
37637 O-48951 2014/30/11 1029 37638 O-48953 2014/30/11 1010 37639 O-48960 2014-11-30 1031 37640 O-48972 2014-11-30 1031 37641 O-48977 2014-11-30 1000 37642 rows × 13 columns	50.09 Standard Class Corporate Office Supplies 21.12 Same Day Consumer Office Supplies 33.57 First Class Consumer Office Supplies 20.72 Second Class Consumer Office Supplies 1.2 Standard Class Consumer Office Supplies	14 3% 14 5% 14 5% 30 3% 14 0%	No Yes Yes Yes Yes	No 2014-12-31 37637 No 2014-12-31 37638 No 2014-12-16 37639 No 2014-12-15 37640 No 2014-12-30 37641		
<pre>(37642, 13) df.isnull().sum() OrderID</pre>						
ExistingPurchaseOrder 0 FirstCustomerOrder 0 CashInDate 0 value 0 dtype: int64 df['OrderVolume'].value_counts() 12.96 47 25.92 39 32.4 32 19.44 29 15.55 29						
212.85 1 862.46 1 90.99 1 1007.74 1 421.28 1 Name: OrderVolume, Length: 16952, dty #Extracting column from dataframe df_test = df[["InvoiceDate","CashInDa df_test InvoiceDate CashInDate value						
0 2011-01-03 2011-02-18 0 1 2011-01-03 2011-02-18 1 2 2011-01-05 2011-02-06 2 3 2011-01-06 2011-02-17 3 4 2011-01-06 2011-02-10 4 37637 2014/30/11 2014-12-31 37638 2014/30/11 2014-12-31 37638						
37639 2014-11-30 2014-12-16 37639 37640 2014-11-30 2014-12-15 37640 37641 2014-11-30 2014-12-30 37641 37642 rows × 3 columns #Coverting column to numeric format df['OrderVolume'] = pd.to_numeric(df df.OrderVolume.dtypes dtype('float64')	'OrderVolume'], errors = 'coerce')					
<pre>fill = df['OrderVolume'].mean() #filling null values with mean df['OrderVolume'] = df['OrderVolume'] df.OrderVolume 0</pre>	.fillna(df['OrderVolume'].mean())					
37637 50.09 37638 21.12 37639 33.57 37640 20.72 37641 1.20 Name: OrderVolume, Length: 37642, dty df_test.dtypes InvoiceDate object CashInDate object dtype: object df.loc[df['Discount'] == '5%', 'Discount']						
<pre>df.loc[df['Discount'] == '3%', 'Discount'] df.loc[df['Discount'] == '0%', 'Discount'] df.Discount.value_counts() 5 18748 3 9440 0 9438 Name: Discount, dtype: int64 mode_dis = df['Discount'].mode()</pre>	nt'] = 3					
<pre>0 5 dtype: object mode_dis = 5 df['Discount'] = df['Discount'].fills df = df.drop(['OrderID'],axis=1) df</pre>	a(mode_dis)					
OrderID InvoiceDate CustomerID Ord 0 O-100 2011-01-03 1024 1 O-101 2011-01-03 1024 2 O-102 2011-01-05 1006 3 O-104 2011-01-06 1023 4 O-105 2011-01-06 1009 37637 O-48951 2014/30/11 1029	276.1 Same Day Corporate Office Supplies 35.88 Same Day Corporate Office Supplies 66.12 Second Class Consumer Office Supplies 408.3 Standard Class Consumer Office Supplies 314.22 Standard Class Consumer Technology 50.09 Standard Class Corporate Office Supplies	30 5% 30 5% 30 3% 45 3% 14 3%	Yes Yes Yes No Yes No	StomerOrder CashInDate value Yes 2011-02-18 0 No 2011-02-18 1 Yes 2011-02-06 2 Yes 2011-02-17 3 Yes 2011-02-10 4 No 2014-12-31 37637		
37638 O-48953 2014/30/11 1010 37639 O-48960 2014-11-30 1031 37640 O-48972 2014-11-30 1031 37641 O-48977 2014-11-30 1000 37642 rows × 13 columns #transforming dataframe to numeric df.FirstCustomerOrder = df.FirstCustomerOrder = df.ExistingPurchaseOrder = df.Existi	21.12 Same Day Consumer Office Supplies 33.57 First Class Consumer Office Supplies 20.72 Second Class Consumer Office Supplies 1.2 Standard Class Consumer Office Supplies merOrder.replace(['No', 'Yes'], [0, 1]) gPurchaseOrder.replace(['No', 'Yes'], [0, 1])	14 5% 14 5% 30 3% 14 0%	Yes Yes Yes Yes	No 2014-12-31 37638 No 2014-12-16 37639 No 2014-12-15 37640 No 2014-12-30 37641		
<pre>from sklearn.preprocessing import Lab #encoding few colmns for future needs label = LabelEncoder() df.ShipMode = label.fit_transform(df df.Segment = label.fit_transform(df[df.Category = label.fit_transform(df)</pre>	elEncoder 'ShipMode']) Segment'])	ExistingPurchaseOrder First	CustomerOrder Cashl	nDate value		
0 2011-01-03 1024 276.10 1 2011-01-03 1024 35.88 2 2011-01-05 1006 66.12 3 2011-01-06 1023 408.30 4 2011-01-06 1009 314.22 37637 2014/30/11 1029 50.09 37638 2014/30/11 1010 21.12	1 1 1 30 5 1 1 1 30 5 2 0 1 30 3 3 0 1 30 5 3 0 2 45 3 3 1 1 14 3 1 0 1 14 5	1 1 1 0 1 0		02-18 1 02-06 2 02-17 3		
37639 2014-11-30 1031 33.57 37640 2014-11-30 1031 20.72 37641 2014-11-30 1000 1.20 37642 rows × 12 columns	0 0 1 14 5 2 0 1 30 3 3 0 1 14 0	1 1 1	0 2014- 0 2014- 0 2014-	12-16 37639 12-15 37640 12-30 37641		
InvoiceDate CashInDate 0 2011-01-03 2011-02-18 1 2011-01-03 2011-02-18 2 2011-01-05 2011-02-06 3 2011-01-06 2011-02-17 4 2011-01-06 2011-02-10 37637 NaT 2014-12-31 37638 NaT 2014-12-31						
37639 2014-11-30 2014-12-16 37640 2014-11-30 2014-12-15 37641 2014-11-30 2014-12-30 37642 rows × 2 columns df_test.dtypes InvoiceDate datetime64[ns] CashInDate datetime64[ns]						
<pre>CashInDate datetime64[ns] dtype: object df_test['Difference'] = df_test['Casl df_test['Difference'] = df_test['Difference'] = df_test['Difference'] df_test InvoiceDate CashInDate Difference 0 2011-01-03 2011-02-18 46.0 1 2011-01-03 2011-02-18 46.0 2 2011-01-05 2011-02-06 32.0</pre>	<pre>InDate'].sub(df_test['InvoiceDate'], axis=0) erence']/np.timedelta64(1,'D')</pre>					
3 2011-01-06 2011-02-17 42.0 4 2011-01-06 2011-02-10 35.0 37637 NaT 2014-12-31 NaN 37638 NaT 2014-12-31 NaN 37639 2014-11-30 2014-12-16 16.0 37640 2014-11-30 2014-12-15 15.0 37641 2014-11-30 2014-12-30 30.0						
37642 rows × 3 columns df_test.dropna(axis=0 , inplace=True) df_test Difference 0 46.0 1 46.0 2 32.0						
3 42.0 4 35.0 37631 29.0 37632 22.0 37639 16.0 37640 15.0 37641 30.0						
37628 rows × 1 columns result = pd.concat([df, df_test], axidisplay(result) OrderID InvoiceDate CustomerID Ord 0	s=1, join='inner') erVolume ShipMode Segment Category PaymentTerm Di 276.1 1 1 1 30 35.88 1 1 1 30 66.12 2 0 1 30 408.3 3 0 1 30	Discount ExistingPurchaseOr 5% 5% 3% 5%	1 1 1	er CashInDate value Difference 1 2011-02-18 0 46.0 0 2011-02-18 1 46.0 1 2011-02-06 2 32.0 1 2011-02-17 3 42.0		
4 O-105 2011-01-06 1009 37631 O-48922 2014-11-30 1013 37632 O-48929 2014-11-30 1020 37639 O-48960 2014-11-30 1031 37640 O-48972 2014-11-30 1031 37641 O-48977 2014-11-30 1000	314.22 3 0 2 45 79.9 0 0 1 14 151.88 3 0 1 14 33.57 0 0 1 14 20.72 2 0 1 30 1.2 3 0 1 14	3% 5% 3% 5% 3% 0%	1 1 1 1	1 2011-02-10 4 35.0 0 2014-12-29 37631 29.0 0 2014-12-22 37632 22.0 0 2014-12-16 37639 16.0 0 2014-12-15 37640 15.0 0 2014-12-30 37641 30.0		
37628 rows × 14 columns data_final = result[["InvoiceDate","] data_final InvoiceDate value Difference 0 2011-01-03 0 46.0 1 2011-01-03 1 46.0 2 2011-01-05 2 32.0 3 2011-01-06 3 42.0	alue", "Difference"]]					
4 2011-01-06 4 35.0 37631 2014-11-30 37631 29.0 37632 2014-11-30 37632 22.0 37639 2014-11-30 37639 16.0 37640 2014-11-30 37640 15.0 37641 2014-11-30 37641 30.0 37628 rows × 3 columns						
<pre>data_final.isnull().sum() InvoiceDate</pre>	iceDate')					
InvoiceDate 2011-01-03 0 46.0 2011-01-03 1 46.0 2011-01-05 2 32.0 2011-01-06 3 42.0 2011-01-06 4 35.0 2014-11-30 37631 29.0						
2014-11-30 37632 22.0 2014-11-30 37639 16.0 2014-11-30 37640 15.0 2014-11-30 37641 30.0 37628 rows × 2 columns df_index['Difference1'] = df_index['Idf_index['Difference2'] = df_index['Idf_index]'Idf_index['Idf_index	<pre>ifference'].shift(-1) ifference'].shift(-2)</pre>					
value Difference Difference1 Difference1 2011-01-03 0 46.0 46.0 2011-01-03 1 46.0 32.0 2011-01-05 2 32.0 42.0 2011-01-06 3 42.0 35.0 2011-01-06 4 35.0 33.0 2014-11-30 37631 29.0 22.0	32.0 42.0 35.0 33.0 40.0 16.0					
2014-11-30 37632 22.0 16.0 2014-11-30 37639 16.0 15.0 2014-11-30 37640 15.0 30.0 2014-11-30 37641 30.0 NaN 37628 rows × 4 columns df_index.dropna df_index	15.0 30.0 NaN NaN					
value Difference Difference1 Difference1 2011-01-03 0 46.0 46.0 2011-01-03 1 46.0 32.0 2011-01-05 2 32.0 42.0 2011-01-06 3 42.0 35.0 2011-01-06 4 35.0 33.0 2014-11-30 37631 29.0 22.0	32.0 42.0 35.0 33.0 40.0 16.0					
2014-11-30 37632 22.0 16.0 2014-11-30 37639 16.0 15.0 2014-11-30 37640 15.0 30.0 2014-11-30 37641 30.0 NaN 37628 rows × 4 columns df_index.dropna	15.0 30.0 NaN NaN					
<pre><bound 2011-01-03<="" dataframe.dropna="" invoicedate="" method="" of="" td=""><td>value Difference Difference1 Diffe 46.0 32.0 32.0 32.0 32.0 33.0</td><td>rence2</td><td></td><td></td><td></td><td></td></bound></pre>	value Difference Difference1 Diffe 46.0 32.0 32.0 32.0 32.0 33.0	rence2				
<pre>[37628 rows x 4 columns]> df_index.isnull().sum() value</pre>						
<pre>import numpy as np x,y = df_index['value'],df_index['Dir x,y = np.array(x),np.array(y) x,y = x.reshape(-1,1), y.reshape(-1,1) print(x) [[0] [1] [2] [37639]</pre>	<pre>ference']</pre>					
[37640] [37641]] print(y) [[46.] [46.] [32.] [16.] [15.] [30.]] from sklearn.model_selection import in the sklear strain steet with a sklear strain steet in the sklear strain str						
<pre>from sklearn.linear_model import Line reg = LinearRegression() reg.fit(xtrain,ytrain) LinearRegression() reg.fit(xtrain,ytrain) LinearRegression() pred = reg.predict(xtest)</pre>	t_split(x,y,test_size = 20,random_state = 1) arRegression					
pred array([[35.68473249],						
[35.68643519], [35.68653034], [35.68653538], [35.68623538], [35.68422725], [35.68493402], [35.68394804], [35.68493783]]) reg.predict([[1], [7], [50]]) array([[35.68653518], [35.68653466], [35.68653094]])						
<pre>from sklearn.metrics import mean_squafrom math import sqrt rmse_lr = sqrt(mean_squared_error(pre) rmse_lr = ('Mean Squared Error for Li rmse_lr ('Mean Squared Error for Linear Regre</pre>	d,ytest)) near Regression Model is :', rmse_lr)					
-0.008000784154470919 num = 20 trange = pd.date_range('2014-12-1', trange DatetimeIndex(['2014-12-01', '2014-12	02', '2014-12-03', '2014-12-04', 06', '2014-12-07', '2014-12-08', 10', '2014-12-11', '2014-12-12', 14', '2014-12-15', '2014-12-16', 18', '2014-12-19', '2014-12-20'],					
Forecast = reg.predict(xtest) print(Forecast) [[35.68473249] [35.68572946] [35.68414179] [35.68440241] [35.68440241] [35.6849374] [35.68629835] [35.68497606] [35.68437811] [35.68585514] [35.68610607]						
[35.6846081] [35.68643519] [35.68653034] [35.68354142] [35.68623538] [35.68422725] [35.68493402] [35.68394804] [35.68493783]] Predict_df = pd.DataFrame(Forecast, : Predict_df.columns = ['forecast'] Predict_df	ndex=trange)					
forecast 2014-12-01 35.684732 2014-12-02 35.685729 2014-12-03 35.684142 2014-12-04 35.684402 2014-12-05 35.684104 2014-12-06 35.686298 2014-12-07 35.684976 2014-12-08 35.684378						
2014-12-08 35.684378 2014-12-09 35.685855 2014-12-10 35.684949 2014-12-11 35.686106 2014-12-12 35.684608 2014-12-13 35.686435 2014-12-14 35.686530 2014-12-15 35.683541 2014-12-16 35.686235						
2014-12-17 35.684227 2014-12-18 35.684934 2014-12-19 35.683948 2014-12-20 35.684938 from sklearn import ensemble from sklearn.model_selection import	<pre>rain_test_split t_split(x,y,test_size = 20,random_state = 1)</pre>					
<pre>from sklearn.linear_model import Logs model = LogisticRegression() model.fit(xtrain,ytrain) C:\Users\shahidul Alam\anaconda3\envs for example using ravel(). return f(**kwargs) C:\Users\shahidul Alam\anaconda3\envs STOP: TOTAL NO. of ITERATIONS REACHED Increase the number of iterations (ma</pre>	sticRegression Data_Science\lib\site-packages\sklearn\utils\v Data_Science\lib\site-packages\sklearn\linear_ LIMIT. C_iter) or scale the data as shown in:					nge the shape of y to (n_sa
https://scikit-learn.org/stable/m Please also refer to the documentatio	odules/preprocessing.html for alternative solver options: odules/linear_model.html#logistic-regression 36., 36., 36., 36., 36., 36., 36.,					
<pre>from sklearn.ensemble import RandomForestClassifier() Rclf = RandomForestClassifier() Rclf.fit(xtrain,ytrain)</pre>	restClassifier dtaConversionWarning: A column-vector y was pas	ssed when a 1d array w	as expected. Pleas	e change the shape of y to (n_samples,)), for example using ravel().	
	3., 52., 52., 34., 39., 28., 25., 8.])					
<pre>RandomForestClassifier() Rclf.predict(xtest) array([42., 44., 50., 36., 60., 38., 28., 6., 46., 61., 41., 28., Rclf.score(xtest, ytest) 0.0 reg.predict(xtest)</pre>						
RandomForestClassifier() Rclf.predict(xtest) array([42., 44., 50., 36., 60., 38., 28., 6., 46., 61., 41., 28., Rclf.score(xtest,ytest) 0.0 reg.predict(xtest) array([[35.68473249], [35.68572946], [35.68414179], [35.68440241], [35.6849835], [35.68497606], [35.68497606], [35.68497606], [35.68497606], [35.68494916], [35.68610607], [35.68610607], [35.68610607], [35.68610607], [35.68643519], [35.68653034],						
RandomForestClassifier() Rclf.predict(xtest) array([42., 44., 50., 36., 60., 38., 28., 6., 46., 61., 41., 28., 8.] Rclf.score(xtest, ytest) 0.0 reg.predict(xtest) array([[35.68473249], [35.68572946], [35.68414179], [35.68440241], [35.68440374], [35.68497606], [35.68497606], [35.68497606], [35.68497606], [35.68497606], [35.68494916], [35.68610607], [35.68610607], [35.68610607], [35.68643519],						