Data Analysis for Average Domestic Itinerary Fare based by Origin City ranked by Passengers in 2021

This project organizes data and visualize the average domestic airfares by US major origin cities. Data is taken from Bureau of US transportation for all the four quarters in 2021. Averages are computed using data from the Bureau of Transportation Statistics' Passenger Origin and Destination (O&D) Survey, a 10% sample of all airline tickets for U.S. carriers

In this project, data was downloaded and prepared for cleaning using python libraries numpy and pandas. Data exploration was done to get the meaningful results and then with the help of matplotlib and seaborn , data visulization of comparing all four quarter's average itinerary prices.

I implement various numpy and pandas functions like sort_values(), duplicated(),unique() and many more, which I learned from Data Analysis with Python: Zero To Pandas.

```
import numpy as np
import pandas as pd
```

```
!pip install matplotlib seaborn --upgrade --quiet
```

```
#read excel file
airfareq1_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q1_2021.x
airfareq2_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q2_2021.x
airfareq3_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q3_2021.x
airfareq4_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q4_2021.x
```

Data Preparation and Cleaning

In this section, loaded data was checked if there is any missing data, duplicate etc. All the different files were merged and the final data frame was ready for further exploration.

```
# display the data frame
airfareq1_df
```

	PassengerRank	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passenger
0	1	LAX	Los Angeles International	Los Angeles	CA	246.229990	265.846412	
1	2	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.528892	262.930125	
2	3	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.116554	273.281608	
3	4	DEN	Denver International	Denver	CO	230.591183	248.961707	
4	5	EWR	Newark Liberty International	Newark	NJ	238.211927	257.189573	

	PassengerRank	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passenger
...	...	...	...	...	...	...	...	
417	417	LNY	Lanai Airport	Lanai	HI	582.625000	629.041026	
418	419	MKT	Mankato Regional	Mankato	MN	154.500000	166.808562	
419	420	CPX	Benjamin Rivera Noriega	Culebra	PR	364.000000	392.998813	
420	421	JHM	Kapalua Airport	Kapalua	HI	484.000000	522.558861	
421	422	SNP	St. Paul Island	St. Paul	AK	929.000000	1003.010707	

422 rows × 8 columns

```
#to know the basic info of data frame
airfareq1_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 422 entries, 0 to 421
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   PassengerRank        422 non-null    int64  
 1   AirportCode          422 non-null    object 
 2   AirportName          422 non-null    object 
 3   CityName             422 non-null    object 
 4   StateName            422 non-null    object 
 5   AverageFare          422 non-null    float64
 6   AdjustedAverageFare  422 non-null    float64
 7   Passengerssample     422 non-null    int64  
dtypes: float64(2), int64(2), object(4)
memory usage: 26.5+ KB
```

```
#to find a particular record of an airport
airfareq1_df[airfareq1_df.AirportCode =='EWR']
```

	PassengerRank	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passengerssa
4	5	EWR	Newark Liberty International	Newark	NJ	238.211927	257.189573	65

```
# to find any missing data in whole data frame
for col_name in airfareq1_df.columns:
    permiss = np.mean(airfareq1_df[col_name].isnull())
    print('{}      {}%'.format(col_name,permiss))
```

```
PassengerRank       0.0%
AirportCode       0.0%
AirportName       0.0%
CityName       0.0%
StateName       0.0%
AverageFare       0.0%
 AdjustedAverageFare       0.0%
 Passengerssample       0.0%
```

```
# to find any duplicates
airfareq1_df.duplicated()
```

```
0      False
1      False
2      False
3      False
4      False
       ...
417    False
418    False
419    False
420    False
421    False
Length: 422, dtype: bool
```

```
# Round upto 2 decimal places for AverageFare  & AdjustedAvergeFare Column

airfareq1_df['AverageFare'] = np.round(airfareq1_df['AverageFare'],2)
```

```
airfareq1_df['AdjustedAverageFare'] = np.round(airfareq1_df['AdjustedAverageFare'],2)
```

```
airfareq1_df
```

	PassengerRank	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passenger
0	1	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	
1	2	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	
2	3	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	
3	4	DEN	Denver International	Denver	CO	230.59	248.96	

	PassengerRank	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passenger
4	5	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	
...	...	...	...	...	...	...	...	
417	417	LNY	Lanai Airport	Lanai	HI	582.62	629.04	
418	419	MKT	Mankato Regional	Mankato	MN	154.50	166.81	
419	420	CPX	Benjamin Rivera Noriega	Culebra	PR	364.00	393.00	
420	421	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	
421	422	SNP	St. Paul Island	St. Paul	AK	929.00	1003.01	

422 rows × 8 columns

```
# to find the average of fares in Quarter 1

print('The Average Fare for Quarter 1  in 2021 : {}'.format(np.mean(airfareq1_df.Averag
```

```
The Average Fare for Quarter 1  in 2021 : 341.3374407582939
```

```
# to drop column in data frame
airfareq1_df.drop('PassengerRank', axis = 1, inplace = True)
```

```
airfareq1_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
417	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
418	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
419	CPX	Benjamin Rivera Noriega	Culebra	PR	364.00	393.00	7
420	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6
421	SNP	St. Paul Island	St. Paul	AK	929.00	1003.01	3

422 rows × 7 columns

```
# Merge files for Quarter 2, 3 and 4

airfareq2_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q2_2021.x
```

```
airfareq2_df.drop('PassengerRank', axis = 1, inplace = True)
```

```
airfareq2_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare	AdjustedAverageFare	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	314.67	331.97	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	272.57	287.56	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	295.73	311.98	768501
3	DEN	Denver International	Denver	CO	265.67	280.28	739773
4	EWR	Newark Liberty International	Newark	NJ	301.78	318.37	652199
...	...	...	...	...	...	...	...
430	MCN	Middle Georgia Regional	Macon	GA	390.00	411.44	6
431	JHM	Kapalua Airport	Kapalua	HI	483.50	510.08	6
432	SNP	St. Paul Island	St. Paul	AK	1316.00	1388.36	3
433	CEC	Jack McNamara Field	Crescent City	CA	328.00	346.03	1
434	OLF	L. M. Clayton	Wolf Point	MT	770.00	812.34	1

435 rows × 7 columns

```
mergedairfare_df = airfareq1_df.merge(airfareq2_df,  on =['AirportCode','AirportName','
```

```
mergedairfare_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare_x	AdjustedAverageFare_x	Passengerssample	Ave
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228	3
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315	2
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501	2
3	DEN	Denver International	Denver	CO	230.59	248.96	739773	2

	AirportCode	AirportName	CityName	StateName	AverageFare_x	AdjustedAverageFare_x	Passengerssample	Ave
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199	3
...	...	...	...	...	...	...	...	
417	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13	4
418	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9	2
419	CPX	Benjamin Rivera Noriega	Culebra	PR	364.00	393.00	7	5
420	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6	4
421	SNP	St. Paul Island	St. Paul	AK	929.00	1003.01	3	13

422 rows × 9 columns

```
mergedairfare_df.rename({'AverageFare_x' : 'AverageFare_Q1' ,'AdjustedAverageFare_x' :
```

```
mergedairfare_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
417	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
418	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
419	CPX	Benjamin Rivera Noriega	Culebra	PR	364.00	393.00	7
420	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6
421	SNP	St. Paul Island	St. Paul	AK	929.00	1003.01	3

422 rows × 9 columns

```
airfareq3_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q3_2021.x
```

```
airfareq3_df.drop('PassengerRank', axis = 1, inplace = True)
```

```
mergedairfare1_df = mergedairfare_df.merge(airfareq3_df,  on =['AirportCode','AirportNa
```

```
mergedairfare1_df.rename({'AverageFare' : 'AverageFare_Q3' ,'AdjustedAverageFare' : 'Ad
```

```
mergedairfare1_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
416	GGW	Wokal Field Glasgow Valley County	Glasgow	MT	461.00	497.73	13
417	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
418	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
419	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6
420	SNP	St. Paul Island	St. Paul	AK	929.00	1003.01	3

421 rows × 11 columns

```
airfareq4_df = pd.read_excel (r'C:\Users\psinghw\Desktop\Practice\AverageFare_Q4_2021.x
```

```
combined_df = mergedairfare1_df.merge(airfareq4_df,  on =['AirportCode','AirportName','
```

```
combined_df.rename({'AverageFare' : 'AverageFare_Q4' ,'AdjustedAverageFare' : 'Adjusted
```

```
# combined data frame with all the 4 quarters
combined_df['AverageFare_Q1'] = np.round(combined_df['AverageFare_Q1'],2)
combined_df['AverageFare_Q2'] = np.round(combined_df['AverageFare_Q2'],2)
combined_df['AverageFare_Q3'] = np.round(combined_df['AverageFare_Q3'],2)
combined_df['AverageFare_Q4'] = np.round(combined_df['AverageFare_Q4'],2)
combined_df['AdjustedAverageFare_Q1'] = np.round(combined_df['AdjustedAverageFare_Q1'],
combined_df['AdjustedAverageFare_Q2'] = np.round(combined_df['AdjustedAverageFare_Q2'],
combined_df['AdjustedAverageFare_Q3'] = np.round(combined_df['AdjustedAverageFare_Q3'],
combined_df['AdjustedAverageFare_Q4'] = np.round(combined_df['AdjustedAverageFare_Q4'],
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
414	HVR	Havre City-County	Havre	MT	266.62	287.87	16
415	GGW	Wokal Field Glasgow Valley County	Glasgow	MT	461.00	497.73	13
416	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
417	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
418	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6

419 rows × 15 columns

```
combined_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
414	HVR	Havre City-County	Havre	MT	266.62	287.87	16
415	GGW	Wokal Field Glasgow Valley County	Glasgow	MT	461.00	497.73	13
416	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
417	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
418	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6

419 rows × 15 columns

Exploratory Analysis and Visualization

With all the individual 4 data frames for each quarter and one combined data frame which include all the information plus all the average prices and adjustment prices for all quarters.

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
# Which Quarter has the  highest average in 2021?

Av_df = combined_df[['AverageFare_Q1','AverageFare_Q2','AverageFare_Q3','AverageFare_Q4

print('        {}  {}  {}  {} and the Quarter with max average fare is {}'.format(combi
```

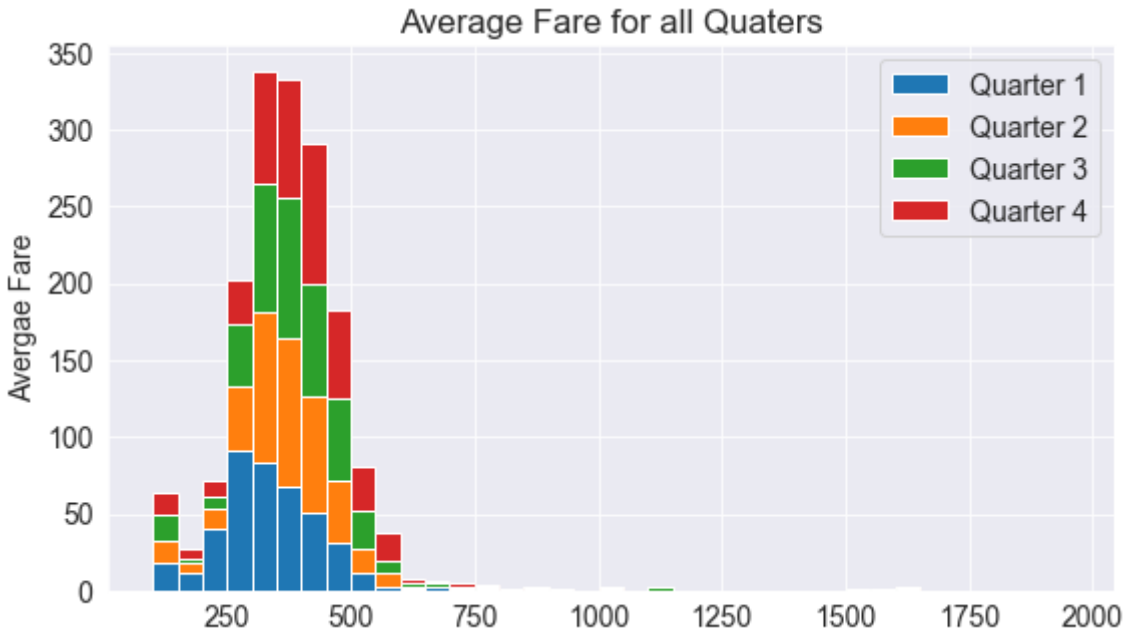
```
        AverageFare_Q1    339.951146
dtype: float64  AverageFare_Q2    371.055513
dtype: float64  AverageFare_Q3    386.726492
dtype: float64  AverageFare_Q4    398.883246
dtype: float64 and the Quarter with max average fare is 398.8832458233892
```

```
# A histogram plot explains the comparsion of Average airfare for all 4 quarter's

plt.title('Average Fare for all Quaters')
plt.ylabel('Avergae Fare')
plt.hist([combined_df.AverageFare_Q1,combined_df.AverageFare_Q2,combined_df.AverageFare

plt.legend(['Quarter 1','Quarter 2','Quarter 3','Quarter 4'])
```

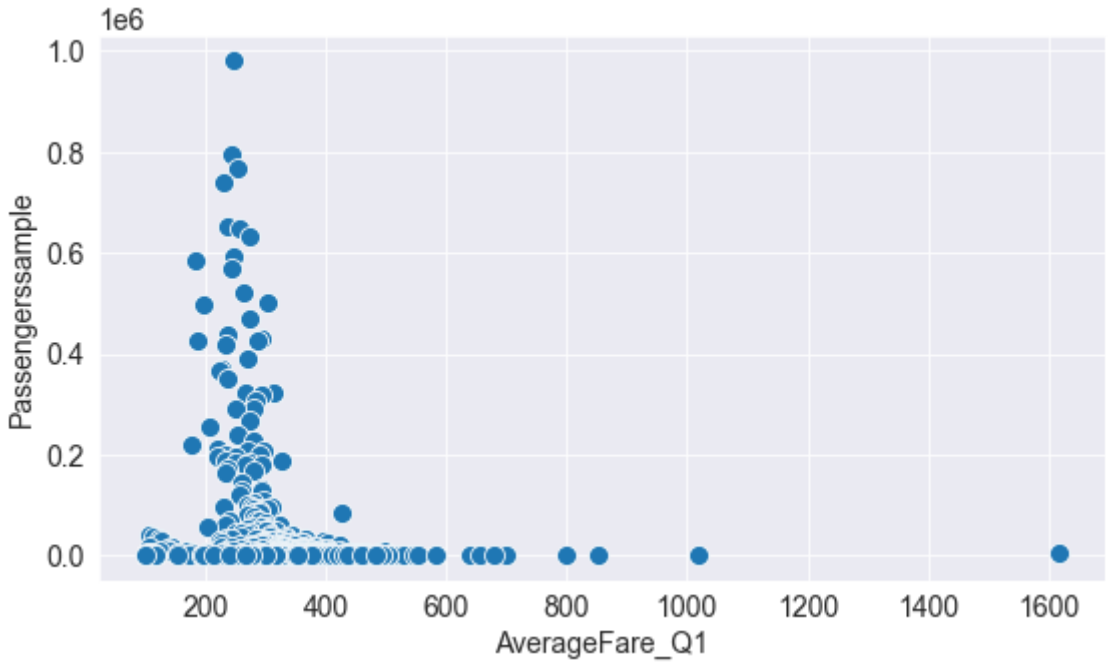
```
<matplotlib.legend.Legend at 0x16217a2b460>
```



The histogram explains that quarter 4 has the highest average fare.

```
#Scatterplot diagram shows average fare for quarter 1 and passenger's 10% survey sample
sns.scatterplot(x = 'AverageFare_Q1', y = 'Passengerssample',sizes=(200,100), hue_norm=
```

```
<AxesSubplot: xlabel='AverageFare_Q1', ylabel='Passengerssample'>
```

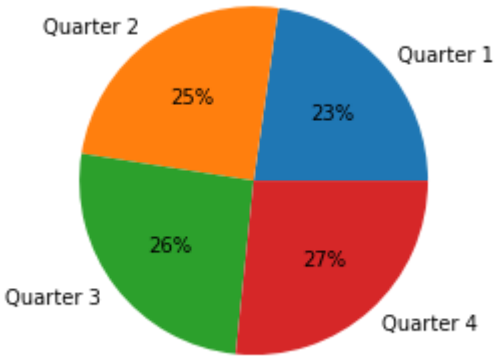


Based on passenger's sample survey, average fare lies betwwen $200.0 to 400.00$ except one at 1600.00 which might be an data entry error.

```
# A pie chart of all the 4 quarters
data = [combined_df.AverageFare_Q1.mean(),combined_df.AverageFare_Q2.mean(),combined_df
keys = ['Quarter 1','Quarter 2', 'Quarter 3','Quarter 4']

plt.pie(data,labels = keys,autopct= '%.0f%%')
```

```
([<matplotlib.patches.Wedge at 0x28c9aaa6790>,
  <matplotlib.patches.Wedge at 0x28c9aaa6e50>,
  <matplotlib.patches.Wedge at 0x28c9aab35b0>,
  <matplotlib.patches.Wedge at 0x28c9aab3cd0>],
 [Text(0.8316101427436979, 0.7200170626351895, 'Quarter 1'),
  Text(-0.6527615375345166, 0.8853826150967581, 'Quarter 2'),
  Text(-0.8722271919807352, -0.6702385587008568, 'Quarter 3'),
  Text(0.7364116186043715, -0.8171278528997099, 'Quarter 4')],
 [Text(0.4536055324056534, 0.3927365796191942, '23%'),
  Text(-0.35605174774609993, 0.4829359718709589, '25%'),
  Text(-0.4757602865349464, -0.3655846683822855, '26%'),
  Text(0.4016790646932935, -0.4457061015816599, '27%')])
```



Pie chart explains that quarter 4 has highest average fare of 27%

```
# Seaborn's pairplot which gives us a variety of graphs with different comparsions of c
sns.pairplot(combined_df, height = 2)
```

```
<seaborn.axisgrid.PairGrid at 0x16212b80d00>
```

combined_df

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
...	...	...	...	...	...	...	...
414	HVR	Havre City-County	Havre	MT	266.62	287.87	16
415	GGW	Wokal Field Glasgow Valley County	Glasgow	MT	461.00	497.73	13
416	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
417	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
418	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6

419 rows × 15 columns

```
# Data frame of Statewise Adjustment fares

Statewise_df = combined_df.groupby('StateName')[['AdjustedAverageFare_Q1', 'AdjustedAve
```

```
Statewise_df
```

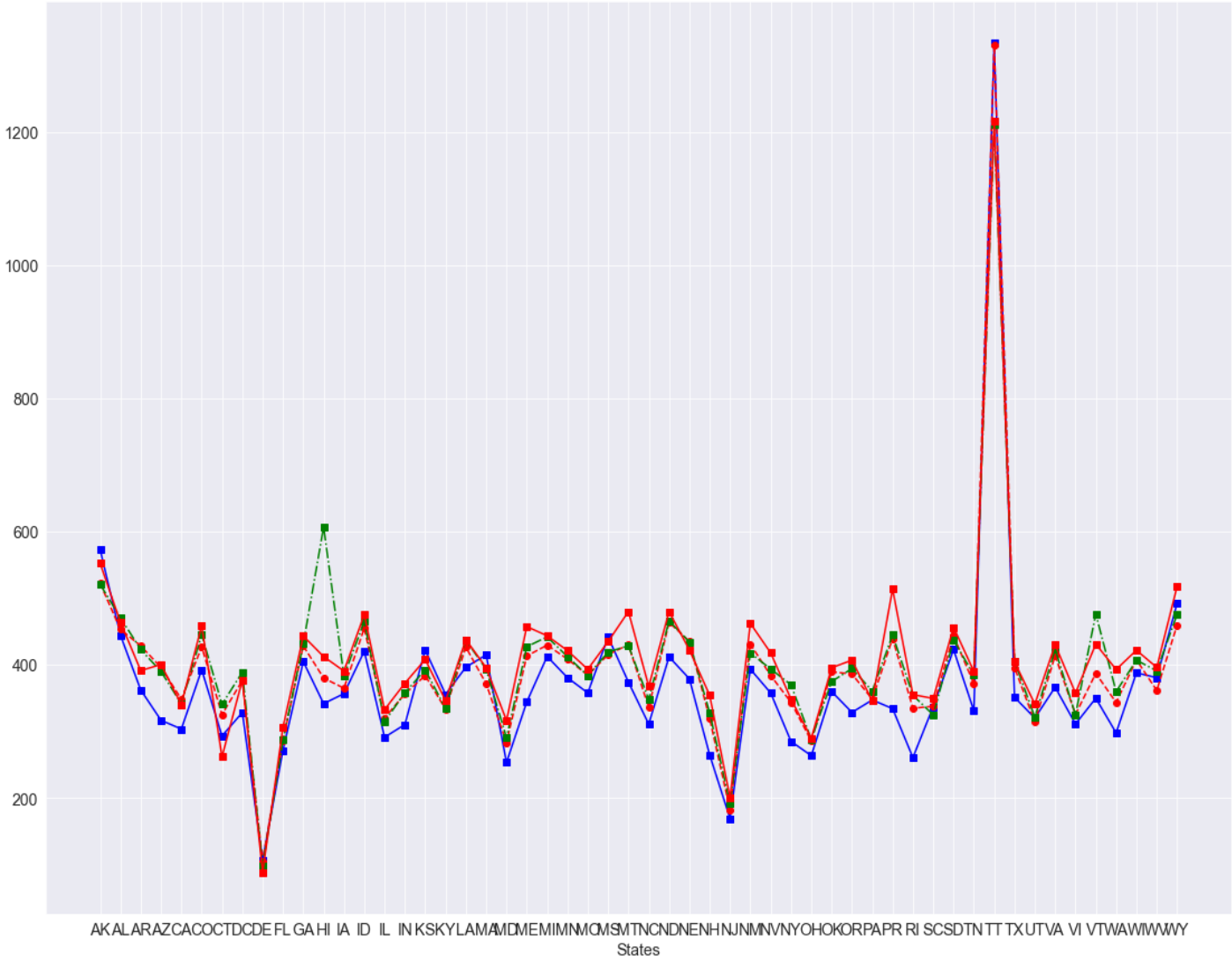
StateName	AdjustedAverageFare_Q1	AdjustedAverageFare_Q2	AdjustedAverageFare_Q3	AdjustedAverageFare_Q4
AK	573.659545	522.545455	521.077273	552.682273
AL	444.678333	453.268333	470.821667	464.203333
AR	362.325714	428.555714	422.980000	391.405714
AZ	316.660000	396.027500	389.907500	400.565000
CA	303.650435	348.546957	341.430870	340.229565
CO	392.640000	427.062500	445.348333	459.018333
CT	293.185000	325.435000	340.695000	261.845000
DC	328.435000	376.520000	389.150000	377.040000
DE	106.860000	105.140000	97.260000	88.250000
FL	271.742105	286.444211	287.063158	306.325789
GA	404.881429	428.504286	431.535714	443.271429
HI	341.708750	380.183750	607.208750	412.445000
IA	356.732857	364.902857	383.384286	390.720000
ID	420.426667	454.950000	465.116667	475.455000
IL	291.263333	320.251667	314.736667	332.813333
IN	310.380000	356.502500	358.210000	372.405000
KS	422.342857	383.828571	392.435714	408.024286
KY	355.730000	333.515000	334.752500	346.532500
LA	396.491429	426.632857	437.308571	438.017143
MA	415.012000	371.600000	393.282000	395.692000

StateName	AdjustedAverageFare_Q1	AdjustedAverageFare_Q2	AdjustedAverageFare_Q3	AdjustedAverageFare_Q4
MD	253.370000	283.366667	290.626667	316.470000
ME	344.403333	413.430000	427.263333	457.446667
MI	412.639412	429.632353	442.035294	443.724706
MN	380.730000	408.619000	412.405000	421.411000
MO	357.891250	384.487500	383.955000	393.303750
MS	441.887143	415.387143	419.167143	434.904286
MT	374.094000	430.233000	428.813000	480.077000
NC	311.008000	335.797000	348.295000	367.661000
ND	411.957500	463.896250	466.168750	479.986250
NE	379.101111	435.282222	434.305556	422.435556
NH	264.533333	319.016667	328.076667	354.153333
NJ	169.223333	182.423333	191.550000	200.470000
NM	393.510000	431.161667	417.703333	462.316667
NV	357.826667	384.026667	393.846667	418.030000
NY	285.108889	343.770000	369.995000	348.615000
OH	263.528571	285.751429	287.841429	289.927143
OK	359.925000	389.585000	374.890000	395.612500
OR	328.620000	386.228333	394.510000	407.050000
PA	347.233846	348.442308	360.701538	345.859231
PR	334.316667	439.700000	445.013333	513.833333
RI	260.730000	334.720000	354.360000	355.260000
SC	337.238333	337.906667	324.823333	349.565000
SD	422.976000	443.762000	436.832000	456.510000
TN	331.072000	371.582000	384.898000	390.094000
TT	1333.605000	1330.845000	1211.545000	1217.620000
TX	351.303846	395.888462	405.366923	404.500000
UT	320.062857	314.232857	321.270000	341.075714
VA	366.477143	413.647143	423.154286	430.315714
VI	311.105000	325.505000	325.050000	357.615000
VT	349.885000	387.260000	476.455000	431.235000
WA	297.792222	343.496667	359.801111	392.978889
WI	388.020000	407.585000	407.022500	421.958750
WV	380.734286	361.582857	390.950000	396.157143
WY	492.116667	459.744444	475.556667	517.622222

```
# A line graph showing statewise adjusted fares by airlines
plt.figure(figsize=(19,15))
plt.xlabel('States')
```

```
plt.plot( Statewise_df.AdjustedAverageFare_Q1, 's-b')
plt.plot( Statewise_df.AdjustedAverageFare_Q2, 'o--r')
plt.plot( Statewise_df.AdjustedAverageFare_Q3, 's-.g')
plt.plot( Statewise_df.AdjustedAverageFare_Q4, 'r-s')
```

```
[<matplotlib.lines.Line2D at 0x1621aff2400>]
```



This line chart explains that Texas state has hihest adjustment fares. Adjustment fares are fares with some discount on base fare.

Asking and Answering Questions

Some interesting data explorations which were performed on different data frames.

```
# to find top 5 airports which  has the best average fares for quarter 1

combined_df.sort_values(by ='AverageFare_Q1', ascending = True).head(5)
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
279	SMX	Santa Maria Public/Capt. G. Allan Hancock Field	Santa Maria	CA	98.56	106.41	1639
351	ILG	New Castle	Wilmington	DE	98.98	106.86	457

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
300	OGD	Ogden-Hinckley	Ogden	UT	100.06	108.03	1229
196	USA	Concord Padgett Regional	Concord	NC	102.60	110.77	6728
201	LBE	Arnold Palmer Regional	Latrobe	PA	104.20	112.50	6136

```
# Which State has max airports
airports_count = airfareq1_df.groupby('StateName')['StateName'].count().sort_values(asc

airports_count.sort_values(ascending = False).head(1)
```

```
StateName
TX    26
Name: StateName, dtype: int64
```

```
# to add a new column in data frame to get combined average fare for each airport for a

combined_df['CombinedAverageFare'] = combined_df.AverageFare_Q1 + combined_df.AverageFa
combined_df
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
0	LAX	Los Angeles International	Los Angeles	CA	246.23	265.85	981228
1	ORD	Chicago O'Hare International	Chicago-O'Hare	IL	243.53	262.93	795315
2	ATL	Hartsfield-Jackson Atlanta International	Atlanta	GA	253.12	273.28	768501
3	DEN	Denver International	Denver	CO	230.59	248.96	739773
4	EWR	Newark Liberty International	Newark	NJ	238.21	257.19	652199
...	...	...	...	...	...	...	...
414	HVR	Havre City-County	Havre	MT	266.62	287.87	16
415	GGW	Wokal Field Glasgow Valley County	Glasgow	MT	461.00	497.73	13
416	LNY	Lanai Airport	Lanai	HI	582.62	629.04	13
417	MKT	Mankato Regional	Mankato	MN	154.50	166.81	9
418	JHM	Kapalua Airport	Kapalua	HI	484.00	522.56	6

419 rows × 15 columns

```
# Which airport has cheap fare from  all 4 quarters
combined_df.sort_values(by= 'CombinedAverageFare', ascending = True ).head(1)
```

	AirportCode	AirportName	CityName	StateName	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample
300	OGD	Ogden-Hinckley	Ogden	UT	100.06	108.03	1229

```
# To find correlation between columns
combined_df.corr(method='pearson')
```

	AverageFare_Q1	AdjustedAverageFare_Q1	Passengerssample	AverageFare_Q2	AdjustedAvera
AverageFare_Q1	1.000000	1.000000	-0.265796	0.893685	
AdjustedAverageFare_Q1	1.000000	1.000000	-0.265796	0.893686	
Passengerssample	-0.265796	-0.265796	1.000000	-0.244614	
AverageFare_Q2	0.893685	0.893686	-0.244614	1.000000	
AdjustedAverageFare_Q2	0.893685	0.893686	-0.244616	1.000000	
AverageFare_Q3	0.805195	0.805196	-0.218215	0.873281	
AdjustedAverageFare_Q3	0.805194	0.805196	-0.218216	0.873281	
PassengerRank	0.409879	0.409880	-0.602113	0.392735	
AverageFare_Q4	0.870689	0.870690	-0.235005	0.930975	
AdjustedAverageFare_Q4	0.870690	0.870691	-0.235006	0.930974	
CombinedAverageFare	0.934218	0.934219	-0.252198	0.968703	

Inferences and Conclusion

Based on the passengers survey sample, we came to conculsion that during quarter 4, the fares are little high than rest of the year.

Texas is the state which has maximum airports in US

Based on the passengers 10 % survey sample, the average fare lies between $200.00 - 300.00$

Ogden-Hinckley airport in Utah has the cheapest average fare for all quarters.

References and Future Work

Check out the following resources to learn more about the dataset and tools used in this notebook:

Pandas user guide: https://pandas.pydata.org/docs/user_guide/index.html Matplotlib user guide: https://matplotlib.org/3.3.1/users/index.html Seaborn user guide & tutorial: https://seaborn.pydata.org/tutorial.html

For future predictions, I can extend this project by adding other discounts offered by airlines to passengers.Mileage plus airfares can also be included.