Capstone Project – Applied Data Science

Introduction: A travel agency which mainly deals with hotel bookings has multiple locations across the United States is expanding. The management team is looking into the options and trying to decide which city would be best suitable to target for business opportunities.
Good starting point is the explore the cities which have following characteristics:

- City has tourist attractions
- City is a commercial hub
- City has airports
- City has big volume of business and leisure travelers

Number of hotels around any airport seems like a good indicator of the tourism/commercial activity in that city which means it is also a good location to open a new office. In this project we will explore whether this hunch that Socio Economic measures such as GDP and Annual Enplanements at airports is a good measure to estimate the density of hotels around the airport of that city is true or not.

Data: Since the aim is to look the big cities through which a lot of commuters pass by on daily basis. Following sets of data will be explored to solve the problem.

- Annual enplanement information of Top 30 US airports
- GDP information corresponding to cities with high annual enplanements
- Number of Hotels/Motels/Resorts around the airport

  Annual Enplanements : Airport which has high annual enplanements signify that a lot of commuters pass through this destination. It also means that this location is either a famous tourist hub or a major commercial hub. Annual enplanements information of Top 30 US Airports is available on Wikipedia here. The data will be scraped from the website to use it in the analysis.

  GDP Information : Logically within the cities which host top 30 airports, cities who have higher GDPs a probably the business hot spots and a preferred travel destination for many. Therefore, GDP information of the cities which host top 30 airports can give even more insights and facilitate in our decision making. GDP data is also available of Wikipedia here. This data will be scraped and matched with the annual enplanements data.

  Hotels Around Airport : This is another indicator of high traveler through a certain city. Data regarding number of hotels/motels/resorts within 10 miles of the airport will be fetched through Foursquare API. Foursquare API documentation shows different 'category_ids' belonging to different hotel types. This will come in handy to make intelligent API calls instead of fetching all the venues first then filtering it later. Following are few category_id example from this page.

Methodology: The first step is to prepare the data for the analysis. To do that Airport data and GDP data is scraped from the Wikipedia and stored in data frames. Figure 1 and

Figure 2 show these data frames respectively.

Figure 1: Top -30 US Airports Dataset Scraped from Wikipedia

In [10]:
```
top_airports
```
Out[10]:

	Rank(2018)	Airports (large hubs)	IATACode	Major city served	State	2019	2018[3]	2017[4]	2016[5]	2015[6]	2014[7]	2013[8]	2012[9]	2011[1
0	1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	GA	NaN	51866464	50251964	50501858	49340732	46604273	45308407	45798809	444141
1	2	Los Angeles International Airport	LAX	Los Angeles	CA	NaN	42626783	41232432	39636042	36351226	34314197	32425892	31326268	305287
2	3	O'Hare International Airport	ORD	Chicago	IL	NaN	39874879	38593028	37589899	36305668	33686811	32317835	32171743	318923
3	4	Dallas/Fort Worth International Airport	DFW	Dallas	TX	NaN	32800721	31816933	31283579	31589832	30766940	29038128	28022877	275183

Figure 2: GDP Data scraped from Wikipedia

In [11]:
```
top_gdp
```
Out[11]:

	2017 Rank	Metropolitan area	2018	2017	2016	2015	2014	2013	2012
0	1	New York-Newark-Jersey City, NY-NJ-PA (Metropo...	1772319	1698122	1634671	1577366	1511763	1439043	1401233
1	2	Los Angeles-Long Beach-Anaheim, CA (Metropolit...	1047661	995114	945600	912384	858170	820353	788081
2	3	Chicago-Naperville-Elgin, IL-IN-WI (Metropolit...	689464	659855	641589	627033	599805	577948	561016
3	4	San Francisco-Oakland-Berkeley, CA (Metropolit...	548613	509382	469472	446344	413519	383254	364594
4	5	Washington-Arlington-Alexandria, DC-VA-MD-WV (...	540684	515553	500084	481861	460254	448268	442224
5	6	Dallas-Fort Worth-Arlington, TX (Metropolitan ...	512509	482218	458973	442879	420929	394178	375065
6	7	Houston-The Woodlands-Sugar Land, TX (Metropol...	478778	447521	430444	446486	430726	423766	404431
7	8	Boston-Cambridge-Newton, MA-NH (Metropolitan S...	463570	439144	421783	406675	381353	365670	357087
8	9	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD (M...	444148	422539	416110	406605	388621	374787	364052
9	10	Atlanta-Sandy Springs-Roswell, GA (Metropolita...	397261	385542	369806	347604	326502	307750	291481
10	11	Seattle-Tacoma-Bellevue, WA (Metropolitan Stat...	392036	356572	334411	317153	296028	280291	267472
11	12	Miami-Fort Lauderdale-West Palm Beach, FL (Met...	354740	344882	330784	315624	292308	268198	266563
12	13	San Jose-Sunnyvale-Santa Clara, CA (Metropolit...	331020	275293	253900	237832	210690	193716	180996
13	14	Detroit-Warren-Dearborn, MI (Metropolitan Stat...	267731	260612	250430	242296	229348	220277	215614
	15	Minneapolis-St. Paul-Bloomington, MN-WI (Metro...	263690	260106	250376	242053	232416	220938	213855

ww.mozilla.org/en-US/firefox/central/

Now the idea is to create a data frame step by step which contains all the relevant information we need. In two data frames shown above. Data frame in Figure 1 has information regarding top airports and annual enplanements from that airport while
Figure 2 has information regarding GDPs of metropolitan areas. From Figure 1 we know which major city is served by which airport. This information can be used to look up the city name in Metropolitan Area column of
Figure 2 and find the appropriate row number to return the GDP of that city. Note that in this project we are only concerned with 2018 data. Code snipped shown in
Figure 3 does the lookup and data frame with desired columns put together in one place can also be seen.

Figure 3: Lookup Function

In [16]:
```python
#Looking up GDP data and storing it in gdp_airport df
for i in range(len(gdp_airport)):
    lookedup_row = 0
    lookup_city = gdp_airport.iloc[i,1]
    lookup_city = lookup_city.split(',')[0]
    lookedup_row = top_gdp[top_gdp['Metropolitan area'].str.contains(lookup_city,case=False)]

    if lookedup_row.empty == False:
        gdp_airport.iloc[i,2] = lookedup_row.iloc[0,1]
        gdp_airport.iloc[i,4] = lookedup_row.iloc[0,2]

gdp_airport
```

Out[16]:

	Airport	City	Area	AirTraffic	GDP
0	Hartsfield–Jackson Atlanta International Airport	Atlanta	Atlanta-Sandy Springs-Roswell, GA (Metropolita...	51866464	397261
1	Los Angeles International Airport	Los Angeles	Los Angeles-Long Beach-Anaheim, CA (Metropolit...	42626783	1047661
2	O'Hare International Airport	Chicago	Chicago-Naperville-Elgin, IL-IN-WI (Metropolit...	39874879	689464
3	Dallas/Fort Worth International Airport	Dallas	Dallas-Fort Worth-Arlington, TX (Metropolitan ...	32800721	512509
4	Denver International Airport	Denver	Denver-Aurora-Lakewood, CO (Metropolitan Stati...	31363573	214157
5	John F. Kennedy International Airport	New York	New York-Newark-Jersey City, NY-NJ-PA (Metropo...	30541459	1772319

Now, its time to fetch the geographical coordinates of these airports. Geopy library is used to do the task. Airport names are provided as search input. Latitude and Longitude values returned from the function are stored in data frame shown in Figure 4. It can seen in the figure below that GeoPy was not able to search the coordinates of some airports. This happened with exactly two airports; George Bush Intercontinental Airport and General Edward Lawrence Logan International. There coordinates values were put manually.

Figure 4: Latitude and Longitude Values Returned from GeoPy

```python
coords
```

Out[17]:

	Latitude	Longitude
0	33.637799	-84.429271
1	33.942168	-118.421376
2	41.977985	-87.909321
3	32.896519	-97.046522
4	39.850188	-104.675328
5	40.642948	-73.779373
6	37.622452	-122.384072
7	47.447567	-122.308016
8	36.086103	-115.161120
9	28.412904	-81.309443
10	40.689064	-74.177255
11	35.210741	-80.945744
12	33.432908	-112.009422
13	NaN	NaN
14	25.794979	-80.286723

Joining this data frame with the desired data frame we have preparing so far for the data analysis now looks like Figure 5

Figure 5: Desired Data Frame with Geographical Coordinates

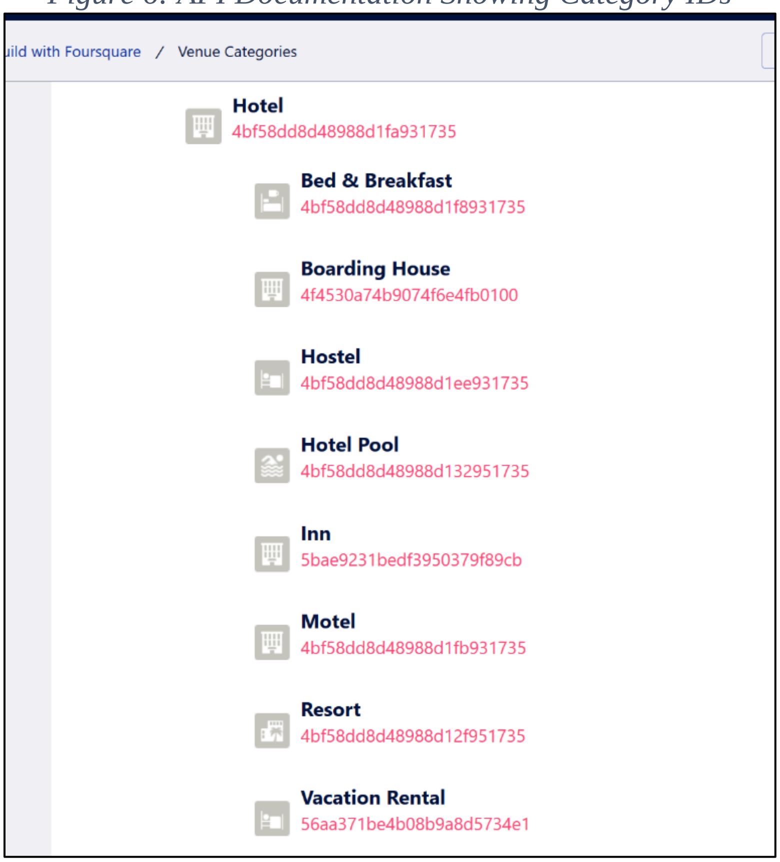
In [19]:
```
#Attach the coordinate values to gdp Airport dataframe
gdp_airport = gdp_airport.join(coords)
gdp_airport
```

Out[19]:

	Airport	City	Area	AirTraffic	GDP	Latitude	Longitude
0	Hartsfield–Jackson Atlanta International Airport	Atlanta	Atlanta-Sandy Springs-Roswell, GA (Metropolita...	51866464	397261	33.637799	-84.429271
1	Los Angeles International Airport	Los Angeles	Los Angeles-Long Beach-Anaheim, CA (Metropolit...	42626783	1047661	33.942168	-118.421376
2	O'Hare International Airport	Chicago	Chicago-Naperville-Elgin, IL-IN-WI (Metropolit...	39874879	689464	41.977985	-87.909321
3	Dallas/Fort Worth International Airport	Dallas	Dallas-Fort Worth-Arlington, TX (Metropolitan ...	32800721	512509	32.896519	-97.046522
4	Denver International Airport	Denver	Denver-Aurora-Lakewood, CO (Metropolitan Stati...	31363573	214157	39.850188	-104.675328
5	John F. Kennedy International Airport	New York	New York-Newark-Jersey City, NY-NJ-PA (Metropo...	30541459	1772319	40.642948	-73.779373
6	San Francisco International Airport	San Francisco	San Francisco-Oakland-Berkeley, CA (Metropolit...	27794154	548613	37.622452	-122.384072
7	Seattle–Tacoma International Airport	Seattle	Seattle-Tacoma-Bellevue, WA (Metropolitan Stat...	24894338	392036	47.447567	-122.308016
8	McCarran International Airport	Las Vegas	Las Vegas-Henderson-Paradise, NV (Metropolitan...	23655285	122423	36.086103	-115.161120
9	Orlando International Airport	Orlando	Orlando-Kissimmee-Sanford, FL (Metropolitan St...	23184634	138947	28.412904	-81.309443
10	Newark Liberty International Airport	New York	New York-Newark-Jersey City, NY-NJ-PA (Metropo...	22798354	1772319	40.689064	-74.177255
11	Charlotte Douglas International Airport	Charlotte	Charlotte-Concord-Gastonia, NC-SC (Metropolita...	22283574	169862	35.210741	-80.945744

Now only this left in preparing our data frame is the number of hotels within 10 miles radius of the top airports shown in Figure 5. To accomplish this take Foursquare API will be used. We are only concerned with how many hotels are around 10 miles radius. To make our search easy and save us from the post processing of the response JSON file, Foursquare API actually allows us to search for specific category. In request URL category_id of desired category to search into is passed which takes care of this problem. Foursqaure API documentation shown in

Figure 6: API Documentation Showing Category IDs



Different category ids are passed as a list in the request URL. The API call with category filters is shown in Figure 7. Notice how ids are passed in categoryId field.

Figure 7: API request URL

```
Out[4]: 'https://api.foursquare.com/v2/venues/explore?&client_id=SOACOHUBSDFDNPMEO3V0WHJE3PI43LY4TOTSEDIST3L3UPLD&client_se
        cret=OU3ELTQ1SYEGB31NJRVRM0A1GWG20YMQTB2SRH5EEKFFN1KL&v=20200130&ll=32.7818135,-96.8144203&radius=16100&limit=500&c
        ategoryId=4bf58dd8d48988d1f8931735,4f4530a74b9074f6e4fb0100,4bf58dd8d48988d1ee931735,4bf58dd8d48988d132951735,5bae9
        231bedf3950379f89cb,4bf58dd8d48988d1fb931735,4bf58dd8d48988d12f951735,56aa371be4b08b9a8d5734e1'

[5]:    results = requests.get(url).json()
        results
```

Now, the next task is to figure out number of responses from the API for each airport. This requires understanding the structure of the JSON file sent by Foursquare. Figure 8 is a screenshot of JSON file. If you look closely, the response key has another key called totalResults which stores the information we need. Now its just the matter of indexing the JSON file in a way which can take us to totalResults value. Which in-fact is a simple indexing of the result file as total_results = results['response']['totalResults'].

Figure 8: JSON File from Foursquare

```
{'meta': {'code': 200, 'requestId': '5efe62bea0a468438fc19db7'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Dallas',
  'headerFullLocation': 'Dallas',
  'headerLocationGranularity': 'city',
  'query': 'b b',
  'totalResults': 74,
  'suggestedBounds': {'ne': {'lat': 32.92671364490014,
    'lng': -96.64239344678614},
   'sw': {'lat': 32.63691335509986, 'lng': -96.98644715321385}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4af2f201f964a52056e921e3',
       'name': 'The Ritz-Carlton, Dallas',
```

Multiple API calls are made for different locations and the total results returned were saved in a data frame. Figure 9 shows the for loop which makes multiple API calls and stores the values we are seeking in a data frame shown in Figure 10. Last two lines of this code save the totalResults in numHotel data frame.

Figure 9: For Loop to Make API Calls

```python
#df to store the API response
numHotels = pd.DataFrame(columns=['Airport', 'NumHotels'])

#Search for hotels in 10 miles radius for various airports

for i in range(len(gdp_airport)):
    total_results = 0
    LIMIT = 500
    radius = 16100
    latitude = gdp_airport['Latitude'][i]
    longitude = gdp_airport['Longitude'][i]
    airport = gdp_airport['Airport'][i]

    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        latitude,
        longitude,
        radius,
        LIMIT,
        hotel_cat)
    results = requests.get(url).json()
    total_results = results['response']['totalResults']
    numHotels = numHotels.append({'Airport':airport,'NumHotels':total_results}, ignore_index = True)
```

Figure 10: Number of Hotels Fetched via API

3]:

	Airport	NumHotels
0	Hartsfield–Jackson Atlanta International Airport	169
1	Los Angeles International Airport	200
2	O'Hare International Airport	146
3	Dallas/Fort Worth International Airport	188
4	Denver International Airport	81
5	John F. Kennedy International Airport	97
6	San Francisco International Airport	128
7	Seattle–Tacoma International Airport	108
8	McCarran International Airport	254
9	Orlando International Airport	140
10	Newark Liberty International Airport	159
11	Charlotte Douglas International Airport	172
12	Phoenix Sky Harbor International Airport	214
13	George Bush Intercontinental Airport	152
14	Miami International Airport	244
15	General Edward Lawrence Logan International Ai...	190

Now we can append these columns to the data frame we have been trying to prepare so far. This is the final step in the data preparation and our data frame looks like the one shown in Figure 11.

Figure 11: Final Data Frame which will be Used in Analysis

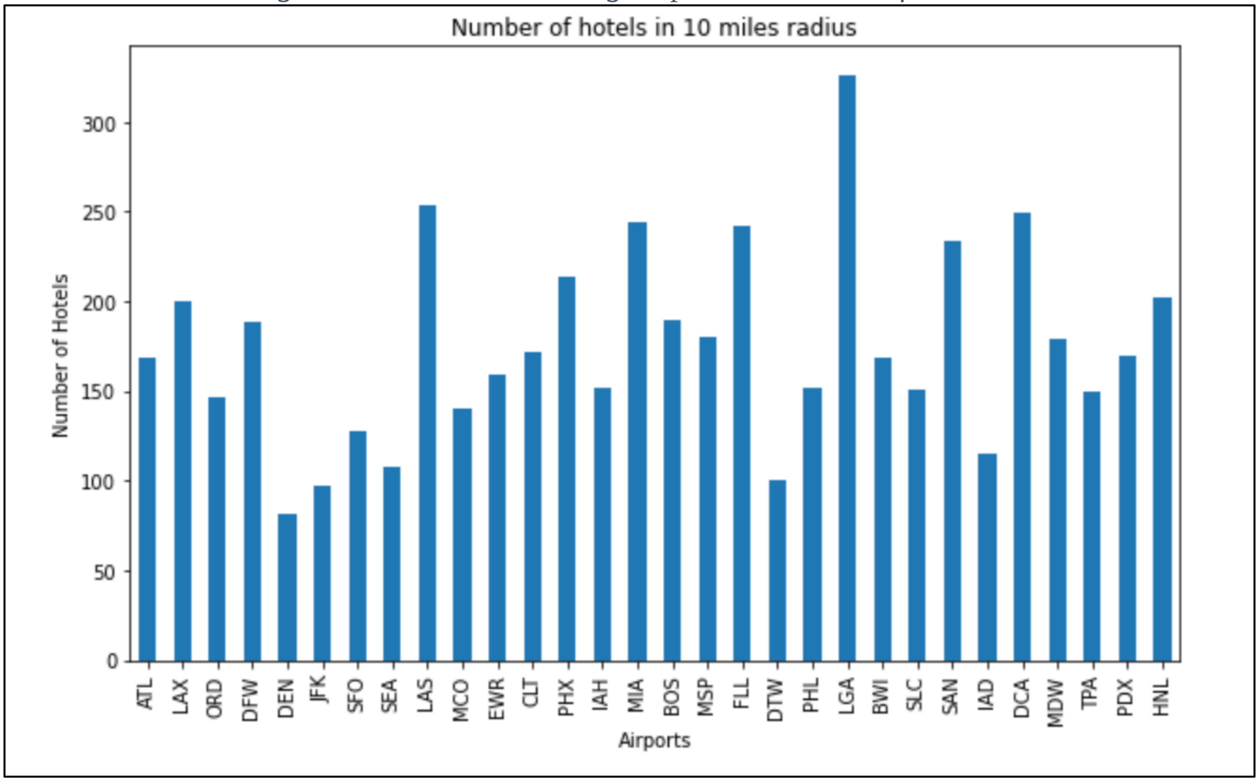
```
gdp_airport = gdp_airport.join(top_airports['IATACode'])
gdp_airport
```

Out[24]:

	Airport	City	Area	AirTraffic	GDP	Latitude	Longitude	NumHotels	IATACode
0	Hartsfield–Jackson Atlanta International Airport	Atlanta	Atlanta-Sandy Springs-Roswell, GA (Metropolita...	51866464	397261	33.637799	-84.429271	169	ATL
1	Los Angeles International Airport	Los Angeles	Los Angeles-Long Beach-Anaheim, CA (Metropolit...	42626783	1047661	33.942168	-118.421376	200	LAX
2	O'Hare International Airport	Chicago	Chicago-Naperville-Elgin, IL-IN-WI (Metropolit...	39874879	689464	41.977985	-87.909321	146	ORD
3	Dallas/Fort Worth International Airport	Dallas	Dallas-Fort Worth-Arlington, TX (Metropolitan ...	32800721	512509	32.896519	-97.046522	188	DFW
4	Denver International Airport	Denver	Denver-Aurora-Lakewood, CO (Metropolitan Stati...	31363573	214157	39.850188	-104.675328	81	DEN
5	John F. Kennedy International Airport	New York	New York-Newark-Jersey City, NY-NJ-PA (Metropo...	30541459	1772319	40.642948	-73.779373	97	JFK
6	San Francisco International Airport	San Francisco	San Francisco-Oakland-Berkeley, CA (Metropolit...	27794154	548613	37.622452	-122.384072	128	SFO
7	Seattle–Tacoma International Airport	Seattle	Seattle-Tacoma-Bellevue, WA (Metropolitan Stat...	24894338	392036	47.447567	-122.308016	108	SEA
[illegible]	[illegible]	[illegible]	Las Vegas-Henderson-Paradise, NV	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

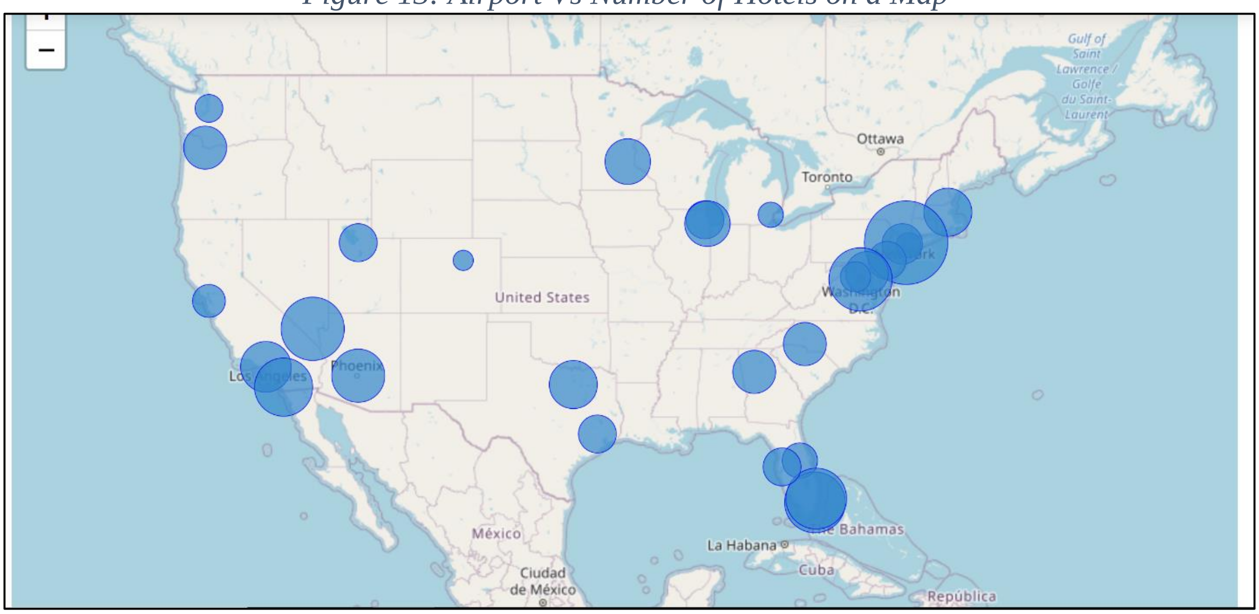
Now we can plot the some figures to see the spread of hotels around each airport.

Figure 12: Bar Plot Showing Airport Vs Number of Hotels



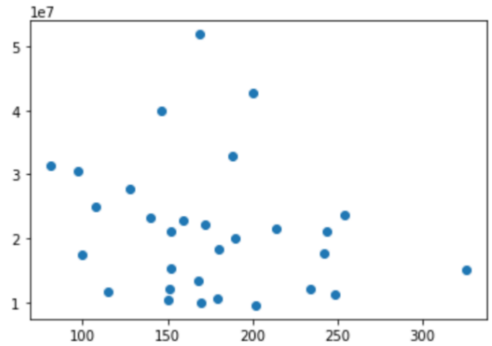
Map shown in Figure 13 is plotted using Folium. It shows the same information but on a map. Size of bubble indicates how many airports are there within 10 miles radius. Large bubble size means more hotels.

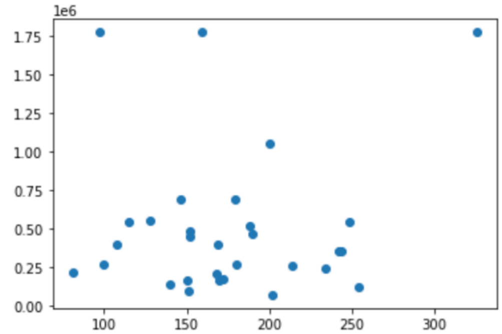
Figure 13: Airport Vs Number of Hotels on a Map



Now let us go back to initial assumptions which states that somehow Socio-Economic measures such as GDP, Annual Enplanements at an airport are a good indicator of predicting how many hotels could be around that airport. In order to do this, we will build a Multi Linear Regression model with GDP and Annual Enplanements as independent variable and Number of Hotels as dependent variable. But first lets see what scatter plots look like for these independent variables.

Figure 14: Scatter Plot of Independent Variables Vs Dependent Variable





Hotel Vs Enplanements	Hotel Vs GDP

From the look it seems like the relationship does not look linear at all. It rather seems random in both cases. A MLR model can confirm or dismiss this doubt. MLR shown in Figure 15 is built using scikitlearn library.

Figure 15: MLR Model

```
X = pd.DataFrame([gdp_airport['GDP'], gdp_airport['AirTraffic']])
X = X.T
y = pd.DataFrame([gdp_airport['NumHotels']])
y = y.T

lm = LinearRegression()
lm.fit(X,y)
lm.score(X,y)
```

```
2]:  0.0747813627408721
```

It can be seen that R^2 value is pretty low in this case which confirms that the relation between dependent and independent variables is not linear at all.

Results: Management team of the travel agency was trying to decide weather the travel agency, which mainly concerns with hotel bookings was trying to figure out weather Socio Economic measures such as GDP and Annual Enplanements at airports is a good measure to estimate the density of hotels around the airport of that city. The analysis of Top-30 US airports shows that it is not a good measure for desired prediction. This could be explained by following reasons:

- In some cities the nearest major airports are within the city limits and in others they are in outskirts. If airport is in outskirts then hotels are probably going to be far from the airport even if the airport is busy and city has strong GDP.
- The high enplanement rates could also mean that the airport is only serving as a hub, and a lot of passenger traffic is only connecting passengers.
- Sometimes a single city can have multiple airports. In our data set New York has two major airports; John F. Kennedy (JFK) and La Guardia (LGA). JFK has high annual

enplanements as compared to LGA but Foursquare data shows that LGA has 326 hotels around it while JFK only has 97.

Conclusion: The travel agency should explore other criteria like number of big companies in the area, employment rates of the city, GDP per capita, major tourist attractions, number of universities, conference centers etc. to be used as good predictors to estimate hotel density around any airport.