BATTLE OF NEIGHBORHOODS
-To determine best neighborhood for a Restaurant

I] INTRODUCTION

Discussion of the Background

Silicon Valley is the heart of global innovation, technology and social media. It is home to big tech giants like Apple, Facebook, Google, Microsoft, LinkedIn, Uber and many more. With such big names come tremendous job opportunities which pulls people all over the world to migrate and make this Silicon Valley their home. But this huge influx of people eventually causes its own problems like, high property rates, cost of living and crazy traffic. While this fast-paced life has its allure, it does not suite everyone and some prefer to have the advantages of Silicon Valley balanced with a somewhat quiet and relaxing life closer to the nature.

For these individuals the "SILICON FOREST" is a godsent! The silicon forest is a group of tech companies established in the Willamette Valley in Oregon. Since 1980's to date the area around silicon forest has seen a rise in population, infrastructure growth and development, expansion of cities, growth in housing and night life. It has almost all the pros of Silicon Valley with a more chilled out lifestyle. The Silicon Forest has drawn a crowd similar to Silicon Valley to its midst.

Description of Problem

Let us consider an individual in interested in starting a Restaurant with a unique idea which has not been tested before. They are not certain if it will be embraced by the masses or be a big failure. They want to try running a successful business at a smaller scale, to test the waters before investing huge amounts and opening at a large scale in heavily populous cities like San Francisco and San Jose. They hear about the silicon forest and feel as the client base might be similar it would be a good idea to consider opening a business there. They want to explore Oregon to start a business.

So in this Capstone project I would like to address this problem. We will try to find out which county in Oregon has the most venues so as to select that county as the

place of business. To explore the neighborhoods in that county, pitting them each other in the battle of neighborhoods.

This kind of analysis will help not only the people seeking to open a business in popular places but with lower property rates, it will also help people who are seeking options of places similar to Silicon Valley to relocate for a quiet life.

II] DESCRIPTION OF DATA

For the list of Counties
Data Source: https://en.wikipedia.org/wiki/List_of_counties_in_Oregon
Data Description: Will scrape the Wikipedia page with the table containing the information of all the counties of Oregon.

For Latitude and Longitudes
Data Source: Geopy
Data Description: The Wikipedia page does not contain the latitudes and longitude information. So the data was obtained by using the Geocoder class of the Geopy client.

List of Venues and Venue categories.
Data Source: Foursquare API's
Data Description: By using the Foursquare API we can get the information of the venues in the counties. Based on this this data we can further analyze the data by selecting the county with the highest number of venues. We can again use Foursquare API to get data about the venue categories in the neighborhoods of the selected county.

III] METHODOLOGY

Data Preparation

As the State of Oregon is the state of interest, I first scraped the Wikipedia page for information about counties in Oregon. I used pandas to scrape the table on Wikipedia page and convert it to a Pandas data frame.

Step 1:
To carry out the steps mentioned above we need to import and install the following libraries

1. PANDAS 2. NUMPY 3. NOMINATIM from geopy.geocoders

Step 2:
After all the required libraries have been imported and installed we scrape the Wikipedia page "https://en.wikipedia.org/wiki/List_of_counties_in_Oregon" for the relevant data. This above-mentioned page contains a table which lists all the Counties of Oregon and gives other significant information such as Population density, area, county seat, year it was established etc.
We first scrape this page and convert this table into a Pandas dataframe.
Once we have the dataframe we need to clean the dataframe to drop columns which we will not be using for our analysis. After this step our dataframe is ready.

	County	County seat[4]	Est.[4]	Population[6]
0	Baker County	Baker City	1862	16510
1	Benton County	Corvallis	1847	91320
2	Clackamas County	Oregon City	1843	404980
3	Clatsop County	Astoria	1844	38225
4	Columbia County	Saint Helens	1854	50795
5	Coos County	Coquille	1853	63190
6	Crook County	Prineville	1882	21580
7	Curry County	Gold Beach	1855	22600
8	Deschutes County	Bend	1916	176635
9	Douglas County	Roseburg	1852	110395

Step 3:
We can now use the Nominatim from geopy.geocoders library to get the exact co-ordinates for each and every county in our initial dataframe. This step is important as we will be using these co-ordinates to call a Foursquare API function to provide us with venue details, upon which our analysis is based.

]:

	County	County seat[4]	Est.[4]	Latitude	Longitude
0	Baker County	Baker City	1862	44.725964	-117.620482
1	Benton County	Corvallis	1847	44.494937	-123.406568
2	Clackamas County	Oregon City	1843	45.160882	-122.230504
3	Clatsop County	Astoria	1844	45.980728	-123.668750
4	Columbia County	Saint Helens	1854	45.928020	-123.082293
5	Coos County	Coquille	1853	43.218414	-124.109621
6	Crook County	Prineville	1882	44.146029	-120.383948
7	Curry County	Gold Beach	1855	42.426543	-124.216879
8	Deschutes County	Bend	1916	43.819484	-121.166966
9	Douglas County	Roseburg	1852	43.192211	-123.113596
10	Gilliam County	Condon	1885	45.329874	-120.220220
11	Grant County	Canyon City	1864	44.464824	-119.056528
12	Harney County	Burns	1889	43.055195	-119.024026
13	Hood River County	Hood River	1908	45.531164	-121.647559
14	Jackson County	Medford	1852	42.398015	-122.754095
15	Jefferson County	Madras	1914	44.607250	-121.246314

Data Analysis

Step 1:

To proceed with the data analysis, we will now call the Foursquare API function to provide us with Venue, Venue Category, Venue Latitude and Venue Longitude. For the Foursquare API we need to create the URL. To generate the URL we define the Foursquare Client ID, Client Secret, Limit, Radius and it takes the latitude longitude and County information from our dataframe. Once the URL is ready we can call the function to generate a new dataframe with the new data returned by the function. By analyzing this data we realized that only a few Counties in Oregon returned significant amount of venues. One such promising county was the Multnomah County.

Step 2:

Now that we have determined that Multnomah County is the best county in Oregon for our restaurant we decide to get data for the neighborhoods of Multnomah county and analyze this data further.

```
Baker County: 0
Benton County: 5
Clackamas County: 2
Clatsop County: 4
Columbia County: 0
Coos County: 2
Crook County: 0
Curry County: 0
Deschutes County: 0
Douglas County: 1
Gilliam County: 0
Grant County: 0
Harney County: 0
Hood River County: 9
Jackson County: 6
Jefferson County: 6
Josephine County: 5
Klamath County: 0
Lake County: 0
Lane County: 10
Lincoln County: 9
Linn County: 2
Malheur County: 0
Marion County: 4
Morrow County: 2
Multnomah County: 100
Polk County: 4
Sherman County: 1
Tillamook County: 4
Umatilla County: 43
Union County: 5
Wallowa County: 0
Wasco County: 0
Washington County: 38
Wheeler County: 0
Yamhill County: 46
```

Step 3:
We now have to repeat Step 1 and Step 2 to get the venue data for all the neighborhoods in Multnomah county.

Step 4:
Once we have venue data on all neighborhoods we determine these venues fall under which categories. We find that there are 260+ unique categories which include categories like trails, food trucks, parks etc. But these categories are just making our data huge without providing us with any wanted information as we are interested in the restaurants in the neighborhoods. So this step includes removing the categories which are not restaurants.

```
262 Unique Categories in Multnomah County
```

The Foursquare returned 262 unique categories. We are interested only in restaurants. So let us clean our dataframe and consider only those venues which are Restaurants.

```
[15]: df_multresto = mult_venues[mult_venues['Venue Category'].astype(str).str.contains('Restaurant')]

      df_multresto.head()
```

[15]:

	County	County Latitude	County Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Alameda	45.548631	-122.636481	Pine State Biscuits	45.558981	-122.642697	Southern / Soul Food Restaurant
2	Alameda	45.548631	-122.636481	Kargi Gogo	45.559113	-122.634038	Dumpling Restaurant
4	Alameda	45.548631	-122.636481	Bollywood Theater	45.559270	-122.644011	Indian Restaurant
8	Alameda	45.548631	-122.636481	Urdaneta	45.559136	-122.634119	Tapas Restaurant
15	Alameda	45.548631	-122.636481	Beast	45.562462	-122.634985	French Restaurant

Step 5:
Now we have a dataframe which includes all venues which are restaurants of different type. We find out there are 43 unique type of restaurants. This is better data and we are much closer to our end result.

Step 6:

For further analysis let us convert our dataframe into Pandas one hot dataframe and calculate the frequency of occurrence of each venue category.

	Neighborhoods	American Restaurant	Argentinian Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Cambodian Restaurant	Chinese Restaurant	Cuban Restaurant	...	Restaurant	Russian Restaura
1	Alameda	0	0	0	0	0	0	0	0	0	...	0	0
2	Alameda	0	0	0	0	0	0	0	0	0	...	0	0
4	Alameda	0	0	0	0	0	0	0	0	0	...	0	0
8	Alameda	0	0	0	0	0	0	0	0	0	...	0	0
15	Alameda	0	0	0	0	0	0	0	0	0	...	0	0

	Neighborhoods	American Restaurant	Argentinian Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Cambodian Restaurant	Chinese Restaurant	Cuban Restaurant	...	Restaurant	Russian Restaurar
0	Alameda	0.034483	0.034483	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.0
1	Albina	0.041667	0.041667	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.0
2	Arleta	0.086957	0.000000	0.043478	0.0	0.0	0.043478	0.0	0.086957	0.0	...	0.000000	0.0
3	Barnes Heights	0.058824	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.058824	0.0
4	Bonneville	0.500000	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.500000	0.0

Step 7:

On the basis of the frequency calculations we determine the Top 5 Most common restaurants in all neighborhoods. This will help us narrow down on possible neighborhoods for our restaurant.

```
Alameda
                                venue  freq
0                  Thai Restaurant  0.14
1                Mexican Restaurant  0.10
2  Southern / Soul Food Restaurant  0.07
3          New American Restaurant  0.07
4       Middle Eastern Restaurant  0.07


Albina
                              venue  freq
0            Thai Restaurant  0.12
1          Sushi Restaurant  0.08
2  New American Restaurant  0.08
3      American Restaurant  0.04
4        Indian Restaurant  0.04
```

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alameda	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Southern / Soul Food Restaurant	French Restaurant
1	Albina	Thai Restaurant	Sushi Restaurant	New American Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant
2	Arleta	Vietnamese Restaurant	Mexican Restaurant	Chinese Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant
3	Barnes Heights	Italian Restaurant	Mexican Restaurant	Restaurant	French Restaurant	Hawaiian Restaurant
4	Bonneville	American Restaurant	Restaurant	Dumpling Restaurant	Indian Restaurant	Hawaiian Restaurant
...	...	...	...	...	...	...
56	West Portland Park	Mexican Restaurant	American Restaurant	Sushi Restaurant	Seafood Restaurant	German Restaurant
57	Westmoreland	Italian Restaurant	Sushi Restaurant	Seafood Restaurant	Vietnamese Restaurant	American Restaurant
58	Whitwood Court	Mexican Restaurant	Thai Restaurant	Italian Restaurant	Chinese Restaurant	Falafel Restaurant
59	Willamette Heights	Italian Restaurant	Restaurant	French Restaurant	Korean Restaurant	Mediterranean Restaurant
60	Woodstock	Vietnamese Restaurant	Mexican Restaurant	Vegetarian / Vegan Restaurant	American Restaurant	Sushi Restaurant

Step 8:

As mentioned in the business problem, we are helping an individual determine which neighborhood would be best for his restaurant. The idea of restaurant is a

fusion of Thai and American cuisines. It is safe to assume that a neighborhood which has very high number of Thai restaurants would be a good neighborhood as many Thai restaurants are successfully thriving in the neighborhood. Lets in this step narrow down our dataframe by creating a new dataframe and adding only those neighborhoods which have Thai restaurants as its 1st most common venue.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alameda	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Southern / Soul Food Restaurant	French Restaurant
1	Albina	Thai Restaurant	Sushi Restaurant	New American Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant
25	Irvington	Thai Restaurant	Middle Eastern Restaurant	Southern / Soul Food Restaurant	French Restaurant	New American Restaurant
28	Kenton	Thai Restaurant	Mexican Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
30	Laurelhurst	Thai Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Modern European Restaurant	Cuban Restaurant
47	Saint Johns	Thai Restaurant	Mexican Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant
52	University Park	Thai Restaurant	Mexican Restaurant	Seafood Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant

Step 9:

In this Final step we have a dataframe with only those neighborhoods which have Thai restaurants as their most common venue. But we still have a lot of neighborhoods as probable candidates. So we can reduce them further by considering the second half of the idea. It is a fusion restaurant with spices and flavors of the Thai Cuisine with the tasty American dishes from American Cuisine. So let us further analyze our dataframe by checking if our neighborhoods have American Cuisine in their most common venues. After doing this we arrive at 3 neighborhoods with Thai and American Cuisines as their most common venues.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Albina	Thai Restaurant	Sushi Restaurant	New American Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant
28	Kenton	Thai Restaurant	Mexican Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
28	Kenton	Thai Restaurant	Mexican Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
25	Irvington	Thai Restaurant	Middle Eastern Restaurant	Southern / Soul Food Restaurant	French Restaurant	New American Restaurant

RESULTS

We arrive at 3 neighborhoods Albina, Irvington and Kenton, which satisfy both our conditions of Thai and American Cuisines.

RESULTS

```
print ('The best neighborhoods for Thai/American fusion restaurants are :')
print(np.unique(df_final['Neighborhoods']))
```

```
The best neighborhoods for Thai/American fusion restaurants are :
['Albina' 'Irvington' 'Kenton']
```

By this analysis we got an idea about which restaurants are most liked in Portland, the county seat of the Multnomah County.

We were also able to determine by this analysis that Mexican Restaurants top the list in most number of restaurants in the Multnomah County. This can be of interest to someone who is interested to open a Mexican Restaurant or give a business idea to someone who wants to open a restaurant in Multnomah County but wants a good analysis of which restaurants are most liked.

```
43 Unique Restaurant Categories in Multnomah County
```

```
18]: print (df_multresto['Venue Category'].value_counts()[
```

```
Mexican Restaurant                148
Thai Restaurant                    96
American Restaurant                95
Vietnamese Restaurant              84
Italian Restaurant                 75
Seafood Restaurant                 60
Sushi Restaurant                   46
Vegetarian / Vegan Restaurant      44
Restaurant                         39
Chinese Restaurant                 34
Name: Venue Category, dtype: int64
```

DISCUSSIONS

Data analysis is a process of inspecting, cleaning, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively. [1]

We carried out this analysis only on the basis of data gathered from Foursquare API. Foursquare is a location-based social networking service that can be accessed via both computer and mobile phones. As a mobile application, it allows people to check in venues in real time. It's important for businesses to take advantage of Foursquare because it gives a company the exposure and visibility it needs. [2]

We have not considered other factors like demographics, foot traffic, property rates, restaurant prices, if the neighborhood has predominantly businesses or residential complexes, schools, proximity to public transportation system etc. Analysis taking all these factors into consideration will yield a better result of the best neighborhood.

This analysis is a very basic one but it gives us a good idea of what businesses are operating in Portland, what kind of food people prefer and which would be good neighborhoods to start a restaurant depending on its Cuisine.

CONCLUSION

Analysis was done on data to determine the best neighborhoods for a Fusion Thai-American Restaurant in Oregon, USA. This analysis involved data collection by web scraping, Geopy, Foursquare API. Pandas dataframes were used for the analysis. The result of the analysis gave us 3 good candidate neighborhoods for the Restaurant. We were also able to determine the most loved cuisine. The analysis can be improved and made more accurate by considering more parameters rather than just venue data.

However, in conclusion our business problem was solved. We were able to determine the best neighborhood. This would be a good trial run for our customer who wants to test the waters in a city not as large, costly, populated as San Francisco but comes close with regards to demographics and has potential for tremendous growth.

References:

1. https://en.wikipedia.org/wiki/Data_analysis

2. https://www.leapfroggr.com/what-is-foursquare/#:~:text=A%20Quick%20Recap%3A-,Foursquare%20is%20a%20location%2Dbased%20social%20networking%20service%20that%20can,exposure%20and%20visibility%20it%20needs.