Capstone Project - The Battle of Neighborhoods (Week 2)

In the pursue of completing the data capstone project in IBM courser specialization for data science, other than then complete working codes and graphs, we also are needed to submit this report containing all the relevant and required fields of a good, comprehensive and complete technical documentation.

We have to use data from foursquare using the APIs by applying our knowledge of cleansing data through web, clustering it though kNN or k-means methods, and then utilizing it to provide the results that are required by the user of this project.

Following is my report for this project submission:

Project Title:

To Open a Restaurant Business in Neighborhood Of Sargodha, Pakistan

Introduction/Business Idea:

Sargodha is a not a very well known city of Pakistan, but it is very important due to its military importance especially airforce. But regardless of that, it is not counted among the top big metropolitan cities of the country.

And perhaps, this is exactly the reason to start the restaurant business in Sargodha. The big cities have established brands and merchandizes and it is really hard to sneak into the big guns. But here, even McDonalds opened its branch in year 2019!

So, the is the wake up call. As the city is gaining recognition, the bigger and strong competitors are coming to deploy their roots in this region too. This is the peak time to prosper for any new restaurant business to open in city like Sargodha otherwise it will be all too late.

The people here have good taste for food and like to have premium quality cuisines. Given the fact that there aren't many in the region, we can cluster the existing ones, apply our data cleansing techniques and then come up with a powerful yet efficient solution to this amazing opportunity in hand.

Target Audience

My target audience will mainly be any company that is interested in opening up a new facility in the region. I will provide them will valuable data predictions and comprehensive solutions my looking into the trends of the past and present, analyzing though technology and knowledge and then help them in proper implementation though the AI ladder techniques:(collect, Organize, Analyze, Infuse) in order to be successful in their business.

Data Preparation:

Scraping Sargodha Wards Table from Wikipedia:

```
]: html = wp.page("List of restaurants of Sargodha: M").html().encode("UTF-8")
   df = pd.read_html(html)[0]

   df = df[df.Borough!= "Not assigned"]
   df = df.groupby(['Food type', 'Ratings'])['Neighbourhood'].apply(list).apply(lambda x:', '.join(x)).to_frame().reset_index()
   for index, row in df.iterrows():
       if row['Neighbourhood'] == 'Not assigned':
           row['Neighbourhood'] = row['Food Type']
   column_names = ['Food Type','Ratings', 'Neighborhood']

   df.columns = column_names
   df.head()
```

First of all we need to get the data of Sargodha city from Wikipedia and then start to implement the analysis or to be specific clustering techniques onto it for cleansing and extracting meaning,

Getting Coordinates of Major Districts

Adding longitudes and latitudes

```
]: import pandas as pd
   import io
   import requests
   url="https://cocl.us/Geospatial_data"
   s=requests.get(url).content
   c=pd.read_csv(io.StringIO(s.decode('utf-8')))

   dfc = df.join(c.set_index('Food Type'), on='Food Type')
   dfc.head()
```

Then we will have the coordinates from districts.

And can further analyze the data.

Using Foursquare Location Data:

```
In [ ]: map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)
        for lat, lng, borough, neighborhood in zip(dfc['Latitude'], dfc['Longitude'], dfc['food types'], dfc['ratings']):
            label = '{}, {}'.format(neighborhood, borough)
            label = folium.Popup(label, parse_html=True)
            folium.CircleMarker(
                [lat, lng],
                radius=5,
                popup=label,
                color='blue',
                fill=True,
                fill_color='#3186cc',
                fill_opacity=0.7,
                parse_html=False).add_to(map_toronto)

        map_toronto
```

Adding my foursqaure credentials

```
In [ ]: CLIENT_ID = 'RGKM2OLTCUTH0Y4SM2YWZS4MF10GXERJPRPOMKW4W12NV5EL'
        CLIENT_SECRET = '2BXTDEIOEWBQUCCFPLJ1J43JWSWR5NFKSXLJT4PMBLBLHZ3H'
        VERSION = '20180604'
        LIMIT = 100
        print('Your credentails:')
        print('CLIENT_ID: ' + CLIENT_ID)
        print('CLIENT_SECRET:' + CLIENT_SECRET)
        radius = 500
```

Then we need to add our foursquare credentials and using the APIs we can add locations that are needed f


```
In [ ]: map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)
        for lat, lng, borough, neighborhood in zip(dfc['Latitude'], dfc['Longitude'], dfc['food types'], dfc['ratings']):
            label = '{}, {}'.format(neighborhood, borough)
            label = folium.Popup(label, parse_html=True)
            folium.CircleMarker(
                [lat, lng],
                radius=5,
                popup=label,
                color='blue',
                fill=True,
                fill_color='#3186cc',
                fill_opacity=0.7,
                parse_html=False).add_to(map_toronto)

        map_toronto
```

Adding my foursqaure credentials

```
In [ ]: CLIENT_ID = 'RGKM2OLTCUTH0Y4SM2YWZS4MF10GXERJPRPOMKW4W12NV5EL'
        CLIENT_SECRET = '2BXTDEIOEWBQUCCFPLJ1J43JWSWR5NFKSXLJT4PMBLBLHZ3H'
        VERSION = '20180604'
        LIMIT = 100
        print('Your credentails:')
        print('CLIENT_ID: ' + CLIENT_ID)
        print('CLIENT_SECRET:' + CLIENT_SECRET)
        radius = 500
```


or further data analysis.

3.Clustering the Districts:

We will divide the data into various clusters and then apply k means techniques to extract the best value of k

```
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

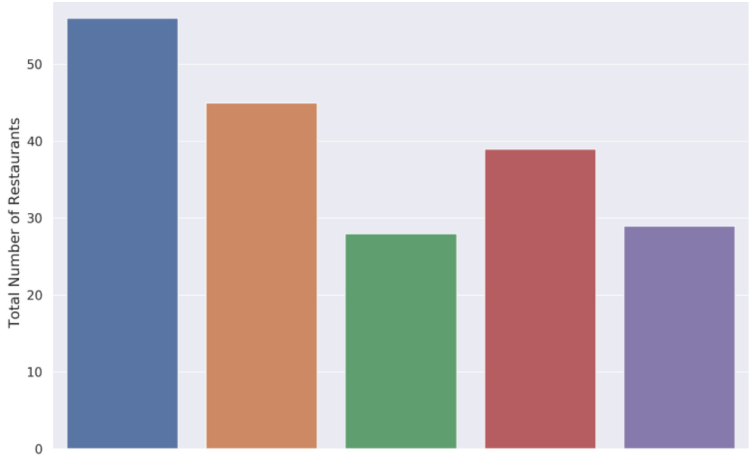
[ ]:
```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)
return row_categories_sorted.index.values[0:num_top_venues]
```

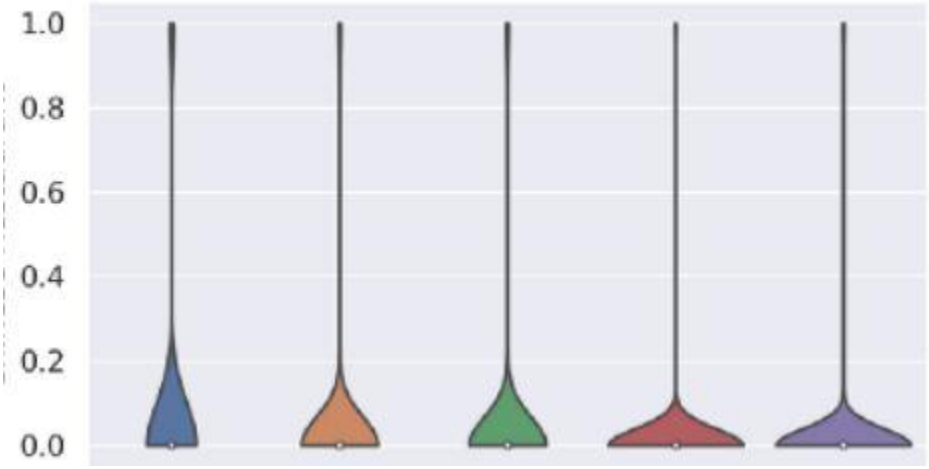
[ ]:
```
num_top_venues = 5
indicators = ['st', 'nd', 'rd']
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = toronto_grouped['Neighborhood']
for ind in np.arange(toronto_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(toronto_grouped.iloc[ind, :

neighborhoods_venues_sorted.head()
```

```
    nearby_venues.columns = [ Neighborhood , Neighborhood Latitude , Neighborhood Longitude , Ve
    return(nearby_venues)
```

4. Visualization and Data Exploration:

The visualized results will make us more clear about the situation of our data.





5. Results and Discussion:

We reached at the end of the analysis, using comprehensive data from Wikipedia and foursquare, it came out that the restaurants with traditional cuisines are mostly preferred in this region and opening of fast food will result in disaster. Therefore, we need to open a traditional restaurant giving staple food like flavor and which is more closer to their roots.

If you have any further discuss point, im always open to discussions and constructive feedback.

6. Conclusion:

Finally to conclude this project, We have got a small insight of how real life data-science projects look like. I've made use of some frequently used python libraries to scrap web-data, use Foursquare API to explore the major districts of Sargodha region and saw the results of segmentation of districts using Folium leaflet map.

It was a great ride and learned a lot.