IBM DATA SCIENCE: DATA VISUALIZATION CAPSTONE

PROMPT

A survey was conducted to gauge an audience interest in different data science topics, namely:

1. Big Data (Spark / Hadoop)
2. Data Analysis / Statistics
3. Data Journalism
4. Data Visualization
5. Deep Learning
6. Machine Learning

The participants had three options for each topic: **Very Interested**, **Somewhat interested**, and **Not interested**. **2,233**respondents completed the survey.

The survey results have been saved in a csv file and can be accessed through this link: https://cocl.us/datascience_survey_data.

If you examine the csv file, you will find that the first column represents the data science topics and the first row represents the choices for each topic.

Use the *pandas* **read_csv** method to read the csv file into a *pandas* dataframe, that looks like the following:

	Very interested	Somewhat interested	Not interested
Big Data (Spark / Hadoop)			
Data Analysis / Statistics			
Data Journalism			
Data Visualization			
Deep Learning			
Machine Learning			

In order to read the data into a dataframe like the above, one way to do that is to use the *index_col* parameter in order to load the first column as the index of the dataframe. Here is the documentation on the *pandas* **read_csv** method: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html

Once you have succeeded in creating the above dataframe, please upload a screenshot of your dataframe with the actual numbers. (**5 marks)**

```
file = 'https://cocl.us/datascience_survey_data'

df_data = pd.read_csv(file, index_col=0)

print('Data downloaded and read into a dataframe!')
```

```
Data downloaded and read into a dataframe!
```

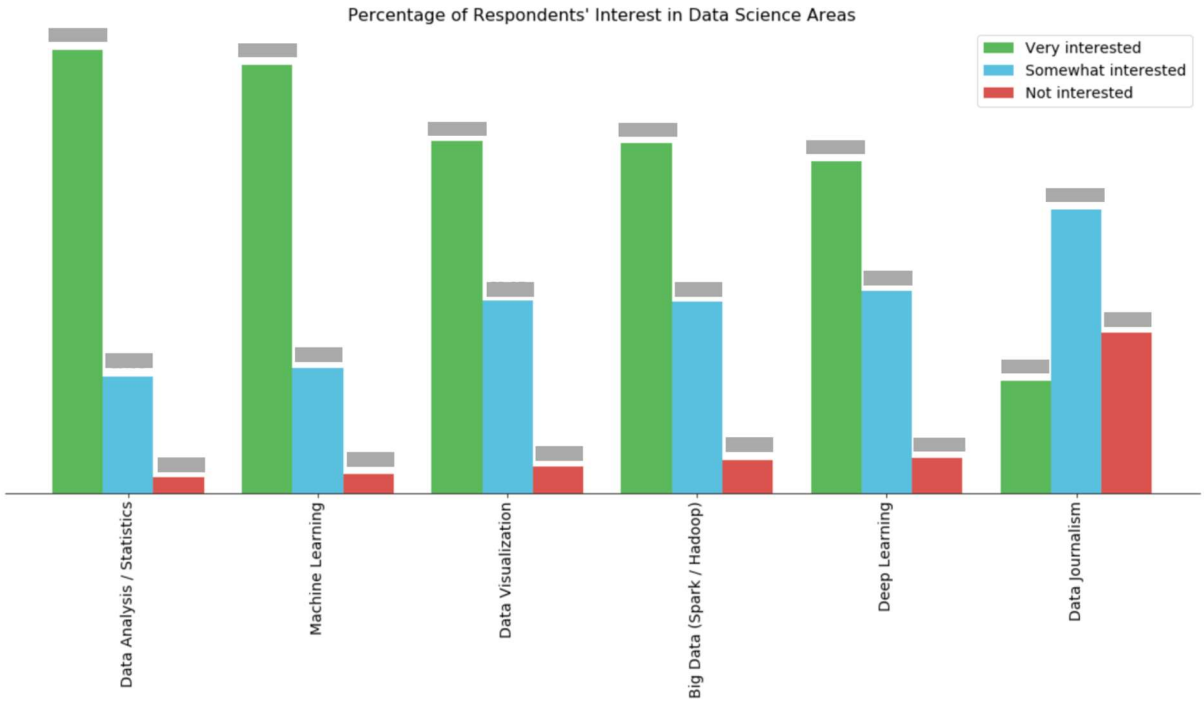
[14]:
```
df_data.head()
```

[14]:

	Very interested	Somewhat interested	Not interested
Big Data (Spark / Hadoop)	1332	729	127
Data Analysis / Statistics	1688	444	60
Data Journalism	429	1081	610
Data Visualization	1340	734	102
Deep Learning	1263	770	136

PROMPT

Use the artist layer of Matplotlib to replicate the bar chart below to visualize the **percentage** of the respondents' interest in the different data science topics surveyed.

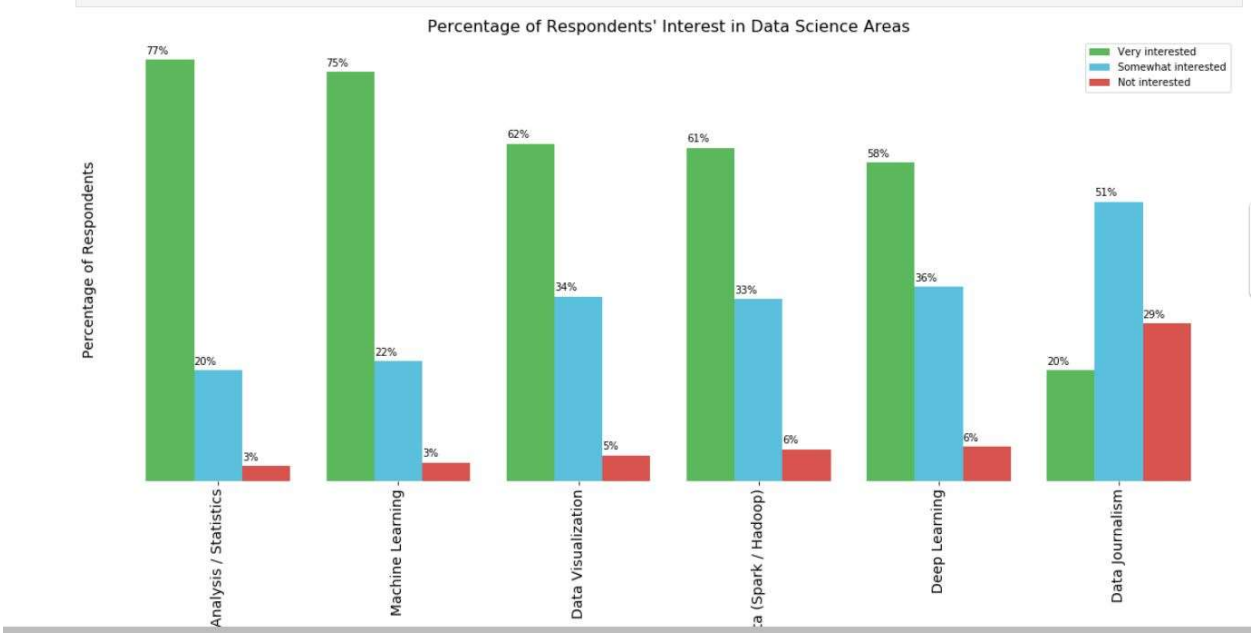


To create this bar chart, you can follow the following steps:

1. Sort the dataframe in descending order of **Very interested.**
2. Convert the numbers into percentages of the total number of respondents. Recall that **2,233** respondents completed the survey. Round percentages to 2 decimal places.
3. As for the chart:

- use a figure size of (20, 8),
- bar width of 0.8,
- use color #5cb85c for the **Very interested** bars, color #5bc0de for the **Somewhat interested** bars, and color #d9534f for the **Not interested** bars,
- use font size 14 for the bar labels, percentages, and legend,
- use font size 16 for the title, and,
- display the percentages above the bars as shown above, and remove the left, top, and right borders.

Once you are satisfied with your chart, please upload a screenshot of your plot. (**5 marks)**



PROMPT

In the final lab, we created a map with markers to explore crime rate in San Francisco, California. In this question, you are required to create a Choropleth map to visualize crime in San Francisco.

Before you are ready to start building the map, let's restructure the data so that it is in the right format for the Choropleth map. Essentially, you will need to create a dataframe that lists each neighborhood in San Francisco along with the corresponding total number of crimes.

Based on the San Francisco crime dataset, you will find that San Francisco consists of 10 main neighborhoods, namely:

1. Central,
2. Southern,
3. Bayview,
4. Mission,
5. Park,
6. Richmond,
7. Ingleside,
8. Taraval,
9. Northern, and,
10. Tenderloin.

Convert the San Francisco dataset, which you can also find here, https://cocl.us/sanfran_crime_dataset, into a *pandas*dataframe, like the one shown below, that represents the total number of crimes in each neighborhood.

	Neighborhood	Count
0	CENTRAL	
1	NORTHERN	
2	PARK	
3	SOUTHERN	
4	MISSION	
5	TENDERLOIN	
6	RICHMOND	
7	TARAVAL	
8	INGLESIDE	
9	BAYVIEW	

Once you are happy with your dataframe, upload a screenshot of your *pandas* dataframe. (**5 marks**)

[32]:

	PdDistrict	Count
0	BAYVIEW	14303
1	CENTRAL	17666
2	INGLESIDE	11594
3	MISSION	19503
4	NORTHERN	20100
5	PARK	8699
6	RICHMOND	8922
7	SOUTHERN	28445
8	TARAVAL	11325
9	TENDERLOIN	9942

PROMPT

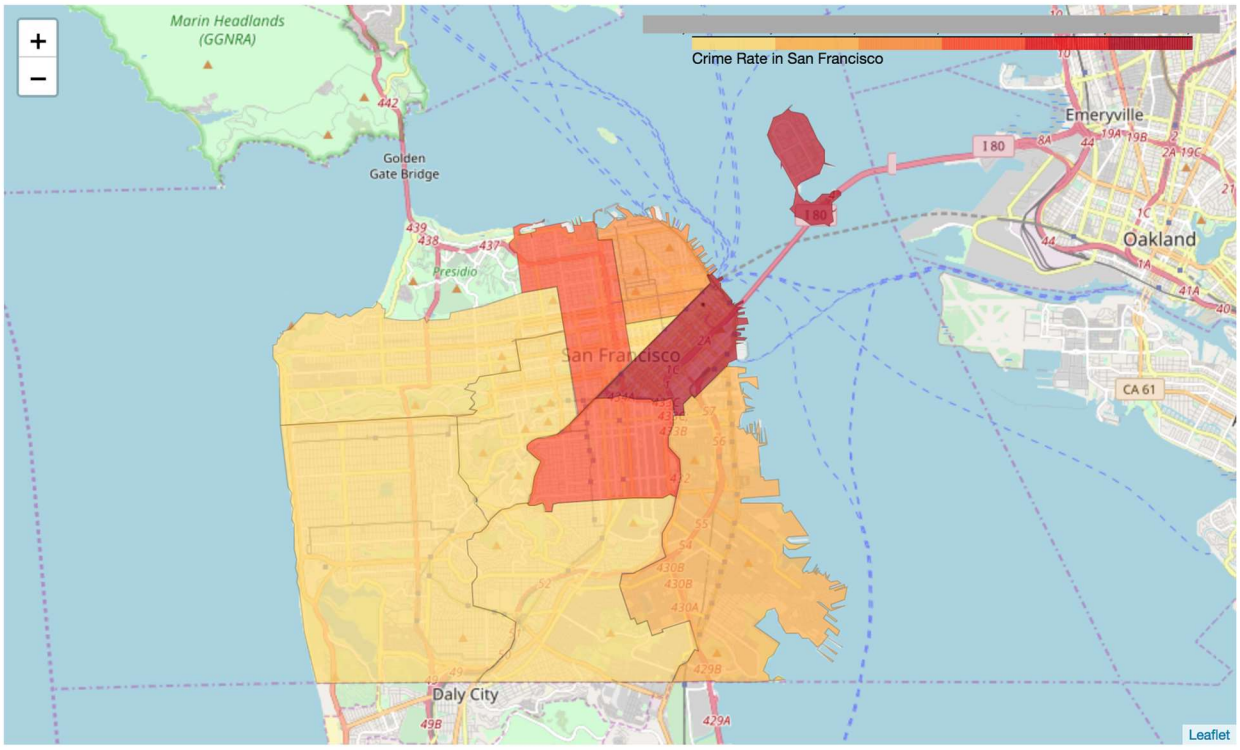
Now you should be ready to proceed with creating the Choropleth map.

As you learned in the Choropleth maps lab, you will need a GeoJSON file that marks the boundaries of the different neighborhoods in San Francisco. In order to save you the hassle of looking for the right file, I already downloaded it for you and I am making it available via this link: https://cocl.us/sanfran_geojson.

For the map, make sure that:

- it is centred around San Francisco,
- you use a zoom level of 12,
- you use fill_color = 'YlOrRd',
- you define fill_opacity = 0.7,
- you define line_opacity=0.2, and,
- you define a legend and use the **default threshold scale**.

If you follow the lab on Choropleth maps and use the GeoJSON correctly, you should be able to generate the following map:



Once you are ready to submit your map, please upload a screenshot of your Choropleth map. (**5 marks)**

```
# display map
sf_map
```

]: