Data Visualization Assignment - IBM DS

May 28, 2020

[14]:
```
#Task 1
import numpy as np
import pandas as pd
```

[15]:
```
df_data = pd.read_csv('https://cocl.us/datascience_survey_data', index_col=0)
df_data.head()
```

[15]:

	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

[21]:
```
%matplotlib inline

import matplotlib as mpl
import matplotlib.pyplot as plt

#Sort the dataframe in descending order of Very interested
df_data=df_data.sort_values(by=["Very interested"], ascending=False)
df_data

#Convert the numbers into percentages
df_data = df_data.div(df_data.sum(1), axis=0)
df_data.head()
```

[21]:

	Very interested	Somewhat interested	\
Data Analysis / Statistics	0.770073	0.202555	
Machine Learning	0.747248	0.218807	

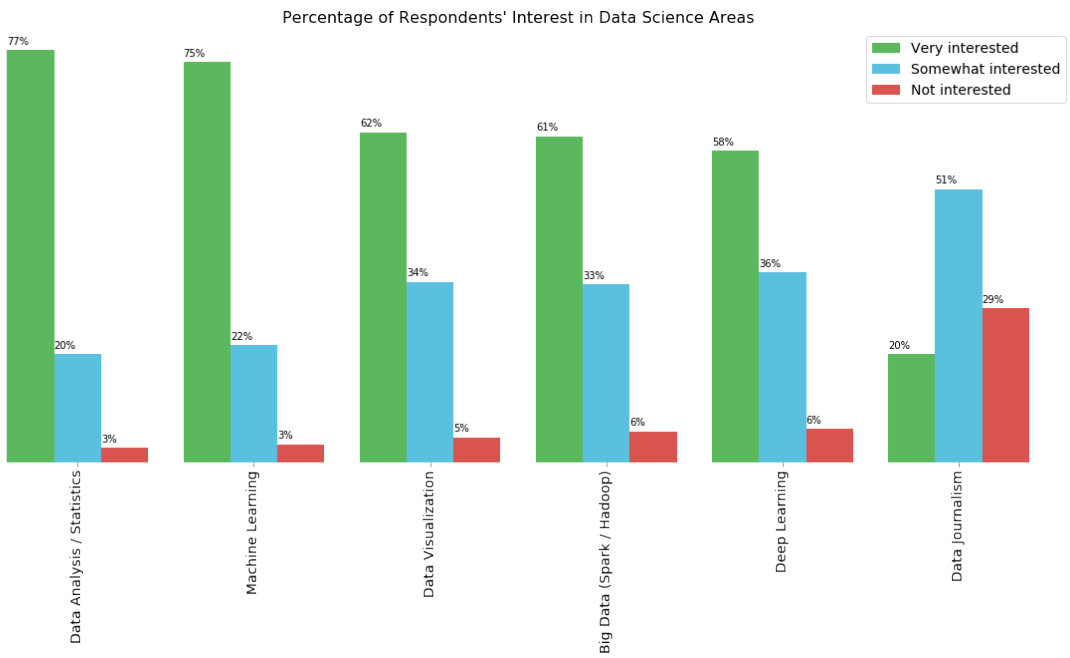
```
Data Visualization                0.615809            0.337316
Big Data (Spark / Hadoop)         0.608775            0.333181
Deep Learning                     0.582296            0.355002

                              Not interested
Data Analysis / Statistics          0.027372
Machine Learning                    0.033945
Data Visualization                  0.046875
Big Data (Spark / Hadoop)           0.058044
Deep Learning                       0.062702
```

[25]:
```
#Draw the barplot
colors = ['#5cb85c', '#5bc0de', '#d9534f']
ax = df_data.plot(kind = 'bar', figsize = (20, 8), width = 0.8, color = colors)
plt.legend(labels = df_data.columns, fontsize = 14)
plt.title("Percentage of Respondents' Interest in Data Science Areas",fontsize=␣
 ↪16)

#Remove the left, top, and right borders
plt.xticks(fontsize = 14)
for spine in plt.gca().spines.values():
    spine.set_visible(False)
plt.yticks([])

#Add annotations
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0%}'.format(height), (x, y + height + 0.01))
```



[29]:
```
#Task 2
df_sf = pd.read_csv('https://cocl.us/sanfran_crime_dataset', index_col=0)
df_sf.head()
```

[29]:

IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Resolution	Address	X	Y
120058272	WEAPON LAWS	POSS OF PROHIBITED WEAPON	Friday	01/29/2016 12:00:00 AM	11:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
120058272	WEAPON LAWS	FIREARM, LOADED, IN VEHICLE, POSSESSION OR USE	Friday	01/29/2016 12:00:00 AM	11:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
141059263	WARRANTS	WARRANT ARREST	Monday	04/25/2016 12:00:00 AM	14:59	BAYVIEW	ARREST, BOOKED	KEITH ST / SHAFTER AV	-122.388856	37.729981
160013662	NON-CRIMINAL	LOST PROPERTY	Tuesday	01/05/2016 12:00:00 AM	23:50	TENDERLOIN	NONE	JONES ST / OFARRELL ST	-122.412971	37.785788
160002740	NON-CRIMINAL	LOST PROPERTY	Friday	01/01/2016 12:00:00 AM	00:30	MISSION				

```
160002740             NONE    16TH ST / MISSION ST -122.419672  37.765050

                                          Location            PdId
IncidntNum
120058272    (37.775420706711, -122.403404791479)  12005827212120
120058272    (37.775420706711, -122.403404791479)  12005827212168
141059263   (37.7299809672996, -122.388856204292)  14105926363010
160013662   (37.7857883766888, -122.412970537591)  16001366271000
160002740   (37.7650501214668, -122.419671780296)  16000274071000
```

[51]:
```
df_neighbor = df_sf.groupby("PdDistrict", as_index= False).count()
df_neighbor.head()
new_df = df_neighbor[["PdDistrict","Category"]]
sf_df = new_df.rename(columns = {"PdDistrict":"Neighborhood", "Category":␣
 ↪"Count"})
sf_df.head(10)
```

[51]:

	Neighborhood	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

[36]:
```
!conda install -c conda-forge folium=0.5.0 --yes
import folium
# San Francisco latitude and longitude values
latitude = 37.77
longitude = -122.42

sf_map = folium.Map(location = [latitude, longitude], zoom_start = 12)
sf_map
```

```
Collecting package metadata (current_repodata.json): done
Solving environment: failed with initial frozen solve. Retrying with flexible
solve.
Collecting package metadata (repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
  current version: 4.8.2
```

```
  latest version: 4.8.3

Please update conda by running

    $ conda update -n base -c defaults conda



## Package Plan ##

  environment location: /opt/anaconda3

  added / updated specs:
    - folium=0.5.0


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    altair-4.1.0               |             py_1         614 KB  conda-forge
    branca-0.4.1               |             py_0          26 KB  conda-forge
    certifi-2019.11.28         |           py37_0         148 KB  conda-forge
    conda-4.8.3                |   py37hc8dfbb8_1         3.0 MB  conda-forge
    folium-0.5.0               |             py_0          45 KB  conda-forge
    python_abi-3.7             |          1_cp37m           4 KB  conda-forge
    vincent-0.4.4              |             py_1          28 KB  conda-forge
    ------------------------------------------------------------
                                           Total:         3.9 MB

The following NEW packages will be INSTALLED:

  altair             conda-forge/noarch::altair-4.1.0-py_1
  branca             conda-forge/noarch::branca-0.4.1-py_0
  folium             conda-forge/noarch::folium-0.5.0-py_0
  python_abi         conda-forge/osx-64::python_abi-3.7-1_cp37m
  vincent            conda-forge/noarch::vincent-0.4.4-py_1

The following packages will be UPDATED:

  conda                       pkgs/main::conda-4.8.2-py37_0 --> conda-
forge::conda-4.8.3-py37hc8dfbb8_1

The following packages will be SUPERSEDED by a higher-priority channel:

  certifi                                        pkgs/main --> conda-forge
```

```
Downloading and Extracting Packages
folium-0.5.0         | 45 KB     | ##################################### | 100%
python_abi-3.7       | 4 KB      | ##################################### | 100%
vincent-0.4.4        | 28 KB     | ##################################### | 100%
altair-4.1.0         | 614 KB    | ##################################### | 100%
conda-4.8.3          | 3.0 MB    | ##################################### | 100%
branca-0.4.1         | 26 KB     | ##################################### | 100%
certifi-2019.11.28   | 148 KB    | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

[36]:
```
<folium.folium.Map at 0x117681dd0>
```

[55]:
```
#Download sf geojson file
!wget --quiet https://cocl.us/sanfran_geojson -O sanf.json

sf_geo = r'sanf.json'
threshold_scale = np.linspace(sf_df['Count'].min(),
                              sf_df['Count'].max(),
                              6, dtype=int)

threshold_scale = threshold_scale.tolist() # change the numpy array to a list
threshold_scale[-1] = threshold_scale[-1] + 1

sf_map.choropleth(
    geo_data = sf_geo,
    data = sf_df,
    columns = ['Neighborhood', 'Count'],
    key_on = 'feature.properties.DISTRICT',
    threshold_scale=threshold_scale,
    fill_color = 'YlOrRd',
    fill_opacity = 0.7,
    line_opacity = 0.2,
    legend_name = 'Crime rate in San Fran'
)

sf_map
```

```
/bin/sh: wget: command not found
```

```
␣
↪---------------------------------------------------------------------------

FileNotFoundError                         Traceback (most recent call␣
↪last)
```

```
<ipython-input-55-e2c49129ff92> in <module>
     19     fill_opacity = 0.7,
     20     line_opacity = 0.2,
---> 21     legend_name = 'Crime rate in San Fran'
     22 )
     23


/opt/anaconda3/lib/python3.7/site-packages/folium/folium.py in␣
↪choropleth(self, geo_data, data, columns, key_on, threshold_scale, fill_color,␣
↪fill_opacity, line_color, line_weight, line_opacity, name, legend_name,␣
↪topojson, reset, smooth_factor, highlight)
    325                 style_function=style_function,
    326                 smooth_factor=smooth_factor,
--> 327                 highlight_function=highlight_function if highlight␣
↪else None)
    328
    329         self.add_child(geo_json)


/opt/anaconda3/lib/python3.7/site-packages/folium/features.py in␣
↪__init__(self, data, style_function, name, overlay, control, smooth_factor,␣
↪highlight_function)
    479                 self.data = json.loads(data)
    480             else:  # This is a filename
--> 481                 with open(data) as f:
    482                     self.data = json.loads(f.read())
    483         elif data.__class__.__name__ in ['GeoDataFrame',␣
↪'GeoSeries']:


FileNotFoundError: [Errno 2] No such file or directory: 'sanf.json'
```

[ ]:

[ ]: