DSC550_Paulovici_Exercise_3_2

March 29, 2020

# Week 3: File: DSC550_Paulovici_Exercise_3_2.py (.ipynb) Name: Kevin Paulovici Date: 3/29/2020 Course: DSC 550 Data Mining (2205-1) Assignment: 3.2 Exercise: Graph Analysis

## Assignment Tasks You can create a new analysis scenario or you can use the tutorials you completed this week. A) Display the same analysis using 3 different charts (ex. A bar chart, a line chart and a pie chart) B) Use appropriate, complete, professional labeling. C) Rank your charts from most effective to least effective. D) Write a 300-word paper justifying your ranking.

### Data I'll be working with Corona Virus data for NY counties. Data was retrieved from https://coronavirus.health.ny.gov/county-county-breakdown-positive-cases for daily updates.

```
In [1]: # Start by importing the csv file into a DataFrame
        import pandas as pd

        data = pd.read_csv("corona_virus_NY2.csv")

        # preview data
        data.head(5)
```

```
Out[1]:         County  3/16/2020  3/17/2020  3/18/2020  3/19/2020  3/20/2020  \
        0       Albany       12.0       23.0       36.0       43.0       61.0
        1     Allegany        2.0        2.0        2.0        2.0        2.0
        2       Broome        1.0        1.0        1.0        2.0        2.0
        3  Cattaraugus        NaN        NaN        NaN        NaN        NaN
        4       Cayuga        NaN        NaN        NaN        NaN        NaN

           3/21/2020  3/22/2020  3/23/2020  3/24/2020  3/25/2020  3/26/2020  \
        0       88.0      123.0      127.0      146.0      152.0      171.0
        1        2.0        2.0        2.0        2.0        2.0        2.0
        2        2.0        3.0        7.0        7.0       11.0       16.0
        3        NaN        NaN        NaN        NaN        NaN        NaN
        4        NaN        NaN        2.0        2.0        2.0        2.0

           3/27/2020  3/28/2020  3/29/2020
        0      187.0      195.0        205
        1        2.0        2.0          6
        2       18.0       23.0         29
        3        NaN        1.0          4
        4        2.0        2.0          2
```

```
In [2]: # I'll be using the latest data row (date) so I'll make a subset of data to graph
        latest_data = data.filter(["County", "3/29/2020"])
        latest_data = latest_data.sort_values("3/29/2020", ascending=False)
        latest_data
```

```
Out[2]:            County  3/29/2020
        29  New York City      33768
        54    Westchester       8519
        27         Nassau       6445
        46        Suffolk       5023
        39       Rockland       2209
        33         Orange       1247
        13           Erie        358
        12       Dutchess        320
        25         Monroe        219
        0          Albany        205
        31       Onondaga        152
        50         Ulster        146
        37         Putnam        144
        40       Saratoga        102
        47       Sullivan         88
        41    Schenectady         76
        49       Tompkins         52
        38     Rensselaer         39
        28        Niagara         38
        2          Broome         29
        30         Oneida         26
        24        Madison         24
        9        Columbia         23
        51         Warren         18
        32        Ontario         18
        45        Steuben         17
        6         Chemung         15
        7        Chenango         15
        8         Clinton         13
        53          Wayne         12
        44   St. Lawrence         12
        20       Herkimer         10
        23     Livingston         10
        36         Otsego         10
        17        Genesee          9
        11       Delaware          8
        55        Wyoming          8
        35         Oswego          8
        21      Jefferson          7
        52     Washington          7
        18         Greene          7
        1        Allegany          6
```

```
10        Cortland          6
26      Montgomery          6
15        Franklin          6
42       Schoharie          5
5       Chautauqua          5
48           Tioga          4
3      Cattaraugus          4
14           Essex          4
34         Orleans          3
4           Cayuga          2
22           Lewis          2
19        Hamilton          2
16          Fulton          1
43        Schuyler          1
```

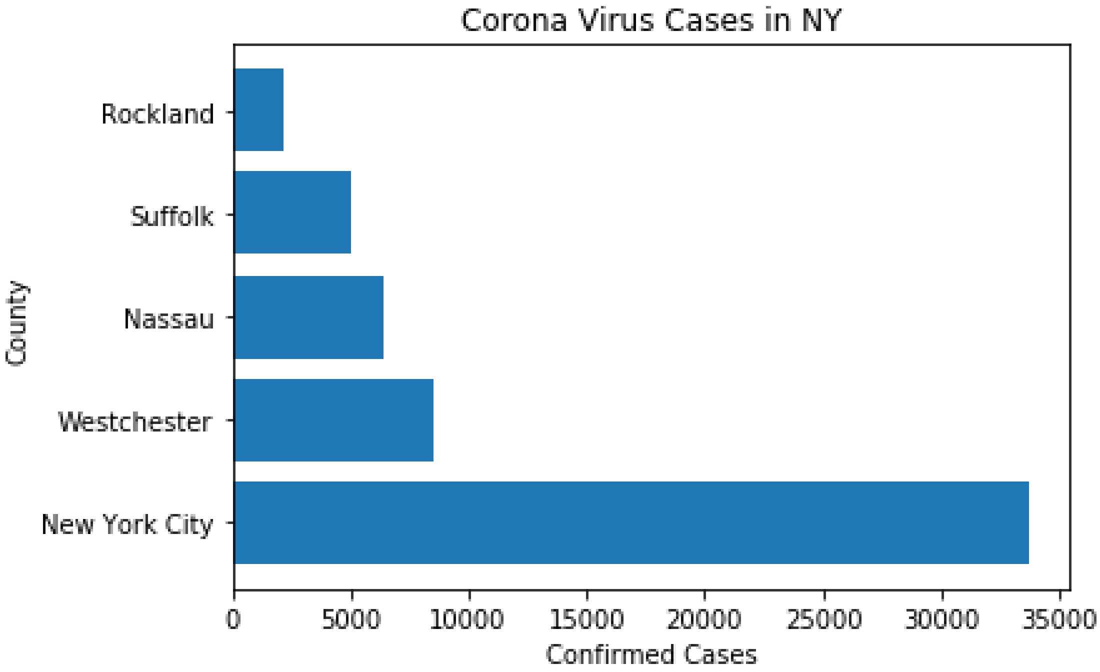
```
In [3]: # We don't need to plot all the counties, it would be too chaotic, so lets select the
        top_5 = latest_data.loc[latest_data["3/29/2020"] > 2000]
        top_5
```

```
Out[3]:            County  3/29/2020
        29  New York City      33768
        54    Westchester       8519
        27         Nassau       6445
        46        Suffolk       5023
        39       Rockland       2209
```

```
In [4]: import matplotlib.pyplot as plt
        import numpy as np
```

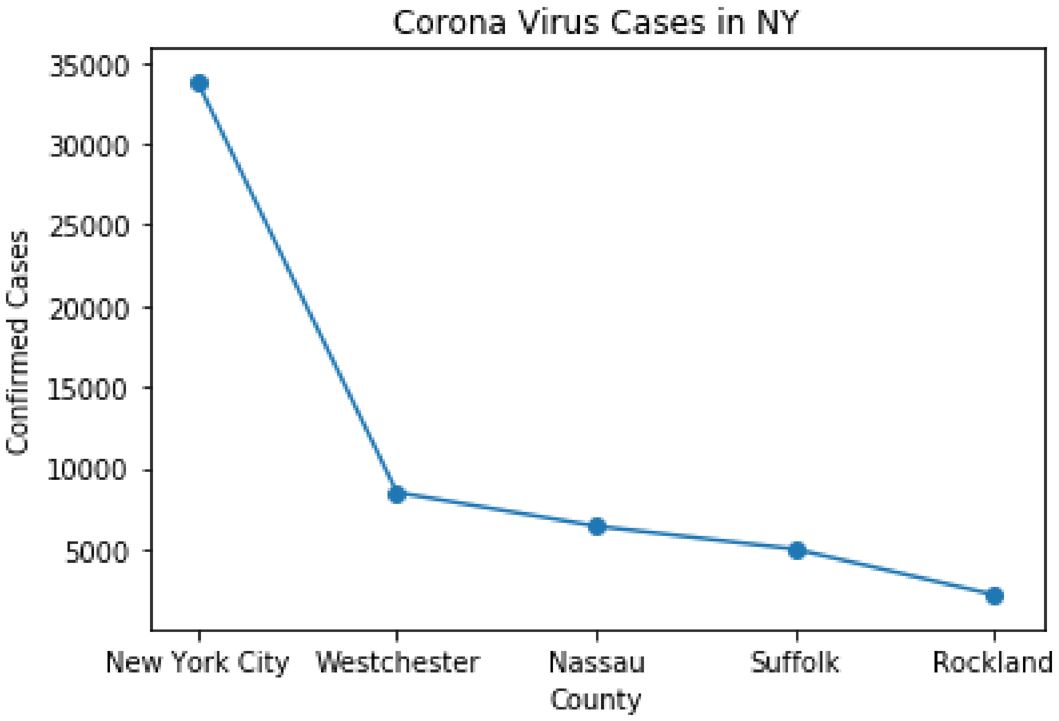
### Bar chart

```
In [5]: # get countries list
        counties = top_5["County"].values

        # get values of countries list
        val = top_5["3/29/2020"].values

        plt.barh(counties, val)
        plt.ylabel("County")
        plt.xlabel("Confirmed Cases")
        plt.title("Corona Virus Cases in NY")
        plt.show()
```



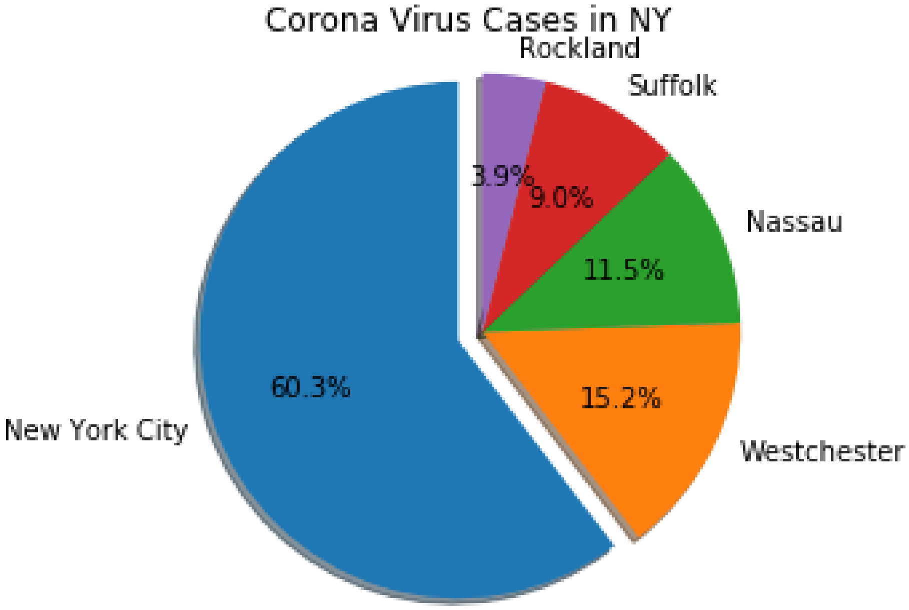
### Line chart

```
In [6]: plt.plot(counties, val)
        plt.scatter(counties, val)
        plt.xlabel("County")
        plt.ylabel("Confirmed Cases")
        plt.title("Corona Virus Cases in NY")
        plt.show()
```



### Pie chart

```
In [7]: explode = (0.1, 0, 0, 0, 0)
        plt.pie(val, explode=explode, labels=counties, autopct='%1.1f%%', shadow=True, startang
        plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
        plt.title("Corona Virus Cases in NY")
        plt.show()
```



### Summary Parts A and B are completed as shown above. For Part C, I would rank these from most effective to least effective as the following: bar chart, pie chart, and the line/scatter. For Part D - see summary below. Visualization is a key component of the data science process. Being able to effectively display analytical results in in a comprehendible way to is a skill and can take some trial-and-error until the right format of the plot is selected. The details, such as color and labels can also play an important role in the process. For this assignment, a bar chart, a scatter/line plot, and a pie chart are used to demonstrate this. Before discussing how I ranked these various plots, it is important to cover what data was used and what is the intent of it. I chose to use confirmed corona cases by counties in NY. Additionally, for plotting purpose I down selected to the most recent data and only focused on the top 5 counties. For my subset of data, I wanted to demonstrate the vast different seen by these counties; this heavily influenced how I ranked these plots. However, some consideration was given if additional counties were included. I ranked the bar chart as the most effective, followed by the pie chart, and the scatter/line plot I found to be least effective. The horizontal bar was most effective to me because it provided a clear and easy comparison with a rough estimation of confirmed cases. While the pie chart does not directly provided the count of cases, it gives a percentage view comparison to the other counties. Again, this is clear and direct. However, by expanding the counties, the pie chart would quickly lose its effectiveness because of too much data, even with the inclusion of a legend it would be chaotic to pin point the counties. In that scenario I would rank the pie chart last. The scatter/line chart was ranked last because I did not think it was as effective as the bar chart. However it does clearly label all points and give an estimation of cases. If this plot was not sorted it could easily become a chaotic by increasing and decreasing cases. Expanding the counties would extend the plot horizontal, being able to scroll vertical (bar chart) is more effective. This exercised demonstrated multiple plots can be used for a given set of data. However, they are not equally effective at demonstrating the intended purposes. Using a variety of plots to fit the data is an import task for a data scientist to reach the intended audience.