CS 590V: Data Visualization and Exploration

Homework 2: March 23rd , 2017

Kriti Shrivastava (Student ID: 31041848)

Data pre-processing:

File Path: /code/main.py

The python code does the following-

1. Facebook:
    - Uses the linear interpolation technique to fill in the missing values.
    - Output file: /data/Facebook_output.csv
2. Wine:
    - Merges the red and white wine datasets into a single csv file.
    - Adds a column "Type" with the values "Red" or "White" in the combined file.
    - Output file: /data/ WineQuality_output.csv

Data Visualization:

Libraries used: ***crossfilter*** *and* ***dc.js***

HTML File Path: /code/index.html
Javascript File Path: /code/script/main.js
CSS File Path: /code/css/style.css

Features implemented:

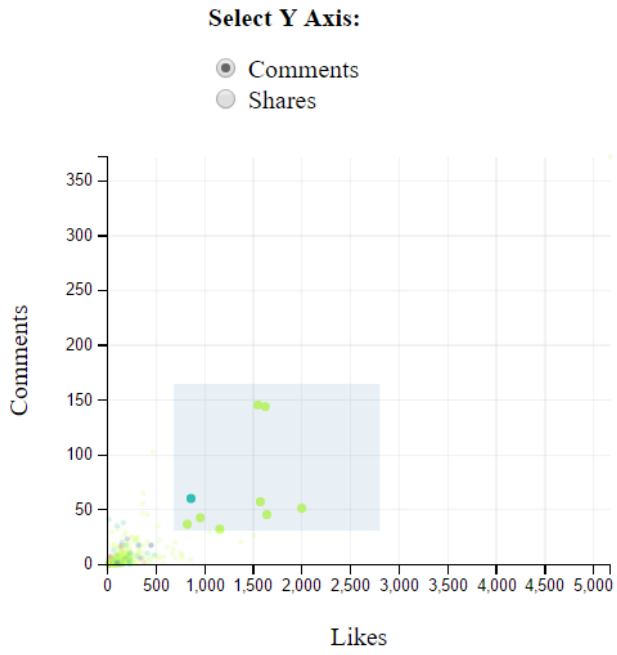
1. Visualizations for both the datasets (Facebook and Wine). Data is read based on the user selection of the dataset via the webpage.
2. All graphs interact with each other.
3. All graphs support selection of multiple records.
4. All graphs (except Scatterplot) support probing.
5. All graphs provide options to change parameters for Facebook data.
6. Parameter change for Pie-Chart and HeatMap changes the binding to a visual attribute (color).
7. Changing parameter for Pie-Chart also changes the color binding for Scatterplot.
8. All graphs have description of their initial state (without filtering and selection). Dynamic change of description on change of parameters or filters is not supported.
9. Reset Link: Link to reset all the filters.

Graphs:

- Facebook: Scatterplot, line chart, histogram, heatmap, pie-chart and bar graph
- Wine: Scatterplot, line chart, pie-chart and bar graph.

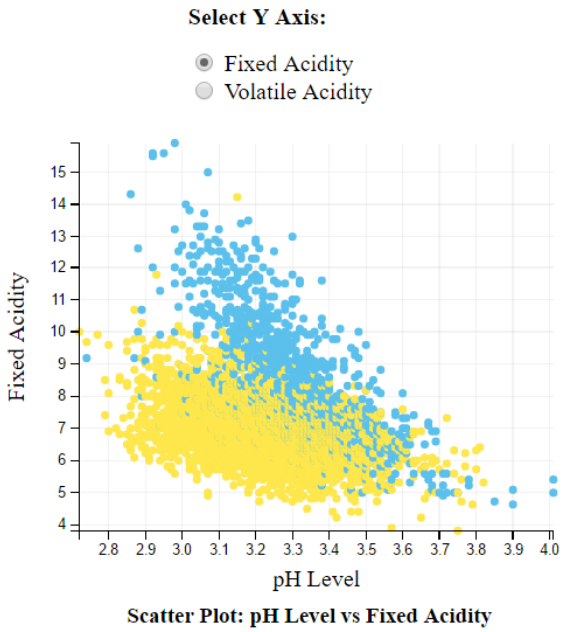
Graph 1: Scatter Plot

1. **Facebook:**
    a. X axis: Number of Likes
    b. Y axis: User can select one of the following two options-
        i. Number of Comments (default)
        ii. Number of Shares
    c. Inference:
        i. Number of likes for a post is significantly more than the number of comments/shares.
        ii. In general, the number of likes/comments/shares for photos and status as compared to videos and links.
        iii. The posts of type photos have significant outliers.



Scatter Plot: Number of Likes vs Number of Comments for a post

2. **Wine dataset:**
    a. X axis: pH level
    b. Y axis: User can select one of the following two options-
        i. Fixed Acidity (default)
        ii. Volatile Acidity
    c. Inference:
        i. Two distinct clusters for white(yellow) and red(blue) wine for acidity vs pH level are visible.
        ii. In general, the acidity in white wine is lesser than the red wine.



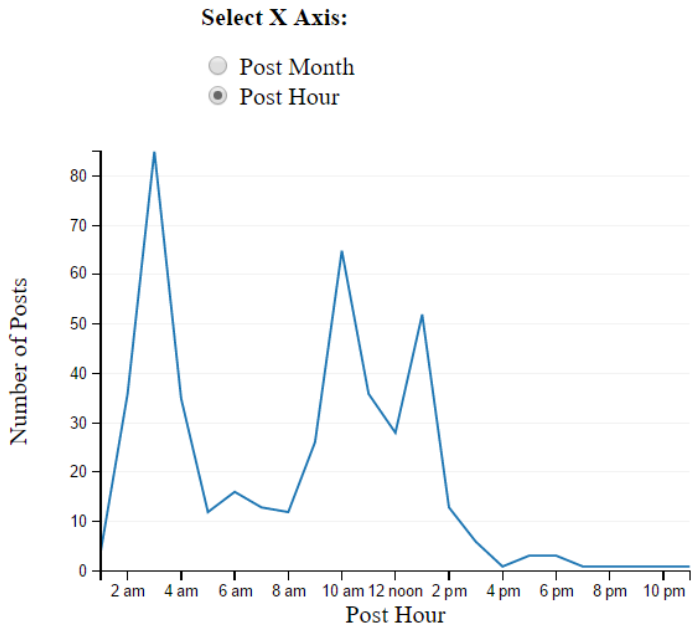
Scatter Plot: pH Level vs Fixed Acidity

Features:

1. Brush to select region
2. Dynamic y axis selection
3. Color change based on the dimension selected for pie-chart.

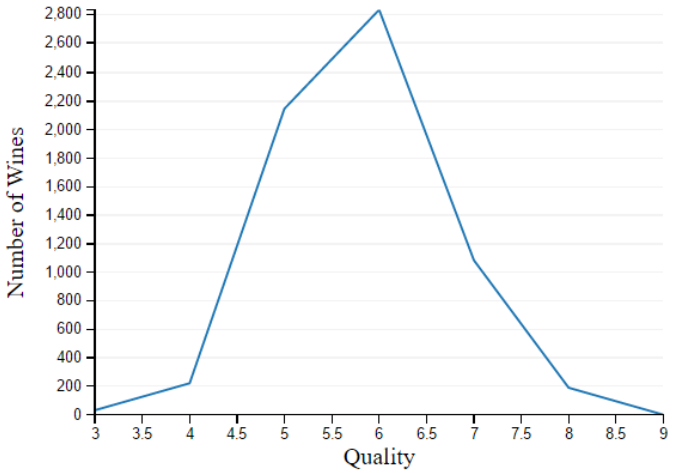
Graph 2: Line Chart

1. **Facebook dataset:**
   a. X axis: User can select one of the following two options-
      i. Post Month (default)
      ii. Post Hour
   b. Y axis: Number of posts
   c. Inference:
      i. Number of the posts put up on the page, steadily increased in the first quarter of the year.
      ii. The number of posts per month almost doubled by the end of the year.
      iii. Month of October saw maximum posts.
      iv. Most posts are put up late night (after mid-night) or mid-day (10 am- 1pm) and it significantly decreases during the evening hours.



Line Chart: Post Hour vs Number of posts

2. **Wine dataset:**
   a. X axis: Wine quality
   b. Y axis: Number of Wines
   c. Inference:
      i. Most of the wines considered in this study have quality between 4-8 with maximum wines (almost 45% of all wines) of the quality 6.
      ii. Very few wines have quality more than 8 and only 5 wines out of all 6497 wines are of the quality 9.



Line Chart: Wine Quality vs Number of Wines

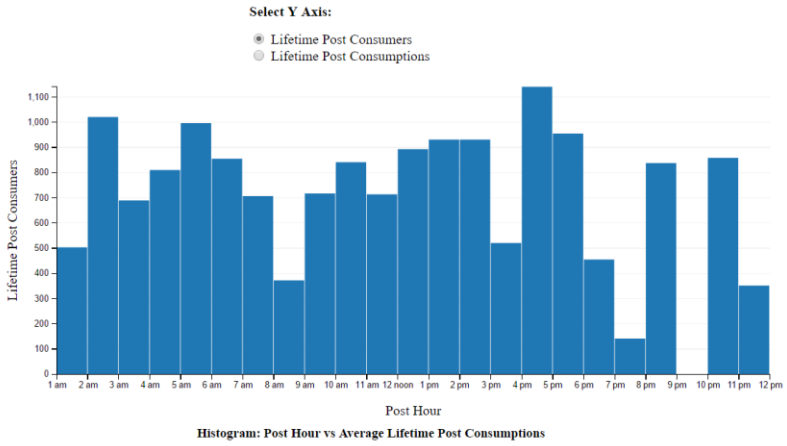
Features:

1. Zoom to select region.

2. Probing.
3. Elastic x and y axis tick marks (changes with zoom).
4. Dynamic x axis selection for Facebook data.

Graph 3: Histogram

1. **Facebook dataset:**
   a. X axis: Post Hour Interval
   b. Y axis: User can select one of the following two options-
      i. Lifetime Post Consumers (default)
      ii. Lifetime Post Consumption
   c. Inference:
      i. While the average number of consumers of a post varies significantly over the day, the post consumption average remains almost constant throughout the day and then shoots up around 5-6 pm in the evening.
      ii. Both, the number of consumers and the post consumption is relatively higher around 4-6 pm, making it a suitable time for putting up the post.
      iii. Whereas, 8-9 am in the morning and 7-8 pm (maybe, travelling time to and from work) in the evening have lesser average consumers as well as post consumption making it the least favorable time for putting up the post.



Histogram: Post Hour vs Average Lifetime Post Consumptions

Description: While the average number of consumers of a post varies significantly over the day, the post consumption average remains almost constant throughout the day and then shoots up around 5-6 pm in the evening. Both, the number of consumers and the post consumption is relatively higher around 4-6 pm, making it a suitable time for putting up the post. Whereas, 8-9 am in the morning and 7-8 pm (maybe, travelling time to and from work) in the evening have lesser average consumers as well as post consumption making it the least favorable time for putting up the post.

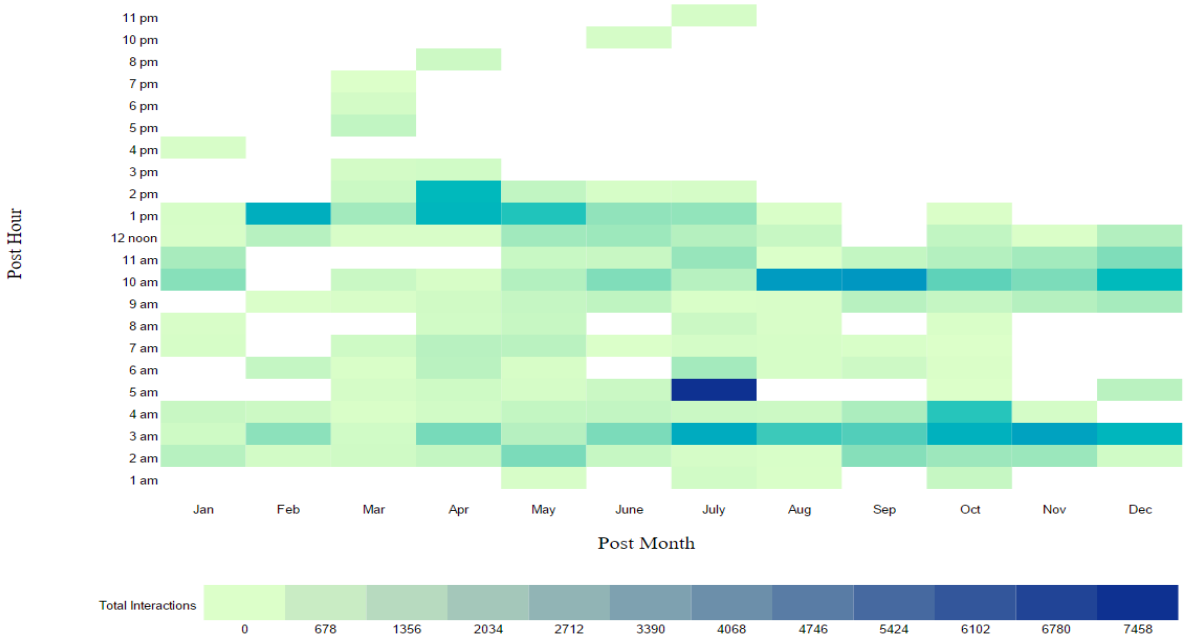
Features:

1. Brushing to select multiple records.
2. Probing.
3. Dynamic y axis selection for Facebook data.

Graph 4: Heatmap

1. **Facebook dataset:**
   a. X axis: Post Month

b. Y axis: Post Hour
c. Dimension for color: User can select one of the following two options-
    i. Total Interactions (default)
    ii. Total Reach
d. Inference:
    i. The absence of color (white-space) in the upper half of the graph suggests that no posts were posted on the page during that time of the day.
    ii. It can be seen that no posts have been put up after 1 pm till mid-night since August. This can be interesting as the histogram (Post Hour vs Average Lifetime Post Consumers/Consumptions) above, shows that the posts put up in the evening (between 4-7 pm) see higher consumers and post consumption.
    iii. Total Interactions- During the first half of the year most interactions on the post were around 1pm in the afternoon whereas the second half of the year saw comparatively more interactions on the posts at 3am and 10am.
    iv. Month of July saw exceptionally high post interactions at 5 am in the morning suggesting the occurrence of a special event.
    v. Total Reach- It can be observed that some months (like May, September) show relatively lesser post reach than others.
    vi. Second half of the year, post reach is higher around 3 am in the morning with November 3 am having the highest post reach ever seen on the page.
    vii. 1 pm saw more reach during the first half of the year as compared to the second half.
    viii. The graph helps in determining which hour is more suitable for putting up a post to increase the reach of the post.



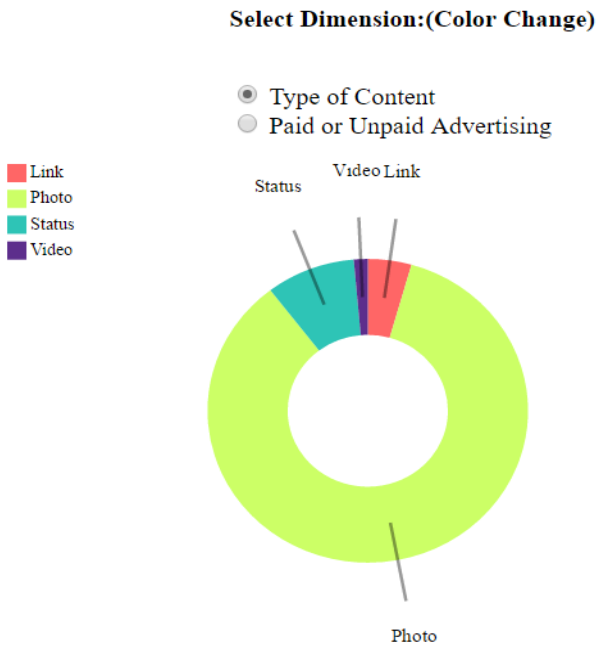
HeatMap: Total Interactions for Month of Post vs Hour of Post

Features:

1. Selection of multiple records (both on x axis and y axis).
2. Probing.
3. Color Legend.
4. Dynamic color dimension selection for Facebook data.

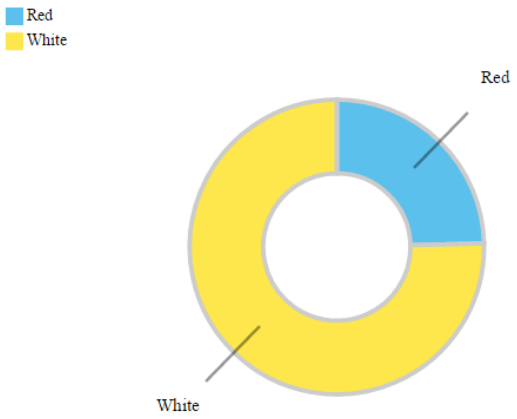
Graph 5 : Pie Chart

1. **Facebook dataset:**
   a. Dimension: User can select one of the following two options-
      i. Type of post content (photo/video/status/link)
      ii. Type of advertising (paid/unpaid)
   b. *Inference:*
      i. Most of the posts are of type photos and least of type videos.
      ii. For majority of the posts, Facebook was not paid for advertising.



Pie Chart: Type of Post Content

2. **Wine dataset:**
   a. Dimension:
   b. *Inference:*
      i. Most of the posts are of type photos and least of type links.
      ii. For majority of the posts, Facebook was not paid for advertising.



Pie Chart: Type of Wine

Features:

1. Option to select dimension.
2. Color change with change in dimension.
3. Option to select multiple records.
4. Probing.

Collaborator:

Nidhi Mundra

References:

[1] https://dc-js.github.io/dc.js/
[2] https://dc-js.github.io/dc.js/docs/stock.html
[3] https://dc-js.github.io/dc.js/examples/heat.html
[4] http://crossfilter.github.io/crossfilter/
[5] http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.interpolate.html