Project 1 – Bar Charts & Histograms

Shreyash Paratkar

YouTube link to project video

February 16, 2020

—

CSE 564 - Visualization

—

Klauss Mueller

Data

This visualization project uses data from 3 different sources described below,

1.) US Pollution Data:
This dataset contains information of every day from 2000 – 2016 on metrics of four major gaseous pollutants NO_2, O_3, SO_2 and CO such as mean, max, hour in which maximum amount of the pollutant was observed by the observation center and the Air Quality Index corresponding to that pollutant. This data has been collected from counties across the United States where the observation centers are located. For the purpose of this project, I have chosen average values of these metrics from 2010 to 2016.

2.) Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2018 (CC-EST2018-ALLDATA)
This dataset gives us the population data and its demographics from each county of the United States. This data is clubbed with the previously filtered data to obtain the number of people directly affected by air pollution in the counties seen before.

3.) County Health Rankings
This dataset gives us a deep insight into some health statistics in each county over the years from 2010 to 2017. Joining this dataset with our previously filtered data was done with the aim of correlating people's health complains with air quality. Hence fields such as 'average physically unhealthy days per month', 'Years of Potential Life Lost Rate', 'Percentage of Adults reporting fair/poor health' have been extracted from this dataset.

The data has been cleaned and joined in python using the 'Pandas' library. (Python script used has been attached with this submission)

Based on the richness of the data distributions, following attributes were finally chosen to be visualized.

Categorical	**Numerical**
State	Average physically unhealthy days per month
County	Mean CO value (Parts per million)
Year	Maximum CO value /day (Parts per million)
Hour of maximum NO2 /day	CO Air Quality Index
Hour of maximum O3 /day	Years of Potential Life Lost
Hour of maximum SO2 /day	Total Population Affected
Hour of maximum CO /day	Mean O3 value (Parts per million)
	Maximum O3 value /day (Parts per million)
	O3 Air Quality Index

Visualization

The data is being loaded using the `d3.csv("VisData.csv")` command and then being using by executing the `data.filter` call on each data item.
For categorical variables, the column of each variable is extracted, and a map is created to store the count of unique value of the variable.
As professor mentioned in the class that for bar charts, 12 bars are generally enough for a bar chart showing unique value, the map built earlier has been sorted and sliced to a size of 12 in order to show the **top 12** values with highest frequencies(count) for each of the categorical variables having more than 12 unique values.

Drawing the bar chart

An 'svg' container is created to hold the chart. Graphic 'g' elements are then appended to the chart.

Element | **Commands**

X-Axis is defined by first defining a horizontal scale using scaleBand and then d3.axisBottom is used to generate a bottom oriented axis.

```
d3.scaleBand()
    .range([0, svgWidth])
    .domain(xList)
    .padding(0.35);

chart.append('g')
    .attr('transform', 'translate(0,'+svgHeight+')')
    .call(d3.axisBottom(xScale))
```

Y-Axis is defined by first defining a horizontal scale using scaleLinear and then d3.axisLeft is used to generate a left oriented axis.

```
d3.scaleLinear()
    .range([svgHeight, 0])
    .domain([0, d3.max(yList)+d3.max(yList)/10])

chart.append('g')
    .call(d3.axisLeft(yScale))
```

A grid is added to the graph by calling the .scale function of d3.axisLeft()
The enter selection represent the elements to be added

```
d3.axisLeft()
    .scale(yScale)

chart.selectAll()
    .data(my_sample)
    .enter()
    .append('g')
```

For every data entered, a 'rect' is appended to the selection returned above and its position, height and width are specified

```
barGroups.append('rect')
    .attr('class', 'bar')
    .attr('x', function(g) {
        return xScale(g.selected_attr);  //Bar width of 20
plus 1 for padding
    })
    .attr('y', function (g) {
        return yScale(g.count);
    })
    .attr('height', function (g) {
        return svgHeight - yScale(g.count);
    })
    .attr('width', xScale.bandwidth())
```

To make the bars higher, wider and display values on top of them, the on function is called to capture 'mouseover' events. The changes in 'mouseover' are reverted back in 'mouseout'

```
.attr('x', function (a) {
    return xScale(a.selected_attr) - 5;
})
.attr('width', xScale.bandwidth() + 10)
.attr('y', function (g) {
    return yScale(g.count+d3.mean(yList)/20);
})
.attr('height', function (g) {
    return svgHeight - yScale(g.count+d3.mean(yList)/20);
});

barGroups.append("text")
```

Labels and title are then added to the graph

```
svg.append('text')
    .attr('class', 'label')
    .attr('x', - (svgHeight / 2) - svgMargin)
    .attr('y', svgMargin / 2.4)
    .attr('transform', 'rotate(-90)')
    .attr('text-anchor', 'middle')
    .text('Count')
```

Higher bars are given lighter colors giving using a color scale

```
d3.scaleLinear()
    .domain([yMin, yMax])
    .range([d3.rgb(color).darker(),
d3.rgb(color).brighter()])
```

Drawing a Histogram

Not many things change from the bar chart while drawing a histogram. Some of the differences are

Element | **Commands**

Data for the histogram

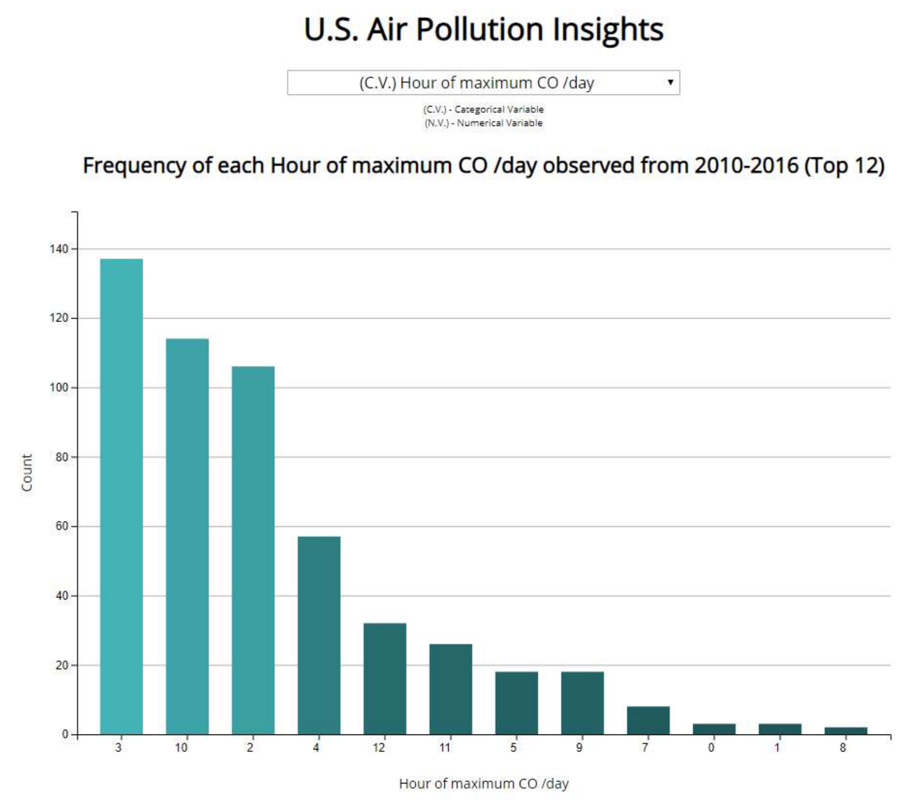
```
histData = d3.histogram()
    .thresholds(xScale.ticks(nBins))
    (yList)
```

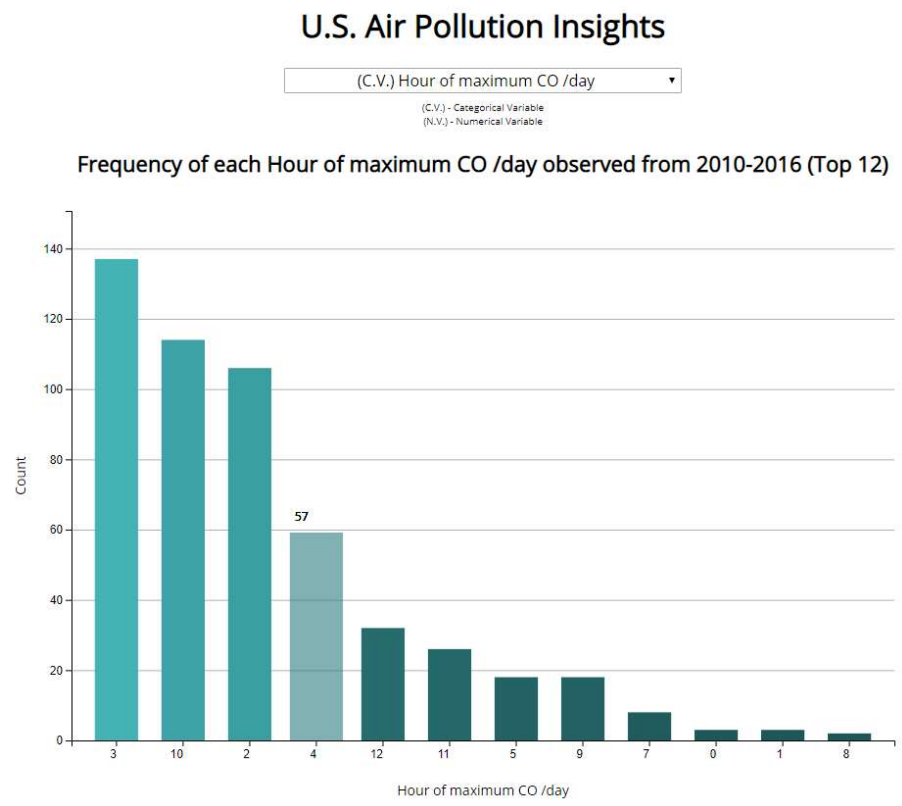
To change bin size of the histogram, d3.drag() is used. In the drag event, the change in x-coordinate of the cursor is identified using d3.event.dx. This value increases while going right and decreases while going left. Hence it is subtracted from the number of bins used to create the histogram data. The chart is then updated with the new data.

```
var dragHandler = d3.drag()
    .on("drag", function () {
        nBins -= (d3.event.dx/10);
        const histData = d3.histogram()
            .thresholds(xScale.ticks(nBins))
            (yList);
        chart.selectAll('g').remove();
        updateChart(histData);

    });
dragHandler(svg);
```

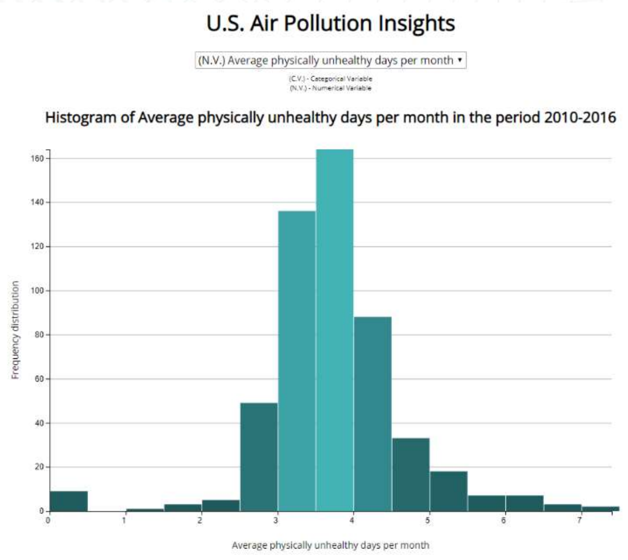
Bar Chart Example

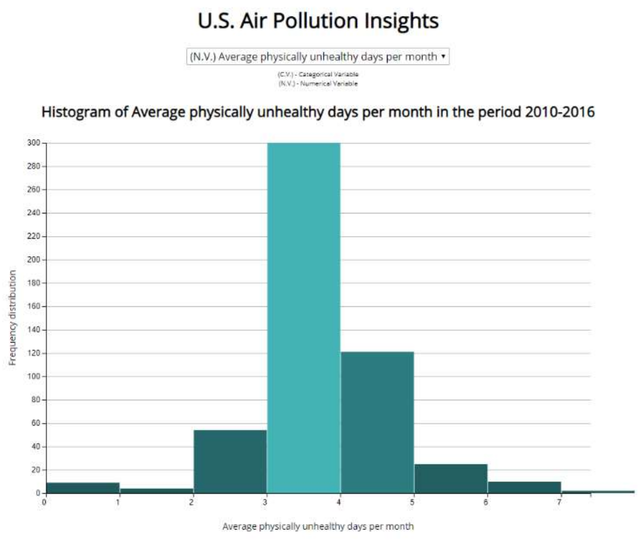

U.S. Air Pollution Insights
(C.V.) Hour of maximum CO /day
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Frequency of each Hour of maximum CO /day observed from 2010-2016 (Top 12)
Count
Hour of maximum CO /day
3
10
2
4
12
11
5
9
7
0
1
8


U.S. Air Pollution Insights
(C.V.) Hour of maximum CO /day
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Frequency of each Hour of maximum CO /day observed from 2010-2016 (Top 12)
57
Count
Hour of maximum CO /day

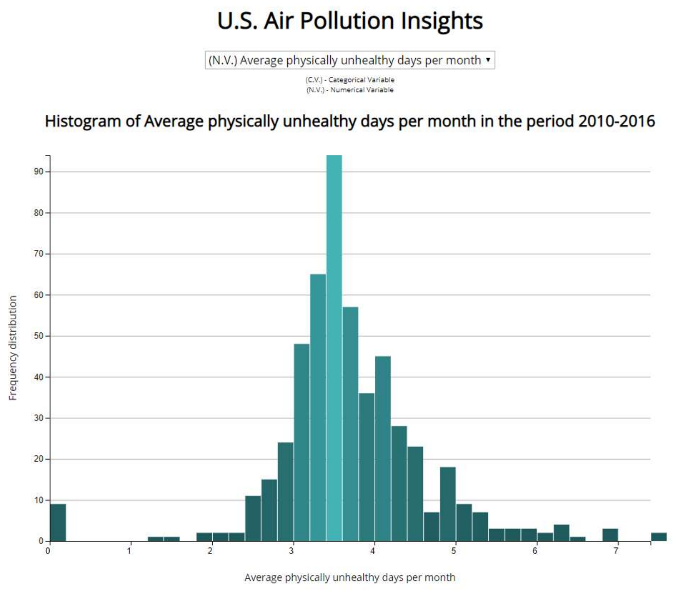
The bar gets wider, higher and decreased opacity on hover with the value being displayed on top of it

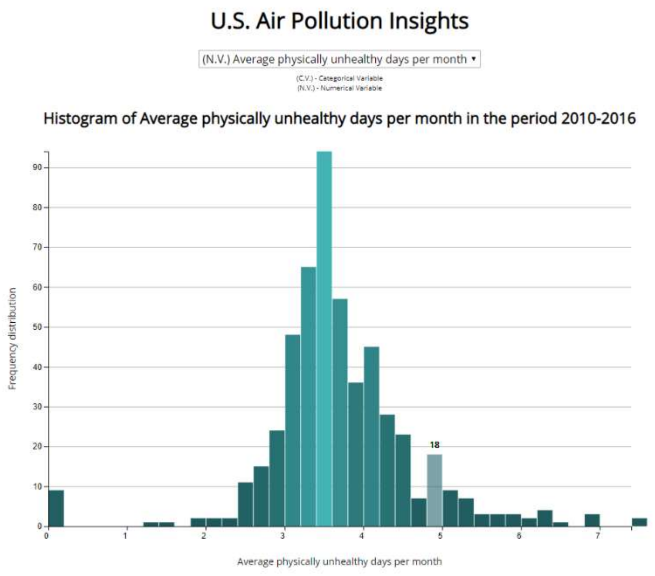
Histogram Example


U.S. Air Pollution Insights
(N.V.) Average physically unhealthy days per month
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Histogram of Average physically unhealthy days per month in the period 2010-2016
Frequency distribution
Average physically unhealthy days per month


U.S. Air Pollution Insights
(N.V.) Average physically unhealthy days per month
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Histogram of Average physically unhealthy days per month in the period 2010-2016
Frequency distribution
Average physically unhealthy days per month

The bar gets decreased opacity on hover with the value being displayed on top of it, dragging mouse right increases histogram bin size and dragging the mouse left decreases the bin size


U.S. Air Pollution Insights
(N.V.) Average physically unhealthy days per month
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Histogram of Average physically unhealthy days per month in the period 2010-2016
Frequency distribution
Average physically unhealthy days per month


U.S. Air Pollution Insights
(N.V.) Average physically unhealthy days per month
(C.V.) - Categorical Variable
(N.V.) - Numerical Variable
Histogram of Average physically unhealthy days per month in the period 2010-2016
18
Frequency distribution
Average physically unhealthy days per month

Link to YouTube video

https://youtu.be/XgDTSwWqafw

References

- https://www.tutorialsteacher.com/d3js/create-bar-chart-using-d3js
- https://jsfiddle.net/x4go8as3/2/
- https://blog.risingstack.com/d3-js-tutorial-bar-charts-with-javascript/
- https://d3js.org/
- https://github.com/d3/d3-drag