**Sri R Sankaranarayanan – Week 7 & 8 - DSC640 – Data Visualization**

1. Power BI -Scatter Plot

Chart, scatter chart

Description automatically generated

**Bubble Plot**

**Chart, bubble chart

Description automatically generated**

**Density Plot**

**Chart, histogram

Description automatically generated**

**Dot Plot**

**Chart, scatter chart

Description automatically generated**

1. **Python**
2. 
3. *# Import libraries*
4. **import** pandas **as** pd
5. **import** matplotlib.pyplot **as** plt
6. **import** numpy **as** np
7. **import** plotly.plotly **as** py
8. **import** cufflinks **as** cf
9. **import** seaborn **as** sns
10. In [58]:
11. 
12. *# Read world population data*
13. dirData **=** 'ex4-2/'
14. f\_crime **=** 'crimerates-by-state-2005.csv'
15. ​
16. dir\_crime **=** dirData**+**f\_crime
17. crime **=** pd.read\_csv(dir\_crime)
18. crime\_state **=** crime[crime['state']**!=**'United States']
19. ​
20. print(crime\_state.head())
21. state murder forcible\_rape robbery aggravated\_assault burglary \
22. 1 Alabama 8.2 34.3 141.4 247.8 953.8
23. 2 Alaska 4.8 81.1 80.9 465.1 622.5
24. 3 Arizona 7.5 33.8 144.4 327.4 948.4
25. 4 Arkansas 6.7 42.9 91.1 386.8 1084.6
26. 5 California 6.9 26.0 176.1 317.3 693.3
27. larceny\_theft motor\_vehicle\_theft population
28. 1 2650.0 288.3 4545049
29. 2 2599.1 391.0 669488
30. 3 2965.2 924.4 5974834
31. 4 2711.2 262.1 2776221
32. 5 1916.5 712.8 35795255

### 1. Scatter plot

1. In [3]:
2. 
3. *# Create a scatter plot showing correlation between murder and robbery*
4. ​
5. plt.scatter(x**=**crime['murder'], y**=**crime['robbery'],alpha**=**0.5)
6. plt.title('Correlation between robbery and murder')
7. plt.xlabel('Murder')
8. plt.ylabel('Robbery')
9. plt.show()
10. Chart, scatter chart

    Description automatically generated

### 2. Bubble Chart

1. In [47]:
2. 
3. *# For this excercise I used the API service from plotly (https://plot.ly/)*
4. *# Cufflinks binds plotly to pandas dataframes in IPython notebook*
5. *# I replaced my username and API for security*
6. ​
7. py.plotly.tools.set\_credentials\_file(username**=**'username', api\_key**=**'apikey')
8. cf.set\_config\_file(offline**=False**, world\_readable**=True**, theme**=**'pearl')
9. ​
10. crime\_state.iplot(kind**=**'bubble', x**=**'murder', y**=**'robbery', size**=**'population', text**=**'state',
11. xTitle**=**'Murder', yTitle**=**'Robbery',
12. filename**=**'Murder vs Robbery by Population')
13. Out[47]:

### 3. Density plot

1. In [68]:
2. 
3. *# Plotting distribution of population*
4. ​
5. sns.distplot(crime\_state['population'], hist **=** **False**, kde **=** **True**,
6. kde\_kws **=** {'shade': **True**, 'linewidth': 3})
7. c:\users\anirban\appdata\local\programs\python\python36-32\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning:
8. Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.
9. Out[68]:
10. <matplotlib.axes.\_subplots.AxesSubplot at 0x14aa0970>
11. A picture containing icon

    Description automatically generated

**3. R – Scatter, Bubble and Density Plots.**

You need to submit 3 scatterplots, 3 bubble charts and 3 density plot charts using Tableau or PowerBI, Python and R using the data below (or your own datasets). You can also submit using D3. You can choose which library to use in Python or R, documentation is provided to help you decide and as you start to play around in the libraries, you will decide which you prefer.

In [38]:



library('magrittr')

*# Set data paths*

dirdata **<-** 'ex4-2/'

file1 **<-** 'crimerates-by-state-2005.csv'

file2 **<-** 'life-expectancy.csv'

file3 **<-** 'states\_detail.xlsx'

​

*# Load crime rate data*

*# Creating additional variable state\_cont to convert the unique state variable from categorical to continuous*

crimerate **<-** read.csv2(paste(dirdata,file1,sep**=**''), header**=**TRUE, sep**=**',', fill**=**TRUE, stringsAsFactors**=**FALSE) **%>%**

dplyr**::**mutate(murder **=** as.numeric(murder),

forcible\_rape **=** as.numeric(forcible\_rape),

robbery **=** as.numeric(robbery),

aggravated\_assault **=** as.numeric(aggravated\_assault),

burglary **=** as.numeric(burglary),

larceny\_theft **=** as.numeric(larceny\_theft),

motor\_vehicle\_theft **=** as.numeric(motor\_vehicle\_theft),

population **=** as.integer(population),

total\_crime **=** murder**+**

forcible\_rape**+**

robbery**+**

aggravated\_assault**+**

burglary**+**

larceny\_theft**+**

motor\_vehicle\_theft,

state\_cont **=** rank(state, ties.method **=** 'first'))

​

*# For the additional data, I used US State to region mapping*

*# I marked the mid west states and rest of USA in the region column in the states\_detail.xlsx file*

*# The file is uploaded in the same github repository*

state\_detail **<-** xlsx**::**read.xlsx2(paste(dirdata,file3,sep**=**''),sheetIndex **=** 1, stringsAsFactors **=** FALSE)

​

*# Remove USA as a state and add region*

crimerate\_states **<-** crimerate **%>%**

dplyr**::**left\_join(state\_detail, by**=**'state') **%>%**

dplyr**::**filter(state **!=**'United States') **%>%**

dplyr**::**mutate(midwest **=** as.integer(midwest))

​

*# Replace NA*

crimerate\_states[is.na(crimerate\_states)] **<-** 0

​

*# Load life expectancy data*

lifeexp **<-** read.csv2(paste(dirdata,file2,sep**=**''), header**=**TRUE, sep**=**',', fill**=**TRUE, stringsAsFactors**=**FALSE)

​

head(crimerate\_states)

head(lifeexp)

| **state** | **murder** | **forcible\_rape** | | **robbery** | | **aggravated\_assault** | **burglary** | | **larceny\_theft** | **motor\_vehicle\_theft** | **population** | **total\_crime** | **state\_cont** | **name\_caps** | **abbr** | **midwest** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alabama | 8.2 | 34.3 | | 141.4 | | 247.8 | 953.8 | | 2650.0 | 288.3 | 4545049 | 4323.8 | 1 | ALABAMA | AL | 0 |
| Alaska | 4.8 | 81.1 | | 80.9 | | 465.1 | 622.5 | | 2599.1 | 391.0 | 669488 | 4244.5 | 2 | ALASKA | AK | 0 |
| Arizona | 7.5 | 33.8 | | 144.4 | | 327.4 | 948.4 | | 2965.2 | 924.4 | 5974834 | 5351.1 | 3 | ARIZONA | AZ | 0 |
| Arkansas | 6.7 | 42.9 | | 91.1 | | 386.8 | 1084.6 | | 2711.2 | 262.1 | 2776221 | 4585.4 | 4 | ARKANSAS | AR | 0 |
| California | 6.9 | 26.0 | | 176.1 | | 317.3 | 693.3 | | 1916.5 | 712.8 | 35795255 | 3848.9 | 5 | CALIFORNIA | CA | 0 |
| Colorado | 3.7 | 43.4 | | 84.6 | | 264.7 | 744.8 | | 2735.2 | 559.5 | 4660780 | 4435.9 | 6 | COLORADO | CO | 0 |
| **country** | | | **year** | | **expectancy** | | |
| Afghanistan | | | 2008 | | 42 | | |
| Albania | | | 2008 | | 73 | | |
| Algeria | | | 2008 | | 71 | | |
| Angola | | | 2008 | | 46 | | |
| Antigua and Barbuda | | | 2008 | | 74 | | |
| Argentina | | | 2008 | | 76 | | |

### Plot 1: Create scatter plot

In [32]:



*# Format graph size*

options(repr.plot.width **=** 5, repr.plot.height **=** 5)

​

*# Plot*

ggplot2**::**ggplot(data **=** crimerate\_states) **+**

ggplot2**::**aes(x **=** murder, y **=** robbery) **+**

ggplot2**::**geom\_point(color **=** "#0c4c8a") **+**

ggplot2**::**labs(title **=** "Correlation between Robbery and Murder",

x **=** "Murder",

y **=** "Robbery") **+**

ggplot2**::**theme\_grey()

**Chart, scatter chart

Description automatically generated**

### Plot 2: Create bubble plot

In [44]:



*# Format graph size*

options(repr.plot.width **=** 7, repr.plot.height **=** 3)

​

*# Plot*

ggplot2**::**ggplot(data **=** crimerate) **+**

ggplot2**::**aes(x **=** murder, y **=** robbery, color **=** total\_crime, size **=** population) **+**

ggplot2**::**geom\_point() **+**

ggplot2**::**labs(title **=** "Relationship between Murder, Robbery, Total Crime and Population") **+**

ggplot2**::**theme\_classic()

**Chart, scatter chart

Description automatically generated**

### Plot 3: Create density plot

In [46]:



*# For this I want to compare total crime rate of mid west states to the rest of the states*

crimerate\_compare **<-** crimerate\_states[,**-**1]

rownames(crimerate\_compare) **<-** crimerate\_states[,1]

​

ggplot2**::**ggplot(crimerate\_compare) **+**

ggplot2**::**aes(total\_crime, fill**=**as.factor(crimerate\_compare**$**midwest)) **+**

ggplot2**::**geom\_density(alpha **=** 0.3) **+**

ggplot2**::**labs(x**=**'Total Crime', y**=**'Density') **+**

ggplot2**::**theme\_classic() **+**

ggplot2**::**guides(fill**=**ggplot2**::**guide\_legend(title**=**"Mid West flag"))

**Chart, histogram

Description automatically generated**

**End of Code**