Data Exploration

Shivani Pal

May 9, 2023

Problem 1

This is a census data set on the incomes (>50K or <=50K) of many adults with 15 different types of attributes. You can download the data from here [link]. The downloadable file already comes with the corresponding names of the attributes. *Note:* Perform appropriate missing data handling procedures if needed.

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?
Ans. The purpose of this data set is to study the incomes of adult individuals and learn about the multiple factors that influence their earnings. This data set can be used to predictive if an person earns more or less than $50,000 based on the given attributes and analyse what factors determine the income. The kinds of data used in this dataset are nominal data like workclass, Marital status, race, gender, ordinal data like education level. The data is primarily categorical in nature, with some numerical data. Data set has no null values, hence no handling of null is required.

R/Python Code

Using R/Python, show code and plots that answers the following questions:

1. What is the distribution of attribute `income` among different education levels? Does it show that highly educated adults have higher incomes? Use proper data handling for the `education` attribute.
   Ans. We notice from the graph that the highly educated people have more chances of having higher salaries than others. Example for Doctorate level, the percentage of people earning more than 50K income is larger. More than 50 Percent of people who did Masters, Doctorate and Prof-school are earning more than 50K income.

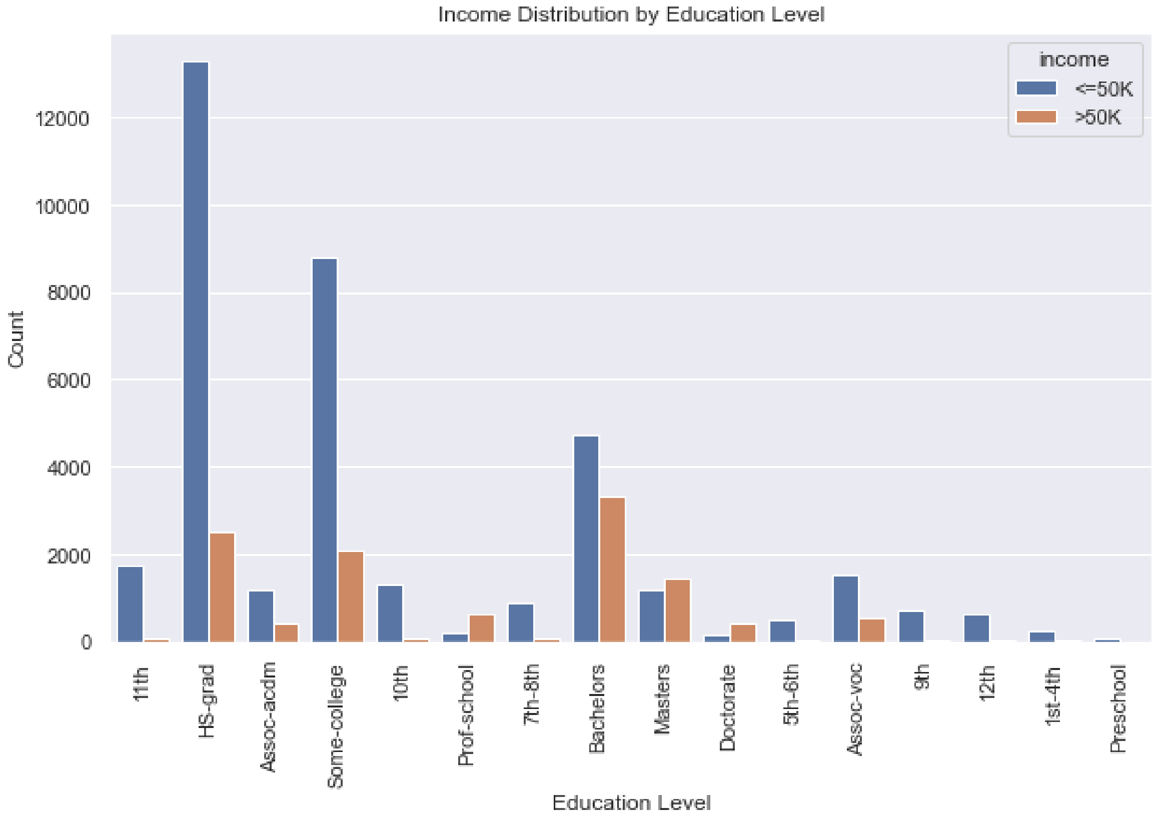
   Write the R or Python code in the box below.

R/Python script

```
plt.figure(figsize=(12,6))
sns.countplot(x=’education’, hue=’income’, data=df)
plt.xlabel("Education Level")
plt.ylabel("Count")
plt.title("Income Distribution by Education Level")
plt.xticks(rotation=90)
plt.show()
```

Problem 1 continued on next page. . .

Histograms/Bar Plots



2. List and show the distribution of occupations from which adults earned more than 50K even after working less than or equal to 40 hours per week? What kind of education level does it require? Write the R or Python code in the box below.

   Ans. People with bachelor degree more likely earn more than 50k anf work less than 40 hours. Following are the top 10 occupation where Occupation Distribution from which adults earned more than 50K even after working less than or equal to 40 hours per week:

   (a) Prof-specialty - 1429

   (b) Exec-managerial - 1149

   (c) Craft-repair - 812

   (d) Sales - 577

   (e) Adm-clerical - 567

   (f) Tech-support - 296

   (g) Machine-op-inspct - 247

   (h) Transport-moving - 202

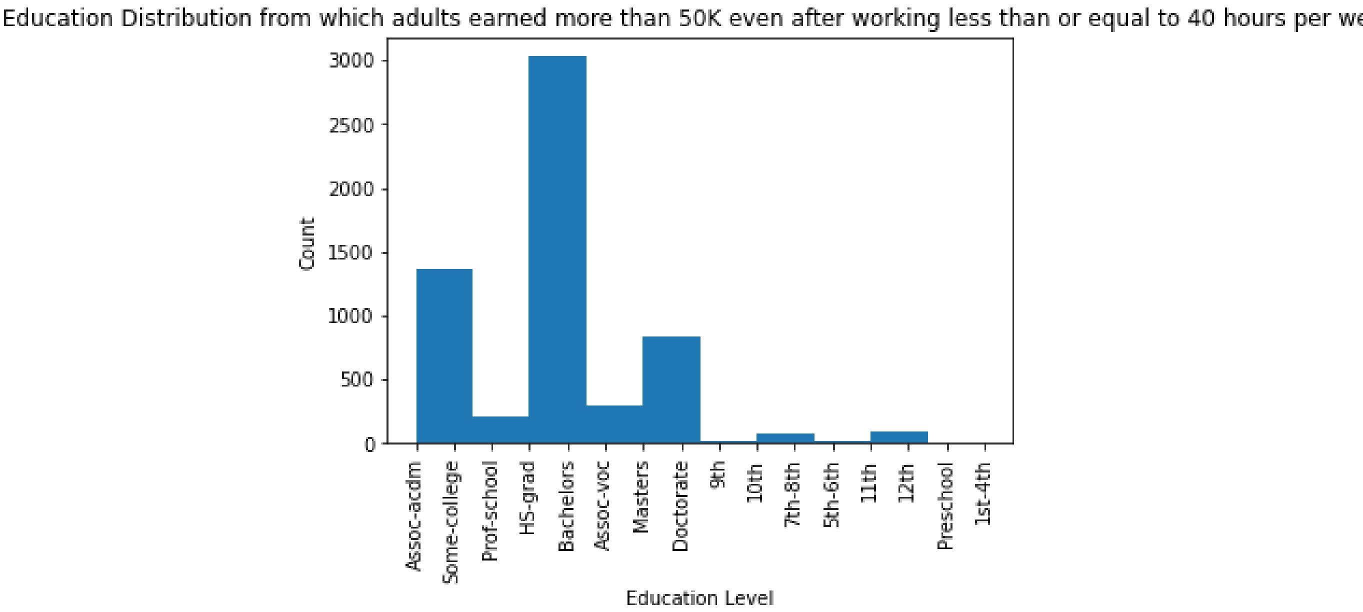
   (i) Protective-serv - 179

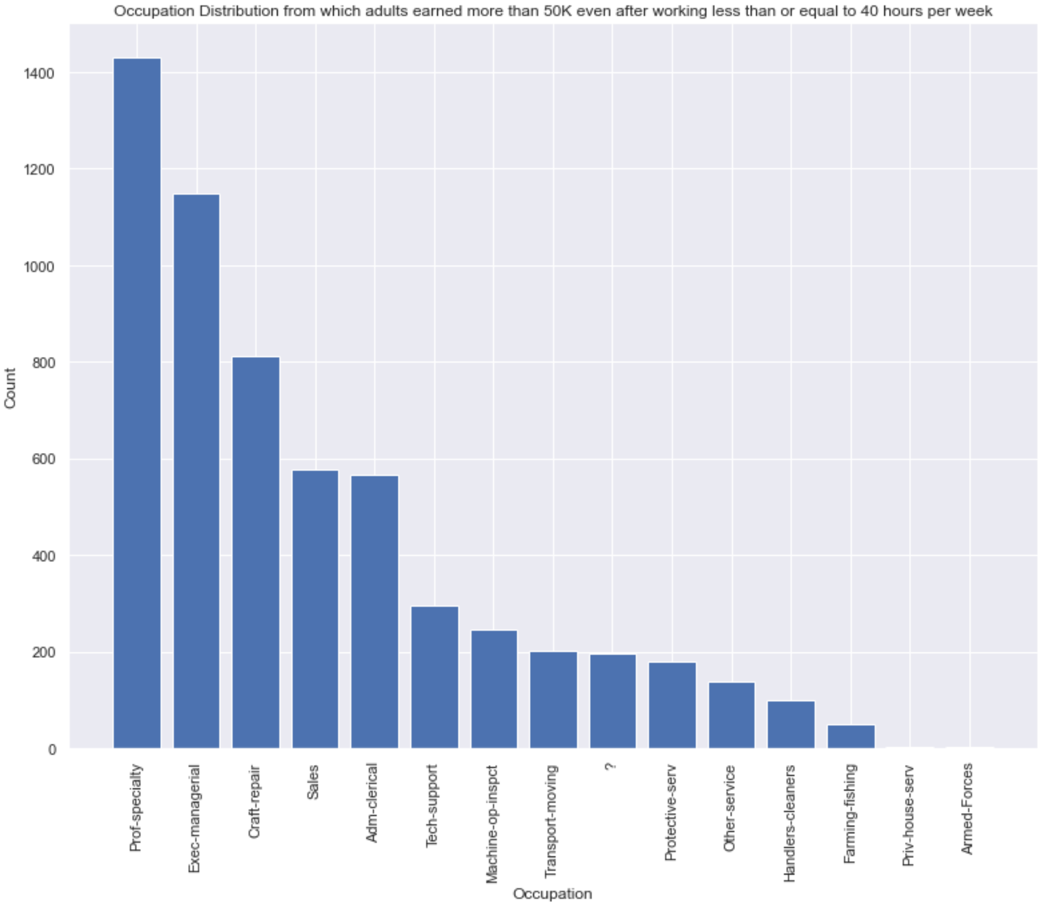
   (j) Other-service - 138

R/Python script

```
# filter
list_of_values = [">50K",">=50K"]
df2 = df[df["income"].isin(list_of_values)]
df2 = df2[df2["hours-per-week"]<=40]

plt.hist(df2[’education’])
plt.xlabel("Education Level")
plt.ylabel("Count")
plt.title("Education Distribution from which adults earned more than 50K
#even after working less than or equal to 40 hours per week")
plt.xticks(rotation=90)
plt.show()

# Count the number of occurrences of each occupation
occ = df2["occupation"].value_counts().sort_values(ascending=False)
# Occupation Distribution from which adults earned more than 50K
# even after working less than or equal to 40 hours per week

# Plot the bar plot
plt.bar(occ.index, occ.values)
plt.xlabel("Occupation")
plt.ylabel("Count")
plt.title("Occupation Distribution from which adults earned more than 50K
even after working less than or equal to 40 hours per week")
plt.xticks(rotation=90)
plt.show()
```

Histograms/Bar Plots





3. “Male adults earned more than female adults during the time the census was taken”. Can you prove this statement with the given data? Why or why not? Describe and illustrate. Write the R or Python code in the box below.

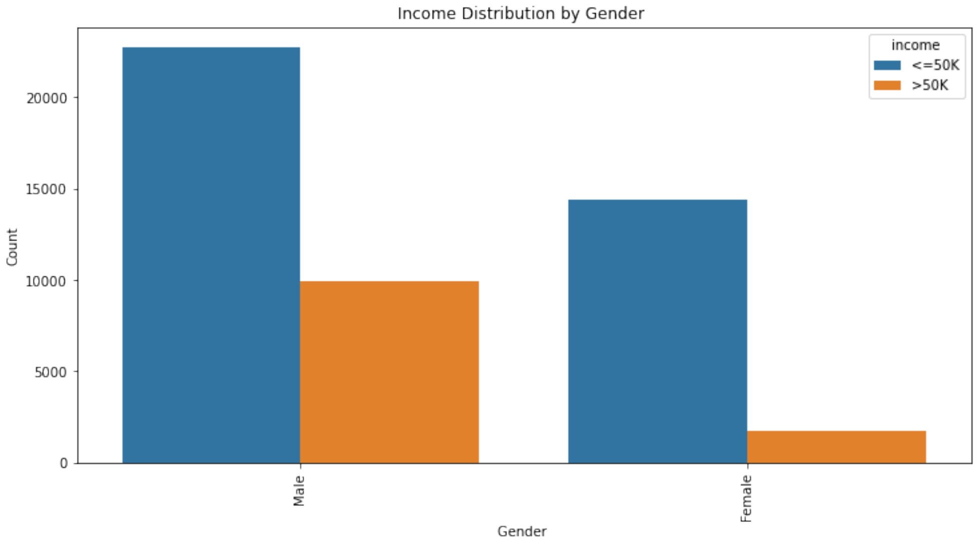
R/Python script

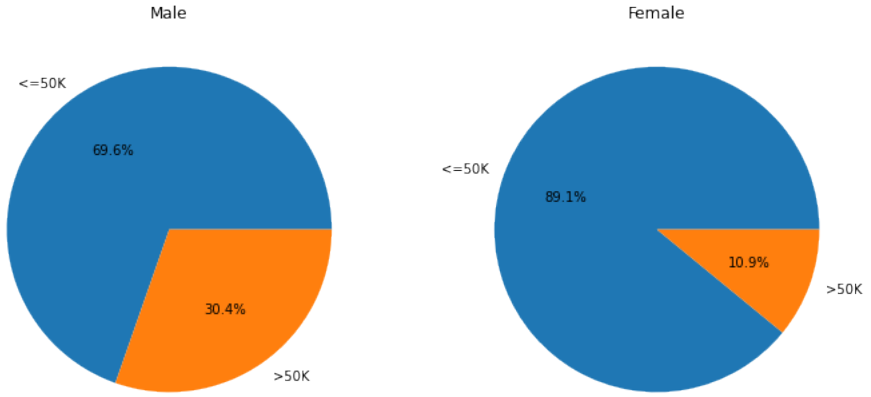
```
plt.figure(figsize=(12,6))
sns.countplot(x='gender', hue='income', data=df)
plt.xlabel("Gender ")
plt.ylabel("Count")
plt.title("Income Distribution by Gender ")
plt.xticks(rotation=90)
plt.show()

# To make more sense of the data, we plot a pie chart
# Calculate the percentage of males and females with different income levels
df_gender_income = df.groupby(['gender',
'income']).size().reset_index(name='counts')
df_gender_income['percentage'] = 100 * df_gender_income['counts'] /
df_gender_income['counts'].sum()

# Plot the pie chart for males and females with different income levels
plt.figure(figsize=(12,6))
for i, gender in enumerate(['Male', 'Female']):
    plt.subplot(1, 2, i+1)
    plt.pie(df_gender_income[df_gender_income['gender'] == gender]['percentage'],
    labels=df_gender_income[df_gender_income['gender'] == gender]['income'],
    autopct='%1.1f\%\%')
    plt.title(gender)

plt.show()
```

Discussion of Findings

The histogram gives us the distribution of income with respect to genders, but the number of males and females are different (Male - 32650, Female - 16192), placing it on a relative scale.
So to find out if Male adults earned more than female adults during the time the census was taken, We plot a pie chart depicting the percentage of each gender correspongin to different income group. We notice that Females adults indeed earn less than male adults during the time census was conducted, with only 10.9% of the females esrning more tahn 50K versus 30.4% men earning more than 50K.

Plot/s





4. Use box plots to show and describe the relation of categorical attributes with the target attribute `income`. *Note:* As `income` is a categorical attribute, be mindful about how you can display the relations in a compact manner.

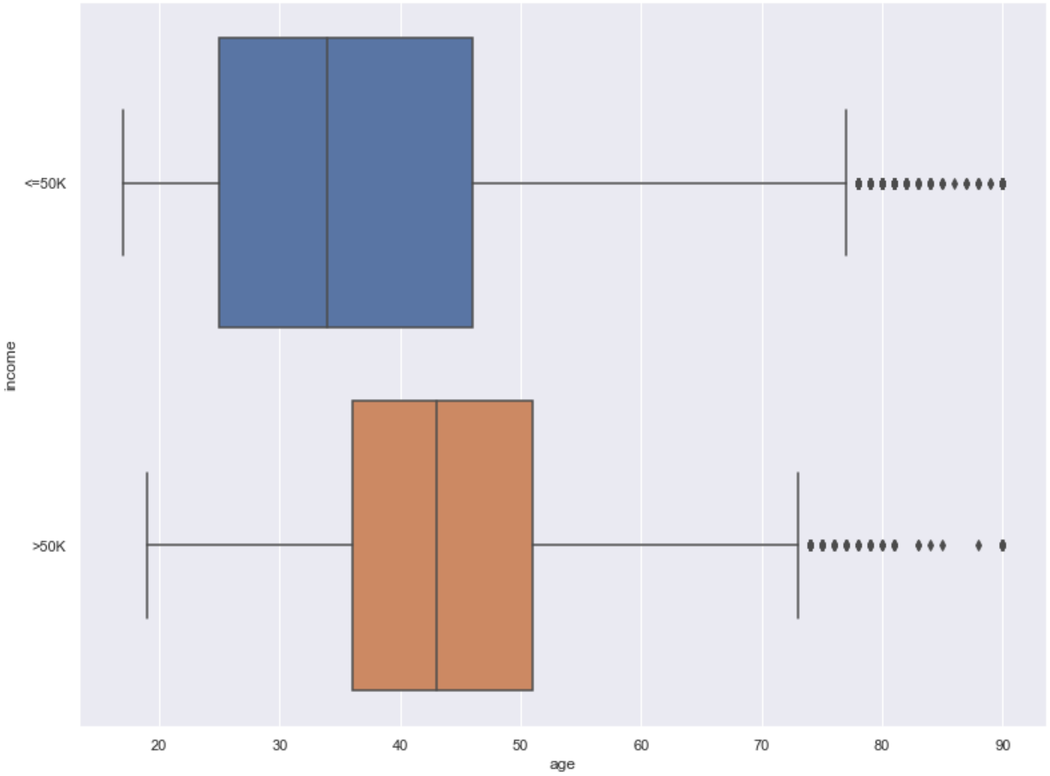
R/Python script

```
box_cols = df.columns.tolist()[:-2]
box_cols

for i in box_cols:
  sns.boxplot(x=df[i],y=df.income)
  plt.show()
```

Discussion

Answer here. . .

Box Plots



Problem 2

This is a data set collected from Boston Police Department with 17 different types of attributes regarding crimes in Boston. You can download the data from here [link]. From this you will download a .zip file which contains two files: `crime.csv` and `offense_codes.csv`. The former contains the actual data and the latter contains description of the offense codes mentioned in the actual data. *Note:* Perform appropriate missing data handling procedures if needed.

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?

R/Python Code

Using R/Python, show code and plots that answers the following questions:

1. Visualize the occurrences of different types of crimes using a bar chart. What is the most common type of crime Boston Police Department has to handle.

   Ans. Crimes.csv has crime data with shape: (319073, 17).It describes the crime via attributes like incident number, offence code, offence description, reporting area, shooting, street, longitude, lattitude details among others. The second file: offense_code has code numbers against offence name. The data set has all kinds of data like categorical data/ Nominal data(incident numbers, offence_codes) and numeric data(Hours). These data-sets can be used to visualize the crime data and patterns it might hold. Example: locations with high crime rate.

R/Python script

```
# Count the number of occurrences of each offense type
offense_counts = df_merged["NAME"].value_counts().sort_values(ascending=False)
# Siince there are around 150ntypes of offences,
#I've plotted most frequently occoring offenses
# Plot the bar plot
plt.bar(offense_counts.index[:20], offense_counts.values[:20])
plt.xlabel("Offense Type")
plt.ylabel("Number of Occurrences")
plt.title("Occurrences of Offenses in Boston")
plt.xticks(rotation=90)
plt.show()
```

Discussion of Findings

Most commom type of crimes that Boston police has to handle is Sick/Injured/Medical Person among many other frequent offenses like INVESTIGATE PERSON, M/V - LEAVING SCENE - PROPERTY DAMAGE, VANDALISM, ASSAULT SIMPLE - BATTERY , VERBAL DISPUTE etc.

Top Offenses with corresponding number of occurrences are as follows:

(a) SICK/INJURED/MEDICAL - PERSON : 18783

(b) INVESTIGATE PERSON : 18754

(c) M/V - LEAVING SCENE - PROPERTY DAMAGE : 16323

(d) VANDALISM : 15154

(e) ASSAULT SIMPLE - BATTERY : 14799

(f) VERBAL DISPUTE : 13099

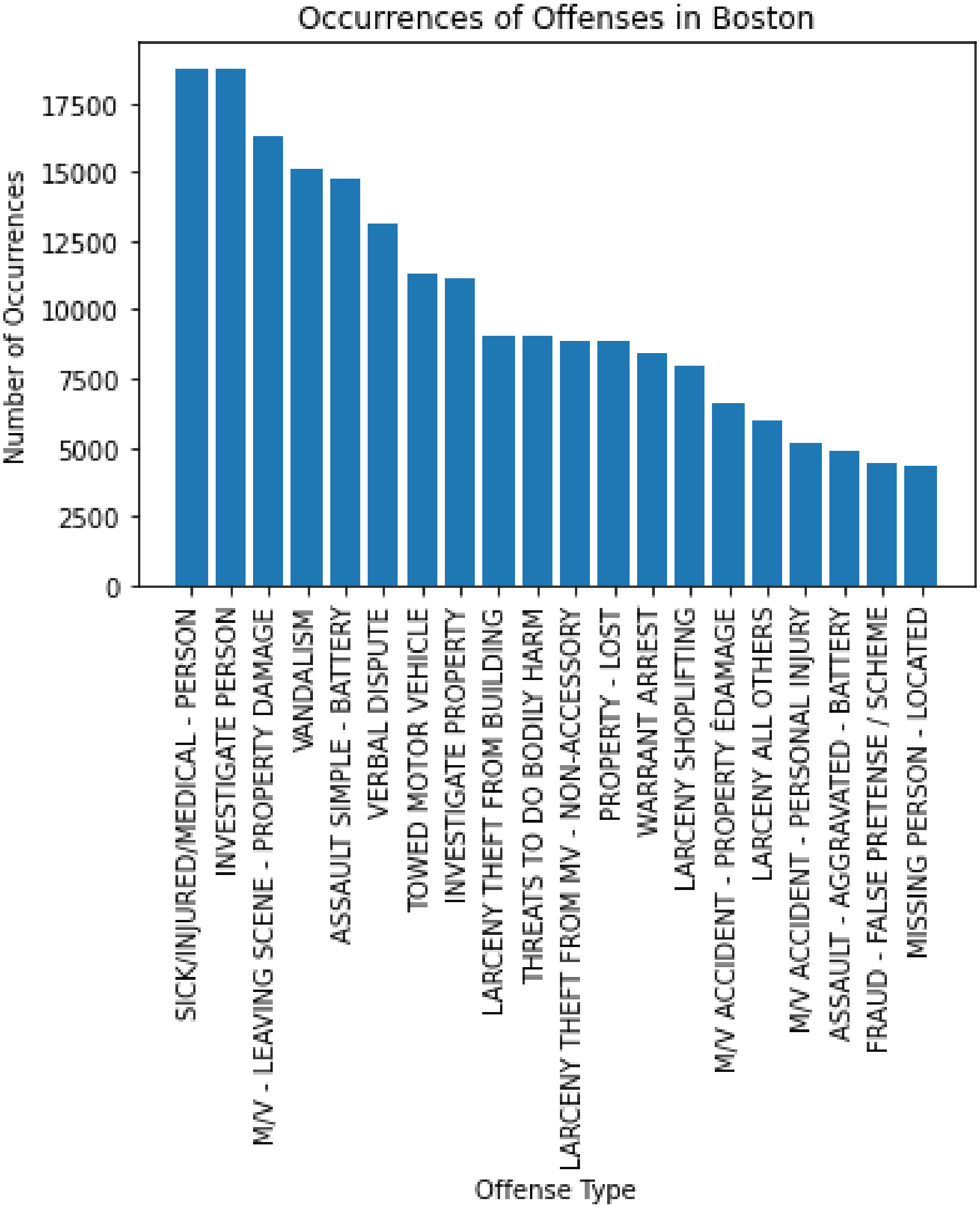
(g) TOWED MOTOR VEHICLE : 11287

(h) INVESTIGATE PROPERTY : 11124

(i) LARCENY THEFT FROM BUILDING : 9074

(j) THREATS TO DO BODILY HARM : 9042

Bar Plots



2. Which street in Boston had the most number of motor vehicle accidents? Find the top 10 streets where motor vehicle accidents happened and show their distribution using violin plots.

R/Python script

```
# Top 10 Streets with M/V accidents
mv_accident_count = df_accidents['STREET'].value_counts()
# plot: observing 100 to understand pattern
plt.violinplot(mv_accident_count.values[:100])
plt.xlabel('Streets')
plt.ylabel('Number of Motor Vehicle Accidents')
plt.title(f'Distribution of Motor Vehicle Accidents')
plt.show()
```

Discussion of Findings

Most number of M/V accidents happened at the Blue Hill Ave - 673, followed by 607 in Washington St. Top 10 streets with highest accidents are:

(a) BLUE HILL AVE : 673

(b) WASHINGTON ST :607

(c) DORCHESTER AVE : 430

(d) COMMONWEALTH AVE : 370

(e) CENTRE ST : 327

(f) COLUMBIA RD : 320

(g) HYDE PARK AVE : 317

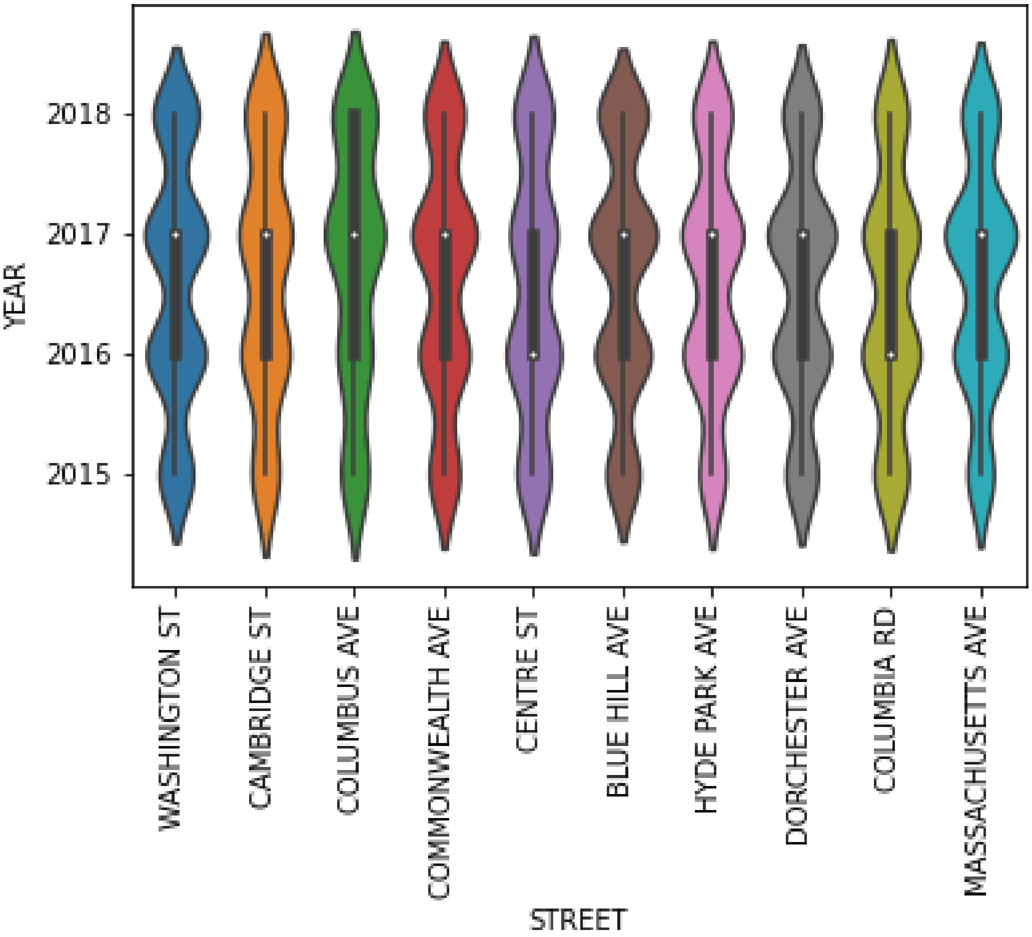
(h) MASSACHUSETTS AVE : 317

(i) CAMBRIDGE ST : 244

(j) COLUMBUS AVE : 230

Violin plot shows the distribution of crime on top 10 street with high crime rates across different years from 2015 to 2018. Columnbus avenue seems to be the thinnest specially through year 2015 to 2016. Blue Hill Avenue seems to be the thickest, showing highest density ie, highest number of crimes . Almost every street seem to have high crime rates specially in the year of 2016/2017 as compared to the other years.

Violin Plot/s



3. Using map, visualize the locations (using attributes `Lat`, `Long`, etc.) of crimes reported in Boston. Do you see a cluster of areas you should avoid to live in Boston? *Note:* You can use a Python library called "folium" for map visualization.

R/Python script

```

# Aggregate the crime data by sector
sector_groups = df_merged.groupby("DISTRICT")
sector_crime_counts = neighborhood_groups["DISTRICT"].count()

# Create a folium map centered on Boston
boston_map = fl.Map(location= [42.290196,-71.071590], zoom_start=12)

# Add markers for each neighborhood with the aggregated crime count
for sector, crime_count in sector_crime_counts.items():
    fl.CircleMarker(
        location=[sector_groups.get_group(sector)["Lat"].mean(),
        sector_groups.get_group(sector)["Long"].mean()],
        # set the marker size based on the crime count
        radius=crime_count/2000,
        fill = True,
        tooltip=f"{sector}: {crime_count} crimes"
    ).add_to(boston_map)

boston_map
```

Discussion of Findings

Since there are a lot of longitude/ combination, the map gets very cluttered, I plot a map with respect to crimes in different sectors. We notice that:

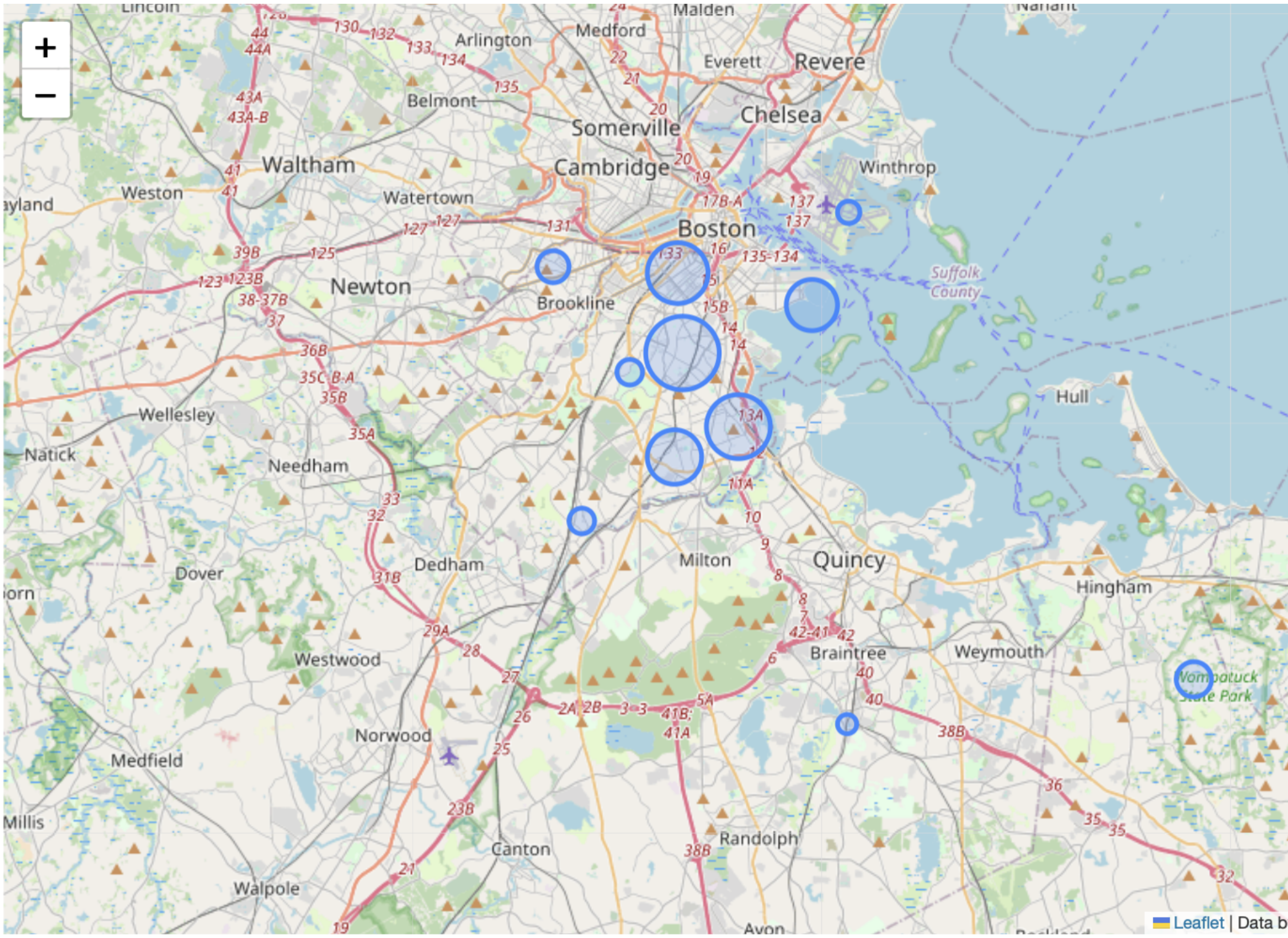
(a) B2 Sector has highest number of crimes - 46207

(b) C11 Sector has second highest number of crimes - 40875

(c) D4, B3 are other sectors with high crime rate, hence unsafe.

The map shows circular marks with radius proportional to the number of crime in that particular sector. So from the graph we understand that we should avoid sectors B2, C11, D4, B3 since the crime rate is higher in those areas.

Map Visualization



4. Using appropriate plotting methods, show the distribution of crime reporting in Boston with respect to time, day, and month. Discuss what you can conclude from those visualizations. What would be the best month/s of the year to visit Boston to avoid any sort of crime?

R/Python script

```
# Count the number of occurrences WRT HOUR
hour_counts = df_merged["HOUR"].value_counts().sort_values(ascending=False)
# Count the number of occurrences WRT MONTH
month_counts = df_merged["MONTH"].value_counts().sort_values(ascending=False)
# Count the number of occurrences WRT Day
day_counts = df_merged["DAY_OF_WEEK"].value_counts().sort_values(ascending=False)

# Plot the number of crimes by hour
plt.figure(figsize=(12,6))
sns.lineplot(x=hour_counts.index, y=hour_counts.values)
plt.xlabel("Hour")
plt.ylabel("Number of Crimes")
plt.title("Number of Crimes by Hour")
```

```

# Plot the number of crimes by Month
plt.figure(figsize=(12,6))
sns.lineplot(x=month_counts.index, y=month_counts.values)
plt.xlabel("Month")
plt.ylabel("Number of Crimes")
plt.title("Number of Crimes by Month")


# Plot the number of crimes by Day
plt.figure(figsize=(12,6))
sns.lineplot(x=day_counts.index, y=day_counts.values)
plt.xlabel("DAY of Week")
plt.ylabel("Number of Crimes")
plt.title("Number of Crimes by Day of Week")
```

Discussion

We observe from the graph that :

(a) The possibility of crime happening peaks at times between 3 pm to 8 pm.

(b) The chance of crime happening at 5 am in the morning is the least

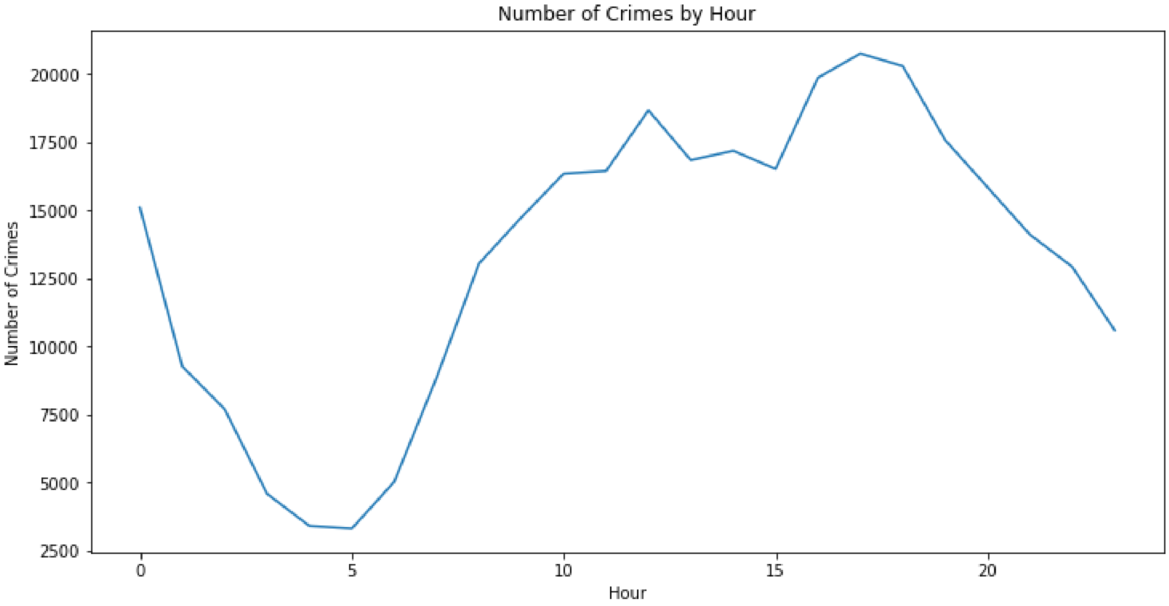
(c) The crime rate is highest during the month of June-August

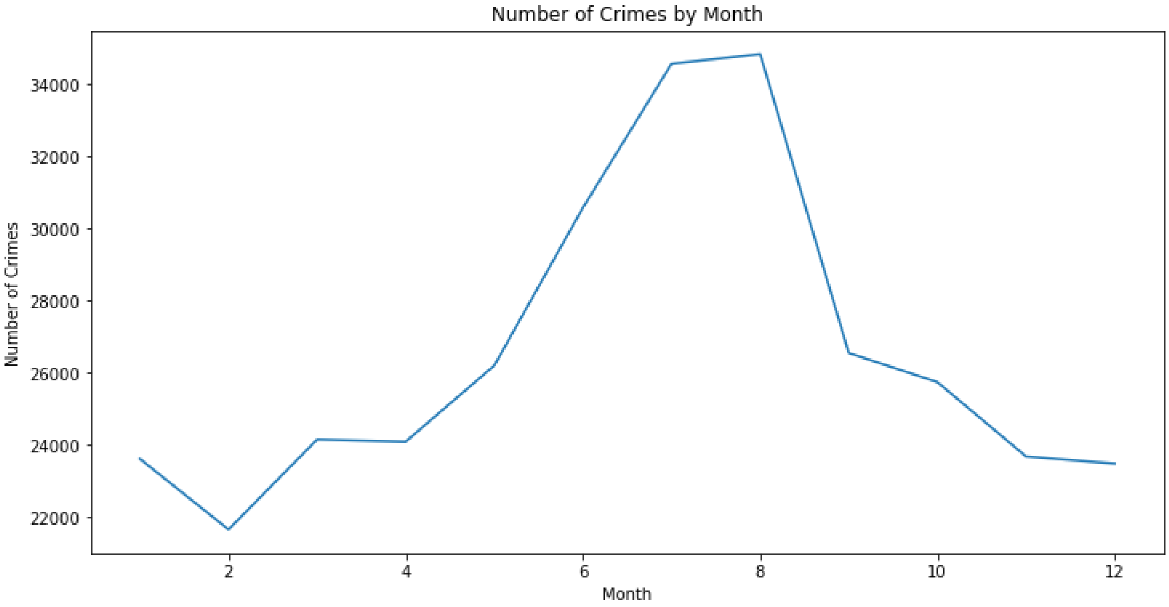
(d) The crime rate is lowest in the month of February

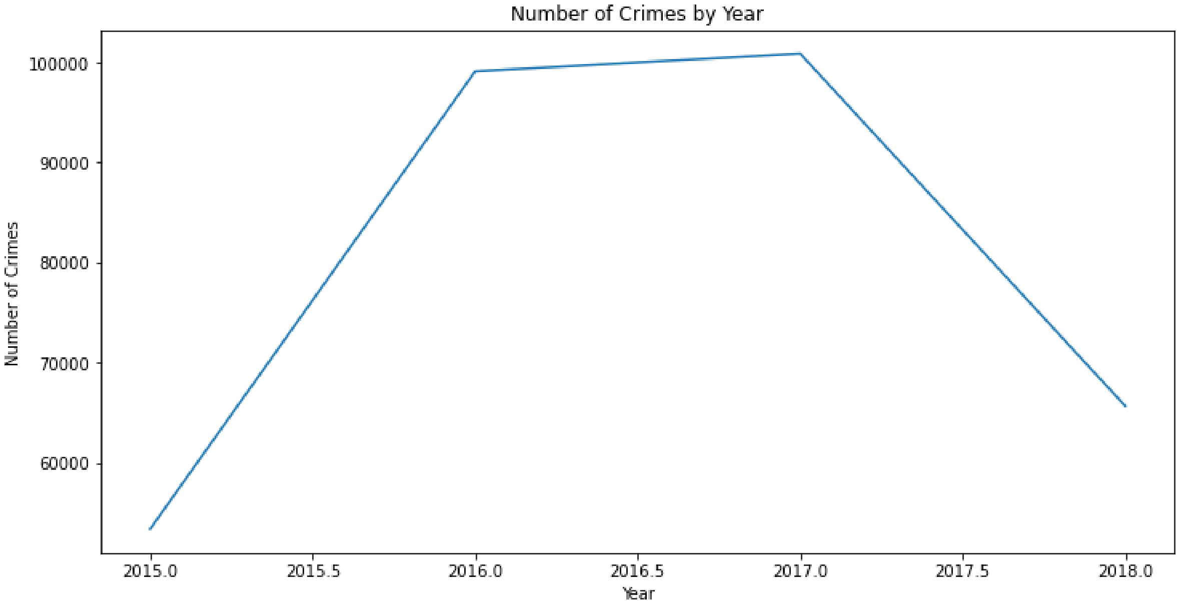
(e) The chances of crime occurring is highest on Fridays and least on Saturdays

The best month to visit Boston would be February in terms of safety to travel, since crime rate seems to be lowest.

Plot/s


Number of Crimes by Hour
Number of Crimes
20000
17500
15000
12500
10000
7500
5000
2500
0
5
10
15
20
Hour


Number of Crimes by Month
Number of Crimes
34000
32000
30000
28000
26000
24000
22000
2
4
6
8
10
12
Month



Problem 3

This is a data set regarding drug consumption of over 1K of human subjects. The participants were asked to rate their use (from never to last day) over 18 different drugs (either legal or illegal). You can download the data from here [link]. The downloadable file already comes with the corresponding names of the attributes. Also a short description of the ratings of drug usage is available here [link]. *Note:* Perform appropriate missing data handling procedures if needed.

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?
In this dataset, we are trying to analyze the variation and dependency of charecters of people with the use of different drugs. There area about 18 drugs we have in the dataset including alcohol and we have afew charectaristics of people's charecters in 5 other columns. These include age, gender, impulsiveness etc

R/Python Code

Using R/Python, show code and plots that answers the following questions:

1. For 5 drugs of your choice, show the distribution of the ratings (`CL0` to `CL6`) in a single image using box plots. Do you find any relations of the drugs you have chosen with the ratings? Which drug is being used the most in recent times?

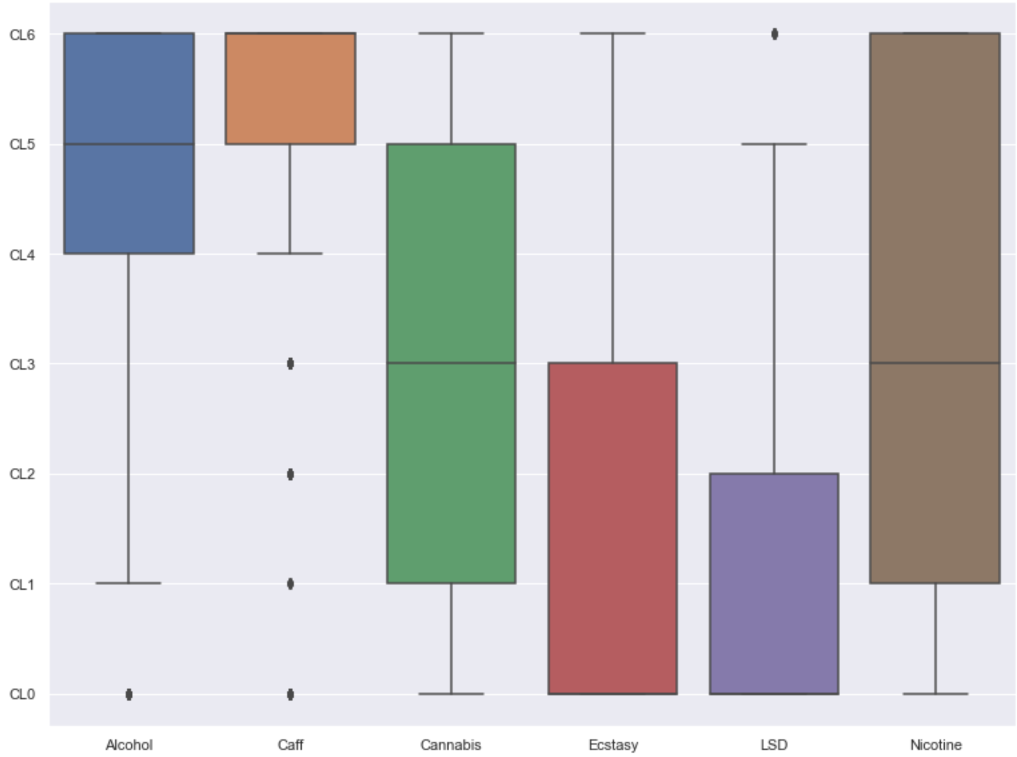
```
drugs = ['Alcohol',
        'Caff', 'Cannabis',
       'Ecstasy',  'LSD',
       'Nicotine']
sns.boxplot(data = df_drugs[drugs])
```

```
plt.yticks(ticks = [0,1,2,3,4,5,6],labels=["CL0","CL1","CL2","CL3","CL4","CL5","CL6"])
plt.show()
```

Discussion of Findings

We notice that Caff has been used the most in recent times, median rating being C6 and majority of the data being in CL5-CL6. Nicotine is being used fronm quite a long time from CL1 to CL6 with median data on CL3. Caff, LSD, Ecstacy are skewed. Caff towards CL6 and Ecstacy and LSD towards CL0. Cannabis is also used majorly from CL1 to CL5, ie, ranging from use since a decade till last week.

Rating Distribution



2. Mine this data to find out if there is a preferred list of drugs for females? Do females take less illegal drugs/no drugs at all than males? Mention your assumption about which drugs you are considering illegal. Describe your answers with suitable visualizations.

R/Python script

```

\begin{lstlisting}[language=Python]
df_drug_f =  df_drugs[df_drugs[’Gender’]==’F’]
df_total_f = (~(df_drug_f[drugcols] == 0)).sum().sort_values(ascending=False)

# Plot the bar plot for preference of women
```

```
plt.bar(df_total_f.index, df_total_f.values)
plt.xlabel("Drug Type")
plt.ylabel("Number of Occurrences")
plt.title("Preference of Women")
plt.xticks(rotation=90)
plt.show()

# consumptionof illegal drug by men vs women
illegal_drugs = [ 'Amyl', 'Benzos', 'Cannabis',  'Coke', 'Crack',
       'Ecstasy', 'Heroin', 'Ketamine', 'Legalh', 'LSD', 'Meth', 'Mushrooms',
       'Nicotine', 'VSA']
# percemtage of females using iklegal drugs
df_illegal_f = ((~(df_drug_f[illegal_drugs] == 0)).sum().sum()/df_total_f.sum())*100
df_drug_m =  df_drugs[df_drugs['Gender']=='M']
df_total_m = (~(df_drug_m[drugcols] == 0)).sum().sort_values(ascending=False)
df_illegal_m = ((~(df_drug_m[illegal_drugs] == 0)).sum().sum()/df_total_m.sum())*100
plt.pie([df_illegal_m, df_illegal_f], labels=['Illegal drugs consumed by Males','Illegal drugs c
plt.title('Illegal drugs consumed by Males vs females')


plt.show()
```

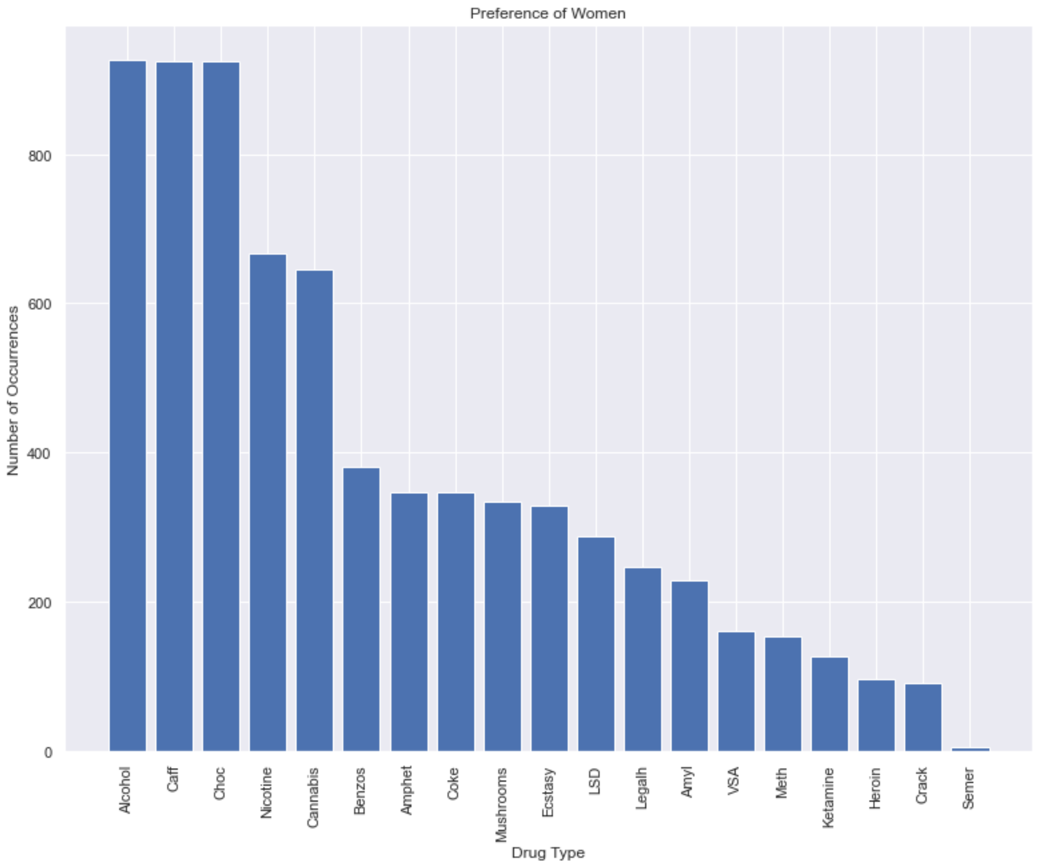
Discussion of Findings

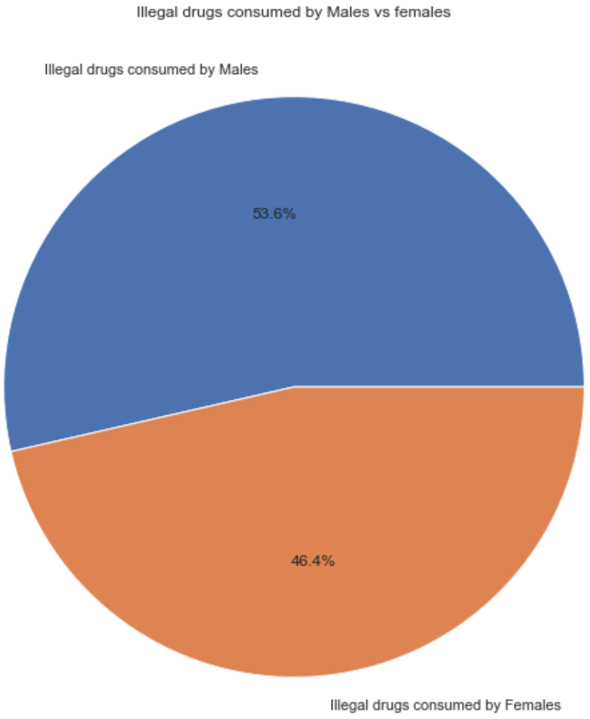
In the bar plot we notice that, Alcohol, Caff, Choc seems to be the most popular choice for female, followed by Nicotine, Cannabis among rest of drugs.

We assume that drugs like alcohol, nicotine, caff, choc, amphet are legal. ALthough other drugs may also be legal depending on the country or state

As we can infer from the pie plot, Females relatitively take less illegal drugs when comapred to men who take around 7.2% more than females.

Visualizations





3. Find out which attributes (among NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking)) are in high correlation with respect to each of the 18 drugs. Show the distributions (using suitable plotting method) of the first two highly correlated attributes for each of the drugs (with all 7 ranking classes).

R/Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

Discussion of Findings

A score and C score have highest correlation with almost all of the 18 drugs with a negative correlation of around -0.4.

Impulsivity has a low negative correlation with alcohol, choc, caff and semer.

SS has a good positive correlation with some of the drugs whereas has a negative correlation with some drugs like semer, heroin, crack, choc and caf

Correlation Map

Place images here with suitable captions.

4. Visualize the correlation map among the drugs and the 12 personality measurement attributes. Do you find any significant correlation among the drugs?

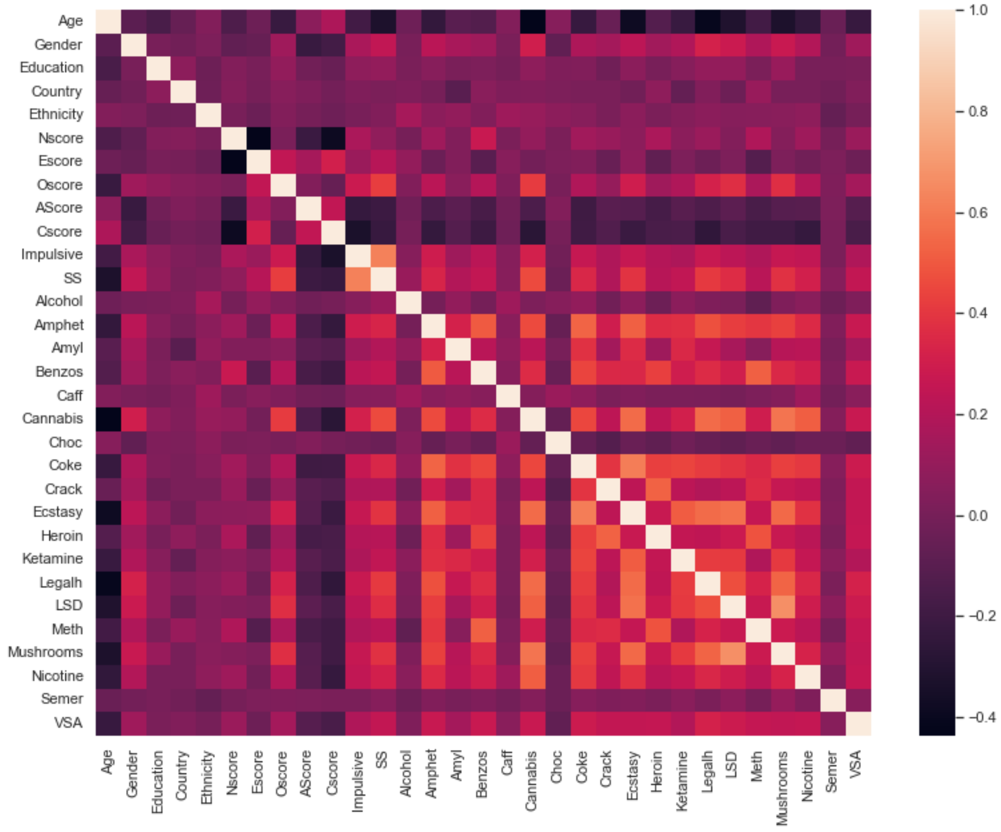
R/Python script

```
le = LabelEncoder()

# Select the columns you want to encode
cols = [’Age’, ’Education’,’Country’,’Gender’,’Ethnicity’]

# Apply the encoding to multiple columns
df_drugs[cols] = df_drugs[cols].apply(lambda col: le.fit_transform(col))


corr = df_drugs.corr()

# Plot the heatmap
sns.heatmap(corr)

# Show the plot
plt.show()
```

Discussion of Findings

Indirect correlation between age and most of drugs whereas direct correlation appears between gender, and most of the drugs. C Score looks to have a slight negative correlation with the drugs. Rest of the personality attributes have a 0 to low positive correlation with the drugs.

Distribution Plots



Problem 4

How might you extend the notion of multidimensional data analysis so that the target variable is a qualitative variable? In other words, what sorts of summary statistics or data visualizations would be of interest?
Ans. If the target variable is qualitative variable, some of the statistics that can be used are 'frequency distribution' and 'chi-squared tests'. Frequency distribution can be used to calculate the total occurrences of each qualitative feature value and plot visualizations like pie charts and histogram.

Problem 5

Construct a data cube from Table 1. Is this a dense or sparse data cube? If it is sparse, identify the cells that are empty.

Ans. Data cube is multidimensional representation of data with all possible aggregates. To represent the table as data cube, we store the data in 3- dimensions, the dimensions being: ProductID, LocationID

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

Table 1: Table for Problem 5.

		LocationID			
		1	2	3	Total
	1	10	0	6	16
ProductID	2	5	22	0	27
	Total	15	22	6	43

This can be considered as a dense data cube, because only the cells where:

1. productID = 1 and LocationID = 2
2. productID = 2 and LocationID = 3

are empty.

and Number Sold. Aggregation can be done accross dimensions like consider ProductID, LocationID, over Number sold over the dimensions.
Consider the table below to represent data cube across 3 dimensions: