Project 1 Report

Team members: Jefferson Roylance, Supratik Chanda

Introduction

This reports analyzes various aspects of crime statistics in Austin, Texas from 2015. The primary purpose of the document is to educate the authors in techniques involving the python pandas library, as well as data analysis and visualization. Secondary purposes include useful statistics that would help with city planning (and be useful if one was walking the streets of Austin in the dark). Our analysis centered around a few important questions:

- What are the most frequent crimes committed in 10 locations?
- How does the amount of crimes vary with unemployment rates?
- Which offense between theft and burglary is more related to unemployment rate?
- Which types of crimes get cleared the most?
- How does median income affect clearance rate of crimes?
- How does the amount of crimes committed per person vary with median income?

Dataset

Our dataset used crime statistics gathered from Austin, Texas in 2015. It included various statistics for the surrounding area as well. The columns of this dataset were as follows:

- Key
- Council District
- Highest Offense Desc
- Highest NIBRS UCR Offense Description
- Report Date
- Location
- Clearance Status
- Clearance Date
- District
- Zip Code of Crime
- Census Tract
- X Coordinate
- Y Coordinate
- Housing Zip Code
- Population below poverty level
- Median household income
- Percentage of Non-White, Non-Hispanic, or Latino population
- Percentage of Hispanic or Latino of any race
- Percentage of Population with disability
- Unemployment

- Percentage of Large households(5+members)
- Homes affordable to people earning less than $50000
- Rentals affordable to people earning less than $25000
- Rent-restricted units
- Housing Choice Voucher holders
- Median rent
- Median home value
- Percentage of rental units in poor condition
- Percent change in number of housing units 2000-2012
- Owner units affordable to average retail/service worker
- Rental units affordable to average retail/service worker
- Rental units affordable to average artist
- Owner units affordable to average artist
- Rental units affordable to average teacher
- Owner units affordable to average teacher
- Rental units affordable to average tech worker
- Owner units affordable to average tech worker
- Change in percentage of population below poverty 2000-2012
- Change in median rent 2000-2012
- Change in median home value 2000-2012
- Percentage of homes within 1/4-mi of transit stop
- Average monthly transportation cost
- Percentage of housing and transportation costs that is transportation-related

This dataset was quite clean, but we had to remove some null values from the table. Furthermore, one primary weakness was the lack of population records. Since this only contained crime statistics, we were unable to utilize quite a few statistics. For example, we couldn't analyze how many crimes were committed per person (outside of the dataset mentioned next), limiting our capabilities.

In addition to the crime dataset, we used another set of statistics downloaded from https://datausa.io/profile/geo/austin-tx/ (https://datausa.io/profile/geo/austin-tx/). It included population counts for various levels of median incomes, and we were able to use this data to analyze how crimes committed per capita varied with median income.

Analysis technique

We preformed six analyses,each of whose description is given in the introduction

Which are the most frequent crimes happened in randomly selected 10 cities?

Analysis by: Supratik Chanda

This analysis is to figure which crimes are the most frequent and which crimes should be stoppped in the respective cities to reduce the crime rate . This in turn will restrict the expansion of crime rate and eventually led to a safer place to live in

How does Crime_Count varies with the unemployment rate in Austin?

Used correlation technique to find out whether crime_frequency is conclusively related to the unemloyment rate or not

Analysis by: Supratik Chanda

A strong sense by analysts explicitly implies that unemployment is the root cause of illegal jobs. Crime is the most obvious among them. So,this analysis is done to strenghten that inference that whether or not unemployment rate is surely the reason for rapid burst of crimes. With the help of statistical analysis such as pearsonr test, we find out the correlation and enough evidence to say that whether with an increase in the mean unemployment rate, no of crimes increases. But according to our findings, we can see that the p_value is close to 9% which is bigger than 5 % there is a 9 % chance that there the data of crimes and unemployment rate is randomly taken . Hence there is no clear evidence that there is a clear relation between no_of_crimes and mean_unemployment. Therefore , we cannot conclusively say that it is because employment that the crime numbers are growing up. There could also be other factors like kleptomania, avengeful natures,racism etc.

But i went to finally crunch the numbers and find out if there are any outliers that's increasing the mean and eventually the p_value . Unfortunately , I found out with the help of box plots that there were no such outliers that can be omitted to find out more conclusive results with the help of statistical difference. Therefore, we cannot say overtly that it's because mean unemployment that no of crimes are expanding

Which offense between theft and burglary is more related to unemployment rate?

Is there a significant difference between the unemployment rate of the two offenses(Theft and Burglary)? Or is just a fluke. Let's find out with the help of T_Test.

Analysis by: Supratik Chanda

First of all, we have normalised the unemployment columns of both the datasets of of theft and burglary .Then we plotted a normalised kde plot showing the mean of both the datasets. It comes out that the meanof unemployment vs theft is more than the unemployment due to burglary . But as we know , this difference can be randomly generated and may not be a conclusive evidence to proof that unemployment because of theft is more than burglary. That's why we perform a T-Test.

And as the outcome: we found out that from the above analysis , we come into a conclusion that the p_value for both of the two datasets is significantly low ,much lower than 0.05. p_value of 0.05 says thet there is a 5 % chance that the data is random and there is a real difference. T value relates the size of the differences. Now the p_value that we are getting is approximately 0.09% which is way lower than 5% . Henceforth we can infer conclusively that since the mean of theft is higher and also the p_value is negligible and has .09% chance of random datas, unemployment is a credible and large cause to perform theft .Whereas burglary is not so much a cause of unemployment

Which types of crimes get cleared most?

Analysis by Jefferson Roylance

For this analysis, we grouped crimes by type and took the count of both total crimes and crimes that were cleared. We then plotted the results on a graph, showing what proportion of crimes of each type were cleared.

How does median income affect crime clearance rate?

Analysis by Jefferson Roylance

We then decided to take a look at how median income affects crime clearance rate. In this case, we grouped crimes by the median income and then plotted that against clearance rate.

How does the average crime rate vary by median income?

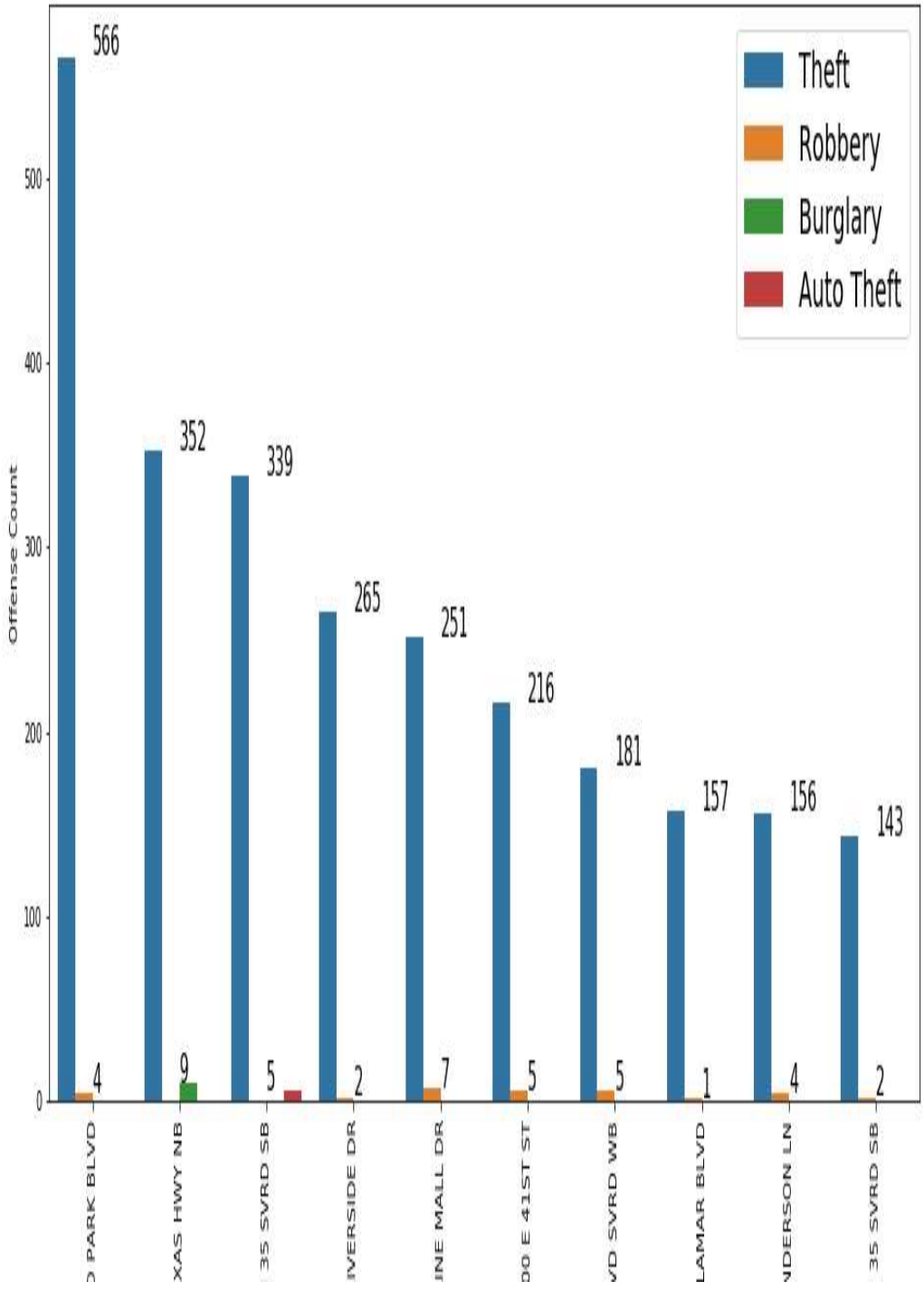
Analysis by Jefferson Roylance

Finally, we examined how the average amount of crime records per person varied with median income. We first downloaded a dataset that showed the population of Austin, Texas for various levels of median income. Then, we cross referenced this with crimes, grouped by levels of median income. Finally, we divided the amount of crimes by the population with a specific median income, resulting in average number of crime records per person.

Results

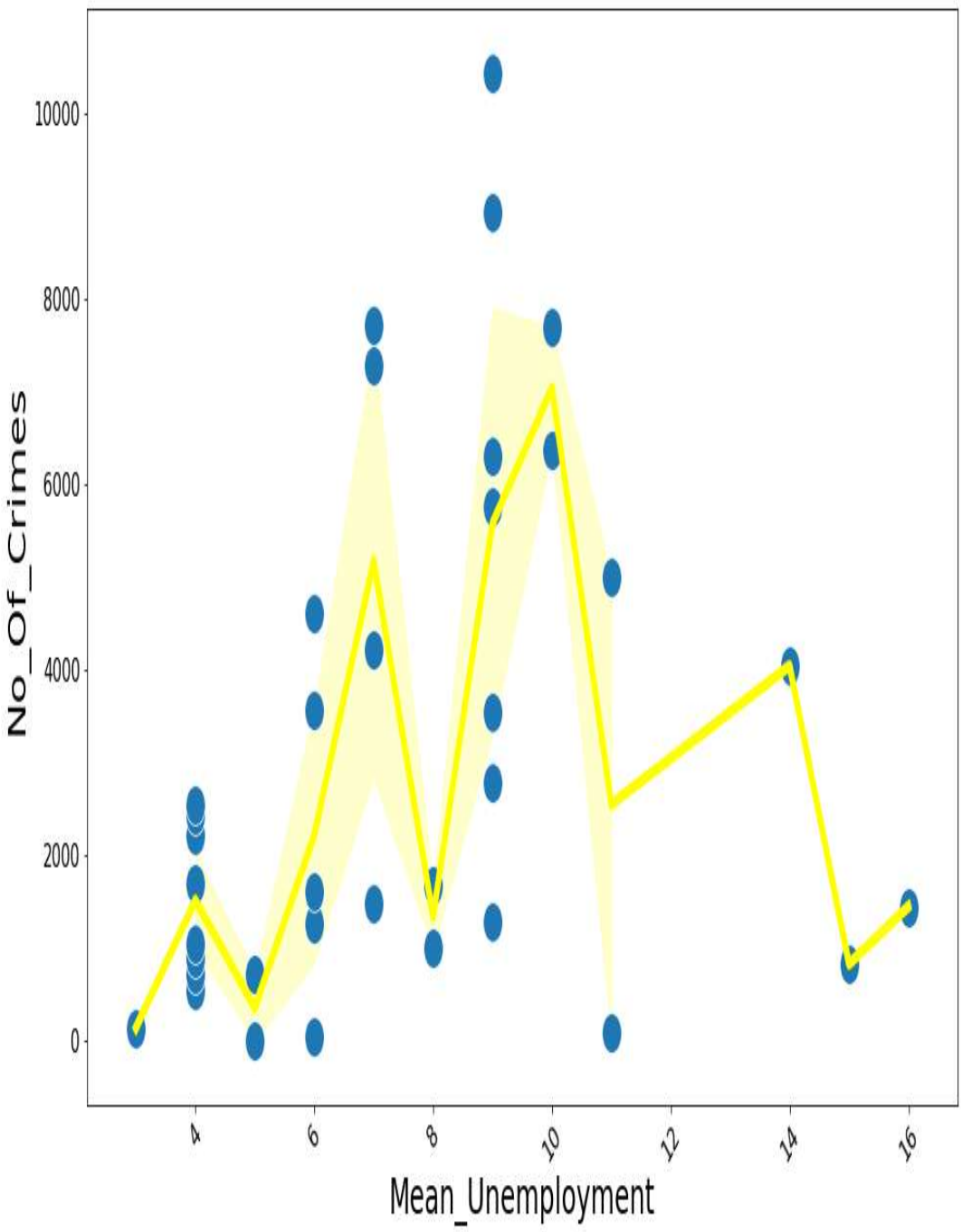
Analysis 1 :Which are the most frequent crimes happened in randomly selected 10 cities?

Figure 1 describes which players are the best among all .

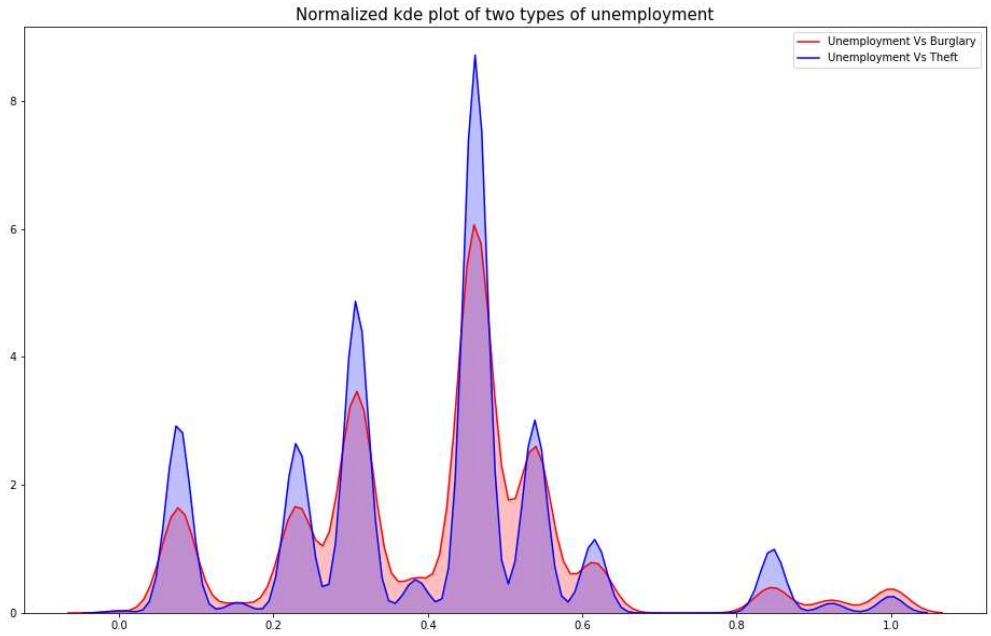


Analysis 2: How does Crime_Count varies with the unemployment rate in Austin?

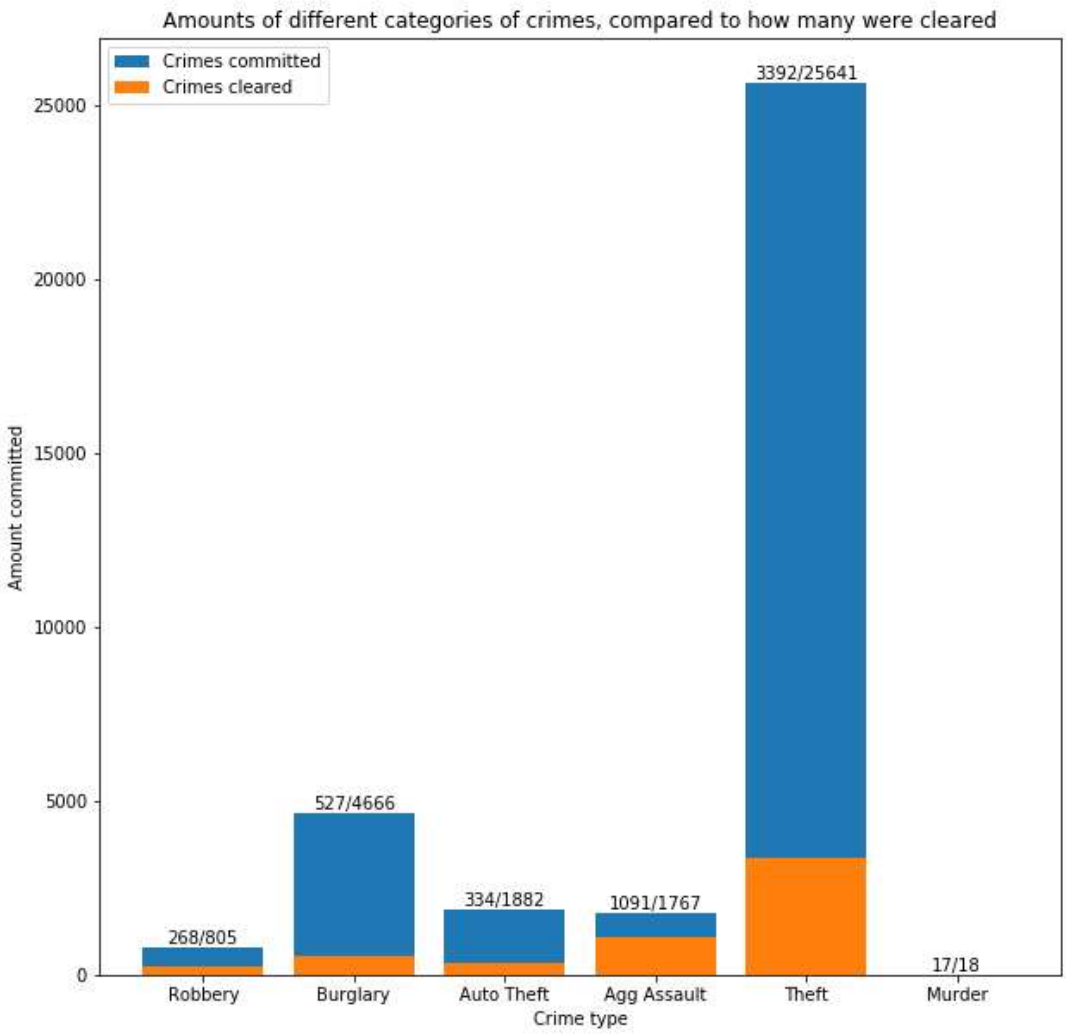
Figure 2 shows whether Crime_count is directly or inversely connected with unemployment rate



Analysis 3: Which offense between theft and burglary is more related to unemployment rate?


Normalized kde plot of two types of unemployment
Unemployment Vs Burglary
Unemployment Vs Theft
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

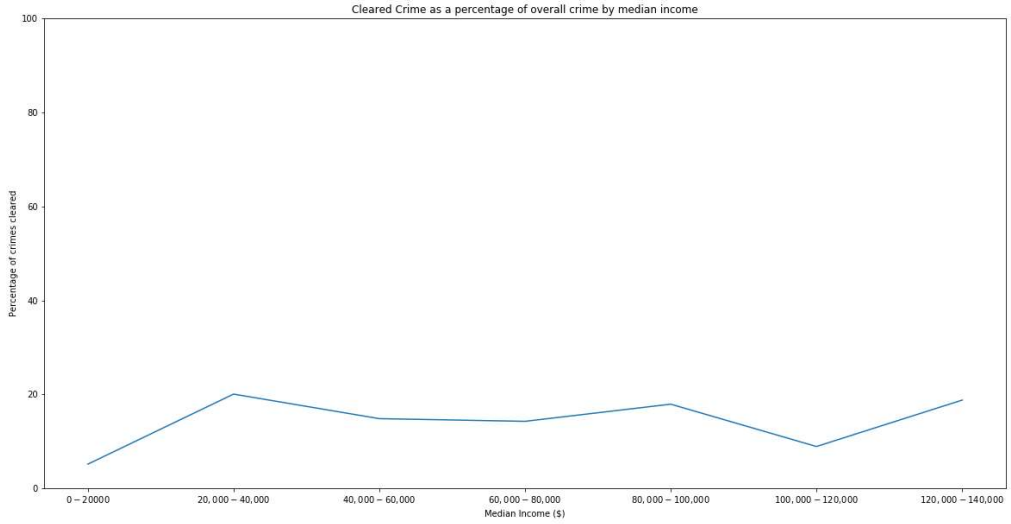
Analysis 4



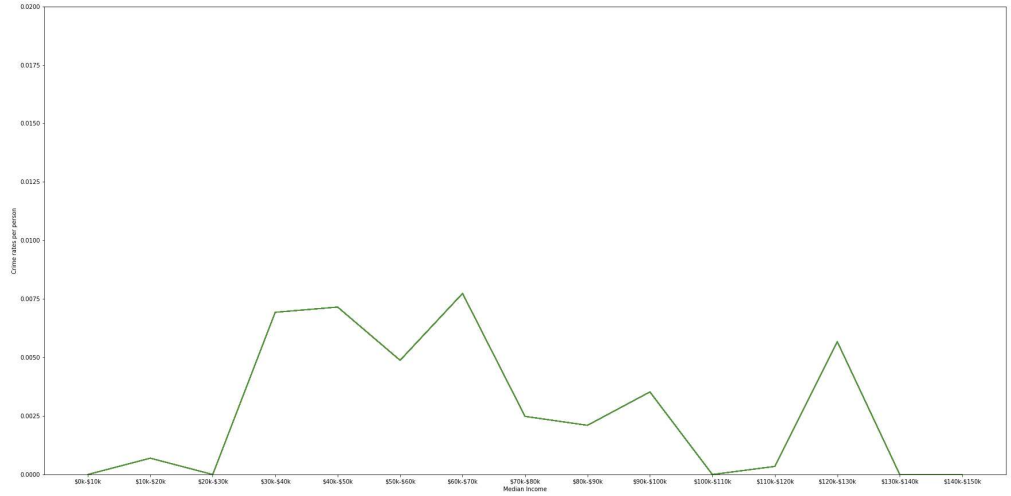
This chart has some interesting data. First off, we can see that theft is far and away the most popular crime. Interestingly enough, theft, along with burglary, are the least cleared crimes. Unfortunately, my knowledge of the U.S. legal system is very limited but this seems counterintuitive. Most people would think that being among the least harmful crimes, they would have the least effort put toward punishment. One hypothesis could be that it's less likely that they have a good defense lawyer, or perhaps they're offered plea deals more often?

This comes in sharp contrast to murders and aggrivated assaults, which get cleared significantly more. With murders, this could make sense because they'd be more high-profile (and therefore more likely to get a good lawyer?). With assaults, however, this is quite confusing because they'd have more witnesses, as well as plenty of evidence against them.

Analysis 5



Analysis 6



Here we unpack the data from files

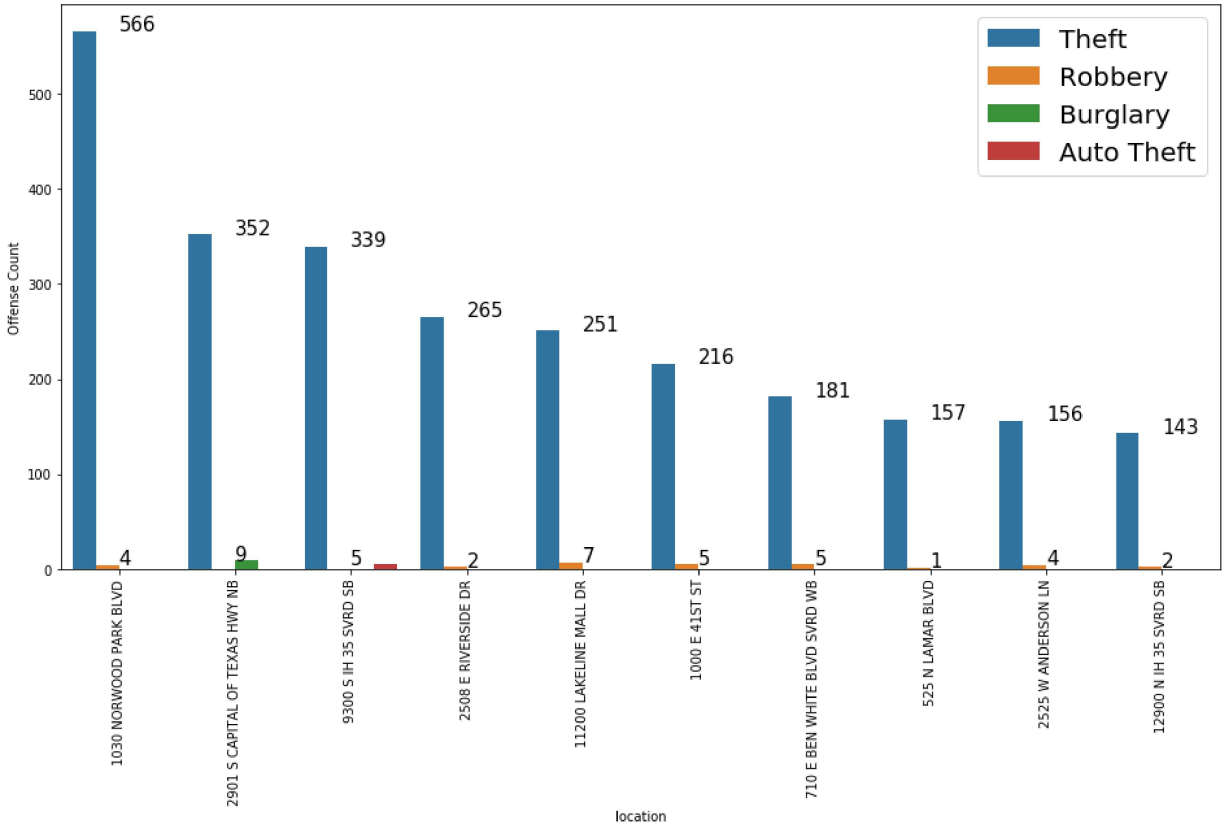
We also clean it up by removing some null values, and include various python packages

Finally, we changed some percentage and dollar values to floats to make data analysis easier.

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display
import math
from scipy import stats

df=pd.read_csv('crime-housing-austin-2015.csv')

#Copy reports for this use
reports1 = df[:]

# Drop blank rows from reports1
reports1 = reports1.dropna()

reports1['HispanicorLatinoofanyrace'] = reports1['HispanicorLatinoofanyrace'].str
reports1['Medianhomevalue'] = reports1['Medianhomevalue'].str.replace('$','').ast
reports1['Largehouseholds(5+members)'] = reports1['Largehouseholds(5+members)'].s
reports1['Changeinpercentageofpopulationbelowpoverty2000-2012'] = reports1['Chang
reports1['Medianhouseholdincome'] = reports1['Medianhouseholdincome'].str.replace
reports1['Populationbelowpovertylevel'] = reports1['Populationbelowpovertylevel']
```

ANALYSIS 1: To find the most frequent crime happening in randomly selected 10 cities in Austin . Why Random? Because the total cities covered in the dataset is 16764

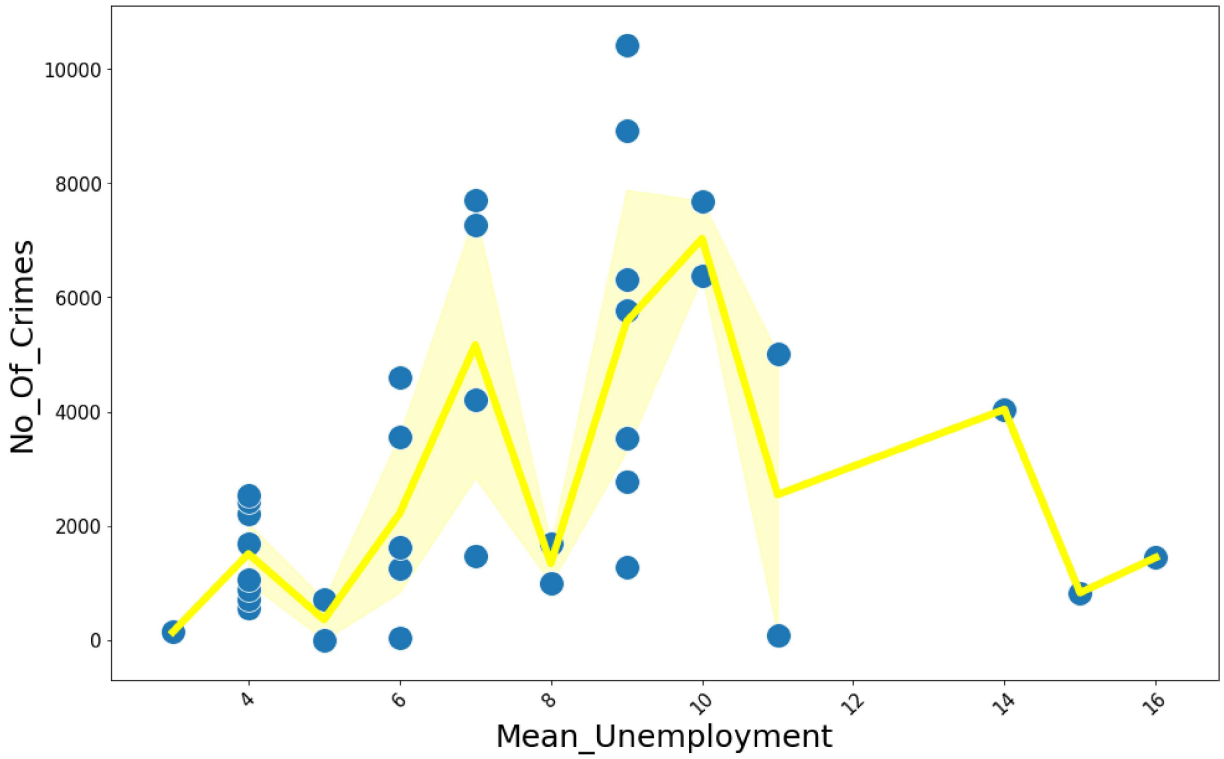
In [3]:
```
k=pd.DataFrame(df.groupby(['location','Highest_NIBRS_UCR_Offense_Description']).s
#display(k)
k.columns=['Offense Count']
k.sort_values(by='Offense Count',ascending=False,inplace=True)
k.reset_index(drop=False,inplace=True)
#display(k[k.Location == '1030 NORWOOD PARK BLVD'])
i=0
overallDTFrame= pd.DataFrame()
for each in k.location.unique():
    overallDTFrame = pd.concat([overallDTFrame,k[k.location == each][0:2]],axis=0
    i+=1
    if i ==10:
        break
overallDTFrame.reset_index(drop=True,inplace=True)
#display(overallDTFrame)
plt.figure(figsize=(16,8))
sns.barplot(y=overallDTFrame['Offense Count'],x=overallDTFrame['location'],hue=ov
plt.legend(loc='upper right',fontsize=20)
j=overallDTFrame['Offense Count'].values
k=0
s=0
elem=0
while k <len(j):
    ind=0
    each=0
    while ind < 2:
        plt.annotate(str(j[k]),xy=(elem,j[k]+0.5),fontsize=15)
        ind+=1
        k+=1
        each-=1.0
    elem+=1
plt.xticks(rotation=90,fontsize=10)
plt.savefig('Chart_1.png')
plt.show()
#overallDTFrame.plot(x='Location',y='Offense Count',kind='bar')









#j= df.groupby(['Location','Highest_NIBRS_UCR_Offense_Description']).agg({'Locati
#j.columns=['Offense Count']
#j.reset_index(drop=False,inplace=True)
#j[j.Location == '1030 NORWOOD PARK BLVD']'''
#Which location has the most crime:Highest_NIBRS_UCR_Offense_Description
#t =measure of the difference of the means relative to the spread
```



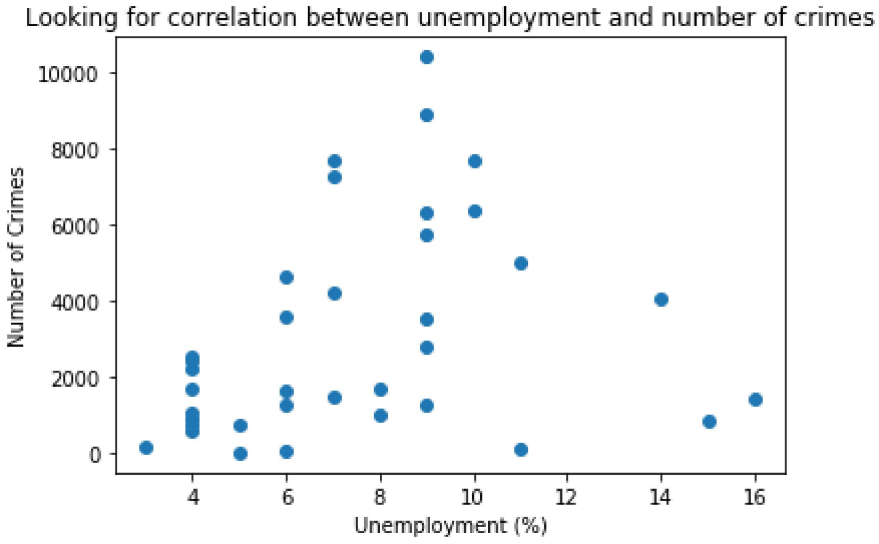
C=Cleared by Arrest O=Cleared by Exception N=Not cleared

Analysis 2: Investigate how no of crimes in particular zip-codes is proportional to the unemployment rates

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display
from scipy.stats import pearsonr as pr
dtSet = pd.read_csv('crime-housing-austin-2015.csv',usecols=['Zip_Code_Crime','Hi
dtSet.dropna(axis=0,inplace=True)
k=0
dtSet.Unemployment = dtSet.Unemployment.str.replace('%','')
dtSet.Unemployment = dtSet.Unemployment.astype(int)
#print(dtSet.dtypes)
df= dtSet.groupby('Zip_Code_Crime')
finalDT = pd.DataFrame()
for k,i in df:
    initDT = pd.concat([pd.DataFrame([k]),pd.DataFrame([i.size]),pd.DataFrame([i.
    finalDT = pd.concat([finalDT,initDT],axis=0)
finalDT.columns=['Zip_Code','No_Of_Crimes','Mean_Unemployment']
finalDT.reset_index(drop=True,inplace=True)
finalDT.sort_values(by=['No_Of_Crimes'],ascending=True,axis=0,inplace=True)
#display(finalDT)
plt.figure(figsize=(16,10))
size=np.random.randint(1,100,1)
sns.lineplot(x=finalDT['Mean_Unemployment'],y=finalDT['No_Of_Crimes'],linewidth=6
sns.scatterplot(x=finalDT['Mean_Unemployment'],y=finalDT['No_Of_Crimes'],s=size*5
plt.xticks(rotation=45,fontsize=15)
plt.yticks(fontsize=15)
plt.xlabel('Mean_Unemployment',fontsize=25)
plt.ylabel('No_Of_Crimes',fontsize=25)
plt.savefig('Chart_2.png')
plt.show()
```



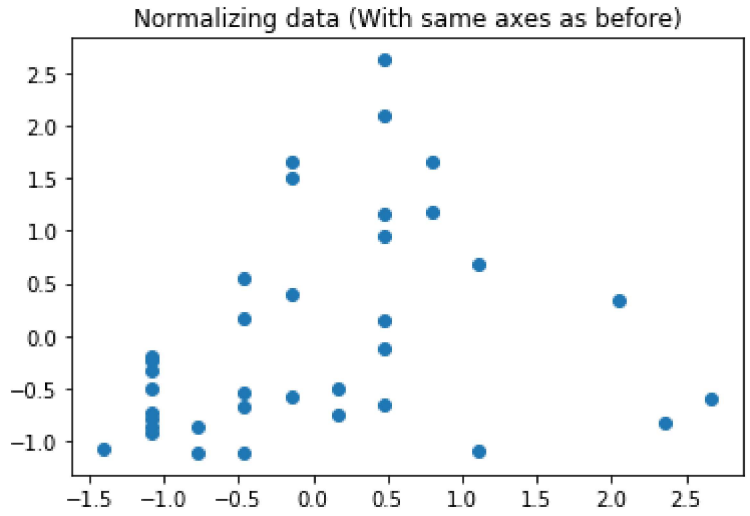
In [6]:
```
x = finalDT['Mean_Unemployment']
y = finalDT['No_Of_Crimes']
pearson_coef,p_value= pr(x,y)

plt.figure()
plt.xlabel("Unemployment (%)")
plt.ylabel("Number of Crimes")
plt.title("Looking for correlation between unemployment and number of crimes")

plt.scatter(x, y)

plt.savefig('unemploymentNumberOfCrimes.png')
plt.show()

print("Averages")
print("x: ", x.mean(), " y: ", y.mean())
print("Standard deviation")
print("x: ", x.std(), " y: ", y.std())
print('pearson_coef: ',pearson_coef, 'p_value: ',p_value)
display(finalDT.corr())

plt.figure()
plt.title("Normalizing data (With same axes as before)")
plt.scatter(stats.zscore(x), stats.zscore(y))
plt.show()
from scipy.stats import pearsonr as pr
pearson_coef,p_value= pr(finalDT['Mean_Unemployment'],finalDT['No_Of_Crimes'])
print('pearson_coef: ',pearson_coef, 'p_value: ',p_value)
#display(finalDT.corr())
plt.figure(figsize=(16,8))
sns.heatmap(finalDT.corr(),annot=True)
plt.title('Correlation Matrix',fontsize=20)
plt.show()
```



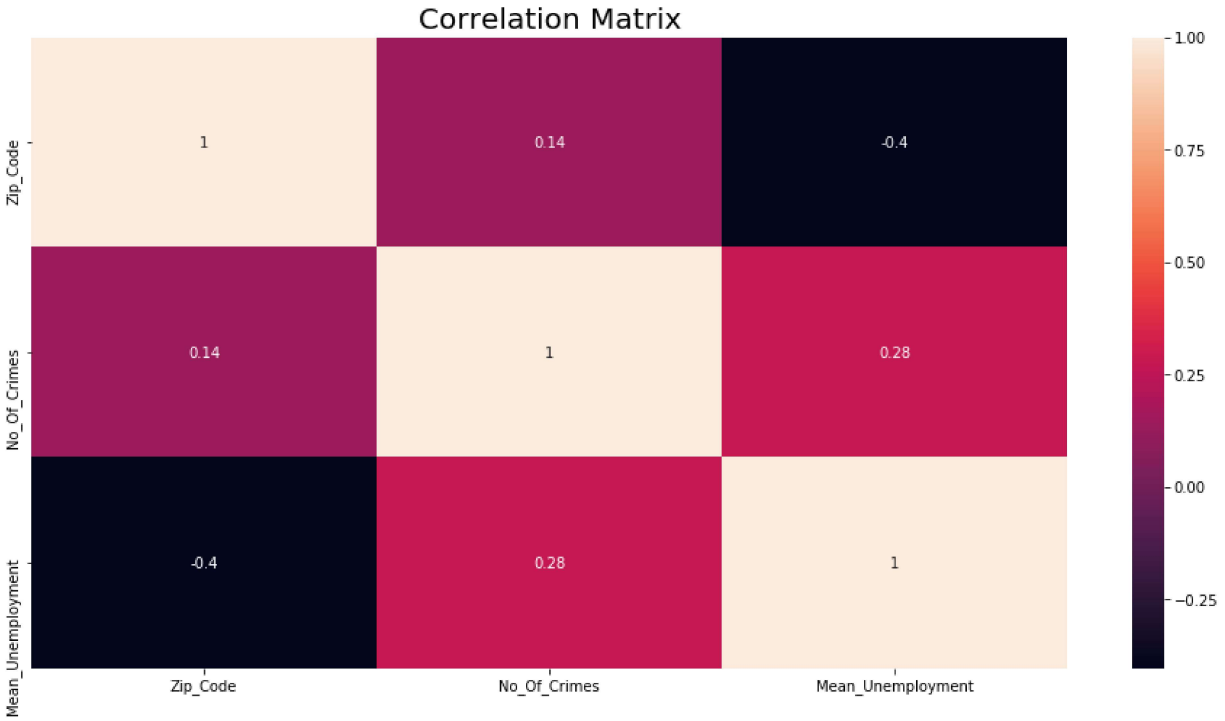

```
Averages
x:  7.472222222222222  y:  3107.0833333333335
Standard deviation
```

```
x:  3.23804271666089  y:  2811.9652240788055
pearson_coef:  0.27917473449148866 p_value:  0.09916755408711679
```

	Zip_Code	No_Of_Crimes	Mean_Unemployment
Zip_Code	1.000000	0.140904	-0.402752
No_Of_Crimes	0.140904	1.000000	0.279175
Mean_Unemployment	-0.402752	0.279175	1.000000



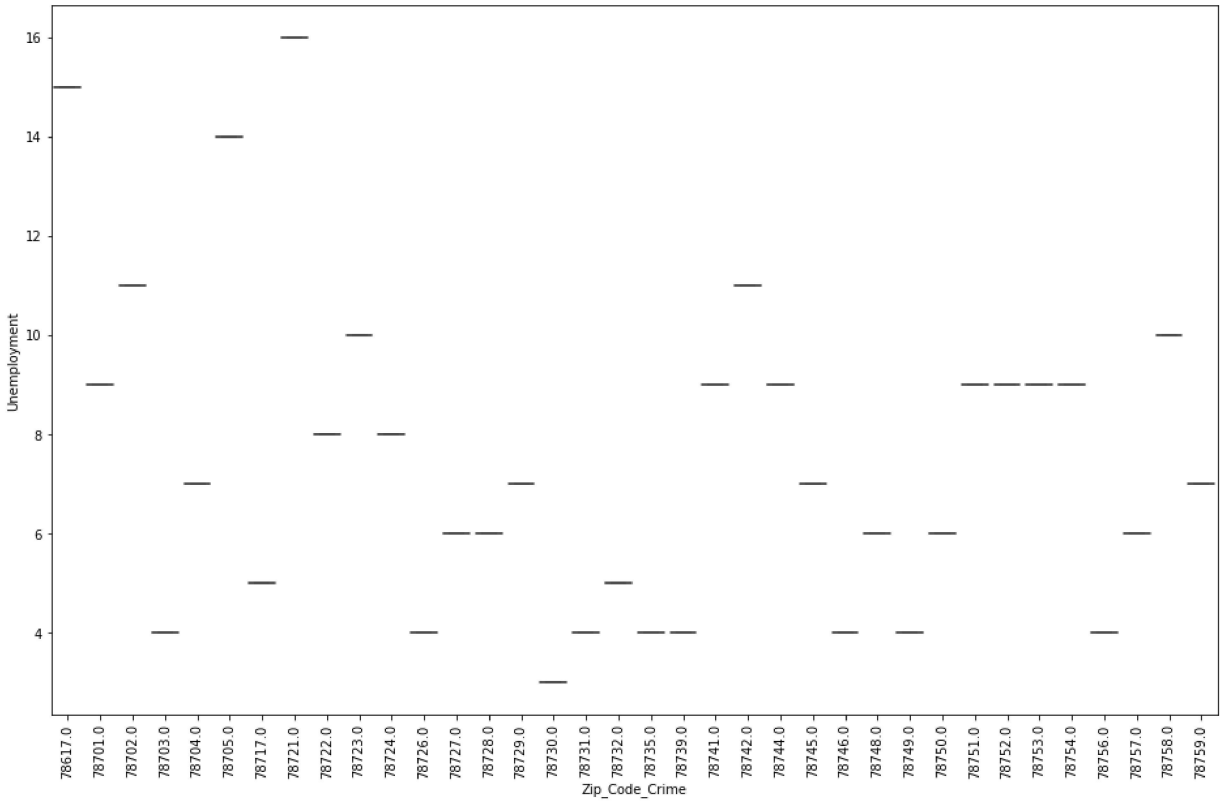

```
pearson_coef:  0.27917473449148866 p_value:  0.09916755408711679
```



As we can see that the p_value is close to 9% which is bigger than 5 % there is a 9 % chance that there the data of crimes and unemployment rate is randomly taken . Hence there is no clear evidence that there is a clear relation between no_of_crimes and mean_unemployment. Therefore , we cannot conclusively say that it is because employment that the crime numbers are growing up. There could also be other factors like kleptomania, avengeful natures,racism etc.

But lets crunch the numbers and find out if there are any outliers that's increasing the mean and eventually the p_value .

In [7]:
```
DT=pd.DataFrame()
for key,val in df:
    val = val.drop(columns=['Highest_Offense_Desc'])
    DT = pd.concat([DT,val],axis=0)
DT.reset_index(drop=True,inplace=True)
#display(DT)
plt.figure(figsize=(16,10))
sns.boxplot(x=DT['Zip_Code_Crime'],y=DT['Unemployment'],palette='spring')
plt.xticks(rotation=90)
plt.show()
```

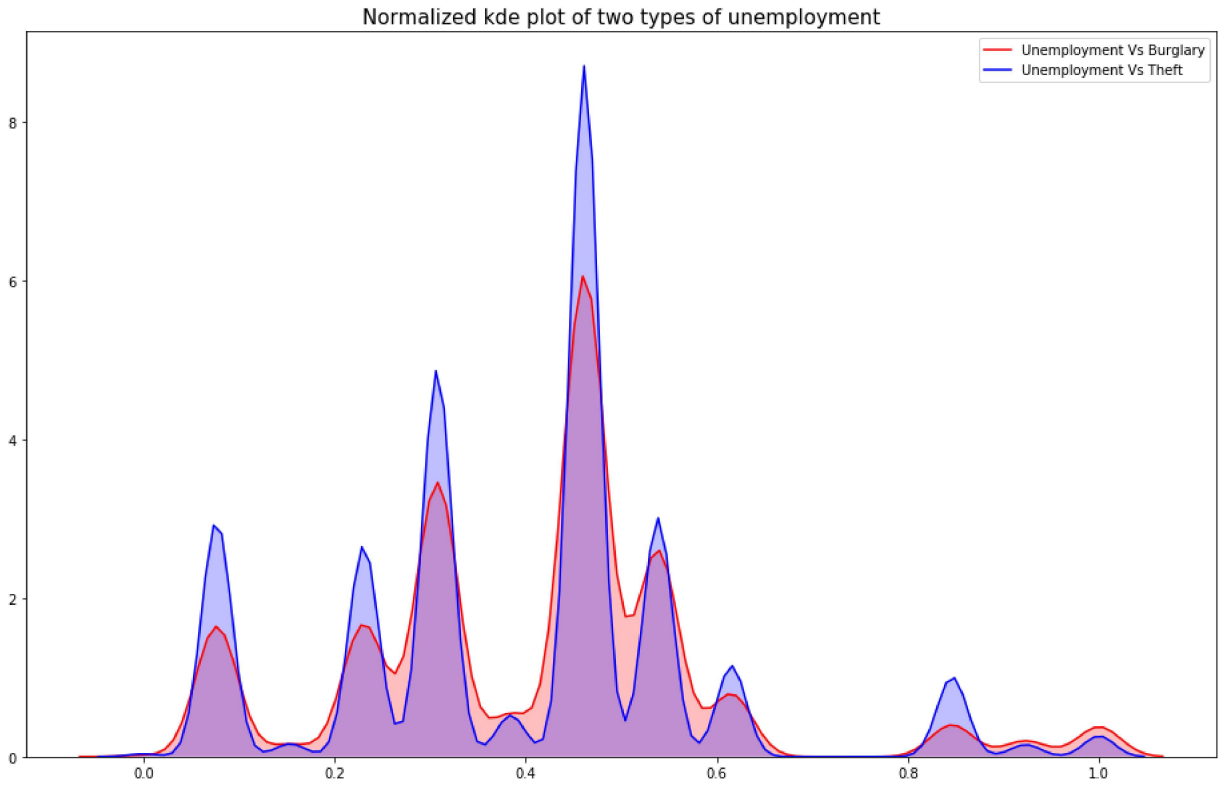


Tried to find out the outliers of the unemployment for every zip_code but eventually landed up to no outliers ,hence the hypothesis that states : rising of unemployment is causing a growth in offensive activities is not quite correct and is misleading. Hence ,we cannot say the causal effect conclusively.

Analysis 3: Which offense between theft and burglary is more related to unemployment rate?

Is there a significant difference between the unemployment rate of the two offenses(Theft and Burglary)? Or is just a fluke. Let's find out with the help of T_Test.

In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df= pd.read_csv('crime-housing-austin-2015.csv',usecols=['Highest_NIBRS_UCR_Offen
df.dropna(inplace=True)
df.head()
count = df.groupby(['Highest_NIBRS_UCR_Offense_Description'])
for key,value in count:
    if key == 'Burglary':
        burgDT = (value)
    elif key == 'Theft':
        theftDT = value
burgDT.Unemployment = burgDT.Unemployment.replace('%','',regex=True).astype(int)
burgDT.reset_index(drop=True,inplace=True)
# Normalizing the Unemployment Data
burgDT.Unemployment = (burgDT.Unemployment - burgDT.Unemployment.min())/(burgDT.U
burgDT.columns=['Highest_NIBRS_UCR_Offense_Description','Unemployment Vs Burglary
theftDT.Unemployment = theftDT.Unemployment.replace('%','',regex=True).astype(int
# Normalizing the Unemployment Data
theftDT.Unemployment = (theftDT.Unemployment - theftDT.Unemployment.min())/(theftI
theftDT.columns=['Highest_NIBRS_UCR_Offense_Description','Unemployment Vs Theft']
theftDT.reset_index(drop=True,inplace=True)
plt.figure(figsize=(16,10))
sns.kdeplot(burgDT['Unemployment Vs Burglary'],shade=True, color="r")
sns.kdeplot(theftDT['Unemployment Vs Theft'],shade=True, color="b")
plt.title('Normalized kde plot of two types of unemployment',fontsize=15)
plt.savefig('Normalized.png')
plt.show()
#display(burgDT)
from scipy.stats import ttest_ind as ti
T_Statistic_Value,P_value=ti(burgDT['Unemployment Vs Burglary'],theftDT['Unemploy
print('T_value : ',T_Statistic_Value, ' p_value: ',P_value)
```

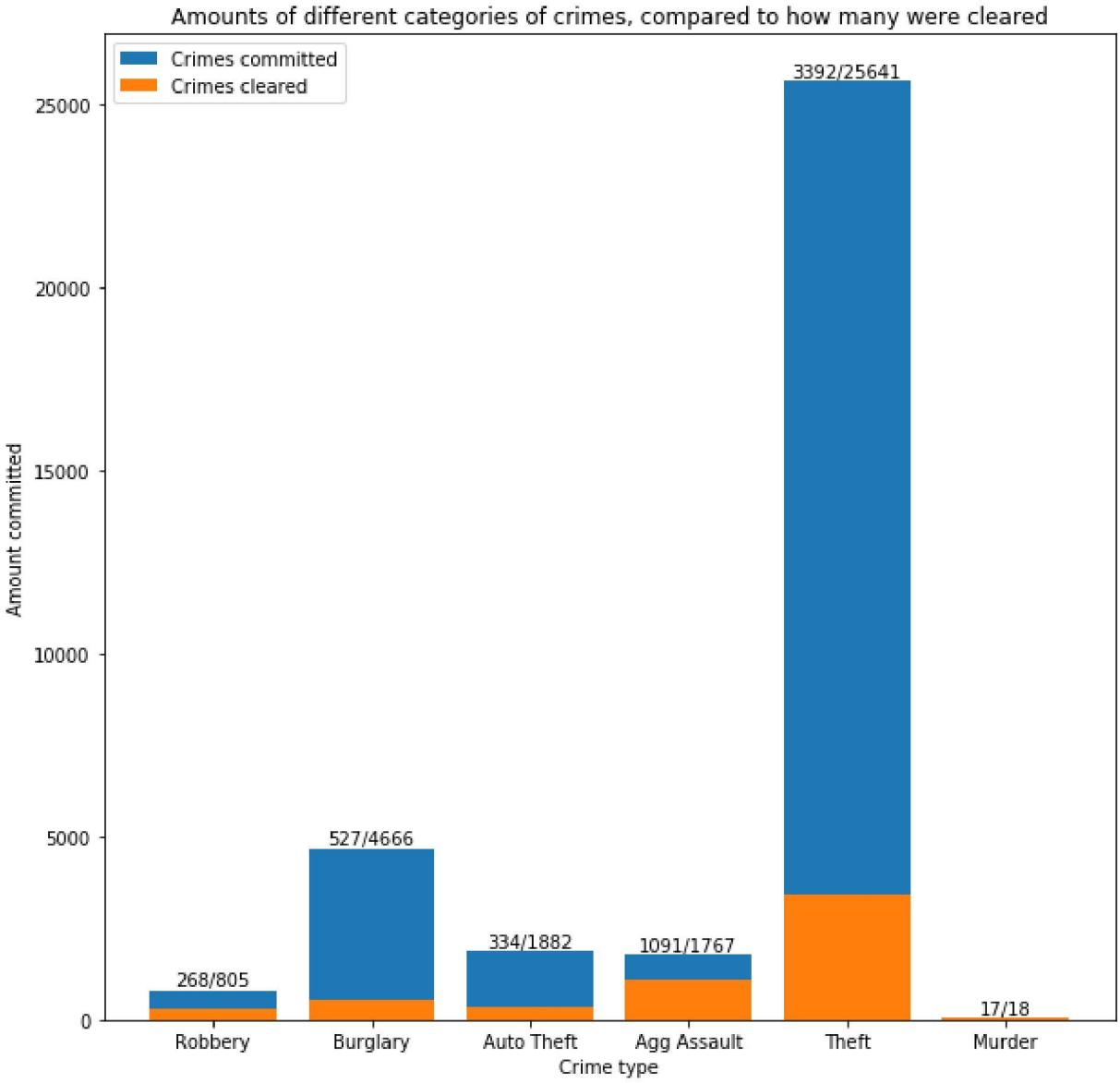



```
T_value :  5.259578252390788  p_value:  1.45307537030827e-07
```

So, from the above analysis , we come into a conclusion that the p_value for both of the two datasets is significantly low ,much lower than 0.05. p_value of 0.05 says thet there is a 5 % chance that the data is random and there is a real difference. T value relates the size of the differences. Now the p_value that we are getting is approximately 0.09% which is way lower than 5% . Henceforth we can infer conclusively that since the mean of theft is higher and also the p_value is negligible and has .09% chance of random datas,unemployment is a credible and large cause to perform theft .Whereas burglary is not so much a cause of unemployment

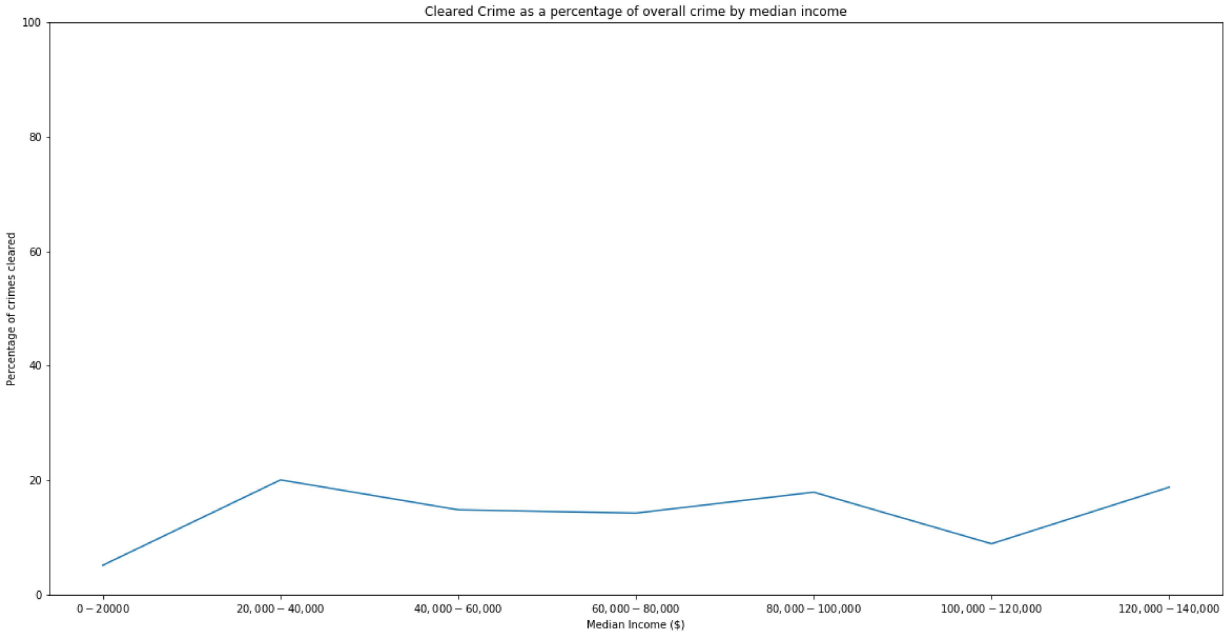
Analysis 4: Here we analyze crime clearance rate by crime type.

In [8]:
```
x = []
y = []
y2 = []
for crime, clearance in zip(reports1.Highest_NIBRS_UCR_Offense_Description, repor
    if crime in x:
        y[x.index(crime)] += 1
        y2[x.index(crime)] += int(clearance != 'N')
    else:
        y.append(1)
        y2.append(int(clearance != 'N'))
        x.append(crime)
plt.figure(figsize=(10, 10))
ax = plt.bar(x, y, width=0.8)
ax2 = plt.bar(x, y2)
plt.title("Amounts of different categories of crimes, compared to how many were c
plt.xlabel("Crime type")
plt.ylabel("Amount committed")
plt.legend(['Crimes committed', 'Crimes cleared'])


for rect, label, label2 in zip(ax.patches, y, y2):
    height = rect.get_height()
    plt.text(rect.get_x() + rect.get_width() / 2, height + 5, str(label2) + "/" +


plt.savefig('clearanceRateByCrimeType.png')
plt.show()
```



Analysis 5: Here we examine the effect that median income has on cleared crime

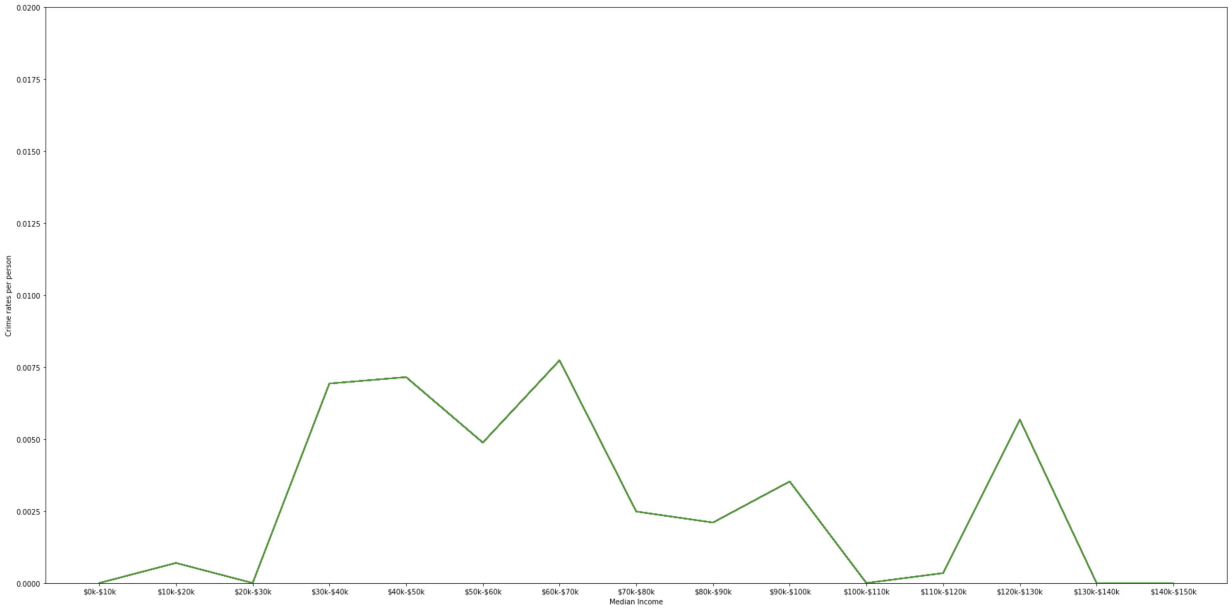
In [9]:
```
x = [0, 0, 0, 0, 0, 0, 0]
x2 = [0, 0, 0, 0, 0, 0, 0]
for income, clearance in zip(reports1.Medianhouseholdincome, reports1.Clearance_S
    x[math.trunc(income / 20000)] += 1
    x2[math.trunc(income / 20000)] += int(clearance != 'N')

plt.figure(figsize=(20, 10))

plt.title("Cleared Crime as a percentage of overall crime by median income")
plt.xlabel("Median Income ($)")
plt.ylabel("Percentage of crimes cleared")
plt.xticks([x for x in range(7)], ["$0-$20000", "$20,000-$40,000", "$40,000-$60,0
plt.ylim(0, 100)

plt.plot([(x2 / x) * 100 for x, x2 in zip(x, x2)])

plt.savefig('medianIncomeOnClearedCrime.png')
plt.show()
```



Analysis 6: Here we examine the effects of median income on crime rates per capita

We start by importing Austin wage information, and link that up with our crime data

In [13]:
```
# Data from https://datausa.io/profile/geo/austin-tx/
austin = pd.read_csv("austin-wage.csv")

x=["\$" + str(i * 10) + "k-\$" + str((i + 1) * 10) + "k" for i in range(15)]
y=[reports1[(reports1.Medianhouseholdincome >= i * 10000) & (reports1.Medianhouse

plt.figure(figsize=(30, 15))
plt.xlabel('Median Income')
plt.ylabel('Crime rates per person')
plt.ylim(0.0, 0.02)

plt.plot(
    x,
    [count / a for a, count in zip(austin.num_ppl, y)]
)

plt.savefig('medianIncomeCrimeRate.png')
plt.show()
```



In [ ]:

In [ ]: