Project 2 Report

Team members: Manish Meshram

Introduction

This analysis focuses on the most occured crime in Austin, Texas in the year 2015 and its underlying causes. As we all know a lot of times the reason behind some things are not visible directly. If we get into the root cause of any problem it is easier to solve the problem from ground up. I tried to find the relation between most problematic crime in Austin, Texas for the year 2015 and the household data given with the dataset. This analysis can be used by the government to solve the problem that we are facing.

Dataset

The dataset includes crime and household data for Austin, Texas in the year 2015. By the philosophy of going slow and tackling one problem at a time, I have chosen one kind of crime to deal with, which is the most occured crime for the year. Based on the crime, the most interesting attributes are chosen for the analysis. Data cleaning is required which includes removing NaN values from the rows and cleaning "$" and "%" sign as required. And finally in some of the analysis we needed to standardize the data. The chosen attributes for this analysis are as follows:

- Highest_NIBRS_UCR_Offense_Description
- Zip_Code_Housing
- Populationbelowpovertylevel
- Populationwithdisability
- Unemployment
- Largehouseholds(5+members)
- Homesaffordabletopeopleearninglessthan$50000
- Rentalsaffordabletopeopleearninglessthan$25000

Analysis technique

Most occured crime

There are 7 categories of crimes as per the dataset referring attribute Highest_NIBRS_UCR_Offense_Description:

- Robbery
- Burglary
- Auto Theft
- Agg Assault
- Theft
- Rape
- Murder

For finding the most problematic crime I have plotted the frequency of each category v/s the names of crimes. If there is a large difference in frequency of the first and second crime category which indeed happend in the analysis, we can focus on that specific crime category for next analysis and find out the reasons for that crime category to see if it lies in the household dataset.

Most important reasons behind the most occured crime

Since we are trying to find the reasons behind the most occured crime, so after this analysis is done we have something to work on and to solve the problem, I tried to find the most important features that contribute to "Theft"(since it is chosen as the most occured crime, refer Results section). These are the steps that I took:

1. Filtered out the rows where Highest_NIBRS_UCR_Offense_Description = "Theft" since we are only concentrating on "Theft" as of now.
2. Then I have calculated mean and standard deviation for all the selected attributes.

Inspecting selected important attributes

For this operation I have first grouped the data by zipcodes to give the frequency of thefts on that zipcode. The other two attributes that are selected from last analysis(please refer result section related to above analysis) bear same value across a unique zip code. For ex: If zipcode = "78745", all the rows with this zipcode will have same values for *Populationbelowpovertylevel*, *Homesaffordabletopeopleearninglessthan$50000* etc. which helped in grouping. The final dataframe had these columns:

1. *ZipCode* - Zip code in Austin, Texas
2. *TheftCount* - The occurances of thefts for the given zip code in year 2015
3. *Populationbelowpovertylevel* - Population below poverty level for given zip code (in %)
4. *Homesaffordabletopeopleearninglessthan$50000* - Homes affordable to people earning less than $50000 for the given zip code (in %)

After organizing this data I have plotted the scatter plots to see if there is any visible correlation between

- TheftCount and Populationbelowpovertylevel
- TheftCount and Homesaffordabletopeopleearninglessthan$50000

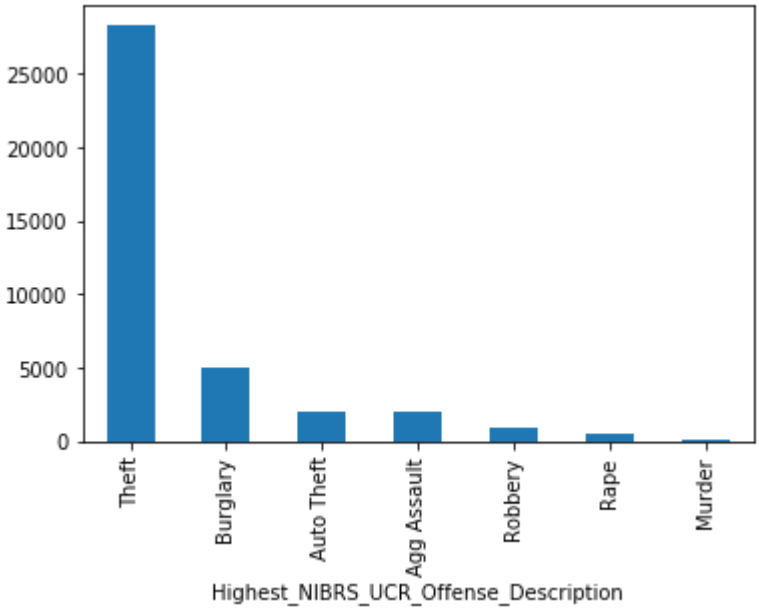
I have also standardized the data to see if there is any improvement on correlation. In addition to this I ran a pearson correlation test to check the randomness of data and make comment on correlation if any.

Comparing the distributions of two selected attributes

For this analysis I have chosen the two most important attributes so far viz. *Populationbelowpovertylevel* and *Homesaffordabletopeopleearninglessthan$50000* to see if there is any significant difference in their distributions. I have used t-test for the same. I also did a kde plot just to visualize the difference in their mean values.

Results

Most occured crime



As per the analysis we can see that **"Theft"** is the most occured crime in the year 2015, infact there is a huge gap between "Theft" and "Burglary". It will be really useful if we can get the reason behind why people steal based on the household dataset given. Most of the times theft happens due to lack of money, keeping that in mind I have selected these columns to work with in futher analysis:

- Populationbelowpovertylevel
- Populationwithdisability
- Unemployment
- Largehouseholds(5+members)
- Homesaffordabletopeopleearninglessthan$50000
- Rentalsaffordabletopeopleearninglessthan$25000

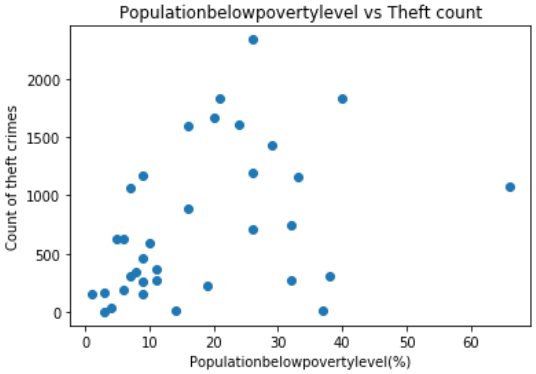
Most important reasons behind the most occured crime

The mean and standard deviations of all the selected attributes are as shown below:

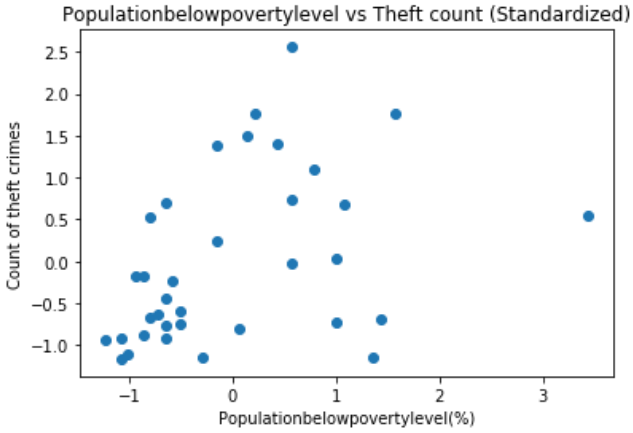
Attribue	Mean	Standard Deviation
Populationbelowpovertylevel	22.828399828399828	13.436202985687384
Populationwithdisability	9.000663000663002	2.280427717436714
Unemployment	8.148122148122148	6.020988100356848
Largehouseholds(5+members)	8.166296166296167	6.020988100356848
Homesaffordabletopeopleearninglessthan$50000	37.5982215982216	28.83318859210238
Rentalsaffordabletopeopleearninglessthan$25000	11.481806481806482	8.575010319427765

From the table we can see that ***Populationbelowpovertylevel*** and ***Homesaffordabletopeopleearninglessthan$50000*** seems to cover more variance than other attributes. Further analysis are done on these attributes.

Inspecting selected important attributes

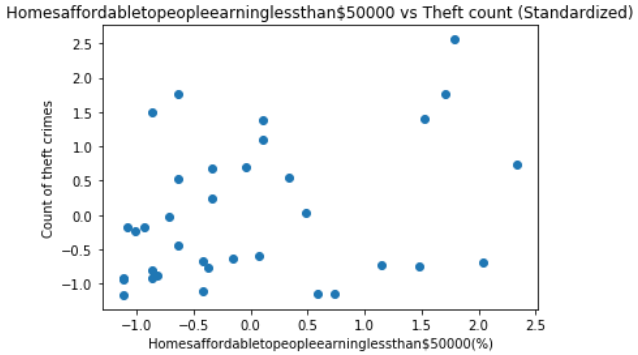


Pearson correlation coefficient: 0.399363685513
p value: 0.0174679610264



Pearson correlation coefficient: 0.399363685513
p value: 0.0174679610264

Pearson correlation coefficient: 0.327312033345
p value: 0.0549435006369



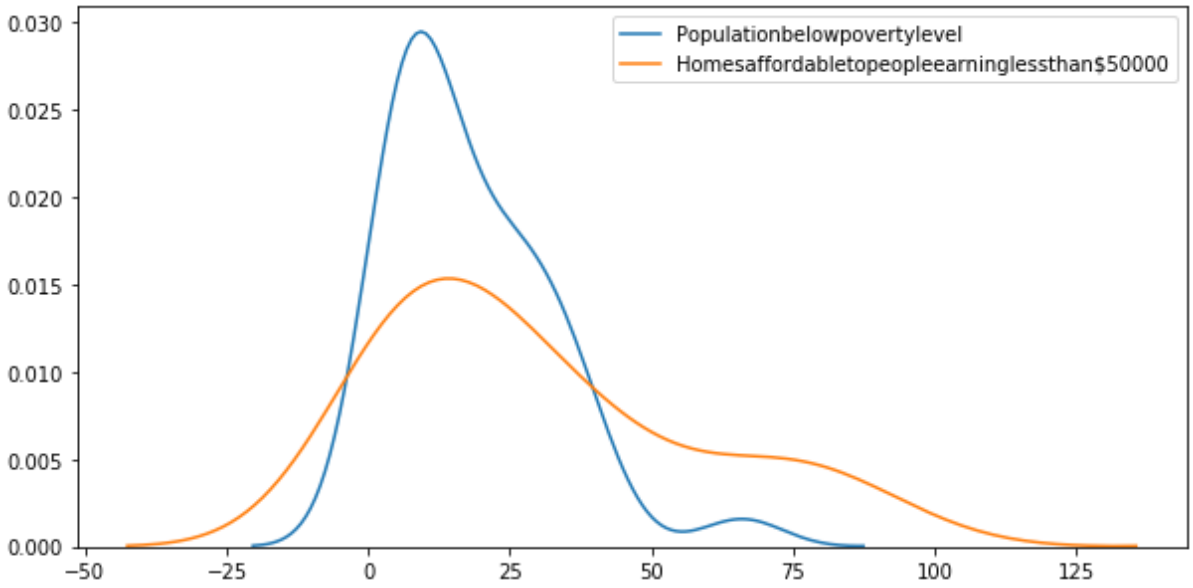
Pearson correlation coefficient: 0.327312033345
p value: 0.0549435006369

From the above graphs and pearson test results we can infer that:

- **For Populationbelowpovertylevel vs TheftCount:** Test result shows the p-value of 0.01 which denotes that there is 1% chance of data being randomly taken (it is less than the typical threshold of 0.05) which implies there is a significant relation between Populationbelowpovertylevel and TheftCount. Also Pearson correlation coefficient value of 0.39 shows it is a good correlation. This infers that Population below poverty level indeed affects the Theft crimes of a location.
- **For Homesaffordabletopeopleearninglessthan$50000 vs TheftCount:** Test result shows the p-value of 0.05 which denotes there is 5% chance of data being randomly taken which is high. Thus we cannot conclusively say that there is a relation between Homesaffordabletopeopleearninglessthan$50000 and TheftCount. In other words we are not sure if Homes affordable to peopl eearning less than $50000 for a particular location contributes in any way to the Theft crimes in that location.

Observation: Standardization didn't change the scatter plots or the pearson correlation coefficient values in both the cases.

Comparing the distributions of two selected attributes



T value: 2.31150686279
p value: 0.023841935982

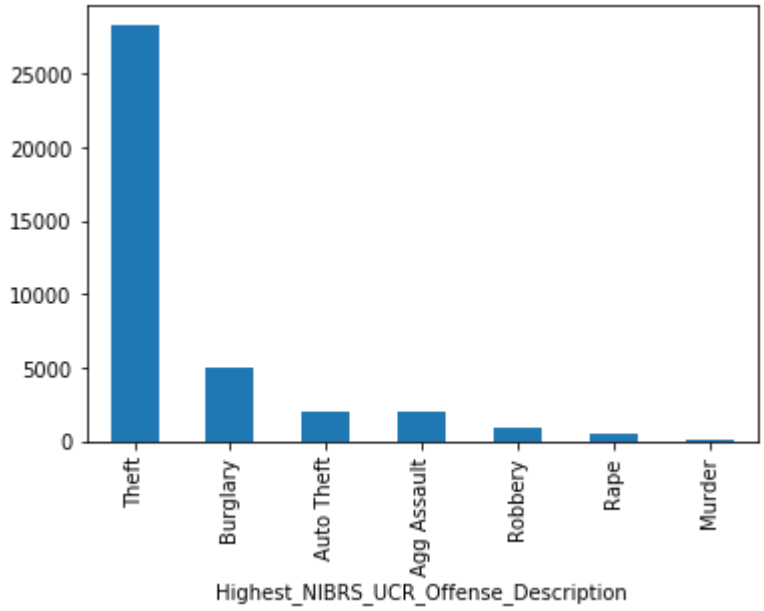
Since p-value for this test is significantly low i.e. 0.02 we can infer that there is a significant difference between these two data attributes. Also as we can see from the graph and T value that the ***Populationbelowpovertylevel*** has higher mean as compared to ***Homesaffordabletopeopleearninglessthan$50000***.

Project 2 Code

Selecting the most occured crime

In [67]:
```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('crime-housing-austin-2015.csv')
crime_count_series = df.groupby('Highest_NIBRS_UCR_Offense_Description').Highest_
crime_count_series.sort_values(ascending=False, inplace=True)
%matplotlib inline
crime_count_series.plot.bar()

### Theft is the biggest problem, lets find out the reasons behind it
```

Out[67]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x246c35c7240>
```



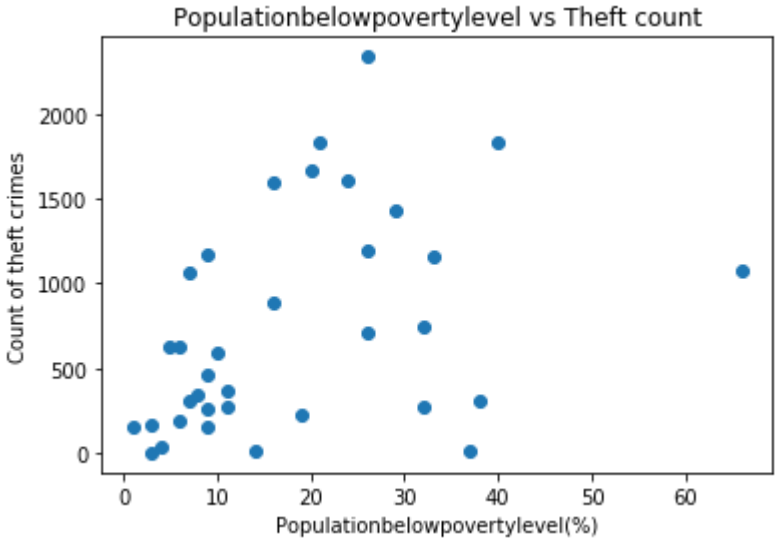
Data Cleaning and calculating standard deviation and mean for attributes

In [186]:
```
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

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

# Choosing only "Theft" crimes
df = df.loc[df['Highest_NIBRS_UCR_Offense_Description'] == 'Theft']

# Dropping NaN values
df = df.dropna()

# Selecting 9 attributes that I found most interesting and saving it to new datafr
# All futher operations will be on this list
zip_column = df[['Zip_Code_Crime']]
df = df[['Populationbelowpovertylevel','Medianhouseholdincome','Populationwithdis

# Cleaning percentages and dollar signs
df['Populationbelowpovertylevel'] = df['Populationbelowpovertylevel'].str.replace
df['Medianhouseholdincome'] = df['Medianhouseholdincome'].str.replace('$', '').as
df['Populationwithdisability'] = df['Populationwithdisability'].str.replace('%',
df['Unemployment'] = df['Unemployment'].str.replace('%', '').astype('float')
df['Largehouseholds(5+members)'] = df['Largehouseholds(5+members)'].str.replace('
df['Homesaffordabletopeopleearninglessthan$50000'] = df['Homesaffordabletopeoplee
df['Rentalsaffordabletopeopleearninglessthan$25000'] = df['Rentalsaffordabletopeo
df['Medianrent'] = df['Medianrent'].str.replace('$', '').astype('float')
df['Averagemonthlytransportationcost'] = df['Averagemonthlytransportationcost'].s


print("Populationbelowpovertylevel:", df['Populationbelowpovertylevel'].mean(), d
print("Populationwithdisability:", df['Populationwithdisability'].mean(), df['Pop
print("Unemployment:", df['Unemployment'].mean(), df['Unemployment'].std())
print("Largehouseholds(5+members):", df['Largehouseholds(5+members)'].mean(), df[
print("Homesaffordabletopeopleearninglessthan$50000:", df['Homesaffordabletopeopl
print("Rentalsaffordabletopeopleearninglessthan$25000:", df['Rentalsaffordabletop

# adding zip_column
df = pd.concat([zip_column, df], axis=1)

# selecting very few final columns
df_min = df.groupby(['Zip_Code_Crime', 'Populationbelowpovertylevel', 'Homesaffor
```

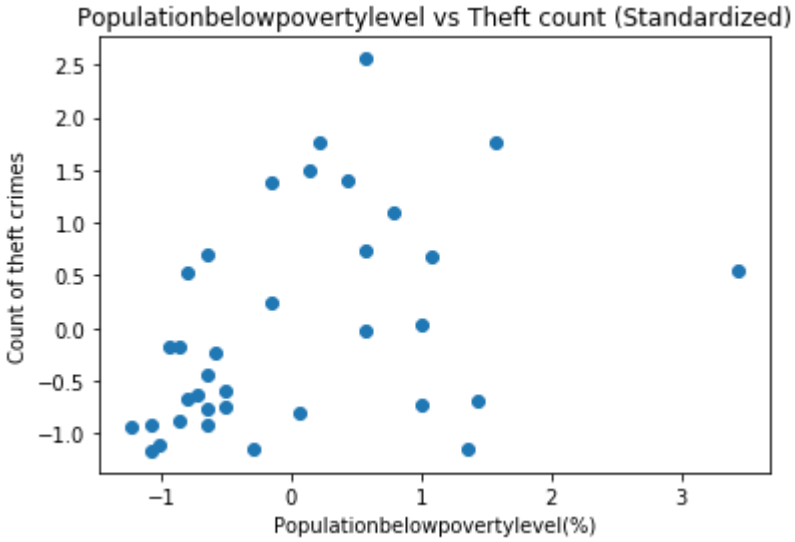
```
Populationbelowpovertylevel: 22.828399828399828 13.436202985687384
Populationwithdisability: 9.000663000663002 2.280427717436714
Unemployment: 8.148122148122148 2.4508223339071495
Largehouseholds(5+members): 8.166296166296167 6.020988100356848
Homesaffordabletopeopleearninglessthan$50000: 37.5982215982216 28.8331885921023
8
Rentalsaffordabletopeopleearninglessthan$25000: 11.481806481806482 8.5750103194
27765
```

Finding the relation between attributes (Scatterplots and Pearson correlation coefficient test)

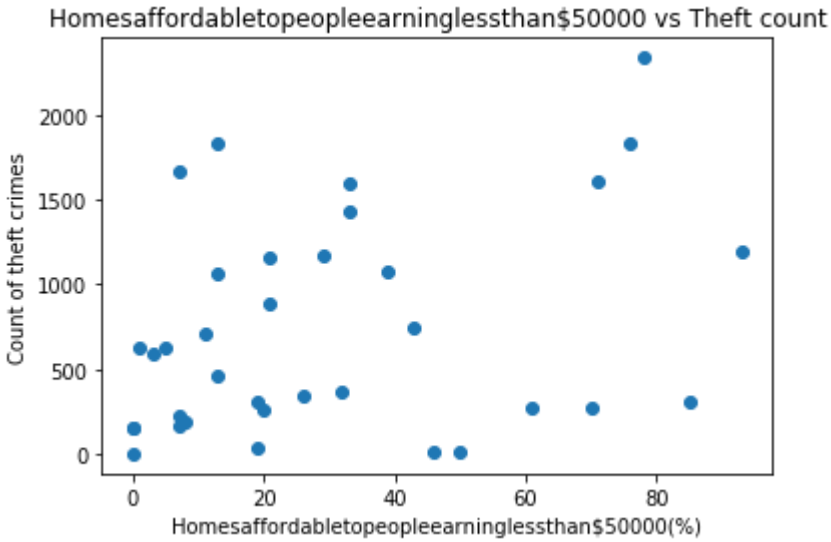
In [187]:
```
import matplotlib.pyplot as plt
from scipy.stats import pearsonr as pr

###### Populationbelowpovertylevel vs Theft count #######

plt.scatter(df_min['Populationbelowpovertylevel'], df_min['Size'])
plt.title("Populationbelowpovertylevel vs Theft count")
plt.xlabel("Populationbelowpovertylevel(%)")
plt.ylabel("Count of theft crimes")
plt.show()
pearson_r, p_value = pr(df_min['Populationbelowpovertylevel'], df_min['Size'])
print("Pearson correlation coefficient: ", pearson_r, "\np value:", p_value)

plt.scatter(stats.zscore(df_min['Populationbelowpovertylevel']), stats.zscore(df_
plt.title("Populationbelowpovertylevel vs Theft count (Standardized)")
plt.xlabel("Populationbelowpovertylevel(%)")
plt.ylabel("Count of theft crimes")
plt.show()
pearson_r, p_value = pr(stats.zscore(df_min['Populationbelowpovertylevel']), stat
print("Pearson correlation coefficient: ", pearson_r, "\np value:", p_value)


############ Homesaffordabletopeopleearninglessthan$50000 vs Theft count ########

plt.scatter(df_min['Homesaffordabletopeopleearninglessthan$50000'], df_min['Size'
plt.title("Homesaffordabletopeopleearninglessthan$50000 vs Theft count")
plt.xlabel("Homesaffordabletopeopleearninglessthan$50000(%)")
plt.ylabel("Count of theft crimes")
plt.show()
pearson_r, p_value = pr(df_min['Homesaffordabletopeopleearninglessthan$50000'], d
print("Pearson correlation coefficient: ", pearson_r, "\np value:", p_value)

plt.scatter(stats.zscore(df_min['Homesaffordabletopeopleearninglessthan$50000']),
plt.title("Homesaffordabletopeopleearninglessthan$50000 vs Theft count (Standardi
plt.xlabel("Homesaffordabletopeopleearninglessthan$50000(%)")
plt.ylabel("Count of theft crimes")
plt.show()
pearson_r, p_value = pr(stats.zscore(df_min['Homesaffordabletopeopleearninglesstha
print("Pearson correlation coefficient: ", pearson_r, "\np value:", p_value)
```



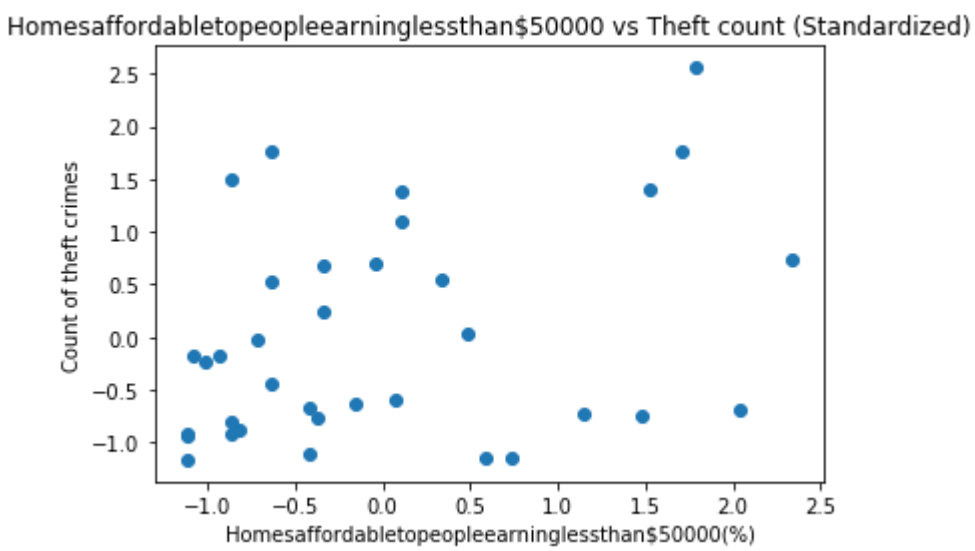
```
Pearson correlation coefficient:  0.399363685513
p value: 0.0174679610264
```




```
Pearson correlation coefficient:  0.399363685513
p value: 0.0174679610264
```

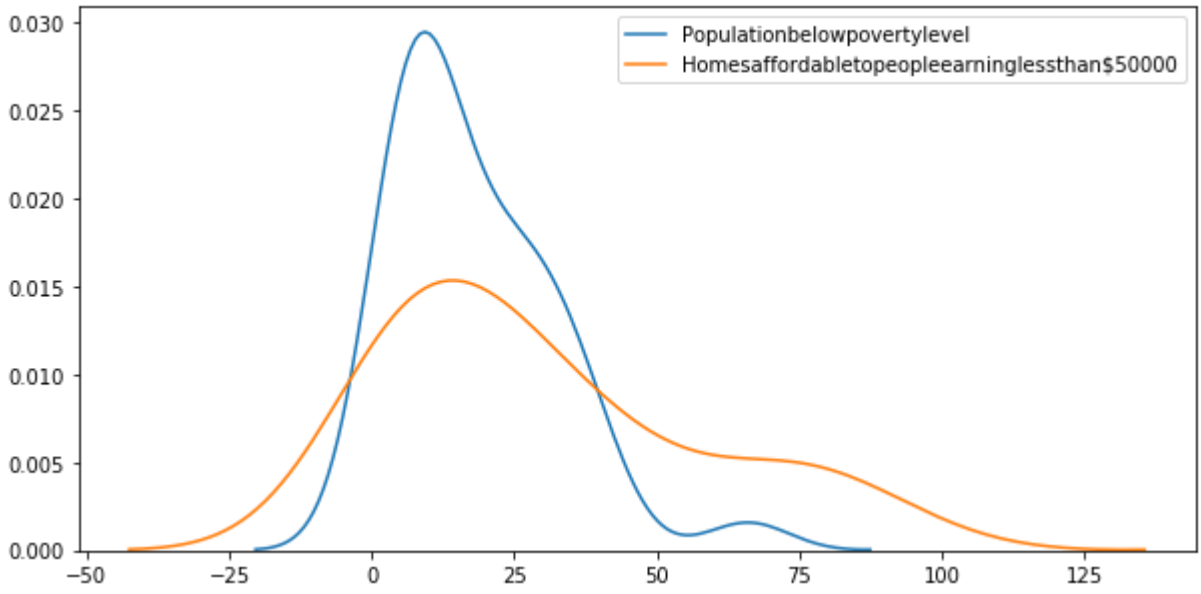

```
Pearson correlation coefficient:  0.327312033345
p value: 0.0549435006369
```




```
Pearson correlation coefficient:  0.327312033345
p value: 0.0549435006369
```

Comparing the distributions (T Test)

In [203]:
```
from scipy.stats import ttest_ind as tt
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10,5))
sns.kdeplot(df_min['Populationbelowpovertylevel'])
sns.kdeplot(df_min['Homesaffordabletopeopleearninglessthan$50000'])
plt.show()

t_value, p_value = tt(df_min['Homesaffordabletopeopleearninglessthan$50000'], df_
print("T value: ", t_value, "\np value:", p_value)
```




```
T value:  2.31150686279
p value: 0.023841935982
```