

Data Science with Python

USA Statistics

Group 3 - Team Members

Jayalakshmi Vaidyanathan
Krutika Ambavane
Neha Narayankar

Importing all the required libraries:

```
In [162]: import pandas as pd
import numpy as np
import seaborn as sns
%pylab inline
import sklearn as sk
import sklearn.tree as tree
from IPython.display import Image
import pydotplus
import plotly.plotly as py
from plotly.graph_objs import *
pd.set_option('display.max_columns', None)
pd.set_option('precision', 2)
from plotly.offline import init_notebook_mode, iplot, plot
import plotly.graph_objs as go
init_notebook_mode(connected=True)
import plotly.tools as tls
import plotly
import nbconvert
```

Populating the interactive namespace from numpy and matplotlib

```
In [163]: # print all the outputs in a cell
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
import warnings
warnings.filterwarnings("ignore")
```

Data set Description

- **df** - USA Real Estate statistic dataset that contains information on Mortgage-Backed Securities, Geographic Business Investment, Real Estate Analysis
- **df2** - Census information dataset for each census tract in the USA
- **df_cen** - Census information dataset aggregated for each state in USA
- **df3** - USA Real Estate statistic dataset aggregated for each state in USA
- **df_all** - Merged dataset of of USA Census(df_cen) and USA Real Estate Statistics(df3)

The columns in df - USA statistics dataset are:

1. State : State reported by the U.S. Census Bureau
2. state_ab : State Abbreviation
3. city : City Name
4. place : The place name reported by the U.S. Census Bureau for the specified geographic location.
5. type : The place Type reported by the U.S. Census Bureau for the specified geographic location
6. primary : Defines whether the location is a tract location or a block group.
7. zip_code : The closest zip code reported by the U.S. Education Department by the closest school relative to the Census Location
8. area_code : The area code reported by the U.S. Census Bureau of the closest geographic location with area code information
9. lat : The latitude of geographic location
10. lng : The longitude of geographic location
11. ALand : The Square area of land at the geographic or track location
12. AWater : The Square area of water at the geographic or track location
13. pop : Male and female population of geographic location
14. male_pop : Male population of geographic location
15. female_pop : female population of geographic location
16. rent_mean : The mean gross rent of the specified geographic location
17. rent_median : The mean gross rent of the specified geographic location
18. rent_stdev : The standard deviation of the gross rent for the specified geographic location
19. rent_sample_weight : The sum of gross rent weight used in calculations
20. rent_samples : The number of gross rent records used in the statistical calculations
21. rent_gt_10 : The empirical distribution value that an individual's rent will be greater than 10% of their household income in the past 12 months
22. rent_gt_15 : The empirical distribution value that an individual's rent will be greater than 15% of their household income in the past 12 months
23. rent_gt_20 : The empirical distribution value that an individual's rent will be greater than 20% of their household income in the past 12 months
24. rent_gt_25 : The empirical distribution value that an individual's rent will be greater than 25% of their household income in the past 12 months
25. rent_gt_30 : The empirical distribution value that an individual's rent will be greater than 30% of their household income in the past 12 months
26. rent_gt_35 : The empirical distribution value that an individual's rent will be greater than 35% of their household income in the past 12 months
27. rent_gt_40 : The empirical distribution value that an individual's rent will be greater than 40% of their household income in the past 12 months

28. **rent_gt_50** : The empirical distribution value that an individual's rent will be greater than 50% of their household income in the past 12 months
29. **universe_samples** : The size of the renter-occupied housing units sampled universe for the calculations
30. **used_samples** : The number of samples used in the household income by gross rent as percentage of income in the past 12 months calculation
31. **hi_mean** : The mean household income of the specified geographic location
32. **hi_median** : The median household income of the specified geographic location
33. **hi_stdev** : The standard deviation of the household income for the specified geographic location.
34. **hi_sample_weight** : The number of households weighted used in the statistical calculations
35. **hi_samples** : The number of households used in the statistical calculations
36. **family_mean** : The mean family income of the specified geographic location
37. **family_median** : The median family income of the specified geographic location
38. **family_stdev** : The standard deviation of the family income for the specified geographic location.
39. **family_sample_weight** : The number of family income weighted used in the statistical calculations
40. **family_samples** : The number of family income used in the statistical calculations
41. **hc_mortgage_mean** : The mean Monthly Mortgage and Owner Costs of specified geographic location
42. **hc_mortgage_median** : The median Monthly Mortgage and Owner Costs of the specified geographic location.
43. **hc_mortgage_stdev** : The standard deviation of the Monthly Mortgage and Owner Costs for a specified geographic location.
44. **hc_mortgage_sample_weight** : The number of samples used in the statistical calculations
45. **hc_mortgage_samples** : The number of samples used in the statistical calculations
46. **hc_mean** : The mean Monthly Owner Costs of specified geographic location
47. **hc_median** : The median Monthly Owner Costs of a specified geographic location
48. **hc_stdev** : The standard deviation of the Monthly Owner Costs of a specified geographic
49. **hc_samples** : The samples used in the calculation of the Monthly Owner Costs statistics
50. **hc_sample_weight** : The samples used in the calculation of the Monthly Owner Costs statistics
51. **home_equity_second_mortgage** : Percentage of homes with a second mortgage and home equity loan
52. **second_mortgage** : percent of houses with a second mortgage
53. **home_equity** : Percentage of homes with a home equity loan.
54. **debt** : Percentage of homes with some type of debt
55. **second_mortgage_cdf** : Cumulative distribution value of one minus the percentage of homes with a second mortgage. The value is used as a performance feature
56. **home_equity_cdf** : Cumulative distribution value of one minus the percentage of homes with a home equity loan. The value is used as a performance feature
57. **debt_cdf** : Cumulative distribution value of one minus the percentage of homes with any home related debt. The value is used as a performance feature.
58. **hs_degree** : Percentage of people with at least high school degree
59. **hs_degree_male** : Percentage of males with at least high school degree

- 60. **hs_degree_female** : Percentage of females with at least high school degree
- 61. **male_age_mean** : The mean male age of specified geographic location
- 62. **male_age_median** : The median male age of specified geographic location
- 63. **male_age_stdev** : The standard male age of specified geographic location
- 64. **male_age_sample_weight** : The samples used in the calculation of the male age of specified geographic location
- 65. **male_age_samples** : The samples used in the calculation of the male age of specified geographic location
- 66. **female_age_mean** : The mean female age of specified geographic location
- 67. **female_age_median** : The median female age of specified geographic location
- 68. **female_age_stdev** : The standard female age of specified geographic location
- 69. **female_age_sample_weight** : The samples used in the calculation of the female age of specified geographic location
- 70. **female_age_samples** : The samples used in the calculation of the female age of specified geographic location
- 71. **pct_own** : Percentage of Owners
- 72. **married** : Percentage of people married
- 73. **married_snp** : Percentage of people married snp
- 74. **separated** : Percentage of people seperated
- 75. **divorced** : Percentage of people divorced

The columns in df2 - USA Census Dataset are:

- 1. State : State, DC, or Puerto Rico, String**
- 2. County: County or county equivalent , String**
- 3. TotalPop: Total population, Numeric**
- 4. Men: Number of men, Numeric**
- 5. Women: Number of women, Numeric**
- 6. Hispanic: % of population that is Hispanic/Latino, Numeric**
- 7. White: % of population that is white, Numeric**
- 8. Black: % of population that is black, Numeric**
- 9. Native: % of population that is Native American or Native Alaskan, Numeric**
- 10. Asian: % of population that is Asian, Numeric**
- 11. Pacific: % of population that is Native Hawaiian or Pacific Islander, Numeric**
- 12. Citizen: Number of citizens, Numeric**
- 13. Income: Median household income (dollars), Numeric**
- 14. IncomeErr: Median household income error (dollars), Numeric**
- 15. IncomePerCap: Income per capita (dollars), Numeric**
- 16. IncomePerCapErr: Income per capita error (dollars), Numeric**
- 17. Poverty: % under poverty level, Numeric**
- 18. ChildPoverty: % of children under poverty level, Numeric**
- 19. Professional: % employed in management, business, science, and arts, Numeric**
- 20. Service: % employed in service jobs, Numeric**
- 21. Office: % employed in sales and office jobs, Numeric**
- 22. Construction: % employed in natural resources, construction, and maintenance, Numeric**
- 23. Production: % employed in production, transportation, and material movement, Numeric**
- 24. Drive: % commuting alone in a car, van, or truck, Numeric**
- 25. Carpool: % carpooling in a car, van, or truck, Numeric**
- 26. Transit: % commuting on public transportation, Numeric**
- 27. Walk: % walking to work, Numeric**
- 28. OtherTransp: % commuting via other means, Numeric**
- 29. WorkAtHome: % working at home, Numeric**
- 30. MeanCommute: Mean commute time (minutes), Numeric**
- 31. Employed: Number employed (16+), Numeric**
- 32. PrivateWork: % employed in private industry, Numeric**
- 33. PublicWork: % employed in public jobs, Numeric**
- 34. SelfEmployed: % self-employed, Numeric**

35. FamilyWork: % in unpaid family work, Numeric

36. Unemployment: Unemployment rate (%). Numeric

Reading the dataset and creating a dataframe for it:

```
In [3]: df = pd.read_csv("real_estate_db.csv", encoding = 'latin-1')
```

```
In [4]: print(df.shape)
```

```
(39030, 79)
```

Cleaning the dataset:

Lets remove the unnecessary columns that we aren't going to analyze on.

```
In [5]: df_useful = df.drop([col for col in df if ('hc_' in col ) or ('sample_weight' in col) or ('ALand' in col) or  
                             ('BLOCKID' in col) or ('cdf' in col) or ('stdev' in col) or ('median' in col) or  
                             ('pct' in col) or ('gt_2' in col) or ('gt_4' in col) or ('gt_35' in col) or  
                             ('gt_15' in col) or ('SUMLEVEL' in col) or ('_snp' in col) or (col.startswith('u'))], axi  
s=1)
```

Handling the NaNs:

```
In [6]: df_useful.fillna(0, inplace = True)
```

Converting to lower case:

```
In [7]: df_useful.columns = [x.lower() for x in df_useful.columns]
```

Removing outliers from required columns:

```
In [8]: h = df_useful["hi_mean"].quantile(0.99)
```

```
In [9]: df_useful = df_useful[df_useful["hi_mean"] < h]
```

```
In [10]: r = df_useful["rent_mean"].quantile(0.99)
```

```
In [11]: df_useful = df_useful[df_useful["rent_mean"] < r]
```

```
In [12]: m = df_useful["married"].quantile(0.99)
```

```
In [13]: df_useful = df_useful[df_useful["married"] < m]
```

```
In [14]: p = df_useful["pop"].quantile(0.99)
```

```
In [15]: df_useful = df_useful[df_useful["pop"] < p]
```

```
In [17]: print(df_useful.shape)
```

```
(37490, 39)
```

Lets visualize an important feature - Type:


```
In [18]: percents = df_useful["type"].value_counts().round(2)

print("Type Count Values: ")
print(percents)

types = df_useful["type"].value_counts() / len(df_useful["type"]) * 100

labels = types.index.values.tolist()
values = types.tolist()

trace1 = go.Pie(labels=labels, values=values, hoverinfo='label+percent', marker=dict(line=dict(color='#000000', width=1)))

layout = go.Layout(title='Distribution of Types', legend=dict(orientation="h"));

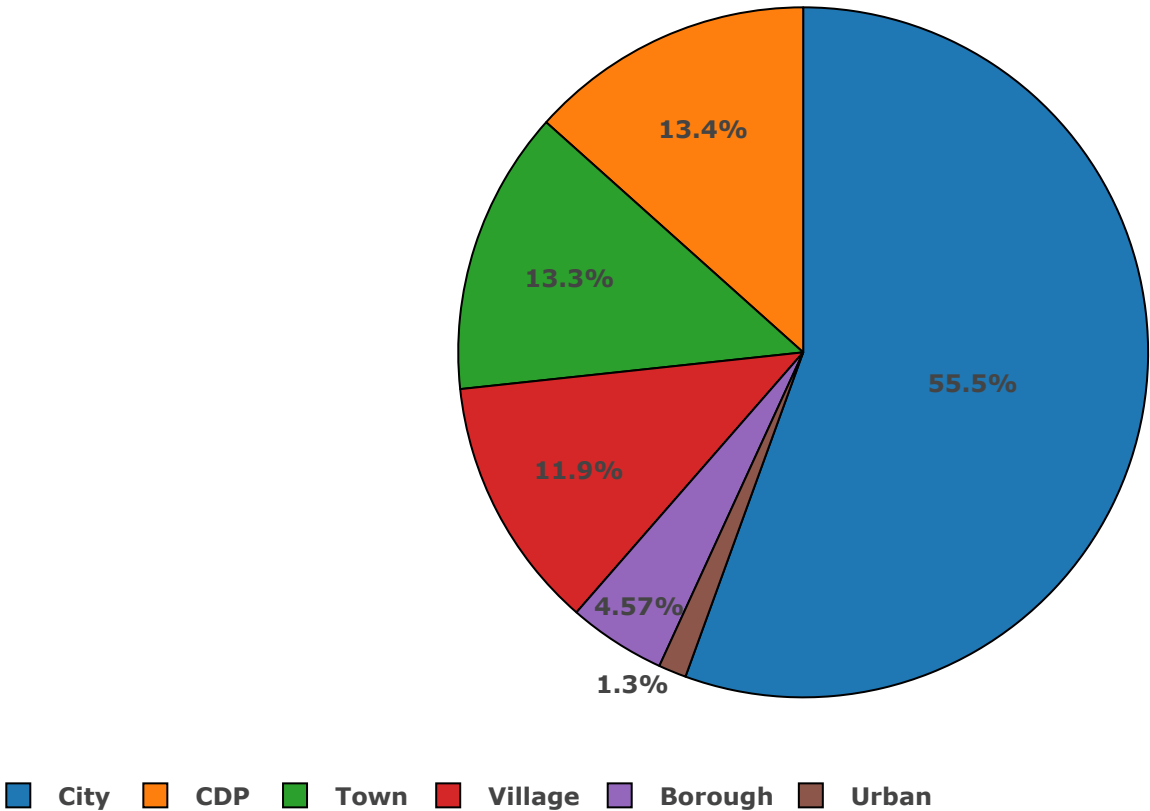
fig = go.Figure(data=[trace1], layout=layout)
iplot(fig)
```

Type Count Values:

City	20823
CDP	5016
Town	4994
Village	4456
Borough	1713
Urban	488

Name: type, dtype: int64

Distribution of Types



As we can see we have major rows of cities and towns so lets take a new dataset containing only town and city records

Creating dataset containing records belonging to cities and towns only:

```
In [19]: df_ct = df_useful[(df_useful.type == 'City') | (df_useful.type == 'Town')]
```

Creating bins:

```
In [20]: df_bins = df_ct.copy()
```

```
In [21]: df_bins['married'] = pd.cut(df_bins.married, bins=4,include_lowest=True)
```

```
In [22]: df_bins['some_type_of_debt'] = pd.cut(df_bins.debt, bins=4,include_lowest=True)
```

```
In [23]: df_bins['high_school_degree'] = pd.cut(df_bins.hs_degree, bins=10,include_lowest=True)
```

```
In [24]: df_bins['rent'] = pd.cut(df_bins.rent_mean, bins=10,include_lowest=True)
```

```
In [25]: df_bins['household_income'] = pd.cut(df_bins.hi_mean, bins=[5000, 30000, 63000, 90000, 300000],include_lowest=True)
```

```
In [26]: df_bins['home_equity'] = pd.cut(df_bins.home_equity, bins=[0.0,0.2,0.4,0.6,0.8,1.0],include_lowest=True)
```

Creating another dataset with only non-string values:

```
In [27]: df_no_strings = df_useful.iloc[:,9:]
```

Analyzing the variations in the rent prices using world map:

According to our dataset , we have latitude and longitude for each record. So lets see the rent prices for each county.

```
In [28]: import plotly.figure_factory as ff
import numpy as np
```

```
In [29]: mapbox_access_token = 'pk.eyJ1Ijoia3J1dGlrYWftYnZhbmUiLCJhIjoiY2poZmoxMjBjMTZ4aTM2bmduYnZtYXlrZCJ9._NLH_EGbJp
qz3VR-rLv1mw'
```

```
In [30]: plotly.tools.set_credentials_file(username='krutika.a', api_key='iSsK12rHGuumSjzRhYDF')
```

```
In [31]: plot = df_useful.copy()
plot['State FIPS Code'] = plot['stateid'].apply(lambda x: str(x).zfill(2))
plot['County FIPS Code'] = plot['countyid'].apply(lambda x: str(x).zfill(3))
plot['FIPS'] = plot['State FIPS Code'] + plot['County FIPS Code']
plot.fillna(0, inplace=True)
```

```
In [32]: map_us = plot.groupby('FIPS')[['FIPS', 'rent_mean']].mean().reset_index()
```

```
In [33]: colorscale1 = ["#f7fbff", "#ebf3fb", "#deebf7", "#d2e3f3", "#c6dbef", "#b3d2e9", "#9ecae1",
                        "#85bcd6", "#6baed6", "#57a0ce", "#4292c6", "#3082be", "#2171b5", "#1361a9",
                        "#08519c", "#0b4083", "#08306b"]

colorscale = [
    'rgb(68.0, 1.0, 84.0)',
    'rgb(66.0, 64.0, 134.0)',
    'rgb(38.0, 130.0, 142.0)',
    'rgb(63.0, 188.0, 115.0)',
    'rgb(216.0, 226.0, 25.0)',
    'rgb(223.0, 223.0, 33.0)',
    'rgb(192.0, 226.0, 57.0)',
    'rgb(226.0, 204.0, 57.0)',
    'rgb(226.0, 169.0, 25.0)',
    'rgb(222.0, 90.0, 25.0)',
]
endpts = list(np.linspace(map_us['rent_mean'].min(), map_us['rent_mean'].max(), len(colorscale) - 1))
fips = map_us['FIPS'].tolist()
values = map_us['rent_mean'].tolist()
```

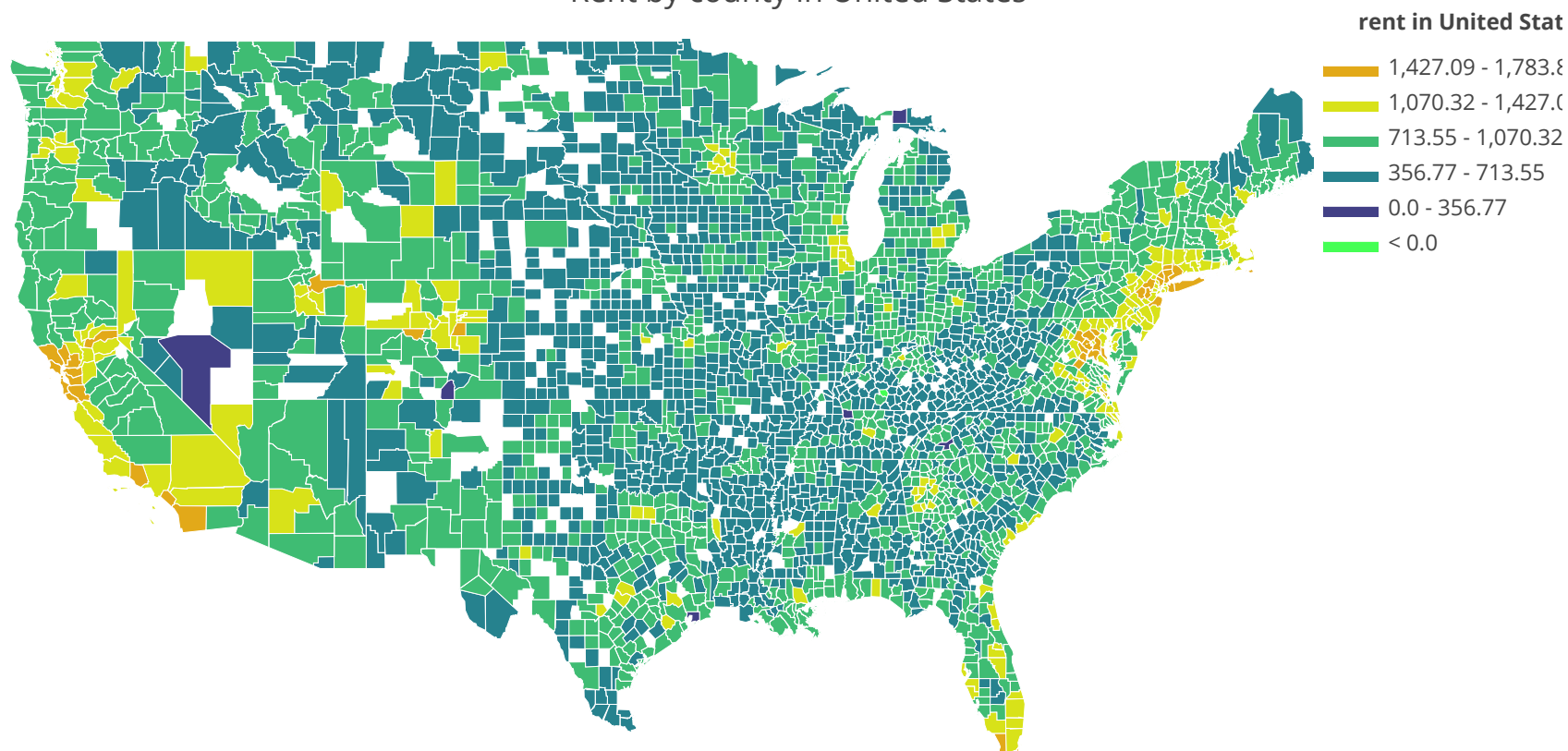
```
In [34]: fig = ff.create_choropleth(
    fips=fips, values=values,
    binning_endpoints=endpts,
    colorscale=colorscale,
    show_state_data=False,
    show_hover=True, centroid_marker={'opacity': 0},
    asp=2.9, title='Rent by county in United States ',
    county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
    legend_title='rent in United States',
)
```

```
In [35]: py.ipplot(fig, filename='choropleth_full_usa')
```

The draw time for this plot will be slow for clients without much RAM.

Out[35]:

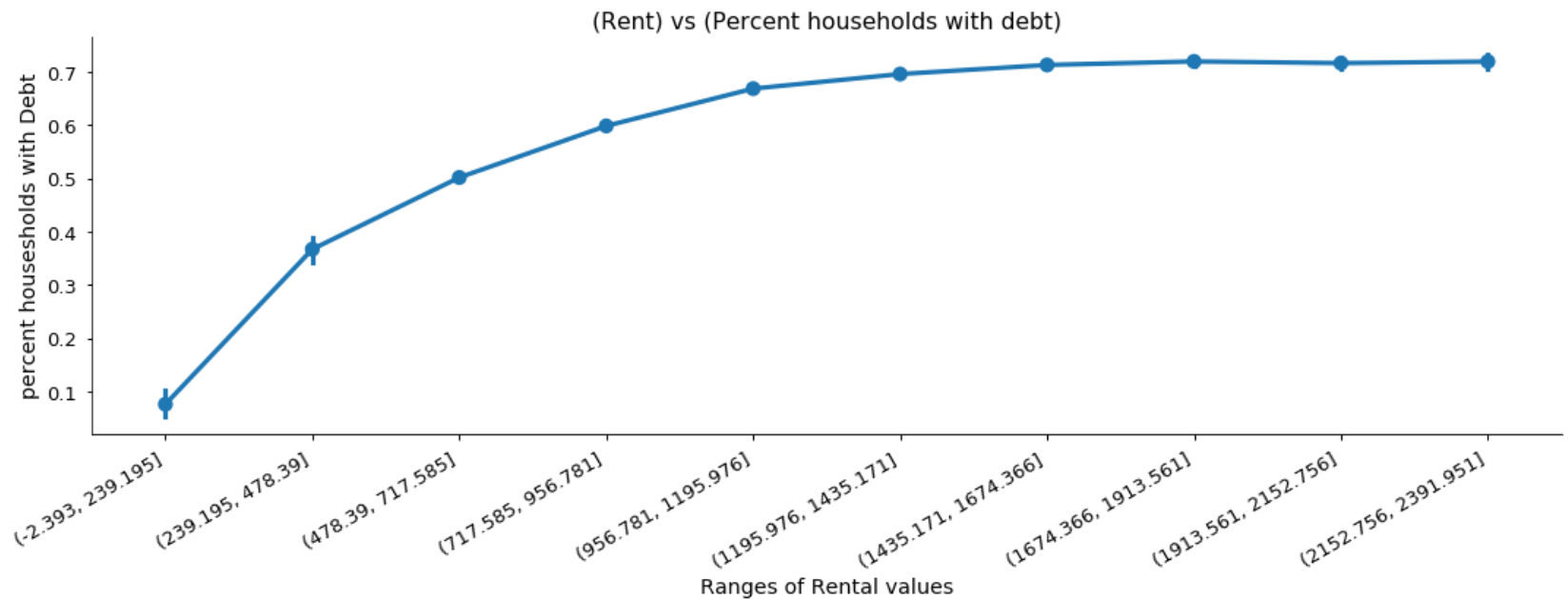
Rent by county in United States



The above map shows the distribution of rent and helps us to visualize which locations have concentration of higher rent prices. So we have found that specific counties in California and NewYork have the highest rent prices eg - Santa Clara County of California

Finding 1: rent vs debt

In [36]:



In real world, we wont expect rent and debt to have any correlation. However, from the above graph we can interpret that locations with higher mean rent have higher percentage of houses with debt.

In [37]:

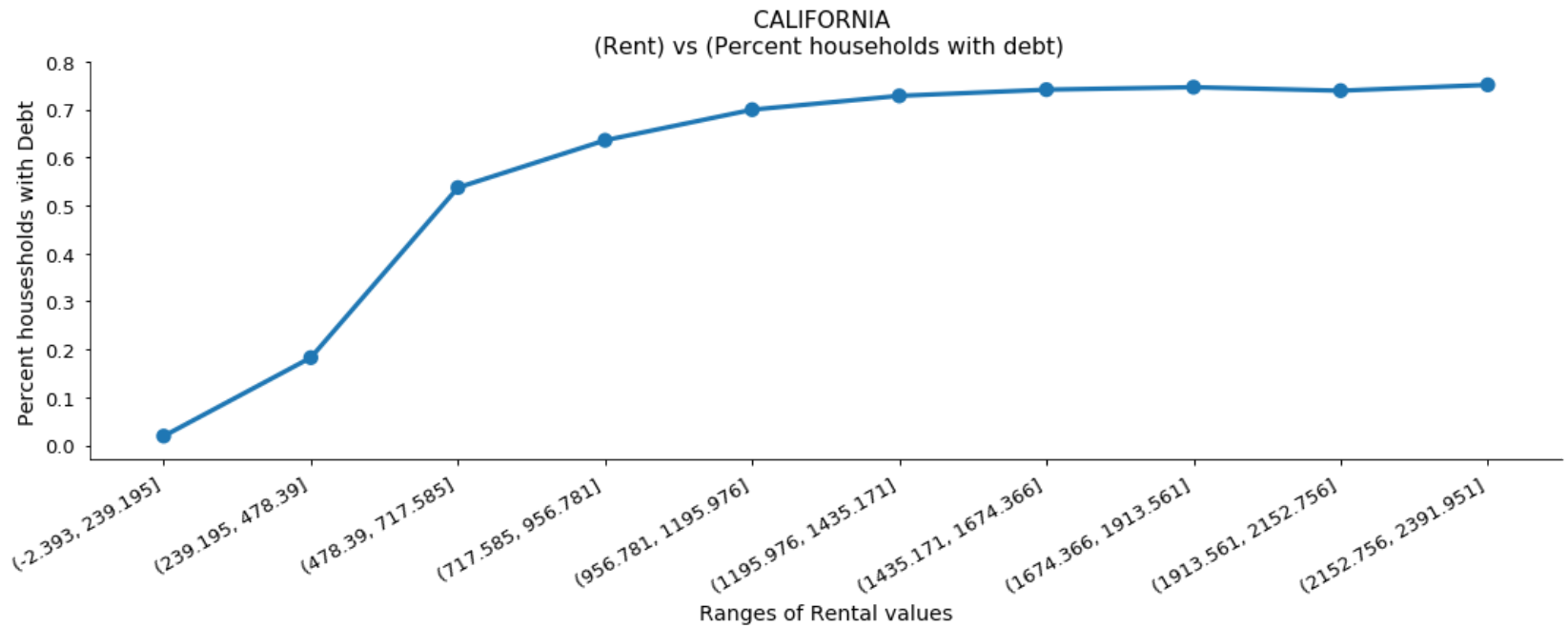
```
Out[37]: state
California    3882
Texas         2595
New York      2467
Florida       2154
Georgia       1023
New Jersey    949
Virginia      911
Indiana       789
Washington    786
Tennessee     770
Name: state, dtype: int64
```

As we can see, this dataset has largest records for the state of California. So lets study a little about this state.

Creating a dataset containing records belonging only to the state of California:

In [38]:

In [39]:



Here we just focused on one state - California. We observe similar correlation between mean rent and percentage of houses having debts.

Lets search for top 5 states with largest mean rent and largest percentage of houses with debt.

In [40]:

```
Out[40]: state
District of Columbia    0.78
California               0.71
Massachusetts           0.70
Colorado                0.69
Washington              0.68
Name: debt, dtype: float64
```

In [41]:

```
Out[41]: state
California              1380.42
New Jersey              1321.75
District of Columbia   1312.28
Massachusetts           1187.72
New York                1185.01
Name: rent_mean, dtype: float64
```

This also supports our conclusion that the states having highest rents also have highest percentages of household with debts.

Now, lets examine the reason behind people having higher rents. Lets see if the type of the place people reside in, matters.

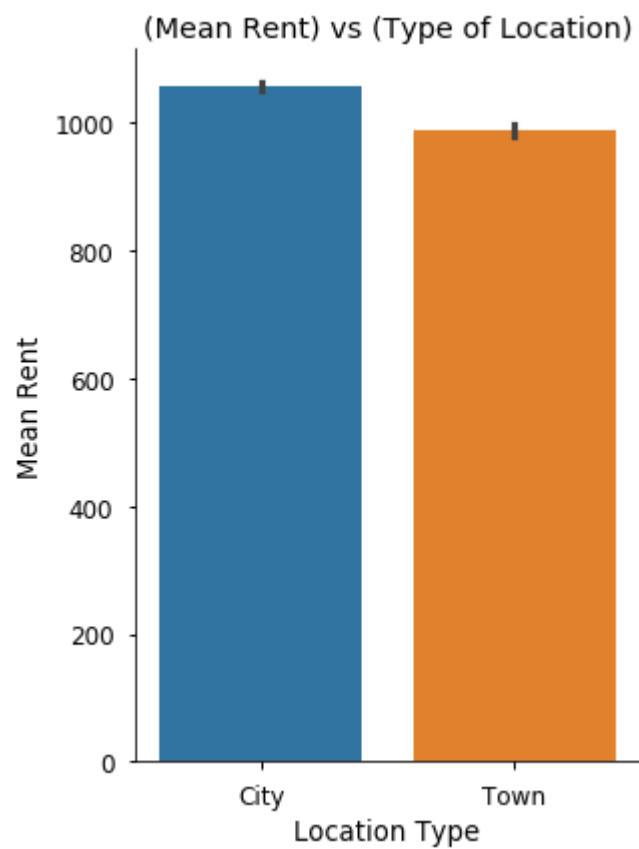
In [42]:

```
Out[42]: type
City      89475826
Town      22643329
Name: pop, dtype: int64
```

In [43]:

```
Out[43]: type
City      1057.75
Town      988.17
Name: rent_mean, dtype: float64
```

In [44]:



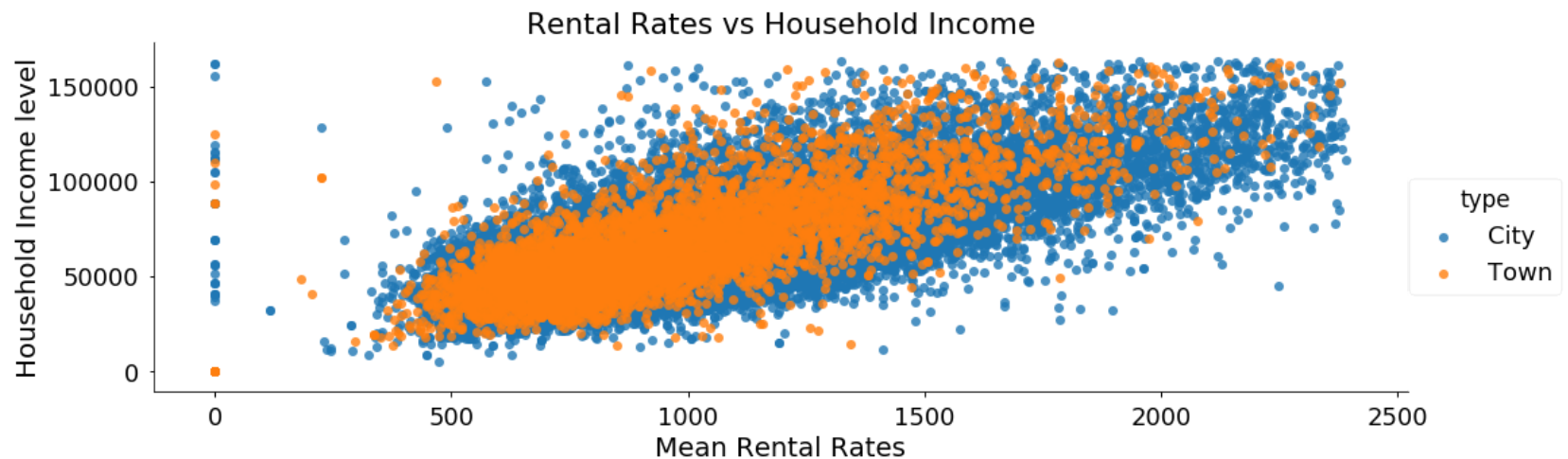
Conclusion :

As we can see from above, the population is highest in cities resulting in an increase in demands, further leading to higher rent prices there compared to towns.

Finding 2: rent vs income

Now lets explore the reason as to why these people from cities can afford higher rents.

In [45]:



In [46]:



From above, we can conclude that as the mean household income goes on increasing, the mean rent prices also go on increasing. We can also deduce that the people from city have more household income and more rent than the town people. This above graph also confirms the same that higher income areas have higher rents.

Lets try to predict the rent prices for cities and towns using Linear Regression:

We will use Simple Regresssion as we are going to predict only one variable.

Step 1. Importing the required libraries:</div>

In [47]:

Step 2. Create X and Y:

Our Y will be the rent prices that we are going to predict.

In [48]:

Step 3. Split the data into train and test data:

We have made our train data as 75% of the original data and test data as 25%. We have randomized the splitting by using random state and given labels to both our data.

In [49]:

In [50]:

Out[50]: (28117, 29)

Step 4. Instantiate and fit our data the model:

In [51]:

Step 5. Check score of our model on test data:

In [52]:

Out[52]: 0.7127132758175054

Step 6. Calculate the error rates:

In [53]:

In [54]:

Mean Absolute Error: 150.62 dollars

In [55]:

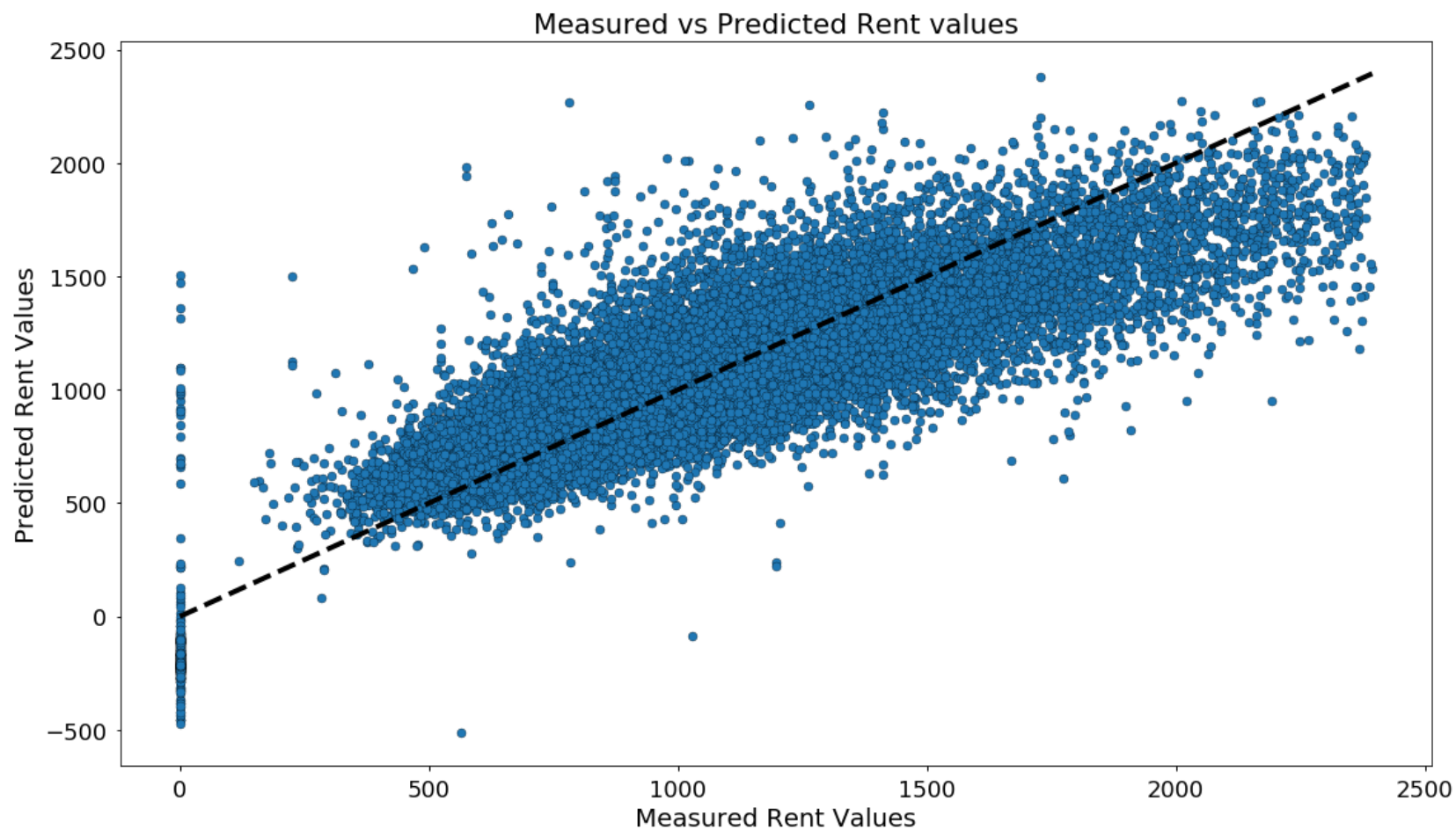
Root Mean square error: 12.27273400673216 dollars

Step 5: Plotting graph:

Let's visualize the predicted values vs the measured values.

In [56]:

In [57]:



The accuracy of the predicted rent is ``0.71`` We can see the upward trend and the points aligned with the linear function line.

Conclusion:

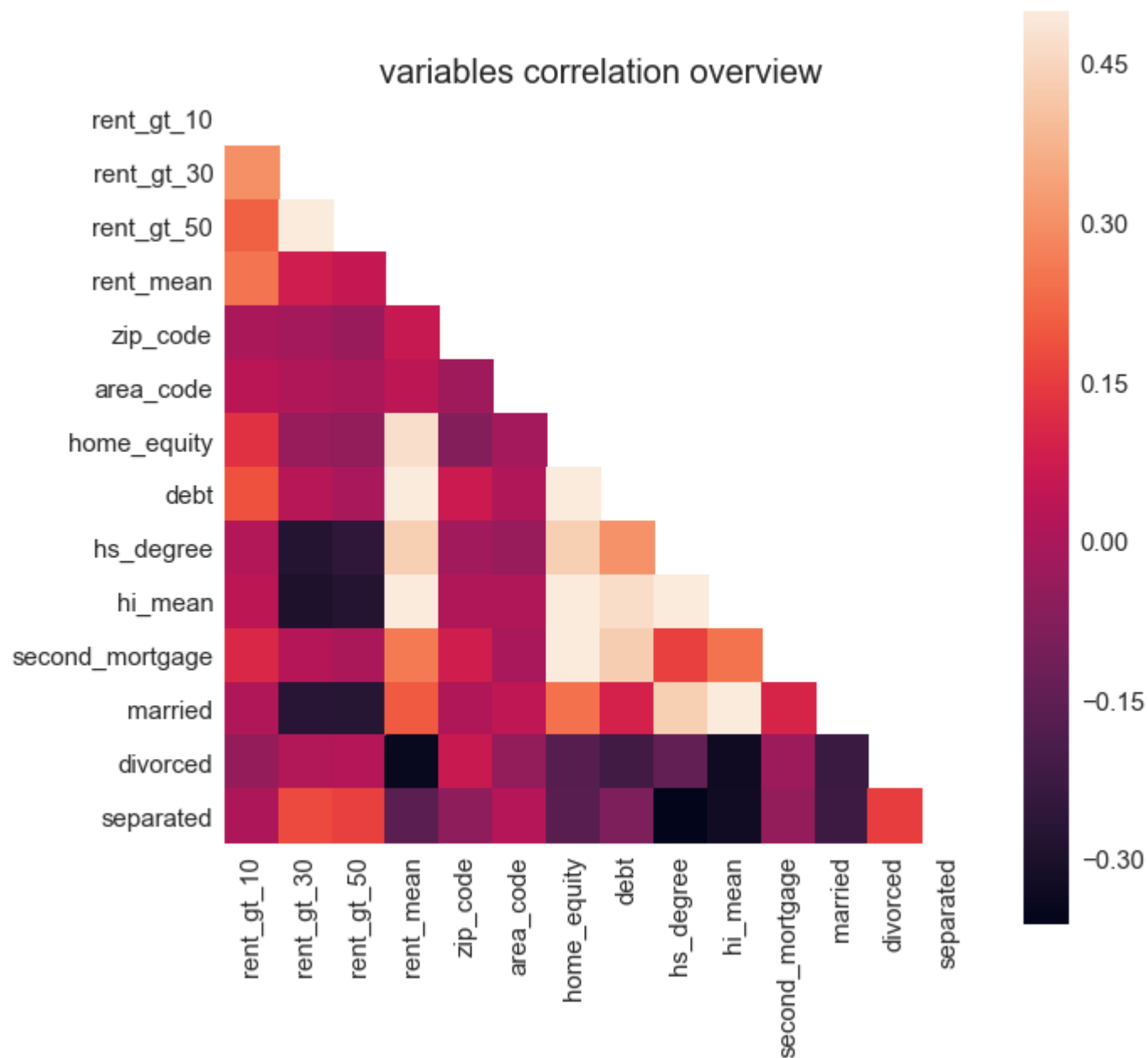
Variables that *highly contribute* to *rent* : *[household income, debt, home equity, highschool degree, second mortgage]*

**We can conclude this because using these factors, we trained the above model and it is able to predict rent with accuracy of ``71%``.
We can come to this conclusion by also looking at the heat map plotted below which shows us the variables with good correlation.**

Finding the correlation and important variables:

In [58]:

In [59]:



Finding 3:

In order to find out the trends based on State, Race and other census variables, we are merging the Real estate data with the USA Census Data

Census Data

Data preparation of USA Census Dataset

In [109]:

In [110]:

Out[110]:

	CensusTract	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Pacific	Citizen	Incom
0	1001020100	Alabama	Autauga	1948	940	1008	0.9	87.4	7.7	0.3	0.6	0.0	1503	61838.
1	1001020200	Alabama	Autauga	2156	1059	1097	0.8	40.4	53.3	0.0	2.3	0.0	1662	32303.
2	1001020300	Alabama	Autauga	2968	1364	1604	0.0	74.5	18.6	0.5	1.4	0.3	2335	44922.
3	1001020400	Alabama	Autauga	4423	2172	2251	10.5	82.8	3.7	1.6	0.0	0.0	3306	54329.
4	1001020500	Alabama	Autauga	10763	4922	5841	0.7	68.5	24.8	0.0	3.8	0.0	7666	51965.

In [111]:

In [112]:

Aggregation: For aggregating the data for each state ,we need to convert the metrics expressed as percentage for each county into counts so they can be summed.

In [113]:

In [114]:

Out[114]:

	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Pacific	Citizen	Income	IncomeErr
0	Alabama	Autauga	1948	940	1008	0.9	87.4	7.7	0.3	0.6	0.0	1503	61838.0	11900.0
1	Alabama	Autauga	2156	1059	1097	0.8	40.4	53.3	0.0	2.3	0.0	1662	32303.0	13538.0
2	Alabama	Autauga	2968	1364	1604	0.0	74.5	18.6	0.5	1.4	0.3	2335	44922.0	5629.0
3	Alabama	Autauga	4423	2172	2251	10.5	82.8	3.7	1.6	0.0	0.0	3306	54329.0	7003.0
4	Alabama	Autauga	10763	4922	5841	0.7	68.5	24.8	0.0	3.8	0.0	7666	51965.0	6935.0

In [115]:

In [116]:

In [117]:

Out[117]:

	State	County	TotalPop	Citizen	Employed	count_hisp	count_white	count_black	count_native	count_asian	cou
0	Alabama	Autauga	1948	1503	943	17.53	1702.55	150.00	5.84	11.69	0.0
1	Alabama	Autauga	2156	1662	753	17.25	871.02	1149.15	0.00	49.59	0.0

In [118]:

In [119]:

Out[119]:

	State	TotalPop	Citizen	count_hisp	count_white	count_black	count_native	count_asian	count_pacifi
State									
Alabama	Alabama	4821879	3612759	1.93e+05	3.20e+06	1.27e+06	22176.93	5.84e+04	1578.01
Alaska	Alaska	729562	520405	4.77e+04	4.54e+05	2.37e+04	98221.23	4.23e+04	8495.72
Arizona	Arizona	6531748	4422184	2.01e+06	3.66e+06	2.61e+05	263963.55	1.94e+05	11255.74
Arkansas	Arkansas	2956316	2162204	2.03e+05	2.18e+06	4.56e+05	16442.01	4.00e+04	6725.93
California	California	38221472	24104414	1.47e+07	1.48e+07	2.13e+06	140378.35	5.18e+06	138308.14

Converting the counts to percent per state population

In [120]:

In [121]:

In [122]:

Out[122]:

	State	Citizen	count_hisp	count_white	count_black	count_native	count_asian	count_pacific	total_inc
State									
Alabama	Alabama	3612759	1.93e+05	3.20e+06	1.27e+06	22176.93	5.84e+04	1578.01	1.16e+11
Alaska	Alaska	520405	4.77e+04	4.54e+05	2.37e+04	98221.23	4.23e+04	8495.72	2.42e+10
Arizona	Arizona	4422184	2.01e+06	3.66e+06	2.61e+05	263963.55	1.94e+05	11255.74	1.68e+11
Arkansas	Arkansas	2162204	2.03e+05	2.18e+06	4.56e+05	16442.01	4.00e+04	6725.93	6.74e+10
California	California	24104414	1.47e+07	1.48e+07	2.13e+06	140378.35	5.18e+06	138308.14	1.16e+12

In [123]:

USA Real Estate Statistics Data

Data preparation of USA Real Estate statistics data so that they can be merged with USA Census data based on states

In [124]:

In [125]:

In [126]:

Out[126]:

	UID	COUNTYID	STATEID	state	state_ab	city	place	type	zip_code	area_code	lat	lng	pop
0	220336	16	2	Alaska	AK	Unalaska	Unalaska City	City	99685	907	53.62	-166.77	4619
1	220342	20	2	Alaska	AK	Eagle River	Anchorage	City	99577	907	61.17	-149.28	3727
2	220343	20	2	Alaska	AK	Jber	Anchorage	City	99505	907	61.28	-149.65	8736
3	220345	20	2	Alaska	AK	Anchorage	Point Mackenzie	City	99501	907	61.23	-149.89	1941
4	220347	20	2	Alaska	AK	Anchorage	Anchorage	City	99504	907	61.22	-149.77	5981

In [127]:

In [128]:

In [129]:

In [130]:

In [131]:

In [132]:

Out[132]:

	state	total_pop	female_age_mean	male_age_mean	rent_mean	hi_mean	count_second_mortgage	count
state								
Alabama	Alabama	2516214	40.69	38.28	761.20	55645.80	620.19	14411
Alaska	Alaska	469126	36.87	36.51	1190.09	83541.33	90.04	2931.0
Arizona	Arizona	3491125	39.97	38.72	1066.94	65558.17	955.49	22941
Arkansas	Arkansas	1540462	40.43	38.35	707.35	54876.13	265.74	8769.0
California	California	20197555	38.92	37.18	1453.74	81251.48	8758.87	14527

Merge : USA Census and USA Real Estate Statistics data

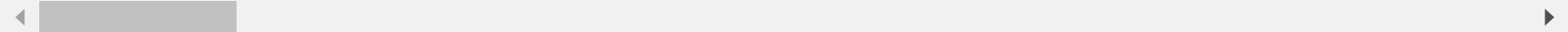
Merging the USA Census Data and USA Real Estate statistics based on State

In [133]:

In [134]:

Out[134]:

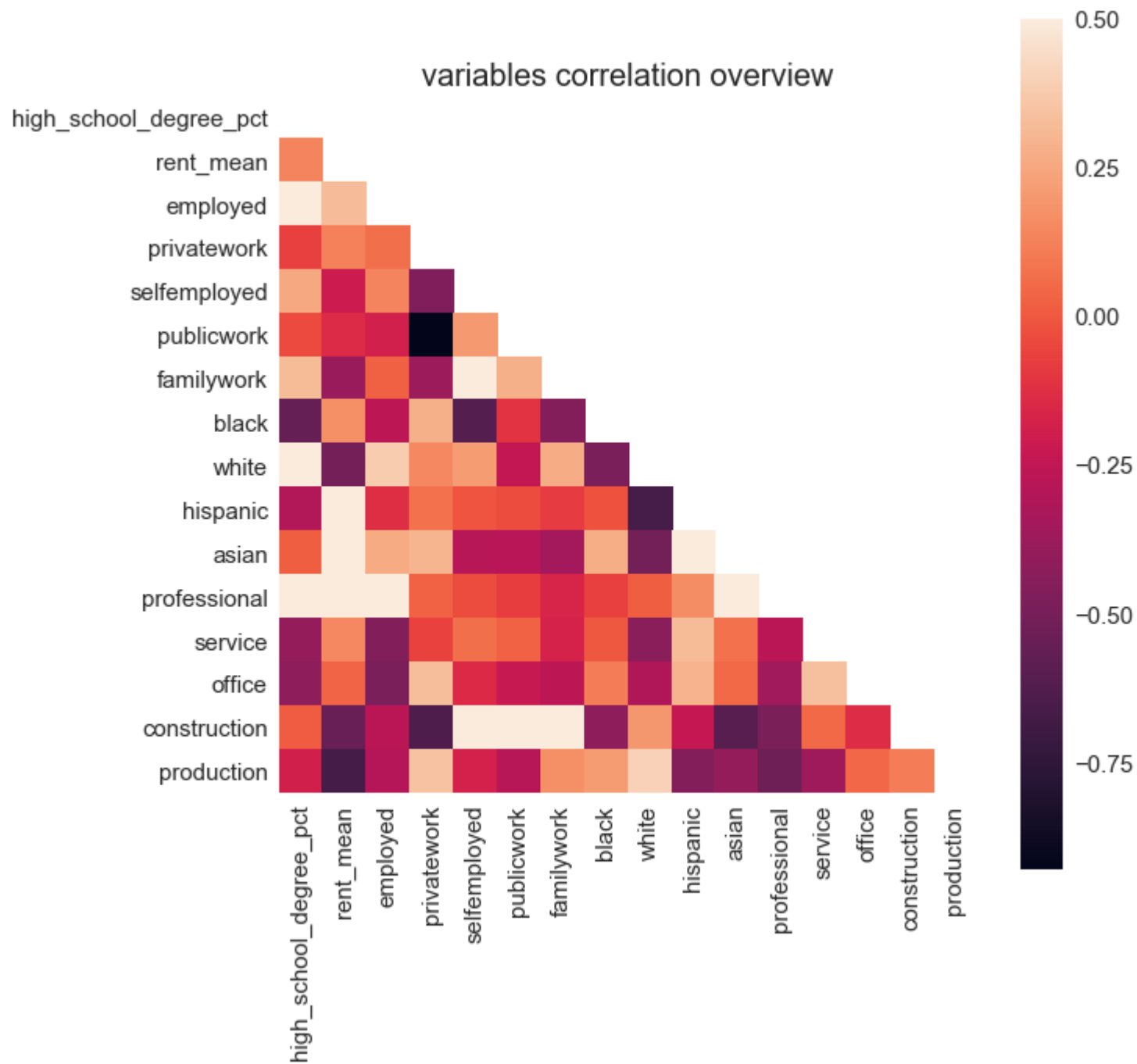
	state	total_pop	female_age_mean	male_age_mean	rent_mean	hi_mean	count_second_mortgage	count_debt	c
0	Alabama	2516214	40.69	38.28	761.20	55645.80	620.19	14411.73	2
1	Alaska	469126	36.87	36.51	1190.09	83541.33	90.04	2931.00	4
2	Arizona	3491125	39.97	38.72	1066.94	65558.17	955.49	22941.82	2
3	Arkansas	1540462	40.43	38.35	707.35	54876.13	265.74	8769.04	1
4	California	20197555	38.92	37.18	1453.74	81251.48	8758.87	145278.61	1



Initial Analysis Plots

Variable correlation

In [135]:

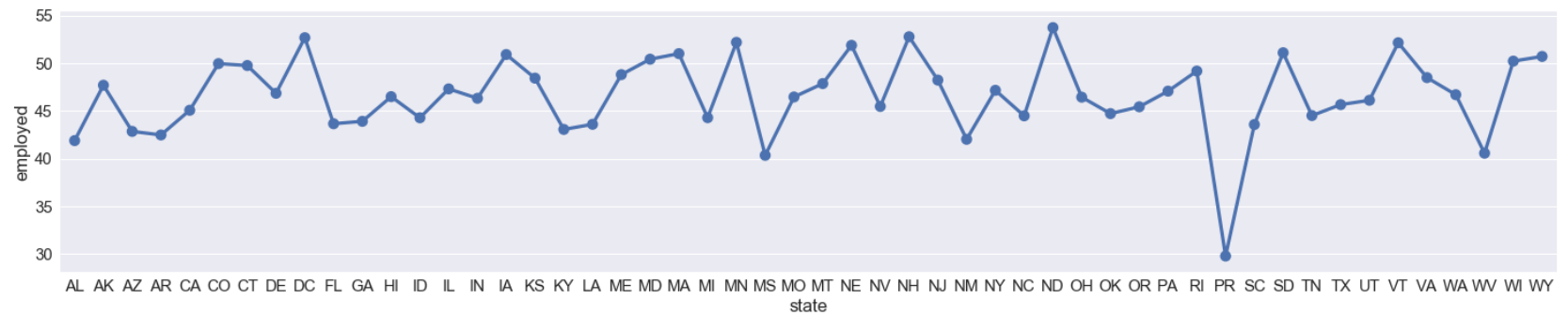


In [136]:

Below graph shows the distribution of employed among all states of USA.

In [137]:

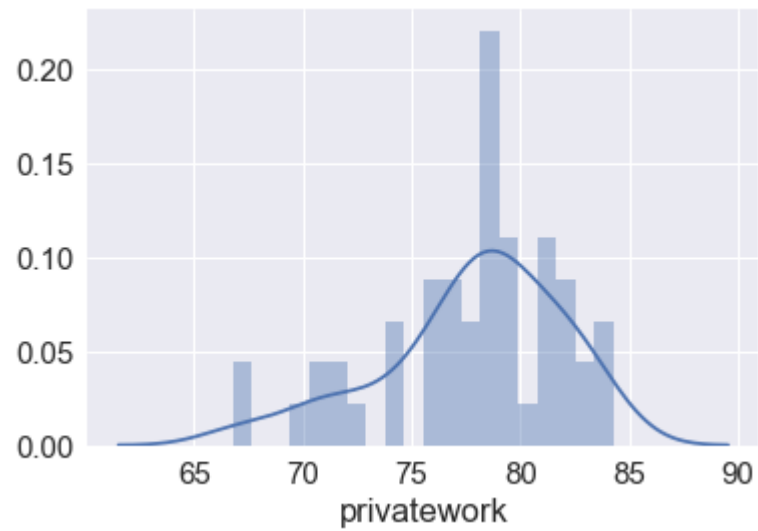
Out[137]: <seaborn.axisgrid.FacetGrid at 0x234c3a296a0>



Lets now identify other metrics that affect the employed

In [138]:

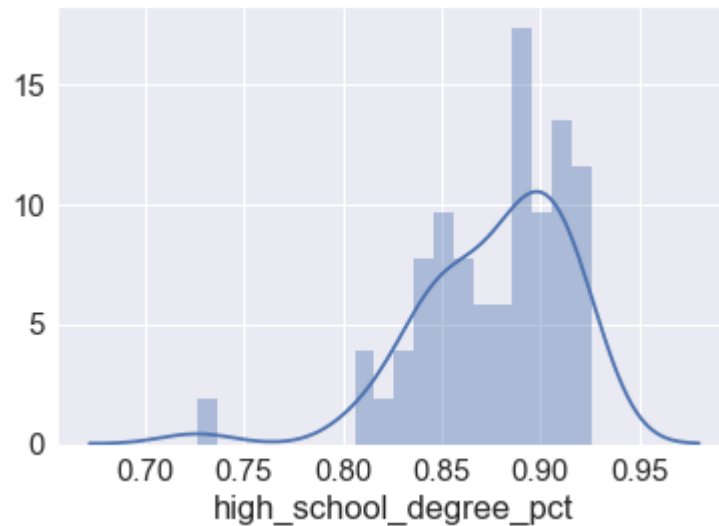
Out[138]: <matplotlib.axes._subplots.AxesSubplot at 0x234c7a67f60>



The Private work sector is skewed towards left

In [139]:

Out[139]: <matplotlib.axes._subplots.AxesSubplot at 0x234c79cca90>



The High school degree is skewed towards left

From the above distplots, it is very clear that the private work and high school degree have similar distribution (skewed left). Now lets relate with other Census metrics and explore further.

Among the employed population, the work type is categorized as Private, Self Employed ,Public work and Family work

Lets analyse their distribution in USA

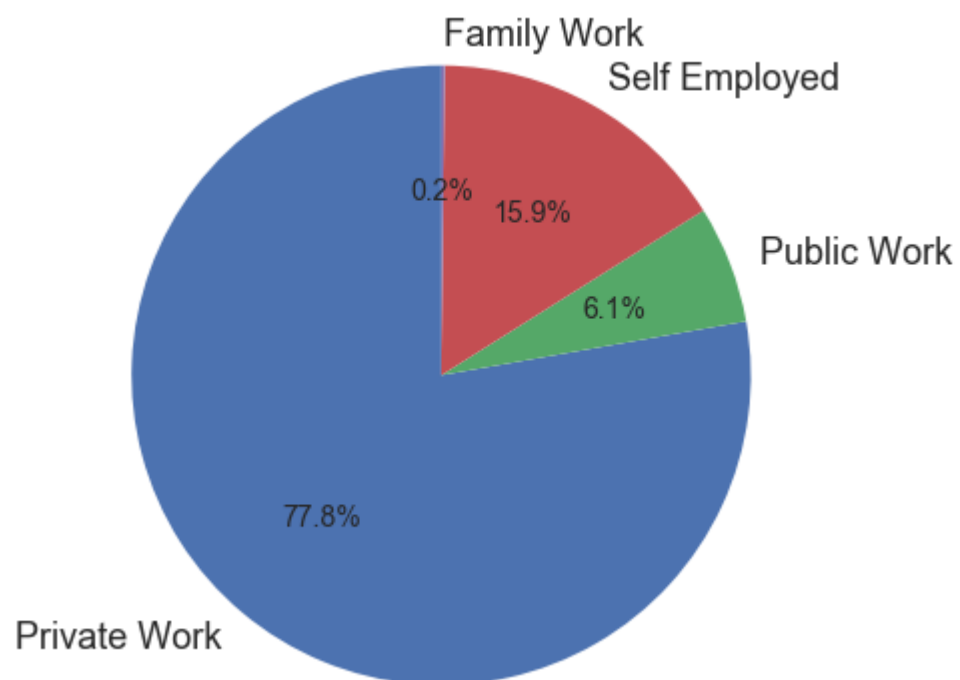
In [140]:

In [141]:

```
Out[141]: privatework    77.79  
          selfemployed    6.14  
          publicwork     15.90  
          familywork      0.18  
          dtype: float64
```

In [142]:

In [143]:



From the above Pic Chart it is very clear that Private work is predominant among all States of USA

Finding 3: Private Work Sector Trend Among Black Population

Lets further Dig into the Private work Sector

In [144]:

In [145]:

Out[145]: 77.78867570273809

Creating Binary column private_high for high and low private work Sector based on its mean

In [146]:

In [147]:

Out[147]:

	state	private_high
0	Alabama	1.0
1	Alaska	0.0
2	Arizona	1.0
3	Arkansas	0.0
4	California	0.0

In [148]:

In [149]:

In [150]:

Out[150]:

	female_age_mean	male_age_mean	rent_mean	hi_mean	second_mortgage_pct	debt_pct	high_school_degree_pct	
0	40.69	38.28	761.20	55645.80	0.02	0.57	0.84	36
1	36.87	36.51	1190.09	83541.33	0.02	0.62	0.92	52

Finding - 3: Machine Learning

In [151]:

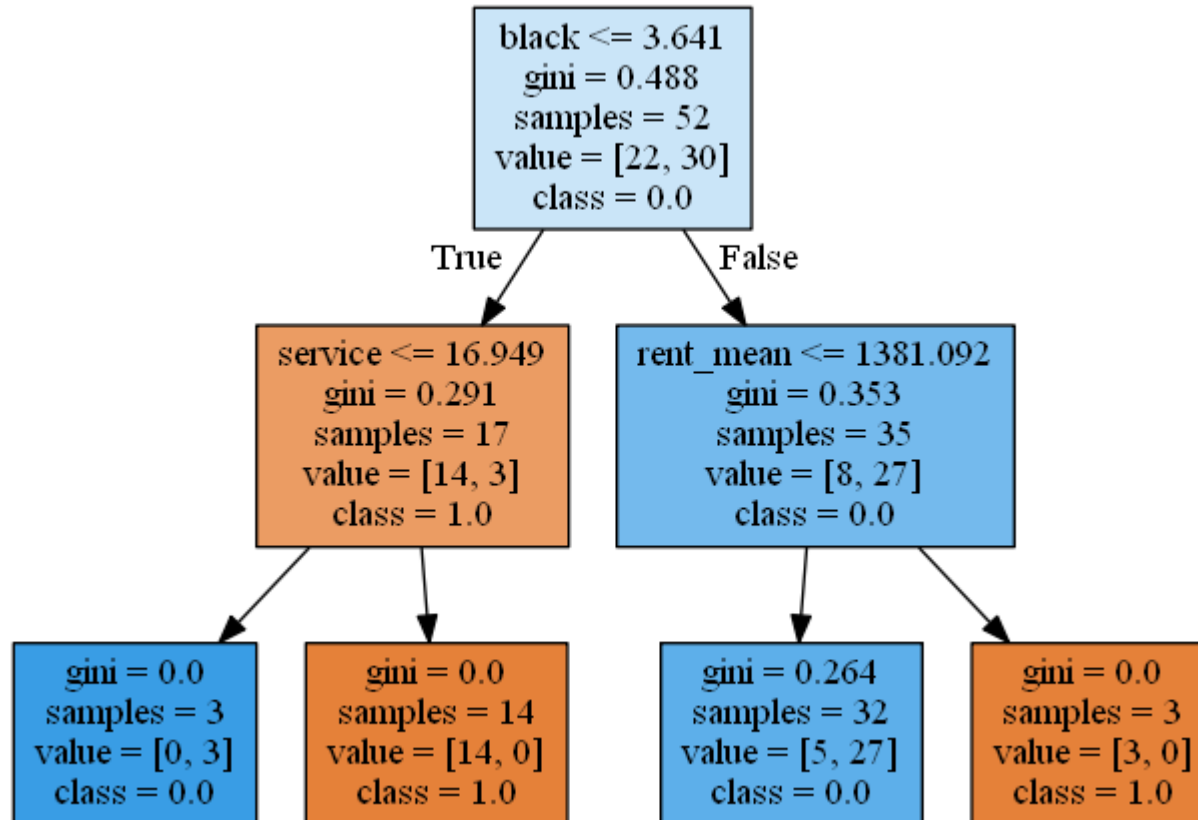
In [152]:

In [153]:

Out[153]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=2, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')

In [154]:

Out[154]:



Within states whose black population is less than 3.6%, and where less than 16.9% have service job , work privately. State whose black population is greater than 3.6% have higher service based private work rate and they pay rent of less than or equal to 1381.0

Finding-3 : Validation

In [156]:

In [157]:

Out[157]: 18.475965217351042

Creating binary Column service_high for high and low Service work based on mean value

In [158]:

In [159]:

Out[159]:

	state	service_high
0	Alabama	0.0
1	Alaska	0.0
2	Arizona	1.0
3	Arkansas	0.0
4	California	1.0

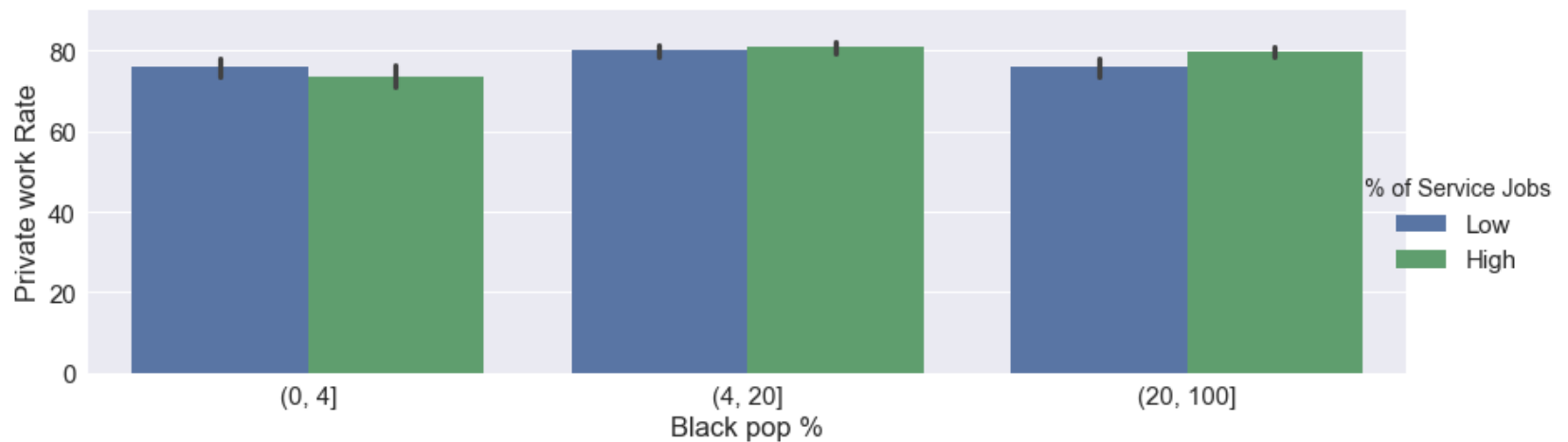
In [160]:

Out[160]: <seaborn.axisgrid.FacetGrid at 0x234c7dd1cc0>

Out[160]: <seaborn.axisgrid.FacetGrid at 0x234c7dd1cc0>

Out[160]: <seaborn.axisgrid.FacetGrid at 0x234c7dd1cc0>

Out[160]: <seaborn.axisgrid.FacetGrid at 0x234c7dd1cc0>



Bar Graph Showing the private work based service job distribution among the Black Population

From the above bar graph it is clear that rate of private work based service job are lower for states with Black population below 4% and for states with population of Balck greater than 4%, have higher rate of private work based service job

The below are the top five states with maximum private work based service job percentage among black population (greater than 3.6%)

In [161]:

Out[161]:

	State	Private_Work	Service	high_school_degree	Black_Pop_Pct	Income	Rent
14	Indiana	84.26	17.40	0.87	9.05	59621.58	806.19
38	Pennsylvania	83.85	18.24	0.89	10.47	68644.26	937.41
22	Michigan	83.69	18.81	0.89	13.74	62692.21	909.09
40	Rhode Island	82.62	20.64	0.87	5.34	74677.45	1045.60
13	Illinois	82.58	18.06	0.87	14.11	72200.29	1024.79

Conclusion

Despite Private work type being predominant in USA, the top 5 state with highest rate of private work based service job have lower income among the Black population, even though their High school degree percentage of the state is high. They are in a position to afford only lower rent. For betterment of their lives, these states should take more efforts in increasing the salary of private work based service jobs among the educated black population.