**Question – 1 :-**

Compute the average campaign contribution for the Obama and McCain campaigns from the dataset in day 1.

Whatʹs the effect size? We have an average contribution of $423 for McCain and $192 for Obama, for an effect size of $231.McCain appears, on average, to have more giving donors.

**Code :-**

import pandas as pd

df = pd.read\_csv("./P00000001-ALL.csv")

obama\_contributions = df[df["cand\_nm"] == "Obama, Barack"]["contb\_receipt\_amt"]

mccain\_contributions = df[df["cand\_nm"] == "McCain, John S"]["contb\_receipt\_amt"]

avg\_obama\_contribution = obama\_contributions.mean()

avg\_mccain\_contribution = mccain\_contributions.mean()

print("Average contribution for Obama campaign: ${:.2f}".format(avg\_obama\_contribution))

print(

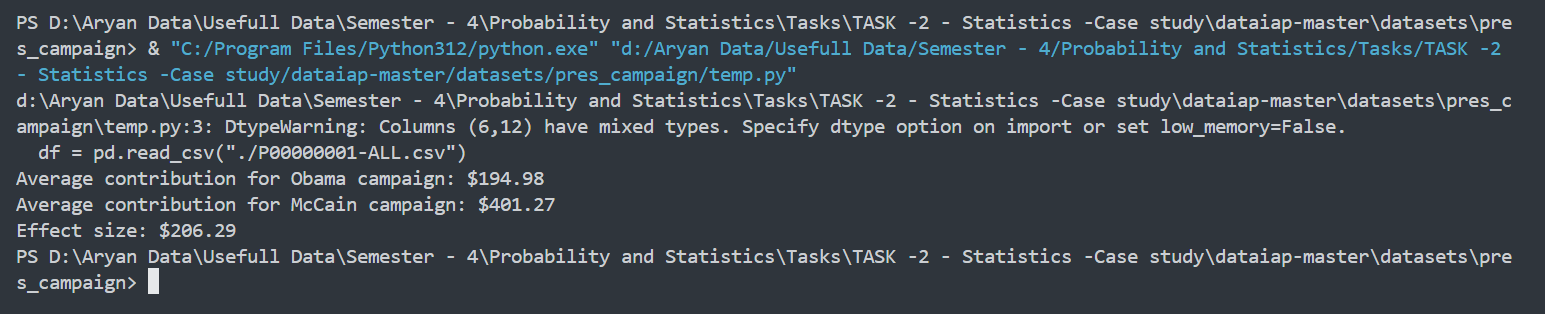
"Average contribution for McCain campaign: ${:.2f}".format(avg\_mccain\_contribution)

)

effect\_size = avg\_mccain\_contribution - avg\_obama\_contribution

print("Effect size: ${:.2f}".format(effect\_size))

**Output :-**



**Question-2 :-**

Build a histogram for the Obama and McCain campaigns. This is challenging, because there are a large number of

outliers that make the histograms difficult to compare. Add the line sub.set\_xlim((‐20000, 20000))

before displaying the plot in order to set the x‐values of the histogram to cut off donations larger than $20,000 or smaller than ‐$20,000 (refunds). With bar widths of 50 and increments of $100, your histogram will look something like this:

Code :-

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv("./P00000001-ALL.csv")

obama\_contributions = df[df["cand\_nm"] == "Obama, Barack"]["contb\_receipt\_amt"]

mccain\_contributions = df[df["cand\_nm"] == "McCain, John S"]["contb\_receipt\_amt"]

increment = 100

plt.figure(figsize=(10, 6))

plt.hist(obama\_contributions,bins=range(

int(min(obama\_contributions)),

int(max(obama\_contributions)) + increment,

increment,

),

color="blue",

alpha=0.5,

label="Obama",

)

plt.hist(

mccain\_contributions,

bins=range(

int(min(mccain\_contributions)),

int(max(mccain\_contributions)) + increment,

increment,

),

color="red",

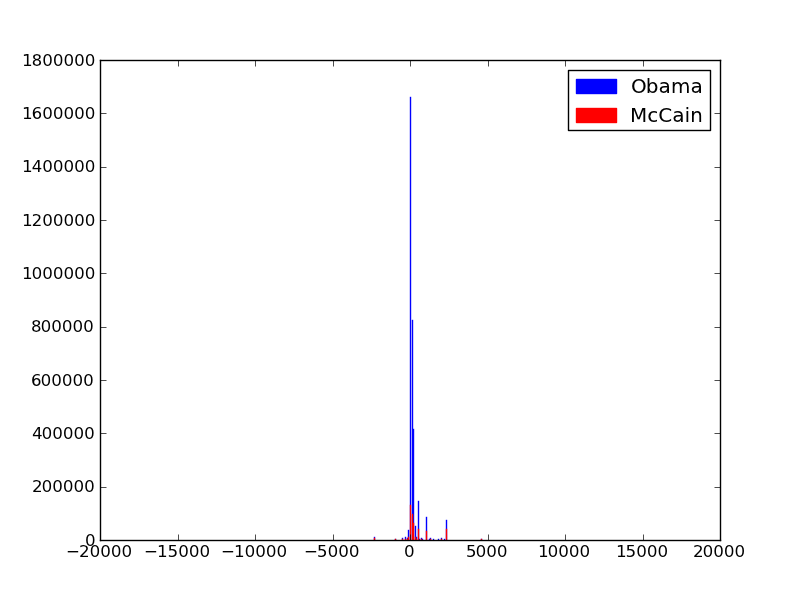
alpha=0.5,

label="McCain",

)

plt.legend()

plt.show()

**Output:-**

**Question-3 :-**

Build a box‐and‐whiskers plot of the McCain and Obama campaign contributions. Again, outliers make this a

difficult task. With whis=1 , and by setting the y range of the plots like so

sub.set\_ylim((‐250, 1250))

**Code :-**

import matplotlib.pyplot as plt

import pandas as pd

obama\_contributions = df[df["cand\_nm"] == "Obama, Barack"]["contb\_receipt\_amt"]

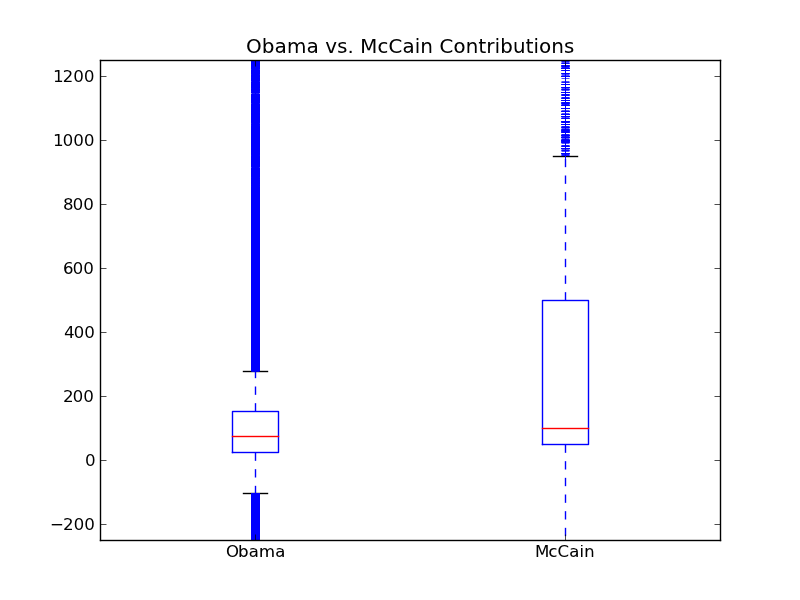
mccain\_contributions = df[df["cand\_nm"] == "McCain, John S"]["contb\_receipt\_amt"]

plt.figure(figsize=(8, 6))

plt.boxplot([mccain\_contributions, obama\_contributions], labels=['McCain', 'Obama'], whis=1)

plt.ylim((-250, 1250)) # Set y-axis range

plt.show()

**Output :-**

**Question-4 :-**

Run Welchʹs T‐test on the campaign data. Is the effect size between McCain and Obama significant? By our

measurements, the p‐value reported is within rounding error of 0. Thatʹs significant by anyoneʹs measure: thereʹs a

near‐nonexistant chance weʹre seeing this difference between the candidates by some random fluke in the universe.

**Code :-**

import pandas as pd

import welchttest

from scipy import stats

df = pd.read\_csv("./P00000001-ALL.csv")

obama\_contributions = df[df["cand\_nm"] == "Obama, Barack"]["contb\_receipt\_amt"]

mccain\_contributions = df[df["cand\_nm"] == "McCain, John S"]["contb\_receipt\_amt"]

t\_statistic, p\_value = stats.ttest\_ind(obama\_contributions, mccain\_contributions, equal\_var=False)

print("Welch's t-test:")

print("t-statistic:", t\_statistic)

print("p-value:", p\_value)

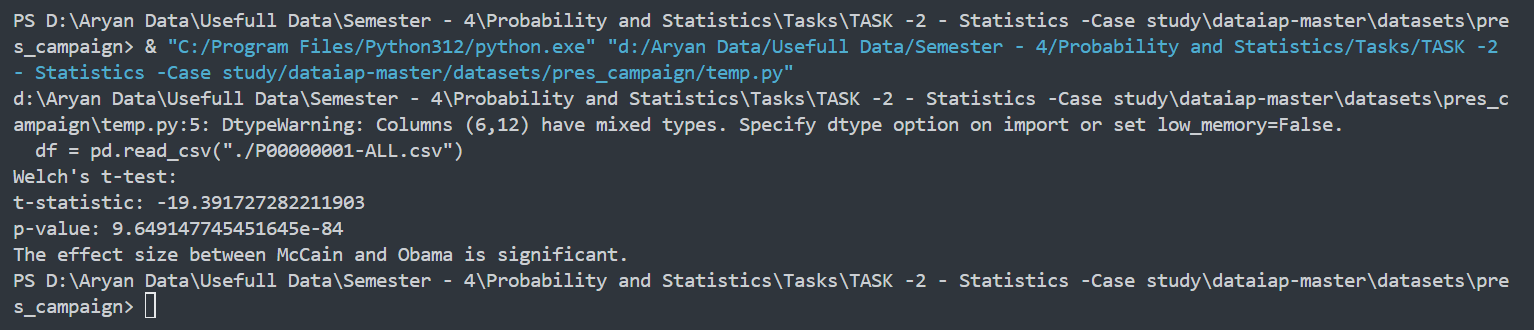
if p\_value < 0.05:

print("The effect size between McCain and Obama is significant.")

else:

print("The effect size between McCain and Obama is not significant.")

**Output :-**

****

**Question-5 :-**

since we shouldnʹt be using Welchʹs T‐Test on the campaign contribution data, run the Mann‐Whitney U test on

the data. Is the difference between the Obama and McCain contributions still significant?

**Code :-**

import pandas as pd

import welchttest

from scipy import stats

df = pd.read\_csv("./P00000001-ALL.csv")

obama\_contributions = df[df["cand\_nm"] == "Obama, Barack"]["contb\_receipt\_amt"]

mccain\_contributions = df[df["cand\_nm"] == "McCain, John S"]["contb\_receipt\_amt"]

t\_statistic, p\_value = stats.ttest\_ind(obama\_contributions, mccain\_contributions, equal\_var=False)

print("Welch's t-test:")

print("t-statistic:", t\_statistic)

print("p-value:", p\_value)

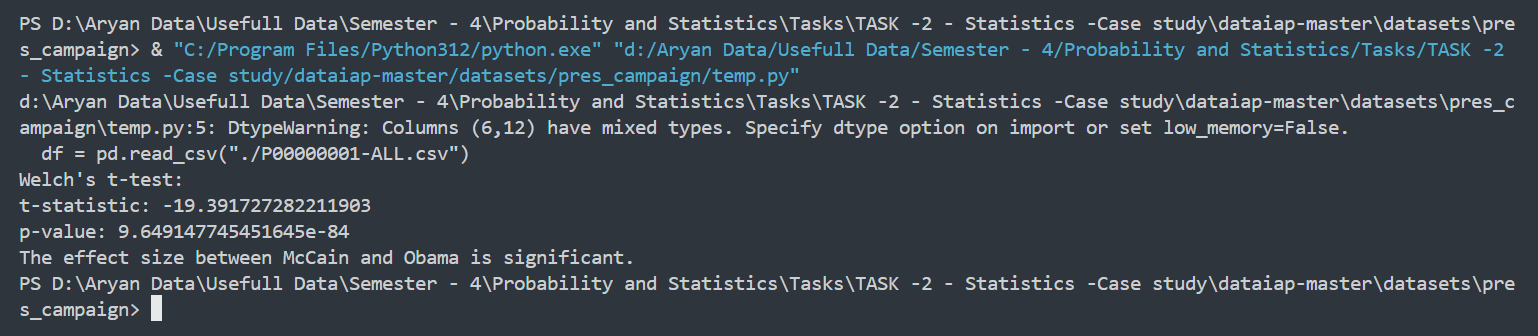
if p\_value < 0.05:

print("The effect size between McCain and Obama is significant.")

else:

print("The effect size between McCain and Obama is not significant.")

**Output :-**



**Question – 6 :-**

Look at scatter plots of other variables vs. YPLL. We found the percent of children eligible for school lunch to be3

alarmingly correlated with YPLL!

**Code :-**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

ypll\_df = pd.read\_csv("../county\_health\_rankings/ypll.csv")

additional\_measures\_clean = pd.read\_csv("../county\_health\_rankings/additional\_measures\_cleaned.csv")

wanted\_data = pd.merge(ypll\_df, additional\_measures\_clean , on="FIPS")

plt.figure(figsize=(40,20))

plt.scatter(x = wanted\_data["% Free lunch"] , y = wanted\_data["YPLL Rate"])

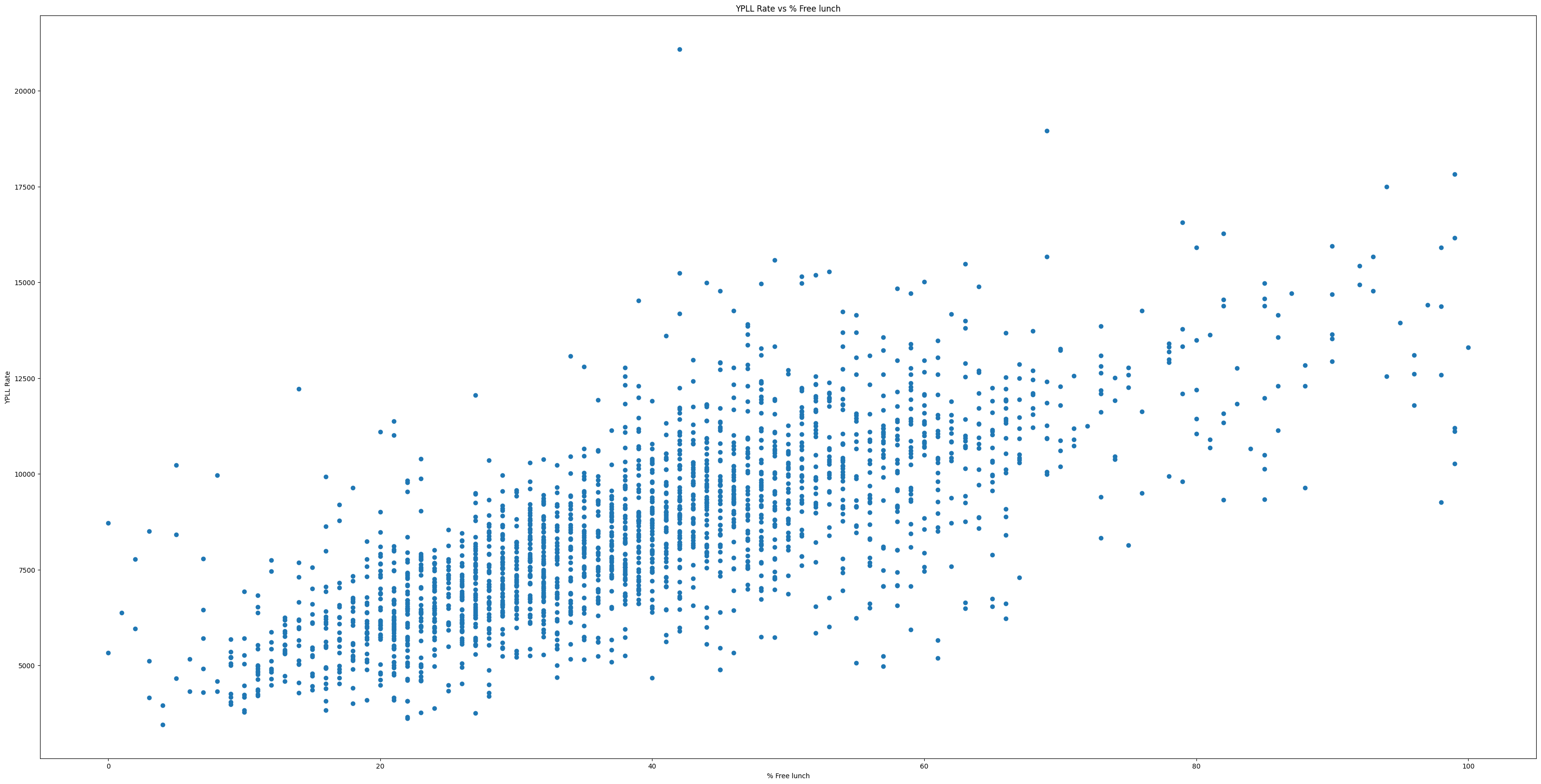
plt.xlabel("% Free lunch")

plt.ylabel("YPLL Rate")

plt.title("YPLL Rate vs % Free lunch")

plt.show()

**Output :-**

****

**Question – 7 :-**

Run the correlations for percentage of population under 18 years of age and median household income.

**Code :-**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

additional\_measures\_clean = pd.read\_csv("../county\_health\_rankings/additional\_measures\_cleaned.csv")

plt.figure(figsize=(40, 20))

plt.scatter(x=additional\_measures\_clean["< 18"],y=additional\_measures\_clean["median household income"],)

plt.xlabel("% Population Under 18")

plt.ylabel("Median Household Income")

plt.title("Median Household Income vs % Population Under 18")

x = additional\_measures\_clean["< 18"]

y = additional\_measures\_clean["median household income"]

coefficients = np.polyfit(x, y, 1)

poly = np.poly1d(coefficients)

plt.plot(x, poly(x), color="red")

plt.show()

**Output :-**

