Marwadi University Marwadi Chandarana Group NAAC A+	**Marwadi University** **Faculty of Technology** **Department of Information and Communication Technology**	
Subject: Probability and Statistics (01CT1401)	**Aim: Hypothesis Testing**	
Task - 3	**Date:-** 19-04-2024	**Enrollment No:-** 92200133030

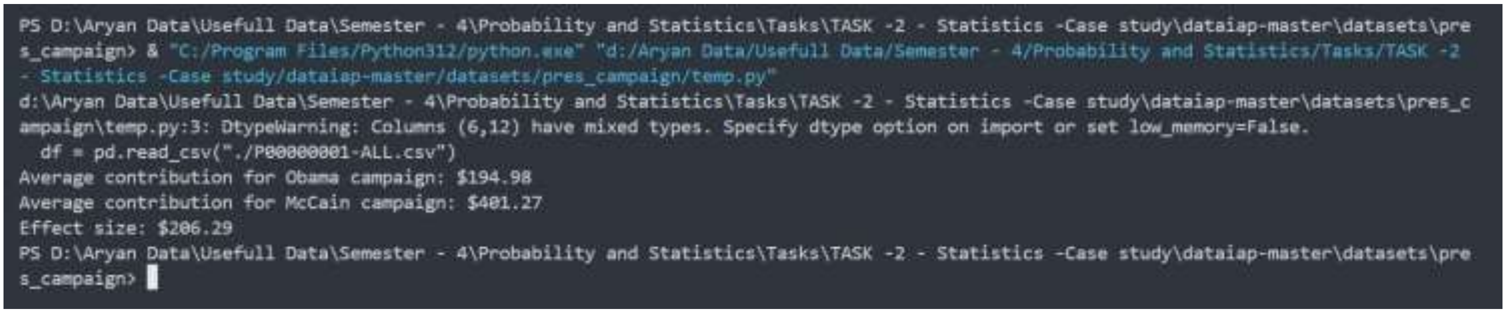
Question – 1 :-

Compute the average campaign contribution for the Obama and McCain campaigns from the dataset in day 1. What′s 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 :-



```
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> & "C:/Program Files/Python312/python.exe" "d:/Aryan Data/Usefull Data/Semester - 4/Probability and Statistics/Tasks/TASK -2
- Statistics -Case study/dataiap-master/datasets/pres_campaign/temp.py"
d:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pres_c
ampaign\temp.py:3: DtypeWarning: Columns (6,12) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("./P00000001-ALL.csv")
Average contribution for Obama campaign: $194.98
Average contribution for McCain campaign: $401.27
Effect size: $206.29
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> 
```

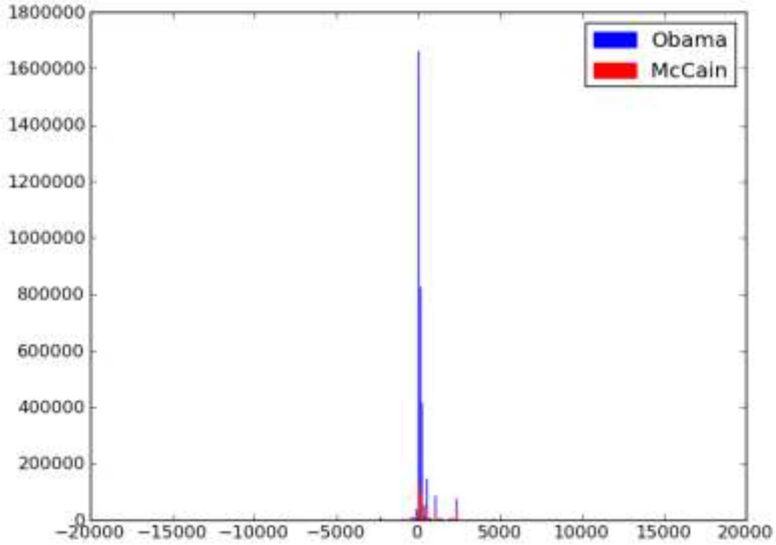
Question-2 :-

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:-



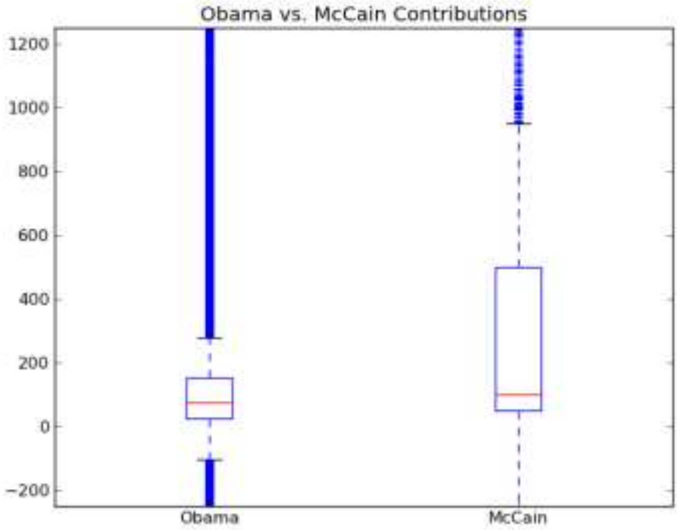
Question-3 :-

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 :-

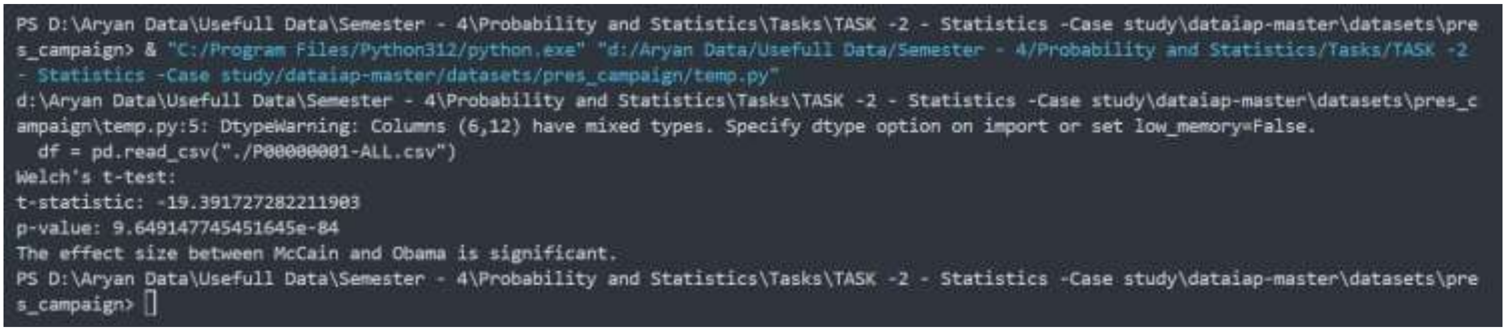
Run Welch′s T-test on the campaign data. Is the effect size between McCain and Obama significant? By our measurements, the p-value reported is within rounding error of 0. That′s significant by anyone′s measure: there′s a near-nonexistant chance we′re seeing this difference between the candidates by some 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 :-



```
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> & "C:/Program Files/Python312/python.exe" "d:/Aryan Data/Usefull Data/Semester - 4/Probability and Statistics/Tasks/TASK -2
- Statistics -Case study/dataiap-master/datasets/pres_campaign/temp.py"
d:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pres_c
ampaign\temp.py:5: DtypeWarning: Columns (6,12) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("./P00000001-ALL.csv")
Welch's t-test:
t-statistic: -19.391727282211903
p-value: 9.649147745451645e-84
The effect size between McCain and Obama is significant.
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> []
```

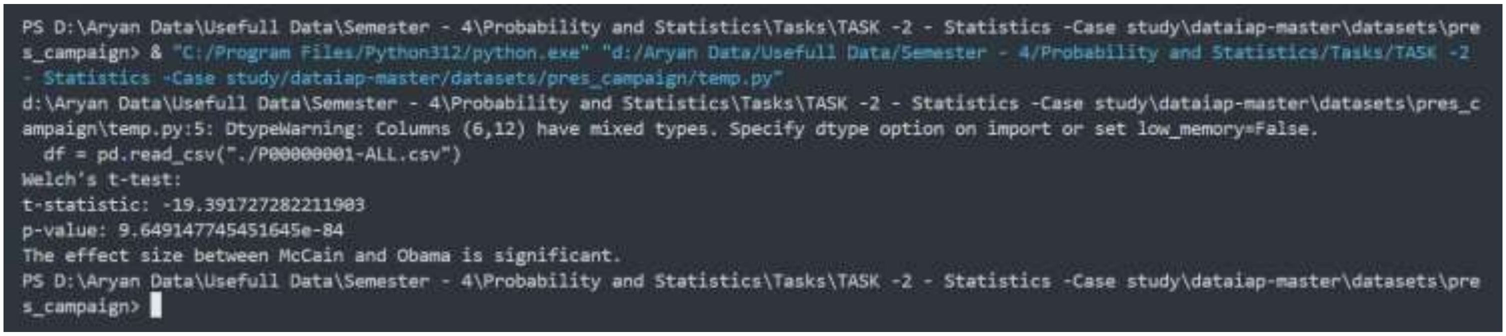
Question-5 :-

since we shouldn′t be using Welch′s 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 :-



```
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> & "C:/Program Files/Python312/python.exe" "d:/Aryan Data/Usefull Data/Semester - 4/Probability and Statistics/Tasks/TASK -2
- Statistics -Case study/dataiap-master/datasets/pres_campaign/temp.py"
d:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pres_c
ampaign\temp.py:5: DtypeWarning: Columns (6,12) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("./P00000001-ALL.csv")
Welch's t-test:
t-statistic: -19.391727282211903
p-value: 9.649147745451645e-84
The effect size between McCain and Obama is significant.
PS D:\Aryan Data\Usefull Data\Semester - 4\Probability and Statistics\Tasks\TASK -2 - Statistics -Case study\dataiap-master\datasets\pre
s_campaign> 
```

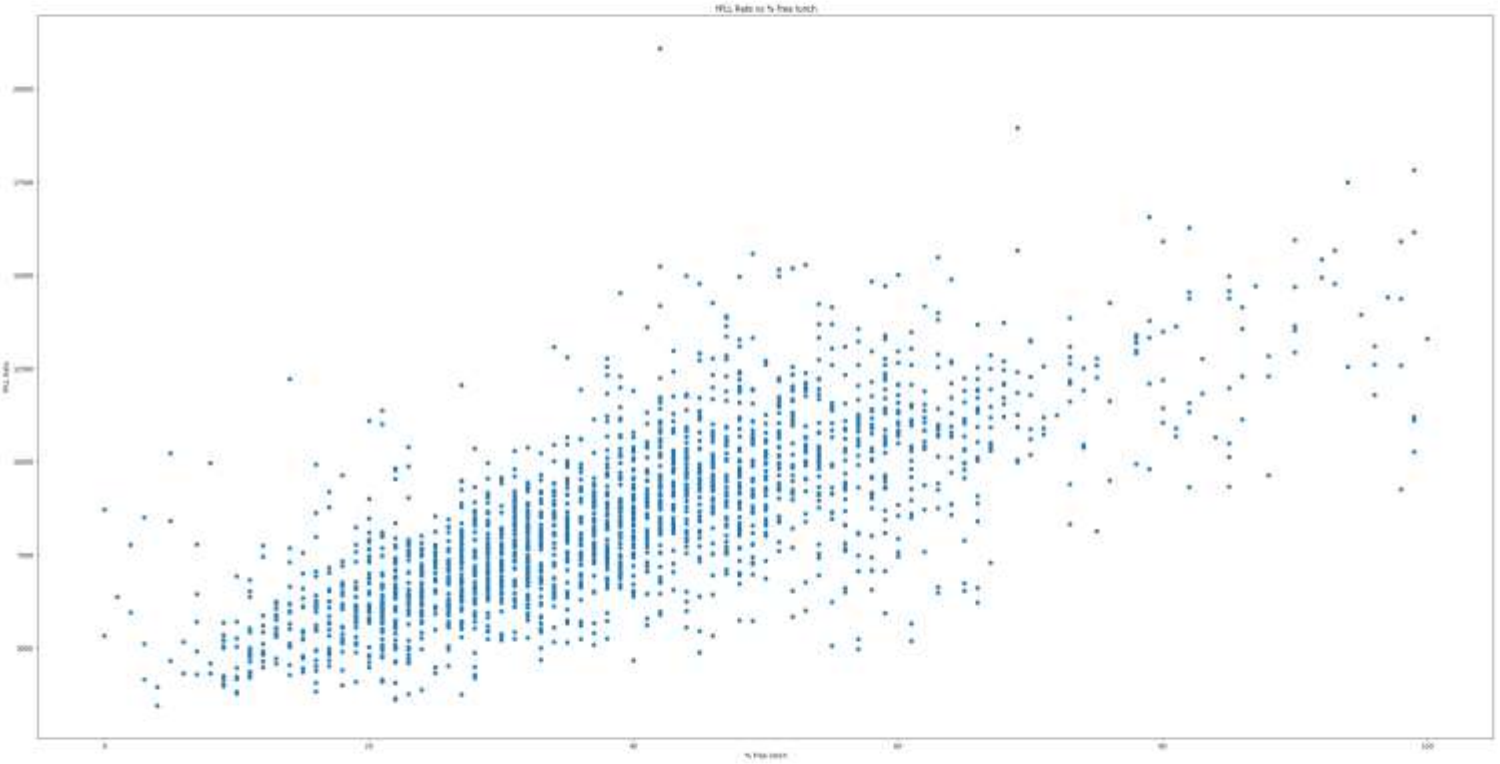
Question – 6 :-

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 :-

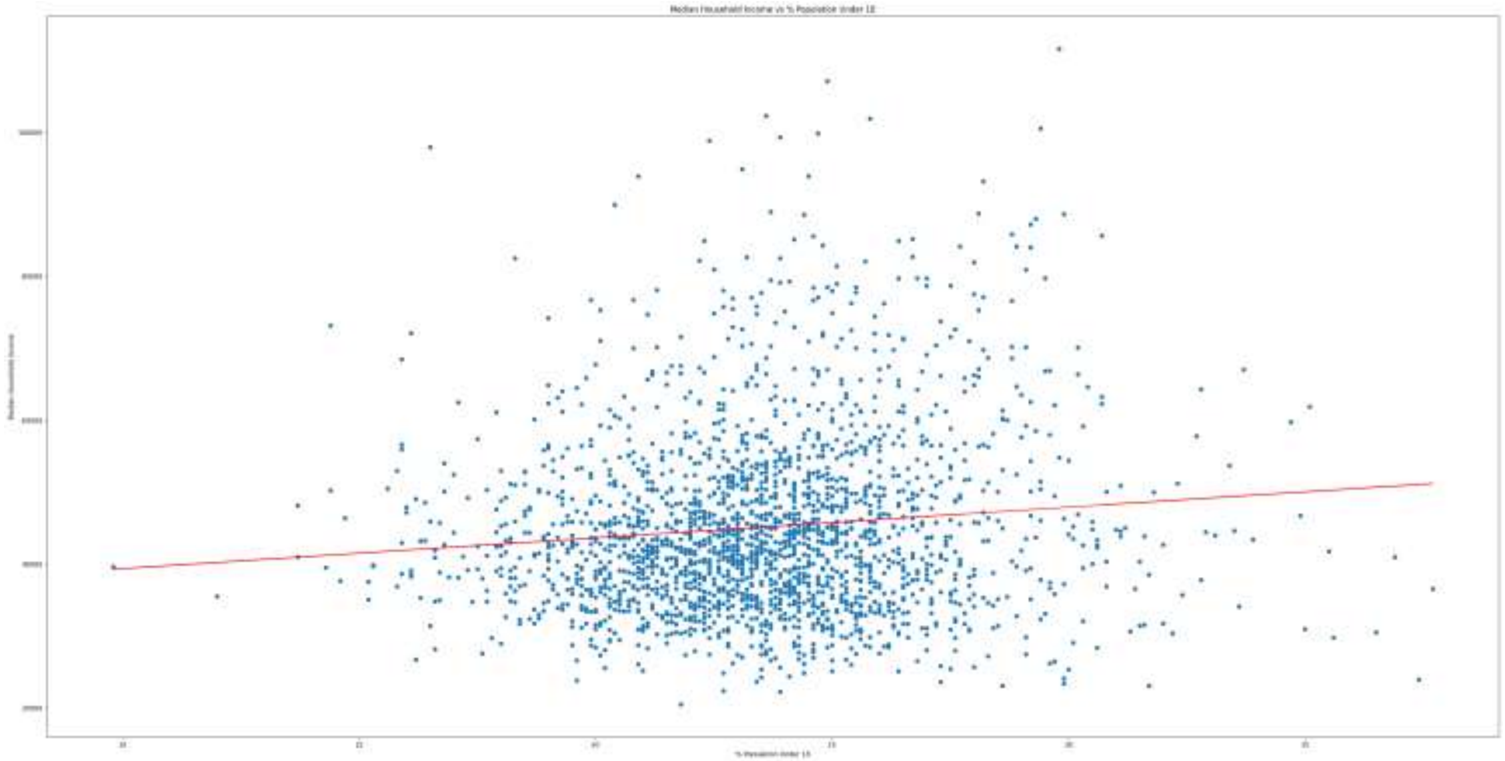
Run the correlations for percentage of population under 18 years of age and median household income.

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 :-


Median Household Income vs % Population Under 18
Median Household Income
% Population Under 18