**Python code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from scipy import stats

df = pd.read\_csv('/content/customer\_data.csv')

education\_gender\_counts = df.groupby(['education', 'gender']).size().reset\_index(name='Count')

education\_gender\_counts

# Create a bar plot

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

plt.bar(education\_gender\_counts['education'] + ' - ' + education\_gender\_counts['gender'],

        education\_gender\_counts['Count'])

plt.xlabel('Education - Gender')

plt.ylabel('Count')

plt.title('Counts of Men and Women from Each Education Type')

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

df = df[['gender','education','income']]

df

df.gender.replace({'Female':0,'Male':1},inplace=True)

filtered\_df = df[df['education'].isin([‘Bachelor’ ,'Master', 'PhD'])]

filtered\_df['income\_in\_k'] = np.round(filtered\_df['income'] / 1000, 1)

filtered\_df.drop('income',inplace=True,axis=1)

filtered\_df.drop('education',inplace=True,axis=1)

filtered\_df

filtered\_df.reset\_index(inplace=True,drop=True)

filtered\_df.info()

filtered\_df.describe()

filtered\_df.gender.value\_counts()

# Splitting the data into male and female groups

male\_income = filtered\_df[filtered\_df['gender'] == 1]['income\_in\_k']

female\_income = filtered\_df[filtered\_df['gender'] == 0]['income\_in\_k']

# Plot histogram for men's income

plt.hist(male\_income, bins=10, density=True, alpha=0.5, label='Men')

plt.xlabel('Income')

plt.ylabel('Density')

plt.title("Histogram of Men's Income")

plt.legend()

plt.show()

# Plot histogram for women's income

plt.hist(female\_income, bins=10, density=True, alpha=0.5, label='Women')

plt.xlabel('Income')

plt.ylabel('Density')

plt.title("Histogram of Women's Income")

plt.legend()

plt.show()

# Create Q-Q plot for men's income

plt.figure()

stats.probplot(male\_income, dist="norm", plot=plt)

plt.title("Q-Q Plot of Men's Income")

plt.show()

# Create Q-Q plot for women's income

plt.figure()

stats.probplot(female\_income, dist="norm", plot=plt)

plt.title("Q-Q Plot of Women's Income")

plt.show()

#checking for normal distribution using KS

kstest\_result\_male = stats.kstest(male\_income, 'norm')

ks\_statistic\_male = kstest\_result\_male.statistic

p\_value\_male = kstest\_result\_male.pvalue

kstest\_result\_female = stats.kstest(female\_income, 'norm')

ks\_statistic\_female = kstest\_result\_female.statistic

p\_value\_female = kstest\_result\_female.pvalue

alpha = 0.05 # significance level

if p\_value\_male > alpha:

print("The data of male appears to be normally distributed.")

else:

print("The data of male does not appear to be normally distributed.")

if p\_value\_female > alpha:

print("The data of female appears to be normally distributed.")

else:

print("The data of female does not appear to be normally distributed.")

# Performing the Mann-Whitney U test for unpaired populations

statistic, p\_value = stats.mannwhitneyu(male\_income, female\_income, alternative='two-sided')

# Checking the results

if p\_value < 0.05:

    print("There is a significant difference between male and female incomes.")

else:

    print("There is no significant difference between male and female incomes.")

print(p\_value)

**שאלת המחקר שלנו:**

**האם קיימים הבדלים בשכר בין נשים וגברים בעלי השכלה אקדמית?**

0: לא קיימים הבדלים בין נשים וגברים בעלי השכלה אקדמיתH

H1: קיימים הבדלים בין נשים וגברים בעלי השכלה אקדמית.

**מקור הנתונים**:

השתמשנו בנתונים שמצאנו באתר KAGGLE

על הדאטה:

מידע של לקוחות של חברה מסוימת, החברה אספה את הנתונים הבאים: הכנסות, הוצאות, מין, השכלה, מדינה, שם לקוח ותדירות רכישה.

**אנאליזה:**

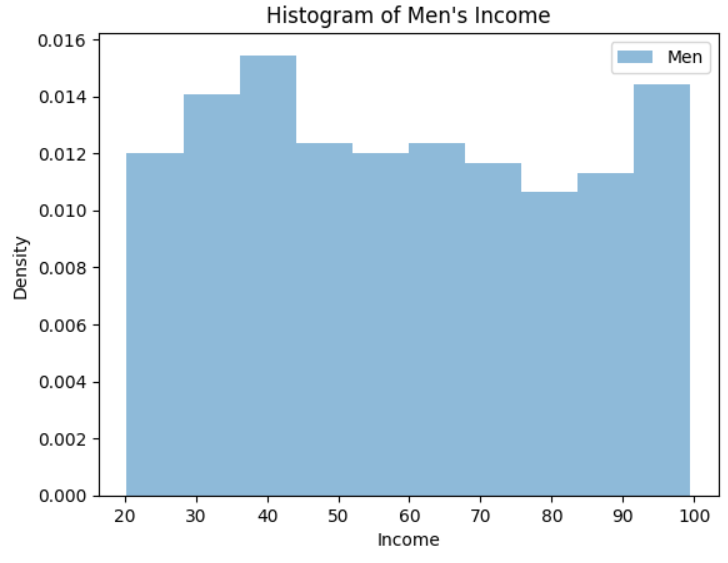
בחרנו להוריד את כל העמודות שלא יעזרו לבדיקת שאלת המחקר ו"הסרה נאיבית" של עמודות כמו שם וכו'.

פיצלנו את הדאטה לגברים ולנשים, לאחר הסרת בעלי השכלה לא אקדמית נותרו סה"כ 755 דגימות, .

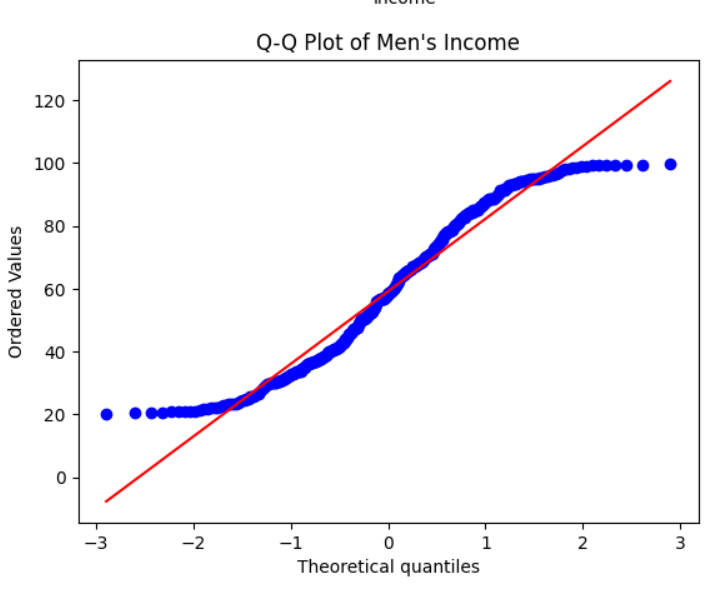
בדקנו הטיות שונות שיכלו לנבוע מכמות לא שווה של גברים ונשים בעלי השכלה אקדמית שונה.

בדקנו נורמליות ע"י היסטוגרמה וביצענו מבחן קוגלנוב – סמירנוב לבדיקת התאמה להתפלגות הנורמלית, אחרי שגילינו שההתפלגות השכר של שתי האוכלוסיות אינה שייכת להתפלגות הנורמלית, המשכנו למבחן שיקבע האם יש הבדל ברמות ההכנסה בין גברים אקדמיים לנשים אקדמאיות.

הרצנו מבחן ווילקוקסון לאוכלוסיות לא מזווגות, קבענו ר.מ 0.05 למבחן.

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, line, diagram, screenshot

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

**תוצאות:**

קיבלנו ערך גבוה מאוד (0.8628) ע"מ לדחות את ההשערה הרבה יותר גדול מאלפא, לא נדחה את השערת האפס כלומר לא ניתן להסיק כי יש הבדל בהכנסות בין גברים ונשים אקדמאיים.

מגישים:

טל קריספין

Talkn1997@gmail.com

דניאל אזולאי

doonil1996@gmail.com