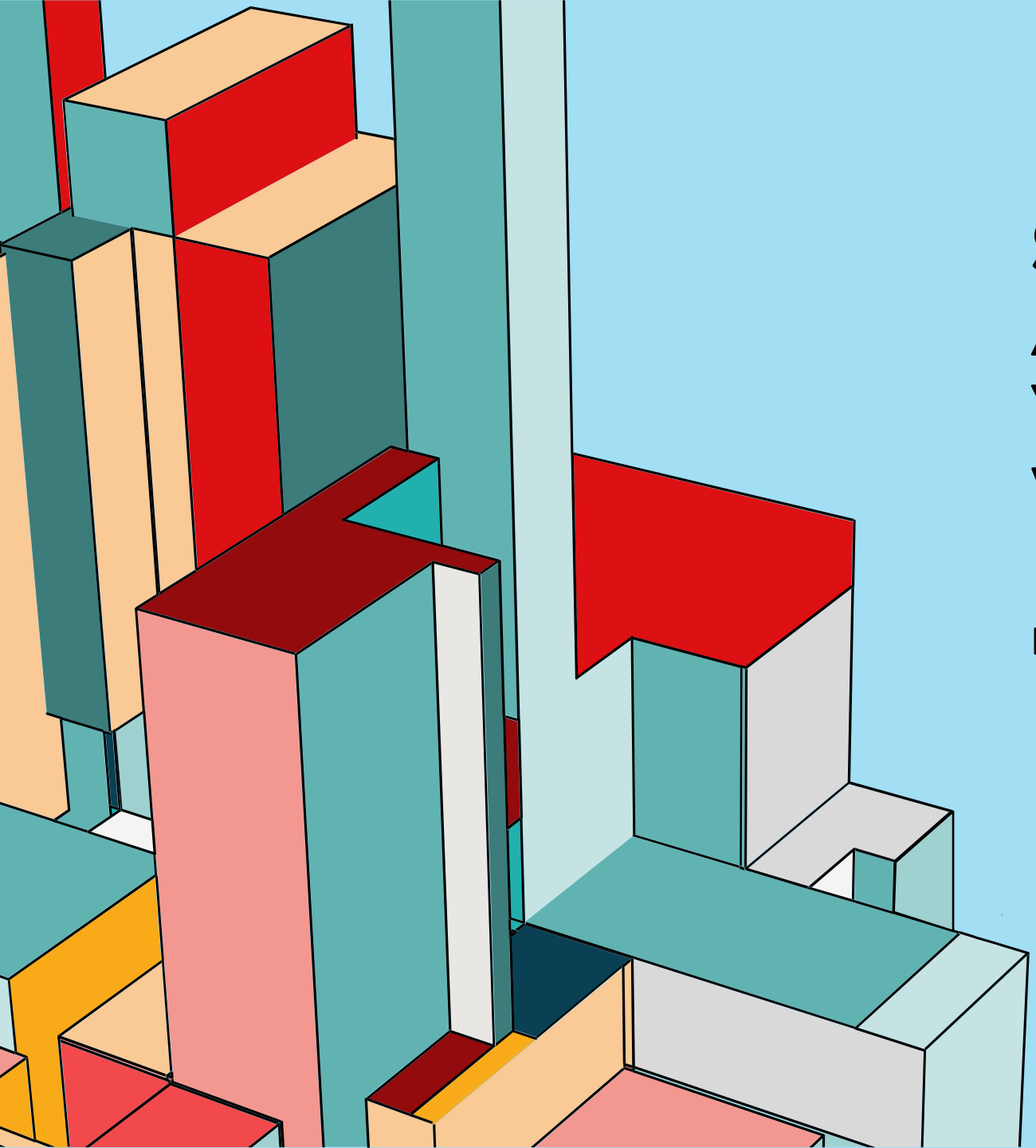
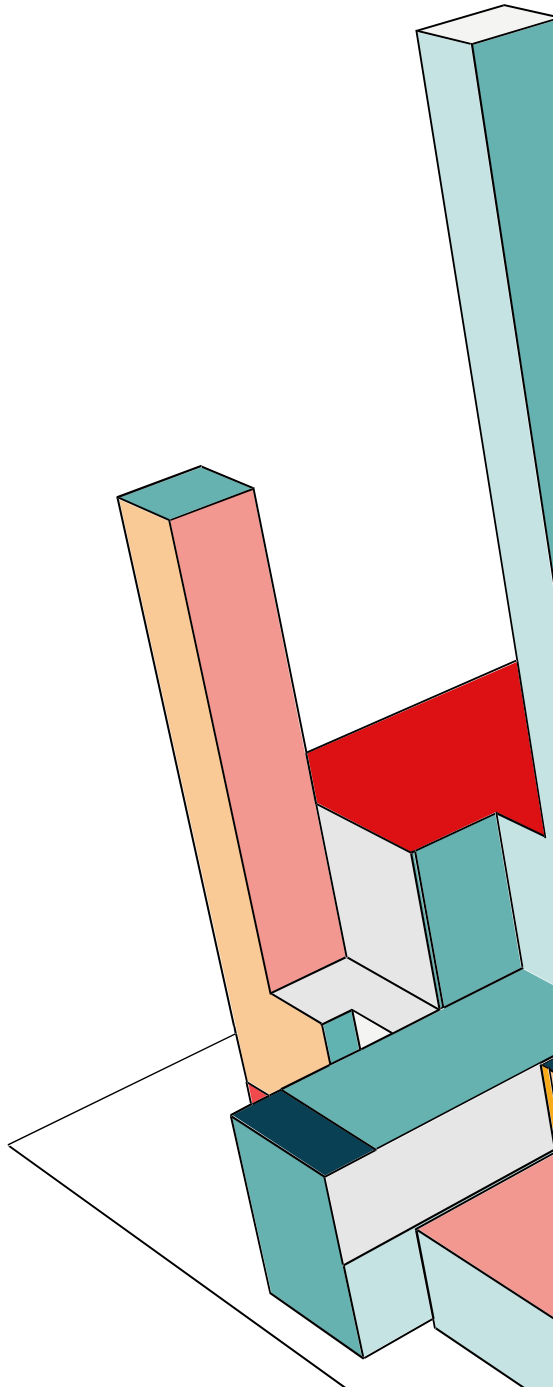


STATISTICAL ANALYSIS AND VISUALIZATION WITH PYTHON

BLOCH CARLO

AGENDA

- Introduction
- Data Generation
- Descriptive statistics
- Data Visualization
- Correlation analysis
- Inferential statistics
- Final considerations



INTRODUCTION

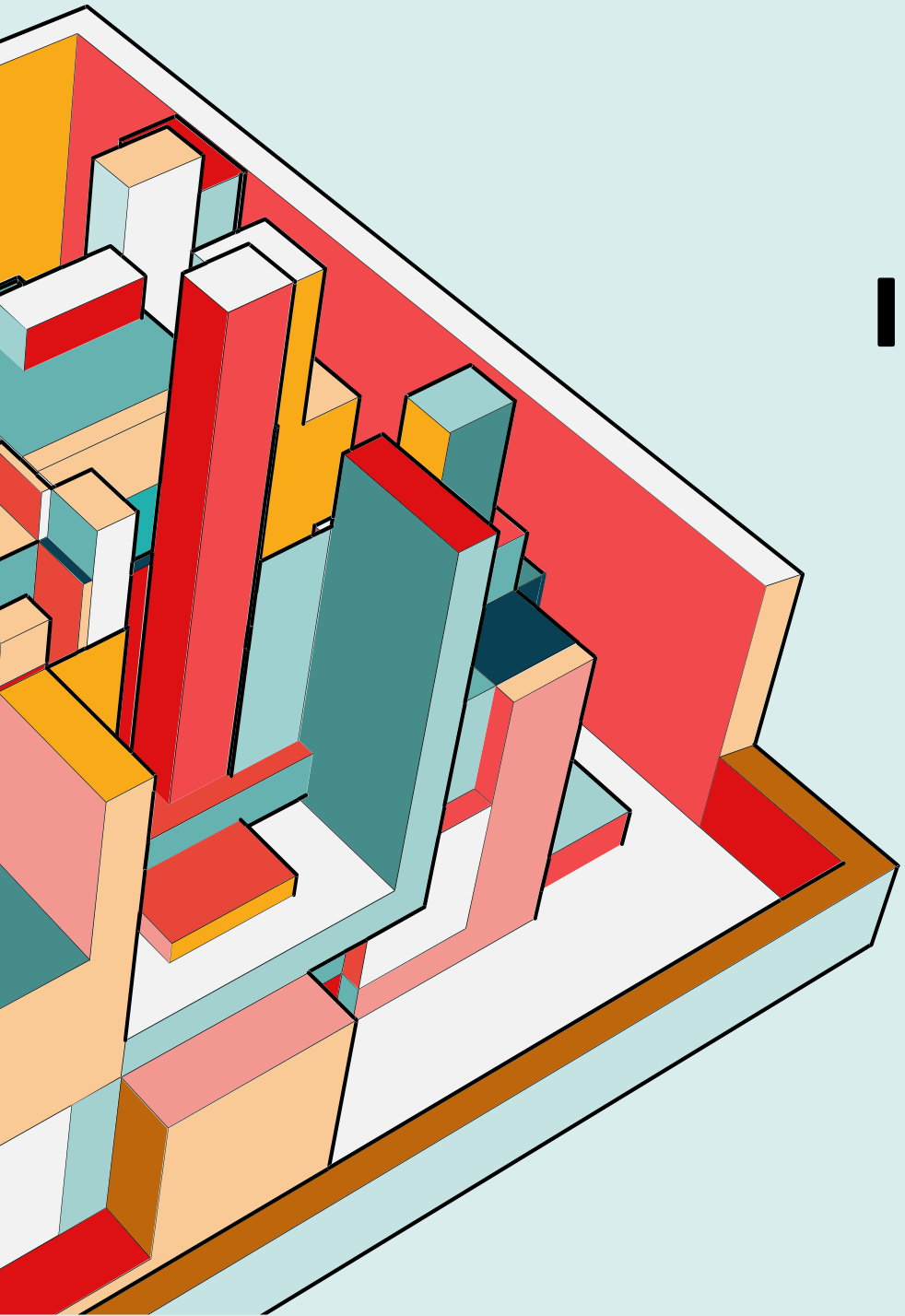
- This project aims to perform statistical analysis and data visualization using Python, generating a synthetic dataset, performing descriptive and inferential statistics, and visualizing the results using Matplotlib and Seaborn.
- Deliverables are a presentation summarizing your findings, visualizations, and interpretations addressing the technical questions provided.

DATA GENERATION

Use Numpy to create a dataset with 1000 samples

The dataset should include the following columns: Age, Height, Weight, Gender, and Income.

Age: Normally distributed with a mean of 35 and a standard deviation of 10
Height: Normally distributed with a mean of 170 cm and a standard deviation of 15 cm
Weight: Normally distributed with a mean of 70 kg and a standard deviation of 10 kg
Gender: Randomly assigned with 50% probability for 'Male' and 'Female'
Income: Normally distributed with a mean of 50,000 and a standard deviation of 15,000



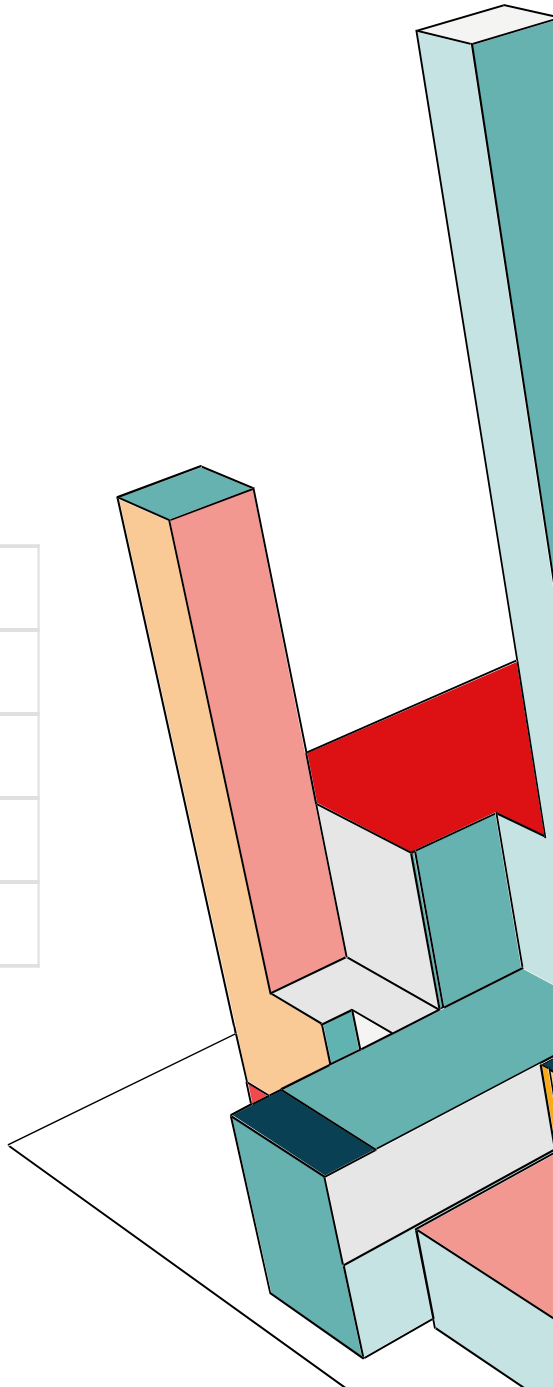
INFERENTIAL STATISTICS

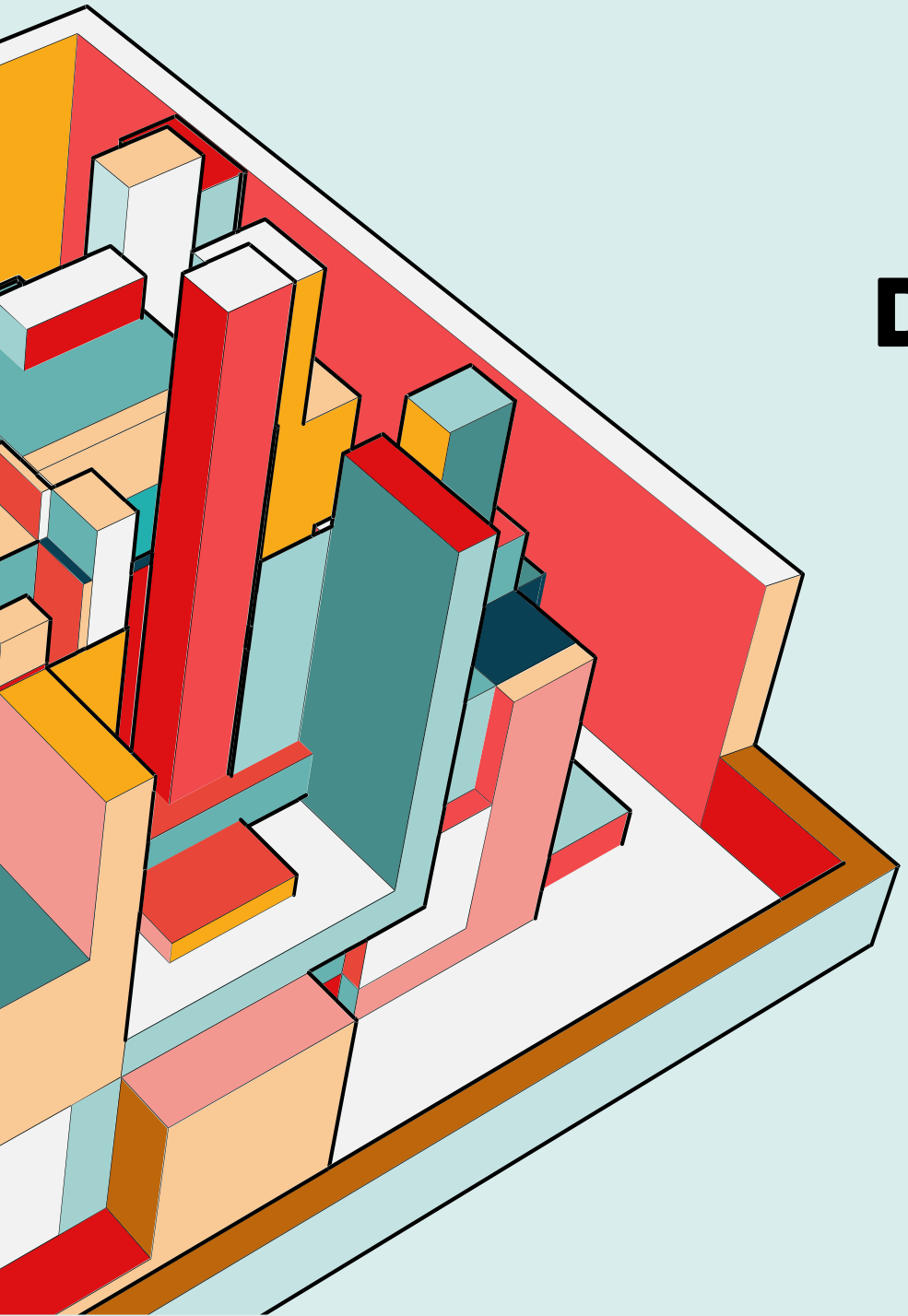
DESCRIPTIVE STATISTICS

Basic descriptive statistics

- Calculate mean, median, standard deviation, and variance for Age, Height, Weight, and Income
- Calculate the mode for Gender.

for column	AGE	Mean=34.46,	Median=34.5,	Standard deviation=9.8,	Variance =96.0	
for column	HEIGHT	Mean=1.7,	Median=1.69,	Standard deviation=0.14,	Variance =0.02	
for column	WEIGHT	Mean=70.17,	Median=69.72,	Standard deviation=10.19,	Variance =103.91	
for column	INCOME	Mean=49583.06,	Median=49065.31,	Standard deviation=15240.22,	Variance =232264164.8	
for column	GENDER	Mode=Male				

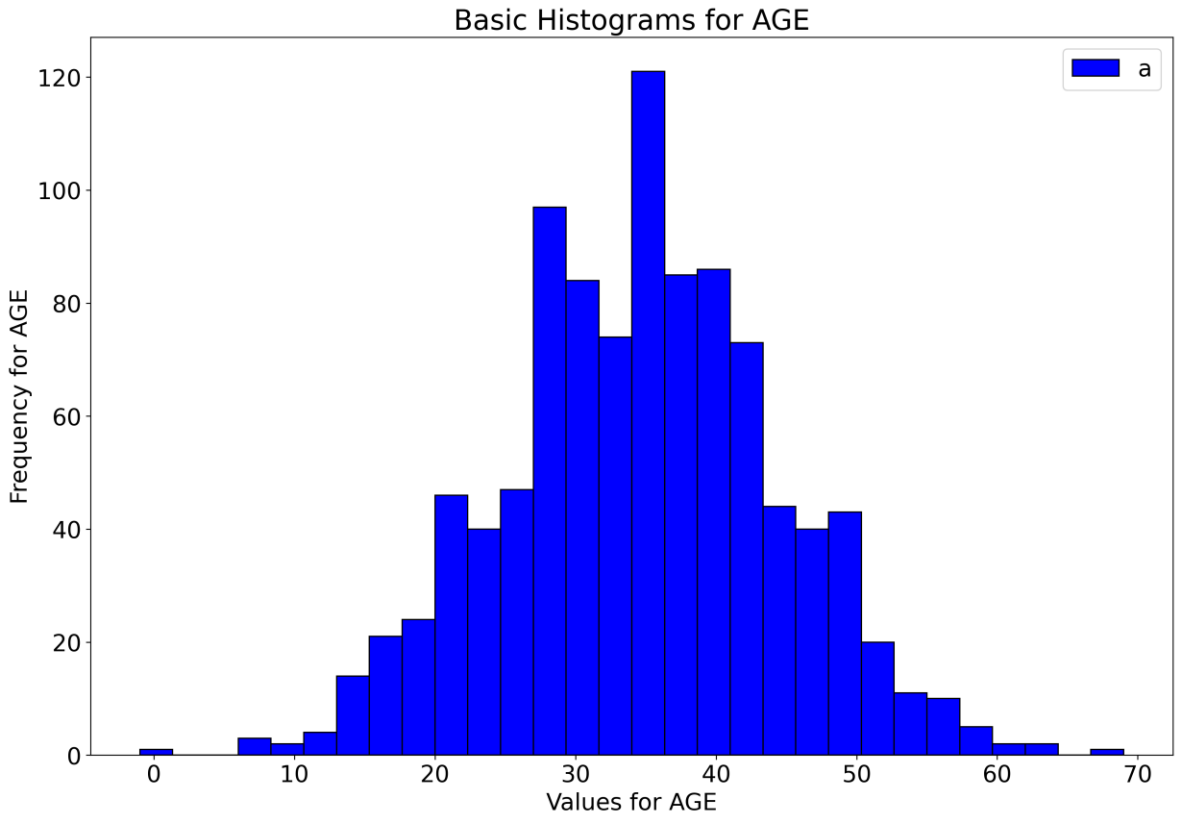


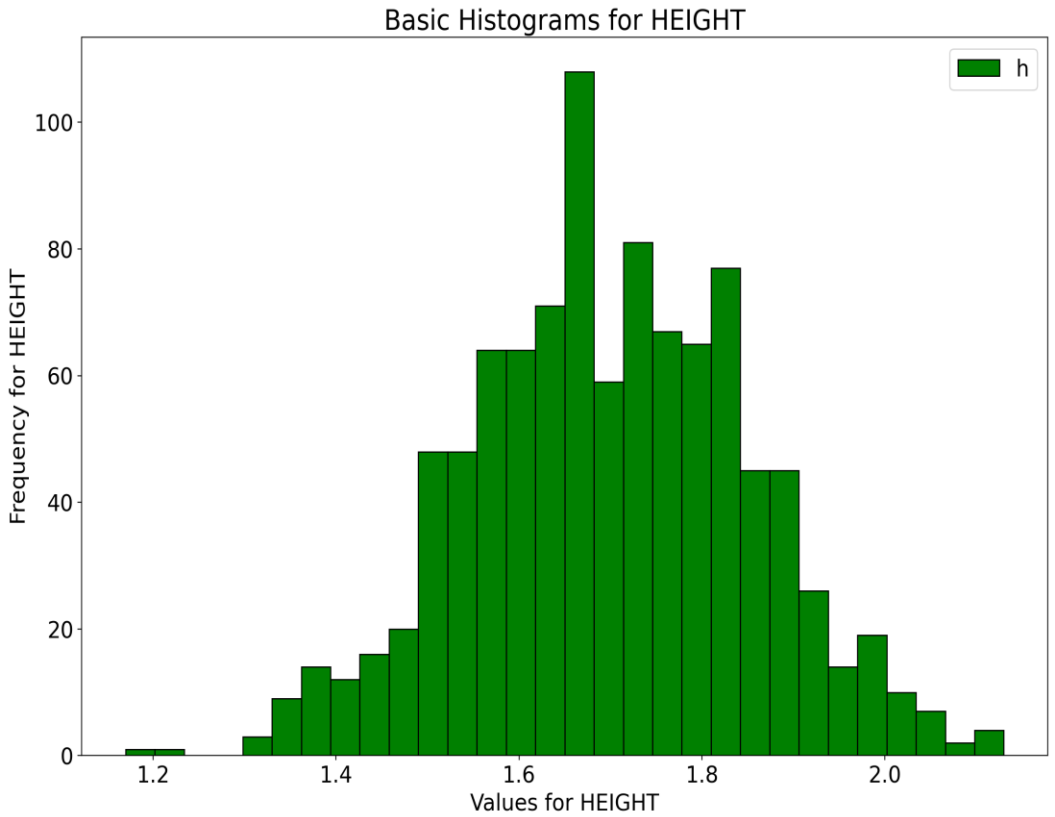


DATA VISUALIZATION

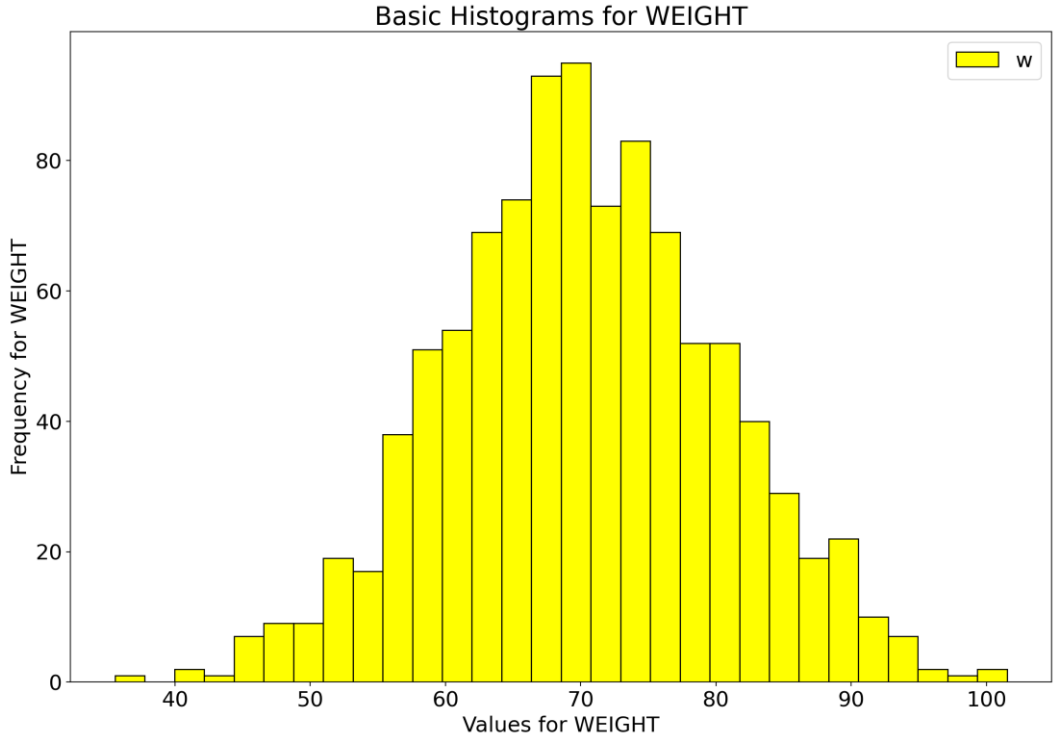
Results following

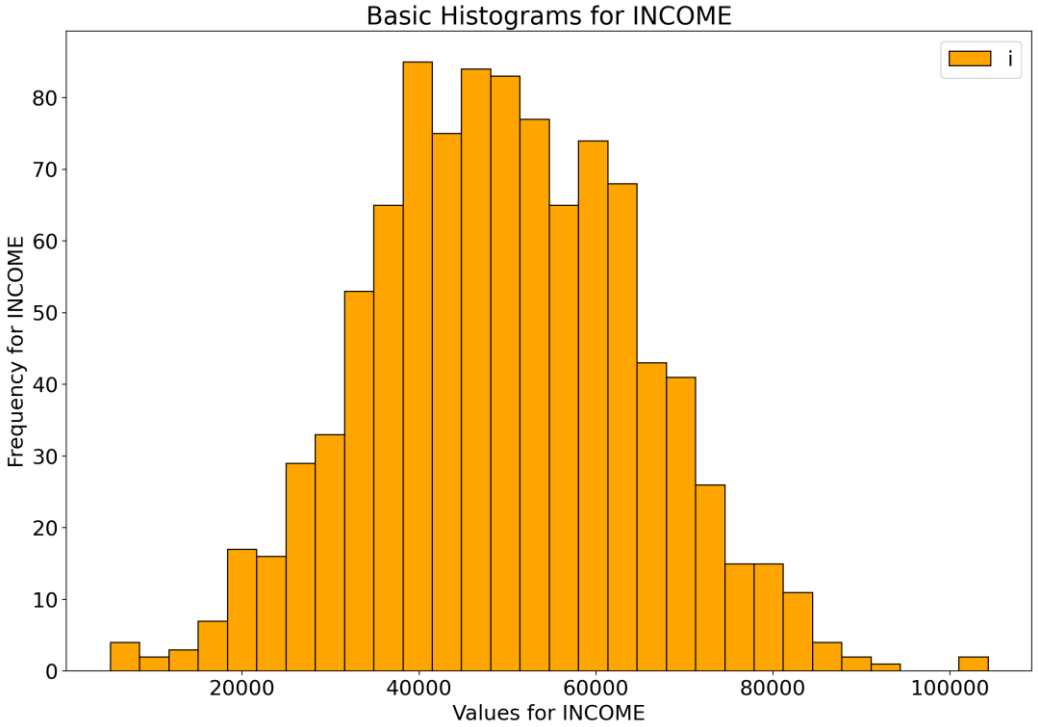
AGE, HEIGHT HISTOGRAMS



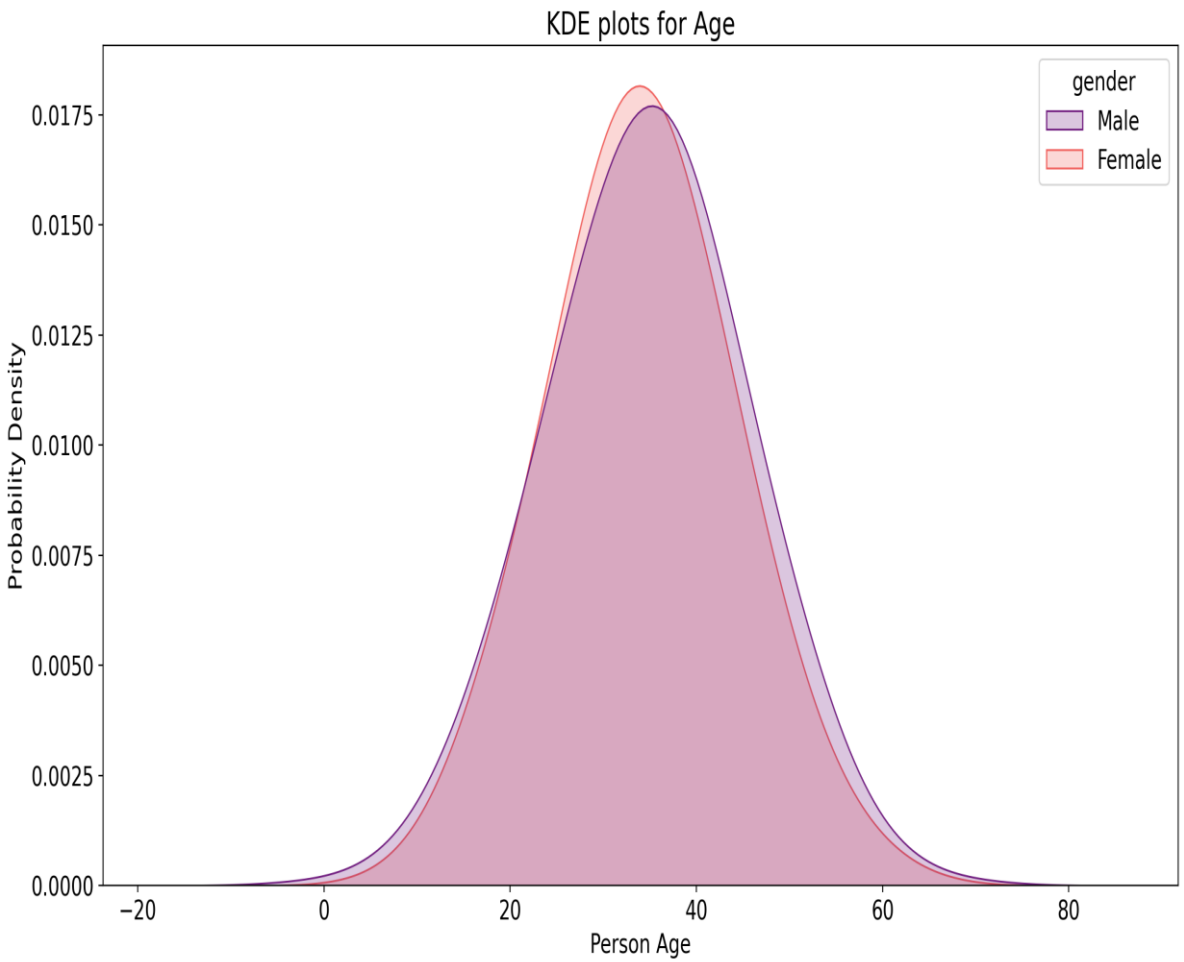


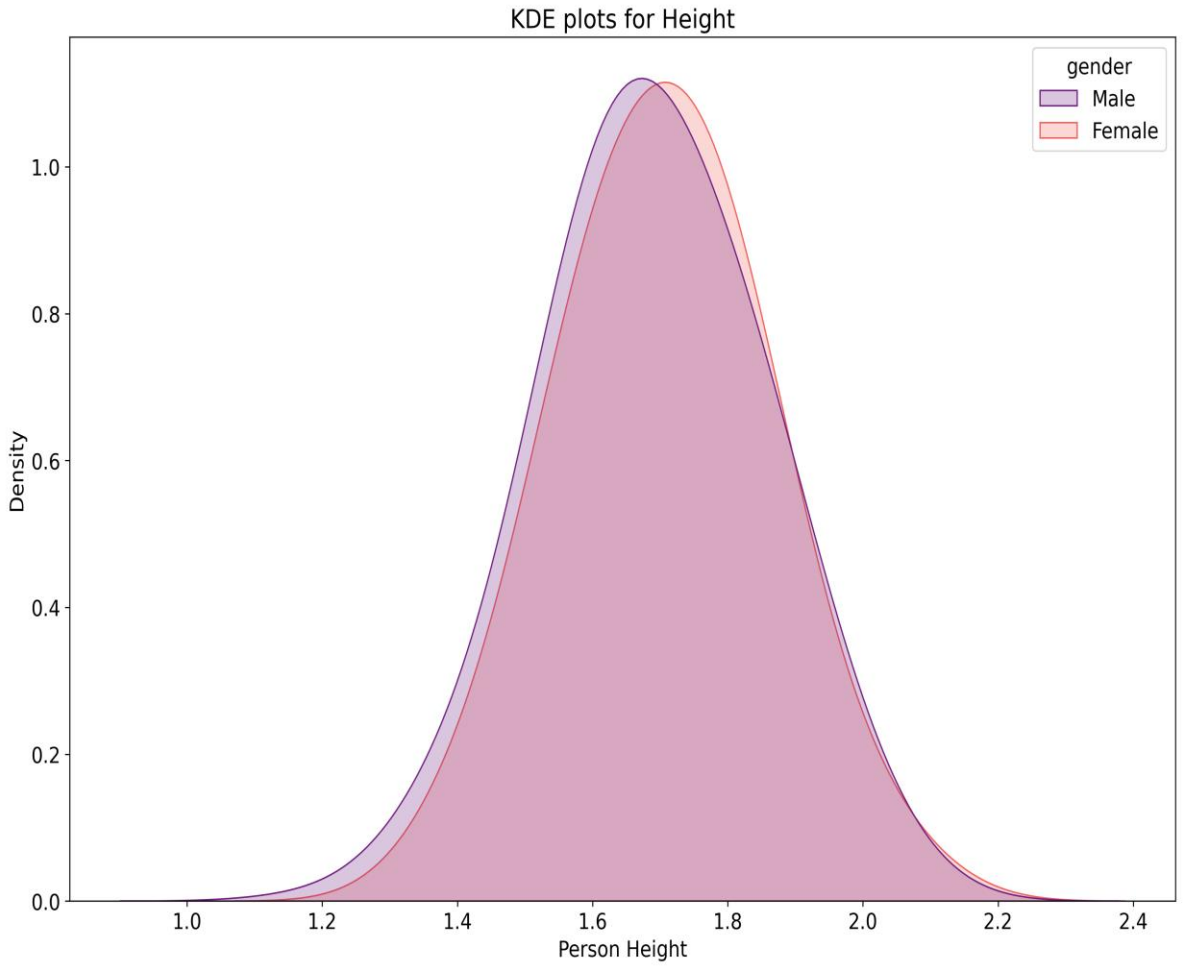
WEIGHT, INCOME HISTOGRAMS



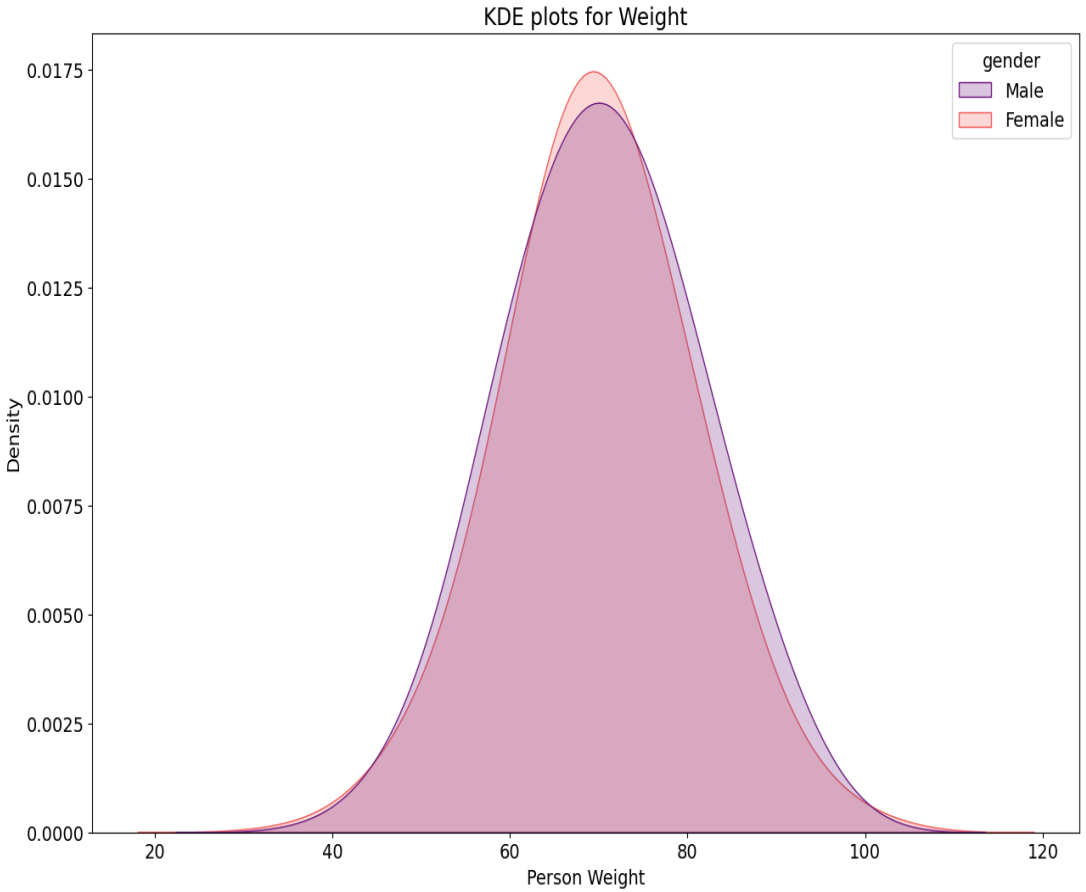


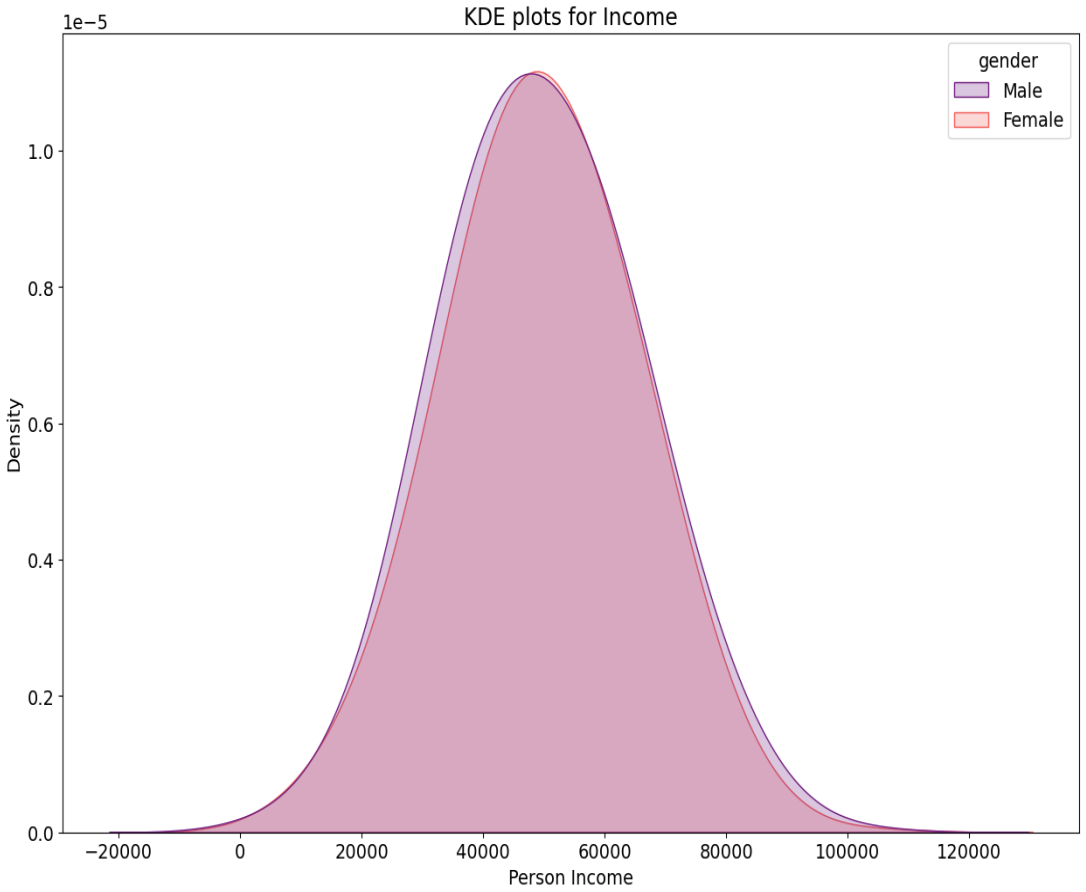
KDE PLOTS


KDE plots for Age
gender
Male
Female
Probability Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−20
0
20
40
60
80
Person Age

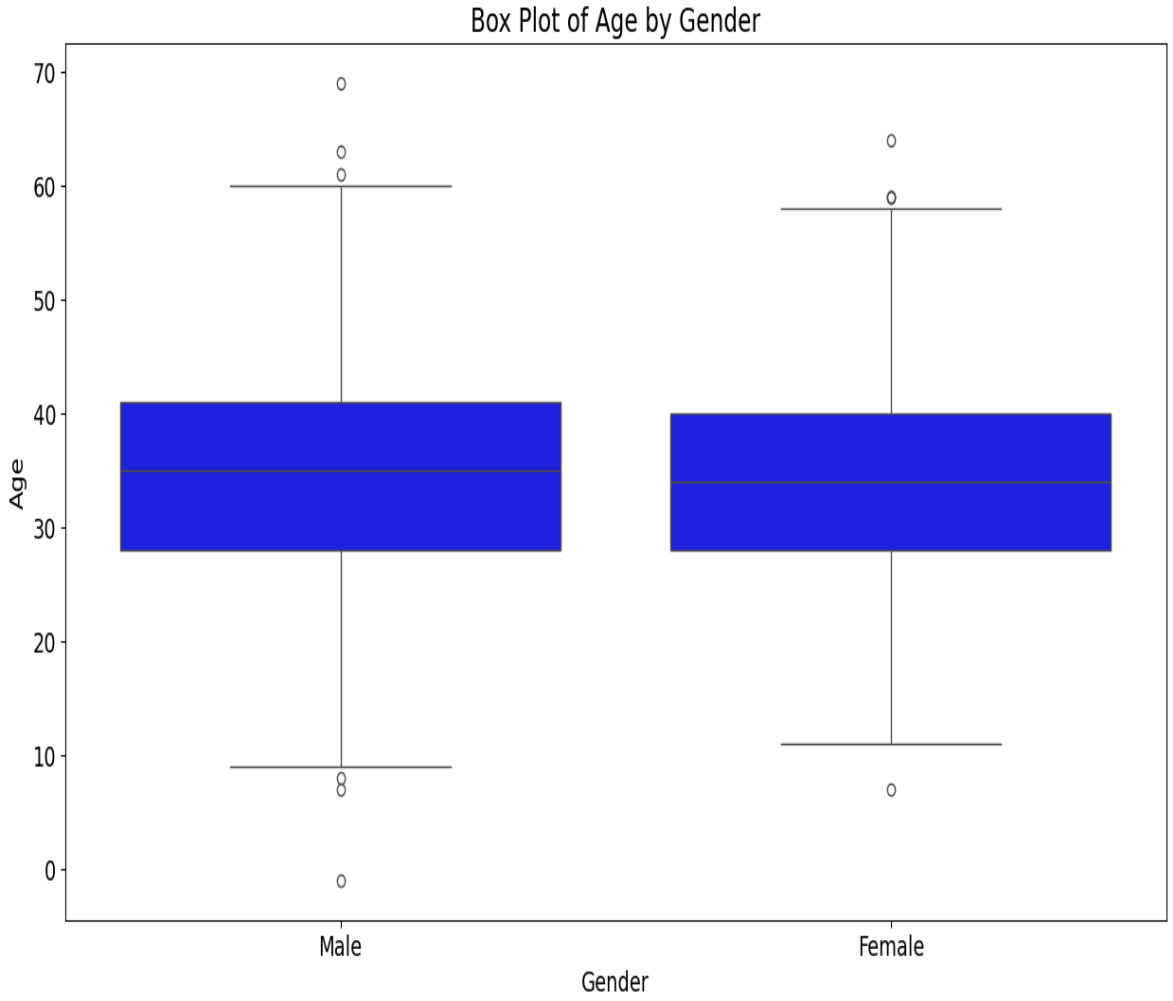

KDE plots for Height
gender
Male
Female
Density
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Person Height

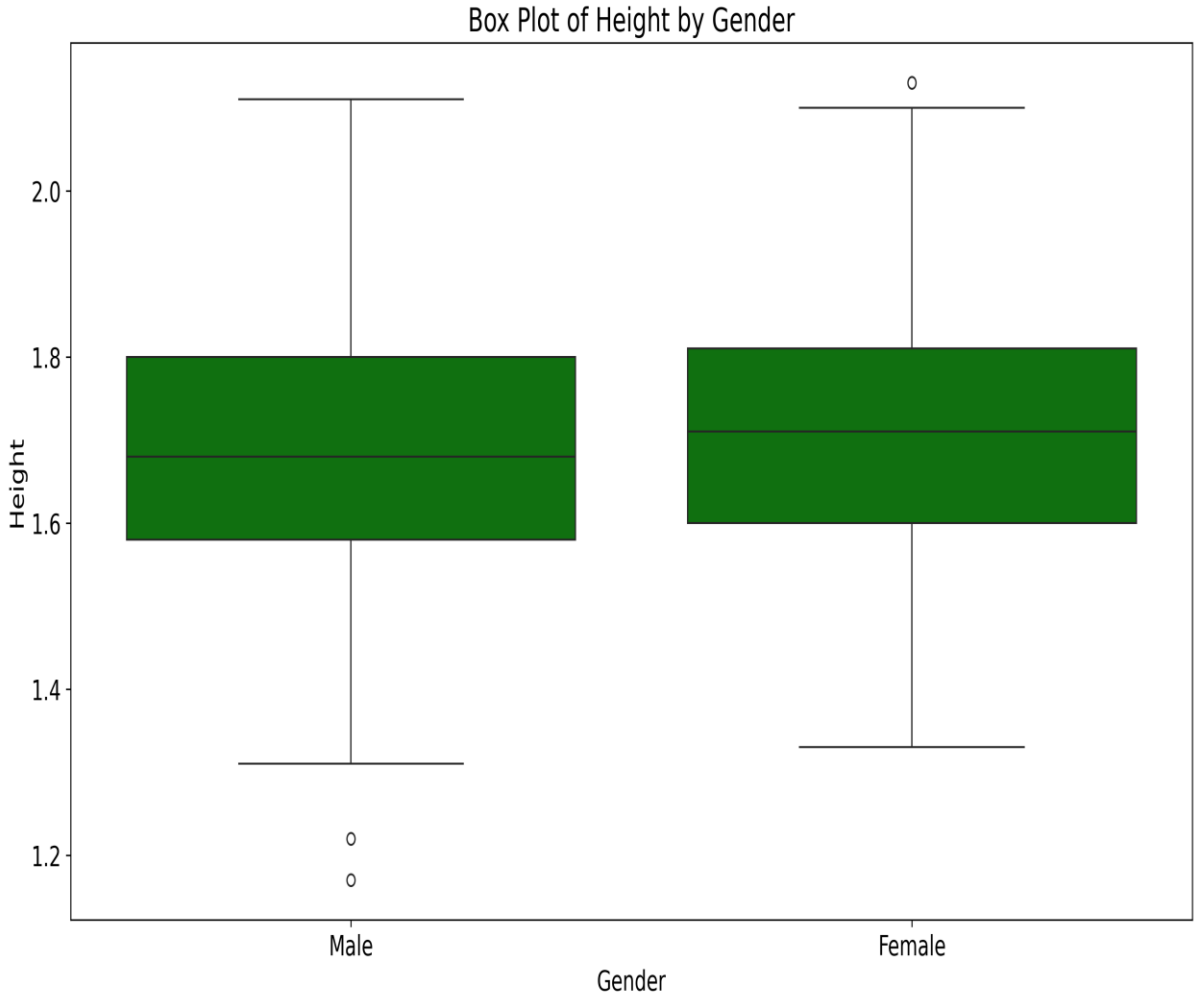
KDE PLOTS


KDE plots for Weight
gender
Male
Female
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
20
40
60
80
100
120
Person Weight

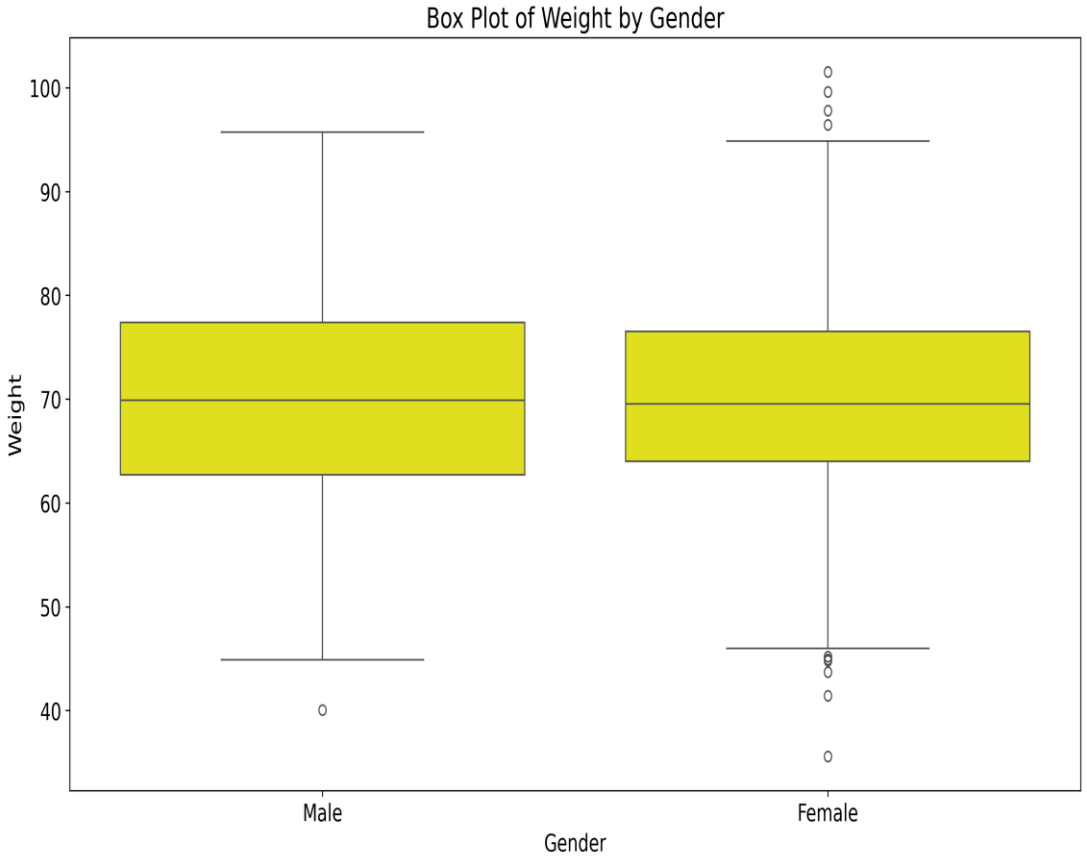

KDE plots for Income
1e−5
gender
Male
Female
Density
1.0
0.8
0.6
0.4
0.2
0.0
−20000
0
20000
40000
60000
80000
100000
120000
Person Income

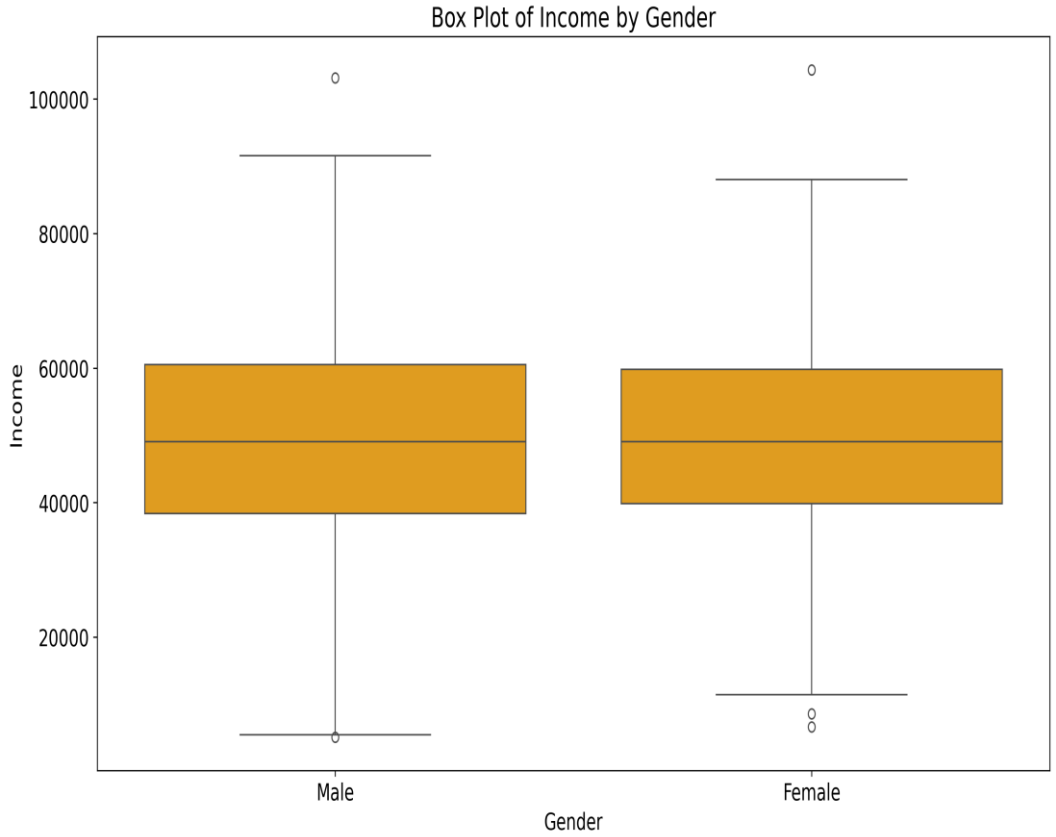
BAR PLOTS – HIGHLIGHTING OUTLIERS

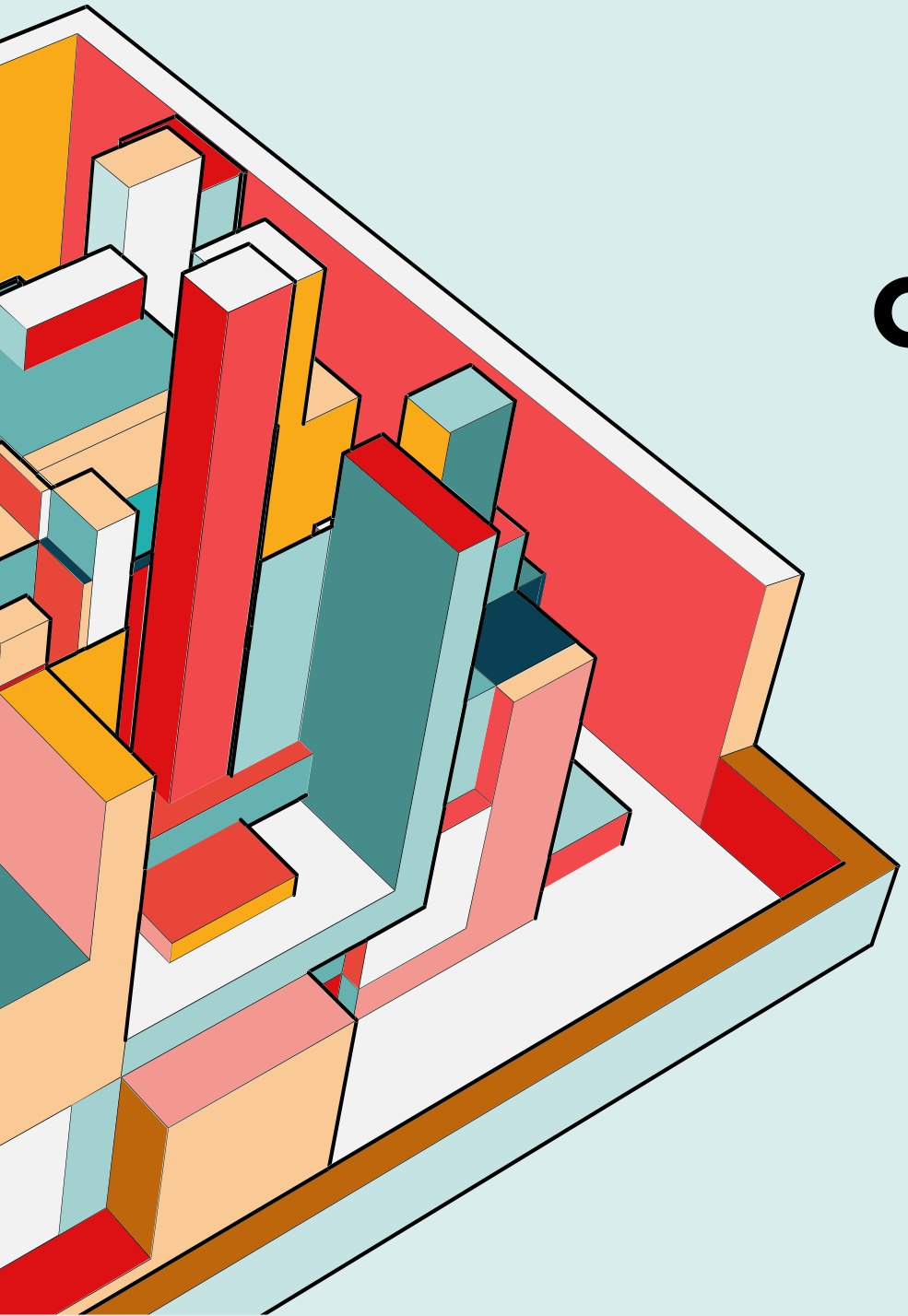




BAR PLOTS – HIGHLIGHTING OUTLIERS

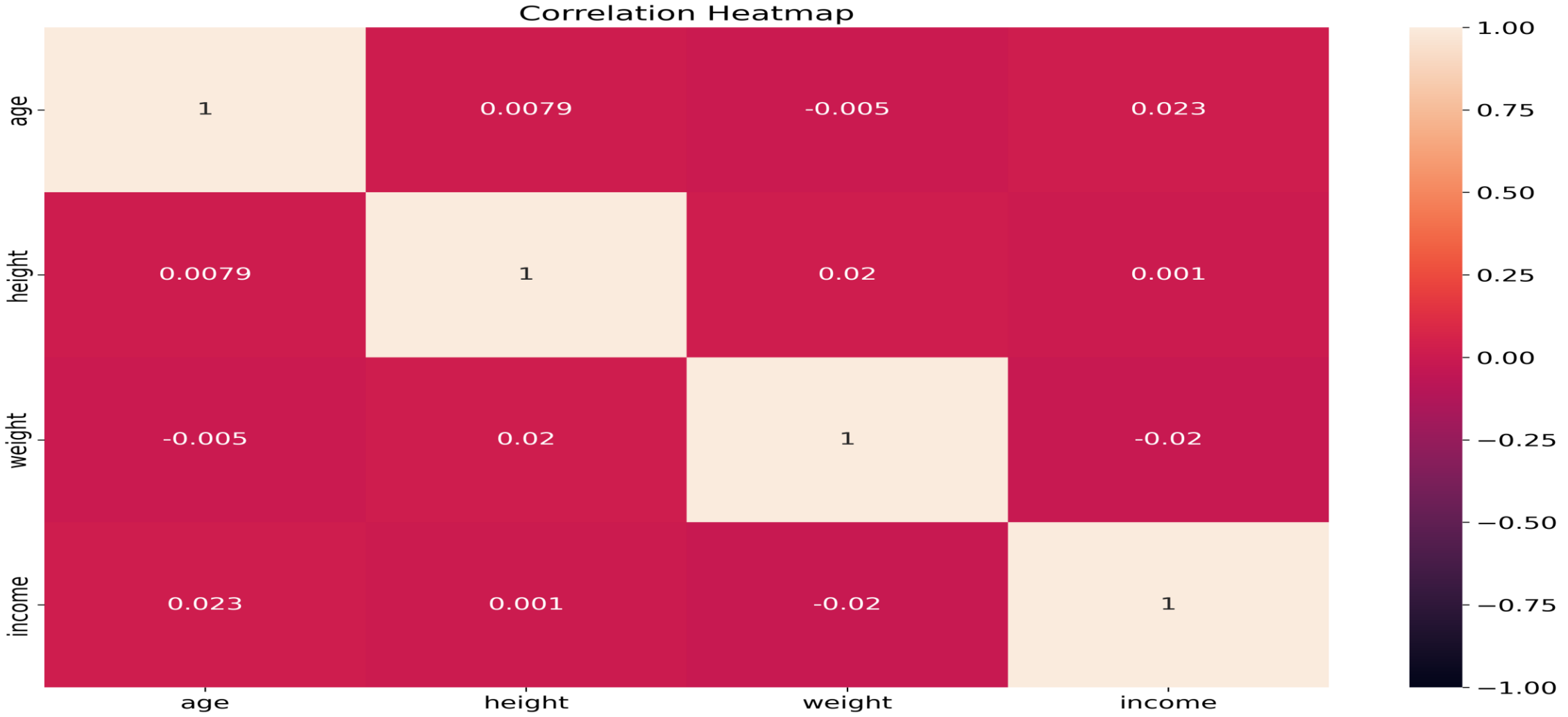


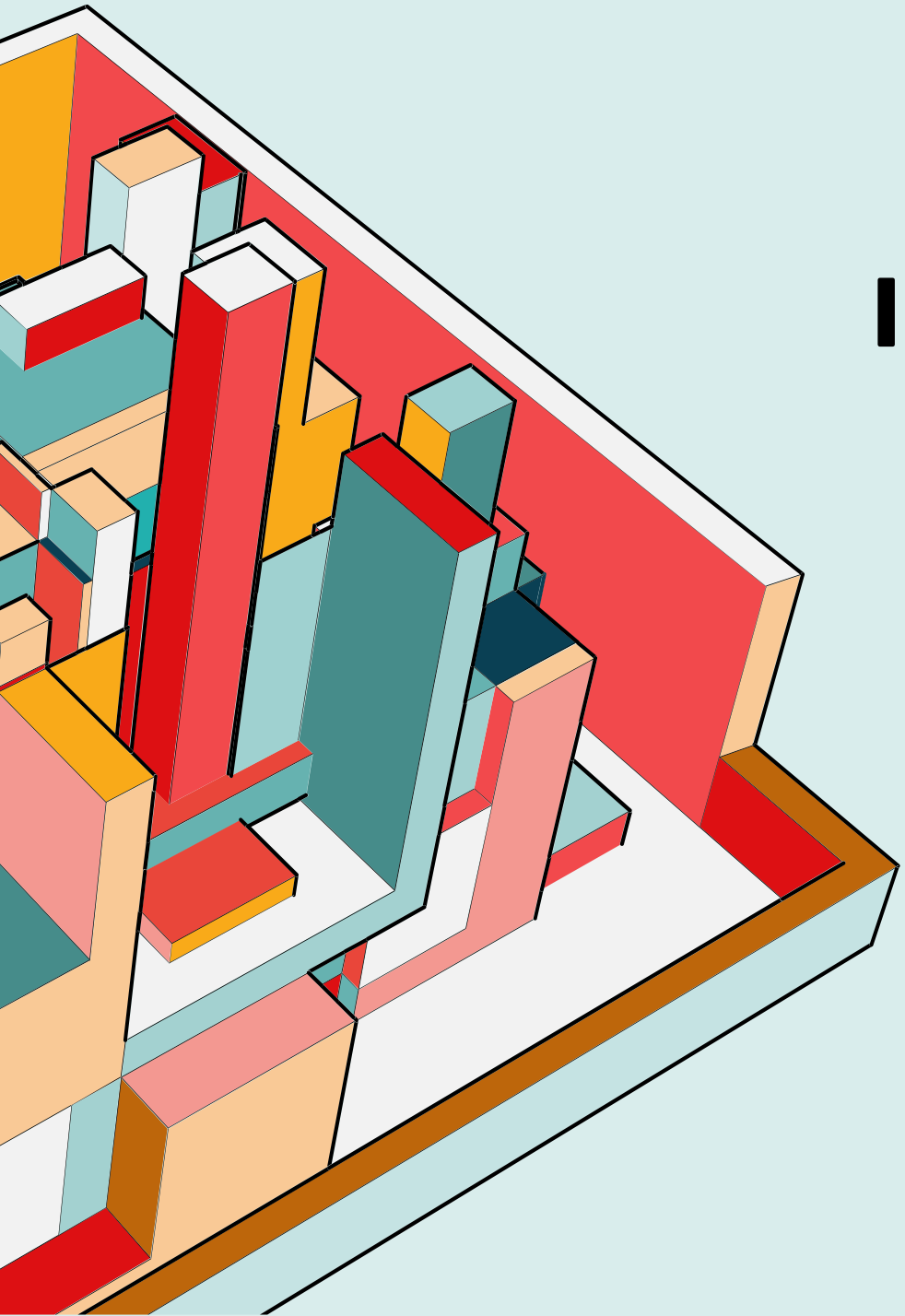




CORRELATION ANALYSIS

PEARSON-CORRELATION MATRIX



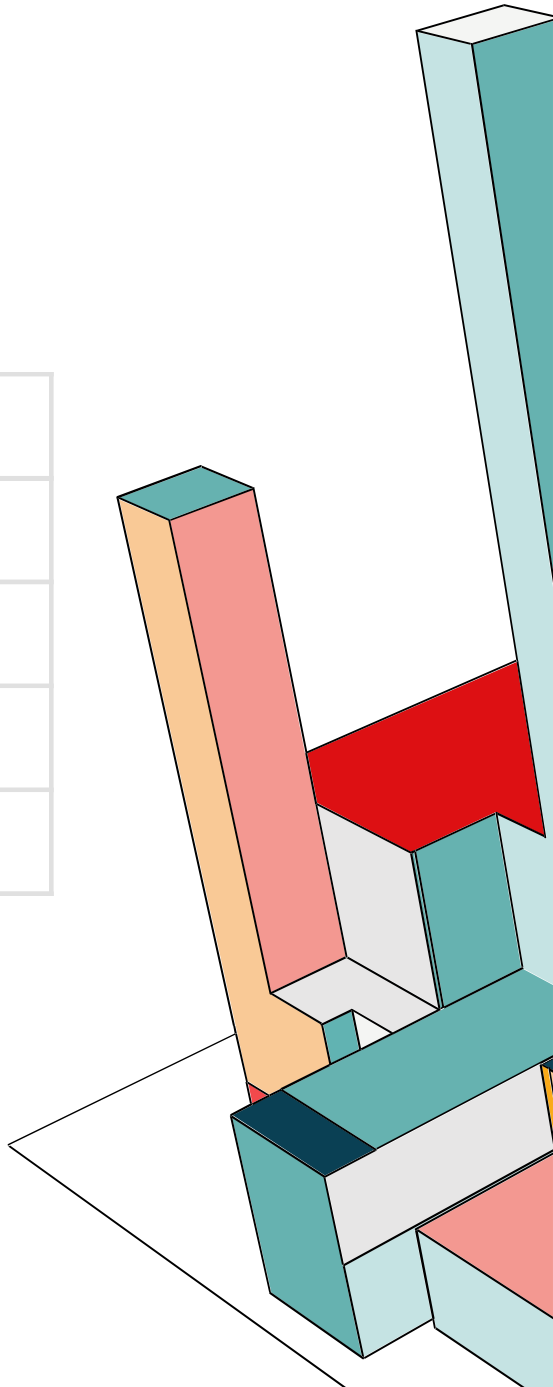


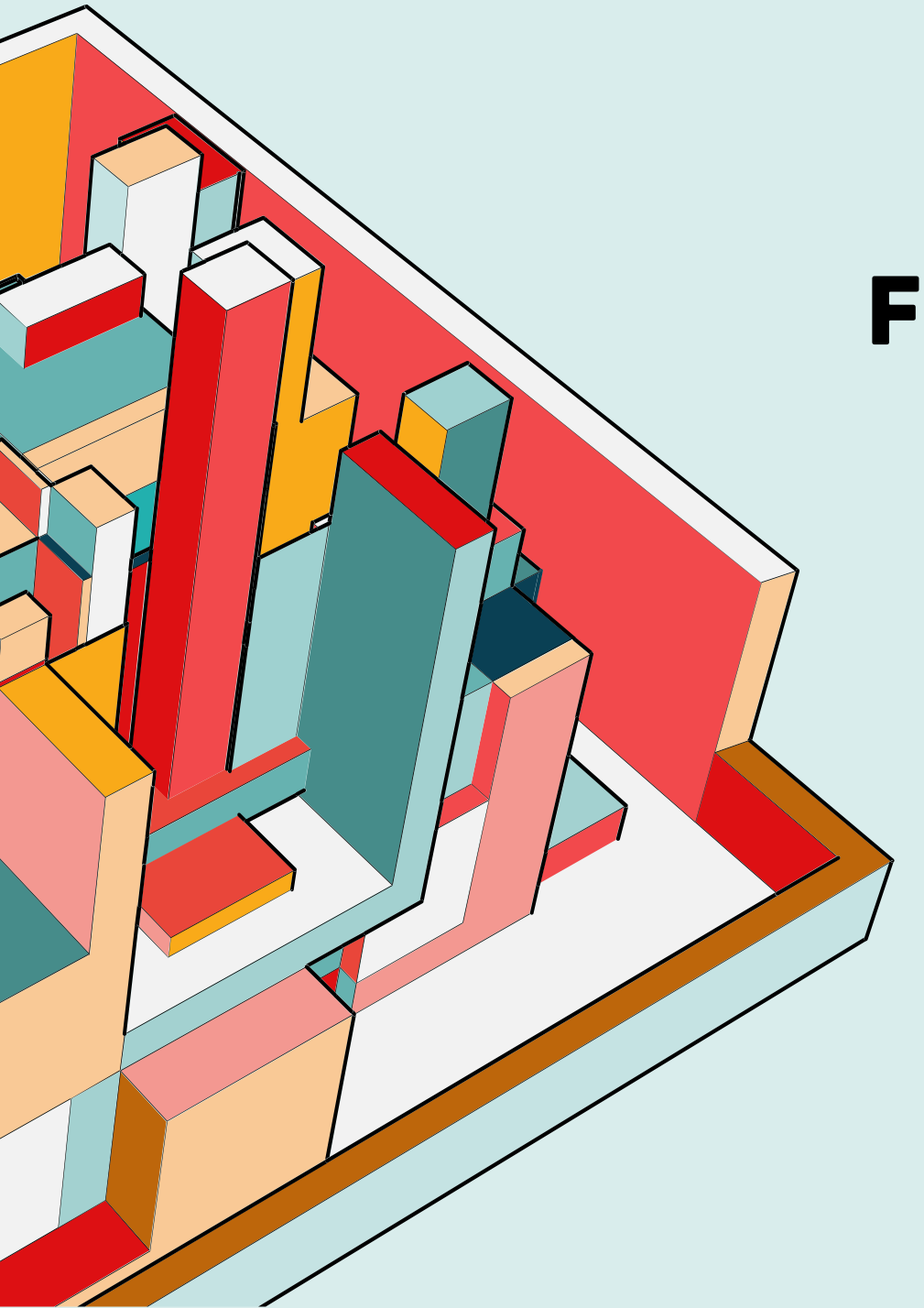
INFERENTIAL STATISTICS

INFERENTIAL STATISTICS

T-test to see if there is a significant difference in Income between Male and Female

income differences between Female vs Male group)
income: t-statistics = 0.0141, p_value = **0.9888**
income differences between Male vs Female group)
income: t-statistics = -0.0141, p_value = **0.9888**





FINAL CONSIDERATIONS

FINAL CONSIDERATIONS

Final questions	Reason	Objective	Test done?
Normal distributions for the variables reason	Having Dataset normalized	Using Numpy to generate random datasets	YES
Descriptive statistics insights	Verify Datasets Requirements	Verify if computed values, for mean, variance, ecc. are correct	YES
Boxplots interpretation, about outliers	N. of outliers very small	Result Expected within gaussain distributions	YES
Correlation Matrix	correlations between data columns	No correlation found	YES
T-test results	The p-value shows the likelihood of your data occurring under the null hypothesis	P-value almost equal to 1, i.e. no income differenceses	YES

THANK YOU

Carlo Bloch