# Assignment 1.1

Name: Parker Christenson

Date: 09-07-2023

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

## Problem # 1.1.

In the 2018 election for Senate in California, a CNN exit poll of 1882 voters stated that 52.5% voted for the Democratic candidate, Diane Feinstein. Of all 11.1 million voters, 54.2% voted for Feinstein.

(a) What was the (i) subject, (ii) sample, (iii) population?

1. subject: all of the adults in the United States
2. sample: 1141 sampled Adults
3. population: 250 million adults in the United States

## Problem # 1.2.

The `Students` data file at http://stat4ds.rwth-aachen.de/data/Students.dat responses of a class of 60 social science graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), cogpa = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* =

belief in life after death (1 = yes, 2 = no, 3 = undecided). You will use this data file for some exercises in this book.

(a) Practice accessing a data file for statistical analysis with your software by going to the book's website and copying and then displaying this data file.

`(a) Your answer goes here`

In [ ]:
```
import pandas as pd

# Brining in the data from the csv file

df = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_501\Assignments\Assignment_1_Sur

df
```

Out[ ]:

| | subject | gender | age | hsgpa | cogpa | dhome | dres | tv | sport | news | aids | veg | affil | ideol | re |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 32 | 2.2 | 3.5 | 0 | 5.00 | 3.0 | 5 | 0 | 0 | 0 | 2 | 6 | |
| **1** | 2 | 1 | 23 | 2.1 | 3.5 | 1200 | 0.30 | 15.0 | 7 | 5 | 6 | 1 | 1 | 2 | |
| **2** | 3 | 1 | 27 | 3.3 | 3.0 | 1300 | 1.50 | 0.0 | 4 | 3 | 0 | 1 | 1 | 2 | |
| **3** | 4 | 1 | 35 | 3.5 | 3.2 | 1500 | 8.00 | 5.0 | 5 | 6 | 3 | 0 | 3 | 4 | |
| **4** | 5 | 0 | 23 | 3.1 | 3.5 | 1600 | 10.00 | 6.0 | 6 | 3 | 0 | 0 | 3 | 1 | |
| **5** | 6 | 0 | 39 | 3.5 | 3.5 | 350 | 3.00 | 4.0 | 5 | 7 | 0 | 1 | 1 | 2 | |
| **6** | 7 | 0 | 24 | 3.6 | 3.7 | 0 | 0.20 | 5.0 | 12 | 4 | 2 | 0 | 3 | 2 | |
| **7** | 8 | 1 | 31 | 3.0 | 3.0 | 5000 | 1.50 | 5.0 | 3 | 3 | 1 | 0 | 3 | 2 | |
| **8** | 9 | 0 | 34 | 3.0 | 3.0 | 5000 | 2.00 | 7.0 | 5 | 3 | 0 | 0 | 3 | 1 | |
| **9** | 10 | 0 | 28 | 4.0 | 3.1 | 900 | 2.00 | 1.0 | 1 | 2 | 1 | 1 | 3 | 3 | |
| **10** | 11 | 0 | 23 | 2.3 | 2.6 | 253 | 1.50 | 10.0 | 15 | 1 | 1 | 0 | 2 | 5 | |
| **11** | 12 | 1 | 27 | 3.5 | 3.6 | 190 | 3.00 | 14.0 | 3 | 7 | 0 | 0 | 1 | 2 | |
| **12** | 13 | 0 | 36 | 3.3 | 3.5 | 245 | 1.50 | 6.0 | 15 | 12 | 5 | 0 | 1 | 1 | |
| **13** | 14 | 0 | 28 | 3.2 | 3.2 | 500 | 6.00 | 3.0 | 10 | 1 | 2 | 0 | 3 | 4 | |
| **14** | 15 | 1 | 28 | 3.0 | 3.5 | 3500 | 1.00 | 4.0 | 3 | 1 | 0 | 0 | 1 | 1 | |
| **15** | 16 | 1 | 25 | 3.8 | 3.3 | 210 | 10.00 | 7.0 | 6 | 1 | 0 | 1 | 3 | 2 | |
| **16** | 17 | 1 | 41 | 4.0 | 3.0 | 1000 | 15.00 | 6.0 | 7 | 3 | 10 | 0 | 3 | 3 | |
| **17** | 18 | 0 | 50 | 3.8 | 3.8 | 0 | 3.00 | 5.0 | 9 | 6 | 10 | 0 | 1 | 2 | |
| **18** | 19 | 0 | 71 | 4.0 | 3.5 | 5000 | 3.00 | 6.0 | 12 | 2 | 2 | 0 | 3 | 2 | |
| **19** | 20 | 1 | 28 | 3.0 | 3.8 | 120 | 1.00 | 25.0 | 0 | 0 | 2 | 1 | 1 | 1 | |
| **20** | 21 | 1 | 26 | 3.7 | 3.7 | 8000 | 8.00 | 4.0 | 4 | 4 | 1 | 0 | 3 | 4 | |
| **21** | 22 | 1 | 27 | 4.0 | 3.7 | 2 | 2.50 | 4.0 | 2 | 7 | 0 | 0 | 3 | 2 | |
| **22** | 23 | 0 | 31 | 2.7 | 3.5 | 1700 | 5.00 | 7.0 | 7 | 2 | 0 | 0 | 2 | 7 | |
| **23** | 24 | 1 | 23 | 3.7 | 3.7 | 2 | 2.00 | 7.0 | 4 | 2 | 0 | 0 | 3 | 4 | |
| **24** | 25 | 0 | 23 | 3.2 | 3.8 | 450 | 4.00 | 0.0 | 7 | 7 | 3 | 0 | 3 | 1 | |
| **25** | 26 | 1 | 44 | 3.0 | 3.0 | 0 | 2.00 | 2.0 | 3 | 2 | 3 | 1 | 3 | 3 | |
| **26** | 27 | 0 | 26 | 3.7 | 3.0 | 1000 | 3.00 | 8.0 | 2 | 7 | 0 | 0 | 1 | 2 | |
| **27** | 28 | 1 | 31 | 3.7 | 3.8 | 850 | 10.00 | 10.0 | 3 | 7 | 0 | 0 | 2 | 5 | |
| **28** | 29 | 0 | 24 | 3.3 | 3.1 | 420 | 2.00 | 10.0 | 6 | 5 | 0 | 0 | 1 | 4 | |
| **29** | 30 | 1 | 26 | 3.3 | 3.3 | 1200 | 0.75 | 10.0 | 0 | 3 | 0 | 0 | 2 | 2 | |
| **30** | 31 | 0 | 26 | 3.3 | 3.5 | 1000 | 1.50 | 0.0 | 3 | 3 | 3 | 1 | 1 | 2 | |
| **31** | 32 | 1 | 32 | 3.5 | 3.9 | 150 | 12.00 | 10.0 | 2 | 0 | 0 | 0 | 1 | 2 | |
| **32** | 33 | 0 | 26 | 3.4 | 3.4 | 2000 | 1.50 | 2.0 | 7 | 14 | 0 | 0 | 1 | 2 | |

| | subject | gender | age | hsgpa | cogpa | dhome | dres | tv | sport | news | aids | veg | affil | ideol | re |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **33** | 34 | 1 | 22 | 3.2 | 2.8 | 316 | 2.00 | 10.0 | 3 | 5 | 2 | 0 | 3 | 2 | |
| **34** | 35 | 1 | 24 | 3.5 | 3.9 | 900 | 1.75 | 8.0 | 0 | 0 | 1 | 0 | 1 | 1 | |
| **35** | 36 | 0 | 24 | 3.6 | 3.3 | 250 | 2.00 | 4.0 | 6 | 3 | 1 | 0 | 2 | 5 | |
| **36** | 37 | 0 | 23 | 3.8 | 3.7 | 180 | 0.50 | 10.0 | 5 | 7 | 0 | 0 | 3 | 2 | |
| **37** | 38 | 0 | 33 | 3.4 | 3.4 | 6000 | 1.50 | 8.0 | 5 | 6 | 2 | 0 | 3 | 2 | |
| **38** | 39 | 0 | 23 | 2.8 | 3.2 | 950 | 2.00 | 37.0 | 1 | 0 5 | 0 | 0 | 2 | 5 | |
| **39** | 40 | 0 | 31 | 3.8 | 3.5 | 1100 | 0.75 | 0.5 | 3 | 5 | 2 | 0 | 2 | 6 | |
| **40** | 41 | 0 | 26 | 3.4 | 3.4 | 1300 | 1.20 | 0.0 | 8 | 2 | 0 | 0 | 3 | 2 | |
| **41** | 42 | 0 | 28 | 2.0 | 3.0 | 360 | 0.25 | 10.0 | 8 | 3 | 0 | 0 | 1 | 3 | |
| **42** | 43 | 1 | 24 | 3.8 | 3.9 | 1800 | 2.00 | 2.0 | 5 | 4 | 1 | 0 | 2 | 6 | |
| **43** | 44 | 0 | 23 | 3.0 | 3.6 | 900 | 15.00 | 12.0 | 0 | 5 | 0 | 0 | 2 | 5 | |
| **44** | 45 | 1 | 25 | 3.0 | 4.0 | 5000 | 5.00 | 1.5 | 0 | 4 | 0 | 0 | 3 | 4 | |
| **45** | 46 | 1 | 24 | 3.0 | 3.5 | 300 | 1.00 | 10.0 | 5 | 5 | 0 | 0 | 1 | 2 | |
| **46** | 47 | 1 | 27 | 3.0 | 3.8 | 2000 | 20.00 | 28.0 | 7 | 14 | 2 | 1 | 2 | 3 | |
| **47** | 48 | 0 | 24 | 3.3 | 3.8 | 630 | 1.30 | 2.0 | 3 | 5 | 0 | 0 | 2 | 7 | |
| **48** | 49 | 1 | 26 | 3.8 | 4.0 | 1200 | 1.00 | 0.0 | 4 | 3 | 1 | 0 | 1 | 2 | |
| **49** | 50 | 1 | 27 | 3.0 | 4.0 | 580 | 2.00 | 5.0 | 15 | 1 | 2 | 0 | 1 | 1 | |
| **50** | 51 | 0 | 32 | 3.0 | 3.0 | 2000 | 5.00 | 5.0 | 5 | 2 | 1 | 0 | 2 | 5 | |
| **51** | 52 | 1 | 41 | 4.0 | 4.0 | 0 | 8.00 | 8.0 | 4 | 2 | 2 | 0 | 2 | 4 | |
| **52** | 53 | 1 | 29 | 3.0 | 3.9 | 300 | 3.70 | 2.0 | 5 | 1 | 11 | 0 | 1 | 2 | |
| **53** | 54 | 1 | 50 | 3.5 | 3.8 | 6 | 6.00 | 7.0 | 3 | 7 | 0 | 0 | 1 | 2 | |
| **54** | 55 | 1 | 22 | 3.4 | 3.7 | 80 | 7.00 | 10.0 | 1 | 2 | 2 | 0 | 3 | 2 | |
| **55** | 56 | 1 | 23 | 3.6 | 3.2 | 375 | 1.50 | 5.0 | 1 | 0 5 | 0 | 0 | 2 | 6 | |
| **56** | 57 | 0 | 26 | 3.5 | 3.6 | 2000 | 0.30 | 16.0 | 8 | 3 | 0 | 0 | 1 | 4 | |
| **57** | 58 | 0 | 30 | 3.0 | 3.0 | 1 | 1.10 | 1.0 | 4 | 3 | 0 | 0 | 3 | 3 | |
| **58** | 59 | 1 | 23 | 3.0 | 3.0 | 112 | 0.50 | 15.0 | 3 | 3 | 0 | 0 | 3 | 4 | |
| **59** | 60 | 1 | 22 | 3.4 | 3.0 | 650 | 4.00 | 8.0 | 16 | 7 | 1 | 0 | 3 | 4 | |

(b) Using responses on *abor*, state a question that could be addressed with (i) descriptive statistics, (ii) inferential statistics.

# Answer

## (i) descriptive statistics:

What is the percentage of students who think abortion should be legal in the first three months of pregnancy?

### (ii) inferential statistics:

What is the percentage of students who think abortion should be legal in the first three months of pregnancy in the population of all students who have went to college?

# Problem # 1.3.

Identify each of the following variables as categorical or quantitative: (a) Number of smartphones that you own; (b) County of residence; (c) Choice of diet (vegetarian, nonvegetarian); (d) Distance, in kilometers, commute to work

## Answer

### Number of smartphones that you own: quantitative

### County of residence: categorical

### Choice of diet (vegetarian, nonvegetarian): categorical

### Distance, in kilometers, commute to work: quantitative

# Problem # 1.4.

Give an example of a variable that is (a) categorical; (b) quantitative; (c) discrete; (d) continuous

## Answer

### (A) Categorical: Gender

### (B) Quantitative: Age

### (C) Discrete: HSGPA, and COGPA

### (D) Continuous: Dhome, and Dres, Sport

# Problem # 1.10.

Analyze the `Carbon_West` (http://stat4ds.rwth-aachen.de/data/Carbon_West.dat) data file at the book's website by **(a)** constructing a frequency distribution and a histogram, **(b)** finding the mean, median, and standard deviation. Interpret each.

`(a)`

In [ ]:

```
# Reading the data in from a CSV file

co2_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_501\Assignments\CO2_Emissi

co2_data
```

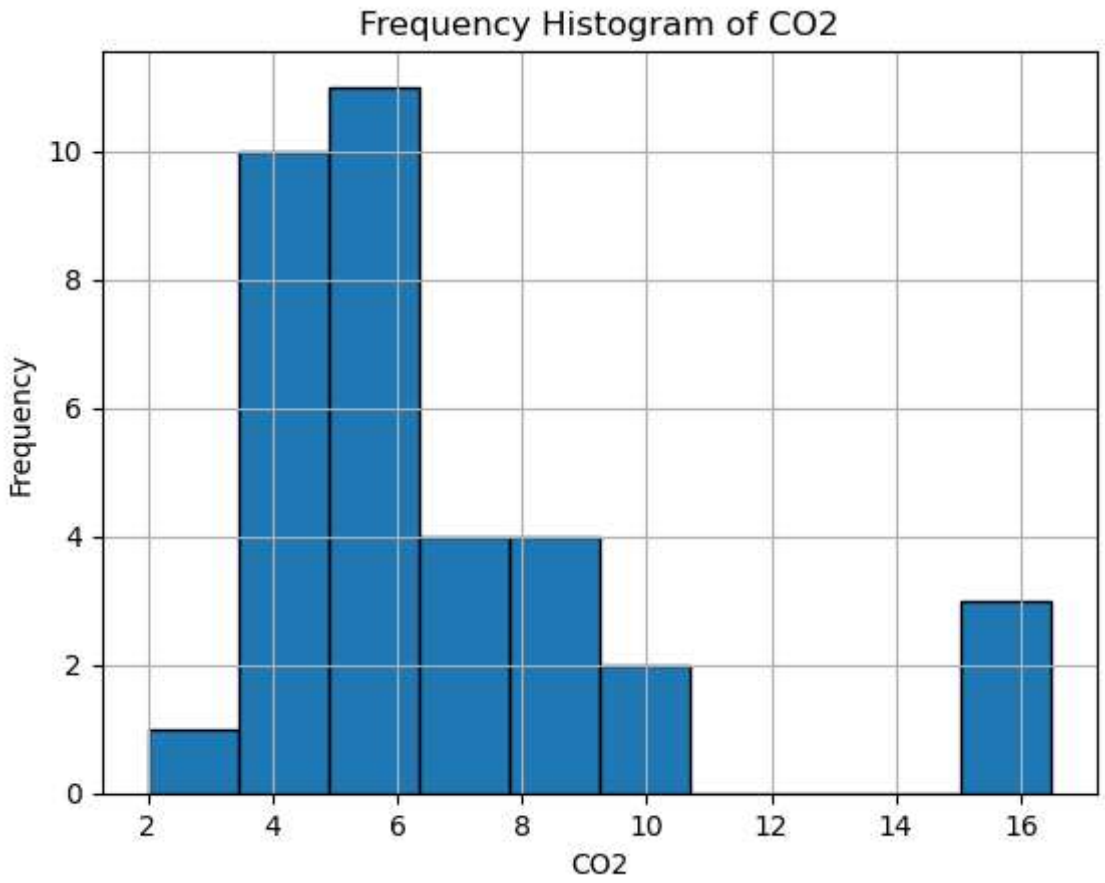
Out[ ]:

| Nation | CO2 |
|---|---|
| Albania | 2.0 |
| Australia | 15.4 |
| Austria | 6.9 |
| Belgium | 8.3 |
| Bosnia | 6.2 |
| Bulgaria | 5.9 |
| Canada | 15.1 |
| Croatia | 4.0 |
| Cyprus | 5.3 |
| Czech | 9.2 |
| Denmark | 5.9 |
| Finland | 8.7 |
| France | 4.6 |
| Germany | 8.9 |
| Greece | 6.2 |
| Hungary | 4.3 |
| Ireland | 7.3 |
| Italy | 5.3 |
| Latvia | 3.5 |
| Lithuania | 4.4 |
| Malta | 5.4 |
| Montenegro | 3.6 |
| Netherlands | 9.9 |
| NewZealand | 7.7 |
| Norway | 9.3 |
| Portugal | 4.3 |
| Romania | 3.5 |
| Serbia | 5.3 |
| Slovak | 5.7 |
| Slovenia | 6.2 |
| Spain | 5.0 |
| Sweden | 4.5 |

| Nation | CO2 |
|---|---|
| Switzerland | 4.3 |
| UK | 6.5 |
| US | 16.5 |

## Frequency Distribution and Histogram

In [ ]:
```
import matplotlib.pyplot as plt

column_name = 'CO2'
data_to_plot = co2_data[column_name]

# Create a frequency histogram
plt.hist(data_to_plot, bins=10, edgecolor='k')  # Adjust the number of bins as needed
plt.xlabel(column_name)
plt.ylabel('Frequency')
plt.title(f'Frequency Histogram of {column_name}')
plt.grid(True)
plt.show()
```




## (b) Mean, Median, and Standard Deviation

In [ ]:
```
Mean = co2_data['CO2'].mean()
Median = co2_data['CO2'].median()
Mode = co2_data['CO2'].mode()
```

```
Standard_Deviation = co2_data['CO2'].std()

print(Mean)
print(Median)
print(Mode)
print(Standard_Deviation)
```

```
6.717142857142858
5.9
0    4.3
1    5.3
2    6.2
Name: CO2, dtype: float64
3.3569493954627134
```

# Problem # 1.11.

According to Statistics Canada, for the Canadian population having income in 2019, annual income had a median of $35,000 and mean of $46,700. What would you predict about the shape of the distribution? Why?

## Answer

The data is right-skewed because the mean is greater than the median.

The data is not symmetric because the mean is not equal to the median.

There very well could be potential outliers. We can't see from the full data set, and we don't get ranges or quartiles to work with.

# Problem # 1.13.

A report indicates that public school teacher's annual salaries in New York city have an approximate mean of $69,000 and standard deviation of $6,000. If the distribution has approximately a bell shape, report intervals that contain about (a) 68%, (b) 95%, (c) all or nearly all salaries. Would a salary of $100,000 be unusual? Why?

## Answer

Using the empirical rule, we can say that 68% of the data is within 1 standard deviation of the mean, 95% of the data is within 2 standard deviations of the mean, and 99.7% of the data is within 3 standard deviations of the mean.

With that being said, it is unusual for a teacher to make $100,000. It is within 3 standard deviations of the mean, and 95% of the data is within two. With that being said, this would be considered to be a high salary for a teacher.

# Problem # 1.17.

From the `Murder` data file (http://stat4ds.rwth-aachen.de/data/Murder.dat) at the book's website, use the variable murder, which is the murder rate (per 100,000 population) for each state in the U.S. in 2017 according to the FBI Uniform Crime Reports. At first, do not use the observation for D.C. (DC). Using software:
(a) Find the mean and standard deviation and interpret their values.
(b) Find the five-number summary, and construct the corresponding box plot. Interpret.
(c) Now include the observation for D.C. What is affected more by this outlier: The mean or the median? The range or the inter-quartile range?

**Answer:**

`(a) Your answer goes here`

In [ ]:
```
murder_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_501\Assignments\Murder_

mstd = murder_data['murder'].std()
mstd

### With the standard Deviation being 3.73 the data points are close to the Mean of th
### Considering the deaths per 100,000 that is a large standard diviation.
```

Out[ ]:
```
3.7253914125190377
```

`(b) Your answer goes here`

In [ ]:
```
# 5 number summary
import numpy as np

murder_column = murder_data['murder']

q1 = np.percentile(murder_column, 25)
median = np.percentile(murder_column, 50)
q3 = np.percentile(murder_column, 75)

minimum = murder_column.min()
maximum = murder_column.max()

print("Minimum:", minimum)
print("1st Quartile (Q1):", q1)
print("Median (2nd Quartile, Q2):", median)
print("3rd Quartile (Q3):", q3)
print("Maximum:", maximum)
```

```
Minimum: 1.0
1st Quartile (Q1): 2.6500000000000004
Median (2nd Quartile, Q2): 5.0
3rd Quartile (Q3): 6.45
Maximum: 24.2
```

`(c) Your answer goes here`

**The Mean is effected more, because the mean was pulled to the higher side, by the outlier being so high. And As for the Quartiles, the Range is more effcted because it directly depends on the max and the minimum.**

# Problem # 1.18.

The `Income` data file (http://stat4ds.rwth-aachen.de/data/Income.dat) at the book's website reports annual income values in the U.S., in thousands of dollars.

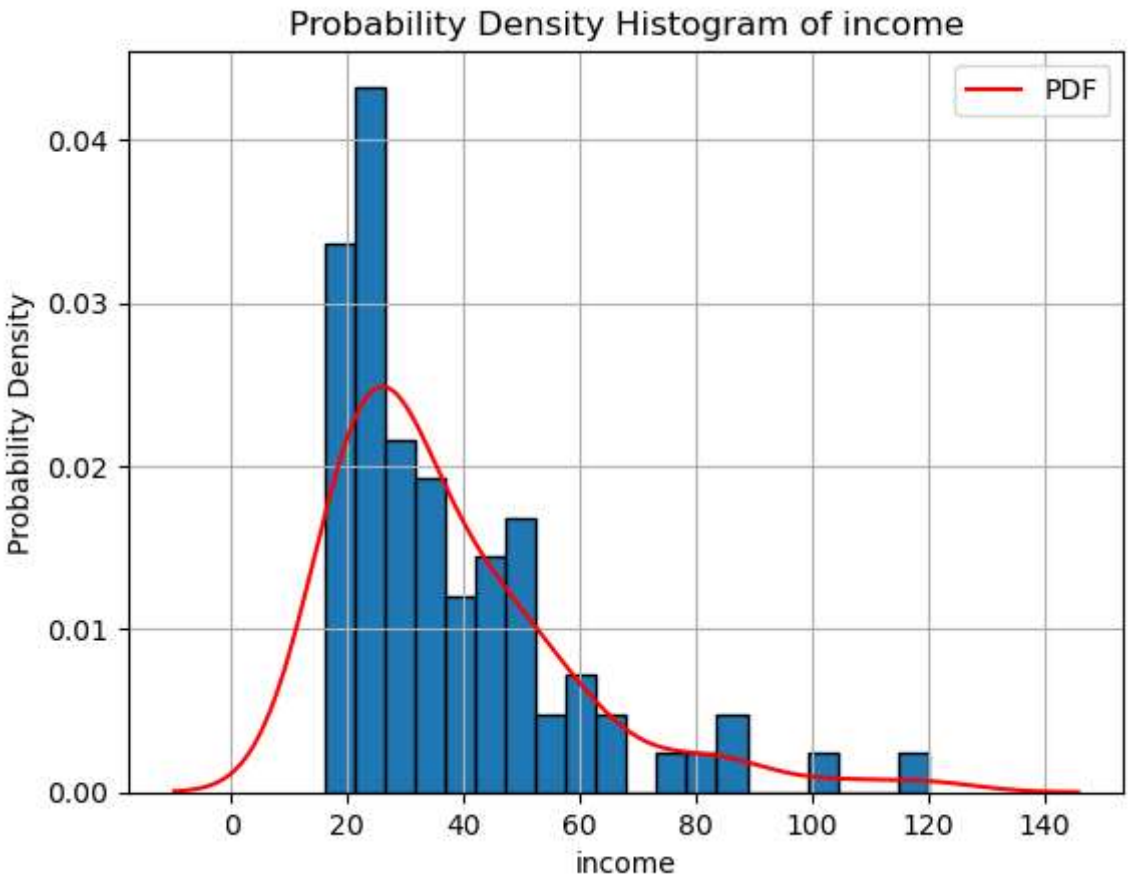
(a) Using software, construct a histogram. Describe its shape.
(b) Find descriptive statistics to summarize the data. Interpret them.
(c) The kernel density estimation method finds a smooth-curve approximation for a histogram. At each value, it takes into account how many observations are nearby and their distance, with more weight given those closer. Increasing the bandwidth increases the influence of observations further away. Plot a smooth-curve approximation for the histogram of income values. Summarize the impact of increasing and of decreasing the bandwidth substantially from the default value.
(d) Construct and interpret side-by-side box plots of income by race (B = Black, H = Hispanic, W = White). Compare the incomes using numerical descriptive statistics

**Answer:**

`(a) Your answer goes here`

In [ ]:
```
import seaborn as sns  # Import seaborn for plotting the PDF

income_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_501\Assignments\Income_
print(income_data.columns)
column_name = 'income'
income_to_plot = income_data[column_name]
plt.hist(income_to_plot, bins=20, edgecolor='k', density=True)  # Use density=True to
plt.xlabel(column_name)
plt.ylabel('Probability Density')
plt.title(f'Probability Density Histogram of {column_name}')
sns.kdeplot(income_to_plot, color='r', label='PDF')
plt.grid(True)
plt.legend()
plt.show()
```

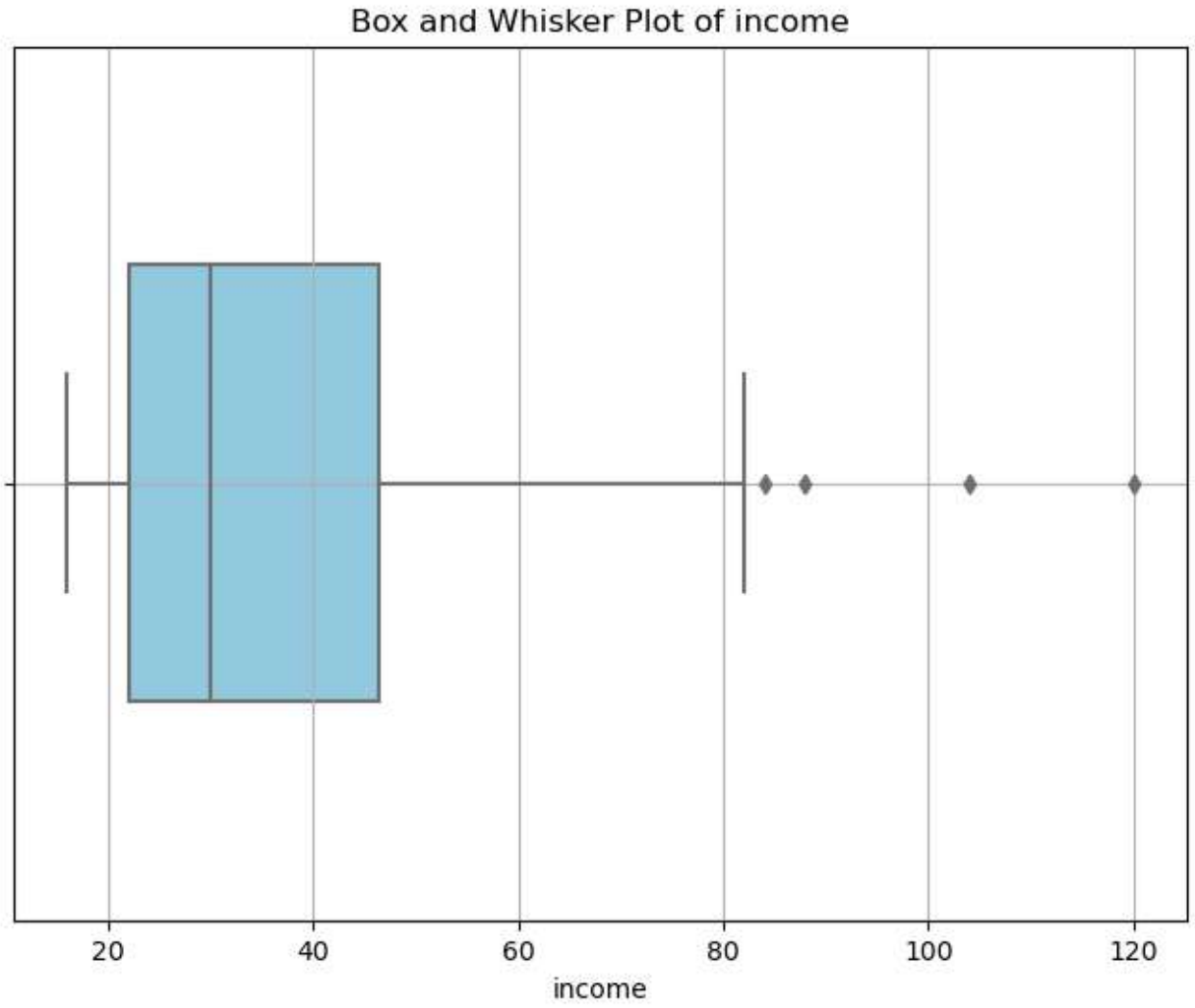
```
Index(['income', 'race'], dtype='object')
```




The shape of the histogram is right skewed.

(b) Your answer goes here

In [ ]:
```
column_name = 'income'
income_column = income_data[column_name]

#Mean, Median, Mode
mean_income = income_column.mean()
median_income = income_column.median()
mode_income = income_column.mode().values[0]

# Range
min_income = income_column.min()
max_income = income_column.max()
income_range = max_income - min_income

# Create a box and whisker plot
plt.figure(figsize=(8, 6))
sns.boxplot(x=income_column, width=0.5, color='skyblue')
plt.xlabel(column_name)
plt.title(f'Box and Whisker Plot of {column_name}')
plt.grid(True)
plt.show()

# Printing the mean, median, mode, and range of the data
print(f"Mean {column_name}: {mean_income:.2f}")
print(f"Median {column_name}: {median_income:.2f}")
print(f"Mode {column_name}: {mode_income:.2f}")
print(f"Range of {column_name}: {income_range:.2f}")
```



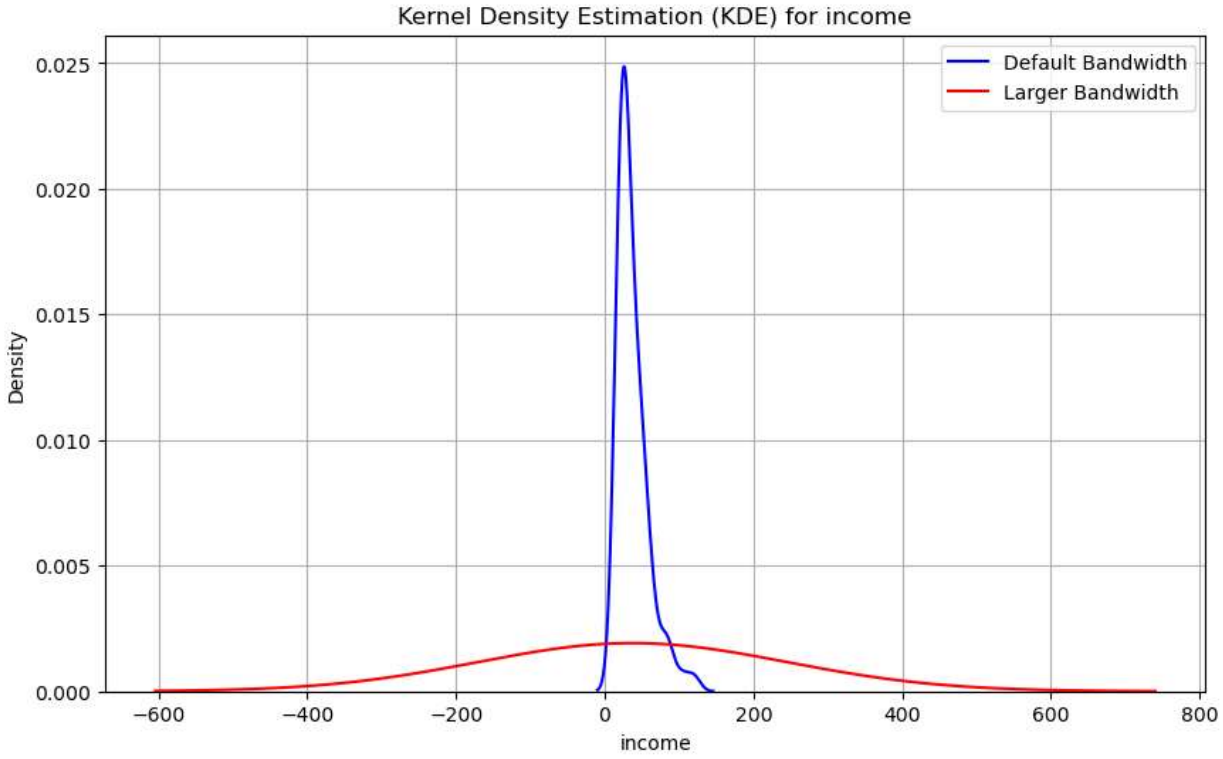

```
Mean income: 37.52
Median income: 30.00
Mode income: 30.00
Range of income: 104.00
```

(c) Your answer goes here

In [ ]:

```python
# Specify the column name for analysis
column_name = 'income'
income_column = income_data[column_name]

# Create a KDE plot with the default bandwidth
plt.figure(figsize=(10, 6))
sns.kdeplot(income_column, color='b', label='Default Bandwidth')

# Create a KDE plot with a larger bandwidth
sns.kdeplot(income_column, color='r', label='Larger Bandwidth', bw=10)  # Adjust the

plt.xlabel(column_name)
plt.ylabel('Density')
plt.title(f'Kernel Density Estimation (KDE) for {column_name}')
plt.legend()
plt.grid(True)

plt.show()
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_18780\4265395061.py:10: UserWarning:

The `bw` parameter is deprecated in favor of `bw_method` and `bw_adjust`.
Setting `bw_method=10`, but please see the docs for the new parameters
and update your code. This will become an error in seaborn v0.13.0.

  sns.kdeplot(income_column, color='r', label='Larger Bandwidth', bw=10)  # Adjust th
e 'bw' parameter as needed
```

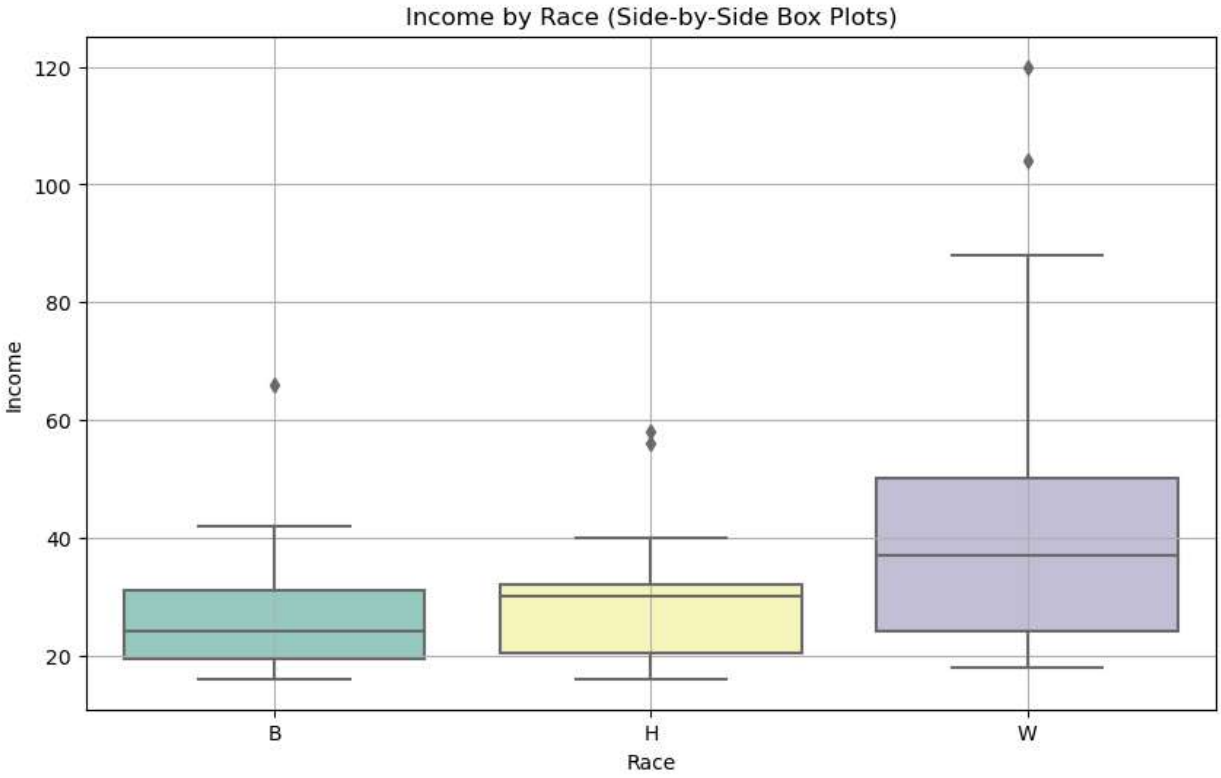



**Increasing the bandwidth increases the influence of observations further away. Decreasing the bandwidth decreases the influence of observations further away. When the bandwidth is increased, the curve becomes more smooth. When the bandwidth is decreased, the curve becomes less smooth.**

(d) Your answer goes here

In [ ]:

```python
# Specify the columns for analysis
columns_to_analyze = ['income', 'race']
selected_data = income_data[columns_to_analyze]

# Create side-by-side box plots
plt.figure(figsize=(10, 6))
sns.boxplot(x='race', y='income', data=selected_data, palette='Set3')
plt.xlabel('Race')
plt.ylabel('Income')
plt.title('Income by Race (Side-by-Side Box Plots)')
plt.grid(True)
plt.show()

race_statistics = selected_data.groupby('race')['income'].describe()
print(race_statistics)
```




| race | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| B | 16.0 | 27.75 | 13.284076 | 16.0 | 19.5 | 24.0 | 31.0 | 66.0 |
| H | 14.0 | 31.00 | 12.812254 | 16.0 | 20.5 | 30.0 | 32.0 | 58.0 |
| W | 50.0 | 42.48 | 22.869854 | 18.0 | 24.0 | 37.0 | 50.0 | 120.0 |

## Description of the data

When looking at the mean, median of the meadian, we see that whites do make more than, blacks or hispanics. We are also able to see that whites have a much larger band, than hispanics and blacks.

# Problem # 1.19.

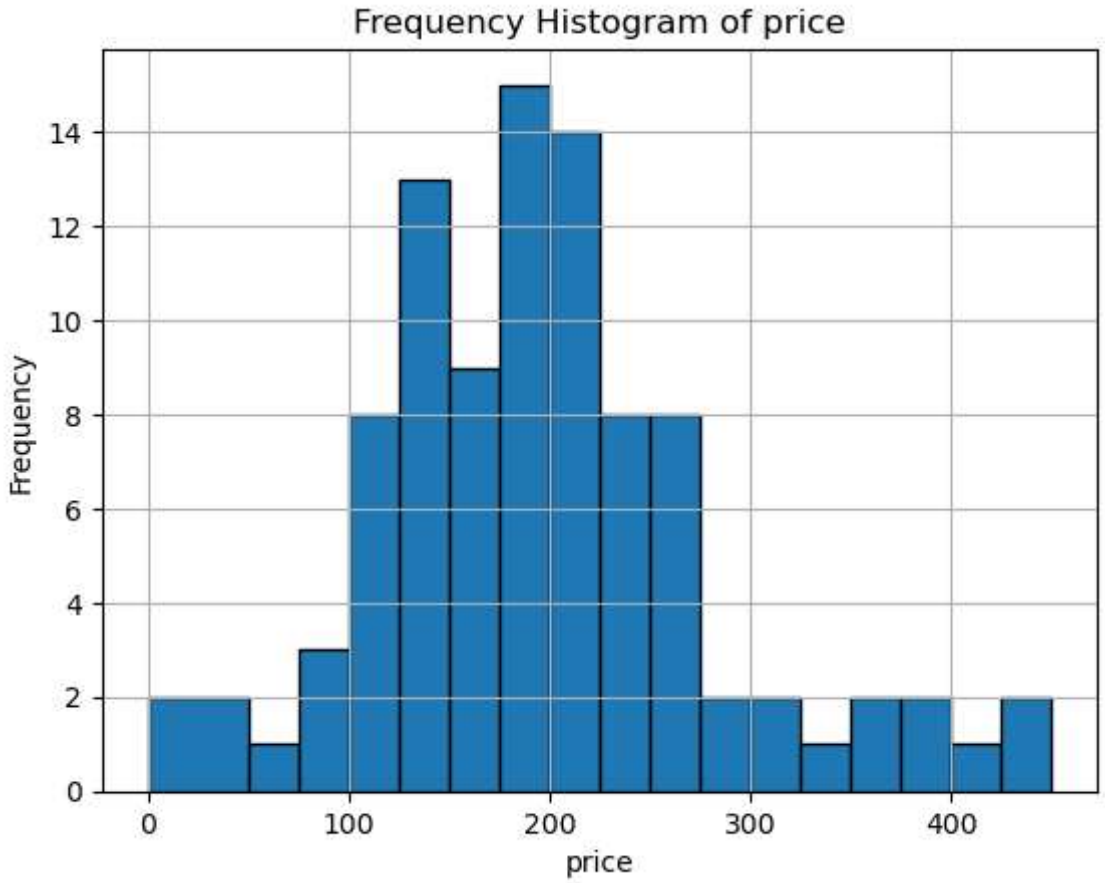
The `Houses` data file (http://stat4ds.rwth-aachen.de/data/Houses.dat) at the book's website lists the selling price (thousands of dollars), size (square feet), tax bill (dollars), number of bathrooms, number of bedrooms, and whether the house is new (1 = yes, 0 = no) for 100 home sales in Gainesville, Florida. Let's analyze the selling prices.

(a) Construct a frequency distribution and a histogram. Describe the shape.
(b) Find the percentage of observations that fall within one standard deviation of the mean. Why is this not close to 68%?
(c) Construct a box plot, and interpret.
(d) Use descriptive statistics to compare selling prices according to whether the house is new.

**Answer:**

`(a) Your answer goes here`

In [ ]:
```
housing_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_501\Assignments\New_Us

# Specify the column name for analysis

column_name = 'price'
data_to_plot = housing_data[column_name]

# Create a frequency histogram with price and new bins

plt.hist(data_to_plot, bins=[0, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 3
plt.xlabel(column_name)
plt.ylabel('Frequency')
plt.title(f'Frequency Histogram of {column_name}')
plt.grid(True)
plt.show()
```




**The Data shows a bell shape to the overall amount of home sales. With that being said, this is just for all home prices, and not just if they are new or not.**

(b) Your answer goes here

In [ ]:
```
mean_price = housing_data[column_name].mean()
std_price = housing_data[column_name].std()
lower_bound = mean_price - std_price
upper_bound = mean_price + std_price
within_std_dev = housing_data[(housing_data[column_name] >= lower_bound) & (housing_da
percentage_within_std_dev = (len(within_std_dev) / len(housing_data)) * 100
```

```
print(f"Percentage of observations within one standard deviation of the mean for '{col
```

```
Percentage of observations within one standard deviation of the mean for 'price': 85.
00%
```

**The percentage of observations that fall within one standard deviation of the mean is 68%. This is not close to 68% because the data is not normally distributed. The data is more foucused in the center of the data, and not spread out evenly. There is also not a lot of data skewing the data to the right or the left. The data is also somewhat tight around the mean of the data.**

(c) Your answer goes here

In [ ]:
```
price_column = 'price'
new_column = 'new'

# Spliting the data up
not_new_group = housing_data[housing_data[new_column] == 0]
new_group = housing_data[housing_data[new_column] == 1]

# Create a box and whisker plot by grouping type
plt.figure(figsize=(10, 6))
plt.boxplot([not_new_group[price_column], new_group[price_column]], labels=['Not New',
plt.xlabel('Condition')
plt.ylabel('Price')
plt.title('Box and Whisker Plot of Price by Condition')
plt.grid(True)

# Show the plot
plt.show()

# Print the statistics for each group
not_new_stats = not_new_group[price_column].describe()
new_stats = new_group[price_column].describe()

print("Statistics for Not New:")
print(not_new_stats)
print("\nStatistics for New:")
print(new_stats)
```


```
Statistics for Not New:
count     89.000000
mean     207.851124
std      121.039149
min       31.500000
25%      135.000000
50%      190.800000
75%      240.000000
max      880.500000
Name: price, dtype: float64

Statistics for New:
count     11.000000
mean     436.445455
std      219.832789
min      158.850000
25%      256.950000
50%      427.500000
75%      519.675000
max      866.250000
Name: price, dtype: float64
```

`(d) Your answer goes here`

## Not new Home Descriptive Statistics

Mean: The mean price for 'Not New' housing units is approximately $207,851.12. This represents the average price of the housing units in this group.

Standard Deviation: The standard deviation of approximately $121,039.15 measures the degree of variation or dispersion in prices within the 'Not New' group. A higher standard deviation indicates greater variability in prices.

Minimum: The minimum price observed in this group is $31,500. This is the lowest price among the 'Not New' housing units.

25th Percentile: The 25th percentile price is approximately $135,000. This means that 25% of the 'Not New' housing units have prices at or below this value.

Median: The median price is approximately $190,800. This is the middle value in the dataset when the prices are ordered from lowest to highest. It divides the dataset into two equal halves.

75th Percentile: The 75th percentile price is approximately $240,000. This means that 75% of the 'Not New' housing units have prices at or below this value.

Maximum: The maximum price observed in this group is $880,500. This is the highest price among the 'Not New' housing units.

## New Home Descriptive Statistics

Mean: The mean price for 'New' housing units is approximately $436,445.45. This represents the average price of the housing units in this group.

Standard Deviation: The standard deviation of approximately $219,830 measures the degree of variation or dispersion in prices within the 'New' group. A higher standard deviation indicates greater variability in prices.

Minimum: The minimum price observed in this group is $158.85. This is the lowest price among the 'New' housing units.

25th Percentile: The 25th percentile price is approximately $256,950. This means that 25% of the 'New' housing units have prices at or below this value.

Median: The median price is approximately $427,500. This is the middle value in the dataset when the prices are ordered from lowest to highest. It divides the dataset into two equal halves.

75th Percentile: The 75th percentile price is approximately $519,680. This means that 75% of the 'New' housing units have prices at or below this value.

Maximum: The maximum price observed in this group is $866,250. This is the highest price among the 'New' housing units.