Question 1

Question: Based on the data for variable total_revenue from the attribute cost. Then calculate for the data together and [inclusive/exclusive] of the 1st and 3rd quartile.

[0, 12, 13, 15, 22, 23, 25, 27, 28, 29, 30, 31, 31, 31, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 50]

a) Calculate the mean and median of the data. Round your answer to two decimal places.

b) Calculate the interquartile range (IQR) of the data. Round your answer to two decimal places.

c) Given the information, describe the skewness of the data. Is the distribution of data skewed to the left or right?

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    import pandas as pd

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   age  sex  bmi  children  smoker  region  charges
0   19    0 27.9         0       1       3  16884.9
1   18    1 33.8         1       0       2   1725.5
2   28    1 33.0         3       0       2   4449.5
3   33    1 22.7         0       0       1  21984.5
4   32    1 28.9         0       0       1   3866.8

Last 5 rows of the dataset:
      age  sex   bmi  children  smoker  region  charges
1333   18    0  31.9         0       0       4   2205.9
1334   18    1  36.9         0       0       2   1629.8
1335   18    1  25.8         0       0       3   2007.9
1336   21    1  26.5         1       0       1   2351.5
1337   61    1  29.1         1       1       4  29141.4

Mean of total_revenue: 13270.422265141257
Median of total_revenue: 9382.033
```

- Question 2

Use the given dataset and perform the operations listed below. Indicate transcription attributes in the dataset:

- Labels: The column that classification
- Average Price: the average price of a single avocado
- Type: conventional or organic
- Year: the year
- Region: the city or region of the observation
- Total Volume: Total number of avocados sold
- 4046: Total number of avocados with PLU# 4046 sold
- 4225: Total number of avocados with PLU# 4225 sold
- 4770: Total number of avocados with PLU# 4770 sold

Introduce a package (pandas, sklearn, etc)?

a. Sort the attribute "Total Volume" in the given dataset and distribute the data into equal-sized frequency bins. Let the number of bins be 200. Identify the values for:

- 25th percentile
- 50 th percentile
- 75 th percentile

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    import pandas as pd

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
0           0      64236.62  1036.74 54454.85  48.16     8696.87     8603.62       93.25         0.00  0     2015  0
1           1      54876.98   674.28 44638.81  58.33     9505.56     9408.07       97.49         0.00  0     2015  0
2           2      118220.22 794.70 109149.67 130.50    8445.35     8042.21      303.14       100.00  0     2015  0
3           3      78992.15  1132.00 71976.41 72.58     5811.16     5677.40      133.76         0.00  0     2015  0
4           4      51039.60  941.48 43838.39  75.78     6183.95     5986.26      197.69         0.00  0     2015  0

Last 5 rows of the dataset:
      Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
18244       7839      149933.10 1351.20 56267.16 159.90    92154.84    29421.44    62733.40         0.00  1     2018  0
18245       7840      191693.90 1655.48 57790.11 1500.13   130747.18   103035.54    27711.64         0.00  1     2018  0
18246       7841      160774.59 2452.79 43325.51 1212.56   113283.73    96358.07    16925.66         0.00  1     2018  0
18247       7842      160529.05 2765.96 75083.67 1083.84   81596.58    80790.77      805.81         0.00  1     2018  0
18248       7843      122357.74 1486.26 43271.92 1364.13   76234.43    75304.31      930.12         0.00  1     2018  0

Mean of total_revenue: 850644.1301318255
Median of total_revenue: 107376.75
```

```
[ ] # Sort the 'Total Volume' column
    df['Total Volume'] = df['Total Volume'].sort_values()

    # Calculate the number of bins
    num_bins = 200

    # Calculate the 25th, 50th, and 75th percentiles
    percentiles = df['Total Volume'].quantile([0.25, 0.50, 0.75])

    # Display the results
    print("25th percentile:", percentiles[0.25])
    print("50th percentile:", percentiles[0.50])
    print("75th percentile:", percentiles[0.75])
```

```
25th percentile: 10838.99
50th percentile: 107376.75
75th percentile: 432962.29
```

b. Calculate the interquartile range (IQR) of the data. Round your answer to two decimal places.

```
[ ] # Calculate the interquartile range (IQR)
    iqr = percentiles[0.75] - percentiles[0.25]

    # Round the IQR to two decimal places
    iqr = round(iqr, 2)

    # Display the IQR
    print("Interquartile range (IQR):", iqr)
```

```
Interquartile range (IQR): 422123.3
```

c. Given the information, describe the skewness of the data. Is the distribution of data skewed to the left or right?

```
[ ] # Calculate the skewness of the data
    skewness = df['Total Volume'].skew()

    # Display the skewness
    print("Skewness:", skewness)

    # Interpretation of skewness
    if skewness > 0:
        print("The distribution is skewed to the right.")
    elif skewness < 0:
        print("The distribution is skewed to the left.")
    else:
        print("The distribution is symmetric.")
```

```
Skewness: 1.8488333333333333
The distribution is skewed to the right.
```

Question 3

Use the given dataset and perform the operations listed below. Indicate transcription attributes in the dataset:

- Labels: The column that classification
- Average Price: the average price of a single avocado
- Type: conventional or organic
- Year: the year
- Region: the city or region of the observation
- Total Volume: Total number of avocados sold
- 4046: Total number of avocados with PLU# 4046 sold
- 4225: Total number of avocados with PLU# 4225 sold
- 4770: Total number of avocados with PLU# 4770 sold

Introduce a package (pandas, sklearn, etc)?

a. Sort the attribute "Total Volume" in the given dataset and distribute the data into equal-sized frequency bins. Let the number of bins be 200.

Identify the values for:

- 25th percentile
- 50 th percentile
- 75 th percentile

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    import pandas as pd

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
0           0      64236.62  1036.74 54454.85  48.16     8696.87     8603.62       93.25         0.00  0     2015  0
1           1      54876.98   674.28 44638.81  58.33     9505.56     9408.07       97.49         0.00  0     2015  0
2           2      118220.22 794.70 109149.67 130.50    8445.35     8042.21      303.14       100.00  0     2015  0
3           3      78992.15  1132.00 71976.41 72.58     5811.16     5677.40      133.76         0.00  0     2015  0
4           4      51039.60  941.48 43838.39  75.78     6183.95     5986.26      197.69         0.00  0     2015  0

Last 5 rows of the dataset:
      Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
18244       7839      149933.10 1351.20 56267.16 159.90    92154.84    29421.44    62733.40         0.00  1     2018  0
18245       7840      191693.90 1655.48 57790.11 1500.13   130747.18   103035.54    27711.64         0.00  1     2018  0
18246       7841      160774.59 2452.79 43325.51 1212.56   113283.73    96358.07    16925.66         0.00  1     2018  0
18247       7842      160529.05 2765.96 75083.67 1083.84   81596.58    80790.77      805.81         0.00  1     2018  0
18248       7843      122357.74 1486.26 43271.92 1364.13   76234.43    75304.31      930.12         0.00  1     2018  0

Mean of total_revenue: 850644.1301318255
Median of total_revenue: 107376.75
```

```
[ ] # Calculate the 25th percentile
    percentile_25 = df['Total Volume'].quantile(0.25)

    # Calculate the 50th percentile (median)
    percentile_50 = df['Total Volume'].quantile(0.50)

    # Calculate the 75th percentile
    percentile_75 = df['Total Volume'].quantile(0.75)

    # Display the results
    print("25th percentile:", percentile_25)
    print("50th percentile:", percentile_50)
    print("75th percentile:", percentile_75)
```

```
25th percentile: 10838.99
50th percentile: 107376.75
75th percentile: 432962.29
```

b. Calculate the interquartile range (IQR) of the data. Round your answer to two decimal places.

```
[ ] # Calculate the interquartile range (IQR)
    iqr = percentile_75 - percentile_25

    # Round the IQR to two decimal places
    iqr = round(iqr, 2)

    # Display the IQR
    print("Interquartile range (IQR):", iqr)
```

```
Interquartile range (IQR): 422123.3
```

c. Given the information, describe the skewness of the data. Is the distribution of data skewed to the left or right?

```
[ ] # Calculate the skewness of the data
    skewness = df['Total Volume'].skew()

    # Display the skewness
    print("Skewness:", skewness)

    # Interpretation of skewness
    if skewness > 0:
        print("The distribution is skewed to the right.")
    elif skewness < 0:
        print("The distribution is skewed to the left.")
    else:
        print("The distribution is symmetric.")
```

```
Skewness: 1.8488333333333333
The distribution is skewed to the right.
```

Question 4

a. Transcribe data for the remaining values of the attribute "Average Price" by averaging all the values of the "Log Price" attribute that lie/cover the quarter "Q1/Q2/Q3" attribute values.

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
0           0      64236.62  1036.74 54454.85  48.16     8696.87     8603.62       93.25         0.00  0     2015  0
1           1      54876.98   674.28 44638.81  58.33     9505.56     9408.07       97.49         0.00  0     2015  0
2           2      118220.22 794.70 109149.67 130.50    8445.35     8042.21      303.14       100.00  0     2015  0
3           3      78992.15  1132.00 71976.41 72.58     5811.16     5677.40      133.76         0.00  0     2015  0
4           4      51039.60  941.48 43838.39  75.78     6183.95     5986.26      197.69         0.00  0     2015  0

Last 5 rows of the dataset:
      Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
18244       7839      149933.10 1351.20 56267.16 159.90    92154.84    29421.44    62733.40         0.00  1     2018  0
18245       7840      191693.90 1655.48 57790.11 1500.13   130747.18   103035.54    27711.64         0.00  1     2018  0
18246       7841      160774.59 2452.79 43325.51 1212.56   113283.73    96358.07    16925.66         0.00  1     2018  0
18247       7842      160529.05 2765.96 75083.67 1083.84   81596.58    80790.77      805.81         0.00  1     2018  0
18248       7843      122357.74 1486.26 43271.92 1364.13   76234.43    75304.31      930.12         0.00  1     2018  0

Mean of total_revenue: 850644.1301318255
Median of total_revenue: 107376.75
```

b. Summarize the number of missing values for each attribute

```
[ ] import pandas as pd

    # Load the dataset
    dataset = fetch_openml(data_id=44065)

    # Convert the dataset to a DataFrame
    df = pd.concat([X, y], axis=1)

    # Calculate the number of missing values for each attribute
    missing_values = df.isnull().sum()

    # Display the results
    print(missing_values)
```

```
Unnamed: 0      0
Total Volume    0
4046            0
4225            0
4770            0
Total Bags      0
Small Bags      0
Large Bags      0
XLarge Bags     0
type            0
year            0
region          0
dtype: int64
```

c. Transcribe data for the remaining values of the attribute "Average Price" by averaging all the values of the "Log Price" attribute that lie/cover the quarter "Q1/Q2/Q3" attribute values.

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
0           0      64236.62  1036.74 54454.85  48.16     8696.87     8603.62       93.25         0.00  0     2015  0
1           1      54876.98   674.28 44638.81  58.33     9505.56     9408.07       97.49         0.00  0     2015  0
2           2      118220.22 794.70 109149.67 130.50    8445.35     8042.21      303.14       100.00  0     2015  0
3           3      78992.15  1132.00 71976.41 72.58     5811.16     5677.40      133.76         0.00  0     2015  0
4           4      51039.60  941.48 43838.39  75.78     6183.95     5986.26      197.69         0.00  0     2015  0

Last 5 rows of the dataset:
      Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
18244       7839      149933.10 1351.20 56267.16 159.90    92154.84    29421.44    62733.40         0.00  1     2018  0
18245       7840      191693.90 1655.48 57790.11 1500.13   130747.18   103035.54    27711.64         0.00  1     2018  0
18246       7841      160774.59 2452.79 43325.51 1212.56   113283.73    96358.07    16925.66         0.00  1     2018  0
18247       7842      160529.05 2765.96 75083.67 1083.84   81596.58    80790.77      805.81         0.00  1     2018  0
18248       7843      122357.74 1486.26 43271.92 1364.13   76234.43    75304.31      930.12         0.00  1     2018  0

Mean of total_revenue: 850644.1301318255
Median of total_revenue: 107376.75
```

d. Describe the attribute: "Date" using concept hierarchy into (Day, Week, Month) (Consider 01-10,20-31 - Day; 10-17 - Week; 18-19- Month)

```
[ ] !rm -r -f data
    !mkdir data
    !cd data
    !ls

    from sklearn.datasets import fetch_openml
    dataset_bunch = fetch_openml(data_id=44065)

    # Extract the data and target
    X = dataset_bunch['data']
    y = dataset_bunch['target']

    # Combine features and target into a single DataFrame
    df = pd.concat([X, y], axis=1)

    # Display the first 5 rows of the dataset
    print("First 5 rows of the dataset:")
    display(df.head())

    # Display the last 5 rows of the dataset
    print("\nLast 5 rows of the dataset:")
    display(df.tail())

    # Calculate the mean of total_revenue
    mean_total_revenue = df['total_revenue'].mean()
    print("\nMean of total_revenue:", mean_total_revenue)

    # Calculate the median of total_revenue
    median_total_revenue = df['total_revenue'].median()
    print("Median of total_revenue:", median_total_revenue)
```

```
First 5 rows of the dataset:
   Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
0           0      64236.62  1036.74 54454.85  48.16     8696.87     8603.62       93.25         0.00  0     2015  0
1           1      54876.98   674.28 44638.81  58.33     9505.56     9408.07       97.49         0.00  0     2015  0
2           2      118220.22 794.70 109149.67 130.50    8445.35     8042.21      303.14       100.00  0     2015  0
3           3      78992.15  1132.00 71976.41 72.58     5811.16     5677.40      133.76         0.00  0     2015  0
4           4      51039.60  941.48 43838.39  75.78     6183.95     5986.26      197.69         0.00  0     2015  0

Last 5 rows of the dataset:
      Unnamed: 0  Total Volume     4046    4225    4770  Total Bags  Small Bags  Large Bags  XLarge Bags  type  year  region
18244       7839      149933.10 1351.20 56267.16 159.90    92154.84    29421.44    62733.40         0.00  1     2018  0
18245       7840      191693.90 1655.48 57790.11 1500.13   130747.18   103035.54    27711.64         0.00  1     2018  0
18246       7841      160774.59 2452.79 43325.51 1212.56   113283.73    96358.07    16925.66         0.00  1     2018  0
18247       7842      160529.05 2765.96 75083.67 1083.84   81596.58    80790.77      805.81         0.00  1     2018  0
18248       7843      122357.74 1486.26 43271.92 1364.13   76234.43    75304.31      930.12         0.00  1     2018  0

Mean of total_revenue: 850644.1301318255
Median of total_revenue: 107376.75
```

Colab paid products - Cancel subscriptions