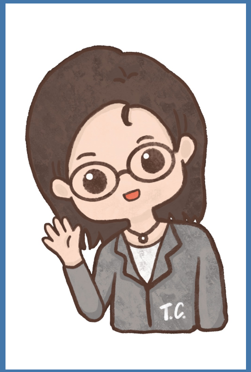
# MTH 208 Exploratory Data Analysis

## Lesson 03: Descriptive Statistics & Data Summarization

Ying-Ju Tessa Chen, PhD

Associate Professor
Department of Mathematics
University of Dayton

@ying-ju
ying-ju
ychen4@udayton.edu



# Learning Objectives

- Measures of Central Tendency: mean, median, mode
- Measures of Spread: range, variance, standard deviation
- Non-Parametric Statistics and Their Significance
- Skewness and Kurtosis
- Measures of Relationship: correlation and covariance
- Interpreting These Statistics in EDA

# Measures of Central Tendency

Central tendency measures are used to identify the center of a data set or its typical value. These measures include the mean, median, and mode, each providing a different perspective on the central value of the data.

### Mean (Arithmetic Average)

- `Definition`: The mean is the sum of all values in a dataset divided by the number of values.
- `Calculation`: Mean = (Sum of all values) / (Number of values)
- `Usage`: Appropriate for interval and ratio data, and when the data does not have extreme outliers.
- `Example`: The average height of a group of people.

# Measures of Central Tendency (Continued)

## Median (Middle Value)

- `Definition`: The median is the middle value in a dataset when it is ordered from smallest to largest. For an even number of observations, it is the average of the two middle numbers.
- `Calculation`: Arrange data in ascending order and identify the middle value.
- `Usage`: Useful for ordinal data or when the dataset contains outliers or is skewed, as it is not affected by extreme values.
- `Example`: The middle income in a list of incomes for a region.

## Mode (Most Frequent Value)

- `Definition`: The mode is the value that appears most frequently in a dataset.
- `Usage`: It can be used for any level of measurement (nominal, ordinal, interval, ratio), and is particularly useful for categorical data.
- `Example`: The most common eye color in a sample of people.

# Measures of Central Tendency (Continued)

## Interpreting Central Tendency in EDA

- `Insights`: These measures help in understanding the general trend or typical value of the data.
- `Contextual Use`: Depending on the nature of the data and its distribution, one measure may be more appropriate than the others.
- `Combination with Other Measures`: Often used alongside measures of spread (like standard deviation) to provide a more complete picture of the data.

# Measures of spread

Measures of spread provide insights into the variability or dispersion within a dataset. They help to understand how much individual data points differ from the central tendency. Key measures include the range, variance, and standard deviation.

**Range**

- `Definition`: The range is the difference between the highest and lowest values in a dataset.
- `Calculation`: Range = Maximum value - Minimum value
- `Usage`: Simplest measure of spread; however, it is sensitive to outliers.
    - `Example`: In a dataset of temperatures over a week, the range is the difference between the highest and lowest recorded temperatures.

# Measures of spread (Continued)

## Variance

- `Definition`: Variance measures the average squared deviation of each number from the mean of the dataset. It gives an idea of how widely the data are spread.
- `Calculation`: The "average" of the squared differences from the Mean.

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  where $X_1, X_2, \ldots, X_n$ are individual observations and $\bar{X}$ is the sample mean.
- `Usage`: More comprehensive than range; used for interval and ratio data. Higher variance indicates greater spread in the data.
- `Example`: Variance in the test scores of a class.

# Measures of spread (Continued)

### Standard Deviation

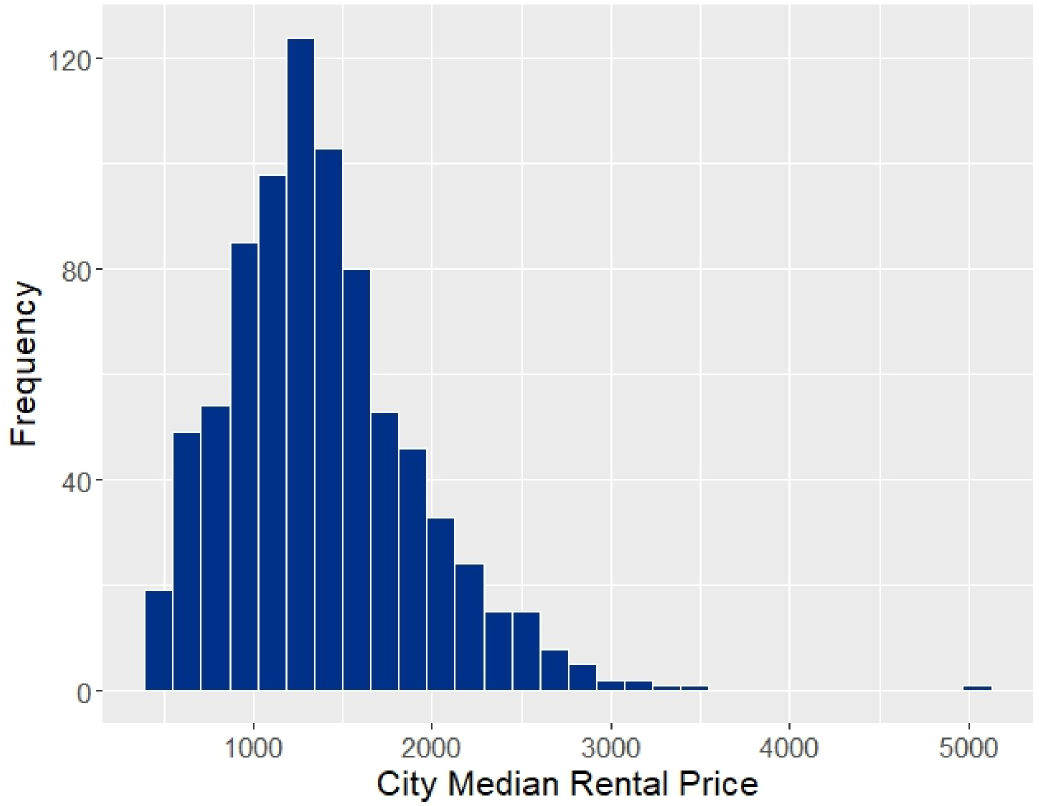
- `Definition`: Standard deviation is the square root of the variance. It is a measure of the amount of variation or dispersion in a set of values.
- `Calculation`: Square root of the variance.
- `Usage`: Widely used because it is in the same unit as the data, making it more interpretable.
- `Example`: Standard deviation in heights within a population.

### Interpreting Spread in EDA

- `Contextual Importance`: Helps in understanding the reliability of the mean. A small spread indicates that the data points tend to be close to the mean, while a large spread indicates more variability.
- `Skewness and Outliers`: These measures can indicate if the data is skewed or if there are outliers affecting the data's spread.
- `Comparative Analysis`: Often used in conjunction with central tendency measures for comprehensive data analysis.

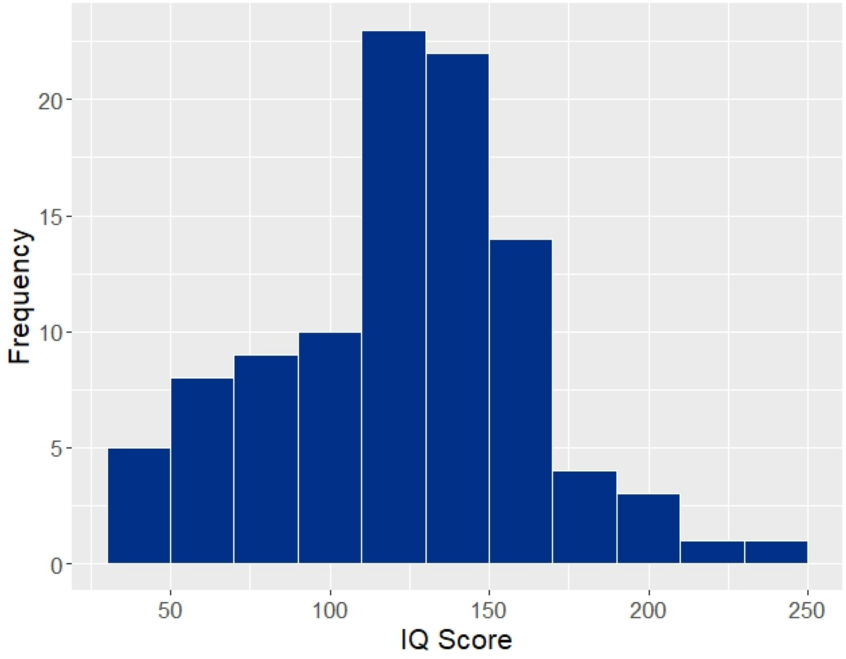
# Case Study I

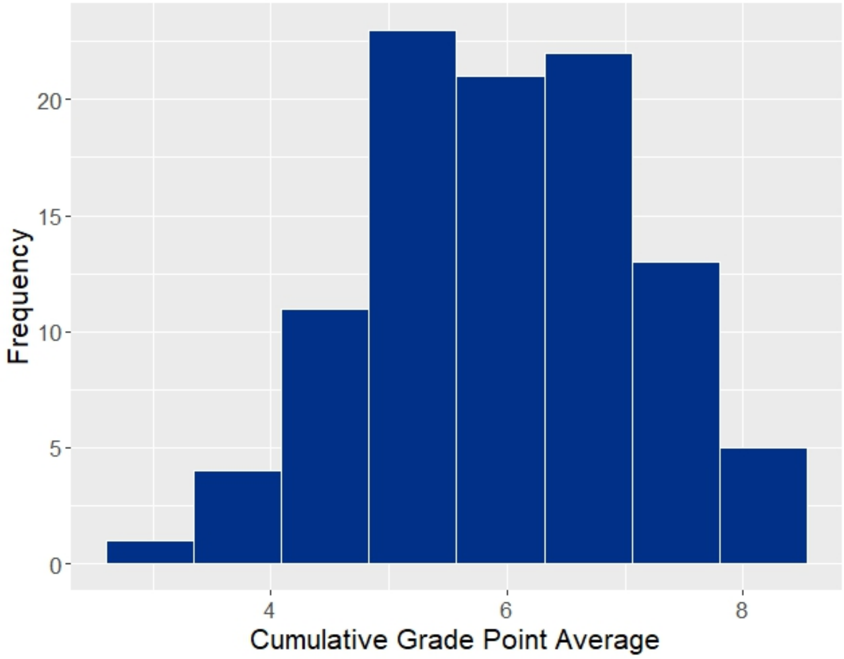
The histogram below shows the City Median Rental Price for a one bedroom home on Zillow in December, 2019. Comment on the distribution of the rental price and discuss the appropriate measures in terms of central tendency and spread.



# Case Study II

The following histograms show the distribution of IQ score and cumulative grade point average from 100 college students respectively. (Source: College Placement Dataset) Comment on the distribution of each histogram and discuss the appropriate measures in terms of central tendency and spread.



# Non-parametric Statistics and Their Significance

Non-parametric statistics are a key area of statistics used for analyzing data that does not assume a specific distribution (like normal distribution). These methods are especially useful when dealing with non-normal datasets or when the data violate the assumptions required for parametric tests.

### Key Concepts of Non-Parametric Statistics

- `Distribution-Free`: Non-parametric methods do not require the data to follow any specific distribution.
- `Types of Data`: Particularly useful for ordinal data or data on a nominal scale. Also applicable to interval or ratio data, especially when it's not normally distributed.
- `Applications`: Commonly used in situations with small sample sizes, heavily skewed data, or data with outliers.

# Non-parametric Statistics and Their Significance (Continued)

## Examples of Non-Parametric Methods

- `Mann-Whitney U Test`: Used to compare differences between two independent groups when the dependent variable is ordinal or continuous but not normally distributed.
- `Kruskal-Wallis Test`: An extension of the Mann-Whitney U Test for comparing more than two groups.
- `Spearman's Rank Correlation Coefficient`: Used to measure the strength and direction of association between two ranked variables.

# Non-parametric Statistics and Their Significance (Continued)

## Significance in EDA

- `Flexibility`: Offers a robust alternative to parametric methods, especially useful in exploratory data analysis where data may not meet parametric assumptions.
- `Handling Skewed Data`: Ideal for analyzing skewed datasets or datasets with outliers where mean and standard deviation might not be appropriate.
- `Insights into Data Structure`: Helps in understanding the underlying structure of the data, which might not be apparent with parametric methods.

# Skewness and Kurtosis

### Introduction to Skewness

- `Definition`: Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. It indicates whether the observations in a dataset are concentrated on one side.
- `Types`:
  - `Positive Skew`: The tail on the right side of the distribution is longer or fatter than the left side.
  - `Negative Skew`: The tail on the left side is longer or fatter than the right side.
- `Interpretation`:
  - Skewness close to 0 indicates a symmetrical distribution.
  - A significantly positive or negative value indicates skewness and potential outliers.

# Skewness and Kurtosis (Continued)

## Introduction to Kurtosis

- `Definition`: Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It describes the peakedness or flatness of the distribution compared to a normal distribution.
- `Types`:
  - `High Kurtosis (>3)`: Indicates a distribution with heavy tails and a sharper peak ("Leptokurtic").
  - `Low Kurtosis (<3)`: Suggests a distribution with light tails and a flatter peak ("Platykurtic").
- `Interpretation`:
  - Kurtosis close to 3 (normal distribution) is considered mesokurtic.
  - Extreme values suggest potential outliers and deviations from the normal distribution.

# Skewness and Kurtosis (Continued)

### Skewness and Kurtosis in EDA

- `Purpose`: Understanding skewness and kurtosis is crucial in EDA to identify the nature of the distribution of the data, which can influence the choice of statistical methods and interpretations.
- `Data Transformation`: Data with high skewness or extreme kurtosis might require transformation to meet the assumptions of various statistical modeling techniques. Activities and Discussion

# Measures of Relationship

Understanding the relationships between variables is a critical aspect of EDA. Two key statistical measures used to assess these relationships are correlation and covariance.

### Covariance

- `Definition`: Covariance is a measure that indicates the extent to which two variables change together. It assesses whether increases in one variable correspond to increases (positive covariance) or decreases (negative covariance) in another.
- `Calculation`: The average of the products of deviations of pairs of observations from their individual means. The covariance between two variables $X$ and $Y$ is given by:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y - \bar{Y})$$

  where:

  - $X_i$ and $Y_i$are the individual values of the variables,
  - $\bar{X}$ and $\bar{Y}$ are the means of the variables,
  - $n$ is the number of data points.

# Measures of relationship (Continue)

## Covariance

- `Interpretation`:
  - `Positive Covariance`: Indicates that as one variable increases, the other tends to increase.
  - `Negative Covariance`: Suggests that as one variable increases, the other tends to decrease.
  - `Zero or Near-Zero Covariance`: Implies no linear relationship between the variables.

# Measures of Relationship (Continued)

## Correlation

- `Definition`: Correlation is a standardized measure of covariance and describes both the strength and direction of the linear relationship between two variables.
- `Types`:
  - `Pearson's Correlation Coefficient`: Measures linear relationship between two interval or ratio variables. The Pearson correlation coefficient between two variables $X$ and $Y$ is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

  - `Spearman’s Rank Correlation`: Used for ordinal variables or when the relationship is not linear.
- `Interpretation`:
  - Values range from -1 to +1.
  - +1: Perfect positive linear relationship.
  - -1: Perfect negative linear relationship.
  - 0: No linear relationship.

# Measures of Relationship (Continued)

### Correlation vs. Covariance

- Covariance provides a directional relationship but not the strength.
- Correlation is a more standardized and interpretable measure, providing both direction and strength of the relationship.

### Significance in EDA

- Understanding these measures helps in identifying potential relationships between variables, which can guide further analysis and modeling.
- They are used to explore data, test hypotheses, and in feature selection for machine learning models.

# Interpreting These Statistics in EDA

**Overview** The interpretation of statistical measures is a critical component of EDA. This process involves understanding what various statistics reveal about a dataset and how this information can inform decision-making, hypothesis testing, and further analysis.

**Key Aspects of Interpretation**

- `Contextual Understanding`: Understanding data within the context of the subject area is crucial. Interpretations should align with the domain knowledge and objectives of the study.
- `Integrative Analysis`:
  - Combine various statistical measures (central tendency, spread, correlation, etc.) to gain a comprehensive understanding of the data.
  - Look for patterns, trends, and anomalies across different measures.

# Interpreting These Statistics in EDA (Continued)

### Key Aspects of Interpretation

- `Correlation vs. Causation`:
    - Distinguish between correlation (two variables moving together) and causation (one variable influencing another). Be cautious about drawing conclusions of causality solely from correlational data.
- `Influence of Skewness and Outliers`:
    - Understand how skewness and outliers impact measures like mean and variance and adjust interpretations accordingly.
    - Use appropriate statistical methods to handle skewed or outlier-heavy data.
- `Role of Non-Parametric Statistics`:
    - Recognize situations where non-parametric methods provide more reliable insights, especially when data do not meet parametric assumptions.

# Interpreting These Statistics in EDA (Continued)

**Practical Application in EDA**

- `Exploratory vs. Confirmatory`: EDA is exploratory, aimed at uncovering insights and forming hypotheses, not confirming them.
- `Visual Representation`: Use graphs and plots alongside numerical measures for a more intuitive understanding of data.
- `Data-Driven Insights`: Use statistical interpretations to guide decisions on further data processing, feature selection, and potential areas for in-depth analysis.

# References

The lectures of this course are based on the ideas from the following references.


- Exploratory Data Analysis by John W. Tukey
- A Course in Exploratory Data Analysis by Jim Albert
- The Visual Display of Quantitative Information by Edward R. Tufte
- Data Science for Business: what you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic