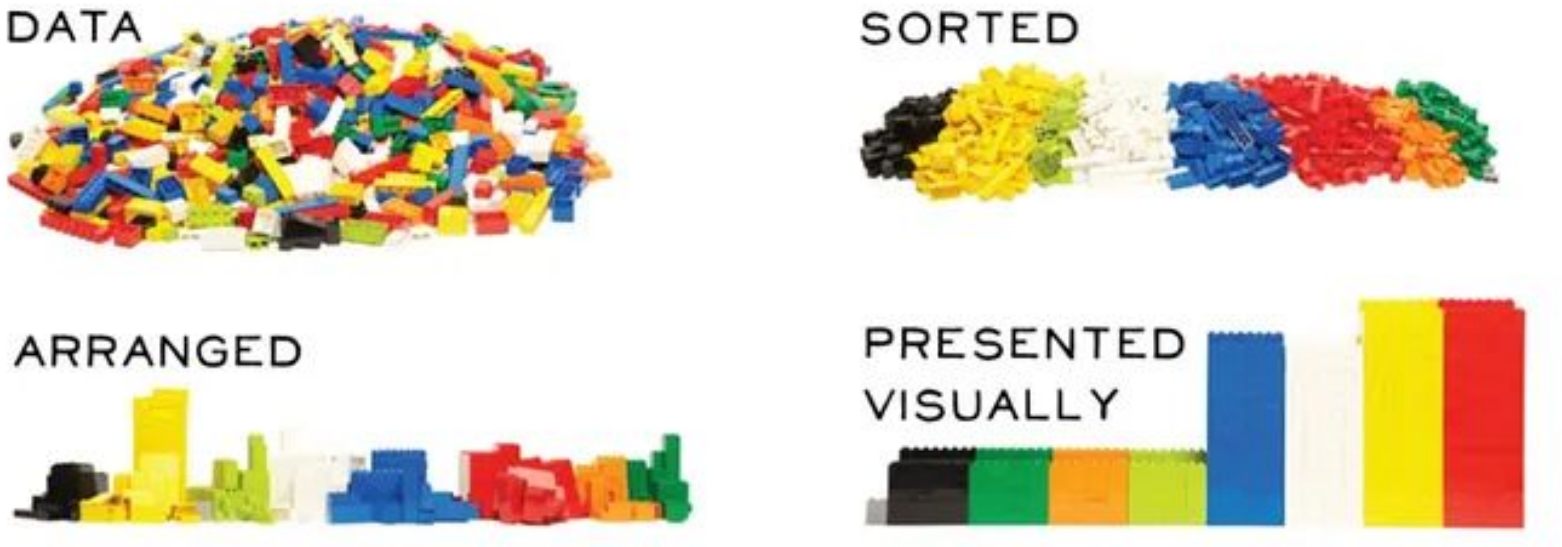
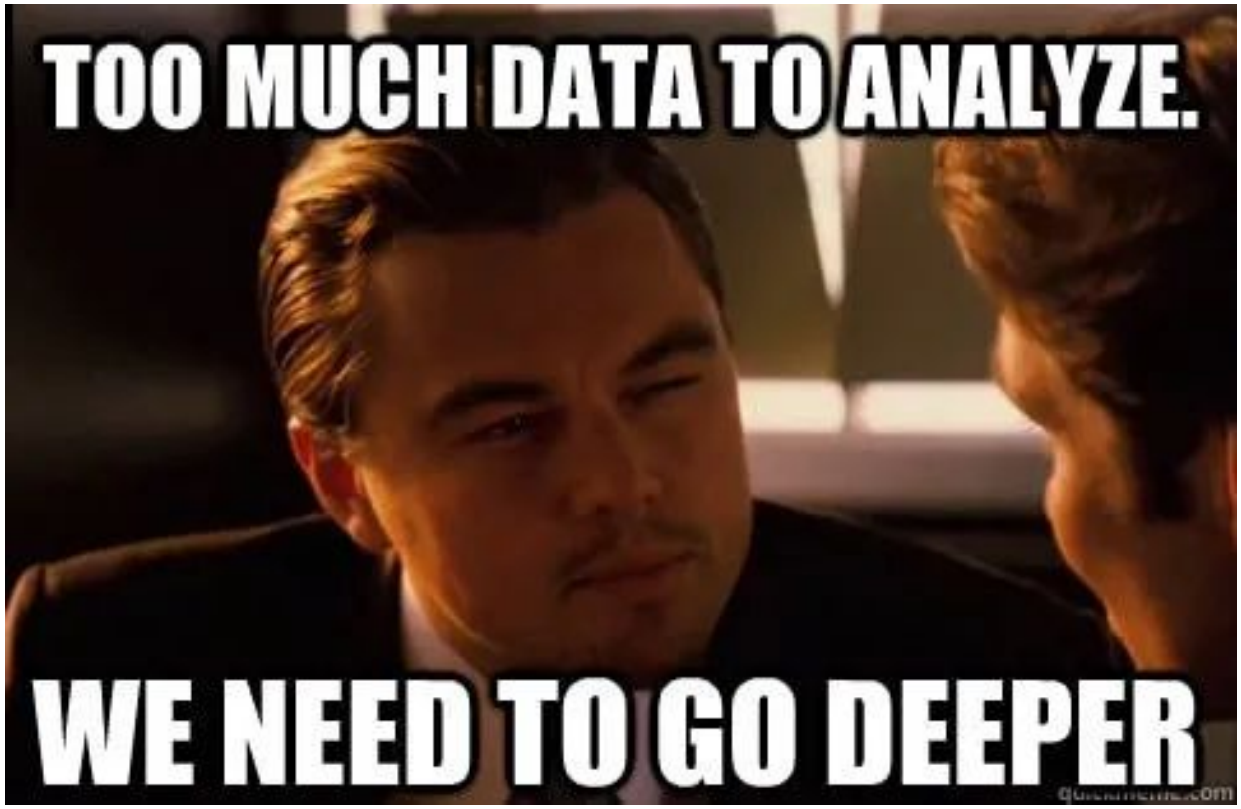
ITDO6014
AI AND DS-1

Module 5: Exploratory Data Analysis

Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

- “Torture the data, and it will confess to anything.”
- — Ronald Coase
- With proper use of data one could rule the entire world as well!
- But, raw, unprocessed data isn’t of much use unless you derive insights from it.
- Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

Exploratory Data Analysis

- With EDA, you can find anomalies in your data, such as outliers or unusual observations, uncover patterns, understand potential relationships among variables, and generate interesting questions or hypotheses that you can test later using more formal statistical methods.
- Exploratory data analysis is like detective work: you're searching for clues and insights that can lead to the identification of potential root causes of the problem you are trying to solve. You explore one variable at a time, then two variables at a time, and then many variables at a time
- EDA is generally classified into two methods, non-graphical or graphical. And each method can be applied to one variable/column (univariate) or a combination of variables/columns(bivariate).

Exploratory Data Analysis

- Typical Data Formats:

https://www.weirdgeek.com/2018/12/common-file-formats-used-in-data-science/

- CSV
- Text Files
- JSON
- Microsoft Excel File
- SAS
- SQL
- Python Pickle File
- Stata
- HDF5
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery

Exploratory Data Analysis

- **Objective of Exploratory Data Analysis:**
- Identifying and removing data outliers
- Identifying trends in time and space
- Uncover patterns related to the target
- Creating hypotheses and testing them through experiments
- Identifying new sources of data

Exploratory Data Analysis

- **Steps Involved in Exploratory Data Analysis (EDA)**
- **1. Data Collection**
- **2. Finding all Variables and Understanding Them**
- **3. Cleaning the Dataset**
- **4. Identify Correlated Variables**
- **5. Choosing the Right Statistical Methods**
- **6. Visualizing and Analyzing Results**

Types of Exploratory Data Analysis

- **1. Univariate Non-Graphical**
- It is the simplest of all types of data analysis used in practice. As the name suggests, uni means only one variable is considered whose data (referred to as population) is compiled and studied. The main aim of univariate non-graphical EDA is to find out the details about the distribution of the population data and to know some specific parameters of statistics. The significant parameters which are estimated from a distribution point of view are as follows:

Types of Exploratory Data Analysis

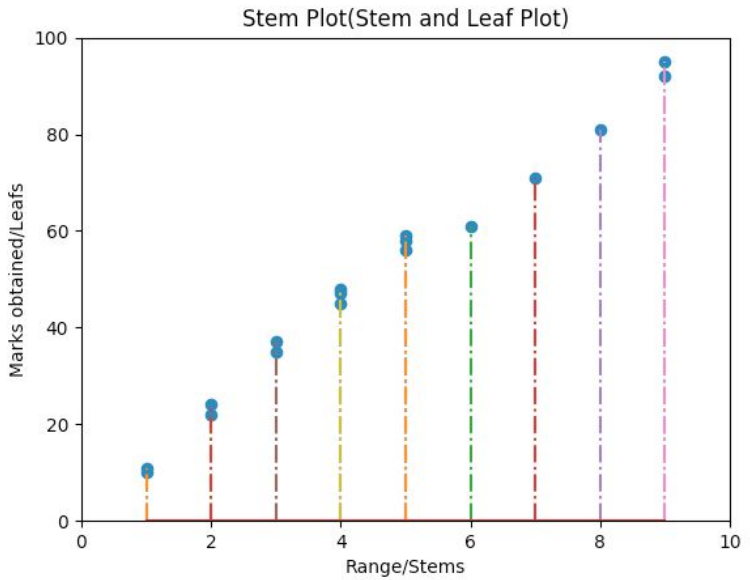
- **Central Tendency:** This term refers to values located at the data's central position or middle zone. The three generally estimated parameters of central tendency are mean, median, and mode. Mean is the average of all values in data, while the mode is the value that occurs the maximum number of times. The Median is the middle value with equal observations to its left and right.

Types of Exploratory Data Analysis

- **Range:** The range is the difference between the maximum and minimum value in the data, thus indicating how much the data is away from the central value on the higher and lower side.
- **Variance and Standard Deviation:** Two more useful parameters are standard deviation and variance. Variance is a measure of dispersion that indicates the spread of all data points in a data set. It is the measure of dispersion mostly used and is the mean squared difference between each data point and mean, while standard deviation is the square root value of it. The larger the value of standard deviation, the farther the spread of data, while a low value indicates more values clustering near the mean.

Types of Exploratory Data Analysis

- **2. Univariate Graphical:**
- **Stem-and-leaf Plots:** This is a very simple but powerful EDA method used to display quantitative data but in a shortened format. It displays the values in the data set, keeping each observation intact but separating them as stem (the leading digits) and remaining or trailing digits as leaves. But histogram is mostly used in its place now.



Types of Exploratory Data Analysis

- **Histograms (Bar Charts):** These plots are used to display both grouped or ungrouped data. On the x-axis, values of variables are plotted, while on the y-axis are the number of observations or frequencies. Histograms are very simple to quickly understand your data, which tell about values of data like central tendency, dispersion, outliers, etc. The simplest fundamental graph is a histogram, which is a bar plot with each bar representing the frequency, i.e., the count or proportion (the ratio of count to the total count of occurrences) for various values.

Types of Exploratory Data Analysis

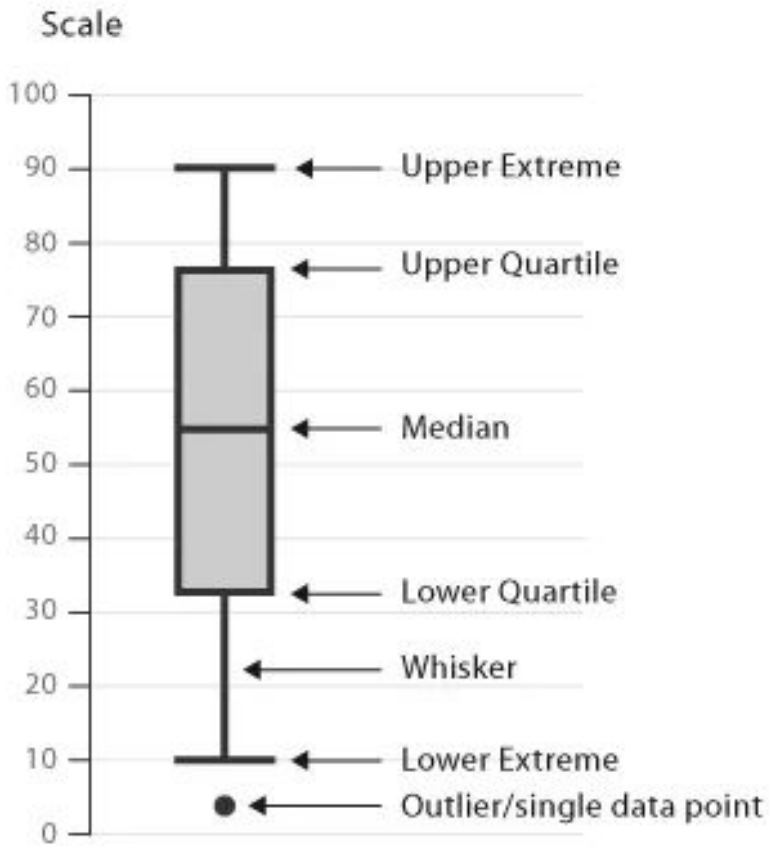
- There are many types of histograms, a few of which are listed below:
- **Simple Bar Charts:** These are used to represent categorical variables with rectangular bars, where the different lengths correspond to the values of the variables.
- **Multiple or Grouped charts:** Grouped bar charts are bar charts representing multiple sets of data items for comparison where a single color is used to denote one specific series in the dataset.

Types of Exploratory Data Analysis

- **Percentage Bar Charts:** These are bar graphs that depict the data in the form of percentages for each observation. The following image shows a percentage bar chart with dummy values.

Types of Exploratory Data Analysis

- **Box plots**
- Box plot shows us the median of the data, which represents where the middle data point is. The upper and lower quartiles represent the 75 and 25 percentile of the data respectively. The upper and lower extremes shows us the extreme ends of the distribution of our data. Finally, it also represents outliers, which occur outside the upper and lower extremes.



Types of Exploratory Data Analysis

- **3. Multivariate Non-Graphical**
- The multivariate non-graphical exploratory data analysis technique is usually used to show the connection between two or more variables with the help of either cross-tabulation or statistics.
- For categorical data, an extension of tabulation called cross-tabulation is extremely useful. For two variables, cross-tabulation is preferred by making a two-way table with column headings that match the amount of one variable and row headings that match the amount of the opposite two variables, then filling the counts with all subjects that share an equivalent pair of levels.

Types of Exploratory Data Analysis

- **4. Multivariate Graphical:**
- Graphics are used in multivariate graphical data to show the relationships between two or more variables. Here the outcome depends on more than two variables, while the change-causing variables can also be multiple.
- Some common types of multivariate graphics include:
- **Scatter Plot**
- The essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis.

Types of Exploratory Data Analysis

- **D) Bubble Chart**
- Bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.
- **E) Heat Map**
- A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among variables. It assists in finding the best features suitable for building accurate Machine Learning models.

Correlation and Covariance

- **Covariance, variance vs. covariance vs. correlation:**
- Variance, covariance, and correlation are all measures used in statistics to describe the relationship between two or more variables.
- **Variance**: Variance measures how much a set of numbers vary from the average (mean) value. It gives an idea of the dispersion or spread of data points. Mathematically, variance is calculated by taking the average of the squared differences between each data point and the mean.

Correlation and Covariance

- **Population Variance**: The variance calculated using the entire population of data points. It is denoted by $\sigma 2$ (sigma squared).
- **Sample Variance**: The variance calculated using a sample of data points from a larger population. It is an estimate of the population variance and is denoted by $s2$.
- **Adjusted Variance**: In the context of sample variance, an adjusted variance may be calculated to correct for bias, especially when using a small sample size. One common adjustment involves dividing by $n-1$ instead of n, where n is the sample size.
- **Conditional Variance**: The variance of a random variable conditional on the occurrence of another event or the value of another variable.

Correlation and Covariance

- **Covariance**: Covariance is a measure of how two variables change together. It indicates the direction of the linear relationship between two variables. A positive covariance means that when one variable increases, the other tends to increase as well, while a negative covariance means that when one variable increases, the other tends to decrease. Covariance is calculated by taking the average of the product of the differences of each variable from their respective means.

Correlation and Covariance

- **Population Covariance**: The covariance calculated using the entire population of data pairs. It is denoted by $\text{Cov}(X,Y)$ or $\sigma(X,Y)$.
- **Sample Covariance**: The covariance calculated using a sample of data pairs from a larger population. It is an estimate of the population covariance and is denoted by $\text{Cov}(X,Y)$ or $s(X,Y)$.
- **Covariance Matrix**: In multivariate statistics, the covariance matrix, also known as the variance-covariance matrix, is a square matrix that contains the variances of variables along the diagonal and covariances between pairs of variables in the off-diagonal elements. It provides a comprehensive summary of the relationships between multiple variables.

Correlation and Covariance

- **Autocovariance**: In time series analysis, autocovariance measures the covariance between a time series and a lagged version of itself at different time lags. It is used to assess the dependence structure of a time series.

Correlation and Covariance

- **Correlation**: Correlation is a standardized measure of the relationship between two variables. Unlike covariance, correlation does not have units and is bounded between -1 and 1. A correlation of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Correlation is calculated by dividing the covariance by the product of the standard deviations of the two variables.

Correlation and Covariance

- **Correlation**:
- **Pearson Correlation Coefficient**: The most common type of correlation coefficient, which measures the linear relationship between two variables. It is denoted by ρ (rho) for the population correlation coefficient and r for the sample correlation coefficient.
- **Spearman's Rank Correlation Coefficient**: A non-parametric measure of correlation that assesses how well the relationship between two variables can be described using a monotonic function. It is based on the ranks of the data points rather than their actual values.
- **Kendall's Tau**: Another non-parametric measure of correlation that assesses the association between two variables based on the ranks of the data points. It measures the ordinal association between two variables.

Correlation and Covariance

- Suppose we have data on the number of hours studied (X) and the corresponding exam scores (Y) for five students:

Hours Studied (X)	Exam Score (Y)
3	75
4	80
6	90
7	85
9	95

Correlation and Covariance

- **Variance**:
- Variance of Hours Studied (X): Mean of X: (3 + 4 + 6 + 7 + 9) / 5 = 5.8 Variance of X = [(3 - 5.8)^2 + (4 - 5.8)^2 + (6 - 5.8)^2 + (7 - 5.8)^2 + (9 - 5.8)^2] / 5 = [8.84 + 4.84 + 0.04 + 3.24 + 12.96] / 5 = 29.92 / 5 = 5.984
- Variance of Exam Score (Y): Mean of Y: (75 + 80 + 90 + 85 + 95) / 5 = 85 Variance of Y = [(75 - 85)^2 + (80 - 85)^2 + (90 - 85)^2 + (85 - 85)^2 + (95 - 85)^2] / 5 = [100 + 25 + 25 + 0 + 100] / 5 = 250 / 5 = 50

Correlation and Covariance

- **Covariance**:
- Covariance between X and Y: Mean of X: 5.8, Mean of Y: 85
  Covariance = [(3 - 5.8)*(75 - 85)* + *(4 - 5.8)*(80 - 85) + (6 - 5.8)*(90 - 85)* + *(7 - 5.8)*(85 - 85) + (9 - 5.8)*(95 - 85)] / 5* = *[(-2.8)*(-10) + (-1.8)*(-5) + (0.2)*5 + (1.2)*0 + (3.2)*10] / 5 = (28 + 9 + 1 + 0 + 32) / 5 = 70 / 5 = 14
- **Correlation**:
- Correlation between X and Y: Correlation = Covariance / (Standard Deviation of X * Standard Deviation of Y) = 14 / (√5.984 * √50) ≈ 14 / (2.445 * 7.071) ≈ 14 / 17.305 ≈ 0.808

Correlation and Covariance

- In this example:
- The variance of hours studied (X) is approximately 5.984.
- The variance of exam scores (Y) is 50.
- The covariance between hours studied (X) and exam scores (Y) is 14.
- The correlation between hours studied (X) and exam scores (Y) is approximately 0.808, indicating a strong positive linear relationship between the two variables.

Degree of Freedom

- In statistics, the degree of freedom (df) refers to the number of independent pieces of information or parameters that are free to vary when estimating a statistic. In simpler terms, it represents the number of values in the final calculation of a statistic that are allowed to vary.
- Suppose you have a sample of size *n* (the number of observations) and you want to calculate the sample variance.
- The formula for the sample variance is:

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

Degree of Freedom

- Where:
- *Xi* represents each individual observation.
- $\bar{X}$ represents the sample mean.
- *n* represents the sample size.
- *s*2 represents the sample variance.

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

Degree of Freedom

- **Total Observations**: In the numerator of the formula, you are calculating the sum of the squared differences between each observation and the sample mean. Since you're calculating the differences, you have n values that are free to vary (i.e., $\bar{X}i - \bar{X}$).
- **Sample Mean**: Once you calculate the sample mean ($\bar{X}$), it is based on all n observations. Therefore, it is not free to vary; its value is fixed based on the data. So, you lose one degree of freedom when you fix the sample mean.

Degree of Freedom

- **Adjustment for Sample Variance**: Finally, in the denominator, you divide by $n-1$. This adjustment is made to correct for the fact that you used the sample mean ($\bar{X}$) instead of the population mean (μ) in calculating the squared differences. Dividing by $n-1$ rather than n helps to provide an unbiased estimate of the population variance. Therefore, you lose one more degree of freedom when making this adjustment.

Statistical Methods for Evaluation

- Statistical methods for evaluation are used to assess the performance, significance, or quality of models, experiments, or processes. Here are some common statistical methods for evaluation:
- **Descriptive Statistics**: Descriptive statistics summarize and describe the main features of a dataset. Measures such as mean, median, mode, standard deviation, and percentiles are often used to provide insights into the central tendency, dispersion, and shape of the data.
- **Hypothesis Testing**: Hypothesis testing is a method for making inferences about population parameters based on sample data. It involves formulating a null hypothesis (H0) and an alternative hypothesis (H1), selecting a significance level (alpha), and using statistical tests such as t-tests, chi-square tests, ANOVA, or z-tests to determine whether to reject or fail to reject the null hypothesis.

Statistical Methods for Evaluation

- **Confidence Intervals**: Confidence intervals provide a range of values within which a population parameter is estimated to lie with a certain level of confidence. They are calculated based on sample data and provide information about the precision of the estimates.
- **Resampling Methods**: Resampling methods involve repeatedly drawing samples from the observed data to estimate the distribution of a statistic or to assess the variability of a model's performance. Common resampling techniques include bootstrap resampling and cross-validation.
- **Regression Analysis**: Regression analysis is used to model the relationship between a dependent variable and one or more independent variables. It helps to understand the strength and direction of the relationships and to make predictions based on the observed data.

Statistical Methods for Evaluation

- **Model Evaluation Metrics**: Model evaluation metrics quantify the performance of predictive models. They include metrics such as accuracy, precision, recall, F1-score, ROC curve, AUC-ROC, and confusion matrix, depending on the type of problem (classification, regression, etc.).
- **ANOVA (Analysis of Variance)**: ANOVA is a statistical technique used to compare means across multiple groups to determine whether there are statistically significant differences between the groups. It assesses the variation within groups and between groups to determine whether the group means are equal.
- **Experimental Design**: Experimental design involves planning and conducting experiments to test hypotheses and make causal inferences. It includes methods such as randomized controlled trials, factorial designs, and Latin squares to control for confounding variables and maximize the

Statistical Methods for Evaluation

- **Time Series Analysis**: Time series analysis is used to analyze data collected over time to identify patterns, trends, and seasonality. It includes methods such as autoregressive integrated moving average (ARIMA) modeling, exponential smoothing, and spectral analysis.

ANOVA Test

- https://www.cuemath.com/anova-formula/
- https://www.statisticshowto.com/tables/f-table/