**Data Analysis Lab Report**

**A Project Report Submitted to**



**Acropolis Institute of Technology and Research Indore Towards Partial Fulfilment for the Award of**

**Bachelor of Technology**

**(Computer Science and Engineering)**

**Submitted By**

**Aryan Gupta(0827CS211043)**

**Under the Supervision of**

**Prof. Anurag Punde**

**Department of Computer Science and Engineering**

**Acropolis Institute of Technology & Research, Indore**

**July-Dec 2023**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Experiment** | **Remarks** |
| 1. | Data Analysis Questions:   1. Data Analysis Principles 2. Statistical Analytics 3. Hypothesis Testing 4. Regression 5. Correlation 6. ANOVA |  |
| 2. | Dashboards:   1. Exploring Car Collection 2. Exploring Order Dataset 3. Shop Sales Data Report 4. Exploring Cookie data: Trends & Analysis Report 5. Store Data Analysis 6. Exploring Loan Dataset 7. Sale Samples: A Detailed Report |  |
| 3. | Reports:   1. Exploring Car Collection Dataset 2. Exploring Order Dataset 3. Exploring Cookie Data: Trends & Analysis Report 4. Exploring Loan Dataset 5. Shop Sales Data Report 6. Sales Samples: A Detailed Report 7. Store Dataset Report |  |
| 4. | Airline Shares Forecast Analysis (2015-2050) |  |

INDEX

Enrolment No. : 0827CS211043

Name: Aryan Gupta

1. Data Analysis Principles:

Data analysis is a comprehensive field that involves various methods, techniques, and principles to collect, process, and interpret data. Here are some key principles and concepts in data analysis:

1. Data Collection: Gathering raw data from relevant sources.
2. Data Cleaning: Removing inaccuracies and inconsistencies to ensure data quality.
3. Data Exploration: Understanding the structure, patterns, and anomalies in the data.
4. Data Modeling: Creating models that represent the data’s underlying structure.
5. Statistical Analysis: Applying statistical methods to infer properties of the population.
6. Machine Learning: Using algorithms to predict outcomes and discover patterns.
7. Data Interpretation: Drawing conclusions and making decisions based on data analysis.
8. Data Visualization: Presenting data in graphical format to communicate information clearly and efficiently.

There are also different types of data analysis, such as:

1. Descriptive Analytics: Describes what has happened using historical data.
2. Diagnostic Analytics: Explains why something happened.
3. Predictive Analytics: Predicts what is likely to happen in the future.
4. Prescriptive Analytics: Suggests actions to achieve desired outcomes.

Each type of analysis serves a different purpose and requires specific techniques and tools. For instance, regression analysis is used to understand relationships between variables, while cluster analysis groups similar data points together.

1. Statistical Analysis:

Statistical analysis is the process of collecting, organizing, interpreting, and presenting data to uncover patterns and trends. It is a fundamental component of data science and research, allowing researchers to make informed decisions based on empirical evidence.

Steps in Statistical Analysis:

1. Define the Research Question: Identify what you want to investigate.
2. Collect Data: Gather data via surveys, experiments, etc.
3. Data Cleaning and Preparation: Handle missing values and outliers.
4. Exploratory Data Analysis (EDA): Summarize data using descriptive stats and visualizations.
5. Choose Statistical Methods: Select appropriate tests and models.
6. Perform Statistical Analysis: Use tools like R, Python, SPSS.
7. Interpret Results: Draw conclusions from the analysis.
8. Report Findings: Present results clearly with tables, charts, and explanations.
9. Hypothesis Testing:

Hypothesis testing is a core component of inferential statistics, used to determine whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis. The process begins with formulating two hypotheses: the null hypothesis (H₀), which represents no effect or status quo, and the alternative hypothesis (H₁), which suggests a significant effect or difference. Researchers collect sample data and calculate a test statistic, which is then compared against a critical value from a probability distribution (such as the t-distribution or chi-square distribution) to determine statistical significance.

A p-value is calculated, representing the probability of obtaining the observed results if the null hypothesis is true. A p-value below a predetermined significance level (usually 0.05) indicates strong evidence against the null hypothesis, leading to its rejection. Common tests include the t-test (for comparing means), ANOVA (for comparing means among multiple groups), chi-square test (for categorical data), and regression analysis (for examining relationships between variables).

Assumptions underlying the chosen test, such as normality, independence, and homoscedasticity, must be validated to ensure accurate results. Hypothesis testing provides a structured method for making data-driven decisions, distinguishing between random variation and genuine effects in scientific and practical research.

1. Regression:

Regression analysis is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. The simplest form, linear regression, models the relationship between two continuous variables with a linear equation, allowing predictions of the dependent variable based on the independent variable. Multiple regression extends this to include multiple predictors.

The regression equation takes the form y = beta0 + beta1x1 + beta2x2 + ... +beta nxn + epsilon ), where ( y ) is the dependent variable, ( beta )s are coefficients, ( x )s are independent variables, and ( epsilon ) represents the error term. The coefficients represent the change in the dependent variable for a one-unit change in the predictor.

Regression analysis helps identify the strength and type of relationships, control for confounding variables, and make forecasts. Assumptions such as linearity, independence, homoscedasticity, and normality of residuals must be checked to validate the model. It's widely used in fields like economics, biology, engineering, and social sciences for predictive and explanatory purposes.

1. Correlation:

Correlation measures the strength and direction of the relationship between two continuous variables. The correlation coefficient, denoted as rrr, ranges from -1 to 1. An r value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. The most common method for calculating correlation is the Pearson correlation coefficient, which assesses linear relationships. For non-linear relationships or ordinal data, Spearman's rank correlation can be used. Interpreting correlation involves examining both the sign (positive or negative) and the magnitude of r.

A positive correlation means that as one variable increases, the other also increases, while a negative correlation means that as one variable increases, the other decreases .It’s important to note that correlation does not imply causation. While it can show that two variables are related, it cannot determine if one variable causes the other to change. Correlation analysis is frequently used in fields like finance, medicine, and social sciences to identify and quantify relationships between variables.

1. ANOVA:

ANOVA is a statistical method used to compare the means of three or more groups to see if there is a significant difference among them. It extends the t-test to multiple groups by examining the variability within groups and between groups. The null hypothesis (H₀) in ANOVA states that all group means are equal, while the alternative hypothesis (H₁) suggests that at least one group mean is different.

The test involves calculating the F-statistic, which is the ratio of the variance between the group means to the variance within the groups. A high F-value indicates a greater likelihood that the observed differences among group means are real and not due to random chance. If the p-value associated with the F-statistic is below a predetermined significance level (commonly 0.05), the null hypothesis is rejected.

ANOVA assumptions include independence of observations, normally distributed groups, and homogeneity of variances. It’s widely used in experimental and observational studies across various fields like psychology, biology, and marketing to compare multiple treatments or conditions.

Assumptions of ANOVA

1. Independence of observations: Each group sample must be independent of the others.
2. Normality: The data within each group should be approximately normally distributed.
3. Homogeneity of variances: The variance among the groups should be roughly equal.