**CS3352- FOUNDATIONS OF DATA SCIENCE**

**UNIT I: INTRODUCTION**

**2-MARK QUESTIONS AND ANSWERS**

1. **What is Data Science?**
   * Data Science is the process of extracting knowledge and insights from structured and unstructured data using scientific methods, algorithms, and systems.
2. **Mention any two benefits of Data Science.**
   * Improved decision-making and better customer experiences.
3. **What are the facets of data?**
   * Volume, variety, velocity, veracity, and value.
4. **What is data mining?**
   * Data mining is the process of discovering patterns and knowledge from large amounts of data.
5. **Define data warehousing.**
   * Data warehousing is the storage of large amounts of data by a business in a data warehouse system.
6. **List two basic statistical descriptions of data.**
   * Mean and standard deviation.
7. **What is the goal of the data science process?**
   * To transform data into actionable insights.
8. **What is exploratory data analysis?**
   * It's the approach of summarizing main characteristics of data often using visual methods.
9. **What is the role of data preparation in Data Science?**
   * It involves cleaning and transforming raw data into a usable format.
10. **What does building the model mean in data science?**

* It refers to applying algorithms to data to create predictive or descriptive models.

**UNIT II: DESCRIBING DATA**

**2-MARK QUESTIONS AND ANSWERS**

1. **What are the two main types of data?**
   * Qualitative and Quantitative.
2. **Define a variable.**
   * A variable is a characteristic or attribute that can assume different values.
3. **What are nominal variables?**
   * Variables that name or categorize without order (e.g., gender, color).
4. **What are ordinal variables?**
   * Variables that have a meaningful order but not evenly spaced (e.g., rankings).
5. **Name any one type of graph used to describe data.**
   * Bar graph.
6. **What is the mean of a dataset?**
   * The average of all values.
7. **Define variability.**
   * Variability refers to how spread out the data is.
8. **What is a standard (z) score?**
   * It measures how many standard deviations a value is from the mean.
9. **What does a normal distribution represent?**
   * A symmetric, bell-shaped distribution where most values cluster around the mean.
10. **Name a table used to describe data.**

* Frequency table.

**UNIT III: DESCRIBING RELATIONSHIPS**

**2-MARK QUESTIONS AND ANSWERS**

1. **What is correlation?**
   * Correlation measures the strength and direction of the relationship between two variables.
2. **What is a scatter plot?**
   * A graph that shows the relationship between two quantitative variables.
3. **What does the correlation coefficient (r) indicate?**
   * It indicates the strength and direction of a linear relationship.
4. **What is regression?**
   * Regression analyzes the relationship between a dependent and independent variable.
5. **What is the least squares regression line?**
   * It's the best-fitting straight line that minimizes the sum of squared errors.
6. **What is the standard error of estimate?**
   * It measures the accuracy of predictions made by a regression line.
7. **What is multiple regression?**
   * A statistical technique that uses multiple variables to predict a dependent variable.
8. **What does r² represent?**
   * It represents the proportion of variance explained by the model.
9. **What does regression towards the mean mean?**
   * It’s the phenomenon that extreme scores tend to move closer to the mean.
10. **Give an example of a positive correlation.**

* As height increases, weight tends to increase.

**UNIT IV: PYTHON LIBRARIES FOR DATA WRANGLING**

**2-MARK QUESTIONS AND ANSWERS**

1. **What is NumPy used for?**
   * NumPy is used for numerical operations on large arrays and matrices.
2. **What is a structured array in NumPy?**
   * It's an array where each element can have named fields of different types.
3. **What is boolean indexing in NumPy?**
   * It allows filtering array elements based on boolean conditions.
4. **What is Pandas used for?**
   * Pandas is used for data manipulation and analysis.
5. **Define hierarchical indexing.**
   * It's a way of indexing data with multiple levels.
6. **What is a pivot table?**
   * A table that summarizes data using aggregation functions.
7. **What is missing data in Pandas?**
   * Data that is not available or is NaN.
8. **What function is used for group-wise aggregation in Pandas?**
   * groupby()
9. **What is fancy indexing?**
   * Indexing using arrays of indices.
10. **Give an example of an aggregation function.**

* mean()

**UNIT V: DATA VISUALIZATION**

**2-MARK QUESTIONS AND ANSWERS**

1. **What is Matplotlib?**
   * A Python library used for creating static, interactive, and animated plots.
2. **What is a line plot?**
   * A graph that connects data points with a continuous line.
3. **What is a scatter plot used for?**
   * To show relationships between two variables.
4. **What does a histogram show?**
   * The distribution of a dataset.
5. **What is the purpose of legends in plots?**
   * To label different data series in a plot.
6. **How are subplots used in Matplotlib?**
   * To display multiple plots in one figure.
7. **What is Seaborn?**
   * A data visualization library based on Matplotlib with higher-level interfaces.
8. **What does annotation mean in a plot?**
   * Adding text or labels to points on a graph.
9. **What is 3D plotting used for?**
   * To visualize data with three variables.
10. **What is Basemap used for?**

* To plot geographic data and maps.