PROJECT DOCUMENTION TEMPLATE

1. Date
2. Project title
3. Content
4. Objective
5. Introduction
6. Methodology
7. Data gathering/source
8. Dataset description
9. Tools
10. Data cleaning and preparation
11. Data model
12. Data insights and recommendations
13. Dashboard
14. Dashboard wireframe
15. Visual colours
16. Dashboard colour
17. Work on the attached data freely and creatively to extract different insights
18. Add at least 5 questions and answer them through your analysis of the data
19. Extract statistical summaries using formulas and functions.
    1. Mean,
    2. Median,
    3. Mode,
    4. Range,
    5. Box plot,
    6. Q1,
    7. Q2,
    8. Q3,
    9. Q4,
    10. Standard deviation,
    11. Variance,
    12. Histogram,
    13. Standardization with Z-test,
    14. Hypothesis testing (Null, alternative, and T-test),
    15. Prediction

Title: Healthcare Cost Analysis Using Power BI

Objective: The goal of this data analysis project is to gain insights into the factors influencing healthcare charges based on a dataset containing information about individuals' age, sex, BMI (Body Mass Index), number of children, smoking status, region, and healthcare charges. We aim to identify patterns, correlations, and trends that can help in understanding the key drivers of healthcare costs.

Data Source: The dataset used for this project includes the following fields:

1. Age: Age of the individual.
2. Sex: Gender of the individual (Male/Female).
3. BMI: Body Mass Index, a measure of body fat based on height and weight.
4. Children: Number of children or dependents covered by health insurance.
5. Smoker: Smoking status of the individual (Smoker/Non-smoker).
6. Region: Geographic region of the individual (e.g., Northeast, Northwest, Southeast, Southwest).
7. Charges: Healthcare charges incurred by the individual.

Steps in the Data Analysis Project:

1. **Data Exploration and Cleaning:**
   * Load the dataset into Power BI.
   * Explore the distribution of each variable.
   * Check for missing values and outliers.
   * Address any data quality issues.
2. **Descriptive Statistics:**
   * Calculate summary statistics for age, BMI, and charges.
   * Explore the distribution of charges based on different categories (e.g., smoker vs. non-smoker, male vs. female).
3. **Correlation Analysis:**
   * Investigate correlations between variables, especially between age, BMI, and charges.
   * Use visualizations like scatter plots or correlation matrices.
4. **Demographic Analysis:**
   * Analyze the distribution of charges based on demographics (age, sex, region).
   * Explore the impact of having children on healthcare charges.
5. **Smoking Status Analysis:**
   * Compare healthcare charges between smokers and non-smokers.
   * Use visualizations to highlight differences in charges.
6. **BMI Analysis:**
   * Explore the relationship between BMI and healthcare charges.
   * Group individuals into BMI categories and analyze charge distributions.
7. **Region-wise Analysis:**
   * Compare healthcare charges across different regions.
   * Identify any regional trends or variations.
8. **Multivariate Analysis:**
   * Conduct multivariate analysis considering multiple factors simultaneously.
   * Use visualizations or statistical models to identify the combined impact of variables on healthcare charges.
9. **Dashboard Creation:**
   * Design a Power BI dashboard to present key findings.
   * Include interactive visualizations for stakeholders to explore data dynamically.
10. **Insights and Recommendations:**
    * Summarize key insights gained from the analysis.
    * Provide recommendations based on the identified patterns and trends.

By conducting this data analysis project, you can offer valuable insights into the factors influencing healthcare charges, helping stakeholders make informed decisions and potentially optimize healthcare costs.

**DATA EXPLORATION** is a crucial initial step in any data analysis project. It involves getting to know your dataset, understanding its structure, and identifying patterns or trends that may inform further analysis. Here are common data exploration steps:

1. **Load the Data:**
   * Import the dataset into your chosen data analysis tool (e.g., Power BI, Python with pandas, R).
2. **Check the Dimensions of the Data:**
   * Examine the number of rows and columns in the dataset to understand its size.
3. **View the Initial Rows:**
   * Display the first few rows of the dataset to get a sense of the data's structure and the types of information it contains.
4. **Understand Column Datatypes:**
   * Review the data types of each column (e.g., numeric, categorical, date) to ensure they are assigned correctly.
5. **Check for Missing Values:**
   * Identify any missing or null values in the dataset and decide on a strategy for handling them (e.g., imputation or removal).
6. **Summary Statistics:**
   * Calculate basic summary statistics (mean, median, standard deviation, min, max) for numerical columns to understand the central tendency and variability of the data.
7. **Explore Categorical Variables:**
   * Count the unique values in categorical columns to understand the diversity of the data and identify potential issues.
8. **Visualize Data Distributions:**
   * Create histograms or box plots to visualize the distribution of numerical variables. This helps identify skewness or outliers.
9. **Identify Outliers:**
   * Use visualizations or statistical methods to detect outliers in numerical data. Decide whether to keep, transform, or remove outliers based on the context of the analysis.
10. **Explore Relationships Between Variables:**
    * Use scatter plots, correlation matrices, or pair plots to investigate relationships between numerical variables.
11. **Time Series Analysis (if applicable):**
    * If the dataset involves time-related data, explore trends, seasonality, and patterns over time.
12. **Check Data Consistency:**
    * Ensure that data values make sense and are consistent with expectations. For example, check for valid ranges in numerical columns.
13. **Address Data Quality Issues:**
    * Correct any errors or inconsistencies identified during exploration, such as typos or inaccuracies in categorical variables.
14. **Feature Engineering (if necessary):**
    * Create new variables or features that may enhance the analysis. For example, derive age from a birthdate column.
15. **Document Findings:**
    * Keep notes on important observations, issues, or decisions made during the exploration phase. This documentation is valuable for communication with stakeholders.

These steps lay the foundation for a thorough understanding of the dataset, guiding subsequent analysis and modeling efforts. Data exploration is an iterative process, and findings may lead to further refinement of the analysis plan.

**DESCRIPTIVE STATISTICS** provide a summary of the main aspects of a dataset, giving insights into its central tendency, dispersion, and shape. Common descriptive statistics include measures of central tendency, measures of dispersion, and measures of shape. Here are some key descriptive statistics:

1. **Measures of Central Tendency:**
   * **Mean (Average):** The sum of all values divided by the number of observations. It represents the center of the distribution.
   * **Median:** The middle value in a dataset when it is ordered. It is less sensitive to extreme values than the mean.
   * **Mode:** The most frequently occurring value(s) in the dataset.
2. **Measures of Dispersion:**
   * **Range:** The difference between the maximum and minimum values in the dataset.
   * **Variance:** The average of the squared differences from the mean. It measures the spread of values.
   * **Standard Deviation:** The square root of the variance. It provides a measure of the average distance of data points from the mean.
   * **Interquartile Range (IQR):** The range between the first quartile (25th percentile) and the third quartile (75th percentile). It is less sensitive to outliers than the range.
3. **Measures of Shape:**
   * **Skewness:** A measure of the asymmetry of the distribution. Positive skewness indicates a longer tail on the right, while negative skewness indicates a longer tail on the left.
   * **Kurtosis:** A measure of the "tailedness" of the distribution. High kurtosis indicates heavy tails and a sharper peak.
4. **Percentiles:**
   * **Percentiles:** Values below which a given percentage of observations fall. The median is the 50th percentile, while the quartiles (Q1 and Q3) represent the 25th and 75th percentiles, respectively.
5. **Crosstabs and Contingency Tables:**
   * **Frequency Tables:** Summarize the number of occurrences of each unique value in a categorical variable.
   * **Cross-tabulations:** Display the distribution of two or more categorical variables.
6. **Correlation:**
   * **Correlation Coefficient:** Measures the strength and direction of a linear relationship between two numerical variables.
7. **Summary Tables and Visualizations:**
   * **Histograms:** Visual representations of the distribution of numerical data.
   * **Box Plots:** Show the distribution of numerical data and highlight outliers.
   * **Summary Tables:** Provide a concise overview of key statistics for different variables.

When conducting data analysis, descriptive statistics help analysts and stakeholders understand the characteristics of the dataset, identify patterns, and make informed decisions about subsequent analyses or actions.