EDA - 1: Bike Details Dataset

1. **8B: What is the range of selling prices in the dataset?**
    - Calculate the difference between the maximum and minimum selling prices.
2. **%B: What is the median selling price for bikes in the dataset?**
    - Find the middle value of the sorted selling prices.
3. **"B: What is the most common seller type?**
    - Determine the seller type with the highest frequency.
4. **B: How many bikes have driven more than 50,000 kilometers?**
    - Count the number of bikes where the "km_driven" value exceeds 50,000.
5. **+B: What is the average km_driven value for each ownership type?**
    - Calculate the average "km_driven" for each unique "ownership" category (e.g., First Owner, Second Owner, etc.).
6. **B: What proportion of bikes are from the year 2015 or older?**
    - Count the number of bikes with a "year" value less than or equal to 2015. Divide this count by the total number of bikes.
7. **B: What is the trend of missing values across the dataset?**
    - Identify columns with missing values.
    - Calculate the percentage of missing values in each column.
    - Visualize this using a bar chart.
8. **B: What is the highest ex-showroom price recorded, and for which bike?**
    - Find the maximum value in the "ex_showroom_price" column and identify the corresponding bike model.
9. **5B: What is the total number of bikes listed by each seller type?**
    - Count the number of bikes listed by each unique "seller_type" (e.g., Dealer, Individual).
10. **8;B: What is the relationship between selling_price and km_driven for first-owner bikes?**
    - Create a scatter plot of "selling_price" vs. "km_driven" for only the first-owner bikes.
    - Analyze the plot to observe any trends or patterns.
    - Calculate the correlation coefficient to quantify the relationship.
11. **88B: Identify and remove outliers in the km_driven column using the IQR method.**
    - Calculate the Interquartile Range (IQR) of the "km_driven" column.
    - Define lower and upper bounds:

        - Lower Bound = Q1 - 1.5 * IQR
        - Upper Bound = Q3 + 1.5 * IQR
    - Identify data points outside these bounds as outliers.
    - Remove the identified outliers from the dataset.

12. **8%B: Perform a bivariate analysis to visualize the relationship between year and selling_price.**
    - Create a scatter plot of "year" vs. "selling_price."
    - Consider adding a trend line to the plot.
13. **8"B: What is the average depreciation in selling price based on the bike's age (current year - manufacturing year)?**
    - Calculate the age of each bike.
    - Analyze the relationship between age and selling price to determine the average depreciation rate per year.
14. **8B: Which bike names are priced significantly above the average price for their manufacturing year?**
    - Calculate the average selling price for each manufacturing year.
    - Identify bike models whose selling price is significantly higher than the average price for their respective manufacturing year. (Define "significantly higher" appropriately, e.g., more than 2 standard deviations).
15. **8+B: Develop a correlation matrix for numeric columns and visualize it using a heatmap.**
    - Calculate the correlation coefficient between all pairs of numeric columns.
    - Create a heatmap where color intensity represents the strength of the correlation.

EDA - 2: Car Sale Dataset

1. **ghat is the average selling price of cars for each dealer, and how does it compare across different dealers?**
    - Calculate the average selling price for each "Dealer_Name."
    - Compare the average selling prices across different dealers.
    - Visualize the distribution of average selling prices for dealers (e.g., box plot).
2. **Ghich car brand (Company) has the highest variation in prices, and what does this tell us about the pricing trends?**
    - Calculate the standard deviation of prices for each "Company."
    - Identify the brand with the highest standard deviation.
    - Analyze the price distribution within that brand.

3. **>hat is the distribution of car prices for each transmission type, and how do the interquartile ranges compare?**
    - Create separate histograms or box plots for car prices for "Manual" and "Automatic" transmissions.
    - Compare the interquartile ranges (IQR) of the two distributions.
4. **=hat is the distribution of car prices across different regions?**
    - Create a histogram or box plot of car prices for each "Dealer_Region."
5. **^hat is the distribution of cars based on body styles?**
    - Create a bar chart or pie chart to visualize the frequency of different "Body Style" categories.
6. **;hat is the average selling price of cars vary by customer gender and annual income?**
    - Calculate the average selling price for cars purchased by "Male" and "Female" customers.
    - Divide customers into income brackets (e.g., low, medium, high) and calculate the average selling price for each income bracket within each gender.
7. **9hat is the distribution of car prices by region, and how does the number of cars sold vary by region?**
    - Create a histogram or box plot of car prices for each "Dealer_Region."
    - Count the number of cars sold in each "Dealer_Region" and create a bar chart to visualize the distribution of sales across regions.
8. **Ihat is the average car price differ between cars with different engine sizes?**
    - Group cars by "Engine" type (e.g., V6, I4) and calculate the average selling price for each engine type.
9. **[hat is the distribution of car prices based on the customer's annual income bracket?**
    - Divide customers into income brackets (e.g., low, medium, high) and create a histogram or box plot of car prices for each income bracket.
10. **_Vhat are the top 5 car models with the highest number of sales, and how does their price distribution look?**
    - Identify the top 5 car models with the highest sales volume.
    - Create separate histograms or box plots for the price distributions of these top 5 models.
11. **__hat is the distribution of car price vary with engine size across different car colors, and which colors have the highest price variation?**
    - Create separate box plots of car prices for each engine size within each color category.

- Compare the interquartile ranges (IQR) of the price distributions for each color to identify colors with the highest price variation.

12. **_Ghat is any seasonal trend in car sales based on the date of sale?**
    - Extract month or quarter from the "Date" column.
    - Count the number of car sales in each month/quarter.
    - Create a line plot or bar chart to visualize any seasonal trends in sales.
13. **_>hat is the distribution of car prices change when considering different combinations of body style and transmission type?**
    - Create separate box plots of car prices for each combination of "Body Style" and "Transmission" (e.g., Sedan-Manual, Sedan-Automatic, SUV-Manual, SUV-Automatic).
14. **_=hat is the correlation between car price, engine size, and annual income of customers, and how do these features interact?**
    - Calculate the correlation coefficients between "Price," "Engine" (if you can quantify engine size numerically), and "Annual Income."
    - Create a correlation matrix or a heatmap to visualize the correlations.
    - Consider performing multivariate analysis (e.g., multiple regression) to understand the combined effect of these features on car price.
15. **_^hat is the average car price vary across different car models and engine types?**
    - Calculate the average car price for each combination of "Model" and "Engine."
    - Create a table or a heatmap to visualize these average prices.

EDA - 3: Spotify Tracks Dataset

1. **Read the dataframe, check null values if present then do the needful, check duplicate rows, if present then do the needful.**
    - Load the dataset into a pandas DataFrame.
    - Check for missing values using df.isnull().sum().
    - Handle missing values (e.g., imputation, removal).
    - Check for duplicate rows using df.duplicated().sum() and remove duplicates if necessary.
2. **What is the distribution of popularity

continue

EDA - 3: Spotify Tracks Dataset

1. **Read the dataframe, check null values if present then do the needful, check duplicate rows, if present then do the needful.**

- Load the dataset into a pandas DataFrame.
- Check for missing values using df.isnull().sum().
- Handle missing values (e.g., imputation, removal).
- Check for duplicate rows using df.duplicated().sum() and remove duplicates if necessary.

2. **What is the distribution of popularity among the tracks in the dataset? Visualize it using a histogram.**
   - Create a histogram of the "Popularity" column to visualize its distribution.
3. **Is there any relationship between the popularity and the duration of tracks? Explore this using a scatter plot.**
   - Create a scatter plot of "Popularity" vs. "Duration (ms)" to visualize any potential relationship.
4. **Which artist has the highest number of tracks in the dataset? Display the count of tracks for each artist using a countplot.**
   - Count the number of tracks for each "Artist."
   - Create a bar chart (countplot) to visualize the count of tracks for each artist.
5. **What are the top 5 least popular tracks in the dataset? Provide the artist name and track name for each.**
   - Sort the dataset by "Popularity" in ascending order.
   - Select the top 5 rows and display the "Artist" and "Track Name" for each.
6. **Among the top 5 most popular artists, which artist has the highest popularity on average? Calculate and display the average popularity for each artist.**
   - Identify the top 5 artists based on the number of tracks.
   - Calculate the average "Popularity" for each of these top 5 artists.
7. **For the top 5 most popular artists, what are their most popular tracks? List the track name for each artist.**
   - For each of the top 5 artists, find the track with the highest "Popularity."
8. **Visualize relationships between multiple numeric variables simultaneously using a pair plot.**
   - Create a pair plot using seaborn to visualize the relationships between "Popularity," "Duration (ms)," and any other relevant numeric features.
9. **Does the duration of tracks vary significantly across different artists? Explore this visually using a box plot or violin plot.**
   - Create a box plot or violin plot of "Duration (ms)" for each "Artist" to visualize the distribution of track durations across different artists.

10. **How does the distribution of track popularity vary for different artists? Visualize this using a swarm plot or a violin plot.**
    - Create a swarm plot or violin plot of "Popularity" for each "Artist" to visualize the distribution of popularity across different artists.