EDA

Assignment Questions

PW SKILLS

EDA - 1

About - Dataset:

Here is the dataset link: **Bike Details Dataset**

The dataset provided focuses on used bikes, capturing various attributes to analyze the second-hand bike market. This dataset helps understand factors influencing resale prices, usage patterns, and market trends for used bikes.

Title: Bike Details Dataset

Source:
The dataset is available on Kaggle, simulating data commonly observed in real-world online bike resale platforms.

Description:
The dataset comprises 1061 records of used bike listings. These records include details about the bike's brand and model, selling price, kilometers driven, ownership history, and other attributes. It is particularly useful for resale value prediction, market analysis, and understanding buyer preferences.

The dataset consists of 7 features, combining both numeric and categorical data, offering a comprehensive view of the second-hand bike market.

Title: Bike Details Dataset

- **=name:** Brand and model name of the bike (Categorical, e.g., "Royal Enfield Classic 350", "Honda Dio").
- **selling_price:** Listed selling price in INR (Numeric, e.g., 175000, 45000).
- **year:** Manufacturing year of the bike (Numeric, e.g., 2019, 2015).
- **seller_type:** Seller category, either "Individual" or "Dealer" (Categorical).
- **owner:** Ownership history, e.g., "1st owner", "2nd owner" (Categorical).
- **km_driven:** Total kilometers driven (Numeric, in km, e.g., 12000, 23000).
- **ex_showroom_price:** Original showroom price in INR (Numeric, e.g., 148114.0, 89643.0, etc,.).

Questions:

1. What is the range of selling prices in the dataset?
2. What is the median selling price for bikes in the dataset?
3. What is the most common seller type?
4. How many bikes have driven more than 50,000 kilometers?
5. What is the average km_driven value for each ownership type?
6. What proportion of bikes are from the year 2015 or older?
7. What is the trend of missing values across the dataset?
8. What is the highest ex_showroom_price recorded, and for which bike?
9. What is the total number of bikes listed by each seller type?
10. What is the relationship between selling_price and km_driven for first-owner bikes?
11. Identify and remove outliers in the km_driven column using the IQR method.
12. Perform a bivariate analysis to visualize the relationship between year and selling_price.
13. What is the average depreciation in selling price based on the bike's age (current year - manufacturing year)?
14. Which bike names are priced significantly above the average price for their manufacturing year?
15. Develop a correlation matrix for numeric columns and visualize it using a heatmap.

EDA - 2

About - Dataset:

Here is the dataset link: **Car Sales**
The dataset provided focuses on used car sales, capturing various attributes to analyze the second-hand car market. This dataset provides insights into factors affecting resale value, trends in the used car industry, and consumer behavior.

Title: Car Sale Dataset

Source:
The dataset is available on Kaggle, representing data commonly seen in online car resale platforms.

Description:
The dataset contains details about used cars listed for sale, such as brand, model, selling price, kilometers driven, fuel type, and transmission type. This information is valuable for predictive modeling, market analysis, and understanding customer preferences.

Features:

- Car_id: A unique identifier for each car in the dataset, helping to track individual car entries.
- Date: The date when the car sale transaction took place, formatted as YYYY-MM-DD.
- Customer Name: The name of the customer who purchased the car, represented as a string.
- Gender: The gender of the customer, categorized as "Male" or "Female."
- Annual Income: The customer's annual income in US dollars, represented as a numeric value.
- Dealer_Name: The name of the dealership selling the car, represented as a string.
- Company: The manufacturer or brand name of the car, such as "Toyota," "Ford," etc.
- Model: The specific model name of the car, such as "Corolla," "Civic," etc.
- Engine: The engine type of the car, such as "V6," "I4," etc.
- Transmission: The type of transmission in the car, either "Manual" or "Automatic."
- Color: The color of the car, represented as a string (e.g., "Red," "Blue").
- Price ($): The selling price of the car in US dollars.
- Dealer_No: A unique identifier for each car dealer in the dataset.
- Body Style: The body style of the car, such as "Sedan," "SUV," etc.
- Phone: The phone number of the customer who purchased the car.
- Dealer_Region: The geographical region of the car dealer, such as "North," "South," etc.

Questions:

1. What is the average selling price of cars for each dealer, and how does it compare across different dealers?
2. Which car brand (Company) has the highest variation in prices, and what does this tell us about the pricing trends?
3. What is the distribution of car prices for each transmission type, and how do the interquartile ranges compare?
4. What is the distribution of car prices across different regions?
5. What is the distribution of cars based on body styles?
6. How does the average selling price of cars vary by customer gender and annual income?
7. What is the distribution of car prices by region, and how does the number of cars sold vary by region?
8. How does the average car price differ between cars with different engine sizes?
9. How do car prices vary based on the customer’s annual income bracket?
10. What are the top 5 car models with the highest number of sales, and how does their price distribution look?
11. How does car price vary with engine size across different car colors, and which colors have the highest price variation?
12. Is there any seasonal trend in car sales based on the date of sale?
13. How does the car price distribution change when considering different combinations of body style and transmission type?
14. What is the correlation between car price, engine size, and annual income of customers, and how do these features interact?
15. How does the average car price vary across different car models and engine types?

EDA - 3

Amazon Sales Data

Description:
This dataset contains information on 1K+ Amazon products, including their ratings, reviews, and other details.

Features:
product_id: Unique identifier for each product
product_name: Name of the product
category: Category of the product
discounted_price: Discounted price of the product
actual_price: Actual price of the product
discount_percentage: Percentage of discount for the product
rating: Rating of the product (1-5)
rating_count: Number of people who voted for the Amazon rating
about_product: Description about the product
user_id: ID of the user who wrote the review
user_name: Name of the user who wrote the review
review_id: ID of the user review
review_title: Short review
review_content: Long review
img_link: Image link of the product
product_link: Official website link of the product

Source: **Amazon Sales**

Questions:

1. What is the average rating for each product category?
2. What are the top rating_count products by category?
3. What is the distribution of discounted prices vs. actual prices?
4. How does the average discount percentage vary across categories?
5. What are the most popular product names?
6. What are the most popular product keywords?
7. What are the most popular product reviews?
8. What is the correlation between discounted_price and rating?
9. What are the Top 5 categories based on the highest ratings?
10. Identify any potential areas for improvement or optimization based on the data analysis.

EDA - 4

Dataset Link: [Spotify Data:Popular Hip-Hop Artists and Tracks🎶]()

Description of the Dataset:
The dataset titled "Spotify Data: Popular Hip-hop Artists and Tracks" provides a curated collection of approximately 500 entries showcasing the vibrant realm of hip-hop music. These entries meticulously compile the most celebrated hip-hop tracks and artists, reflecting their significant influence on the genre's landscape. Each entry not only highlights the popularity and musical composition of the tracks but also underscores the creative prowess of the artists and their profound impact on global listeners.

Application in Data Science:
This dataset serves as a valuable resource for various data science explorations. Analysts can delve into trend analysis to discern the popularity dynamics of hit hip-hop tracks over recent years. Additionally, the dataset enables network analysis to uncover collaborative patterns among top artists, shedding light on the genre's evolving collaborative landscape. Furthermore, it facilitates the development of predictive models aimed at forecasting track popularity based on diverse features, offering insights for artists, producers, and marketers.

Column Descriptors:

Artist: The name of the artist, providing direct attribution to the creative mind behind the track.
Track Name: The title of the track, encapsulating its identity and essence.
Popularity: A numeric score reflecting the track's reception and appeal among Spotify listeners.
Duration (ms): The track's length in milliseconds, detailing the temporal extent of the musical experience.
Track ID: A unique identifier within Spotify's ecosystem, enabling direct access to the track for further exploration.

Questions:

1. Read the dataframe, check null value if present then do the needful, check duplicate row , if present then do the needful.
2. What is the distribution of popularity among the tracks in the dataset? Visualize it using a histogram.
3. Is there any relationship between the popularity and the duration of tracks? Explore this using a scatter plot.
4. Which artist has the highest number of tracks in the dataset? Display the count of tracks for each artist using a countplot.
5. What are the top 5 least popular tracks in the dataset? Provide the artist name 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.
7. For the top 5 most popular artists, what are their most popular tracks? List the track name for each artist.
8. Visualize relationships between multiple numerical variables simultaneously using a pair plot.
9. Does the duration of tracks vary significantly across different artists? Explore this visually using a box plot or violin plot.
10. How does the distribution of track popularity vary for different artists? Visualize this using a swarm plot or a violin plot.