King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT 362 Course Project
Term-2, 1445H

DATA SCIENCE PROJECT REPORT

STUDENT NAME	ID
Shdn alsheddi	443203081
Razan aldakhil	443201096
Munira Almogren	443200856
Dana Alomar	443203037

Instructor : Dr.reem alqifari

1.Problem overview

As the CEOs/Founders of a startup focused on adoption of electric vehicles (EV) to the Saudi market aligning with Saudi Arabia vision 2030 to accelerate the transition to sustainable transportation, We need to understand the factors influencing consumer preferences, adoption rates, and overall market trends for electric vehicles ,but to do so effectively, we must base our strategies on solid data analysis and market insights. Addressing the five key questions derived from the data collected can help us overcome challenges and make informed decisions.

2.Data Source

Source: Electric Vehicle Data API Documentation (adrienpelletierlaroche) | RapidAPI

ElectricVehicleData API (electric-vehicle-data-documentation.netlify.app)

The Electric vehicle data API, available on RapidAPI, provides essential details on electric vehicles, such as VIN, location, model specifics, electric range, and pricing. RapidAPI is a platform marketplace for APIs where developers can find, connect to, and manage thousands of APIs, facilitating the integration of data and services into applications. This API is particularly useful for stakeholders in the electric vehicle sector for strategic planning and operational management.

3.About data:

The Electric Vehicle Data API offers comprehensive data on electric vehicles, which would be essential for our dataset. We will use this API to collect specific details for each electric vehicle.

Each row in the dataset will represent an individual electric vehicle, with columns detailing its characteristics such as :

- **County:** The county in which the vehicle is registered. This can provide insights into regional distribution and adoption rates of electric vehicles.
- **City:** The city of the vehicle's registration. Like county, it offers a more granular view of EV distribution.
- **model year:** The year the vehicle model was manufactured. This can be crucial for understanding trends in EV technology and adoption over time.
- **Make:** The manufacturer or brand of the vehicle. Analyzing this can show which car manufacturers are leading in the EV market.
- **Model:** The specific model of the vehicle. This information, combined with make, helps in understanding consumer preferences and the performance of specific EV models.
- **Electric range:** The maximum distance the vehicle can travel on a single charge. This metric is key for assessing the usability and technological advancement of EVs.
- **Electrical vehicle type:** Specifies the type of electric vehicle, such as Battery Electric Vehicle (BEV) or Plug-in Hybrid Electric Vehicle (PHEV), offering insights into the types of EVs preferred in different regions.
- **Electric utility:**The electric utility provider(s) for the vehicle's charging needs. This can indicate the availability and possibly the type of charging infrastructure.

4.Dataset Overview

The majority of the data in the dataset consists of qualitative attributes, including city, county, model year, make, model, electric vehicle type, and electric utility. However, there is one quantitative attribute, which is the electric range. This attribute allows for measurement and comparison on a ratio scale.

5. Questions

By exploring the available dataset, we can answer the following questions:

1.Which country has the highest adoption of electric vehicles?

2.What are the trends in EV adoption over time?

3.How does EV adoption vary by geographic location?

4.Which manufacturers and models are most popular among electric vehicle owners?

5.What is the relationship between electric vehicle range and vehicle type?

6.Can we build a model that could classify vehicles based on their electric range?

6.Bias and Fairness

After all we have noticed a few favoritisms in our data collection process that can lead to inaccurate or discriminatory outcomes.

for instance, in the dataset there is huge Adoption for tesla model among all models. This bias can be attributed to factors such as the popularity of Tesla vehicles among early adopters of EV technology and the availability of charging infrastructure in certain areas.

Also, the dataset is heavily targeted in Washington nation, in towns like Seattle. This geographic bias may not accurately represent the EV adoption patterns and utility providers in other regions or states.

Additionally, the dataset shows a strong association with specific utility providers, such as Puget Sound Energy Inc, City of Tacoma, Bonneville Power Administration, and others. This bias may reflect the availability of incentives, charging infrastructure, or partnerships with these utility providers, which could influence EV adoption rates.

on the other hand, Fairness is the principle of treating all individuals fairly and without discrimination, addressing bias and fairness aims to identify and address biases in data to ensure equitable and unbiased outcomes, and since we have identified above a few numbers of favoritism, this Bias can cause to implications for the fairness and reliability, this May skew the analysis and restriction the generalizability of the findings. The conclusions drawn from this dataset maybe won't accurately mirror the broader EV adoption trends or offer a comprehensive information of the elements influencing EV adoption.

so, we suggest the following to mitigate biases in future data collection and analysis efforts:

collect data from a wider range of geographic areas, to capture a more representative sample of EV adoption patterns. Include a variety of EV makes and models to account for differences in consumer preferences and market availability.

Collect data on socioeconomic factors such as income levels, education, and access to charging infrastructure. This information can help identify and account for disparities in EV adoption rates among different demographic groups, ensuring a more equitable analysis.

implement random sampling techniques to ensure an extra balanced representation of EV proprietors and their characteristics. This can help reduce the influence of self-selection bias and capture a more diverse sample.

7. Data Processing and Cleaning summary:

1. **Data Typing Adjustments**:

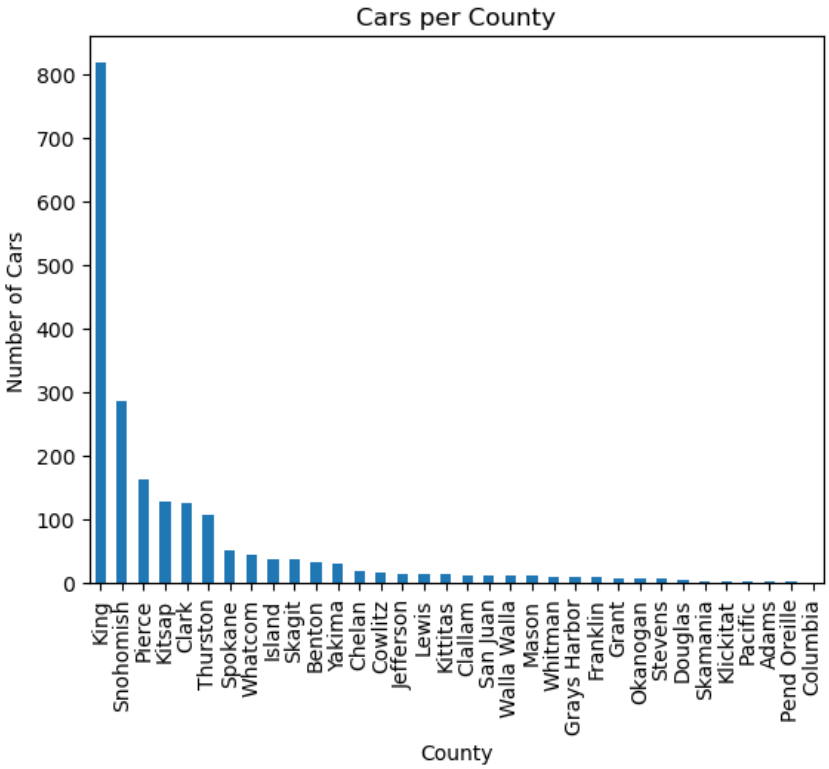
- Changed the data type of the model year from integers to categorical, acknowledging that these values represent specific calendar years and are not used in numerical calculations.

2. **Handling Missing Electric Range Values**:
    - Identified a high standard deviation in the electric range, indicating wide variability. The average electric range was calculated to be approximately 83.92 miles, suggesting vehicles can travel around 84 miles on a single charge on average.
    - Noted a significant number of vehicles (1243) with an electric range value of 0, indicating potential missing data or entry errors. To mitigate this, missing or zero electric range values were imputed with the average values based on vehicle model and type.
    - After imputation, 945 vehicles had their electric range updated, but 298 vehicles still had an electric range of 0. These were considered for further action due to potential reasons like shared model and type attributes with zero range or uniqueness in model/type leading to the inability to impute.
    - Decided to drop the remaining 298 vehicles with an electric range of 0, classifying them as outliers or cases of insufficient data for imputation.
3. **Imputing Missing 'ElectricUtility' Values**:
    - Addressed missing values in the 'electricUtility' column by imputing based on the most common utility for the corresponding city and county. This approach acknowledges the geographic association of utility data.
    - Specific checks were performed to identify patterns or singularities in missing utility data, leading to dropping rows where imputation was not feasible due to lack of sufficient data (e.g., only one row with missing data for a given city/county combination).
4. **Removing Duplicates**:
    - Identified and removed 1013 duplicate entries to ensure data uniqueness and integrity.
5. **Discretizing Electric Range for Classification**:
    - To support the development of a classifier for vehicle electric range, we discretized the electric range into three categories: Short Range (less than 100 miles), Medium Range (100 to 200 miles), and Long Range (greater than 200 miles). This step transformed the problem from regression to classification, facilitating the analysis of vehicles based on their electric range capabilities.

8. EDA:

By applying statistical measures on the dataset, we used the mean, median ,range on the electric range column. these metrics are useful in understanding the overall distribution of electric ranges in the dataset, and the large standard deviation relative to the mean implies substantial variability among the electric ranges of the vehicles studied. We see a variety of categorical variables that help us understand the distribution and characteristics of electric vehicles. 'King' is the most frequently occurring county, which may indicate a higher adoption rate or availability of electric vehicles in that region. Similarly, 'Seattle' tops the city category, which could suggest urban centers are more likely to embrace electric vehicle technology. The dataset also reveals that the most common model year is '2022', indicating that the data might be quite recent. 'TESLA' and 'MODEL 3' appearing as the most common make and model, respectively, highlight the market dominance of this particular manufacturer and model in the electric vehicle sector. When it comes to the type of electric vehicle, 'Battery Electric Vehicle (BEV)' is by far the most common, which aligns with global trends towards fully

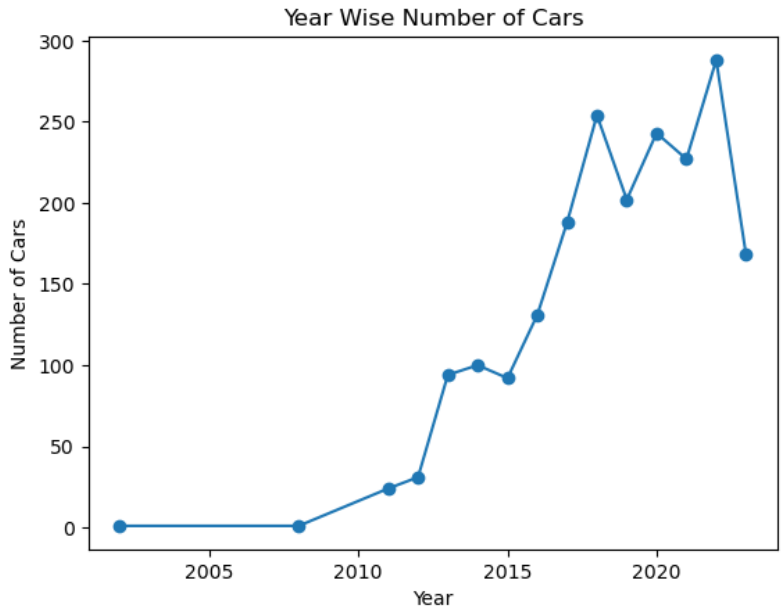
1.Which County has the highest adoption of electric vehicles?



Answer 1:

King county has the highest number of electric cars registered with 818 cars, followed by Snohomish and Pierce County.

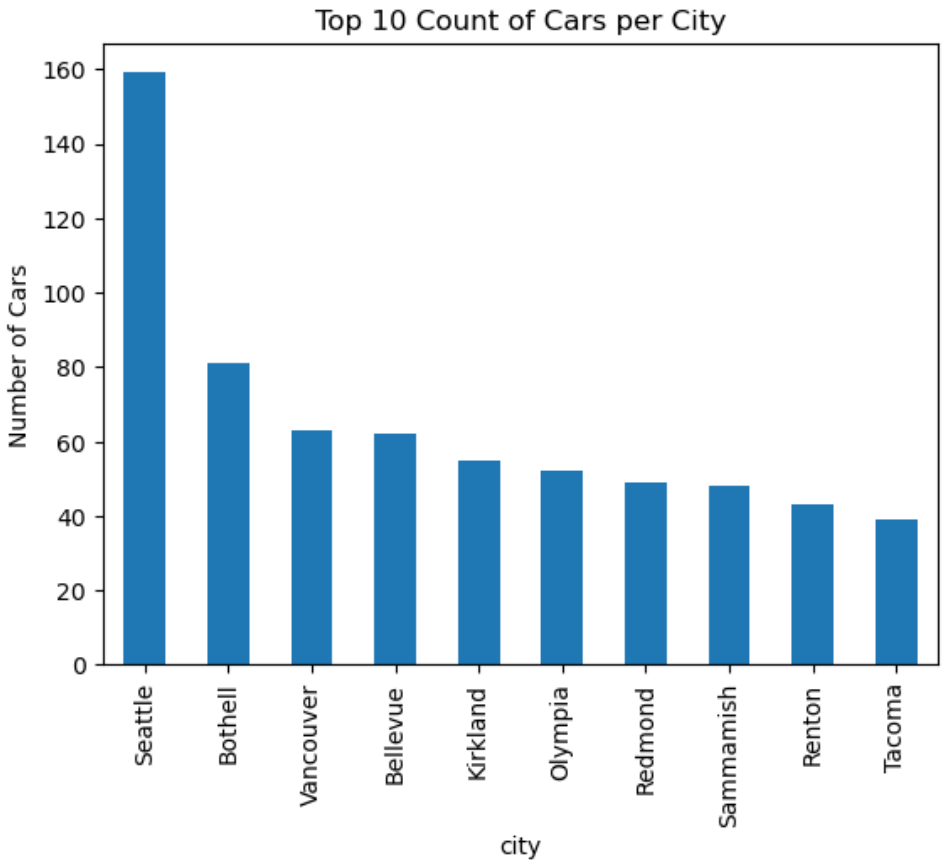
2.What are the trends in EV adoption over time?



Answer 2:

The market's trust in electric vehicles began to grow after 2010, and the demand for electric vehicles has been consistently increasing ever since. However, in 2019, there was a noticeable decline in demand. This drop was primarily caused by the lockdowns imposed due to the COVID-19 pandemic.

3.How does EV adoption vary by geographic location?

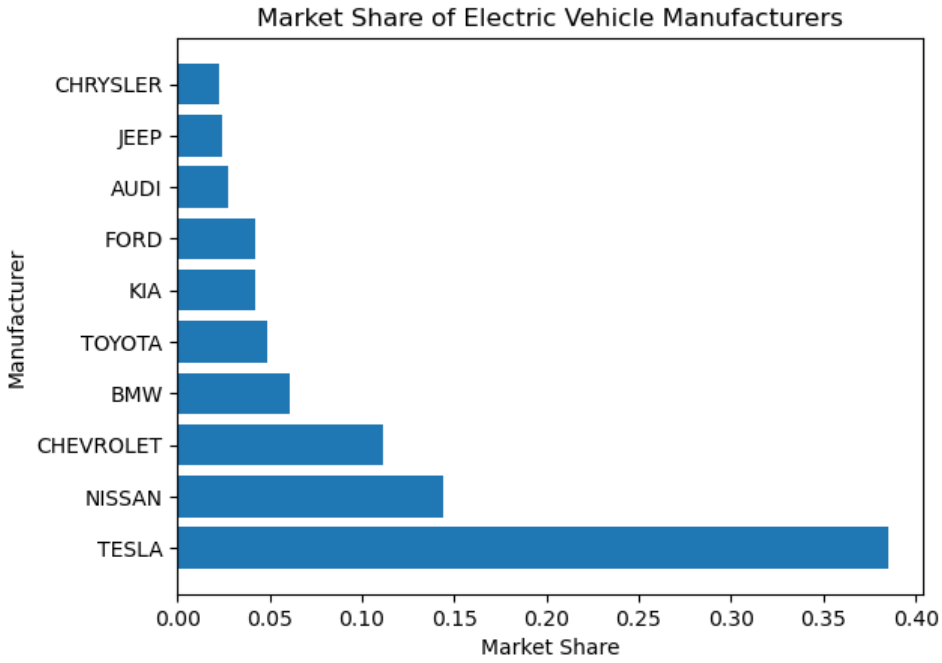


Answer 3:

EV adoption varies by geographic location within the region. Larger cities generally exhibit higher numbers of EVs, but smaller cities are also showcasing a growing interest in electric vehicles.

It's important to note that the factors driving EV adoption can vary between cities. The availability of EV charging stations, financial incentives, local policies promoting sustainable transportation, and community initiatives all play a role in influencing adoption rates in different locations.

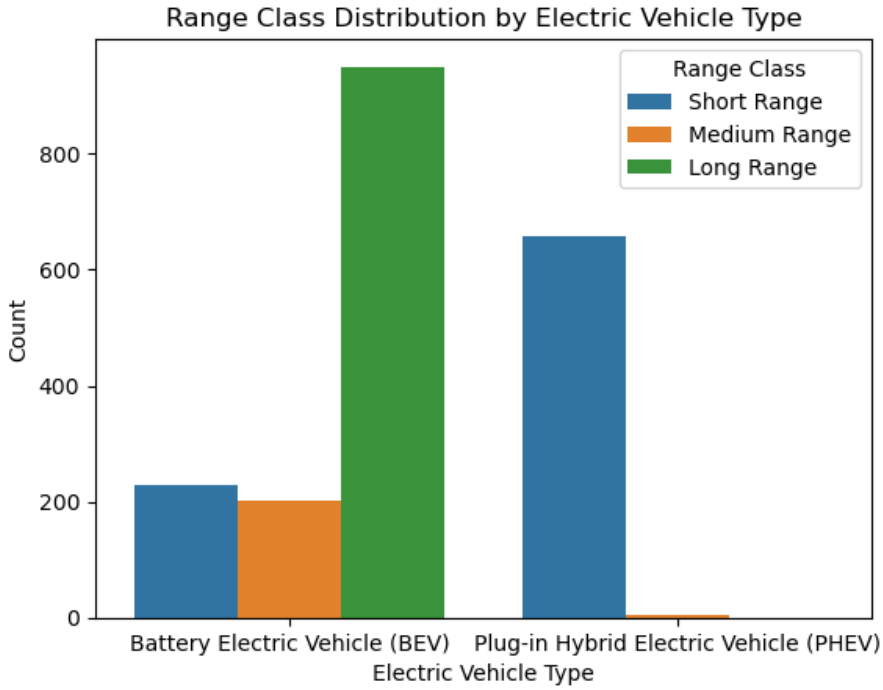
4.Which manufacturers and models are most popular among electric vehicle owners?



Answer 4:

TESLA is dominating the electric vehicle market, along with the presence of other major players like NISSAN and CHEVROLET. Additionally, the inclusion of BMW as a luxury electric vehicle manufacturer adds diversity to the market, catering to consumers who value both sustainability and luxury.

5.What is the relationship between electric vehicle range and vehicle type?



Answer 5:

The relationship between electric vehicle range and vehicle type demonstrates that BEVs generally have a wider range of options, including vehicles with short, medium, and long ranges. On the other hand, PHEVs tend to have a stronger presence in the short-range category. This relationship reflects the different design and usage considerations for these two types of electric vehicles.

Tools and libraries:

Python offers a rich collection of tools and libraries that enable users to create insightful and visually appealing visualizations.

1. matplotlib.pyplot:

matplotlib.pyplot is a widely used plotting library in Python. It provides a MATLAB-like interface for creating a variety of static, animated, and interactive plots.

- plt.plot: This function was used to create line plots. It is commonly used to visualize trends or patterns over a continuous variable, such as the year-wise distribution of cars.
- plot(kind='bar'): was used to create a bar plot showing the count of cars per county,Top 10 Count of Cars per City, And the top 20 counties with top 10 consumed make, Electric Utility Distribution in top 20 cities with highest number of cars

2. seaborn:

`seaborn` is a statistical data visualization library built on top of `matplotlib`. It offers a high-level interface for creating attractive and informative statistical graphics.

- `sns.countplot`: This function creates a bar plot showing the count of occurrences of each category in a categorical variable. It is useful for visualizing the distribution of categorical data, such as the count of different range classes of electric vehicles.
- `sns.heatmap` : was used to create the heatmap that visually represents the relationship between two categorical variables.
- `sns.lineplot()`: creates a line plot that visualizes the range class by model year.

3. plotly.express:

`plotly.express` is a high-level visualization library based on `plotly.py`. It provides a concise and expressive syntax for creating interactive visualizations.

- `px.pie`: This function creates a pie chart, which represents the proportion of different categories in a whole. In the code, it is used to display the distribution of electric vehicle types based on their average range.

4. Pandas:

Pandas is a powerful data manipulation and analysis library in Python. It provides data structures and functions to efficiently work with structured data.

- `pd.crosstab`: This function is part of the panda's library and is used to compute a cross-tabulation table. It calculates the frequency or count of occurrences of combinations between categorical variables.

9. Future Steps

For the upcoming steps, we're going to build a classification model to help predict the vehicle electric range based on its features. We've noticed a lot of cars fall under the Long Range category, and we want to understand if this trend is driven by consumer demand or industry patterns. The Tesla Model 3 is a common sight in our dataset, so we plan to investigate the reasons behind its popularity. Gathering more data on daily driving habits could help us see how they relate to the car's range. Finally, we aim to develop a straightforward model to classify cars into range categories, which could be a big help for car manufacturers and policymakers.