Fuel Consumption Data Analysis

Northwest Missouri State University Maryville, MO

November 20, 2023

Section: 44517-01
Team Name: EnergySync-Explorers
Team Members:
1) Devineni Vasavi (S554779)
2) Venkata Rayudu Adapa (S555576)
3) Pathuri Tejeswara Rao (S555054)
4) Prasanna Boddapati (S554972)
5) Sai Tejaswini Yarapathineni (S554224)

Introduction

The Fuel Consumption dataset contains data about Fuel Consumption for various vehicles from 2000 to 2022. This data will be analyzed using PySpark and visualized using Tableu to gain insights into the patterns and trends of fuel consumption. In this project, we will explore various techniques for analyzing and visualizing the data to understanding fuel consumption patterns, factors affecting it, and emissions associated with different vehicles for making informed decisions towards achieving sustainability and efficiency in the automotive industry.

Metadata

The Fuel Consumption dataset encompasses vehicle information from 2000 to 2022, featuring columns such as YEAR, MAKE, MODEL, VEHICLE CLASS, ENGINE SIZE, CYLINDERS, TRANSMISSION, FUEL, FUEL CONSUMPTION, HWY (L/100 km), COMB (L/100 km), COMB (mpg), and EMISSIONS. These details provide a comprehensive overview of fuel efficiency, engine specifications, and emissions for a diverse range of vehicles. The dataset includes the following variables:

- Year: Records the manufacturing year of the vehicles (ranging from 2000 to 2022).
- Make: Specifies the brand or manufacturer of the vehicles.

- Model: Identifies the specific model or variant of each vehicle.
- Vehicle Class: Categorizes the vehicles into classes based on their characteristics or purpose.
- Engine Size: Indicates the volume of the engine in cubic centimeters.
- Cylinders: Represents the number of cylinders in the vehicle's engine.
- Transmission: Describes the type of transmission system used (e.g., automatic, manual).
- Fuel: Specifies the type of fuel the vehicle utilizes (e.g., gasoline, diesel).
- Fuel Consumption: Provides data on the overall fuel consumption of the vehicle.
- HWY (L/100 km): Indicates the highway fuel consumption in liters per 100 kilometers.
- COMB (L/100 km): Represents the combined (city and highway) fuel consumption in liters per 100 kilometers.
- COMB (mpg): Expresses the combined fuel consumption in miles per gallon.
- Emissions: Records the emissions produced by the vehicle, reflecting its environmental impact.

The Fuel Consumption dataset is utilized to analyze and assess the fuel efficiency, engine specifications, and emissions of vehicles spanning the years 2000 to 2022, providing insights into the environmental impact and performance characteristics of diverse automotive models.

Tools and Technologies:

For this project, we will leverage the following technologies and tools:

1. Jupyter Notebook
2. pySpark
3. Pandas
4. Tableau

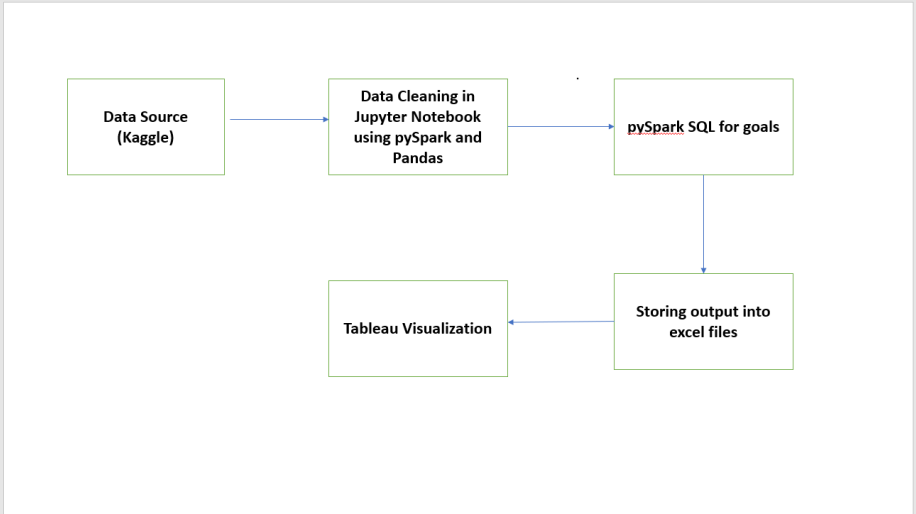

Figure 1: Architecture Diagram

Architecture Summary:

1. **Data Source:** We choose Kaggle as our data source to download Fuel Consumption Trends (2000-2022) data set. The downloaded data source is in csv format with almost 22546 rows and 13 columns namely YEAR, MAKE, MODEL, VEHICLE CLASS, ENGINE SIZE, CYLINDERS, TRANSMISSION, FUEL, FUEL CONSUMPTION, HWY (L/100 km), COMB (L/100 km), COMB (mpg), and EMISSIONS.

2. **Data Cleaning in Jupyter Notebook:** We then loaded the above dataset into PySpark Data Frame. Then to perform data cleaning and preprocessing we used PySpark and Pandas to handle missing values and other issues to make the data ready for analysis.

3. **pySpark SQL Goals:** Then for the goals we want to analyze for the current data set chosen, we have chosen pySpark SQL to get those outputs for the queries related to goals.

4. **Storing output into excel files:** Using pandas we then stored output of each SQL query into an csv file for using it as input for visualization in tableau.

5. **Tableau Visualization:** Next we plan to connect Tableau to csv files where we have the data from the goals to create visualizations and reports for getting our analysis for the goals chosen.

Explanation Of Implementation Steps

Step 1: Importing and Preprocessing the Data

The first step is to import the data into a PySpark dataframe and preprocess it to ensure that it is in a suitable format for analysis. This involves cleaning and transforming the data to ensure that it is consistent, complete, and ready for analysis. We will be leveraging Jupyter Notebook, some of the preprocessing steps include:

- **Handling missing values:** There may be missing values in the data, which can be imputed using various techniques such as mean, median, or mode imputation.
- **Data transformation:** The data may need to be transformed to ensure that it is in the right format for analysis such as Pandas data frame or pySpark data frame.
- **Data normalization:** The data may need to be normalized to ensure that it is on a consistent scale and can be compared across different households or time periods.

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

spark = SparkSession.builder.appName("Fuel_Consumption_2000-2022").getOrCreate()

df = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("Fuel_Consumption_2000-2022.csv")
df.describe().show()

class_mapping = {
    'SUV:Small': 'SUV - SMALL',
    'Station wagon: Mid-size': 'STATION WAGON - MID-SIZE',
    'Station wagon: Small': 'STATION WAGON - SMALL',
    'Van: Passenger': 'VAN - PASSENGER',
    'Pickup truck: Small': 'PICKUP TRUCK - SMALL',
    'Pickup truck: Standard': 'PICKUP TRUCK - STANDARD',
    'SUV: Standard': 'SUV - STANDARD'
}

@udf(StringType())
def standardize_vehicle_class(vehicle_class):
    return class_mapping.get(vehicle_class, vehicle_class)

df_clean = df.withColumn('Vehicle Class', standardize_vehicle_class(df['Vehicle Class']))
```

```
24
25 panda_df = df.toPandas()
26 panda_df.to_csv("Cleaned_Fuel_Consumption_2000-2022.csv")
```

Listing 1: Examining Fuel Efficiency Trends 2000-2022

We started with downloading the dataset from Kaggle and then we established a pySpark session to engage with the Spark engine, and proceed to ingest the data set into a PySpark DataFrame. The code performs data loading, exploration, transformation, and exports the cleaned data into csv format for further analysis.

Goals:

Goal 1: Analyze how fuel efficiency (HWY and COMB) has evolved over time (2000-2022) to identify long-term trends and fluctuations.

Examining the Fuel Consumption dataset from 2000 to 2022 reveals trends in fuel efficiency, specifically highway (HWY) and combined (COMB) consumption, allowing for a comprehensive understanding of how these metrics have evolved over time. This analysis aids in identifying long-term patterns, fluctuations, and potential factors influencing the changing landscape of vehicle fuel efficiency across the years.

Code Snippet:

```
1  filtered_data = df_clean.filter((df_clean['YEAR'] >= 2000) &
       (df_clean['YEAR'] <= 2022))
2  fuel_efficiency_by_year = filtered_data.groupBy('YEAR').agg(
3      round(avg('HWY (L/100 km)'), 2).alias('
       avg_HWY_efficiency'),
4      round(avg('COMB (L/100 km)'), 2).alias('
       avg_COMB_efficiency')
5  )
6  fuel_efficiency_by_year = fuel_efficiency_by_year.orderBy('
       YEAR')
7  fuel_efficiency_by_year.show(fuel_efficiency_by_year.count()
       , truncate=False)
8
9  pandas_df_year = fuel_efficiency_by_year.toPandas()
10 local_path = "Goal1_fuel_efficiency_by_year.csv"
11 pandas_df_year.to_csv(local_path, index=False)
```

Listing 2: Fuel efficiency by year

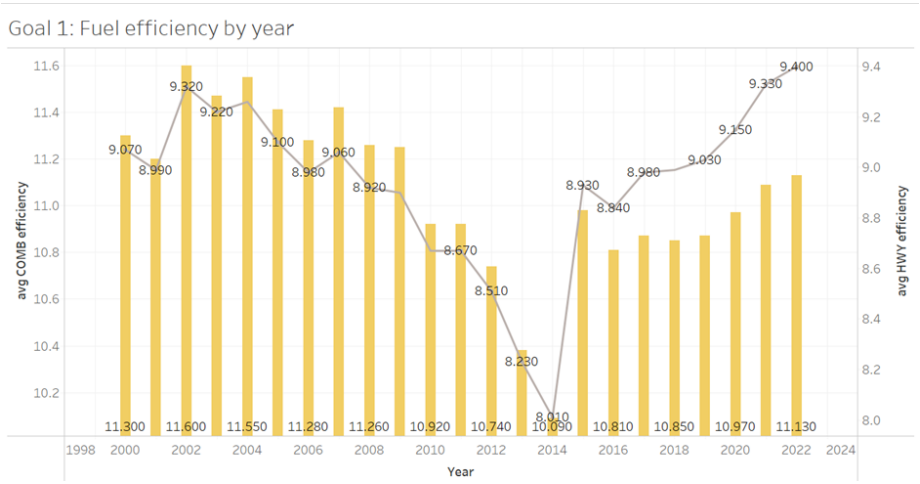

Figure 2: Fuel efficiency by year

Observations:

- From above graph bar chart allows for a clear comparison of average COMB efficiency for each individual year where 2002 has highest average COMB efficiency (11.600) and in 2014 has lowest average COMB efficiency (10.090).
- The Line chart allows for a clear comparison of average HWY efficiency for each individual year where 2022 has Highest average HWY efficiency (9.400) and in 2014 has lowest average HWY efficiency (8.010).
- Both charts can help identify fluctuations and variability in fuel efficiency. Peaks and troughs may correspond to economic conditions, fuel price changes, or shifts in consumer preferences.
- From the graph Initially, from 2000 to 2014 both average highway (avg HWY efficiency) and combined (avg COMB efficiency) fuel efficiencies showed a gradual decrease, indicating a potential period of lower fuel efficiency during this timeframe.
- From 2014 onwards, there is a noticeable reversal in the trend, with both avg HWY efficiency and avg COMB efficiency experiencing an upward trajectory. Avg HWY efficiency increased very high compared to avg COMB efficiency.

Goal 2: Investigate how fuel efficiency varies among different vehicle makes and models to pinpoint high-performing and under performing vehicles.

By exploring the Fuel Consumption data set, researchers can discern variations in fuel efficiency across different vehicle makes and models. This investigation aims to identify high-performing vehicles that exhibit optimal fuel consumption, as well as pinpoint under performing ones with lower fuel efficiency. Such insights contribute to informed decision-making for consumers, policymakers, and the automotive industry, fostering improvements in environmental sustainability and economic considerations.

Code Snippet:

```
# 2. Investigate how fuel efficiency varies among different vehicle makes and models to pinpoint high-performing and underperforming vehicles.
fuel_efficiency_by_make_model = df_clean.groupBy('MAKE', 'MODEL').agg(
    round(avg('COMB (mpg)'), 2).alias('avg_COMB_mpg')
)
top_high_performing_vehicles = fuel_efficiency_by_make_model.orderBy('avg_COMB_mpg', ascending=False).limit(10)
top_low_performing_vehicles = fuel_efficiency_by_make_model.orderBy('avg_COMB_mpg').limit(10)
top_high_performing_vehicles.show()
top_low_performing_vehicles.show()

pandas_top_high_performing_vehicles = top_high_performing_vehicles.toPandas()
local_path = "Goal2_top_high_performing_vehicles.csv"
pandas_top_high_performing_vehicles.to_csv(local_path, index=False)
pandas_top_low_performing_vehicles = top_low_performing_vehicles.toPandas()
local_path = "Goal2_top_low_performing_vehicles.csv"
pandas_top_low_performing_vehicles.to_csv(local_path, index=False)
```

Listing 3: Top high low performing vehicles

Observations:

- The first stacked bar chart visually compares the average combined miles per gallon (avg COMB mpg) for different car models within each make (Top performing vehicles).

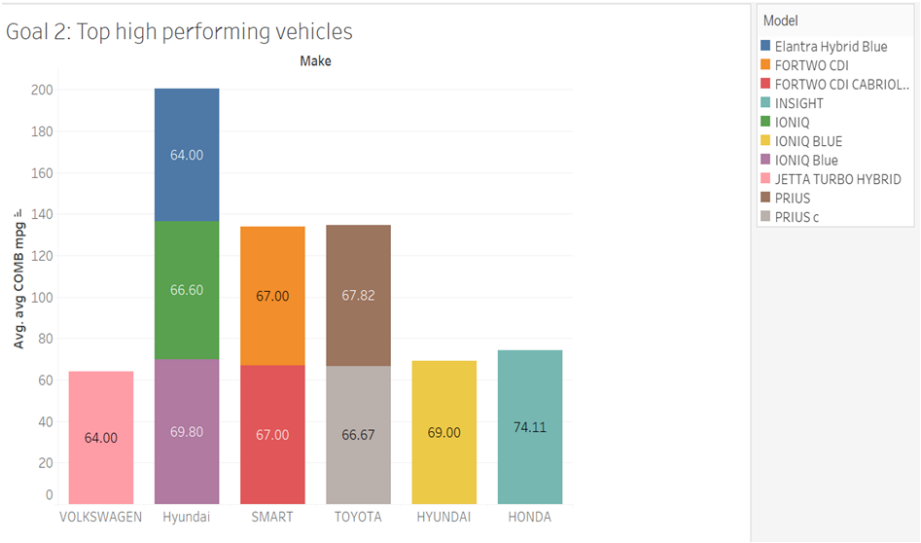

Figure 3: Top High Performing Vehicles

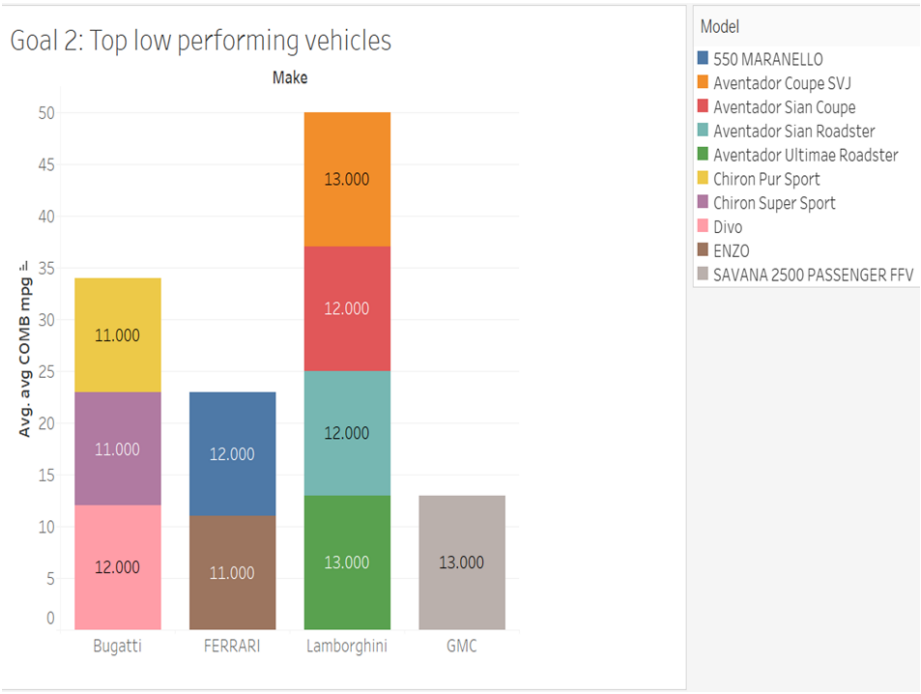

Figure 4: Top Low Performing Vehicles

- Within each make, the tallest segment in the stacked bar represents the model with the highest average combined mpg.
- Among the top 10 high-performing vehicles, the INSIGHT model from the HONDA maker has the highest average combined miles per gallon (avg-COMB-mpg). In contrast, the Elantra Hybrid Blue model from Hyundai and the JETTA TURBO HYBRID model from Volkswagen have the lowest average combined miles per gallon.
- The second stacked bar chart visually compares the average combined miles per gallon (avg COMB mpg) for different car models within each make (lowest performing vehicles).
- From Fig 03, among the bottom 10 performers in terms of fuel efficiency, the SAVANA 2500 PASSENGER FFV model from GMC, Aventador Coupe SVJ and Aventador Ultimae Roadster both the models from Lamborghini has the highest average combined miles per gallon (avg-COMB-mpg). In contrast, Chiron Pur Sport, Chiron Super Sport both models from Bugatti and ENZO model from FERRARI exhibit the lowest average combined miles per gallon.

Goal 3: Assess how engine size, cylinders, and other specifications impact fuel consumption to optimize vehicle design for efficiency.

By scrutinizing the Fuel Consumption data set, we can assess how variations in engine size and cylinder count, along with other specifications, influence the overall fuel consumption of vehicles. This analysis enables the optimization of vehicle design to enhance fuel efficiency, guiding manufacturers toward more resourceful and environmentally friendly engineering solutions.

Code Snippet:

```
df_clean = df_clean.withColumn('YEAR', df_clean['YEAR'].cast
    ('int'))
trend_analysis_by_year = df_clean.groupBy('YEAR', 'ENGINE
    SIZE').agg(
     round(avg('FUEL CONSUMPTION'), 2).alias('
    avg_fuel_consumption')
)
trend_analysis_by_year.count()

pandas_trend_analysis_by_year = trend_analysis_by_year.
    toPandas()
local_path = "Goal3_trend_analysis_by_year.csv"
pandas_trend_analysis_by_year.to_csv(local_path, index=False
    )
```

Listing 4: Trend analysis by year

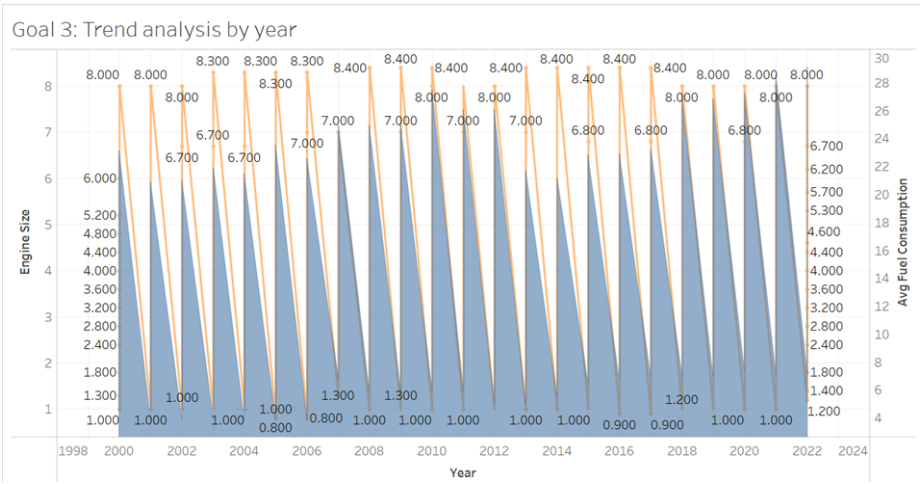

Figure 5: Trend analysis by year

Observations:

- In the above figure, there is a noticeable correlation between engine size and fuel consumption. Generally, larger engine sizes tend to result in higher fuel consumption.
- Vehicles with smaller engine sizes, such as 1.0 to 1.5, show lower average fuel consumption, while larger engine sizes, like 6.0 and above, tend to have higher fuel consumption.
- It's important to note that there are exceptions, and other factors such as cylinder configuration and technological advancements can influence fuel efficiency.

Goal 4: Determine how different types of transmissions (e.g., automatic, manual) influence fuel efficiency to guide transmission selection in vehicle design.

By analyzing the Fuel Consumption data set, we can discern the impact of various transmissions (automatic, manual) on a vehicle's fuel efficiency. This information aids in informed decision-making during the vehicle design process, enabling designers to optimize transmission choices for better overall fuel economy and environmental performance.

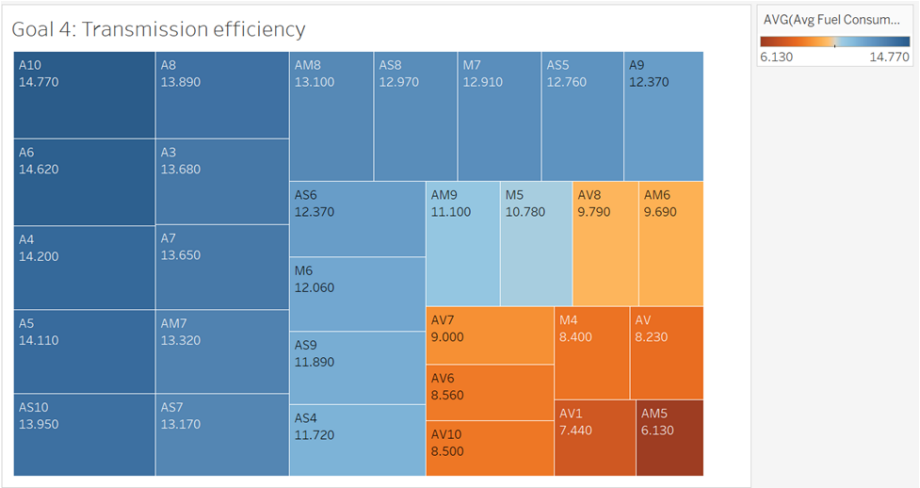

Figure 6: Transmission efficiency

Code Snippet:

```
transmission_efficiency = df_clean.groupBy('TRANSMISSION').
    agg(
     round(avg('FUEL CONSUMPTION'), 2).alias('
    avg_fuel_consumption')
)

# Show the results to compare fuel efficiency by
    transmission type
transmission_efficiency.show(transmission_efficiency.count()
    , truncate=False)

pandas_transmission_efficiency = transmission_efficiency.
    toPandas()
local_path = "Goal4_transmission_efficiency.csv"
pandas_transmission_efficiency.to_csv(local_path, index=
    False)
```

Listing 5: Transmission Impact on Efficiency

Observations:

- The above figure represent the relationship between different types of transmissions and their influence on fuel efficiency in tree map into sections, each representing a different type of transmission (e.g., AS4, AS10, A9, etc.).

- The size of each section corresponds to the frequency or prevalence of that particular transmission type in the data set.
- Specifically, vehicles with the AM5 and M5 transmissions have the lowest fuel consumption values, indicating higher fuel efficiency. Conversely, vehicles with AV10 and A10 transmissions tend to have higher fuel consumption, suggesting lower fuel efficiency in certain automatic transmission types.

Goal 5: To analyze how different fuel types (e.g., gasoline, diesel, electric) affect fuel consumption and identify trends in the popularity and efficiency of different fuel sources.

The data set is employed to examine the impact of various fuel types (gasoline, diesel, electric) on fuel consumption, facilitating a comprehensive analysis of their efficiency. Through this analysis, trends in the popularity and effectiveness of different fuel sources can be identified, offering valuable insights into the evolving landscape of automotive energy consumption.

Code Snippet:

```
# 5. To analyze how different fuel types (e.g., gasoline, diesel, electric) affect fuel consumption and identify trends in the popularity and efficiency of different fuel sources.
selected_df = df_clean.select('FUEL', 'COMB (L/100 km)')
fuel_consumption_by_fuel_type = selected_df.groupBy('FUEL').agg(
    round(avg('COMB (L/100 km)'), 2).alias('avg_fuel_consumption')
)
fuel_consumption_by_fuel_type = fuel_consumption_by_fuel_type.orderBy('avg_fuel_consumption', ascending=True)
fuel_consumption_by_fuel_type.show()

pandas_fuel_consumption_by_fuel_type = fuel_consumption_by_fuel_type.toPandas()
local_path = "Goal5_fuel_consumption_by_fuel_type.csv"
pandas_fuel_consumption_by_fuel_type.to_csv(local_path, index=False)
```

Listing 6: Fuel consumption by fuel type

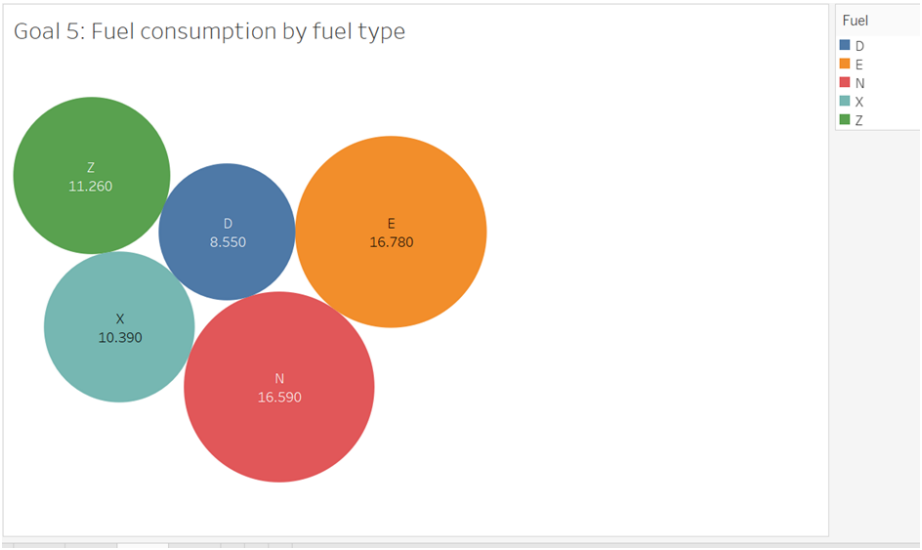

Figure 7: Fuel consumption by fuel type

Observations:

- From the above figure, the analysis of fuel types highlights notable variations in average fuel consumption. Fuel type D stands out with the lowest consumption at 8.55, followed by X and Z. Conversely, fuels N and E exhibit higher average fuel consumption, suggesting potentially lower fuel efficiency.
- The packed bubbles map visually represents these differences, with larger bubbles representing higher fuel consumption and smaller bubbles indicating lower consumption, offering a clear overview for decision-making in fuel selection.

Goal 6: Analyze whether vehicles with more cylinders tend to consume more fuel.

Analyzing the Fuel Consumption data set reveals a positive correlation between the number of cylinders in a vehicle's engine and its fuel consumption. Vehicles with a higher cylinder count generally exhibit increased fuel consumption, suggesting a connection between engine configuration and efficiency. This insight can inform decisions related to fuel efficiency standards and vehicle design.

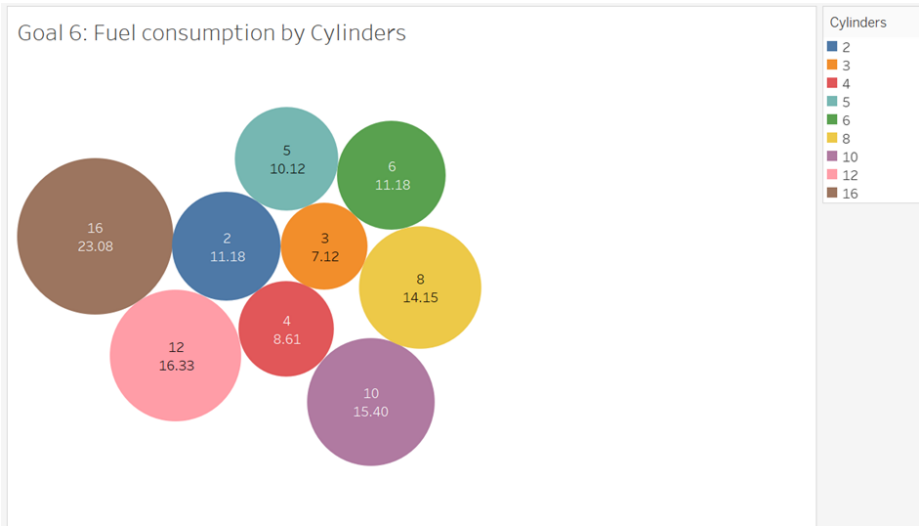

Figure 8: Fuel consumption by Cylinders

Code Snippet:

```
# 6. Analyze whether vehicles with more cylinders tend to consume more fuel.
selected_df = df_clean.select('ENGINE SIZE', 'CYLINDERS', 'COMB (L/100 km)')
fuel_consumption_by_cylinders = selected_df.groupBy('CYLINDERS').agg(
    round(avg('COMB (L/100 km)'), 2).alias('avg_fuel_consumption')
)
fuel_consumption_by_cylinders = fuel_consumption_by_cylinders.orderBy('avg_fuel_consumption', ascending=True)
fuel_consumption_by_cylinders.show()

pandas_fuel_consumption_by_cylinders = fuel_consumption_by_cylinders.toPandas()
local_path = "Goal6_fuel_consumption_by_cylinders.csv"
pandas_fuel_consumption_by_cylinders.to_csv(local_path, index=False)
```

Listing 7: Fuel consumption by Cylinders

Observations:

- From the above figure, the analysis of fuel consumption by cylinders, a clear trend emerges, indicating a positive correlation between the number

of cylinders and average fuel consumption.

- Vehicles with fewer cylinders, such as 3 and 4, exhibit lower fuel consumption (7.12 and 8.61, respectively), while those with more cylinders, such as 8, 10, and 12, show higher fuel consumption, reaching 14.15, 15.4, and 16.33, respectively.
- The highest average fuel consumption is observed in vehicles with 16 cylinders, at 23.08.

Discussions around Relevant Metrics

Data Quality check on the Data set

Conducting thorough data quality checks on the fuel consumption data set for robust analysis and reliable insights. However, like any other data set, it is important to evaluate its quality before using it for analysis.
Here are some aspects of data quality for this data set:
For the Fuel Consumption Data Analysis data set, a thorough Data Quality Check involves several key steps to ensure the reliability and usability of the information.

1. **Completeness:** We verified the dataset's completeness by assessing whether all expected columns, including YEAR, MAKE, MODEL, VEHICLE CLASS, ENGINE SIZE, CYLINDERS, TRANSMISSION, FUEL, FUEL CONSUMPTION, HWY (L/100 km), COMB (L/100 km), COMB (mpg), and EMISSIONS, contain non-null values. Also addressed any missing data to maintain a comprehensive dataset for accurate analysis.

2. **Accuracy:** We checked the accuracy of the data by cross-referencing entries against established standards and specifications. Specifically, ensured that numerical values such as ENGINE SIZE, CYLINDERS, and the fuel efficiency metrics (HWY, COMB) align with real-world expectations and are within plausible ranges. This step aims to identify and rectify any discrepancies that may impact the precision of subsequent analyses.

3. **Consistency:**Enforced consistency in data formatting and units across all columns. Confirmed that categorical variables like TRANSMISSION and FUEL adhere to standardized formats. Consistency checks contributed to the dataset's uniformity, preventing inconsistencies that could lead to misinterpretation during analysis.

4. **Relevance:** Evaluated the relevance of each column with respect to the analysis goals. Ensured that all columns, including less quantitative ones like MAKE and VEHICLE CLASS, contribute meaningfully to the intended insights. This step helps streamline the dataset, focusing on essential information for the fuel consumption analysis.

Overall, conducted a comprehensive Data Quality Check for the Fuel Consumption Data Analysis dataset involving assessing completeness, accuracy, consistency, and relevance across all specified columns. This meticulous process ensures a robust foundation for deriving meaningful insights and making informed decisions based on the dataset.

The 5 Vs of The Fuel Consumption Data Analysis

1. **Volume:** The dataset comprises a substantial volume with around 22546 rows and 13 columns, capturing detailed information about vehicles, including their make, model, specifications, and fuel-related metrics.

2. **Velocity:** The data showcases a consistent velocity, representing multiple entries for each year, make, and model, reflecting the ongoing pace of vehicle data accumulation over time.

3. **Variety:** The dataset exhibits diverse data types, encompassing categorical variables like MAKE, MODEL, and TRANSMISSION, as well as numerical values such as ENGINE SIZE, FUEL CONSUMPTION, and EMISSIONS. This variety enables a comprehensive analysis of different aspects of vehicle characteristics and fuel efficiency.

4. **Veracity:** Ensuring data accuracy is crucial for reliable analysis. Verification processes should be in place to confirm the precision of numerical values like HWY (L/100 km) and COMB (mpg), enhancing the overall veracity of the dataset.

5. **Value:** The dataset provides significant value by offering insights into fuel consumption patterns, emissions, and other key factors for a wide range of vehicles. Extracting valuable information from this dataset can guide decisions in vehicle design, fuel efficiency improvements, and environmental impact assessments.

Latency and Processing time

- **Initial Latency Assessment:** Considering the dataset's size, an initial evaluation of latency during data retrieval and loading is imperative. This involves measuring the time it takes to access and load the entire dataset into memory. Identifying potential bottlenecks at this stage is essential for optimizing subsequent processing steps.

- **Processing Time for Data Analysis:** Once loaded, processing time for data analysis becomes a critical factor. This includes tasks such as aggregations, filtering, and transformations on columns like FUEL CONSUMPTION, HWY (L/100 km), COMB (L/100 km), COMB (mpg), and EMISSIONS. Employing parallel processing and optimized algorithms can significantly reduce the time required for these computations.

- **Latency in Transmission Handling:** Given the dataset's characteristics, the latency associated with data transmission and communication between different components of a distributed system must be considered. Minimizing this latency ensures smooth and timely flow of information, especially in scenarios involving distributed data processing or remote servers.

- **Optimizing Query Response Time:** For interactive data exploration, minimizing query response time is essential. Indexing key columns like YEAR, MAKE, MODEL, and VEHICLE CLASS can enhance the efficiency of filtering and retrieval operations. Additionally, employing caching mechanisms for frequently accessed data can further reduce latency in responding to user queries.

Resource Utilization, Security, and Cost

- **Resource Utilization:** Efficient resource utilization is crucial for handling a dataset like vehicle analysis. Optimizing hardware resources, employing parallel processing, and leveraging scalable cloud solutions can enhance processing speed and manage the computational demands effectively.

- **Security:** Ensuring data security is paramount when dealing with a large dataset, especially one containing sensitive information about vehicles. Implementing robust encryption protocols, access controls, and regular audits helps safeguard against unauthorized access or data breaches, ensuring the confidentiality and integrity of the dataset.

- **Cost:** Managing costs associated with storage and processing is a key consideration for a dataset of this scale. Employing cost-effective storage solutions, optimizing query performance, and exploring serverless computing options can contribute to cost efficiency while maintaining the required processing power for comprehensive analysis of the fuel consumption dataset.

Conclusions

Based on the analysis of the Fuel consumption data, We have drawn several conclusions in the goals sections.
Overall, the analysis of the historical fuel consumption trends of cars unveils insights into evolving usage patterns, influenced by factors like engine efficiency and technological advancements. Understanding these trends is pivotal for assessing environmental impact, pollution levels, and the overall efficiency of automotive technologies.

References

1 https://jupyter.org/
2 https://pandas.pydata.org/
3 https://spark.apache.org/