José G. Portela

In the dedicated GitHub repository, you can access the data used in the analysis. Navigate to the "data" folder, where you will find the two datasets, along with the Jupyter Notebook documenting the analysis process.

Preliminary overall summary report:

1. **Library Import and Dataset Naming:** To gather insights and facilitate the analysis, relevant libraries were imported, encompassing both those already employed and others potentially useful for the upcoming analysis. The datasets were individually loaded and named as measurements1 and measurements2, respectively.

```
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import chardet
from scipy import stats
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.tsa.seasonal import STL
from matplotlib.ticker import ScalarFormatter
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.api import STLForecast
from statsmodels.tsa.arima.model import ARIMA
```

executed in 6.22s, finished 12:56:57 2024-02-09

```
measurements1 = pd.read_csv("../data/measurements.csv")
measurements2 = pd.read_excel("../data/measurements2.xlsx")
```

executed in 3.79s, finished 12:59:27 2024-02-09

2. **Dataset Discrepancy:** It has been observed that although the two datasets share identical columns and exhibit similar values, they are not identical. Discrepancies may exist in the data, and further investigation is necessary to understand the variations between the datasets. The following is the code used to identify the dataset discrepancy:

```
# Comparing data sets to confirm if the two data sets are the same or not

measurements_equal = measurements1.equals(measurements2)

if measurements_equal:
    print("Measurements from both data sets are the same.")
else:
    print("Measurements from both data sets are not the same.")
```

executed in 43ms, finished 13:11:01 2024-02-09

```
Measurements from both data sets are not the same.
```

3. **Outlier Detection using Z-Score:** Applying the z-score method revealed the presence of outliers in the dataset. Notably, one of the datasets (measurements1) contains string values, introducing a data type mismatch that may impact the analysis.

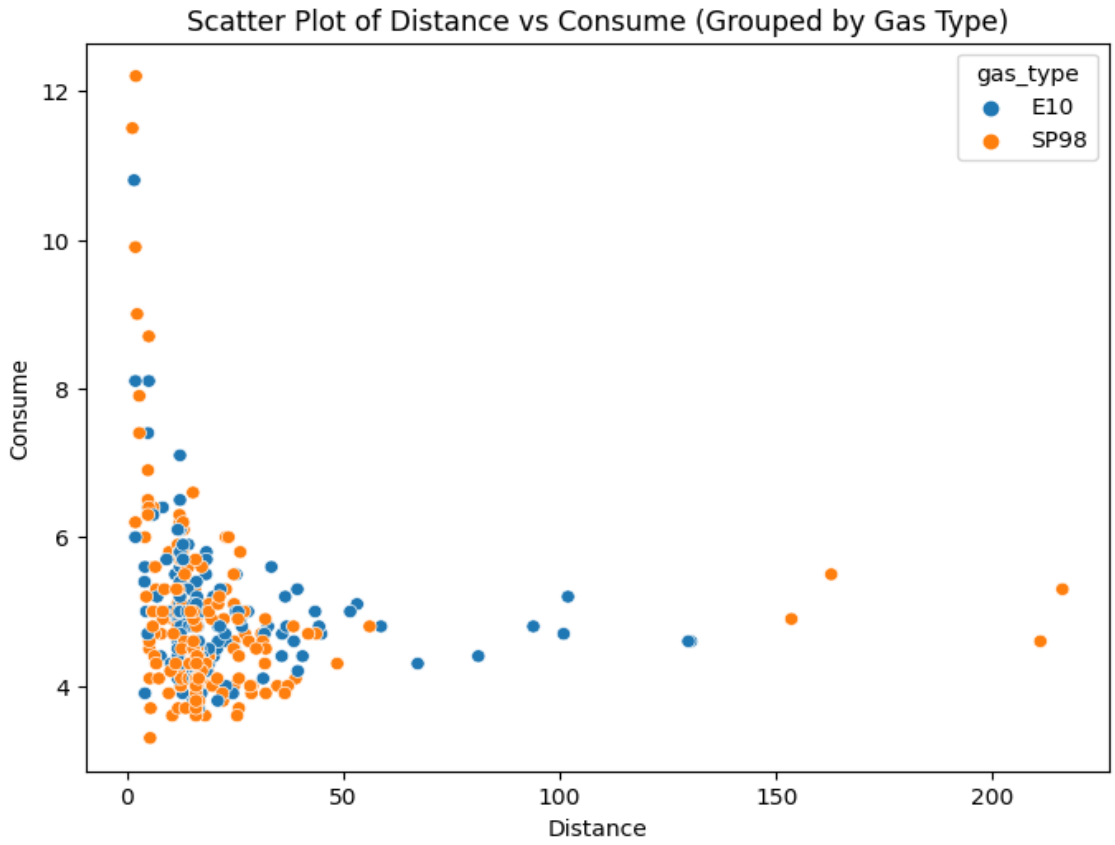
```
from scipy import stats
z_scores = stats.zscore(measurements2[['consume','distance']])
outliers = (z_scores > 3) | (z_scores < -3)
print(measurements2[outliers])

```

executed in 113ms, finished 13:30:56 2024-02-09

	distance	consume	speed	temp_inside	temp_outside	specials	gas_type	AC	rain	sun	refill liters	refill gas
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
..	...	...	...	...	...	...	...	..	...	...	...	...
383	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
384	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
385	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
386	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
387	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

[388 rows x 12 columns]

4. **Data Set Adjustments for Complex Studies:** The preliminary check for missing values within the datasets proved crucial, allowing an evaluation of the potential impact of missing values on the study. Both scenarios with and without missing values were examined to assess how these variations might influence the analysis. The visualization further highlights the need for potential adjustments to the datasets for more intricate analyses. Techniques such as imputing missing values using mean or median values could be considered. It is crucial to address any data inconsistencies or type disparities before proceeding with advanced studies.
5. **Visualization Insights:** Examination of the visualization indicates that, without alterations to the datasets, markers can be employed to identify associations within the second dataset (measurements2). Specifically, the 'distance' and 'consume' columns exhibit noteworthy connections when exploring different gas types. This insight underscores the importance of meticulous dataset preparation for accurate and meaningful visual interpretations. (Next page for plot)


Scatter Plot of Distance vs Consume (Grouped by Gas Type)
gas_type
E10
SP98
Consume
Distance
12
10
8
6
4
0
50
100
150
200