

José G. Portela

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.

```
1 import pandas as pd
2 import os
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 import numpy as np
6 import chardet
7 from scipy import stats
8 from statsmodels.graphics.tsaplots import plot_acf
9 from statsmodels.tsa.seasonal import STL
10 from matplotlib.ticker import ScalarFormatter
11 from statsmodels.tsa.seasonal import seasonal_decompose
12 from statsmodels.tsa.api import STLForecast
13 from statsmodels.tsa.arima.model import ARIMA
```

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

```
1 measurements1 = pd.read_csv("../data/measurements.csv")
2 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:

```
1 # Comparing data sets to confirm if the two data sets are the same or not
2
3 measurements_equal = measurements1.equals(measurements2)
4
5 if measurements_equal:
6     print("Measurements from both data sets are the same.")
7 else:
8     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.

```
1 from scipy import stats
2 z_scores = stats.zscore(measurements2[['consume','distance']])
3 outliers = (z_scores > 3) | (z_scores < -3)
4 print(measurements2[outliers])
5
```

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)

