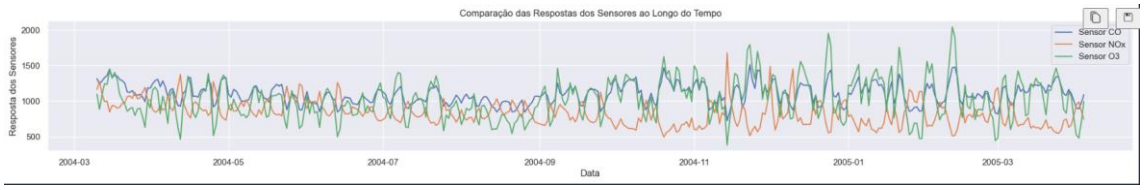
Data Scientist: Marina Lourenço Borges

1. Air Quality

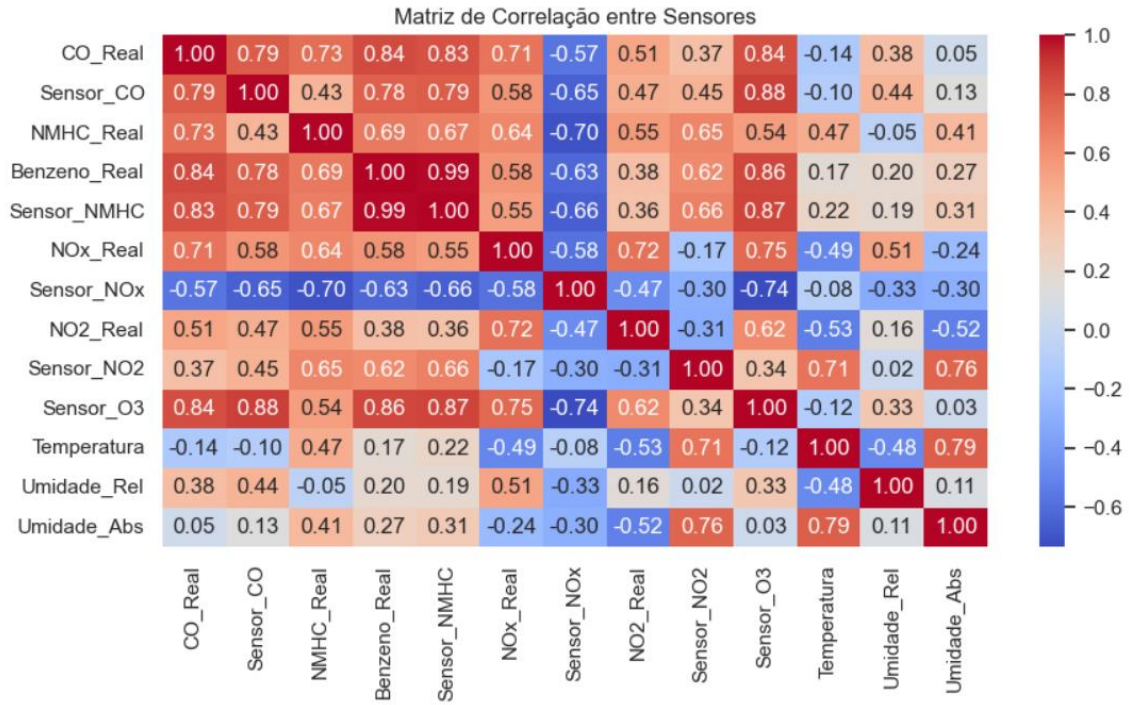
The DataFrame on Air Quality contains information about air quality in a polluted area of an Italian city over the course of a year. It includes data such as the date, time of day, actual and sensor-based hourly average concentrations of various atmospheric pollutants like carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene, nitrogen oxides (NOx), and nitrogen dioxide (NO2), as well as sensor responses to these pollutants. Additionally, the DataFrame also encompasses information on temperature, relative humidity, and absolute humidity.

Initially, we validated that the dataset contained NaN data masked with the number -200. Therefore, we validated the information and removed the lines from the three main columns: Temperature, Relative_Humidity, and Absolute_Humidity.

An evaluation was conducted on the sensor responses, as shown in the figure below, suggesting that there might be some correlation between the sensor responses, as the lines appear to follow similar patterns at certain points, rising and falling together, which could indicate a simultaneous response to certain environmental factors or sources of pollution.



Next, an analysis of the correlation matrix among all available data was conducted:



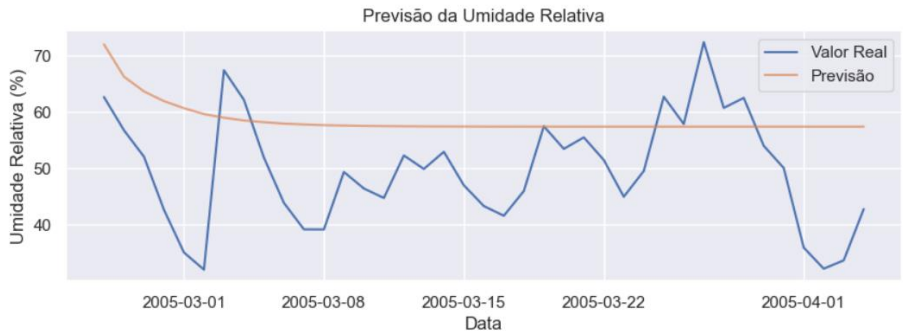
Successfully mapping the correlations above 0.7 of the features:

```
[('Sensor_CO', 'CO_Real'),
 ('NMHC_Real', 'CO_Real'),
 ('Benzeno_Real', 'CO_Real'),
 ('Benzeno_Real', 'Sensor_CO'),
 ('Sensor_NMHC', 'CO_Real'),
 ('Sensor_NMHC', 'Sensor_CO'),
 ('Sensor_NMHC', 'Benzeno_Real'),
 ('NOx_Real', 'CO_Real'),
 ('NO2_Real', 'NOx_Real'),
 ('Sensor_O3', 'CO_Real'),
 ('Sensor_O3', 'Sensor_CO'),
 ('Sensor_O3', 'Benzeno_Real'),
 ('Sensor_O3', 'Sensor_NMHC'),
 ('Sensor_O3', 'NOx_Real'),
 ('Temperatura', 'Sensor_NO2'),
 ('Umidade_Abs', 'Sensor_NO2'),
 ('Umidade_Abs', 'Temperatura')]
```

Therefore, it is possible to conclude that the graph and the correlation matrix reveal an important insight: the presence of strong correlations between the sensor readings and the actual measurements of pollutants, such as CO and NOx, indicate that the sensors are effective in tracking the real levels of these contaminants.

Specifically, the strong correlation between 'Sensor_CO' and 'CO_Real' suggests that the CO sensor is a reliable indicator of the presence of carbon monoxide in the environment. This can be useful for real-time air quality monitoring and for the implementation of more effective pollution control measures.

Continuing with the estimation of Relative Air Humidity, the first ARIMA (Autoregressive Integrated Moving Average) model was developed for estimating Relative Air Humidity, using all available variables in the database.



The graph shows a comparison between the actual values of relative humidity and the predictions generated by a forecasting model. The blue line represents the actual data of relative humidity over time, while the orange line shows the model's predictions.

We can see that there are some significant discrepancies between the predictions and the actual values, suggesting that the model may need to be fine-tuned to better capture the variation in relative humidity. This might involve adjusting the model's parameters, incorporating other predictive variables, or using a different type of model.

For a second test, knowing that relative humidity has a relationship between the amount of water present in the air (absolute humidity) and the maximum amount of water that could exist at the same temperature (saturation point).

A second model was constructed using only the variables Temperature and Absolute Humidity, with the SARIMA (Seasonal Autoregressive Integrated Moving Average) model. SARIMA is an extension of ARIMA that adds support for seasonal patterns in time series data. It

is capable of modeling and forecasting series that exhibit seasonality, in addition to non-seasonal trends and patterns.

The training dataset (in blue) is used to adjust the model, and the predictions (in red) are compared with the actual test values (dark line). The red shaded area represents the confidence interval of the predictions. The model captures the general trend and some fluctuations, but there are deviations, especially at points of peaks and troughs.

The Mean Squared Error (MSE) provided is a measure of the difference between predicted values and actual values, indicating the accuracy of the model's predictions. An MSE of approximately 9.93 suggests that the predictions are, on average, about 3.15 units (the square root of the MSE) percentage points away from the actual values.

The SARIMA model is a suitable choice for starting the study, given the exploratory analysis performed. However, it is important to note that it only utilizes two of the 13 features available in the dataset.

Relative humidity, despite not showing strong correlations with other variables, can be indirectly estimated since absolute humidity and temperature have relationships with other variables.

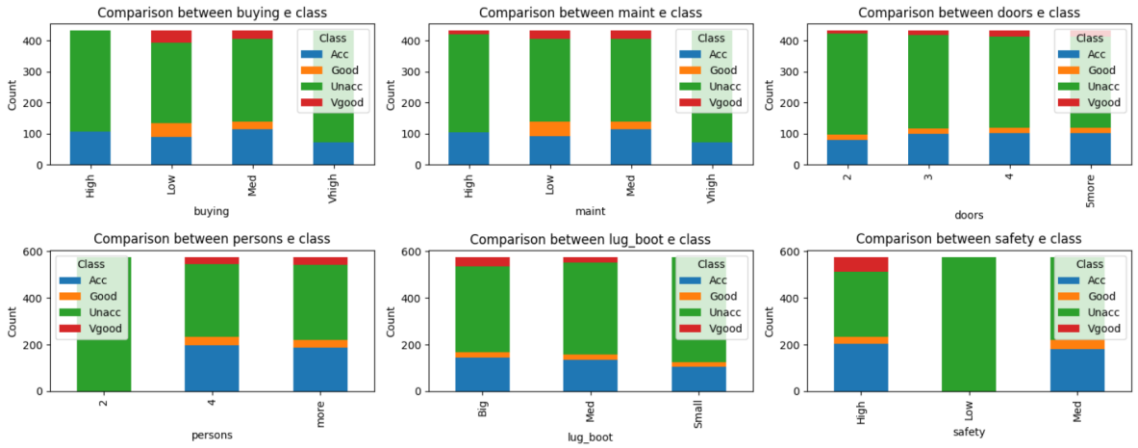
Understanding how temperature and pollutants affect absolute humidity could, therefore, be useful in modeling and estimating air relative humidity. For example, if the temperature increases and the absolute humidity does not change, the relative humidity will decrease, as warmer air can contain more water vapor.

Therefore, the study should proceed to the next sprint with the concept of responding to validations of how temperature and pollutants affect absolute humidity as a direct inference to relative humidity in the air. Moving on to the construction of another model that brings the other available variables as features for the study.

2. Evolution of Cars.

The DataFrame on the evolution of cars contains information on different aspects of vehicles, such as purchase price, maintenance cost, number of doors, passenger capacity, trunk size, and estimated safety level. It includes a column that provides the car evaluation level, classifying them as unacceptable, acceptable, good, or very good.

Firstly, by performing a series of bar graphs as shown below, it was possible to construct an immediate visual comparison between the categories and their respective characteristics.



It is noted that vehicles with high purchase prices and maintenance costs are often associated with negative evaluations, frequently being classified as unacceptable.

On the other hand, characteristics such as the ability to accommodate passengers and a high level of safety are valued attributes, leading to more positive classifications, such as acceptable or good. Similarly, a generously sized trunk is a factor that positively contributes to a vehicle's classification, reinforcing the preference for more spacious and versatile cars.

Interestingly, the number of doors on vehicles, a characteristic that could be considered indicative of the car's size or type, appears to have a lesser influence on evaluations compared to the other factors mentioned, suggesting that consumer priorities may align more with functional and safety aspects rather than design or form.

When evaluating the effectiveness of different machine learning models in classifying vehicles based on characteristics like price, maintenance, capacity, and safety, a significant variation in performance was observed.

- Logistic Regression, despite being a widely used method for classification problems, achieved an accuracy of only 66.47%, with low precision and recall for most classes, indicating limited suitability for this dataset.

- Random Forest demonstrated substantial improvement, with an accuracy of 96.72%, highlighting its ability to capture the complexity and nuances of the data.

- Decision Tree exhibited slightly superior performance, with an accuracy of 97.11%, and showed a particularly higher recall for the class labeled as 'good'.

These results suggest that, for this specific dataset, both Random Forest and Decision Tree are robust models, with Decision Tree offering a marginal advantage in accuracy. The final choice between the two may depend on additional factors such as model complexity and

computational efficiency, but from a strictly quantitative standpoint, the Decision Tree presents a slightly superior performance.

The analysis of feature importance in a Random Forest model revealed that safety was identified as the most crucial feature, followed by passenger capacity (persons) and purchase price (buying).

These findings suggest that aspects of safety, practicality, and cost are key considerations for consumers when evaluating cars. In contrast, trunk size (lug_boot) and the number of doors (doors) were considered the least influential features in determining the car class, indicating that physical characteristics have less impact on class choice.

For the next steps in studying this model, it is suggested to apply class balancing techniques (Oversampling, Undersampling, and SMOTE) due to the observed imbalance, especially in the 'unacceptable' class, which accounts for 70% of the data.

This imbalance can lead to biased results, where the model may struggle to correctly learn the minority class, resulting in less accurate and reliable predictions for that class. Therefore, it is important to address this imbalance to ensure that the model is capable of generalizing well for all car evaluation classes.