Dataset: https://archive.ics.uci.edu/dataset/360/air+quality
GROUP 24-003
474 Milestone 06
Project Title: *How The City Decides When You Can Breathe*

Instructions

0. Prediction Goal/s

1. **Clearly state your prediction goal or prediction goals**

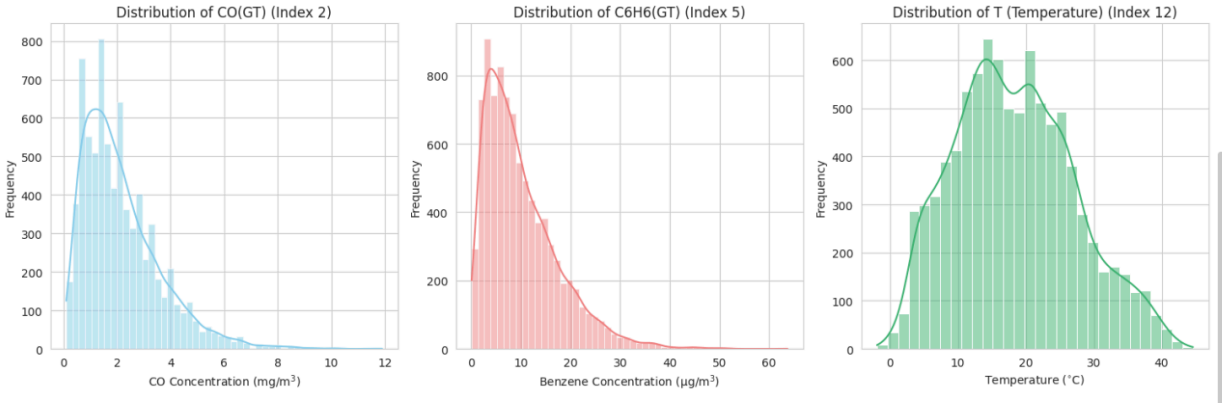
   Our project addresses this challenge by forecasting hourly CO levels using predictive analytics. By identifying when pollution peaks occur throughout the day, we aim to provide actionable insights that help residents-especially those with respiratory conditions-make safer decisions, while equipping policymakers with evidence-based tools to mitigate risks.

1. Dataset Exploration

1. **Dataset Overview**
   - Clearly state the size of the dataset (number of observations and features).
   - Describe each feature briefly (e.g., data type, meaningful range if applicable).
   - Identify the response variable (continuous or categorical).

The UCI dataset showing the different metals within air samples of the surrounding urban environment contains 9,357 total observations, of which 18% or 1,683 values are missing. The data contains 14 predictor variables classifying both numerical and temporal values. The chosen response variable for this analysis is CO(GT)(True CO Concentration), which is a continuous variable measured in mg/m^3.

2. **Descriptive Statistics & Visualizations**
   - Provide basic summary statistics (e.g., mean, median, standard deviation) for the response variable.
   - Generate at least one relevant visualization (e.g., histogram for continuous variables, bar plot for categorical variables).
   - Discuss any notable findings or patterns (e.g., skewness, outliers).




```
--- Summary Statistics (after index-based cleaning) ---
0           CO(GT)  PT08.S1(CO)     NMHC(GT)     C6H6(GT)  PT08.S2(NMHC)
count  7674.000000  8991.000000   914.000000  8991.000000    8991.000000
mean      2.152750  1099.833166   218.811816    10.083105     939.153376
std       1.453252   217.080037   204.459921     7.449820     266.831429
min       0.100000   647.000000     7.000000     0.100000     383.000000
25%       1.100000   937.000000    67.000000     4.400000     734.500000
50%       1.800000  1063.000000   150.000000     8.200000     909.000000
75%       2.900000  1231.000000   297.000000    14.000000    1116.000000
max      11.900000  2040.000000  1189.000000    63.700000    2214.000000

0           NOx(GT)  PT08.S3(NOx)      NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)
count  7718.000000   8991.000000  7715.000000   8991.000000  8991.000000
mean    246.896735    835.493605   113.091251   1456.264598  1022.906128
std     212.979168    256.817320    48.370108    346.206794   398.484288
min       2.000000    322.000000     2.000000    551.000000   221.000000
25%      98.000000    658.000000    78.000000   1227.000000   731.500000
50%     180.000000    806.000000   109.000000   1463.000000   963.000000
75%     326.000000    969.500000   142.000000   1674.000000  1273.500000
max    1479.000000   2683.000000   340.000000   2775.000000  2523.000000

0                T           RH
count  8991.000000  8991.000000
mean     18.317829    49.234201
std       8.832116    17.316892
min      -1.900000     9.200000
25%      11.800000    35.800000
50%      17.800000    49.600000
75%      24.400000    62.500000
max      44.600000    88.700000
```

The distribution is strongly right-skewed. The mean is noticeably higher than the median. This suggests that while most CO(GT) concentrations are low, there are a number of high concentration readings that pull the mean upwards. The long tail suggests the presence of outliers, representing peak pollution events.

3. **Predictor Variables & Response Variable**
    - Classify each predictor's type (numerical, categorical, etc.).
    - Briefly discuss how you expect each predictor to relate to the response variable (e.g., domain knowledge, preliminary correlation).

Predictor Group	Feature(s)	Expected Relationship to CO(GT)	Rationale
Temporal	Date, Time	Cyclic/Non-linear	CO concentration is expected to follow diurnal (high during traffic hours) and seasonal (higher in colder months) patterns.
Pollutants	NMHC(GT), C6H6(GT), NOx(GT), NO2(GT)	Strong Positive Correlation	These pollutants share common sources (e.g., vehicle exhaust, combustion). When the concentration of one is high, the others are also likely to be high.

Sensors	PT08.S1(CO)	Strongest Positive Correlation	This sensor is nominally selective to CO and is designed to measure it directly, making it the most critical predictor.
Sensors (Others)	PT08.S2-S5	Positive/Moderate Correlation	These sensors may exhibit positive correlation due to cross-sensitivity (responding weakly to CO and the strong inter-correlation of the pollutants themselves.
Environmental	T (Temperature)	Negative/Inverse Correlation	Higher temperatures generally aid in the atmospheric dispersion of pollutants, leading to lower ground-level CO

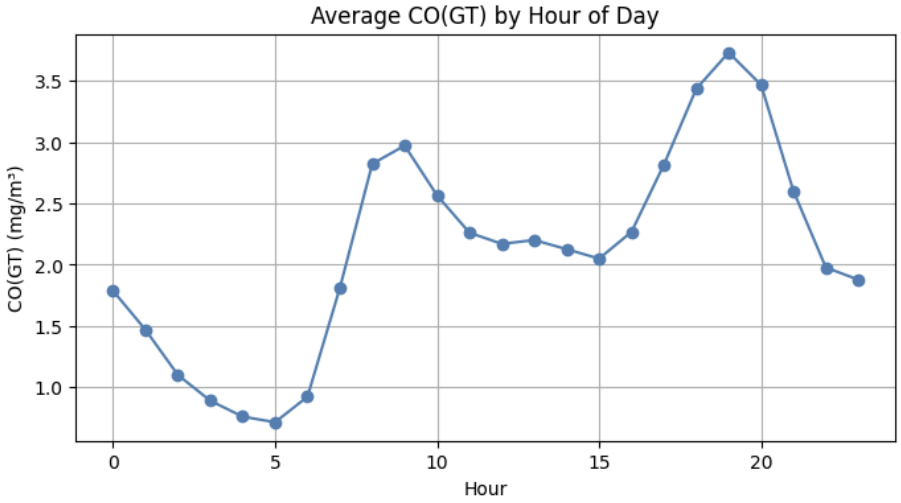
			concentrations.
Environmental	RH, AH (Humidity)	Complex/Non-linear	Humidity can influence CO concentration and significantly affect the response of the metal oxide sensors, introducing a confounding factor.

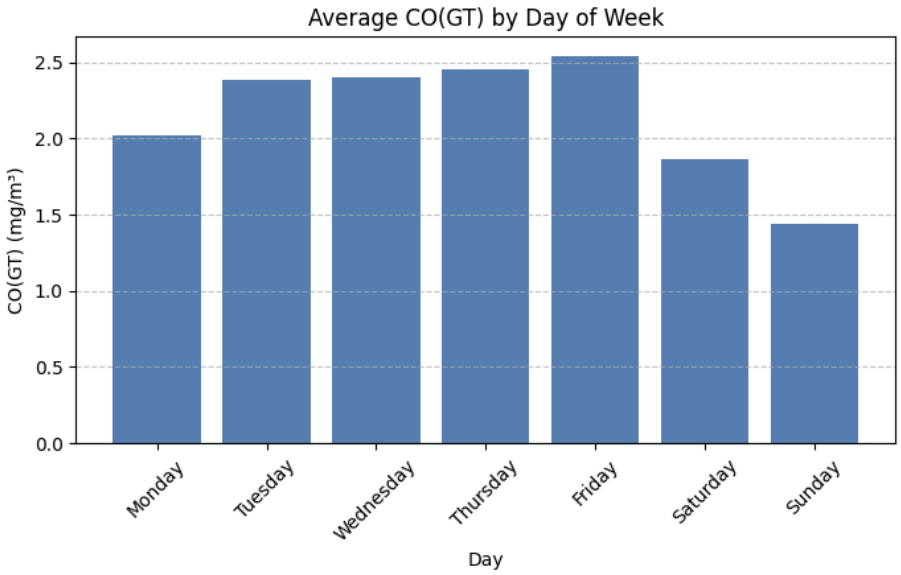
---

2. Feature Engineering

1. **Creation of a New Feature**

We created three new features, all temporal in nature, taking the datetime and converting it to hour of day, day of week, and month of year. We call these features “Hour”, “Month” and “DayOfWeek”. The reason we did this is fairly standard, to capture relationships not seen in the raw date feature which is not exceptionally useful on its own. These three features actually became some of our strongest ones with exceptionally high correlations to the target of carbon monoxide. Of course, this makes sense, as the carbon monoxide concentration is the majority from automobiles. Therefore, the concentration will be higher when there are more cars on the road. The graphs below show the correlation between the features hour and day and the target of carbon monoxide concentration.





2.

3.

4. **Justification**
    - Provide a clear rationale for why this new feature might be important for the predictive task.
    - Relate the new feature to either domain knowledge or data-driven insights.

These features are all incredibly important, because they are important signals for carbon monoxide concentration. The time of day and day of week Of course, it is important to note that day of week is not causal, just because it is Saturday does not mean there is less carbon monoxide, but instead it is the concentration of vehicles, heating systems and all of those things which are actually the root cause. Regardless, this information illustrates some of the strongest relationships to the target in our entire dataset, and will be massive in aiding our prediction combined with the other features in this dataset.

As we can see in the graphs above, carbon monoxide seems highest during peak commuting times (i.e. weekdays & rush hour). The concentration falls throughout the night, as it dissipates and there are less people on the road. One interesting thing about time of day particularly is that it shows the time it takes for the carbon monoxide to dissipate, and illustrates that some concentration of it will remain in the air for a while even if very few people are driving.

---

3. Handling Missing Values

1. **Identification**
    - Determine which features have missing values and the extent of these missing values.
    - Present and justify a clear strategy for dealing with missing data (e.g., dropping observations vs. imputing).

We solved some problems with missing data and cleaned the data set by removing the -200 values which indicate a faulty reading; some columns, such as NMHC(GT), had an overwhelming proportion of missing data, making them unsuitable for most standard analyses without imputation.

---

4. Baseline Model Implementation

1. **Model Choice**
    - Identify the simplest reasonable model type for your prediction goal:
        - **Continuous Response**: Linear regression or another basic regression method.
        - **Categorical Response**: Logistic regression or another basic classification method.

We believe that multiple linear regression is the simplest reasonable model type for our prediction goal. We are predicting a continuous response, and thus it is the best simple model to choose. We may branch out to more advanced and ensemble models later on in our prediction if it is needed.

2. **Implementation**
    - Implement the model combining a feature selection procedure with k-fold (k=5 or k=10) cross validation to evaluate each model performance
    - Using the CV error, select the best model among all

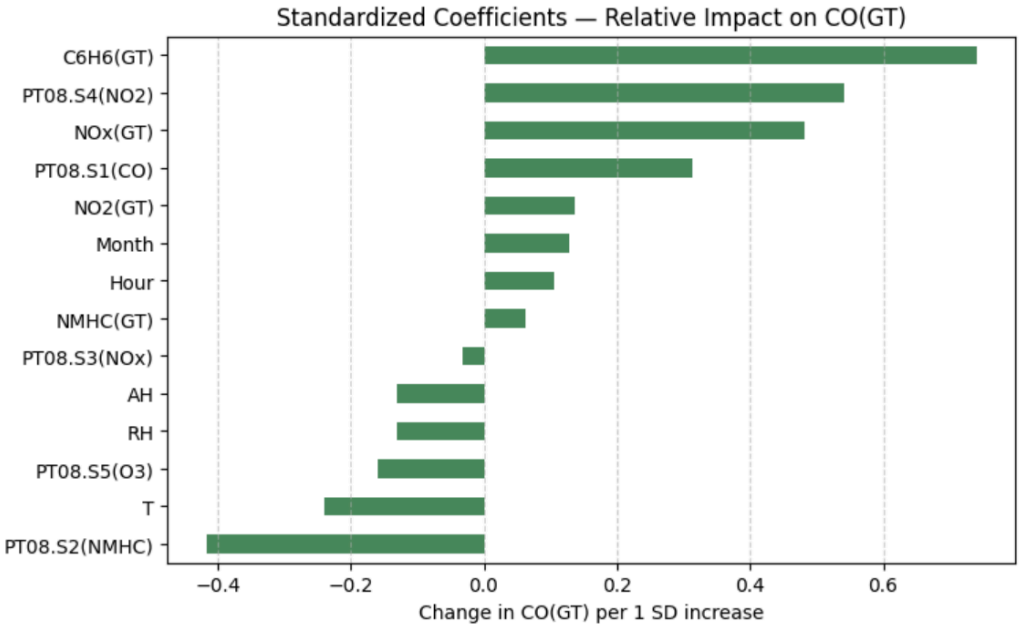
We also used stepwise selection as a feature selection procedure, to try and find the best combination of features for the purposes of the model. Finally, we k-fold with k=10 to evaluate each model performance. The stepwise selection kept almost all of the features, with only one

being cut from the model (MonthName) as it was redundant with month number also being added.

- **Interpretation & Next Steps**
  - Summarize the performance results.
  - Reflect on whether the model results suggest the need for further engineering, feature selection, or tuning.

```
R² on full data: 0.896

Model Coefficients (sorted):
C6H6(GT)           0.101818
Month              0.036275
Hour               0.015237
NO2(GT)            0.002916
NOx(GT)            0.002301
PT08.S4(NO2)       0.001580
PT08.S1(CO)        0.001468
NMHC(GT)           0.000887
PT08.S3(NOx)      -0.000125
PT08.S5(O3)       -0.000404
PT08.S2(NMHC)     -0.001605
RH                -0.007685
T                 -0.027581
AH                -0.333205
dtype: float64
```



As we can see from this visualization, there are plenty of strong predictors, but some such as benzene (C6H6) and Nitrogen Oxide (NO2) in the positive direction are especially strong. On the other hand, Nitrogen Dioxide and Temperature are strong predictors in the

negative direction. Indicating lower carbon monoxide concentration when the temperature is higher. This initial model had an R^2 of .896, which is a very strong start, but it should be noted R^2 is not everything in the scope of this problem. We are not simply trying to predict exactly how much carbon monoxide is in the atmosphere, but instead we are trying to find solutions to make the streets safer to walk in for millions of pedestrians. Therefore, predictors that are actionable and causal can help us make the city safer. As it is clear automobiles and peak commute times are main causal drivers to carbon monoxide concentration, we should focus on cars which are more efficient, or implementing more green spaces in urban areas.

In the future, we will do additional feature engineering to try and judge the interaction between some of our features, such as potentially separating weekends versus weekdays, as there tends to be differences in trends on those days. Furthermore, it is important to focus more on examining which of these predictors are truly causal. For example, benzene is the strongest correlated predictor which indicates automobiles are a factor, but simply lowering the benzene concentration would not necessarily lower carbon monoxide concentration unless gas cars went with it. We need to work to figure out which of these drivers can be implemented throughout the world to make our cities safer for pedestrians at all times of the day.

https://colab.research.google.com/drive/1GcytZTHgkFVx7wiA5k5aRVIGNkMaU_h8?usp=sharing