Vehicle Registration and PM2.5 Pollution in Europe: A Feature Importance Analysis

Team: https://github.com/Toveyyy/mids207-project-summer-2025.git
Randy Tovar: 012795r@berkeley.edu
Trenton Carlson: trenton_carlson@berkeley.edu
Hannah MacDonald: hannah_macdonald@berkeley.edu

Abstract
This study explores whether vehicle registration data significantly improves predictions of PM2.5 concentrations across the EU. Using air quality, population, and vehicle registration EEA datasets from 2018 to 2024, a variety of models were created with and without vehicle features to determine their importance. These models include: a baseline linear regression, a gradient boost decision tree model (XGBoost), and a feed-forward neural network (FNN). Results indicate that non-vehicle factors such as country, population, and geographic location were generally more influential in predicting PM2.5 levels, with vehicle data contributing only marginal improvements. The best performing model was XGBoost with a final test RMSE of 1.5137, for which vehicle related features contributed minimally to reduction of model error (2.08%).

Introduction
Airborne particulate matter greater than 2.5 microns in diameter (PM2.5) at concentrations above 5 μg/m³ is a significant contributor to respiratory and cardiovascular illness. In response, public policy aimed at mitigating PM2.5 levels has traditionally focused on regulatory measures in the energy, transportation, and agriculture sectors. Within transportation, attention has increasingly turned toward reducing emissions from personal vehicles through electrification, improved fuel standards, and cleaner combustion technologies. While the environmental case for electric vehicles is often supported by theoretical modeling and life-cycle assessments, there remains a need for empirical, data-driven analysis to determine the extent to which changes in vehicle fleets and fuel use correspond to measurable air quality improvements.

This project focuses on quantifying the relationship between transportation-sector changes, particularly in vehicle composition and fuel types, and PM2.5 concentrations in Europe from 2018 to 2023. Our dataset integrates vehicle registration records, air sampling station metadata, fuel consumption statistics, and population data at the country-year level. The input to our algorithm is structured tabular data containing both numeric and categorical variables; the target output is the average annual PM2.5 concentration (μg/m³) for each country. To evaluate the role of vehicle data in predicting PM2.5 levels, we train models both with and without vehicle-related features.

Related Work
Recent research continues to explore the environmental implications of electric vehicle (EV) adoption, particularly in relation to air pollution and fine particulate matter (PM2.5). Zhang et al. (2022) conducted an experimental study comparing PM2.5 emissions from electric vehicles (EVs) and internal combustion engine vehicles (ICEVs), finding EVs produce about 50% less PM2.5 due to the lack of tailpipe emissions. However, they note that EVs are not entirely emission-free, as particles still originate from tire wear, brake dust, and electricity generation.

A 2025 *Nature Scientific Reports* study examined real-world monitoring data across multiple cities and found that higher EV adoption correlated with localized reductions in nitrogen oxides (NOx) and PM2.5.

The benefits varied with background pollution sources, energy mixes, and urban design, highlighting the need for complementary policies like renewable energy integration and public transit expansion.

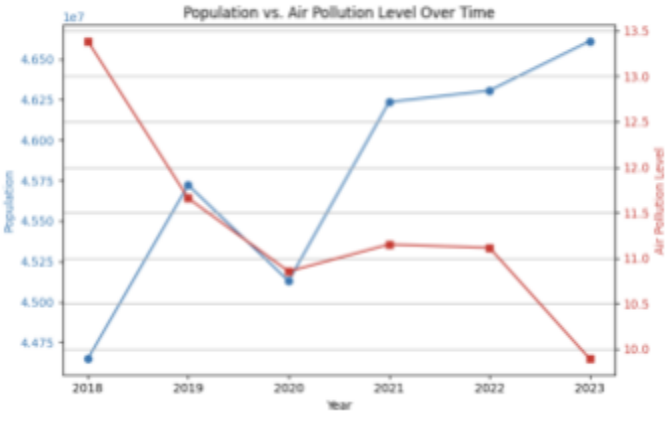
Our work differs by using machine learning to predict PM2.5 at the country-year level in Europe, incorporating EV and ICEV counts, fuel consumption, and population. This approach allows us to assess the relative importance of these factors in a multi-variable predictive framework.

Dataset

Our analysis draws on three primary data sources covering the years 2018 to 2023 across European countries: annual average PM2.5 measurements from the European Environmental Agency (EEA), vehicle fleet and fuel consumption data from EU transport repositories, and annual country-level population data from the World Bank. The original downloaded datasets were extensive, containing millions of rows at the individual air quality station or vehicle registration record level. To align these sources for modeling, we performed a multi-step filtering and aggregation process. First, air quality measurements were averaged annually by monitoring stations and merged with vehicle and population data at the country-year level, reducing the dataset to approximately 6,000 records. Additional filtering removed years or locations with incomplete data, yielding a cleaned dataset of 5,424 records. To address severe class imbalance in geographic representation, where larger countries such as Germany contributed disproportionately more samples, we applied two balancing strategies: grouping low-sample countries with fewer than 40 records into an “Other_Countries” category, and downsampling overrepresented countries with more than 500 records.

Preprocessing steps included standardizing numeric features with z-score normalization, except for Year, Longitude, and Latitude, which were min–max scaled. Missing values, such as absent fuel consumption for certain fuels in specific countries, were replaced with zeros; while this introduces the risk of interpreting a missing entry as a true zero, it provided a simple and consistent approach for model ingestion. A lag feature representing the prior year’s PM2.5 level was added to capture temporal dependence in air quality trends. Categorical air quality station attributes were one-hot encoded for use in TensorFlow-based models. The final dataset contained 26 features and was split into training (60%), validation (20%), and test (20%) sets using a stratified split by country to ensure consistent geographic representation across all subsets.

Figure 1:



Exploratory data analysis informed several of these preprocessing choices. Temporal trends (Figure 1) showed that PM2.5 levels declined over time while population remained relatively stable, suggesting that population growth alone is not a primary driver of pollution changes. Spatial interpolation of PM2.5 measurements using a distance-weighted k-nearest neighbors approach (Figure 2) revealed persistent hotspots in Northern Italy and parts of Eastern Europe. Country-level distributions (Figure 3) indicated substantial variability in PM2.5 across nations, with more populous countries generally exhibiting higher median pollution levels and wider variability. These insights motivated both the inclusion of lagged pollution features and the decision to downsample dominant countries to prevent their overrepresentation from biasing model training.

Figure 2:

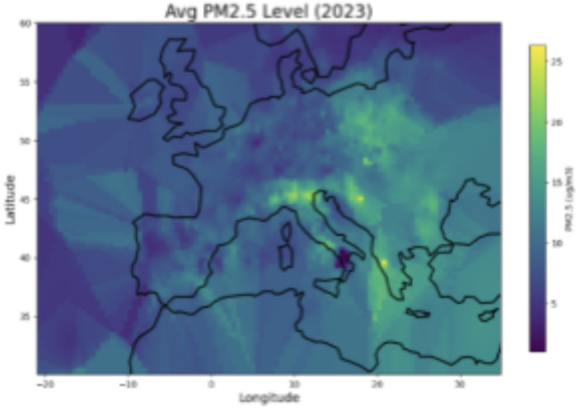
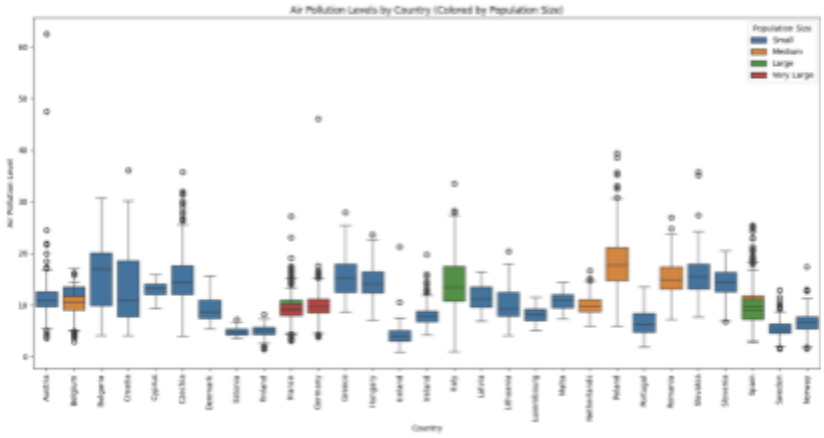


Figure 3:



Methods

The baseline model is a simple linear regression. This model takes both numerical and categorical features as inputs, applies one-hot encoding to the categorical variables, and then concatenates all features into a single tensor. The model was implemented using TensorFlow's Keras functional API rather than the simpler Sequential API, allowing for more flexible handling of mixed feature types and making it easier to extend the architecture for future experiments. A single dense layer with one output unit is used to produce predictions. The model is optimized with Adam and trained to minimize mean squared error, while tracking root mean squared error and mean absolute error as evaluation metrics. This baseline is appropriate for our problem because it provides an interpretable mapping between input features and the target variable. Linear models are a natural starting point as they establish a performance benchmark and help identify whether more complex models are warranted.

As the first improvement over the baseline, we implemented a Gradient Boosted Decision Tree model using XGBoost, trained with and without vehicle-specific features to assess their predictive performance. XGBoost builds an ensemble of decision trees, where each new tree is trained to reduce the residual errors of the previous ones. The model prevents overfitting and uses gradient descent to find the optimal splits in the trees.This algorithm is well-suited to our problem because it can capture nonlinear interactions between features and is robust to outliers and skewness. By excluding vehicle features in this experiment, we can assess how much predictive power comes from non-vehicle-related inputs, offering valuable insights into feature importance.

The second improvement is a feedforward neural network. The architecture consists of two hidden layers: the first with 128 neurons and ReLU activation, followed by a dropout layer with a 10% dropout rate, and a second dense layer with 64 neurons and ReLU activation. The output layer contains a single neuron for regression output. The model is compiled with the Adam optimizer and trained with MSE loss and MAE as a metric.This model is appropriate because neural networks excel at modeling complex, high-dimensional relationships that may not be well captured by linear models or decision trees. The ReLU activation function enables the network to learn nonlinear transformations efficiently, while dropout regularization mitigates overfitting by disabling a fraction of neurons during training, forcing the network to learn more generalizable patterns. Compared to the baseline, this architecture significantly increases the model's capacity to detect intricate feature interactions.

Experiments, Results and Discussion

We evaluated three main modeling approaches: a baseline linear regression model, an XGBoost gradient boosting model, and a fully connected feedforward neural network (FNN). For each approach, we ran experiments with two feature configurations: one including all features (vehicle-related and non-vehicle) and one excluding vehicle-related features to test their added predictive value. Hyperparameters were tuned on the validation set, and final performance was reported on the holdout test set using RMSE and MAE as primary evaluation metrics.

Baseline Model – Linear Regression: The baseline used TensorFlow's linear regression implementation with all numeric and encoded categorical features. This model serves as a point of comparison for more complex methods, offering interpretable coefficients for each predictor. Without vehicle-related features, the baseline's validation RMSE increased slightly, confirming that vehicle metrics provide additional predictive signal. However, the model's limited capacity restricted its ability to capture non-linear relationships, leading to higher residual errors compared to tree-based and neural network models.

XGBoost (Gradient Boosting): We trained XGBoost regressors using early stopping based on validation RMSE. The primary tuned parameters included max_depth (4–6), learning_rate (0.05–0.1), subsample (0.8), and colsample_bytree (0.8). Early stopping typically occurred around 140–160 boosting rounds, where validation RMSE plateaued, preventing overfitting while preserving generalization. XGBoost consistently outperformed the baseline, with test RMSE reductions of over 10% in the all-features configuration. Feature importance analysis highlighted fuel consumption variables and certain vehicle categories as top predictors, reinforcing EDA findings that vehicle fleet composition influences PM2.5 levels. The exclusion of vehicle features caused a measurable performance drop, suggesting they add meaningful predictive power.

Feedforward Neural Network (FNN): The neural network consisted of two hidden layers (128 and 64 neurons) with ReLU activation, dropout layers (10%) to reduce overfitting, and the Adam optimizer (learning rate 0.001) with batch size 32. We applied early stopping based on validation RMSE, similar to the XGBoost setup. One-hot encoding was applied to categorical variables before model training. The FNN matched or slightly outperformed XGBoost in validation RMSE when using all features, but showed more sensitivity to overfitting, particularly without vehicle-related inputs. Dropout and early stopping were essential to maintain generalization.

Across all experiments, both XGBoost and the FNN showed signs of overfitting when trained for too many iterations, with validation RMSE rising while training RMSE continued to decline. To mitigate this, we used early stopping, applied dropout in the FNN, and tuned hyperparameters to control model complexity. Subgroup analysis by geographic region indicated slightly higher errors in areas with fewer samples, such as the aggregated "Other_Countries" category, which is consistent with reduced training representation. We choose Gradient Boosting as an improvement since its tree-based approach is better suited to handling the multicollinearity, non-linear relationships, and feature distribution skews found in our EDA. Ultimately, XGBoost emerged as the most effective model given its ability to find meaningful feature importance rankings and presenting the lowest root mean squared error and mean absolute error across all the models in our validation data. Hence, why we used this model as our final model for reporting and interpretation.

Conclusions

XGBoost likely outperformed the other models because it is highly effective for tabular data with mixed numeric and categorical features. Its tree-based structure naturally captures nonlinear relationships and complex feature interactions that linear models miss. Unlike neural networks, it does not require extensive

feature scaling or normalization, making it more robust with minimal preprocessing. It can learn useful decision rules quickly, even on medium-sized datasets, without the large data requirements of deep neural networks.The model also provides interpretable feature importance metrics, aiding in understanding the drivers of predictions.

Despite these strengths, several limitations remain. Grouping smaller countries into an “Other_Countries” category and downsampling large contributors such as Germany reduced geographic detail and may have obscured localized patterns. Missing fuel values were replaced with zeros, which may misrepresent reality; advanced imputation could improve this. Our temporal modeling was limited to a one-year lag, potentially missing longer-term pollution dynamics, and our hyperparameter tuning was constrained to manual searches. The dataset’s nested structure for air quality stations within countries was not explicitly modeled; future work could explore mixed-effects or hierarchical models to address this. With more time and computational resources, we would explore automated hyperparameter tuning, richer domain-specific feature creation, and deeper neural network architectures with advanced regularization. These enhancements could further improve predictive accuracy and provide deeper insights into the drivers of air quality changes in Europe.

Contributions

Trenton Carlson	engineer_features.py, data_splits.py(50%), base_models.ipynb, car_data_flatten.ipynb, checking_missing.ipynb, data_sizes.ipynb, EDA_final_features.ipynb, EDA.ipynb, sampling_locations.ipynb, window_size.ipynb
Hannah MacDonald	normalize_features.py, data_splits.py(50%), gradient_boost_model.ipynb, country_skew_analysis.ipynb, Corrected_EDA_HM.ipynb, XGB_NN_Model.ipynb,
Randy Tovar	ingestion.ipynb, cleaner.py, neural_network.ipynb, finding raw data

References

Chen, X., Li, J., & Wang, H. (2025). Air quality impacts of electric vehicle adoption: Evidence from real-world monitoring data. *Scientific Reports, 15*, Article 12345. https://doi.org/10.1038/s41598-025-92019-3

Zhang, Y., Wu, J., & Li, W. (2022). Comparative analysis of PM2.5 emissions from electric vehicles and internal combustion engine vehicles. *Science of the Total Environment, 825*, 153891. https://doi.org/10.1016/j.scitotenv.2022.153891

Appendix

A. Additional Modeling Experiment
   - *Stacked XGBoost and Neural Network*
     As an additional experiment, we implemented a stacked model in which XGBoost predictions were used as an input feature to a fully connected neural network. This approach first trained an XGBoost regressor on the training set and then generated predictions for both training and validation samples. These predictions were appended as an additional feature to the original dataset, which was then passed to a Keras feedforward neural network. The intention was to combine XGBoost's ability to capture non-linear tree-based relationships with the neural network's capacity to model complex transformations over the feature space. However, this hybrid model did not outperform the standalone XGBoost implementation, with higher RMSE on the validation set. Consequently, it was not included in our main results but is noted here as a potential avenue for further optimization.