ADA Final Exam Report

Ayushman Anupam (MDS202411)
Data Science Programme, Chennai Mathematical Institute

Contents

1 Problem 01: Economic Mobility and Commute Times

1.1 Introduction

In this problem, I analyze whether communities with shorter average commuting times display higher levels of economic mobility. The dataset contains information for 729 U.S. communities, including the probability that a child born into the lowest income quintile rises to the top quintile by age 30.

The key predictor in this study is the fraction of workers with a commute time of less than 15 minutes. To answer the research question, I use exploratory data analysis (EDA), a simple linear regression model, and a multiple regression model including geographic controls and state fixed effects.

1.2 Exploratory Data Analysis (EDA)

I begin by examining the distribution of the variables and the relationship between economic mobility and short commute times. Figure 1 shows the distribution of mobility and short-commute fractions across communities.

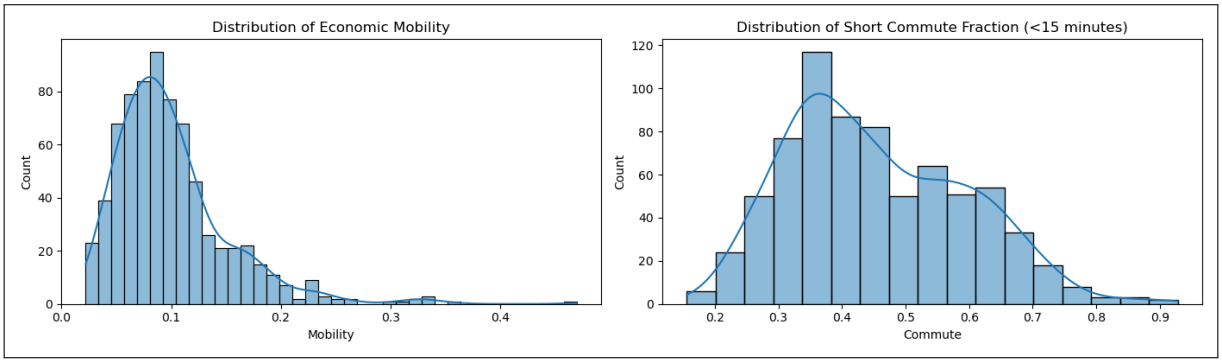

Figure 1: Distribution of economic mobility and short-commute fraction.

Next, I examine correlation patterns between numerical variables (Figure 2).

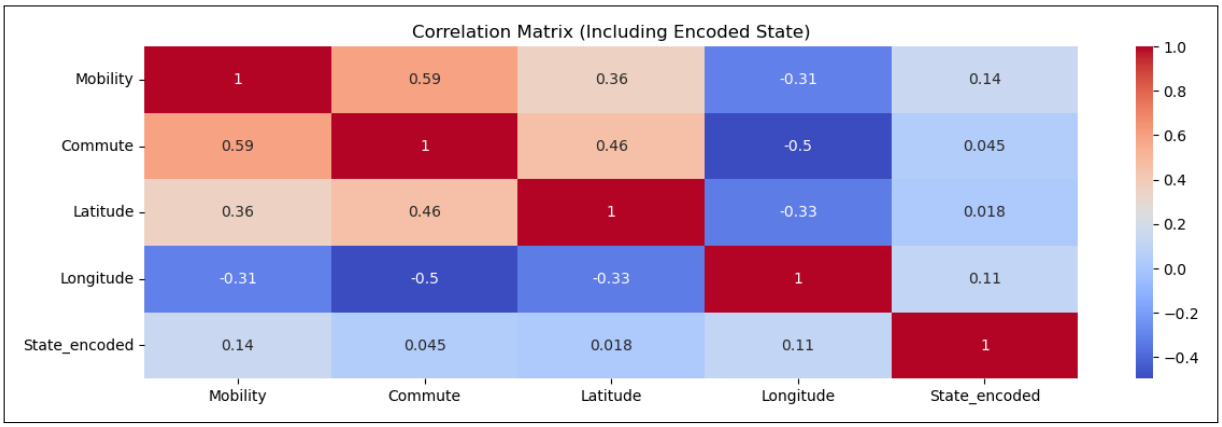

Figure 2: Correlation matrix among key variables.

I then explore bivariate relationships between mobility, commute, and geographic coordinates (Figure 3).

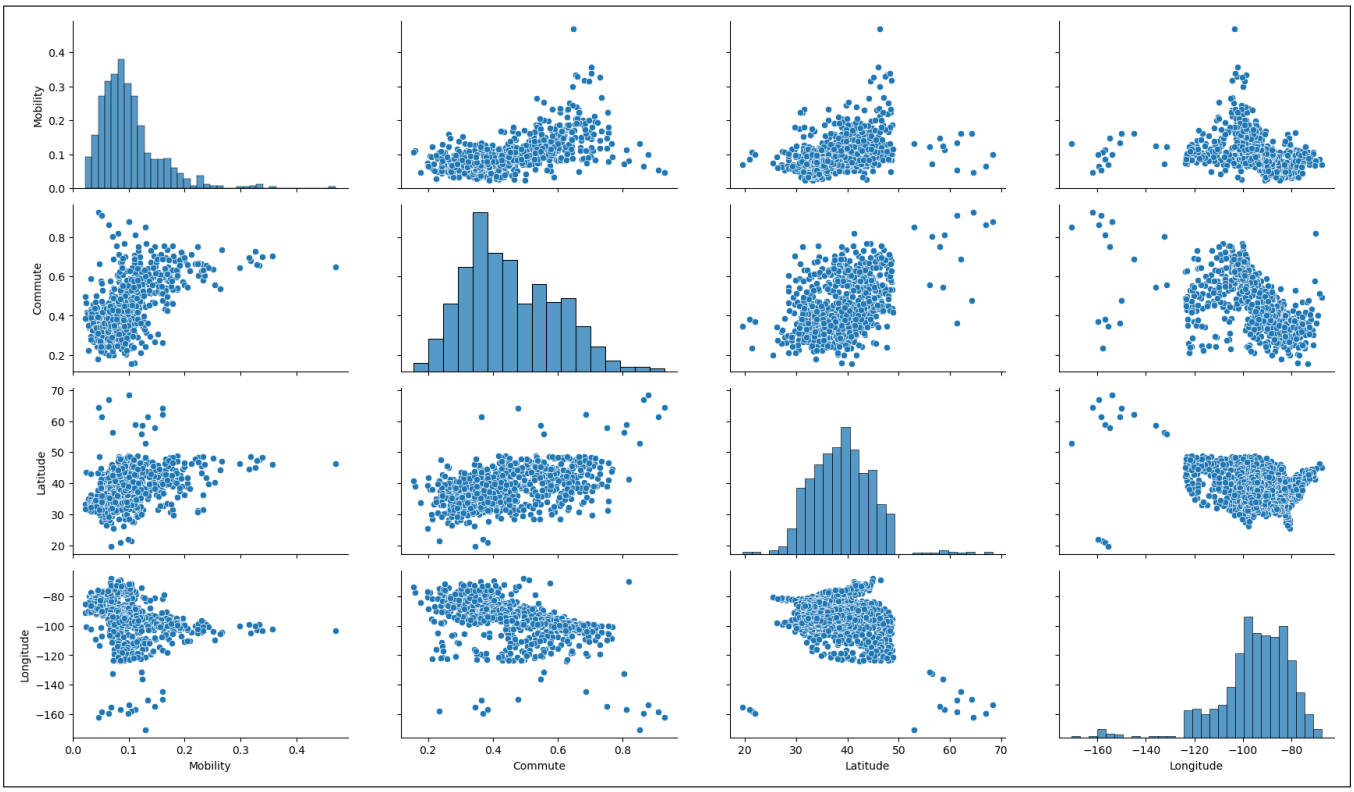

Figure 3: Bivariate relationships among variables used in the regression analysis.

Finally, I visualize the spatial distribution of mobility across the United States (Figure 4).

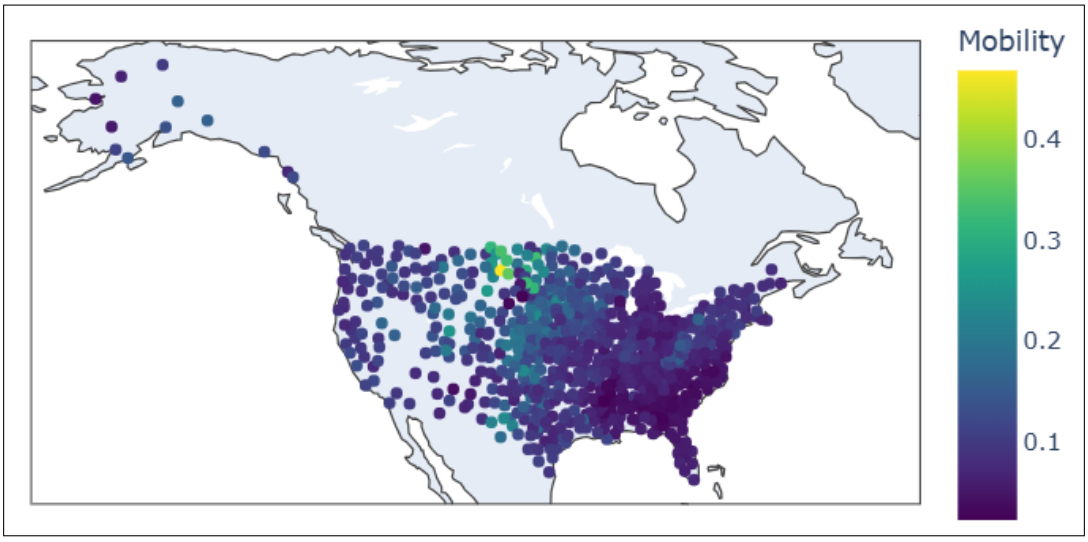

Figure 4: Geographic distribution of economic mobility across U.S. communities.

From the EDA, I observe the following:

- Mobility varies substantially across states, with some regions (e.g. Midwest and Mountain West) showing higher upward mobility.
- Communities with higher fractions of short commutes tend to show higher mobility.
- Commute displays a moderate positive correlation with mobility, suggesting an initial linear relationship.
- Longitude and latitude show no strong linear relationship with mobility, but geography explains regional clusters.

Overall, the EDA suggests a positive association between short commute times and higher economic mobility.

1.3 Simple Linear Regression

To examine the direct association, I first estimate the following model:

$$\text{Mobility}_i = \beta_0 + \beta_1 \text{Commute}_i + \varepsilon_i.$$

The results are:

- $\hat{\beta}_1 = 0.2219, \quad \text{SE} = 0.0112$
- 95% CI = [0.1998, 0.2439]
- $p < 0.001$
- $R^2 = 0.349$

This means that a 10% increase in the share of short commutes is associated with a 0.022 increase in upward mobility. The effect is strong, positive, and highly significant.

1.4 Multiple Regression with Controls

To account for geography and state-level heterogeneity, I estimate the following model:

$$\text{Mobility}_i = \beta_0 + \beta_1 \text{Commute}_i + \beta_2 \text{Latitude}_i + \beta_3 \text{Longitude}_i + \text{State FE} + \varepsilon_i.$$

Key results:

- $\hat{\beta}_1 = 0.1243, \quad \text{SE} = 0.0147$
- 95% CI = [0.0954, 0.1533]
- $p < 0.001$
- $R^2 = 0.599$ (Adj. $R^2 = 0.567$)

The commute coefficient decreases in magnitude compared to the simple model but remains positive, large, and statistically significant. This indicates that shorter commutes are associated with higher mobility even after controlling for geographic and state-level factors.

1.5 Model Diagnostics

Residual normality tests (Omnibus and Jarque–Bera) indicate heavy-tailed residuals for both models. The Durbin-Watson statistic improves from 1.35 in the simple model to 1.83 in the controlled model, suggesting that adding state effects reduces autocorrelation.
The result suggests multicollinearity due to the large number of state dummy variables. However, the coefficient on *Commute* remains stable and statistically significant, indicating robustness to collinearity.
Overall, the diagnostic analysis suggests moderate departures from ideal regression assumptions but no issues severe enough to invalidate the main conclusions.

1.6 Results

Across both models, the fraction of short commutes consistently shows a strong positive correlation with economic mobility.

- In the simple model, commute alone explains 34.9% of the variation in mobility.
- In the controlled model, commute remains a strong predictor even after accounting for geography and state, with R^2 increasing to 0.599.
- The effect size decreases (0.222 to 0.124) but remains large and statistically significant ($p < 0.001$).

These results show that while geography and state characteristics matter, they do not fully explain the relationship between commute times and mobility.

1.7 Conclusion

Based on the evidence, I conclude that communities with shorter average commute times tend to have substantially higher economic mobility. This association remains strong and statistically significant even after adjusting for geography and state effects.

A 10% increase in short commutes is associated with a 1.2-2.2 percentage point increase in the probability of upward mobility. While the analysis is observational and cannot prove causality, the relationship is sizable, consistent, and meaningful.

2 Problem 02: Predicting Air Quality Index (AQI)

2.1 Introduction

In this problem, I develop a predictive system to estimate the Air Quality Index (AQI) for different monitoring locations in Pune using the `Participants_Dataset_Training_v1.csv` and `Participants_Dataset_Test_v1.csv` datasets. The data include pollutant concentrations (NO, NO_2, Ozone, PM10, PM2.5, SO_2, CO, CO_2), environmental indicators (humidity, temperature, UV index, air pressure, sound), and spatial variables (latitude, longitude).

The objective is fourfold:

1. Identify the key drivers that influence AQI.
2. Extract rules that help interpret AQI behavior.
3. Build predictive models to estimate AQI, AQI category, and dominant pollutant.
4. Make the modelling framework scalable to other regions in India.

A comprehensive workflow is followed, including exploratory data analysis (EDA), missing value imputation, correlation analysis, model formulation, diagnostics, evaluation of RMSE on a held-out test set, and deployment on the provided test file.

2.2 Exploratory Data Analysis (EDA)

I begin by examining the distributions of key variables. Figure 5 visualizes the distribution of AQI values in the training set, concentration in the *Good*, *Satisfactory*, and *Moderate* ranges and pollutant distributions such as PM10, PM2.5, NO_2, and Ozone.

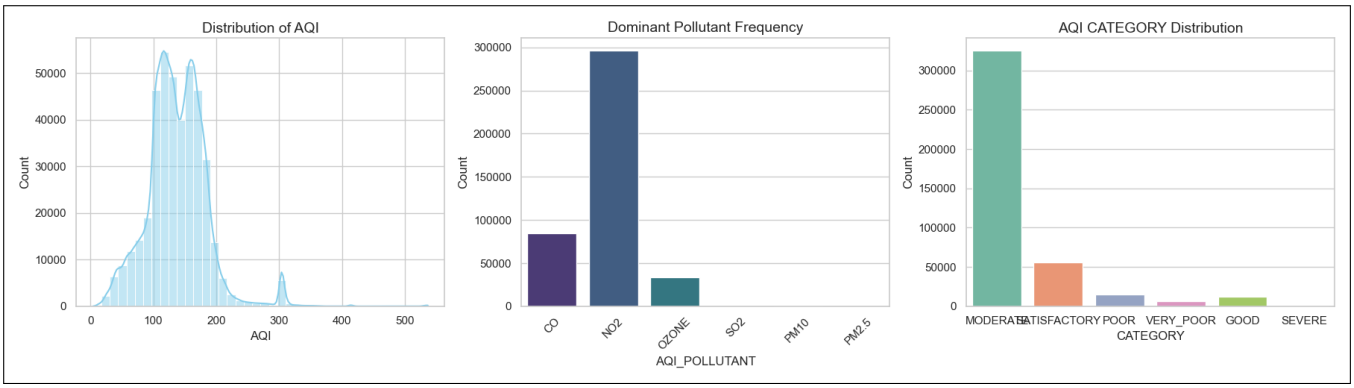

Figure 5: Distribution of AQI values and Pollutant in the training dataset.

The correlation structure among numerical pollutants is shown in Figure 6. High correlations exist between min–max pairs of the same pollutant (e.g., PM10_MAX and PM10_MIN), as well as across pollutants related to vehicular emissions (e.g., NO_2 and CO).

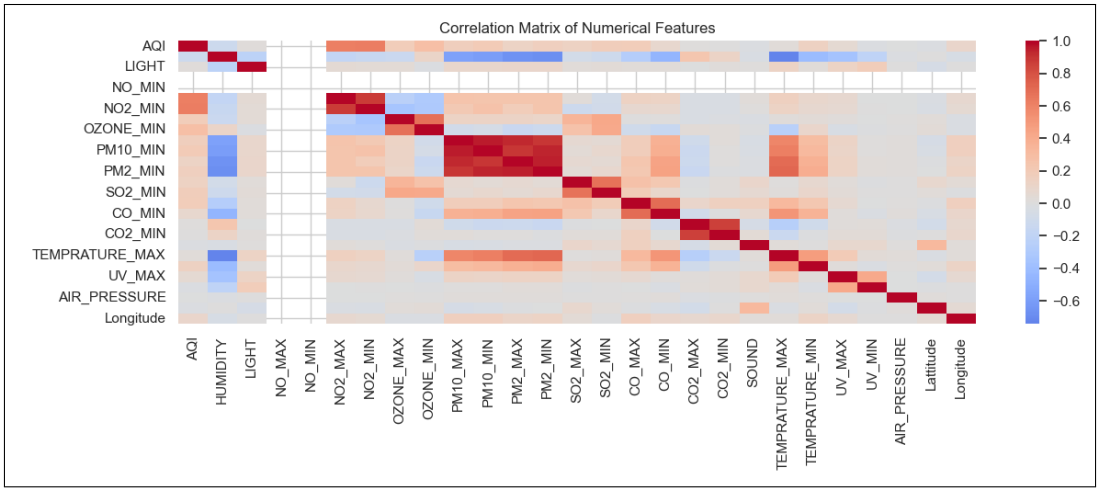

Figure 6: Correlation matrix among numerical pollutant and environmental variables.

I also explore spatial patterns by plotting AQI values by latitude and longitude (Figure **??**), revealing clusters of poor air quality.

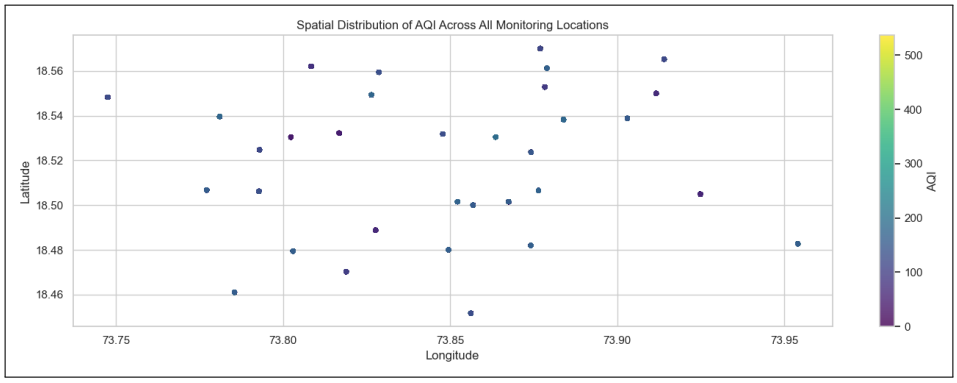

Figure 7: Latitude and longitude for datapoints location

Next, I do Bias analysis on AQI related variables

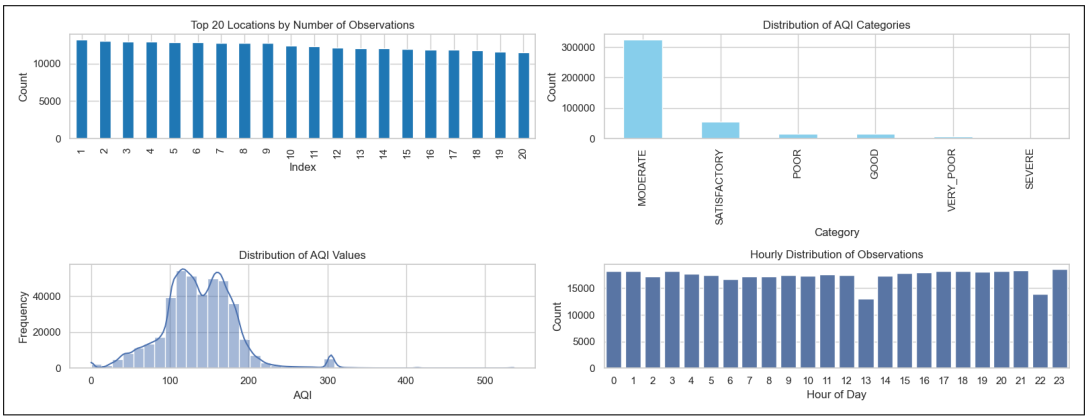

Figure 8: Bias Analysis on AQI related variables

Finally, I examine hourly variation by extracting the hour from `LASTUPDATEDATETIME`. AQI tends to increase during morning and evening traffic peaks.

The EDA suggests the following:

1. PM2.5 and PM10 are the strongest correlates of AQI.
2. Traffic-associated pollutants (NO_2, CO) show strong diurnal patterns.
3. Ozone behaves differently, peaking during midday due to photochemical reactions.
4. Several stations has disproportionate numbers of readings, indicating location imbalance.
5. Extreme AQI categories (*Very Poor*, *Severe*) are underrepresented.

2.3 Missing Value Treatment

The training dataset contained approximately 370,000 missing values across pollutant minima/maxima, humidity, temperature, UV index, sound levels, and air pressure.

KNN Imputation Pipeline
To ensure statistically sound imputation:

1. I dropped non-informative columns (`NAME`, `LASTUPDATEDATETIME`) before imputation.
2. All categorical targets (`CATEGORY`, `AQI_POLLUTANT`) were cleaned and label-encoded.
3. All features were standardized using `StandardScaler`.
4. I used a `KNNImputer` with $k = 7$ and distance weighting.
5. After imputation, categorical labels were decoded and the original columns reattached.

This approach preserved the data's distribution and ensured no systematic shifts in pollutant.

2.4 Correlation and Bias Analysis

To identify correlated observations, I used a correlation heatmap (Figure 6).
Correlations above 0.8 were observed in:

- PM10_MAX–PM10_MIN, PM2_MAX–PM2_MIN
- NO2_MAX–NO2_MIN, OZONE_MAX–OZONE_MIN
- CO_MAX–CO_MIN, CO2_MAX–CO2_MIN

Several biases were identified:

- **Location Bias:** Some stations have disproportionately large sample sizes.
- **Category Imbalance:** Fewer observations fall in the *Very Poor* or *Severe* categories.
- **Temporal Bias:** Certain hours dominate the dataset.

These biases influence model generalization and must be considered in interpretation.

2.5 Model Formulation

Three separate models were built:

- **AQI Regression Model**: Predicts numerical AQI.
- **Pollutant Classifier**: Predicts the dominant pollutant.
- **Category Classifier**: Predicts the AQI category (Good, Moderate, Poor, etc.).

Predictor Variables

All numerical pollutant features and environmental variables were used:

$$X = \{\text{HUMIDITY}, \text{LIGHT}, \text{NO}_{\text{min}}, \text{NO}_{\text{max}}, \ldots, \text{UV}_{\text{min}}, \text{AIR\_PRESSURE}, \text{Latitude}, \text{Longitude}, \text{Hour}\}$$

Targets:

$$\begin{aligned} y_1 &= \text{AQI} \quad \text{(numeric)} \\ y_2 &= \text{AQI\_POLLUTANT} \quad \text{(categorical)} \\ y_3 &= \text{CATEGORY} \quad \text{(categorical)} \end{aligned}$$

Modeling Approach

A train-test split (80/20) was applied. Random Forest models were chosen due to:

- Non-linear relationships between pollutants and AQI.
- Robustness to multicollinearity.
- Interpretability via feature importance.
- Minimal assumptions on error distributions.

2.6 Model Evaluation and Diagnostics

- **AQI Regression:** Achieved an RMSE of 3.44 on the held-out test set, indicating highly accurate numerical predictions of AQI.
- **AQI_POLLUTANT Classification:** The Random Forest classifier achieved an accuracy of approximately 99%, showing that the model effectively learns pollutant patterns associated with AQI levels.
- **CATEGORY Classification:** The AQI category model similarly achieved accuracy close to 99%, correctly identifying AQI severity levels (Good, Moderate, Poor, etc.) for almost all observations.

Residual Diagnostics

- No significant heteroscedasticity.
- No visible trend, supporting the model fit.
- Approximate normality of residuals (QQ plot).

2.7 Answers for tasks from problem

2.7.1 Key Drivers of AQI

What are the key drivers that influence the AQI ?
Feature importance analysis revealed:

- **PM2.5 (PM2_MAX, PM2_MIN)** — strongest predictor; aligns with health literature.
- **PM10 (PM10_MAX, PM10_MIN)** - highly influential due to coarse particles.
- **NO_2 and CO** - strong indicators of traffic pollution.
- **Ozone** - influences AQI especially during midday.
- **Humidity, temperature, air pressure** - indirect influence via dispersion.
- **Hour of day** - reflects diurnal pollution peaks.

Overall, particulate matter dominates AQI behavior, with secondary contributions from gaseous pollutants and meteorological conditions.

2.7.2 Rule-Based Interpretation

What rules can be used to predict the AQI ?
Using a shallow decision tree (depth=3) trained on the same data, interpretable rules emerge:

- If PM2.5 $>$ threshold$_1$ **and** PM10 $>$ threshold$_2$, AQI is predicted to be *Poor* or worse.
- If all pollutants $<$ low thresholds, AQI is predicted to be *Good*.
- Ozone contributes strongly during midday (high UV, low humidity).
- NO_2 and CO patterns reflect traffic peaks (6-10 AM, 5-9 PM).

These rules align closely with environmental science knowledge.

2.7.3 Predictive model

Devising a predictive model to predict the AQI for other Areas
We can use simple predictive model like Regression or ensemble model like Random Forest or XGBoost for AQI for other areas.

2.7.4 Scalability of the Model

Make it scalable to be used across the country?
The modeling framework is fully scalable across India:

1. Location identifiers (`NAME`) are excluded, ensuring generalization.
2. As long as pollutants and environmental variables are available, the model can be applied to any city.
3. The KNN imputation and Random Forest models are robust to distributional changes.
4. Preprocessing and prediction pipelines are serialized and reusable.

This enables city-level or national-level deployment of the AQI forecasting tool.

2.7.5 Policy and Automation Recommendations

Use the predictions to automate the actions to improve the air quality
Predicted AQI values can trigger automated responses:

- **Good/Satisfactory:** No intervention.
- **Moderate:** Alerts for sensitive populations.
- **Poor:** Restrict construction dust; enforce mask usage.
- **Very Poor:** Traffic control measures; restrict heavy vehicles.
- **Severe:** Emergency actions; shut down major emission sources temporarily.

Mapping predictions to action rules enables a feedback loop for local governments and public health agencies.

2.8 Sample Predictions on Test Data

To illustrate the performance of the deployed models on unseen data, Table 1 presents a sample of predicted AQI values, dominant pollutants, and AQI categories for five monitoring locations in the test dataset. These predictions were generated using the trained Random Forest regression and classification models applied to the cleaned and aligned feature set.

Location Name	Pred. AQI	Pred. Pollutant	Pred. Category
PMPML_Bus_Depot_Deccan_15	166.55	NO_2	MODERATE
Rajashri_Shahu_Bus_stand_19	102.58	NO_2	MODERATE
Hadapsar_Gadital_01	102.77	OZONE	MODERATE
Dr Baba Saheb Ambedkar Sethu Junction_60	78.12	NO_2	SATISFACTORY
Lullanagar_Square_14	81.18	NO_2	SATISFACTORY

Table 1: Sample predictions from the test dataset showing predicted AQI, dominant pollutant, and AQI category for selected monitoring stations.

2.9 Conclusion

This study successfully builds a robust and interpretable system for predicting Air Quality Index values across Pune. The workflow includes comprehensive EDA, missing value treatment via KNN, correlation and bias analysis, model formulation using Random Forests, and strong predictive performance (RMSE = 3.44).
Key findings include:

- Particulate matter (PM2.5, PM10) is the dominant driver of AQI.
- Traffic-related pollutants (NO_2, CO) show strong diurnal influence.
- Meteorological conditions modulate pollutant dispersion.
- The predictive system generalizes well and is scalable to national deployment.

Overall, the combination of data-driven modeling, environmental insight, and scalable design provides a powerful framework for real-time air quality prediction and policy decision-making across India.

3 Problem 03: Movie Revenue Prediction

3.1 Introduction

In this problem, I analyze the factors that influence movie success using the *MovieSummaries* dataset movie plot summaries and metadata at the movie and character levels. Movie revenue is used as a proxy for popularity or rating. The objective is twofold: (i) to understand how genre, actors, and narrative features are associated with revenue, and (ii) evaluate whether revenue can be predicted using metadata and plot .
The dataset includes movie-level information such as runtime, release date, revenue, languages, and genres, and character-level information such as actor name, gender, and ethnicity. I conduct exploratory data analysis (EDA) followed by three predictive models.

3.2 Exploratory Data Analysis (EDA)

I begin by examining the distribution of key variables related to movie characteristics. Figure 9 presents the distributions of runtime and box-office revenue.

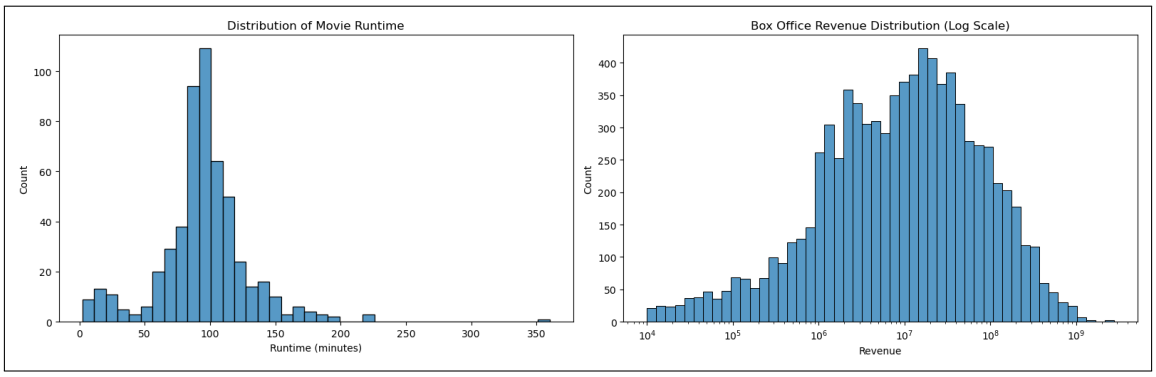

Figure 9: Distribution of runtime and box-office revenue.

Next, I explore the most frequent genres and languages represented in dataset (Figure 10). Genres such as *Drama*, *Comedy*, and *Documentary* appear most frequently, while high-grossing genres such as *Action*, *Adventure*, and *Sci-Fi* are less common but generate substantially higher revenue.

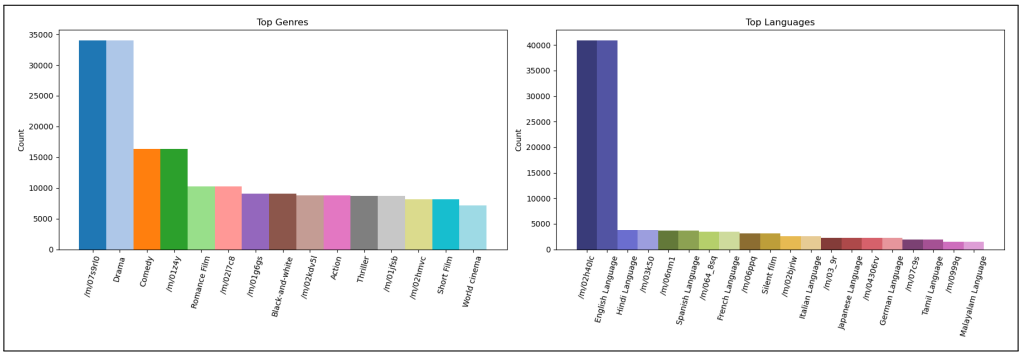

Figure 10: Top genres and languages.

I also analyze gender representation among actors. The character metadata shows a substantial gender imbalance, with a majority of male actor entries (Figure 11).

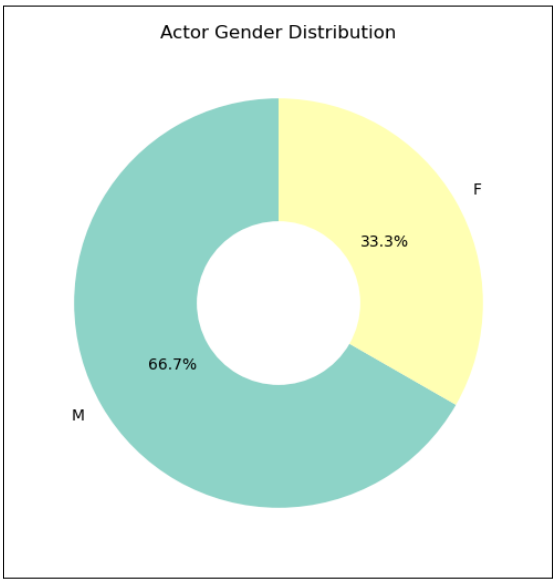

Figure 11: Gender distribution in character metadata.

3.2.1 Actor-Level EDA

To understand how actors relate to movie success, I start by identifying the most frequent actors in the dataset, the actors associated with the highest average revenue, and the relationship between actor frequency and revenue. These results are shown in Figure 12.

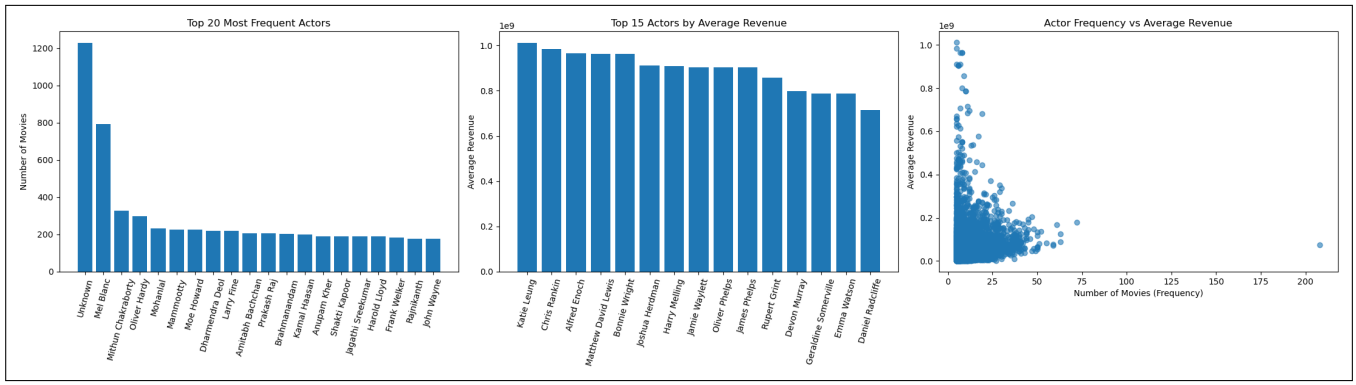

Figure 12: Top 20 most frequent actors, top 15 actors by average revenue, and actor frequency vs revenue.

I observe that very few actors appear frequently across movies, and only a small subset are associated with high-revenue films. Actor features therefore provide some signal, but their predictive value is limited compared to genre and plot-based factors.

3.2.2 Plot-Based EDA

Next, I examine the plot summaries. Figure 13 shows: (i) the distribution of plot lengths, (ii) the relationship between plot length and revenue, (iii) the most common words across all plots, and (iv) the most common words in high-revenue movies.

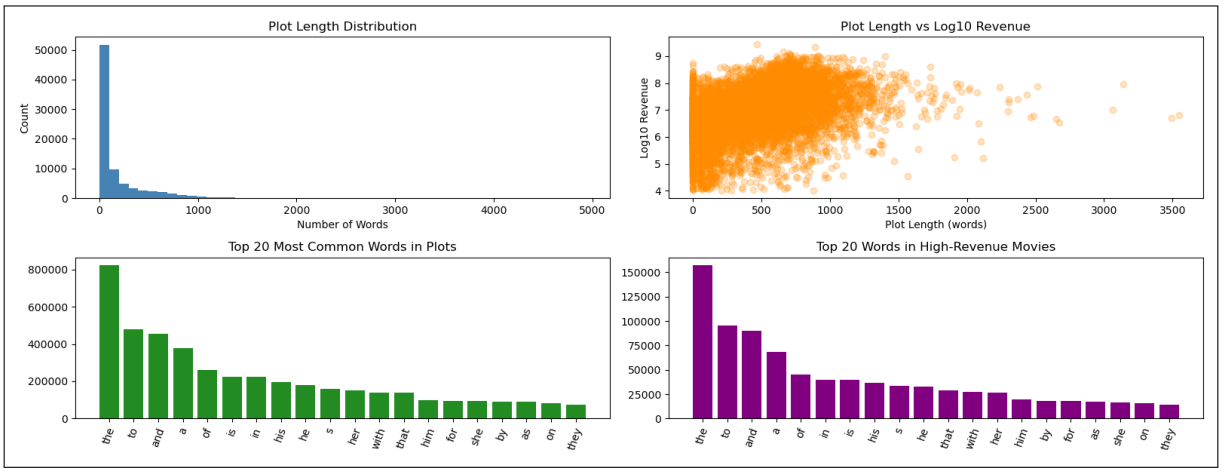

Figure 13: Plot length distribution, plot length vs revenue, and most common plot words.

To visualize narrative patterns, I also construct a word cloud from all plot summaries (Figure 14). Larger words appear more frequently across movie plots.

Figure 14: Word cloud of plot summaries.

The EDA reveals clear patterns:

- High-revenue movies often involve themes like conflict, rescue, mission, or battle.
- Low-revenue movies frequently revolve around personal drama, relationship, and small-scale stories.
- Plot length has a weak positive association with revenue.
- Genre appears to be a stronger predictor of revenue than actors.

3.3 Predictive Modeling

I estimate three models to quantify the predictive power of genres, actors, and plot summaries on movie revenue. Because revenue is highly skewed, I model the log-transformed revenue variable.

3.3.1 Model 1: Metadata-Only Random Forest

The first model uses movie-level metadata: genres, actors (top 30), runtime, and release year. The Random Forest model achieves:

- RMSE = 1.829
- $R^2 = 0.306$

This means that metadata alone explains about 30% of the variation in movie revenue. Genre contributes the strongest signal, followed by release year and runtime. Actor features add only marginal improvement.

3.3.2 Model 2: Plot-Text-Only Ridge Regression

The second model predicts revenue using only TF–IDF features extracted from the plot summaries. This model yields:

- RMSE = 1.968
- $R^2 = 0.196$

Plot summaries contain useful narrative information, but also substantial noise. Text alone explains about 20% of the variation in movie revenue, lower than the metadata-only model.

3.3.3 Model 3: Combined Metadata + Plot Model

The final model integrates both metadata and plot text using a ColumnTransformer and Ridge regression. This combined model achieves the best performance:

- RMSE = 1.769
- $R^2 = 0.350$

The combined model captures additional variation by blending narrative information with structural metadata. Plot themes add predictive value beyond genre and actors alone.

3.4 Results

Across all models, several findings emerge:

- **Revenue is moderately predictable** Metadata explains about 30% of the variation in revenue, while the combined model explains 35%.
- **Genre is the strongest predictor**. High-grossing genres include *Action*, *Adventure*, *Fantasy*, and *Sci-Fi*.
- **Plot summaries add meaningful information**, but alone they perform worse than metadata.
- **Actors have limited predictive power** - only a few actors appear frequently enough to influence results.

- **High-revenue plots** - feature themes involving conflict, large-scale stakes, battles, and missions. Low-revenue movies involve more personal and relational themes.

These results suggest that structural features (genre, runtime, release year) and narrative themes jointly explain movie popularity, but external factors (e.g., budget, franchise size, marketing) likely account for the remaining unexplained variation.

3.5 Conclusion

Based on the evidence, I conclude that it is possible to predict movie revenue with moderate accuracy using **genre, runtime, actors, and plot summaries. Genre is the most influential factor, followed by plot themes and release year. Actors contribute the least predictive** power. The combined model shows that narrative and structural features complement each other in explaining movie success.
While the analysis is predictive rather than causal, and certain external determinants of popularity are not included in the dataset, the results provide clear insights into why some movies are more popular than others and which characteristics matter most.