# Flight Data:
# Predicting Speed From Altitude

Justin Short • 02.17.2025

# Overview

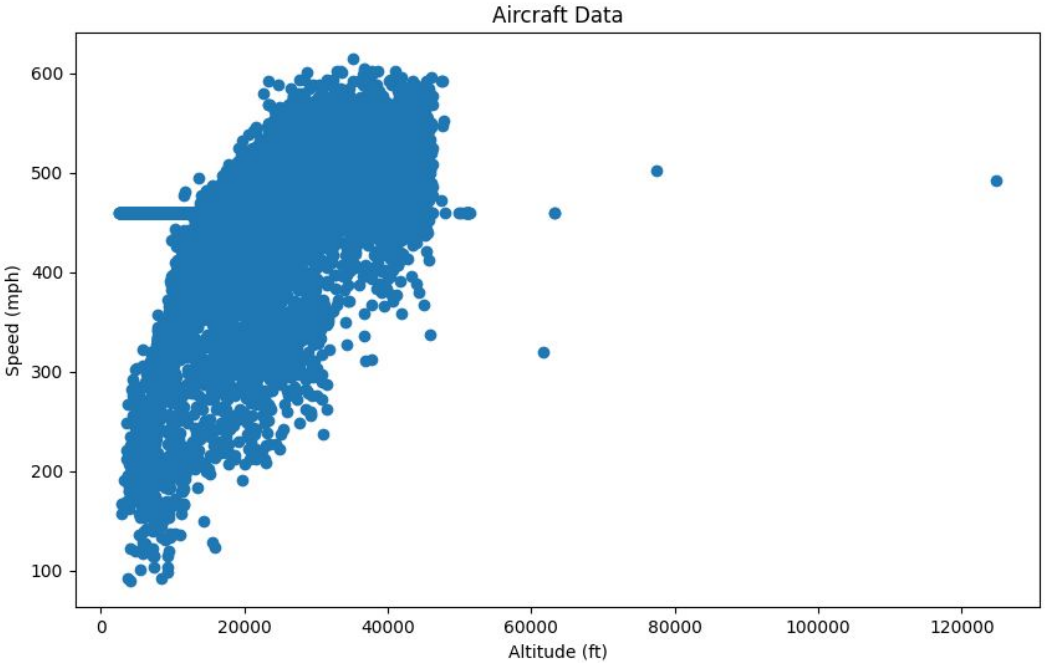
**Dataset**

https://www.kaggle.com/datasets/brianwarner/aircraft-data-from-nov-2022-through-dec-31-2022

**Key Points**

- Data Processing
- Model Performance
- Final Decisions

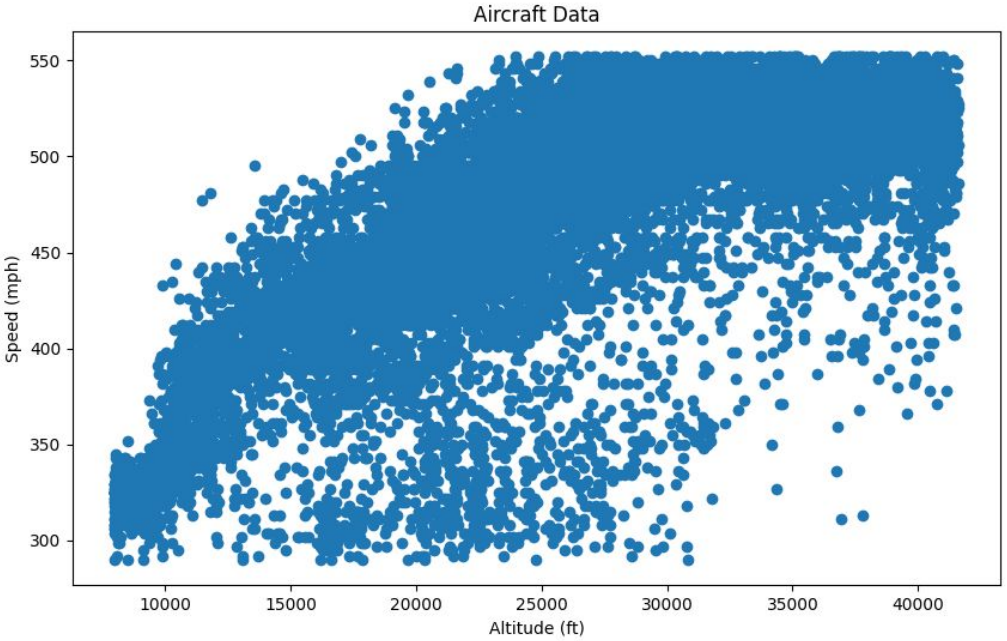
# Data Before Cleaning




Issues:

- Outliers
  - Linear Regression is sensitive to outliers
- Odd line of data points near 460 miles per hour
  - Likely a missing value replacement (close to mean)
  - Omitting these values significantly improved model performance

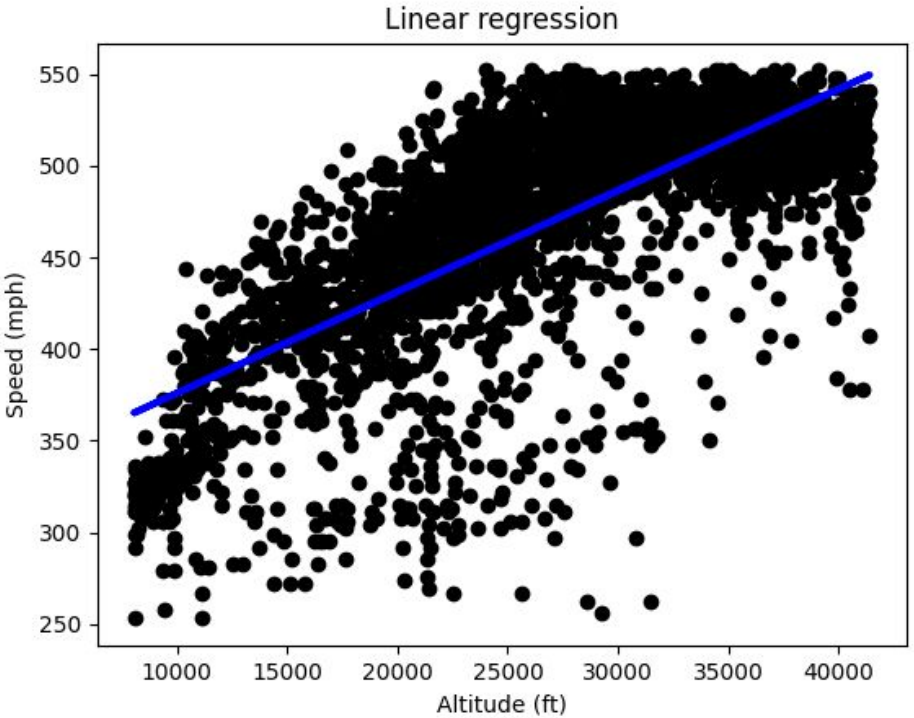
# Processed Data




Observations:

- A clear trend in the data
- A curve can be fitted to this
  - Polynomial of degree 2 or 3 should be able to produce a decent best fit line

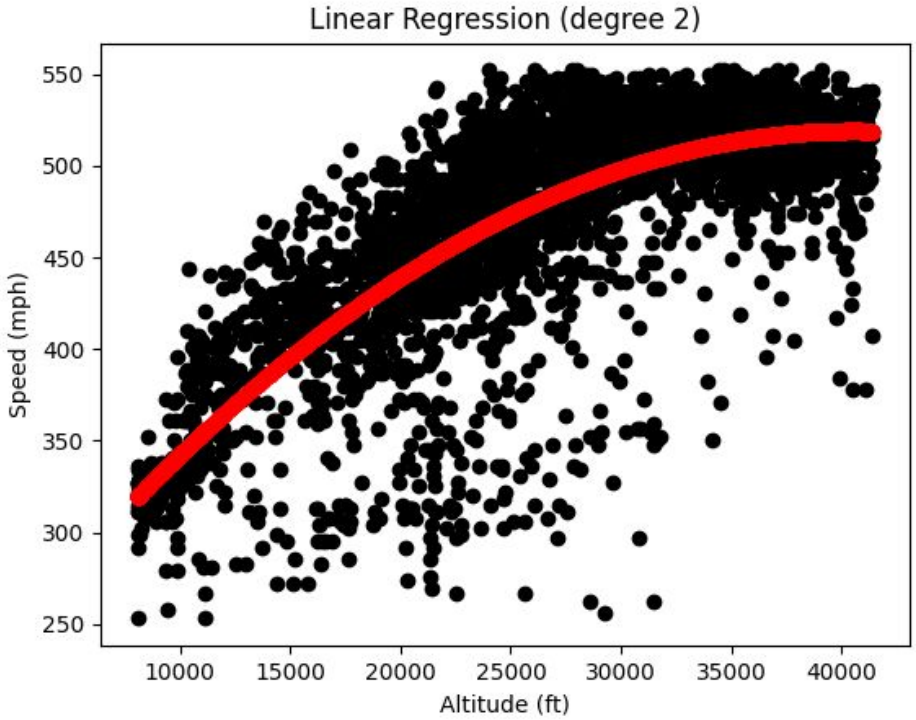
# Linear Regression (Degree 1 Polynomial)




Metrics:

- RMSE: 40.8
- R2: 0.575

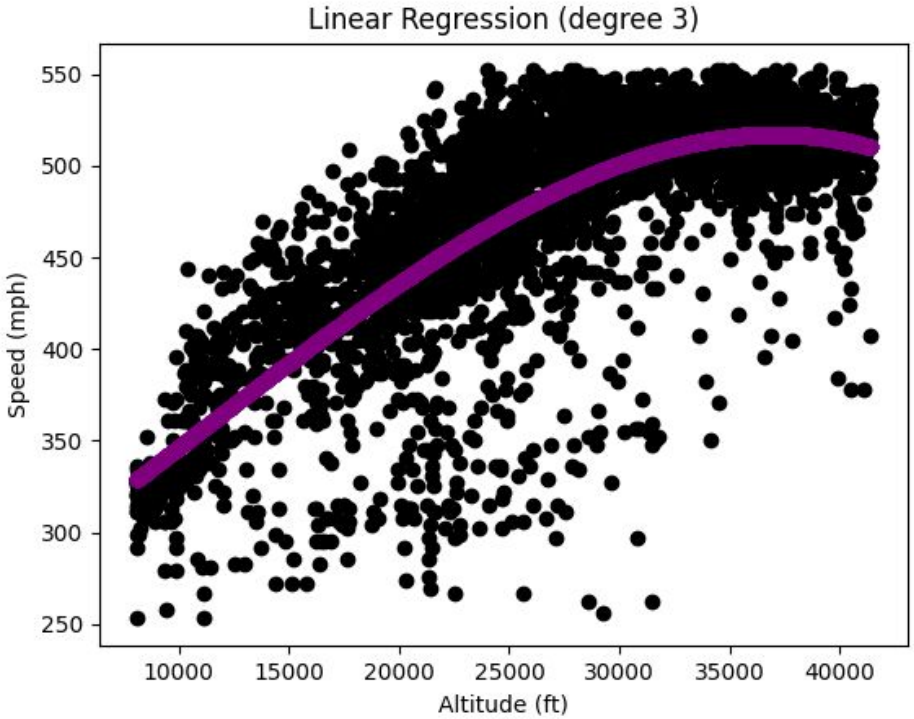
# Linear Regression (Degree 2 Polynomial)




Metrics:

- RMSE: 37.84
- R2: 0.635

# Linear Regression (Degree 3 Polynomial)




Metrics:

- RMSE: 37.68
- R2: 0.638

# Failed Improvement Techniques

Scaling/Normalization:

- Since there was only a single input feature, scaling

Regularization:

- Neither L1 (Ridge) or L2 (Lasso)  regularization improved the performance

Ensemble Models:

- Gradient Boost and Random Forest models were tested,
- But improvements were minimal
    - And were more likely to be overfitted to the dataset

# Results

Data Cleaning

- Removed outer 2% of data points
  - Both for mph and altitude
- Removed all data points where mph = 460
  - While some of these would have been real data points
  - Many of them were likely replacing missing values

Model

- Polynomial of degree 3

Github

- https://github.com/jnshort/FlightDataML