Executive Summary

NYC Taxi Fare Prediction — A Deployable Linear Regression Model for Real-Time Fare Estimates

🧭 Project Context

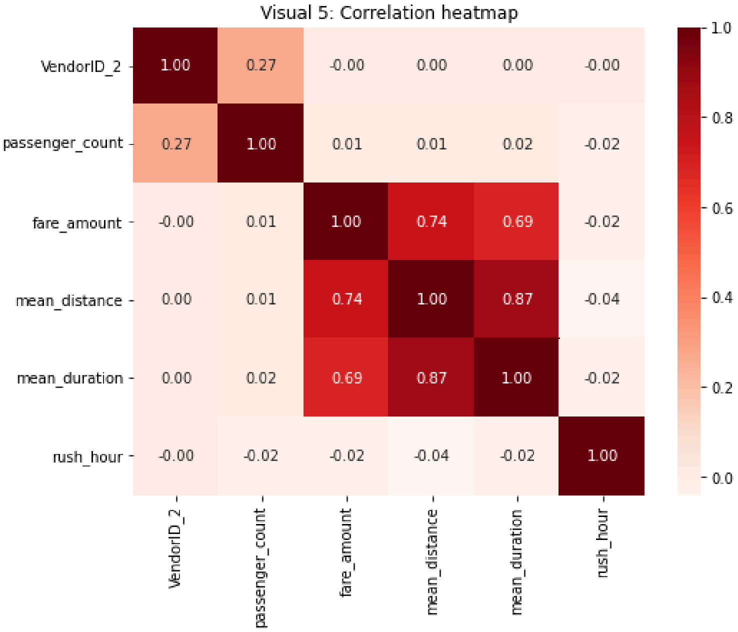
The NYC TLC commissioned a model to predict fares **before each ride**, aiming to improve transparency and enable smarter pricing and auditing..

🛠️ Model Strategy

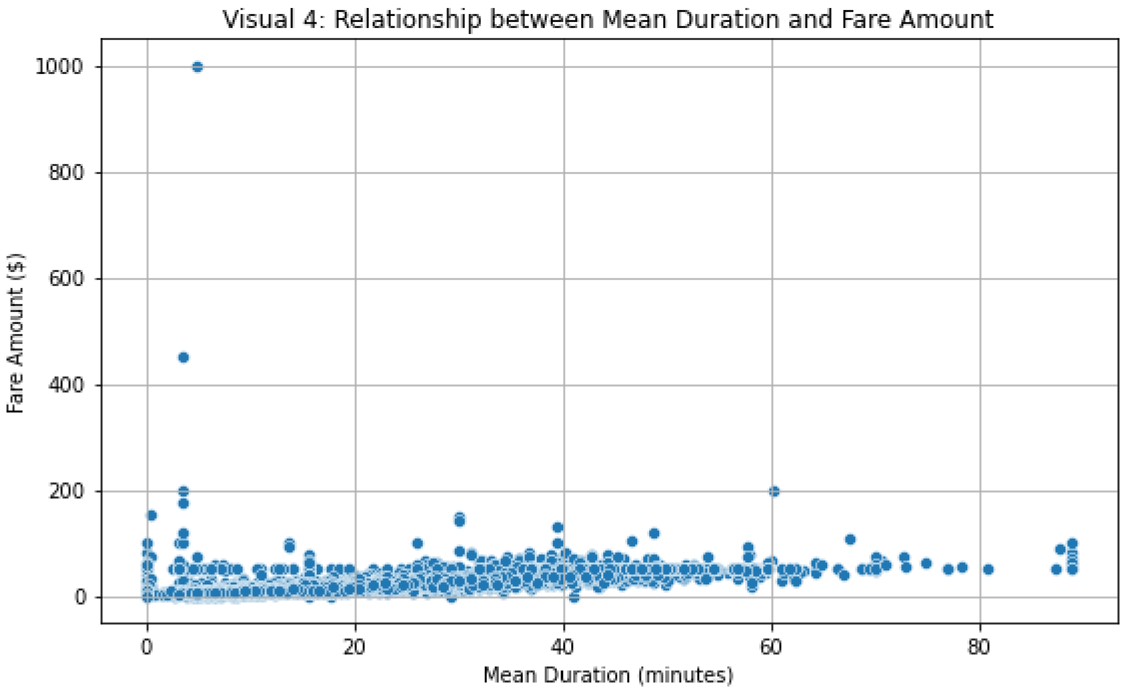
A multiple linear regression model was developed using historical TLC trip data. After data cleaning, feature engineering, and collinearity checks, five pre-ride predictors were selected: `mean_distance`, `mean_duration`, `rush_hour`, `passenger_count`, and `VendorID_2`.

Post-ride features (e.g., tip_amount, total_amount) were excluded to keep the model production-ready. Model evaluation used R² and residual diagnostics.

A correlation heatmap (**Visual 5**) confirmed the relationships between selected features and the fare amount, justifying their inclusion in the model.

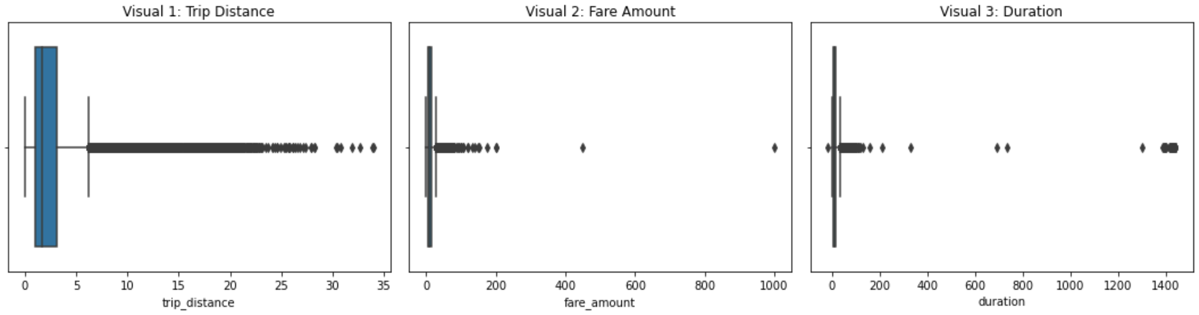


For example, mean_duration showed a strong positive trend with fare amount (**Visual 4**), making it a key predictor.


Visual 4: Relationship between Mean Duration and Fare Amount

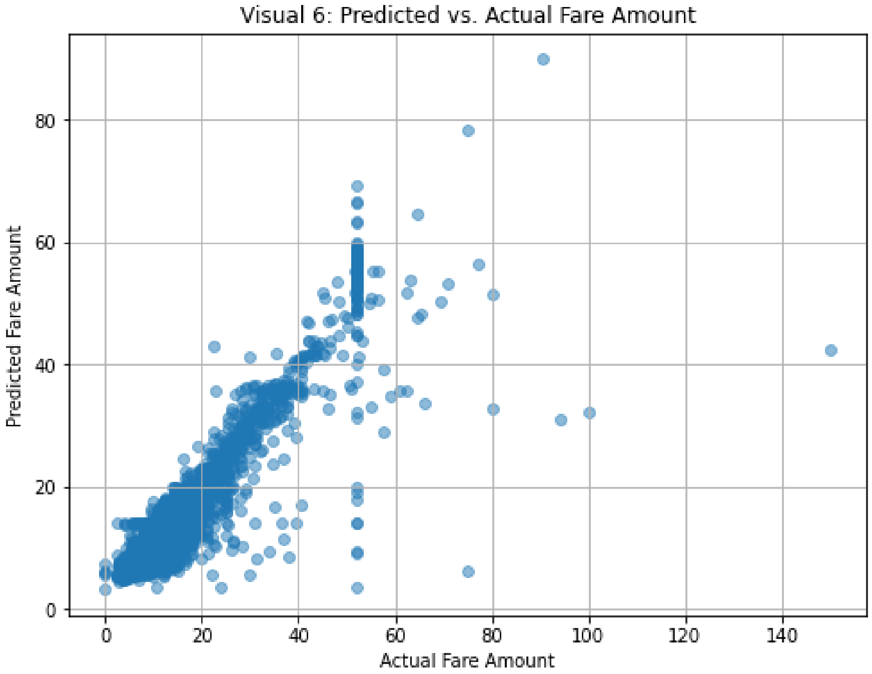
📊 Key Findings

- Distance and duration were the most influential fare predictors



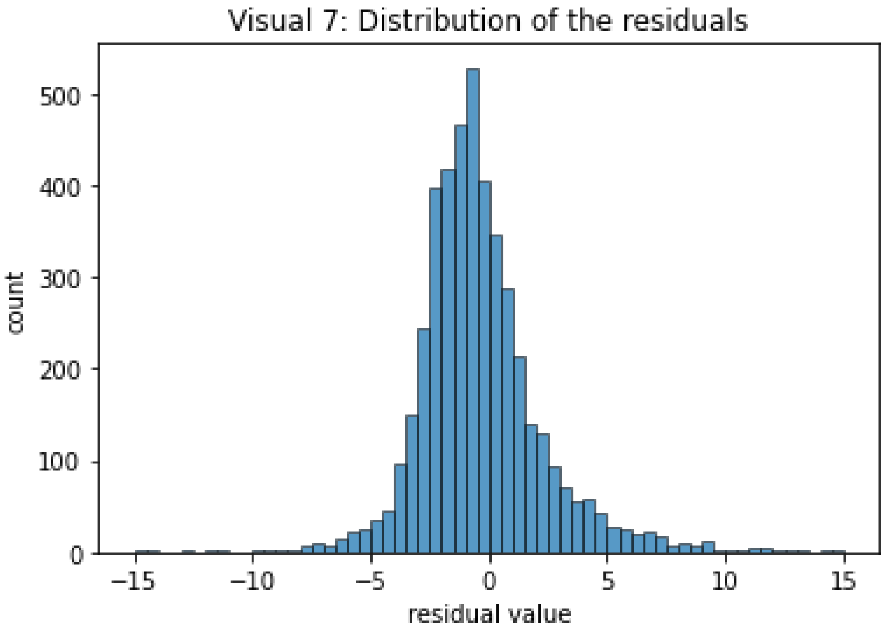
➡️ This aligns with initial distribution insights (**Visuals** 1–3), where skewed patterns in **trip distance** (Visual 1), **fare amount** (Visual 2), and **duration** (Visual 3) highlighted the need for transformation and cleaning.

- The model achieved an R² of 0.87 on the test set

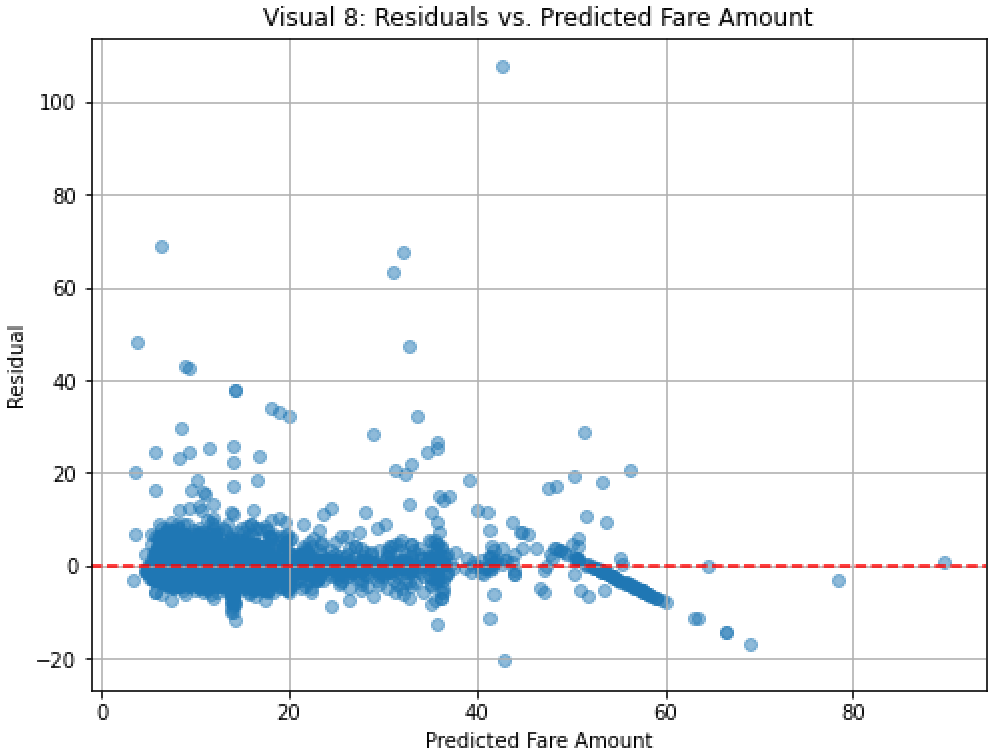


➡️**Visual 6** illustrates the strong alignment between predicted and actual fares, supporting the model's reliability.

- Residuals were approximately normally distributed (**Visual 7**), supporting assumptions of linearity and unbiased error terms.



- Feature redundancy (e.g., between trip distance and mean distance) was avoided to reduce multicollinearity
- Fares increased moderately during rush hours, while passenger count had negligible impact.



➡️ Finally, **Visual 8** shows the residuals spread evenly across predicted values, which supports the assumptions of linearity and constant variance.

🏁 Business Impact

This model can power pre-ride fare estimates across NYC taxis, helping passengers make informed decisions and enabling TLC to detect inefficiencies. Its linear structure enhances transparency, making it easy to audit and justify fare estimates in regulatory contexts.

📌 Next Steps:

- Deploy the model to estimate fares in real time
- Monitor for data drift and retrain periodically
- Add contextual data like weather or traffic to improve accuracy
- Use the model to simulate pricing policy changes before rollout

👤 Author

Jonatan Zemedebrhan

Google Advanced Data Analytics Certificate – Course 5

NYC Taxi Fare Prediction *Case Study*