TAXI FARES

Objective: To develop a predictive model that accurately forecasts yellow taxi fare amounts based on key determinants. This analysis aims to understand the influence of certain factors on the overall taxi fare, thereby providing insights for taxi companies to optimize pricing strategies and enhance service efficiency.

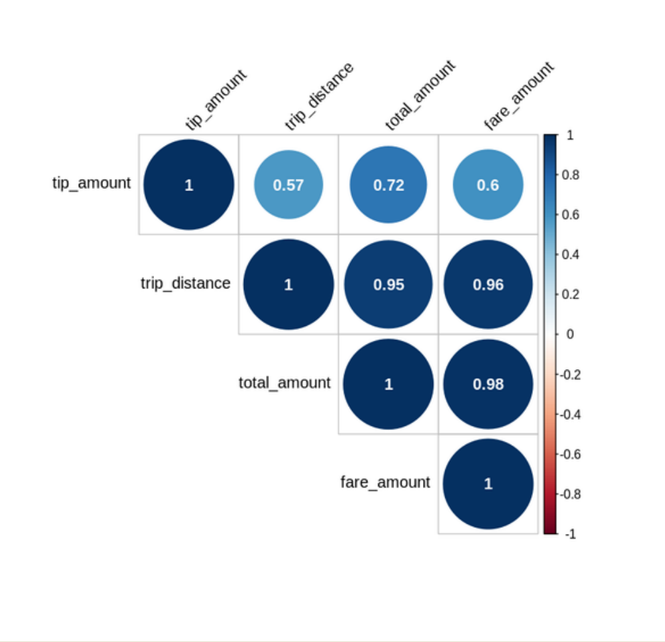
KARAN KARTHIK

DATASET OVERVIEW AND EDA INSIGHTS

The dataset contains detailed records of Yellow Taxi trips in New York City from the year 2023. It includes information about trip distances, pickup and dropoff locations, times, fare amounts, tips, and additional charges, among other variables. By examining these key factors, the primary influencers of taxi fare amounts are determined.

"VendorID"	"tpep_pickup_datetime"	"tpep_dropoff_datetime"
"passenger_count"	"trip_distance"	"RatecodeID"
"store_and_fwd_flag"	"PULocationID"	"DOLocationID"
"payment_type"	"fare_amount"	"extra"
"mta_tax"	"tip_amount"	"tolls_amount"
"improvement_surcharge"	"total_amount"	"congestion_surcharge"
"Airport_fee"		

DATA SET FEATURES



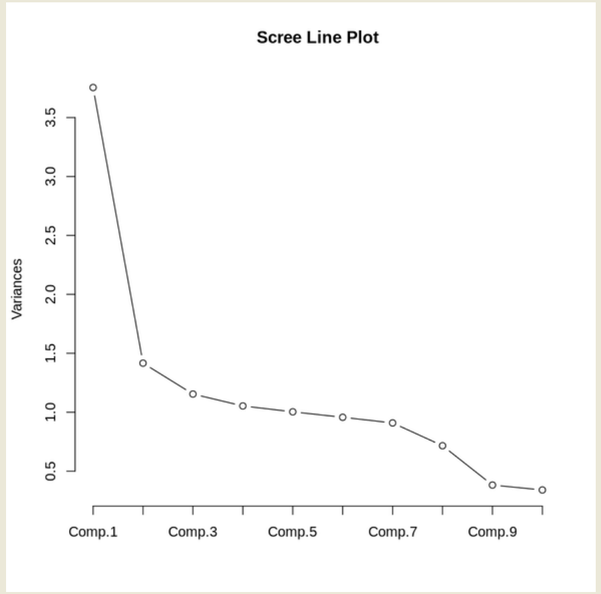
CORRELATION MAP

The correlation heatmap highlights the strong positive correlation between the trip distance and total amount (**0.95**), emphasizing that longer trips typically incur higher charges. Similarly, fare and total amount share a significant positive correlation (**0.98**), pointing to the fare's substantial role in determining the total charge. The tip amount's positive correlation with the total amount (**0.72**) suggests that as fares increase, so do tips, but with a less pronounced relationship compared to fare and total amount. Moreover, the correlation between the trip distance and tip amount (**0.57**) indicates a moderate association, suggesting that the distance of the trip influences tipping behavior to a certain extent.

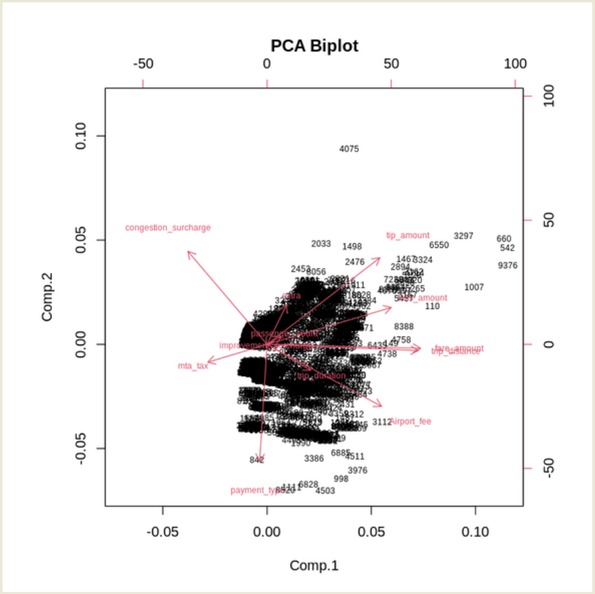
INSIGHTS FROM PCA AND INITIAL REGRESSION MODEL

Performing Principal Component Analysis (PCA), the screen line plot revealed that cumulative variance significantly plateaus after the first five components. This suggested that the dimensionality of our dataset could effectively be reduced to these principal components to capture the most substantial variance within the data.

Additionally, examining the biplot revealed that certain original variables, specifically *mta_tax*, *fare_amount*, *tip_amount*, *congestion_surcharge*, and *Airport_fee*, exhibit high variance and are thus likely to contribute significantly to these principal components.



PCA SCREE PLOT



PCA BI PLOT

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.0004903  0.0005169  -0.949    0.343
mta_tax            0.0081930  0.0006860  11.943   <2e-16 ***
tip_amount         0.1791649  0.0006679 268.250   <2e-16 ***
fare_amount        0.8056534  0.0023170 347.713   <2e-16 ***
tolls_amount       0.1082760  0.0007618 142.138   <2e-16 ***
trip_distance      0.0000305  0.0023264   0.013    0.990
congestion_surcharge  0.0321350  0.0006338  50.703   <2e-16 ***
Airport_fee        0.0180231  0.0007441  24.220   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04247 on 6744 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9982
F-statistic: 5.397e+05 on 7 and 6744 DF,  p-value: < 2.2e-16
```

INITIAL REGRESSION MODEL

After analyzing the initial regression model informed by the PCA insights, it became evident that certain variables, notably *trip_distance*, had minimal influence on predicting *total_amount* (p-value = **0.990**), suggesting its insignificance in the model. Despite this, the model exhibited a high R-squared value of **0.9982**, indicating that it could explain a vast majority of the variance in *total_amount*. This highlighted the need for model refinement. While the overall fit was strong, not all predictors were contributing equally.

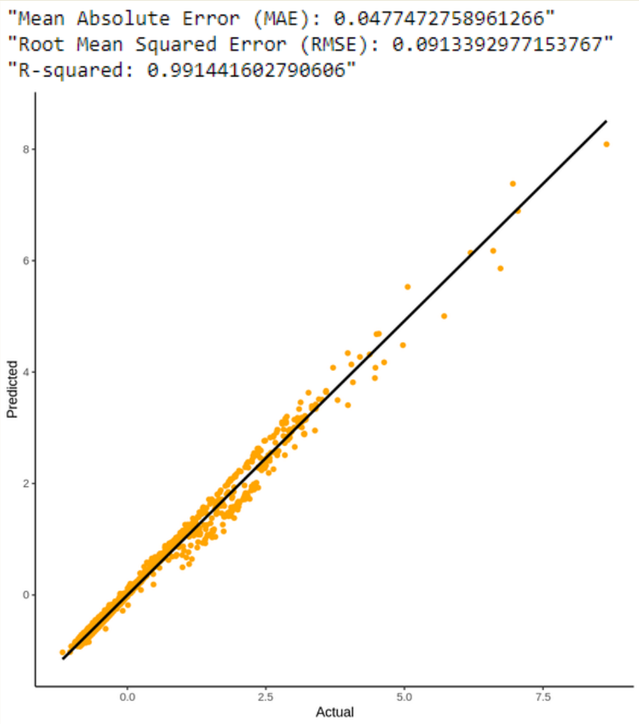
RESULTS

The final regression model, focusing on *trip_distance*, *fare_amount*, *tip_amount*, and *trip_duration* as predictors for *total_amount*, showcased robust predictive capabilities, as evidenced by their p-values being well below the threshold of **0.05**. Key findings include statistically significant relationships for all predictors, with *fare_amount* and *tip_amount* demonstrating particularly strong positive impacts on the total fare. The model achieved a remarkable fit, with a Multiple R-squared value of **0.9914**, indicating it explains nearly **99.14%** of the variance in taxi fares.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.000783   0.001136  -0.689  0.49082
trip_distance  0.072852   0.004278  17.029  < 2e-16 ***
fare_amount    0.786105   0.004396 178.834  < 2e-16 ***
tip_amount     0.206541   0.001404 147.132  < 2e-16 ***
trip_duration -0.003315   0.001110  -2.987  0.00283 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09337 on 6747 degrees of freedom
Multiple R-squared:  0.9914,    Adjusted R-squared:  0.9914
F-statistic: 1.941e+05 on 4 and 6747 DF,  p-value: < 2.2e-16
```

FINAL REGRESSION MODEL



TEST DATA VALUES PLOT

The plot on the model’s test values comparing actual to predicted taxi fares demonstrates the model's precision, with a linear fit line closely mirroring the data points and indicating very few discrepancies. Performance metrics further validate the model's accuracy with a Mean Absolute Error (MAE) of **0.0477**, a Root Mean Squared Error (RMSE) of **0.0913**, and an R-squared value of **0.9914**. Overall, these metrics underscore the model's utility in yellow taxi fare prediction, highlighting its effectiveness in real-world applications.