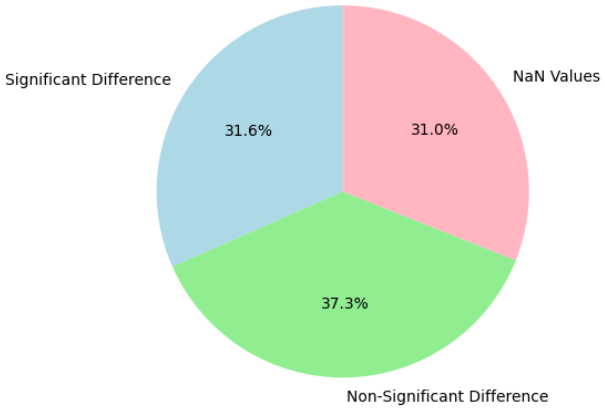
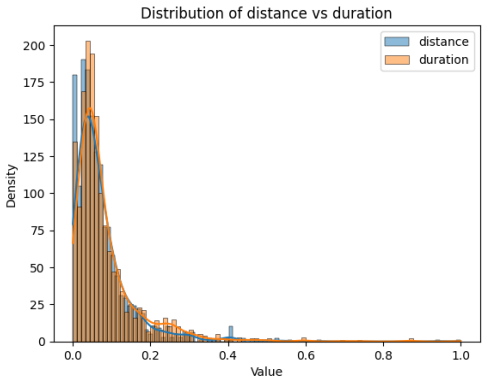
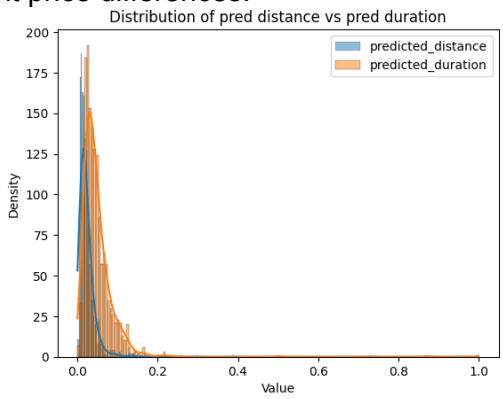
Ride Data Analysis

1) The dataset spans a period of one month, from **February 2, 2020**, to **March 13, 2020** with **4943** rows and **26** columns.
2) The entire dataset can be divided into three major segments.
    a) The first segment comprises **non-significant price difference**, which accounts for approximately **37.3%** of the data.
    b) The second segment represents **significant price differences**, exceeding 20%, making up around **31.6%** of the dataset.
    c) Finally, the third segment consists of missing values in the price, totaling **~31.0%** of the data.



3) The variables **distances/duration** and **predicted distance/duration** exhibit a similar distribution resembling a skewed normal distribution with a pointed peak. This suggests a correlation between the two variables in terms of their behavior. Additionally, these variables play a significant role in predicting upfront price differences.



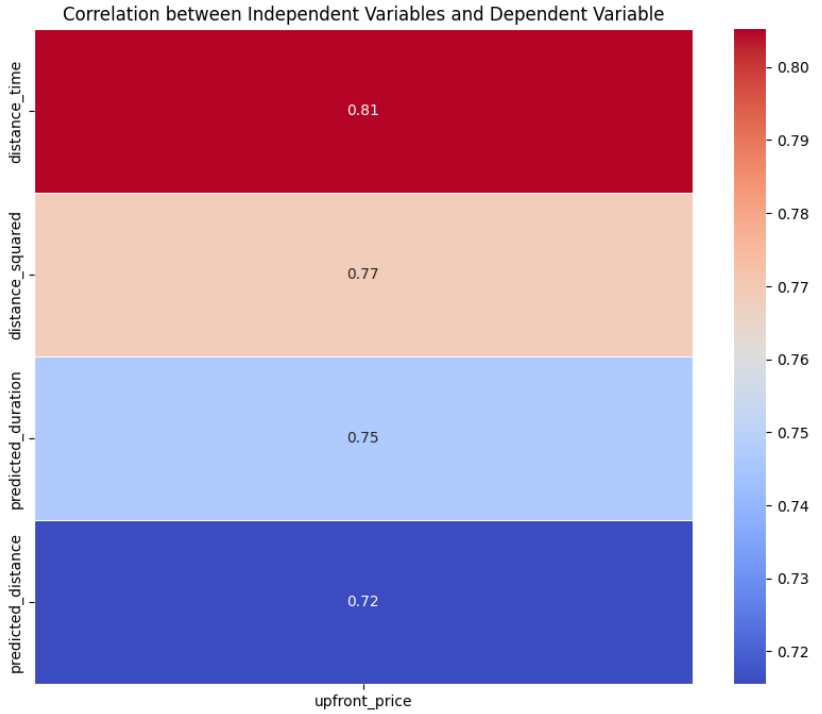


Opportunities to improve the upfront price

1) **Feature Engineering for Upfront_Price:**
   One way to enhance the prediction of upfront_price is by constructing or utilizing new features derived from existing ones. In our analysis, we focused on engineering three new features: distance_time, time_distance, and distance_squared, using the predicted_distance and predicted_duration variables.

To assess the impact of these newly created features on upfront_price prediction, we conducted regression and correlation analyses. Based on our analysis, we have reached the conclusion that the new features exhibit a correlation with upfront_price and can account for the variability in its prediction. A heatmap is employed to visually understand the correlations between the variables, while a table represents the p-values.

	Variable	R-squared	P-value
0	All Combined	0.843141	5.710689e-57
1	distance_time	0.648295	0.000000e+00
2	distance_squared	0.590489	0.000000e+00
3	predicted_duration	0.557765	0.000000e+00
4	predicted_distance	0.512000	0.000000e+00
5	eu_indicator_1	0.170775	9.482035e-141
6	gps_confidence_1	0.042844	2.637585e-34
7	dest_change_number_2	0.023298	3.210145e-19
8	time_distance	0.000669	1.310676e-01
9	dest_change_number_5	0.000052	6.725229e-01
10	dest_change_number_4	0.000048	6.849410e-01
11	dest_change_number_7	0.000035	7.304898e-01
12	dest_change_number_3	0.000033	7.364677e-01



We can also improve upfront_price accuracy by taking into account variables like traffic pattern, number of stops during the ride, weather and visibility details, road blockages etc.

2) Improving Predicted Duration and Distance:

An ANOVA test was conducted to investigate the most influential variables for predicted distance and duration. The results indicated that **"gps_confidence"** was a highly influential variable, with statistically significant p-values (< 0.05).

This variable is associated with the mobile devices used by drivers. Further analysis revealed that a majority of the devices with bad gps connection are from the Techno brand. As a recommendation, drivers could be encouraged to utilize mobile devices from other brands or consider upgrading to the latest models, as this may lead to improved gps_confidence and consequently enhance the accuracy of upfront price predictions.