Uber

FarePath: Navigating the Dynamics of Uber Pricing

Morgan Tucker, Ali Khan, Alicia Wilson

Introduction

This report delves into the factors influencing rideshare fare prices, specifically examining the role of geographical distance and its interaction with other variables. By employing a robust dataset we have conducted a series of multivariate regression models and detailed analyses to refine our understanding and build robust predictive models for fare variations.

Data Preparation

The table below shows the data directly downloaded from Kaggle. The data was not usable and required adding/subtracting columns.

Table 1: *Raw Data*

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1

The following changes were made to the Raw Data:

- Dropped 'key'/'Unnamed: 0'/'pickup_datetime'/
- New columns: Year'/'Month'/'Day of Week' (0 = Monday, 1 = Tuesday…)/'Hour' (military time) /'Quarter' (Binary)
- Renamed column headings uniform for clarity and ease of use
- Removed remaining missing values (0.0065% of the data)

Uber calculates its fare considering several factors. One of these factors is the distance from the origin to the destination. While we did not have this data readily available, we converted the latitudes and longitudes into distance. The Haversine formula calculates the distance between two points, using spherical trigonometry to calculate distances from the longitude and latitude coordinates.

Haversine Formula:

$$a = sin^2(\frac{\Delta lat}{2}) + cos(pickup\ latitude) \cdot cos(dropoff\ latitude) \cdot sin^2(\frac{\Delta long}{2})$$

$$c = 2 \cdot atan(\sqrt{a}, \sqrt{1 - a})$$

$$d = R \cdot c$$

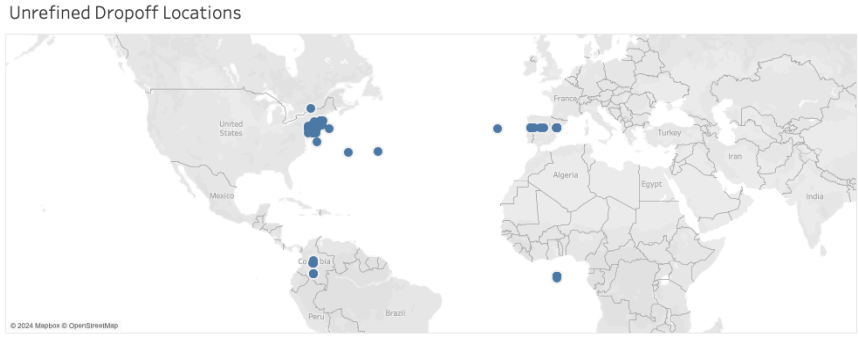
Where:

- Δlat is the dropoff latitude - pickup latitude
- $\Delta long$ is the dropoff longitude - pickup longitude
- R is the Earth's radius (taken as 6,371km)
- d is the distance between the two points (measured in km)

This formula was applied in Python (created as a define function and mapped over every row, then divided by 1600) to produce a 'distance_miles.'

Next, we wanted to understand the longitude and latitude coordinates. Using Tableau we plotted the Drop Off coordinates (Figure 1).

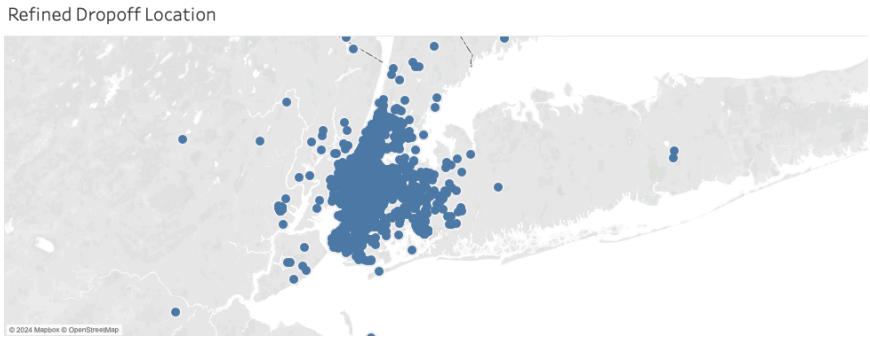
Figure 1: *Using Tableau to Plot the Dropoff Longitude and Latitudes*



The dropoff locations span across globally with concentrations in the NY and NJ area (USA) and Spain. We assume pickup locations are similar to drop off as the average Uber ride is 5.41 miles (Forbes 2015).

Different countries have varying economic climates, regulatory differences, cultural preferences, competition, and more. A predictive pricing model that combines different locations will ignore these considerations. Thus, we have filtered the data to include only the New York/New Jersey area. This was done by looking at the longitudes/latitudes, creating a perimeter, and dropping any points outside of this perimeter. Figure 2 shows the data points of the refined data we are using. Refining this area shrunk the dataset from 999,999 points to 23,310.

Figure 2: *Using Tableau to Plot the Dropoff Longitude and Latitudes for the NY/NJ Area*



After consolidating the data frame to the New York/New Jersey region, we then removed outliers. This process was done using RStudio and removed the bottom 2.5% and the top 2.5% of the data (from fare and distance variables), under the assumption that they represented outliers.

Figure 3: *Function to remove outliers for Fare and Distance variables in RStudio*

```
# function to remove outliers based on quantiles
remove_outliers <- function(data, varname) {
  # calculate lower and upper bounds
  bounds <- quantile(data[[varname]], probs = c(0.025, 0.975))
  # filter out outliers
  data %>%
    filter(get(varname) >= bounds[1] & get(varname) <= bounds[2])
}

# remove outliers from fare_amount and distance_miles
final_uber <- remove_outliers(final_uber, "fare_amount")
final_uber <- remove_outliers(final_uber, "distance_miles")
```

Outliers can distort the results, skewing the model and resulting in less accurate and reliable outcomes. Moreover, they disproportionately influence the model's parameters, moving them away from what would be representative of the general trend within the data. The goal of excluding these extreme data points is to achieve a more robust analysis where the resulting model parameters genuinely reflect the typical patterns present in the data. The refined and clean version of our dataframe now consists of 20,067 observations and 15 columns (Table 2).

Table 2: *Final, Cleaned and Refined Data*

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	number_of_passengers	year	month	day_of_week	hour	distance_miles	quarter_of_year_Q1	quarter_of_year_Q2	quarter_of_year_Q3	quarter_of_year_Q4
0	7.5	-73.999817	40.738354	-73.999512	40.723217	1	2015	5	3	19	1.052077	0	1	0	0
1	7.7	-73.994355	40.728225	-73.994710	40.750325	1	2009	7	4	20	1.535994	0	0	1	0
2	12.9	-74.005043	40.740770	-73.962565	40.772647	1	2009	8	0	21	3.147736	0	0	1	0
3	5.3	-73.976124	40.790844	-73.965316	40.803349	3	2009	6	4	8	1.038552	0	1	0	0

Exploratory Data Analysis

This report contains some plots and tables created. For a more comprehensive EDA review this link.

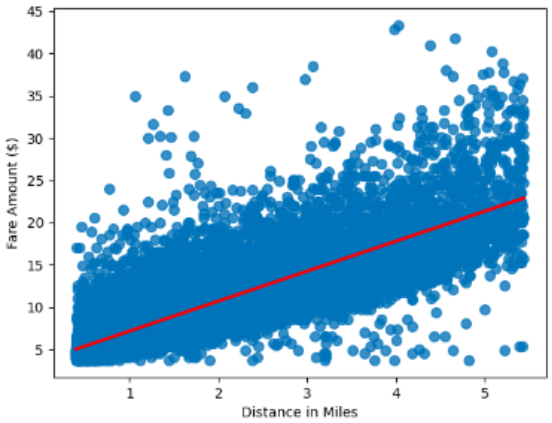
Table 3 shows the summary statistics for fare, number of passengers, and distance traveled. The majority of rides involve only one passenger, as shown by the median (50%) value, even though the mean number of passengers is slightly higher at 1.68, suggesting occasional higher passenger counts. The average trip covers a distance of approximately 1.69 miles, but there is a range, with the longest trips reaching up to 5.44 miles.

Table 4: *Summary Statistics for Fare and Number of Passengers*

	Fare Amount ($)	Number of Passengers	Distance (Miles)
Mean	9.60	1.68	1.69
Std	4.81	1.30	1.11
Min	3.70	0.00	0.40
50%	8.50	1.00	1.34
Max	43.33	6.00	5.44

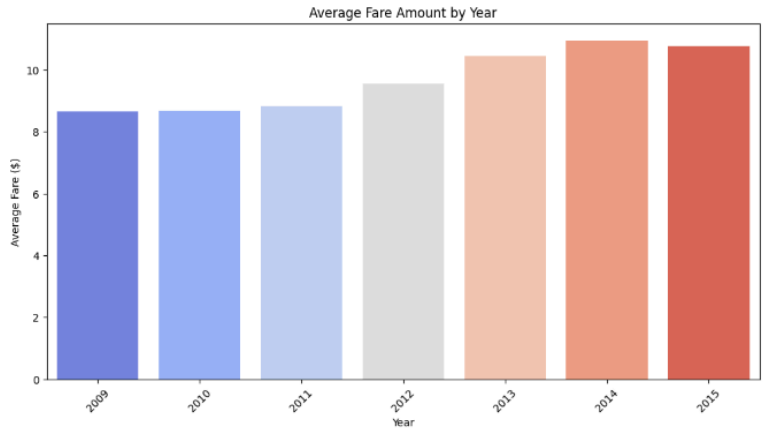
Aside from high correlations between longitude and latitude figures (which will be dropped in the regression model since we have distance), none of the dependent variables are highly correlated. The only variable with a high correlation to fare (independent variable) is distance (0.82). From this we can predict that distance is going to be a very strong predictor of fare, and will likely explain a significant portion of the variance. Figure 4 visualizes this relationship by plotting fare (y-axis) and distance (x-axis).

Figure 4: *Fare versus Distance*



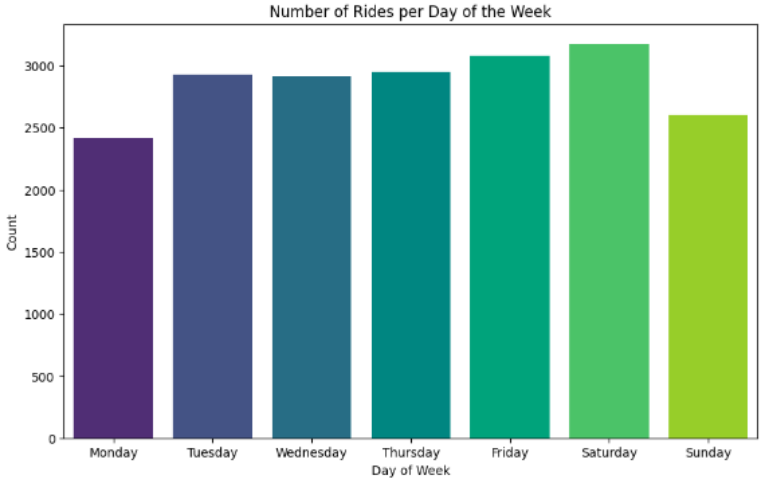
This next figure (Figure 5) demonstrates how fare prices have steadily increased over the years. This is going to be another important factor in our regression model.

Figure 5: *Fare versus Year*



Lastly, we look at the number of rides across the days of the week. Saturday has the highest demand, with the most rides on this day. Both Monday and Sunday have the lowest counts.

Figure 6: *Number of Rides taken throughout Week*



Again, for a more comprehensive review and plots, view the Colab file previously linked.

Multivariate Regression Methodology

We employed regression analysis to identify the key determinants influencing fares. Our method involved juxtaposing conventional multivariate regression techniques with models incorporating interaction effects and feature engineering to ascertain which provided the best insights.

We hypothesize that fluctuations in ride fares are influenced by a combination of direct and interactive factors. Direct factors such as distance and time are well-known determinants of fare pricing. However, we also propose that interactions between these factors—specifically the time of day combined with the day of the week—play a critical role by reflecting higher fares. Additionally, we suggest that the proximity to significant locations, like airports, impacts fare prices. We expect that incorporating 'distance to the airport' as a feature will show that rides closer to the airport are priced higher. By enhancing our feature engineering in this way, we aim to capture the complex dynamics of fare structuring more effectively and improve the predictive accuracy of our model.

Baseline Approach

The baseline approach included regressing fare price on all of the initial variables. The table below shows the coefficients from this approach:

Table 5: *Coefficients for Baseline Regression*

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)	p-values signification
Intercept	-915.603	20.028	-45.715	**<0.0001**	-954.861	-876.346	***
number_of_passengers	0.025	0.014	1.763	0.078	-0.003	0.053	.
year	0.457	0.010	45.866	**<0.0001**	0.437	0.476	***
month	0.075	0.022	3.377	**0.001**	0.032	0.119	***
day_of_week	-0.046	0.010	-4.786	**<0.0001**	-0.064	-0.027	***
hour	0.015	0.003	5.395	**<0.0001**	0.010	0.021	***
distance_miles	3.545	0.017	214.794	**<0.0001**	3.512	3.577	***
quarter_of_year_Q1	0.060	0.207	0.289	0.773	-0.346	0.466	*
quarter_of_year_Q2	-0.039	0.144	-0.269	0.788	-0.320	0.243	*
quarter_of_year_Q3	-0.093	0.086	-1.082	0.279	-0.260	0.075	*

*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 <. < 0.1 < ° < 1*

- Year: positive and statistically significant effect on fare price. Fare prices have steadily increased over years. All else equal, each year is associated with an increase of 46 cents in the fare price.
- Month: positive and statistically significant, suggesting that there is a slight increase in fare prices with each advancing month (holding other variables constant), possibly due to seasonality effects. This could suggest that as we enter the Winter season, rideshare fares are likely higher.
- Distance: large positive coefficient, which is statistically significant. This confirms that distance is a key factor for fare price; for every additional mile traveled, fare price increases by $3.55 (holding variables constant).

The analysis yielded an R-squared value of 0.708, indicating that 70.80% of the variation in fare prices is accounted for by the model. The adjusted R-squared is 0.708, confirming that the model's explanatory power is not enhanced by superfluous variables. While this provides a solid foundation, our objective is to enhance the R-squared value.

Interaction Approach

Subsequently, we explored enhancing the predictive model through interaction terms. We investigated whether the interplay between certain variables could yield a more accurate and nuanced representation of the determinants affecting fare prices. To do so, we constructed and analyzed several models, each incorporating a unique set of interaction terms, to evaluate their contribution to the model's performance:

- Interaction 1: $Distance\ Traveled\ x\ Day\ of\ Week$
- Interaction 2: $Distance\ Traveled\ x\ Day\ of\ Week\ x\ Hour$
- Interaction 3: $Distance\ Traveled\ x\ Day\ of\ Week\ x\ Hour\ x\ Month$
- Interaction 4: $Distance\ Traveled\ x\ Hour$

Preliminary visual examinations of scatterplots didn't suggest non-linear patterns in the data. However, we hypothesized that there might be underlying interactions between the day of the week and the time of day, influenced by seasonal factors. For instance, in NY's winter months, when the evenings are colder and darker there could be a tendency for people to book rides earlier. This prompted us to consider these potential interactions and their implications on fare prices more closely in our subsequent modeling

efforts. Table 6 shows the results from these interactions, when compared to the baseline model:

Table 6: *Results from Interactions*

Interaction	**R-Squared**	**Adjusted R^2**	**RMSE**	**↑↓ vs. Baseline**
Baseline	70.80%	70.80%	2.602	
Interaction 1	70.80%	70.78%	2.601	**Lower**
Interaction 2	70.80%	70.79%	2.601	**Lower**
Interaction 3	70.78%	70.77%	2.602	**Lower**
Interaction 4	70.82%	70.80%	2.601	**Higher**

The R-squared and adjusted R-squared values do not show substantial improvements vs. the baseline model. All the interaction models, except Interaction 5, resulted in a marginally lower adjusted R-squared, indicating that they may not be adding explanatory value beyond what the baseline model provides. Additionally, the RMSEs for Interactions 1 to 4 are slightly lower than the baseline but might not represent a practical improvement in predictive performance. Interaction 5 has a slightly higher R-squared, yet when adjusted it returns to the baseline level. This indicates that the interaction may be unnecessary.

Table 7: *Coefficients for Interaction 4 Regression*

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)	p-values signification
Intercept	-914.935	20.018	-45.706	**<0.0001**	-954.172	-875.698	***
number_of_passenger	0.025	0.014	1.773	0.076	-0.003	0.053	.
year	0.456	0.010	45.872	**<0.0001**	0.437	0.476	***
month	0.075	0.022	3.373	**0.001**	0.032	0.119	***
day_of_week	-0.045	0.010	-4.726	**<0.0001**	-0.064	-0.026	***
hour	-0.006	0.005	-1.105	0.269	-0.016	0.004	*
distance_miles	3.397	0.035	97.501	**<0.0001**	3.329	3.465	***
quarter_of_year_Q1	0.058	0.207	0.281	0.778	-0.347	0.464	*
quarter_of_year_Q2	-0.036	0.143	-0.251	0.801	-0.317	0.245	*
quarter_of_year_Q3	-0.093	0.086	-1.085	0.278	-0.261	0.075	*
distance x hour	0.011	0.002	4.806	**<0.0001**	0.007	0.016	***

*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1 < ° < 1*

- Distance x hour is statistically significant, but has a very small positive coefficient. As the interaction increases, so does the fare (holding all else constant).
- Hour becomes statistically insignificant after adding this interaction, indicating hour should be dropped. The same does not apply to the distance variable.

Airport Approach

Upon observing that interaction terms did not confidently enhance the model's predictive capability, and with the assumption that polynomial transformations would likely follow suit, we pivoted our strategy toward augmenting the dataset with an additional variable. Informed by personal experience that Uber fares are typically higher when ordering from airports, we tried to quantify this factor. We pinpointed the geographical coordinates of John F. Kennedy International Airport (JFK), a major transit hub in New York,

and used the Haversine Formula to compute the distances from this airport to the pickup and drop-off locations. The two new columns were:

1. Distance from JFK and Pickup Point
2. Distance from JFK and Drop Off Point

We added these variables to the baseline regression equation to produce the following results:

Table 8: *Results from Airport Approach*

Additional Variables	**R-Squared**	**Adjusted R^2**	**RMSE**	**↑↓ vs. Baseline**
Baseline	70.80%	70.80%	2.602	
Dist. to JFK (from Drop Off) Dist. to JFK (from Pickup)	71.2%	71.2%	2.583	**Higher**
Dist. to JFK (from Drop Off)	71.2%	71.2%	2.583	**Higher**
Dist. to JFK (from Pickup)	71.0%	70.9%	2.594	**Higher**

While this improvement in the R-squared is only slight, it is the best result. Including both the dropoff and pickup distances yield the same results as the dropoff→JFK alone. The pickup→JFK distance (though significant alone) becomes insignificant when combined with the dropoff→JFK distance. This is unsurprising given the correlation between the dropoff and pickup distances is fairly high (0.665), and it is likely that the dropoff distance to JFK is a more dominant factor, overshadowing the effect of the pickup distance. The Figure below shows the coefficients from the best Airport Analysis Model (Baseline + JFK to Drop Off Distance).

Table 9: *Coefficients from the Best Airport Approach*

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)	p-values signification
Intercept	-908.177	19.889	-45.662	**<0.0001**	-947.162	-869.192	***
number_of_passengers	0.026	0.014	1.838	0.066	-0.002	0.053	.
year	0.455	0.010	45.994	**<0.0001**	0.435	0.474	***
month	0.082	0.022	3.702	**0.000**	0.039	0.126	***
day_of_week	-0.049	0.009	-5.146	**<0.0001**	-0.067	-0.030	***
hour	0.018	0.003	6.340	**<0.0001**	0.012	0.023	***
distance_miles	3.501	0.017	211.221	**<0.0001**	3.469	3.534	***
quarter_of_year_Q1	0.124	0.206	0.603	0.546	-0.279	0.527	
quarter_of_year_Q2	0.009	0.143	0.061	0.952	-0.271	0.288	
quarter_of_year_Q3	-0.078	0.085	-0.912	0.362	-0.244	0.089	
distance_miles_from_JFK_dropoff	-0.265	0.015	-17.099	**<0.0001**	-0.295	-0.234	***

*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 <. < 0.1 < ° < 1*

- For every additional mile the dropoff location gets to JFK, the price increases by roughly 27 cents.
- The ride distance coefficient remains at $3.50.

These two coefficients highlight that distance is the pivotal factor in determining Uber fares. The cost of providing the service, including fuel or electric charge and time spent driving, directly correlates with the length of the trip. Furthermore, considering the distances between high-demand areas such as international airports and various dropoff or pickup locations is crucial. Areas closer to these high-demand points often experience increased fare prices. This price adjustment serves to push the price up to shift supply and demand back into equilibrium.

To cement the importance of distance we have produced a regression model including only distance_miles and distance from JFK to pickup. This model predicted has an R^2 of 67.7% which is incredibly close to all of our other models. Below are the corresponding coefficients.

Table 10: *Coefficients from just Distance measurements*

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)	p-values signification
Intercept	6.111	0.233	26.194	**<0.0001**	5.654	6.568	***
distance_miles	3.530	0.017	203.030	**<0.0001**	3.496	3.564	***
distance_miles_from_JFK_pickup	-0.190	0.017	-10.901	**<0.0001**	-0.224	-0.156	***

*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 <. < 0.1 < ° < 1*

- Both coefficients are statistically significant and similar in magnitude to Table 9.

Combined and Comparative Analysis

Drawing on the results from the baseline, interactive, and airport approaches we have combined an optimal multivariate regression model. This model drops all statistically insignificant variables (quarterly variables, hour (becomes insignificant when paralleled with hour x distance), and number of passengers). The variables year, month, day of the week, and distance from the baseline model are included, alongside the interaction of distance x hour and the distance between JFK and pickup. The table below shows a direct comparison between our best models from each analysis and this combined model.

Table 11: *Results from Best Models*

Models	R-Squared	Adjusted R-Squared	RMSE	# Variables	↑↓ vs. Baseline
Baseline	70.80%	70.80%	2.602	10	
Interaction 4	70.82%	70.80%	2.601	11	**Higher**
Dist. to JFK (from Drop Off)	71.2%	71.2%	2.583	11	**Higher**
Optimal Model	71.2%	71.2%	2.582	6	**Higher**

Table 11 shows that only marginal improvements are achieved, but the model's variance can be preserved with far fewer variables. The final table (Table 12) below shows the coefficients of this optimized model:

Table 12: *Coefficients from Optimized Model*

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)	p-values signification
Intercept	-909.366	19.839	-45.837	**<0.0001**	-948.252	-870.479	***
year	0.455	0.010	46.201	**<0.0001**	0.436	0.475	***
month	0.067	0.005	12.647	**<0.0001**	0.057	0.078	***
day_of_week	-0.047	0.009	-4.927	**<0.0001**	-0.065	-0.028	***
distance_miles	3.362	0.023	144.532	**<0.0001**	3.316	3.408	***
distance x hour	0.011	0.001	8.388	**<0.0001**	0.008	0.013	***
distance_miles_from_JFK_drop	-0.268	0.015	-17.292	**<0.0001**	-0.298	-0.237	***

*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 <. < 0.1 < ° < 1*

All coefficients are statistically significant. The largest coefficient is distance, followed by year (considered to be inflation) and distance from JFK. The equation below transforms these coefficients into the final equation predicting fare prices:

Fare Prices = -909.37 + 3.362(Trip Distance) + 0.455(Year) + 0.067(Month) - 0.047(Day) - 0.268(Distance$_{[JFK \rightarrow Pickup]}$) + 0.011(Distance x Hour)

Where:

- Trip Distance is measured in miles from pick up to drop off
- Month is entered as: 0 = January, 1 = February, …, 11 = December
- Day is entered as: 0 = Monday, 2 = Tuesday, …, 6 = Sunday
- Hour is entered in military time (e.g. 8:15pm is entered as 20)

Main Takeaways

- **Distance is the prime determinant:** The distance of the journey is the main predictor for fare. This is intuitive; the longer the journey, the more fuel used and time spent driving.
- **Importance in refining initial location:** Focusing on the New York/New Jersey area allowed for a more targeted analysis, acknowledging the specific economic and logistic dynamics of this densely populated region.
- **Distance can be used to analyze high traffic areas:** Incorporating the distance to JFK improved the model's accuracy, underscoring the higher fare associated with airport pickups.
- **Temporal Dynamics are not as important as initially anticipated:** Time-related factors (e.g. Hour), only subtly impact fares. As long as the supply of drivers is not too different from the demand for drivers, time becomes almost insignificant.

Project Improvements

- **Extend airport locations to include all airports:** The model only includes JFK. Include all airports in NJ and NY region (e.g. Newark Airport) to solidify the notion that fares are higher when closer to airports.
- **Extend airport approach to high-traffic areas:** Look at proximities to the busiest areas in areas

(e.g. Central Station and Fifth Avenue, NY) to incorporate higher fares in higher demand areas.

- **Continue to explore interactions:** More thoroughly investigate interaction and nonlinear variable relationships. Even explore endogeneity.
- **Extend across global locations:** Apply model in the same vein to Spain and other countries. Assess how models fluctuate based on location.

Conclusion

Our comprehensive analysis within "FarePath: Navigating the Dynamics of Uber Pricing" provides a key insight into the factors affecting rideshare pricing, with a particular focus on the geographical and distance elements that influence fare costs. Our project's evolution from a broad dataset to a focused analysis reiterates the criticality of context-specific research. This project may be a testament to the power of targeted data analysis and model refinement, but, in the face of complex, real-world challenges, "FarePath" has endless scope to improve its horizon and become even more comprehensive.

Project File Links

- Presentation Slides:
    - This file is the slide deck used in the class presentation
    - https://docs.google.com/presentation/d/1pOsJjZFl3mLvka76FVu36Lw8pn52Awh7LNEAKttezas/edit?usp=sharing
- GitHub Page:
    - Excel file with all regression models and results
    - R Studio file for data cleaning (taking IQR out)
    - Colab files with further links for EDA, refining geographical locations and longitude/latitudes
    - https://github.com/aliwilson2000/demand_analytics

Works Cited

Campbell, Harry. "Just How Far Is Your Uber Driver Willing to Take You?" Forbes, 24 Mar. 2015, www.forbes.com/sites/harrycampbell/2015/03/24/just-how-far-is-your-uber-driver-willing-to-take-you/. Date Accessed: 04-10-2024

"Largest Airports in North America." Airport Technology, www.airport-technology.com/features/largest-airports-north-america/. Date Accessed: 04-10-2024