Pace University

UBER FARE PREDICTION MODEL

NAME: SUMIT NALWADE
EMAIL: SUMIT.NALWADE@PACE.EDU
INSTRUCTOR NAME: STEPHANIE LANGELAND

PRACTICAL DATA SCIENCE(CS-667)
MASTERS IN DATA SCIENCE FALL- 2023

OCT 31, 2023

Agenda

1. Summary slide
2. Project plan recap
3. Data selection
4. EDA
5. Modeling
6. Findings
7. Recommendations
8. Appendix

Problem Statement

Uber delivers service to lakhs of customers daily. It’s really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

Solution

Build a Regression model to predict the price of the Uber
ride from a given pickup point to the agreed dropoff location.

Project Plan

Deliverable	Details	Due Date	Status
Data & EDA	Data visualization and EDA of Ube ride dataset.	**31-0ct-2023**	**COMPLETE**
Methods, findings, & Recmmendations	Model building, results, recommendations for improvements	**14-Nov-2023**	**COMPLETE**
Final Presentation		**5-Dec-2023**	**Not Started**

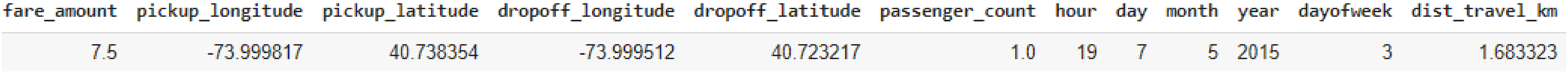
Dataset

Dataset: The dataset “uber-fares-dataset” referred from Kaggle (https://www.kaggle.com/datasets/yasserh/uber-fares-dataset)

Description: The dataset consists of information from 2009 to 2015 regarding the fare amount charged for a trip along with information such as location and pickup time etc. Each row represents a single Uber ride detail such as pickup time, location, fare etc.
Below are the fields that the dataset contains.

- key - a unique identifier for each trip
- fare_amount - the cost of each trip in USD
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged

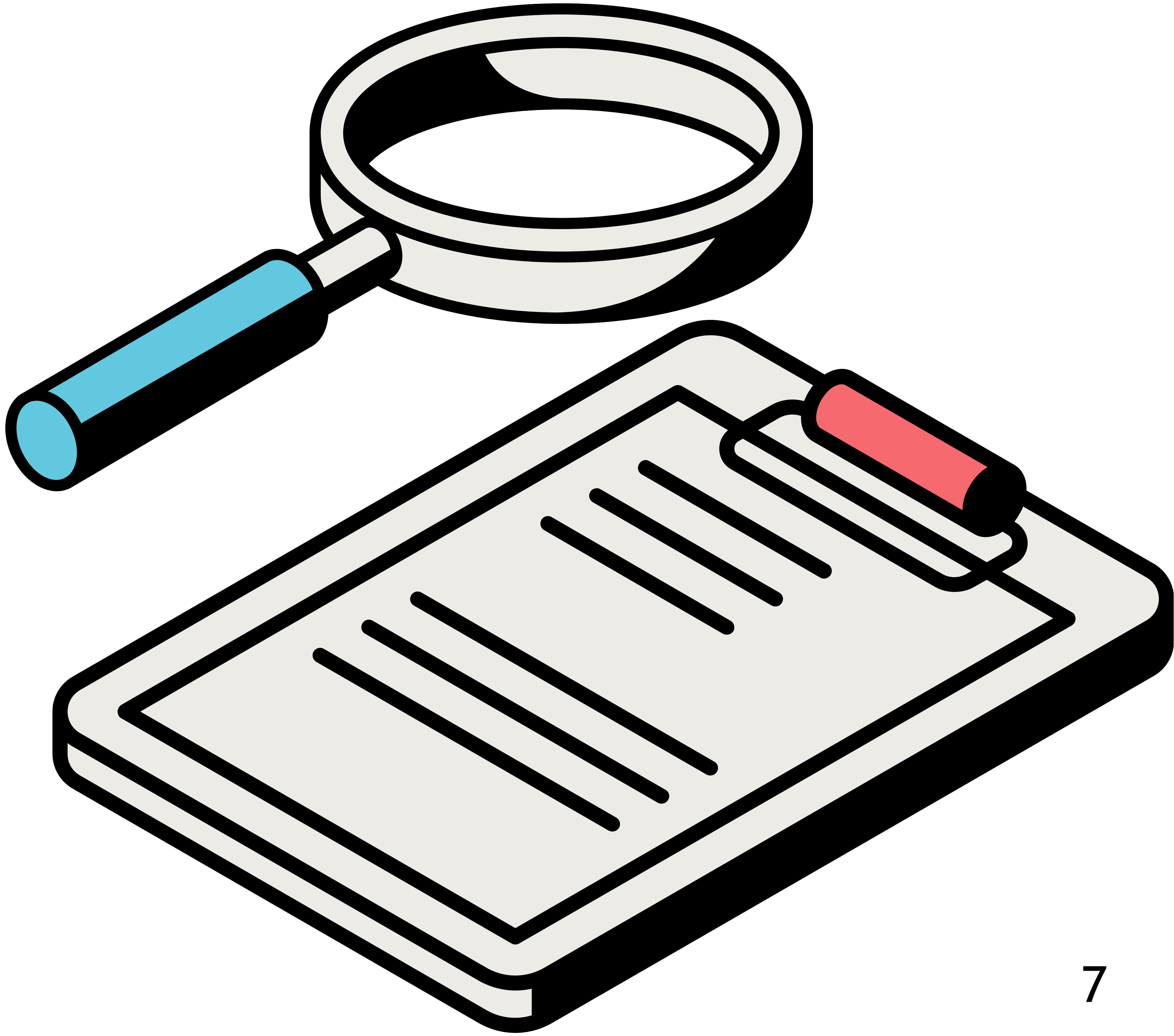
Data Size: 20000 rows, 8 Columns
Period: 2009 to 2015
Data Sample:



fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	hour	day	month	year	dayofweek	dist_travel_km
7.5	-73.999817	40.738354	-73.999512	40.723217	1.0	19	7	5	2015	3	1.683323

Assumptions: All the rides' minimum fare is 2.5$ because that is what standard charges for uber and each ride at least has 1 passenger

EDA & Data Visualization

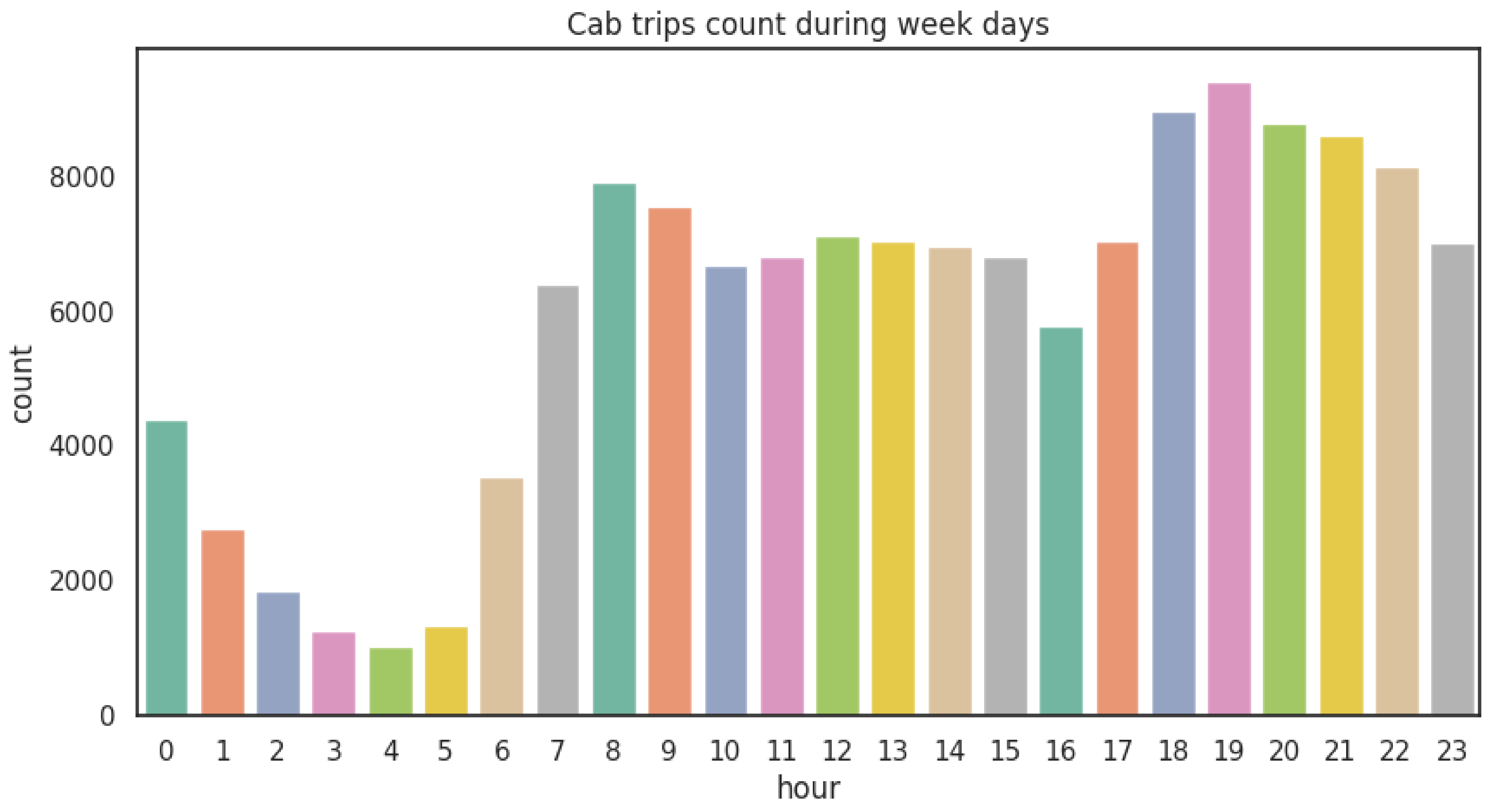


Strategies

- Data Cleaning
- Handing outlier
- Heatmap
- Data Visulization

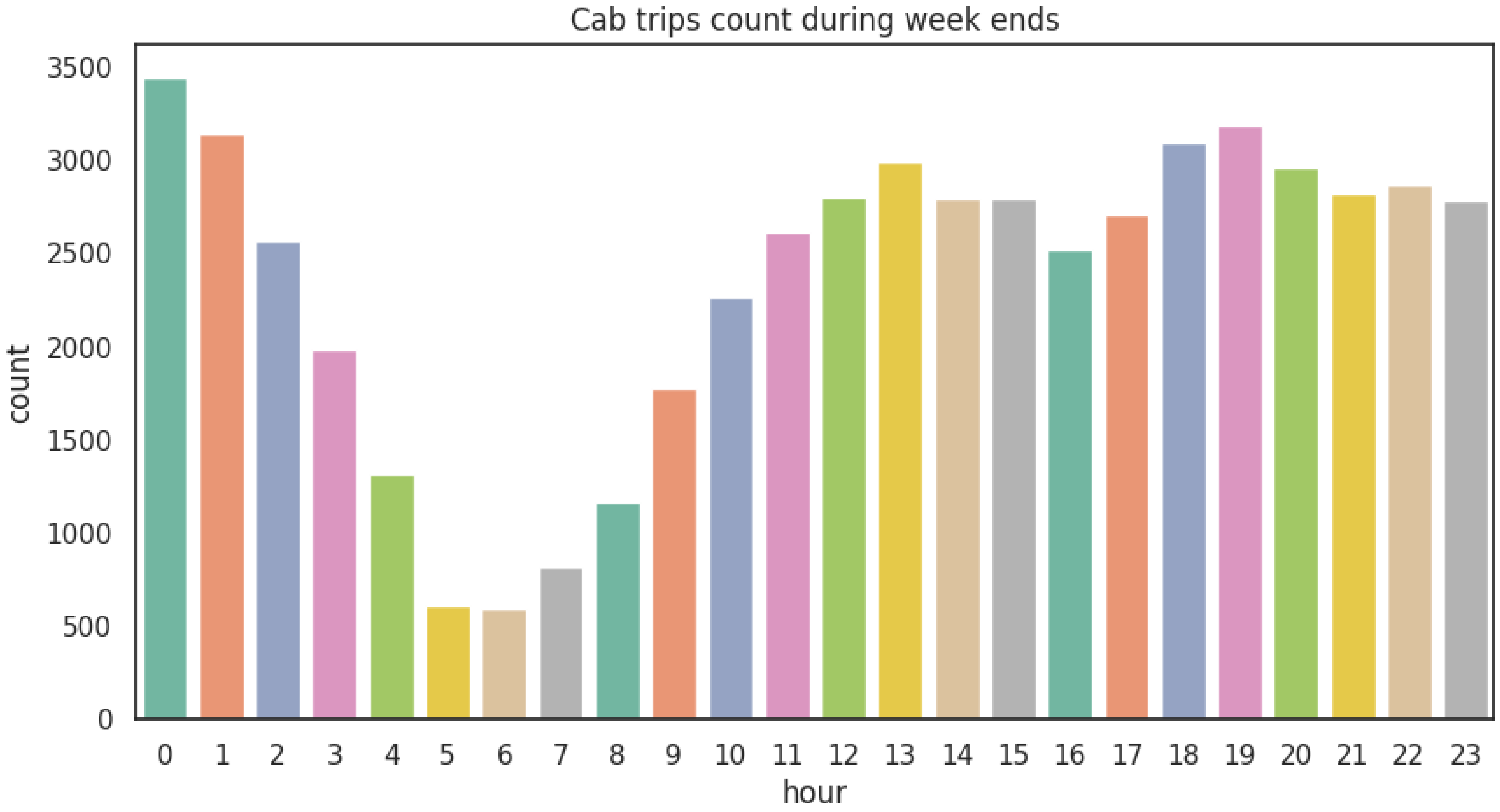
Number of cab trips on weekday

- From the chart we can see that cabs are in **high demand** during **evening** hours in **weekdays**.
- There is **decrease** in cab rides from **morning 8 am to 11 am**. probably because of office hours.
- This chart can help us to model our fare prediction model according to the number of trips for an hour.



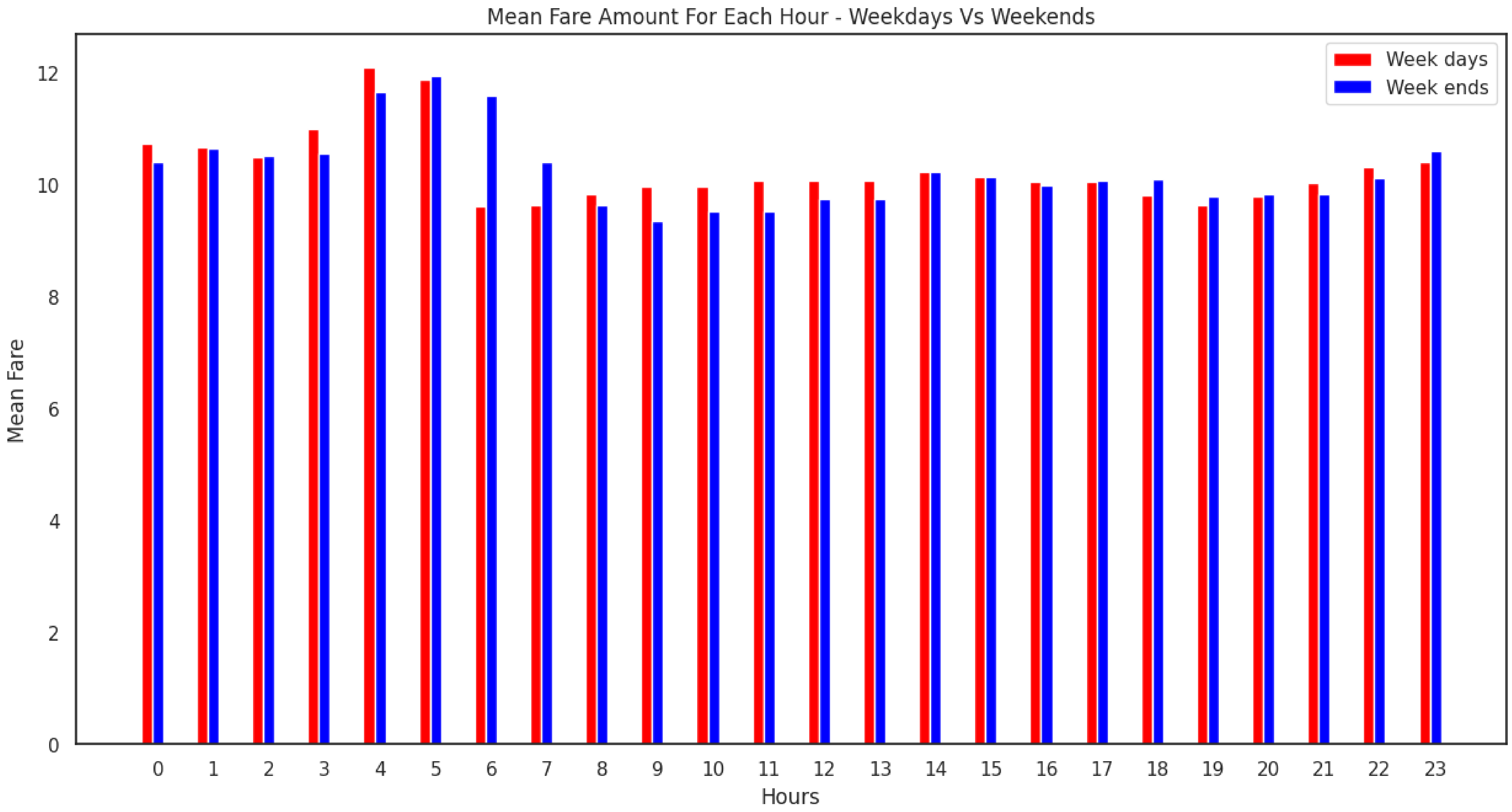
Number of cab trips on weekend

- Unlike **weekdays** during **weekends** cab trips are **high in demand** all the time except late in night(5am - 6am)



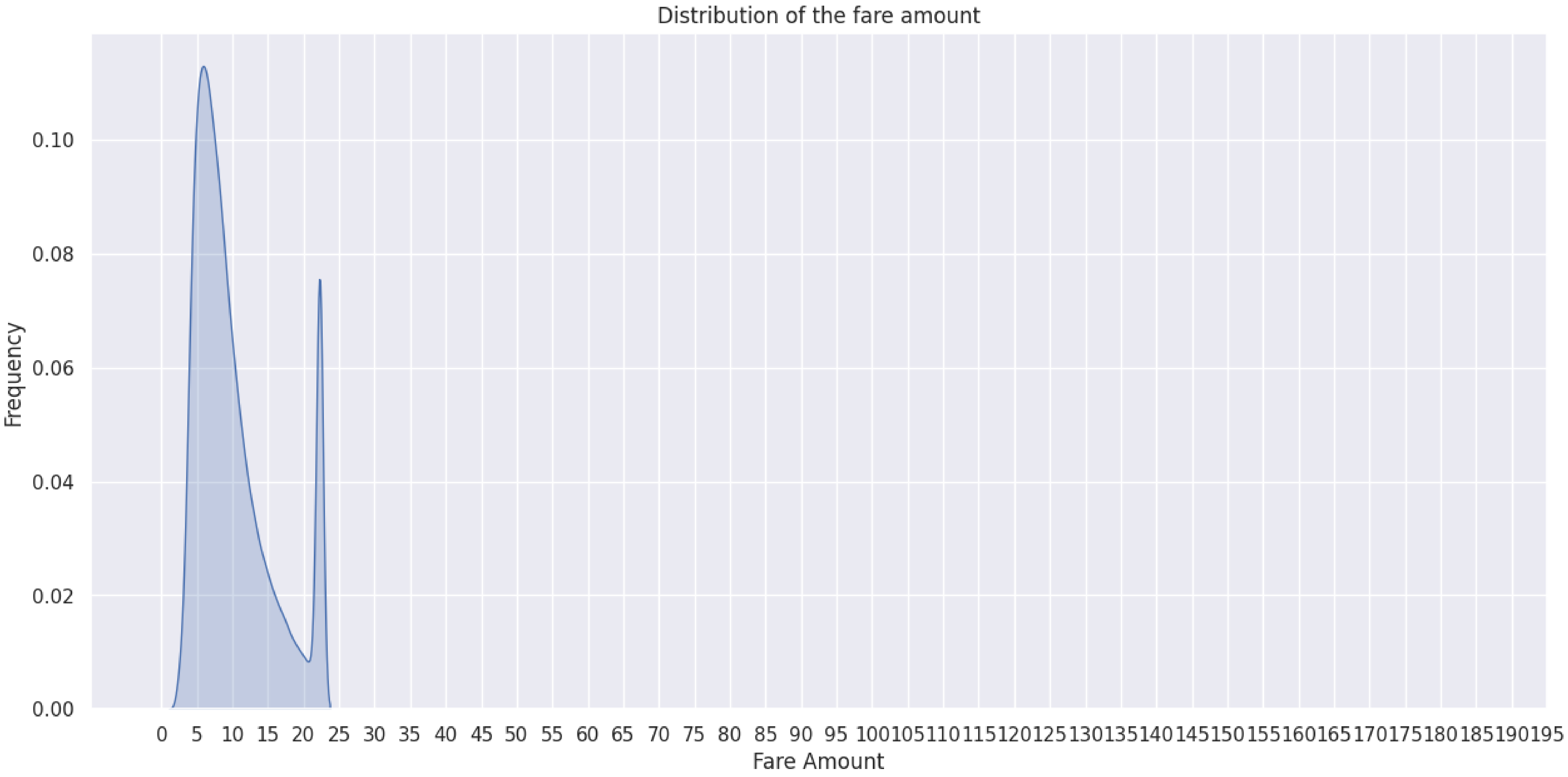
The average cost of cab trips on weekday Vs weekend

- The chart illustrates that the Average fare for a ride during each hour fluctuates for weekdays and weekends.
- We can see a pattern of high mean fares for morning rides during weekdays.
- With this information, we can tune our model to predict fare based on the hour and day of the week.

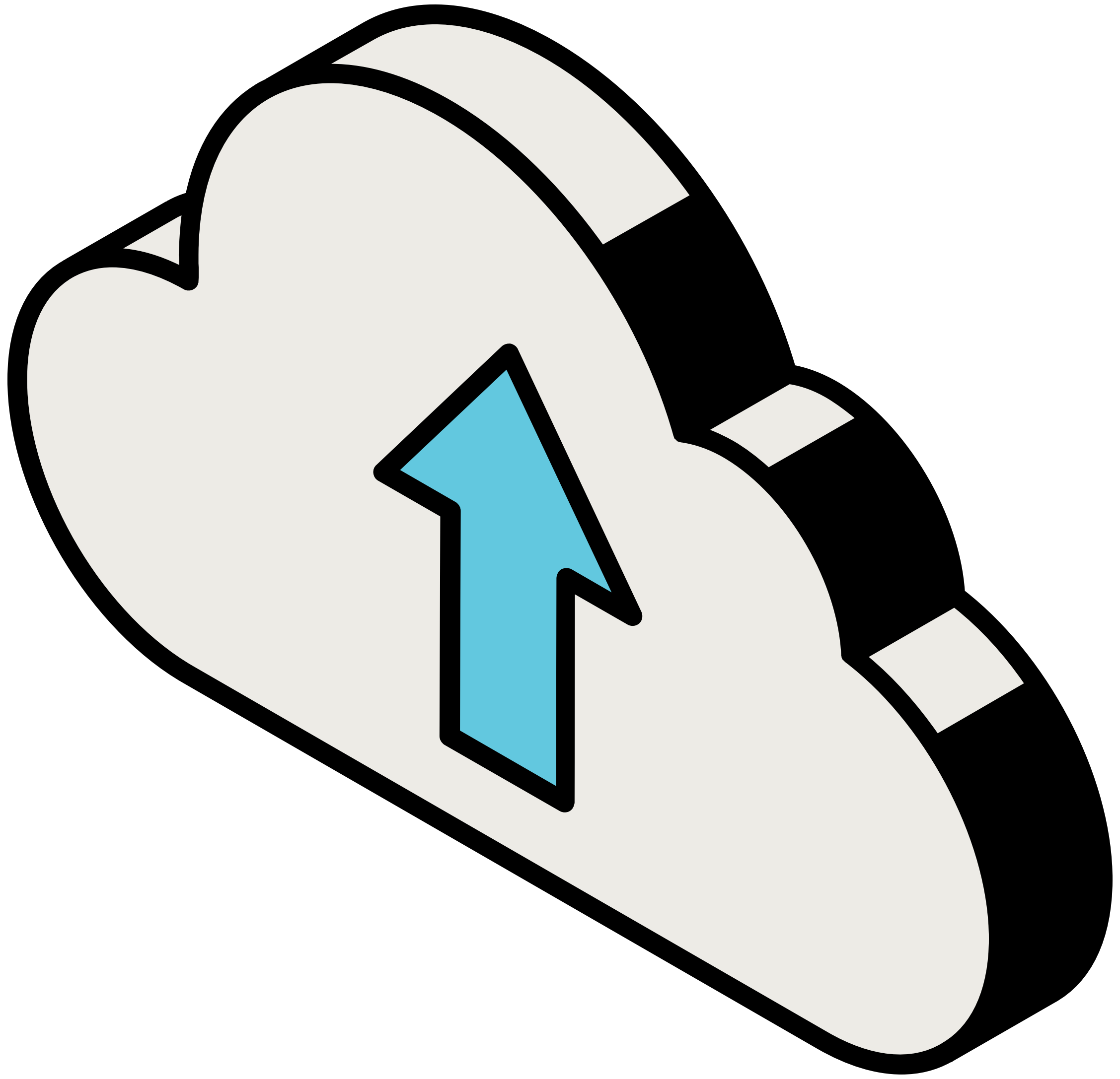


Range of Fare Amounts for Cab Rides

- We can see that most of the trips have a fare amount of 2 to 20 dollars.
- We can keep this range as a reference to train our model which can help achieve higher accuracy.



Modeling



Uber Fare Amount Prediction with Random Forest

Introduction:

- We've employed a cutting-edge approach known as a "Random Forest" to enhance our Uber fare prediction system.
- Imagine this model as a super-smart virtual Uber driver who calculates your fare based on factors: **‘pickup longitude', ’pickup latitude', 'dropoff longitude', 'dropoff latitude', 'hour', 'month', 'year', 'day of the week', 'distance travel (km)'**.

Features selection explanation:

If it's rush hour or you're travelling a long distance, these factors influence the fare. The Random Forest takes all these elements into account simultaneously, ensuring your fare estimate reflects the complexity of real-world scenarios.

Contunue...

Explanation:

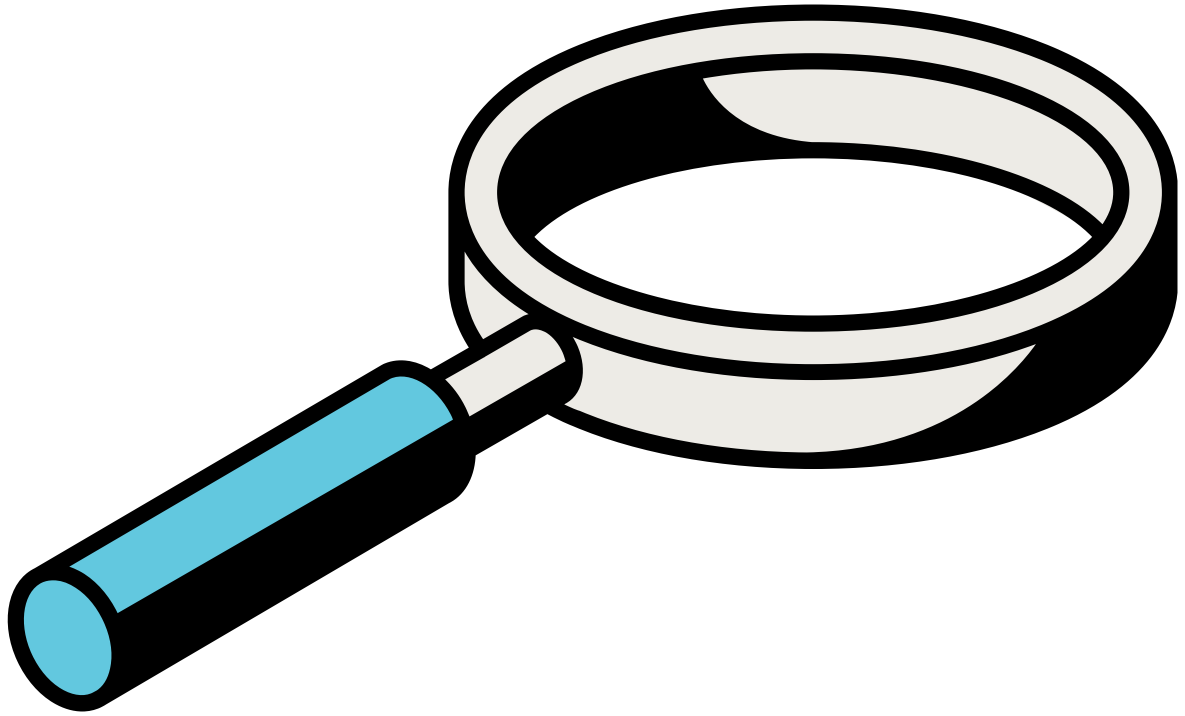
Imagine our prediction system as a team of experts, each with a unique skill in estimating Uber fares. These experts (or "trees") work together, combining their insights to provide the most accurate fare predictions possible.

Why Random Forest?

Think of it like making important decisions in your life. You wouldn't rely on just one person's opinion; you'd gather insights from various experts. Similarly, our Random Forest considers multiple factors like distance, time, and traffic, making it a robust and reliable predictor.

Learn More: For those interested in the technical details, we have a detailed breakdown in our appendix with all the nitty-gritty information.

Findings



Model Accuracy in the prediction of ride fare amount

Methodology to calculate accuracy:

- To check the performance of our prediction model we have looked for two statistical factors i.e. Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Mean Absolute Error (MAE):

- MAE is on the same scale as the original data.
- Smaller values of MAE are better. The MAE represents the average absolute difference between predicted and actual values.
- In our case, an MAE of 1.479 indicates, on average, That means our model's predictions are off by approximately 1.48 units from the actual values.

Mean Squared Error (MSE):

- MSE is in the squared units of the original data.
- Smaller values of MSE are better. The MSE penalizes larger errors more heavily than smaller errors due to the squaring operation.
- In your case, an MSE of 5.513 means that, on average, the squared differences between your predicted and actual values are 5.513.

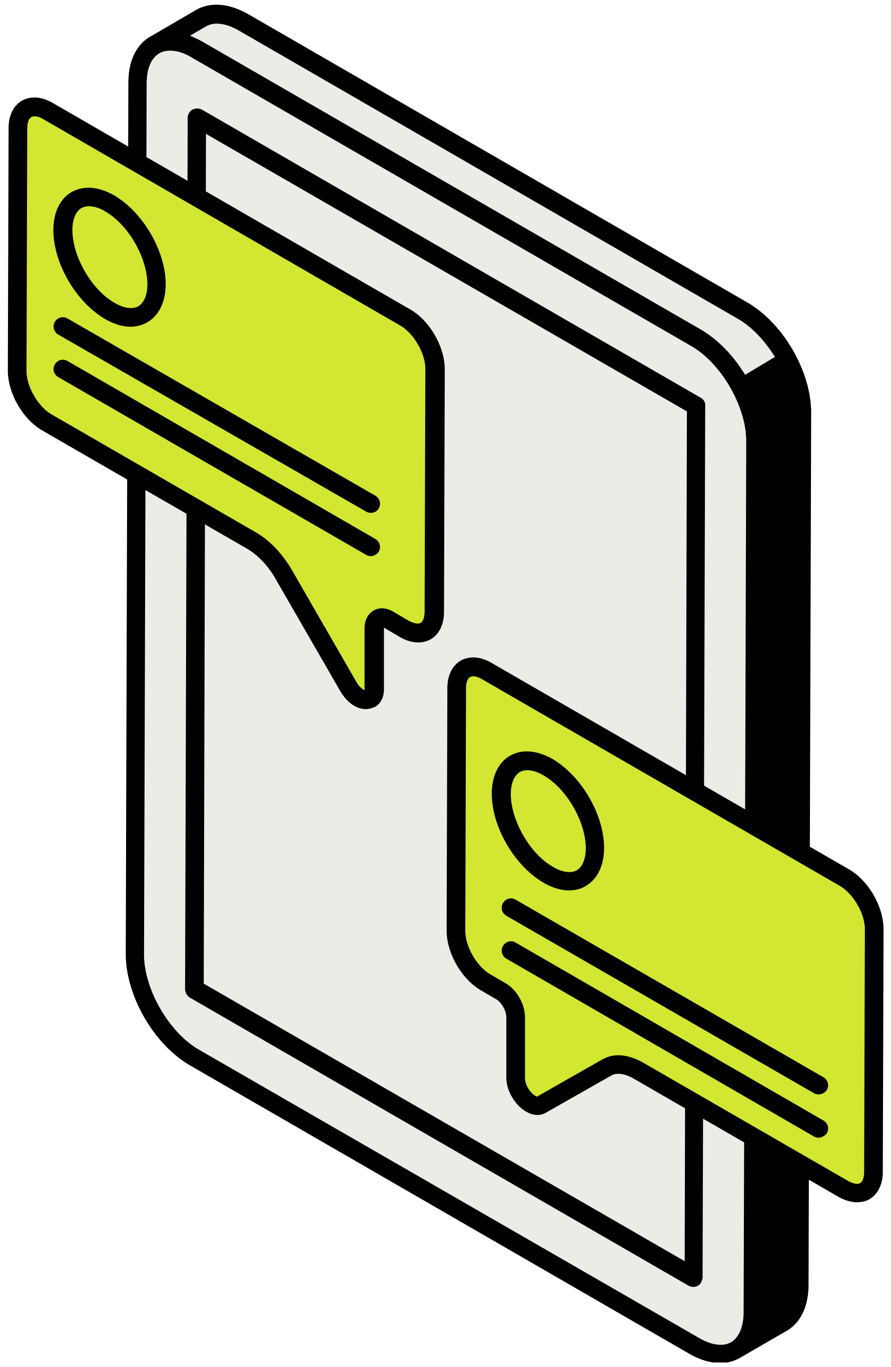
Findings:

- During our model training, we analyzed and found out that the Distance travel feature contribute the most in fare amount prediction.
- Pickup location and dropoff location also contribute significantly to fare prediction.
- We found out that late-night cab rides have mostly higher fare rates than day rides.
- The day of the week feature does not contribute much to fare prediction but anyway we keep it in out model training for insights.

Comparison of actual fare amount and model prediction value

Actual Fare amount	Model Prediction
2.5$	3.98$
5$	6.48$
10$	11.48$

Recommendations



How to increase model accuracy

1. Get More Data:

Collect a larger variety of information about Uber rides to help the computer learn better.

2. Think About Important Details:

Pay attention to details that might affect the price, like what Kind of weather it is when you're travelling and does it affects the fare amount.

3. Make Sure Data is Clean:

Check that the information you have is accurate and complete. Clean up any mistakes or missing info.

4. Adjust Model Settings:

Tweak the computer's settings to make it better at predicting prices. It's like adjusting the settings on your phone for better performance.

5. Keep Improving:

Regularly check how well the predictions are working. If you get more information or notice something new, update the computer program to keep making better predictions.

Next Step

1. Focus on improving model accuracy by selecting more relevant features to Target (fare amount) during training.
2. Adjust the parameters of the model and try different values to boost the performance of the existing model.
3. Train different types of models like Logistic regression, XGboost or maybe a deep learning model and select the one with higher accuracy at minimum compute cost.
4. Collect more real-time data for Uber rides fair amount.
5. Add more features like weather conditions, public holidays and big events around the location.

Appendix