SWIFTPICK

CS-GY 6513 Big Data Final Project
Professor Juan Rodriguez

Name:
Ping Chang (ypc231),
Kevin Lee (kl3642),
Guan Cherng Lin (gl2547),
Wei-heng Lin (whl318)

Table of contents

What are we proposing?

The motivation behind this project stems from the increasing demand for data-driven decision-making in the ride-sharing industry. As the volume of data generated by these platforms continues to grow, there is a compelling need for efficient tools that can handle large datasets and extract meaningful insights.

This project focuses on creating a web application that empowers users to input specific time windows and geographical locations to obtain insights into average trip costs and strategic recommendations for drivers to maximize profitability.

01

Intro & Methodology

What is SwiftPick?

Proposal

In the dynamic landscape of urban transportation, the advent of big data has opened unprecedented avenues for enhancing the efficiency and customer experience of taxi services.

Web App

Interactive big data web application

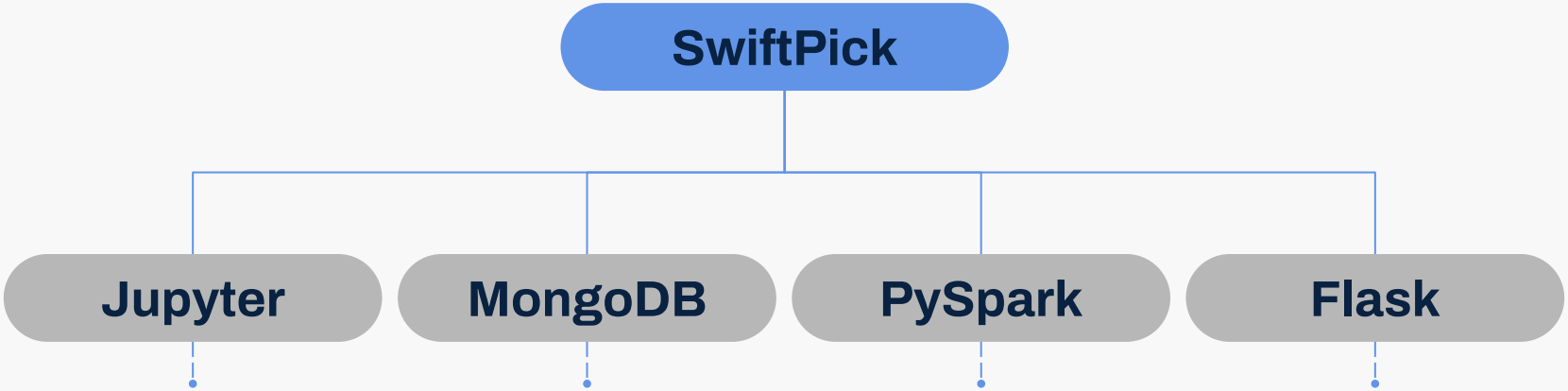
Predictions

Accurately predicting the cost of taxi rides

Insights

Intelligently guiding drivers to optimal pickup locations

Technologies Used



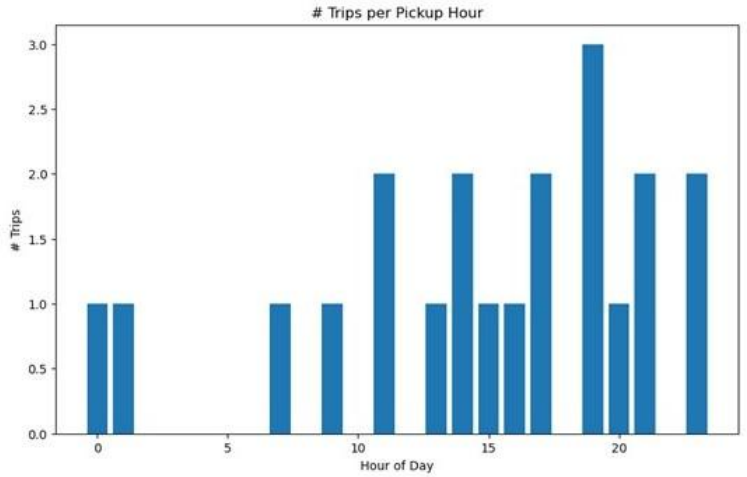
02

EDA
& Results

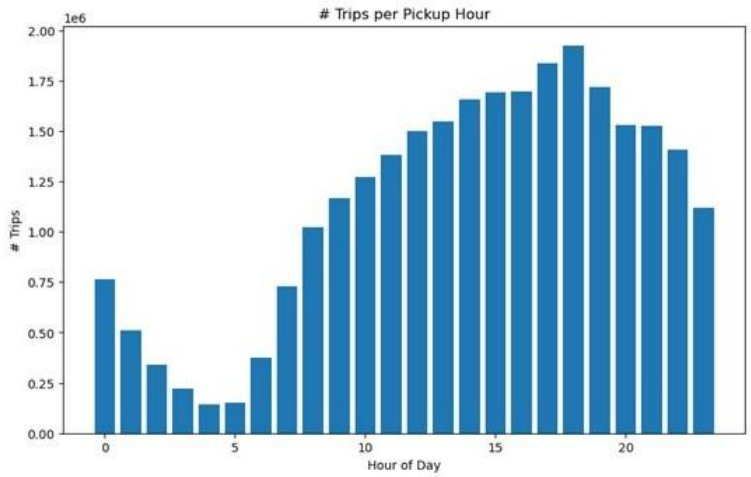
Data Exploration

- Utilizing data from 2022 onwards

2020

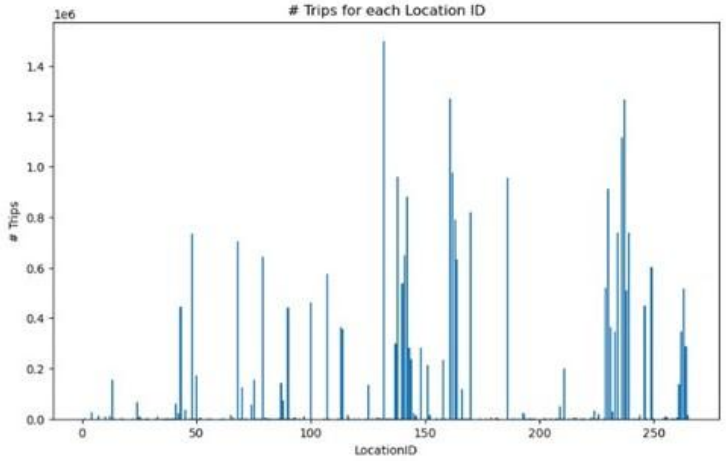


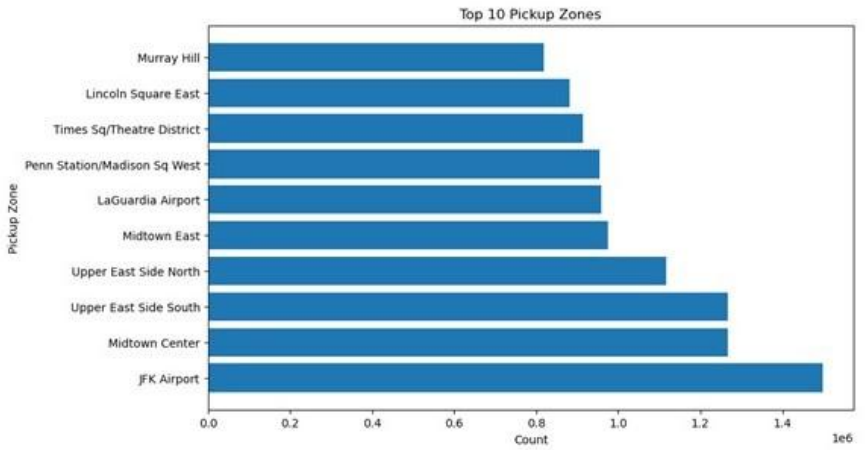
2023



Data Exploration

- Focussing on yellow taxi data



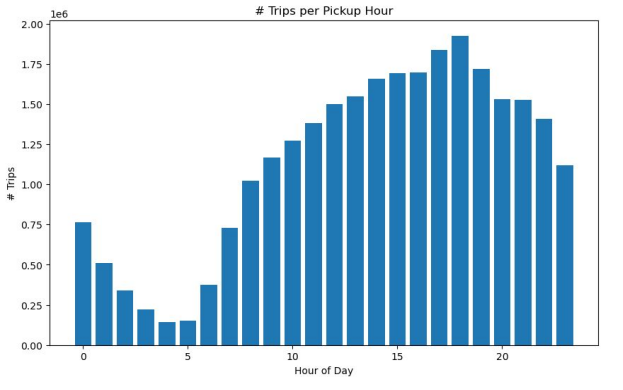


Machine Learning

K-Means Clustering

- To identify most favorable areas for driver positioning
- Clustering geographical coordinates
- Incorporated time-specific data analysis

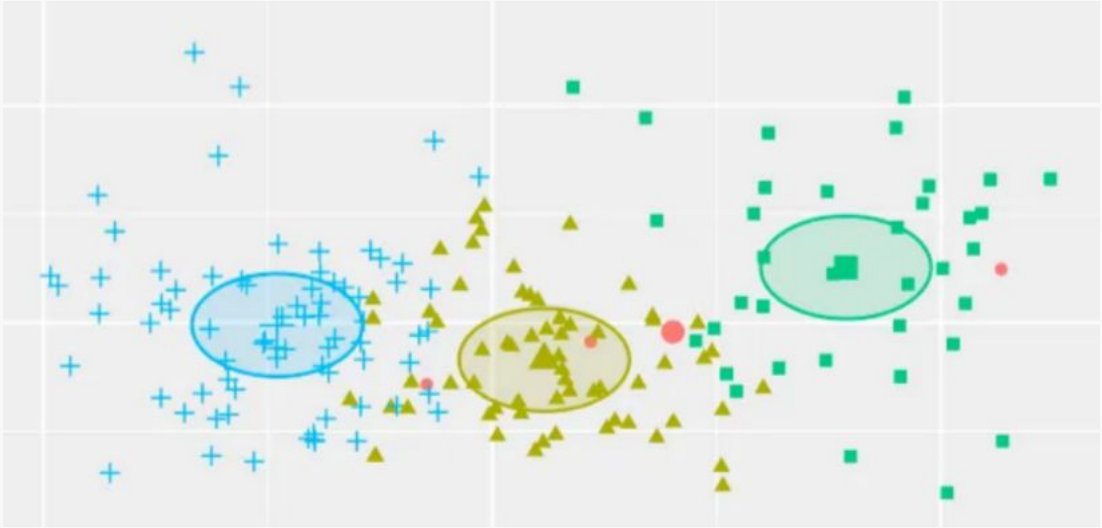
	tpep_pickup_datetime	tpep_dropoff_datetime	time_diff	Trip_distance	Passenger_count	PULocationID	DOLocationID	Zone_PU
0	2022-12-31 14:39:43	2022-12-31 14:43:37	-21	0.54	1	163	162	Midtown North
1	2022-12-31 15:14:12	2022-12-31 15:25:34	14	1.54	1	193	7	Queensbridge/Ravenswood
2	2023-01-01 14:39:59	2023-01-01 14:52:58	-21	2.20	1	246	249	West Chelsea/Hudson Yards
3	2023-01-01 14:31:15	2023-01-01 15:06:57	-29	11.69	1	140	177	Lenox Hill East
4	2023-01-01 14:43:04	2023-01-01 14:50:40	-17	1.61	2	151	239	Manhattan Valley



	Cluster	Centroid Longitude	Centroid Latitude
0	1	-73.784329	40.640388
1	3	-73.871681	40.773241
2	4	-73.958292	40.785653
3	2	-73.996985	40.733642
4	0	-73.975715	40.762317

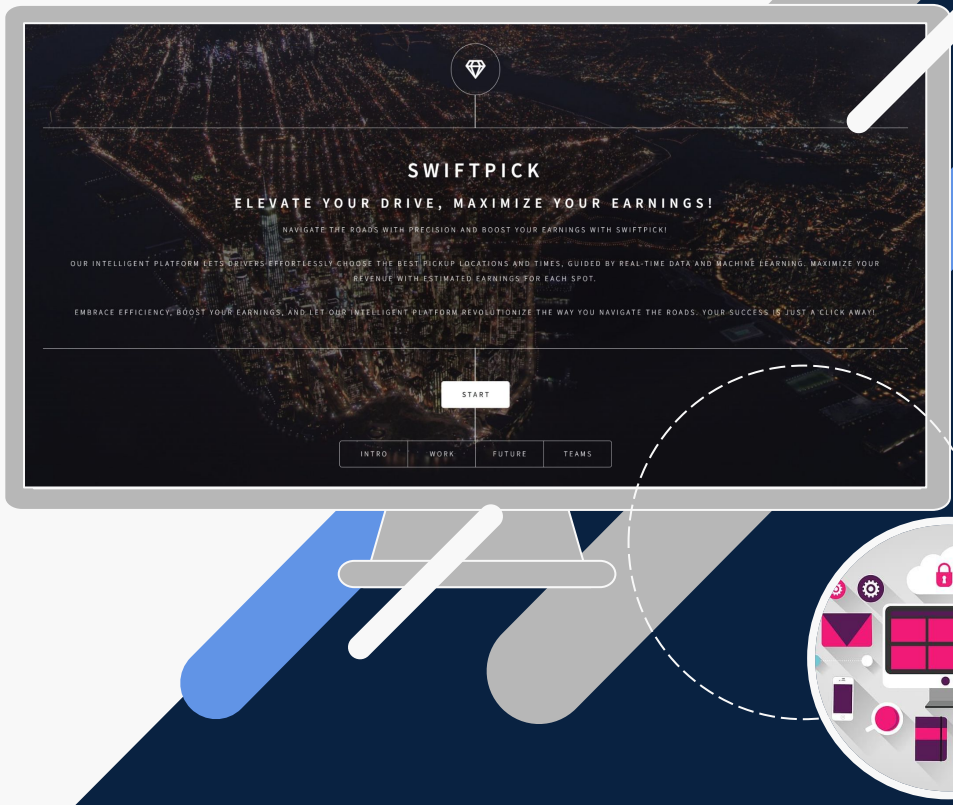
K-Means Processing

- Identifies high-demand area clusters
- Generate location specific recommendations
- Position to reduce idle time (maximize profitability)

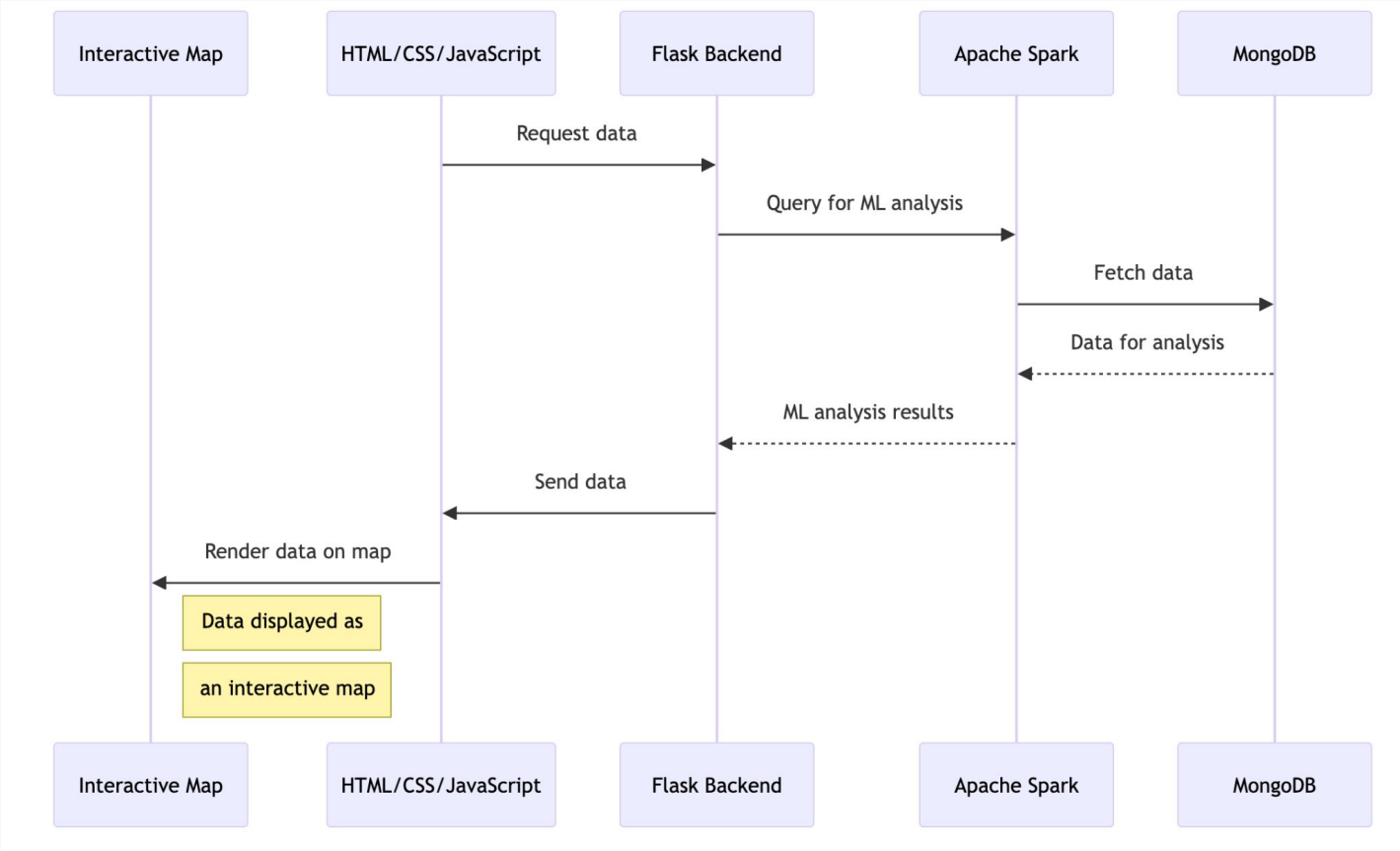


03

Architecture & Demo



Web App Architecture



04

Future & Conclusion

“The future depends on what we do in the present.”

Future Implementations

Real-time Data Collection

Collaborate with real-time data to present real-time estimations

Platform Integration

Apache Kafka
MongoDB

Expand Scopes in ML

More details added, more effective data processing

User-centric Upgrades

More comprehensive and beneficial for both drivers and passengers

Conclusion

What is SWIFTPICK?

The application pioneers the use of predictive technology in urban transportation.

Technology

- Pyspark/ Matplotlib
- Machine Learning
- Web development

Our Solutions

- Analyzing data of taxi ride in the past
- Integrating location data in real time
- Predict fares and guiding drivers to optimal pick-up locations.

The Future

What We Have Learned?

Thanks!

Q & A