ALI KAYA, 2024-01-22

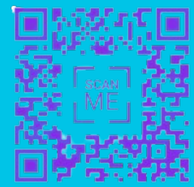
DATA ALAYSIS ON ORDER INFO.

Summer Intern Assignment Presentation

ali.kaya@abo.fi

PORTFOLIO

Wolt



RAW DATA OVERVIEW

- Consisting of 18,706 entries and 13 columns
- Including Order details spanning from 2020-08-01 to 2020-09-30, like time stamp, item count per order, user and venue locations, weather conditions, etc.
- Slightly containing missing value for weather conditions on 2020-09-10

1

PREPROCESSING

- Imputing Missing Values for the Date September 10, 2020
- Crafting Additional Features through Feature Engineering including DAY_OF_MONTH, HOUR_OF_DAY, DATE, DAY_OF_WEEK, DISTANCE(METERS)
- Detecting and Correcting Outliers if exiting

2

3

SOME STATISTICS

- User and venue locations have mean latitudes around 60.18 and longitudes around 24.94 (e.g. Helsinki area)
- Average estimated delivery time is 33.81 minutes, with actual delivery averaging at 32.61 minutes.
- Distances range from 0 to 4.67 meters, with a mean of 1.02 meters.
- Weather conditions like temperature, wind speed, and cloud coverage vary.

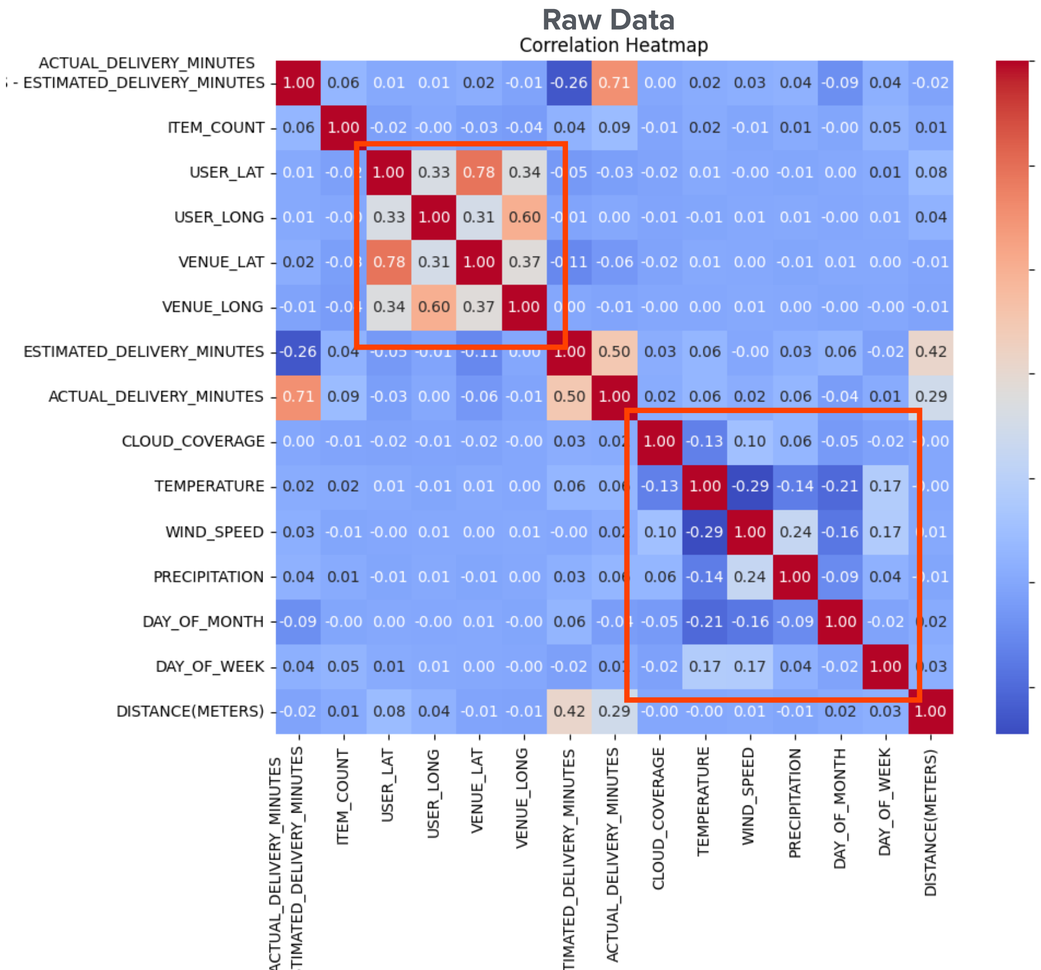
4

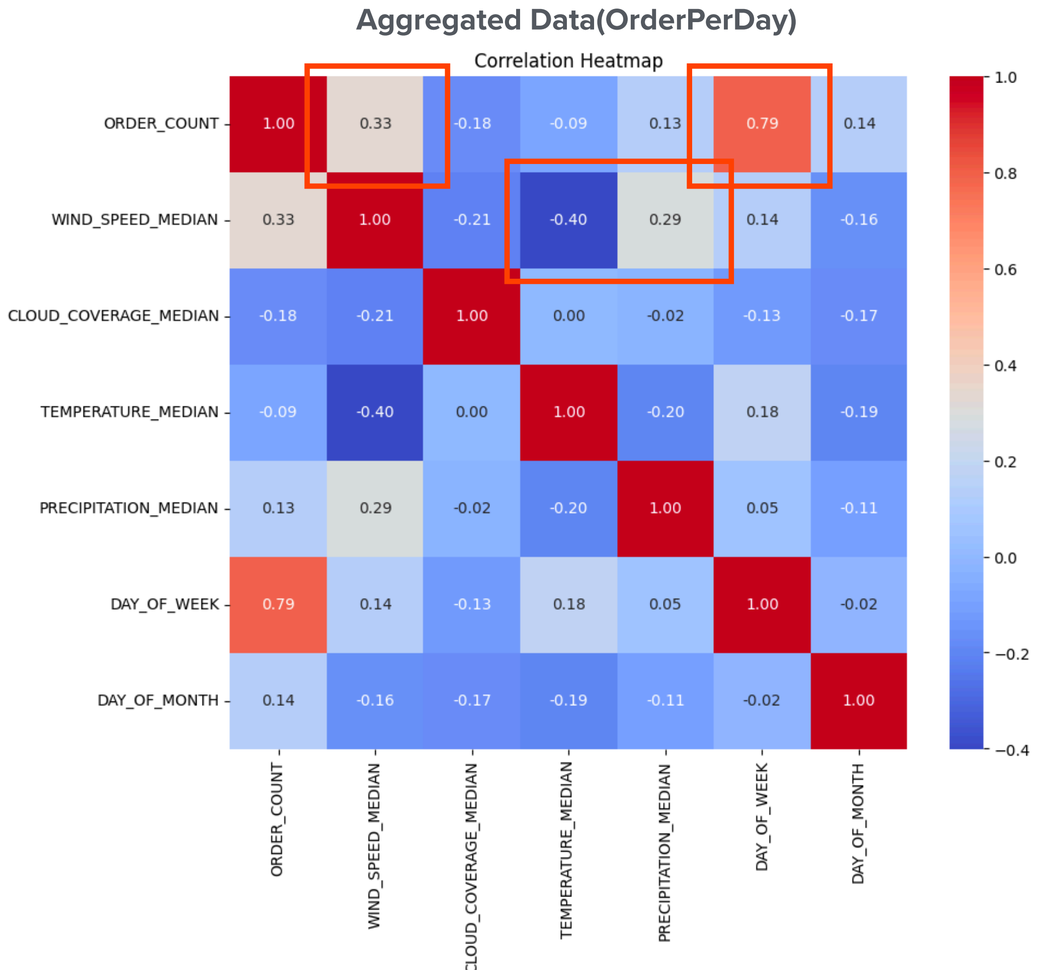
ANALYTIC DIRECTION

- Baseline: Exploring Correlations
- Assessing the Current Estimation Model for Delivery Time
- Geospatial Examination through Spatial locations for User and Venue
- Temporal Investigation aligned with User and Venue Clusters
- Forecasting Orders: Employing ARIMA and Prophet Models with Evaluation

CORRELATION INSIGHTS

- Customers generally prefer ordering from nearby venues, as evidenced by a strong correlation between user and venue latitude, as well as user and venue longitude.
- Weather patterns exhibit a consistent trend in alignment with established facts. E.g., Towards the end of the year, temperatures tend to decrease, and there is a proportional increase in wind strength as temperatures drop.
- The number of orders per day shows a strong correlation with the day of the week, indicating a need for further investigation during order forecasting.
- There is a mild correlation between the number of orders per day and wind speed, which is an intriguing observation.

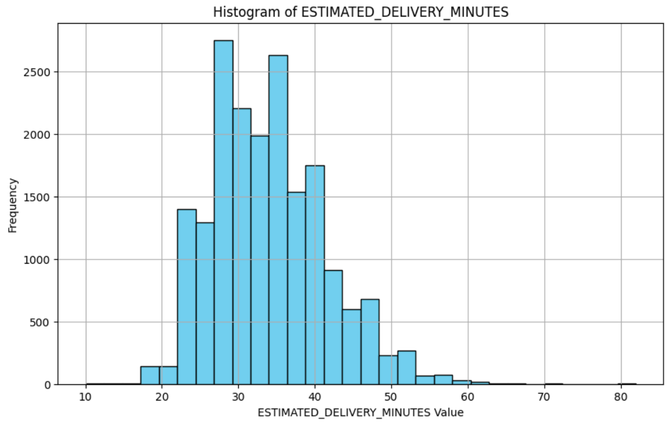


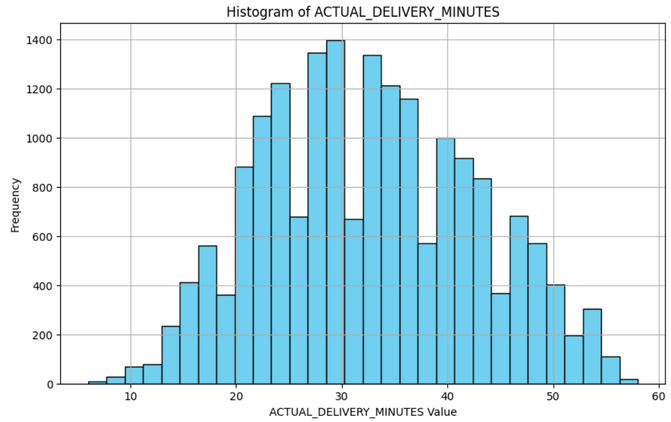


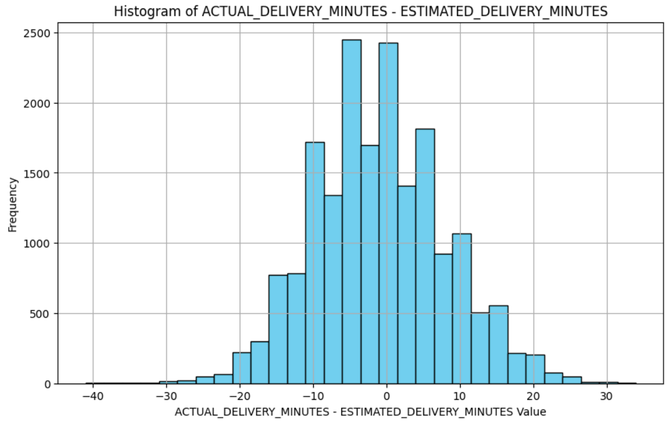
ASSESSMENT OF CURRENT ESTIMATED DELIVERY TIME

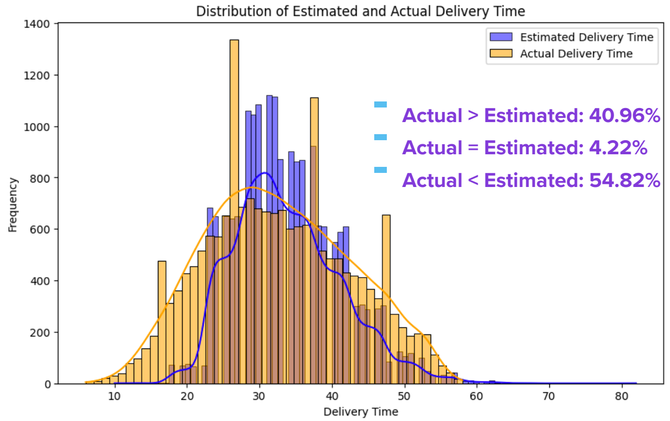
- The estimated delivery time exhibits a **slight leftward skew,** suggesting that the current estimation algorithm tends to be somewhat **aggressive**
- In contrast, the actual delivery time aligns closely with a **normal distribution.**
- As a result, the disparity between the estimated and actual delivery times leans slightly towards the negative side.

- As estimated delivery time increases, the lower bound of actual delivery time rises. Meanwhile, the upper bound of actual delivery time stays consistently high and **doesn't change with estimated time increments.**
- I suppose there is an '**under a one-hour delivery restriction**' in regards to commission, riders tend to **adopt a more leisurely approach,** particularly for nearby deliveries, aligning with individual **strategies** influenced by platform rules.





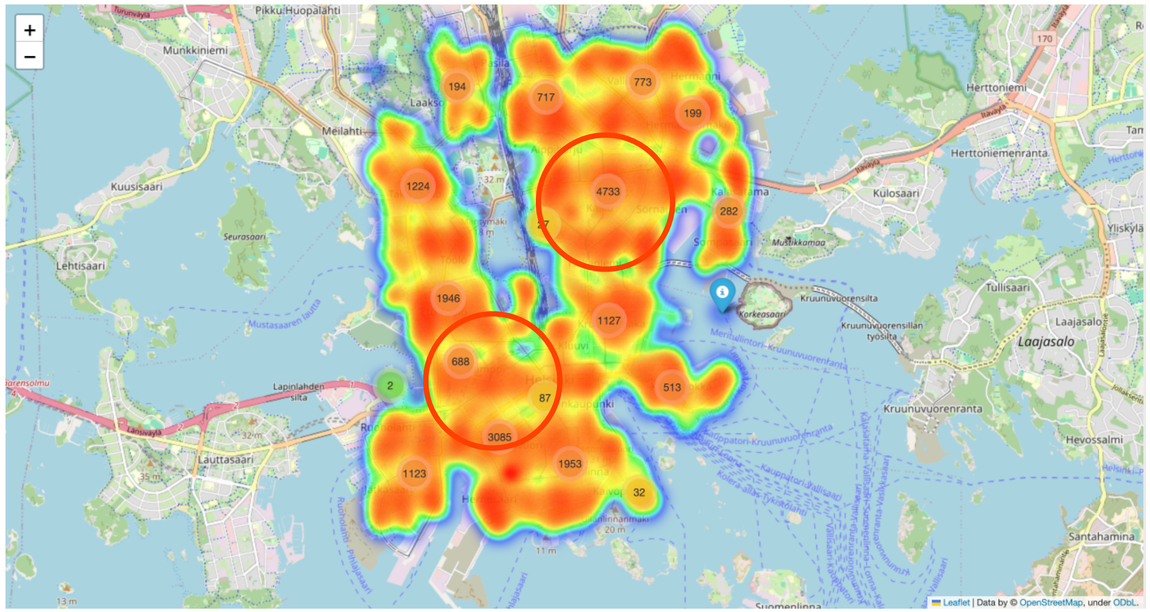




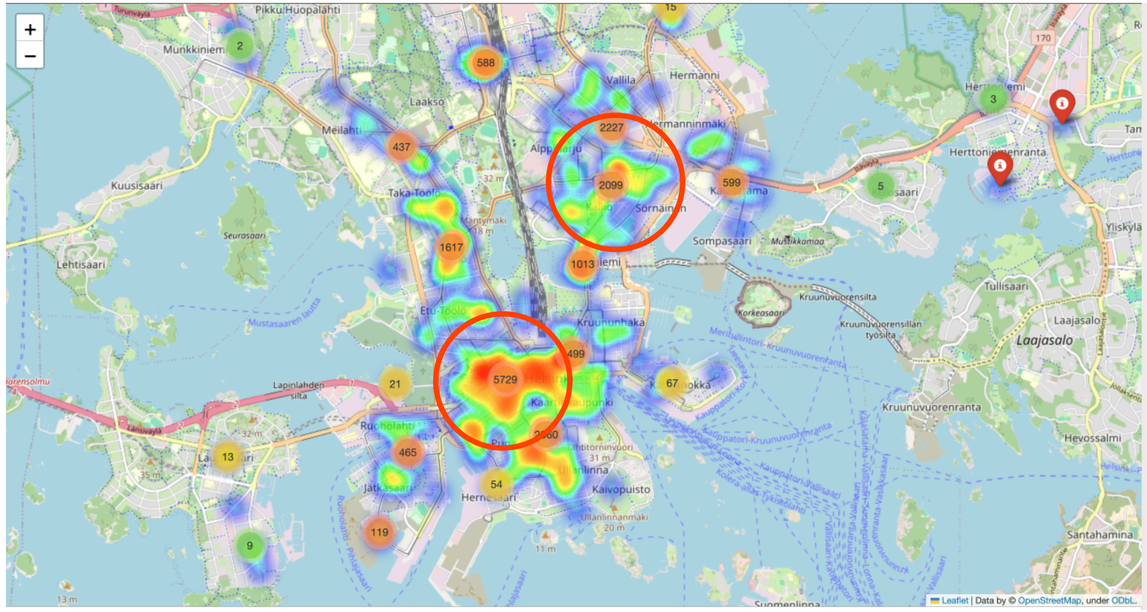
GEOSPATIAL EXAMINATION THROUGH SPATIAL LOCATION

- In terms of geography, users are spread across a broader and more expansive area, suggesting widespread acceptance of Wolts.
- Conversely, venues are densely clustered. This may be due to the actual distribution of restaurants (objectively) or could signify the need for Wolts to intensify efforts in areas with fewer registered restaurants for increased merchants recruitment(subjectively).
- Kamppi and Kallio emerge as the two busiest areas – the former renowned for its shopping centers and the latter for its abundance of restaurants and cafes.

USER AGGREGATION

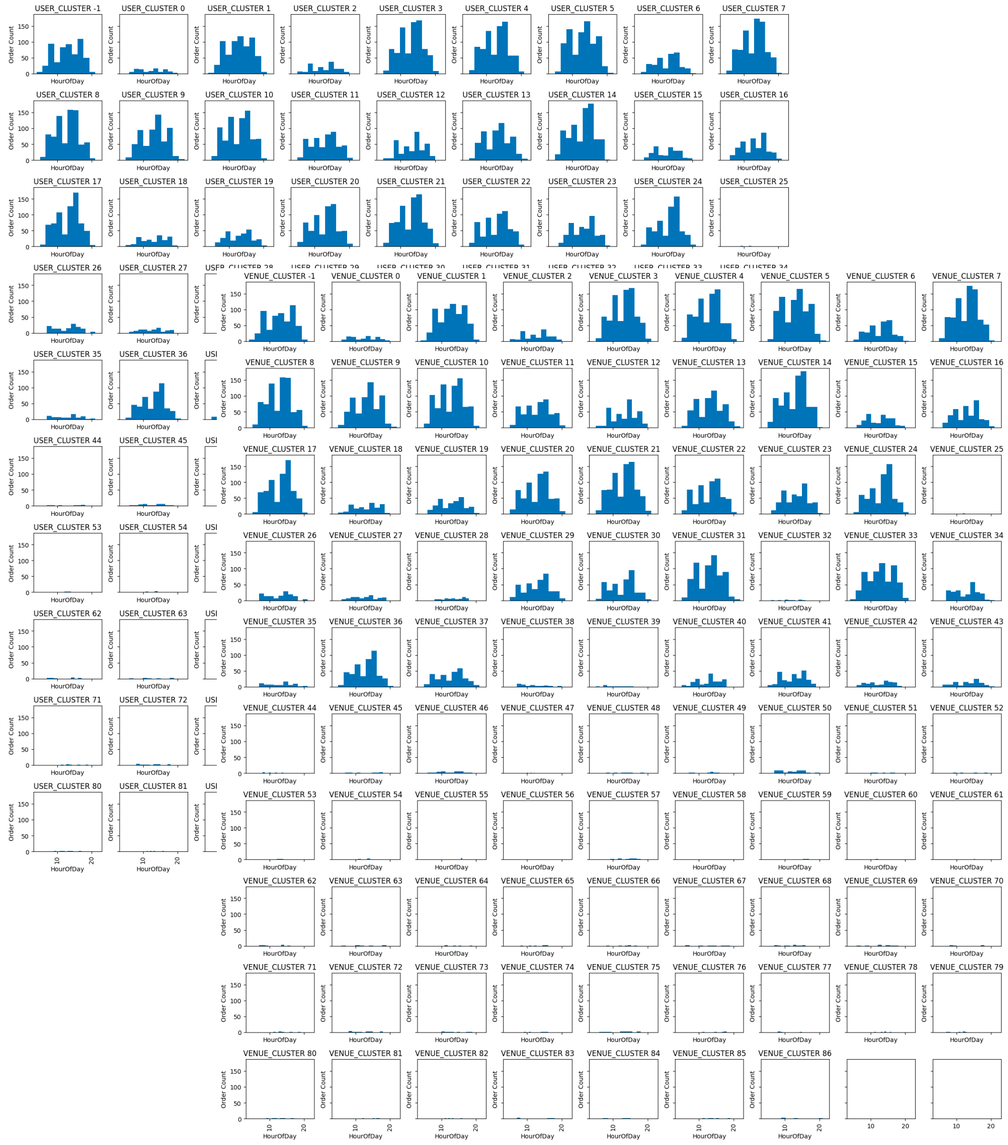


VENUE AGGREGATION



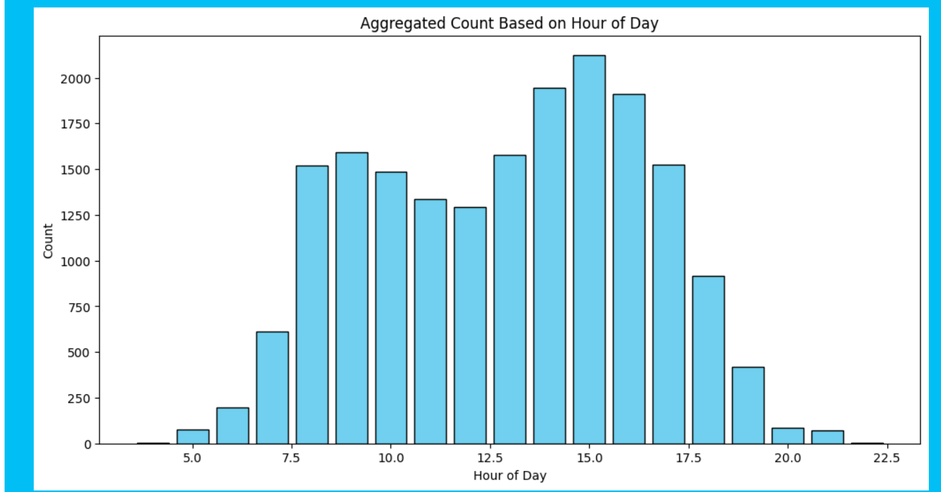
TEMPORAL INVESTIGATION ALIGNED WITH USER AND VENUE CLUSTER

USERS CLUSTERS AGAINST HOUR



Venues Clusters agains Hour

- The clusters are formed using **latitude**, **longitude**, **order distance**, and **disparities** between actual and estimated delivery times for both users and venues, individually, to capture more complex patterns if exists.
- Aligned with the spatial distribution of orders, certain clusters **consistently exhibit high order volumes**, while others **experience infrequent orders.**



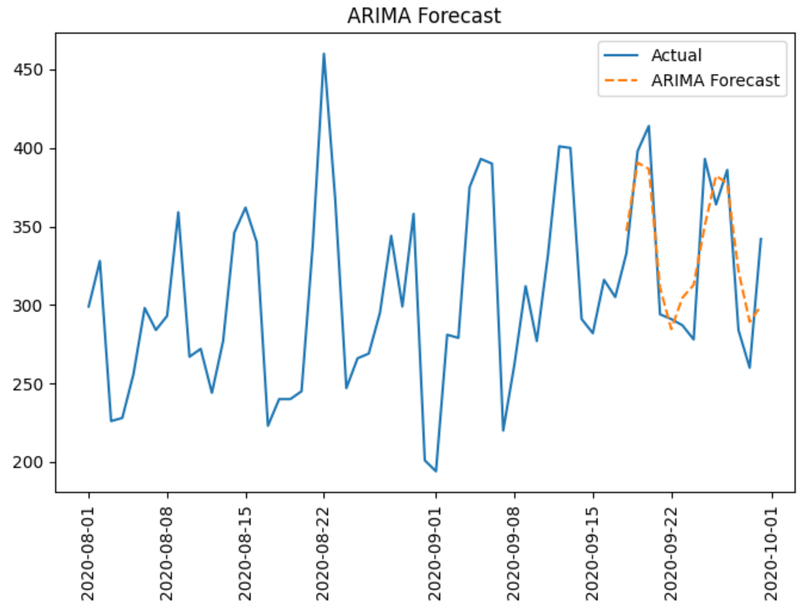
This represents the aggregated data without any clustering

- **Upon thorough examination of the clustered outcomes, we can infer that every cluster aligns with the global pattern.**
- **With a greater time span in the data, intriguing patterns may emerge, such as certain groups of users or venues displaying a higher frequency of orders in the morning instead of afternoon, which will be useful for ad promotion.**

FORECASTING ORDERS

- Exploring two popular forecasting models: ARIMA and Prophet
- The order data exhibits a prominent weekly pattern, as evident from the previous correlation analysis. This pattern should be taken into account both during training and testing phases.
- Weather-related variables are omitted from these models due to their low correlation with the order counts.
- The training and testing split is set at 80/20.

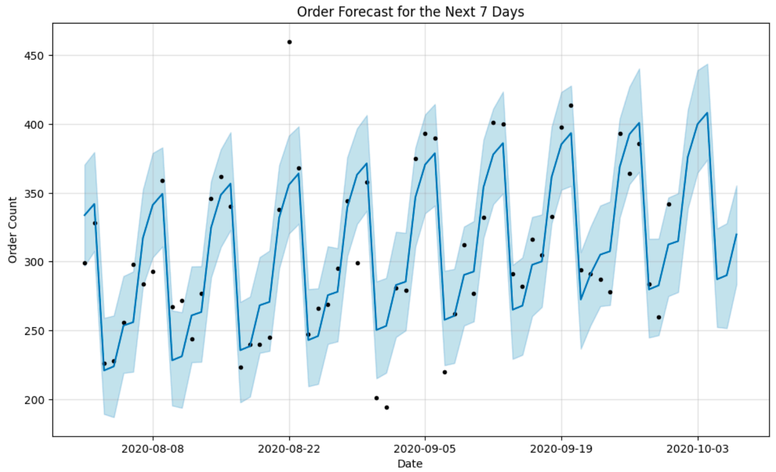
1



2

- Grid Search for best Hyperparameters
- Best ARIMA Order is (6, 1, 4)
- Mean Squared Error (ARIMA): 703.46

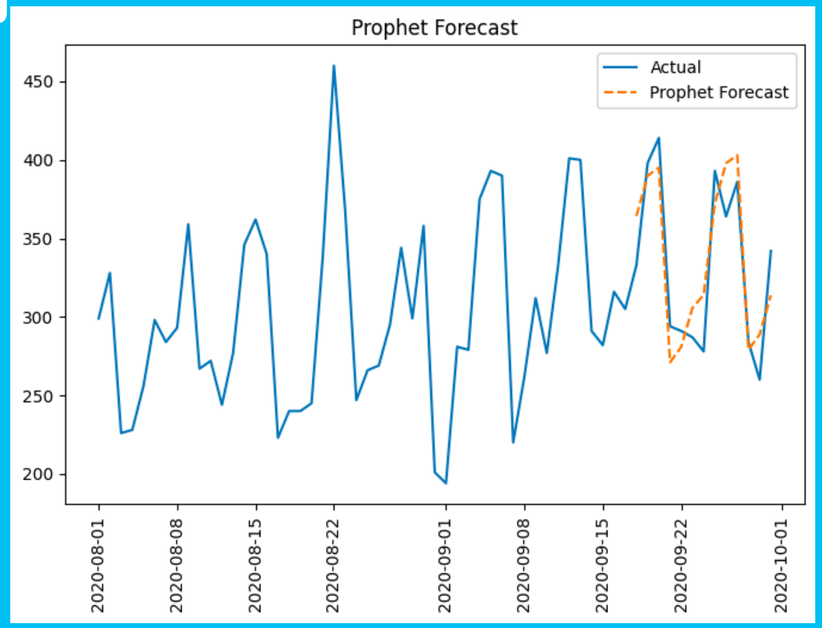
4



Using Prophet for forecasting because of the lower MSE

2020-10-01	315.0
2020-10-02	376.0
2020-10-03	400.0
2020-10-04	408.0
2020-10-05	287.0
2020-10-06	290.0
2020-10-07	320.0

3



- Grid Search for best Hyperparameters
- Best Paras(partial): yearly_seasonality: False; 'weekly_seasonality' : True, 'daily_seasonality': True, 'seasonality_mode': 'additive'...
- Mean Squared Error (Prophet): 554.45

FURTHER DISCUSSION

- The analyses are conducted based on limited data, both vertically and horizontally, introducing certain limitations. There are potential enhancements, including but not limited to:
  - Intuitively, with a year's worth of order data, more correlated patterns may emerge, such as those related to orders in different seasons or influenced by weather conditions.
  - The current estimation algorithm for delivery time exhibits a certain level of aggressiveness (or rationality), but real-world riders often adhere subjectively to their own methods of delivery within time constraints(i.e. Rules setup by Wolts). To delve into this behaviors, A/B testing can be implemented to enhance delivery efficiency and adjust platform delivery rules accordingly if necessary .
  - Geographically, areas with low order density relative to venues should be further investigated to determine whether additional efforts are needed in merchant recruitments.
  - Customer and venue portraits can be refined by employing clustering techniques with multidimensional data, facilitating precision marketing strategies.
  - Order forecasting can be improved through more sophisticated methods with the inclusion of additional data, particularly vertically. For instance, weather conditions should be considered a crucial factor influencing users' decisions to place online orders.