Exploring Instacart Data

Analyzing user preferences to predict when they will place an order

General Assembly Data Science

Capstone Project

Darshan Donthi

Background

- Instacart is a grocery delivery and pick-up service
- Released dataset in 2017 [1,2]
  - 3 million grocery orders
  - 200,000 users
- Details about items and users

Goals

What products are reordered the most?

When are these products ordered?

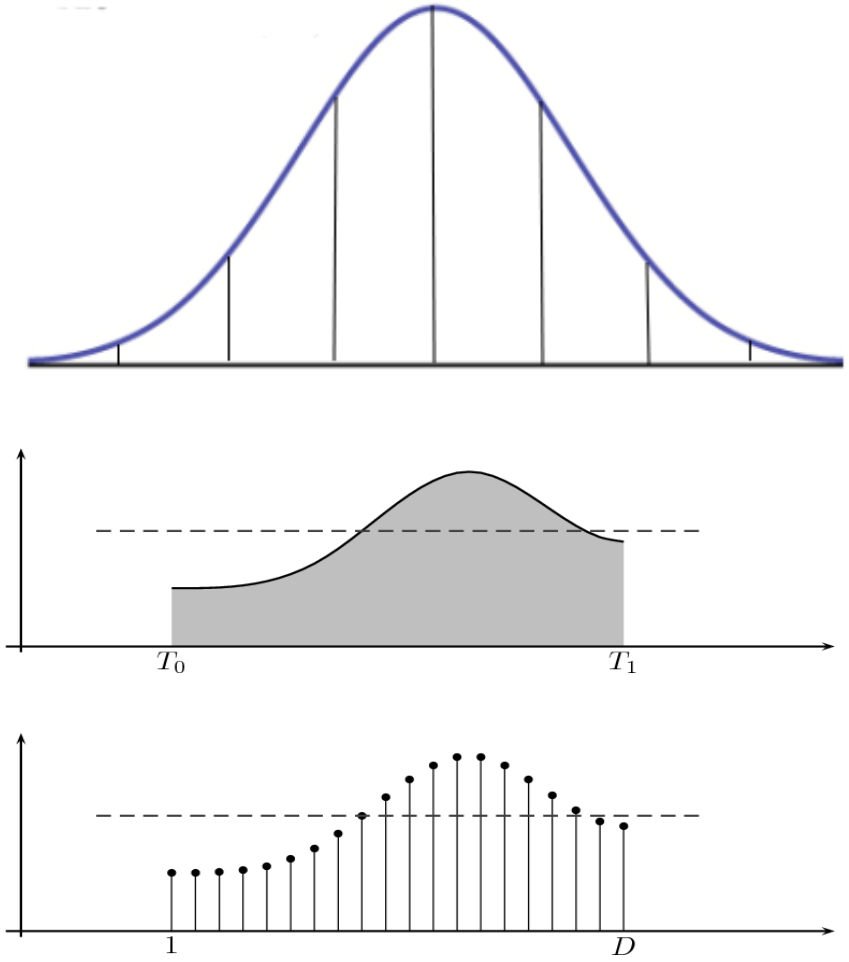
Predict the hour of day a user will place an order.

Motivation

- Improve sales revenue and profit-per-order for Instacart
- Recommend certain products during surge pricing hours [3]
- Suggest online bundle pricing deals rather than in-store

Metrics and Assumptions

- Using Mean Squared Error (MSE)
  - Lower error = higher accuracy
- Data is normally distributed
  - Central Limit Theorem
  - Use statistical functions
  - Identify outliers
- Time of day is continuous value



Approach and Process

Read, Clean, Explore

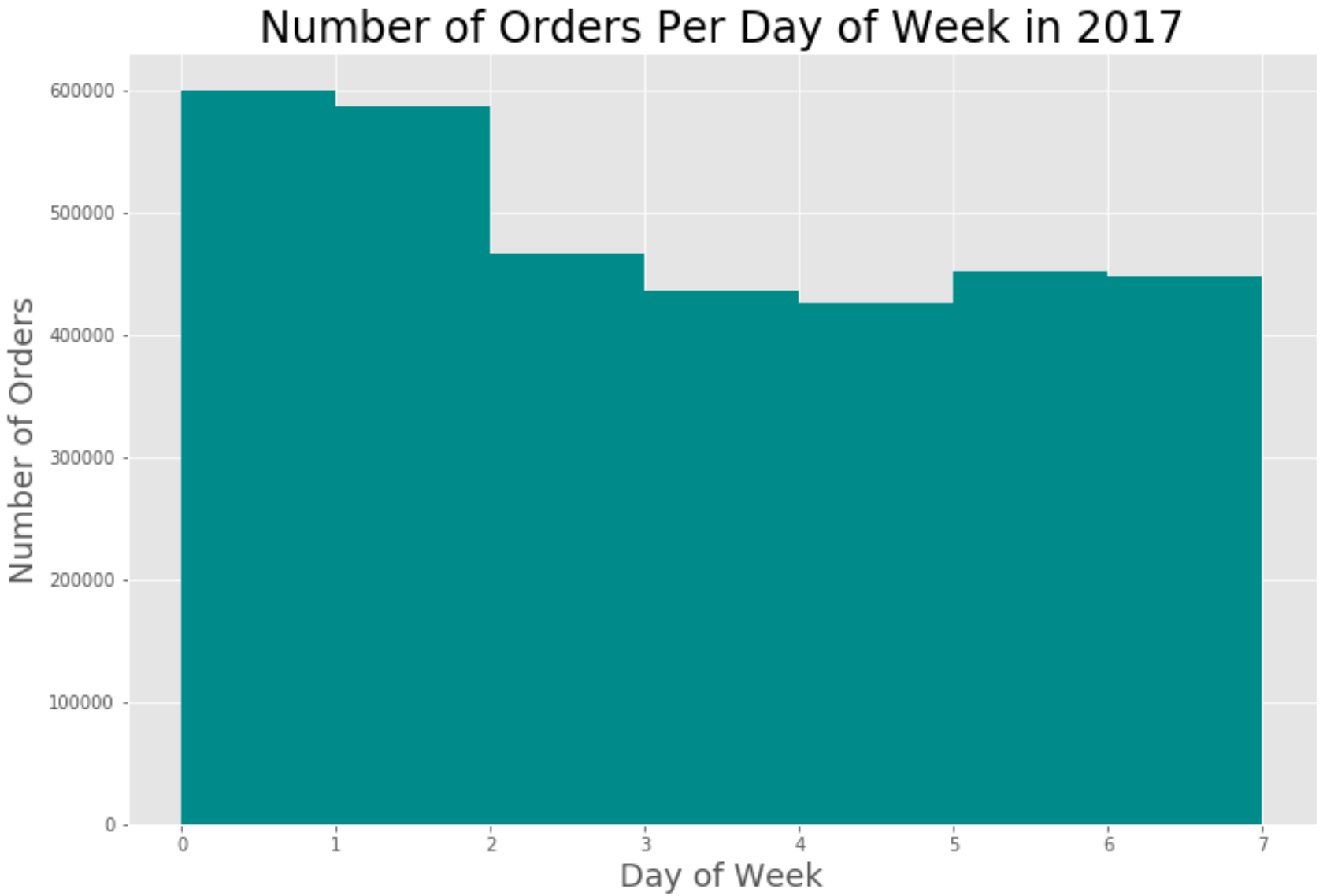
Answer Key Business Questions

Feature Engineering

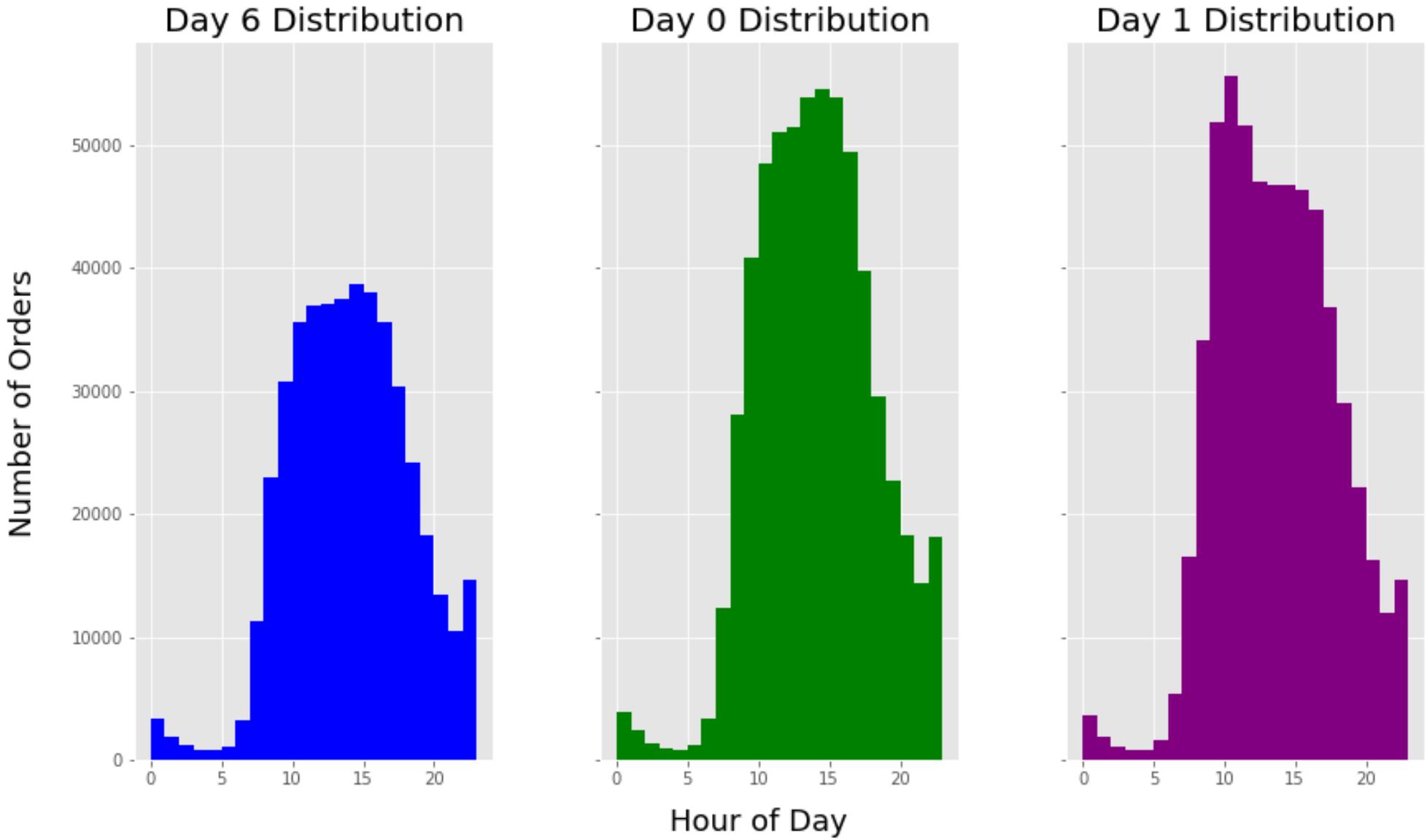
Predictive Models and Analysis

Future Work

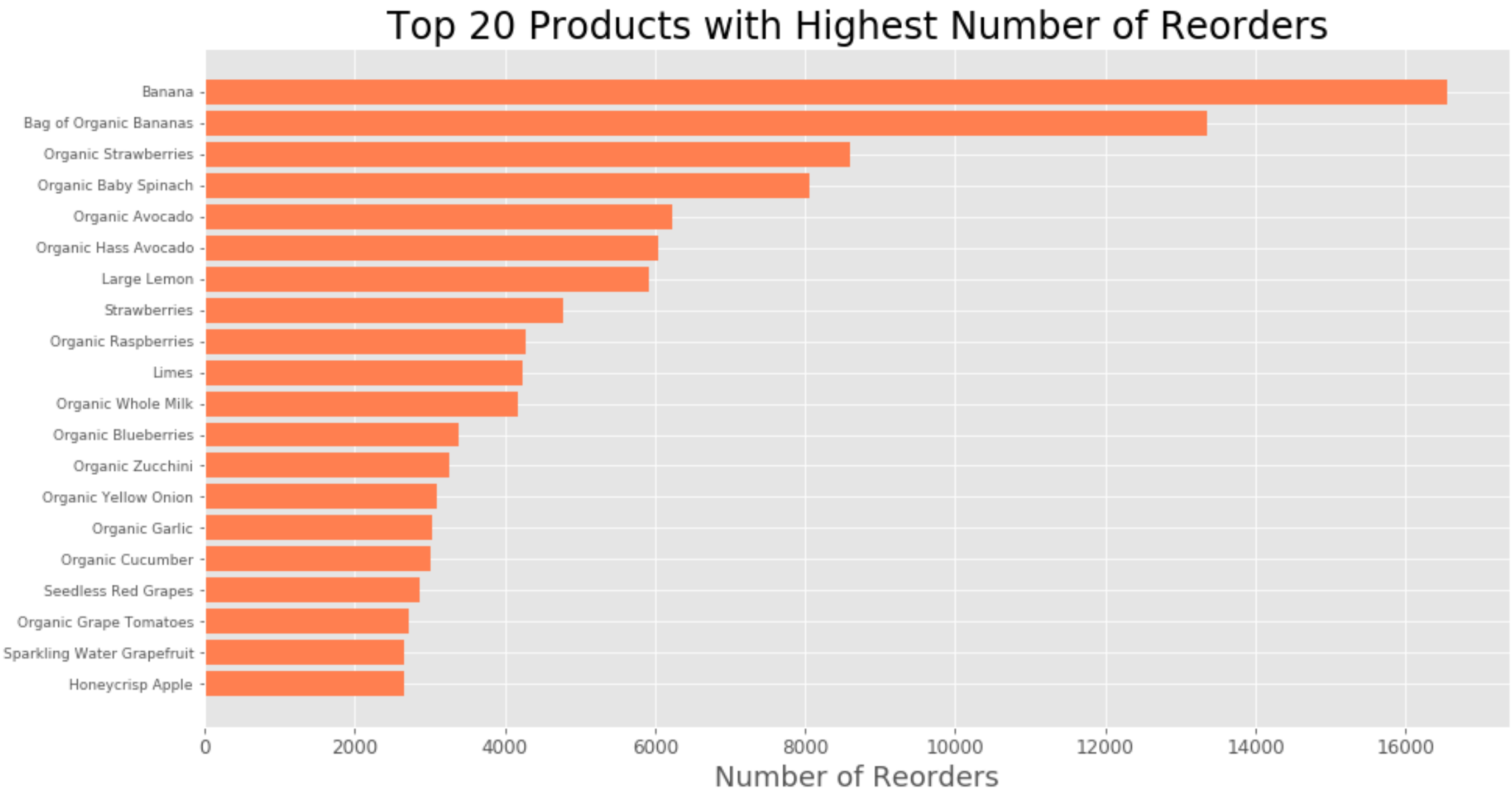
Exploration: What is Day 0 ?



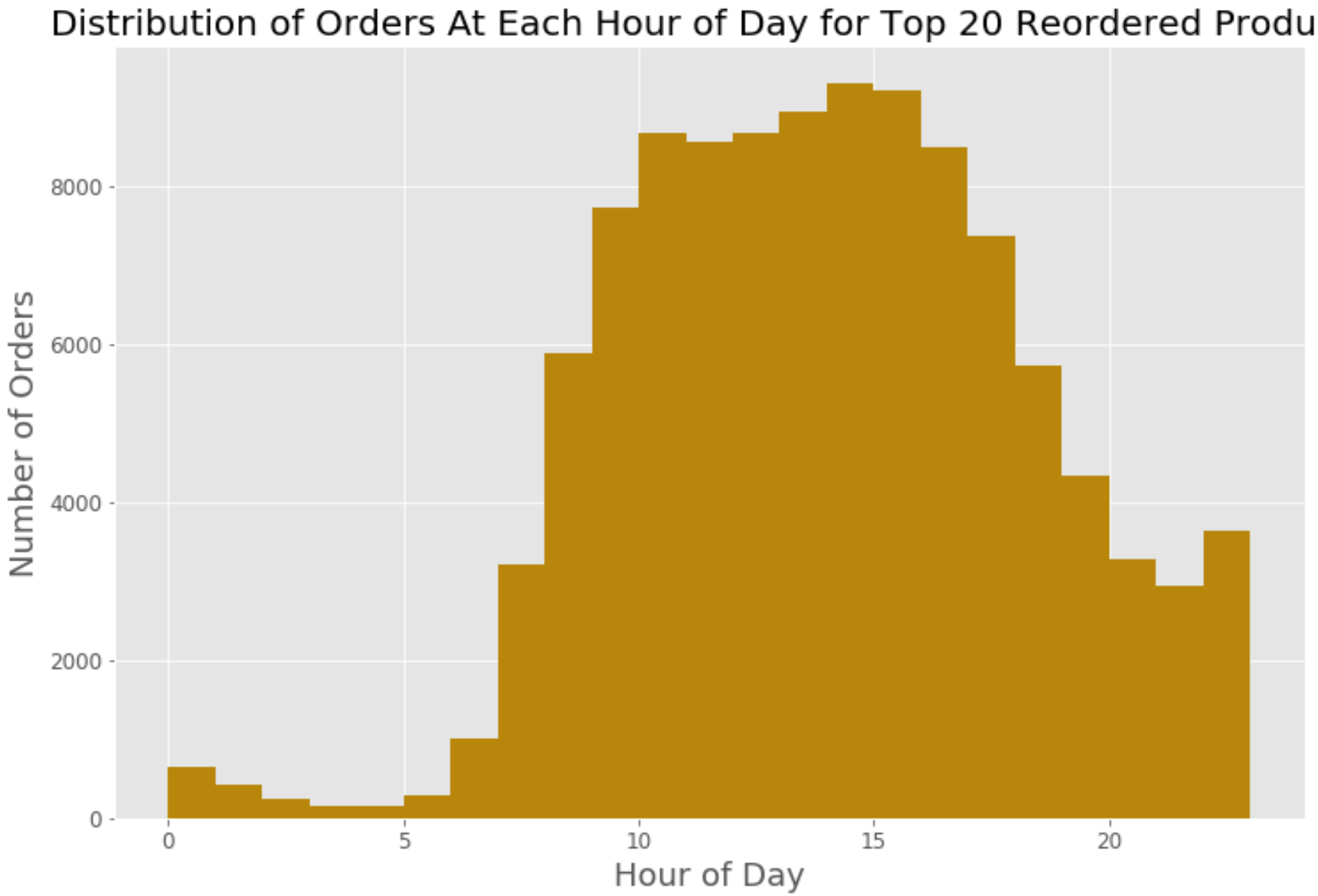
Exploration: What is Day 0 ?



What products are reordered the most?



What times of day are top products ordered?



Suggestions

New users: actively suggest produce or organic produce to establish interest

Repeat users: suggest bundling with a banana at peak hours

Feature Engineering

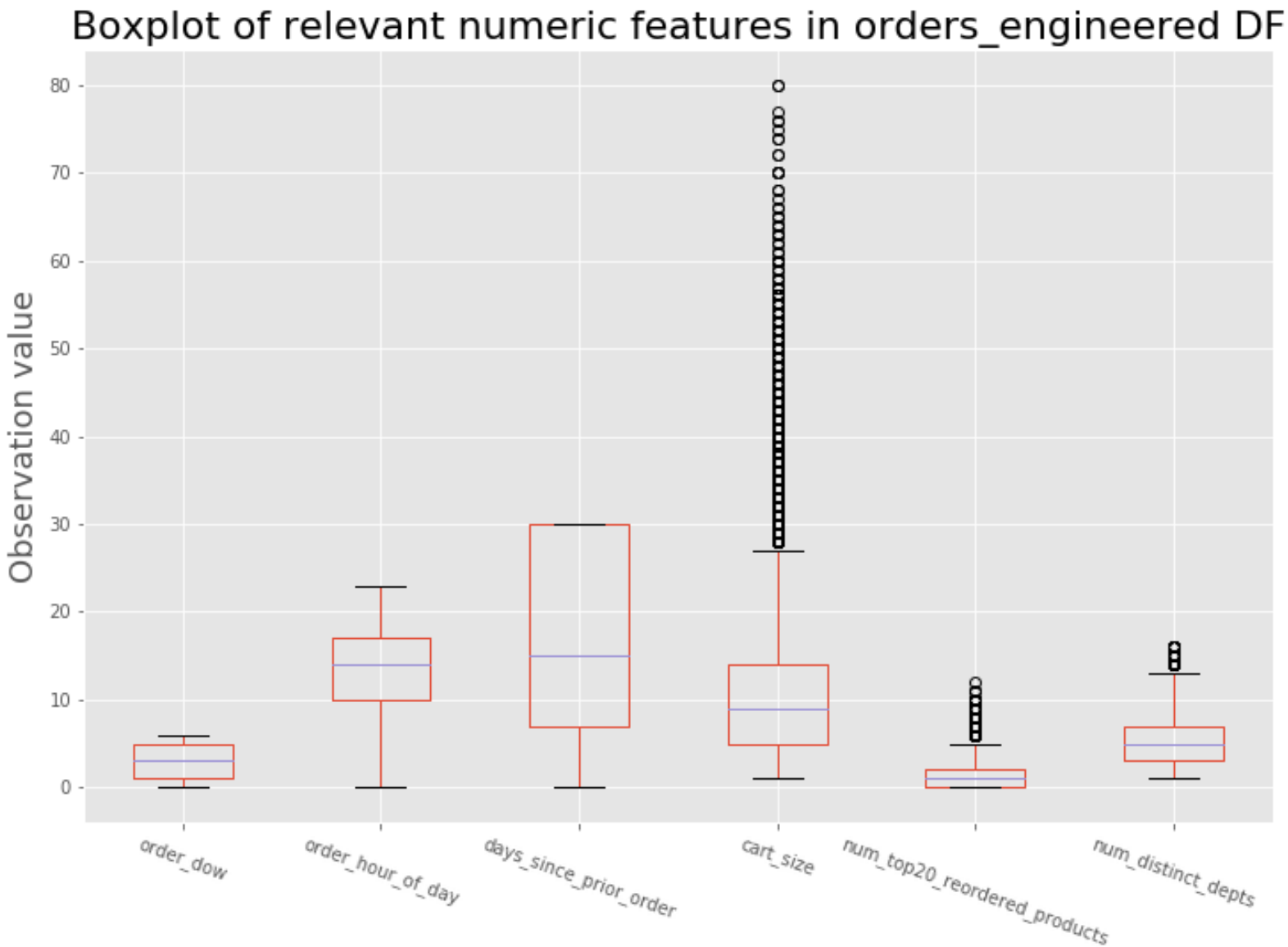
Build new features from existing data

1. Cart size (number of items)

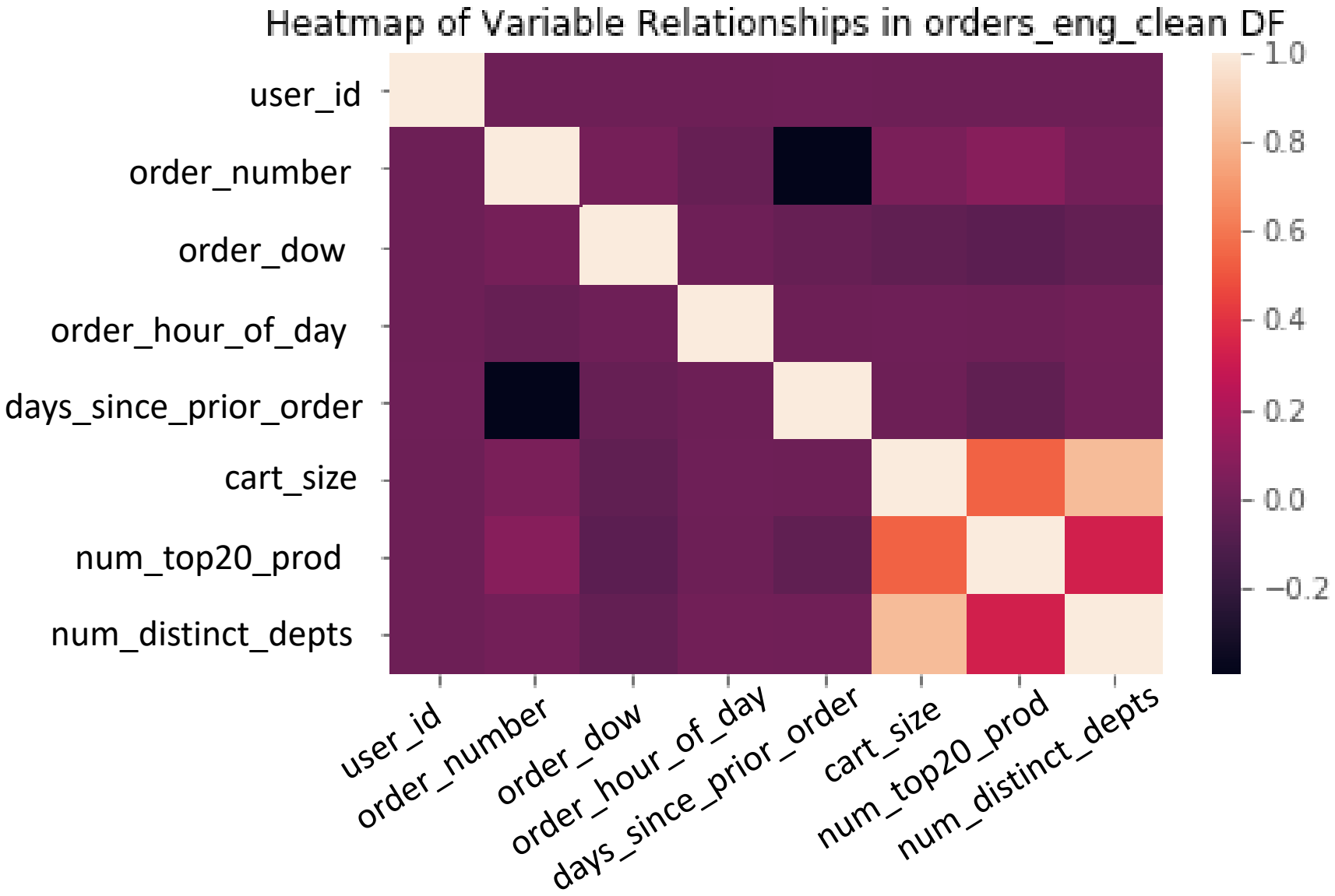
2. Number of top 20 products in cart

3. Number of distinct depts in cart

Handling Outliers

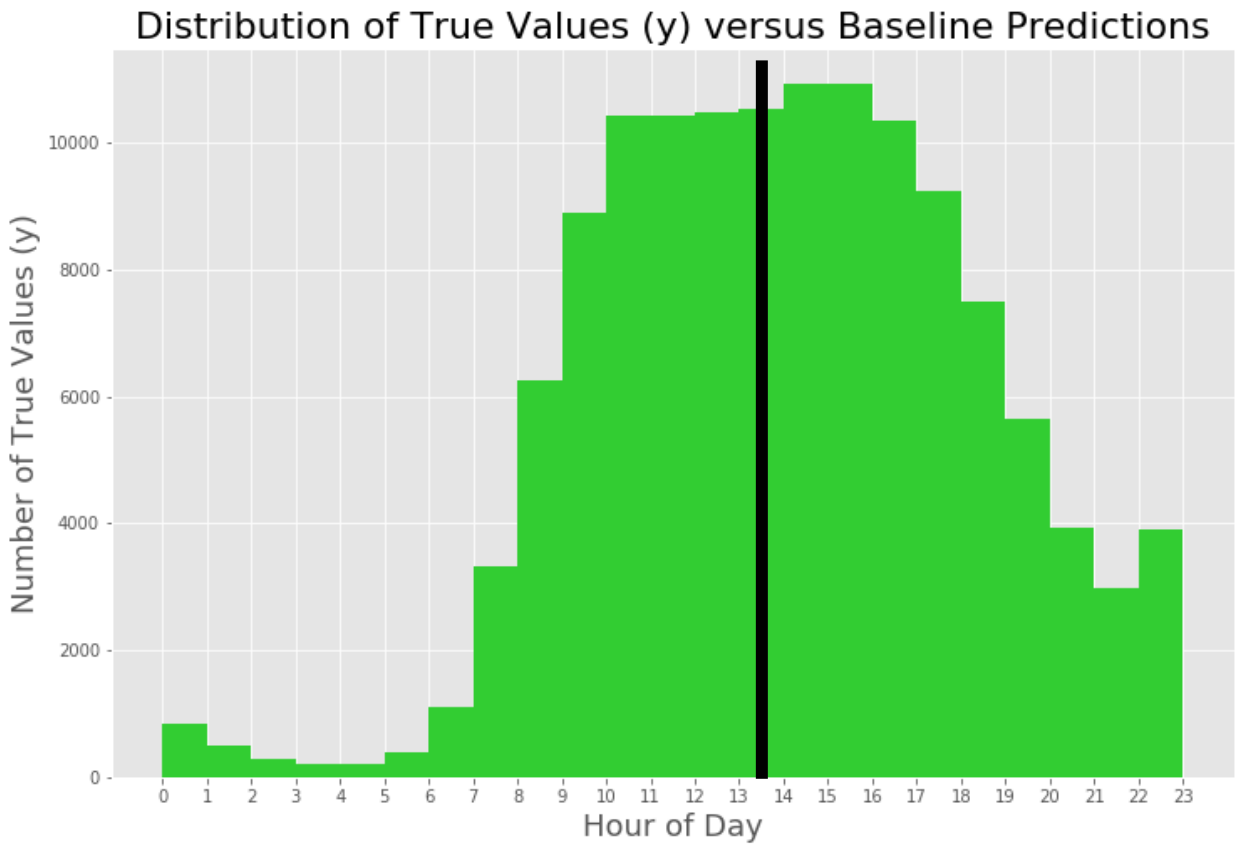


Variable Relationships



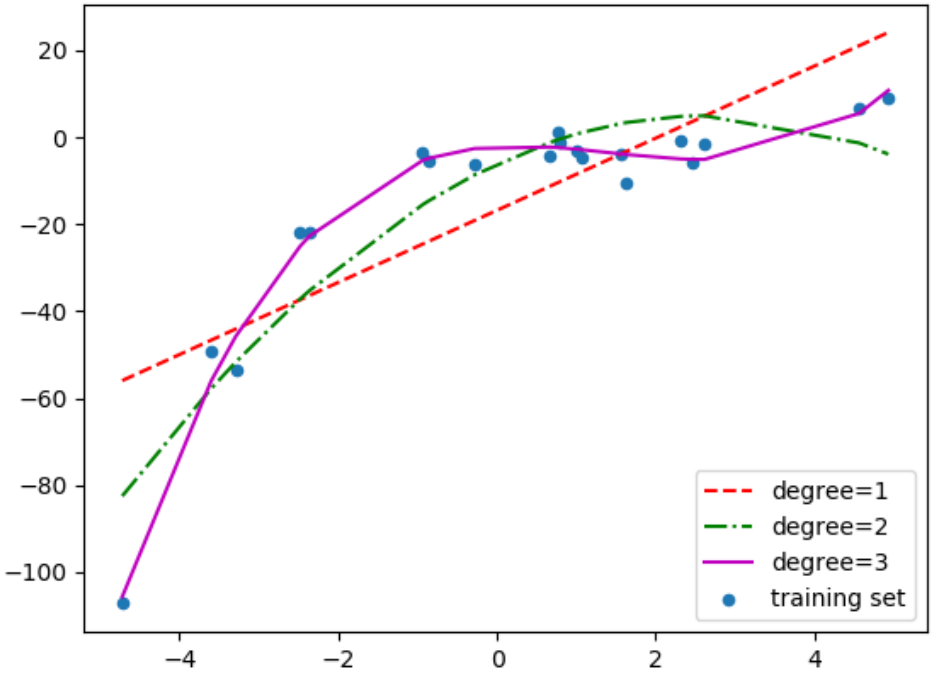
Predictive Modeling: Baseline

- Target Vector(y): order_hour_of_day
- Predictor Matrix (X): not needed*
- Prediction Values (y_hats):
  - Mean order_hour_of_day ≈ 13.5
  - All predictions use the mean value (“monkey throwing darts”)
- Calculate MSE between true values (y) and predicted values (y_hats)
  - Baseline MSE = 17.805



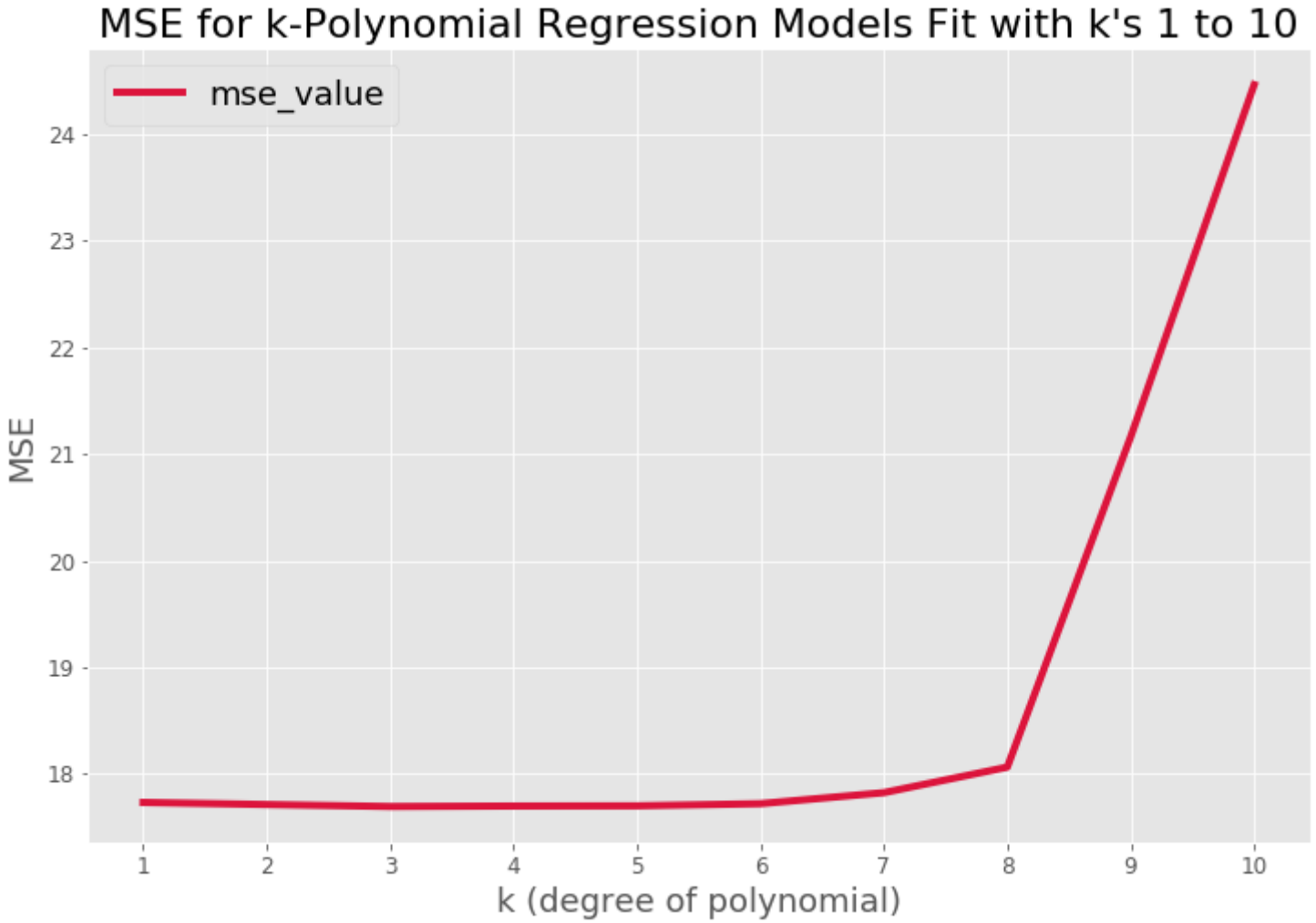
Predictive Modeling: Polynomial Regression

- Fit models with poly features for degrees 1-10
- Predictor Matrix (X):
  - order_number
  - days_since_ prior order
  - cart_size
  - num_distinct_depts
- Train/test split of 70/30, and set random seed
- Fit, transform, train, and test
- Prediction Values (y_hats): created from test set
- Calculate MSE between true values (y) and predicted values (y_hats)



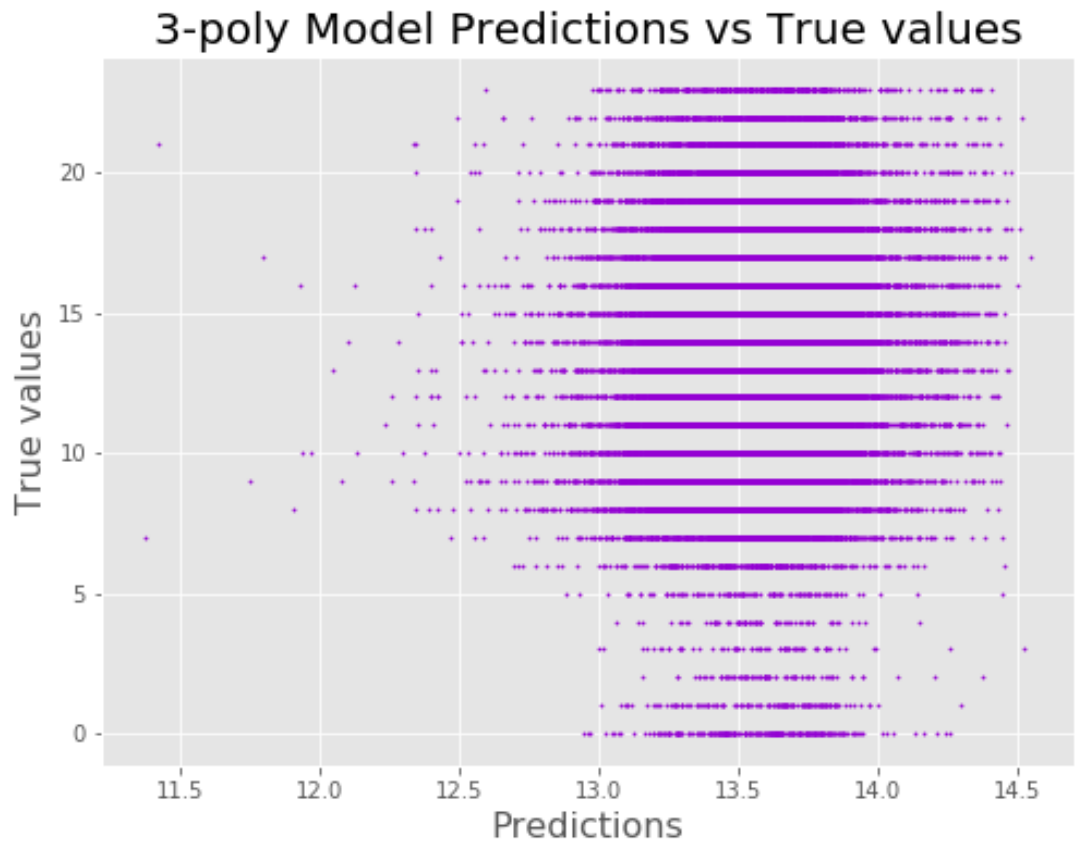
Predictive Modeling: Polynomial Regression

k	mse_value
1	17.733797
2	17.714342
3	17.695641
4	17.700164
5	17.701941
6	17.722147
7	17.82531
8	18.064848
9	21.160661
10	24.459326



Polynomial Regression Analysis

- Degree-3 model has decrease of 0.11 MSE compared to Baseline
- Translates to only 0.6% improvement
- Is this model actually better? → No
- Contributing factors:
  - Little correlation to order_hour_of_day
  - Train/test split



<u>Conclusion:</u>

This Degree-3 model is **not** useful for predicting order hour of day.

Future Work

Dataset has limited predictive capability for order hour of day

Map ranges of hours to intervals of day (e.g. Morning, Evening)

Find supplemental dataset with features indicating seasonality

Thank you for your attention!
Any Questions?

References

1) "3 Million Instacart Orders, Open Sourced", Accessed from https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2.

2) "The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <6/16/20>.

3) "How does Instacart make money?", Accessed from https://vator.tv/news/2016-08-02-how-does-instacart-make-money.

Images Used (Links)

- https://images.app.goo.gl/8rBxLrV2FDZf1jrF8
- https://images.app.goo.gl/uzJcsi6Uk6J6TdMo9
- https://images.app.goo.gl/pfdPoAWmCHcFG5286
- https://images.app.goo.gl/ESwhmhaWSunw3D3c8
- https://images.app.goo.gl/TD6gYpQB7NmFK35A8
- https://images.app.goo.gl/tViZyaZvktGddq1B6
- https://images.app.goo.gl/K9Tv2kN4XZ4iGjXQ8
- https://images.app.goo.gl/3rnXWmWMz1mkkgF86
- https://images.app.goo.gl/xrPwwjE5FuyDue1J6
- https://images.app.goo.gl/LmtyAYCJ16WRjXNbA
- https://images.app.goo.gl/shGsmBWHfGLoE4Lt5
- https://images.app.goo.gl/7P7PbdvYrQAVoJWC9
- https://images.app.goo.gl/YZXfAzLRBiDYnCEZ8
- https://images.app.goo.gl/RHtnPUJZU9jxj8HY7
- https://images.app.goo.gl/giRrWUdp3ymMjhaD7
- https://images.app.goo.gl/rBbiegNaDYKSCTNq9