REORDER PREDICTIONS:
MARKET BASED ANALYSIS OF INSTACART ORDERS

Maithili Joshi

WHAT IS INSTACART

Grocery delivery service

Founded in 2012

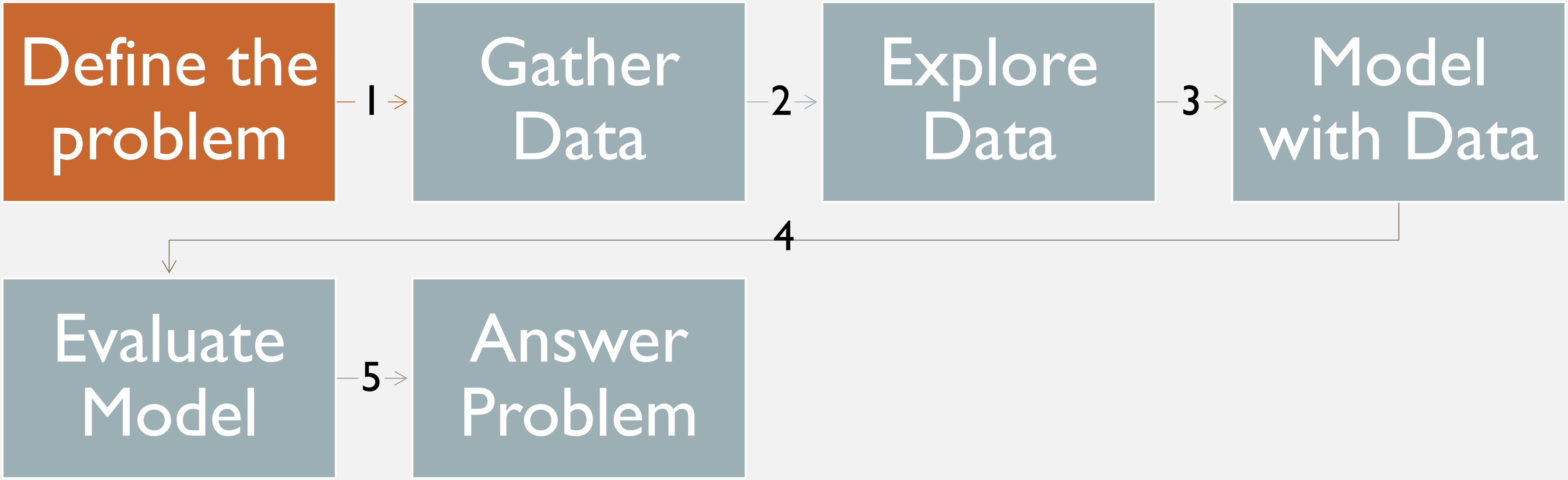
50k Shoppers

20k Stores

Serves 5500 cities

Valued at $7.6 billion

DATA SCIENCE PROCESS



Define the problem

1

Gather Data

2

Explore Data

3

Model with Data

4

Evaluate Model

5

Answer Problem

THE PROBLEM

- “In this competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order. ”
- Classification problem
- Evaluation metric is the F1 score

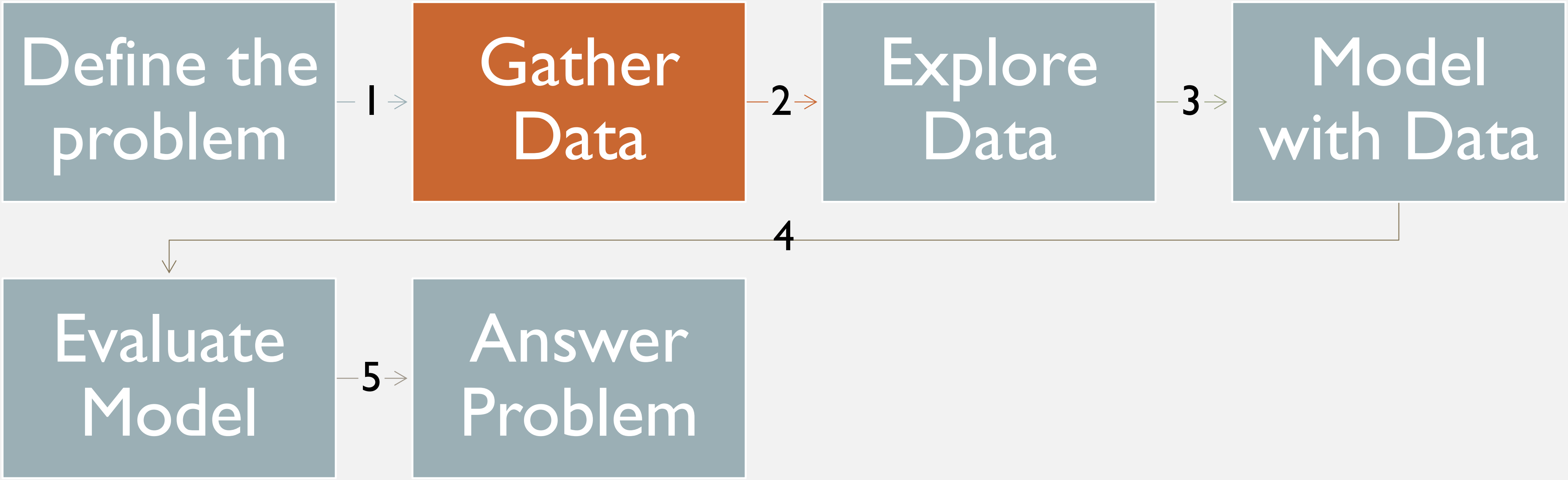
F1 SCORE

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$

DATA SCIENCE PROCESS



Define the problem

1

Gather Data

2

Explore Data

3

Model with Data

4

Evaluate Model

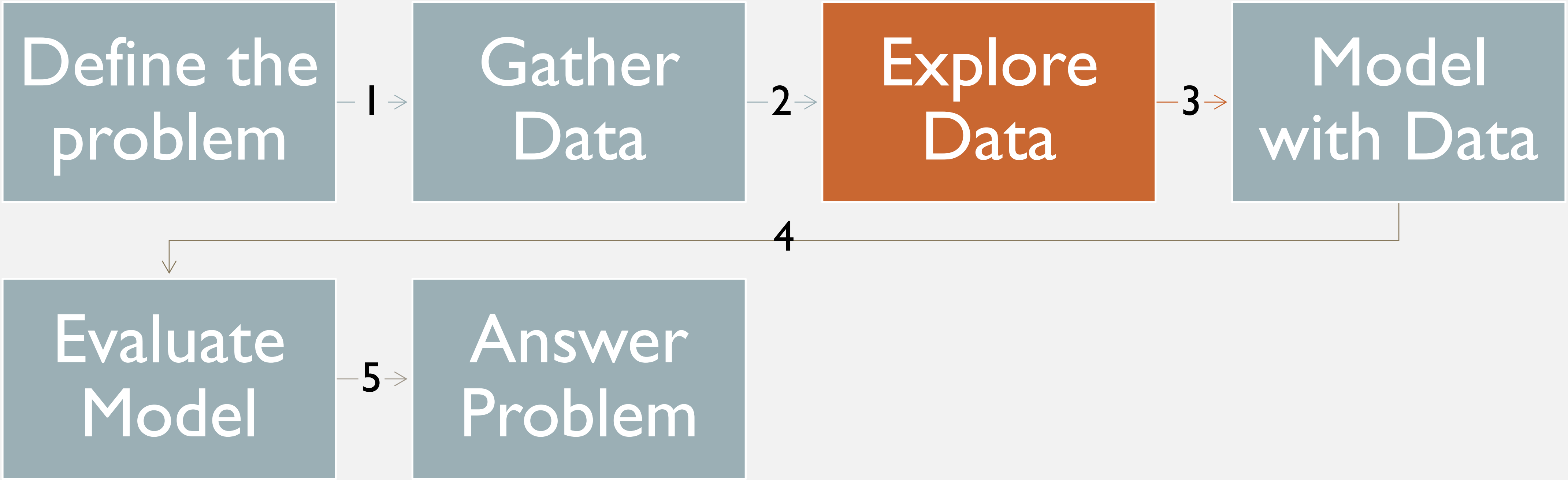
5

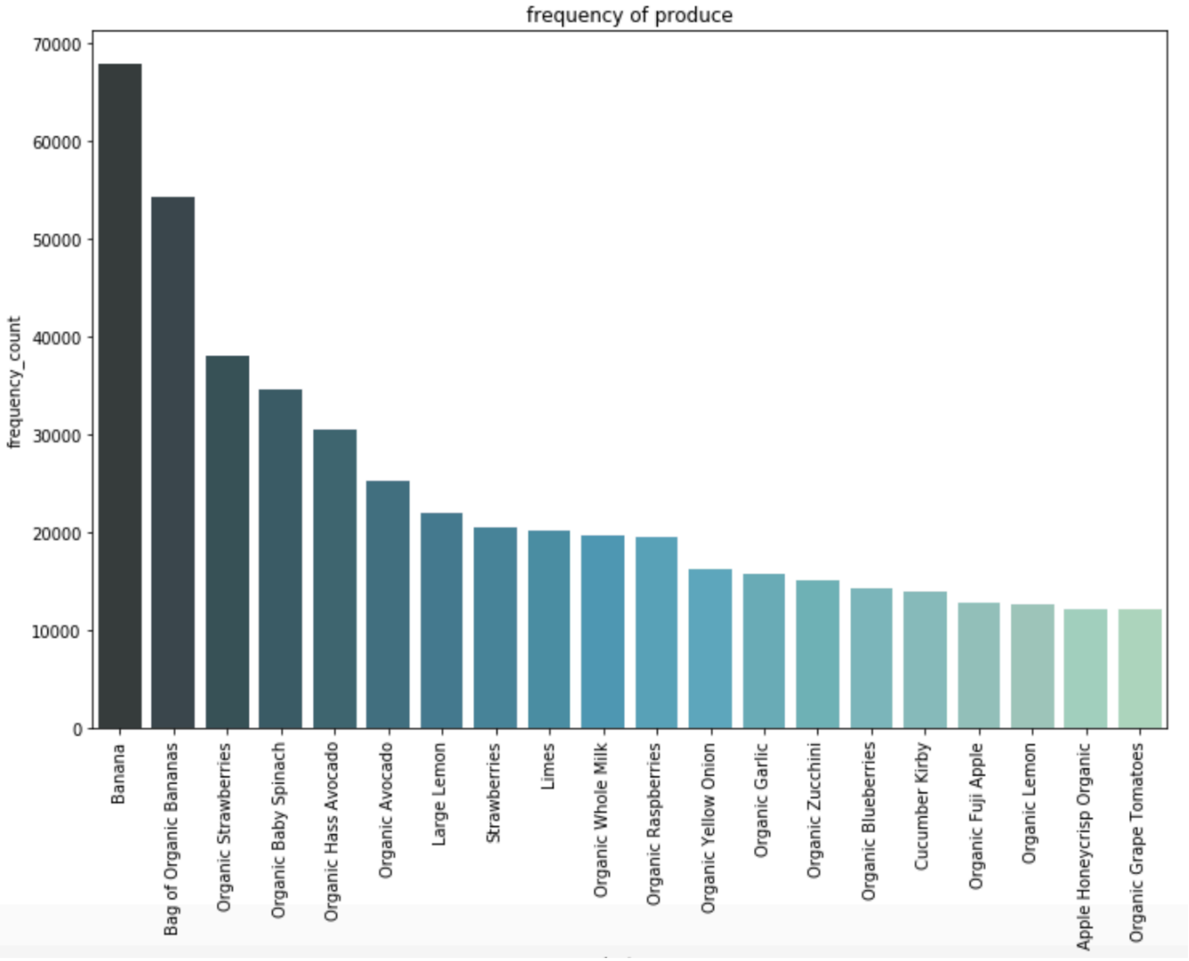
Answer Problem

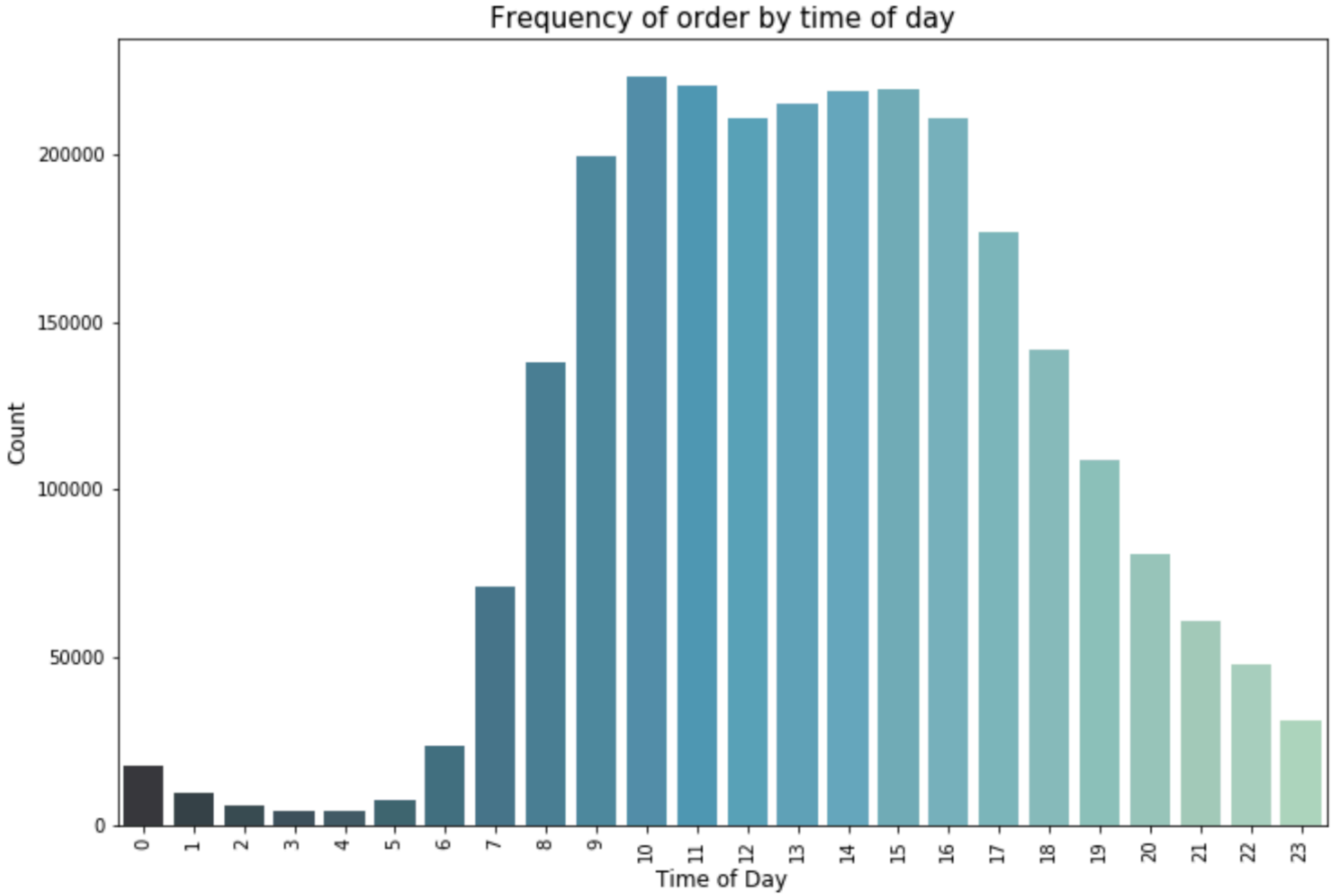
THE DATA

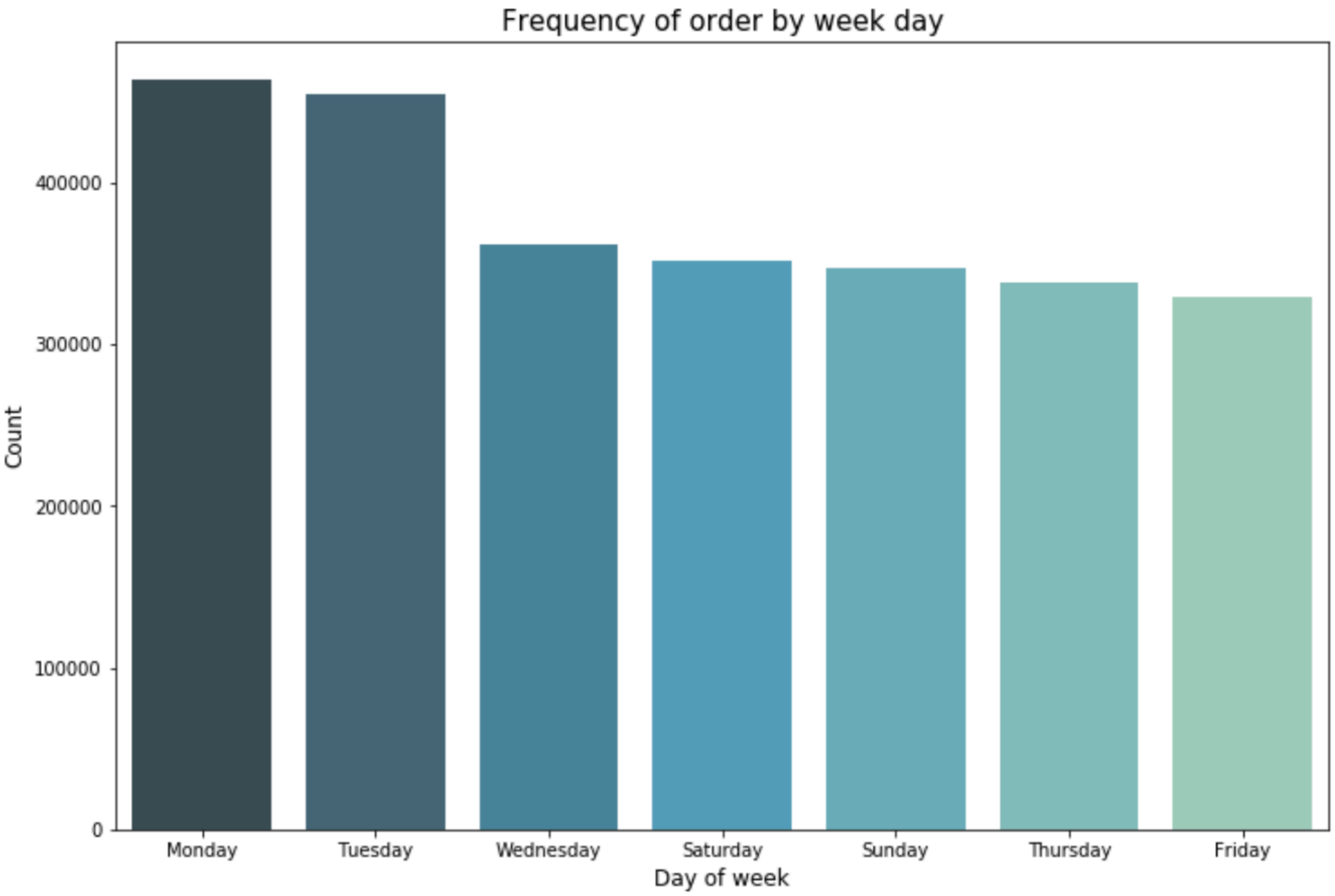
- sample of > 3 million grocery orders from > 200,000 Instacart users
- Aisles
- Departments
- Products
- Orders
- Orders_prior
- Orders_train

DATA SCIENCE PROCESS

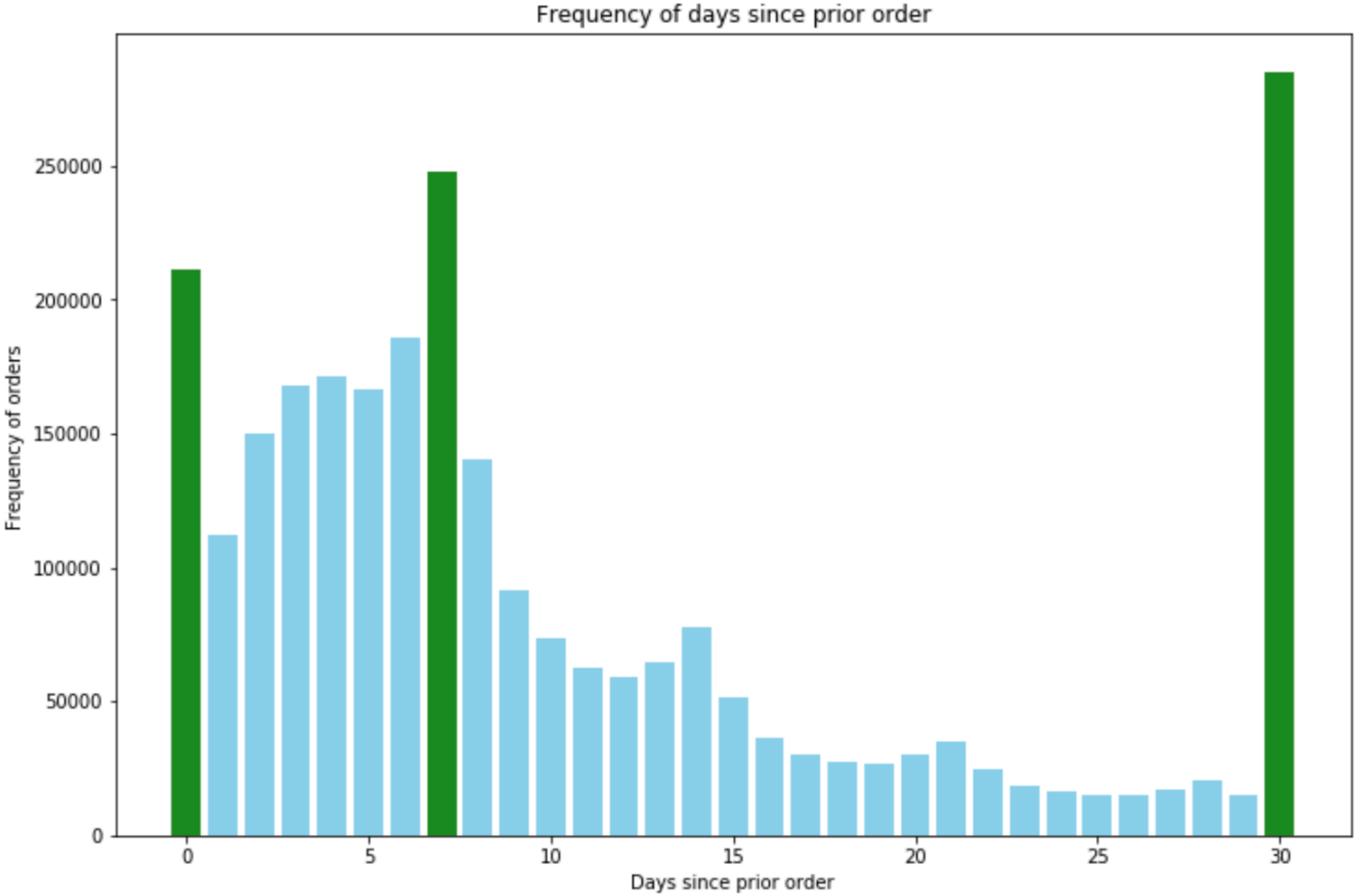


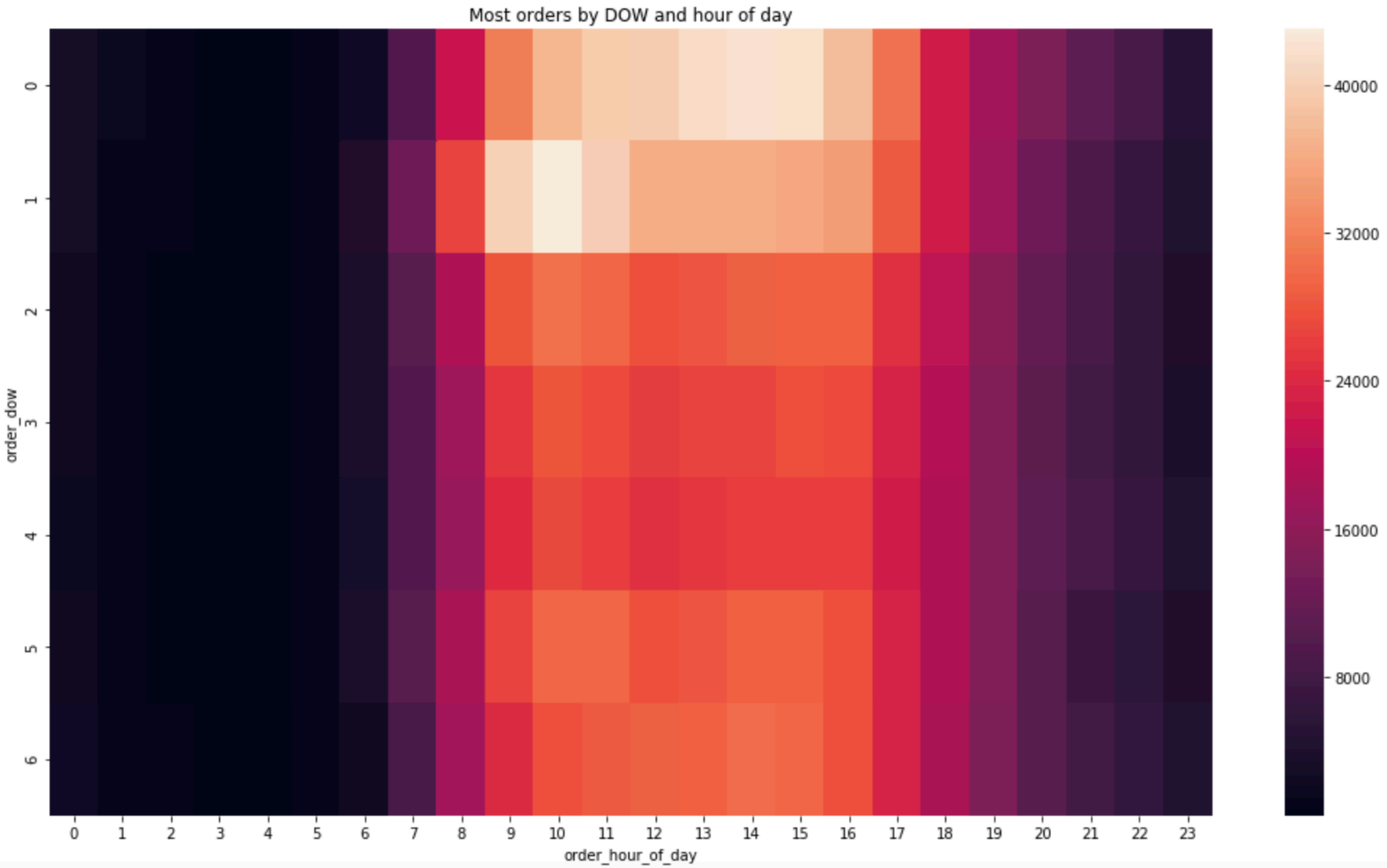

frequency of produce
frequency_count
0
10000
20000
30000
40000
50000
60000
70000
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Apple Honeycrisp Organic
Organic Grape Tomatoes


Frequency of order by time of day
Count
200000
150000
100000
50000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Time of Day



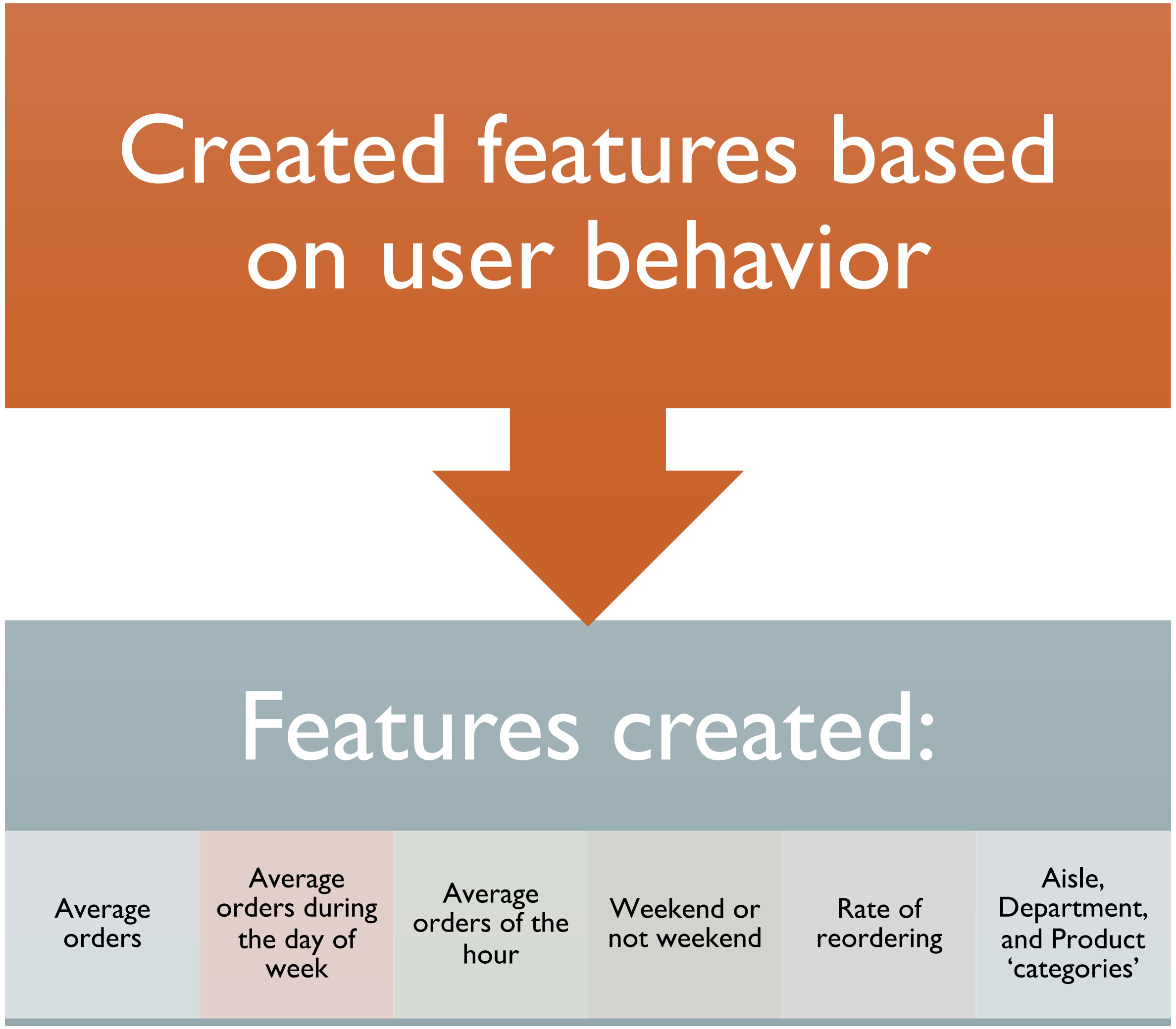
Frequency of order by week day


Frequency of days since prior order
Frequency of orders
250000
200000
150000
100000
50000
0
0
5
10
15
20
25
30
Days since prior order



Most orders by DOW and hour of day

FEATURE ENGINEERING

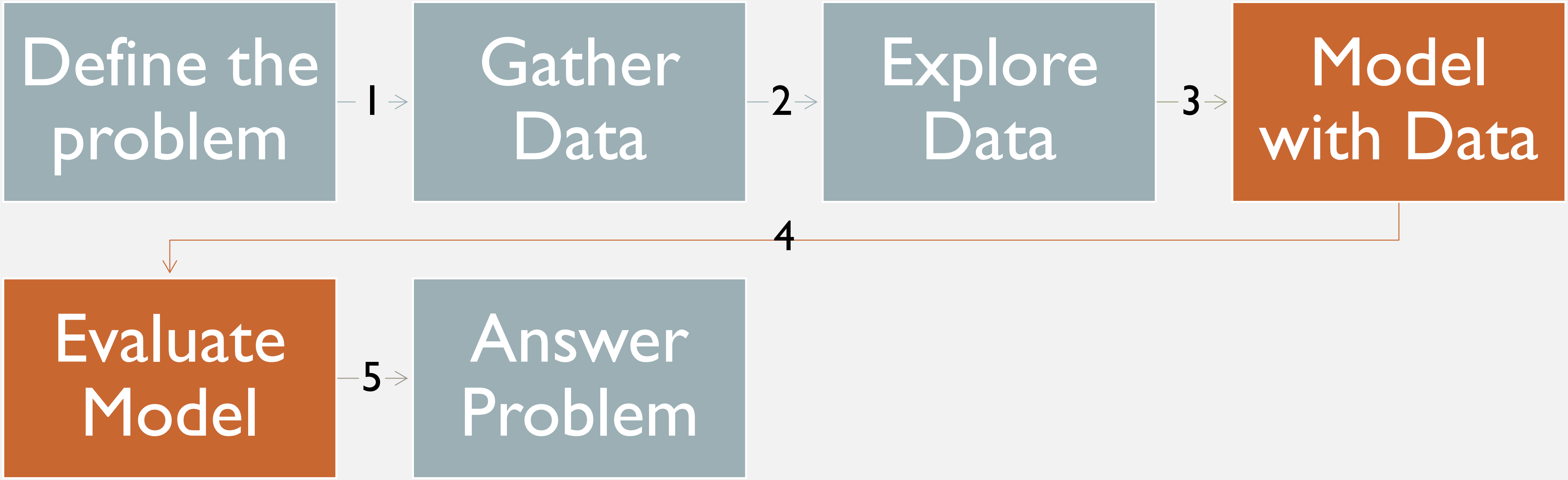


Created features based on user behavior

Features created:

- Average orders
- Average orders during the day of week
- Average orders of the hour
- Weekend or not weekend
- Rate of reordering
- Aisle, Department, and Product 'categories'

DATA SCIENCE PROCESS



Define the problem

1

Gather Data

2

Explore Data

3

Model with Data

4

Evaluate Model

5

Answer Problem

3 MODELS

LOGISTIC REGRESSION

RANDOM FORESTS

XGBOOST

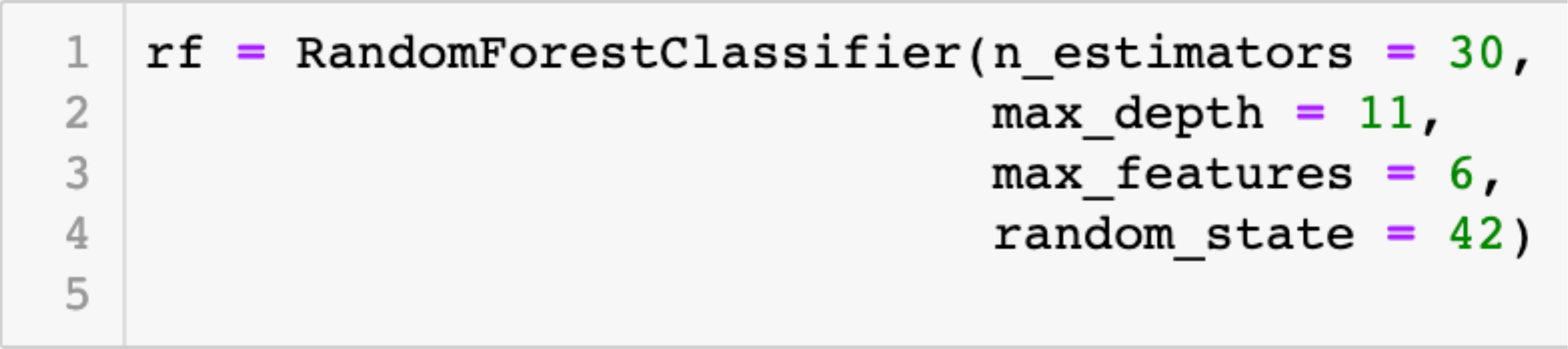
59%

- Baseline accuracy score

LOGISTIC REGRESSION

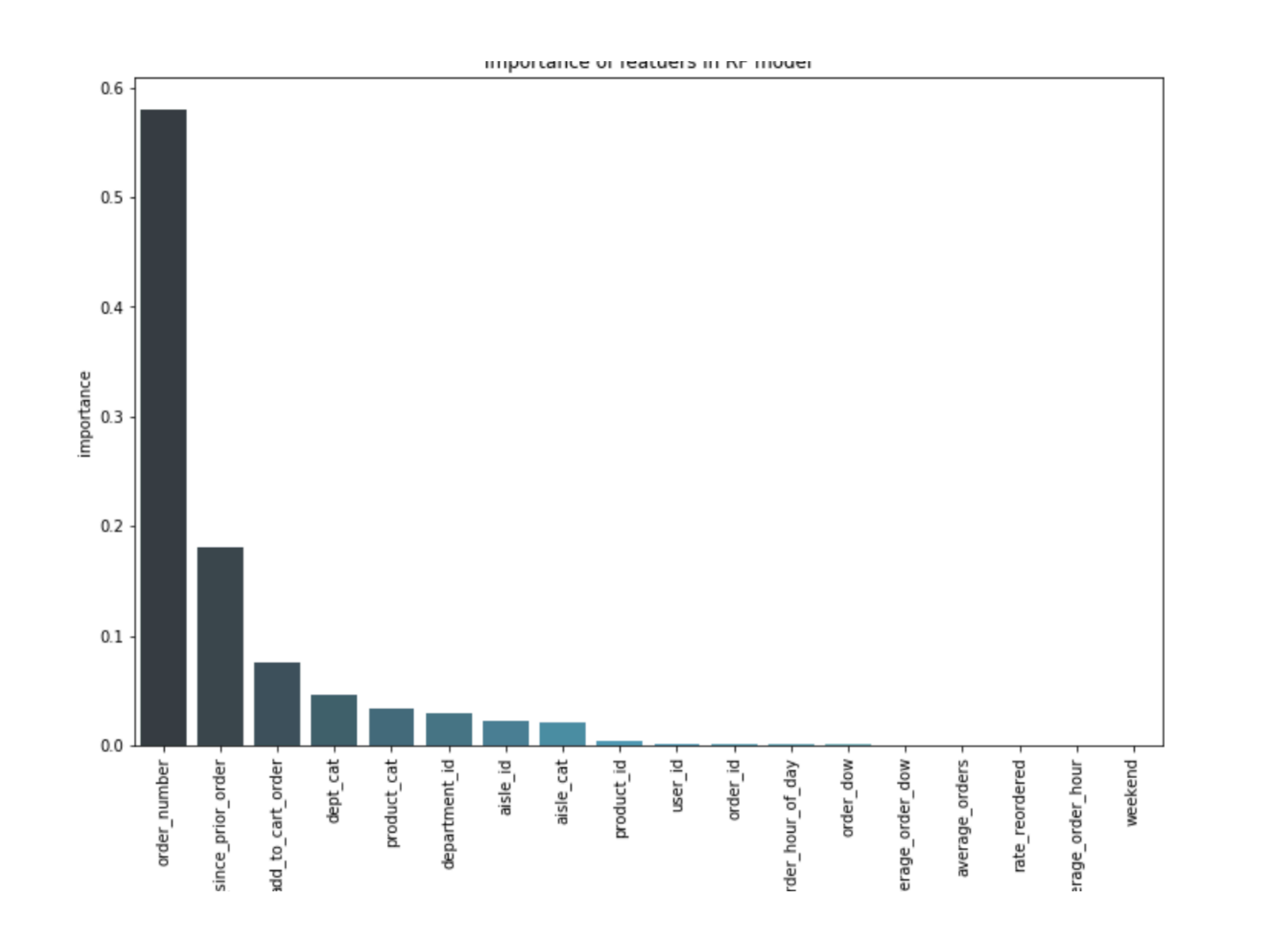
- F1 Score = 75.78%
- Why?
  - Classic regression
  - Ease of interpretability
  - Why not

RANDOM FORESTS



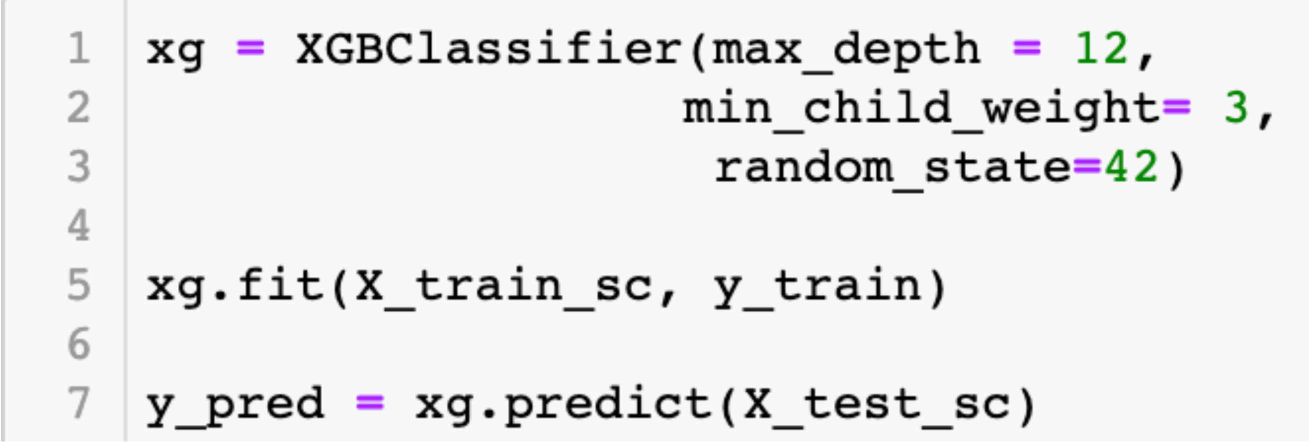
```python
rf = RandomForestClassifier(n_estimators = 30,
                            max_depth = 11,
                            max_features = 6,
                            random_state = 42)

```

RANDOM FORESTS



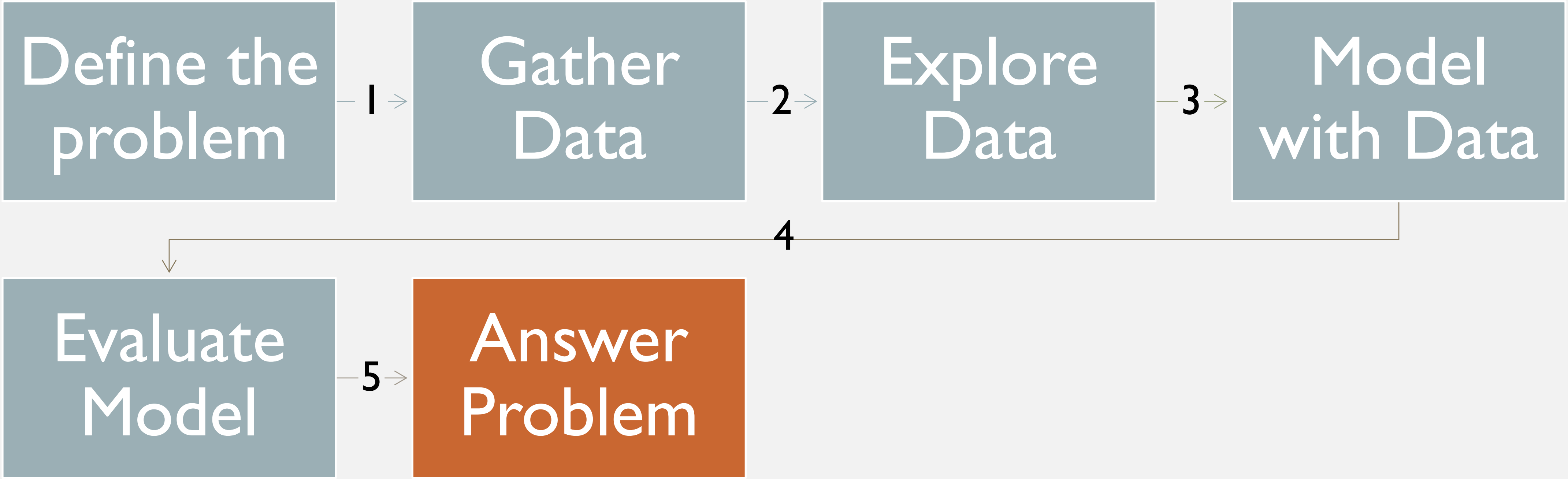
XG BOOST

- F1 score = 80%



```
xg = XGBClassifier(max_depth = 12,
                  min_child_weight= 3,
                   random_state=42)

xg.fit(X_train_sc, y_train)

y_pred = xg.predict(X_test_sc)
```

DATA SCIENCE PROCESS



Define the problem

1

Gather Data

2

Explore Data

3

Model with Data

4

Evaluate Model

5

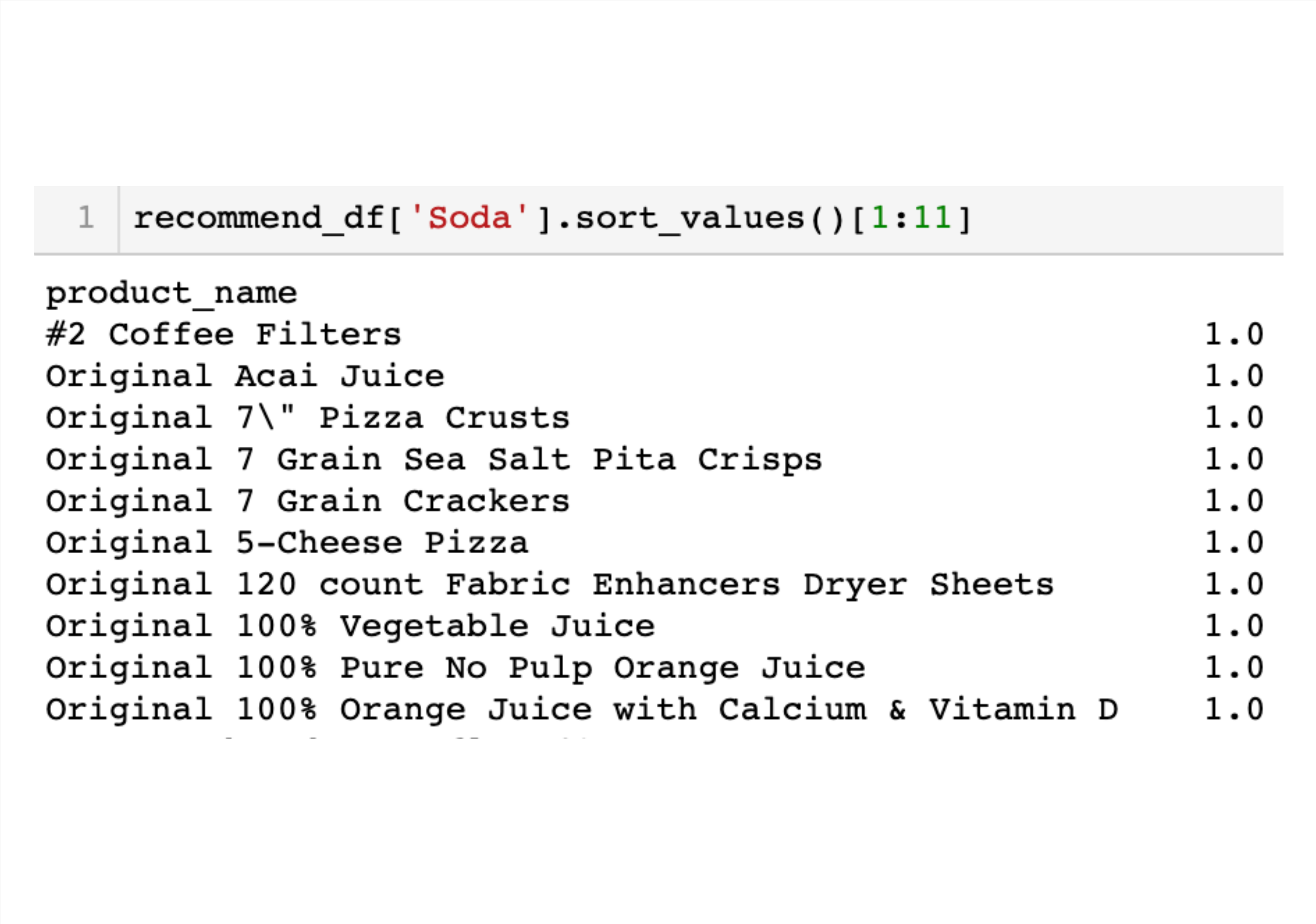
Answer Problem

THE PROBLEM

- “In this competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order. ”

RECOMMENDER SYSTEM

- By product ID and whether or not it was re-ordered:



```python
recommend_df['Soda'].sort_values()[1:11]
```

```
product_name
#2 Coffee Filters                                             1.0
Original Acai Juice                                           1.0
Original 7\" Pizza Crusts                                     1.0
Original 7 Grain Sea Salt Pita Crisps                         1.0
Original 7 Grain Crackers                                     1.0
Original 5-Cheese Pizza                                       1.0
Original 120 count Fabric Enhancers Dryer Sheets              1.0
Original 100% Vegetable Juice                                 1.0
Original 100% Pure No Pulp Orange Juice                       1.0
Original 100% Orange Juice with Calcium & Vitamin D           1.0
```

NEXT STEPS

- More product features
- More order features
- Words2Vec
    - Creating a more accurate recommender system

THANK YOU

RESOURCES

- https://www.kaggle.com/c/instacart-market-basket-analysis
- https://deepai.org/machine-learning-glossary-and-terms/f-score
- https://medium.com/syncedreview/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283