Next Product Buy

Analysis and Prediction

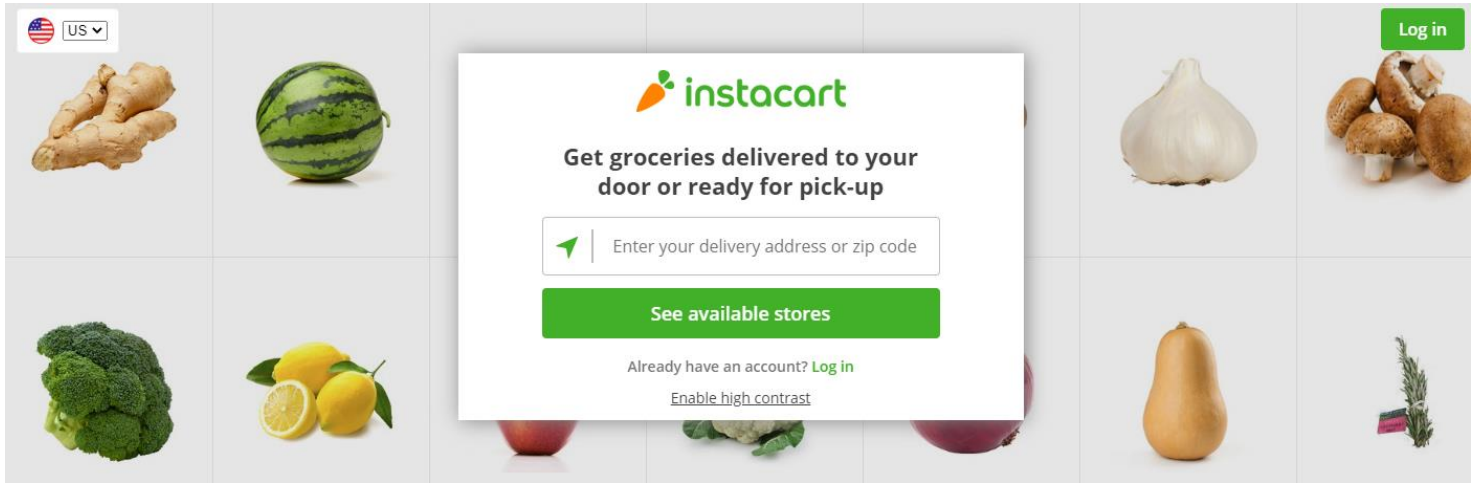
by Alex Dance

Agenda

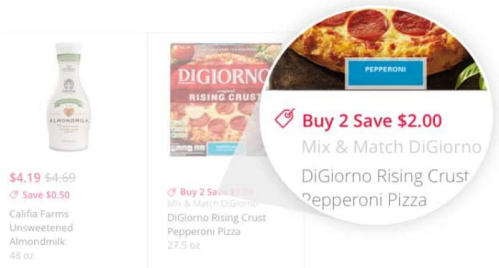
- Setting the scene / problem
- Initial observations / recommendations
- Analytic approach / recommendations – using Python
- My time management

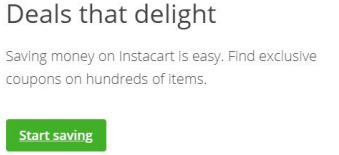
Instacart – The uber eats of grocery shopping in the US and Canada

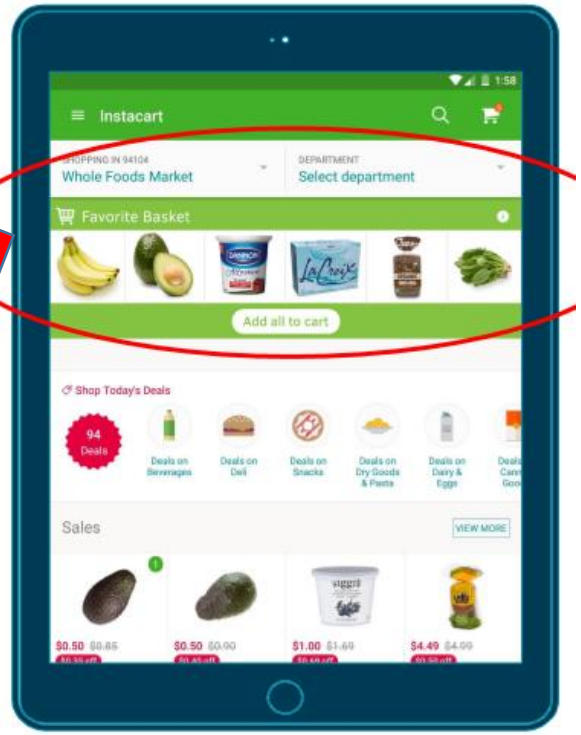
- Online grocery delivery / pick up
- Revenue $2B USD
- Valuation >$4B US
- Have partners for (eg Aldi) for inventory management



Brief







Criteria of assessment

1. Insights Presentation
   a. Structure of presentation
   b. Variety of insights determined from the data
   c. **A determination of an item most likely added to a customer's next order**
   d. Technical principles articulated appropriately

Remarketing + email marketing with a "discount" on these products

2.

Time Management

   a. Structured approach to the process
   b. Breakdown of high-level coverage in how the time was spent working on the scope
   c. This should demonstrate your methodology of approaching a request

https://www.kaggle.com/c/instacart-market-basket-analysis/data

For a better, more personalised customer experience

DATA OVERVIEW – if interested – to answer the question "A determination of an item most likely added to a customer's next order?"

Unique

```
order_products_train.head()
```

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0
3	1	49683	4	0
4	1	43633	5	1

```
order_products_prior .head()
```

Out[10]:

	order_id	product_id	add_to_cart_order	reordered
0	2	33120	1	1
1	2	28985	2	1
2	2	9327	3	0
3	2	45918	4	1
4	2	30035	5	0

50K products
134 aisles
21 departments

3.4M orders
206K users

```
orders.head()
```

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

```
aisles.head()
```

[4]:

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation

```
departments.head()
```

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce
4	5	alcohol

```
sample.head()
```

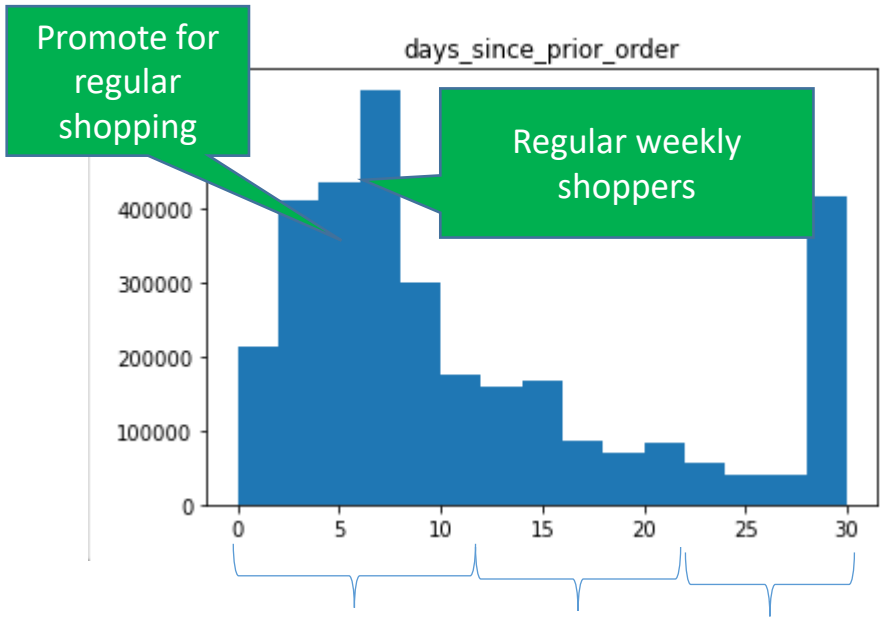
	order_id	products
0	17	39276 29259
1	34	39276 29259
2	137	39276 29259
3	182	39276 29259
4	257	39276 29259

```
products.head()
```

Out[7]:

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13

High # days since last order highlights 1) Some customers are not buying regularly 2)Potential Data Issues



When > 30 days it is recorded as 30 days (PROBLEM)

- What is the real length
- Is this a data quality issue?
- Is there an opportunity to get them back?

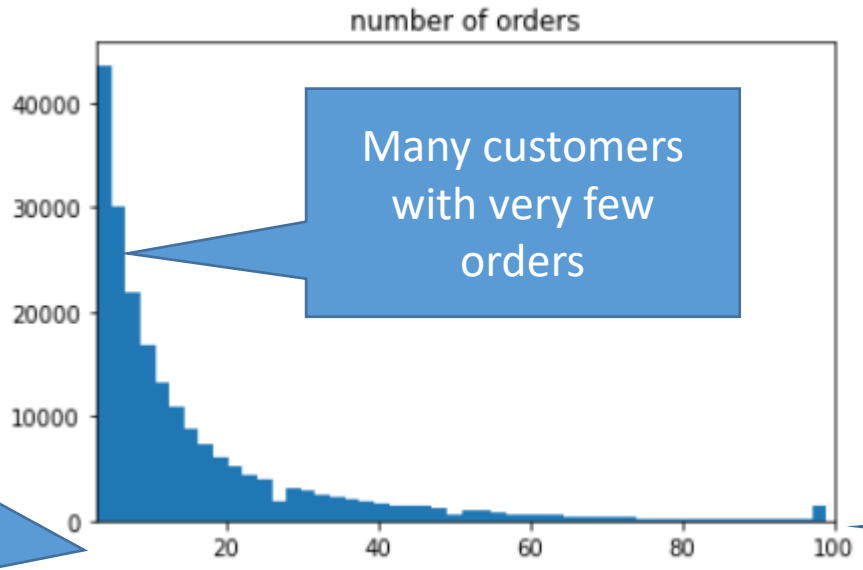
Some customers are not buying products often enough

Is this a data capture / card usage issue?

- Are they going to competitor
- Are they living in a rural area
- Is the shopping purchase not being picked up properly

Why not go in 15 days

There are a Lot of CUSTOMERS who aren't buying very often

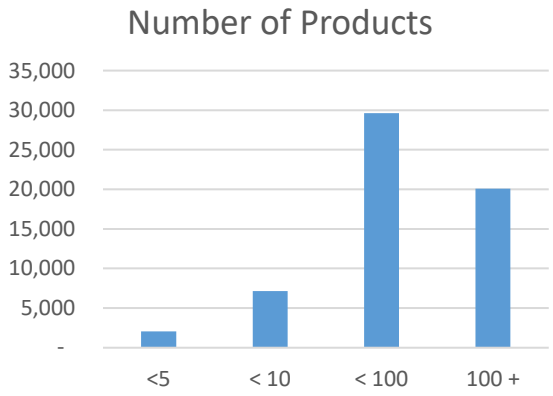


Many customers with very few orders

There is no data with 1 or 2 orders which may be intentionally left out.

Is there a maximum of 100

There are a LOT of products (possibly suppliers) with VERY LOW sales



Over 2,000 products had < 5 sales

End of Life

- If already ended then OK
- If not ended then perhaps should end

Beginning of life

- If right at beginning – OK
- If not getting traction - review

Low sales

- Why
- Potential Review

Data issue

Wastage

Storage problem

Transport problem

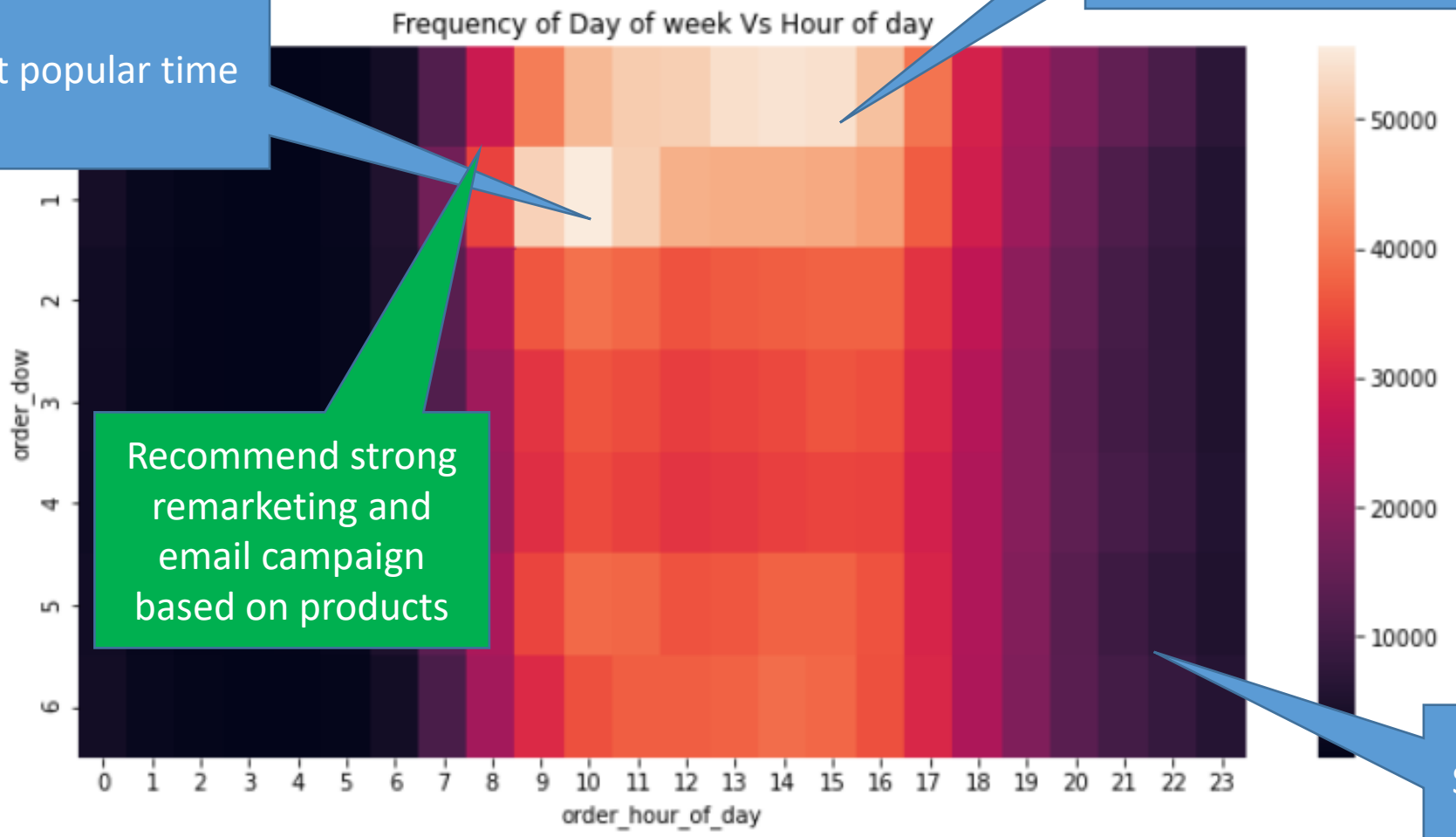
NOT Promoting products well

OR "SMALL RETAILERS" – with multiple low selling products

Promote small retailers more

OR OK as business strategy is to take on small sales items / customers

Time of day and day of week



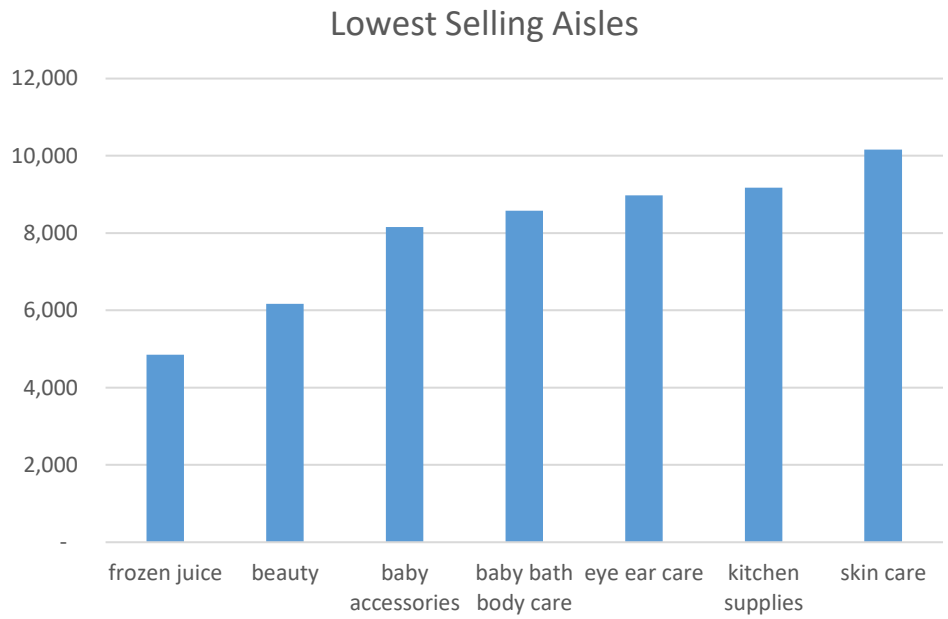
Most popular time

Most popular period

Recommend strong remarketing and email campaign based on products

Sales continue into the night

Overall 134 aisles. Fresh fruit and vegies are the biggest sellers. The bottom sellers are very low.



Extremely low volumes

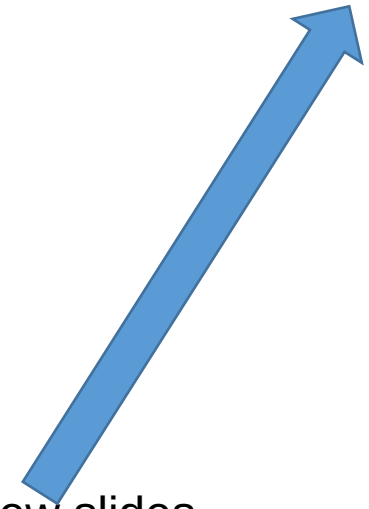
Analytic Approach using Python and Excel for some graphs

Initial

- Looked at the question
- Downloaded the data
- Looked at data tables and links
- Shared with fellow students
- Initial EDA (30 minutes)
- Looked at the SQL question

Mid review

- Looked at approach of others
- More EDA – found issues
- Presentation
  - Worked out the “story”
  - Put some meat behind a few slides
- Finalised approach
- Kicked off the SQL answers



Modelling Prepare

- More analysis
- Collated all info on user id, orders, by product (reordered count + ratio)
- First order from a customer
- Found reorder ratio
- For each user ID – product ID, order ratio and other
- Modelled on Train data
- Results in submission file

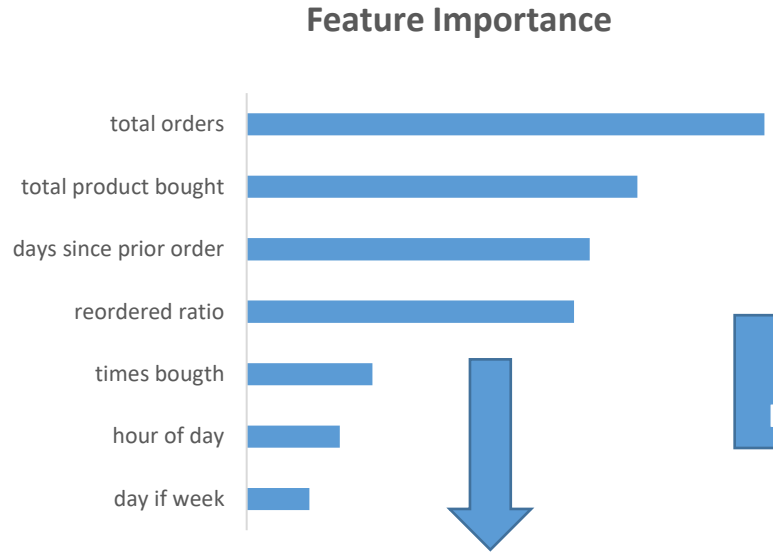
Finalised

- SQL
- Analysis / modelling
- Presentation

Ran a XG boost - Binary Logistic model

Train Data

XG BOOST



For 66K unique recommendations

For 75K orders

Order ID	Recommended Products
2774568	17668 18599 21903 23650 24810 39190 43961 47766
1528013	8424 21903 38293
1376945	8309 13176 14947 27959 28465 33572 34658 35948...
1356845	7076 10863 11520 13176 14992
2161313	196 10441 11266 12427 14715 27839 37710
1416320	5134 17948 21137 21616 21903 24561 24852 28985...
1735923	196 12108 15131 17008 31487 35123
1980631	6184 9387 13575 13914 22362 41400 46061

Order id 2774568 Barbeque Kettle Cooked Flavor Potato Chips +Garlic Couscous +Organic Baby Spinach + others

$22M / pa benefits of implementing

Plug numbers

	Assumptions	Benefit / pa
More Sales from recommendation engine	12% more product sales $10 a sale	$5,000,000
More relevant products	2% more product sales $10 a sale	$2,000,000
Better reactivation of customers	5% more orders $100 an order	$15,000,000
		$22,000,000

RESULTING IN

A BETTER

CUSTOMER EXPERIENCE

Many ‘optimistic’ Guesses to show thought about the issue

$22M / pa benefits of implementing

Plug numbers

	Assumptions	Benefit / pa
More sales from recommendation analysis	12% more product sales $10 a sale	$5,000,000
More relevant products so happier customers	2% more product sales $10 a sale	$2,000,000
Better reactivation of customers	5% more orders $100 an order	$15,000,000
		$22,000,000

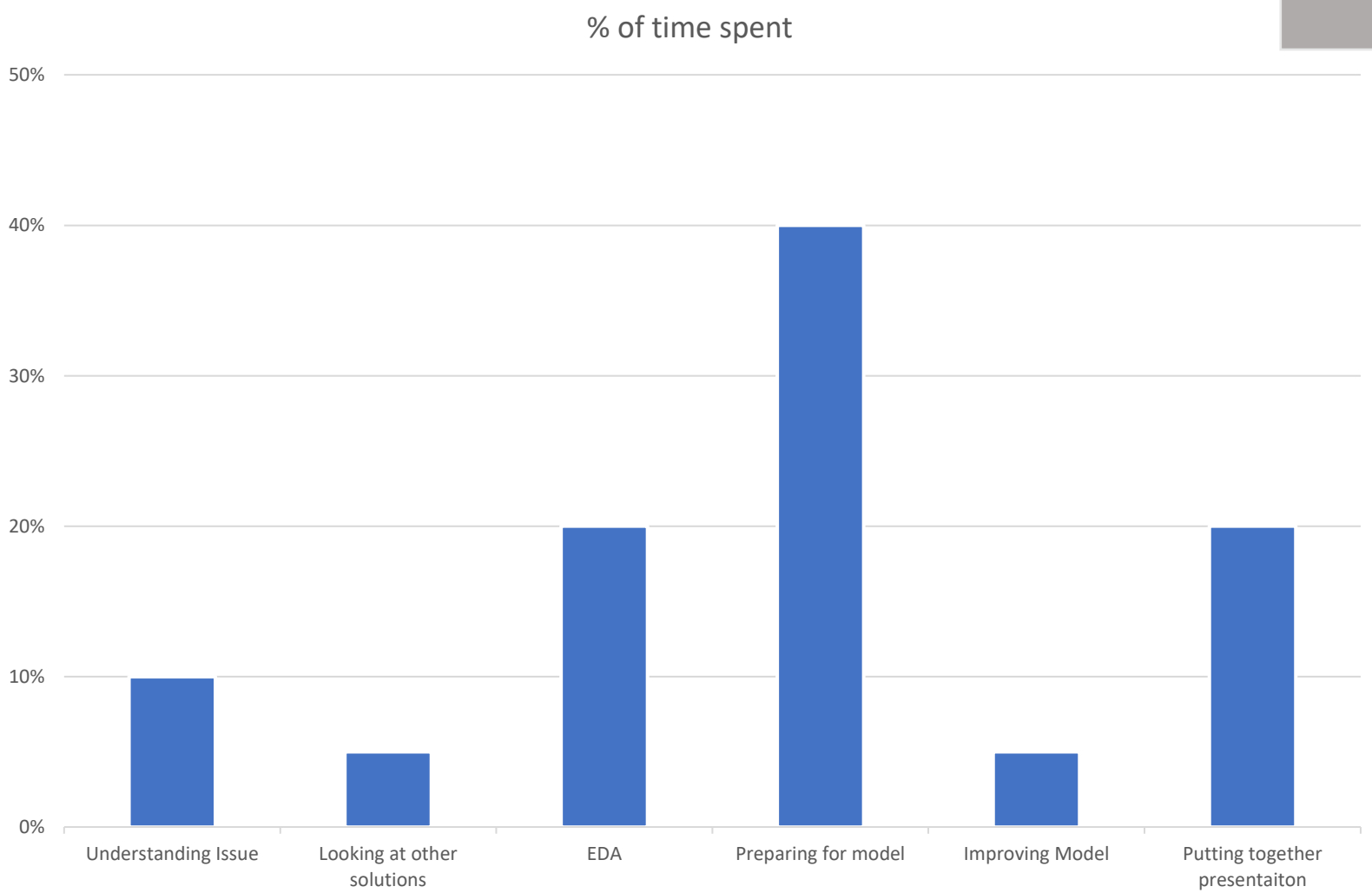
Many
'optimistic'
Guesses to show thought about the issue

What not included, to keep the presentation succinct.

- Products with most sales (banana, bag of organic bananas, organic strawberries….)
- Reorder ratios of products (results vary and already covered)
- Reorder ratios of customers (varies vary and already covered)
- Market Size of Instacart
- Standard statistical analysis (mean, standard deviations)

HOW TIME WAS SPENT

Plus time writing the SQL



Thanks