3 MILLION INSTACART ORDERS

AGUSTIN LASTRA
GENERAL ASSEMBLY DATA SCIENCE
JUNE 22 2017

01

INSTACART DATASET

instacart

Data Set:

3,421,083 Orders from Users from Instacart
206,209 Users
49,688 Products
134 Aisles
21 Departments

First Released Data thru Medium Article, later became a Kaggle Competition.

02

ORIGINAL HYPOTHESIS

Users with healthier lifestyle and habits order during certain days of the week.

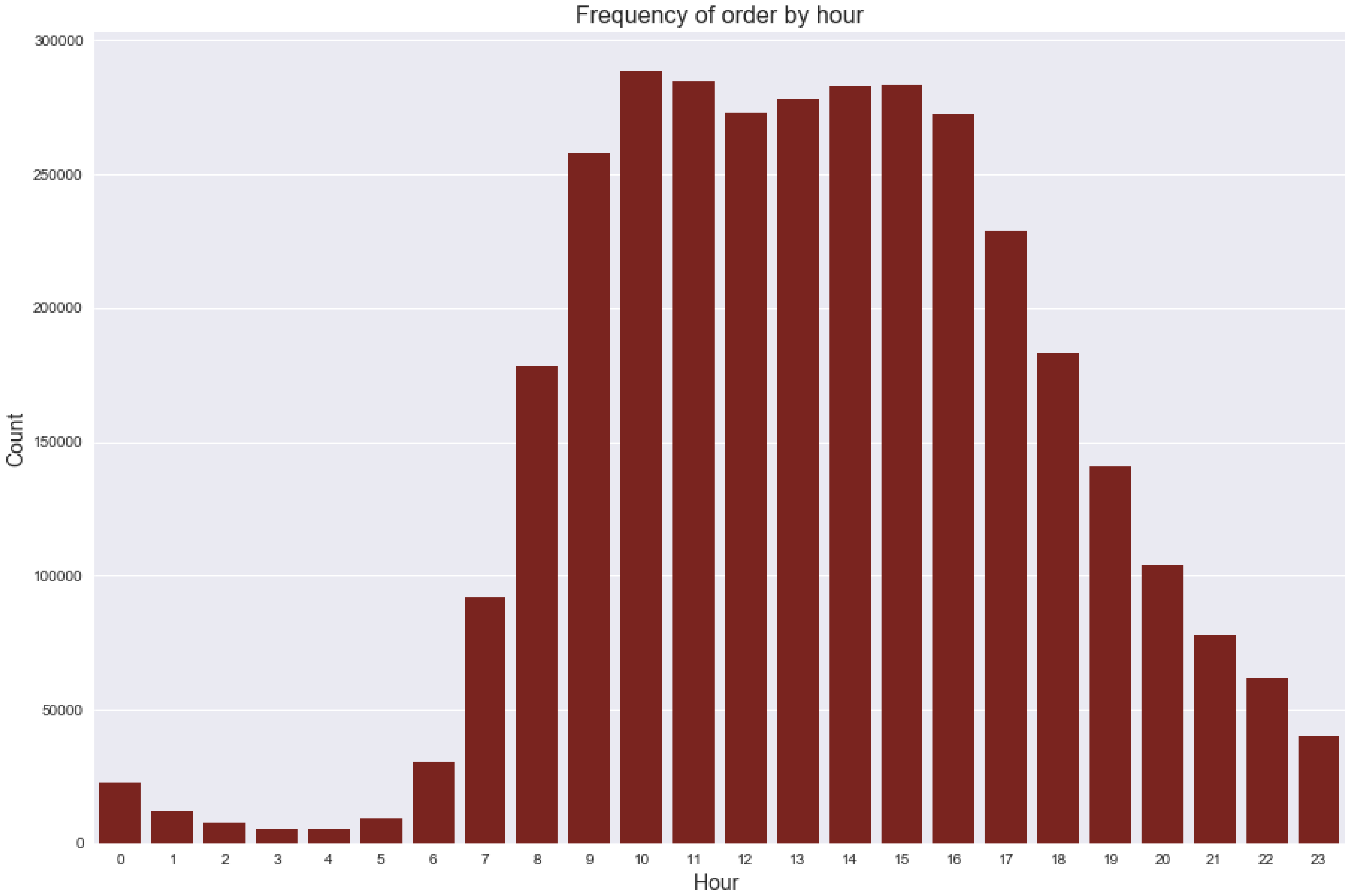
Larger orders of products are ordered during certain times of the week.

Certain products will lead to frequent re-ordering, driving sales.

03

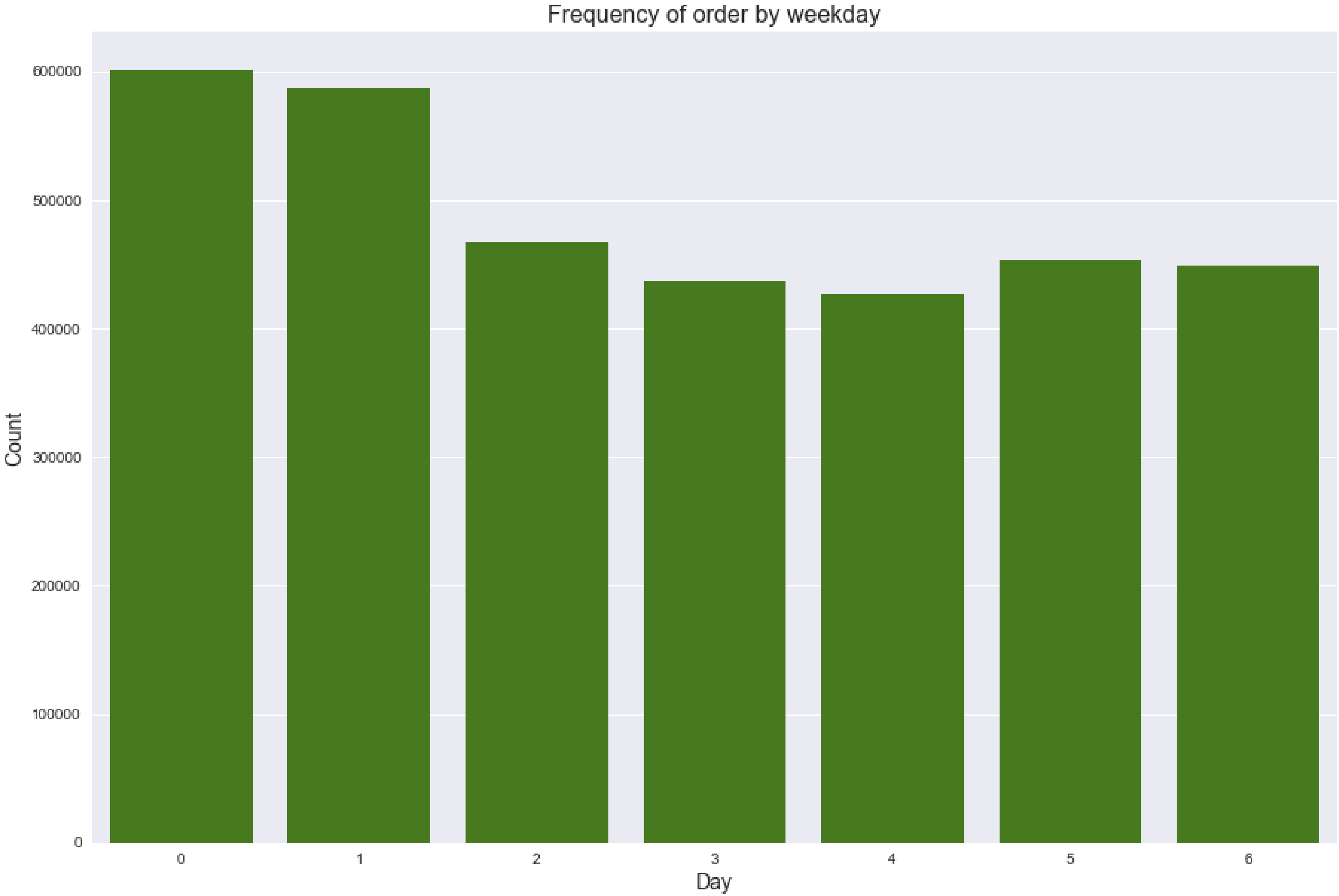
EXPLORATORY DATA ANALYSIS

What does the data look like?



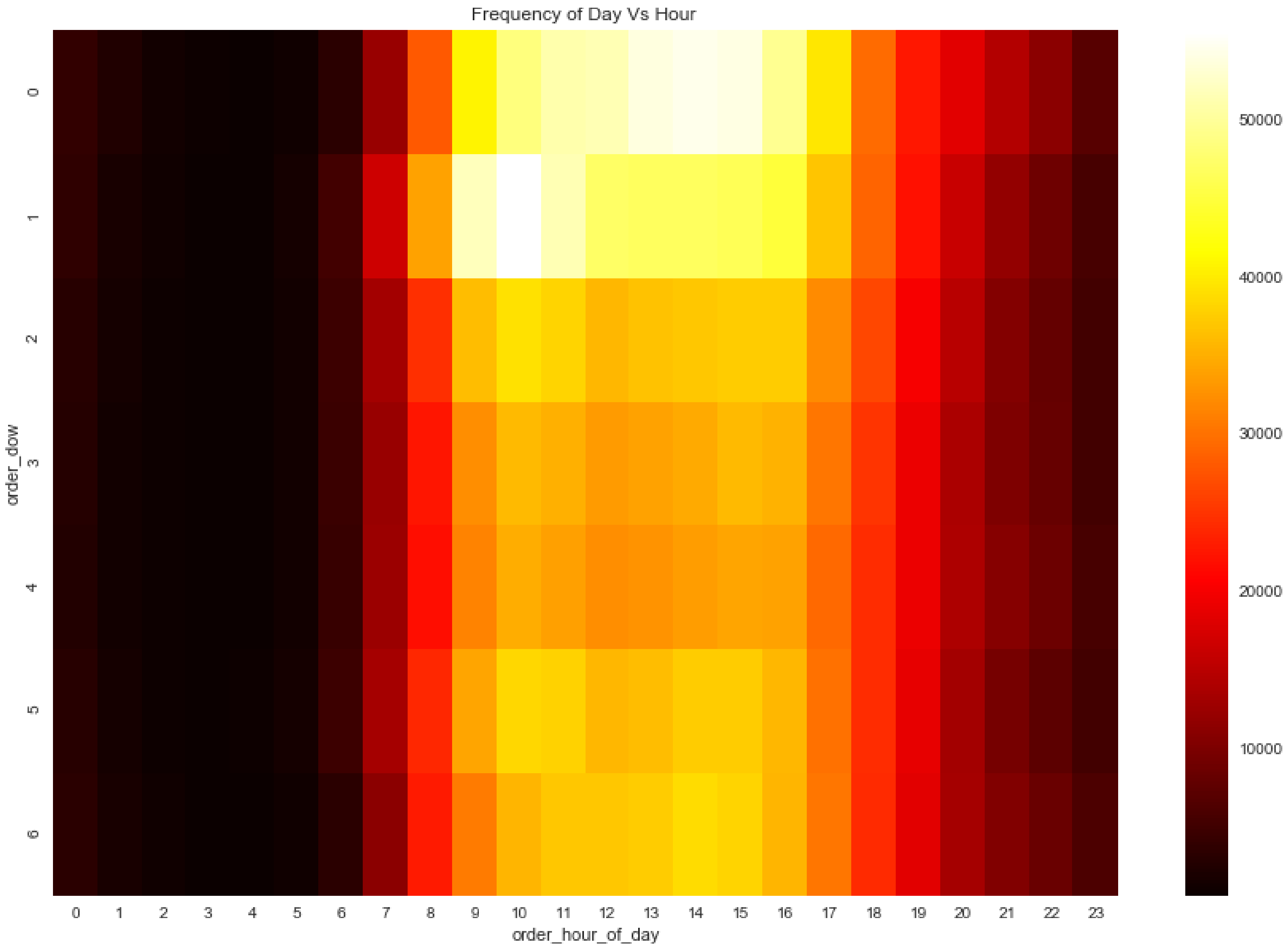
04

EXPLORATORY DATA ANALYSIS



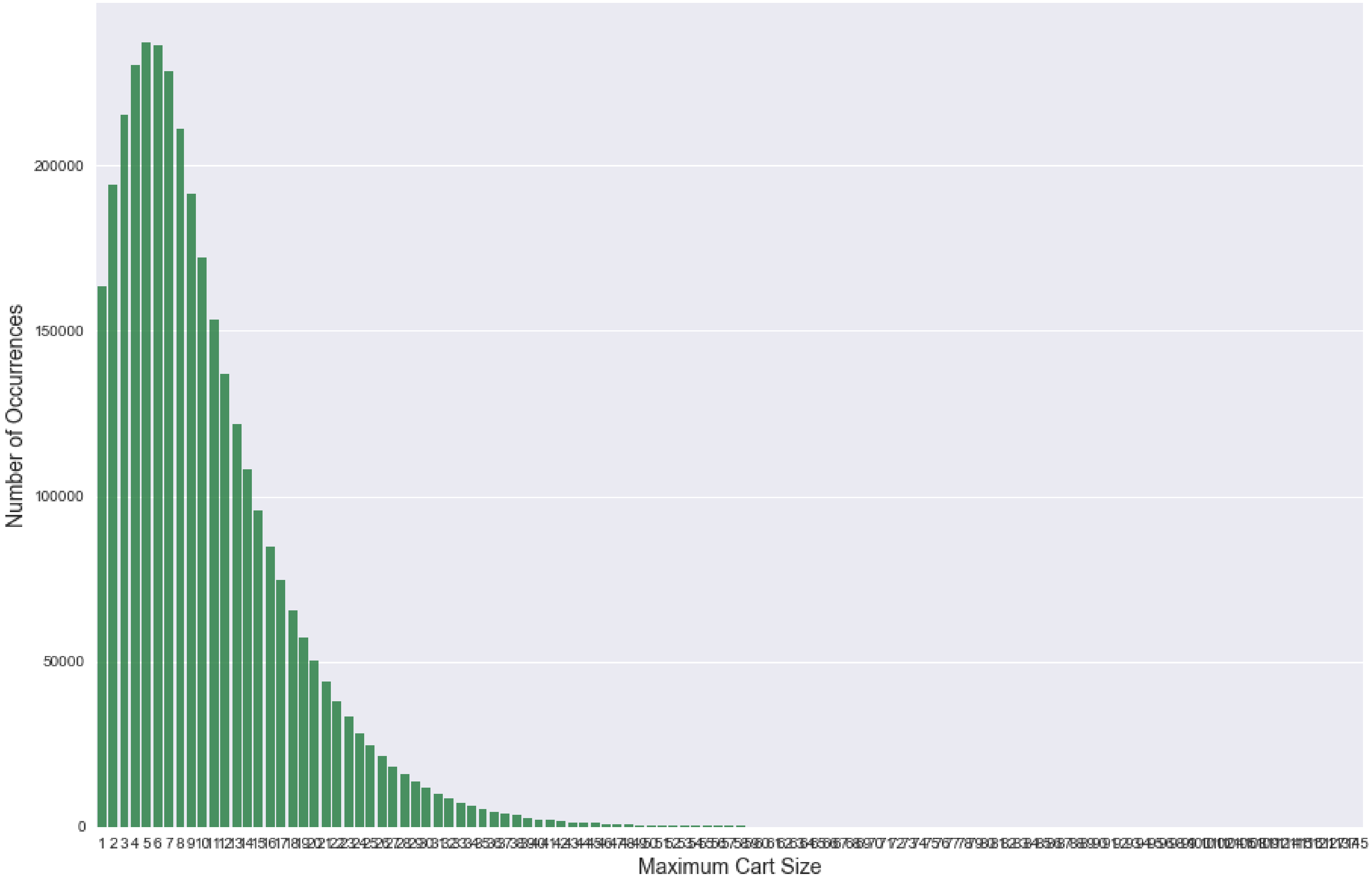
05

EXPLORATORY DATA ANALYSIS

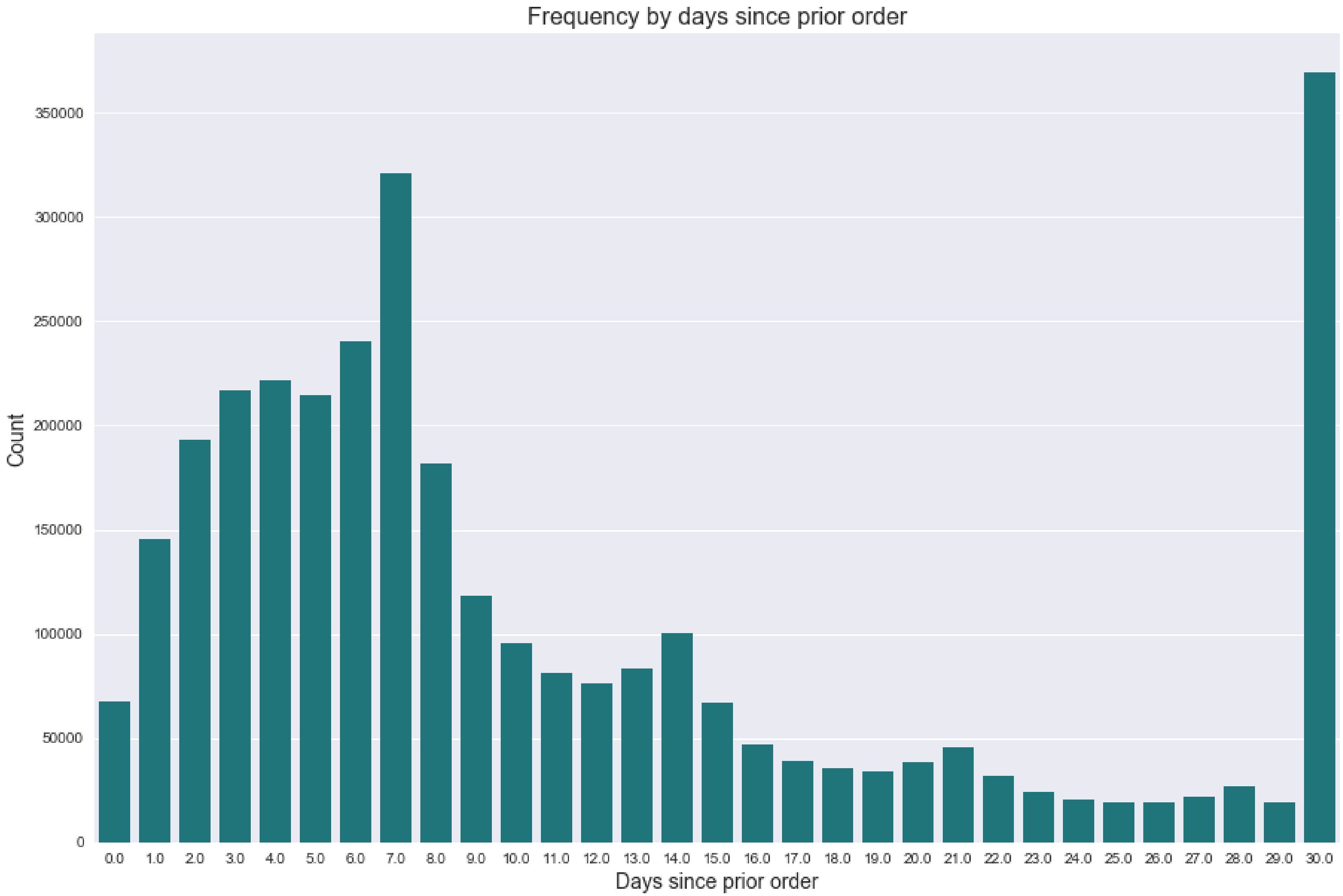


06

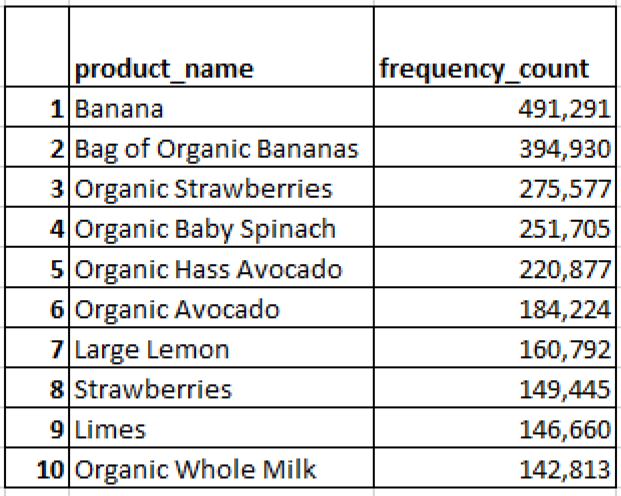
EXPLORATORY DATA ANALYSIS



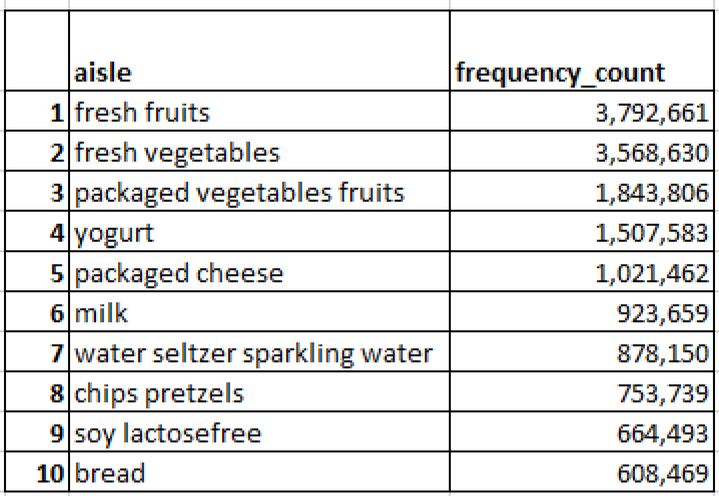
EXPLORATORY DATA ANALYSIS



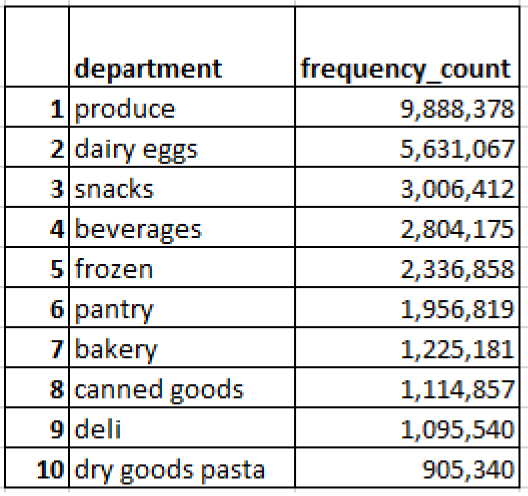
EXPLORATORY DATA ANALYSIS



	product_name	frequency_count
1	Banana	491,291
2	Bag of Organic Bananas	394,930
3	Organic Strawberries	275,577
4	Organic Baby Spinach	251,705
5	Organic Hass Avocado	220,877
6	Organic Avocado	184,224
7	Large Lemon	160,792
8	Strawberries	149,445
9	Limes	146,660
10	Organic Whole Milk	142,813

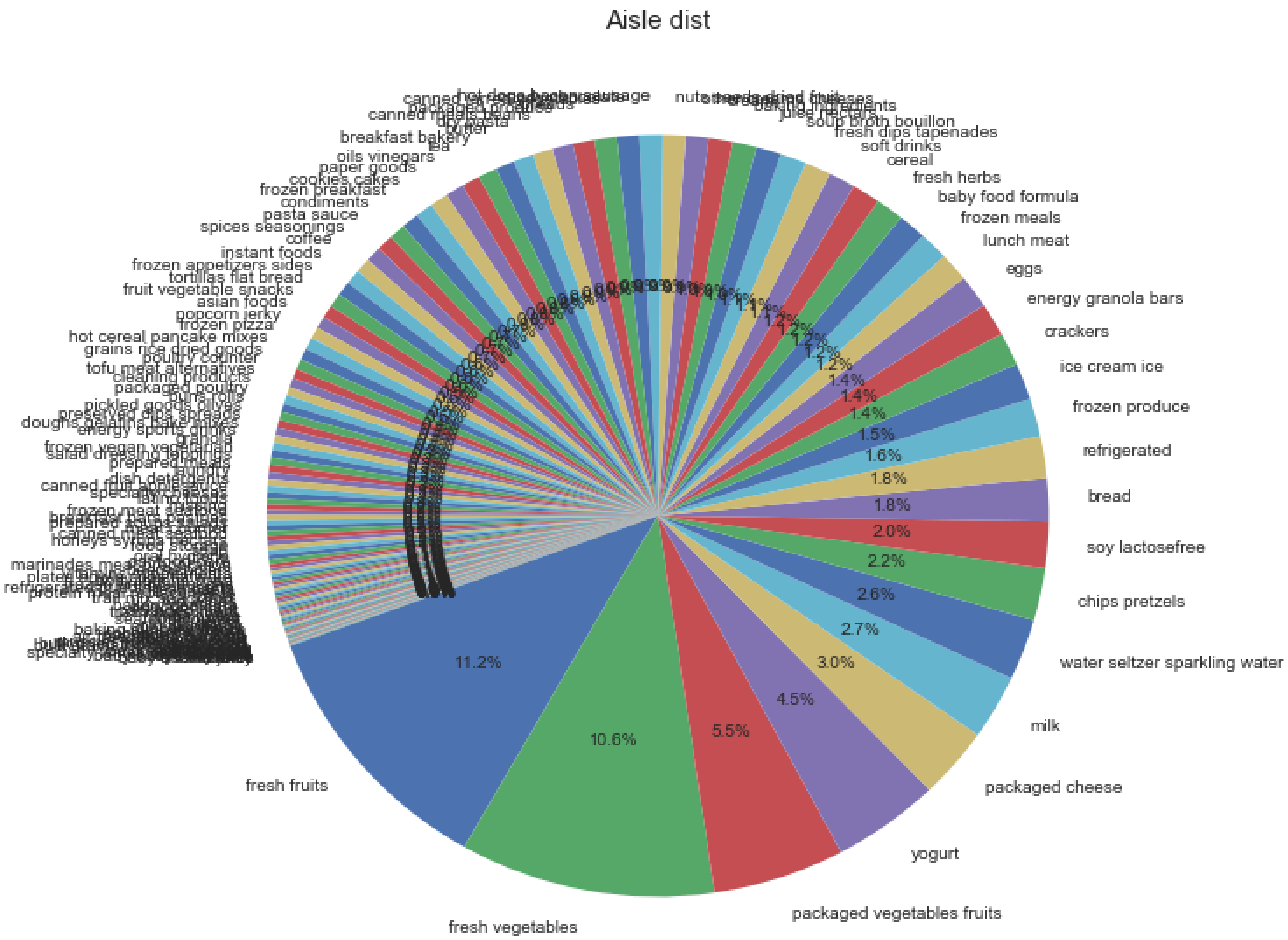


	aisle	frequency_count
1	fresh fruits	3,792,661
2	fresh vegetables	3,568,630
3	packaged vegetables fruits	1,843,806
4	yogurt	1,507,583
5	packaged cheese	1,021,462
6	milk	923,659
7	water seltzer sparkling water	878,150
8	chips pretzels	753,739
9	soy lactosefree	664,493
10	bread	608,469

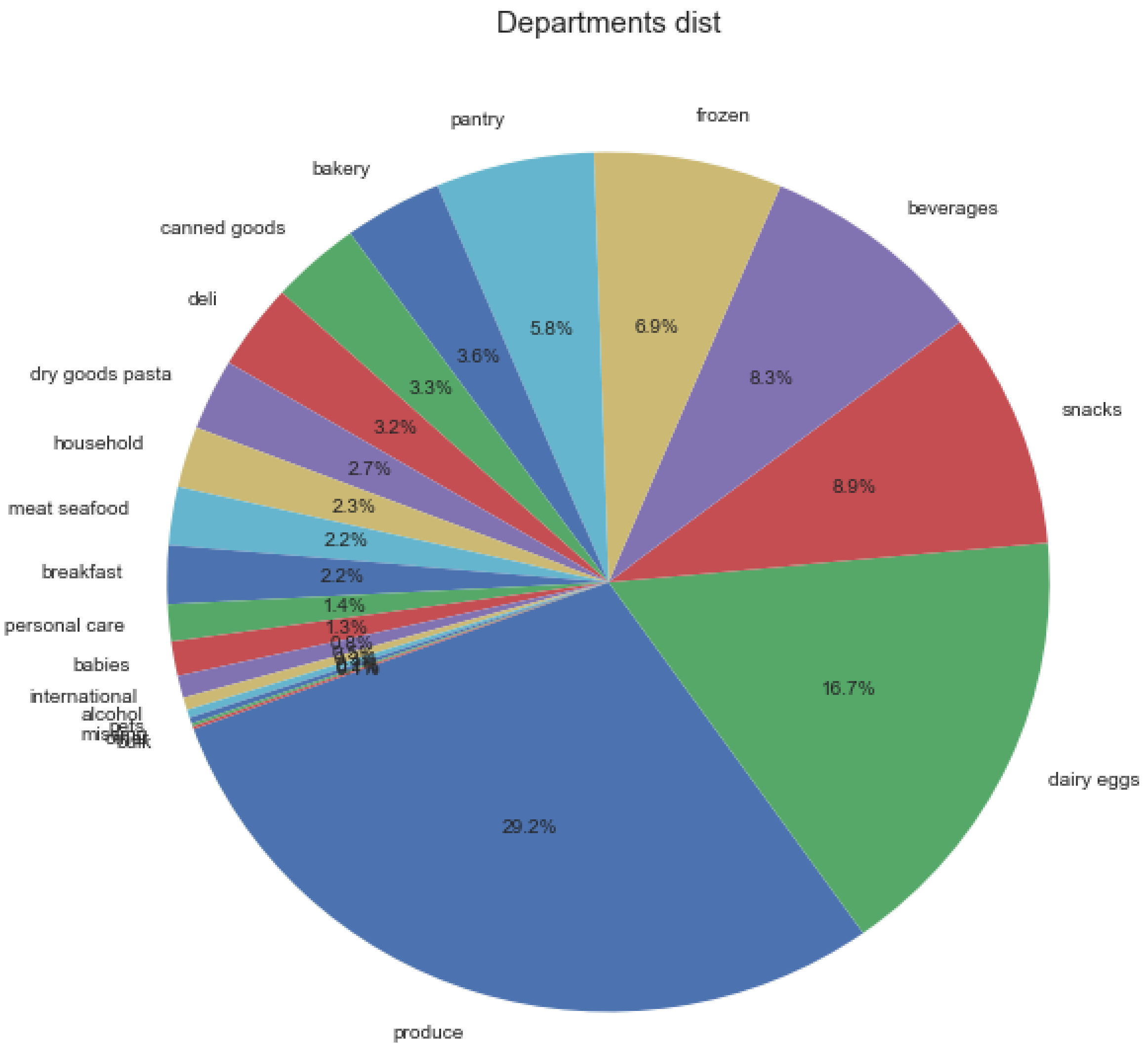


	department	frequency_count
1	produce	9,888,378
2	dairy eggs	5,631,067
3	snacks	3,006,412
4	beverages	2,804,175
5	frozen	2,336,858
6	pantry	1,956,819
7	bakery	1,225,181
8	canned goods	1,114,857
9	deli	1,095,540
10	dry goods pasta	905,340

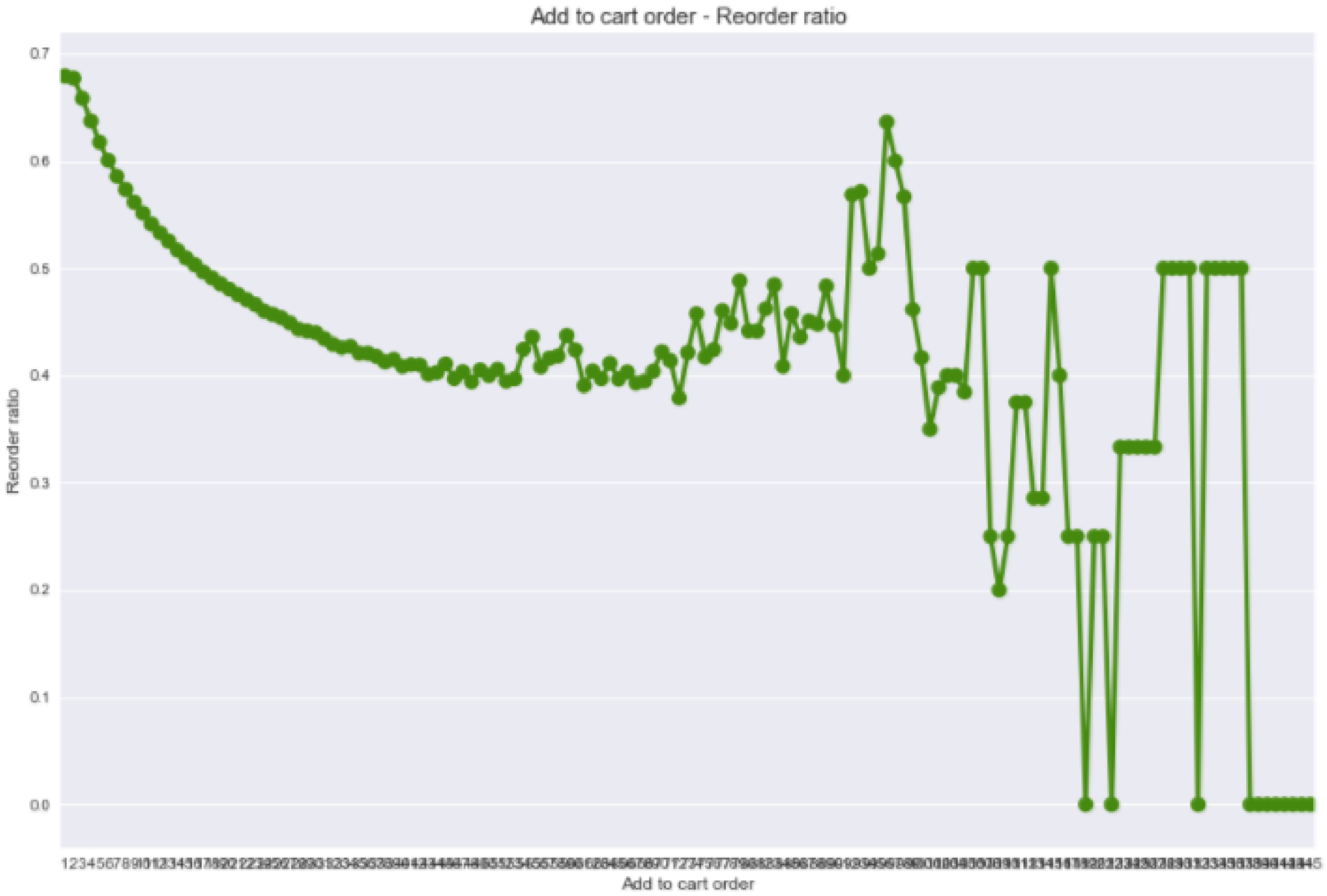
EXPLORATORY DATA ANALYSIS



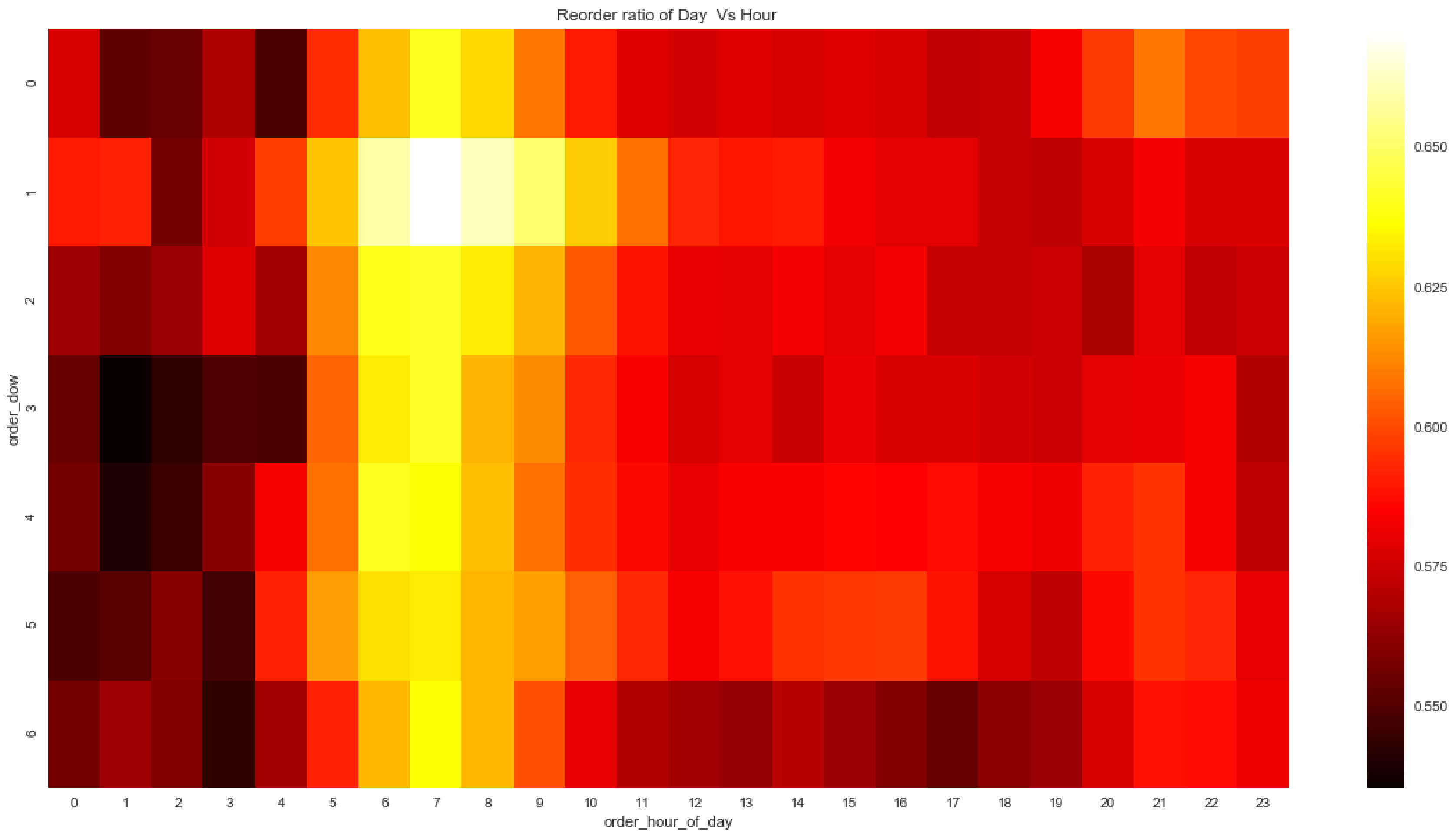
EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



ORIGINAL HYPOTHESIS'S

Users with healthier lifestyle and habits order during certain days of the week.

Needed more data to easily classify 50K Products into Healthy or not.

Larger orders of products are ordered during certain times of the week.

This didn't really require prediction or classification, it required analysis discovered in the data exploration.

Certain products will lead to frequent re-ordering, driving sales.

Similar case to the second hypothesis, the question can be answer thru analysis.

KAGGLE HYPOTHESIS

Use data on customer orders over time to predict which previously purchased products will be in a user's next order.

Possible solutions: Multiclass and multilabel algorithms, Neural Network with Softmax.

Solutions above my current skill set.

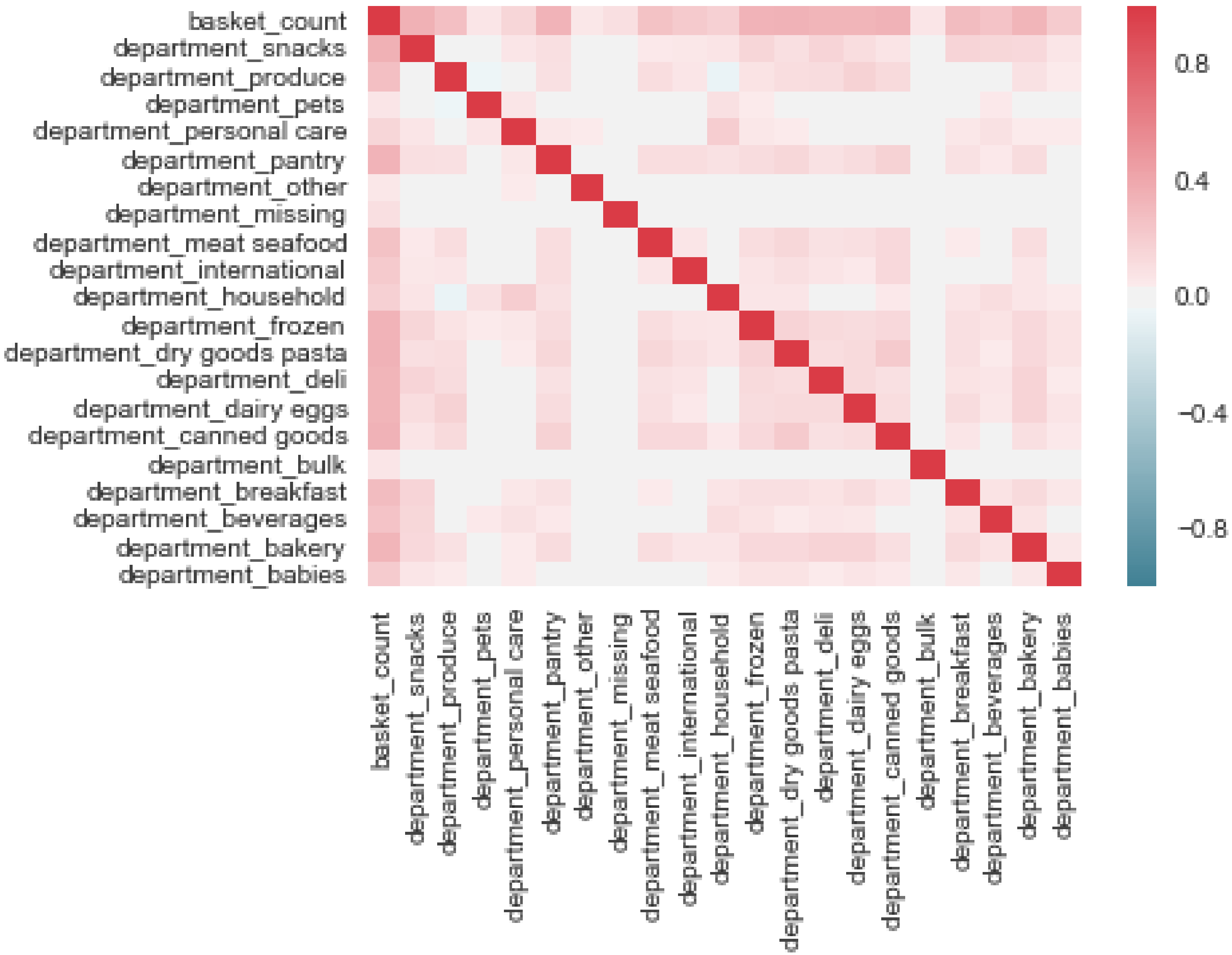
PIVOT

New Hypothesis:
Can we predict order size by which departments are present in the order?

Manipulated data to get dummies for departments,
days of the week, and hours of the day,
and also looked at order number, and days since
prior order.

MODEL: VARIABLES

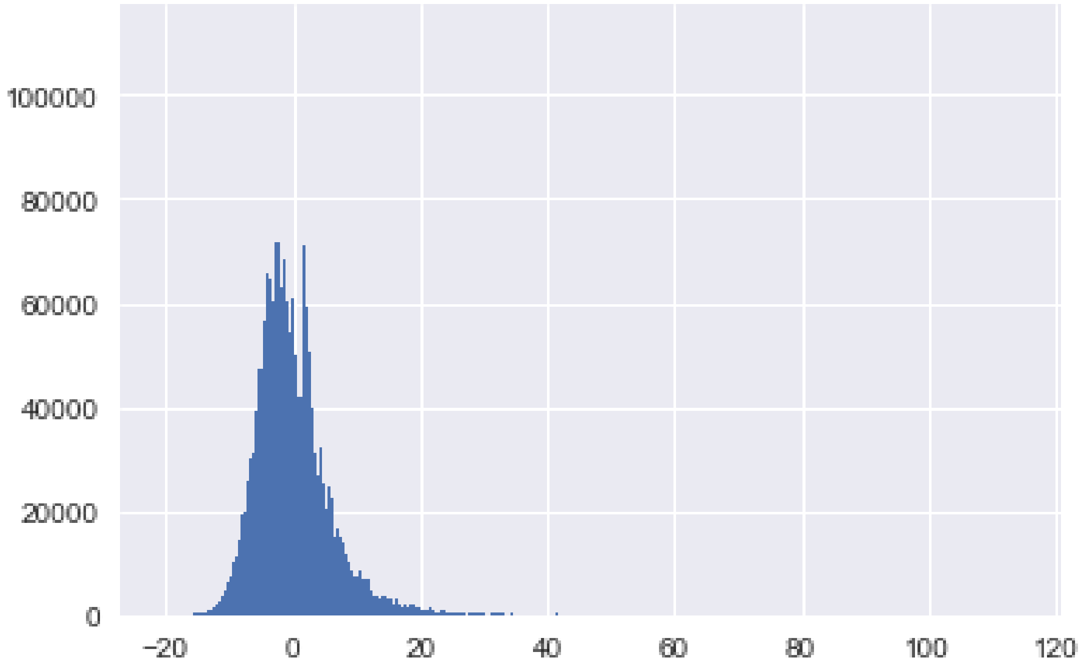
DEPARTMENTS



MODEL: VARIABLES

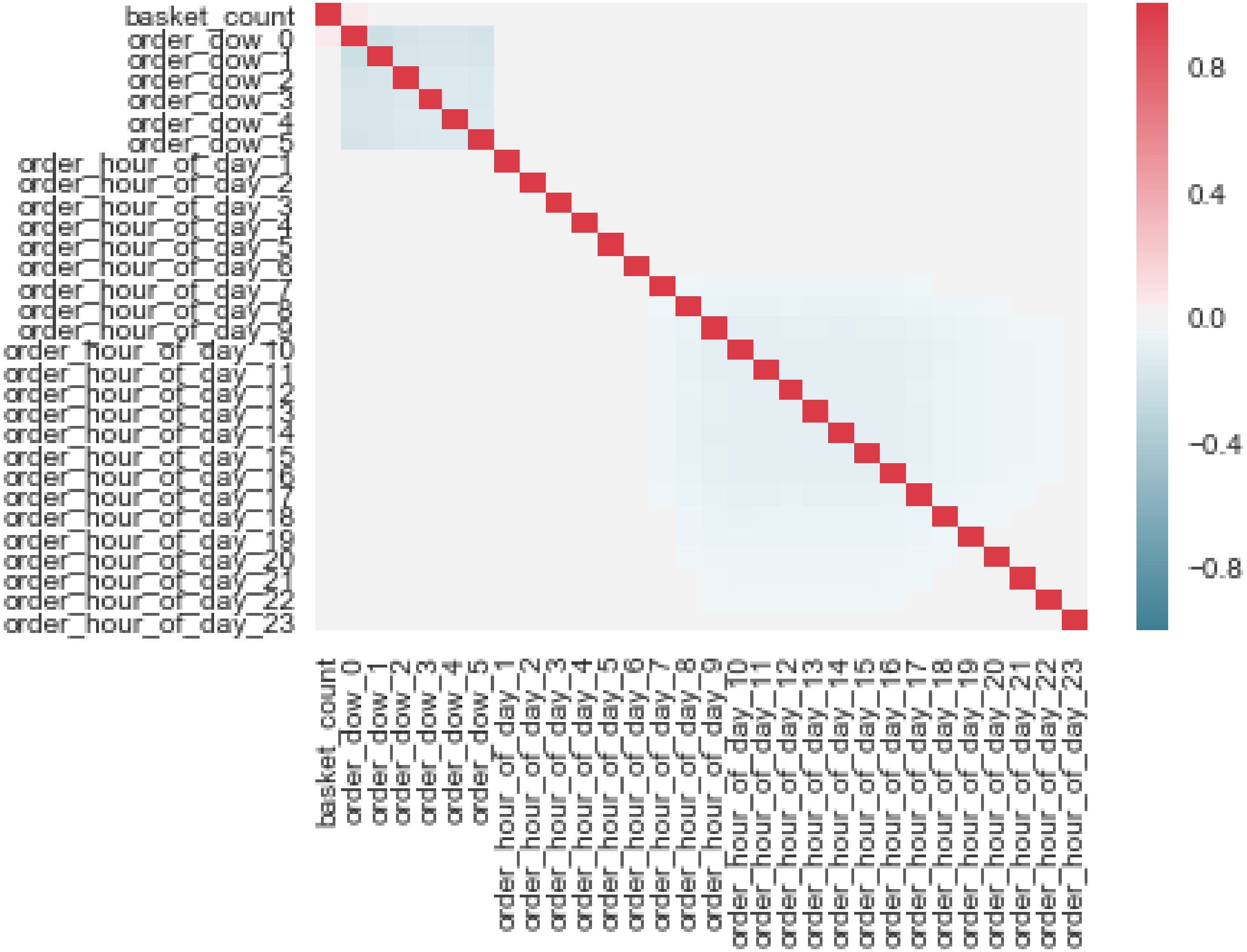
DEPARTMENTS

```
P Values: [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0. 0.  0.]
Coefficients: [ 3.17303724  4.0682438   1.47020394  1.86471505 2.84678067
 2.34308557 3.42033061  2.28185628  2.81118552  1.90402229  2.41523817
 2.61605598 2.9319955   3.33939645  2.96195038  3.03207047  2.85292723
 2.72306267 2.39198866  3.63027709]
Y-intercept: -2.92604698141
R-Squared: 0.601097163272
Residuals:
```



MODEL: VARIABLES

DAYS AND HOURS



MODEL: VARIABLES

DAYS AND HOURS

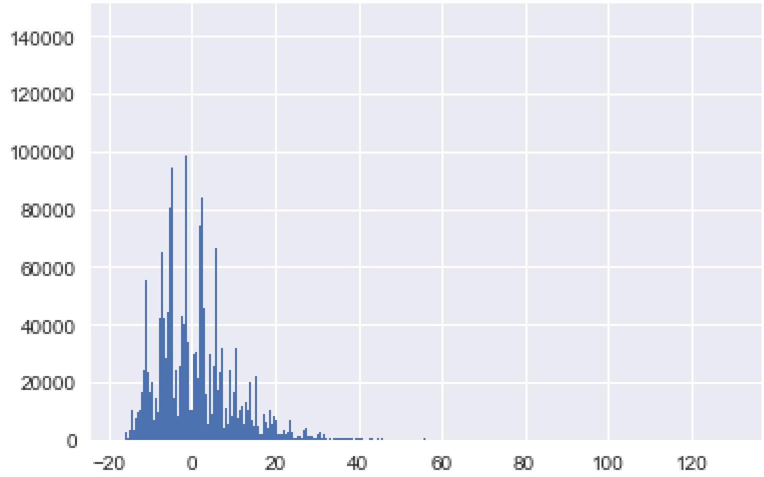
P Values: [0.00 0.00 0.00 0.000 0.00 1.01300131e-031 1.46367775e-050 8.85541979e-017 6.64647646e-025 4.87051971e-202 1.94868594e-006 1.89831559e-109 8.66467556e-234 0.00000000e+000 0.00 0.00 0.00 3.33228485e-287 6.56410039e-082 3.73085240e-034 0.00 0.00 0.00 0.00 0.00 1.55078127e-134 0.00 0.00 0.00]

Coefficients: [0.18418994 -0.45837851 -1.1376072 -1.45265595 -1.23680608 -0.642327 -0.09237608 -0.2030593 -0.04829538 0.79733612 -0.25169836 0.01043609 -0.12029103 -0.13520175 -0.1746029 -0.10250025 -0.15359368 -0.31604865 -0.4166791 -0.59403056 -0.78505558 -0.96429715 -1.26071563 -1.5879151 -1.57603259 -0.70513684 0.56704196 0.97138459 0.68269736]

y-intercept: 16.7917886052

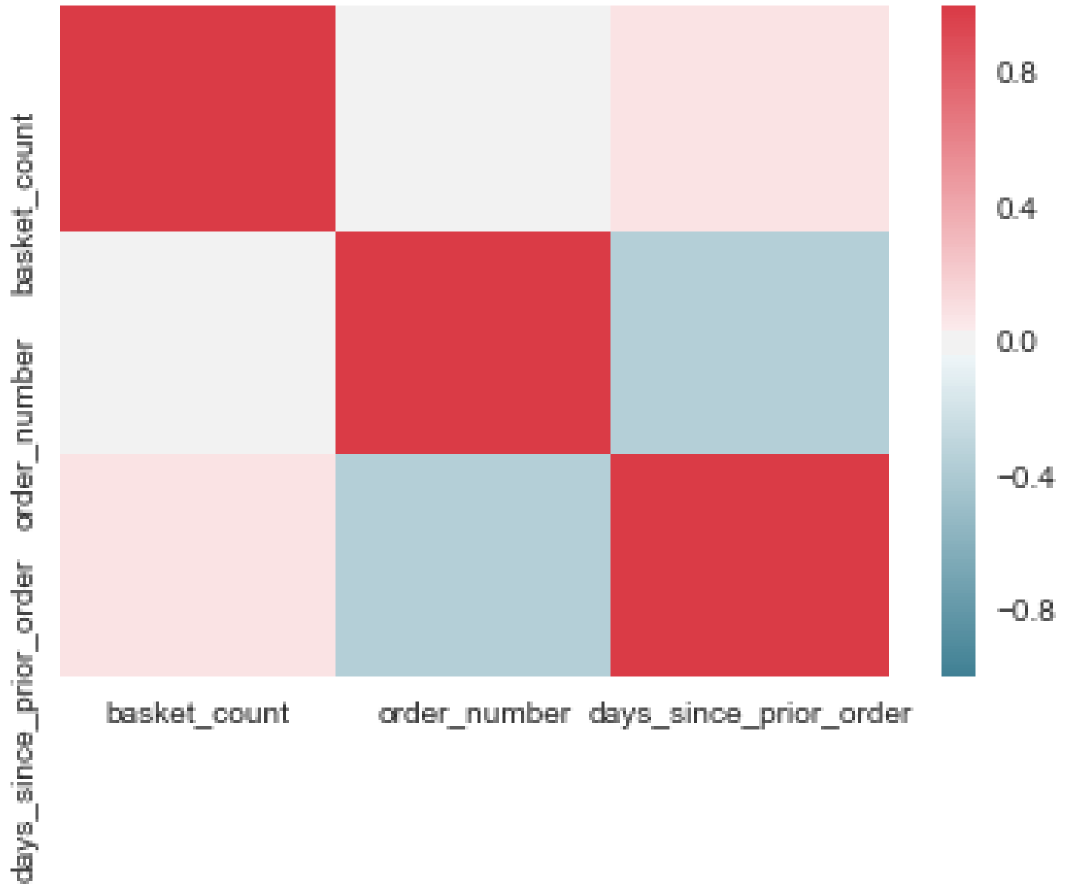
R-Squared: 0.00720666247426

Residuals:



MODEL: VARIABLES

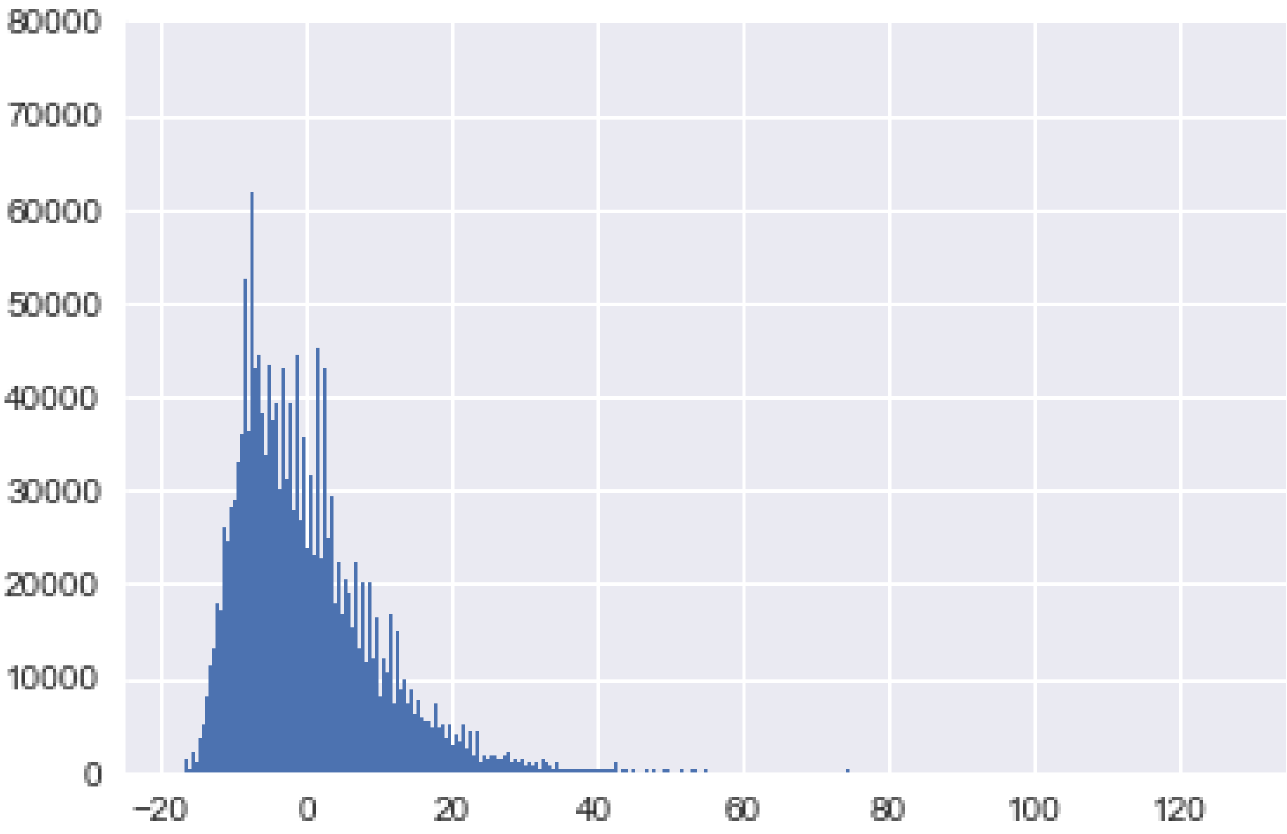
ORDER NUMBER AND DAYS SINCE PRIOR ORDER



MODEL: VARIABLES

ORDER NUMBER AND DAYS SINCE PRIOR ORDER

P Values: [0. 0.]
Coefficients: [0.01341501 0.09731443]
y-intercept: 14.3829273289
R-Squared: 0.00702686915257
Residuals:



MODEL: APPROACH

Training Set 500K
Testing Set 500K

Methods:
Ridge Regression
Cross Validation using Grid Search

Model A
Data - not normally distributed

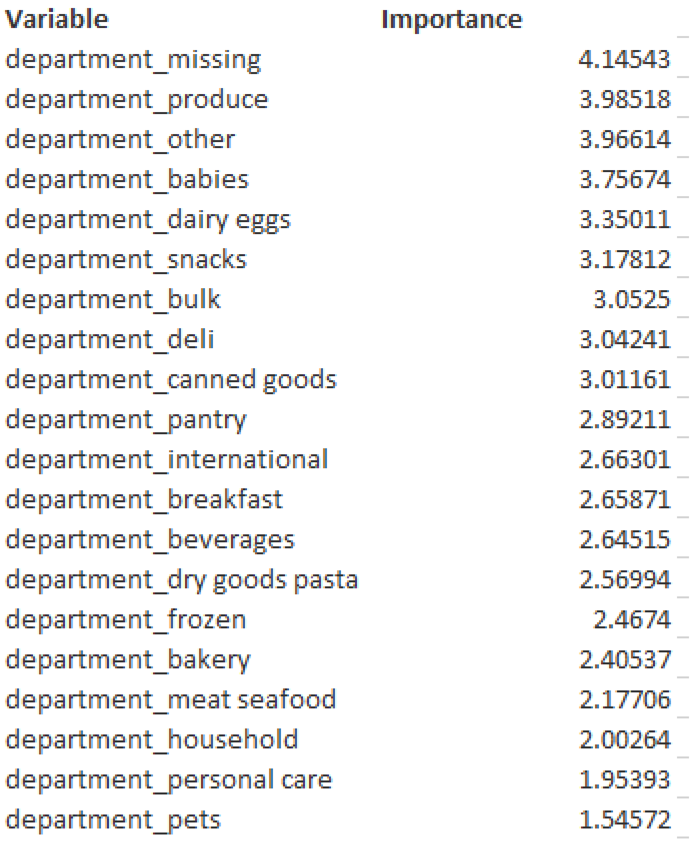
Model B
Normalized Data - Log of Basket Size

MODEL: RESULTS

DEPARTMENTS

Model A
Data - not normally distributed
Best Estimator:
Alpha 100
Best Mean Squared Error:
35.840503476209619

Mean Squared Error for Test:
39.800885883718564

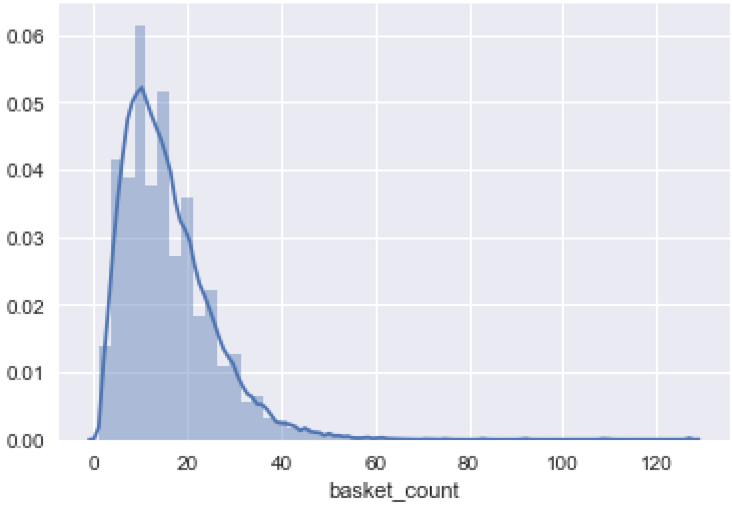


Variable	Importance
department_missing	4.14543
department_produce	3.98518
department_other	3.96614
department_babies	3.75674
department_dairy eggs	3.35011
department_snacks	3.17812
department_bulk	3.0525
department_deli	3.04241
department_canned goods	3.01161
department_pantry	2.89211
department_international	2.66301
department_breakfast	2.65871
department_beverages	2.64515
department_dry goods pasta	2.56994
department_frozen	2.4674
department_bakery	2.40537
department_meat seafood	2.17706
department_household	2.00264
department_personal care	1.95393
department_pets	1.54572

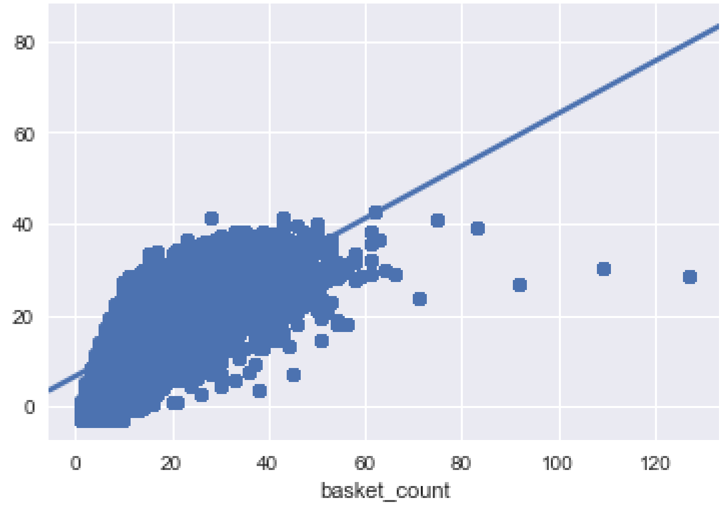
MODEL: RESULTS

DEPARTMENTS

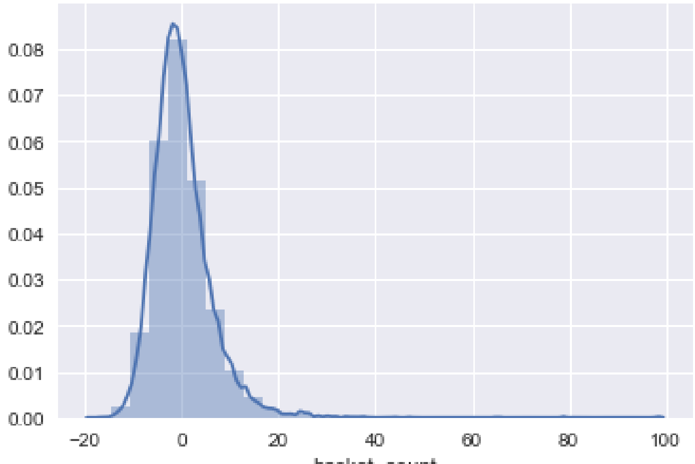
Y Test Set Distribution



Predictions



Predictions Distribution

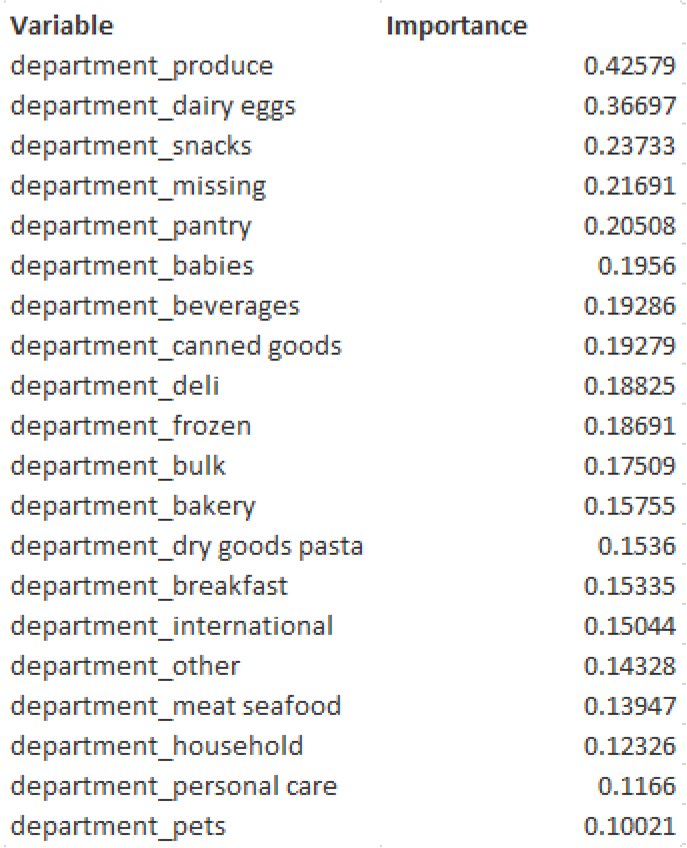


MODEL: RESULTS

DEPARTMENTS

Model B
Normalized Data - Log of Basket Size
Best Estimator:
Alpha 100
Best Mean Squared Error:
0.14410141396282597

Mean Squared Error for Test:
0.14740478309119431

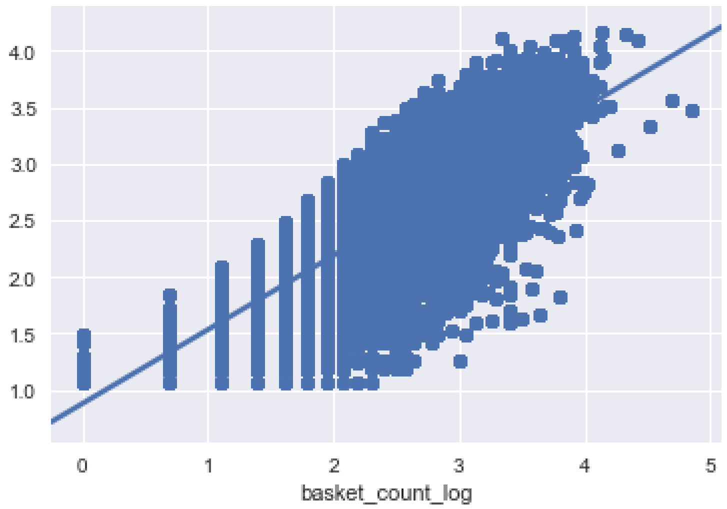


Variable	Importance
department_produce	0.42579
department_dairy eggs	0.36697
department_snacks	0.23733
department_missing	0.21691
department_pantry	0.20508
department_babies	0.1956
department_beverages	0.19286
department_canned goods	0.19279
department_deli	0.18825
department_frozen	0.18691
department_bulk	0.17509
department_bakery	0.15755
department_dry goods pasta	0.1536
department_breakfast	0.15335
department_international	0.15044
department_other	0.14328
department_meat seafood	0.13947
department_household	0.12326
department_personal care	0.1166
department_pets	0.10021

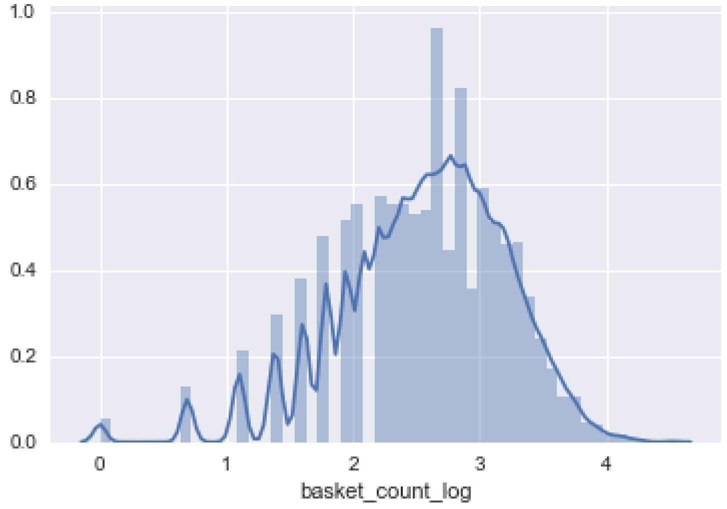
MODEL: RESULTS

DEPARTMENTS

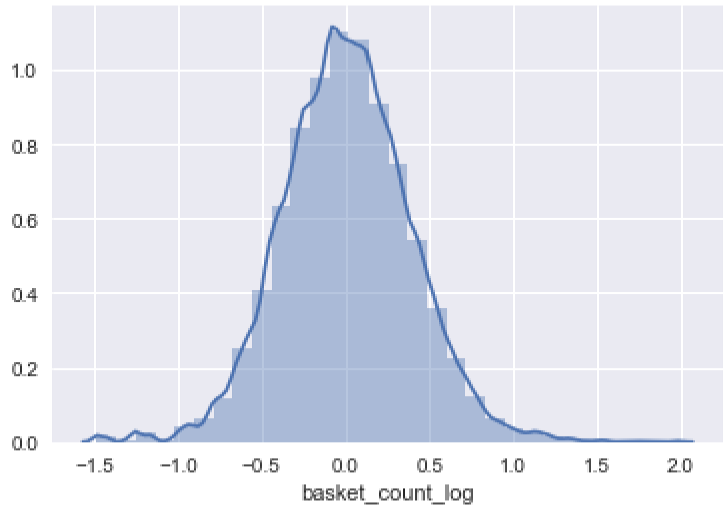
Y Test Set Distribution



Predictions

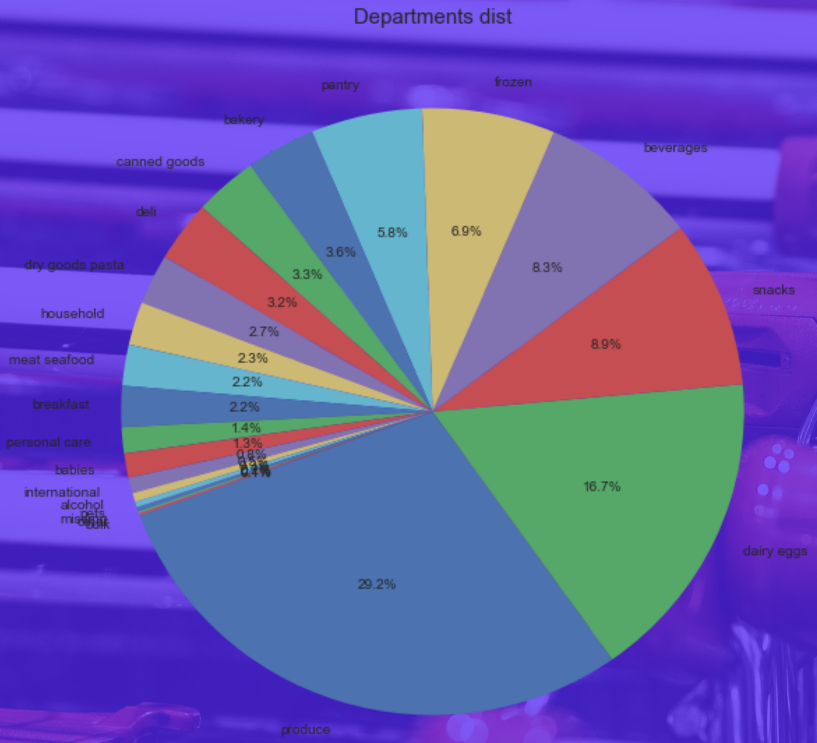


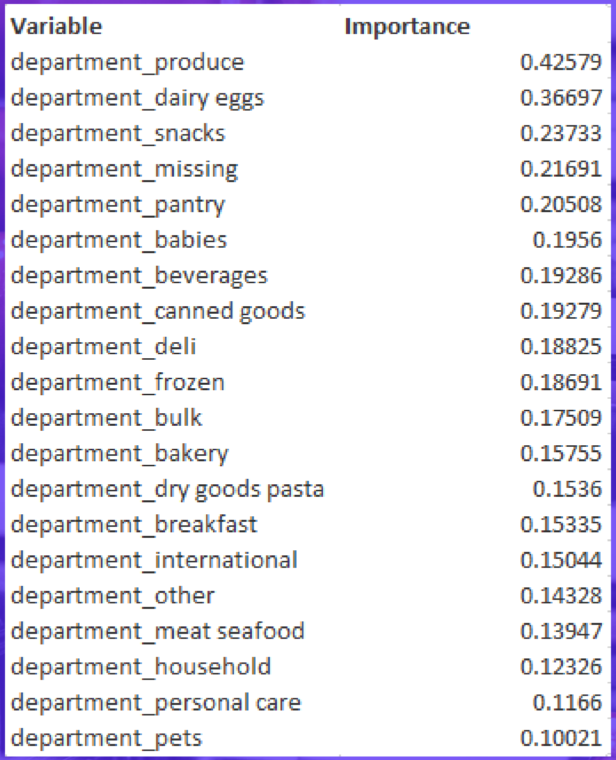
Predictions Distribution



CONCLUSION

Our second model supports the importance of the different departments and their relation to the size of an order.





Variable	Importance
department_produce	0.42579
department_dairy eggs	0.36697
department_snacks	0.23733
department_missing	0.21691
department_pantry	0.20508
department_babies	0.1956
department_beverages	0.19286
department_canned goods	0.19279
department_deli	0.18825
department_frozen	0.18691
department_bulk	0.17509
department_bakery	0.15755
department_dry goods pasta	0.1536
department_breakfast	0.15335
department_international	0.15044
department_other	0.14328
department_meat seafood	0.13947
department_household	0.12326
department_personal care	0.1166
department_pets	0.10021

Valuable Lessons:

Large Data Sets

Data size vs Computation Power

Persistence vs Flexibility

NEXT STEPS

Grow skill set in order to be able to use models that would be able to take on the Kaggle Hypothesis.

Create more complicated models that would include more variables and perhaps have greater precision.

Look for more data that would allow to draw better conclusions around the products being ordered and their healthiness.

A/B Test on the website the order and display of products according to departments using the information from our model, to see if the suggestions lead to more higher quantity orders.

THANK YOU!
THANK YOU!
THANK YOU!

Acknowledgements
Stefan Jansen
Samir Poonawala
Anthony Sorrentino
Marc Weisi

APPENDIX

https://github.com/alastra32/hw-datascience/tree/master/Final%20Project