Walmart sales forecasting

Part I: Annual sales forecasting with regression models

Qiwei Lu

Oct 2020

WANT TO BECOME WALMART'S VENDOR? THE LINE STARTS HERE…

- Everyone wants to sell through Walmart, the question is: can you keep up with the most advanced inventory management system Walmart has?
- Do you know how much inventory is enough for next year?

DATA OVERVIEW: SOURCE

- Source: M5 Walmart sales forecasting competition on Kaggle

DATA SET OVERVIEW: DETAILS

- 3 files including 30490 records for 3049 unique products in 10 different stores

For 3 categories(FOODS, HOBBIES, HOUSEHOLD),

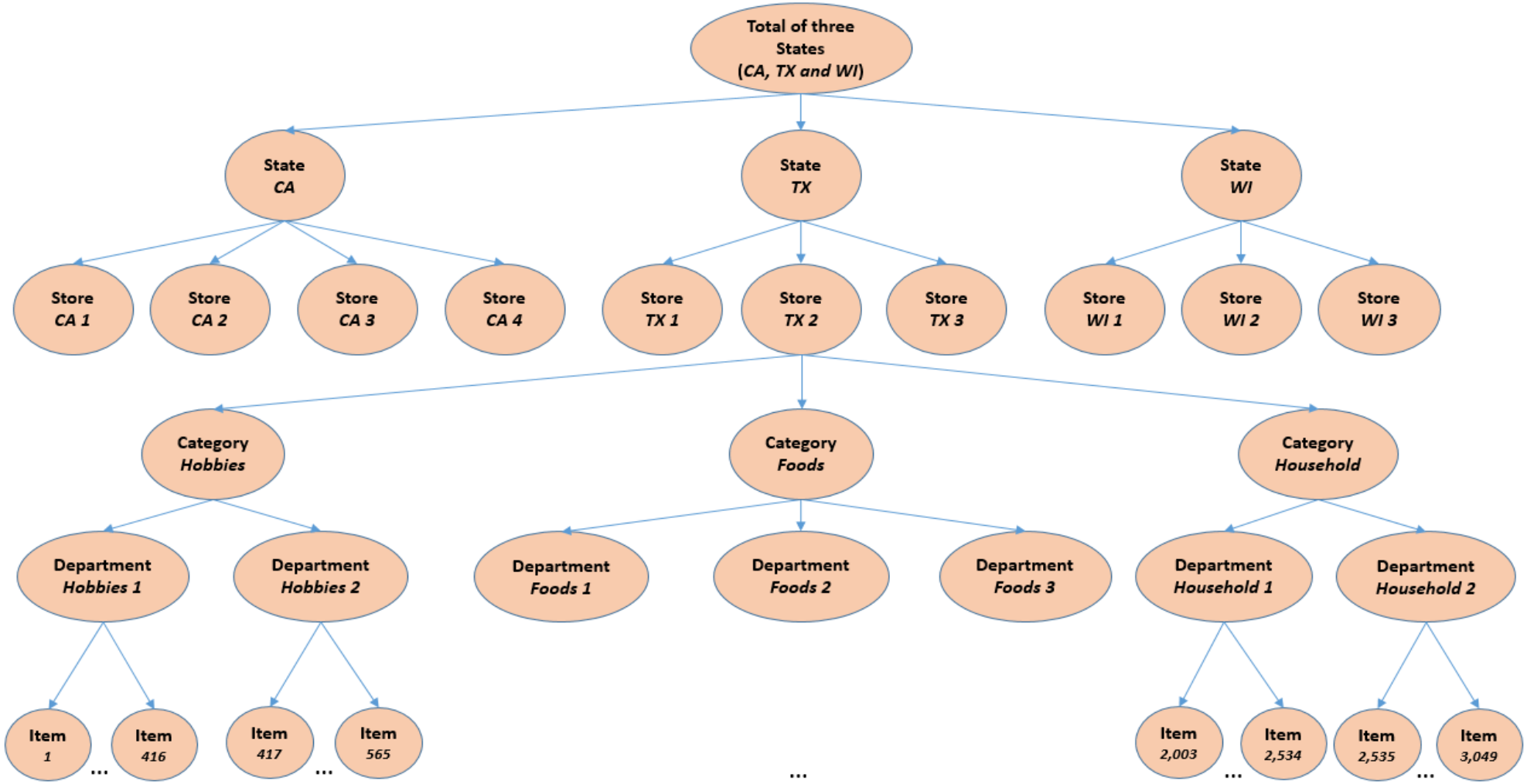
In 7 departments(FOODS1-4, HOBBIES1-3, HOUSEHOLD 1-3

Daily sale number and weekly price are given

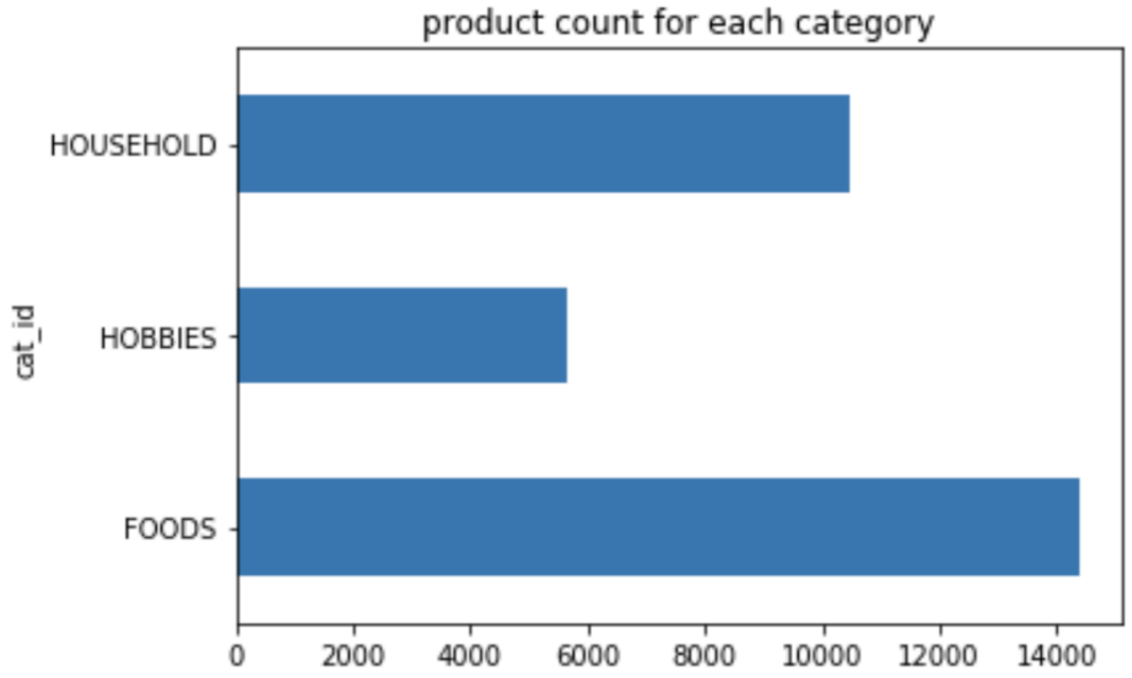
Across 3 states: CA, TX, WI

Historical data range from **2011-01-29** to **2016-06-19(5.5 years)**

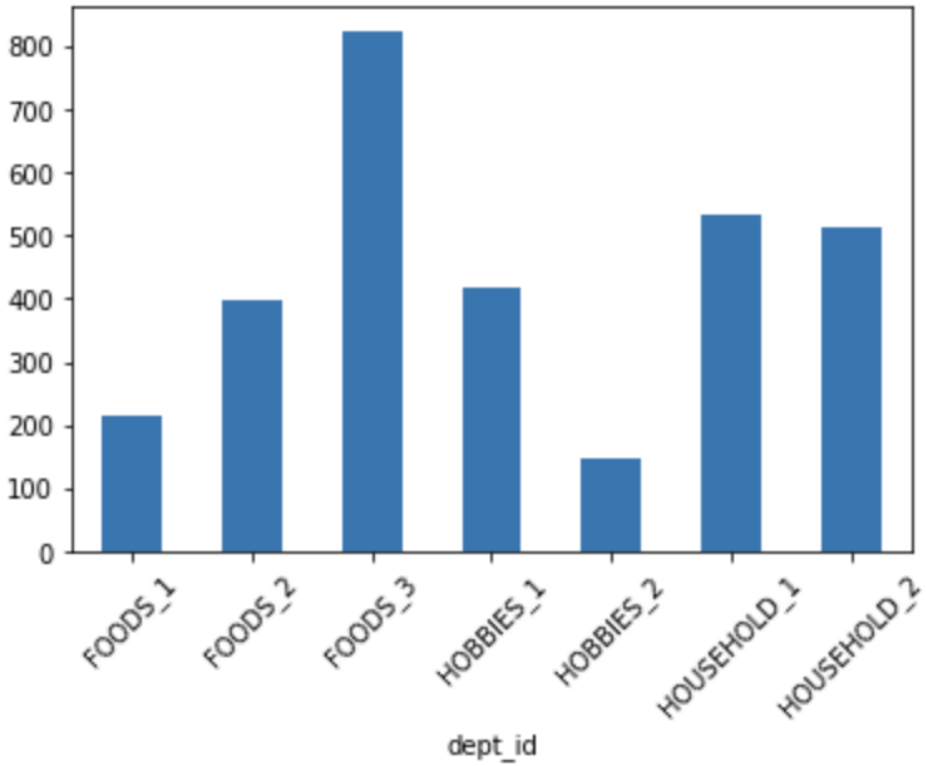
DATA SET OVERVIEW: STRUCTURE



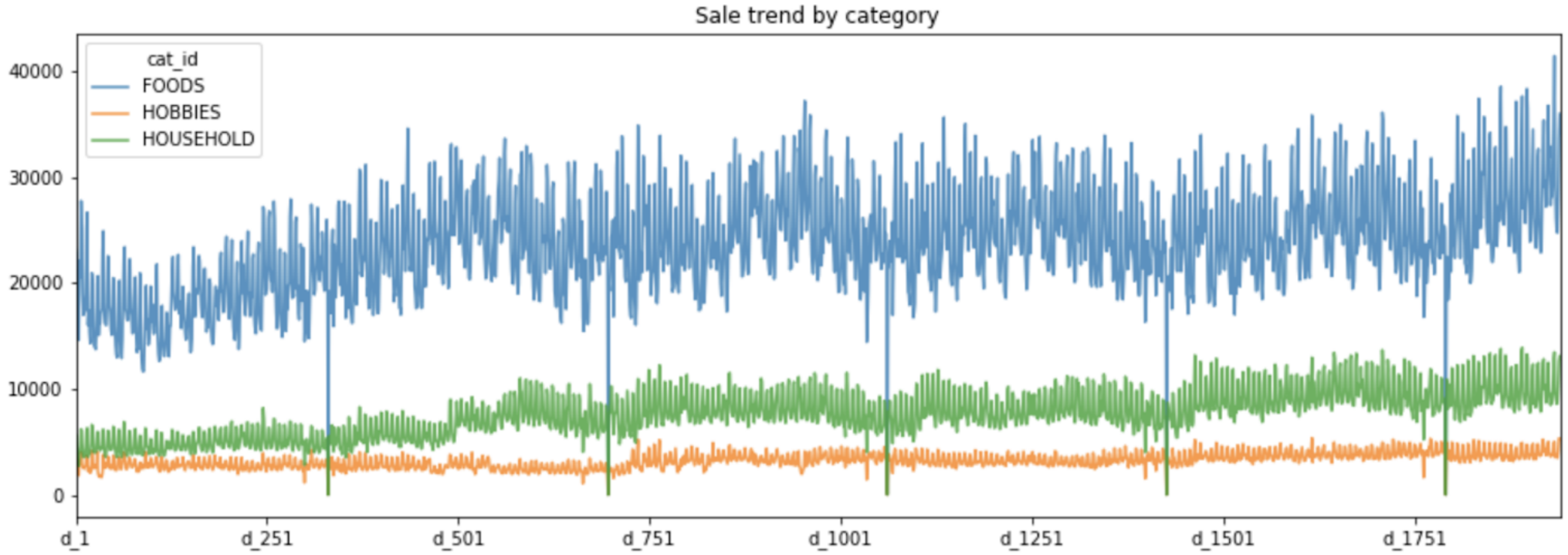
DATA WRANGLING AND VISUALIZATION – PRODUCT COUNTS



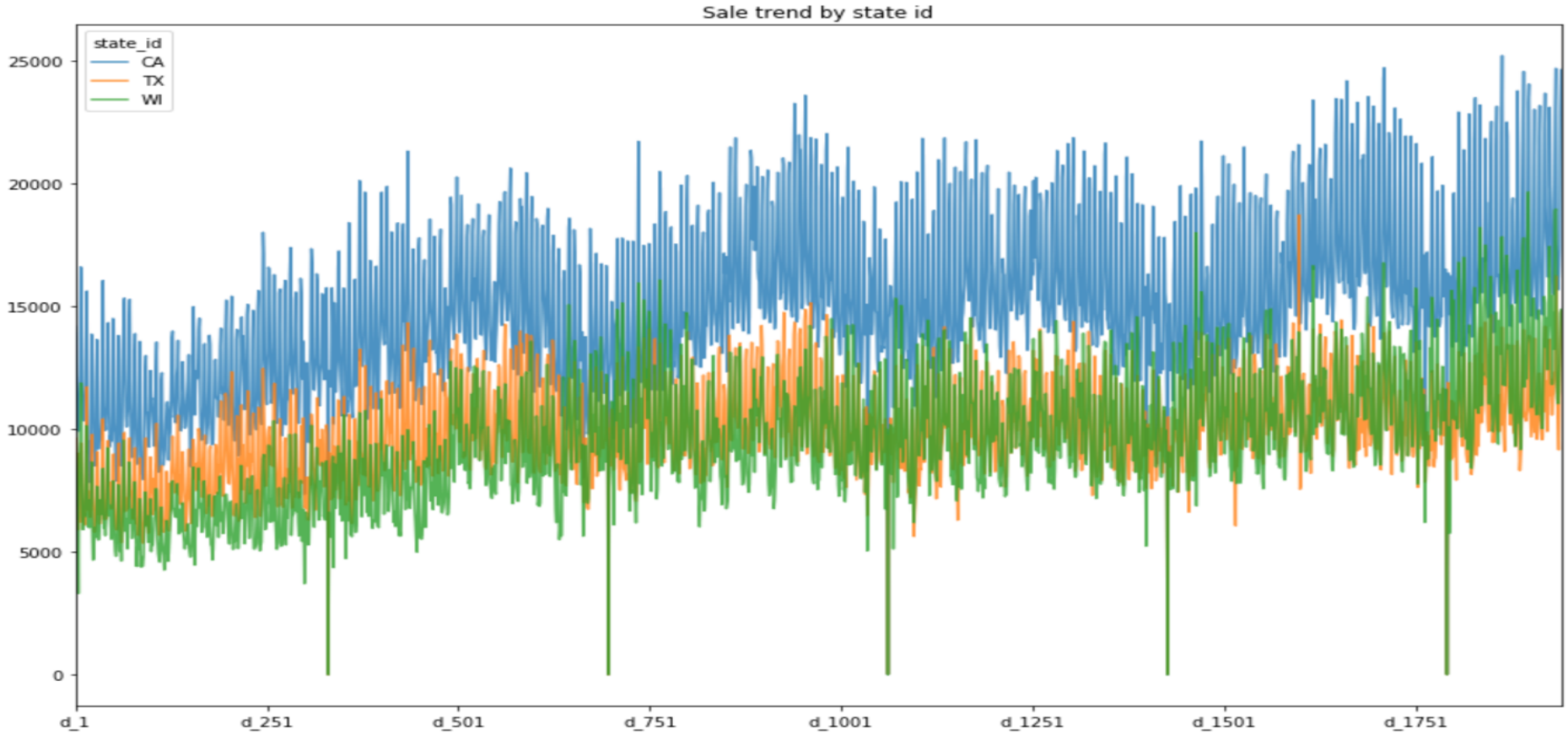
DATA WRANGLING AND VISUALIZATION – PRODUCT COUNTS



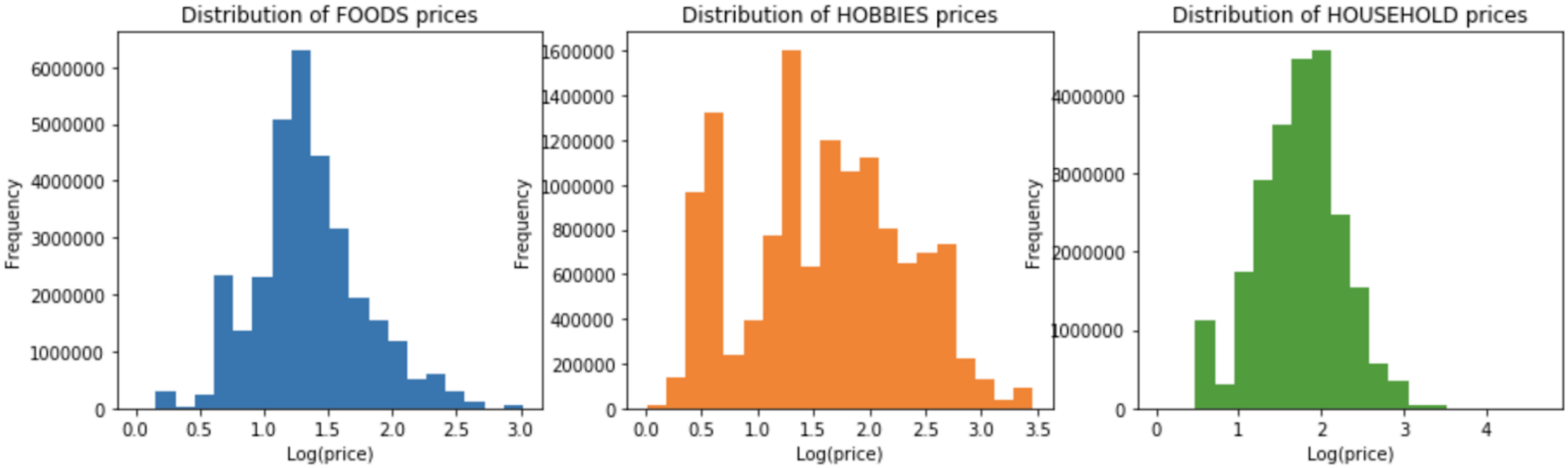
DATA WRANGLING AND VISUALIZATION – SALES TREND



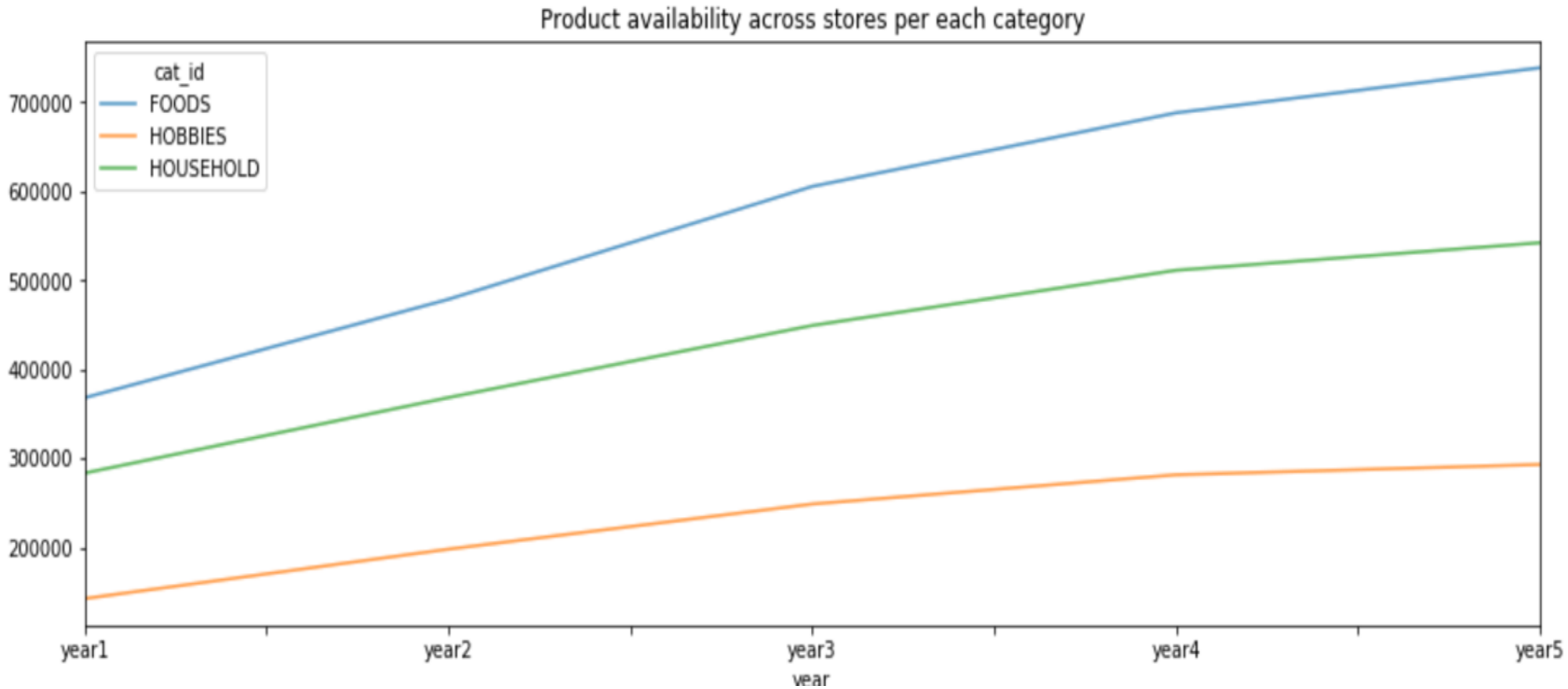
DATA WRANGLING AND VISUALIZATION – SALES TREND



DATA WRANGLING AND VISUALIZATION – PRICE DISTRIBUTION



DATA WRANGLING AND VISUALIZATION – PRODUCT AVAILABILITY



ASSUMPTIONS

- 1. About SNAP: we assume the product is eligible to be purchased by SNAP as long as it is FOODS
- 2. About sale of '0': we assume restock is always on time. Customers can have access as long as there is inventory
- 3. About when the price is set: We will assume that the prices are set on the previous year end and the suppliers know if their products will be sold at each store by that time as well.

FEATURE AGGREGATION

- current year' min, max and average price and previous year's min, avg, max price
- previous year's total selling days for each store for the year (which is certain and happened already), and if in current years, the product will be continue (if the price count for current year is over 0 then yes, equals 0 then no)
- previous year's sale by store_id and current year's sales number as target.
- department id and category id
- if the item can be purchased by SNAP (we assume all the food item is eligible)

MODEL SELECTION AND TRAINING

- Linear Regression
- Ridge Linear Regression
- Lasso Regression
- Support Vector Linear Regression & Support Vector Polynomial Regression
- Decision Tree Regression
- Random Forest Regression
- Adaboost regression
- Neural networks
- **Metrics for performance evaluation**: Root-mean-square error (RMSE) & R^2 (coefficient of determination) score as reference for linear models

CROSS VALIDATION RESULT– DROP THE WEAK MODELS

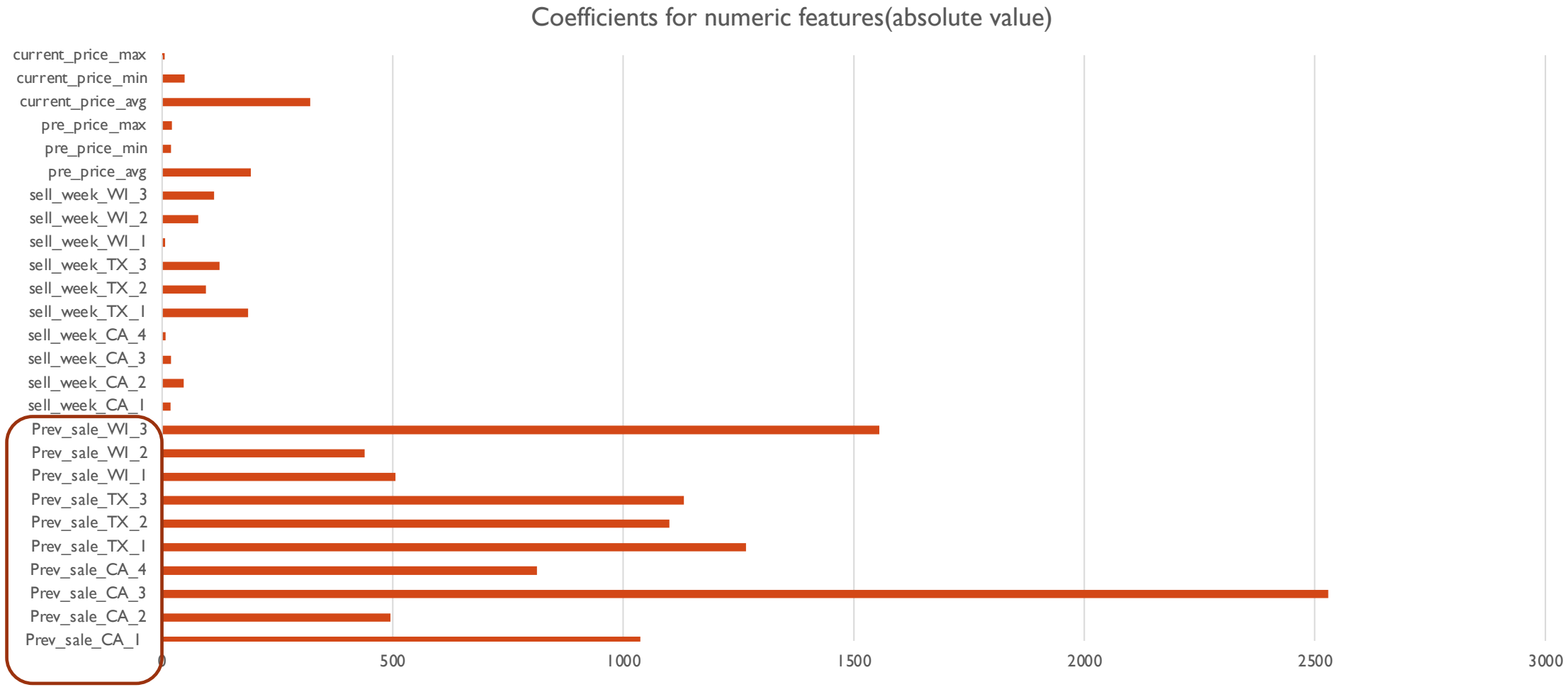
	Linear	Ridge	Lasso	SVM Linear	SVM Polynomial 👎	Decision Tree 👎	Random Forest	AdaBoost 👎	Neural Networks
rmse_ave	1945.56	1945.42	1945.25	2774.45	6661.81	2695.86	2014.28	4794.52	2079.84
rmse_std	1730.40	1730.17	1730.81	1497.19	941.20	1651.74	1864.87	1304.00	1669.69
R2	0.94	0.94	0.94	0.91	0.58	0.91	0.93	0.77	0.93

Benchmark: we assume the target is the same as previous year's sales, the rmse is 2545.92

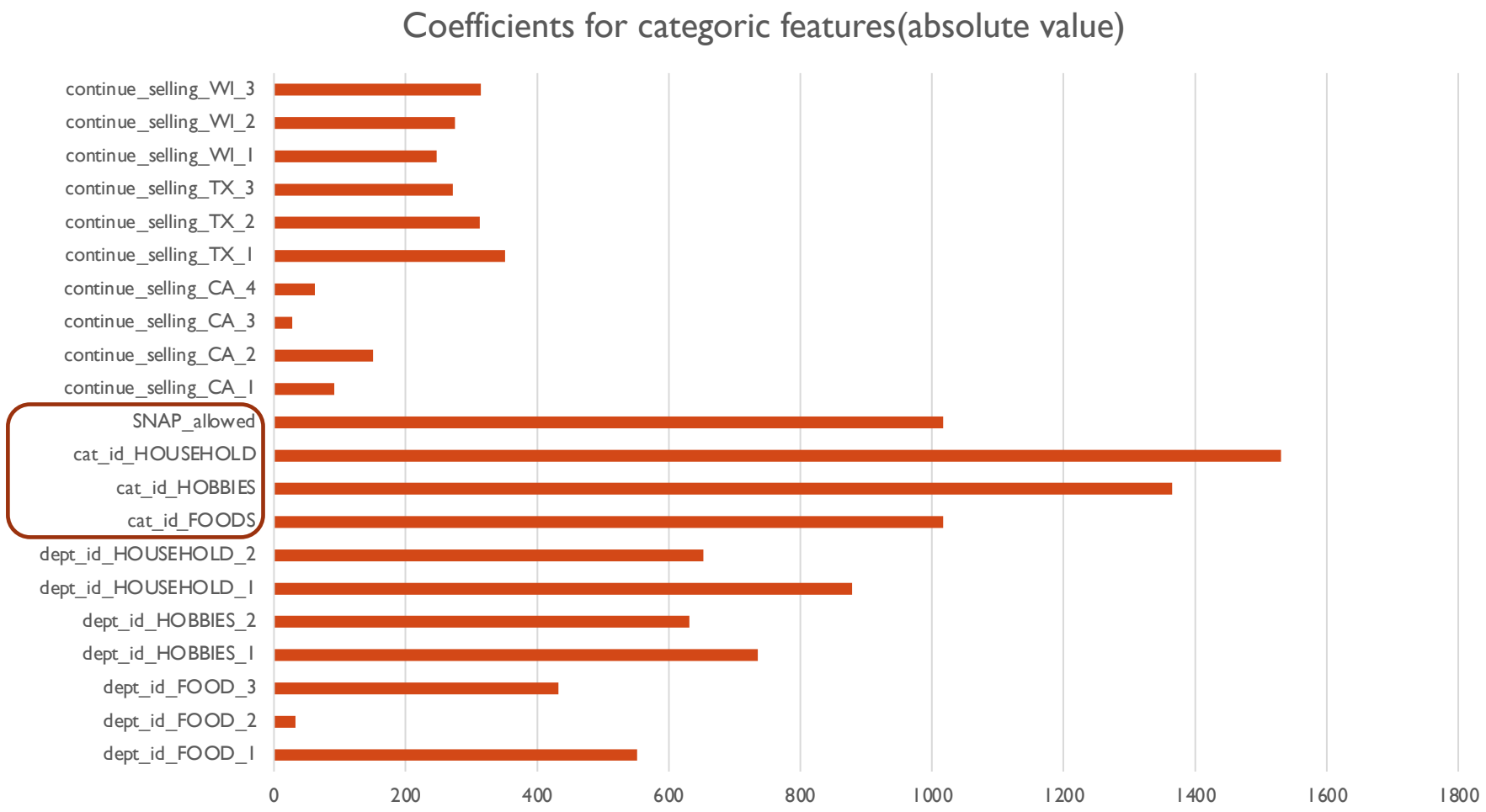
TESTING RESULT

	Linear	Ridge	Lasso	SVM linear	Random Forest	Neural Networks
RMSE	2211.48	2209.94	2213.66	2254.70	2571.69	2211.14
R^2	0.942433	0.942513	0.942320	0.940161	0.922153	0.942451

FEATURE IMPORTANCE – NUMERIC FEATURES



FEATURE IMPORTANCE – CATEGORIC FEATURES



APPLICATION TO THE REAL WORLD

- Potential and current Walmart's suppliers should make production plan based on the category of their products and use previous sales number as reference.

Q&A