**Introduction**

The purpose of the project is to predict trip types for each visit for Walmart. To improve customer’s shopping experience, what Walmart has already done is segmenting their store visits into different trip types. Whether the customers are purchasing stuff for parties or making their way through a weekly grocery list, Walmart wants to target them with different marketing strategies. Therefore, accurately predicting trip types with machine learning algorithm is extremely important for Walmart. Currently, Walmart’s customer trip type is classified into 38 categories. We want to employ multiclass classification to solve the problem of classifying each visit into one single type.

**Data Preparation**

The dataset is from Kaggle and contains 1,300,700 rows and 7 columns. Each row is an item purchased corresponding to a single trip by a single customer. The features are TripType, Upc, ScanCount, DepartmentDescription, FinelineNumber, VisitNumber, and Weekday.

|  |  |
| --- | --- |
| Feature | Description |
| DepartmentDescription | Description of the product’s department |
| FinelineNumber | A refined category for the product purchased |
| ScanCount | The number of the given item that was purchased. A negative value indicates a product returned |
| TripType | The 38 original trip types. |
| Upc | The Universal Product Code of the product purchased |
| VisitNumber | A unique id corresponding to a single trip by a single customer |
| Weekday | The weekday of the customer visit |

We selected 5 columns including ScanCount, DepartmentDescription, FinelineNumber, VisitNumber, Weekday as our primary features. We didn’t include Upc as the Upc we were provided with was not complete and accessing to Walmart API to extract product information based on Upc is not available. Our data preparation includes:

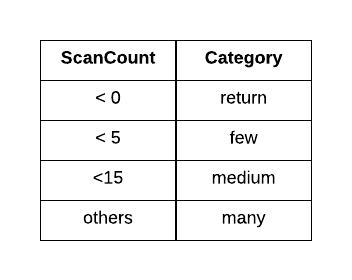
* Missing values imputation
* Categorize some features
* One Hot Encoding for each feature

Missing Values Imputation

All missing values are in DepartmentDescription and FinelineNumber. We found that lots of missing value in FinelineNumber are from Pharmacy department. Instead of dropping these instances, we created a new FinelineNumber – 9999 for them.

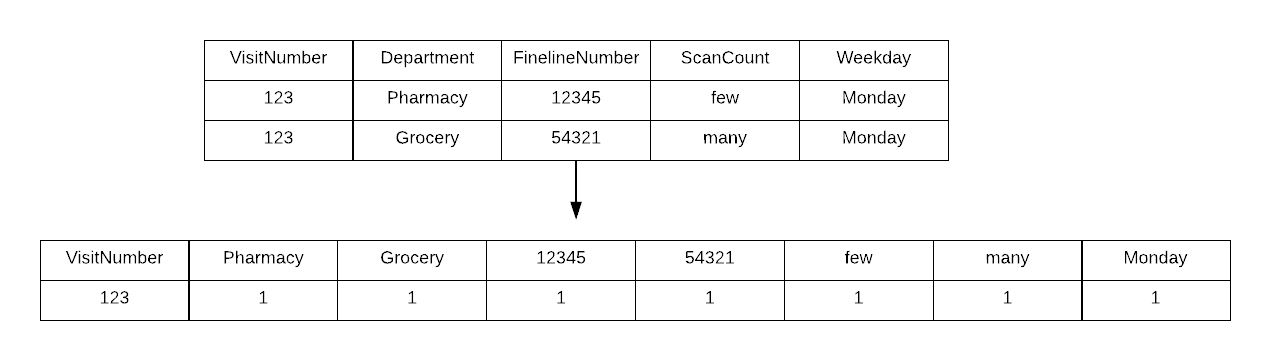
Categorize some features

Since the range of ScanCount is large, we decided to category this feature into three categories based on the number of the ScanCount.



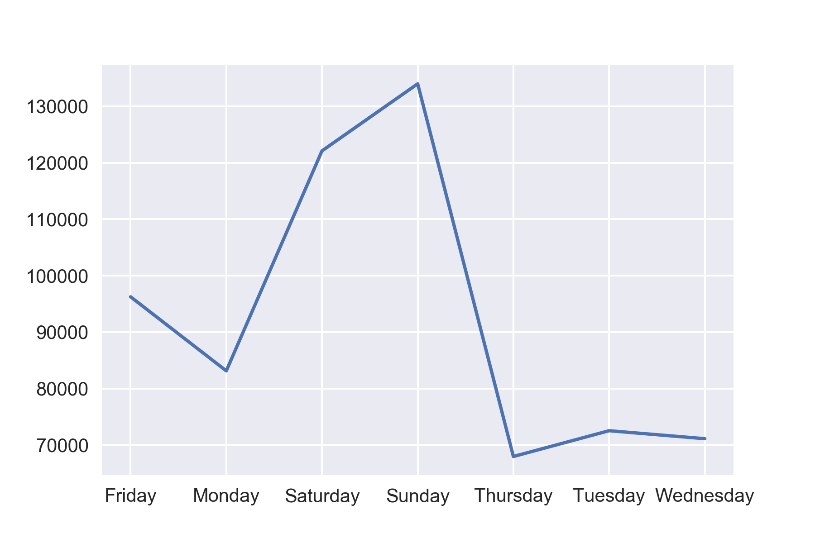
One Hot Encoding for each feature

We spread the FinelineNumber, Weekday, and DepartmentDescription, ScanCount into columns, turning our data frame into a large sparse matrix. The shape of our matrix now becomes: 95,516 rows and 5,275 columns.

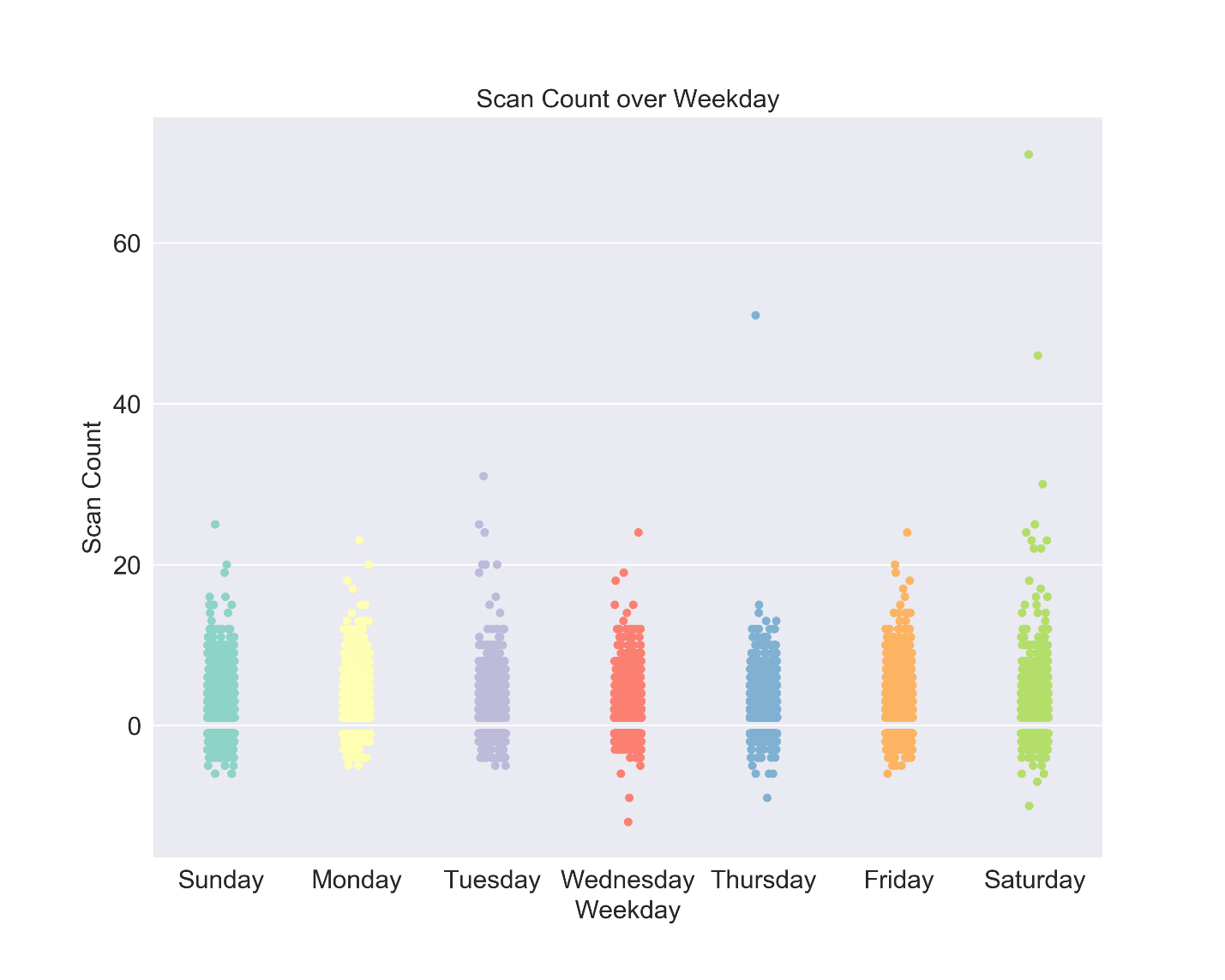


**Exploratory Data Analysis**

We first want to know visit frequency over weekdays. Most customers come to Walmart during the weekday.



However, the number of product purchased is not different across the weekdays. However, from the boxplot we could see that the range on Saturday is larger than other days. (scan count smaller than 0 means product returned)



We then want to know the most common things that customers purchase in Walmart