# Predicting LTV and Repeat Customers for Shopify Stores with Unsupervised Machine Learning

BY BENJAMIN BELLMAN

# Outline:

- Introduction
- Data Exploration
- Data Wrangling
- Exploratory Data Analysis
- Modelling
- Business Recommendations

# Outline:

- Introduction
- Data Exploration
- Data Wrangling
- Exploratory Data Analysis
- Modelling
- Business Recommendations

# Introduction

- E-commerce is **$4.89 trillion** industry.
- Shopify a leading platform with **1.7 million** active stores.
- Project aims to get most of data Shopify provides.
- Supervised Machine Learning models (ML) to predict:

- *Post-First Week Customer Spending (Regression)*

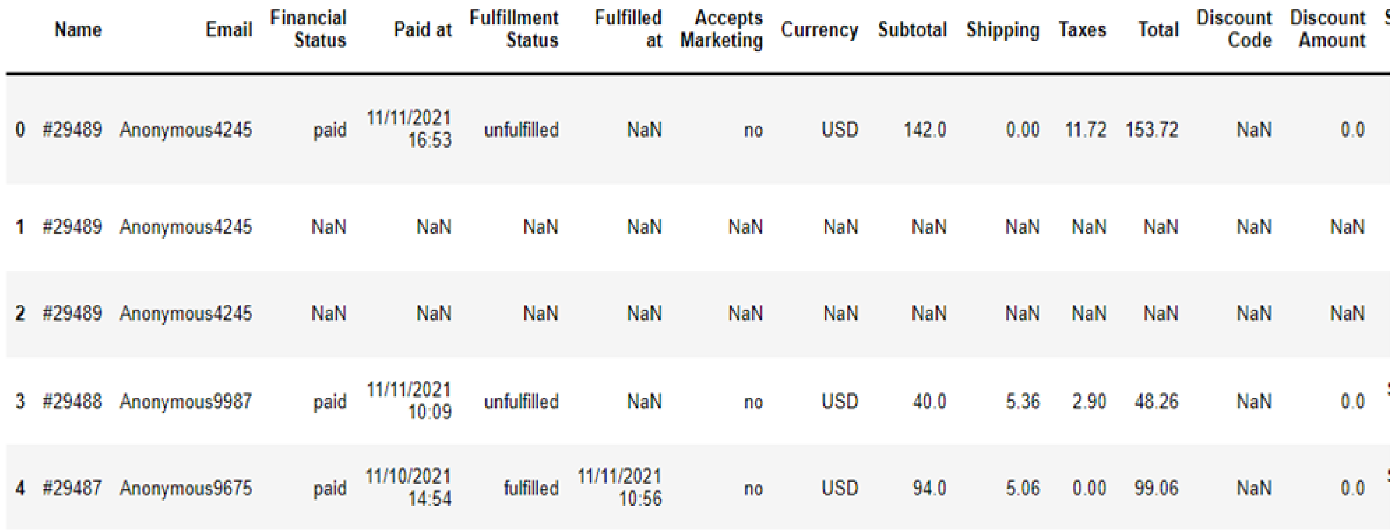
- *Repeat Orders Past The First Week (Classification)*

# Data Exploration and Wrangling

# Data Exploration

- Shopify Order Exports standardized; each row represents item sold with **73** columns.
- **Feature Columns**:

- *Name*
- *Email*
- *Paid at*
- *Subtotal*
- *Line-Item Name*
- *Line-Item Quality*



| | Name | Email | Financial Status | Paid at | Fulfillment Status | Fulfilled at | Accepts Marketing | Currency | Subtotal | Shipping | Taxes | Total | Discount Code | Discount Amount |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | #29489 | Anonymous4245 | paid | 11/11/2021 16:53 | unfulfilled | NaN | no | USD | 142.0 | 0.00 | 11.72 | 153.72 | NaN | 0.0 |
| 1 | #29489 | Anonymous4245 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | #29489 | Anonymous4245 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | #29488 | Anonymous9987 | paid | 11/11/2021 10:09 | unfulfilled | NaN | no | USD | 40.0 | 5.36 | 2.90 | 48.26 | NaN | 0.0 |
| 4 | #29487 | Anonymous9675 | paid | 11/10/2021 14:54 | fulfilled | 11/11/2021 10:56 | no | USD | 94.0 | 5.06 | 0.00 | 99.06 | NaN | 0.0 |

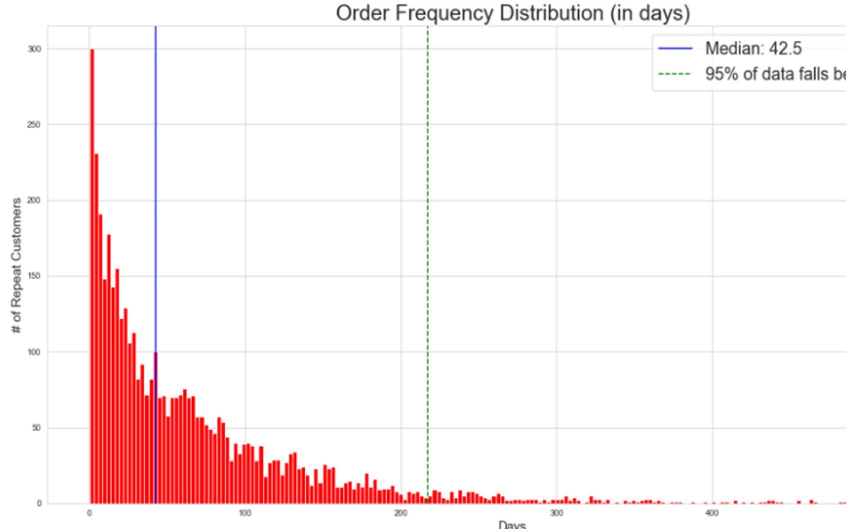
# Data Exploration (2)

- Dataset from a Women’s Sports Fitness Apparel Store.
- Generated $**1.8 million** in YTD revenue since June 2018.
- Contains **50,418** items purchased since inception.
- **16,180** unique customers ; *28%* are repeat
- **25,188** unique orders; *53%* made by repeat customers.
- **560** unique item skus.

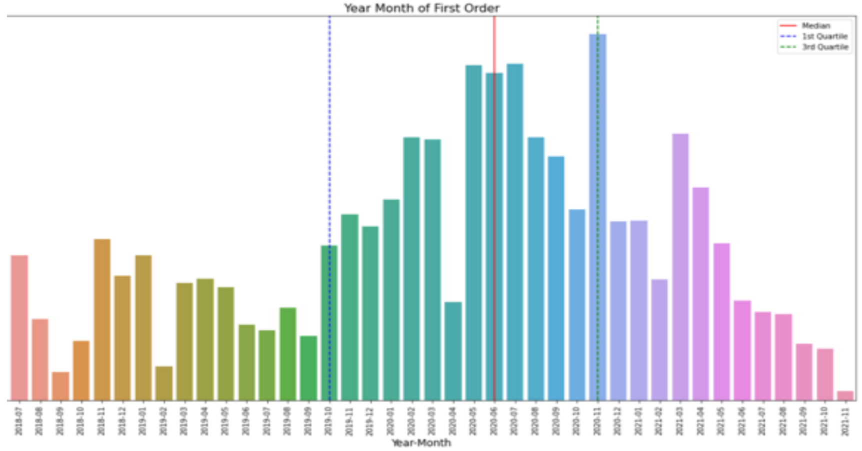
# Data Exploration Summary:

- Median Order frequency is **43** days.
- Unlikely to make an order after **218** days.
- **75%** first prior to *<u>2020-11</u>*, final data ends in *<u>2021-11</u>*.
- 1-year LTV Threshold.

  - Lose only ~ **25%** of our data



Distribution of First Orders by month since June 2018 .

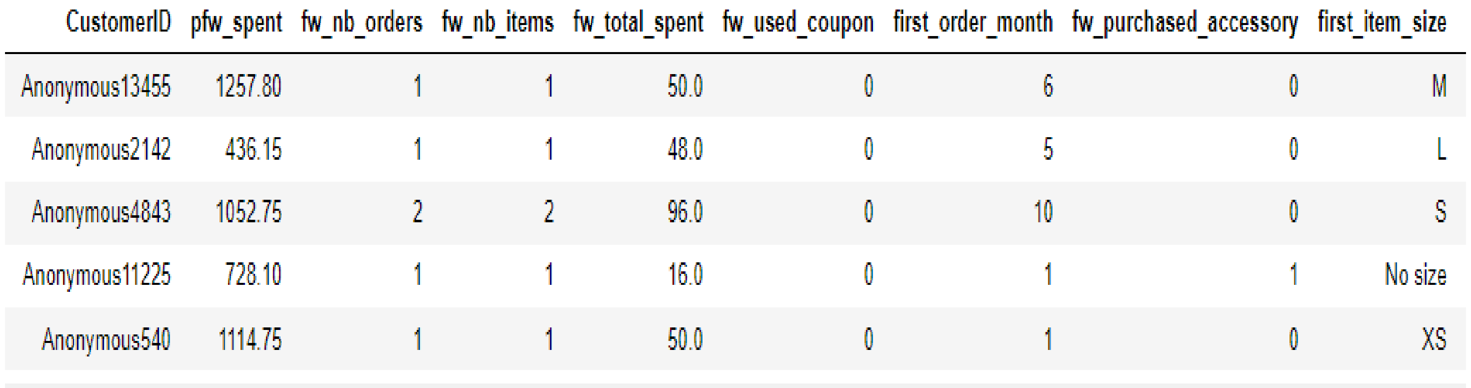


# Data Wrangling

- **Objective**: Predict Behavior for *358* days following the week of their first order.

- Reshape the DataFrame by customer:

- *First Order Date*
- *Most Recent Order Date*
- *LTV start*
- *LTV end*

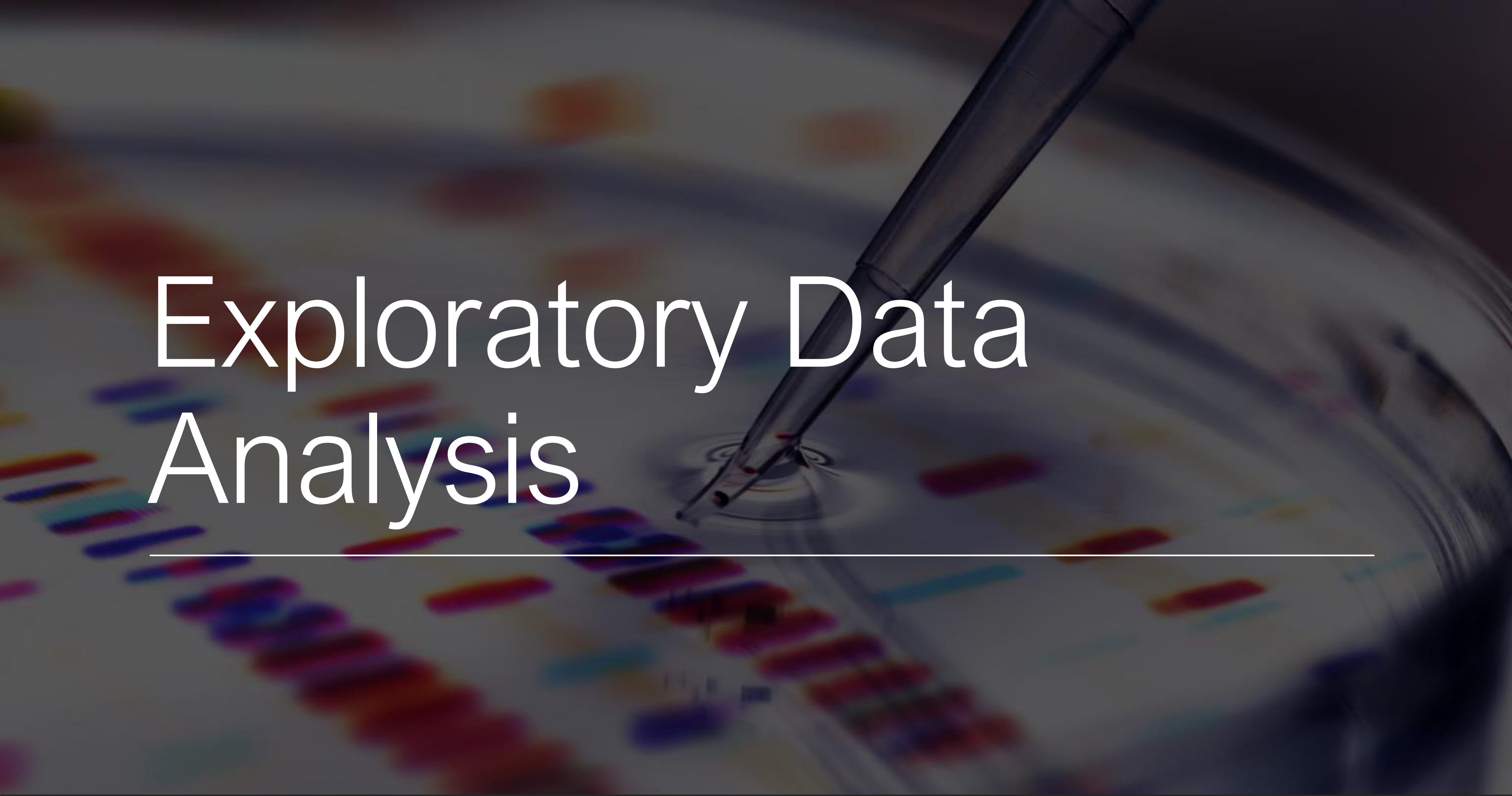


| CustomerID | pfw_spent | fw_nb_orders | fw_nb_items | fw_total_spent | fw_used_coupon | first_order_month | fw_purchased_accessory | first_item_size |
|---|---|---|---|---|---|---|---|---|
| Anonymous13455 | 1257.80 | 1 | 1 | 50.0 | 0 | 6 | 0 | M |
| Anonymous2142 | 436.15 | 1 | 1 | 48.0 | 0 | 5 | 0 | L |
| Anonymous4843 | 1052.75 | 2 | 2 | 96.0 | 0 | 10 | 0 | S |
| Anonymous11225 | 728.10 | 1 | 1 | 16.0 | 0 | 1 | 1 | No size |
| Anonymous540 | 1114.75 | 1 | 1 | 50.0 | 0 | 1 | 0 | XS |

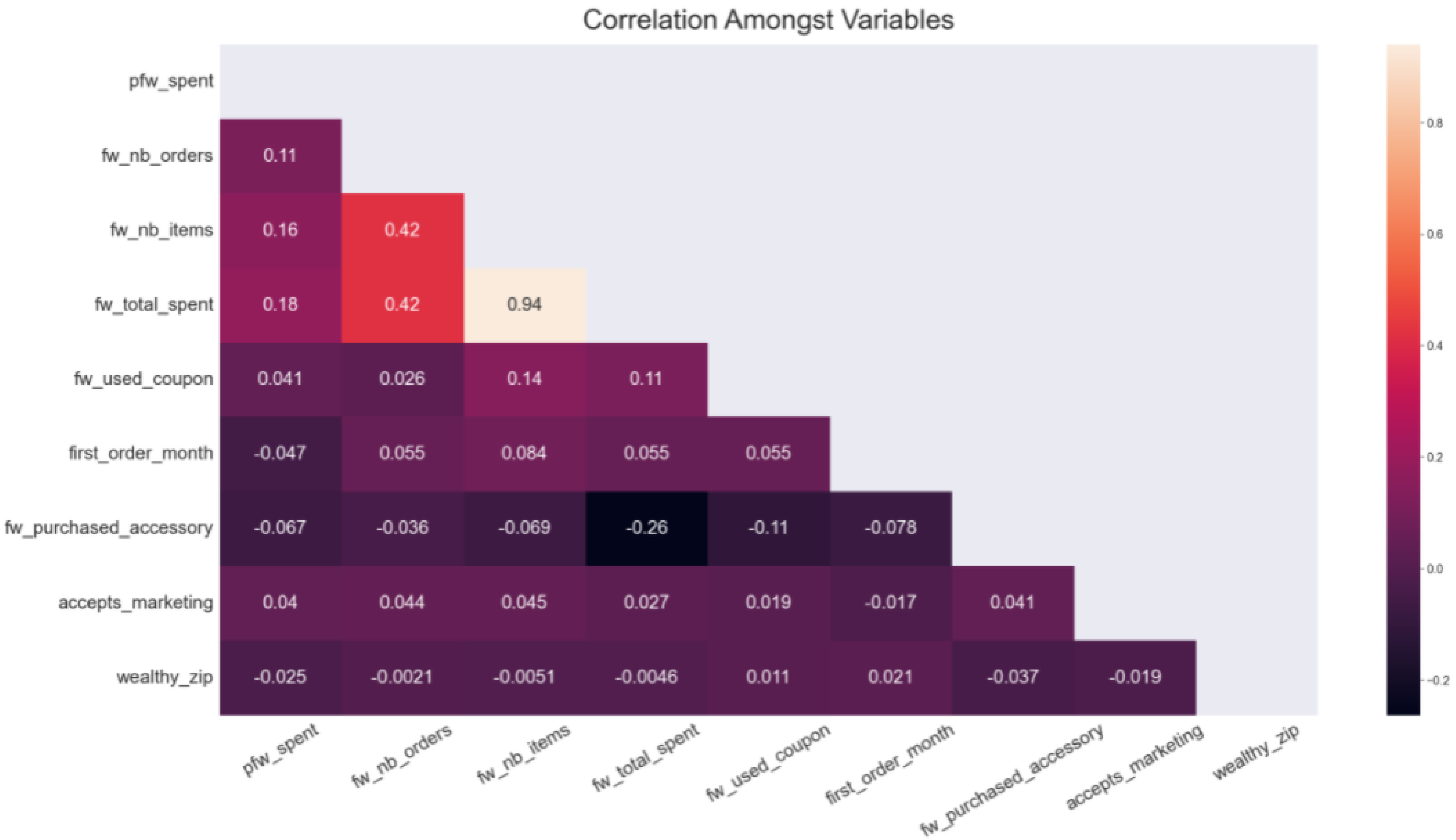
# Data Wrangling: Variables to test with *pfw_spend*

- fw_nb_orders
- fw_nb_items
- fw_total_spent
- fw_used_coupon
- fw_purchased_accessory
- location
- in_wealthiest zip code
- first_item_size
- accepts_marketing
- first_order_month
- first_purchase_price

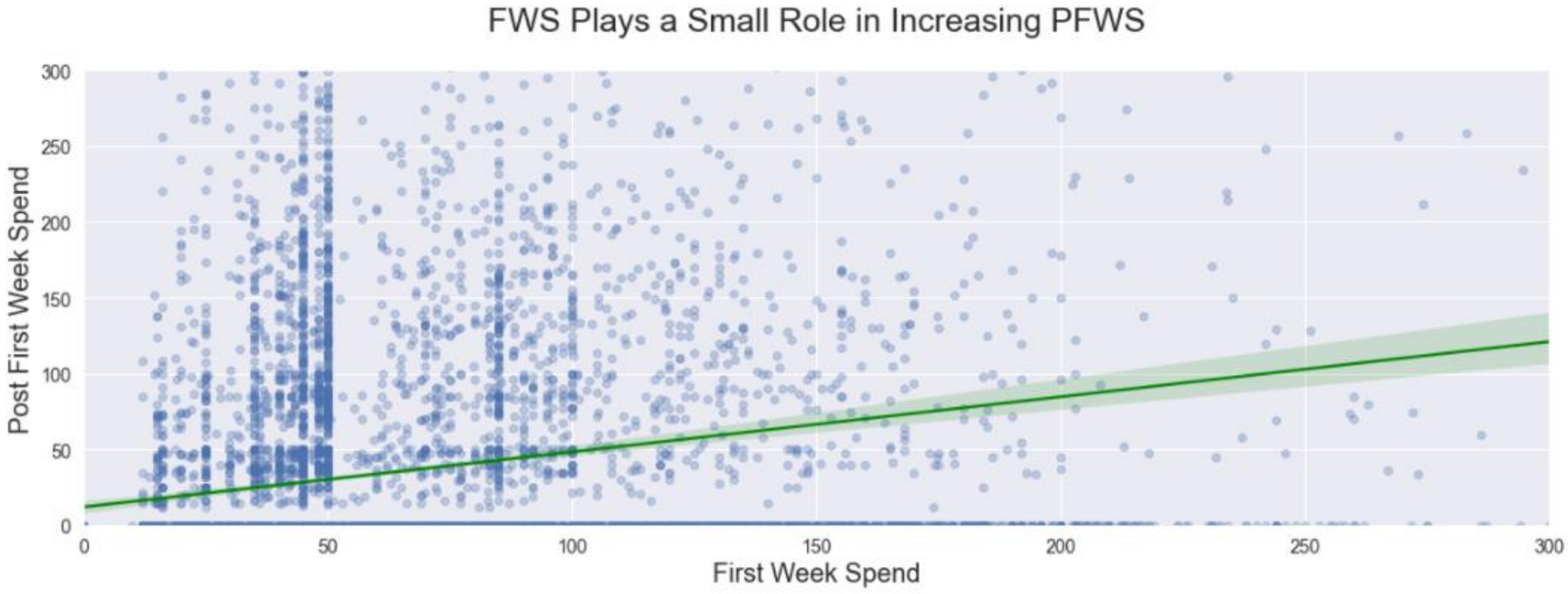
# Exploratory Data Analysis



# Correlation Amongst Variables



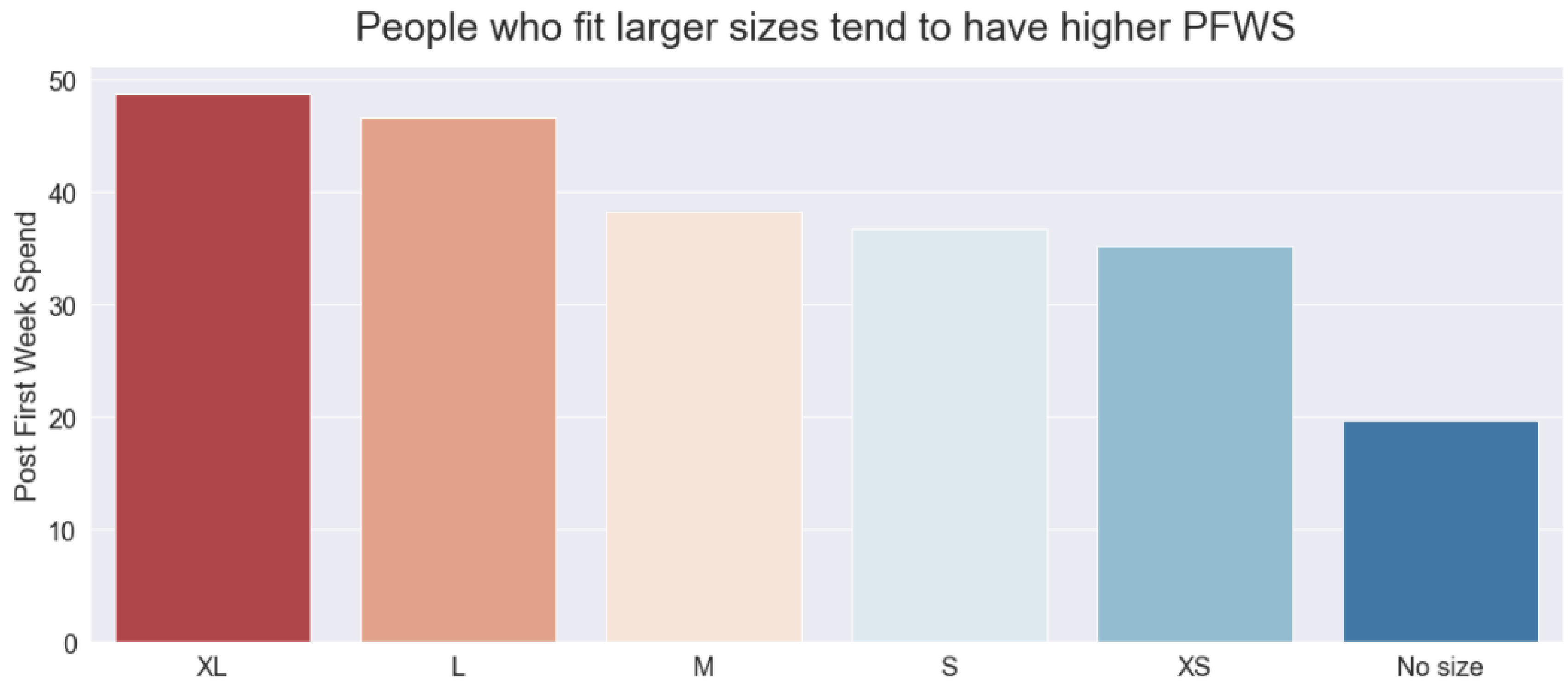
| | pfw_spent | fw_nb_orders | fw_nb_items | fw_total_spent | fw_used_coupon | first_order_month | fw_purchased_accessory | accepts_marketing | wealthy_zip |
|---|---|---|---|---|---|---|---|---|---|
| pfw_spent | | | | | | | | | |
| fw_nb_orders | 0.11 | | | | | | | | |
| fw_nb_items | 0.16 | 0.42 | | | | | | | |
| fw_total_spent | 0.18 | 0.42 | 0.94 | | | | | | |
| fw_used_coupon | 0.041 | 0.026 | 0.14 | 0.11 | | | | | |
| first_order_month | -0.047 | 0.055 | 0.084 | 0.055 | 0.055 | | | | |
| fw_purchased_accessory | -0.067 | -0.036 | -0.069 | -0.26 | -0.11 | -0.078 | | | |
| accepts_marketing | 0.04 | 0.044 | 0.045 | 0.027 | 0.019 | -0.017 | 0.041 | | |
| wealthy_zip | -0.025 | -0.0021 | -0.0051 | -0.0046 | 0.011 | 0.021 | -0.037 | -0.019 | |



# Post First Week Spend vs. First Week Spend

A CUSTOMER SPENDING MORE IN THEIR FIRST WEEK TENDS TO SPEND MORE 358 DAYS LATER.

XL Mean:
**$48**

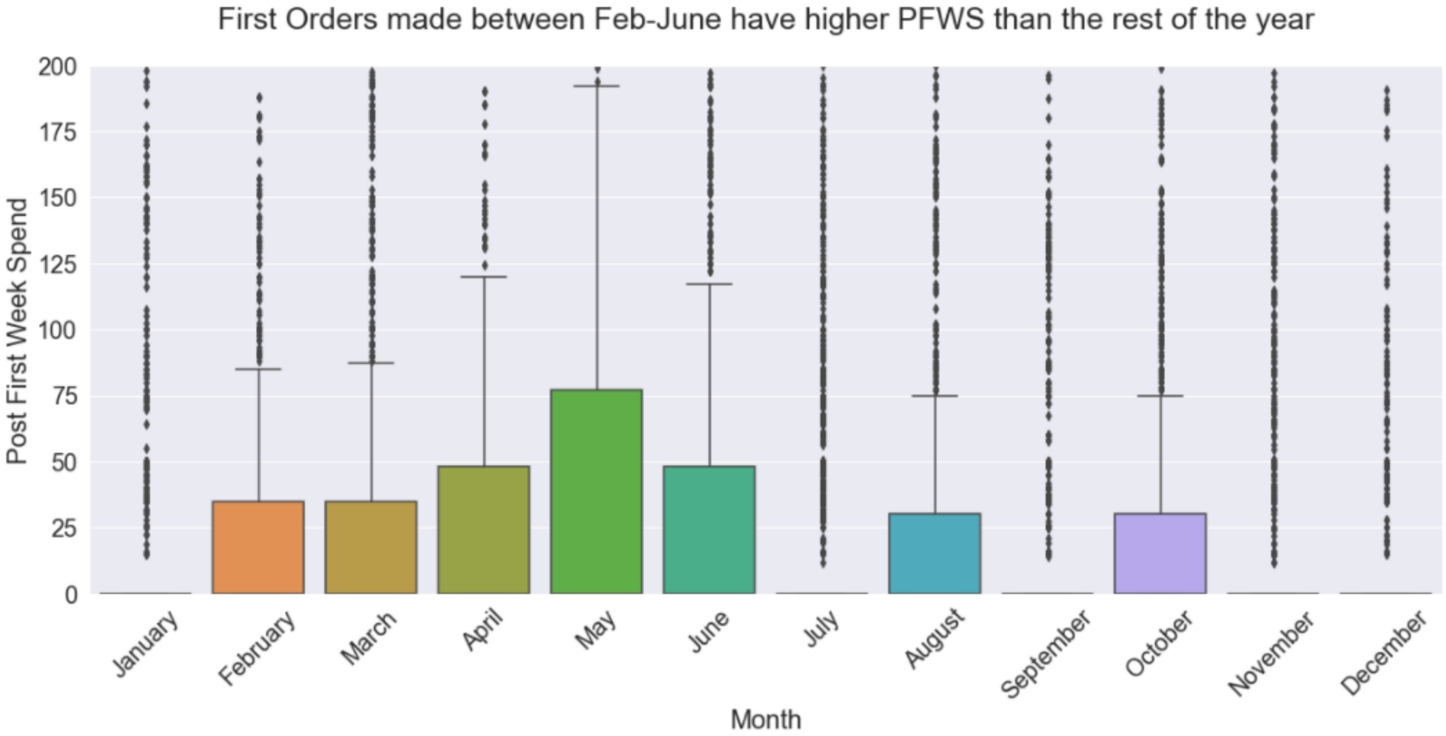


XS Mean:
$**38**

# Post First Week Spend vs. First Item Size

PFWS IS LOWER FOR CUSTOMERS WHO ORDER SMALLER SIZES.

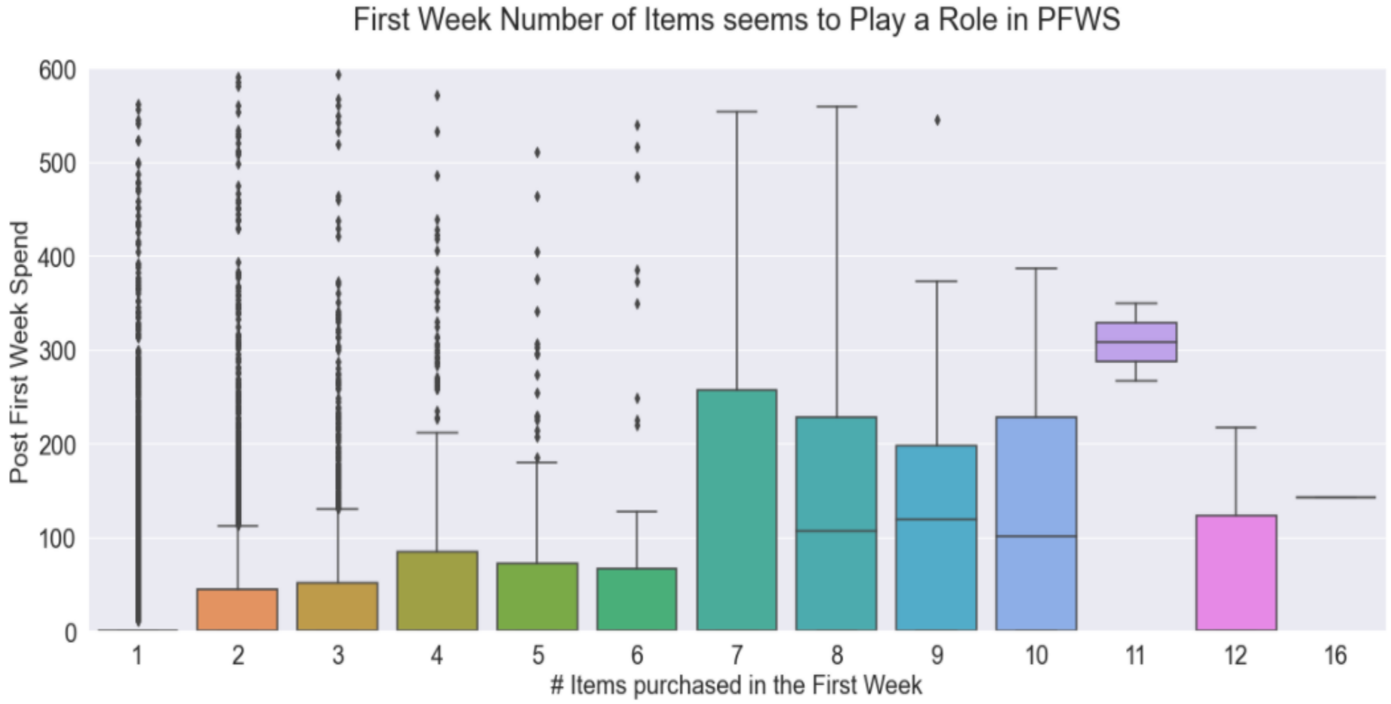
Highest Month:
Mean (May):
**$59**



Lowest Month
Mean (Nov):
**$24**

# Post First Week Spend vs. **Month First Order**

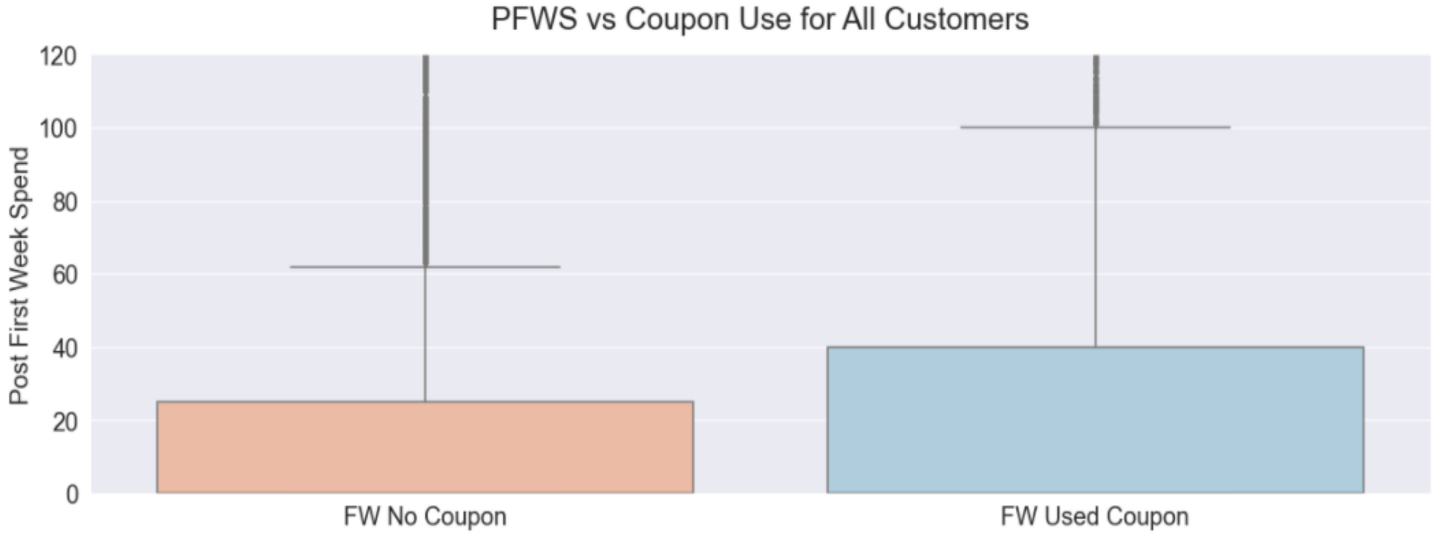
CUSTOMERS MAKING THEIR FIRST PURCHASE THROUGH FEB-JUNE TEND TO SPEND MORE 358 DAYS LATER



# Post First Week Spend vs. First Week Items Purchased

CUSTOMERS WHO PURCHASE MORE ITEMS PURCHASED MORE IN THE PRECEDING 358 DAYS.

No Coupon
Mean:
**$33**

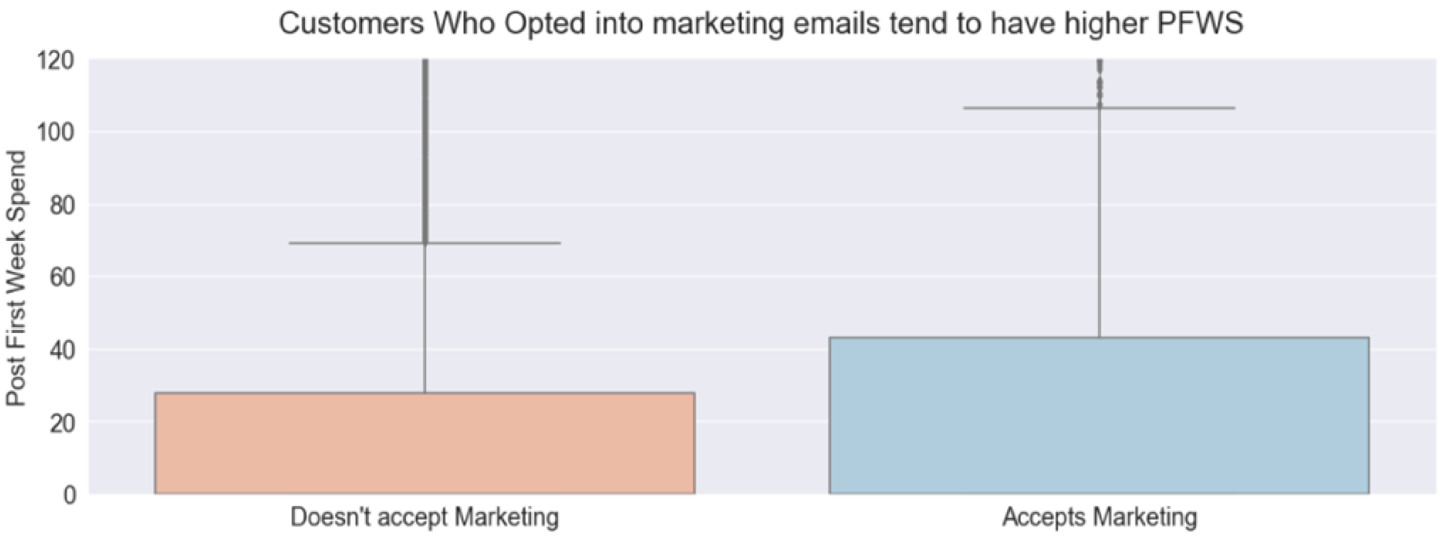


Used Coupon
Mean:
**$42**

# Post First Week Spend vs. Coupon Use in First Week

CUSTOMERS WHO USE COUPONS HAVE HIGHER PFWS

Opted Out
Mean:
**$35**



Opted In
Mean:
**$45**

# Post First Week Spend vs. Opting in Marketing Emails

CUSTOMERS WHO OPT INTO MARKETING EMAILS HAVE HIGHER PFWS.

# Preprocessing and Modelling

# Pre-Processing

- **Regression Models used**: *Linear Regression* and *Random Forest Regressor*
- **Drop columns** in EDA deemed not relevant and those that are redundant.
- **Label Encoding:** Changed *Month of First Order*.
- **One-Hot Encoding** on Categorical Variables
- **Scale** Data for *First Week Spend Data*.
- **Train Test Split** at 0.75 /0.25 ; training data has *7942* , testing has *2648*
- **Grid Search CV** to perform hyperparameter tuning with 5-fold cross validation.

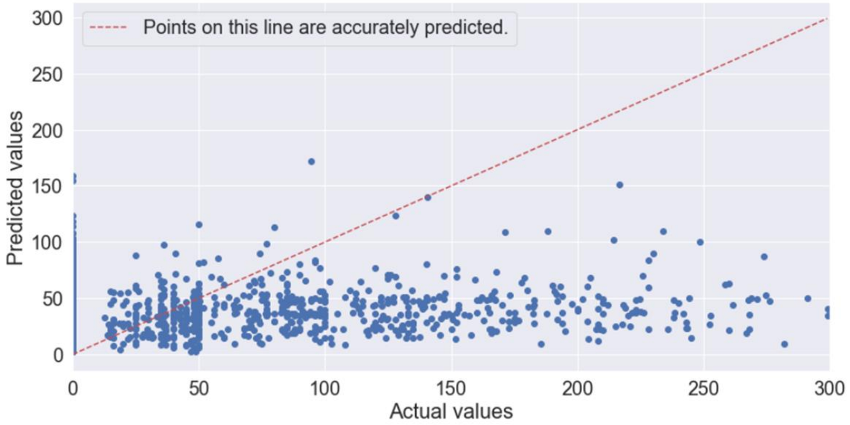
# Regression Model Results

- *Multiple Linear Regression Results*:

- **R squared** of ~ 0.04 on testing data.

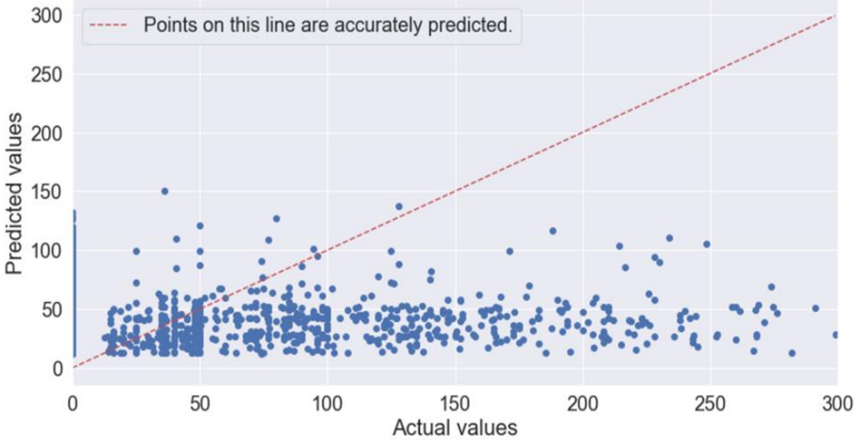
- Only 4% of the variance can be explained

- Linear Model not best at explaining the relationship.



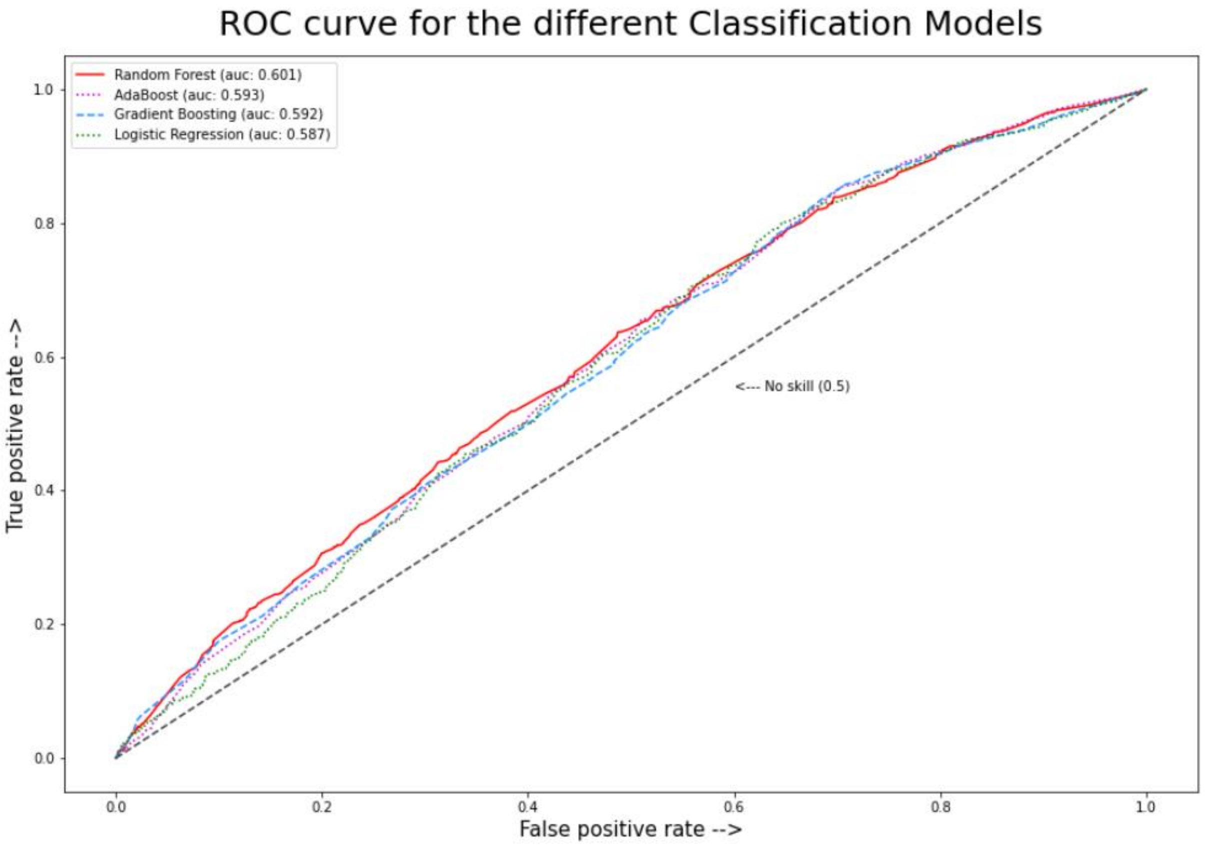
- *Random Forest Regressor:*

- **R squared** of 0.036 on testing data.



# Switch to Classification

- Reformulate our initial question “Can we predict the post first week spend?” to “ Can we predict if a customer will purchase again after the first week ?”
- New target column ***repeat***
- Drop ***post_first_week_spend***
- Address Class imbalance with resampling of training data 50/50.



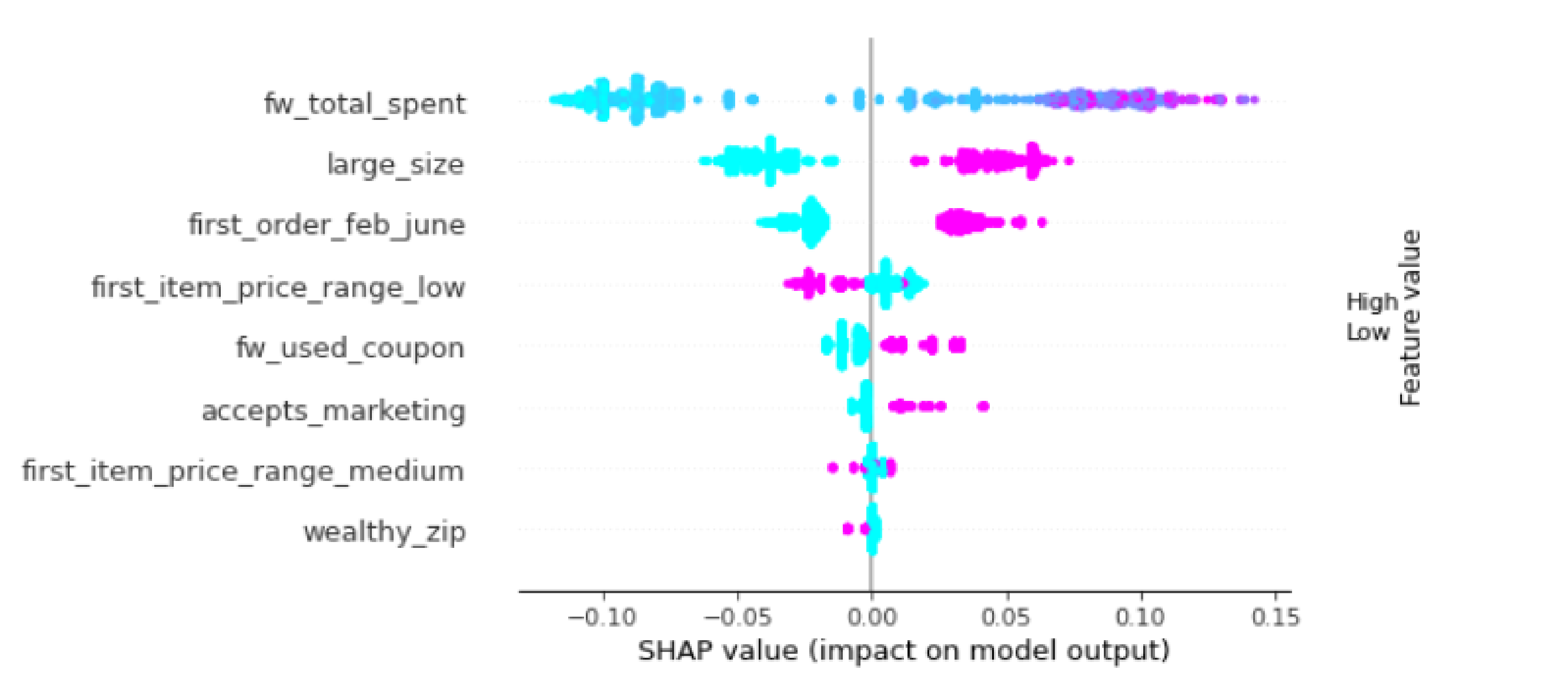
# Classification Results

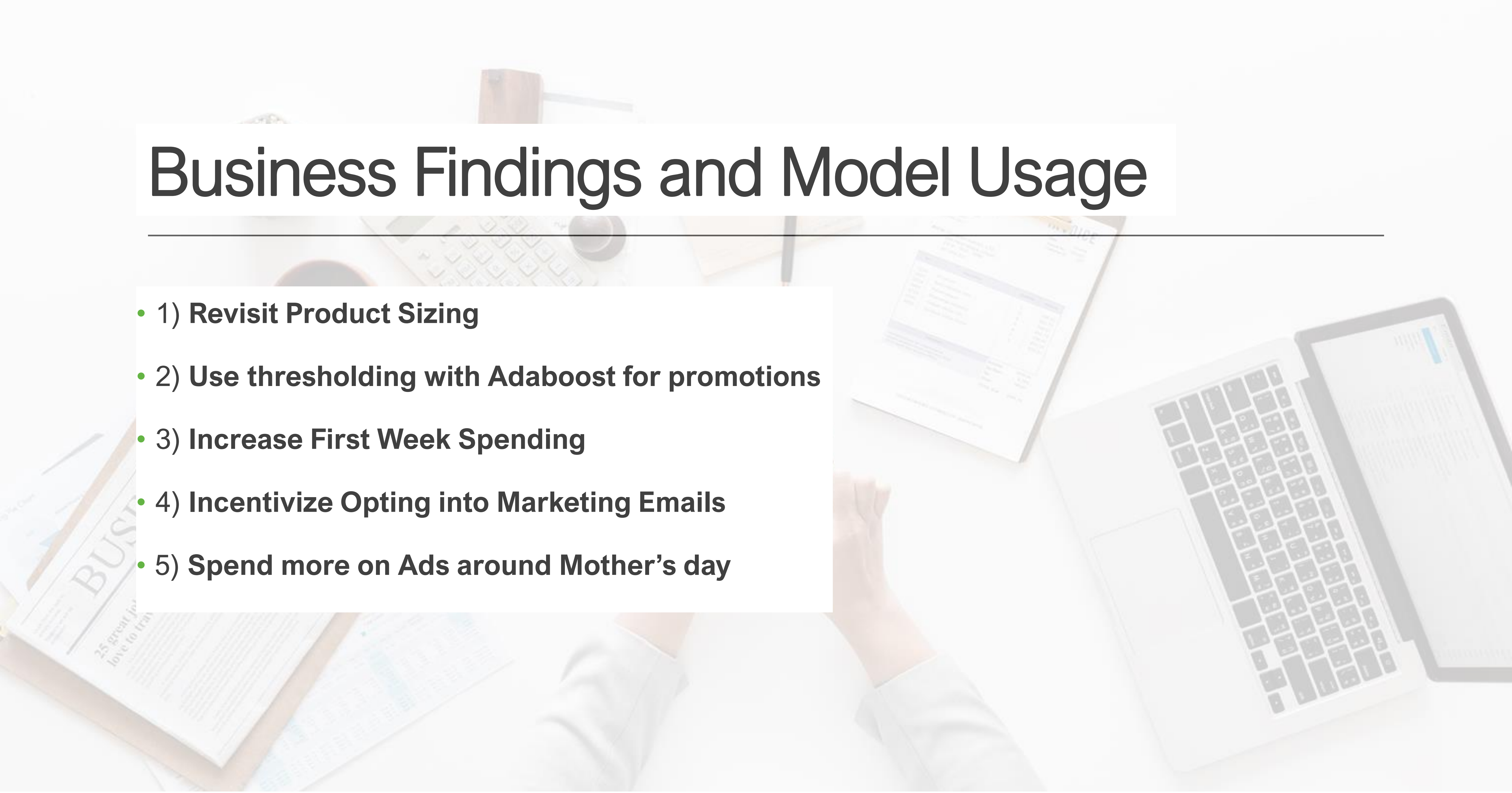
- Use the ***Adaboost*** model to get the best *precision (+20%) and good recall (+31%)* to predict repeat customers.
- Use ***Logistic Regression*** model to predict non-repeat customers.

| Model | Best Hyperparameters | ROC_AUC | F1_Score (weighted) | Precision | Recall | Accuracy |
|---|---|---|---|---|---|---|
| **Logistic Regression** | {'C': 0.001, 'l1_ratio': 0, 'penalty': 'l2'} | **0.587** | **0.384** | 0.319 | **0.484** | **0.579** |
| **Random Forest Classifier** | {'n_estimators': 10, 'max_features': 'sqrt', 'max_depth': 4, ' criterion': 'gini', 'bootstrap': 'False'} | **0.601** | **0.437** | 0.318 | **0.697** | **0.512** |
| **AdaBoost Classifier** | {'n_estimators': 250, 'learning_rate': 0.01, 'min_samples_leaf': 10, be_max_depth': 2} | 0.593 | 0.433 | **0.323** | 0.656 | 0.534 |
| **Gradient Boosting Classifier** | {'n_estimators': 5, 'max_depth': 3, 'loss': 'deviance', 'criterion': 'mae'} | 0.592 | 0.419 | **0.315** | 0.624 | 0.530 |
| **No Skill Classifier** | NA | **0.5** | **0.35** | **0.27** | **0.5** | **0.5** |

# Business Cases

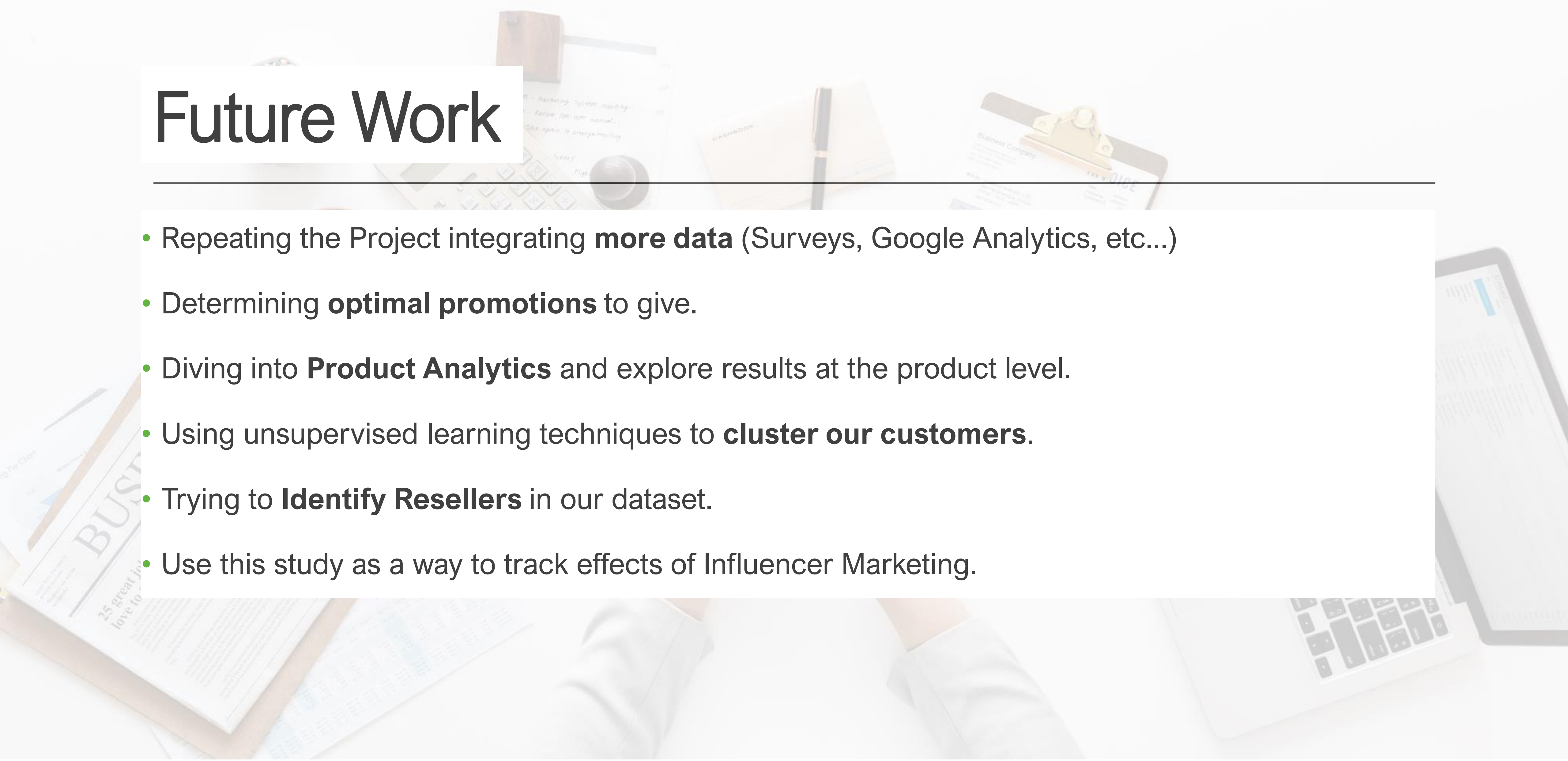
# Feature Importance in the Model





# Business Findings and Model Usage

- 1) **Revisit Product Sizing**
- 2) **Use thresholding with Adaboost for promotions**
- 3) **Increase First Week Spending**
- 4) **Incentivize Opting into Marketing Emails**
- 5) **Spend more on Ads around Mother’s day**



# Future Work

- Repeating the Project integrating **more data** (Surveys, Google Analytics, etc...)
- Determining **optimal promotions** to give.
- Diving into **Product Analytics** and explore results at the product level.
- Using unsupervised learning techniques to **cluster our customers**.
- Trying to **Identify Resellers** in our dataset.
- Use this study as a way to track effects of Influencer Marketing.

any questioms?
30
40
50