

Acquire Valued Customers

predict repeat buyers

Mudassir Syed

Project

➤ Predict Repeat Customer

- Businesses run discounts on select items to increase sales and make the customers repeat their purchase.
- Data used in this project is of customers transactions, history and offers.
- Original data comprises of almost 350 million rows of completely anonymized transactional data from over 300,000 shoppers.
- This project deals with a subset of original data.

➤ Outcome from this Project

- Provide insights into shopping patterns of customers.
- Predict the repeat buyers based on different features.

Clients

➤ Subscription Businesses

- Companies with Business Model revolving around Subscriptions/renewals
- Companies interested in building customer base

➤ Grocery Stores

- Which products when offered discount sell more
- Help convert customers to repeat buyers
- Retain existing customers

Data Acquisition

- Retrieved dataset from <https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data>
- This data comprises anonymized fields with transactions for 1 year along with Offer and Offer History information.
- Acquired dataset is reduced to manageable size for EDA and Model building.

Dataset Exploration

- 16000 Customer records
- 1.6 GB (transactions)
- 875 KB (history)
- 1.8KB (offers)

```
In [329]: offers.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37 entries, 0 to 36
Data columns (total 6 columns):
offer          37 non-null int64
category       37 non-null int64
quantity       37 non-null int64
company        37 non-null int64
offervalue     37 non-null float64
brand          37 non-null int64
dtypes: float64(1), int64(5)
memory usage: 1.8 KB
```

```
In [334]: trans.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17359709 entries, 0 to 17359708
Data columns (total 12 columns):
index          int64
id             int64
chain          int64
dept          int64
category       int64
company        int64
brand          int64
date           object
productsize    float64
productmeasure object
purchasequantity int64
purchaseamount float64
dtypes: float64(2), int64(8), object(2)
memory usage: 1.6+ GB
```

```
In [330]: hist.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16000 entries, 0 to 15999
Data columns (total 7 columns):
id              16000 non-null int64
chain          16000 non-null int64
offer          16000 non-null int64
market         16000 non-null int64
repeattrips    16000 non-null int64
repeater       16000 non-null object
offerdate      16000 non-null object
dtypes: int64(5), object(2)
memory usage: 875.1+ KB
```

Data Dictionary

- ***transactions***

id - A unique id representing a customer

chain - An integer representing a store chain dept - An aggregate grouping of the Category (e.g. water)

category - The product category (e.g. sparkling water)

company - An id of the company that sells the item

brand - An id of the brand to which the item belongs

date - The date of purchase

productsize - The amount of the product purchase (e.g. 16 oz of water)

productmeasure - The units of the product purchase (e.g. ounces)

purchasequantity - The number of units purchased

purchaseamount - The dollar amount of the purchase

Data Dictionary

- ***offers***

offer - A unique id representing an offer

category - The product category (e.g. sparkling water)

quantity - The number of units one must purchase to get the discount

company - An id of the company that sells the item

offervalue - The dollar value of the offer

brand - An id of the brand to which the item belongs

- ***history***

id - A unique id representing a customer

chain - An integer representing a store chain

offer - An id representing a certain offer

market - An id representing a geographical region

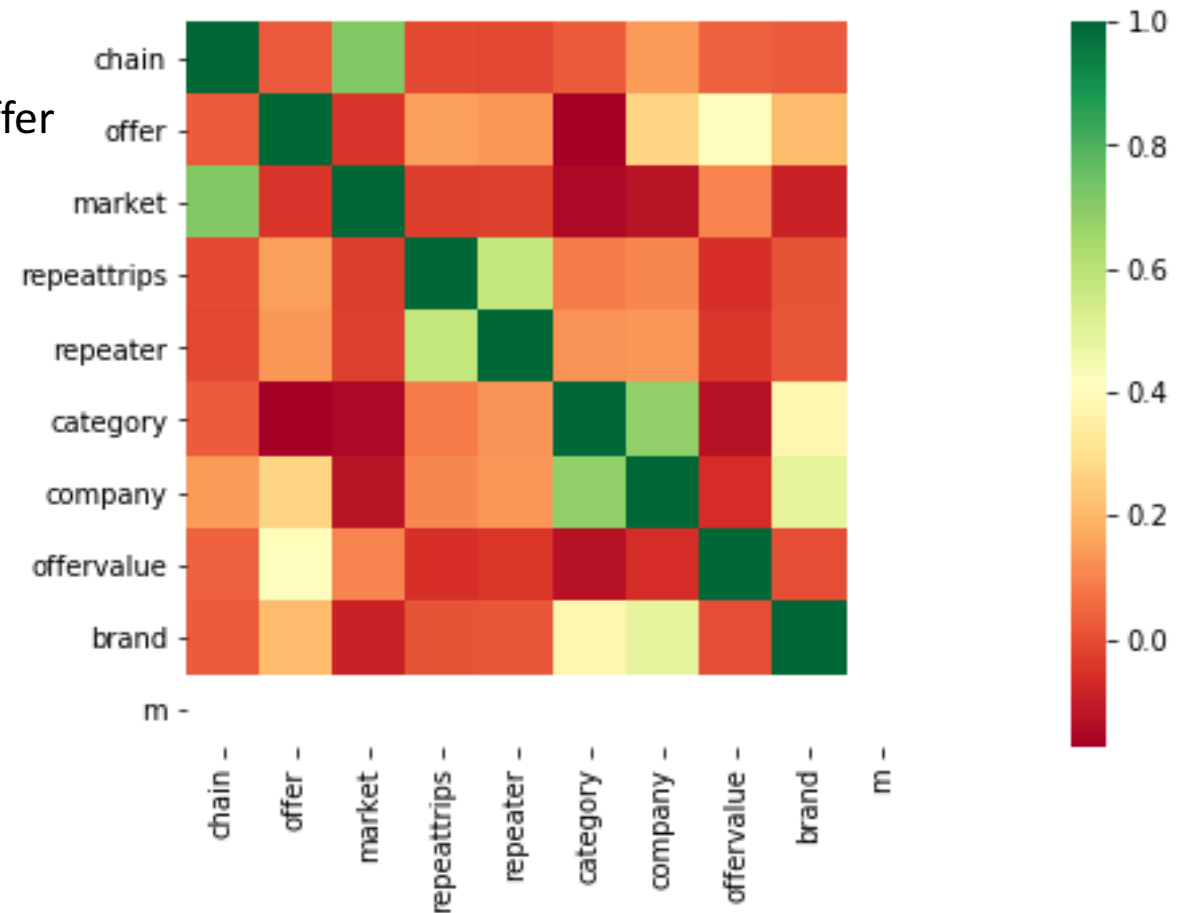
repeattrips - The number of times the customer made a repeat purchase

repeater - A boolean, equal to repeattrips > 0

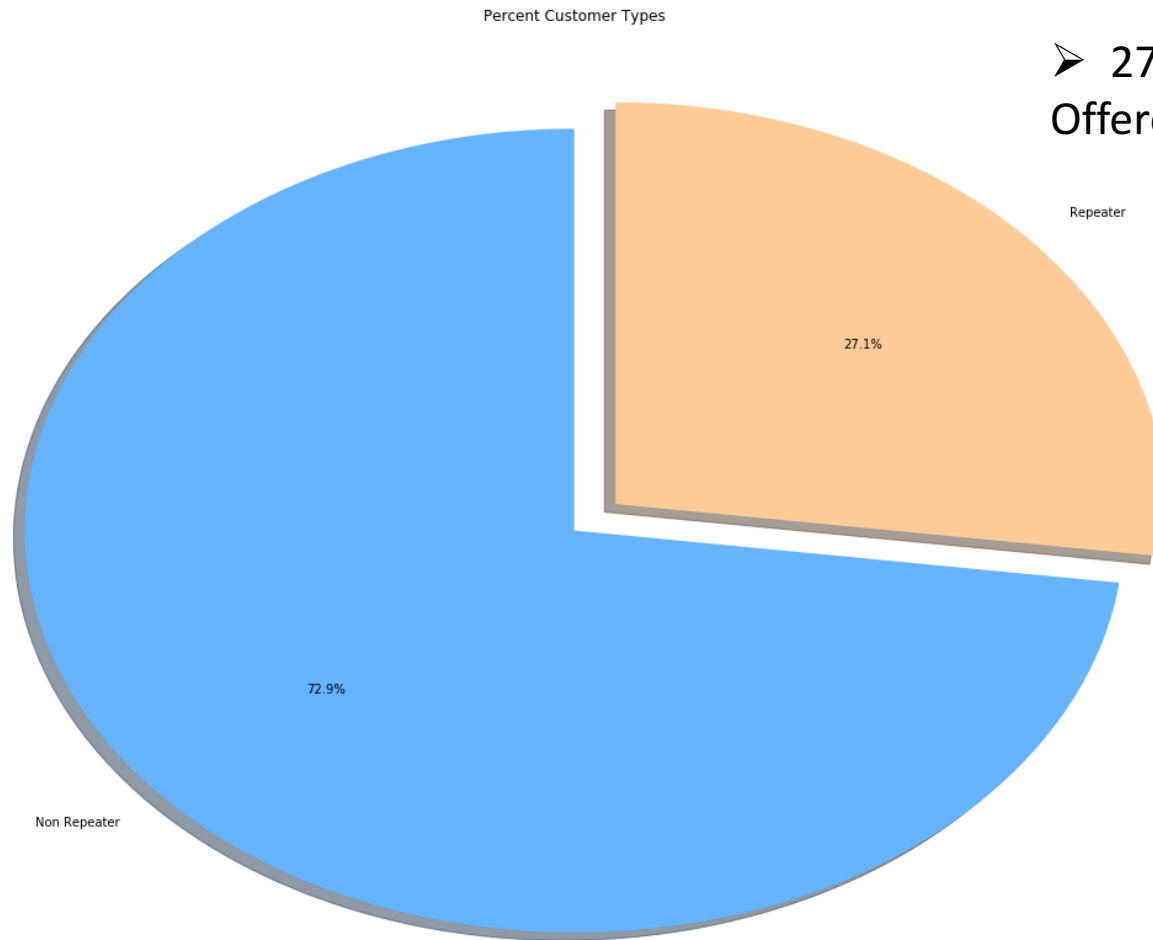
offerdate - The date a customer received the offer

Data Aggregation & Feature Selection

- Merge Offers and History data with OfferId into Hist_Offer
- Category, Company, Offer and Brand are correlated to target variable 'repeater'
- Chain and Market have little correlation to target variable

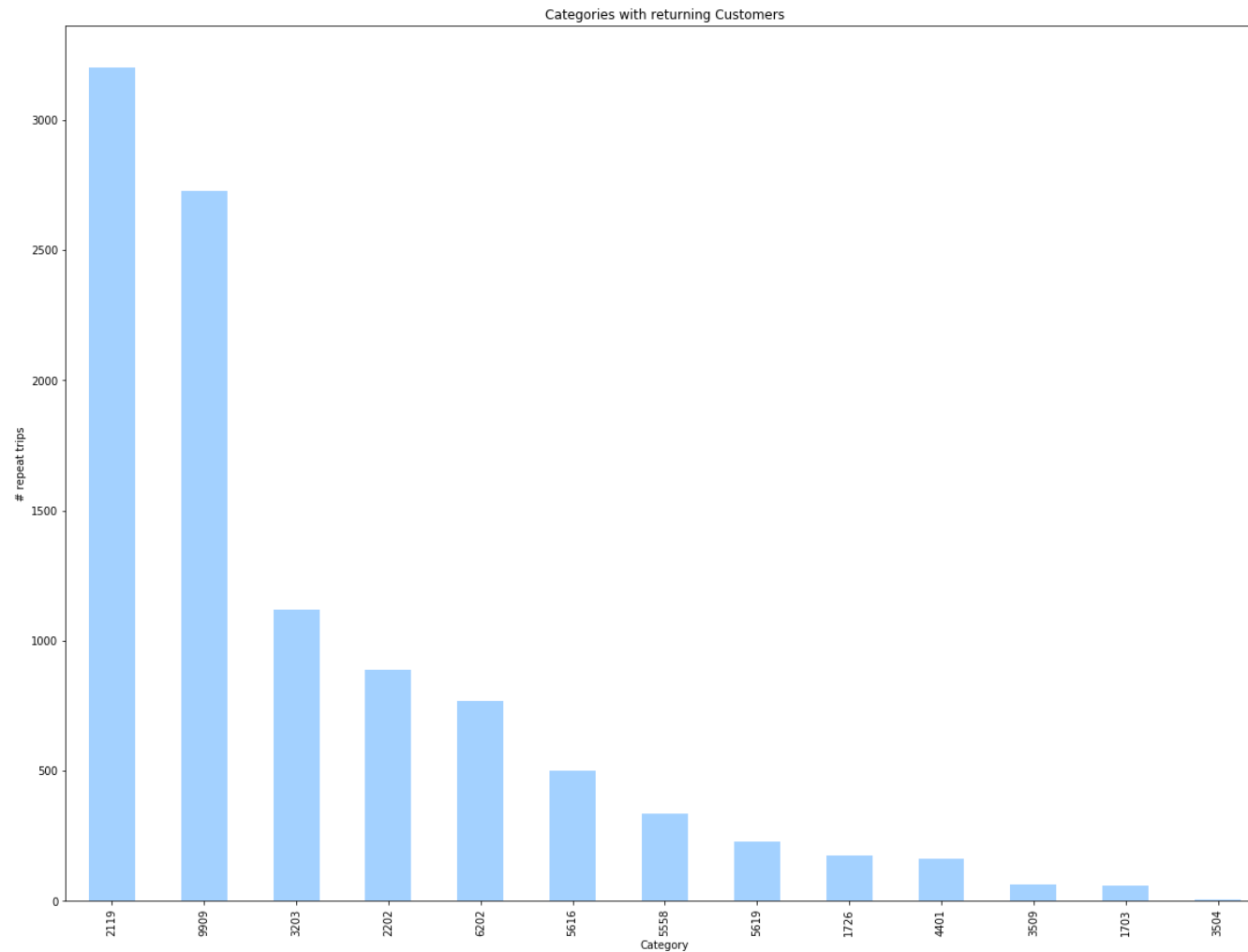


Explanatory Data Analysis



➤ 27 % of Customers repeat purchase when Offered discount coupon

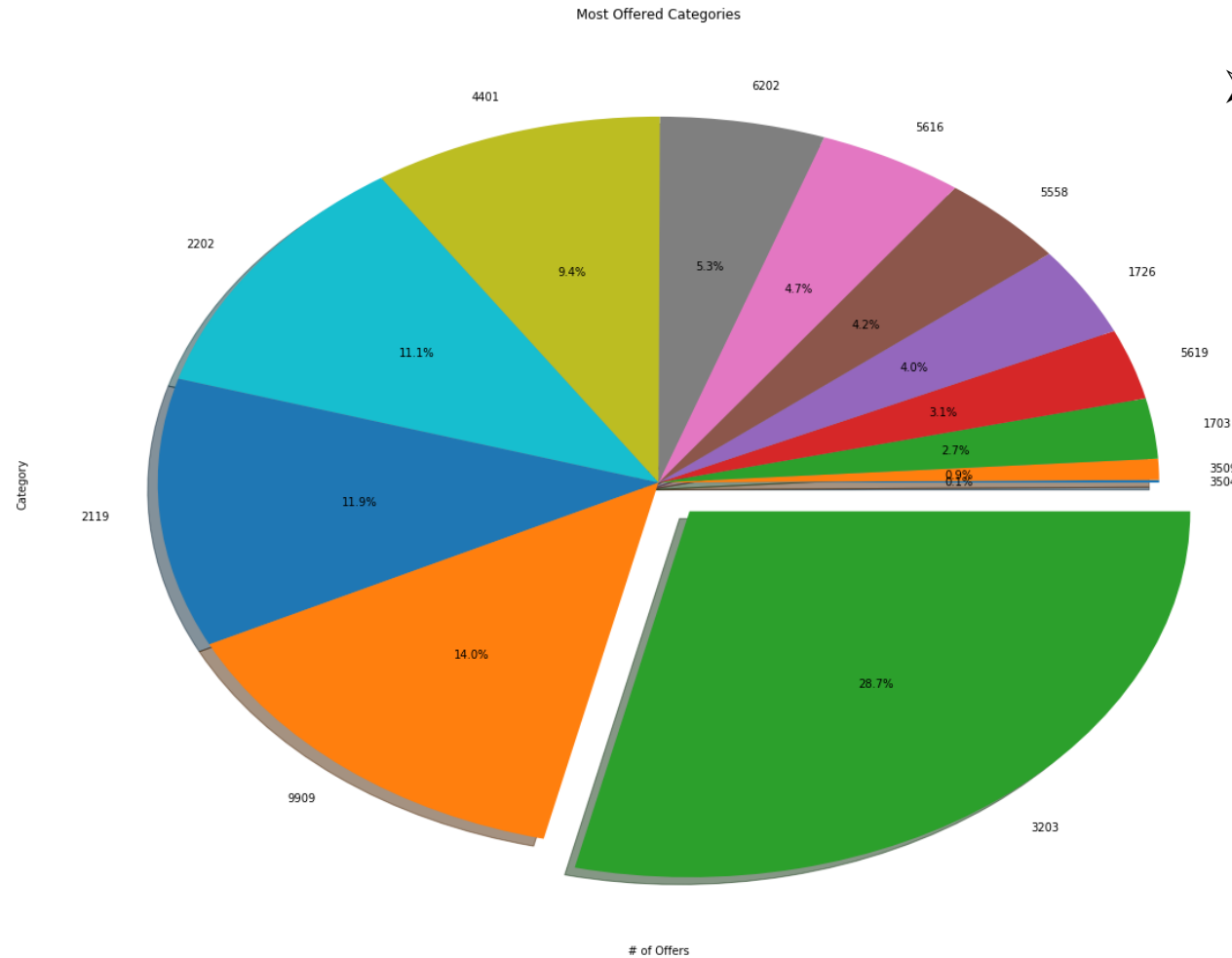
Explanatory Data Analysis



➤ Returntrip count for each category

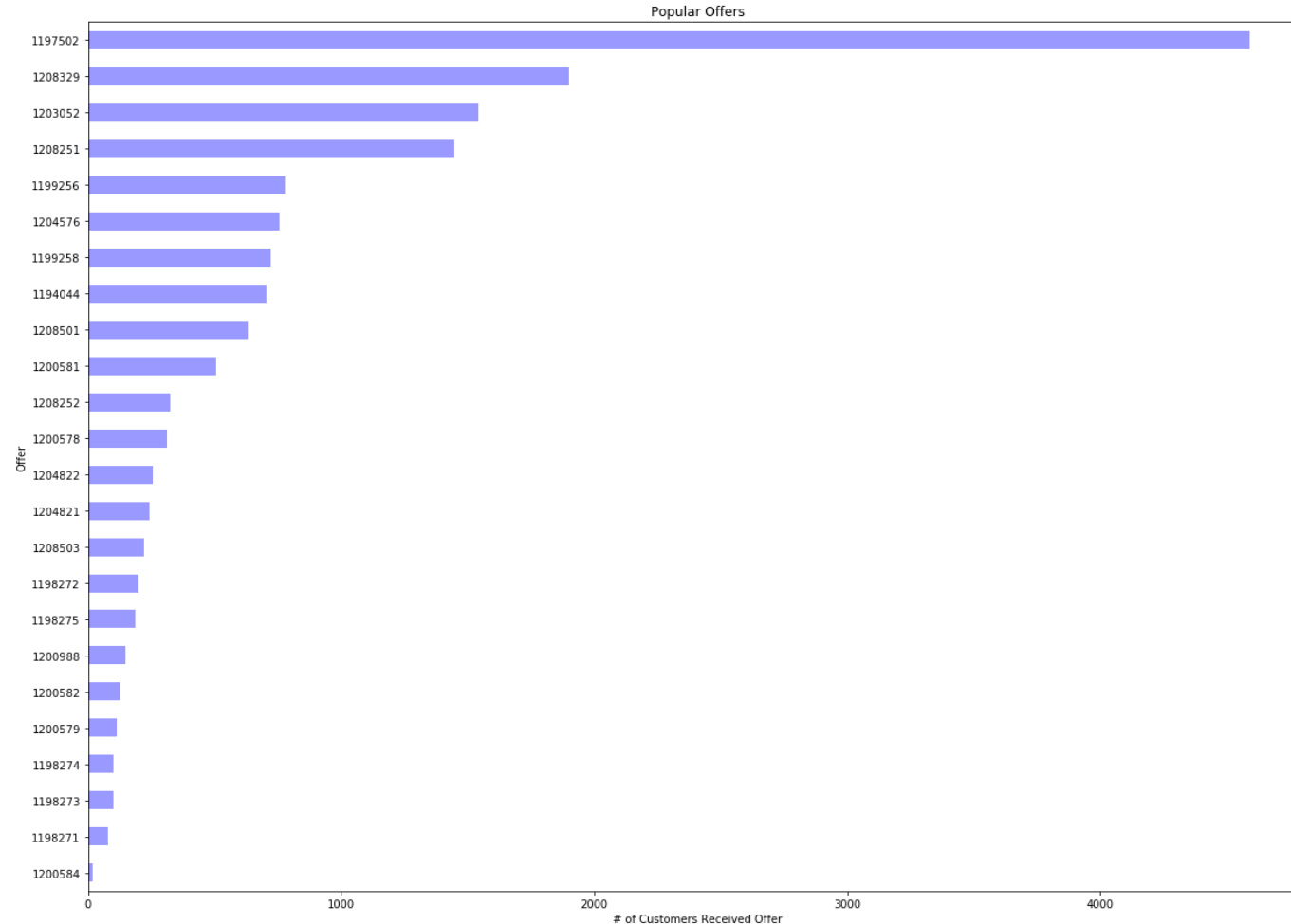
Acquire Valued Customers

Explanatory Data Analysis



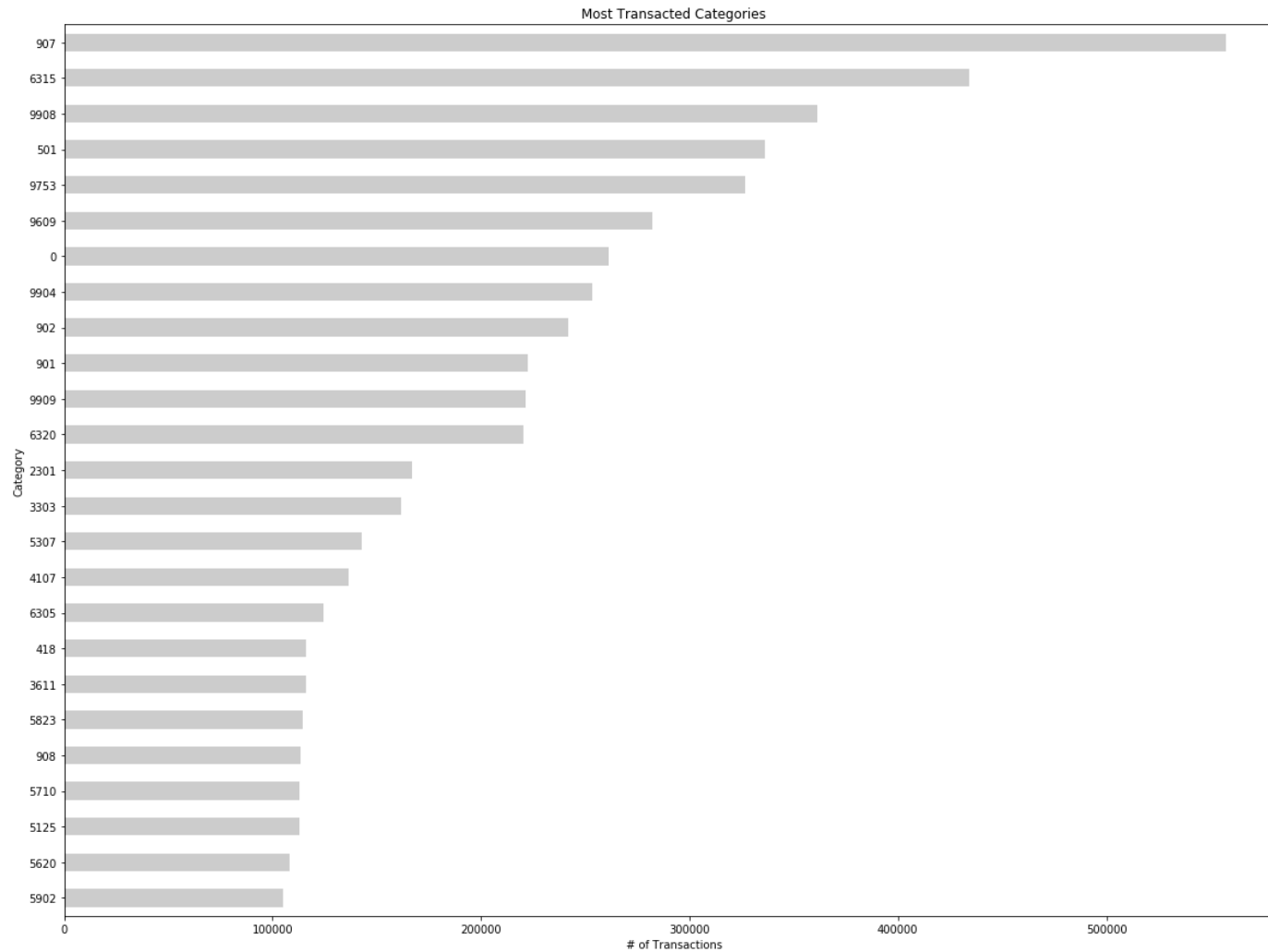
➤ Category 3203 was offered to 27 % Customers

Explanatory Data Analysis



➤ More than 4000 customers were Offered coupon 1197502

Explanatory Data Analysis



- Top 25 highly transacted Categories which could have impact on prediction

Data Aggregation

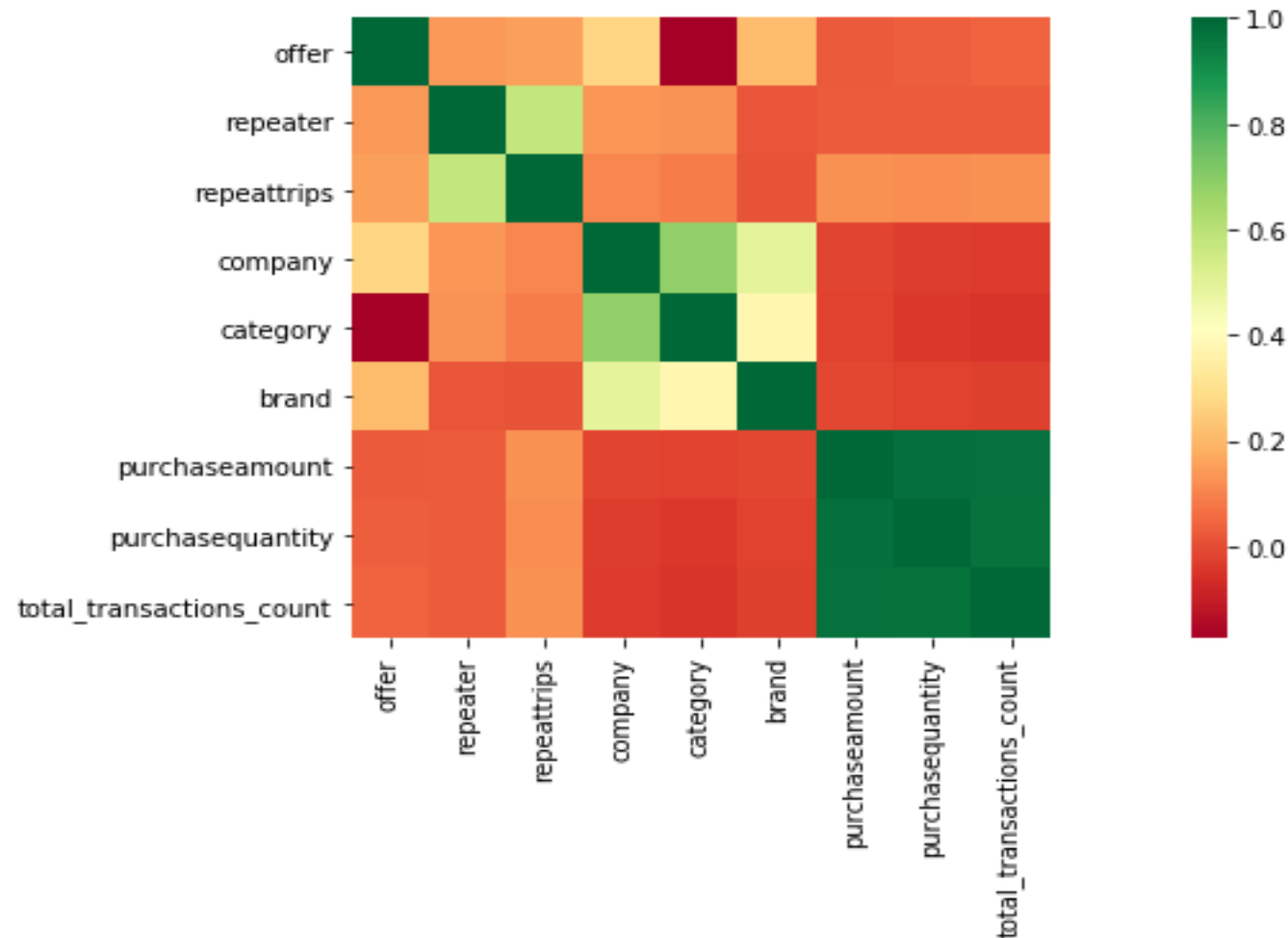
➤ Aggregate purchase amount and quantity of transactions data for each customer.

➤ Merge aggregated transactions and Hist_Offer.

```
1 trans_hist_offer.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 16000 entries, 0 to 15999  
Data columns (total 10 columns):  
id                16000 non-null int64  
offer             16000 non-null int64  
repeater          16000 non-null int64  
repeattrips       16000 non-null int64  
company           16000 non-null int64  
category          16000 non-null int64  
brand             16000 non-null int64  
purchaseamount    16000 non-null float64  
purchasequantity  16000 non-null int64  
total_transactions_count  16000 non-null int64  
dtypes: float64(1), int64(9)  
memory usage: 2.0 MB
```

Explanatory Data Analysis



➤ purchaseamount, purchasequantity and total_transaction_count is found correlated to repeattrips

Feature Extraction

- offer
repeater
repeattrips
company
category
brand
- offervalue
offeredmonth
quantity
- purchaseamount
purchasequantity
- total_trans_purchaseamount_avg
purchaseamount_category_avg
- total_transactions_count
- total_purchse_company_count
total_purchse_category_count
total_purchse_brand_count
- [1,3,6,9,12]_month_total_purchase_amt
- category_purchased_amt_[30,90,180,270]days
category_purchased_qty_[30,90,180,270]days
- company_purchased_amt_[30,90,180,270]days
company_purchased_qty_[30,90,180,270]days
- brand_purchased_amt_[30,90,180,270]days
brand_purchased_qty_[30,90,180,270]days

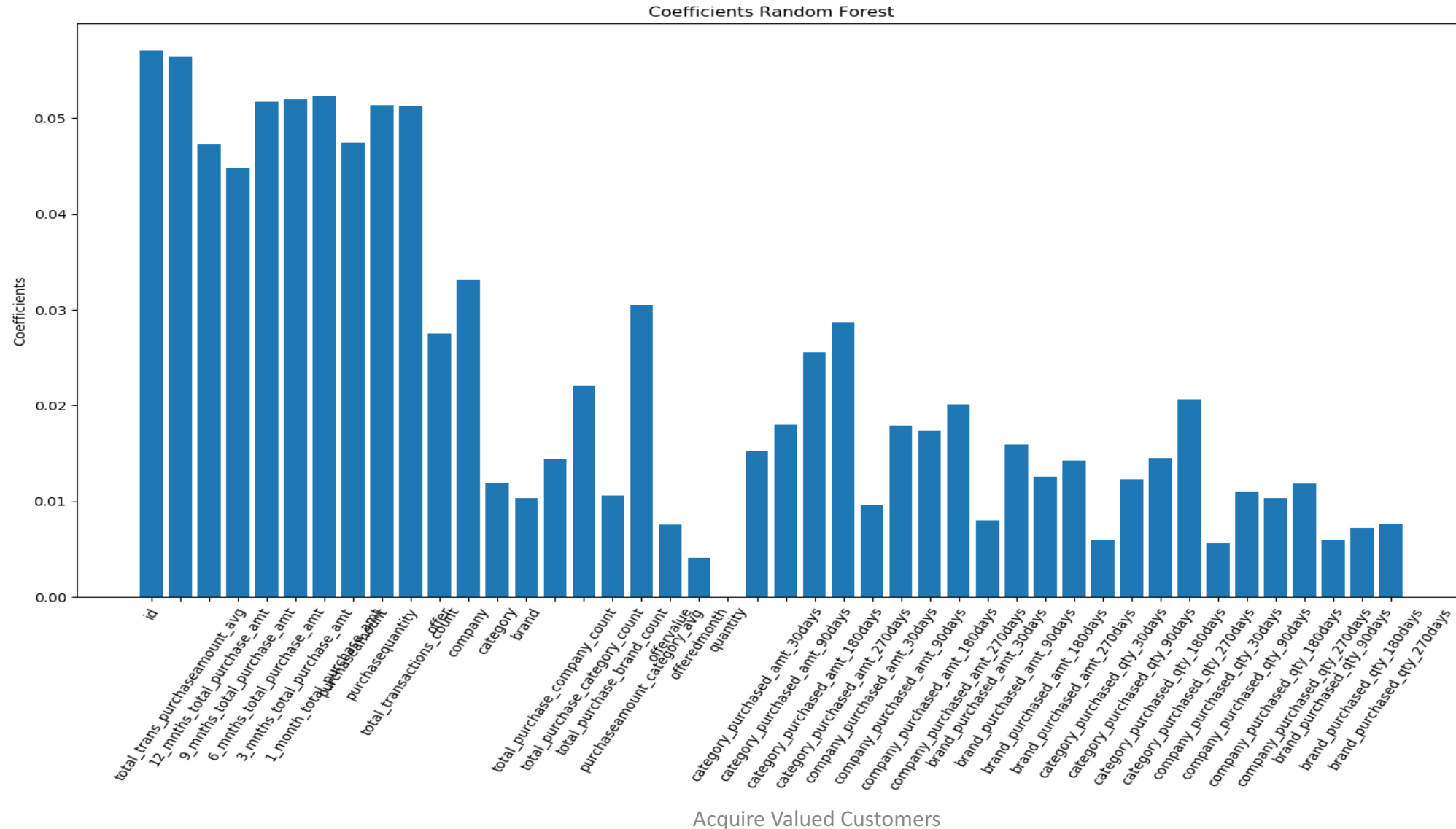
Model Evaluation & Selection

- Evaluated below 3 Classification Models

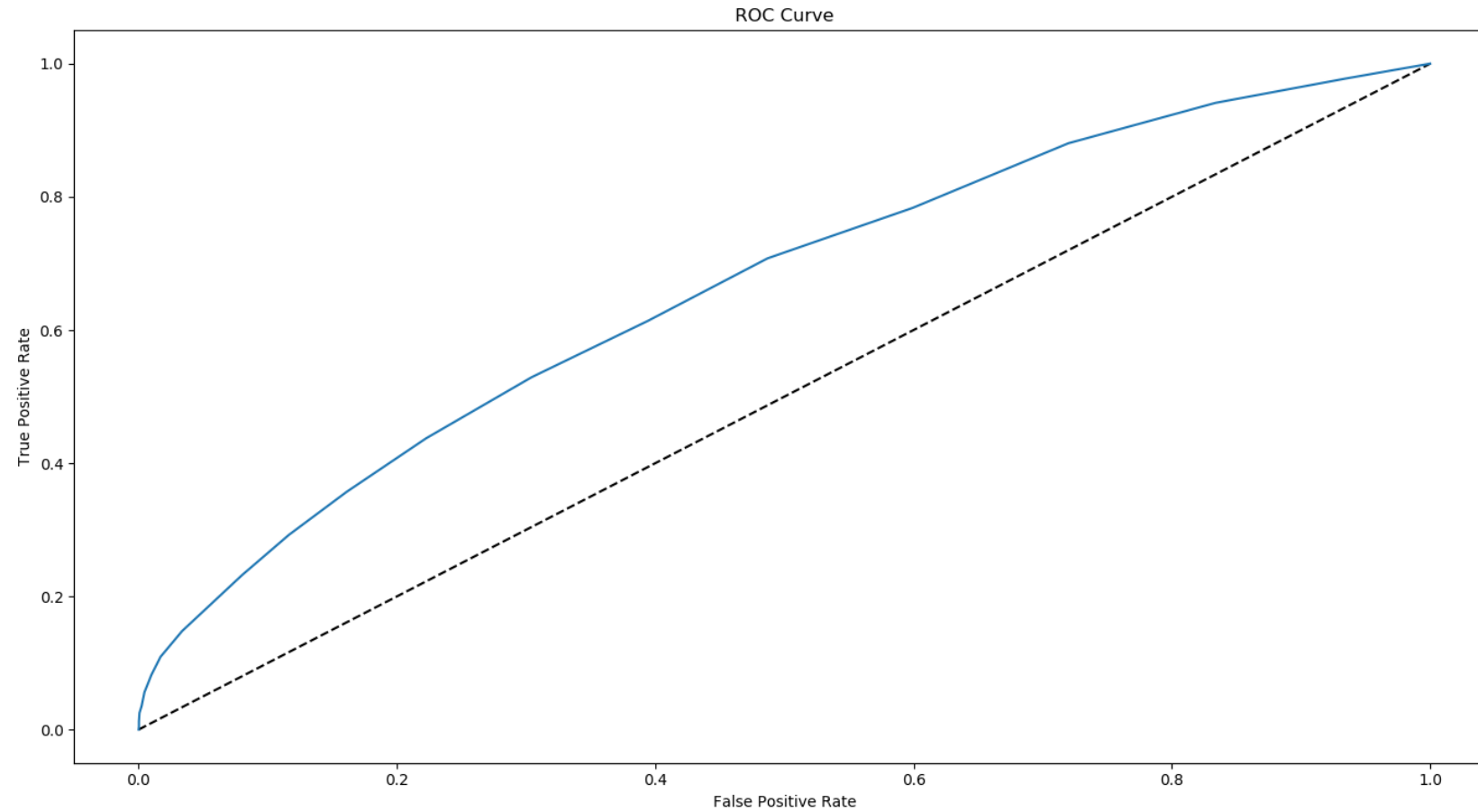
	Logistic Regression	Decision Tree	Random Forest
AUC Score	0.5225	0.5161	0.6118 ←

- Its evident from above metrics, that logistic regression is not the best model for this project
- Random Forest have produced better results for $n_estimators = 20$

Model Evaluation & Selection



Model Evaluation & Selection



Acquire Valued Customers

Recommendations to Clients

➤ Factors Influencing customers to repeat purchases

- Average customer spending
- Recent Purchase trends of customer
- History of Category, Company, Brand purchases by customer

➤ Recommendations to Companies

- Offers are effective way of improving sales and retaining customers
- Choose right Offer for products of Category, Company, Brand

Thank You.