INSTACART MARKET BASED ANALYSIS

A CASE STUDY REPORT

PRESENTED BY: PRATHAM NAWA

pratham.nawal21@gmail.com

TABLE OF CONTENTS

INSTACART MARKET BASED ANALYSIS

1. PROBLEM STATEMENT

Build a classification model which will predict which previously purchased products will be in user's next order.

2. INTRODUCTION TO PROJECT

A) CASE STUDY EXPLANATION

This report is an analysis on the behaviour of over 200,000 customers, predicting insights of their orders at Instacart Grocery Store . Every month, thousands of customers buy their products from various grocery stores of Instacart. While customers may range from first timers to the ones who frequently purchase from here . We have given datasets of to analyze the customer segments and behaviour of buying goods. The aim of the project is to predict the products which are ordered frequently by the customers so that we can work on better products, optimizing the cost on purchase of products and therefore increasing the overall sales of the Instacart using the power of Analytics.

B) PAIN AND GAIN ANALYSIS

The Pain and Gain analysis is really a tricky question here. As we are dealing with customer behaviour and goods purchased, which is a subjective opinion and not objective. As it also depends on the time of the year whether it is a festival or any special occasion and more similar factors which could also affect customer's order, reaching a fair accuracy is a pain in the neck.

Prediction of re-ordered products gives us the edge to keep more stock of those products which are ordered frequently. As marketing experts are involved in this, they can also devise good business and sale's strategies for better consumer experience and increasing profits .

Using Analytics for e-commerce sales gives us a mathematical approach for predicting products. This approach saves us lot of time and resources, it saves money and effort. But using Analytics for predicting re-orders has few shortcomings, we don't get any idea about what's currently trending product in the market and also we are not aware of the age group of the customers which majorly effects the type of products purchased.

C) DOMAIN KNOWLEDGE

Instacart, a grocery ordering and delivery app, aim is to make it easy to fill your refrigerator and pantry with your personal favourites and staples when you need them. After selecting products through the Instacart app,personal shoppers review your order and do the in-store shopping and delivery for you.

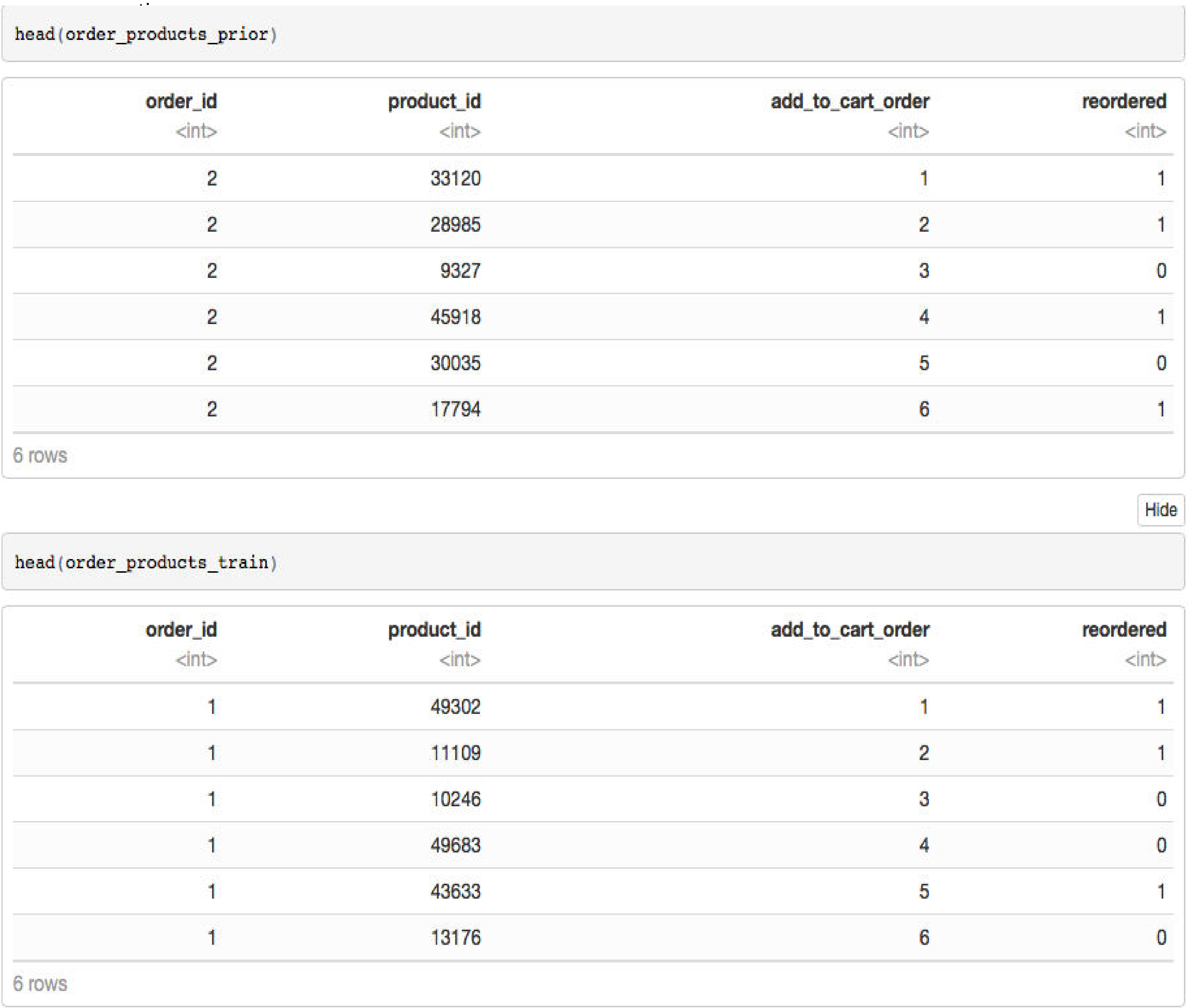
Instacart's data science team plays a big part in providing this delightful shopping experience. They use transactional datato develop models that predict which products a user will buy again, try for their first time or add to their cart again.

3. EXPLORATORY DATA ANALYSIS

We have 3 major Data sets for predicition of re-orders. We merge 2 data sets(order_products_prior and order_products_train) with the main one(orders) using product id. Orders data set contains 3421083 Observations and 7 Variables ,order_products_prior and order_products_train data sets contains (32434489,4) – Observations,variables and (1384617,4) Observations, Variablesrespectively.

A) UNDERSTANDING OF DATA

The dataset for this competition is a relational set of files describing customers' orders over



```
head(order_products_prior)
```

order_id <int>	product_id <int>	add_to_cart_order <int>	reordered <int>
2	33120	1	1
2	28985	2	1
2	9327	3	0
2	45918	4	1
2	30035	5	0
2	17794	6	1

6 rows

```
head(order_products_train)
```

order_id <int>	product_id <int>	add_to_cart_order <int>	reordered <int>
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1
1	13176	6	0

6 rows

We discuss each major variable here.

1. **order_dow**: It is denoted by a number which ranges from 0 to 7. It represents the day of the week on which the order was placed. By plotting a histogram for the same to understand customer's preferred day of order. We observed that they mostly order on day 0 or day 1. Since no tag is mentioned we can assume these days to be weekend. Ie Saturday and Sunday

2. **Order_hour_of_the_day:** It is denoted by an integer. It is a categorical variable with 24 categories. It represents at what hour of the 24 hour clock, a customer purchases his groceries or materials. It ranges from 0-24.

3. **Days_since_prior_order**: It denotes the number of days that has been passed since the day a customer has placed his/her previous order. It is denoted by an integer. It is a continuos variable where the value is 0 for the first order of any customer.

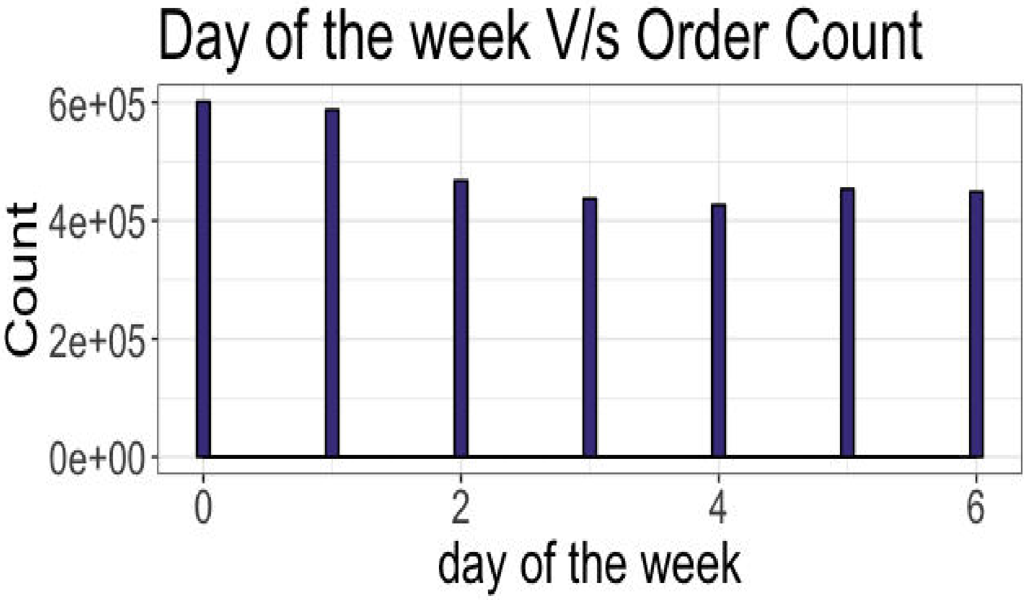
4. **Add_to_cart_order**: It denotes how many number of products were purchased by a customer during a single order. It is represented by an integer.

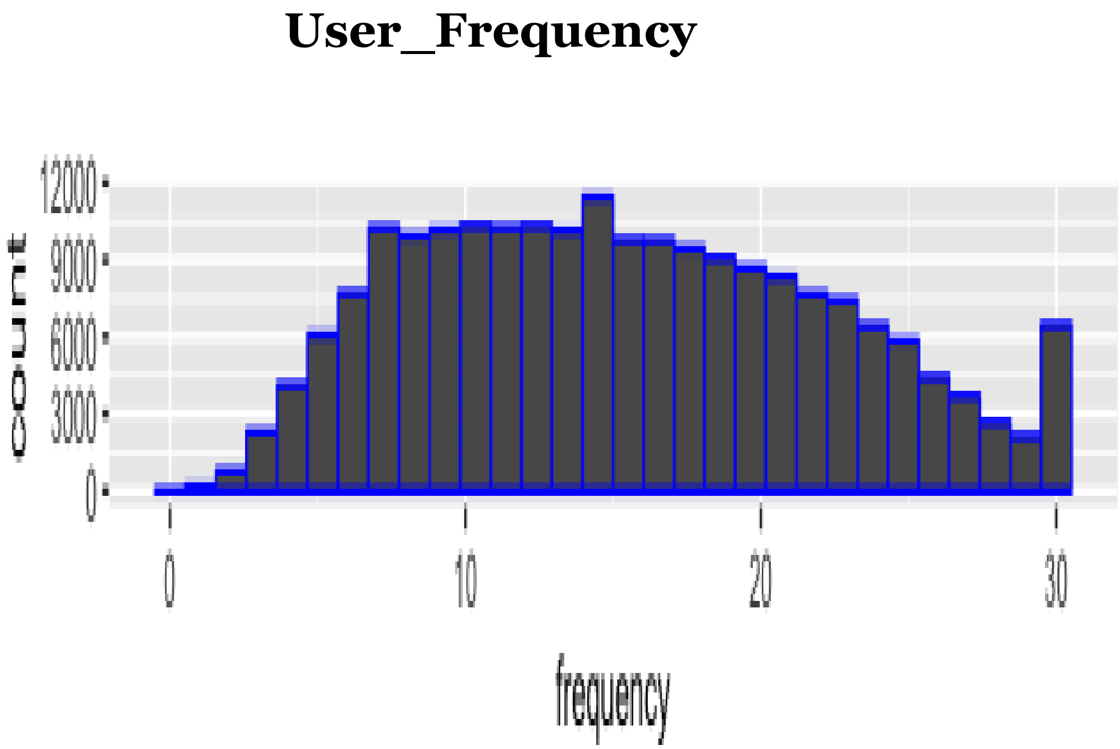
5. **Aisle :-** It refers to the product type or parent form of the product. It excludes all the brand names attached to a particular food. There are 134 different aisles in the data. Each associated with aisle id.

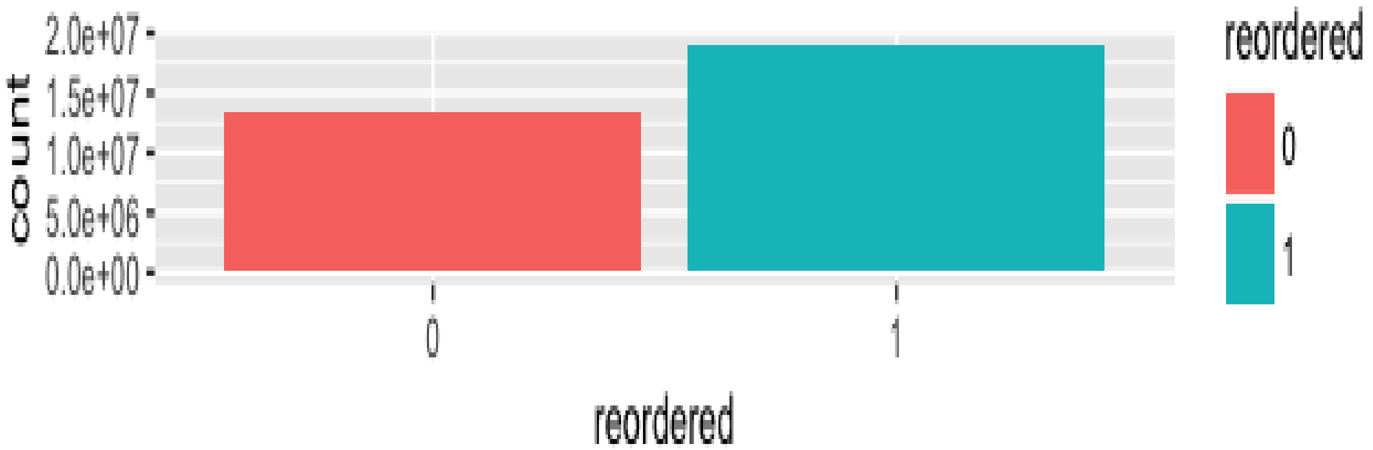
6. **Department** : It is a categorical variable. Here it denotes the department from where the food or grocery ordered by the customer belongs to. There are 21 different departments for a particular food ranging from pantry, frozen,snacks etc. Each associated with a departmental id.

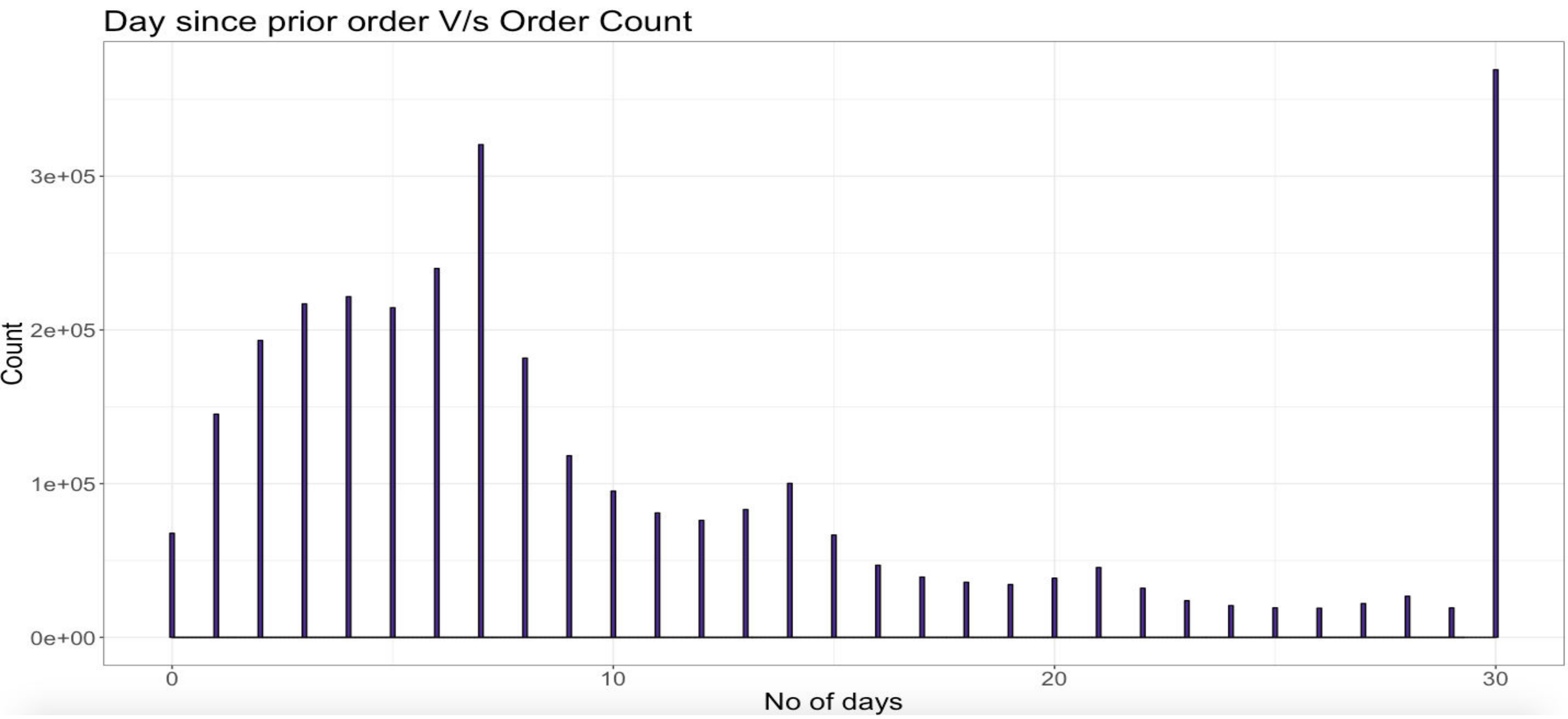
ii) UNDERSTANDING ON DATA BASIS:

To explain the Distribution of each variable, we are using Histogram and bar graphical representation:









The above Data Visualizations are prepared using ggplot2 package. Above visualizations on data distribution tells us many important things. i.e., Known Insights

KNOWN INSIGHTS OR FIRST LEVEL OF INSIGHTS:

1. Data distribution is not normal for all the variables. They all are mostly positive skewed variables.
2. From the previous data sets, nearly 59% were re-ordering some products.
3. Mostly people prefer to buy products between 8am to 9 am
4. Customers usually prefer to buy products on Saturday and Sundays.
5. Normally people re-order usally after a week or after a month(a bulk order).

6. Banana is the highest selling product.

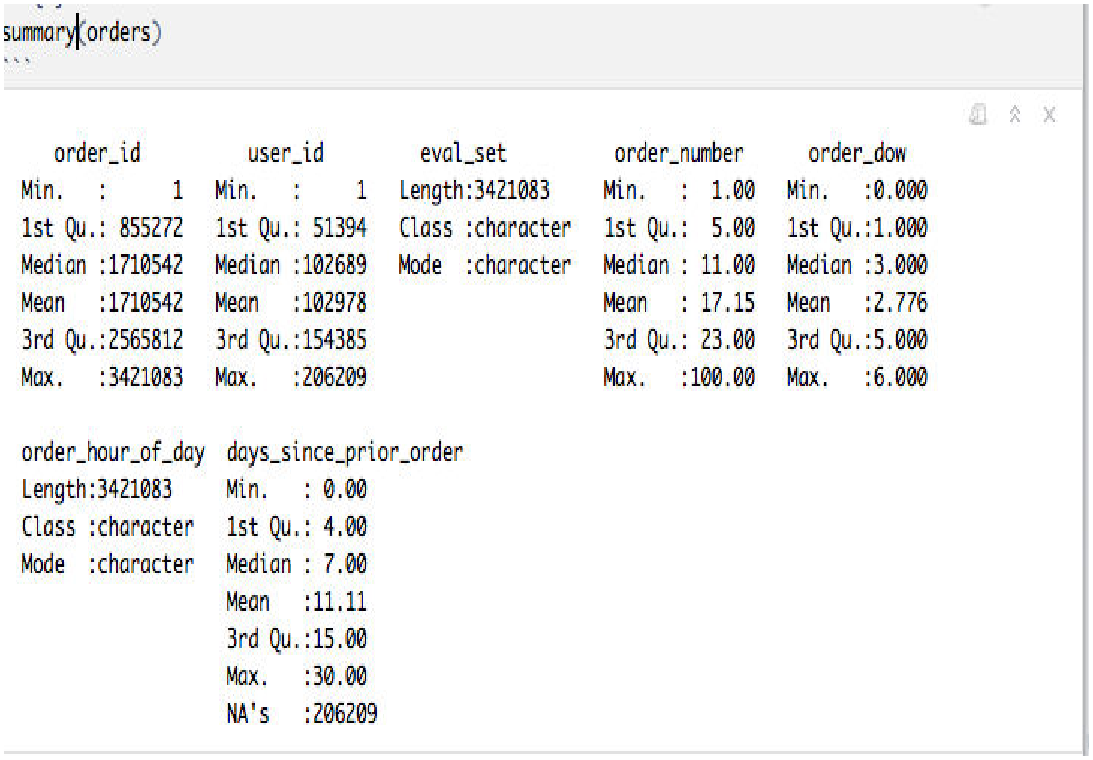
B) Preprocessing and Cleaning

A) MISSING VALUE ANALYSIS

Our Data contains which is in the variable days_since_prior_order. It was due to the fact that the first order of every customer got a NA. We have imputed all the missing values with 0.

B) Summary:- Orders

Our orders Data has many independent variables. Let us see



```
summary(orders)
```

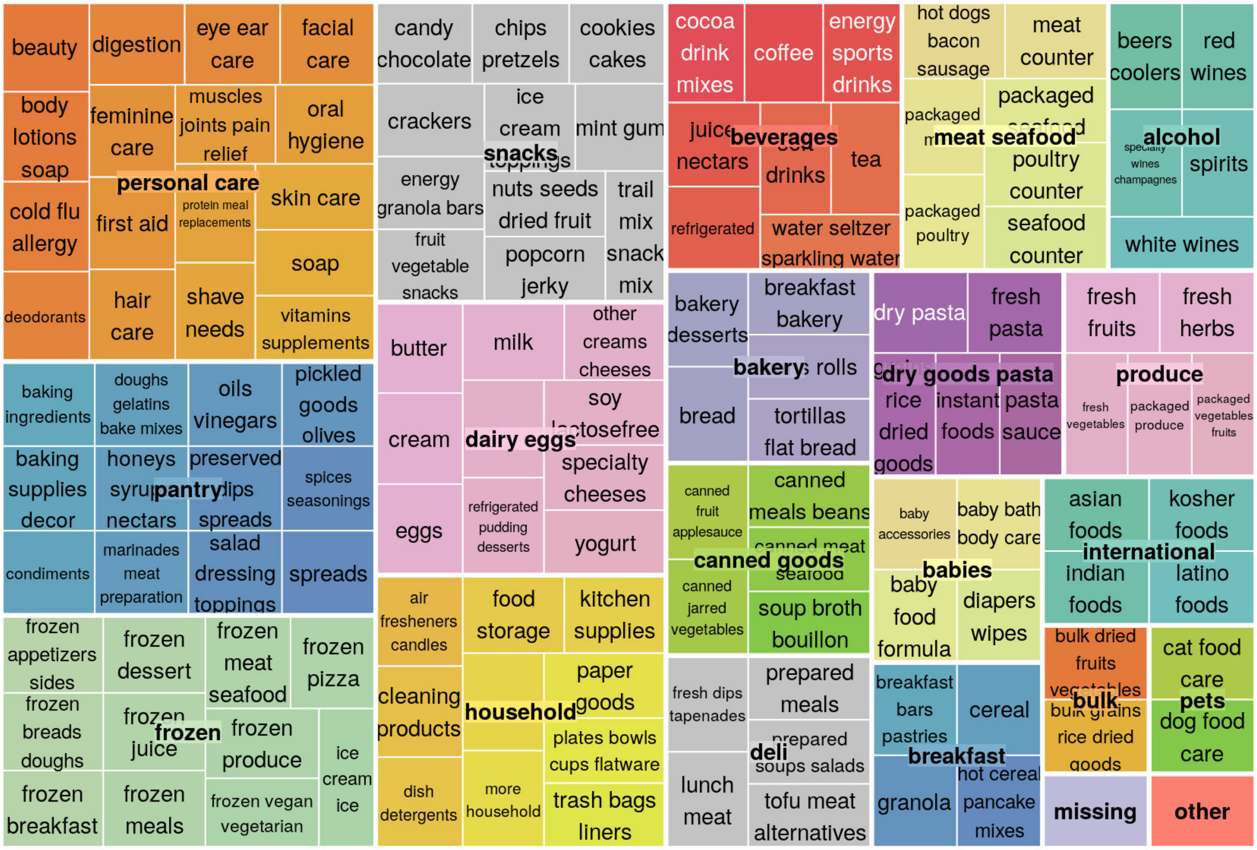
```
    order_id          user_id          eval_set          order_number       order_dow
 Min.   :      1   Min.   :     1   Length:3421083     Min.   :  1.00   Min.   :0.000
 1st Qu.: 855272   1st Qu.: 51394   Class :character   1st Qu.:  5.00   1st Qu.:1.000
 Median :1710542   Median :102689   Mode  :character   Median : 11.00   Median :3.000
 Mean   :1710542   Mean   :102978                      Mean   : 17.15   Mean   :2.776
 3rd Qu.:2565812   3rd Qu.:154385                      3rd Qu.: 23.00   3rd Qu.:5.000
 Max.   :3421083   Max.   :206209                      Max.   :100.00   Max.   :6.000

 order_hour_of_day  days_since_prior_order
 Length:3421083     Min.   : 0.00
 Class :character   1st Qu.: 4.00
 Mode  :character   Median : 7.00
                    Mean   :11.11
                    3rd Qu.:15.00
                    Max.   :30.00
                    NA's   :206209
```

Order_id	: 1- 3421083
User_id	: 1- 206209
Order_dow	: 0- 6
Hour of day	: 0- 23
Order_number	: 1- 100
Days_since_prior_	: 0- 30
order	:
Reorder_probabilit	: 0- 1

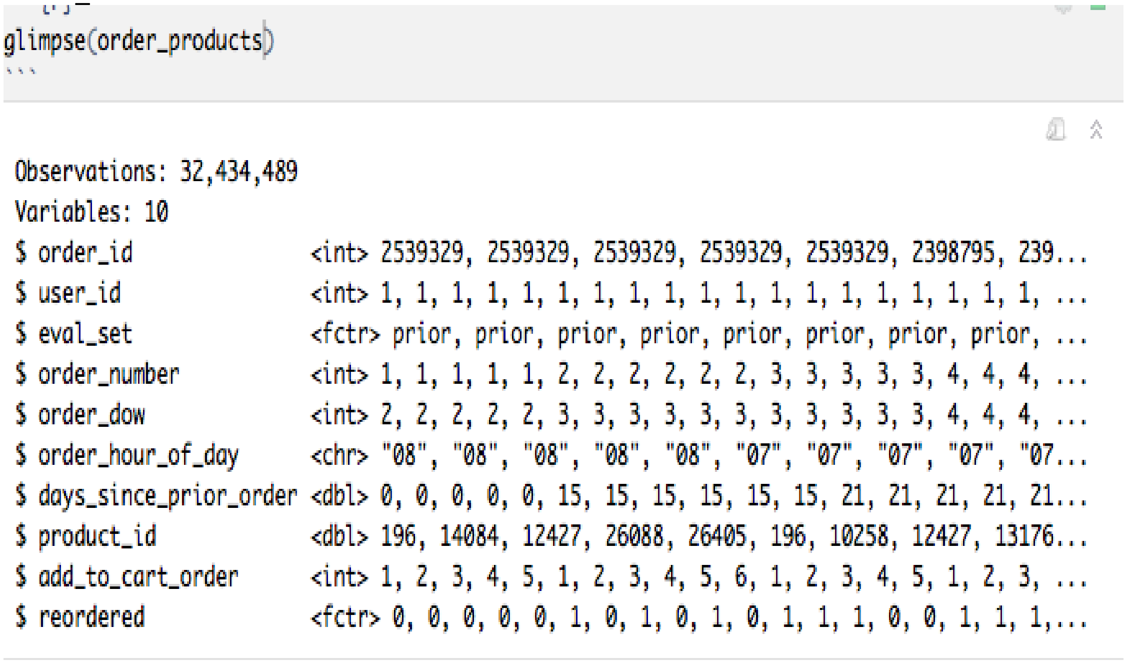
Target Variable

Re-Ordered (levels)	: 0 and 1.

Let us Analyze products sold by each department using treemap



A Data frame ‘order_products’ is created by merging orders and order_products by ’order_id’



```
glimpse(order_products)
```

```
Observations: 32,434,489
Variables: 10
$ order_id               <int> 2539329, 2539329, 2539329, 2539329, 2539329, 2398795, 239...
$ user_id                <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ eval_set               <fctr> prior, prior, prior, prior, prior, prior, prior, prior, ...
$ order_number           <int> 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, ...
$ order_dow              <int> 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, ...
$ order_hour_of_day      <chr> "08", "08", "08", "08", "08", "07", "07", "07", "07", "07...
$ days_since_prior_order <dbl> 0, 0, 0, 0, 0, 15, 15, 15, 15, 15, 15, 21, 21, 21, 21, 21...
$ product_id             <dbl> 196, 14084, 12427, 26088, 26405, 196, 10258, 12427, 13176...
$ add_to_cart_order      <int> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 1, 2, 3, ...
$ reordered              <fctr> 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1,...
```

First we will replace outliers with NAs, thereafter we will impute them.
Here, we have to consider domain knowledge and use it carefully. The domain knowledge says that the missing value in the variable days_since_prior_order is NA only for every customer because the first order of every customer has no prior order. So it goes NA by default and we have imputed it by 0 considering no prior orders.

D) VARIABLE IMPORTANCE AND FEATURE ENGINEERING

Our Master data set has 6497 Observations and 13 Variables including dependent variable (quality) and variables (type) added during analysis. Feature selection or Variable importance is a crucial step in building our model. **Not all variables carry equal information to explain our Target variable**. So, we go through feature engineering and take those variables which explains Target variable's variance more clearly.

Here we have analyzed variable importance and considering the fact that we have previous orders from our customers from data sets-order_products_prior and order_product_train. We have improved our probability of predicting the customer's re-ordering products. Thus after moulding some features from the above mentioned datasets we have introduced features like :-

- Reorderchance – Chances in terms of probability that the customer will re-order
- Usertimesreordered – no of times the customer re-ordered products
- Prodreorderchance- Chances of getting a reorder of a particular product (in terms of probability
- Prodtimesreordered- No of times a particular product got reordered.

C) MODEL BUILDING

We are building our model using Statistical Methods and Machine learning algorithms for this analysis. Logistic regression for ordinal variables, Decision tree algorithm, Random forest and KNN classification algorithms are implemented.

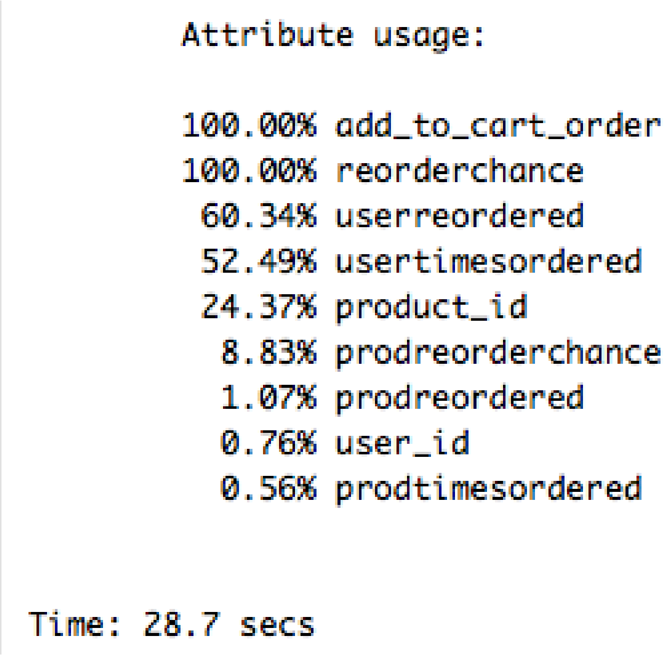
A) DECISION TREE MODEL

Decision Tree is our **Base Model**. Logistic regression model is build using 'C50' package and C50() function. C50() function is used as our target categorical variable is ordinal data type.

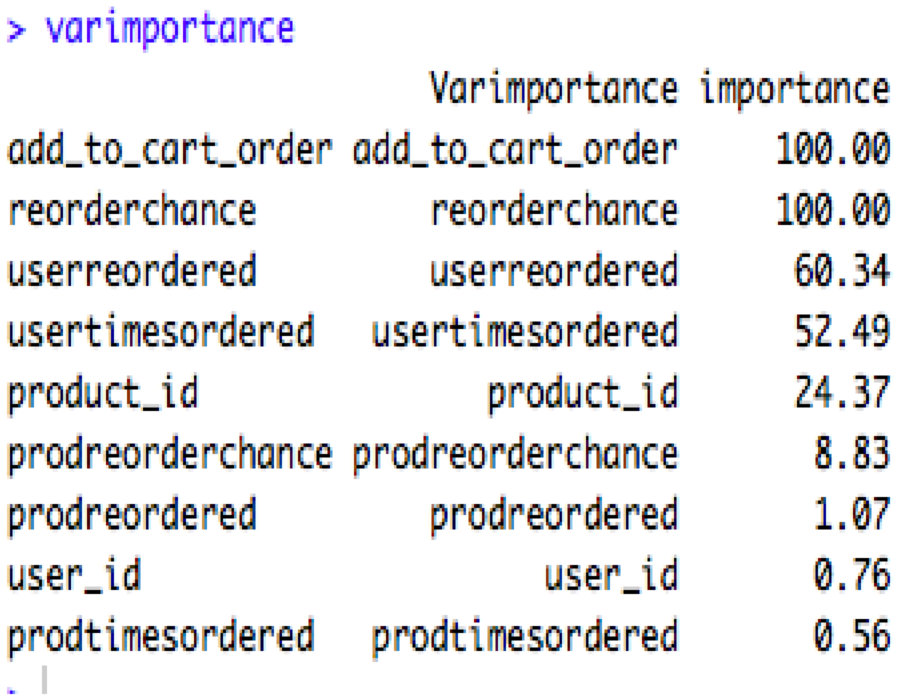
Here we are using the cart method to build our decision tree.

Decision Tree method:

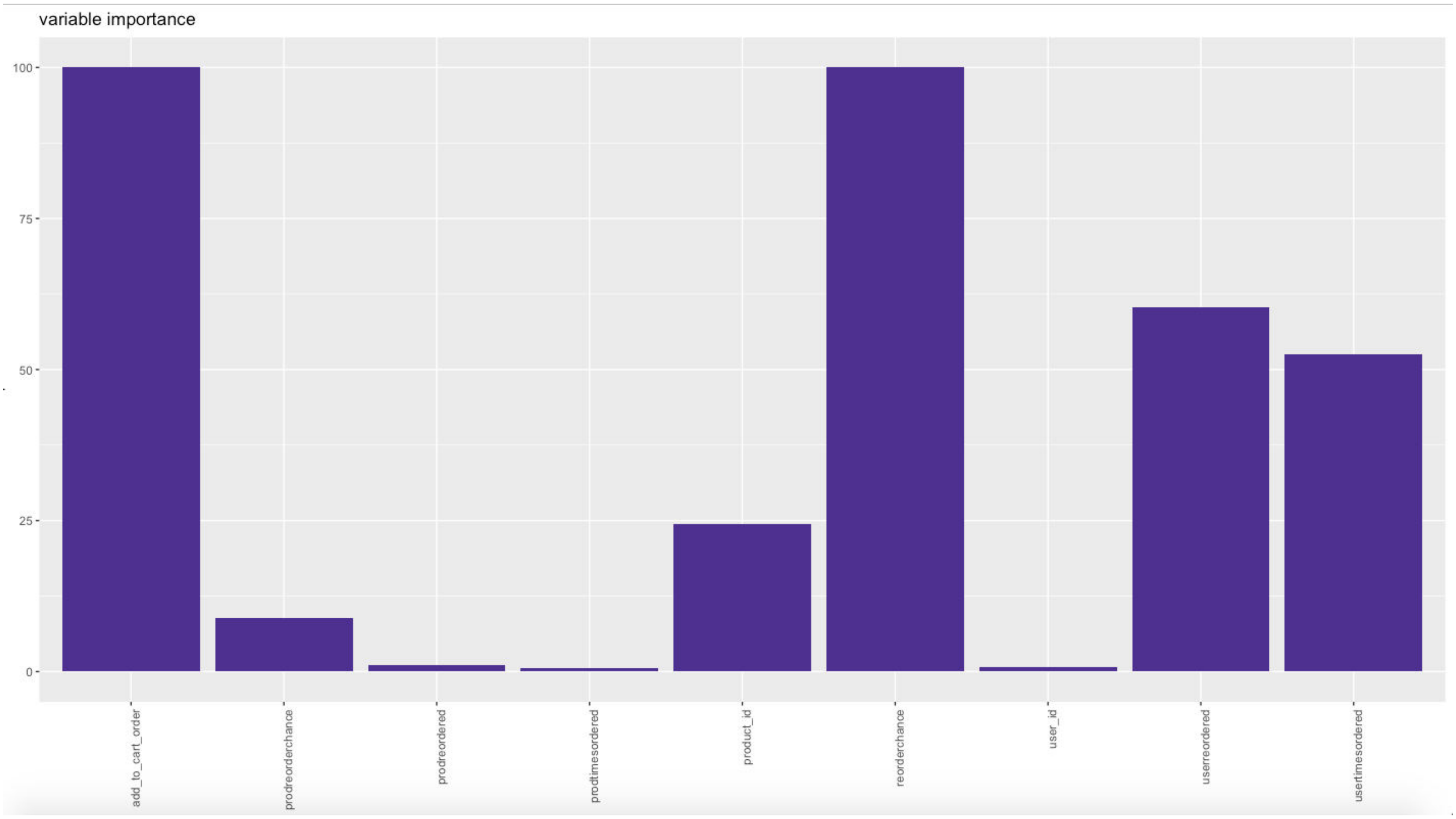
We get **0.76** F1 score with all variables and highest **70.52** accuracy with removing the least significant variables type, density and fixed acidity in both normal and normalized data sets i.e., **trainmodel**. Statistics and summary for this is given below :-



```
        Attribute usage:

        100.00% add_to_cart_order
        100.00% reorderchance
         60.34% userreordered
         52.49% usertimesordered
         24.37% product_id
          8.83% prodreorderchance
          1.07% prodreordered
          0.76% user_id
          0.56% prodtimesordered


Time: 28.7 secs
```



```
> varimportance
                     Varimportance importance
add_to_cart_order add_to_cart_order     100.00
reorderchance         reorderchance     100.00
userreordered         userreordered      60.34
usertimesordered   usertimesordered      52.49
product_id               product_id      24.37
prodreorderchance prodreorderchance       8.83
prodreordered         prodreordered       1.07
user_id                     user_id       0.76
prodtimesordered   prodtimesordered       0.56
```



ERROR METRIC

In Instacart Market Based Analysis, We are talking about predicting products which is likely to be in the user's cart in his next order. We need to look for the products which are revelant to him/her. So we need to talk in terms of precision and recall I.e not taking negative score into account . So, error metric chosen here is **F-1 score**.

So, the model which gives **least recall and highest precision** will be frozen for Deployment.

D) CONCLUSION

Instacart Market Based Analysis is a **subjective opinion which varies from customer to customer** depending on their preferences of **Customer buying's pattern & service provided by the store**. It also depends on many influencing factors like time of the day, day of the week and many other factors . Yet, this is analysis to show the **Power of Analytics**. There are many other influencing factors involved, with the given data, we have built our model to its peak many techniques right from variable importance based on probabilities and re-ordering chance.

Thank you!