Machine Learning (regression) using Python

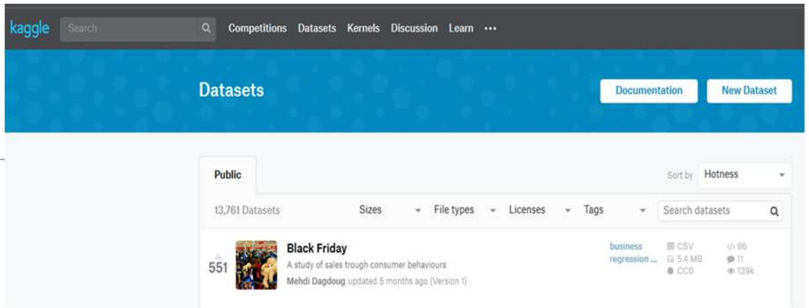
MEHUL KALIA

Objectives

- Apply knowledge learnt in AP Statistics course on a real world dataset
- Learn coding statistical diagnostics in a programming language
- Visualize data to explore and apply statistical analysis framework

About the dataset

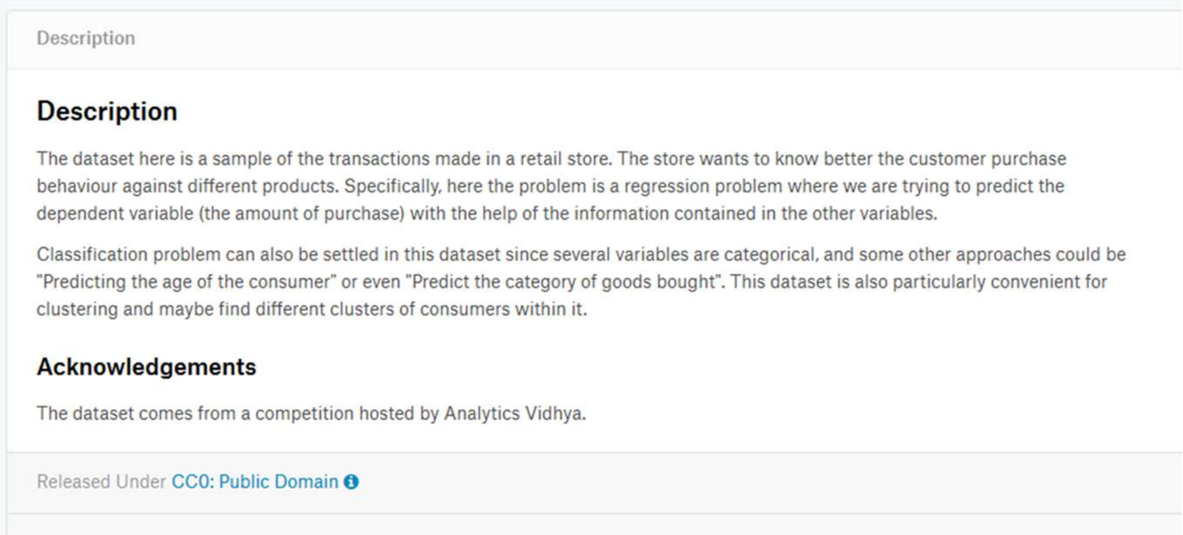
- Data: Black Friday retail store transactions dataset uploaded on Kaggle.com
- Objective : Predict purchase amount based on other given variables.



Datasets

13,761 Datasets

551 **Black Friday**
A study of sales trough consumer behaviours
Mehdi Dagdoug updated 5 months ago (Version 1)
business regression ... CSV 5.4 MB CC0 96 11 139k



Description

Description

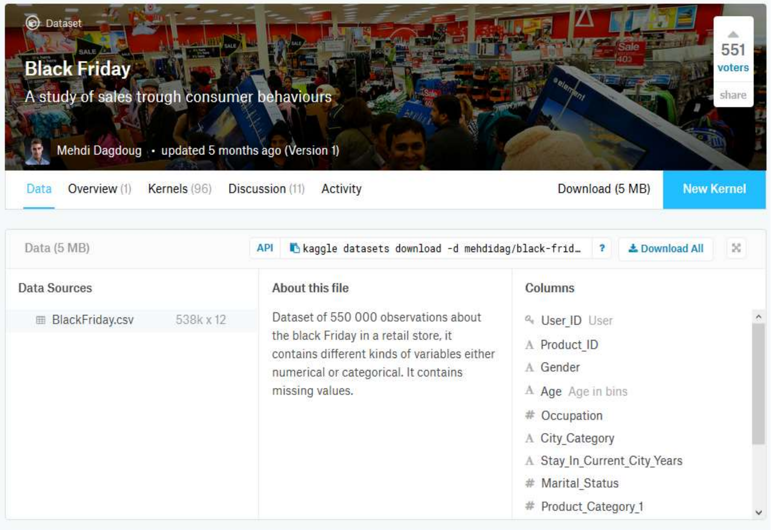
The dataset here is a sample of the transactions made in a retail store. The store wants to know better the customer purchase behaviour against different products. Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

Classification problem can also be settled in this dataset since several variables are categorical, and some other approaches could be "Predicting the age of the consumer" or even "Predict the category of goods bought". This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it.

Acknowledgements

The dataset comes from a competition hosted by Analytics Vidhya.

Released Under CC0: Public Domain



Black Friday

A study of sales trough consumer behaviours

Mehdi Dagdoug • updated 5 months ago (Version 1)

Data (5 MB)

```
kaggle datasets download -d mehdidag/black-frid…
```

Data Sources

BlackFriday.csv 538k x 12

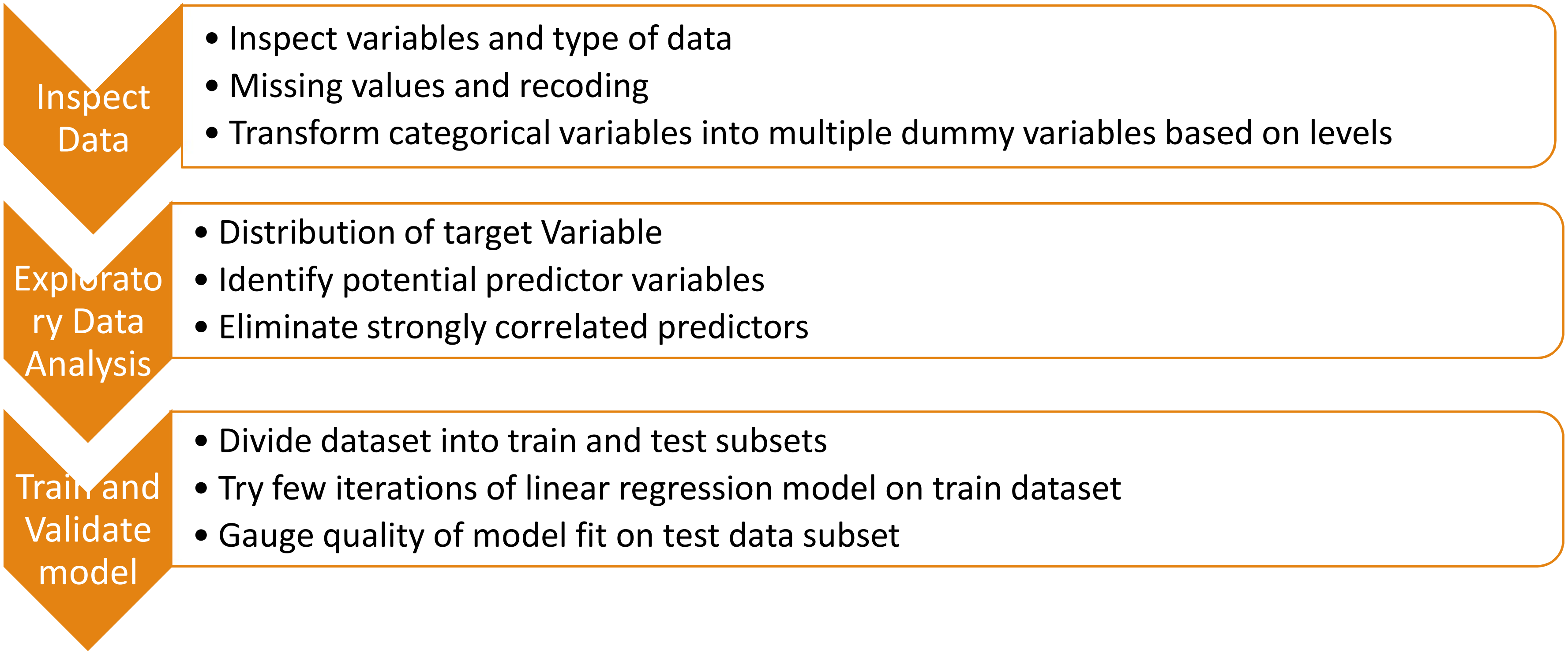
About this file

Dataset of 550 000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. It contains missing values.

Columns

- User_ID User
- Product_ID
- Gender
- Age Age in bins
- Occupation
- City_Category
- Stay_In_Current_City_Years
- Marital_Status
- Product_Category_1

Framework



Inspect Data

- Inspect variables and type of data
- Missing values and recoding
- Transform categorical variables into multiple dummy variables based on levels

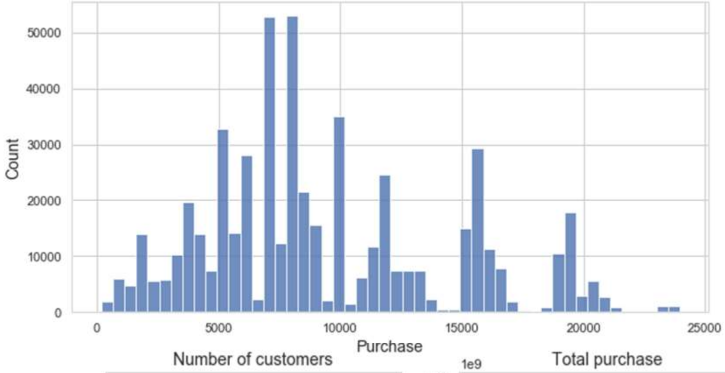
Exploratory Data Analysis

- Distribution of target Variable
- Identify potential predictor variables
- Eliminate strongly correlated predictors

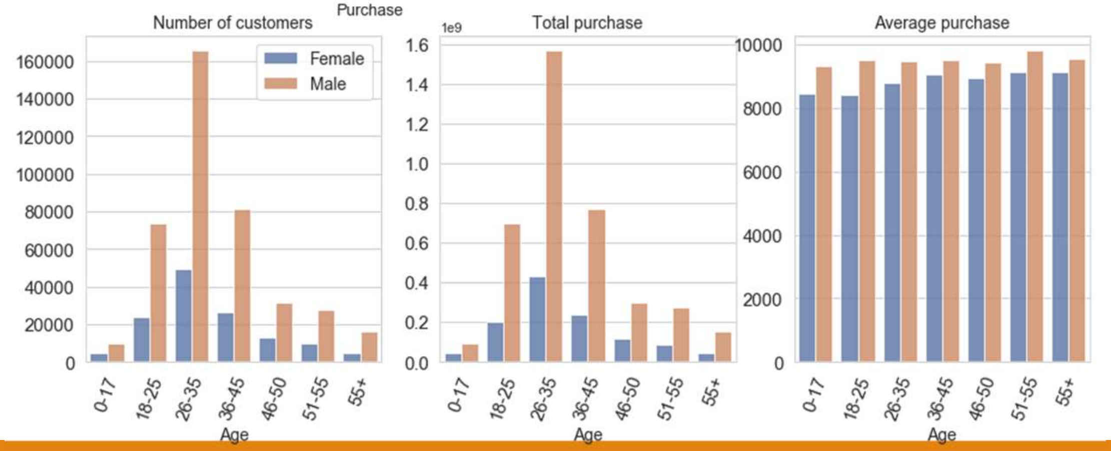
Train and Validate model

- Divide dataset into train and test subsets
- Try few iterations of linear regression model on train dataset
- Gauge quality of model fit on test data subset

Exploratory Data Analysis (EDA)- Purchase Amount, Gender, and Age

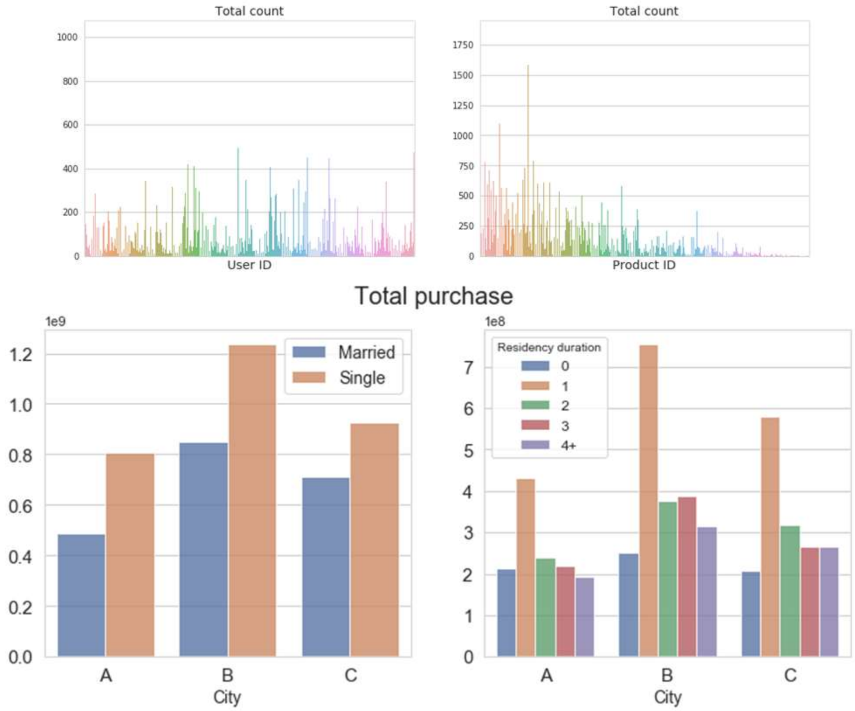


Target Variable : Purchase amount appears to have a normal distribution with right skew



Males are spending more than females, but on average all age groups spend around the same.

Exploratory Data Analysis (EDA)- Customers, City, and Relationship Status

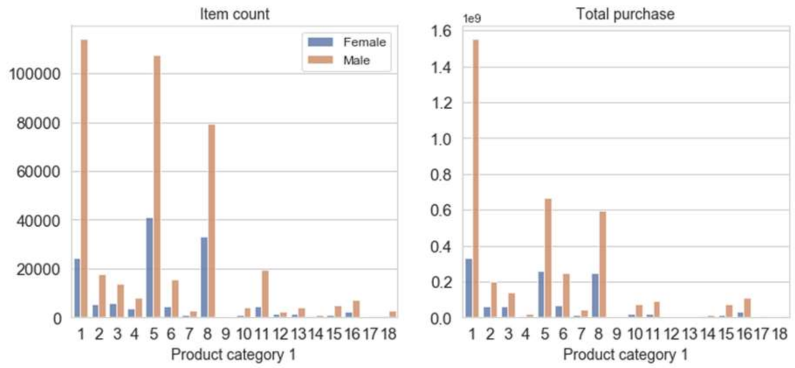


Customers and products are spread out, which is helpful for regression.

Singles are spending more than married. Customers who lived in their city for 1 year are spending more than other groups. City B inhabitants spend the most.

Variables like city and product category are encoded by letters and numbers.

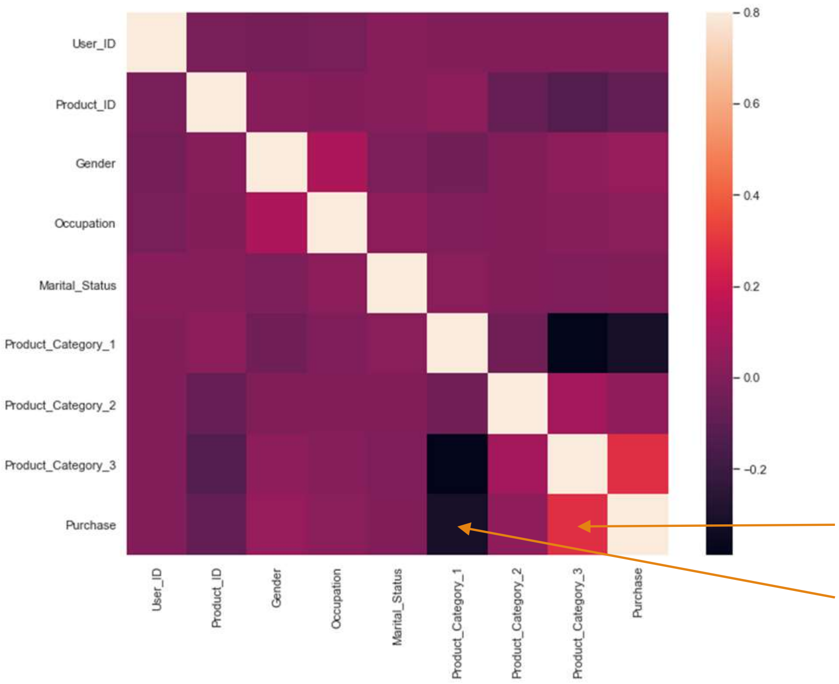
Exploratory Data Analysis (EDA)-Product Category



Product_category_1 #1,5,8 sold the most items while Product_category_1 #1 had highest revenue

Product category is hierarchical with product category 1 above product category 2, which is above product category 3.

Exploratory Data Analysis (EDA)- Correlation Between Variables



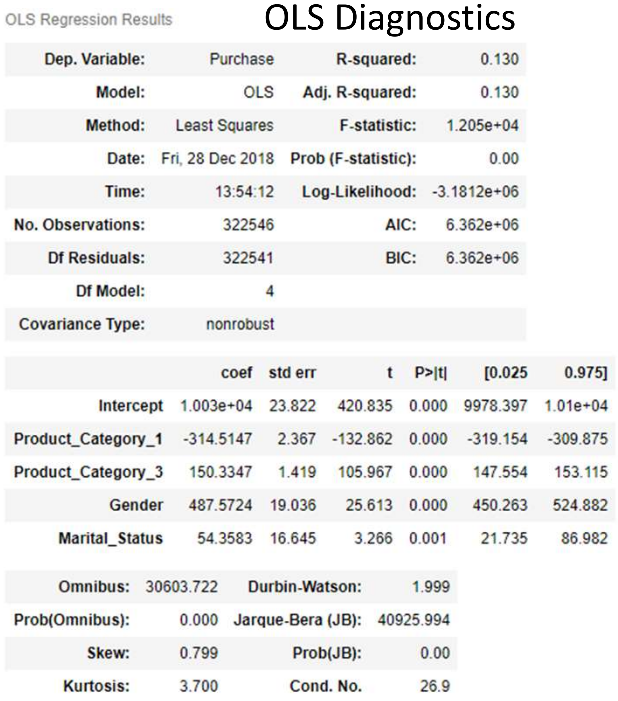
This is a correlation matrix, which shows how correlated variables are. Extremely light or dark squares show that the variables on the row and column of that square have high correlation and should not be put in a regression equation together.

No predictor variables are too correlated with each other.

Product_Category_1 and Product_Category_3 appear to have strong correlation with purchase amount.

Model Diagnostics

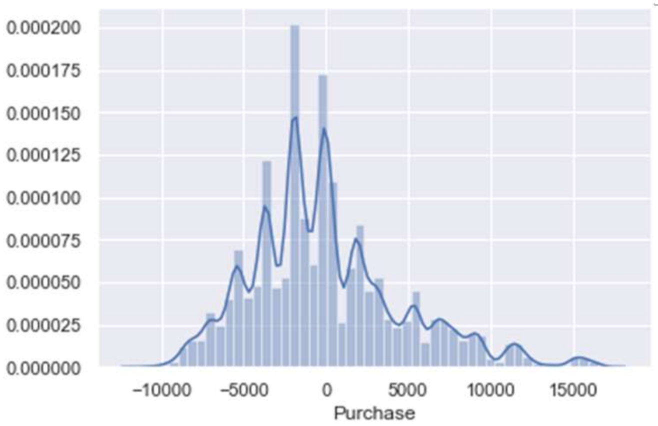
OLS Diagnostics



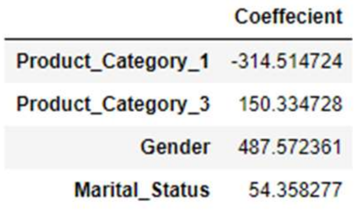
```
OLS Regression Results

   Dep. Variable:        Purchase          R-squared:        0.130
           Model:             OLS      Adj. R-squared:        0.130
          Method:   Least Squares         F-statistic:    1.205e+04
            Date: Fri, 28 Dec 2018  Prob (F-statistic):        0.00
            Time:        13:54:12     Log-Likelihood:  -3.1812e+06
No. Observations:          322546                AIC:    6.362e+06
    Df Residuals:          322541                BIC:    6.362e+06
        Df Model:               4
 Covariance Type:       nonrobust

                         coef   std err          t   P>|t|     [0.025     0.975]
          Intercept  1.003e+04    23.822    420.835   0.000   9978.397   1.01e+04
 Product_Category_1  -314.5147     2.367   -132.862   0.000   -319.154   -309.875
 Product_Category_3   150.3347     1.419    105.967   0.000    147.554    153.115
             Gender   487.5724    19.036     25.613   0.000    450.263    524.882
     Marital_Status    54.3583    16.645      3.266   0.001     21.735     86.982

      Omnibus: 30603.722    Durbin-Watson:      1.999
Prob(Omnibus):     0.000  Jarque-Bera (JB):  40925.994
         Skew:     0.799         Prob(JB):       0.00
     Kurtosis:     3.700         Cond. No.       26.9
```

Adjusted R Square of 13% is pointing not to a strong fit.

Residuals



Residuals spread is not strongly normal distributed.



	Coeffecient
Product_Category_1	-314.514724
Product_Category_3	150.334728
Gender	487.572361
Marital_Status	54.358277

Strongest predictor variables are : Gender, Product Category 1, Product Category 3, and Marital Status

Takeaways

The prediction model fit (based upon adj. R sq) has room for improvement based upon traditional statistical techniques such as linear regression. Also, this real life data is not picture perfect as textbook data

Nevertheless, I have learned how to come up with a framework to analyze data, employ analytical tools, and examine the results.

In the future, newer techniques in machine learning could have alternative approaches to problem solving and data sets.