**Capstone project 1: Women’s shoe price analysis**

**Milestone report**

**Dataset:** This dataset is from Kaggle’s competitions a sample list of 10,000 women's shoes and their product information provided by Datafiniti's Product Database. Dataset includes 33801 observations and 52 features. Some of those are shoe name, brand, price, and more. That includes categorical, numerical and Boolean values.

**Objective:**

The project goal is to analyses the pricing strategies of women’s shoes and how the price distribution varies across brands. Is there specific product features influence the pricing.

**Questions:**

What prices and brands sell more?

Which product features increases the value?

What is the average price of each distinct brand listed?

Which brands have the highest prices? Which ones have the widest distribution of prices? Is there a typical price distribution across brands?

**Environment**

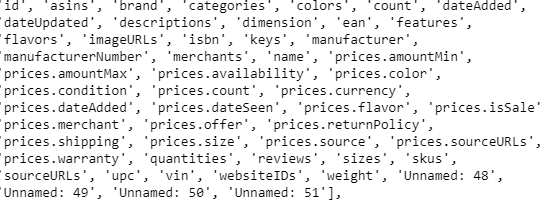
Python- Jupyter notebook

**Data acquisition**

The dataset is a csv file acquired from Kaggles competition Women's shoe prices.This is a list of 10,000 women's shoes and their product information provided by Datafiniti's Product Database.This is a sample of a large dataset available through Datafiniti. We use pandas pd.read\_csv() to load data.

**Exploratory data analysis of shoe price data**

Column names in dataset:

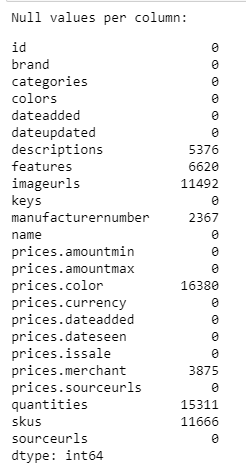


**Data cleaning and preprocessing**

Data cleaning includes changing the column names and values to lowercase, eliminating columns that contains more than 50% of null values. Also created a new column prices.amountavg using the values of minimum and maximum price values. After removing the columns with null values over 50% new dataset contains 24 columns only.

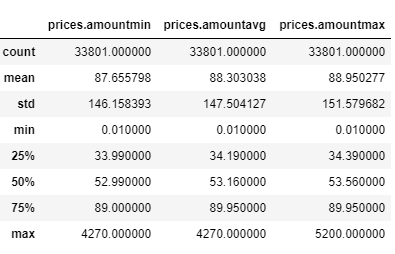
**Dealing the null values in data**

Even after removing columns contains null values using the threshold 0.5 some columns in our data set still contains null values. For further analysis we change the null values in the categorical columns brand and colors to no brand and no color using fillna. The price features in our dataset doesn’t have any null values.



**Data preprocessing**

Created a new column prices.amountavg using values of minimum price and maximum price. The prices min, max and avg is not showing much significant difference in our dataset.



Customers in our data purchased from different countries and paid different currencies. In our analysis we need data of customers who paid by US dollar. Created a new data frame with prices.currency=USD.

The new data frame has 32680 values and 25 features.

Our column names and categorical values contains both lowercase and uppercase letters. To avoid further confusions changed the column names to lowercase and all categorical values to lower case.

Exploring the brand feature shows same brand names are grouped differently in our data frame because of upper case and lower case letters in the values. e.g. A2 BY AEROSOLES 2, A2 By Aerosoles 6, A2 by Aerosoles 188

We are going to convert it to all lowercase letters. To converting the dataset first step is to create a data frame with only datatypes objects. Then change all the string datatype to lowercase using lamda()function. Create another data frame contains the number and float datatypes. Third data frame for binary values and combine all the data frames using pd.concat().

### New columns

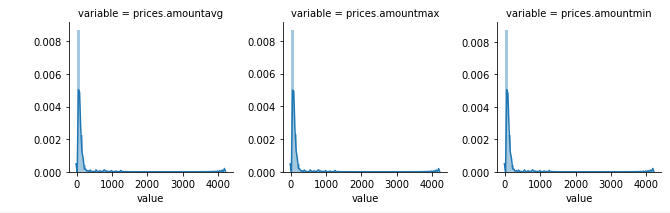
For more detailed analysis we created 3 more columns.

Column 'Price.stat' contains boolean value 'True' if the shoe price is over 100 or 'False' if the price is under 100

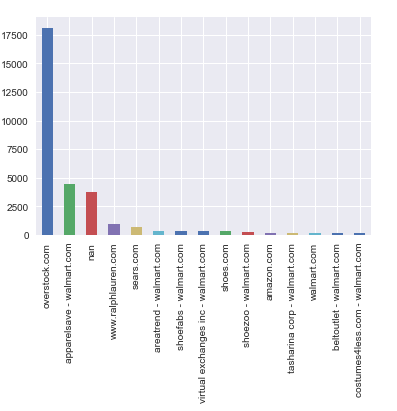
Column 'brand.count' is the number of shoes that brand sold

Column 'price\_band' has values 0,1,2. The price of the shoe is <=30 column value 0, >30 and <=100 column value 1 and >100 column value 2.

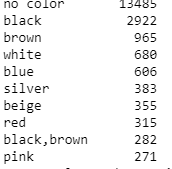
**Univariate analysis of numerical features**



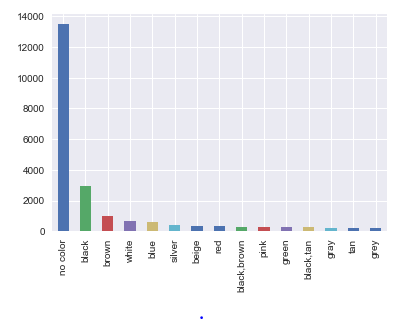
**Merchant plot**



**Popular colors**

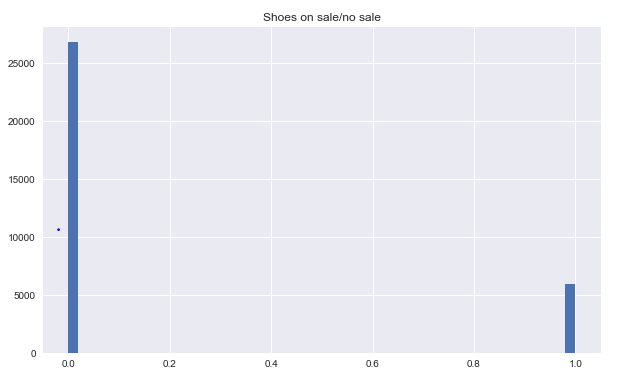


Black and brown are most popular colors.



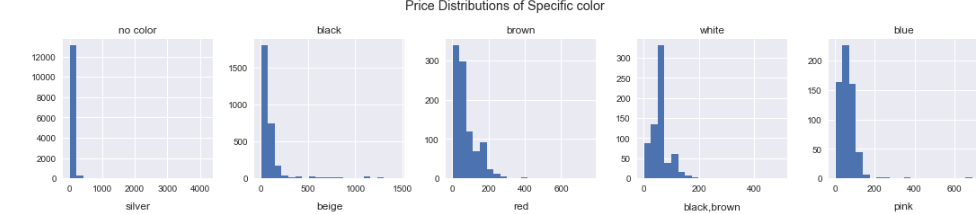
**Histogram of shoes on sale**

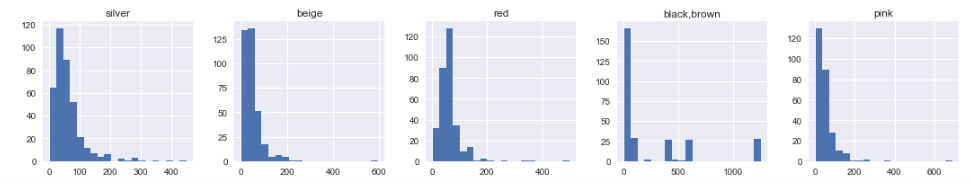
26766 shoes not on sale and 5794 on sale.

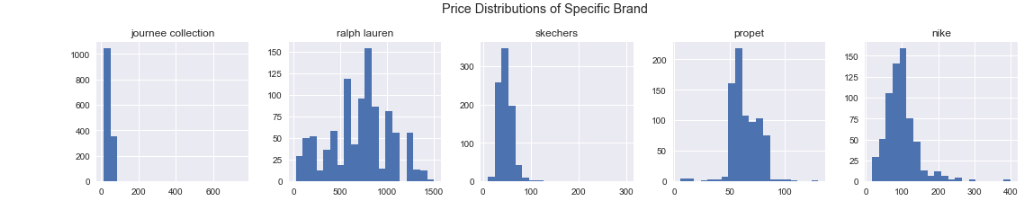


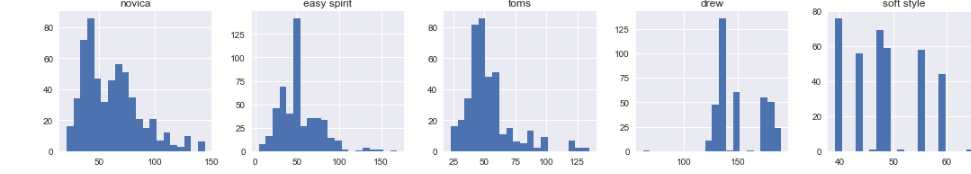
**Bivariate analysis**



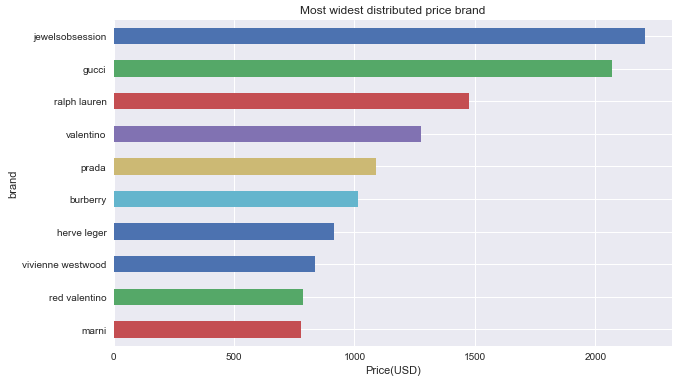








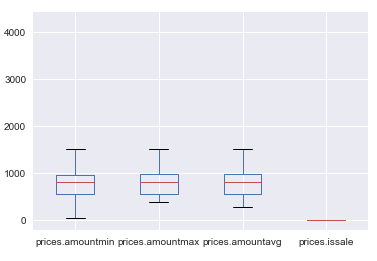
**Prices across brands not showing normal distribution.**



Checking the most expensive shoes in the data shows ,some of them are not shoes, in fact they are diamond wedding bands.

Looks like the average shoe price is 88.24, median is 51.99. The price distribution chart on popular brands shows their most selling items are priced under dollar 150.

**Boxplot of outliers**



### Label encoding

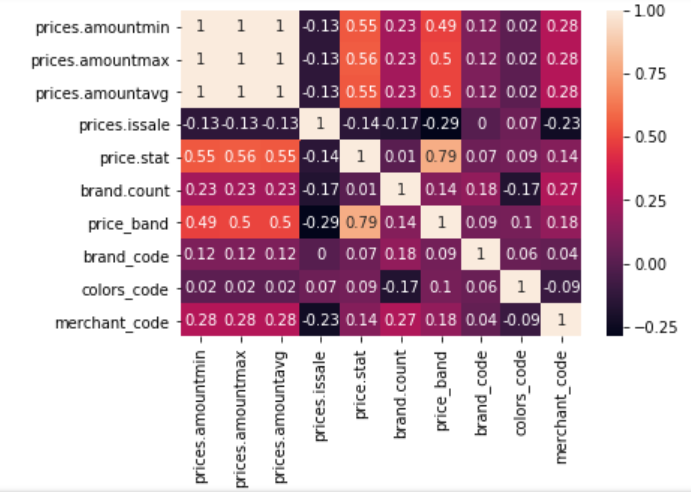
The next step in data preparation is label encoding. Our data set contain many categorical variables and the values are stored in text format. To convert these values text to numbers we use a technique called label encoding. Here we change the 'brand','colors','prices.merchant' columns values from words to numbers.

The encoded values are stored in to column names 'brand\_code','colors\_code','merchant\_code'.

Also changed the boolean values to 0 and 1 by simply changing their type to integer.

**Correlation matrix of data**

Next we creat the correlation matrix to shows the correlation between variables in the data set. The correlation matrix not showing any strong correlation between the variables in our data.



### Removing outliers

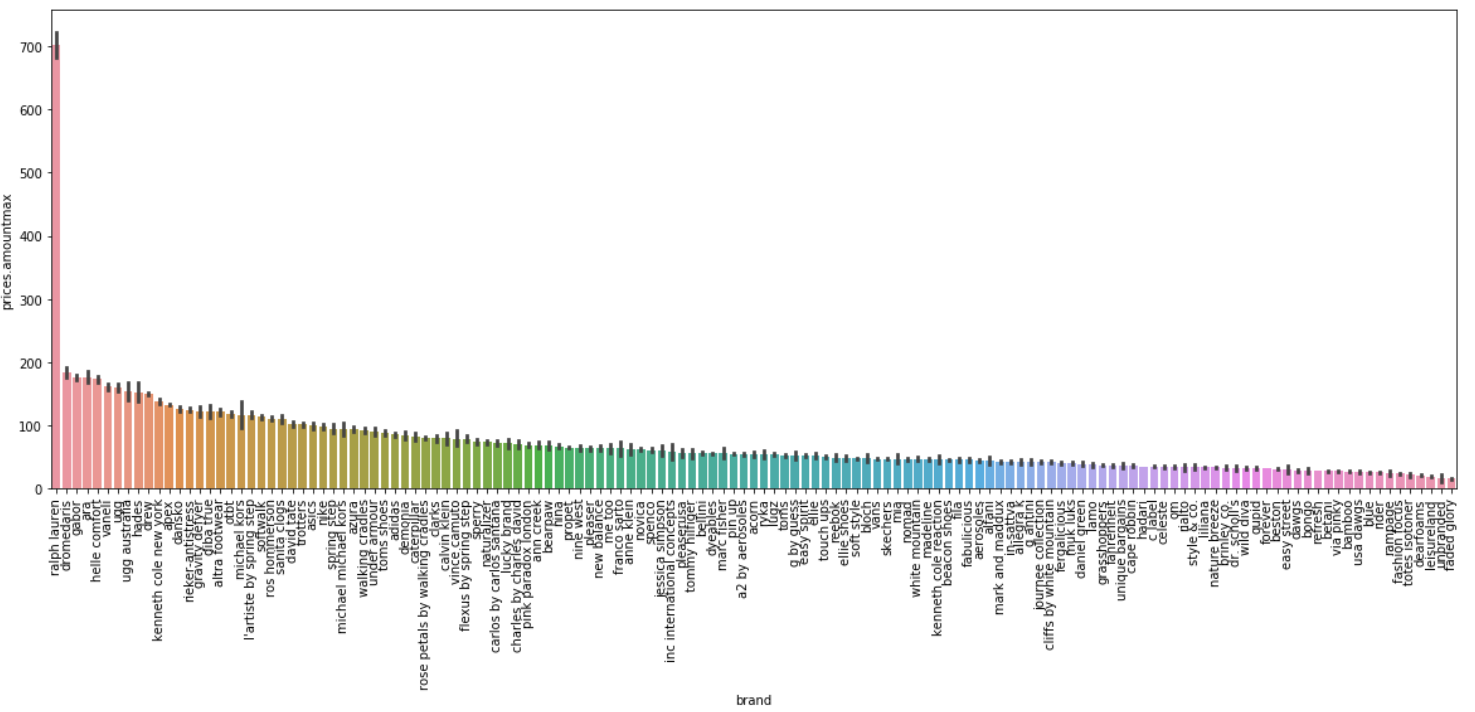
Removing outliers using Z score or IQR will cause loosing a big amount of data and we will end up with a small dataset. So we will remove outliers depend on the visualization on our initial boxplot.

We will remove all the data price below $10 and over 1000 to reduce the variation. Also we are only keeping the brands who sold more than 50 pairs of shoes for our analysis. The dataset now got 24685 values and 31 features.

### Creating new data frame for machine learning model

We will create a new data frame called "shoedf" with only relevant features for building the machine learning models to predict the brand. Based on the exploration analysis we choose features: 'brand','prices.amountmax','prices.issale','price.stat','brand.count','price\_band','brand\_code','colors\_code','merchant\_code' to create our final dataframe.

### Visualizations on final data frame

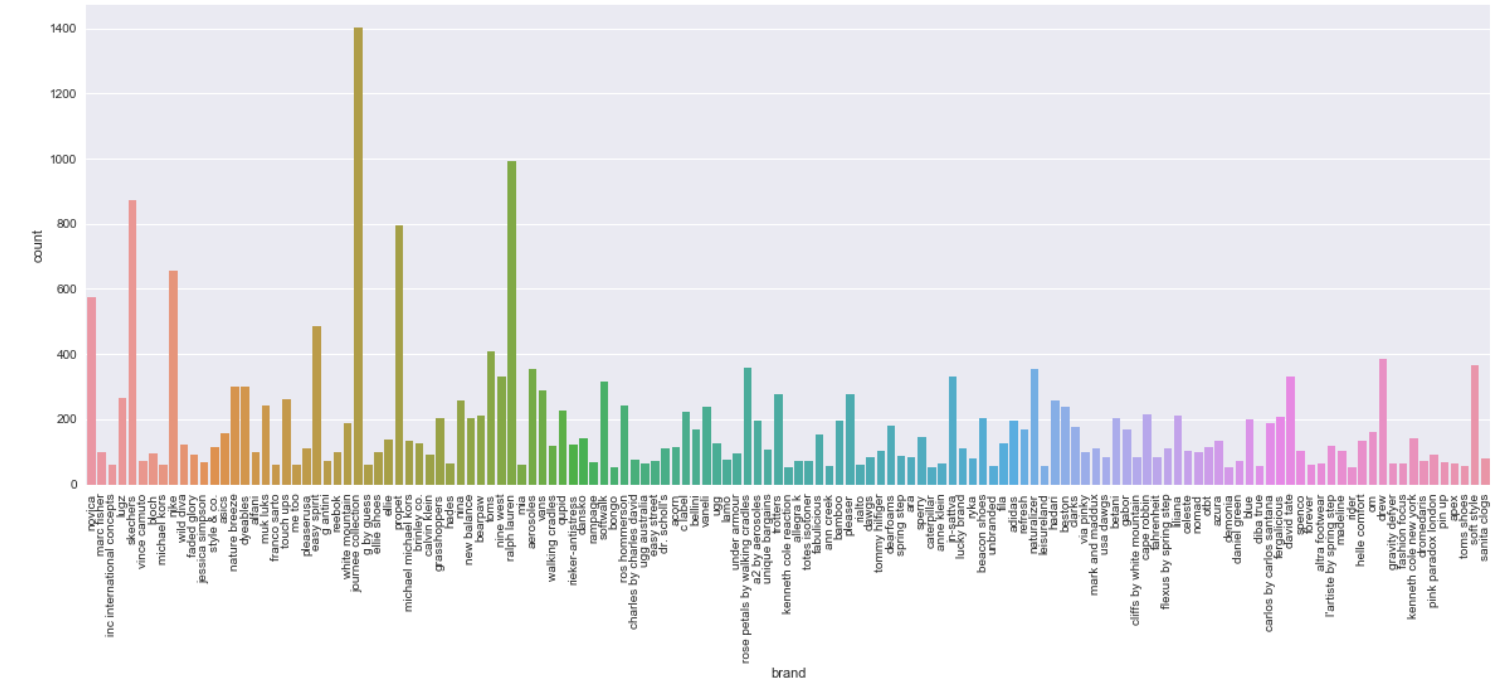


#### **Average price by brand**

Except' Ralph Lauren' all brands in the final dataset got average price under $200.

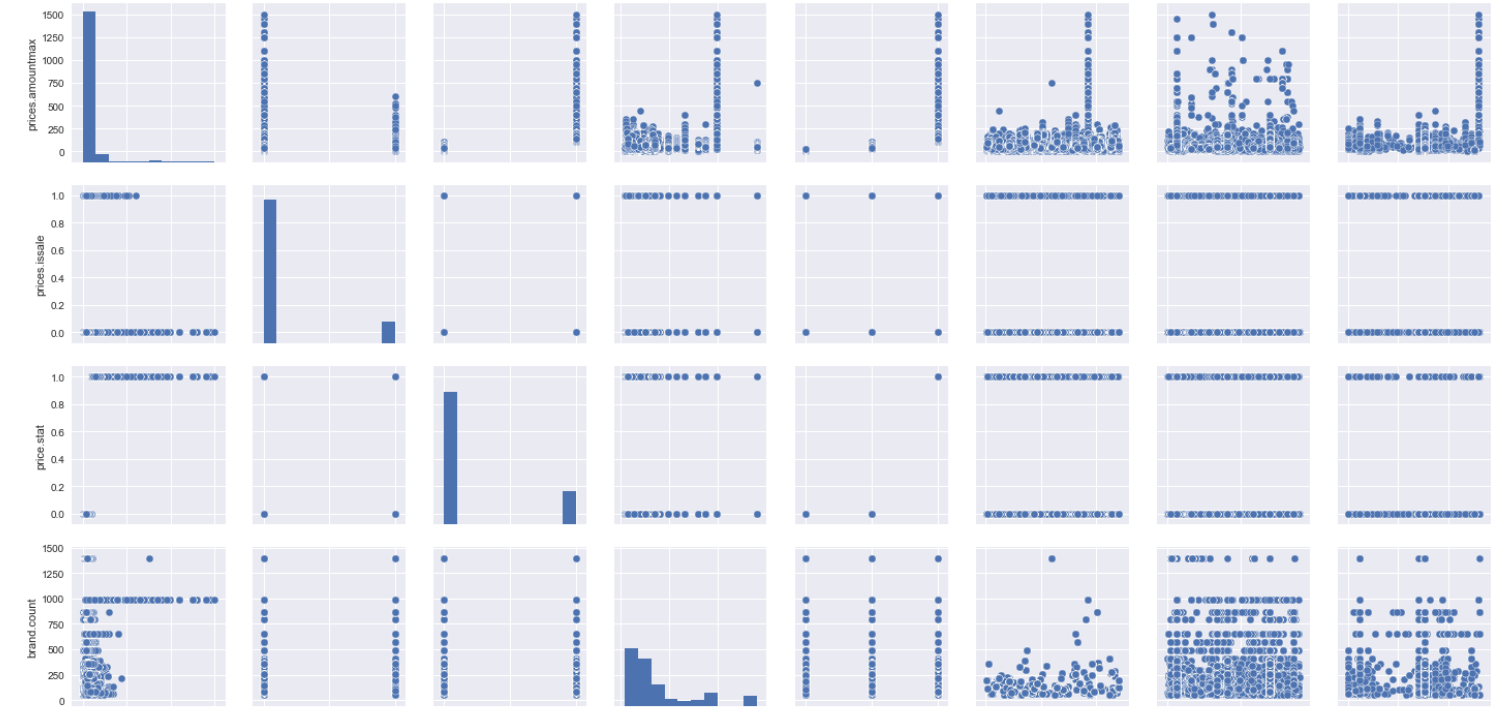
#### **Count plot of brand**

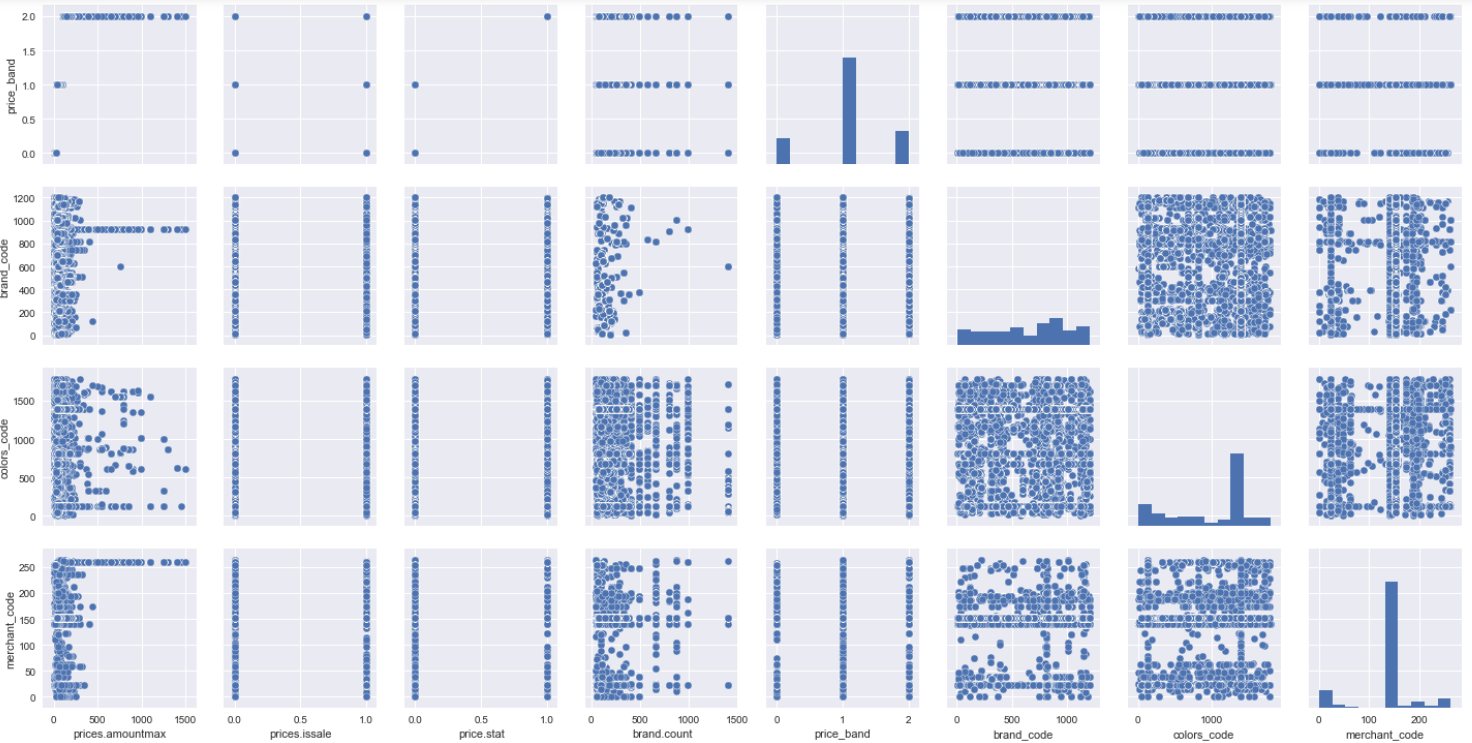
Journee collection is the top selling brand got a low average price.Ralph lauren is the second top selling brand got the highest average price.

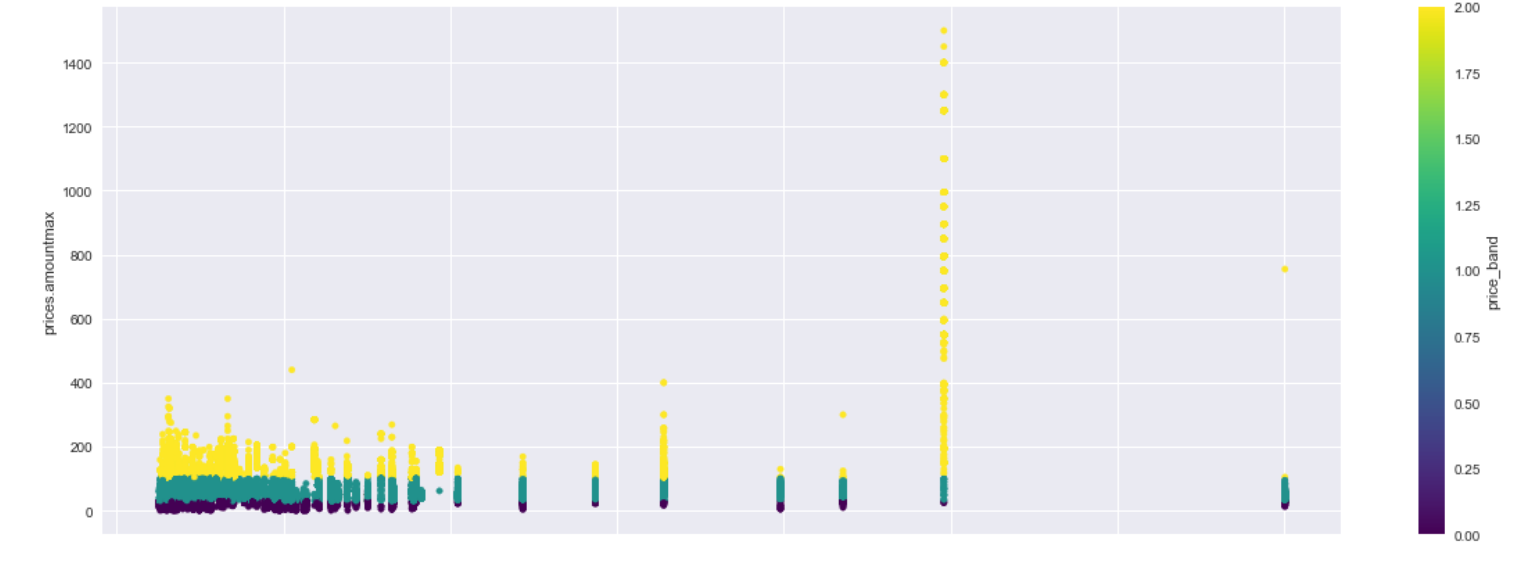


#### **Pairplots**

The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.







#### **Price that sell**

Above scatter plot displays shoes that sold more is in the price range over $30 and up. Women really do prefer to spend reasonable or high amount of money for their shoes.Cheaper price is not an attractive factor here.

#### **Data Modeling**

This step involves splitting data set to train and test. We split the dataset into 70% train data and 30% test data. The dependent variable is' brand\_code'( We use label encoding to change the brand value text to number and create the feature brand\_code ).

### Normalization

The next step is to normalize the train data set before applying machine learning. We use the StandardScaler from sklearn to achieve this.

The models we use SVM, K-nearest neighbors, and logistic regression require features to be normalized for better results

#### **Evaluate 6 different algorithams**

Next step is to try different machine learning algorithams in our model and use accuracy as evaluation metric. We are going to use 6 different algorithams and choose the one with high accuracy rate for cross validation to evaluate the model.

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

#### **Machine learning models**

1. Logistic Regression (LR)
2. Linear Discriminant Analysis (LDA)
3. K-Nearest Neighbors (KNN)
4. Classification and Regression Trees (CART)
5. Gaussian Naive Bayes (NB)
6. Support Vector Machines (SVM)

### Model Accuracy score

Logistic Regression (LR) - 22%

Linear Discriminant Analysis (LDA) - 83%

K-Nearest Neighbors (KNN) - 80%

Classification and Regression Trees (CART) - 97%

Gaussian Naive Bayes (NB) - 92%

Support Vector Machines (SVM) - 41%

###### It looks like CART has the largest estimated accuracy score. Linear regression has the worst.

##### **Accuracy of the CART model on validation set**

The accuracy is 0.97 or 97%. The classification report provides a breakdown of each class by precision, recall, f1-score and support showing excellent results.

### Insights from the Model and Business Recommendations

Brands do not follow any particular price distribution. The top selling brands sell more shoes in the price range between 30 and 150. 49.99 seems like a magic price sells more. Black, brown, white and blue are the popular colors.

Here is what i recomment based on the data : special promotions and sale for the shoes those priced over 200 to attract more customers. Second customers who is willing to spend money on shoes on particular brand and price range be interested in other products within the price range or brand. Improve cross selling and upselling techniques by creating personalized recommendations.