Using Data Science to Determine Harley Davidson's Most Valuable Assets
Capstone by Allen Edge

Problem Formulation

What factors are most responsible for Harley Davidson's profits? As the most recognizable motorcycle company in the world, Harley Davidson has proven to be a staple of American culture and has also successfully established an international presence. The purpose of this capstone is to use machine learning to find the most vital components (features) of Harley's sales in an effort to make executive recommendations to increase profits. The codes for every step of this report are available in an accompanying Jupyter notebook.
Before doing so, however, we have to download the dataset, available here.

Data Wrangling

After downloading the data from Kaggle, I then uploaded it via Jupyter notebook and took the following steps to clean it:

1) I converted the original dataset from an XLSX spreadsheet to a CSV file via pandas
2) Once the CSV file uploaded, I used .info() to view the indices, column names, null counts and data types
3) I dropped columns deemed unnecessary, and created my target variable, Order_Profit
4) Only acceptable data types (int64 and int32) exist in the dataset, and I have reaffirmed that no null values exist

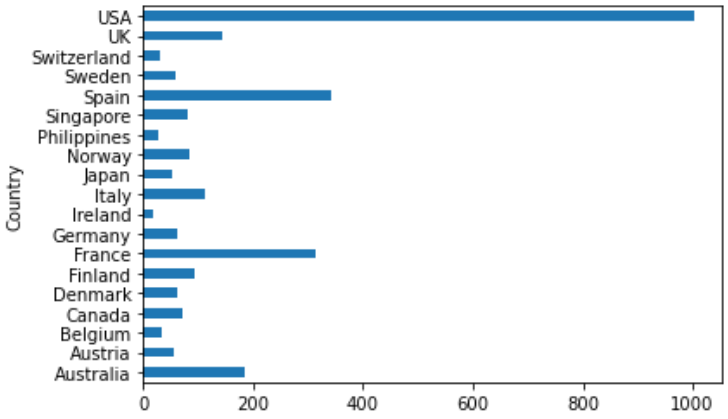
Now that the dataset has been cleaned, I'm ready to begin Exploratory Data Analysis.

Exploratory Data Analysis

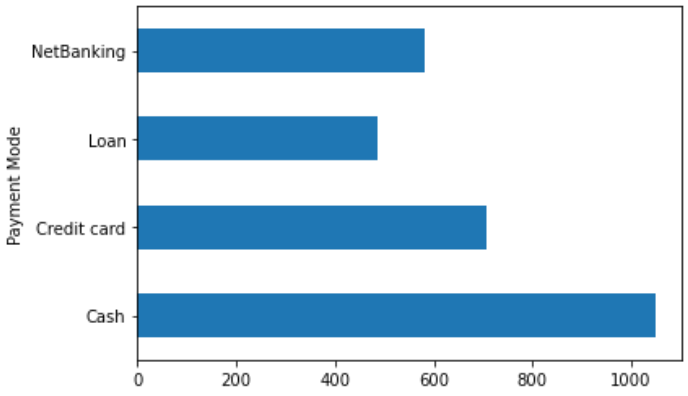
Before starting EDA, I've provided a list of the feature names and their respective meanings as a reference:

- Product Name: The name of Harley Davidson motorcycle
- Quantity: The number of motorcycles included per order
- Price: In US dollars, the cost per motorcycle
- Payment: The method of payment
- City: Name of the city of the store receiving the order
- Country: Name of the country that receives the order
- Year: Year in which the order took place
- Month: Month in which the order took place
- Order_Profit: earnings per order, defined as Quantity x Price

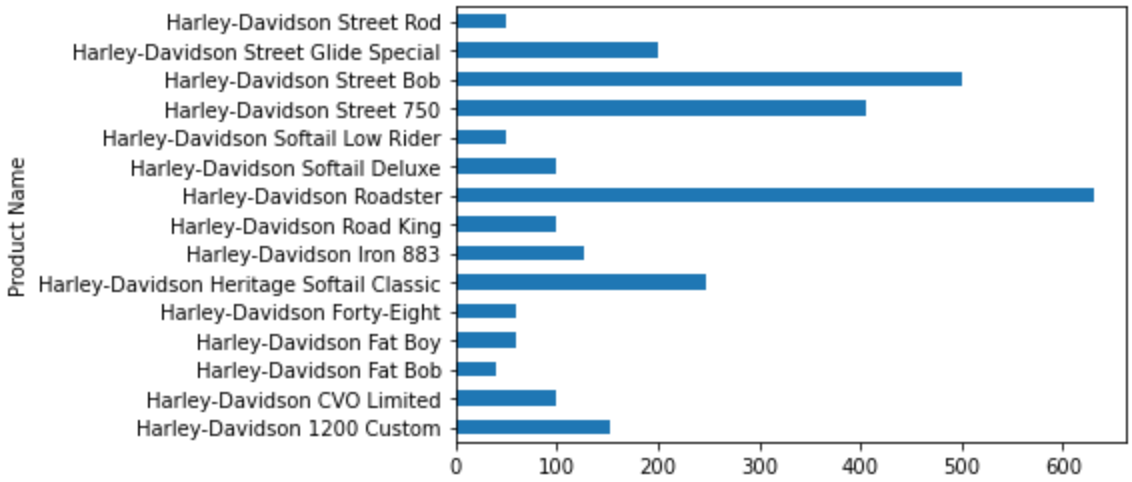
There are 2,823 orders containing 15 series of motorcycles spanning 73 cities in 19 countries. The first part of this EDA consists of analyses by order amount:



The US far surpasses the other 18 countries, responsible for 35.57% of motorcycle orders; conversely, Ireland accounts for just 0.57% of motorcycle orders.

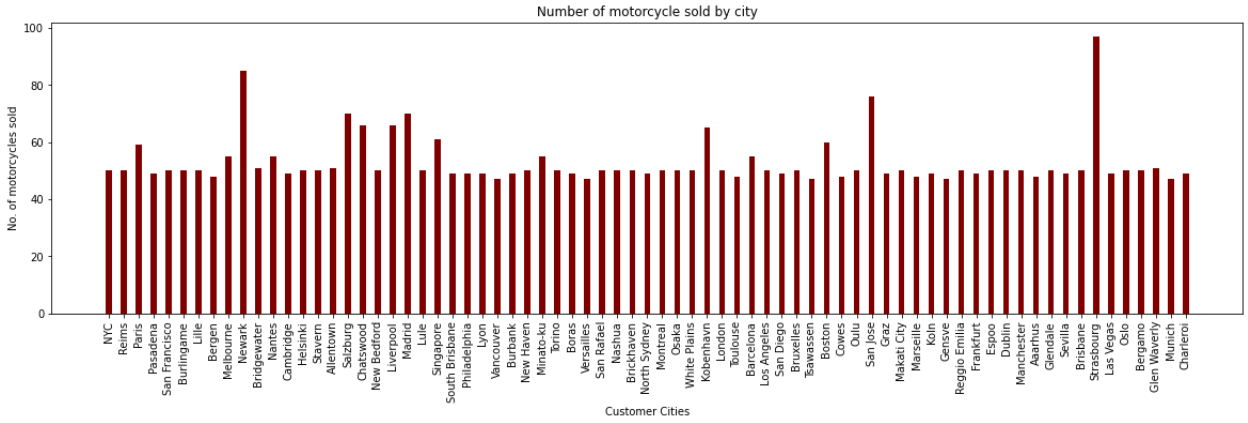


Of the four accepted payment methods, cash is the most common, accounting for 37.19% of all transactions. NetBanking, Loans and Credit Cards account for 20.62%, 17.21% and 24.97%, respectively.



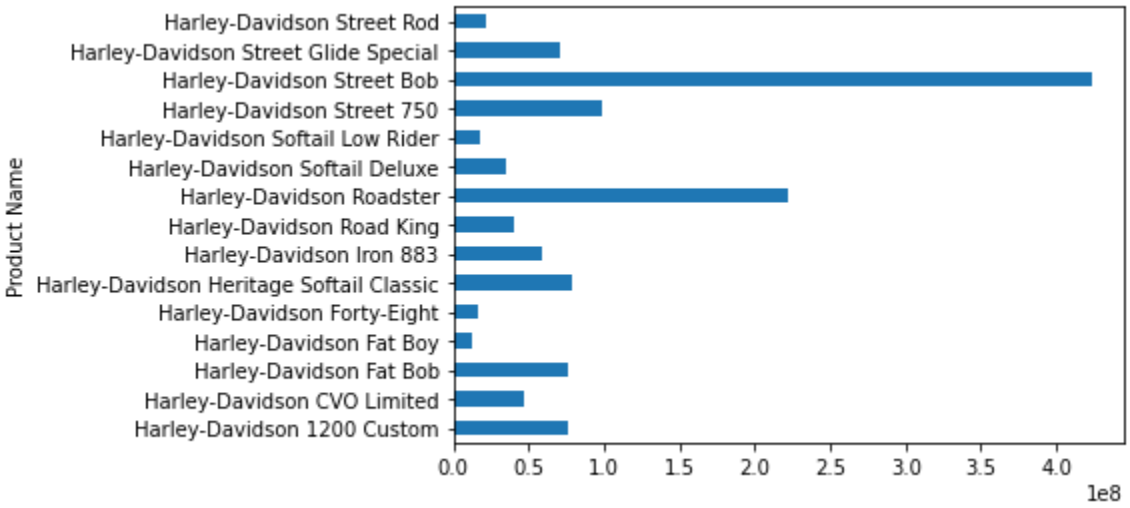
Heritage Softail Classic: 247
Harley-Davidson Roadster: 630
Harley-Davidson Street 750: 406
Harley-Davidson Street Bob: 500

The four aforementioned motorcycles are collectively responsible for nearly ⅔ of all sales, accounting for 63.16% of orders.



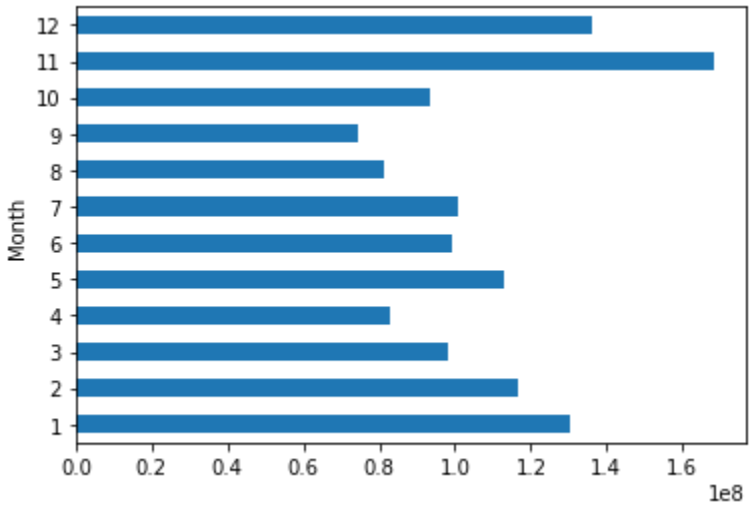
Newark, USA; Copenhagen, Denmark; San Jose, USA; and Strasbourg, France make up the most motorcycles received of any city

The second part of this EDA contains analyses by profitability. In the timespan of this dataset, Harley Davidson amassed $1,294,443,868:



Product Name	
Harley-Davidson Street 750	$12,810,000
Harley-Davidson Roadster	$221,490,000
Harley-Davidson Street Bob	$423,072,000

Two series of motorcycles, namely the Roadster and Street Bob, account for nearly 50% of Harley's profits; the Fat Boy, on the other hand, accounts for a mere 0.99%.



The above graph indicates profits by month. November and December account for the highest profits at nearly 23% combined.

Dummy Features, Scaling, Fitting, Test/Train Sets

There are four columns from this dataset which I converted into dummy features: Product Name, Payment Mode, City, and Country. That way, I can perform analyses on these categories, as they are unanalyzable as strings.

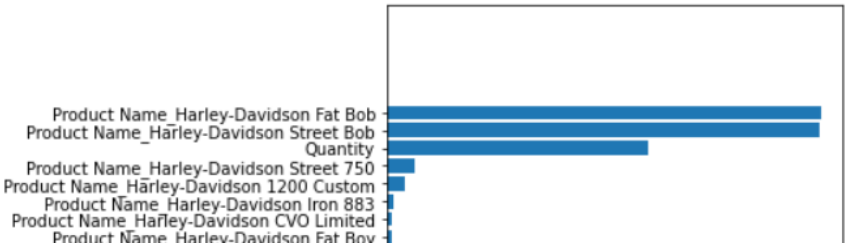
After successfully implementing the dummy variables, I then scaled and fitted the data using Panda's StandardScaler() method. As for the test and train sets, I subdivided the dataset into the dependent x variable, and the dependent y variable. I assigned every column other than the final Order_Profit to the x's as the independent and predicting columns, and the Order_Profit column to y. I want to use the given columns to predict the profit per order, and as such, assign y as the dependent column, Order_Profit. Following the standard 80/20 split, 80% of the data was used for training, and 20% of the data was used for testing.

Modeling

The goal of this project is to predict the profitability per order. Therefore, in the Modeling part of this project, I explored the use of different regression metrics as well as their importance features; the importance features reveal what are responsible for profits:

Model 1: Random Forest

The first model was a Random Forest. Scoring at 99.69%, it performed the best. In addition to excellent scoring, it also provided an easily interpretable importance feature: every feature from this dataset is a percentage value, and the sum of every importance feature from this dataset is 100%; the higher the percentage, the more important of a feature it is. The results are shown below:



- Based on the Random Forest Importance features, the Fat Bob, Street Bob, Quantity, Street 750, 1200 Custom, Iron 883, CVO Limited and Fat Boy are the largest determiners of profits
- By city: Kobenhavn, Reggio Emilia, Liverpool, Salzburg, Bruxelles, New Bedford, Chatswood, Las Vegas, San Jose, Strasbourg, Newark, Philadelphia, Melbourne, Madrid, Manchester, Brickhaven, Gensve, New York City, Boston, Nantes, Glen Waverly, Barcelona, Bergen, Marseille, and Lyon
- By country: Denmark, Italy, Austria, USA, Australia, UK, Belgium, France, Spain, and Singapore

Model 2: OLS Regression

The second model was the OLS Regression. With an r-squared value of 0.95, 95% of the profit's variance is explained by the independent variables. The importance features I used from OLS Regression came from the 'coef' (coefficient) column, where bigger coefficients result in a bigger response variable, or, in the context of this question, a bigger feature coefficient means a bigger profit. The results are as follows, grouped by category:

- Quantity, Year, Month
- Motorcycles: 1200 Custom, Fat Bob and Street Bob
- Cities: Barcelona, Bergamo, Bergen, Boston, Brisbane, Bruxelles, Burlingame, Chatswood, Dublin, Frankfurt, Helsinki, Kobenhavn, Koln, Lule, Madrid, Manchester, Marseille, Melbourne,Minato-Ku, Montreal, NYC,Nashua, New Haven, North Sydney,Oslo, Oulu, Paris, Reims,San Diego, San Francisco, San Jose, Ran Rafael, Singapore, South Brisbane, Torino, Vancouver, and White Plains
- Countries: Denmark, Ireland and Singapore

Model 3: LASSO Regression

The third and final model, the LASSO Regression scored 94.04%. Its importance features were similar to those of the OLS Regression, in that a bigger coefficient value implies more importance in predicting the results. Its results are listed below by category:

- By city: Barcelona, Bergamo, Boston, Brisbane, Bruxelles, Burlingame, Chatswood, Dublin, Frankfurt, Kobenhavn, Koln, Madrid, Makati City, Marseille, Melbourne, Minato-ku, Montreal, New York City, Paris, San Diego, San Francisco, San Jose, Singapore, South Brisbane, Torino, Vancouver, and White Plains
- Loans and cash were the only positive coefficients from payment methods
- Street Bob, Fat Bob, and 1200 Custom were the only motorcycles with positive coefficients
- Quantity is a very high coefficient

Client Recommendations

Through both Exploratory Data Analysis and Feature Importance, I was able to prove both visually and algorithmically what features are most vital and responsible for Harley Davidson's profits. Based on the findings above, I will list my own recommendations below which I think will allow for Harley Davidson's earnings to continue thriving:

- Continue focusing on the optimal quantity amount per shipment, as this has proven vital to the profits earned per shipment.
- Continue regularly producing the Street Bob, Fat Bob, and 1200 Custom. Their notable feature importance from all 3 tests indicates their value in the Harley Davidson industry.
- The 9 cities from the previous feature importance come from 6 countries; specifically we have the following: Kobnhavn, Denmark; Bruxelles, Belgium; San Jose, New York City and Boston, USA; Madrid and Barcelona, Spain; Melbourne, Australia; and Marseille, France. Continue shipping to them and regularly maintaining business relations with

them, as these are your most valuable customers. Also focus on the other cities and countries mentioned in the importance features, but prioritize your customer base found in the overlap of the three sets of importance features.

- Emphasize sales in November and December. These 2 months alone account for nearly 25% and 22% of profits, respectively.
- Consider limiting the production of all motorcycles that don't yield significant earnings and/or didn't appear on importance features.