Analysis of Adventure Works Cycles Company's Customer Data

Matthew Yeo, Apr 2016

Executive Summary

This document presents an analysis of data concerning the customers of Adventure Works Cycles Company. Based on 18355 observations of customer data, this document aims to analyze the features of each customer and their,

1. **Average Monthly Expenditure**
2. **Bike Purchase Label** (Customers are labeled 1 if they purchased a bike and 0 if they did not purchase a bike)

Before carrying out the analysis, the data first, undergoes preprocessing, meta-data editing and normalization.

Summary and descriptive statistics are calculated to understand the trend in the customer's average monthly expenditure and bike purchase label. Several potential relationships between the customer's average monthly expenditure, bike purchase labels and their corresponding features are also obtained using various data visualizations.

To model the data, a Decision Tree Classifier is used to classify customers into two categories depending on their bike purchase label. Finally, a Gradient Boosting Regression model is used to predict a customer's average monthly expenditure from their unique features.

After performing the analysis, the following conclusions can be drawn from the data.

While the customer data has many features that help in indicating a customer's average monthly expenditure and bike purchase label, through calculating feature importance, it is found that the following features are particularly significant.

- **Yearly Income** – the annual income of the customer.
    - *Average Monthly Expenditure* – There appears to be a positive correlation. A positive trend is observed where the higher the annual income, the higher the average monthly expenditure.
    - *Bike Purchase Label* – The median yearly income for customers who bought a bike is greater than those who did not.

- **Number of Children at Home** – the number of children the customer has who live at home
    - *Average Monthly Expenditure* – Customers with no children at home have a lower median average monthly expenditure than customers with one or more children at home
    - *Bike Purchase Label* – Customers who purchased a bike have a higher median number of children at home as compared to those who did not purchase a bike
- **Number of Cars Owned** – number of cars owned by the customer
    - *Average Monthly Expenditure* – There appears to be a positive correlation. The greater the number of cars owned, the greater the average monthly expenditure.
    - *Bike Purchase Label* – The median number of cars owned by customers who bought a bike is greater than customers who did not buy one

Initial Data Exploration

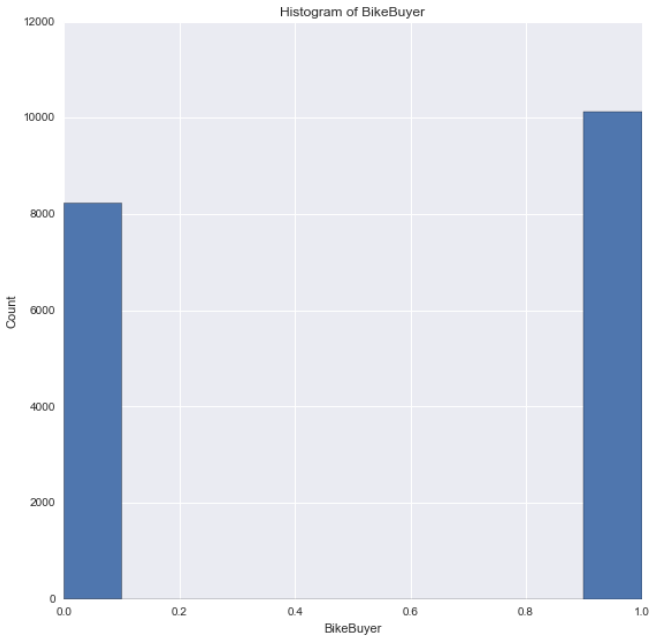
To explore each feature of the customer data, summary and descriptive statistics were calculated.

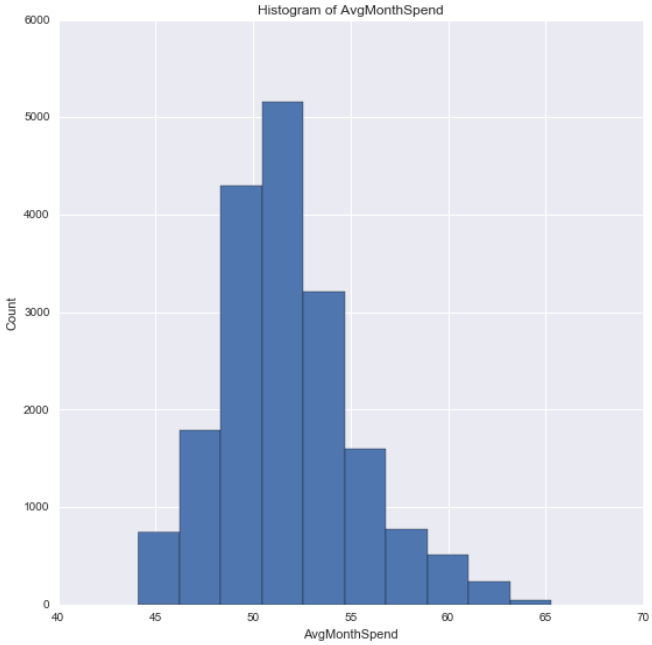
Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median and standard deviation were calculated for the numeric columns.

Column	Min	Max	Mean	Median	Std Dev	DCount
HomeOwnerFlag	0	1	0.610569	1	0.487634	2
NumberCarsOwned	0	5	1.270390	1	0.913887	5
NumberChildrenAtHome	0	3	0.338218	0	0.569001	3
TotalChildren	0	3	0.850449	0	0.927363	3
YearlyIncome	25435.0	139115.0	72758.95	61851.0	30687.7	15355
BikeBuyer	0	1	0.55173	1	0.49733	2
AvgMonthSpend	44.10	53.60	51.78	51.42	3.438	1803
Age	17	87	35.43	34	11.24	70

AvgMonthSpend and **BikeBuyer** are the two features that are of interest in this analysis.





With two distinct values, the histogram of the BikeBuyer column of the customers show that the a majority of the customers have purchased the bike.

Plotting the histogram of the AvgMonthSpend of the customers show that the AvgMonthSpend values appear to follow a normal distribution.

In addition to the numeric values, the customer data also includes categorical features:

- **Gender** – Male or Female
- **Marital Status** – Married or Single
- **Occupation** – Manual, Skilled Manual, Clerical, Management, Professional
- **Education** – Partial High School, High School, Partial College, Bachelors, Graduate Degree
- **Number Children At Home** – Number of children at home ranges from 0 to 3
- **Total Children** – Total children ranges from 0 to 3
- **Number Cars Owned** – Number of cars owned ranges from 0 to 5

Bar Charts were created to show frequency of features with more than two distinct values. The following observations were made.

- Males are more common than females
- Most customers are employed under the Skilled Manual Occupation, followed by Clerical, Manual, Management and finally Professional. The trend is largely a linear decrease.
- Most customers have a bachelor's degree. The trend is similarly largely a linear decrease, with the next majority being Partial College, High School, Graduate Degree and finally Partial High School.
- Most customers are married as compared to being single.
- Number of children at home has a left skewed distribution with a majority of customers having no children at home
- Number of cars owned seems to have a normal distribution with a majority of customers owning one car
- Total children do not seem to show any apparent distribution with a majority of customers having no children

Correlation and Apparent Relationships

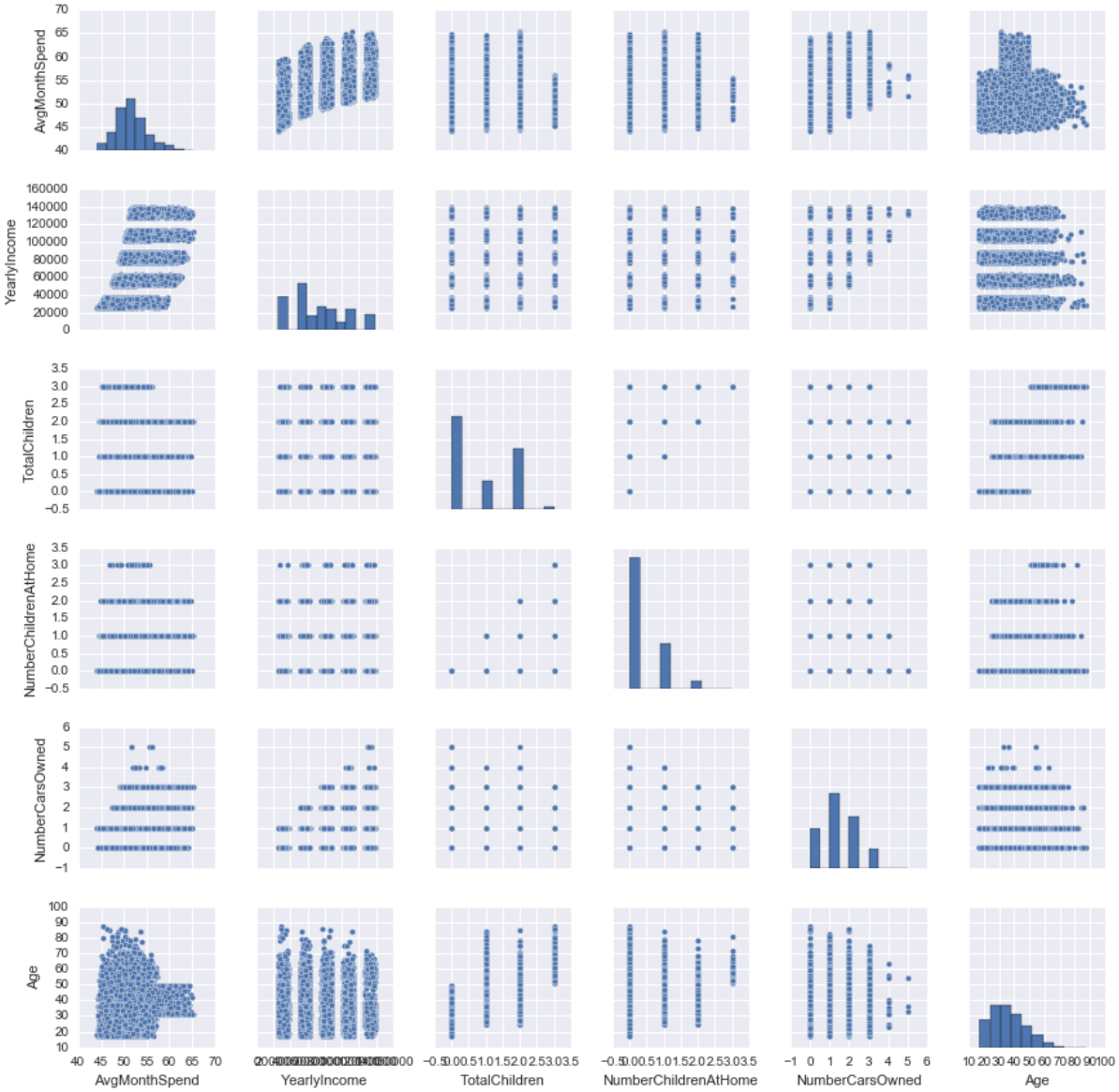
To have a better understanding of how customer features each influenced Average Month Expenditure and the Bike Purchase Label,

- **Average Month Expenditure**
  - Scatter-plot matrices were constructed to understand numeric relationships between Average Month Expenditure and other numeric features
  - Box-plots were used to understand categorical relationships and categorical feature
- **Bike Purchase Label**
  - Condition bar charts and scatter plots were used to understand both numeric and categorical relationships between the Bike Purchase Label and the other customer features

Average Month Expenditure

Numeric Relationships

As mentioned, a scatter plot matrix was used to compare each numeric feature with the Average Month Expenditure.



The scatter plot matrix of the numerical features of the customer data allows us to identify the more apparent relationships between features. In particular, there are several relationships that indicate a largely positive trend.

- **Relationship 1:** Average Month Expenditure and Yearly Income share a largely positive relationship.
- **Relationship 2:** Number of cars owned and the Yearly Income of customers also share a largely positive relationship.
- **Relationship 3:** Number of children at home also seems to have a positive relationship with the total children a customer has.

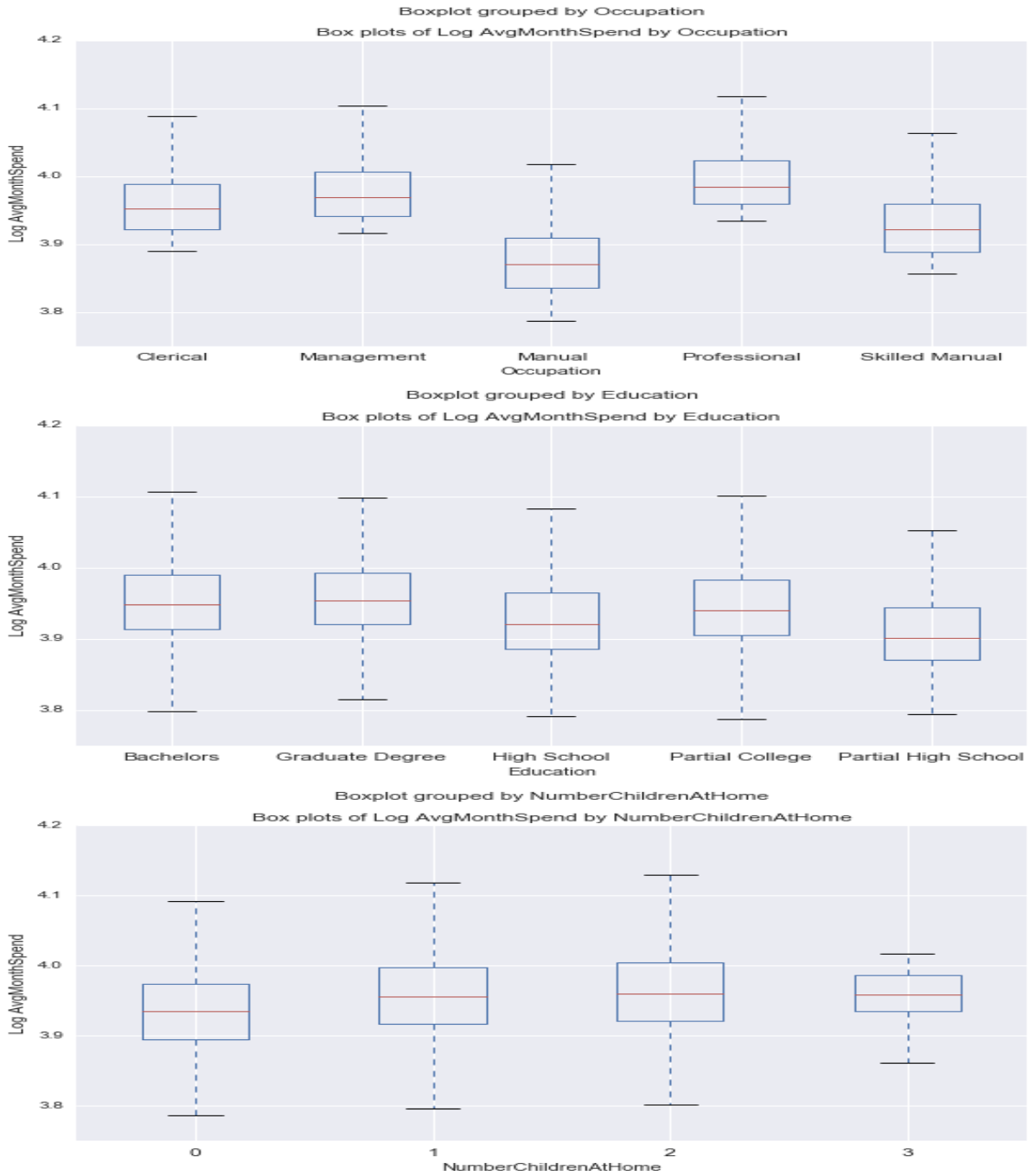
Upon calculating the correlation matrix, the above three relationships have the largest Pearson Correlation coefficients

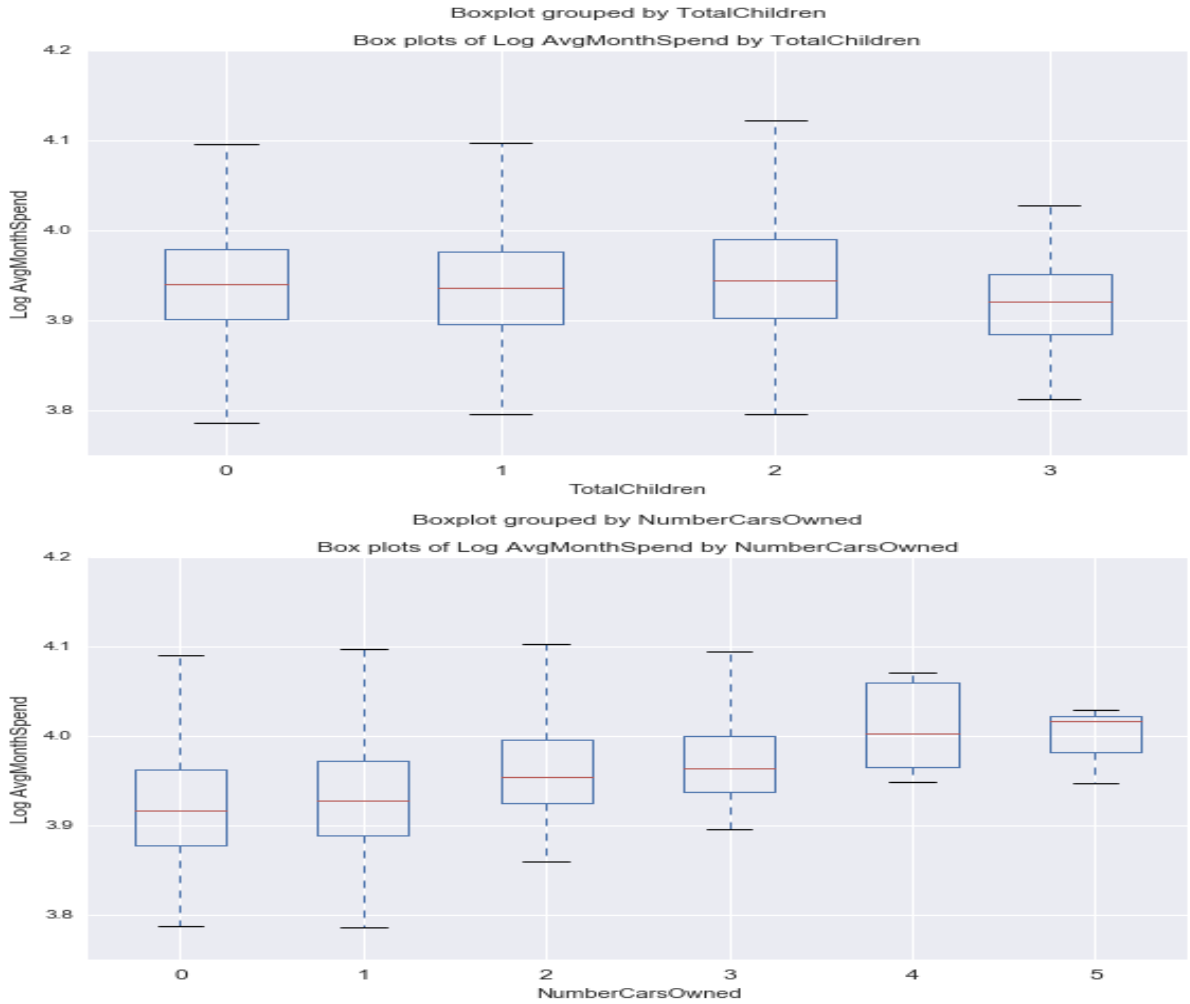
- **Relationship 1:** 0.543319
- **Relationship 2:** 0.477301
- **Relationship 3:** 0.606142

Due to the largely categorical nature of numerical features such as **Total Children**, **Number Children At Home** and **Number Cars Owned**, relationships may not be as apparent and box plots may be more appropriate.

Categorical Relationships

To accentuate any relationships between the categorical features and the average month expenditure, a logarithm function was imposed on the average month expenditure.





Similar to the scatter plot matrix constructed above, the most notable positive trend is shared between the number of cars owned and the average month expenditure of the customer.

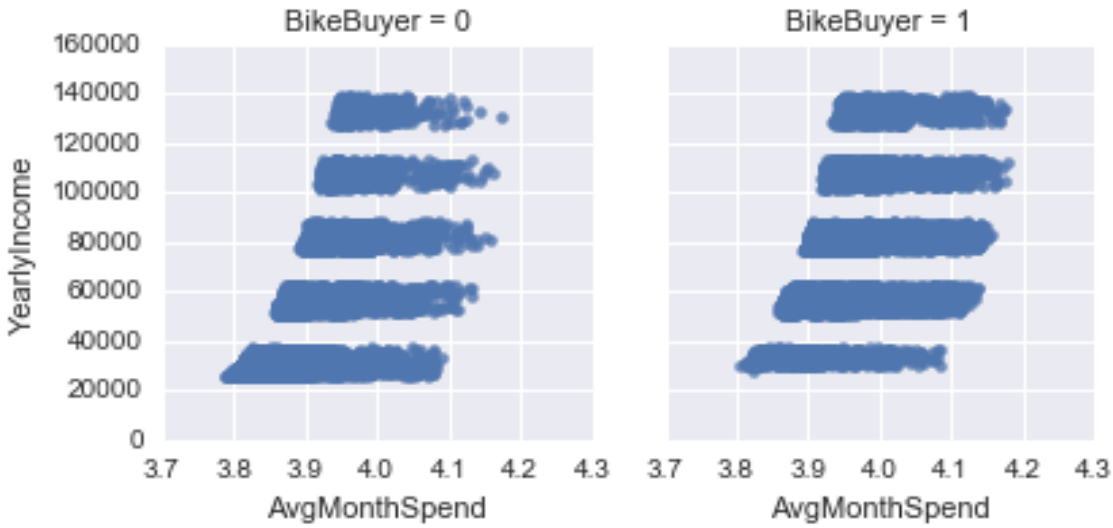
For the more explicit categorical features **Occupation** and **Education**, the median of the average month expenditure of the customer shows fluctuations depending on the type of **Occupation** or **Education**. **Occupation**, however, seems to cause more pronounced differences in the median of average month expenditure as compared to **Education**.

In the order of Manual Occupation, Skilled Manual, Clerical, Management and Professional, the median of average month expenditure increases.

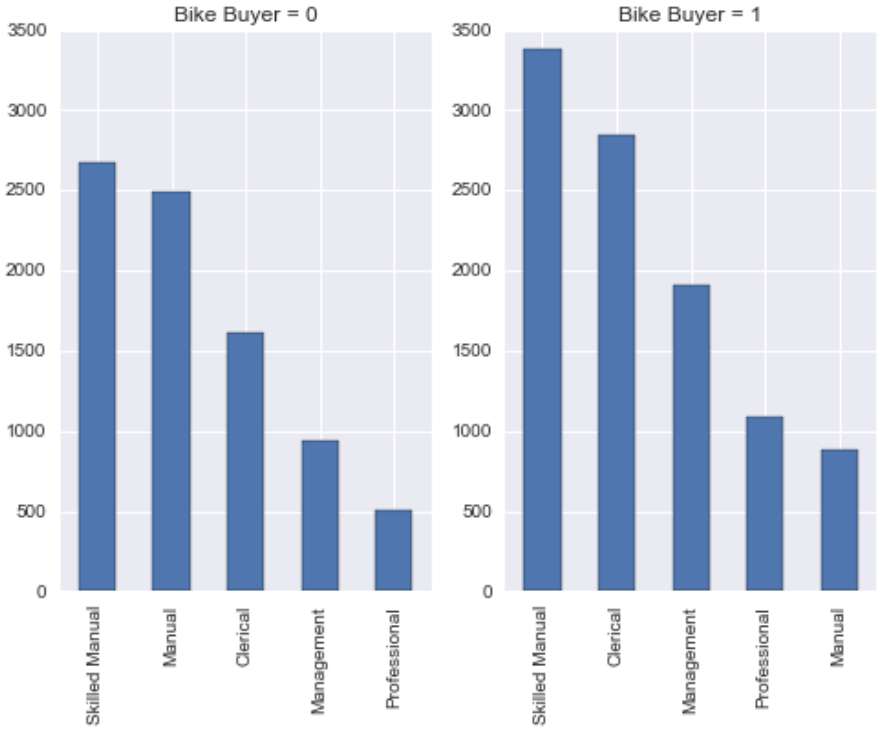
In the order of Partial High School, High School, Partial College, Bachelor's Degree and Graduate Degree, the median of the average month expenditure of customers increases.

Bike Purchase Label

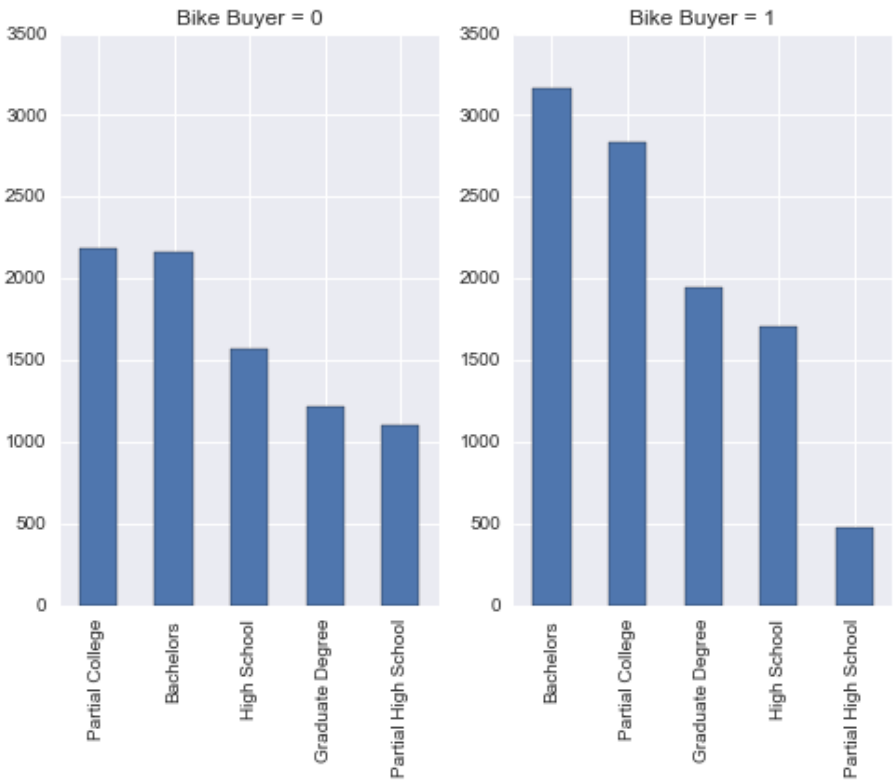
Multi-Faceted Relationships



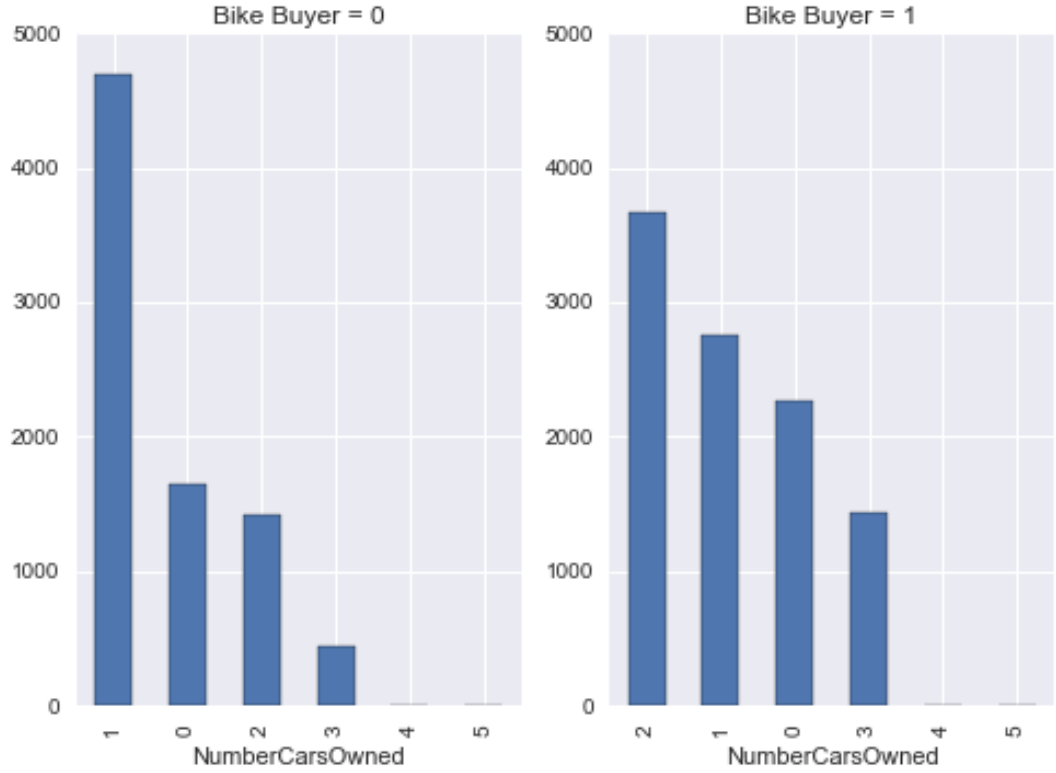
While there is no apparent relationship between average monthly expenditure and the bike purchase label, there is a positive correlation between the bike purchase label and the yearly income.
More specifically, customers who have purchased a bike enjoy a greater yearly income.



The distribution of bike buyers differs depending on their occupation. Most notably, manual workers are more likely to have not purchased a bike, while clerical, professional and management workers are more likely to have purchased one.



Education does not seem to have a very pronounced effect on the bike purchase label, as the distribution of customers remains largely similar across customers regardless of whether they purchased a bike. Most customers have either a Partial College or Bachelor's education, while the least customers have only a Partial High School education.



Finally, the number of cars owned appears to have a distinctive effect on whether a customer purchases a bike. Customers who own more than one car are more likely to have purchased a bike than customers who do not purchase a bike.

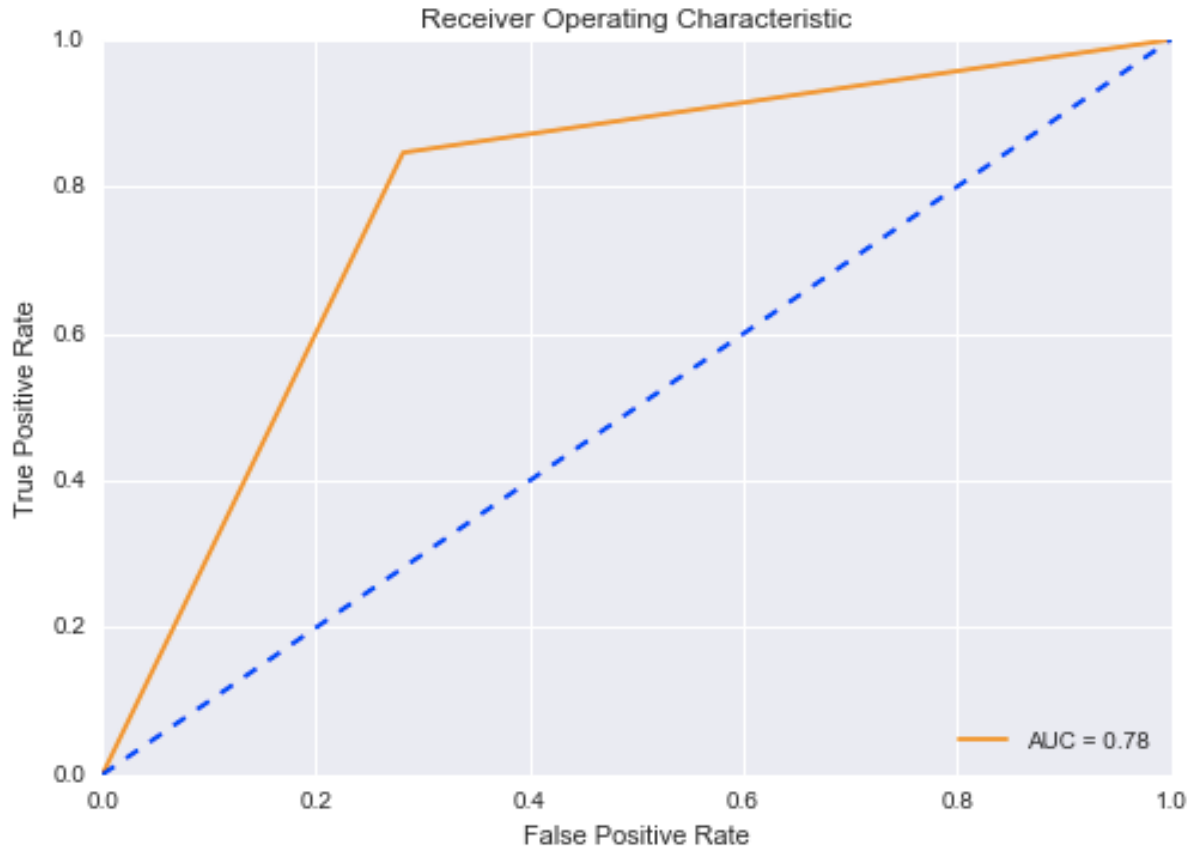
Classification of Customers based on their Bike Purchase Label

Through analyzing the customer data, a predictive model is constructed to classify customers into two Bike Purchase Labels: 0 (Customers who have not purchased a bike) and 1 (Customers have purchased a bike).

After iterating the data and scoring it based on various algorithms, the algorithm was used to create a model and trained with 70% of the data. The remaining 30% was then used to test the model. A GridSearchCV was run across the model's hyper-parameters, to determine which parameters could result in the highest accuracy.

After which, scoring the model against the test data again yielded

- True Positives: 3429
- True Negatives: 2364
- False Positives: 926
- False Negatives: 623



The Receiver Operating Characteristic curve of the model is also plotted, with the orange line indicating the model's performance at varying thresholds, and the blue line showing the expected results of a random guess.

From this curve, the following metrics can also be inferred.

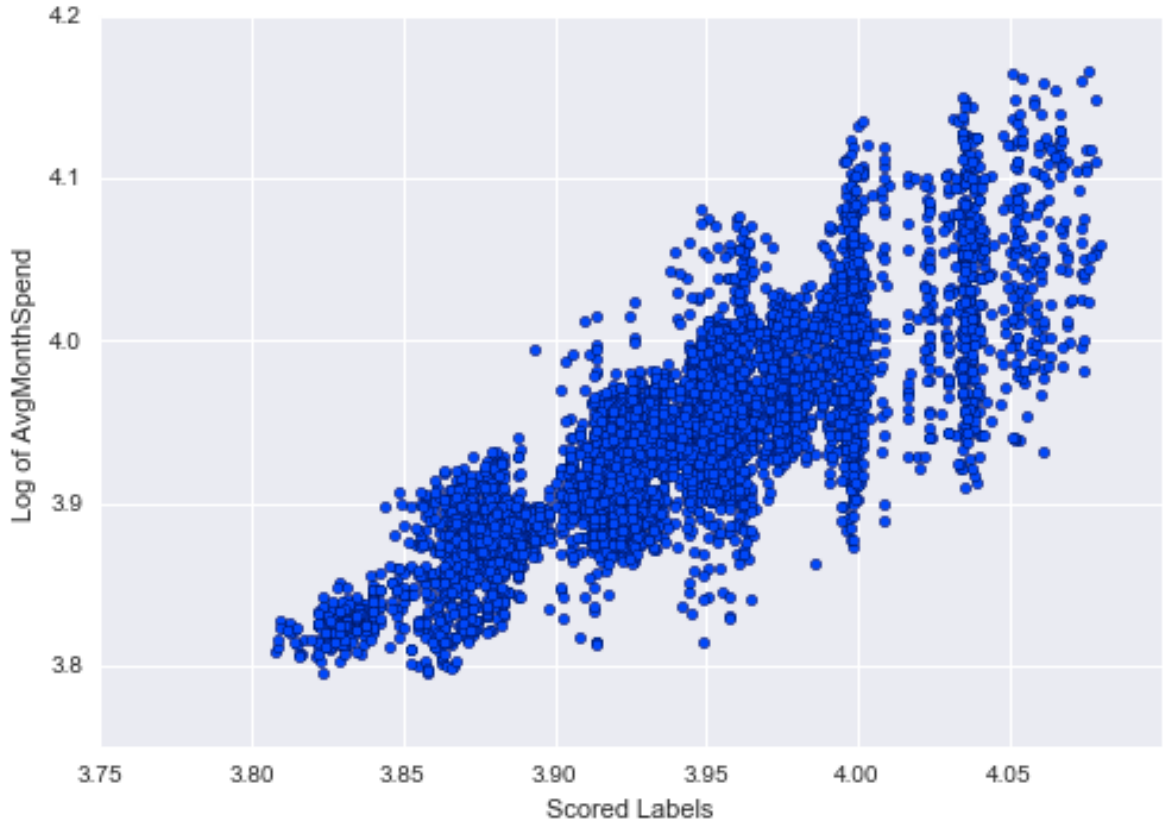
- Recall Score: 84.63%
- Accuracy Score: 78.90%
- Precision: 78.74%
- F1 Score: 81.58%

The model has a higher recall score than its precision, indicating that the model is better at identifying customers who have not purchased a bike then identifying those who have.

Regression of Customer's Average Monthly Expenditure

Finally, a regression model is used to predict the average monthly expenditure of the customers. Similarly, the data was iterated over various regression algorithms and the model with the highest accuracy was selected. A GridSearchCV was also used to tune the model's hyper-parameters to find parameters that will lead to the greatest accuracy score.

The model was trained with 70% of the data, while the remaining 30% was used to test the model's accuracy.



The plot indicates a generally positive linear relationship between the scored labels and the logarithm of the average monthly expenditure, showing that the model is able to make reasonable predictions. However, due to the relatively high Root Mean Squared Error of 0.036356, also shown by the spare distribution of points, the model suffers from lower accuracy. As a result, the model's accuracy score is only 68.3%.

Feature Importance

Through calculating feature importance, we are able to identify the features that are most important in the determining a customer's Bike Purchase Label.

Feature	*Feature Importance*
Gender	0.049144
Marital Status	0.028191
Home Owner Flag	0.018253
Number of Cars Owned	**0.177078**
Number of Children At Home	**0.266166**
Total Children	0.035968
Yearly Income	**0.203470**
Age	0.077262
Education (Bachelor's)	0.006679
Education (Graduate Degree)	0.004869
Education (High School)	0.004919
Education (Partial College)	0.005890
Education (Partial High School)	0.035872
Occupation (Clerical)	0.004751
Occupation (Management)	0.004904
Occupation (Manual)	0.056973
Occupation (Professional)	0.004622
Occupation (Skilled Manual)	0.014991

As seen, the features with the greatest feature importance values are the Number of Cars Owned, Number of Children At Home and finally Yearly Income of the customers. Age seems to also influence the decision on whether a customer purchases a bike.

Conclusion

In summary, the analysis done has allowed us to conclude that the Bike Purchase Label and Average Monthly Expenditure of customers can be predicted from its features. The Bike Purchase Label can be predicted with higher accuracy as compared to the Average Monthly Expenditure. Although many of the features seem to share some relationship with both the Average Monthly Expenditure and Bike Purchase Label, a further study on the feature importance has allowed us to narrow down the features to the four most important, namely being the Total Children, Yearly Income, Number of Children at Home and Number of Cars Owned.

However, in order to further increase the accuracy of predicting the Average Monthly Expenditure of customers, more customer features with stronger correlations with the Average Monthly Expenditure could be sourced and introduced to the data set.