Analysis of Adventure Work Cycles Customers

Alessandro Corradini, April 2017, Modena (Italy)

Executive Summary

This document presents an analysis of data concerning customers, their demographic features and purchases they have made. The analysis is based on 18361 observations of customer data.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between customer characteristics and their purchases were identified. After exploring the data, a predictive model to classify customers into two categories was created, bike buyer and not buyer, and finally a regression model to predict the average monthly spend of new customers based from its features was created.

After performing the analysis, the author presents the following conclusions:

While many factors can help to predict if the new customer will purchase a bike or not, significant features found in this analysis were:

- **Age**: Most Customers that bought a bike, are into 30-50 age range.
- **Gender**: Male Customers bought a Bike more frequently than Female Customers.
- **Marital Status**: Married customers bought a Bike more frequently than Single customers.
- **Home Owner Flag**: Customers with own Home, bought a Bike more frequently than Customers without a Home.
- **Number Children at Home**: Customers with children have more probability for purchasing a Bike.
- **Occupation**: Skilled Manual, Clerical and Manager Jobs have more probability for purchasing a Bike.
- **Education**: High level of education such a Bachelors or Partial College have more probability for a Customer to buy a Bike.
- **Yearly Income**: High Income increase the probability for a Customer to buy a bike.

While many factors can help estimate the Average Monthly Spend, significant features found in this analysis were:

- **Age**: Most Customers that are into 30-50 range, have a higher Average Monthly Spend.
- **Gender**: Male Customers have a higher Average Monthly Spend than Female Customers.
- **Home Owner Flag**: Customers with own house have a higher Average Monthly Spend than Customers without a home.
- **Numbers Cars Owned**: Customers with at least one car, have a higher Average Monthly Spend than Customers without a car.
- Number Children at Home:
- **Occupation**: Skilled Manual, Clerical and Management Jobs, with a greater spending capability have a higher Average Monthly Spend than less profitable jobs.
- **Education**: High level of education such a Bachelors or Partial College, have a higher Average Monthly Spend than Customers with low level of education.
- **Yearly Income**: Greater spending capability is translated into a higher Average Monthly Spends.

Initial Data Exploration

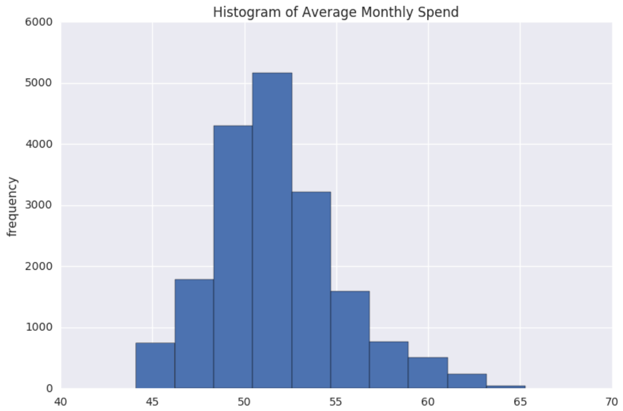
The initial exploration of the data began with some summary and descriptive statistics.

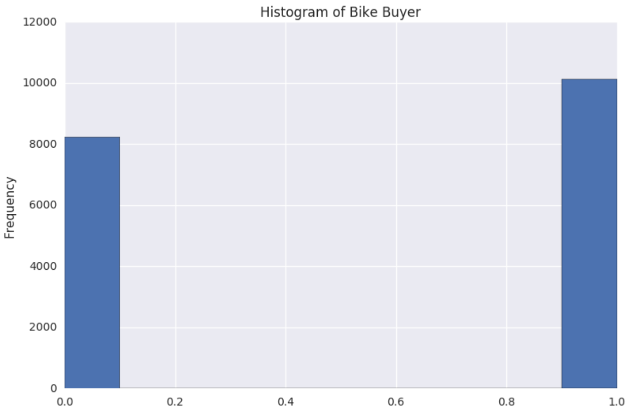
Individual Feature Statistics

Individual Feature Statistics Summary statistics for minimum, maximum, mean, median, standard deviation, were calculated for numeric columns, and the results taken from 18361 observations are shown here:

	Min	**Max**	**Mean**	**Median**	**Std. Dev.**
Number Cars Owned	0	5	1.270301	1	0.913989
Number Children At Home	0	3	0.338162	0	0.568957
Total Children	0	3	0.850389	0	0.927315
Yearly Income	25,435.00	139,115.00	72,754.78	61,851.00	30,686.014
Average Month Spend	44.10	65.29	51.766744	51.42	3.437684
Age*	17	87	35.427428	34	11.24848

***Age** is an approximate measure, it's calculated simply subtracting the current year by customer's birth year.





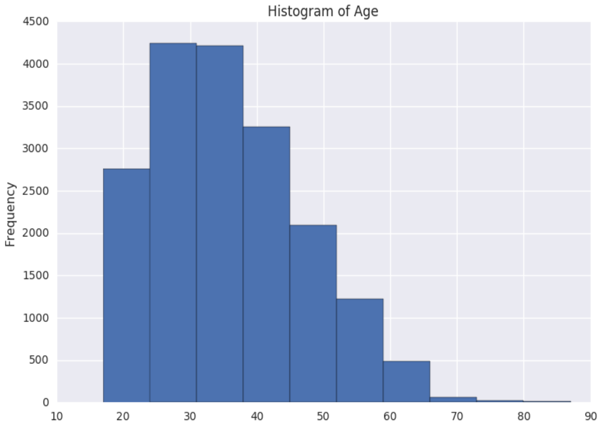
Above, we can appreciate the histogram and the distribution of the Average Monthly Spend and the Bike Buyer. The Average Monthly Spend appears right skewed, and the customers than bought a bike, are more than they didn't. In addition to the numeric values, the observations include categorical and demographic features, including:

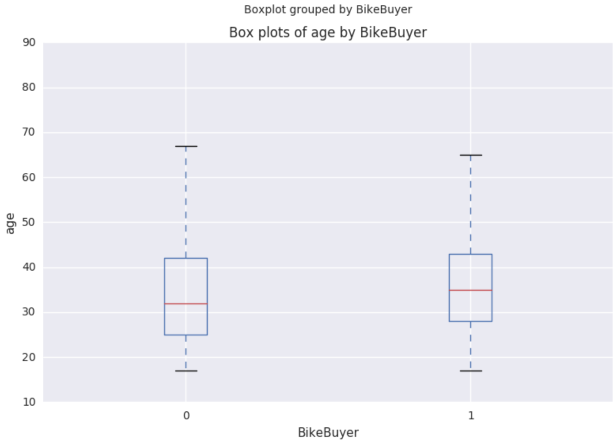
- **Title**: The customer's formal title (Mr, Mrs, Ms, Miss Dr)
- **City**: The city where the customer lives.
- **StateProvince**: The state or province where the customer lives.
- **CountryRegion**: The country or region where the customer lives.
- **Education**: The maximum level of education achieved by the customer:
    1. Partial High School
    2. High School
    3. Partial College
    4. Bachelors
    5. Graduate Degree
- **Occupation**: The type of job in which the customer is employed:
    1. Manual
    2. Skilled Manual
    3. Clerical
    4. Management
    5. Professional
- **Gender***:* The customer's gender (for example, M for male, F for female, etc.)
- **MaritalStatus***:* Whether the customer is married (M) or single (S).
- **HomeOwnerFlag***:* A Boolean flag indicating whether the customer owns their own home (1) or not (0).

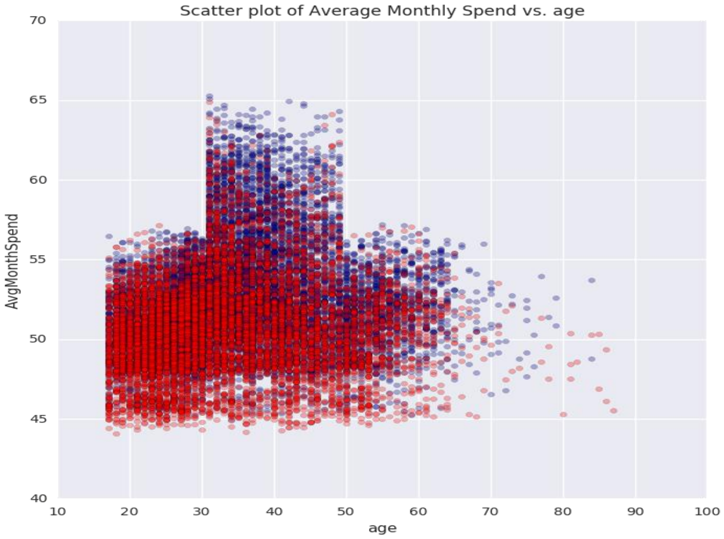
Correlation and Apparent Relationships

Different Views of the data features will help us to find correlations and relationships.

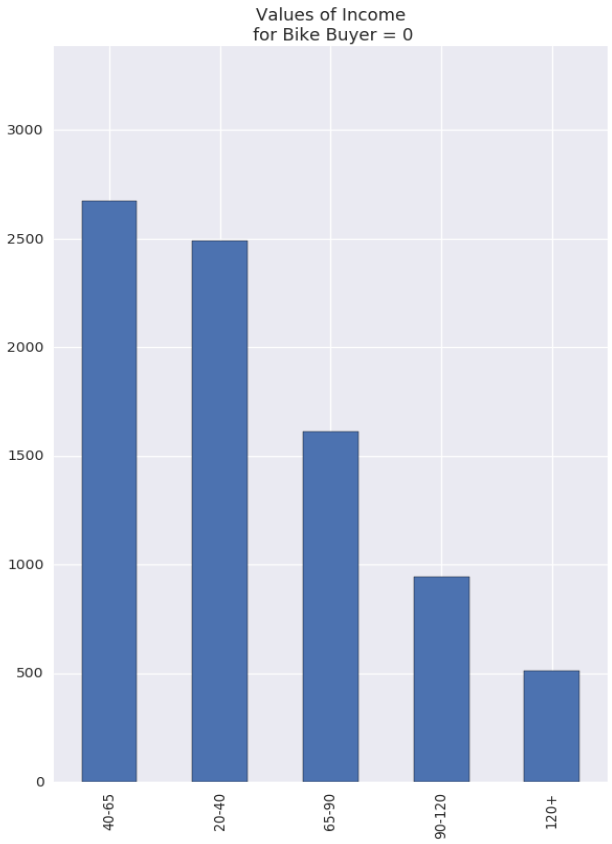
Numeric Relationships

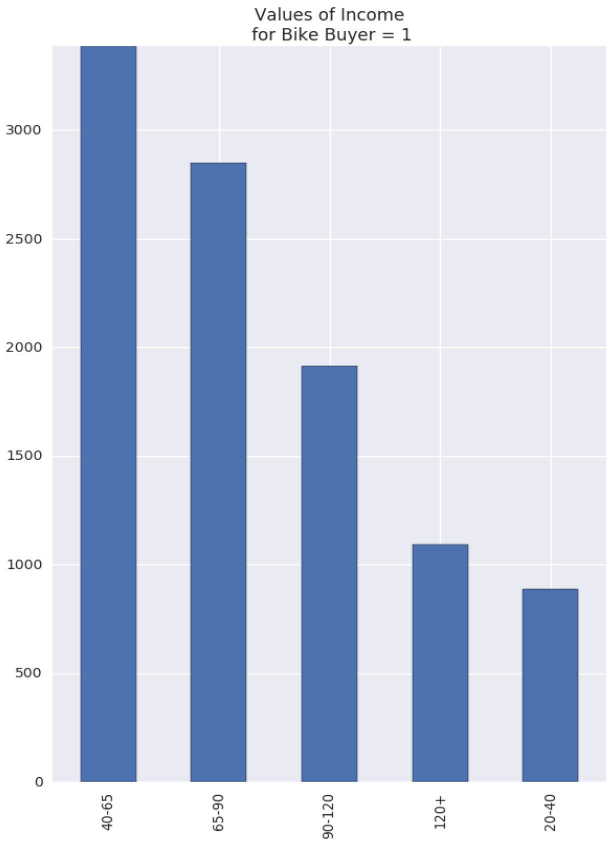


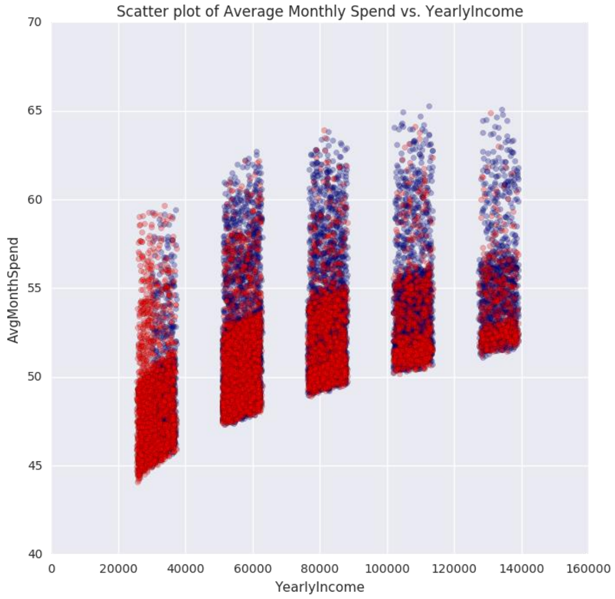


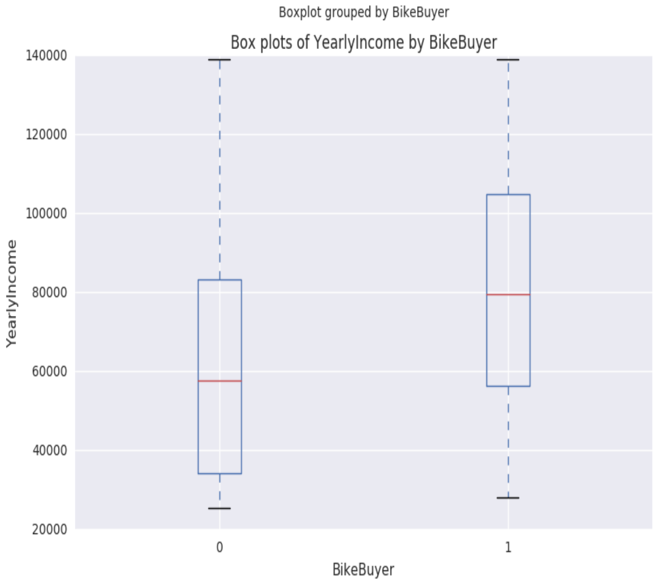


Age is an approximate measure, it's calculated simply subtracting the current year by customer's birth year. We can appreciate a right skewed histogram and the scatter plot about Average Monthly Spend and the age with a distinction from customers that have purchased a bike in **BLUE** and customers they didn't in **RED**. **This color distinction is applied for the other scatter plots.** Most customers are between 30 and 50 years old. We can identify some outliers for customers with 65 years old and older.


Values of Income
for Bike Buyer = 0

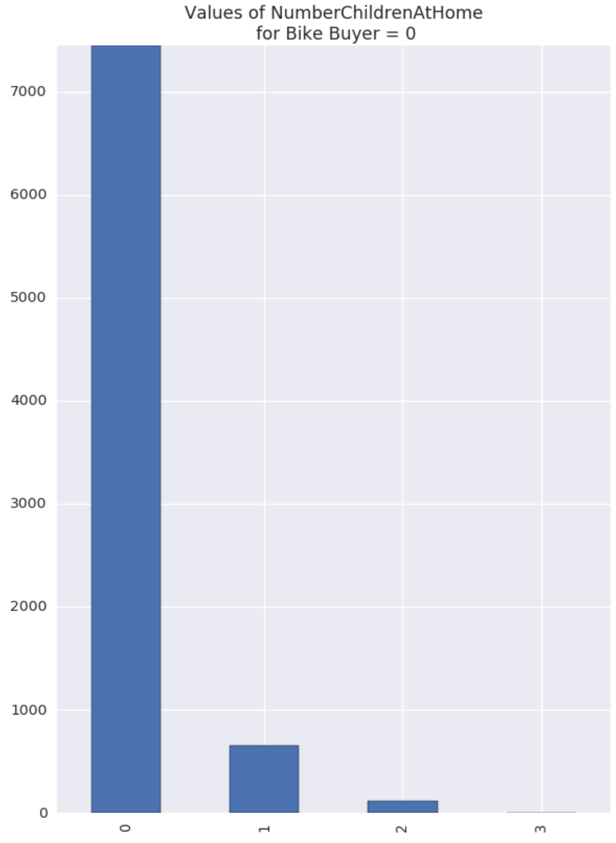

Values of Income
for Bike Buyer = 1

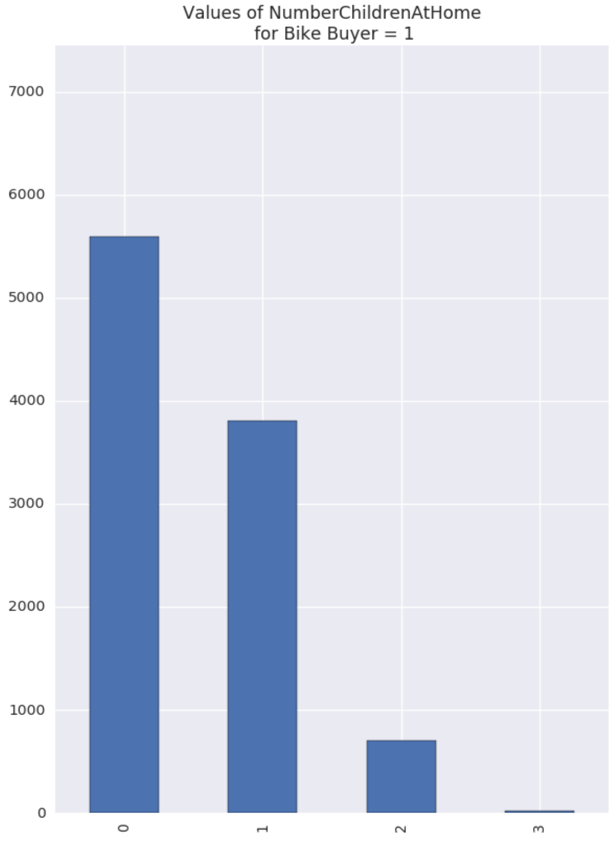

Scatter plot of Average Monthly Spend vs. YearlyIncome

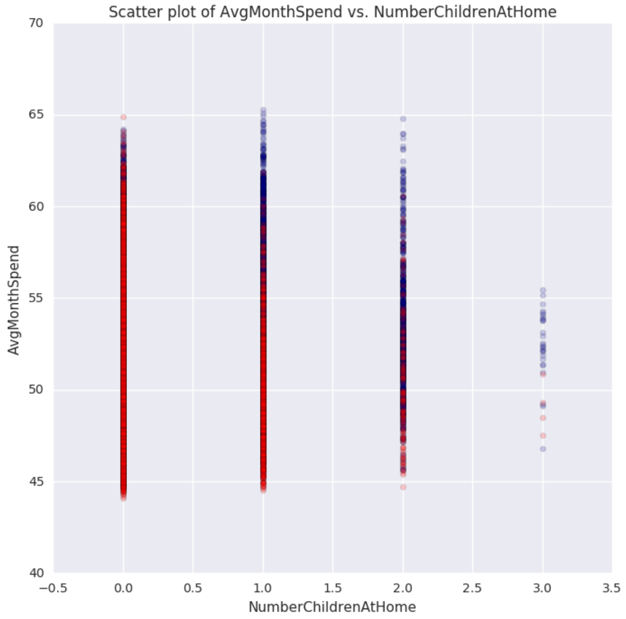

Boxplot grouped by BikeBuyer

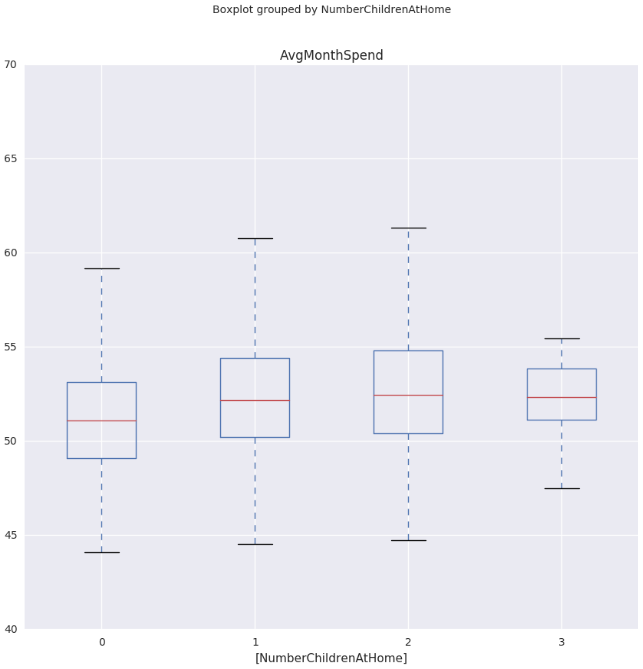
Box plots of YearlyIncome by BikeBuyer

YearlyIncome shows than are five groups of Incomes with a linear relation for its and Average Monthly Spend. Also, Bike Buyers have a high median of YearlyIncome. Customers with a medium-high Income, has more probabilities for purchasing a bike.

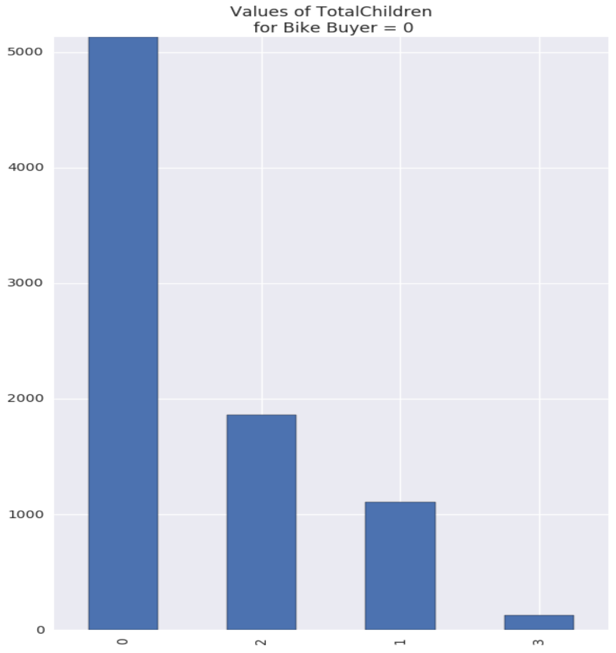

Values of NumberChildrenAtHome
for Bike Buyer = 0

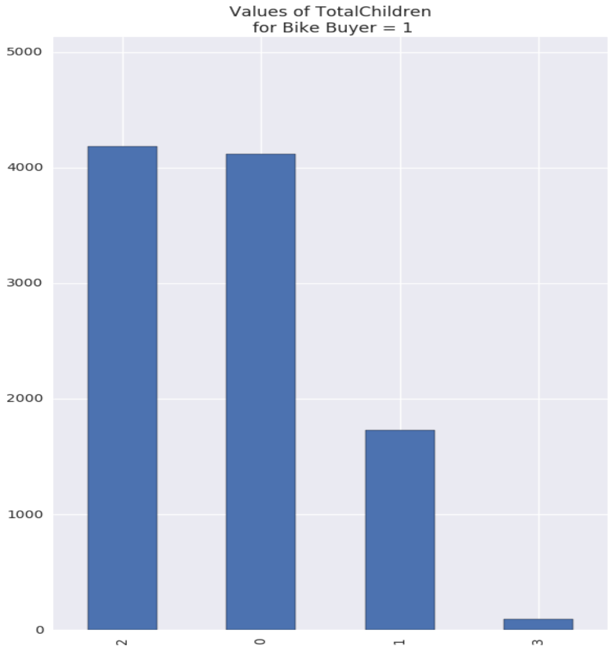

Values of NumberChildrenAtHome
for Bike Buyer = 1

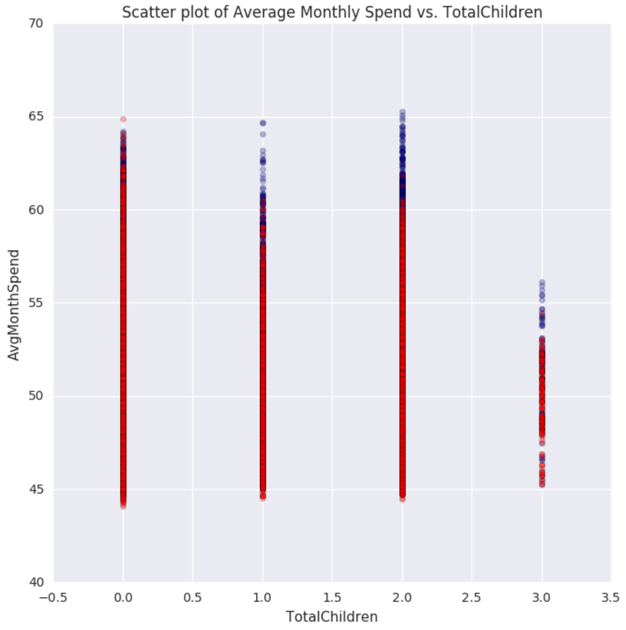

Scatter plot of AvgMonthSpend vs. NumberChildrenAtHome

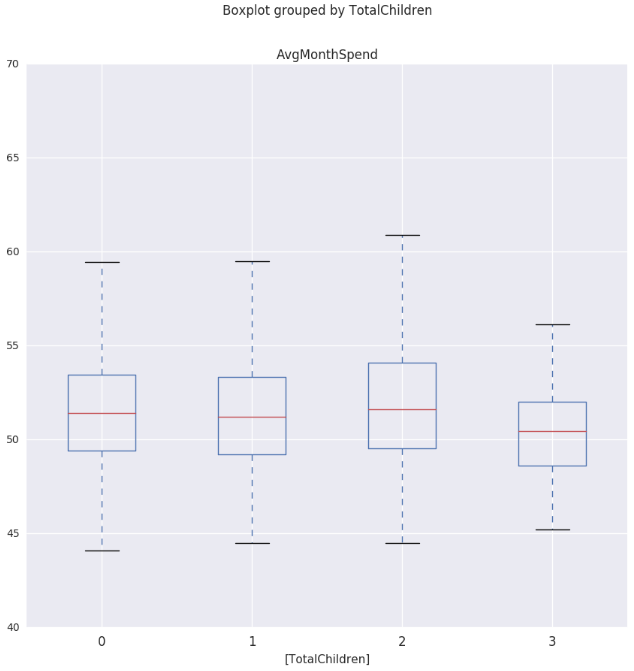

Boxplot grouped by NumberChildrenAtHome
AvgMonthSpend

NumberChildrenAtHome shows a relation between the number of children at home and purchases. For Customers with at least one children the probability to purchasing a bike, increase, and also increase the Average Month Spend. We can identify outliers for Customers with 3 children at home. For the predictions will be cleaned and transformed in Customers with 2 children at home.

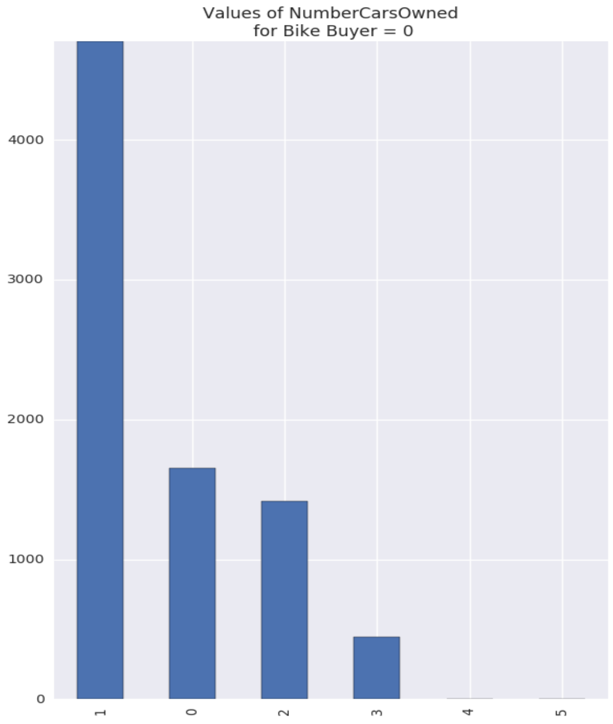


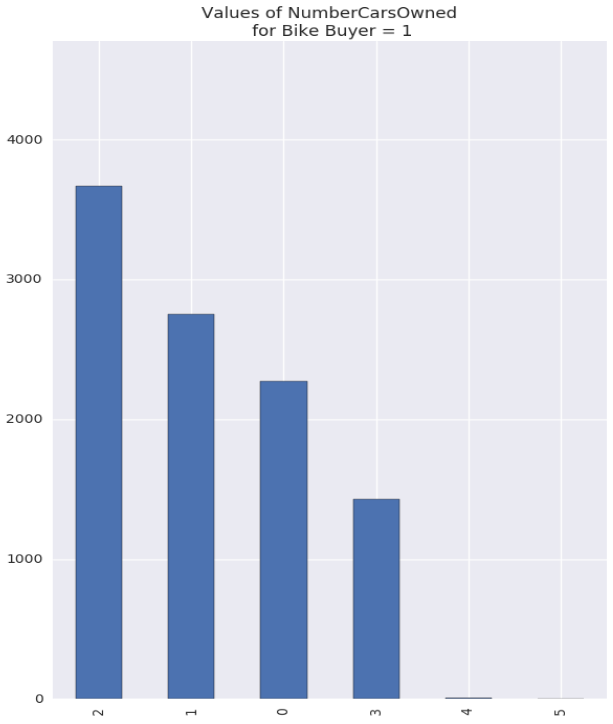


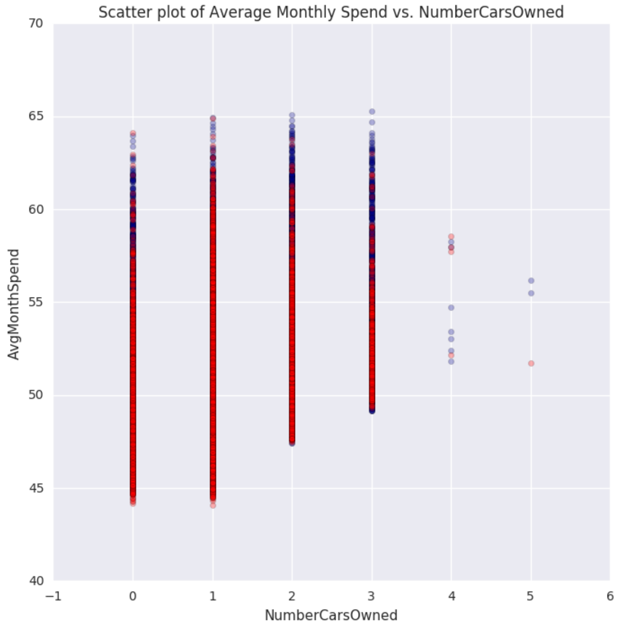


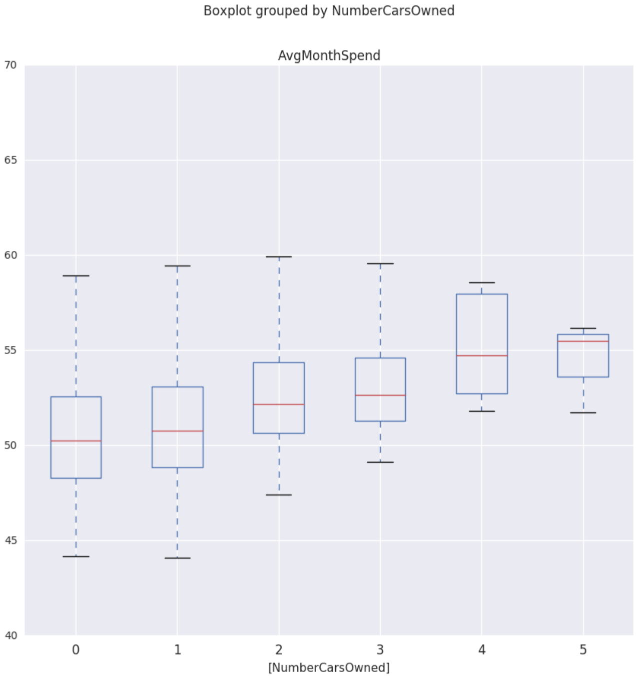


TotalChildren shows that Customers with at least one children bought a Bike. Most of No-BikeBuyers Customers haven't a child. We can identify outliers for Customers with 3 children in total. For the predictions will be cleaned and transformed in Customers with 2 children in total.



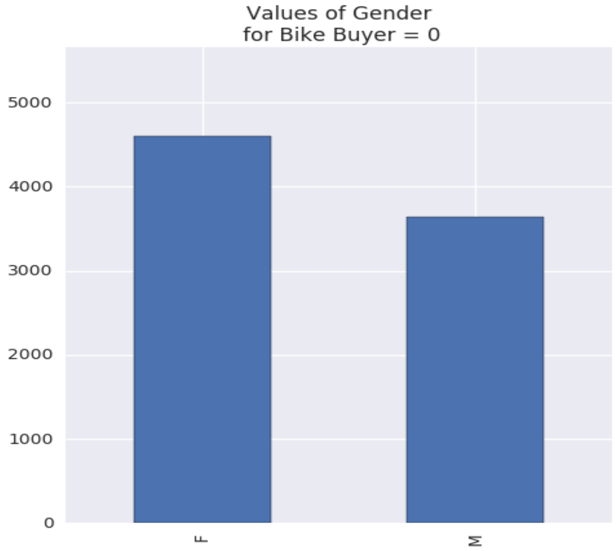


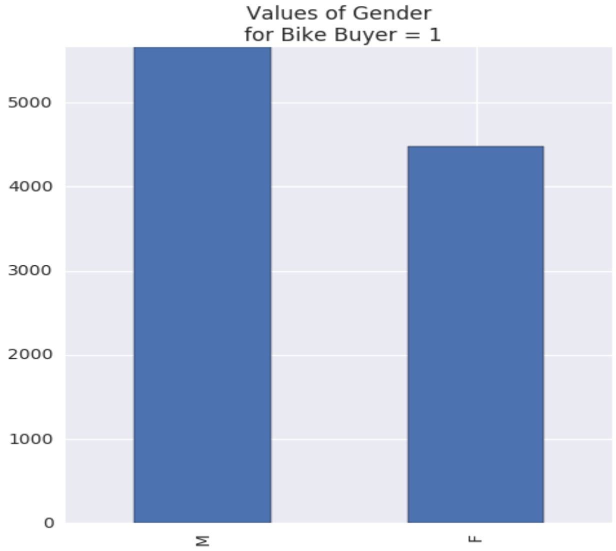


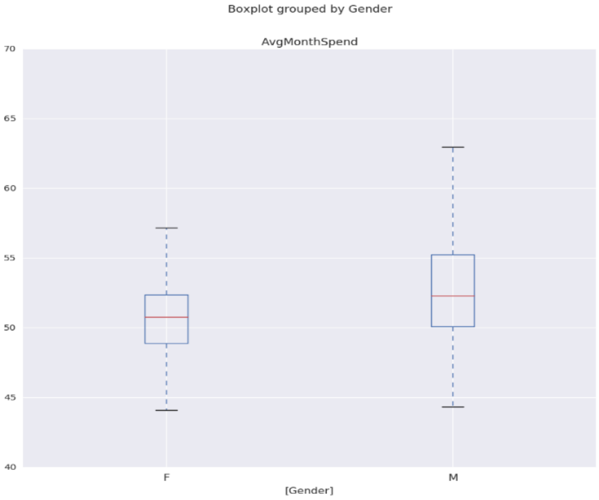


NumberCarsOwned is particularly useful for the Average Monthly Spend. This is a feature that indicates the level of economic wellbeing for customers and obviously, high level of wellbeing, translate into a higher Average Monthly Spend. We can identify outliers for Customers with 3 or more cars. For the predictions will be cleaned and transformed in Customers with 3 cars.

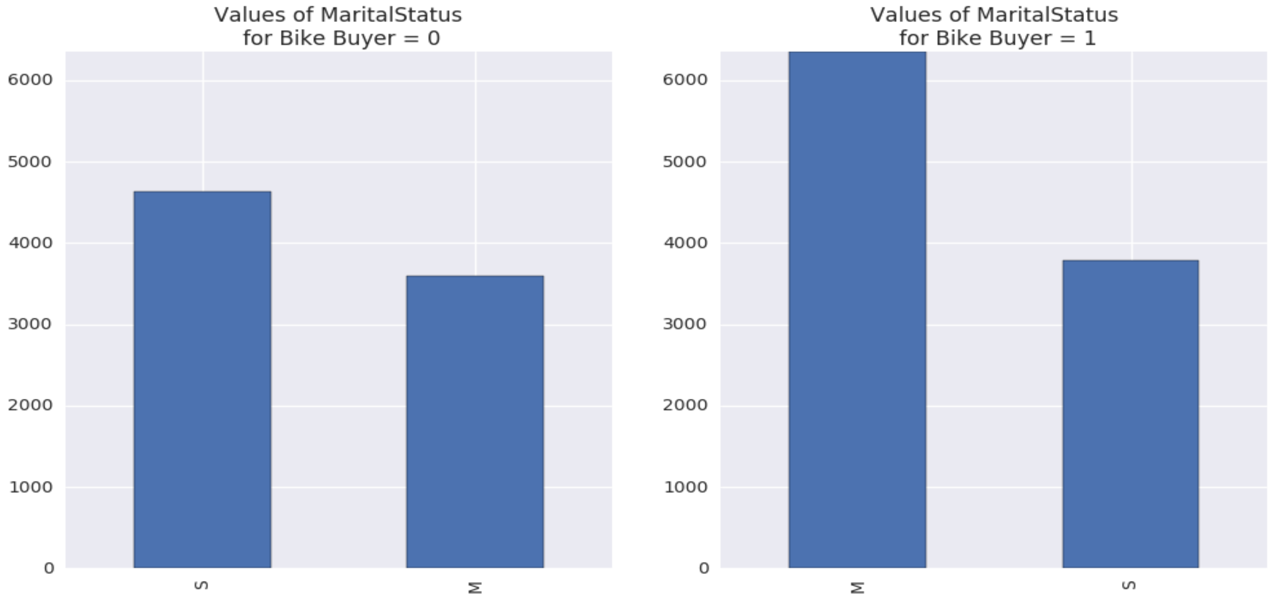
Categorical Relationships

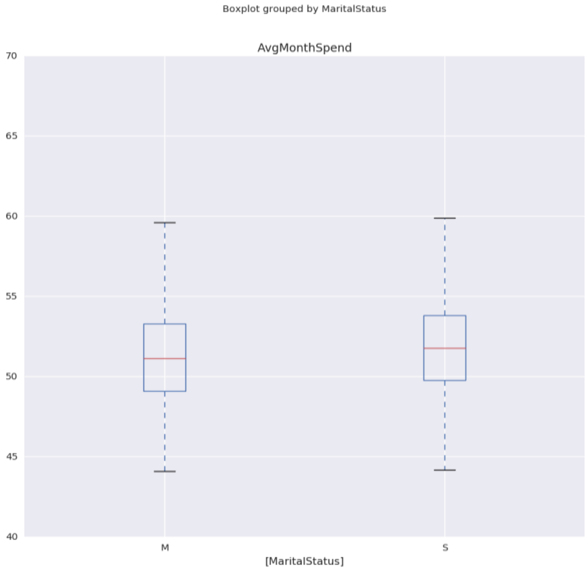




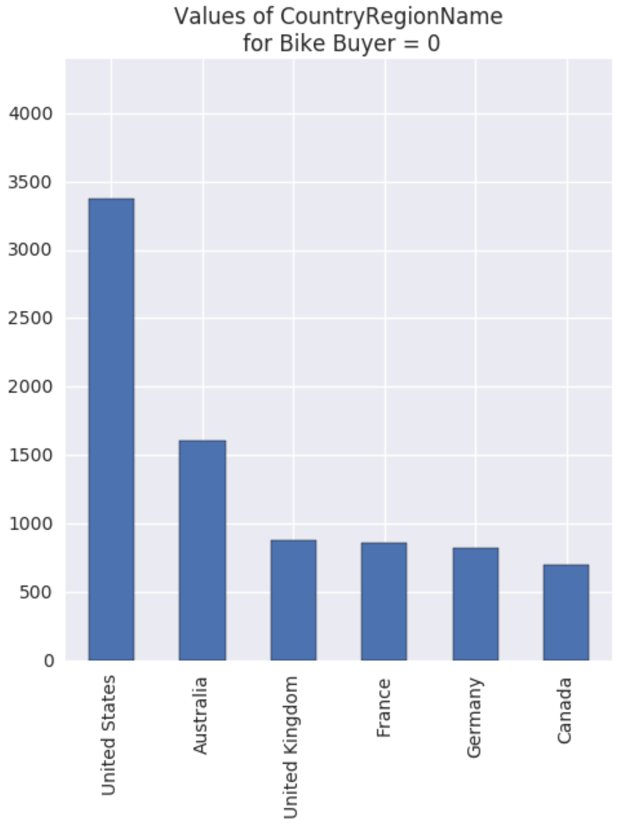


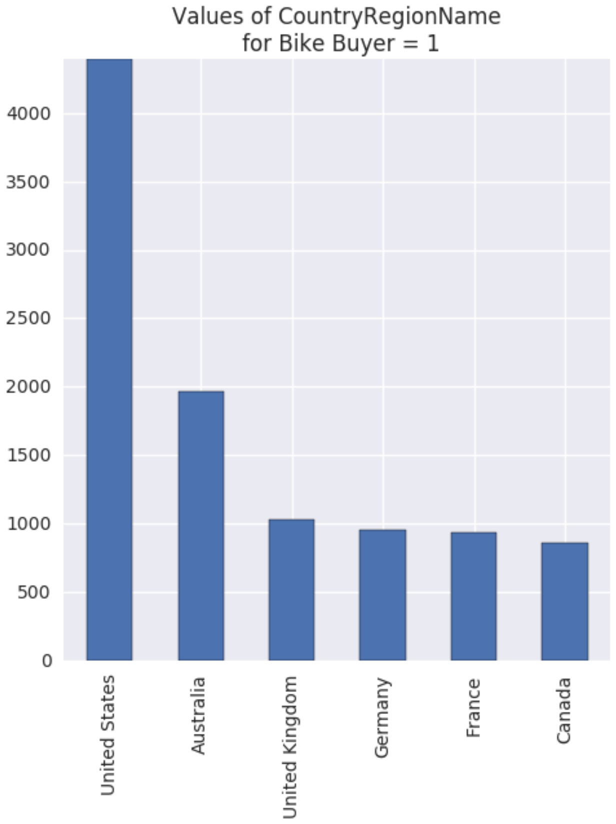
Gender is particularly useful for predictions, it clearly shows than Male Customers than bought a bike are more than Female Customers. It’s also useful to estimate the Average Monthly Spend for new customers because the boxplot shows a high variance and median for the Male Customers.

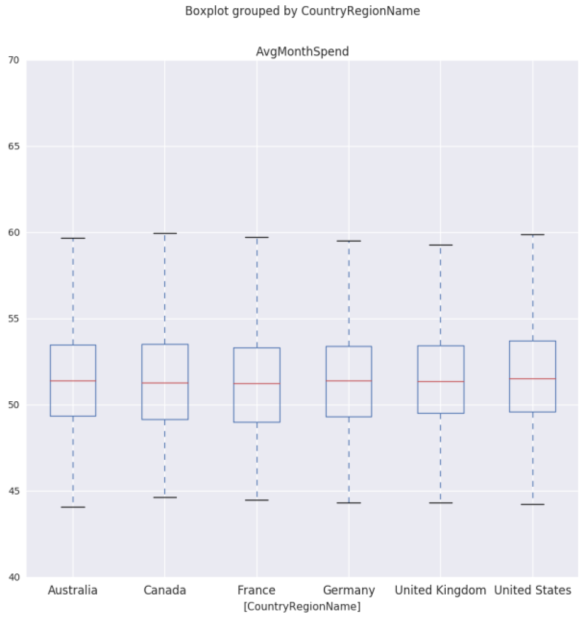




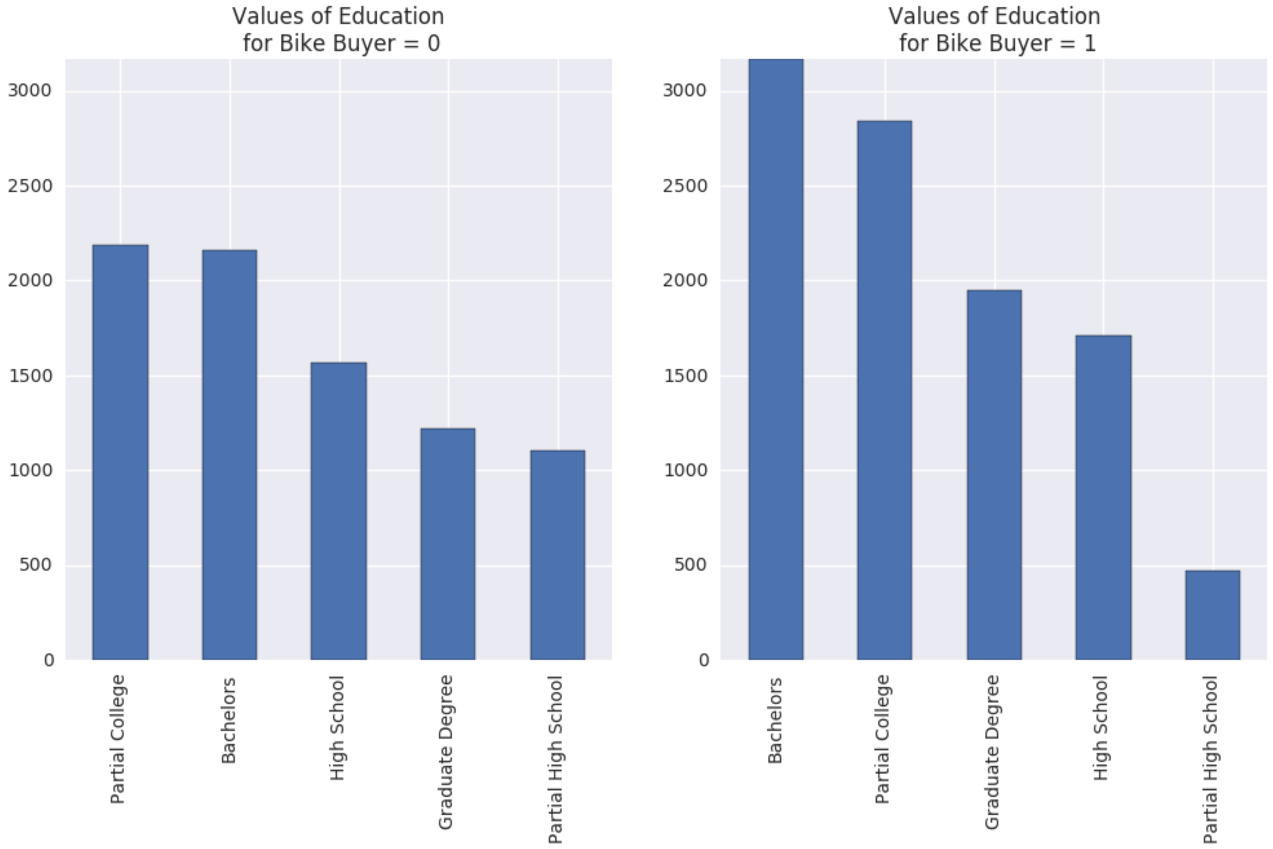
MaritalStatus shows than the Married customers that have purchased a bike are more than Singles customers, and Singles has as a tiny higher median of Average Month Spend.

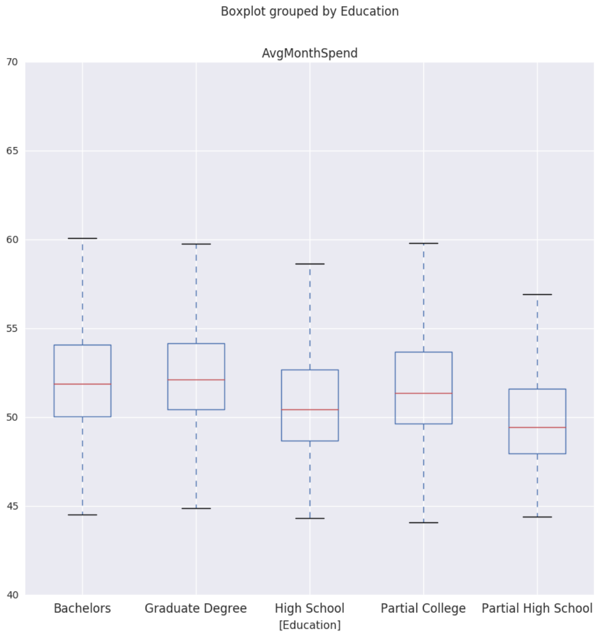




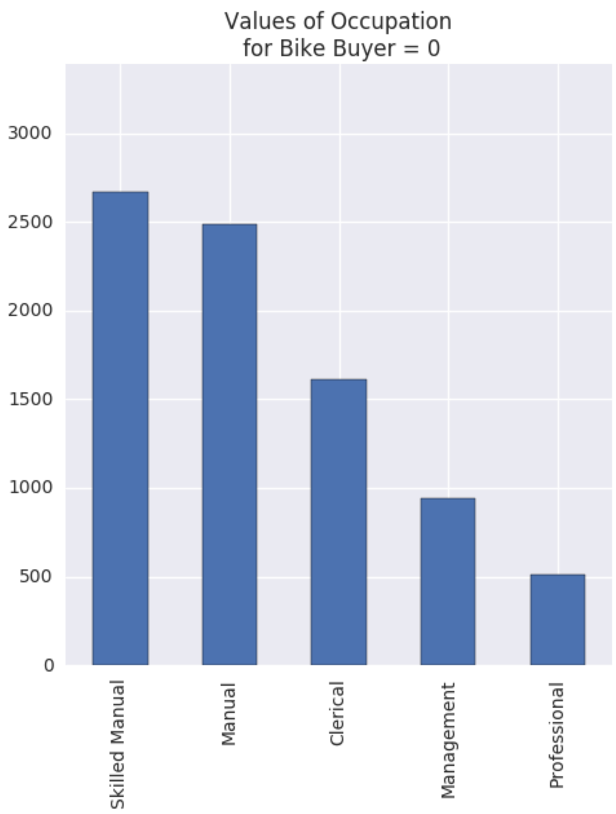


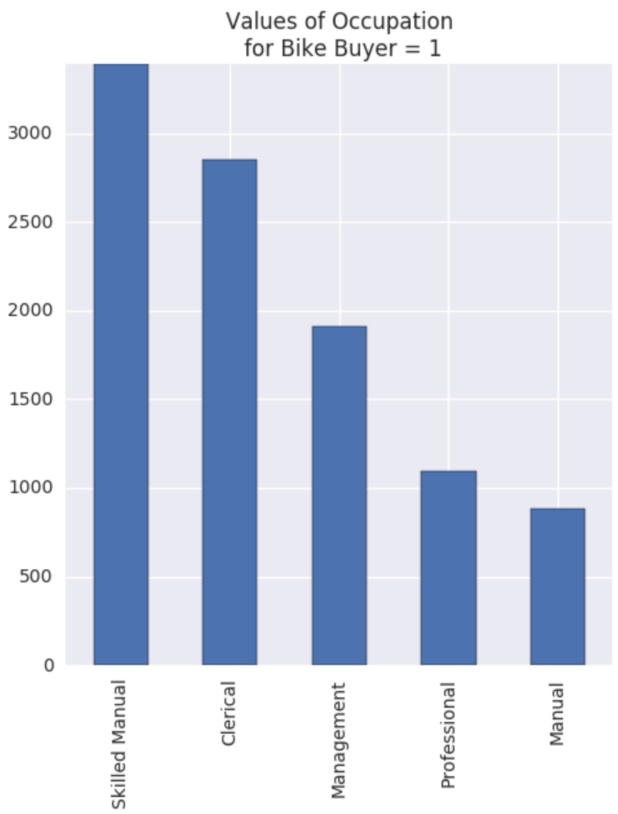
CountryRegion shows than the most of Customers are from United States and Australia, however, the proportion of BikeBuyer and No-BikeBuyer are so similar, and the Average Monthly Spend are so close relate to each other countries. This feature is useless for predict new customers purchase and also average monthly spend.

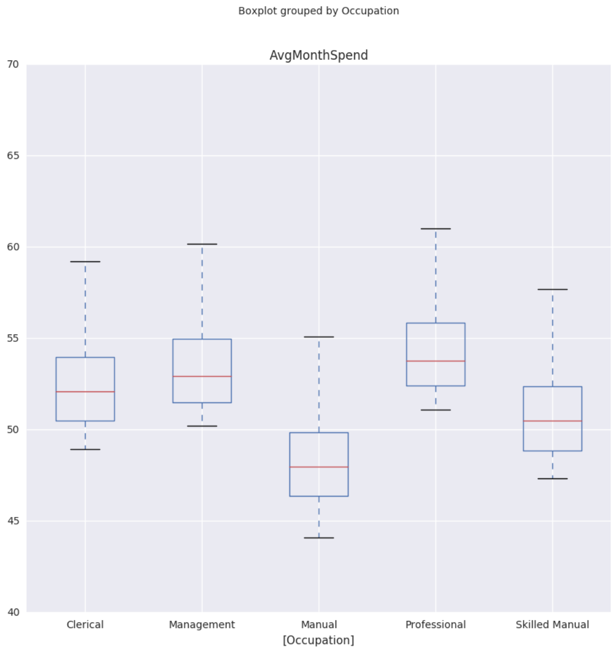




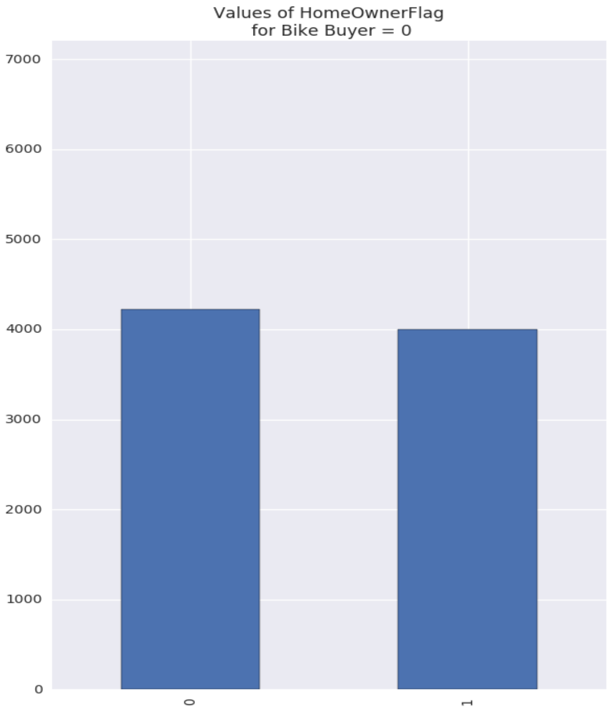
Education views show than the Customers with High and Medium levels of education have more probabilities to purchase a bike and they spend more than Customer with Low level of education. This intuition will be confirmed with the Occupation's views.

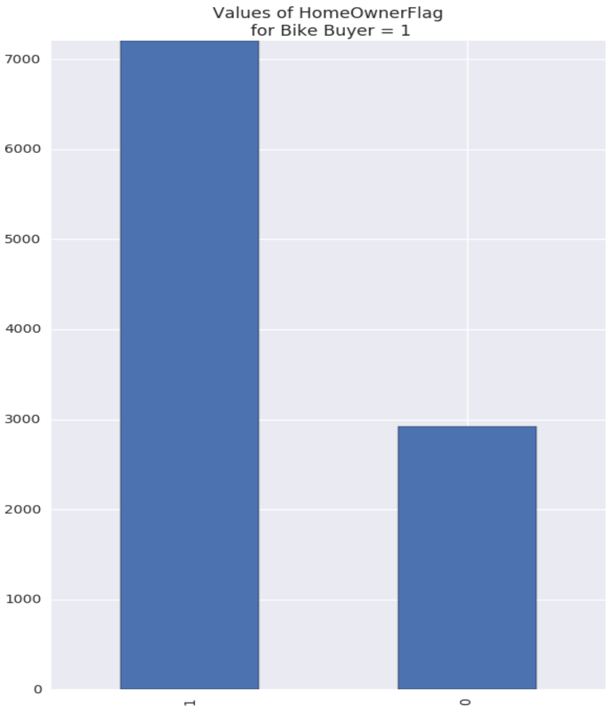


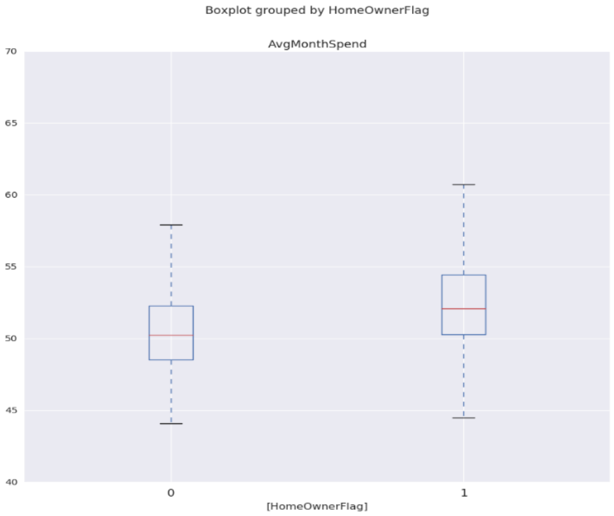




Occupation views show a higher BikeBuyer rate for Skilled Manual, Clerical and Manager Customers. It confirms A greater spending capability translates into a higher Average Monthly Spend. This feature is useful for new Customers prediction and for the estimate for the Average Monthly Spend.







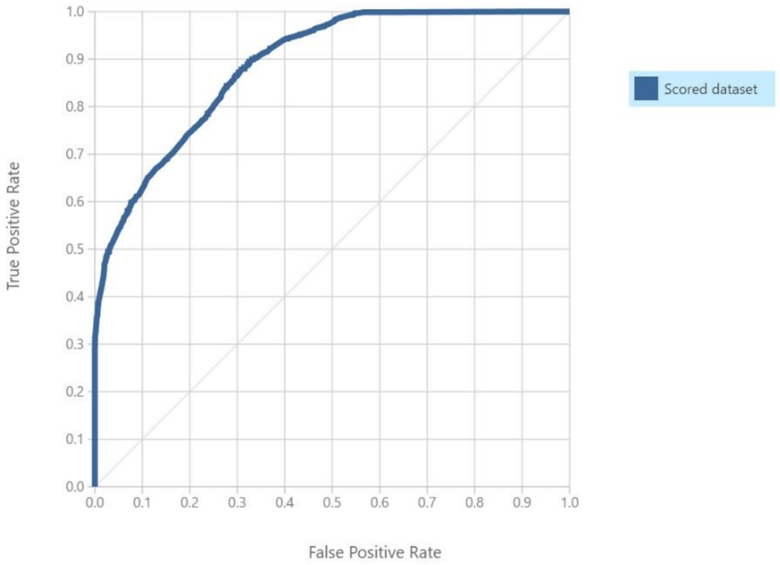
HomeOwnerFlag shows Customer with a house, have much more probabilities for purchase a bike and as it suggests the boxplot, they spend more than Customers without a house. This feature is useful for purchase prediction and average spend estimation.

Classifications of new potential Bike Buyer

Based on the analysis of the customer data, a predictive model to classify new customers purchase. The model was created using the **Two-Class Boosted Decision Trees** algorithm and trained with 70% of the data. Testing the model with the remaining 30% of data yelded the following results:

- **True Positives**: 2712
- **True Negatives**: 1678
- **False Positives**: 797
- **False Negatives**: 319

The **Received Operator Characteristic** (ROC) curve for the model is shown here, with the blue line indicating the model's performance at varying classification threshold values, and the diagonal line showing the expected results of a random guess:



This translates in to the following standard performance metrics for classification:

- **Accuracy**: 79.7%
- **Precision**: 77.3%
- **Recall**: 89.5%
- **F1 Score**: 82.9%
- **AUC**: 88.7%

Regression of Average Monthly Spend

After creating a classification model to predict if the new customers purchase a bike, I build a regression model to predict the average monthly spend for them. Based on the apparent relationships identified when analyzing data, a **Boosted Decision Tree Regression** model was created to predict the average monthly spend. The model was trained with 70% of the data, and tested with the remaining 30%.

This translates in to the following standard performance metrics for regression:

- **Mean Absolute Error:** 1.41947
- **Root mean Squared Error:** 1.88696
- **Relative Absolute Error:** 0.547801
- **Relative Squared Error:** 0.325004
- **Coefficient of Determination:** 0.674996

Conclusion

This analysis has shown that the purchasing of new customers and the average monthly spend, can be predicted from its characteristics. In particular, **Gender**, **MaritalStatus**, **Occupation**, **Education**, **HomeOwnerFlag** and **YearlyIncome** are useful for both the predictions. Features that suggest the degree of economic wellbeing such **Occupation**, **HomeOwnerFlag** or **NumberCarsOwned** are useful for the estimation of Average Monthly Spend.