Analysis of customer spend on bikes

Seema G, April, 2017

Overview

The Adventure Works Cycles Company collected a large volume of data about their existing customers, including demographic features and information about purchases they have made. The company is interested in analyzing customer data to determine any apparent relationships between demographic features and the likelihood of a customer purchasing a bike. Additionally, the company wants to know whether a customer's average monthly spend on bikes with the company can be predicted from known customer characteristics.

Executive summary

This document presents an analysis of the likelihood of a customer buying a bike, and the average monthly spend on bikes based on 18355 observations of customer data relating to yearly income, education, occupation, gender, marital status, homeowner flag, number of cars owned, number of children, and age information.

Initial step involved exploring data, gathering summary statistics and creating visualizations of data to identify potential relationships between customer characteristics and customers who bought a bike as well as average monthly spend on bikes. After exploring the data, a classification model was created to predict whether or not a customer will purchase a bike. Then a regression model was created to predict the average monthly spend on bikes.

After performing the analysis, in conclusion, while many factors can help predict the likelihood of a customer buying a bike, and the average monthly spend on bikes, significant features found in this analysis were:

- **Yearly Income**: Customers who bought a bike tend to have higher yearly income than those who did not buy a bike. Customers with higher yearly income tend to have higher average monthly spend on bikes.
- **Gender**: Male customers bought more bikes than female customers. Male customers tend to have higher average monthly spend on bikes than female customers.
- **Age**: Male customers in the 30-50 age groups have higher average monthly spend on bikes than customers in the other age groups.
- **Number of Children**: Most non-bike buyers did not have any children at home. Customers with one or more children at home tend to have higher average monthly spend than customers without any children.
- **Marital Status**: Married customers bought more bikes than single customers.
- **Home Owner**: Home owners tend to have higher average monthly spend than non-home-owners.
- **Number Cars Owned**: Bike-buyers tend to own more cars than non-bike-buyers. Customers that owned more cars tend to have higher average monthly spend on bikes.

- **Occupation**: Customers in the clerical and skilled manual occupations tend to buy more bikes than customers in other occupations. Customers in the manual occupation tend to have lower average monthly spend than customers in other occupations.

Overview of steps

Appendix A1 has a description of data. Dependent variables are Bike Buyer flag (also referred to as BikeBuyer) for classification and Average Monthly Spend on bikes (also referred to as AvgMonthSpend) for regression.

1. **Data preparation**: This step involved reviewing and cleaning the data by replacing any missing values and removing duplicate rows. In this dataset, each customer is identified by a unique customer ID.
2. **Data exploration**: This step involved exploring the data by calculating summary and descriptive statistics for the features in the dataset, identifying correlations between features, and creating data visualizations to determine apparent relationships in the data.
3. **Feature engineering**: This involved calculating age from birthdate information and exploring correlations of dependent variables with age.
4. **Feature selection**: This step involved identifying the set of independent variables that the dependent variables BikeBuyer and AvgMonthSpend are correlated to in order to utilize these features for classification and regression modelling in predicting the values of dependent variables BikeBuyer and AvgMonthSpend.
5. **Feature scaling**: Normalizing numeric features to ensure the coefficients are on the same scale.
6. **Classification modeling**: Creating classification models that predict whether or not a customer will purchase a bike.
7. **Regression modeling**: Creating regression models to predict the value of the dependent variable Average Monthly Spend on bikes.
8. **Comparing and tuning model**: Comparing models and choosing the one with the highest accuracy for classification, and the one with the lowest root mean square error for regression. Finally analyzing and tuning the highest performing model to predict the dependent variable values for the test data.

Data preparation

This step involved fixing missing values, removing duplicates, and ensuring data was classified correctly.

Initial Data Exploration

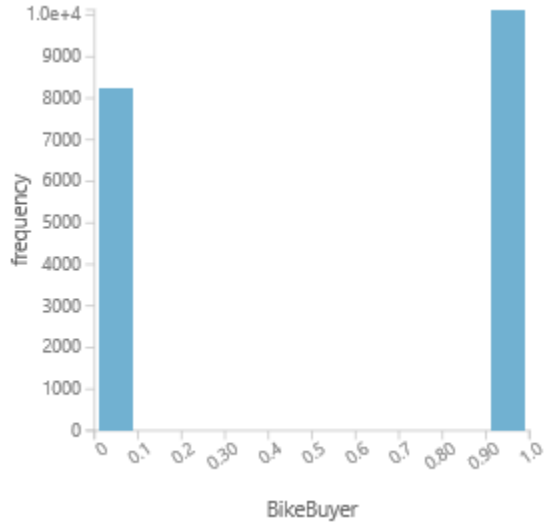
Appendix A1 has a description of data.

Initial data exploration began with summary and descriptive statistics. Summary statistics of the dependent variables BikeBuyer and AvgMonthSpend is as follows. BikeBuyer refers to customers who bought a bike. AvgMonthSpend refers to average monthly spend on bikes.

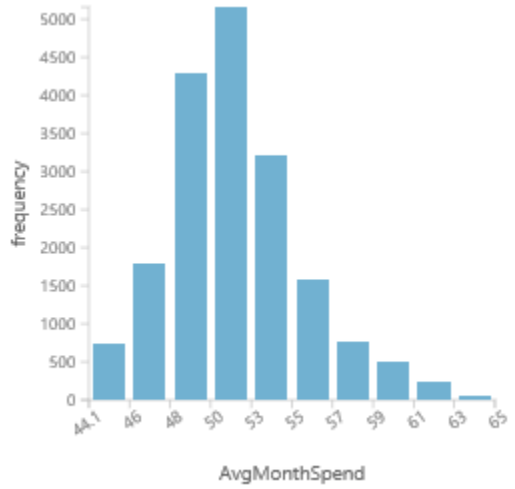
Dependent	Column	Type	Min	Max	Mean	Median	Std Dev

variable for							
Classification	Bike Buyer	Flag (0 or 1)	0	1	0.55	1	0.49
Regression	Avg Monthly Spend	Numeric continuous	44.1	65.29	51.7672	51.42	3.438

Histogram of BikeBuyer



Histogram of AvgMonthSpend



Following tables have summary statistics for independent features.

Column	Type	Min	Max	Mean	Median	Std Dev	DCount
YearlyIncome	Numeric	25435	139115	72759	61851	30687	15355
NumberCarsOwned	Numeric	0	5	1.27	1	0.91	6
NumberChildrenAtHome	Numeric	0	3	0.33	0	0.56	4
TotalChildren	Numeric	0	3	0.85	0	0.92	4

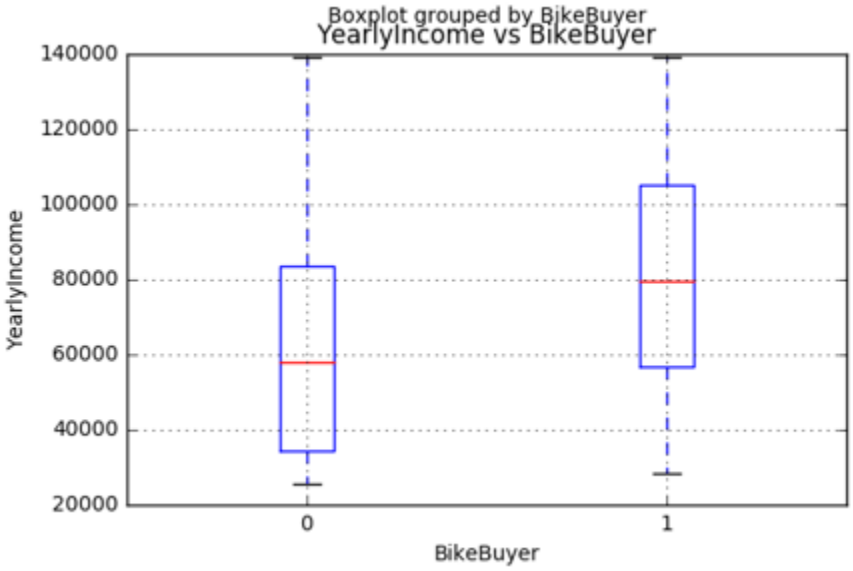
Age	Numeric	16	87	35.2	34	11.2	71

Column	Type	Number of Categories	Values
Gender	Categorical	2	M (51%), F (49%)
Marital status	Categorical	2	M (54%), S (46%)
Occupation	Categorical	5	Skilled Manual (33%), Clerical (24%), Manual (18%), Management (16%), Professional (9%)
Home Owner Flag	Categorical	2	Yes (61%), No (39%)

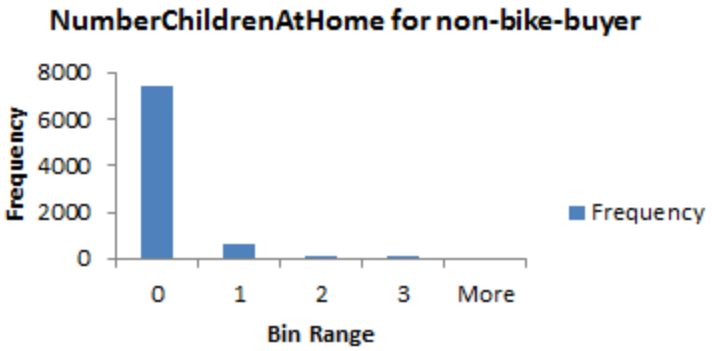
Bike Buyer Profile and Correlations

After exploring the individual features, an attempt was made to identify relationships between features in the data, in particular, between BikeBuyer and the other features.

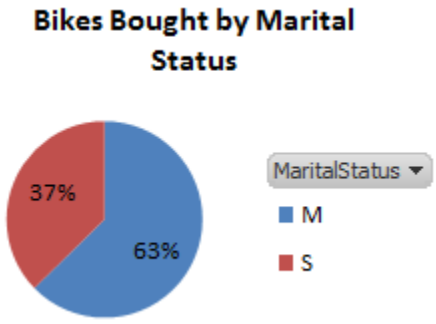
- Based on the Bike Buyer histogram above, more customers have bought bikes than have not bought bikes.
- The median yearly income for customers who bought a bike is higher than the median yearly income for customers who did not buy a bike.



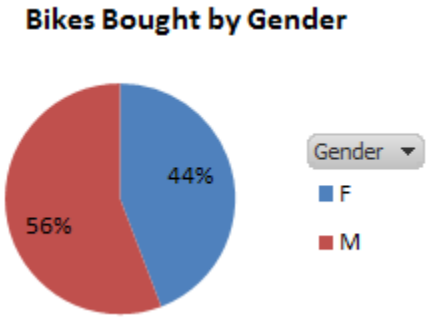
- Most non-bike-buyers did not have any children at home.



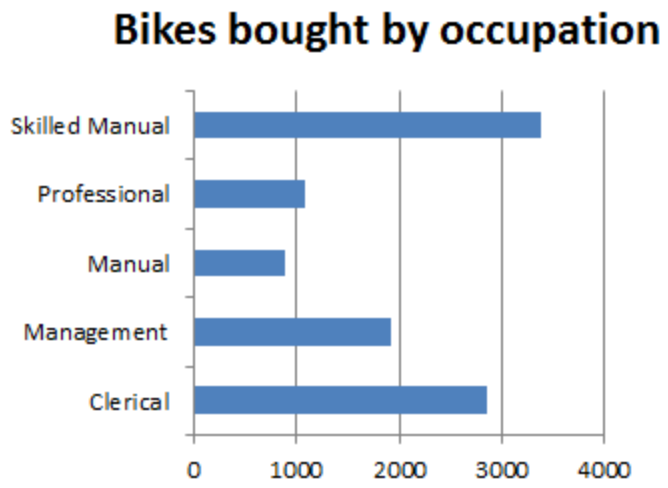
- Married customers bought more bikes than single customers.



- Male customers bought more bikes than female customers.



- When reviewing number of bikes bought by occupation, it was found that customers in skilled manual and clerical professions bought more bikes than others.



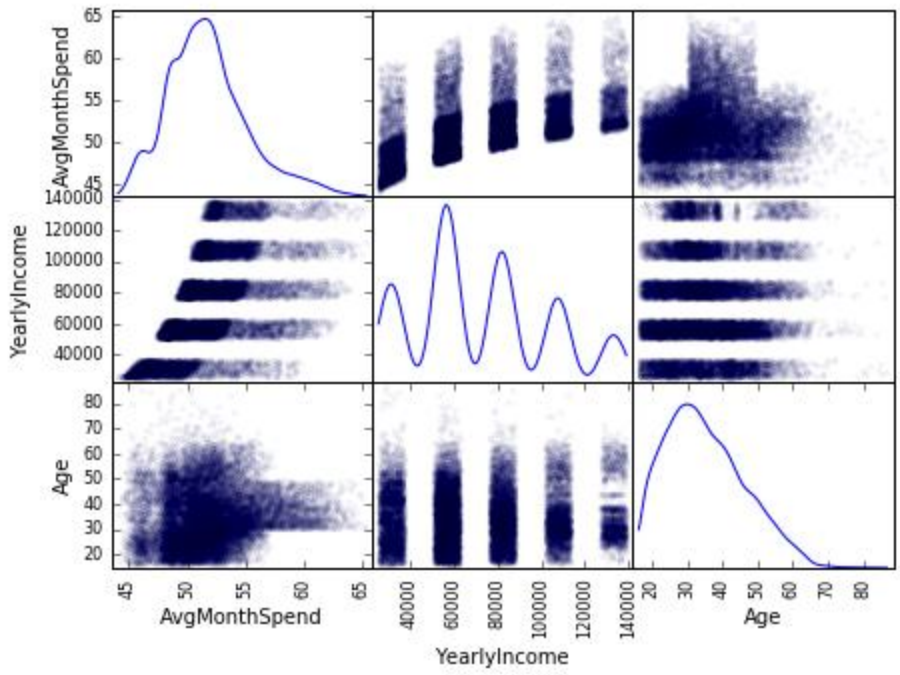
Summary of Bike Buyer Relationships

- Most non-bike-buyers did not have any children at home.
- The median yearly income is higher for customers who bought a bike than for customers who didn't.
- The most common occupation for customers who bought a bike is skilled manual.
- Male customers bought more bikes than female customers.
- Married customers bought more bikes than single customers.
- The median number of cars owned by customers who bought a bike is higher than for customers who didn't buy a bike.

Average Monthly Spend Correlations

This step identifies relationships between features in the data, in particular, between AvgMonthSpend and the other features.

- A scatter plot matrix of numerical features AvgMonthSpend, YearlyIncome, and Age shows that AvgMonthSpend and YearlyIncome are correlated. Based on the correlation matrix AvgMonthSpend has a relatively smaller correlation with NumberCarsOwned, NumberChildrenAtHome, and Age.



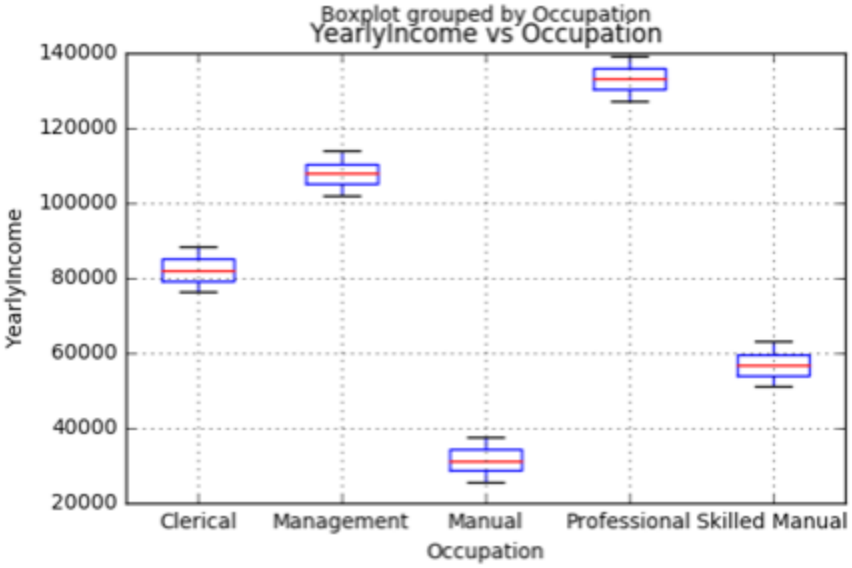
Correlation coefficients for numeric features are as follows:

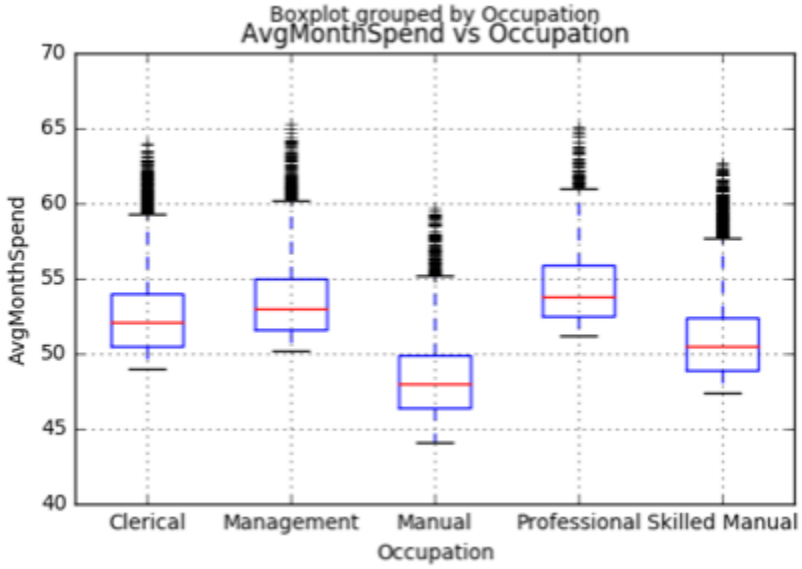
	AvgMonthSpend	YearlyIncome	NumberCarsOwned
AvgMonthSpend	1.000000	0.530120	0.275498
YearlyIncome	0.530120	1.000000	0.477301
NumberCarsOwned	0.275498	0.477301	1.000000
NumberChildrenAtHome	0.145095	0.005967	0.020476
TotalChildren	0.026108	0.021885	0.030165
Age	0.111714	0.026656	0.042027

	NumberChildrenAtHome	TotalChildren	Age
AvgMonthSpend	0.145095	0.026108	0.111714
YearlyIncome	0.005967	0.021885	0.026656
NumberCarsOwned	0.020476	0.030165	0.042027
NumberChildrenAtHome	1.000000	0.606142	0.326599
TotalChildren	0.606142	1.000000	0.548607
Age	0.326599	0.548607	1.000000

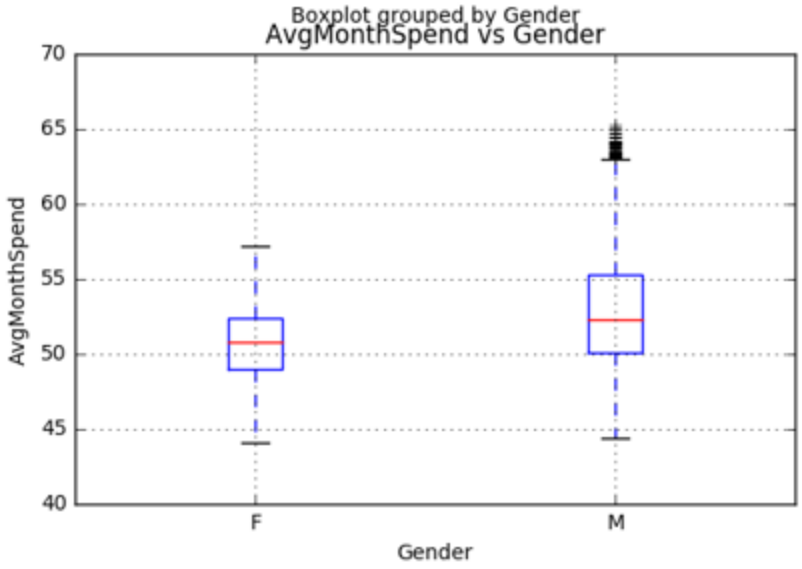
Following plots were generated to identify correlations between AvgMonthSpend and features like yearly income, occupation, gender, number of cars owned, marital status, number of children, and age.

- Median YearlyIncome and AvgMonthSpend by occupation vary as follows from lowest to highest: Manual, Skilled Manual, Clerical, Management, and Professional. Customers in the manual occupation had lower AvgMonthSpend compared to others.

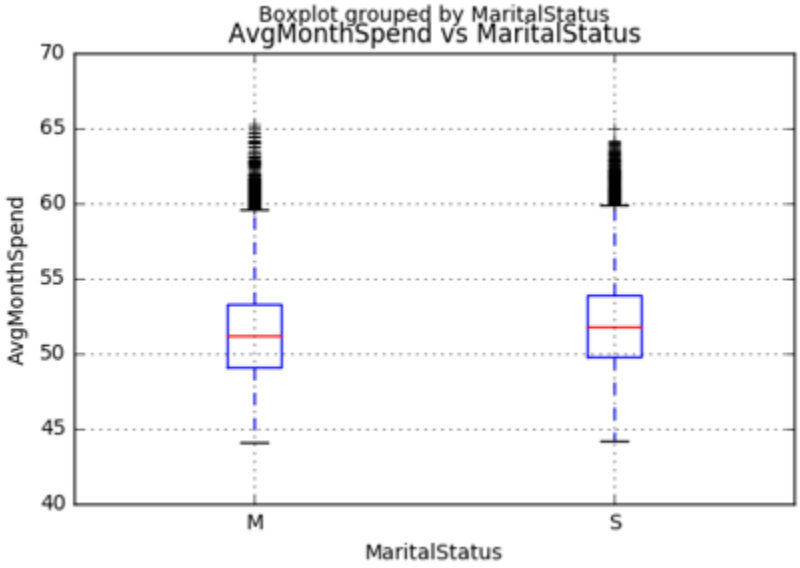




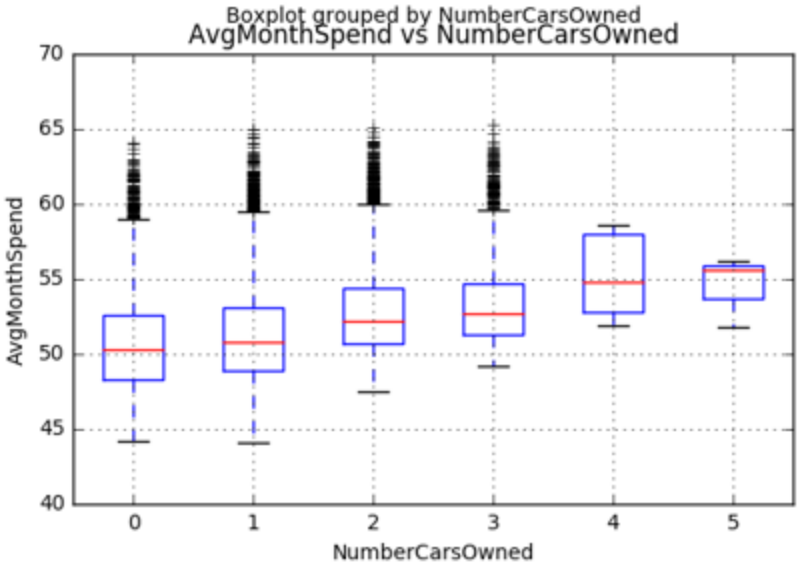
- Male customers have higher median AvgMonthSpend and a wider spread than female customers. The maximum AvgMonthSpend for female customers was 57 whereas it was over 65 for male customers.



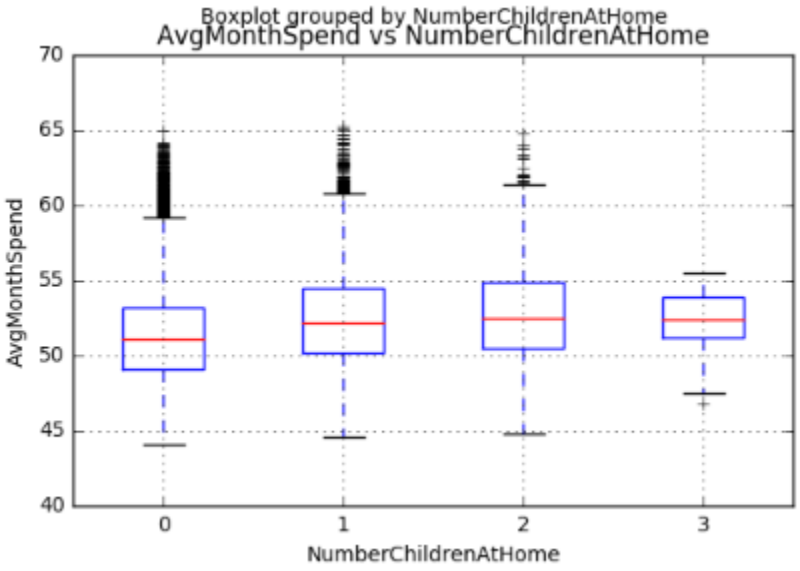
- Single customers have a slightly higher median AvgMonthSpend than married customers.



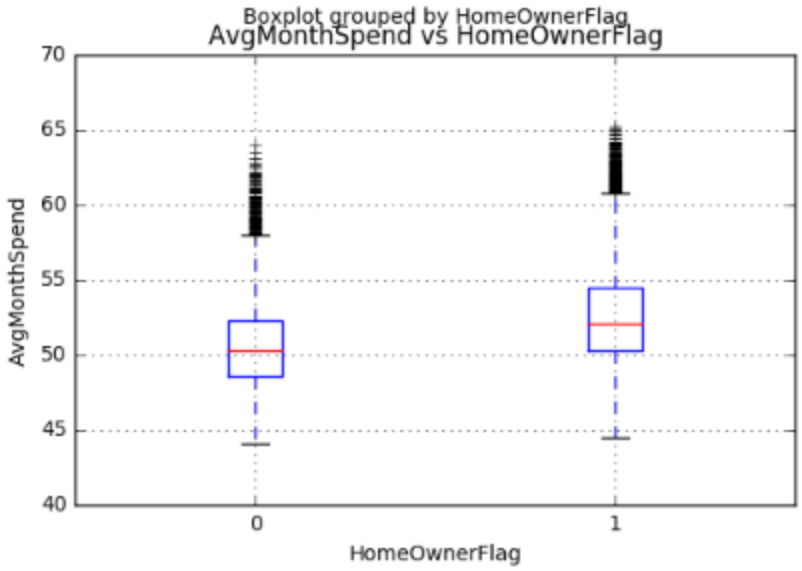
- Customers who owned more cars have higher median AvgMonthSpend on bikes.



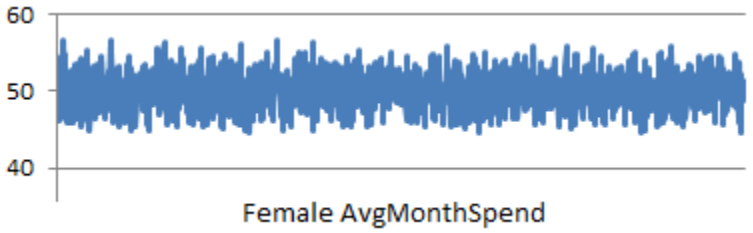
- Customers with one or more children at home have higher median AvgMonthSpend than customers with no children.

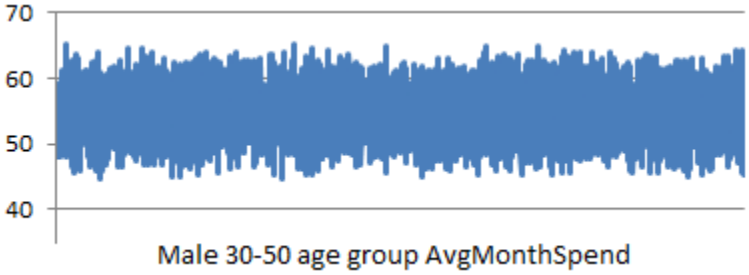


- Home owners have higher median AvgMonthSpend than non-home owners.



- When reviewing AvgMonthSpend by gender and age group, it was observed that females AvgMonthSpend is in the 45-57 range approximately. Males aged less than 30 years and more than 50 years have similar spending pattern. However, AvgMonthSpend is distinctly higher for males in the 30-50 age groups. Based on this insight, age was added as a feature for the dependent variables.





Summary of Average Monthly Spend Relationships

- Male customers have higher median AvgMonthSpend and a wider spread than female customers. AvgMonthSpend is distinctly higher for males in the 30-50 age groups. The maximum AvgMonthSpend for female customers was 57 whereas it was over 65 for male customers.
- Single customers have higher median AvgMonthSpend than married customers.
- Average Monthly Spend is directly correlated to yearly income. Median yearly incomes as well as AvgMonthSpend by occupation are in the following order from lowest to highest: Manual, Skilled

Manual, Clerical, Management, and Professional. Customers in the manual occupation had lower AvgMonthSpend compared to others.

- Home owners have higher median AvgMonthSpend than non-home owners.
- Customers with cars have higher median AvgMonthSpend than customers with no cars.
- Customers with one or more children at home have higher median AvgMonthSpend values than customers with no children.

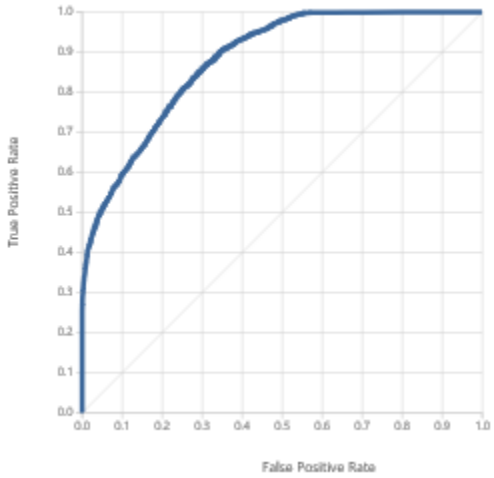
Bike Buyer Classification for new customers

Based on the analysis of bike buyer, a classification model was created that predicts whether or not a customer will purchase a bike. The model predicts bike purchasing for new customers for whom no information about average monthly spend or previous bike purchases is available. Based on the apparent relationships identified when analyzing the data, a classification model was created to predict Bike Buyer information from yearly income, number of children at home, age, number of cars owned, gender, marital status, homeowner flag, and occupation. Of the several machine learning algorithms considered, the model with the highest accuracy was chosen.

The model was created using Two-Class Boosted Decision Trees algorithm and trained with 60% of data. Testing the model with the remaining 40% of data yielded the following results:

True Positive	False Negative	Accuracy	Precision
3363	661	0.784	0.785
False Positive	**True Negative**	**Recall**	**F1 Score**
923	2395	0.836	0.809

The classification model predicted BikeBuyer labels and scored probabilities for the test data. The Receiver Operator Characteristic (ROC) curve for the model is shown below with the blue line indicating the model's performance. The area under the curve (AUC) is 0.880. Accuracy is 78.4%.



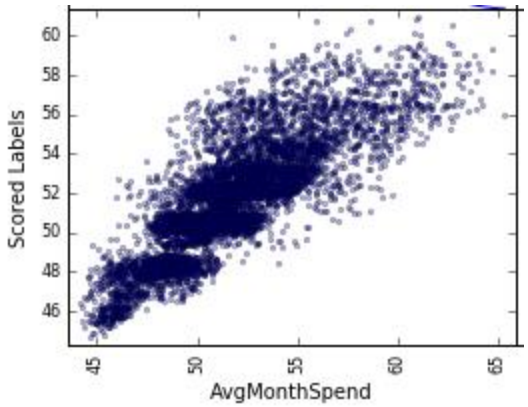
A comparative analysis was done using Two-Class Logistic Regression to identify feature coefficients for classification. Accuracy for Two-Class Logistic Regression was 70%. Two-Class Boosted Decision Trees classification model is used for BikeBuyer predictions as it has higher accuracy of 78.4% compared to 70% for Two-Class Logistic Regression. Following feature coefficient matrix for Two-Class Logistic Regression shows that the dominant features for classification were NumberChildrenAtHome and YearlyIncome which is supported by the observations in the data.

Feature	Weight
NumberChildrenAtHome	5.90561
YearlyIncome	2.38311
Age	-1.05841
Bias	-0.903847
NumberCarsOwned	0.828696
Occupation_Professional_3	-0.68513
Gender_F_0	-0.560591
Occupation_Skilled Manual_4	0.513162
Occupation_Manual_2	-0.470715
Occupation_Management_1	-0.265193
HomeOwnerFlag_0_0	-0.261768
Occupation_Clerical_0	0.239284
MaritalStatus_M_0	-0.0596182
HomeOwnerFlag_1_1	0.0382288
Gender_M_1	0.0118879
MaritalStatus_S_1	-0.000321735

Regression to predict Average Monthly Spend for new customers

After creating a classification model to predict whether or not a customer will purchase a bike, a regression model to predict the AvgMonthSpend on bikes was created. The model predicts average monthly spend for new customers for whom no information about average monthly spends or previous bike purchases are available. Of the several machine learning algorithms considered, the algorithm with the lowest root mean squared error was considered. Based on the apparent relationships identified when analyzing the data, a Boosted Decision Tree Regression model was created to predict AvgMonthSpend from yearly income, age, gender, marital status, homeowner flag, occupation, number of children at home, and number of cars owned.

The model was trained with 70% of data and tested with the remaining 30%. A scatter plot showing the actual Avg Monthly Spend and the predicted monthly spend is shown below.



The plot shows a linear relationship between the predicted and actual values in the test data. The Root Mean Square Error (RMSE) for the test results is 1.93. This indicates that on average the model prediction is within 1.93 units of the actual monthly spend value. The coefficient of determination is 0.68. Model evaluation metrics are as follows:

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
1.428768	1.93585	0.541505	0.319939	0.680061

A comparative analysis using Linear Regression was done to analyze feature coefficients. Linear Regression yielded a Root Mean Square Error (RMSE) of 2.48 and a coefficient of Determination of 0.47 using an L2 regularization weight of 0.001. Boosted Decision tree regression algorithm is preferred and used for predictions as it has lower Root Mean Square Error of 1.93 compared to 2.48 for Linear Regression. Since male customers in the 30-50 age groups had higher average monthly spend compared to others, a new age-group category called Catage_x-y (for x <= age <= y) was added and is referred to below. Following table has feature coefficients information for linear regression. These are consistent with the observations in the data.

Feature	Weight
Gender_M_1	9.77126
MaritalStatus_S_1	9.51003
HomeOwnerFlag_1_1	8.70138
HomeOwnerFlag_0_0	8.65875
MaritalStatus_M_0	7.85009
Gender_F_0	7.58887
Catage_30-50_3	5.11204
Occupation_Skilled Manual_4	4.31923
Occupation_Clerical_0	4.00057
Occupation_Manual_2	3.62318
Catage_26-29_2	3.22095
Occupation_Management_1	3.19356
Catage_19-25_1	3.16956
Catage_18orless_0	3.06719
Catage_51ormore_4	2.79039
YearlyIncome	2.28772
Occupation_Professional_3	2.2236
NumberChildrenAtHome	1.3142
Age	0.0503319
NumberCarsOwned	0.0195824

Conclusion

This analysis has shown that the likelihood of a customer purchasing a bike and average monthly spend on bikes can be predicted from customer information. In particular customer's yearly income, gender, marital status, homeowner flag, number of children at home, age, number of cars owned, and occupation can determine whether a customer will buy a bike or not, and the amount of average monthly spend on bikes.

Appendix

A1: Description of data

This data consists of customer demographic data consisting of the following fields:

- **CustomerID** (*integer*): A unique customer identifier.
- **Title** (*string*): The customer's formal title (Mr, Mrs, Ms, Miss Dr, etc.)
- **FirstName** (*string*): The customer's first name.
- **MiddleName** (*string*): The customer's middle name.
- **LastName** (*string*): The customer's last name.
- **Suffix** (*string*): A suffix for the customer name (Jr, Sr, etc.)
- **AddressLine1** (*string*): The first line of the customer's home address.
- **AddressLine2** (*string*): The second line of the customer's home address.
- **City** (*string*): The city where the customer lives.

- **StateProvince** (*string*): The state or province where the customer lives.
- **CountryRegion** (*string*): The country or region where the customer lives.
- **PostalCode** (*string*): The postal code for the customer's address.
- **PhoneNumber** (*string*): The customer's telephone number.
- **BirthDate** (*date*): The customer's date of birth in the format YYYY-MM-DD.
- **Education** (*string*): The maximum level of education achieved by the customer:
    - Partial High School
    - High School
    - Partial College
    - Bachelors
    - Graduate Degree
- **Occupation** (*string*): The type of job in which the customer is employed:
    - Manual
    - Skilled Manual
    - Clerical
    - Management
    - Professional
- **Gender** *(string):* The customer's gender (for example, M for male, F for female, etc.)
- **MaritalStatus** *(string):* Whether the customer is married (M) or single (S).
- **HomeOwnerFlag** *(integer):* A Boolean flag indicating whether the customer owns their own home (1) or not (0).
- **NumberCarsOwned** *(integer):* The number of cars owned by the customer.
- **NumberChildrenAtHome** *(integer):* The number of children the customer has who live at home.
- **TotalChildren** *(integer):* The total number of children the customer has.
- **YearlyIncome** *(decimal):* The annual income of the customer.
- **LastUpdated** *(date):* The date when the customer record was last modified.

Sales data for existing customers, consisting of the following fields:

- **CustomerID** *(integer):* The unique identifier for the customer.
- **BikeBuyer** *(integer):* A Boolean flag indicating whether a customer has previously purchased a bike (1) or not (0).
- **AvgMonthSpend** *(decimal):* The amount of money the customer has spent with Adventure Works Cycles on average each month.