Analysis of customer data

Will new customer likely buy a bike?
Predict average monthly spend of a new customer?

Executive Summary

This document presents an analysis of Adventure Works Cycles internal database concerning customers in respect to their likelihood of purchasing a bike as well as expected average monthly spend. The analysis is based on 18355 unique records of customers. Each record describes a customer in multiple demographic dimensions.

The initial data analysis revealed some noteworthy characteristics, namely a very strong relationship between Yearly Income and Occupation groups with not only zero overlap but in fact clear separation between the fields (several thousands of dollars on the income scale are not used between neighboring groups). Further, part of the dataset seems to be effected by unknown preconditions. This results in less accurate results in this assessment.

The most likely predictors for whether a customer will buy a bike are:

- Number Cars Owned
- Number Children At Home
- Occupation
- Total Children
- Marital Status
- Home Owner

The most likely predictors for the average monthly spend of a customer are:

- Yearly Income
- Age
- Gender
- Marital Status
- Number Children At Home
- Number Cars Owned.

The estimated monthly spend in this assessment appears to be generally lower then the actual monthly spend. This distortion is due to the existing and unknown precondition. Nonetheless, it is advised to use the results of this assessment with ample caution.

Initial Data Exploration

The initial data exploration began with some summary and statistics. I feature ‘age’ was added to the dataset in which the age of the customer based on the DOB was calculated.

Individual Numerical Feature Statistics

Statistical summary for the numeric columns over all 18355 unique observations:

Column	Min	Max	Mean	Median	Std Dev	DCount
# Cars Owned	0	5	1.270390	1	0.913887	6
# Children at home	0	3	0.338218	0	0.569001	4
# Total Children	0	3	0.850449	0	0.927363	4
Yearly Income	25435	139115	72758.950041	61851	30687.664358	15355
Avg. Monthly Spend	44.10	65.29	51.767207	51.42	3.438024	1803
Age	16	86	34.729719	33	11.255742	70

Flag Feature Statistic

Additionally, two columns contain numerical values that are used as indicators, summarized here:

Column	Min	Max	Mean	Median	Std Dev	DCount
Home Owner	0	1	0.610569	1	0.487634	2
Bike Buyer	0	1	0.55173	1	0.49733	2

Categorical Features:

Additional to the numeric features, the data includes categorical features, including:

- **Education** - Partial High School, High School, Partial College, Bachelors, Graduate Degree
- **Occupation** – Manual, Skilled Manual, Clerical, Management, Professional
- **Gender** - M for male or F for female
- **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.
- **PostalCode** - The postal code for the customer's address.

Informal Features:

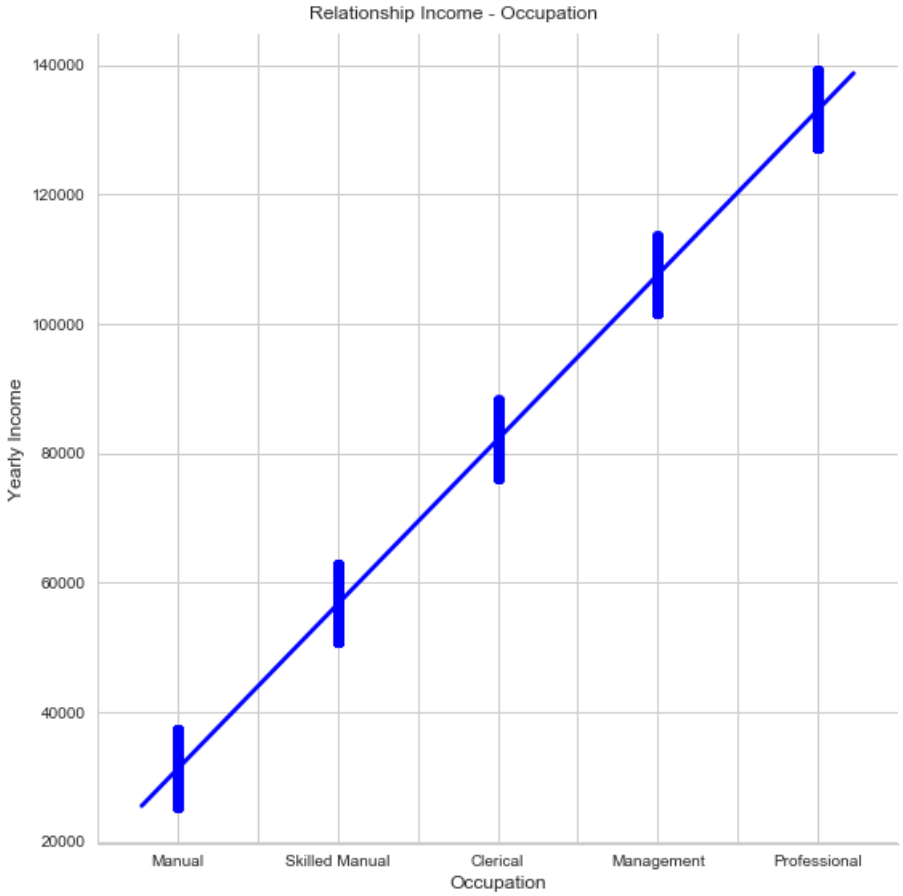
Further, the data contained columns of informal character for this assessment:

- **CustomerID** - A unique customer identifier.
- **Title** - The customer's formal title (Mr, Mrs, Ms, Miss Dr, etc.)
- **FirstName** - The customer's first name.
- **MiddleName** - The customer's middle name.
- **LastName** - The customer's last name.
- **Suffix** - A suffix for the customer name (Jr, Sr, etc.)
- **AddressLine1** - The first line of the customer's home address.
- **AddressLine2** - The second line of the customer's home address.
- **LastUpdated** - The date when the customer record was last modified.

Apparent Relationships

The dataset is remarkable in respect to some very distinct relationships between features.

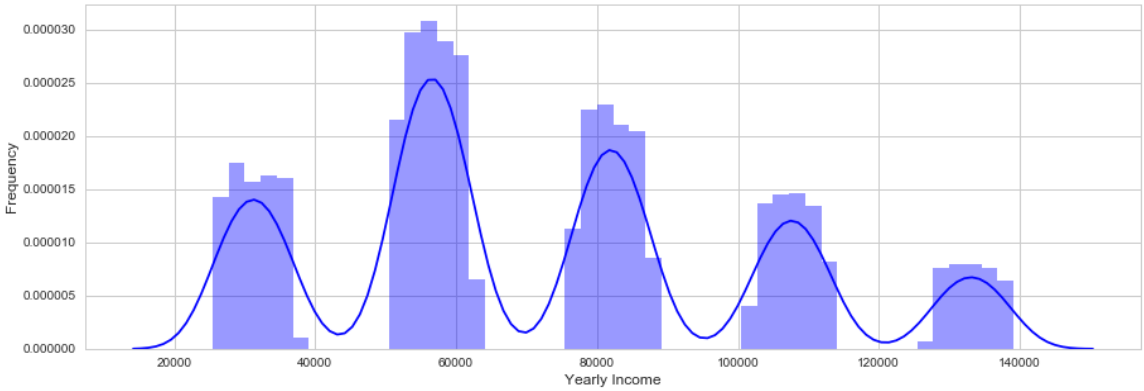
Linear distinct relationship Occupation / Yearly Income



Not only exists a linear relationship, but also, the "income classes: are very distinctly separated. This shows in the statistical summary for grouped-by Occupation data:

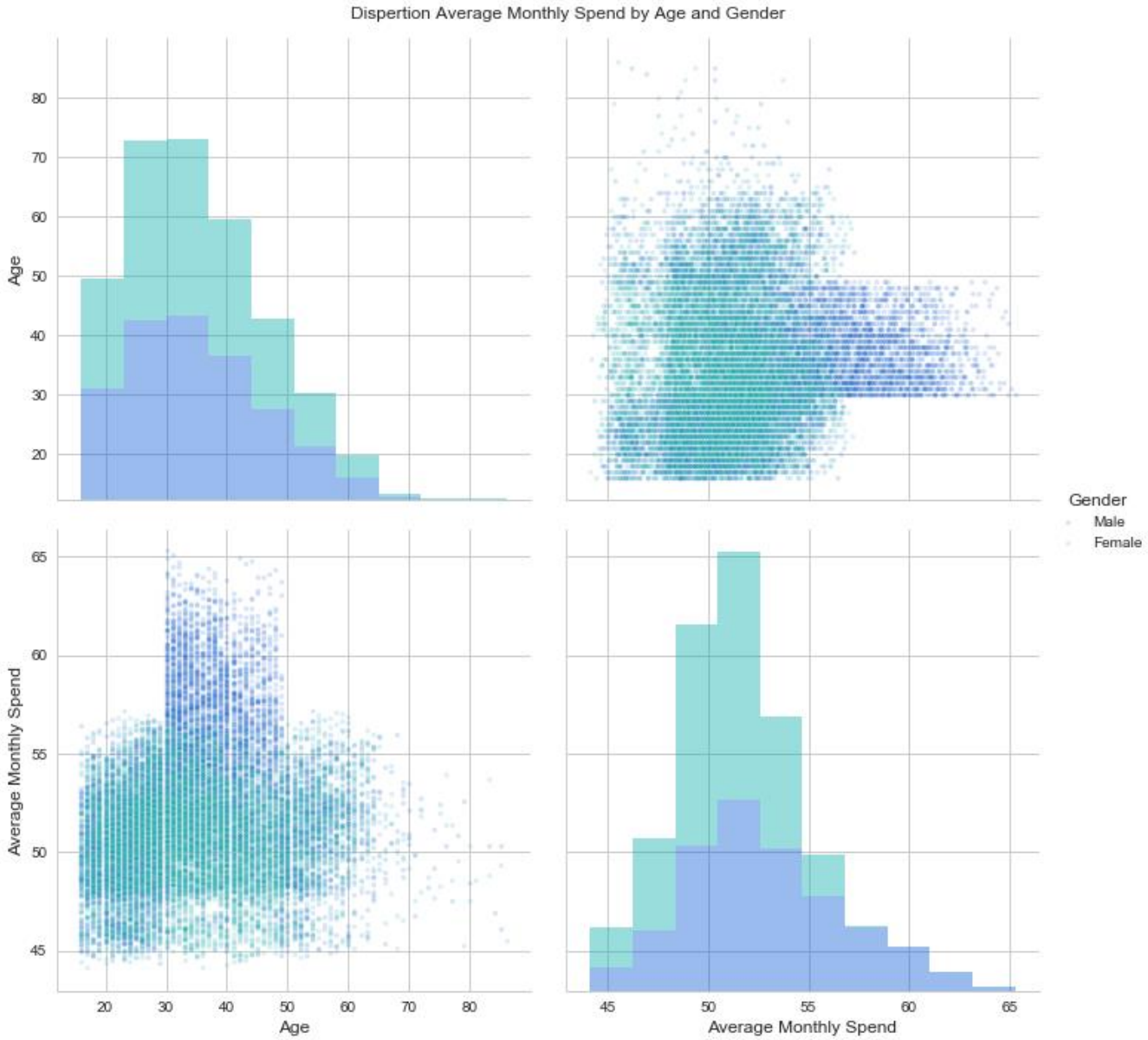
Occupation	Min	Max	Mean	Median	Std. Deviation	Count
Manual	25435	37374	31255.593481	31151	3269.406185	3375
Skilled Manual	50869	62806	56623.889402	56547	3231.097494	6058
Clerical	76294	88226	82036.469401	81964	3234.142349	4461
Management	101730	113674	107624.820854	107633	3200.647674	2858
Professional	127166	139115	133137.211478	133057	3223.119676	1603

Visualizing this gets very apparent:

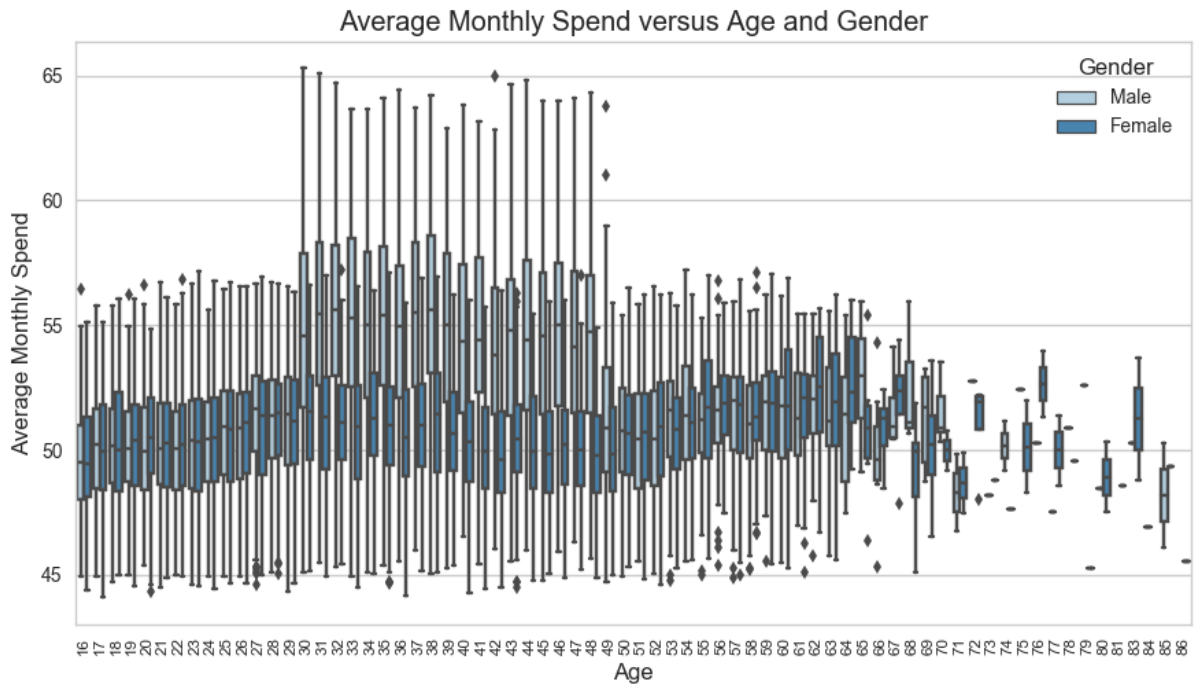


Average Monthly Spend / age / gender

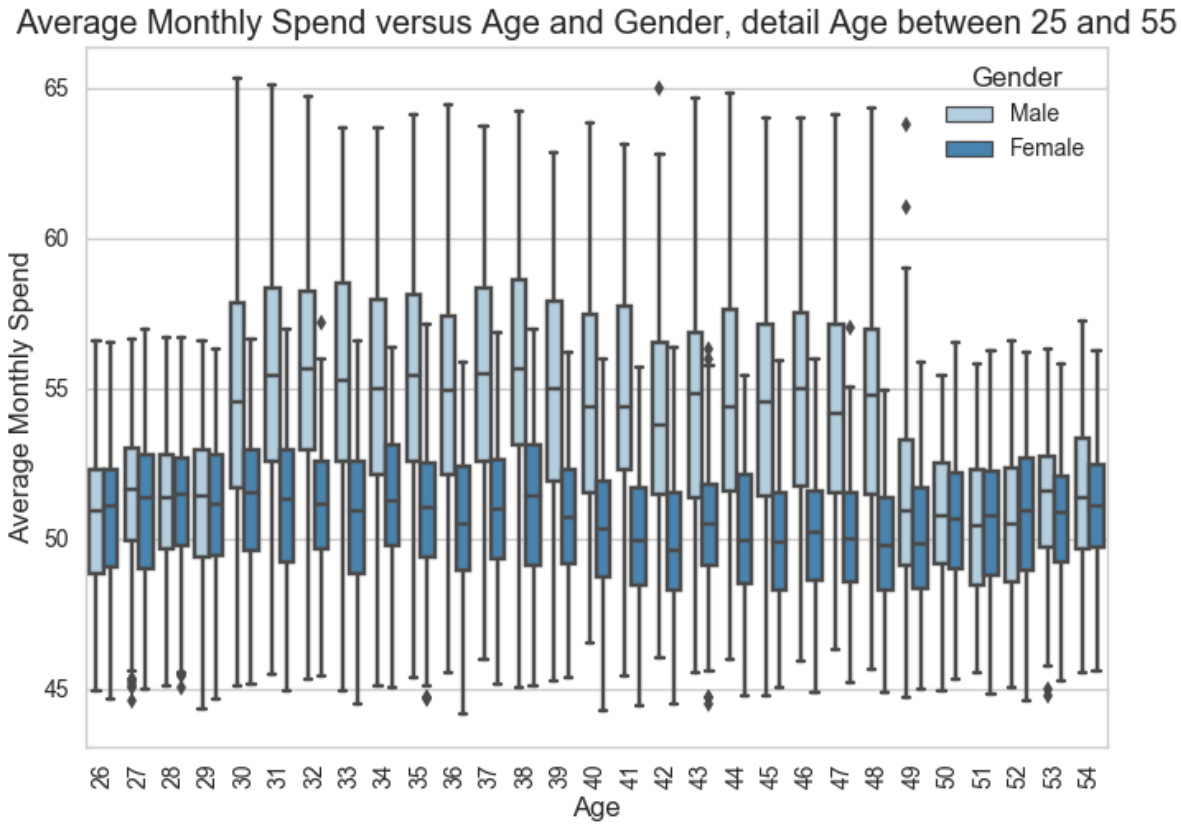
Age / gender / Average Monthly spend



Apparently, some distinct pattern for males, 30-50 yrs age group



Boxplot visualization of the same data

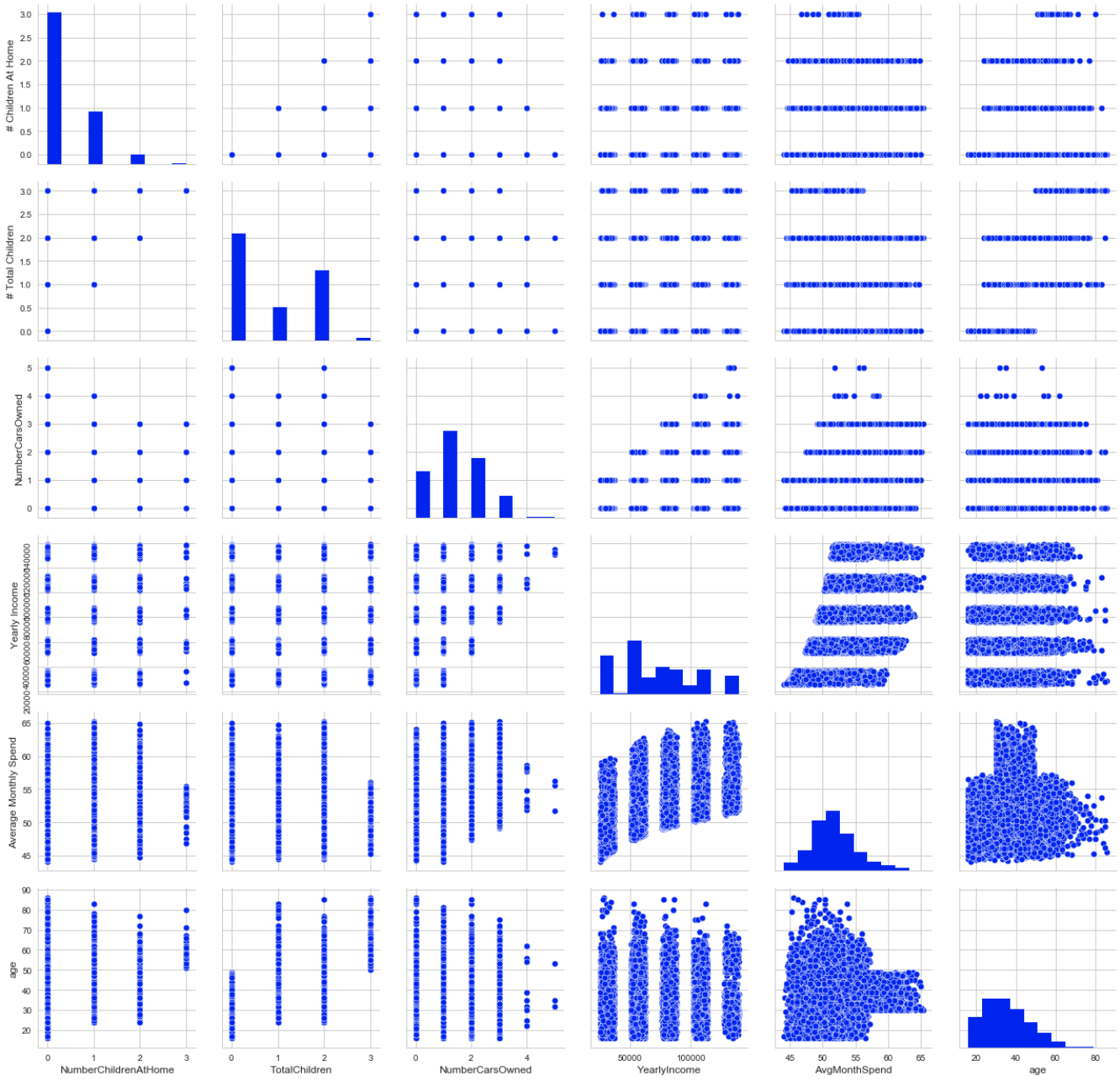


Out of a total of 4627 Males between (including) age 30 and (including) age 48, 1856 have an Average Monthly Spend in excess of 56. This group has a higher probability of being home owners, has more cars, more children at home (while the max children at home is lower at 2), slightly higher total Children (while the max is 2), a higher yearly income and a higher ratio of bike buyers than the average.

A Hypothesis for explaining this distinct pattern could include target advertisement in the past that could have targeted male, higher income, lower children and car owners. Another hypotheses could be a group or club purchase. In order to establish proof for any hypothesis / establish any correlation between the observed data and cause, additional information would be required. As no additional information is available, I see no reliable way to identify the complete set of effected datasets. Because of this, there is no reliable and trustworthy method to isolate these two datasets completely and use them separately. Ergo, I left the dataset complete and consequently must expect lower yields. This assessment is outside the current assessment; hence this concludes the remarks in this regard.

Yearly Income / has bought bike

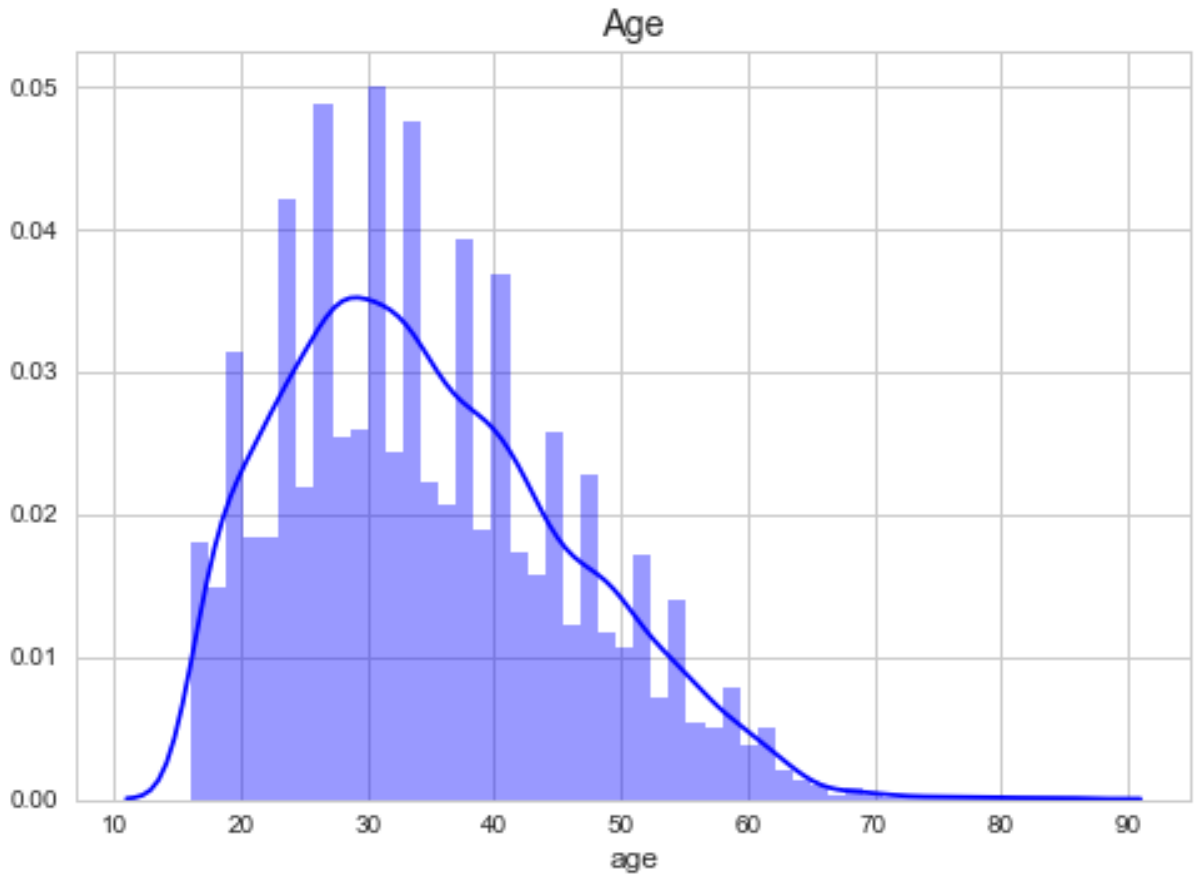
Other relationships:



No significant relationships between Average Monthly Spend and Marital Status or Country has been identified.

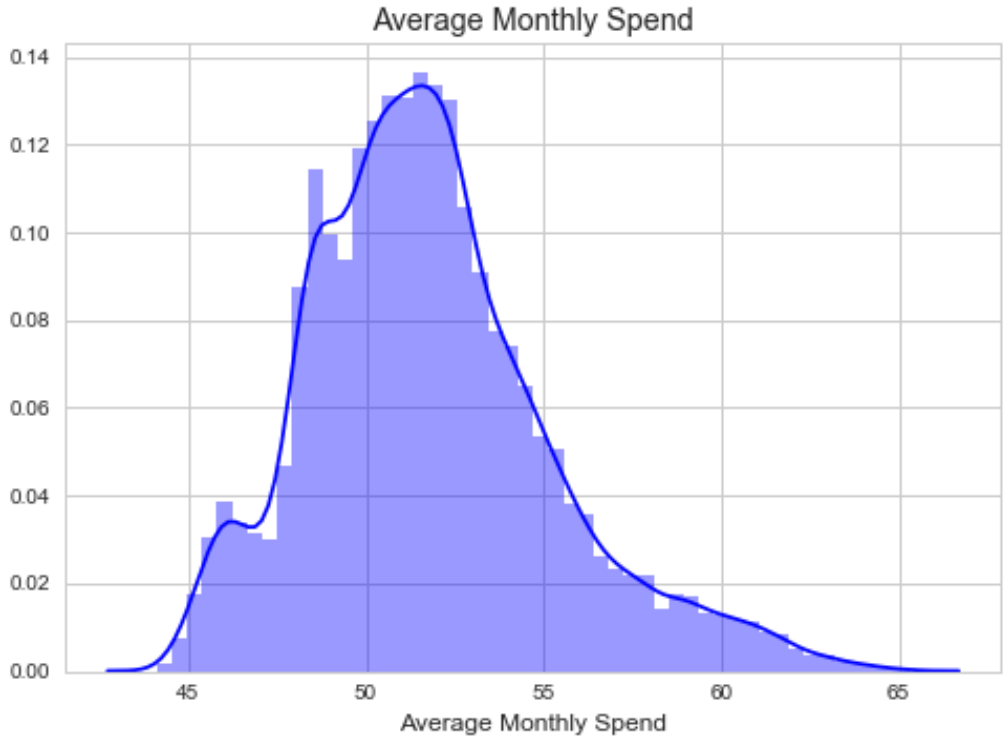
Distributions

Age:

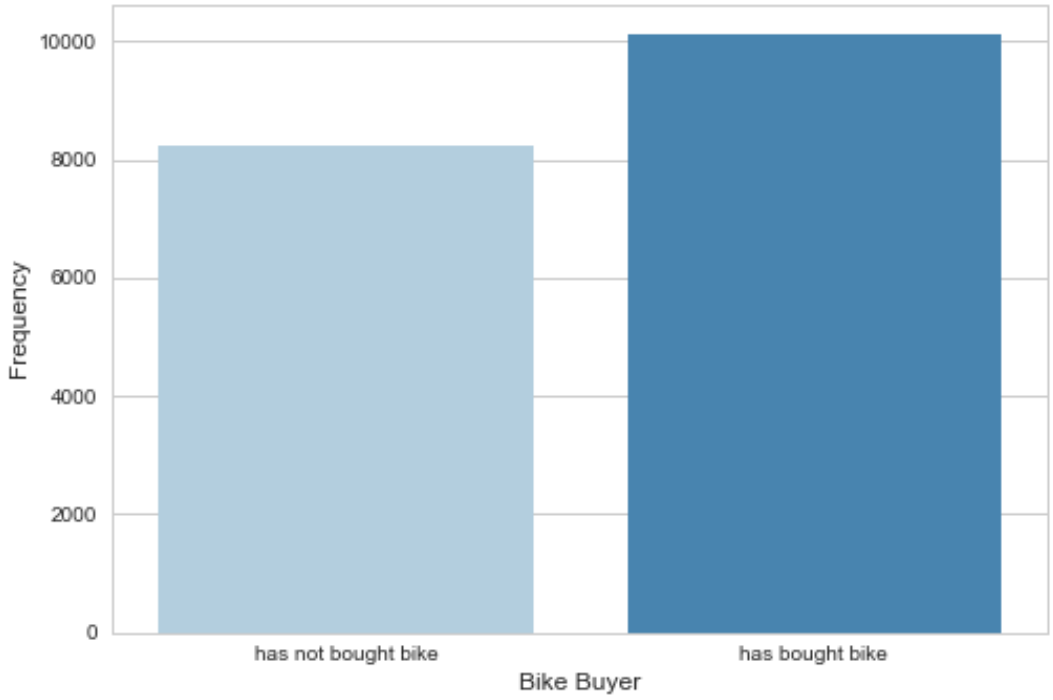


Noteworthy are the numerous distinct spikes with approx. double the occurrences then neighbouring data.

Average Monthly Spend

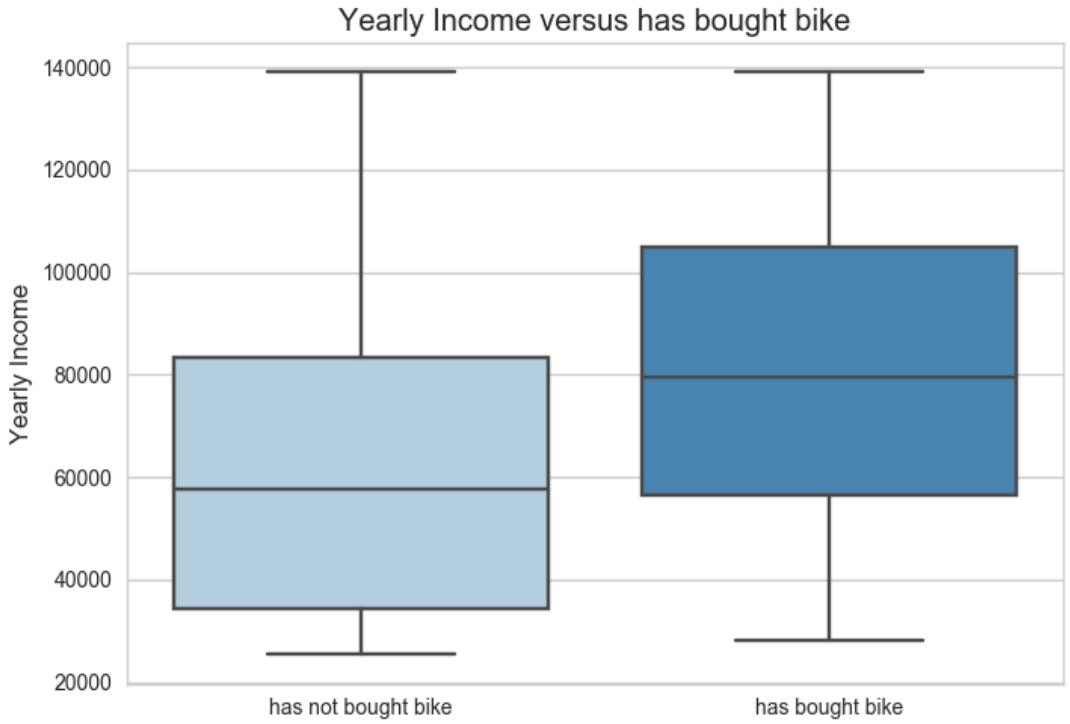


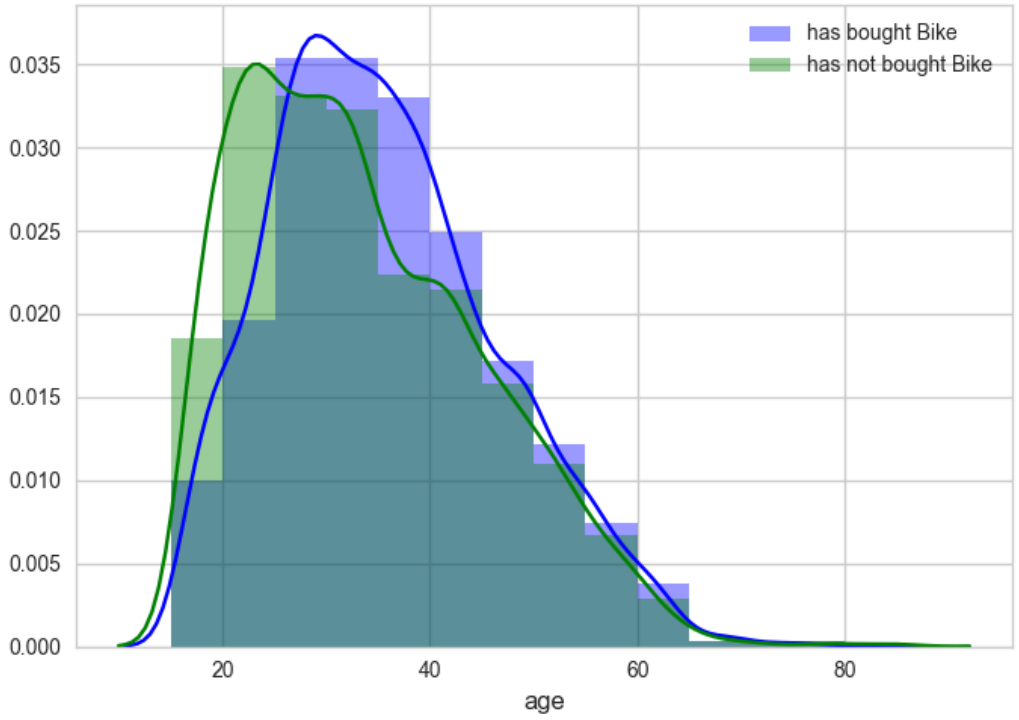
Bike Buyer



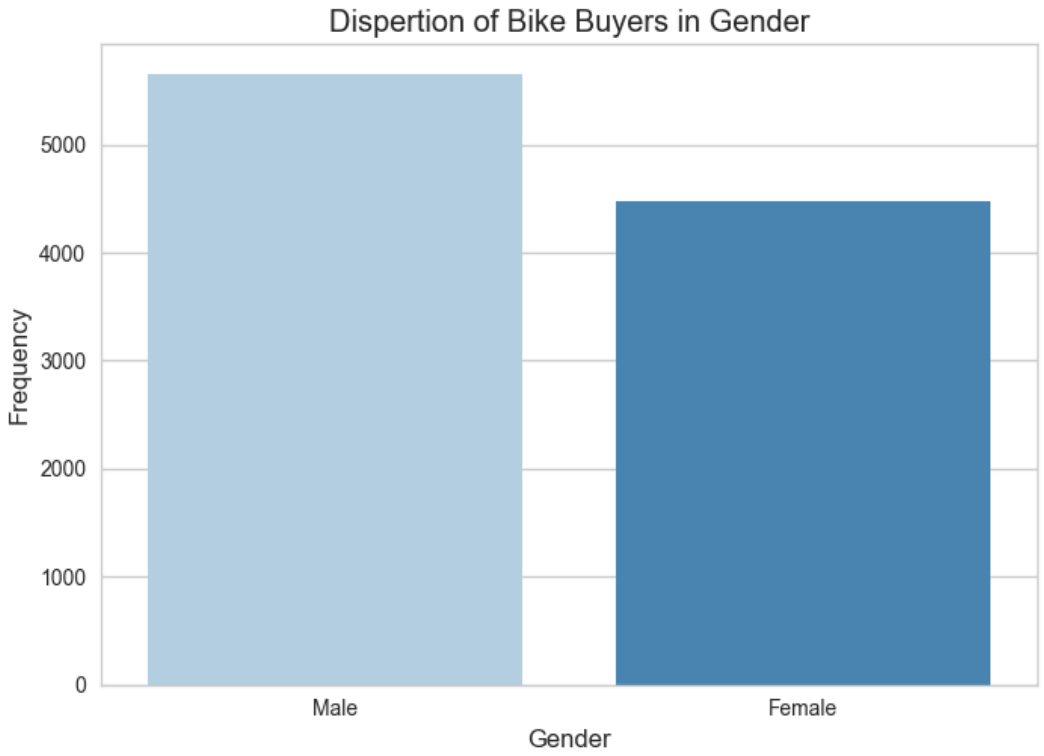
Correlation and Apparent Relationships relative to Likelihood of Bike Purchase

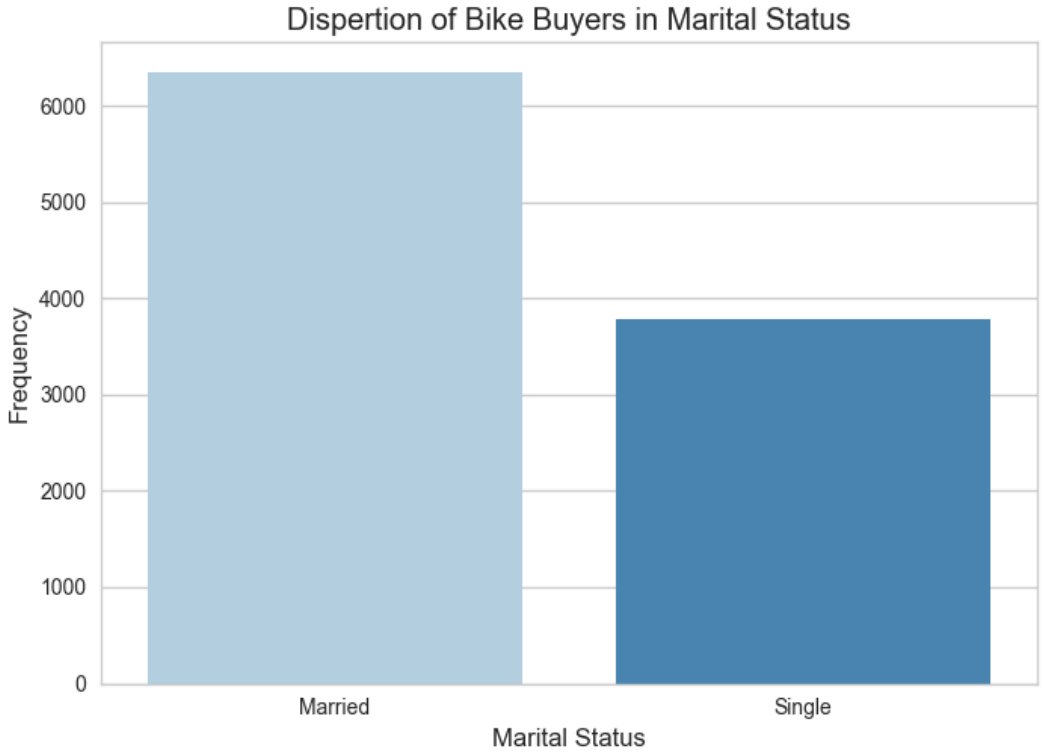
Numeric Relationships

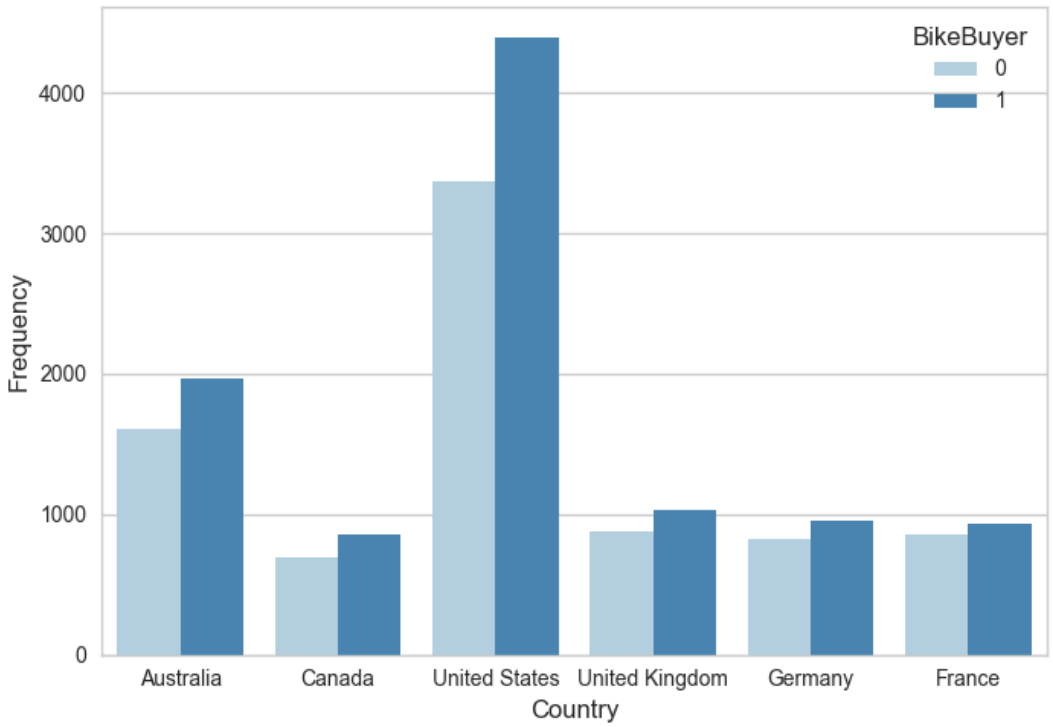




Categorical Relationships







Classification of Customers Based on likelihood to purchase a bike

Based on the initial data analysis datasets with NumberCarsOwned >= 4, Total Children = 3 or NumberChildrenAtHome = 3 were removed as outliers. The resulting data contains 18115 out of the original 18355 datasets. This constitutes an approx. 1.31% loss and appears reasonable. At train/test split of 70/30 was used to train a boosted tree classifier.

As a distinctly separated dataset on the feature of occupation is observed, a model considering different classifiers based on the occupation feature has been considered, but did not yield substantive better results. It was determined, that not dividing the dataset and observation realm into sections are the more robust and future proof approach.

The Features promising the most relevant correlation are: Number Cars Owned, Number Children At Home, Occupation, Total Children, Marital Status, Home Owner.

The achieved accuracy is:0.76669733

The achieved precision is:0.771073

Detailed Results (test data = 5435 datasets):

True Positives: 2479

True Negatives: 1688

False Positives: 736

False Negatives: 532

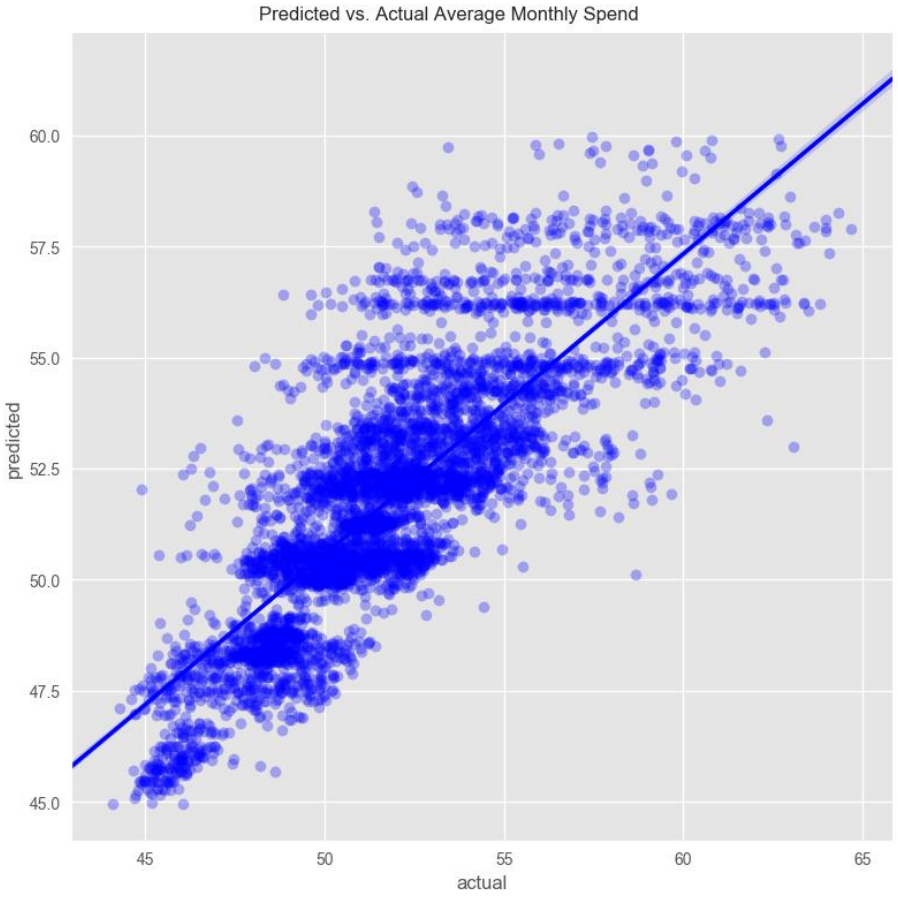
Regression of Average Monthly Spend

Based on the initial data analysis datasets with NumberCarsOwned >= 4, Total Children = 3 or NumberChildrenAtHome = 3 were removed as outliers. The resulting data contains 18115 out of the original 18355 datasets. This constitutes an approx. 1.31% loss and appears reasonable. At train/test split of 70/30 was used to train a random forest regressor.

The features most correlated with a successful prediction of Average Monthly Spend are: Yearly Income, age, Gender, Marital Status, Number Children At Home, Number Cars Owned.

MSE: 3.8995

RMSE: 1.9747



It seems that the earlier mentioned group of above $56 average monthly spend creates some distinct interferences. The tendency is that the actual monthly spend is higher than the predicted value, based on the existing data. As no further information is available in regard to what caused the earlier mentioned outlier group, with reasonable confidence and caution, the predicted values seem trustworthy.

Conclusion

The dataset provided offered some challenges due to its distribution and dependencies of values. Further, it seems that a portion of the dataset is skewed due to some external, unknown precondition. The data itself seems legit, within stated limitations and reasonable low additional noise or outliers.

It was shown that the likelihood of a customer to purchase a bike can be predicted, although with a relative high margin of uncertainty. The features most likely to be successful predictors are: Number Cars Owned, Number Children At Home, Occupation, Total Children, Marital Status, Home Owner

It was shown that the average monthly spend can be predicted with reasonable accuracy given the overlays in the dataset. The predicted values in general appear to be conservative estimates, at least considering the available dataset, hence can be deemed suitable for forecasting with appropriate caution. The most successful predictors are: Yearly Income, age, Gender, Marital Status, Number Children At Home, Number Cars Owned.