Executive Summary

This report summarizes the analysis of a dataset relating to customer info and sales from a bicycle company. The dataset contains 18,356 unique records with many features that help to provide a clear picture of the customer.

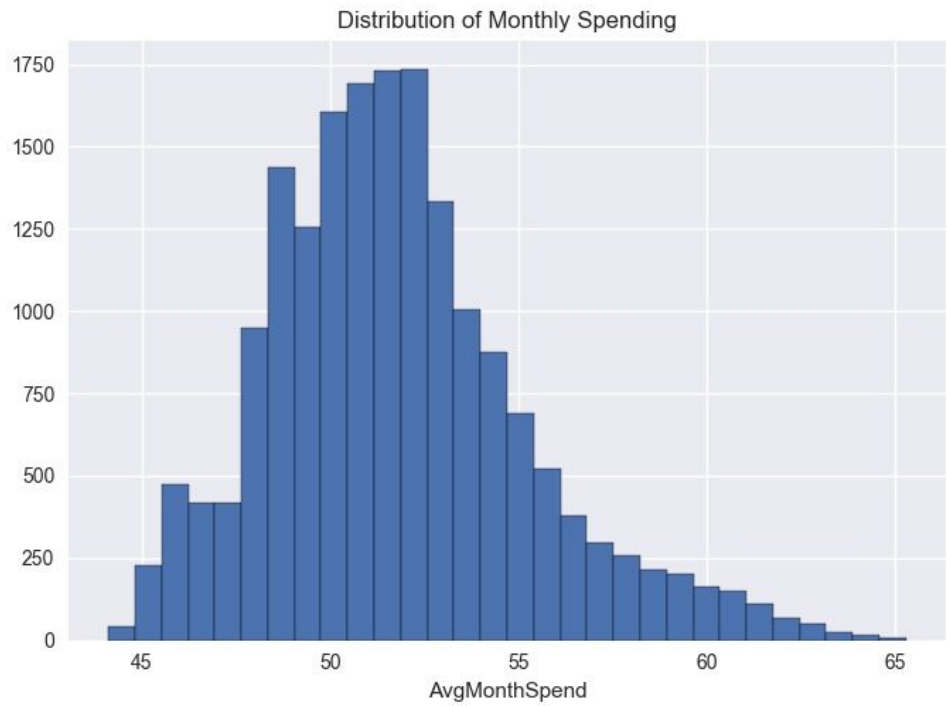
The feature names are as follows:

- CountryRegionName
- BirthDate
- Education
- Occupation
- Gender
- MaritalStatus
- HomeOwnerFlag
- NumCarsOwned
- NumChildrenAtHome
- TotalChildren
- YearlyIncome
- BikeBuyer

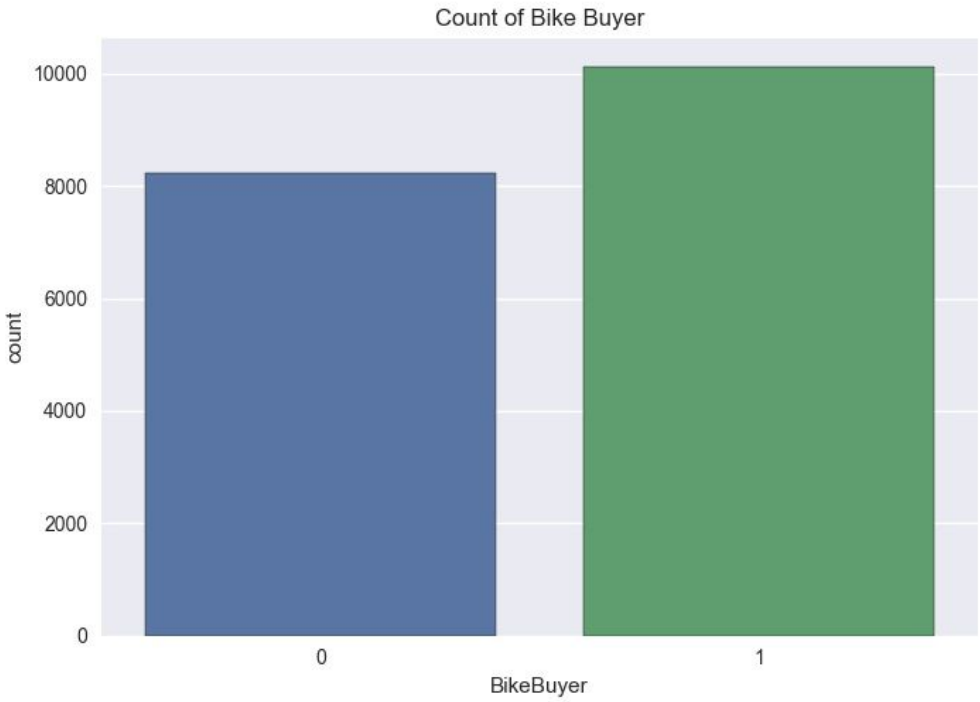
Below is a summary of the numerical columns in the dataset, after converting the **BirthDate** column to one showing **Age**. You will notice that the **HomeOwnerFlag** and **BikeBuyer** columns are both a binary indicator column, with each value as either a 0 or 1. It is also of note that the mean age of a customer is 34.7 and the average spent by a typical customer each month is $51.77.

	HomeOwnerFlag	NumCarsOwned	NumChildrenAtHome	TotalChildren	YearlyIncome	BikeBuyer	AvgMonthSpend	Age
count	18356.000000	18356.000000	18356.000000	18356.000000	18356.000000	18356.000000	18356.000000	18356.000000
mean	0.610591	1.270484	0.338200	0.850512	72759.308128	0.551700	51.767100	34.730551
std	0.487630	0.913951	0.568991	0.927377	30686.866749	0.497334	3.437961	11.255865
min	0.000000	0.000000	0.000000	0.000000	25435.000000	0.000000	44.100000	16.000000
25%	0.000000	1.000000	0.000000	0.000000	53312.750000	0.000000	49.410000	26.000000
50%	1.000000	1.000000	0.000000	0.000000	61851.500000	1.000000	51.420000	33.000000
75%	1.000000	2.000000	1.000000	2.000000	87412.000000	1.000000	53.600000	42.000000
max	1.000000	5.000000	3.000000	3.000000	139115.000000	1.000000	65.290000	86.000000

As the **AvgMonthSpend** and **BikeBuyer** are the primary concerns of the experiment, a preliminary analysis of the two are shown below:



The histogram shows a right-tailed distribution, with a small number of outliers reaching up to a maximum of about $65 spent per month. The majority of customers, though, are spending between $50 and $53.

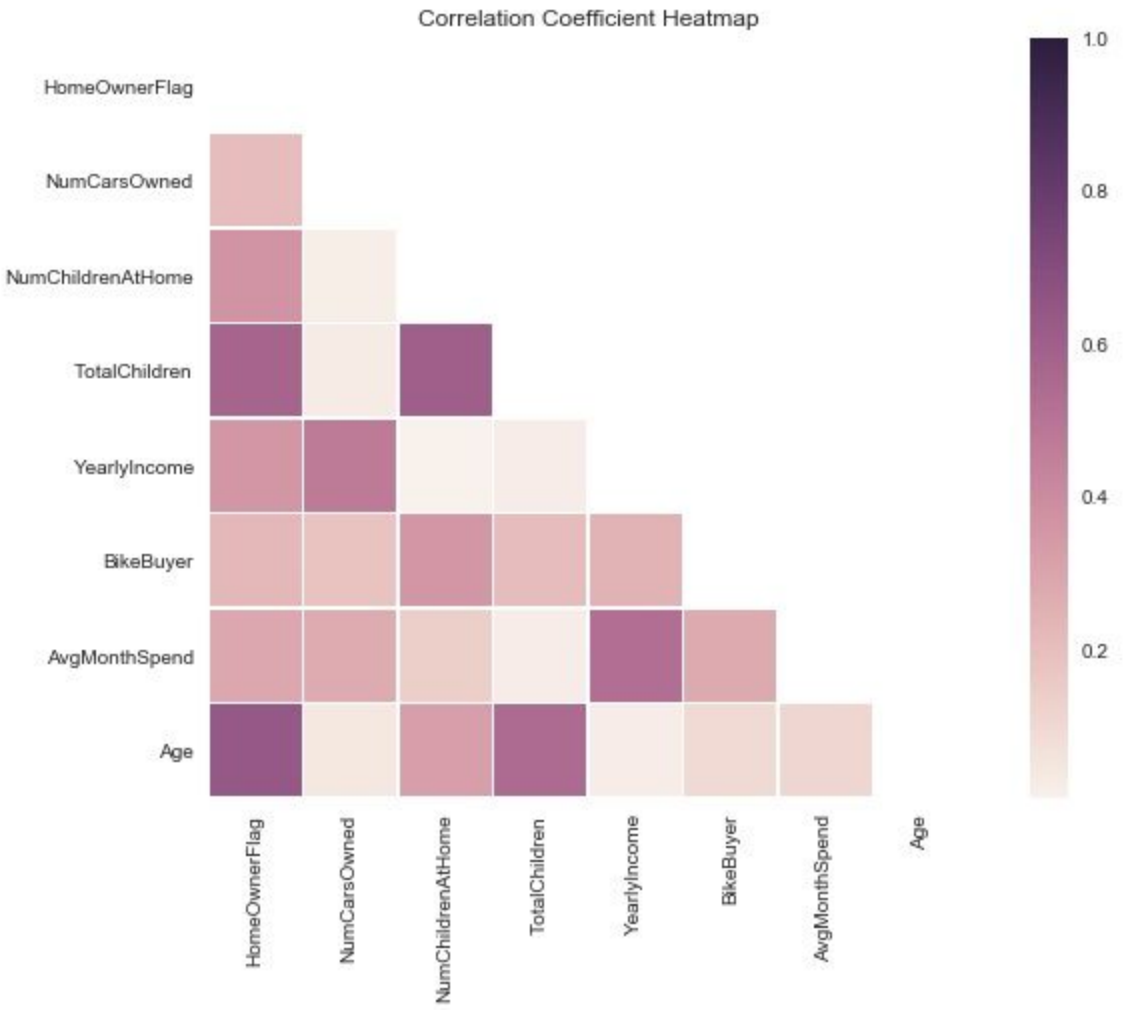


The countplot of **BikeBuyer** highlights a difference of approximately 2,000 customers between bike buyers and non-bike buyers, showing a fairly balanced

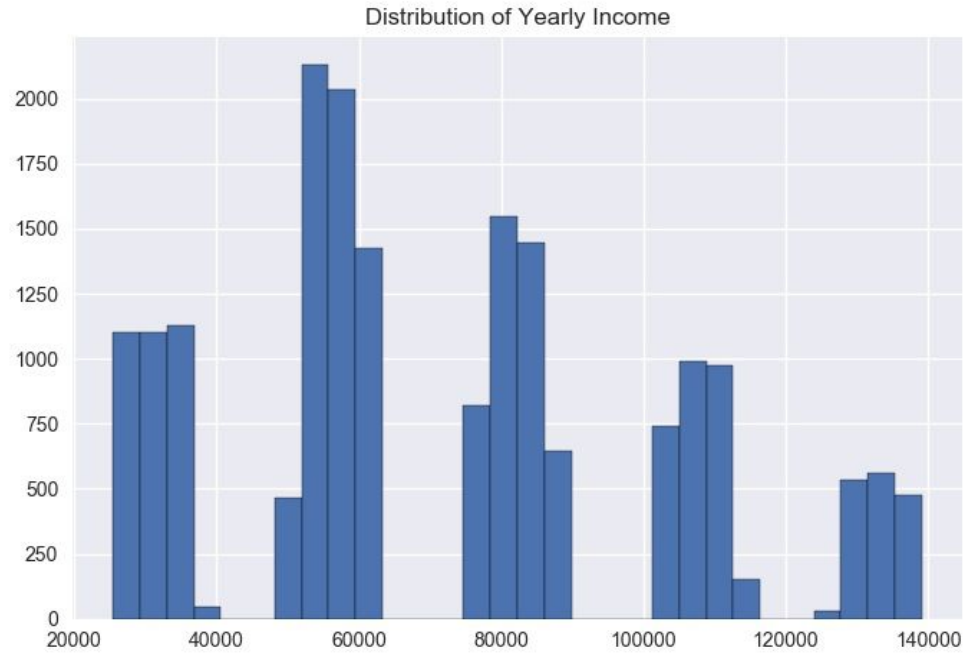
dataset. Further analysis will hone in on some key factors that create a larger discrepancy between the two classes.

Numerical Data

A heatmap displaying the correlation coefficients for the numerical columns shows that **NumChildrenAtHome** has the greatest direct impact on our **BikeBuyer** column and the **AvgMonthSpend** column is most influenced by **YearlyIncome**.



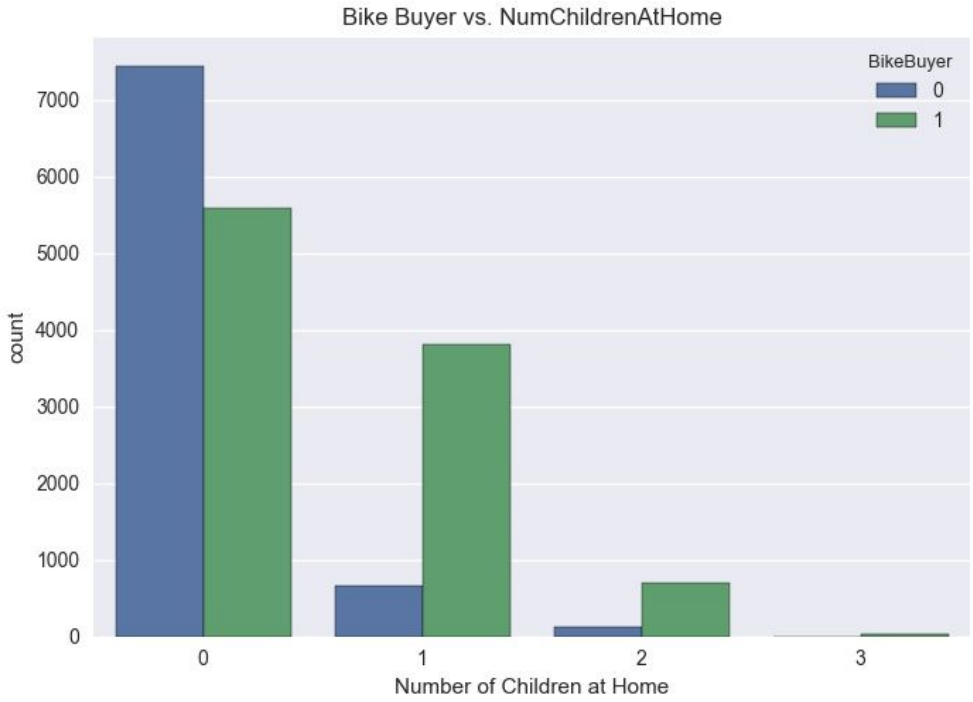
A histogram shows 5 distinct levels of income, with virtually no overlap. The shape is right-tailed, driving up the mean ($72,759.30) from the median ($61,851.50).



The scatterplot between **YearlyIncome** and **AvgMonthSpend** reveals a positive linear correlation between the two features: As a customer's yearly income increases, so too does the amount spent per month.

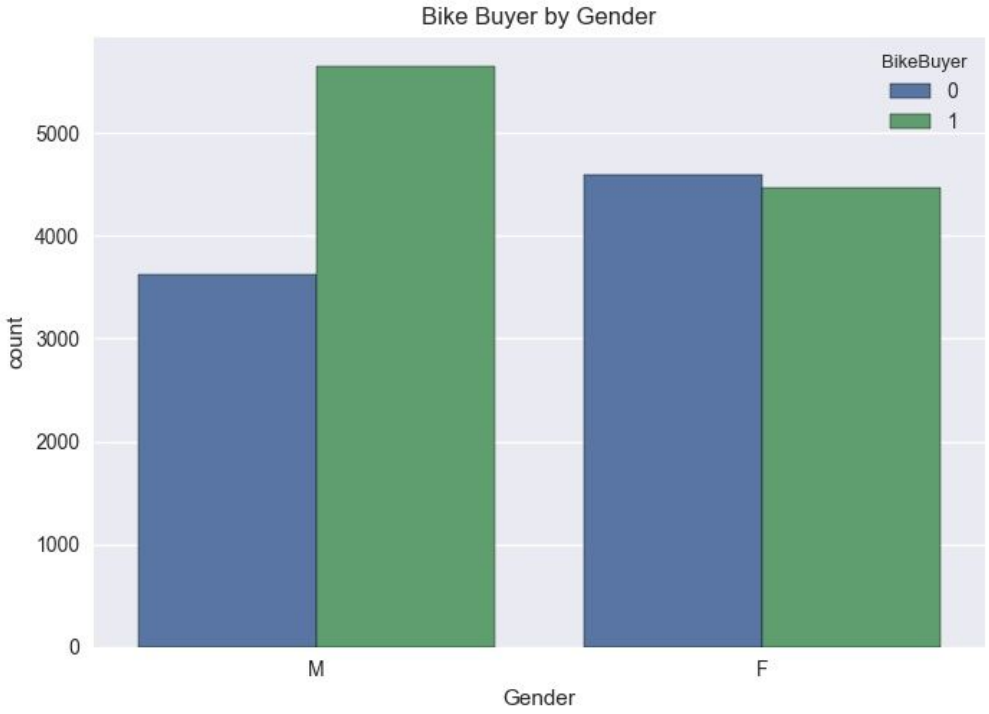
The following chart provides valuable insight into relationship between how many children still living at home, and whether or not that customer buys a bike. For instance, a customer with 1 child at home is more than 5 times more likely to buy a

bike than to not buy a bike. The numbers are very similar for those customers with 2 children at home.

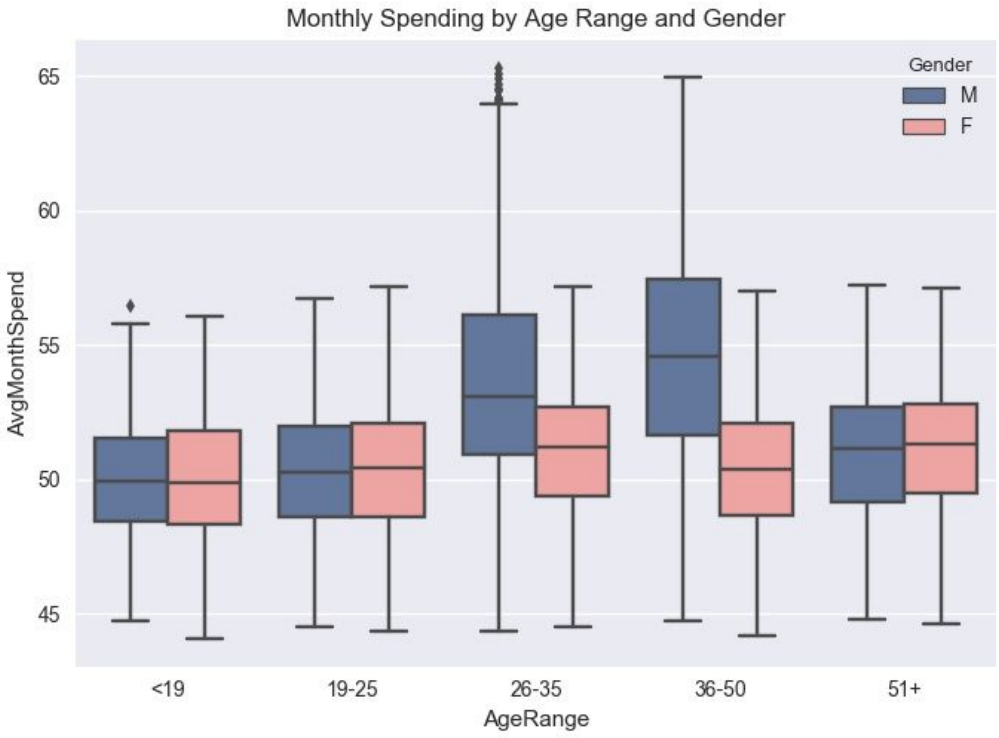


Categorical Data

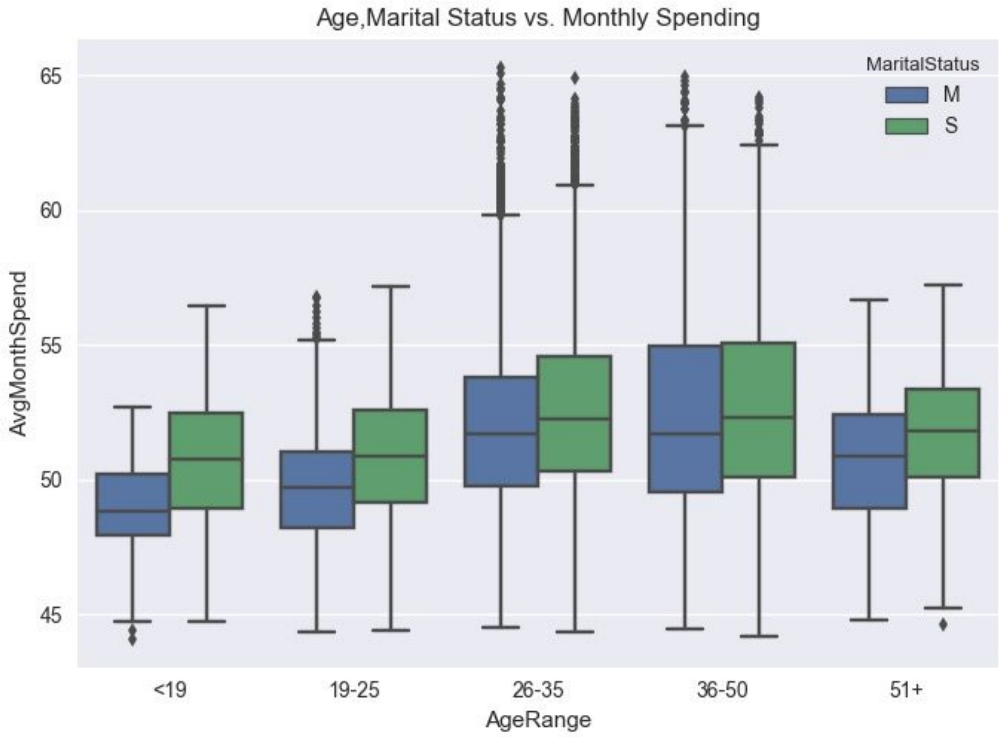
A cursory peek at the **BikeBuyer** column broken down by gender reveals that men are more likely to purchase a bike than women, and that women, in general, are slightly more likely to not purchase a bike than to purchase one.



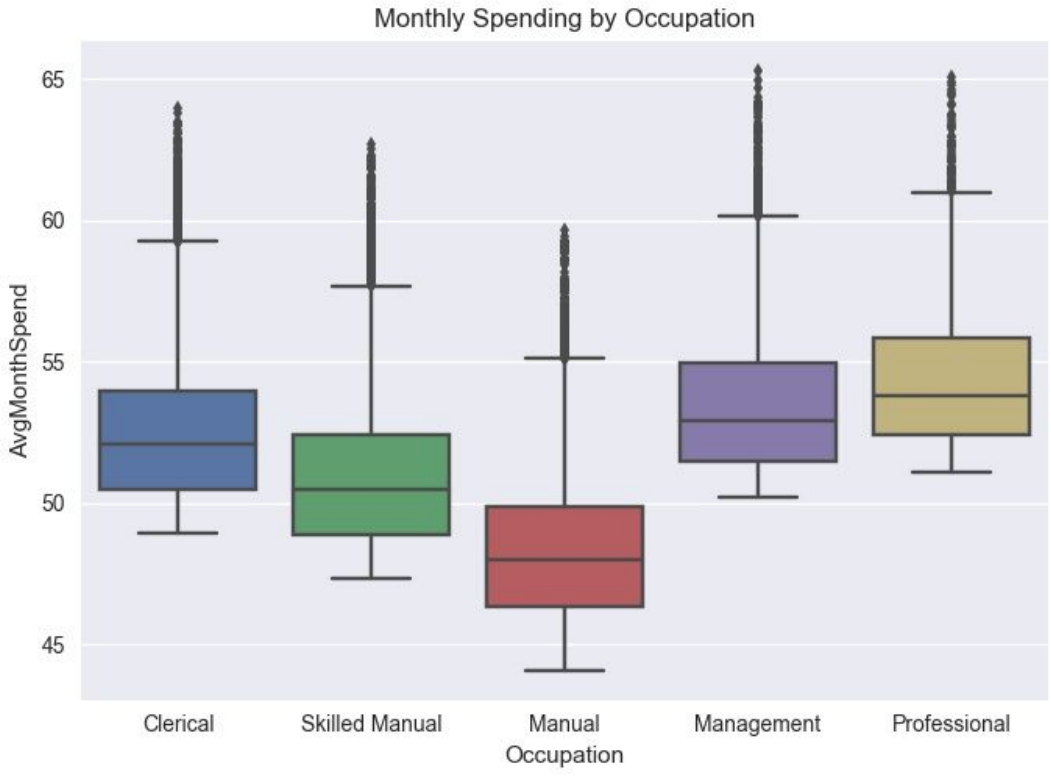
Turning back to monthly spending, the following visualization reveals an interesting trend. When broken down into **age ranges**, it is apparent that men and women within three of those range categories spend virtually the same amount per month. However, men spend significantly more than women in both the 26-35 and the 36-50 categories, with the most drastic difference in the latter.

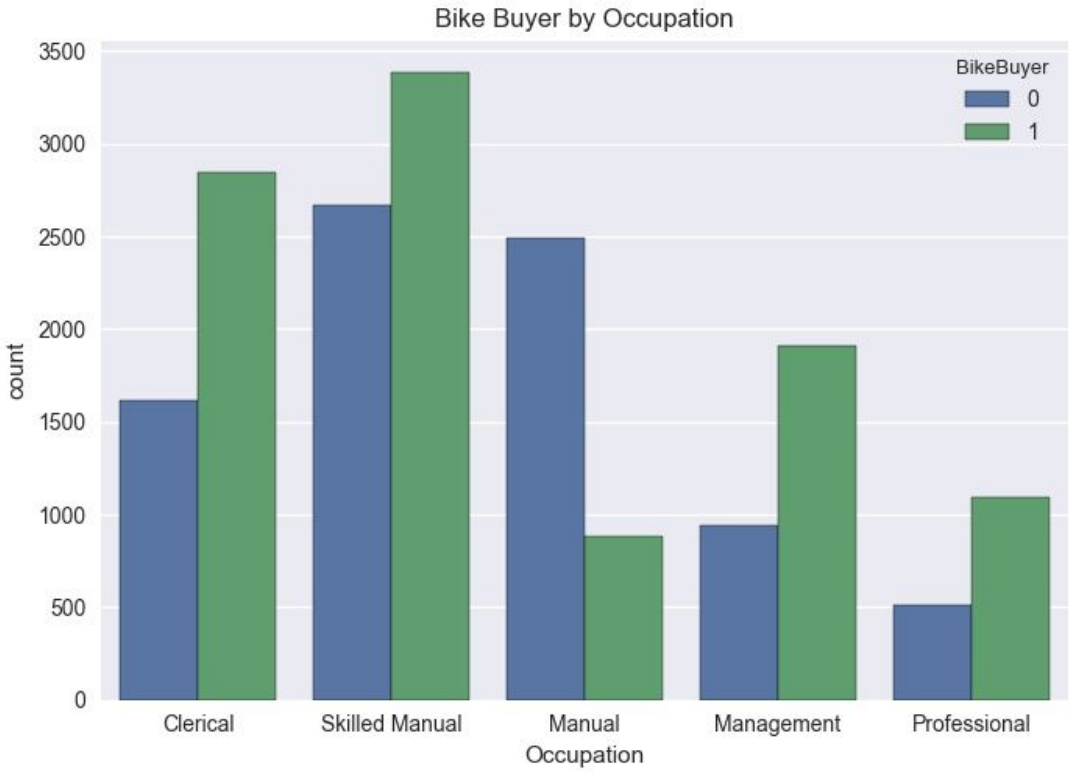


A similar boxplot shows that single customers tend to spend, on average, more than single customers for each age range:

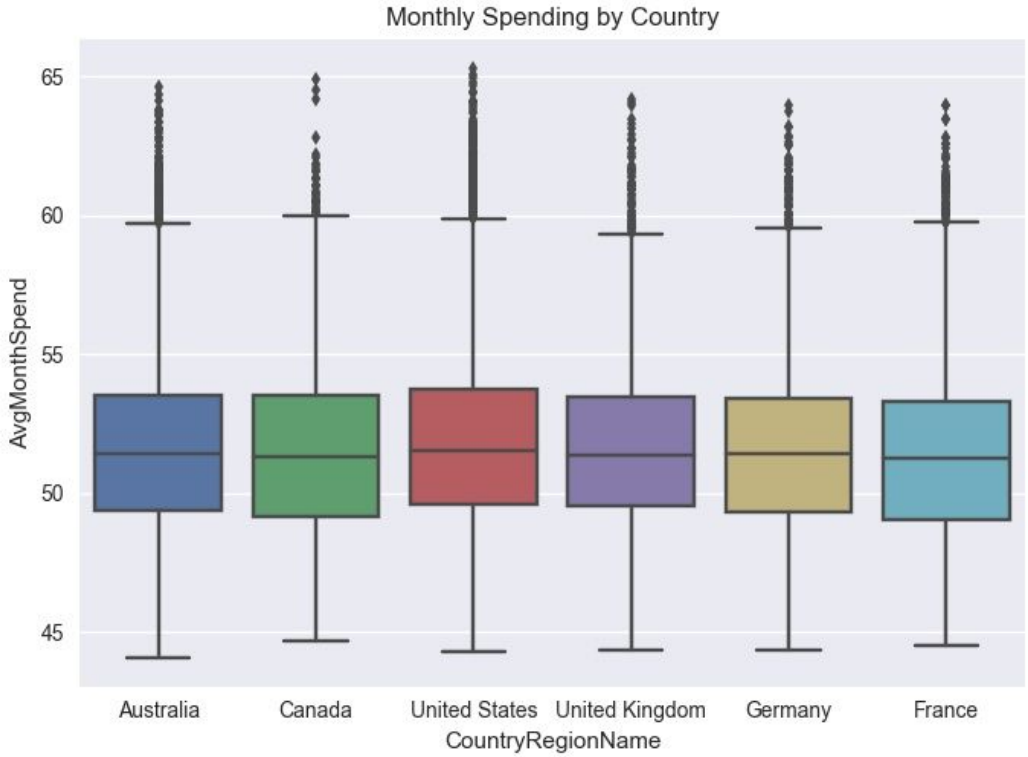


Occupation, too, tends to influence both the average amount spent per month and whether or not a customer purchases a bike. For instance, a customer whose job is in the manual labor category tends to spend significantly less than customers in other occupational categories. Those in the "Professional" category, though, typically spend more and are much more likely to purchase a bike than not.





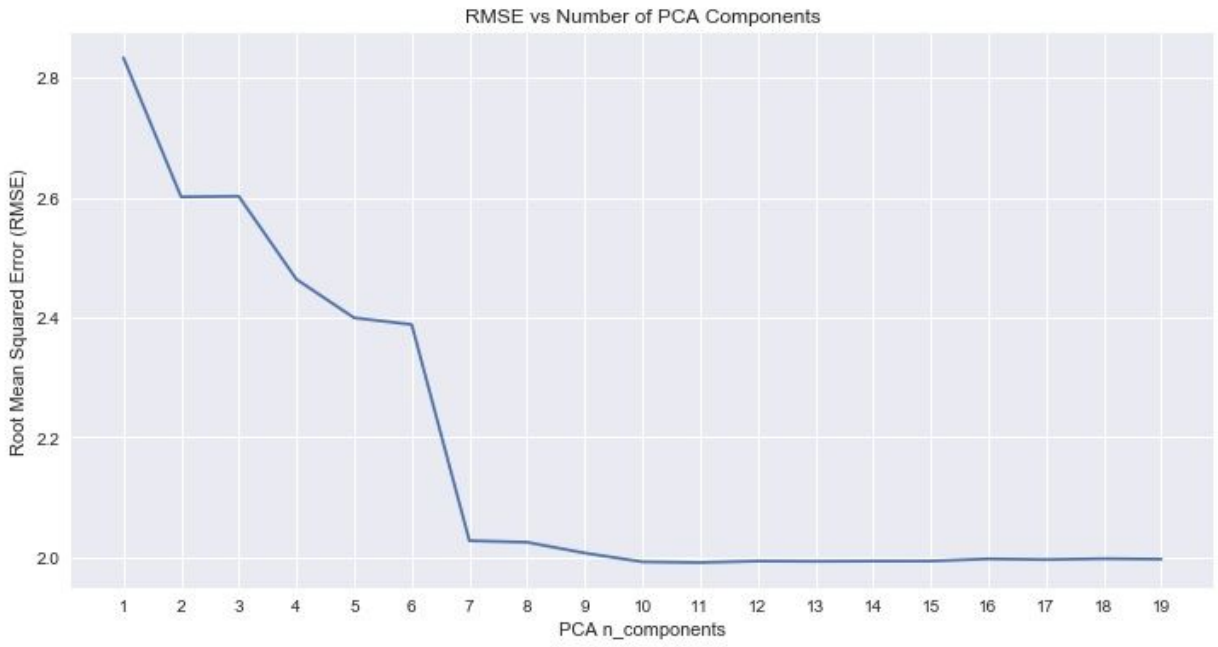
The last feature of note is **CountryRegionName**. The following boxplot shows that, interestingly, the country a customer is in likely has little influence over the amount spent each month.



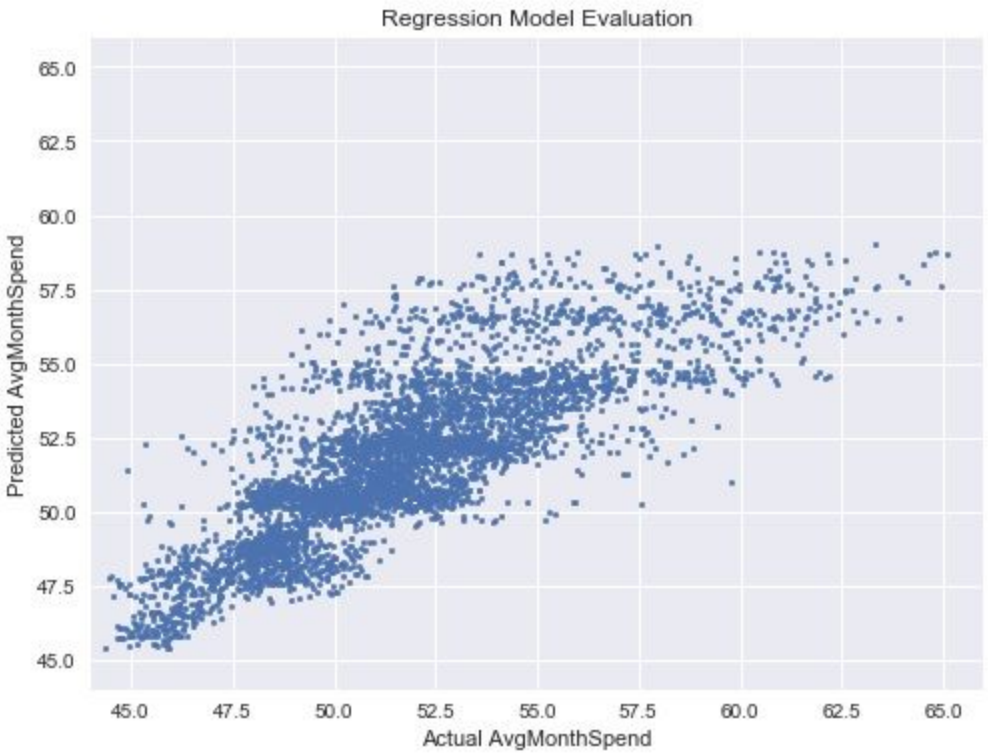
After the data has been cleaned, analyzed, and visualized, two predictive models were created to 1) predict the average amount a customer will spend at the shop monthly, and 2) classify a customer into a BikeBuyer or a non-BikeBuyer. It was determined that a boosted decision tree performed well for both models.

Regression Model

Before fitting the data to a machine learning model, Principal Component Analysis showed that at n_components=11 (see chart below), there was a definite decrease in error.



Using scikit-learn's GradientBoostingRegressor paired with its grid search capabilities, the following hyperparameters were used for the model: learning_rate= 0.1, max_depth= 3, n_estimators= 100. Using a training size of 75%, the model achieved a root mean squared error (RMSE) of 1.996. The following graph shows the predictions plotted against the actual values:

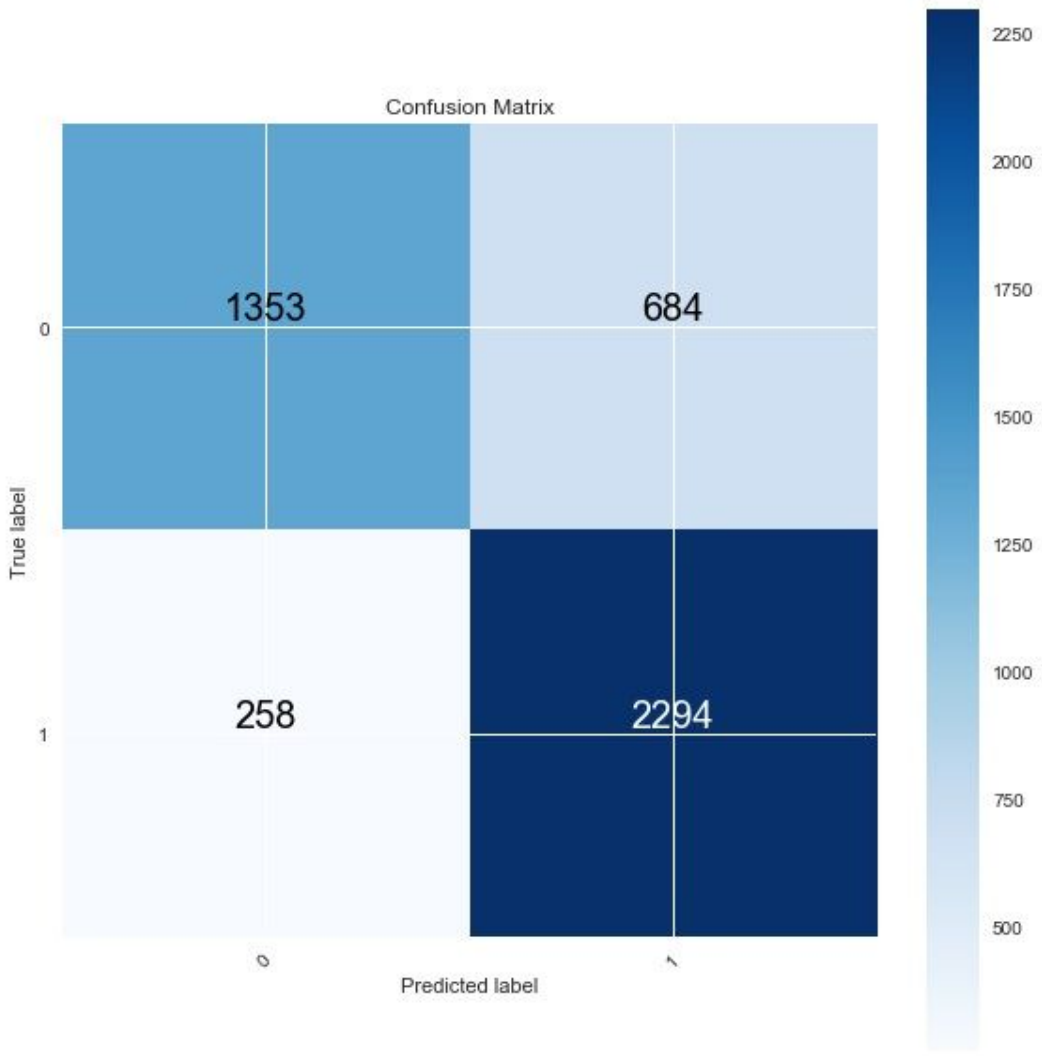


A perfect model would show all the points lined up in a precise 45 degree angle. As seen on the visualization, the model performs quite well up until the x-value reaches about 55. Then you can see larger errors begin to appear. Further analysis on the higher-paying customers may prove useful in this case.

Classification Model

A classification model was creating using the GradientBoostingClassifier from scikit-learn with the following hyperparameters: learning_rate=0.001, max_depth=4, n_estimators=500. Like the regression model, the training size was 75% of the data and the test size was 25%. Below is a classification report and a confusion matrix for the model:

	precision	recall	f1-score	support
0	0.84	0.66	0.74	2037
1	0.77	0.90	0.83	2552
avg / total	0.80	0.79	0.79	4589



The total precision of the model is .80, with recall having the highest deviation (.66 vs .90). Like the training data, the test data shows that there are slightly more bike buyers than non-bike buyers.

Conclusion

A model can be successfully created to predict how much a customer will spend on an average month, as well as whether or not that customer will purchase a bike. The primary features that influence these factors include yearly income, how many children a customer has at home, gender, marital status, and age range.