Trivalley Bike Store

Predictive Analytics

Content

Business Case: Analysis of Trivalley bike store's data and build a

- Classification model to predict whether a customer will buy a bike
- Regression model to predict a customer's Average monthly spending in the bike store

Data Acquisition: Bike store dataset was obtained in CSV format

Data Visualization: Descriptive Statistics using Tableau

Exploratory Data Analysis and Predictive Analysis done using Azure ML and Python

Data Preparation: Clean all noise, missing data (Qualitative and Quantitative analysis)

Exploratory Data Analysis: Find the most impacting features on dependent variable and find linear correlation between variables

Predictive Analysis with Machine Learning: Find the suitable machine learning algorithm, train, score and evaluate the prediction model

Data Visualization - Tableau

Tableau is an interactive data visualization tool used for Exploratory Data Analysis (EDA), where charts/graphs are plotted for dimensions (qualitative values) against measures (quantitative values) and dependent variables (readmit30) to get insights and understand their data. Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their statistical characteristics, often with visual methods.

Tableau is quick, simple, user-friendly, intuitive, can handle lot of data, provide statistical calculations on datasets

EDA:

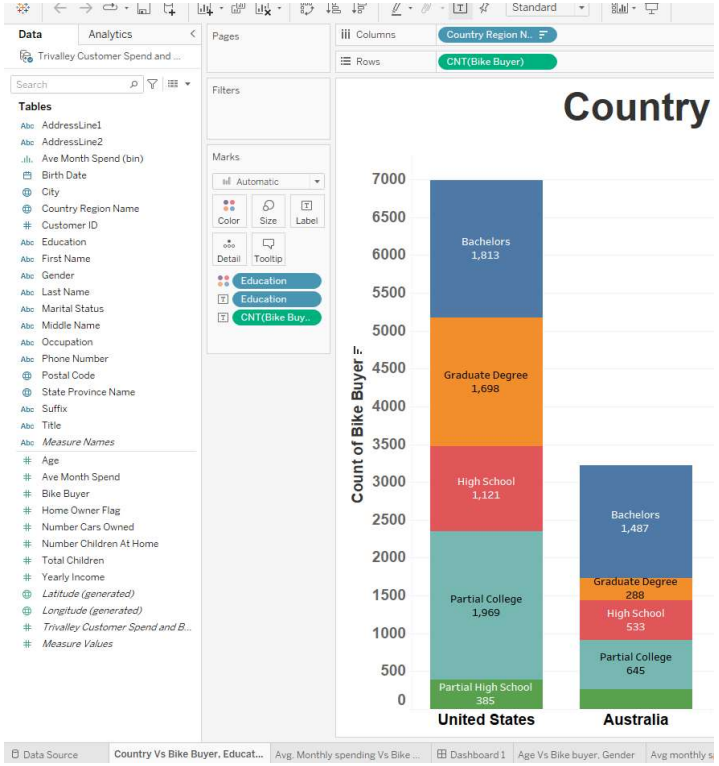
- Get a better understanding of data that may not be analyzed by standard data science algorithms.
- Understanding data patterns that may be skipped by typical machine learning algorithms.
- Drawing charts and graphs for better understanding from different angles and projects the results.
- To get a better understanding of the problem statement, visually.
- To find the hidden trends and relationship between variables.
- Assess and validate your assumptions on the variables, whether the variables help answer business problem or not.
- Screen for noise variables, missing data, outliers, etc. Find which variables need imputation, preprocessing

Tableau

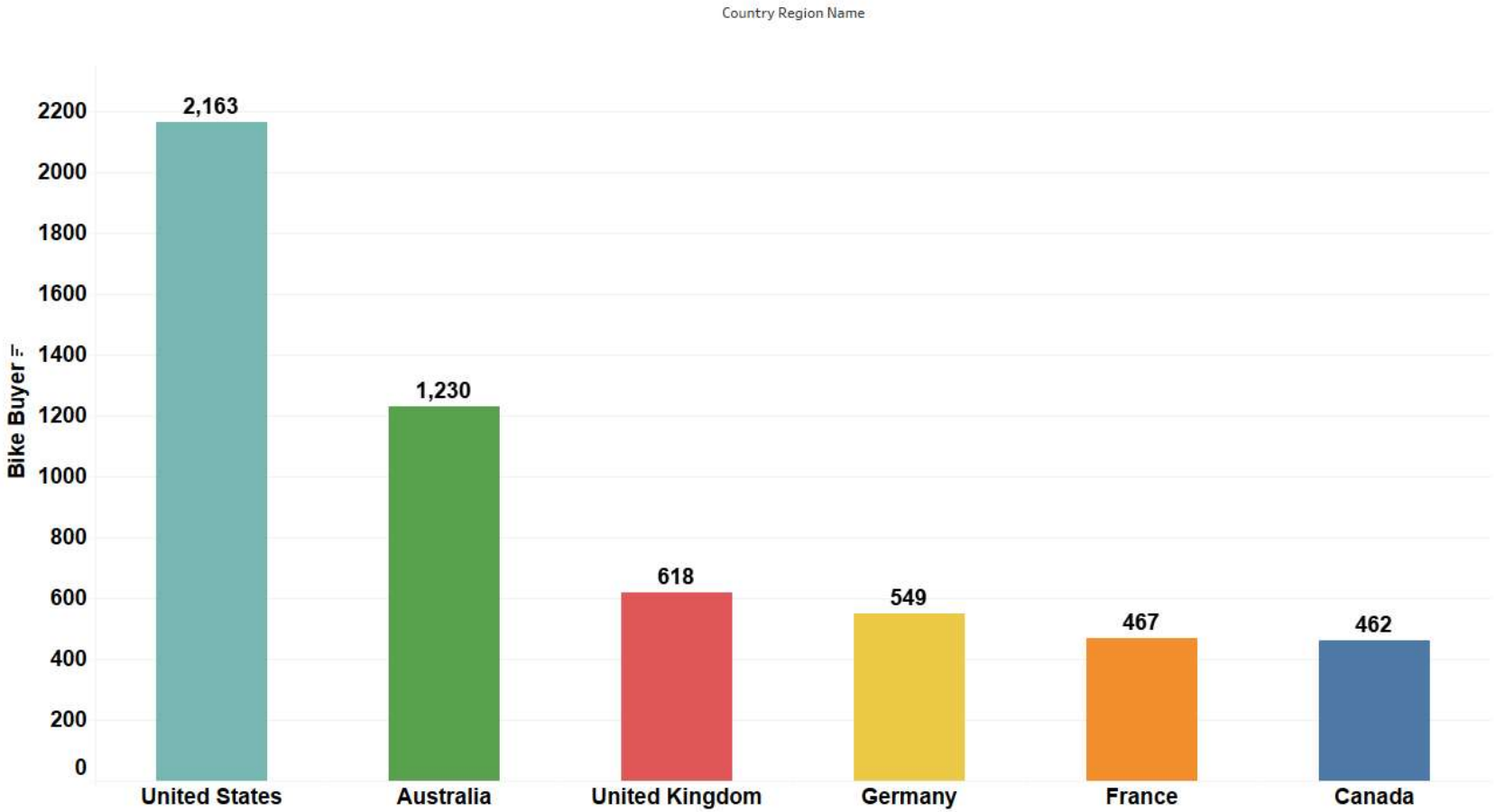
Chart Views

1. Text tables
2. Heat maps
3. Highlight tables
4. Symbol maps
5. Maps
6. Pie charts
7. Horizontal bars
8. Stacked bars
9. Side-by-side bars
10. Tree maps
11. Circle views
12. Side-by-side circles
13. Lines (continuous)
14. Lines (discrete)
15. Dual lines
16. Area charts (continuous)
17. Area charts (discrete)
18. Dual combination
19. Scatter plots
20. Histogram
21. Box and whisker plots
22. Gantt
23. Bullet graphs
24. Packed bubbles

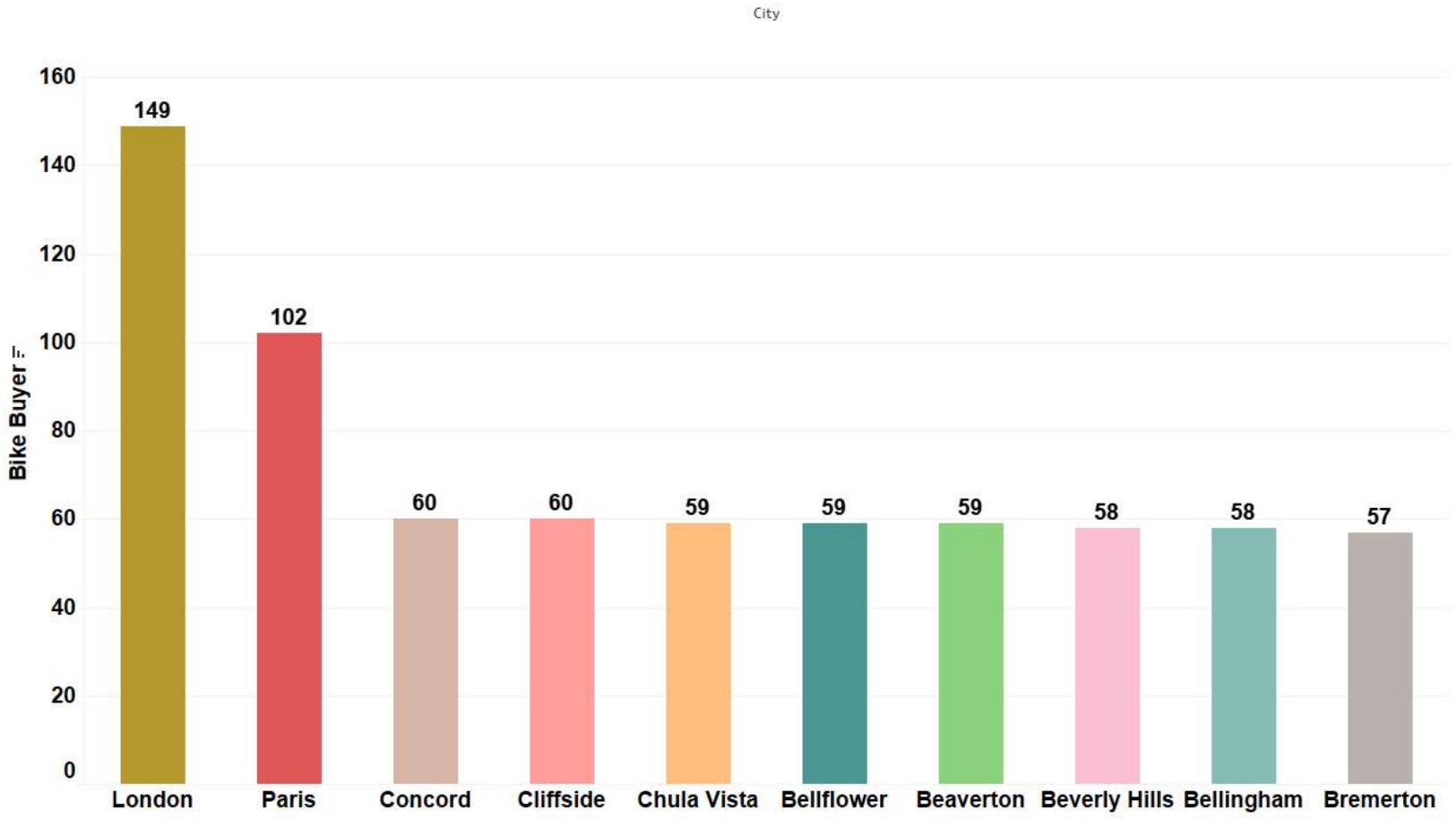
Data Pane, Marks card and Worksheet



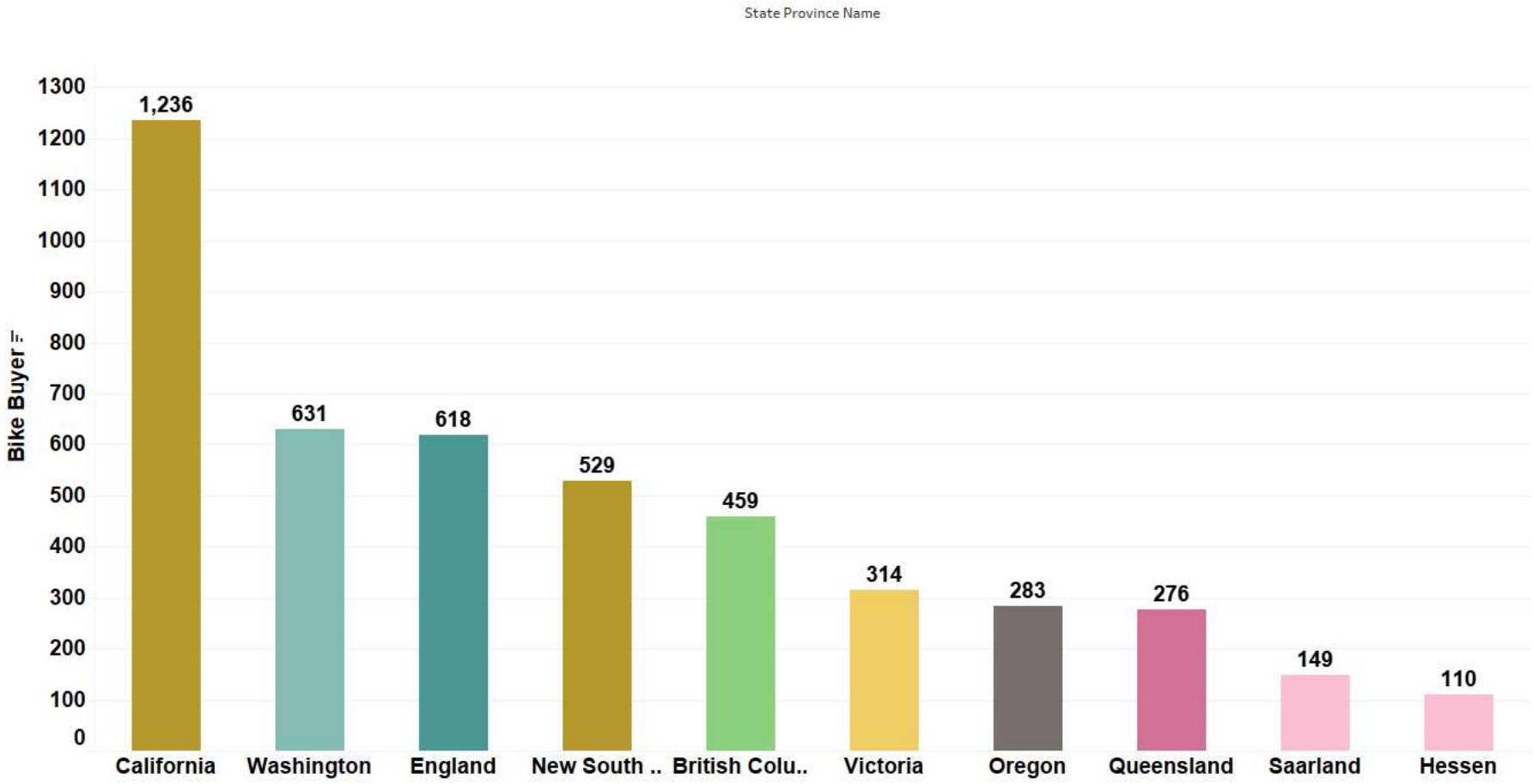
Countries Vs Bike owners



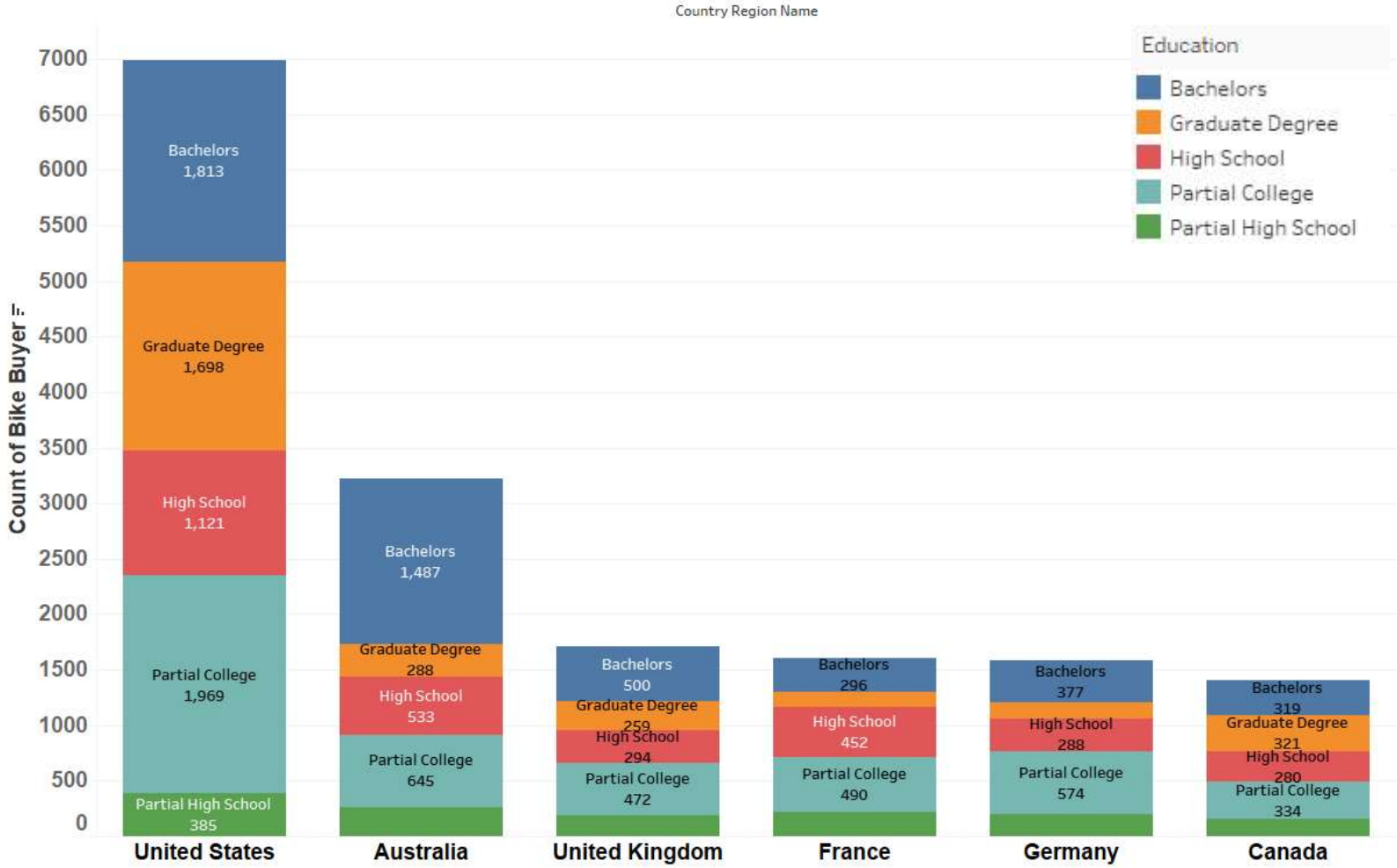
Top 10 cities with more Bike owners



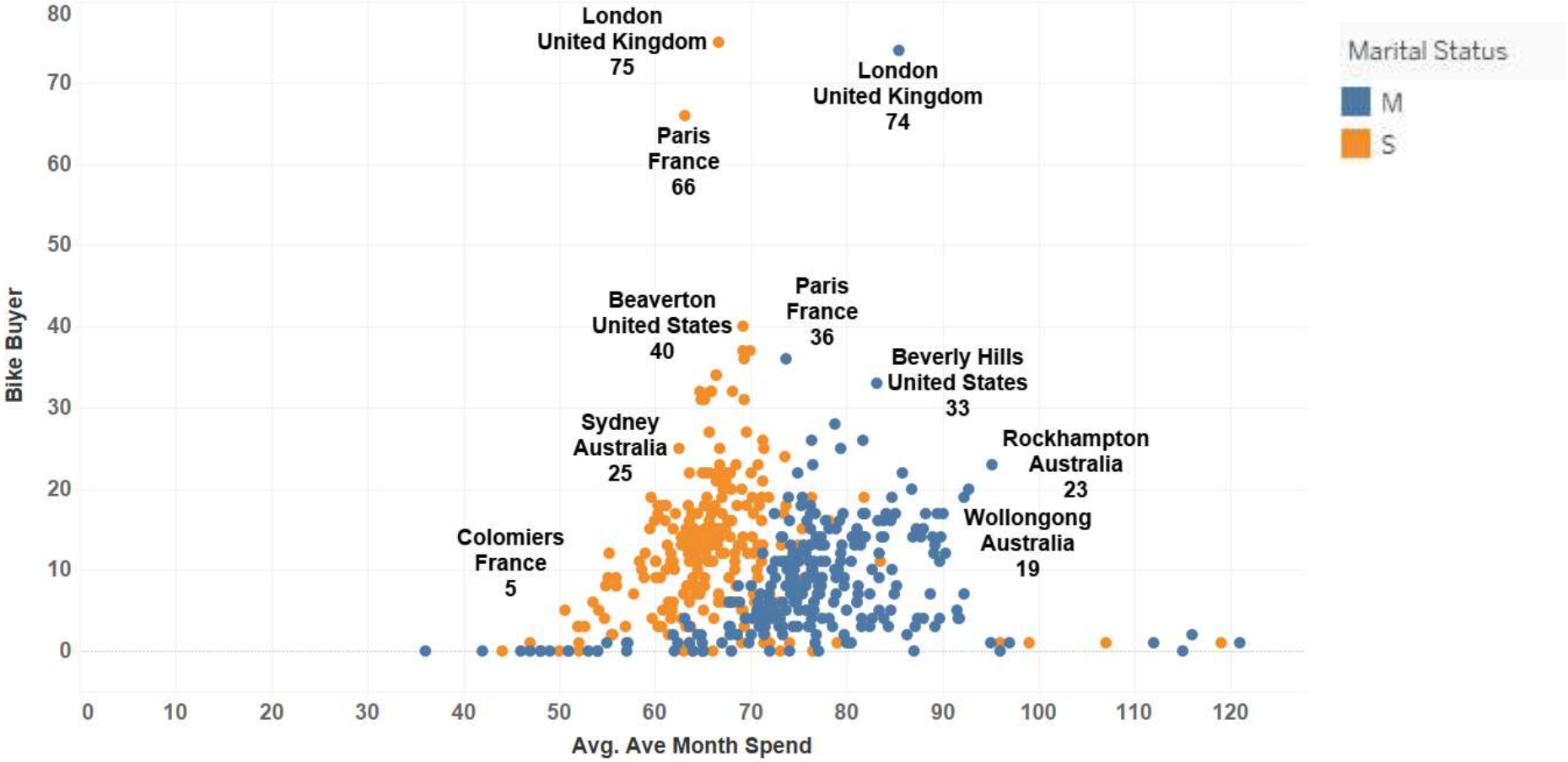
Top 10 State Province with Bike owners



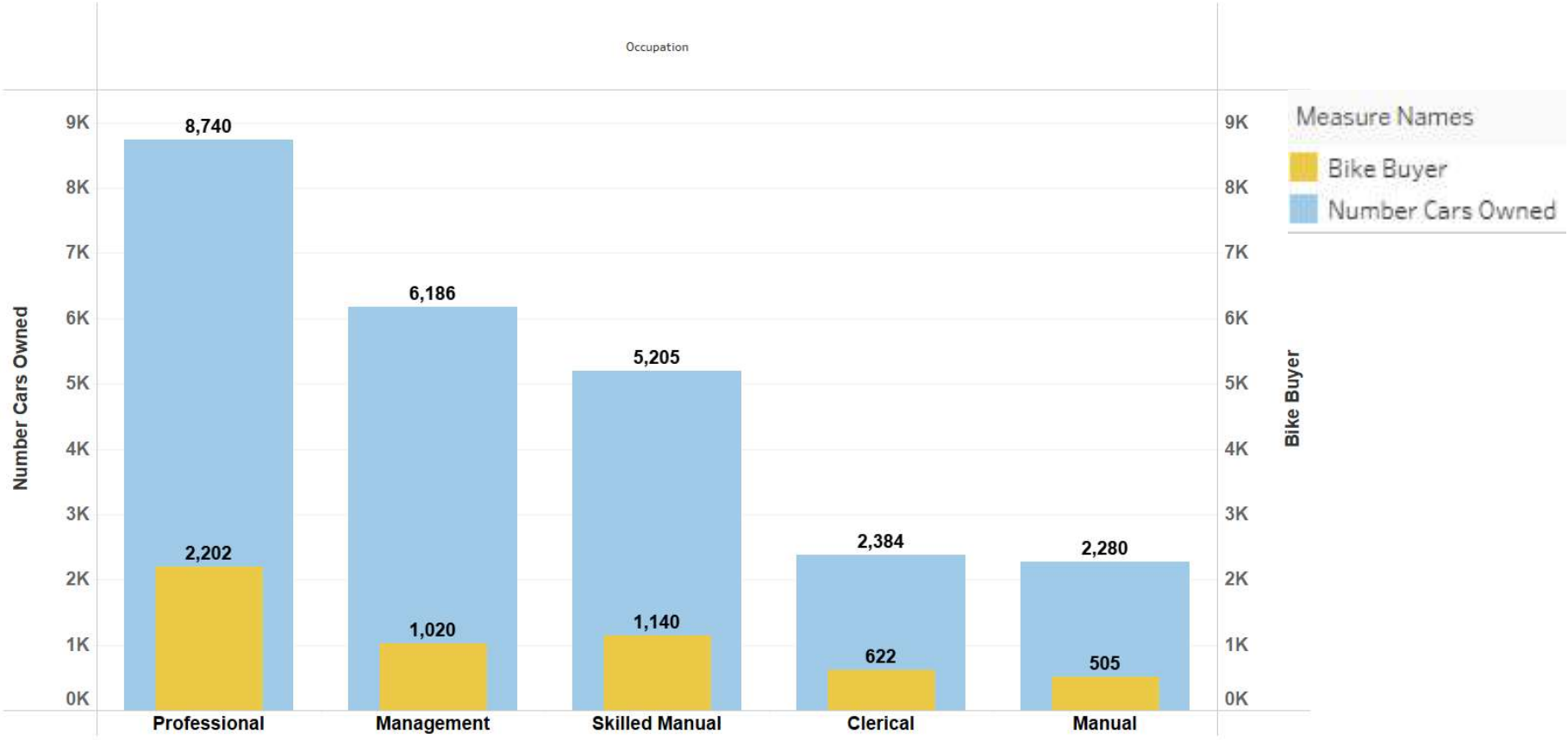
Country Vs Bike Buyer, Education



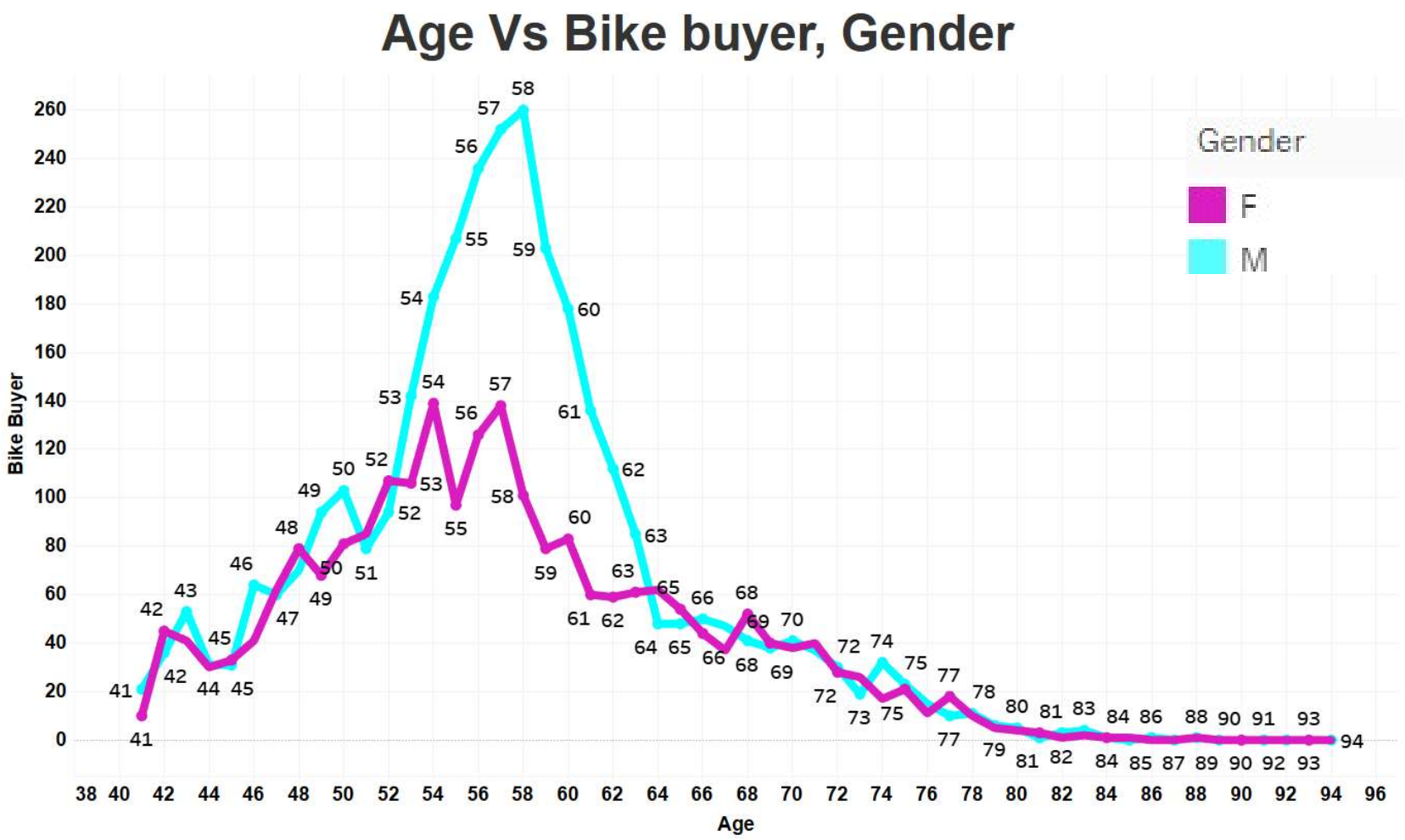
Avg. Monthly spending Vs Bike Buyer, City, Marital Status



Occupation Vs No. of cars, bikes owned

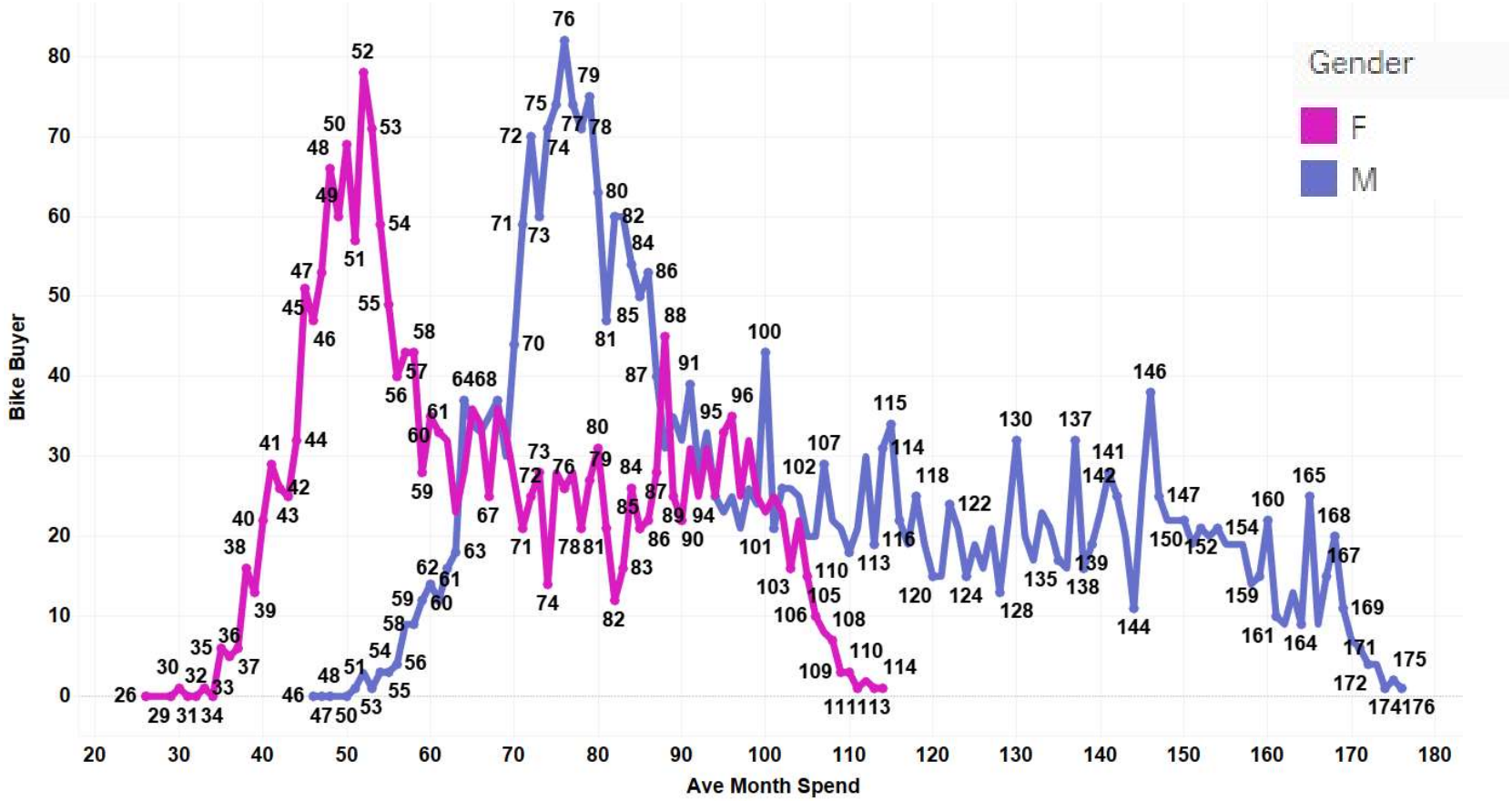


Occupation influences the number of car and bike buyers. Professionals buy more cars and bikes than other occupations.

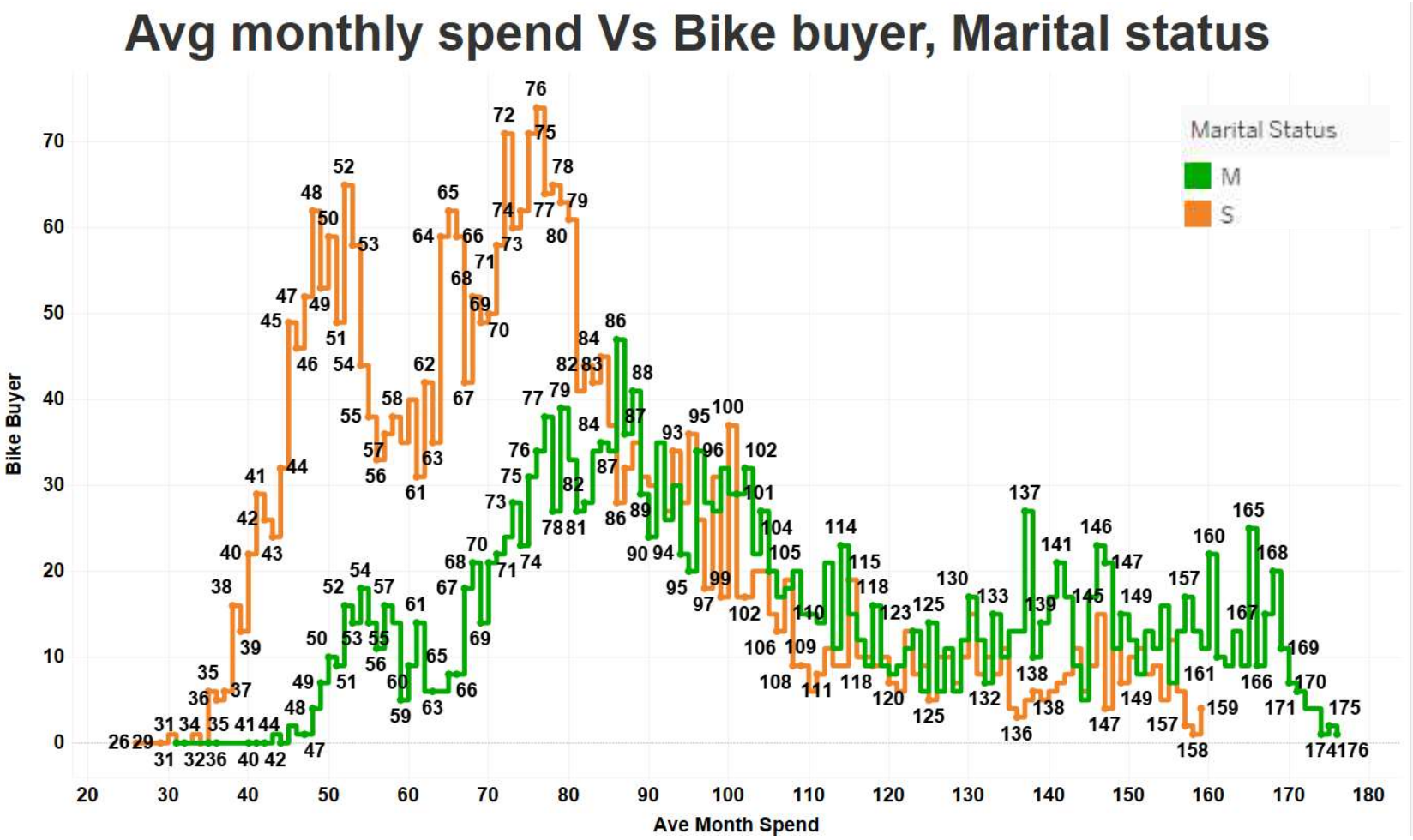


More male bike owners than female bike owners. Most bikes are bought between 40 to 70 years.

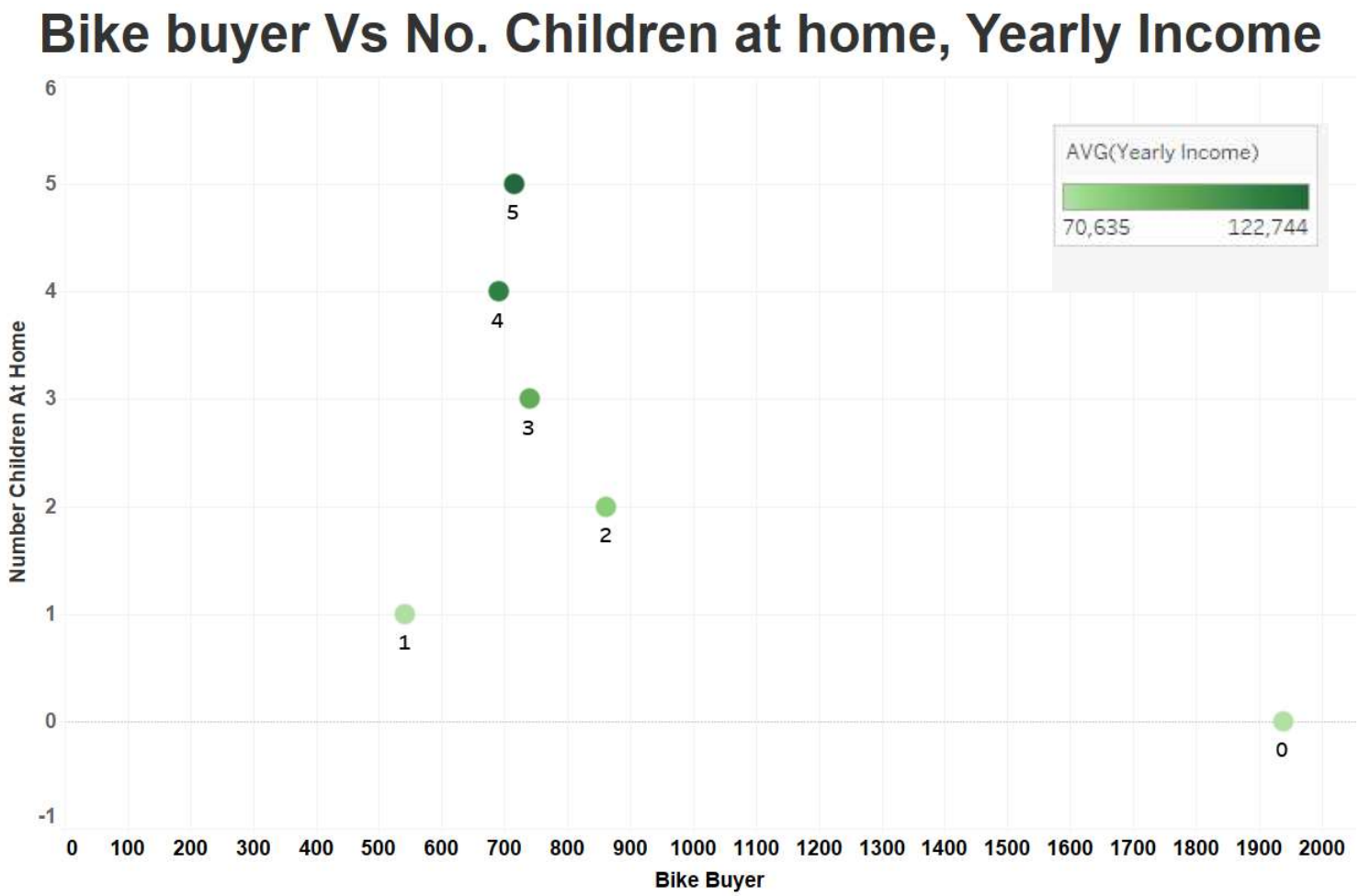
Avg monthly spend Vs Bike buyer, Gender



Men spend more on bikes, than women.

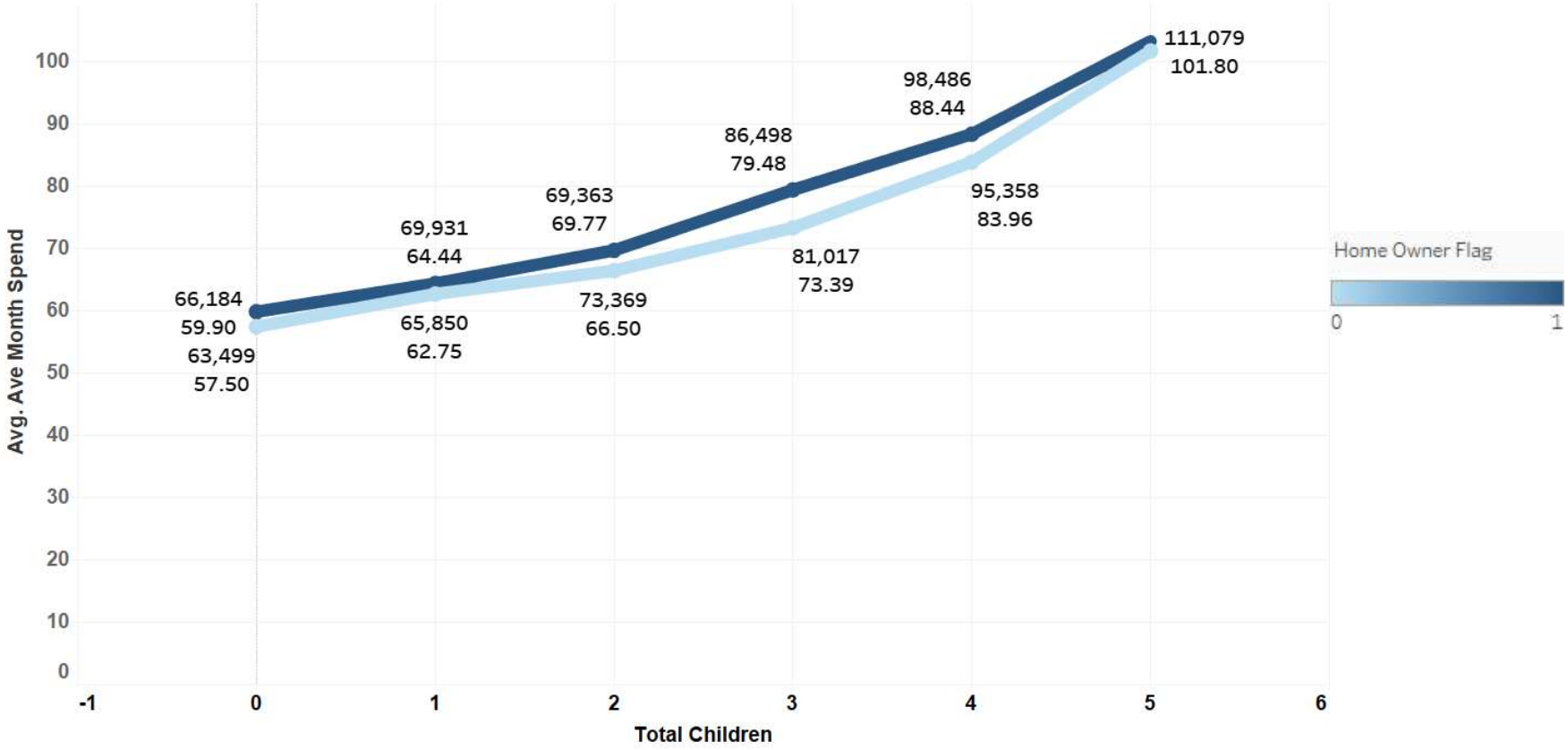


Married people spend less on bikes, than single people.



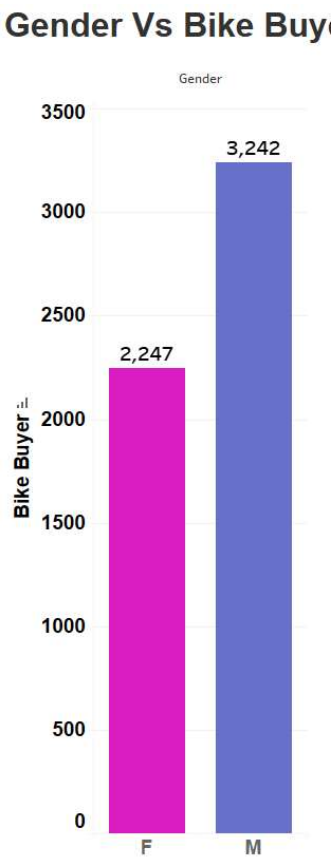
People with high yearly income have more children, but people with 0 children buy most bikes

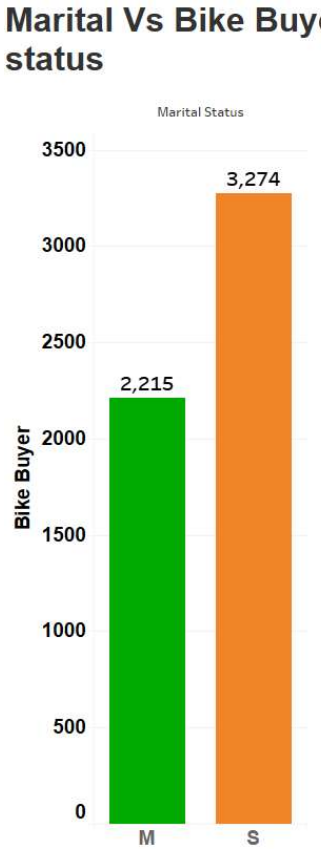
Total Children Vs Avg month spend, Yearly Income, Home Owner flag

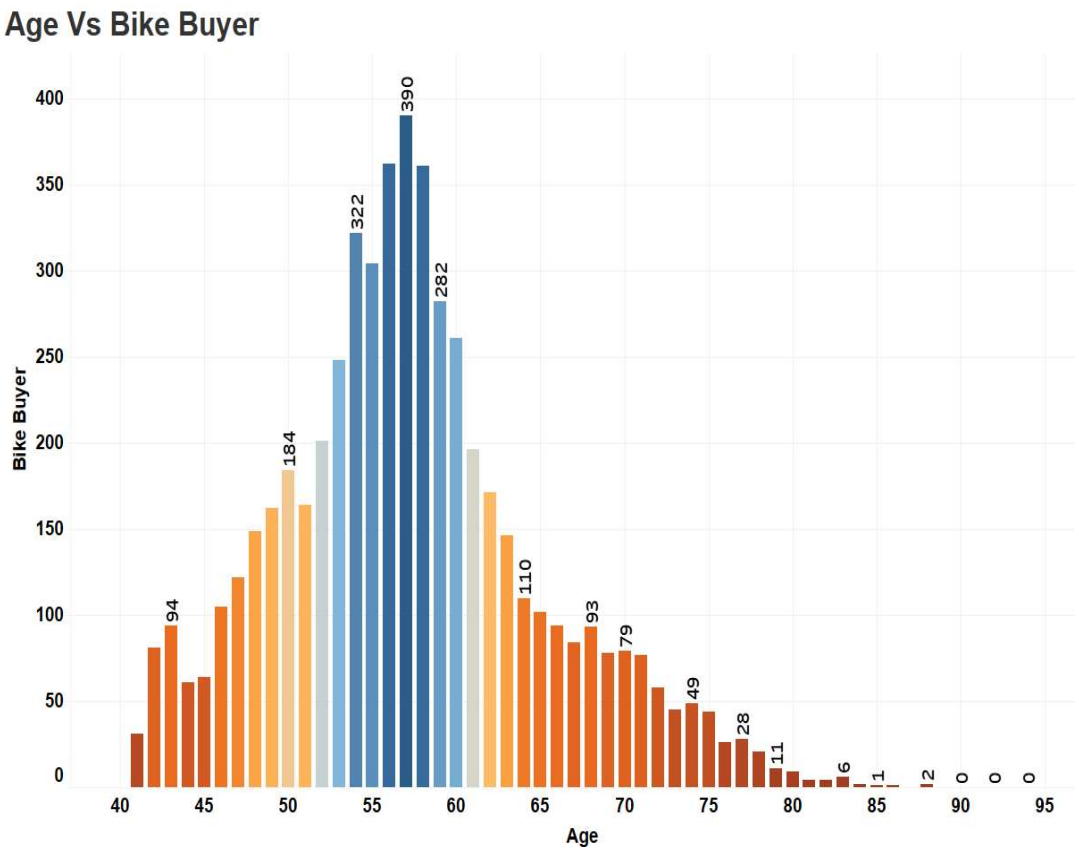


People with more children have high annual income and spend more monthly in bike store.Also people owns a house spend more monthly in bike store, even if they have the same number of children.

Features Vs Bike buyer status

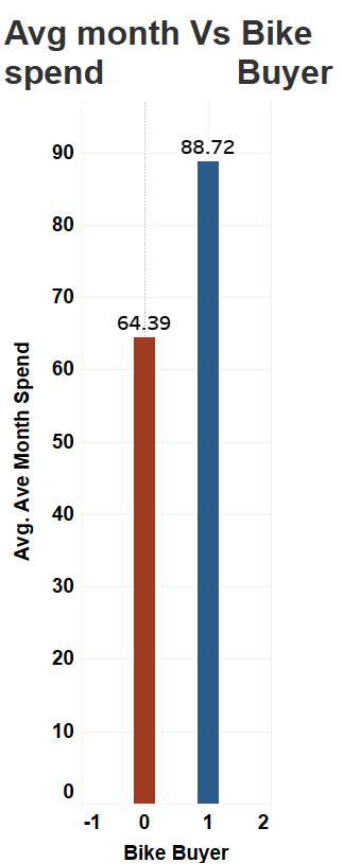


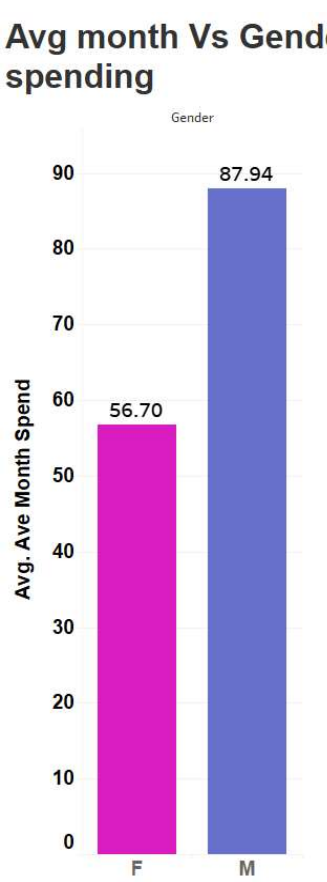


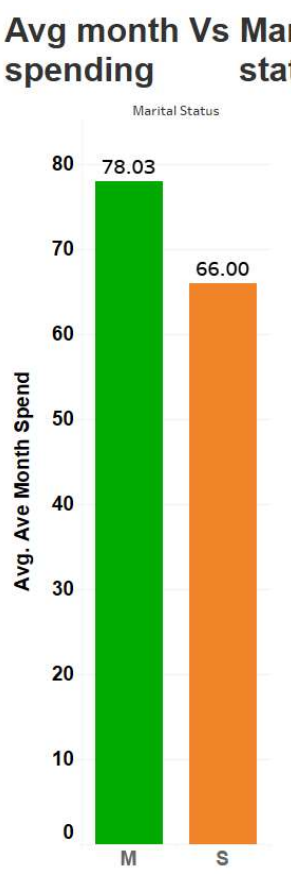


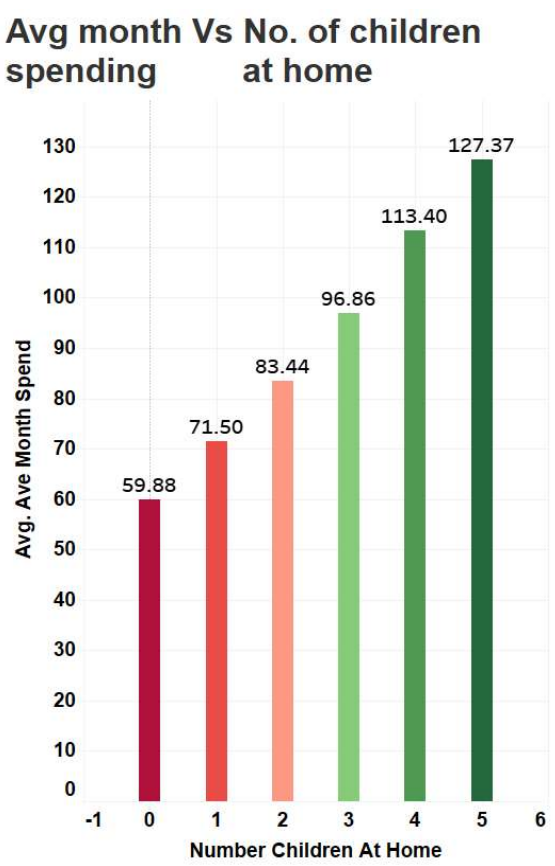
Single, men, people in 50's buy more bikes

Features Vs Avg monthly spending









Men, married people, people who buy more bikes, people with more children at home, spend more monthly in bike store

Data Preparation

Total number of features = 26
Total number of records = 16519

No duplicate values found

Missing values
Title
MiddleName
Suffix
AddressLine2

All the missing values are noise variables - ignore them

Convert DateBirth to Age, added Age as a feature to dataset

Binary/ Boolean
BikeBuyer, HomeOwnerFlag: (0,1)

Exploratory Data Analysis (EDA)

The Features (i.e., Variables) are segregated into three categories:

Dependent Variable (Y): Variable that is being measured in the experiment. It changes as a result of the changes to the independent variables. Y values to predict:

Y : BikeBuyer Status – Classification model

Y : AveMonthSpend – Regression model

Noise: Variable that does not affect the dependent variable.

Independent Variable or Predictor Variable (X): Variable whose change isn't affected by any other variable in the experiment.

Independent variable is the cause, and dependent variable is the effect.

Exploratory Data Analysis - continued

Noise variables

- ❑ Customer ID
- ❑ Title
- ❑ FirstName
- ❑ MiddleName
- ❑ LastName
- ❑ Suffix
- ❑ AddressLine1
- ❑ AddressLine2
- ❑ PhoneNumber
- ❑ BirthDate

Independent Variables

❑ **AveMonthSpend**	❑ Occupation
❑ City	❑ Gender
❑ StateProvinceName	❑ MaritalStatus
❑ CountryRegionName	❑ HomeOwnerFlag
❑ PostalCode	❑ NumberCarsOwned
❑ Age	❑ NumberChildrenAtHome
❑ Education	❑ TotalChildren
❑ **BikeBuyer**	❑ YearlyIncome

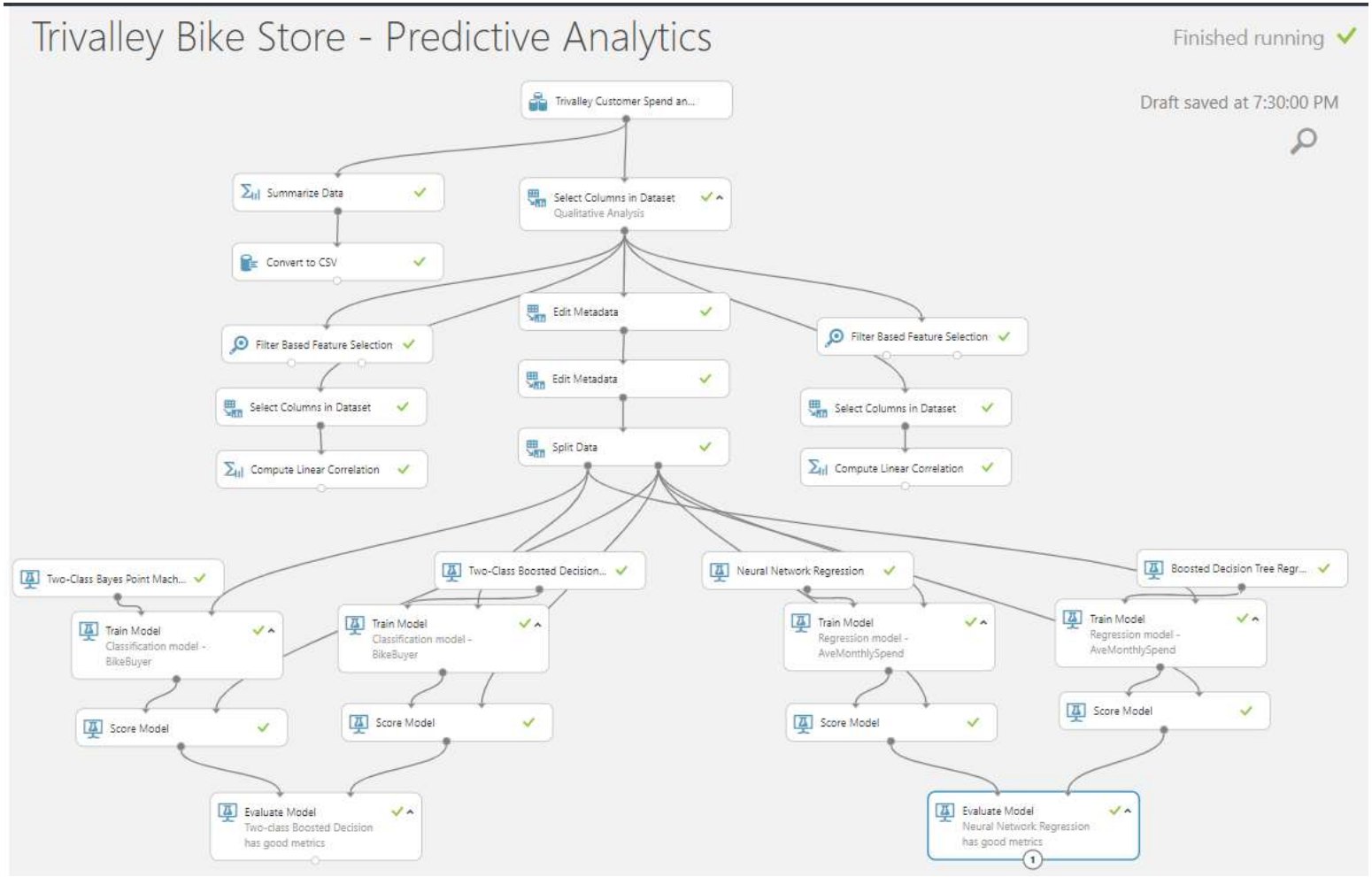
Exploratory Data Analysis - Azure ML

- Qualitative and Quantitative analysis
- Find features impacting **Y (BikeBuyer Status, AveMonthSpend)**
- Find linear relationship between numeric variables
- Edit Metadata – specify/modify new datatype for the column(s)
- Split data - divide dataset into two distinct sets. Use random data, 70% for training and 30% for testing
- Train model - using suitable Azure ML algorithm
- Score model - generate predictions using a trained classification or regression model

  Classification model – gives probability of the predicted value

  Regression model - generates the predicted numeric value
- Evaluate model - measure the accuracy of a trained model using metrics

Azure Predictive model



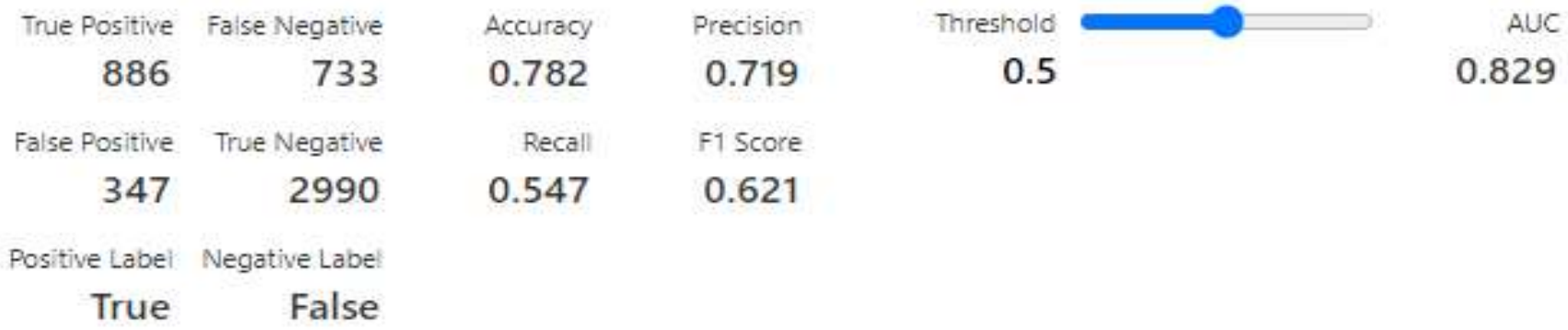
Performance metrics - Classification model

- **Accuracy** measures the proportion of correctly classified results from the total number of cases.
- **Precision** is the proportion of true results over all positive results.
- **Recall** is the ability of a model to detect all positive results.
- **F-score** is the weighted average of precision and recall, where the ideal F-score value is 1.
- **AUC** (Area Under Curve) measures the area under the curve plotted with true positives on y axis and false positives on x axis.
- **True Positives (TP)** are the instances in which the model *correctly* predicted a *positive* results.
- **True Negatives (TN)** are the instances in which the model *correctly* predicted a *negative* result.
- **False Positives (FP)** are the instances in which the model *incorrectly* predicted a *positive* result

  FP: Type 1 error
- **False Negatives (FN)** are the instances in which the model *incorrectly* predicted a *negative* result (

  FN: Type 2 error

Performance metrics - Classification model



Two-Class Bayes point machine and Two-Class Boosted Decision Tree algorithms are trained, scored and evaluated.

Two-Class Boosted Decision Tree algorithm has the best performance metrics.

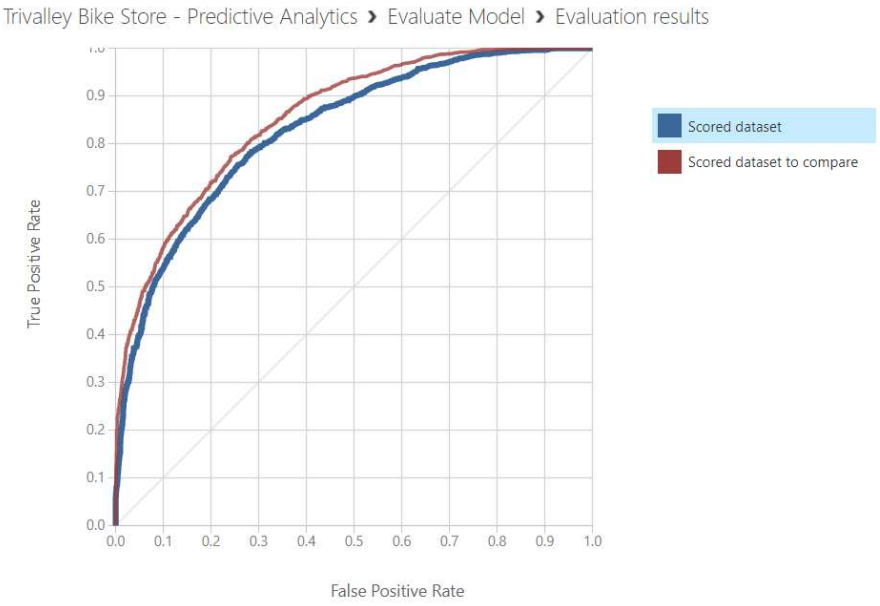
Results show high accuracy (78.2%), high precision (71.9%) values.

ROC Curve – Classification model

ROC (**Receiver Operating Characteristic**) curve describes the binary classifier model's performance.

This plot is most useful when the dataset is balanced.

The ideal plot should arc close to the top-left corner of the chart.



Performance metrics - Regression model

- **Mean absolute error (MAE)** measures how close the predictions are to the actual outcomes; thus, a lower score is better.
- **Root mean squared error (RMSE)** is the square root of the mean of the square of all of the error. It is always non-negative, and a value of 0, would indicate a perfect fit to the data. RMSE values **between 0.2 and 0.5** shows that the model can relatively predict the data accurately.
- **Relative absolute error (RAE)** is the relative absolute difference between expected and actual values
- **Relative squared error (RSE)** normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.
- **Coefficient of determination**, often referred to as R^2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit.

Performance metrics– Regression model

Trivalley Bike Store - Predictive Analytics › Evaluate Model › Evaluation results

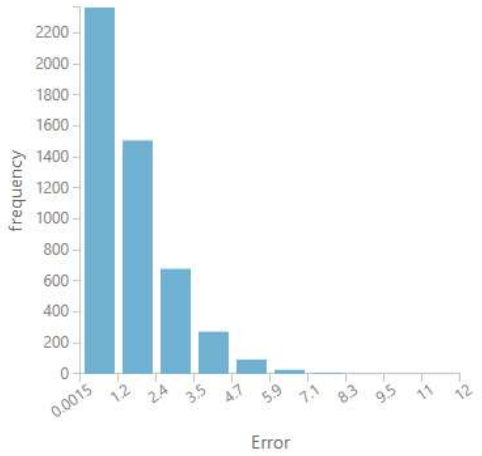
Metrics

Mean Absolute Error	1.56714
Root Mean Squared Error	2.027632
Relative Absolute Error	0.076098
Relative Squared Error	0.005581
Coefficient of Determination	0.994419

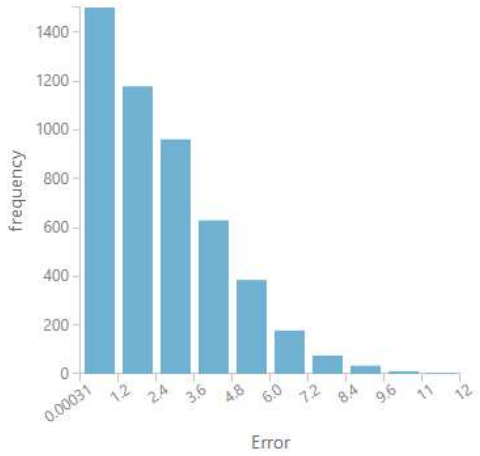
Metrics

Mean Absolute Error	2.554118
Root Mean Squared Error	3.21582
Relative Absolute Error	0.124023
Relative Squared Error	0.014038
Coefficient of Determination	0.985962

Error Histogram



Error Histogram



Neural Network Regression and Boosted Decision Tree Regression algorithms are trained, scored and evaluated.

Neural Network Regression algorithm has the best performance metrics with R^2 value 99.4% and RMSE 2.02

Chi-Squared Test Analysis - BikeBuyer

Chi-squared test is a statistical method that measures how close expected values are to actual results.

Top 5 impacting features on BikeBuyer: **NumberChildrenAtHome, AveMonthSpend, Age, YearlyIncome, NumberCarsOwned**

Bottom 3 impacting features on employee retention: **HomeOwnerFlag, CountryRegionName, StateProvinceName**

Independent variable	Chi-squared test value
NumberChildrenAtHome	3647.0139
AveMonthSpend	2957.0945
Age	1946.4329
YearlyIncome	1405.7185
NumberCarsOwned	1201.2517
TotalChildren	1041.8744
MaritalStatus	614.4861
Occupation	542.3577
Postalcode	381.5548

Linear Correlation Tests Analysis - BikeBuyer

The correlation coefficient **r** measures the strength and direction of a linear relationship between two variables. r is always between +1 (Strong positive) and –1 (Strong negative).

Strong correlation: r > 0.7, Moderate correlation: 0.6 to 0.4 , Weak correlation: r < 0.4

Top 2 features that have moderate linear relationship with **BikeBuyer status**: NumberChildrenAtHome, AveMonthSpend.

All other correlations are weak (< ±0.3)

Independent Variable	R (Independent variable, BikeBuyer)
NumberChildrenAtHome	0.456377
AveMonthSpend	0.421783
YearlyIncome	0.28708
TotalChildren	0.233467
NumberCarsOwned	0.164751
Age	-0.137715
HomeOwnerFlag	0.000552

Chi-Squared Test Analysis - AveMonthSpend

Chi-squared test is a statistical method that measures how close expected values are to actual results.

Top 5 impacting features on BikeBuyer: **NumberChildrenAtHome, YearlyIncome, Gender, TotalChildren, NumberCarsOwned**

Bottom 3 impacting features on employee retention: **CountryRegionName, HomeOwnerFlag, StateProvinceName**

Independent variable	Chi-squared test value
NumberChildrenAtHome	10821.3795
YearlyIncome	9234.7757
Gender	8651.7349
TotalChildren	5025.7259
NumberCarsOwned	3876.6647
Occupation	3708.4833
PostalCode	3545.5967
Age	3097.4688
City	3042.6127

Linear Correlation Analysis - AveMonthSpend

The correlation coefficient **R** measures the strength and direction of a linear relationship between two numeric variables. R is always between +1 (Strong positive) and –1 (Strong negative).

Top 3 features that have strong/moderate linear relationship with **AveMonthSpend**: NumberChildrenAtHome, YearlyIncome, TotalChildren

All other correlations are moderate or weak

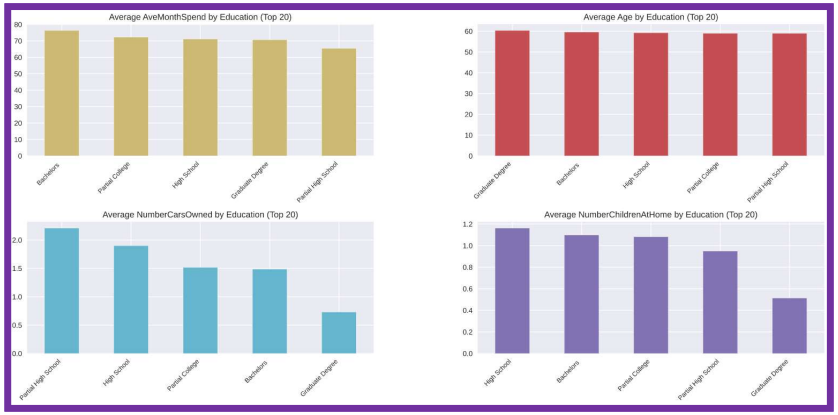
Independent Variable	R (Independent variable, AveMonthSpend)
NumberChildrenAtHome	0.730421
YearlyIncome	0.607616
TotalChildren	0.500159
BikeBuyer	0.421783
NumberCarsOwned	0.34667
HomeOwnerFlag	0.134242
Age	0.014858

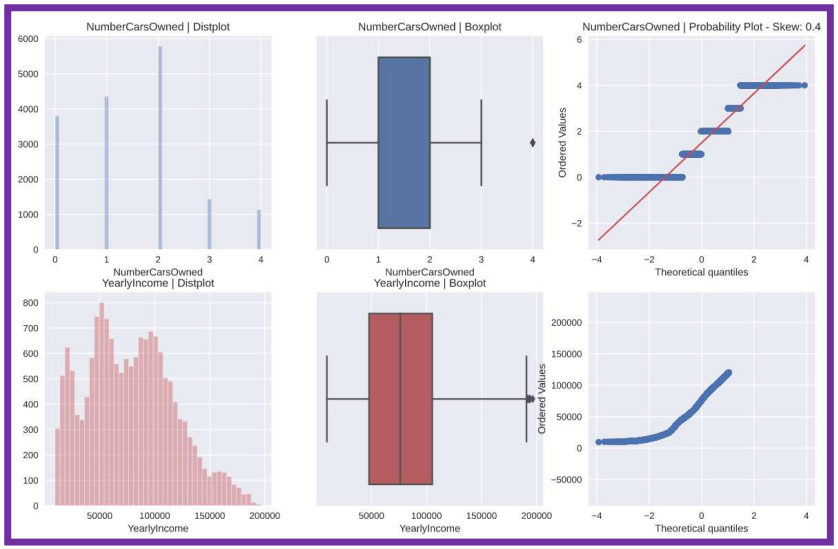
Exploratory Data Analysis - Python

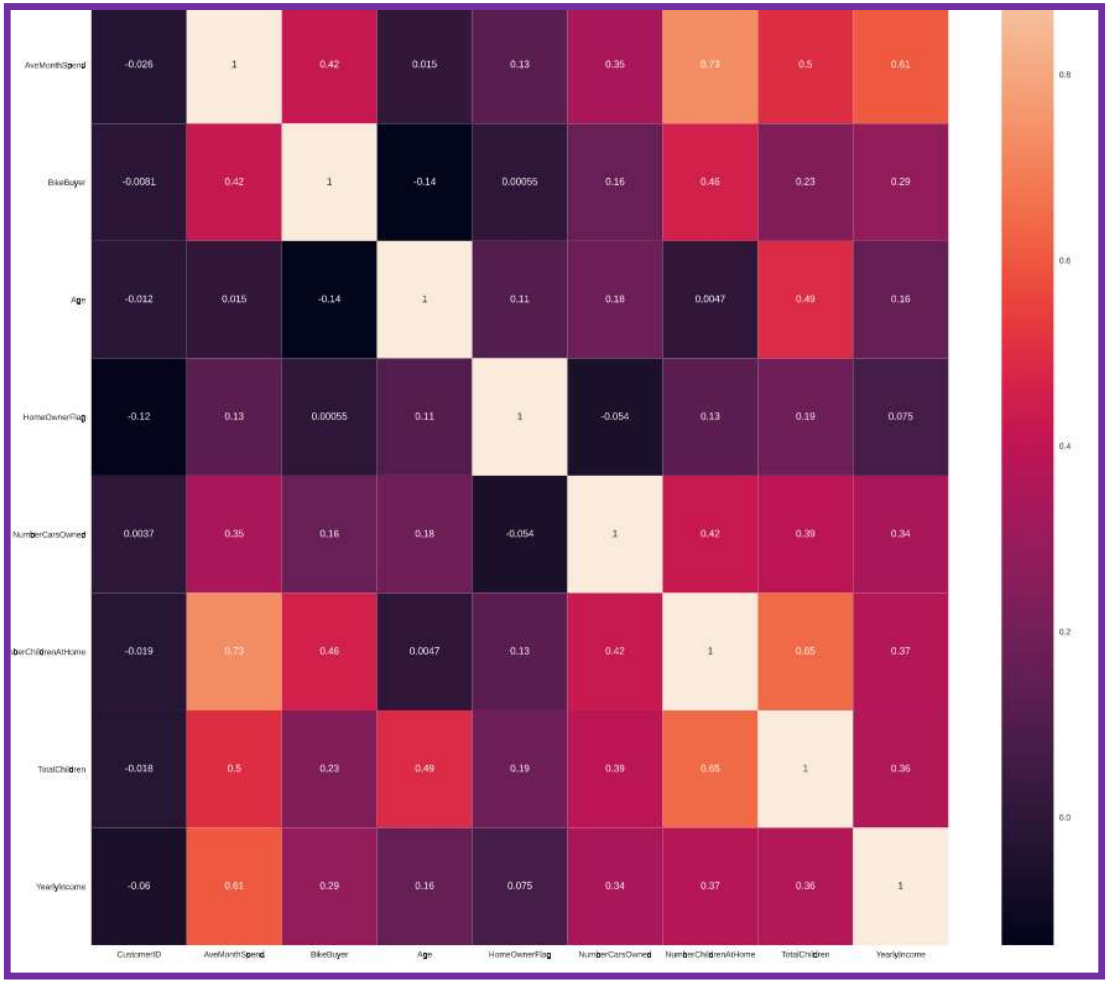
Python has open-source libraries that can automate the whole process of Exploratory Data Analysis and save a lot of time. Some of these popular EDA libraries are:

- **Autoviz** - performs automatic visualization of any dataset with one line
- **Pandas Profiling -** generates interactive HTML reports and describes various aspects of the dataset. Key functionalities include handling missing values, statistics of dataset like mean, mode, median, skewness, standard deviation etc,, charts like histograms.
- **Sweetviz** - generates visualizations which is useful in exploratory data analysis with just a few lines of codes. The library can be used to visualize the variables and comparing the dataset.
- **Pycaret** – is an end-to-end machine learning and model management tool that speeds up the experiment cycle exponentially and makes you more productive.
- **H2O** - supports the most widely used statistical & machine learning algorithms with AutoML functionality that automatically runs through all the algorithms and their hyperparameters to produce a leaderboard of the best models.

AutoViz EDA


Average AveMonthSpend by Education (Top 20)
Average Age by Education (Top 20)
Average NumberCarsOwned by Education (Top 20)
Average NumberChildrenAtHome by Education (Top 20)

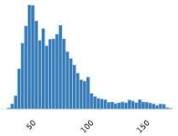

NumberCarsOwned | Distplot
NumberCarsOwned | Boxplot
NumberCarsOwned | Probability Plot - Skew: 0.4
Ordered Values
Theoretical quantiles
NumberCarsOwned
YearlyIncome | Distplot
YearlyIncome | Boxplot
YearlyIncome
Ordered Values
Theoretical quantiles


AveMonthSpend
BikeBuyer
Age
HomeOwnerFlag
NumberCarsOwned
TotalChildren
YearlyIncome
CustomerID

Pandas Profiling EDA

AveMonthSpend
Real number (ℝ≥0)

Distinct	152	Minimum	22
Distinct (%)	0.9%	Maximum	176
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	72.40595678	Memory size	129.2 KiB



Toggle details

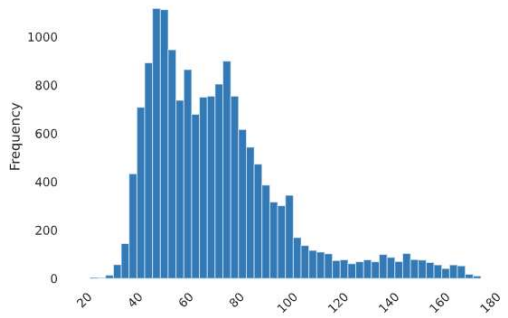
Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	22
5-th percentile	41
Q1	52
median	68
Q3	84
95-th percentile	135
Maximum	176
Range	154
Interquartile range (IQR)	32

Descriptive statistics

Standard deviation	27.28537049
Coefficient of variation (CV)	0.3768387534
Kurtosis	1.738203009
Mean	72.40595678
Median Absolute Deviation (MAD)	16
Skewness	1.312897404
Sum	1196074
Variance	744.4914429
Monotonicity	Not monotonic



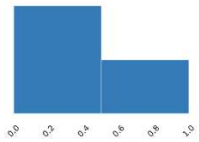
Minimum 5 values | Maximum 5 values

Value	Count	Frequency (%)
0	11030	66.8%
1	5489	33.2%

BikeBuyer
Real number (ℝ≥0)

ZEROS

Distinct	2	Minimum	0
Distinct (%)	< 0.1%	Maximum	1
Missing	0	Zeros	11030
Missing (%)	0.0%	Zeros (%)	66.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.3322840366	Memory size	129.2 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

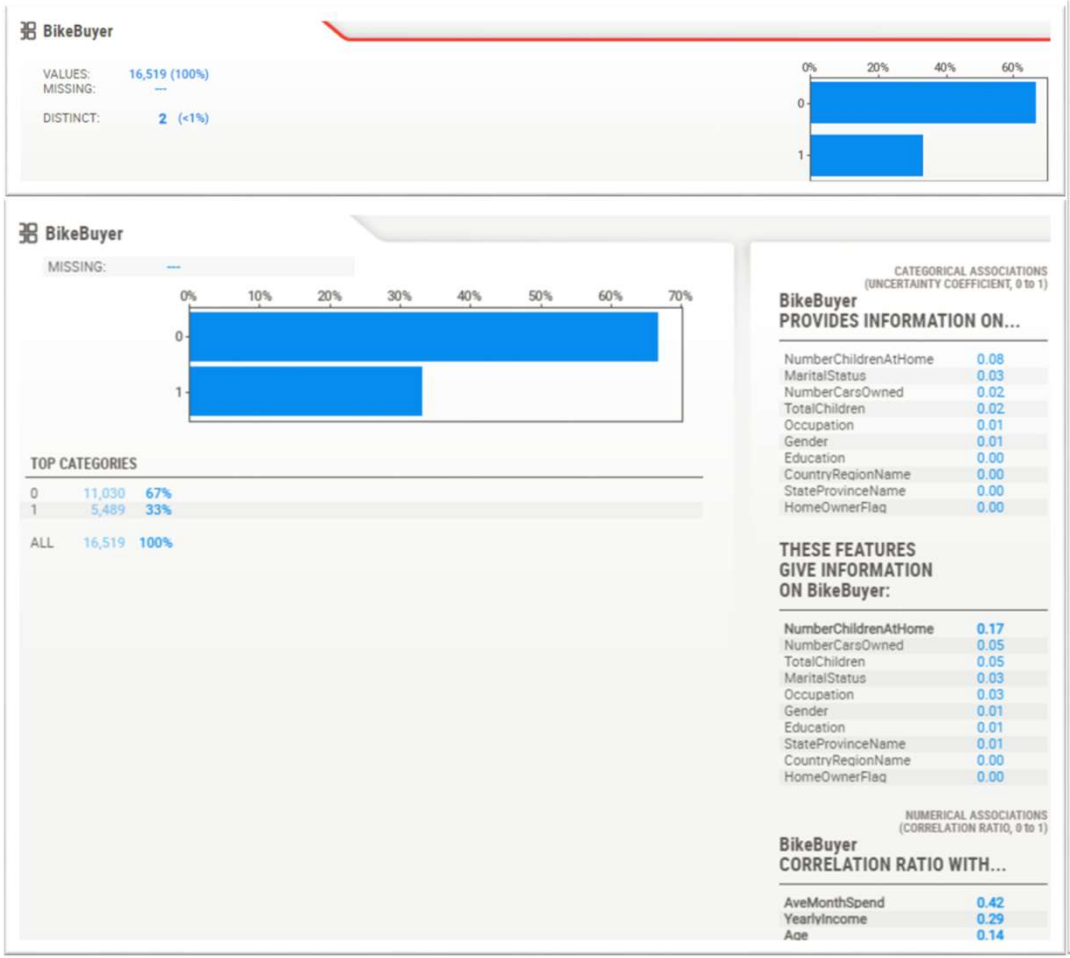
Quantile statistics

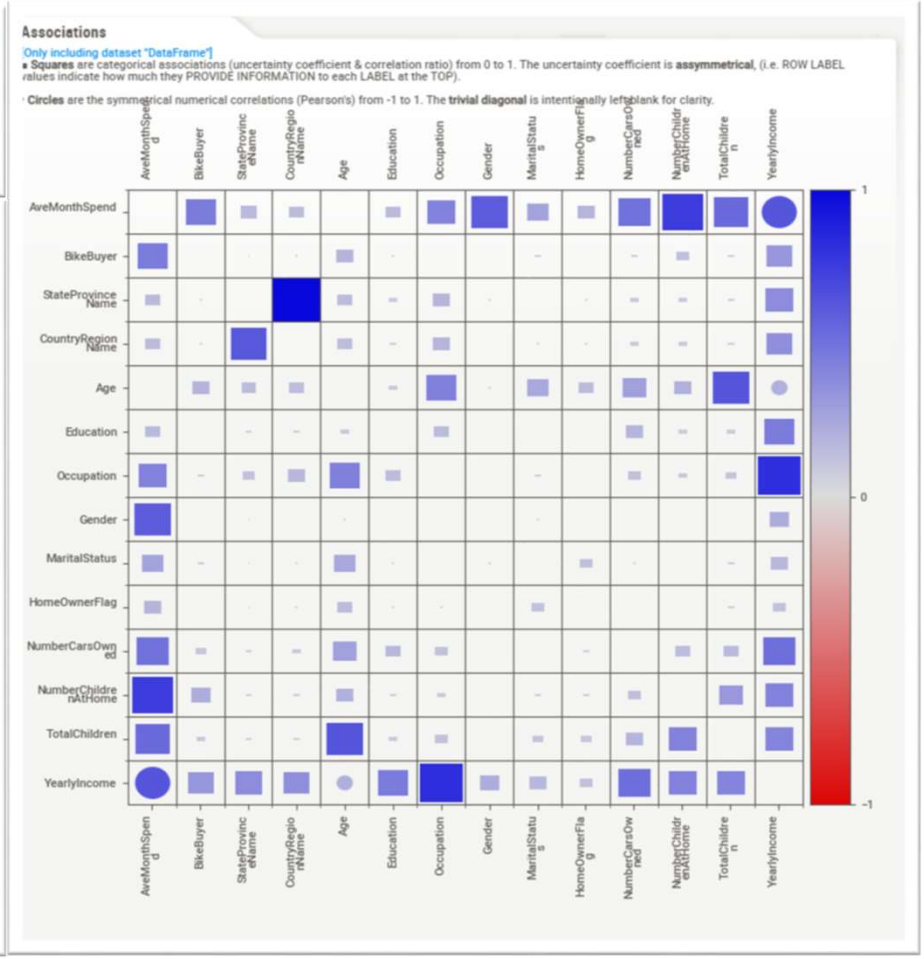
Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	1
95-th percentile	1
Maximum	1
Range	1
Interquartile range (IQR)	1

Descriptive statistics

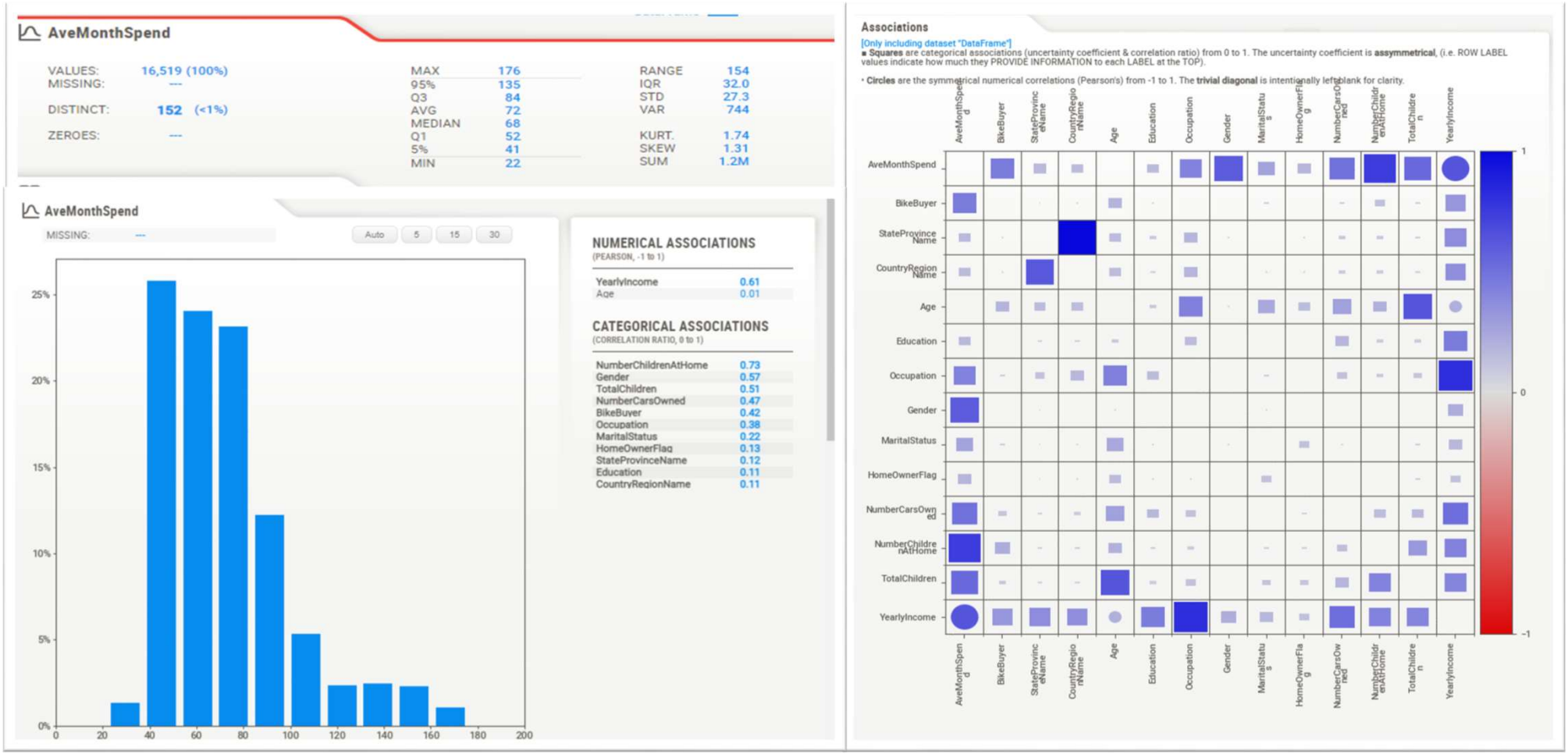
Standard deviation	0.4710464815
Coefficient of variation (CV)	1.4176019
Kurtosis	-1.492972376
Mean	0.3322840366
Median Absolute Deviation (MAD)	0
Skewness	0.7121856476
Sum	5489
Variance	0.2218847877
Monotonicity	Not monotonic

Sweetviz EDA (Classification model)





Sweetviz EDA (Regression model)

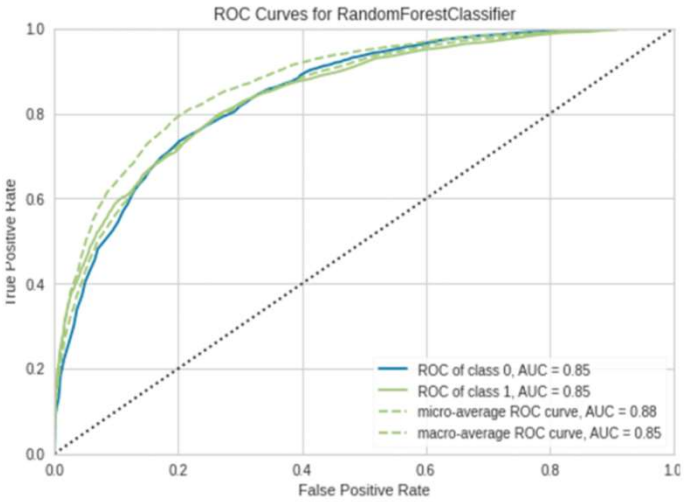


Pycaret EDA (Classification model)

```
1 # Code snippet 25
2 # Compare and train all Classification models to evaluate performance
3 compare_models(budget_time=1) # time limit of 1 min
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7843	0.8374	0.5624	0.7264	0.6336	0.4843	0.4925	3.358
ridge	Ridge Classifier	0.7776	0.0000	0.5369	0.7220	0.6155	0.4638	0.4741	0.129
nb	Naive Bayes	0.7649	0.8055	0.4485	0.7416	0.5582	0.4112	0.4356	0.095
dt	Decision Tree Classifier	0.7441	0.7072	0.5976	0.6187	0.6079	0.4180	0.4183	0.393
lr	Logistic Regression	0.7056	0.7182	0.3192	0.6029	0.4162	0.2451	0.2670	0.284
knn	K Neighbors Classifier	0.6587	0.6079	0.3599	0.4807	0.4113	0.1782	0.1820	0.810
svm	SVM - Linear Kernel	0.5105	0.0000	0.5567	0.2239	0.3088	0.0414	0.0578	1.828

```
1 # Code snippet 21
2 plot_model(trained_model)
```



Random Forest Classifier is the best classification model with best performance metrics

Pycaret EDA (Regression model)

```
# Code snippet 19
# Compare and train all Regression models to evaluate performance
compare_models(budget_time=1) # time limit of 1 min
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dt	Decision Tree Regressor	3.3049	17.8348	4.2225	0.9762	0.0677	0.0515	0.272
omp	Orthogonal Matching Pursuit	4.8766	40.6277	6.3710	0.9460	0.0917	0.0707	0.075
br	Bayesian Ridge	4.8762	40.7206	6.3785	0.9459	0.0911	0.0706	0.604
ridge	Ridge Regression	4.9053	41.1002	6.4081	0.9454	0.0921	0.0711	0.069
lr	Linear Regression	5.3789	48.7310	6.9434	0.9345	0.1020	0.0788	0.375
lasso	Lasso Regression	5.1883	50.3044	7.0862	0.9333	0.0869	0.0705	0.055
en	Elastic Net	10.0649	153.1868	12.3722	0.7964	0.1582	0.1421	0.247
huber	Huber Regressor	14.3708	361.5639	18.4441	0.5188	0.2502	0.2063	1.854
knn	K Neighbors Regressor	17.2705	512.8132	22.6353	0.3188	0.2900	0.2517	0.490
llar	Lasso Least Angle Regression	20.7699	754.9038	27.4558	-0.0008	0.3467	0.3078	0.287
par	Passive Aggressive Regressor	23.2764	877.0547	29.2954	-0.1916	0.4173	0.3276	1.152

```
plot_model(trained_model)
```

Decision Tree Regressor is the best regression model with best performance metrics

H2O EDA (Classification model)

The current version of AutoML in H2O, trains and cross-validates a default Random Forest, an Extremely-Randomized Forest, a random grid of Gradient Boosting Machines (GBMs), a random grid of Deep Neural Nets, a fixed grid of Generalized Linear Model (GLMs), and then trains two Stacked Ensemble models at the end.

The H2O AutoML interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time constraint or limit on the number of total models trained.

Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
0	max f1	0.291442	0.692308	260.0
1	max f2	0.141998	0.792655	346.0
2	max f0point5	0.569330	0.728564	131.0
3	max accuracy	0.569330	0.797546	131.0
4	max precision	0.906448	1.000000	0.0
5	max recall	0.086393	1.000000	390.0
6	max specificity	0.906448	1.000000	0.0
7	max absolute_mcc	0.462732	0.524586	176.0
8	max min_per_class_accuracy	0.265559	0.767890	273.0
9	max mean_per_class_accuracy	0.291442	0.770389	260.0
10	max tns	0.906448	2170.000000	0.0
11	max fns	0.906448	1088.000000	0.0
12	max fps	0.075288	2170.000000	399.0
13	max tps	0.086393	1090.000000	390.0
14	max tnr	0.906448	1.000000	0.0
15	max fnr	0.906448	0.998165	0.0
16	max fpr	0.075288	1.000000	399.0
17	max tpr	0.086393	1.000000	390.0

H2O EDA (Regression model)

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on test data. **

MSE: 10.566096627427786
RMSE: 3.2505532802013546
MAE: 2.566272855495790 8
RMSLE: 0.05321449189077472
R^2: 0.9857956781174668
Mean Residual Deviance: 10.566096627427786
Null degrees of freedom: 3259
Residual degrees of freedom: 3254
Null deviance: 2427054.4419247042
Residual deviance: 34445.475005414584
AIC: 16951.419683376396
```

Questions?