Denver 2016 B-cycle

Data Exploration
Regression Machine Learning
Classification Machine Learning

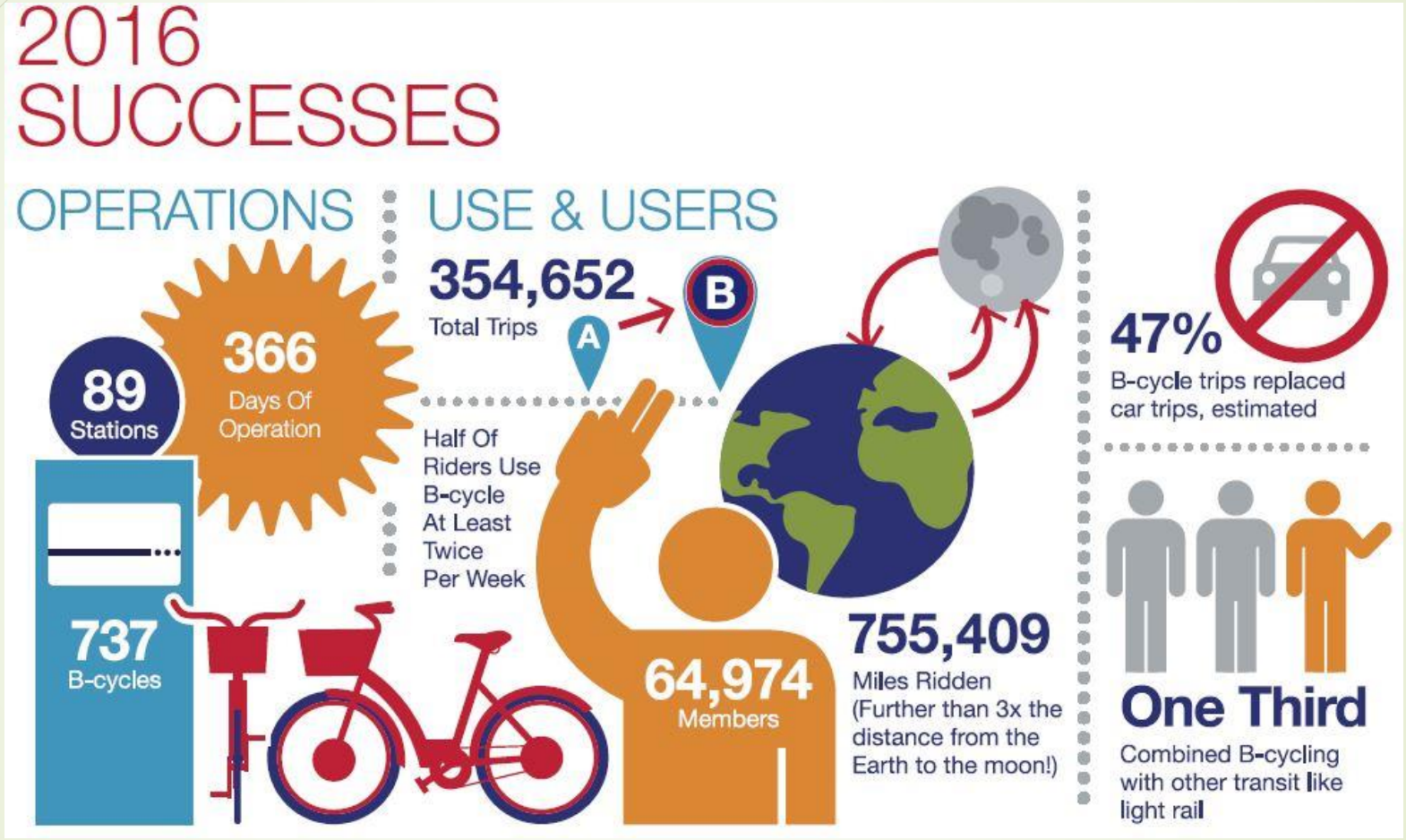
Harpreet Bhasin

Denver B-cycle

- Non-profit public organization
- Owns and operates an automated public bike sharing system
- Has 737 bicycles and 89 kiosks located throughout downtown Denver and nearby areas
- Complements and integrates with Denver's comprehensive metropolitan transportation
- Contributes to Denver becoming the healthiest and greenest city in America
- Encourages the replacement of short car trips for recreational, social and functional purposes

Source: https://www.denverbcycle.com/company

Denver Bike Share

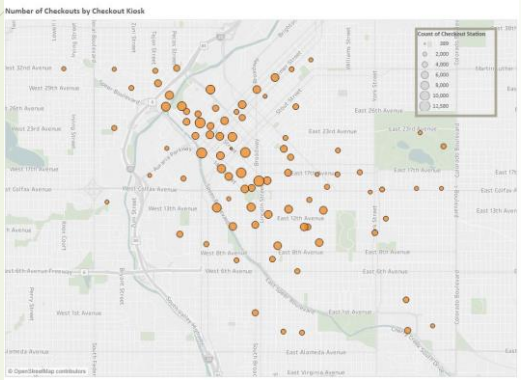


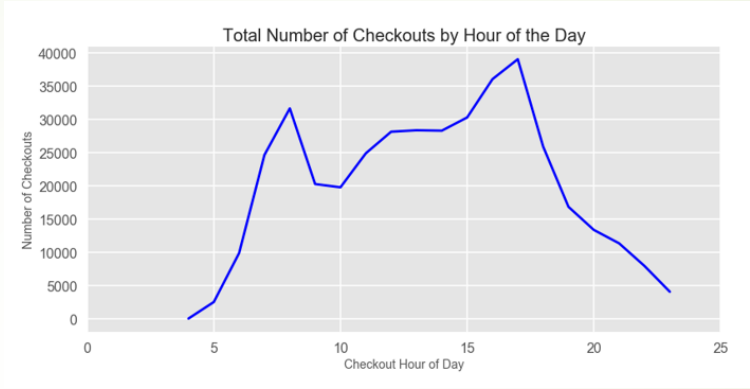
Source: https://denver.bcycle.com/docs/librariesprovider34/default-document-library/dbs_annualreport_2016_05.pdf

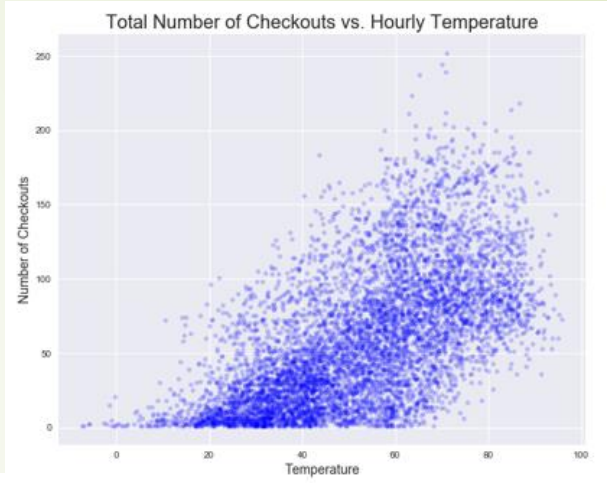
The Objective

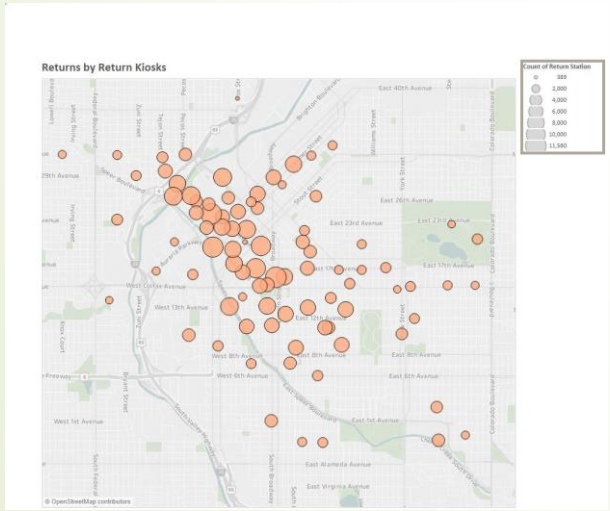
- Explore the publicly available 2016 Trips dataset and visualize the data to provide useful and interesting information
- Deploy a variety of regression machine learning models to predict number of bike checkouts using a combination of calendar, clock and weather attributes
- Deploy variety of classification machine learning models to predict number of bike checkouts using a combination of calendar, clock and weather attributes
- Provide and/or present findings to Denver B-cycle executives to improve future ridership

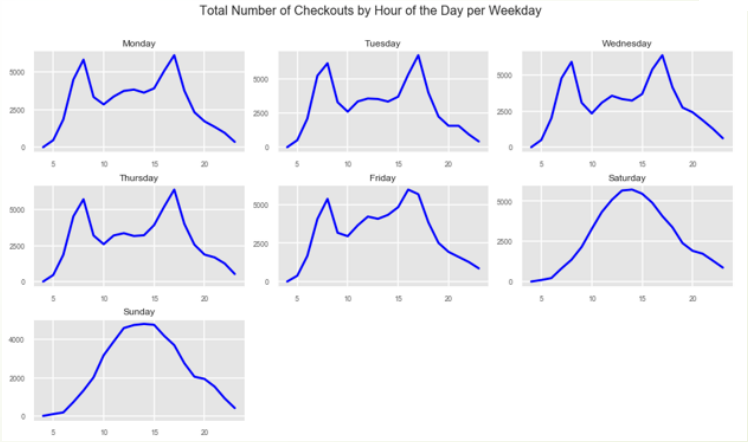
Step 1- Data Exploration

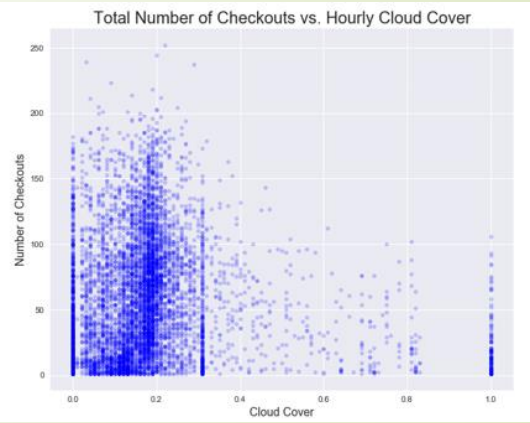












Step 2 – Regression Models

- Predict number of bike checkouts using the following models
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Bayesian Ridge Regression
  - Decision Tree Regression
  - Random Forest Regression
  - Extra Trees Regression
  - Nearest Neighbors Regression

Step 3 – Classification Models

- Predict number of bike checkouts using the following models
  - Logistic Regression Classification
  - Decision Tree Classification
  - Random Forest Classification
  - Extra Trees Classification
  - Naïve Bayes Classification
  - Gradient Boosting Classification
  - Nearest Neighbors Regression
  - Multi-Layer Perceptron Classification

Results – Regression Models

Regression Modeling Summary – Categorical Feature Set

	Linear	Lasso	Ridge	Bayesian Ridge	Decision Tree	Random Forest	Extra Trees	Nearest Neighbors
Training Test Score	0.676	0.676	0.676	0.676	1.000	0.969	1.000	0.575
Test Set Score	0.696	0.696	0.696	0.696	0.718	0.825	0.840	0.476
R Squared	0.834519	0.834457	0.834457	0.834448	0.847276	0.908443	0.916278	0.690249
RMSE	627.95439	628.16826	628.16826	628.19832	583.57445	361.43485	331.86035	1082.98114

- Extra Trees is best performing regressor with 44 features
- Random Forest is pretty close

Regression Modeling Summary – Numerical Feature Set

	Linear	Lasso	Ridge	Bayesian Ridge	Decision Tree	Random Forest	Extra Trees	Nearest Neighbors
Training Test Score	0.433	0.433	0.433	0.433	1.000	0.975	1.000	0.880
Test Set Score	0.448	0.447	0.447	0.447	0.741	0.854	0.838	0.646
R Squared	0.669090	0.668243	0.668243	0.668785	0.861079	0.924077	0.915609	0.803447
RMSE	1142.475	1144.818	1144.818	1143.319	534.800	302.172	334.397	733.229

- Random Forest is best performing regressor with 9 features
- Either Extra Trees or Random Forest regressor is recommended for 44 or 9 features

Results – Classification Models

Classification Modeling Summary – Categorical Feature Set

	Logistic	Decision Tree	Random Forest	Extra Trees	Naïve Bayes	Nearest Neighbors	Gradient Boosting	Multi-Layer Perceptron
Accuracy	0.671898	0.639051	0.660949	0.69080	0.360584	0.549635	0.697080	0.713869
F1 (macro)	0.496528	0.524387	0.486226	0.576138	0.290211	0.322821	0.564894	0.556326
F1 (micro)	0.671898	0.639051	0.660949	0.697080	0.360584	0.549635	0.697080	0.713869
Precision (macro)	0.565525	0.524363	0.597933	0.603859	0.355717	0.393611	0.595219	0.620760
Precision (micro)	0.671898	0.639051	0.660949	0.697080	0.360584	0.549635	0.697080	0.713869
Recall (macro)	0.487240	0.524940	0.464222	0.557927	0.408047	0.319211	0.547159	0.550908
Recall (micro)	0.671898	0.639051	0.660949	0.697080	0.360584	0.549635	0.697080	0.713869

Classification Modeling Summary – Numerical Feature Set

	Logistic	Decision Tree	Random Forest	Extra Trees	Naïve Bayes	Nearest Neighbors	Gradient Boosting	Multi-Layer Perceptron
Accuracy	0.577007	0.656934	0.670073	0.665328	0.500365	0.589781	0.70146	0.606569
F1 (macro)	0.325555	0.552091	0.50567	0.508443	0.299163	0.447329	0.571128	0.328199
F1 (micro)	0.577007	0.656934	0.670073	0.665328	0.500365	0.589781	0.70146	0.606569
Precision (macro)	0.337289	0.559901	0.5497	0.554362	0.388027	0.449879	0.617866	0.372159
Precision (micro)	0.577007	0.656934	0.670073	0.665328	0.500365	0.589781	0.70146	0.606569
Recall (macro)	0.334286	0.545906	0.489289	0.488079	0.348306	0.445305	0.550013	0.374967
Recall (micro)	0.577007	0.656934	0.670073	0.665328	0.500365	0.589781	0.70146	0.606569

- Multi-Layer Perceptron is best performing classifier with 44 features
- Gradient Boosting is pretty close

Gradient Boosting is the recommended classifier with 44 or 9 features

Next Steps

- Undertake similar project for Boulder B-cycle
- Longmont, CO has just introduced its bike sharing system – this study could be useful