PREDICTING THE COUNT OF BIKE RENTALS BASED ON VARIOUS FACTORS

BIKE SHARING DEMAND - DATA ANALYSIS

- **Adedamola Adebamiro**
- **Adjoa Light Myra Dagba**
- **Jesse Annan**
- **Prince Osei Bonsu**

PROBLEM DESCRIPTION

- Africa is one of the largest continents in the world, with an estimated 1.2 billion people. This massive population causes major transportation problems.
- **Motivation:** Experience, providing alternative faster, cheaper, and easier transport, maximizing business profit.
- **Solution:** BUILD A PREDICTIVE MODEL THAT HELPS BIKE RENTAL COMPANIES PREDICT WHEN TO INCREASE THEIR BIKE INVENTORY FOR RENTAL THEREBY MAXIMIZING PROFIT WHILE SOLVING PROBLEM

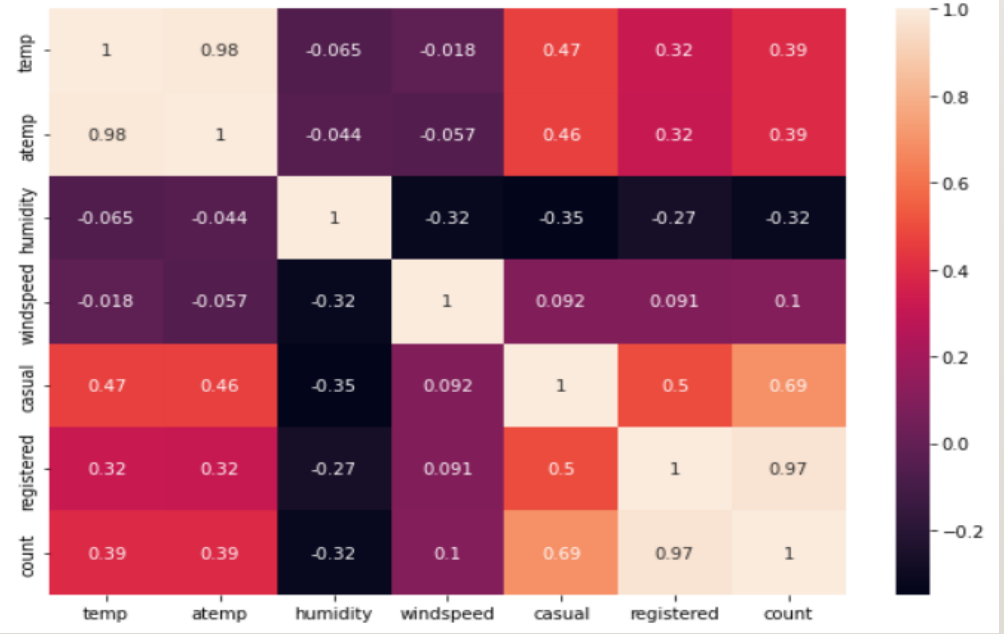
DATASET

Year
Month
Day
Dayofweek
Hour

Data Fields	Data Type	Description
Datetime	datetime64(ns)	Hourly date + Timestamp
Season	object	1 = Spring, 2 = Summer, 3 = Fall, 4 = Winter
Holiday	object	Whether the day is a holiday or not
Working Day	object	Whether the day is a working day or not
Weather	object	1 = Clear, 2 = Cloudy, 3 = Light Rain, 4 = Snow
Temp	float64	Temperature in Celsius
Atemp	float64	'Feels like' temperature in Celsius
Humidity	int64	Relative humidity
Windspeed	float64	Wind speed
Casual	int64	Number of non-registered bike rentals initiated
Registered	int64	Number of registered bike rentals initiated
Count	int64	Total number of bike rentals

10,886 entries with 4 continuous features, 9 categorical features, and 3 target features

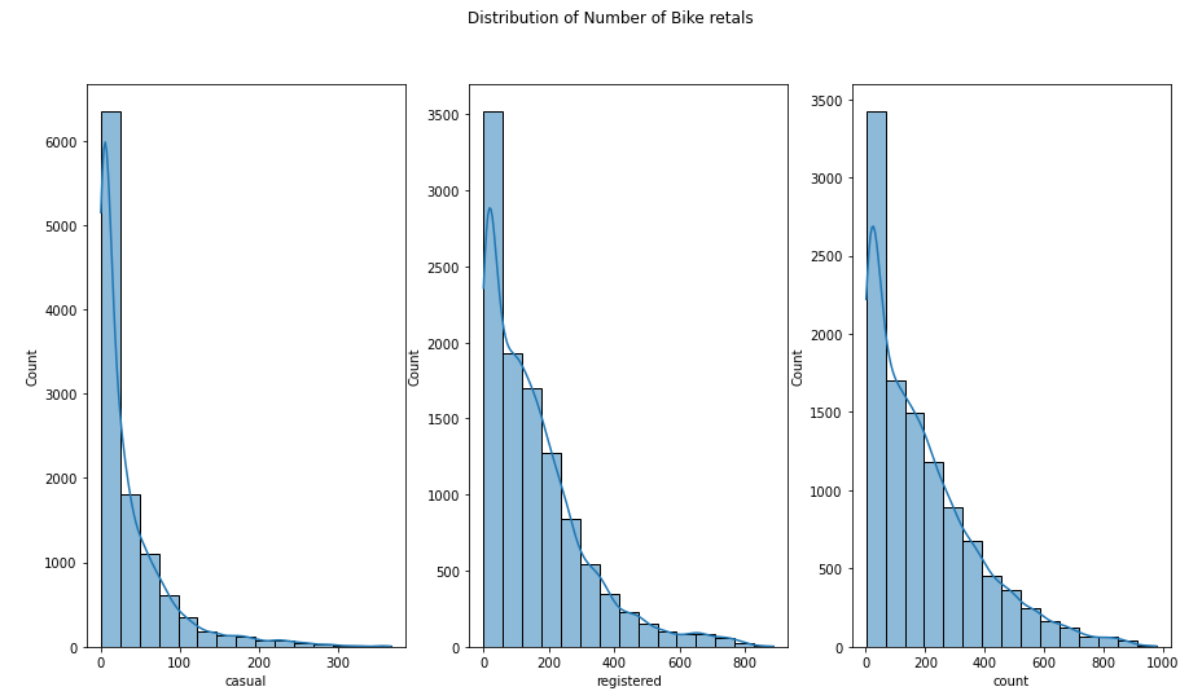
Kaggle.com: 'Bike Sharing Demand'



DATA EXPLORATION AND PREPROCESSING

- Heatmaps used for feature selection
- Features ‘temp’ and ‘humidity’ are kept from the continuous features
- Features ‘year’, ‘hour’, ‘season’, ‘workingday’, and ‘weather’ are kept from categorical features

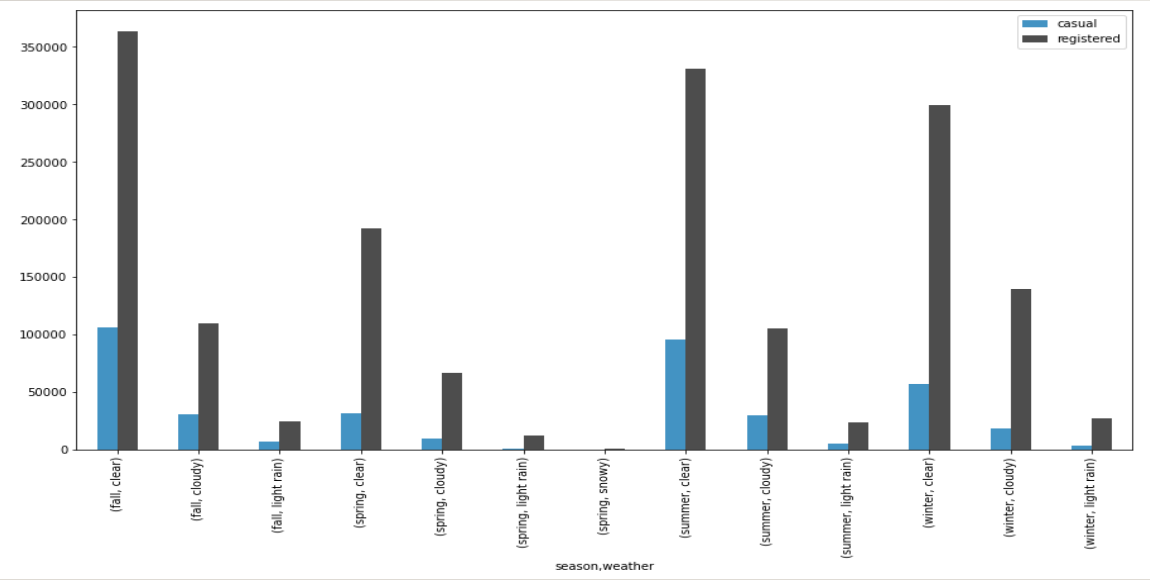
DATA EXPLORATION AND PROCESSING (CONT'D)

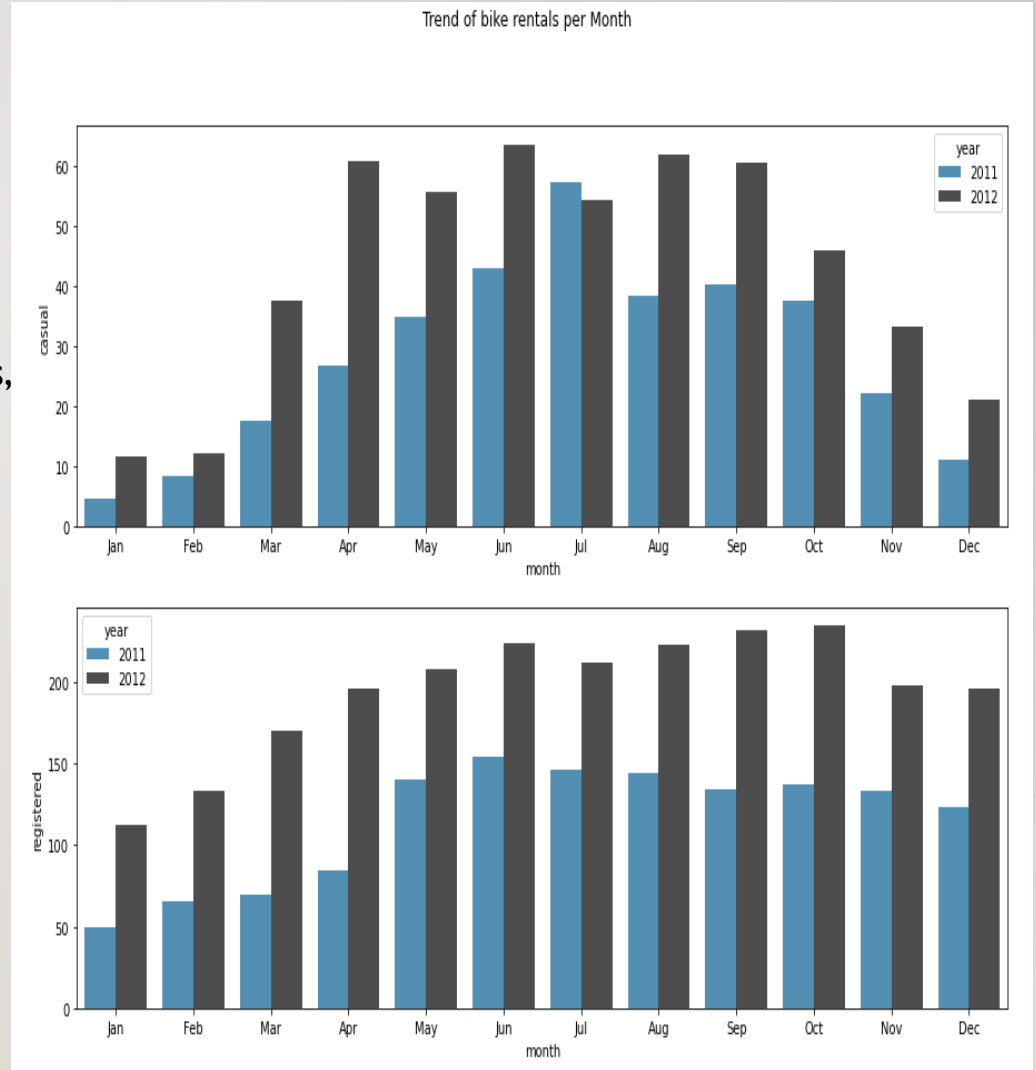


- Highly skewed distributions for all target features
- Therefore, we use the IQR method for binning when converting our regression problem into a classification problem

DATA EXPLORATION AND PROCESSING (CONT'D)

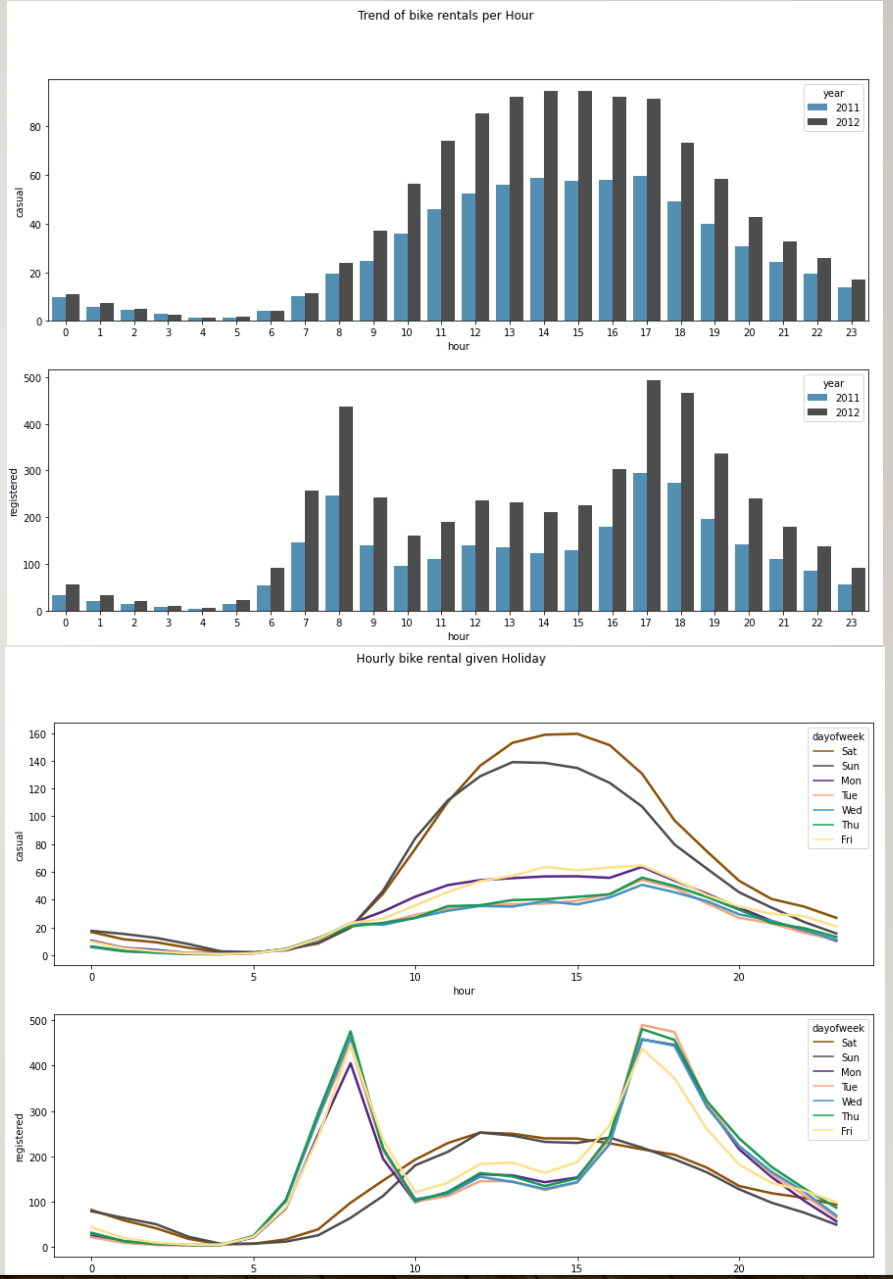
- Bike demand increased from 2011 to 2012: casual and registered
- Highest demand for all seasons occur in clear weather conditions, lowest demand in snow and rain





DATA EXPLORATION AND PROCESSING (CONT'D)

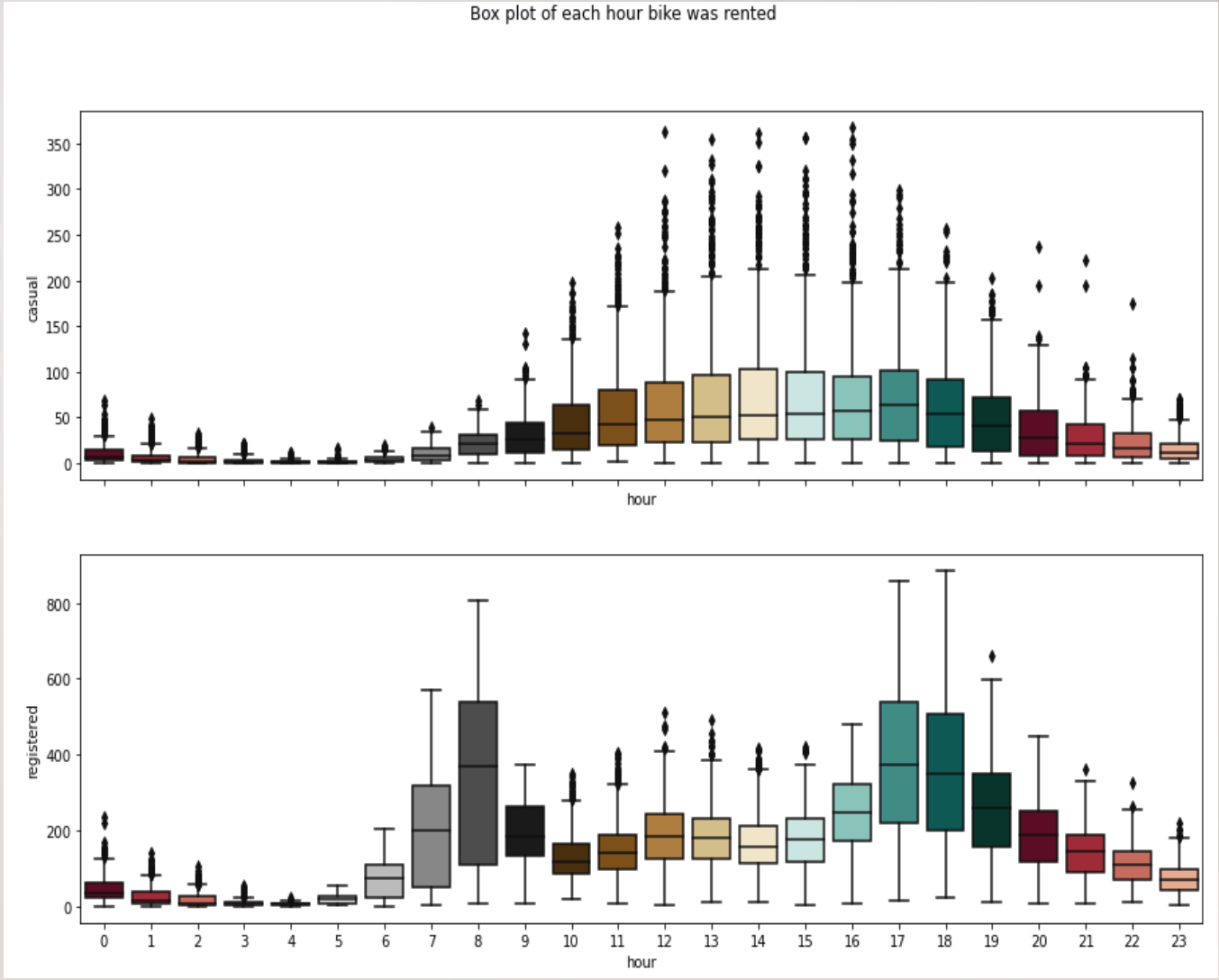
- Casual users rent bikes most in the middle of the day
- Registered users rent bikes the most in the mornings and evenings, suggesting that these users use bikes as a means of transport to and from work
- Casual users rent bikes most on the weekends, in the middle of the day
- Registered users rent bikes the most during the weekdays, in the morning and evening hours



HANDLING POTENTIAL OUTLIERS

- Boxplots show possible outliers
- 'windspeeds' **over 29mph (dangerous)** are removed
- Data set trimmed by removing the top and bottom 2%.

	windspeed			
	min	mean	std	max
season				
fall	0.0	11.508862	7.160605	56.9969
spring	0.0	14.636978	9.150257	51.9987
summer	0.0	13.405607	7.992956	47.9988
winter	0.0	11.678147	7.842632	43.0006



DATA QUALITY REPORTS

Features	Description	Count	% of Missing	Card.	Min #	Q1	Median	Q3	Max #	Std. Dev.
temp	Temperature in Celsius	10886	0.0	49	0.82	13.9400	20.500	26.2400	41.0000	7.791590
atemp	Feels like temperature in Celsius	10886	0.0	60	0.76	16.6650	24.240	31.0600	45.4550	8.474601
humidity	Relative humidity	10886	0.0	89	0.00	47.0000	62.000	77.0000	100.0000	19.245033
windspeed	Wind speed	10886	0.0	28	0.00	7.0015	12.998	16.9979	56.9969	8.164537

Continuous

Features	Description	Count	% of Missing	Card.	1st Mode	1st Mode Freq.	1st Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
year	Year of rental	10886	0.0	2	2012	5464	50.19	2011	5422	49.81
month	Month of rental	10886	0.0	12	Aug	912	8.38	Dec	912	8.38
day	Day of rental	10886	0.0	19	1	575	5.28	5	575	5.28
dayofweek	Day of week	10886	0.0	7	Sat	1584	14.55	Sun	1579	14.50
hour	Hour of day	10886	0.0	24	12	456	4.19	13	456	4.19
season	Current Season	10886	0.0	4	winter	2734	25.11	fall	2733	25.11
holiday	Day is holiday or not	10886	0.0	2	0	10575	97.14	1	311	2.86
workingday	Day is working day or not	10886	0.0	2	1	7412	68.09	0	3474	31.91
weather	Current weather	10886	0.0	4	clear	7192	66.07	cloudy	2834	26.03

Categorical

No missing values!

NORMALIZATION AND TRANSFORMATION

atemp	humidity	count	year_2012	hour_1	hour_2	hour_3	hour_4	hour_5
0.293127	0.81	16	0	0	0	0	0	0
0.275831	0.80	40	0	1	0	0	0	0
0.275831	0.80	32	0	0	1	0	0	0
0.293127	0.75	13	0	0	0	1	0	0
0.275831	0.80	2	0	0	0	0	0	0

New dataset size: 10155 x 34 columns

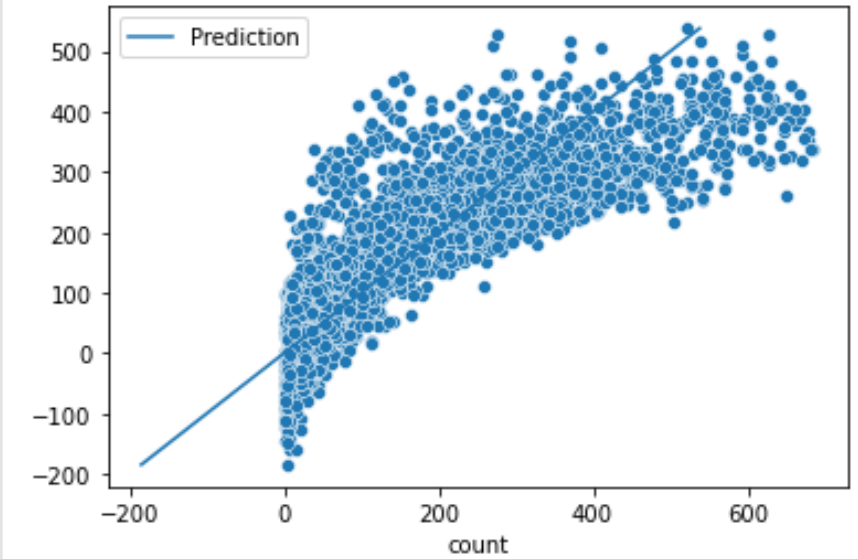
- Continuous features were normalized using range normalization, **range [0, 1]**
- Categorical features were transformed using **one-hot-encoding**
- Snippet of the dataset after normalization and transformation

MODEL SELECTION AND EVALUATION METRICS

- **Regression models:** Linear regression and Decision tree regression
- **Classification models:** Decision tree classifier and K-NN classifier
- **Evaluation metrics:** accuracy score, confusion matrix, f1-score, mean squared log error, and R squared
- Train-test split: 70% for training and 30% for testing
- Random state = 2

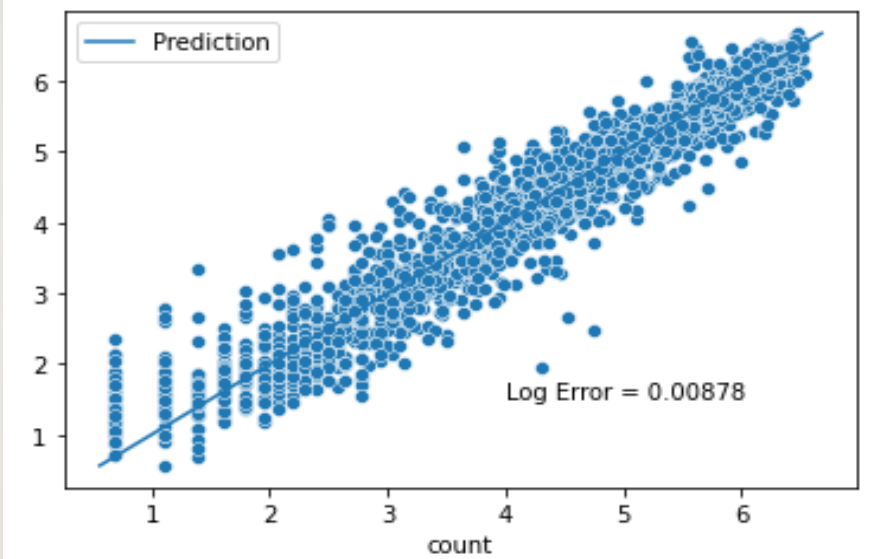
Linear Regression

```
R2 accuracy:
Train data = 0.68144249871614
Test data = 0.6682709105027007
```



Polynomial Regression

```
R2 accuracy:
Train data = 0.9435116308754387
Test data = 0.9418677877945212
```



RESULTS

- Baseline model: Linear regression predicts negative values and shows a non-linear relationship
- **Fix:** Targets features are **log transformed;** Polynomial model (with degree = 2)
- Linear regression test accuracy = 66.8%; **Polynomial regression** test accuracy = **94.2%**

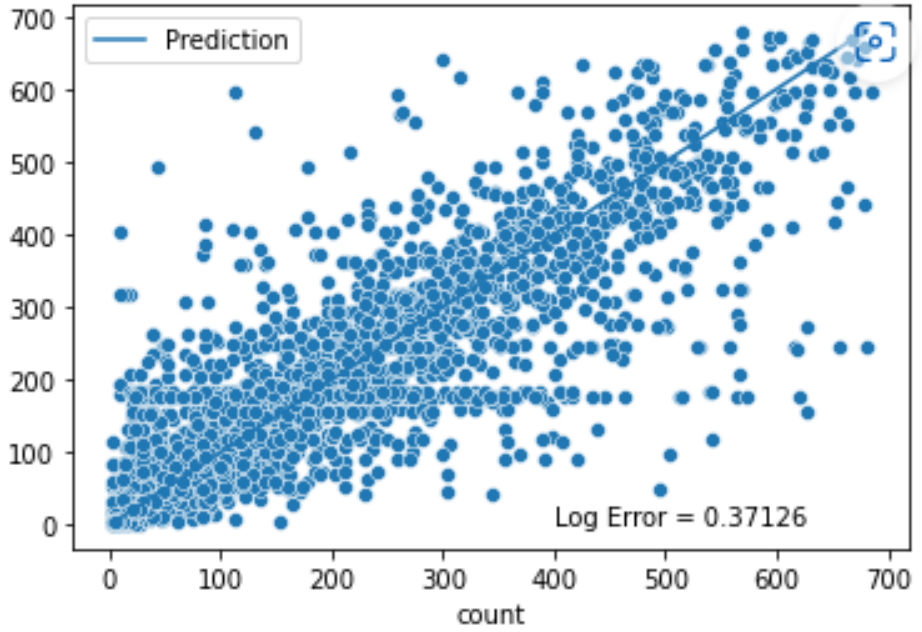
RESULTS (CONT'D)

- Hyperparameter settings:
- Criterion = 'absolute error'
- Max depth = 20
- Random state = 2

- Test accuracy = **74.29%**

Decision Tree Regression

```
R2 accuracy:
Train data = 0.8915584585105459
Test data = 0.7429224021201706
```



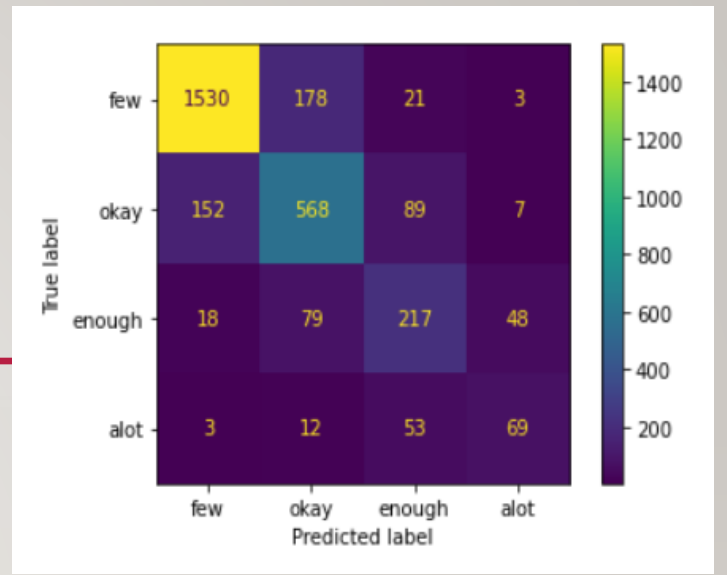
RESULTS (CONT'D)

- Decision tree: test accuracy = 78%
  - Hyperparameters:
  - Criterion = 'gini'
  - Max depth = 20
  - Random state = 2
- **K-NN**: test accuracy = **81%**
  - Hyperparameters:
  - K = 9
  - Distance metric = 'Euclidean'

Decision Tree classifier

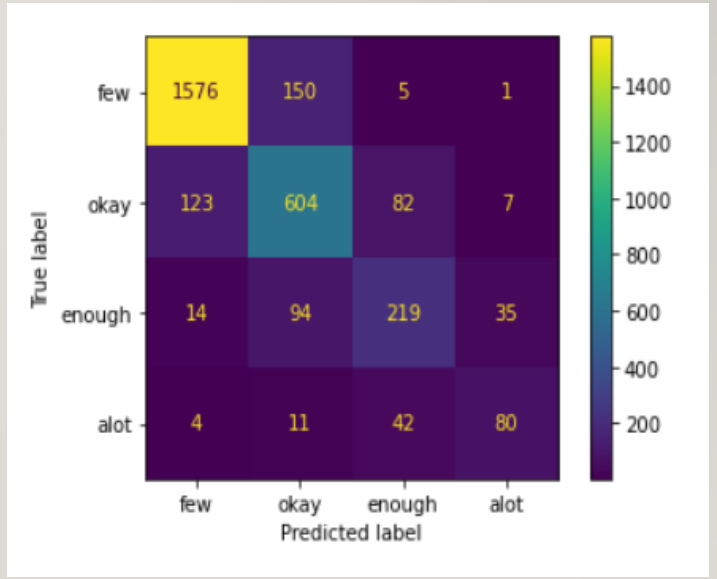
	precision	recall	f1-score
few	0.90	0.88	0.89
okay	0.68	0.70	0.69
enough	0.57	0.60	0.58
alot	0.54	0.50	0.52
accuracy			0.78

Binned targets:
Few < 171
170 < okay < 342
341 < enough < 513
512 < A lot



K-NN classifier

	precision	recall	f1-score
few	0.92	0.91	0.91
okay	0.70	0.74	0.72
enough	0.63	0.60	0.62
alot	0.65	0.58	0.62
accuracy			0.81



RESULTS (CONT'D)

- Decision tree: test accuracy = 74%
  - Hyperparameters:
  - Criterion = 'gini'
  - Max depth = 20
  - Random state = 2
- **K-NN**: test accuracy = **79%**
  - Hyperparameters:
  - K = 9
  - Distance metric = 'Euclidean'

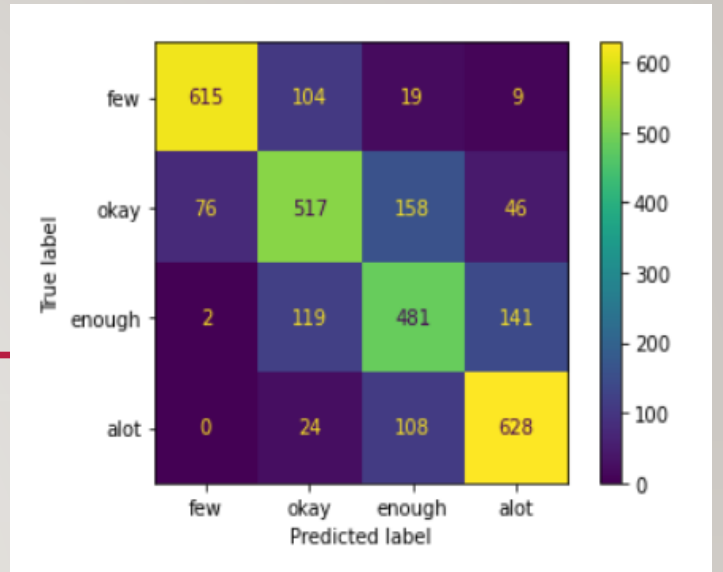
Decision Tree classifier

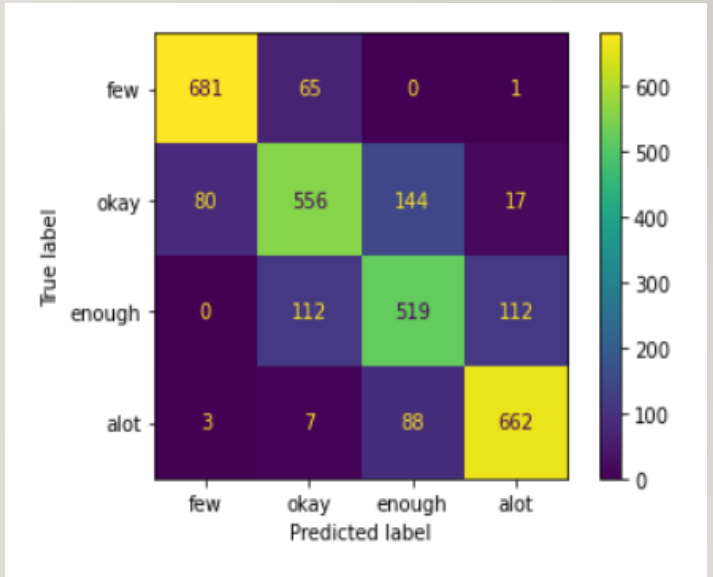
	precision	recall	f1-score
few	0.89	0.82	0.85
okay	0.68	0.65	0.66
enough	0.63	0.65	0.64
alot	0.76	0.83	0.79
accuracy			0.74

IQR targets:
Few <= Q1
Q1 < okay <= Q2
Q2 < enough <= Q3
Q3 < A lot

K-NN classifier

	precision	recall	f1-score
few	0.89	0.91	0.90
okay	0.75	0.70	0.72
enough	0.69	0.70	0.69
alot	0.84	0.87	0.85
accuracy			0.79





CONCLUSION

EDA suggests that it is best for businesses to make more bikes available for rental when the weather is clear in all seasons, especially in the Fall and Summer seasons.

The polynomial regression model had the **best accuracy (94%)** and least error in predicting bike demand (count).

The KNN model is also useful if a company needs a range of number of bikes in demand. Results produced by the IQR method has better **overall accuracy**.

Q & A

THANK YOU!