BIKE SHARING DEMAND

- DATA ANALYSIS

DSCI 4780/6780 Final Project Report

BY

Adedamola Adebamiro

Adjoa Myra Light Dagba

Jesse Annan

Prince Osei Bonsu

April 22, 2023

PROBLEM DESCRIPTION

Africa is the world's second largest and second-most-populous continent, after Asia in both aspects, with over 1.2 billion people, most of whom would need some means of transportation to go about their daily lives. Unfortunately, the continent is plagued with major challenges of transportation, largely attributed to the massive population, in addition to insufficient and sometimes unreliable transport systems (SSATP, 2015).

Problem

Living in a developed country is often taken for granted, the luxury of having good roads, affordable transportation, and quick and easy access to vital services. Imagine having to travel many miles to work, school, and the market on foot every day, instead of being able to hop in your car and go where you need to when you need to. It would be exhausting, and you would waste many hours each day just commuting. You would also miss out on countless opportunities for education, economic growth, and healthcare. Unfortunately, for some West African countries like Sierra Leone and Ghana, this is not a hypothetical scenario - it's their daily reality. Lack of transportation is one of the biggest challenges facing rural residents in these West African nations. The distances they need to travel are often long, the roads are mostly unpaved and bumpy, and public transportation is either prohibitively expensive or nonexistent.

These are complex issues that can seem impossible to solve, however, something as simple as a bike can make a huge difference. With a bike, inhabitants of these rural areas can cover greater distances faster and easily, providing much-needed access to new opportunities for education, work, and healthcare. It's a small solution that can have a big impact on improving the quality of life for those who are struggling. Growing up in Africa and experiencing some of these problems first-hand has been a major motivation for picking this project. Another motivation is the potential positive impact of this innovation on the environment and public health. Biking is a sustainable and eco-friendly mode of transportation that can help reduce carbon emissions and air pollution. By assisting biking companies in optimizing their services, more people may be encouraged to choose biking as a means of getting around, leading to improved public health and a cleaner environment.

Dataset

The dataset used in this project was sourced from Kaggle.com. Below is a list of all the features in the dataset, as well as a brief description of each feature:

Feature	***Description***
datetime	Hourly date + Timestamp
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	Whether the day is considered a holiday
workingday	Whether the day is neither a weekend nor a holiday, i.e., a working day of the week
weather	1: Clear, Few clouds, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Temperature in Celsius
atemp	'feels like' temperature in Celsius
humidity	Relative Humidity
windspeed	Wind Speed
casual	Number of Non-registered user rentals initiated
registered	Number of Registered user rentals initiated
count	Number of total rentals

The target features are *casual, registered,* and *count* (a summation of *casual* and *registered* features).

Proposed Analytics Solution

The objective of this project is to build a machine learning model that will assist bike companies predict the demand for their services. Through data analysis, this project will aid business owners in identifying when to provide bike rentals to consumers to maximize profitability. The findings of this project will also enable bike sharing systems to determine the optimal times to increase their bike inventory for rental, to meet the demands of consumers during peak seasons. By utilizing data-driven insights, bike companies can maximize their profits while providing their customers with the convenience and accessibility they need.

DATA EXPLORATION AND PREPROCESSING

The dataset, as seen above, consists of 7 descriptive features and 3 targets. To begin the data exploration process, the *datetime* column was first transformed into 5 components namely: *'year', 'month', 'day', 'dayofweek', and 'hour'*. To begin our exploratory data analysis, we first observe the linear relationship between descriptive features and our target features.

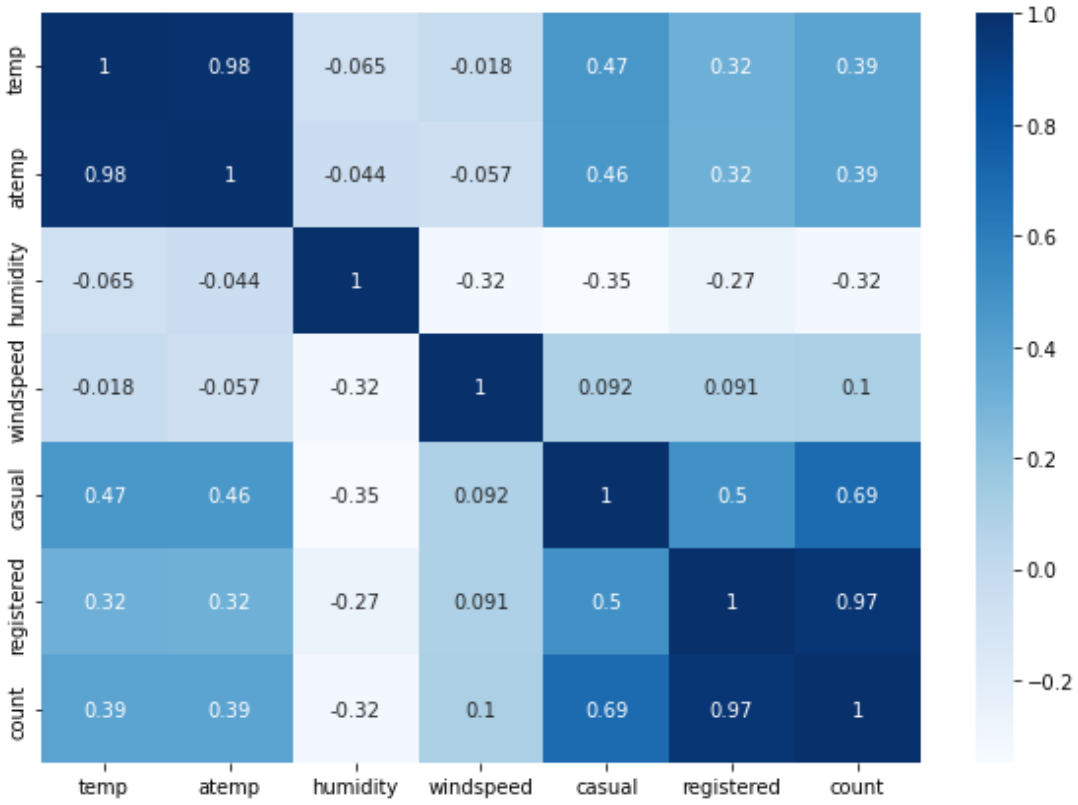


Figure 1: *A heatmap of the continuous features and the 'count' target feature.*

The correlation matrix of the continuous features and targets shows that as *humidity* increases the demand of bikes by both *casual* and *registered* members reduced, it also shows *temp and atemp's* increase, increased the demand of bikes for both *casuals* and *registered* users. However, both features are highly correlated. This makes sense because the actual temperature should not differ significantly from the 'feels like' temperature.

Observing the linear relationship between categorical features and targets. We observed a general increase in features like *year, month, hour, season,* and *weather* had a positive impact on the demand for bikes. It is also very clear to observe that *month* and *season* are highly correlated thus using both in any predictive analysis could negatively affect the predictive power of any model. Same could be said about *workingday* and *dayofweek.* In view of that, we chose to select *seasons* over the *month.*

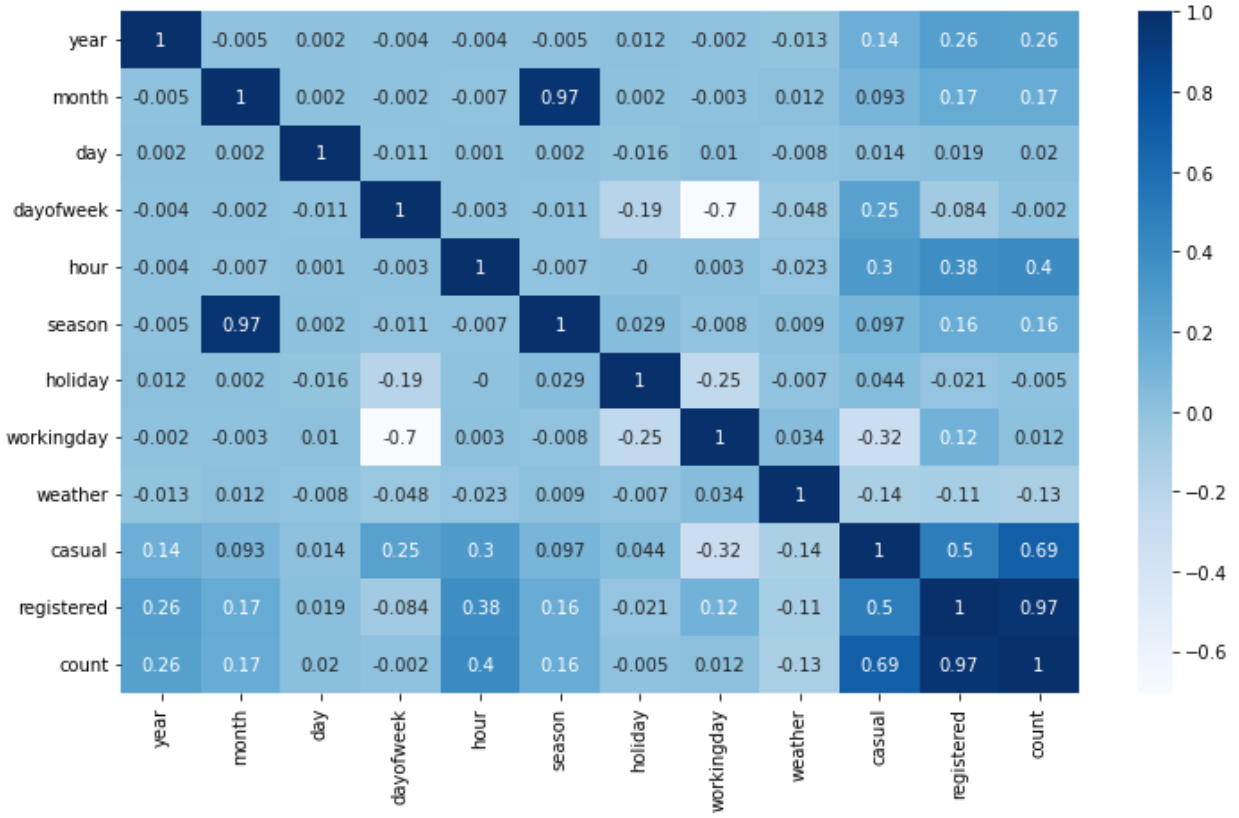


Figure 2: *A heatmap of the categorical features and the 'count' target feature.*

The *windspeed* feature although it has very weak linear relationships with the target features. We carried out some descriptive analysis. Average and maximum *windspeeds* for the *year 2011* were both greater than *year 2012*. Riding bikes at high *windspeed* is generally dangerous and therefore we deduce it could be the reason for the fewer bikes demanded in some seasons.

		windspeed				
		min	mean	std	max	size
year	season					
2011	fall	0.0	11.981981	7.126475	56.9969	1365
	spring	0.0	15.112946	9.233609	51.9987	1323
	summer	0.0	13.374759	8.218433	40.9973	1367
	winter	0.0	11.295965	8.235174	43.0006	1367
2012	fall	0.0	11.036781	7.166016	43.0006	1368
	spring	0.0	14.174978	9.048076	47.9988	1363
	summer	0.0	13.436477	7.763650	47.9988	1366
	winter	0.0	12.060328	7.412716	43.0006	1367

Figure 3: *Windspeed discriptive statistics grouped by year and seasons.*

The skewed distribution of our target features shows that more often than not, few people of groups of people ride bikes for any given day.

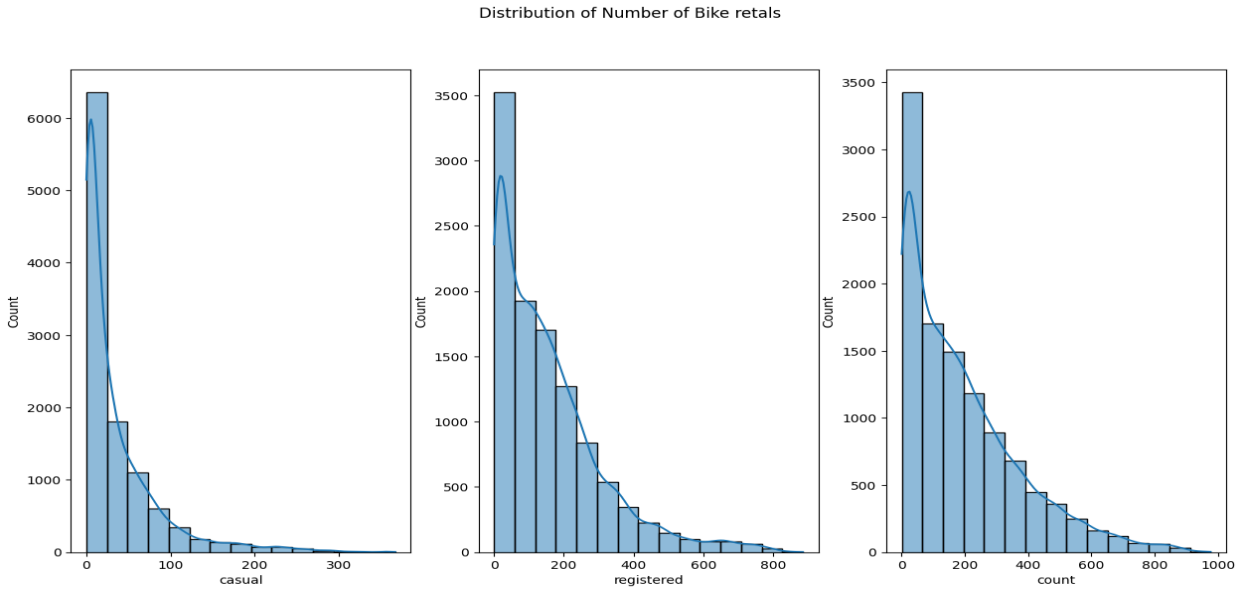


Figure 4: *Distribution of target features.*

The trends of bike demand for categorical features like *month, hour, season,* and *weather* were explored for the *casual* and *registered* bike users. In ***Figure 5****,* in 2012 we can observe a general increase in bike demand (for both *casual* and *registered* bikers) for any given *month* except July 2012. The slight decrease in July 2012, could be because of favorable humid condition during the month of July 2011. Nonetheless, the general increase also suggests that people are becoming more welcoming of the idea of using bikes as an alternative means of transportation.

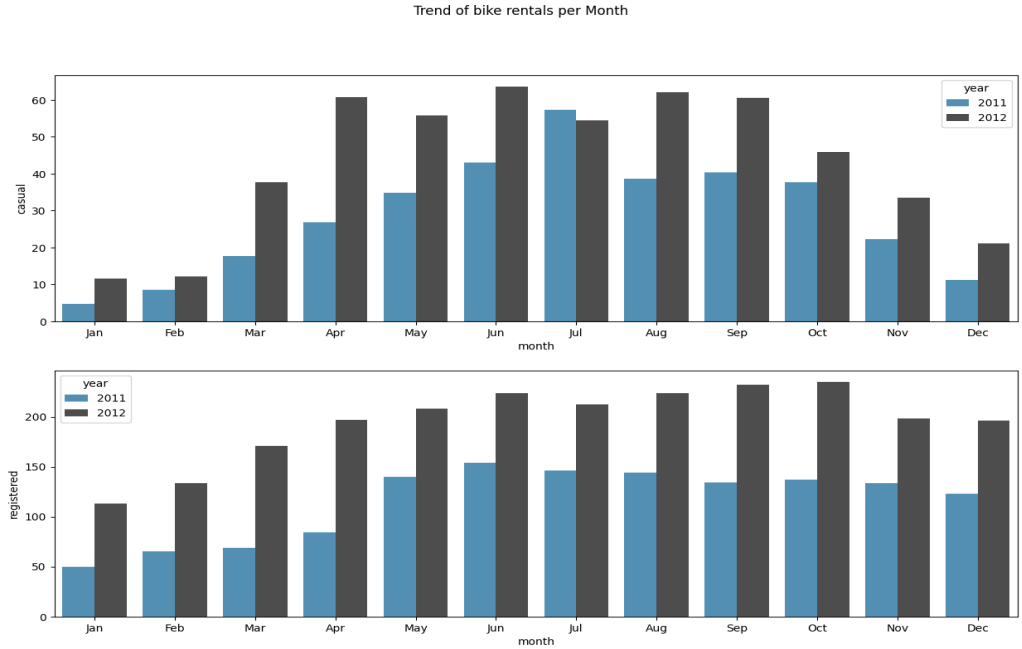


Figure 5: *Trend of Bike rentals per month for casual and registered bike users.*

Observing the trends of bikes hourly and for a given day of the week, In ***Figure 6,*** *casual* bike users tend to use the bikes the most after 8am, peak during the afternoon, and slowly reduces around 5pm. This trend, backed by the trend in ***Figure 7,*** which also shows an increase in demand for *casuals* during the same periods. However, this trend of increased demand happens during the weekends. On the contrary, *registered* bike users usually demanded the bikes (***Figure 6)*** in the morning between 6am and 9am and also in the evening between 4pm and 8pm. This trend, complemented by the fact that the demands occurred during the weekdays as shown in ***Figure 7***. This led us to believe that there is a good possibility the *registered members* may be renting the bikes as an alternative means for transportation - to get to and from work, while the *casuals* may also be renting it for fitness purposes.

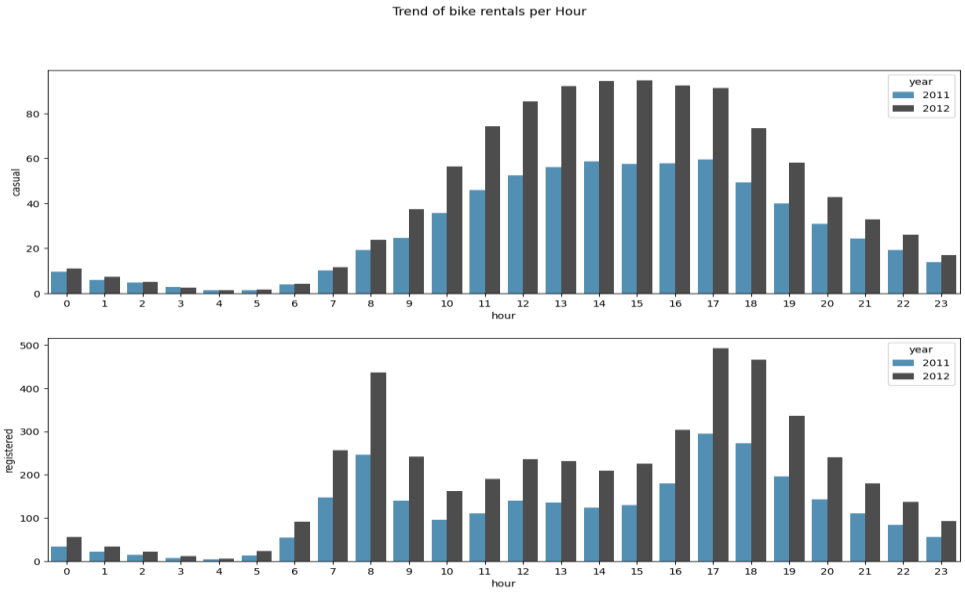


Figure 6: *Trend of bike rentals per hour for both years.*

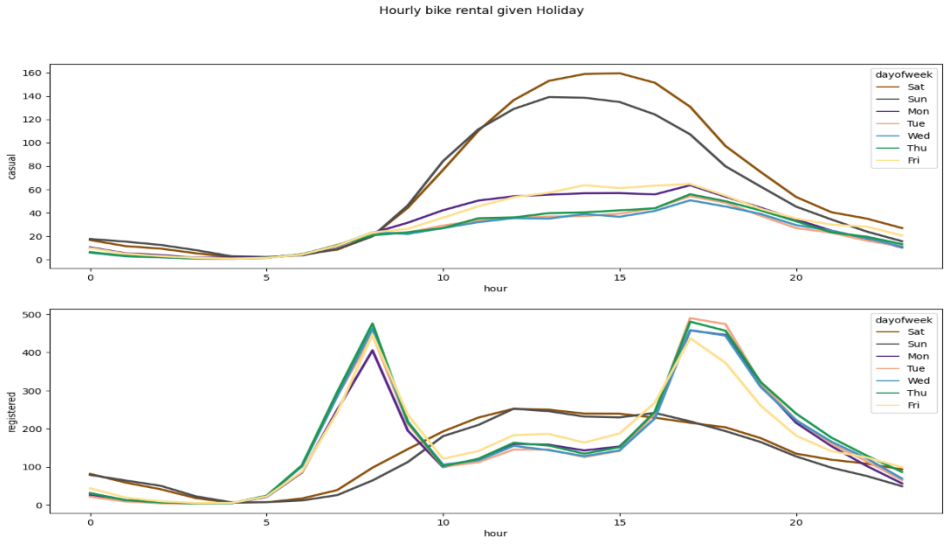


Figure 7: *Trend of bike rentals for a given day of the week.*

Figure 8 shows an increase in bike demand for fall and summer seasons and clear weather. The increase in bike demand when the weather is clear makes sense since during such weather conditions the temperatures are more favorable. In that same sense, there are low bike rentals during when it snowy and in the winter due to hash temperatures and windspeeds.

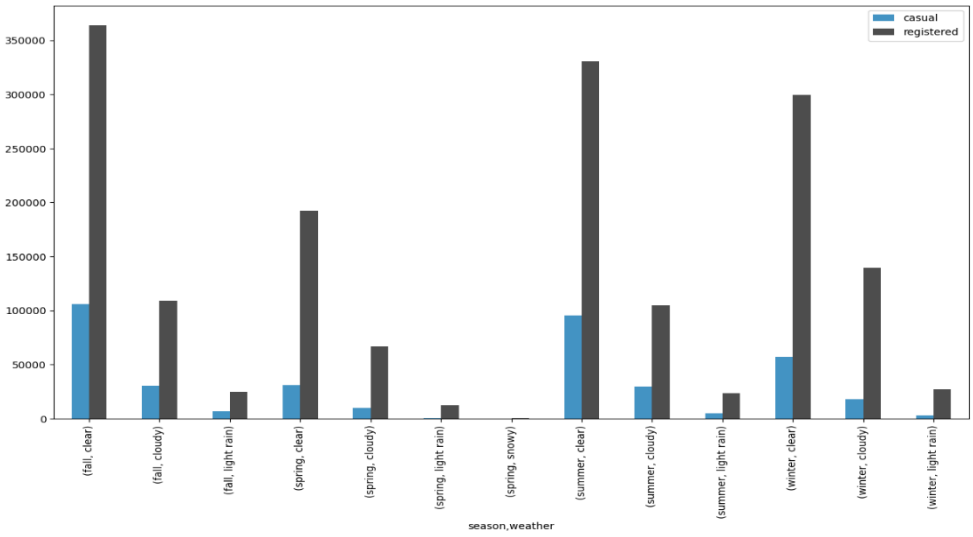


Figure 8: *Bike trend given a given season and weather.*

Data Quality Reports

The data quality reports for both the categorical and continuous features are displayed below.

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

Table 1: *Data Quality Report for Continuous Features.*

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

Table 2: *Data Quality Report for Categorical Features.*

Handling Missing Values and Outliers

As seen in the data quality report tables above, there are no missing values in the dataset. Hence, there were no missing values to handle. Observing ***Figure 9***, the box plot shows that there are potential outliers.

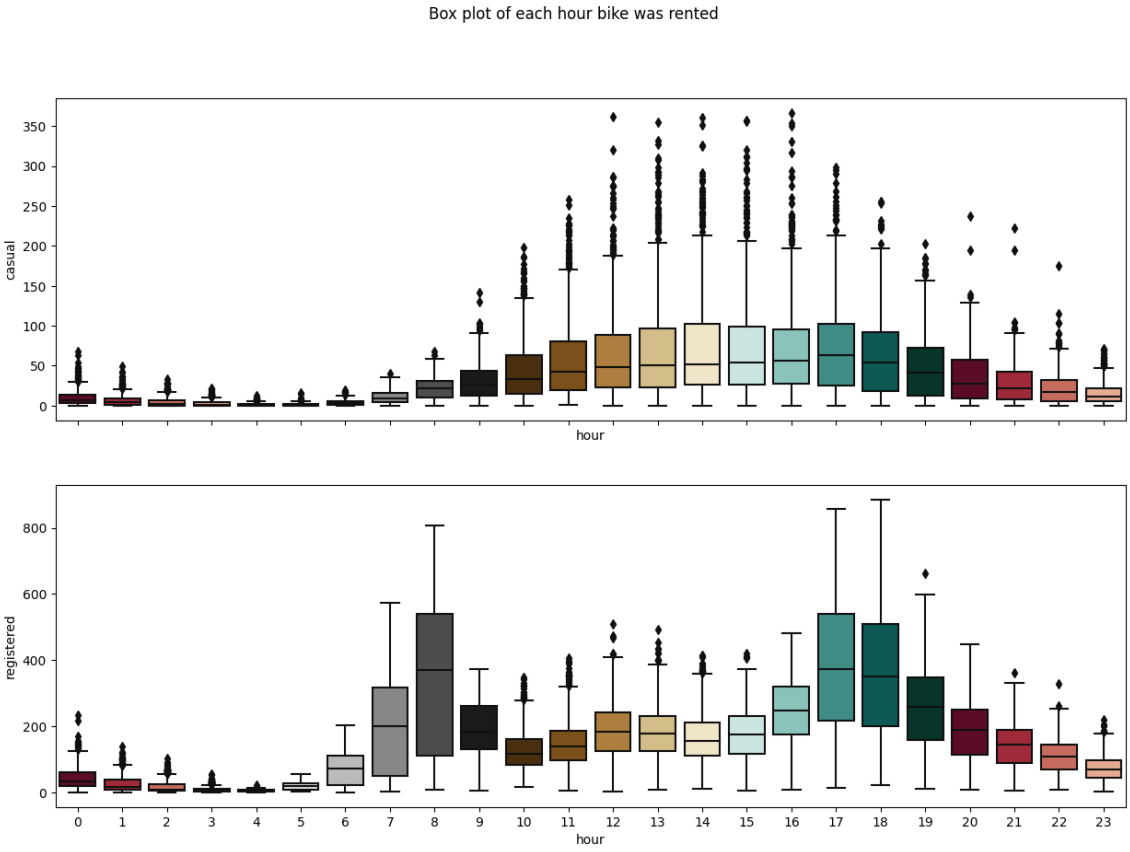


Figure 9: *Hourly variation of bike rentals.*

However, removing all these high demands of bikes regardless of the season or weather condition as an outlier would be expensive, therefore we decided to trim our dataset. The top and bottom 2% of the dataset was dropped. We also observed our *windspeed* feature, shows very dangerous values when bikes were used. It will be very difficult for even pro athletes to use bikes in *windspeeds* above *30mph* therefore we removed instances where *windspeed* were above 29mph. After these methods had been applied, the initial 10,886 instances of bike demands were reduced to 10,155 instances, implying that approximately 7% of our original dataset were considered outliers.

Normalization

The continuous features, *temp, atemp, windspeed,* and *humidity* were normalized into the range of [0,1]. This is important because huge differences in ranges of values in respective features may affect the way our machine learning algorithm(s) think a feature is worth.

Feature Selection and Transformations

	year	hour	season	workingday	weather	temp	humidity	count
0	2011	0	spring	0	clear	0.224490	0.81	16
1	2011	1	spring	0	clear	0.204082	0.80	40
2	2011	2	spring	0	clear	0.204082	0.80	32
3	2011	3	spring	0	clear	0.224490	0.75	13
6	2011	6	spring	0	clear	0.204082	0.80	2

Table 3: *Dataset before transforming categorical features.*

Since implementing the wrapper method was deemed expensive, the filter method was used for selecting relevant features for our machine learning algorithm. After the exploratory data analysis, the features deemed fit for selection from the correlation heatmap are *temp, humidity, year, hour, season, weather,* and *workingday*. Although *workingday* has a bad correlation with all our target features, we observed a general increase in model accuracy when the feature was added. Machine Learning algorithms, inspired by some mathematical concepts usually need numerical values to make sense of the data, therefore, our categorical features were transformed using the one-hot-encoding method (with the first value dropped). After this, *Season* was selected over *month* because month increased the dimensionality of our data without causing any significant increase in our model performances. Transforming regression problem to a classification problem, we took advantage of the symmetric nature of our target features (***Figure 4***) and implemented two methods. In the first method, the targets were grouped into four groups using equal width binning and in the second method, the targets were grouped based on their interquartile positions.

MODEL SELECTION AND EVALUATION

For this project we explored two problems, a regression case – predicting the number of total bike rentals (*count)* and a classification problem – predicting a range of bike demand. The first case of the classification problem used equi-width binning to create 4 groups *{few:[min, 170], okay:[171, 341], enough:[342, 512], a lot:[513, max]}.* The second classified the 4 groups based on their interquartile position.

Evaluation Metrics

The evaluation metrics used for the regression models were coefficient of determination (R^2) and root mean squared log error (RMSLE). The R^2 is used to measure the proportion of variance in the dependent variable that can be explained by the independent variable while the RMSLE measure the error produced by prediction. The RMSLE is robust to outliers, which is very useful in this case since we don't want to conclude that high values of bike demand were all necessarily outliers.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(log(x_i+1) - log(y_i+1)\big)^2} \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

The evaluation metrics used for the classification models were the accuracy scores and the f1-scroe. The accuracy score is the equivalent of R^2 for a classification problem and the f1-score is used to calculate the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by the accuracy score. It is simply the geometric means of precision and recall scores.

$$Accuracy\ \ score = \frac{TP+TN}{TP+TN+FP+FN} \qquad F1\ score = \frac{2TP}{2TP+FP+FN}$$

Where: TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative

Models

We first implemented Simple Linear Regression as a baseline model. Polynomial Regression was used to fix two issues (predicting negative values and non-linear relationship observed between the predicted and actual values) raised by the simple linear regression. The polynomial regression fixes the non-linear relationship, and we also transformed the target feature, *count* using log transform to prevent the negative predictions from occurring. Decision tree regression was also implemented to compare the predictive power between other linear models. For the classification problem, we implemented Decision tree classification and KNN algorithms for both groups.

Sampling and Evaluation Settings

The best evaluation scores for the models selected above occurred when train-test-split were 70% for the train data set and 30% for the test data set.

Hyper-Parameter Optimization

For both classification problems setting the different hyper-parameters affected the accuracy of the predictive model on both the train and test data sets.

Equi-width binning method:

Decision Tree Class.	**Max depth**		
Accuracy	10	15	20
Test	.7210	.7388	.7824
Trian	.7641	.8547	.9395
k-Nearest Neigh.	**k**		
Accuracy	3	9	15
Test	.8188	.8136	.8034
Trian	.9048	.8530	.8355

IQR binning method:

Decision Tree Class.	**Max depth**		
Accuracy	10	15	20
Test	.6058	.6974	.7355
Trian	.6216	.7724	.8848
k-Nearest Neigh.	**k**		
Accuracy	3	9	15
Test	.7831	.7936	.7795
Trian	.8820	.8277	.8104

RESULTS AND CONCLUSION

Regression Models: Linear Regression, Polynomial Regression, Decision Tree Regression

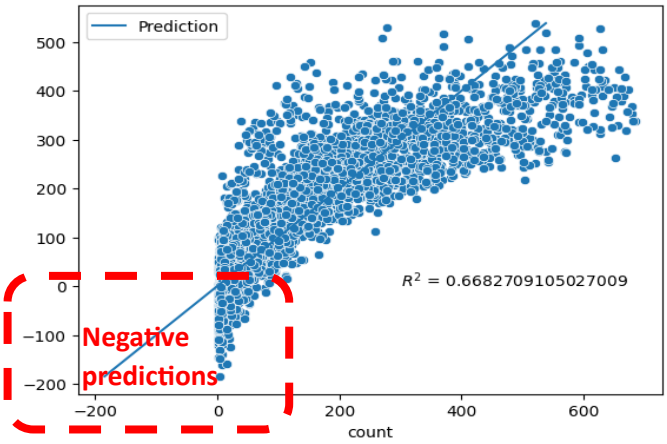


Figure 10: *Simple Linear Regression Model.*

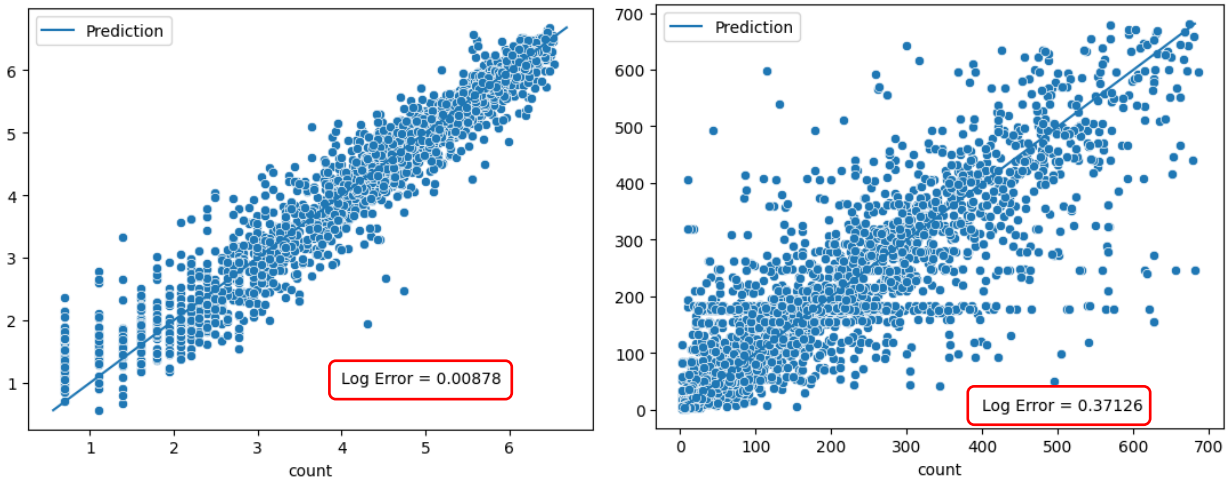


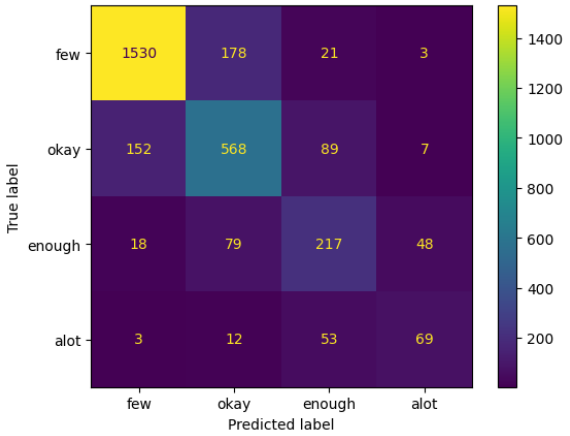
Figure 11: *Polynomial Regression Model (left) and Decision Tree Regression Model (right)*

Evaluation		Linear Reg.	Polynomial Reg.	D. Tree Reg
R2	Train	.68144	.94351	.89156
	Test	.66827	.94190	.74292
RMSLE		-	.00878	.37126

Classification Models: Decision Tree Regression and K-Nearest Neighborhood

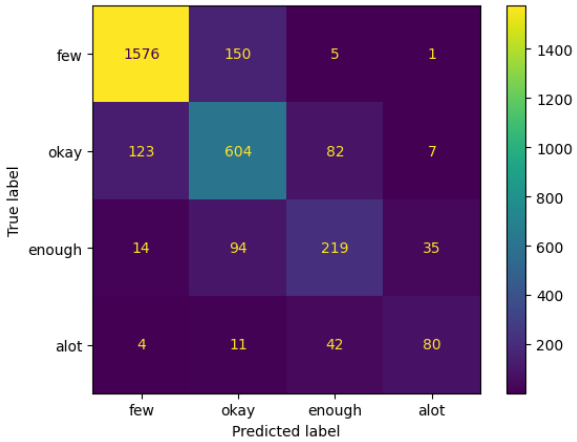
1. Using Equi-width binning:

[few ≤ 170.75 ; 170.75 < okay ≤ 341.5 ; 341.5 < enough ≤ 512.25 ; 512.25 < a lot]



	precision	recall	f1-score	support
few	0.90	0.88	0.89	1732
okay	0.68	0.70	0.69	816
enough	0.57	0.60	0.58	362
alot	0.54	0.50	0.52	137
accuracy			0.78	3047
macro avg	0.67	0.67	0.67	3047
weighted avg	0.78	0.78	0.78	3047

Figure 12: *Decision Tree with max depth = 20*

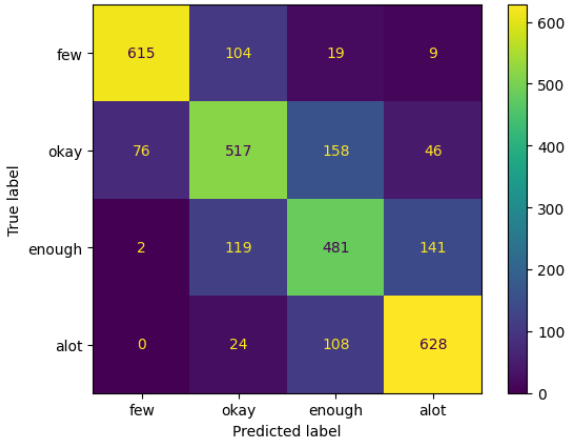


	precision	recall	f1-score	support
few	0.92	0.91	0.91	1732
okay	0.70	0.74	0.72	816
enough	0.63	0.60	0.62	362
alot	0.65	0.58	0.62	137
accuracy			0.81	3047
macro avg	0.73	0.71	0.72	3047
weighted avg	0.81	0.81	0.81	3047

Figure 13: *k-NN with k = 9*

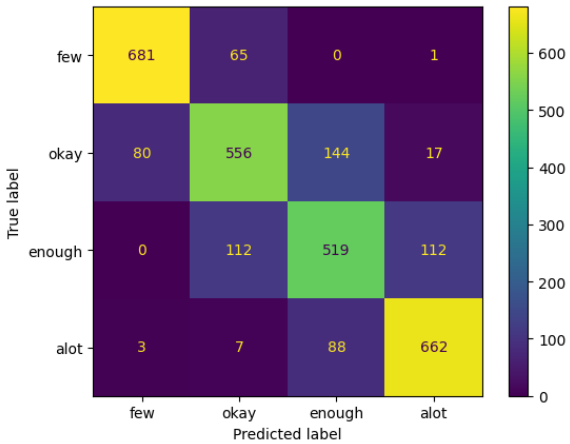
2. Using IQR binning method:

[few ≤ Q1 ; Q1 < okay ≤ Q2 ; Q2 < enough ≤ Q3 ; Q3 < a lot]



	precision	recall	f1-score	support
few	0.89	0.82	0.85	747
okay	0.68	0.65	0.66	797
enough	0.63	0.65	0.64	743
alot	0.76	0.83	0.79	760
accuracy			0.74	3047
macro avg	0.74	0.74	0.74	3047
weighted avg	0.74	0.74	0.74	3047

Figure 14: *Decision Tree with max depth = 20*



	precision	recall	f1-score	support
few	0.89	0.91	0.90	747
okay	0.75	0.70	0.72	797
enough	0.69	0.70	0.69	743
alot	0.84	0.87	0.85	760
accuracy			0.79	3047
macro avg	0.79	0.79	0.79	3047
weighted avg	0.79	0.79	0.79	3047

Figure 15: *k-NN with k = 9*

From the evaluation metrics presented above we conclude that the polynomial regression is the best regression model to use. It has the best accuracy among the regression models with R^2 of 94% and also the least log error of .00878. ***Figure 8*** suggests that bike demands are most in the fall and summer seasons therefore, we'd recommend that bike sharing companies increase the number of bikes available for rentals.

Moreover, the k-NN models show best accuracies for predicting ranges of bikes with both 81% and 79%, which are better than the predictions made by the decision trees, 78% and 74%. However, the best overall (classification) model is the model produced by the IQR binning method. We select this model with an overall accuracy lesser than the equi-width binning method because of the more

superior f1-score produced by the IQR binning method and the values of the macro avg and weighted avg are very different for the equi-width method. This difference is because of class imbalance (number of *few* are far greater than number of *a lot*) which is very clear looking at the confusion matrix in ***Figure 13,*** but class imbalance is not an issue for IQR method (***Figure 15).***

Furthermore, our models did not account for the effect of *windspeed* and *weather.* We believe both features affects the demand of bike greatly and suggest future works explore models which are able to incorporate the information hidden in both features.

CREDITS AND REFERENCES

1. **https://www.kaggle.com/competitions/bike-sharing-demand/code**
2. **https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/**
3. **https://www.youtube.com/watch?v=XiUlqN1AyoU**
4. **Storytelling with data – Cole Nussbaumer Knaflic**
5. **Fundamentals of Machine Learning for Predictive Data Analytics – John D. Kelleher, Brian Mac Namee, Aoife D'Arcy**