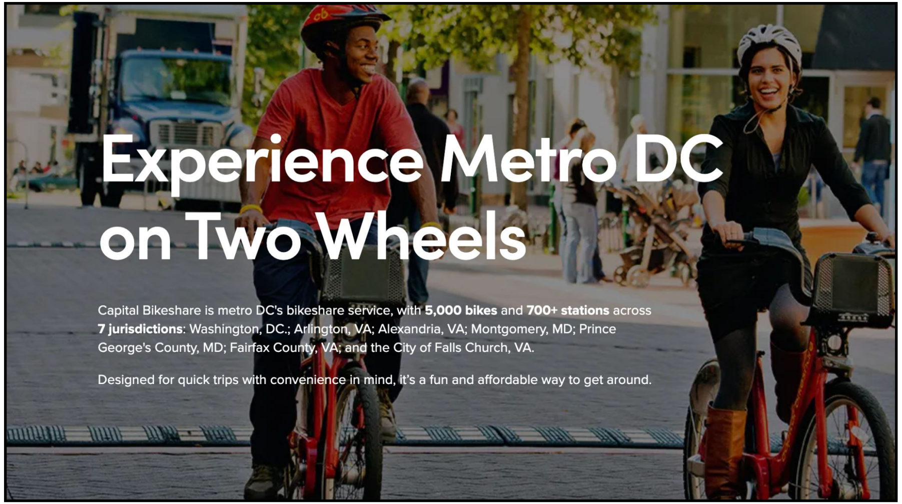
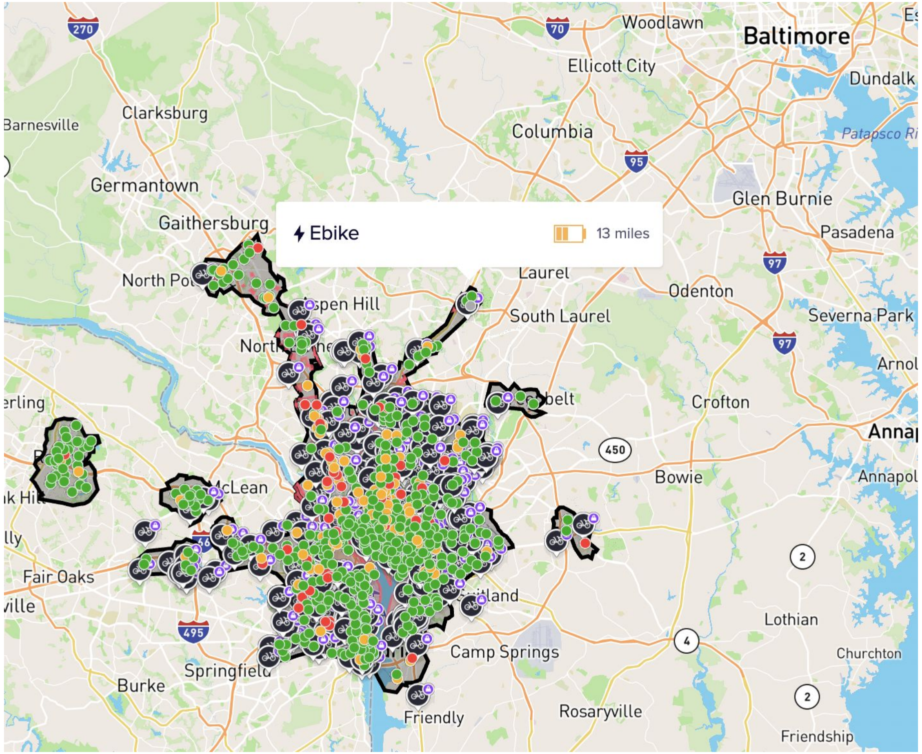
Cycles of Demand: Predicting Usage of Washington DC's E-Bike Rentals

Brian Miller

Background:

- Capital Bikeshare is a DC based company that provides rentable Ebikes
- Their Ebike usage data is public
- Additionally weather data for the area is publicly available





Research Question:

Can we predict the number of riders per hour given the time and weather conditions?

- With a better understanding of bike usage times and locations, the bike network can be optimized to increase user activity and user satisfaction.
- With this information new bike storage locations and inventory quantities can be adjusted to help optimize the network.

Project Takeaways:

- We were able to build a model that explains 88% of the variance in E-Bike usage, a promising result (Random Forest Regressor).

- Temperature, humidity, and if it was rush hour were the three most important variables in predicting E-Bike usage.

- <u>Given these results, some potential changes to improve business:</u>
  - Increase E-Bike redistribution before rush hour to capitalize on commuter usage.
  - Add E-Bike charging stations near dense employment locations.
  - Offer discounts or incentives during no-optimal weather conditions.
  - If E-Bike rollbacks or updates need to be done, perform them during non-peak months.

Variable to Predict: Ebike Rides per Hour

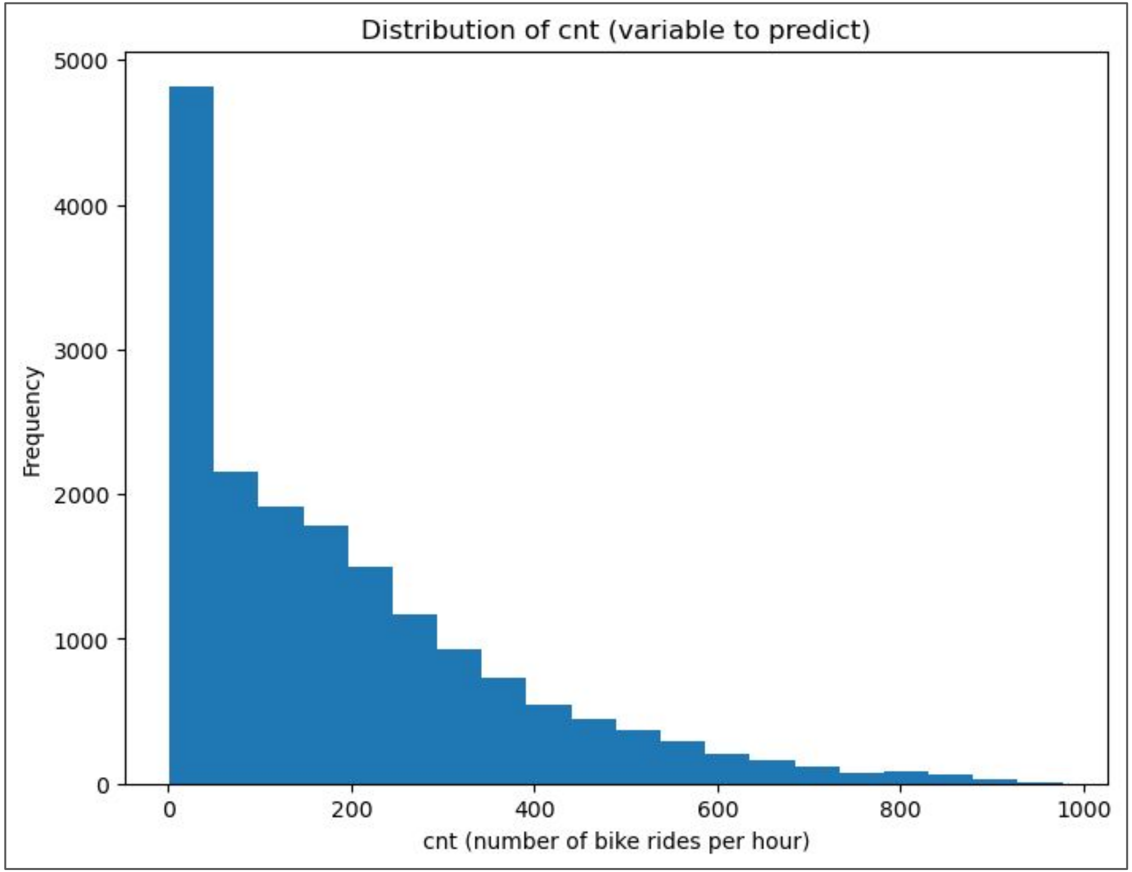
Explanatory Variables:

Numerical Variables:

- temperature
- humidity
- wind speed

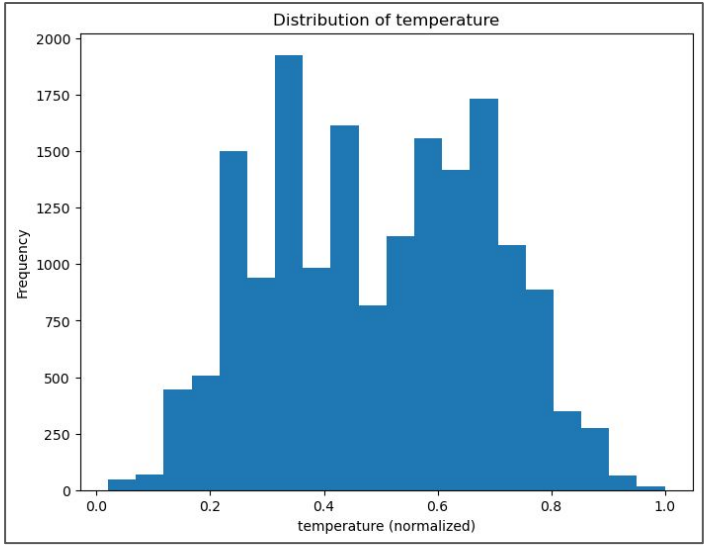
Categorical Variables:

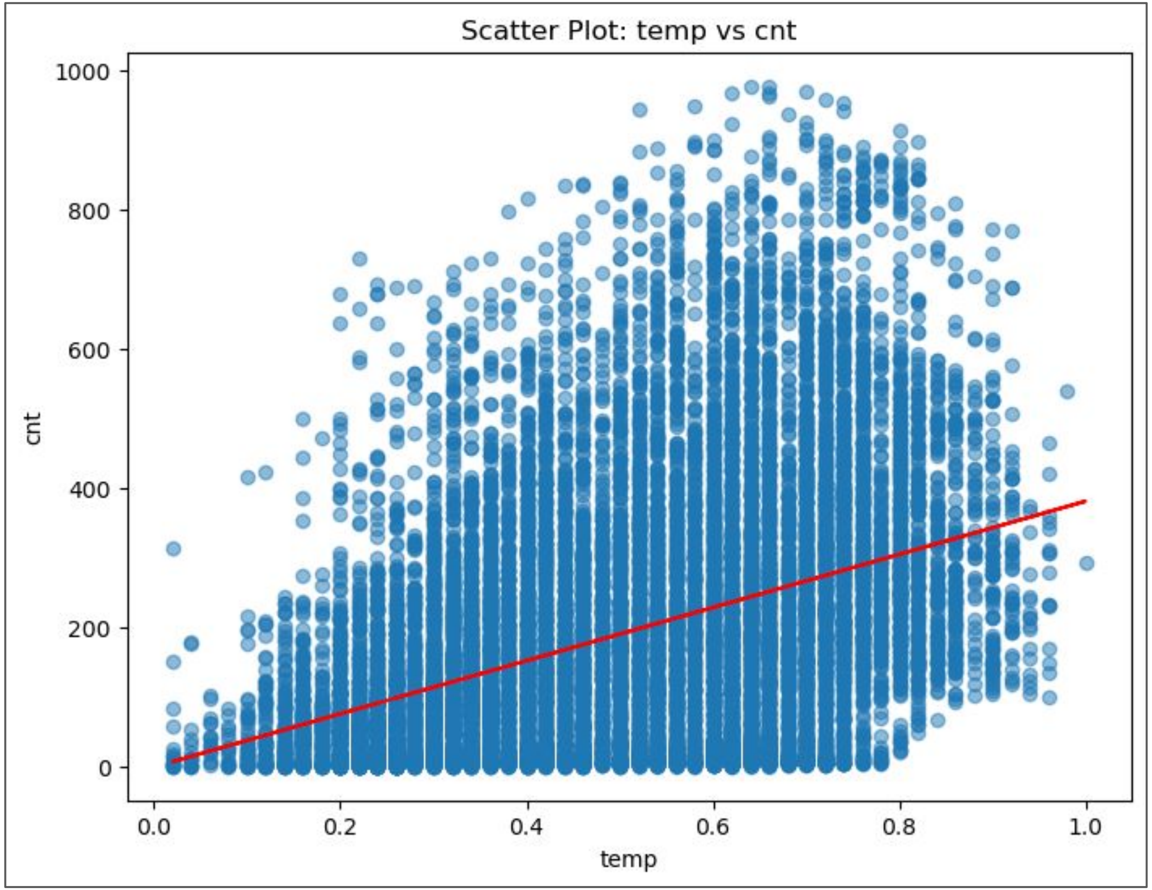
- weather score
- weekend or not
- season of year



More Riders in Warmer Temperature:

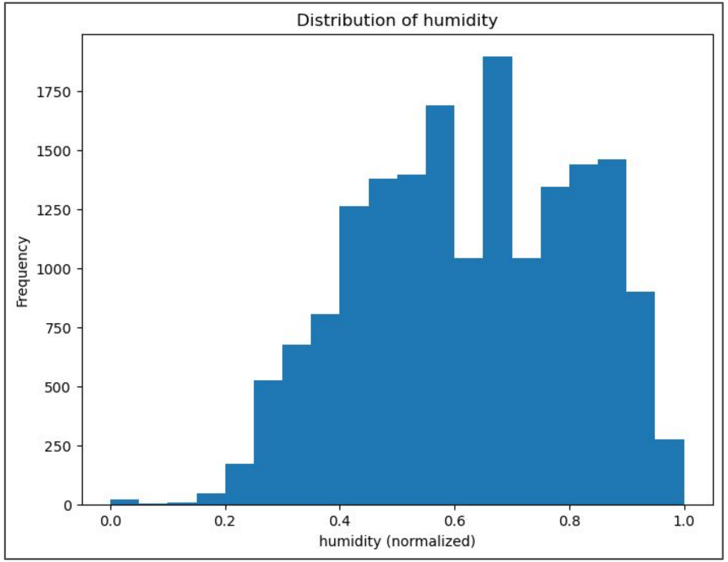
- Temperature normalized by max (105°F)
- Decent positive correlation ($R^2 = 0.16$)

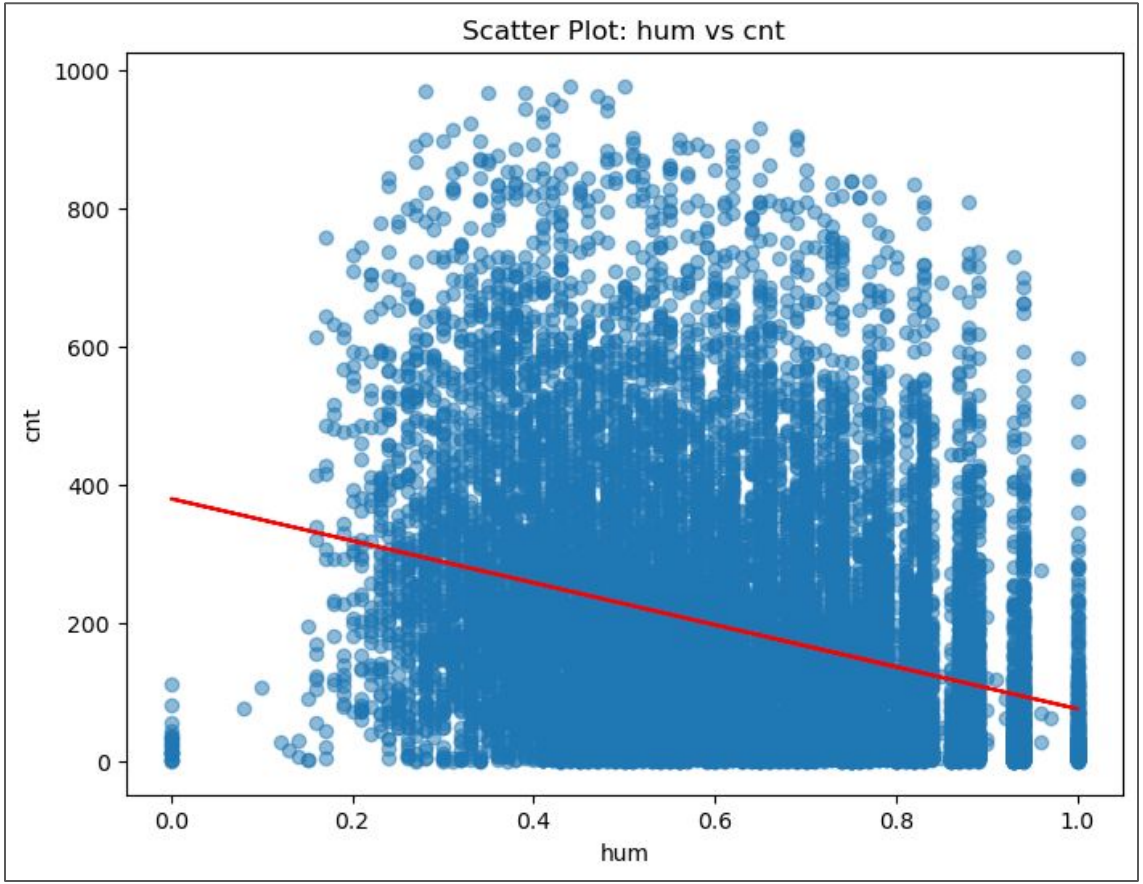




Less Riders in Higher Humidity:

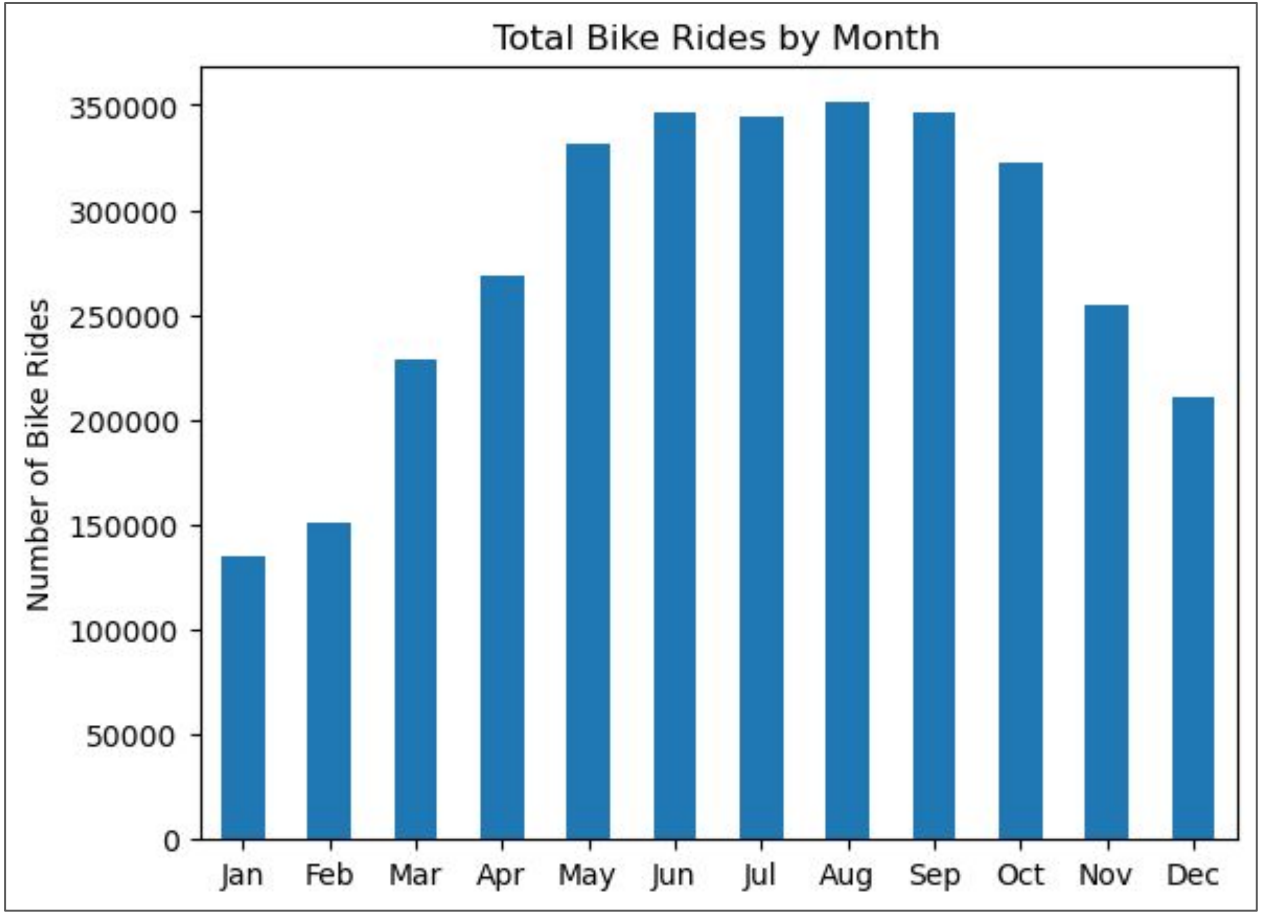
- Humidity normalized by max
- Decent negative correlation ($R^2 = 0.10$)





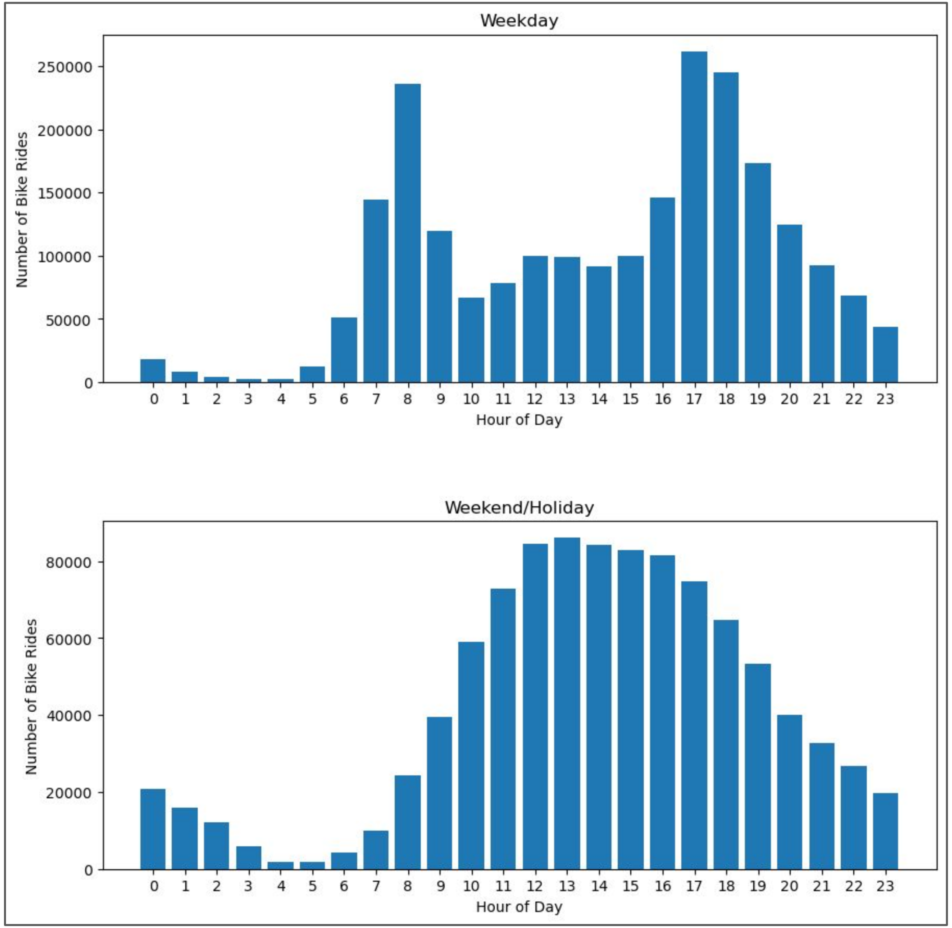
More Rides in the Warmer Months:

- May - Oct are the most popular months
- Peak usage is in the summer months when it is nice out



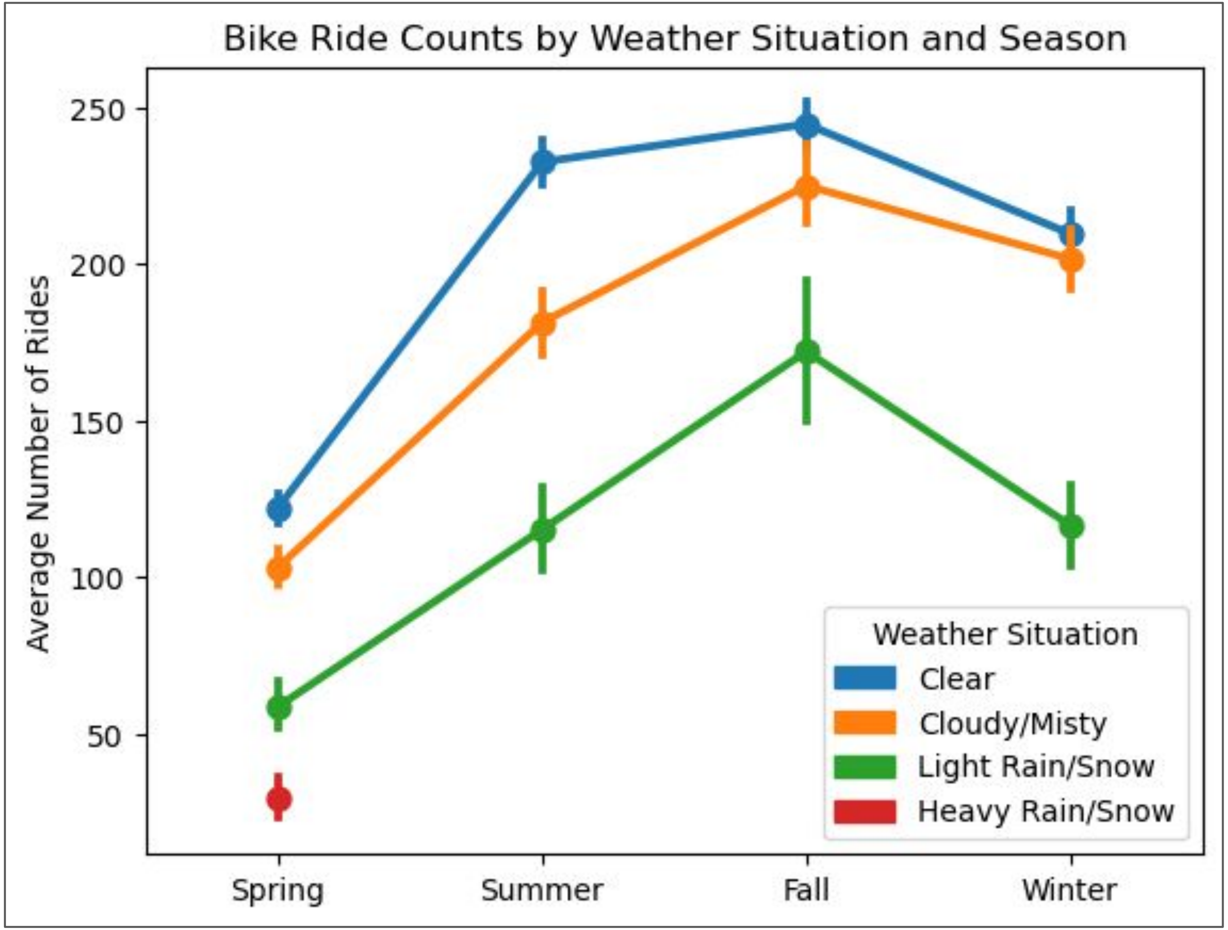
More Rides During Rush Hour:

- 8am and 5pm are peak Ebike usage hours, most likely for commuters
- Usage in general tends to follow normal daytime hours



More Rides With Better Weather Conditions:

- Clear skies during the Summer or Fall had the most rides
- Again, only a small number of rides were completed with terrible weather



Data Modeling Summary:

- Many different model types were attempted
- Random Forest Regressor had the best results
- Tuning the RF slightly improved the models performance

Model	RMSE (cv=6)	R^2 (cv=6)
Linear Regression	0.56	0.69
SGD Regression	0.57	0.68
Ridge Regression	0.56	0.68
Decision Tree Regression	0.40	0.83
Bagging Regression	0.31	0.90
Random Forest Regression	0.294	0.915
Random Forest Regression (tuned)	0.289	0.917
Random Forest Regression (tuned)	0.333 (**Test**)	0.885 (**Test**)

Data Modeling:

- Training R^2 = 0.917
- Testing R^2 = 0.885
- All R^2 and RMSE values use 6-fold cross validation

```
rf_tuned_mse_scores = -cross_val_score(random_search.best_estimator_, x_train, y_train, cv=6, scoring='neg_mean_squared_error')
rf_tuned_r2_scores = cross_val_score(random_search.best_estimator_, x_train, y_train, cv=6, scoring='r2')

print("RandomForestRegressor (tuned with random search)")
print("Train Xval R^2  ", np.mean(rf_tuned_r2_scores))
print("Train Xval RMSE ", np.sqrt(np.mean(rf_tuned_mse_scores)))
```

✓ 17m 55.4s

```
RandomForestRegressor (tuned with random search)
Xval R^2   0.9170254090835693
Xval RMSE  0.28978735730887784
```

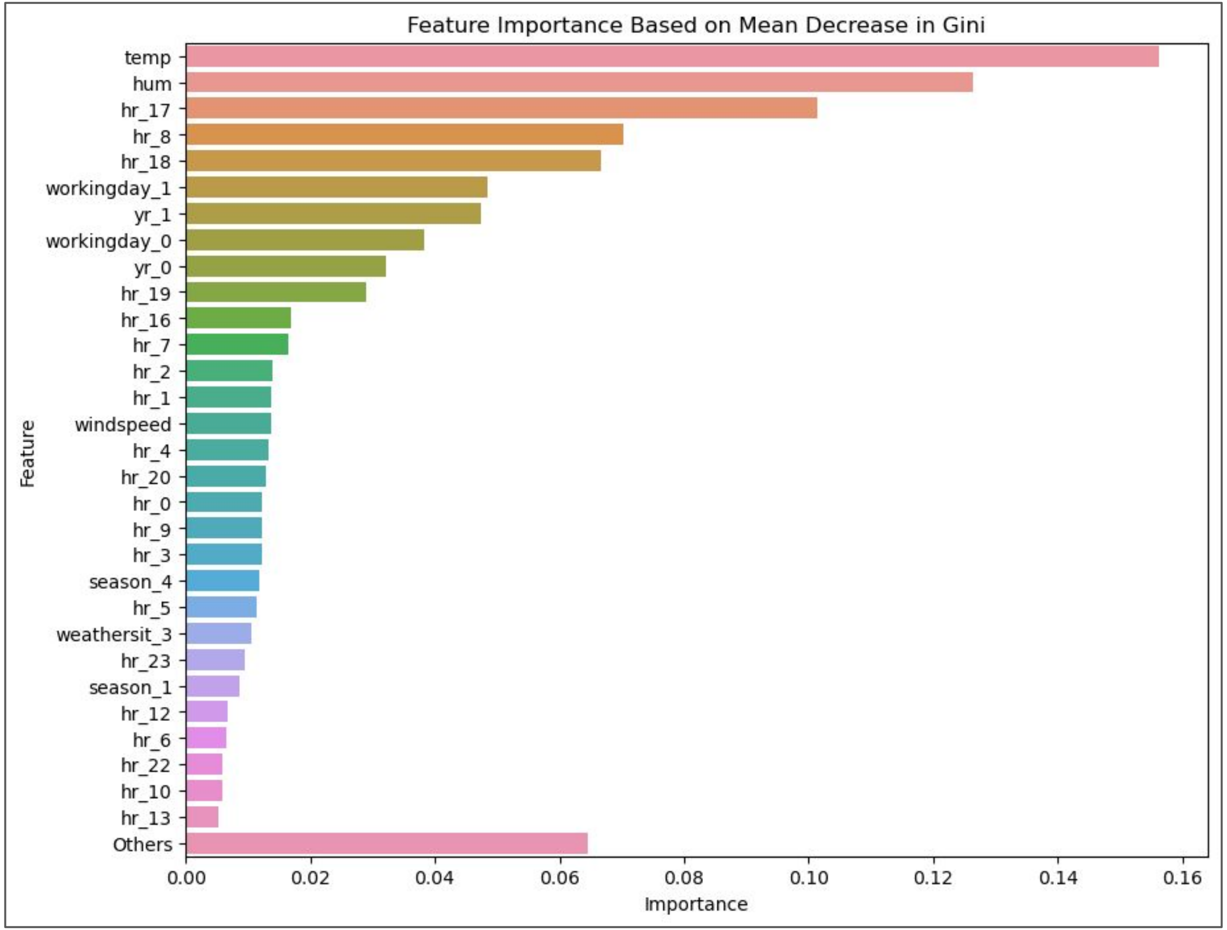
```
rf_tuned_mse_scores = -cross_val_score(random_search.best_estimator_, x_test, y_test, cv=6, scoring='neg_mean_squared_error')
rf_tuned_r2_scores = cross_val_score(random_search.best_estimator_, x_test, y_test, cv=6, scoring='r2')

print("RandomForestRegressor (tuned with random search)")
print("Test Xval R^2  ", np.mean(rf_tuned_r2_scores))
print("Test Xval RMSE ", np.sqrt(np.mean(rf_tuned_mse_scores)))
```

✓ 8m 34.9s

```
RandomForestRegressor (tuned with random search)
Test Xval R^2    0.8853303580216282
Test Xval RMSE   0.33349899770253133
```

Feature Importance:



Conclusion:

- Given the weather conditions and time of day, using a Random Forest Regression yielded the best results.

- We obtained an R^2 of 0.885 on the test set. Thus our predictor variables explain a large portion of the variability in the number of Ebike rides per hour.

- Temperature, humidity, and if it was rush hour were the three most important variables.