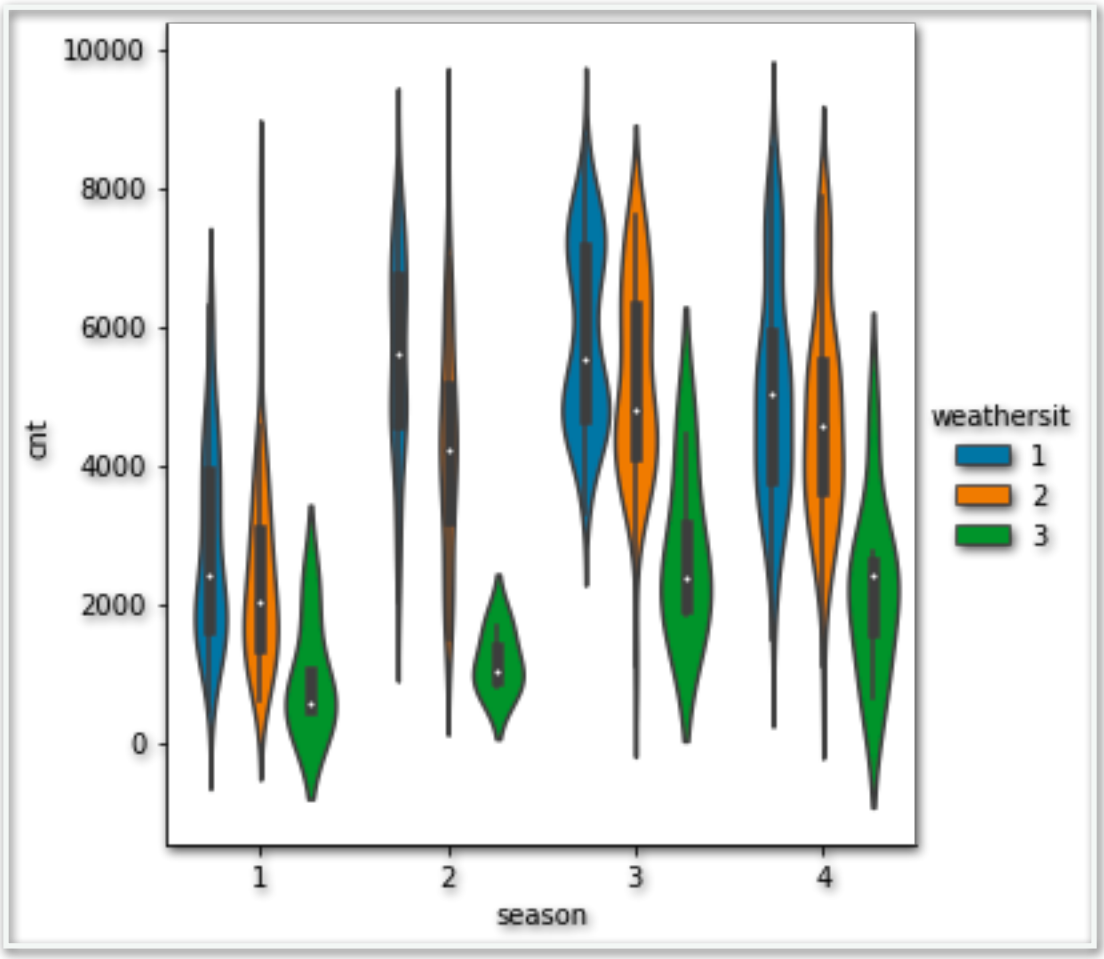
Capstone_2_Final_Report

The Problem was about Capital Bike Share System - aDC-based Bike Sharing Company, which launches a new branch in the 8th jurisdiction, Philadelphia next year (2024). I needed to predict the average bike rental count of its Philadelphia based customers for next year so that they can guarantee their 10 percent of profit like in other jurisdictions based on weather predictions and population data of Philadelphia?

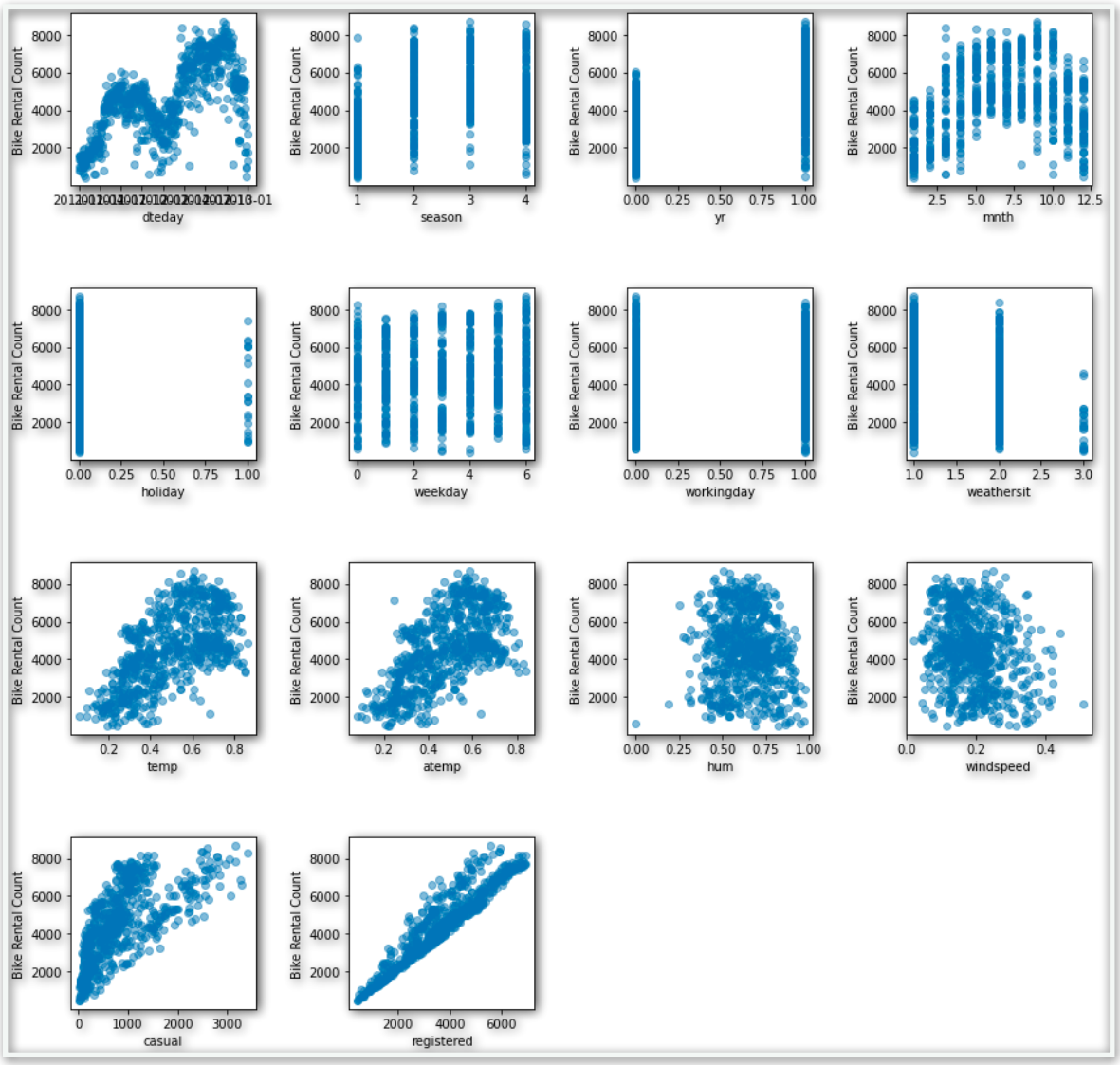
For this task, I did a quick **data wrangling**, there was no missing or duplicate values. I had daily and hourly data for two years, 2011 and 2012. After a quick look I decided to go forward with the day data, since it would be easy to forecast daily weather for Philadelphia.

I had a quick look to seasons, temperature, months, and precipitation to get a sense of important features.

In the **exploratory data analysis** the analysis of features continued using heat map, and later cat plotting and count-plotting bike rentals per each season, weather situation, humidity, temperature etc. I also used scatterplot to understand why low humidity and windspeed would also result in the decrease of bike rentals. (In all cases the temperature was very low and cold was the main reason.)



I also created a scatterplot function to plot each feature's relation to the bike rental count as subplots.



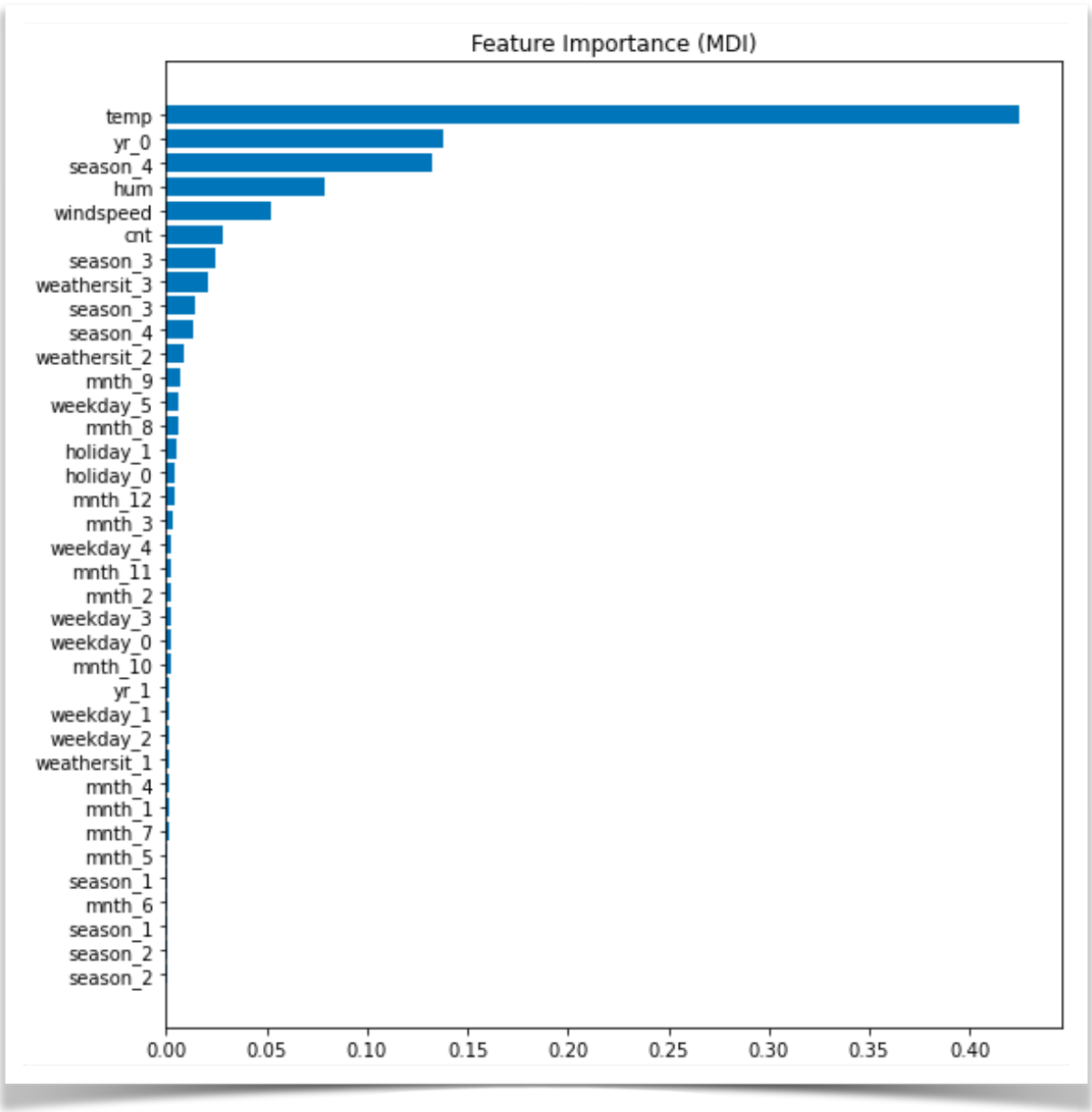
After these exploration I wanted to stick only to the data of year 2011(the first year of Capital Bike as a business in DC) because that would include more accurate data for the Philadelphia branch's 1st year forecast. However, this choice made the accuracy of all the models I tried very low so I went back to have both datasets to come up with a more accurate model. So I had two attempts to measure models. I used mean deviation and R squared and .score function to measure accuracy.

When I tested **modeling after preprocessing**, both Random Forest Regressor and Gradient Boosting Regressor performed well using the dataset from both years. I tuned both of their hyperparameters and at the end I decided to use Gradient Boosting Regressor with hyperparameters that tuned. The hyperparameters are below:

```
learning_rate': 0.01,
'max_depth': 4,
'n_estimators': 1000,
'random_state': 1,
'subsample': 0.5
```

I used GridSearchCV to find these parameters. Before using the model, I made sure to get dummies for the category features like weekday, holiday, season, weathers etc. I also used Standard Scaler to scale the data.

After finalizing the Gradient Boosting Regressor model, it was obvious that year, temperature, season, humidity, windspeed, weather situation were the most important features.



Finally I created a function that would predict the bike count for each day (each row) of Philadelphia using the population ration of Philadelphia to DC.

Because I was not provided with a dataset of annual weather of Philadelphia, I just used the weather-related values of Philadelphia on January 1st, 2023 as a sample to try the function. And if Capital Bike was launched on January 1st 2023 in Philadelphia, there would be approximately 2968 bike rental in total. When I created a data frame with January 1st weather details (after undoing StandardScaler), I also marked 'yr' as 0 since this would be the first year of Philadelphia as well.

The function to predict Philly numbers is below (worded here with more simplified vocabulary) :

```
def predict_philly (data):
    y_pred_philly = pd.DataFrame(model.predict(data) * population_ratio)
     return round ((y_pred_philly * bike_std) + bike_mean)
```

FUTURE IMPROVEMENTS

If I was provided with a dataset for Philadelphia I would use it to forecast the next year weather with ARIMA model and later applied the function on it. However, that was not the case.

Besides the lack of dataset for Philadelphia weather, there are certain variables specific to the case of Philadelphia or year 2024 and they might also influence the accuracy of our model. For instance, DC and Philadelphia attract different number of tourists who might be also a big percentage of bike users. There might be local events that can impact the number of bikes rented. The city landscape is also factor that I can't measure. While DC is generally flat, the northwestern area of Philadelphia is not. Also the inconvenience of public transportation can be also a factor for the increase of bike rentals and we don't have data to compare that as well.

Overall, because our task is marketing oriented, an accuracy of 91 percent is a sufficient number to take actions to guarantee that 10 percent profit. If the bike counts are not enough to assure 10 percent, I would advise Tham to install bike stations near college campuses and business centers to attract registered users who need to use bikes even when the season or the weather is not great.