Capstone Project

On

Bike sharing Dataset

Nastaran Mahmoodzadeh

Summer 2020

Data Collection

- From where can the data be obtained?

  The Bike sharing dataset has been collected from UCI website. The data is based on two separate variable: Day & Hour

  We focused on Day and hour dataset for this particular analysis.

- How must the data be cleansed and validated?

  We used pandas library in Jupyter Notebook to clean the dataset.

Business Question

Main Business Question

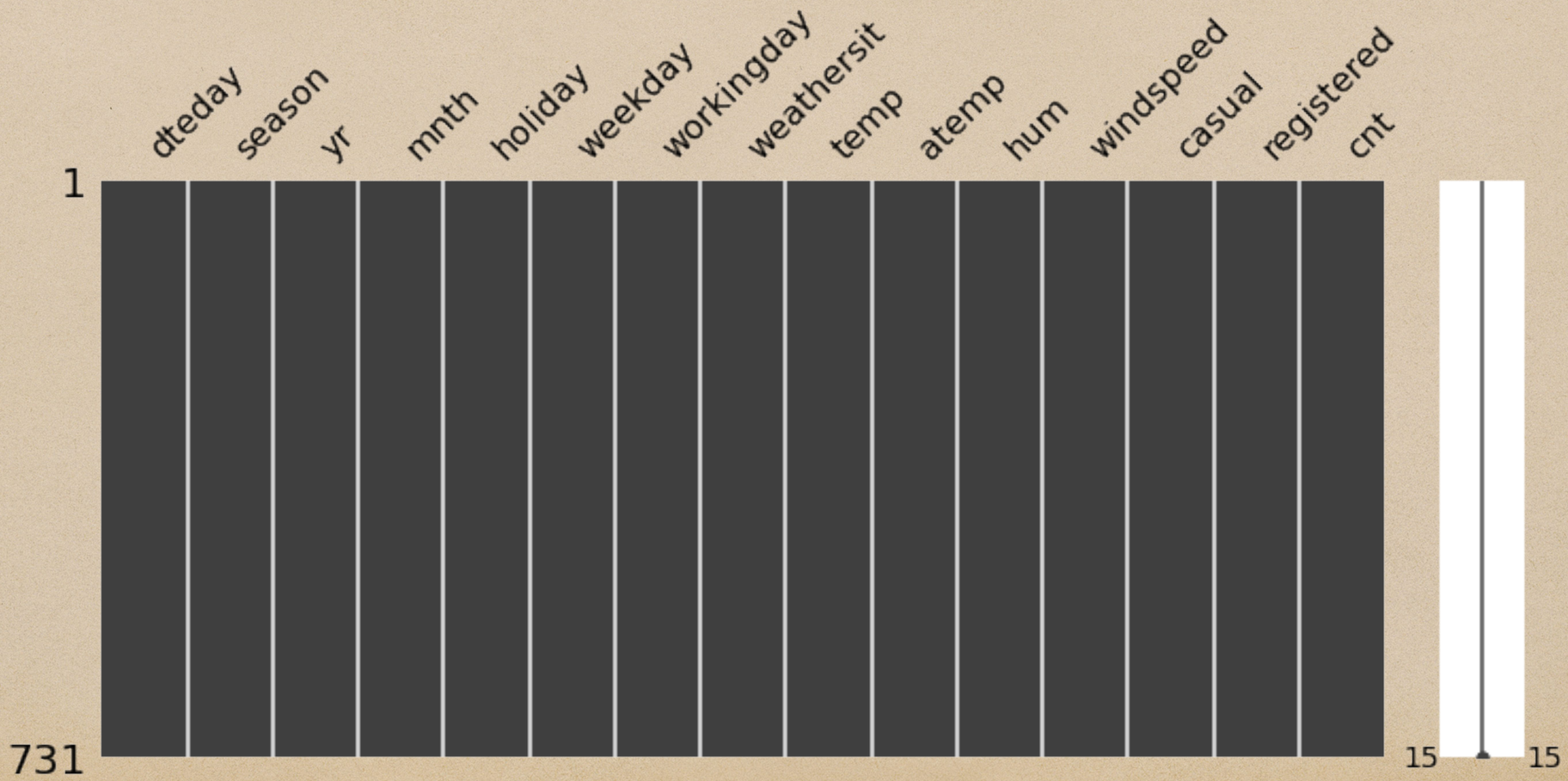
How to predict the bike rental numbers with the highest accuracy based on weather situation, and time of the year, and other crucial factors?

Sub-questions

- Daily Trend: Registered users demand more bike on weekdays as compared to weekend or holiday.
- Rain: The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.

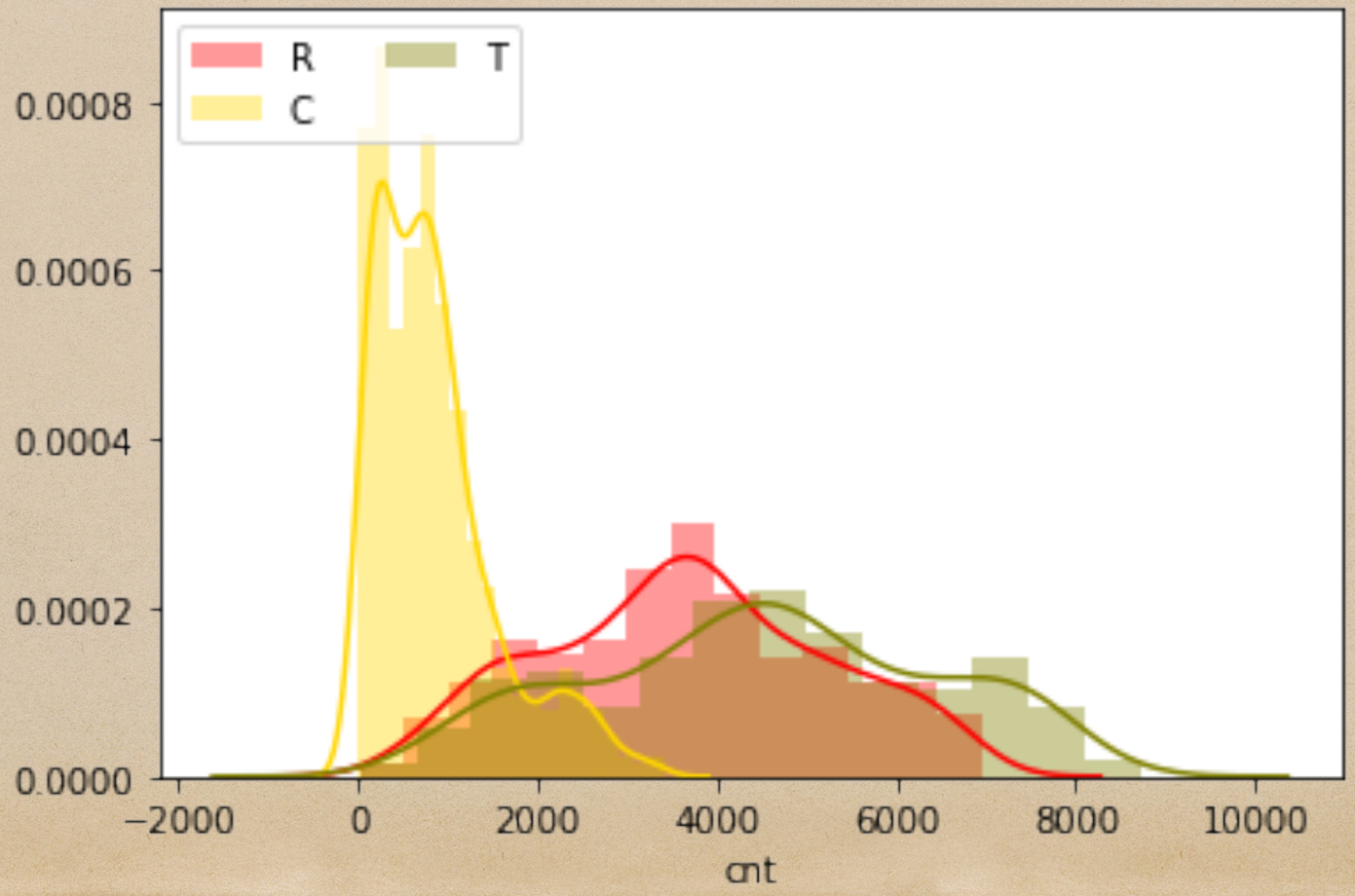
Explanatory Data Analysis

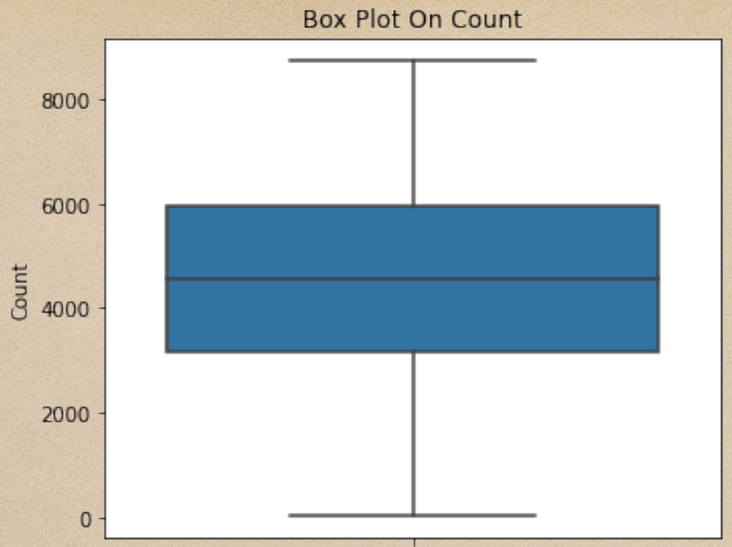
Missing Value Check

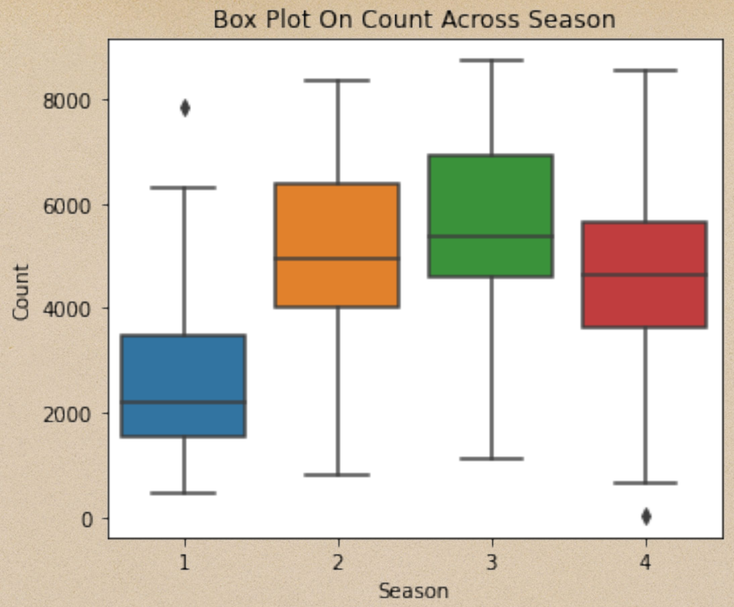


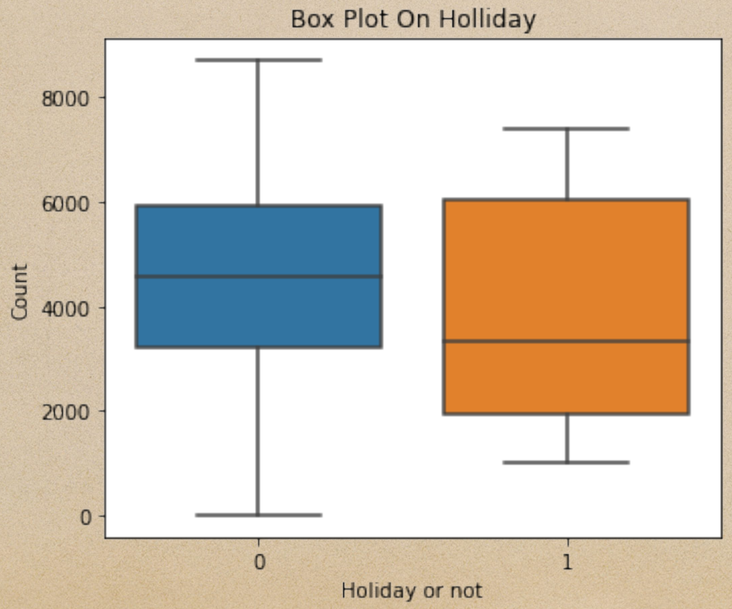
Comparing Registered (R), Casual (C), and Total (T)

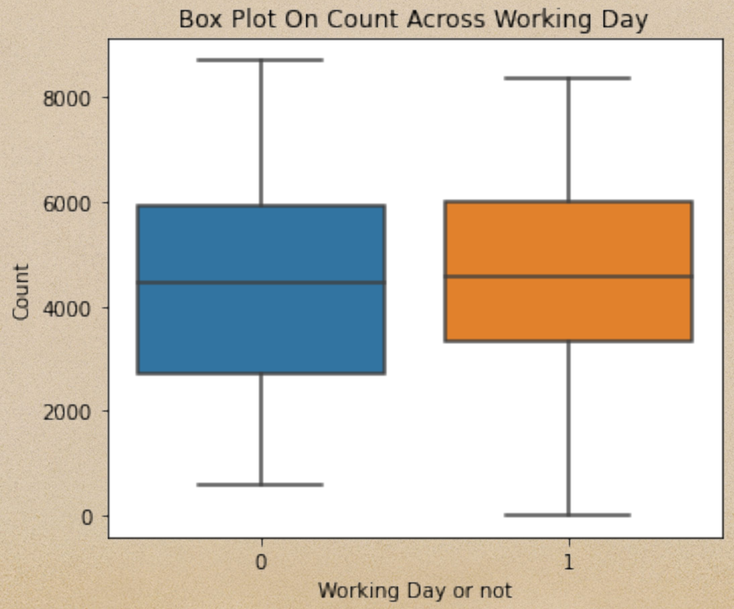
Registered and Total are normal distributions while Casual is highly skewed to the right.





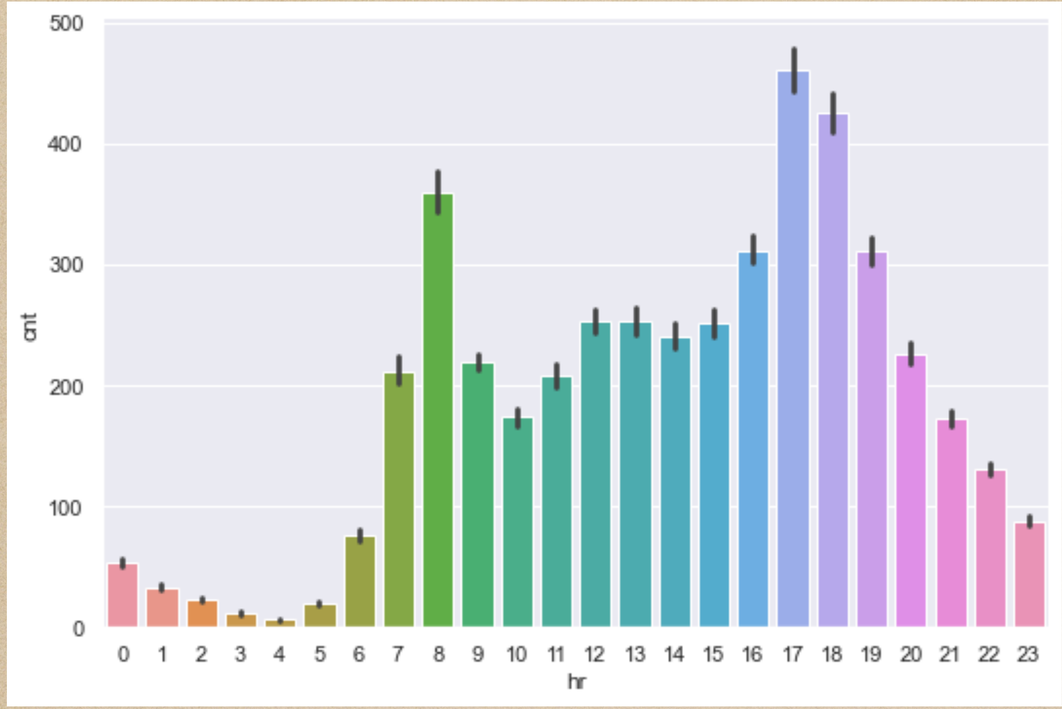




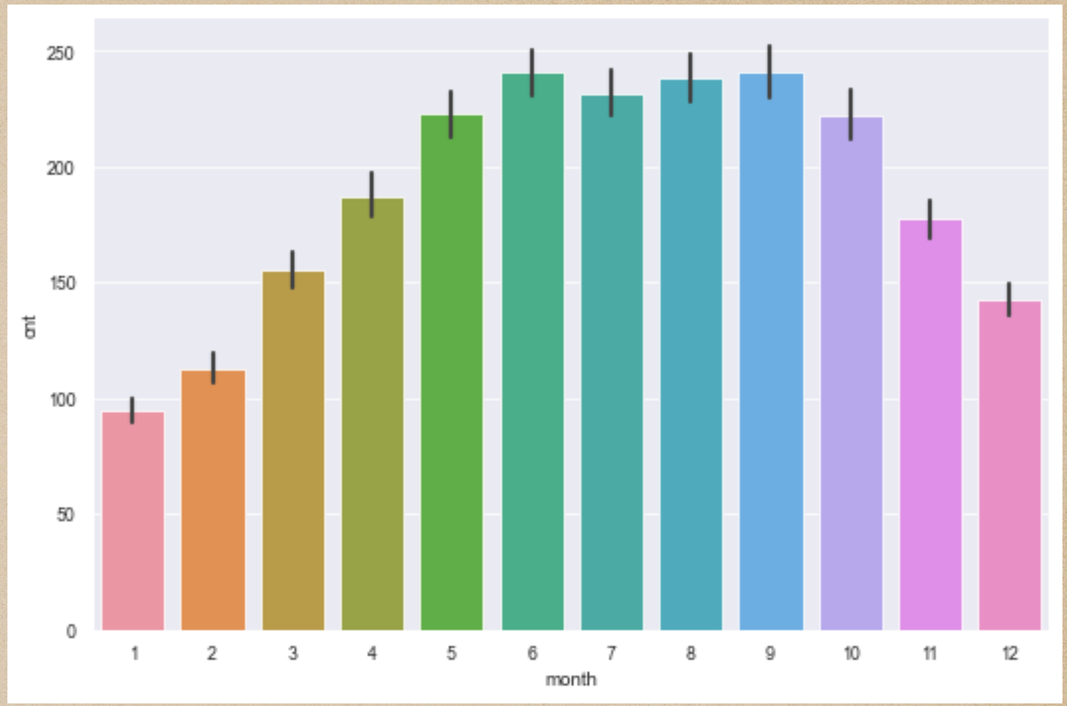


Outlier Check

Count Variation with Hour

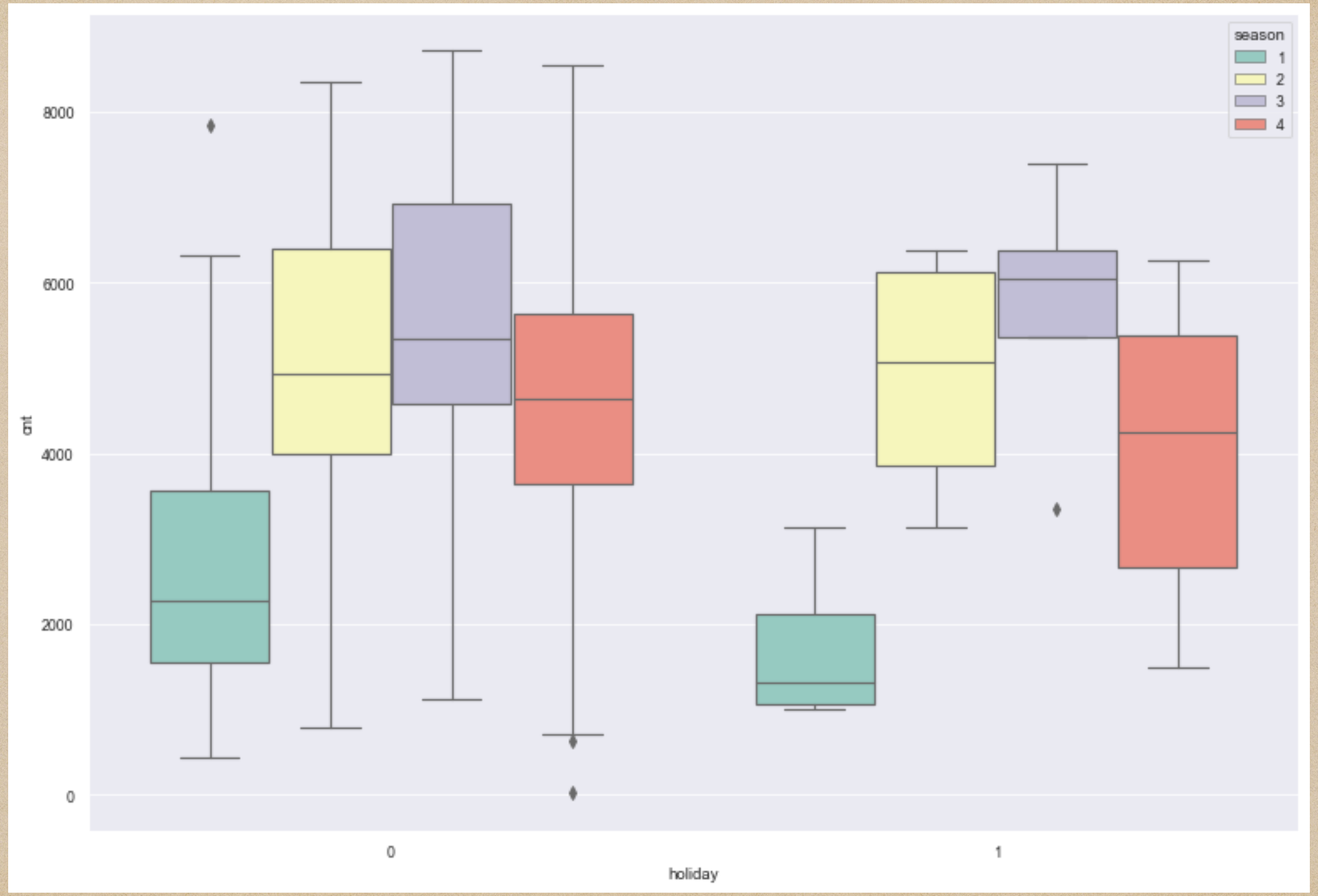


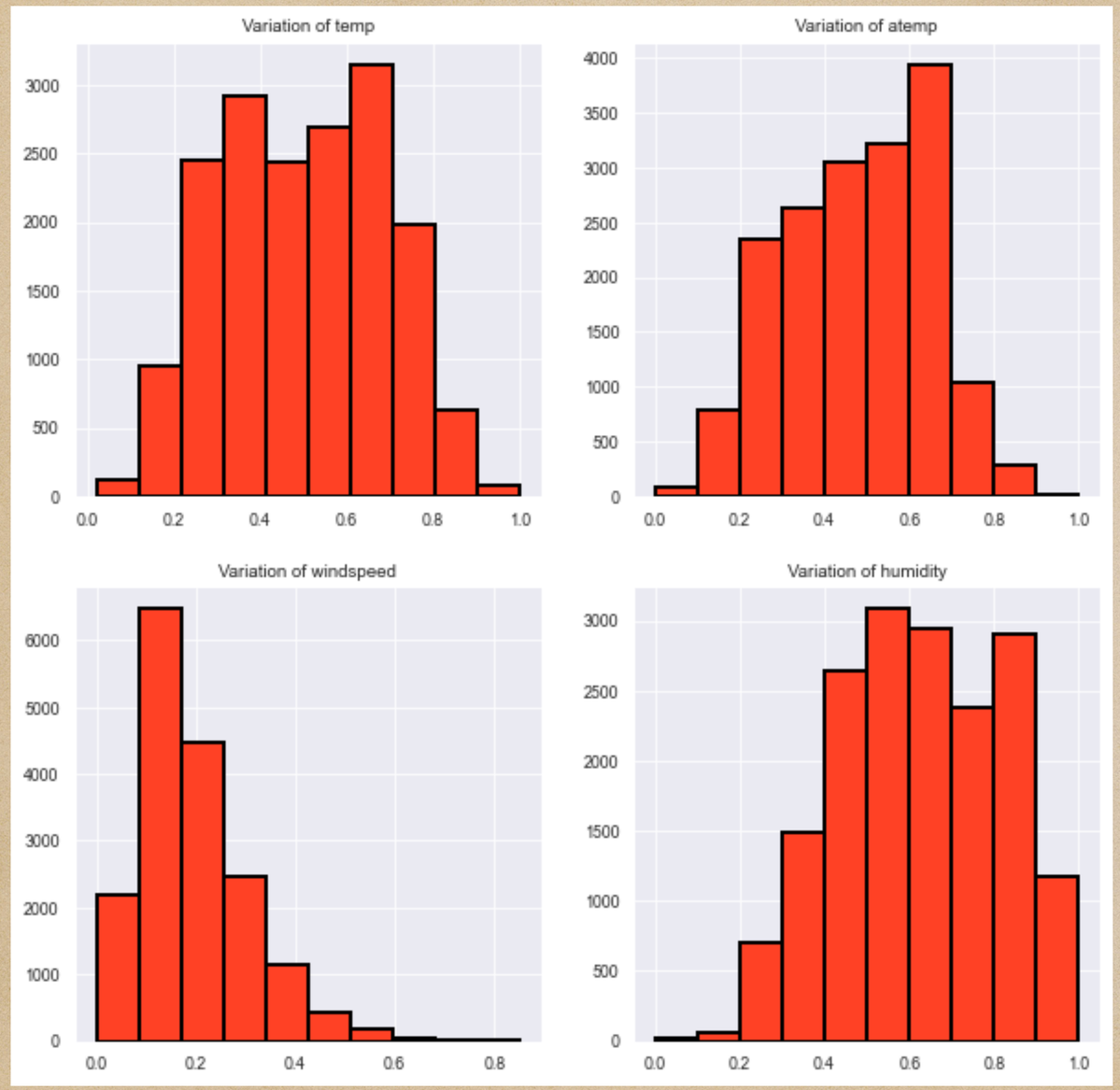
Count Variation with Month



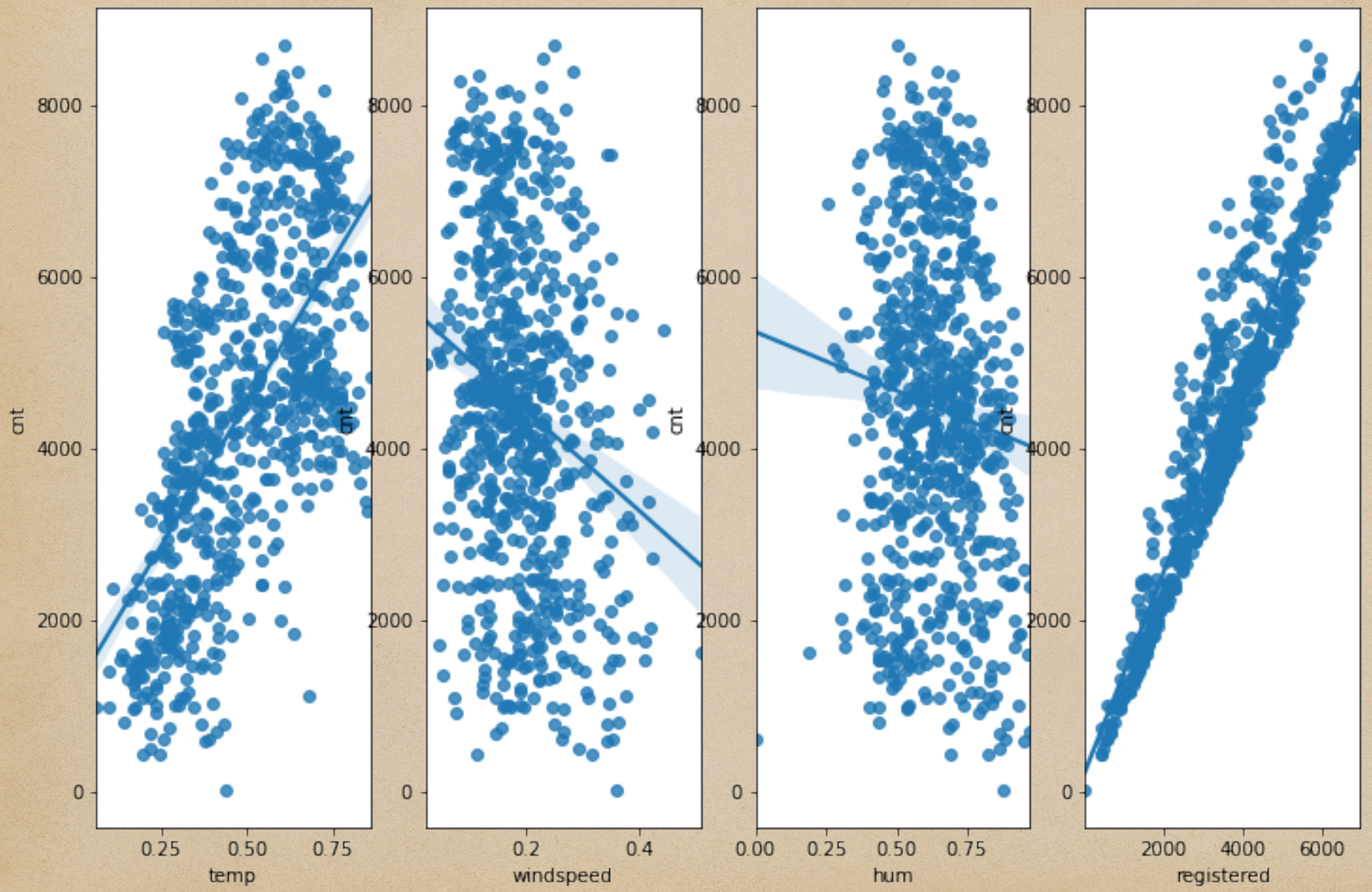
We see that the count of bike rental gets at its highest at rush hours: 7-9AM and 4-8PM.
Also, month affects season and that affects whether people take bike or not.

Box plot of holiday bike usage in each Season (Season 1: Winter)



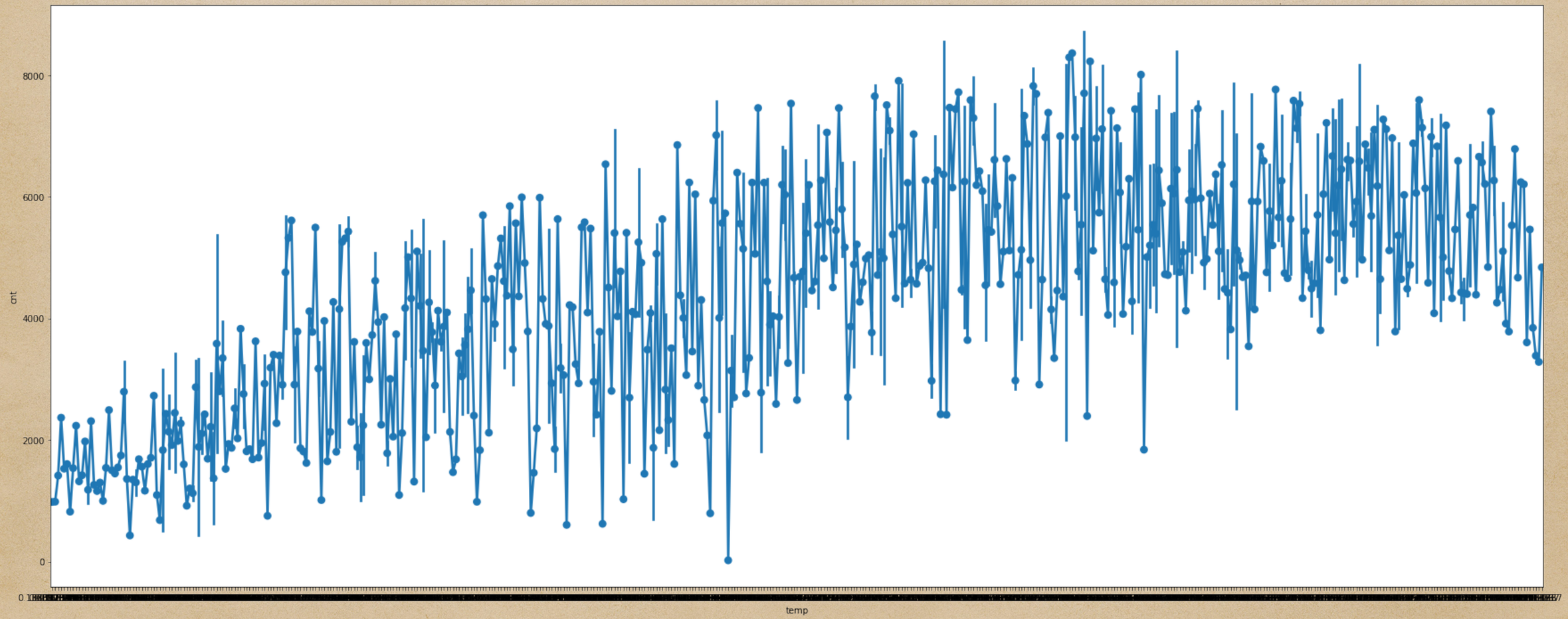


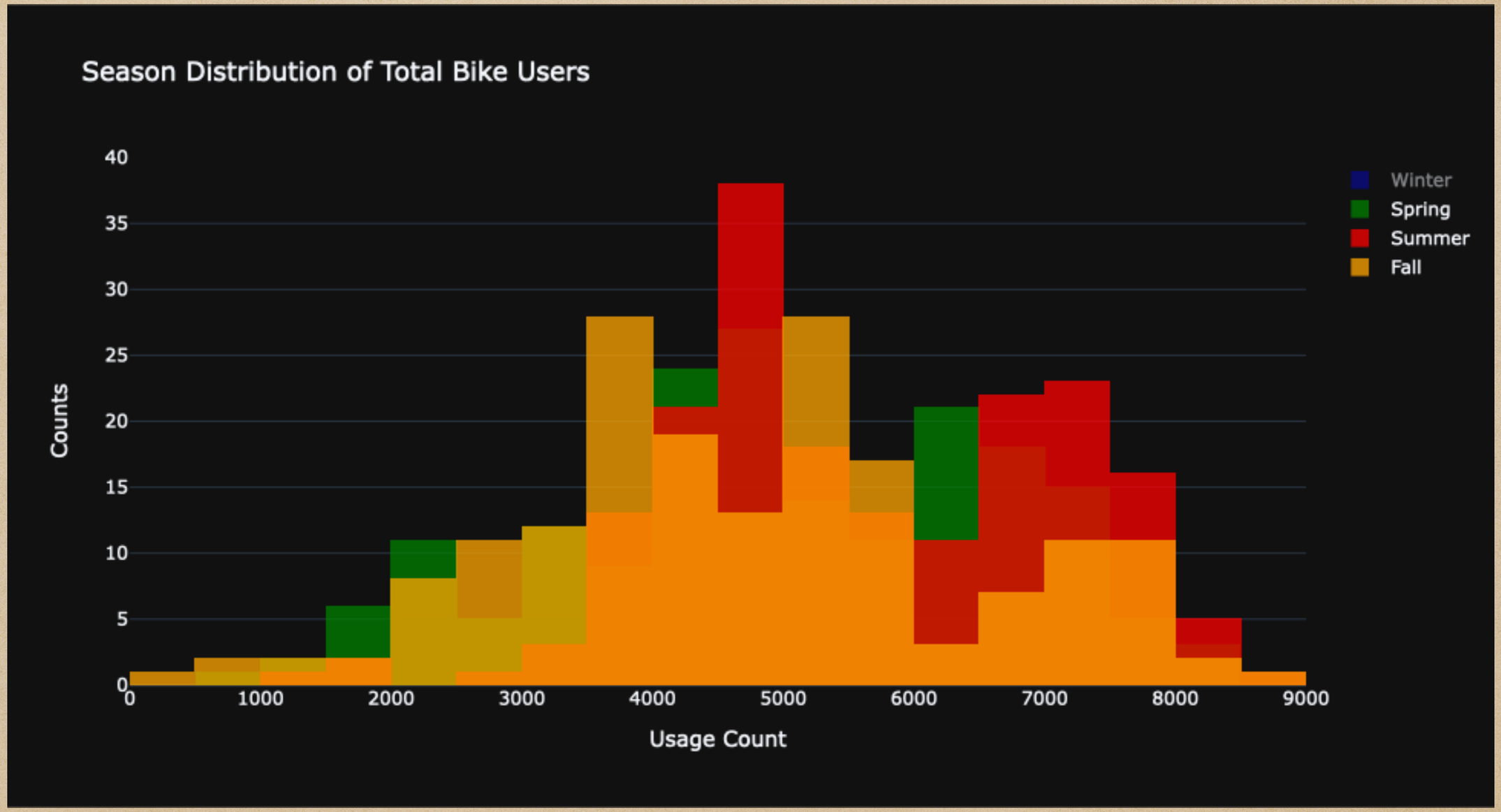
Visualizing all continues variables with Histogram

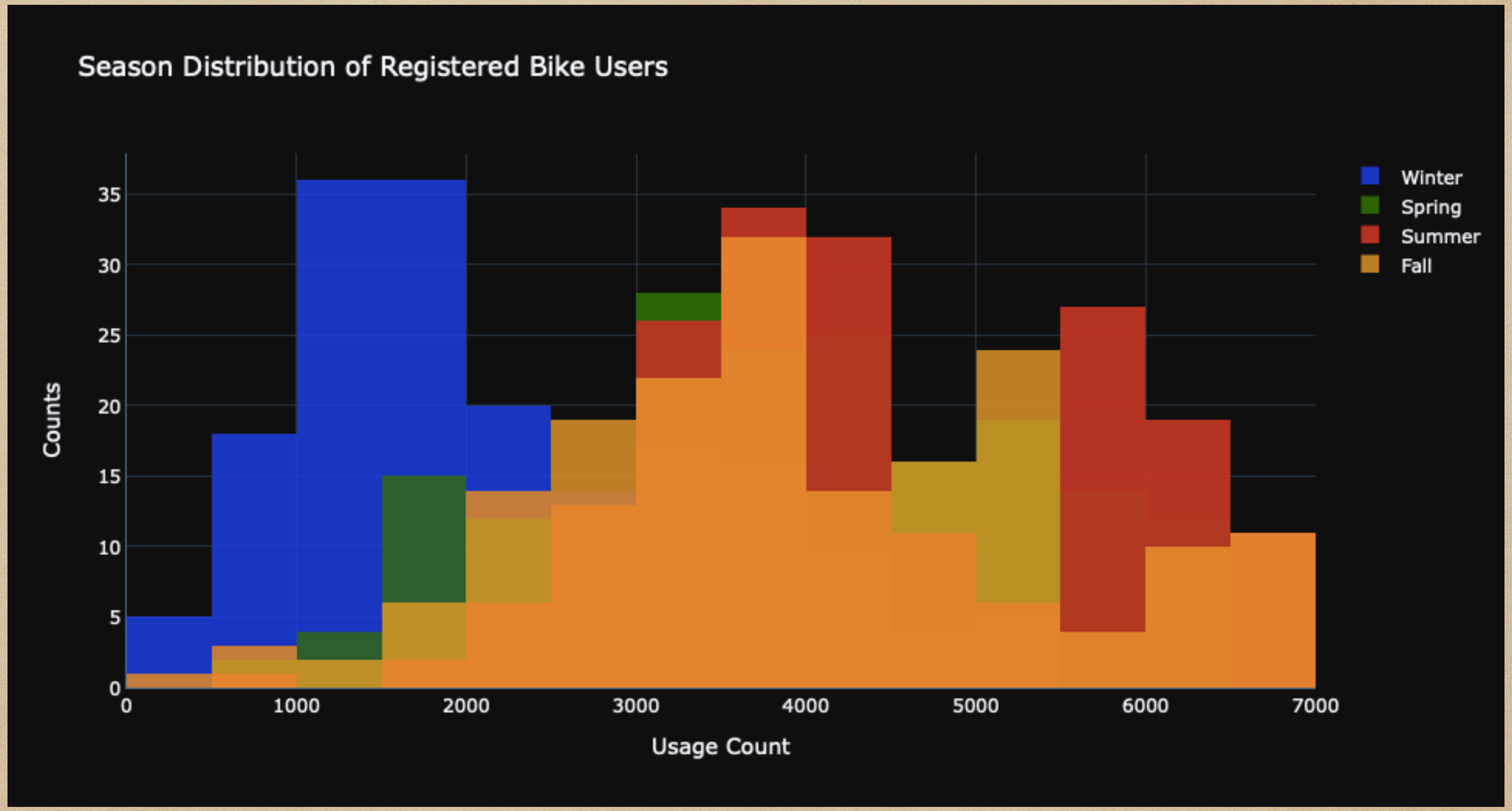


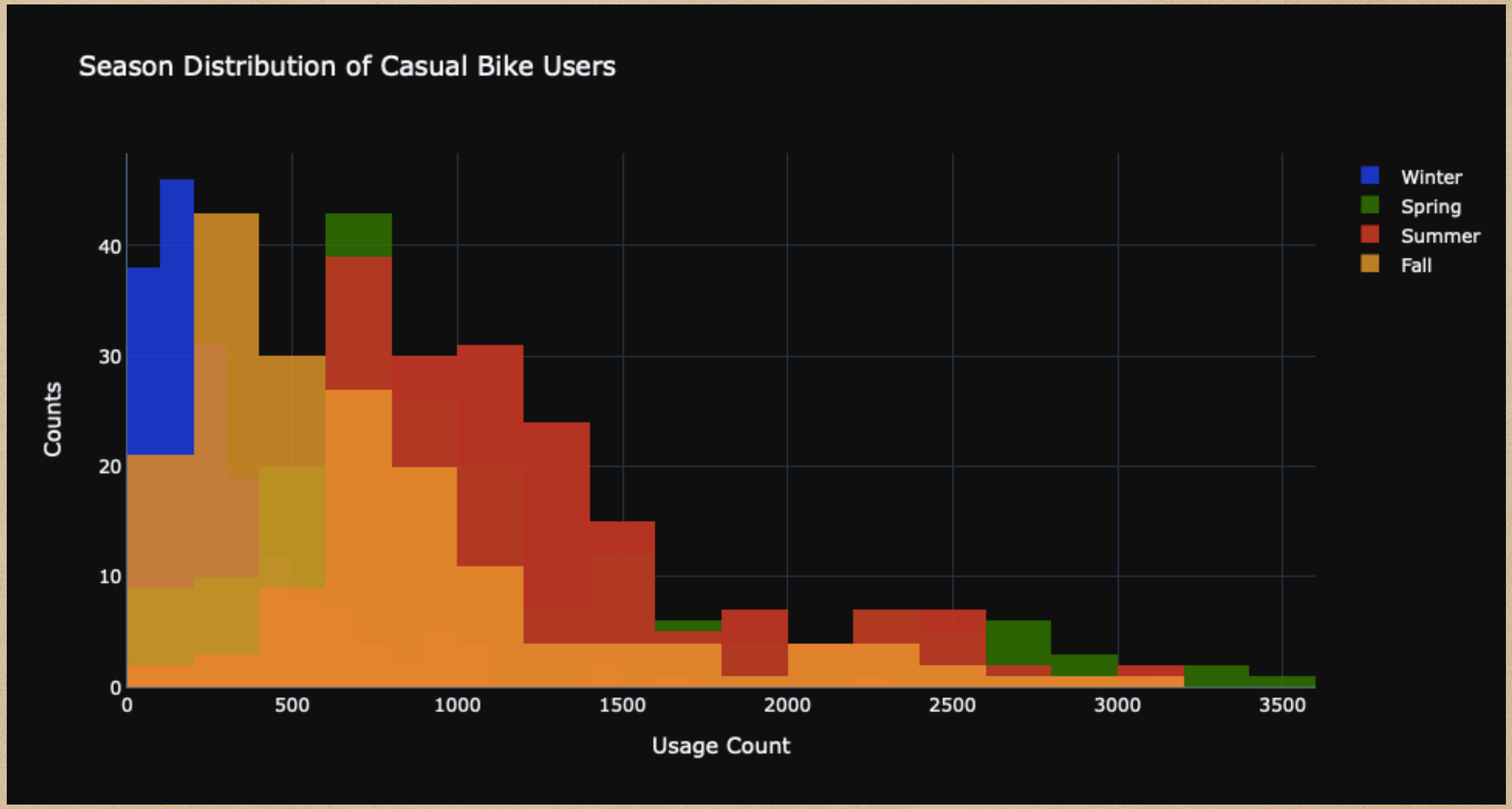
Scatter Plots

Positive and Strong Relationship of Count and Temp shown in a Point Plot




Season Distribution of Total Bike Users
Counts
Usage Count
Winter
Spring
Summer
Fall
0
5
10
15
20
25
30
35
40
0
1000
2000
3000
4000
5000
6000
7000
8000
9000


Season Distribution of Registered Bike Users
Winter
Spring
Summer
Fall
Counts
0
5
10
15
20
25
30
35
0
1000
2000
3000
4000
5000
6000
7000
Usage Count


Season Distribution of Casual Bike Users
Winter
Spring
Summer
Fall
Counts
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500
Usage Count

Correlation Heat Map Matrix

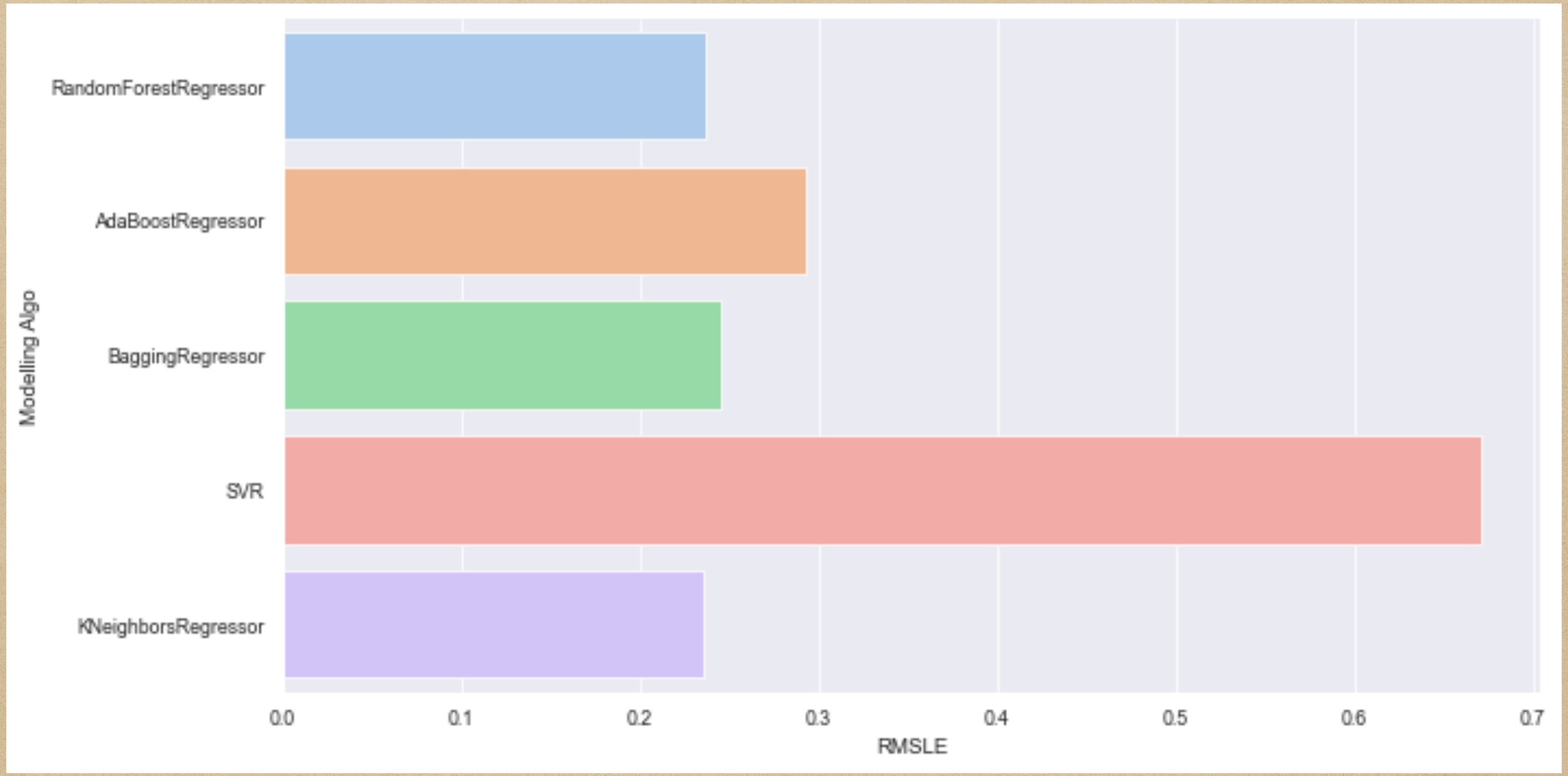
Inferences from the Correlation Heat-map:

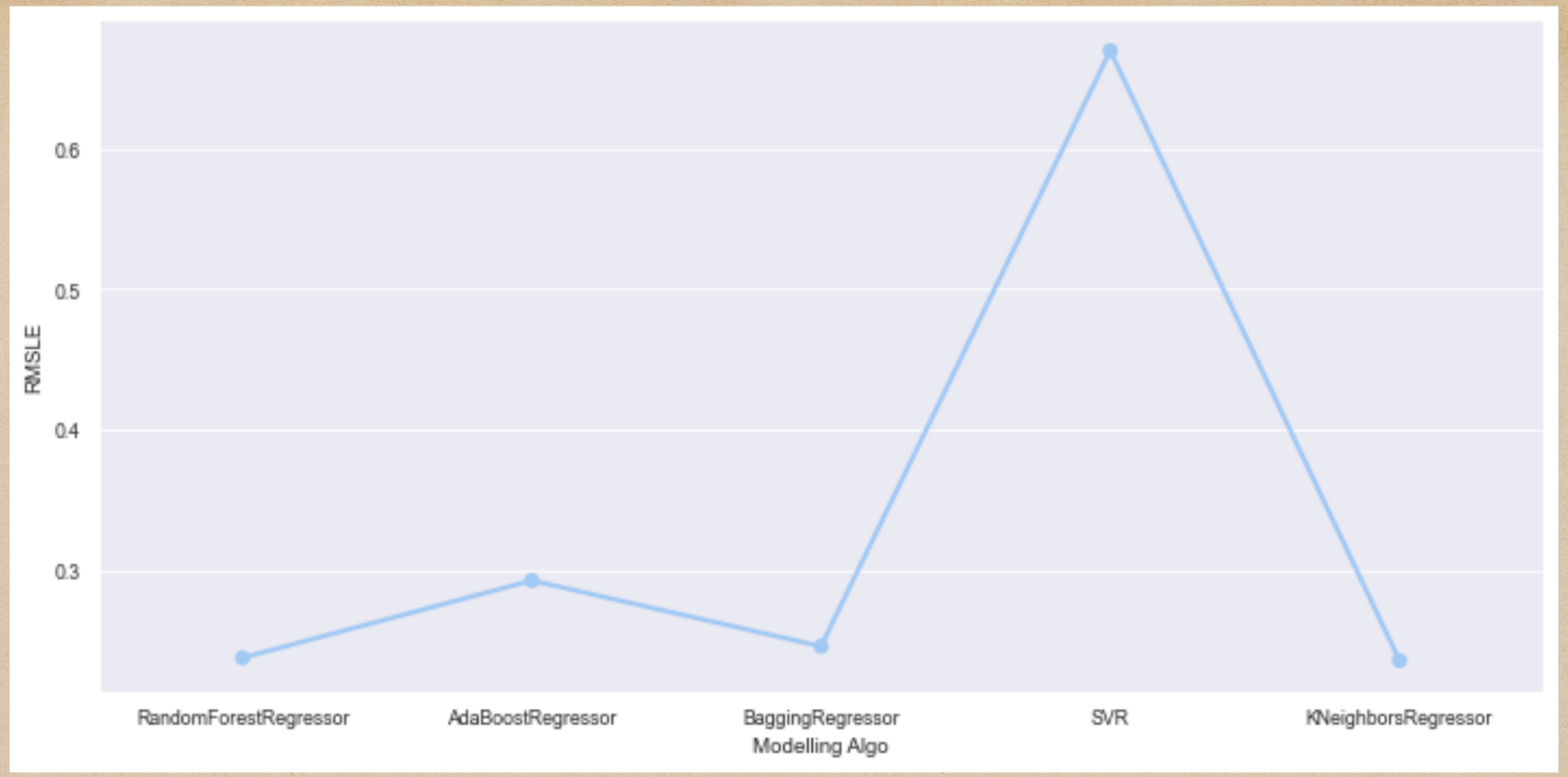
- temp and atemp are highly related as expected so we must omit one from our modeling.
- humidity is inversely related to count as expected: meaning as the weather gets more humid, people will not like to travel on a bike.
- casual and working day are highly inversely related.
- count and holiday are highly inversely related.
- temp and/or atemp highly effect the count.
- weather and count are highly inversely related. This is because in our data as weather increases from 1 to 4, it implies that weather is getting worse, so people are less likely to rent bikes.
- registered/casual and count are highly related which indicates that most of the bikes that are rented are registered.

Analysis Plan

- Analysis Goal: We want to predict the number of bike rentals.
- Methodology(-ies): Random Forest, SVR, KNeighborsRegressor
- Prioritization: Bike Rental Prediction, Weather Pattern, Timeframe pattern

Modeling Results


RandomForestRegressor
AdaBoostRegressor
BaggingRegressor
SVR
KNeighborsRegressor
Modelling Algo
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
RMSLE



The Random Forest Regressor gives us the least RMSE, hence we will use this to make predictions for the future bike renting demand.