Predicting Demand for a Bicycle Sharing Company

Debadrito Saha (DS-July'20-Praxis)

| CRISP-DM |

Introduction:

Bicycle Sharing Systems are a means of renting bicycles where the process of obtaining membership, rental,and bike return is automated via a network of bike-stations located throughout a city. The users of bike sharing systems can pick up bicycles from a bike-station in one location and return them to another in possibly any location of the city.Bike Sharing System ensures that pollution is reduced with use of bicycles there is reduction in use of motor vehicles which leads to reduction in emission of pollutants in the air.

Project-Description:

This project work, focuses on which algorithm can work better for the real world problem of bicycle sharing demand prediction and there after recommending which are the important factors for predicting the demand.

Problem-Statement:

In a bicycle sharing systems it is very important for the administrators to know how many cycles will be needed in each bicycle station, knowing this count, it enables them to arrange proper number of cycles at the stations and decide whether a particular station needs to have extra number of bicycle at that stands. Utilizing a seperate weather data and not the usage pattern here, its needed to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Business Goal:

To predict the demand of the bicycles for rent on a given day on the basis of the data set provided so that utilization of the cycles are maximum there by resulting in maximising the revenue.Also Identifying the major features for predicting.

Data-Source:

The dataset in this project is provided by Kaggle and is an open dataset hosted at UCI Machine Learning Repository " **Bike Sharing Data Set**"

Data:

Independent Variables:

datetime - hourly date + timestamp (2011-12)
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 holiday
weather -
1: Clear, Few clouds, Partly cloudy, 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

Dependent Variable:

count - number of total rentals

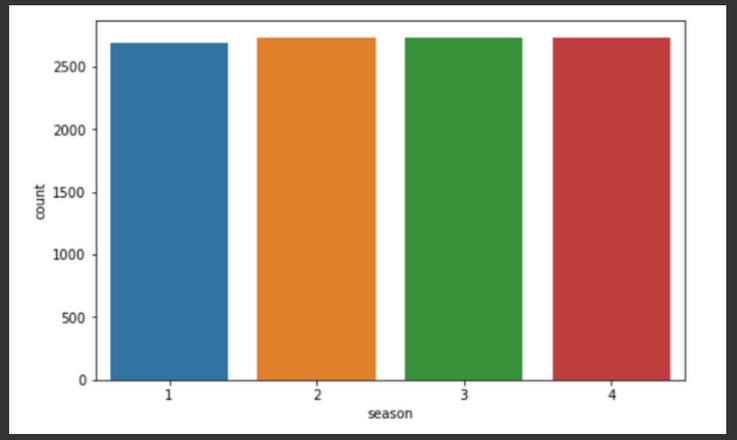
Feature definition

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
```

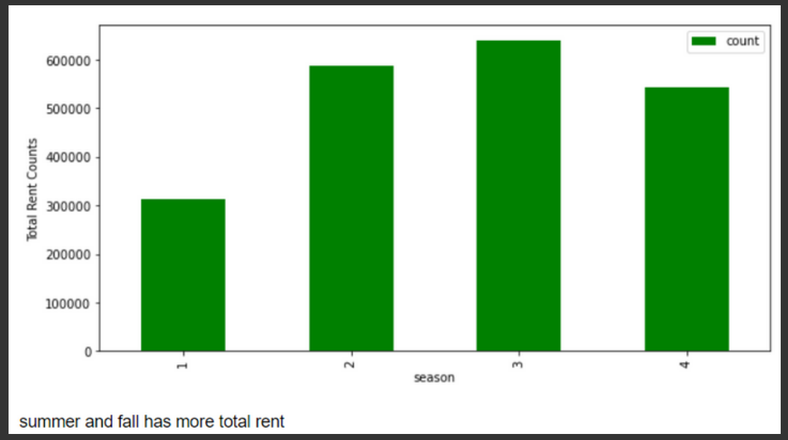
Data Types

```
datetime      0
season        0
holiday       0
workingday    0
weather       0
temp          0
atemp         0
humidity      0
windspeed     0
casual        0
registered    0
count         0
dtype: int64
```

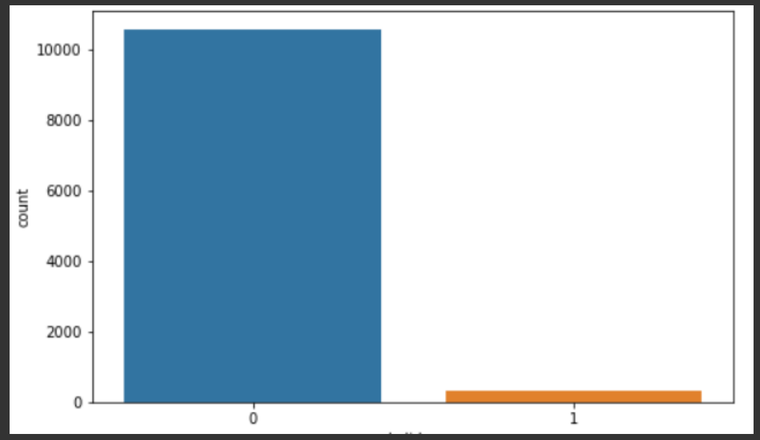
Null Count



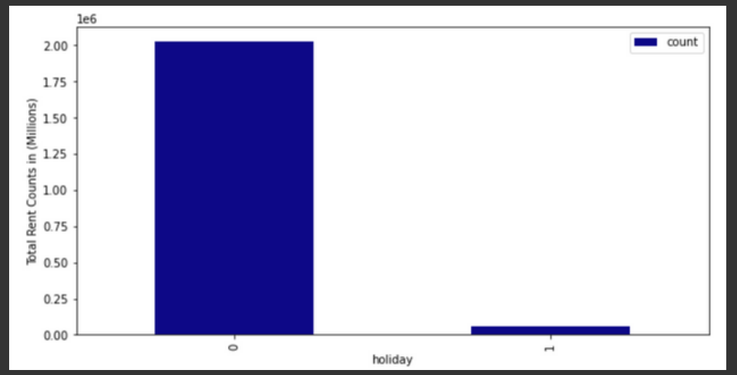
data counts of seasons



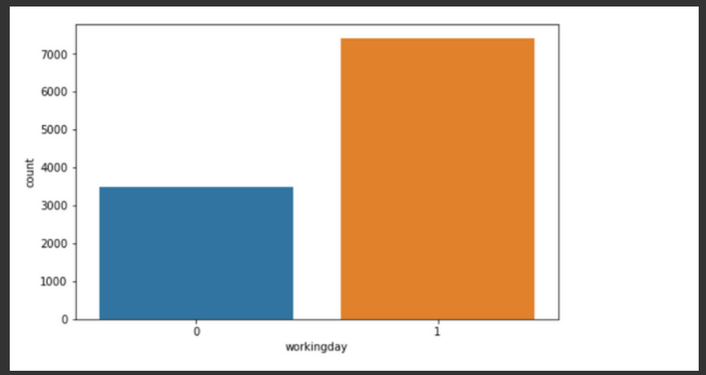
Number of rent counts group by season



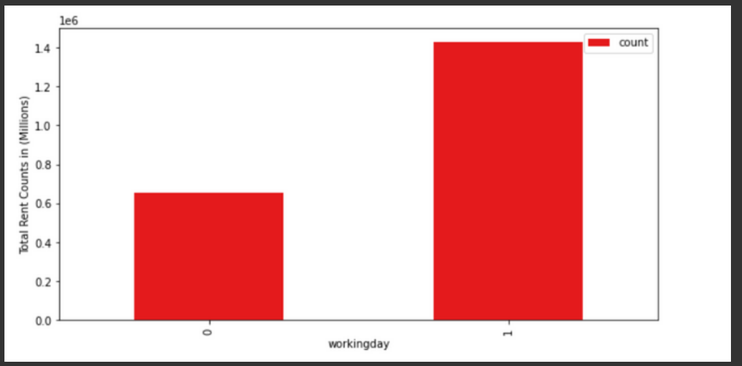
holiday vs not holiday data counts



Number of rent counts on holiday, not holiday basis



Working vs not workingday data counts



Number of rent counts on working day, not workingday basis

summer and fall has more total rent

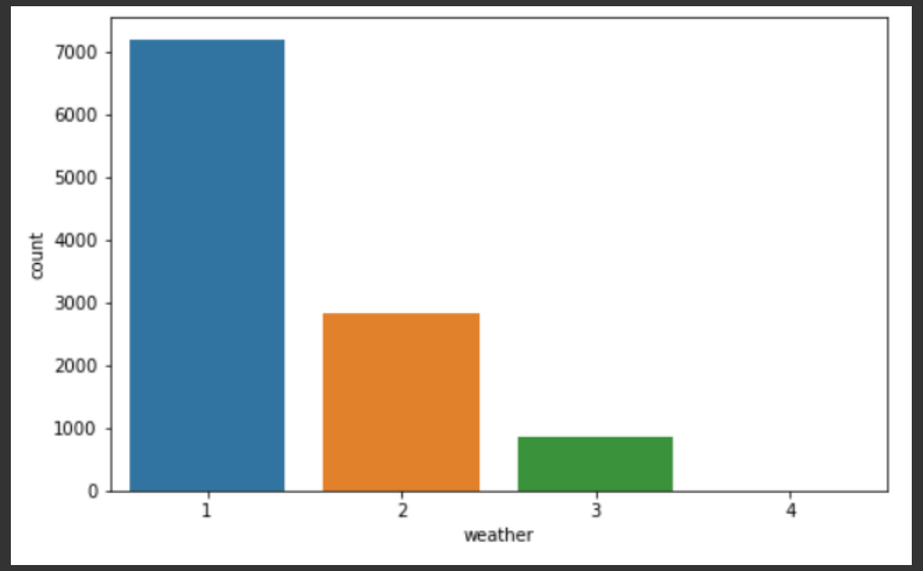
Total Holiday is very less

When its not a holiday total rent count is more

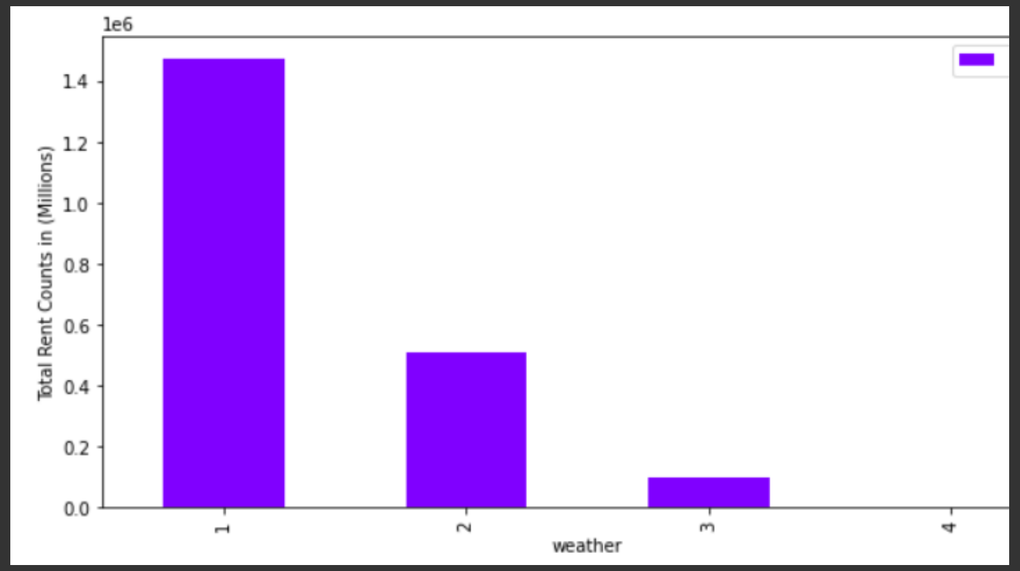
When its not a working day its weekend which may include a part of a holiday but not fully. Working day is neither a holiday nor a weekend.

When its not a workingday total rent count is less

POINTS



Data Counts weather basis

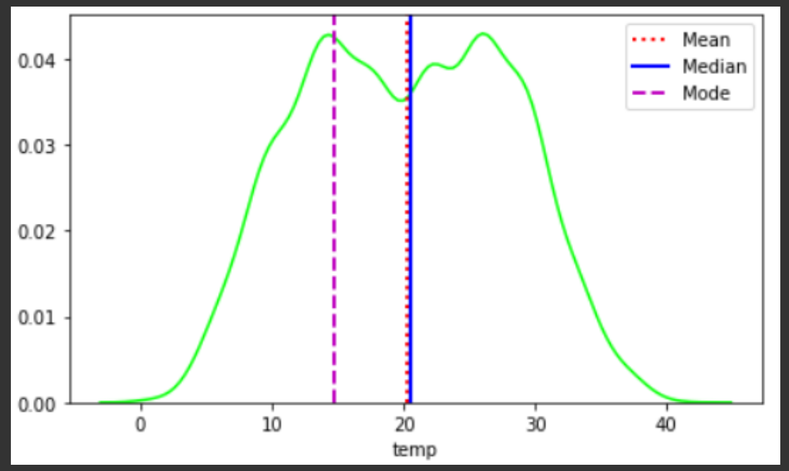


Rent Count weather basis

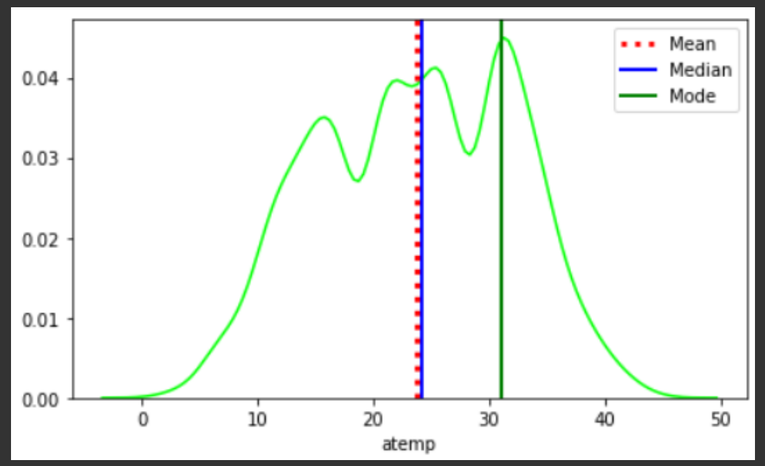
Spring and Summer has more number of records

Spring and summer having more number of total rents count

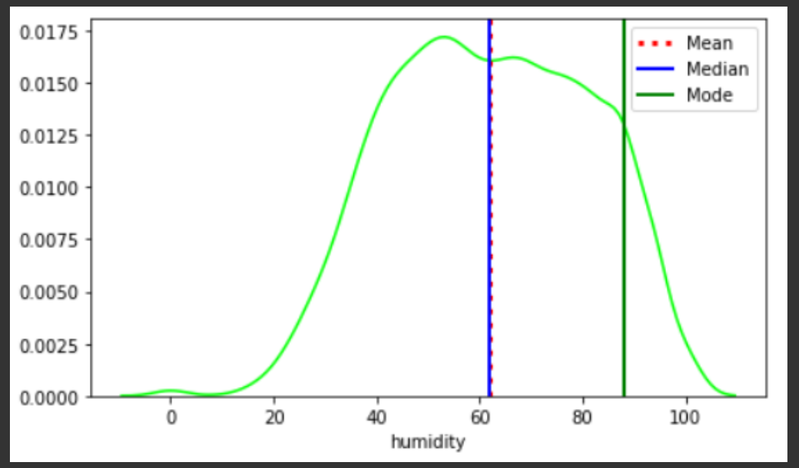
POINTS



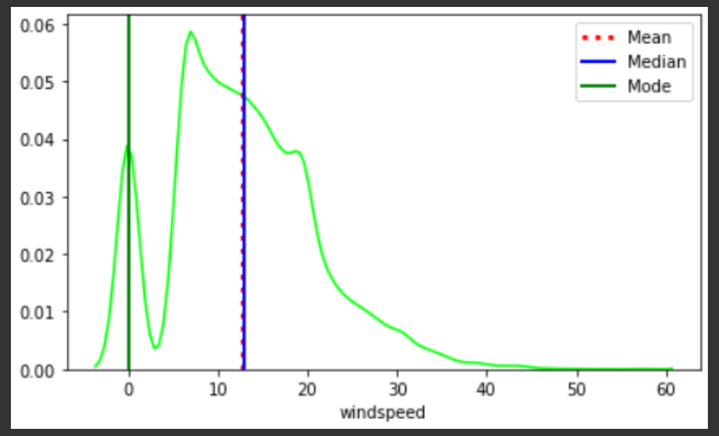
Temperature



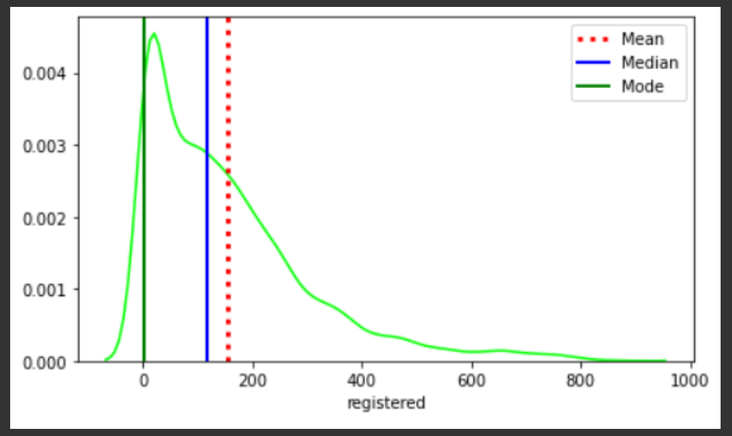
Atemp



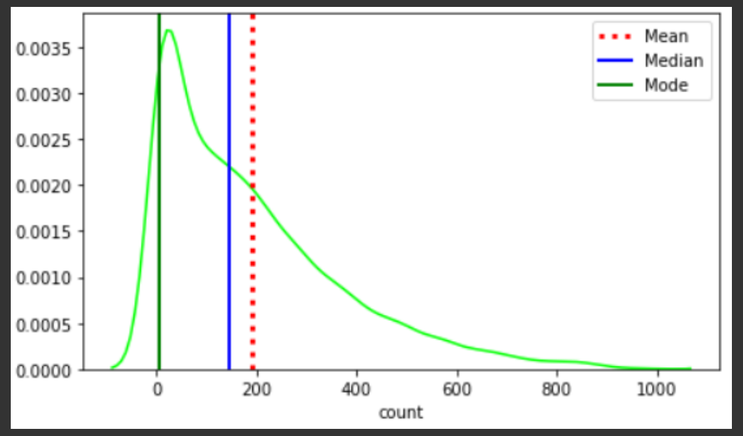
Humidity



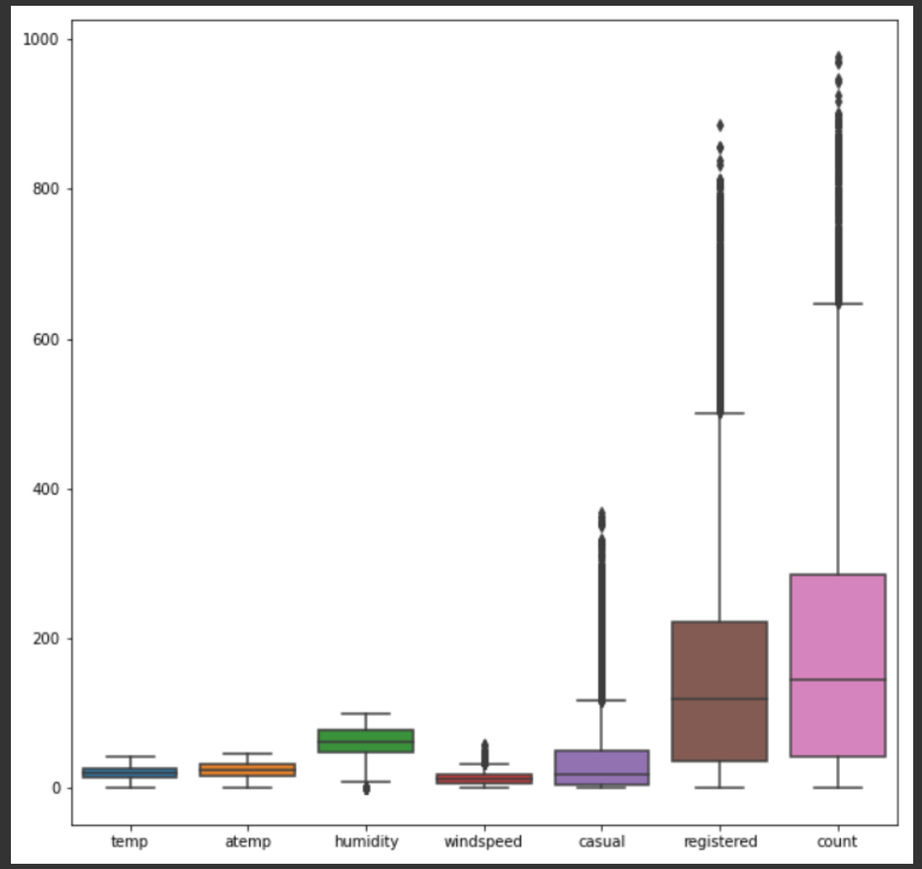
wind-speed



registered

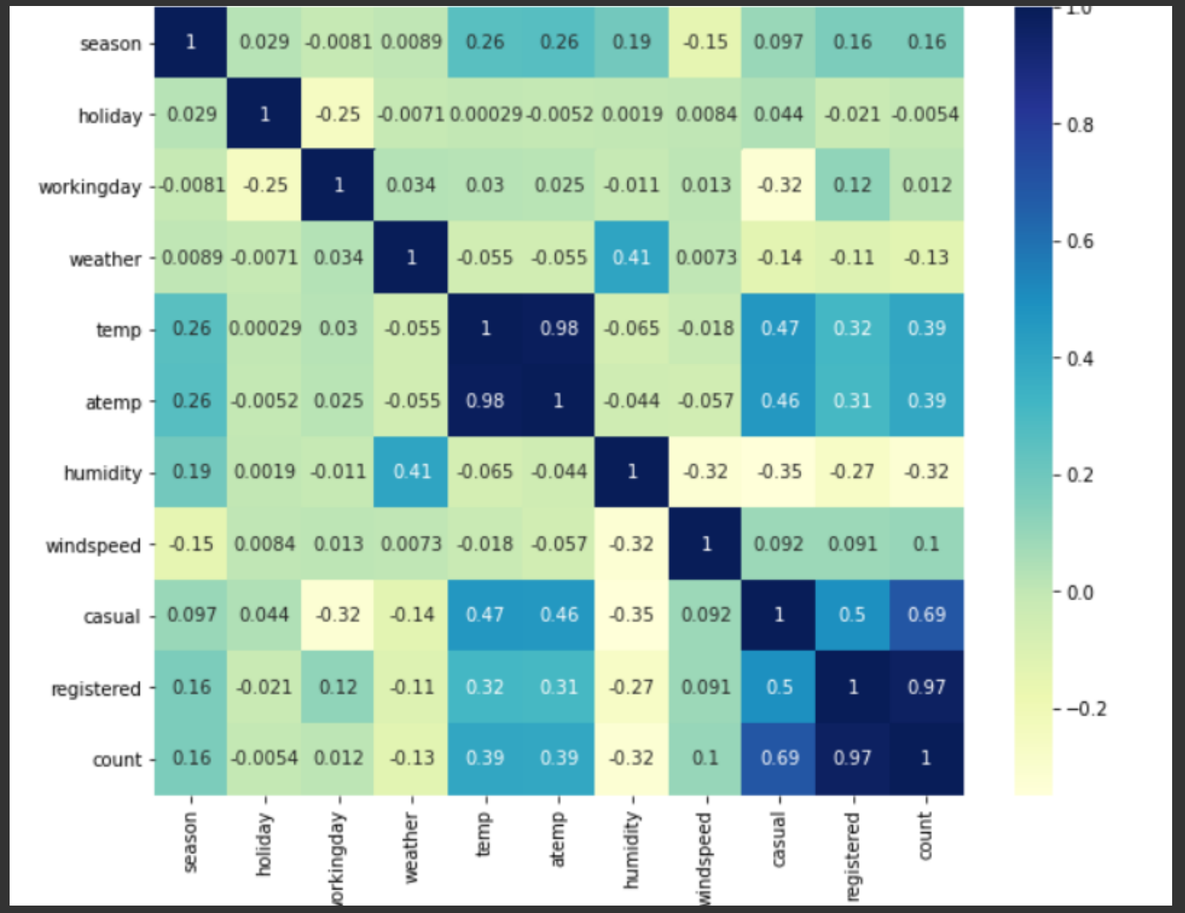


count

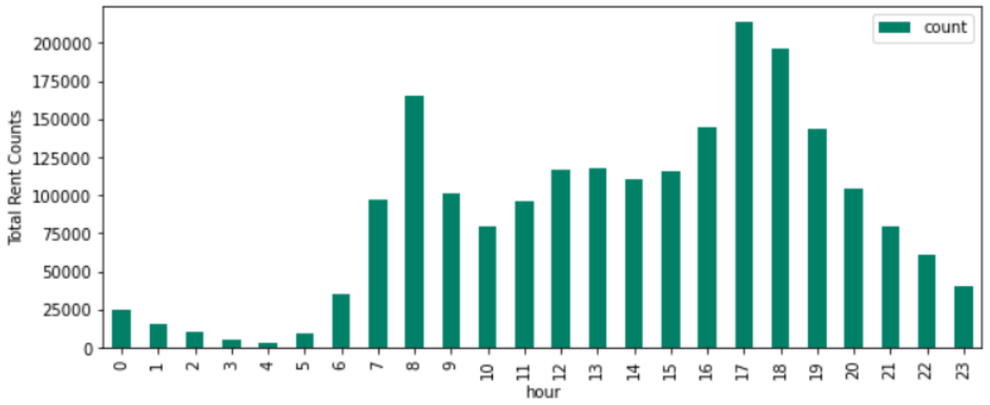


Left to Right

- temp
- atemp
- humidity
- winspeed
- casual
- registered
- count



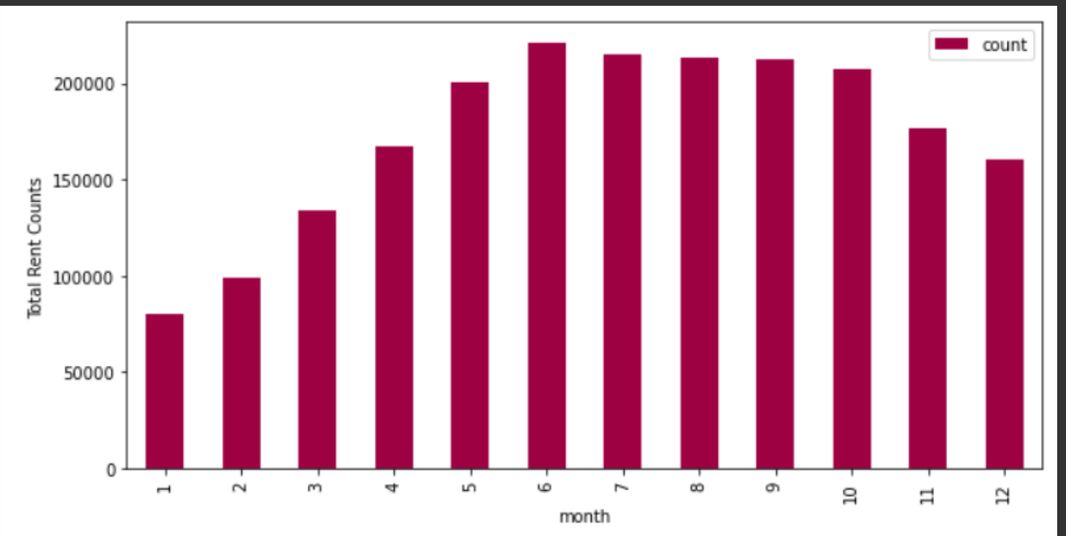
Heatmap



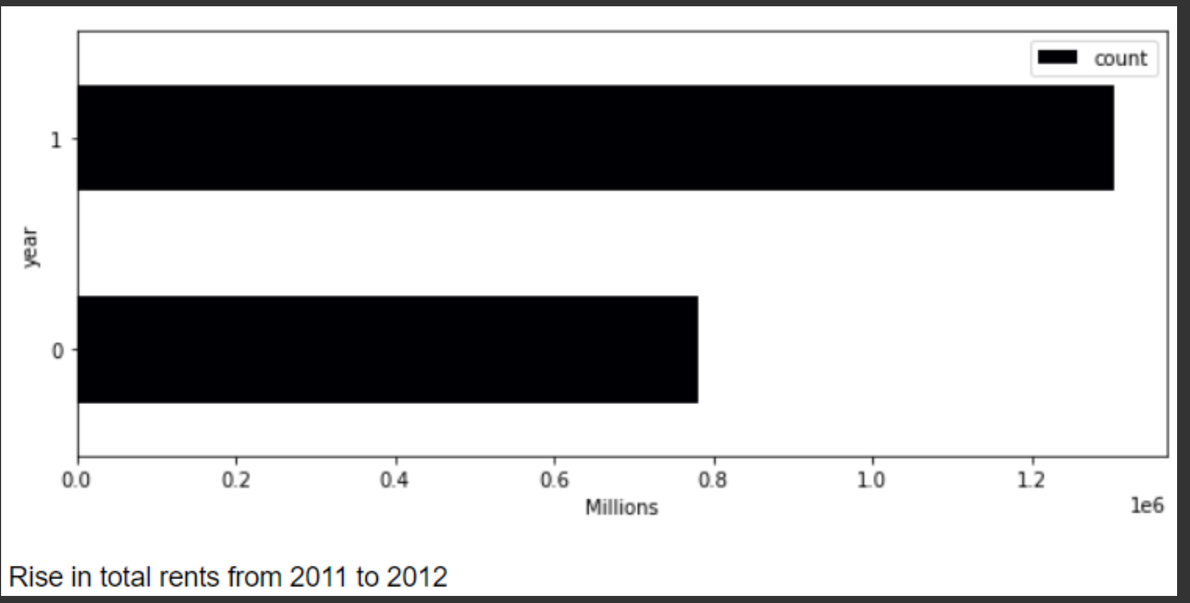
Here we can observe that around 7-9am in the morning and around 4pm-7pm there is a peak in count might because of office/duty etc hours

Dec-Feb there is a lesser rent may be because of winter season

There is a growth in the numbers of rent from 2011 to 2014

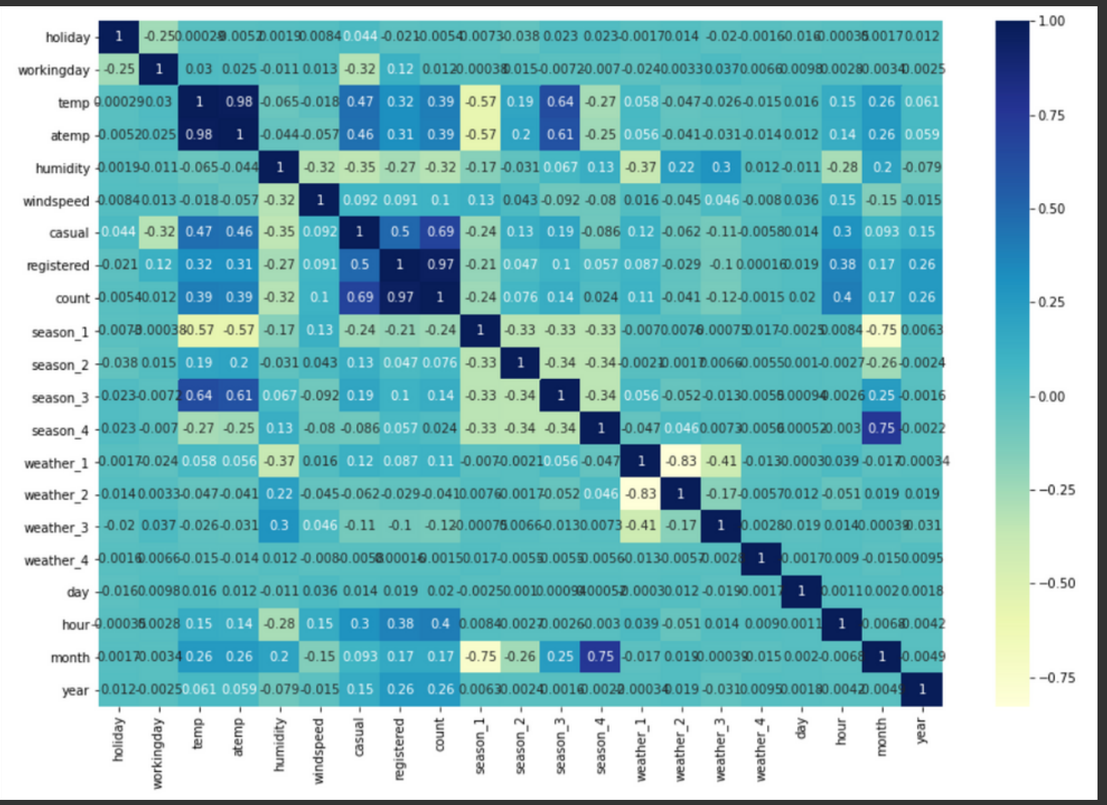


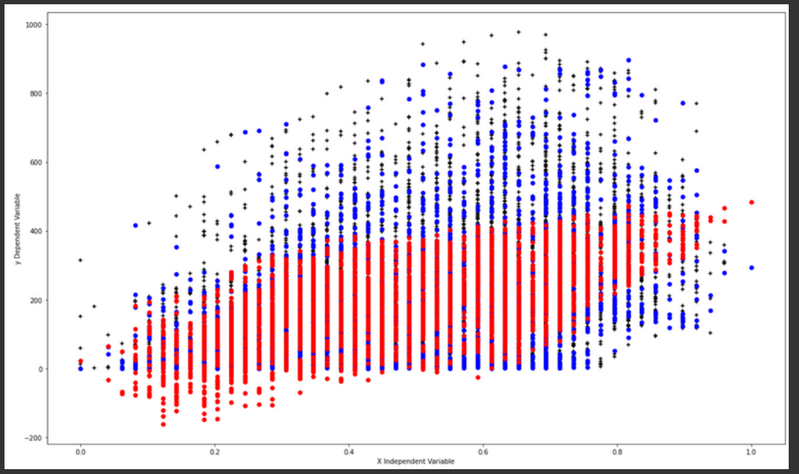
Total rents count follows a cyclic pattern peaking during summers and dipping in winters



Rise in total rents from 2011 to 2012

Heatmap post dummification,min-max scaling and removing dummy-trap --->



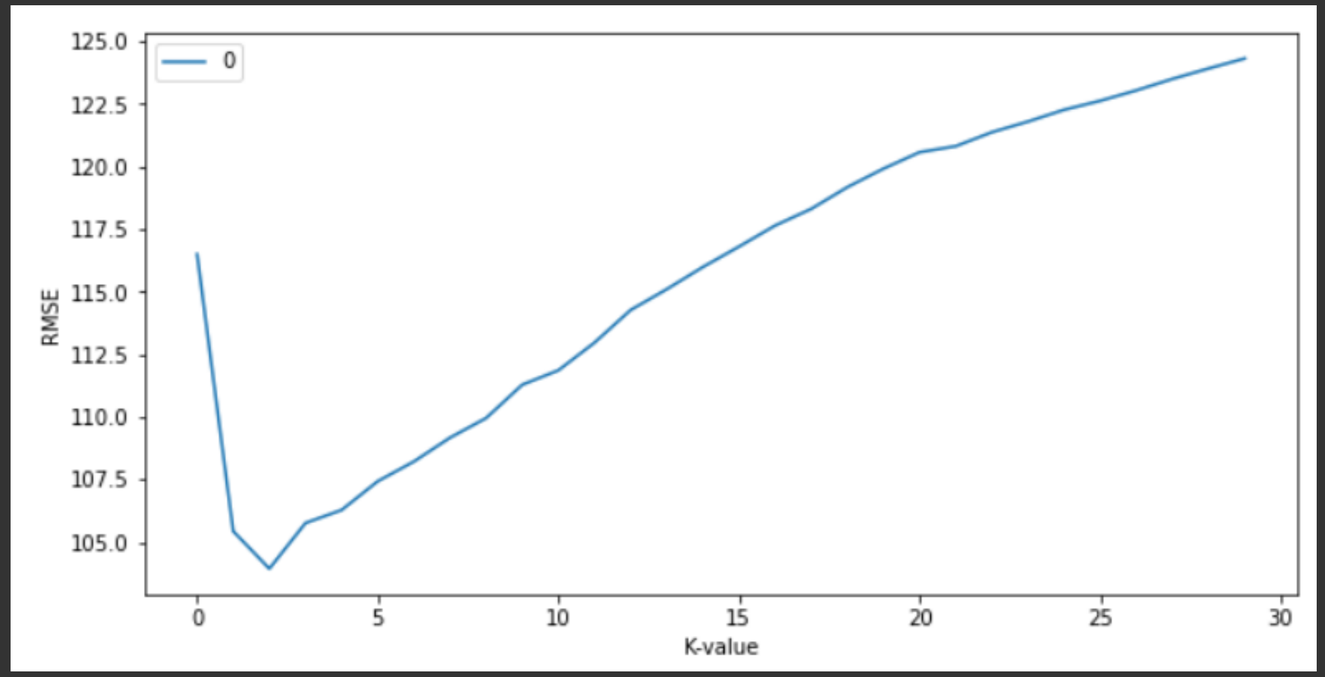


red-train
blue-test
black-predicted

	Coefficient
holiday	-13.715647
workingday	-0.185160
temp	240.827537
atemp	102.426020
humidity	-190.054567
windspeed	30.187555
season_1	14.005039
season_2	7.639326
season_3	-41.576393
weather_1	7.711626
weather_2	15.504847
weather_3	-23.216473
day	6.457380
hour	175.059023
month	101.791394
year	82.593440

Linear Regression Outcomes

MAE: 105.6315
MSE: 19347.67
RMSE: 139.095

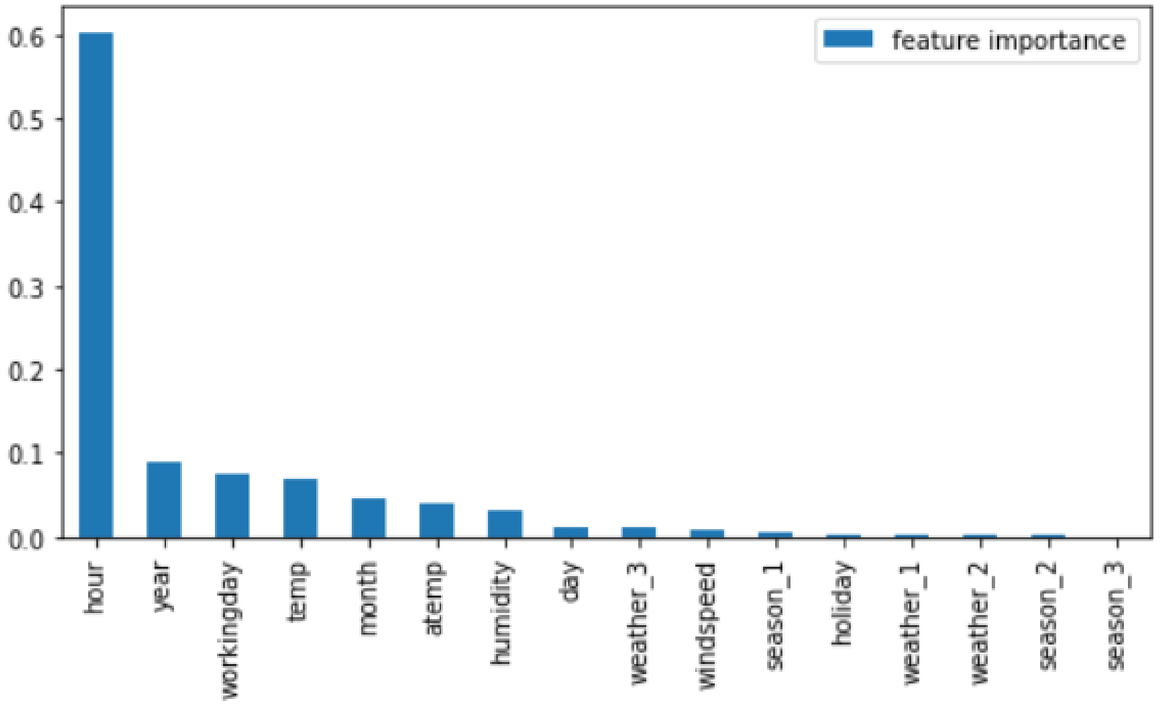


KNN-Regressor Outcomes

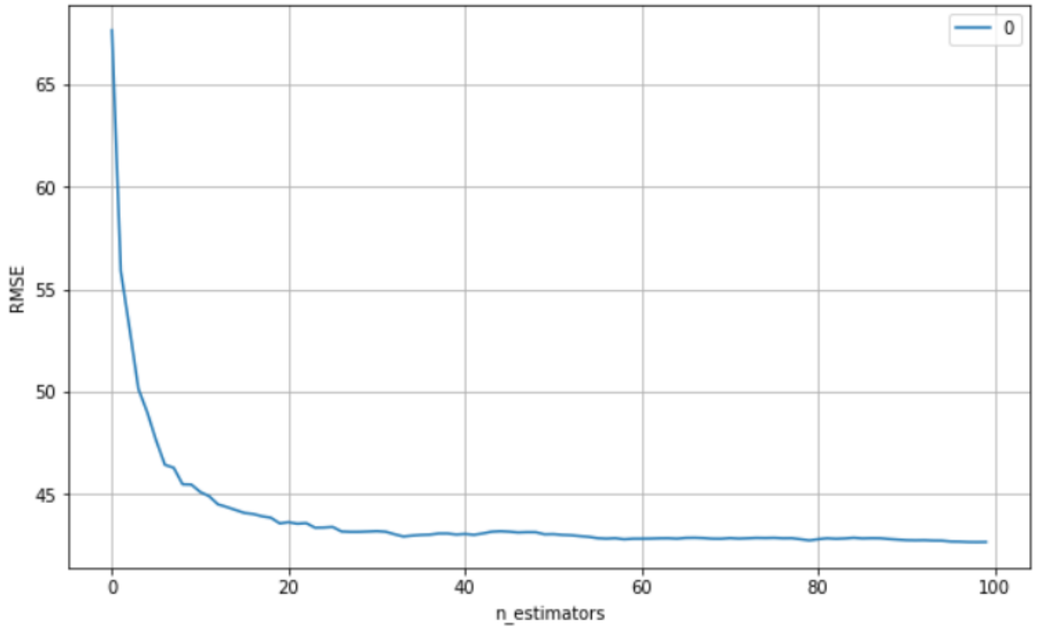
RMSE: 124.31780880156097

DT-Regression Outcomes

Selected Model



Feature Importances



Hyper-parameter Base learners (n_estimators)

We can observe that there is drastic dip in the RMSE values as we have increased the number of trees and and then the RMSE remains consistent from n_estimators=20 and there on.

RF-Regressor Outcomes

Hyperparameters

```
Splitting Criteria:  mse
Nos. of Trees:  100
max depth:  None
Base Estimator:  DecisionTreeRegressor()
Max Features:  auto
Min Sample Split:  2
Bootstrap:  True
```

RMSE: 42.6820

Conclusion

```
Random Forest RMSE: 42.682001391383174
Decision Tree RMSE: 124.31780880156097
KNN Regressor RMSE: 103.93610253853174
Linear Regression RMSE: 139.09592253007418
```

Random Forest has least RMSE

Thank You