Bike sharing demand

Abstract

The goal of this project was to use a multiple linear regression model for the prediction of demand for shared bikes.

Business Goal : Model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Design

The project originates from the Data Science Bootcamp T5 the data is provided by UCI Machine Learning Repository We would be interested in prediction the rentals on various factors including season , temperature , weather and building a model that can successfully predict the number of rentals on relevant factors **.**

Data

This dataset contains the seasonal and weekly count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding temperature and humidity information. Bike sharing systems are a new way of traditional bike rentals. The whole process to from membership to rental and return back has become automatic. Given below is the description of the data which is a (17379,17) shaped data, The variables are

The table represent the features used in the training and analysis:

Features	Description
rec_id	Daily customer index
datetime	The date index for both years
season	Season type (1-winter, 2-spring, 3- summer, 4-fall)
year	The year (0-2011, 1-2012)
month	The months (1-12)
Is_holiday	0 – not holiday, 1-holiday
weekday	Weekdays 0(Monday) – 6(Sunday)
Is_workingday	0- not a working day, 1- workingday
weather	Weather type(1-Clear, 2- Cloudy, 3- Rian, 4-Storm
temp	Normalized value of temperatures at every rec_id
atemp	Normalized value of the absolute temperature
humidity	Contains the normalized value for the humidity
windspeed	Contains the normalized value for the windspeed
casual	Has the number of unregistered users at a given day
registered	Has the number of registered users
Total_count	Total rentals with both casual and registered users

Algorithms

- **Feature Engineering**

The provided data in Its raw form wasn't directly use as an input to the model serval future engineering was carried out where a few features we modified few were dropped

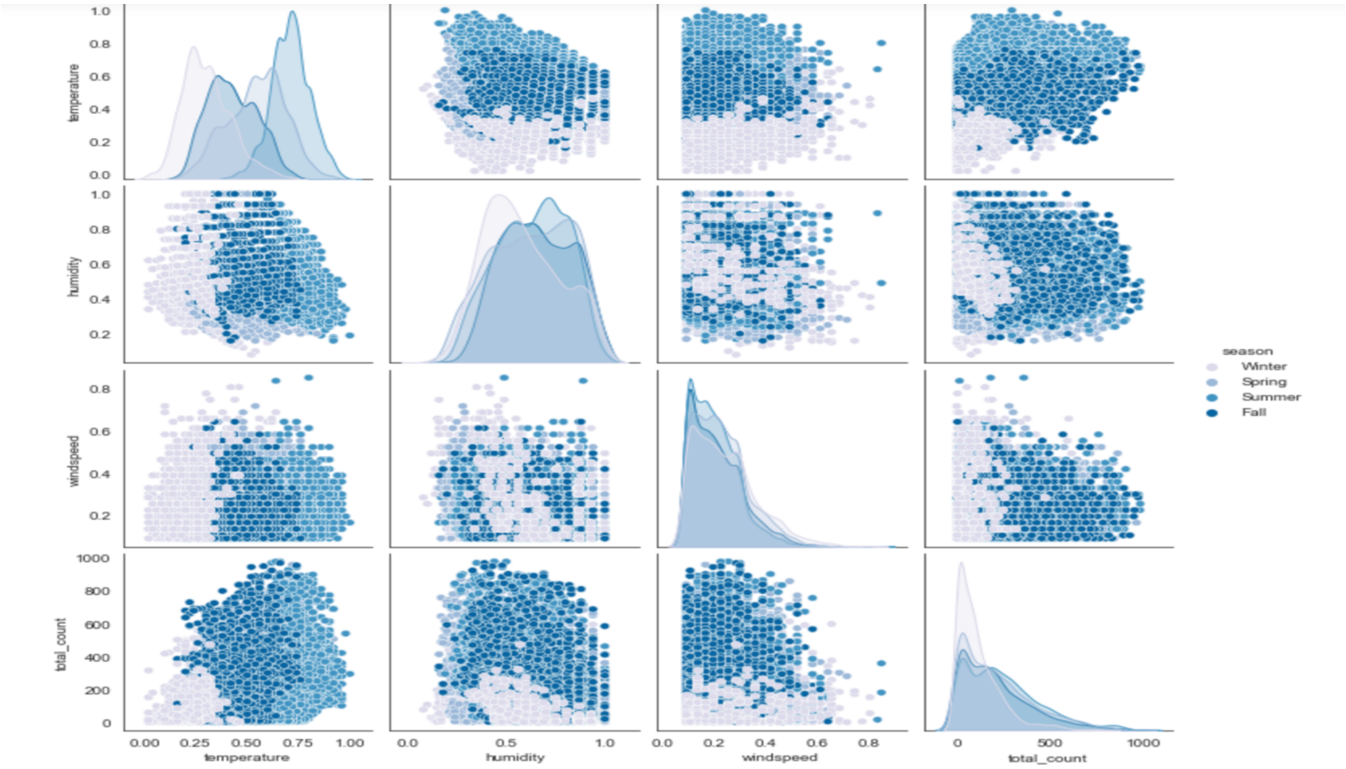
You can observe in the dataset that some of the variables like 'weathersit' and 'season' have values as 1, 2, 3, 4 which have specific labels associated with them (as can be seen in the data dictionary). These numeric values associated with the labels may indicate that there is some order to them - which is actually not the case (Check the data dictionary and think why). So, it is advisable to convert such feature values into categorical string values before proceeding with model building. Please refer the data dictionary to get a better understanding of all the independent variables. You might notice the column 'yr' with two values 0 and 1 indicating the years 2012 and 2012 respectively. At the first instinct, you might think it is a good idea to drop this column as it only has two values so it might not be a value-add to the model. But in reality, since these bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing every year proving that the column 'yr' might be a good variable for prediction. So think twice before dropping it.

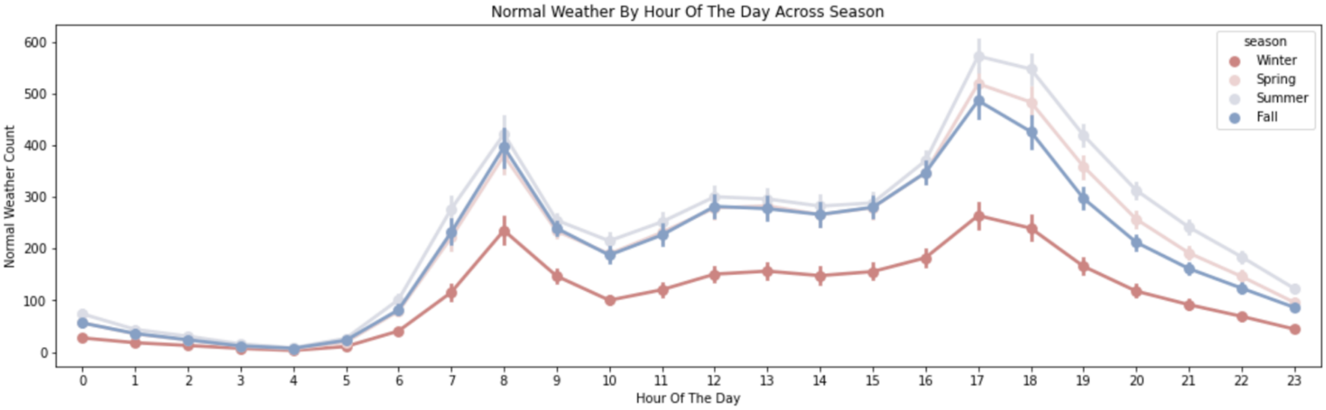
- **Models**
  linear regression , polynomial , Ridge regression , lasso regression
- **Model Evaluation and selection**
  The entire training dataset of 17379 was split 25 test/75
  The official metric for Bike Sharing Data was regression rate (accuracy); however, class count were included to improve performance against F1score and provide a more useful real-world application where regression of the minority calss (functional needs repair) would be essential.
  - Final linear regression 5-fold CV scores: 34 features(8 numaric) with class cout ACCURCY(0.69)

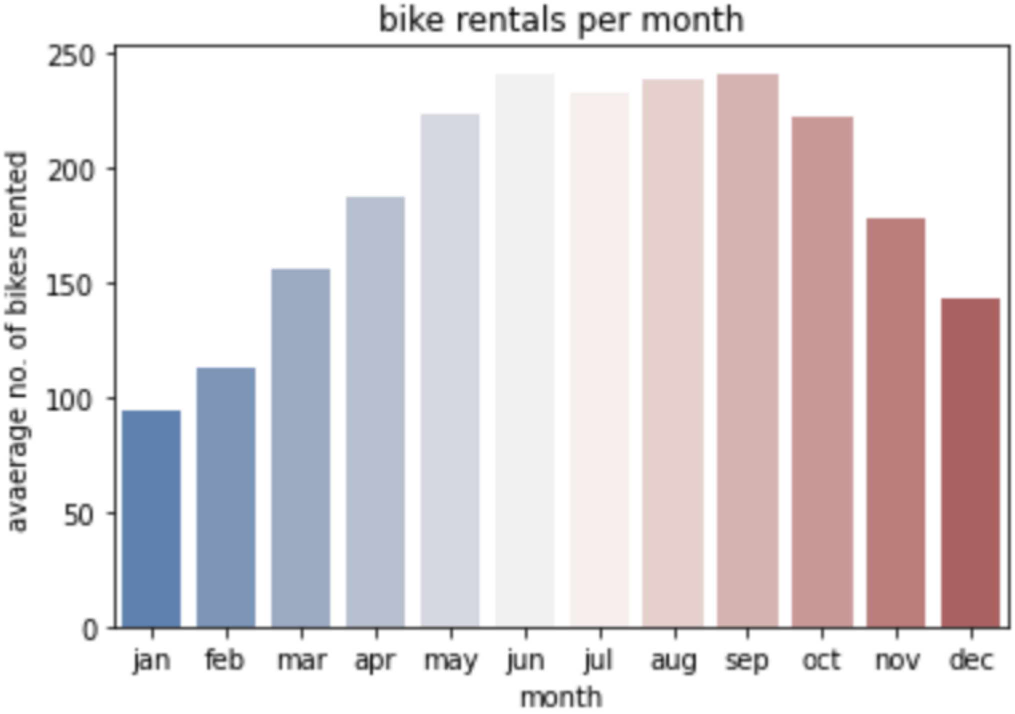
Tools

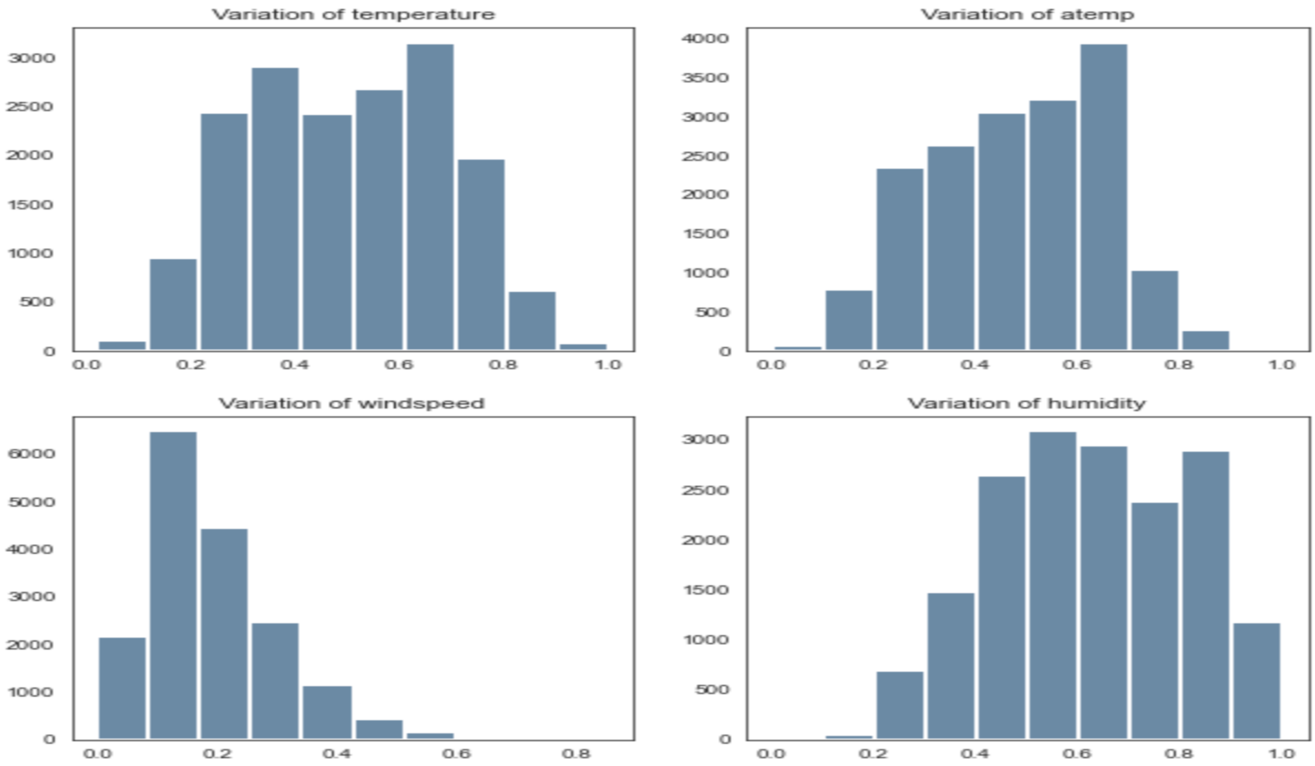
- Numpay and pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and seaborn for plotting

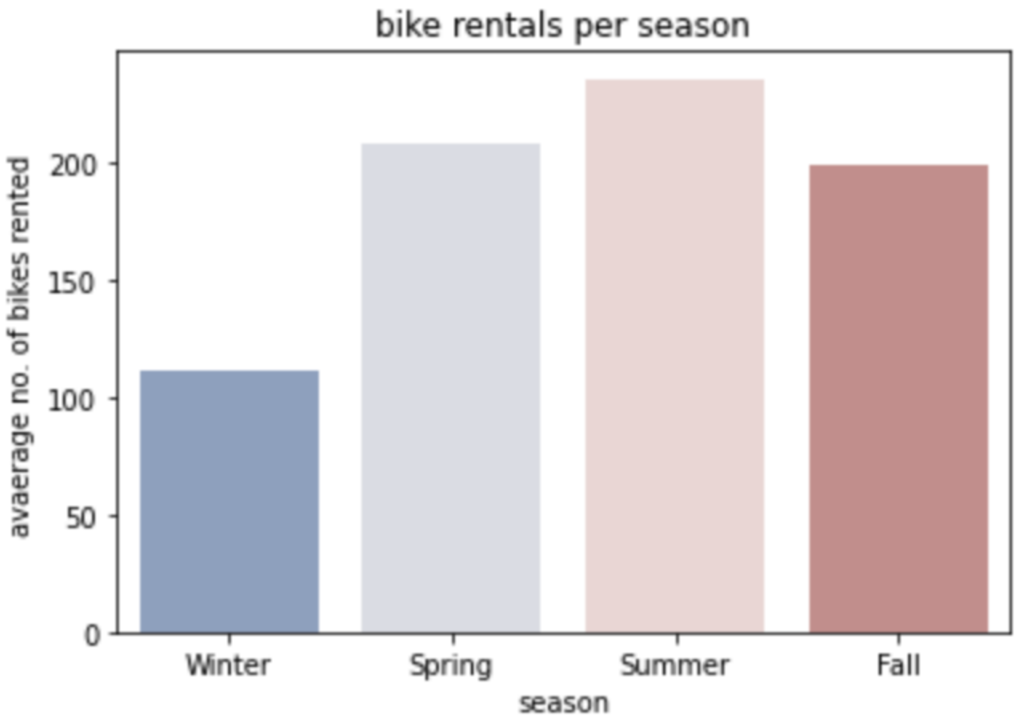
Communication

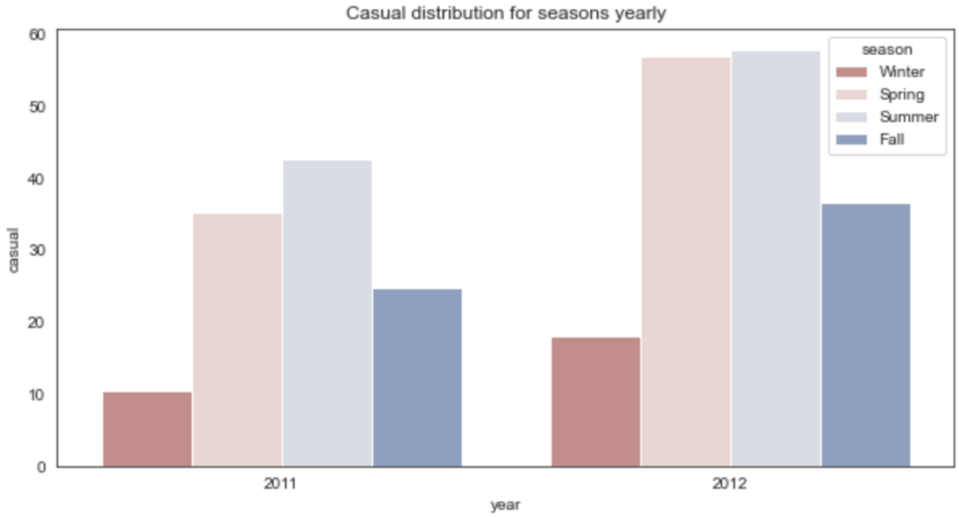


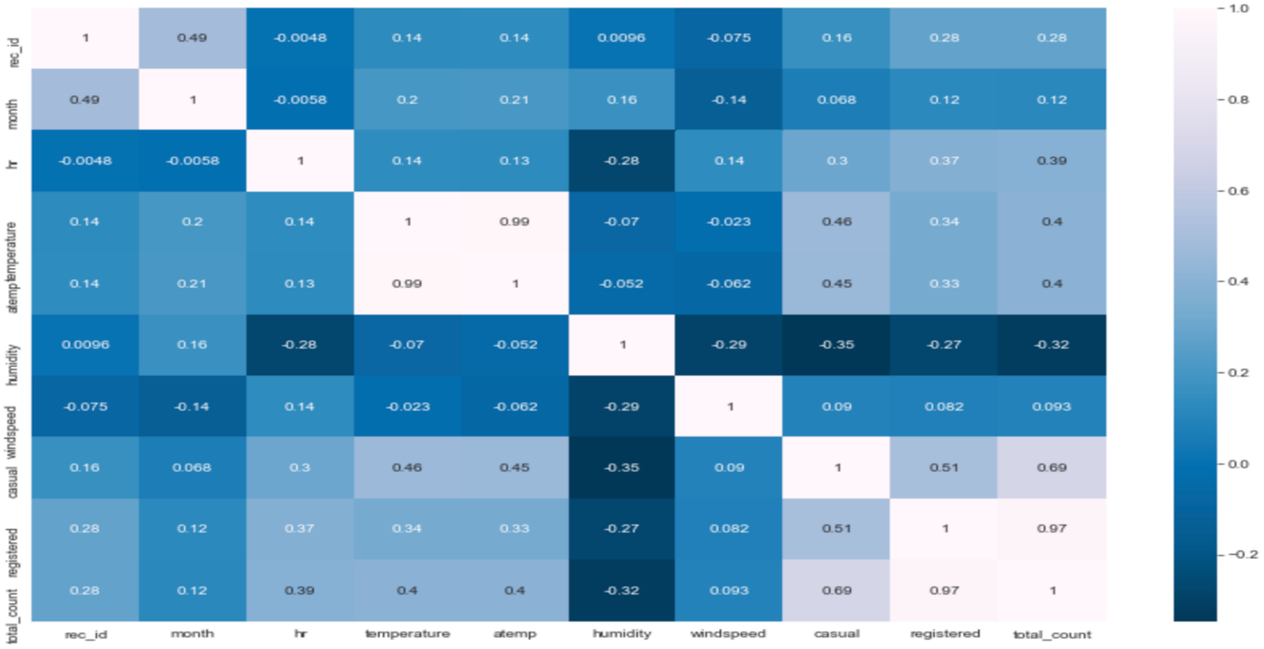

Normal Weather By Hour Of The Day Across Season
Normal Weather Count
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour Of The Day
season
Winter
Spring
Summer
Fall

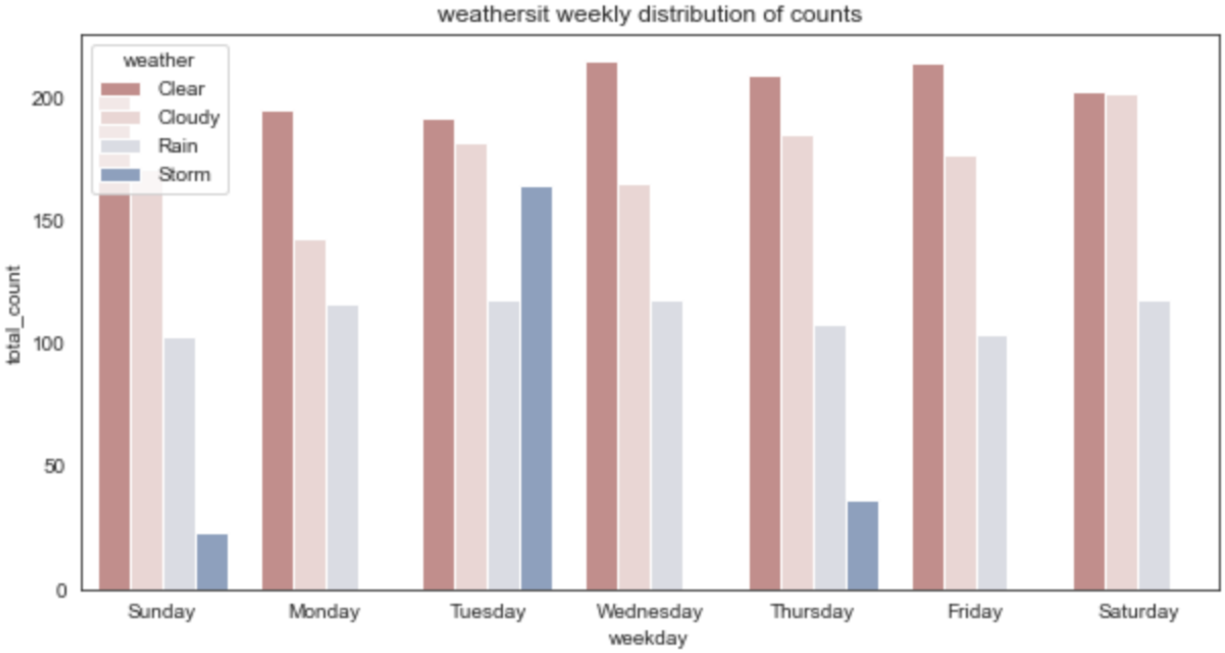

bike rentals per month
avaerage no. of bikes rented
250
200
150
100
50
0
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month


Variation of temperature
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
Variation of atemp
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.2
0.4
0.6
0.8
1.0
Variation of windspeed
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
Variation of humidity
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0


bike rentals per season
avaerage no. of bikes rented
0
50
100
150
200
Winter
Spring
Summer
Fall
season


Casual distribution for seasons yearly
season
Winter
Spring
Summer
Fall
casual
0
10
20
30
40
50
60
2011
2012
year


rec_id 1 0.49 -0.0048 0.14 0.14 0.0096 -0.075 0.16 0.28 0.28
month 0.49 1 -0.0058 0.2 0.21 0.16 -0.14 0.068 0.12 0.12
hr -0.0048 -0.0058 1 0.14 0.13 -0.28 0.14 0.3 0.37 0.39
temperature 0.14 0.2 0.14 1 0.99 -0.07 -0.023 0.46 0.34 0.4
atemp 0.14 0.21 0.13 0.99 1 -0.052 -0.062 0.45 0.33 0.4
humidity 0.0096 0.16 -0.28 -0.07 -0.052 1 -0.29 -0.35 -0.27 -0.32
windspeed -0.075 -0.14 0.14 -0.023 -0.062 -0.29 1 0.09 0.082 0.093
casual 0.16 0.068 0.3 0.46 0.45 -0.35 0.09 1 0.51 0.69
registered 0.28 0.12 0.37 0.34 0.33 -0.27 0.082 0.51 1 0.97
total_count 0.28 0.12 0.39 0.4 0.4 -0.32 0.093 0.69 0.97 1
rec_id
month
hr
temperature
atemp
humidity
windspeed
casual
registered
total_count
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


weathersit weekly distribution of counts
weather
Clear
Cloudy
Rain
Storm
total_count
0
50
100
150
200
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
weekday