Statistical Data Analysis on Bike Rental Dataset

Team Members

- Akshay Kumar S20170010006
- Nived Gupta S20170010104
- Venkata Sri Mukesh Inturi S20170010056
- Shivam Gupta S20170010149

Dataset

- Bike Rental
- 17,379 Rows
- 17 Columns

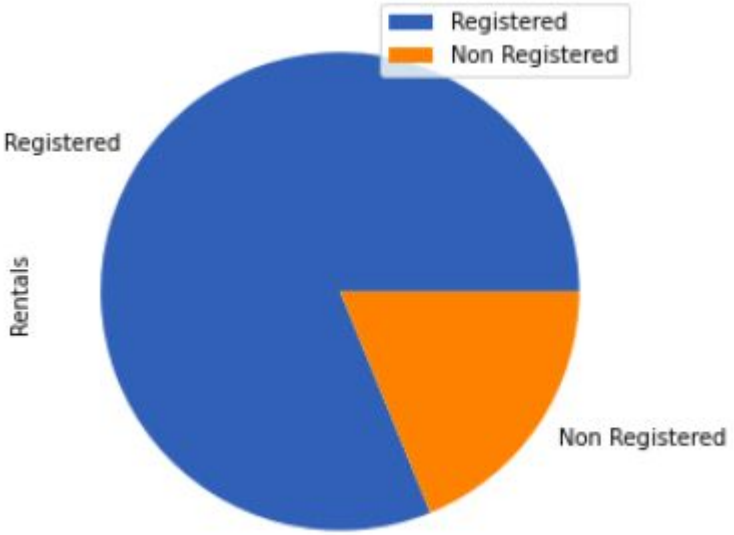
	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1
5	6	2011-01-01	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
6	7	2011-01-01	1	0	1	6	0	6	0	1	0.22	0.2727	0.80	0.0000	2	0	2
7	8	2011-01-01	1	0	1	7	0	6	0	1	0.20	0.2576	0.86	0.0000	1	2	3
8	9	2011-01-01	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0.0000	1	7	8
9	10	2011-01-01	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0.0000	8	6	14
10	11	2011-01-01	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
11	12	2011-01-01	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56
12	13	2011-01-01	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84
13	14	2011-01-01	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94
14	15	2011-01-01	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106
15	16	2011-01-01	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110
16	17	2011-01-01	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93
17	18	2011-01-01	1	0	1	17	0	6	0	2	0.44	0.4394	0.82	0.2836	15	52	67

Details about some features

	temp	atemp	hum	windspeed
count	17379.000000	17379.000000	17379.000000	17379.000000
mean	20.376474	23.788484	62.722884	12.736233
std	7.894801	8.592587	19.292983	8.196891
min	0.820000	0.000000	0.000000	0.000000
25%	13.940000	16.660000	48.000000	7.000000
50%	20.500000	24.240000	63.000000	13.000000
75%	27.060000	31.060000	78.000000	17.000000
max	41.000000	50.000000	100.000000	57.000000

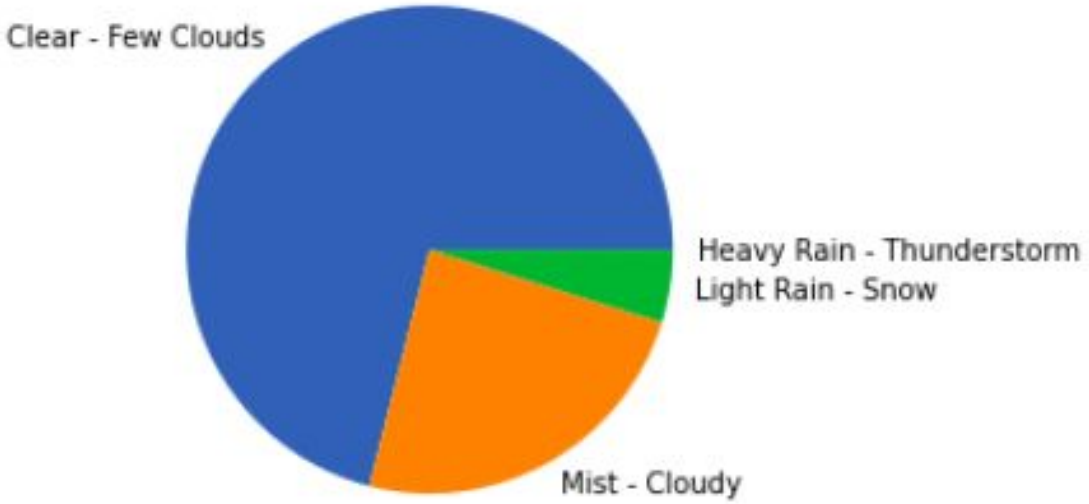
Registered vs Non Registered Users

- Registered 81.17 %
- Non Registered 18.83 %



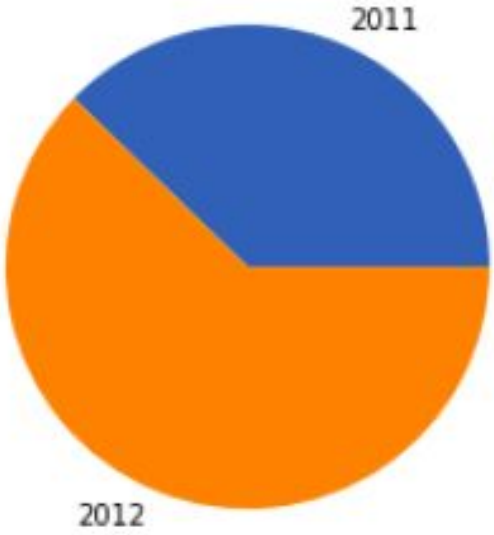
Rentals Depending on Weather

Clear - Few Clouds	2338173
Mist - Cloudy	795952
Light Rain - Snow	158331
Heavy Rain - Thunderstorm	223

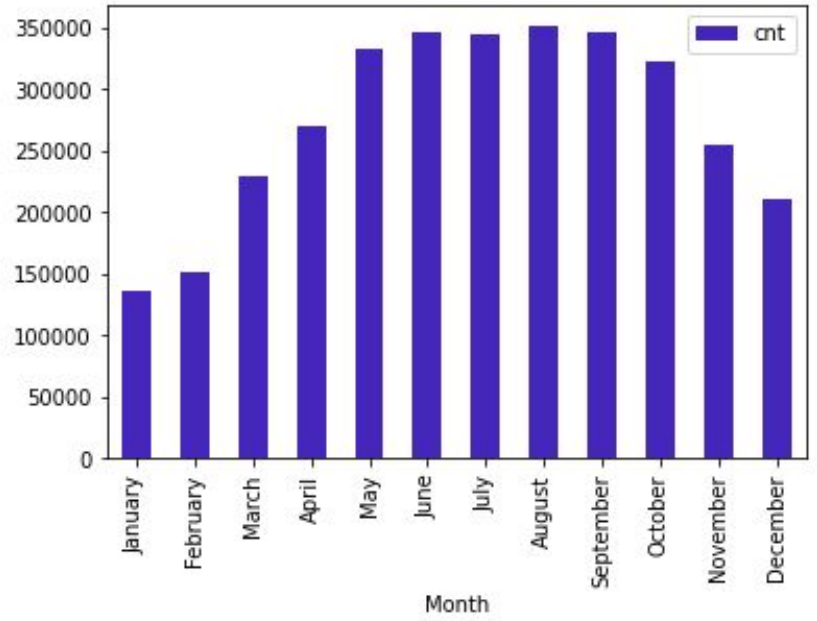


Year wise rentals

Biking rental was up by 64.88 % from 2011 to 2012

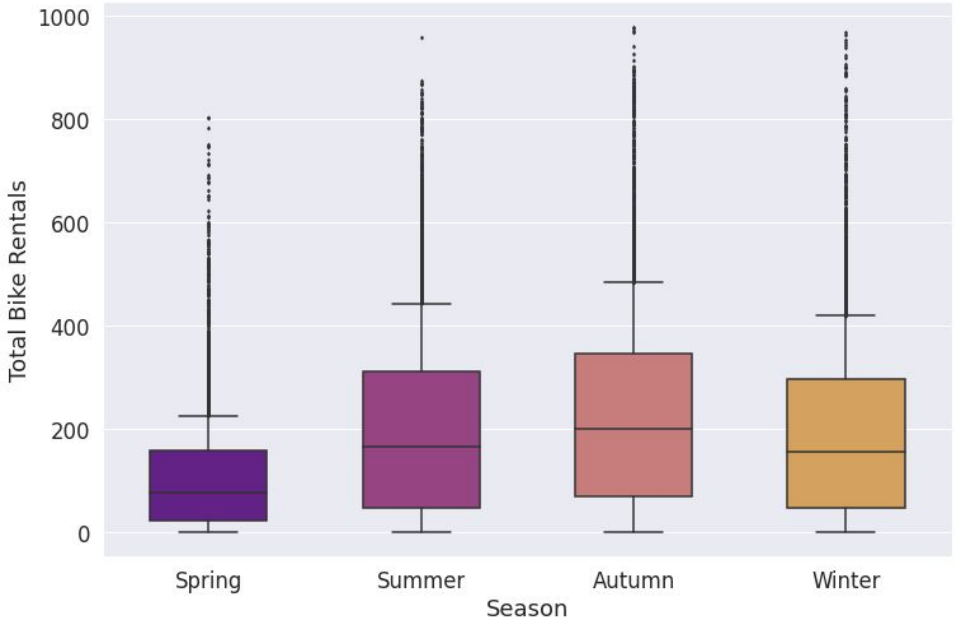


Month Wise Total Rentals



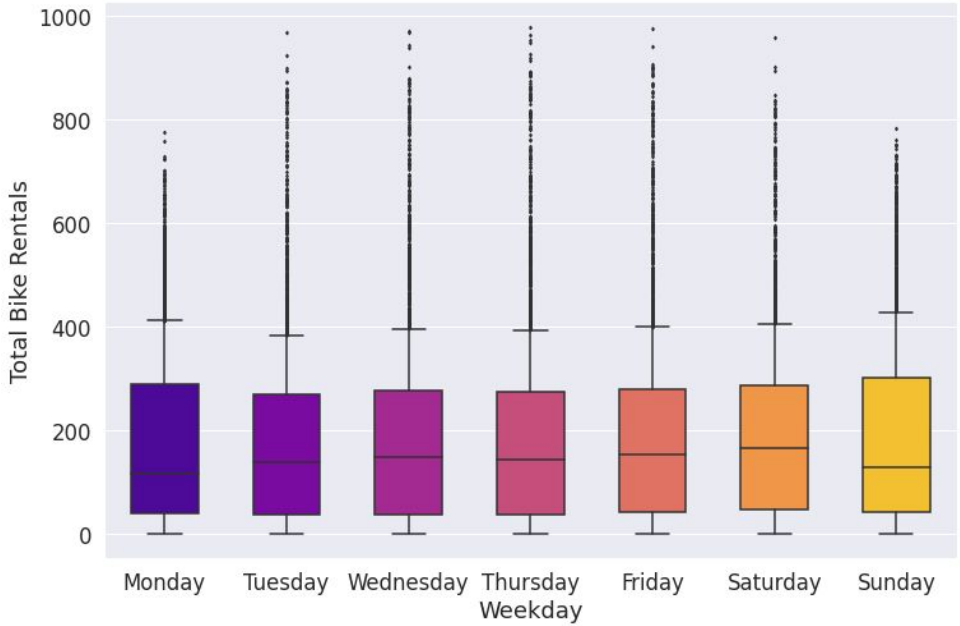
January	134933
February	151352
March	228920
April	269094
May	331686
June	346342
July	344948
August	351194
September	345991
October	322352
November	254831
December	211036

Boxplot: Season-wise bike rentals



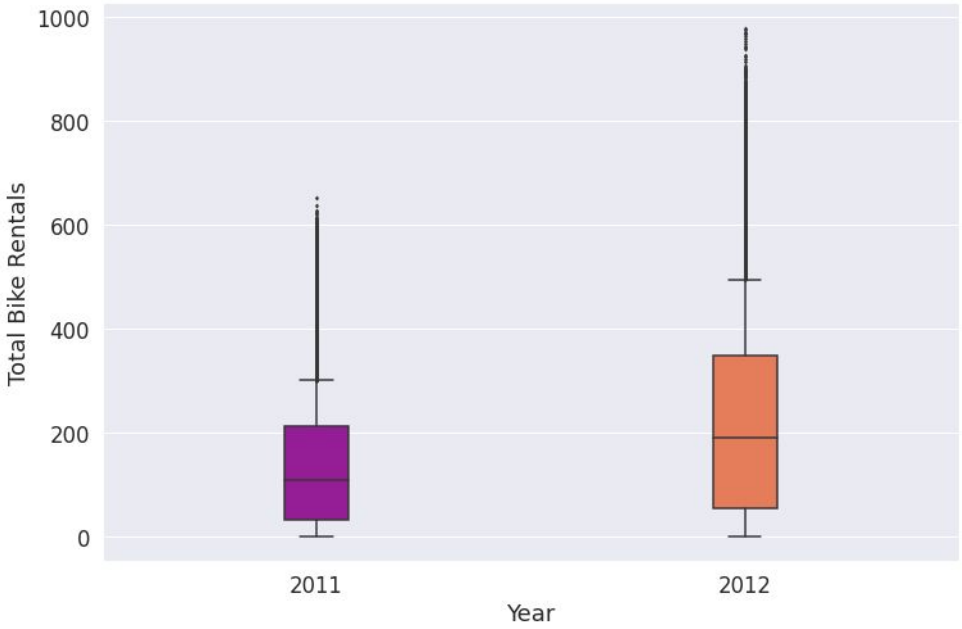
Summer and Autumn are the most preferred seasons for bike rentals.

Boxplot: Weekday-wise bike rentals



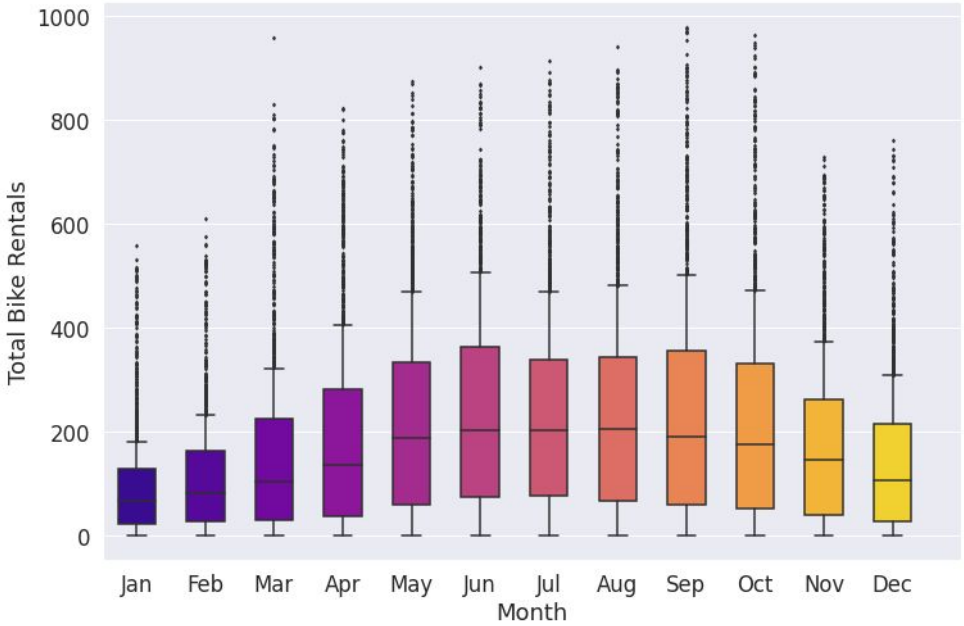
We can see that saturday has highest median of bike rentals followed by wednesday.

Boxplot: 2011 vs 2012



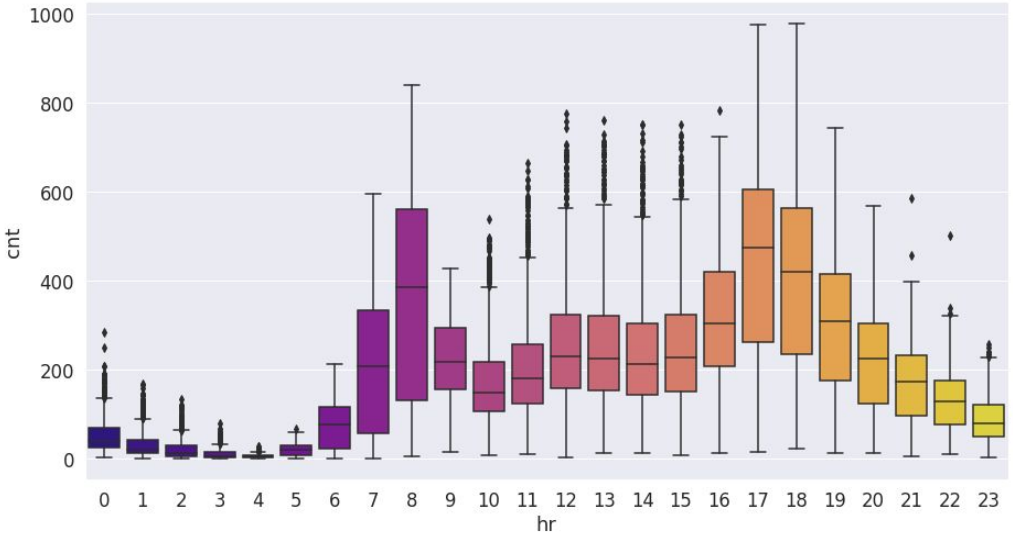
2012 has almost the double the number of rentals compared to 2011.

Boxplot: Rentals on a monthly basis



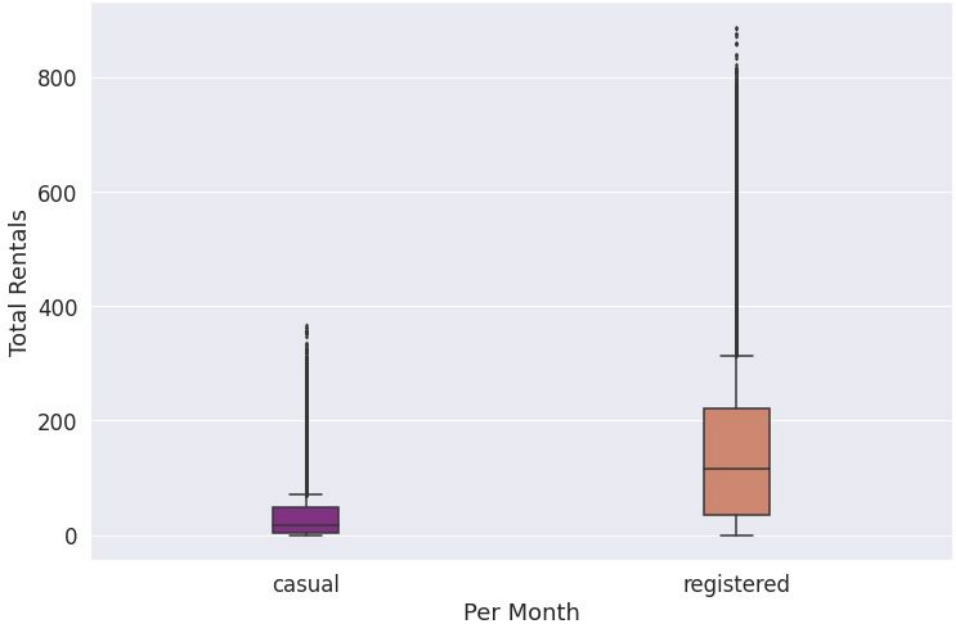
June, July and August have higher median number of bike rentals than other months.

Boxplot: Rentals on an hourly basis



Most of the bikes are rented at 8 in the morning and 5-6 in the evening.

Boxplot: Casual vs Registered users

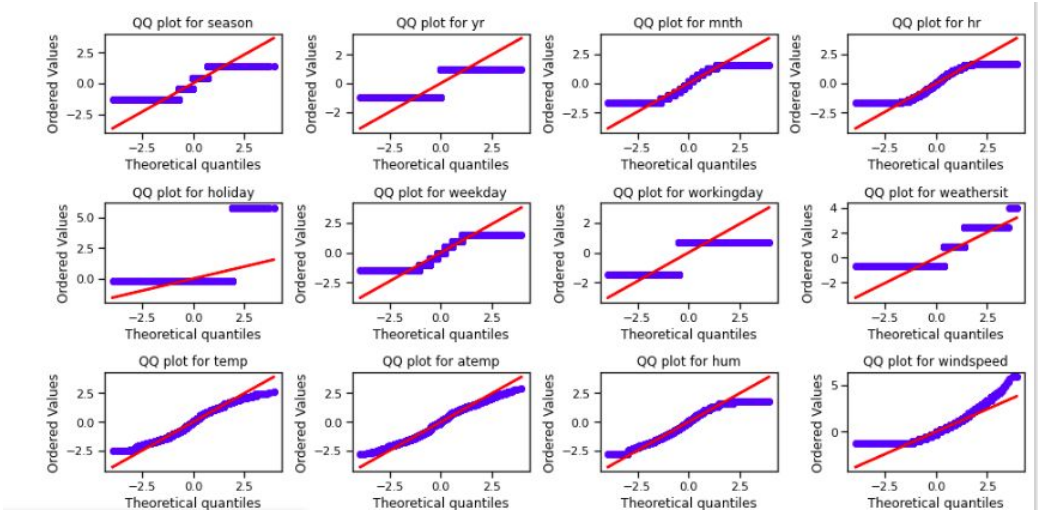


Has good customer base.

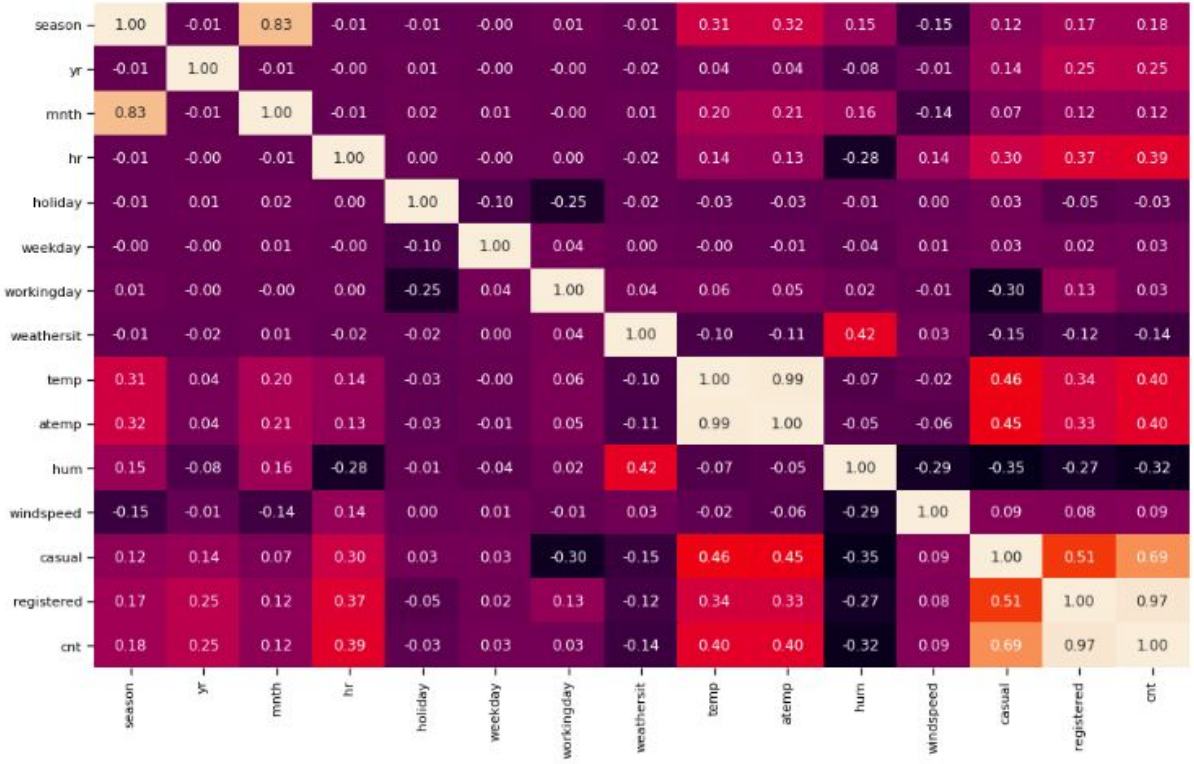
The box plot is taken per month.

Univariate Analysis QQ Plots

For a distribution to be normal QQ Plot is straight line



Correlation Matrix



Model building and evaluation

- We build model using our linear regression equation $\theta = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$
- We will predict the value for the target variable by using our model parameter for the test data set.
- Then compare the predicted value with the actual value in the test set. We compute Mean Square Error using formula $\mathbf{J}(\theta) = \frac{1}{\mathbf{m}} \sum_{\mathbf{i=1}}^{\mathbf{m}} (\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_\mathbf{i})^\mathbf{2}$
- R^2 is a statistical measure of how close data are to the fitted regression line.
- $$\mathbf{R^2} = \mathbf{1} - \frac{\mathbf{SSE}}{\mathbf{SST}}$$

  SSE = Sum of Square Error $= \sum_{\mathbf{i=1}}^{\mathbf{m}} (\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_\mathbf{i})^\mathbf{2}$

  SST = Sum of Square Total $= \sum_{\mathbf{i=1}}^{\mathbf{m}} (\mathbf{y}_\mathbf{i} - \bar{\mathbf{y}}_\mathbf{i})^\mathbf{2}$

Regression Analysis

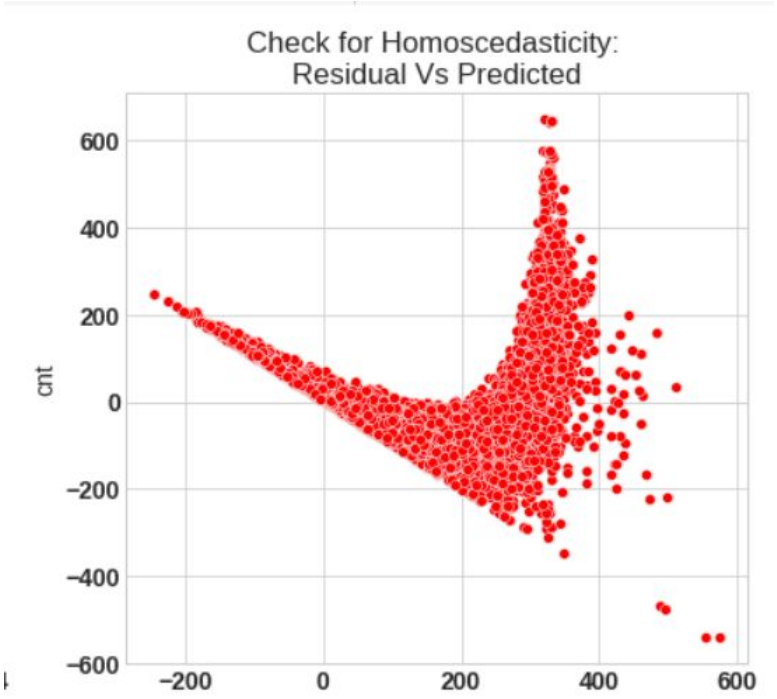
- The Mean Square Error(MSE) or J(theta) is: 15037.758709640962
- R square obtain for normal equation method is : 0.5403064650055478
- VIF = 2.1753623313672845

Test of Assumption

- Homoscedasticity
- Normality
- Autocorrelation
- Multicollinearity

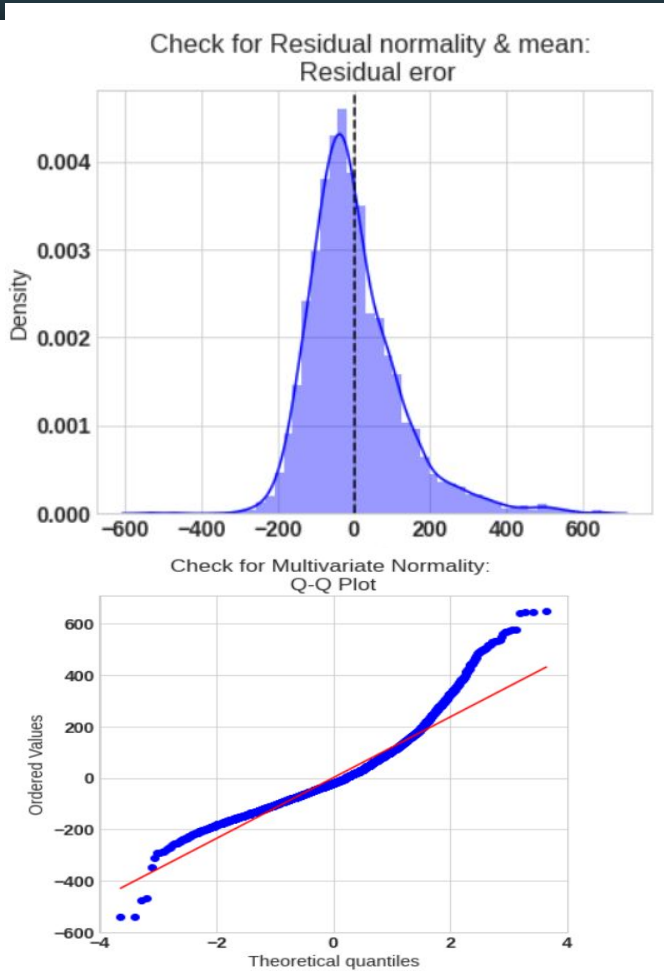
Homoscedasticity

- We are getting the funnel shape.
- Variance not constant
- Heteroscedastic



Normality

- The residual mean is zero
- Residual error plot is slightly right skewed
- Q-Q Plot not showing a straight line
- All variables are not multivariate normal



Autocorrelation

- Occurs when the residuals are not independent from each other
- For checking autocorrelation, used Durbin-Watson test
- This test gives value between 0 to 4
- If value between 1.5 and 2.5, no autocorrelation
-

Multicollinearity

- Multicollinearity represents the dependency among the features
- Calculated Variation Inflation Factor(VIF) for each and every dependent variable
- For a VIF value greater than 10, represents multicollinearity among the variables
- Variables can be dropped with VIF > 10
- VIF = 2.1753623313672845

Factor Analysis

- Factor Analysis is a dimensionality Reduction Technique
- Get Factors (unobserved variables) from the observed variables in data

  Tested Data with

  - Bartell's test of sphericity and Kaiser-Meyer-Olkin (KMO) Test

Bartell’s test of sphericity

- Bartlett’s test checks whether the correlation is present in the given data.
- Observed P value 0
- Observed correlation matrix is not an identity Matrix

Kaiser-Meyer-Olkin (KMO) Test

- Measures suitability of data for Factor Analysis
- KMO score is always between 0 to 1.
- Observed Value 0.594

Factor Analysis

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
SS Loadings	7.342814	3.354377	2.687979	0.968784	0.932514	0.816750	0.672255
Proportional Var	0.293713	0.134175	0.107519	0.038751	0.037301	0.032670	0.026890
Cumulative Var	0.293713	0.427888	0.535407	0.574158	0.611459	0.644129	0.671019

- In our case, the 7 factors together are able to explain 67.1% of the total variance.

Number of Factors

- Scree Plot
- 7 factors have eigenvalue > 1