**Predicting Bike Rental Counts**

Goldie Sahni

15 August 2018

Contents

1. Problem Definition & Data Distributions
   1. Problem Statement………………………………………………………………………………2
   2. Data………………………………………………………………………………………………………2

**Modelling in Python**

1. Exploratory Data Analysis
   1. Variable Distributions………………………………………………………………………….5
   2. Outlier Analysis……………………………………………………………………………………7
   3. Missing Value Analysis…………………………………………………………………………8
   4. Correlation Analysis…………………………………………………………………………….9
2. Modelling
   1. Linear Regression……………………………………………………………………………….10
   2. Decision Tree………………………………………………………………………………………11
   3. Random Forest……………………………………………………………………………………11
   4. KNN……………………………………………………………………………………………………..11
   5. Gradient Boosted Trees………………………………………………………………………11

**Modelling in R**

1. Exploratory Data Analysis
   1. Variable Distributions………………………………………………………………………..13
   2. Outlier Analysis………………………………………………………………………………….15
   3. Missing Value Analysis……………………………………………………………………….16
   4. Correlation Analysis…………………………………………………………………………..16
2. Modelling
   1. Linear Regression………………………………………………………………………………21
   2. Decision Tree……………………………………………………………………………………..22
   3. Random Forest…………………………………………………………………………………..22
   4. KNN……………………………………………………………………………………………………..22
   5. Gradient Boosted Trees……………………………………………………………………..22
3. Conclusion…………………………………………………………………………………………………….24

Chapter 1

Problem Definition & Data Distributions

* 1. Problem Statement

The problem is how to predict the daily total count (casual + registered) of bikes that will be rented taking into consideration the historical data, day of week, weather and temperature on the particular day. We can ask several questions in this problem definition as under:

1. Do we have a pattern in past historical data that can be used to predict the total count of bike rentals?
2. Does count of bike rentals depend on the day of week?
3. Does holiday have any effect on the count of bike rentals?
4. Does season have any effect on the count of bike rentals?
5. Does weather have any effect on the count of bike rentals?
6. Does temperature have any effect on the count of bike rentals?
7. Does humidity have any effect on the count of bike rentals?
8. Does wind speed have any effect on the count of bike rentals?
   1. Data

The data given has 15 variables (not considering instant). A sample is shown below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| dteday | season | yr | mnth | holiday | weekday | workingday | weathersit | temp | atemp |
| 1/1/2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 | 0.344167 | 0.363625 |
| 1/2/2011 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0.363478 | 0.353739 |
| 1/3/2011 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0.196364 | 0.189405 |
| 1/4/2011 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 0.2 | 0.212122 |
| 1/5/2011 | 1 | 0 | 1 | 0 | 3 | 1 | 1 | 0.226957 | 0.22927 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| hum | windspeed | casual | registered | cnt |
| 0.805833 | 0.160446 | 331 | 654 | 985 |
| 0.696087 | 0.248539 | 131 | 670 | 801 |
| 0.437273 | 0.248309 | 120 | 1229 | 1349 |
| 0.590435 | 0.160296 | 108 | 1454 | 1562 |
| 0.436957 | 0.1869 | 82 | 1518 | 1600 |

Variable description is as under:

1. **dteday**: Date
2. **season**: Season (1:springer, 2:summer, 3:fall, 4:winter)
3. **yr**: Year (0: 2011, 1:2012)
4. **mnth**: Month (1 to 12)
5. **holiday**: whether day is holiday or not (extracted from Holiday Schedule)
6. **weekday**: Day of the week
7. **workingday**: If day is neither weekend nor holiday is 1, otherwise is 0.
8. **weathersit**: (extracted from Freemeteo)

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

1. **temp**: Normalized temperature in Celsius
2. **atemp**: Normalized feeling temperature in Celsius
3. **hum**: Normalized humidity
4. **windspeed**: Normalized wind speed
5. **casual**: count of casual users
6. **registered**: count of registered users
7. **cnt**: count of total rental bikes including both casual and registered

There are 3 dependent variables here – casual, registered and cnt. Casual and registered variables will be dropped since cnt is present which can be used as dependent variable.

All other variables are independent variables.

This is a regression problem since dependent variable is continuous.

We will use these machine learning algorithms to see which algorithm performs best out of these for regression in this problem:

1. Linear Regression
2. Decision Tree
3. Random Forest
4. K Nearest Neighbors
5. Gradient Boosted Trees

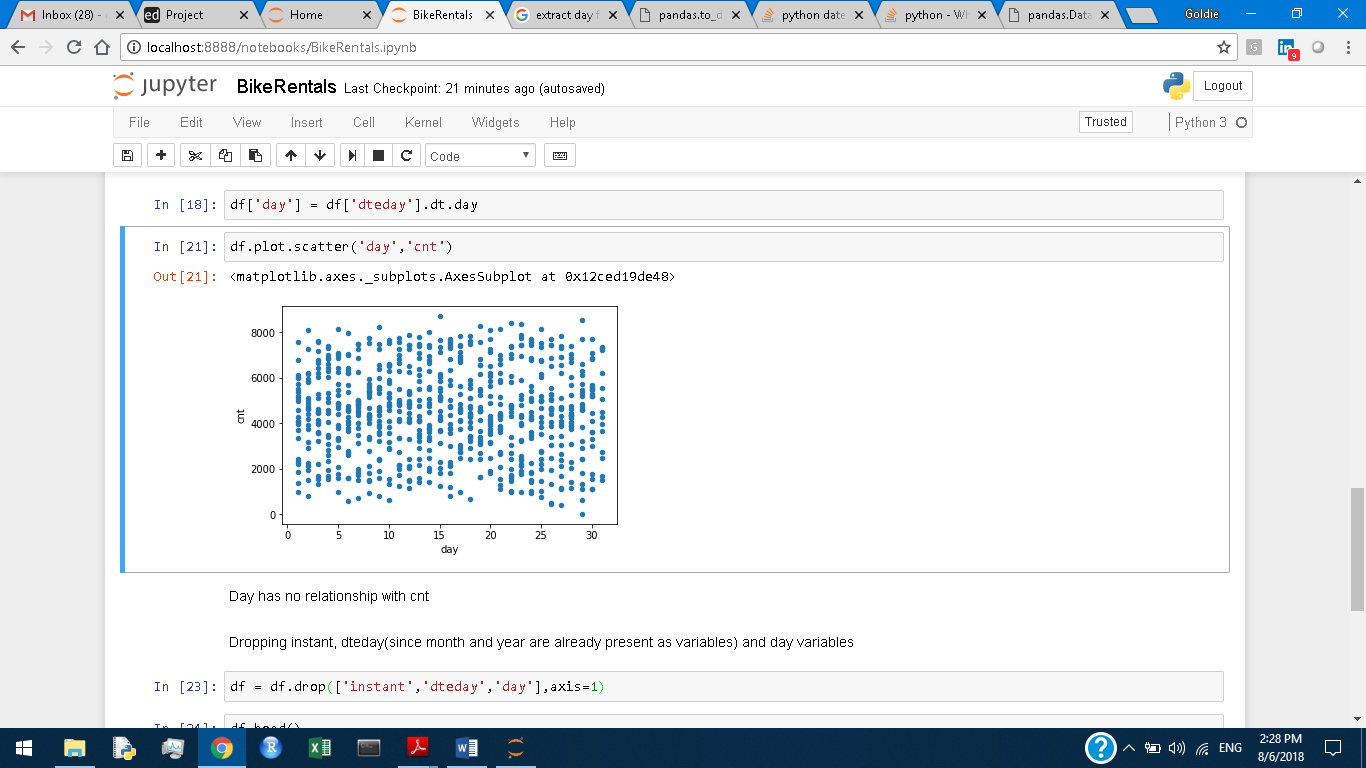
Modelling in Python

Chapter 2

Exploratory Data Analysis

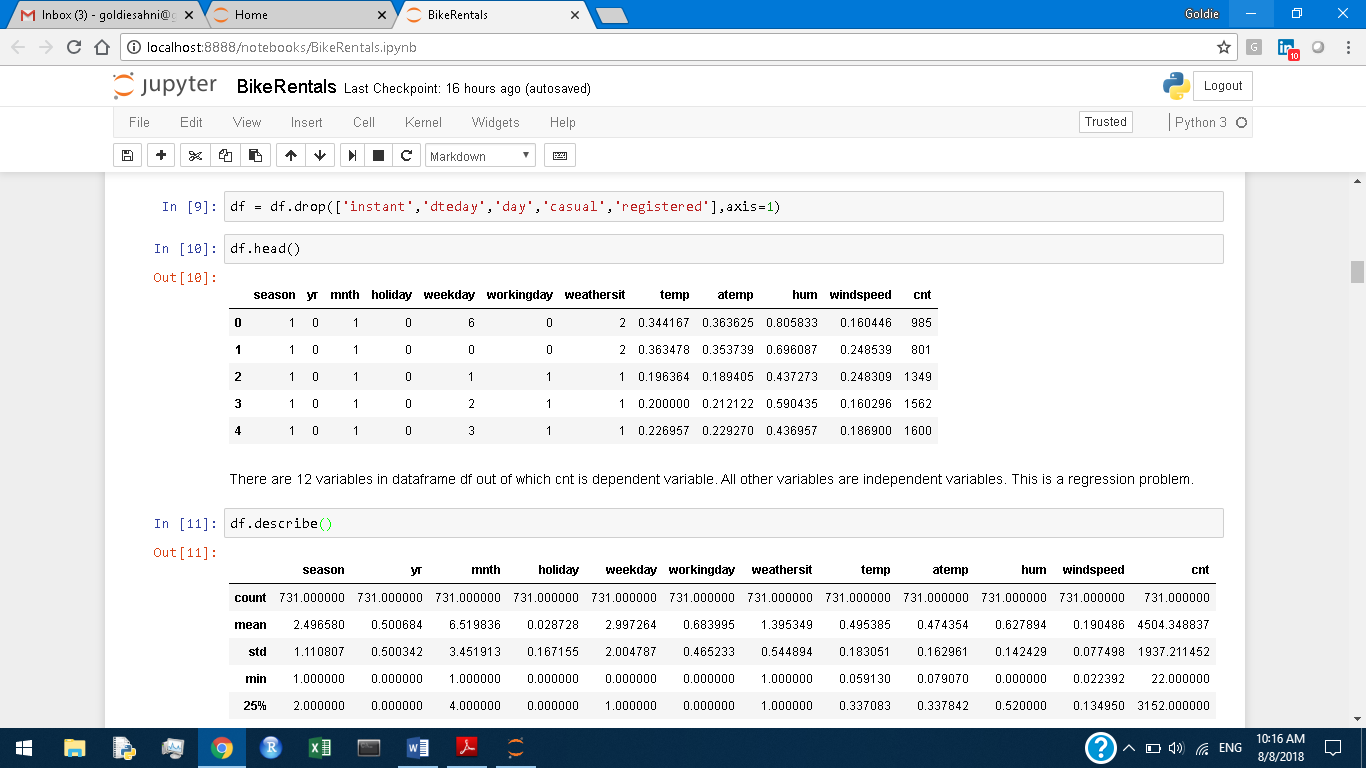
**Day Relationship with cnt**

Day was extracted from dteday variable to check if day has any relationship with cnt. But it was found that no relationship exists between day and cnt as shown by scatter plot of day vs cnt.



So, we are dropping instant, dteday (mnth and yr are present), day, casual & registered from the dataframe df (data has been read as df).

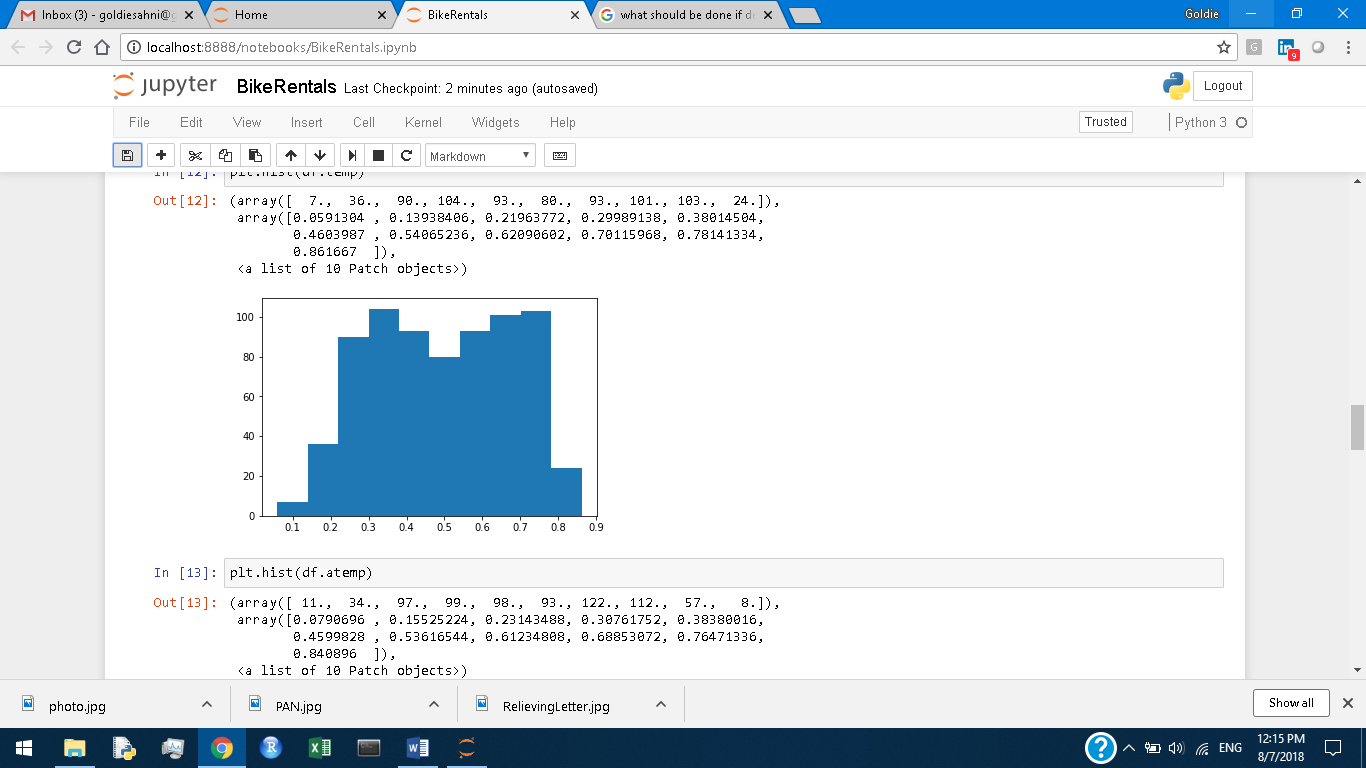
Now df has 12 variables.



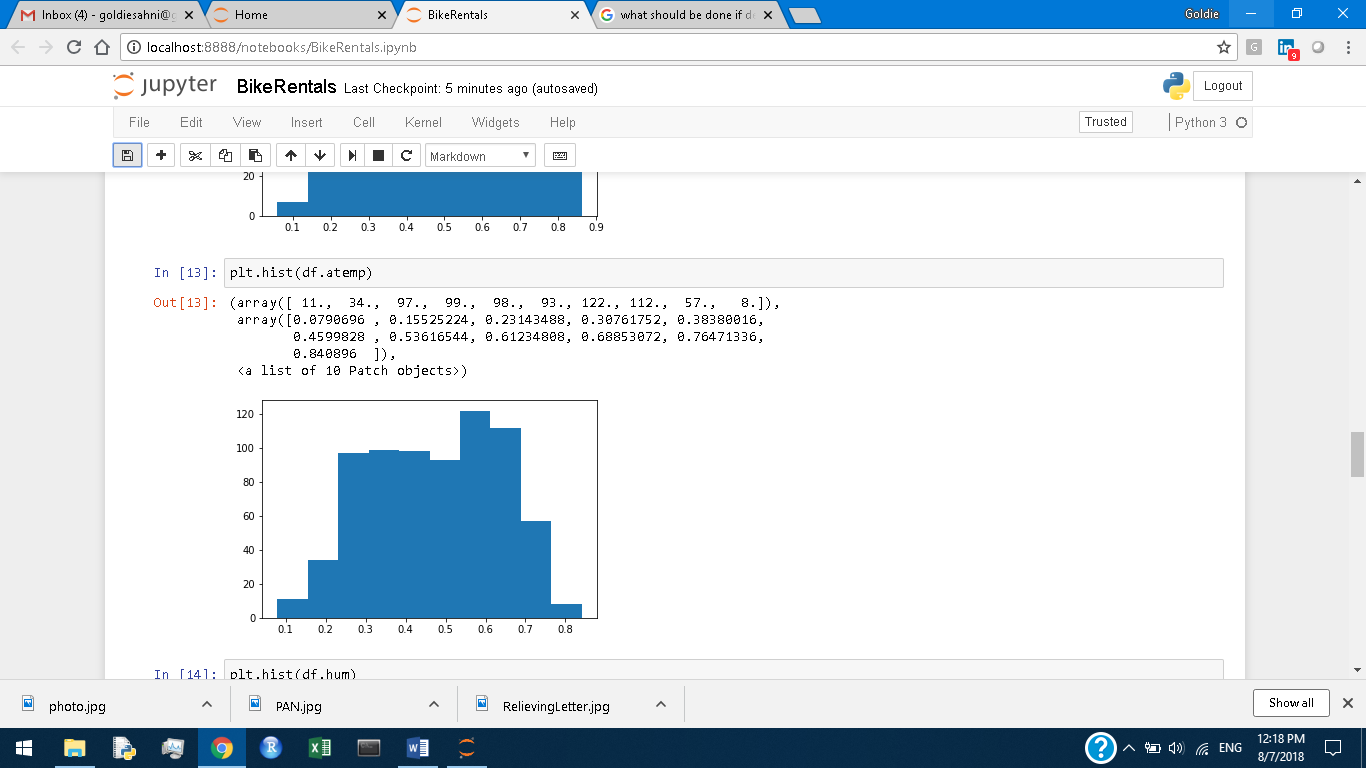
**2.1 Variables distributions**

Variables - season, yr, mnth, holiday, weekday, workingday, weathersit - are categorical so there is no need of distribution visualization.

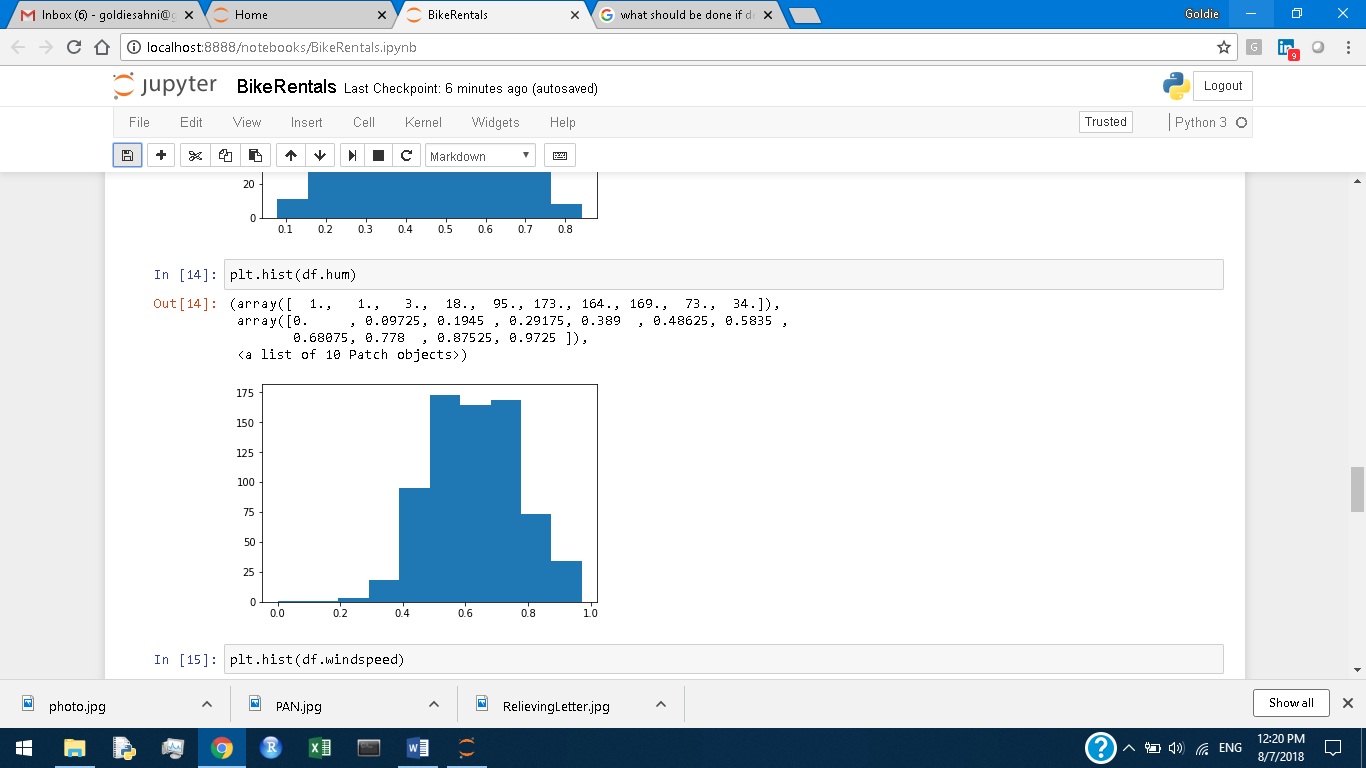
1. Variable temp distribution



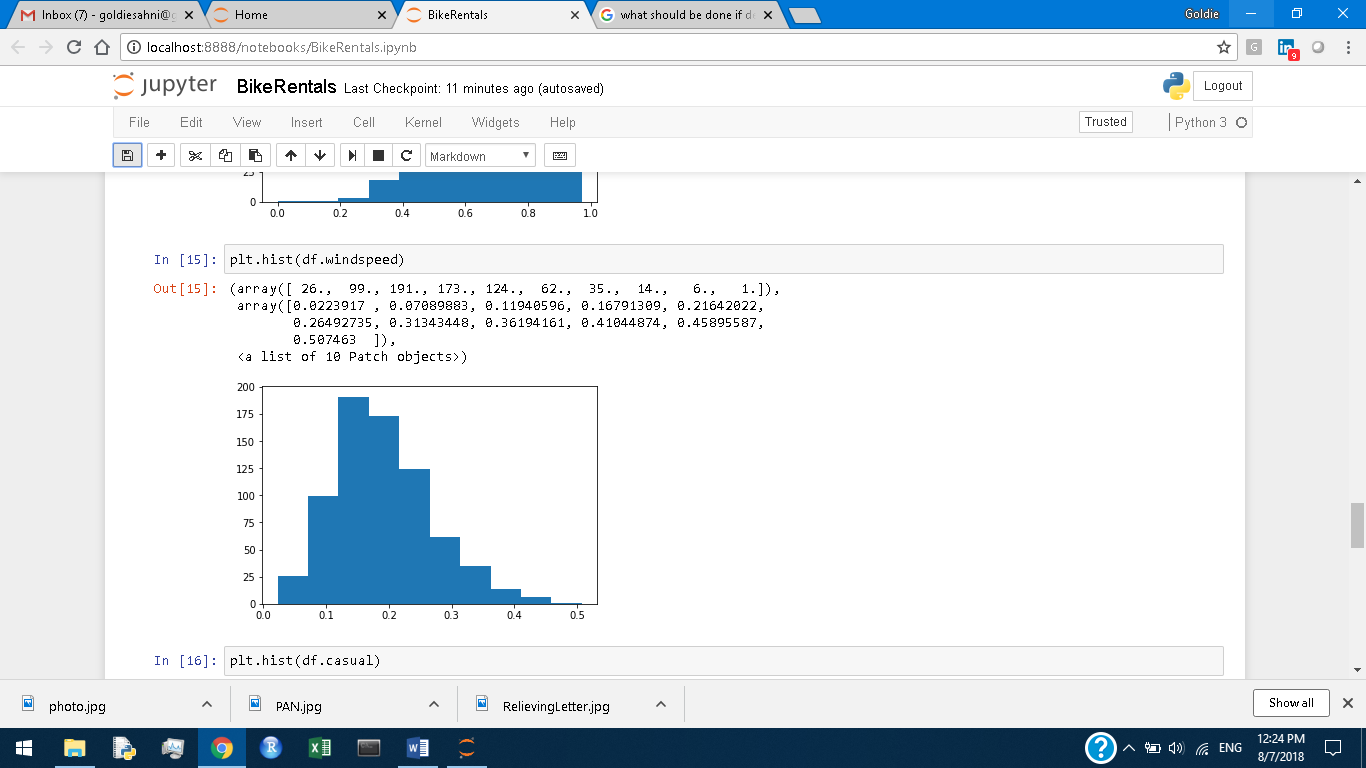
1. Variable atemp distribution



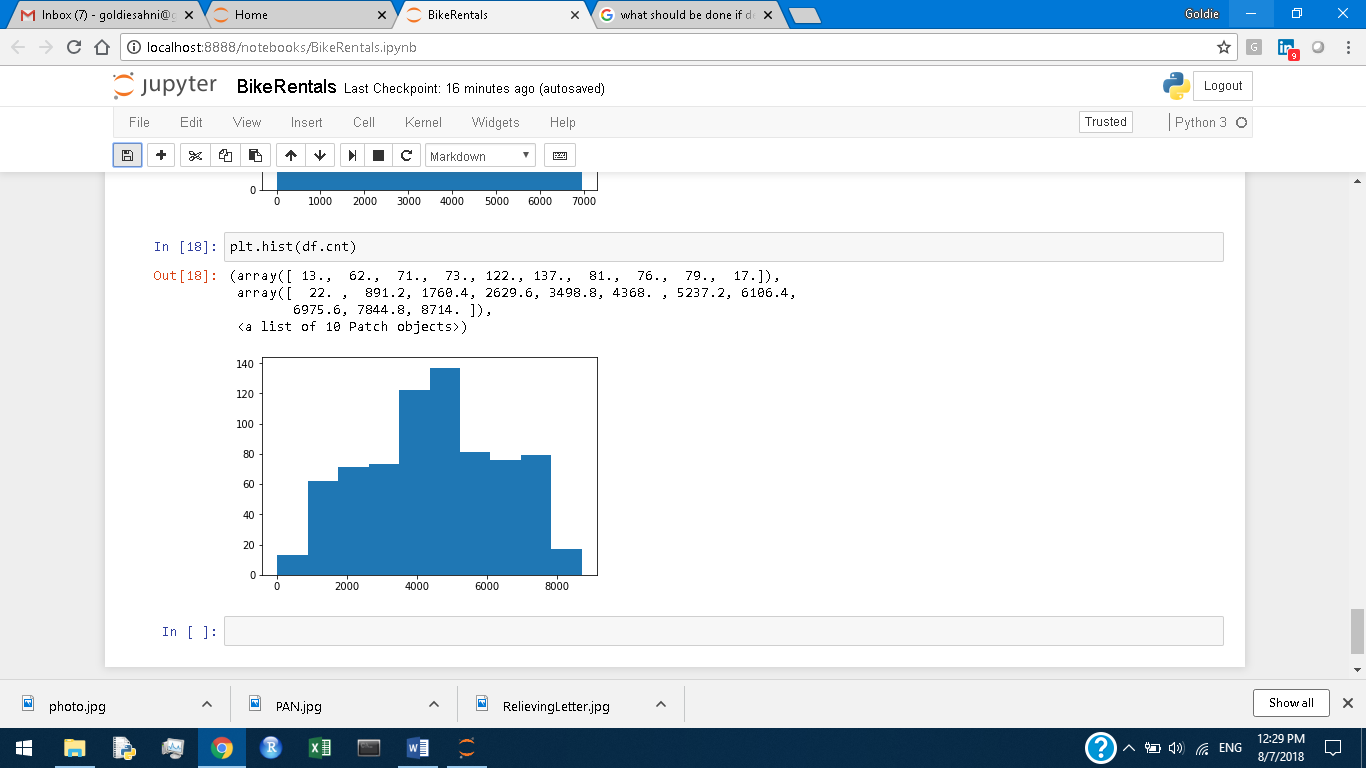
1. Variable hum distribution



1. Variable windspeed distribution



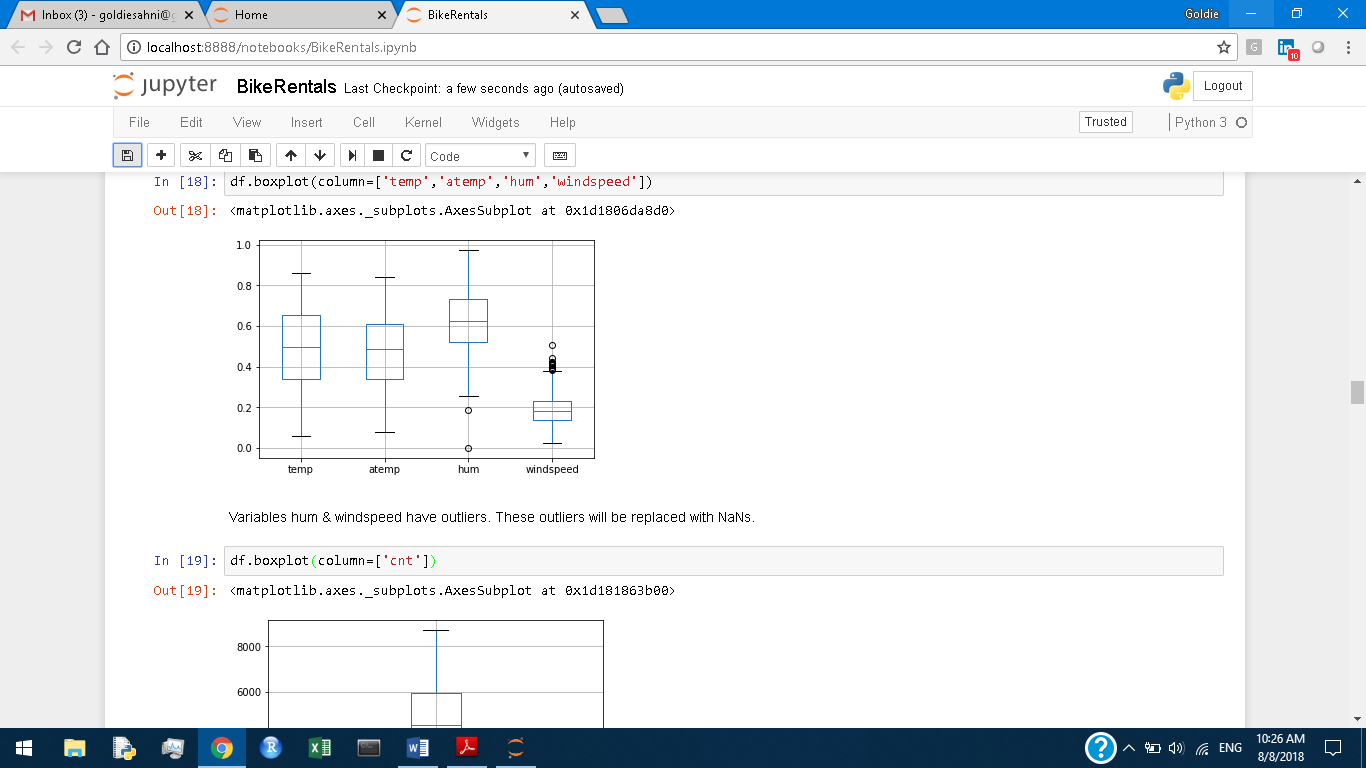
1. Variable cnt distribution



Variables registered & cnt are close to normal. All other continuous variables are skewed.

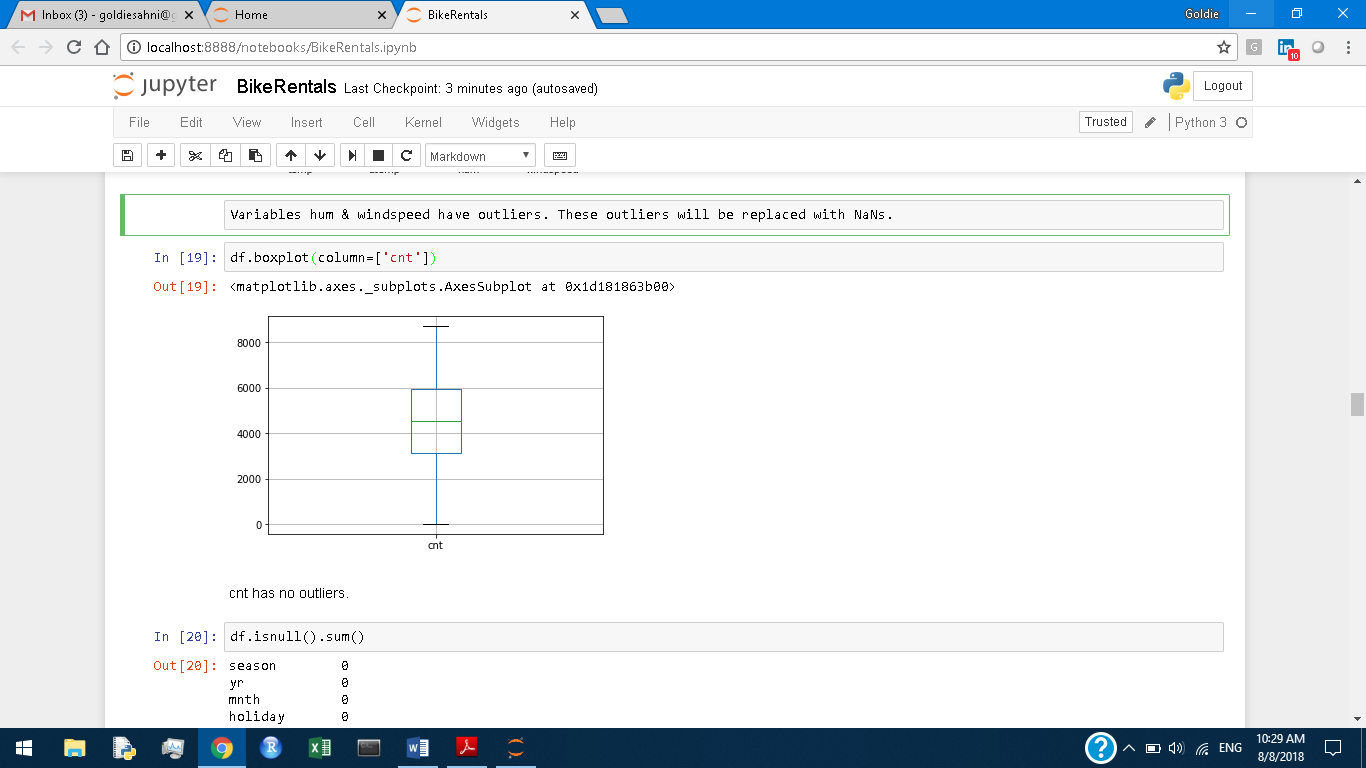
**2.2 Outliers Analysis**

Boxplots for variables – temp, atemp, hum, windspeed



Variables hum & windspeed have outliers. These outliers will be replaced with NaNs.

Boxplot for variable cnt



Variable cnt has no outliers.

Variable hum has got 2 null values and variable windspeed has got 13 null values after outliers replacement with NaNs.

**2.3 Missing Values Analysis**

There are several methods for missing values replacement as under:

1. Mean Substitution – Missing values for a variable column are replaced by the mean of all values in the column.
2. Median Substitution – Missing values are replaced by the median of all values in the column.
3. KNN imputation – Missing values are imputed by K Nearest Neighbours algorithm. This algorithm finds the K nearest observations to the observation having missing value and takes the mean of K nearest observations for the column having missing values to find the missing value replacement.

We apply the 3 methods above to find the replacement for the missing value.

We take one observation with non-missing values and put the value equal to np.nan for the column (whose missing values have to be replaced).

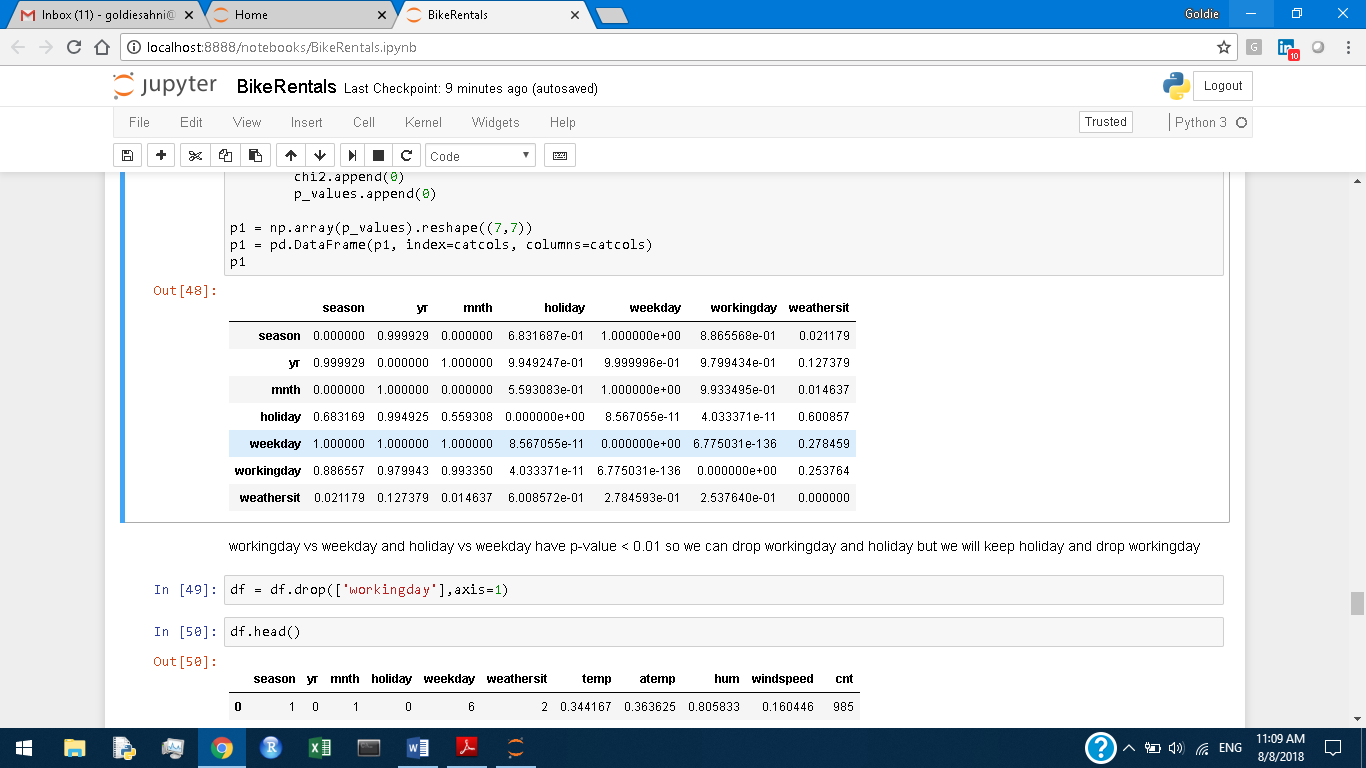
The method which gives the replacement closest to original value of the column in the observation is taken as the best method for missing values replacement.

Median substitution has been found to the best method for imputing missing values in the variables hum & windspeed.

**2.4 Correlation Analysis**

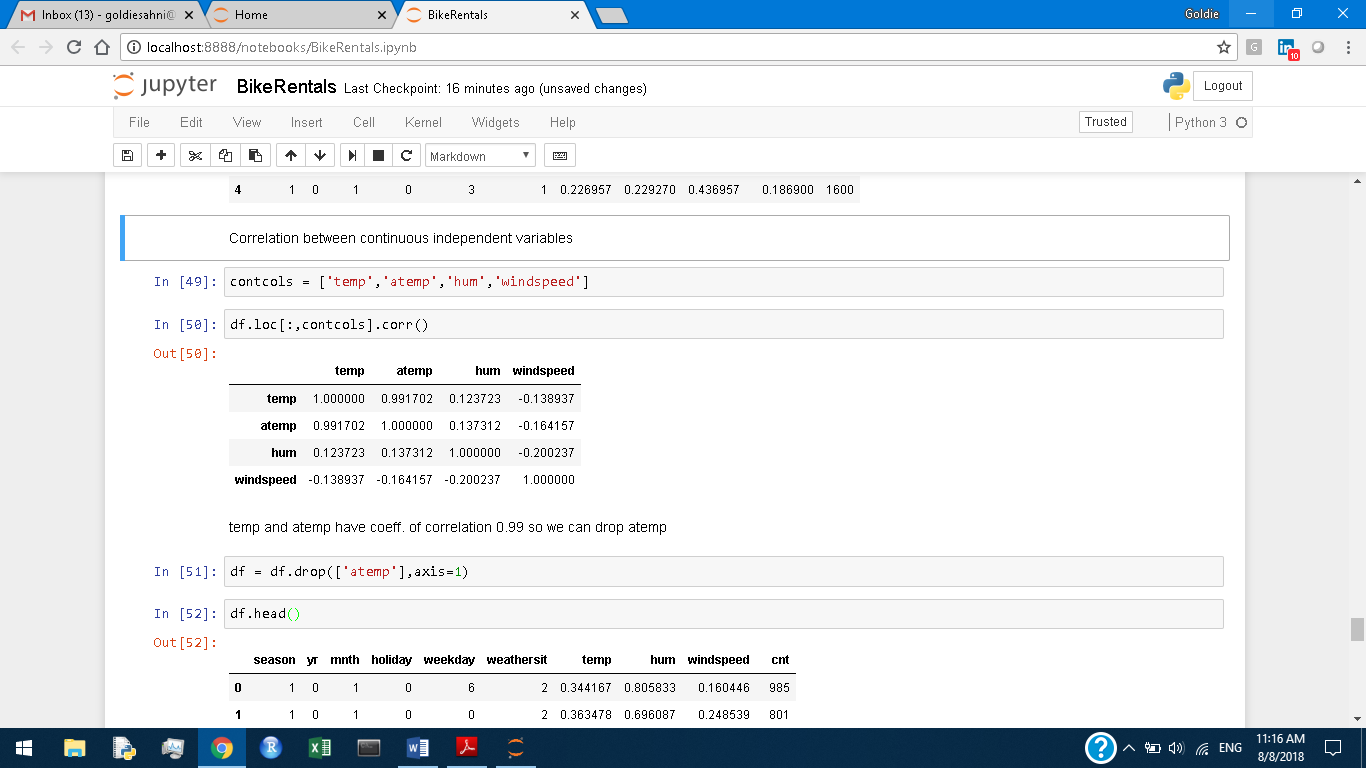
Variables - season, yr, mnth, holiday, weekday, workingday, weathersit have been converted to category type.

Chi-square test was done for correlation between categorical variables:



workingday vs weekday and holiday vs weekday have p-value < 0.01 so we can drop workingday and holiday but we will keep holiday and drop workingday.

Correlation test was done for continuous independent variables:



temp and atemp have coeff. of correlation 0.99 so we will drop atemp.

Chapter 3

Modelling

Categorical variables have been converted to dummy variables for modelling.

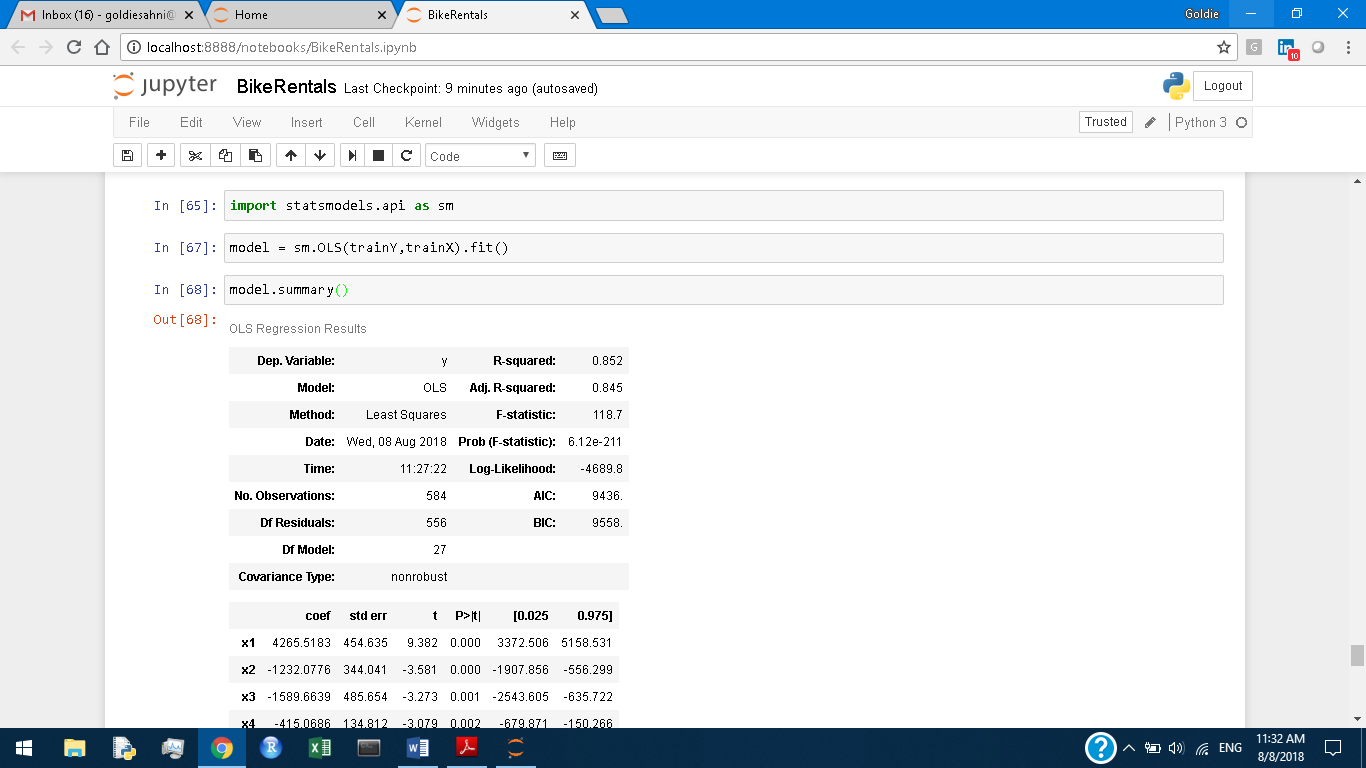
After dummy variables conversion, dataframe df has 33 independent variables and one dependent variable.

Dataframe df has been split into train & test sets in the ratio of 80:20.

**3.1 Linear Regression**

Ordinary Least Squares Regression has been applied to the data to build a model.

Model Summary is:



R-squared is 0.852 which means that the model is good. R-squared of 0.852 means that 85% of variance in dependent variable may be explained by variance in independent variables.

temp and windspeed are the variables having most effect on bike counts. temp has positive effect and windspeed has negative effect.

Prediction was done for test set. Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) was calculated.

MAPE was 139 % and MAE was 598.19

**3.2 Decision Tree**

A decision tree was fit to the train data with maximum depth of tree equal to 4.

MAPE was 159% and MAE was 739.63

**3.3 Random Forest**

A random forest of 570 trees was fit to train data with max. depth of tree equal to 9.

MAPE was 159% and MAE was 565.61

**3.4 K Nearest Neighbors**

KNN was fit to train data with no. of nearest neighbors to be considered as 8.

MAPE was 132% and MAE was 665.80

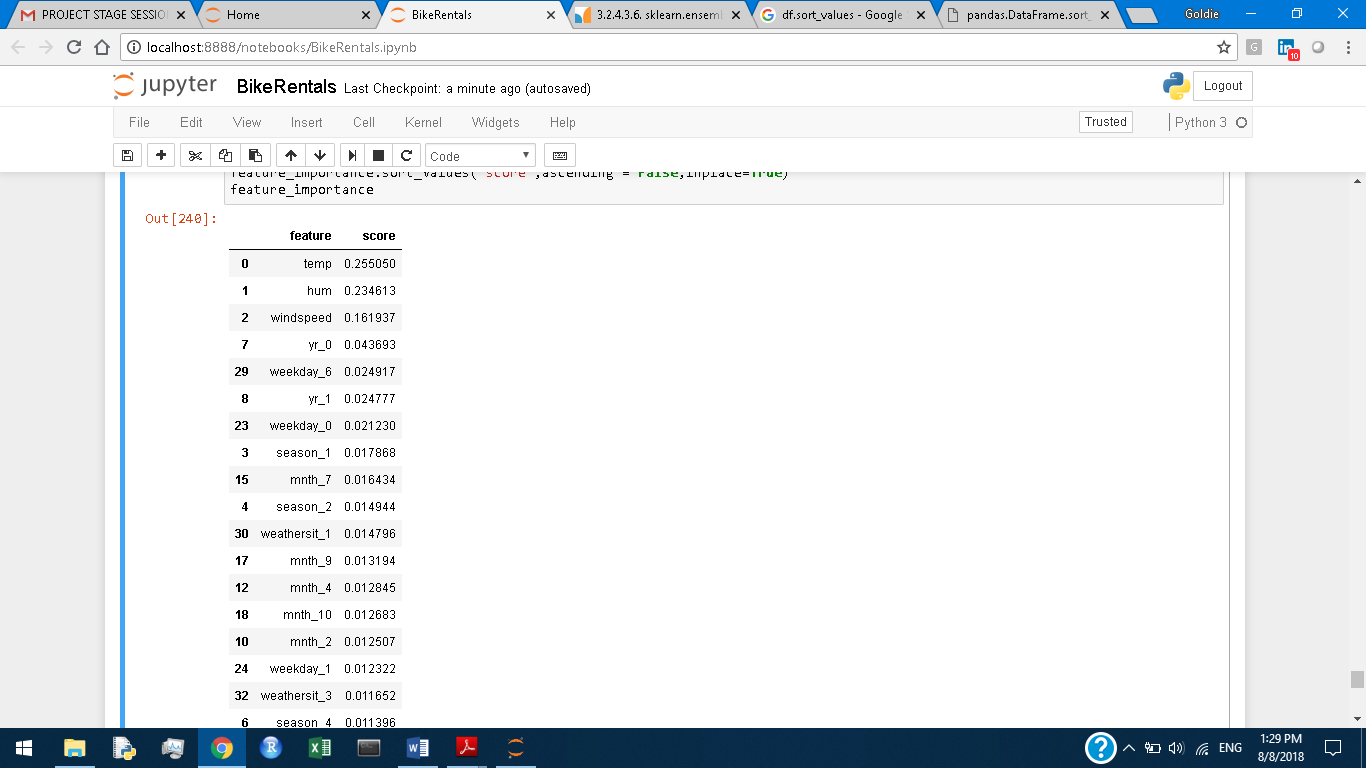
**3.5 Gradient Boosted Trees**

140 GBT were fit to train data with max. depth of trees equal to 4.

MAPE was 112% and MAE was 520.76

So, Gradient Boosted Trees algorithm has given us the best result for this problem.

Feature importance is as under:



temp, hum and windspeed are the three most important features.

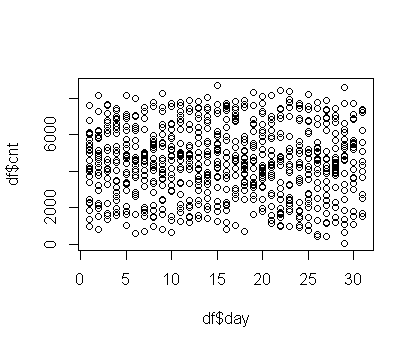
Modelling in R

Chapter 4

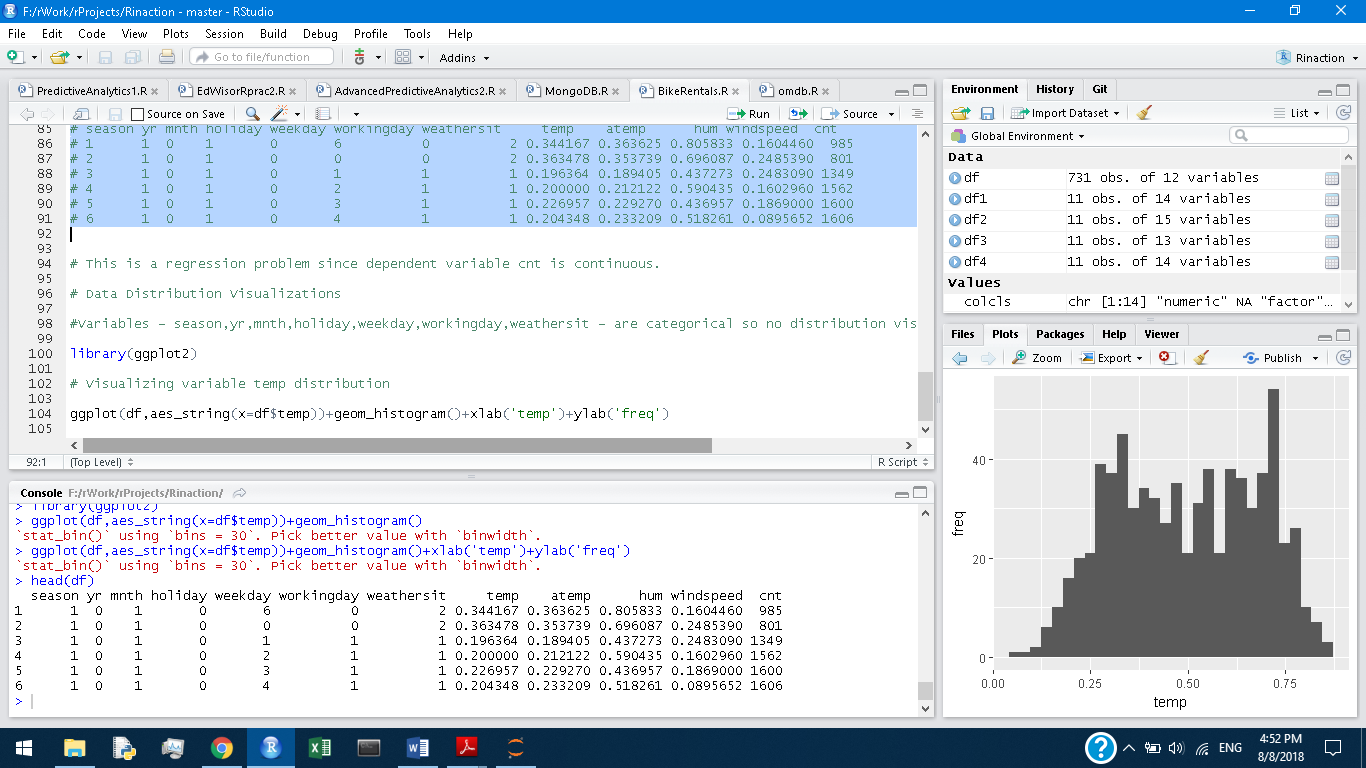
Exploratory Data Analysis

**4.1 Day Relationship with cnt**

First, dteday was converted to date type and then day was extracted from dteday variable to check if day has any relationship with cnt. But it was found that no relationship exists between day and cnt as shown by scatter plot of day vs cnt.



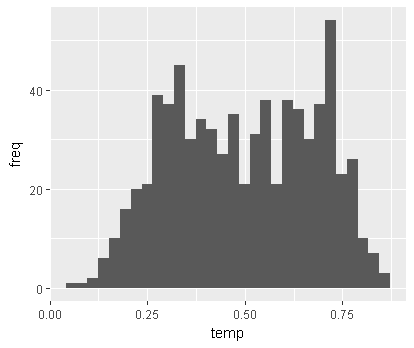
We are dropping instant, dteday (mnth and yr are present), day, casual & registered. Now, data has 12 variables only.

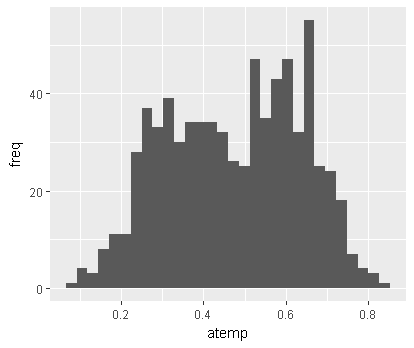


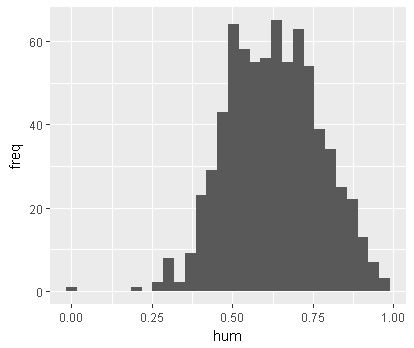
**4.2 Variables distributions**

Variables - season, yr, mnth, holiday, weekday, workingday, weathersit - are categorical so there is no need of distribution visualization.

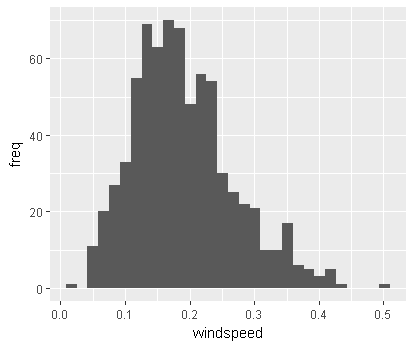
1. Variable temp distribution



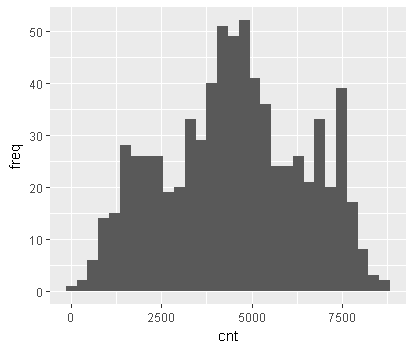
1. Variable atemp distribution
2. Variable hum distribution



1. Variable windspeed distribution



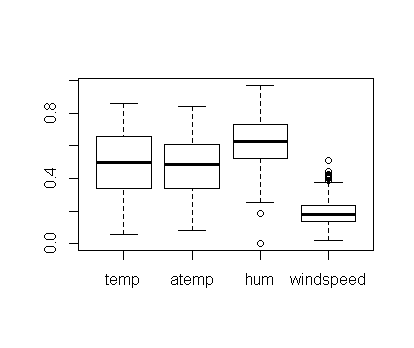
1. Variable cnt distribution



Variables hum & cnt are close to normal. All other continuous variables are skewed.

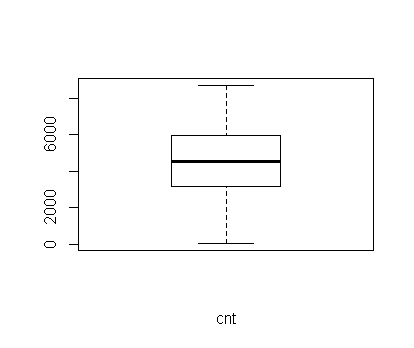
**4.2 Outliers Analysis**

Boxplots for variables – temp, atemp, hum, windspeed



hum & windspeed have outliers. These outliers will be replaced with NAs.

Boxplot for cnt



cnt has no outliers.

Variable hum has 2 missing values and windspeed has 13 missing values after outliers replacement with NAs.

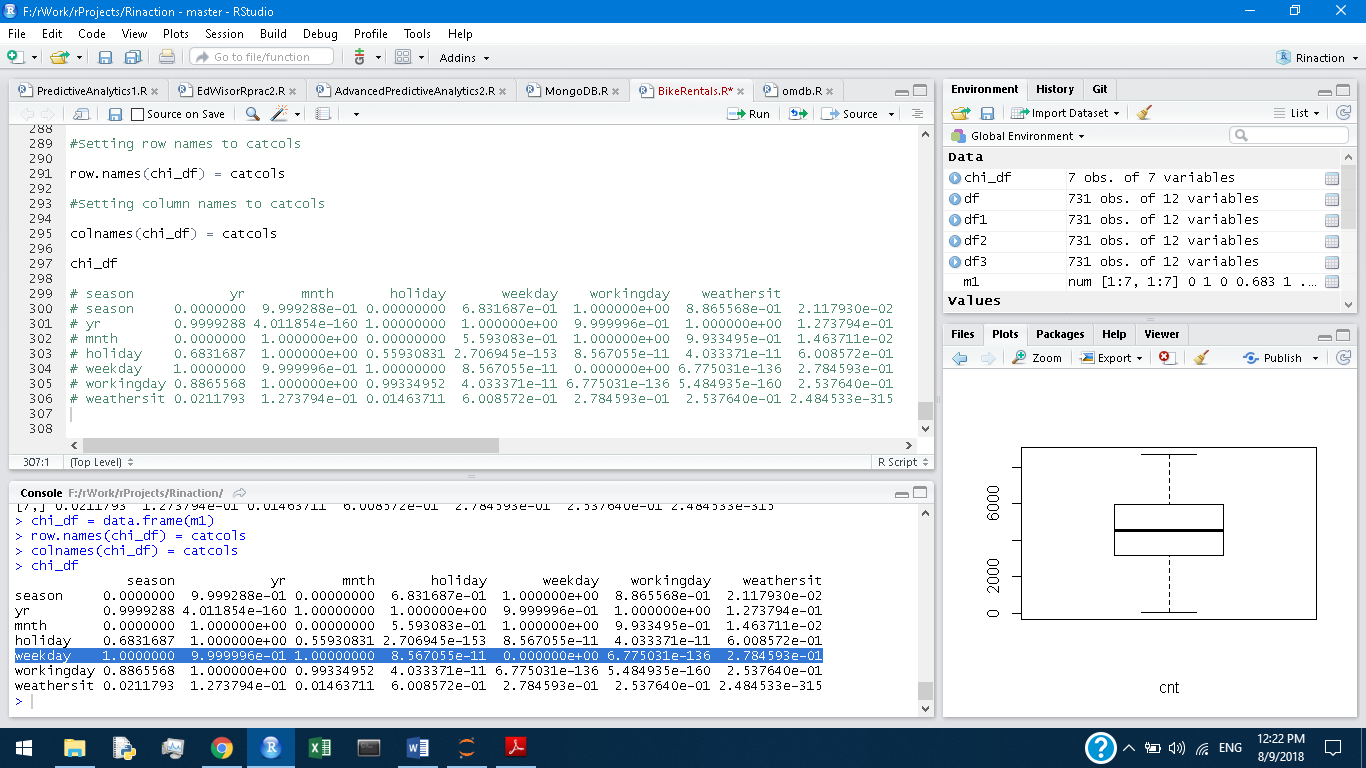
**4.3 Missing Values Analysis**

Knn Imputation has been found to the best method for imputing missing values in the variables hum & windspeed.

**4.4 Correlation Analysis**

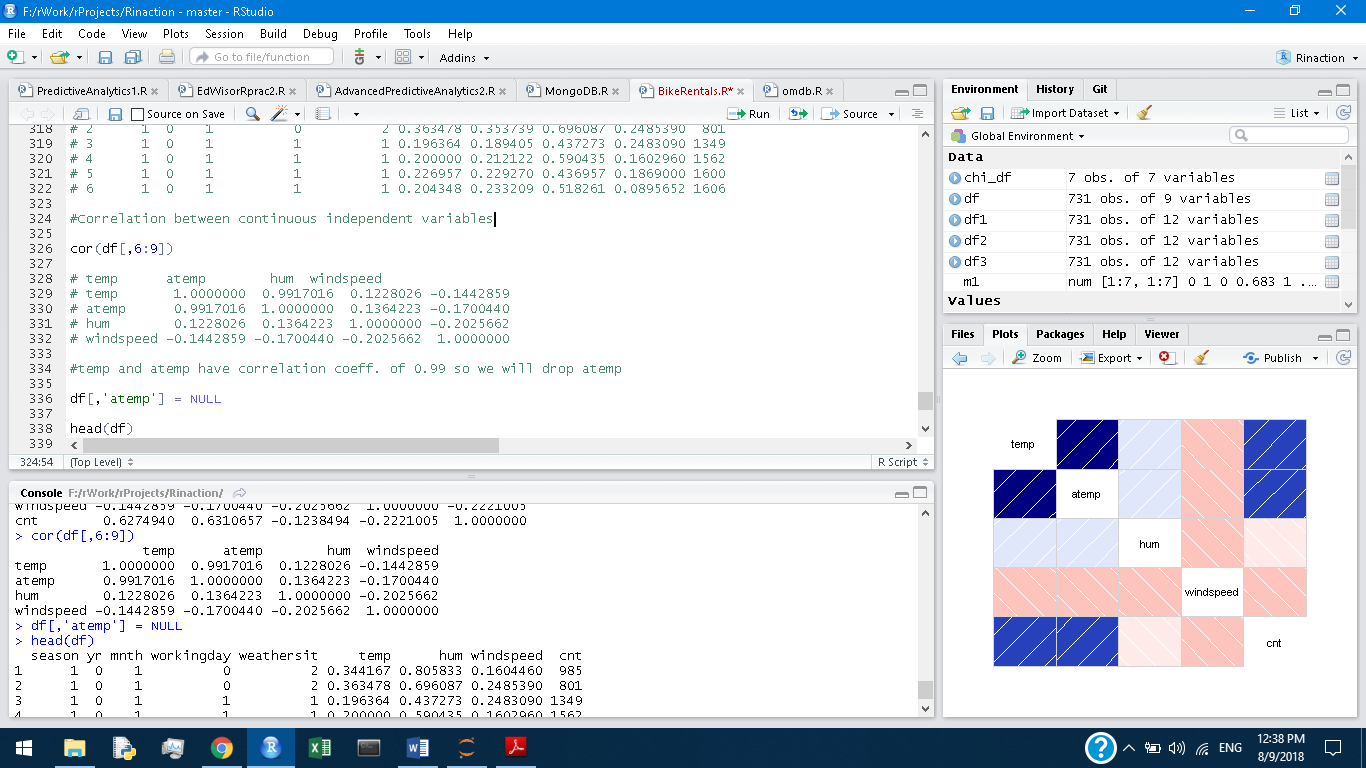
Variables - season, yr, mnth, holiday, weekday, workingday, weathersit have been converted to factor.

Chi-square test has been done for finding correlation between factors:



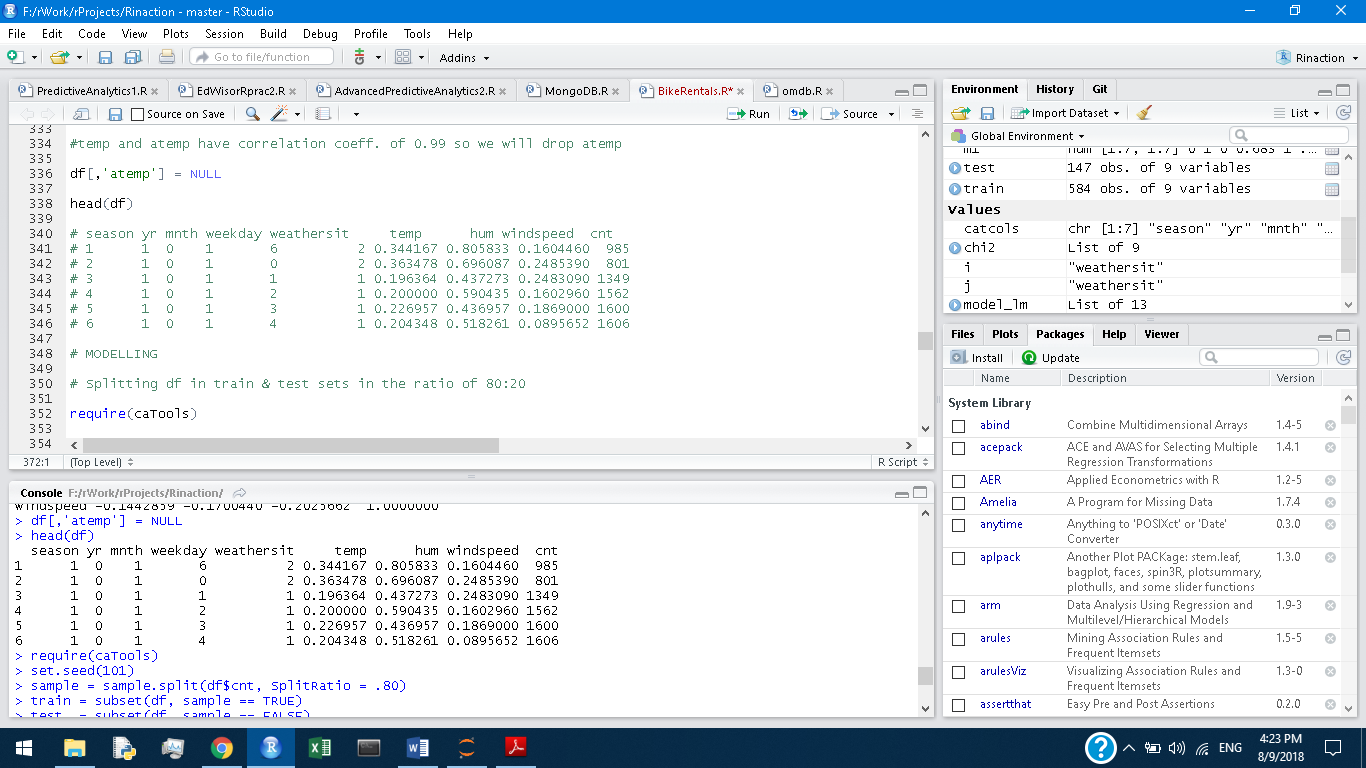
workingday vs weekday and holiday vs weekday have p-value < 0.01 so we can drop workingday and holiday.

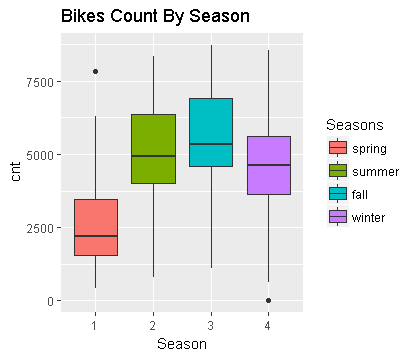
Correlation between continuous independent variables



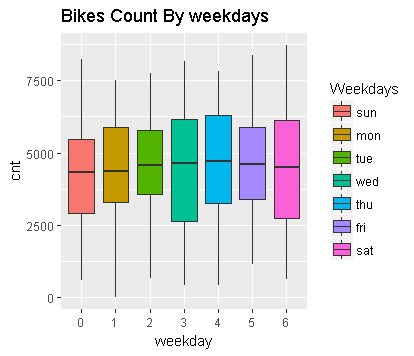
temp and atemp have correlation coeff. of 0.99 so we will drop atemp.

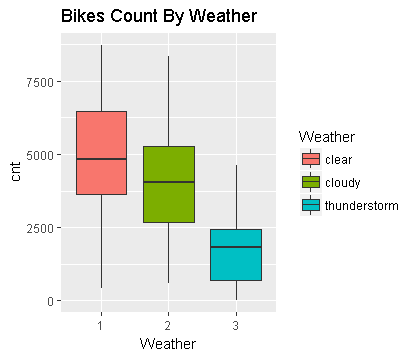
Now, df has 9 variables.



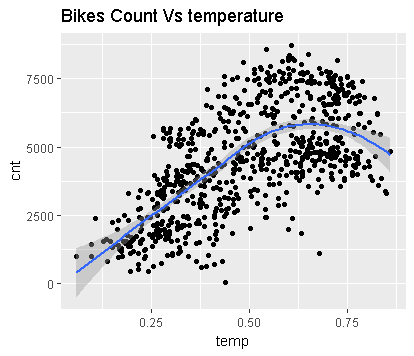


Most bike rentals occur in Fall followed by Summer.

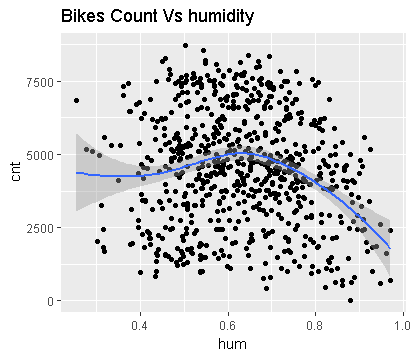




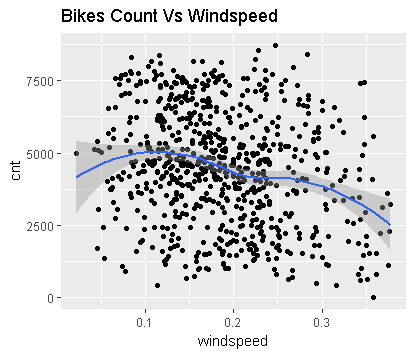
Most bikes are rented during clear weather



As temp. increases, bikes count increases initially then decreases.



As humidity increases, bike count decreases.



As windspeed increases, bike count decreases.

Chapter 5

Modelling

Dataframe df has 8 independent variables and one dependent variable.

Dataframe df has been split into train & test sets in the ratio of 80:20.

**5.1 Linear Regression**

lm function has been applied to the data to build a model.

Model summary:

Residuals:

Min 1Q Median 3Q Max

-3992.2 -352.7 63.0 462.5 2207.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1356.56 274.87 4.935 1.06e-06 \*\*\*

season2 958.08 209.12 4.581 5.70e-06 \*\*\*

season3 753.00 246.48 3.055 0.002358 \*\*

season4 1719.00 202.45 8.491 < 2e-16 \*\*\*

yr1 2051.86 64.40 31.859 < 2e-16 \*\*\*

mnth2 240.34 160.00 1.502 0.133642

mnth3 586.55 192.15 3.053 0.002377 \*\*

mnth4 473.05 278.03 1.701 0.089421 .

mnth5 775.60 298.92 2.595 0.009718 \*\*

mnth6 730.50 314.11 2.326 0.020396 \*

mnth7 296.89 349.51 0.849 0.395998

mnth8 757.21 336.07 2.253 0.024640 \*

mnth9 1235.58 296.59 4.166 3.59e-05 \*\*\*

mnth10 511.47 264.25 1.936 0.053429 .

mnth11 -103.16 255.08 -0.404 0.686062

mnth12 -79.52 199.53 -0.399 0.690394

weekday1 94.49 117.96 0.801 0.423447

weekday2 270.91 117.81 2.300 0.021840 \*

weekday3 293.50 120.15 2.443 0.014880 \*

weekday4 388.10 117.24 3.310 0.000992 \*\*\*

weekday5 440.77 121.24 3.636 0.000303 \*\*\*

weekday6 393.51 116.60 3.375 0.000790 \*\*\*

weathersit2 -561.56 87.10 -6.448 2.47e-10 \*\*\*

weathersit3 -1889.02 235.89 -8.008 6.83e-15 \*\*\*

temp 4196.15 452.74 9.268 < 2e-16 \*\*\*

hum -1361.91 342.93 -3.971 8.08e-05 \*\*\*

windspeed -2211.15 483.82 -4.570 6.01e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 756.7 on 557 degrees of freedom

Multiple R-squared: 0.854, Adjusted R-squared: 0.8472

F-statistic: 125.3 on 26 and 557 DF, p-value: < 2.2e-16

R-squared is 0.854 which means that about 85% of variance in dependent variable can be explained by variance in independent variables.

temp has coeff. of 4196.15 which means that temp is having the most(positive) effect on cnt.

windspeed has coeff. of -2211.15 which means that windspeed is having the second most effect(negative) on cnt.

MAPE was 133 % and MAE was 602.36

**5.2 Decision Tree**

Rpart package was used to fit a decision tree to the train data.

MAPE was 196 % and MAE was 678.02

**5.3 Random Forest**

randomForest package was used to fit 500 trees.

MAPE was 140 % and MAE was 554.63

**5.4 K Nearest Neighbors**

Knn.reg from FNN was used to fit to train data.

MAPE was 171 % and MAE was 738.44

**5.5 Gradient Boosted Trees**

50000 Gradient Boosted Trees from gbm library was used to fit to train data.

MAPE was 99 % and MAE was 506

Variables relative influence:

|  |  |
| --- | --- |
| **Variable** | **Rel. Influence** |
| temp | 35.419354 |
| yr | 25.271824 |
| mnth | 11.605649 |
| hum | 8.668744 |
| season | 6.797198 |
| windspeed | 4.612378 |
| weekday | 4.406589 |
| weathersit | 3.218264 |

Gradient Boosted Trees have the lowest MAPE and lowest MAE so GBT is the best algorithm for this problem.

Chapter 6

Conclusion

We have applied following algorithms to the problem in both R and Python:

1. Linear Regression
2. Decision Tree
3. Random Forest Regression
4. K Nearest Neighbors
5. Gradient Boosted Trees

Gradient Boosted Trees have been found to be the best algorithm for this problem in both R and Python.

140 Gradient Boosted Trees with maximum depth of tree as 4 were fit to training data in Python. Then, prediction for test data was done. Mean Absolute Percent Error (MAPE) was 1.12(112%) and Mean Absolute Error (MAE) was 520.76

50000 Gradient Boosted Trees with maximum depth of tree as 4 were fit to training data in R. Then, prediction for test data was done. Mean Absolute Percent Error (MAPE) was 0.99(99%) and Mean Absolute Error (MAE) was 506.48

Gradient Boosted trees model in Python is similar to that in R.

Gradient Boosted Trees model in Python gives temp(temperature), hum(humidity) and windspeed as top three influencers of dependent variable cnt.

Temperature has a positive influence on cnt whereas humidity and windspeed have negative influence on cnt.

We can also quantify their effects if we consider Linear Regression models in both Python and R as under:

|  |  |  |
| --- | --- | --- |
| Variable | Coefficient in Python | Coefficient in R |
| temp | 4265.5183 | 4196.15 |
| hum | -1232.0776 | -1361.91 |
| windspeed | -1589.6639 | -2211.15 |

Regression Coefficients of temp(temperature) and hum(humidity) are almost similar in both Python & R.