# Bike Lending Company – Linear Regression assignment

Member Name - Pamela Roy

Batch – MLC51 EPGP ML&AI

# Agenda

- Problem Statement
- Reading and Understanding the data
- Data preparation for Modelling
- Building and Training the Model
- Residual data analysis
- Prediction and evaluation on test data

# Problem Statement

BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

They want to understand the factors affecting the demand for these shared bikes in the American market.

The company wants to know:

-Which variables are significant in predicting the demand for shared bikes.

-How well those variables describe the bike demands

- **Reading and Understanding the Data**

Data files provided – day.csv , Readme

# Reading and Understanding the Data

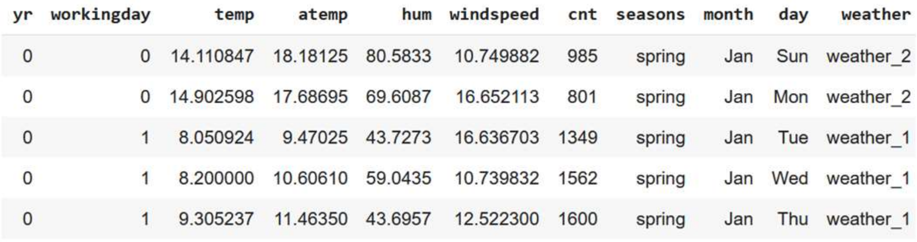
**On analysing day.csv file below are the observations :**

- The data file ‘day.csv’ is uploaded in G-drive & code is written in google.colab to read the file
- On checking the detailed file information we find -
  - Total number of rows : 730
  - Total number of columns – 16
  - There are NO null values in any rows
  - Data types used are float64 , int64 and object
- Below are the statistical information on the file were computed – count , mean , std , min , max and percentiles – 25 , 50 & 75
- Column details are analysed from ‘Readme’ file .
  - Column ‘Instant’ : Record Index , which is of no use for analysis
  - Column ‘dteday’ : Date - Individual dates wont be of much use in the analysis
  - Column ‘holiday’ : weather day is a holiday or not is already captured in column ‘working’
  - Column ‘cnt’ is the total number of bikes which includes casual & registered users
  - Data in Column ‘casual’ and ‘registered’ is redundant
- Columns ‘instant’ , ‘dteday’ are of no use & columns holiday , ‘casual’ and ‘registered’ are redundant information . Hence these columns are dropped .

# Reading and Understanding the Data

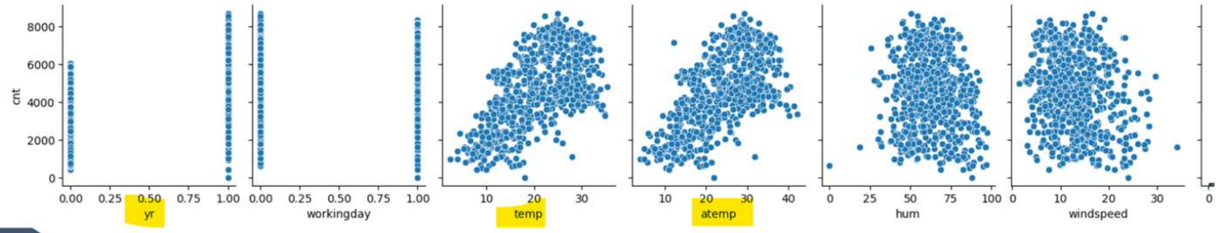
**Data understanding continued… :**

- Check for duplicate data is done . NO duplicates were found
  The non binary categorical variables like – season , mnth , weekday , weather which had numeric values were mapped back to original values . Eg : 1:spring, 2:summer, 3:fall, 4:winter
- Post changes the data lookd like as below : 730 rows & 11 columns



| yr | workingday | temp | atemp | hum | windspeed | cnt | seasons | month | day | weather |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 14.110847 | 18.18125 | 80.5833 | 10.749882 | 985 | spring | Jan | Sun | weather_2 |
| 0 | 0 | 14.902598 | 17.68695 | 69.6087 | 16.652113 | 801 | spring | Jan | Mon | weather_2 |
| 0 | 1 | 8.050924 | 9.47025 | 43.7273 | 16.636703 | 1349 | spring | Jan | Tue | weather_1 |
| 0 | 1 | 8.200000 | 10.60610 | 59.0435 | 10.739832 | 1562 | spring | Jan | Wed | weather_1 |
| 0 | 1 | 9.305237 | 11.46350 | 43.6957 | 12.522300 | 1600 | spring | Jan | Thu | weather_1 |

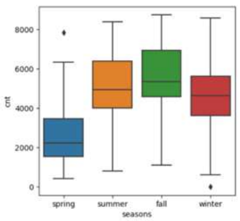
- Pair plotting is done for the numeric variables to see the corelations . Co relation is observed between the target variable ‘cnt’ and variables like ‘temp’ , ‘atemp’ , ‘yr’



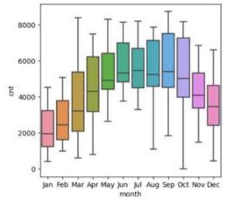
# Reading and Understanding the Data

**Data understanding continued… :**

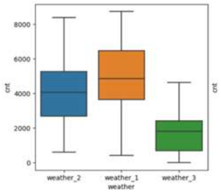
- Analysing the categorical variables against the target variable ‘cnt’ we see that for some seasons like Summer & fall the demand has increased & during spring there is substantial decrease in demand of bikes .



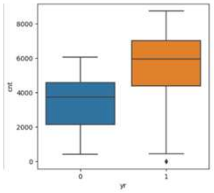
- The ‘month’ variables show that there high demands in July , Sept where as very low demand in Jan ,Feb



- The ‘weathersit ‘ variable shows there is no demand for bikes where value is 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog  and there is a substantial decrease in demand where weathersit is 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds .



- The ‘yr’ column shows demand has increased in 2019  .

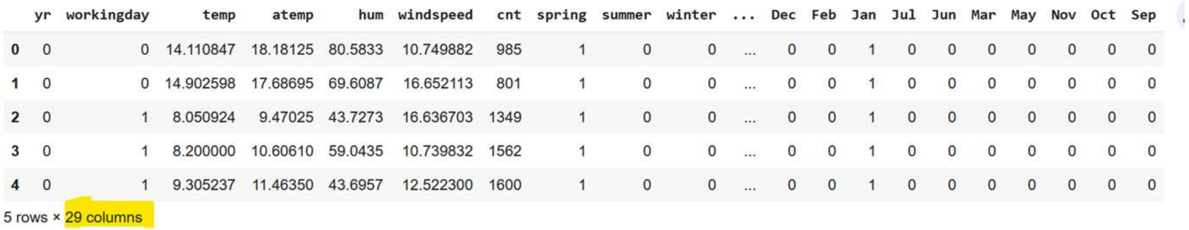


- Other categorical variables like  ‘workingday’ or ‘day’  , don’t show much effect on target variable .

- Data preparation for Modelling

# Data preparation for Modelling

- Dummy variables added for non binary categorical variables
  - Dummies added for season and original 'season' column dropped
  - Dummies added for 'weathersit' and original 'weather' column dropped
  - Dummies added for 'weekday' and original 'day' column dropped
  - Dummies added for 'mnth' and original 'month' column dropped
- After adding dummies the number of columns grow to 29

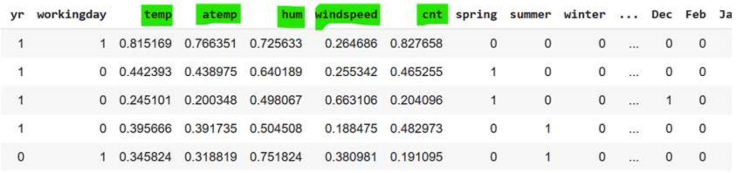


| | yr | workingday | temp | atemp | hum | windspeed | cnt | spring | summer | winter | ... | Dec | Feb | Jan | Jul | Jun | Mar | May | Nov | Oct | Sep |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 14.110847 | 18.18125 | 80.5833 | 10.749882 | 985 | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 14.902598 | 17.68695 | 69.6087 | 16.652113 | 801 | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 1 | 8.050924 | 9.47025 | 43.7273 | 16.636703 | 1349 | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 1 | 8.200000 | 10.60610 | 59.0435 | 10.739832 | 1562 | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 1 | 9.305237 | 11.46350 | 43.6957 | 12.522300 | 1600 | 1 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

5 rows × 29 columns

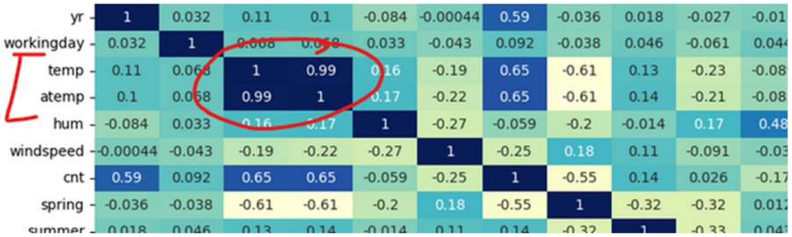
- The data set is then split to TRAIN & TEST datasets in 70% and 30% ratio
  The shape of the sets are as below :
  Train set : Rows – 510 ,Col – 29
  Test Set : Rows – 220 , Col – 29
- Rescaling done using the MinMax Scaler for the numeric non – binary variables
  - temp , atemp , hum , windspeed , cnt

# Data preparation for Modelling



| yr | workingday | temp | atemp | hum | windspeed | cnt | spring | summer | winter | ... | Dec | Feb | Ja |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.815169 | 0.766351 | 0.725633 | 0.264686 | 0.827658 | 0 | 0 | 0 | ... | 0 | 0 | |
| 1 | 0 | 0.442393 | 0.438975 | 0.640189 | 0.255342 | 0.465255 | 1 | 0 | 0 | ... | 0 | 0 | |
| 1 | 0 | 0.245101 | 0.200348 | 0.498067 | 0.663106 | 0.204096 | 1 | 0 | 0 | ... | 1 | 0 | |
| 1 | 0 | 0.395666 | 0.391735 | 0.504508 | 0.188475 | 0.482973 | 0 | 1 | 0 | ... | 0 | 0 | |
| 0 | 1 | 0.345824 | 0.318819 | 0.751824 | 0.380981 | 0.191095 | 0 | 1 | 0 | ... | 0 | 0 | |

- Post scaling all values are within range of 0 and 1 .
- Next we created a heatmap with the data available (all numeric)
  - the heatmap shows a very high correlation between predictor variables – 'atemp'& 'temp'



- 'temp' column is dropped based on the above observation .

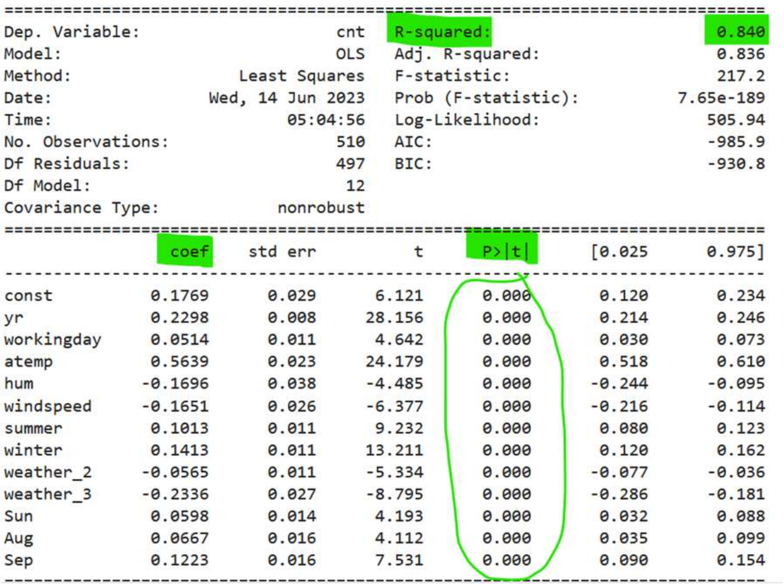
- The train dataset is next split into X & y sets for model building
  - Only target variable 'cnt' is stored in y & rest all columns are stored in X

- Building and Training the Model

# Building & Training  the Model

- Using Linear Regression model the X_train and Y_train dataset is fit
- Using RFE  - Recursive feature Elimination method we bring down the number of columns from 29 to 12
- The RFE model ranks the 12 variables as 1
- The **selected predictor variables** are : 'yr', 'workingday', 'atemp', 'hum', 'windspeed', 'summer', 'winter', 'weather_2', 'weather_3', 'Sun', 'Aug', 'Sep’
- Add a constant to the model using add_constant method from StatsModels library
- Then we run the Linear regression using the OLS method from same library on the TRAIN dataset
- Next we check the summary  :
  - The R2 value is .84
  - Coefficients of the 12 variables are given
  - For all the 12 variables the P value is Zero .

## LM Summary

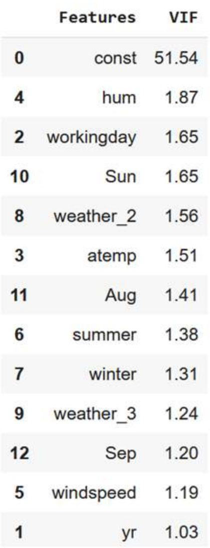


```
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.840
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     217.2
Date:                Wed, 14 Jun 2023   Prob (F-statistic):          7.65e-189
Time:                        05:04:56   Log-Likelihood:                 505.94
No. Observations:                 510   AIC:                            -985.9
Df Residuals:                     497   BIC:                            -930.8
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1769      0.029      6.121      0.000       0.120       0.234
yr             0.2298      0.008     28.156      0.000       0.214       0.246
workingday     0.0514      0.011      4.642      0.000       0.030       0.073
atemp          0.5639      0.023     24.179      0.000       0.518       0.610
hum           -0.1696      0.038     -4.485      0.000      -0.244      -0.095
windspeed     -0.1651      0.026     -6.377      0.000      -0.216      -0.114
summer         0.1013      0.011      9.232      0.000       0.080       0.123
winter         0.1413      0.011     13.211      0.000       0.120       0.162
weather_2     -0.0565      0.011     -5.334      0.000      -0.077      -0.036
weather_3     -0.2336      0.027     -8.795      0.000      -0.286      -0.181
Sun            0.0598      0.014      4.193      0.000       0.032       0.088
Aug            0.0667      0.016      4.112      0.000       0.035       0.099
Sep            0.1223      0.016      7.531      0.000       0.090       0.154
------------------------------------------------------------------------------
```

# Building & Training  the Model

- To check for multicollinearity we check the Variance Inflation factor – VIF for the 12 chosen variables . The pic in right shows that none of the predictor variables have a high VIF value
- No further variables need to be dropped & the current model is our final model .

## VIF values

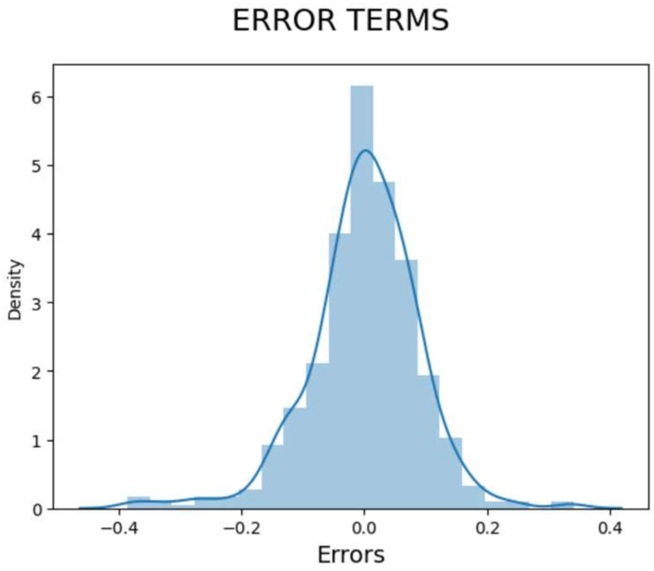


| | Features | VIF |
|---|---|---|
| 0 | const | 51.54 |
| 4 | hum | 1.87 |
| 2 | workingday | 1.65 |
| 10 | Sun | 1.65 |
| 8 | weather_2 | 1.56 |
| 3 | atemp | 1.51 |
| 11 | Aug | 1.41 |
| 6 | summer | 1.38 |
| 7 | winter | 1.31 |
| 9 | weather_3 | 1.24 |
| 12 | Sep | 1.20 |
| 5 | windspeed | 1.19 |
| 1 | yr | 1.03 |

- **Residual data analysis**

# Residual data analysis

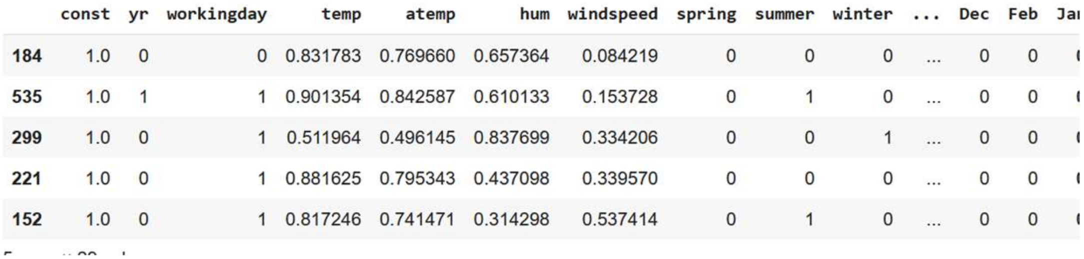
- The target variable Y is predicted using the new model
- Next we plot the differences of Y_train  & Y_predict i.e. the Residuals
- Observation – The graph shows NORMAL distribution   as shown below .



# Prediction and evaluation on test data

# Prediction and evaluation on test data

- Next step is to do the prediction & evaluation on the TEST dataset using the model that was built on TRAIN dataset.
- The scaler is used on test dataset to transform
- Post scaling we add constant to the data .



| | const | yr | workingday | temp | atemp | hum | windspeed | spring | summer | winter | ... | Dec | Feb | Ja |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **184** | 1.0 | 0 | 0 | 0.831783 | 0.769660 | 0.657364 | 0.084219 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **535** | 1.0 | 1 | 1 | 0.901354 | 0.842587 | 0.610133 | 0.153728 | 0 | 1 | 0 | ... | 0 | 0 | 0 |
| **299** | 1.0 | 0 | 1 | 0.511964 | 0.496145 | 0.837699 | 0.334206 | 0 | 0 | 1 | ... | 0 | 0 | 0 |
| **221** | 1.0 | 0 | 1 | 0.881625 | 0.795343 | 0.437098 | 0.339570 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **152** | 1.0 | 0 | 1 | 0.817246 | 0.741471 | 0.314298 | 0.537414 | 0 | 1 | 0 | ... | 0 | 0 | 0 |

- Drop the ‘temp’ column in test dataset same like we did on the train dataset .
- Make predictions on test dataset using the Model
- Evaluate the model checking on the R2 – score .
- R2 on test data is .81 where as the R2 on Train dataset was .84
- The model is behaving good as we get less than 5% variance between Train & Test sets using the Multiple Linear regression model

# Thank you

Pamela Roy

Email : bhattacharya.pam@gmail.com