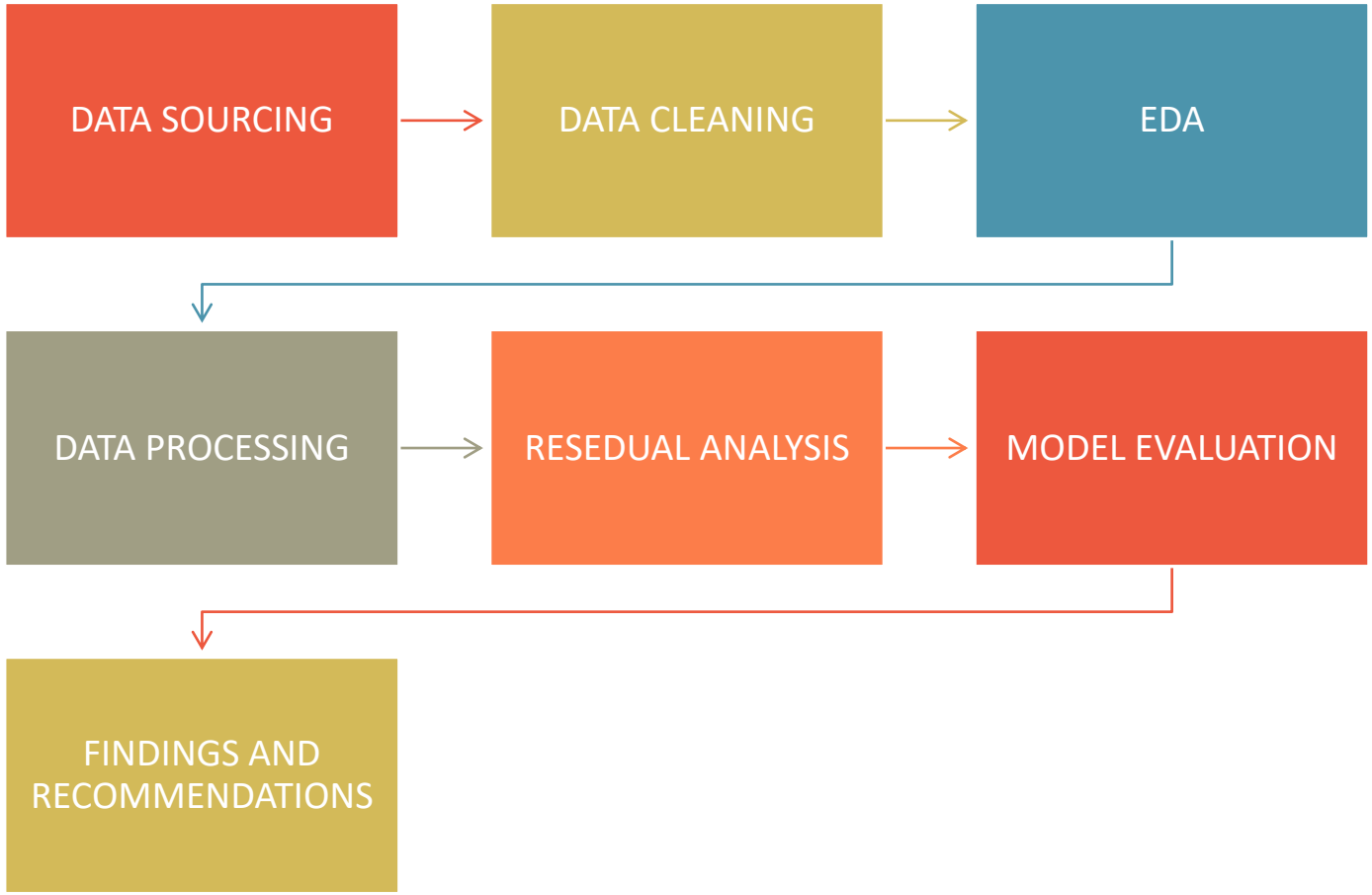
# Bike Sharing Assignment

Linear Regression

NAMRATA SHIVTARKAR

# Road Map for Methodology



# DATA SOURCING

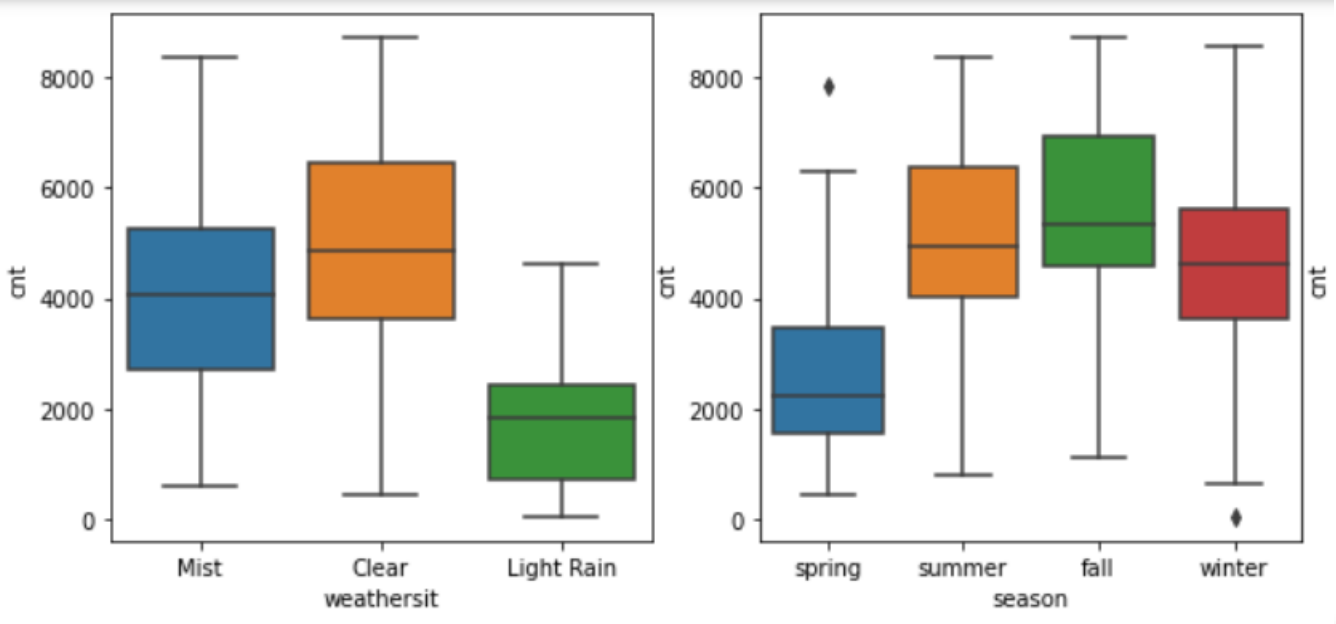
- day.csv file provided by upgrad is used for data analysis
- Each variables mentioned in Data dictionary excel is used for better understanding provided by upgrad

# DATA CLEANING

- Checking the datatypes of data
- Checking mean, min, max, median
- Checking for missing values and null
- Dropping columns where only one value presented because It don't give any explanation
- Concerting variables to categorical string variables

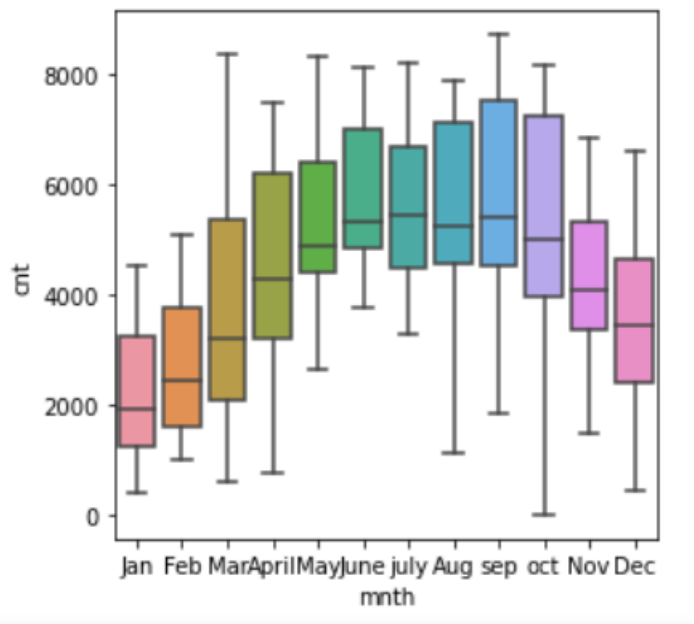
# EDA

- categorical vs continuous variable
  - High demand for bikes when weather is clear and mist
  - High demand for bikes in fall and summer season
  - Demand raises gradually till sep its on high peak



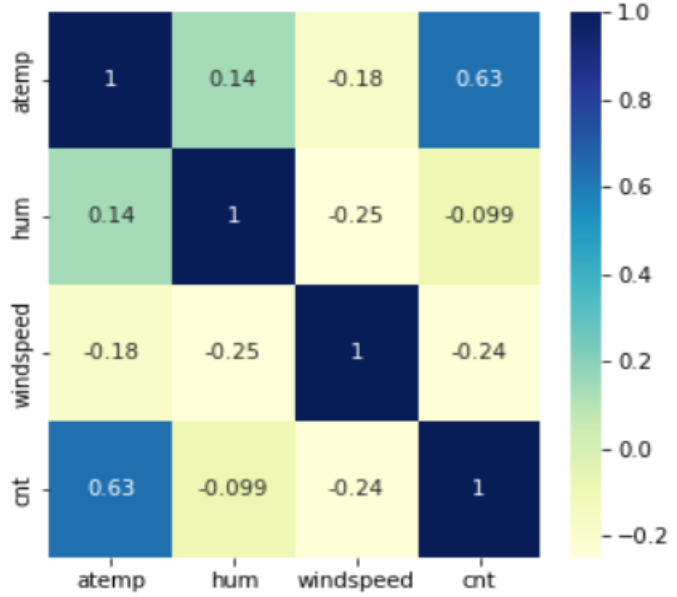


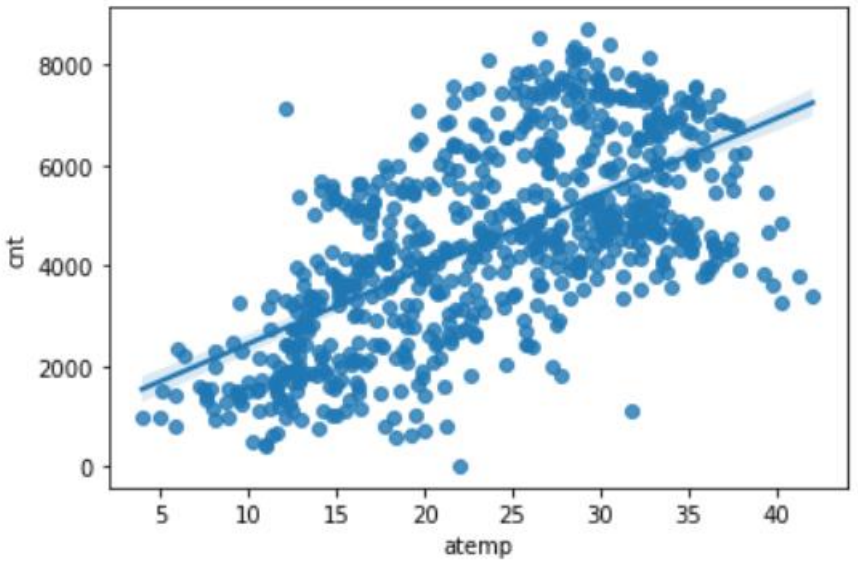




# EDA

- Continuous vs continuous variables
  - Bike demand increases as temperature increases
  - Heatmap also tells same thing



# DATA PROCESSING

- Create dummy variables for all categorical variables like weather, season, weekday and month
- Train – Test split: dividing the data into train set(70%) and test set(30%)
- Rescaling continuous X variables between 0-1
- Train the model by adding constant and using OLS method
- Checking the R2 and p- values of all variables
- But we will use REF and manual elimination method here

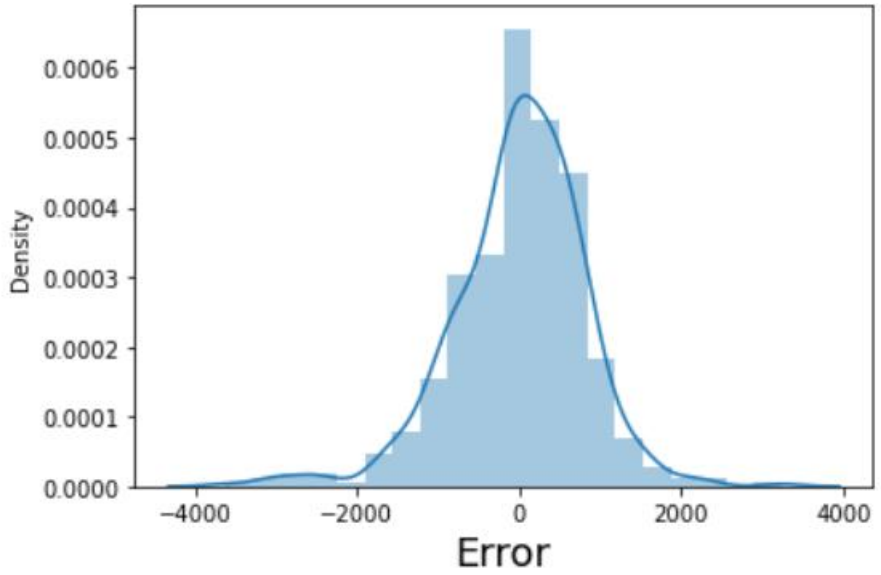
# RFE

- Using REF we found 15 important variables like year, workingday, atemp, windspeed, weather_clear, weather_light_rain, weather_mist, mnth_dec, mnth_feb, mnth_jan, mnth_nov, mnth_sep, season_spring, season_winter, weekday_sat
- After fitting the model with above variables we found R2 with 0.83
- Now we manually eliminate variables based on non-zero p-value
- After deleting 5 variable we found final 9 variables. We had to delete weathersit_clear because it have VIF 15, hence we have 9 variables

| | Features | VIF |
|---|---|---|
| 2 | atemp | 5.34 |
| 1 | workingday | 4.29 |
| 3 | windspeed | 3.88 |
| 0 | yr | 2.03 |
| 8 | weekday_Sat | 1.74 |
| 7 | season_spring | 1.62 |
| 5 | weathersit_Mist | 1.51 |
| 6 | mnth_sep | 1.15 |
| 4 | weathersit_Light Rain | 1.07 |

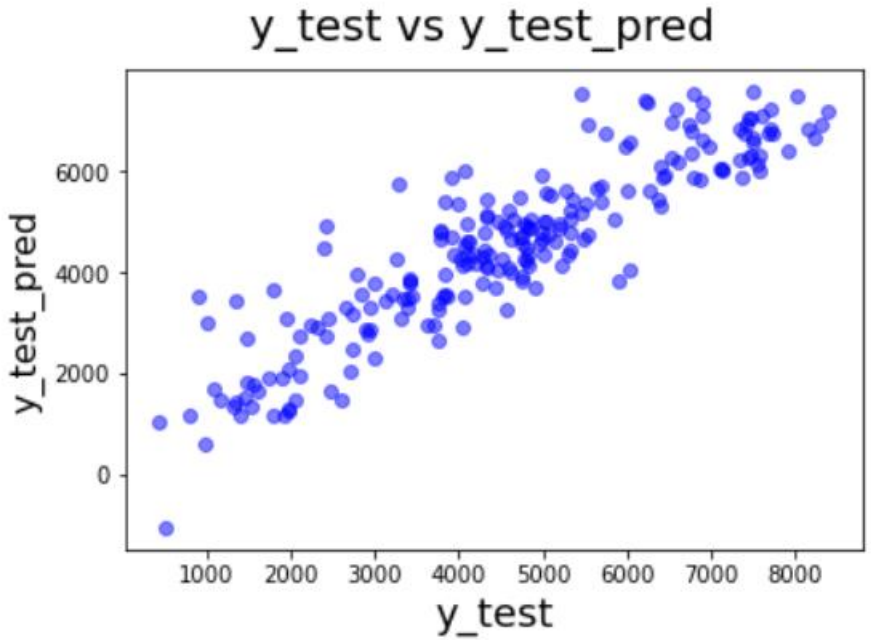
# RESIDUAL ANALYSIS ON TRAIN DATA

- Difference between y_train and y_train_pred is the residual
- Below distplot shows normal distribution of error terms with mean 0



# PREDICTION AND EVALUATION ON TEST SET

- similarly transform, adding constant and predict methods are used on test set
- Variance of residual is constant is seen below
- R2 value for test set is 0.80
- R2 for train set 0.82



# FINDING AND RECOMMENDATIONS

Observation:

- Company can focus during summer and fall season by introducing new marketing schemes like promoting brand through youtube ads, instagrams, facebook etc
- In the other season when market is down, you can start introducing shelter bikes which you can drive in light rains or in emergencies. Company can introduce shelters just before rain starts with offers so that demand dont decrease.
- when people book bike or registered themselves they can get special discounts, vouchers, scratch cards
- during sep demand is at peak, company can do promotion with above suggestion at this time

# FINDING AND RECOMMENDATIONS

Variables that help predict demand of bikes

- year
- workingday
- atemp
- windspeed
- weathersit_Light Rain
- weathersit_Mist
- mnth_sep
- season_spring
- weekday_Sat