Assignment-based Subjective Questions & Answers:

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Season, yr,mnth,holiday,workingday,weathersit are the categorical variables in the dataset.

- Bike Share count is the dependent variable .

- Season 3 that is on Autumn/fall has highest no of Bike share.

- Year-2019 has most no of bike share

- May to September are the months where customers have used more no of Bikes.

- Non-Holiday shows less no bikes are shared.

- Each weekday has almost 13-14% of bikes are shared.

- No of bikes shared are more on Working day

- "Clear, Few clouds, Partly cloudy, Partly cloudy" day shows more no of bikes are shared.

2. Why is it important to use DROP_FIRST=True during dummy variable creation?

- DROP_FIRST=TRUE is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Looking at the pair plot among numerical variable,Temp has highest correlation with target variable.

This can also be observed from correlation with heatmap.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Before making the prediction, we should to be certain, that the model is reliable.

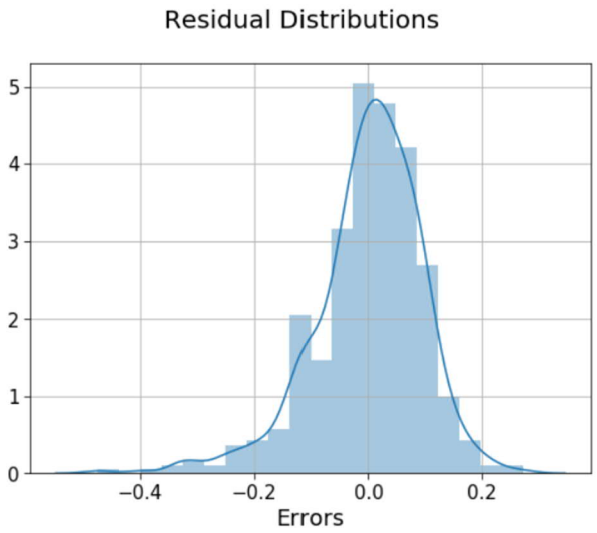
- Hence, we need to first perform Residual Analysis of error terms .

- Residual analysis evaluates the goodness of fitted model.

- Residuals can be calculated as → Observed value – Fitted value.

- The distributions of error term or residuals, should be centered around 0, signifies that it follows a normal distribution.

- We can plot a distribution plot/Histogram to visualize the residuals as below.



5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top 3 features whose contributions are most to the model

- Temperature
- Year (year -1)
- Weathersit (**weathersit_3)**

General Subjective Questions & Answers:

1.Explain the linear regression algorithm in detail.

→Linear regression is a modelling technique between a dependent variable and one or more Independent variable .

→ Linear regression models can be classified into :

1.simple linear regression (One dependent and one Independent variable)

2.Multiple linear regression.(One dependent and more than 1 Independent variable)

→ The equation of best fit regression line can be found out as

Y = B0 +B1X by minimizing the cost function using Differentiation method and Gradient descent method.

→The strength of linear regression model can be found out from R^2 where

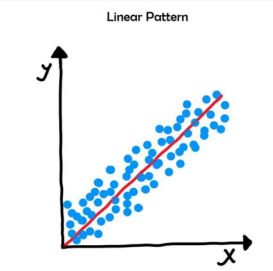
R-squared = 1 – (RSS/TSS)

RSS = Residual sum of squared

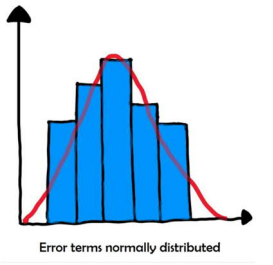
TSS = Total sum of squared.

→The assumptions in linear regression :

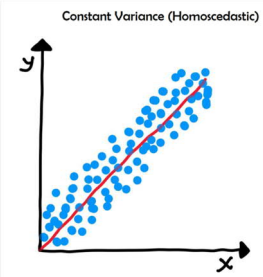
- There should be a linear relationship between predictor and predicted variable.



- Error terms should be normally distributed with mean 0



- Error terms should have a constant variance which is called as Homoscedasticity.

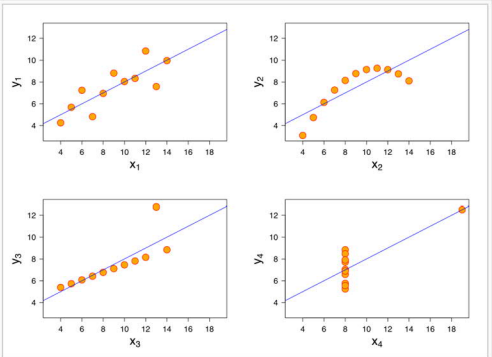


2. Explain the Anscombe's quartet in detail

- Anscombe's quartet is a group of four datasets.

- It appears to be similar when we use summary statistics to explain these datasets but

- They have very different distributions and appear very different when plot the data.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- It tells us that we cannot simply rely in statistics without checking the overall distributions.

- The mean, standard deviation, variance, correlation, R-squared and regression equation are same for all the dataset , but when we graph them the datasets are completely different.

- It reminds us that it is very important to visualize the data to get a clear picture what is going with the data.

3. What is Pearson's R?

- Pearson's R is Pearson correlation coefficient.

- It measures the linear association between two variables let say X and y.

- It has values range between -1 to +1 .

- +1 is highly Positively correlated and -1 is highly Negatively correlated.

- 0 is no correlation

- Mathematically it can be calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

n is sample size

x_i, y_i are the individual sample points indexed with *i*

$\bar{x}$ $\bar{y}$ are the sample mean.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is a process which is applied to variables /columns to bring them to a same level of magnitude.

- Scaling is performed as most of the times the dataset contains features with highly varying magnitudes.

- Let say in our dataset has data points range in few Thousand and some are in Tens. During model training, model assumes higher ranging number has superiority. Hence more significant number starts playing more decisive role. In that way we cannot generalize a model very well.

- Scaling helps in faster gradient descent in optimization process and also it helps in faster convergence.

Normalized Scaling :

- Normalization scaling put our data in a range of either [0, 1] or [1 , -1].

- Normalized scaling through Min-Max scaler can be represented as below.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Normalization scaling loses some information during process if the dataset has outliers.

Standardized scaling :

-Standardized scaling puts the data to have Zero mean and One Variance.

- Standardized scaling can be represented as

$$x_{new} = \frac{x - \mu}{\sigma}$$

mu – mean

Sigma – Standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF infinitive means it shows the perfect correlation among independent variables.

For highly correlated variables R-squared is 1.

VIF is 1/ (1- R_squared). R-squared 1 means denominator of VIF is 0 and 1/0 is infinitive.

Hence if we get VIF as infinitive we need to drop that variable for model training.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

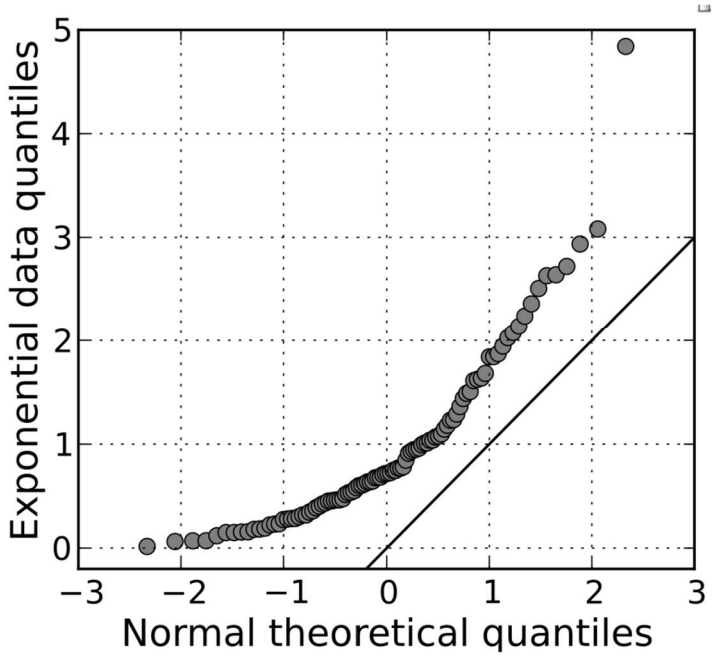
- Q-Q plot is quantile -quantile plot.

- It is used to determine if two sets of data come from same distribution .

- This helps in linear regression when we have training and test dataset received separately.

- Through Q-Q plot we can check if the training and test data has same distributions, so that model can be generalized on train and test dataset.

-A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.



-Q-Q plot can be plotted via STATSMODELS.API in python as below .

```
import numpy as np
import statsmodels.api as sm
test = np.random.normal(0,1, 1000) # Normal distribution data
sm.qqplot(test, line='45')
```

]: