# Assignment based Subjective Questions and Answers

Q1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variables?

Answer:- In season column, most of the booking were happen in season 2, 3 over 5000 and in season 4 above 4000 to 5000 . In month column, May (5) to October (10) had booking over 4000 which was also a good sign.

In Weathersit column, weathersit1:- Clear, Few clouds, partly cloudy, partly cloudy had more booking compared to other situation. In holiday column most of bike booking happen when it's not holiday. Weekday variables value were same in all days so it has no significant effect. Year 2019 had more booking then 2018. In Working day column, high booking percentage had 2 years of data.

Q2. Why is it important to use drop _first=True during dummy variable creation?

Answer:- in more than two variables we have N numbers of dummy variables which is increase model redundancy ,less efficient , less effectiveness in model ,and if we can intercept data using N-1 dummy variables why we need N variables, that's why we use droup_first=True .

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with target variable?

Answer:- Temp and Atemp has highest correlation between target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: - Based on P-value (0.00) which show all significant variable.

F-statistic value is high (272.9) which show best fit model. Adjusted R2 and R2 value has less than 5 % gap and shows above 80% accuracy. No sign of multicollinearity between variables. AIC value is less which shows best fit for model. Durban-Watson value is 2.0 which shows no autocorrelation between variables. Based this factors I validate the assumptions.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:- Temp, yr shows positive significantly change on demanding bikes means increase those variables increase bike demand while weathersit_3 shows show negative sign when increase this situation decrease bike demand. Thus Temp, yr, weathersit_3 shows significantly change in bike demand.

# General Subjective Questions and Answer

Q1. Explain the linear regression algorithm in detail.

Answer:- linear regression algorithm based on the supervised learning method.

It has two type linear regression

1. Simple linear regression
2. Multiple linear regression

SLR method: - this method has 1 independent variables. We can evaluate data in SLR with help of best fit line , Residual sum of squared (R2) value y actual – y predicted values , cost function (Rmse) . Cost function indicates the sum of error between y actual and y predicted value. The equation of regression line found minimize the cost function using ordinary least squared method. Which is done using the different ion and gradient descent function. After doing eda and understating the data divide the data in to two parts training and testing the data.

MLR method: - this method has more than 1 independent variables. Above process are same in MLR but some difference in it. Overfitting, multicollinearity, feature selection are the new consideration for mlr model. Creating dummy variables, r2 and adjusted R2 value, feature selection, aic, bic manual feature selection and automated feature selection, model evolution this are use in mlr model.

Thus, linear regression algorithm predict dependent variables based on independent variables.so, this help linear relationship between variables.

## Q2. Explain the Anscombe's quartet in detail.

Answer:-Anscombe's quartet comprise four data sets which have nearly same stasticaly property. Which shows correlation between x and y. which comprise eleven points of (x, y). It demonstrate both graphical data before analyzing and statistical property of data. With help of we can found mean, standard deviation, correlation between x and y. this also shows varies anomalies of data like diversity, linearseparability of data , data is fit for linear relation with other data or its incapable of handling any other data sets.

Q3. What is Pearson's R?

Answer: - Pearson's R correlation coefficient referred to as Pearson's R. is linear correlation between two sets of data, bivariate analysis. Covariance of two variables, divide by the product of their standard deviation. Covariance has normalized measurement such result value has -1 to 1.

This pcc varies between -1 to 1:-

r=1 means data is perfectly positive linear relationship with slope.

r=-1 means data is negative linear relationship with slope.

r=0 means there is no relationship.

r>0<5 means there is weak association.

r>5<8 means there is moderate association.

r>8 means there is a strong association.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: - scaling is a pre-processing data of independent variables to normalize the data within particular range. It also help for better efficiency in algorithm.

Most of the times data set collected features are highly varying in magnitudes. Because of this attribute of features we end up with incorrect model. So, we need the features varying magnitude in same level of range. So, we can end correct modeling and accurate insights from data sets.

Normalized scaling vs Standardized scaling

| Normalized scaling:- | Standardized Scaling:- |
|---|---|
| 1. Min and max value is use. | 1. Mean and std deviation is use |
| 2. used when features are Different. | 2. Used when want zero mean STD deviation |
| 3. Scale [-1, 1] | 3. It is not bounded on range. |
| 4. Called scaling normalization | 4. Calles z- score normalization. |
| 5. Minmaxscaler for normalization . | 5. Standard scaler for Standardization. |

Q5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: - This indicates perfect correlation between independent variables. Where R2=1. To solve this problem we need to drop one of the variables from data sets which causing multicollinerity between variables.

This infinite value indicates exactly linear relationship between variables.

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

Answer: - Q-Q plot means quantile –quantile plots are two quantiles among each other. For example median is quantiles where 50% of data fall below that point and 50% data fall above that point. Purpose of plot is to find out of two kind of dataset fall from same distribution. If two data set come from common distribution then it will fall on that reference line.

Q-Q plot is used to compared the shapes of distribution, providing a graphical view of properties like location, scale etc.