

# Assignment-based Subjective Questions

**1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

I have plotted the categorical variables with the target variables on boxplot and has Inferred following effect on target:

Season: 3: fall has highest demand for rental bikes

I see that demand for next year has grown, Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing. When there is a holiday, demand has decreased. Week-day is not giving clear picture about demand.

The clear weathers hit has highest demand

**2. Why is it important to use drop first=True during dummy variable creation? (2 marks)**

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.

If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

“temp” has highest correlation. It is very well linearly related with target “cnt”.

**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Error terms are normally distributed with mean 0.

Error Terms do not follow any pattern.

Multicollinearity check using VIF(s).

Linearity Check.

Ensured the overfitting by looking the R2 value and Adjusted R2.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Features “holiday”, “temp” and season “hum” are highly related with target column, so these are top contributing features in model building.

## **General Subjective Questions**

**1.Explain the linear regression algorithm in detail. (4 marks)**

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, we can write a simple linear regression equation as follow  $y \sim b_0 + b_1 * x$  Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.

Here, x and y are two variables on the regression line. b1 = Slope of the line. b0 = yintercept of the line. x = Independent variable from dataset

Here, x and y are two variables on the regression line. b1 = Slope of the line. b0 = yintercept of the line. x = Independent variable from dataset y = Dependent variable from dataset.

**2. Explain the Anscombe’s quartet in detail. (3 marks)**

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the

Data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled

By linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled

By linear regression model.

### **3. What is Pearson's R? (3 marks)**

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

Pearson's r measures the strength of the linear relationship between two variables. Pearson's r always between -1 and 1.

If data lie on a perfect straight line with negative slope, then  $r = -1$ .

Positive correlation indicates the both the variable increase and decrease together.

Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

### **4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

It is important to note that scaling just affects the coefficients and none of the

Other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Machine Learning algorithm works on numbers not units. So, before regression on a Data set it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scale a variable in range 0 and

1. Standardization: It transforms data to have a mean of 0 and standard deviation of 1.

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

When there is a perfect relationship then  $VIF = \text{Infinity}$  whereas if all the independent variables are orthogonal then to each other then  $VIF = 1.0$ . Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution

It is used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few advantages:

- a) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and
- b) presence of outliers can all be detected from this plot.

It can be used with sample sizes also. It is used to check following scenarios: If two data sets —come from populations with a common distribution

have common location and scale

have similar distributional shapes

have similar tail behaviour

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset.

Below are the possible interpretations for two data sets.

- a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45degree from x -axis
- b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c)  $X\text{-values} < Y\text{-values}$ : If  $x$ -quantiles are lower than the  $y$ -quantiles.

d) distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from  $x$  -axis.