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?**

Answer: Found following categorical variables in the dataset:

- Season: Bike demand declines during Spring. Seen highest demand during fall season.
- Year: Demand is higher in 2019 as compared to 2018.
- Month: Seen high demand during month of May to October.
- Weather Condition: Demand is high on clear and mist cloudy weather. Seeing decline in light rain or snow condition.
- Weekday: There is no effect in bike's demand.
- Workingday: Demand doesn't change whether day is working or not.
- Holiday: There is no effect in bike's demand.

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

Answer: It prevents multicollinearity while creating dummy variables from categorical variable. It helps in reducing extra column generated during dummy variable. E.g. let's say we have season data with values like fall, spring winter and summer. When generating dummy variable for season, we just need 3 columns instead of 4. We can define season values as below:

Fall	=	1 0 0
Spring	=	0 1 0
Winter	=	0 0 1
Summer	=	0 0 0

If we include 4 columns instead of 3, one extra column will have linear correlation with other 3 columns which is not desirable.

So, if we have K discrete values in categorical column, we create K-1 dummy variable and use drop_first=True to delete redundant variable.

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

Answer: temp and atemp both are having similar and highest correlation with target variable cnt.

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

Answer: Executed followings to validate linear regression assumptions:

- Normally distributed error terms with 0 mean: Plotted histogram of the error terms. Found normally distributed error terms with mean 0. Error terms are calculated as below:
  Error terms = (y_train – y_train_predicted)
- Linear relationship between X and Y: Plotted pair-plots for continuous variables and found linear relationship between continuous variables like temp and atemp with predicted variable cnt.

- Error terms are independent and have constant variance: Plotted scatter plot between X_train(training data) and residual. Could not find any pattern in the plot.

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

Answer: Top 3 contributing features are given below:

- Temperature(temp) - Coefficient value of 0.3743 indicates that a unit increase in temperature increases the bike booking by 0.3743 units
- Year(yr) - Coefficient value of 0.2359 indicates that a unit increase in yr variable increases the bike booking by 0.2359 units
- Weather(lightrainsnow) - Coefficient value of -0.1992 indicates that a unit increase in lightrainsnow variable decreases the bike booking by 0.1992 units

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is an algorithm which provides a linear relationship between dependent variable and independent variable to predict the outcome of future events. It mathematically models the dependent variable and independent variable as linear equation. E.g. We can use linear regression to predict the house price(y) using the number of rooms(x1), or we can use it to predict the sales(y) based on marketing spend(x1).

Linear regression is of two types:

- Simple Linear Regression: When there is only one independent feature or variable we can use simple linear regression. The equation for simple linear regression is given below:

  $y = \beta 0 + \beta 1 X$

  y = Independent variable
  X = Dependent variable
  β1 = Slope
  β0 = Intercept

- Multiple Linear Regression: When there is more than one independent feature or variable we can use multiple linear regression. The equation for multiple linear regression is given below:

  $y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots\ldots\ldots \beta n Xn$

  y = Independent variable
  X1, X2 ,.., Xn are dependent variable
  β0 = Intercept
  $\beta 1$, $\beta 2$,.., βn are slopes

2. Explain the Anscombe's quartet in detail

Answer: Anscombe's quartet proves the importance of visualizing the data before applying various algorithms to build models. It comprises of four datasets with identical summary statistics but have very different distribution and appearance when plotted and those plots are not interpretable by any regression algorithms. Four datasets show following characteristics when plot using scatterplot:

Set1: Follows linear regression model
Set2: Data is non-linear
Set3: There are multiple outliers and cannot be fitted in linear regression
Set4: There are multiple outliers and cannot be fitted in linear regression

So based on the metrics summary, we may decide to apply linear regression which will be grossly inaccurate. That is why Anscombe's quartet emphasize to visualize the data set before attempting to model the data or implement any machine learning algorithm.

3. What is Pearson's R?

Answer: Pearson's R or Pearson correlation coefficient is the common way to measure a linear correlation. It is a measure of the strength of the association between two continuous variables. Value of this can lie from -1 to +1. A value of +1 is the perfect positive relationship between two variables. Positive correlations indicate that both variables move in the same direction. E.g. Sales increase with marketing spends. On the other hand, a value of -1 is the perfect negative relationship. Negative correlation indicates that both variables move in opposite direction. E.g. Car rental demand decreases in snowy weather. A zero indicates no relationship between variables.

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

Answer: Scaling is a data pre-processing step which is applied to independent variables to normalize the data to a particular range. It is applied to handle highly varying values or units.

Scaling is performed for the following reasons:

a) Scaling brings values of all independent variables on a comparable scale and ranges. If it is not done, then machine learning algorithm tend to give higher weightage to bigger values. So, we will get incorrect modelling.
b) When scaling is done, several machine learning methods including gradient descent based algorithms perform better and converge more quickly.

Differences between normalized and standardized scaling are given below:

Normalized Scaling	Standardized Scaling
Min and max values of the feature variables are used for this scaling	Mean and standard deviation are used for this scaling
Scale values between 0 and 1	It is not bounded to specific range
Outliers affect values	It is less affected by outliers

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

Answer: A variance inflation factor (VIF) is a measure of the multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in regression model. The value of VIF is calculated as below:

$$VIF_i = 1/(1 - Ri^2)$$

i refer to ith variable

Now if Ri^2 =1, then denominator of above formula become 0 and value of VIF will be infinity. Ri^2 =1 denotes that this variable is perfectly correlated with another independent variable.

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

Answer: A quantile-quantile (Q-Q) plot is a visual method to determine if a dataset follows a certain probability distribution or two samples of data came from the same population or not. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of second data set.

It can be used to check the followings for two data sets:

- Come from same population with same distribution
- Have similar distribution shape
- Have similar tail behavior

In linear regression, we can use Q-Q plot to check if the residuals of the model are normally distributed which is an assumption for linear regression model. Also, from the Q-Q plot we can check if residuals have constant variance which is an assumption for the homoscedasticity of the model.

****