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)

- We plotted a box plot for categorical variables with the target variables, as per the plot, Season 3 Fall has the highest demand for bike hires.
- The demands were higher in 2019 compared to 2018.
- There is no significant difference in the demand on working day versus non-working day.

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

- Using drop_first = True is important because it helps in reducing the extra column while creating dummy variables.
- Reason for this is to avoid multi-collinearity getting added to the model if all dummy variables are included. The reference category can be easily found out where 0 is present in a single row for all the other variables in a particular category.

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

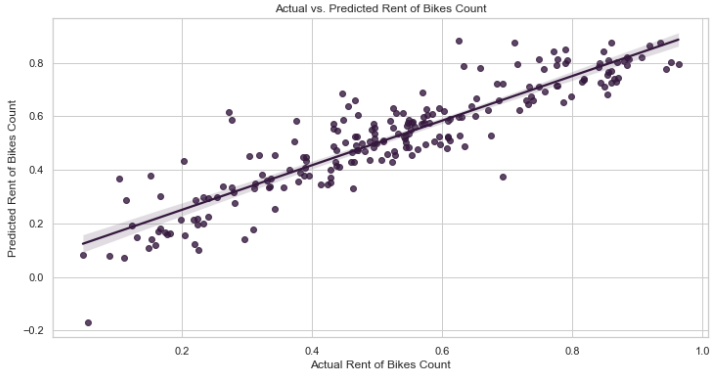
- 'temp' has the highest correlation with the target variable 'cnt'. It is linearly related to the target variable.

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

After building the model on the training set, we checked for various assumptions, including:

- Multicollinearity check with Variance Inflation Factor (VIF)
- Checking R^2 and adjusted R^2 values for training and test set (both did not have much difference 84.2% and ~82% respectively)
- Linear Relationship between independent and dependent variables as shown in the below image

```
In [174]: # Plotting y_test and y_pred to understand the spread.
          plt.figure(figsize = (12,6))
          sns.set_style("whitegrid")
          sns.regplot(x=y_test, y=y_test_pred)
          plt.title('Actual vs. Predicted Rent of Bikes Count')
          plt.xlabel('Actual Rent of Bikes Count')
          plt.ylabel('Predicted Rent of Bikes Count')
          plt.show()
```



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

Top 3 variables contributing significantly towards explaining the demand of the shared bikes are:

- Temperature 'temp'
- Year 'yr'
- Season 'season'
  These variables have the highest significance and are highly correlated to the target column 'cnt'.

General Subjective Questions

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

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

It is one of the very basic forms of machine learning where we train a model to predict the behaviour of our 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.

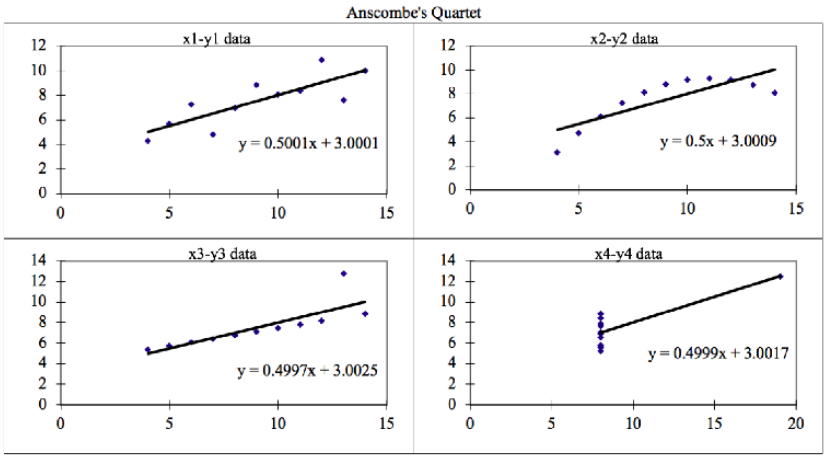
There are two types of linear regression algorithms:

- Simple Linear Regression: Single independent variable is used.
  - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
- Multiple Linear Regression: Multiple independent variables are used.
  - $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ is the line equation for MLR.

  Here, β_0 = value of Y when X = 0 (Y intercept); $\beta_1, \beta_2, .., \beta_p$ = Slope or the gradient.

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

Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



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.

Some points to remember:

- Plotting the data is very important and a good practice before analysing the data.
- Outliers should be removed while analysing the data.
- Descriptive statistics do not fully depict the data set in its entirety.

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

- The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between
- the different variables and the relation with each other.
- Pearson's R always lies between -1 and 1.
- If data lies on a perfect straight line with negative slope, then R = -1.

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.

The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling. There are two types of Scaling:

1. Min-Max scaling (aka Normalization): The MinMax scaling normalizes the data within the range of 0 and 1. The MinMax scaling helps to normalize the outliers as well.

   MinMaxScaling: $x = \frac{x - \min(x)}{\max(x) - \min(x)}$

2. Standardization: It converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

   Standardization: $x = \frac{x - \text{mean}(x)}{sd(x)}$

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

If there is perfect correlation, then VIF is infinite whereas if all the independent variables are orthogonal to each other then VIF = 1.

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which leads : $VIF = \frac{1}{1 - R\text{^}2}$ equal to infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

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.

Advantages of Q-Q plots:

- Many distributional aspects like shifts in locations, scale shifts, symmetry changes and presence of outliers, all can be identified from this plot.
- The plot has a provision to mention the sample size as well.