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)

Ans 1) From the given dataset, let’s consider the effect of the categorical variable ‘weathersit’ on the target variable ‘cnt’. During EDA I observed the relationship between the categorical variables and the target variable. It was seen that during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, there was a decreases in the bike hires numbers by 0.3331 units approximately.

Similarly, we see upward trend in Summer and Fall seasons.

Also, during model building on inclusion of categorical features such as yr,season etc we saw a significant change in the value of R-squared and adjusted R-squared.

With this I concluded that categorical features were helpful in explaining a greater proportion of variances in the data sets

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

Ans 2) It is important to use drop_first=True during dummy variable creation because normal get_dummy() method creates n variables for n distinct values in the field. Where as we can represent all the data with n-1 fields hence that one column is redundant hence need to be dropped in order to minimize computing.

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

Ans 3) Registered had the highest correlation of .95 with cnt.

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

Ans 4) We ploted distplot between the error (i.e. actual y – predicted y) using following:

```
res = y_train-y_train_pred

# Plot the histogram of the error terms

fig = plt.figure()

sns.distplot((res), bins = 20)

fig.suptitle('Error Terms', fontsize = 20)          # Plot heading

plt.xlabel('Errors', fontsize = 18)                 # X-label
```

and verified that the plot is forming a normal distribution centered around 0.0

- Verified there is a linear relation between X and Y.
- Residual Analysis of training data also proves residuals are normally distributed.

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

Ans 5) Based on the final model I found following top contributers in the predictions.
- Temperature (temp) coeff 0.4156
- Weather Situation coeff -.4016
- Year (yr) Coeff .2313

General Subjective Questions

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

Ans 1) Linear regression is a data analysis technique used to predict some target variable base o the past values of other variables. It finds it use in almost industry verticals.

In this we follow following steps:

- Gather data
- Understand data
- Identify insignificant data such missing values outliers etc
- Correct data wherever necessary
- Introduce new data based on the current fields like day, month year form date type fields etc
- Identify redundant and unnecessary data and drop it.
- Univariate and multivariate analysis
- Analyse the effect of various variables on target variable using R-Squared, Adjusted R-Square and P-Value
- Eliminate variables one by one based on above analysis and repeat these steps till desired number of most significant variables remain.

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

Ans 2) **Anscombe's quartet** comprises of a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

It underlines the importance of EDA and the drawback of only depending on summary stats.

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

Ans 3) The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

Ans 4) Scaling is the process of ensuring all the variable values are on similar scale so that they can be plotted on a graph and visualise easily.

Normalized VS Standardized Scaling:

Normalization is a technique used to scale numerical data in the range of 0 to 1. This technique is useful when the distribution of the data is not known or when the data is not normally distributed.

Normalization formula is: Xnorm = (X – Xmin) / (Xmax – Xmin)

standardization is a technique used to transform data into a standard normal distribution. This technique is useful when the distribution of the data is known and when the data is normally distributed.

Standardization: Z = (x-mean)/standard deviation

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

Ans 5) A large value of VIF i.e. (1/ (1 – Ri ^2)) indicate high correlation between variables if there

We normally drop such variables from our analysis, but if there is a perfect correlation between 2 variables that is Ri^2 becomes 1 then the VIF value will become infinite as the denominator in the above equation becomes 0.

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

Ans 6) Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable we are testing the hypothesis for and the second one is the actual distribution we are testing it against.

In Linear regression A Q-Q plot helps us to compare the sample distribution of the variable at hand against any other possible distributions graphically.