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?

Ans: Firstly, except for a single value in the "spring" category and one value in "winter" category of "season" column, we don't have any outliers in our data.

"season" column:

- "min" value in all the four seasons lies in the same range: 0-2000
- "max" value in "spring" is in between 6000 and 8000 where as in other three seasons, "summer", "fall" and "winter", it is above 8000
- Fewer values in "spring" compared to other seasons.
- "cnt" has maximum value during "fall"

"mnth" column:

- "cnt" observed it's least value in the month of October and highest value inthe month of September
- For most of the months (8), average "cnt" value is in between 4000 and 6000

"weekday" column:

- "cnt" is least on Tuesday and highest on Sunday
- Irrespective of the day of the week, the average "cnt" lies in the range 4000-6000

"weathersit" column:

- Interquartile range of "Light Snow/ Light Rain" is comparatively very low w.r.t other weather conditions
- More number of "cnt" values are observed when the weather is "Clear/ Partly Cloudy "

---

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

Ans: Logically, if we have "n" levels in a categorical variable, we need "n-1" columns to represent the dummy variable.
"drop_first = True" is useful in achieving that.
It reduces the extra columns created when creating a dummy variable. Thus, in turn reduces the correlations created among the dummy variables.

----

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

Ans: "temp" has the highest correlation with the target variable (cnt).

----

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

Ans: The assumptions of Linear Regression are validated from the results obtained by displaying the summary of the best model (last one, here) and by plotting a histogram for obtained Error values.

- The error terms are normally distributed.
- The error values are independent of each other
- Every independent variable and the target variable are related with some standard error and a coefficient. Which can be written in the slope intercept form:

  $$y = m1X1 + m2X2 + \ldots +miXi + c$$

  where,
  y is the target variable
  X1,X2,…,Xi are the independent variables and
  m1, m2,…, mi are their corresponding intercepts.
  c is the coefficient

This shows the linear relationship between the independent and target variables.

Thus, all the assumptions of Linear Regression are validated.

----

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

Ans: As Spring season, January and February months are the top three features with VIF values 3.16, 2.12 and 1.84, respectively, they contribute significantly towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression algorithm is a machine learning algorithm, based on the predictive modelling technique that helps us to find the relationship between input and the target variable. It is used to find the effect of input on the target variable and predict the upcoming trends. There are two types of Linear regression based on the number of independent variables.

1. Simple Linear Regression: where there is only one independent variable (x) and a target variable (y)
2. Multiple Linear Regression: where there can be two or more independent variables (x1, x2, x3, …, xn) and a target variable (y).

The independent variables are known as "predictor variables" and the dependent variables are known as "output" or "target" variables.

Linear regression at each X finds the best estimate for Y. As the model predicts a single value, there is a distribution of error terms.

As we are making inferences on the population using a sample, the assumption that variables are linearly dependent is not enough to generalize the results from sample to the population. So, we should have some assumptions to make inferences.

Assumptions of Linear regression:

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero.
3. Error terms are independent of each other
4. Error terms have constant variance. (homoscedasticity)

When you fit a straight line through the data, you will obviously get the two parameters of the straight line: intercept (β0) and the slope (β1). You start by saying that β1 is not significant, i.e. there is no relationship between X and y.

So in order to perform the hypothesis test, we first propose the null hypothesis that β1 is 0.

And the alternative hypothesis thus becomes β1 is not zero.

- **Null Hypothesis (**H0**):** β1=0
- **Alternate Hypothesis (**HA**):** β1≠0

If you fail to reject the null hypothesis that would mean that β1 is zero which would simply mean that β1ia insignificant and of no use in the model. Similarly, if you reject the null hypothesis, it would mean that β1 is not zero and the line fitted is a significant one.

To perform the hypothesis test, you need to derive the p-value for the given β. If the p-value turns out to be less than **0.05**, you can reject the null hypothesis and state that β1 is indeed significant.

The first important step before building a model is to perform the train_test_split. To split the model, you use the train_test_split function and check the summary statistics that was outputted by the model: F-statistic, r-squared, coefficients and their p-values.

Then perform the Residual analysis by plotting histogram of the error terms to check the normality and independence. Then make predictions on the test set

----

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet can be defined as a group of 4 datasets which look nearly identical in simple descriptive statistics but have different distributions and appear differently when plotted on scatter plots. These peculiarities in the data should be identified and handled without doing which these peculiarities fool the built regression model.

This signifies the importance of plotting graphs and visualizing the data before analysing it and building a model. Anscombe's quartet suggests that the variables in the dataset must be plotted to observe the sample distribution which help in identifying the anomalies in the data.

----

3. What is Pearson's R?

Ans: Pearson's R is the correlation coefficient that lies between -1 and +1.
R = -1 and +1 means the data is perfectly linear with negative and positive slopes respectively.
R = 0 means there is no linear correlation in the data
0<R<5 means there is a weak association
5<R<8 means the association in data is moderate and
R>8 means a strong association

----

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

Ans: Scaling is a data pre-processing step applied on predictor variables to normalize the data within certain range, and thus help in speeding up the calculations in an algorithm.

Scaling is very important to bring all the variables to the same magnitude level. This issue arises when the dataset contains variables with high variations in units, ranges and magnitude. If scaling is not done, then the algorithm only takes magnitude into account and not units. This results in incorrect modelling. So scaling is very important to address this issue.

“Normalized scaling” means to scale a variable to have values between 0 and 1.
“Standardised Scaling” means the data is transformed 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?

Ans: 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.

Ans: Q-Q plot is a tool used to determine if two data sets come from populations with a common distribution. It is useful in a case of linear regression when train and test data set are received separately and we need to confirm that both these datasets are from populations with same distribution. It can also be used with sample sizes and also can detect outliers if any, shifts in location, scale, and changes in symmetry.

Interpretations of Q-Q plot:

1. If all the points of quantiles lie in or close to straight line at an angle of 45 degrees from x-axis
2. If all points lie away from the straight line at an angle of 45degrees from x-axis
3. If y- quantiles < x-quantiles
4. If x-quantiles < y-quantiles, if any