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?**

   Based on the data we found 7 categorical variables, which was used for EDA.
   Here are the info based on our analysis:

   - **Season** (season): The demand for bikes was less during spring season, compared to other
   - seasons
   - **Year** (yr): The demand for bikes was quite high in 2019 compared to 2018, with a wider spread
   - **Month** (mnth): The demand for bike shows an increasing trend form Jan to May, followed by decreasing trend from Sep to Dec. And June to Aug can we terms as the peak periods
   - **Holiday** (holiday): Demand for bikes were quite high on holidays with a wide spread upto 9K, with a median of 4.5K approx
   - **Weekday** (weekday): weekday variable shows a very close trend with medians between 4K to 5K. This variable can have some or no influence on the predictor.
   - **Working Day** (workingday): Working day has no major impact on demand
   - **Weather Sit** (weathersit): Demand for weathersit rain_snow is low compared to others, which make sense. It can be a good factor

   Out these variables which was used for modeling based on the analysis were:
   Year, Season, Month, Weather Sit

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

   It is important to use drop_first=True because when we use get_dummy method, it creates all the categorical values for a feature as a new feature with values as 0 and 1. Which leads to creation on n new feature (where n is the number of categories) so it leads to a dummy variable trap. To get rid of that we use this (basically to get n-1 new features), where if all the 0 for n-1 features indicates the n^{th} features

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

   Based on the pair-plot among the numerical variables, we can say that the variables “temp” and “atemp” has the highest correlation with the target variable cnt (excluded casual and registered as cnt is the sum of these two)

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

   To validate the assumptions of linear regression after building the model, we used residual analysis.
   1. The residuals plot should follow a normal distribution with a mean equal to zero. So we plotted a distribution plot of residuals which followed the same behaviour.
   2. No auto correlation, so using Durbin Watson Test were value close to zero means no auto correlation
   3. Calculated VIF to make sure there is no multi-co-linearity.
   4. Ploted a QQ plot, where a straight line means that residual follows a normal distribution
   5. The residuals vs the predicted values should not follow any trend. So, based on that, there was no significant trend visible

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

   The top three features that contribute significantly are:

   - Temperature (temp): p-value is 0, and coeff is 0.3865
   - Year (yr): p-value is 0, and coeff is 0.2280
   - Rain or Snow (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds): p-value is 0, and coeff is -0.2306

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear Regression is a fundamental supervised learning algorithm used to model the relationships between two or more variables, where one is a dependent and others are independent. The goal is to find a best fit relationship for the dependant variable using the independent ones.

   To achieve this, there are mainly 7 assumptions taken into account for linear regression:

   a) Linear Relationship: There is a linear relationship between the dependant and independent variables, we can use a scatter plot to very this relation ship

   b) No Multicolinearlity: There should not be any correlation between any of the independent variables. As this can lead to redundancy in the dataset. We check the correlation matrix to verify this info.

   c) Homosedasticity: It states that the residuals that we get post building the model homogeneous and should follow a normal distribution. We can have a residual density curve to verify this

   d) No Autocorrelation in residuals: When residual are dependent on each other it leads to auto-correlation. And there should not be any pattern to it. We can use the residuals Vs predicted plot to check this

   e) Higher training data compared to Test data: The training data should always be more than the testing dat set. We can have a fair split like 70-30 or 80-20 or 60-40 based on the size of the entire data

   f) Each observation is unique: each observation is independent of the other observation. Meaning each observation in the data set should be measured separately as a unique occurrence of the event that caused the observation.

   g) Independent variables are normally distributed: It ensures that we have a equally distributed observations for each of the independent variables.

   There are mainly two types of linear regression:

   a) Simple Linear Regression: When we have one dependant and one independent variable

   b) Multiple Linear Regression: When we have one dependant and two or more independent variables

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical like 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.The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics.It was used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

3. **What is Pearson's R?**

   Pearson's R also known as Pearson Correlation Coefficient are used to measure how strong a relationship is between two variables. There are different types of formulas to get a correlation coefficient, one of the most popular is Pearson's correlation which is commonly used for linear regression. It is a statistical measure that calculates the strength and direction of the linear relationship between two variables denoted as **r**, where the value of r ranges from -1 to 1.

   Where:

   0 < r <= 1 indicates a positive linear relationship

   -1 =< r < 0 indicates a negative linear relationship

   r = 0 indicates no linear relationship

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

   Scaling is a data pre-processing step, which is applied to bring all the numerical values on a same scale. Which helps to speed up the calculations. In an actual use case the dataset will have multiple numerical variables and each of them can have a different scale.
   For example:We have a data set with variables A, B and C. With there ranges as.
   Variable A: 0 to 10
   Variable B: -100 to 100
   Variable C: 1000 to 100000
   It is important to bring al this values on same scale to to speed up and get accurate the model. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

   Normalized scaling brings all the data in the range of 0 to 1 while the standard scaling brings all the data to a normal distribution with mean 0 and standard deviation 1. Normalized scaling are sensitive to outliers where as standard scaling are less sensitive.

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

   It can happen due to multicolinearity (when independent variables are highly correlated with each other).
   As we the value of VIF is calculated using R2, to be precise, VIF = 1/(1-R2).
   So when the correlation is very high, which makes R2 = 1 and it lead to make the denominator 0 which in return gives as inf, which can be inffered as a very high values (hence very high multicolinearity)

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

   Q-Q plot (quantile-quantile plot) are graphical representation method for determining whether a dataset follows a certain probability distribution or not. Can also be used to check whether two samples comes from same distribution of not.
   It can be used to check whether the data is normally distributed or not. In linear regression we can plot the q-q plot for the residuals. And if it provides a straight line we can say that the residuals are normally distributed.