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?**

A: By performing EDA, we found that the categorical variables have an effect on our dependent variable 'cnt' which represents demand of bike sharing in the following ways:

a. **Season:** The demand (cnt) is lowest in Spring and increases in summer and fall and then again decreases slightly in winter. Now, this can be due to the fact that during spring weather is unpredictable with rain and windy conditions. The same can be said for winter due to cold weather and snowfall as well as less daylight hours. Summer and Fall on the other hand have stable weather conditions and promotes tourism and fun activities which in turn increases bike sharing demand.
b. **Year:** Compared to 2018 the demand has increased significantly in 2019. This may be simply due to the rising popularity of BoomBikes.
c. **Month:** The demand of bike sharing is low in January and it increases till September and again decreases till December. This is due to the seasonal change effect as explained earlier.
d. **Holiday:** The maximum demand is lower on a holiday compared to days which are not holiday. This may be due to the fact that people tend to be indoors on holidays.
e. **Weekday:** The demand throughout the week seems to be somewhat same however the spread slightly decreases till Thursday and increases on a Friday and stays somewhat same till Sunday. The increase in the lower spread of boxplot during Friday may be due to dip of demand in certain time/places and increase of demand in other time/places. In other words, the demand is a bit unpredictable and varies largely.
f. **Working day:** The demand is almost same for both working day and non-working day. Please note the non-working day also includes holidays.
g. **Weather conditions:** The demand is very low during Light Snow, Light Rain, Thunderstorm and high during a clear day. This is because people don't like to ride on a bad weather.

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

A: When we set drop_first=True during dummy variable creation, it drops the first category in each categorical variable. Now this is important because if we have let's say 3 categorical variables then creation of 2 dummy variables will be sufficient to represent all three categorical variables but by default get_dummies without drop_first=True will create 3 dummy variables. So, this way we can prevent redundancy and correlation 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?**

A: Looking at the pair-plot among the numerical variables we can observe that Temperature (temp) column has the highest correlation with the target variable.

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

A: We validated the assumptions of Linear Regression in the following ways:

a. **Distribution Plot (distplot) of Residuals:** Using this plot we found that residuals are almost centered around 0 and so the **residuals are following a normal distribution.**
b. **Residual vs fitted value(y_train_predicted) plot:** Using this plot we validated the assumption of **Linearity** because the residuals were randomly scattered around the horizontal axis with no clear pattern. Also, we validated the assumption of **homoscedasticity** because the spread of residuals across all fitted values were roughly constant which indicated there was no increasing or decreasing variance.
c. **VIF values**: Using the VIF values obtained from our final model, we found that all the VIF values were less than 5 which validated the assumption that there is no multicollinearity among the independent variables.

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

A: Based on our final model the top 3 features which contribute significantly towards explaining the demand of shared bikes are:

1. Temperature
2. Weather condition of Light Snow, Light Rain, Thunderstorm.
3. Windspeed

General Subjective Questions

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

A: Linear regression is a statistical technique that uses one or more independent variables also called predictor variables to predict their relationship with an output variable or dependent variable. The primary goal of linear regression is to understand and quantify how changes in the independent variables are associated with changes in the dependent variable.

A linear regression equation can be written in the form of

Y = B0 + B1*x1 + B2*x2 … +Bn*xn + e

Here,

Y : Dependent variable or response variable

B0 : Constant or intercept.

B1,B2,..,Bn : Coefficients of independent variables x1,x2,…,xn respectively

x1,x2,…,xn : Independent variables or predictor variables.

e : Error term or the variability in Y which cannot be explained by the predictor variables.

There are certain assumptions which must be satisfied for using linear regression model for prediction. These assumptions are:

- **Linearity:** There must be a linear relationship between predictor and response variable.
- **Homoscedasticity:** The residuals(y_actual – y_predicted) must be spread uniformly across the fitted values indicating constant variance.
- **Normality**: The residuals must follow a normal distribution.
- No Multicollinearity: There must be no correlation among the predictor variables.

Also, Linear regression is of 2 types:

a. **Simple Linear Regression:** When the response variable can be explained by only one predictor variable.
   In other words, the equation becomes:
   Y = B0 + B1*x1
b. **Multiple Linear Regression:** This involves more than one predictor variables.

2. Explain the Anscombe's quartet in detail.

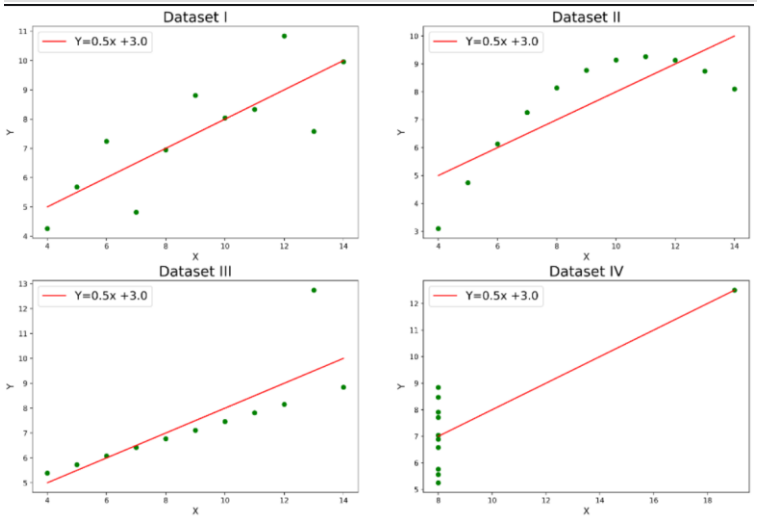
A: Anscombe's quartet is a set of 4 datasets in which each dataset contains 11 x-y pairs. These 4 datasets have identical descriptive statistics such as sum, mean, std deviation but all four show very different relationships between x and y when the datasets are plotted visually.

This quartet was developed by statistician Francis Anscombe in 1973 to show the importance of visualization in data analysis and that summary statistics alone can sometimes be not enough.

In the below images we can see that the summary statistics of the 4 datasets are almost identical however the x-y plots of each dataset show very different relationships between x and y

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II appears to have a nonlinear relationship between x-y.

- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier producing a high correlation coefficient.

3. What is Pearson's R?

A: Pearson's R or Pearson correlation coefficient R is a way in which we can measure the strength of linear correlation between two variables.

The value of Pearson's R lies between -1 to +1. The magnitude determines the strength and +/- sign determines the direction of correlation.

0 < R value <=1: In this case there is positive correlation between the two variables. In other words, if there is an increase in one variable the other variable also increases linearly and same when either one decreases. Also, a R value of 1 means perfect linear correlation and as the value decreases till 0, the strength of correlation will also decrease.

Ex: As inflation rises, Loan interest rates also rise.

R value = 0: In this case there is no correlation between the variables. Increase or decrease of one variable will not affect the other.

Ex: Height of a student and their GPA in exam.

-1<= R value < 0: In this case, there is negative correlation between the two variables. So if there is an increase in one variable there will be decrease in the value of the other and vice versa. Also R=-1 means perfect negative correlation and as R value increases from -1 till 0 the strength of correlation decreases.

Ex: Smoking and lung capacity. As a person increases his/her smoking frequency his/her lung capacity will decrease.

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

A: Scaling is a technique used to bring all features under a common scale so that the model does not favor higher scales over lower scales. This is also used so as to make sure that all the features contribute equally in the model's performance

Scaling is performed because of the following:

a. **Equal contribution:** Features with larger scales can dominate the model's learning process. Scaling ensures that this does not happen.

b. **Interpretability:** It is easier to interpret linear regression models when scaling is done as the coefficients reflect the importance of each feature much more accurately.

Differences between normalized scaling and Standardized:

Normalized Scaling	Standardized Scaling
Used when features are of different scales.	Used when we want 0 mean and unit standard deviation
Scales features between [0,1] or [-1,1]	There is no such bounded range in standardized scaling.
Gets affected by outliers	Much less affected by outliers.
Uses Minimum and Maximum values of features to scale.	Uses mean and standard deviation to scale

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

A: VIF (Variance Inflation Factor) calculates how well one independent variable is explained by all the other independent variables combined.

Now VIF is given by:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

Now VIF value can sometimes become infinite because of something called perfect multicollinearity. Perfect multicollinearity occurs when a predictor variable is an exact linear combination of one or more other predictor variables. In this case, R-i-th Square value becomes 1, leading to the denominator of the VIF formula becoming zero, which results in an infinite VIF value.

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

A: A quantile-quantile (Q-Q) plot is a visual tool used to assess if a dataset adheres to a specific probability distribution or to compare if two data samples originate from the same population. This type of plot is especially beneficial for verifying if a dataset is normally distributed or follows another

recognized distribution. Q-Q plots are frequently utilized in statistics, data analysis, and quality control to validate assumptions and detect deviations from anticipated distributions.

Uses:

- **a.** Assess whether the residuals from a regression model are normally distributed.
- **b.** Compare the distribution of a sample dataset to a theoretical distribution (e.g., normal, exponential) or another sample dataset.
- **c.** Assess how well a theoretical distribution fits the observed data.

Importance:

a. Can be used to validate the assumption that residuals (errors) in linear regression are normally distributed.
b. Q-Q plots can detect patterns in residuals that indicate heteroscedasticity (non-constant variance), which violates another key assumption of linear regression.
c. Q-Q plots provide a visual tool to validate the linear regression model. Deviations from the expected straight line suggest potential issues with the model's assumptions.