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?

    - The demand of renting shared bikes is more in fall and summer season compared to spring and winter.
    - There is increasing trend of renting bikes from January to July and decreasing trend from July to December.
    - The demand of renting shared bike is almost same in both working and non-working day.
    - The demand of renting bike is increased in 2019 as compared to 2018.
    - The demand of bike throughout the weekdays is almost same.
    - people tend to rent bike in clear sky more compared to light snow

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

    The drop_first parameter reduces the **extra unnecessary column** creation during dummy variable creation. The column is unnecessary because when all the other columns are zero that automatically means the first column is 1 and this ultimately **minimize multicollinearity** in the model.

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

    In my case, "**atemp**" (feeling temperature in Celsius) is having highest correlation with the target variable.

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

    I. Assumption: The dependent variable (target variable) and independent variable (predictors) has a linear relationship.
    This is validated by plotting actual data points versus predicted data points by model on the scatter chart.

    II. Assumption: The residuals (error terms) are independent of each other (Little or No autocorrelation in the residuals)
    This is validated using "**durbin_watson**" test

    III. Assumption: Homoscedasticity

This is validated by plotting **residual values (error terms) vs predicted values** on to scatter plot.

IV. Assumption: Little or no Multicollinearity between the features
This is validated using **VIF** (Variance Inflation Factor) method.

V. Assumption: Distribution of residuals (error terms) forms normal distribution.
This is validated using **histogram** of residual values.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - atemp (feeling temperature in Celsius)
    - light_snow
    - yr (year)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

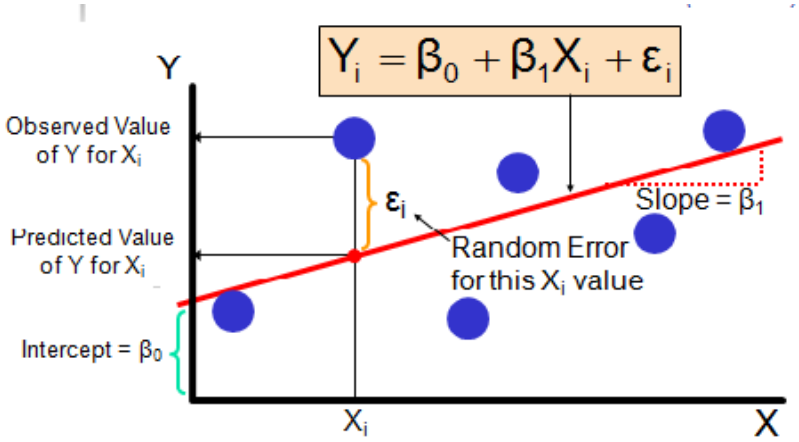
   It is Machine Learning Algorithm based on supervised learning that is statistical way of measuring the relationship between one or more independent variables versus one dependent variable. It is used to predict continuous variable. There are two types of linear regression:
   A. **Simple Linear Regression**: Regression for 1 independent and 1 dependent variable.
   B. **Multiple Linear Regression**: Regression for > 1 independent and 1 dependent variable.

   Linear Regression Equation for bet fit line.

   Line of best fit refers to al line through a scatter plot of data points that best expresses the relationship between those points and minimizes the distance between the line and data points. Ordinary Least Square method (OLS) is being used to arrive at the equation for the line.

   Simple Linear Regression: $y = \beta 0 + \beta 1 * x + \epsilon$

   y = predictor
   $\beta 0$ = y intercept
   $\beta 1$ = slope



Assumptions in Simple Linear Regression:

1. The dependent variable (target variable) and independent variable (predictors) has a linear relationship.
2. Distribution of residuals (error terms) forms normal distribution
3. Homoscedasticity
4. The residuals (error terms) are independent of each other

Residuals:

It is a measure of how far away an observed data point is vertically from the regression line (predicted y value).

Residual (ε)= observed y value – predicted y value.

If residual value is negative that means the predicted value is high and positive residual value says predicted value is low. The aim of a regression line is to minimize the sum of residuals.

Ordinary Least Square Method (OLS):

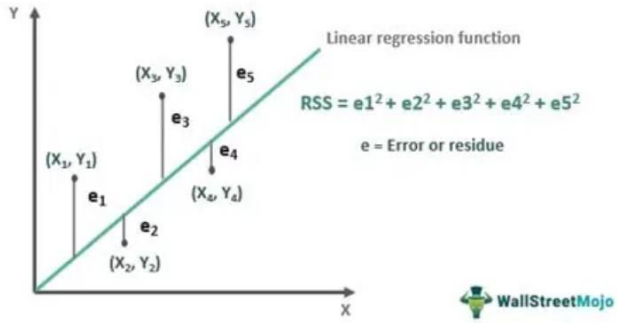
The OLS method is used to estimate β0 and β1 coefficients. The OLS method used to minimize the sum of the squared residuals.

$\beta 1$ (Slope) = $\sum (X - \bar{X}) * (y - \bar{y}) / \sum ((X - \bar{X})^2$

$\beta 0$ (Intercept) = $\bar{y} - (\beta 1 * \bar{X})$

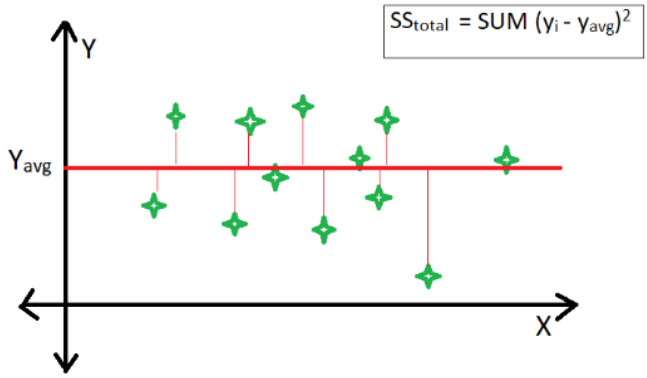
Residual Sum of Squares (RSS):

It is a statistical method used to measure the deviation in a dataset not explained by the regression model. The smaller the RSS the better your model fits your data and the greater the RSS, the poorer your model fits your data.



Total Sum of Squares (TSS):

It is squared difference between the observed dependent variable and its mean. This is nothing but a dispersion of the observed variables around the mean.



$\sum (y - \bar{y})^2$

Strength of Simple Linear Regression:

R^2 is used to measure strength of the linear regression model. R^2 is a statistical measure that represents the goodness of fit of regression model. The ideal value for R^2 is 1. The closer the value to 1, the better is the model fitted.

R^2 = 1 – (RSS / TSS)

Ways to Minimize Cost Function (RSS):

- Differentiation
- Gradient Descent Approach: It is an iterative first-order optimization algorithm used to find a local minimum / maximum of a given function. It is used in ML to minimize a cost function in a linear regression.

Multiple Linear Regression:

Multiple linear regression is needed when one variable is not sufficient to create a good model to make accurate predictions.

$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots \beta p * Xp + \epsilon$

- The model fits a hyperplane instead of a line.
- The coefficients are still obtained by minimizing the sum of squared errors, least squares criteria.
- All the assumptions from simple linear regression still hold for multiple linear regression

The new aspects to consider when moving from simple to multiple linear regression are as follows:

Multicollinearity: If the model is built using several interrelated independent variables that affects interpretation of the model and reliability of p-value. Using Variance Inflation Factor (VIF) and pair plot of independent variables, multicollinearity is detected.

VIF = 1 / 1 – R^2

VIF > 10: then variable should be eliminated

VIF > 5: can be okay, but it is worth inspecting

VIF < 5: VIF is good, no need to eliminate the variable.

Feature Selection: Can be used mixed approach – automated (Recursive Feature Elimination (RFE) + manual dropping insignificant variables based on VIF and p-value.

Adjusted R^2: The value of R^2 always increases or remains the same as new variables are added to the model, without detecting the significance of the newly added variable. As a result, non-significant attributes can also be added to the model. So, using adjusted R^2, we can determine whether adding new variables to the model increases to model fit.

Adding a random independent variable did not help in explaining the variation in the target variable, Adjusted R^2 value decreases that indicates the variable is insignificant for the model.

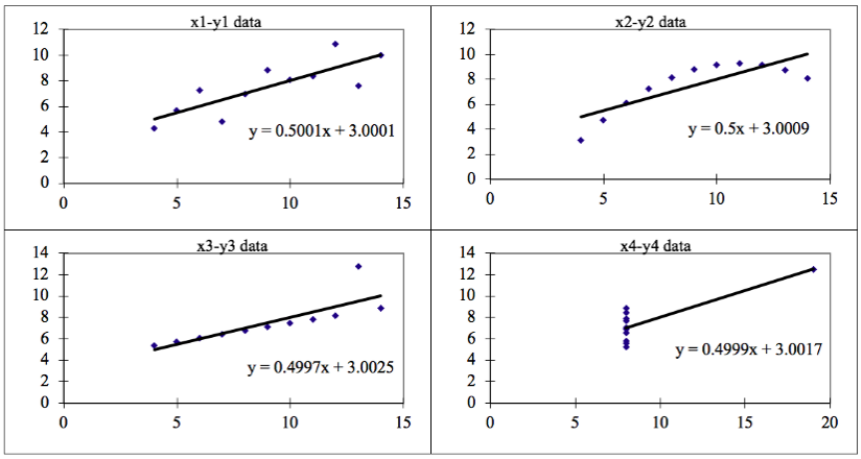
1 – (1 – R^2) (n – 1) / (n – k – 1)

- n represents the number of data points in our dataset
- k represents the number of independent variables, and
- R represents the R-squared values determined by the model.

2. Explain the Anscombe's quartet in detail
   Anscombe's Quartet describes the importance of data visualization before building a well-fit model. This was developed by the statistician Francis Anscombe in 1973. Anscombe's

Quartet comprises four datasets that have nearly identical simple descriptive statistics such as mean, variance, standard deviation etc., yet they have very different distributions and graphical representation that fools the regression model if built.

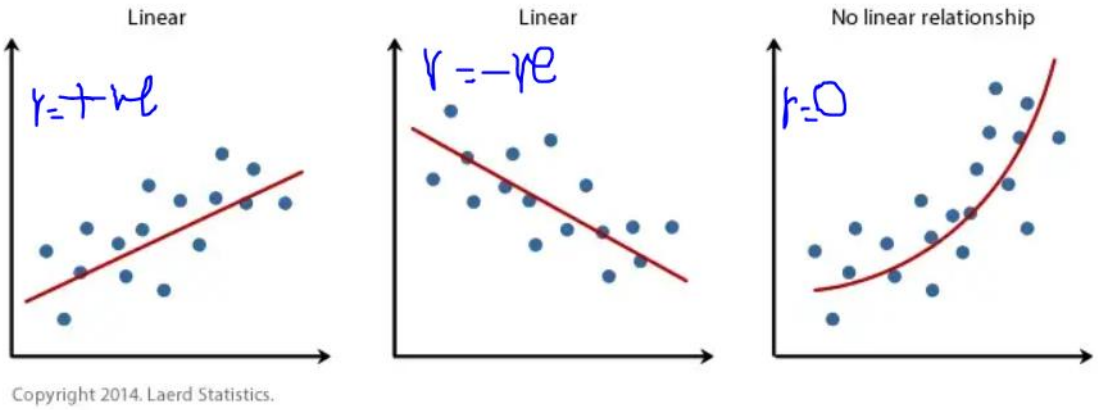
Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3	x4	y4
1	10	8.04		10	9.14		10	7.46	8	6.58
2	8	6.95		8	8.14		8	6.77	8	5.76
3	13	7.58		13	8.74		13	12.74	8	7.71
4	9	8.81		9	8.77		9	7.11	8	8.84
5	11	8.33		11	9.26		11	7.81	8	8.47
6	14	9.96		14	8.1		14	8.84	8	7.04
7	6	7.24		6	6.13		6	6.08	8	5.25
8	4	4.26		4	3.1		4	5.39	19	12.5
9	12	10.84		12	9.13		12	8.15	8	5.56
10	7	4.82		7	7.26		7	6.42	8	7.91
11	5	5.68		5	4.74		5	5.73	8	6.89
				Summary Statistics						
N	11	11		11	11		11	11	11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94	3.16	1.94
r	0.82			0.82			0.82		0.82	



Here, the summary statistics are misleading so visualizing the data allows us to revisit our summary statistics. Anscombe's Quartet suggest that the data must be plotted in order to see the distribution, identify outliers, visualize diversity of the data and data linearity.

3. What is Pearson's R?

Pearson's R is the most popular type of correlation coefficient. Correlation coefficients are used to measure the strength of association between two variables and the direction of the relationship. It is commonly used in linear regression.
Correlation coefficient values varies between +1 and -1. A value of 1 indicates a perfect degree of association between two variables. A value of 0 indicates there is no linear correlation. The direction of the relationship is indicated by the sign of the coefficient e.g., a `+` sign indicates a positive relationship and `−` sign indicates a negative relationship.



There are few assumptions in the Pearson's R correlation:

1. Both variables are normally distributed
2. Both variables are not having significant outliers.
3. Both variables are continuous
4. Both variable follow linear relationship
5. Homoscedasticity

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

It is one of the most important data pre-processing steps in machine learning applied to all numeric variables (not binary categorical variables) to normalize the data in single unit or range.

Gradient Descent optimization technique require data to be scaled. The difference in ranges of features will cause different step sizes for each feature so to ensure that the gradient descent moves smoothly towards the minima and the learning rate are updated at the same rate for all the features. It also helps in speeding up the calculations in an algorithm.
Scaling is required to build the correct model by bringing all the variables to the same level of unit.
The most popular feature scaling techniques are Normalization and Standardization.
Normalization: It also known as Min-Max Scaling. This scales the data in the range between 0 and 1 or -1 and 1. It is being used when features are of different scales.

$X\_new = (X - \min(X))/ (\max(X) - \min(X))$

Minimum and Maximum value of features are used for scaling.
It is affected by outliers. It is useful when we don't know about the distribution.

Standardization: It also known as Z-score Normalization. It brings all the data into a standard normal distribution which as mean 0 and standard deviation 1. This technique is helpful in cases where the data follows a Normal Distribution. It is being used to ensure zero mean and 1 standard deviation.

X_new = (X - Mean)/SD

Mean and Standard Deviation is used for scaling. The scale values are not bounded to a certain range. It is much less affected by outliers.

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

   If there is perfect correlation between two independent variables where R^2 = 1, then VIF is infinity because VIF = 1 / (1 – R^2). To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   It is also known as Quantile-Quantile plot. It is nothing but a scatter plot created by plotting 2 different quantiles against each other. It is graphical tool for determining if two datasets come from populations with a common distribution. Using Q-Q plot, we plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
   It helps us to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
   Q-Q plot helps to verify one of the linear regression assumptions – residual follow normal distribution.

   Q-Q plot from normal distribution