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)

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

If you don't drop the first column then your dummy variables will be. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance's may be distorted. If you have a small number of dummies, I suggest removing the first dummy. For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1then the person is a male and if male=0then the person is female. However, if you have a category with hundreds of values, I suggest not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible)

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

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

- Train R^2 : 0.826

- Train Adjusted R^2 : 0.82

- Test R^2: 0.8115

- Test Adjusted R^2: 0.790564

- Difference in R^2 between train and test: 1.5%

- Difference in adjusted R^2 between Train and test: 3.15% which is less than 5%

Yes! Its a best model

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

- We arrived at a very decent model for the demand for shared bikes with the significant variables

- We can see that temperature variable is having the highest coefficient 0.4914, which means if the temperature increases by one unit the number of bike rentals increases by 0.4914 units.

Similarly we can see coefficients of other variables in the equation for best fitted line.

We also see there are some variables with negative coefficients, A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease. We have spring, mist cloudy, light snow variables with negative coefficient. The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant

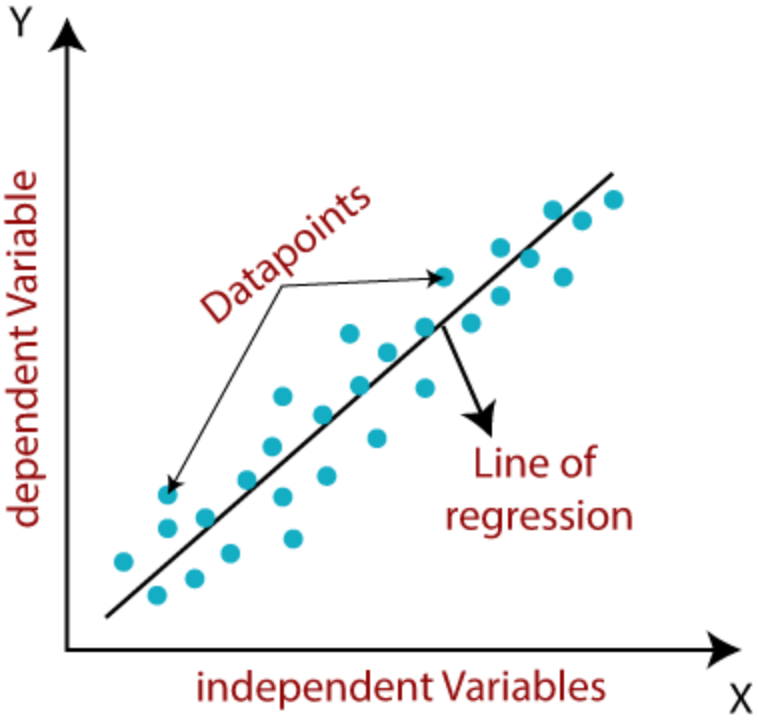
General Subjective Questions

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

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$$y= a_0+a_1x+ \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a

numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. Explain the Anscombe's quartet in detail. (3 marks)
   Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
   Simple understanding:
   Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

3. After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

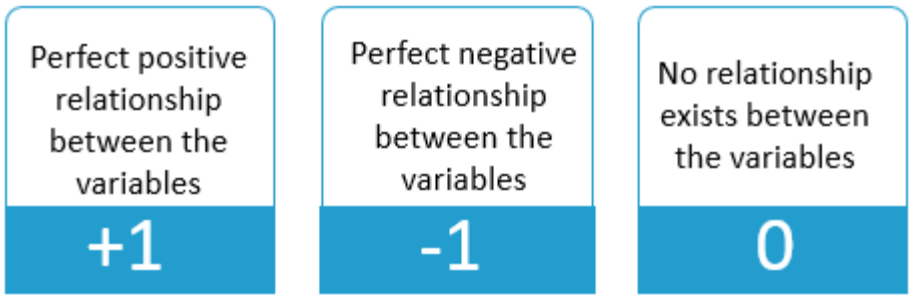
4. What is Pearson's R? (3 marks)

Pearson correlation coefficient, also known as Pearson R statistical test, measures strength between the different variables and their relationships. Whenever any statistical test is conducted between the two variables, then it is always a good idea for the person doing analysis to calculate the value of the correlation coefficient for knowing that how strong the relationship between the two variables is.

Pearson's correlation coefficient returns a value between -1 and 1. The interpretation of the correlation coefficient is as under:

- If the correlation coefficient is -1, it indicates a strong negative relationship. It implies a perfect negative relationship between the variables.
- If the correlation coefficient is 0, it indicates no relationship.
- If the correlation coefficient is 1, it indicates a strong positive relationship. It implies a perfect positive relationship between the variables.

A higher absolute value of the correlation coefficient indicates a stronger relationship between variables. Thus, a correlation coefficient of 0.78 indicates a stronger positive correlation as compared to a value of say 0.36. Similarly, a correlation coefficient of -0.87 indicates a stronger negative correlation as compared to a correlation coefficient of say -0.40.



Values of the Correlation Coefficient

In other words, if the value is in the positive range, then it shows that the relationship between variables is correlated positively, and both the values decrease or increase together. On the other hand, if the value is in the negative range, then it shows that the relationship between variables is correlated negatively, and both the values will go in the opposite direction.

Pearson Correlation Coefficient Formula

Pearson's Correlation Coefficient formula is as follows,

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
6. Scaling is a personal choice about making the numbers feel right, e.g. between zero and one, or one and a hundred. For example converting data given in millimeters to meters because it's more convenient, or imperial to metric.

While normalisation is about scaling to an external 'standard' - the local norm - such as removing the mean value and dividing by the sample standard deviation, e.g. so that your sorted data can be compared with a cummulative normal, or a cummulative Poisson, or whatever

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

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

Looking at the equation above, this happens when R^2 approaches 1.
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).

Here's the formula for calculating the VIF for X_1:

$$VIF_1 = \frac{1}{1 - R^2}$$

R^2 in this formula is the coefficient of determination from the linear regression model which has:

- X_1 as dependent variable
- X_2 and X_3 as independent variables

In other words, R^2 comes from the following linear regression model:
$X_1 = \beta_0 + \beta_1 \times X_2 + \beta_2 \times X_3 + \varepsilon$

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

Q-Q plot stands for a "quantile-quantile plot".
It is a plot where the axes are **purposely transformed** in order to make a **normal (or Gaussian) distribution appear in a straight line**. In other words, a perfectly normal distribution would exactly follow a line with slope = 1 and intercept = 0. Therefore, if the plot does not appear to be - roughly - a straight line, then the underlying distribution is not normal. If it bends up, then there are more "high flyer" values than expected, for instance.
Importance in Linear Regression. Quantile-Quantile (Q-Q) plot, is a graphical tool for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution. This helps in a scenario of linear regression when we have the training and test data set

received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.