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)**

ANSWER:
I have analyzed season, month, weather and weekdays categorical variables with target variable and identified the following:

Season:

- Highest average ride counts under Fall season followed by Summer and Winter seasons.
- Spring season has the least average ride count.

Month:

- There is a sudden increase in average count after April and maintaining consistently from May to October and then falls in November.
- We have less average ride count for January and February months compared with other months.

Weather:

- We have very good average ride count in Clear and Mist weathers and there is a huge dip in light rainy weather.
- There are no ride counts present in heavy rain weather which makes sense.

Weekdays:

- Here we can see that average ride count is higher in weekends (Saturday and Sunday) and on Thursday and Friday.
- We have less ride count on Monday, Tuesday and Wednesday compared with other days but still not much significant difference.
- We can safely say that people prefer renting on weekends.

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

ANSWER:
When you have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels so that n-1 variables can be able to explain all the levels.
drop_first=True is important to use because it helps to reduce correlation amongst dummy variables which may lead to multicollinearity during model build. So, it is always recommended to drop one variable and maintain n-1 variables for n levels.

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

ANSWER:
'Temp' independent variable has the highest correlation with target variable with value of 0.63. It has good positive correlation with target variable.

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

ANSWER:

I have checked the following:

- Error terms are normally distributed with centered to 0 as mean.
- Homoscedasticity – Cluster of data points are scattered with same width and of same variance.
- R-Squared and Adjusted R-Squared – Contains good accuracy of data prediction (0.836 and 0.805 R2 scores for train and test data sets respectively).
- Mean Squared Error: 0.08 and 0.09 for train and test data sets respectively which are very closer to zero.

Based on the above validations, we can say that the model isn't fit by chance and well generalized for prediction.

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

ANSWER:

The top three features identified from my model built are:

1. **'temp'** with coefficient as 0.491508.
2. **'year'** with coefficient as 0.233482.
3. **'Winter'** (derived from 'season' as dummy variable) with coefficient as 0.083084.

General Subjective Questions

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

ANSWER:

Linear regression is one of the statistical models that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slope/coefficient of the regression line which represents the effect X has on Y

c is a constant value, known as the Y-intercept.

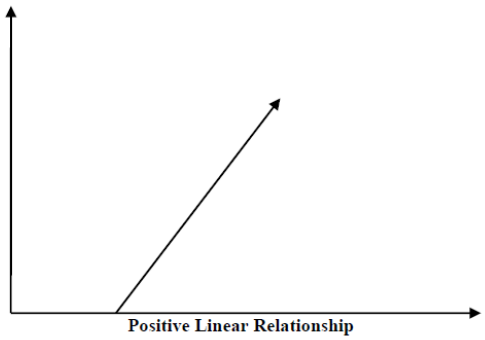
In Machine learning we used to represent as:

$$Y = b0 + b1X$$

Where b0 is a constant value, known as the Y-intercept and b1 is slope/coefficient of the regression line.

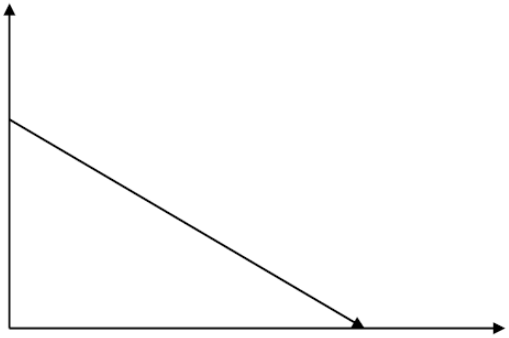
There are two types of Linear relationships:

1. **Positive Linear Relationship:** If a target variable increases with an increase in independent variable, then it is called Positive linear relationship.



Positive Linear Relationship

2. **Negative Linear Relationship:** If a target variable decreases with an increase in independent variable, then it is called Negative linear relationship.



Negative Linear Relationship

Types of Linear Regression:

Linear Regression is of two types:

1. **Simple Linear Regression:**

   When the model built using only one independent variable to predict target variable then it is called Simple Linear Regression.

   Linear Equation: **Y = b0 + b1X**

   Where b0 is a constant value, known as the Y-intercept and b1 is slope/coefficient of the regression line.

   X – Independent variable and Y – Target variable

2. **Multiple Linear Regression:**

   When the model built with more than one independent variable to predict target variable then it is called Multiple Linear Regression.

   Linear Equation: **Y = b0 + b1X1 + b2X2 + b3X3 + …. + bnXn**

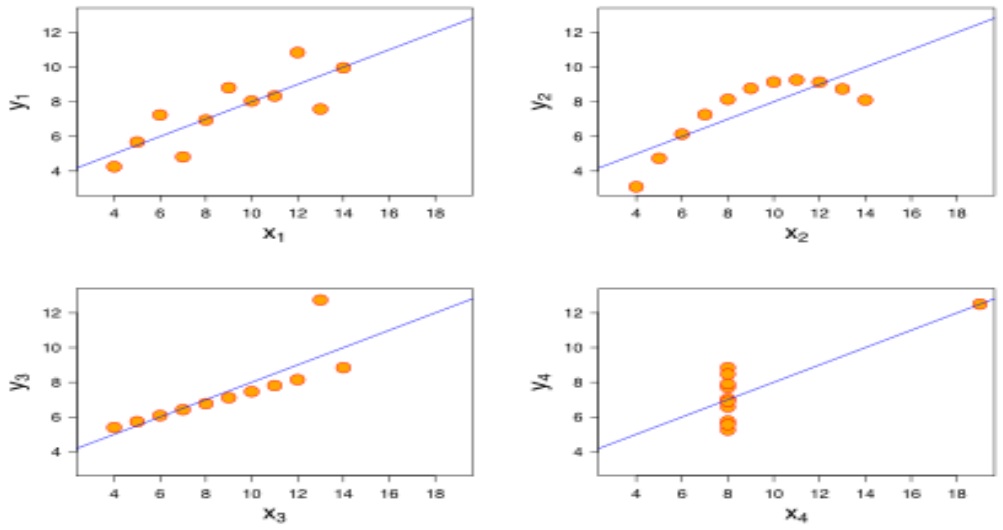
   Where b0 is constant value and b1, b2, b3…. are coefficients.

   X1, X2, X3…. are independent variables and Y is a Target variable.

2. **Explain the Anscombe's quartet in detail.** **(3 marks)**

ANSWER:

Anscombe's quartet comprises four data sets with nearly identical simple descriptive statistics, but they have unique distributions and appear very different when graphed. Let's say each dataset consists of eleven (x, y) points and the respective scatterplots are shown below.



- **Dataset-1 (X1, Y1):** The scatter plot appears to be a simple linear relationship.
- **Dataset-2 (X2, Y2):** The second graph while a relationship between the two variables is not a linear relationship.
- **Dataset-3 (X3, Y3):** Perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- **Dataset-4 (X4, Y4):** This shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

ANSWER:

Pearson's R is a statistical measure that defines correlation (or) linearity between two numerical variables. Correlation value ranges between -1 and +1.

Mathematical formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores (x and y variables)

Σx = the sum of x variable data points

Σy = the sum of y variable data points

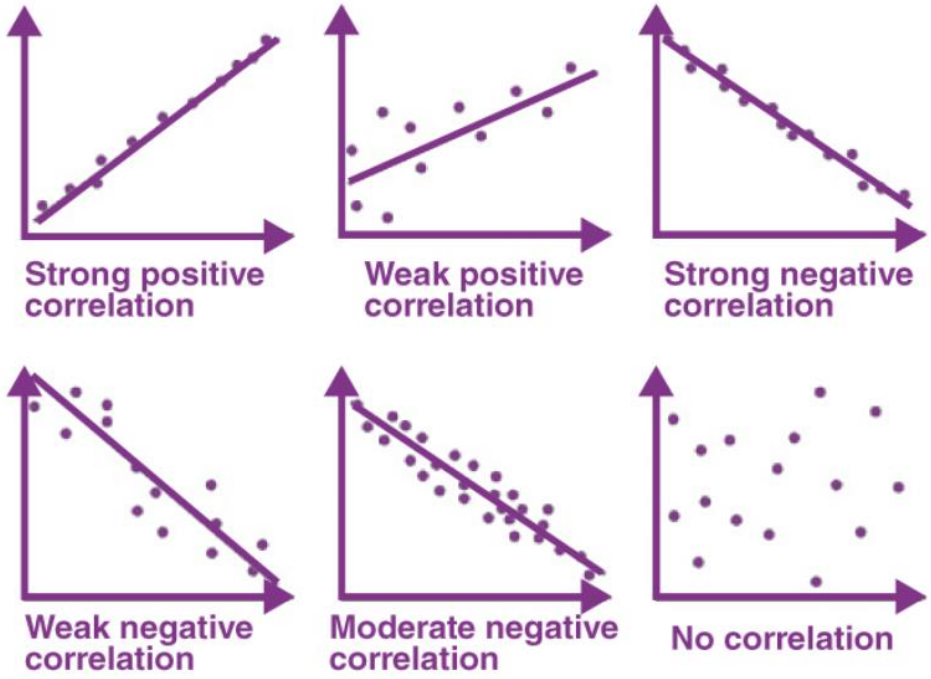
Σx2 = the sum of squared x data points

Σy2 = the sum of squared y data points

In data science, we use either scatterplot or regplot (linear regression model fit) to check the linear relationship between two numerical variables.

There are three types of correlation:

1. **Positive Correlation:** This occurs when two variables move in same direction i.e., both X and Y increases. Any value nearer to +1 represents highly positive correlation
2. **Negative Correlation:** This occurs when two variables move in opposite direction i.e., when X increases then Y decreases. Any value nearer to -1 represents highly negative correlation.
3. No Correlation: This represents that there is no linear relationship between two numerical variables. Correlation value will be 0.



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

ANSWER:

Scaling is a data pre-processing step to standardize the data for all the numerical variables to bring and contain same units and range across the variables.

Scaling of variables are important before we proceed for model building. This is because let's say in the source dataset we have different units for each variable and when our model try to predict the outcome it usually checks the value of independent variable but not it's units which may end up predicting wrong value and produce incorrect result. So, in this case we need to standardize the data for all the variables to the same unit and then proceed with model building which yields better model creation and prediction accuracy will be high.

There are two types of Scaling:

1. **Normalization/Min-Max Scaling:** It brings all the data in the range of 0 and 1 based on min and max value from the variable.

Min-Max Scaling: x = (x – min(x)) / (max(x) – min(x))

2. **Standardization Scaling:** *It brings all the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

Standardization: x = (x – mean(x)) / std(x)

Difference between Normalization and Standardization Scaling:

Normalization Scaling	Standardization Scaling
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is really affected by outliers.	It is much less affected by outliers.
It is useful when we don't know about the distribution.	It is useful when the feature distribution is Normal or Gaussian.

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

ANSWER:

Variance Inflation Factor (VIF) is a statistical measure that checks the severity of multicollinearity in the regression analysis. It calculates how well one independent variable is explained by all the other independent variables combined.

VIF Formula:

$$VIF = 1/(1-R2)$$

where R2 denotes the correlation between independent variables.

R2 = 1 represents perfect correlation. If there is a perfect correlation, then VIF becomes Infinite which also indicates of perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

To solve this problem, we need to drop one of the independent variables from the dataset which causes this perfect multicollinearity.

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

ANSWER:

Q-Q (Quantile-Quantile) plot is a probability plot that helps us to assess if a set of data that comes from theoretical distributions such as Normal, Exponential or Uniform distributions. It takes two data sets and compare their probability distributions by plotting their quantiles against each other.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model.

Interpretation:

Below are the possible interpretations for two data sets.

- **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.

- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.
- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.
- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.

Q-Q plots are also used to find the skewness of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution, we want to know on the y-axis then we see a very peculiar shape of a Normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is left-skewed (negatively skewed) but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is right-skewed (positively skewed).