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?**
    - Season fall has seen the highest usage of bikes, followed by summer
    - Users increased in the year 2019 significantly
    - Month of Sept has more users compared to other months
    - There are more users when weather is clear
    - Day of the week does not have a major impact

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

For a categorical variable with n different values, n-1 dummy variables are enough. Since we don't need extra col, we can drop that.

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

Registered has the highest correlation with cnt

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

- Checked the R2 value to see if the variance in data is explained properly
- Checked the VIF to see if there is any multi collinearity
- Checked the F statistic to see the extent of the fit
- Checked the p-values of all variables to see their significance
- Checked the adjusted r2 to see the impact of multiple variables as it is multi linear regression

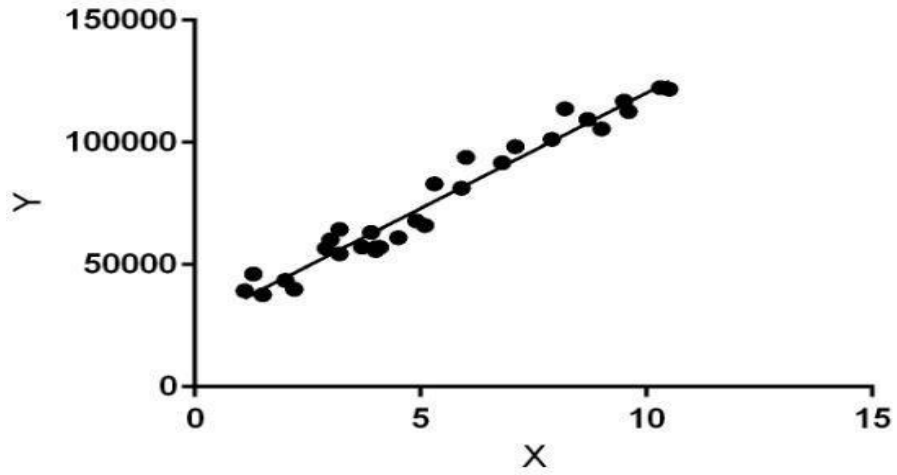
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Atemp
- Hum
- casual

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



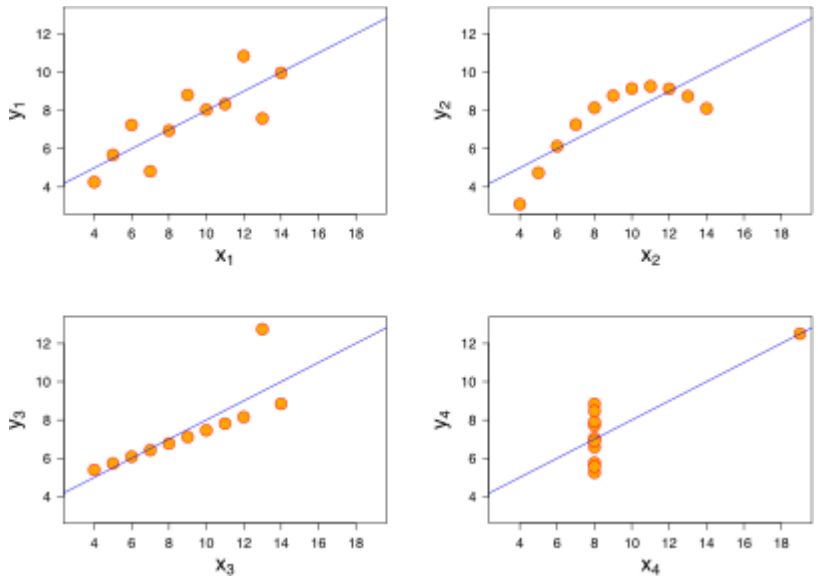
Linear regression performs the task to predict a dependent variable value y based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics , yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*, *y*) points.



All the summary statistics you'd think to compute are close to identical:

- The average *x* value is 9 for each dataset
- The average *y* value is 7.50 for each dataset
- The variance for *x* is 11 and the variance for *y* is 4.12
- The correlation between *x* and *y* is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation *y = 0.5x + 3*
  So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the above Picture results

Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV looks like *x* remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

3. What is Pearson's R?

Pearson`s Correlation coefficient (R) is a measure of the strength of the relationship between variables

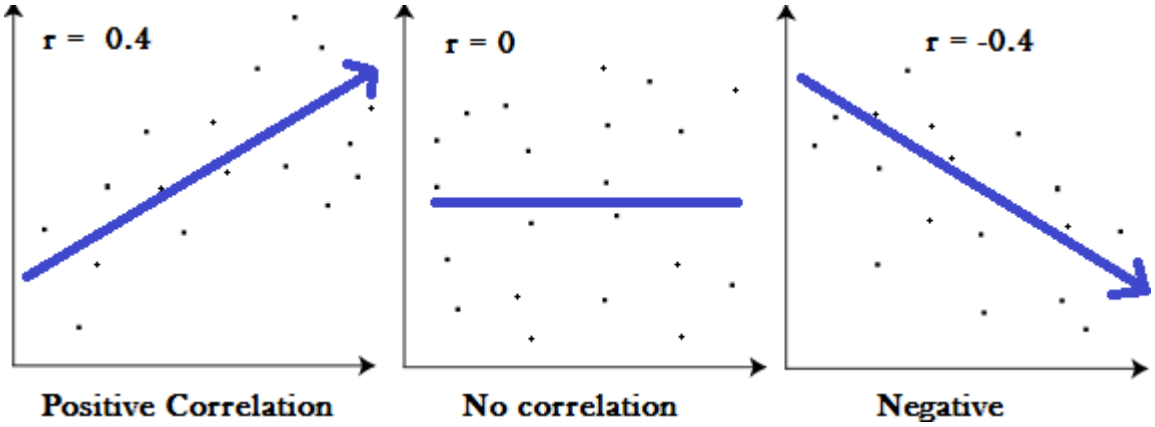
To get relationship between continuous variables is to draw a scatter plot of the variables to check linearity. The correlation coefficient should not calculated if the relationship is not linear

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

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

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

a. 1 indicates a strong positive relationship.
b. -1 indicates a strong negative relationship.
c. A result of zero indicates no relationship at all.



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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

Reasons to perform scaling are:

1. Ease of interpretation.
2. Faster convergence of gradient descent methods.

Scaling affects the coefficients
There are two types of scaling 1. Min-max scaling 2. Standardisation scaling.

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

A variance inflation factor (VIF) detects multicollinearity in regression analysis. Multicollinearity is when there is correlation between predictors (i.e. independent variables) in a model; it is presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

The thumb rule for interpreting VIF

a. Between 1 and 5 = moderately correlated.
b. Greater than 5 = highly correlated.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

1. It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

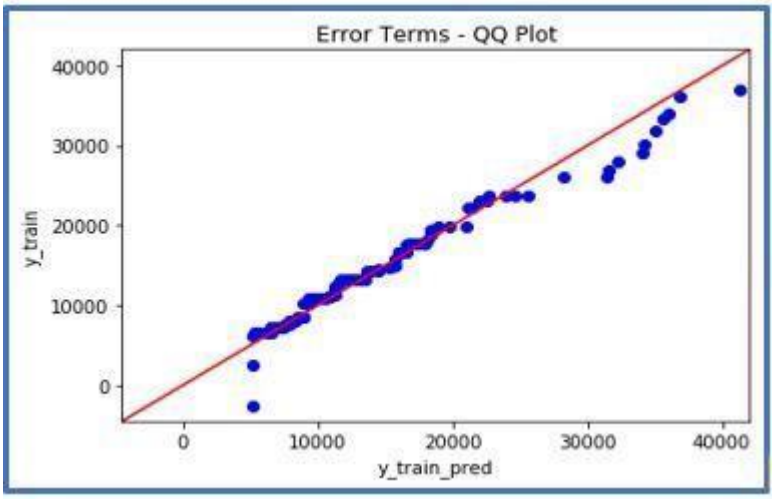
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

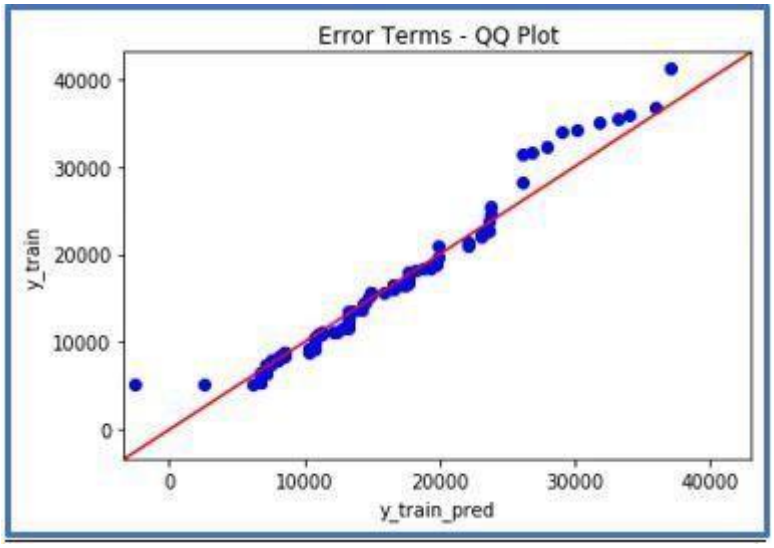
Below are the possible interpretations for two data sets.

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

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis

Ref: wikipedia