Assignment-based Subjective Question

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

From the analysis on categorical variable, we can clearly see that certain variables have drastic effect on the target variable.

1. season: Majority of the bookings were happening on season3 with the median of over 5000 booking followed by season2 with median of around 5000. This indicates that the season can be a good dependent variable.

2. mnth: We can see that there is a trend for this variable and in the month 6,7,8,9 and 10 the median is around 5000. So, this variable can be a good dependent variable.

3. weathersit: Around 65% of the booking were happening during weather1 followed by weather2 with the median of 5000 and 4000 respectively. Thus, we can say that weathersit can be a good dependent variable.

4. holiday: Almost 95% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday cannot be a good predictor for the dependent variable.

5. workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

This is done to prevent multicollinearity in the dataset. It occurs when variables are highly correlated to each other.

The reason we drop the first dummy variable is because we can predict each outcome with n-1 where n is the total number of dummy variable.

Let's say for example we have a dummy variable furniture_status which has 3 values which are furnished, semi-furnished and unfurnished. Now, when we create dummy variable, we create 3 new columns. But in this case, we can remove any one of the columns still we would be able to analyze the data.

1 0 → semi furnished

0 1 → unfurnished

0 0 → furnished

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

atemp has highest correlation with the target variable cnt.

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

I validated the assumptions of linear regression using the residual analysis which can tell us whether the residuals have mean close to zero. A non-zero mean could indicate that there's a problem with the model.

Also, I checked the multicollinearity by checking the VIF values. High VIF (typically above 10) indicates multicollinearity issue.

In the end I created scatter plot to test the linear relation between X_train and X_train_pred.

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

temp, season_4 and year are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Question

Explain the linear regression algorithm in detail?

Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.

It can extend to multiple linear regression involving several independent variables and logistic regression, suitable for binary classification problems

Linear regression is basically divided into two categories.

1. **Simple linear regression**: Here we have only one dependent and one independent variable.
2. **Multiple Linear regression**: Here we can have multiple dependent variables.

Linear Regression model representation

Linear regression is such a useful and established algorithm, that it is both a statistical model and a machine learning model. Linear regression tries a draw a best fit line that is close to the data by finding the slope and intercept.

Linear regression equation is,

Y=a+bx

In this equation:

- y is the output variable. It is also called the target variable in machine learning or the dependent variable.
- x is the input variable. It is also referred to as the feature in machine learning or it is called the independent variable.
- a is the constant
- b is the coefficient of independent variable

Multiple linear regression

- Multiple Linear Regression assumes there is a linear relationship between two or more independent variables and one dependent variable.
- The Formula for multiple linear regression:
- Y=B0+B0X1+B2X2+……+BnXn+e**Y** = the predicted value of the dependent variable
    - **B0** = the y-intercept (value of y when all other parameters are set to 0)
    - **B1X1**= the regression coefficient (B1) of the first independent variable (**X1**)
    - **BnXn** = the regression coefficient of the last independent variable
    - **e** = model error

The multiple linear regression model can be represented as a plane (in 2-dimensions) or a hyperplane (in higher dimensions).

Explain the Anscombe’s quartet in detail

Anscombe's Quartet is a clever set of four datasets designed by a statistician named Francis Anscombe. What's fascinating about these datasets is that they all share the same basic numerical properties—like averages and correlations—yet they look completely different when you plot them.

Each dataset has two sets of numbers (X and Y), and despite their similar arithmetic characteristics, they form very distinct patterns when graphed. Some look like simple straight lines, others show curves, and one even has a clear outlier.

This challenges the notion that relying solely on summary statistics, which are those numerical averages and measures, might not give you the whole picture.

The point Anscombe wanted to make is that looking at data visually, in addition to crunching the numbers, is crucial. Different datasets can have identical statistical summaries but tell very different stories when you actually see them in a graph.

It's a reminder to not blindly trust numbers and always explore your data visually to capture its full story.

Imagine you have four sets of student data with nearly identical averages and correlations between study hours and exam scores.

Sounds similar, right? But when you plot the data, one group may show a linear relationship, another a curve, and one might have a standout outlier.

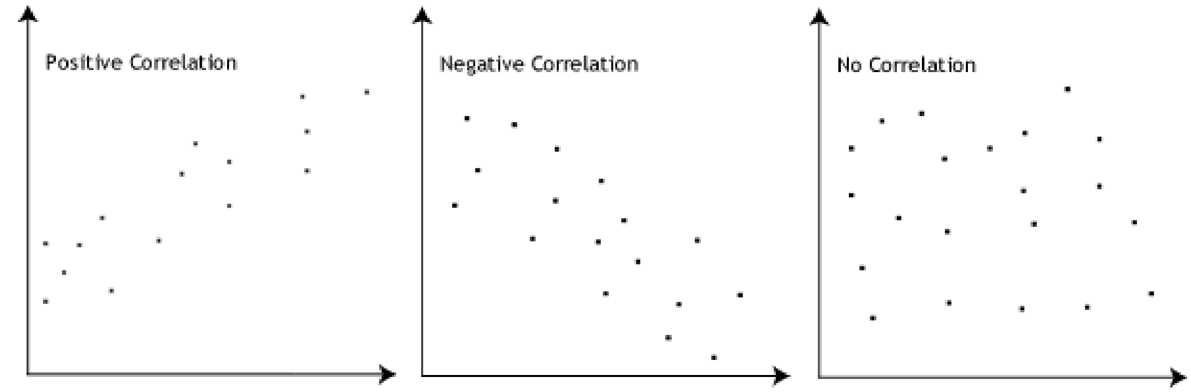
This is Anscombe's Quartet, teaching us that visualizing data is crucial—similar numbers can tell different stories, emphasizing the importance of not relying solely on statistics but also exploring data visually.

What is Pearson's R

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

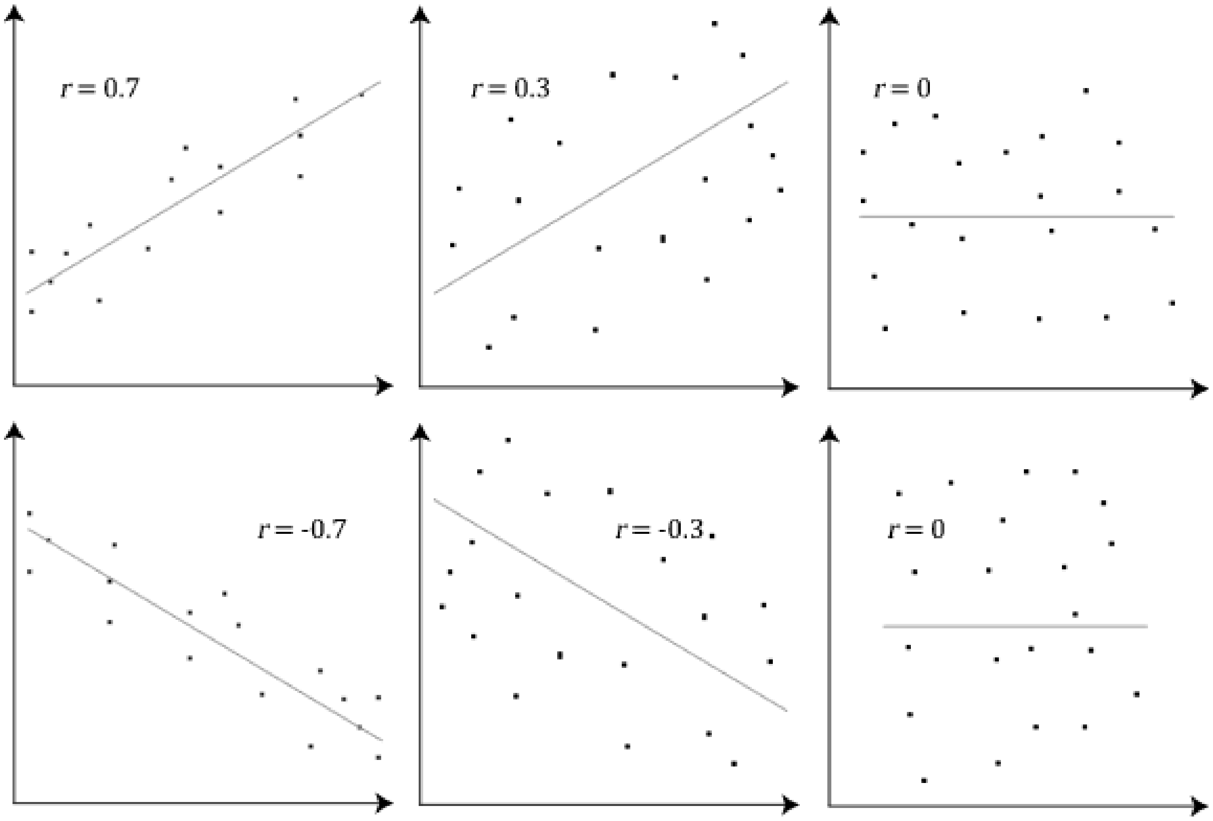
What values can the Pearson correlation coefficient take?

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



How can we determine the strength of association based on the Pearson correlation coefficient?

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Scaling is the process of transforming numerical features of different scales into a common scale. It involves adjusting the range of the variables so that they can be more easily compared and contribute equally to the analysis.

Scaling is particularly important in machine learning algorithms that are sensitive to the magnitude of input features.

Why Scaling is Performed:

Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of input features. Features with larger scales might dominate the learning process.

Distance-Based Algorithms:

Algorithms that rely on distance metrics, such as k-nearest neighbors or support vector machines, can be significantly affected by different scales.

Convergence Speed:

Optimization algorithms, like gradient descent, may converge faster when features are on a similar scale.

Regularization:

Regularization techniques, such as L1 and L2 regularization, assume features are on similar scales for fair penalization.

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

The Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient increases when the predictors in a regression model are correlated. A high VIF indicates that the variance of the coefficient estimate is inflated, suggesting potential multicollinearity issues.

Now, the issue of VIF being infinite typically arises when there is perfect multicollinearity. Perfect multicollinearity occurs when one predictor variable in the model can be exactly predicted by a linear combination of the other predictor variables. In this situation, the correlation matrix of the predictor variables is singular, and the inverse of this matrix, which is needed to calculate VIF, does not exist.

Here are some scenarios that can lead to perfect multicollinearity and an infinite VIF:

Duplicate Predictor Variables:

If two or more predictor variables are identical or have a perfect linear relationship, the correlation matrix becomes singular.

Linearly Dependent Variables:

When one predictor variable is a perfect linear combination of others, it causes a singularity in the correlation matrix.

Overparameterization:

If the model is overparameterized (more parameters than observations), perfect multicollinearity can occur.

To address the issue of infinite VIF, it's essential to identify and resolve multicollinearity in the dataset. This can involve removing redundant predictor variables, combining correlated variables, or applying regularization techniques if appropriate for the modeling task. Resolving multicollinearity not only stabilizes VIF but also improves the overall stability and interpretability of the regression model.

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.

Few advantages:

- It can be used with sample sizes.
- 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 —

- Come from populations with a common distribution
- Have common location and scale

- Have similar distributional shapes

- Have similar tail behavior

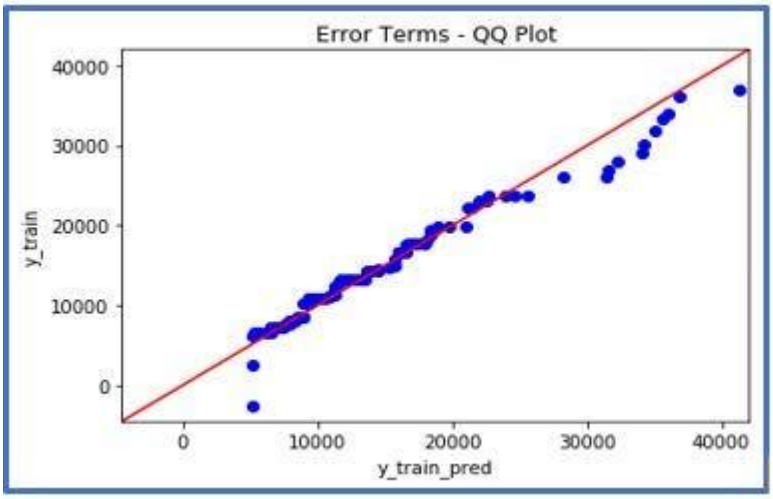
Interpretation:

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

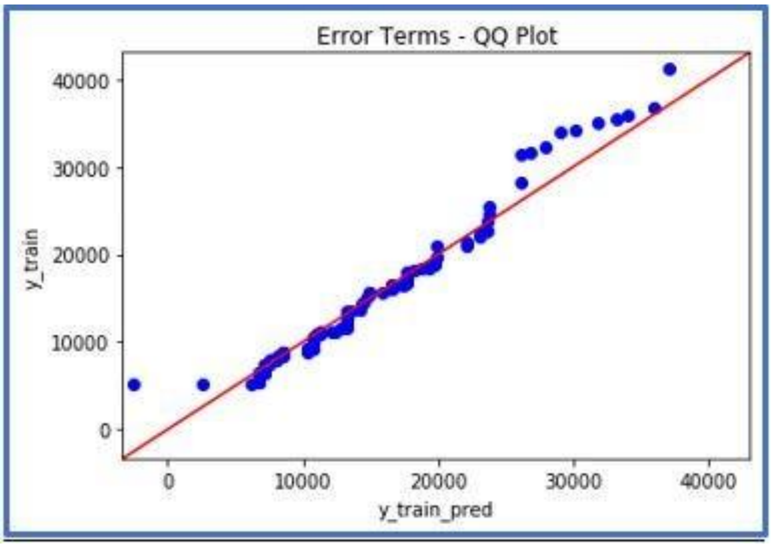
Below are the possible interpretations for two data sets.

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