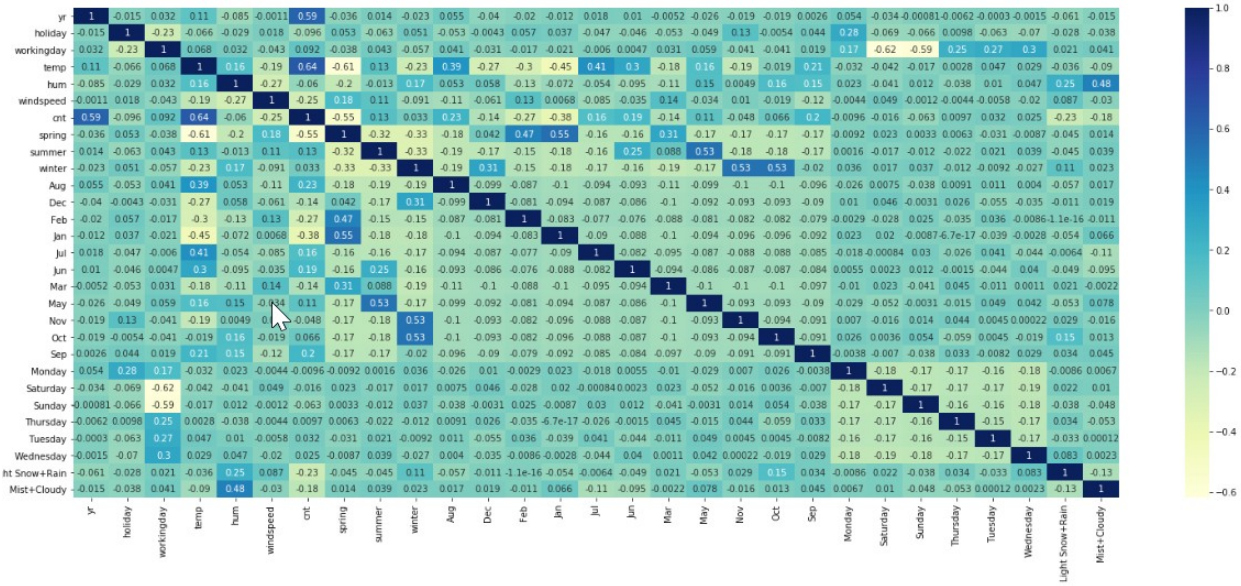
Assignment-based Subjective Questions

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



Among the categorical variables i.e. Season, month, weekday, weather sit , if the weather conditions are like snow or rain, the demand reduces.

And if the Season is pleasant the demand of bikes increases.

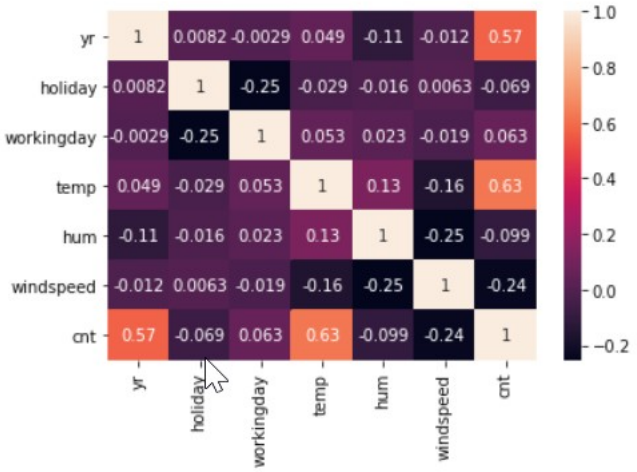
On Mondays the demand is less, as compared to weekends and holidays

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

This is used so that whether to get k-1 dummies out of k categorical levels by removing the first level.

"A categorical variable of K categories, or levels, usually enters a regression as a sequence of K-1 dummy variables. This amounts to a linear hypothesis on the level means."

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



The target variable is count, and windspeed and temperature are highly correlated with the target variable.

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

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No autocorrelation
- Homoscedasticity

To check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression), let us plot the histogram of the error terms and see what it looks like.

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

The variables which play an important role in demand graph are **temperature**, whether it is a working day or not. Temperature also plays a vital role in determining the demand for bikes.

We can also tell when there **are weather conditions** like snow, rain, the demand goes low

During the **holiday and during weekends** the demand is high as compared to other days.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A linear regression is one of the statistical models in machine learning that is used to show the linear relationship between a dependent variable and one or more independent variables.

Let's say we have a dataset which contains information about the relationship between 'number of hours studied' and 'marks obtained'. Several students have been observed and their hours of study along with their grades are recorded. This will be our training data. Our goal is to design a model that can predict the marks if number of hours studied is provided. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used to apply for a new data. That is, if we give the number of hours studied by a student as an input, our model should be able to predict their mark with minimum error.

For a model with one predictor,

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The best fit line is the line for which the error between the predicted values and the observed values is minimum. It is also called the ***regression line*** and the errors are also known as ***residuals***. The figure shown below shows the residuals. It can be visualized by the vertical lines from the observed data value to the regression line.

Linear Regression's power lies in its simplicity, which means that it can be used to solve problems across various fields. At first, the data collected from the observations need to be collected and plotted along a line. If the difference between the predicted value and the result is almost the same, we can use linear regression for the problem.

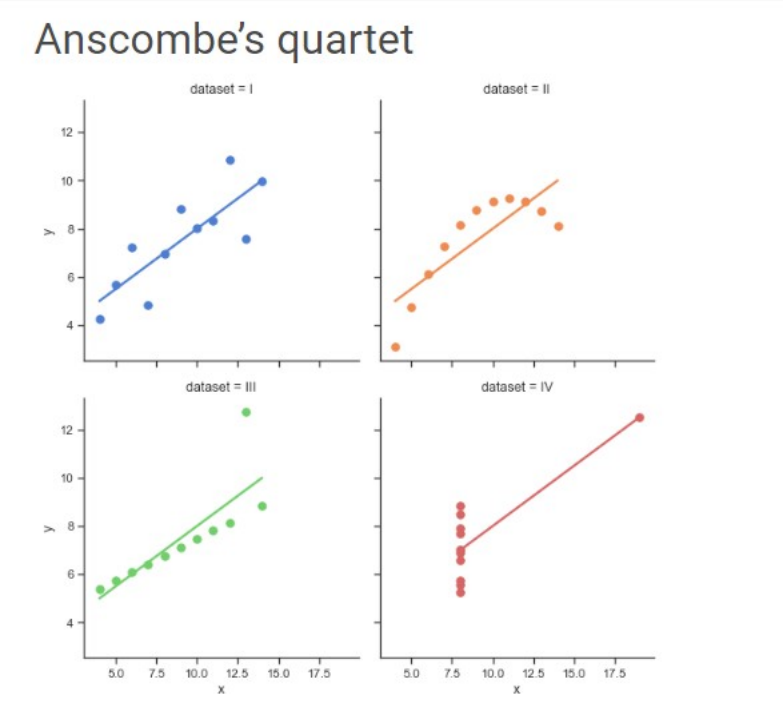
Assumptions in linear regression

If you are planning to use linear regression for your problem, then there are some assumptions you need to consider:

- The relation between the dependent and independent variables should be almost linear.
- The data is homoscedastic, meaning the variance between the results should not be too much.
- The results obtained from an observation should not be influenced by the results obtained from the previous observation.
- The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

2. Explain the Anscombe's quartet in detail.

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 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



3. What is Pearson's R?

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The t-test is used to establish if the correlation coefficient is significantly different from zero, and, hence that there is evidence of an association between the two variables. There is then the underlying assumption that the data is from a normal distribution sampled randomly. If this is not true, the conclusions may well be invalidated. If this is the case, then it is better to use Spearman's coefficient of rank correlation (for non-parametric variables).

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

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.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values N.

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

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the

variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity 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.

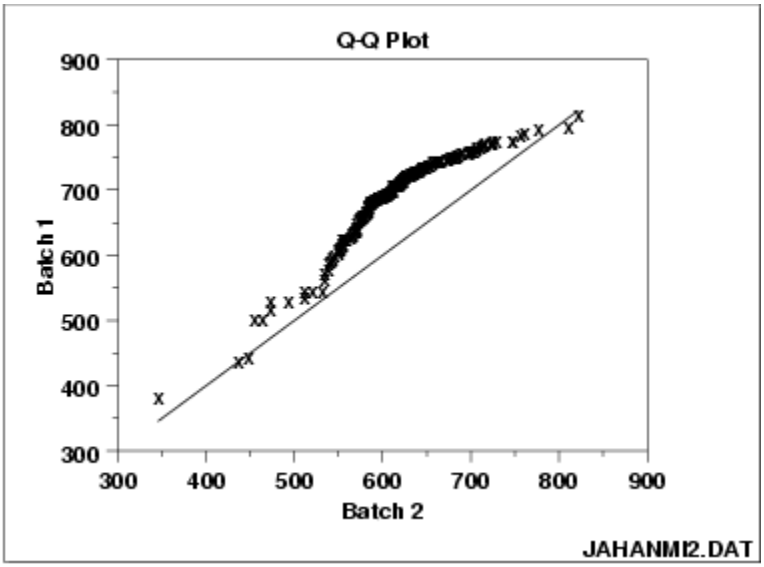
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

The sample sizes do not need to be equal.The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples



are replaced with the quantiles of a theoretical distribution.

Definition:

Quantiles for Data Set 1 Versus Quantiles of Data Set 2 The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1

Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

==========