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

The categorical variables present in dataset are season, year, month, holiday, weekday, workingday, weather situation. To analyze categorical variables box plot was used.

In boxplot for season variable it is seen that there was more demand in winter, less in spring and almost equal in summer and fall

For year column it is observed that demand was significantly increased in the year 2019

For month column the demand was increasing till mid of year & again started decreasing

Demand was seen to bit lower in case of holiday

There was not much variation seen on weekend or weekday. Same is the case for working day

When there is rainfall or thunderstorm the demand is less as compared to the demand when weather situation is few clouds or partly cloudy.

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

While converting categorical columns to numeric we create dummy variables. drop_first=True will drop first category thereby reducing the extra column that is created during dummy variable creation. This will eventually reduce the correlation between dummy variables. Hence the parameter is important

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

The numerical variables are temp, atemp, hum, windspeed, casual, registered. Among these registered variable has highest correlation with target variable i.e. cnt. Then we have casual, temp and atemp after registered column which are more correlated with cnt.

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

First assumption is that there should be linear relationship between dependent & independent variable which is verified by plotting pair plot.

Second assumption is that the error terms are normally distributed, which is verified by plotting the graph for error terms.

Third assumption is error terms should not be dependent on each other, that is also verified using graph

Fourth assumption is error terms should have constant variance across values of dependent variable which can be verified from residual plot.

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

Looking at the coefficients of final 6 features the top 3 features are year, temperature and season spring. These features significantly explain whether the demand will increase or decrease. It can be seen that there positive correlation i.e. demand will increase with respect to year and temperature. There is negative correlation i.e. demand will decrease in spring season.

1. Explain the linear regression algorithm in detail.

Machine learning algorithms are broadly classified into supervised and unsupervised algorithms. Linear regression comes under supervised learning algorithm which has associated labels in the dataset. The model predicts target variable based on independent variables. The labels have continuous numeric values, so it finds out relation between dependent & independent variable. Linear relationship between variables means that when value of one or more independent variables changes, value of dependent variable will also change accordingly. There are 2 types of linear regression –

1. Simple Linear Regression (SLR) – Finds relation between single dependent & single independent variable.

   The equation for SLR is : $y=\beta_0 + \beta_1 x$

   where,

   y = predicted value of dependent variable

   x = independent/predictor variable

   β_0 β_1 = regression coefficient

2. Multiple Linear Regression (MLR) – Finds relation between single dependent & multiple independent variables.

   The equation for MLR is : $y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n$

   where,

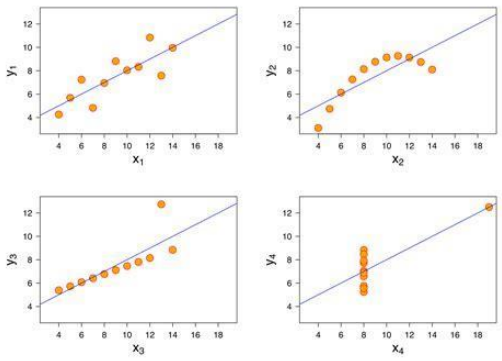
   y = predicted value of dependent variable

   x_1, x_2..x_n = independent/predictor variable

   β_0,β_1..β_n = regression coefficient. Each regression coefficient indicates change in y relative to one-unit change in respective x variable

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was developed by Francis Anscombe in the year 1973 to signify importance of plotting data i.e. visualization before analyzing and effect of outliers on the statistical properties. It consists of 4 datasets and each dataset has 11 datapoints. These datasets have same statistical properties i.e. mean, variance, SD etc. but different graphical representation. So, each graph shows different behaviour irrespective of statistical properties.



Dataset 1 consists of a set of points that represent linear relationship with some variance.

Dataset 2 shows doesn't show a linear relationship i.e. it is curve shape.

Dataset 3 shows tight linear relationship between points *x* and *y*, except for one outlier.

Dataset 4 shows the value of *x* remains constant, except for one outlier.

3. What is Pearson's R?

Pearson's R is a measurement of how two variables are dependent on each other. It is used for measuring linear correlation. It describes the direction of linear relationship between two variables. It can values ranging between -1 and 1. Values between 0 to 1 represents positive correlation, 0 represents no correlation and 0 to -1 represents negative correlation. We can also determine whether the slope of best fit line is positive or negative from Pearson's R (r) value.

Ex- When *r* is 1 or –1, all the data points fall exactly on the best fit line

When *r* is greater than 0.5 or less than -0.5, the points are close to best fit line

When *r* is 0, a best fit line is not helpful in describing the relationship between variables

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

Scaling is a process which is applied on independent variables to convert values in specified range(Ex -0-1 or 10-20). The dataset contains features having varying magnitude or units. The algorithm will take magnitude without knowing the units which would result in forming wrong model. So scaling will bring the magnitudes of feature at same level. Scaling just affects the coefficient and not other parameters like p-value, r-squared, f-statistic etc. Normalized scaling and standardized scaling is used for scaling.

Normalization/ Min-Max scaling – It converts values in between 0 and 1. It is calculated as -

x = x-min(x)/max(x)-min(x)

Normalization is not recommended when there are large outliers in dataset, otherwise the information will be lost. It is also called as scaling normalization. It is useful when we don't know the distribution.

Standardization replaces values by z-score. It converts all the data points such that their mean comes out to be 0 & standard deviation as 1. It is calculated as -

x = x-mean(x)/SD(x)

Standardization scaling is less affected by outliers. It is not bounded to certain range like normalization. It is often called as z-score normalization. It is useful when data is normally distributed.

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

Variance Inflation Factor(VIF) is used to find multicollinearity between variables. The equation used to calculate VIF is : 1/(1-Rsquared). When we get R-squared value equal to 1, VIF comes out to be infinite. This means there is perfect correlation between two independent variables. An infinite VIF indicates that the corresponding variable can be expressed exactly by another variable. We can solve this problem by dropping one of variable from dataset that is causing the multicollinearity.

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

The Q-Q plot or quantile-quantile plots are plots of 2 quantiles against each other. A quantile is fraction in which certain values fall below that quantile. If both sets of quantiles came from the same distribution, we should see points forming a line that's roughly straight.

It is graphical tool that helps to assess if data points came from same distribution. Ex – if we assume our data points are normally distributed, we can use Q-Q plot to check assumption. If assumption is wrong it helps us to understand how the assumption is violated & what points contribute to it.

This tool is used to compare shapes of distribution, thereby provides graphical view of how properties are similar or different in 2 distributions.