Assignment based subjective questions

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

- We observe that the maximum demand is driven in the fall season and is followed by winter and summer. We see a steep decline in demand in the fall season.
- The business seems to have grown in 2019 since we see a significantly higher demand in 2019 vs 2018.
- We observe high demand in the months of May, Jun, Jul, Aug, Sep and Oct. These months of seem to have significant impact on the demand.
- Demand with respect to weekdays appears to be mixed with highest demand on Thu, Fri and Sat. Other days of the week are fairly consistent with little fluctuations.
- Demand seems to be lowers on holidays vs demand on weekdays.
- Demand for bikes are highest on Clear/Partly cloudy days.

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

If we do not use drop_first = True, then n dummy variables will be created instead of (n-1). These predictors will correlate with each other which is known as multicollinearity and this in turn leads to dummy variable trap.

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

Registered Users has the highest correlation with the dependent variable.

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

After building the model on the training set, we check for multicollinearity using the variance inflation factor (VIF). Variables with high VIF (above 10) are dropped until we are sure that there is no multicollinearity in the model. Then we do a residual analysis to check for the distribution of the residuals if they are normally distributed or not.

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

- Temperature is an important factor affecting demand. With a unit change in temperature the demand will increase by 43%. Therefore, the recommendation would be to build up to required capacity during hotter months of the year.
- Light Rain or Snow impacts the demand greatly, i.e., if there a unit change in Rainfall/Snowfall the demand falls by 30%.
- Another important factor is Year. Given other factors remain unchanged the company is expected to see a growth of 23%.

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear Regression algorithm is a supervised machine learning algorithm which finds a linear equation which best describes the variation of a target variable using explanatory variables. It tries to fit a line which minimizes the sum squared of errors between the dependent and the predicted values we get from the model.

The linear regression equation is:

$Y = b_0 + (b_1 * x_1) + (b_2 * x_2) + (b_3 * x_3) + \ldots + (b_n * x_n) + (\text{error})$

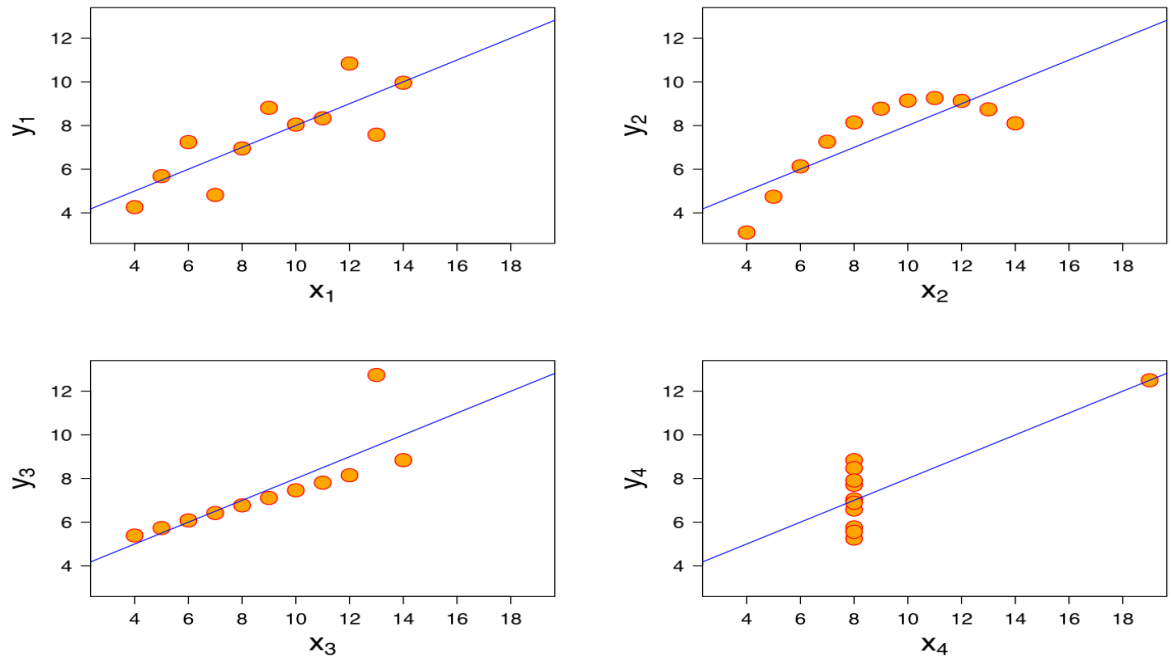
Where,

- Y = dependent variable
- $X_{1, 2, 3,,n}$ = the independent variables
- n = number of variables
- b_0 = Constant
- $b_{1, 2, 3,\ldots,n}$ = coefficients

A linear regression model helps in prediction of dependent variables. It helps us understand what are the import factors effecting the dependent variable and how much it will change if there is a unit change in the independent variables. The explanatory of a regression model is given by R-Squared whose value lies between 0 and 1. A value of 0.9 means that the model is able to explain 90% variation of the dependent variable.

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.



1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
2. The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

What is Pearson's R?

Pearson's R or Pearson Correlation Coefficient measures the strength between different variables. The value of Pearson Correlation Coefficient lies between -1 meaning a robust negative relationship and 1 meaning a robust positive relationship. 0 means there is no relationship between the variables.

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

Scaling is a data pre-processing step which is applied to normalize the data within a particular range. It also helps in speeding up the calculations of algorithms.

The data collected varies in magnitude. When an algorithm is run on the unscaled data it only takes the magnitude in account and not units and therefore modelling is done incorrectly. To solve this issue scaling is done to bring all the variables in the same magnitude.

Normalized Scaling bring all of the data within a range of 0 and 1.

Standardized Scaling replaces the value by their Z-scores, It all of the data into a standard normal distribution with mean = 0 and standard deviation = 1.

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

The value of VIF is calculated by $\mathbf{1/(1-R_i^2)}$, where 'i' refers to the ith variable. If R-Square becomes 1 then the denominator becomes 0 and overall value becomes infinite. This mean perfect correlation between independent variables.

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.