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

Season plot indicates more bikes are rent during fall season

Month plot indicates there are more sales on rented bikes during September month

Week day plot indicates that there are more sales on rented bikes during Saturday

Working day and Holiday plot indicates that here are more bikes on working day compared to holiday

Weathersit plot indicates that there are more rental plot compared to Clear, Cloudy and Partially Cloudy Days

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

Helps to reduce extra column created using dummy variable

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

From the analysis done in pair plot temp and a temp receive very high correlation with the target variable

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

Establish a Linear Relationship between Dependent and Independent Variable

No Multicollinearity

Normal Distribution of Error Terms

No Auto correlation

Homoscedasticity

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

weathersit_2 (Negative Correlation), yr (Positive Correlation), temp (Positive Correlation)

General Subjective Questions

Explain the linear regression algorithm in detail

Linear Regression is based on Supervised Machine Learning Technique and performs Regression Task. Useful for finding out relationship between variables and forecasting. Establish a Linear Relationship and kind of relationship between Dependent and Independent Variable. Used to predict dependent variable (y) based on independent variable (x)

Y = mX + c

X = Input Training Data Y = Labels of Data

m = Slope c = Intercept

Cost Function (J):

By achieving the best-fit regression line and the model is used to predict y in such a way that the error difference between the predicted and true value is minimum

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Explain the Anscombe's quartet in detail

Founded by Anscombe in 1973 used 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56

7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

After that mean standard deviation are found, correlation between x and y are found.

What is Pearson's R?

Pearson Correlation is used to find the strength of linear relationship between two variable.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

x and y are two vectors of length n. m, x and m,y corresponds to the means of x and y, respectively

It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and +1 meaning a total positive correlation.

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

Scaling is a step of data processing which is applied to independent variable to normalize the data to a particular range. Mainly used for speeding up calculations in algorithm

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude**.**

Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Normalization/Min-Max Scaling:

Nomalisation brings back the data to zero and one

Normalisation: x = x – min(x) / (max(x)-min(x))

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

Standardisation: x = x – mean(x) / std(x)

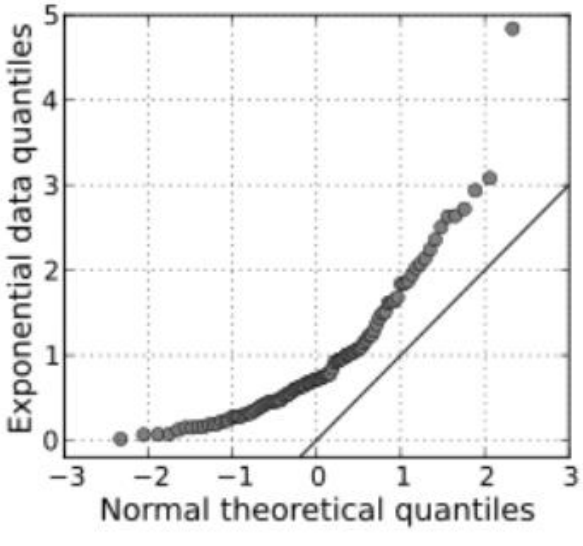
Disadvantage of normalisation over standardisation is it loses some information in data especially outliers.

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

If the value of VIF is infinity and there would be perfect correlation between 2 independent variables. If R2 is one 1/ (1-R2) which leads to infinity. In order to solve the problem we first need to remove variable which causes high Multicollinearity. An Infinite VIF value can be expressed exactly the linear combination of the variables

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

Q-Q plots are the plots of two quantiles against each other. Purpose of Q-Q plot is to find the two sets of data that comes under same distribution. A 45 degree angle is plotted to Q-Q plot if the two datasets forms under same distribution.



If two distributions are similar the points in the Q-Q plot lie on the line y = x. If the distributions are linearly related thus the points in Q-Q plot approximately lie on line but not necessarily y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.