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? (3 marks)

In the bike sharing dataset, lets consider the effect of the categorical variable ‘weathersit’ on the target variable ‘cnt’. While performing EDA, I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately. Similarly, certain inferences could be made ‘season’ and ‘yr’ as well.Also, during model building on inclusion of categorical features such as yr,season etc we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

----

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

During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding

----

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt’.

----

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
2. Multivariate Normality–Multiple regression assumes that the residuals are normally distributed. Here we can plot the bar graph of errors between observed and predicted values (i.e., the residuals of the regression) and it should be normally distributed.
3. No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
4. Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

----

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Based on final model top three features contributing significantly towards explaining the demand are:
a.) Temperature (0.5124)
b.) weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2843)
c.)year (0.2331)
So it recomended to give these variables utmost importance while planning to achieve maximum demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.
Basically regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Hence we can say that, Regression analysis is a technique of predictive modelling that helps you to find out the relationship between input and the target variable.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of the data based on some variables.

There are two types of linear regression models;
1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression

As the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

For eg: In order to increase the sales, we increased the expenditure on advertisements.We now can witness that, as the cost of advertisements increases, sales of the department also get increases.
In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.
Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a simple linear regression equation as:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

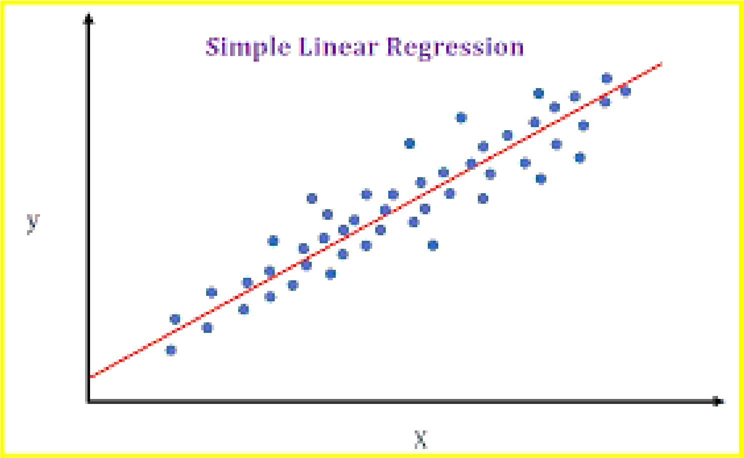
y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).

B0 is the intercept, the predicted value of y when the x is 0.

B1 is the regression coefficient – how much we expect y to change as x increases.

x is the independent variable (the variable we expect is influencing y).

e is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

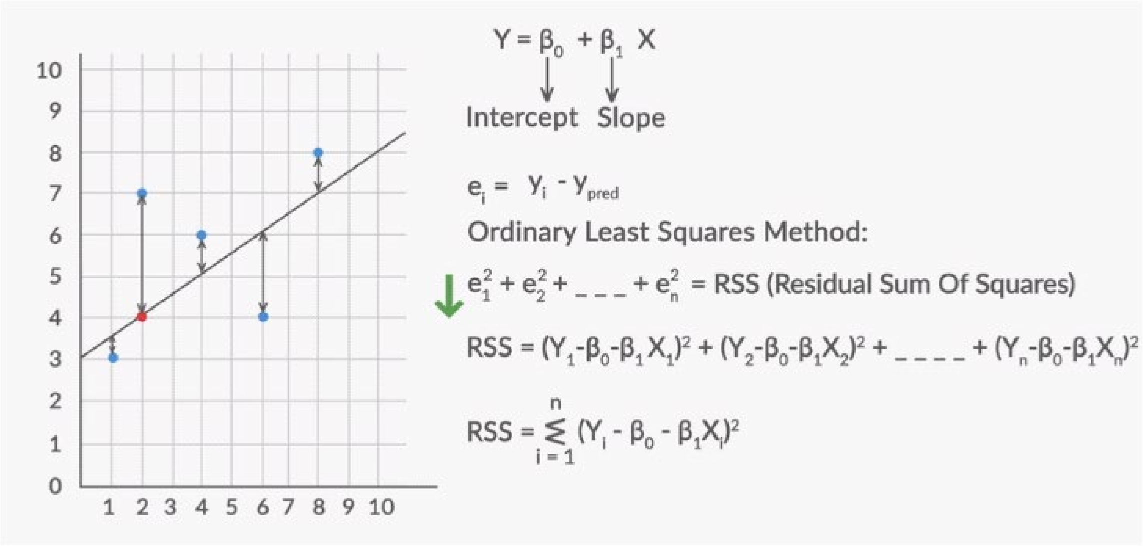


Assumptions of Simple Linear Regression

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.
4.The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Best Fit Line

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.



Strength of Linear Regression Model can be assessed using 2 metrics:

1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

1. R2 or Coefficient of Determination (takes a value between 0 & 1.)
It provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as: R2 = 1 - (RSS / TSS)

$$\mathrm{R}^2 = \frac{TSS - RSS}{TSS}$$

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma\left(y - \bar{y}\right)^2}$$

RSS (Residual Sum of Sqaures) the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

TSS (Total Sum of Squares) the sum of errors of the data points from mean of response variable.

Importance of RSS/TSS:
Think about it for a second. If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points. This is the worst possible approximation.

Multiple Linear Regression

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables.
The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

$$\boxed{\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ \ldots + \beta_i \mathrm{X}_i}$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

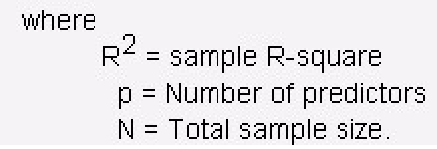
Assumptions of Multiple Linear Regression

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.
In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated ,then it is the case of Multicollinearity , which can be removed by using only one variable b/w them in the regression model.
3. Normality: The data follows a normal distribution.
4. Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

Adjusted R-squared

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.
The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$



where
R^2 = sample R-square
p = Number of predictors
N = Total sample size.

In order to detect and deal with Multicollinearity , VIF (Variance Inflation Factor) can be used.
A **variance inflation factor** is a tool to help identify the degree of multicollinearity. A multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. Multicollinearity creates a problem in the multiple regression because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.
In statistical terms, a multiple regression model where there is high multicollinearity will make it more difficult to estimate the relationship between each of the independent variables and the dependent variable. Small changes in the data used or in the structure of the model equation can produce large and erratic changes in the estimated coefficients on the independent variables.
To ensure the model is properly specified and functioning correctly, there are tests that can be run for multicollinearity. Variance inflation factor is one such measuring tool.
Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is

calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

$$VIF_i = \frac{1}{1 - R_i^2}$$

---

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

As we know, statistics have long been used to describe data in general terms.
For example, things like variance and standard deviation allow us to understand how much variation there was in some data without having to look at every data point individually. They give us a rough idea as to how consistent data is. However, knowing variance alone does not give you the full picture as to what the data truly is in its native form.
So in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties, Anscombe’s quartet was constructed.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

---

3. What is Pearson’s R? (3 marks)

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^{2}- (\Sigma x)^{2}][N\Sigma y^{2} - (\Sigma y)^{2}]}}$$

Assumptions:

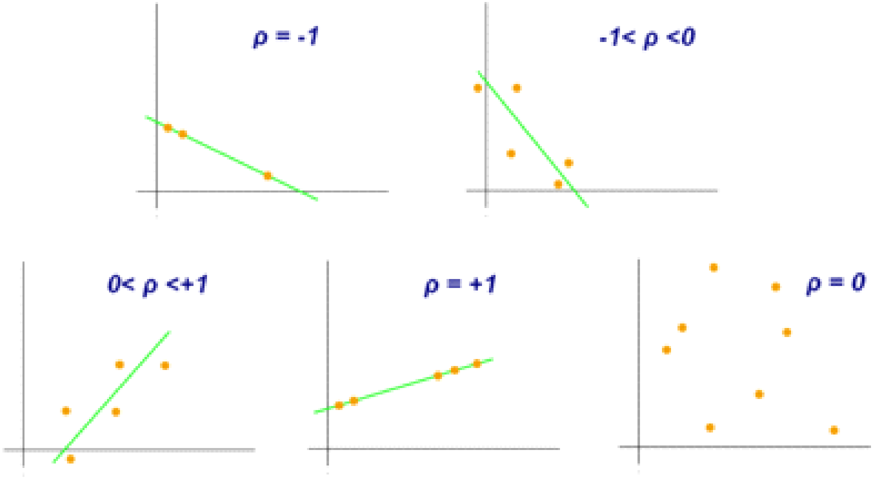
1. Independent of case: Cases should be independent to each other.
2. Linear relationship: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
3. Homoscedasticity: the residuals scatterplot should be roughly rectangular-shaped.

Properties:

1. Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
2. Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

1. Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. Low degree: When the value lies below + .29, then it is said to be a small correlation.
5. No correlation: When the value is zero.



----

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an 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 is done at the time of data pre-processing.

It is important to note that 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:
It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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).
sklearn.preprocessing.scale helps to implement standardization in python.

$$X_{new} = \frac{X_i - X_{mean}}{\textbf{Standard Deviation}}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

----

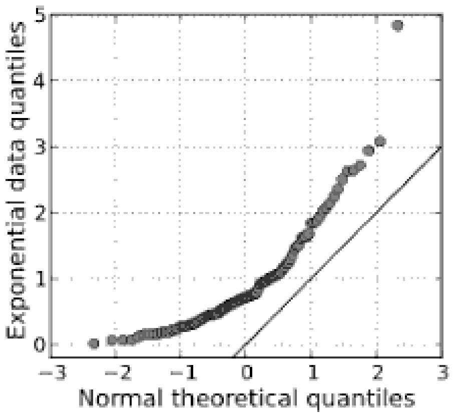
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If there is perfect correlation, then VIF = infinity.
This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

----

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



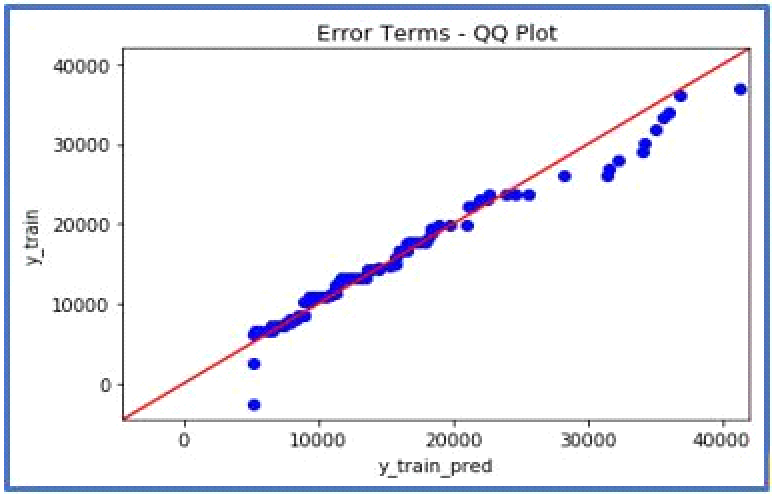
Q-Q Plot helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) 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.

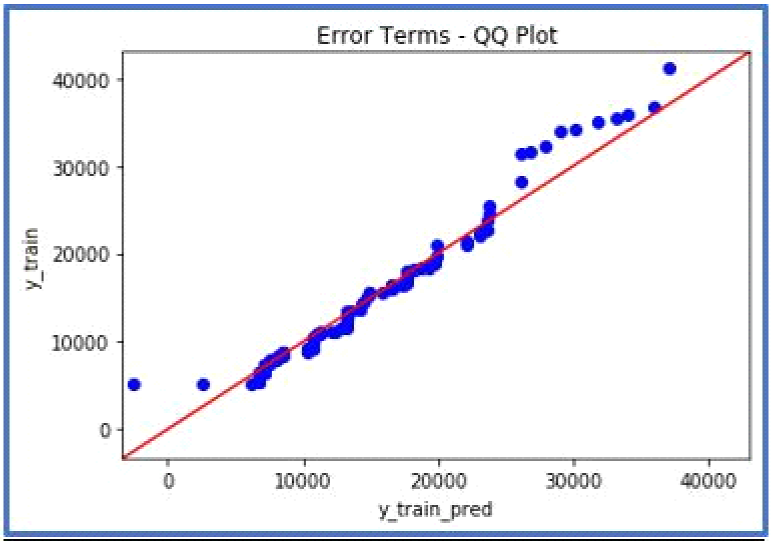
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

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

In Python: **statsmodels.api** provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.
