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?

Following categorical variables effect the dependant variable as

- Season: With a median of over 5000 bookings, season 3 saw the majority of bike reservations. Then came seasons 2 and 4 after this. The season can therefore be an effective predictor of the dependent variable.
- Weathersit: Most of the bike booking happened during 'weathersit1 with a median of close to 5000 booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- Mnth: With a median of approximately 4000 bookings every month, the majority of bike reservations took place in the months 5, 6, 7, 8, & 9. This suggests that mnth has a clear trend and can serve as a reliable predictor of the dependent variable.
- Workingday: 'Workingday' saw the majority of bike bookings, with a median of over 5000 bookings. This suggests that the working day can serve as a reliable predictor of the dependent variable.
- Holiday: Most of the bike booking were happened when it is not a holiday which means this data is not reliable. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- Weekday: The weekday variable exhibits a very close trend, with independent median reservations ranging from 4,000 to 5,000. This variable may or may not have any bearing on the predictor.

2. 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, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Thus using drop_first=true, give n-1 dummies out of n discrete categorical levels by removing the first level.

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

Between 'temp' and 'atemp' variable, there is high correlation

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

We validated the assumptions by checking

- **Error terms are normally distributed with mean zero**

we can see that the Residuals are normally distributed from the histogram plotted. Hence our assumption for Linear Regression is valid

- **There is a linear relationship between X and Y**

Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor 'cnt'.

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

As per the evaluation following 3 variables are top 3 contributing features

- Temperature (temp) - A coefficient value of '0.535253' indicated that a unit increase in temp variable increases the bike hire numbers by 0.535253 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.294276' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.294276 units.
- Year (yr) - A coefficient value of '0.235262' indicated that a unit increase in yr variable increases the bike hire numbers by 0.235262 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds

how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a linear regression as:

$Y = b_0 + b_1 x + \varepsilon$

where

y = dependant Variable

x = Independent Variable

b_0 = intercept of the line

b_1 = Linear regression coefficient

ε = random error

There are two types of linear regression:

- Simple linear regression
- Multiple linear regression

1. Simple Linear Regression

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

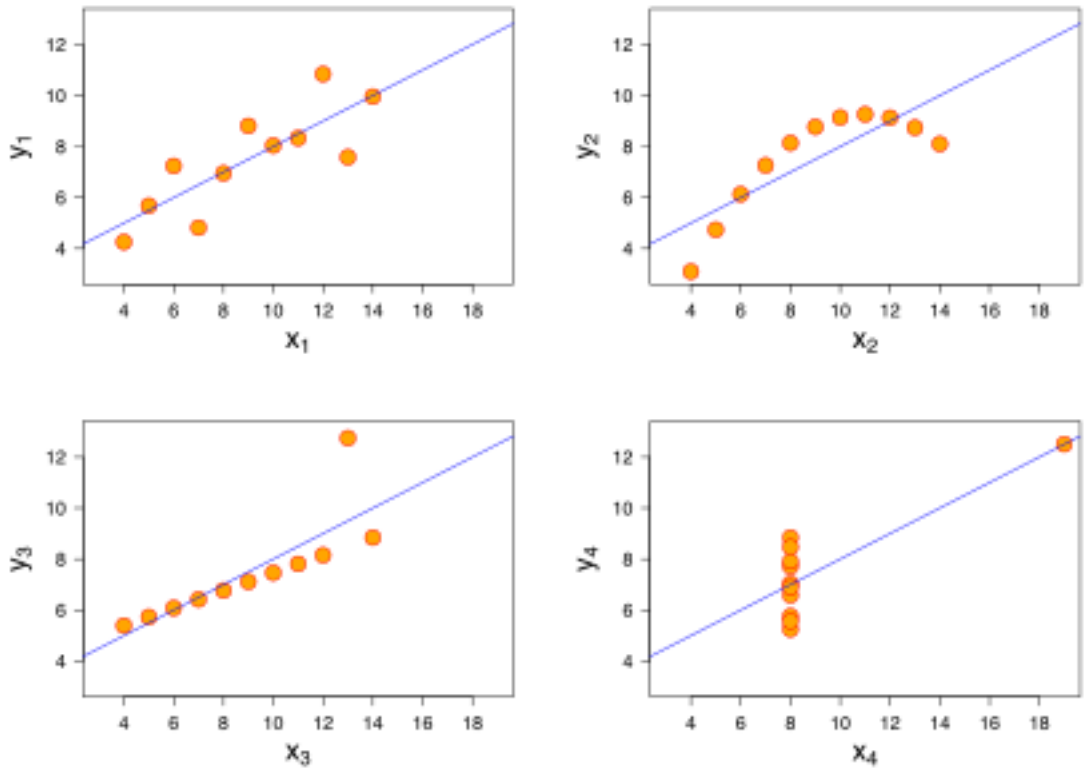
2. Multiple Linear Regression

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. Explain Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that, when plotted as a scatter plot on a graph, have different representations despite having similar descriptive statistical qualities in terms of means, variance, R-Squared, correlations, and linear regression lines. The datasets were developed by statistician Francis Anscombe in 1973 to highlight the value of data visualisation and to illustrate how summary statistics by themselves can be deceptive.

Anscombe's quartet consists of four datasets, each of which has 11 x-y pairings of data. Each dataset appears to have a distinct relationship between x and y when it is plotted, with various variability patterns and correlation strengths. The x and y mean and variance, x and y correlation coefficient, and other summary statistics are the same for each dataset despite these differences.



- 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*.
- The second graph (top right); while a relationship between the two variables is obvious, 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.
- 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.
- 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.

3. What is Pearson's R?

- Pearson correlation coefficient, also known as Pearson R, is a statistical test that estimates the strength between the different variables and their relationships. Hence, whenever any statistical test is performed between the two variables, it is always a good idea for the person to estimate the correlation

coefficient value to know the strong relationship between them.

- The correlation coefficient of -1 means a robust negative relationship. Therefore, it imposes a perfect negative relationship between the variables. If the correlation coefficient is 0, it displays no relationship. Moreover, if the correlation coefficient is 1, it means a strong positive relationship. Therefore, it implies a perfect positive relationship between the variables.
- The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, It estimates the relationship strength between the two continuous variables.
- It is independent of the unit of measurement of the variables. For example, suppose the unit of measurement of one variable is in years while the unit of measurement of the second variable is in kilograms. In that case, even then, the value of this coefficient does not change.

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

Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Scaling is performed When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret, hence scaling is performed.

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1.*

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**).*

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

It indicates that the relation between the variables is high i.e they have data that are similar to each other.
For eg: 2 dataset regarding weight, one is in Kgs and other in lbs then data is about to duplicate, thus to avoid this situation, either we can drop these columns and create another new columns or drop one of the columns or can add some more data.

The higher the VIF value for a variable, the more it contributes to multicollinearity. Removing variables with high VIF values can help reduce multicollinearity and improve the accuracy and stability of the regression model.

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

Q-Q plots are also known as Quantile-Quantile plots. It is the plot of the quantiles of a sample distribution against the quantiles of a theoretical distribution. It helps in determing if a dataset follows any particular type of probability distribution like normal, uniform, or exponential.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in post-deployment scenarios to identify covariate shift visually.