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?

For the season as a categorical value, from the box plot we can observe that summer, fall has the highest median, while the spring has least median.

From the year it is clearly seen that 2019 has more bike demands compared to 2018

From the weather situation, we can say that during the rainy season the bike demand is very low when compared to other weather conditions.

From the month variable, through the boxplot we can say that the median gradually increasing from Jan and then decrease when it comes to December

From the holiday boxplot, we can say that bike demands are more on non-holiday days than the holidays

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

If a variable has n levels, then we consider only n-1 levels, there by dropping the first column.

Even if we drop the first variable, we can represent the complete data. If we do not drop the column that leads to redundant. While building the model we generally avoid the redundant or duplicate columns there by to get the perfect model.

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

atemp and temp 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?

The first assumption is that the there is a linear relationship between dependent and independent variables and this is found by creating scatterplots.

The residuals are normally distributed and can be found by displot

Checking for the Homoscedasticity

Scatter plot between y_train_pred and res we can say that there is no pattern y predictions and the overall residuals

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

The variables : weathersit, temp, windspeed are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. 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 getting used.

There are two types of linear regression :

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression :

It 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.

The standard equation for regression line is : $Y = \beta_0 + \beta_1 X$

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

$Y = \beta_0 + \beta_1 X1+ \beta 2X2+ \beta 3X3+\ldots. \beta pXp + e$

Assumptions of simple linear regression were:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

2.Explain the Anscombe's quartet in detail

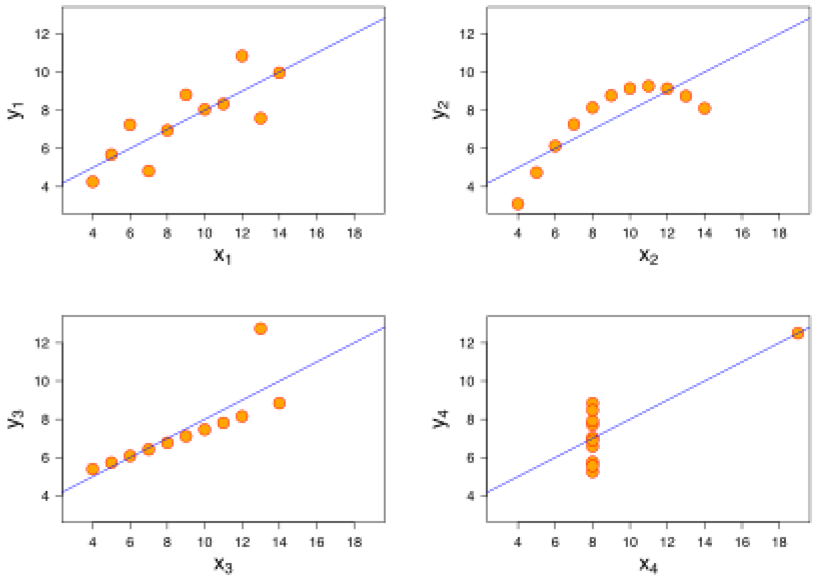
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points.

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, 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.

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 or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other

It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

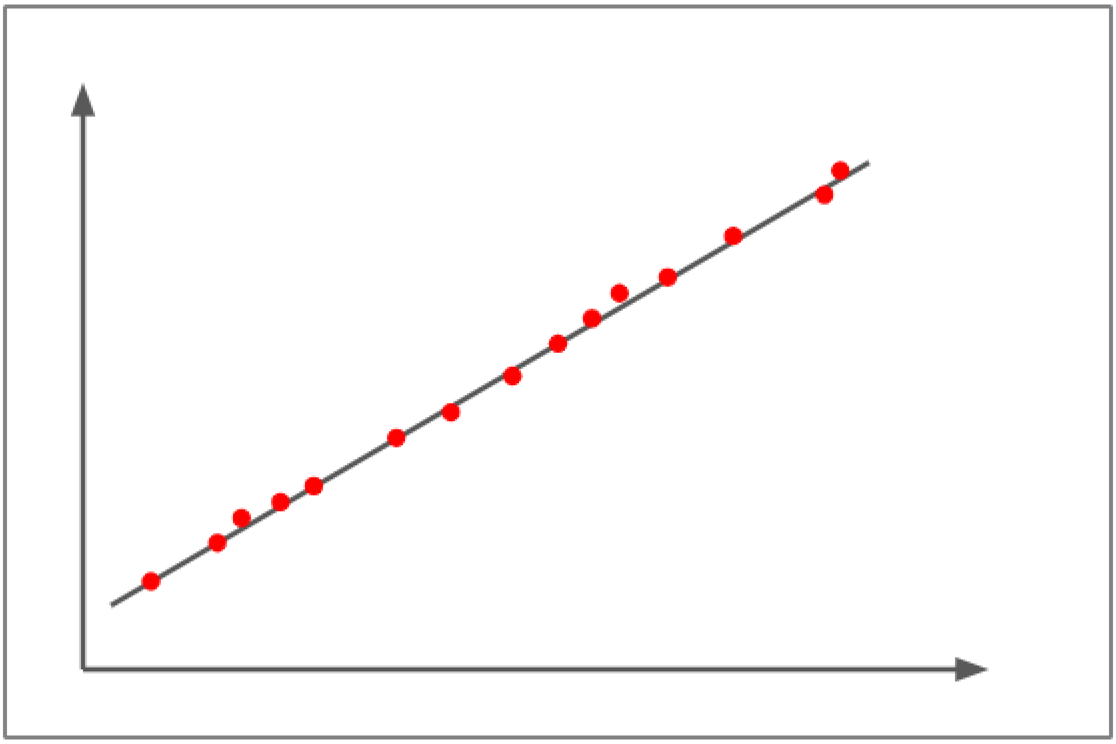
For example:

- **Positive linear relationship:** In most cases, universally, the income of a person increases as his/her age increases.
- **Negative linear relationship:** If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

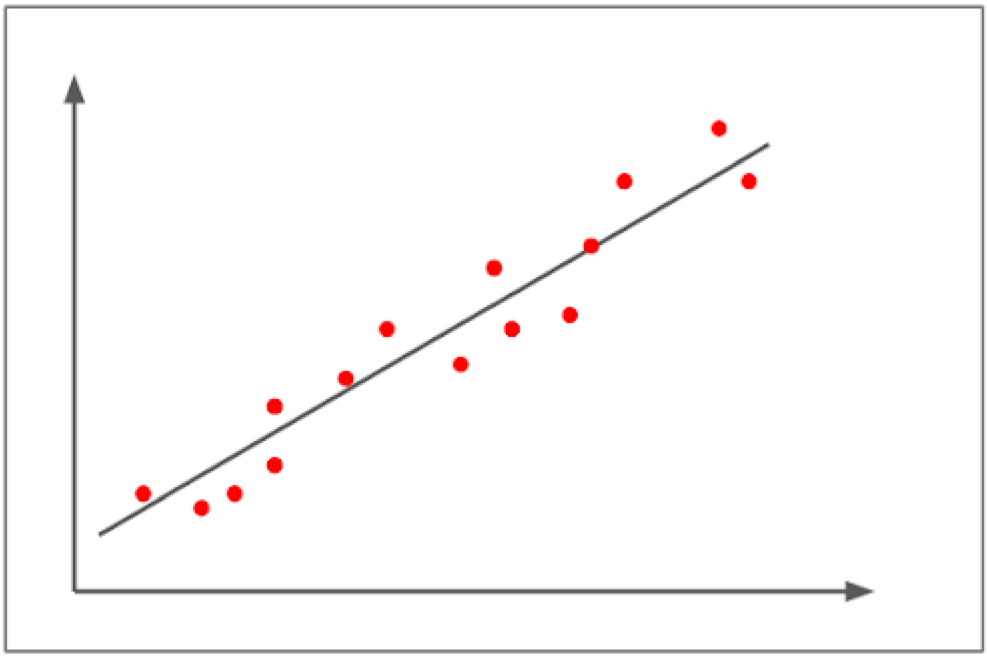
Strength signifies the relationship correlation between two variables. It means how consistently one variable will change due to the change in the other. Values that are close to +1 or -1 indicate a strong relationship. These values are attained if the data points fall on or very close to the line. The further the data points move away, the weaker the strength of the linear relationship. When there is no practical way to draw a straight line because the data points are scattered, the strength of the linear relationship is the weakest

The direction of the line indicates a positive linear or negative linear relationship between variables. If the line has an upward slope, the variables have a positive relationship. This means an increase in the value of one variable will lead to an increase in the value of the other variable. A negative correlation depicts a downward slope. This means an increase in the amount of one variable leads to a decrease in the value of another variable.

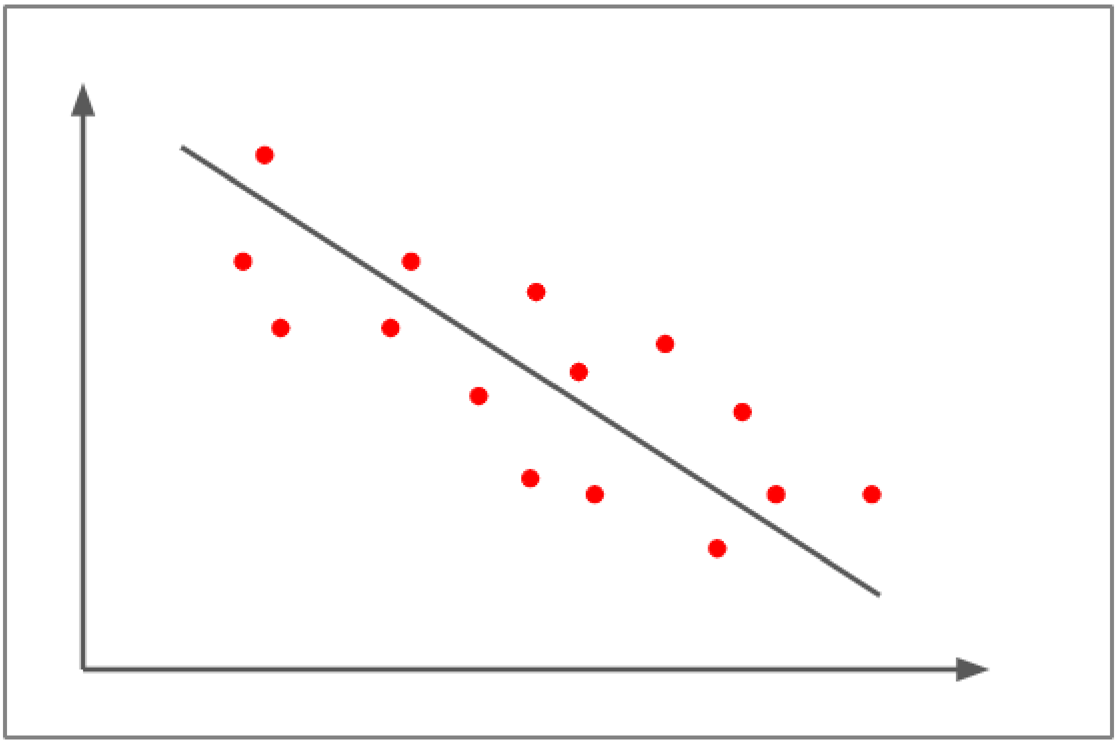
- **Large positive correlation**



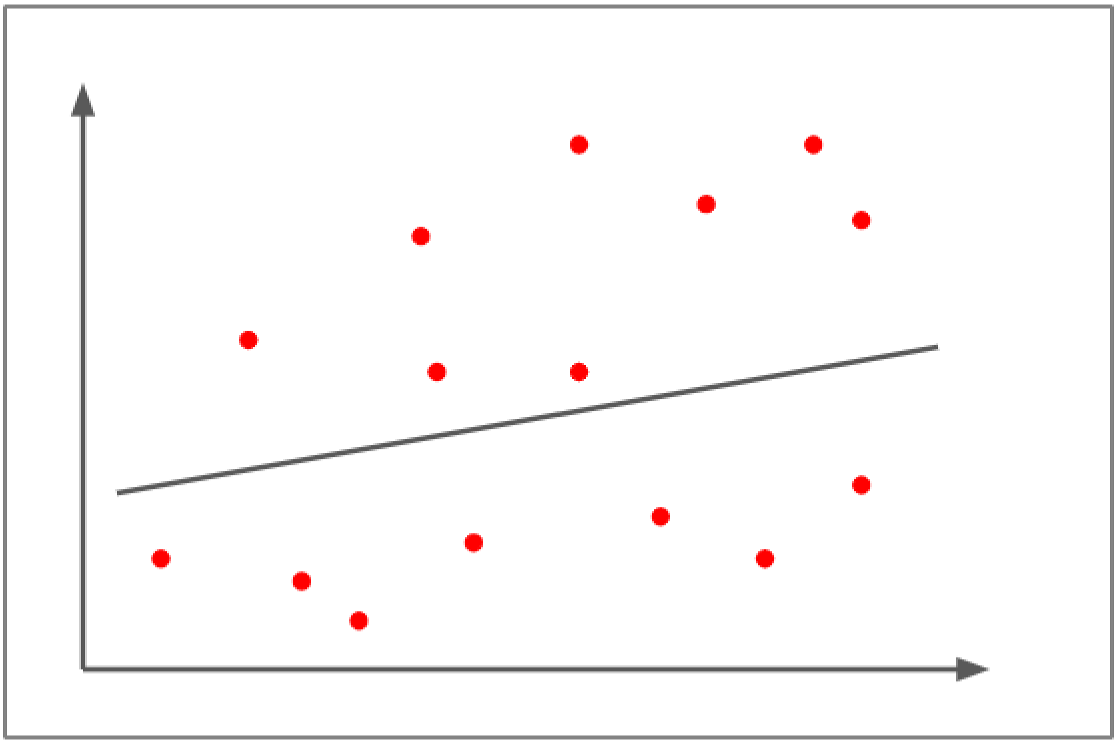
- **Medium positive correlation**



- **Small negative correlation**



- **Weak / no correlation**



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

Feature scaling is one of the most important data pre-processing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected. Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

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.

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

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line 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.