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?

Season 3 has highest median which shows : Highest sales in fall followed by summer

- yr 1 : Median sales in 2019 higher than 2018
- Weathersit 1 : Highest median sales in Clear weather and lowest sales in thunderstorm
- Highest sales in July and June. This is inline with season variable
- Not much difference if this is a working day or not

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

Suppose you have a categorical variable with three levels: A, B, and C.

- If you create dummy variables without drop_first=True, you'll get three variables: D_A, D_B, and D_C.
- Using drop_first=True, you might drop D_A and only keep D_B and D_C.

In the second case, the absence of both D_B and D_C implies the presence of the reference category A. This avoids multicollinearity and keeps the model interpretable.

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

- Atemp and temp have 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?

- There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not
- linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated

with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another

- Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.

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

Temp has the highest positive coefficient

- Temp : 4614.624
- Yr_1 : 1992.562
- Weathersit_3 : -2146.795

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, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.
The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
$y = \theta 0 + \theta 1x$
where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

$y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

1. **First Dataset**: Appears to be a well-behaved linear relationship.
2. **Second Dataset**: Shows a clear curve, indicating a non-linear relationship.
3. **Third Dataset**: Contains an outlier that strongly influences the linear regression line.
4. **Fourth Dataset**: Consists of a vertical line with one outlier, giving the illusion of a linear relationship.

Anscombe's quartet illustrates several important concepts:

1. **The Importance of Visualization**: Simple numerical summaries can be misleading; visualizing data helps to understand its true nature.
2. **Outlier Impact**: Outliers can heavily influence statistical summaries and regression results.
3. **Linear Regression Limitations**: Regression analysis might suggest a relationship that isn't actually there or miss non-linear relationships.

3. **What is Pearson's R?**
   Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear

relationship between two sets of data. In simple terms, it tells us “can we draw a line graph to represent the data?”

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

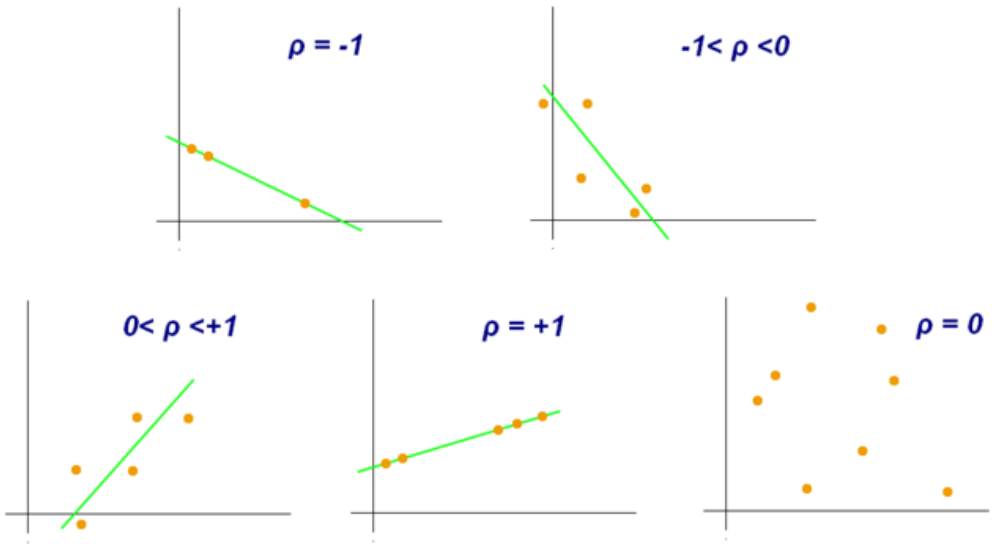
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.
A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?