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?**

Answer:

Below is the effect explained for each of the categorical variables :

- **Year** - Demand has significantly increased in the year 2019 compared to 2018
- **Season** - Demand is high during Fall, followed by Summer, Winter and least during Spring
- **Weather Situation** - Demand is high when weather condition is good i.e. clear sky, and decreases as it gets mist and cloudy and further decreases during rain and snow.
- **Month** - Demand is high during the months of Jun, July, Aug and Sept.
- **Holiday** - Demand decreases during a holiday. It is high during working days.

Note: Here demand refers to the target variable cnt.

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

Answer:

Drop_first helps in reducing one extra variable which gets created while creating dummy variables for categorical features. The dropped variable can be considered as absence of other options. If the variable is not dropped, it will cause multicollinearity i.e. one variable can be predicted from other.

e.g. We have a feature citizen having three possible categories – Indian, American & European.

So 3 dummy variables are initially created – citizen_Indian, citizen_American and citizen_European.

	citizen_Indian	**citizen_American**	**citizen_European**
Indian	1	0	0
American	0	1	0
European	0	0	1

As citizen_Indian and citizen_American both being 0 represents obviously the third one i.e. European we can drop the dummy variable.

	citizen_Indian	**citizen_American**
Indian	1	0
American	0	1
European	0	0

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

   Answer:

   'atemp' has the highest correlation with target variable cnt of 0.630685.

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

   Answer:

   - Predictions are made using the model on the training dataset. The error or residual terms (y_train - y_train_pred) are plotted and they are normally distributed. Hence validates the assumption – error terms are normally distributed.
   - There are no visible patterns in the error terms in the plot, which ensures they are independent of each other. Hence validating the assumption the error terms are independent of each other.
   - The error terms just appear to be evenly distributed noise around zero hence having constant variance which validates the assumption that error terms have constant variance(homoscedasticity).
   - The plots between various independent variables with dependent target variable cnt during EDA, we observe linear relationship.

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

   Answer:

   The top 3 features contributing significantly towards the demand of the shared bikes are:

   Yr, weathersit & season.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer:

   Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression

shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

E.g. X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

$$y = \boldsymbol{\theta 1} + \boldsymbol{\theta 2}.x$$

While training the model we are given:

x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept
θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail.

Answer:

Summary statistics are helps in giving important insights about the data, however there's risk of missing out on some important points and alone summary statistics can be misleading. Hence data visualizations using graphs and plots are essential along with summary statistics. This is beautifully demonstrated using Anscombe's Quartet dataset. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.

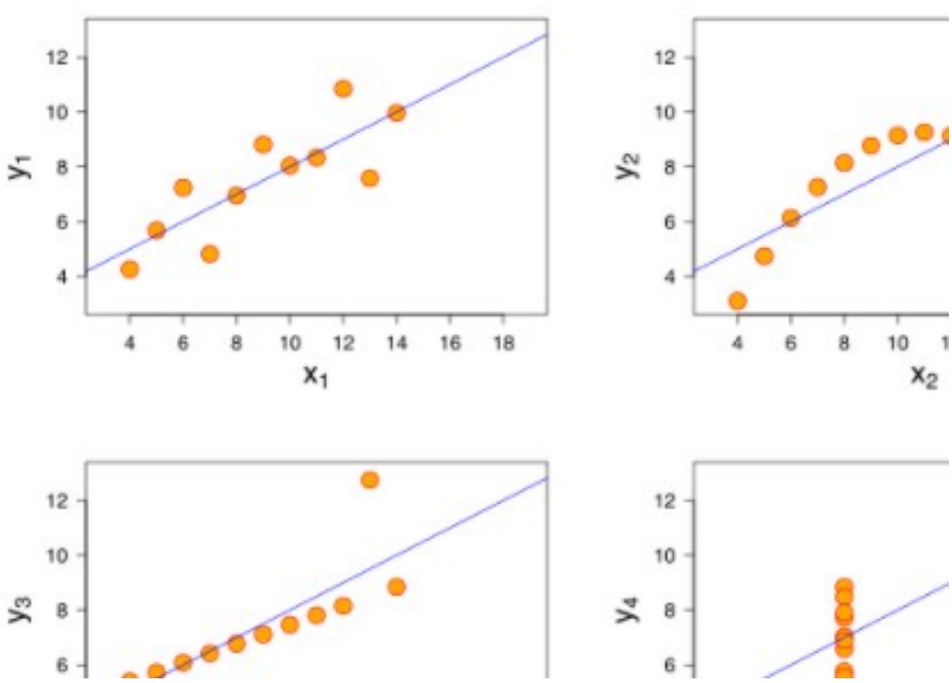
All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset

- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar.

But when we plot these four data sets on an x/y coordinate plane, we get the following results:



- Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance.
- Dataset II fits a neat curve but doesn't follow a linear relationship .
- Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Dataset IV looks like *x* remains constant, except for one outlier as well.

Hence we can conclude, summary statistics alone can be misleading hence data visualization is very important.

3. **What is Pearson's R?**

Answer:

Pearson's correlation (also called Pearson's R) is a statistic that measures the linear correlation between two variables. It also has a numerical value that lies between -1.0 and +1.0. The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative.

For example:
Positive linear relationship: Generally, 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.

Also, it is the covariance of the two variables divided by the product of their standard deviations.

Pearson correlation coefficient formula:

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Answer:

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. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

X' = (X-Xmin)/(Xmax-Xmin)

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.

X'=(X-Mu)/Sigma

In case of standardized scaling the values are not restricted to a particular range.
Another key difference is, Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution and Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However both can be used for either case interchangeably.

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

Answer:

In case of perfect correlation between two independent variables (or the corresponding variable may be expressed exactly by a linear combination of other variables) the VIF is infinite. The R2 = 1, so Vif=1/1-R2 = infinity. In order to solve this, we would need to drop one of those variable from the dataset.

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

Answer:

Quantile-Quantile (Q-Q) plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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

This 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.