## Assignment-based Subjective Questions

**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
**Total Marks**: 3 marks (Do not edit)
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

After plotting categorical variables in boxplot, one could infer that:

- Seasons have an impact on the dependent variable and there is more demand for bikes in summer and fall, followed by winter and spring
- Month of year also has an impact on dependent variable. Demand increases consistently every month and peaks between April to September, with a slight dip in August.
- Weekday doesn't seem to have much effect on the target variable
- Weather situation has a big impact on demand. There is higher demand on clear days whereas demand drops when the weather is misty and drops further when there is snow
- There is lower demand during holidays
- There was significantly higher demand in 2019 when compared to 2018

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

Dummy variables are used to help convert categorical variables to useable numerical values in linear regression models. Dummy variables represent values in ones and zeros to convey all of the necessary information.
When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels.
The reason we use drop_first=True when creating dummy variables is to avoid dummy variable trap. By setting drop_first=True, we drop one dummy variable to reduce its perfect multicollinearity with the others. This removes redundant information in the data.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
**Total Marks:** 1 mark (Do not edit)
**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

'temp' has the highest correlation with target variable 'cnt'

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

These following assumptions should be met before we draw inferences from the model estimates or before we make predictions with the model:

- The relationship between predictor and target is linear
    - Plotted a graph to visualize the linear relationship between target and predictor variables. The plot shows a fitted line through the data points
- Errors are normally distributed
    - Plotted a histogram with the error residuals and it shows a normal distribution
- Homoscedasticity of errors (constant variance around the fitted line)
    - There is a visibly uniform variance between error terms and the fitted line
- Independence of the observations
    - The error terms are randomly scattered around the fitted line, with no specific pattern

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

'Temp', 'Light-snow' and 'Yr' contribute significantly towards explaining the demand of the shared bikes?

---

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression algorithm basically learns from a sample dataset, for which the target value is already known, and maps the data points to the most optimized linear function, which can be then used for prediction of target value on new datasets.

It is a type of machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent variables by fitting a linear equation to the observed data.

The algorithm can be used for prediction of a) continuous and b)categorical/binary variables

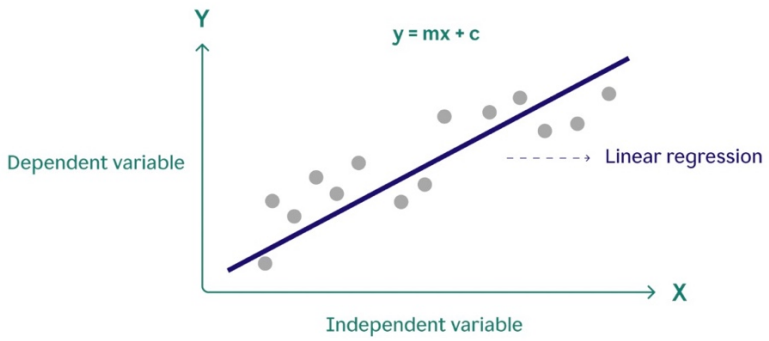
With linear regression we try to find the best-fit line through known data points such that the error between the predicted and actual values is minimum. One commonly used way to do this is through 'gradient descent' which iteratively calculates the best-fit line through the data points by reducing the errors each time until we find the line with the least errors.

If we consider Y as the target/dependent variable and X as the independent variable that explains Y then the relationship between X and Y can be explained by the following linear function:

Y= mX + c

Where Y is the dependent variable that we are trying to explain/predict,
X is the independent variable that defines Y
m is the co-efficient that specifies the significance of X
c is a constant value (baseline of the model) which explains Y when X = 0.

Finding the coefficients of independent variables that best fits the sample/training data is the objective of linear regression.




When there is one independent variable we called it a simple linear regression and for more than one dependent variables we call it multiple linear regression.

There are some assumptions we make with linear regression:

1- The independent and dependent variables have a linear relationship
2- The data points in the dataset are independent of each other
3- The error terms are in a normal distribution
4- Error terms have a constant variance
5- No multicollinearity/high correlation of between variables

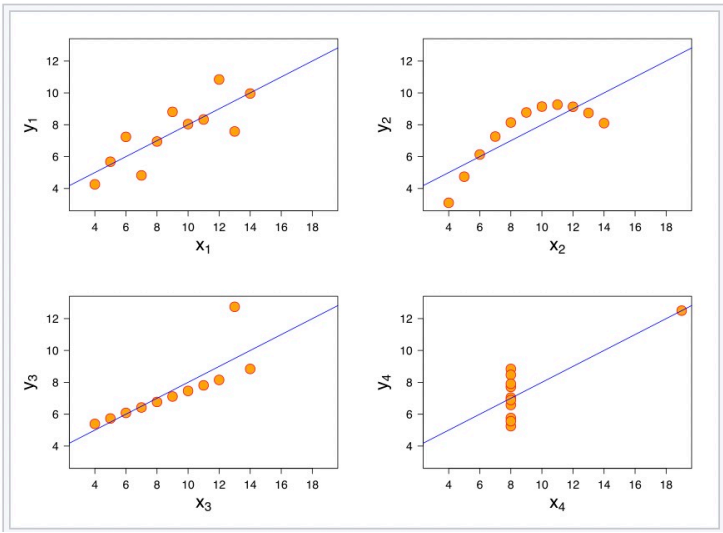
---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar but look very different when plotted.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and exploratory data analysis. Here's how different the seemingly similar datasets look when plotted.




The summary statistics (means and variances) for the four datasets were identical across the datasets. The correlation is high and the regression line is pretty similar as well. The graphs however seem to tell a different story.

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson correlation 'r' is used to measure the strength and direction of the linear relationship between two continuous variables. It is a standard method to determine whether the correlation coefficient is statistically significant or not.

Pearson correlation coefficient is calculated by dividing the covariance of the variables by their standard deviations.

Range of r:

$-1\ r \leq 1$

When r= 1: Perfect positive correlation
When r= −1: Perfect negative correlation
When r=0: There is no correlation

Positive correlation: As one variable increases the other tends to increase
Negative correlation: As one variable increases the other decreases
When r is closer to -1 or 1, it is a strong linear relationship
When r is close to 0, the relationship is weak

It is assumed that the variables are continuous, the relationship between them is linear, the data is free from outliers and they are normally distributed.

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Scaling is a step during data preparation which is applied to independent variables to normalize the data within a particular range. This makes the values comparable to be able to derive meaningful insights. It also helps in speeding up the calculations in an algorithm.

Usually, a given data set contains values that are highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units. To solve this, we do scaling to bring all the variables to the same level of magnitude.
Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

Normalization rescales the values into a range of [0,1]. also called min-max scaled.
Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1.
Some key differences are as follows:

| Normalization | Standardization |
|---|---|
| Scales data to a fixed range (e.g., [0, 1]) | Transforms data to mean = 0, std = 1 |
| Sensitive to outliers | Less sensitive to outliers |
| Rescale to a range | Standardize for unitless comparison |

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) can be infinite when there is perfect multicollinearity between the independent variables in a model. Perfect multicollinearity is when one independent variable can be predicted exactly by another independent variable. For example, if there are multiple identical columns in the input dataset, there will be perfect multicollinearity.

When the correlation coefficients between the independent variables are close to 1 or -1, the VIF can be very high and lead to infinite values. A large value of VIF indicates that there is a correlation between the variables.

The value of VIF is calculated by the formula:

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

Where '*i*' refers to the *i* th variable.

If R-squared value in the above formula is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against each other. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Uses:

Q-Q plot is primarily used to check the normality assumption of the residuals. The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Normally distributed.

The plot helps in detecting Deviations.

Helps identify outliers.

Importance:

In linear regression, QQ plot helps in validating the model assumptions such as the normal distribution of residuals. It helps improve the model and make it reliable.

---