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? (3 marks)**

   There are few dependent variables Season, Workingday, Weathersit, Mnth, weekday
   Effect on the dependent variable
   $mnth\_Feb$: A coefficient value of '0.1441' indicated that a unit increase in $mnth\_Feb$ variable decreases the bike hire numbers by 0.1441 units.
   $mnth\_Jan$: A coefficient value of '0.1935' indicated that a unit increase in $mnth\_Jan$ variable decreases the bike hire numbers by 0.1935 units
   $mnth\_Mar$: A coefficient value of '0.0890' indicated that a unit increase in $mnth\_Mar$ variable decreases the bike hire numbers by 0.0890 units
   $mnth\_Jul$: A coefficient value of '0.0298' indicated that a unit increase in $mnth\_Jul$ variable increases the bike hire numbers by 0.0298 units
   $mnth\_Jun$: A coefficient value of '0.0632' indicated that a unit increase in $mnth\_Jun$ variable increases the bike hire numbers by 0.0632 units
   $mnth\_Sep$: A coefficient value of '0.0776' indicated that a unit increase in $mnth\_Sep$ variable increases the bike hire numbers by 0.0776 units
   $weekday\_Saturday$: A coefficient value of '0.0634' indicated that a unit increase in $weekday_Saturday\ dnesday$variable increases the bike hire numbers by 0.0634 units
   $weekday\_Thursday$: A coefficient value of '0.1057' indicated that a unit increase in $weekday\_Thursday$ variable increases the bike hire numbers by 0.1057 units
   $weekday\_Wednesday$: A coefficient value of '0.0678' indicated that a unit increase in $weekday\_Wednesday$ variable increases the bike hire numbers by 0.0678 units

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Temp and atemp

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   Steps to validate assumptions of linear regression model are:
   •Linear Relationship: A scatter plot was plotted between one independent and one dependent variable, a straight line passing through the points could be observed.
   •Homoscedasticity: Variance of error terms was observed and found that the variance of error terms is constant.
   •building and deciding model: P-value and VIF was used.
   •Normality of Errors: Histogram.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Yr
mnth_Jan
temp
mnth_Feb

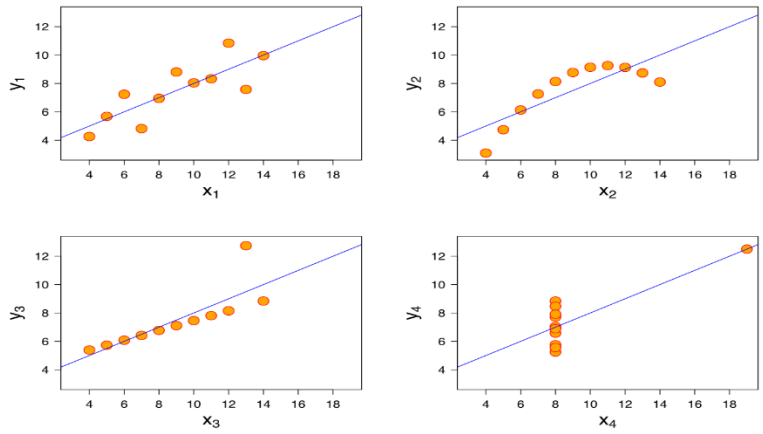
General Subjective Questions

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**. Regression models a target prediction value based on independent variables.
The linear regression algorithm consists of following steps:
1. Analysis and conversion of variables: Variables must be converted to required format, ie, Conversion of Categorical variables. Analysis of variables, to understand correlation and directionality of the data.
2. Dividing the model into test and train dataset: The data set must be divided ideally in 70-30 proportion. This is done to check the predictive capacity of final regression model.
3. Estimating the model, i.e., fitting the line: A final model is estimated which has the best representation of maximum points in a linear line. After developing the model, we check the assumptions of linear regression model to determine usefulness of the model.
4. Evaluating the validity and accuracy of the model: The model is run of the test dataset to obtain the R 2 and other factors.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



The four datasets can be described as:
Dataset 1**:** this **fits** the linear regression model pretty well.
Dataset 2**:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
Dataset 3**:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
Dataset 4**:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value **r** = 1 means a perfect positive correlation and the value **r** = -1 means a perfect negative correlation.

Requirements for Pearson's correlation coefficient

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

What is Scaling:
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Why is scaling performed:

Most of the times, collected data set contains features highly varying in magnitudes, units and range.

If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect

modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of

magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters

like t-statistic, F-statistic, p-values, R-squared, etc.

What is the difference between normalized scalingand standardized scaling:

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to

implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

(3 marks)

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity. 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"

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

(3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. 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.