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)

- Ridership shows that autumn season seems to have more bookings.
- There is a sharp increase in ridership across the two years 2018 and 2019
- There is more booking in the month of May, June, Jul, Aug and Sept. The beginning and end of the year show lesser bookings
- Booking seemed to be almost equal either on working day or non-working day.
- Rain impacted the ridership adversely

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)

It eliminates the extra column created during dummy variable creation thereby reducing the correlations created among dummy variables. If there a m categorical levels for a variable they can be represented by m-1 dummy variables. drop_first=True helps us achieve this.

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 and atemp variables have the highest correlation among numerical variables

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)

Residual analysis was done on the train data. The checks below were performed:

- Normality of error terms: Error terms were normally distributed
- Linear relationship validation: There was visible linearity among variables
- Multicollinearity check: There was no significant multicollinearity among variables.
- Homoscedasticity: There was no specific pattern in residual values.
- Independence of residuals: Autocorrelation does not exist

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)

The features having the largest absolute coefficient values considered to be the most important.

Based on their absolute values we get the features significantly contributing as:

- Temp
- Year
- Windspeed

Year: While the coefficient for Year is higher, I have excluded them since we need more data across multiple years to conclude further.

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)

Linear Regression is a statistical model that analyses the linear relationship between a dependent variable (target) with given set of independent variables (features). The dependent variable increases or decreases based on the independent variable in a proportional manner.

A liner regression can be mathematically represented as
y = mX + c
Here,
y is the dependent variable being predicted.
X is the independent variable used to make predictions.
m is the slope of the line which relationship between X and y
c is a constant- also, known as the Y-intercept. If X = 0, Y would be equal to c. Which is the point at which the line intersects the y-axis.

There are certain assumptions on the dataset considered for liner regression models:

1. **Linearity**: That is the relationship between variables between target and feature variables must be linear.
2. **Normality** of error terms: Error terms should be normally distributed
3. **Multi-collinearity**: There is very little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables have dependency between them
4. **Homoscedasticity**: There should be no visible pattern in residual values.
5**. Autocorrelation**: Auto-correlation occurs when there is dependency between them. Linear regression model assumes that there is very little or no autocorrelation in the data

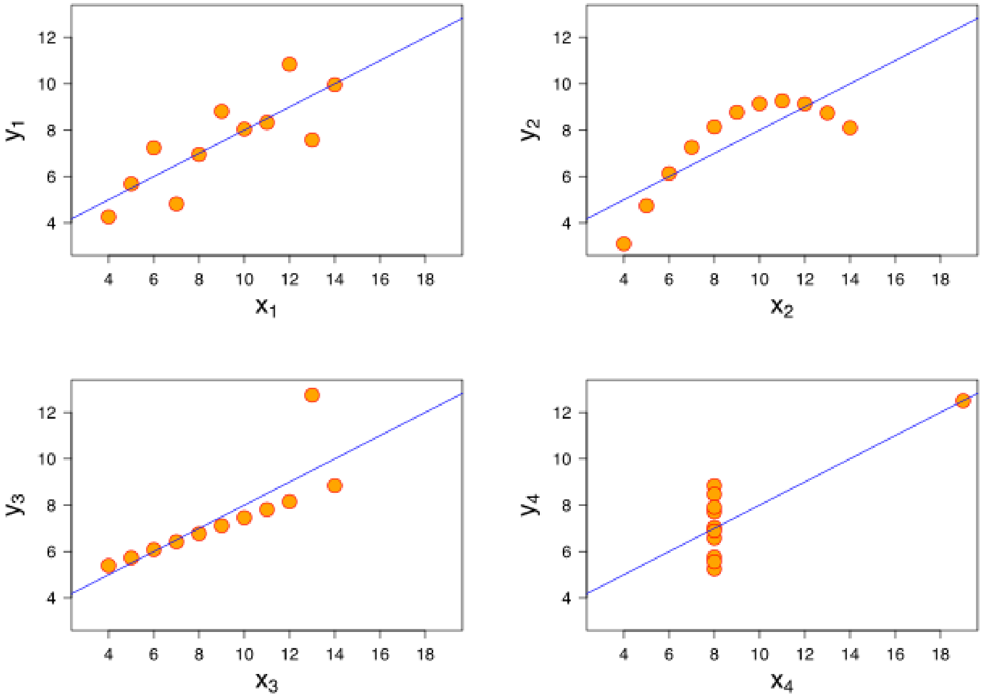
Linear regression is of the following two types:

1. Simple Linear Regression
2. Multiple Linear Regression

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)

Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when plotted on a graph. Each dataset consists of eleven (x, y) points. It conveys the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

(Source of the below figure is Wikipedia)



- The first scatter plot (top left) - It appears to be a simple linear relationship, corresponding to two correlated variables, 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 exists 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.
- The third graph (bottom left) - The modelled relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- 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.

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)

Pearson's r represents the summary of strength of linear association between variables. If the variables go up or down together then their correlation coefficient is positive. If the variables tends to go up or down in opposing directions that is when one value goes up the other goes down and vice versa, then the correlation coefficient will be negative.

Its value ranges between -1 to +1.

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

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)

Scaling or Feature scaling specifically is a technique of standardizing the independent features present in the data into a fixed range. It is performed during the pre-processing step to handle highly varying magnitudes of values or units. If feature scaling is not performed, then the machine learning algorithm will attempt to weigh greater values, higher and consider smaller values as the lower values, irrespective of the unit of the values.

Differences between Normalized and Standardized scaling is given below:

Normalized Scaling (Min Max Method)	**Standardized Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between (0, 1) or (-1, 1)	Not bounded to a certain range
Highly impacted by outliers in the data	Less impacted by outliers.
Used when features are of different scales	Used when we want to ensure mean is zero and standard deviation is one.
(x-min(x)) / (max(x) -min(x))	(x-mean(x))/ sd(x)

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)

VIF - Is the variance inflation factor -It indicates how much the variance of the coefficient estimate is being inflated by collinearity.

VIF (X) = 1/ (1- R^2)

If the VIF is 2.5, this means that the variance of the model coefficient is inflated by a factor of 2.5 due to the presence of multicollinearity.

VIF of infinity shows a perfect correlation between two independent variables. If **R^2** tends to 1 then the denominator tends to zero and hence VIF tends to infinity. This can be solved by dropping one of the variables from the dataset.

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)

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 dataset

Use of Q-Q plot:
It is used to establish evidence that the two data sets have come from populations with different distributions.
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.
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences.

The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.