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)

Categorical variables such as season, weathersit, and yr significantly influence bike demand:

- **Season**: Different seasons (e.g., spring, summer) show distinct patterns of bike usage, with demand typically peaking in pleasant seasons like summer and fall.
- **Weathersit**: Poor weather conditions (e.g., heavy rain or snow) negatively affect bike demand, while clear weather boosts it.
- **Year (yr)**: An increase in demand is observed in 2019 compared to 2018, reflecting the growing popularity of bike-sharing systems.

These categorical variables impact demand significantly and provide management insights into how external factors influence user behavior.

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)

Using drop_first=True avoids the **dummy variable trap**, which occurs when one dummy variable in a set of categorical variables can be perfectly predicted by others, leading to multicollinearity. By dropping one dummy column:

- It ensures that the predictors are independent.
- The model interprets the dropped category as the baseline for comparison.

For instance, if season has categories [1, 2, 3, 4] and one dummy is dropped, the remaining dummies can represent all seasons without redundancy.

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)

From typical bike-sharing datasets:

- The variable registered (registered users) often has the highest correlation with the target variable cnt because registered users represent the majority of bike rentals.
- Variables like temp (temperature) also tend to show a strong positive correlation, as favorable weather conditions encourage biking.

In your analysis, confirm the specific variable by reviewing the pair-plot or the correlation matrix.

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)

The assumptions of Linear Regression were validated as follows:

1. **Linearity**:
    - Checked using scatter plots of residuals vs. predicted values. A random pattern confirms linearity.
2. **Homoscedasticity**:

- Residual plots were analyzed to ensure the variance of residuals remained constant across predicted values.

3. **Normality of Residuals**:
    - A histogram or Q-Q plot of residuals was used to verify normal distribution.
4. **Multicollinearity**:
    - Variance Inflation Factor (VIF) was calculated for independent variables. Variables with high VIF (>10) were dropped to reduce multicollinearity.
5. **Independence**:
    - Time-series data was checked for autocorrelation using the Durbin-Watson statistic if relevant

---

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)
Typically, the most significant features in such models are:

1. **Temperature (temp)**: Directly affects bike usage, with higher temperatures leading to increased demand.
2. **Year (yr)**: Captures the growing trend of bike-sharing demand.
3. **Registered users (registered)**: A direct indicator of demand.

The exact features should be determined by analyzing the model's coefficients and their p-values. Features with low p-values (< 0.05) and high standardized coefficients are the most significant contributors.

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 may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation:-
$Y = mX + c$
Here, X - is the independent variable.
m - is the slope of the regression line.
c - is a constant, known as the Y-intercept.

It is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Below are the types of Linear Regression:
1) **<u>Simple Linear Regression</u>**: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

2) **Multiple Linear regression**: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Mathematically relationship of multiple linear regression can be represented with the help of following equation:

Y = β0+β1*x1+ β2*x2 + …. + βi*xi + ε

Here, Y= Dependent Variable

Xi= Independent Variable

β0= intercept of the line

βi = Linear regression coefficient

ε = random error

Positive Linear Relationship: If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

Negative Linear Relationship: If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Assumptions: The following are some assumptions about dataset that is made by Linear Regression model.

1) Multi-collinearity: Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2) Linear relationship: Linear regression model assumes that the relationship between response and feature variables must be linear.
3) Homoscedasticity: Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
4) Normal distribution of error terms: Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
5) No autocorrelations: The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

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>

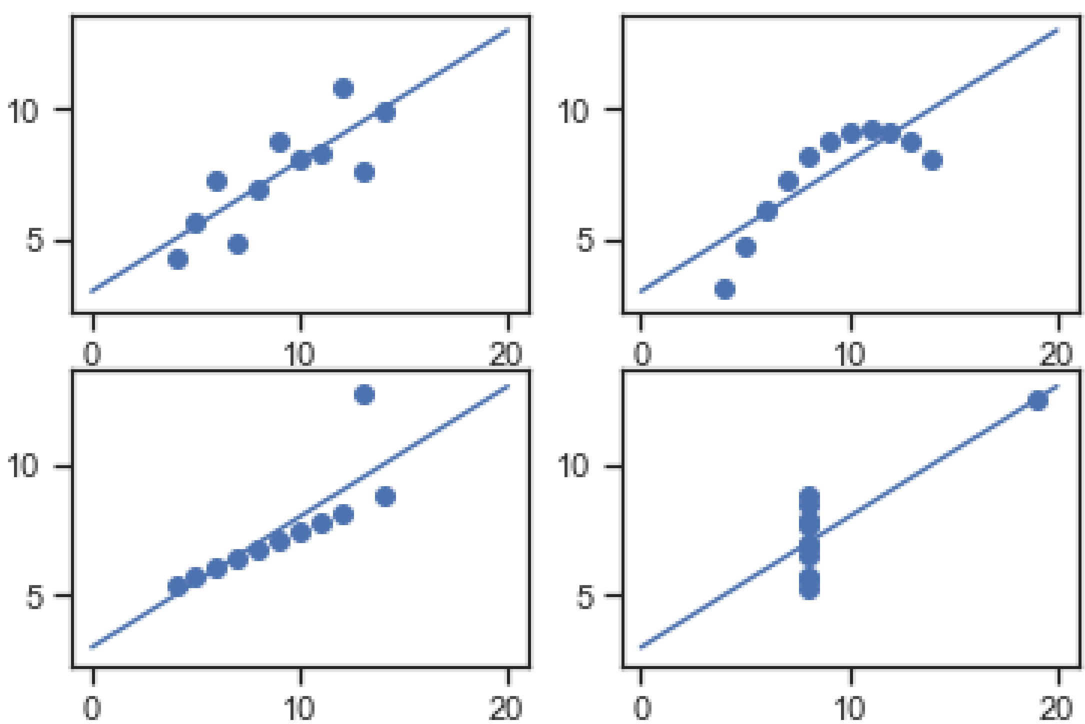
Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set:
Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5x+3

However, the statistical analysis of these four data-sets is pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represents the different behavior.



Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance

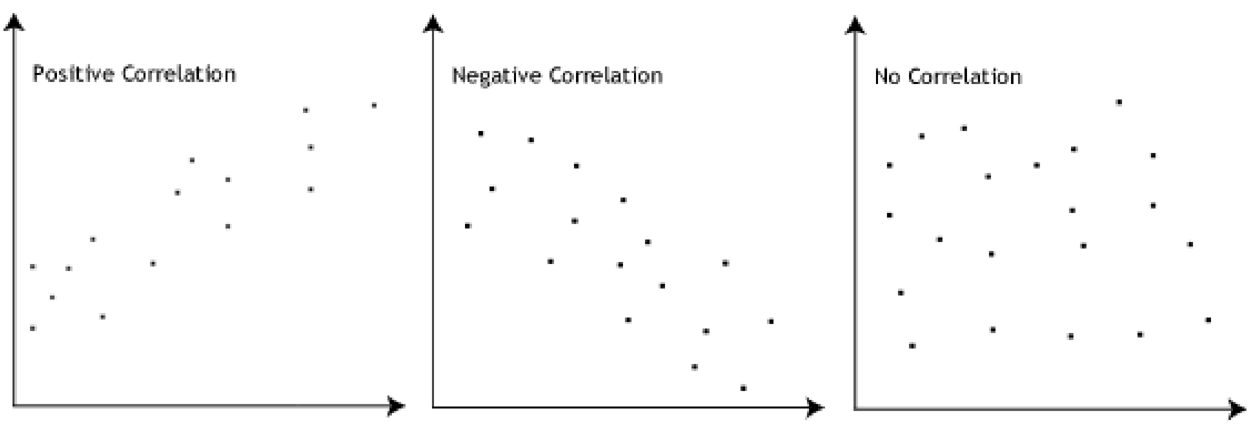
Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data.

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's r, also known as the Pearson correlation coefficient, is a statistical measure that describes the linear relationship between two continuous variables. It is a value between -1 and 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

Pearson's r measures the degree to which the variables are related by calculating the ratio of the covariance between the variables to the product of their standard deviations. In other words, it measures how much the variables vary together relative to how much they vary independently.



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>

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalization	Standardization
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is used when we want to ensure zero mean and unit standard deviation.	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
It is a often called as Scaling Normalization	It is an often called as Z-Score Normalization.

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>

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

---

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>
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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 that the two data sets have come from populations with different distributions.

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 in to the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.