

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)

- **'yr'** (Year): The bike bookings in 2019 (yr = 1) are noticeably higher compared to 2018 (yr = 0), reflecting the company's steady progress and the increasing demand for bike-sharing services over time.
- **'season'** (Seasonality): The Fall season stands out with the highest number of bike bookings, indicating that this time of the year is particularly favorable for bike-sharing activities.
- **'mnth'** (Month): Bike bookings follow a seasonal trend, with a gradual increase at the beginning of the year, peaking during the summer months of June, July, August, and September. Afterward, bookings experience a gradual decline, reflecting the typical behavior of bike usage as the weather changes.
- **'weekday'** (Weekday Usage): While the overall pattern of bike bookings across weekdays is fairly consistent, Thursday, Friday, and Saturday show a slight increase in demand. This suggests that people may be more inclined to use bikes towards the end of the workweek and during the weekend.
- **'weathersit'** (Weather Situation): Clear and partly-cloudy weather correlates with the highest number of bike bookings, with misty/cloudy conditions following behind. Conversely, snow/rain/thunderstorm weather results in the lowest bookings, which is understandable as adverse weather conditions deter people from biking.
- **'workingday'** (Working Day vs Non-working Day): On average, working days see slightly higher bookings compared to non-working days. This may indicate that people are more likely to use bikes as an alternative mode of transportation during weekdays, while non-working days have a more relaxed biking pattern.
- **'holiday'** (Holiday Effect): Bike bookings on holidays appear to be lower than on non-holidays, possibly due to people preferring to engage in other activities or travel during their time off, rather than biking.

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)

In multiple linear regression (MLR), it is crucial to use **drop_first=True** when creating dummy variables to prevent a problem known as multicollinearity, specifically the dummy variable trap.

What is the Dummy Variable Trap?

The dummy variable trap occurs when all categories of a categorical variable (converted into dummy variables) are included in the regression model, leading to perfect multicollinearity. This happens because one of the dummy variables can be perfectly predicted by the others and the intercept term, creating redundancy in the data. This redundancy can result in:

The inability of the regression algorithm to compute unique estimates for coefficients.

Numerical instability, leading to unreliable results in your regression model.

How Does drop_first=True Solve This

By setting drop_first=True, one category of the categorical variable (referred to as the reference category) is excluded when creating dummy variables. This exclusion effectively addresses the issue of multicollinearity in the following ways:

Prevents Multicollinearity: The excluded category serves as the baseline, and the coefficients of the remaining dummy variables represent differences relative to this baseline. This ensures that the model does not suffer from perfect collinearity.

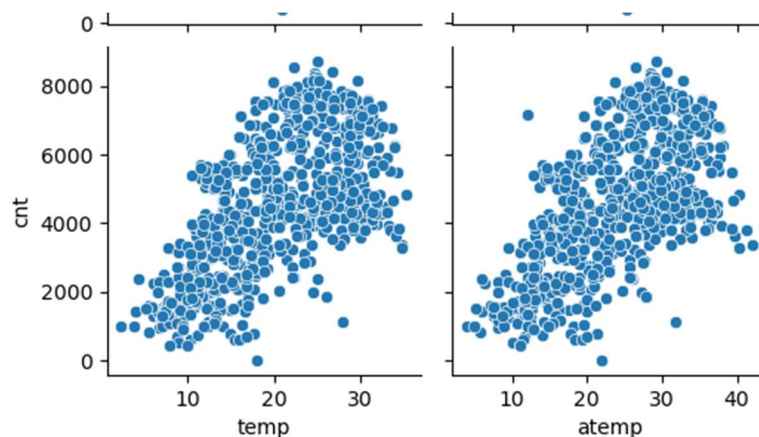
Improves Interpretability: Each coefficient now represents the effect of being in a specific category compared to the reference category, making the model's output easier to understand and interpret.

Ensures Model Stability: By removing redundant information, drop_first=True makes the computation more efficient and ensures the model remains numerically stable, improving its overall reliability.

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' variable has the highest correlation with the target variable 'cnt'. They are positively correlated.

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)

- **Linearity:** By noting that the CCPR (Component and component plus residual) plots of all the features in the final model are producing straight lines, I validated that the predictors have linear relationship with target variable.
- **Normality of Residuals:** by plotting the distribution of residuals I validated that they are normally distributed.
- **Homoscedasticity:** By plotting the scatter plot of residuals, I validated that there is no

significant pattern in them.

- **Independence of Residuals:** By noting the Durbin-Watson value 2 (actually 2.072) of final model, I validated that there is no dependency on the residuals.
 - **Multicollinearity:** By ensuring that VIF (Variance Inflation Factor) values of all the predictors in final model are below the permissible value 5, I validated that there is no multicollinearity.
-

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)

As per the absolute values of the coefficients or weights of the linear regression model, the top 3 features contributing significantly towards the demand of shared bikes are as follows:

- Temp
 - Yr
 - snow/rain/thunderstorm (weathersit 3)
-

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 is a supervised learning algorithm used to predict a continuous dependent variable (output) based on one or more independent variables (inputs). The goal is to find the linear relationship between the variables. Below is a detailed explanation of the algorithm: 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, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y. c is a constant, known as the Y-intercept. If $X = 0$, Y would be equal to c. Furthermore, the linear relationship can be positive or negative in nature as explained below–

Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.

Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decreases.

Linear regression is of the following two types –

1. Simple Linear Regression
2. Multiple Linear Regression

Assumptions

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.

Auto-correlation :-

Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables :-

Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms :-

Error terms should be normally distributed

Homoscedasticity:-

There should be no visible pattern in residual values.

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 that were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization and how different datasets with identical descriptive statistics (such as mean, variance, correlation, etc.) can have very different distributions and relationships. It shows that summary statistics alone (like the mean or correlation) are not sufficient to fully understand the nature of the data, and that visualization is essential for proper data analysis.

The data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4	
1	10	8.04	10	9.14	10	7.46	8	6.58	
2	8	6.95	8	8.14	8	6.77	8	5.76	
3	13	7.58	13	8.74	13	12.74	8	7.71	
4	9	8.81	9	8.77	9	7.11	8	8.84	
5	11	8.33	11	9.26	11	7.81	8	8.47	
6	14	9.96	14	8.1	14	8.84	8	7.04	
7	6	7.24	6	6.13	6	6.08	8	5.25	
8	4	4.26	4	3.1	4	5.39	19	12.5	
9	12	10.84	12	9.13	12	8.15	8	5.56	
10	7	4.82	7	7.26	7	6.42	8	7.91	
11	5	5.68	5	4.74	5	5.73	8	6.89	
Summary Statistics									
N	11	11	11	11	11	11	11	11	
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50	
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94	
r	0.82		0.82		0.82		0.82		

Anscombe's Quartet Four Datasets

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to deceive a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

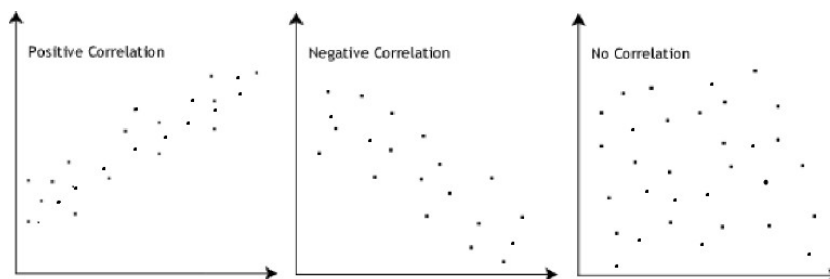
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also called the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It indicates both the direction and strength of the relationship. It is the numerical summary of the linear association between the variables.

Formula to calculate Pearson's R:

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



Interpretation of Pearson's rrr:

- $r=1$: Perfect positive linear correlation. As x increases, y increases in perfect proportion (i.e., the points lie exactly on an upward-sloping straight line).
- $r=-1$: Perfect negative linear correlation. As x increases, y decreases in perfect proportion (i.e., the points lie exactly on a downward-sloping straight line).
- $r=0$: No linear relationship between the variables. The variables do not have any discernible linear pattern, though there might still be some other kind of relationship (e.g., non-linear).
- $0 < r < 1$: A positive linear relationship exists. As x increases, y tends to increase as well. The closer r is to 1, the stronger the positive relationship.
- $-1 < r < 0$: A negative linear relationship exists. As x increases, y tends to decrease. The closer r is to -1, the stronger the negative relationship.

Strength of the Relationship:

- Strong positive correlation: r between 0.7 to 1.
- Weak positive correlation: r between 0.3 and 0.7.
- Very weak / almost no correlation: r between -0.3 and 0.3.

- Weak negative correlation: r between -0.3 and -0.7.
- Strong negative correlation: r between -0.7 and -1.

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 the pre-processing technique of transforming feature values to a similar scale either by standardizing or normalizing the range of independent variables (features) in a dataset, so that each feature contributes equally to the model. Common methods of feature scaling include min-max scaling, which rescales the data to a specific range (usually 0 to 1), and standardization (z- score normalization), which centers the data around zero with a unit variance.

Scaling is especially important for algorithms that rely on distance metrics (like k-NN or SVM) or gradient-based optimization (like linear regression, logistic regression or neural networks), where features with larger values can dominate the model.

For gradient based algorithms: The difference in the ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

For distance based algorithms: Let's say we have data containing high school CGPA scores of students (ranging from 0 to 5) and their future incomes (in thousands Rupees). Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitudes. This will impact the performance of the machine learning algorithm making it biased towards one feature. Therefore, we scale our data before employing a distance based algorithm so that all the features contribute equally to the result.

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean 0 and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: $(x - \min)/(\max - \min)$	Equation: $(x - \text{mean})/\text{standard deviation}$

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>

VIF is given by:

$$VIF = \frac{1}{1-R^2}$$

Variance Inflation Factor (VIF) specifically measures the degree of multicollinearity among the independent variables in a regression model. It tells how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Suppose we have 4 feature variables x_1, x_2, x_3, x_4 . If VIF of variable x_1 is infinite, it means that

$$x_1 = a_2x_2 + a_3x_3 + a_4x_4$$

Where a_2, a_3, a_4 are constants.

This means that x_1 can be perfectly predicted from the other variables, and no additional unique information is being provided by x_1 in relation to the outcome variable.

In such cases of infinite VIF, we have to either remove the perfectly correlated variable or do Principal Component Analysis.

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. For example, if the points are curved, it means that the data are skewed or have heavy tails. If the points are scattered or have gaps, it means that the data have outliers or are multimodal.

Use of Q-Q plot in linear regression:

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.

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.
