Assignment-based Subjective Questions

Student name : Atul Kumar Srivastava

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer.

Following inferences can be drawn

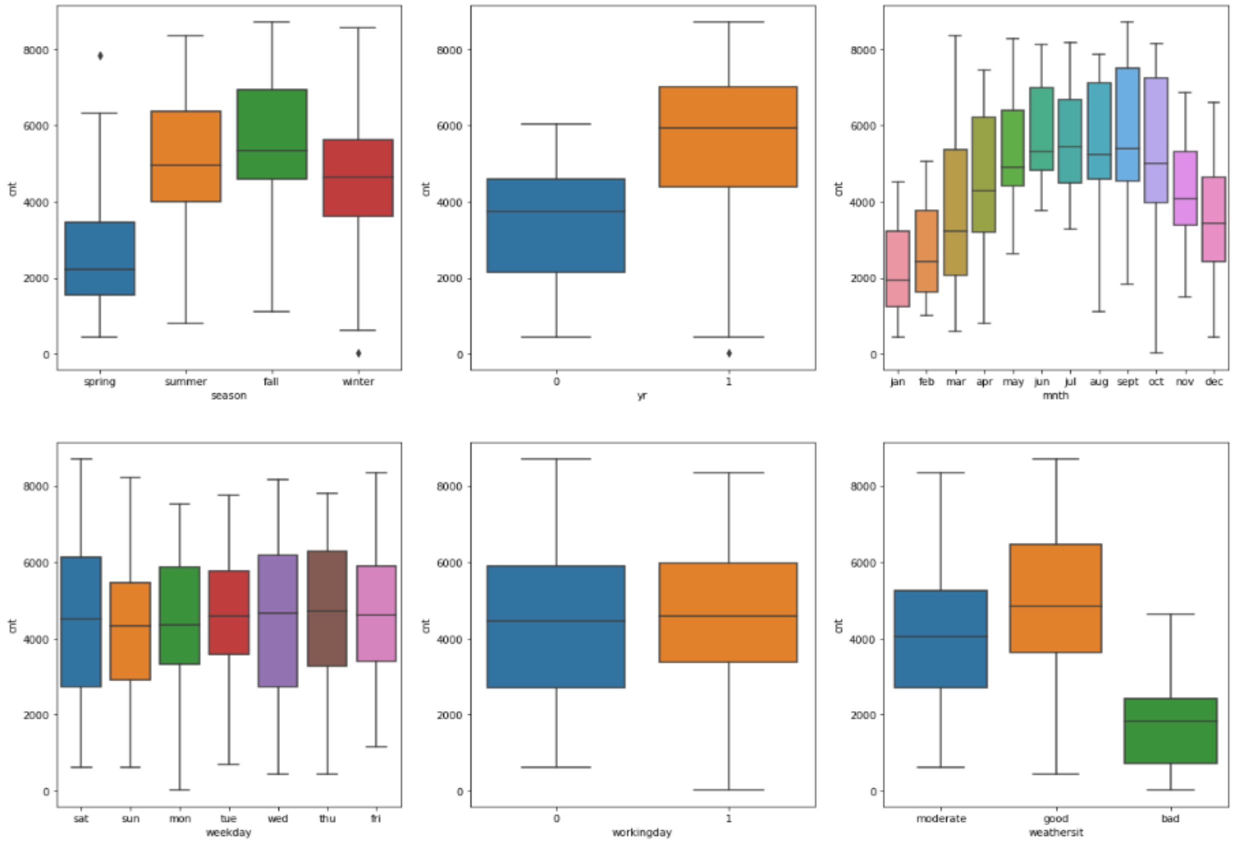
Season : Season can be good predictor, as it has been observed that season3 had 32% booking and season 2 and season 4 were after that.

Mnth : Month is also a good predictor as it was observed that 10% booking were in the month 5,6,7,8,9.

Weathersit : It was observed that 67% booking were during weathersit1 followed by weather2. Therefore a good predictor

Holiday : As more that 97 percent booking were happening on a non holiday. It is not a good predictor.

Workingday : 69% booking were happening on weekday. Meaning working day is good predictor.

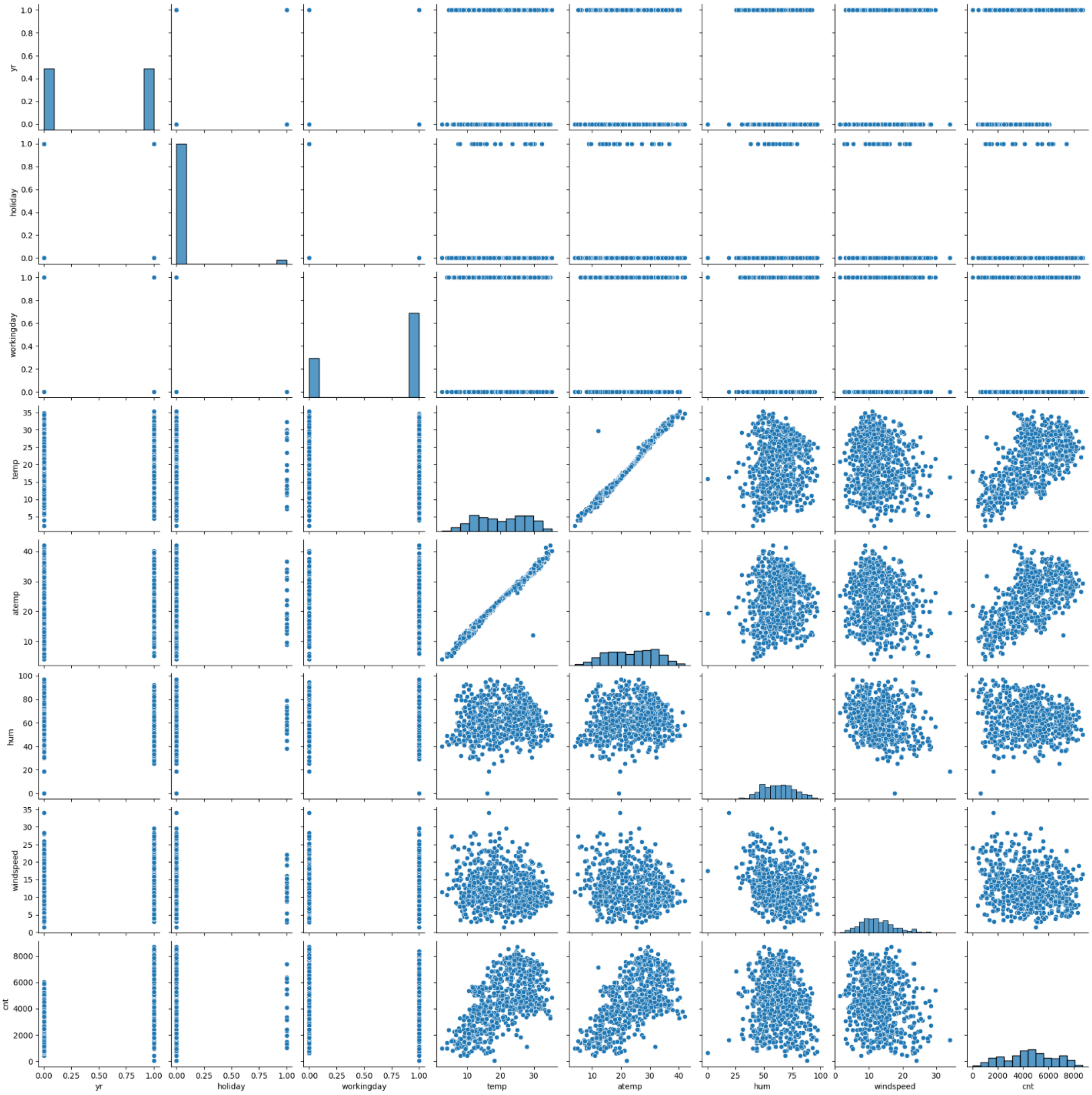


Q2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True while creating dummy variables helps to prevent multicollinearity which means that when independent variables in a model are highly correlated, which can cause problems in estimating coefficients and interpreting the model. By dropping the first dummy variable, we can avoid linear dependence among dummy variable.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer : It can be observed that temp and atemp has the highest correlation.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

i) Error terms should be normally distributed with mean zero
ii) There should be insignificant multicollinearity among variables and There should be Linear relationship validation
iii) Error terms is constant all of x.
iv) No auto-correlation

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

i) temp
ii) weathersit
iii) yr

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer

Linear regression is a fundamental statistical method which is used to model the relationship between a dependent variable (also called the response or target variable) and one or more independent variables (also called predictors or features). The goal is to find the best-fitting straight line through the data points that can predict the dependent variable based on the values of the independent variables.

1. Simple Linear Regression

Model

In simple linear regression, the relationship between the dependent variable y and a single independent variable x is modeled as a straight line:

$$y=\beta_0+\beta_1x+\epsilon$$

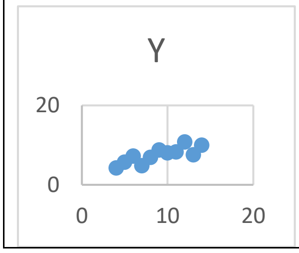
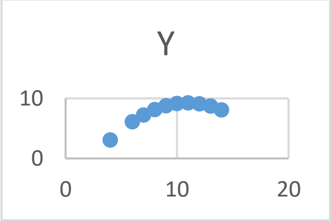
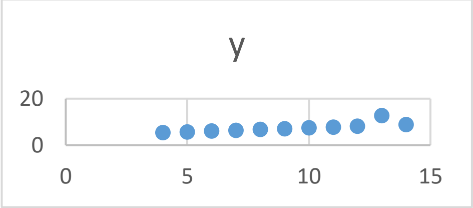
y is dependent variable
x is independent variable
β_0 is the intercept, the value of y when x is zero.

β1 is the slope (the change in y for a one-unit change in x).
ϵ epsilon ϵ is the error term (the difference between the observed and predicted values of y).

Assumptions
Linearity: The relationship between x and y is linear.
Independence: The observations are independent of each other.
Homoscedasticity: The variance of the error terms is constant across all levels of x
Normality: The error terms are normally distributed.

Q2. Explain the Anscombe's quartet in detail.

Answer :
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. The quartet demonstrates the importance of graphing data before analyzing it and highlights how summary statistics alone can be misleading. It was developed by statistician Francis Acncombe.

Dataset1		Dataset2		Dataset3	
X	Y	X	Y	x	y
10	8.04	10	9.14	10	7.46
8	6.95	8	8.14	8	6.77
13	7.58	13	8.74	13	12.74
9	8.81	9	8.77	9	7.11
11	8.33	11	9.26	11	7.81
14	9.96	14	8.10	14	8.84
6	7.24	6	6.13	6	6.08
4	4.26	4	3.10	4	5.39
12	10.84	12	9.13	12	8.15
7	4.82	7	7.26	7	6.42
5	5.68	5	4.74	5	5.73
					

Same parameters for descriptive statistics

□ **Mean of x:** 9
□ **Variance of x:** 11
□ **Mean of y:** 7.5
□ **Variance of y:** 4.125

□ **Correlation between x and y:** ~0.816
□ **Linear regression line:** y=3+0.5x
But graphs drawn are different.

Q3. What is Pearson's R?

Answer:
Pearson's r, also known as Pearson's correlation coefficient, is a measure of the linear relationship between two continuous variables. It is a widely used statistical measure that quantifies the strength and direction of the linear association between two variables. Pearson's r ranges from -1 to 1.

$$r = \frac{\sum_{i=1}^{n}(xi - xbar)(yi - ybar)}{\sqrt{\sum_{i=1}^{n}(xi - xbar)2}\ \sum_{i=1}^{n}(yi - ybar)2}$$

Where:

n is the number of pairs of scores.
xi and yi are the individual sample points indexed with i.
xbar and ybar are the mean values of x and y

r=1: Perfect positive linear relationship. As x x increases, y y increases proportionally.
r= −1 : Perfect negative linear relationship. As x increases, y decreases proportionally.
r= 0 : No linear relationship. The variables do not have a linear association.

Pearson's r is a unit free measure.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer

Scaling is the process of transforming the features of your data so that they lie within a specified range or have specific statistical properties. This is often an essential step in data preprocessing for many machine learning algorithms, as it ensures that all features contribute equally to the result and can improve the performance and convergence speed of algorithms.

Scaling is Performed as it

i) improves Convergence:
ii) Equal Contribution: Ensures that all features contribute equally to the result.
iii) Prevents Numerical Issues: Large differences in feature scales can lead to numerical instability and issues in matrix operations, particularly for algorithms like linear regression and principal component analysis (PCA).

Types of Scaling

Normalized Scaling (Min-Max Scaling)
Normalization scales the data to a fixed range, typically [0, 1]. This method is sensitive to outliers since it uses the minimum and maximum values of the data.

Standardized Scaling (Z-score Scaling)
Standardization scales the data to have a mean of 0 and a standard deviation of 1. This method is less sensitive to outliers and is particularly useful when the data follows a Gaussian distribution.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer :
Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity in a set of multiple regression variables. Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy.

VIF becomes infinite when $R\mathrm{i}^{2}=1$. This happens when a predictor variable Xi can be perfectly explained by linear combination of other predictor variable. Or we can say there is perfect multicollinearity.
Division by Zero: The formula for VIF involves dividing by (1-Ri^{2}). When Ri^{2} is 1.
This causes model instability. It can also means that predictors are duplicate.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer 6.
A Q-Q (quantile-quantile) plot is a graphical tool to assess if a set of data follows a particular theoretical distribution. It plots the quantiles of the data against the quantiles of a specified theoretical distribution (often the normal distribution). If the data follows the specified distribution, the points on the Q-Q plot will approximately lie on a straight line.

Construction of a Q-Q Plot
Order the Data: Sort the sample data in ascending order.
Compute Quantiles: Calculate the theoretical quantiles from the specified distribution.
Plot the Points: Plot the sample quantiles (y-axis) against the theoretical quantiles (x-axis).
Use and Importance in Linear Regression
In the context of linear regression, a Q-Q plot is primarily used to assess the normality of the residuals (errors). One of the assumptions of linear regression is that the residuals are normally distributed. Checking this assumption is crucial because:

Validates Inferences: Many statistical tests and confidence intervals in linear regression rely on the normality of residuals. If the residuals are normally distributed, the estimates and tests based on the model are reliable.
Identifies Skewness and Kurtosis: A Q-Q plot can reveal deviations from normality, such as skewness (asymmetry) and kurtosis (peakedness or flatness) in the residuals.

Detects Outliers: Points that deviate significantly from the straight line in a Q-Q plot may indicate outliers in the data.

Interpretation of a Q-Q Plot

Straight Line: If the points roughly follow a straight line, the data is approximately normally distributed.

S-shaped Curve: This indicates heavy tails in the data, suggesting the presence of outliers or a leptokurtic distribution.

Inverted S-shaped Curve: This suggests light tails in the data, indicating a platykurtic distribution.

Upward/Downward Curve: Indicates skewness. An upward curve indicates right skewness, while a downward curve indicates left skewness.