Assignment-based Subjective Questions

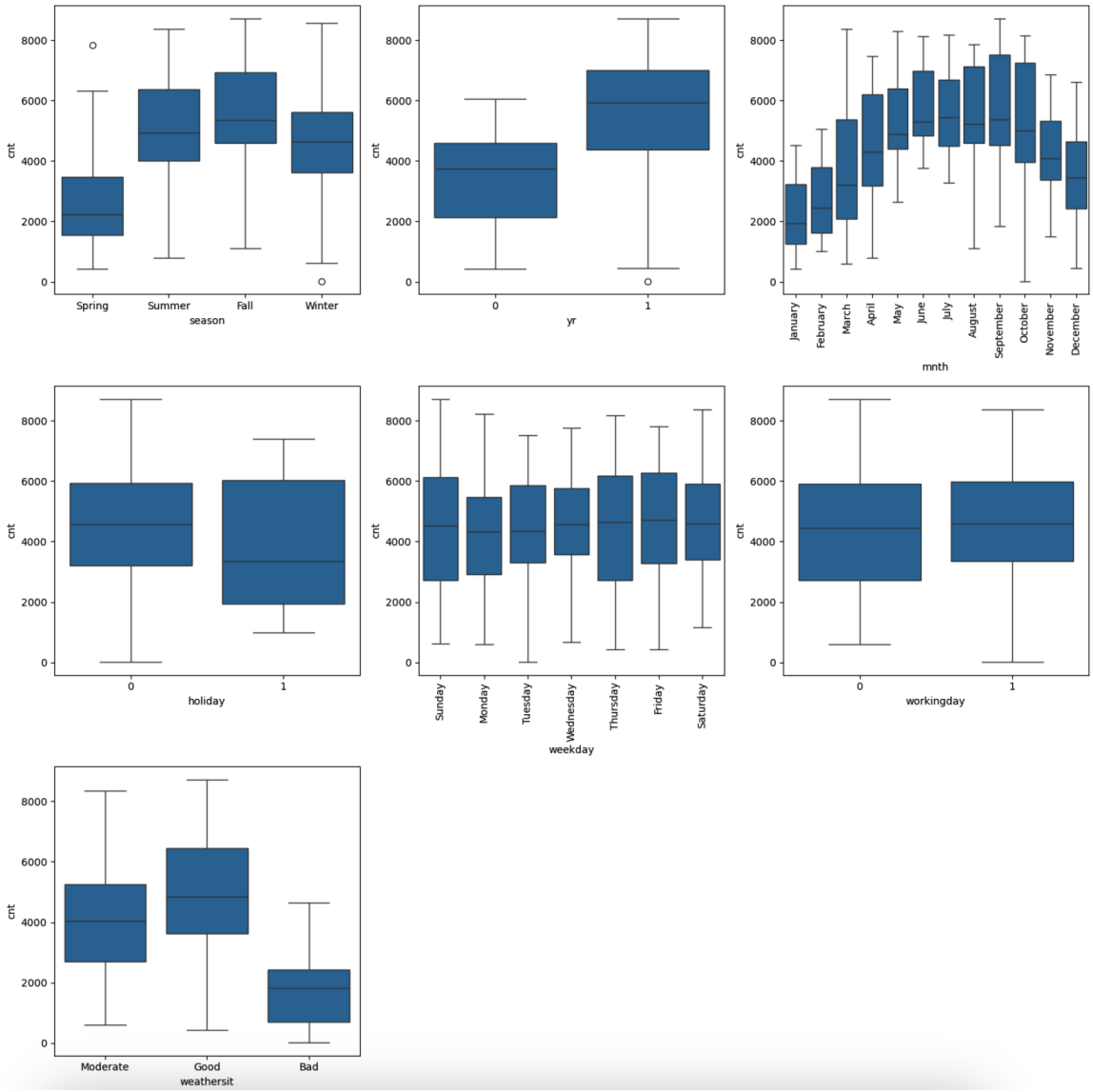
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

There are a couple of categorical variables namely season,mnth,yr,weekday,
working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'

. The below fig shows the correlation among the same



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)

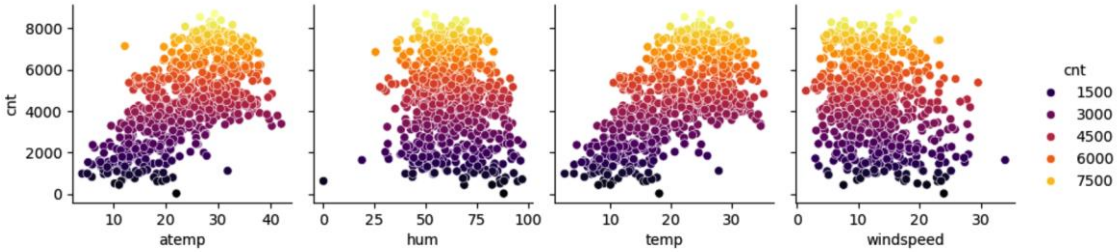
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.
Eg: If there are 3 levels, the drop_first will drop the first column.

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 the plot , we can see that temp has the highest correlation with target variable

---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

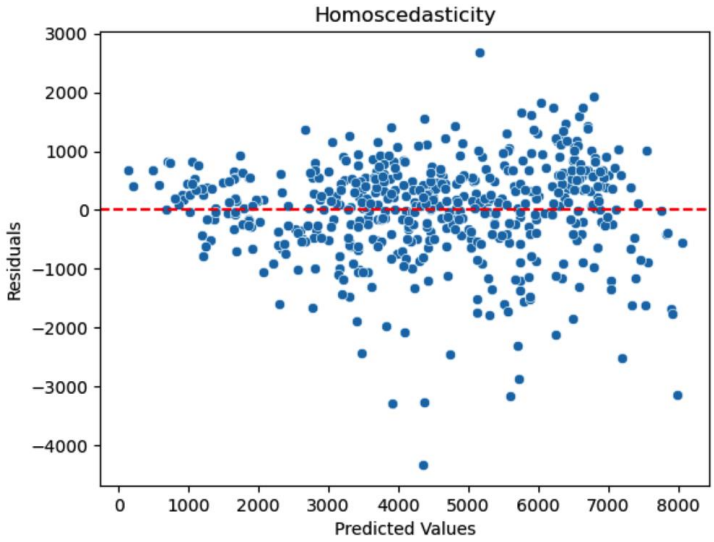
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

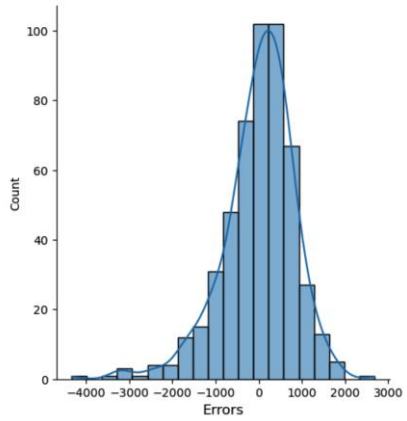
1. **Linearity:** The relationship between the independent and dependent variables is assumed to be linear.

2. **Independence:** Residuals (the differences between observed and predicted values) should be independent of each other.

3. **Homoscedasticity:** The variance of residuals should be constant across all levels of the independent variables, indicating consistent levels of variability.



4. **Normality of Residuals:** The residuals should be approximately normally distributed.



- *The error terms follow the principle of a normal distribution curve.*

5. **No Perfect Multicollinearity:** Independent variables should not exhibit high correlation with each other, avoiding multicollinearity issues.

22]:

	Features	VIF
1	workingday	1.63
8	Sunday	1.63
6	Winter	1.24
2	temp	1.20
5	Summer	1.19
3	hum	1.17
4	windspeed	1.13
7	September	1.11
0	yr	1.02

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)

cnt = 4491.30 + 998.75 x ***yr*** + 178.28 x ***workingday*** + 1174.49 x ***temp*** - 429.07 x ***hum*** - 349.15 x ***windspeed*** + 344.84 x ***Summer*** + 526.80 x ***Winter*** + 234.70 x ***September*** + 159.98 x ***Sunday***

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

---

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 form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.
In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

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 group of four datasets that contain summary statistics (such as mean, variance, correlation, and regression line), but have very different visual styles. Created by Francis Anscombe in 1973, these datasets emphasize to the limitations of relying only on summary statistics for data analysis Main lessons: Importance of Visualization: Data visualization reveals patterns, relationships, and outliers that statistics alone miss.
1. Limitations of summary statistics: Statistics such as mean, variance, and correlation. may cause misunderstanding Especially with outliers or non-linear relationships.
2. Exploratory Data Analysis (EDA): Anscombe's Quartet emphasizes the value of EDA, including data visualization. for accurate data interpretation

Essentially, Anscombe's Quartet emphasizes the need for visualization and statistical analysis to ensure a true understanding of the 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>

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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>

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modelling.
Difference between Normalizing Scaling and Standardize Scaling:
In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
Normalized scaling is called as scaling normalization whereas standardized scaling is 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>

VIF (VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value 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>

Q–Q plot is a probability plot, which is a graphical method for comparing two probabilities distributions by plotting their quantiles against each other.
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.
Importance of QQ Plot in Linear Regression :
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.
Advantages:
● It can be used with sample size also
● Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot
Q-Q plot use on two datasets to check
● If both datasets came from population with common distribution
● If both datasets have common location and common scale
● If both datasets have similar type of distribution shape
● If both datasets have tail behaviour