Assignment-based Subjective Questions

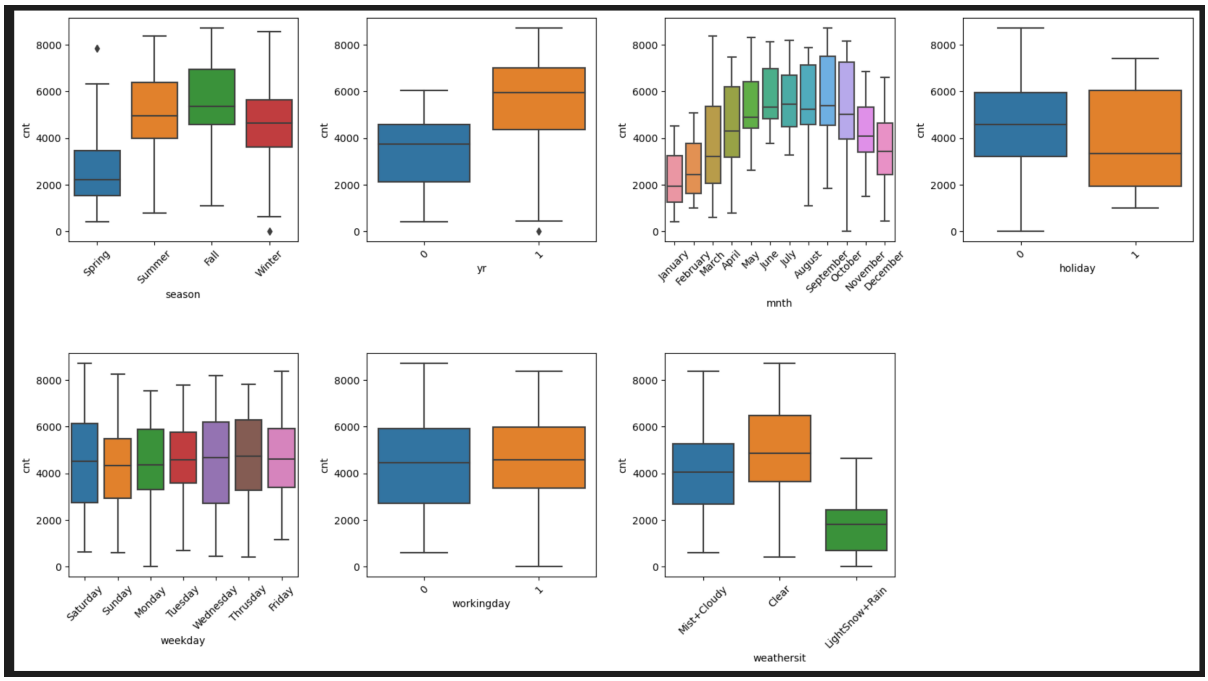
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

In the dataset, we have 7 categorical variables, namely, season, yr mnth, holiday, weekday, workingday and weathersit.

The following is the box plot for the categorical variables against the target variable (cnt).



The following are the inferences

1. We see higher demand in the 'Fall' season.
2. The yr with value 1, which is year 2019 have more demand as compared to 2018.
3. There is a increase in demand between months August, September and October. There is sharp decline in months November and December.
4. On Saturdays, there is a slight increase in demand as compared to other days.
5. The weather situation 'Clear' have significantly positive impact on the bike demand.
6. During holidays, there is a reduced demand of bikes.
7. On working days, the median value of bike sharing is almost 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)

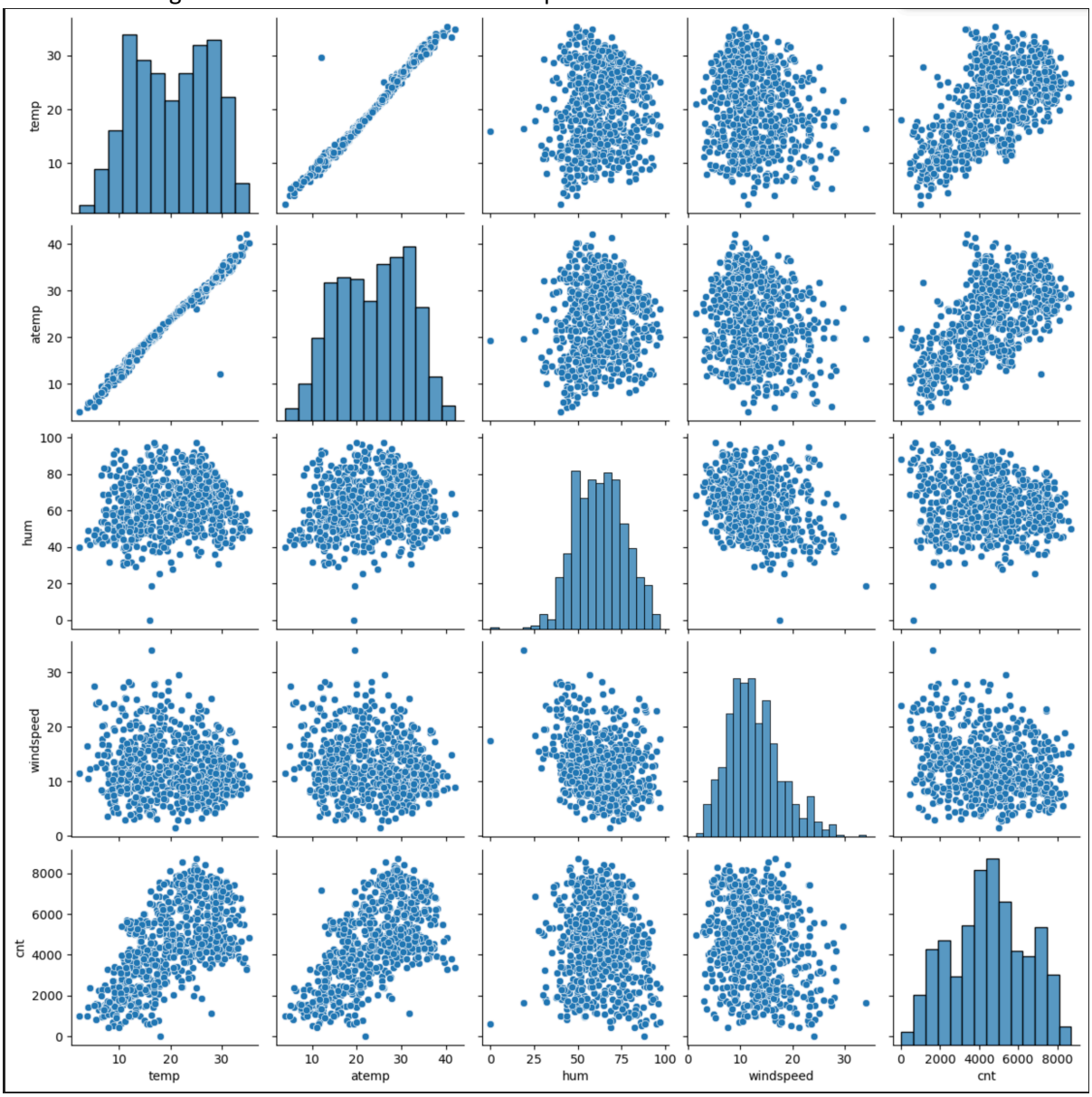
We use get_dummies function to convert the categorical variables which have more than 3 values to convert into binary form. While using the function we use the parameter drop_first=True, so that multicollinearity can be prevented in the regression model.

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)

There is a strong correlation between cnt and temp.



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

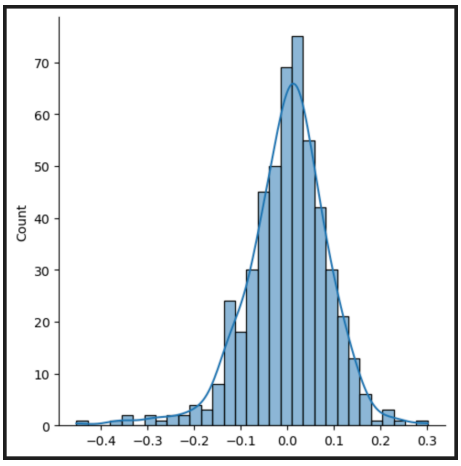
Total Marks: 3 marks (Do not edit)

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

To validate the assumptions of Linear regression we have performed the Residual Analysis. In this method we have found the error values between training data of cnt and predicted data of cnt. The error data is then plotted, and we observed that

1. The error data is normally distributed.
2. The mean of error data is 0.
3. Multicollinearity should be 0. We have ensured that the VIF is calculated to ascertain correlation of the variables. VIF < 5.

4. The pairplot of the numerical variables provided a strong collinear relationship between temp and atemp with the target variable.



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)
1. temp
2. yr
3. weathersit_LightSnow+Rain

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 algorithm is supervised algorithm used in statistics and machine learning. The algorithm assumes there is a Linear relationship between the Target variable (Y) and Independent variable (X). The statistically it is represented as

$$Y=\beta_0+\beta_1X+\epsilon$$

where
Y = Target variable
X = Independent variable
β_0 = Intercept
β_1 = Slope
ϵ = Error Term.

When we have multiple independent variables, the equation converts to
$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\epsilon$

In Linear regression algorithm, following assumptions are made

1. The independent and Target variable have Linear relationship
2. There is a constant variance of error terms.
3. Error terms are normally distributed.
4. There is no multicollinearity within the features.eib

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 comprises of 4 distinct datasets where the statistical properties are identical across the datasets. Each of the dataset comprises of identical values of R-squared, mean, variance and Linear regression, however when plotted on graph they exhibit distinct patterns. It thereby proves that Visual representations is very important along with Statistical representation.
The four datasets of the quartet contains each quartet containing 11 pairs of x-y value. When plotted, these represent different relationships, variability and correlations, however the statistical properties of each quartet are identical.

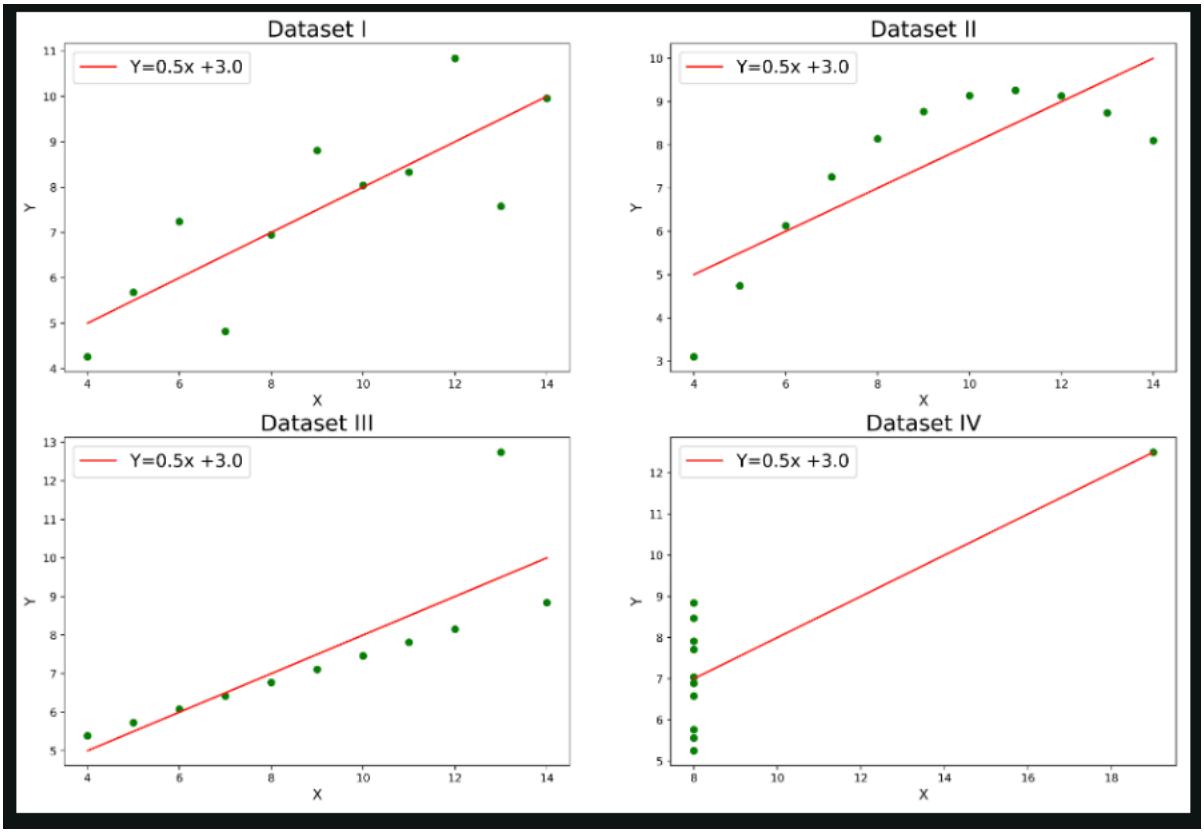
Consider the following data

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

The following are the statistical properties.

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
slope	0.500091	0.500000	0.499727	0.499909
intercept	3.000091	3.000909	3.002455	3.001727

Scatter plot gives a very different interpretation



As we can see that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when plotted praphically.

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 or Pearson's coefficient is used to measure the relationship between two variables. The coefficient identifies the how strong the relation is between the two variables when plotted on scatter plot. It's value ranges from -1 to +1, which indicates that it not only provides the magnitude of the strength of relationship but also the direction.

The values indicate as follows

+1 indicates a perfect positive linear regression.

-1 indicates a perfect negative linear regression.

0 indicates no linear relationship.

Formula for calculation

$r = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$

In python, the function "pearsonr()" of "scipy.stats" library provides the value.

```
from scipy.stats import pearsonr
r, p_value = pearsonr(X, Y)
```

```
## X and Y are numpy array, r is pearson's coeff
```

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 is a technique used to standardize or normalize the data within a specific range, so that the independent variables have a similar scale thereby ensuring that they don't have coefficients which are unnecessarily bloated. Scaling effects only the coefficients and the other statistical values like R-squared, t-statistic are not affected.
Normalized Scaling – Scales data between 0 and 1. It takes care of outliers, but it sensitive to them.

$$X' = (X - X_{min}) / (X_{max} - X_{min})$$

Standardized scaling – Scales data in such a way that mean is 0 and standard deviation is 1. It is less sensitive to outliers.

$$X' = (X - \mu) / \sigma$$

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 defines the multicollinearity between the variables, and it is denoted by the below equation

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

When VIF is infinite, this means the denominator (1 – R^2) is zero, thereby R-squared is 1. This indicates that there is a perfect multicollinearity between the independent variables. It can be in cases when there is duplicate or highly correlated variables.

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 Q-Q Plot or Quantile- Quantile plot is a visual representation to understand if the dataset follows certain probability distribution. It is used in statistics, data analysis to check if the assumptions made in the model are correct and the calculated values don't deviate from their

actual values. The quantiles from one dataset are plotted against quantiles from another. If the points closely align along a diagonal line, then it suggests a similarity between distributions.
The following is an example of Q-Q plot