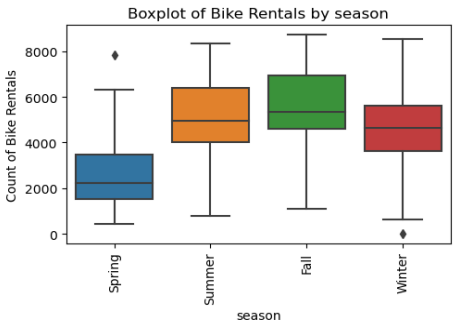
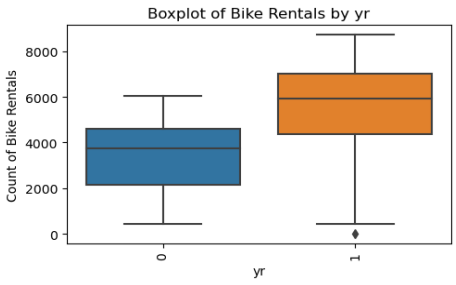
Assignment Subjective Questions

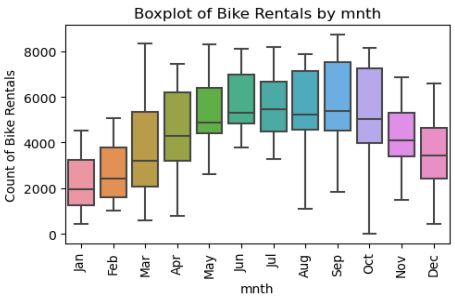
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

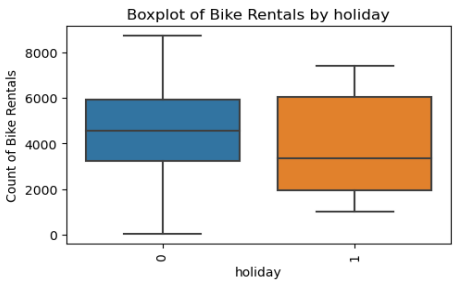
Answer: The dependent variable was plotted in boxplot with categorical variables and have found the following insights:

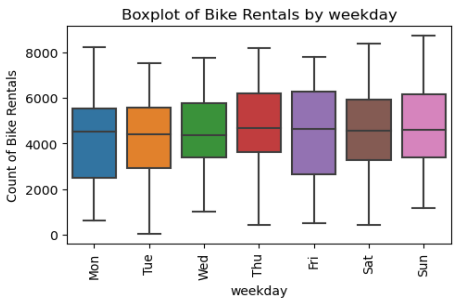
- ❖ Fall has the highest average rental, followed by Summer.
- ❖ Demand is continuously increasing from January to June and after September it decreases. September tops the monthly rentals.
- ❖ No significant difference in rentals is seen across the weekdays.
- ❖ When there is holiday, bike demand gets decreased.
- ❖ Clear weather attracted more bookings which is obvious.
- ❖ No. of booking seems to be equal on working days as well as on non-working days.
- ❖ In 2019, the number of bookings increased significantly, hence showing a growth in business.

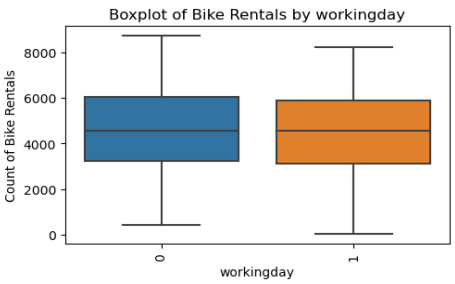


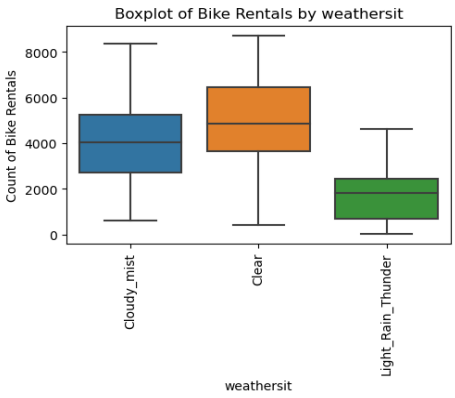










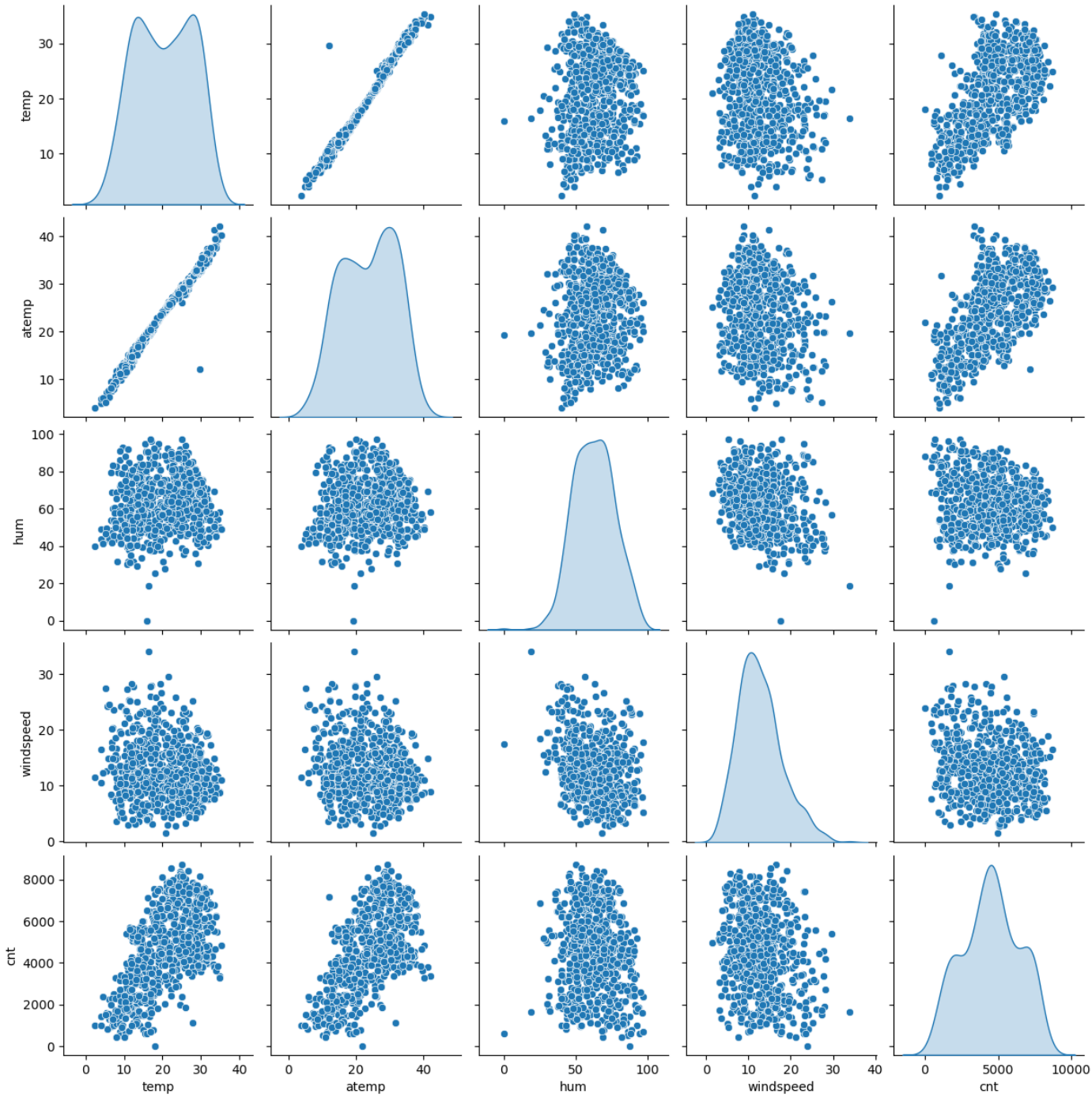


2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   Answer: drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Answer: 'temp' variable has the highest correlation with the target variable.



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

Answer: I have checked the following assumptions:

- ❖ Error terms are normally distributed with mean 0.
- ❖ Error Terms do not follow any pattern.
- ❖ Multicollinearity check using VIF(s).
- ❖ Linearity Check.
- ❖ Ensured the overfitting by looking the R2 value and Adjusted R2.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Answer: Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

   - ‘yr’ as it has a positive coefficient indicating an increase in number of rentals from 2018 to 2019
   - ‘temp’ also has a positive coefficient which means that it impacts the number of rentals
   - ‘weathersit_Light_Rain_Thunder’ indicates that during the Rains and Thunderstorm weather, the demand is very less as who will use bikes in that weather condition. Though this feature has ne gative coefficient but it explains when bike is less consumed.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Answer: Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The goal of linear regression is to find the best-fit linear relationship between the independent variables and the dependent variable. This relationship is represented by a linear equation in the form:

$y=b0+b1*x1+b2*x2+\ldots+bn*xn$

Here,
- y is the dependent variable (the variable we want to predict),
- $x1,x2,\ldots,xn$ are the independent variables,
- $b0$ is the y-intercept (the value of y when all x values are 0),
- $b1,b2,\ldots,bn$ are the coefficients (representing the change in y for a unit change in the corresponding x).

The linear regression algorithm involves two main steps:

1. **Training the Model:** Given a dataset with known values of both the independent and dependent variables, the algorithm aims to find the optimal values for the coefficients that minimize the difference between the predicted values and the actual values. This process is often done using the method of least squares, where the sum of the squared differences between the observed and predicted values is minimized.

2. **Making Predictions:** Once the model is trained, it can be used to predict the dependent variable for new, unseen data. The linear equation, with the learned coefficients, is applied to the new input data to calculate the predicted outcome.

The performance of a linear regression model is often evaluated using metrics like R^2.

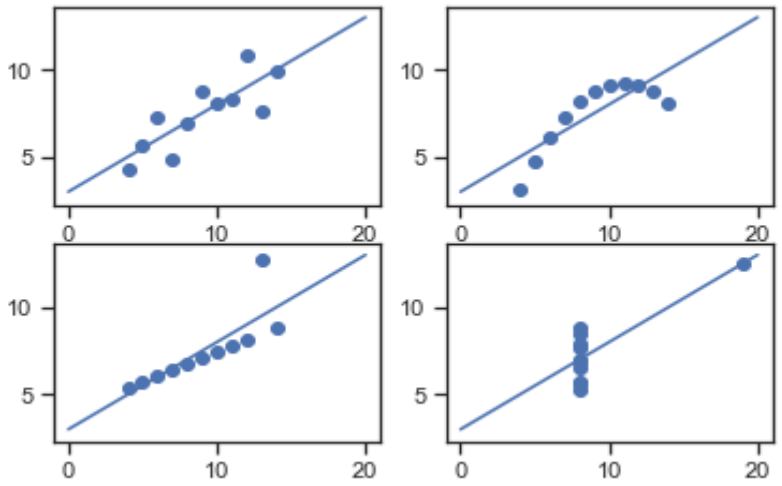
2. **Explain the Anscombe’s quartet in detail. (3 marks)**

Answer: 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

When we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



- 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.

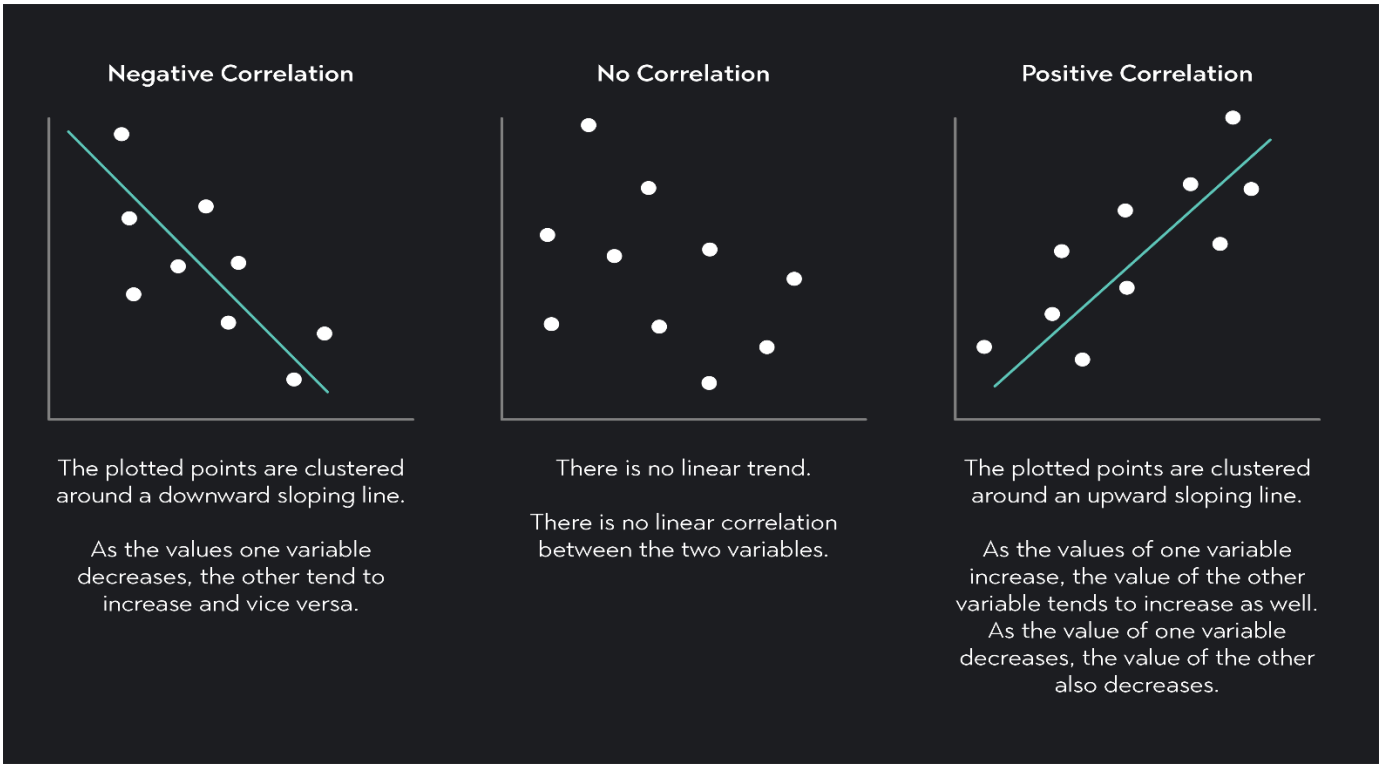
3. What is Pearson's R? (3 marks)

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

Pearson's r measures the strength of the linear relationship between two variables.

Pearson's r always between -1 and 1.

If data lie on a perfect straight line with negative slope, then r = -1.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: 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 al gorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not perfor med than algorithm tends to weigh high values magnitudes and ignore other parameters which will resul t in incorrect modelling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is us ed to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by o utliers.
5. Normalized scaling is used when we don’t know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Sc ore Normalization

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The value of VIF is calculated by the below formula:

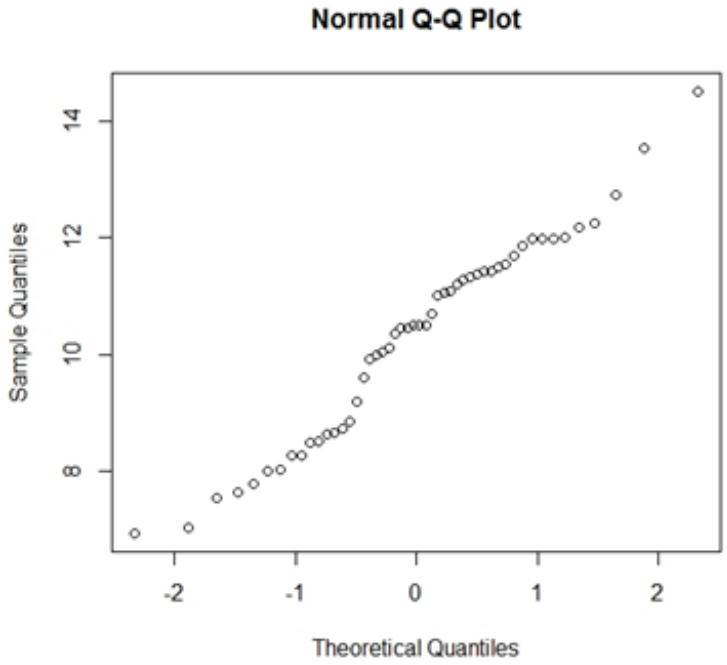
$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall va lue become infinite. It denotes perfect correlation in variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression:

The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

i. The sample sizes do not need to be equal.
ii. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
iii. The q-q plot can provide more insight into the nature of the difference than analytical methods.