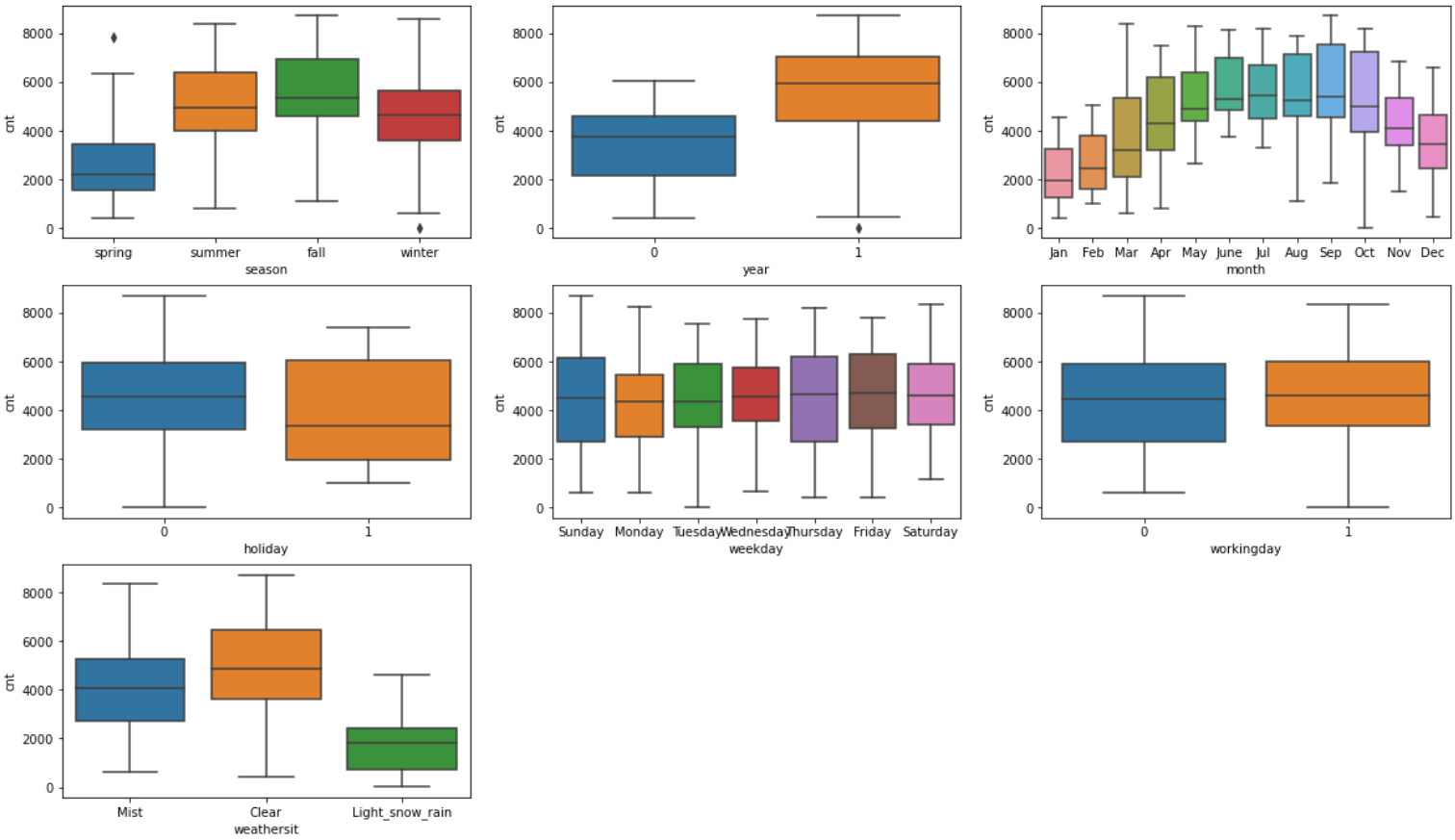
Assignment-based Subjective Questions

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

A: Based on the analysis of categorical values using box plots and bar plots. Below are the few points that can be inferred.

- Categorical variables season,month,year,holiday,weekday, working day and weathersit have a major effect on the dependent variable ‘cnt’. The below fig shows the correlation among the same



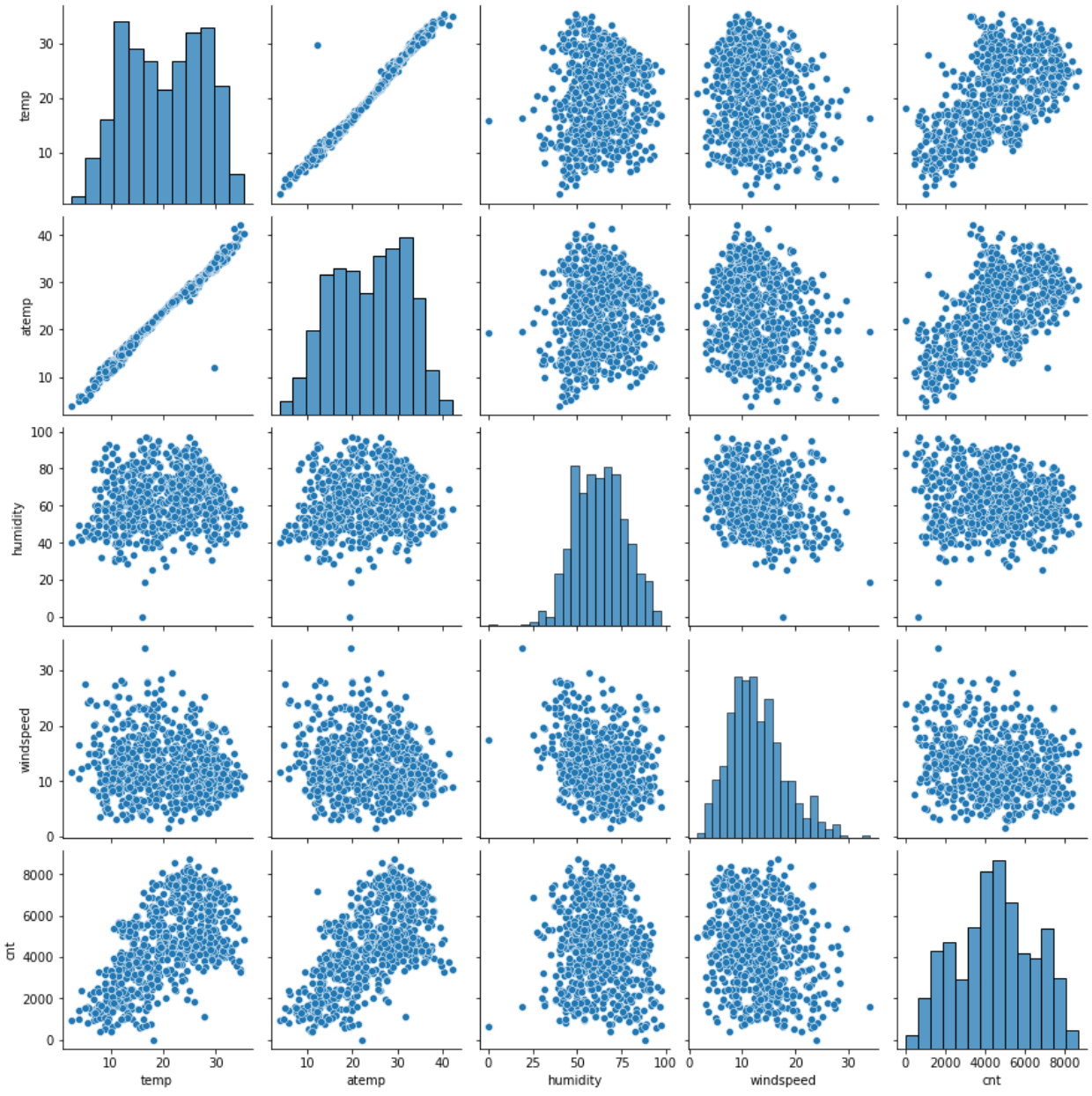
- Season: Fall has highest demand for rental bikes
- Every month the demand for rental bikes increases till June. September has the highest number of demand then it is decreasing.
- During weekdays and working days demands don't have that much variation.
- Clear weathersit has highest demand
- When there is a holiday, demand has decreased.
- I see that demand for next year has grown

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

A. drop_first = True is used so that we can reduce the extra columns created using dummy variables. In addition it also reduces the correlation created among the dummy variables. If we don't drop the existing variable, then correlation would increase and our model will not lean.

Ex: If there are 3 levels, the drop_first will drop the first column

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



A.

The 'temp' and 'atemp' variables have the highest correlation when compared to the rest with the target variable as 'cnt'.

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

A. Linear Regression models are validated based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, Multicollinearity.

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

A. Top 3 features that has significant impact towards explaining the demand of the shared bikes are **temp,year and winter**

General Subjective Questions

1. Explain the linear regression algorithm in detail?

A. Linear regression is a form of predictive modeling 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.

2. **Explain the Anscombe's quartet in detail?**

A. Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed.

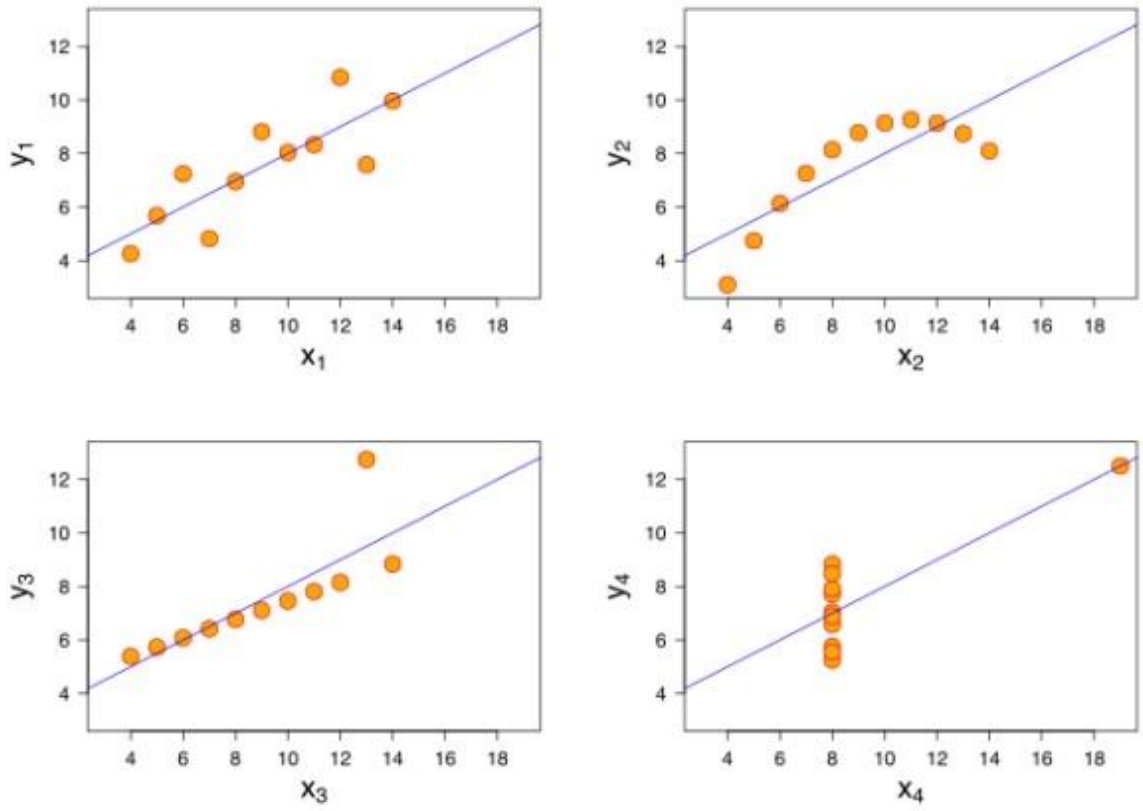
Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- **1st data set fits linear regression model as it seems to be linear relationship between X and y**
- **The 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.**
- **The 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.**
- **The 4th data set has a high leverage point means it produces a high correlation coeff.**
- **Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before building a machine learning model.**

3. What is Pearson's R?

A. 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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

A. 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 a 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 then the algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardizing Scaling:

1. In normalized scaling minimum and maximum value of features are 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 used 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 outliers.
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 scaling normalization whereas standardized scaling is called Z Score Normalization.

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

A. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

A. Q–Q plot is a probability plot, which is a graphical method for comparing two probability 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 a 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 behavior