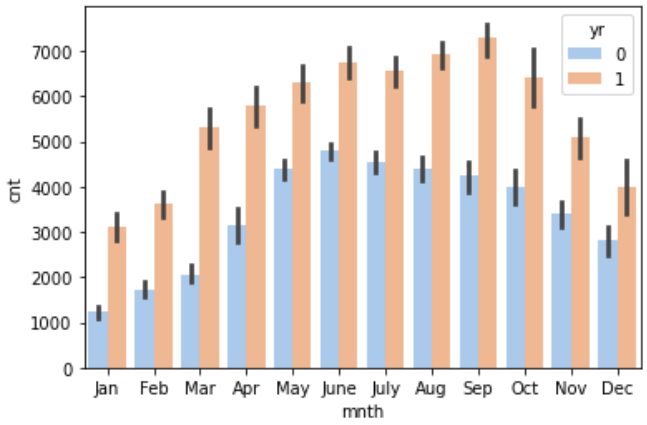
Assignment-based Subjective Questions

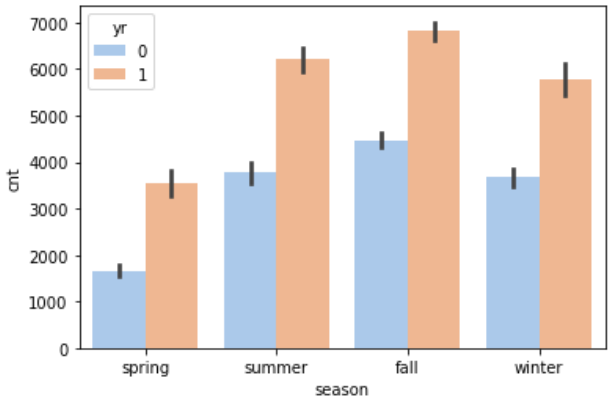
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

We can say that:

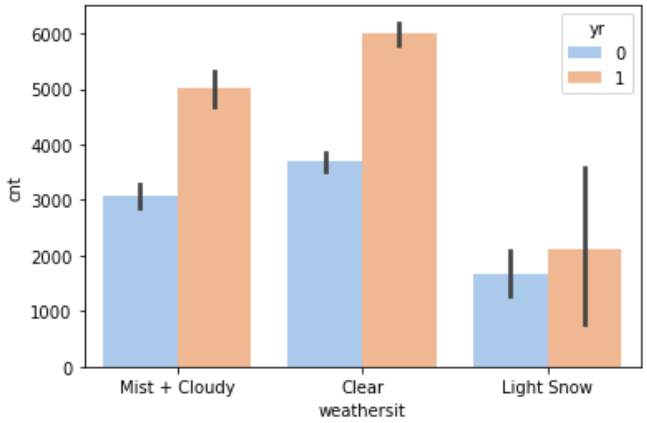
- In year 2019, the count is very high compared to previous year.



- Summer and fall have high number of count than spring and winter



- In clear weather, we have more count compared to others.



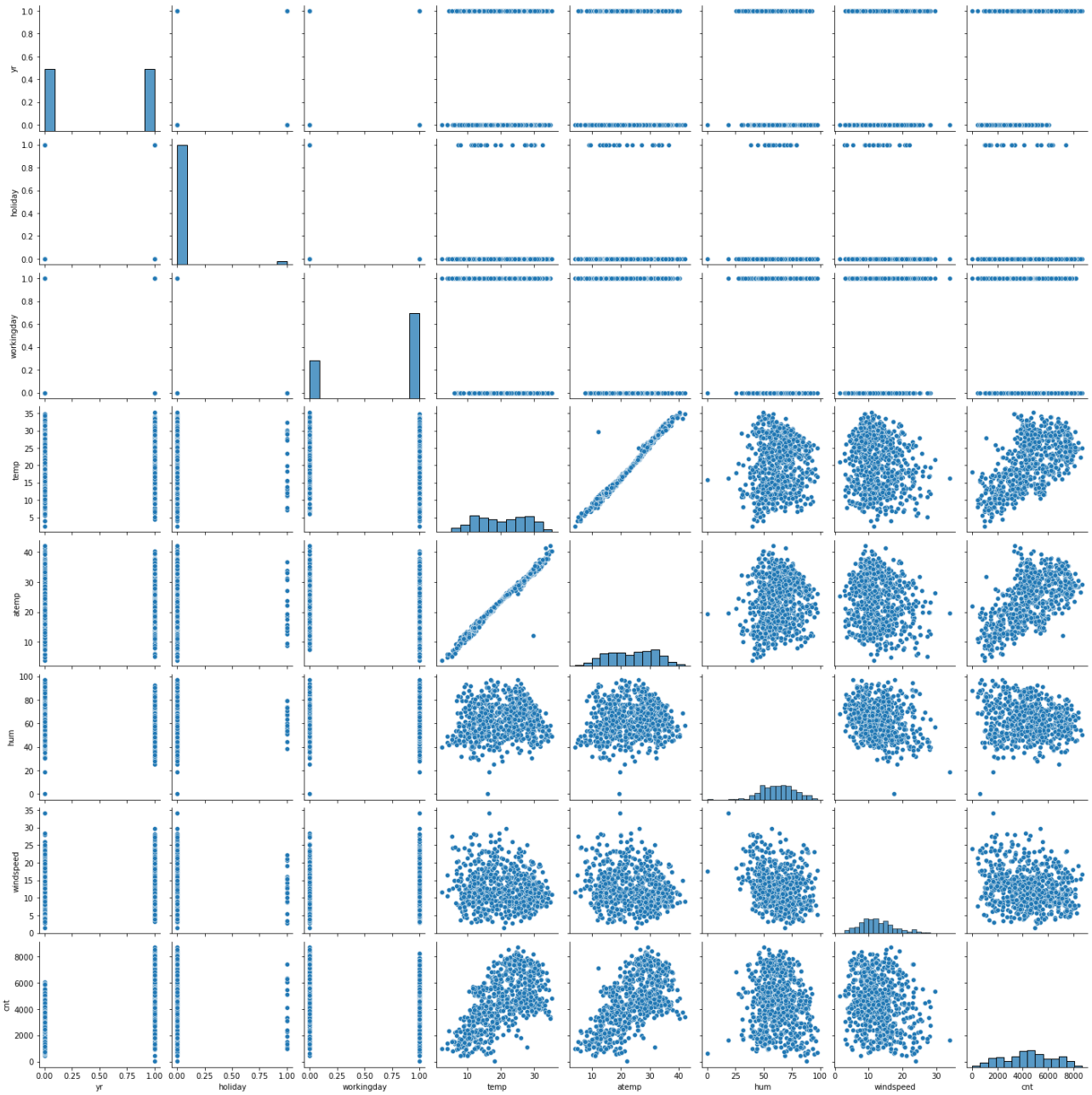
-

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

Drop first is necessary to remove the redundant variables. Drop first is essential when creating dummy variables because it helps to avoid the problem of data multicollinearity. If the first column is not removed, the model estimates will be unstable and biased due to perfect correlation with the other dummy variables.

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

Temp and atemp are showing highest correlation, we can observe from the pair plot.

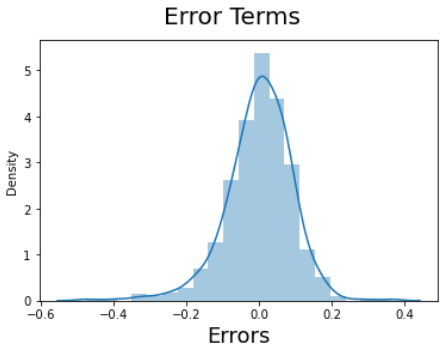


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

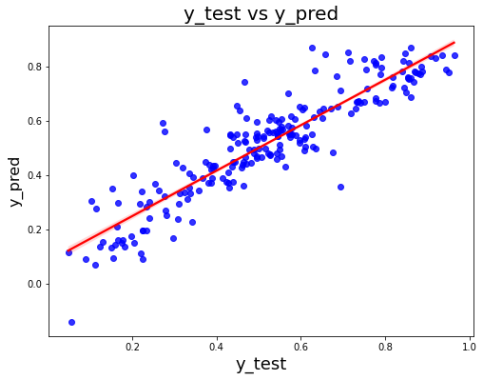
- We can validate the assumptions by looking at the pair plot, the relation between the dependent and independent variables.
- Multicollinearity with the help of heatmap and VIF

	Features	VIF
16	temp	4.49
11	spring	3.69
13	winter	2.90
12	summer	2.41
2	Jan	2.21
14	yr	2.09
1	Feb	1.88
5	Nov	1.80
3	July	1.61
10	Mist + Cloudy	1.57
4	May	1.56
0	Dec	1.55
6	Sep	1.35
7	Mon	1.23
8	Tue	1.23
9	Light Snow	1.08
15	holiday	1.07

- Errors are normally distributed.



- Homoscedastic (constant variance)



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

General Subjective Questions

Q: Explain the linear regression algorithm in detail.

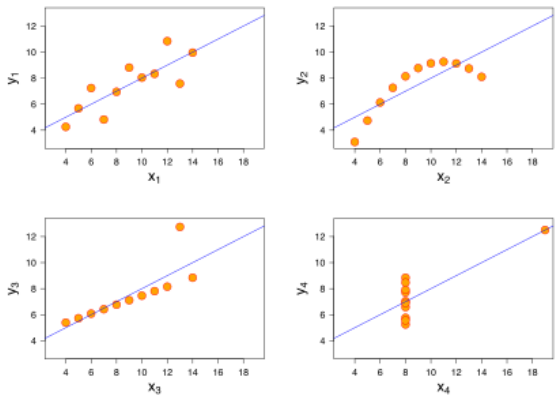
Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$y = a + bx$

Q: Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



Q: What is Pearson's R?

The Pearson correlation coefficient is a descriptive statistic, which means it summarizes a dataset's characteristics.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson’s r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient.

It describes the strength and direction of a linear relationship between two quantitative variables in particular. Although different disciplines have different interpretations of relationship strength (also known as effect size). In addition, the Pearson correlation coefficient is an inferential statistic, which means it can be used to test statistical hypotheses. We can specifically test for a significant relationship between two variables.

Q: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a data Pre-Processing step that is applied to independent variables in order to normalize the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the collected data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the algorithm only considers magnitude rather than units, resulting in incorrect modelling. To solve this problem, we must scale all of the variables to the same magnitude level.

Normalized Scaling - It gathers all data between 0 and 1. sklearn.preprocessing. MinMaxScaler aids in the implementation of normalization in Python. Values are replaced by their Z scores after standardization.

Standardized Scaling - It transforms the data into a standard normal distribution with a mean () of zero and a standard deviation of one (). sklearn.preprocessing. Python's scale aids in the implementation of standardization. One disadvantage of normalization over standardization is that it removes some data information, particularly about outliers.

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

The value of VIF is infinite when there is perfect multicollinearity in the data, meaning that the predictor variables are perfectly correlated with each other, making it impossible to determine the unique effect of each variable on the outcome.

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

A Q-Q plot (Quantile-Quantile Plot) is a graphical method to check if a set of data is approximately normally distributed. It plots the sample quantiles against the theoretical quantiles of a normal distribution.

In linear regression, a Q-Q plot is used to check the assumption of normality of residuals, which is important for valid inference. If the residuals are not normally distributed, it can affect the validity of statistical tests and confidence intervals for the regression coefficients. A Q-Q plot helps to visually inspect if the residuals are close to a straight line, indicating normality, or if there are deviations, suggesting non-normality.