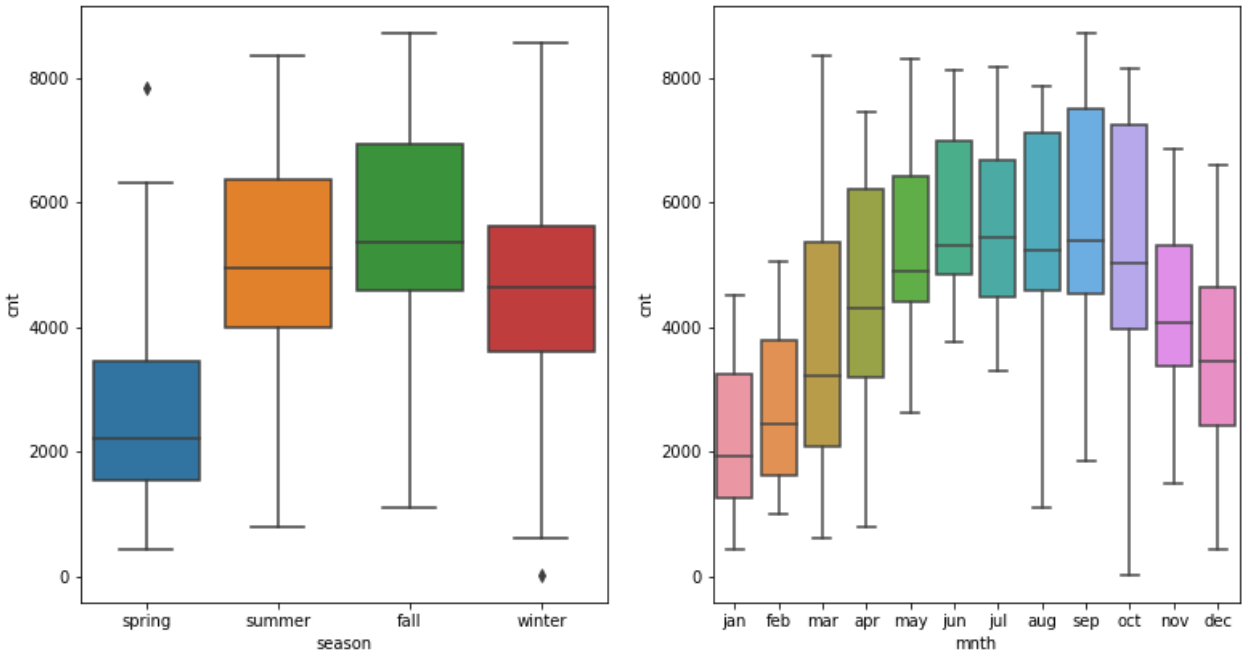
Assignment-based Subjective Questions

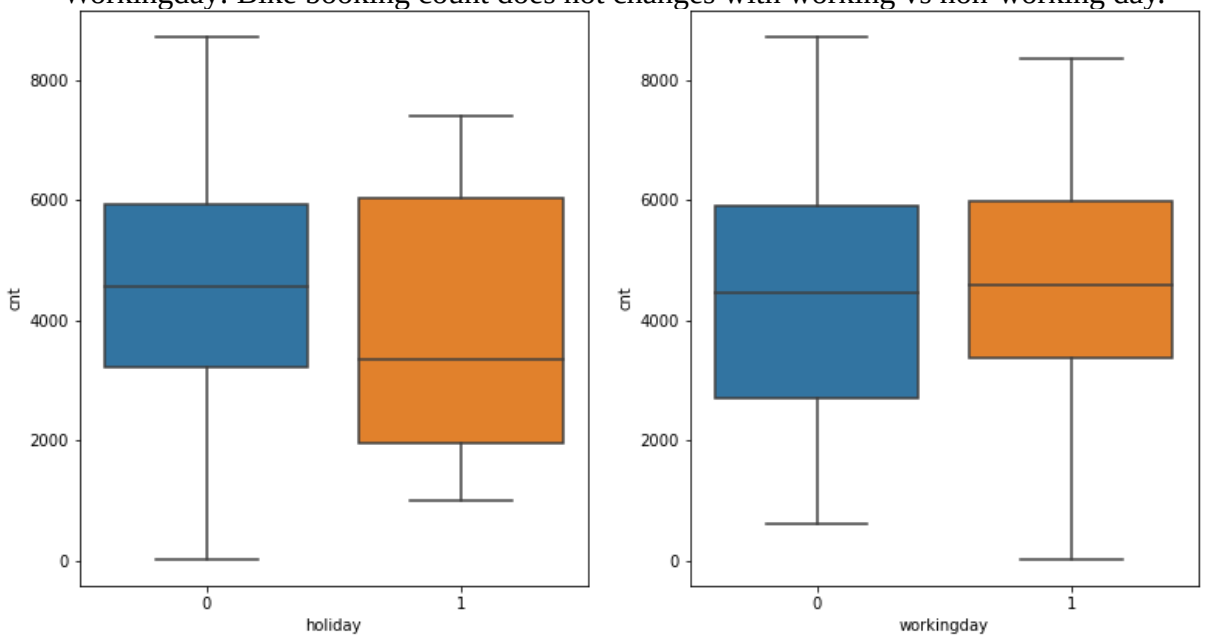
1. **From your analysis of the catagorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer: We are having the following catagorical features in the dataset:

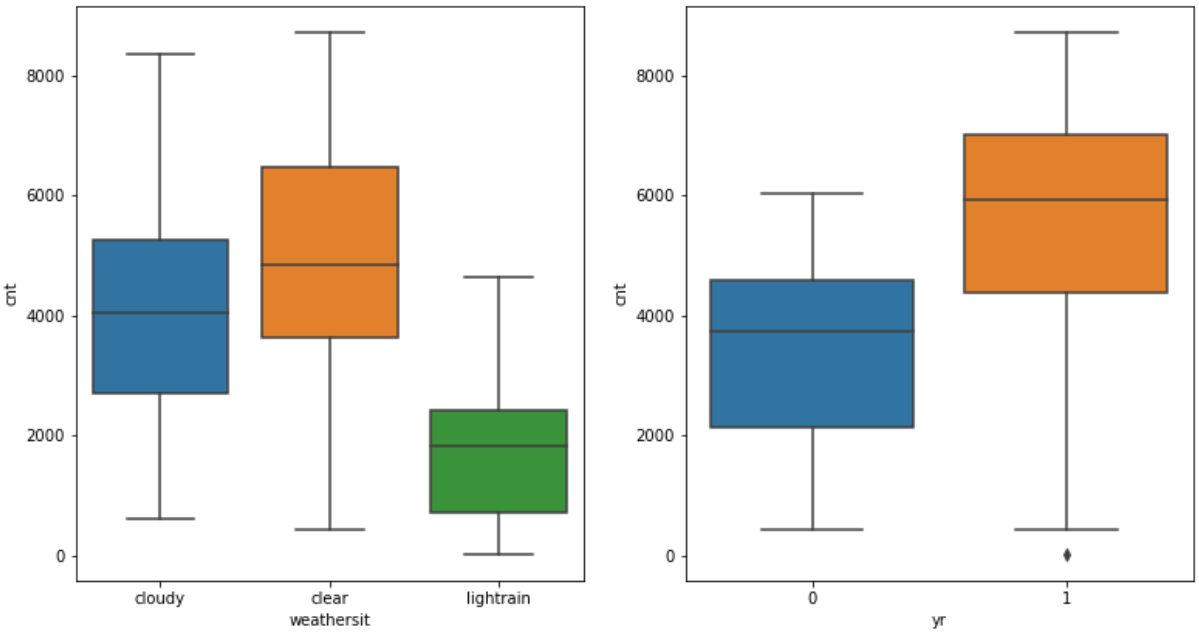
- Season: Counts of rented bikes depends on season, cnt increases in Fall and Summer season and it gets decreases with winter and spring
- mnth: Cnt varies with months as we are having higher booking with May, June, July , August, and September. And booking decreases with Winters month.



- Holiday: Cnt depends on holiday. On non-holiday we are having higher cnt value.
- Workingday: Bike booking count does not changes with working vs non-working day.



- Yr: Count of booking increases with year. We are having higher no of bike renting in 2019 year compared to 2018.



- Weathersit: Weathersit affects the bike booking. Clear weather attracts the customers and has higher booking than cloudy and then light rain.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
Answer: When we are converting the catagorical features to numeric feature. So that these can be scalarized. We break the single column into all the possible probable columns and fill the values between 0 and 1.
For example: We are having the following categorical feature:

Season
summer
winter
fall
spring

And it will be converted to following table:

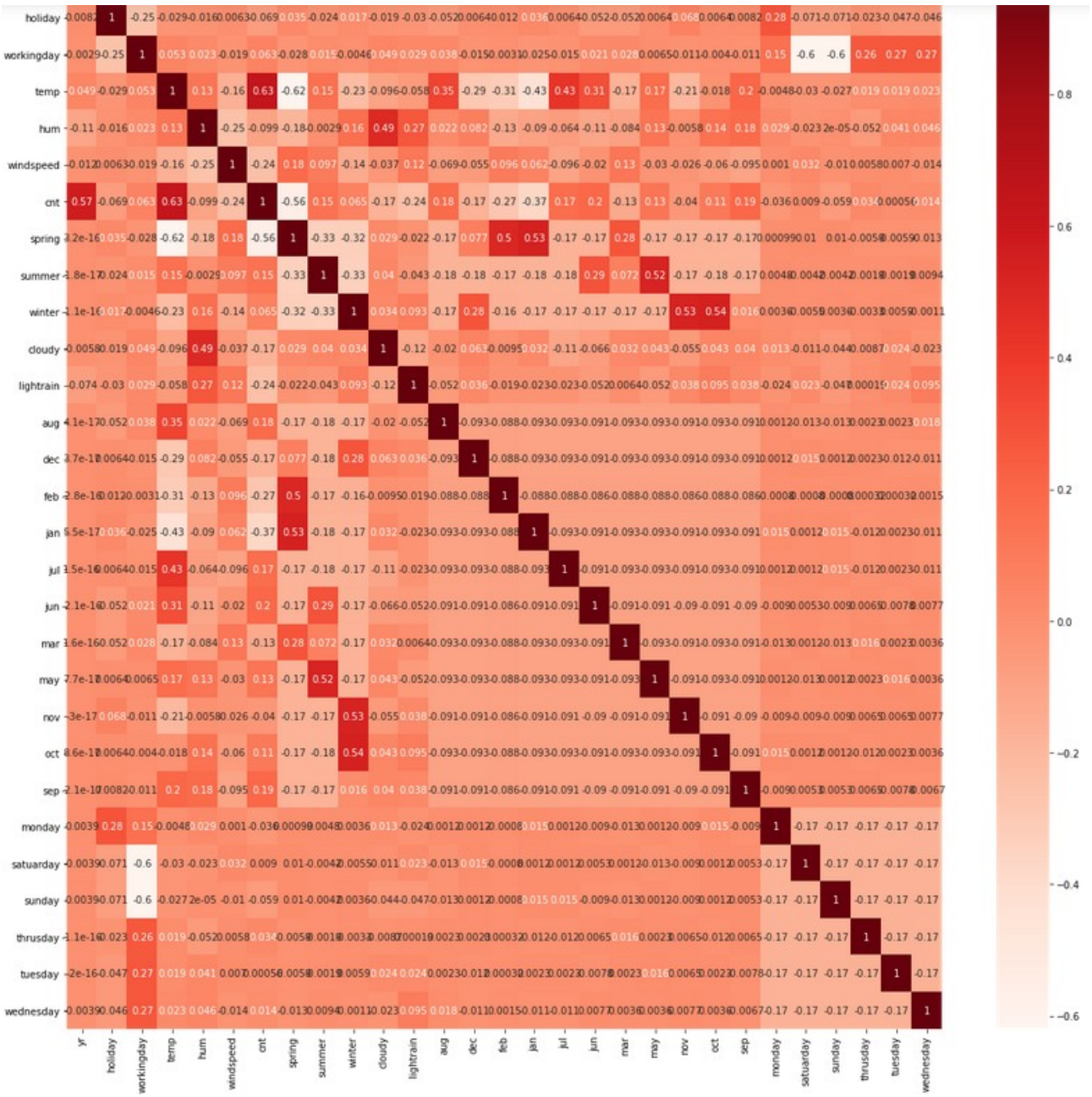
summer	Winter	Fall	spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

But we can extract the information with 1 less column like if all the values are zero then we can assume it would be the final option and we can have the following table:

Winter	Fall	spring
0	0	0
1	0	0
0	1	0
0	0	1

As we can see 4 colums are reduced to 3 columns.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
Answer: temp attribute is having the highest correlation with cnt (target variable) with 0.63 value.



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

Answer: We can validate the model on the following scenarios.

a) Low p-value --> p-value should be low (less then 0.05 is acceptable)

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	187.7
Date:	Wed, 22 Dec 2021	Prob (F-statistic):	4.31e-154
Time:	19:38:41	Log-Likelihood:	415.03
No. Observations:	510	AIC:	-810.1
Df Residuals:	500	BIC:	-767.7
Df Model:	9		
Covariance Type:	nonrobust		

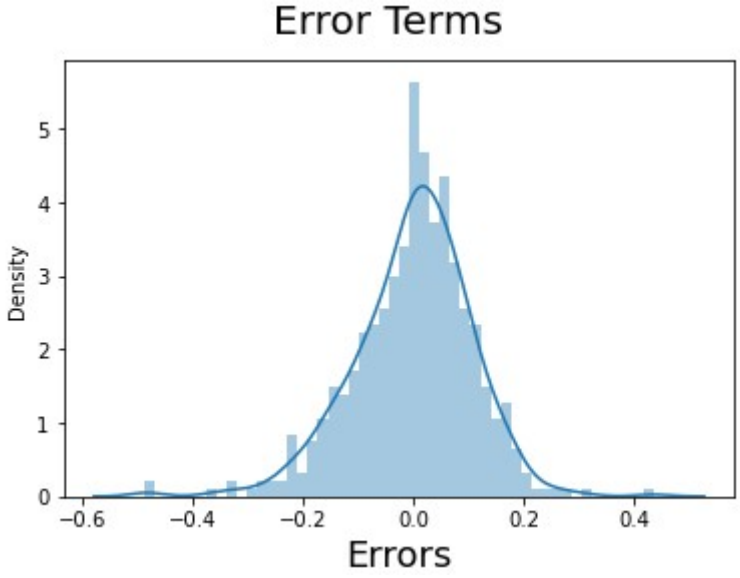
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5723	0.013	43.029	0.000	0.546	0.598
yr	0.2486	0.010	25.831	0.000	0.230	0.268
windspeed	-0.2032	0.029	-6.948	0.000	-0.261	-0.146
spring	-0.2368	0.014	-16.511	0.000	-0.265	-0.209
winter	-0.0546	0.012	-4.589	0.000	-0.078	-0.031
cloudy	-0.0901	0.010	-8.790	0.000	-0.110	-0.070
lightrain	-0.2990	0.029	-10.279	0.000	-0.356	-0.242
jan	-0.1040	0.020	-5.097	0.000	-0.144	-0.064
sep	0.0862	0.018	4.764	0.000	0.051	0.122
sunday	-0.0465	0.014	-3.383	0.001	-0.074	-0.019

Omnibus:	47.207	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	99.456
Skew:	-0.530	Prob(JB):	2.53e-22
Kurtosis:	4.886	Cond. No.	8.67

b) Variance inflation factor (VIF) - VIF should be low (lesser then 5 is acceptable).
as VIF = 1/(1-R-square)

	Features	VIF
1	windspeed	2.58
2	spring	2.12
0	yr	1.74
6	jan	1.59
4	cloudy	1.45
3	winter	1.37
8	sunday	1.15
7	sep	1.09
5	lightrain	1.08

c) Error Rate: Normalised Error rate with zero centralized.



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: Following features contribute the most:

- Light rain
- Yr
- Spring

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
Answer: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into catagories (e.g. cat, dog). There are two main types:

1. **Simple regression** : Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y = mx + b$$

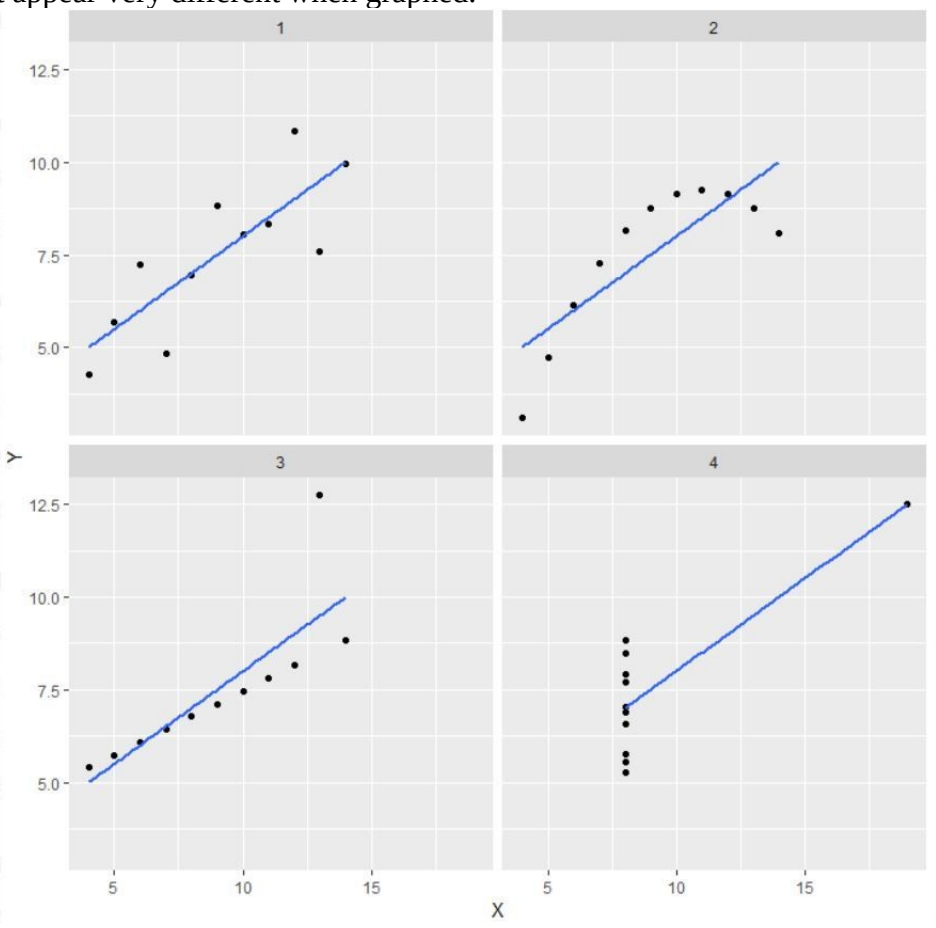
2. **Multivariable regression** : A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z) = w1x + w2y + w3z$$

For example: For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$Sales = w1Radio + w2TV + w3News$$

2. **Explain the Anscombe's quartet in detail. (3 marks)**
Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



These above datasets have same Mean(x), std(x), Mean(y), std(y) and Cor (x,y) but their prediction are completely different.
Conclusion: It is strongly recommended to look at data first then start performing linear regression or any other analysis.

3. What is Pearson's R? (3 marks)
Answer: This is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
Answer: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Normalized scaling: This is also called Min-Max scaling, This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardized saling: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
Answer: Formula for VIF is 1/(1-R-square). If VIF is infinite means denominator is zero, R-square is equal to 1 or close to 1. Whenever there is high correlation between independent variables then there 'RSquare' value will be high (approaching to 1).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
Answer: a **Q–Q (quantile-quantile) plot** is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested.