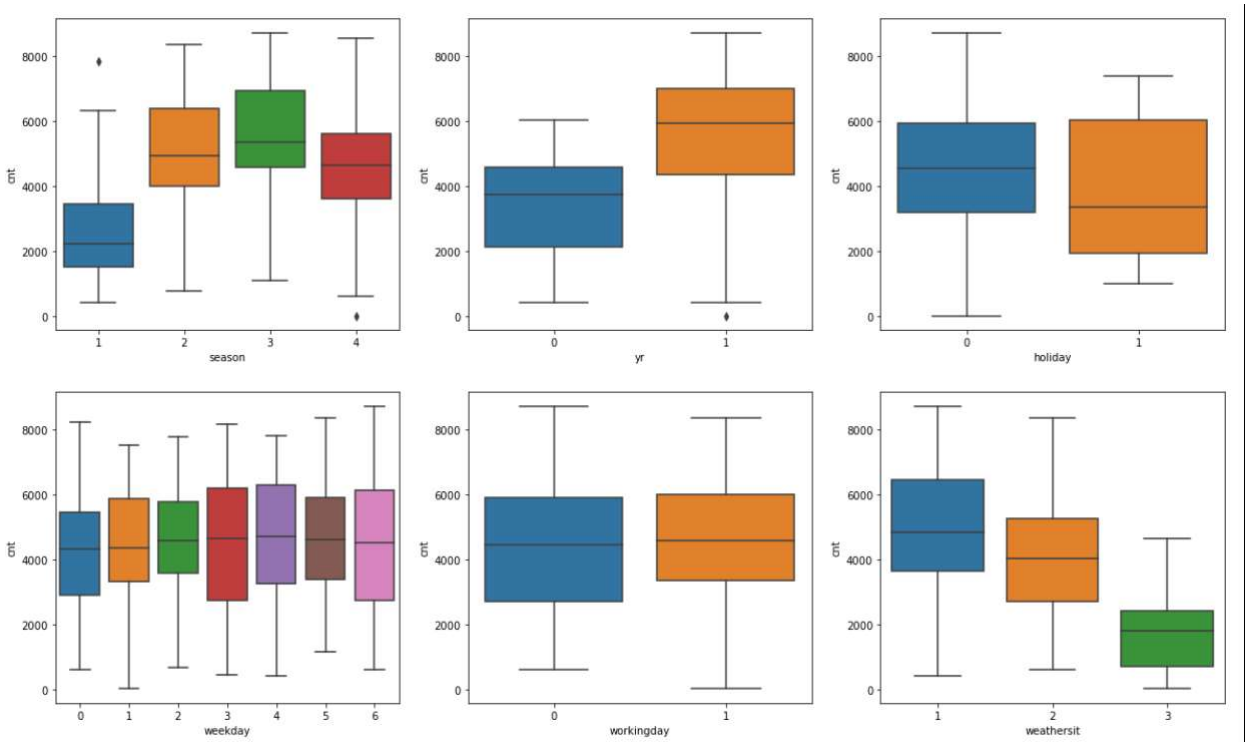
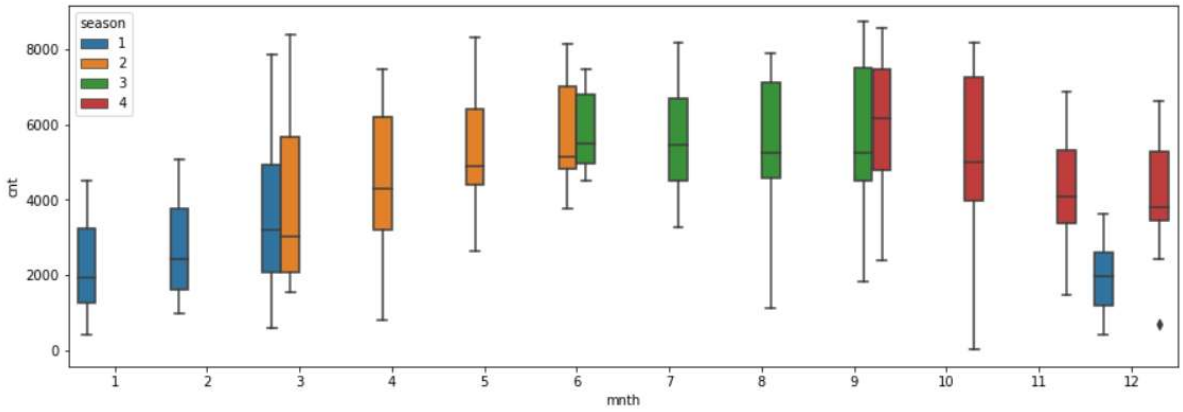
Assignment-based Subjective Questions

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



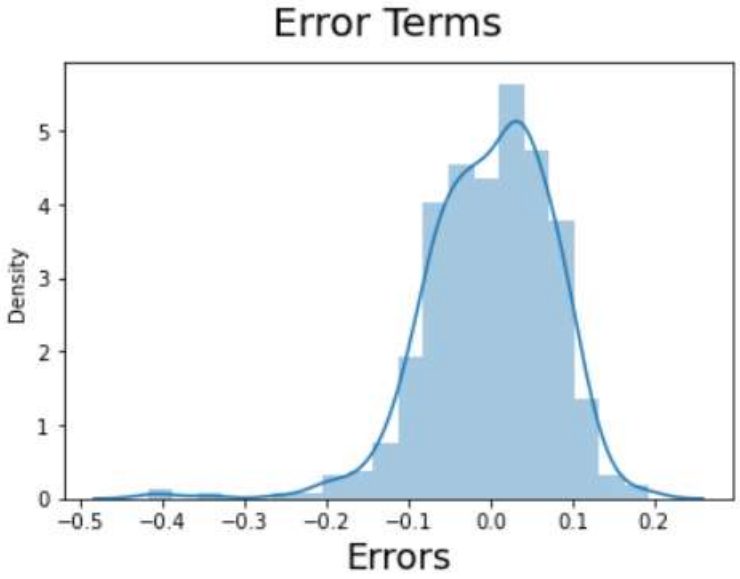


- Season - based on season there is trend effect on 'cnt' variable. Season 3 has got higher number of bookings
- Yr - 2019 has got highest booking. Increase in sales for YOY
- Holiday – do not have much of impact on 'cnt'
- Weekday - 4th day of the week is getting higher bookings
- Weathersit – Based on weather situation 'cnt' seems to be declining. i.e. 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- Mnth - 9th month has got the highest booking for 'cnt' during season 3 & 4 (fall & winter)

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

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Atemp – is highly correlated with 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   By using Residual analysis approach, we examine the error terms if it is normally distributed. In our case it is normally distributed.



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

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.884
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	422.5
Date:	Wed, 09 Feb 2022	Prob (F-statistic):	2.97e-227
Time:	19:39:29	Log-Likelihood:	587.34
No. Observations:	510	AIC:	-1155.
Df Residuals:	500	BIC:	-1112.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0467	0.016	2.891	0.004	0.015	0.078
season	0.1245	0.018	6.791	0.000	0.088	0.160
yr	0.1738	0.007	23.849	0.000	0.159	0.188
mnth	0.0207	0.021	0.984	0.325	-0.021	0.062
holiday	0.0216	0.023	0.948	0.343	-0.023	0.066
weekday	0.0183	0.010	1.771	0.077	-0.002	0.039
workingday	0.2141	0.009	23.044	0.000	0.196	0.232
weathersit	-0.1149	0.013	-8.778	0.000	-0.141	-0.089
windspeed	-0.0679	0.021	-3.207	0.001	-0.110	-0.026
casual	0.7684	0.023	33.717	0.000	0.724	0.813

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear regression is a supervised learning approach. It is used to predict independent variable which is continuous.
   - If there is a single input variable it is called simple linear regression.
   - If there are multiple input variables, then it is called as multiple linear regression

   There are certain assumptions about Linear regression

   - Dependent and independent variables will show a linear relationship
   - Error terms are normally distributed

   Positive Linear Relationship:

   - If the dependent variable expands on the Y-axis and the independent variable progress on X-axis

   Negative Linear relationship:

   - If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that tricks the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This educates us concerning the significance of imagining the information prior to applying different calculations out there to construct models out of them which proposes that the information highlights should be plotted to see the circulation of the examples that can assist you with recognizing the different abnormalities present in the information like anomalies, variety of the information, direct distinctness of the information, and so forth Additionally, the Linear Regression can be just be viewed as a fit for the information with straight connections and is unequipped for taking care of some other sort of datasets

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3	x4	y4
1	10	8.04		10	9.14		10	7.46	8	6.58
2	8	6.95		8	8.14		8	6.77	8	5.76
3	13	7.58		13	8.74		13	12.74	8	7.71
4	9	8.81		9	8.77		9	7.11	8	8.84
5	11	8.33		11	9.26		11	7.81	8	8.47
6	14	9.96		14	8.1		14	8.84	8	7.04
7	6	7.24		6	6.13		6	6.08	8	5.25
8	4	4.26		4	3.1		4	5.39	19	12.5
9	12	10.84		12	9.13		12	8.15	8	5.56
10	7	4.82		7	7.26		7	6.42	8	7.91
11	5	5.68		5	4.74		5	5.73	8	6.89
				Summary Statistics						
N	11	11		11	11		11	11	11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94	3.16	1.94
r	0.82			0.82			0.82		0.82	

3. What is Pearson's R? (3 marks)
Pearson's R is the correlation coefficient which explains the correlation between continuous dependent variables & independent variables. It is based on the method of covariance.
Highly correlated variables have to be dropped to get the better results.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Scaling is required to bring all the variable sin to same level of magnitude.
Most of the times when we use data for modelling will have different variables with different unit & ranges to bring all these on to same scale we use scaling.
    - Normalized scaling – brings all of the data in the range of 0 & 1

- Standardized Scaling - replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   If there is a perfect correlation between the variables, then VIF is infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Q-Q plot is the probability plot. Q-Q plots are used to find the type of distribution for a random variable.
   With regards to linear regression, Q-Q plot will help us to check if there is a normal distribution of error terms & a linear relationship between X&Y variable.