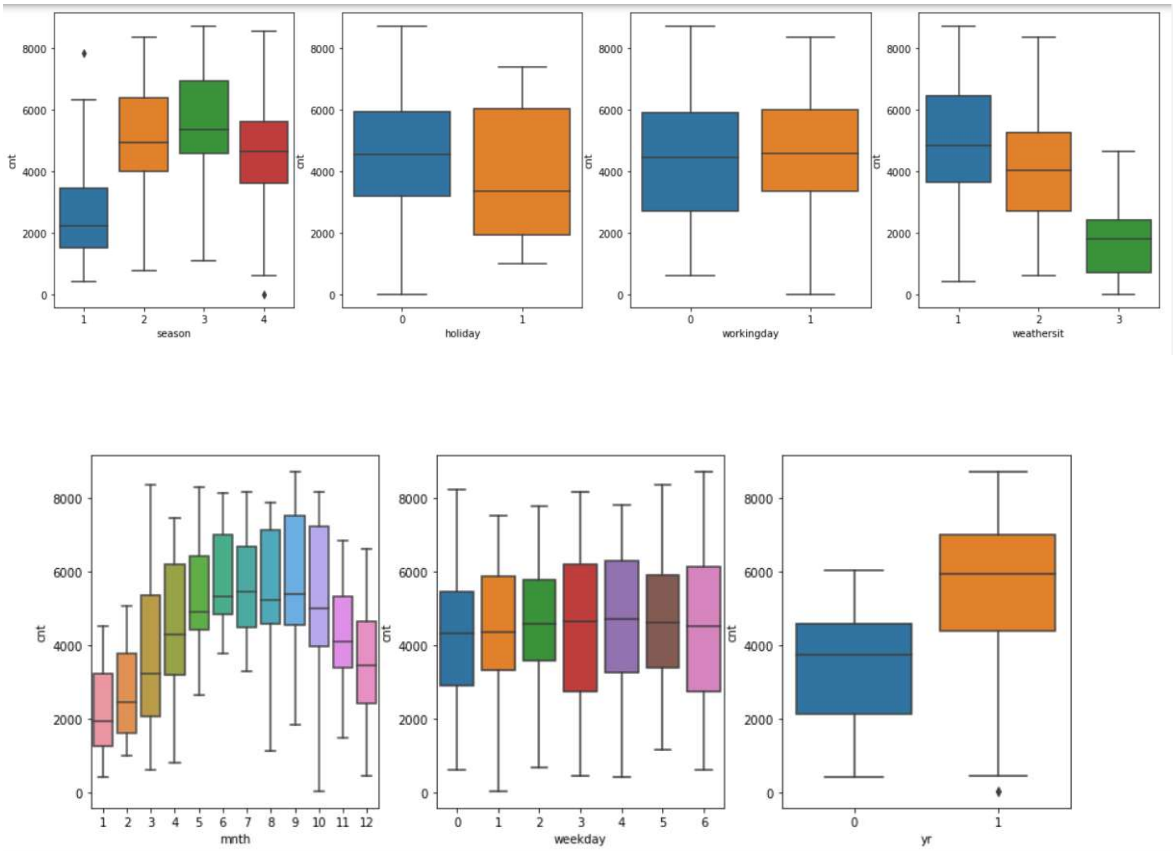
Assignment-based Subjective Questions

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



From the plots above I can say that:-

- Season has impact on count of bike rentals which is our target variable.
- Holiday has impact on target variable.
- Working day and holiday gives same observation.
- Weather has impact on target variable
- High number of bike rentals when month is September.
- From weekday I can't say anything clear for now.
- Bike rentals increases with year.

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

This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). Dummy encoding uses N-1 features to represent N labels/categories.

Dummy variable creation leads to a drawback, where features are highly correlated. That means using the other variables, we can easily predict the value of a variable. This is called Dummy variable trap. So to deal with trap we use the drop_first=True.

For example:- see below in image.

Column	Code
A	10
B	01
C	00

Dummy Code

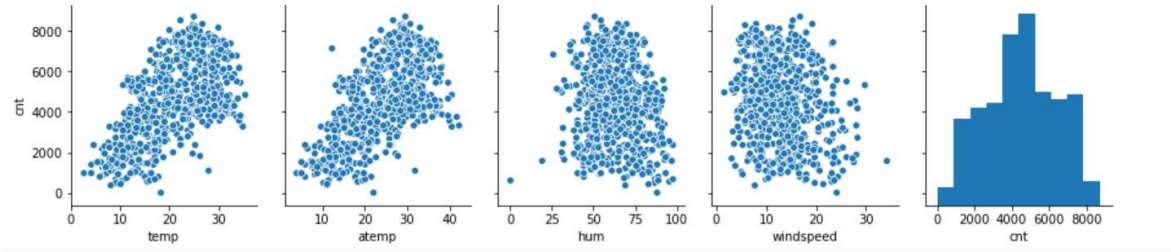
if there are A,B and C are categorical variables .When we convert these into dummies A ,B and C will be the extra columns created. If we see A has 10, B has 01 and C has 00, so if drop first value which is A, then we still can say that A has 10.

B	01
C	00

Dummy Code

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

Temp has the highest correlation with target variable.
(temp and atemp are collinear so considering only one temp.)



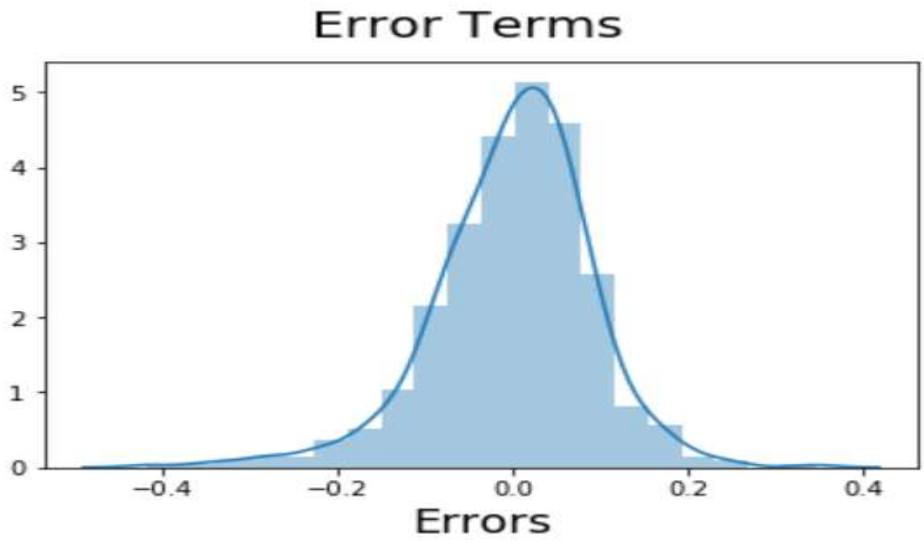
---

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   I draw the below plots to check the assumptions made by Linear Regression.

   1. Error terms are normally distributed (not X, Y)
      To Check this assumptions I draw the below histogram for residuals.

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

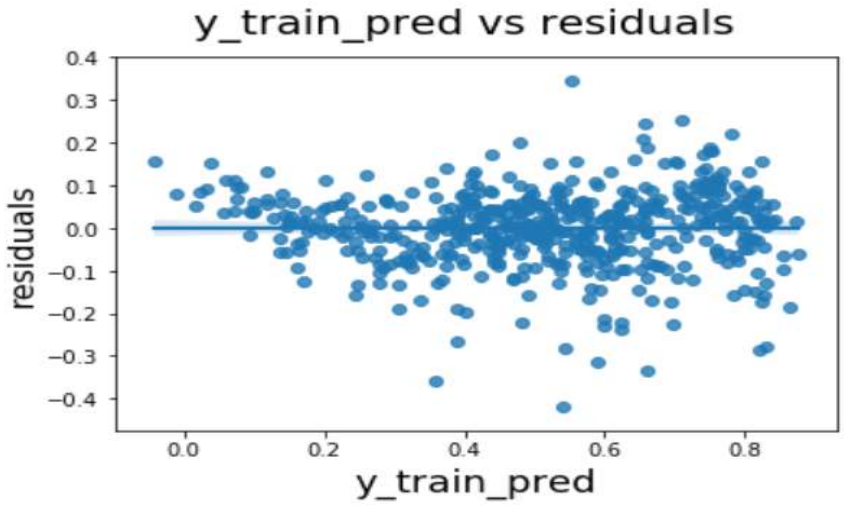
```
Text(0.5, 0, 'Errors')
```



Observations:-
Error terms are normally distributed

2. Error terms have constant variance (homoscedasticity) and independent of each other.
   To Check this assumptions I draw the below scatter plot in between residuals and y_train_pred.

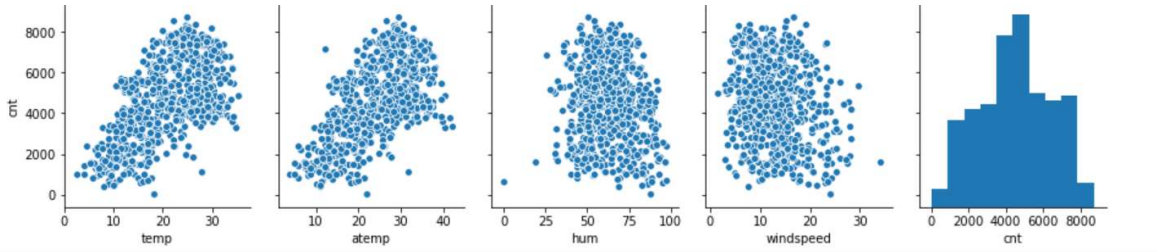
```
# plotting residuals and y_train_pred to understand the spread.
fig = plt.figure()
sns.regplot(y_train_pred,res)
fig.suptitle('y_train_pred vs residuals', fontsize = 20)
plt.xlabel('y_train_pred', fontsize = 18)
plt.ylabel('residuals', fontsize = 16)
plt.show()
```



Observation:-

- Error Terms do not show any pattern.
- Variance is constant.

3. Linear relationship between X and Y
   To Check this assumption I draw the scatter plot between variables and I can see that there is linear relationship between X and y.



Observations:-

- There is linear relationship between X and y.

After observing the above three plots, I concluded that

- Error Terms do not show any pattern.
- Variance is constant.
- X and y has the liner relationship.

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

   After observing the model summary below, I concluded that:-

   Temp feature has a coefficient of 0.4721 .
   yr feature has a coefficient of 0.2343 .
   LightSnow which is a weather feature category has a coefficient of -0.2854 (absolute value is 0.2854).

   Therefore I can conclude below points.

   - **Temperature (temp), year(yr) and weather(weathersit:- LightSnow)** features has the high impact on bikes rental count.
   - **When the weather is clear** we are seeing lot of bike rentals.
   - **Bike rentals increases as the temperature increases**.
   - **Bike rentals increases with the year.**

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.837
Model:	OLS	Adj. R-squared:	0.833
Method:	Least Squares	F-statistic:	212.3
Date:	Sat, 03 Oct 2020	Prob (F-statistic):	8.14e-187
Time:	16:30:04	Log-Likelihood:	501.13
No. Observations:	510	AIC:	-976.3
Df Residuals:	497	BIC:	-921.2
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2154	0.030	7.088	0.000	0.156	0.275
yr	0.2343	0.008	28.517	0.000	0.218	0.250
holiday	-0.0968	0.026	-3.722	0.000	-0.148	-0.046
temp	0.4721	0.034	13.772	0.000	0.405	0.539
windspeed	-0.1549	0.025	-6.135	0.000	-0.205	-0.105
spring	-0.0617	0.021	-2.905	0.004	-0.103	-0.020
summer	0.0434	0.015	2.845	0.005	0.013	0.073
winter	0.0757	0.017	4.335	0.000	0.041	0.110
Jan	-0.0383	0.018	-2.138	0.033	-0.073	-0.003
Jul	-0.0503	0.019	-2.706	0.007	-0.087	-0.014
Sep	0.0764	0.017	4.506	0.000	0.043	0.110
Cloudy	-0.0793	0.009	-9.067	0.000	-0.096	-0.062
LightSnow	-0.2854	0.025	-11.575	0.000	-0.334	-0.237

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression

This is a machine learning algorithm based on supervised learning. This works on finding the best linear relationship between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method. The objective is to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible. Error is the distance between the points to the regression line.

Mathematically the relationship can be expressed by below formula:

$$Y = mX + b$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

b is a constant, known as the YY-intercept. If X = 0,Y would be equal to bb.

Type of Linear Regression

1. Simple Linear Regression
2. Multiple Linear Regression

Assumptions made by Linear Regression model

1. Error terms are normally distributed (not X, Y)
2. Error terms are independent of each other
3. Error terms have constant variance (homoscedasticity)

---

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

Anscombe's quartet

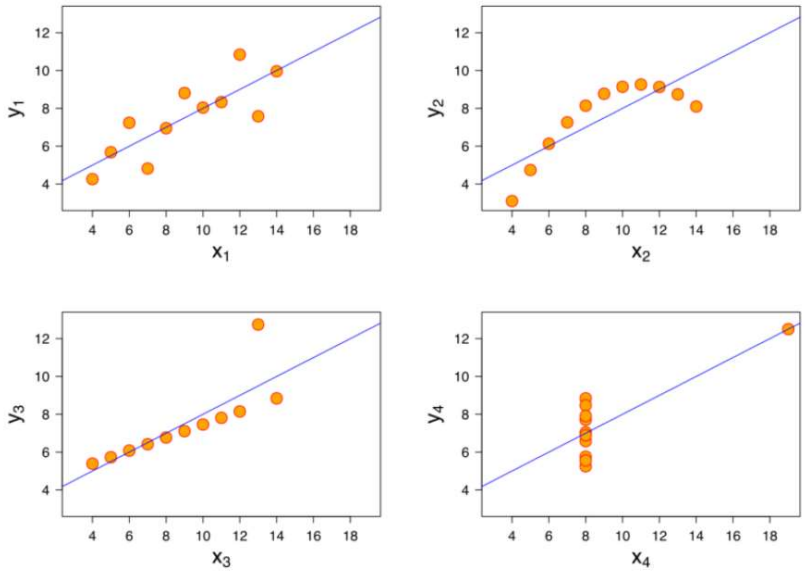
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

1. Mean of x is 9 and mean of y is 7.50 for each dataset.
2. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
3. The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

Statistics wise these looks same, but when we plot it, it changes completely.



1. Dataset I appears to have clean and well-fitting linear models.
2. Dataset II is not distributed normally.

3. In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
4. Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

---

3. What is Pearson's R?

Pearson's R

This is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1.

Important points about R:-

1. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

2. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

3. r = 0 means there is no linear association

4. r > 0 < 5 means there is a weak association

5. r > 5 < 8 means there is a moderate association

6. r > 8 means there is a strong association

---

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

What is Scaling and why

Feature Scaling is a technique to formalize 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. Also, the model convergence takes more time.

Normalized scaling and standardized scaling

- Normalized scaling rescales the value into a range of [0,1]. This is a good technique to use when the distribution of the data is unknown or when the distribution is not Gaussian.
- Standardized scaling rescales data to have a mean of 0 and a standard deviation of 1. This scaling assumes that the data has a Gaussian distribution, however, this does not have to be true but the technique is more effective if the attribute distribution is Gaussian.

Formula of Min-Max Normalization

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

Formula of Standardization

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

---

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

VIF

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model.

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

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. Therefore, To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Hence we can say that, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

---

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

Q-Q plot and Importance

Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. It is used to compute the theoretically expected value for every data point based on the distribution in question.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

In Linear Regression, Q-Q plot is used to assess if your residuals are normally distributed. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed then they should form an approximately straight line.

Image below are showing how the Q-Q plot changes with data distributions.