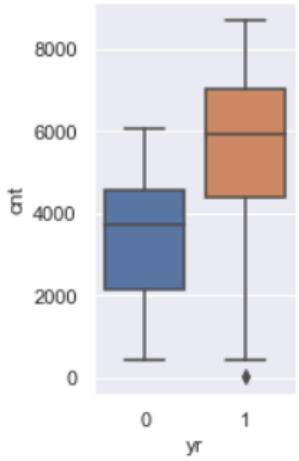
Assignment-based Subjective Questions

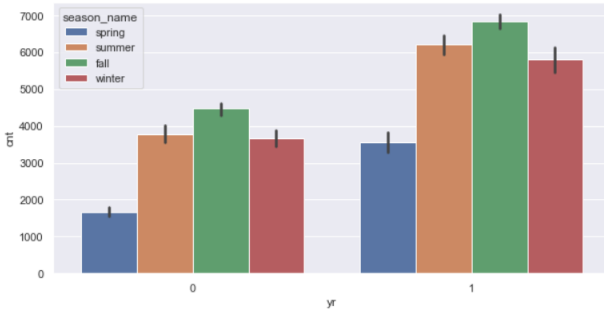
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

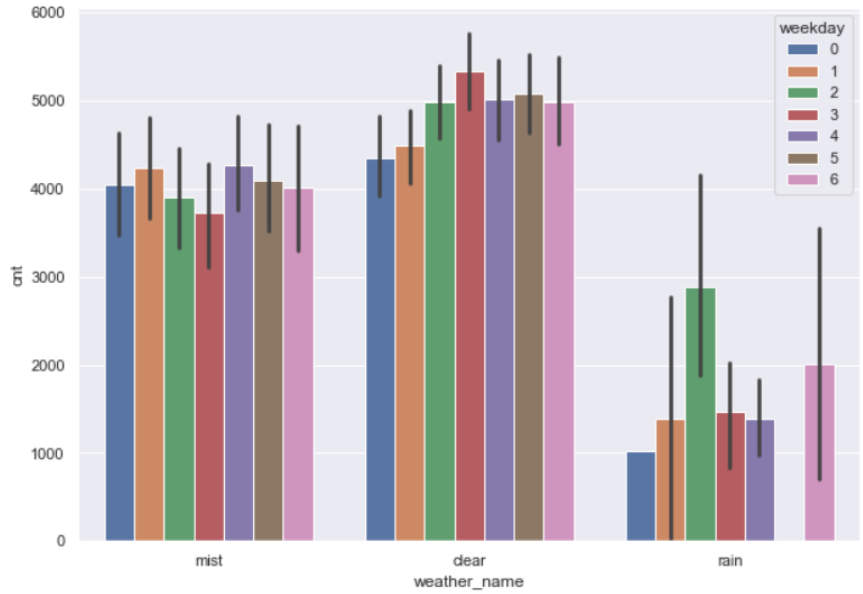
Year – There is a steady increase in the User base/Usage of the bikes by both the casual and registered users year on year.

Season – Except for Spring rest of the 3 seasons show a better usage of the bikes

Weather – When the weather is Snow or heavy rain the usage of bikes is greatly affected.





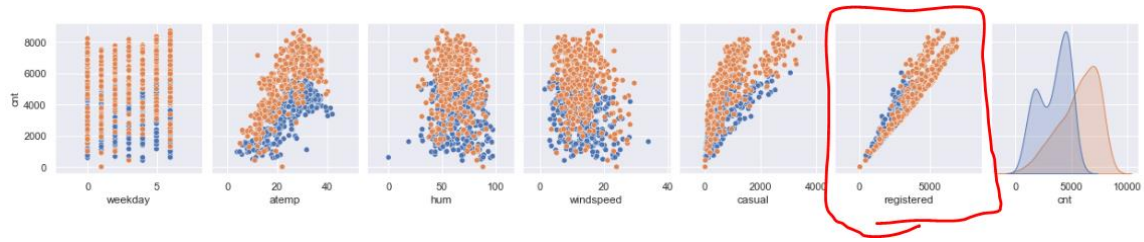


2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

The aim is to reduce the number of variables to the minimum required to effectively build the prediction model. When there are N values for a variable, we are able to create N-1 dummy variables and represent the same data. So we are dropping one of the dummy variable to reduce the cost of the prediction model.

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

Registered Bike Users Count - Variable



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

There is a **linear relationship between Independent variables and Target variable**. When Temp, Weather, Wind speed are favourable ie. Not too cold, Not snowing, Not too windy the Bike User Count increases.

The **error is distributed normally with mean ZERO**

Error has constant variance. **Homoscedastic**

No pattern in error. The error values are independent.

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

Temperature, Year, Casual/Registered User Count Ratio

	coef	std err	t	P>|t|
const	0.1236	0.029	4.210	0.000
yr	0.2443	0.008	29.527	0.000
workingday	0.1328	0.016	8.137	0.000
atemp	0.3781	0.031	12.201	0.000
windspeed	-0.1212	0.025	-4.857	0.000
mist	-0.0707	0.009	-8.099	0.000
rain	-0.2579	0.025	-10.378	0.000
spring	-0.1105	0.015	-7.373	0.000
winter	0.0533	0.012	4.344	0.000
sat	0.0668	0.014	4.671	0.000
_CR_ratio	0.2601	0.040	6.476	0.000

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is an algorithm for **Supervised** learning in which we build a model by training on a representative dataset and then use that model as a **forecast** mechanism to predict (the dependent variable) on future datasets.

The model tries to establish the **linear relationship** between the independent variables X and the dependant target variable Y.

Y = m1X1 + m2X2 + m3X3 + … + C

Where C is the **intercept** (value of Y when X =0) and X1,X2,X3 etc are the **independent** variables and m1,m2,m3, etc are the coefficients of X1,X2,X3,… etc respectively. The **coefficients** will describe how the value of Y will change for **EACH unit** of the independent variables X1,X2,X3,….

The aim of the model is to **draw a line which best fits the regression** for a given set of X values.

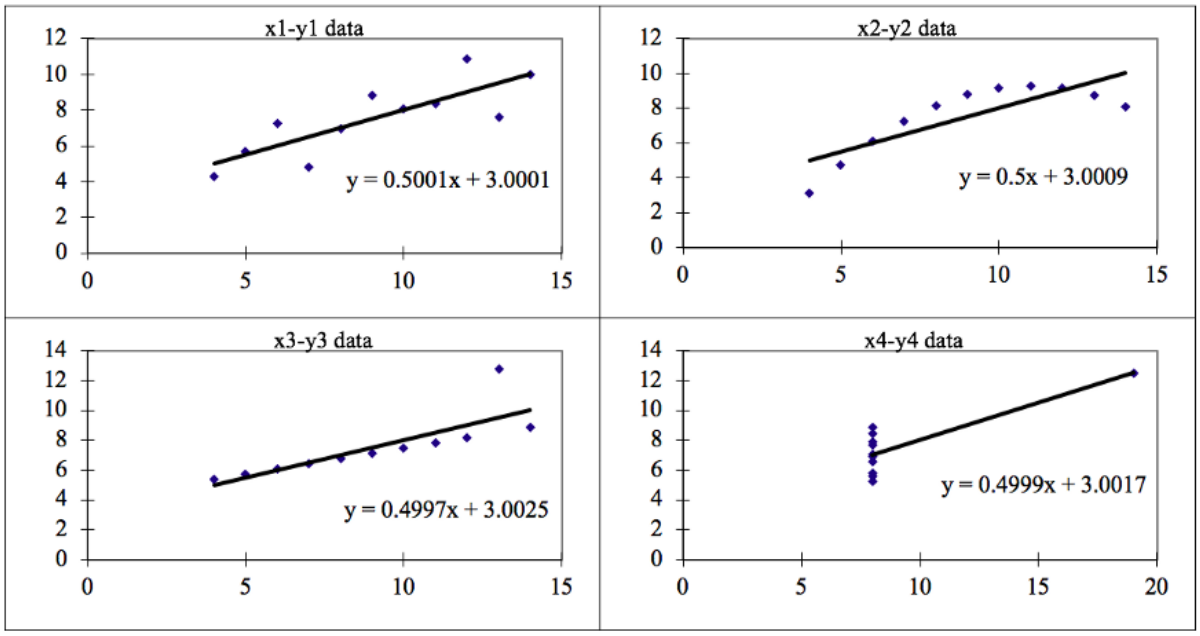
It should also hold true that the **ERROR** i.e. the difference between the True Value and Predicted value is **minimum**. This is MSE or **RMSE** Root Mean Squared Error.

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

Anscombe's Quartet is a group of four data sets which are **nearly identical with basic statistics like mean, standard deviation, R2 being identical**, but there are some peculiarities in the dataset that **fools the regression model** when built.
They have very different distributions and **appear differently** when plotted on scatter plots.

For example consider these 4 datasets (x,y)

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	



The four datasets can be described as:

1. **Dataset 1: fits** the linear regression model well.
2. **Dataset 2: could not fit** linear regression model as the data is non-linear.
3. **Dataset 3:** the **outlier** involved in the dataset **is not handled** by linear regression model
4. **Dataset 4:** the **outliers** involved in the dataset **cannot be handled** by linear regression model

3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient, R, is a value between -1 and 1. Further the value of R is from zero, stronger the linear relationship between those two variables.

- If R is positive, then when one independent variable increases, the dependant variable tends to increase.
- If R is negative, then when one independent variable increases, the dependant variable tends to decrease.
- When R=-1 or R=1 it means that the 2 variables are exactly linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature scaling is a data pre-processing step used **to normalize the range of independent variables** in the dataset.

For example — Assume we have a dataset with the following 3 independent variables

- Age : 18–100 Years
- Salary : 1 Lakh–5 Lakhs INR
- Height : 150 – 190 CMS

After applying Feature scaling technique the ranges of the above 3 variables would all be in the same range, for example- centred around 0 or in a much smaller range like (0,1), (-1,1) depending on the scaling technique used.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1].

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. Also called as Z-score.

X_new = (X - mean)/Std Deviation

Normalization	Standardization
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
Used when features are of different scales.	Used when we want to ensure zero mean and unit standard deviation.
Scale ranges between [0, 1] or [-1, 1].	Not bounded to a certain range.
Affected by Outliers.	Less Affected by Outliers.
Used when we don't know the distribution	Used when the feature distribution is Normal or Gaussian.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF = 1 / (1 -R^2)

where, R-squared (R^2) is the coefficient of determination in linear regression. Its value lies between 0 and 1.
If all the independent variables are orthogonal (right angle) to each other, then VIF = 1.0.
In the case of **perfect correlation, we get R2 =1**, then VIF = infinity. To solve this problem we need to drop one of the independent 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. (3 marks)

Q-Q Plots (**Quantile-Quantile plots**) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

QQ plots are used to assess the similarity of the distributions of two datasets. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

- If the data is **normally distributed,** the points will fall on the 45-degree reference line.
- If the data is not normally distributed, the points will **deviate from the reference line.**

Points on the Q-Q plot provide an indication of univariate normality of the dataset.