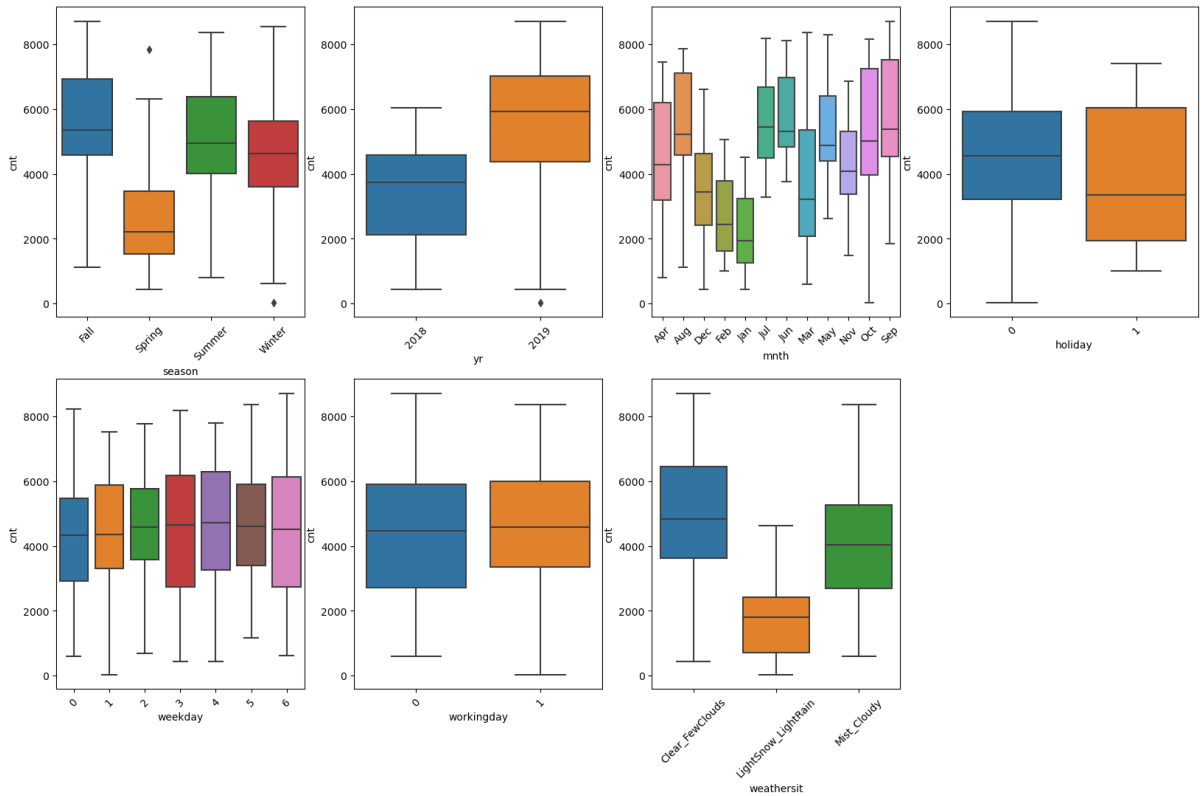
Name: Sandeep Pathak

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)**

ANSWER: Based on the boxplots below for the categorical variables



a. **Season**: Bike demand is highest in the fall followed by summer, and lowest in the spring. In winter bike demand is slightly lower than that of summer
b. **Yr** (year): Bike demand is considerably more in the year 2019 as compared to 2018
c. **Mnth** (month): Bike demand is considerably more in the months of Aug, Sep and Oct, as compared to lowest in Jan, Feb and Dec
d. **Holiday**: Maximum demand for bikes is similar whether it's a holiday or not, but minimum demand is much higher when its not a holiday
e. **Weekday**: There is more demand for bikes on weekdays 3,4 and 6
f. **Workingday**: Demands for bike are comparable whether it's a working day or not
g. **Weathersit** (weather situation): As expected, demand for bikes is much higher when weather is clear/few clouds as compared to when its Light snow/Light rain

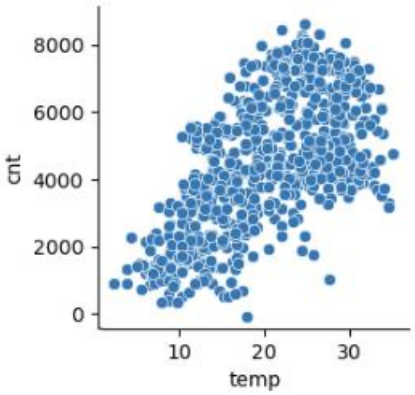
2. **Why is it important to use *drop_first=True* during dummy variable creation? (2 mark)**

ANSWER: It helps removing the extra column, thereby reducing multicollinearity amongst the dummy variables. If this is not done and all the *n* dummy variables instead of *n -1* (*after*

drop_first=True) are used, then these dummy variables will themselves be highly correlated, causing multicollinearity. The predictions for such a model will be thus incorrect due to high multicollinearity.

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

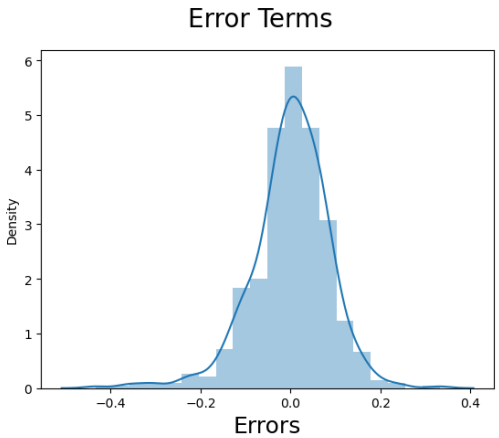
ANSWER: *temp* (independent variable) is highly correlated with with the *cnt* target/dependent variable, with a correlation of 0.64



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

ANSWER: The basic assumptions of Linear Regression are:

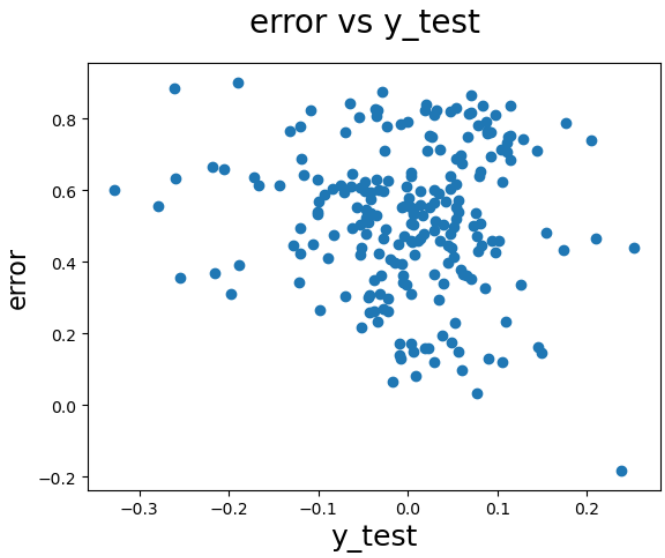
1. Error terms are normally distributed with mean zero



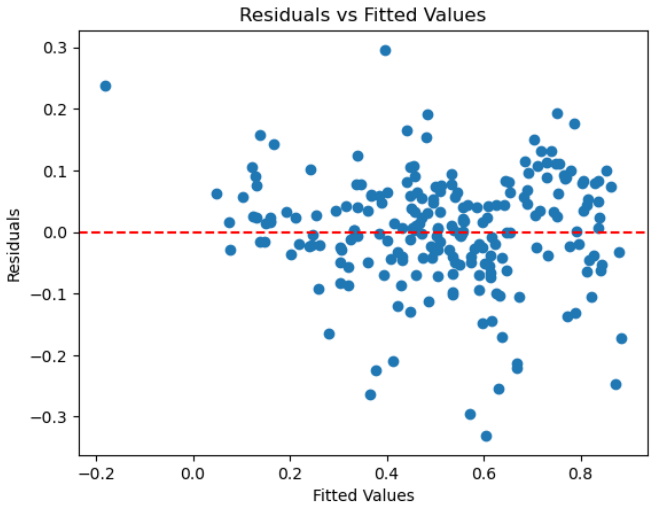
2. There is a linear relationship between input features X and output y – this is verified by final output linear model, showing linear relationship –
*y = 0.2694 + (0.4729 * temp) − (0.1459 * hum) − (0.1880 * windspeed) − (0.0628 * season_Spring) + (0.0406 * season_Summer) + (0.1052 * season_Winter) + (0.2310 * yr_2019) − (0.0415 * mnth_Dec) − (0.0468 * mnth_Jan) − (0.0516 * mnth_Jul) − (0.0462 * mnth_Nov) + (0.0708 * mnth_Sep) + (0.0627 * weekday_6) + (0.0522 * workingday_1) − (0.0415 * mnth_Dec) − (0.0468 * mnth_Jan) − (0.0516 * mnth_Jul) − (0.0462 * mnth_Nov) +*

*(0.0708 * mnth_Sep) + (0.0627 * weekday_6) + (0.0522 * workingday_1) − (0.2568 * weathersit_LightSnow_LightRain) − (0.0596 * weathersit_Mist_Cloudy) − (0.2568 * weathersit_LightSnow_LightRain) − (0.0596 * weathersit_Mist_Cloudy)*

3. Error terms (residuals) are independent of each other



4. Error terms have constant variance (homoscedasticity) - The variance should not increase (or decrease) as the error values change. Also, the variance should not follow any pattern as the error terms change.



5. Independent variables should not be highly correlated (no Multicollinearity) – We check the VIF (Variance Inflation Factor) score, which should be < 5 for all variables or close to 5

	Features	VIF
0	const	89.12
4	season_Spring	5.26
1	temp	4.42
6	season_Winter	3.83
5	season_Summer	2.76
2	hum	1.94
11	mnth_Nov	1.73
9	mnth_Jan	1.68
14	workingday_1	1.66
13	weekday_6	1.65
16	weathersit_Mist_Cloudy	1.58
8	mnth_Dec	1.50
10	mnth_Jul	1.49
12	mnth_Sep	1.33
15	weathersit_LightSnow_LightRain	1.27
3	windspeed	1.22
7	yr_2019	1.04

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: This is observed by seeing the coefficients of the features selected by the model. Higher the coefficients value (+ve or -ve, i.e absolute value), more it affects the final predicted output for any change in the features value. Thus, as per the below, top 3 coefficients are:

1. ***temp*** : Temperature – Coefficient = 0.4729
2. ***weathersit_LightSnow_LightRain***: Weather situation – Light Snow / Light Rain – Coefficient = -0.2568
3. ***yr_2019***: Year 2019 - Coefficient = 0.2310

General Subjective Questions

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

ANSWER: Linear regression is a fundamental machine learning algorithm that uses a mathematical equation/model to calculate the value of an unknown variable using known variable(s). For example, if you have data on the prices of a specific agricultural commodity, like a tomato crop and the amount of rainfall in the region over the past year, the soil composition, temperature and the season, Linear regression can analyze this data and determine that the commodity price increases by a certain amount with each additional inch of rainfall or affected accordingly by changes in soil nutrient value and temperature or by season of the year. It can then predict future commodity prices based on expected inputs of rainfall, soil composition, season and temperature.

In the above example, one may find via Linear Regression that the output (dependent) variable commodity (crop) price may or may not be dependent on the various features (independent)

– rainfall, temperature, soil composition. Please note that the independent features can be either numerical (eg temperature) or categorical (eg season – winter, summer etc)

So, essentially, Linear regression uses **historical data** for a continuous output variable to predict future output values. It is one of the simplest data modelling techniques that should be utilized first before moving to higher level (non-linear) models

While using Linear Regression, the basic mathematical concept is of *minimizing* the predicted error, which is –

a. Split the historical data into training and test set, ideally 80-20 or 70-30 composition.
b. Create the model using training data set.
c. Use the created model to predict the value of the inputs in the test dataset
d. Minimize the error of the above with respect to the actual value in the test dataset

While doing this, one needs to convert the categorical variables into their dummy equivalent so that they can be used in a mathematical formulae of linear equations, and remove the first dummy variable to avoid multicollinearity.

For the earlier example of commodity pricing of the agricultural crop, the model should result in a linear relationship like –

*Price = Constant + w1*rainfall + w2*temperature + w3*season_1 + w4*season_2 + w5*soil*

Above is a linear relationship of input independent variables with output *Price* variable. Once the model is built, i.e, the weights (*w1, w2..)* are determined, one can rule out variables that introduce high multicollinearity, thus finetuning the model till only the required features are left, which are good enough to predict the *Price*

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

ANSWER: Anscombe's quartet points to the fact that before analysing the data, one should first plot/visualize them. This is because, it can happen that the dataset has same summary statistics, so one might think that they are exactly the same or have some underlying relationship, but when visualized, they might be completely different

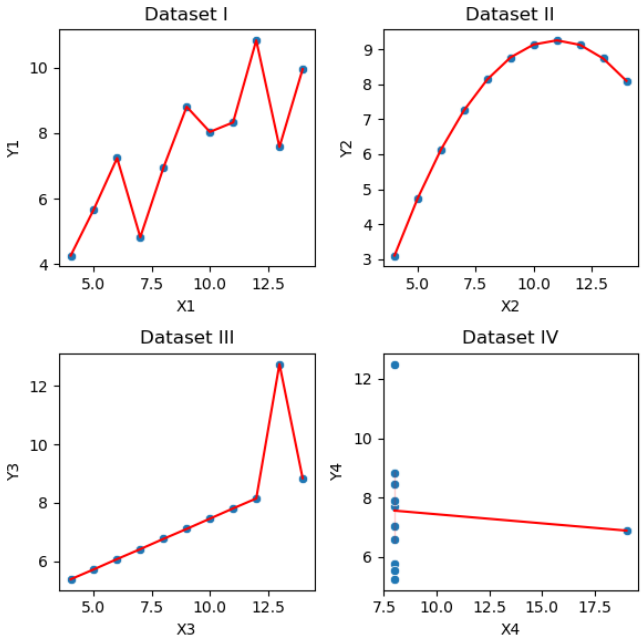
Consider example below – a dataset with for sets of (X1, Y1), (X2, Y2)...

	X1	Y1	X2	Y2	X3	Y3	X4	Y4
0	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
1	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
2	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
3	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
4	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47

When we see the summary of above, notice that the mean, standard deviation etc are same for Xs and Ys (using *pandas describe* function)

	X1	Y1	X2	Y2	X3	Y3	X4	Y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	9.000000	7.500909	9.000000	7.500000	9.000000	7.500909
std	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
min	4.000000	4.260000	4.000000	3.100000	4.000000	5.390000	8.000000	5.250000
25%	6.500000	6.315000	6.500000	6.695000	6.500000	6.250000	8.000000	6.170000
50%	9.000000	7.580000	9.000000	8.140000	9.000000	7.110000	8.000000	7.040000
75%	11.500000	8.570000	11.500000	8.950000	11.500000	7.980000	8.000000	8.190000
max	14.000000	10.840000	14.000000	9.260000	14.000000	12.740000	19.000000	12.500000

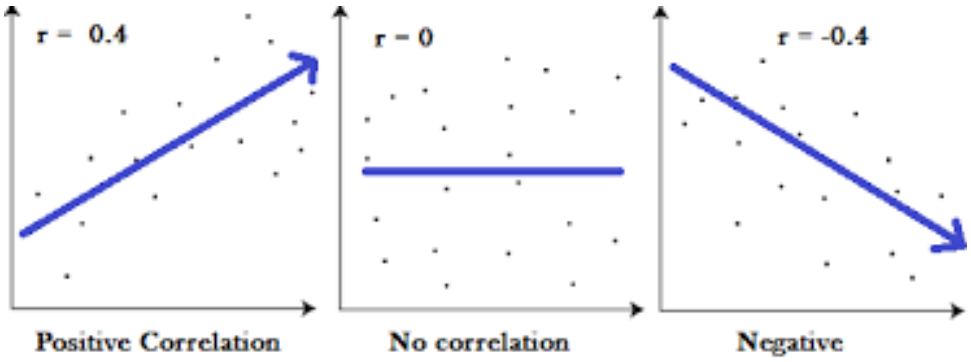
However, when visualized, they are completely different –



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

ANSWER: Pearson's correlation coefficient (R) tells you how strongly two variables are related and whether they increase or decrease together. It ranges from -1 to 1.

- **+1**: Perfect positive correlation – as one variable increases, other also increases in perfectly linear way. Eg, height and weight often have a positive correlation - people tend to weigh more as their height grows
- **-1:** Perfect negative correlation - as one variable increases, the other decreases in a perfectly linear way. Eg, the amount of exercise and body fat percentage might have a negative correlation - the more one exercises, the lower their body fat tends to be
- **0:** No correlation - there is no linear relationship between the variables. Eg, body fat and body strength have no correlation.



It is, thus, used to provide a simple numerical way to understand how two variables are related, thus predicting trends on how one variable might change as another one changes.

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

ANSWER: Scaling is adjusting the sizes/values, to make sure all numerical data (features) that are used in a Linear Regression model are on a similar scale/level, usually between 0 and 1 (normalized). When scaling is done on features, it makes sure they're all on the same level, so they work together smoothly. Scaling is done only for numerical data and not for dummy variables created in place of categorical data, as the dummy variables are already either 0 or 1.

Without scaling, the cost function (difference between the model's predicted values and the actual values from your data in Linear Regression) may become difficult to minimize due to large gradients. On the other hand, when scaled, ensures the model/algorithm takes uniform steps towards all dimensions/features, thus converging faster.

Normalized Scaling: This involves adjusting the values so they all fit within a range, usually between 0 and 1. Like resizing images, so that they become the same size and are easy to compare/manage. This is used when model is sensitive to outliers or the features have different wide ranges.

Standardized Scaling: This involves adjusting the values so they have a common average (mean) and spread (standard deviation), i.e, everything is balanced around the same centre point. Like making sure all the pictures are not only the same size but also centred and evenly spread out. This is used to make sure data follows normal distribution and is generally better in dealing with outliers than normalization.

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

ANSWER: VIF (Variance Inflation Factor) is used to detect multicollinearity between independent variables in Linear Regression. So, in a model with independent variables x1, x2, x3, x4...., can we build a model that predicts say output y using x2, x3, x4..., i.e, build a model to explain *one predictor* using *other predictors*.

VIF = 1 (1 – R_i^2), where R_i is the R-squared for the model built for the i^{th} variable.

Relationship between VIF and Multicollinearity:

- VIF of 1: Indicates no multicollinearity among the predictor variables.
- VIF between 1 and 5: Suggests moderate multicollinearity, but it may not be problematic. One needs to consider low p-value (< 0.05)
- VIF above 5 or 10: Indicates high multicollinearity, which could be problematic and affect the stability and interpretation of the regression coefficients.
- **VIF = Infinity** indicates perfect multicollinearity. This means that one predictor variable is a perfect linear combination of one or more other predictor variables, i.e, one variable can be exactly predicted from the others

  Causes of VIF = Infinity: It shows either duplicate variables (same variable used twice in the model, like *x* and *2*x*), or one predictor variable is an exact linear function of another variable.

One should remove the variables with VIF=Infinity first, going for the one with the highest p-value first. Then rebuild the model and recheck the VIFs – they might change from Infinity. If not, then repeat the above process

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWER: Quantile are values that divide the data into equal-sized portions. Eg., the median is the 50th percentile or 0.5 quantile. A **Q-Q plot** (**Quantile-Quantile plot**) shows the quantiles (like medians, quartiles) of the data on the y-axis and quantiles of the distribution (usually normal distribution) on the x-axis.

In Linear Regression, one of the key assumptions is that the residuals (errors or difference between observed and predicted values) should be normally distributed. Q-Q plot is used to evaluate/assess this assumption by plotting the residuals. If it is a –

- Straight Line: points that lie close to a straight line suggest that the residuals are approximately normally distributed
- Curved Line: model might have outliers or non-linear data as residuals are not normally distributed.