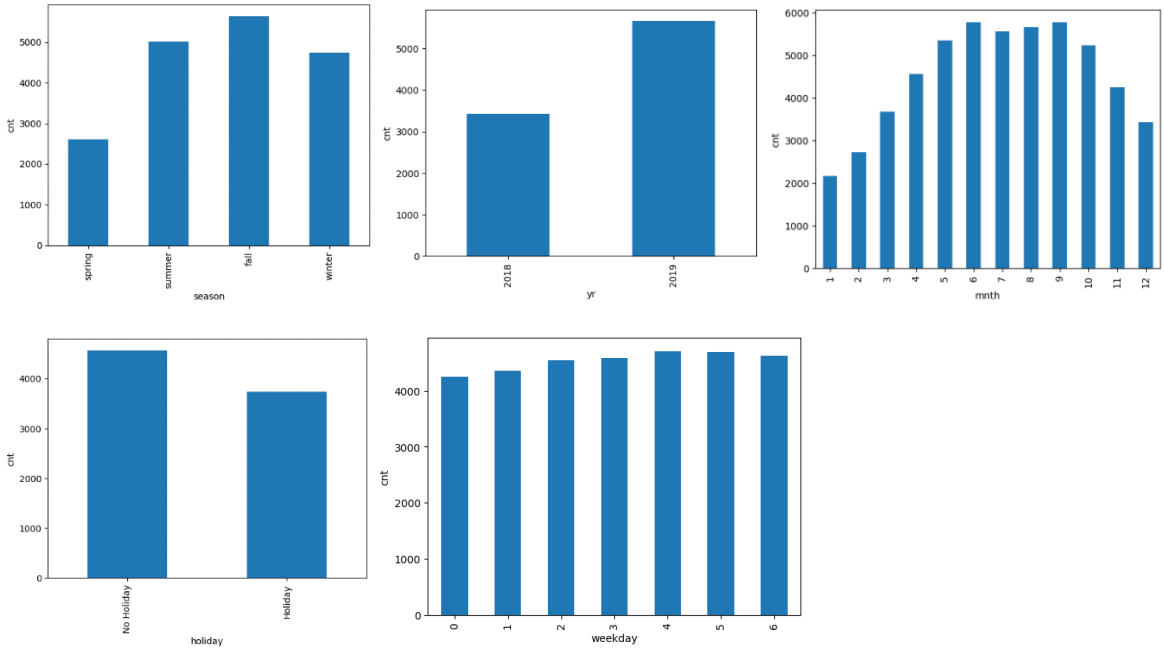
Assignment-based Subjective Questions

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

Ans: We infer that some of the categorical variables such as season, year, month, holiday and weathersit have significant impact on the output variable cnt but certain other categorical variables such as weekday, workingday do not have much impact. Same can be seen from the below plots.



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

Ans: while creating dummy variables using pandas

pd.get_dummies()

we get same number of new columns as the number of categories in the feature. For Ex- if we have 4 categories in season, we will get 4 new columns. But we don't need 4 new columns to express all 4 categories of season and we need only 3 (n-1). This option of **drop_first=True** removes one redundant column thus reducing the complexity and time needed to model the data and make predictions.

drop_first=False

	season_1	season_2	season_3	season_4
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1

1,0,0,0->season_1
0,1,0,0->season_2
0,0,1,0->season_3
0,0,0,1->season_4

drop_first=True

	season_2	season_3	season_4
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

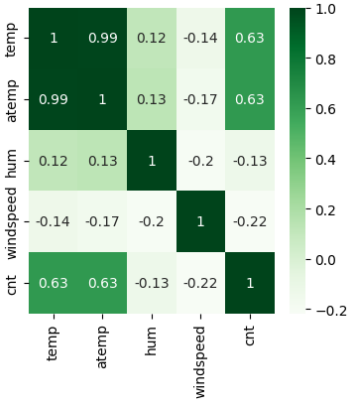
0,0,0->season_1
1,0,0->season_2
0,1,0->season_3
0,0,1->season_4

Thus, we can see drop_first represents same number of categories in one less column by removing the redundant column

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

Ans: The variable *atemp* has the highest correlation but *temp* is very close second as shown below

	temp	atemp	hum	windspeed	cnt
temp	1.000000	0.991696	0.128565	-0.158186	0.627044
atemp	0.991696	1.000000	0.141512	-0.183876	0.630685
hum	0.128565	0.141512	1.000000	-0.248506	-0.098543
windspeed	-0.158186	-0.183876	-0.248506	1.000000	-0.235132
cnt	0.627044	0.630685	-0.098543	-0.235132	1.000000



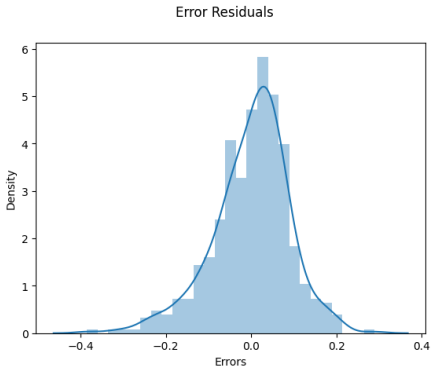
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: There are 4 assumptions of linear regression model

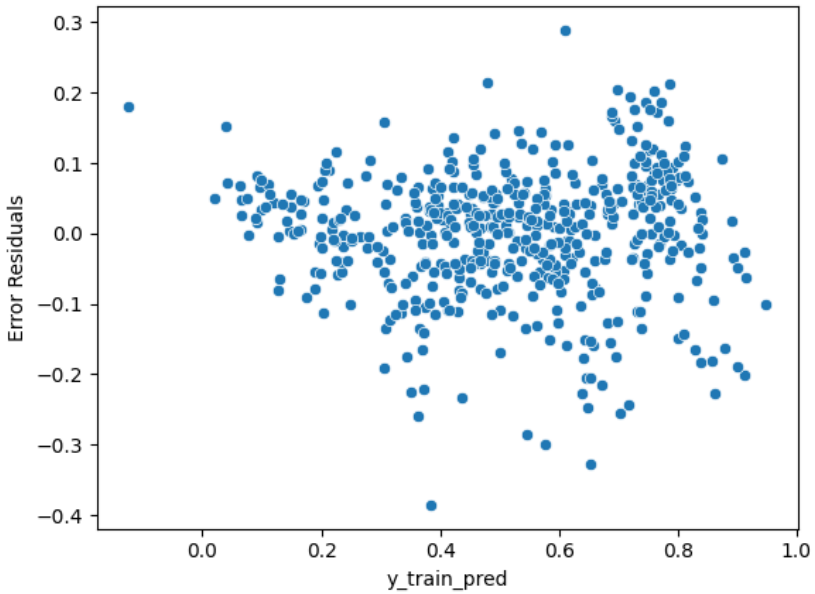
1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero(not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

The first assumption is something that has to be looked at before building the model by looking at the scatter and bar plots to check if the linear model can be used. But the question talks about AFTER building the model where the remaining 3 assumptions need to be validated.

To validate the 2nd assumption, we calculate the residual and plot the histogram of residuals as shown below. This looks very close to a normal distribution with mean 0.



The second assumption is validated by plotting the residual against the y value to see if there is any trend, and we don't see any major trend.



For the homoscedasticity assumption, we see the scatter plot of error residuals vs predicted y and we see that the data points are not having a lot of increase in variance and thus can be assumed to follow the constant variance assumption and not a heteroscedastic one.

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

Ans: The top 3 features contributing significantly towards the demand are:

S.No	Variable name	Variable Explanation	Coefficient value
1.	temp	Temperature	0.601
2.	yr	Year	0.224
3.	hum	Humidity	-0.202

The remaining variables in the model and their coefficients are show below

```
temp            0.601675
yr              0.223747
hum            -0.202301
weathersit_3   -0.145656
season_4        0.143844
windspeed      -0.137630
mnth_9          0.110944
holiday        -0.086901
season_2        0.078834
mnth_3          0.031604
dtype: float64
```

General Subjective Questions

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

Ans: Linear regression algorithm is an algorithm which is used to model the target variable with one or more input or independent variables via a linear relation. If the number of input variables is one then the regression is called Simple Linear Regression and more than one input variables are present then it is called Multiple Linear Regression Analysis. The main assumption here is that the input and output are linearly related. This allows us to create a model of the form

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\epsilon$$

Here the one unit increase in one variable will result in beta times increase in Y. There are other assumptions which are used to fit a Linear Regression model (given below)

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero(not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

In this algorithm, the model is developed by minimizing the cost function, Residual Sum of Squares(RSS) using the gradient descent optimization algorithm. The coefficients which give the least RSS is chosen as the best coefficients. This method is same for both simple and Multiple linear regression. But, for multiple linear regression we have to look at a few aspects before fitting a model

1. Overfitting – This is the tendency of the model to memorize data points rather than generalizing the data. This can happen when we keep on adding many variables and the model becomes too complex
2. Multicollinearity – In this analysis, it is assumed that the input variables are independent of each other but that normally doesn't happen in real life data. We try to minimize the collinearity of variables by using VIF and drop the variables which have high VIF value of more than 5.
3. Feature Selection – Selecting the right set of features based on business use case and to reduce overfitting and multicollinearity.

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

Ans:Anscombe's quarted is a set of 4 data points that have nearly identical simple statistical properties such as mean, standard deviation, correlation but are totally different datasets once visualized.

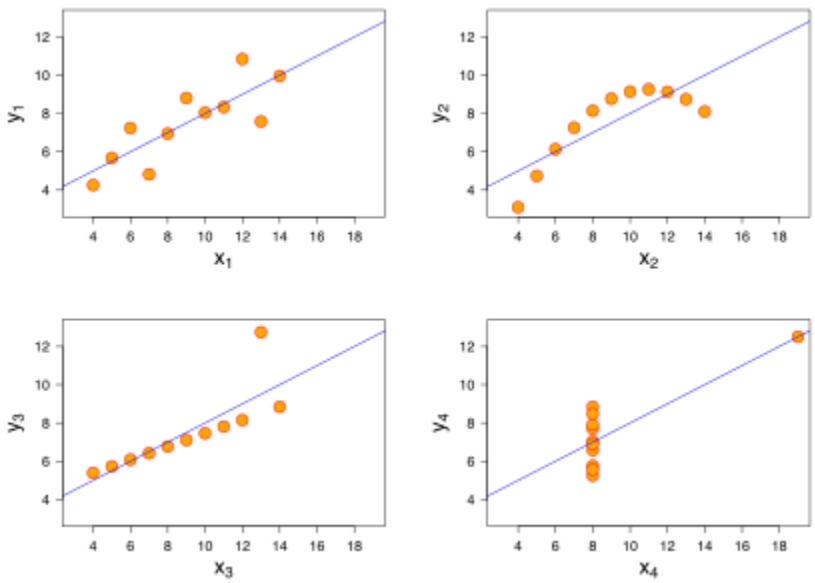


Image Source:
https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg

This quartet demonstrates the importance of visualizing the data when analyzing it and shows the effect of outliers on statistical properties.

For all 4 datasets

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

(source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

1. The first dataset has linear relationship between input and output.
2. The second dataset is non-linear in nature and a linear line cannot be fitted even though it has same correlation coefficient as the first data
3. The third data is linear but the entire relationship gets affected by just one outlier point and we can see that the fitted line is not representing the actual dataset
4. The fourth dataset shows that there is no relationship between input and output, but a highly skewed data points results in a high correlation and simply fitting a linear model is meaningless.

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

Ans: This is a method of finding the strength and direction of relationship between 2 variables i.e. correlation. It has a value from 1 to -1. In other words, if the value of Pearson's R has high magnitude

closer to 1 then we can say the two variables are highly correlated and increase in would definitely result in increase of other or decrease of other. The minus sign indicates whether the first variable(X) would increase of decrease with second variable(Y) i.e. "+" values of R denotes than Y would increase with X and "–" value of R denotes that y would reduce with increase in X.

Some example images showing the different Pearson correlation is shown below

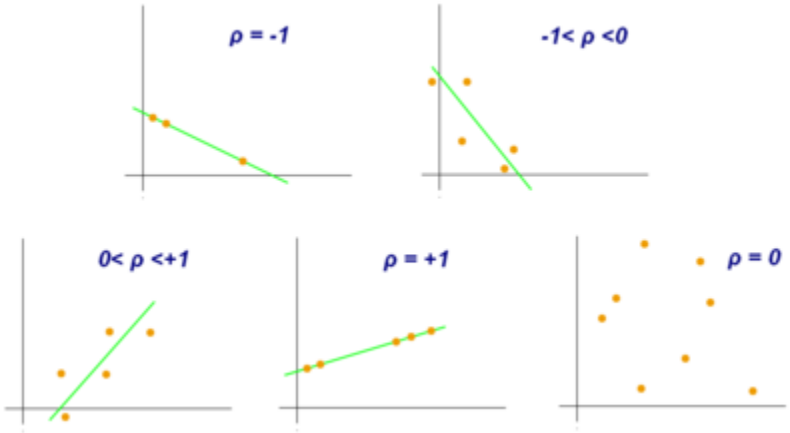


Image Source:
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_coefficient.png

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

Ans: scaling is the process of bringing variables in similar range of values or similar scales. It may happen that different variables are of different scale like one variable may range from 1 to 10 and other may vary from 0 to 10000. When we fit a linear model, the coefficients, beta, will also be of very different scales. Because of this it becomes difficult to tell which feature is more important based on the coefficients. When we scale the values in the data, the convergence of the gradient descent optimization or any other optimizer to get the final coefficients becomes faster.

In normalized scaling or MinMax scaling the data points are scaled in the range of 0 to 1 using the formula (Xi-Xmin)/(Xmax-Xmin). In standardized scaling, the data is converted into a new data with mean 0 and sigma of 1. The formula used is (Xi-mean)/sigma.

The advantage of standised scaling over normalized scaling is that it doesn't compress the data between a particular range and this is useful when we extreme data point outliers. In case of extreme data points, MinMax scaling can result in most of the data points getting compressed very close together.

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

Ans: The value of VIF is given by the formula $1/(1-R_i^2)$. This value would be infinite when Ri is 1 or -1. It means that the given feature can be explained completely by all the other features combined. It could be due to feature being a derived parameter calculated from a few other features. For Ex- if we are

given GDP, population and per capita income as 3 features then VIF of per capita income would be infinite or really high as per capita income = GDP/population.

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

(3 marks)

Ans: A Q-Q plot is a probability plot to compare two different distributions with each other by plotting their quantiles against each other. In this plot each quantile of first distribution is taken as x and each quantile of second distribution as y. If two distributions are identical, then the Q-Q plot would be a straight line y=x. This kind of plot is used to compare various aspects like shifts in location, shifts in scale presence of outliers etc. of two distributions simultaneously. This kind of plot is useful in linear regression to determine if two different features come from same distribution. This is useful in scenarios when we have training and test data given separately and we can confirm if both the datasets are from populations with same distribution using the Q-Q plot.