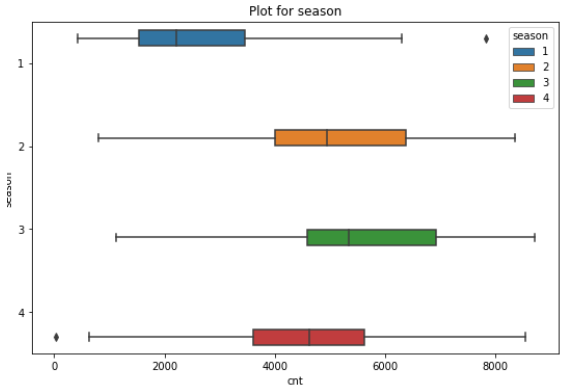
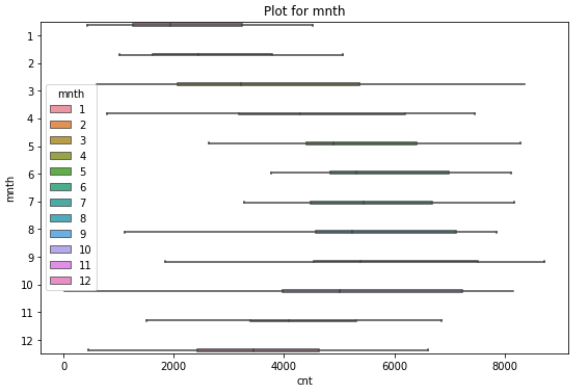
Assignment-based Subjective Questions

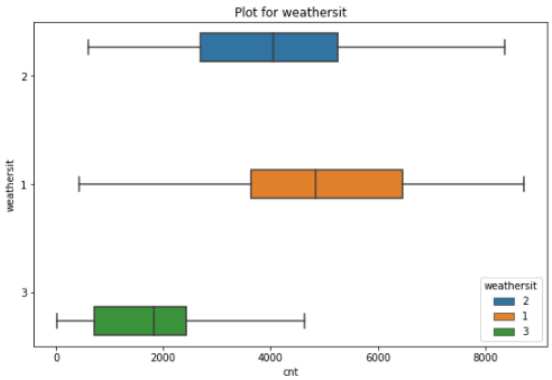
1. ***<u>From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)</u>***

   Answer - Here is the analysis of categorical variables with target variable –

   - Variables – **Season, Month and weathersit**



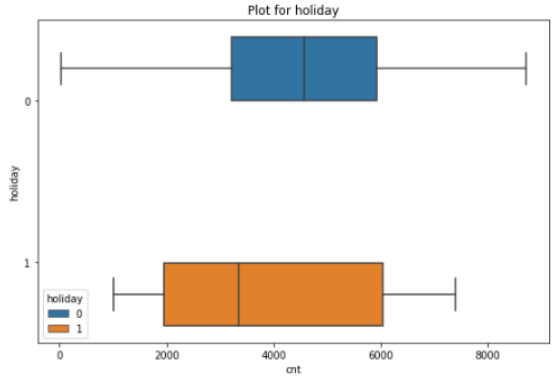


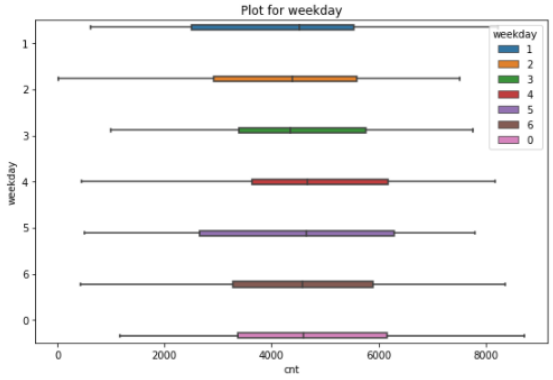


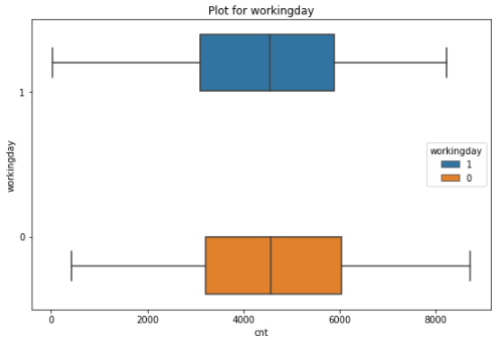
As can be seen these variables have an impact on the target variables, looking at the trend people are more likely to showup and the demand grows in Summers (Season 2/3 and Months 5/6/7) compared to other months/seasons. This also suggests the data is collected from a place with cold climate where people enjoy cycling more in a warm weather.

The variable “weathersit” equally explains the variation in demand based on the weather situation on a given day. As can be clearly seen in the graph above, the demand is higher with value=1 i.e. Clear, Few clouds, Partly cloudy, Partly cloudy. The demand decreases a bit at value=2 i.e. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist. And finally for value=3 , a very low demand is expected in a Snow/Rain/thunderstorm weather situation.

- Variables – **Holiday, Weekday and workingday**







It can be seen that the demand on a holiday is more stochastic (IQR - 2000-6000), however is more concentrated and predictable (IQR - 4000-6000) on a non-holiday.

The demand is slightly higher on Day 4, however, the other days do not show high variability. The same can also be looked in with variable “workingday”, where it does not differ a lot if the day is a weekend/holiday or a regular working day.

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

Answer –

N- levels in a categorical variable can be represented by N-1 dummy variables, for e.g. variable “weathersit” in the data had 3 distinct values 1,2 and 3. We can represent them using 2 variables only i.e.

Value	Dummy_var1	Dummy_var2
1	0	0
2	0	1
3	1	0

However if we look at how pandas pd.get_dummies() works, it create N variables (one-hot encoded) for representing N levels i.e.

Value	Dummy_var1	Dummy_var2	Dummy_var3
1	1	0	0
2	0	1	0
3	0	0	1

Therefore, drop_first= True will remove the first dummy variable from the output generated by get_dummies(). Not doing so will unnecessary keep an additional variable in the data that eventually represents the same information as by remaining dummy variables created when they are all set to value 0.

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

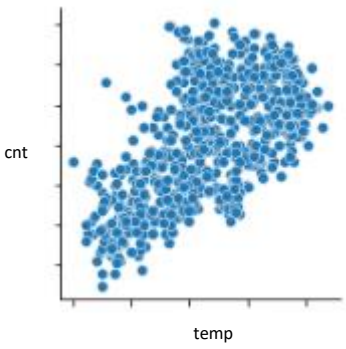
Answer –

Variable “temp” is highly correlated to target variable. It is also highly correlates to “atemp” therefore both “temp” and “atemp” eventually correlates high to target variable.

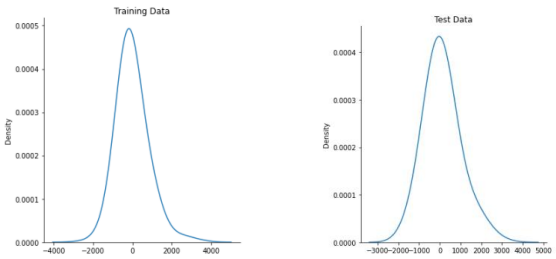
4. ***How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)***

Answer –

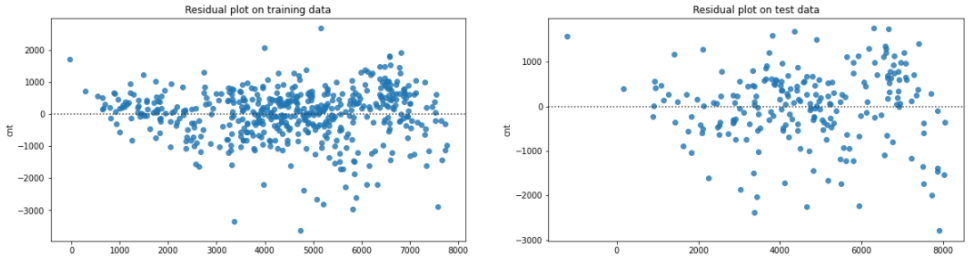
- *Linearity* - Atleast one variable shows a linear relationship to target variable during EDA - Yes, we saw variable temp has a positive linear relationship



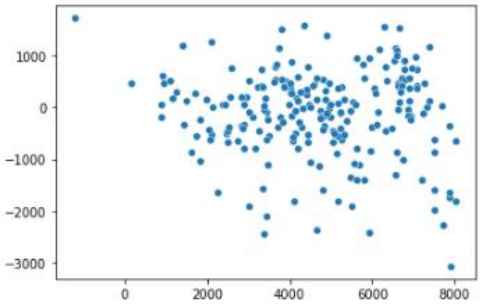
- *Error terms are normally distributed* - Yes, visible through the density plots, Plot is centered as zero and seems like a normal distribution.



- *Homoscedasticity* - Yes, Error term seems to have almost constant variance (No cone shape) as seen in the residual plot.



- *Error terms are independent* - Yes, Can be verified from plot of error vs predicted, as there is no relationship seen between them



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

Answer –

1. "temp" – Has the highest positive coefficient and p value nearly 0, higher the temperature, higher the demand
2. "hum" – Has a high coefficient and p value nearly 0, but in negative direction to target variable which means higher the humidity lower the demand
3. "yr" – Has a high coefficient and p-value nearly, it means business is becoming popular as it is getting older, there is high changes of coming years to see more demand

	coef	std err	t	P>|t|	[0.025	0.975]
const	1990.6743	195.052	10.206	0.000	1607.452	2373.897
yr	2024.1545	72.500	27.919	0.000	1881.712	2166.597
temp	5098.6614	170.310	29.938	0.000	4764.050	5433.273
hum	-2039.5256	225.019	-9.064	0.000	-2481.624	-1597.427

General Subjective Questions

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

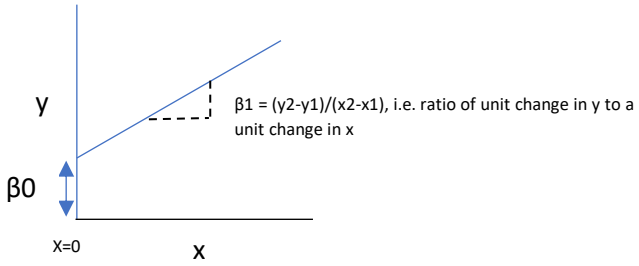
Answer –

Linear regression is an approach to define a linear relationship between independent (driver) variables and a single dependent(target) variable. It is one of the oldest and commonly used supervised predictive analysis technique.

Types- There are two types of linear regression models

1. Single Linear regression- When there is only one independent variable
2. Multi Linear regression – When there are more than one independent variables

Simple linear regression is represented with the equation y= β0+ β1*x, where β0 represents the intercept and β1 represents the slope of a linear regression line on a 2-D plot of single independent and dependent variable.



Similarly, for a multi linear regression with N dependent variables, the equation is

represented as -

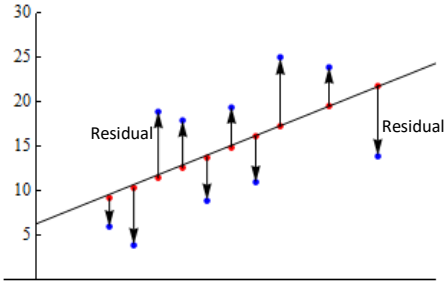
y= β0+ β1*X1 + β2*X2 + β3*X3 + β4*X4 + ---------- + βN*XN

Best fit line in linear regression –

The best fit line is the one that fits in a way that minimizes residual sum of squares (RSS).

Residual for a given point = (Actual value – Predicted Value) = y-(β0+ β1*X)

RSS = $\sum_{i=1}^{n}(y(i) - \beta - \beta(i) * X(i))^2$



Hypothesis - The null hypothesis for a linear regression is H0 : β1 = 0

and the alternate hypothesis is HA: β1 ≠ 0

, the variable coefficient β1 has no significance. We try to find out variables in a linear regression model that have low p-value < 0.05 so as to reject the null hypothesis.

Assumptions – Here are the assumptions for a linear regression

1. Linearity - Atleast one variable shows a linear relationship to target variable
2. Error terms are normally distributed
3. Homoscedasticity - Error term seems to have a constant variance.
4. Error terms are independent – Error terms have no relation to either dependent or independent variables.

Calculations – Generally Ordinary least square (OLS) method is used to calculate the variable coefficients in a linear regression.

Evaluation – The goodness of a linear regression fit can be determined by

1. RSE(Residual Standard error)
2. R2 (Coefficient of determination)

R2 = 1- (RSS/TSS), where

RSS = $\sum_{i=1}^{n}(\mathrm{y(i)} - \beta - \beta(\mathrm{i}) * \mathrm{X(i)})^2$

TSS = $\sum_{i=1}^{n}(\mathrm{y(i)} - \mathrm{y(mean)})^2$

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

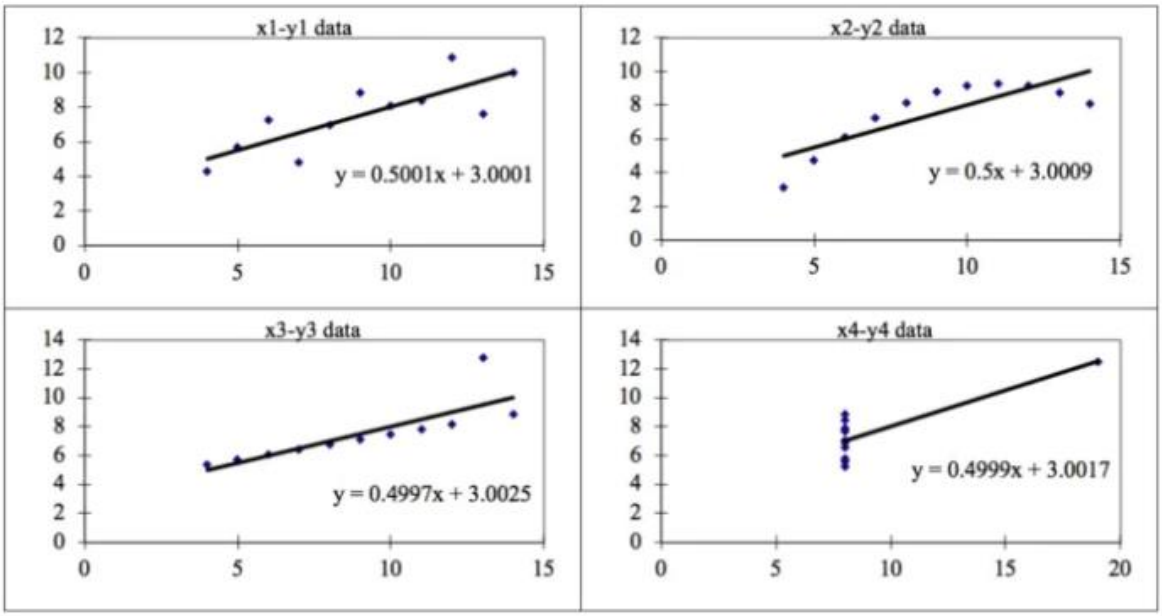
Answer –

It is named after statistician Fransic Anscombe in 1973, It consists of 4 datasets which look nearly same on simple descriptive statistics; however they look very different when plotted. This quartet is used to describe the limitation of descriptive statistics for explaining realistic datasets. This eventually tells us the importance of plotting the data to check the distribution before deciding the algorithm to build models.

The image below shows the statistical summary of the 4 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		

When the 4 datasets are plotted on scatter plots, they show different patterns



Dataset 1 - Seems like a good fit for a linear regression model

Dataset 2 – Data points seems non-linear

Dataset 3 – It has an outlier which will divert a linear regression model

Dataset 4 - It has an outlier which will divert a linear regression model

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

Answer –

Definition -

It is a measure of strength of linear correlation between two continuous sets of data.

Formula -

It is a normalized measure of covariance calculated by dividing covariance in two data cov (X,Y) by the product of their standard deviation SD(X) *SD(Y).

$R_{xy} = S_{xy} / S_x * S_y$, where S_x and S_y are standard deviations for X and Y.

Interpretation-

It ranges between -1 and 1. Both the magnitude and direction are important to see a correlation between two data/variables. If there is a positive value (direction) this means increase in one variable increase the other and similarly for the decrease.

A negative value or direction indicates when one variable increases the other would decrease.

A higher value is considered a strong correlation between the variables/data. A zero value mean there is no correlation between the variables.

It is one of the ways to check multicollinearity in data variables.

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

Answer – Scaling is a technique to normalize the range of independent variables. This is performed as a part of data preprocessing in machine learning.

Scaling is required because

1. It nullifies the impact of magnitude and units of the data, which if not done could lead to incorrect model results. As an example, if temperature variable is mentioned in Fahrenheit vs. degree Celsius it would have an impact on the model however when scaled both normalizes between a standard range and the impact is minimalized. This is a very common scenario with realistic data, another example is currency, USD vs INR Million vs Billions etc.
2. Scaling improves model training process by limiting the calculations in a defined range and does not impact the model performance/accuracy.
3. It is easy to relates and visualize multiple variables when they are on same scale.

Types –

1. Normalized scaling – Also known as MinMax scaling, It is the simplest method to normalize data in 0-1 range. The formula for scaling is given as –

   $x = (x-x_{min})/ (x_{max} – x_{min})$

2. Standardized scaling – In this approach variable values are replaced by their Z scores, thereby making all the values into a standard normal distribution with mean zero. The formula for this scaling is given as -

   $x = x-x_{mean} / SD(x)$

Differences in Normalized vs standardized scaling –

Normalized scaling	**Standardized scaling**
Scaled values are in range 0-1	Scaled value are Z score of original values
It is simplest approach to apply	This approach needs calculation of mean and standard deviation
It may lose information about outliers in the scaled data in some cases	It keeps outlier information in the scaled data
Can be done with sklearn.preprocessing.MinMaxScaler in python	Can be done with sklearn.preprocessing.scale in python

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

Answer – Infinite VIF represents perfect correlation. It may happen that a variable in data can be perfectly explained by a linear relationship with other independent variables,

for example, if there are two independent variables A and B and B = A/2, In this case we can fit a perfect linear line between A and B or in other words B can be perfectly be calculated using

variable A. This situation results in high collinearity between A and B and therefore the VIF observed will be infinity.

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

Answer – Q-Q stands for Quantile to Quantile plot, it is plotted between quantiles of two distributions with respect to each other. If both distributions are similar we will get a 45 degree straight line on a Q-Q plot, A deviation from this reference line depicts skewness between distributions. This graph is usually used to test

1. If the two datasets come from the same distribution.
2. Type of data distribution such as Normal, exponential or Uniform etc.

Use in linear regression – Q-Q plots are used when modelling linear regression to

1. See that error terms in a linear regression model are normally distributed
2. Check that Train and unknown Test data belongs to same population