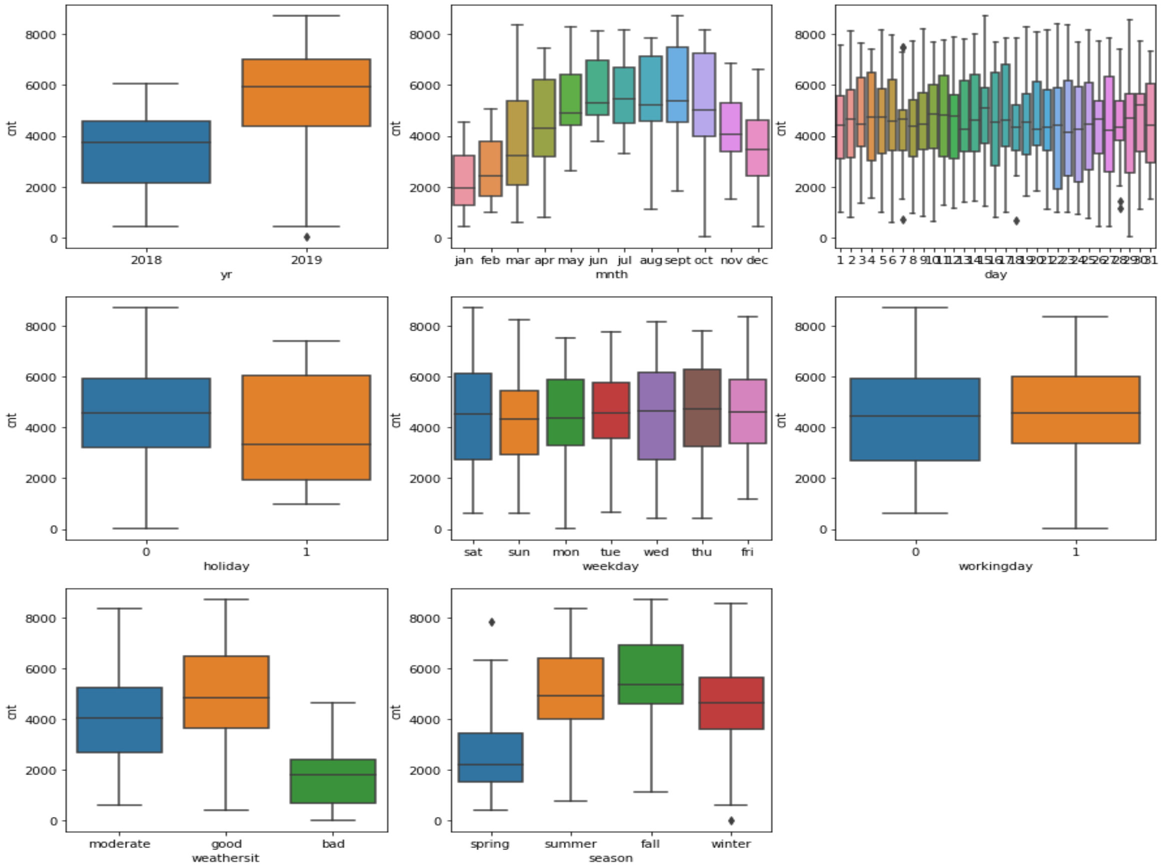
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



- Counts vary a lot during the months of March, April, August and October.
- Number of rides are actually more during the period between May till July.
- Count is more in 2018 compared to 2018.
- Rides on holidays are more as obviously expected.
- Day of the week does not create much impact.
- Having a good Weathersit significantly affects the target variable.
- Rides are more during Fall, followed by summer and significantly less during spring and winter.

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

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

For example, in the given bike sharing dataset we have 4 values in Categorical column-season and we want to create dummy variable for that column. If 3 dummy variables are created for summer, winter and spring, it is obvious we need the left-over data is of fall season and we may not one variable to define the same.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Variable cnt has linear relationship with temp and atemp.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set?

Validation was done using following assumptions:

1. Linear Relationship among variables
2. Less to no-multi collinearity between variables
3. Normality of error terms

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and windspeed.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

In Machine Learning, Linear Regression is a supervised machine learning algorithm, that computes the linear relationship between the independent variable(target) and the dependent variables. Linear relationship among variables, define how a change in a given independent variable can influence the behaviour of dependent variable.

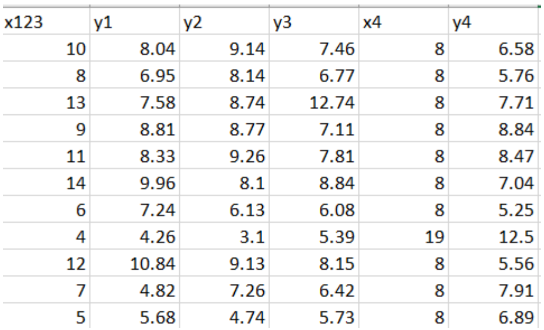
Depending on the number of dependent variables, linear Regression can be classified into two types:

1. Simple Linear Regression – one dependent variable
   Linear equation for SLR is given as: $y= \beta o+\beta 1X$
   y – Independent variable
   βo – Intercept
   $\beta 1$ - slope

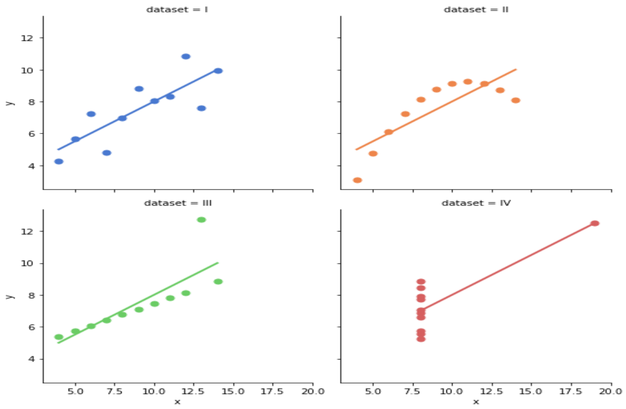
2. Multiple Linear Regression – multiple dependent variables.
   Linear equation for MLR is given as: $y= \beta o+\beta 1X1+ \beta 2X2+\beta 3X3+\ldots.. \beta nXn$
   y – Independent variable
   βo – Intercept
   $\beta 1, \beta 2, \beta 3\ldots\beta n$ - slopes

2. Explain the Anscombe's quartet in detail

Anscombe's quartet is a group of datasets which are nearly identical in descriptive statistics, but shows difference in linear regression when plotted. These datasets are helpful in stressing the importance of visualising the data for exploratory analysis.



x123	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.1	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.1	5.39	19	12.5
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89



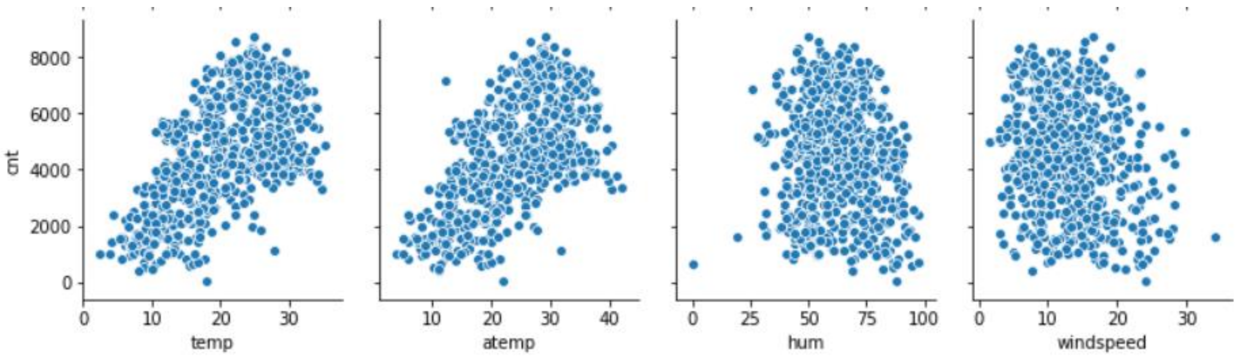
As shown in above datasets and plots:

- Dataset 1 – shows linear relationship with respect to few features
- Dataset 2 – does not have a linear relationship
- Dataset 3 – shows presence of outliers
- Dataset 4 – shows collinearity among features

3. What is Pearson's R?

Pearson's R is a correlation coefficient that defines the strength and direction of a linear relationship among continuous variables. Value ranges from -1 to 1.

In our given dataset the relationship of target with various numeric variable as plotted below

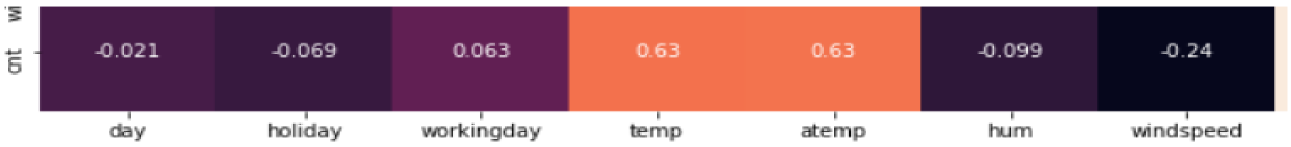


Here we observe 'cnt' increases whenever there is an increase in temp and atemp and no such relationship can be defined with respect to hum and windspeed.

Pearson correlation draws a line of best fit through two variables, indicating the distance of data points from this line. A 'r' value near +1 or -1 implies all data points are close to the line. An 'r' value close to '0' suggests data points are scattered around the line.

R value for our dataset can be observed as

1. 0.63 for both temp and atemp – indicates a strong positive relationship with temp and atemp
2. hum is very low -0.099 – indicates a very low negative relationship
3. windspeed is -0.24 – indicates negative relationship



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

and standardized scaling?

Feature Scaling is a pre-processing technique which helps us to fit the given data into a similar scale of values. This becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

S.No	**Normalisation**	**Standardization**
1	It is useful when features are of different scales. Scales values between [0,1]	It is used when we need to ensure zero mean and unit standard deviation
2	Retains the shape of original distribution	Changes the shape of original distribution
3	Sensitive to outliers	Least sensitive to outliers
4	$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	$X' = \frac{X - \mu}{\sigma}$

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

VIF(Variance Inflation Factor) is a factor that helps us understand collinearity among dependent variables.

Value can be calculated by VIF=1/(1-R^2), where R^2 gives the relationship between linear model and the dependent variable.

When R^2 = 1 means that the model is absolutely linear or behaves exactly the same as dependent variable.

This makes the VIF to be infinity.

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

'Q' stands for quantile and the Q-Q plot, as the name implies, is the comparison of distributions of two different datasets. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

Importance of QQ Plot in Linear Regression: In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape