Assignment Based Subjective questions:

1. Effect of categorical variables on dependent variable

I

- Year:

  If the year is 2019, count of total bike rentals or cnt is higher

- Holiday:

  if its a holiday, count of total bike rentals or cnt is lower

- Working day:

  working day doesn't seem to have any effect on total bike rentals

- Season:

  Count of total rental bikes (cnt) is highest in the fall season followed by summer and winter. It's the lowest in spring season

- Month:

  Count of total rental bikes (cnt) tends to increase in the months June to September while it's low in the months from January to March.

- Weekday:

  Count of total rental bikes (cnt) is higher on Sundays and Thursdays while it's low on Mondays and Tuesdays.

- Weathersit_relablled:

  Count of total rental bikes (cnt) is higher on days with clear weather and few clouds followed by misty days or days with broken clouds. However, cnt is low on days with light rain or thunderstorms and scattered clouds

2. Importance of using drop_first= True while dummy variable creation:

   We use drop_first= True while creating dummy variables in order to avoid creating unnecessary independent variables which would otherwise unnecessarily increase the number of predictors in our prediction model and thus making it more complex than it's needed.

3. Numerical variable with highest correlation with target variable is temp (temperature in Celsius)

4. Validating the Linear Regression Assumptions:

a. The first assumption i.e. X and Y have a linear relationship, can be validated by simply plotting a scatter plot between X and Y. If the x and y points show a clear positive relationship without turning backwards or anything then the assumption is validated.
b. The 2nd assumption i.e. the errors (or residuals) are normally distributed can be easily validated by plotting a distplot for the residuals using seaborn. By looking at the distplot, if the distribution resembles a normal or Gaussian distribution i.e. a bell shaped curve, then our assumption is validated.
c. The 3rd assumption is, the errors are homoscedastic i.e. errors have constant variance. This can be validated by simply plotting a scatter plot of the residuals vs the y hat values. If the scatterplot shows absolutely no trend whatsoever, then our assumption is validated.

5. Top 3 features from our final model:

The top 3 features of our temp, yr and winter.

- Temp stands for temperature in celsius. It's a numeric variable and has proved to be significant. It's coefficient turned out to be 4089.863659. This means there is an increase of 4089(approx) new bike rentals with a every 1 unit increase in the temperature.

- yr variable is a binary variable indicating the year (0: 2018, 1: 2019). It proved to be significant and the coefficient turned out to be 1978.412291. This means that there is an increase of 1978 total bike rentals(approx) if the year increased by one.

- winter is a dummy variable we created out of the season variable given in the original dataset. It's a binary variable (0: bike not rented in the month of winter, 1: bike rented in the month of winter). It also proved to be significant and its coefficient turned out to be 1097.585083. This means there is an increase of 1097 total rentals (approx) everytime when the season is winter.

General Subjective questions:

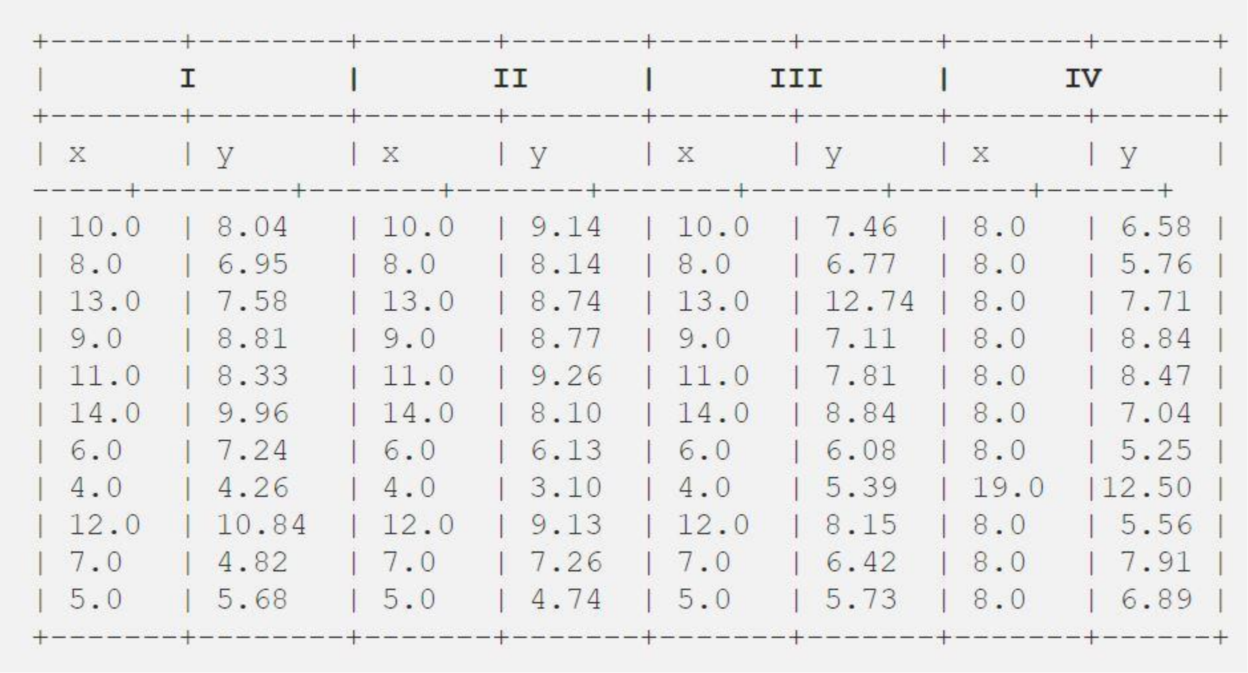
1. Explain the Linear Regression algorithm in detail:

Linear Regression is a linear model that defines a linear relationship between a dependent variable (say y) and an independent variable x (or independent variables x1, x2, ... in case of multivariate linear regression).
The model is used to calculate or predict the value of y or the target variable given an intercept c, the slope of the independent variable x (or variables x1, x2, ...) and the values of the independent variables.
Linear Regression has 2 types:
1. Simple Linear Regression where there is only one independent or predictor variable x:
y= c+ mx+ e
where: y= target/independent variable
x= independent or predictor variable
c= intercept
m= slope of x
e= errors
Assumptions: a. There is a linear relationship between x and y
b. Error terms are normally distributed
c. Error terms are independent of each other
d. Error terms have constant variance i.e. they are homoscedastic
2. Multivariate Linear Regression where there are more than one predictor or independent variable variable:
y= c+ m1x1+ m2x2+ .....+ mnxn+ e
where: y= target/independent variable
xn= nth independent or predictor variable
c= intercept
mn= slope of nth x (predictor) variable
e= errors
Key points about MLR:
a. The model now fits a hyperplane instead of a line.
b. Coefficients are still obtained by minimizing the sum of squared errors, the least squares criteria.
c. assumptions from simple linear regression still hold: zero mean, independent and normally distributed error terms with constant variance.

2. Explain the Anscombe's quartet in detail:

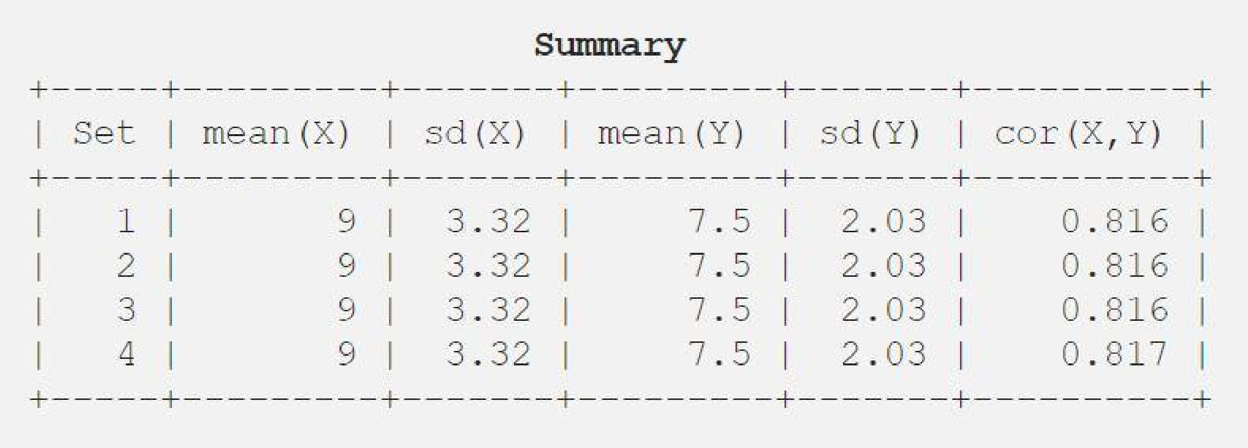
According to Andscombe's quartet, there are 4 datasets with x-y pairs of values that result in very similar simple statistical summaries i.e. all have the same mean(x), sd(x), mean(y) and sd(y). However, when graphed these 4 datasets appear very different. This was discovered by Francis Andscombe.

The 4 datasets are as follows:



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

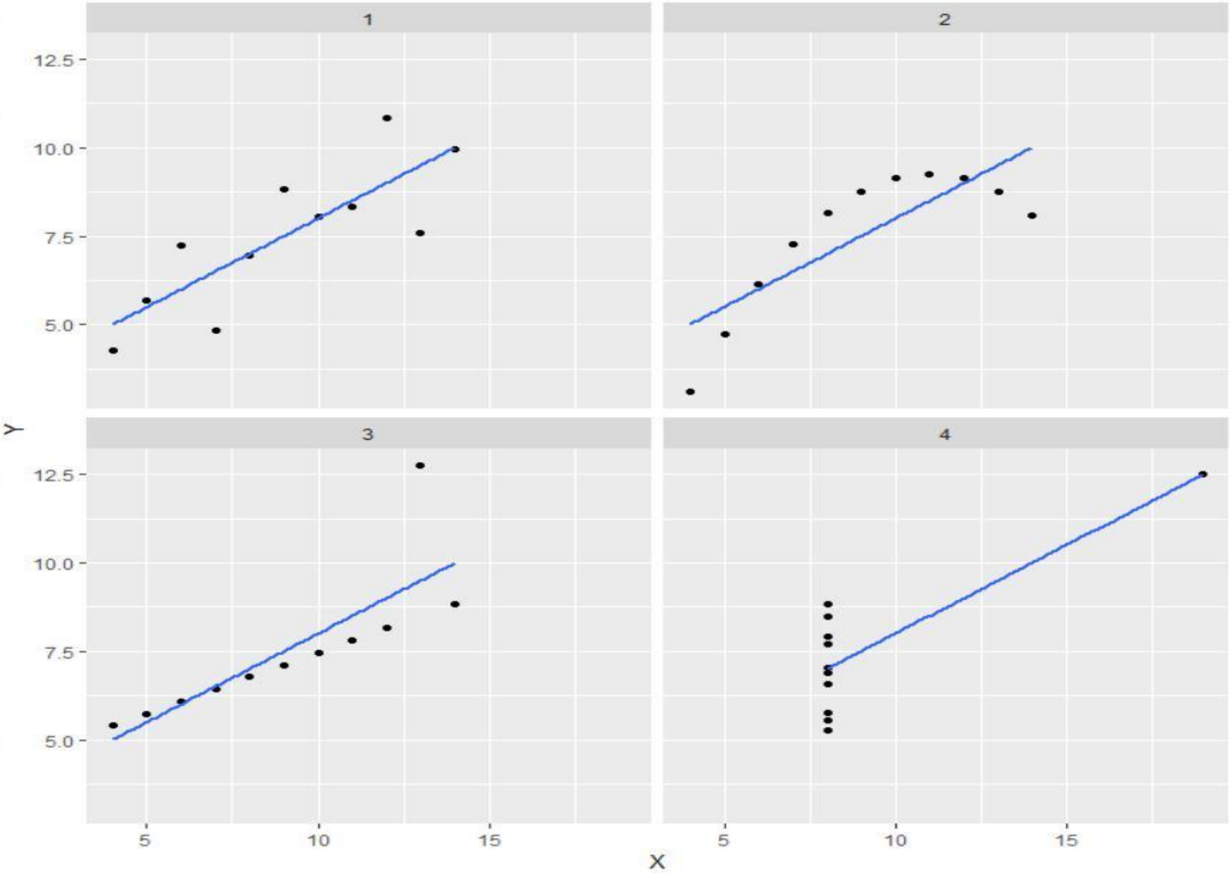
The statistical summaries of these 4 datasets are follows:



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

The graphs of these 4 datasets showing scatterplots and regression lines fitted through each of them appear as follows:



So as we can see, from the above that all the 4 datasets have the same simple statistical summaries i.e. all have the same mean(x), sd(x), mean(y) and sd(y).

Despite this, the graphs of these 4 datasets look very different:
1. In the first graph, we can see that x and y seem to have a linear relationship.
2. In the 2nd graph, there seems to be a non-linear relationship between the x and y.
3. In the 3rd graph, there is a clear linear relationship between x and y except one datapoint that seems to be an outlier.
4. In the 4th graph we can see x and y clearly don't have a linear relationship but one outlying datapoint alone is enough to make the regression line seem linear and produce a high correlation coefficient.

Thus, Andscombe's quartet proves to us that we should always look at our data graphically before making any conclusions based on the simple statistics like mean and sd.

3. What is Pearson’s R?

The Pearson’s R or the correlation coefficient is a measure of the linear relationship between 2 datasets (say x and y). It measures how strong or weak a linear relationship is between 2 variables and it lies between -1 and 1.

A. A correlation coefficient of 1 indicates a strong positive correlation..
B. A correlation coefficient of -1 on the other hand indicates a strong negative correlation.
C. A correlation coefficient of 0 means there is no relationship at all.

The formula for calculating Pearson’s R or the correlation coefficient is as follows:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

4. What is Scaling? Why is it performed? What is the difference between normalization and standardization methods of scaling?

Scaling is the process of transforming the independent predictor variables before building the linear regression model such that all predictors are on the same scale.

Scaling is performed for the sake of improving interpretability of the model e.g. the coefficients. A model usually consists of many independent variables where some are numeric and some are categorical. Some variables may be storing values on a much bigger scale or completely different units than the others which will make the interpretation of their coefficients needlessly complicated. So it's always better to have the variables on the same scale as it makes it much easier for anybody to understand the model..

Normalization vs Standardization:

Normalization and standardization are the 2 popular methods of scaling.

Normalization can be done for a given variable X by using the following formula:

(X- Xmin)/(Xmax-Xmin)

Standardization on the other hand can be done for a variable X by doing the following:

(X- Xmean)/Standard Deviation of X

a. Hence, normalization uses the maximum and minimum values of the given variable while Standardization uses mean and standard deviation

b. Standardization is better suited when the distribution at hand follows a gaussian distribution.
   Normalization on the other hand, is more suited when we are not sure of the distribution of the given variable.

c. If the X value lies between the maximum X value and minimum X value (which it normally should), then normalization will give us a value between 0 and 1. However, when there are outliers, normalization gets impacted.
   This however, is not the case in Standardization as it doesn't have any preset range for the values it results in and so it's better in dealing with outliers.

5. Why is sometimes the value of VIF infinite?

We know that variance inflation factor or VIF is a way to measure the degree of multicollinearity of a given independent variable with respect to the other independent variables of a model.

The variance inflation factor is calculated by the formula: 1/(1-R Squared).

Here R Squared is simply the square of R or the correlation coefficient.

We know that when there is perfect positive correlation, R=1.

Hence, R Squared is also 1.

Therefore, if we plug this value of R squared into the VIF formula we will get infinity i.e

1/(1-R Squared)

= 1/(1-1)

= 1/0

= infinity

Therefore, when there is perfect correlation, the variance inflation factor or VIF becomes infinity.

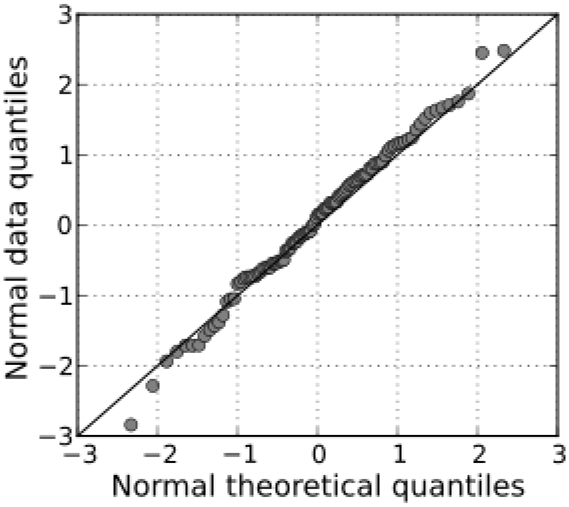
6. Q-Q plot and its use in Linear Regression

Q-Q plot is a type of graph that allows us to assess whether a given dataset came from a normal, exponential or uniform distribution or not.

It also makes it possible for us to assess whether 2 given datasets from different populations share a common distribution or not.

Basically in a QQ plot, we plot the quantiles of the first dataset against the quantiles of the 2nd dataset in question.

Example of Q-Q plot



The way we interpret this plot is: if the data points of the quantiles plotted here lie on or are close to the straight 45 degree line then it means the datasets in question are of a similar distribution.

However, if all the quantile points are lying away from the 45 degree straight line then the datasets are of a different distribution.

Use in Linear Regression: In linear regression, if we receive training and test datasets separately and doubt that they may be of different distributions, then we can find it out for ourselves using the Q-Q plot.