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?

Ans: There are 7 categorical variables present in data namely season, year, month, holiday, weekday, workingday, weathersit and there effect on dependent variable i.e. count is as below:

1) Average count of total rental bikes is less for spring season, for other seasons it is between 4000-6000
2) Average count of total rental bikes has increased in year 2019 as compared to year 2018
3) Average count of total rental bikes gradually increase from month 1 till month 7, and then decreases
4) Average count of total rental bikes is less for holidays as compared to non-holidays
5) While weekdays and workings days have no impact on average count of total rental bikes
6) Average count of total rental bikes is higher when weather conditions are (Clear Few clouds, Partly cloudy, Partly cloudy) and it reduces as weather condition gets worse, while there is no any count when weather conditions are (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)

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

Ans: To explain this let's take an example of gender, so there are three values which can be present in gender i.e male, female, neutral. Since it is a categorical column we create dummy variables for this as below:

Person	Male	Female	Neutral
1	0	1	0
2	1	0	0
3	0	0	1

Now for 1st person binary value 1 belongs to female column so 1st person is female, similarly 2nd and 3rd person is male and neutral. Let's say we drop first column male then,

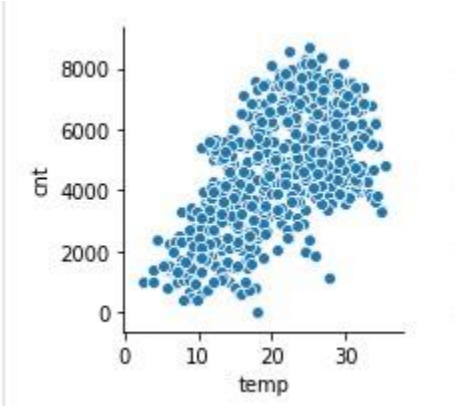
Person	Female	Neutral
1	1	0
2	0	0
3	0	1

Person	Values	Gender
1	10	Since female has value 1, so it is female
2	00	Since binary value 1 doesn't belong to female and neutral so it has to be male
3	01	Since binary value 1 belongs to neutral, so it is neutral

So even after dropping first column results are same, so it is redundant to use first column so we drop it.
Also, we keep the first column then there will be high correlation among these columns and VIF will reach to infinity

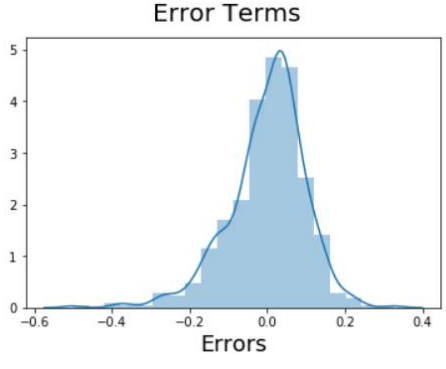
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans: If you see in below pairplot, temp has highest correlation with count.

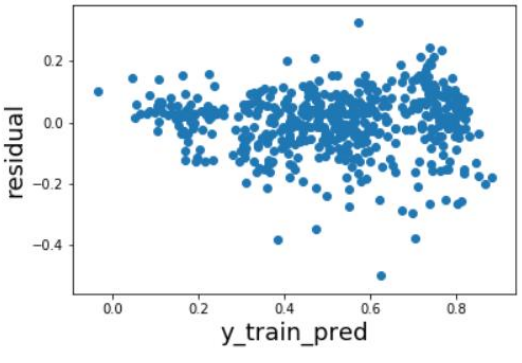


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: To validate assumption of Linear Regression model,



1) We check whether error term is normally distributed or not, also we check average mean of error term is zero.
2) We check for Homoscedasticity(Constant Variance), whether distribution is randomly sampled and there is no relation between residual and predicted value of target variable



3) Also there has to be linear relation between target variable and independent variables, for that p value for independent variables has to be less than 0.05 in order to reject null hypothesis $B_1 = 0$
4) We check for multicollinearity, whether there is any correlation between independent variable. This is checked by using VIF. Features with High VIF greater than 5 are removed

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

Ans: Top 3 features contributing significantly towards explaining the demand of the shared bikes with their coefficient:

1) yr 0.236430
2) holiday -0.080102
3) temp 0.417794

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a model which helps finding the relationship between dependent and independent variables with linear equation. Here we minimize the error between predicted values and actual values by using best fit line, so the task is to define the best fit line which can be divided into following steps

a) Step 1: Reading and understanding data
   First step is to read the dataset given, understanding information and columns present. Understanding target variable and dependent variable.
b) Step 2: Performing EDA
   Next step is to perform EDA and get inferences about variables and their correlation within them.
c) Step 3: Data preparation
   Here we read dtype of data if there any categorical data is present we convert into the dummy variables, Also yes and no values in columns are converted to binary variables
d) Step 4: Splitting data into train and test
   Here we split data into train and test set. We create model based on train set and perform it on test set. Also we scale features so as to standardize the values and making efficient model
e) Step 5: Building linear model
   Here we build linear model using X and y train sets. Then we select top most features with the help of R square, adjusted r square and VIF, p value.
   - We select features with low p value, as low p value(<0.05) means feature issignificant
   - There will be the case of multicollinearity i.e. dependent variables will be having high correlation within them. This features are detected by VIF value, if VIF value is greater (>5) then that feature is dropped.
   - Then we also look for r square and adjusted r square value, greater the values are better will be the model
f) Step 6: Residual analysis
   We then predict values of target variables and then calculate residual also by subtracting predicted value of target variable from actual value of target variable
   We then check assumptions on linear regression which are:
   I. Error term has to be normally distributed (done by plotting distplot)
   II. Error terms are independent of each other (there shouldn't be any patter in residual
g) Step 7: Prediction on test set
   We then calculate predicted value for target variable on test set and also calculate r square.
   R square should be close to what we have got in train set. Then we find the equation of best fit line.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet tells us we should just rely on descriptive statistics. We should graph it before analysing the dataset and also helps us understand the effect of outliers.

We have four data sets that have nearly identical simple summary statistics yet have very different distributions and appear very different when graphed.

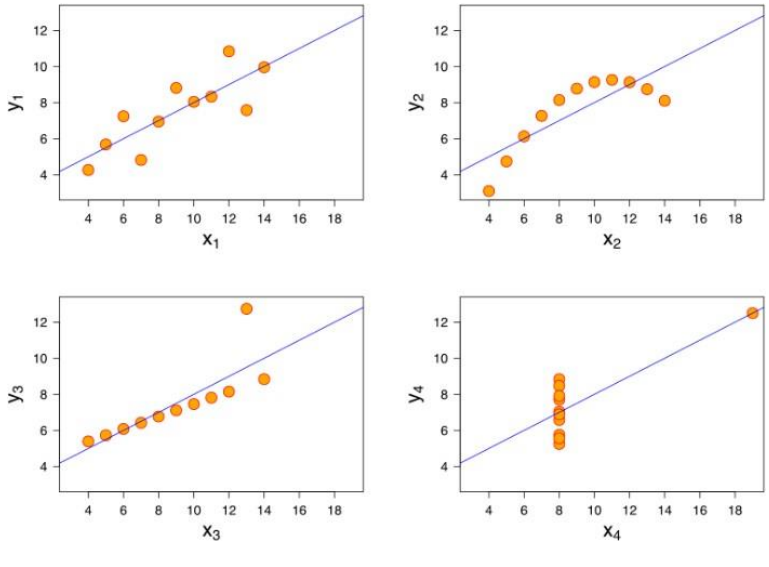
Anscombe's quartet

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

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^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

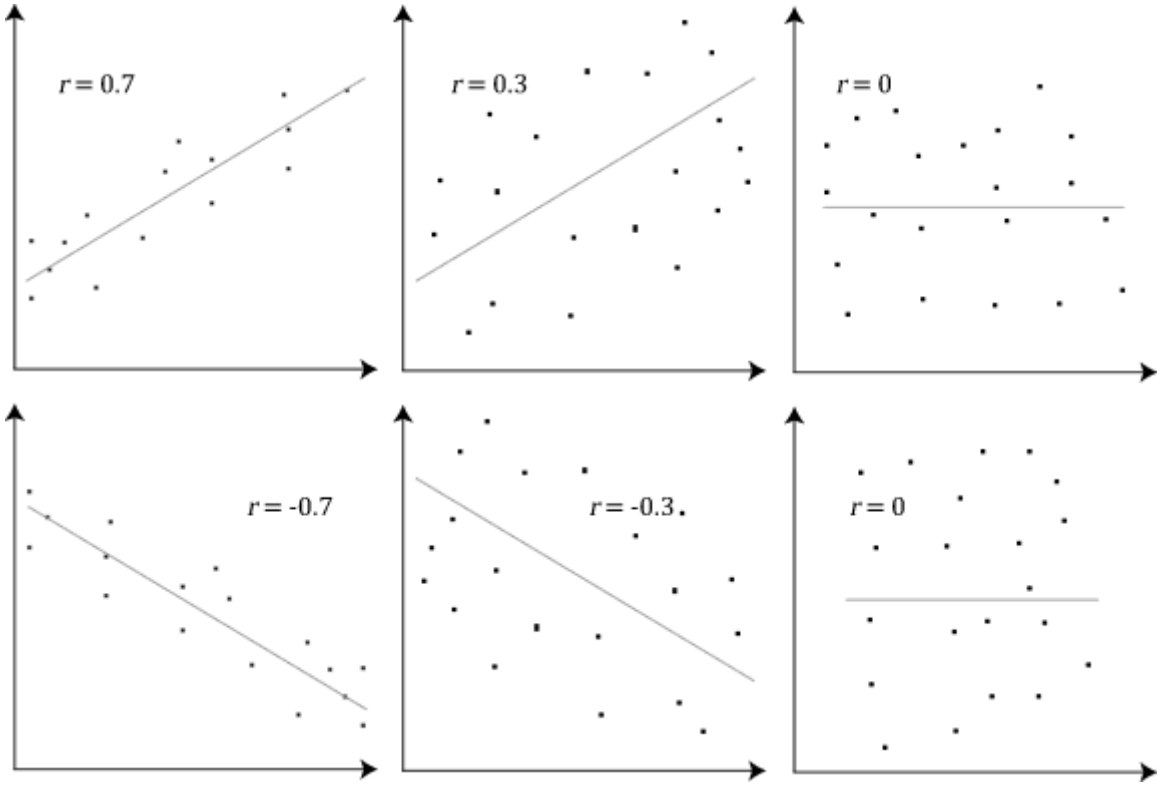
If you see the properties of these datasets they are almost identical but when you graph it you will see the difference so it tells you the importance of plotting data



3. What is Pearson's R?

Ans: The Pearson correlation coefficient is used to define the strength of a linear association between two variables and is denoted by *r*. Pearson correlation coefficient, *r*, indicates how far away all the data points of two variables are to best fit line (i.e., how well the data points fit new model/line of best fit)

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



Suppose there are two variables x and y then Pearson's r is calculated by

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

The value of the coefficient of correlation (r) always **lies between ±1**. Such as:
r=+1, perfect positive correlation
r=-1, perfect negative correlation
r=0, no correlation

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value of height 120 cm to be greater than 2 m but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

The Variance Inflation Factor (VIF) is a measure of co-linearity among predictor variables within a multiple regression. That is it is a measure of multicollinearity

A large value of VIF indicates that there is a correlation between the variables. If there is perfect correlation, then VIF = infinity. Also while creating dummy variable f you don't drop first column then VIF becomes infinity as there exist strong correlation between them. VIF of value greater than 10 is considered as dangerous, while anything below 5 is considered as safe. VIF is calculated as

$$VIF = \frac{1}{1 - R^2}$$

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical technique to help determining if two data sets come from populations with a common distribution such as a Normal, exponential or Uniform

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

i. The sample sizes do not need to be equal.
ii. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from thisplot.
iii. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions