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)**

   Answer:
   The categorical columns from the data set can be set as the following fields - **season, yr, mnth, holiday, weekday, workingday, weathersit.**
   While performing the univariate analysis of these categorical field against the field **cnt** following observations can be concluded
   Categorical columns insights
   1. The fall season observed to have a higher demand of rental bikes
   2. Clearly the year 2019 have more usage of bikes as compared to year 2018
   3. For the months may to oct the demand of bikes is more as compared to other months
   4. Bookings are more for a non-holiday as compared to holiday
   5. Equal demand can be seen for all the days of the week
   6. Books are nearly equal for a working and a non-working day

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer:
   Below are few points that explains the importance of dropping the first during dummy variable creation
   a. It does not create all the categories in a field as different columns, rather it creates 1 less column of the total categories, thus extra column is not created. Suppose there are 3 categories of a scenario execution – passes, failed and pending. While declaring the dummy variable only 2 new fields are created for passed and failed. For pending no new field is created as it is understood if both passed and failed are 0 it shall be pending
   b. It reduces the multi-collinearity which can occur between the new dummy variables. All the columns created as dummy variable should be linearly independent that means no column can define other columns. In the above example taken in point – 1, clearly passed and failed column values can define the column pending so pending column is linearly dependent. Therefore, dropping it makes the other two columns passed and failed as linearly independent

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Answer:
   temp column seems to have highest correlation with the target variable **cnt**

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

   Answer:

   a. Check whether the Error terms are normally distributed: The error terms should be normally distributed having the mean equal to 0
   b. Check the variables do not follow multicollinearity: All the driving fields selected for the model there should not be any correlation among them
   c. Check the Linearity of the fields: There should be linear relationship exists among the driving variables
   d. The error terms should not follow any pattern of their distribution which is termed as Homoscedasticity
   e. The residues should be independent of each other and there should not be any correlation.

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

   Answer:

   After selecting the model and observing the coefficients we can take top 3 features as

   a. temp
   b. spring
   c. sep

General Subjective Questions

1. Explain the linear regression algorithm in detail
   Answer:
   Linear Regression is one of the machine learning techniques applied on the data set providing insights on the predictions. This method is applied on the supervised data in which we know the independent and the dependent variables. In this algorithm we find the relationship between the dependent variable with each of independent variables. It also explains what amount of the dependent variable changes with change in an independent variable keeping all other independent variables as constant.

   The basic equation for linear Regression is provided as:

   y = mX + c

   Where –
   Y -> Dependent variable used for the prediction
   m -> slope, also termed as the coefficient of the independent variable which implies what amount of dependent variable y changes when m amount of independent variable x changes
   X -> Independent variable
   c -> constant or the y-intercept, the value of y when the regression line passed in the y axis or where X=0

   The linearity of the X with y can be categorized under –

   a. **Positive Linearity**:
   This happens when X and Y increase simultaneously, i.e., when X increases y also increases
   b. **Negative Linearity**:
   This occurs when one variable increases other variable decreases. When X increases y decreases and vice-versa

   Types of Linear Regression applied in Machine Learning model are

   a. **Simple Linear Regression**:
   When there is a single independent variable
   b. **Multiple Linear Regression**:
   When there are more than one independent variable deciding the effect on the dependent variable

   While using this algorithm for building model, below assumptions are being taken –

   a. Check whether the Error terms are normally distributed: The error terms should be normally distributed having the mean equal to 0

b. Check the variables do not follow multicollinearity: All the driving fields selected for the model there should not be any correlation among them
c. Check the Linearity of the fields: There should be linear relationship exists among the driving variables
d. The error terms should not follow any pattern of their distribution which is termed as Homoscedasticity
e. The residues should be independent of each other and there should not be any correlation.

2. **Explain the Anscombe's quartet in detail**.

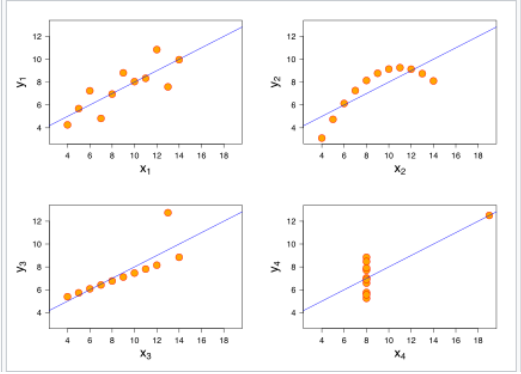
Answer:

Anscombe's quartet is a statistical analysis performed by Francis Anscombe which to prove it is always important to plot the data rather relying on the mathematical statistics. It's a group of four data sets, each sharing the same summary statistics like mean, variance, correlation coefficient standard deviation. So, all the 4 data sets have the exact same summary statistics. We can get the line of best fit. However, while plotting these data sets and visualize them in four different graphs, each of them have the data set distributed very differently

Below is the data set used to represent the above condition

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 these data sets are plotted, we observe the below –



In the first graph x1 – regression line seems to be fine – got the best fit line

In the second graph x2 – the data seems to be not liner and follows a curve pattern

In the third graph x3 – due to 1 outlier the line of best fit move towards the outlier which could have been removed through some techniques

In the fourth graph x4 – Also due to the outlier the linear regression tries to pass through the outlier point and does not fit as per the data set

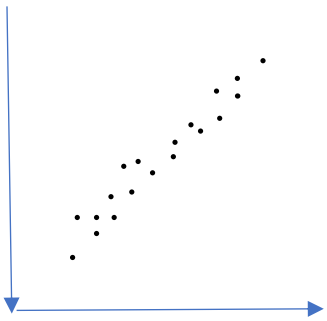
3. What is Pearson's R?

   Answer:

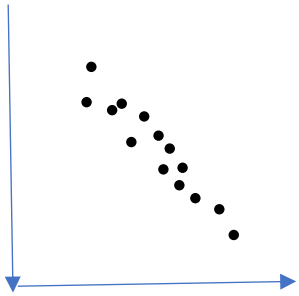
   Pearson's R is the measure of degree of linearity between two variables or measurement of linear correlation. The value ranges from a scale of -1 to 1.

   a. If the correlation coefficient ranges from **0 to 1**, the correlation type is referred as **positive correlation**. That means when one variable increases the other variable too increases and vice versa. **The two variables change in the same direction**
   b. If the correlation coefficient ranges from **0 to -1**, the correlation type is referred as **negative correlation**. That means when one variable increases the other variable too decreases and vice versa. **When one variable change in one direction the other variable changes in opposite direction**
   c. If the correlation coefficient is **0,** there is no correlation. The data points are randomly distributed.

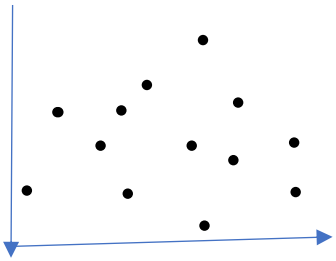
Positive correlation



Negative correlation



No correlation



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer:
   Scaling basically termed as Feature scaling is the process of normalizing the independent variables in the data set. It is method applied in the data preparation step before building the model.

   Scaling is performed to bring a uniformity across the different independent features. The data set may contain values that might differ in units, categorical ranges etc. To bring them all to a same scale, scaling is performed.
   If scaling not performed, while building model and predicting the values we end up getting the larger coefficients for the values which may be at a lower scale. For example distance travelled for one column it may be in km and for the other it may be in meters. Suppose the value lies as 5km and 300m respectively, while modelling it calculates the coefficient of 300m as higher value and the modelling goes towards incorrectness. So it is important to scale all the values to a common scale.

Normalized Scaling	Standardized scaling
min max values of the feature data is used for scaling	Mean and standard deviation used for scaling
scales the values between 0 and 1 or -1 to 1	No such boundary conditions are present
Outliers may impact	Not much impacted by outliers

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   VIF is defined as –

   VIF = 1/(1-R-square)

   R-square = 1-(RSS/TSS)
   = (TSS-RSS)/TSS

   For VIF to be infinite,
   r-square value needs to be 1 so that 1/(1-1) = 1/0 = infinite

   For r-square to be 1,
   RSS should be 0, so that (TSS-0)/TSS = TSS/TSS = 1

   RSS is the measure of how close the actual value is from the fitted line.
   If RSS to be 0, the actual values lies exactly on the fitted line.

This means that there is a highest or perfect correlation between the two variables. For this we need to drop one of the columns to avoid multi-collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
   Answer:
   Q-Q plot also termed as Quantile-Quantile plots, is a way to determine two datasets originates from the same population data set with a common distribution.

   A q-q plot is the plotting of data sets which takes the data sets from one quantile and then the other data set falling under different quantile. Suppose if we take the one data set falling under 40% quantile or 0.4 quantile and 60% of the remaining data set falls above the value. A reference line is kept at 45 degrees.
   If the above two data sets come from the same population with same distribution, the points should fall along the reference line. The far of the points from this reference line represent how well the two sets come from different population

   Importance of q-q plot:
   When we get two sample for analysis, We assume that these data sets comes from the same distribution, to verify this assumption, q-q plot is useful. If the two samples differ it also provide information of the same. It also provide more insights on the difference of the data set.