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: The demand of bike is less in the month of spring when compared with other seasons

- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

Ans: Not using drop_first=True would make the dummy variables correlated to each other and hence, redundant, which is not expected of our analysis.

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

Ans: atemp and temp has the highest correlation with the target variable.

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

Ans: One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram.

Have validated the assumption of Linear Regression Model based on below 5 assumptions -

○Normality of error terms

- ■ Error terms should be normally distributed

○Multicollinearity check

- ■ There should be insignificant multicollinearity among variables.

○Linear relationship validation

- ■ Linearity should be visible among variables

○Homoscedasticity

- ■ There should be no visible pattern in residual values.

○Independence of residuals

- ■ No auto-correlation

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

Ans: The top 3 features directly influencing the count are the features with highest coefficients. These are: Temp, Year (positively influencing) and snowy and rainy weather (negatively influencing).

General subjective questions:

1. Explain the linear regression algorithm in detail?

Ans: An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s) is linear regression.

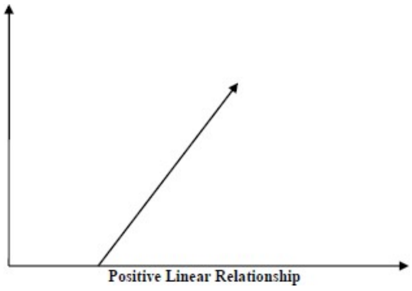
After looking into the data and cleaning it with exploratory data analysis, we split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output). After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.

According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



- Negative Linear relationship:

A linear relationship will be called positive if independent increases and dependent variable decreases.

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

✔ Multi-collinearity –

o Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

✔ Auto-correlation –

o Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

✔ Relationship between variables –

o Linear regression model assumes that the relationship between response and feature variables must be linear.

✔ Normality of error terms –

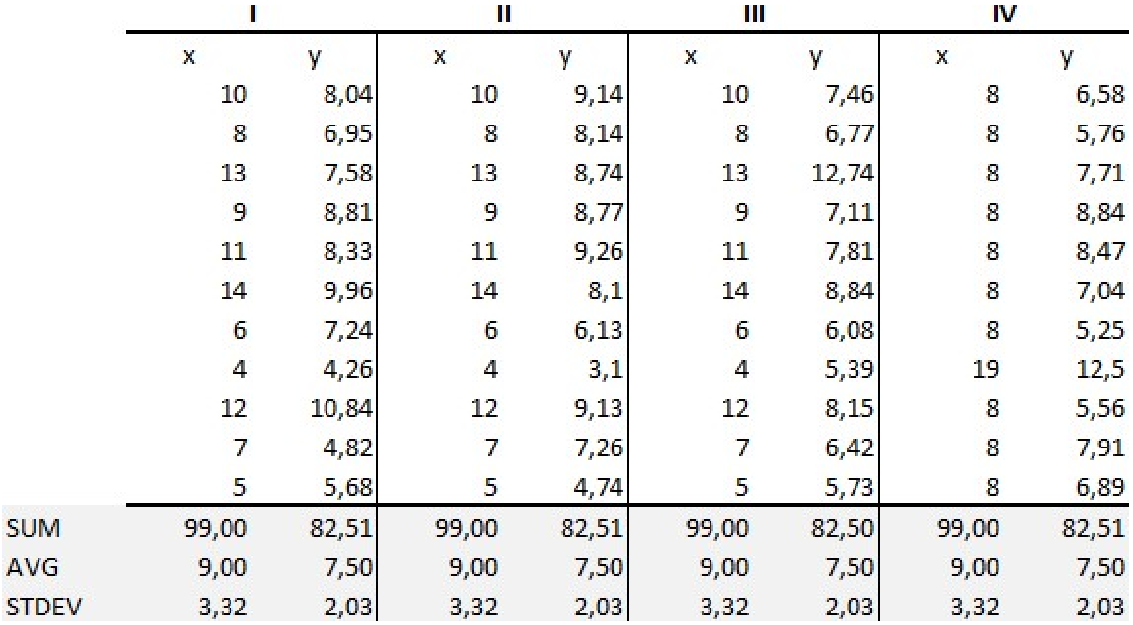
o Error terms should be normally distributed

✔ Homoscedasticity –

o There should be no visible pattern in residual values.

2. Explain the Anscombe’s quartet in detail?

Ans: Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

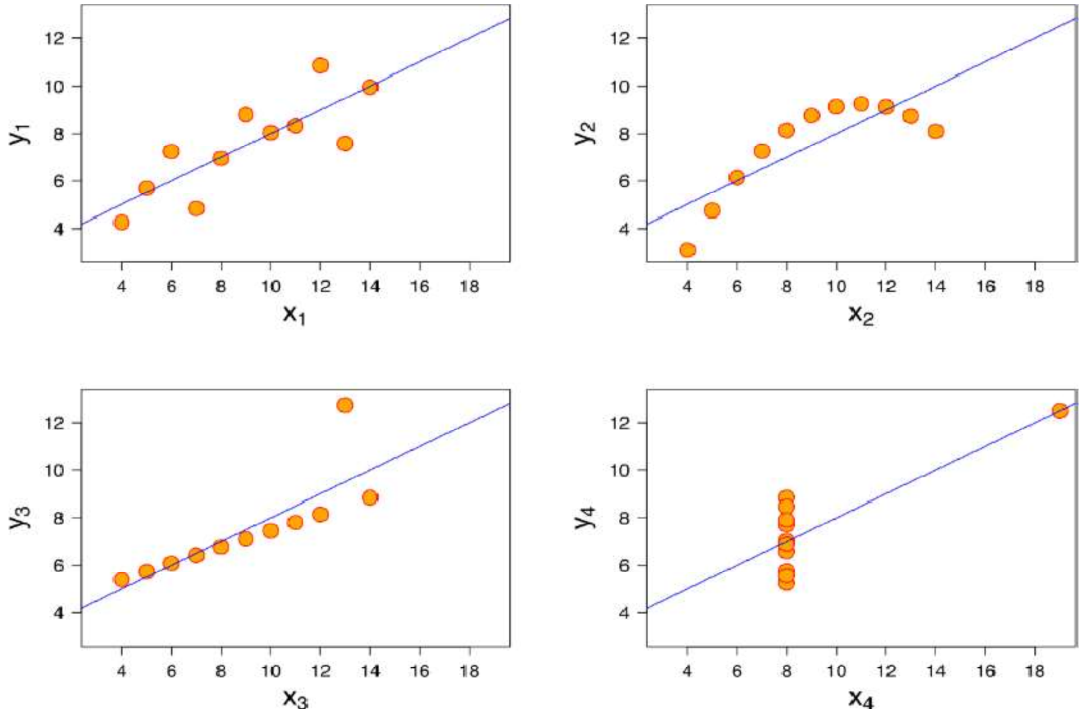
The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



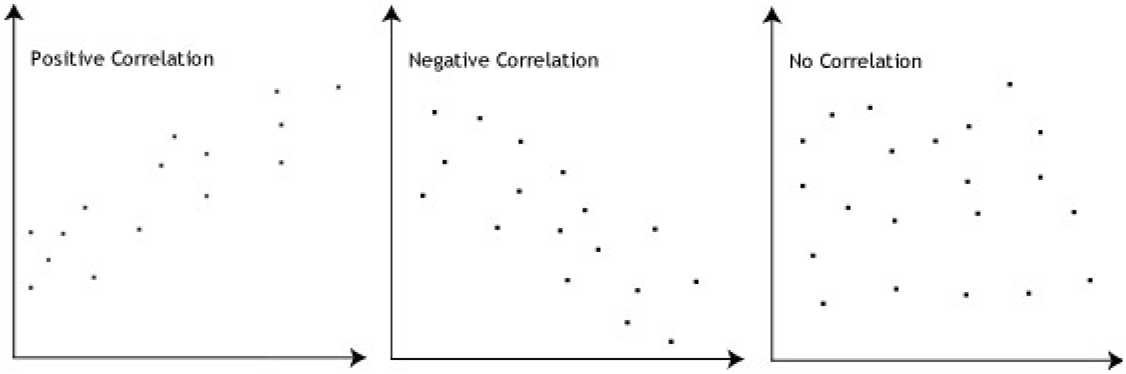
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an Outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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 nit of the values.

- Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km 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.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans: If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: quantile-quantile (Q-Q) plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. 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