# Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

→

The following inference were made about the categorical variables effect on the dependent variable

a) Fall Season has highest demand for rental bikes.
b) Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
c) More people prefer to rent bike on working days and during holiday people prefer to stay home.
d) Demand is continuously growing each weekday till 5th weekday and demand has decreased in 6th weekday.
e) Clear weather or few clouds weather attracted more booking.
f) For workingday there is slight more demand in bikes than a non workingday but noting significant can be inferred.
g) 2019 attracted more number of booking from the previous year, which shows good progress in terms of business

Q2. Why is it important to use drop_first=True during dummy variable creation?

→

During dummy variables the attribute drop_first = True is very important to use because as it helps in reducing the extra column and make the model less complex. And hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

→

Temp and atemp variable are more correlated with target variable cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

→

The assumption of Linear Regression were validated based on the following parameters:

a. Error terms are normally distributed with mean 0.
b. Error terms do not follow any patter.
c. Multicolinearity Check among variables using VIF.
d. Linearity Check.
e. Compared the R2 value and Adjusted R2 value.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

→

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes based on my final model.

a) Temp.
b) Year.
c) Winter season.

# General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

→

Linear regression is a statistical regression method used for predictive analysis and observe the relationship between the continuous variables. It shows the relationship between the independent variable (X) also known as predictor variable and the dependent variable (Y-axis) also called target variable.

Depending on the number of input variables or predictor variables, there are two types:

- **Simple linear regression**: When the number of independent variables is 1
- **Multiple linear regression**: When the number of independent variables is more than 1

Linear regression estimates the relationship between a dependent variable and an independent variable using the line equation:

**y = mx+c**

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c

The above equation could be also written as y = b1x+b0, which implies that for every rise/fall in the value of x, the y is rising or falling b1 times provided b0 is kept constant.

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. It can be understood with the help of following graph.

**The main aim of the linear regression algorithm is to get the best values for b0 and b1 to find the best fit line for the model.**

Assumptions

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

1. Multi-collinearity

- 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.

2. Auto-correlation

- 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.

3. Relationship between variables

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

4. Normality of error terms

- Error terms should be normally distributed

5. Homoscedasticity

- There should be no visible pattern in residual values

Q2. Explain the Anscombe's quartet in detail.

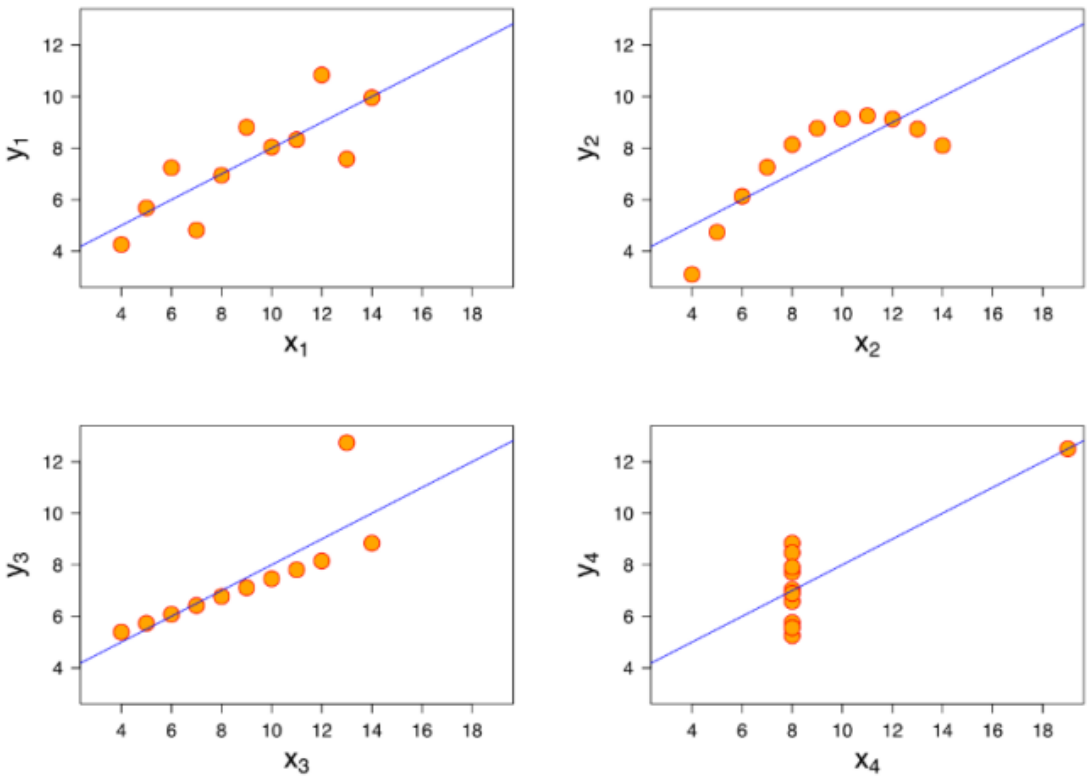
→

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



This quartet emphasizes the importance of visualization in Data Analysis. Datasets which are indistinguishable over a number of statistical properties, still produce different graphs. Sometimes statistics summary of the data are misleading on their own. So it's important to use graphical or visualization of data for larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed

## Q3. What is Pearson's R?

→

The Pearson's R referred to as Pearson's Correlation Coefficient in statistics. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between - 1.0 and +1.0.

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

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

## Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

→

Scaling is a step of data processing in data analyse using models, it is applied to independent variables to normalize the data within a particular range. It also helps in faster calculations in an algorithm.

The data collected often contains features/variables which are highly varying in magnitudes or units or range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared.

| Normalized scaling | Standardized scaling |
|---|---|
| Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation |
| Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| It is really affected by outliers. | It is much less affected by outliers. |
| Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization |

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

→

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. A large value of VIF indicates that there is a correlation between the variables.

If the VIF is infinite which implies that there is a perfect correlation between two predictor variables. It is the case of perfect correlation.

In this case we get Rsquared value equal to 1. Due to which the term 1/ (1-Rsquared) reached infinity.

For this we need to identify the feature which is causing this perfect correlation and should be dropped to get a best model.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

→

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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 for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.