# 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?

Answer: The analysis of categorical variables is done using boxplot.
Following are the inferences after the analysis:

- Overall in all variables, 2019 got more bookings as compared to 2018.
- Bookings are low during the Spring season, the rest seasons get good amount of bookings.
- The weather has a significant impact on the bookings, there are no bookings during heavy rain and Snow, however significant bookings are attracted when it's clear sky or misty.
- Month Wise bookings are seen high in count from April till October, there after due to severe weather condition booking count dips.
- Weekday does not have any impact on bookings.

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

Answer: By using drop_first=True, it drops the one variable while creating dummy variables. This helps in reducing issues related to multicollinearity.

For ex. If a column Color has 3 values,  Red, blue, and Black. When we create the dummy variable on this column, we will get two new columns Color_red and Color_Blue. When the boolean value of Color_red and Color_Blue is 0 then it is obvious that value is Black.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: The ‘temp’ variable has the highest correlation with the target variable.

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

Answer: The following steps are performed while validating the Linear Regression model

- Multicollinearity check - There should not be significant multicollinearity among variables.
- Error terms should be normally distributed
- Linear relationships should be visible in the plot.
- Homoscedasticity - No visible pattern in the residual values.
- Independence of residuals - no auto-correlations

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

Answer: The following are the top 3 features of the model:

1. Year
2. Winter
3. september

## General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

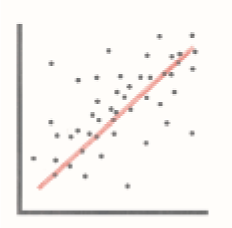
Answer:
Linear regression is defined as statistical model that analyses the linear relationship between a dependent variable for a given set of independent variables. This means that the value of one variable changes then the dependent variables value will also get change with respect to independent variable. This change can be positive or negative.

Linear quation which explain the linear regression:

Y = mX + C

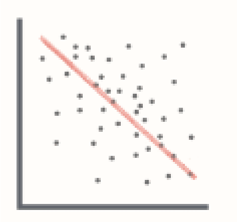
Where
Y = dependent variable
m = slope of the regression
X = independent variable
c= constant

A positive linear relation wil have Y and X as positive relation, which means if independent variable X increases then Y will also increase.



Positive Correlation

On the other hand a negative linear relation will have Y and X as negative relation, which means if independent variable X increases then Y will decrease.



Negative Correlation

There two types of Linear Regression:

1. Simple Linear Regression - this has only one independent variable
2. Multiple linear regression - this has more than one independent variable.

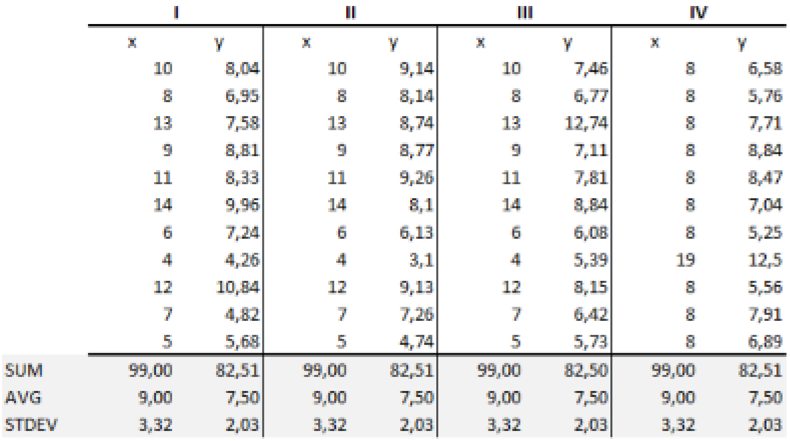
$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$

Where :
B are the coefficient of the independent variables.

Q2. Explain the Anscombe's quartet in detail.

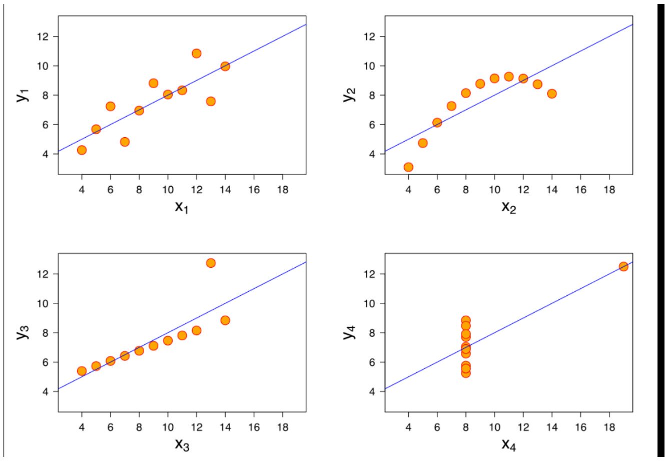
Answer:
**Anscombe's quartet** comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted in the graph. Each dataset consists of eleven (x,y) points.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



| | 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 we same for x and y across the groups:



- Mean of x is 9 and mean of y is 7.5 for each dataset.
- The variance of x is 11 and variance of y is 4.13 for each data set.
- The correlation coefficient between x and y is 0.816 for each dataset.

However while plotting on graph we observe that they show same regression line but different plots.

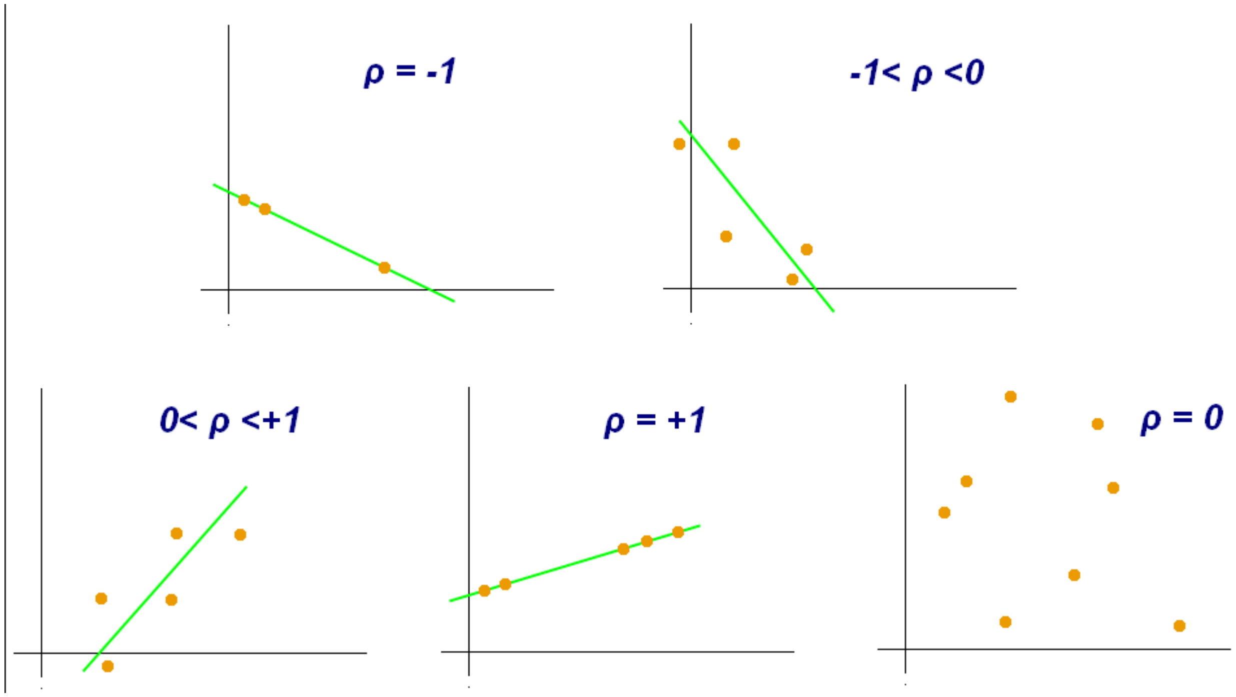
Q3. What is Pearson’s R?

Answer :

Pearson’s R is correlation coefficient, which tells about 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 the opposite direction then the correlation coefficient will be negative.

The Pearson’s correlation coefficient, r, can take a range of values from +1 to -1.
0 - means no relationship between two variables
Between 0 and 1 - positive correlation
Between -1 and 0 - negative correlation.

It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).



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

Answer:
Scaling - It is a process of tramsforming data to fit whitin a specific range.Since in linear regression different feature contribute proportionally to the model’s training process, if scaling is not conducted then feature with larger values could dominate th learning process which may result in the wrong prediction.

Scaling is performed for some of the reasons mentioned below:

- □ Model performance - Many ML algorithms perform better when features are on same scale.
- □ Convergence - Gradient descent optimization algorithm converge faster with scaled features.
- □ Equal distribution - This also ensures that all feature contribute equal to the model predictions and not dominate a few hence reducing bias prediction.
- □ Easy interpretation

Difference between normalized scaling and standardized scalling:
Normalized scaling, also known as Min-Max scaling scaled te data to a fixed range usually between 0 and 1.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized scaling:
Sandardization: The variables are scalled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

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-square, etc.

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

Answer:
VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.
A VIF value becomes infinite when R2=1. This situation arises when there is a perfect multicollinearity among the predictors. There is perfect linear combination of one or more of the other predictor variables.

Reasons for the multicollinearity:

1. Duplicate variables
2. Linearly dependent variable, instead of independent variables which is need for linear regression.

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

Answer:
A Q-Q plot is quantile - quantile plot is a probability plot, and a graphical method for comparing two probability distribution by plotting their quantiles against eash other.

Use and importane of Q-Q plot:

1. In linear regression the assumption is residuals are normally distributed. A Q-Q plot can help to explain this assumption and prove if the assumption is correction or not.

2. QQ plot help to detect the outliers or deviations from normality.Points that deviate significantly from the straight line suggest that the residuals are not not normally distributed indicating outliers or skewness in data.
3. By ensuring residuals are normally distributed, it confirms that the model is a good fit for the data.

Interpretation of QQ plot:
Straight line- if the points on the QQ plot lie close to the straight line, it indicates that the residuals are approximately normally distributed.
S-Shaped curve - it indicates data is skewed.
Upward or downward curve : if suggests that the data has heavier or lighter tails than the normal distribution.