# Assignment-based Subjective Questions and Answers

**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: --**

From my analysis on categorical variables through box plot I inferred,

- In "fall season" we can see demand and bookings were increased compared to remaining 3 seasons.
- Compared to the year 2018, year 2019 has more bookings. As it was started in 2018 the business got expanded and increased.
- In months we can see a curve type structure where the bookings are increasing in May, June, July, August, September and October. In all the months we can see more bookings in September and least in January.
- During holidays bookings were less. So, public may have used the transport for college or offices etc.
- In the last weekdays we can see a greater number of bookings.
- On working days bookings are a bit more than non-working days.
- In clear weather bookings are more and it's obvious.

**2) Why is it important to use drop_first=True during dummy variable creation? (2 marks)**

**Answer: --**

During dummy variable creation, 'n' categorical levels are created for 'n' columns. When we use **drop_first=True,** the levels will reduce to 'n-1' means one column is dropped from the dataset.

Initially in default "drop_first = False".

This may also reduce the correlations created among dummy variables.

For example: Suppose we have A, B, C as categorical variables

We can represent all the 3 with just 2 levels as

| / | B | C |
|---|---|---|
| A | 0 | 0 |
| B | 1 | 0 |
| C | 0 | 1 |

Similarly, we can use the same method for 'n' number of categorical columns to create 'n-1' levels.

**Syntax:**

Status = pd.get_dummies(['column 1', 'column 2',……., 'column n'] , drop_first = True)

**3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

**Answer: --**

The ‘temp’ and ‘atemp’ variables have high correlation with ‘0.627044’ and ‘0.630685’.

But as we are dropping ‘atemp’ we consider **‘temp’** as highly correlated variable with target.

**4) How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

**Answer: --**

- All the VIF’s of the variables are <5 only.
- P-values of the coefficients are zero. Hence, Significant.
- F-Statistics are high. Hence, Significant.
- Here we can reject null hypothesis.
- Multicollinearity is very low.
- Linear relationship among the variables.
- No auto-correlation.
- No visible pattern in residual values.
- The residual terms are normally distributed with mean zero.

Hence, with this we can validate the assumptions of Linear Regression for the model.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

**Answer: --**

1) Temp :- As coefficient value of temp is very highly positive i.e '0.470173'. It play's a vital role in prediction.
2) weathersit_Rain :- As coefficient value of weathersit_Rain is highly negative i.e '-0.299970'. As we know during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds people on roads are less. It is also important for prediction.
3) yr :- As it shows positive coefficient i.e '0.233713' it indicates the business in a profitable state compared to the previous year.

# General Subjective Questions

**1) Explain the linear regression algorithm in detail. (4 marks)**

**Answer: --**

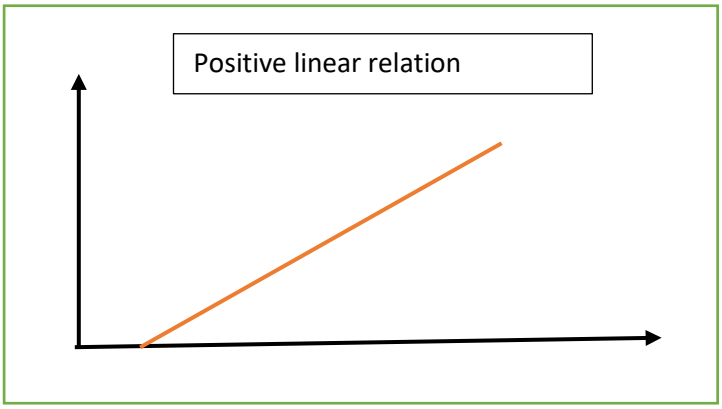
Linear regression is defined as the statistical model that analyses the linear relationship between dependent variables with one or more independent variables.

Here, linear relationship means when one or more independent variable value change either in positively or negatively i.e either increase or decrease, the dependent variable value may also change w.r.t to the independent variables.

Linear relationship may be positive and negative in nature.

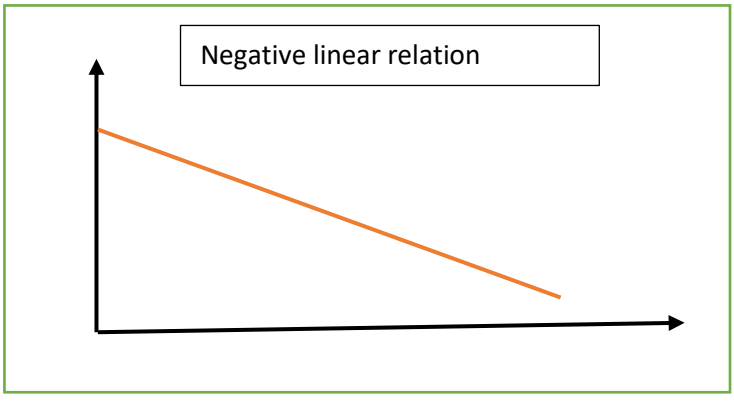
Positive Linear Relationship:

When dependent variable increases with increase in the independent variable it is know as positive linear relationship.




Negative Linear Relationship:

When dependent variable decreases with increase in the independent variable it is know as positive linear relationship

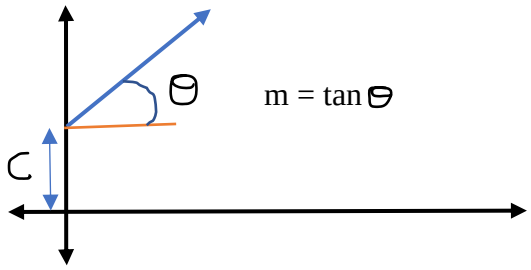


There are generally 2 types of linear regressions

1) Simple Linear Regression: In this type of regression where the dependent variable is depended on only one independent variable. It is given by

$$Y = m X + C$$

Graphically,




Here,

C = It is a constant called Y- Intercept. It is said to be the value of Y when X = 0

M = slope = tan$\theta$ = dY/dX

X = Independent variable used to make predictions

Y = Dependent variable.

2) Multiple Linear Regression : In this type of regression where the dependent variable is depended on more than one independent variable. It is given by

$$Y = b0 + b1\ X1 + b2\ X2 + \text{---------} + bn\ Xn$$

Here,

X1, X2, ----, Xn = These are independent variables

Y = Dependent variable

b0 = constant

b1, b2, ---, bn = Respective coefficients of independent variables

**Assumptions in linear regression model:**

- All the VIF's of the variables <5 are recommended.
- P-values of the coefficients should be as low as possible to become significant.
- F-Statistics should be as high as possible to become significant.
- Multicollinearity must be very low as possible.
- There must be linear relationship among the variables.
- No auto-correlation should be present.
- No visible pattern in residual values should present.
- The residual terms should be normally distributed with mean zero.

**2) Explain the Anscombe's quartet in detail. (3 marks)**

**Answer: --**

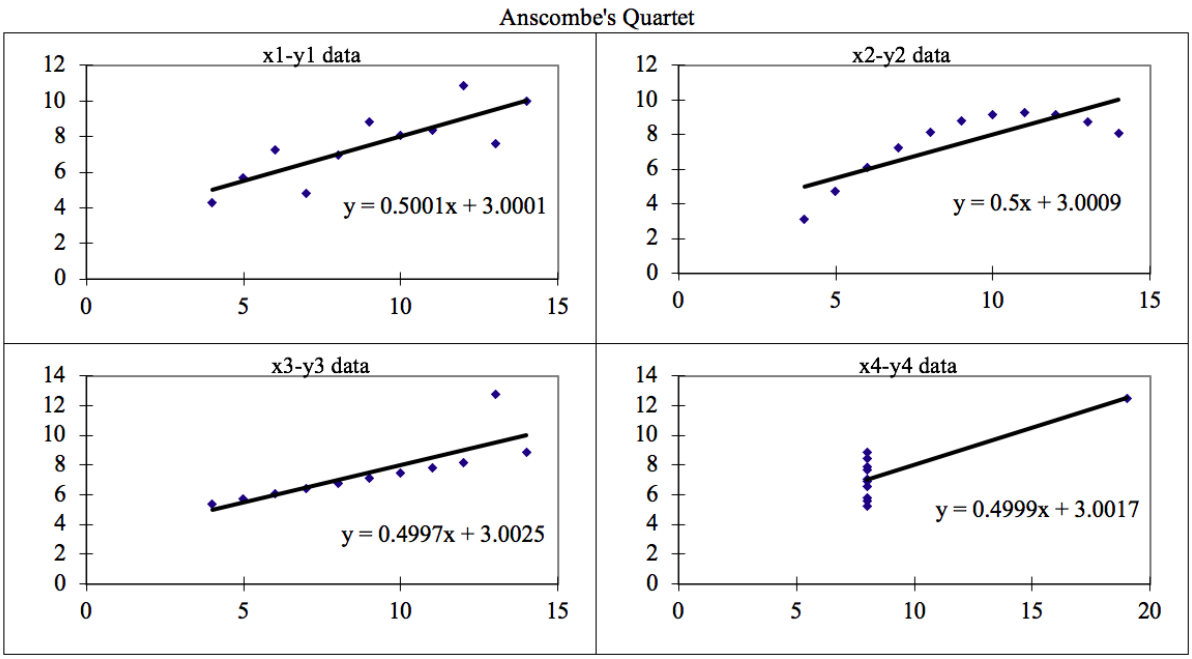
Anscombe's Quartet can be defined as a group of four data sets containing eleven (x, y) pairs which are nearly identical in simple descriptive statistics**,** such as mean, average, standard deviation, variance and etc.

Though they have same statistics when they are plotted in scatter plot, all the data sets show different distributions and appear differently.

This tells us the importance of the visualising data through plots.

The different data sets are shown below: --

| 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 | |




Dataset 1: This linear regression model is well fitting.

Dataset 2: This will not fit as it is not distributed normally well as the data is non-linear.

Dataset 3: This shows the linear model, but outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: Here there is one outlier which may produce the high correlation coefficient and where the outliers involved in the dataset which cannot be handled by linear regression model

**3) What is Pearson's R? (3 marks)**

**Answer: --**

The Pearson correlation coefficient describes the strength and direction of the linear relationship between two quantitative variables.

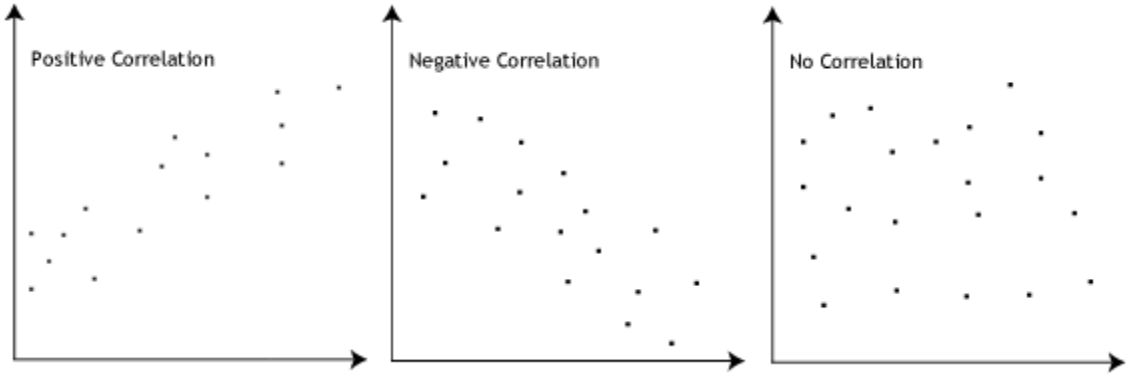
If the variable tend to go up and down together, the correlation coefficient is positive.

If the variable tend to go up and down in opposite that is when one variable is at low value other variable at high value then the correlation coefficient is negative.

The Pearson's correlation coefficient varies between -1 and +1.

When,

- r = 1 means the data is perfectly linear with a positive slope. This means both variables tend to change in the same direction.
- r = -1 means the data is perfectly linear with a negative slope. This meas both variables tend to change in different directions.
- r = 0 means there is no linear association.
- r > 0 < 5 means there is a weak association.
- r > 5 < 8 means there is a moderate association.
- r > 8 means there is a strong association.




**4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

**Answer: --**

Scaling is a process which is applied to normalize the data within a particular range.

Suppose in a data set we have many number of variables which are varying with magnitude, units, range etc. Where the units of the variables may differ, if we create a model using that values we may end up with false predictions. So, here we apply scaling and bring all the values to the same level of magnitude to get correct models.

Scaling will only effect the correlation coefficients but not other parameters like P-Value, T-Statistics, F-Statistics, R-Squared, Adjusted R-squared etc.

For Example: When there are 2 variables with values as 2Kg and 400grams, without scaling algorithm may consider 400grams as greater value which is incorrect. So scaling is important.

**Difference between normalized scaling and standardized scaling: --**

| 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. | Effected with outliers | Not effected with outliers |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 6. | It is useful when we don't know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |

**5) You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

**Answer: --**

When VIF is high, this means that variable is associated with another variable. If VIF is low, that means that variable is not really related to other variables. So, by this we can say that this happens when correlation is perfect i.e VIF = Infinite.

If VIF = 3, this means that the variance of the model coefficient is inflated by the factor of 3 due to the presence of multicollinearity

Generally preferred VIF is <5 which is good and no need to eliminate that particular variable from the data set.

To resolve this issue we drop one of the variables from the data set which is causing this perfect multicollinearity.

From the formula of VIF i.e

$$VIF = 1/ (1-R^2)$$

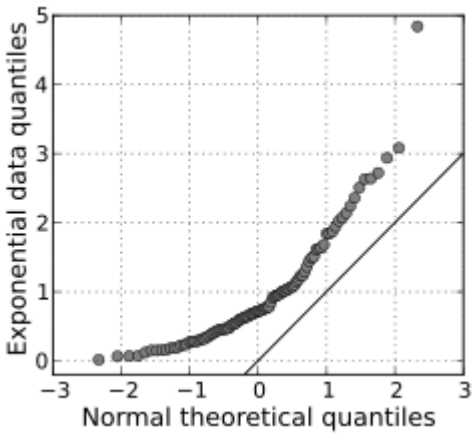
Here, R = correlation coefficient

If, r = 1 then VIF = 1/ 0 = Infinite.

**6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

**Answers: --**

Q-Q plots means Quantile-Quantile plots. These are the plots of two quantiles against each other. For example, the median is the quantile where 50% of the data is below that point and 50% of the data is above that point**.** The purpose of the Q-Q chart is to see if the two datasets are from the same distribution. A 45-degree angle is plotted on the QQ plot**.** If the two datasets are from a common distribution, the point is on this reference line.




If the two distributions being compared are similar, the points on the Q-Q plot are approximately on the y = x line**.** If the distributions are linearly related, the points on the Q-Q plot are almost on the line, but not necessarily on the y = x line. Q-Q plots can also be used as a graphical means of estimating the parameters of a family of site-scale distributions.

When there are 2 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 2 samples do differ, it is also useful to gain some understanding of the differences.