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?**

Answer:
Below are few inferences–

- The majority of reservations were made during the months of May, June, July, August, September, and October. The booking trend saw an increase from the beginning of the year until mid-year, after which it began to decline towards the year's end.
- The fall season experienced the highest booking demand, with winter and summer closely following, whereas spring witnessed significantly lower demand compared to other seasons
- There is a strong correlation between months and seasons. Clear weather conditions tend to attract more bookings.
- The latter part of the week, specifically Thursday, Friday, Saturday, and Sunday, sees a higher volume of bookings compared to the earlier days.
- Bookings tend to dip during holidays.
- The year 2019 observed an uptick in demand across all categories when compared to the previous year.
- Additionally, the demand appears to be independent of whether it is a working or non-working day.

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

Answer:
Dummy variables have high multicollinearity. This is because their values are mutually exclusive. Setting drop_first = True reduces the number of variables by one hence reducing the correlations created among dummy variables. (not eliminating though)

Example: If we have 3 distinct values (say A, B, C) in a Categorical variable and we want to create dummy variable for that column. If we just have two dummy variables A and B. Below table illustrate it.

OriginalVariable	DummyA	DummyB
A	1	0
B	0	1
C	0	0

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

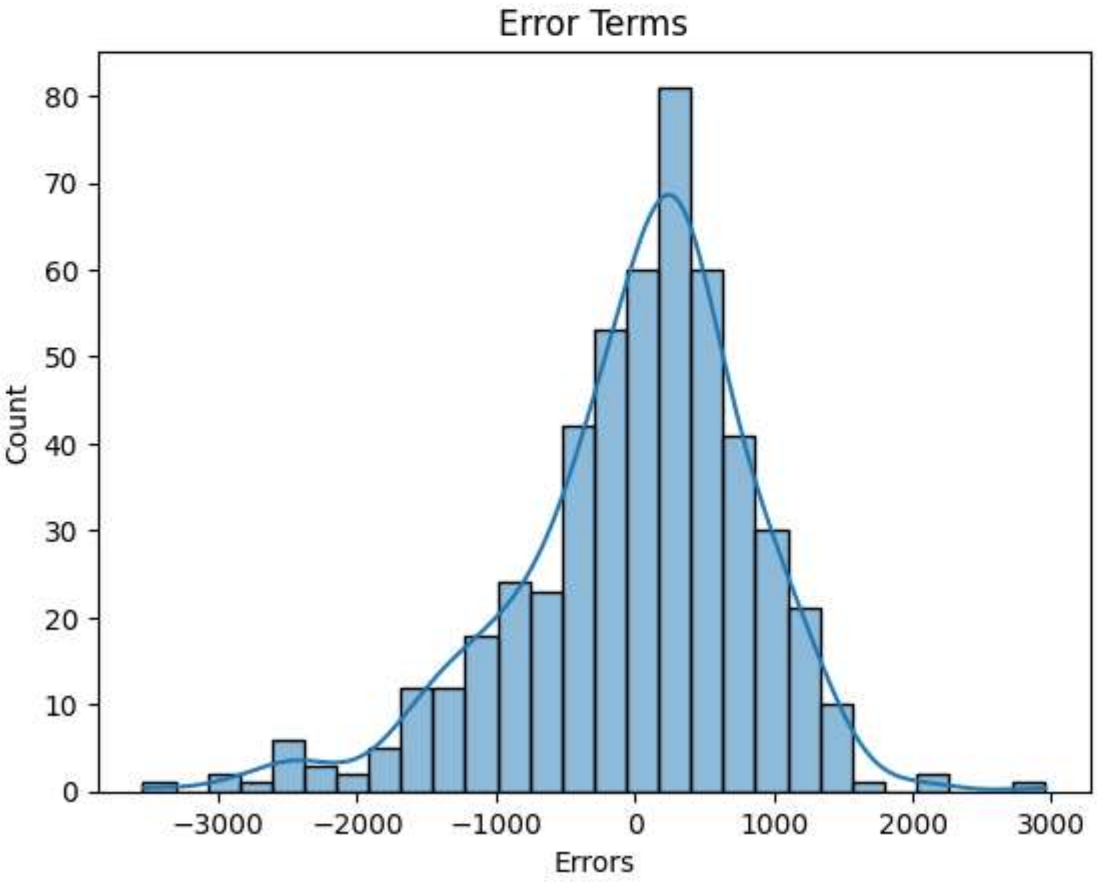
Answer:
‘temp’ variable has the highest correlation with the target variable.

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

Answer:
Below are details of validations used for linear regression:

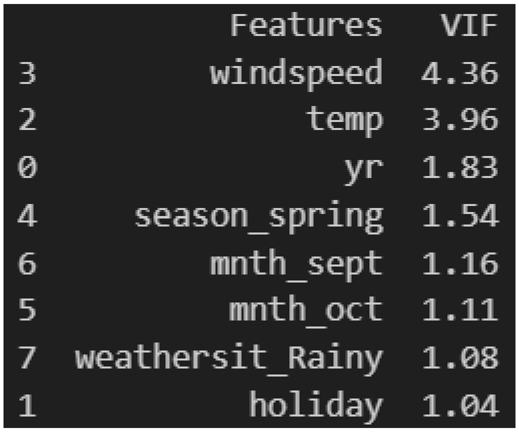
- Normality of error terms

Use seaborn to plot distribution plot of error in prediction at every term.
Verified that it is normally distributed.



- Multicollinearity check

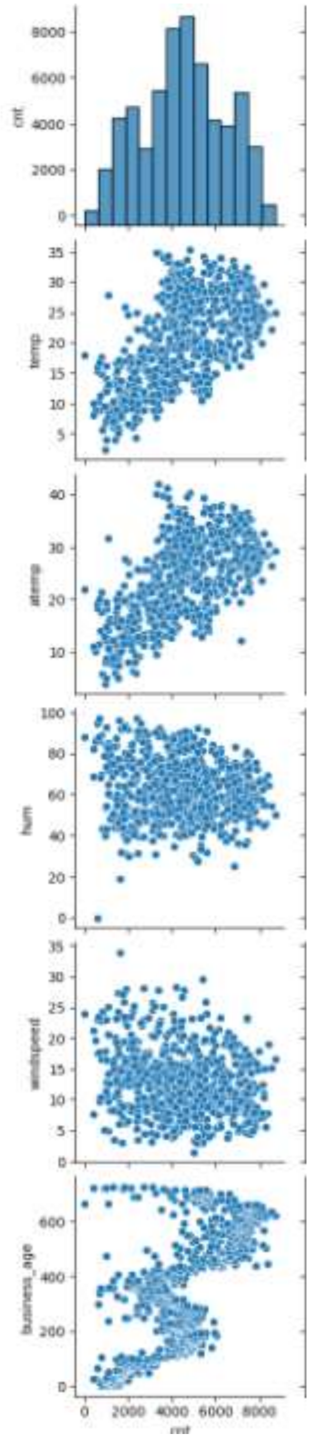
Used variance_inflation_factor method from statsmodels to calculate VIF for all features present in model and verified that they are all less than 5.



```
          Features   VIF
3        windspeed  4.36
2             temp  3.96
0               yr  1.83
4    season_spring  1.54
6        mnth_sept  1.16
5         mnth_oct  1.11
7  weathersit_Rainy  1.08
1          holiday  1.04
```

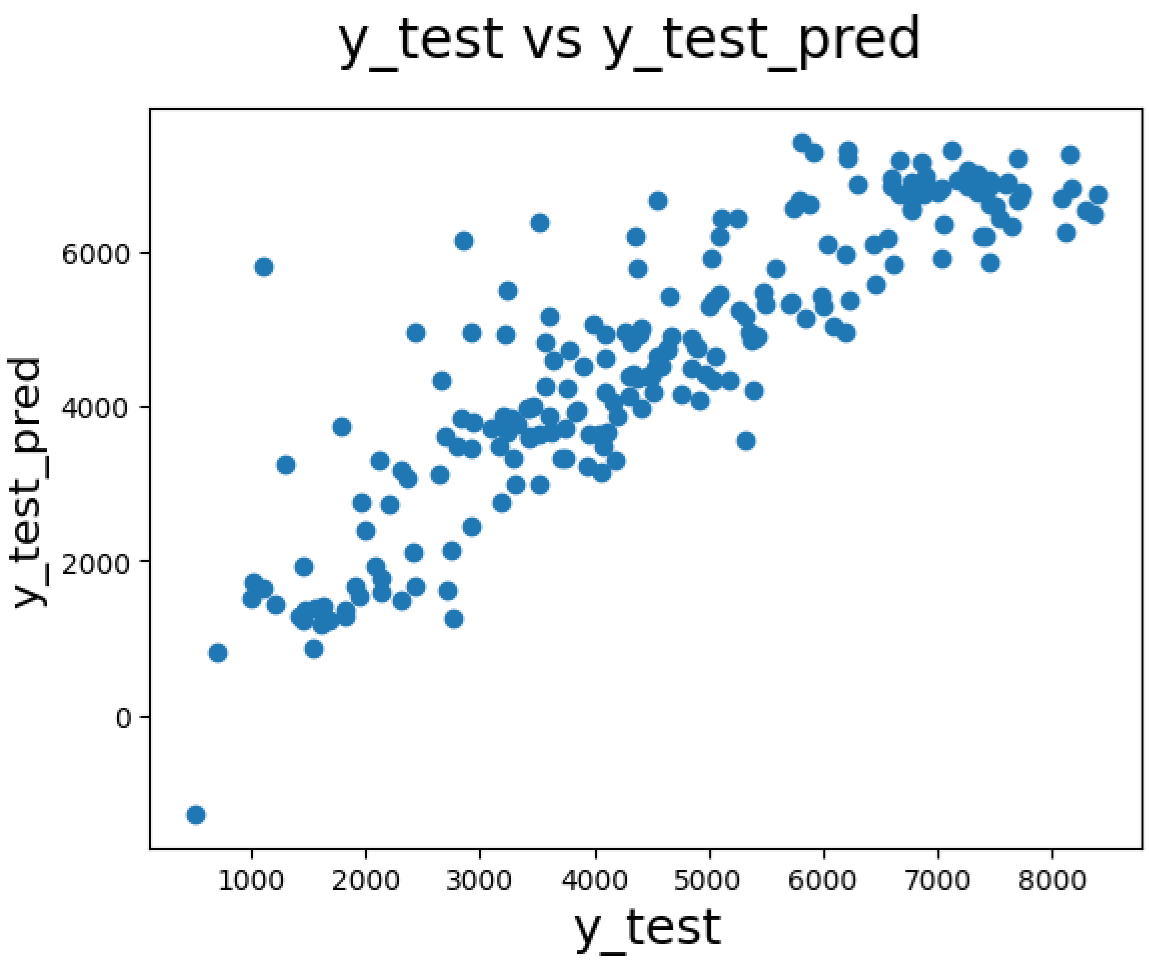
- Linear relationship validation

Validated linear relationship between some of independent variables with dependent variable.



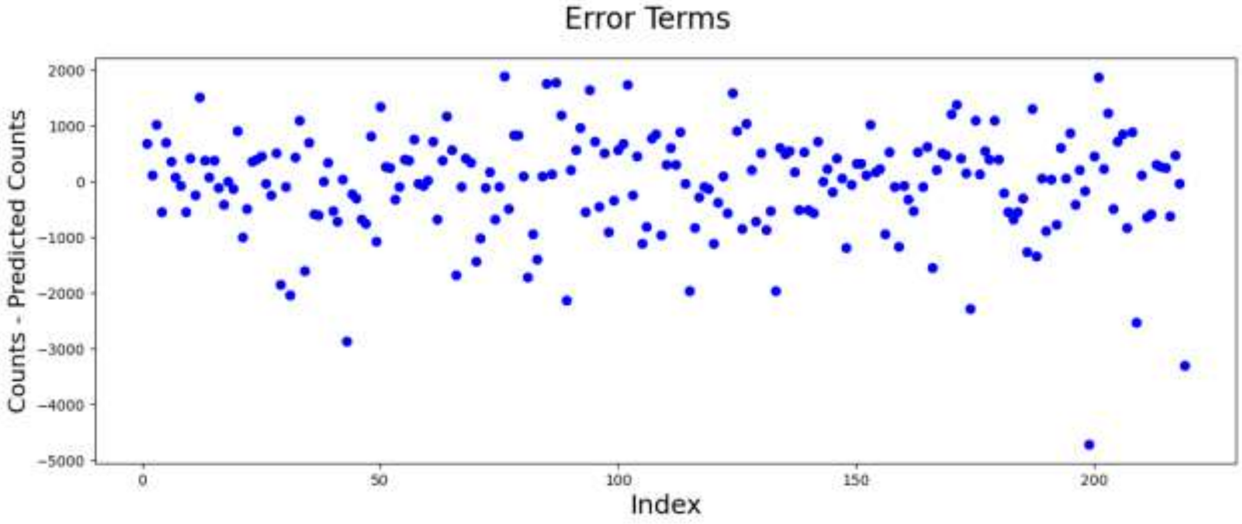
- Homoscedasticity

Plotted y_test and y_test_pred to understand the spread. Validated that it is linearly correlated.



- Independence of residuals

Plotted residual in sequence of data points to validate that there is no pattern.



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

Answer:

Looking at corresponding coefficients, below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- weathersit = 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is predicated on the assumption of linear relationships between the dependent variable and one or more independent variables. Essentially, the algorithm iteratively estimates the coefficients and intercepts of the linear equations to minimize the overall prediction error.

Mathematically the relationship can be represented with the help of following equation −
There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y=\beta0+\beta1X$y=β0+β1X
where: Y is the dependent variable

X is the independent variable

β0 is the intercept

β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$y=\beta0+\beta1X+\beta2X+\ldots\ldots\ldots\beta nX$y=β0+β1X+β2X+………βnX
where:

Y is the dependent variable, X1, X2, …, Xp are the independent variables

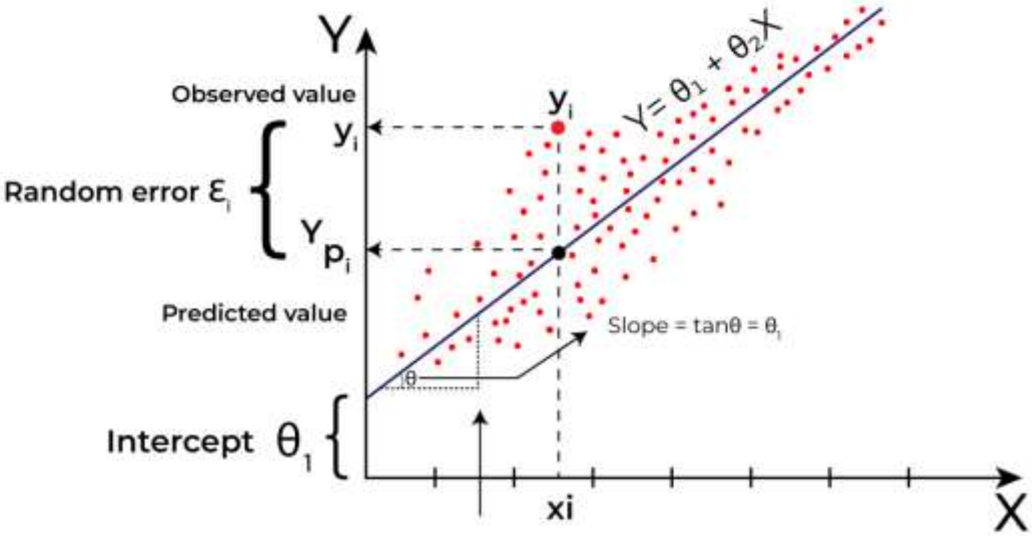
β0 is the intercept

β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Within regression analysis, one works with a dataset that includes an independent variable X and a dependent variable Y. The aim is to find a function capable of predicting Y's value for a new, unseen X. This process involves identifying a function that can predict a continuous outcome of Y given X as the predictor.

The formulation for the line of best fit depicts a linear correlation between the dependent and independent variables. The slope of the line signifies the extent of variation in the dependent variable with each single-unit alteration in the independent variable(s).


Y
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
Slope = tanθ = θ_1
θ
Intercept θ_1
xi
X

Linear Regression

In this scenario, Y is known as the dependent or target variable, and X represents the independent variable, also referred to as the predictor of Y. There are several functions or models that can be applied to regression analysis. The simplest form is often a linear function. In such a context, X may be a singular feature or a combination of features defining the problem.

The purpose of linear regression is to predict the value of a dependent variable (y) using an independent variable (x), which is why it's called Linear Regression. As shown in the figure provided, X (input) denotes the work experience while Y (output) stands for the individuals salary. The best-fit line on the graph is known as the regression line and represents our model's prediction.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values
Sum of Residuals/Errors = Sum(Actual- Predicted Values)
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e

Rsq, AdjRsq, MSE,RMSE,MAE – 5 evaluation metrics

Deep dive to R Squared (R2) approach:

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

As our regression line moves towards perfection, R2 score move towards one. And the model performance improves.

The normal case is when the R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

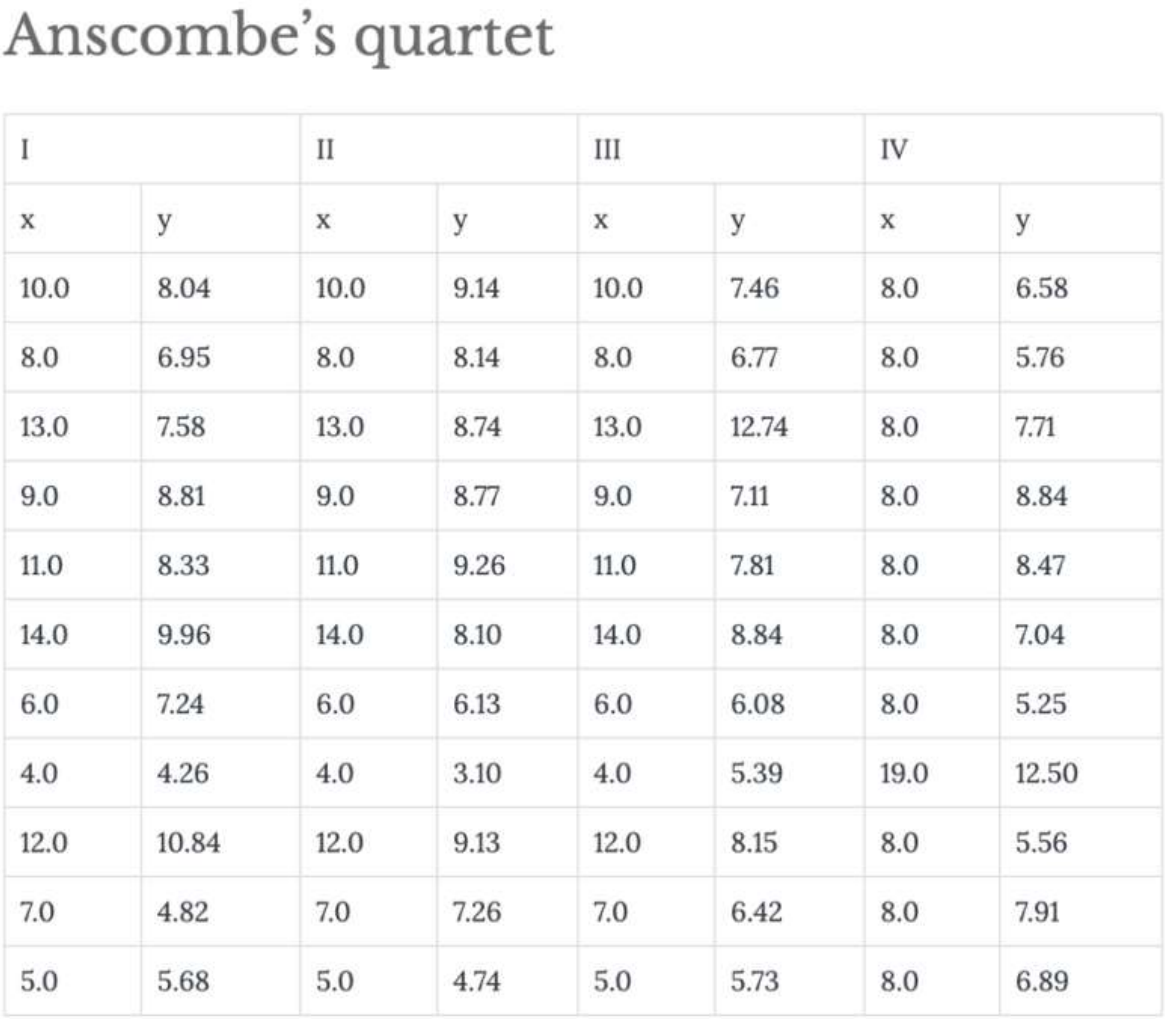
```
from sklearn.metrics import r2_score

r2 = r2_score(y_test,y_pred)

print(r2)
```

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet comprises four different datasets that have identical statistical properties—mean, variance, correlation, and regression lines—but show varied patterns when plotted. Introduced by Francis Anscombe in 1973, these datasets demonstrate the importance of visualizing data along with numerical analysis, highlighting that relying only on statistical summaries can be misleading. Each dataset in the quartet has 11 x-y point pairs, with their scatter plots revealing distinct trends and degrees of correlation between the x and y variables. Despite their graphical differences, they all produce the same summary statistics, with matching means, variances, correlation coefficients, and lines of best fit.



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Purpose of Anscombe's Quartet

Anscombe's Quartet serves as a compelling example of why exploratory data analysis is crucial and the pitfalls of relying solely on summary statistics. It further accentuates the significance of employing data visualization techniques to identify patterns, anomalies, and other essential aspects that may not be immediately apparent through summary statistics.

3. What is Pearson’s R?

The Pearson correlation coefficient (r) is widely used to assess a linear relationship, providing a value from –1 to 1 to gauge the intensity and trend of the link between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Spending on Marketing and Revenue. More we spend on marketing more we get more revenue.
0	No correlation	There is **no relationship** between the variables.	House price & Exterior color of house.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Limitation of Pearson correlation coefficient:

The Pearson correlation coefficient (*r*) *is one of the many correlation measures available for identifying a correlation. It's appropriate to use the Pearson correlation coefficient when each of the following conditions is met:*

- *Both variables under consideration are quantitative.*
- *The variables follow a normal distribution.*
- *There are no outliers present in the data.*
- *There exists a linear relationship between the variables.*

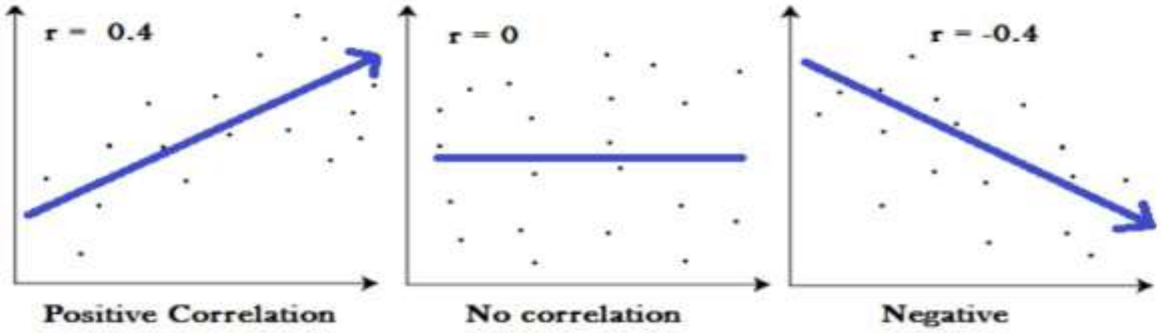
Calculating the Pearson correlation coefficient

Below is a formula for calculating the Pearson correlation coefficient (*r*):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Correlation coefficient calculations determine the strength of a relationship between datasets, yielding a value from -1 to 1:

- 1 denotes a strong positive correlation.
- -1 denotes a strong negative correlation.
- Zero suggests no correlation whatsoever.



Graphs showing a correlation of -1, 0 and +1

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

Scaling, or data normalization, is a method used to standardize the range of features in datasets. This is crucial for preparing data for machine learning algorithms because it adjusts diverse value ranges to fit within a certain scale, usually [0, 1], ensuring each feature contributes equally to the model's calculations. Without scaling, the raw magnitude of features could skew the algorithm's performance, especially since many algorithms use Euclidean distance, which would be influenced by different units and scales.

The difference between normalized scaling and standardized scaling is as follows:

- Normalization adjusts feature values into a [0, 1] range by assigning the minimum and maximum values to 0 and 1, respectively.
- Standardization changes data so that its mean is 0 and standard deviation is 1, without a fixed range.
- Outliers impact normalization more, but standardization resists such effects.
- They differ in that normalization depends on the variable's minimum and maximum, whereas standardization relies on its mean and variance.

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

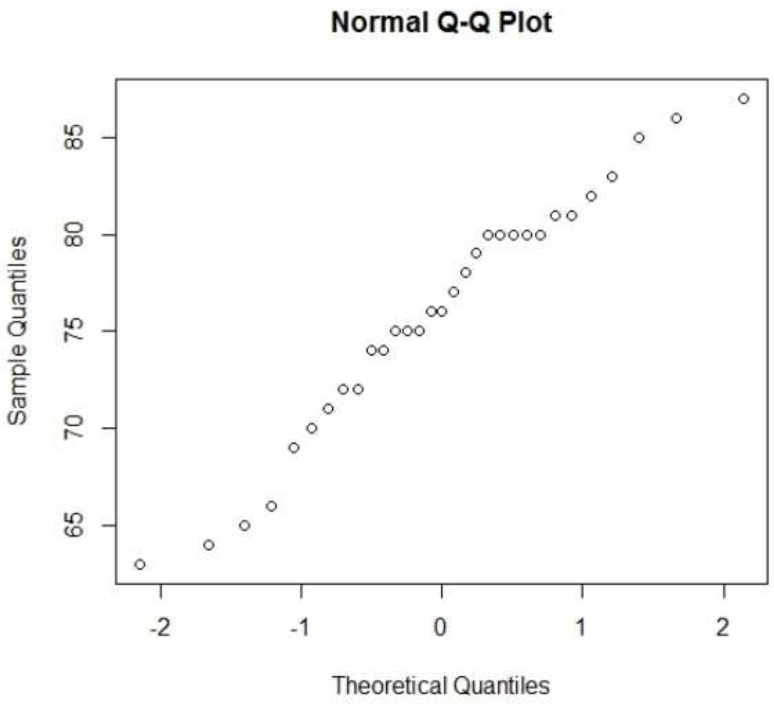
The R-squared statistic measures how much a predictor variable correlates with other predictors in a linear regression model by running a regression of the selected predictor against all the others. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

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

A quantile-quantile plot, commonly referred to as a Q-Q plot, is a graphical tool used to assess if a dataset is likely derived from a particular theoretical distribution like the Normal or exponential distribution. For example, when an analysis hinges on the normality of the dependent variable, a Normal Q-Q plot can help check this assumption. Although it's not conclusive evidence but rather a subjective gauge, it offers a quick visual assessment of whether our assumption holds water. When an assumption is found to be flawed, the Q-Q plot can highlight how and where our data deviates, pinpointing the specific data points at fault. In essence, the Q-Q plot is a scatterplot made by plotting pairs of quantiles from two different sets against one another. If both are sourced from similar distributions, the points should fall around a straight line. Take a Normal Q-Q plot, for instance; if both sets of quantiles come from Normal distributions, it can suggest if residuals follow a normal distribution, with an ideal alignment being a straight dashed line, whereas notable deviations would indicate that residuals might not be normally distributed.



Q-Q plots arrange your data in order and compare it to expected quantiles from a theoretical distribution adjusted for sample size. Though often based on the normal distribution, any distribution can be used. They help confirm if a linear regression model's residuals fit a

Gaussian distribution; misalignment indicates non-normality, which can invalidate standard statistical inferences for small samples.