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?

(3 marks)

Answer:

I have done analysis on categorical columns using the boxplot and bar plot.

Below are the few points we can infer from the visualization –

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend some quality time at home and enjoy with family and friends.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

(2 mark)

Answer:

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax -

drop_first : bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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 A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

(1 mark)

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? *(3 marks)*

Answer:

I 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?

(2 marks)

Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the
shared bikes –

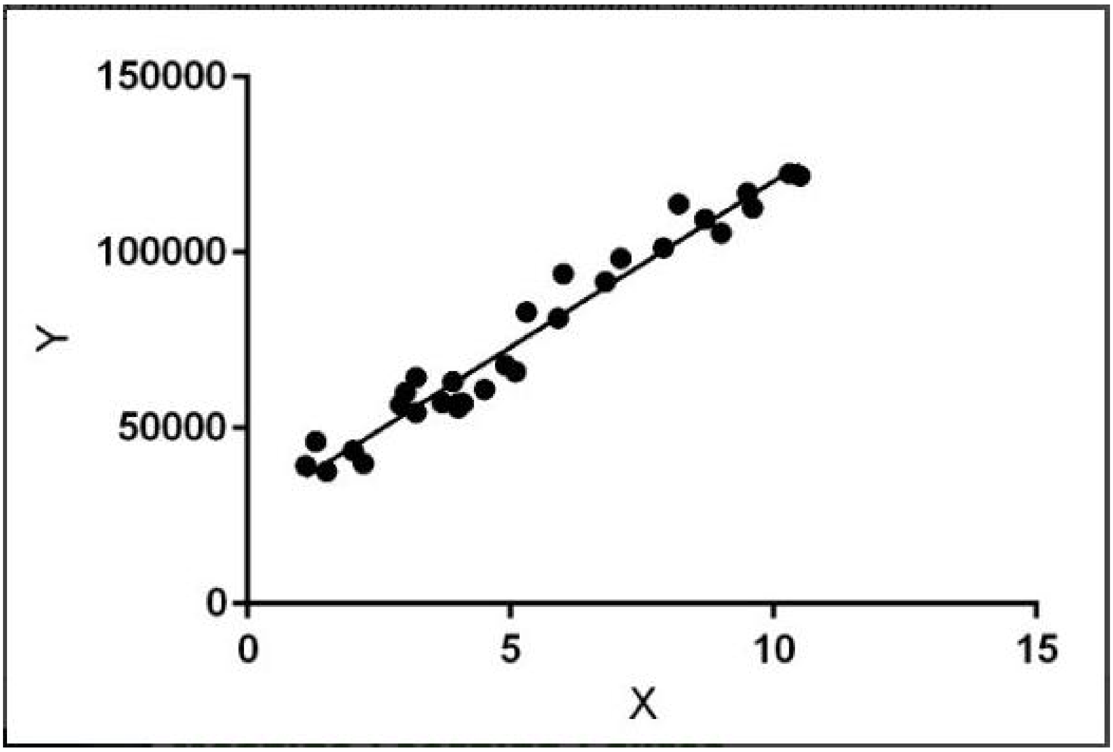
✓ temp
✓ winter
✓ sep

General Subjective Questions

Q1.} Explain the linear regression algorithm in detail.

Ans:-

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Mathematically, we can represent a linear regression as:

$y= a_0+a_1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a

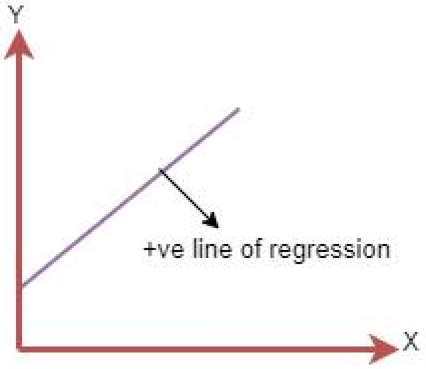
numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

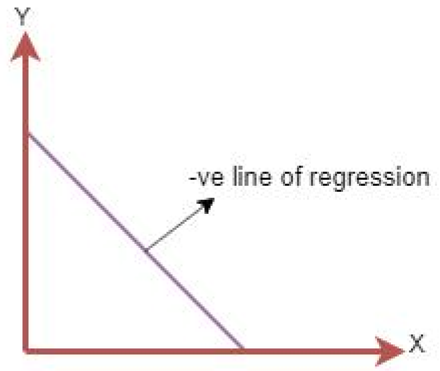
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: **Y= a_0+a_1x**

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: **$Y= -a_0+a_1x$**

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

- Multi-collinearity-

✓ Linear regression Linear regresssion model assumes is 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 -

✓ 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 dependencybetween residual errors.

- Relationship between variables -

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

- Normality of error terms -

✓ Error terms should be normally ditributed.

- Homoscedasticity -

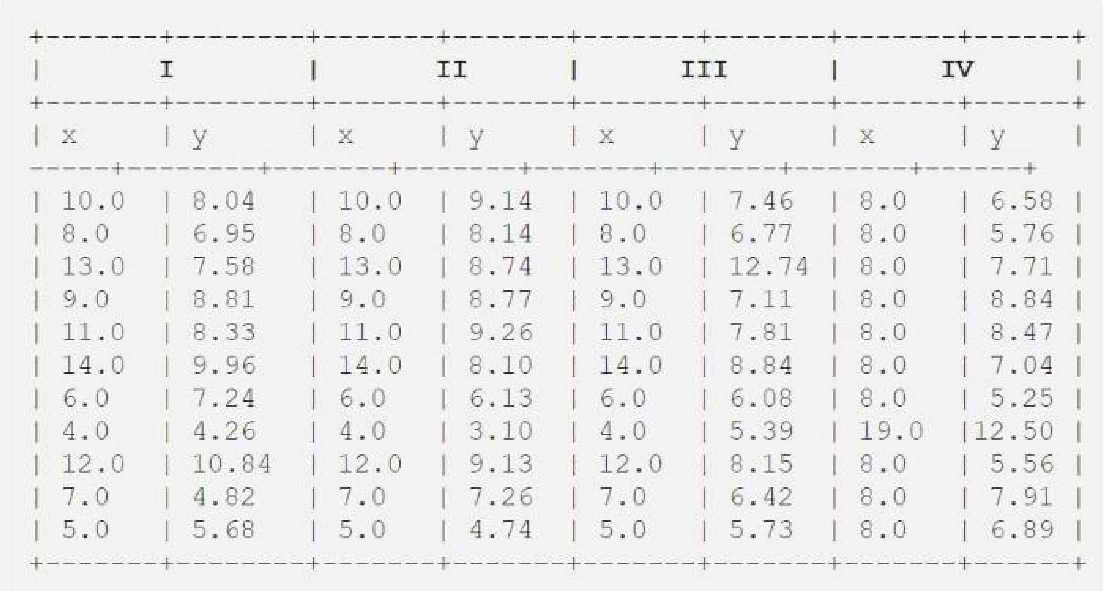
✓ There should be no visible pattern in residual values.

Q2} . Explain the Anscombe's quartet in detail. *(3 marks)*
Answer:

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

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 before analyzing it and the effect of outliers on statistical properties.



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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

Variance of x = 11

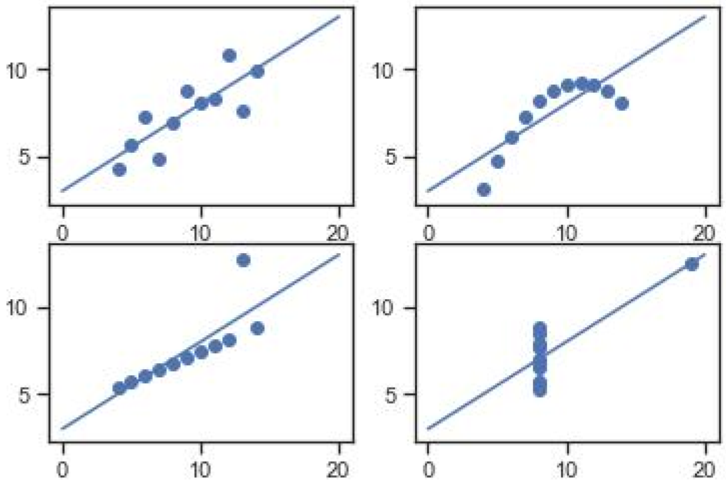
Variance of y = 4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar.

But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

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

Correlation coefficients are used to measure how strong a relationship is between two variables.

There are several types of correlation coefficient, but the most popular is Pearson's.

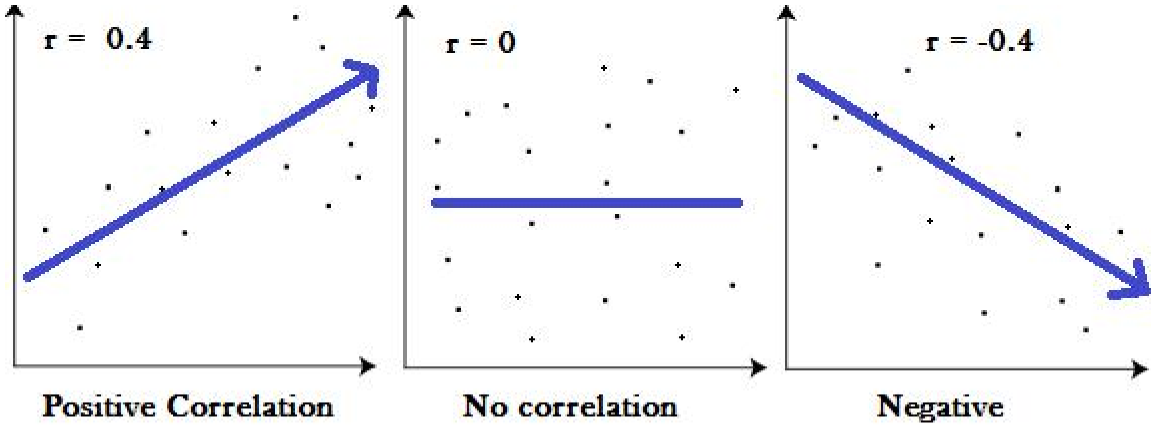
Pearson's correlation (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression.

If you're starting out in statistics, you'll probably learn about Pearson's *R* first.

In fact, when anyone refers to **the** correlation coefficient, they are usually talking about Pearson's.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



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

Types of correlation coefficient formulas.
There are several types of correlation coefficient formulas.

One of the most commonly used formulas is Pearson’s correlation coefficient formula. If you’re taking a basic stats class, this is the one you’ll probably use:

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

Pearson correlation coefficient

Two other formulas are commonly used: the sample correlation coefficient and the population correlation coefficient.

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Sx and sy are the sample standard deviations, and sxy is the sample covariance.

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The population correlation coefficient uses σx and σy as the population standard deviations, and σxy as the population covariance.

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

(3 marks)

Answer:

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 unit of the values.

Example:

If an algorithm is not using the feature scaling method then it can consider the value 3000 meters 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 the same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling
Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

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

Q5}. You might have observed that sometimes the value of VIF is infinite. Why does this happen? *(3 marks)*

Answer:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

[(VIF)]_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$[(VIF)]_2=1/(1-R_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. 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.

This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

The standard error of the coefficient determines the confidence interval of the model coefficients.

If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

A general rule of thumb is that if VIF > 10 then there is multicollinearity.

Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

Q6}. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. *(3 marks)*

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

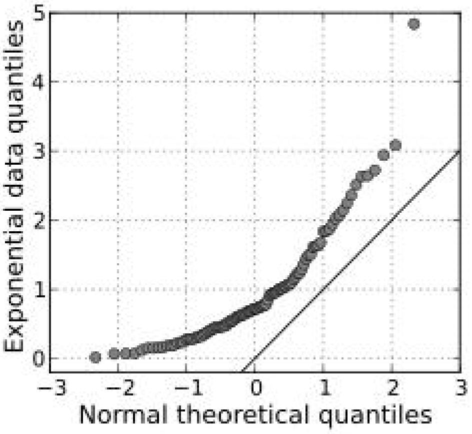
The purpose of Q Q plots is to find out if two sets of data com from the same distribution.

A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is to find out if two sets of data com from the same distribution.

From the same distribution. A 45 degree angle is plotted on the Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

Also, it helps to determine 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 .

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.