

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)

Ans:

From Boxplot visualization I have done below analysis.

- Demand increased in fall season followed with summer and Winter.
- Most of the booking made on Saturday followed by Sunday and Friday compare to other days in week.
- Booking trend increases from start of year till June then from July to Sept we can see there is stability but more booking in this period and then there is fall in demand till year end.
- Clear weather attracted more booking which seems obvious.
- Working day or non-working day has no impact on bike booking.
- 2019 has more booking compare to last year booking.

2. Why is it important to use drop first=True during dummy variable creation? (2 mark)

Ans: When we use drop_first = True, it drops first column. It is important to avoid data redundancy and co-relation between variables. As this column is well explained by other columns so there is not any missing of information. So we prefer to create n-1 dummy variables for categorical columns with n values.

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

Ans: Looking at pair plot, we can tell '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)

Ans: Below are assumptions of Linear Regression that I have validated with graph visualization on actual and predicted values.

- Zero distribution of error - This validation is done with distribution plot on error values.
 - Multi collinearity - This validation is done by calculating VIF for each variable and making sure value is less than 5.
 - Linear relationship - Check linear relationship with pair graph.
 - Homoscedasticity - No visible pattern in residual values.
 - Independence of residuals - No autocorrelation - Difference between R2 and adjusted R2 is very less.
-

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

Ans: Based on final model, below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

- temp
 - spring
 - sept
-

General Subjective Questions

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

Ans:

Linear regression- It is the model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

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 .

Linear regression is of the following two types –

1. Simple Linear Regression
2. Multiple Linear Regression

Assumptions -

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

- 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.
- Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.
- Zero distribution of errors – Error terms should be normally distributed.

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

Ans:

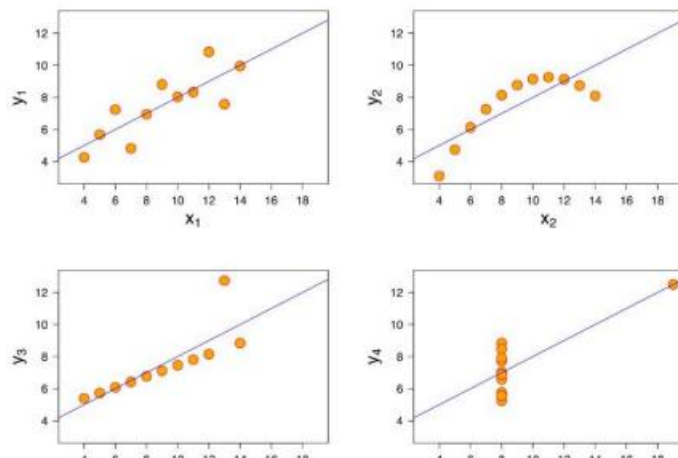
Anscombe's quartet:-Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R -squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

	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:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Ans:

Pearson's r is a numerical summary of 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 opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. Its value ranges between -1 to +1.

$r = 1$ means the data is perfectly linear with a positive slope

$r = -1$ means the data is perfectly linear with a negative slope

$r = 0$ means there is no linear association

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

Ans:

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 feature scaling method then it can consider the value 10000 meter to be greater than 5 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue

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

Ans:

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. $(VIF) = 1 / (1 - R^2)$. If there is perfect correlation, then $VIF = \text{infinity}$. A large value of VIF indicates that there is a correlation between the variables.

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

Ans:

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.
