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:

   There are categorical variables such as season, yr, mnth, holiday, weekday, workingday and weathersit. From the analysis performed, it is understood that all variables have effect on the dependent variable ‘cnt’ but the effect of the individual categorical variables on the dependent variable is different. The followings are inferred from the boxplot and barplot prepared in the analysis

   From the plots for both 2018 & 2019, it is understood for bike sharing that

   - Bike sharing Bike sharing appears to be highest in fall season and lowest in spring season
   - Bike sharing increased significantly from 2018 to 2019.
   - Bike sharing appears to be highest when sky is clear or partly clouded and lowest when there is rain & light snow.
   - Bike sharing is high during the month of May to October and is decreasing from November to January month. From February month, Bike sharing is increasing till June month.
   - Bike sharing is more or less same during all working days and non-working day.
   - Bike sharing is less on Tuesday and this could be due to people spending time with family on Tuesday.

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

   Answer:

   It is used while creating dummy variables. When creating a categorical variable with 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. This will reduce one redundant level/column. This avoids duplication of calculation/analysis. The ‘drop_first = True” is used to drop column/level which is redundant

   Example.

   There are 3 levels below to define relationship. Here the column single can be avoided as this condition can be defined based on information from other two columns (“In a relationship” and “Married”). Here when “In a relationship” and “Married” are zero, it means it is “Single”. “drop_first=True” is used to drop the column.

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

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

   Answer:

Variable “temp” has the highest correlation with the target variable.

(Note that from the correlation chart, both temp and atemp has correlation factor of 0.65. Both temp and atemp are closely representing (collinear) each other).

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

Answer:

There are four assumption of linear regression:

1. Linear relationship between independent variable and target variable
2. Error terms have constant variance (homoscedasticity)
3. Error terms are normally distributed with mean value of zero
4. Error terms are independent of each other

-Error terms are plotted to check whether they are normally distributed and they were found to be normally distributed.

-Also the residuals are plotted and they are found to be independent. There is no visible pattern.

-The model is based on linear relationship between independent and target variables.

Error

-From the plot, it is found that the variance do not increase or decrease as the error value changes. The variance do not follow any pattern when error term changes.

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:

The below top 3 features are contributing significantly for positive Bike sharing count.

1. Temp
2. winter
3. Sep

Note: If both positive and negative features are considered, the features impacting Bike sharing are 1.temp 2.Lightsnow+ rain 3.windspeed

General Subjective Questions

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

Answer:

Linear regression is a statistical method to analyse the linear relationship between a dependent variable with given set of independent variables. In the Linear relationship when the value of one or more independent variables changes, the value of dependent variable also changes in linear proportion

The relationship can be given as below

Y = mX + c

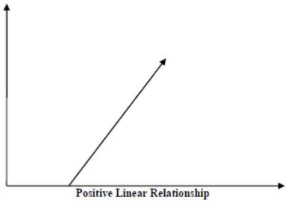
Where Y is the dependent variable and X is independent variable

m is the slope of the linear line and it indicates relationship (strength of X on Y) between X and Y.

c is a constant, known as the Y-intercept. It is fixed irrespective of X value.

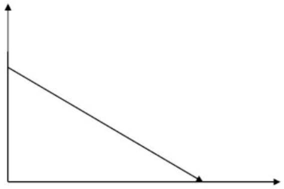
The linear relationship can be positive or negative:

- ❖ Positive Linear Relationship: A linear relationship where dependent variable(Y) value increases with increase in independent variable (X) value and vice versa. It shown pictorially below.



Positive Linear Relationship

- ❖ Negative Linear Relationship: A linear relationship where dependent variable(Y) value decrease with decrease in independent variable (X) value. It shown pictorially below.



Negative Linear Relationship

Linear regression analysis can be classified into two types –

- ➢ Simple Linear Regression: Here only one independent variable is used to predict dependent variable.
- ➢ Multiple Linear Regression: Here more than one independent variable are used to predict dependent variable.

In regression, there is a notion of a best-fit line, the line which fits the given scatter-plot in the best way.

Residual: Residual is used to find the best fit line. Every data point has a residual value which is the difference between the actual value and the predicted value (the value of point on line). Let's denote this by E(error)

E = Actual – Predicted (for every data point)

Minimize the total error square (RSS) i.e. minimize e12 + e22 + …… + en2.

This is also called as Residual Sum of Squares (RSS). So, choose the value of m and c in such a way that it reduces the value of RSS

Let's write E in terms of m and c.

E = ei = yi (actual) – ypred

ei = yi – mxi – c

In linear regression, a cost function is defined for a problem and then it is either minimized or maximized according to the requirement.

Minimizing cost function:

Ways to minimize cost function: 1. Differentiation method 2.Gradient descent method

Gradient Descent is an optimisation algorithm which optimises the objective function (for linear regression it's cost function) to reach to the optimal solution.

R2 (R2 = 1 – RSS/TSS) is used to explain how the linear regression model fits with data well.

Where, TSS is the sum of square of difference of each data point from the mean value of all the values of target variable

Assumptions:
The below assumptions are considered in linear regression:

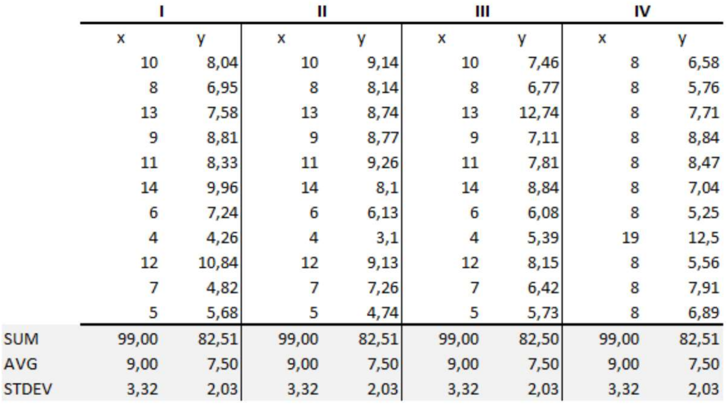
1. Linear relationship between independent variable and target variable
2. Error terms have constant variance (homoscedasticity)
3. Error terms are normally distributed with mean value of zero
4. Error terms are independent of each other

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

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises of four datasets, each containing eleven (x, y) pairs that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
They were constructed to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. It was intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

.

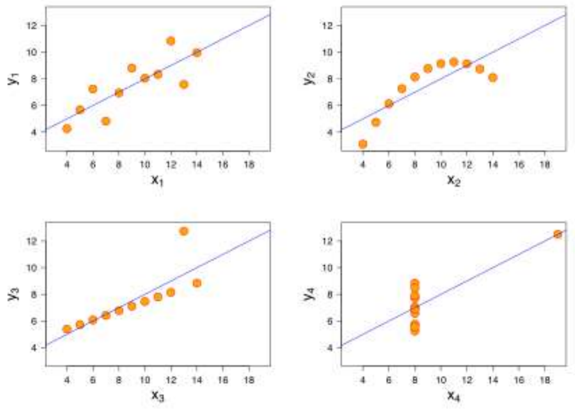


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 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 4.125 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset and R1 is 0.67.

When we plot these four datasets on X and Y coordinate planes as shown below, we can observe that they show the same regression lines as well but each dataset is telling a different perspective.



- The Top left graph appears to be a simple linear relationship, corresponding to two variables (X1 and Y1) correlated where Y1 is modelled as Gaussian with mean linearly dependent on X1.
- In the top right graph, while a relationship between the two variables (X2 and Y2) is clear, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the bottom left graph, the modelled relationship is linear between variables(X3 and Y3), but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- The bottom right graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is often used to illustrate the importance of checking dataset graphically before starting to analyse, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R? (3 marks)

Answer:

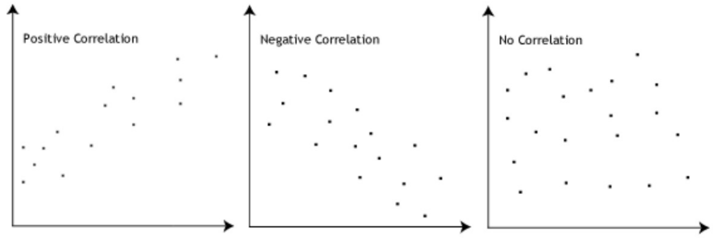
Pearson's r gives numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

A value of 0 indicates that there is no association between the two variables.

A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer:

Scaling is a technique to standardize the independent features/variables 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 of the features/variables. If scaling is not done then algorithm only takes magnitude in account and not units, it leads to incorrect modelling

Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 gram to be greater than 4 kg 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 same magnitudes.

S. No	N**ormalisation**	**Standardisation**
1.	Minimum and maximum value of variables are used for scaling	Mean and standard deviation is used for scaling.
2.	Used when features are of different scales.	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 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.
7.	It is useful when we don't know about the distribution	It is useful when the variable distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Answer:

If there is perfect correlation between variables, then VIF will be infinity. A large value of VIF indicates that there is a perfect correlation between the variables.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To resolve this incident, one of the variables from the dataset which is causing this perfect multi-collinearity can be dropped.

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

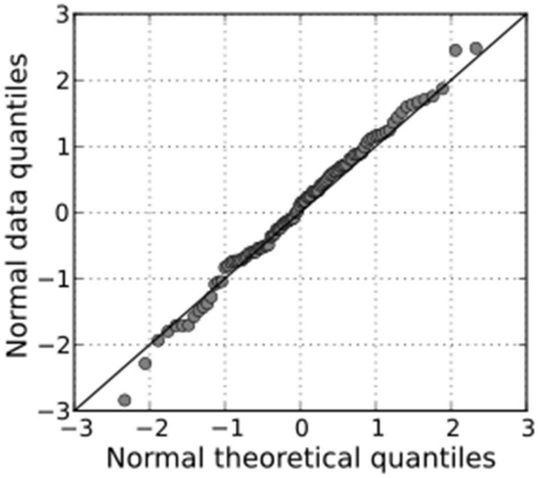
Answer:

Q-Q plot compares the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

The quantile-quantile (Q-Q) plot is used 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.