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 my analysis of the categorical variables from the dataset, I could infer their effect on the dependent variable (Count of bookings 'cnt') as follows –

- **season:** Fall season seems to have attracted more bookings.
- **yr:** 2019 attracted a greater number of bookings as compared to 2018, which shows good progress in terms of business.
- **mnth:** Most of the bookings have been done between the month of may to oct, with aug recoding highest bookings.
- **holiday:** Less bookings when it's not a holiday as people may want to spend time at home and enjoy with family.
- **weekday:** Number of bookings look identical across the weekdays.
- **workingday:** More bookings on workingday.
- **weathersit:** Clear weather attracted more bookings which seems obvious.

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

Ans: It is important to use drop_first=True, as it helps in reducing the extra column created during dummy variable creation thereby reducing the correlations created among dummy variables.
If we set drop_first = True, then it will drop the first category. So, if we have K categories, it will only produce K – 1 dummy variables.

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 the pair-plot among the numerical variables**, '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: After building the model on the training set, I validated the assumption of Linear Regression in the following ways:

a. **Linear Relationship**:
   - Plotted CCPR Plot to validate the linear relationship between the independent and the dependent variables.

b. **Multivariate Normality:**
   - Plotted histogram to validate that the errors between observed and predicted values (i.e., the residuals of the regression) are normally distributed.

c. **No Multicollinearity:**
   - Computed a matrix of Pearson's bivariate correlations among all independent variables by plotting a heatmap and the VIF to validate no multicollinearity.

d. **Homoscedasticity:**
   - Plotted regplot of residuals versus predicted values to confirm that there is no clear pattern in the distribution.

e. **No autocorrelation:**
   - Looked for Durbin – Watson (DW) statistic which was within the accepted range implying no autocorrelation.

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 the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are **yr**, **holiday** and **temp**

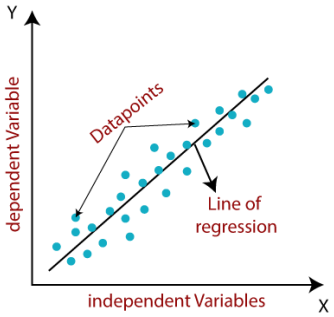
General Subjective Questions

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

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It is a statistical method that is used for predictive analysis. 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.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables as shown in the image alongside:



Mathematically, we can represent a linear regression as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where,

y = Dependent Variable (Target Variable)
x = Independent Variable (Predictor Variable)
β_0 = intercept of the line (Gives an additional degree of freedom)
β_1 = Linear regression coefficient (scale factor to each input value)
ε = random error

- **Types of Linear Regression**

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

  a. **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.

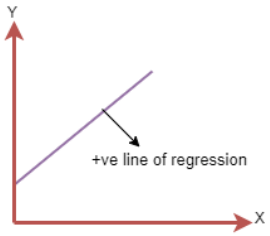
  b. **Multiple Linear Regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

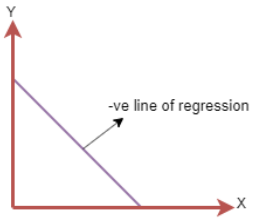
- **Linear Regression Line**

  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:

  a. **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. Linear Regression in Machine Learning

  b. **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.





- **Finding the best fit line**

  When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

  The different values for weights or the coefficient of lines (β_0, β_1) gives a different line of regression, so we need to calculate the best values for β_0 *and* β_1 to find the best fit line, so to calculate this we use cost function.

a. Cost function:

- It is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\boldsymbol{MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y_i}\right)^2}$$

where,

N=Total number of observations
Y_i = Observed values
$\widehat{Y_i}$ = Predicted values

b. Residuals:

The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

c. Gradient Descent:

- It is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

➢ **Model Performance**

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.
- It can be calculated from the below formula:

$$\boldsymbol{R^2 = 1 - \frac{RSS}{TSS}}$$

where,

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

➢ **Assumptions of Linear Regression**

For a successful regression analysis, it's essential to validate the assumptions mentioned below-

a. **Linear relationship**: There should be a linear relationship between the dependent and independent variables.
b. **Multivariate normality**: The error terms should follow the normal distribution pattern.
c. **No multicollinearity**: The independent variables should not be correlated.
d. **Homoscedasticity**: The error terms must have constant variance and should be no clear pattern distribution of data in the scatter plot.
e. **No autocorrelation:** There should be no correlation between the residual (error) terms.

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

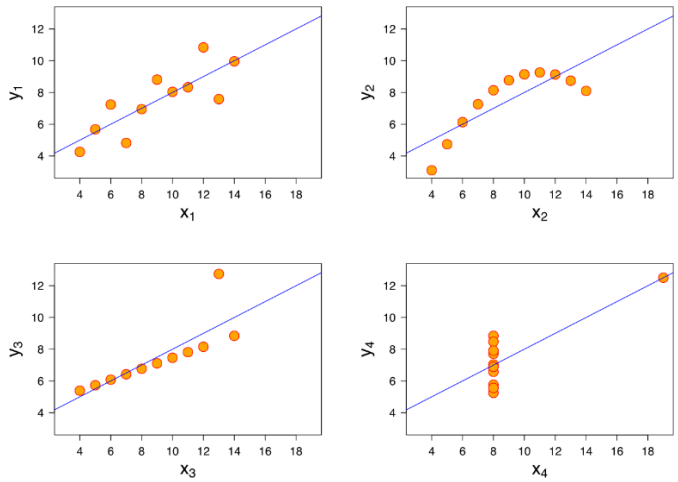
Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

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

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed:



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, 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 third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) 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 looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

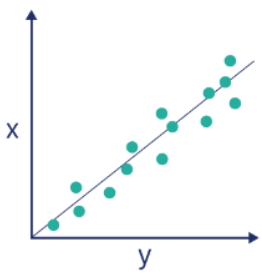
3. What is Pearson's R? (3 marks)

Ans: The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

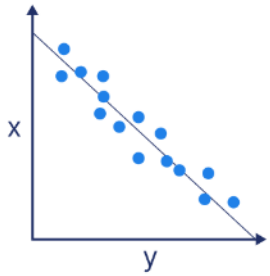
The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

The Pearson correlation coefficient, r, is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

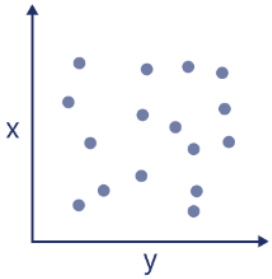
Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.



Positive Correlation



Negative Correlation



No Correlation

It is calculated as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson correlation attempts to draw a line of best fit through the spread of two variables. Hence, it specifies how far away all these data points are from the line of best fit. Value of 'r' equal to near to +1 or -1 that means all the data points are included on or near to the line of best fit respectively. Value of 'r' closer to the '0' data points is around the line of best fit.

- **Assumptions for a Pearson Correlation**:
  a. Both variables are quantitative
  b. The variables are normally distributed
  c. The data have no outliers
  d. The relationship is linear

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

Ans: What is scaling?

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.

Why is scaling performed?

When there are many independent variables in a model, a lot of them might be on very different scales. This leads to a model with very weird coefficients, which are difficult to interpret. Thus, one needs to scale the features for ease of interpretation of the coefficients as well as for the faster convergence of gradient descent methods.

Difference between normalized and standardized scaling

One can scale the features using the following methods:

a. **Standardization**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x' = \frac{x - \bar{x}}{\sigma} \quad where \ \bar{x} \ is \ mean, \ \sigma \ is \ standard \ deviation$$

b. **Min-Max Scaling**: The variables are scaled in such a way that all the values lie between zero and one.

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

S.NO.	Normalization	Standardization
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.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

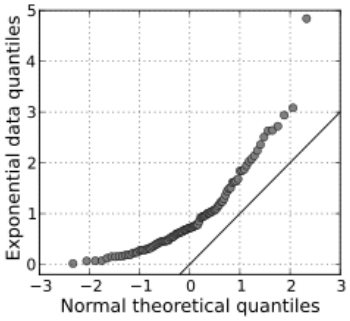
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans: Q-Q plots or Quantile-Quantile plots, plot the quantiles of a sample distribution against quantiles of a theoretical distribution to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.



Use and importance:

Q-Q plots have the ability to summarize any distribution visually.

They are very useful to determine-

- If two populations are of the same distribution.
- If residuals follow a normal distribution.
- Skewness of distribution

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Some machine learning models work best under some distribution assumptions. Knowing which distribution we are working with can help us select the best model.