Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- The year 2019 witnessed a higher number of bookings compared to the previous year, indicating positive progress in terms of business.
- The Autumn(fall) season has experienced a notable increase bookings. Additionally, across all seasons, there has been a substantial rise in booking counts from 2018 to 2019.
- On non-holidays, the booking count tends to be lower, which is reasonable as people may prefer spending time at home with family during holidays.
- It's evident that clear weather conditions (labelled as Good in the notebook) played a significant role in attracting more bookings.
- There appears to be a relatively equal distribution of bookings between working days and non-working days.
- Bookings were more prevalent on Thursday, Friday, Saturday, and Sunday compared to the early days of the week.
- Majority of bookings occurred in May, June, July, August, September, and October. The trend exhibited an increase from the beginning of the year until mid-year, followed by a decrease towards the year's end.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

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.

Dummy variable creation is a technique used in statistical modelling and machine learning to represent categorical variables with binary values (0 or 1). It involves creating new binary (dummy) variables for each category of the original categorical variable. These dummy variables serve as indicators for the presence or absence of a specific category.

The number of dummy variables to create depends on the number of categories in the original categorical variable. For a categorical variable with categories, you typically create n-1 dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

'temp' variable has the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I have validated the assumption of Linear Regression Model based on below assumptions:

- Normality of error terms
- Linear relationship validation
- Homoscedasticity
- Independence of residuals
- Multicollinearity check

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)
There are below top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. temp
2. windspeed
3. summer

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Linear regression may be defined as the statistical 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).
Mathematically the relationship can be represented with the help of following equation:
Y=mX+c
Y is the dependent variable we are trying to predict.
m is the slope of the regression line with 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.

Types of Linear Regression:

1. Simple Linear Regression: One independent variable(X) and one dependent variable(Y). Example explained above.
2. Multiple Linear Regression: More than one independent variables.
   Equations: Y= B0+B1X1+B2X2+B3X3+---------+BnXn
   Here, X1,X2,X3,----,Xn are multiple independent variables.
   Each independent variable has a coefficient Bi representing its impact on Y.

There are below concept behind Linear Regression work:

- Hypothesis function (Equal of a line)
- Cost Function (Mean Squared Error – MSE)
- Optimization using Gradient Descent

The Linear relationship can be positive or negative in nature. If both independent and dependent variable increases it will be positive. If both variables decrease it will become negative.

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 auto-correlation 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.
Normality of error terms: Error terms should be normally distributed.
Homoscedasticity: There should be no visible pattern in residual values.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

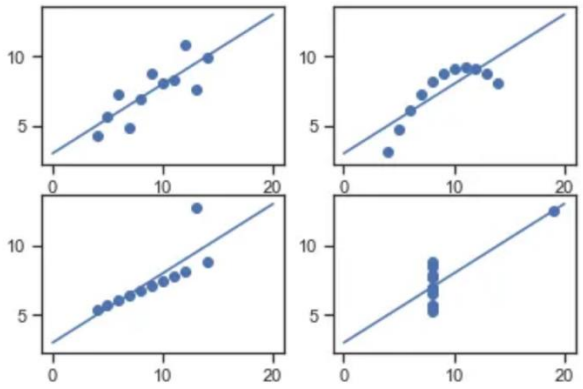
Anscombe's Quartet is the modal example to demonstrate the importance of data visualisation which was developed by the statistician Francis Anscombe to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x, y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc.) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
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

Four Data-sets

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 behaviour.



Graphical Representation of Anscombe's Quartet

Understand each dataset with visualization.
Dataset1: Linear relationship
Dataset2: Nonlinear relationship
Dataset3: Influence of outliers
Dataset4: Vertical outliers

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
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.
The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

Value of r	Interpretation
+1.0	Perfect positive correlation (both increase together)
+0.7 to +0.9	Strong positive correlation
+0.4 to +0.6	Moderate positive correlation
+0.1 to +0.3	Weak positive correlation
0	No correlation (no linear relationship)
-0.1 to -0.3	Weak negative correlation
-0.4 to -0.6	Moderate negative correlation
-0.7 to -0.9	Strong negative correlation
-1.0	Perfect negative correlation (one increases, the other decreases)

For example in Bike Sharing: If temperature and bike demand (cnt) have Pearson's R = +0.85, this means as temperature increases, bike demand also increases.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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.

Scaling is the process of transforming numerical features so that they have a consistent range or distribution. This is crucial in machine learning models where different features have different units (e.g., temperature in °C, windspeed in km/h, bike demand in counts).

Scaling importance:

- Prevents larger values from dominating the model.
- Improves convergence speed in gradient-based optimization (e.g., linear regression, logistic regression, neural networks).
- Enhances model performance in distance-based algorithms (KNN, SVM, PCA, clustering)

For example, in the Bike Sharing System dataset, features like: Temperature (temp) ranges from 0 to 40°C, Humidity (humidity) is between 0 and 100% and Bike Demand (cnt) varies from hundreds to thousands etc. Since these variables have different scales, applying scaling ensures fair comparison and improves model efficiency.

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 svals	It is used when we want to insure 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 ooutliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provide a transformer called StandardScaler for standardization.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. It tells us how much the variance of a regression coefficient is inflated due to multicollinearity.

VIF=1/(1-R^2)

IF R^2 is close to 1, VIF become very high. High Multicollinearity.

If R^2 =1, then VIF become infinite.

IF becomes infinite (or very high) when:

1. Perfect Multicollinearity Exists.
2. Duplicate or Highly Correlated Features.
3. Dummy Variable Trap in One-Hot Encoding.
4. Including an Interaction Term without the Original Features.

For example, If 'Temperature (°F)' and 'Temperature (°C)' are both in the dataset, one can be exactly derived from the other. In the Bike Sharing dataset,
if temp and feels_like_temp are highly correlated, VIF may be infinite.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
In your Bike Sharing dataset, a Q-Q plot can be used to check if residuals follow a normal distribution.
As its Normal residuals -- > Regression model is valid

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.