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)

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 time at home and enjoy with family.
- 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.

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.

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.

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 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

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)

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

temp
winter
sep

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)

<Your answer for Question 6 goes here>

What is Linear Regression?

Linear Regression is a **supervised machine learning algorithm** used for **predictive modeling** where the relationship between **independent variables (X)** and **dependent variable (Y)** is modeled using a straight line.

It assumes that there is a **linear relationship** between input features (**X**) and the target variable (**Y**). The equation for this relationship is:

Y=mX+cY = mX + cY=mX+c

For multiple variables:

Y=b0+b1X1+b2X2+...+bnXnY = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_nY=b0+b1X1+b2X2+...+bnXn

where:

- YYY → Dependent Variable (Target)
- X1,X2,...,XnX_1, X_2, ..., X_nX1,X2,...,Xn → Independent Variables (Features)
- b0b_0b0 → Intercept (Bias)
- b1,b2,...,bnb_1, b_2, ..., b_nb1,b2,...,bn → Coefficients (Weights)
- ε\varepsilonε → Error term

Types of Linear Regression

1. Simple Linear Regression

- Uses only **one independent variable**.
- Equation: Y=b0+b1X+εY = b_0 + b_1X + \varepsilonY=b0+b1X+ε
- Example: Predicting house price (Y) based on **square footage (X)**.

2. Multiple Linear Regression

- Uses **multiple independent variables**.
- Equation: Y=b0+b1X1+b2X2+...+bnXn+εY = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n + \varepsilonY=b0+b1X1+b2X2+...+bnXn+ε
- Example: Predicting house price (Y) based on **square footage (X_1), number of bedrooms (X_2), and location (X_3)**.

How Linear Regression Works

Linear Regression estimates the best-fit **line** (hyperplane in multiple dimensions) by minimizing the difference between actual values (**y_actual**) and predicted values (**y_predicted**).

1. Cost Function (Mean Squared Error - MSE)

To find the best coefficients b0,b1,...bnb_0, b_1, ... b_nb0,b1,...bn, we minimize the **cost function**:

MSE=1N∑i=1N(yi−y^i)2MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2MSE=N1i=1∑N (yi−y^i)2

where:

- yiy_iyi = Actual value
- y^i\hat{y}_iy^i = Predicted value
- NNN = Number of data points

MSE measures the **average squared error** and penalizes larger deviations.

2. Optimization using Gradient Descent

Gradient Descent is an optimization algorithm that **iteratively updates** the coefficients to minimize MSE.

1. **Initialize coefficients** b0,b1,...bnb_0, b_1, ... b_nb0,b1,...bn with random values.
2. Compute **gradient (derivative of cost function)**: bj=bj−α∂MSE∂bjb_j = b_j - \alpha \frac{\partial MSE}{\partial b_j}bj=bj−α∂bj∂MSE where α\alphaα is the **learning rate**.
3. Update **parameters** iteratively until **convergence** (i.e., cost function stops decreasing).

Advantages of Linear Regression

☑**Easy to interpret** - Coefficients show how much each feature contributes to prediction.
☑**Computationally efficient** - Works well for large datasets.
☑**Useful for understanding relationships** between variables.

Disadvantages of Linear Regression

❌**Assumes linearity** - May not work well for non-linear relationships.
❌**Sensitive to outliers** - A few extreme values can distort predictions.
❌**Multicollinearity issue** - If independent variables are highly correlated, model accuracy decreases.

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)

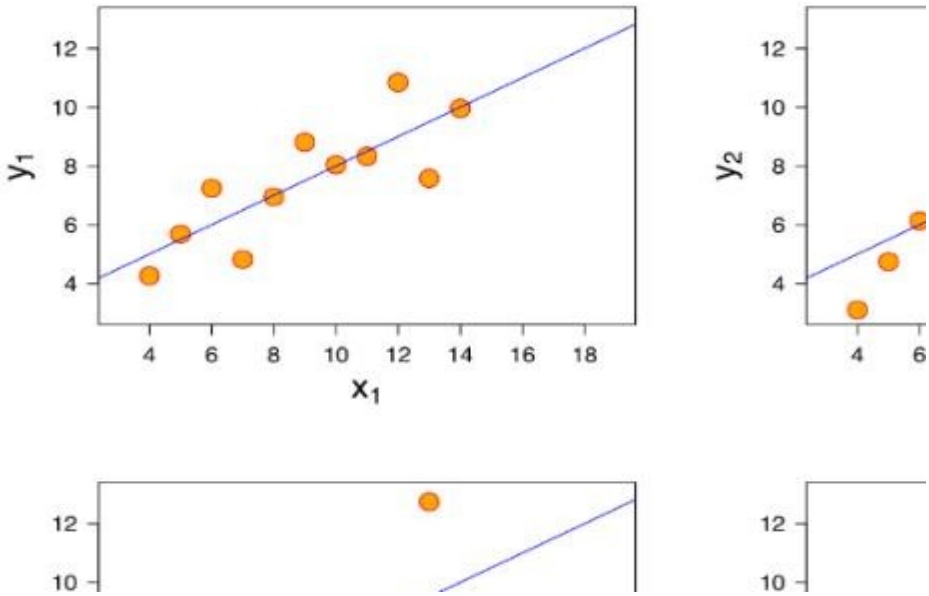
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

I		II		III
x	y	x	y	x
10	8,04	10	9,14	10
8	6,95	8	8,14	8
13	7,58	13	8,74	13
9	8,81	9	8,77	9
11	8,33	11	9,26	11
14	9,96	14	8,1	14
6	7,24	6	6,13	6
4	4,26	4	3,1	4
12	10,84	12	9,13	12

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 appears 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.

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. 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.

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.

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

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.

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)

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.

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 solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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.

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.