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:

Based on the analysis of categorical variables using boxplots and bar plots, the following inferences can be made:

1. **Seasonal Trends**: The fall season saw a significant increase in bookings, with a noticeable rise from 2018 to 2019. This indicates that fall is a peak period for bookings.
2. **Monthly Trends**: Bookings peaked during the months of May, June, July, August, September, and October. There was an upward trend in bookings from the start of the year, reaching a high in mid-year, and then tapering off towards the end.
3. **Weather Conditions**: Clear weather conditions were associated with a higher number of bookings, which is understandable as people are more likely to engage in activities during pleasant weather.
4. **Day of the Week**: Bookings were higher on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week, suggesting a preference for leisure activities closer to or during weekends.
5. **Holidays**: There were fewer bookings on holidays, likely because people prefer to stay at home and spend time with family rather than go out.
6. **Working Days**: The number of bookings did not differ significantly between working days and non-working days, indicating a steady demand across the week.
7. **Yearly Comparison**: There was a noticeable increase in bookings in 2019 compared to the previous year, reflecting positive growth in business.

These insights highlight patterns and preferences among users, offering valuable information for business planning and strategy.

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

   Using drop_first=True during dummy variable creation is important because it helps eliminate redundancy by removing one of the dummy variables for each categorical feature. This reduction prevents multicollinearity, which occurs when dummy variables are highly correlated.

   In technical terms, setting drop_first=True means that if a categorical feature has *k* unique values, only *k-1* dummy variables will be created. The first category is dropped, serving as a reference level. For instance, if a categorical variable has three categories (A, B, C), only two dummy variables are created (for A and B). The third category, C, can be inferred if both A and B are 0. This practice avoids the "dummy variable trap" and ensures that the model is not overfitting due to perfectly collinear features.

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

I have validated the assumptions of the Linear Regression Model based on the following five criteria:

1. **Normality of Error Terms:**
   The error terms should be normally distributed.
2. **Multicollinearity Check:**
   There should be insignificant multicollinearity among the independent variables.
3. **Linear Relationship Validation:**
   There should be a linear relationship between the independent and dependent variables.
4. **Homoscedasticity:**
   The residuals should have constant variance, with no visible pattern in their distribution.
5. **Independence of Residuals:**
   The residuals should be independent, with 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)**
   Answer:
   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
   - temp
   - winter
   - sep

General Subjective Questions

Explain the linear regression algorithm in detail. (4 marks)Answer:
Linear regression is a statistical model used to analyze the linear relationship between a dependent variable and a given set of independent variables. This linear relationship implies that changes in the independent variables (whether an increase or decrease) will result in corresponding changes in the dependent variable. The relationship can be represented mathematically by the equation:

$Y=mX+c$

Y represents the dependent variable we aim to predict.

- **X** represents the independent variable used for prediction.
- **m** is the slope of the regression line, indicating the influence of X on Y.
- **c** is the constant term, known as the Y-intercept, representing the value of Y when X equals zero.

The nature of the linear relationship can be either positive or negative:

- **Positive Linear Relationship:**
  When both the independent and dependent variables increase together. The graph of such a relationship slopes upwards.
- **Negative Linear Relationship:**
  When the independent variable increases while the dependent variable decreases. The graph of such a relationship slopes downwards.

Types of Linear Regression:

1. **Simple Linear Regression:** Involves one independent variable.
2. **Multiple Linear Regression:** Involves two or more independent variables.

Assumptions of Linear Regression:

1. **Multi-collinearity:**
   The model assumes minimal or no multicollinearity among the independent variables, meaning there should be little or no dependency among them.
2. **Auto-correlation:**
   The model assumes minimal or no autocorrelation in the data, indicating that the residual errors should not be dependent on each other.
3. **Linear Relationship:**
   The relationship between the response variable and the independent variables must be linear.
4. **Normality of Error Terms:**
   The error terms should follow a normal distribution.
5. **Homoscedasticity:**
   The variance of the residuals should be constant across all levels of the independent variables, with no visible patterns in the residuals.

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

Answer:

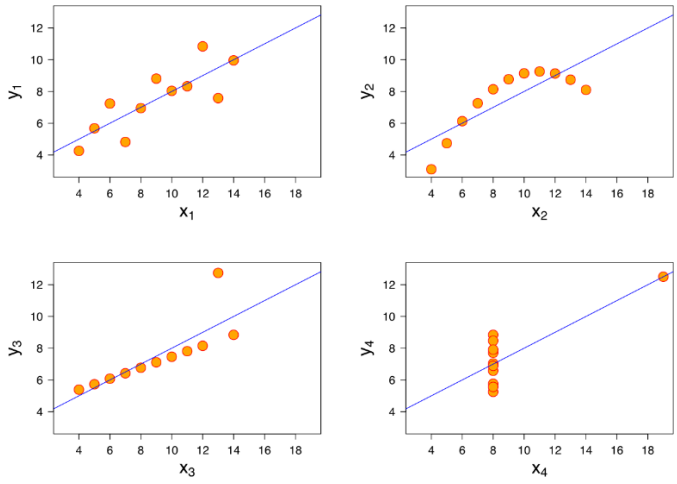
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when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

	I		II		III		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 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.

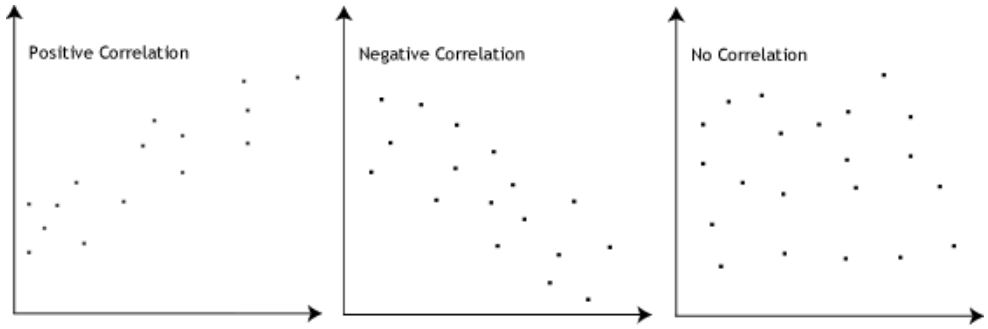
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

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. This is shown in the diagram below:



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

Answer:

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.

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

Answer:

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.

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

Answer:

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.