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:

I have conducted the univariate and segmented univariate analysis on the categorical variables using Boxplot and bar plot. Following is the inference that concluded from those visualizations:

1. Bike sharing count increases more in fallen and summer seasons compared to other seasons, and as the year increases from 2018 to 2019, the demand for bike sharing also increases
2. Bike sharing count is highest in the months of May, June, July, August, September, and October Trend increased starting the year till mid of the year, and then it decreased as we approached the end of the year. For the same month, the number of bike sharing counts is more in 2019 compared to 2018. This indicates that as the year increases, the demand for bike sharing is also increasing
3. Clear weather leads to more bike-sharing counts
4. Bike-sharing counts increase on Monday, Friday, and Saturday
5. When it's not a holiday, bike sharing counts seem to be less in number, which seems reasonable as on holidays, people may want to spend time at home and enjoy with family
6. Bike sharing count is almost equal on working and non-working days
7. 2019 attracted more bookings than the previous year, showing promising business progress.

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

Answer:

drop_first = True is useful since it reduces the unnecessary column produced during dummy variable construction. As a result, it lowers the correlations formed between dummy variables.
drop_first : bool, the default value is False When True, it signifies whether to extract k-1 dummies from k category levels by deleting the first level.
Assume we have three different sorts of values in the Categorical column and want to construct a dummy variable for that column. If one of the variables is neither A or B, the obvious answer is C. As a result, we do not require a third variable to identify the C.
drop_first = True avoids the curse of dimensionality if our dataset has a large number of categorical variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
'temp' and ‘atemp’ 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 validated the Linear Regression Model assumptions using the five assumptions listed below:

- **Normality of error terms:**
  It is assumed that the error terms, ε(i), are normally distributed.

- **Multicollinearity:**
  There should be insignificant multicollinearity among variables.
- **Linear relationship:**
  It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.
- **Homoscedasticity:**
  There should be no visible pattern in residual values, and it is assumed that the residual terms have the same variance.
- **Independent error term:**
  It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

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 top three factors that contribute significantly to explaining the demand for shared bikes are as follows:

- Temp
- Year
- Sep

General Subjective Questions

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

Answer:

Linear regression is a statistical model that examines the linear relationship between a dependent variable and a collection of independent variables. When there is a linear relationship between variables, it indicates that when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes as well (increases or decreases).

Mathematically the relationship can be represented with the help of the following equation -

$$Y = mX + c$$

Here,

Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line, which 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 relationship:

- Positive Linear Relationship
- Negative Linear relationship

Linear regression is of the following two types:

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

The following are some assumptions about the dataset that is made by the Linear Regression model -

- Multi-collinearity:

  The 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 a dependency in them.
- Auto-correlation:

  Another assumption the Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is a dependency between residual errors.
- Relationship between variables:

  The 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.

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

Answer:

Francis Anscombe, a statistician, created Anscombe's Quartet. It consists of four datasets, each with eleven (x, y) pairings. The most important thing to remember about these datasets is that they all use the same descriptive statistics. But when things are graphed, they shift totally, and I mean absolutely. Regardless of their identical summary data, each graph conveys a different story.

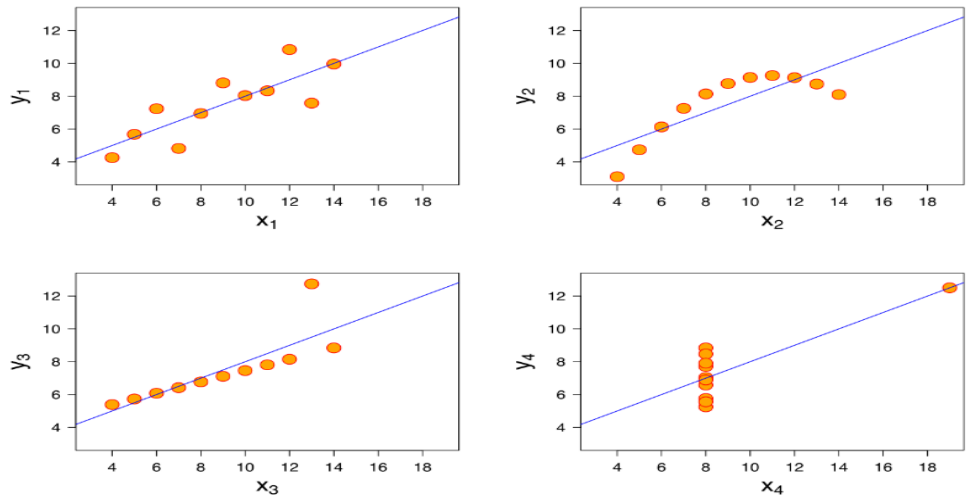
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 see that they all have the same regression lines, but they each tell a different story:



• Dataset I appear 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.

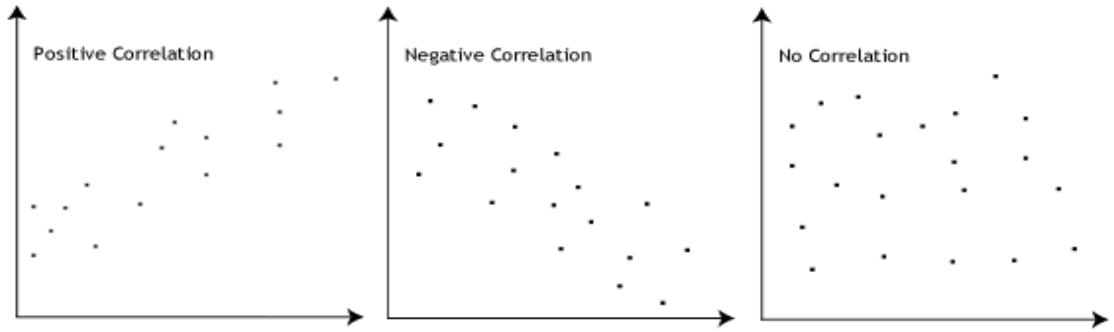
This Anscombe's Quartet underlines the significance of visualization in data analysis. Looking at the data provides much of the structure as well as a clear image of the dataset.

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

Answer:

Pearson's r is a numerical measure of the strength of the linear relationship between the variables. If the variables tend to rise and fall together, the correlation coefficient will be positive. If the variables tend to move in opposite directions, with low values of one variable correlated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can vary from +1 to -1. A value of 0 shows that there is no relationship between the two variables. A number larger than 0 shows a positive correlation; that is, when the value of one variable grows, so does the value of the other variable. A number less than 0 implies a negative. A number less than 0 implies a negative relationship; that is, when the value of one variable rises, the value of the other variable falls. This is seen in the figure below:



4. **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 for standardizing the independent features included in data within a specific range. It is used during data pre-processing to deal with highly varying magnitudes, values, or units. If feature scaling is not performed, a machine learning algorithm will tend to weight bigger values higher and consider smaller values to be lower, regardless of the unit of measurement.

For example, if an algorithm does not use the feature scaling approach, it may perceive the value 3000 meter to be more than 5 km, which is not accurate and causes the algorithm to make incorrect predictions. As a result, we apply Feature Scaling to bring all values to the same magnitude and so address this issue.

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.

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

Answer:

The Variance Inflation Factor (VIF) is a statistic that calculates how much the variance of an estimated regression coefficient rises as a result of collinearity. It is determined by dividing the variance of all the betas in a given model by the variance of a single beta if it were fit alone. The higher the VIF score, the stronger the variable's association with other variables. Values more than 4 or 5 are commonly considered moderate to high, whereas values greater than 10 are considered extremely high. VIF = infinite if there is a perfect correlation. A variable with an infinite VIF value is a perfect linear combination of other variables. If the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”

When the value of VIF is infinite, the correlation between two independent variables is perfect. In the event of perfect correlation, R-squared (R2) = 1, resulting in 1/ (1-R2) infinite. To address this, we must remove one of the variables from the dataset that is producing the perfect multicollinearity.

6. 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 tool that may be used to determine if a collection of data is likely to have come from a theoretical distribution such as a Normal, Exponential, or Uniform distribution. It also aids in determining whether two data sets are from populations with a similar distribution. A Q-Q plot is used to compare the shapes of distributions, offering a graphical representation of how features like location, size, and skewness change between the two distributions. The strength of Q-Q charts rests in their capacity to graphically summarize any distribution.

Use of Q-Q plot:

A Q-Q plot is a comparison of the first data set's quantiles to the quantiles of the second data set. A quantile is the proportion (or percentage) of points that fall below a specific number. That is, the 0.3 (or 30%) quantile is the number at which 30% of the data falls below and 70% falls above. There is also a 45-degree reference line drawn. If the two sets are drawn from the same population, the points should fall roughly along this Reference line. The larger the deviation from this reference line, the stronger the evidence that the two data sets came from populations with distinct distributions.

Importance of Q-Q plot:

It is frequently desirable to determine if the assumption of a common distribution is warranted when there are two data samples. If this is the case, then location and scale estimators can combine both data sets to generate estimates of the common location and scale. If two samples differ, it is also beneficial to get an understanding of the discrepancies. Analytical approaches such as the chi-square and Kolmogorov-Smirnov 2-sample tests can give greater insight into the nature of the difference than the Q- Q plot.