Contents

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?

Answer:

From my analysis of the categorical variables from the dataset, I could infer few points about their effect on the dependent variable.
By using boxplot and bar plot, I found couple of observations which are as follows:
The categorical variables from the dataset are: “season”, “year”, “month”, “holiday”, “weekday”, “workingday”, “weathersit”.

- Fall and Summer season are the most favorable season for biking.
- It is observed that bike booking increased from the year 2018 to 2019.
- Bike booking are very less in Spring season.
- 2019 attracted a greater number of booking as compared to 2018 which is a good progress in terms of the business.
- The bike booking is greater in the month of “June”, “Aug”, “Sep” and “Oct”.
- From the graph it is observed that clean weather attracted more booking as compared to the mist and cloudy, light snow.
- The number of bookings on working day and holiday seemed to be equal.
- Thu, Fri, Sat and Sun seems to have a greater number of bookings as compared to the start of the week.

These are the all inferences that can be made.
Hence these categorical variables show a good trend and can be chosen for the model building.

---

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

Answer:

- The creation of dummy variables to convert a categorical variable into a numeric variable is an important step in data preparation.
- During dummy variable creation it is important to use drop_first = True because it removes the first column which is created for the first unique value of a column.
- It reduces the correlations created among dummy variables and make the model significant. Hence drop_first = True is an important in dummy variable creation.
- When you have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

- By looking at the pair-plot among the numerical variables, 'temp' variable is the one which has the highest correlation with the target variable.
- The numerical variable 'atemp' also shows high correlation with the target variable but due to the multicollinearity both the variables cannot be used in the model.
- Hence not considering 'atemp' as it is dropped in the model preparation.

---

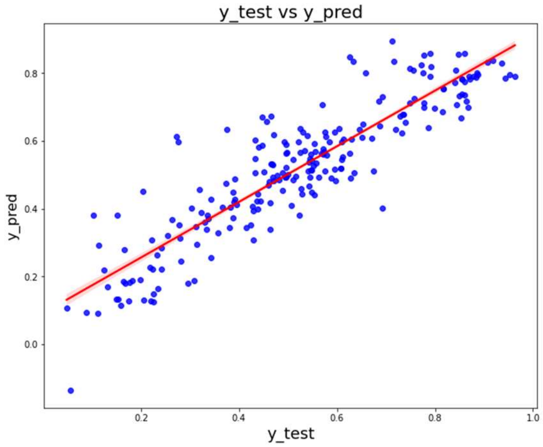
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Following are the assumptions of Linear Regression after building the model on the training set:

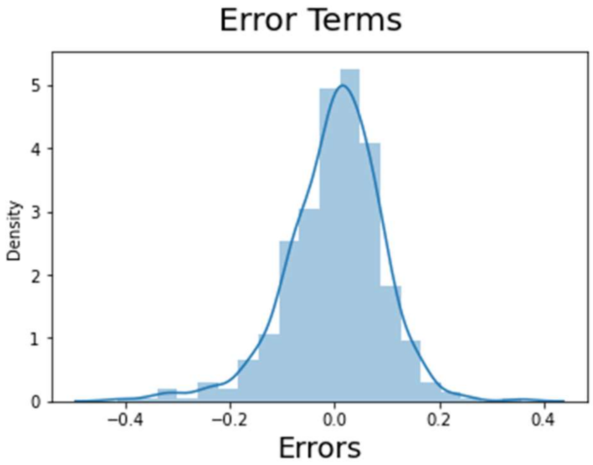
1. **There is a *linear relationship* between independent and dependent variable**

   By plotting the actual vs predicted graph it can be inferred that the points are distributed symmetrically around the diagonal.



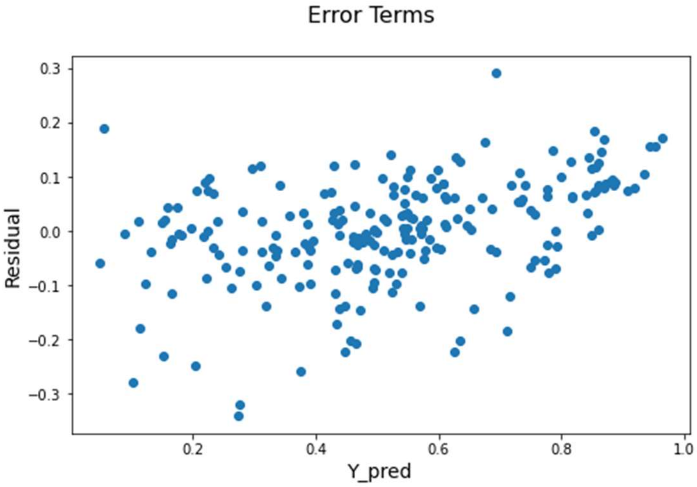
2. **Error terms are *normally distributed* with mean zero**

By plotting the histogram of the error terms, it can be seen that Error terms are normally distributed



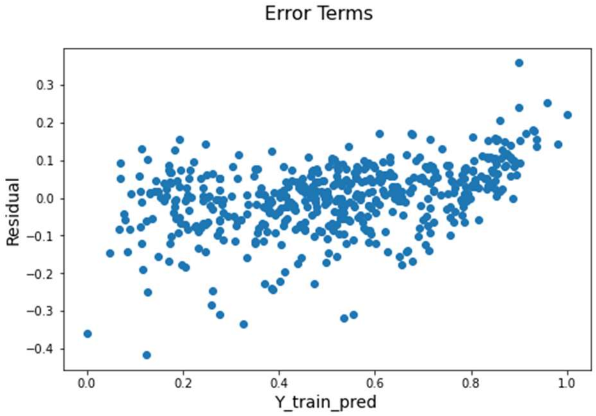
3. **Error terms are independent of each other**

We can see there is no specific Pattern observed in the Error Terms with respect to Prediction.



4. **Error terms have *constant variance* (homoscedasticity)**

   It can be seen that the plots are randomly distributed.



---

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are determined by the magnitude of the coefficients and the variable are as follows:

a) temp: It is most important feature which is affect the business positively.
b) year: the growth in the year plays important role.
c) Winter: winter season is paying important role in the demand of bike.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

- Linear regression may be defined as the statistical model that analyzes 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).
- Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: This is used when the number of independent variables is 1. The equation of the best fit regression line is:

  $Y = \beta_0 + \beta_1 X$

  It can be found by minimising the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:

  - ✓ Differentiation
  - ✓ Gradient descent

- Multiple linear regression: This is used when the number of independent variables is more than 1. The formulation for predicting the response variable now becomes this:

  $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

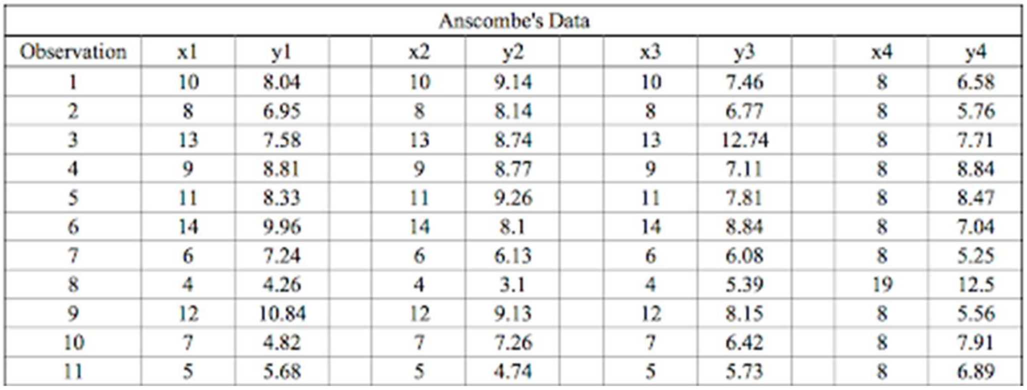
- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.
  - RSS: Residual sum of squares
  - TSS: Total sum of squares
- Building a linear model
  - OLS (Ordinary Least Squares) method in statsmodels to fit a line.
  - Summary statistics
  - F-statistic, R-squared, coefficients and their p-values.

---

2. Explain the Anscombe's quartet in detail.

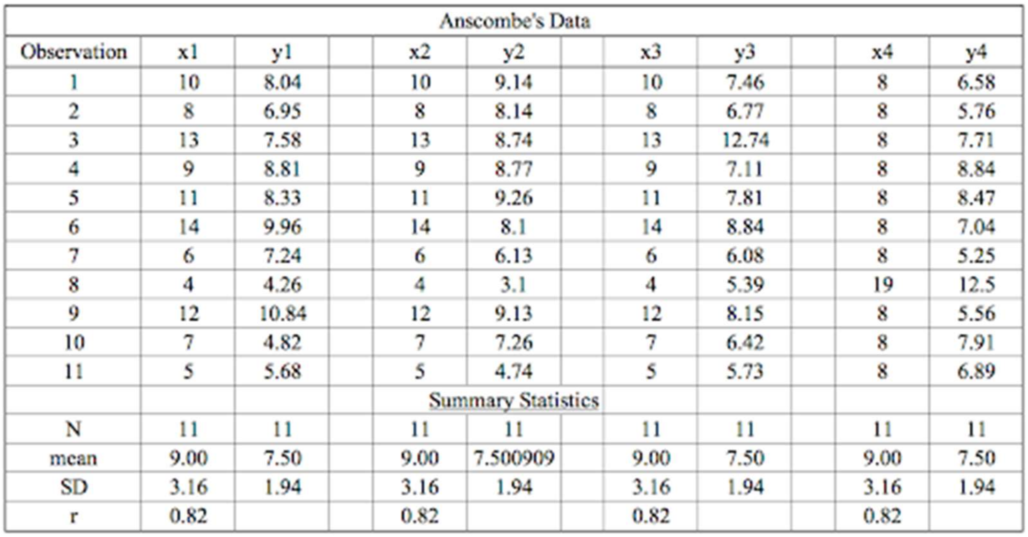
Answer:

- **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet 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 before analyzing it and the effect of outliers on statistical properties.
- The basic thing to analyze 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 behavior irrespective of statistical analysis.
- We can define these four plots as follows:



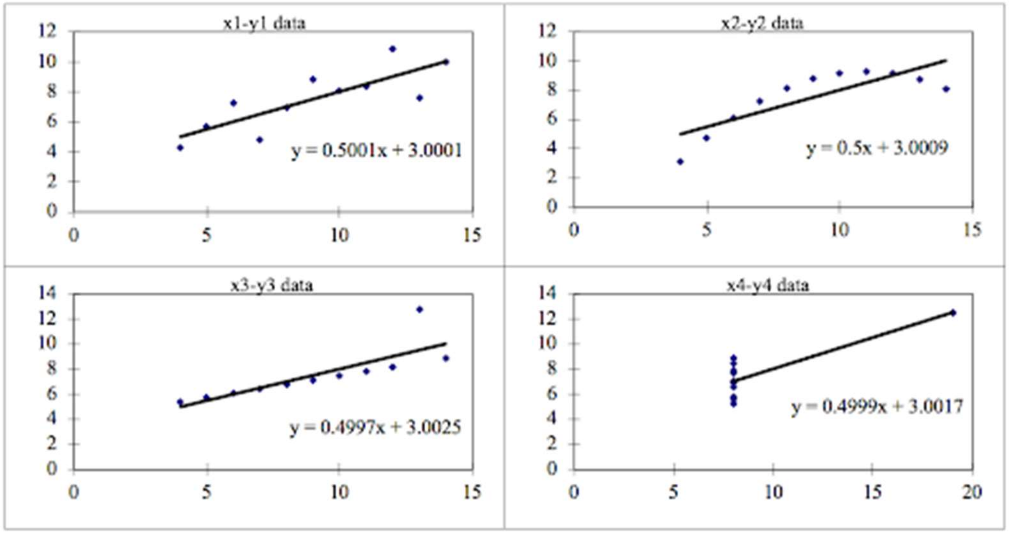
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

- However, the statistical analysis of these four data-sets are pretty much similar. We can compute them as follows:



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

- 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 behavior.



- Data Set 1: fits the linear regression model pretty well
- Data Set 2: cannot fit the linear regression model because the data is non-linear
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

Hence it is **importance of visualizing data before applying various algorithms to build models**.

---

3. What is Pearson's R?

Answer:

- In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's R,** measures the linear correlation between two variables.
- It also mentions whether there is a statistically significance relationship between any two variables.
- It also mentions about how 2variables are strongly related to each other and it is sensitive to outliers.

Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

- If the correlation coefficient is -1, it indicates a strong negative relationship. It implies a perfect negative relationship between the variables.
- If the correlation coefficient is 0, it indicates no relationship.
- If the correlation coefficient is 1, it indicates a strong positive relationship. It implies a perfect positive relationship between the variables.

Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range.
- If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.
- To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- Normalization typically means rescales the values into a range of [0,1].
- Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Answer:

- Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating `VIF` is:

$$\text{VIF} = \frac{1}{1-R^2}$$

- 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

---

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

Answer:

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile.
- For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.
- A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
- Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
