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:

There are 5 categorical variables present in the dataset = **"weathersit, season, weekday, mnth, yr"**

Category1:
Weathersit 1: Clear, Few clouds, Partly cloudy, Partly cloudy This above set up had highest number of bike rentals, whereas weathersit 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) had less number of bike rentals

Category2:
season3: Fall had the highest number of bike rentals
season1: Spring had the lowest number of bike rentals.

Category3:
Weekday: More bikes were rented out on Friday, and second highest was seen on Thursday.

Category4:
Mnth: August, September and October had the highest bike rentals. Also these are the months where Fall season occurs in the US.

Category5:
Yr: Highest number of bikes were rented in 2019

---

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

Answer:

- It is to avoid the multicollinearity among the datasets.
- We create dummies for the different features for which 1 and 0 valued columns are assigned.

- Removing the first column with value “1” is dropped to maintain the uniformity in the dummy variables created.

---

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

Answer:
“atemp” and **“temp”** are the two numerical variables that are highly correlated with target variables

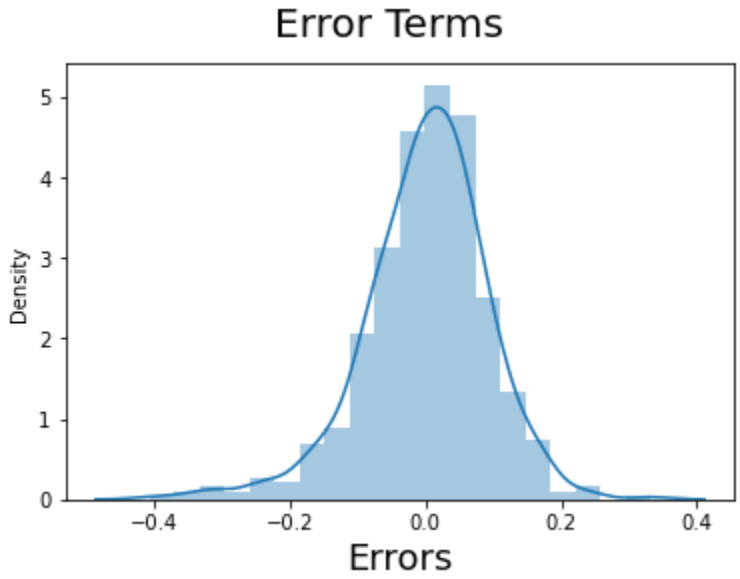
---

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

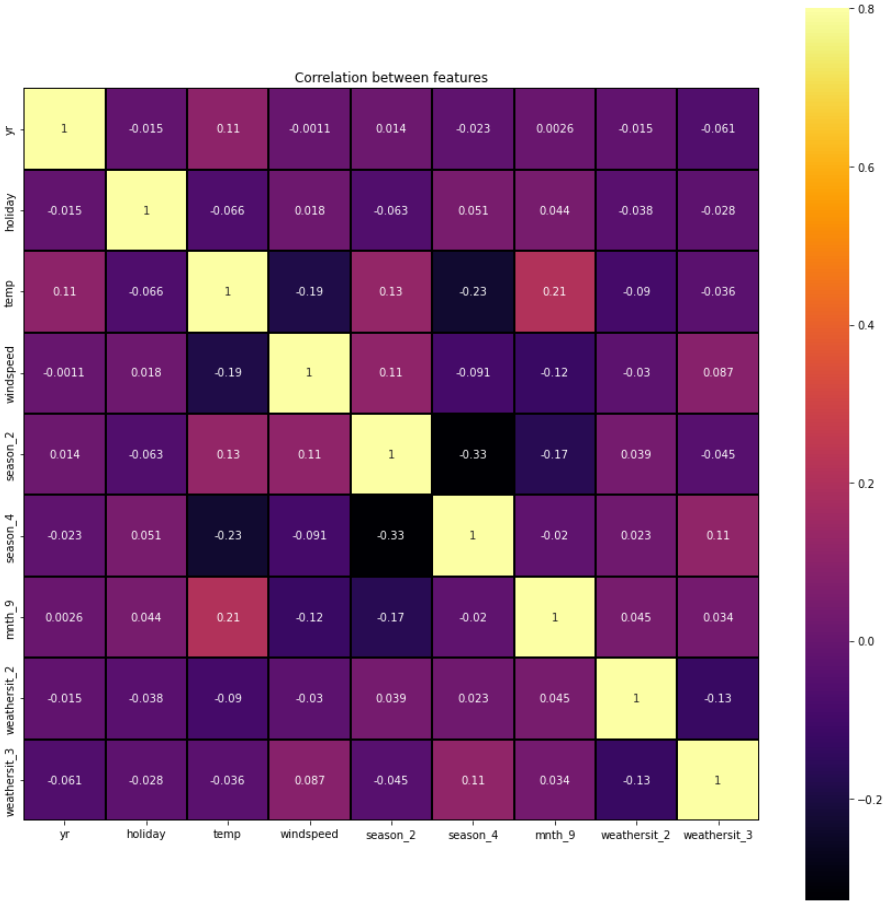
Checking R2 squared: **83.1%**

Higher the F-Statistic score, better the model. **245.1**

Checking the residual is normally distributed, **Residual are normally distributed**



Checking the multicollinearity, it should not be present in the trained model



Durbin watson: Autocorrelation check

Value **2.08** shows there is 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)

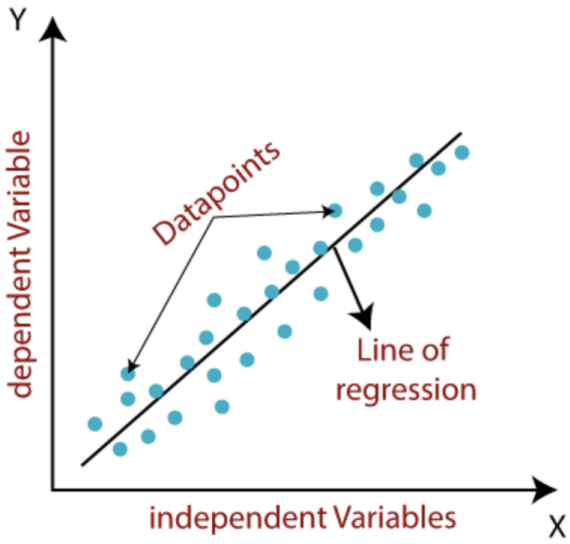
Answer:

1. Year
2. Temp
3. Windspeed

General Subjective Questions

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

Linear regression is a method in the Machine Learning process used majorly to find out the relationship between dependent variable and single/multiple independent variables.



Y = mX+C

Y = is the dependent variable
X = is the independent variable
m= slope
C = intercept

Type of linear regression:

- Simple Linear Regression
- Multiple Linear Regression

Linear regression line:

Positive Linear Relationship: If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship

Negative Linear Relationship: If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Finding the best fit Line:

- The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

Cost function: An optimization problem seeks to minimize a loss function

Gradient descent: Gradient descent (GD) is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given functio

Assumption of Linear regression:

Linear regression assumes the linear relationship between the dependent and independent variables

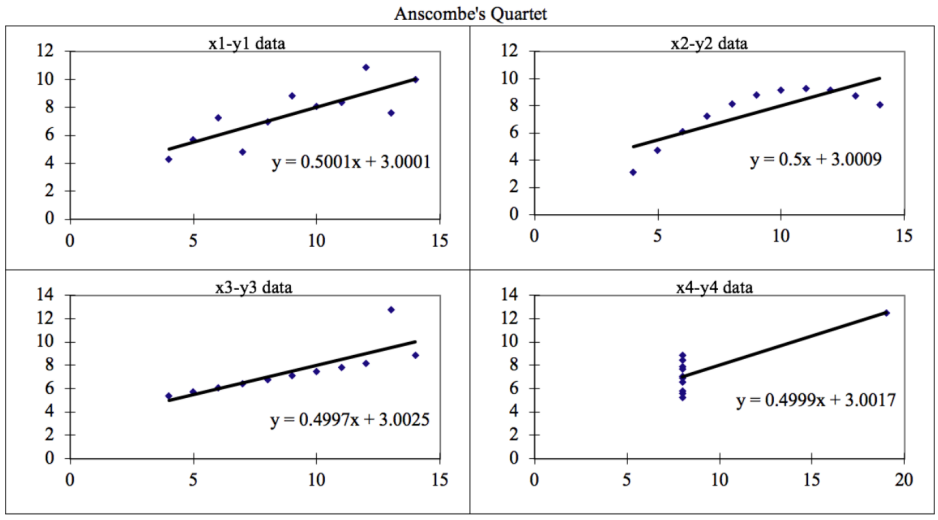
- Small or no multicollinearity between the features
- Homoscedasticity is a situation when the error term is the same for all the values of independent variables
- Normal distribution of error terms
- No autocorrelations

---

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

It contains four groups of datasets with 11 points each in (x,y). Datasets are simple and descriptive. But when plotted, it shows a different distribution and appears different in the graph

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	



---

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

Answer:

Pearson's R is the measure of linear correlation between two sets of data.

P's R value varies between -1 and +1,

Negative correlation: -1

No correlation: 0

Positive correlation: +1

Formulae: $\frac{\Sigma (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})2 (y_i - \bar{y})2}}$

r = correlation coefficient
Xi = x-values in the sample
$\bar{x}$ = mean
Yi = x-values in the sample
$\bar{y}$ = mean of y variable

---

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

Answer:

Scaling is a pre-processing step used to generalize the value of independent variables. The value of the variables ranges from low-high causing the model to take any value. Hence these variables must be scaled to one value eg., mean, median, mode or between the particular range.
This affects coefficients and not any other parameters of the variables.

Difference:
Normalization: The data depicted in the range between 0 to 1

(Min-max scaling) Normalization: $x = \frac{x - min(x)}{max(x) - min(x}$

Standardization: This replaces the value by Z-scores. This standardized data depicts the normal distribution with mean and SD

Standardization: $x = \frac{x - mean(x)}{sd(x)}$

---

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

Answer:

VIF is Variance inflation factor. It is used to measure the multicollinearity of independent variables in multiple regression analysis.

It is defined as, VIF = 1/1-Ri2

Variables that are infinite in VIF show that they are highly correlated. Such variables must be removed.
VIF less than 5 are acceptable, greater than 5 is not acceptable.

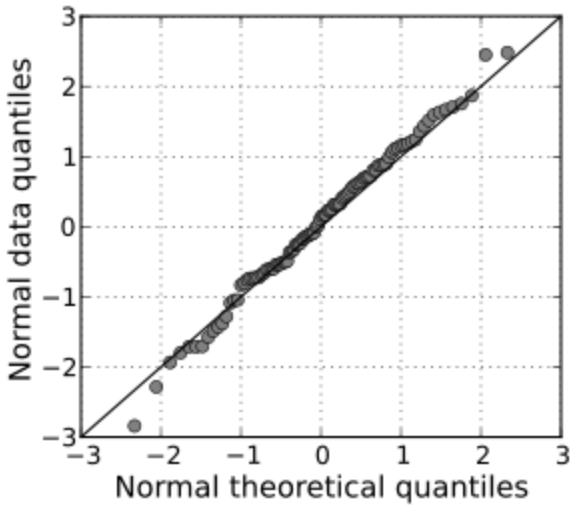
---

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

Answer:

The Q-Q plot is defined as a quantile-quantile (q-q) plot. It is basically used in machine learning to find out the datasets come from the common distribution.

Q-Q plot is the quantiles of a sample distribution against quantiles of a theoretical distribution in a graphical manner. This shows whether the dataset follows any particular type of probability distribution like normal, uniform, exponential.


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3