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 EDA, below is the inference :

- Bike demand is higher in fall season
- Bike demand is higher in the year 2019
- Bike demand is higher between the months June – Oct
- Bike demand is lower on a holiday. Indicates that people don’t have any reason to rent a bike on a holiday.
- Bike demand is similar throughout the weekdays.
- Bike demand is same whether working day or not
- Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow

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

Answer:

After we create dummy_variables, the first column can be dropped because the first column values can be derived from the other dummy columns.

For example, weekday column has 7 values, 1 for each day in a week. When we use one hot coding to get dummy variables, we create a column each for each day. If it is the day of the column, then column has value 1 else 0. In such a scenario, we can derive whether it is day 1 if it is not any other day (day 2 – day 7). By doing this we can reduce the number of features which would reduce the number of features. It is important to do this because this column would have had high collinearity if retained.

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

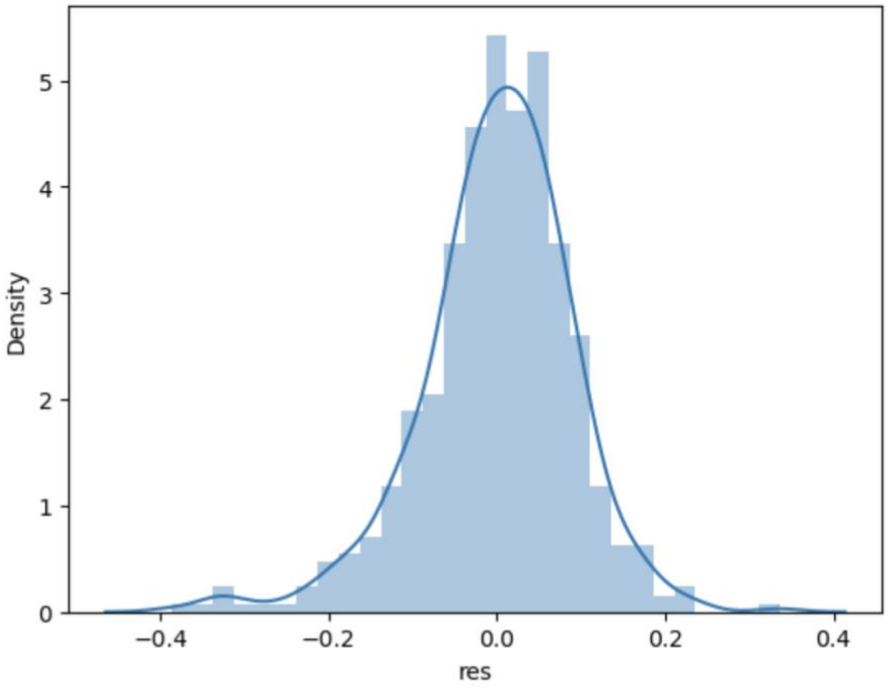
Answer:

The numerical variables “atemp” and “temp” have the highest correlation of 0.63 with the target variable “cnt”

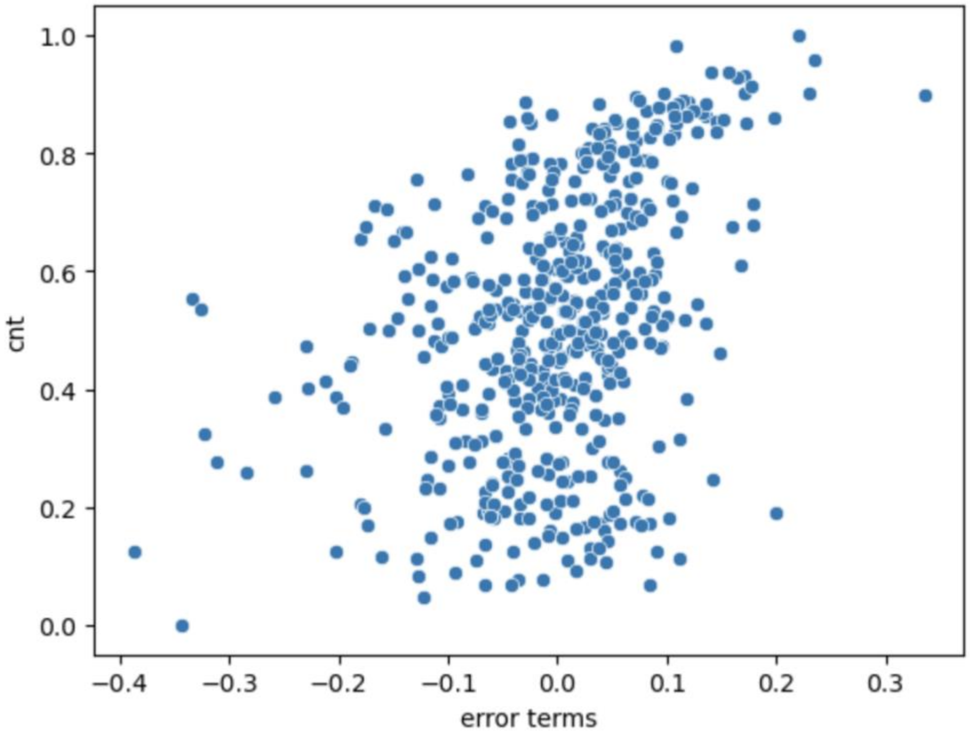
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Did residual analysis to validate that the error terms are normally distributed and has a mean of approximately 0.



There is no corelation between the residuals and the target variable



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 model equation:

cnt = 0.23*const + 0.54*temp - 0.18*hum - 0.18*windspeed + 0.10*season_2 + 0.14*season_4 + 0.23*yr_1 + 0.05*mnth_8 + 0.12*mnth_9 - 0.10*holiday_1 - 0.054*weathersit_2 - 0.24*weathersit_3

Top 3 features are:

- Temp => a unit increase in temperature, increases the bike demand by 0.54
- Weathersit_3 (Light snow, light rain+ thunderstorm) => a unit increase in weathersit_3, decreases the bike demand by 0.24
- Yr_1 (the year 2019) => if the year is 2019, the bike demand increases by 0.23

General Subjective Questions

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

Answer: Linear regression is a statistical algorithm that predicts values based on the model learnt by assuming a linear relationship between the feature variables and the target variable. There are certain assumptions :

- There should be a linear relationship between the feature variables and target variable
- Error terms (residuals) are normally distributed
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

If the number of features are more than 1, it is called multiple linear regression.

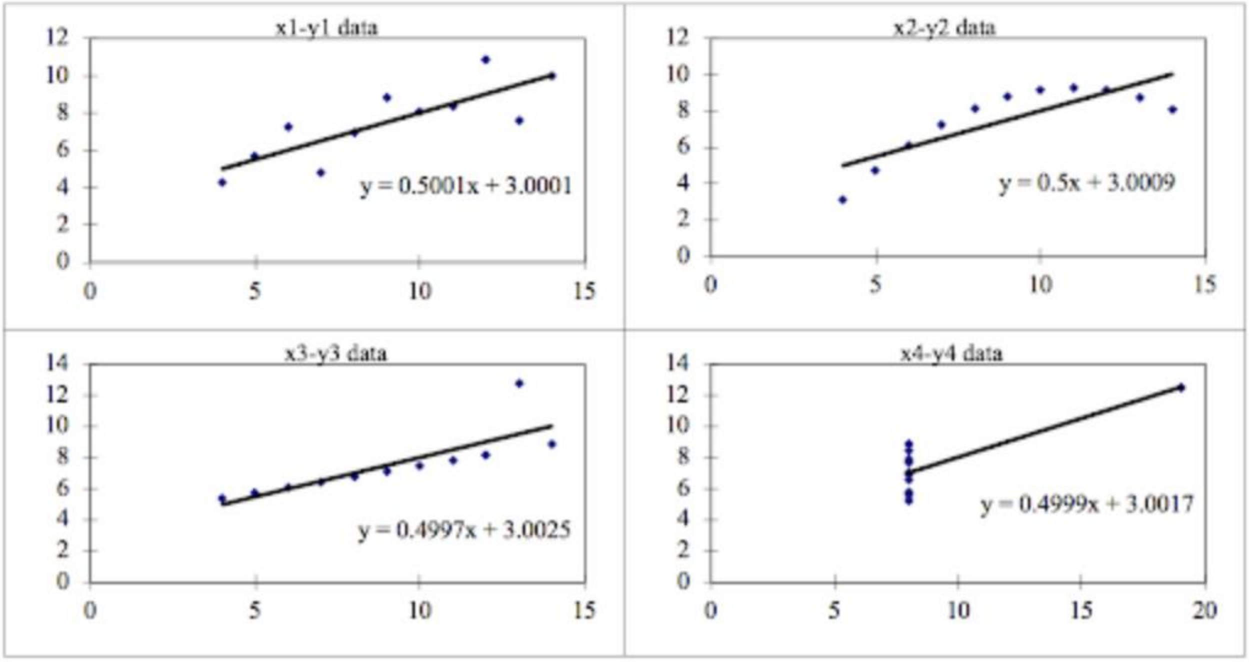
Linear regression algorithm :

- Load data
- Pre-processing of data (check data types, null values, outliers, derived variables)
- EDA (analyse the feature variables to better understand the data we are going to model)
- Handling categorical features using one hot encoding
- Scaling of features so that the coefficients are interpretable and also for faster convergence of the gradient descent algorithm.

- Feature selection & modelling using sklearn and statsmodel library to drop features which have high p value (> 0.05) or high VIF (> 5)
- Residual analysis to test if the error terms are normally distributed to validate that our assumption is correct
- Model interpretation by checking which feature positively or negatively affects the target variable by interpreting the coefficients.

2. Explain the Anscombe's quartet in detail. (4 marks)

Answer: It is a set of 4 datasets which illustrates the importance of why graphing the data is important before analysing it. The datasets have the same mean, standard deviation and regression line, but however there distributions are qualitatively different.



- **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.

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

Answer: Pearson corelation coefficient R summarises the strength and direction of linear relationship between 2 variables.

For example a R > .5 indicates a strong positive corelation, whereas a R < .5 indicates a strong negative corelation. Between .3 and .5, it is a moderate strength of corelation, and if R is between 0 and .3, is a weak strength of corelation.

A corelation of 0 indicates that there is no dependency of one variable with the other. The variables are related completely randomly.

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

Answer: Scaling is a process of getting all the feature variables values within one common range. It is specifically performed in multiple linear regression to make it easy to interpret the coefficients of the variable in the final model. If we don't scale for example, if a column has all high values, obviously, the coefficient will be a little small in value compared to other columns. This doesn't indicate that it is of lesser importance though. To avoid this confusion, all variables are scaled so that the coefficients of the final model can be compared.

Another mathematical reason to do scaling is that it helps the gradient descent function to converge faster. This helps in faster model building.

There are 2 types of scaling

- Normalised scaling or min-max scaling puts all data between 0 & 1 ((x - xmin)/(xmax - xmin))
- standardization – puts all data such that mean=0, sigma=1. ((x-mu)/sigma)

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

Answer:

Where there is a perfect corelation between 2 variables, the VIF will be infinite. This can be understood by the VIF formula

VIF = 1/(1-R^2).

If there is a perfect corelation , R^2 = 1

VIF = 1/(1-1) = 1 / 0 = infinity

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

Answer:

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

- If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.