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 the final model “yr” has +ve coefficient. Which means improving sales as year passes.
- Season: The ‘fall’ season attracts more rides while ‘spring’ the least.
- Mnth: If we observe the median number of rides across different months, It peaks around the time July-September. Which also confirms our previous point.
- Holiday: As expected, the median number of rides on a holiday are less compared to Non-holiday.
- Workingday, weekday: There is not much difference in the median number of rides on a working day compared to a Non- working day. This is also evident from the coefficient of the ` Workingday` in the final model.
- Weathersit: There is a clear pattern here. Rainy/ thunderstorm days attract considerably low number of rides. This is also evident from the negative coefficient values in the final model.

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

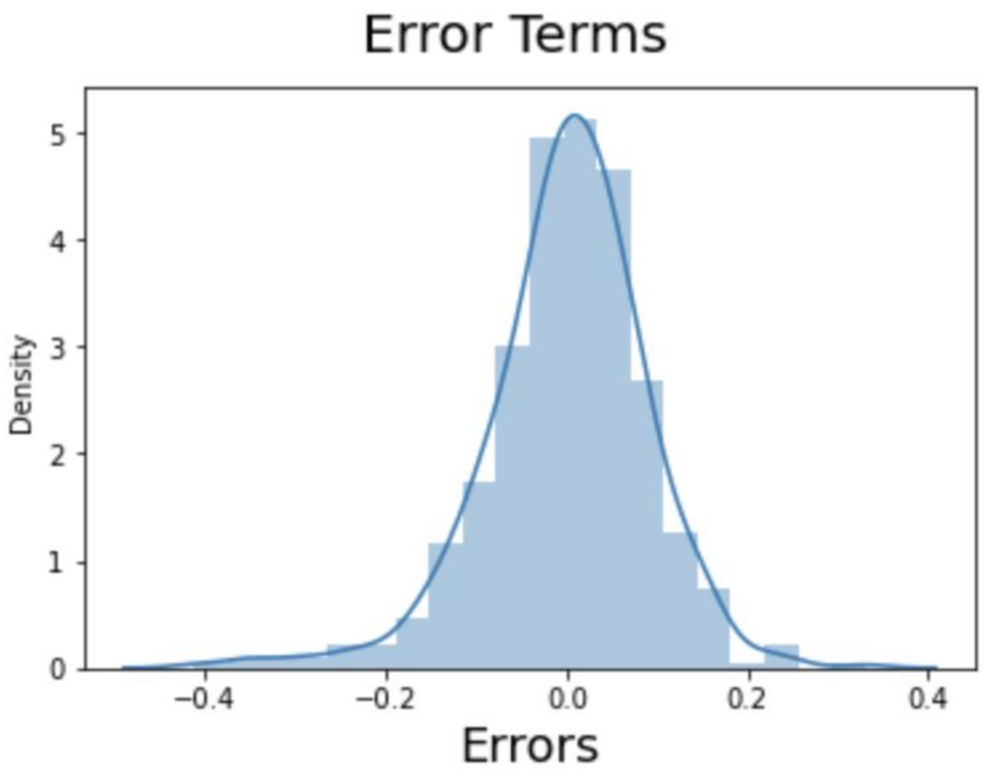
Answer:
For a categorical variable with n possible values, we only need n-1 dummy variables to represent. To achieve the same, we use **drop_first=True.**
We can observe the same being followed for all categorical variable dummy creation in our model building.

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

Answer: “temp” and “atemp” have the highest correlation with the target variable. Their correlation values are “0.64” and “0.65” respectively.

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

Answer: The main assumption of Linear Regression is that `error terms are normally distributed and centered around 0.` We did validate this after our model building.
Below is the screenshot:



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

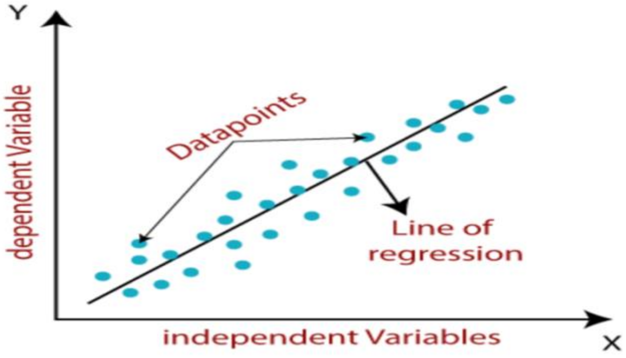
Answer:

By observing the coefficients of the parameters from the final model, we can see `temp` , `weathersit` and `yr` are contributing the most.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Regression is a supervised learning technique to statistically model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. Linear regression models this relationship in a linear manner, hence called linear regression. Below is the figure depicting a simple linear regression involving one independent variable:



As you can see in the above picture, the goal is to find the best fit line which minimizes the error between predicted values and actual. Here the cost function is Residual Sum of Squares. “Gradient descent” algorithm is used to minimize the cost function and finally arrive at the optimized coefficients.

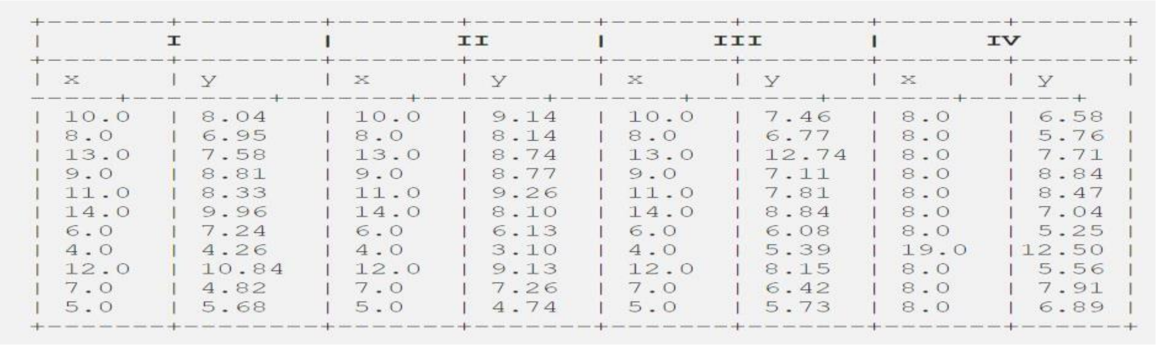
Assumptions of Linear Regression:

- Linear relationship between the features and target.
- Small or no multicollinearity between the features.
- Normal distribution of error terms.
- No clear pattern in the distribution of error terms.

We finally evaluate the arrived model using various statistical parameters (Adjusted R-squared, p-value of coefficients, Prob(F-stat) etc.)

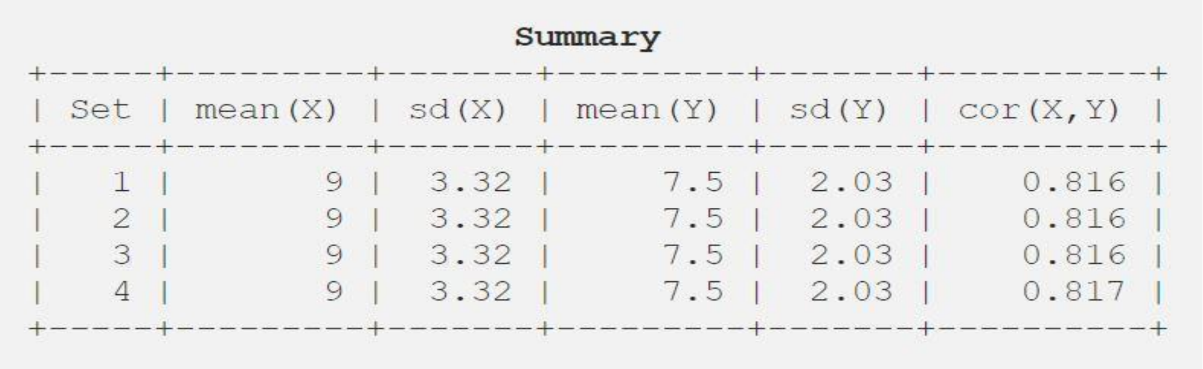
2. Explain the Anscombe's quartet in detail?

Answer: Anscombe’s quartet comprises four datasets of 11-data points each that have identical simple statistical properties, yet appear very different when graphed. Below are those data points:



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

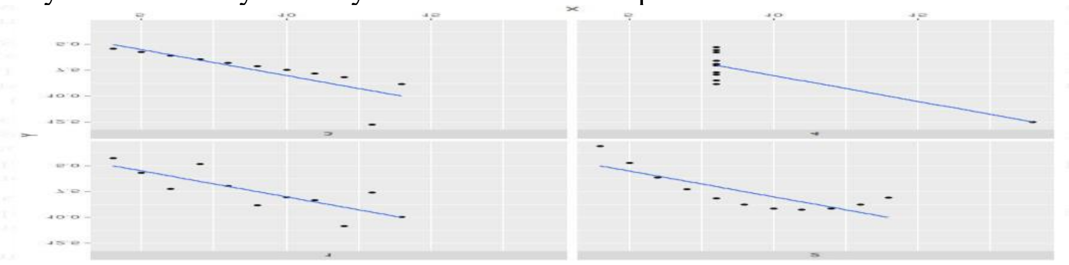
Below are the statistical properties of the above data sets:



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

As you can see they are very identical. Let’s now plot the dataset and see their distribution:



Below are some observations:

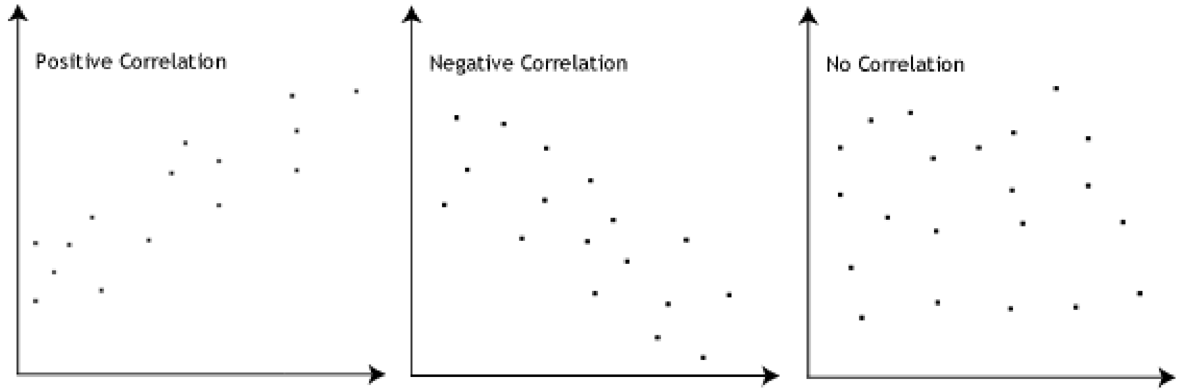
- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

This illustrates the importance of looking at a dataset graphically before starting to analyse.

3. What is Pearson's R?

Answer: The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. It attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

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. A value less than 0 indicates a negative association.



This correlation makes below assumptions about the data:

1. Two variables should be measured on a continuous scale.
2. Two variables should be paired.
3. Observations should be independent.
4. There should be a linear relationship.

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

Answer: Scaling is generally applied on independent numerical variables. The data set for this variable can have huge variation in the ranges. Likewise multiple independent variables can have very different ranges.

If model building is performed on this. It can lead to below things:

1. It makes it hard for the model training algorithm (like Gradient Descent) to converge on optimized set of model parameters.
2. It becomes difficult to assess which parameters affect the most if they follow different ranges.

So this method of optimizing different ranges is called as scaling.

Below are the two popular types of scaling methods:

1. Min-Max scaling.
2. Standard scaling

Min-Max scaling compresses the variable range between [0,1] or [-1, 1] depending on the data. It is useful when there are no outliers in the data.

While Standard scaling uses mean and standard deviation.

Normalized scaling:

X_scaled = (X - X_min)/(X_max - X_min)

Standardized Scaling:

X_ scaled = (X – X_mean)/X_Std

Apart from these there are other below less used scalars as well:

1. Max Abs Scalar
2. Robust Scalar
3. Quantile Transformer Scalar
4. Unit Vector Scalar

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

Answer: It indicates that there is perfect correlation between the variables. This means you can exactly derives the value of this independent variable with others. This also indicates high multicollinearity. One of the best ways to solve this is to drop another variable and re-evaluate VIF.

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

Answer: Q–Q plot (quantile-quantile plot) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. By quantile, it means the fraction (or percent) of points below the given 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 that the two data sets have come from populations with different distributions.

We can plot y_train and y_test in linear regression to compare their distributions.