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 1. The Categorical variables identified from the data set are

‘season’ – Seasons throughout the year (1:spring, 2:summer, 3:fall, 4:winter)

‘mnth’- Month of the year.

'weathersit' – Different weather situation defined as follows

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

'weekday' - Day of the week.

Effect on dependent variable –

- **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

- **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

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

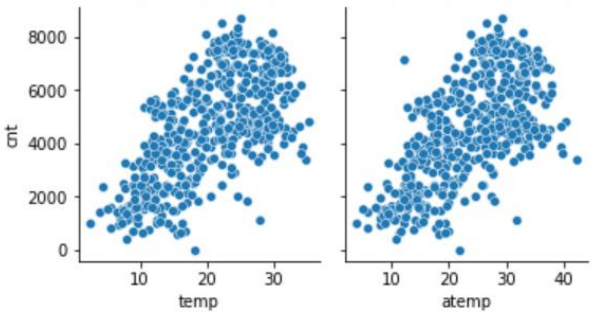
Answer 2. When creating dummy variables, we only need n-1 dummy variables to represent n variables.

So to reduce the number of dummy variable we use drop_first = True.

Suppose you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

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

Answer 3. Variable with the highest correlation with target variable ‘cnt ’ are ‘temp’ and ‘atemp’



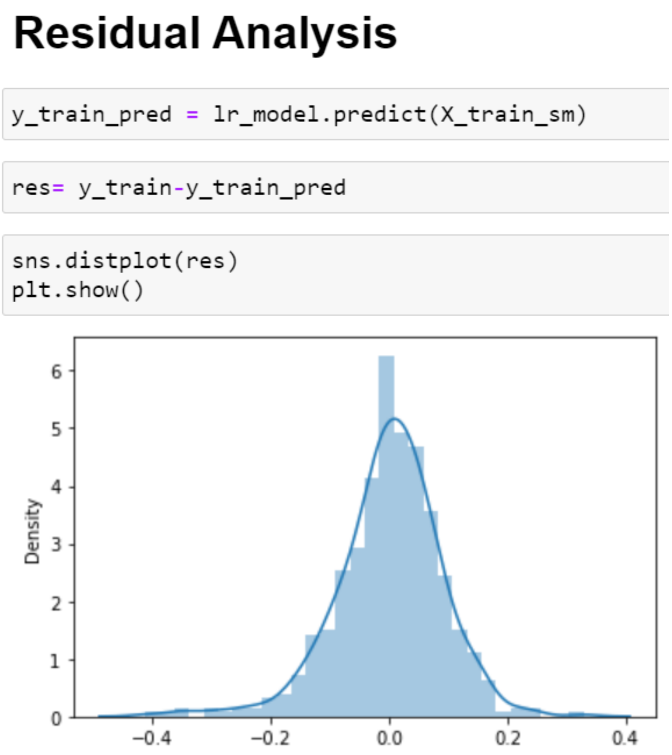
When we found out the correlation on training data set the correlation value with ‘cnt’ for both ‘temp’ & ‘atemp’ was 0.63

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

Answer 4. The most important assumption of a linear regression model is that the ***errors are independent and normally distributed****.*

Assumption of Linear Regression after building the model on the training set was validated by calculating residual points (y actual – y predicted) and plotting their distribution.

The plot was normally distributed with mean centered around zero, thus validating our linear regression model.



Residual Analysis

```
y_train_pred = lr_model.predict(X_train_sm)
```

```
res= y_train-y_train_pred
```

```
sns.distplot(res)
plt.show()
```

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

Answer 5.

The final model equation is

cnt = 0.08 + 0.23 yr + 0.05 workingday + 0.52 temp - 0.15 windspeed + 0.10 season_2 + 0.13 season_4 + 0.05 mnth_8 + 0.11 mnth_9 + 0.06 weekday_6 - 0.08 weathersit_2 - 0.28 weathersit_3

So from the above equation we can validate that **‘temp’, ‘yr’ & ‘weathersit’** are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

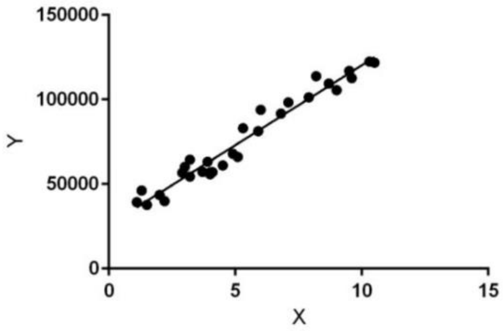
Answer-1

Regression is a statistical technique that shows an algebraic relationship between two or more variables.

Based on this algebraic relationship, one can estimate the value of a variable, given the values of the other variables.

When this algebraic relationship is of linear nature such that it is defined by, **y = mx + c**, we follow linear regression algorithm.

Where, (x,y) are co-ordinates on X and Y- axis respectively, m is the slope of the line and c is the interception of line on Y axis.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Linear Regression may further be classified into -

- Simple Linear Regression/ Univariate Linear regression (single independent variable).
- Multivariate Linear Regression (Multiple independent variable).

To perform Linear Regression –

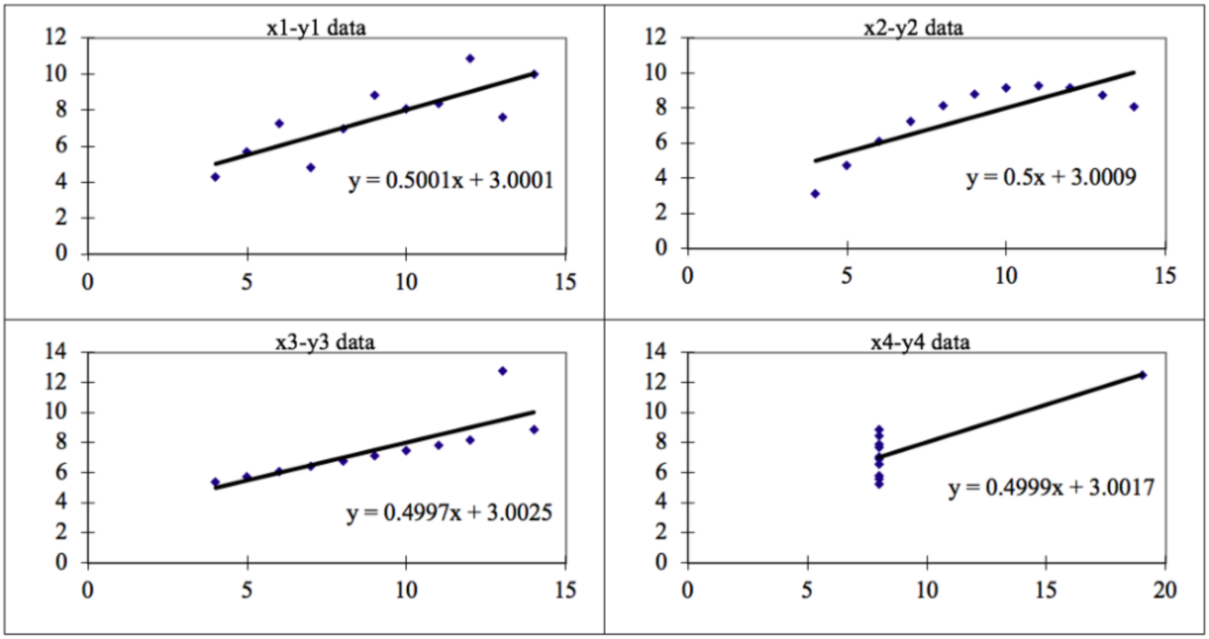
- We divide the given data set in train and test data set.
- With the help of train data set we find the best-fit line using suitable model, which passes through maximum number of data points.
- With the equation of the best fit line (y = mx + c) we can predict dependent variables and compare them with given values of test data set to find the credibility of our best-fit line. And can use it to make further predictions.

2. Explain the Anscombe's quartet in detail.

Answer 2.

Anscombe's Quartet can be defined as **a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model** if built. They have very different distributions and appear differently when plotted on scatter plots.

This tells us about the **importance of visualizing the data before applying various algorithms** out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



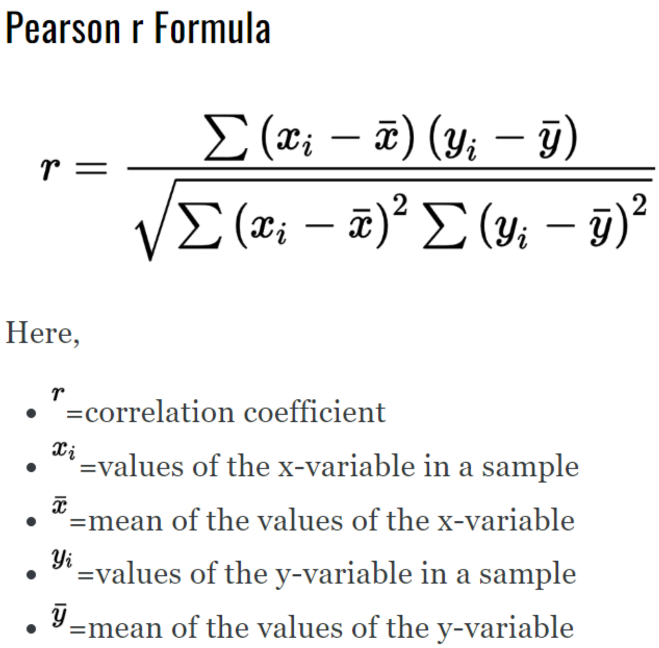
1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson’s R?

Answer 3.

The Pearson's R also called Pearson correlation coefficient is a **Product Moment Correlation** which gives numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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



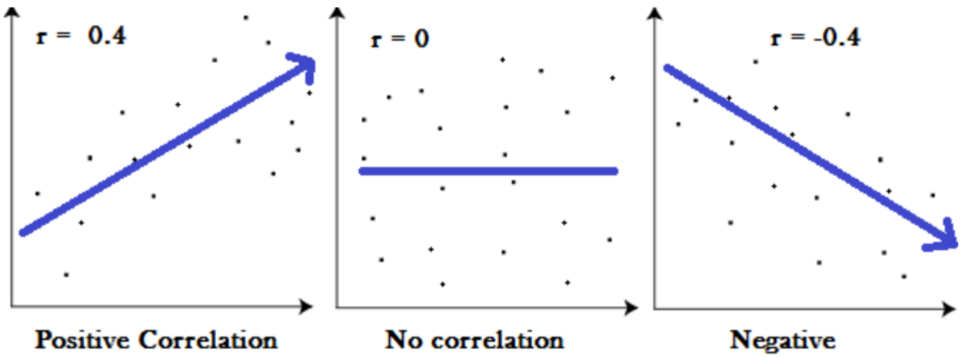
Pearson r Formula

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

- If R is 1, indicates a strong positive relationship.
- If R is -1, indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



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

Answer 4.

The Raw data set provided may have huge variations between the ranges of different variables, if the model is built on such variables the coefficients for these variables will be having huge difference as well.

The method of optimizing this ranges difference of variables is know as scaling.

There are two different types of scaling-

Normalized scaling and Standardized scaling, the difference in both these scaling is that the Normalized scaling compresses the variable range between [0,1] and sometimes also in range [-1, 1] and is more effective when there are no outliers, whereas Standardized scaling doesn’t exactly compress the range but centers the mean around zero with s standard deviation of 1.

Normalized scaling
X_new = (X - X_min)/(X_max - X_min)

Standardized Scaling

X_new = (X - mean)/Std

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

Answer 5.

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 6.

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

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.

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.

In linear regression we can plot y_test_predict and y_test to check how effective is the linear model that we have built.