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: Done the analysis using box plot and bar plot for categorical variables. Here are few observations:

A) Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.

B) Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

C) Clear weather attracted more booking which seems obvious.

D) Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.

E) When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.

F) Booking seemed to be almost equal either on working day or non-working day.

G) 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

H) The working day and holiday plot indicates that more bikes are rent during normal working days than on weekends or holidays.

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Answer: The 'temp' variable 'registered' has the highest correlation with the target variable 'cnt' .

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

Answer: I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- ✓ Normality of error terms
  - Error terms should be normally distributed
- ✓ Multicollinearity check
  - There should be insignificant multicollinearity among variables.
- ✓ Linear relationship validation
  - Linearity should be visible among variables
- ✓ Homoscedasticity
  - There should be no visible pattern in residual values.
- ✓ Independence of residuals
  - No auto-correlation

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

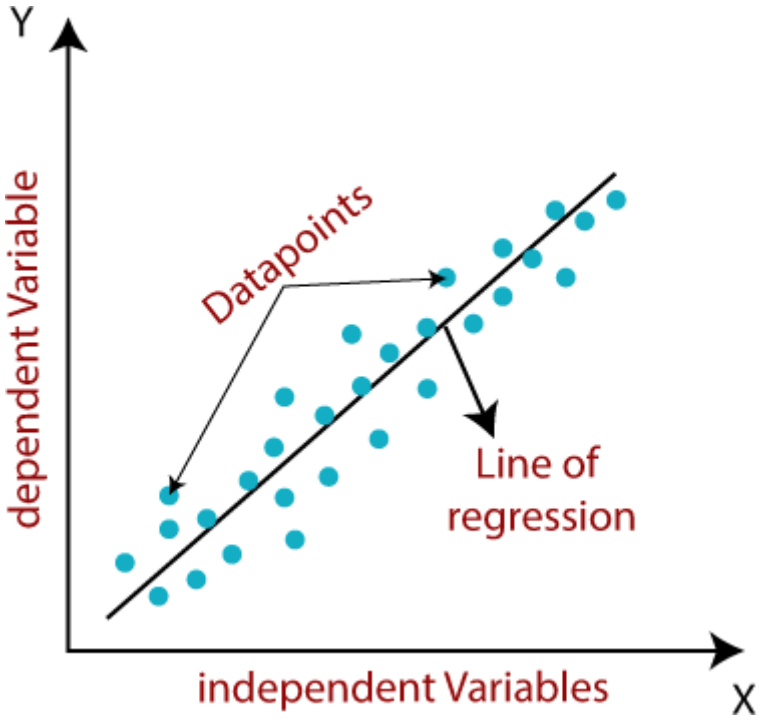
Answer: The Top 3 features contributing significantly towards the demands of share bikes are: temp, season & weathersit

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer: Linear regression algorithms is supervised learning algorithm which is applicable for continuous target variable. The basic condition of this algo is there must be some linear relation between independent and dependent variables. There are few more assumption which are for this algorithm as part of answer in abo

above question answer .Linear regression is one of the basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. It is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.



We represent the simple line of regression by an equation:
Y= B0 + B1X
B0 is constant and it will be zero when line passes through origin.
B1 is coefficient of X which is also known as slope of line.

If there are multiple lines in a plane then same equation will be represented as
Y = Y= B0 + B1X1 + B2X2 +……

The cost function of regression line RSS (Residual Sum of Squares) which we try to minimize to keep the residual as minimum as possible. This is reason we try to find the B coefficient for the minimum RSS.

Basically when we train the dataset then we are allowing this regression line to learn and create the optimum B coefficient for regression line..

Use Cases of Linear Regression:

- Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
- Price Prediction – Using regression to predict the change in price of stock or product.
- Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

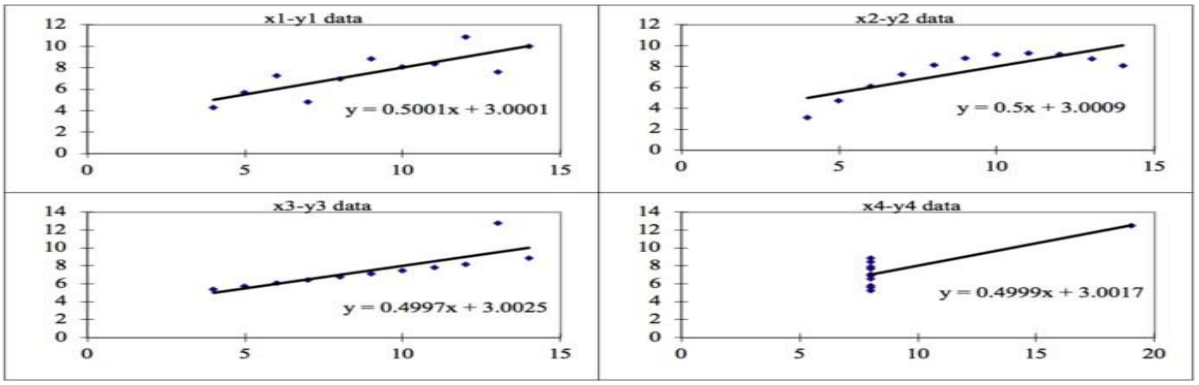
2.Explain the Anscombe's quartet in detail.

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 analysing it and the effect of outliers on statistical properties ,diversity of the data, linear separability of the data, etc.

When these models are plotted on a scatter plot, all datasets generates a different kind of plot thatis not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

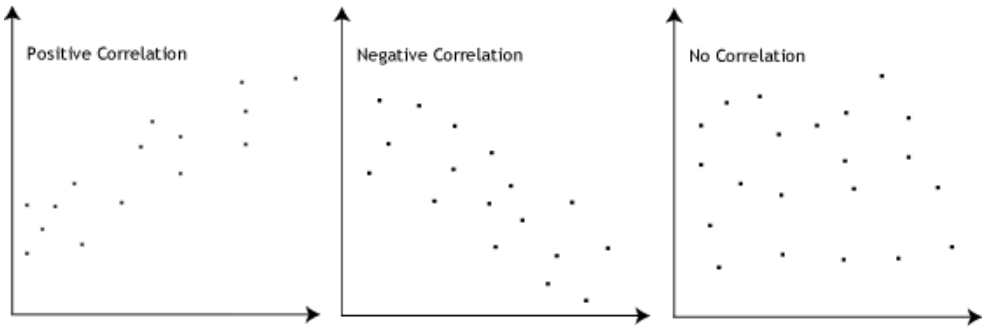
2. Dataset 1: this fits the linear regression model pretty well.
3. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
4. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
5. Dataset 4: shows the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

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Answer: Pearson's r is a 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.

The Pearson correlation coefficient, 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; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.

r = 1 means the data is perfectly linear with a positive slope

(i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope

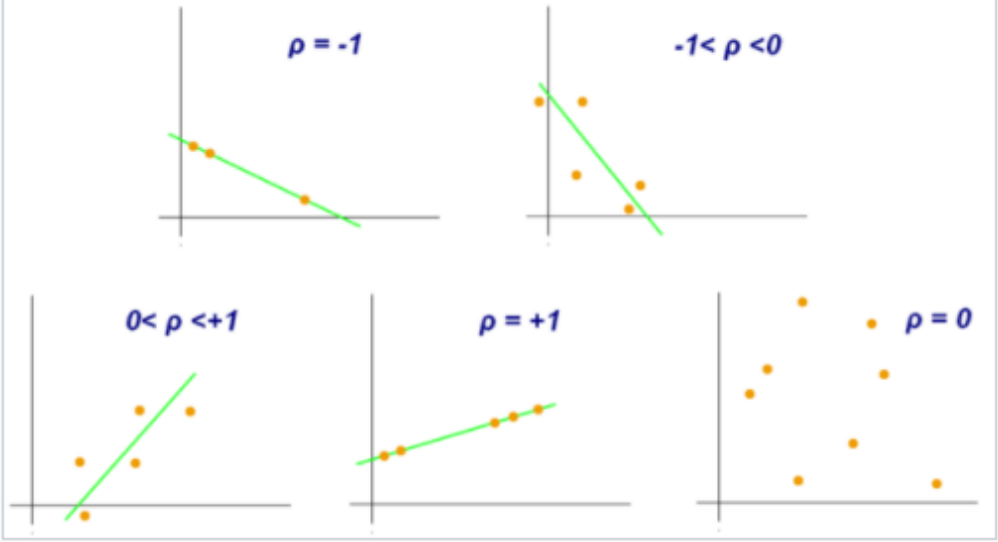
(i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < .5 means there is a weak association

r > .5 < .8 means there is a moderate association

r > .8 means there is a strong association Formula



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Using the formula proposed by Karl Pearson, we can calculate a **linear relationship** between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- ☐ Scale of measurement should be interval or ratio
- ☐ Variables should be approximately normally distributed
- ☐ The association should be linear
- ☐ There should be no outliers in the data

Formula given is:

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

Wheíe,

N = the numbeí of paiís of scoíes

Σxy = the sum of the píoducts of paiíed scoíes

Σx = the sum of x scoíes

Σy = the sum of y scoíes

Σx2 = the sum of squaíed x scoíes

Σy2 = the sum of squaíed y scoíes

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

Answer: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range or on the same scale. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
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. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

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

This scales the range to [0, 1] or sometimes [-1, 1]

Normalization is useful when there are no outliers as it cannot cope up with them.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std
Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

Answer: 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 (which show an infinite VIF as well).

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. A Q Q plot showing the 45 degree 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.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that 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.

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.

The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution or two data sets have come from populations with different distributions.