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?**

The demand for shared bikes during spring season is more compared to other seasons

The demand for shared bikes is more during months 3, 4, 5, 6, 7, 8, 9, 10 compared to other months

The demand for bikes decreases with windspeed, spring, Light Snow, (LightRain+Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) conditions

The demand for bikes is more on Saturday and working day

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

drop_first=True drops the extra column created during dummy variable creation, thereby it reduces the correlations created among dummy variables.

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

temp and atemp are having highest correlation as observed from pair plot.

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

Observing at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others and calculating a score as follows:
VIF = 1/(1-R^2).
Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large R-squared for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multi co-linearity (above 10 it can definitely be considered an issue).

- One more method is checking the P value, if P value is high we can say that it is statistically less significant.

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

Weathersit (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) has a negative correlation
Working day and Saturday has Positive correlation
Spring Season has nnegative correlation

General Subjective Questions

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

A: Linear Regression is a machine learning algorithm based on supervised learning. It

Performs a regression task. Regression models a target prediction value based on independent Variables. It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and

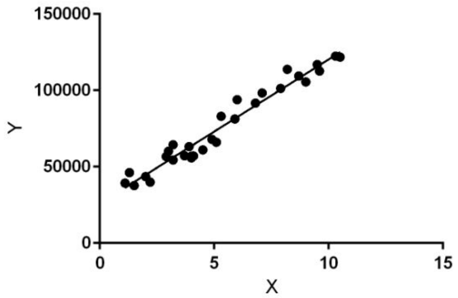
Independent variables, they are considering and the number of independent variables being

Used.

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.

y = a1 +a2.x

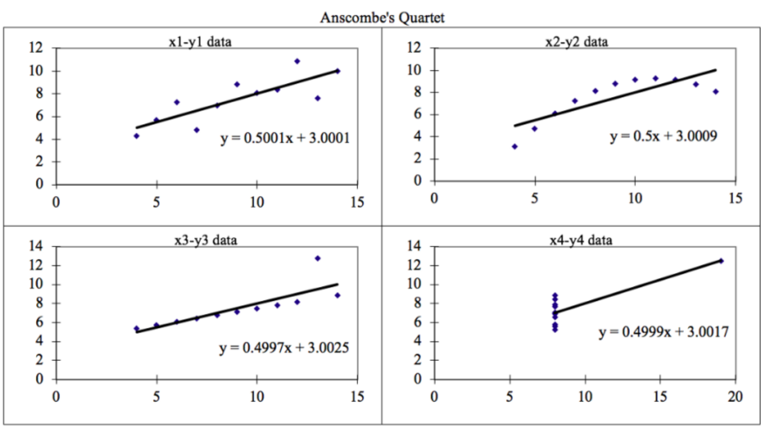
Here, a1 is intercept, a2 is the coefficient of x, x: input training data, y: labels to data

2. **Explain the Anscombe's quartet in detail.**

A: 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.

Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties There are these four data set plots which have nearly same statistical observations, which provides same
statistical information that involves variance, and mean of all x,y points in all four datasets.This
tells us about the importance of visualising 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

The four datasets can be described as



Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

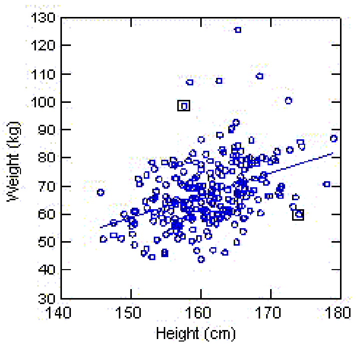
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson’s R?

A: 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.



"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women

shows. The correlation coefficient is positive and height and weight tend to go up and down together. yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrates.

The Pearson's correlation coefficient varies between -1 and +1 where:

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

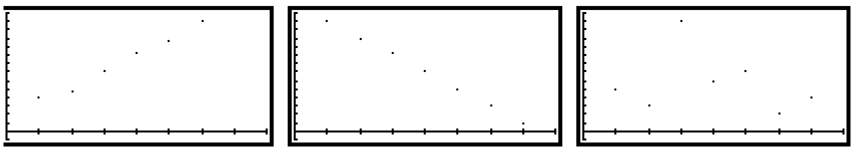
r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

The figure below shows some data sets and their correlation coefficients. The first data set has

an r=0.996, the second has an r = -0.999 and the third has an r= -0.233



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

A: normalizing the independent variables within a particular range for ease and speeding of algorithm calculations is called scaling

some times data has higher values or magnitudes compared to other variables, to bring all the variables into same magnitude scaling is performed because
it does not affect parameters like p-values, R-squared etc.

min-max scaling brings the total available data between 0 and 1
Minimum and maximum value of features are used for scaling
it is affected by outliers

standardisation brings all of the data into standard normal distribution with mean 0 and standard deviation 1

Mean and standard deviation is used for scaling

it is not affected by outliers

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

A: The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

VIF = 1/1-R^2

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 . The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution