Explain the linear regression algorithm in detail
Liner regression is a statistical technique used for finding the existence of an association relation ship between a dependent variable and an independent varaible. We can only establish that chnage in the value of the outcome varaible(Y) is associated with chnage in the value of feature X, i.e., regression technique cannot be used for establishing casual relation ship between two variables.

Regression is one of the most popular supervised learning algorithms in the predictive analytics. A regression model requires the knowledge of the both outcome and the feature variables in the training dataset.

Explain the Anscombe's quartet in detail.
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 analyzing it and the effect of outliers on statistical properties.
Sample Code:

```
library(xtable)
# Create the interactive table
ansombe.table <- xtable(anscombe)
print(ansombe.table, "html")
```

What is Pearson's R?
Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.
The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent variable is plotted on the x-axis (horizontally) and the dependent variable is plotted on the y-axis (vertically).
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1

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

scaling is an important technique in Machine Learning and it is one of the most important steps during the preprocessing of data before creating a machine learning model. This can make a difference between a weak machine learning model and a strong one. They two most important scaling techniques is Standardization and Normalization.
Normalization: Normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.
Standardization: typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
A good reason to perform features scaling is to ensure one feature doesn't dominate others.

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

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).

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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance of Q-Q plot:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets -

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.