Question 1:
Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train accuracies, and how can this problem be solved?

Solution:
1. Model accuracy is too high on the training dataset because the model generated has memorised the dataset by generating a complex equation which passes through almost all the points on the training dataset.
2. Generated equation is too complex and doesn't account for general trend in data.
3. The test dataset needs a more generic equation for prediction.
4. In this case, the variance is too high while the bias is too low.
5. This problem can be solved by, reducing the complexity of equation by using Regularisation with optimum hyper parameter and removing multi-collinearity.

Question 2:
List at least four differences in detail between L1 and L2 regularisation in regression.

Solution:
1. L1 also know as 'Lasso Regression' adds "*squared magnitude*" of coefficient as penalty term to the cost function. While L2 also known as 'Ridge Regression' adds "*absolute value of magnitude*" of coefficient as penalty term to the cost function.

2. L1 regression removes redundant variables automatically. **While** L2 regression do not help in reducing multicollinearity automatically.

3. L1 regression is computationally more intense. **While** L2 regression is computationally less intense.

4. L1 regression is used when the dataset is not too large. **While** L2 can be used in case of very large dataset.

Question 3:
Consider two linear models:
L1: y = 39.76x + 32.648628
And
L2: y = 43.2x + 19.8
Given the fact that both the models perform equally well on the test data set, which one would you prefer and why?

Solution:
1. L2: y = 43.2x + 19.8 shall be preferred more than L1: y = 39.76x + 32.648628 because L1 is more complex than L2 and is likely to give less accuracy on a test dataset while L2 which is more general and less complex equation will give better accuracy when subjected to unknown test dataset.
2. L1: y = 39.76x + 32.648628 has been generated by such that I will pass through most of the datapoint of the train dataset(which makes it complex but good for train dataset score) while L2: y = 43.2x + 19.8 has been generated by such that it captures the general trend of the dataset (which makes it less complex but good for test dataset score)

Question 4:
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Solution:
1. A robust and generalisable model is not too complex.
2. A robust and generalisable model always have balance in Variance and bias.
3. A robust and generalisable model performs well on both train and test dataset while a complex and non generalisable model performs well on train data but not on test data.
4. A robust and generalisable model may have low accuracy score for the train dataset as compared to the complex model but it will always have more accuracy score on test dataset than a complex model.

Question 5:
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Solution:
1. Since, Lasso Regression is computationally more intense. If the dataset is not very large then we shall choose Lasso regression as it automatically removes the redundant variables and removes multicollinearity also.
2. Since, Ridge Regression is computationally less intense. If the dataset is too large then we shall choose Ridge regression as it is capable of handling large data.