1. Project details: Name of your project, your team, and team members.
   Name: Predict the house price using the regression techniques.
   Team: Feng Yongxiang, Ashutosh Anand
2. Problem Statement: What problem are you tackling, and what's the setting you're considering?
   I want to solve the problem that predict the future price based on some samples. There many features in the datasets we can choose. So I consider whether choose the all features or choose some of them. Because some of features may not have equal important to other ones. Such as roof style is not important as the area. But if we don't choose all features in the datasets, it will not represent all respects about the house.
   I will first try to choose linear regression to separate features, see what perfrom of this method. Then I will try some advanced methods to predict the house price. I will try ridge regression to predict the house price, because it adds penalty parameter, so I think it will perform better than just linear regression. Last I will try random forest to predict the house price. It will divide the data to some unit and predict the result independently. So I think this method will result a reliable result.
3. Method: What machine learning techniques (definitely you should use some of the techniques that you are learning in class. However, you can also show results with other methods that you may feel, would be appropriate for your problem) have you tried and why?
   Techniques: we will use random forest to predict the future house price in the dataset. Firstly we choose the features in the datasets. The train data has 80 features which includes area, building year, utilities and etc. We can take these features into forest tree model. The model will random choose some samples belong to the whole datasets and build its own tree. And when some features lack of samples, it will replace average data to the empty one. In this way, we can get many trees separately and it will choose the best fit tree's outcome as its result. But the random forest may lead to overfitting. So we can apply another method to compare with it. It's the ridge regression. In this way, we can catch up the features belong to train data and put them into the ridge regression function. Because of penalty term. It can avoid large coefficients hurt the ultimate result.
   Refer to the test method. We have the test data to test the result of the train model. We can use the cross validation method to test the model's errors. We can divide the test data into some parts, and use the train model to predict the result of them. At last we can get the mean squared errors through this way.
4. Preliminary Results: Results in the state of the art methods and your preliminary achievements
   I have loaded the train data and test data to the jupyter notebook. And I analysis them. I think all features are useful. So I decide to use all features to train the model.
5. Dataset: What are the datasets that you have used, and why did you decide to choose them?

I will use datasets in the Kaggle competition. Because it’s a public datasets and include many data and features I can use. It has the test data and the train data. They have the same features. So we can use the train data to train the model we choose and use the test data to test what preform the model we choose.
The dataset is on this website:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

6. Plan: Given your preliminary results, what are the next steps that you’re considering?
   I will train them using the scilearn library.