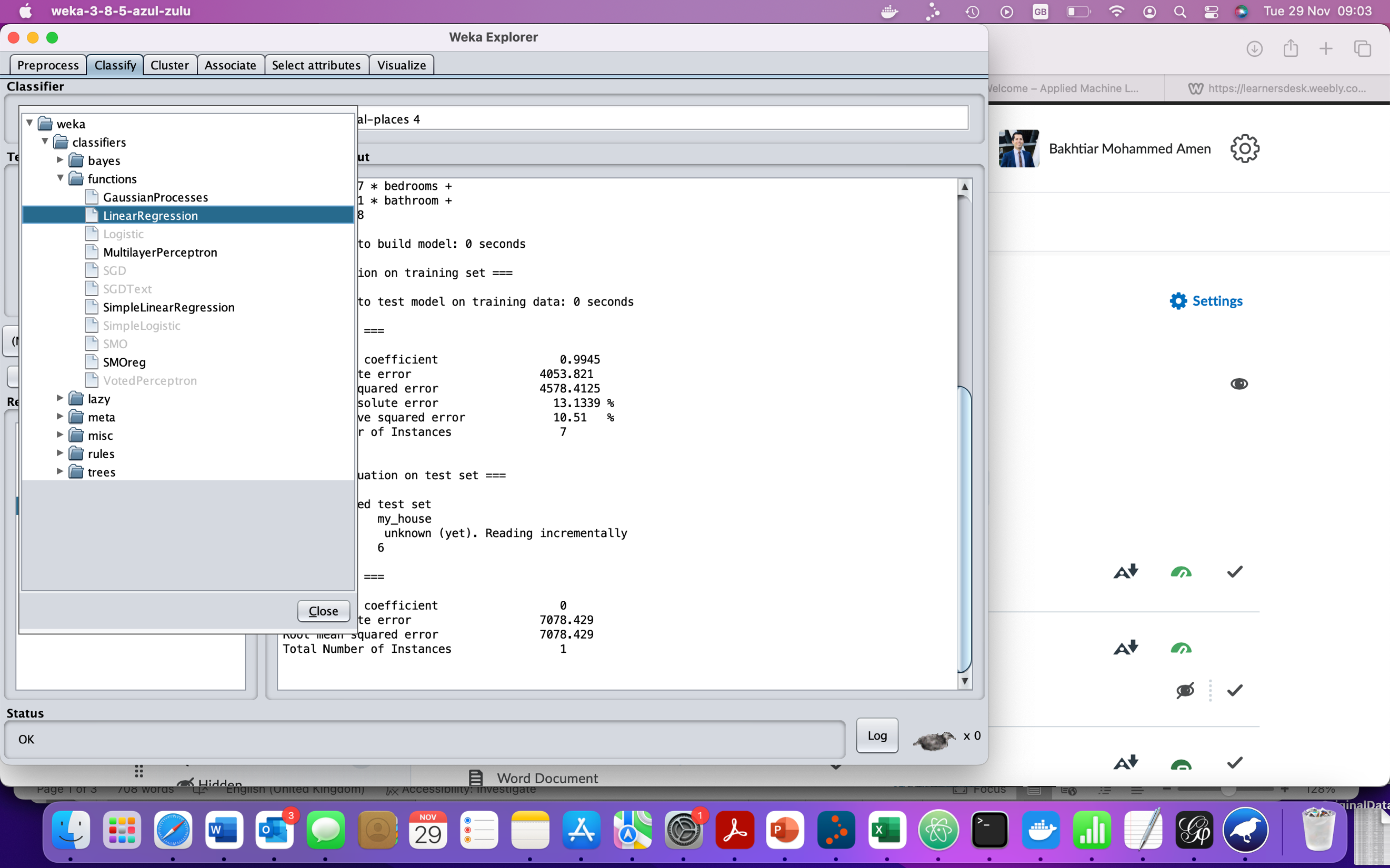
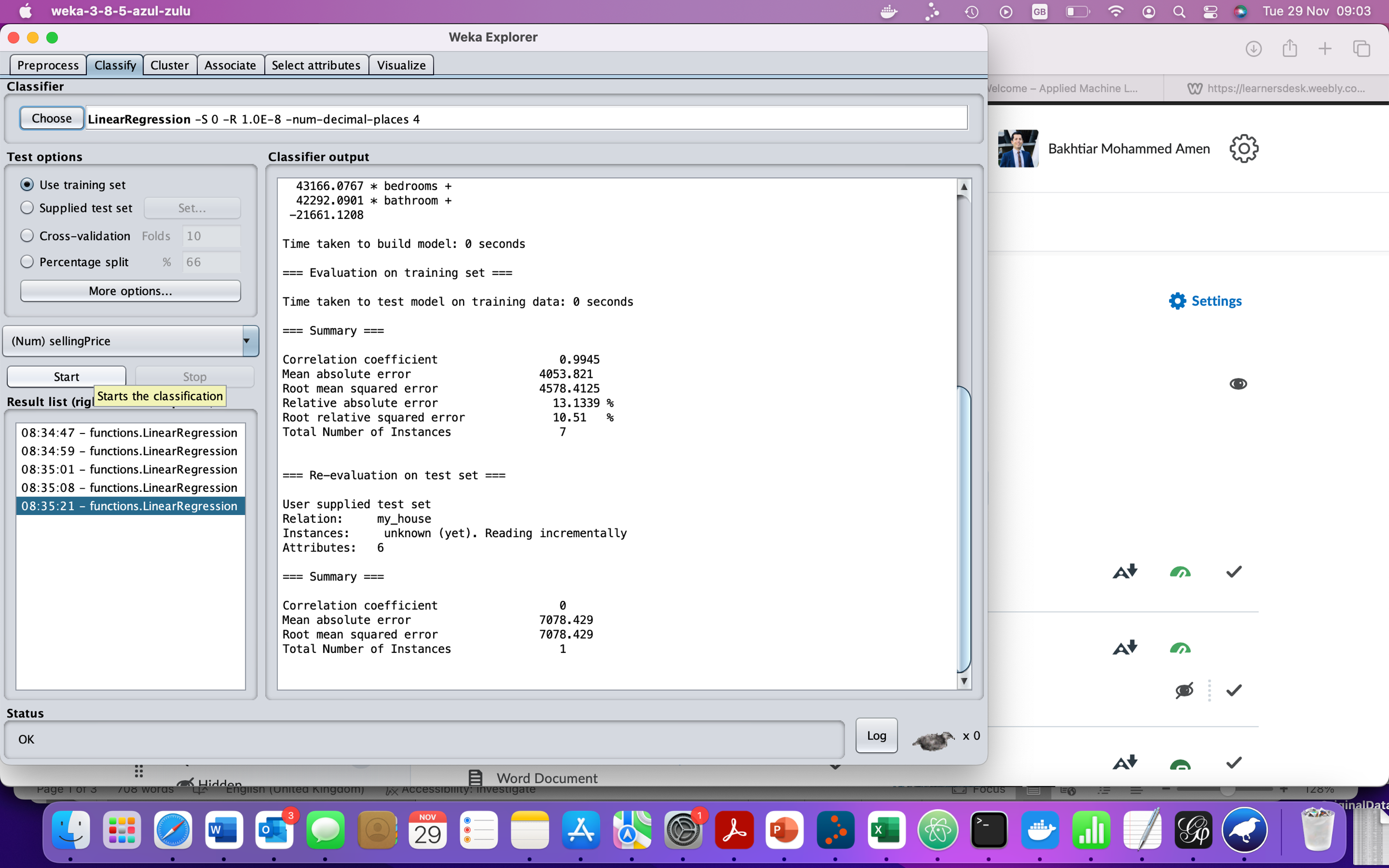
# CHA2555 Artificial Intelligence

# **Practical 16: Linear Regression**

# **Activity 1: Building the Linear Regression Model**

**Objective**: In this lab, you will examine **Linear Regression** classifier to predict house price.

1. Click the “***Open file,*** button and open house.arff data set (from Brightspace Week10\_Unit-1), which consists of house prices. Read through the description available for this data to get a feel for what you're dealing with. The data has been converted to ARFF format for you:



Given the data set we've just loaded, let’s now train **a Linear Regression** classifier to build your model from train give dataset. Under the *Classify* tab:

1. Select *Choose* in the *Classifier* frame at the top and select ***classifiers > functions > Linear Regression model*.**
2. In your test option choose training set (to build your modelon the train set)
3. Click *Start* to build the classifier.
   * Examine the classifier models produced by Weka for the outputs of Correlation, RMSE others

For the final part of this section we will now pretend we are interested in sale our house based on using Linear Regression prediction model. We will now use all of the training data to train our classifier and apply the learnt classifier to a dedicated test set. Load the (myhouse) test set in Weka.

1. Under the *Classify tab*, select *supplied* ***test set > set > open file*** and set the test file to the supplied***myhouse.arff.***  This ARFF file contains the numerical values. Run the Linear Regression classifier on this test set; use ***right click*** > re-evaludate Linear Regression model

# Graphical user interface, text, application Description automatically generated

Interpret the patterns and conclusions that our model tells us

* + - **Granite** doesn't matter :It will throw out and ignore columns that don't help in creating a good model. So this regression model is telling us that granite in your kitchen doesn't affect the house's value.
    - **Bathrooms** do matter: Since we use a simple 0 or 1 value for an upgraded bathroom, we can use the coefficient from the regression model to determine the value of an upgraded bathroom on the house value.
    - **Bigger houses** reduce the value: Model is telling us that the bigger our house is, the lower the selling price? This can be seen by the negative coefficient in front of the **houseSize** variable. The model is telling us that every additional square foot of the house reduces its price. That doesn't make any sense at all. How should we interpret this? The house size, unfortunately, isn't an independent variable because it's related to the bedrooms variable, which makes sense, since bigger houses tend to have more bedrooms. So our model isn't perfect. But we can fix this. On ***the Preprocess tab***, you can remove columns from the data set. Remove the **houseSize** column and create another model (*repeat step 1,2,3, 4*). How does it affect the price of my house? How does this new model make more sense?

**Activity 2:**

When a bank receives a loan application, based on the applicant’s profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the bank’s decision :

If the applicant is a **good credit risk**, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank If the applicant is a **bad credit risk**, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

Objective of Analysis:

**Minimization of risk and maximization of profit on behalf of the bank.**

To minimize loss from the bank’s perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant’s demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants.

1. Click the “***Open file,*** button and open **Trainset.csv** data set (from Brightspace Week10\_Unit-1). Read through the description available for this data to get a feel for what you're dealing with.
2. Repeat step activity 1 > step 1, 2, 3, 4, but this time use **Testset.csv** data to test your Linear Regression mode.