Quantifying the influence of imputing missing values.

Find a moderately large dataset with the following properties:
I'm thinking at least 1000 rows at least 10 features preferably 20. The features include: continuous, discrete and qualitative variables and one numeric variable which can be used as an outcome variable for regression.
The data should have no missing values, or at least complete data is available on enough features to satisfy the conditions above.

The general outline for each "run through" is
1 Simulate the missing values according to specific properties.
2 Apply an imputation method.
3 Assess the distributional statistics for the imputed features in comparison to the complete feature.
4 Compare a regression model using the imputed data compared to the reference regression model, which is fitted to the original complete data.

More details for the above parts
1 To start with simulate just missing completely at random, MCAR, with different proportions of missing eg, 5%, 10%, 20% … First univariate, then multivariate.
2 Univariate imputation: remove rows with missings, mean replacement, mean-variance replacement, sampling. Multivariate imputation: Regression imputation and MICE. Optional extra would be another approach that you might have read about.
4 You can choose an appropriate regression method to apply. I suggest something simple such as a linear model, Ridge or Lasso regression, as you don't want to spend too much time on this part, It would be interesting to see if the MSE is critical to the imputation methods and if so how does this compare to the findings from 3.

Other points
You should read more about the background theory on this subject than in ML2 and include this in the thesis. In particular I would like you to include a subsection on the MCMC/Gibbs sampling used in the MICE package.

Train/Test split: Need to be careful here.
Split the data into train and test. Keep track of which rows are train and which are test, this will stay fixed throughout the project.
Simulate missing values on all the rows: training and test.
Impute on the training data, train the regression model. For the test model, you can use all the data for imputation, but use the same imputation procedure. Evaluate the model using just the test data.

Optional extras not mentioned above
Investigate MAR or MNAR missingness.
As this is simulation, the bias in the regression estimates and predicted values can be obtained, by repeating the missing value simulation process.
Apply the method on a dataset with real missing values and see if the results are consistent.
Find and implement imputation methods not covered in ML2.