Exercise 1

You need to use the financial dataset `SPY.csv` and its Adjusted Close Price (`Adj Close`) to complete the following Exploratory Data Analysis.

- Plot raw time series data for `Adj Close` price.
- Make a histogram of the adjusted close 1-day percent difference (use *pct_change()*).
- Create the 5-day future price (as *5d_future_close*).
- Create the % price change 5 days in the future (*5d_close_future_pct*), and the current 5-day % price change (5d_close_pct).
- Examine correlations between the two 5-day percent price change columns.
- Make a scatterplot of *5d_close_pct* vs *5d_close_future_pct*.

Exercise 2

You need to use the Ex1data.csv to fit multiple regression models which will be used to estimate the target variable in column V1.

- Split `data` into a training set `data.tr` and a testing set `data.te`. The 75% of dataset should be used of training, the remaining are used for testing. In each of the following points, the models are to be fit to `data.tr`, the accuracy is to be analysed on both `data.tr` and `data.te`.
- Train a cross-validated lasso regression. Compute the sum of squares of error for both the training and testing set.
- Given the best fit models, list the relevant variables (the ones with non-zero coefficients). Run a simple linear regression using these input variables only. Compare the sum of squares of errors of this model to the best fit model of the previous part.
- Train the cross-validated Lasso model once with standardisation and once without it. Investigate the accuracy of the fits, and compare the lists relevant variables of the two models. Interpret the outcome.
- Train the cross-validated ridge-regression model, investigate the value of coefficients for the whole range of the penalisation parameter, investigate the accuracy, display and compare the results.
- Plot figures showing how the selected coefficients for both lasso and ridge regressions vary when penalisation parameter Lambda is increasing. Interpret the plots by comparing the shrinkage effect of Lambda for both models.

Exercise 3

Henry would like to buy a property. He performs an in depth survey in the area where he likes to buy. He quantifies, normalizes, and maps the desired features of houses to values

ranging from 0 to 1. However, he was only able to find the price per square foot for certain houses. Thus you are asked to develop a model to estimate the price per square foot of the houses for which he has compiled feature data but no pricing.

He observed features and assembled them into a table inclusing both training and testing datasets.

TRAINING DATASET

The first line contains two space separated integers, F and N. Over here, F is the number of observed features. N is the number of rows for which features as well as price per square-foot have been noted. This is followed by a table having $F+1$ columns separated by a single space and N rows separated by a newline. The last column is the price per square foot.

TESTING DATASET

After the table, is a single integer T. T lines follow, each being a row entry of the table with F columns separated by a single space.

OUTPUT FORMAT

T lines. Each line i contains the predicted price for the ith test case.

```
CONSTRAINTS
1 <= F <= 10
5 <= N <= 100
1 <= T <= 100
0 <= Price Per Square Foot <= 200
0 <= Factor Values <= 1
```

```
SAMPLE INPUT:
2 7
0.18 0.89 109.85
1.0 0.26 155.72
0.92 0.11 137.66
0.07 0.37 76.17
0.85 0.16 139.75
0.99 0.41 162.6
0.87 0.47 151.77
4
0.49 0.18
0.57 0.83
0.56 0.64
0.76 0.18

SAMPLE OUTPUT:
105.22
142.68
132.94
```

```
129.71
```