Estimated percentage contribution:

Stella Ramirez: 47.5%
Andrew Cheng: 5%
I(Liuqian Bao): 47.5%

Stella is a great partner and she has contributed significantly ever since she joined at step 2. She always summarises/organises what we need to do from the instructions and keeps track of what we are left to do. She and I did all the coding, interpretations, and formatting of the final deliverables. Whenever we are trying to debug our code or interpreting the results, she is helpful and makes a lot of contributions. She put a lot of time into our project.

Andrew, frankly, made little contribution during the whole quarter. In step 1, he helped a little bit with creating the graphs and writing the interesting features but I ended up correcting most of them when writing the report. In step 2, he did not do anything in R because he said he was not good at it. He only offered to write an introduction and a conclusion, both of which Stella and I ended up rewriting because they made very little sense(since he did not do any actual stuff). In step 3, I followed your advice and tried to give him small tasks to do in R, but he claimed that his R was empty(whatever that meant), thus it was inconvenient for him to code in R. He, again, wrote a conclusion that did not make sense before we even finished the actual contents. In step 4, he did not do any coding or write any interpretations in the report. He also did not join the discussion when we were trying to decide on the new method, so Stella and I ended up doing everything in that part. He tried to help us debug at some point, but only by sending us random google results related to the problem we were talking about. When it came to the conclusion, he again offered to write it, but when Stella was trying to communicate with him the content needed to be included in it, he did not care and said “It’s ok.” When we finished the summary, instead of checking for things the TA said she expects to see, he added meaningless sentences to our summary, for example, after I summarised the effectiveness of the shrinkage methods, he added:’The method really provided us a good way to work with our data, and the data provided us mostly with data that we expected.’

Data:

Anna Montoya, DataCanary (2016). House Prices - Advanced Regression Techniques. Kaggle. Obtained from <https://kaggle.com/competitions/house-prices-advanced-regression-techniques>

Rating: 8/10 (Clean, no missing values, albeit too many variables that are potentially confounding)

Difficulties:

Step 1: We have 80 variables in the original data set so we have to choose only the useful ones to present. We looked at the correlation matrix for numerical variables and boxplots for categorical variables to find ones that have strong correlation with the response.

Step 2: We had some difficulty with the poly() function used to fit quadratic terms in the model. Fitting the model was easy but we do not know how to construct PI's and CI's with the poly() function because the usual predict() function breaks with the poly() term. It also makes interpreting the model very challenging. We solved this by using log transformation to address our linearity assumption and avoid using the poly() function.

Step 3: We had a hard time getting the ggpair() to give us clear eligible pair graphs(we used too many variables at first). We solved the issue by further reducing the number of variables and grouping them to make separate pair plots. Also, we encountered difficulty when making CI's for the mean response using our statistical model because we did not know what to do with categorical variables(the mean() function clearly does not work for categorical variables). We figured to use the most frequent level for each categorical variable and construct CI based on that.

Step 4: We had some difficulty with plotting the three models in a single graph and adding a legend to it. The augment() function does not work for the models made by the glmnet() function but we finally figured it out by looking up examples on stack overflow. It is also challenging to put together a report for a client/manager in terms of what to include and we have to make sure it flows. We compiled information that we think is helpful to the client and tried to present it in a concise way. Stella had his father, who has a basic understanding of statistics but did not know anything about our data, read our summary and we used his feedback to make sure we explained everything clearly and excluded any unnecessary information.

In general, it is somewhat challenging to knit the graphs to where they meant to be in the pdf reports. We usually just put \newpage everywhere so that they come in order albeit occupying more pages, and also put descriptive fig.cap so that they are clearly labelled even when the order is messed up.

Approximate # of hours allocated to the project

About 8 hours/week overall. It varies with each step and sometimes we finish a step all in one week even when we are given one and a half or two weeks to do it.

General feedback

I wish we could have more time to do step 4. It would be more fun if we could explore different new methods and make comparisons. It could also be paced better so we could have the information we need for the project in time. Overall, it is more time-consuming than I thought it would be(sometimes it took us so long to figure out how to make a single graph or just one line of code). Even just formatting everything in rmd is more time-consuming than I expected. However, I think we had fun and learnt a lot!