Good afternoon, everyone. I am Laney with majoring in Data Science. Dr Guido Zuccon is my supervisor.

Today, I will introduce my data analysis project: Forecasting Zestimate Errors for Zillow.

Zillow is a leading real estate agent in the United States. Zestimate is their price prediction model to estimate market value for each individual property in the country. So the error I will work on is the difference between Zestimate and the actual sales price of the property. There are four sections I would like to discuss about the project plan today. They include project scopes and deliverables, project motivation and significance, project methodologies together with the constrains and risks management.

Let’s start from project scopes and deliverables. The scope stories are relevant to two product owners Zillow and QUT. For Zillow, they require participants to improve the accuracy of Zestimate model by predicting errors for 6 timepoints: October, November and December of both 2016 and 17. They also require the participant not to use any external data except the data they supply. The training data is through 2016 with 3millions of properties in 3 counties of California. Each property has around 57 variable data such as the room number and the location. Then the test data of 2017 will be available on the beginning of October. Therefore, the deliverables will be one CSV file which contains the errors data of 6 timepoints for 3million properties and one R markdown file which contains the analysis progress and programming codes. Regarding to QUT, the relevant project proposal, project presentation and final report will be required accordingly.

The motivations to apply this project are to implement my educational skills and enrich my experience about the financial data analysis. As to the significance of this project, I would like to explain it with some data. Zestimate has been released for 11 years based on millions statistics data points and machine learning models. By continually developing, Zestimate’s accuracy has been improved from a median error rate of 14% to 5%. However, according to the statistical meaning of the median error rate, there are still half of the home values are off more than 5% even though half of them is close to 5%. A home is often the largest and most expensive purchase a person makes in his or her life. It is not only a lovely place to live in. It is also an asset for the investment. It is important for the homeowners to have a trusted platform to monitor their investment. Therefore, there is a strong significance for this project to minimize the Zestimate error with efficient machine learning algorithms.

Let’s move to project methodologies. My project is a development project as some artefacts of error forecast are required as explained in the second slide. I would like to use SCRUM as my project management method. Because error predication is really a complicated job basing on statistics and machine learning. No one knows what result could largely present the accuracy of the Zesitmate with available resources. Scrum provides the sprint and increment factors to enable me to repeat execution and product review effectively until getting the most satisfactory outcomes. Here is my two-week Sprint blog. A meeting with supervisor is always at the beginning of each sprint to review what I have done and plan what I am gonna do. The underlined parts should be the deliverables in the end of each sprint. I am right now in W3 to present my project plan.

The last section today is about the project constrains and risks management. As we can see from the table, there is one constrain for Zillow to limit file submission. What I need to do is to think and check carefully before submission. And there are 3 main risks. The first one is about the unavailable supervisor. For example, he might be in a sick leave or an oversea conference. Alternative communication channels such as Slack, and other group students are the way to keep information updated. The second one is about the loss of the work. Then a stable backup platform like GitHub is good solution for that. The last one is about the limited computing power of private laptop or university computer when processing large amount of data and their visualization. The best way is to discuss this problem with supervisor to look for more powerful resources.

Finally, I would like to finish my project plan introduction here. Thanks so much for your time.