**Linni Qin**

**Student ID: n9632981**

**Supervisor: Dr. Guido Zuccon**

**Unit Code: IFN701**

**Project Category: A Development Project**

**Queensland University of Technology**

**Forecasting zestimate error for zillow: a data analysis project Proposal**

**Table of Contents**

1. **Project Introduction1**

1.1 Background1

1.2 Scopes1

1.3 Approach Overview3

1.4 Outcomes3

1. **Project Methodology4**

Phase One: Data Preparation4

Phase Two: Data Analysis5

Phase Three: Results Reflection5

Phase Four: Results Dissemination5

1. **Project Management Approach6**

3.1 Project Environment6

3.2 Project Management Plan6

1. **Communication Plan7**
2. **Risks Management Plan7**

**Reference**

**Appendix: Project Proposal Statement – Supervisor Sign-Off**

1. **Project Introduction  
   1.1 Background**

Zillow (zillow.com) is one of the popular information sites for real estate in the United States. It plays the role as realestate.com.au and domain.com.au in Australia to offer an online ecosystem for real estate professionals such as agents and mortgage bankers, home buyers, sellers and renters. There are 110 million properties across the nation have been served on Zillow’s platform (Kaggle Inc., n.d.). Zestimate is their price prediction model to forecast the market value of the properties and it is also an online tool to assist their website visitors making rational home-relevant decisions. According to the official data, the current median error rate of Zestimate is 5% (Zillow Inc., n.d.). This percentage seems that there is solely a minor disparity between the forecasting value and the actual sales price. However, it means greatly for the most expensive property people purchases in their lives. For example, a home with actual value of $700,000, its 5% will be $35,000 that approximately equals to an annual minimum income for a full-time employee in California (California Payroll, n.d.). As a result, the prediction home value on Zestimate might cause the potential home buyers to change their purchase decision or to be overconfident in their purchase capability. Furthermore, given the statistical concept, median error means there are a half of the properties’ estimation error rates exceed 5%. Consequently, the less visitor resulting from the inaccurate home value estimate service will lead to the decrease of Zillow’s revenue resulting from the less commercial advertiser on its site. Therefore, it is a critical issue for Zillow to improve the accuracy of Zestimate by finding a more effective home value prediction model. This is also the reason why Zillow posted this problem a challenge with a prize up to $1.2millions on the data science platform Kaggle.com to allow global data scientists for competing (Kaggle Inc., n.d.). Likewise, this project largely denotes the value of research relevant to value prediction models as the current Zestimate model has relied on more than 7millions of machine learning algorithms and statics data points (Kaggle Inc., n.d.).

**1.2 Scopes**

Given the competition rules of Kaggle, the expectation of Zillow, the table *“MoSCow Prioritised Requirement List”* below will represent the significance of the requirements through the whole project with applying prioritisation tool MoSCoW. “Must”, “Should” and “Could” requirements would be determined in the scope and “Would” items should be out of the scope as the limited time factor for this data analysis task.

**MoSCow Prioritised Requirement List**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***ID*** | ***Requirement List***  ***(Kaggle Inc., n.d.)*** | ***Deliverable Priority*** | ***Effort Points*** | ***Reason*** |
| 1 | As a contest sponsor, Zillow requires the participants to submit the outcomes with using their supplied data only. Any external data is prohibited. | Must  (Guaranteed) | 2 | This is compulsory requirement. Only the valid data could be fitted for the project. |
| 2 | As a competition organiser, Kaggle requires the participants to predict the error rate and store the value in a CSV file for 6 time points: October, November and December for both 2016 and 2017. R markdown file or Python file is required for producing the data analysis report and prediction model. | Must  (Guaranteed) | 28 | This is compulsory requirement. Without these outputs, the progress of data analysis and error rate prediction cannot be shown. Neither does the data analysis reproduction. |
| 3 | Kaggle wants the participants to submit the CSV and data analysis files on October 16, 2017. | Should  (Expected) | 2 | Outputs should be submitted on time. |
| 4 | As a data supplier, Zillow wants the participants to analyse the value of their collected data. | Should  (Expected) | 23 | Data analysis is aimed to exploring the depth insight of the datasets. |
| 5 | As a real estate service provider, Zillow wants the participants to find out one effective prediction model to increase Zestimate’s accuracy. | Could  (Possible) | 25 | The most effective prediction model could improve Zestimate’s performance definitely. However, the level of effectiveness should be evaluated accordingly to the time factor of this project. |
| ***In-Scope Points:*** | | | ***80*** | |
| 6 | As a data science community, Kaggle wants the participants to do contribution on their coding sharing platform and forum. | Would  (Maybe) | 2 | This is based on the willingness of participants. |
| 7 | Kaggle expects the participants to predict the home values for Zillow. | Would  (Maybe) | 18 | Data analysis and error prediction are two current core section. Home value prediction could be done in next round of competition. |
| ***Out-Scope Points:*** | | | ***20*** | |
| ***Project Total Points:*** | | | ***100*** | |

**1.3 Approach Overviews**

The table *Data Analysis Workflow* below indicates the commonly four phases with totally eight steps involved in the workflow for facilitating data analysis project. The four phases are data preparation, data analysis, results reflection and results dissemination. Data preparation is the time-consuming part in the workflow. However, it plays as the fundamental base for the further analysis. Data analysis is the core activity to execute the parameter and analyse the data for obtaining the insightful information. The phases of results reflection and dissemination will determine the quality for the final outcomes by continuing adjusting the experiments with collecting the helpful feedbacks from the supervisor and comparing various outputs value.

As to the details of the project approach, it can be referred to section two *Project Method*.

**Data Analysis Workflow**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Phase One***  ***Data Preparation*** | ***Phase Two***  ***Data Analysis*** | ***Phase Three***  ***Results Reflection*** | ***Phase Four***  ***Results Dissemination*** |
| **Step 1: Defining problem**  **Step 2: Identifying ideal datasets to answer analysis problems**  **Step 3: Acquiring data**  **Step 4: Cleaning data** | **Step 5: Exploring data**  **Step 6: Statistical prediction and modelling data** | **Step 7: Interpreting results** | **Step8: Communicating and distributing results** |

**1.4 Outcome**

Two main tangible outcomes will be returned in the end of this competition task. One of them is a data analysis report that represents the depth insight of the Zillow datasets building on a considerable large amount of house data across three counties in California of the United States. It is going to demonstrate the relationship patterns between the estimated price of each property and their respective over-fifty physical variables. The other one is a rational prediction model that generates the estimated error value relating to 3millions of properties during the last quarter of both 2016 and 2017, depending on the training data and test data supplied on Kaggle.

1. **Project Methodology**

Dr. Guido Zuccon’s lecture of unit IFN509 – Data Manipulation at Queensland University of Technology discussed the data analysis methodology (February, 2017). The justification of project methodology below builds on the knowledge obtained from IFN509.

**Phase one: Preparation**

**Step 1: Defining the problem**

Before proceeding, it needs to figure out the categories of data analysis questions that are to be solved. As per the supplied property data, there are around 3millions of properties needing predicted on their error rate, then there are 57 variables such as the room number, square footage and location for each house. The goal to understand this large scale of data set is to define the relationships among these variables and confirm whether physical factors of each property correlates with their estimated price. Meanwhile, the training dataset of 2016 and the test data set of 2017 are used to predict the error rate between the estimated home value and the actual sales price for each individual property. Thus, this project is going to solve a hybrid analysis problem - an inferential and predictive analysis.

**Step 2: Identifying the ideal dataset to answer the analysis problem**

Although there are 3millions of properties listed, certain of them are provided high level of detail information for their physical elements from the home owner, and parts of them are with little data. Similarly, 90thousands of the training data with different error rates, some larger than 5%, others smaller than that. It is essential to identify the ideal dataset to implement the data searching.

**Step 3: Acquiring the data**

Data will be acquired from the Kaggle competition site as this is the only official and approved channel to obtain the valid data. Then, the data will be stored in working computer hard disk and other stable backup working station. A couple of risks related to data storage will be explained in detail in section five Risk Management Plan. Finally, import the data into the experimental environments such as RStudio for further proceeding.

**Step 4: Cleaning the data**

Raw datasets downloaded from Kaggle are all CSV files that might contain certain value formats like inconsistence that cannot be executed and computed while a specific analysis needs running. Meanwhile, different datasets might need to integrate into one file for analysis convenience. Therefore, after importing into experimental platform, the data should be cleaned. The executing tool for data munging can be R packages such as plyr and dplyr.

**Phase two: Analysis**

**Step 5: Exploring the data**

Starting from this step, it is getting to understand the data with developing hypotheses for the relationships between the variables that to be correlated beforehand. Secondly, identify the hypotheses with check if there are positive or negative relationships between the targeted variables then explore the relevant pattern so as to find out the most valuable data. The tool can be R with its packages, like ggplot to plot or cluster the datasets.

**Step 6: Statistical prediction and modelling the data**

Implement the research to find out which prediction model, for example, the logistic regression model and machine learning models, are appropriate for this project. Deploy the models into the experiment and monitor their performance. Repeat the cycle of model building, model deploying and model monitor until finding the most effective one within the timebox.

**Phase three: Reflection**

**Step 7: Interpreting the results**

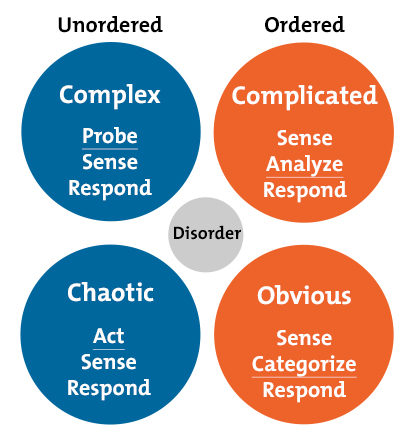
Analysis and reflection phases should be frequently alternative, especially when making comparisons among the various outputs. Adjust the code and parameters to rerun the processes with testing data. Also, discuss the outputs with the supervisor to get their helpful feedbacks.

**Phase four: Dissemination**

**Step8: Communicating and distributing the results**

The final step is to disseminate the results. The required forms of result representation from Kaggle are one CSV file that will contain the predicted error value of 3millions of properties for 6 timepoints. The other representation is one programming file such as R markdown file to include the execution code and written report that will allow Kaggle examiners and other data analysists to reproduce the experiment. Those two communicating channels are going to display the depth insights of the Zillow property data and the power of prediction model to improve the performance of Zestimate.

1. **Project Management Approach  
   3.1 Project Environment**

 As voted on MindTools.com, Cynefin is most popular tool to identify the problem-solving approach in a particular environment (n.d.). The table *“Project Environment Framework”* below will represent the problem types and justify the reasons concerning five key factors with applying Cynefin theory in this data analysis task.

Source: MindTools

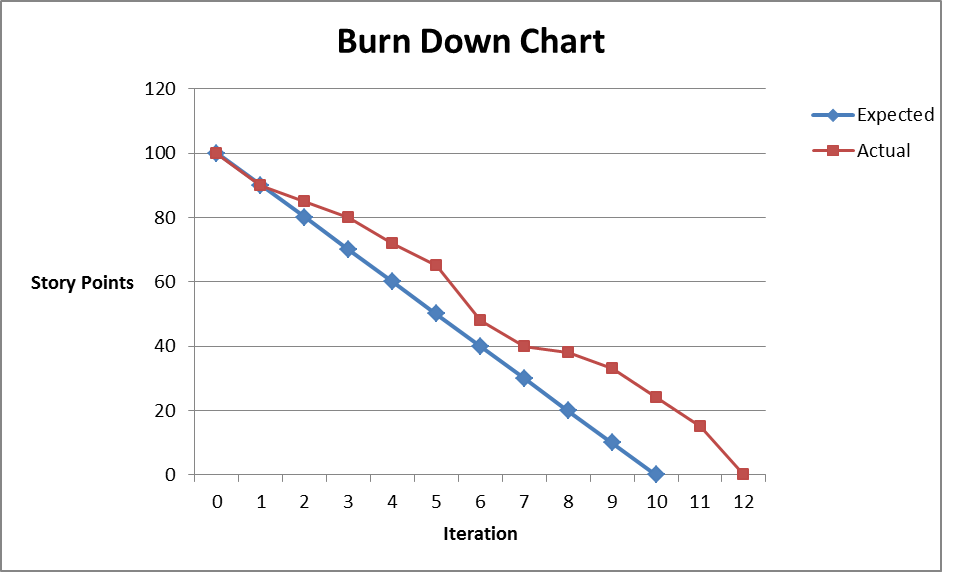
**Project Environment Framework**

|  |  |  |
| --- | --- | --- |
| **Factors** | **Types** | **Reason** |
| **Domain** | Complex | Considering the home value prediction model, it might be challenging to determine which model is the correct solution for shooting Zillow’s problem. The best way is to “Probe-Sense-Respond” with frequent reflections and data analysis. |
| **Cause & Effect** | Perceived in retrospect |
| **Risks** | Unknown - unknown |  |
| **Cost of Change** | Low | Given a data analysis project, the relevant changes of using new data, new statistical prediction and machine learning model |
| **Project Management Approach** | Scrum |  |

* 1. Project Management Approach

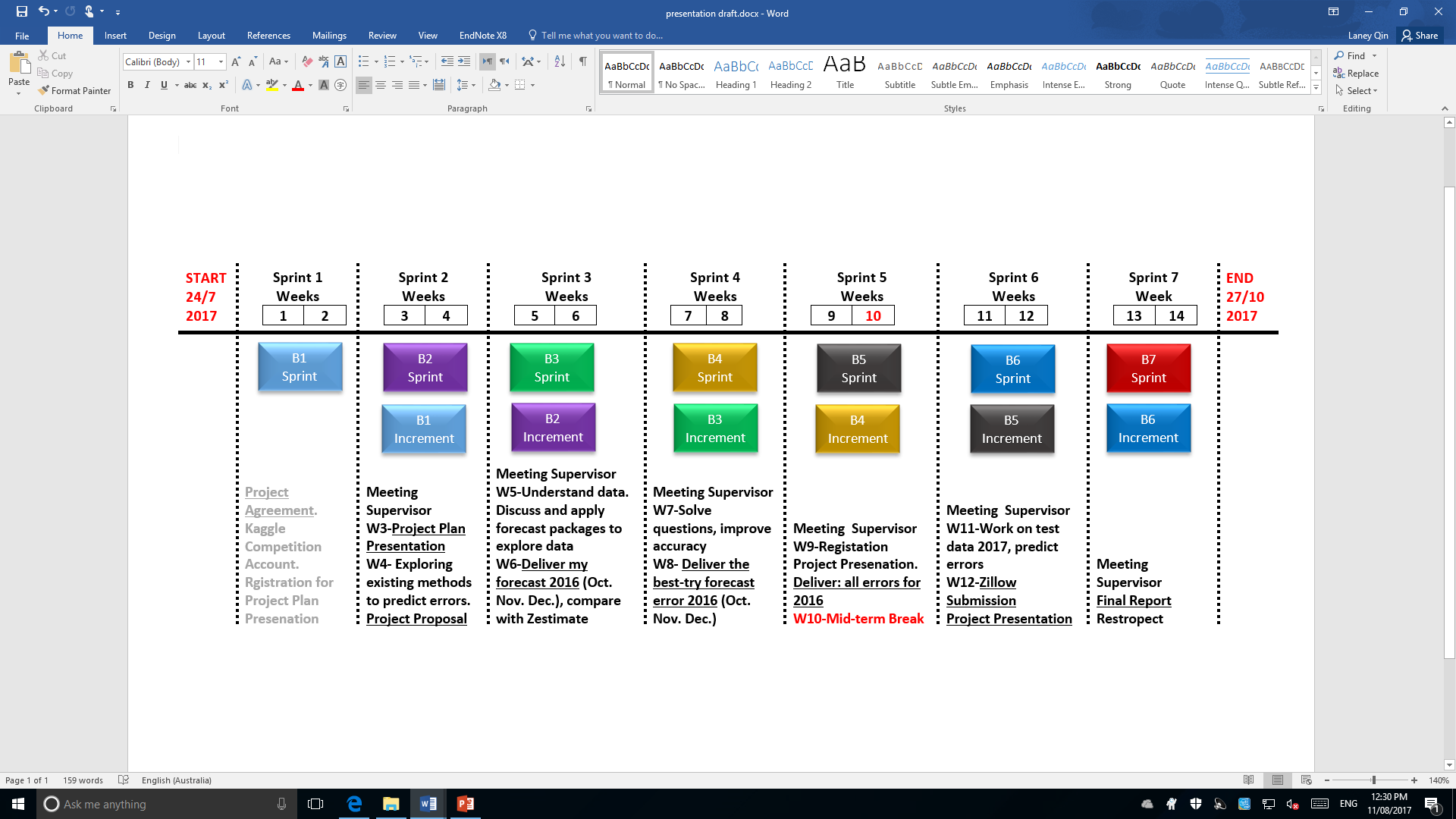
Referring to the table *“MoSCow Prioritised Requirement List”* in section 1.2, the requirement list will be transferred into the product backlog and the estimated effort points will be the measurement scale for Scrum management. 80 story points of the project backlog will be allocated into each sprint until they are completed as indicated in the *Burn Down Chart* below.

Burn Down Chart (will be changed)

[](https://www.google.com.au/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjak4P58OLVAhXFX5QKHVJOB3kQjRwIBw&url=http%3A%2F%2Fblog.caplin.com%2F2012%2F05%2F31%2Fgo-up-not-down%2F&psig=AFQjCNE9O-bEjkzzJzqKyoxqoe8gZNxbLQ&ust=1503217805421637)

The sprint plan to complete the targeted backlogs is demonstrated as the following table “Two-Week Sprints” (the table will be adjusted later).

Two-Week Sprints



1. Communication Plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phases** | **Aim** | **Type of Communication** | **Location** | **Frequency** | **Duration** |
| **Preparation** | Ensure the right scope and method | Online chat | Slack | Depends | Depends |
| Email | Depends | Depends |
| Face-to-Face | Supervisor’s office | Once | 30-45 minutes |
| **Data Analysis** |  |  |  |  |  |
| **Reflection** |  |  |  |  |  |
| **Dissemination** |  |  |  |  |  |

1. Risks Management Plan

With considering the risk occurrence probability and the consequence severity level, the *Risk Assessment Matrix* below will be used to access the serious level of the potential risks and define the relevant mitigations. The green tables with low 1 represent the risk and consequence are acceptable without doing any response. The yellow tables with medium 2 show the risk and consequence are also acceptable but need certain kind of monitoring to keep it from being worse. For example, the updated version of Zillow training data and property data is required to download frequently on Kaggle. This might affect the continual accuracy of the value prediction or the pattern of relationships during the data understanding phase. However, if the data might be often updated daily, it is not necessary to download the large scales of data daily. The better way to deal with this situation is to download the latest one when starting the work and assign the unique and convenient names for effective files management. The orange tables with high 3 mean the risk and the results are unacceptable then it needs to stop the project execution until finding the workable solution. The last red table with extreme 5 show the immediate project close is a must to prevent the catastrophe happening. For instance, the data available on Kaggle is hacked and replaced with malware programs. This will cause catastrophe for the private workstation and together QUT public network environment. When encountering this kind of extreme situation, the project should be stop immediately. The table *Risk Level 3 and Mitigation* below the matrix will indicate mainly the orange unacceptable risks together with their mitigations.

**Risk Assessment Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Likelihood to happen | Very likely  (90% to occur) | Acceptable  Risk  Medium 2 | Unacceptable Risk  High 3 | Unacceptable  Risk  Extreme 5 |
| Likely  (50% to occur) | Acceptable  Risk  Low 1 | Acceptable  Risk  Medium 2 | Unacceptable Risk  High 3 |
| Unlikely  (20% to occur) | Acceptable  Risk  Low 1 | Acceptable  Risk  Low 1 | Acceptable  Risk  Medium 2 |
|  | | Minor | Moderate | Major |
| The severity of the risk | | |

**Risks of Level 3 and Mitigation**

|  |  |
| --- | --- |
| **Risks** | **Mitigation** |
| **Loss of the work**  The hard disk of the private workstation or the public workstation at QUT might break down without any notice. | To prevent losing the efforts and the analysis results, it needs to store the work in a stable repository such as GitHub. GitHub does not only offer the free repository for the public account user, but also provide the users the tools and bridges to manage and share the code. Supervisor can also check regularly my progress of data analysis there. |
| **Limited computing power**  The main workstation for the task is on private laptop and public computer at QUT. Their hardware specifications, such as the graphic processor and flash memory, are not capable to deal with GB level of data calculation and complex 3D plots. | Discuss with supervisor or together with faculty to get more powerful resources and rational working environment. |
| **Unavailable of Supervisor**  Supervisor might be on the sick leave or an oversea conference for two weeks that equals a whole sprint timebox. When encounter this hardship especially on the key data analysis phase and refection phase, it is hardly to finish the task with high quality without the supports from the supervisor. | Alternative communication channels, such as Slack and Skype, are able to increase the chances of getting supports from supervisor.  The other way is to get the support from the project coordinator who holding the rich resources to solve the problem. |

**Reference**

California Payroll (n.d.). 2017 California Minimum Wage Table. Retrieved from https://californiapayroll.com/minimum-wage-changes-effective-july-1-2016/

Guido, Z. (Feburary, 2017). IFN509: Data Manipulation – Week 1 Lecture. Retrieved from https://blackboard.qut.edu.au/bbcswebdav/pid-6726760-dt-content-rid-7967236\_1/courses/IFN509\_17se1/lecture\_w1.pdf

Kaggle Inc. (n.d.). Zillow Prize: Zillow’s Home Value Prediction (Zestimate). Retrieved from https://www.kaggle.com/c/zillow-prize-1#Competition Overview

MindTools (n.d.). The Cynefin Framework. Retrieved from https://www.mindtools.com/pages/article/cynefin-framework.htm

Zillow Inc. (n.d.). Zestimate. Retrieved from https://www.zillow.com/zestimate/

**Appendix: Project Proposal Statement – Supervisor Sign-Off**

I, Dr. Guido Zuccon *<name of supervisor>,* confirm that I have gone through the project plan made by Linni Qin *<student name>* holding student ID number: n9632981 on the project titled: “IT MASTER (DS): Perform a Data Science Analysis of a Dataset/Task – Dr Guido Zuccon” for IFN701 <unit code>.

I confirm that I have been consulted in deriving this project proposal and that I approve of the suggested scope and tasks described in this project plan and that I am satisfied with the identified risk mitigation and communication plans articulated here.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor signature Date