**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 dataset analysis REPORT**

**Executive Summary**

This report explains the methodology to predict the individual logarithm error, which is the difference between Zillow’s estimation price and the actual sales price, for 3millions of properties in California for the last quarter of 2017. The available data supplied from the Kaggle platform is containing solely 57 physical features for those properties such as the room number and square footage, together with certain historical log error value for the transacted properties through the year of 2016 and the first nine months of 2017 for model training purpose. The multiple linear regression model is employed to predict the log error, then the proper size of the variable is filtered to fit the prediction model.

**Table of Contents**

**Executive Summary**

1. **Project Introduction1**

1.1 Background and Context1

1.2 Problem1

1.3 Purpose1

1.4 Approach Overview2

1.5 Scopes2

1.6 Outcomes and Significance4

1. **Environmental scan and review of prior related work4**

2.1 Competition Timelines4

2.2 Available Data5

2.3 Project Requirements6

2.4 Difficulties for Pre-Process6

2.5 Core “TO DO” Tasks7

1. **Project Methodology7**

3.1 Phase One: Data Preparation8

3.2 Phase Two: Data Analysis9

3.3 Phase Three: Results Reflection10

3.4 Phase Four: Results Dissemination10

1. **Project Management Approach11**

4.1 Scrum11

4.2 Burn Down Chart and Sprint Backlog11

1. **Outcomes**1**4**

5.1 “Done” List14

5.2 A Data Analytics Report5

5.3 A Rational Multiple Linear Regression Model5

1. **Discussion9**
2. **Conclusion10**

**Reference12**

**Appendix: Reflections on your learning13**

1. **Project Introduction**

**1.1 Background and context**

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 Zillow’s 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. The Logarithm error value is predicted within this project is the logarithm difference between the Zestimate price and the actual price for each property (Kaggle Inc., n.d.). Another relevant platform for this project is Kaggle.com. It is a global data science competition platform for predictive modelling and data analytics. Zillow posted their data on Kaggle for crowdsourcing the most effective prediction techniques in two-round competitions starting from May of 2017 to January of 2019. The final winners’ models might be applied in Zestimate model to improve their prediction accuracy.

**1.2 Problem**

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 that people purchase 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.). This error rate on Zestimate might cause the potential home buyers and seller to change their decisions or to overestimate or underestimate their financial capability. 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 precision of Zestimate by finding more effective prediction techniques. This is the reason why Zillow posted this problem a challenge with a prize up to $1.2millions on Kaggle.com to allow global data scientists for competing (Kaggle Inc., n.d.).

**1.3 Purpose**

The purpose of this project can be explained in two ways. For Zillow, their terminal aim is to find more effective methods to improve Zestimate by crowdsourcing on Kaggle. Meanwhile, this is also a kind of marketing strategy to claim their ambitions to deliver the best real estate data services for the site users. Regarding the purpose of my supervisor of unit 701, the project participant can take this opportunity to explore the real world raw data and perform a data science analysis then to be able to build a rational prediction model.

**1.4 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 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 determine the quality for the final outcomes by continually 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 three ***Project Methodology***.

**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** |

**1.5 Scopes**

Given the competition rules of Kaggle and the expectation of Zillow, the table *“MoSCow Prioritised Requirement List”* below represents the significance of the requirements through the whole project with applying prioritisation tool MoSCoW (Business Analyst Learnings, 2013). “Must”, “Should” and “Could” requirements were determined in the scope and “Would” items were out of the scope as the limited time factor for this task. Those items were scored as well, then they were allocated into the weekly “TO DO” tasks. The details about “TO DO” and “DONE” tasks can be referred to section four ***Project Management Approach***. In general, the driver behind determining the scope can be referred to section two ***Environmental Scan and Review of Prior Related Work.***

**MoSCow Prioritised Requirement List**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***ID*** | ***Requirement List***  ***(Kaggle Inc., n.d.)*** | ***Deliverables Priority*** | ***Effort Points*** | ***Reason*** |
| 1 | As a contest sponsor, Zillow requires the participants to submit the predicted errors 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) | 10 | 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. The submission will be evaluated by mean absolute error. | Could  (Possible) | 17 | 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 in second round of competition. | Would  (Maybe) | 18 | Data analysis and log error prediction are current core works. Home value prediction could be done in next project. |
| ***Out-Scope Points:*** | | | ***20*** | |
| ***Project Total Points:*** | | | ***100*** | |

**1.6 Outcome and Significance**

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

The analytics of data and the predictive model were uploaded online such as the GitHub.com and Kaggle.com for all the global readers, programmers and data scientists who fell interested in this topic to read and communicate. The generated predictive log error for the last quarter of 2016 and 2017 is being evaluated by Kaggle and Zillow with applying Zillow’s self-defined scoring mechanism so that they are able to choose the most effective and reliable techniques to improve Zestimate in the future.

1. **Environment Scan and Review of Prior Related Work**

Even though the purpose of my project of unit 701 is not to take part in the Kaggle competition but to use solely its datasets for analytics and predictive modelling, to understand well the environment of the whole project, the tabular form is used mainly in this section to present the available datasets and requirements obtained from Kaggle. Then, the relating difficulties and strategies to overcome those hardships as well as the methods to execute the analysis are discussed parallelly until finding the proper prediction model to handle these datasets for generating the desired value.

**2.1 Competition Timelines**

|  |  |
| --- | --- |
| **Name of Prediction Competition** | **Zillow Prize: Zillow’s Home Value Prediction (Zestimate)** |
| ***Contents*** | ***Timelines*** |
| First round competition:  Log error prediction | May 24, 2017 – January 17, 2018 |
| Release of 2017 training data: | October 2, 2017 |
| First round submission deadline: | October 16, 2017 |
| First round sales tracking period: | October 17, 2017 – December 15, 2017 |
| Second Round Competition: | February 1, 2018 – January 15, 2019 |

*(Sources from Kaggle Inc., n.d.)*

**2.2 Available Data**

|  |  |  |  |
| --- | --- | --- | --- |
| ***No.*** | ***Name of Datasets*** | ***Forms*** | ***File Description*** |
| 1 | Properties\_2016 | csv | A table lists full parcelids of 3millions of real estate properties in three counties of California in 2016 (Orange, Los Angeles and Ventura), with 57 physical features of properties such as the room number, year of built and square footage and so on. |
| 2 | Train\_2016\_v2 | csv | A transaction file contains the actual log error for around 90thousands properties which have been transacted through 2016 ***(0101/2016 – 31/12/2016)*** and their detail transaction date. |
| 3 | Properties\_2017 | csv | A table lists full data of 3millions real estate properties in three counties in California in 2017 (Orange, Los Angeles and Ventura), including 57 physical features of properties such as the room number, year of built and square footage and so on.  *(same format with properies\_2016)* |
| 4 | Train\_2017 | csv | A transaction file contains the actual log error for around 80thousnads of properties which have been transacted from January to September of 2017 ***(0101/2017 – 15/09/2017)*** and their detail transaction date.  *(same format with train\_2016\_v2)* |
| 5 | Zillow\_data\_dictionary | csv | A dictionary to explain the field name of 57 property features. |
| 6 | Sample\_submission | csv | A sample file to ensure the final submission is in the correct format. |

*(Sources from Kaggle Inc., n.d.)*

**2.3 Project Requirements**

|  |  |
| --- | --- |
| ***No.*** | ***Requirement Detail*** |
| 1 | The submission file of predicted value will be evaluated on Mean Absolute Error. |
| 2 | Data Analysing tools: R or Python |
| 3 | The log error must be predicted for each property and for each time point. There are six timepoints: October 2016, November 2016, December 2016, October 2017, November 2017 and December 2017. Then, the submission file must be a csv file. |
| 4 | Participants are required to submit the predicted errors with using supplied data only on Kaggle.com. Any external data is prohibited. |

*(Sources from Kaggle Inc., n.d.)*

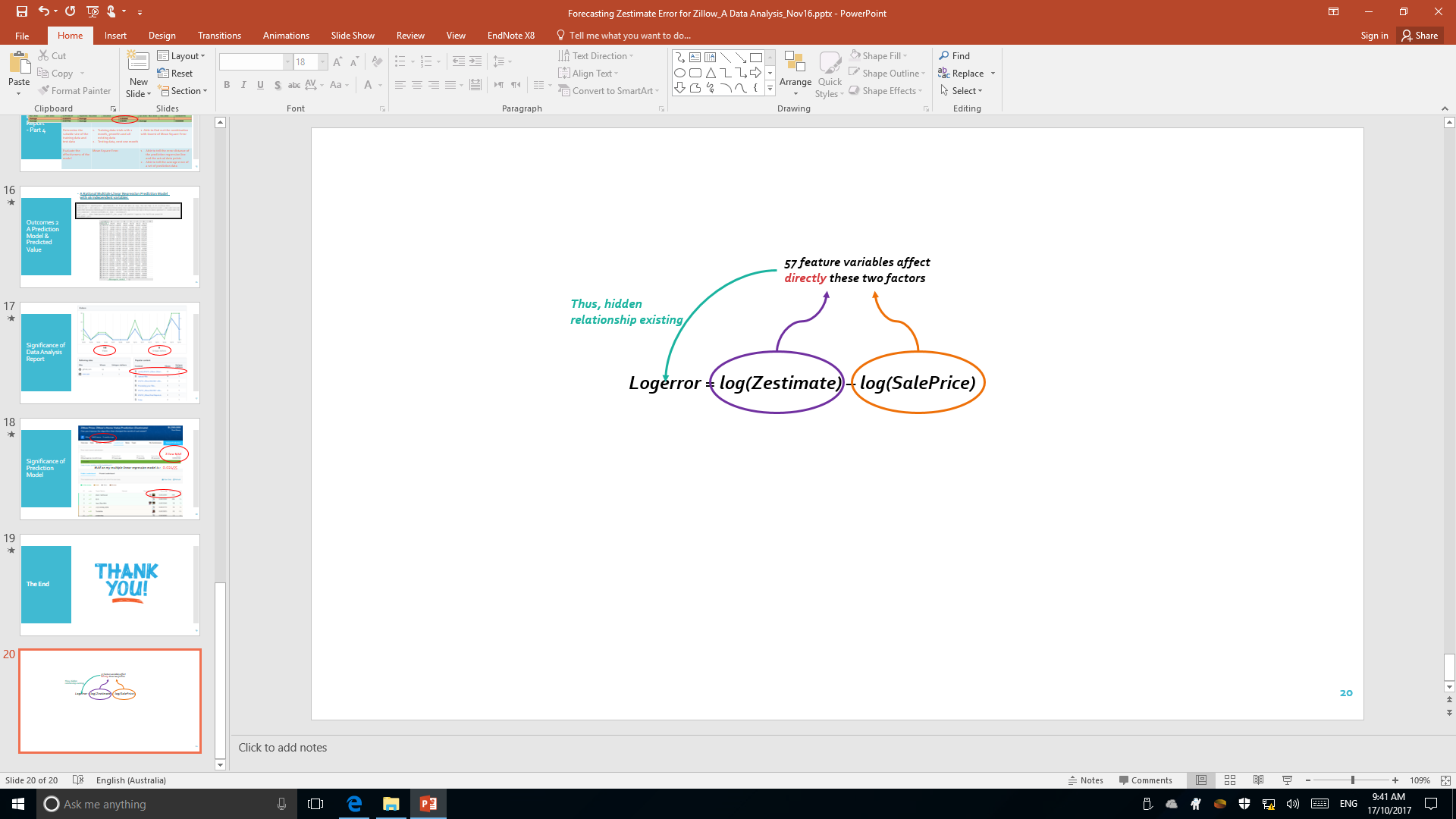
**2.4 Difficulties for Pre-process**

There are two difficulties need to be overcome during the phase one of project workflow.

As mentioned in the overview section of project approaches, there are four steps. Among them, step two is to identify the ideal datasets to answer analysis problems. The first hardship is that there are only two main datasets from 2016 and 2017 can be used for training. One set is the transaction file and another one is the property features file. There is lack of testing dataset. Furthermore, without enough historical data, the trends of log error development cannot be explored easily.

The second hardship is that there is no direct relationship between log error and the 57 physical property features. As stated in the first section and demonstrated as the function below, the log error required to be predicted is the difference between the Zestimate and the actual sale price of the property. However, there is no data relating to the Zestimate and the actual sales price within the available data. Instead, the 57 property variables have the direct effect on the Zestimate price and the actual sales price.

With the purpose to figure out these two hardships, the prototype of major “TO DO” tasks for phase one – data preparation and phase two – data analysis was identified in the next section 2.5.



**2.5 Core “TO DO” Tasks**

|  |  |  |
| --- | --- | --- |
| ***No.*** | ***Phases*** | ***TO DO*** |
| 1 | One | Merge transaction dataset and property dataset by common primary key that is the unique “parcelid” for each property. |
| 2 | One | Clean the data with filtering the missing data, non-numeric data and duplicated data |
| 3 | One | Identify the ideal data for testing purpose |
| 4 | Two | Explore the hidden relationship between the log error and 57 variables |
| 5 | Two | Determine a suitable prediction model |
| 6 | Two | Build the function to fit the prediction model |
| 7 | Two | Predict the log error |

This “TO DO” list are the important tasks that were allocated into the sprint logs. They will be compared carefully with the results on “DONE” list stated in section five ***Outcomes***.

1. **Project Methodology**

Considering the four major phases integrating with eight steps to accomplish this whole project, together with the necessary “To Do” tasks listed in the last section, the data analysis methodologies and tools discussed and taught by Dr. Guido Zuccon in his lecture of unit IFN509 – Data Manipulation in 2017 at Queensland University of Technology was applied in facilitating this project. The following subsections will explain the methodologies in detail and present the relating justification.

**3.1** **Phase one: Data Preparation**

**Step 1: Defining the problem**

Before proceeding, the categories of data analysis questions need to be determined. As per the supplied property data, there are around 3millions of properties needing predicted on their error rate. Then, the hidden relationship between the log error and 57 variables such as the room number, square footage and location for each house needs to be defined. Thus, this project solves a hybrid analysis problem - an inferential and predictive analysis.

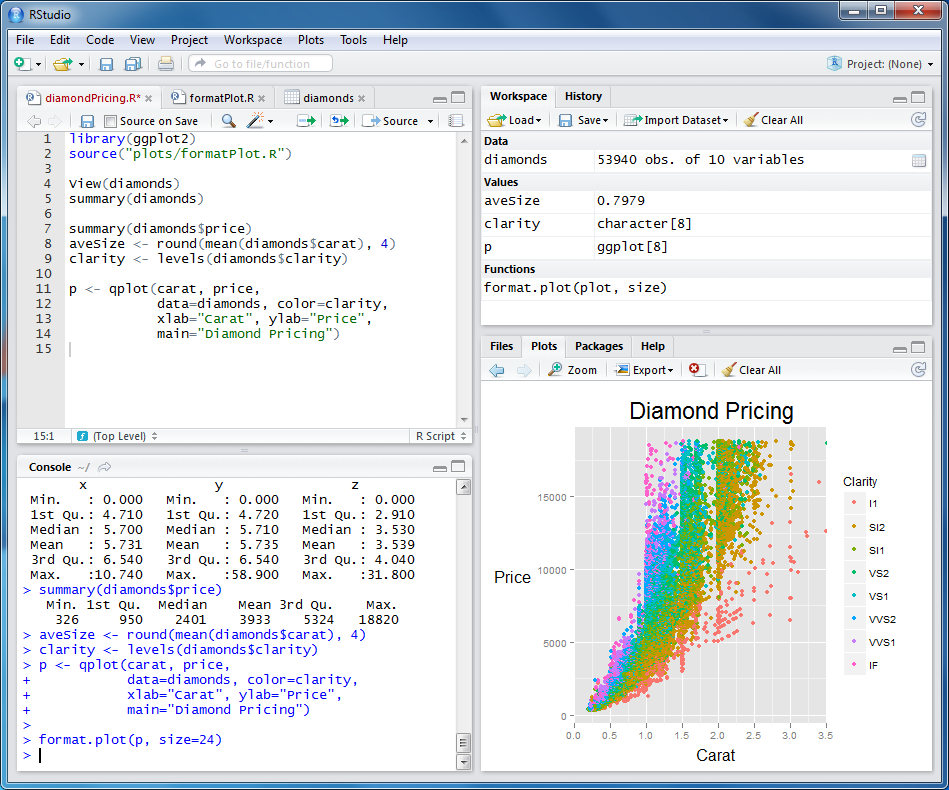
**Step 2: Identifying the ideal dataset to answer the analysis problem**

The ideal first action is to merge the transaction dataset and the property dataset by common primary key which is the unique “parcelid” for each property. There are two reasons. The first, transaction dataset cannot predict the full 3millions log error for the property without knowing their variables. The second, the property dataset cannot predict the log error by itself because there is no actual log error value involved in this file. The following ideal action is to separate the merged file of 2016 into two parts. One part involves the log error and 57 variables from January to September as the training data, and the rest part including the relevant data from October to December works as the test data.

**Step 3: Acquiring the data**

Programming Tool and Programming Environment: R version 3.3 and RStudio 1.1.3

R language and RStudio were chosen as the data analysing and code compiling tool because this meets the requirement of the competition. R is an open source computing software for statistical analysis and visualization (R, n.d.).



As shown as left figure, RStudio is a free source intergrade development environment as well to make programming language R more productive, because it is integrated with four windows with console, editor and a wide range of visual plotting tool (RStudio, n.d.). Furthermore, these two environments are the analysis software I am able to use proficient currently.

**Step 4: Cleaning the data**

R packages: plyr and dplyr

Although there are data of 57 variables listed for 3millions of properties, a certain percentage of them are provided high level of detail information from the home owner, and other fields are missing a significant amount of data. Similarly, there are some duplicated data stored under different names of the field (Kaggle Inc., n.d.). Likewise, there are some non-numeric datatype that the prediction model cannot handle when calculating the specific log error number. It is essential to clean the data for a clean and light analysis environment. The executing functions for data munging are R packages such as plyr and dplyr (Hadley, W., n.d.).

**3.2** **Phase two: Data Analysis**

**Step 5: Exploring the data**

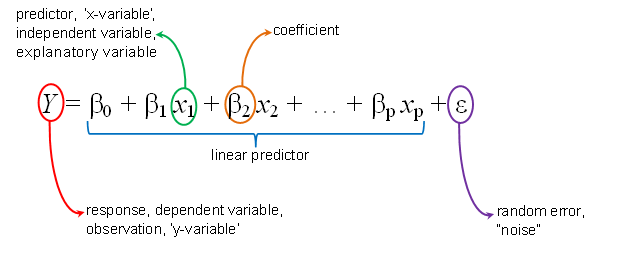
R packages: ggplot

Starting from this step, it began to explore the data with justifying the hypotheses for the correlations between the log error and the 57 variables. Then, identify the most valuable variables to fit the prediction model. The functions were used like R package ggplot which can plot or cluster the datasets (Hadley, W., n.d.).

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

Prediction Model: Multiple Linear Regression

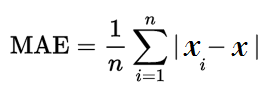
*Source: Cross Validated (2014)*

 Multiple linear regression is a type of machine learning model that is able to identify the strength of the effect that the independent variables have on a dependant variable, as well as it can forecast future value and return a specific value (Statistics Solutions, n.d.). As the function shown on the right side, X1, X2, X… denote the valuable feature variables of the property, Y denotes the log error, then βs are the coefficients calculated by the training model for predicting the data on the test dataset. The specific log error value predicted from this model is the value to be filled into the submission file.

**3.3** **Phase three: Results Reflection**

**Step 7: Interpreting the results**

R package: Metrics, mae



Package “Metrics” was used to supervise and evaluate the performance of the multiple linear model. Mean Absolute Error (MAE), one of the functions in package “Metrics”, is a metric to score the model (R-Project, n.d.). MAE is also required to be used for the model evaluation by Kaggle. During this reflection phase, the comparisons among the various outputs were frequently made. The parameters were adjusted and the processes were rerun so as to train the best model. Also, the outputs were discussed with the supervisor for getting his feedbacks.

**3.4 Phase four: Results Dissemination**

**Step8: Communicating and distributing the results**

Formats of Data Analysis Report: Rmarkdown and html

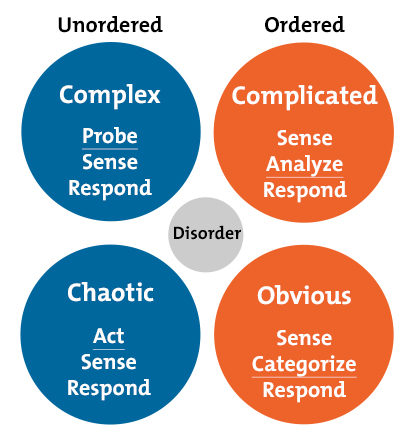
Formats of predictive log error file: csv

The final step is to disseminate the results. The final file predicted log error of 3millions of properties for 6 time-points were submitted on Kaggle for their evaluation with the purpose to prove the significance of the created model. The Rmarkdown and html format analysis are the outcomes when using the R and RStudio to implement the statistic analysis. These files were upload on GitHub and sent to supervisor for various feedbacks. Likewise, this final report is also one type of result dissemination to prove what I have achieved.

1. **Project Management Approach**

**4.1 Scrum**

Scrum was adopted as project management approach for this project. The essential reasons are highlighted as followings:

* With applying Cynefin framework to analyse the project context, it is a complex project producing products that require “Probe-Sense-Respond”.

*Source: MindTools*

* Transparent plan and visible progress made the communication effective between supervisor and me.
* “To do”, “done” “undo” backlogs enabled me to control the progress on the right track.
* Sprint retrospect allowed me to adjust progress as an increment.
* It enabled the project to be completed on time.
* It helped to reduce the risk of submitting failure outcomes.

**4.2 Burn Down Chart and Sprint Backlog**

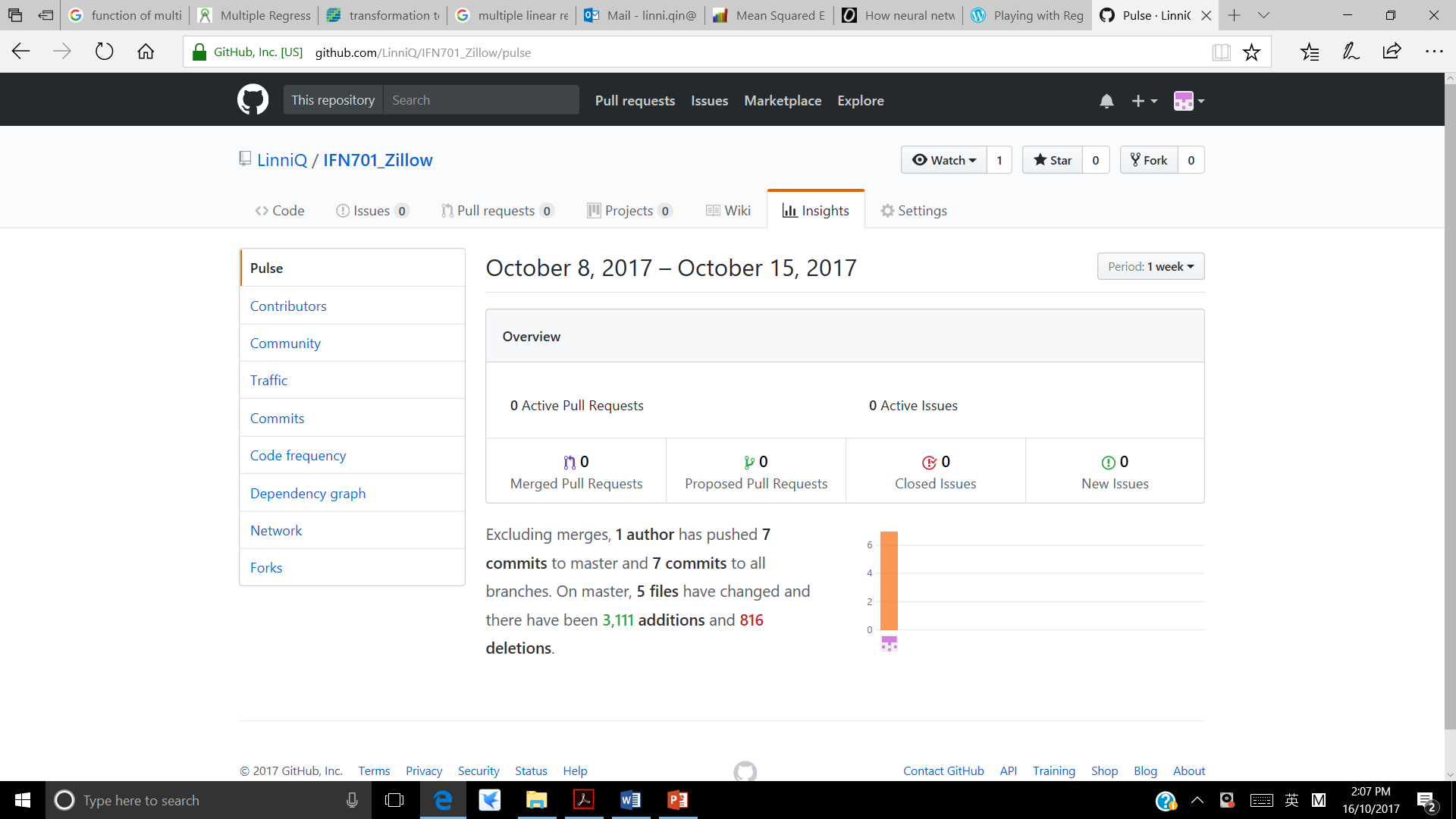
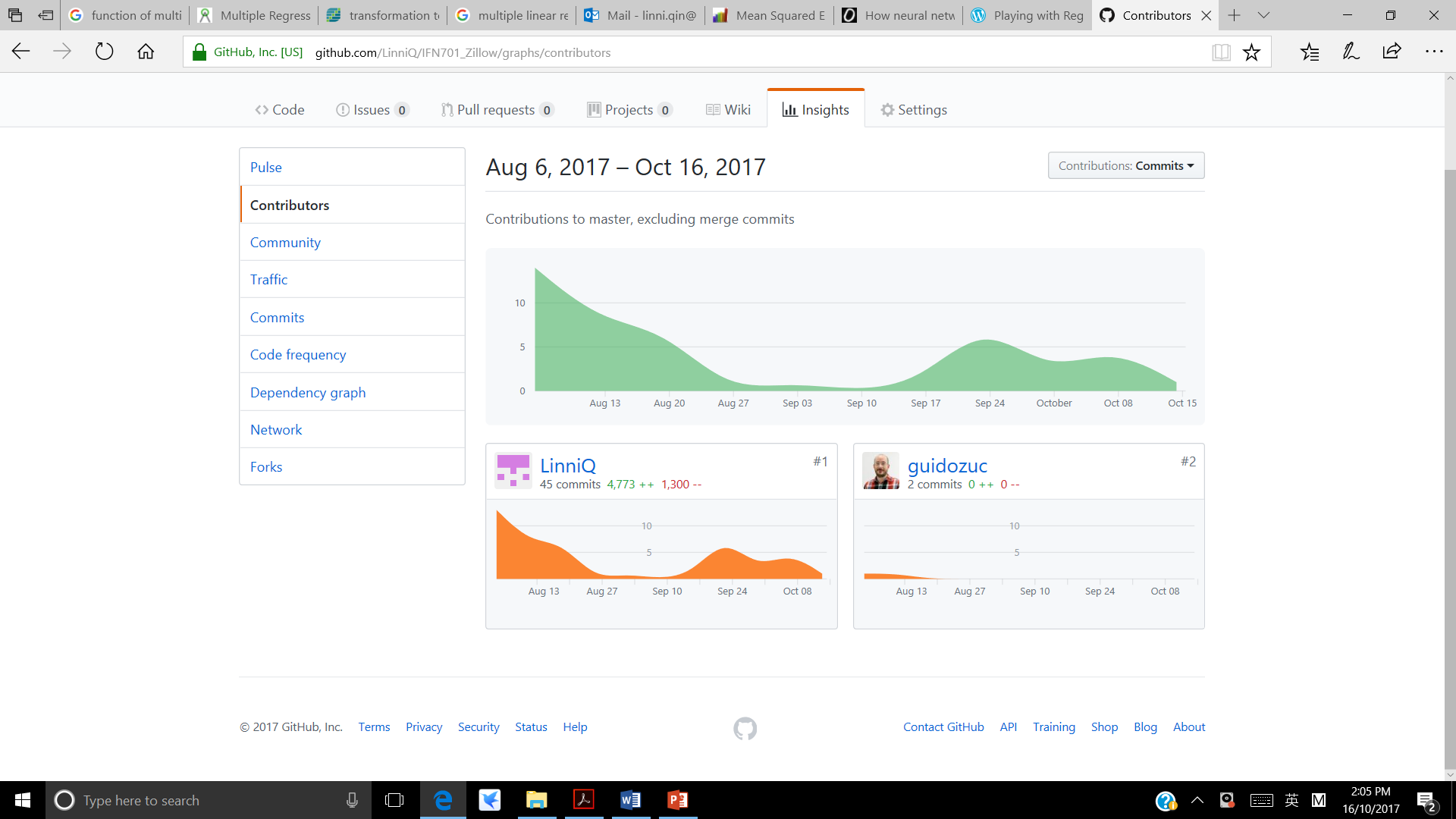
Referring to the table *“MoSCow Prioritised Requirement List”* in section 1.5, the requirement list was transferred into the product backlog and the estimated effort points were the measurement scale for Scrum management. 80 story points of the project backlog were allocated into each two-week sprint backlog until they were completed on time as indicated in the *Burn Down Chart* below.

**Week 10   
Mid-Term Break**

weeks

Story Points

The 3rd party platform GitHub was used as well to save the project processes and works. The below statics insight from GitHub can verify the whole project were executed perfectly as scheduled and that Scrum works properly as the management approaches for this project.



**Phase 1:**

**Data Preparation**

**Phase 2:**

**Data Analysis**

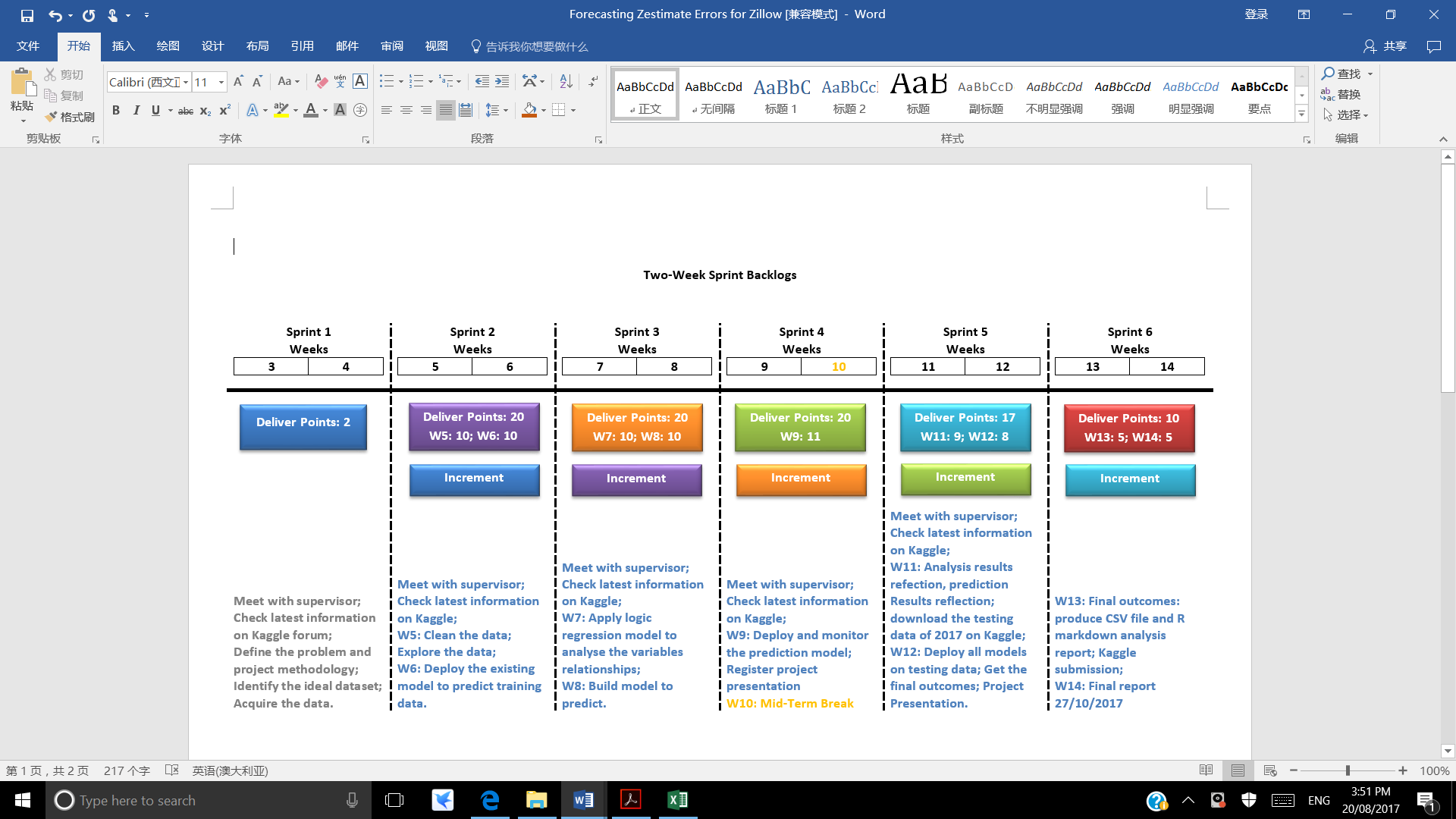
**Phase 3:**

**Result Reflection**

**Oct. 16 ~ Oct. 27:**

**Phase 4: Result Dissemination**

The “Two-Week Sprint Backlogs” below allowed me to check the “To Do”, “Done” and “Undo” tasks effectively.



**Phase 1: Data Preparation**

**Phase 2: Data Analysis**

**Phase 3:**

**Result Reflection**

**Phase 4:**

**Result**

**Dissemination**

1. **Outcomes**

**5.1 “Done” List**

As there are about 2thousands of data analysists globally enrolled as the competitors for Zillow’s project on Kaggle including myself, forum communication might work as a kind of team communication. Therefore, the stakeholders involved in this project will be Kaggle, other competitors, supervisor and me.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***No.*** | ***Phases*** | ***TO DO*** | ***DONE*** | ***UNDO*** |
| 1 |  | Merge transaction dataset and property dataset by common primary key that is the unique “parcelid” for each property. |  |  |
|  |  | Clean the data with filtering the missing data, non-numeric data and duplicated data |  |  |
|  |  | Identify the ideal data for testing purpose |  |  |
|  |  | Explore the hidden relationship between the log error and 57 variables |  |  |
|  |  | Determine a suitable prediction model |  |  |
|  |  | Build the function to fit the prediction model |  |  |
|  |  | Predict the log error |  |  |

**5.2 A Data Analytics Report**

As there are about 2thousands of data analysists globally enrolled as the competitors for Zillow’s project on Kaggle including myself, forum communication might work as a kind of team communication. Therefore, the stakeholders involved in this project will be Kaggle, other competitors, supervisor and me.

**5.3 A Rational Multiple Linear Prediction Model**

As there are about 2thousands of data analysists globally enrolled as the competitors for Zillow’s project on Kaggle including myself, forum communication might work as a kind of team communication. Therefore, the stakeholders involved in this project will be Kaggle, other competitors, supervisor and me.

1. **Discussion**

With considering the risk occurrence probability and the consequence severity level,

1. **Conclusion**

With considering the risk occurrence probability and the consequence severity level,

**Reference**

Business Analyst Learnings (2013, March 5). MoSCoW: Requirements Prioritization Technique. Retrieved from https://businessanalystlearnings.com/ba-techniques/2013/3/5/moscow-technique-requirements-prioritization

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

Cross Validated (2014). Transformation to normality of the dependent variable in multiple regression. Retrieved from https://stats.stackexchange.com/questions/86830/transformation-to-normality-of-the-dependent-variable-in-multiple-regression

Guido, Z. (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

Hadley, W. (n.d.). ggplot2 v2.2.1. Create Elegant Data Visualisations Using the Grammar of Graphics. Retrieved from https://www.rdocumentation.org/packages/ggplot2/versions/2.2.1

Hadley, W. (n.d.). plyr v1.8.4. Tools for Splitting, Applying and Combining Data. Retrieved from https://www.rdocumentation.org/packages/plyr/versions/1.8.4

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

R (n.d). The R Project for Statistical Computing. Retrieved from https://www.r-project.org/

R-Project (n.d.). Package ‘Metrics’. Retrieved from https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

RStudio (n.d.). Why RStudio? The technology to amass data exceeds our abilities to make use of it. Retrieved from https://www.rstudio.com/about/

Statistics Solutions (n.d.). What is Multiple Linear Regression? Retrieved from http://www.statisticssolutions.com/what-is-multiple-linear-regression/

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

**Appendix: Reflection on your learning**