

FORECASTING ZESTIMATE ERROR FOR ZILLOW: A DATASET ANALYSIS REPORT

Linni Qin

Student ID: n9632981

Supervisor: Dr. Guido Zuccon

Unit Code: IFN701

Project Category: A Research Project

QUEENSLAND UNIVERSITY OF TECHNOLOGY

Executive Summary

Based on the dataset analysis for Los Angeles, Orange and Ventura, this report explains mainly the methodologies to build a predictive model for forecasting the individual logarithm error covering 3millions of real estate properties in California for the last quarter of 2017. The purpose of this data analytics is trying to find out a formula to reduce the error margin of Zillow between their estimated market value and the actual sales price. The limited raw data supplied from the Kaggle data science competition platform contain the value relating to 57 physical features for targeted 3millions of properties such as the room number, year of built and square footage, together with historical log error for the transacted properties through the year of 2016 and the first nine months of 2017. Therefore, one tangible outcome is delivered based on this report. That is a well-trained multiple linear regression model which can generate Zillow's log error for a specific in the future.

This analysis report represents the depth insight of the Zillow datasets with a group of statistic data and multiple types of visualization, as well as discusses how to apply a collection of relevant techniques of data mining and machine learning into building a predictive model. Additionally, Scrum is utilized as the approach of project management to ensure these two outcomes are successfully delivered as scheduled within eleven weeks. A data analysis workflow with four phrases and eight steps is conducted to execute the project, which will be discussed in detail in section *Project Methodology*. Then in the section of *Outcomes*, to fit properly the multiple linear regression model, a set of approaches for data filtering, such as identifying missing data, cleaning duplicated data and non-numeric data, are claimed as the major contribution to determining the most valuable. Likewise, the method to classify the proper training data and test data for model training is considered after building the linear model. In particular, as a statistic metric, mean squared error is recommended for evaluating the performance of the forecasting model comparing with other estimators. The data mining and prediction techniques applied in this project might be useful as well for analysing similar datasets and solving similar predictive problems when facing the difficulty to identify the valuable independent variables for fitting the multiple linear regression model.

Table of Contents

Executive Summary	
I.	Project Introduction 1
1.1	Background and Context..... 1
1.2	Problem 1
1.3	Purpose..... 1
1.4	Approach Overview 2
1.5	Scopes..... 2
1.6	Outcomes and Expected Significance 4
II.	Review of Prior Related Work 4
2.1	Zestimate Accuracy Table in LA, OR, VE..... 4
2.2	Available and Required Data..... 5
2.3	Difficulties for Pre-Process..... 6
2.4	Core “TO DO” Tasks..... 7
III.	Project Methodology..... 7
3.1	Phase One: Data Preparation..... 8
3.2	Phase Two: Data Analysis 9
3.3	Phase Three: Results Reflection 10
3.4	Phase Four: Results Dissemination..... 10
IV.	Project Management Approach..... 11
4.1	Scrum..... 11
4.2	Burn Down Chart and Sprint Backlog 11
V.	Outcomes..... 14
5.1	“Done” List..... 14
5.2	Data Analytics Report 15
5.2.1	Significance of the Report 15
5.2.2	Major findings in the Report 16
5.3	Rational Multiple Linear Regression Model 19
5.3.1	Significance of the Model 19
5.3.1	Details of the Model and Relating Outcomes 20
VI.	Discussion 24
6.1	Limits of Applied Techniques 24
6.2	Limits within Project Environment 25
VII.	Conclusion..... 25
Reference	
Appendix 1: Reflections on my learning	
Appendix 2: Project Log Sheets	

I. 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 (Zillow Inc., 2016). 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 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 solely their data on Kaggle for crowdsourcing the more 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. This analysis is based on the public data for round one competition – forecasting the log error for Zestimate.

1.2 Problem

According to the official data, the current median error rate of Zestimate is 5% since the Zestimate was launched in 2006 (Zillow Inc., 2016). 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 \$800,000, its 5% will be \$40,000 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 purposes of analysing the dataset of Zillow is to explore the real world raw data and perform a data science analysis then to build a rational prediction model that would examine the possible improvement on the accuracy of Zestimate. The main objectives of this project are:

- To claim an effective project management approach, a data analysis workflow and a collection of data mining tools and skills for data manipulation;
- To examine the correlation between the log error and the multiple real estate features;
- To identify a well-fit prediction model then the proper size of training test dataset for model training;
- To evaluate the importance of multiple variables for a linear model and model performance;

1.4 Approach Overviews

The *Table 1: 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 per the details of the project approach, it can be referred to section three *Project Methodology*.

Table 1: Data Analysis Workflow

<i>Phase One</i> <i>Data Preparation</i>	<i>Phase Two</i> <i>Data Analysis</i>	<i>Phase Three</i> <i>Results Reflection</i>	<i>Phase Four</i> <i>Results Dissemination</i>
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 2: 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 that can be referred to section four *Project Management Approach*.

Table 2: MoSCow Prioritised Requirement List

<i>ID</i>	<i>Requirement List (Kaggle Inc., n.d.)</i>	<i>Deliverables Priority</i>	<i>Effort Points</i>	<i>Reason</i>
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 contest sponsor, Zillow expects the participants to identify a suitable prediction model to fit their data, to examine the most valuable fields of data that affect the accuracy of the prediction model, to identify the effective size of training and test data.	Must (Guaranteed)	28	These factors are essential to a prediction model so as to estimate the value with higher accuracy.
3	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.	Should (Expected)	10	Outputs should be evaluated.
4	As a data supplier, Zillow wants the participants to analyse the value of their data with techniques of data mining.	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 build one effective prediction model to increase Zestimate’s accuracy.	Could (Possible)	17	The ideal prediction model could definitely improve Zestimate’s performance. However, the level of effectiveness should be evaluated accordingly.
<i>In-Scope Points:</i>			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 mission. Home value prediction could be done in next project.
<i>Out-Scope Points:</i>			20	
<i>Project Total Points:</i>			100	

1.6 Outcome and Expected Significance

A well-trained multiple linear regression model which can generate Zillow's log error for a specific in the future is building with analysing the considerably large amount of real estate data across three counties of California. This analysis report represents not only the depth insight of the Zillow datasets with a group of statistic data and multiple types of visualization but also provide a mini-thesis to conduct how to apply a collection of relevant techniques of data mining and machine learning into building a predictive model. In the section of *Outcomes*, the relevant details of those techniques will be explained.

The data mining and prediction techniques applied in this project is expected to be useful for those data scientists and programmer with R language and so on, when they are analysing similar datasets and solving similar predictive problems with facing the difficulty to identify the valuable independent variables for fitting the multiple linear regression model.

II. Review of Prior Related Work

To understand well the prior related work for this project, the tabular form the tabular form is used mainly in this section to present the current accuracy data of Zestimate in the targeted three counties of California, as well as, the available and required datasets information 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 Zestimate Accuracy Table in LA, OR, VE

Table 3: Date Coverage and Zestimate Accuracy Table (Zillow Inc.,2017)

Data Coverage and Zestimate Accuracy Table							
California Counties	Zestimate Accuracy	Homes on Zillow	Homes with Zestimates	Within 5% of Sales Prices	Within 10% of Sales Prices	Within 20% of Sales Prices	Median Error
Los Angeles	Best	868.3K	777.7K	72.0%	88.3%	95.6%	2.6%
Orange	Best	2.3M	2.1M	62.0%	82.1%	92.6%	3.6%
Ventura	Best	244.5K	227.2K	67.2%	86.0%	97.7%	3.0%

Based on the above table 3, the Zestimate accuracy level on the listed three counties are varied according to the location and the relevant facts data relating to the properties. Comparing the accuracy with other counties in California or other states on Zillow.com, these three counties are with

the best rating level (2017). Their median error rate is better than the Zestimate's general median rate 5%. For example, in Los Angeles, the error margin of Zestimate for a half of the properties is within 2.6% of the actual selling price over 72.0% of the time. On the other side, the other half of the properties are off and more than 2.6%. The error rate will cause the overestimation and underestimation of the real estate transaction as discussed in the section of Introduction. Considering, other counties with worse rating levels in California or other states, it is a critical problem to be addressed.

2.2 Available Data and Required Data

Table 4: Available Data and Required Data (Kaggle Inc., n.d.)

No.	Name of Datasets	Forms	File Description																																																																																																																																																																																																																												
1	Properties_2016	csv	<p>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.</p> <table><tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th><th>G</th><th>H</th><th>I</th><th>J</th><th>K</th><th>L</th><th>M</th><th>N</th><th>O</th><th>P</th><th>Q</th><th>R</th><th>S</th></tr><tr><td>1</td><td>parcelid</td><td>airconditic</td><td>architectu</td><td>basementi</td><td>bathroom</td><td>bedroom</td><td>buildingcla</td><td>buildinggs</td><td>calculated</td><td>decktype</td><td>epic</td><td>finishedflo</td><td>calculated</td><td>finishedsq</td><td>finishedsq</td><td>finishedsq</td><td>finishedsq</td><td>finishedsq</td><td>fireplacec</td></tr><tr><td>2</td><td>10754147</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>6037</td></tr><tr><td>3</td><td>10759547</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>6037</td></tr><tr><td>4</td><td>10843547</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td>73026</td><td></td><td></td><td>73026</td><td></td><td></td><td></td><td>6037</td></tr><tr><td>5</td><td>10859147</td><td></td><td></td><td></td><td>0</td><td>0</td><td>3</td><td>7</td><td></td><td></td><td></td><td>5068</td><td></td><td></td><td>5068</td><td></td><td></td><td></td><td>6037</td></tr><tr><td>6</td><td>10879947</td><td></td><td></td><td></td><td>0</td><td>0</td><td>4</td><td>7</td><td></td><td></td><td></td><td>1776</td><td></td><td></td><td>1776</td><td></td><td></td><td></td><td>6037</td></tr><tr><td>7</td><td>10898347</td><td></td><td></td><td></td><td>0</td><td>0</td><td>4</td><td>7</td><td></td><td></td><td></td><td>2400</td><td></td><td></td><td>2400</td><td></td><td></td><td></td><td>6037</td></tr><tr><td>8</td><td>10933547</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>6037</td></tr><tr><td>9</td><td>10940747</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td>3611</td><td></td><td></td><td>3611</td><td></td><td></td><td></td><td>6037</td></tr><tr><td>10</td><td>10954547</td><td></td><td></td><td></td><td>0</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>6037</td></tr></table>		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	1	parcelid	airconditic	architectu	basementi	bathroom	bedroom	buildingcla	buildinggs	calculated	decktype	epic	finishedflo	calculated	finishedsq	finishedsq	finishedsq	finishedsq	finishedsq	fireplacec	2	10754147				0	0													6037	3	10759547				0	0													6037	4	10843547				0	0						73026			73026				6037	5	10859147				0	0	3	7				5068			5068				6037	6	10879947				0	0	4	7				1776			1776				6037	7	10898347				0	0	4	7				2400			2400				6037	8	10933547				0	0													6037	9	10940747				0	0						3611			3611				6037	10	10954547				0	0													6037
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S																																																																																																																																																																																																												
1	parcelid	airconditic	architectu	basementi	bathroom	bedroom	buildingcla	buildinggs	calculated	decktype	epic	finishedflo	calculated	finishedsq	finishedsq	finishedsq	finishedsq	finishedsq	fireplacec																																																																																																																																																																																																												
2	10754147				0	0													6037																																																																																																																																																																																																												
3	10759547				0	0													6037																																																																																																																																																																																																												
4	10843547				0	0						73026			73026				6037																																																																																																																																																																																																												
5	10859147				0	0	3	7				5068			5068				6037																																																																																																																																																																																																												
6	10879947				0	0	4	7				1776			1776				6037																																																																																																																																																																																																												
7	10898347				0	0	4	7				2400			2400				6037																																																																																																																																																																																																												
8	10933547				0	0													6037																																																																																																																																																																																																												
9	10940747				0	0						3611			3611				6037																																																																																																																																																																																																												
10	10954547				0	0													6037																																																																																																																																																																																																												
2	Train_2016_v2	csv	<p>A transaction file contains the actual log error for around 90thousands properties which have been transacted through 2016 (01/01/2016 – 31/12/2016) and their detail transaction date.</p> <table><tr><th></th><th>A</th><th>B</th><th>C</th></tr><tr><td>1</td><td>parcelid</td><td>logerror</td><td>transactiondate</td></tr><tr><td>2</td><td>11016594</td><td>0.0276</td><td>1/01/2016</td></tr><tr><td>3</td><td>14366692</td><td>-0.1684</td><td>1/01/2016</td></tr><tr><td>4</td><td>12098116</td><td>-0.004</td><td>1/01/2016</td></tr><tr><td>5</td><td>12643413</td><td>0.0218</td><td>2/01/2016</td></tr><tr><td>6</td><td>14432541</td><td>-0.005</td><td>2/01/2016</td></tr><tr><td>7</td><td>11509835</td><td>-0.2705</td><td>2/01/2016</td></tr><tr><td>8</td><td>12286022</td><td>0.044</td><td>2/01/2016</td></tr><tr><td>9</td><td>17177301</td><td>0.1638</td><td>2/01/2016</td></tr><tr><td>10</td><td>14739064</td><td>-0.003</td><td>2/01/2016</td></tr></table>		A	B	C	1	parcelid	logerror	transactiondate	2	11016594	0.0276	1/01/2016	3	14366692	-0.1684	1/01/2016	4	12098116	-0.004	1/01/2016	5	12643413	0.0218	2/01/2016	6	14432541	-0.005	2/01/2016	7	11509835	-0.2705	2/01/2016	8	12286022	0.044	2/01/2016	9	17177301	0.1638	2/01/2016	10	14739064	-0.003	2/01/2016																																																																																																																																																																																
	A	B	C																																																																																																																																																																																																																												
1	parcelid	logerror	transactiondate																																																																																																																																																																																																																												
2	11016594	0.0276	1/01/2016																																																																																																																																																																																																																												
3	14366692	-0.1684	1/01/2016																																																																																																																																																																																																																												
4	12098116	-0.004	1/01/2016																																																																																																																																																																																																																												
5	12643413	0.0218	2/01/2016																																																																																																																																																																																																																												
6	14432541	-0.005	2/01/2016																																																																																																																																																																																																																												
7	11509835	-0.2705	2/01/2016																																																																																																																																																																																																																												
8	12286022	0.044	2/01/2016																																																																																																																																																																																																																												
9	17177301	0.1638	2/01/2016																																																																																																																																																																																																																												
10	14739064	-0.003	2/01/2016																																																																																																																																																																																																																												
3	Properties_2017	csv	<p>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.</p> <p>(same format with properties_2016)</p>																																																																																																																																																																																																																												

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	Required Log error	csv	<p>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.</p> <table><tr><th>Parcelld</th><th>201610</th><th>201611</th><th>201612</th><th>201710</th><th>201711</th><th>201712</th></tr><tr><td>10754147</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10759547</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10843547</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10859147</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10879947</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10898347</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10933547</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10940747</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>10954547</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr></table>	Parcelld	201610	201611	201612	201710	201711	201712	10754147	0	0	0	0	0	0	10759547	0	0	0	0	0	0	10843547	0	0	0	0	0	0	10859147	0	0	0	0	0	0	10879947	0	0	0	0	0	0	10898347	0	0	0	0	0	0	10933547	0	0	0	0	0	0	10940747	0	0	0	0	0	0	10954547	0	0	0	0	0	0
Parcelld	201610	201611	201612	201710	201711	201712																																																																			
10754147	0	0	0	0	0	0																																																																			
10759547	0	0	0	0	0	0																																																																			
10843547	0	0	0	0	0	0																																																																			
10859147	0	0	0	0	0	0																																																																			
10879947	0	0	0	0	0	0																																																																			
10898347	0	0	0	0	0	0																																																																			
10933547	0	0	0	0	0	0																																																																			
10940747	0	0	0	0	0	0																																																																			
10954547	0	0	0	0	0	0																																																																			
6	Zillow_data_dictionary	csv	A dictionary to explain the field name of 57 property features.																																																																						
7	Sample_submission	csv	A sample file to ensure the final submission is in the correct format.																																																																						

2.3 Difficulties for Pre-process

There are two difficulties need to be overcome during the phase one of data analysis workflow. As mentioned in the overview section of project approaches, there are four steps. At 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 at the right-hand side Figure 1, the log error required to be predicted is the difference between

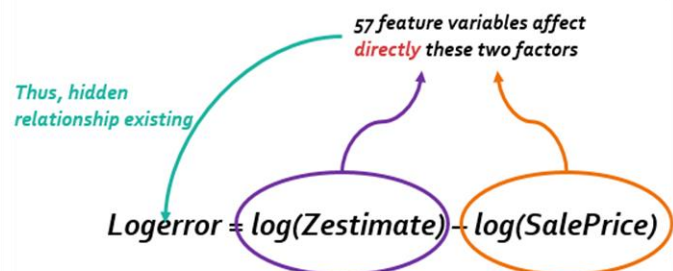


Figure 1: Log Error Function

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 2.4.

2.4 Core “TO DO” Tasks

Table 5: To Do List

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	Examine the hidden relationship between the log error and 57 variables
4	Two	Determine a suitable prediction model
5	Two	Identify the ideal data for model training and testing
6	Two	Build the function with proper size of variables to fit the prediction model
7	Two	Predict the log error
8	Three	Evaluate the accuracy of the model, reflect the results
9	Four	Present the deliverables: a data analytics, a rational prediction model and the estimated log error value

This *Table 5: To Do list* contains 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*.

III. Project Methodology

Considering the four major phases integrating with eight steps to accomplish this whole project as well as 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

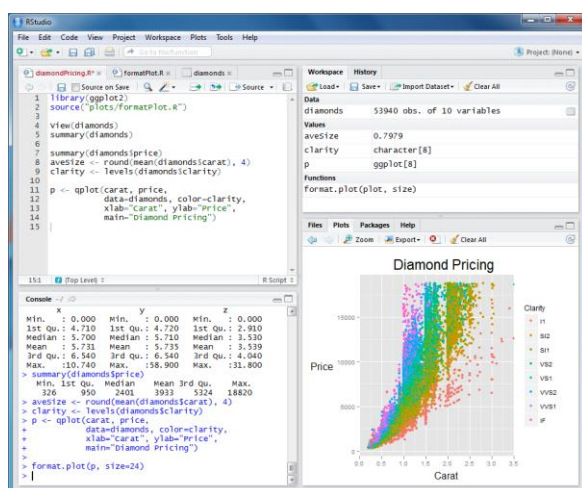
The ideal first action is to merge the transaction dataset and the property dataset by common primary key which is the unique field of “parcelid” for each property. There are two reasons. The first, transaction dataset cannot predict the full 3millions log error for the property without knowing their 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 3.3.2 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.).

Figure 2: RStudio Interface



As shown as left-hand side *Figure 2*, 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 dataR packages: plyr and dplyr

Although there are data of 57 variables listed for 3 millions 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 and corrplot

Starting from this step, it began to investigate 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 visualization functions were used like R package ggplot which can plot or cluster the datasets (Hadley, W., n.d.). Package of corrplot is able to display the graphics for the correlation matrix with various methods such as “circle”, “number” and “pie” and so on (R-Project, 2016).

Step 6: Statistical prediction and modelling the dataPrediction Model: Multiple Linear Regression*Figure 3: Linear Regression Function (Cross Validated, 2014)*

There are many effective and popular prediction models such as decision tree, logistic regression and support vector machine. However, the predicted value classified by these three models are binary

values. That means they can identify the log error can increase or decline, but cannot return a specific value. Multiple linear regression is a type of machine learning model that is not only able to identify

The diagram shows the linear regression equation:
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$
 Annotations include:

- A red circle around Y with a red arrow pointing to the text: "response, dependent variable, observation, 'y-variable'"
- A green circle around x_1 with a green arrow pointing to the text: "predictor, 'x-variable', independent variable, explanatory variable"
- An orange circle around β_2 with an orange arrow pointing to the text: "coefficient"
- A blue bracket under the terms $\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ with the text: "linear predictor"
- A purple circle around ϵ with a purple arrow pointing to the text: "random error, 'noise'"

the strength of the effect that the independent variables have on a dependant variable but also it can forecast future value and return a specific value (Statistics Solutions, n.d.). As the function shown on the *Figure 3*, $X_1, X_2, X...$ denote the valuable feature variables of the property, Y denotes the log error, then β_s 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, mse

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

Package “Metrics” was used to supervise and evaluate the performance of the multiple linear model. Mean Squared Error (MSE), one of the functions in package “Metrics”, is a metric to score the accuracy of model with averaging the distance between the estimated value and the actual value as indicated in the above function (R-Project, 2015). Comparing to other scale-dependent estimator like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), MSE is more effective to calculate the gradient loss (R-Project, 2017). On the other hand, the predictive task concerns much more about the large error and small error then MSE is able to make the larger error large and small error smaller. MSE 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.

IV. 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”.

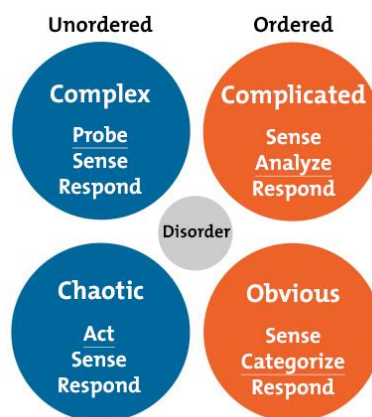
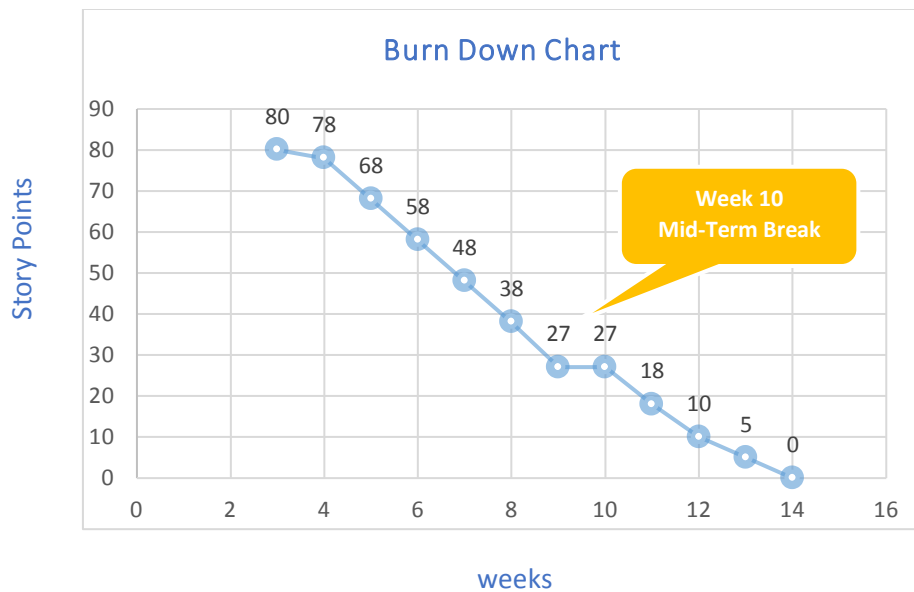


Figure 4: MindTools

- 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.



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.

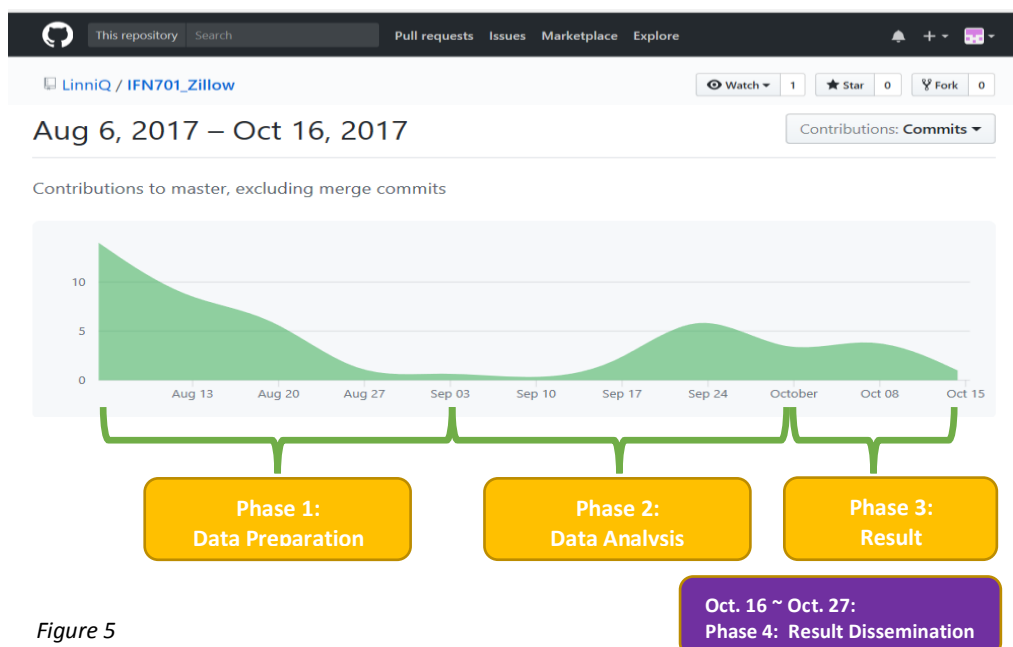
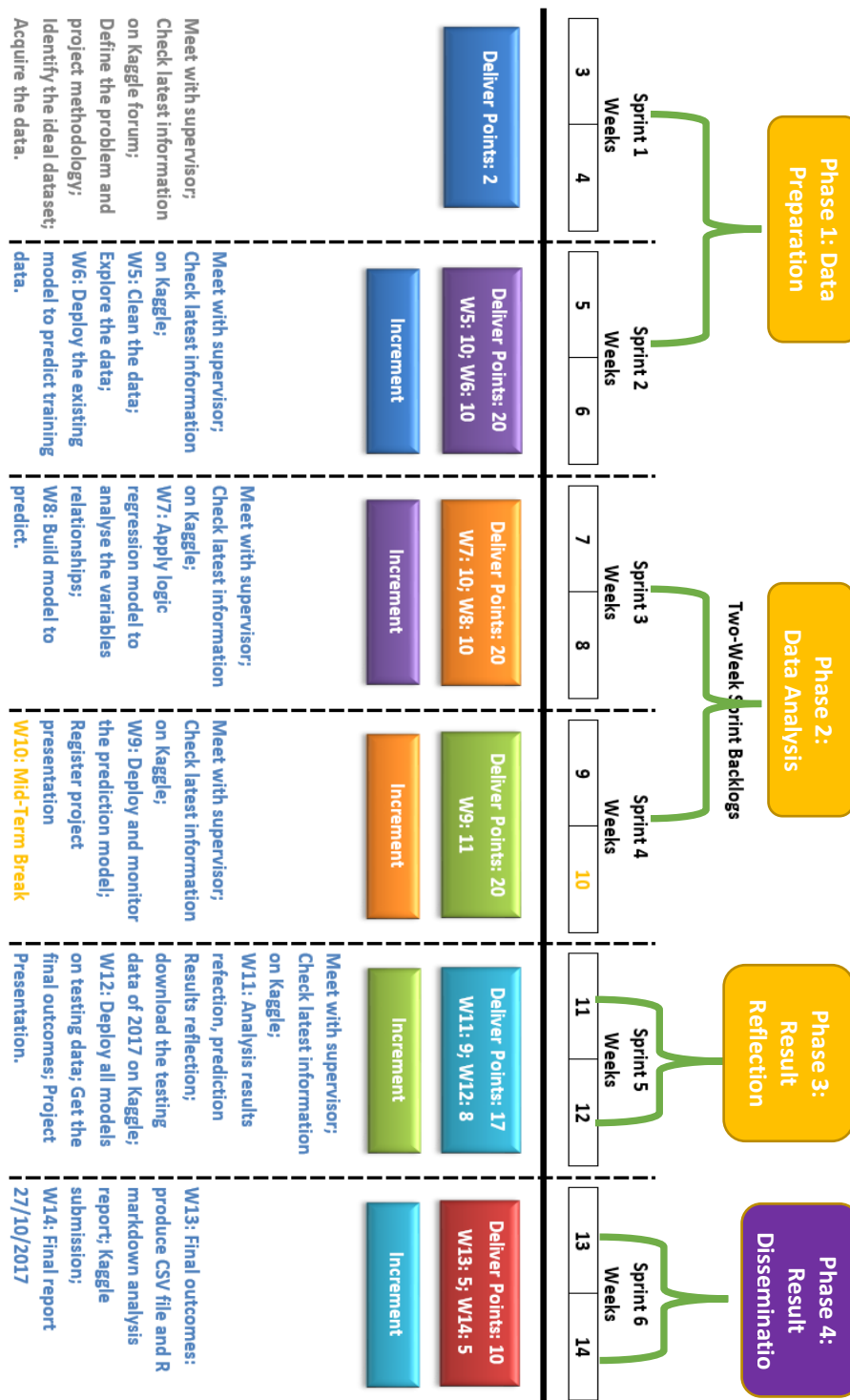


Figure 5

The Figure 6: Two-Week Sprint Backlogs below allowed me to check the “To Do”, “Done” and “Undo” tasks effectively.

Figure 6



V. Outcomes

5.1 “Done” List

Table 6: Done List contains a general list of “Done” works for checking whether the assumed “To do” tasks were completed successfully or not. In the meantime, the significance of the “Done” items can be referred to section 5.2 and 5.3. The particular reasons for those tasks with limited commits will be discussed in section six. Two outcomes were delivered after getting all tasks done, including a data analytics report in the forms of Rmarkdown and html, together with a rational multiple linear regression model with its calculated predicted log error.

Table 6: Done List

No.	Phases	TO DO	DONE	Note
1	One	Merge transaction dataset and property dataset by common primary key that is the unique “parcelid” for each property.	Two merged datasets for 2016 and 2017 transacted properties. They play as the essential fundamental datasets for exploring the correlations between the log error and 57 variables and other types of data patterns.	
2	One	Clean the data with filtering the missing data, non-numeric data and duplicated data	27 variables after filtering the missing data, duplicated data and non-numeric data.	
3	One	Examine the hidden relationship between the log error and 57 variables	No direct correlation existing between and log error and other variables	Obstacles
4	Two	Determine a suitable prediction model	Multiple Linear Regression Model	
5	Two	Identify the ideal data for model training and testing	Training data: trials with 1-month data, 3months and all existing data; Testing data: data of next one month	
6	Two	Build the function with proper size of variables to fit the prediction model	Final 16 variables after filtering	
7	Two	Predict the log error	3millions of log error for properties in 10/2016, 11/2016, 12/2016 3millions of log error for properties in 10/2017, 11/2017, 12/2017	
8	Three	Evaluate the accuracy of the model, reflect the results	Mean Squared Error with 25 times of reflection	
9	Four	Present the deliverables: a data analytics, a rational prediction model and the estimated log error value	GitHub.com, Kaggle.com Presentation and Final Report	Future research

5.2 Data Analytics Report

5.2.1 Significance of the Report

The outstanding of the data analytics report can be implied from the increasing viewer number of GitHub users. Since the report was coded and generated in the forms of Rmarkdown and html respectively, they are saved and uploaded in the assigned GitHub public repository with meaningful names containing “Zillow” and “Prediction”. The GitHub users or other data scientists who are interested in analysing Zillow’s data with language R are able to access to my reports and original codes freely via internet.

It is noted from the two statics screenshots (*Figure 7 and Figure 8*) of my GitHub repository at the right-hand side that the number of views for my Zillow data analytics reports increased from 118 on October 16, 2017 to 187 on October 22 (GitHub, n.d.). The number of views increased by 70 in one week with average 12 views in

one day. The high density of hitting on my report happened during the last two days before the submission deadline (October 16, 2017, PTC) on Kaggle.com. The highest number of views reached 40 in one day. These meaningful data imply the referring valuable of this data manipulation thesis.

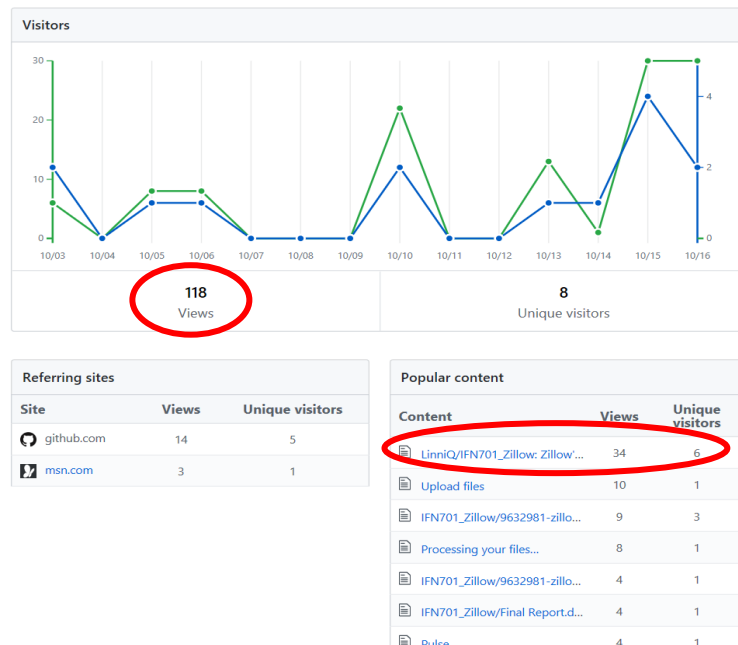


Figure 7

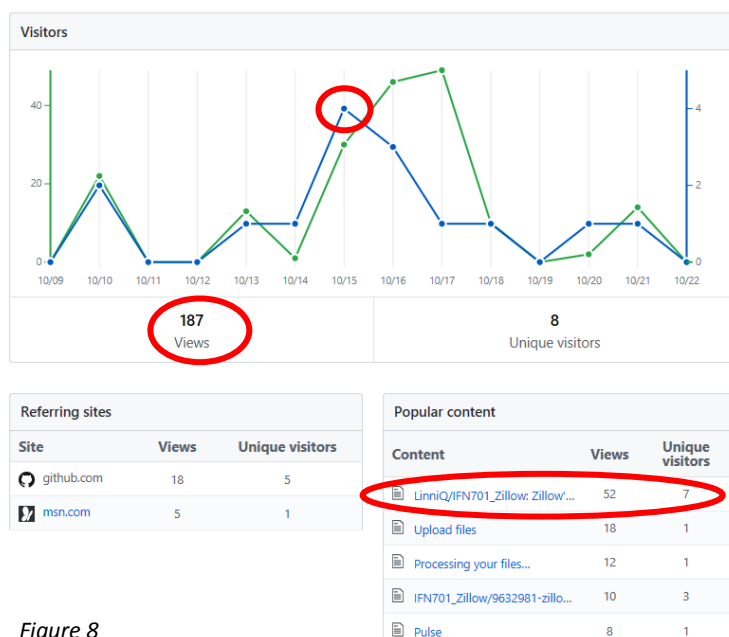


Figure 8

5.2.2 Major findings in the Report

Look into log error itself

First of all, the log error distribution in both 2016 and 2017 were explored in both forms of visualization and statistics in order to find any hidden pattern among the log error value itself. The left jitter plot (*Figure 9*) is the log error distribution in 2016 and the right jitter plot (*Figure 10*) below is the log error distribution in 2017. However, they appear without any significant difference, excepting both groups of the log error show a high tense of distribution around ± 0.1 . It cannot tell too much from the distribution of these two years.

Figure 9: Log Error Distribution in 2016

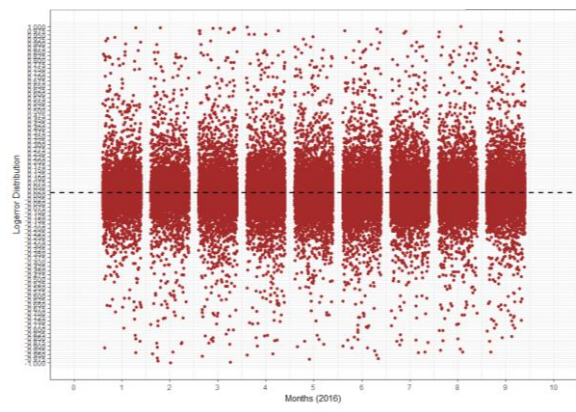
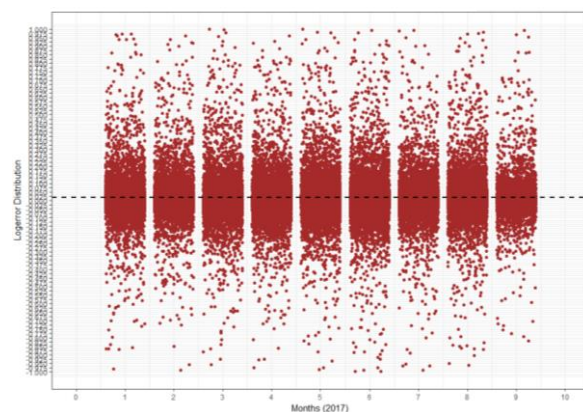


Figure 10: Log Error Distribution in 2017



The other way was used to look into the log error itself is by statistics. The calculated statistics metrics involved the values for Median, Mean, Max value, Min value and Standard Deviation.

Table 7: Statistic Value

Year	Median	Mean	Max	Min	Std.
2016	0.005	0.010929	4.737	-4.605	0.1623305
2017	0.0067	0.016755	5.262	-4.655	0.1708873

Studying from the above statics values in *Table 7*, the accuracy of Zestimate of 2017 appears a bit worse than 2016 because the median, mean and standard deviation of 2017 perform slightly higher than that of 2016. Meanwhile, the outliers of 2017's log error goes a bit far away of its median value while comparing that with 2016.

The plotting image (*Figure 11*) on the right side showing the change of log error mean for both 2016 and 2017 by monthly. The blue line denotes log error monthly mean for 2017 and the red line denote log error monthly mean. They are compared within the same period, from January to September. This plot provides more confidence to state that the performance of Zestimate in the first nine months of 2017 did slightly worse than what it did in the same period of 2016. However, this finding cannot imply strongly that the log error of last quarter of 2017 will go worse. More factors to influence the log error demand to be investigated.

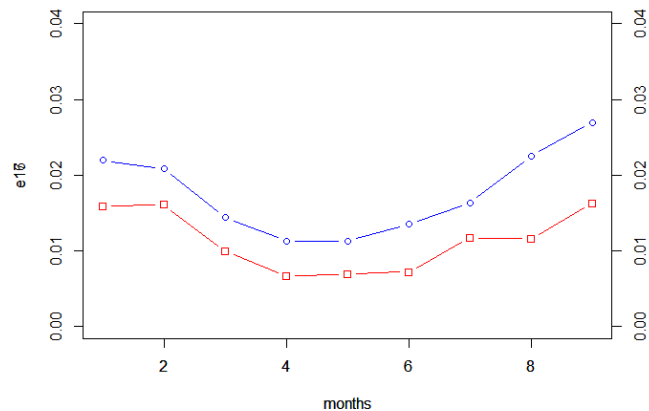
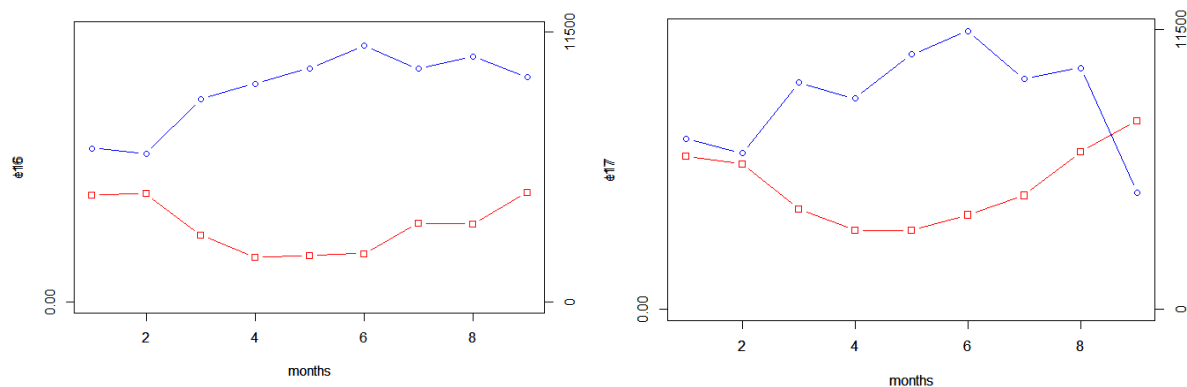


Figure 11: Mean of Log Error in 2016 / 2017

Look into the correlation between the transaction frequency and the log error

The correlation between the log error and the transaction date of the purchased property is the next factor being examined. As indicated as two plots below (*Figure 12*), the blue lines denote the monthly transaction frequencies. The red lines denote the monthly log error means. The high log errors happened when the transaction frequency stayed low. A stable fluctuation of the low error happened when the transaction frequency stayed with high volume. These two variables run in the opposite directions as shown. Meanwhile, their correlation rates are -0.6387566 and -0.7372831 respectively for year 2016 and 2017.

Figure 12: Correlation between the log error and transaction frequency in 2016 / 2017



Regarding to the statistics explanation, the correlation is represented within the range of -1, 0 and 1. 0 indicates no correlation and 1 denotes a positive correlation (Investopedia, n.d.). The negative correlation as found above between the log error and the transaction frequency denotes that

while transaction number increases the log error decreases. There hides certain high negative relationship between these two factors. However, there is transaction data for the whole year of 2016 only. It is lack of strong evidence to prove that the transaction frequency of 2017 stays low in the last season as 2016 did. Furthermore, the transaction date is the information for the one third of properties that have been transacted only. For the large amount of the rest properties required to be predicted, there is only the value of physical features. Therefore, the transaction value might be one important factor to affect the log error, but the error might not be predicted based on the transaction trends.

Look into the correlation between log error and the property features

According to Zillow, the accuracy of Zestimate depends on the available data supplied by the homeowners. Datasets of Prop_16 and Prop_17 contain those value from the property owners. Thus, to investigate the relationship between the log error and the property variables, transaction files and the property files were combined respectively by the primary key – the field of parcelid. Unfortunately, the output file that is plotted as the correlation matrix below (Figure 13) tells that there is no correlation existing among the log error and the 57 elements completely.

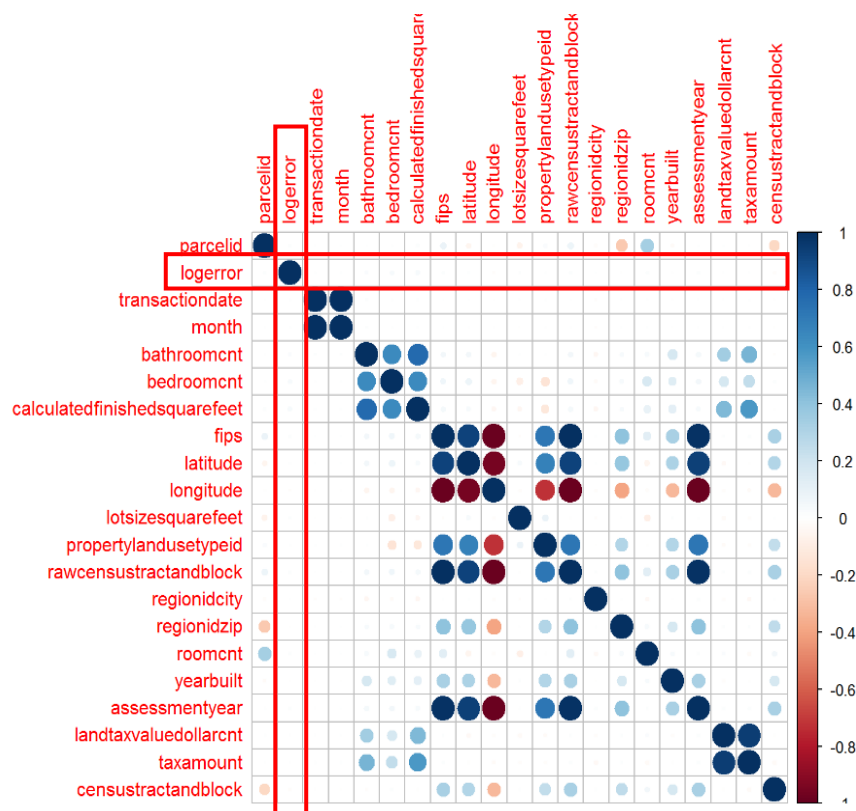


Figure 13: Correlation between the log error and property factors

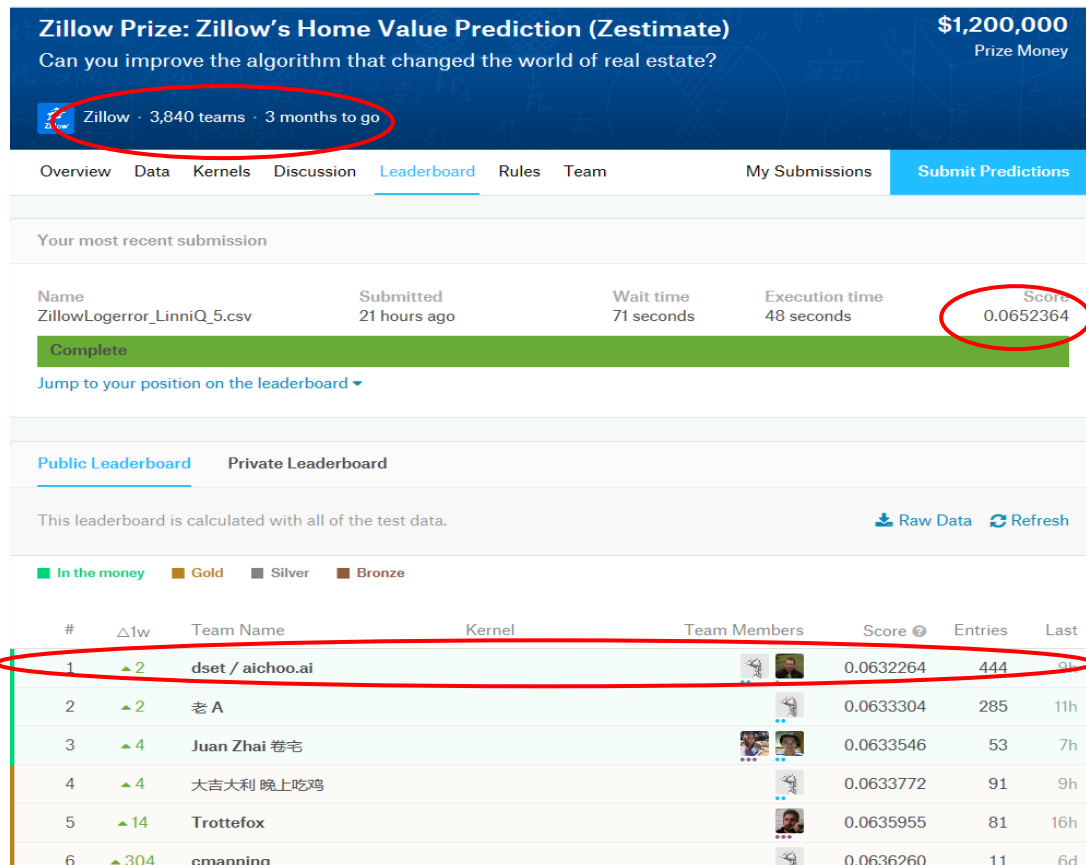
Till now, all the available data and elements that were assumed to affect the log error have been investigated through. Even though there is no clue which elements might affect greatly the accuracy of log error, it still demands a prediction model that can handle these datasets and generate the estimated value.

5.3 Rational Multiple Linear Prediction Model

5.3.1 Significance of the Model

With the purpose to check the significance of the outcome model, the csv file generated by it with containing the log error value for 3millions of properties in California in 10/2016, 11/2016, 12/2016, 10/2017, 11/2017 and 12/2017 were submitted on Kaggle.com. As told by the screenshot from Kaggle.com below (Figure 14), my temporary score is 0.0652364 which is evaluated by Zillow's self-defined scoring mechanism. This is really closed to the first place's score with a tiny gap of 0.002. The final score will be released in three months as October, November and December of 2017 which are the period for sales tracking according to the competition rules.

Figure 14: Kaggle Score



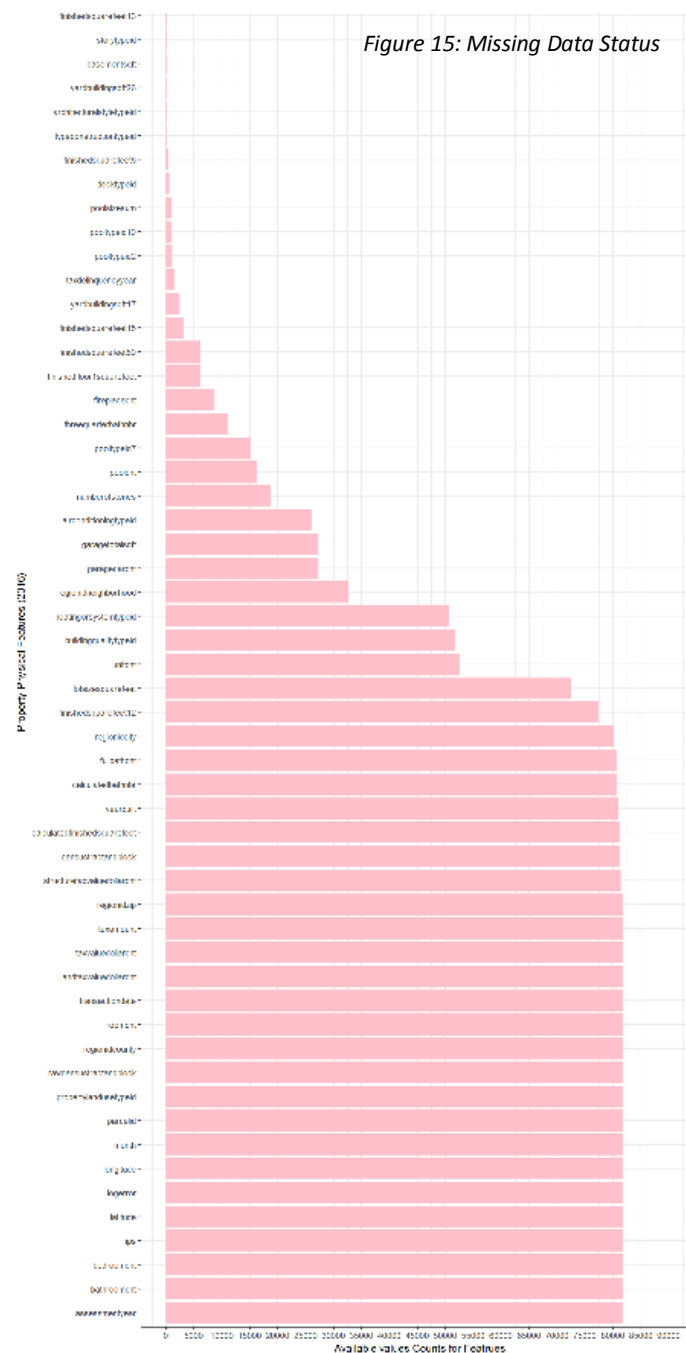
5.3.2 Details of the Model and Relating Outcomes

Valuable variables for the linear model

The justification to choose multiple linear regression prediction model has been discussed in section 3.2. After the choice of prediction model has been made, the model needs a rational size of variable to fit the model. Otherwise, the dependent value of log error might be overfit if all 57 variables are used within the model function. Learning from whole section 5.2.2, the most valuable data in this data frame cannot be selected by the way of identifying the correlations. Filtering is the method used in this analysis progress to identify the elite independent variables.

Three proceeds were employed in the operation of filtering. The first step is to explore the status of missing data. As shown on the right-side plot (Figure 15), there are 27 features containing nearly and or full values in the datasets.

With the information supplied from Andrew Martin who is a Data Scientist at Zillow, there are some duplicate data for the features of the properties (Kaggle, n.d.). For example, fields of bathroomcnt, calculatedbathnbr and fullbathcnt mean one same thing. Fields of calculatedbathnbr and fullbathcnt are calculated by the property assessor. Meanwhile, bathroomcnt is calculated repeatedly by Zillow. There are some similar types of duplicated value for other fields as well such as the regionidzip number and the regionidcounty. They all denotes the three counties (Los Angeles, Orange and Ventura) in California in different ways. Thus, cleaning the repeated data is the second essential proceed in data filtering.



The last step is to clean the non-numeric data. Since the model needs all the data in numeric type to do the calculation then returns a specific number of estimated log error after model training. Therefore, the character value or logical value demands to be cleaned as well. So far, after the filtering proceeding, there are 16 valuable variables can be applied into the function of prediction linear model which is shown in the function below. The trials of taking out of some variables like latitude and longitude, the evaluation metric MSE (mean squared error) for the model went worse. As a result, this model can be considered rationally.

```
Model = lm(logerror ~ bathroomcnt + bedroomcnt + calculatedfinishedsquarefeet
+ fips + latitude + longitude + lotsizesquarefeet + propertylandusetypeid +
rawcensustractandblock + regionidcity + regionidzip + roomcnt + yearbuilt +
taxvaluedollarcnt + taxamount + censustractandblock, data = traindata)
```

Quality of training and test data

After building the linear model to execute the prediction, the size of the training data and test data demand optimized because the accuracy of the model depends heavily on the quality and quantity of the training dataset. At the same time, since the target of this project is to predict the log error month by month for the last three months of the year 2016 and 2017, there are three solutions to separate the data into training set and testing set until the ideal combination is found. What are supposed to do is to use one-month data as training data to predict the next month data, or use three-month data to predict the next month data, or use the existing data to predict the next month data. Accuracy metric of mean squared error was applied in evaluating the performance of each group. As the result, the combination with the lowest averaging score was chosen as the most proper sizes to apply in the final prediction task. As shown in the *Table 8: Avergae Mean Squared Error for Training*, the average MSE scores were calculated. The result by using one-month data to predict the next month data is 0.027734. The average MSE score for using three-month data to predict the next month data is 0.02455 and the average MSE score for using all the previous data to predict the next month data is 0.024566. The score of the third group is very close to the score of the second group. There difference is tiny with 0.000016. However, considering the large scale of the forecasting data, the minor difference might mean greatly in proceeding the estimation. Finally, the best combination of the training data and test data is to use three-month data to predict the next month data. With using this combination of the training and test data, the MSE score is the best one 0.02455.

Table 8: Average Mean Squared Error for Training

Group 1: One-month data to predict next month data			Group2: Three-month data to predict next month data			Group3: All previous data to predict next month data		
Training Data	Testing Data	MSE	Training Data	Testing Data	MSE	Training Data	Testing Data	MSE
01/2016	02/2016	0.03943015						
02/2016	03/2016	0.02967089						
03/2016	04/2016	0.02761392	01/2016 -	04/2016	0.02763571			
			03/2016					
04/2016	05/2016	0.02275491	02/2016 -	05/2016	0.02266364			
			04/2016					
05/2016	06/2016	0.02305983	03/2016 -	06/2016	0.02296181			
			04/2016					
06/2016	07/2016	0.02296013	04/2016 -	07/2016	0.02295128			
			06/2016					
07/2016	08/2016	0.02606143	05/2016 -	08/2016	0.02595377			
			07/2016					
08/2016	09/2016	0.02271136	06/2016 -	09/2016	0.02255886			
			08/2016					
09/2016	10/2016	0.01816775	07/2016 -	10/2019	0.01799051	01/2016 -	10/2016	0.01799051
			09/2016			09/2016		
10/2016	11/2016	0.01945185	08/2016 -	11/2016	0.01941113	01/2016 -	11/2016	0.01942081
			10/2016			10/2016		
11/2016	12/2016	0.04558319	09/2016 -	12/2016	0.03624948	01/2016 -	12/2016	0.03628793
			11/2016			11/2016		
Average		0.025374	Average		0.024264			
Last Quarter Average		0.027734	Last Quarter Average		0.02455	Last Quarter Average		0.024566

Narrow the Standard Derivation

So far, the log error of October has been estimated. With the assumption that there would be bigger log error variation in the last quarter of 2017 as similar as the one in 2016, the method to estimate the log error value for the next November and December is to take narrow down the dispersion of the predicted value. The purpose of this step is to make the forecasted value much closer to the expected actual data. Reviewing the standard derivation for 2016 and 2017 in section 5.2.2, they are around 0.16 and 0.17 respectively. Therefore, the estimated log error for October and multiply it by the difference between one and standard derivation. Thus, the multiple value should be around 0.845. Resulting from this

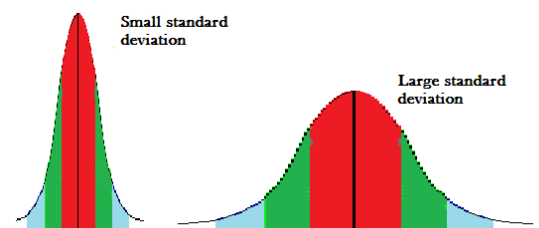


Figure 16: Standard Deviation

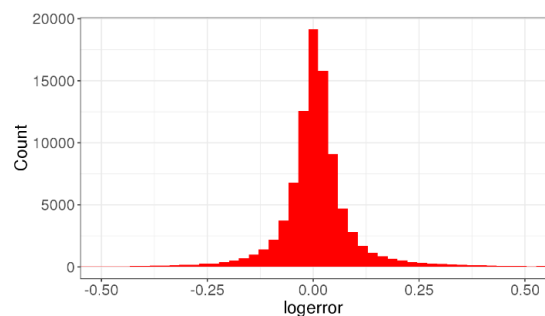


Figure 17: Zillow Log Error Distribution

operation, the whole prediction dataset for the last quarter will match the total distribution trend of the previous log error data as demonstrated in the *Figure 17*.

This method was proved to be effective while Zillow's scoring mechanism on Kaggle.com with using different scale for this value. Five output csv files with 3millions of log error data multiplying by varying values: 0.983, 0.83, 0.6, 0.7, 0.845 to predict the value for November and December, the best score of the whole prediction task on Kaggle results from narrowing down the standard derivation by 15.5%. The best score is 0. 0.0652364 as listed below in *Table 9: Five Trials of Submission*, also as released in *Figure 14: Kaggle Score*.

Table 9: Five Trials of Submission

File No.	Trial Parameter	Multiple by	Narrow Down %	Kaggle Score
1	1 – 0.017	0.983	2%	0.0653246
2	1 – 0.17	0.83	20%	0.0652873
3	1 – 0.4	0.6	40%	0.0653492
4	1 – 0.3	0.7	30%	0.0653218
5	1 – 0.155	0.845	15.5%	0.0652364

Example of the submission data in csv

Finally, all the required data are obtained and created into the csv file for submission as shown in the right-hand *Figure 18: Final CSV File*. As stated previously, the purpose of this project is not for competition. The action to submit the final file is for evaluating and supporting the significance the outcomes. This final file contains forecasted log error for 3millions of properties in three counties (Orange, Los Angeles and Ventura) in California for six timepoints: 10/2016, 11/2016, 12/2016, 10/2017, 11/2017, 12/2017. The effectiveness of the output values has been justified and reflected by the above techniques thoroughly and seriously.

parcelid	201610	201611	201612	201710	201711	201712
10711725	0.012524	0.009769	0.00762	0.010667	0.00704	0.010667
10711726	0.02068	0.016131	0.012582	0.01988	0.013121	0.01988
10711727	0.01681	0.013112	0.010227	0.015419	0.010177	0.015419
10711728	0.022791	0.017777	0.013866	0.018082	0.011934	0.018082
10711729	0.021237	0.016565	0.012921	0.015363	0.01014	0.015363
10711730	0.023108	0.018024	0.014059	0.018205	0.012015	0.018205
10711731	0.023436	0.01828	0.014258	0.018478	0.012195	0.018478
10711732	0.017066	0.013311	0.010383	0.013222	0.008726	0.013222
10711733	0.017197	0.013414	0.010463	0.018797	0.012406	0.018797
10711734	0.020258	0.015802	0.012325	0.016741	0.011049	0.016741
10711735	0.023352	0.018214	0.014207	0.018291	0.012072	0.018291
10711736	0.022289	0.017385	0.013561	0.015691	0.010356	0.015691
10711737	0.019082	0.014884	0.011609	0.02087	0.013774	0.02087
10711738	0.019848	0.015481	0.012075	0.017805	0.011751	0.017805
10711739	0.016348	0.012752	0.009946	0.015631	0.010317	0.015631
10711740	0.02098	0.016364	0.012764	0.017101	0.011287	0.017101
10711741	0.019083	0.014885	0.01161	0.019728	0.013021	0.019728
10711742	0.023485	0.018318	0.014288	0.018415	0.012154	0.018415
10711743	0.004115	0.00321	0.002503	0.010244	0.006761	0.010244
10711744	0.015844	0.012359	0.00964	0.018846	0.012438	0.018846
10711745	0.020455	0.015955	0.012445	0.01692	0.011167	0.01692
10711746	0.019585	0.015276	0.011916	0.015075	0.00995	0.015075
10711747	0.027052	0.0211	0.016458	0.02945	0.019437	0.02945
10711748	0.022208	0.017322	0.013511	0.015698	0.010361	0.015698
10711749	0.026463	0.020641	0.0161	0.017809	0.011754	0.017809
10711750	0.018593	0.014502	0.011312	0.00429	0.002832	0.00429
10711751	0.003248	0.002533	0.001976	0.000204	0.000135	0.000204
10711752	0.020752	0.016186	0.012625	0.010391	0.006858	0.010391

Figure 18: Final CSV File

VI. Discussion

Reviewing the objectives discussed in section 1.3, this data analytic conducts the method to build the formula for predictive modelling. The increasing views on GitHub about the data analysis report indicates the usefulness of suggested data mining strategies, tools and data manipulation skills. The created linear model works well so far with achieving an enlightened score on Kaggle.com for its output – 3 millions of predicted log error. However, as per the information on Kaggle.com, there are near 4,000 teams taking part in Zillow's competition. The temporary ranking after evaluating by Zillow's proprietary scoring mechanism still means the rough performance of these 4,000 prediction models. Even though there is only a minor gap of 0.002 between my best score and the first place's score, which has been discussed in section 5.3.1 *Significance of the Model*, my rough ranking on Kaggle.com is 3020th place. This denotes that there exists a limitation on my prediction model. In other words, there are other prediction and analysis skills that I have not realized and applied in my work. Also, there exists a space to allow me to improve the accuracy of my model by reducing 0.002 points.

In general, the accuracy of the forecasting accounts heavily on the historical internal data, the external factors and the techniques to handle those elements. The limits of discussed data manipulation techniques will be analysed firstly, then the limits of the whole project environment are investigated after that.

6.1 Limits of Applied Techniques

Independents Variables for Prediction Model

After realising there is no direct correlations between the log error and the individual 57 property variables, two solutions were adopted to examine the correlations between the log error and groups of the variables. The first one is the one discussed in section 5.3.2 with filtering the missing data, duplicated data and the non-numerical data to get the current 16 variables used in the linear model. The other solution is to use the machine learning packages in R such as "caret" and "randomForest" to classify the value of variables (R-Project, 2015). At that time, there were obstacles to the second solution. It caused the breakdown of RStudio installed on my laptop and the stop-working on computers at the university. When this problem was reported to the supervisor, the further work about this method was permitted to skip. Similarly, there might be other solutions to determine the most suitable variables in the model, however, because of the time limits, the further work cannot be proceeded.

Training Data and Test Data

The adopted methods to divide the training data and test data is based on the calendar month order. The predicted value is based on the previous data. Meanwhile, there is another method named cross-validation which can examine each data partition by rotating them into training set and testing set till finding the ideal partition of the training data and test data. However, this method will cause much workload based on the large scale of Zillow's dataset. After the communication with the supervisor, this way was skipped also. Otherwise, the MSE score might be better than 0.02455

Method to Predict the Log Error for October afterward

When predict the log value for November and December of 2017, the stand deviation was narrowed with a hypothesis that there would be bigger log error dispersion in the last quarter of 2017 as similar as the one in 2016. This hypothesis might be false as there is shortage of enough historical data to prove this point. Although the contribution of this method was identified in section 5.3.2 with temporary score on Kaggle.com, it requires waiting until the final score released in January of 2018 to confirm again the effectiveness of this model.

6.2 Limits within Project Environment

Shortage of internal data and external factors

First of all, the available historical transaction data covers the range of 01/2016 – 12/2016 and 01/2017 – 09/2017 solely. There is lack of enough data to indicate the direction of sales always decline and the log error often increase to opposite direction to appear the negative correlation that is found in section 5.2.2.

Secondly, external data is prohibited to be considered in this project. External factor might affect the actual real estate sales price. For example, seasonal festivals events like Halloween and Christmas in the end of every year might cool down the real estate market because most people are enjoying the holiday and travelling with families. If any buyer in a cool market might control the sales price easier than the buyer in a hot market. Therefore, such kind of external factor influence considerably the difference between the estimated price and the actual price. Meanwhile, Zestimate is a proprietary model. It is not able to access its techniques and models. In another word, project participants have to build a model without a baseline sample for advantages or weaknesses comparison.

VII. Conclusion

To sum up, with analysing the data about three counties of Los Angeles, Orange and Ventura, the delivered data analytics and the predictive modelling generally match the purpose of this project

to offer a mini thesis for data mining, data manipulation and data prediction. In the view of the dataset analysis, this task has been performed successfully with applying Scrum for project management and being conducted by a data analysis workflow with essential four phases for project execution. The delivered analysis thesis is evidenced by GitHub that it is able to provide a certain kind of guideline for those data scientists who feel interested in reducing Zestimate error margin. In terms of the possibility of improving the Zestimate model, the produced linear model is going to be verified because the final score and ranking will be released in January of 2018 after the sales tracking period. Significantly, the average mean squared error of the created model has already stayed around 0.02455 and it achieved enlightened score on Kaggle.com.

Within this report, the rational statistic data and persuasive visualisations are utilized. The values of Median, Mean, Max value, Min value and Standard Deviation were calculated separately for year 2016 and 2017 with finding that the accuracy of Zestimate of 2017 appears a bit worse than 2016 and the outliers of log error goes a bit far away of its median value while comparing that with 2016. Meanwhile, the negative correlation has been found between the log error and the transaction status. These two findings are used to support the hypothesis that there would be bigger log error dispersion in the last quarter of 2017 based on the data of 2016. Therefore, the stand deviation should be narrowed down while predict the log error for November and December of 2017 to improve the accuracy prediction. Similarly, the predictive model is built based on the effective data mining approaches and machine learning techniques: correlation matrix was used to explore the relationship between the log error and the 57 property variables; data filtering method was applied in identifying the valuable factors for the linear model; groups classification was applied while determine the ideal combination for training data and test data; mean squared error was used to evaluate the accuracy of the created multiple linear regression model and so on.

During the reflection phase of the project, however, a couple of obstacles and limits have been recognized based on the utilized data mining techniques and machine learning while creating the predictive model. These particular proceeds involve the strategies of determining the importance of independent variables for the linear model, classifying the training data and test data for model training purpose, and identifying the noisy factors for forecasting the log value for November and December of 2017. Although certain machine learning methods such as cross-validation have already been realized as the alternative technique to improve quality of the training data, other methods and or solutions still demand to be explored to find out the possibility to overcome the existing problems and improve the accuracy of the current multiple linear regression model.

Reference

Ballot Pedia (2016). Presidential election in California, 2016. Retrieved from https://ballotpedia.org/Presidential_election_in_California,_2016

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>

GitHub (n.d.). LinniQ/IFN701_Zillow. Retrieved from https://github.com/LinniQ/IFN701_Zillow/graphs/traffic

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>

Investopedia (n.d.). Negative Correlation. Retrieved from <http://www.investopedia.com/terms/n/negative-correlation.asp>

Kaggle Inc. (n.d.). Data fields description problem. Retrieved from <https://www.kaggle.com/c/zillow-prize-1/discussion/34168>

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 (2017). Measuring the error in calculating z. Retrieved from <https://cran.r-project.org/web/packages/zFactor/vignettes/statistics.html>

R-Project (2016). An Introduction to corrplot Package. Retrieved from <https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>

R-Project (2015). cPackage 'Metrics'. Retrieved from <https://cran.r-project.org/web/packages/Metrics/Metrics.pdf>

R-Project (2015). Package 'randomForest'. Retrieved from <https://cran.r-project.org/web/packages/randomForest/randomForest.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 How To (n.d.). Standard Deviation: Simple Definition, Step by Step Video. Retrieved from <http://www.statisticshowto.com/probability-and-statistics/standard-deviation/>

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

Zillow Inc. (2017). Data Coverage and Zestimate Accuracy. Retrieved from <https://www.zillow.com/howto/DataCoverageZestimateAccuracyCA.htm>

Zillow Inc. (2016). Zillow Boosts Accuracy with Update to Zestimate Algorithm. Retrieved from <http://zillow.mediaroom.com/2016-06-08-Zillow-Boosts-Accuracy-with-Update-to-Zestimate-Algorithm>

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

Appendix 1: Reflection on my learning

- 1. What were the things/activities you thought you did best in this project? What have you learnt about project management and research within professional practice?**

First of all, adopting Scrum as the approach of project management is one of the enlightened things I did in this project. The 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. All those enabled the project to be completed on time, even though the data of Zillow in 2017 was updated in the late period of the project.

Using GitHub as working progress repository and another communication platform between supervisor and me is another significantly right thing I did. Its stable and effectiveness made me feel free to manipulate the datasets with any trial techniques I found, without worrying the loss of the work. Its statics insight as shown in section 4.2 Burn Down Chart and Sprint Backlog help me to realize how perfectly the project was executed as scheduled.

Applying the data analysis workflow with clear purposes as the guideline to conduct the work of data analysis is one more hi-lighted thing I did. All the skills of data mining I learned from unit 509 were utilized in the project proficiently as explained in the outcomes section, including considering the data in the way of statistic, displaying the patterns with visualizations and identifying the suitable package and function to obtain and evaluate the expected results.

- 2. What were the things/activities you thought you did least well in this project?**

I have to admit that I have not tried my best to finish this project. Because of the time limited as a mother who takes care a 5-year old boy on my own, and the distraction of the other two advanced units this semester, I have not tried every possible technique I learned during the research to improve the model as effective as possible. This makes me expect the further research and analysis on this project as mentioned in the conclusion section.

- 3. Were there any specific problems or challenges you encountered? How did you handle them? What did you find as the hardest part of this project?**

There are two significant obstacles I encountered in this project. The first hardship is that there are only two main datasets from 2016 and 2017 can be used for training. One dataset is the transaction file and another one is the property features file. There is shortage of testing dataset. With

consulting to the supervisor, I divided the dataset into multiple groups to identify the proper training and test data. The second hardship is that there is no direct relationship between log error and the 57 physical property features. That means the formula cannot be determined with its most important variable apparently. After realising this difficulty, two solutions were adopted to examine the correlations between the log error and groups of the variables. One solution is to use the machine learning packages in R such as “caret” and “randomForest” to classify the value of variables. Unfortunately, this method caused the breakdown of RStudio installed on my laptop and the stop-working on computers at the university. After reporting this problem to the supervisor, I got the permission to skip this hardship. Another solution works well. It is the one discussed in section 5.3.2 with filtering the missing data, duplicated data and the non-numerical data to get the current 16 variables used in the linear model.

- 4. What was the most important thing you learned doing this project? What kind of opportunities or next steps you see based on your learnings doing this project? Which areas of your own professional knowledge and skills, or your personal attributes do you feel require further development?**

The most important thing I learned from this project is the techniques and experience to build a value predictive model. There are a couple of essential factors needs to be considered before and after the model building. Before building the formula, it needs to identify the proper type of model to fit the analysing dataset, then it needs to identify the important variables to build the formula. After the function is built, it needs to examine the training and testing data so as to the model can be trained and evaluated. Also, it requires to reflect the outputs and justify the model until it meets the purpose.

After finishing this project, I feel confident to apply a data scientist temporary job opportunity released by Queensland Rail last two weeks, even though it is not known yet if I am qualified for that currently. The confidence comes from I am able to understand my ability to manage a project of building a predictive model and to evaluate the data in the way of statistic, as well as knowing my capacity to manipulate the data and explore their hidden patterns with visualisations. This encourages me to consider a CEED project for my project next semester, because I believe more experience with dealing the real industrial data will enhance my professional knowledge and skills in data mining.

Appendix 2: Project Log Sheets

(Serial Number)

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name:Linni Qin..... Date: ...18/08/2017.....Meeting No: 1...

Project title: IT MASTERS (DS): Perform a Data Science Analysis of a Dataset/Task UNIT:IFN701.....

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Guido Zuccon Supervisor's Signature: *Guido Zuccon*

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

N/A ...
The first formal meeting.

Items for discussion (noted by student before mandatory supervisory meeting):

Project Plan
Project problem
Data Analysis Methodology

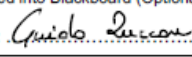
Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Finish the project plan
2. Import the raw data
3. Explore the data and get some insights about the data

Project Log Sheet – Supervisory Session
--

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name:Linni Qin.....	Date: ...08/09/2017..... Meeting No: 2...
Project title: IT MASTERS (DS): Perform a Data Science Analysis of a Dataset/Task UNIT:IFN701.....	
<input type="checkbox"/> Journal entry logged into Blackboard (Optional)	
Supervisor's Name: Guido Zucco	Supervisor's Signature: 
Update on progress since last meeting, and challenges faced if any (noted by student <u>before</u> mandatory supervisory meeting): Update: Presentation and the project plan have been finished. Challenge: What type of prediction model will suit this project?	
Items for discussion (noted by student <u>before</u> mandatory supervisory meeting): <ol style="list-style-type: none"> 1. Zillow's log error distribution in 2016 2. The necessary to analyze the data by different counties (LA, OR, VE) 3. Zillow's transaction monthly pattern in 2016 4. The negative correlation between the log error and transaction frequency in 2016 5. Binary Prediction models: Decision Tree, Logistic Regression, Support Vector Machines 	
Action List (to be attempted or completed by student by the <u>next</u> mandatory supervisory meeting): <ol style="list-style-type: none"> 1. Merge the transaction data and the property data 2. Missing data 3. Apply the binary prediction models 4. Randomly choose the training data from different data groups: LA, OR,VE, different months 5. Explore the correlation between log error and other feature 	

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

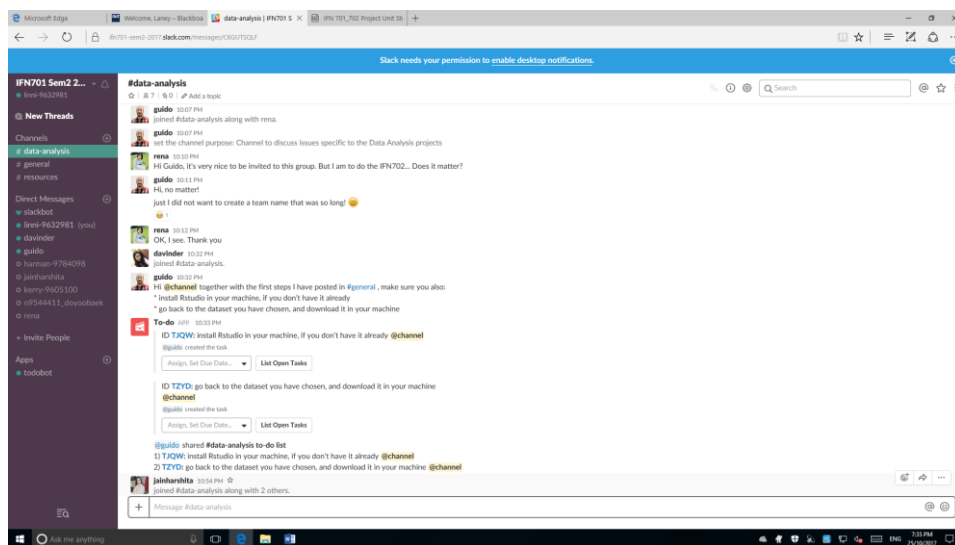
1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name:Linni Qin.....		Date: ...29/09/2017.....		Meeting No: 3...	
Project title: IT MASTERS (DS): Perform a Data Science Analysis of a Dataset/Task UNIT:IFN701.....					
<input type="checkbox"/> Journal entry logged into Blackboard (Optional)					
Supervisor's Name: Guido Zuccon			Supervisor's Signature: <i>Guido Zuccon</i>		
Update on progress since last meeting, and challenges faced if any (noted by student <u>before</u> mandatory supervisory meeting): <ol style="list-style-type: none"> 1. Transaction data and the property data have been merged 2. Missing data status have been identified 3. Correlation between log error and each property feature has been explored 4. Binary Prediction models were applied. <p>Challenge:</p> <ol style="list-style-type: none"> 1. No correlation exists between the log error and the 57 property features 2. Binary prediction models do not fit Zillow's dataset 					
Items for discussion (noted by student <u>before</u> mandatory supervisory meeting): <ol style="list-style-type: none"> 1. Will the linear regression suit this project? 2. The way to identify the proper combination of training data and test data 3. The introduction about the cross-validation 4. Mean Squared Error is recommended to evaluate the prediction model 					
Action List (to be attempted or completed by student by the <u>next</u> mandatory supervisory meeting): <ol style="list-style-type: none"> 1. Use different groups of training data and test data to identify the best combination dataset for model training. For example, use the first three-week data to predict the last week value, use three-months data to predict the value for a future week. 2. Apply Mean Squared Error to evaluate the performance of the training model 3. Find other R techniques to identify the correlation between the log error and the variable by groups 					

As displayed at the above project log sheets, there are only three formal meetings with Dr. Guido Zuccon who is my supervisor of unit 701 this semester. The less chance of meeting is caused by his busy conference schedules and his sick leaves. The following nine important Slack chatting screenshots are used to prove the consistent communications with the supervisor centring on the project.

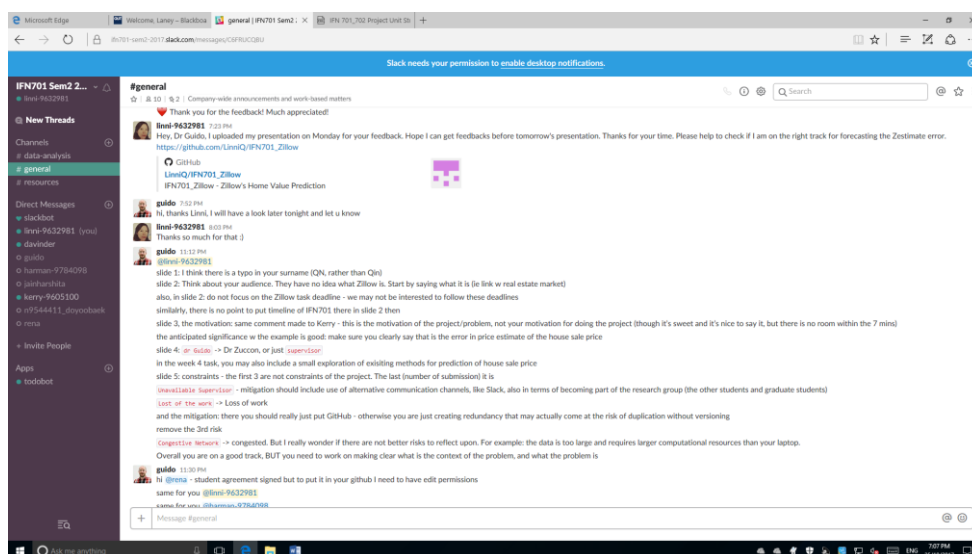
July 31, 2017 Project start-up “To do” List:

1. Prepared the R and RStudio on the laptop
2. Got the data ready
3. Got a GitHub account to save the works for supervisor’s convenience to provide the feedback
4. Introduced individual project in the team



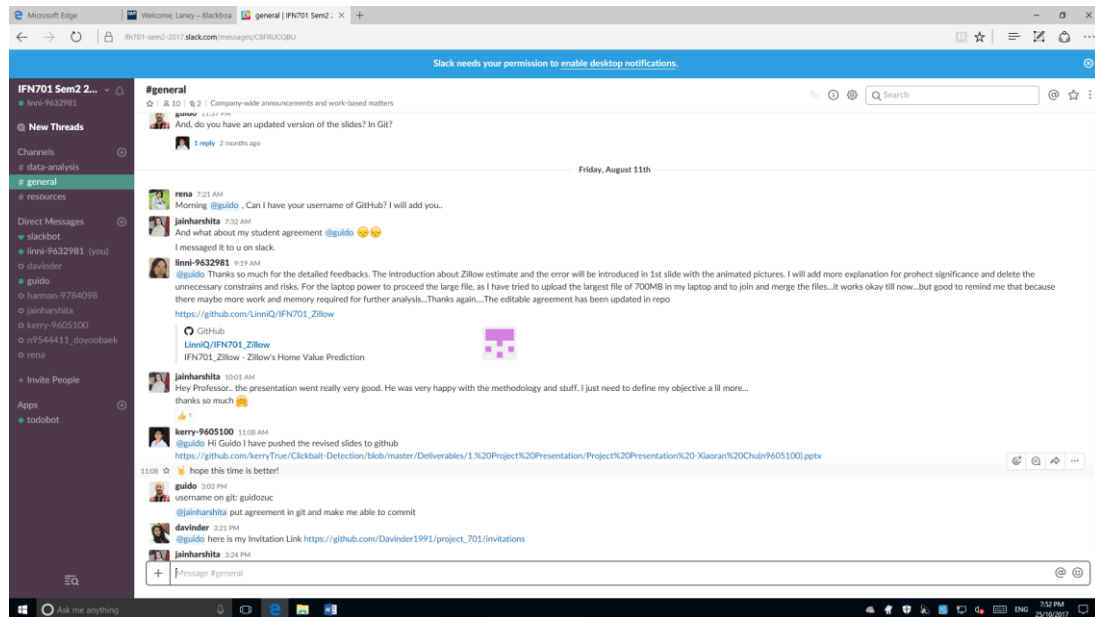
August 10, 2017 Proposal Presentation Discussion

1. Corrected the typo
2. Re-edited the motivation of the project



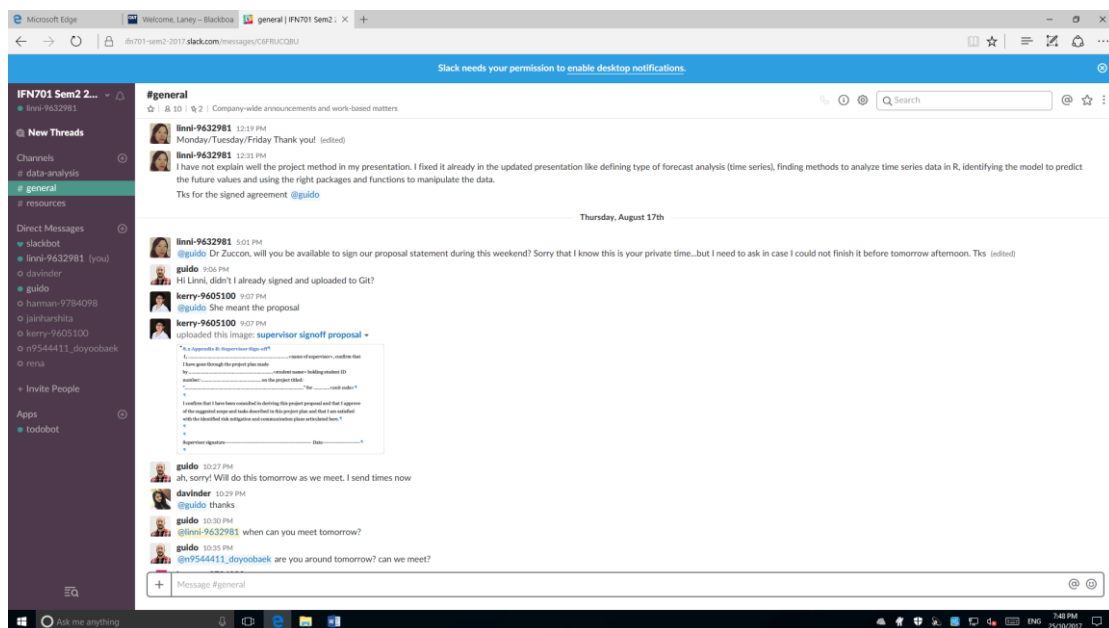
August 11, 2017 Discussion on Project Agreement

1. Update the agreement on GitHub for supervisor's signature



August 17, 2017 Discussion on project plan

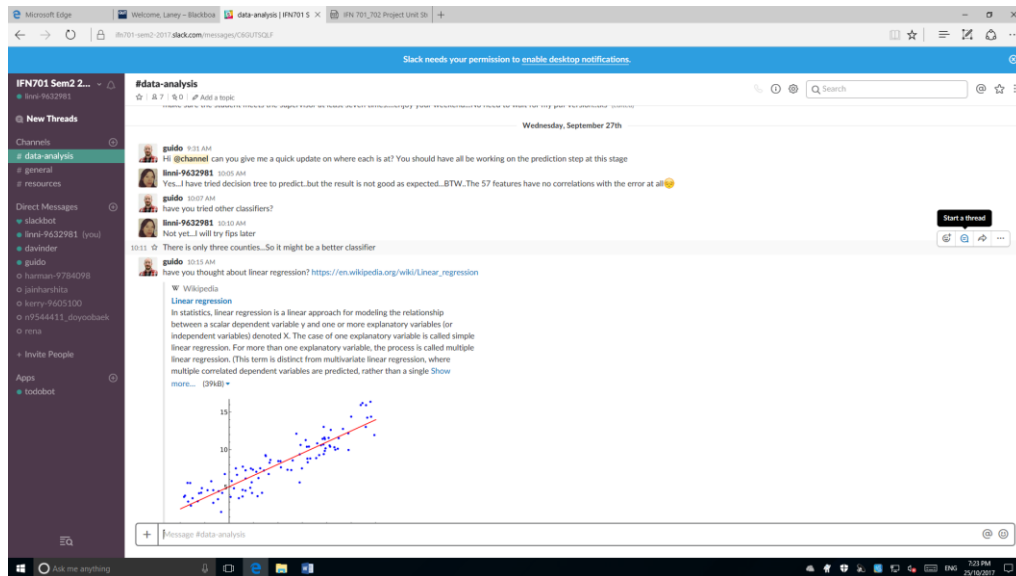
1. Got the project plan draft ready for supervisor's feedbacks



September 27, 2017

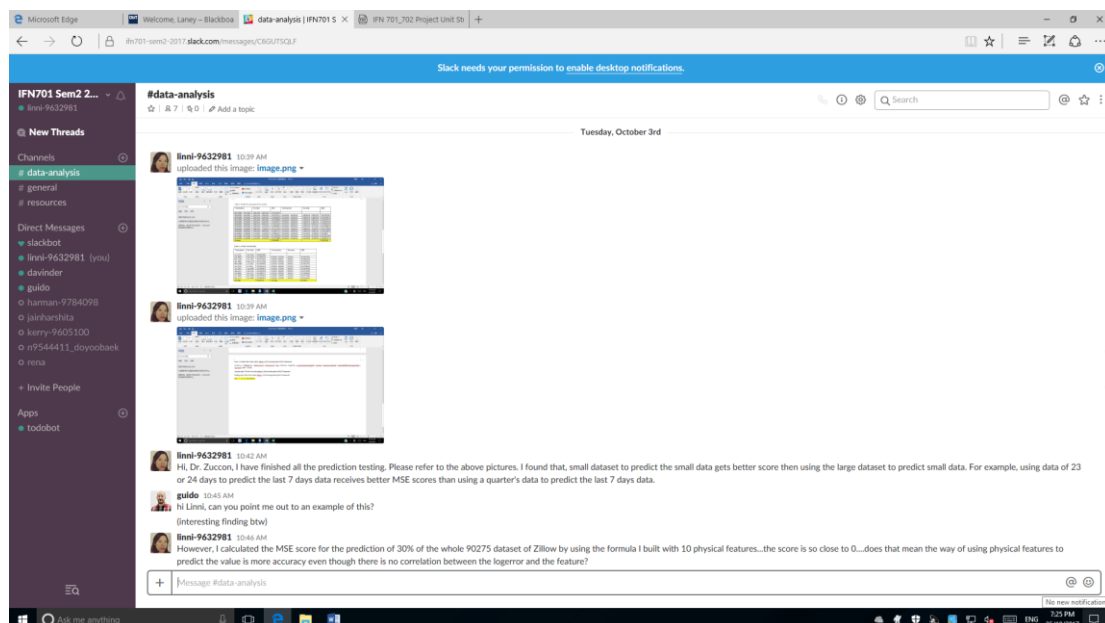
Discussion on the prediction model

1. Binary prediction models have been proved not suitable for this project
2. Linear Regression is better for this project
3. Multiple linear regression model is the proper model for this project



October 3, 2017

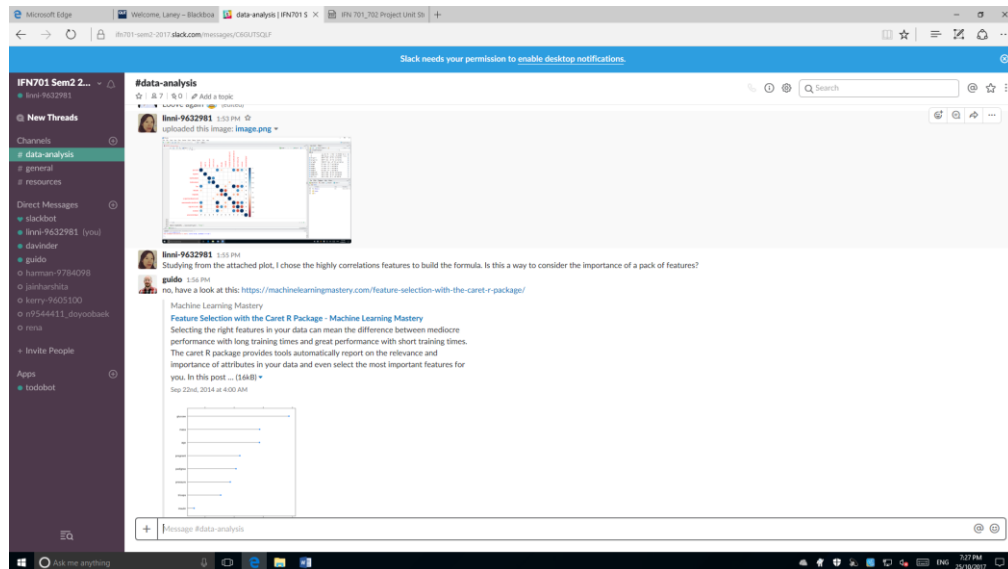
Update the Mean Squared Error Value based on the testing on different groups of training and test datasets.



October 5, 2017

Discussion on the challenge of zero correlation between the log error and the property features.

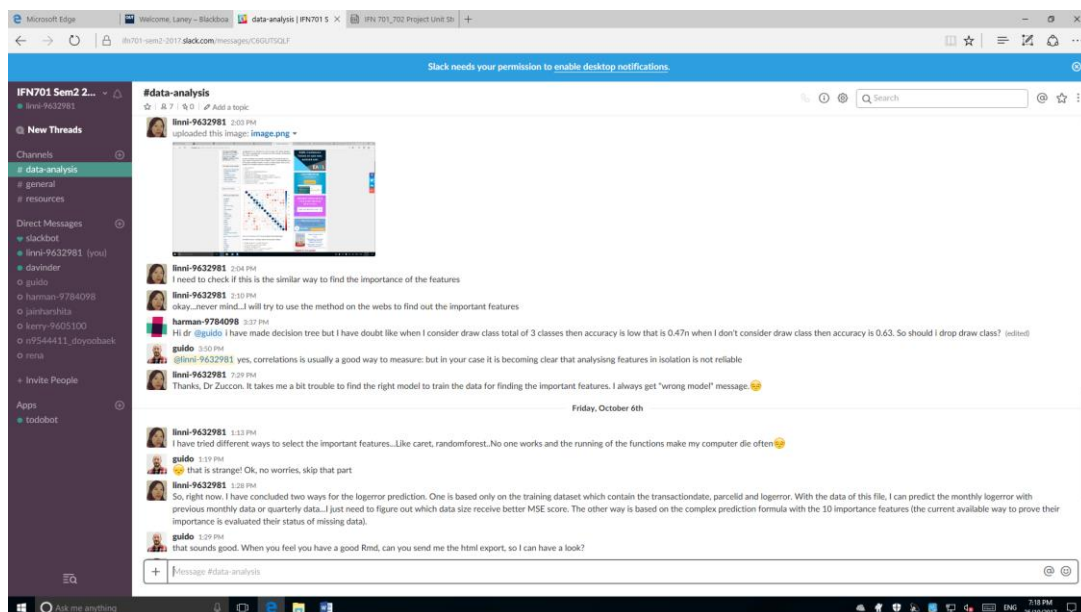
Discussion on the obstacle of applying the package “caret” and “randomForest” to identify the importance of the variables.



October 6, 2017

Updated the results based the discussion on October 5.

Supervisor approved to skip some hardships.



October 7 – October 13, 2017

As the new data of Zillow in 2017 was updated on October 3, I needed to do more work on exploring the data of 2017 as similar as what I did for 2016. Then, the findings were combined together to build the prediction model.

Updated the new findings to supervisor.

Two deliverables were committed as scheduled for supervisor's feedbacks, including a Rmarkdown data analysis report and a multiple linear regression prediction model.

