Statistics for Business Assignment 1

Understanding and Predicting the Sale Price of Houses in Ames

Byron Graham

12345678

word count (Excluding tables, figures, references and appendices): maximum 2000 + 10%

Contents

[1.0 Introduction and Background (n.b. this can be split into two sections if you prefer). 2](#_Toc117583810)

[2.0 Methodology 2](#_Toc117583811)

[3.0 Results and Discussion (if you prefer, the results and discussion can be two separate sections). 3](#_Toc117583812)

[4.0 Conclusions 3](#_Toc117583813)

[5.0 Reflective Commentary 3](#_Toc117583814)

[6.0 References 4](#_Toc117583815)

[Appendix 1: R Code Used 4](#_Toc117583816)

This document provides some guidance on the structure and content of the statistics for business assignment 1. The structure of the assignment follows quite closely the structure of a typical quantitative academic paper, so if you are unsure about what to include in different sections it is worth reading through a few academic papers, in addition to the guidance below. You can also include any other information you feel is relevant, and can deviate from the suggested structure – the key thing is that you meet the requirements of the assignment task that are specified in the module handbook. The module handbook provides very detailed steps on the assignment tasks that you will need to undertake, as well as the content of the report, and should be read alongside the additional guidance below.

The following provides some guidance on the overall structure (section headings), as well as some guidance on the content of each section.

# Introduction and Background (n.b. this can be split into two sections if you prefer).

The introduction should present a general overview of the topic and area – in particular a brief introduction to the topic, the general problem or question, and a brief description of what you are going to do.

Once you have introduced the topic you should review the relevant literature – in this section you can discuss past research that has focused on understanding the prices of houses as well as on papers that have taken a machine learning approach to predict house prices. You can focus on areas such as, what theories did they use, what methods were used, how accurate were the models, and what were the key findings. You will need to synthesise this well and focus on the most important areas to stay within the overall word count.

This should be followed (and ideally linked to) the five hypotheses (i.e. the relationships that you expect to find). To formulate the hypotheses, consider the factors that you believe will be most important in predicting the sale price of the houses – this can be based on a combination of your own logic and intuition and on the past literature.

Hypotheses can be worded as follows:

H1: There will be a positive/negative relationship between variable X and sale price.

It is important to explain why you expect the relationship to exist, drawing on the wider literature and logical arguments.

# Methodology

The methodology section should describe the tasks you undertook to analyse the data. It is also important to explain why you undertook certain tasks. You can include references to the methodological literature in this section (or to textbooks).

This section should also include information on the data quality issues that were identified and how you addressed these.

In this section, you can also include some general commentary on the information contained in the overall dataset, the techniques that were used to produce the visualisations, the techniques used for the measures of association, and techniques for the regression models. You can also talk through how you assess the model accuracy and any assumptions that were checked. N.B. the focus here is on what you did and why – the results of this process are presented in the following results and discussion section.

# Results and Discussion (if you prefer, the results and discussion can be two separate sections).

The results section should include the key outputs from the analysis (any non key outputs can be included in the appendix and referred to in the main text). The discussion section then focuses on the interpretation of the results, and in particular how the results relate to the wider literature. You can also include explanations for interesting findings from the results.

The results section should include the key output from the analysis:

* A table of descriptive statistics
* The final visualisations
* The measures of association
* The (final) regression model and accuracy measures
* Any other results you feel are useful.

In the written part of the results section you should talk through these results – essentially describing the content

The results of the regression model are crucial to this section and should take up the majority of the results section. You should build at least one regression model, but you can build more. A common approach is to start with a ‘baseline’ model, and then build more to see if you can enhance the model by including more variables. For this assignment it’s probably best to use less than 15 or 20 variables at a maximum. You can focus on talking through the most accurate regression model although others can be included in the appendix). When talking through the regression model you should talk through each variable in turn mentioning whether it is statistically significant, as well as the size of the coefficient (the effect size). You should pay particular attention to the variables that are used to test your hypotheses. N.B. the hypotheses are tested using the regression model, rather than the measures of association.

In the discussion section you should refer back to the wider literature – how do your findings compare to those of other research published in the academic literature? Are there any theories that you found that could be used to explain the results, or other explanations for why variables are/are not significant (/important) in the models?

# Conclusions

In the conclusion section, you should summarise the main findings. Here you should also talk about the contributions of the project – what were the most interesting findings? You should also talk about how the results can be used in practice – for example, how can businesses use the model in practice to more accurately value property, and what would be the benefits of this?

In this section, you should also talk about any limitations of the project, or any areas that could be improved. For example, how could you make the model more accurate?

# Reflective Commentary

This is a personal reflection about how you feel you are progressing with statistics and R programming. How have you progressed in statistics and R programming during the module? Are there any areas you find challenging – if so how will you address these? Are there any areas you find particularly interesting? How will you apply what you have learned in this module in your career?

# References

The usual recommendation is around 10 references per 1000 words, so you may have around 20 references for this assignment.

# Appendix 1: R Code Used

Copy and paste the R code into an appendix.

Code should include comments that allow the reader to understand what the code is doing.

It is important to reference any code that you have used from elsewhere – it is expected for the assignments that you write your own code. If you have used functions from elsewhere (e.g. if you find a piece of code online to accomplish a particular task then this should be referenced). Standard functions do not need to be referenced (e.g. you do not need to add references for individual packages such as ggplot2, psych etc. as these are standard approaches). It is understood that there are standard approaches for some tasks such as setting the working directory, reading data, and loading packages which will be highlighted by turnitin.