MA26620: Practical 4

Housing Data and Assignment One

1 Housing data

1.1 The Data

The dataset 'houses', on the module webpages (Practical worksheets), contains a subset of data concerning sold houses collected by estate agents in the town of Douglasville, Ohio, USA. Below is a description of fields in the dataset:

home	A unique reference number for each property assigned by the estate agent
nbhd	Notes in which of Douglasville's three main neighbourhoods the house resides. *Key: N=Normsville, R=Realtown, S=Spectralia*
offers	Number of offers that were received on the properties before it sold
sqft	Floor area of the property (measured in square feet)
brick	Is the house constructed of brick? (Yes/No)
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
price	Final selling price ($)
ptype	Type of property. *Key: Semi=Semi-detached, Bung=Bungalow, Det=Detached, Apt=Apartment*

1.2 Error checking

Begin by loading the data into a dataset called `houses` and attach it using the usual method (be sure to indicate that the first row of data gives column headings). Data input errors are fairly common in real world data, with the likelihood of errors increasing with datasets size. It's therefore a good idea when given new data to inspect the data carefully and look for any anomalous values. A good approach for beginning this task is to use the `summary(houses)` command to see some summary statistics. For instance, the field '`brick`' has a meaningless value of 'Yed'. This is close enough to 'Yes' for us to be fairly confident that this was just a typo in data entry (although it's good practice to note that such corrections have been made in any report on the data). See if you can find another error – there should be one more.

So how do we fix these errors in the data? It's possible but surprisingly long-winded to fix it in RStudio itself by running various commands, or we could instead open the csv file in some spreadsheet software (e.g. MS Excel) or a text editor (e.g. Notepad) and edit it there, then save as a csv. You might also want to change codes (N, R, S, Semi, Bung etc.) with their meanings as this will save you having to do lots of plot-editing later, as good plots don't have codes as axis labels or in legends).

1.2.1 Using Excel

Open Excel, then within Excel open `houses.csv`. Find and fix the two input errors, then Save As 'housescorrected.csv', being sure to have chosen 'CSV (Comma delimited)' in the 'Save as type' box. Now come back to RStudio, detach `houses` (you may even want to remove it with `rm(houses)`), import `housescorrected` and attach that. Run a `summary` command of your new dataset to check that everything is sensible now.

1.2.2 Save it!

Once it's been corrected, save your workspace (Session > Save Workspace As...) somewhere on your `M:` drive as this will enable you to reopen the corrected data later when you're working on the assignment.

1.3 A sensible way of working

Until now, we've mainly been typing and running commands in the lower-left area of the screen called the Console. This is fine if you're only going to be running a few commands, but as you'll be delving into this data quite deeply, let's use an R Script file.

Click File > New File > R Script. The upper left panel becomes a notepad where you can type commands. So in today's practical and while you're working on the assignment, type your commands in *here* rather than in the Console pane. When you've typed a command, clicking the 'Run' button will run everything that's on the currently selected line. Equivalently, you can press Ctrl+Enter to run the command on the currently selected line. If you want to run more than one command at once, highlight several commands and then click 'Run' or press Ctrl+Enter.

When you make a typo or your command isn't quite perfect, edit it in the notepad and then run it again.

So what's the advantage of working using this notepad rather than just typing things into the Console window? Well, you can save your notepad at the end of a session (just click the Save icon in the notepad pane - it's the one immediately above the first line of your notepad). When you do, be sure to save it to your `M:` drive, as stuff saved there will be available on any university computer. Then, the next time you work on your assignment, all you'll need to do is load the data (if it isn't already there), enable ggplot in the Packages tab (a step which isn't even necessary if you put `library(ggplot2)` at the start of your notepad), load the notepad that's full of your lovely, polished commands, highlight everything and click run, and R will remake every subset and produce every plot that you've made.

1.4 Tables and barcharts

Let's begin to analyse the data by using the tables and barcharts methods that we have learned in previous practicals. To begin, we should get an overview of our dataset. Make a table that shows how many houses in the dataset have 2, 3, 4 or 5 bedrooms.

Transfer your table into Word and format it in a visually appealing way as just copying and pasting the text with no modification does not look very professional.

Do different neighbourhoods have a different proportion of 2, 3, 4, 5 bedroom houses? Make a proportional table to answer this and consequently make a stacked barchart to display this graphically. (consult your notes from previous practicals if you can't remember how, or ask for help in the practical). As a reminder: in order to make a barplot in R, you must first make a table (a proportional table in this case); the values in the columns determine the height of the bars.

Label and comment on your barplot. Are the neighbourhoods similar or different? How?

Clearly many other tables could be produced by looking at different combinations of variables. In your assignment you will be given credit for conducting your own investigations, so you are encouraged to experiment with delving into the data to see what you can find!

1.5 Boxplots

Let's use the `ggplot2` package that we used in the last practical to create a boxplot of house prices with separate boxes for each neighbourhood. (Note: you will first have to tick "ggplot2" in the "Packages" tab. If it doesn't appear in the Packages list, repeat the installation instructions from the previous practical).

To get you started, a basic boxplot can be generated using the command:

```
ggplot(houses2,aes(x=factor(nbhd),y=price))+geom_boxplot()
```

Again, copy into Word after you've edited titles, axis labels etc. and comment on what you observe. Maybe altering box widths to represent sample sizes would be a good idea? If you've forgotten how to do any of these things, consult earlier practical notes or search the web for ggplot2 help – there's lots out there. In fact searching the web might show you how to make even fancier plots than we've met in the practicals – feel free to use any methods/commands you find so long as you think they're helping you explain features of the data.

How about a few other boxplots to show how house price is distributed among the other categorical variables (not just neighbourhood)? Remember the `factor` command if you encounter one big fat boxplot when you were expecting several.

1.6 Scatterplots

Make a plot of house price vs floor area. How strong is the correlation? Is the relationship linear? Using the colouring techniques we met last week, investigate whether neighbourhood has an effect on price. How about some of the other categorical variables? Copy any insightful plots into Word and comment about what they tell you.

According to the linear regression model, how much does each additional square foot add to the price? Does this change much if you consider each neighbourhood separately? Is there anything surprising about these results and if so can you explain it?

1.7 Further investigations

You're now free to investigate the data further in whatever way you wish. The plots we've made so far are a good start, but you should now investigate some other aspects of the data of your choosing. The `subset` command may be useful for investigating subsets of the data individually.

2 MA26620 Assignment 1

Your first assessed assignment (worth 20% of the module mark) consists of two parts. For the first part of the assignment (counting for 20% of the assignment's marks), you should hand in your solution (typed up) to Question P3.Q1 from Practical 3 on Regression.

For the second, larger, part of the assignment (making up the other 80% of the marks), you should write a report, in Word or similar, on what you have learned about the Douglasville houses data based on a selection of the material you have gathered in this practical and any further investigations into the data that you conduct. The report should be written for a reader who is a competent statistician but is unfamiliar with the dataset (so for instance, you should refer to house types as detached rather than Det).

The assignment must be handed in via Blackboard under the Assignments content item as a single file (Word or PDF).

Your Report must include:

(i) a title for the report;

(ii) an introductory paragraph explaining briefly the nature of the investigation, where the data came from i.e. the nature of the data and the aims of your analysis;

(iii) the results of your R analysis, in a sensible order (quite probably different from the order you did them in the practicals). Explain the purpose of each element and give your comments; use sub-headings where it helps to make the structure of your report clearer.

(iv) a conclusion where you summarise briefly what you have learned and any questions that remain unanswered.

So if you have not done so already, give the report a title and write an introductory paragraph.

You should include *at least* one of each of the following: table, barplot, boxplot, scatterplot. Marks will be awarded for choice, presentation, editing, and commenting on each of these components, as well as the general structure and flow of your project. You must also give some thought to the order and structure of the content. Consider whether you have used the same style (e.g. font, type size) when you typed the project, that you have organised the text into suitable paragraphs and that the page breaks occur in natural places.

Consider also whether the plots or statistics you have included do indeed illustrate the points you wish to make or whether there are any others which would do the job better. There may be other questions you wish to follow up about this data and some credit (20% of the overall mark for Assignment One) will be given for any relevant additional investigations. Make sure that all plots are well labelled and well presented.

Please note that unnecessary or uninformative graphics will receive no credit and, if excessive, may be penalised. Be selective and always have a reason for your choices. There is no word (or page) limit but in the past, most assignments have been fewer than 10 pages including plots.

Assignments must be handed in via Blackboard by **5pm on Thursday 28th November**, giving you just less than two weeks to complete the assignment.

2.1 Notes on policies

You should also be aware of the university's Unfair Academic Practice (plagiarism) policy. Assignments will be handed in via TurnItIn on Blackboard which very effectively makes it clear if your work is not substantively your own. Plagiarism is one of the easiest ways to be excluded from the university or at the very least score zero on the assignment or the whole module. Moreover and perhaps most importantly, it is unfair to your fellow students, so don't do it. Both the copied and copier will be awarded zero marks, so don't send your work to others.

You should also be aware of the university's extension request policy. Any extension requests must have supporting evidence and be made to your year tutor (Dr Adil Mughal [aqm]) at least three working days prior to the deadline.

Good luck!