STA302/1001: Assignment 2

Craig Burkett

Nov 1, 2015

Due at the beginning of lecture on Thursday, Nov 12^{th}*. Please hand it in on 8.5 x 11 inch paper, stapled in the upper left, with no other packaging and no title page. Please try to make this assignment look like something you might hand in to your boss at a job. In particular, it is inappropriate to hand in pages of R output without explanation or interpretation. Quote relevant numbers from your R output as part of your solutions. The only direct R output you should submit with the assignment are relevant plots.* ***You must append your R program file to the end of the assignment, formatted nicely with a fixed-width font.*** *No assignment will be marked without a program file, and marks will be deducted if the instructions above are not followed.*

In this assignment, you will examine some more datasets from the Toronto Open Data Portal, containing statistical information on building permits, as well as some grades from a 'hypothetical' class very similar to ours. Any time that I use the words {Present, State, Give, Show, Predict, Display}, you must supply that plot/table/output/prediction in your submission. If I say {Produce, Make}, you do not need to show what you produced or made, but you still need to do it.

The data dictionary and a link to the datasets are available on Portal. To answer most of these questions you can use the sample code from lecture, but you will also have to search for some functions online. This is how 99%* of all useRs learn R. You can always use the forum on Portal if you get frustrated, so start early! Benchmarks of 95% confidence and 5% significance can be used unless otherwise specified.

A Building Permits (30 marks)

We will consider some demographical information on building permits in Toronto. Please try to contain your excitement.

1. Let's format this new building data. Unless otherwise specified, you can ignore all missing data here (ie. just imagine as though there weren't any and code that way)

   (a) Read in the building data CSV file, and format the application and issued dates as R Date objects (as in the last assignment).

   (b) Make new factors called *Year* and *Month* for the application dates only, to match the format from Assignment 1. These should be new columns in your data frame, of the same length as the raw data.

   (c) Let's see how stagnant the municipal bureaucracy can be. Make a new numeric variable that measures the number of days between the application date and the issue date.

   (d) Looks like some odd data here. Subset the data by records that have a non-negative number of days between application and issue. Keep only the columns:

   `Year, Month, Number of Days, STATUS, POSTAL, CURRENT_USE, PROPOSED_USE, DWELLING_UNITS_CREATED, DWELLING_UNITS_LOST`

2. Let's explore this new building data.

   (a) How many Single Family Dwelling (SFD) homes are there in the dataset? How many office buildings? We'd like some bar charts or something showing the city's makeup, but we should probably look at the strings before doing this naively. Please follow these steps to clean up the **current usage** column only:

      i. Look at a one-way table of current usage, ordered by number of records, for usage classifications with at least 30 records. Do not print this table for your submission. To be clear, you are not subsetting your data at this point, just the table that shows the usage classifications (otherwise it would be quite long).

ii. What a mess. Look at all of those codes that are basically the same thing! You could go through one at a time, changing text to clean things up. Or try to do it programmatically with regular expressions. It's up to you, really.

iii. First, make the strings all uppercase, just in case.
Hint: `toupper()`

iv. Convert anything that looks like an apartment (or apt.) building or unit, or condo, into a common code "Apartment". Go through the list (of codes that occur at least 30 times - use the table you already looked at including spelling mistakes) for all of these parts.

v. Convert anything that looks like an educational institution into "Educational".

vi. Convert anything that looks like a restaurant (but not grocery store) into "Restaurant".

vii. Convert anything that looks like a place where you'd park a car into "Parking Lot". Include 'garage' here, but not 'repair garage' nor 'car dealership'.

viii. Convert anything with the word 'retail' in it (including 'ret') to 'Retail', even if vacant. Those places never stay vacant for long!

ix. Convert anything with the word 'office' in it (including 'off') to 'Office', even if vacant.

x. Convert anything that looks like mixed use or multi use or multiple use to the code 'Mixed'.

xi. Roll 'Church' into 'Place of Worship', 'Lab' into 'Laboratory', 'res' into 'Residential', 'ind' into 'Industrial', and 'subway station' or 'union station' into 'Transit Station'.

xii. Convert anything that looks vacant into 'Vacant', whether it's residential, commercial or industrial.

xiii. Convert 'N/A' and 'not known' to a blank string "".

xiv. Now the fun part. We'd like to combine all the SFD Detached homes into one code. Convert all permutations of SFD Detached into one code. Leave 'SFD' as a separate code.
You can assume that all detached and semi-detached dwellings are SFDs (even if they don't say that) from here on

xv. Do the same thing with SFD semi-detached homes.

xvi. Do the same thing with SFD rowhouses/townhouses (treat them as the same thing).

xvii. Clean up anything else that is a SFD but with no other specification as 'SFD'.

xviii. Almost there! Now let's deal with all of those multi-unit dwellings that aren't apartments or condo. Put them all under a code called 'Multi-Unit'. This should include duplexes and triplexes as well.

xix. Let's combine any sort of plant, lumber yard, or manufacturing lot into our existing 'Industrial' code.

xx. Nursing homes, group homes, long term care facilities, homes for the aged; those are all really just 'Nursing Homes'. *One might argue that bowling alleys are as well, but we'll leave those out.*

xxi. Let's classify all bowling alleys, parks, arenas, stadiums, theaters, clubs and fitness centers with 'Recreational'.

xxii. That's pretty good; we got most lot types classified. The city probably would have paid you for an entire Summer to do that. Present a table showing the counts of your remaining Current Use types (for counts of at least 30 only!), displayed as a vertical list in descending order. It should fit on one page. I have 41 items in my list, including the blank. Your list should have between 35-45 items, and this is the only thing you need to display for this question. It looks something like this:

```
          Usage Count
                37454
            SFD 37105
SFD Detached 19577
            ...
```

(10 mks)

(b) Produce a histogram of the number of days it takes to have a permit issued.

(c) That's not very helpful. It looks like the number of days is highly right-skewed. Add one day to the number of days and take the

natural log, and present the new histogram of the log-transformed data. Set an informative title and label the axes appropriately. (2 mks)
Voilà! Zero-inflated Normal. Use this log-transformed delay time for the next two plots.

(d) Maybe the delay in issuing depends on whether any residential units are gained or lost? Present a grouped histogram (well, three histograms on one plot) showing this case. Do not plot the cases with missing data. (2 mks)

(e) Looks like it takes a bit longer when residential status is changing, but maybe side-by-side boxplots would better help us quantify this difference. Show these, for the same question above. (2 mks)

(f) The city wants to know which permit "Status" levels have the longest delays between permit application and issue. Show a table listing the top 10 Status levels, ranked by decreasing (original, untransformed) average delay in permit issue (rounded to nearest day). (2 mks)

3. Let's see what's going on in (y)our neighbourhood! Present a **word-cloud** of the (cleaned) current uses for the permits in your postal code. If you don't live in Toronto or if your postal code doesn't show anything interesting, just pick a postal code at random from the city.
Hint: Use the {`wordcloud`} *package - this part should only take 3 lines of R code.* (5 mks)

4. Join the new building data to the license data (marriage and business) by Year and Month. The new column you are adding to the previous dataset is the number of building permits **applied for** in each month. You should have records only up to August, 2015.

   (a) Present a scatterplot of the Business Licenses vs. Building Permits, fit a linear regression model to the same data, and plot the regression line on the scatterplot. (3 mks)

   (b) Show both the residual plot and the Normal QQ plot in a 1 x 2 grid (ie. plot both of them on the same 'line' of your report, side by side). (2 mks)

(c) Do you see any problems with these plots? If so, state what the problem is (but don't fix it). If not, state why you think the plots are OK. (2 mks)

B Assignment Submission **(15 marks)**

Consider the timing data from Assignment #1. We seek to predict a student's assignment grade using their submission time.

1. Present a histogram of submission times (in number of hours before deadline - that is what is in the file) and a histogram of A1 grades, in a 2x1 grid. Do you see any potential problems with high leverage points? (2 mks)

2. Present a plot of assignment grade vs. submission time and fit an SLR model, showing the fitted line on the plot. (1 mk)

3. Show the residual and Normal QQ plots in a 2x1 grid. Do you see any problems? (2 mks)

4. Take the natural logarithm of submission time (add 1 hour to everyone's time to avoid the zero problems) and take the square of A1 grade, and redo the scatterplot and regression model with these transformed variables. Show the new scatterplot with the fitted regression line, and the new residual plots. (3 mks)

5. Give an interpretation of the fitted slope estimate in a plain English sentence. Your answer is something about how the submission time is related to assignment grade. (2 mks)

6. Predict the assignment grade for students who submitted their assignment 12 hours early, and give an appropriate interval around this value. (4 mks)

7. Why do you think the residual plot looks the way it does? Do you feel comfortable using this regression model? Why or why not? (1 mk)

C Format **(5 marks)**

Please make your submission look nice. This means:

- Proper sentences free of typing errors
- Graphs and tables should be in the **body of the report**, not thrown in at the end
- R Code should be appended to the assignment, in a small fixed-width font like `courier new`

  Missing any of these items will forfeit all of the format marks.