STAC32 data analysis project

November 7, 2019

For this project, you may work alone or in a group, chosen by you, of up to 6 students. If you work as a group, all members of the group receive the same mark, so it is up to you to ensure that all members of your group make an equal contribution.

The project is due on the last day of classes, Monday December 2, 2019, at 11:59pm. Hand an electronic version of your report in on Quercus, either HTML or Word or PDF format. I will create an "assignment" called "Data Analysis Project". Remind me if I forget.

Your report should begin with a cover page clearly showing the names and student numbers of all the students in your group. If you work as a group, only *one* of you should hand in the project (it doesn't matter which one of you: I will make sure that you all get credit).

It is *not* possible to extend the due date. The last day of classes is the University-wide due date for term work like this.

Choose *one* of the numbered data sets described below, or use a data set of your own choosing. (If you use your own data, it is a good idea to discuss your plans with me before you start, to ensure that you are doing something suitable for this course.) Your job is write a report describing a statistical analysis of the data set you choose. Your report should follow the lines described in the lecture slides (in the section with the example academic paper), or the mini-report that you did for Assignment 6: basically Introduction, Analysis and Conclusion with the Analysis subdivided if it has several steps. You should consider exploratory analysis (graphs, calculations of summary statistics), checking of assumptions and possible transformations as appropriate. Your aim is to produce a complete analysis that addresses what you see as the most important research question or questions posed. Your report needs to stand alone, so that it can be read by itself (without the data set description below). You might, therefore, need to repeat in your Introduction some of what is in the data description. Assume that your reader is an "intelligent layperson" who is interested in what you have to say, but who won't understand the subject matter unless you explain what is going on.

In marking your project, I will take into consideration the difficulty of the analysis that you attempt. A simple analysis with few errors may be worth less than a more ambitious analysis with more errors.

You may use R or SAS or a combination of both, as you deem appropriate.

If you have questions about any of these data sets, you should treat me as the "subject matter expert". You should note, in your report, that you sought clarification from me, as well as noting any other assumptions that you made about the data. If you have questions about the *analysis*, it is up to you to figure them out.

I will give your report an overall grade out of 10, on this scale:

Mark	Meaning
9	A truly impressive report, going well beyond what I would expect in this course.
8	Work that illustrates an excellent command of the material in the course, as it applies to your chosen data set.
7	Work that is about what I would expect in this course, with a few small errors or some lack of clarity in thinking.
6	Work that is below the standard I would expect, with clear errors in the analysis and/or a clear lack of understanding.
5 or less	Work that falls far below the level required, being hard to understand and/or having serious problems in the analysis or explanation.

I don't give 10 out of 10, because there is always something in work like this that can be done better.

In coming up with the mark, I am guided by the quality of your analysis (about 50% weight), introduction and data description (20%), conclusions and discussion (20%), and the overall quality of your writing (10%). Grading projects is not an exact science; every project is different, but I will endeavour to be consistent in my standards. There may not be time for much feedback. This is primarily a "summative assessment" whose purpose is to give you a grade, the grade reflecting the understanding that you are able to demonstrate in your work.

For your data set, you are ultimately responsible for deciding upon the research question or questions of interest. In some cases, I have made suggestions, which you may use if you wish, but you don't have to. Ultimately, it is up to you to decide what you think are "interesting" questions to answer. If appropriate, you should assess the assumptions behind your analysis and critically assess what you have done.

As for length, your report should be as long as is needed to tell your story, whatever it is, but not excessively long. When I read your report, it should "flow" nicely, so that things are in the report if and only if they add to your story, and that story is told in a logical way. If you want to include supplementary material that would interrupt the flow of your report, feel free to add an

Appendix to the end of your report to contain this. (I want to see your code somewhere, but if you don't want it to be in the body of your report, you can put that in an Appendix.)

If you use R for your analysis, you may (but do not have to) use an R Markdown document (R Notebook). My major interest is in seeing a coherent, well-thought-out analysis with appropriate conclusions; the closer you get to that, the better.

The data sets follow:

1. A number of different wines of the Pinot Noir type were rated by experts. A number of different qualities were assessed, "clarity", "aroma", "body", "flavour" and "oakiness", as well as an overall "quality". In addition, the region from which the wine came was recorded (labelled 1, 2 or 3). The data can be found in `http://www.utsc.utoronto.ca/%7ebutler/c32/wine.txt`.

   There are two principal questions of interest: does the overall quality of wine differ between the regions, and can the overall quality be predicted from the other quality ratings? Going along with the second question is whether high or low values of clarity, aroma, body, flavour and oakiness are associated with a high quality rating.

2. In professional football, a lot of statistics are collected on player and team performance. The data in `http://www.utsc.utoronto.ca/%7ebutler/c32/football.txt` are for the teams of the National Football League in 1976. Specifically, the variables recorded are: team name, games won, rushing yards, passing yards, punting average, field goal percentage (made divided by attempted times 100%), turnover differential (turnovers acquired minus turnovers lost), penalty yards, percent of rushing plays, opponents' rushing yds, opponents' passing yards.

   The important thing for any football team is winning. The research question here is what variables are associated with the number of games won, and in particular whether high or low values of those variables are associated with winning a lot of games.

3. In professional football (again), one of the ways of assessing the performance of a quarterback is via a "efficiency rating". The web page `http://www.nfl.com/stats/categorystats?statisticCategory=PASSING` shows up-to-date efficiency ratings of the quarterbacks in the National Football League (for this season). I have organized the same data into a file `http://www.utsc.utoronto.ca/%7ebutler/c32/qbrating.txt`. The variables are "rank" (ignore), quarterback's name, his team, his position ("QB" always, of course), the number of completions (passes caught), number of attempted passes, the percentage of passes that were caught, number of passing attempts per game, total yards gained by passing, average yards per completion, average passing yards per game, number of touchdown passes, number of interceptions, number of passes resulting in a first

down, percentage of that out of number of attempts, length of longest pass (a "T" on the end means that the pass was for a touchdown), number of passes for 20 yards or more, number of passes for 40 yards or more, number of sacks, quarterback rating.

Can you figure out the formula that the NFL uses to calculate quarterback efficiency ratings? (It is a fairly simple formula, and it depends only on the information in this data file, though you might need to calculate some new variables from the variables in the data file. You probably won't get an *exact* fit, because there are some "tweaks" that the NFL puts in.) I am more interested in the procedure that you use to attempt to discover the formula, than in the formula you come up with. Note that the formula does not depend on *how much* the quarterback has played. This might guide your choice of variables.

4. How do features of a house or apartment influence the asking price when that house or apartment is put up for sale? The website `realtor.ca` contains "MLS listings" of properties that were for sale in the fall of 2017. I randomly chose just over 50 properties in the area bounded roughly by Scarborough Town Centre on the northwest, Morningside on the east and Lawrence on the south. The properties are a mixture of houses, apartments and townhouses. The data are in a spreadsheet at `http://www.utsc.utoronto.ca/~butler/c32/houses.xlsx`. The columns are:

   - MLS listing number (an identifier for the property)
   - the asking price (dollars)
   - the number of bedrooms (a "den" in an apartment or an extra room in a townhouse that could be used as a bedroom counts 0.5)
   - the number of bathrooms
   - the type of property (apartment, house or townhouse).

   Develop a model for predicting asking price from the other variables (as many of them as seem necessary). Assess the assumptions of your model. How are changes in the explanatory variables reflected in the asking price?

   If you like, visit `realtor.ca` and find some properties in the same area currently for sale. How well are the asking prices predicted for these properties, using your model? If necessary, how might you modify your model to improve your predictions?

5. Sociologists like to speculate on the relationship between "crowding" and crime rates: that is, whether there is more crime in cities where the population density is higher, other things being equal. The data in `http://www.utsc.utoronto.ca/%7ebutler/c32/crowding.txt` are for a number of US cities. The variables recorded are: the name of the city, the population (in thousands), the percent of the population that is nonwhite, the population density (people per square mile), and the crime rate. Some of

the values are missing, denoted in the data file by `?`; you may need to take appropriate action to handle these in your analysis (and in your report, you should say what you did). Does crime rate depend on population density, or any of the other factors?

6. The data in `http://www.utsc.utoronto.ca/%7ebutler/c32/vocab.txt` come from over 20,000 individuals. For each individual, the following were recorded: the year data collected, whether respondent was male or female, years of education, score on 10-item vocabulary test. Come up with at least two interesting hypotheses that these data will enable you to assess, and then assess them. Note: there are only a few possible values of years of education and vocabulary test score, so you might want to investigate the R function `jitter` to improve your plots.

7. What affects the tip that a restaurant server receives? The owner of a bistro called *First Crush* in Potsdam, New York, collected a representative sample of 157 bills at the bistro over a two-week period. He recorded: the total amount of the bill, the amount of the tip, whether or not the bill was paid with a credit card (vs. being paid in cash), the number of guests in the party, the day of the week, the server (coded as A, B or C), and the size of the tip as a percentage of the total bill.

   The data are at `http://www.utsc.utoronto.ca/~butler/c32/restaurant-tips.txt`.

   A bistro is defined as "a small restaurant or cafe, often serving simple food."

   Investigate whether there is any relationship between any of these variables, most of which are factors, and the size of the tip, measured in what you think is an appropriate way. What seem to be the most important of these variables in determining how big a tip will be left?

8. Computer memory has gotten cheaper over time. At `https://jcmit.net/memoryprice.htm` you will find a table of memory price per megabyte at various times from 1957 to the present. (The first column is fractional year: year with appropriate number of months as given in the 4th column. The second column is the cost of the memory in US$ per megabyte.) Your challenge is to develop a satisfactory model for predicting memory cost as a function of year. You should also describe the process by which you got the data from the website or the spreadsheet linked there into your chosen software. What can you say about the evolution of memory price: can you come up with a good description of how memory price has changed, or do a prediction of memory price in some future year?