STAC32

Assignment 8

Due Thursday November 14 at 11:59pm

Question 2 below gives you instructions for setting up and structuring your SAS assignments. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

See `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html` for instructions on handing in assignments in Quercus. You can (and for SAS assignments, *should*) hand in a Word document.

Reminder: *there are no extensions due to failure to access software.* It is *your* responsibility to make sure that you allow yourself enough time to get connected to SAS Studio and to get your work done. You will be competing with a lot of other people, here and around the world, for access to SAS's servers, so it is up to you to allow enough time.

This is the last Assignment that you'll need to hand in. I will point you towards PASI-SAS for practice problems on the remaining material, and a reminder that your Project will be due at 11:59pm on the last day of classes (December 2).

1. Work through Section 2 of PASI-SAS, `http://ritsokiguess.site/STAC32/pasi-sas.pdf`. (You can also get to this from the course website, using the link at the top.) You *definitely* need to do question 2.1, or else you won't be able to do anything else.

2. To prepare yourself to hand in questions on SAS, work through Section 1 of PASI-SAS. (I probably have sections 1 and 2 the wrong way around, but so it is.) You need to *do* the first set of seven bullet points, and you need to *read and understand* the second set of three bullet points. If you don't do the first set of bullet points, you won't be able to hand in this Assignment properly.

   Hand the next question in:

3. A statistics course had two lecture sections, taught by the same instructor. One section used a typical variety of examples from different subject areas, while the other section used only sports-themed examples. The lecture sections were labelled as such, so that a student knew which type of section they were enrolling in, and could choose the one they preferred. The sections are labelled `Regular` and `Sports` in the data file. The sports-themed section had its lectures earlier in the day than the regular section.

   The data are in `https://www.utsc.utoronto.ca/~butler/c32/SportsExamples.csv`. For each student, the total number of points earned in the course was recorded, as well as which section they were in and their letter grade.

   (a) (2 marks) Read the data into SAS, and display the dataset. (It has 57 observations, which is rather longer than I like you to display normally, but the grader will check only that you have apparently the right thing.)

   Solution: This is a `.csv` file (making life easier for you). It's a file on the web, so do the `filename` thing first:

   ```
   filename myurl url
     "https://www.utsc.utoronto.ca/˜butler/c32/SportsExamples.csv";
   ```

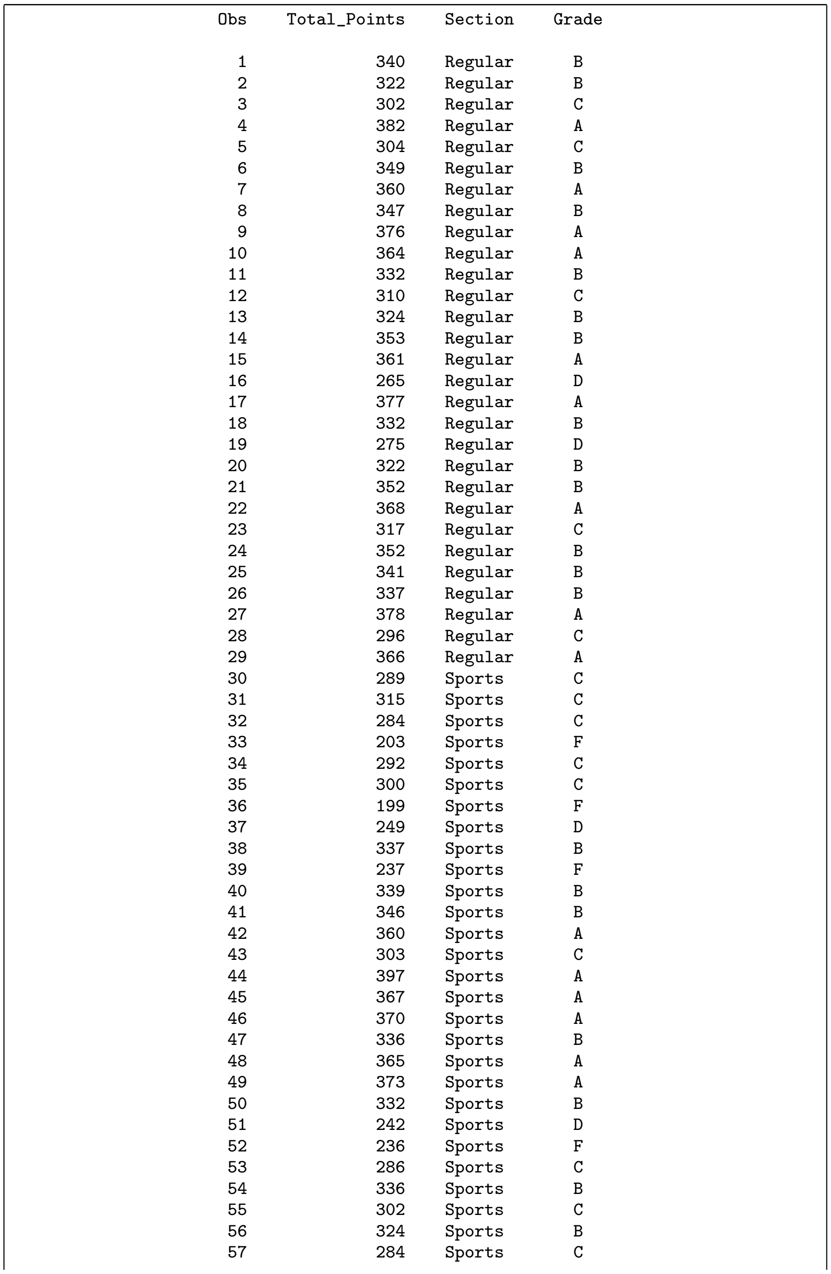
```
proc import
  datafile=myurl
  out=sections
  dbms=csv
  replace;
  getnames=yes;
```

This produces *no output.*

Then

```
proc print;
```

with output



```
Obs    Total_Points    Section    Grade

  1             340    Regular      B
  2             322    Regular      B
  3             302    Regular      C
  4             382    Regular      A
  5             304    Regular      C
  6             349    Regular      B
  7             360    Regular      A
  8             347    Regular      B
  9             376    Regular      A
 10             364    Regular      A
 11             332    Regular      B
 12             310    Regular      C
 13             324    Regular      B
 14             353    Regular      B
 15             361    Regular      A
 16             265    Regular      D
 17             377    Regular      A
 18             332    Regular      B
 19             275    Regular      D
 20             322    Regular      B
 21             352    Regular      B
 22             368    Regular      A
 23             317    Regular      C
 24             352    Regular      B
 25             341    Regular      B
 26             337    Regular      B
 27             378    Regular      A
 28             296    Regular      C
 29             366    Regular      A
 30             289    Sports       C
 31             315    Sports       C
 32             284    Sports       C
 33             203    Sports       F
 34             292    Sports       C
 35             300    Sports       C
 36             199    Sports       F
 37             249    Sports       D
 38             337    Sports       B
 39             237    Sports       F
 40             339    Sports       B
 41             346    Sports       B
 42             360    Sports       A
 43             303    Sports       C
 44             397    Sports       A
 45             367    Sports       A
 46             370    Sports       A
 47             336    Sports       B
 48             365    Sports       A
 49             373    Sports       A
 50             332    Sports       B
 51             242    Sports       D
 52             236    Sports       F
 53             286    Sports       C
 54             336    Sports       B
 55             302    Sports       C
 56             324    Sports       B
 57             284    Sports       C
```

That looks like the right kind of thing. I don't know what the Total Points is out of. Looks like it might be 400. (This is actually all of the students; there were 57 of them.)

Expect the grader to check that you have the right columns, and apparently the right values in each column (by glancing at the first few).

(b) (2 marks) Display the mean `TotalPoints` for each lecture section.

Solution:

Somewhere you need to discover that "Total Points" is actually `Total_Points` with uppercase and an underscore in it. (Actually, SAS is not case-sensitive, so the uppercase doesn't matter, but the underscore *does*.)

OK, the mean "total points" for each section. This is `proc means`. Glue the code below onto the end of your `proc import` and `proc print`, and run again:

```
proc means;
  var Total_Points;
  class Section;
```

I was careful about Uppercase, but it works just fine without (and so is fine for you).

Results:

```
                              The MEANS Procedure

                        Analysis Variable : Total_Points

              N
 Section    Obs    N            Mean         Std Dev         Minimum         Maximum
 ----------------------------------------------------------------------------------
 Regular     29   29     338.0689655      30.6045410     265.0000000     382.0000000

 Sports      28   28     307.2500000      52.5107573     199.0000000     397.0000000
 ----------------------------------------------------------------------------------
```

The highest score observed overall is 397, which further supports the maximum possible score being 400.

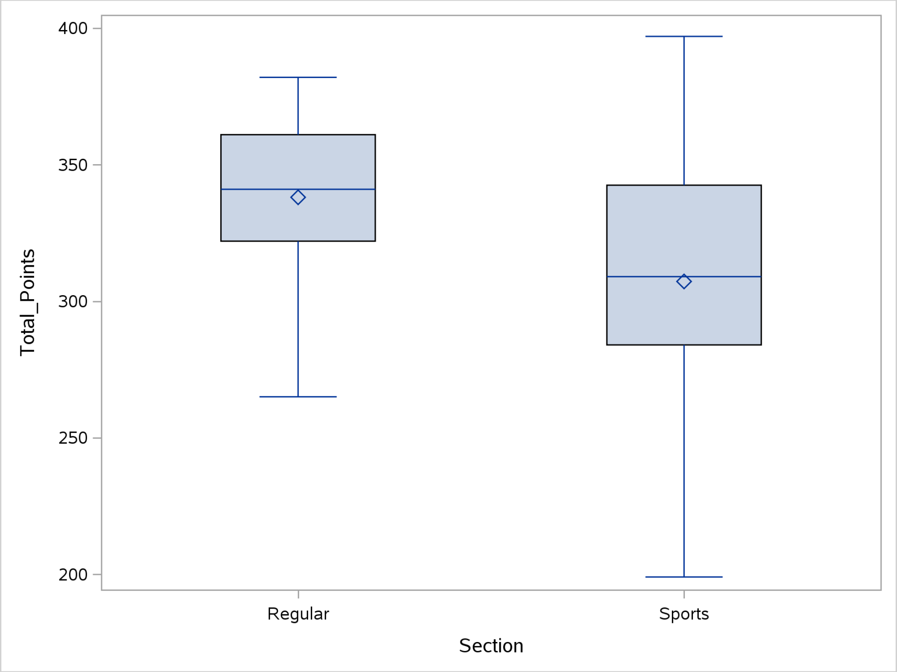
Extra: the reason for giving you this is mainly to give you practice at getting the means, but if you like you can comment that the mean score for the `Regular` group seems a fair bit higher than for the `Sports` group. (Then you can compare this conclusion with the one from the boxplot in the next part.)

(c) (3 marks) Obtain a boxplot of total points for each section. Does one of the sections have a clearly higher average than the other? Explain briefly.

Solution: `proc sgplot`. The syntax needs to be figured out, but once you have it, boxplots in SAS are really all the same. Glue the code below onto the end of what you have so far, and run again:

```
proc sgplot;
  vbox Total_Points / category = Section;
```

Here's the plot:



I would say that the average (mean or median) is clearly higher for the Regular section than for the Sports section, because most of the left-hand boxplot is higher than the right-hand one (or the median and quartiles are all noticeably higher). Have an opinion. If the opinion is "there is not much difference between the two sections because there is a lot of variability", I'm OK with that too. (An opinion plus a reason for that opinion is what I want to see.)

(d) (2 marks) What is it about the design of this study that would make you hesitant to say that having only sports examples is a bad idea? Explain briefly.

Solution: The key thing I'm after here is that the students were *not* randomly assigned to sections; the students could choose which section they enrolled in, knowing what kind of examples they would see. It could be that the weaker students chose the section with sports examples because they thought this would be more interesting, for example (in which case if you randomly assigned students to sections, the `Sports` section wouldn't look so bad).

Another way to argue it is that the `Sports` section was earlier in the day, and so students in that section were sleep-deprived, or were otherwise less capable of taking in the material. This would be another reason, independent of the subject matter, why students would do less well.

Come up with an argument that expresses a possible reason for the difference in means that is *not* that sports examples cause lower grades. You want another reason why the kind of grade difference seen might have been observed. I'm fairly relaxed about what that is, as long as it makes sense. But I want you to "join the dots": whatever you come up with has to be a convincing explanation. Expect only one point if it's not convincing enough.

If you said in the previous part that there was no great difference between the two sections, then you can say something like "having sports examples didn't help" and then tackle it as above (or kick off with whatever you would have epected to happen, and then say why it didn't). Or you can discuss something, like early vs. late, that you would have expected to make a difference (saying why) and note that it didn't. Be consistent with yourself. Whatever you said when you compared the boxplots, talk about that.

I did say that the same instructor taught both sections, so this is not a reason why the sections may have come out differently. (If the instructors had been different, that could have been another reason for the observed difference.)

If the students had been randomly assigned to sections that were *only* different because of the kind of examples, and this kind of difference in mean/median had been observed, *that* would have been a reason to say that the kind of examples makes a difference. That's why we design experiments the way we do: have only *one* thing be different, and *randomize* subjects to experimental conditions. That did not happen here.

Extra: this part is an argument for not doing any kind of inference, but with 28 and 29 students in each section, we could otherwise justify doing a two-sample t-test (there are no outliers and there is only moderate skewness). If we had done that test, and it had come out significant (which I think it will), it's not at all clear what that would have told us: yes, there is a difference in mean Total Points between the two sections, but that difference could be for any number of different reasons, not only because one kind of examples is better than the other.