STA 303H1S / 1002 HS Winter 2019 Assignment # 2
"Factoring GPA"

Posted: Sunday, February 3, 2019

Due: In Crowdmark by 10pm on Saturday, February 16, 2019.

Late assignments will be subjected to a penalty of 20% per day late. Submissions will not be allowed beyond 48 hours of the due date. Email submissions are not accepted.

Instructions:

- Use R (or R Studio) to do the analysis for the following questions.

- Use a benchmark significant level of 5%.

- Compile your solution as a PDF document (Word, LaTeXor Rmarkdown can be your base).

- Presentation of solutions is very important. Your assignment should have two main sections-Solutions and Appendix. Include relevant plots, and quote relevant numbers from your R output for your solutions. Unless asked otherwise, include all R codes and output in your Appendix. Marks will be awarded for excellent presentation.

- Write and submit **your own work**. For instance, personalized your code as much as possible, using your first name. **All plots produced must be given a title with the last 4 digits of your student number**.

- Where appropriate, your answers are expected to be written in plain English.

Grading: The grand total for this assignment is 100 marks. A general marking scheme for each part is given below:

Per Question Part

- 100%: complete and correct answers

- 80%: answers with minor problems

- 60%: good answers that are unclear, contain some mistakes, missing components

- 40%: poor answers with some value

- 0: incorrect or unanswered questions

Presentation and Appendix

- 10 points: well presented, easy to read, proper English used, R code and extra output in Appendix

- 6 points: good presentation, some R code in main write-up.

- 2 points: poor presentation, handwritten, hand-drawn diagrams, unnecessary R code in main section.

- 0 point: illegible, missing R-codes/output

The Data

The data is based on an optional, online class survey done on January 31. For the purposes of this assignment some data values were edited for correctness and privacy. The data file - "data2.csv" can be found in Quercus.

Our data consists of values from 399 students. We want to investigate factors that are related to a student's GPA. Specifically,

1. is one's expected grade related to their GPA?
2. do students who play video or computer games possess a different GPA from who do not play?

The variables in the dataset are:

- `Play`- the number of hours spent playing video or computer games in the three days prior to January 31,

- `GPA`- GPAs on the scale 0 to 4, and

- `Grade`- an expected grade (A+, A or B).

1. (15 marks) Create two new variables: (1) `Player`- by converting play time to a factor with 2 levels; 1 if a student spent some time playing video or computer games, and 0 otherwise, and (2) `Glay`- a variable that combines player status and expected grade. You can use the following R code to do this:

```
Player=array(0,399)
Glay<-NULL
for (i in 1:399)
{ if (Play[i]>0)
    {Player[i]=1}
else {Player[i]=0}
}
for (i in 1:399)
{ if (Player[i]==0 & Grade[i]=="B ")
  {Glay[i]="NonplayerNA"}
  else if (Player[i]==0 & Grade[i]=="A ")
  {Glay[i]="NonplayerA"}
  else if (Player[i]==0 & Grade[i]=="A+ ")
  {Glay[i]="NonplayerAP"}
  else if (Player[i]==1 & Grade[i]=="B " )
  {Glay[i]="PlayerNA"}
  else if (Player[i]==1 & Grade[i]=="A ")
  {Glay[i]="PlayerA"}
  else {Glay[i]="PlayerAP"}
}
Player=as.factor(Player)
Glay=as.factor(Glay)
```

Construct three sets of side-by-side boxplots:

   i. to compare GPA between players and non-players,
  ii. to compare GPA by expected grade and
 iii. to compare GPA among the 6 categories of the new factor- `Glay`.

Do there appear to be any differences? Explain.

2. (10 marks) Using the R pooled `t.test` procedure, investigate whether or not there is a difference in the GPA between players and non-players of video and/or computer games.

3. (15 marks) Investigate whether or not there is a difference in GPA among students classified by expected grade, using a one-way analysis of variance. If there is a difference among the levels of Grade, carry out an appropriate analysis to see which levels of Grade differ.

4. (15 marks) Use one-way analysis of variance to investigate whether or not there is a difference in GPA among the six categories of students classified by the combination of their player status and expected grade. If there is evidence of differences among the six categories of students, carry out an appropriate analysis to see which differ.

5. (15 marks) Do you trust the results of the statistical tests carried out in question 4? Assess whether the necessary assumptions of the model hold.
   Should we be concerned that the data contained different numbers of students in the three grade levels? Why or why not?

6. (10 marks) Instead of the one-way classification model used in question 4, a two-way analysis of variance model could have been used with player status, expected grade and their interaction. WITHOUT fitting this model, answer the following questions.

   (a) Write a mathematical equation to describe an interaction two-way analysis of variance model.

   (b) Would the number of predictor variables be the same as in the model used in question 4? Why or why not?

   (c) Would the F-test for the presence of interaction between expected grade and player status be statistically significant? How do you know from your results of question 4?

7. (5 marks) Discuss the use of Play as a quantitative explanatory variable rather than as a factor in an **additive** linear model for GPA. Include mathematical equations to describe the difference in models for GPA.

8. (5 marks) Name two additional potential factors of GPA and briefly describe their levels.