STA303H1S / 1002 HS Winter 2018 Assignment # 3
Who plays video games?

Due: In Crowdmark via Blackboard by 10pm on Thursday, March 22, 2018.
Late assignments will be subjected to a penalty of 5% per hour late.

Grading: The grand total for this assignment is 100 marks.

Instructions:

- Use R (or R Studio) to do the data analysis.

- Use a benchmark significance level of 10%. Report p-values to 4 decimal places.

- 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. In the Appendix, include your R code and other output. A maximum of 5 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.

PART 1-Research Article Review
(10 marks) Using the virtual Assignment #3 Library Guide under Assignments Section in the class website, identify a recent research article in your field of interest (for example, Finance, Health, Psychology) that includes a section with statistical analysis. The article must have been published by at least one University of Toronto author within the past five years (that is, 2013 to present). Note, STA1002 students are allowed to write about their current research (or proposal).
Based on your article of choice, answer the following.

1. What was your selected field of interest?

2. Write a proper reference for the article, including the author(s), title, journal, year of publication, volume and page indices.

3. Which UofT department was the UofT author affiliated with?

4. Provide a link to the article or a soft copy of the article.

5. Which statistical software was used for the data analysis?

6. Was the data derived from an observational study or experiment?

7. Did the article present summary statistics, tables and/or plots? Explain.

8. Did the article present test statistics, their distributions under H_0, p-values and/or confidence intervals? Explain.

9. To how many decimal places were values reported? Explain.

10. Identify at least one statistical method used to analyze the data.

- *Grading Notes:*

  - No more than 2 students are allowed to choose the same article. If this occurs, the marks for this part will be scaled by the number of persons with the same article.
  - If you work with another student on this part, then indicate the name of the student on your solutions.
  - For Participation 6 mark, give the reference of your article in the online Participation 6 Forum. This can be done by stating the title and author, or providing a link to the text or providing the article itself. This forum will be used by the grader (and can be used by student) to identify the first person or persons with the same article.

- *Further assistance* in acquiring suitable articles can be sought from Math Librarian, Bruce Garrod at the Math Library.

PART 2 - Contingency Tables
The Data
The data to be considered for this part is from *Stat Labs* by Nolan and Speed. The data was collected from a survey of introductory Statistics students at a US University in 1994. We will investigate difference between those who like to play video games and those who do not.

The file video.csv on Blackboard contains the data. The variables in the dataset are:

- like- whether the student liked to play video games or not (yes or no)

- sex- the sex (male or female) of a student

- grade- grade student expected in the Statistics course (A or not A (coded as nA))

1. *Analysis comparing proportions and using contingency tables:*

   (a) (10 marks) Construct a 2×2 table of sex by like. Is there evidence that sex is independent of a student's preference for playing video games? Quote 2 different p-values to support your answer. If there is evidence of association between the variables, explain in practical terms, with illustrative numbers, the nature of the association.

   (b) (15 marks) Examine the sex and like relationship separately for each grade type expected. Is there evidence that the association between sex and student's preference for playing video games changes with grade expected? Quote relevant p-values to support your answers.

2. *Analysis using Logistic Regression:*
   Since we are interested in whether or not students like to play video games, like can be considered as a response variable for these data and a logistic regression analysis could be carried out to determine the effect of sex and grade expected on the odds of liking video gaming.

   Fit two logistic regression models to these data, both with sex and grade as predictor variables. Let

   - Model 2.1 be the one to include interaction between sex and grade, and
   - Model 2.2 be the one without interaction.

   (a) (20 marks) Write the models being fit; clearly define all terms. Which of the two model should you use? Give the results of two tests that support your choice of logistic regression model. Explain clearly what is being tested for each test.

   (b) (10 marks) Give practical implications of the model selected in part (a). What do you conclude? Does it agree with your answer to question 1(b)?

3. *Analysis using Poisson Regression:*

The following table expresses the data as counts of independent students.

count	like	sex	grade
5	no	female	A
7	no	female	nA
1	no	male	A
7	no	male	nA
4	yes	female	A
22	yes	female	nA
21	yes	male	A
23	yes	male	nA

(a) (10 marks) Model the counts as Poisson variables and fit two models:

- Model 3.1 with explanatory variables `sex`, `grade` and `like`, the three two-way terms and the three-way interaction, and
- Model 3.2 - Model 3.1 with the three-way interaction term removed

Write the models being fit; clearly define all terms.

(b) (20 marks) Describe how the results from the Poisson regression models compare to the results in part 2 under Logistic regression modelling, in terms of:

   i. (5 marks) Deviance

  ii. (5 marks) Wald tests

 iii. (10 marks) Interpretation