Higher Diploma in Science in Data Analytics
Probability and Statistics
2019-2020 Semester 2
Assessment 2
26th March 2020
Examiner: Tom Corcoran

Overview

- The assessment is due to on Sunday 19th April at 23:59 and is worth 30%.

Submission details

- This is an individual assignment. You may not work in groups. Submission includes a signed statement saying it is your own work.
- Your submission will comprise of 4 files zipped, the project writeup, the project code, a video and a 1 page signed student plagiarism disclaimer form.
- The project writeup will take the form of a .pdf, you can choose to generate this .pdf with any tool including R Markdown, Word, Google Docs, etc.
- The project code can be a .R file or a .Rmd if you are using R Markdown to generate your pdf. The R code should be setup to read the data file from the same directory but do not submit the data file.
- The project video is mandatory for your assessment to get graded; it consists of you reflecting to the camera about your experience doing this project in detail and is a maximum of 3 minutes in .mp4 form
- Form A1: Signed copy of Form A1 from the AIT Plagiarism Policy (Available here: https://www.ait.ie/uploads/downloads/Plagiarism_Policy_Revised_Jan_201811.pdf)
- File names should be of the form:
    - Writeup: Assessment 2 <student_no>.pdf
    - Code: Assessment 2 <student_no>.R or .Rmd
    - Video: Assessment 2 <student_no>.mp4
    - Disclaimer: Assessment 2 <student_no> A1.pdf
    - where <student_no> is replaced by your student number

Grading criteria

The following criteria will be considered when marking your work:

- The R code should not generate any errors, e.g. all libraries used need to loaded. Comment out and explain any code that does not work but is your best attempt if you wish but very minimal marks will be given for commented out code.
- The Data should only be read in once in the project.
- Data cleaning is not required but if doing any (see extra credit), do not manually perform any data cleaning steps - all data cleaning should be done programmatically and be reproducible by running your code.
- Comment the code well, not every line needs to be commented but your code needs to be explained. This is especially true if you use R packages not covered in our class.
- All plots must have a title and labelled axes.
- Do not include multiple plots that all convey the same information.
- Correctness - are statistical procedures carried out and explained correctly?
- Each question will be graded based on both the analysis process and included visualisations where requested.
- Clearly written answers, that communicate the answers unambiguously.

Specification

The data file titled *AppUsage.csv* contains the registration and session dates and times for users of a mobile App as recorded on the server. The data is incomplete for the period after 31 March 2018.

- Email (anonymized)
- Gender (1-Male, 2-Female)
- Date of birth
- Location of user (1-US, 2-Europe, 3-Africa, 4-Asia, 5-South America, 6–Australasia)
- App registration – date and time.
- App session – date and time at start of session.

Analysis

1. **Gender Effect**. Using appropriate analysis investigate whether there is a gender difference in registered users. Display the results visually and include a single sentence interpretation.
2. **Location Effect**. Investigate the location of registered users. Display the results visually and include a short interpretation.
3. **Age Effect**. Using appropriate analysis investigate the age distribution of users. Is there a gender difference in the age distribution? You should measure the average and spread of the age values and you should also produce a histogram for the age of the users. Display the results visually and include a short interpretation.
4. **Times of Sessions**. Investigate the times of sessions. Is there any useful information for the App's product manager in this data? Maintenance and updates can take up to one hour and so when should the company schedule this work to minimize disruption?
5. **Growth in Users**. Investigate the trend and produce a rough prediction for the number of new registered users that there are likely to be in the three-month period from 1 April through to 30 June 2018. You will need to use the regression line for the prediction. Comment on your prediction.
6. **Usage frequency**. How often do users use the App? Produce metrics to analyse this and investigate if there is a difference between genders.
7. **App Retention**. Assuming a user is defined as "Active" if they have used the app in the last 4 weeks, produce a plot of the proportion of people who are still active at regular times since the release of the App.

Hints

- *dplyr* package - mutate can be used to add columns, e.g. the hour of the session.
- *lubridate* package for parsing *App Session*
- The usage value can be found as either the Sessions per Day or as the time between Sessions.

Extra credit

The maximum is 30%, but if you have dropped points elsewhere in this assessment then extra credit can be gained as follows:

- There is no requirement to check and clean the data, but you are welcome to do any as you see fit. Any such work needs to be fully reproducible by running your R code.
- For Problem #5 Growth in users, you can use the *predict()* function to give a more exact prediction.
- Originality and creativity, e.g. expressive & effective visualisations.