STAT 6240 - HW 0 - Due 1/24

Due Date: January 24th, Thursday, 6:10 PM

Your homework submission should contain:

1. Your R code,
2. Outputs from R (including figures),
3. Answers to the questions.

You should use *R Markdown* or *R Notebooks*. If you want to learn more about *R Markdown*, R Studio has an online tutorial at http://rmarkdown.rstudio.com/lesson-1.html. However, the best way to learn is probably by creating an R Notebook file and then adjusting it to change the output.

Do not print out your homework. Save it as a pdf file and upload it to Blackboard.

This week's homework is the assignment given to data analyst applicants at the Wikimedia Foundation. You can find the full assignment at https://github.com/wikimedia-research/Discovery-Hiring-Analyst-2016.

This is a fairly standard data wrangling assignment. You are also asked to use your statistics knowledge and perform some analyses.

Background

Wikimedia Foundation relies on *event logging* (EL) to track a variety of performance and usage metrics to help them make decisions. Specifically, they are interested in:

- *clickthrough rate*: the proportion of search sessions where the user clicked on one of the results displayed
- *zero results rate*: the proportion of searches that yielded 0 results

and other metrics outside the scope of this task.

Task

You must create a **reproducible report** answering the following questions:

1. What is their daily overall clickthrough rate? How does it vary between the groups?
2. Which results do people tend to try first? How does it change day-to-day?
3. What is their daily overall zero results rate? How does it vary between the groups?
4. Let *session length* be approximately the time between the first event and the last event in a session. Choose a variable from the dataset and describe its relationship to session length.
5. Summarize your findings in an *executive summary*. **We will discuss this part in class. Be prepared to present your results**.

Data

The dataset comes from a tracking schema for assessing user satisfaction. Desktop users are randomly sampled to be anonymously tracked by this schema which uses a "I'm alive" pinging system that we can use to estimate how long the users stay on the pages they visit. The dataset contains just a little more than a week of EL data.

Column	Value	Description
uuid	string	Universally unique identifier (UUID) for backend event handling.
timestamp	integer	The date and time (UTC) of the event, formatted as YYYYMMDDhhmmss.
session_id	string	A unique ID identifying individual sessions.
group	string	A label ("a" or "b").
action	string	Identifies in which the event was created. See below.
checkin	integer	How many seconds the page has been open for.
page_id	string	A unique identifier for correlating page visits and check-ins.
n_results	integer	Number of hits returned to the user. Only shown for searchResultPage events.
result_position	integer	The position of the visited page's link on the search engine results page (SERP).

The following are possible values for an event's action field:

- **searchResultPage**: when a new search is performed and the user is shown a SERP.
- **visitPage**: when the user clicks a link in the results.
- **checkin**: when the user has remained on the page for a pre-specified amount of time.

Example Session

uuid	timestamp	session_id	group	action	checkin	page_id	n_results	result_position
4f..	20160305195246	..efc	b	searchResultPage	NA	1b..	7	NA
75..	20160305195302	..efc	b	visitPage	NA	5a..	NA	1
77..	20160305195312	..efc	b	checkin	10	5a..	NA	1
42..	20160305195322	..efc	b	checkin	20	5a..	NA	1
8f..	20160305195332	..efc	b	checkin	30	5a..	NA	1
29..	20160305195342	..efc	b	checkin	40	5a..	NA	1

This user's search query returned 7 results, they clicked on the first result, and stayed on the page between 40 and 50 seconds. (The next check-in would have happened at 50s.)