Assignment 1 - Obtaining Data

Assignment Questions

1. **What are the types of data available to you?**
   The data is a log of all the user clicks and action on a prototype application developed for a previous class. The data was logged in a couple of SQL tables: a user information table and a log table for all users.

2. **For data sets: how many records are in the data set?**
   41 Users, 2980 Log Entries

3. **For API: what are the limits on fetching data?**
   Not Applicable. I tried initially to get data from Google Places. However, the API limits the number of results to 60. The script is available in the folder.

4. **Provide an "interesting" record, explain its properties and why it is interesting**
   The data was to be used to evaluate the features and UI elements of a web application developed for another class. The application provides an interface that helps users deliver speeches better. All the user actions are logged (button clicks and setting values. The application also captures demographics information (gender, age...etc.). The logging functionality was put in place to allow remote user testing using MobileWorks. The analysis would help recommend the best default configuration for the users based on their profile among other things. The product research work and previous experiences show that font size and speech pace vary depending on age and gender.
   Example:
   {"user_id": 39,"**gender**": "**Male**","**age**": **28**,"education": "Graduate Student","**location**": "Berkeley","temp_id": 725165,"date_added": "2013-12-13 15:15:57","log_id": 2525,"user_id": 39,"target_id": null,"**target**": "**Change Theme**","value": "**BW**","action": 0,"timestamp": "2013-12-13 15:18:33"},

5. **What are 3 questions you could answer using your data?**
   a. How many people participated? what are their genders, ages?
   b. What is the total number of clicks? by gender? by Age?
   c. What are the most actions or objects clicked overall? by gender?
   d. Who are the top users?
   e. What was the favorite test file?

Activities:

- Cleaned up user data and add random values for missing information (i.e. missing Age and Gender information).
- Standardized gender to (Female, Male).
- Ran a joint SQL to generate one table
- Exported the output to a json file
- Cleaned up the JSON file to make it 1 row per line. I could have used CSV but that would remove column titles from each row are useful for command line analysis.
- Conducted Analysis

Notes: Some results are superficial due to the randomly added demographics values. Some of the questions below are there just to explore the technique and don't necessarily serve a purpose.

Data Questions:

1. How many people participated? what are their genders, ages?

41 (21 Male, 20 Female).

Different ages. Highest frequency is age 25. Average age is 29.1.

2. What is the total number of clicks? by gender? by Age?

2980 clicks (1663 "Male", 1317 "Female"). Males seem to have more clicks per person (79 and 65). Average is 72.7 clicks.

Ages 25, 26, 27 have the highest clicks followed by ages 31 and 32.

3. What are the most actions or objects clicked overall? by gender?

"Change Speed", "Highlighted Word", and "Bigger Font". The trend is the same for both genders.

4. Who are the top users?

Users 10, 4, 8, 6 and 38 respectively.

5. What was the favorite test file?

Speech 1.

Console Output

Number of users

```
egrep -o '\{"user_id": [0-9]*' podium_users.json  | uniq | wc -l
41
```

Users by Gender

```
egrep -o '\{"user_id": [0-9]*,[^:]*: "[^"]*"' podium_users.json | uniq -c | sort -nr | egrep -
o '"* "[^0]*"'|sort | uniq -c
     20  "Female"
     21  "Male"
```

Users by Age

```
grep -o '\{"user_id": [0-9]*,[^$]*,"age": [0-9]*' podium_users.json  | uniq | egrep -o '"age":
[0-9]*'| sort | uniq -c
      3 "age": 20
      1 "age": 21
      2 "age": 22
      1 "age": 23
      6 "age": 25
      1 "age": 26
      3 "age": 27
      3 "age": 28
      1 "age": 29
      1 "age": 30
      3 "age": 31
      3 "age": 32
      4 "age": 33
      1 "age": 34
      4 "age": 35
      1 "age": 36
      1 "age": 37
      1 "age": 38
      1 "age": 39
```

Average Age

```
grep -o '\{"user_id": [0-9]*,[^$]*,"age": [0-9]*' podium_users.json  | uniq | egrep -o '"age":
[0-9]*'|awk '{sum+=$NF+0} END{print "average " sum/NR}'
average 29.0976
```

Total and Average Clicks

```
grep -o '\{"user_id": [0-9]*' podium_users.json  | sort | uniq -c |  awk '{sum+=$1+0; }
END{print "Total Clicks=" sum "  Average=" sum/NR}'
Total Clicks=2980  Average=72.6829
```

Clicks by Gender

```
egrep -o '"gender": "[^"]*"' podium_users.json | egrep -o ' "[^"]*"' | sort | uniq -c | sort -
nr
   1663  "Male"
   1317  "Female"
```

Average Clicks by Gender

```
echo '1663/21' | bc
79
echo '1317/20' | bc
65
```

Clicks by Age

```
egrep -o '("age":) [0-9]*' podium_users.json | sort|uniq -c | sort -nr
    686 "age": 25
    645 "age": 26
    321 "age": 27
    266 "age": 31
    194 "age": 32
    184 "age": 30
    166 "age": 35
    118 "age": 20
     82 "age": 22
     78 "age": 28
     77 "age": 33
     75 "age": 21
     57 "age": 37
     24 "age": 29
      2 "age": 36
      2 "age": 23
      1 "age": 39
      1 "age": 38
      1 "age": 34
```

Clicks by Action Type

```
egrep -o '"target": "[^"]*"' podium_users.json | sort | egrep -o '* "[^"]*"' | uniq -c | sort
-nr
   2130  "Change Speed"
    223  "Highlighted Word"
    113  "Bigger Font"
     90  "Play Button"
     61  "Pause Button"
     49  "Test File"
     48  "Change Text Mode"
     48  "Back"
     34  "Change Theme"
     33  "Change Font"
     31  "Change Line Height"
     21  "Smaller Font"
     20  "Pause Overlay"
     19  "Text Mode"
     15  "Line Height"
     14  "Themes"
     12  "Fonts"
      8  "New File"
      7  "Open File"
      2  "File Opened"
      2  "Feedback"
```

Top Users

```
grep -o '\{"user_id": [0-9]*' podium_users.json  | sort | uniq -c |  sort -nr | head -5
    645 {"user_id": 10
    345 {"user_id": 4
    254 {"user_id": 8
    201 {"user_id": 6
    191 {"user_id": 38
```

Female Clicks by Action Type

```
grep 'Female' podium_users.json | egrep -o '"target": "[^"]*"' | sort | egrep -o '* "[^"]*"' |
uniq -c | sort -nr
    852  "Change Speed"
    113  "Highlighted Word"
     62  "Bigger Font"
     59  "Play Button"
     44  "Pause Button"
     29  "Change Text Mode"
     28  "Test File"
     28  "Back"
     25  "Change Theme"
     12  "Change Line Height"
     11  "Text Mode"
     11  "Pause Overlay"
     10  "Themes"
      8  "Line Height"
      7  "Fonts"
      7  "Change Font"
      3  "Smaller Font"
      3  "Open File"
      3  "New File"
      1  "File Opened"
      1  "Feedback"
```

Male Clicks by Action type

```
grep 'Male' podium_users.json | egrep -o '"target": "[^"]*"' | sort | egrep -o '* "[^"]*"' |
uniq -c | sort -nr
   1278  "Change Speed"
    110  "Highlighted Word"
     51  "Bigger Font"
     31  "Play Button"
     26  "Change Font"
     21  "Test File"
     20  "Back"
     19  "Change Text Mode"
     19  "Change Line Height"
     18  "Smaller Font"
     17  "Pause Button"
      9  "Pause Overlay"
      9  "Change Theme"
      8  "Text Mode"
      7  "Line Height"
      5  "New File"
      5  "Fonts"
      4  "Themes"
      4  "Open File"
      1  "File Opened"
      1  "Feedback"
```

Favorite Test File

```
egrep -o '"target": "[^"]*","value": [^,]*"' podium_users.json | grep "Test File" | sort |
uniq -c | sort -nr
     23 "target": "Test File","value": "Speech 1"
     16 "target": "Test File","value": "Speech 2"
     10 "target": "Test File","value": "Speech 3"
```