Spring Health Assignment

Stephen Hanna

July 14, 2019

The total number of unique member id hashes is 1166, given by the number of rows in a unique member id hashed dataframe.

```
data <-read.csv("spring_health_take_home_df.csv",header=T)
#Reads in excel file
nrow(unique(data[2]))
```

```
## [1] 1166
```

```
#Gets number of rows for unique member id hashes dataframe
```

The average frquency of occurrences for each member id hash is 2.8, but since each member should fill in both questionnaire each time perhaps this number is better off halved to 1.4.

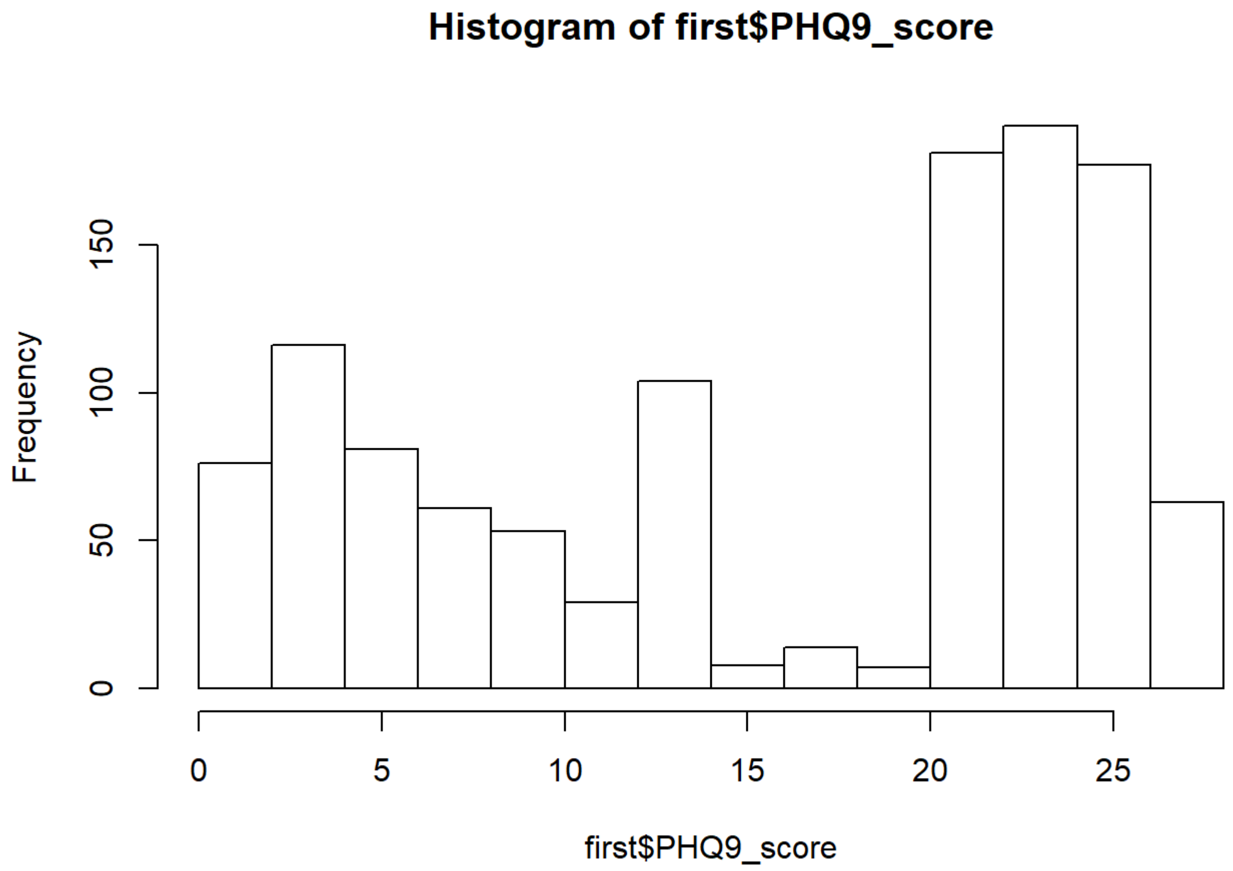
```
frequency <- table(data[2])
#Makes table of each member hash and the frequency of its occurence
mean(frequency)
```

```
## [1] 2.803602
```

```
#Gives average frequency of occurence of a member hash, which is 2.8
```

The PHQ9 scores do not seem to have a normal distribution, but most people do seem to have a rather high PHQ9 score. This suggests a biased population where most people who are members likely are depressed. This makes sense, since the platform is meant to be used, and provide benefits to, those with depression.

```
data$assessment_created_at <- substr(data$assessment_created_at, 1, 10)
#Gets first 10 characters of date for when the assessment was created, then replaces the origina
l timestamp with that
sorted <- data[order(data$member_id_hashed, as.Date(data$assessment_created_at, format="%Y-%m-%
d")),]
#Sorts the data by member hash, then by date in ascending order (later dates are lower)
PHQ9 <- sorted[sorted$questionnaire_kind == 'PHQ9',]
#Extracts only the rows which have PHQ9 as the questionnaire used
first <- PHQ9[match(unique(PHQ9$member_id_hashed), PHQ9$member_id_hashed),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the baseline PHQ9 score
first$PHQ9_score <- as.numeric(first$PHQ9_score)
#Changes the PHQ9 score from a string to an integer
hist(first$PHQ9_score)
```



```
#Makes histogram visualization of PHQ9 baseline score
summary(first$PHQ9_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     1.0     7.0    21.0    16.1    24.0    27.0
```

```
#Gives summary statistics of PHQ9 baseline score
```

The average difference of PHQ9 score from last use and first use is .64. Since intervals of 5 is usually used for each grouping of PHQ9 scores, this suggests a significant change does not often occur for symptoms of depression in users.

```
PHQ9_frequency <- table(PHQ9$member_id_hashed)
#Makes table of each member hash and the frequency of its occurence, but only for PHQ9 questionn
aire
hashed_more_than_once <- subset(PHQ9, member_id_hashed %in% names(PHQ9_frequency[PHQ9_frequency
 > 1]))
#Makes dataframe that only contains member hashes that occurred more than once for PHQ9 question
naire
sorted_asc <- hashed_more_than_once[order(hashed_more_than_once$member_id_hashed, as.Date(hashed
_more_than_once$assessment_created_at, format="%Y-%m-%d")),]
#Sorts the data by member hash, then by date in ascending order (later dates are lower)
sorted_des <- hashed_more_than_once[rev(order(hashed_more_than_once$member_id_hashed,as.Date(has
hed_more_than_once$assessment_created_at, format="%Y-%m-%d"))),]
#Sorts the data by member hash, then by date in descending order (later dates are higher)
sorted_first <- sorted_asc[match(unique(sorted_asc$member_id_hashed), sorted_asc$member_id_hashe
d),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the baseline PHQ9 score
sorted_last <- sorted_des[match(unique(sorted_des$member_id_hashed), sorted_des$member_id_hashe
d),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the final PHQ9 score
sorted_first_hash <- sorted_first[order(sorted_first$member_id_hashed),]
#Orders the rows by member hash
sorted_last_hash <- sorted_last[order(sorted_last$member_id_hashed),]
#Orders the rows by member hash, this sorted_first_hash and sorted_last_hash can be combined in
 a matching hash order
avg_change_PHQ9 <- data.frame(sorted_first_hash$PHQ9_score, sorted_last_hash$PHQ9_score)
#Combines the first and last PHQ9 scores into a single row for each member with more than one sc
ore
avg_change_PHQ9$sorted_first_hash.PHQ9_score <- as.numeric(avg_change_PHQ9$sorted_first_hash.PHQ
9_score)
#Changes the PHQ9 score from a string to an integer
avg_change_PHQ9$sorted_last_hash.PHQ9_score <- as.numeric(avg_change_PHQ9$sorted_last_hash.PHQ9_
score)
#Changes the PHQ9 score from a string to an integer
avg_change_PHQ9$diff <- (avg_change_PHQ9$sorted_first_hash.PHQ9_score - avg_change_PHQ9$sorted_l
ast_hash.PHQ9_score)
#Makes new column with difference between last and first PHQ9 score
mean(avg_change_PHQ9$diff)
```

```
## [1] 0.6428571
```

```
#Gets average difference between first and last PHQ9 scores
```

The average difference of PHQ9 score from last use and first use is .64. Since intervals of 5 is usually used for each grouping of PHQ9 scores, this suggests a significant change does not often occur for symptoms of depression in users who are suspected to have depression.

```
depressed <- hashed_more_than_once[hashed_more_than_once$PHQ9_positive == 'TRUE',]
#Extracts rows only where PHQ9 scores positively for depression
sorted_asc2 <- depressed[order(depressed$member_id_hashed, as.Date(depressed$assessment_created_
at, format="%Y-%m-%d")),]
#Sorts the data by member hash, then by date in ascending order (later dates are lower)
sorted_des2 <- depressed[rev(order(depressed$member_id_hashed,as.Date(depressed$assessment_creat
ed_at, format="%Y-%m-%d"))),]
#Sorts the data by member hash, then by date in descending order (later dates are higher)
sorted_first2 <- sorted_asc2[match(unique(sorted_asc2$member_id_hashed), sorted_asc2$member_id_h
ashed),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the baseline PHQ9 score
sorted_last2 <- sorted_des2[match(unique(sorted_des2$member_id_hashed), sorted_des2$member_id_ha
shed),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the final PHQ9 score
sorted_first_hash2 <- sorted_first2[order(sorted_first2$member_id_hashed),]
#Orders the rows by member hash
sorted_last_hash2 <- sorted_last2[order(sorted_last2$member_id_hashed),]
#Orders the rows by member hash, this sorted_first_hash and sorted_last_hash can be combined in
 a matching hash order
avg_change_PHQ9_2 <- data.frame(sorted_first_hash2$PHQ9_score, sorted_last_hash2$PHQ9_score)
#Combines the first and last PHQ9 scores into a single row for each member with more than one sc
ore
avg_change_PHQ9_2$sorted_first_hash2.PHQ9_score <- as.numeric(avg_change_PHQ9_2$sorted_first_has
h2.PHQ9_score)
#Changes the PHQ9 score from a string to an integer
avg_change_PHQ9_2$sorted_last_hash2.PHQ9_score <- as.numeric(avg_change_PHQ9_2$sorted_last_hash
2.PHQ9_score)
#Changes the PHQ9 score from a string to an integer
avg_change_PHQ9_2$diff <- (avg_change_PHQ9_2$sorted_first_hash2.PHQ9_score - avg_change_PHQ9_2$s
orted_last_hash2.PHQ9_score)
#Makes new column with difference between last and first PHQ9 score
mean(avg_change_PHQ9_2$diff)
```

```
## [1] 0.1646091
```

```
#Gets average difference between first and last PHQ9 scores
```

The number of unproductive days goes down by an average of one day for from the time of first use to last use of the platform, suggesting the platform is effective at reducing the number of unproductive days.

```
SDS <- sorted[sorted$questionnaire_kind == 'SDS',]
#Extracts only the rows which have SDS as the questionnaire used
SDS_frequency <- table(SDS$member_id_hashed)
#Makes table of each member hash and the frequency of its occurence, but only for SDS questionna
ire
hashed_more_than_once_SDS <- subset(SDS, member_id_hashed %in% names(SDS_frequency[SDS_frequency
> 1]))
#Makes dataframe that only contains member hashes that occurred more than once for SDS questionn
aire
sorted_asc3 <- hashed_more_than_once_SDS[order(hashed_more_than_once_SDS$member_id_hashed, as.Da
te(hashed_more_than_once_SDS$assessment_created_at, format="%Y-%m-%d")),]
#Sorts the data by member hash, then by date in ascending order (later dates are lower)
sorted_des3 <- hashed_more_than_once_SDS[rev(order(hashed_more_than_once_SDS$member_id_hashed,a
s.Date(hashed_more_than_once_SDS$assessment_created_at, format="%Y-%m-%d"))),]
#Sorts the data by member hash, then by date in descending order (later dates are higher)
sorted_first3 <- sorted_asc3[match(unique(sorted_asc3$member_id_hashed), sorted_asc3$member_id_h
ashed),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the baseline SDS score
sorted_last3 <- sorted_des3[match(unique(sorted_des3$member_id_hashed), sorted_des3$member_id_ha
shed),]
#Extacts only the rows which are the first time a member hash has occurred, which is presumed to
be the final SDS score
sorted_first_hash3 <- sorted_first3[order(sorted_first3$member_id_hashed),]
#Orders the rows by member hash
sorted_last_hash3 <- sorted_last3[order(sorted_last3$member_id_hashed),]
#Orders the rows by member hash, this sorted_first_hash and sorted_last_hash can be combined in
 a matching hash order
avg_change_SDS <- data.frame(sorted_first_hash3$SDS_days_unproductive, sorted_last_hash3$SDS_day
s_unproductive)
#Combines the first and last SDS scores into a single row for each member with more than one sco
re
avg_change_SDS$sorted_first_hash3.SDS_days_unproductive <- as.numeric(avg_change_SDS$sorted_firs
t_hash3.SDS_days_unproductive)
#Changes the SDS score from a string to an integer
avg_change_SDS$sorted_last_hash3.SDS_days_unproductive <- as.numeric(avg_change_SDS$sorted_last_
hash3.SDS_days_unproductive)
#Changes the SDS score from a string to an integer
avg_change_SDS$diff <- (avg_change_SDS$sorted_last_hash3.SDS_days_unproductive - avg_change_SDS
$sorted_first_hash3.SDS_days_unproductive)
#Makes new column with difference between last and first SDS score
mean(avg_change_SDS$diff)
```

```
## [1] -1.063973
```

```
#Gets average difference between first and last SDS scores
```

A correlation analysis was conducted for the average difference in PHQ9 scores and number of unproductive days from first and last use. Given a value of -.14, it seems that there is little relation between depressive symptom reduction and reduction of number of unproductive days. Since the platform seems effective at reducing number of unproductive days but not symptoms of depression, it could be that the platform helps people cope with the fatigue that comes with depression, but does not actually make their life more pleasant or fulfilling.

A paired t test was used to see if the platform is effective at mitigating symptoms of depression and number of unproductive days. A p value less than .01 for number of unproductive days strongly suggests, with greater than 99% certainty, that there is a significantly lower number of unproductive days after using the platform. A p value of .34 for symptoms of depression suggests that the platform likely does not cause a change in depressive symptoms. Put more numerically, there is a 65% chance there is a significantly lower number of depressive symptoms after using the platform, which is far from a statistically strong confidence threshold.

```
SDS_diff <- data.frame(sorted_first_hash3$member_id_hashed, avg_change_SDS$diff)
#Makes dataframe just using member hashes and the difference in SDS unproductive day values from
last and first interaction
PHQ9_diff <- data.frame(sorted_first_hash$member_id_hashed, avg_change_PHQ9$diff)
#Makes dataframe just using member hashes and the difference in PHQ9 values from last and first
 interaction
colnames(SDS_diff)[1] <- "member_id_hashed"
#Changes column name for member id hashed
colnames(PHQ9_diff)[1] <- "member_id_hashed"
#Changes column name for member id hashed to match that for the SDS diff dataframe so an inner j
oin can be done
merged <- merge(SDS_diff,PHQ9_diff, by = "member_id_hashed")
#Performs an inner join of the SDS diff and PHQ9 diff values based on member id hashed
cor(merged$avg_change_SDS.diff,merged$avg_change_PHQ9.diff)
```

```
## [1] -0.1400205
```

```
#Gets correlation between the SDS unproductive day values and the PHQ9 values
t.test(avg_change_SDS$sorted_last_hash3.SDS_days_unproductive, avg_change_SDS$sorted_first_hash
3.SDS_days_unproductive, paired = TRUE, alternative = "two.sided")
```

```
## 
##  Paired t-test
## 
## data:  avg_change_SDS$sorted_last_hash3.SDS_days_unproductive and avg_change_SDS$sorted_first
_hash3.SDS_days_unproductive
## t = -6.9141, df = 296, p-value = 2.902e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3668187 -0.7611275
## sample estimates:
## mean of the differences
##               -1.063973
```

```
#Performs paired t test to see if there is a significant change in number of days unproductive
t.test(avg_change_PHQ9$sorted_last_hash.PHQ9_score, avg_change_PHQ9$sorted_first_hash.PHQ9_scor
e, paired = TRUE, alternative = "two.sided")
```

```
## 
##  Paired t-test
## 
## data:  avg_change_PHQ9$sorted_last_hash.PHQ9_score and avg_change_PHQ9$sorted_first_hash.PHQ9
_score
## t = -0.96241, df = 265, p-value = 0.3367
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9580518  0.6723375
## sample estimates:
## mean of the differences 
##              -0.6428571
```

```
#Performs paired t test to see if there is a significant change in PHQ9 score
```