HOME WORK-II
RE-IDENTIFICATION AND ANONYMIZATION IN PRACTICE

SUMMARY REPORT

SUBMITTED BY: Agasthya Harekal(adh5677@psu.edu)

RE-IDENTIFICATION OF DATA

The re-identification of data is a major concern which has to be looked into in any dataset that has been made public. In the current dataset, we have app_usage, calendar, call_log that have a csv file for each user with device id. This device id can be used to re-identify the user by having data that has details of device id and corresponding username. This data could be easily available for employees of company that offered mobile sensing technology as a service to the users/ students.The device id should not have been made public. Sensitive information can be obtained of users such as overall mobile device usage, call patterns. Further, after re-identification of the user, dinning folder has text files of each user which has information regarding the time and date of each meal ate by the user. This again exposes the eating patterns of the users. Further education folder has class csv file that exposes courses taken by each user, deadlines.csv file that exposes all the deadlines of the user, grades.csv that shows the grades of each user and finally the piazza.csv file that exposes the usage patterns of the piazza form of all the users like the days online, views etc. In many of the json file responses of the users the location cooridnates are exposed. In sensing based data, it is ensured that absolute time is not exposed, but only timestamp is exposed. Although, in gps csv files, the location of each user is exposed. Again, in the sms data, each sms csv file belonging to the user exposes the device id of the device used by the user. Further, in the surveys, lot of information is collected about the user. In the BigFive survey, each user's behaviour is assessed and each user's reponses is stored. This could be personal to each user but is getting exposed. In the Flourishing scale survey each user has to rate about how one feels about various aspects of one's life. Each user's response is recorded and each user's response is mentioned as a record in the csv file. In Loneliness survey similarly the user responses regarding loneliness based questions is collected. In the PercievedStressScale survey, each user's responses on stress experienced is collected in a file. In PHQ-9 survey, user responses regarding basic health is recorded. Further, in pspi survey each user's mental health is recorded. In the last survey, health condition of user is directly questioned as a response and further responses are taken about health. These surveys question about health, behaviour, routine which can be co-related to a person whom one knows and the person can be re-identified. By getting to know the user's private responses all of these information can be traced to an individual by the process of re-identification. This also may be done using the device id or any other attribute that exposes the user indirectly. Hence, these are some of re-identification risks.

ANONYMIZATION OF DATA

For anonymization, we select the vr_12.csv, where we consider health of the student as sensitive/ confidential attribute as present in the column name "In general, would say your health is". There are many quasi-identifiers in the data which might be able to help identify a known person apart from the userid. Here, we select the most efficient of the quasi-identifiers that do not cause lot of wastage of records, which can be beneficial from utility perspective. Here we choose the column with names "How much of the time during the past 4 weeks: Have you felt calm and peaceful? " , "How much of the time during the last 4 weeks: Have you felt downhearted and blue"and "Compared to one year ago, how would you rate your physical health in general now?" which have responses like "A good bit of the time", "Most of the time", "Yes, a little of the time","None of the time".

Further we implement MinGen Algorithm, with some variations or changes.

- First step is generalisation , all these responses present in these columns are further generalised as FeltBetter, FeltLessBetter where responses like "A good bit of time", "None of time" where judged as FeltBetter, FeltLessBetter based on the corresponding columns.
- Further, the confidential attribute "In general, would say your health is", which has values like "Very Good", "Excellent", "Good","Fair" is converted into equivalence classes "Good Health" and "Bad Health".
- In the further approach, we group the data using quasi identifiers into rows.
- Next, we consolidate all the rows based on whether k-anonmyity condition is met, that is number of rows in the grouping based on quasi-identifiers is greater than k, t-closeness of each group with entire data is within the expected threshold and l-diversity of the group is greater than expected value. We reject all the group of rows that do not meet the criterion.

From this algorithm we get anonymised data and further we can consider the utility aspect of the data anonymised.

UTILITY ANALYSIS

The generalisation approach utilised for all quasi-identifiers ensured that the data will have only two values, “FeltBetter”, FeltLessBetter” as responses. But, this does not reduce the utility of the dataset since the co-relation between the sensitive attribute and quasi-identifiers is maintained and dataset can be further utilised to understand about the sensitive attribute using quasi-identifiers. But, attackers cannot directly identify the sensitive attribute using information about the person and data that was revealed initially by the quasi-identifiers. Later, after process of anonymization of data using k=5,t=0.3,l=1, we find that the anonymised data has lost only 10.8 percent of the records from the initial data. This was possible since we considered only two columns as quasi-identifiers and did not many quasi-identifiers for generalisation. As k value and t value is increased, more number of rows get rejected, since there are only two equivalence classes we cannot increase l greater than 2. As more quasi-identifiers are used, more rows get rejected and further decreases the overall utility of the data. Without user data being sensing data can be used for analysis except gps. General analysis can be done on app usage of students, caller trends without infringement on the privacy of users. The surveys can also be used to do analysis of general physical and mental health of students.

OUTPUT

```
Intial Data Frame
Very good    29
Good         25
Excellent    18
Fair         11
Name: In general, would you say your health is, dtype: int64
No, none of the time         49
Yes, a little of the time    21
Yes, some of the time        11
Yes, all of the time          1
Yes, most of the time         1
Name: Accomplished less than you would like., dtype: int64
A little of the time      32
Some of the time          26
A good bit of the time    14
None of the time           9
Most of the time           2
Name: How much of the time during the past 4 weeks: Have you felt downhearted and blue?, dtype: int64
Some of the time          27
A good bit of the time    25
Most of the time          20
A little of the time       8
All of the time            2
None of the time           1
Name: How much of the time during the past 4 weeks: Have you felt calm and peaceful?, dtype: int64
A little of the time    37
None of the time        26
Some of the time        15
Most of the time         4
All of the time          1
Name: During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?, dty
A good bit of the time    34
Some of the time          23
Most of the time          12
A little of the time       8
All of the time            4
None of the time           2
Name: How much of the time during the past 4 weeks: Did you have a lot of energy?, dtype: int64
```

```
About the same     37
Slightly better    23
Slightly worse     15
Much worse          5
Much better         3
Name: Compared to one year ago, how would you rate your physical health in general now?, dtype: int64
About the same     30
Slightly worse     21
Slightly better    18
Much better         8
Much worse          6
Name: Compared to one year ago, how would you rate your emotional problems (such as feeling anxious, depressed or irritable) now?, dtype: int64
Data after generalization
    uid  ... Compared to one year ago, how would you rate your emotional problems (such as feeling anxious, depressed or irritable) now?
0   u00  ...                                    About the same
1   u01  ...                               FeltBetter(EmoProb)
2   u02  ...                               FeltBetter(EmoProb)
3   u03  ...                           FeltLessBetter(EmoProb)
4   u04  ...                                    About the same
..  ...  ...                                               ...
78  u51  ...                                    About the same
79  u52  ...                           FeltLessBetter(EmoProb)
80  u53  ...                           FeltLessBetter(EmoProb)
81  u56  ...                                    About the same
82  u59  ...                                    About the same

[83 rows x 16 columns]
    uid  ... equivalence_class
2   u02  ...        Bad Health
5   u05  ...       Good health
8   u09  ...       Good health
10  u12  ...       Good health
15  u17  ...        Bad Health
18  u20  ...        Bad Health
23  u30  ...       Good health
28  u35  ...        Bad Health
29  u36  ...        Bad Health
33  u44  ...       Good health
39  u51  ...        Bad Health
53  u09  ...       Good health
```

```
53  u09  ...       Good health
54  u10  ...       Good health
65  u30  ...       Good health
66  u31  ...       Good health
67  u32  ...       Good health
71  u36  ...       Good health
72  u42  ...       Good health
76  u47  ...       Good health
78  u51  ...        Bad Health
81  u56  ...       Good health

[21 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Good health is 0.6666666666666666

Group Frequency of the equivalence class variable Bad Health is 0.3333333333333333

Process of T-Closeness Has Ended
    uid  ... equivalence_class
4   u04  ...       Good health
12  u14  ...       Good health
25  u32  ...       Good health
27  u34  ...       Good health
40  u52  ...        Bad Health
42  u56  ...       Good health
43  u57  ...       Good health
44  u58  ...       Good health
48  u02  ...        Bad Health
50  u04  ...       Good health
51  u05  ...       Good health
55  u14  ...       Good health
69  u34  ...       Good health
75  u45  ...       Good health

[14 rows x 17 columns]
```

```
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Good health is 0.8571428571428571

Group Frequency of the equivalence class variable Bad Health is 0.14285714285714285

Process of T-Closeness Has Ended
    uid  ... equivalence_class
30  u39  ...        Bad Health
46  u00  ...        Bad Health
56  u15  ...       Good health
73  u43  ...        Bad Health
77  u49  ...       Good health

[5 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Bad Health is 0.6

Group Frequency of the equivalence class variable Good health is 0.4

Process of T-Closeness Has Ended
    uid  ... equivalence_class
24  u31  ...        Bad Health

[1 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Bad Health is 1.0

Process of T-Closeness Has Ended
```

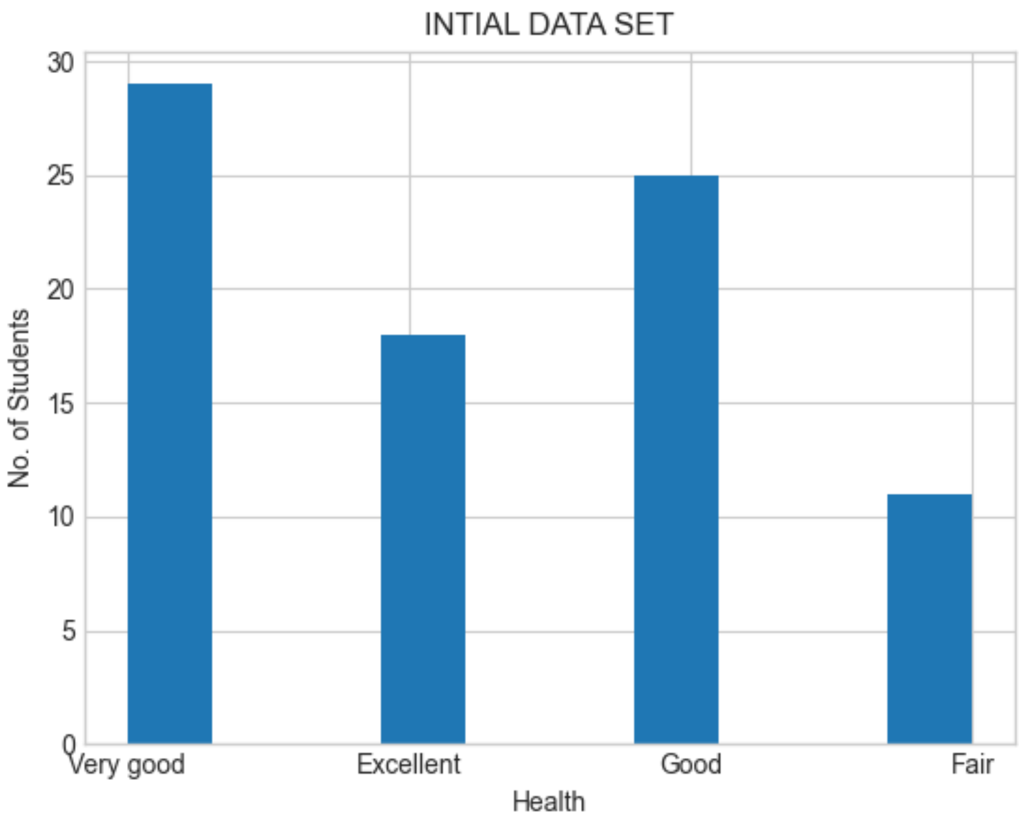
```
Process of T-Closeness Has Ended
    uid  ... equivalence_class
1   u01  ...       Good health
9   u10  ...       Good health
16  u18  ...        Bad Health
17  u19  ...        Bad Health
22  u27  ...        Bad Health
31  u42  ...       Good health
47  u01  ...       Good health
57  u16  ...        Bad Health
60  u19  ...       Good health
70  u35  ...        Bad Health
82  u59  ...       Good health

[11 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Good health is 0.5454545454545454

Group Frequency of the equivalence class variable Bad Health is 0.45454545454545453

Process of T-Closeness Has Ended
    uid  ... equivalence_class
11  u13  ...        Bad Health
26  u33  ...        Bad Health
41  u53  ...        Bad Health
59  u18  ...        Bad Health
64  u27  ...        Bad Health

[5 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Bad Health is 1.0
```

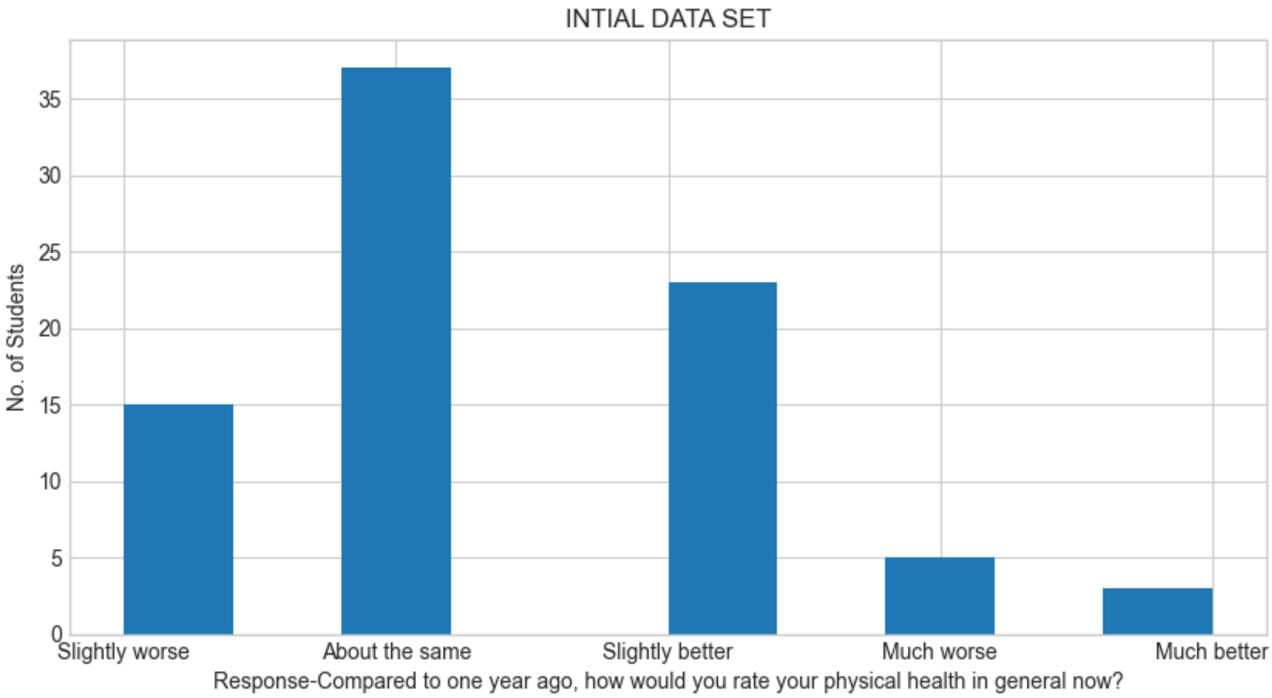
```
Process of T-Closeness Has Ended
    uid  ... equivalence_class
3   u03  ...       Good health
13  u15  ...       Good health
14  u16  ...        Bad Health
19  u22  ...       Good health
21  u24  ...       Good health
36  u47  ...       Good health
37  u49  ...       Good health
38  u50  ...        Bad Health
45  u59  ...        Bad Health
49  u03  ...       Good health
61  u20  ...       Good health
79  u52  ...        Bad Health

[12 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Good health is 0.6666666666666666

Group Frequency of the equivalence class variable Bad Health is 0.3333333333333333

Process of T-Closeness Has Ended
    uid  ... equivalence_class
0   u00  ...       Good health
6   u07  ...        Bad Health
7   u08  ...       Good health
20  u23  ...        Bad Health
32  u43  ...        Bad Health
34  u45  ...       Good health
35  u46  ...       Good health
52  u07  ...        Bad Health
63  u24  ...        Bad Health
68  u33  ...        Bad Health
74  u44  ...       Good health
```

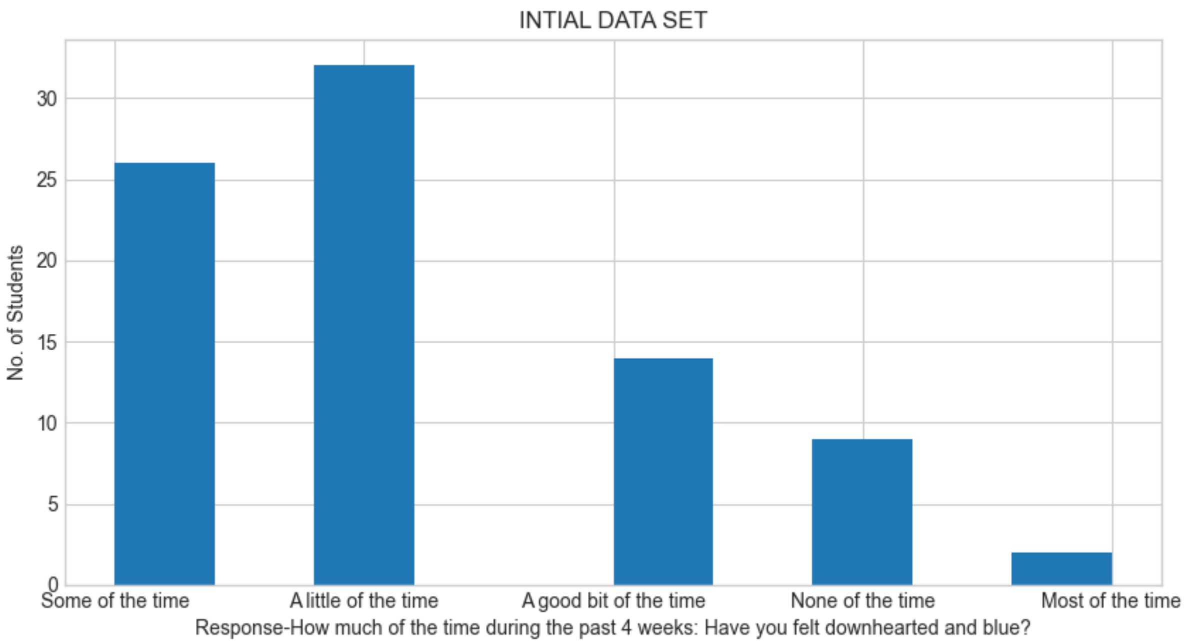
```
[11 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Bad Health is 0.5454545454545454

Group Frequency of the equivalence class variable Good health is 0.45454545454545453

Process of T-Closeness Has Ended
    uid  ... equivalence_class
58  u17  ...        Bad Health
62  u23  ...        Bad Health
80  u53  ...        Bad Health

[3 rows x 17 columns]
Process of T-Closeness Started
DataFrame Frequency of the equivalence class variable Good health is 0.5662650602409639

DataFrame Frequency of the equivalence class variable Bad Health is 0.43373493975903615

Group Frequency of the equivalence class variable Bad Health is 1.0

Process of T-Closeness Has Ended
Final Dataframe
    uid  ... equivalence_class
0   u02  ...        Bad Health
1   u05  ...       Good health
2   u09  ...       Good health
3   u12  ...       Good health
4   u17  ...        Bad Health
..  ...  ...               ...
69  u46  ...       Good health
70  u07  ...        Bad Health
71  u24  ...        Bad Health
72  u33  ...        Bad Health
73  u44  ...       Good health
```

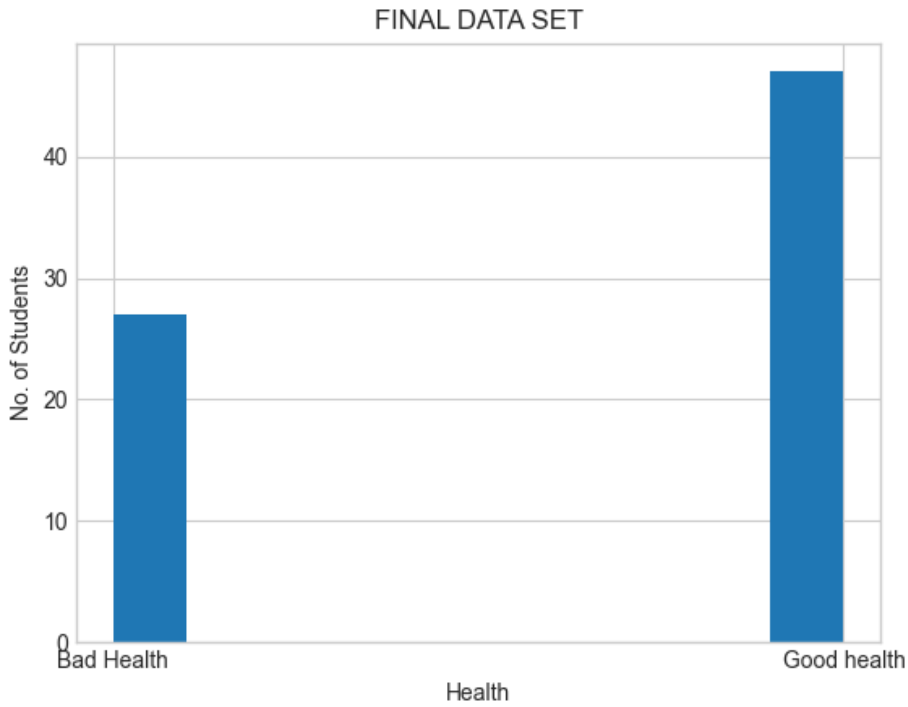
```
[74 rows x 17 columns]
Good health    47
Bad Health     27
Name: equivalence_class, dtype: int64
FeltBetterAcc        64
FeltLessBetterAcc    10
Name: Accomplished less than you would like., dtype: int64
FeltBetter(Down)      40
FeltLessBetterDown    34
Name: How much of the time during the past 4 weeks: Have you felt downhearted and blue?, dtype: int64
FeltBetter(Int)        59
FeltLessBetter(Int)    15
Name: During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.
FeltBetter(Ener)      68
FeltLessBetterEner     6
Name: How much of the time during the past 4 weeks: Did you have a lot of energy?, dtype: int64
About the same              32
FeltBetter(CompPhyH)        26
FeltLessBetter(CompPhyH)    16
Name: Compared to one year ago, how would you rate your physical health in general now?, dtype: int64
About the same             30
FeltLessBetter(EmoProb)    27
FeltBetter(EmoProb)        26
Name: Compared to one year ago, how would you rate your emotional problems (such as feeling anxious, depressed or irritable) now?, dtype: int64
Data Set is k-anonymised

Process finished with exit code 0
```

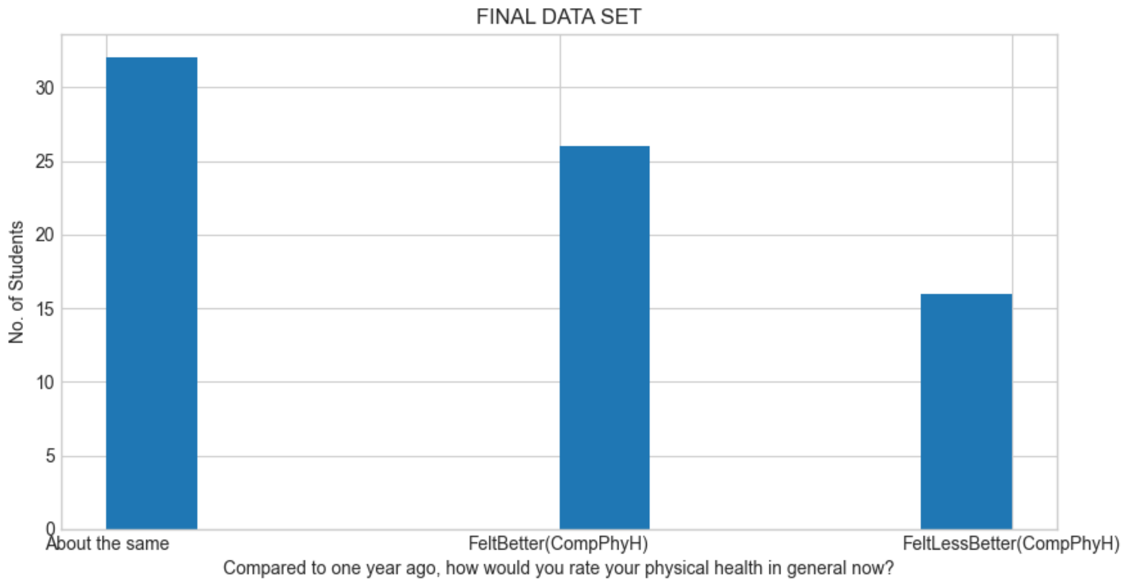
PLOTS

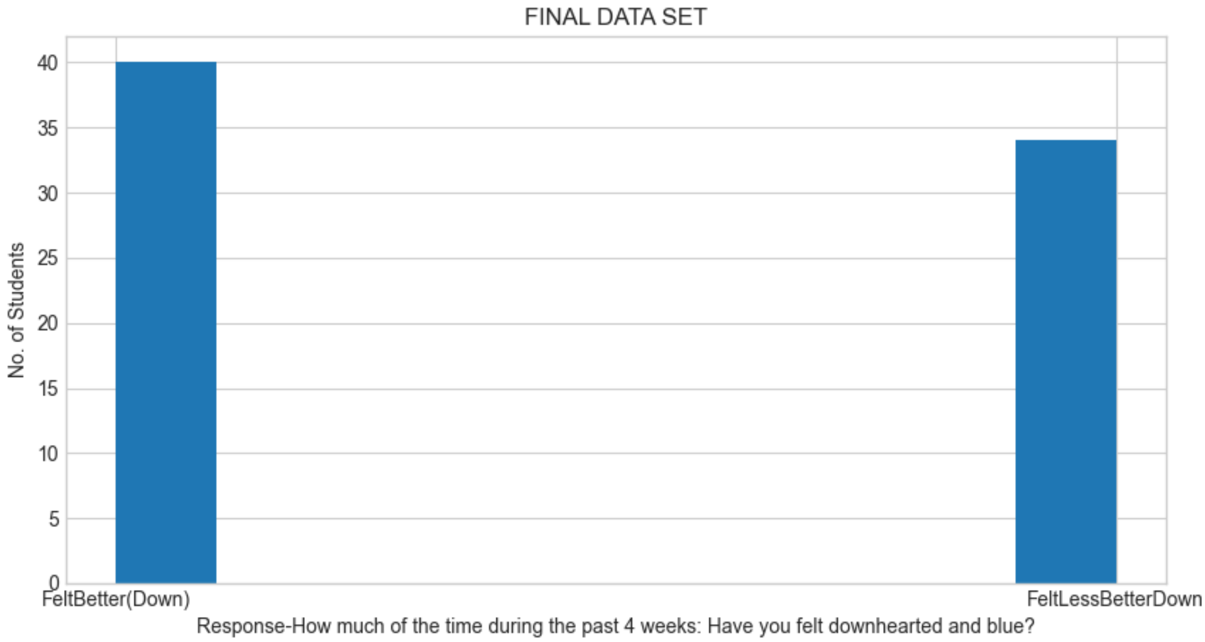



INTIAL DATA SET
No. of Students
0
5
10
15
20
25
30
35
Slightly worse
About the same
Slightly better
Much worse
Much better
Response-Compared to one year ago, how would you rate your physical health in general now?


INTIAL DATA SET
No. of Students
0
5
10
15
20
25
30
Some of the time
A little of the time
A good bit of the time
None of the time
Most of the time
Response-How much of the time during the past 4 weeks: Have you felt downhearted and blue?


FINAL DATA SET
No. of Students
0
10
20
30
40
Bad Health
Good health
Health


FINAL DATA SET
No. of Students
0
5
10
15
20
25
30
About the same
FeltBetter(CompPhyH)
FeltLessBetter(CompPhyH)
Compared to one year ago, how would you rate your physical health in general now?


FINAL DATA SET
No. of Students
0
5
10
15
20
25
30
35
40
FeltBetter(Down)
FeltLessBetterDown
Response-How much of the time during the past 4 weeks: Have you felt downhearted and blue?