6 Oct 2020

FYP 4th meeting

Contents

1 Exploratory Data Analysis

2 Feature Engineering

DIAMANTE Student dataset

- Total of 121 columns, 3770 rows
- 29 columns has missing values

Variable:	Number of objects:	Number of missing values:
eth_other	90	3680
income_ladder	3725	45
nat_lang_able	1749	2021
country	1620	2150
years_us	1575	2195

DIAMANTE Student dataset

Variable:	Number of objects:	Number of missing values:
start_phq8_prd	3725	25
start_gad_prd	3680	90
time_msg	2942	828
days.since.F0	3165	605
days.since.F1	3155	615
days.since.F2	3152	648
days.since.F3	3134	636

DIAMANTE Student dataset

Variable:	Number of objects:	Number of missing values:
days.since.F4	3224	546
days.since.M0	3066	704
days.since.M1	3190	580
days.since.M2	3267	503
days.since.M3	3294	476
days.since.T1	3274	496
days.since.T2	3275	495

DIAMANTE Student dataset

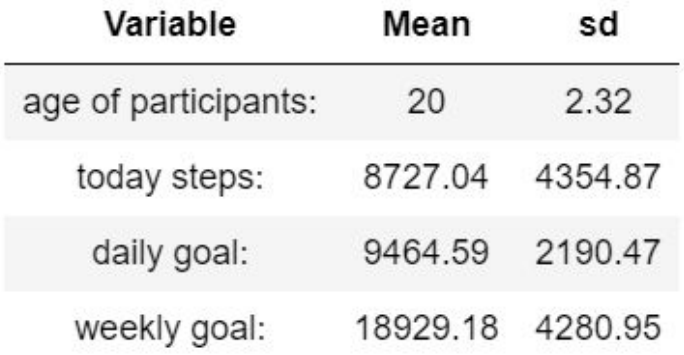
Variable:	Number of objects:	Number of missing values:
days.since.T3	3250	520
days.since.T4	3242	528
days.since.ind	3199	571
days.since.soc	3097	673
days.since.M3	3294	476
yesterday_progess	3412	358
yesterday_steps	3248	522

DIAMANTE Student dataset

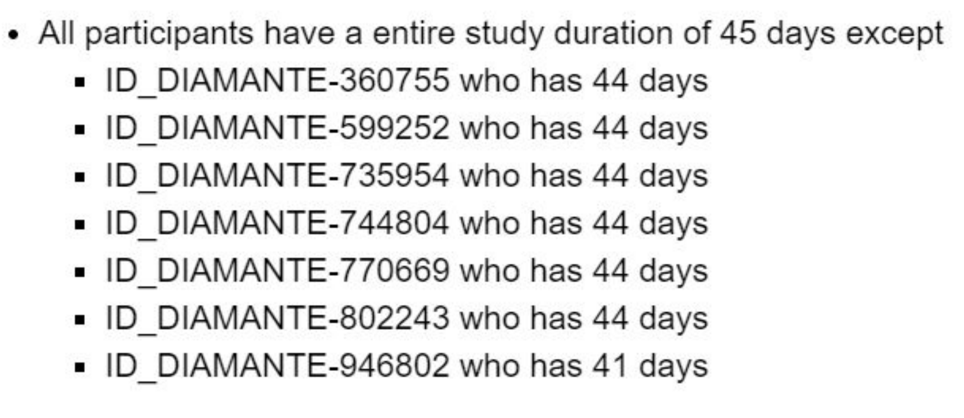
Variable:	Number of objects:	Number of missing values:
today_steps	3210	560
daybefyest_steps	3196	574
step_change_yest	3055	715
step_change_today	3095	675

- Out of these 29 variables, 4 variables **eth_other**, **nat_lang_able**, **country**, **years_us** has more than 2000 missing values

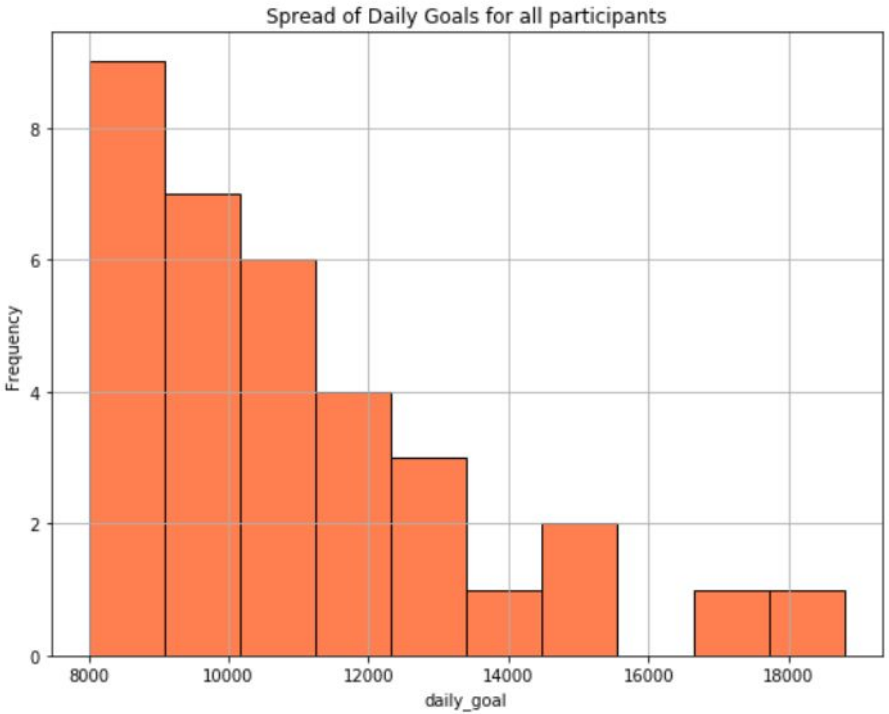
Some descriptive statistics (n = 84)

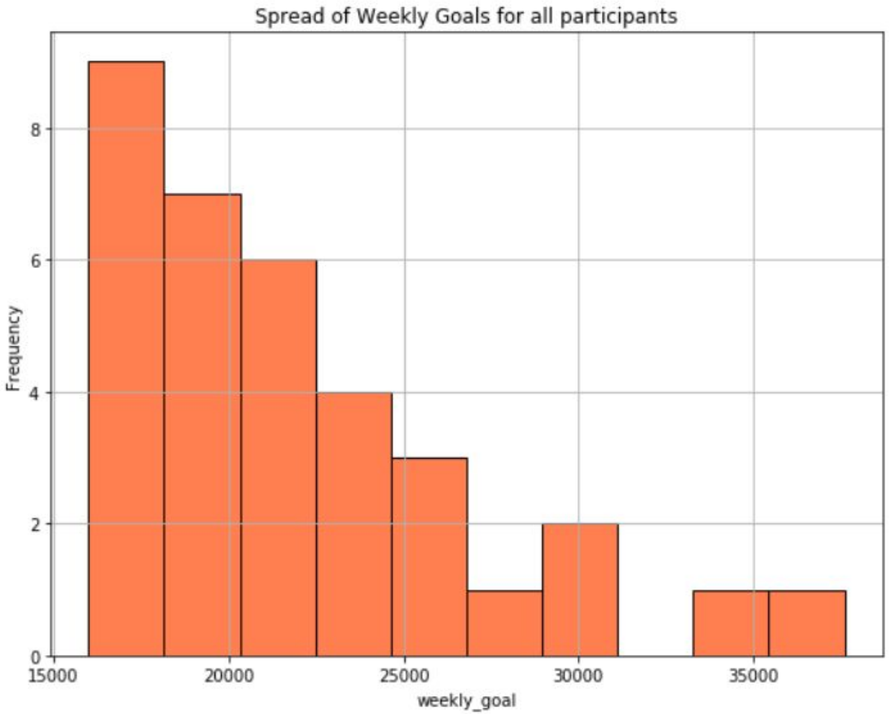


Variable	Mean	sd
age of participants:	20	2.32
today steps:	8727.04	4354.87
daily goal:	9464.59	2190.47
weekly goal:	18929.18	4280.95



- All participants have a entire study duration of 45 days except
  - ID_DIAMANTE-360755 who has 44 days
  - ID_DIAMANTE-599252 who has 44 days
  - ID_DIAMANTE-735954 who has 44 days
  - ID_DIAMANTE-744804 who has 44 days
  - ID_DIAMANTE-770669 who has 44 days
  - ID_DIAMANTE-802243 who has 44 days
  - ID_DIAMANTE-946802 who has 41 days

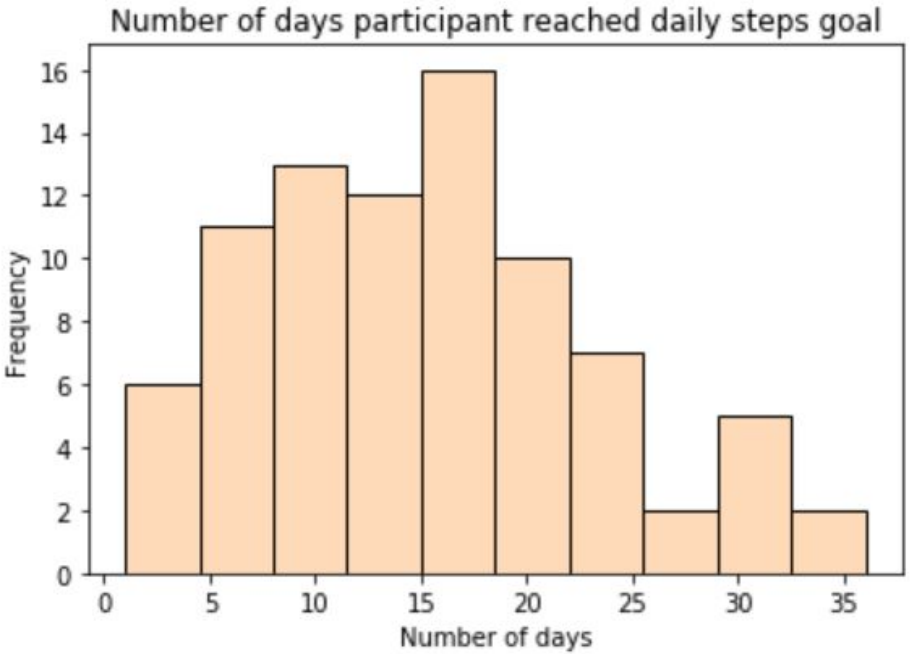




- Majority of participants indicated their daily steps goal of 8000 to 10000
- Majority of participants indicated their weekly steps goal of 15000 to 22500

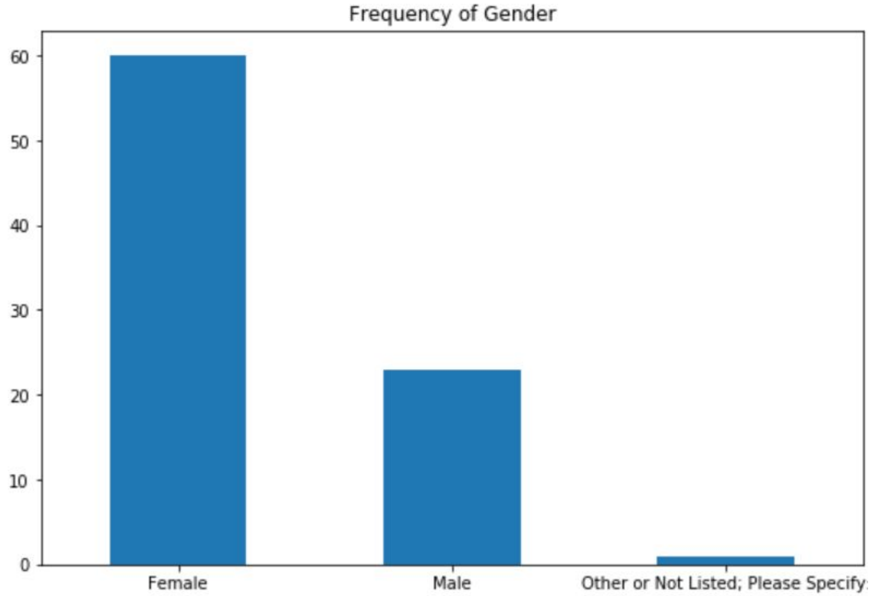
Distribution of variables:

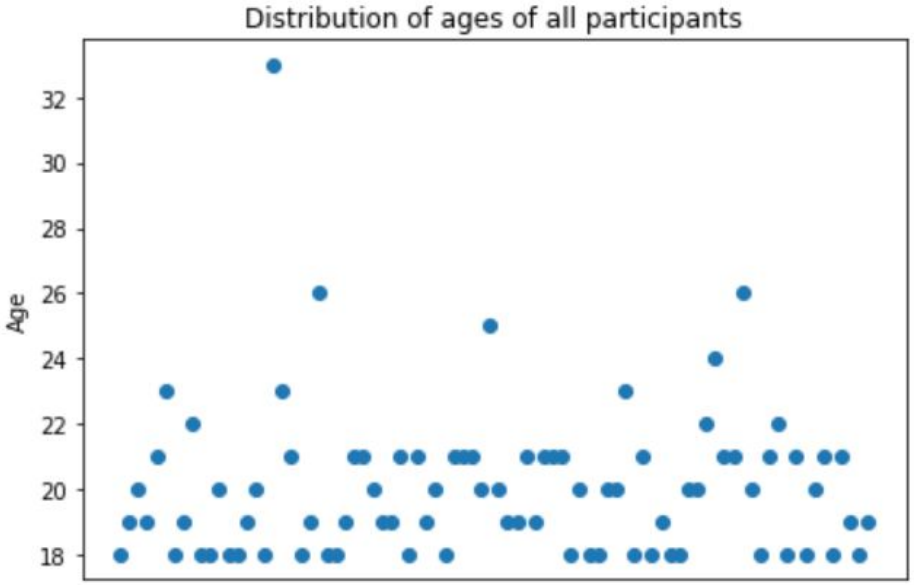
Number of days participant reached daily steps goal



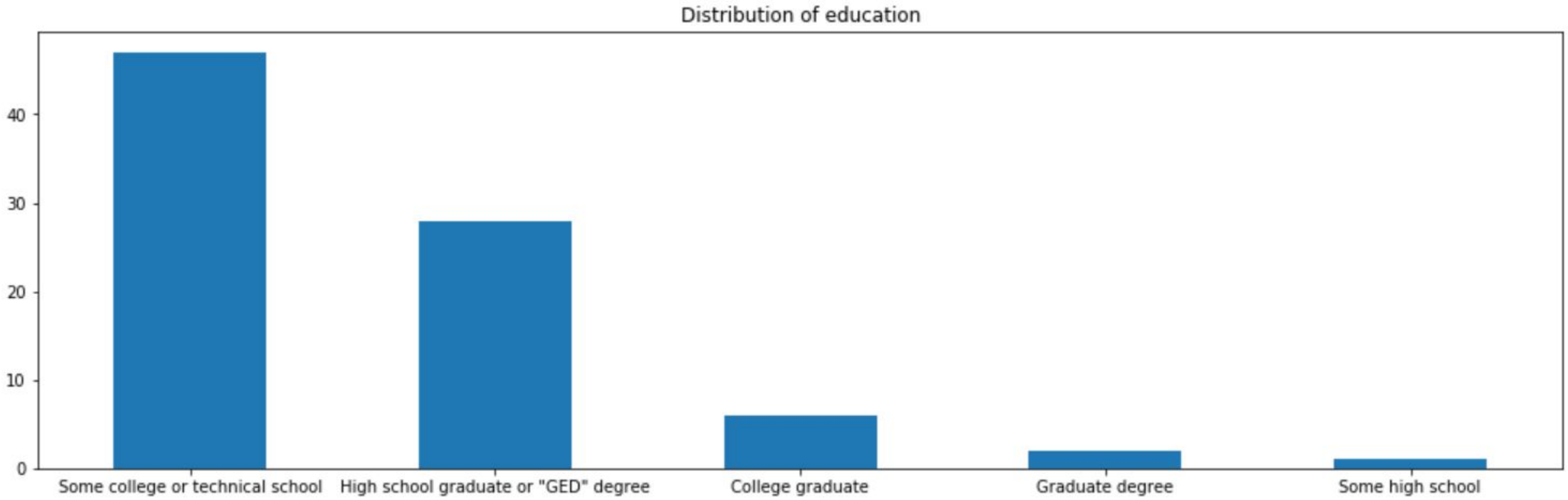
- Out of the whole duration of 45 days for each participants, most participants reached their daily steps goals for about 15 to 19 days
- No participants hit their daily steps goals for all 45 days

Some barcharts for different variables

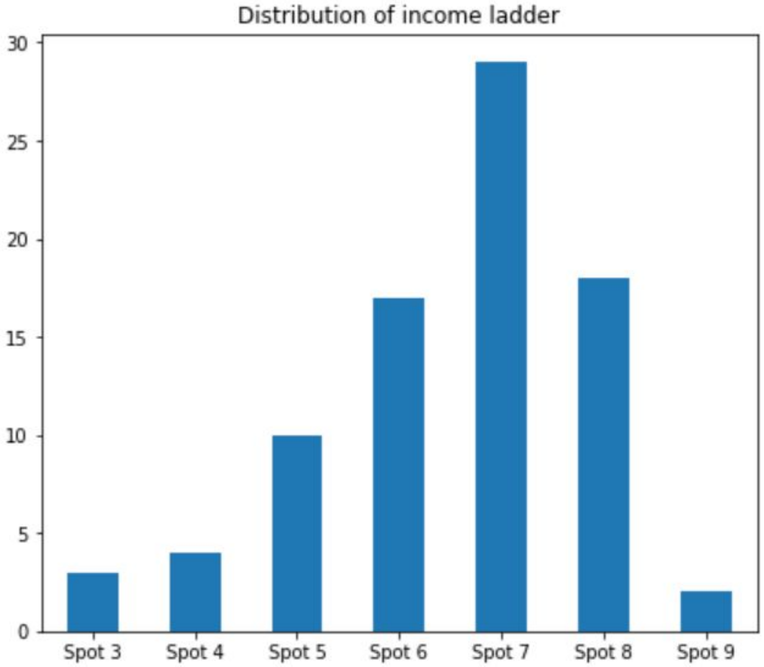


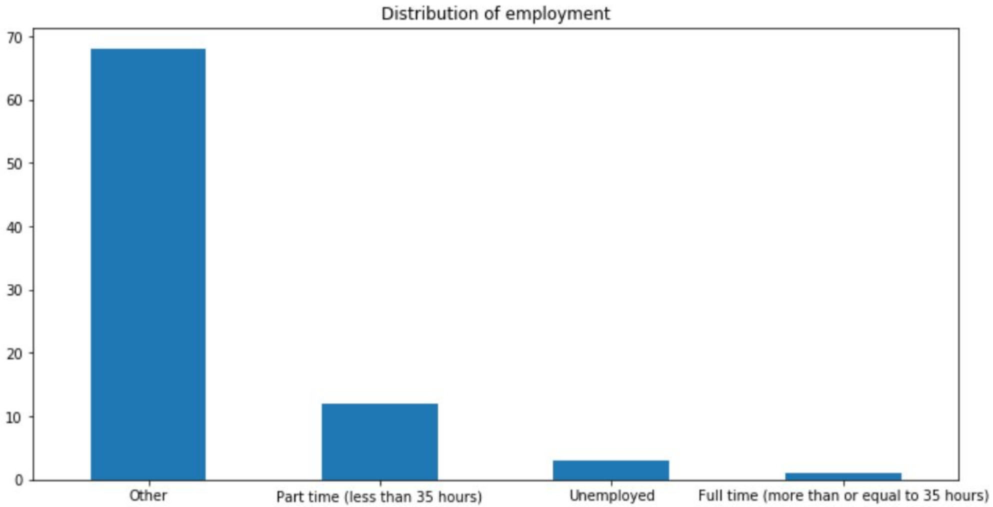


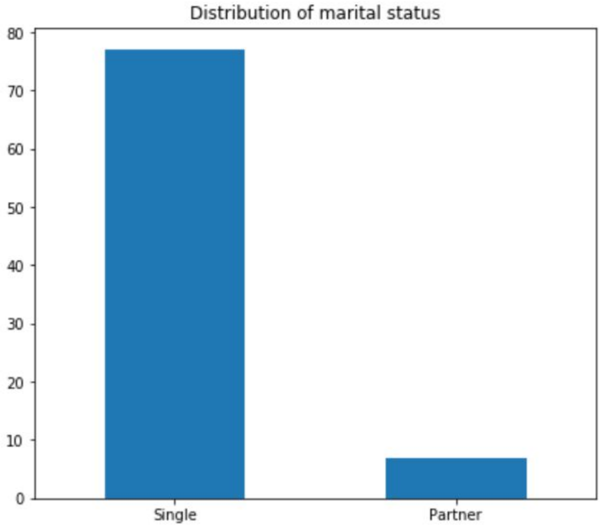
- Majority of the participants are female
- Most of the participants age range from 18 to 21



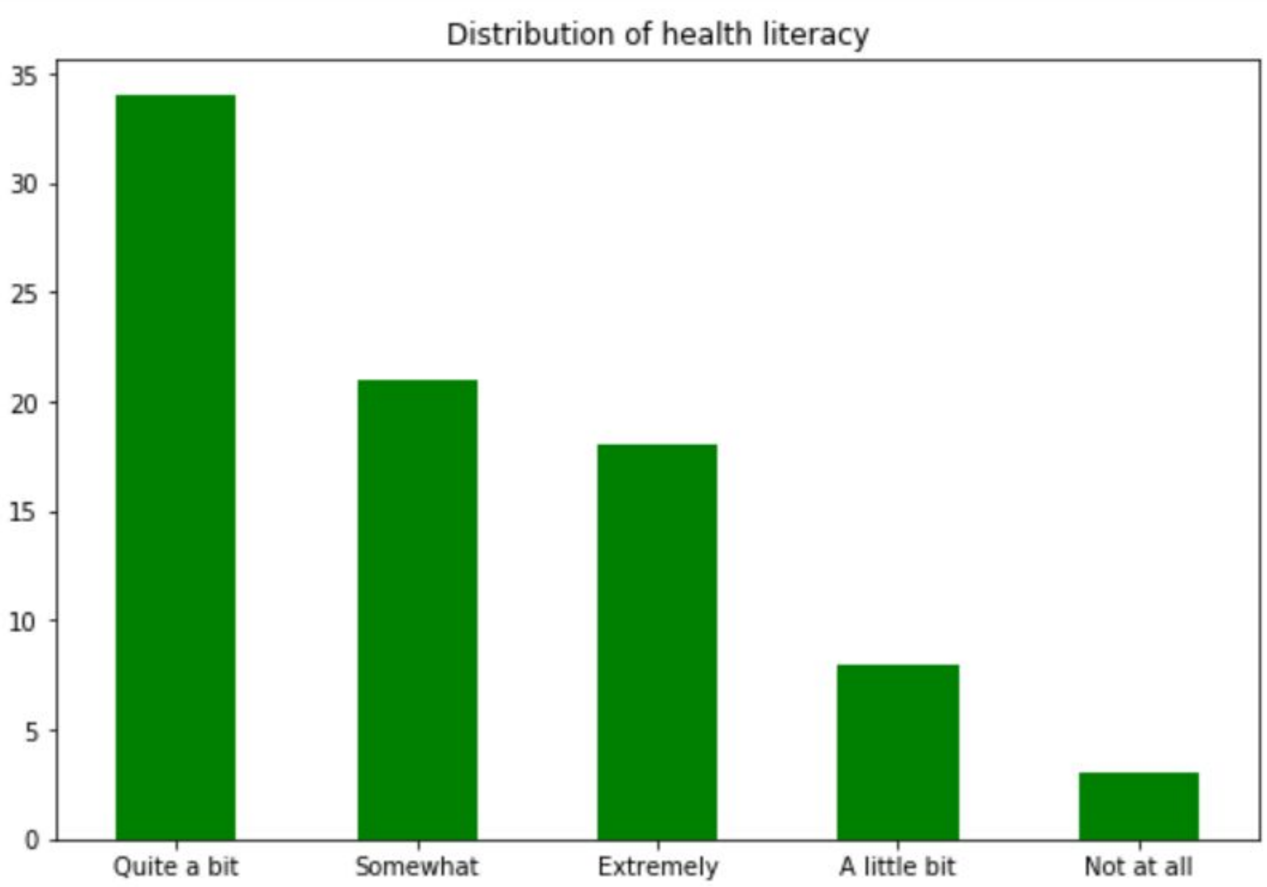
- Majority of the participants are from college or technical school and are high school graduate or has a “GED” degree



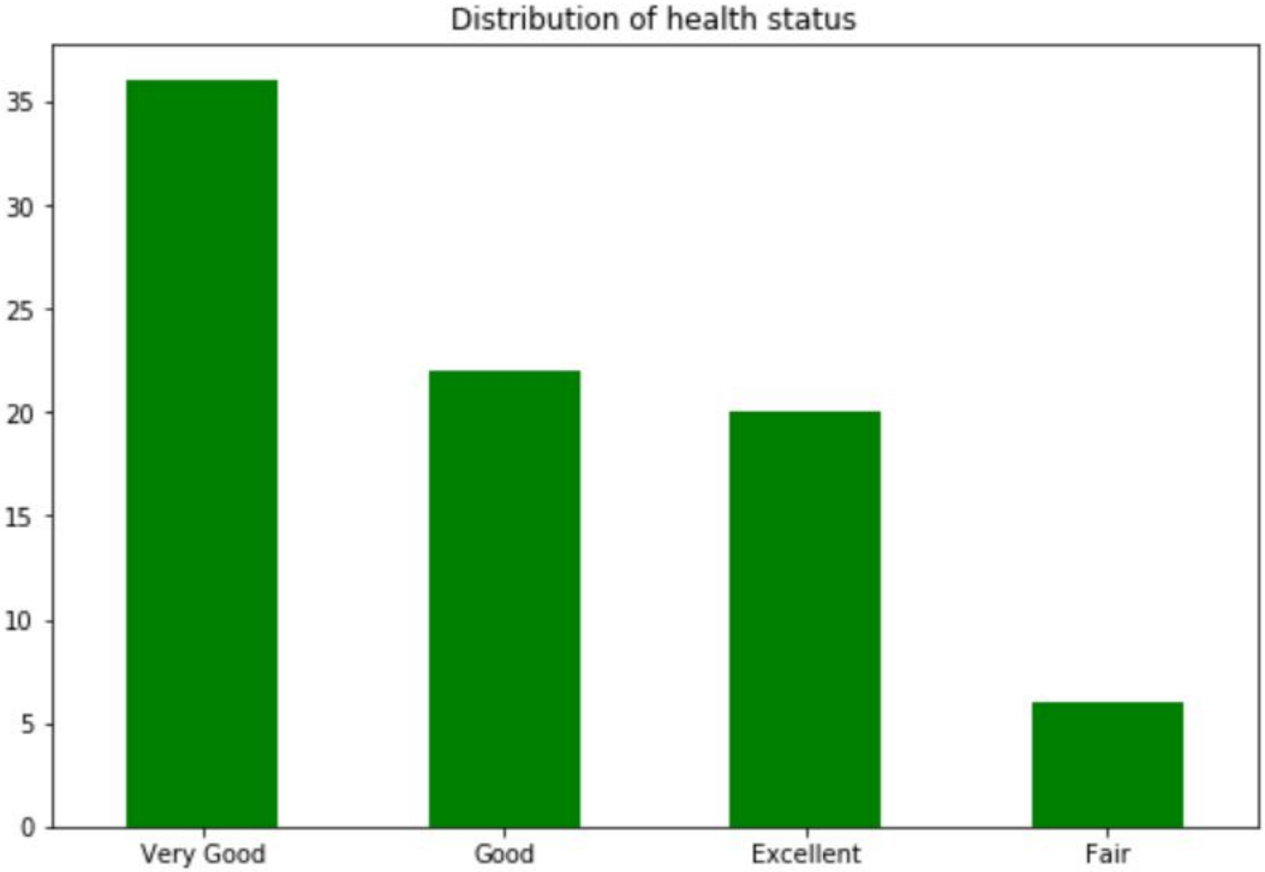




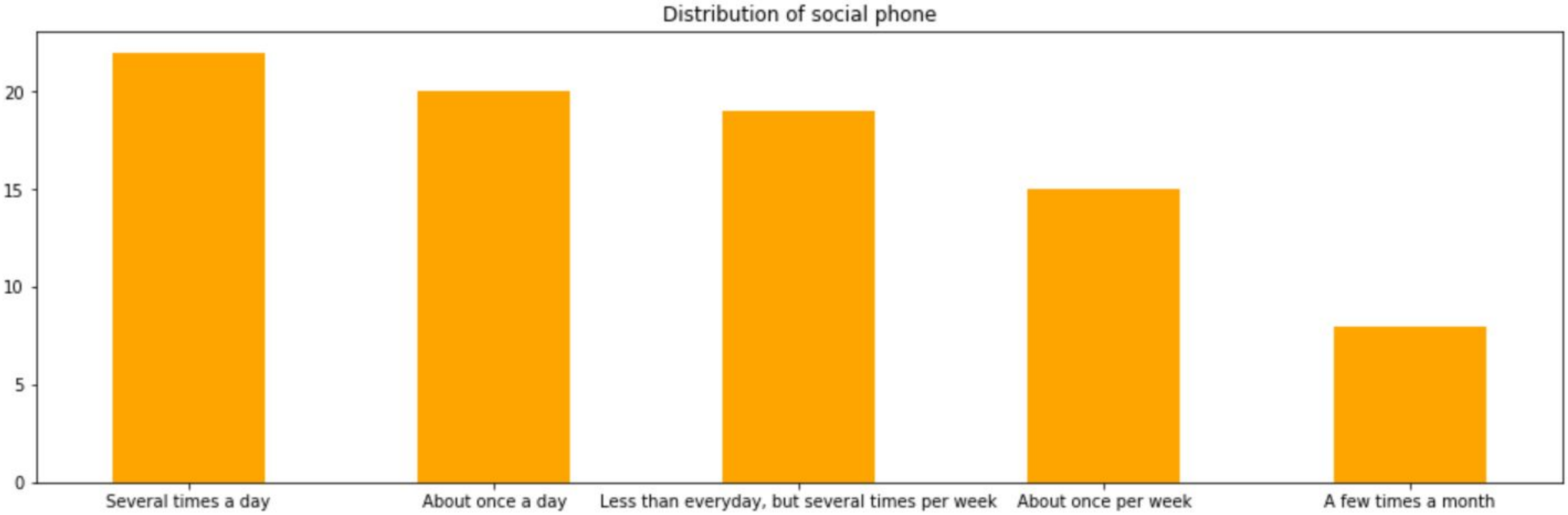
with 10 being the most
well off for income ladder



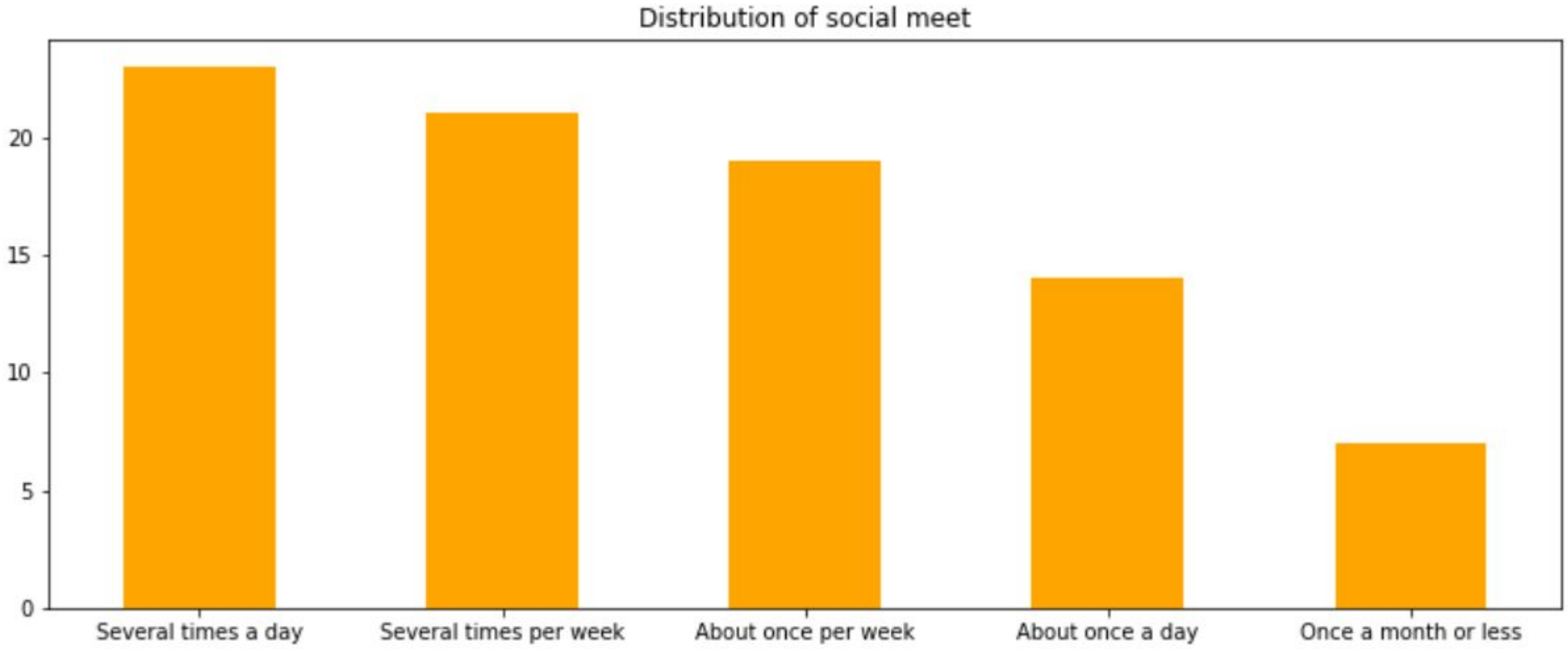
- Most participants are quite comfortable with filling medical forms by themselves



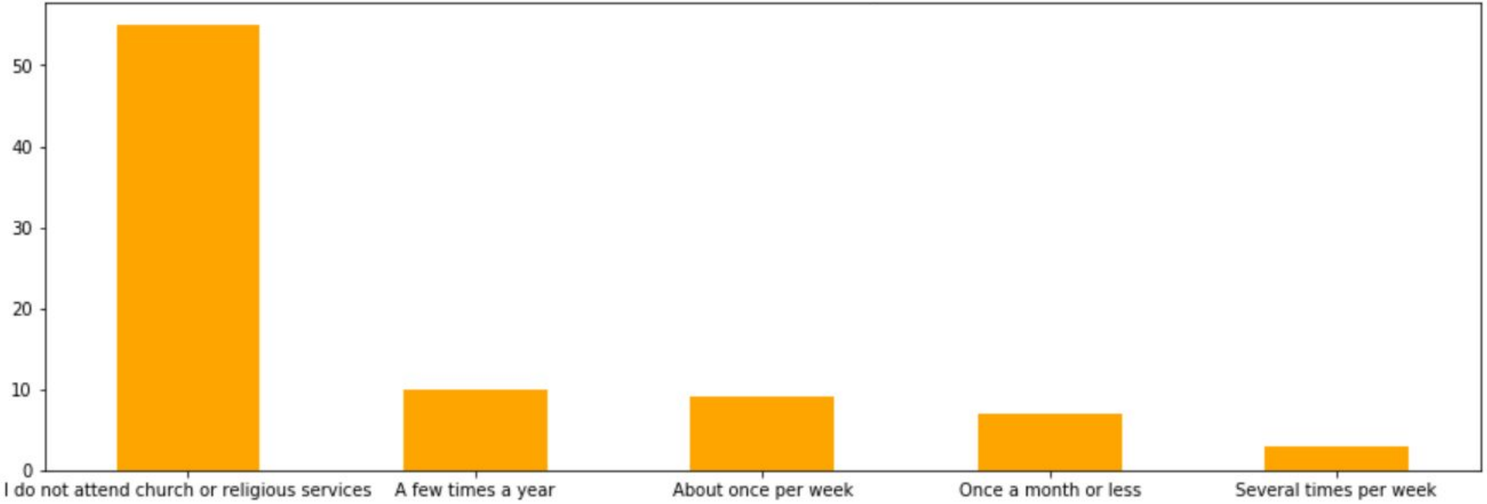
- Majority of participants rate their health status as good



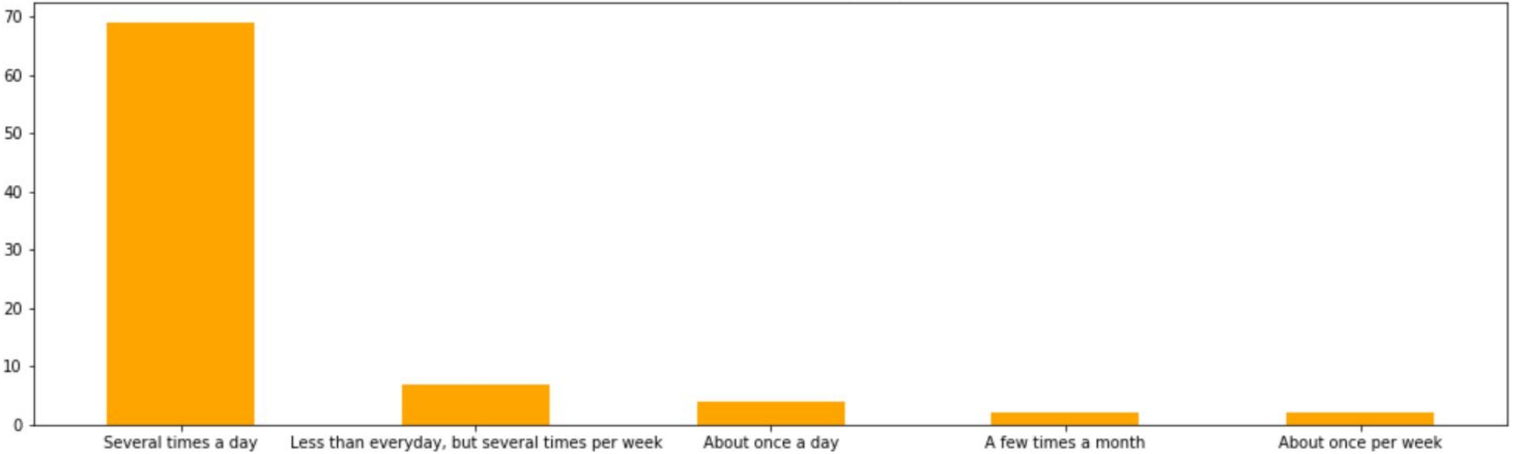
- Distribution of participants talking on their phone with family, friends, or neighbors is rather similar with only a minority talk only a few times a month



- Majority of participants meet together with their friends or relatives at least once a week. Only a minority meet once a month or less

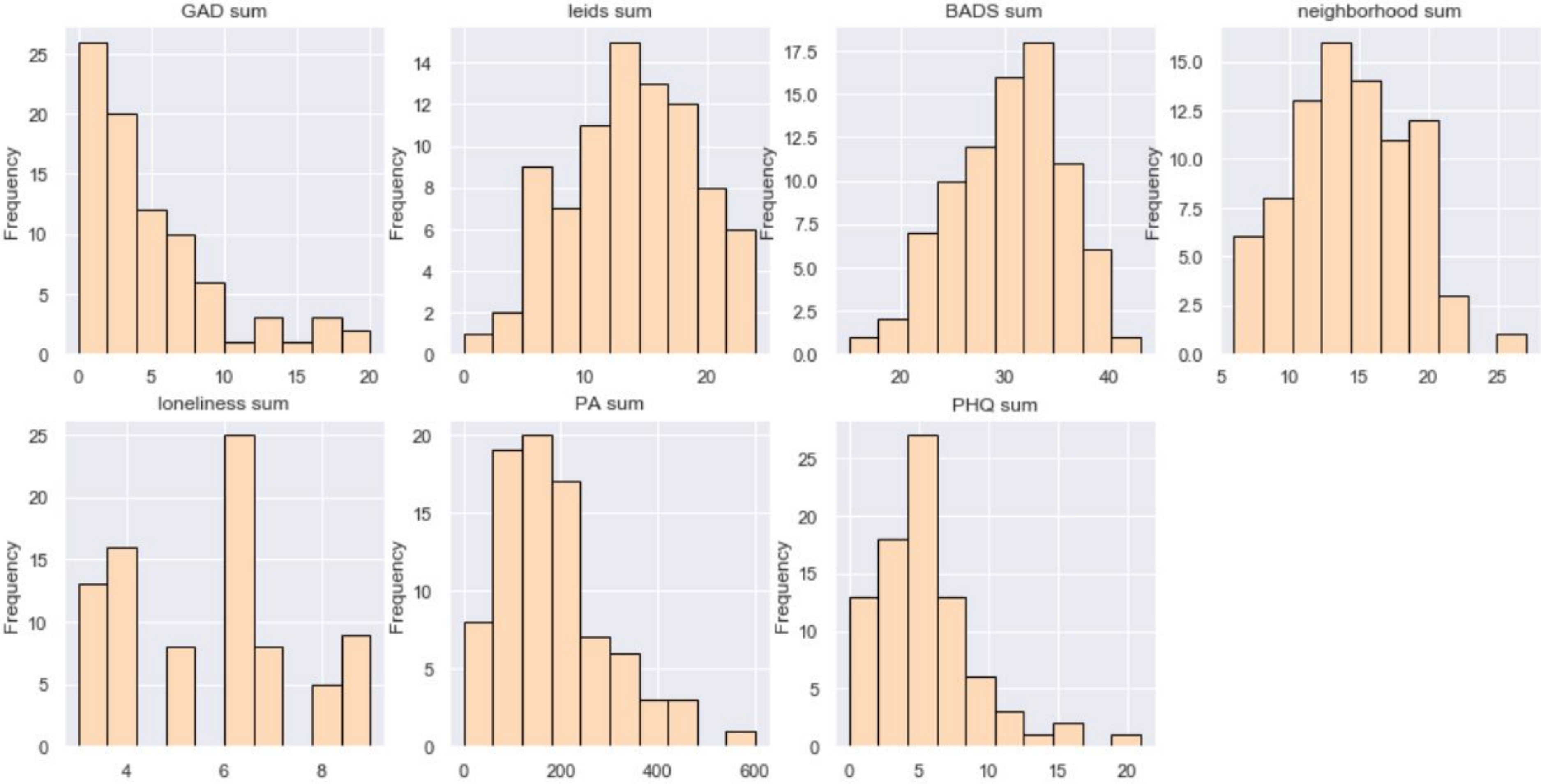


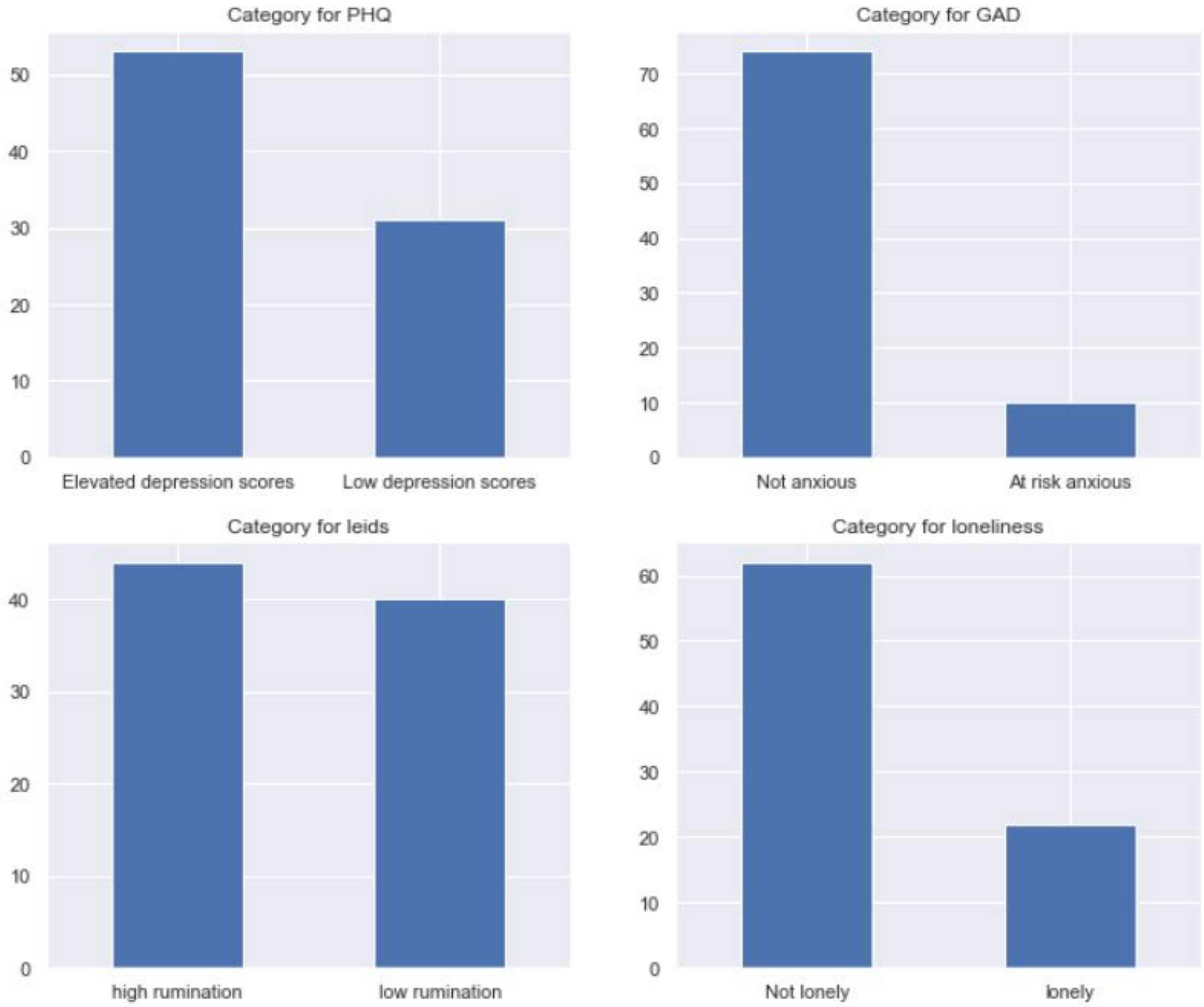
Distribution of social religion



Distribution of text frequency

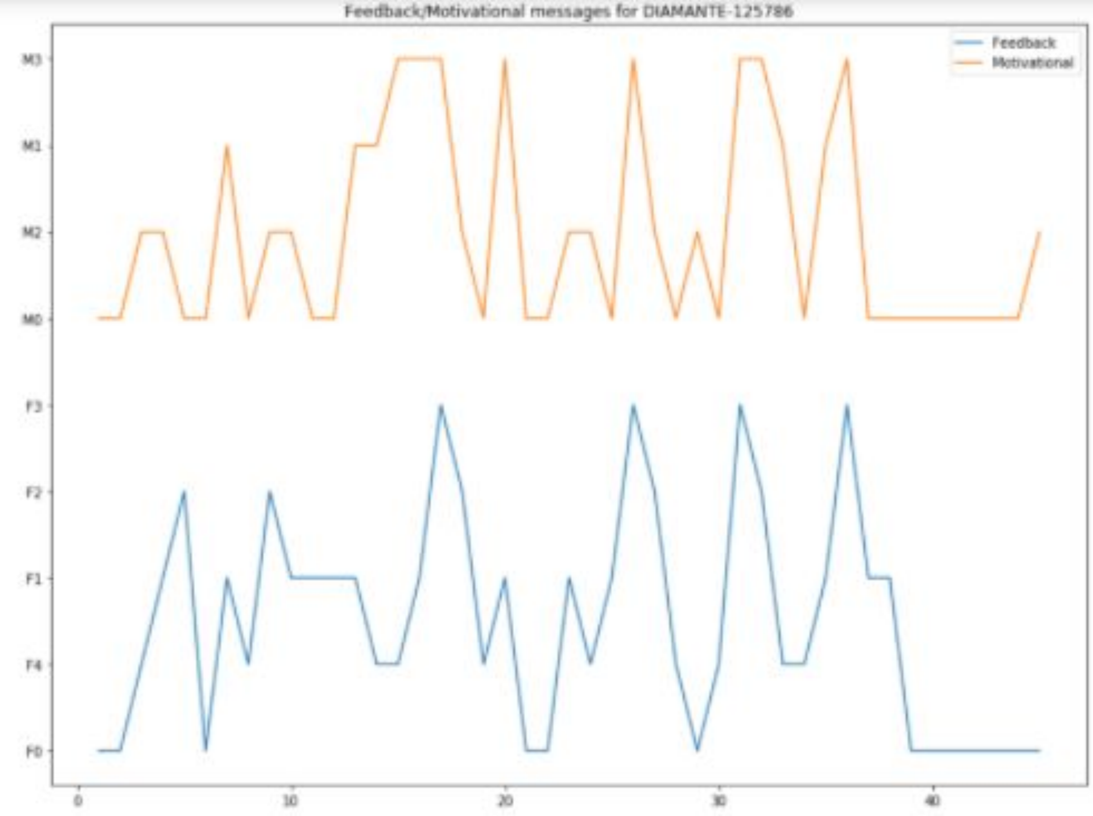
Majority of participants do not attend religious services and text at least several times a day


GAD sum
Frequency
0
5
10
15
20
25
0
5
10
15
20
leids sum
Frequency
0
2
4
6
8
10
12
14
0
10
20
BADS sum
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20
30
40
neighborhood sum
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
5
10
15
20
25
loneliness sum
Frequency
0
5
10
15
20
25
4
6
8
PA sum
Frequency
0
5
10
15
20
0
200
400
600
PHQ sum
Frequency
0
5
10
15
20
25
0
5
10
15
20



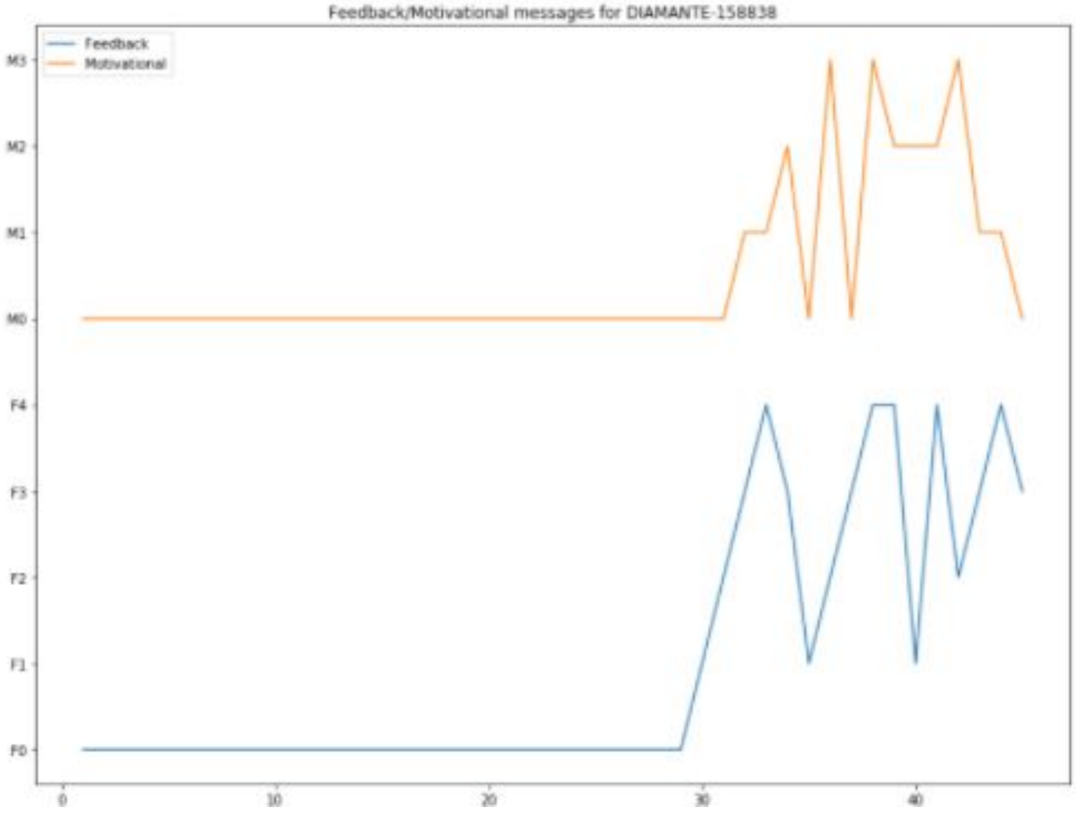
- **Majority of the students have an elevated depression score, do not have generalized anxiety disorder and are not lonely**
- **There is roughly equal number of students who has high and low rumination**

Feedback/motivational messages for each individual participants



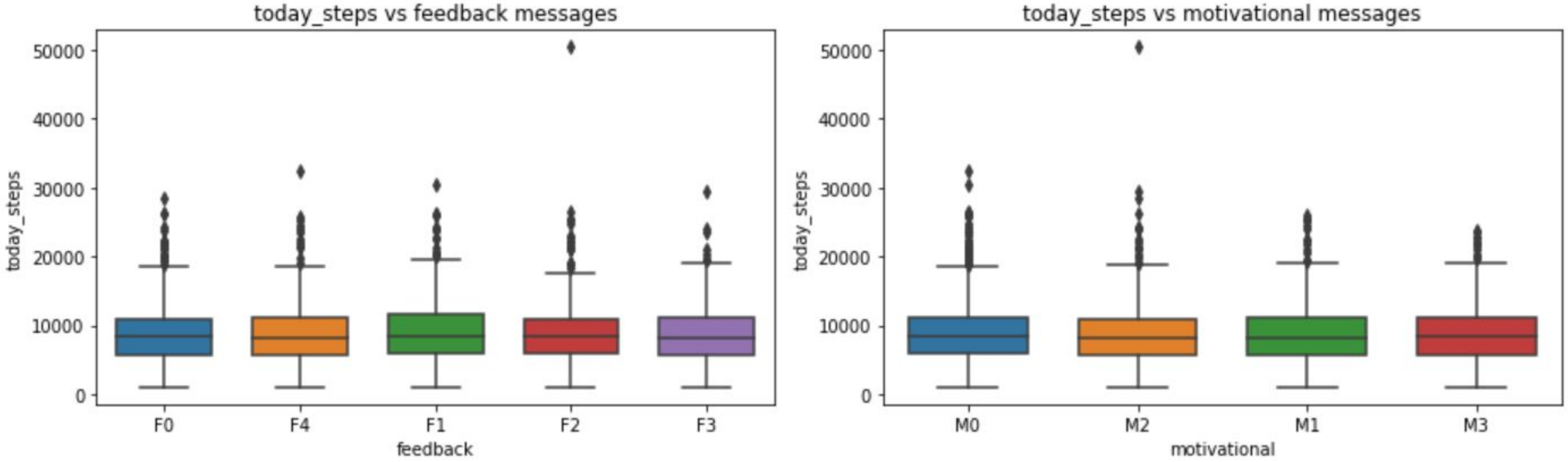
- Feedback and motivational messages for participant ID 125786
- daily step counts for the entire 45 days duration

Feedback/motivational messages for each individual participants



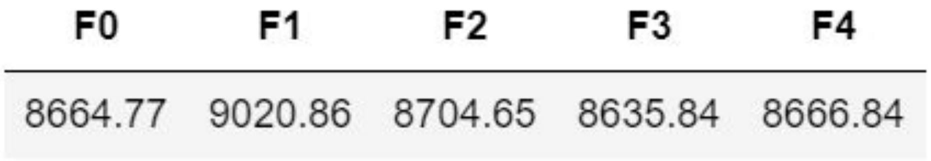
- Feedback and motivational messages for participant ID 158838
- The first close to 30 days this participant does not receive any feedback or motivational messages
- Does this mean the first part is a control?

Boxplots of today_steps vs feedback and motivational

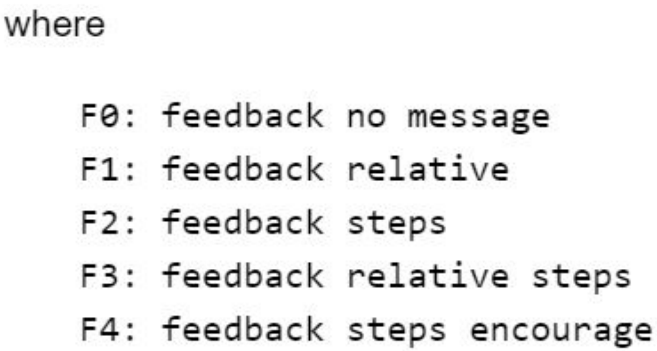


- **For these results, I removed rows with NA's for today_steps**

Table of average steps for each feedback messages for all participants



F0	F1	F2	F3	F4
8664.77	9020.86	8704.65	8635.84	8666.84

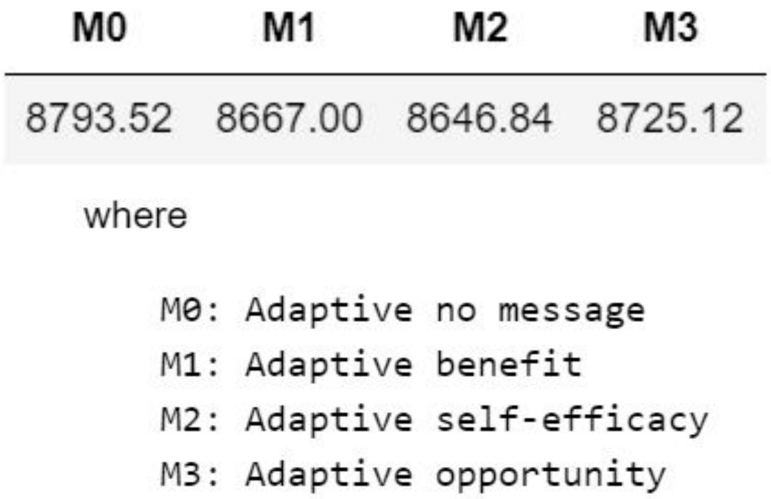


where

```
F0: feedback no message
F1: feedback relative
F2: feedback steps
F3: feedback relative steps
F4: feedback steps encourage
```

- **For these results, I removed rows with NA’s for today_steps**

Table of average steps for each motivational messages for all participants



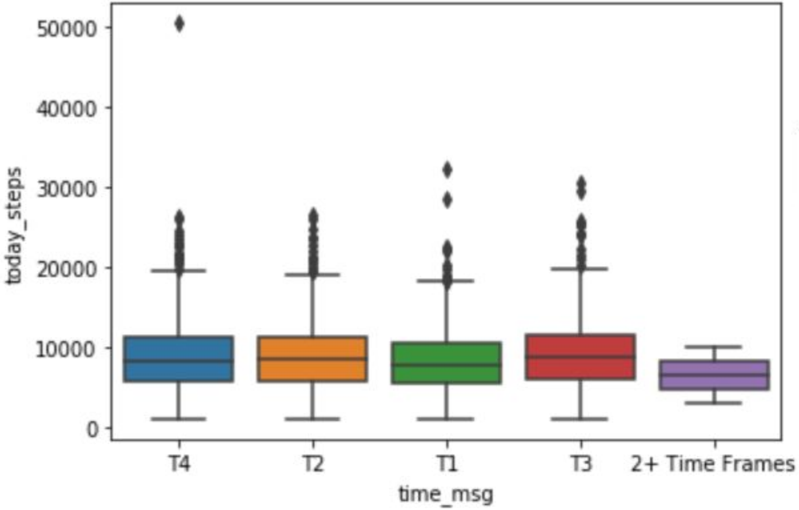
M0	M1	M2	M3
8793.52	8667.00	8646.84	8725.12

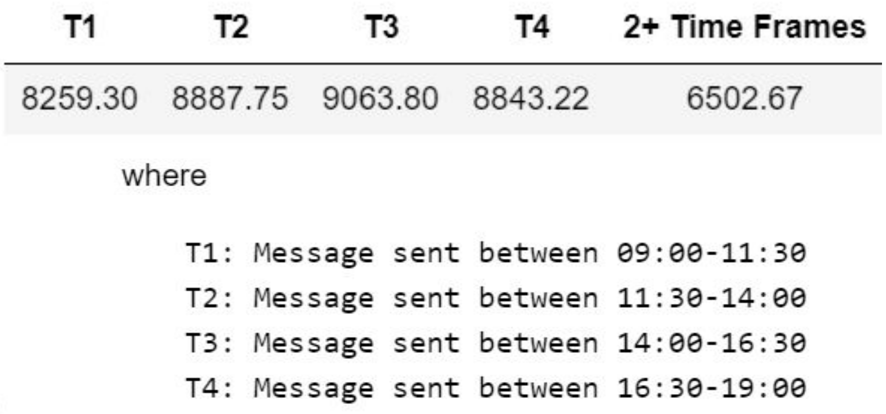
where

```
M0: Adaptive no message
M1: Adaptive benefit
M2: Adaptive self-efficacy
M3: Adaptive opportunity
```

- **For these results, I removed rows with NA's for today_steps**

Boxplot and table of average steps for each message timing for all participants





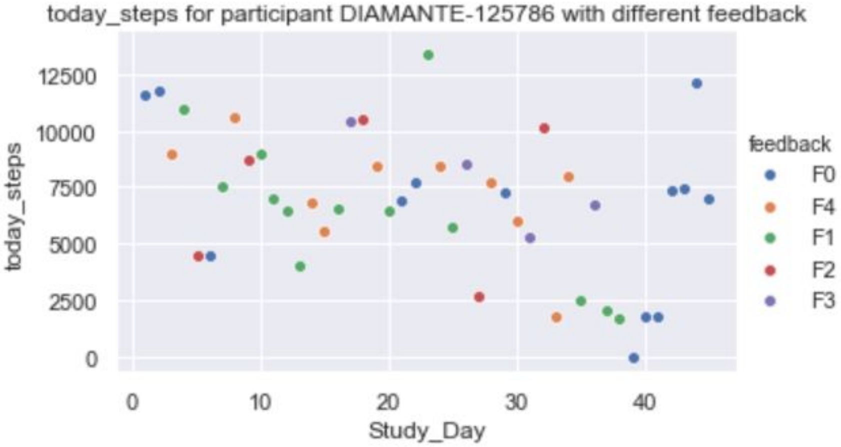
T1	T2	T3	T4	2+ Time Frames
8259.30	8887.75	9063.80	8843.22	6502.67

where

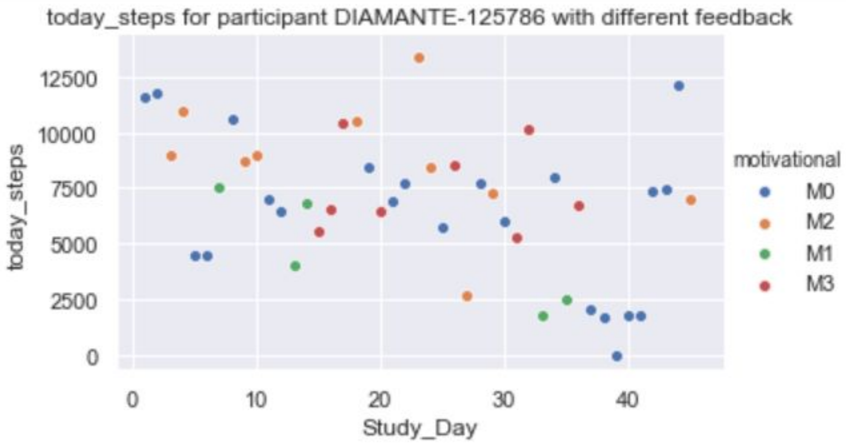
```
T1: Message sent between 09:00-11:30
T2: Message sent between 11:30-14:00
T3: Message sent between 14:00-16:30
T4: Message sent between 16:30-19:00
```

- **For these results, I removed rows with NA’s for today_steps**

today steps vs study day for different feedback/motivational messages

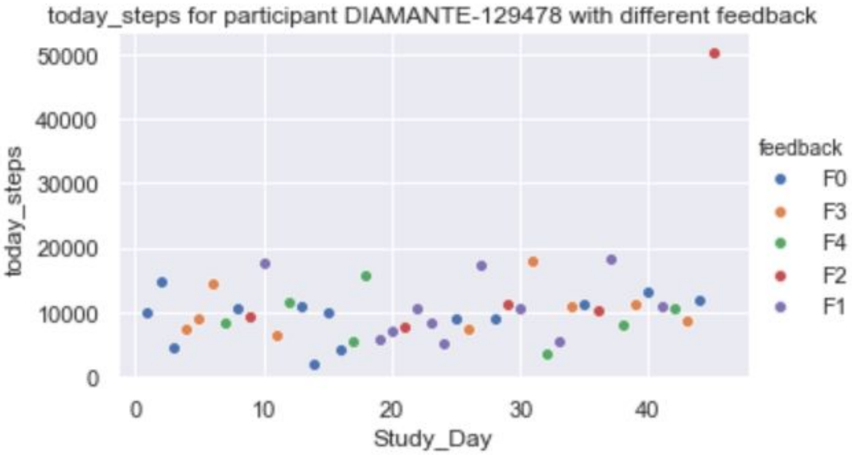


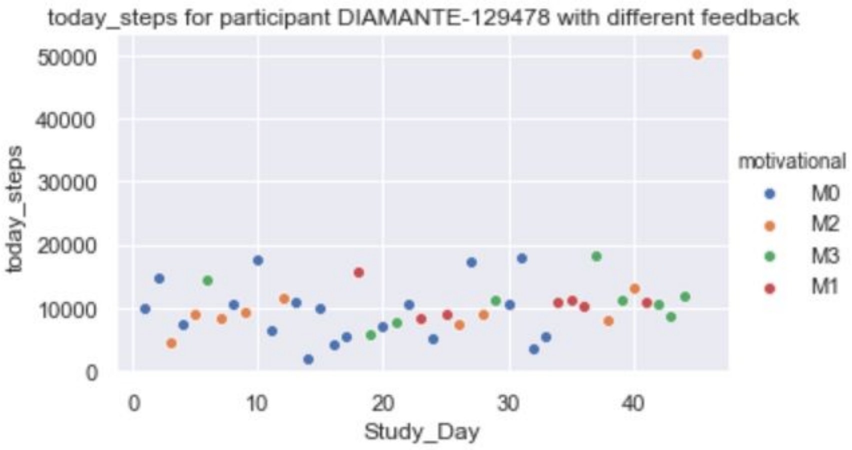
today_steps for participant DIAMANTE-125786 with different feedback



today_steps for participant DIAMANTE-125786 with different feedback

today steps vs study day for different feedback/motivational messages





Feature engineering

Feature Engineering

Imputation

- Drop rows or columns with missing values
- Or can use an threshold for dropping (for e.g. drop rows which have mean missing values with higher than a certain threshold)

Numerical Imputation

- Preferred choice over just dropping as it preserves data size
- Replace NA's with 0's as long as it make sense
- Could replace NA's with mean of columns, median or most frequent value

Feature Engineering

Log transformation

- Allows skewed data to become approximately normal and makes statistical analysis more valid

One-hot encoding

- Converts categorical data into multiple binary columns of 1’s or 0’s

Feature Engineering

Normalization

- Scales all values in a fixed range between 0 and 1

Standardization

- Transforms data to have mean 0 and standard deviation 1

Feature Engineering

Binning

- Group continuous numerical numbers into smaller number of bins
- Can be used to reduce the minor observation errors

Thank you!