WEEK 3 (16.1.2016 to 23.1.2017)

planned work:
1. identify subjects with at least two visits.
2. make a list of variables defining the unhealthy environment
3. identify subjects who have all the data for those variables
4.make a table of the descriptive statistics for the chosen list of variables
5. start identifying subjects persistently exposed to unhealthy environment by constructing a diet score:
a) commonly used score called healthy diet score
b) commonly used score called Nordic nutrition score
c) combination
*multiply the outcome with the weight association outcome

1. there are 111505 unique subject ids in the Multimodality/VIP_161102.csv, some of those are repeated more that once through the 168330 rows.
65937 subjects(65938 -1 for head), have only one visit, 34348 subjects have two visits, 11183 subjects have three and 37 subjects have four visits.
Those that have three and four will have two visits as well so I added them up and got **45568** subjects with at least two visits. Here you can see how I checked that, if you want to make sure it is correct:

```
[jerneja_m@purple Private]$ cut -d"," -f1 VIP_161102.csv | wc -l
168331
```

-1 for head=168330

```
[jerneja_m@purple Private]$ cut -d"," -f1 VIP_161102.csv | sort -n | uniq | wc -l
111506
```

-1 for head= 111505

```
[jerneja_m@purple Private]$ cat VIP_161102.csv | cut -d"," -f1 |sort -n |uniq -c | cut -d" " -f7 | sort -n | uniq -c
  65938 1
  34348 2
  11183 3
     37 4
```

I checked the time difference of these, to make sure the right visits will be identified. It turns out it is a little messy. I had checked the ordered years of the visits, then counted the number of subjects for each year, for each visit, then for all the visit combination(2 visits-1combination, 3 visits-3 combinations, 4 visits-6 combinations) I looked at the time difference and counted the number of subjects. If you want to repeat to check I did everything correct, run the python script(on git in the code/data_exploration) in the Multimodality folder and then run the below commands which use the output file of the script.

Years of the visits:

```
jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep "first visit:"|
sort -n | uniq -c
    143      first visit: 1985
    189      first visit: 1986
    155      first visit: 1987
    612      first visit: 1988
   1262      first visit: 1989
   2060      first visit: 1990
   3535      first visit: 1991
   4309      first visit: 1992
   4510      first visit: 1993
   3995      first visit: 1994
   4198      first visit: 1995
   3338      first visit: 1996
   3185      first visit: 1997
   3020      first visit: 1998
   2644      first visit: 1999
   1841      first visit: 2000
   1554      first visit: 2001
   1129      first visit: 2002
    992      first visit: 2003
   1150      first visit: 2004
   1254      first visit: 2005
    478      first visit: 2006
      5      first visit: 2007
      2      first visit: 2008
      1      first visit: 2009
      1      first visit: 2010
      3      first visit: 2011
      2      first visit: 2012
      1      first visit: 2014
[jerneja_m@purple Private]$
```

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep "second visit:"|
sort -n | uniq -c
      5      second visit: 1990
      1      second visit: 1991
      4      second visit: 1992
      7      second visit: 1993
      5      second visit: 1994
    117      second visit: 1995
    177      second visit: 1996
    157      second visit: 1997
    537      second visit: 1998
    456      second visit: 1999
   2244      second visit: 2000
   3147      second visit: 2001
   3911      second visit: 2002
   4088      second visit: 2003
   3710      second visit: 2004
   3920      second visit: 2005
   2880      second visit: 2006
   3228      second visit: 2007
   3023      second visit: 2008
```

```
   2738     second visit: 2009
   2379     second visit: 2010
   2030     second visit: 2011
   1569     second visit: 2012
   1505     second visit: 2013
   1589     second visit: 2014
   1591     second visit: 2015
    550     second visit: 2016
```

```
[jerneja_m@purple Private]$
```

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep "third visit:"|
sort -n | uniq -c
      1     third visit: 1999
      6     third visit: 2000
      4     third visit: 2001
      3     third visit: 2002
      3     third visit: 2003
      3     third visit: 2004
     68     third visit: 2005
     46     third visit: 2006
     58     third visit: 2007
    235     third visit: 2008
    228     third visit: 2009
   1179     third visit: 2010
   1635     third visit: 2011
   2030     third visit: 2012
   1883     third visit: 2013
   1783     third visit: 2014
   1688     third visit: 2015
    367     third visit: 2016
```

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep "fourth visit:"|
sort -n | uniq -c
      1     fourth visit: 2007
      1     fourth visit: 2009
      2     fourth visit: 2010
      3     fourth visit: 2011
      1     fourth visit: 2012
      1     fourth visit: 2013
      1     fourth visit: 2014
     22     fourth visit: 2015
      5     fourth visit: 2016
```

So, the first visit can also be quite late and sometimes this will result in little time difference, sometimes the first and third, or first and fourth, or second and third etc. will have the time difference of around 10 years we want, but they might be shifted in time and I am not sure if we want to allow that?
Here are all the differences in the visits:

First two visits:

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the first two" | cut -d":" -f2 | sort -n | uniq -c
     13  0
     11  1
      3  2
     13  3
     15  4
```

```
     16  5
     17  6
     28  7
    108  8
   1884  9
  38984  10
   1376  11
     50  12
      7  13
     10  14
      6  15
      6  16
      2  17
     10  18
    256  19
   2530  20
    214  21
      4  22
      1  25
      1  26
      3  30
[jerneja_m@purple Private]$
```

} Majority is around 10 years difference

Second two visits(difference between the second visit and the third):

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the second two" | cut -d":" -f2 | sort -n | unig -c
      7  0
      1  1
      3  2
      5  3
      2  4
      2  5
      3  6
      3  7
      8  8
    240  9
  10752  10
    181  11
      3  12
      1  13
      8  20
      1  21
[jerneja_m@purple Private]$
```

} Majority is around 10 years difference

Third two visits(difference between the third visit and the fourth, with 31+1 candidates:

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the third two" | cut -d":" -f2 | sort -n | unig -c
      4  0
      1  1
      1  9
```

```
     31  10
[jerneja_m@purple Private]$
```

Difference between the first visit and third, with 1+14+5 candidates:

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the first and third" | cut -d":" -f2 | sort -n | uniq -c
      1  9
     14  10
      5  11
      4  12
      6  13
      9  14
      3  15
      7  16
      5  17
     36  18
    790  19
   9739  20
    575  21
      9  22
      1  23
      1  25
      1  26
     14  30
[jerneja_m@purple Private]$
```

Difference between the second visit and fourth, with 4+1 candidates:

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the second and fourth" | cut -d":" -f2 | sort -n | uniq -c
      4  10
      1  11
      1  13
      2  19
     29  20
[jerneja_m@purple Private]$
```

Difference between the first visit and fourth, without any candidates:

```
[jerneja_m@purple Private]$ cat temporary_file_checking_visits_time | grep
"between the first and fourth" | cut -d":" -f2 | sort -n | uniq -c
      6  20
      3  21
      1  22
      1  25
     26  30
[jerneja_m@purple Private]$
```

So here we need to discuss whether we want to be safe and take just those subjects who had their first visit in a certain period around [1985,**?**] and second visit 10 +-**?** years later.
Or take any two visits as long as they have 10 +-**?** years in between?
And then I will save a list of enummers for the visit pairs we want.

2. I started putting together a list of variables used to identify the lean phenotype in unhealthy environment. I checked if any of the described variables are missing in the dataset and other way around. In the described variables there are Date of diagnosis, fasta_prov and l_v_uppskattad, which are not in the dataset. For fasta_prov, it says it is for when fasta_enk has a missing value and for the last one it says, that if it is missing the weight is measured at baseline, instead of using a self reported weight.
In the dataset there are these variables, which dont have the description. Most of them look like diet variables and that makes sense since there are no diet variables description in the variables description list. But I am not sure for all of them and are we going to include all the diet variables?

besok
agr10
BMR
enkver
enkver2
antfrag
year
exclude
missport
missproc
FIL
potport
kottport
gronport
da01-da84
dat01-DAT66
gramlong1-gramlong84
gramshort1-gramshort64
ensum1
protsum1
protsum1_anim
protsum1_veg
kolhsum1
sacksum1
DISAsum1
MOSAsum1
fibesum1
FULLKsum1
alkosum1
fettsum1

mfetsum1
MONOsum1
POLYsum1
TRANSsum1
kolesum1
FA140_sum1
FA160_sum1
FA182_sum1
FA183_sum1
FA204_sum1
FA205_sum1
FA226_sum1
FA150_sum1
FA170_sum1
MAGNsum1
NATRsum1
FOSFsum1
NIACsum1
selesum1
ZINCsum1
retisum1
karosum1
TIAMsum1
Folasum1
B2sum1
B6sum1
B12sum1
askosum1
Dsum1
tokosum1
VITKsum1
jernsum1
JODIsum1
kalcsum1
KALIsum1
l2a1
l2a2
l2a3
l2a4
l2a5
l2a6
l2b1
l2b2
l2b3
l2b4
l2b5
l2b6
l1
l2

l3
l5a
l5b
l5c
kostdata