IDENTIFYING SUBJECTS WITH 2+ VISITS IN THE VIP DATASET

Code for all in the bottom.

1.Number of subjects:

A. Based on repetition of the Subject_id. Out of 168330 entries, there are 111505 unique subject ids, out which some are repeated once, twice or four times:

	Number of subjects
Occurring once	65937
Occurring twice	34348
Occurring three times	11183
Occurring four times	37

B. Based on the variable besok, the value can be 1 for the first visit, 2 for second and so on, but a lot of subjects have the value for besok missing:

	Number of subjects
besok==1	100835 (100835-38506=62329)
besok==2	38506 (38506-5126=33380)
besok==3	5126 (5126-3=5123)
besok==4	3
besok is missing	23858*

> *this number is from bash, when looking with R and counting the number of is.na() there are two less, so 23856. Another weird thing is that with bash, I get 2 occurrences(subject id 78303 and 79138) with the value 8888, but that is just because I cut on the 228th column which is the besok column for all subjects except these two. Might be that some lines are actually longer or shorter, since there are 605 variables and with the header 168331 rows, there should be 168331*604=101671924 commas in the file, but I get 101671929.
> But when checking the length of rows in R they are all 605 long.
> Within the missing variable besok, there are 23621 subjects with only one occurrence, 116 with two occurrences, 1 with three and none with four.

2. The distribution of the years for the first, second, third and fourth visit:

A. Based on the repetition of Subject_id:

year	Number of subjects having the FIRST visit in year
1985	143
1986	189
1987	155
1988	612
1989	1262
1990	2060
1991	3535
1992	4309
1993	4510
1994	3995
1995	4198
1996	3338
1997	3185
1998	3020
1999	2644
2000	1841
2001	1554
2002	1129
2003	992
2004	1150
2005	1254
2006	478
2007	5
2008	2
2009	1
2010	1
2011	3
2012	2
2014	1

year	Number of subjects having the SECOND visit in year
1990	5
1991	1
1992	4
1993	7
1994	5
1995	117
1996	177
1997	157
1998	537
1999	456
2000	2244
2001	3147
2002	3911
2003	4088
2004	3710
2005	3920
2006	2880
2007	3228
2008	3023
2009	2738
2010	2379
2011	2030
2012	1569
2013	1505
2014	1589
2015	1591
2016	550

year	Number of subjects having the THIRD visit in year
1999	1
2000	6
2001	4
2002	3
2003	3
2004	3
2005	68
2006	46
2007	58
2008	235
2009	228
2010	1179
2011	1635
2012	2030
2013	1883
2014	1783
2015	1688
2016	367

year	Number of subjects having the THIRD visit in year
2007	1
2009	1
2010	2
2011	3
2012	1
2013	1
2014	1
2015	22
2016	5

B. Based on the variable besok:

year	Number of subjects having besok == 1 in year
1990	3
1991	3843
1992	6737
1993	7290
1994	6575
1995	6828
1996	6265
1997	6050
1998	6223
1999	5584
2000	5500
2001	3567
2002	2185
2003	2292
2004	2551
2005	2862
2006	3740
2007	3558
2008	3326
2009	3495
2010	3657
2011	3178
2012	2964
2013	2545
2014	17

year	Number of subjects having besok == 2 in year
1992	3
1993	2
1994	2
1995	7
1996	4
1997	5
1998	3
1999	6
2000	25
2001	2274
2002	3910
2003	3975
2004	3636
2005	3960
2006	2916
2007	3267
2008	3190
2009	2829
2010	3097
2011	2311
2012	1570
2013	1506
2014	8

year	Number of subjects having besok == 3 in year
2001	1
2002	1
2003	3
2004	2
2005	7
2006	2
2007	4
2008	3
2009	5
2010	9
2011	1257
2012	2018
2013	1807
2014	7

year	Number of subjects having besok == 4 in year
2011	2
2013	1

year	Number of subjects having besok == NA in year
1985	270
1986	249
1987	251
1988	1042
1989	1988
1990	3680
1991	1887

1992	184
1993	191
1994	183
1999	1
2000	1
2009	1
2011	3
2012	3
2013	112
2014	6148
2015	6108
2016	1556

3. Difference in years between the visits:

A. Based on the repetition of the variable Subject_id:

Difference in years	Number of subjects having the FIRST and SECOND visit apart by
0	13
1	11
2	3
3	13
4	15
5	16
6	17
7	28
8	108
9	1884
10	38984
11	1376
12	50
13	7

14	10
15	6
16	6
17	2
18	10
19	256
20	2530
21	214
22	4
25	1
26	1
30	3

Difference in years	Number of subjects having the SECOND and THIRD visit apart by
0	7
1	1
2	3
3	5
4	2
5	2
6	3
7	3
8	8
9	240
10	10752
11	181
12	3
13	1
20	8
21	1

Difference in years	Number of subjects having the THIRD and FOURTH visit apart by
0	4
1	1
9	1
10	31

Difference in years	Number of subjects having the FIRST and THIRD visit apart by
9	1
10	14
11	5
12	4
13	6
14	9
15	3
16	7
17	5
18	36
19	790
20	9739
21	575
22	9
23	1
25	1
26	1
30	14

Difference in years	Number of subjects having the SECOND and FOURTH visit apart by
10	4
11	1
13	1
19	2
20	29

Difference in years	Number of subjects having the SECOND and FOURTH visit apart by
20	6
21	3
22	1
25	1
30	26

3. Based on the variable besok:

Difference in years	Number of subjects having the besok==1 and besok==2 apart by
0	14
1	6
2	3
3	10
4	11
5	14
6	16
7	26
8	102
9	1817
10	34234

11	703
12	26
13	4
14	3
16	3
17	1
18	5
19	103
20	1155
21	26

Difference in years	Number of subjects having the besok==2 and besok==3 apart by
0	5
2	2
3	3
4	2
5	1
6	1
7	1
8	2
9	68
10	4940
11	84
12	1

Difference in years	Number of subjects having the besok==3 and besok==4 apart by
0	2
10	1

Difference in years	Number of subjects having the besok==1 and besok==3 apart by
9	1
10	10
11	2
12	2
13	5
14	7
16	6
17	1
18	22
19	330
20	4561
21	100

Difference in years	Number of subjects having the besok==1 and besok==4 apart by
20	3

Difference in years	Number of subjects having the besok==2 and besok==4 apart by
10	2
19	1

4. Missing variable besok:
Looking into the subset of those that have the variable besok missing I noticed that a lot of other values are missing as well. There are 11600 subjects within the missing besok subset that seem to have the majority of diary data missing, like even the basic variables like year, exclude and kostdata. The total of missing value in a separate table, based on output of 4. below.
I can calculate the besok number for the missing besok values, based on the variable datum, but it might also be a good idea to ask the biobank whats up with these and why do they have so much of other variables values missing as well. Probably a lot of those subjects with the missing besok will be excluded from the analysis, since there are too many missing data for other variables.

1.A

In bash:

```
cat VIP_161102.csv | cut -d"," -f1 | sort -n | uniq -c | cut -d" " -f7| sort -n | uniq -c
```

In R:

```
VIP_data <- read.csv("VIP_161102.csv", header = TRUE, sep = ",", row.names = NULL, fill=TRUE)
Subject_id_occurrences <- aggregate(enummer~Subject_id,data=VIP_data,FUN=length
length(Subject_id_occurrences$enummer[Subject_id_occurrences$enummer==1,1])
length(Subject_id_occurrences$enummer[Subject_id_occurrences$enummer==2,1])
length(Subject_id_occurrences$enummer[Subject_id_occurrences$enummer==3,1])
length(Subject_id_occurrences$enummer[Subject_id_occurrences$enummer==4,1])
```

1.B

In bash:

```
cat VIP_161102.csv | cut -d"," -f228 | sort -n | uniq -c

*(cat VIP_161102.csv | grep -o "," | wc -l)
```

In R:

```
VIP_data <- read.csv("VIP_161102.csv", header = TRUE, sep = ",", row.names = NULL, fill=TRUE)
VIP_data_missing_besok<-VIP_data[is.na(VIP_data$besok),]
length(VIP_data_missing_besok[,1])
VIP_data_not_missing_besok<-VIP_data[!is.na(VIP_data$besok),]
length(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,1])
length(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,1])
length(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,1])
length(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,1])
```

*

```
repeated_subjects<-aggregate(enummer~Subject_id, data=VIP_data_missing_besok, FUN=length)
length(repeated_subjects[repeated_subjects[,2]==1,1])
length(repeated_subjects[repeated_subjects[,2]==2,1])
length(repeated_subjects[repeated_subjects[,2]==3,1])
length(repeated_subjects[repeated_subjects[,2]==4,1])
```

2.A done in Python

2.B

In bash:

```
cat VIP_161102.csv | cut -d"," -f3,228 | grep ",1" | cut -d"," -f1| cut -d"/" -f3| sort -n| uniq -c
cat VIP_161102.csv | cut -d"," -f3,228 | grep ",2" | cut -d"," -f1| cut -d"/" -f3| sort -n| uniq -c
cat VIP_161102.csv | cut -d"," -f3,228 | grep ",3" | cut -d"," -f1| cut -d"/" -f3| sort -n| uniq -c
cat VIP_161102.csv | cut -d"," -f3,228 | grep ",4" | cut -d"," -f1| cut -d"/" -f3| sort -n| uniq -c
cat VIP_161102.csv | cut -d"," -f3,228 | grep -v ",1" | grep -v ",2" | grep -v ",3" | grep -v ",8888" | grep -v ",4" | cut -d"," -f1| cut -d"/" -f3| sort -n| uniq -c
```

In R:

```
occurences<-aggregate(1:length(VIP_data_missing_besok$datum)~substr(VIP_data_missing_besok$datum,7,10),FUN=length)
colnames(occurences)<-c("year","number of subjects")
occurences

occurences<-
aggregate(1:length(VIP_data_not_missing_besok$datum[VIP_data_not_missing_besok$besok==1])~substr(VIP_data_not_missing_
besok$datum[VIP_data_not_missing_besok$besok==1],7,10),FUN=length)
colnames(occurences)<-c("year","number of subjects")
occurences

occurences<-
aggregate(1:length(VIP_data_not_missing_besok$datum[VIP_data_not_missing_besok$besok==2])~substr(VIP_data_not_missing_
besok$datum[VIP_data_not_missing_besok$besok==2],7,10),FUN=length)
colnames(occurences)<-c("year","number of subjects")
occurences

occurences<-
aggregate(1:length(VIP_data_not_missing_besok$datum[VIP_data_not_missing_besok$besok==3])~substr(VIP_data_not_missing_
besok$datum[VIP_data_not_missing_besok$besok==3],7,10),FUN=length)
colnames(occurences)<-c("year","number of subjects")
occurences

occurences<-
aggregate(1:length(VIP_data_not_missing_besok$datum[VIP_data_not_missing_besok$besok==4])~substr(VIP_data_not_missing_
besok$datum[VIP_data_not_missing_besok$besok==4],7,10),FUN=length)
colnames(occurences)<-c("year","number of subjects")
occurences
```

3.A done in Python

3.B

```
#1-2
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences



#2-3
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences


#3-4
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences


#1-3
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==3,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences


#1-4
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==1,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences
```

```
#2-4
year_differences<-as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,3],7,10))-
                as.numeric(substr(merge(VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==2,c(1,3)],

        VIP_data_not_missing_besok[VIP_data_not_missing_besok$besok==4,c(1,3)],by="Subject_id")[,2],7,10))
count_year_differences<-aggregate(1:length(year_differences)~year_differences,FUN=length)
count_year_differences
```

4.

```
VIP_data <- read.csv("VIP_161102.csv", header = TRUE, sep = ",", row.names = NULL, fill=TRUE)

VIP_data_missing_besok<-VIP_data[is.na(VIP_data$besok),]

VIP_data_missing_besok_copy<-VIP_data_missing_besok

for (variable in c(4,6:length(colnames(VIP_data_missing_besok)))) {

        missing_values=length(VIP_data_missing_besok[is.na(VIP_data_missing_besok[,variable]),1])
        non_missing_VIP_data_missing_besok=VIP_data_missing_besok[!is.na(VIP_data_missing_besok[,variable]),variable]

missing_values=missing_values+length(non_missing_VIP_data_missing_besok[non_missing_VIP_data_missing_besok=='5555'])+
                        length(non_missing_VIP_data_missing_besok[non_missing_VIP_data_missing_besok=='6666'])+
                        length(non_missing_VIP_data_missing_besok[non_missing_VIP_data_missing_besok=='7777'])+
                        length(non_missing_VIP_data_missing_besok[non_missing_VIP_data_missing_besok=='8888'])+
                        length(non_missing_VIP_data_missing_besok[non_missing_VIP_data_missing_besok=='9999'])
        message(paste(colnames(VIP_data_missing_besok)[variable],":",missing_values, "missing values "))

}
```