Load tidyverse libraries.

```
library(ggplot2)
library(tidyr)
library(dplyr)
```

Read in data, which is just the Excel sheet saved as a CSV.

```
data <- read.csv("Data624_project1_data.csv",header=TRUE)
```

Remove last rows where all fields are blank.

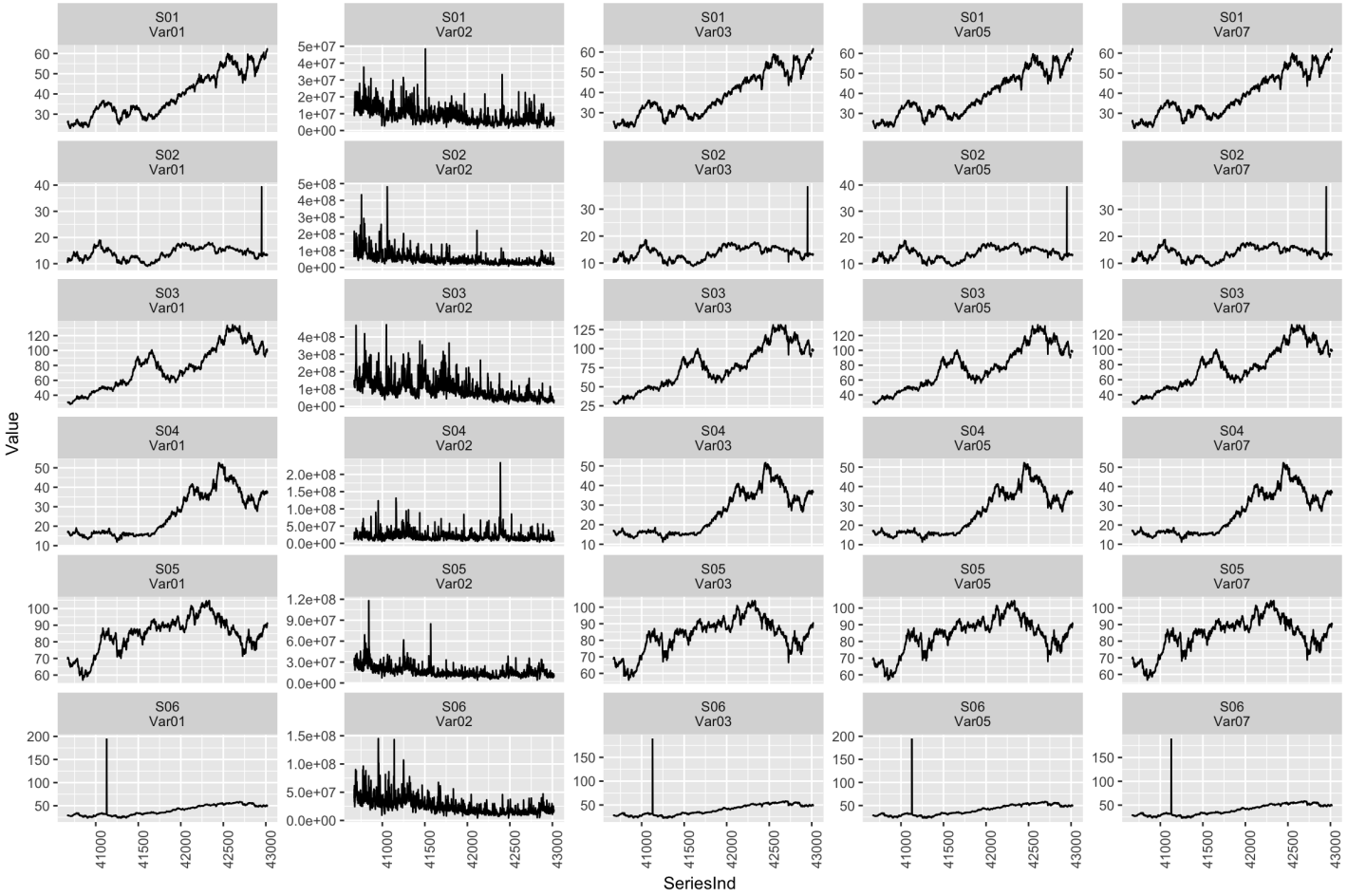
```
data <- data[1:9732,]
```

Convert to long format.

```
data_gathered <- gather(data,
    key="Variable",
    value="Value",
    -SeriesInd,-group)
data_gathered <- data.frame(data_gathered,
    Group.plus.var = paste0(data_gathered$group,"\n",data_gathered$Variable),
    stringsAsFactors=FALSE)
data_gathered$Group.plus.var <- factor(data_gathered$Group.plus.var,
    levels=paste0(rep(paste0("S0",1:6),each=5),"\n",rep(c("Var01","Var02","Var03","Var0
5","Var07"),times=6)))
```

Make line plots.

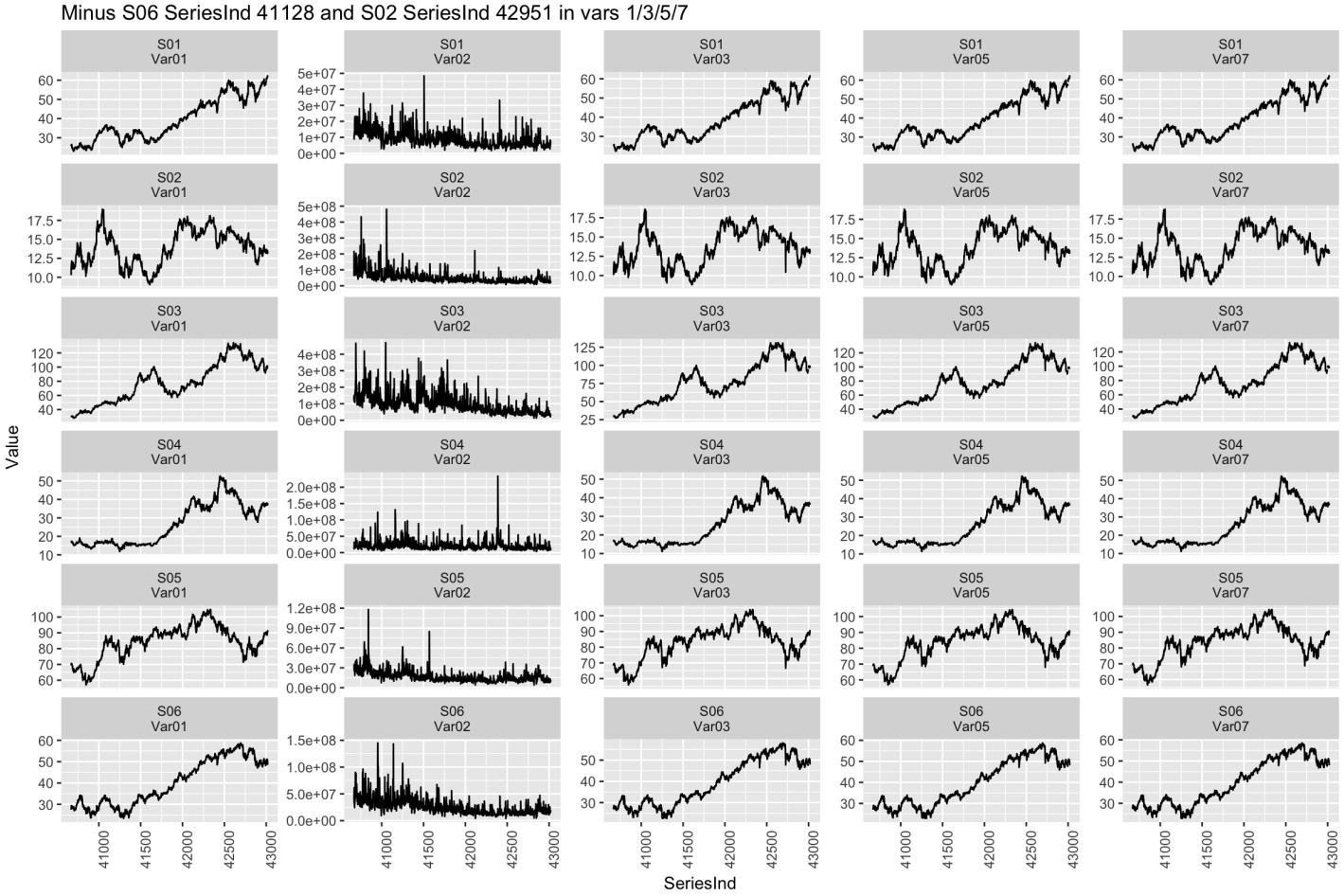
```
ggplot(data_gathered,
    aes(SeriesInd,Value)) +
    geom_line() +
    facet_wrap(~Group.plus.var,scales="free_y",nrow=6,ncol=5) +
    theme(axis.text.x=element_text(angle=90, hjust=1))
```



Plot minus S06 consistent outliers (SeriesInd 41128, variables 1,3, 5 and 7) and minus S02 consistent outliers (SeriesInd 42951, variables 1, 3, 5, and 7).

```
extreme_outliers <- which(data_gathered$SeriesInd == 41128 & data_gathered$group == "S0
6" & data_gathered$Variable != "Var02")
extreme_outliers <- c(extreme_outliers,
    which(data_gathered$SeriesInd == 42951 & data_gathered$group == "S02" & data_gathere
d$Variable != "Var02"))
data_gathered_minus_outliers <- data_gathered[setdiff(1:nrow(data_gathered),extreme_outl
iers),]
```

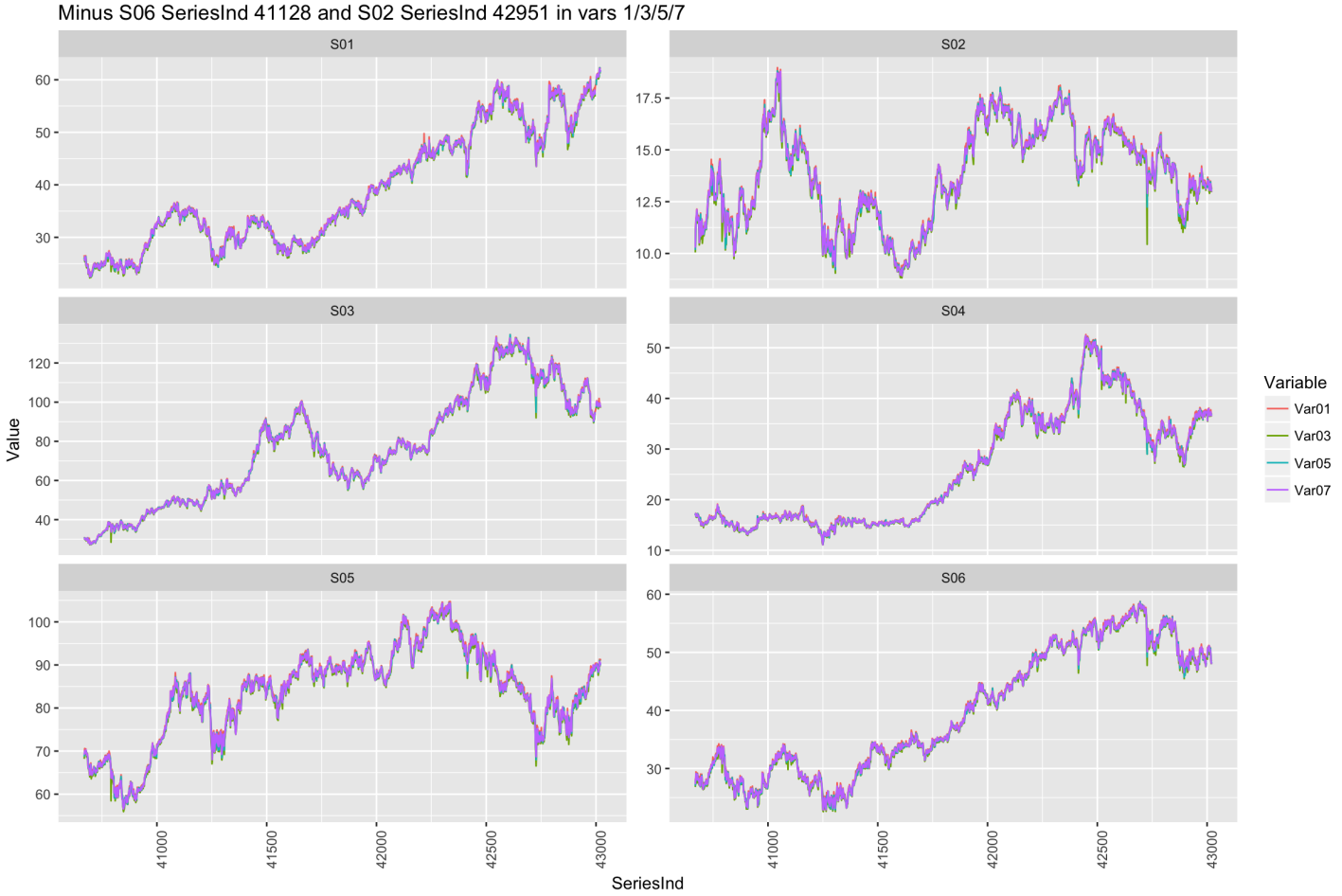
```
ggplot(data_gathered_minus_outliers,
    aes(SeriesInd,Value)) +
    geom_line() +
    facet_wrap(~Group.plus.var,scales="free_y",nrow=6,ncol=5) +
    theme(axis.text.x=element_text(angle=90, hjust=1)) +
    ggtitle("Minus S06 SeriesInd 41128 and S02 SeriesInd 42951 in vars 1/3/5/7")
```



We now see more clearly that variables 1, 3, 5, and 7 tend to have a similar time pattern within each series.

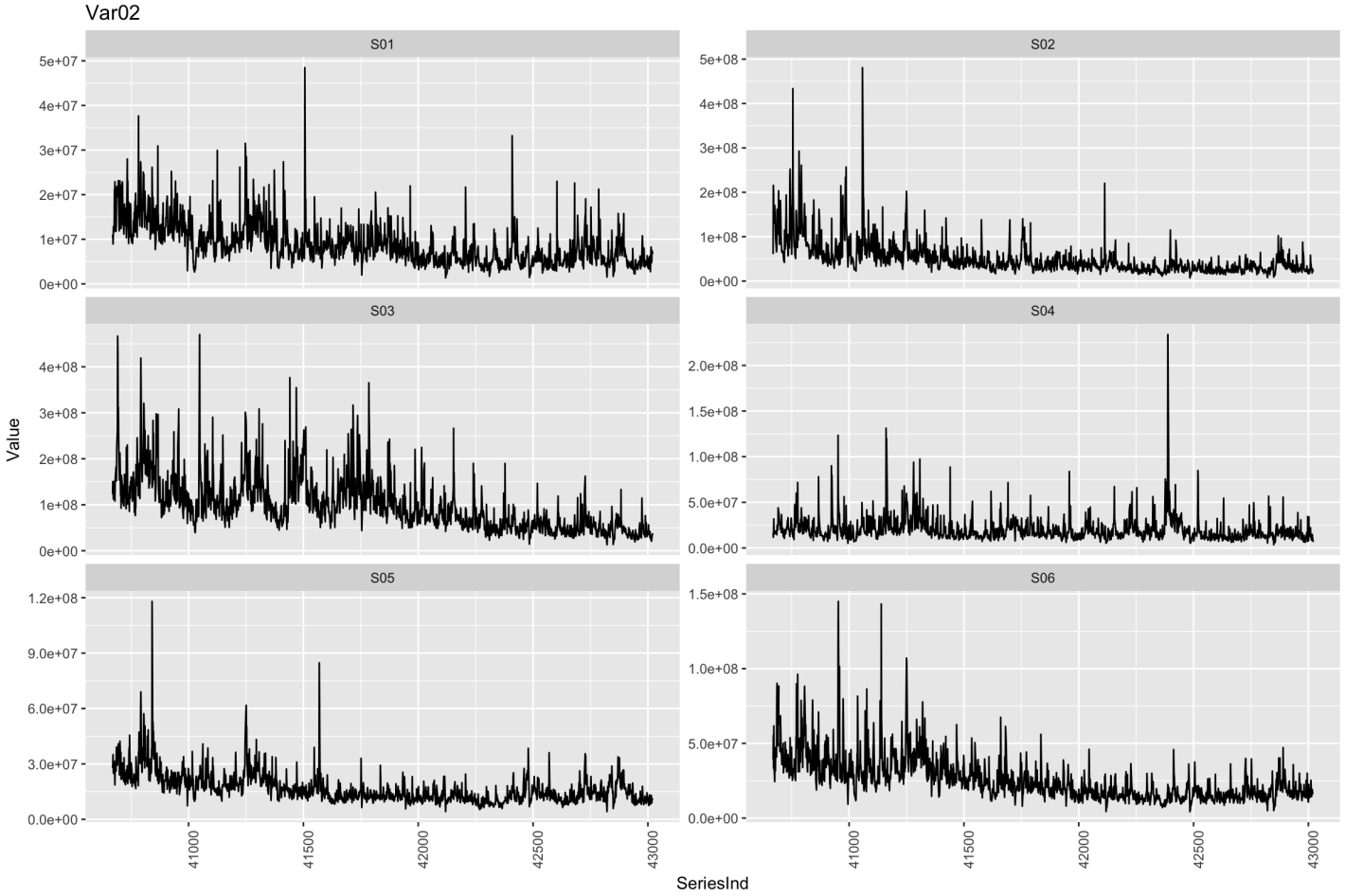
Let's try plotting these variables on the same plot for each series.

```
ggplot(data_gathered_minus_outliers[data_gathered_minus_outliers$Variable != "Var02",],
    aes(SeriesInd,Value,colour=Variable)) +
    geom_line() +
    facet_wrap(~group,scales="free_y",nrow=3,ncol=2) +
    theme(axis.text.x=element_text(angle=90, hjust=1)) +
    ggtitle("Minus S06 SeriesInd 41128 and S02 SeriesInd 42951 in vars 1/3/5/7")
```



Now, plot Var02 alone.

```
ggplot(data_gathered[data_gathered$Variable == "Var02",],
    aes(SeriesInd,Value)) +
    geom_line() +
    facet_wrap(~group,scales="free_y",nrow=3,ncol=2) +
    theme(axis.text.x=element_text(angle=90, hjust=1)) +
    ggtitle("Var02")
```



Next, let's see if there are any missing values in the data.

```
length(which(is.na(data_gathered) == TRUE))
```

```
## [1] 94
```

```
data_gathered <- data_gathered[order(data_gathered$SeriesInd,data_gathered$group),]

table(data_gathered[which(is.na(data_gathered$Value) == TRUE),"SeriesInd"])
```

```
##
## 40697 41821 42897 42898 42997 43000
##     5     5    24    24    18    18
```

```
data_gathered[which(is.na(data_gathered$Value) == TRUE & (data_gathered$SeriesInd == 406
97 | data_gathered$SeriesInd == 41821)),]
```

```
##       SeriesInd group Variable Value Group.plus.var
## 118       40697   S06    Var01    NA     S06\nVar01
## 9850      40697   S06    Var02    NA     S06\nVar02
## 19582     40697   S06    Var03    NA     S06\nVar03
## 29314     40697   S06    Var05    NA     S06\nVar05
## 39046     40697   S06    Var07    NA     S06\nVar07
## 4769      41821   S05    Var01    NA     S05\nVar01
## 14501     41821   S05    Var02    NA     S05\nVar02
## 24233     41821   S05    Var03    NA     S05\nVar03
## 33965     41821   S05    Var05    NA     S05\nVar05
## 43697     41821   S05    Var07    NA     S05\nVar07
```

```
data_gathered[which(is.na(data_gathered$Value) == TRUE & (data_gathered$SeriesInd == 428
97 | data_gathered$SeriesInd == 42898)),]
```

```
##       SeriesInd group Variable Value Group.plus.var
## 9219      42897   S01    Var01    NA     S01\nVar01
## 28683     42897   S01    Var03    NA     S01\nVar03
## 38415     42897   S01    Var05    NA     S01\nVar05
## 48147     42897   S01    Var07    NA     S01\nVar07
## 9218      42897   S02    Var01    NA     S02\nVar01
## 28682     42897   S02    Var03    NA     S02\nVar03
## 38414     42897   S02    Var05    NA     S02\nVar05
## 48146     42897   S02    Var07    NA     S02\nVar07
## 9217      42897   S03    Var01    NA     S03\nVar01
## 28681     42897   S03    Var03    NA     S03\nVar03
## 38413     42897   S03    Var05    NA     S03\nVar05
## 48145     42897   S03    Var07    NA     S03\nVar07
## 9222      42897   S04    Var01    NA     S04\nVar01
## 28686     42897   S04    Var03    NA     S04\nVar03
## 38418     42897   S04    Var05    NA     S04\nVar05
## 48150     42897   S04    Var07    NA     S04\nVar07
## 9221      42897   S05    Var01    NA     S05\nVar01
## 28685     42897   S05    Var03    NA     S05\nVar03
## 38417     42897   S05    Var05    NA     S05\nVar05
## 48149     42897   S05    Var07    NA     S05\nVar07
## 9220      42897   S06    Var01    NA     S06\nVar01
## 28684     42897   S06    Var03    NA     S06\nVar03
## 38416     42897   S06    Var05    NA     S06\nVar05
## 48148     42897   S06    Var07    NA     S06\nVar07
## 9225      42898   S01    Var01    NA     S01\nVar01
## 28689     42898   S01    Var03    NA     S01\nVar03
## 38421     42898   S01    Var05    NA     S01\nVar05
## 48153     42898   S01    Var07    NA     S01\nVar07
## 9224      42898   S02    Var01    NA     S02\nVar01
## 28688     42898   S02    Var03    NA     S02\nVar03
## 38420     42898   S02    Var05    NA     S02\nVar05
## 48152     42898   S02    Var07    NA     S02\nVar07
## 9223      42898   S03    Var01    NA     S03\nVar01
## 28687     42898   S03    Var03    NA     S03\nVar03
## 38419     42898   S03    Var05    NA     S03\nVar05
## 48151     42898   S03    Var07    NA     S03\nVar07
## 9228      42898   S04    Var01    NA     S04\nVar01
## 28692     42898   S04    Var03    NA     S04\nVar03
## 38424     42898   S04    Var05    NA     S04\nVar05
## 48156     42898   S04    Var07    NA     S04\nVar07
## 9227      42898   S05    Var01    NA     S05\nVar01
## 28691     42898   S05    Var03    NA     S05\nVar03
## 38423     42898   S05    Var05    NA     S05\nVar05
## 48155     42898   S05    Var07    NA     S05\nVar07
## 9226      42898   S06    Var01    NA     S06\nVar01
## 28690     42898   S06    Var03    NA     S06\nVar03
## 38422     42898   S06    Var05    NA     S06\nVar05
## 48154     42898   S06    Var07    NA     S06\nVar07
```

```
data_gathered[which(is.na(data_gathered$Value) == TRUE & (data_gathered$SeriesInd == 429
97 | data_gathered$SeriesInd == 43000)),]
```

```
##       SeriesInd group Variable Value Group.plus.var
## 29103     42997   S01    Var03    NA     S01\nVar03
## 38835     42997   S01    Var05    NA     S01\nVar05
## 48567     42997   S01    Var07    NA     S01\nVar07
## 29102     42997   S02    Var03    NA     S02\nVar03
## 38834     42997   S02    Var05    NA     S02\nVar05
## 48566     42997   S02    Var07    NA     S02\nVar07
## 29101     42997   S03    Var03    NA     S03\nVar03
## 38833     42997   S03    Var05    NA     S03\nVar05
## 48565     42997   S03    Var07    NA     S03\nVar07
## 29106     42997   S04    Var03    NA     S04\nVar03
## 38838     42997   S04    Var05    NA     S04\nVar05
## 48570     42997   S04    Var07    NA     S04\nVar07
## 29105     42997   S05    Var03    NA     S05\nVar03
## 38837     42997   S05    Var05    NA     S05\nVar05
## 48569     42997   S05    Var07    NA     S05\nVar07
## 29104     42997   S06    Var03    NA     S06\nVar03
## 38836     42997   S06    Var05    NA     S06\nVar05
## 48568     42997   S06    Var07    NA     S06\nVar07
## 29109     43000   S01    Var03    NA     S01\nVar03
## 38841     43000   S01    Var05    NA     S01\nVar05
## 48573     43000   S01    Var07    NA     S01\nVar07
## 29108     43000   S02    Var03    NA     S02\nVar03
## 38840     43000   S02    Var05    NA     S02\nVar05
## 48572     43000   S02    Var07    NA     S02\nVar07
## 29107     43000   S03    Var03    NA     S03\nVar03
## 38839     43000   S03    Var05    NA     S03\nVar05
## 48571     43000   S03    Var07    NA     S03\nVar07
## 29112     43000   S04    Var03    NA     S04\nVar03
## 38844     43000   S04    Var05    NA     S04\nVar05
## 48576     43000   S04    Var07    NA     S04\nVar07
## 29111     43000   S05    Var03    NA     S05\nVar03
## 38843     43000   S05    Var05    NA     S05\nVar05
## 48575     43000   S05    Var07    NA     S05\nVar07
## 29110     43000   S06    Var03    NA     S06\nVar03
## 38842     43000   S06    Var05    NA     S06\nVar05
## 48574     43000   S06    Var07    NA     S06\nVar07
```

We find the following missing values.

- SeriesInd 40697, all five variables of S06. Simply remove or use the nearby timepoints within the same series and variable to fill in.
- SeriesInd 41821, all five variables of S05. Simply remove or use the nearby timepoints within the same series and variable to fill in.
- SeriesInd 42897 and 42898, all six series and all four correlated variables (1/3/5/7). Simply remove or use the nearby timepoints within the same series and variable to fill in.
- SeriesInd 42997 and 43000, all six series and 3/4 correlated variables (3/5/7). Can either use nearby timepoints or the correlated variable without missing values (Var01).