Project 1

Rosa Juan (rij87)

3/15/2020

Introduction

```
library(tidyverse)
library(dbplyr)
library(ggplot2)
getwd()
```

```
## [1] "/Users/Rosa/Downloads/SDS 348"
```

```
use <- read.csv(file = "MotionUsage_P12-18.csv")
acc <- read.csv(file = "MotionAcceleration_P12-18.csv")
glimpse(use)
```

```
## Observations: 1,112
## Variables: 8
## $ Participant         <fct> P12, P12, P12, P12, P12, P12, P12, P12, P12, P1...
## $ Day                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,...
## $ Hour.Count          <int> 17, 18, 19, 20, 21, 22, 23, 24, 1, 2, 3, 4, 5, ...
## $ Wake.Percentage     <dbl> 1.000, 0.717, 0.667, 0.517, 0.867, 0.050, 0.100...
## $ Sleep.Percentage    <dbl> 0.000, 0.283, 0.333, 0.483, 0.133, 0.950, 0.900...
## $ Not.Worn.Percentage <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,...
## $ hrvIsValid.Sum      <int> 23, 52, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60,...
## $ Hr.1.min.           <dbl> 159.2242, 137.1555, 125.5858, 131.5647, 130.831...
```

```
glimpse(acc)
```

```
## Observations: 1,112
## Variables: 5
## $ Participant           <fct> P12, P12, P12, P12, P12, P12, P12, P12, P12, ...
## $ Absolute.Time         <fct> 4/2/18 17:00, 4/2/18 18:00, 4/2/18 19:00, 4/2...
## $ Minute.Count          <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 6...
## $ Acceleration.Sum      <dbl> 4.7041027, 4.1347807, 0.8929123, 1.8587045, 1...
## $ Acceleration.Variance <dbl> 0.012115565, 0.002782706, 0.000238694, 0.0009...
```

The two datasets chosen are from the Daily Activity Lab (DAL) research project at UT Austin. The purpose of the lab is to study daily interactions between moms and their infants. The study is trying to determine how infants' soioemotional development is impacted through various objective data measures such as movement

and heartbeat. These datasets were obtained from 6 participants who used a Movisense sensor at home over the span of a week. One dataset recorded the use of the monitor meaning on which day and on what hour the sensor was used as well as how many hours of that data were usable (Day, Hour.Count, Hr.1.min, Absolute.Time, Minute.Count, hrvlsValid.Sum). Relative to a baseline assesment, the other dataset calculated the baby's activeness percentage, including the time when the participant did not put the sensor on the baby (Wake.Percentage, Sleep.Percentage, Not.Worn.Percentage). Also, the acceleration sum and variance of the baby's motion was determined relative to the baseline (Acceleration.Sum, Acceleration Variance). This data is interesting because it is just one parameter that the DAL research center is trying to incorporate into a fitbit program or other related device that will tell mom, relatively to other babies, how their baby is developing.

Tidying: Rearranging Wide/Long

```
accNew <- unique(acc, by = "Participant") %>% pivot_wider(names_from = "Participant",
    values_from = "Absolute.Time") %>% glimpse
```

```
## Observations: 556
## Variables: 10
## $ Minute.Count          <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 6...
## $ Acceleration.Sum      <dbl> 4.7041027, 4.1347807, 0.8929123, 1.8587045, 1...
## $ Acceleration.Variance <dbl> 0.012115565, 0.002782706, 0.000238694, 0.0009...
## $ P12                   <fct> 4/2/18 17:00, 4/2/18 18:00, 4/2/18 19:00, 4/2...
## $ P13                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ P14                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ P15                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ P16                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ P17                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ P18                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
accNew %>% pivot_longer(c("P12", "P13", "P14", "P15", "P16",
    "P17", "P18"), names_to = "Participant", values_to = "Absolute.Time") %>%
    na.omit %>% glimpse
```

```
## Observations: 556
## Variables: 5
## $ Minute.Count          <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 6...
## $ Acceleration.Sum      <dbl> 4.7041027, 4.1347807, 0.8929123, 1.8587045, 1...
## $ Acceleration.Variance <dbl> 0.012115565, 0.002782706, 0.000238694, 0.0009...
## $ Participant           <chr> "P12", "P12", "P12", "P12", "P12", "P12", "P1...
## $ Absolute.Time         <fct> 4/2/18 17:00, 4/2/18 18:00, 4/2/18 19:00, 4/2...
```

Both of the datasets were already tidy and rearranging them in any other way did not make sense. Additionally, both of these datasets were hard to rearrange as the data included a lot of duplicates which makes sense since the majority of infants will fall within the same ranges of motion and participants recorded around the same dates. The 'acc' data was rearranged to be in a wider format so that each participant was a column with its Absolute.Times in the rows. Even after selecting only the unique values, there was still alot of NAs, both when pivoting wider and returning it to its original format by using pivot_longer.

Joining/Merging

```
alldata <- left_join(use, acc, by = "Participant")
```

The function left-join was used to combine 'acc' and 'use' datasets to combine one table to columns from another, matching values with the corresponded rows in this case the rows being under the column of 'Participant'. This join was chosen because it joined all matching values of the second dataset to the first dataset in which both dataset had the exact same 'Participant' column. Any other merge would have been confusing and uncertain as whether the data had combined correctly since there was a lot of duplicate data. Again, this duplicate data is because it is expected that the infants would give around the same type of data.

Wrangling

```
alldata %>% summarize_all(n_distinct)
```

```
##   Participant Day Hour.Count Wake.Percentage Sleep.Percentage
## 1           7   9         24              55               51
##   Not.Worn.Percentage hrvIsValid.Sum Hr.1.min. Absolute.Time Minute.Count
## 1                  33            40       266           556            6
##   Acceleration.Sum Acceleration.Variance
## 1              556                  397
```

```
alldata %>% summarize_if(is.numeric, mean, na.rm = T)
```

```
##        Day Hour.Count Wake.Percentage Sleep.Percentage Not.Worn.Percentage
## 1 3.911553   12.52544       0.1406528       0.06742402           0.7919244
##   hrvIsValid.Sum Hr.1.min. Minute.Count Acceleration.Sum Acceleration.Variance
## 1        9.44108  150.1155     59.89254        0.7108883            0.00045246
```

```
alldata %>% summarize_if(is.numeric, sd, na.rm = T)
```

```
##        Day Hour.Count Wake.Percentage Sleep.Percentage Not.Worn.Percentage
## 1 2.248899   6.922575       0.3065477         0.217132            0.391539
##   hrvIsValid.Sum Hr.1.min. Minute.Count Acceleration.Sum Acceleration.Variance
## 1       20.68176  28.97232     1.900433         1.219025           0.001820869
```

```
alldata %>% summarize_if(is.numeric, list(min = min, max = max),
    na.rm = T)
```

```
##   Day_min Hour.Count_min Wake.Percentage_min Sleep.Percentage_min
## 1       1              1                   0                    0
##   Not.Worn.Percentage_min hrvIsValid.Sum_min Hr.1.min._min Minute.Count_min
## 1                       0                 0      95.07498                9
##   Acceleration.Sum_min Acceleration.Variance_min Day_max Hour.Count_max
## 1            0.1356928                  9.28e-10       9             24
##   Wake.Percentage_max Sleep.Percentage_max Not.Worn.Percentage_max
## 1                   1                    1                       1
##   hrvIsValid.Sum_max Hr.1.min._max Minute.Count_max Acceleration.Sum_max
## 1                 60      237.9404               60             7.663717
##   Acceleration.Variance_max
## 1                0.02100429
```

```
alldata %>% select(Wake.Percentage, Sleep.Percentage) %>% cor
```

```
##                   Wake.Percentage Sleep.Percentage
## Wake.Percentage       1.00000000       0.09154185
## Sleep.Percentage      0.09154185       1.00000000
```

```
alldata %>% select(Wake.Percentage, Not.Worn.Percentage) %>%
    cor
```

```
##                     Wake.Percentage Not.Worn.Percentage
## Wake.Percentage           1.0000000          -0.8336933
## Not.Worn.Percentage      -0.8336933           1.0000000
```

```
alldata %>% select(Sleep.Percentage, Not.Worn.Percentage) %>%
    cor
```

```
##                     Sleep.Percentage Not.Worn.Percentage
## Sleep.Percentage           1.0000000          -0.6262269
## Not.Worn.Percentage       -0.6262269           1.0000000
```

```
alldata %>% group_by(Day) %>% summarize(var(Acceleration.Sum,
    na.rm = T))
```

```
## # A tibble: 9 x 2
##     Day `var(Acceleration.Sum, na.rm = T)`
##   <int>                              <dbl>
## 1     1                               1.83
## 2     2                               1.96
## 3     3                               1.90
## 4     4                               1.53
## 5     5                              0.638
## 6     6                              0.638
## 7     7                              0.638
## 8     8                              0.638
## 9     9                              0.638
```

```
alldata %>% group_by(Hour.Count) %>% summarize(`25%` = quantile(Acceleration.Sum,
    probs = 0.25), `50%` = quantile(Acceleration.Sum, probs = 0.5),
    `75%` = quantile(Acceleration.Sum, probs = 0.75), n = n())
```

```
## # A tibble: 24 x 5
##    Hour.Count `25%` `50%` `75%`     n
##         <int> <dbl> <dbl> <dbl> <int>
##  1          1 0.207 0.209 0.273  9920
##  2          2 0.207 0.209 0.273  9920
##  3          3 0.207 0.209 0.273  9920
##  4          4 0.207 0.209 0.273  9920
##  5          5 0.207 0.209 0.273  9920
##  6          6 0.207 0.209 0.273  9920
##  7          7 0.207 0.209 0.273  9920
```

```
##  8          8 0.207 0.209 0.273  9920
##  9          9 0.207 0.209 0.273  9920
## 10         10 0.207 0.209 0.273  9920
## # ... with 14 more rows
```

```
alldata %>% group_by(Participant = Participant) %>% select(-c("Hr.1.min.",
    "Absolute.Time", "Minute.Count")) %>% summarize_if(is.numeric,
    mean, na.rm = T)
```

```
## # A tibble: 7 x 9
##   Participant   Day Hour.Count Wake.Percentage Sleep.Percentage Not.Worn.Percen~
##   <fct>       <dbl>      <dbl>           <dbl>            <dbl>            <dbl>
## 1 P12          2.13       12.4          0.371           0.189             0.440
## 2 P13          2.06       12.5          0.648           0.292             0.0606
## 3 P14          2.58       12.5          0.0296          0.308             0.662
## 4 P15          4.83       12.5          0.0456          0.00972           0.945
## 5 P16          2.62       12.5          0.488           0.150             0.363
## 6 P17          2.25       13.1          0.176           0.0695            0.754
## 7 P18          2.46       12.5          0.233           0                 0.767
## # ... with 3 more variables: hrvIsValid.Sum <dbl>, Acceleration.Sum <dbl>,
## #   Acceleration.Variance <dbl>
```

```
alldata %>% group_by(Participant = Participant) %>% summarize(sd(Acceleration.Sum,
    na.rm = T), n(), n_distinct(Day))
```

```
## # A tibble: 7 x 4
##   Participant `sd(Acceleration.Sum, na.rm = T)`  `n()` `n_distinct(Day)`
##   <fct>                                   <dbl>  <int>             <int>
## 1 P12                                     1.42    8464                 3
## 2 P13                                     1.41    9604                 3
## 3 P14                                     0.615  20736                 4
## 4 P15                                     0.798 147456                 9
## 5 P16                                     1.84   20736                 4
## 6 P17                                     0.979  11236                 3
## 7 P18                                     1.80   20736                 4
```

```
alldata %>% group_by(Day) %>% select(-c("Hr.1.min.", "Absolute.Time",
    "Minute.Count")) %>% summarize_if(is.numeric, mean, na.rm = T)
```

```
## # A tibble: 9 x 8
##     Day Hour.Count Wake.Percentage Sleep.Percentage Not.Worn.Percen~
##   <int>      <dbl>           <dbl>            <dbl>            <dbl>
## 1     1       18.0         0.265             0.187            0.547
## 2     2       12.5         0.350             0.147            0.502
## 3     3       11.8         0.104             0.0365           0.860
## 4     4       10.5         0.00721           0                0.993
## 5     5       12.5         0                 0                1
## 6     6       12.5         0.1               0.0778           0.822
## 7     7       12.5         0.0354            0                0.965
## 8     8       12.5         0                 0                1
## 9     9        4.5         0                 0                1
## # ... with 3 more variables: hrvIsValid.Sum <dbl>, Acceleration.Sum <dbl>,
## #   Acceleration.Variance <dbl>
```

```
alldata %>% group_by(Hour.Count) %>% summarize(sd(Acceleration.Sum,
    na.rm = T))
```

```
## # A tibble: 24 x 2
##    Hour.Count `sd(Acceleration.Sum, na.rm = T)`
##         <int>                             <dbl>
##  1          1                              1.22
##  2          2                              1.22
##  3          3                              1.22
##  4          4                              1.22
##  5          5                              1.22
##  6          6                              1.22
##  7          7                              1.22
##  8          8                              1.22
##  9          9                              1.22
## 10         10                              1.22
## # ... with 14 more rows
```

```
alldata %>% filter(Participant == "P12") %>% arrange(Hour.Count) %>%
    mutate(HrsValid.Percentage = hrvIsValid.Sum/72) %>% glimpse
```

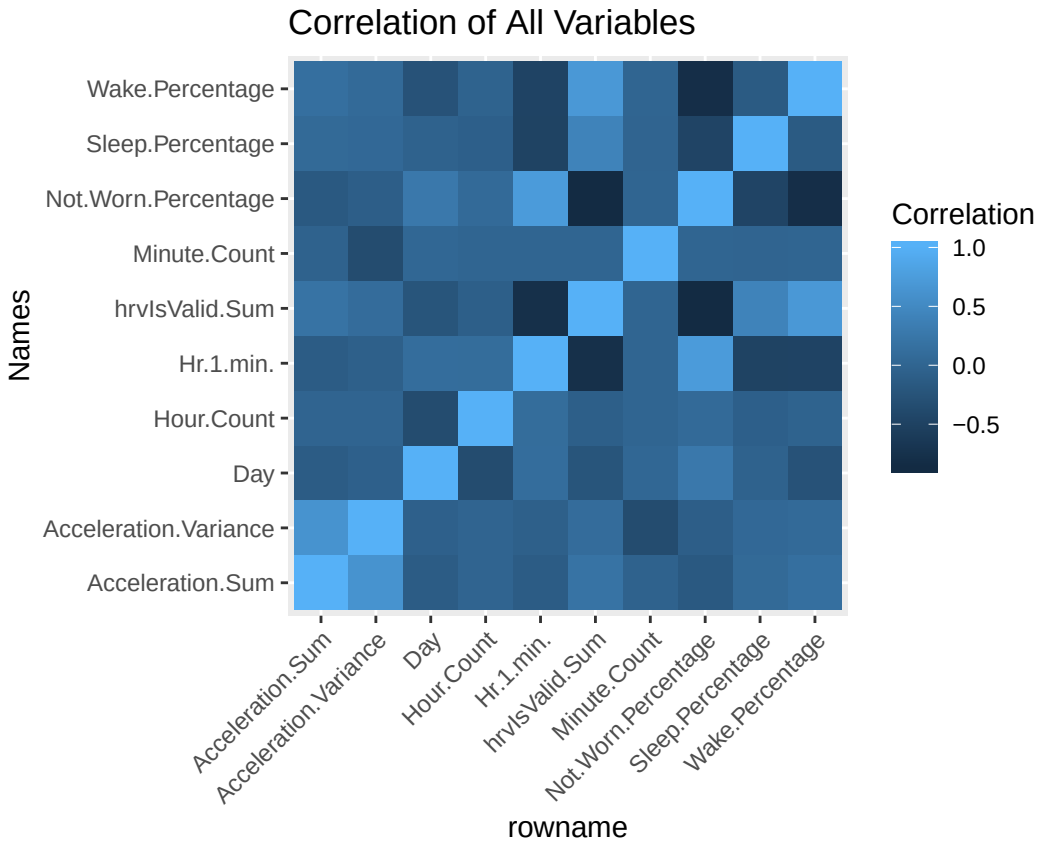
```
## Observations: 8,464
## Variables: 13
## $ Participant           <fct> P12, P12, P12, P12, P12, P12, P12, P12, P12, ...
## $ Day                   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ Hour.Count            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Wake.Percentage       <dbl> 0.467, 0.467, 0.467, 0.467, 0.467, 0.467, 0.4...
## $ Sleep.Percentage      <dbl> 0.533, 0.533, 0.533, 0.533, 0.533, 0.533, 0.5...
## $ Not.Worn.Percentage   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ hrvIsValid.Sum        <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 6...
## $ Hr.1.min.             <dbl> 116.1844, 116.1844, 116.1844, 116.1844, 116.1...
## $ Absolute.Time         <fct> 4/2/18 17:00, 4/2/18 18:00, 4/2/18 19:00, 4/2...
## $ Minute.Count          <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 6...
## $ Acceleration.Sum      <dbl> 4.7041027, 4.1347807, 0.8929123, 1.8587045, 1...
## $ Acceleration.Variance <dbl> 0.012115565, 0.002782706, 0.000238694, 0.0009...
## $ HrsValid.Percentage   <dbl> 0.8333333, 0.8333333, 0.8333333, 0.8333333, 0...
```

The data contains zeros as well as a wide range of data which could make it difficult to plot and interpret. The correlation between Wake.Percentage and Sleep.Percentage is very low but this is probably because there was more time that the participants did not use the Movisense than the time they were using it. Given that, as expected, the correlation between Wake.Percentage and Sleep.Percentage with Not.Worn.Percentage is high. As expected, the mean and standard deviation are most similar when the data is grouped by participants than by day which were most likely at different days of the week. This gives space for confounding variables in mom's ability to respond to baby and the amount of time they have to interact with baby.

Visualizing

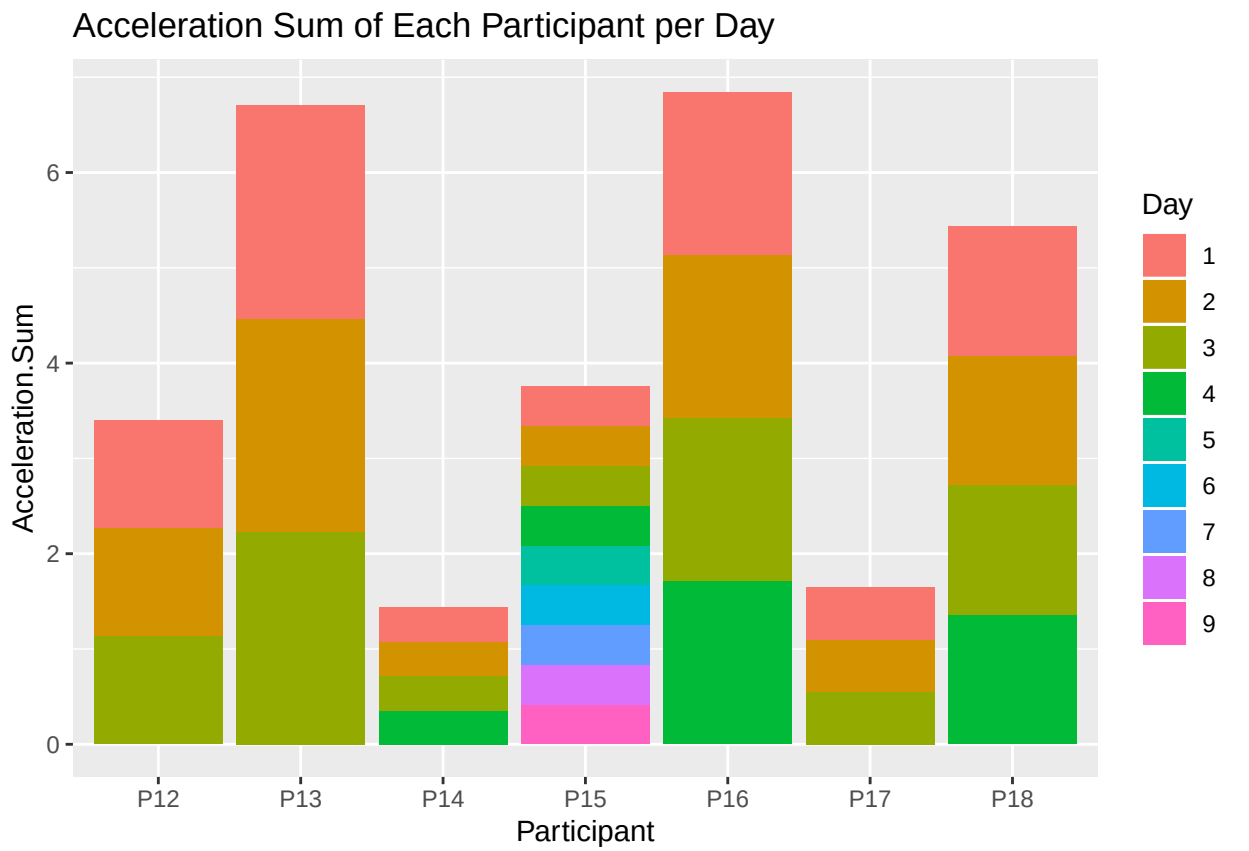
```
df <- alldata %>% na.omit %>% select_if(is.numeric) %>% cor %>%
    as.data.frame %>% rownames_to_column %>% pivot_longer(-1,
    names_to = "Names", values_to = "Correlation")
```

```
df %>% ggplot(aes(rowname, Names, fill = Correlation)) + geom_tile() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    coord_fixed() + labs(title = "Correlation of All Variables")
```



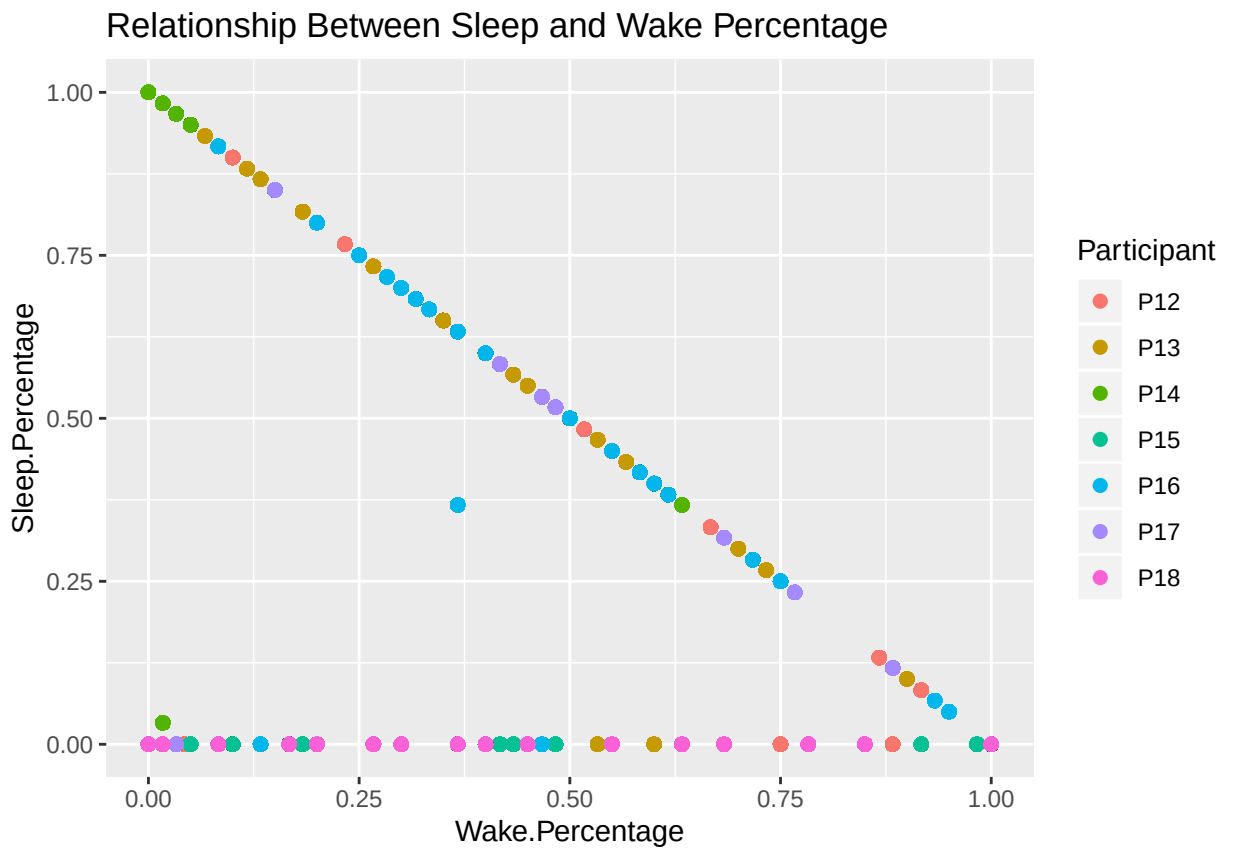
Based on this heatmap, there is not a lot of strong correlations as most of the shades are in the medium range near zero and not in the polar ends of one. However, there is a strong correlation between hrvIsValid.Sum and Hr.1.min as well as with Not.Worn.Percentage which makes sense since when the sensor was not being used, participants were asked to place the sensor in a box they were given so the data was most similar then as opposed to when infants did have the sensor on them.

```
newdata <- alldata %>% mutate(Day = as.character(Day))

ggplot(newdata, aes(Participant, Acceleration.Sum, fill = Day)) +
    geom_bar(stat = "summary", fun.y = "mean") + labs(title = "Acceleration Sum of Each Participant per
```



The Acceleration.Sum is a measure of the infant's movements. This looked different across all participants, event among those that recorded the same amount of days such as P12 and P13. Interestingly, even though P15 recorded over the span of 9 days, it just had an Acceleration.Sum mean a little over that of P12. Perhaps this indicates that data over larger sets of days is needed to get a more accurate picture of what motion looks like thoughout a week.

```
ggplot(newdata, aes(Wake.Percentage, Sleep.Percentage, color = Participant)) +
    geom_point(size = 2) + labs(title = "Relationship Between Sleep and Wake Percentage")
```



As expected, the relationship between sleep and wake percentage has a negative correlation meaning that as an infant has a higher wake percentage then that means that they have less of the sleep percantage because an infant cannot be both sleeping and be awake at the same time. However, there is a straight line where even though wake percentage increases, sleep percentage does not change. This is likely due to the not=worn percentage.

Dimensionality Reduction

```
library(cluster)
view <- alldata %>% select(-Participant, -Hr.1.min., -Absolute.Time) %>%
    mutate_all(as.numeric) %>% cor %>% glimpse
```

```
##  num [1:9, 1:9] 1 -0.177 -0.335 -0.243 0.397 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:9] "Day" "Hour.Count" "Wake.Percentage" "Sleep.Percentage" ...
##   ..$ : chr [1:9] "Day" "Hour.Count" "Wake.Percentage" "Sleep.Percentage" ...
```

```
pdata <- alldata %>% select(-Hr.1.min., -Participant, -Absolute.Time) %>%
    mutate_all(function(x) log(x + 2))
mypca <- prcomp(pdata, center = TRUE, scale. = TRUE)
summary(mypca)
```

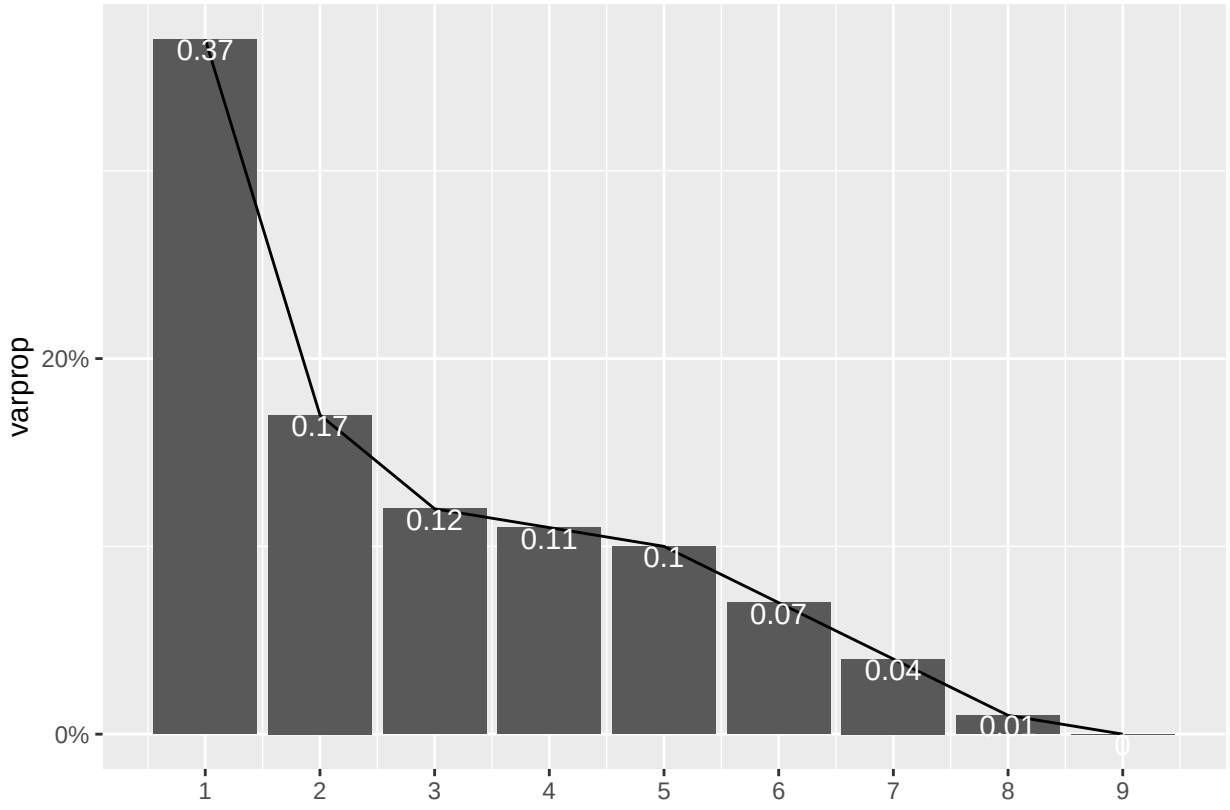
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
```

```
## Standard deviation     1.8340 1.2397 1.0466 0.9881 0.9595 0.80969 0.5724
## Proportion of Variance 0.3737 0.1708 0.1217 0.1085 0.1023 0.07284 0.0364
## Cumulative Proportion  0.3737 0.5445 0.6662 0.7747 0.8770 0.94983 0.9862
##                            PC8     PC9
## Standard deviation     0.35001 0.03816
## Proportion of Variance 0.01361 0.00016
## Cumulative Proportion  0.99984 1.00000
```

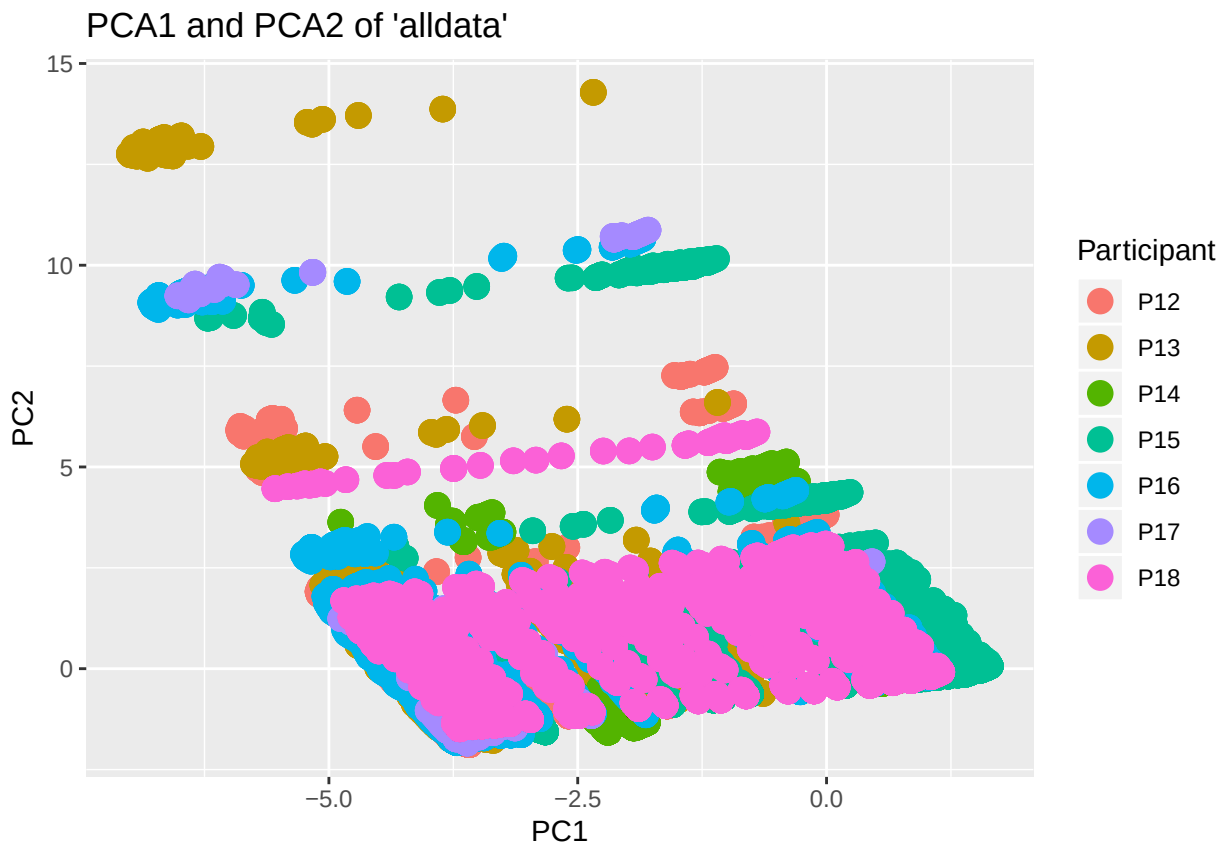
```
str(mypca)
```

```
## List of 5
##  $ sdev    : num [1:9] 1.834 1.24 1.047 0.988 0.959 ...
##  $ rotation: num [1:9, 1:9] 0.3028 -0.0453 -0.465 -0.2921 0.5174 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:9] "Day" "Hour.Count" "Wake.Percentage" "Sleep.Percentage" ...
##   .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:9] 1.706 2.526 0.752 0.722 1.015 ...
##   ..- attr(*, "names")= chr [1:9] "Day" "Hour.Count" "Wake.Percentage" "Sleep.Percentage" ...
##  $ scale   : Named num [1:9] 0.375 0.597 0.1265 0.0907 0.1583 ...
##   ..- attr(*, "names")= chr [1:9] "Day" "Hour.Count" "Wake.Percentage" "Sleep.Percentage" ...
##  $ x       : num [1:238968, 1:9] -5.29 -4.41 -3.59 -3.88 -3.91 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"
```

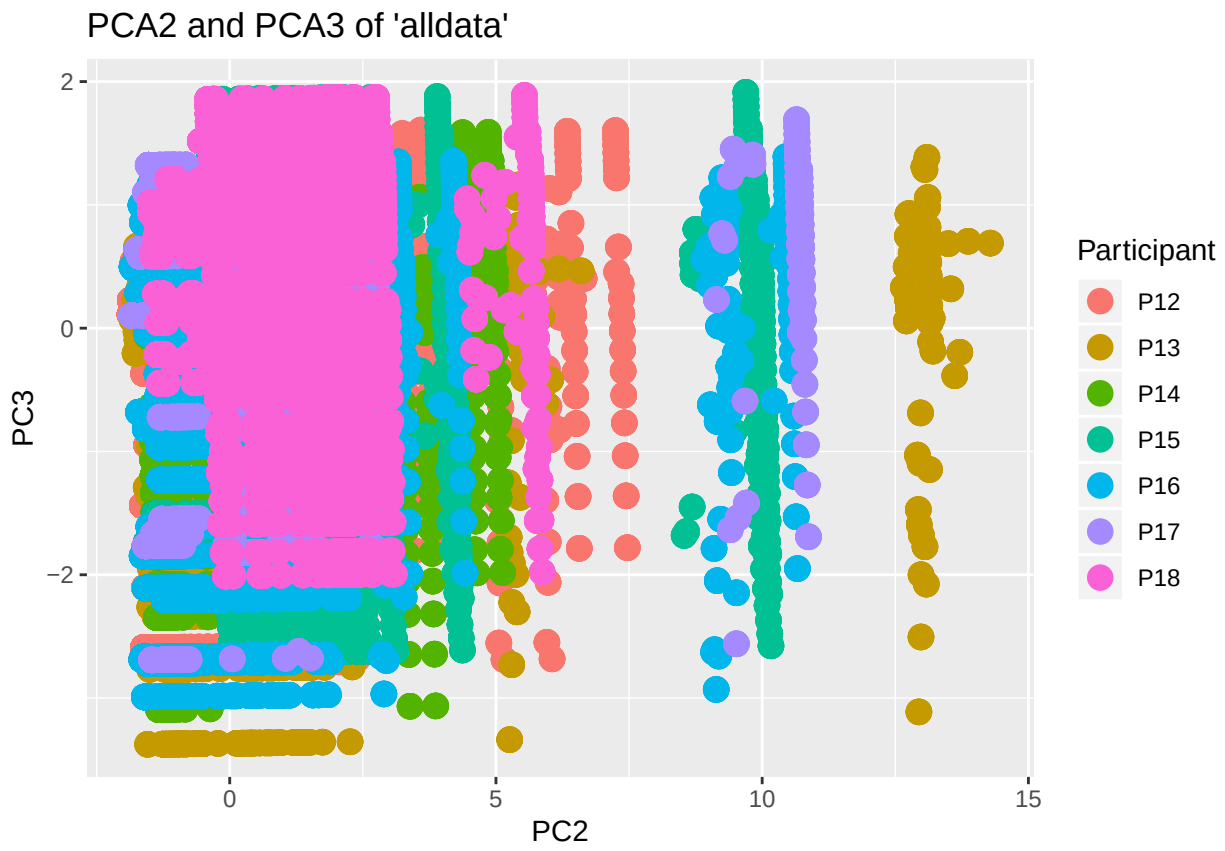
```
eigval <- mypca$sdev^2
varprop = round(eigval/sum(eigval), 2)
ggplot() + geom_bar(aes(y = varprop, x = 1:9), stat = "identity") +
    xlab("") + geom_path(aes(y = varprop, x = 1:9)) + geom_text(aes(x = 1:9,
    y = varprop, label = round(varprop, 2)), vjust = 1, col = "white",
    size = 4) + scale_y_continuous(breaks = seq(0, 0.6, 0.2),
    labels = scales::percent) + scale_x_continuous(breaks = 1:9)
```



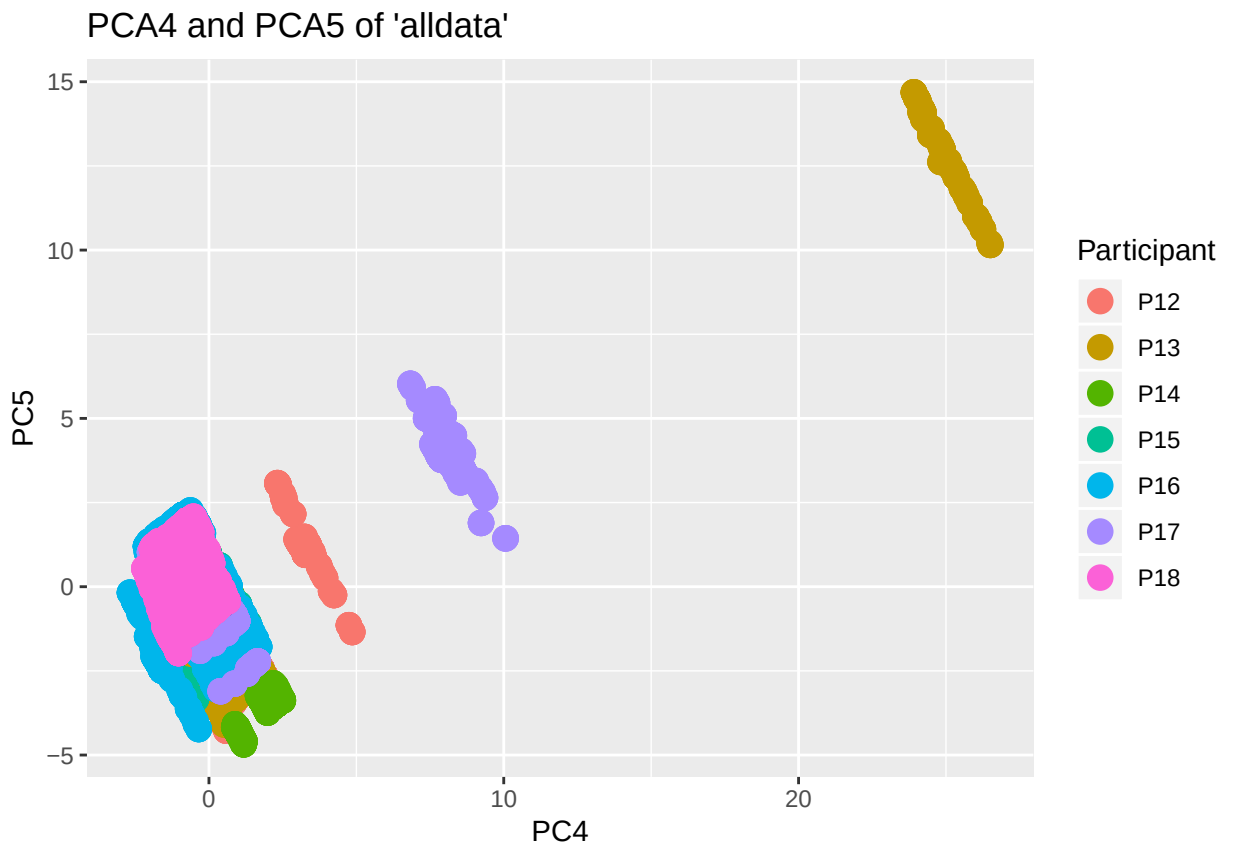

```
results <- alldata %>% select(-Hr.1.min., -Absolute.Time) %>%
    as.data.frame %>% mutate(PC1 = mypca$x[, 1], PC2 = mypca$x[,
    2], PC3 = mypca$x[, 3], PC4 = mypca$x[, 4], PC5 = mypca$x[,
    5])

results %>% ggplot(aes(PC1, PC2, color = Participant)) + geom_point(size = 4) +
    labs(title = "PCA1 and PCA2 of 'alldata'")
```




```
results %>% ggplot(aes(PC2, PC3, color = Participant)) + geom_point(size = 4) +
    labs(title = "PCA2 and PCA3 of 'alldata'")
```




```
results %>% ggplot(aes(PC4, PC5, color = Participant)) + geom_point(size = 4) +
    labs(title = "PCA4 and PCA5 of 'alldata'")
```



According to the scree plot and going with the rule of thumb to pick PCs until cumulative proportion of variance is greater than 80%, 5 PCs should be retained which account for 87% of the data. This PCA shows very spread out data where the majority of PC1 is mostly influenced by P18 and where PC2 is mostly influences by P13. The PC1 value shows that all componentsd have a similar sign and magnitude which is true since a lot of data are duplicated. PC2 is the correlation between HrvIsValid and Not.Worn.Percentage meaning that the more the sensor was not worn, the more valid hours were calculated since there was only one possible value and we knew that the value meant it was not being worn. However, as the Not.Worn.Percentage increased, the data got more and more noise and the amount of valid/usable hours decreased. PC3 is the correlation between Wake.Percentage and HrvIsValid, which is close to the value of PC2 which means that the data is similarly reliable. However, it was a positive rather than a negative correlation which is good since data when infant is awake is desired. As infant spends more time awake, the number of valid hours also increases. PC4 means that the higher the Acceleration.Sum, the higher the Acceleration variance. PC5 means that the higher the sleep percentage, the more valid hours there is though it has a significantly weaker correlation than that between wake percentage and valid hours. The data was transformed with log so as to minimize the inflation by zeros. As expected, the data is still quite messy since the infants all had similar data with very small and very big numbers which makes it incredibly difficult to cluser. Another categorical variable would have given more insight such as the infant's age which could have eliminated some of the confouding variables such as how much time an infant spends asleep or awake as well as the range of motion.
