SYDE 780 Assignment 2: R Module on data visualizations

Winter 2025

The **RSA dataset** on the course portal page offers 1071 sample implant migration data observations (j is the notation for observations) from 137 subjects (n is the notation for subjects). In this module, we will be examining features related to Maximum Total Point Migration (MTPM, measured in millimeters) longitudinally among patients.

Getting Started

```
rm(list=ls())  # Clears R Envionrment

library(ggplot2) # Load plotting library
```

Read in the dataset

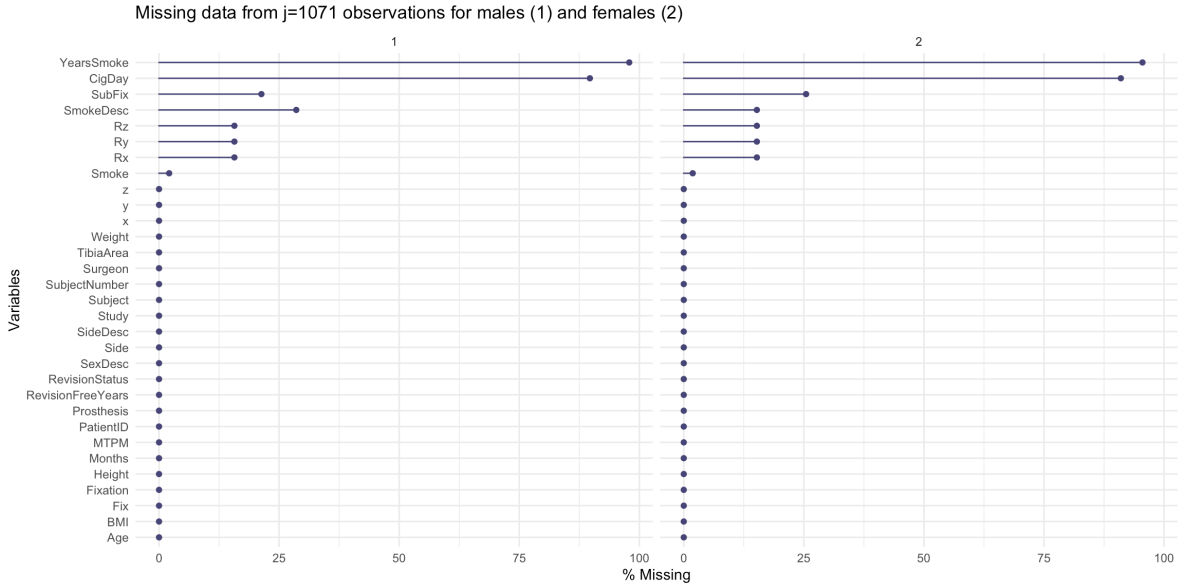
```
#Read in data
  fileToImport = "./dataset.csv"
  Data.RSA = read.csv(fileToImport, header=TRUE) #n=1071
```

Visualize your data

Data completeness

It is important to understand missingness in any dataset. For example, if there is poor data representation among a certain demographic (such as females), analysis conducted using your sample population may not accurately represent the true population. Because we are working with a longitudinal dataset (more than one observation per person), a lack of follow-up response among a certain demographic, or those who go on to have poorer outcomes, may lend to bias findings and interpretation of results. Because this sort of missingness may be *not random*, we need to investigate and understand patterns in missing data, as most statistical tools are only valid under conditions of a truly random population sample.

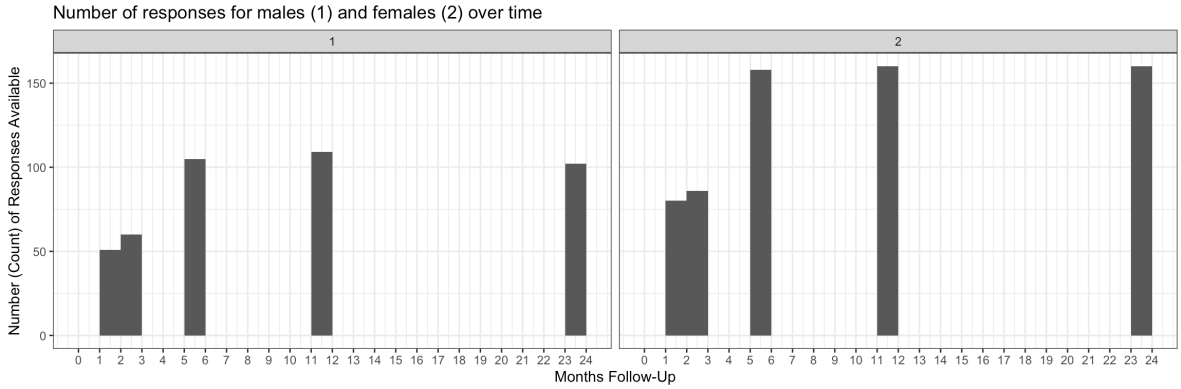
```
#Plot completeness of the dataset (you don't need to re-create this plot)
  library(naniar) #CRAN library for feature completeness plotting
  gg_miss_var(Data.RSA, facet=Sex, show_pct = TRUE) + labs(title = "Missing data from j=1071 observations for mal
es (1) and females (2)")
```



As we can see from the plot above, we have a fairly complete dataset, with 32 features. There is also a fairly equal response completeness among males and females. Descriptive information on smoking is the category with the most missing information.

Now let's plot the number of responses by males and females at each follow-up time:

```
#Plot a histogram of responses by follow up time for males and females
  ggplot(data=Data.RSA, aes(x=Months)) + theme_bw() +
    facet_grid(. ~ Sex) +
    geom_histogram(bins = 25, breaks=seq(0,24,by=1)) +
    scale_x_continuous(breaks=seq(0,24,by=1)) +
    labs(title = "Number of responses for males (1) and females (2) over time", x="Months Follow-Up", y="Number
(Count) of Responses Available")
```



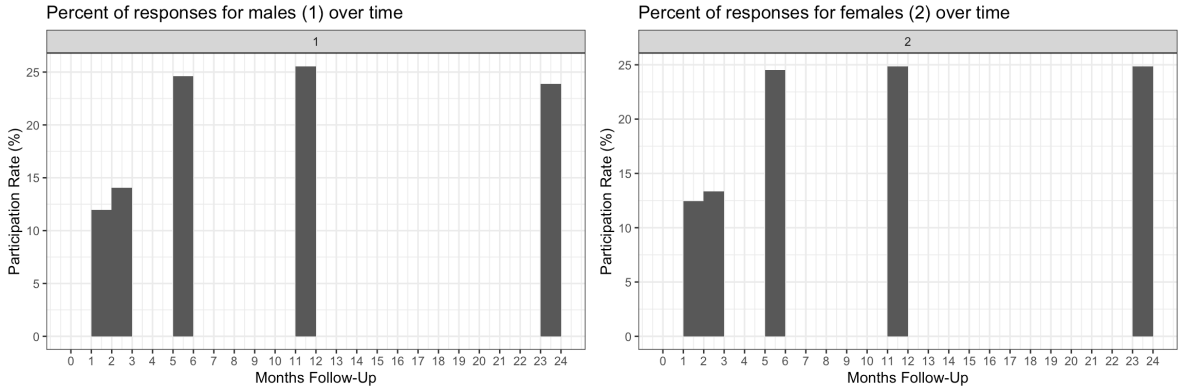
This plot tells us two things; i) if we have an equal amount of data between males and females, and ii) if we have an equal number of data points at all follow-ups between males and females.

From above, it is clear that there are more female participants (i.e., the dataset simply contains more female patients), and there is more data at later follow-up times (i.e., six months or later).

However, to be sure that the *follow-up participation rate* ((number of patients who participated at time x / total number of patients who participated) * 100) is equivalent between males and females at each follow-up, we can then plot the participation rate (%) at each time point, for each sex separately. From the plot below, both sexes have similar data available at each time point (i.e., males and females were equally likely to attend each follow-up).

We may need to keep this is mind during interpretation, particularly if sex is found to be a meaningful factor in our results. We could say something like “This dataset had a higher ratio of females over males, yet each sex was equally likely to participate at each follow-up.”

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



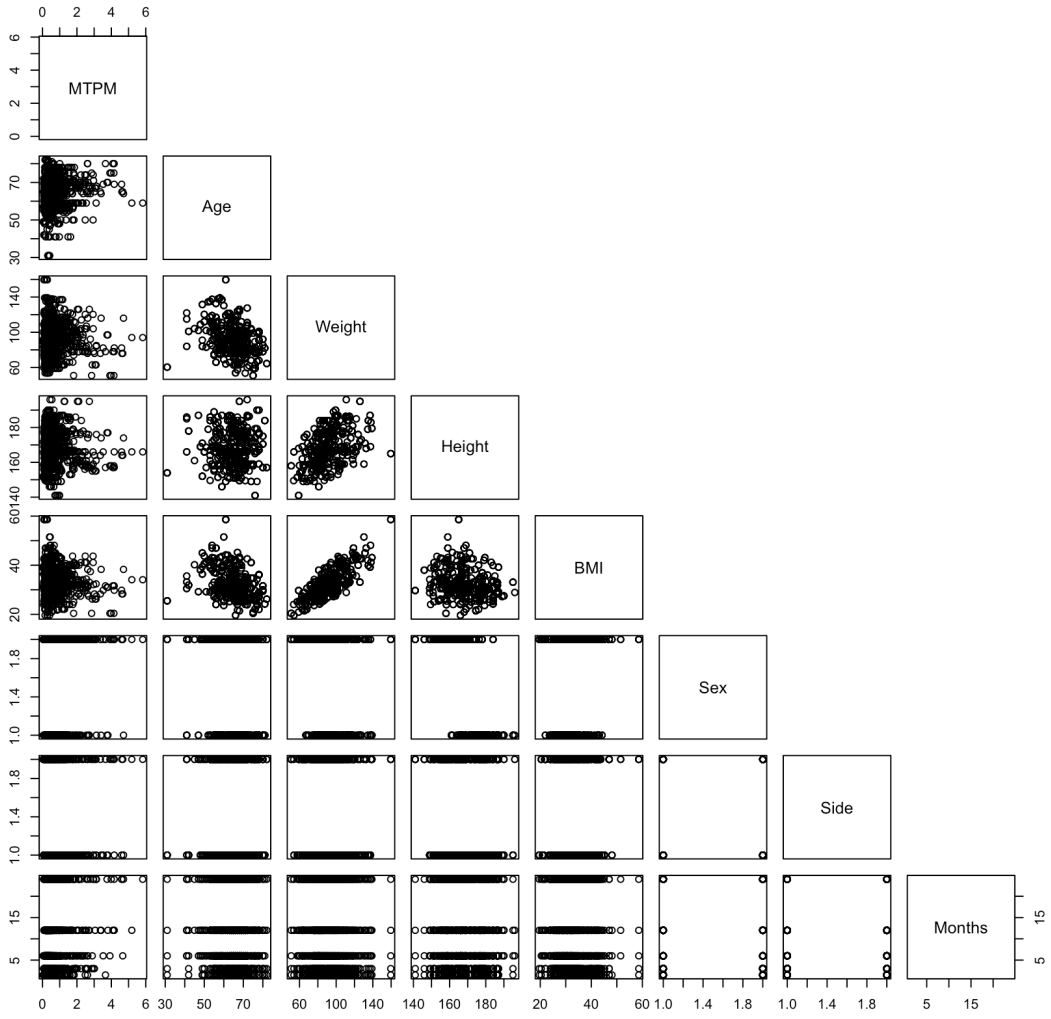
Features associations

Pairs (pair-wise relationship) plots offer a quick check of any associations between features (see https://en.wikipedia.org/wiki/Correlation_and_dependence (https://en.wikipedia.org/wiki/Correlation_and_dependence) for plot interpretation for continuous data). Pairs plots for patient demographic features, prosthesis features, and smoking features are included below. Females are represented in red, and males in blue. Sample code to re-create these plots has been provided.

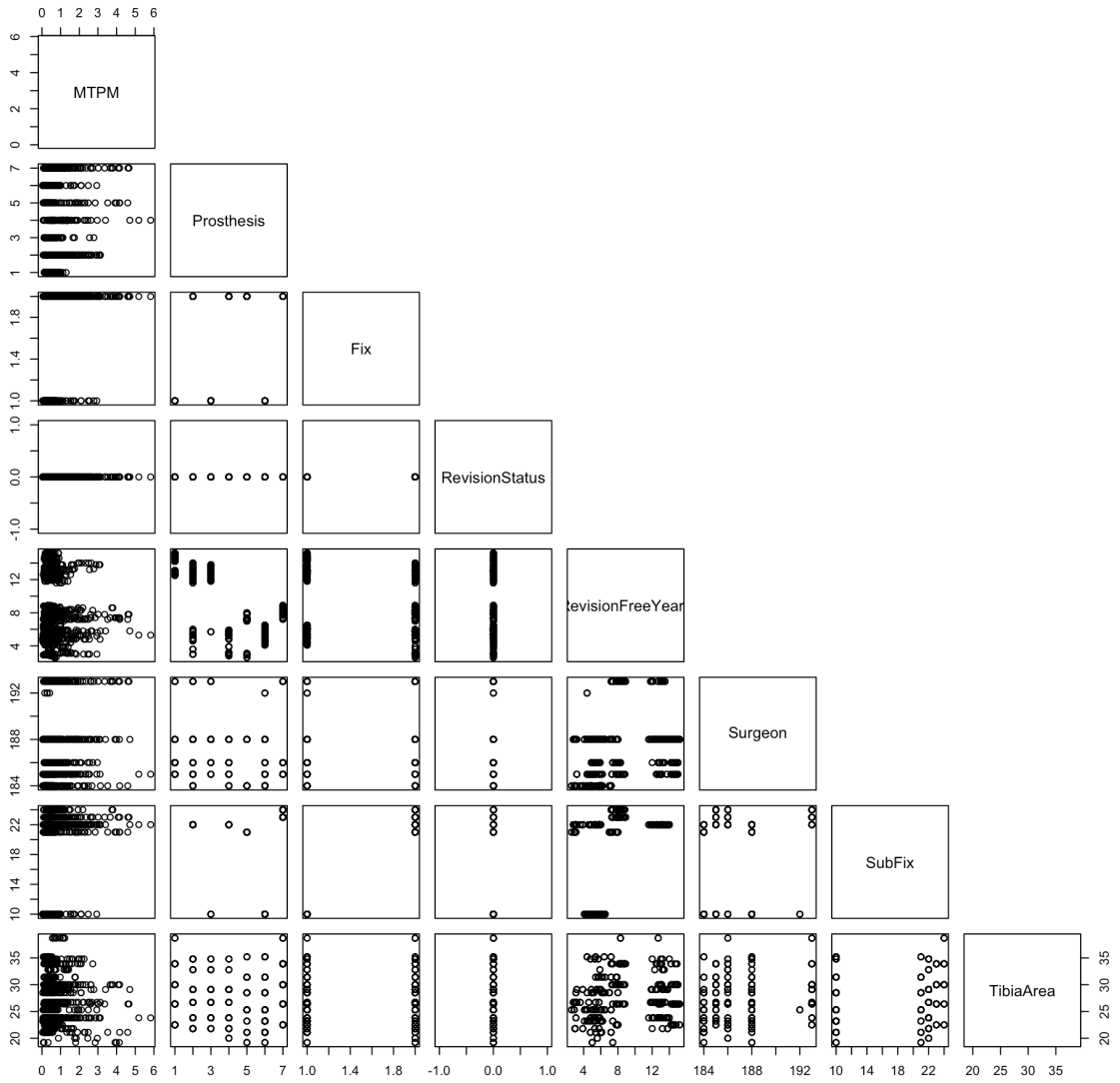
If we find high correlations between our response variable of interest (MTMP) and a feature, that feature may be a significant predictor in our response. If we find a high correlation between two other features (such as `BMI` and `Weight`), it might suggest **multi-collinearity** of features. In this case, only one of the two features would typically be applied to any multi-variable analysis, as to limit redundant information being applied to any model. (We will learn more about multi-variable regression later in the module).

```
#Demographic & Collection Features
  #Select features to plot
  Data.RSA.pairs = Data.RSA[, c("MTPM", "Age", "Weight", "Height", "BMI", "Sex", "Side", "Months")]
  #Plot pairs data
  pairs(Data.RSA.pairs, labels = colnames(Data.RSA.pairs),
        main = "Demographic Features Pairs Matrix", pch = 21, bg = c("red", "blue")[unclass(unclass(Data.RSA$SexD
esc))],
        upper.panel = NULL, cex.labels=1.2)
```

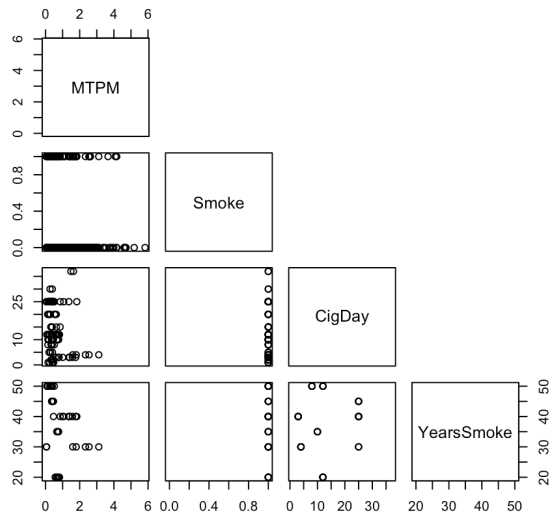
Demographic Features Pairs Matrix



Prosthesis & Surgical Features Pairs Matrix



Smoking Features Pairs Matrix



Assignment Question

Question 1

Question 1

a. Describe the terms correlation and multicollinearity. What might be a consequence of using highly correlated data during any multi-variable modeling?

b. Using the Help tab (or Google) and the function cor.test, calculate the Pearson's correlation coefficient and report the corresponding p-value between Weight and BMI.

c. Based on your findings in part b, which demographic feature(s) (if any) might you remove during analysis?

a. Describe the terms correlation and multicollinearity. What might be a consequence of using highly correlated data during any multi-variable modeling?

- **Correlation**: A statistical relationship between variables where changes in one leads to a similar change in another. For example, the sea level rises with global age.
- **Multicollinearity**: When multiple variables (>2) are correlated with each other in a statistical model
- **Highly correlated data** may lead to misleading results during multivariate studies For example, a researcher studies the effect of age on typing speed, finds a negative correlation, then concludes aging decreases our ability to type. However, age is also correlated with muscle and bone health, and it would be more accurate to link these traits to typing ability rather than age. With multicollinearity, the correlations may become confused as causations because it's unclear which variable(s) are the most tied to the dependent variable.

b. Using the Help tab (or Google) and the function cor.test, calculate the Pearson's correlation coefficient and report the corresponding p-value between Weight and BMI.

```
cor_test_result <- cor.test(Data.RSA$Weight, Data.RSA$BMI, method="pearson")
cat("p-value: ", cor_test_result$p.value, "\n")
```

```
## p-value:  1.087925e-236
```

```
cat("Pearson's correlation coefficient:", cor_test_result$estimate)
```

```
## Pearson's correlation coefficient: 0.797356
```

c. Based on your findings in part b, which demographic feature(s) (if any) might you remove during analysis?

Weight and BMI have a strong positive correlation which is justified given that weight is a component of BMI. Since height (another component of BMI) is also given in the dataset, I would remove BMI for the analysis. Though, it would also make sense to remove both weight and height in favor of BMI since it provides a more holistic measure, at the cost of granularity.

Assignment Question

Question 2

a. Using the pairs plots, discuss any observations between fixation method and MTPM.

b. Using the pairs plots, discuss any observations between sex and MTPM.

c. Using the pairs plots, discuss any observations between prosthesis and MTPM.

d. Using the pairs plots, discuss any observations between month follow-up and MTPM.

a. Using the pairs plots, discuss any observations between fixation method and MTPM.

- Uncemented fixtures (Fix=2) have a higher range (and possibly mean) of MTPM
- Both types of fixtures follow a similar left-skewed distribution, but the uncemented one is wider
- It appears that there might be more uncemented fixtures (there are 45 (~8.8%) more)

b. Using the pairs plots, discuss any observations between sex and MTPM.

- Female (Sex=2) participants have a larger MTPM range compared to males (Sex=1)
- Female participants may have a larger MTPM mean value
- Both distributions are left-skewed, but females have a wider distribution

c. Using the pairs plots, discuss any observations between prosthesis and MTPM.

- Each of the prosthesis v MTPM distributions are visibly different
- Prosthesis #1 and #2 appear to be uniformly distributed
- Prosthesis #3-#7 are left skewed
- Prosthesis #3-#6 have very long tails (ie large ranges)
- Prosthesis #1 has the smallest MTPM range
- All prosthesis appear to have roughly the same number of datapoints

d. Using the pairs plots, discuss any observations between month follow-up and MTPM.

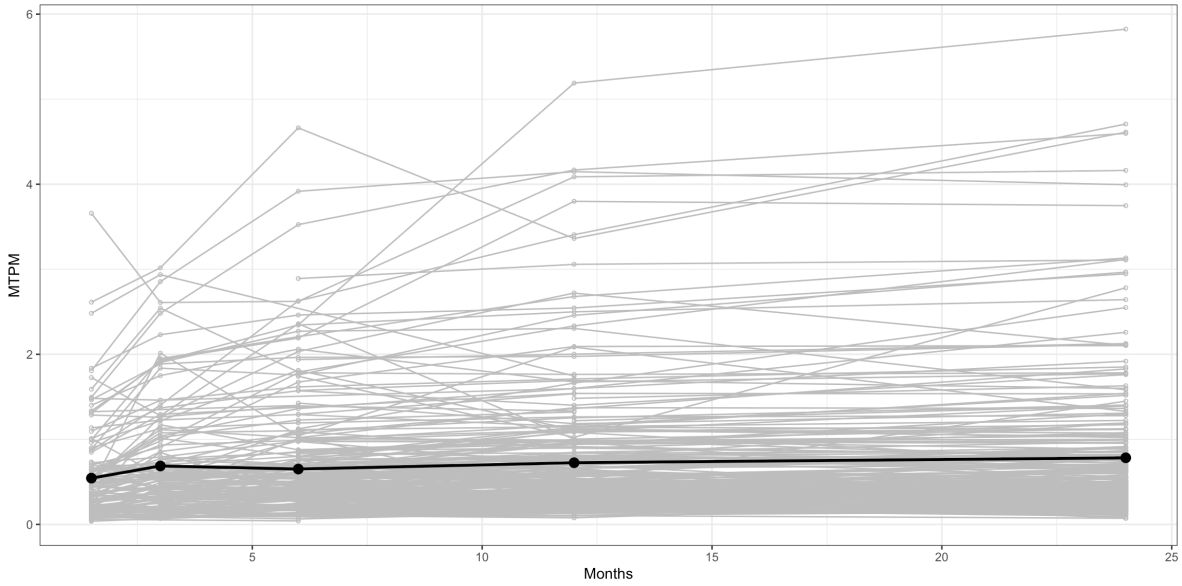
- Follow-up time ("Months") appears to be positively correlated with MTPM
- The datapoints become more sparse as the follow-up time increases (increasing standard deviation)
- All distributions appear left-skewed

Data Visualizations for longitudinal data

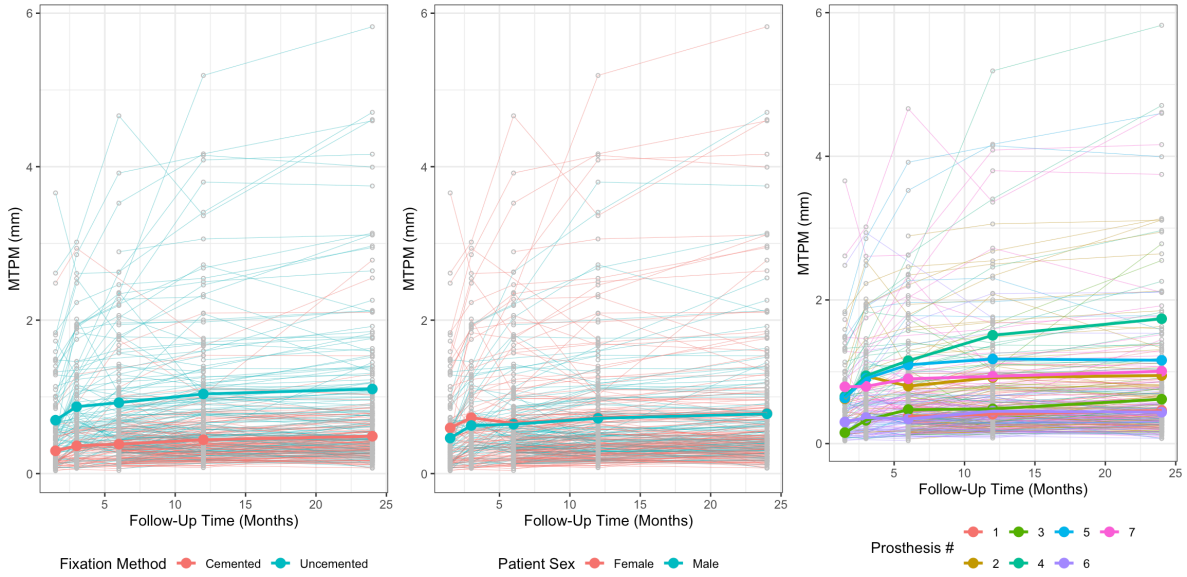
Visualize the data longitudinally

From Question 2d, we know there may be a relationship between MTPM and follow-up time. As with any longitudinal dataset, it's important to visualize patterns in our data over time. We can use a "spaghetti" plot to do this, which shows MTPM on the y-axis, and follow-up time in months on the x-axis. Further, we can compute the mean MTPM for all patients at each time point (black line).

```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## ℹ Please use the `fun` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



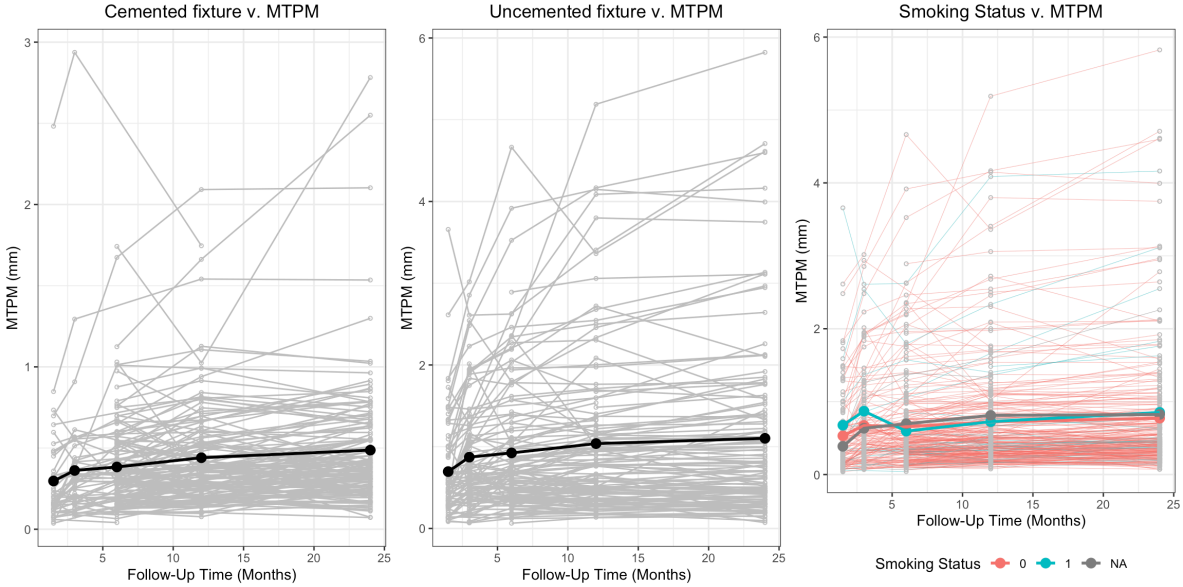
We can also plot data longitudinally by some of the explanatory variables (bold lines are group means)



Subsetting Data

Note that the line `Data.RSA.Cemented <- Data.RSA[which(Data.RSA$Fixation=="Cemented"), ]` pulls all rows when column column `Fixation==Cemented` .

Recreate the plots looking at sex and prosthesis above for Cemented and Uncemented implants separately and generate a 3rd plot examining an additional variable of your choosing.



Assignment Question

Question 3

a. Using your longitudinal plots for cemented data, provide some observations about the cemented implants.

b. Using your longitudinal plots for uncemented data, provide some observations about the uncemented implants

c. Discuss any differences between the two types of fixation.

a. Using your longitudinal plots for cemented data, provide some observations about the cemented implants.

- The average cemented implant starts at ~0.3 MTPM at 0 months and goes to ~0.5 at ~24 months
- The MTPM growth appears to be logarithmic
- The MTPM lines are concentrated in the 0.1-0.5 mm range

b. Using your longitudinal plots for uncemented data, provide some observations about the uncemented implants

- The average uncemented implant starts at ~0.7 MTPM at 0 months and goes to ~1.0 MTPM at ~24 months
- The MTPM growth appears to be logarithmic
- The MTPM lines are concentrated in the 0.1-1.0 mm range and are fairly sparse between 1.0-4.5 mm

c. Discuss any differences between the two types of fixation.

- Cemented fixations have lower average MTPM (~0.5 mm) over time
- Cemented fixation MTPM lines are ‘denser’, especially near the lower values
- Both fixations have a similar MTPM growth pattern according to the mean line