

Data Wrangling

2024-08-02

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.1      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

```
# Set a random seed for reproducibility
set.seed(42)

# Number of samples
n <- 10000

# Generate data frame for required variables
randomdata <- data.frame(
  'Age' = trunc(runif(n, min = 18, max = 36)), # Age uniformly distributed between 18 and 35
  'InfantSex' = factor(rbinom(n, size = 1, prob = 0.5), labels = c("Male", "Female")) # Infant sex with 50% probability
)

# Generate Glucose1 and Glucose2 based on InfantSex
randomdata$Glucose1 <- ifelse(randomdata$InfantSex == "Male",
  rnorm(n, mean = 85, sd = 6), #normalized distribution
  rnorm(n, mean = 80, sd = 6))
randomdata$Glucose2 <- ifelse(randomdata$InfantSex == "Male",
  rnorm(n, mean = 165, sd = 9),
  rnorm(n, mean = 155, sd = 9))
randomdata$Diagnosis <- ifelse(randomdata$Glucose1 > 95 | randomdata$Glucose2 > 180, #define diagnosis
  "Gestational Diabetes", "Healthy")

# Subset the data for male infants
# Subset using https://www.statmethods.net/management/subset.html
male_data <- subset(randomdata, InfantSex == "Male")
# Subset the data for female infants
female_data <- subset(randomdata, InfantSex == "Female")

#Male infants
print("Summary for Male Infants")
```

```
## [1] "Summary for Male Infants"
```

```
summary(male_data)
```

```
##      Age      InfantSex      Glucose1      Glucose2
## Min.   :18.00   Male   :5000   Min.    : 64.28   Min.    :132.4
## 1st Qu.:22.00   Female:  0   1st Qu.: 80.91   1st Qu.:158.7
## Median :27.00                      Median : 85.02   Median :165.0
## Mean   :26.47                      Mean    : 85.01   Mean    :164.9
## 3rd Qu.:31.00                      3rd Qu.: 89.08   3rd Qu.:171.1
## Max.   :35.00                      Max.    :106.22   Max.    :199.9
##   Diagnosis
## Length:5000
## Class :character
## Mode  :character
##
##
##
```

```
#Female Infants
```

```
print("Summary for Female Infants")
```

```
## [1] "Summary for Female Infants"
```

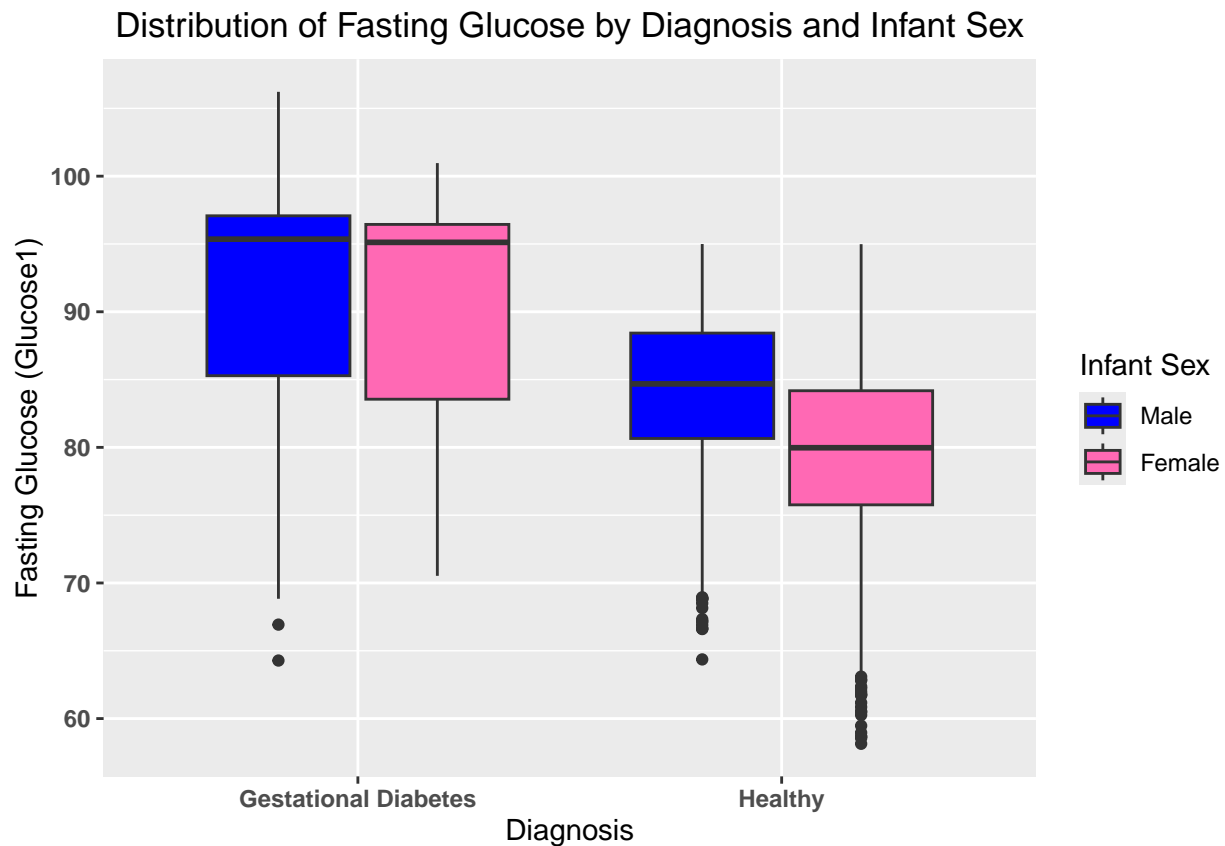
```
summary(female_data)
```

```
##      Age      InfantSex      Glucose1      Glucose2
## Min.   :18.0   Male    :  0   Min.    : 58.14   Min.    :123.2
## 1st Qu.:22.0   Female:5000   1st Qu.: 75.78   1st Qu.:148.9
## Median :26.0                      Median : 80.01   Median :154.9
## Mean   :26.5                      Mean    : 79.95   Mean    :154.9
## 3rd Qu.:31.0                      3rd Qu.: 84.24   3rd Qu.:160.9
## Max.   :35.0                      Max.    :100.96   Max.    :187.6
##   Diagnosis
## Length:5000
## Class :character
## Mode  :character
##
##
##
```

```
color_palette <- c("Male" = "blue", "Female" = "hotpink")
```

```
ggplot(randomdata, aes(x = Diagnosis, y = Glucose1, fill = InfantSex)) +
  geom_boxplot() +
  scale_fill_manual(values = color_palette) + # Apply the custom color palette
  labs(title = "Distribution of Fasting Glucose by Diagnosis and Infant Sex",
       x = "Diagnosis",
       y = "Fasting Glucose (Glucose1)",
       fill = "Infant Sex") +
  theme_gray() + #adds gray background
```

```
theme(
  plot.title = element_text(hjust = 0.5),
  axis.text = element_text(face = "bold"))
```



```
randomdata$Subject <- 1:n #adding in Subject to call
```

```
longData <- randomdata %>%
  pivot_longer( #https://tidyr.tidyverse.org/reference/pivot_longer.html - additional explanation
    cols = c(Glucose1, Glucose2),
    names_to = "Timepoint",
    values_to = "Glucose") %>%
  mutate(Timepoint = ifelse(Timepoint == "Glucose1", "Baseline", "One Hour"))
```

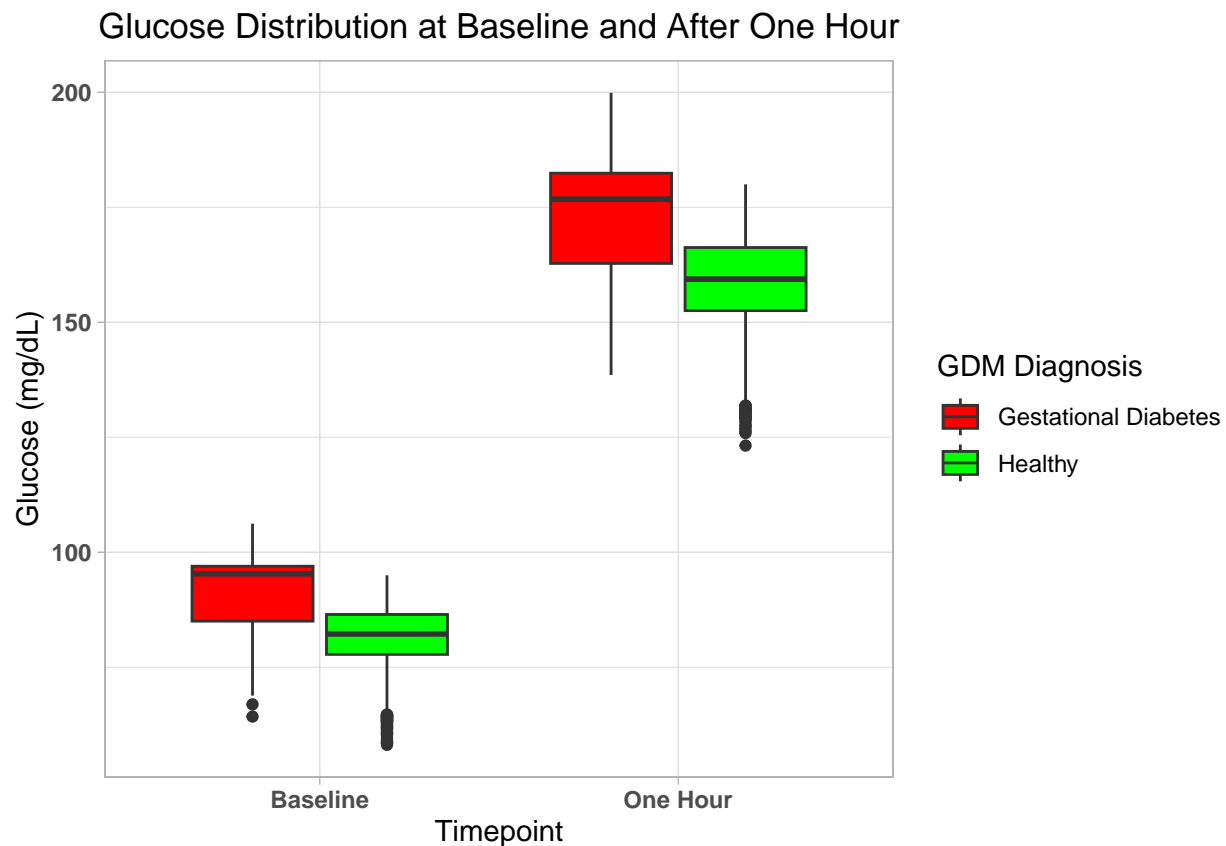
```
print(longData[longData$Subject == 1, ]) #will have 2 outputs, one for Baseline (baseline) and one for
```

```
## # A tibble: 2 x 6
##   Age InfantSex Diagnosis Subject Timepoint Glucose
##   <dbl> <fct>    <chr>      <int> <chr>      <dbl>
## 1    34 Female    Healthy        1 Baseline    76.4
## 2    34 Female    Healthy        1 One Hour    157.
```

```
color_scale <- c("Healthy" = "green", "Gestational Diabetes" = "red") #another color palette
```

```
ggplot(longData, aes(x = Timepoint, y = Glucose, fill = Diagnosis)) +
```

```
geom_boxplot() +
scale_fill_manual(values = color_scale) + # Apply the custom color palette
labs(
  x = "Timepoint",
  y = "Glucose (mg/dL)",
  fill = "GDM Diagnosis",
  title = "Glucose Distribution at Baseline and After One Hour") +
theme_light() + #light theme
theme(
  plot.title = element_text(hjust = 0.5),
  axis.text = element_text(face = "bold"))
```



```
library(ggpubr)
```

```
# Define color palette
color_plot <- c("Baseline" = "orange", "One Hour" = "purple")

# Function to calculate mean and standard deviation for each group
# used https://www.carlislerainey.com/teaching/pols-209/files/notes-10-average-sd-r.pdf
calc_stats <- function(data) {
  data %>%
    group_by(Timepoint) %>% #find groups of interest
    summarise( #https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/summarise
      Mean = mean(Glucose),
      SD = sd(Glucose))
}
```

```

    )
  }

# Calculate statistics for female and male infants
female_stats <- calc_stats(subset(longData, InfantSex == "Female")) #store statistics for female
male_stats <- calc_stats(subset(longData, InfantSex == "Male")) #store statistics for male

# Add text annotation for female_plot
female_plot <- ggplot(subset(longData, InfantSex == "Female"), aes(x = Age, y = Glucose, color = Timepoint)) +
  geom_point() +
  scale_color_manual(values = color_plot) +
  labs(
    title = "Mothers of Female Infants",
    x = "Maternal Age (yrs)",
    y = "Glucose (mg/dL)",
    color = "Timepoint"
  ) +
  theme_light() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(face = "bold")
  ) +
  #How to label plots:
  #https://ggplot2.tidyverse.org/reference/geom_text.html
  #https://www.rdocumentation.org/packages/ggplot2/versions/0.9.1/topics/geom_text
  #https://r-graph-gallery.com/275-add-text-labels-with-ggplot2.html
  geom_text(data = female_stats %>% filter(Timepoint == "Baseline"),
    aes(x = position_Baseline_female[1], y = position_Baseline_female[2], #defining so the move
      label = sprintf("Baseline: Mean = %.1f (SD = %.1f)", Mean, SD)), #label is what text yo
    color = color_plot["Baseline"],
    size = 3) + #have to make text smaller to be seen on combined graph
  geom_text(data = female_stats %>% filter(Timepoint == "One Hour"),
    aes(x = position_onehour_female[1], y = position_onehour_female[2],
      label = sprintf("One Hour: Mean = %.1f (SD = %.1f)", Mean, SD)),
    color = color_plot["One Hour"],
    size = 3)

# How you move the annotation around the graph, adjusted to the VALUES on the table, not pixels
position_Baseline_female <- c(x = 26.5, y = 110)
position_onehour_female <- c(x = 26.5, y = 120)

# Add text annotation for male_plot, similar plot and notes to above just altered for male infant sex
male_plot <- ggplot(subset(longData, InfantSex == "Male"), aes(x = Age, y = Glucose, color = Timepoint)) +
  geom_point() +
  scale_color_manual(values = color_plot) +
  labs(
    title = "Mothers of Male Infants",
    x = "Maternal Age (yrs)",
    y = "Glucose (mg/dL)",
    color = "Timepoint"
  ) +
  theme_light() +

```

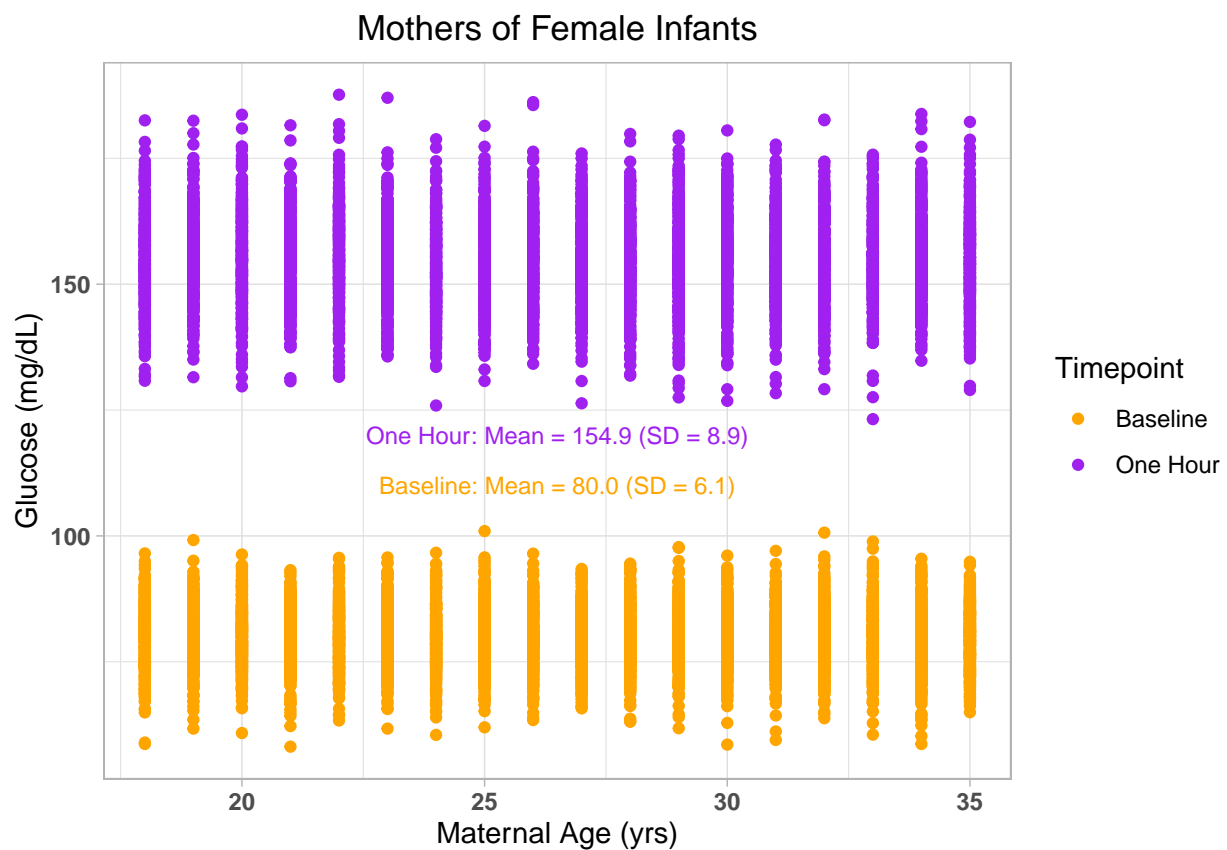
```

theme(
  plot.title = element_text(hjust = 0.5),
  axis.text = element_text(face = "bold")
) +
geom_text(data = male_stats %>% filter(Timepoint == "Baseline"),
  aes(x = position_Baseline_male[1], y = position_Baseline_male[2], #defining so the movement
    label = sprintf("Baseline: Mean = %.1f (SD = %.1f)", Mean, SD)), #label is what text yo
    color = color_plot["Baseline"],
    size = 3) +
geom_text(data = male_stats %>% filter(Timepoint == "One Hour"),
  aes(x = position_onehour_male[1], y = position_onehour_male[2],
    label = sprintf("One Hour: Mean = %.1f (SD = %.1f)", Mean, SD)),
    color = color_plot["One Hour"],
    size = 3)

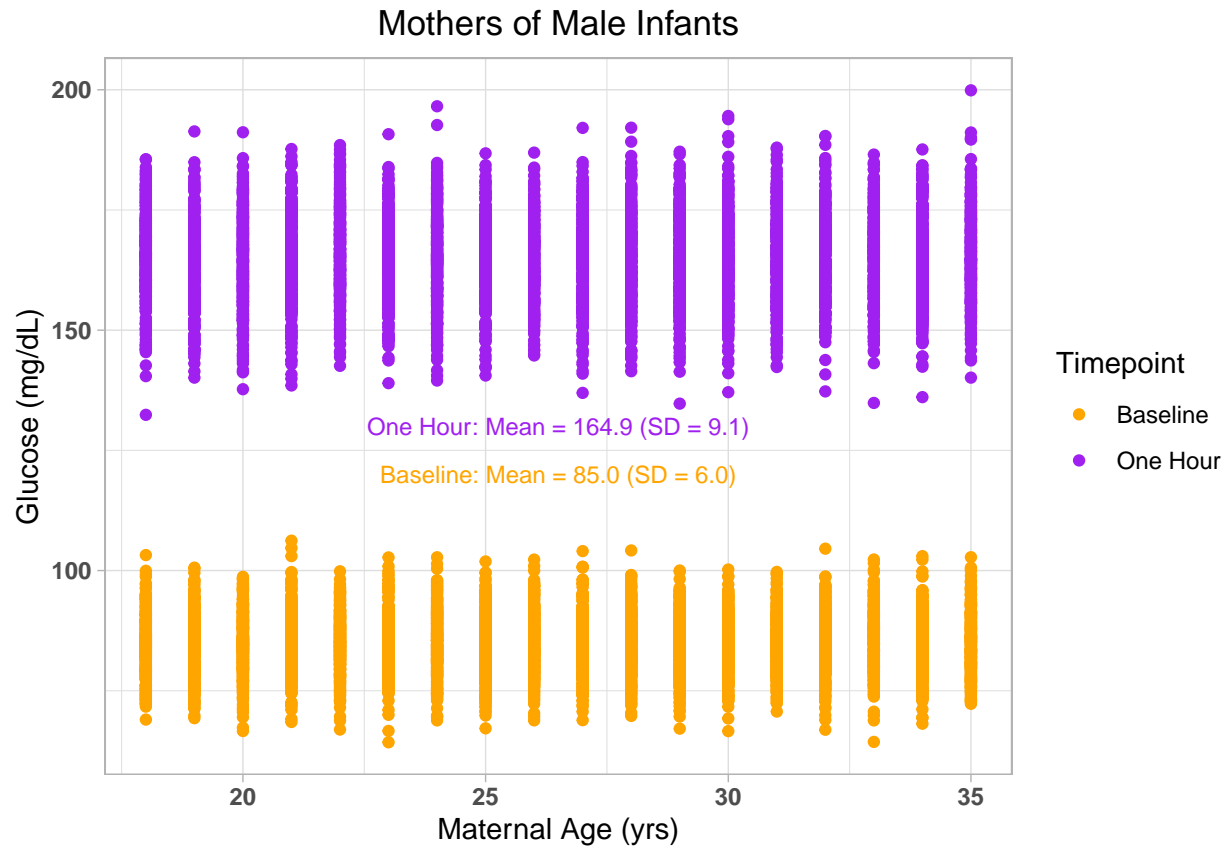
# How you move the annotation around the graph, adjusted to the VALUES on the table, not pixels
position_Baseline_male <- c(x = 26.5, y = 120)
position_onehour_male <- c(x = 26.5, y = 130)

# Print plots
print(female_plot)

```

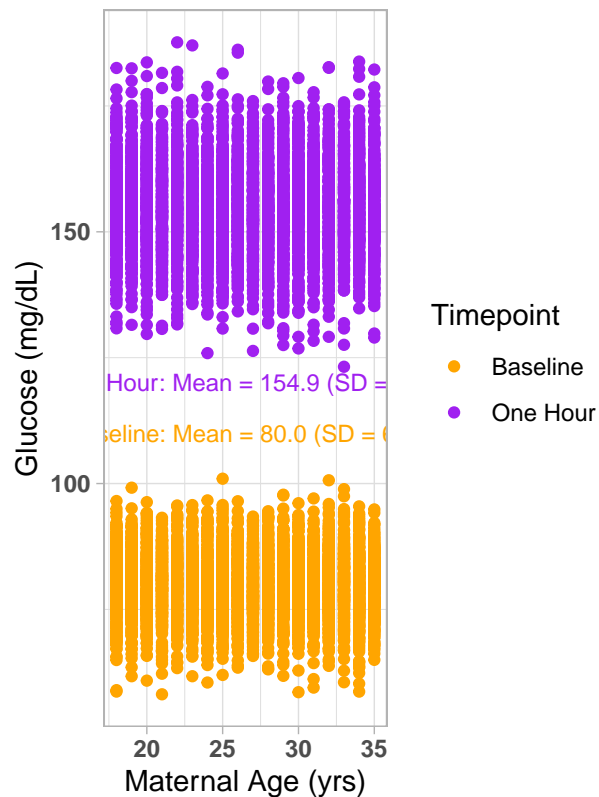


```
print(male_plot)
```

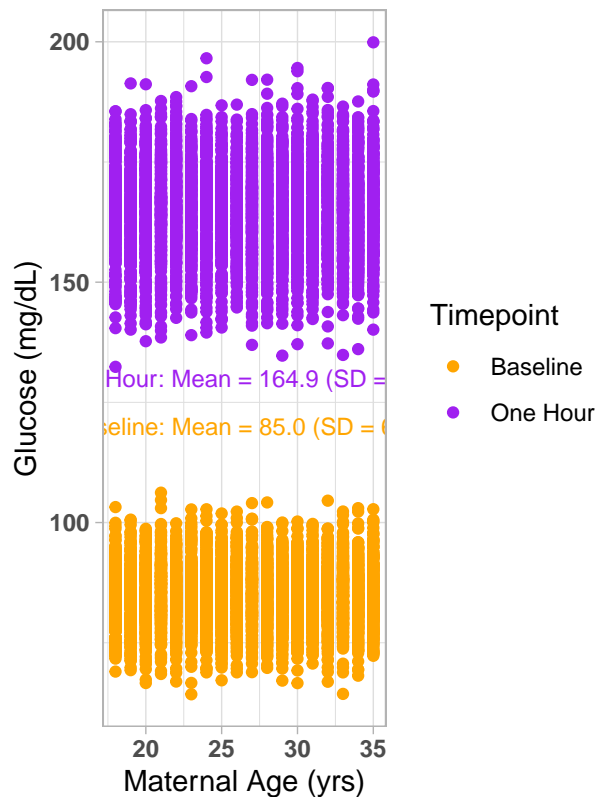


```
# Arrange the plots using https://www.rdocumentation.org/packages/ggpubr/versions/0.6.0/topics/ggarrange
ggarrange(female_plot, male_plot,
  labels = c("A", "B"),
  ncol = 2, nrow = 1) #two plots right next to each other, makes 2 columns and 1 row
```

A Mothers of Female Infants



B Mothers of Male Infants



```
table_wide <- randomdata %>%
  group_by(InfantSex, Diagnosis) %>% #Healthy Female/Gestational Diabetes Female
  summarise( #calculations for material i want included in the table
    Mean_Age = mean(Age),
    Mean_Fasting_Glucose = mean(Glucose1),
    SD_Fasting_Glucose = sd(Glucose1),
    Mean_One_Hour_Glucose = mean(Glucose2),
    SD_One_Hour_Glucose = sd(Glucose2),
  ) %>%
  # Combine Diagnosis and Infant sex
  mutate(Group = paste(Diagnosis, InfantSex)) %>%
  # Ensure the rows are in the required order https://www.rdocumentation.org/packages/dplyr/versions/1.
  arrange(factor(Group, levels = c(
    "Healthy Female",
    "Gestational Diabetes Female",
    "Healthy Male",
    "Gestational Diabetes Male"
  ))) %>%
  # Selecting the columns I want included in the table
  select(
    Group,
    Mean_Age,
    Mean_Fasting_Glucose,
    SD_Fasting_Glucose,
    Mean_One_Hour_Glucose,
    SD_One_Hour_Glucose
  )
```



```

)

## 'summarise()' has grouped output by 'InfantSex'. You can override using the
## '.groups' argument.
## Adding missing grouping variables: 'InfantSex'

# Print the summary table
print(table_wide)

## # A tibble: 4 x 7
## # Groups:   InfantSex [2]
##   InfantSex Group          Mean_Age Mean_Fasting_Glucose SD_Fasting_Glucose
##   <fct>      <chr>          <dbl>          <dbl>          <dbl>
## 1 Female    Healthy Female      26.5            79.9            6.03
## 2 Female    Gestational Diabet~  26.4            89.8            8.72
## 3 Male      Healthy Male          26.5            84.4            5.43
## 4 Male      Gestational Diabet~  26.5            91.5            7.81
## # i 2 more variables: Mean_One_Hour_Glucose <dbl>, SD_One_Hour_Glucose <dbl>

```