

# Slide 15 - R Output

Elvan Ceyhan

11/25/2024

```
#Set Working Directory to Source File Location  
library("rstudioapi") # Load rstudioapi package  
#setwd(dirname(getActiveDocumentContext()$path)) # Set working directory to source file location  
#getwd()
```

Some required packages:

```
# Load packages  
library(rstan)  
library(bayesrules) # For Bayesian statistical modeling  
library(tidyverse) # For data manipulation and visualization  
library(rstanarm) # For Bayesian regression models  
library(broom.mixed) # For tidying model outputs  
library(ggplot2)
```

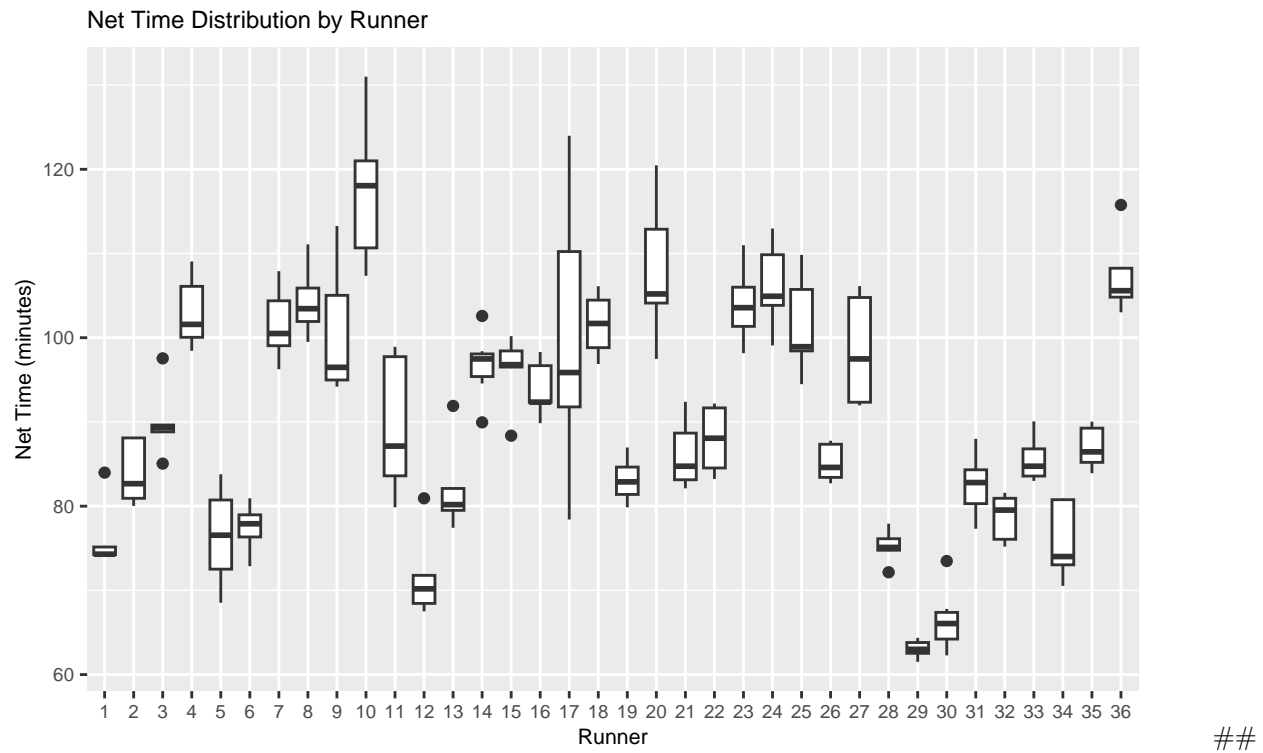
## Cherry Blossom Data Analysis:

This script performs exploratory data analysis and Bayesian regression modeling on the Cherry Blossom dataset to examine the relationship between age and race net time across runners.

```
# Load the dataset and extract relevant columns  
data(cherry_blossom_sample)  
running <- cherry_blossom_sample %>%  
  select(runner, age, net) # Keep only runner ID, age, and net time  
cat("Number of observations in the dataset:", nrow(running), "\n")
```

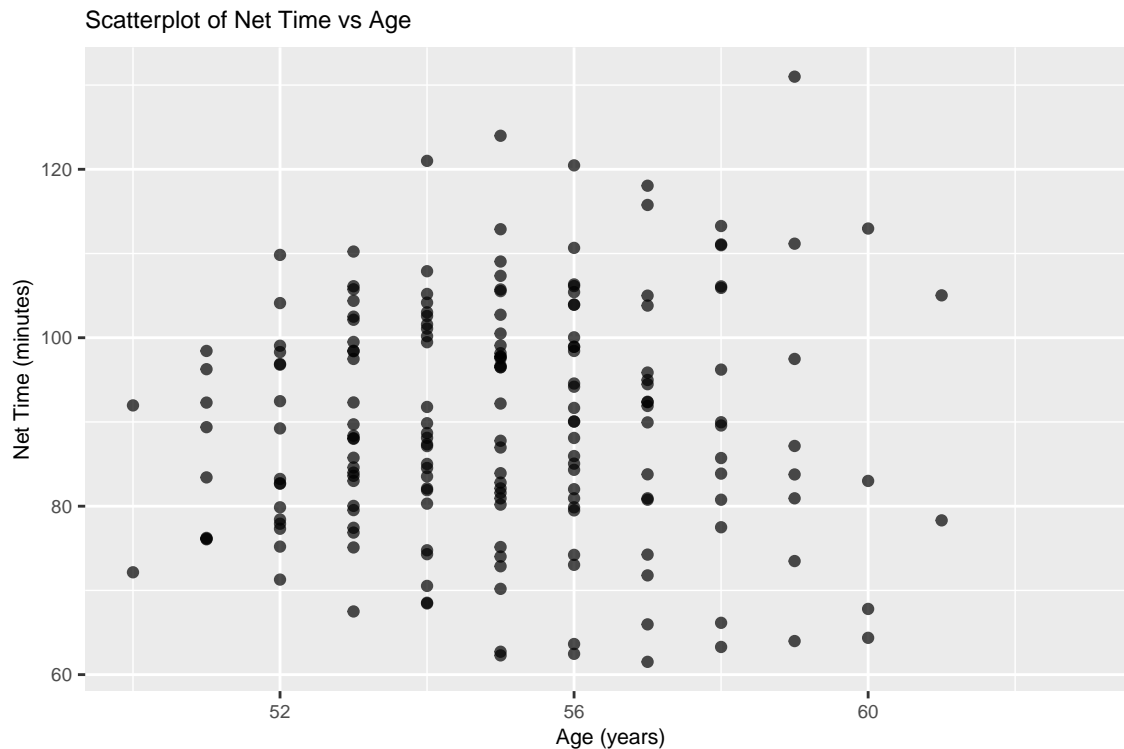
```
## Number of observations in the dataset: 252
```

```
# Boxplot of net times across runners  
ggplot(running, aes(x = runner, y = net)) +  
  geom_boxplot() +  
  labs(title = "Net Time Distribution by Runner",  
        x = "Runner", y = "Net Time (minutes)")
```



Complete pooling model: Examining the relationship between age and net time

```
# Scatterplot of net time against age
ggplot(running, aes(x = age, y = net)) +
  geom_point(alpha = 0.7) + # Add transparency for better visualization
  labs(title = "Scatterplot of Net Time vs Age",
        x = "Age (years)", y = "Net Time (minutes)")
```



Bayesian Regression Model:  $Y = \text{net time}$ ,  $X = \text{age}$

```
complete_pooled_model <- stan_glm(
  net ~ age,
  data = running,
  family = gaussian,
  prior_intercept = normal(0, 2.5, autoscale = TRUE),
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_aux = exponential(1, autoscale = TRUE),
  chains = 4, iter = 10000, seed = 84735 # Adjusted iterations for clarity
)

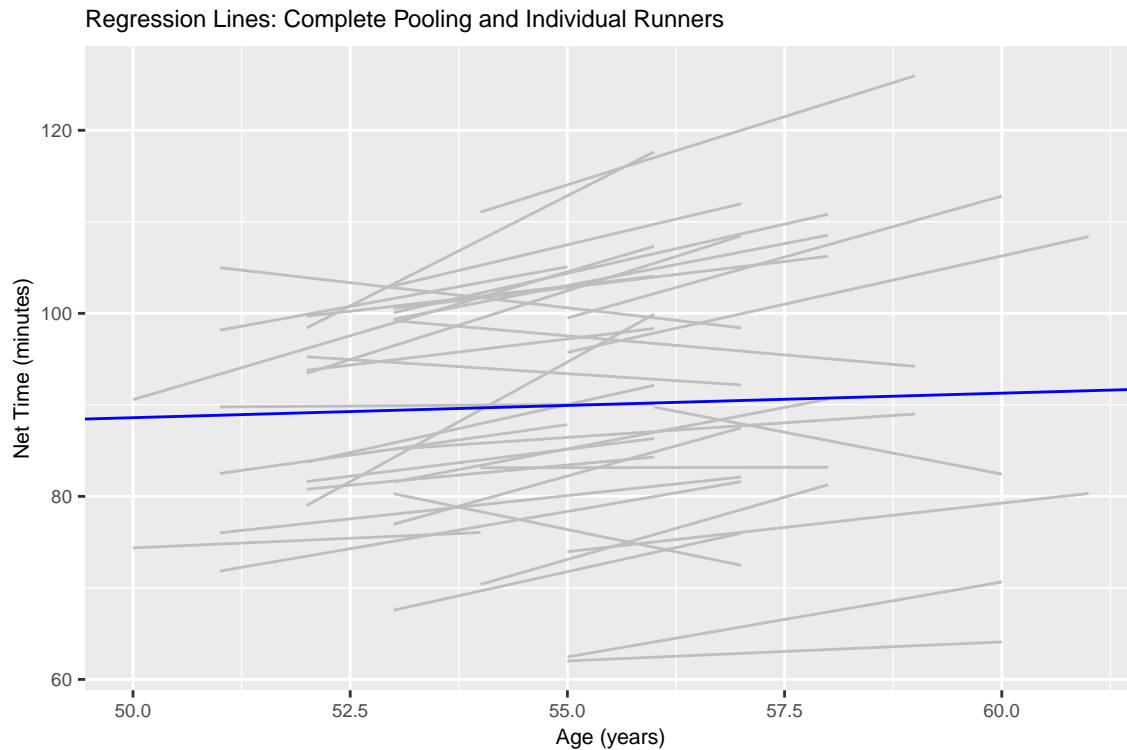
# Summarize posterior statistics with 80% credible intervals
pooled_summary <- tidy(complete_pooled_model, conf.int = TRUE, conf.level = 0.80)
print(pooled_summary)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error conf.low conf.high
##   <chr>      <dbl>    <dbl>   <dbl>   <dbl>
## 1 (Intercept)  75.2      24.6    43.7    106.
## 2 age          0.268     0.446   -0.302    0.842
```

```
# Plotting regression line for the pooled model
ggplot(running, aes(x = age, y = net, group = runner)) +
  geom_smooth(method = "lm", se = FALSE, color = "gray", linewidth = 0.5) +
  geom_abline(aes(intercept = pooled_summary$estimate[1],
```

```
slope = pooled_summary$estimate[2]), color = "blue") +
labs(title = "Regression Lines: Complete Pooling and Individual Runners",
x = "Age (years)", y = "Net Time (minutes)")
```

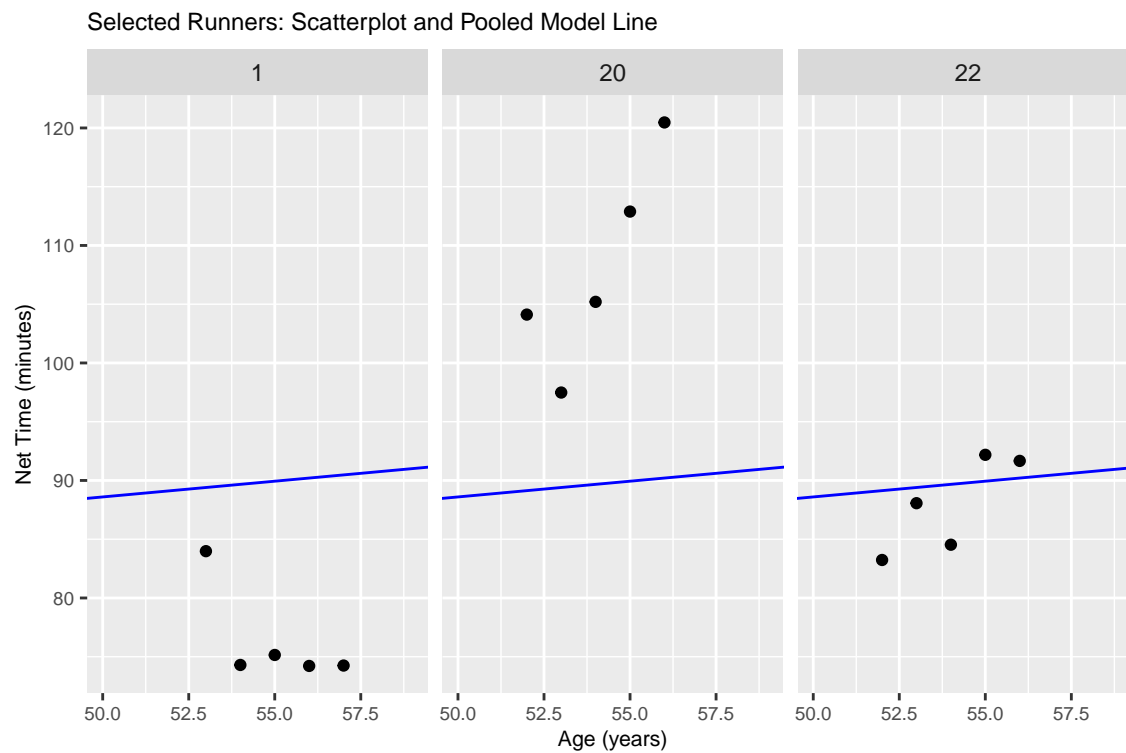
```
## 'geom_smooth()' using formula = 'y ~ x'
```



Analyze data for selected runners: Runners 1, 20, and 22

```
# Filter data for specific runners
examples <- running %>%
  filter(runner %in% c("1", "20", "22"))

# Scatterplot with regression line for selected runners
ggplot(examples, aes(x = age, y = net)) +
  geom_point() +
  facet_wrap(~ runner) +
  geom_abline(aes(intercept = pooled_summary$estimate[1],
                  slope = pooled_summary$estimate[2]), color = "blue") +
  labs(title = "Selected Runners: Scatterplot and Pooled Model Line",
x = "Age (years)", y = "Net Time (minutes)")
```



```
# No pooling model: Regression lines for individual runners
ggplot(examples, aes(x = age, y = net)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, fullrange = TRUE) +
  facet_wrap(~ runner) +
  xlim(52, 62) +
  labs(title = "Individual Runner Regression Lines (No Pooling)",
       x = "Age (years)", y = "Net Time (minutes)")
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

