---

title: "House of Excellence"

output-file: titulo do projeto

---

# ANALISE 1

{r}

library(kableExtra)

library(DT)

library(knitr)

library(pacman)

library(tidyverse)

library(ggplot2)

library(dplyr)

library(readxl)

library(plotly)

athenas = read\_excel("OlimpiadasPS.xlsx",

sheet = "Athina")

rj = read\_excel("OlimpiadasPS.xlsx",

sheet = "Rio de Janeiro")

londres = read\_excel("OlimpiadasPS.xlsx",

sheet = "London")

sydney = read\_excel("OlimpiadasPS.xlsx",

sheet = "Sydney")

pequim = read\_excel("OlimpiadasPS.xlsx",

sheet = "Beijing")

# Organizando o banco de dados

colnames(sydney) = c("Nome", "Sexo", "Idade", "Altura", "Peso", "Pais", "Esporte", "Evento", "Medalha")

colnames(rj) = c("Nome", "Sexo", "Idade", "Altura", "Peso", "Pais", "Esporte", "Evento", "Medalha")

colnames(londres) = c("Nome", "Sexo", "Idade", "Altura", "Peso", "Pais", "Esporte", "Evento", "Medalha")

colnames(pequim) = c("Nome", "Sexo", "Idade", "Altura", "Peso", "Pais", "Esporte", "Evento", "Medalha")

colnames(athenas) = c("Nome", "Sexo", "Idade", "Altura", "Peso", "Pais", "Esporte", "Evento", "Medalha")

primeiro = rbind(athenas,rj,londres,sydney,pequim)

olimpiadasG = subset(primeiro, Medalha== "Gold")

olimpiadasS = subset(primeiro, Medalha== "Silver")

olimpiadasB = subset(primeiro, Medalha== "Bronze")

olimpiadas = rbind(olimpiadasG,olimpiadasS,olimpiadasB)

estat\_colors <- c(

"#A11D21", "#003366", "#CC9900",

"#663333", "#FF6600", "#CC9966",

"#999966", "#006606", "#008091",

"#041835", "#666666" )

theme\_estat <- function(...) {

theme <- ggplot2::theme\_bw() +

ggplot2::theme(

axis.title.y = ggplot2::element\_text(colour = "black", size = 12),

axis.title.x = ggplot2::element\_text(colour = "black", size = 12),

axis.text = ggplot2::element\_text(colour = "black", size = 9.5)

,

panel.border = ggplot2::element\_blank(),

axis.line = ggplot2::element\_line(colour = "black"),

legend.position = "top",

...

)

return(

list(

theme,

scale\_fill\_manual(values = estat\_colors),

scale\_colour\_manual(values = estat\_colors)

)

)

}

{r}

library(dplyr)

library(ggplot2)

medalW = subset(olimpiadas, Sexo == "F")

medalW$Pais[medalW$Pais == "United States"] <- "Estados Unidos"

medalW$Pais[medalW$Pais == "Germany"] <- "Alemanha"

atletasW <- medalW %>%

distinct(Nome, .keep\_all = TRUE) %>%

group\_by(Pais) %>%

summarise(Numero\_Atletas = n())

top5\_atletas <- atletasW %>%

arrange(desc(Numero\_Atletas)) %>%

head(5)

total\_atletas <- sum(atletasW$Numero\_Atletas)

top5\_atletas <- top5\_atletas %>%

mutate(Porcentagem = round(Numero\_Atletas / total\_atletas \* 100, 1))

ggplot(top5\_atletas, aes(x = reorder(Pais, -Numero\_Atletas), y = Numero\_Atletas)) +

geom\_bar(stat = "identity", fill = "#A11D21") +

geom\_text(aes(label = paste0(Numero\_Atletas, " (", Porcentagem, "%)")),

vjust = -0.5, size = 2.5) +

labs(title = "Histograma sobre Top 5 países com mais atletas medalhistas únicos entre as mulheres",

x = "Países", y = "Número de Atletas Medalhistas Únicos") +

scale\_y\_continuous(limits = c(0, 600), breaks = seq(0, 600, by = 200))+

theme\_estat()

# ANALISE 2

{r}

olimpiadas$Peso <- as.numeric(olimpiadas$Peso)

olimpiadas$pesokg = olimpiadas$Peso \* 0.453592

olimpiadas$alturam = olimpiadas$Altura / 100

olimpiadas$imc <- olimpiadas$pesokg / (olimpiadas$alturam^2)

judo = subset(olimpiadas, Esporte == "Judo")

badminton = subset(olimpiadas, Esporte == "Badminton")

futebol = subset(olimpiadas, Esporte == "Football")

ginastica = subset(olimpiadas, Esporte == "Gymnastics")

atletismo = subset(olimpiadas, Esporte == "Athletics")

espotes=rbind(judo,badminton,futebol,ginastica,atletismo)

espotes$Esporte[espotes$Esporte == "Football"] = "Futebol"

espotes$Esporte[espotes$Esporte == "Gymnastics"] = "Ginastica"

espotes$Esporte[espotes$Esporte == "Athletics"] = "Atletismo"

library(forcats)

ggplot(espotes) +

aes(x = fct\_reorder(Esporte, imc, .fun = median), y = imc) + # Organizando pela mediana

geom\_boxplot(fill = c("#A11D21"), width = 0.5) +

stat\_summary(

fun = "mean", geom = "point", shape = 23, size = 3, fill = "white"

) +

labs(x = "Esporte", y = "IMC") +

ggtitle("Boxplot sobre os IMCs no esporte") +

theme\_estat()

```{r}

print\_quadro\_resumo <- function(data, var\_name, title = "Medidas resumo da(o) [nome da variável]", label = "quad:quadro\_resumo1") {

var\_name <- substitute(var\_name)

data\_clean <- data %>%

filter(!is.na(!!sym(var\_name)) & !!sym(var\_name) != Inf)

data\_clean <- data\_clean %>%

summarize(

`Média` = round(mean(!!sym(var\_name), na.rm = TRUE), 2),

`Desvio Padrão` = round(sd(!!sym(var\_name), na.rm = TRUE), 2),

`Variância` = round(var(!!sym(var\_name), na.rm = TRUE), 2),

`Mínimo` = round(min(!!sym(var\_name), na.rm = TRUE), 2),

`1º Quartil` = round(quantile(!!sym(var\_name), probs = 0.25, na.rm = TRUE), 2),

`Mediana` = round(quantile(!!sym(var\_name), probs = 0.5, na.rm = TRUE), 2),

`3º Quartil` = round(quantile(!!sym(var\_name), probs = 0.75, na.rm = TRUE), 2),

`Máximo` = round(max(!!sym(var\_name), na.rm = TRUE), 2)

) %>%

t() %>%

as.data.frame() %>%

rownames\_to\_column()

latex <- str\_c("\\begin{quadro}[H]\n",

"\\caption{", title, "}\n",

"\\centering\n",

"\\begin{adjustbox}{max width=\\textwidth}\n",

"\\begin{tabular}{", sep = "")

col\_count <- ncol(data\_clean)

row\_count <- nrow(data\_clean)

latex <- str\_c(latex, "| l |\n", sep = "")

for (i in seq(2, col\_count)) {

numCount <- data\_clean[i, -c(1)] %>%

as.numeric() %>%

{floor(log10(.)) + 1} %>%

max()

latex <- str\_c(latex, "\t\t\tS[table-format = ", numCount, ".2]\n", sep = "")

}

latex <- str\_c(latex, "\t\t\t|}\n\t\\toprule\n\t\t", sep = "")

if (col\_count > 2) {

for (i in seq(1, col\_count)) {

if (i == 1)

latex <- str\_c(latex, "\\textbf{Estatística}", sep = "")

else

latex <- str\_c(latex, " \\textbf{", data\_clean[1, i], "}", sep = "")

if (i < col\_count)

latex <- str\_c(latex, "&", sep = "")

else

latex <- str\_c(latex, "\\\\\n", sep = "")

}

} else {

latex <- str\_c(latex, "\\textbf{Estatística} & \\textbf{Valor} \\\\\n", sep = "")

}

latex <- str\_c(latex, "\t\t\\midrule\n", sep = "")

starting\_number <- ifelse(col\_count > 2, 2, 1)

for (i in seq(starting\_number, row\_count)) {

latex <- str\_c(latex, "\t\t", str\_flatten(t(data\_clean[i, ]), collapse = " & "), " \\\\\n")

}

latex <- str\_c(latex, "\t\\bottomrule\n",

"\t\\end{tabular}\n",

"\t\\label{", label, "}\n",

"\t\\end{adjustbox}\n",

"\\end{quadro}", sep = "")

writeLines(latex)

}

espotes %>%

print\_quadro\_resumo(var\_name = imc, title = "Medidas resumo do imc dos atletas")

```

{r, echo=FALSE, results='asis'}

cat("\\begin{quadro}[H]

\\caption{Medidas resumo dos esportes}

\\centering

\\begin{tabular}{| l |

S[table-format = 2.2]

S[table-format = 1.2]

S[table-format = 2.2]

S[table-format = 2.2]

S[table-format = 2.2]

|}

\\toprule

\\textbf{Estatística} & \\textbf{Ginastica} & \\textbf{Atletismo} & \\textbf{Badminton} & \\textbf{Futebol} & \\textbf{Judo} \\\\

\\midrule

Média & 20,68 & 22,30 & 22,21 & 22,51 & 25,70 \\\\

Desvio Padrão & 2,38 & 3,86 & 1,50 & 1,73 & 5,12 \\\\

Variância & 5,67 & 14,92 & 2,26 & 2,99 & 26,23 \\\\

Mínimo & 15,16 & 15,82 & 18,94 & 16,73 & 18,52 \\\\

1º Quartil & 18,61 & 20,03 & 21,22 & 21,34 & 22,06 \\\\

Mediana & 21,09 & 21,45 & 22,28 & 22,49 & 24,68 \\\\

3º Quartil & 22,48 & 23,67 & 23,21 & 23,71 & 27,70 \\\\

Máximo & 26,45 & 44,38 & 26,73 & 29,07 & 56,50\\\\

\\bottomrule

\\end{tabular}

\\label{quad:quadro\_resumo1}

\\end{quadro}

")

# ANALISE 3

{r}

olimpiadas$Medalha[olimpiadas$Medalha == "Gold"] = "Ouro"

olimpiadas$Medalha[olimpiadas$Medalha == "Silver"] = "Prata"

olimpiadas$Nome[olimpiadas$Nome == "Michael Fred Phelps, II"] = "Michael Phelps"

olimpiadas$Nome[olimpiadas$Nome == "Natalie Anne Coughlin (-Hall)"] = "Natalie Anne Coughlin"

ouro = subset(olimpiadas, Medalha == "Ouro")

prata = subset(olimpiadas, Medalha == "Prata")

bronze = subset(olimpiadas, Medalha == "Bronze")

medalhaalltime = rbind(ouro,prata,bronze)

michaelphelps = subset(medalhaalltime,Nome == "Michael Phelps")

natalie = subset(medalhaalltime, Nome == "Natalie Anne Coughlin")

ryan = subset(medalhaalltime,Nome == "Ryan Steven Lochte")

top3= rbind(michaelphelps, natalie, ryan)

{r}

library(dplyr)

library(forcats)

library(ggplot2)

library(stringr)

trans\_drv <- top3 %>%

mutate(Nome = case\_when(

str\_detect(Nome, "auto") ~ "auto",

str\_detect(Nome, "manual") ~ "manual",

TRUE ~ Nome

)) %>%

group\_by(Nome, Medalha) %>%

summarise(freq = n(), .groups = "drop") %>%

mutate(

Medalha = fct\_relevel(Medalha, "Ouro", "Prata", "Bronze"),

freq\_relativa = round(freq / sum(freq) \* 100, 1)

) %>%

group\_by(Nome) %>%

mutate(total\_medalhas = sum(freq)) %>%

ungroup() %>%

mutate(Nome = fct\_reorder(Nome, total\_medalhas, .desc = TRUE))

porcentagens <- str\_c(trans\_drv$freq\_relativa, "%") %>% str\_replace("\\.", ",")

legendas <- str\_squish(str\_c(trans\_drv$freq, " (", porcentagens, ")"))

ggplot(trans\_drv) +

aes(

x = Nome,

y = freq,

fill = Medalha,

label = legendas

) +

geom\_col(position = position\_dodge2(preserve = "single", padding = 0)) +

geom\_text(

position = position\_dodge(width = 0.9),

vjust = -0.5, hjust = 0.5,

size = 3

) +

labs(x = "Nome do Atleta", y = "Quantidade de Medalhas") +

ggtitle("Histograma sobre a Quantidade de Medalhas por Atleta e Tipo de Medalha") +

scale\_y\_continuous(limits = c(0, 25), breaks = seq(0, 49, by = 10))+

theme\_estat()

# ANALISE 4

{r}

ggplot(olimpiadas, aes(x = Altura, y = pesokg)) +

geom\_point(color = "#A11D21", alpha = 0.2) +

labs(

title = "Gráfico de Dispersão Altura pelo Peso",

x = "Altura",

y = "Peso",

color = "Legenda"

) +

theme\_estat()

correlacao <- cor(olimpiadas$pesokg, olimpiadas$Altura, use = "complete.obs")

```{r}

print\_quadro\_resumo <- function(data, var\_name, title = "Medidas resumo da(o) [nome da variável]", label = "quad:quadro\_resumo1") {

var\_name <- substitute(var\_name)

data\_clean <- data %>%

filter(!is.na(!!sym(var\_name)) & !!sym(var\_name) != Inf)

data\_clean <- data\_clean %>%

summarize(

`Média` = round(mean(!!sym(var\_name), na.rm = TRUE), 2),

`Desvio Padrão` = round(sd(!!sym(var\_name), na.rm = TRUE), 2),

`Variância` = round(var(!!sym(var\_name), na.rm = TRUE), 2),

`Mínimo` = round(min(!!sym(var\_name), na.rm = TRUE), 2),

`1º Quartil` = round(quantile(!!sym(var\_name), probs = 0.25, na.rm = TRUE), 2),

`Mediana` = round(quantile(!!sym(var\_name), probs = 0.5, na.rm = TRUE), 2),

`3º Quartil` = round(quantile(!!sym(var\_name), probs = 0.75, na.rm = TRUE), 2),

`Máximo` = round(max(!!sym(var\_name), na.rm = TRUE), 2)

) %>%

t() %>%

as.data.frame() %>%

rownames\_to\_column()

latex <- str\_c("\\begin{quadro}[H]\n",

"\\caption{", title, "}\n",

"\\centering\n",

"\\begin{adjustbox}{max width=\\textwidth}\n",

"\\begin{tabular}{", sep = "")

col\_count <- ncol(data\_clean)

row\_count <- nrow(data\_clean)

latex <- str\_c(latex, "| l |\n", sep = "")

for (i in seq(2, col\_count)) {

numCount <- data\_clean[i, -c(1)] %>%

as.numeric() %>%

{floor(log10(.)) + 1} %>%

max()

latex <- str\_c(latex, "\t\t\tS[table-format = ", numCount, ".2]\n", sep = "")

}

latex <- str\_c(latex, "\t\t\t|}\n\t\\toprule\n\t\t", sep = "")

if (col\_count > 2) {

for (i in seq(1, col\_count)) {

if (i == 1)

latex <- str\_c(latex, "\\textbf{Estatística}", sep = "")

else

latex <- str\_c(latex, " \\textbf{", data\_clean[1, i], "}", sep = "")

if (i < col\_count)

latex <- str\_c(latex, "&", sep = "")

else

latex <- str\_c(latex, "\\\\\n", sep = "")

}

} else {

latex <- str\_c(latex, "\\textbf{Estatística} & \\textbf{Valor} \\\\\n", sep = "")

}

latex <- str\_c(latex, "\t\t\\midrule\n", sep = "")

starting\_number <- ifelse(col\_count > 2, 2, 1)

for (i in seq(starting\_number, row\_count)) {

latex <- str\_c(latex, "\t\t", str\_flatten(t(data\_clean[i, ]), collapse = " & "), " \\\\\n")

}

latex <- str\_c(latex, "\t\\bottomrule\n",

"\t\\end{tabular}\n",

"\t\\label{", label, "}\n",

"\t\\end{adjustbox}\n",

"\\end{quadro}", sep = "")

writeLines(latex)

}

olimpiadas %>%

print\_quadro\_resumo(var\_name = pesokg, title = "Medidas resumo do peso (kg) dos atletas")

```

{r, echo=FALSE, results='asis'}

cat("\\begin{quadro}[H]

\\caption{Medidas resumo da variável peso}

\\centering

\\begin{tabular}{| l | S[table-format = 2.2] |}

\\toprule

\\textbf{Estatística} & \\textbf{Valor} \\\\

\\midrule

Média & 74.00 \\\\

Desvio Padrão & 16.26 \\\\

Variância & 264.26 \\\\

Mínimo & 28.00 \\\\

1º Quartil & 62.00 \\\\

Mediana & 72.00 \\\\

3º Quartil & 84.00 \\\\

Máximo & 175.00 \\\\

\\bottomrule

\\end{tabular}

\\label{quad:quadro\_resumo1}

\\end{quadro}

")

‘’’

```{r}

print\_quadro\_resumo <- function(data, var\_name, title = "Medidas resumo da(o) [nome da variável]", label = "quad:quadro\_resumo1") {

var\_name <- substitute(var\_name)

data\_clean <- data %>%

filter(!is.na(!!sym(var\_name)) & !!sym(var\_name) != Inf)

data\_clean <- data\_clean %>%

summarize(

`Média` = round(mean(!!sym(var\_name), na.rm = TRUE), 2),

`Desvio Padrão` = round(sd(!!sym(var\_name), na.rm = TRUE), 2),

`Variância` = round(var(!!sym(var\_name), na.rm = TRUE), 2),

`Mínimo` = round(min(!!sym(var\_name), na.rm = TRUE), 2),

`1º Quartil` = round(quantile(!!sym(var\_name), probs = 0.25, na.rm = TRUE), 2),

`Mediana` = round(quantile(!!sym(var\_name), probs = 0.5, na.rm = TRUE), 2),

`3º Quartil` = round(quantile(!!sym(var\_name), probs = 0.75, na.rm = TRUE), 2),

`Máximo` = round(max(!!sym(var\_name), na.rm = TRUE), 2)

) %>%

t() %>%

as.data.frame() %>%

rownames\_to\_column()

latex <- str\_c("\\begin{quadro}[H]\n",

"\\caption{", title, "}\n",

"\\centering\n",

"\\begin{adjustbox}{max width=\\textwidth}\n",

"\\begin{tabular}{", sep = "")

col\_count <- ncol(data\_clean)

row\_count <- nrow(data\_clean)

latex <- str\_c(latex, "| l |\n", sep = "")

for (i in seq(2, col\_count)) {

numCount <- data\_clean[i, -c(1)] %>%

as.numeric() %>%

{floor(log10(.)) + 1} %>%

max()

latex <- str\_c(latex, "\t\t\tS[table-format = ", numCount, ".2]\n", sep = "")

}

latex <- str\_c(latex, "\t\t\t|}\n\t\\toprule\n\t\t", sep = "")

if (col\_count > 2) {

for (i in seq(1, col\_count)) {

if (i == 1)

latex <- str\_c(latex, "\\textbf{Estatística}", sep = "")

else

latex <- str\_c(latex, " \\textbf{", data\_clean[1, i], "}", sep = "")

if (i < col\_count)

latex <- str\_c(latex, "&", sep = "")

else

latex <- str\_c(latex, "\\\\\n", sep = "")

}

} else {

latex <- str\_c(latex, "\\textbf{Estatística} & \\textbf{Valor} \\\\\n", sep = "")

}

latex <- str\_c(latex, "\t\t\\midrule\n", sep = "")

starting\_number <- ifelse(col\_count > 2, 2, 1)

for (i in seq(starting\_number, row\_count)) {

latex <- str\_c(latex, "\t\t", str\_flatten(t(data\_clean[i, ]), collapse = " & "), " \\\\\n")

}

latex <- str\_c(latex, "\t\\bottomrule\n",

"\t\\end{tabular}\n",

"\t\\label{", label, "}\n",

"\t\\end{adjustbox}\n",

"\\end{quadro}", sep = "")

writeLines(latex)

}

olimpiadas %>%

print\_quadro\_resumo(var\_name = Altura, title = "Medidas resumo da Altura dos atletas")

```

{r, echo=FALSE, results='asis'}

cat("\\begin{quadro}[H]

\\caption{Medidas resumo da variável Altura}

\\centering

\\begin{tabular}{| l | S[table-format = 2.2] |}

\\toprule

\\textbf{Estatística} & \\textbf{Valor} \\\\

\\midrule

Média & 178.24 \\\\

Desvio Padrão & 11.80 \\\\

Variância & 139.23 \\\\

Mínimo & 137.00 \\\\

1º Quartil & 170.00 \\\\

Mediana & 178.00 \\\\

3º Quartil & 186.00 \\\\

Máximo & 219.00 \\\\

\\bottomrule

\\end{tabular}

\\label{quad:quadro\_resumo1}

\\end{quadro}

")