ECS111-Preprocessing

Samarth Sridhara

2025-05-13

```
# Load data
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)

# Step 1: Load the data
df <- read_csv("/Users/samarthsridhara/Downloads/valorant_players_processedMay12,2025.csv")
```

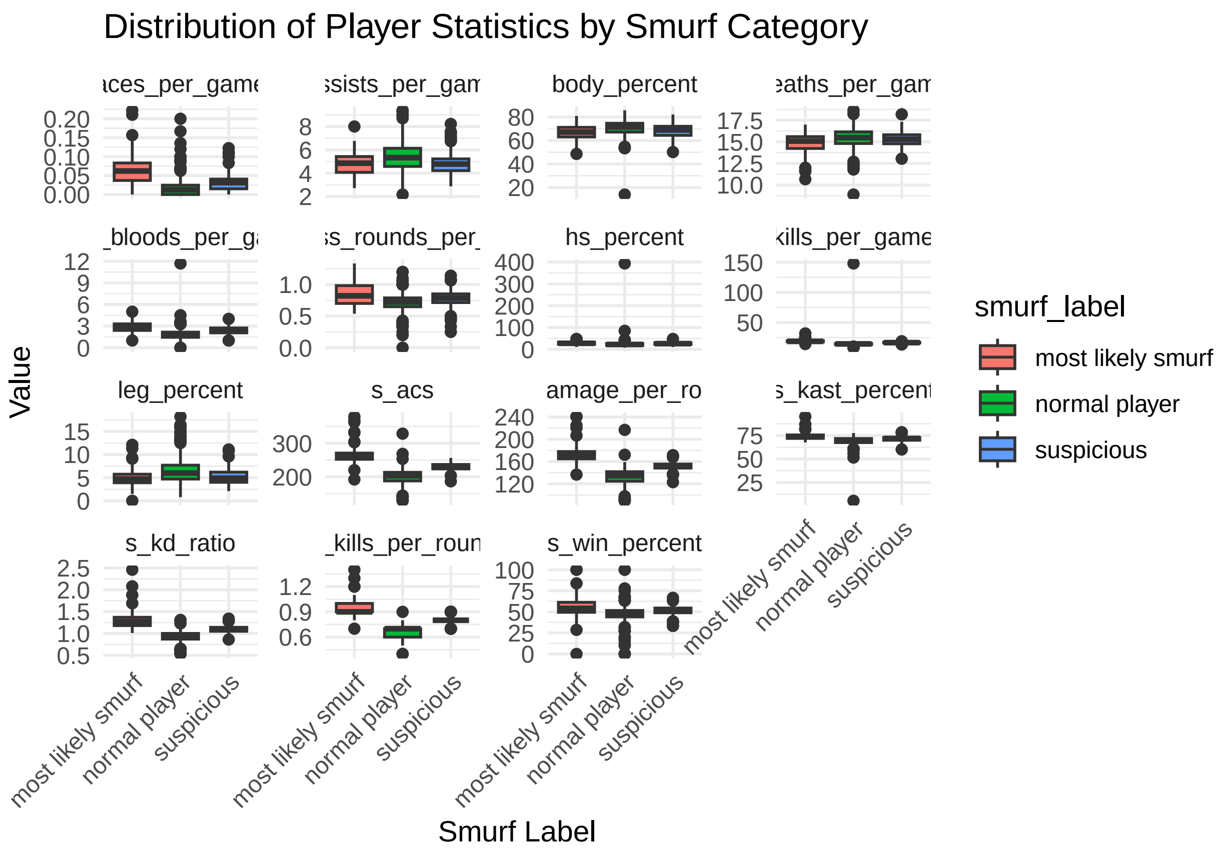
```
## Rows: 553 Columns: 22

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): puuid, user, tag, deaths_per_game, kills_per_game, assists_per_gam...
## dbl (12): hs_percent, body_percent, leg_percent, s_damage_per_round, s_kd_ra...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Step 2: Coerce target columns to numeric using dplyr::across()
cols_to_numeric <- c("hs_percent", "body_percent", "leg_percent", "s_kd_ratio",
                     "s_win_percent", "s_kast_percent", "s_damage_per_round",
                     "s_acs", "s_kills_per_round", "deaths_per_game",
                     "kills_per_game", "assists_per_game", "first_bloods_per_game",
                     "flawless_rounds_per_game", "aces_per_game")

df <- df %>%
  mutate(across(all_of(cols_to_numeric), as.numeric))
```

```
## Warning: There were 6 warnings in ‘mutate()‘.
## The first warning was:
## i In argument: ‘across(all_of(cols_to_numeric), as.numeric)‘.
## Caused by warning:
## ! NAs introduced by coercion
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 5 remaining warnings.
```

```
# Step 3: Pivot longer for faceting
df_long <- pivot_longer(df, cols = all_of(cols_to_numeric),
                        names_to = "statistic", values_to = "value")

# Step 4: Plot with facet_wrap
ggplot(df_long, aes(x = smurf_label, y = value, fill = smurf_label)) +
  geom_boxplot() +
  facet_wrap(~ statistic, scales = "free_y") +
  theme_minimal() +
  labs(title = "Distribution of Player Statistics by Smurf Category",
       x = "Smurf Label", y = "Value") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 30 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



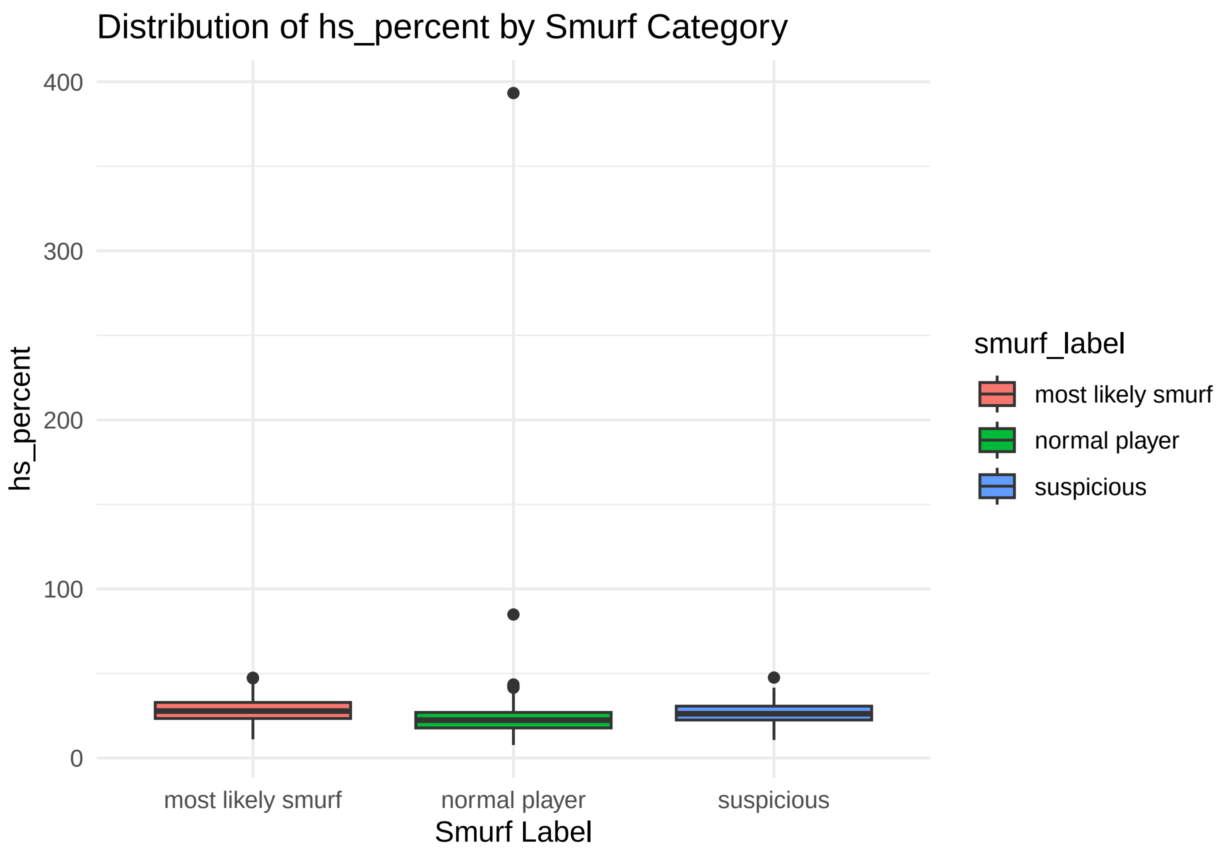
```r
library(ggplot2)
library(dplyr)
library(readr)
library(tidyr)

df <- read_csv("/Users/samarthsridhara/Downloads/valorant_players_processedMay12,2025.csv")
```

```
## Rows: 553 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): puuid, user, tag, deaths_per_game, kills_per_game, assists_per_gam...
## dbl (12): hs_percent, body_percent, leg_percent, s_damage_per_round, s_kd_ra...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

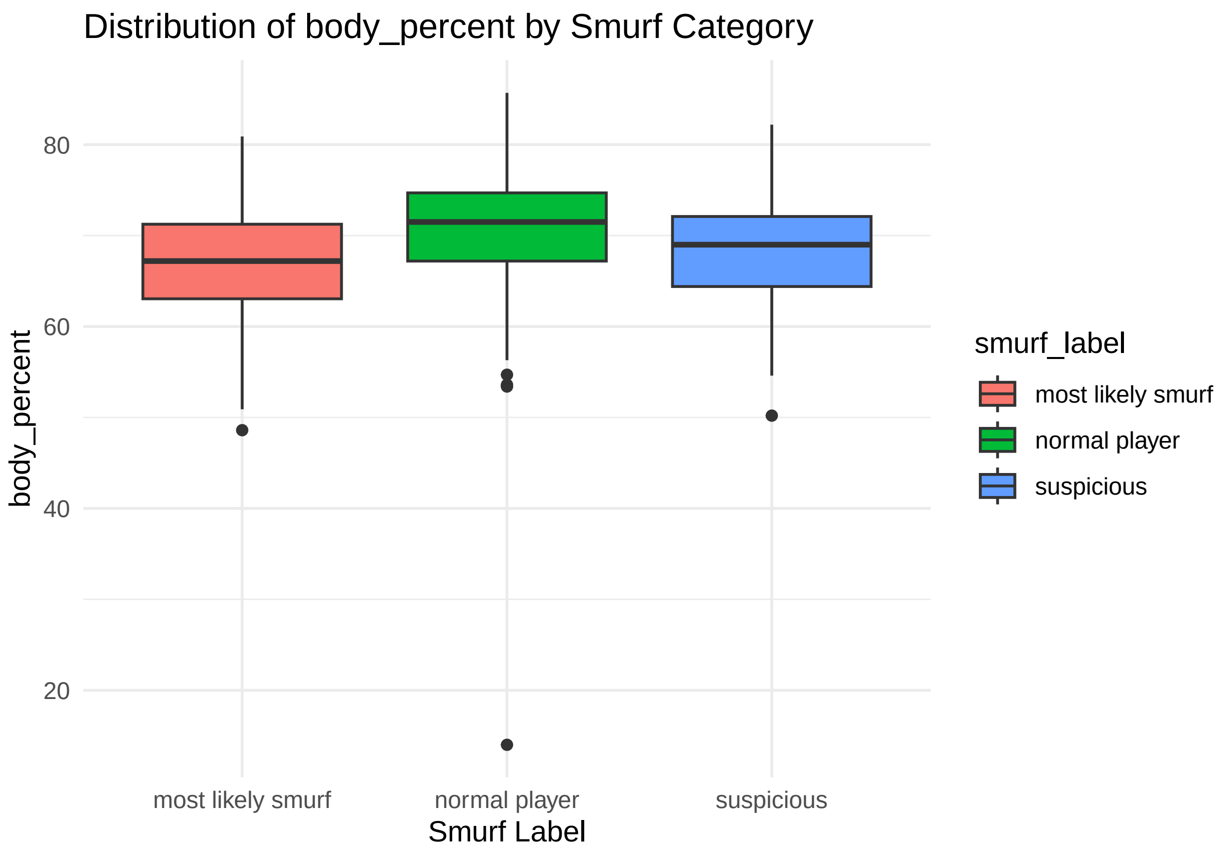
```r
cols_to_numeric <- c("hs_percent", "body_percent", "leg_percent", "s_kd_ratio",
                     "s_win_percent", "s_kast_percent", "s_damage_per_round",
                     "s_acs", "s_kills_per_round", "deaths_per_game",
                     "kills_per_game", "assists_per_game", "first_bloods_per_game",
                     "flawless_rounds_per_game", "aces_per_game")

df <- df %>%
  mutate(across(all_of(cols_to_numeric), as.numeric)) %>%
  pivot_longer(cols = all_of(cols_to_numeric), names_to = "statistic", values_to = "value")
```

```
## Warning: There were 6 warnings in ‘mutate()‘.
## The first warning was:
## i In argument: ‘across(all_of(cols_to_numeric), as.numeric)‘.
## Caused by warning:
## ! NAs introduced by coercion
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 5 remaining warnings.
```

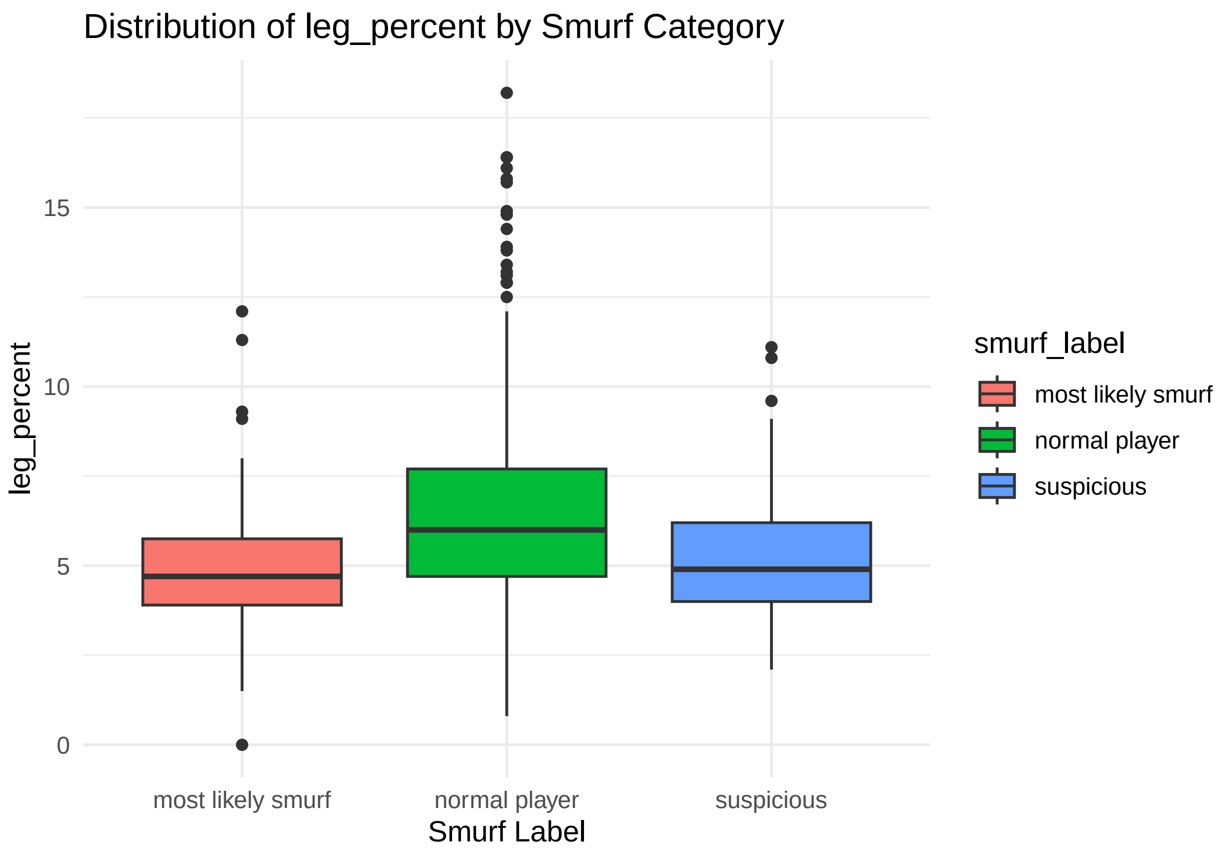
```
unique_stats <- unique(df$statistic)

for (stat in unique_stats) {
  p <- ggplot(filter(df, statistic == stat), aes(x = smurf_label, y = value, fill = smurf_label)) +
    geom_boxplot() +
    theme_minimal() +
    labs(title = paste("Distribution of", stat, "by Smurf Category"),
         x = "Smurf Label", y = stat)
  print(p)
  readline(prompt = "Press [enter] to continue to next plot")
}
```



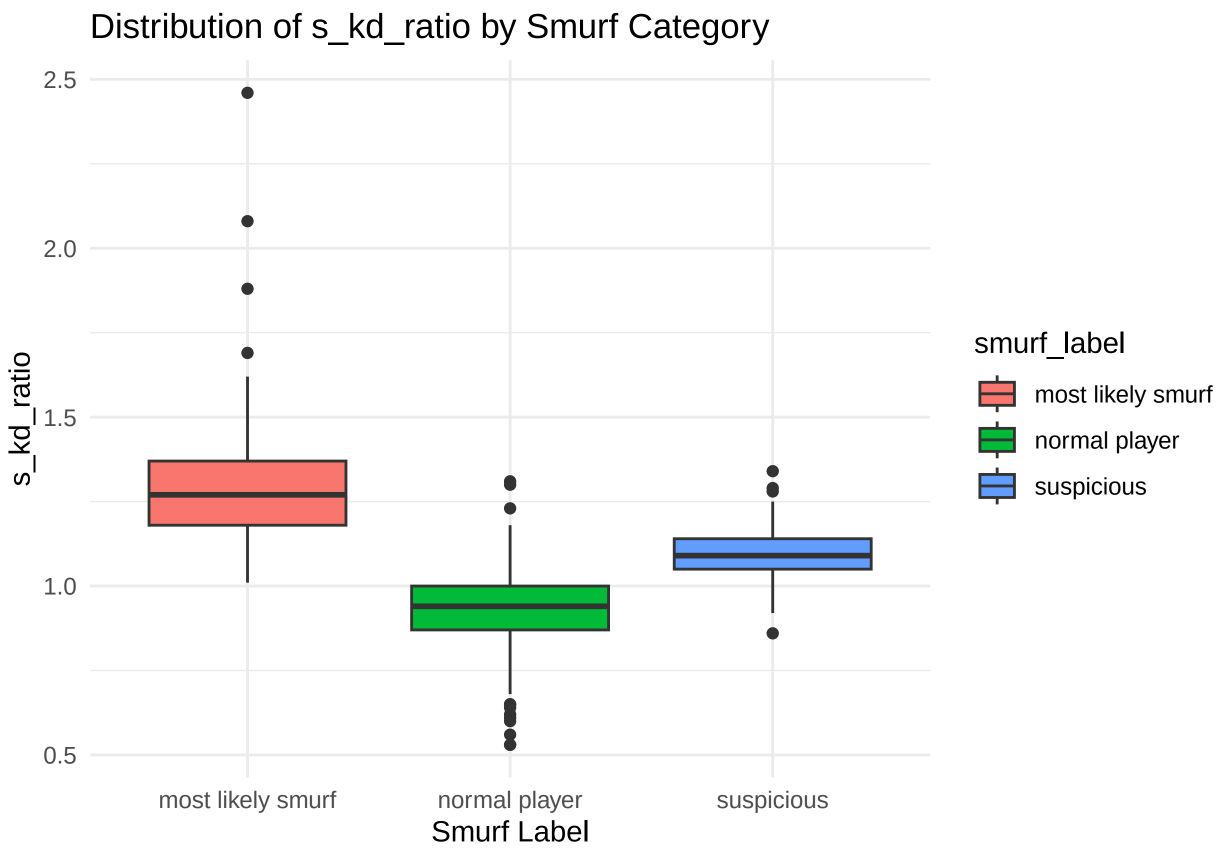
```
## Press [enter] to continue to next plot
```



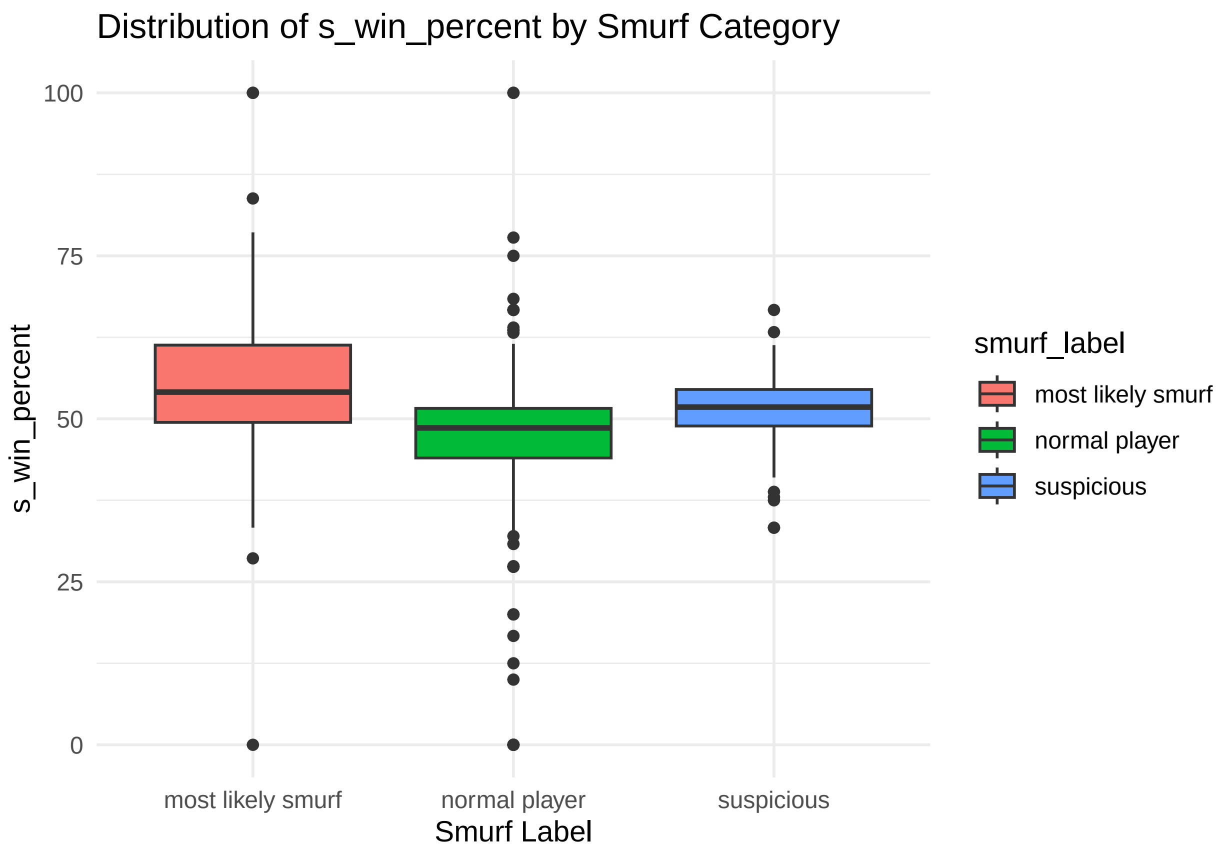
```
## Press [enter] to continue to next plot
```



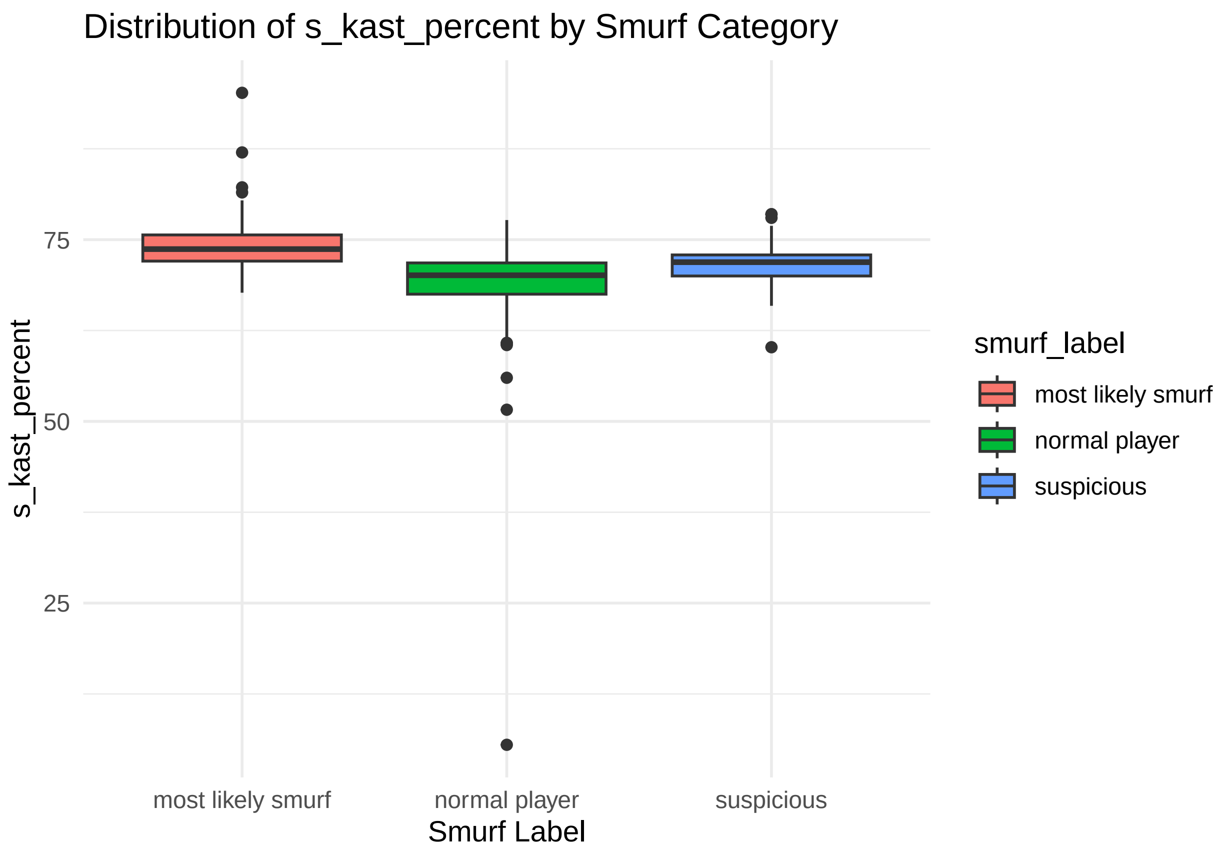
```
## Press [enter] to continue to next plot
```



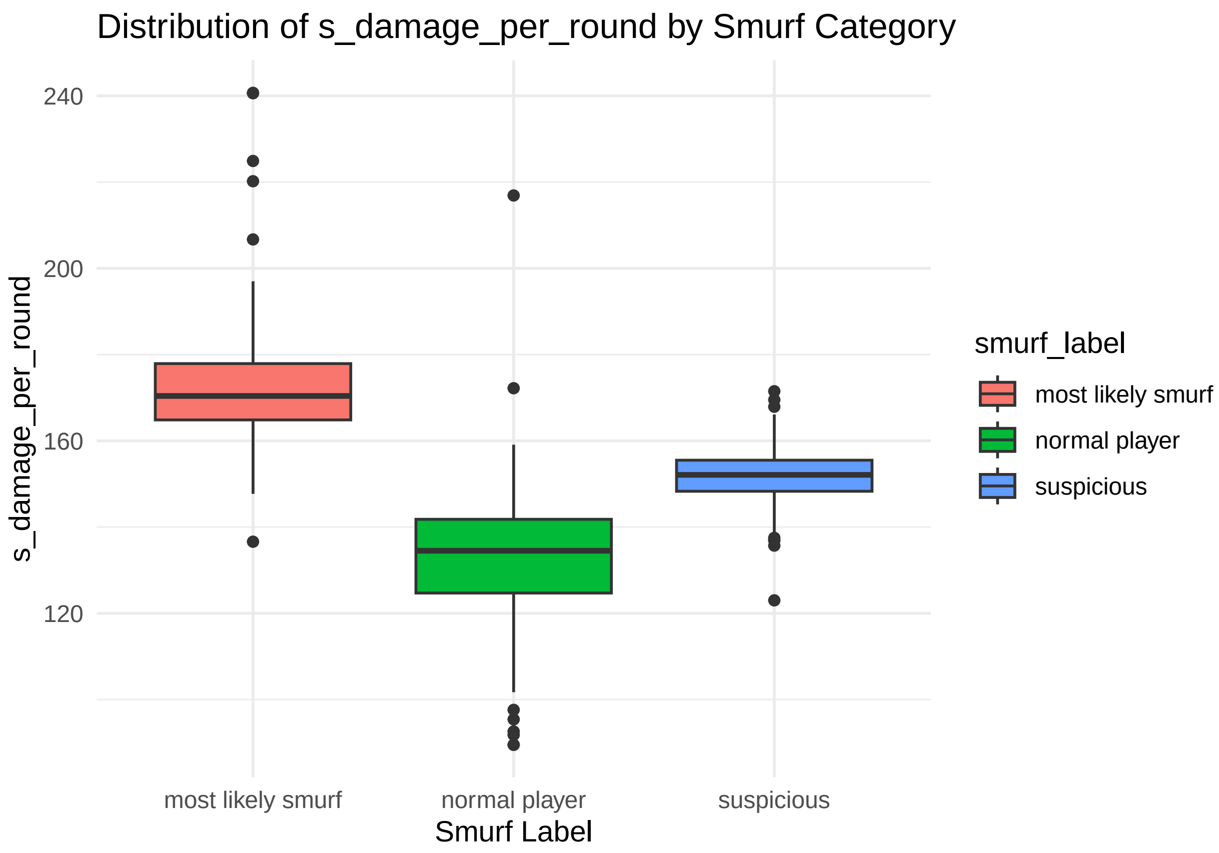
```
## Press [enter] to continue to next plot
```



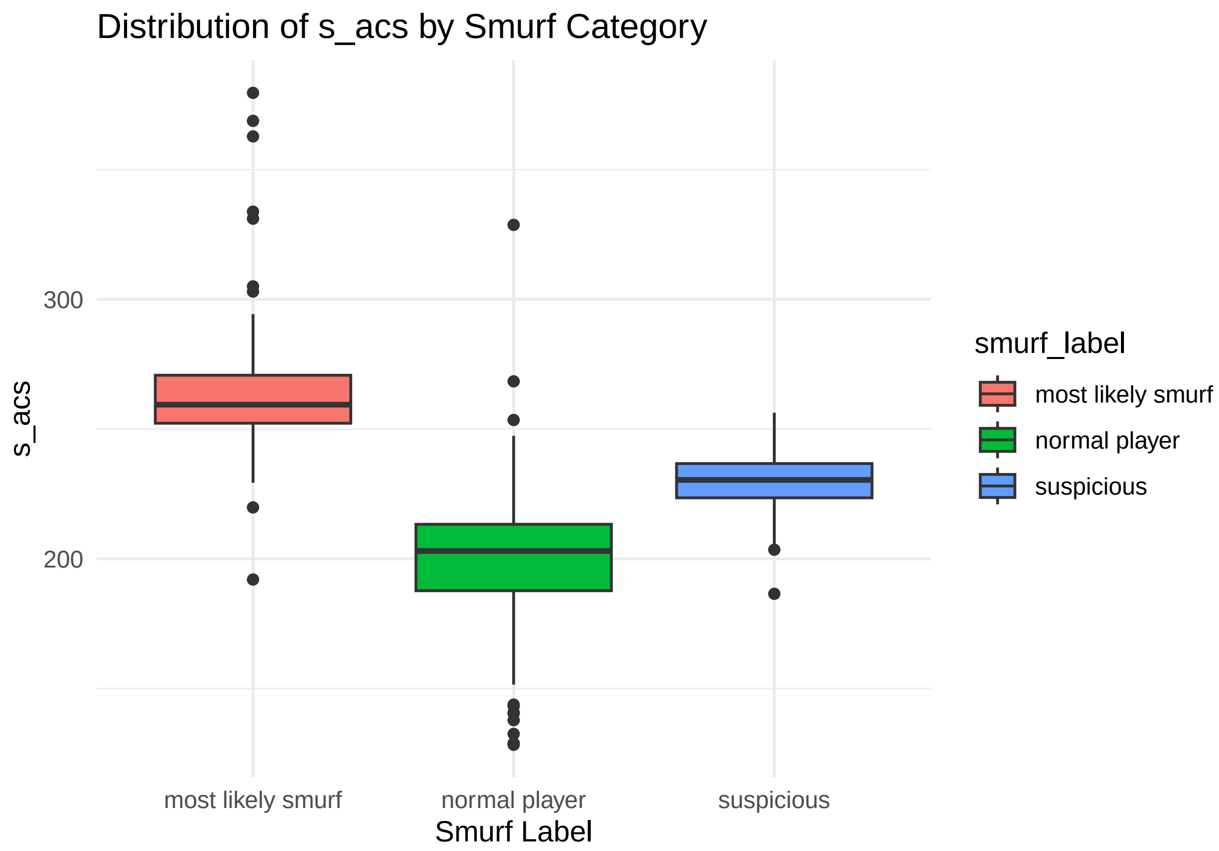
```
## Press [enter] to continue to next plot
```



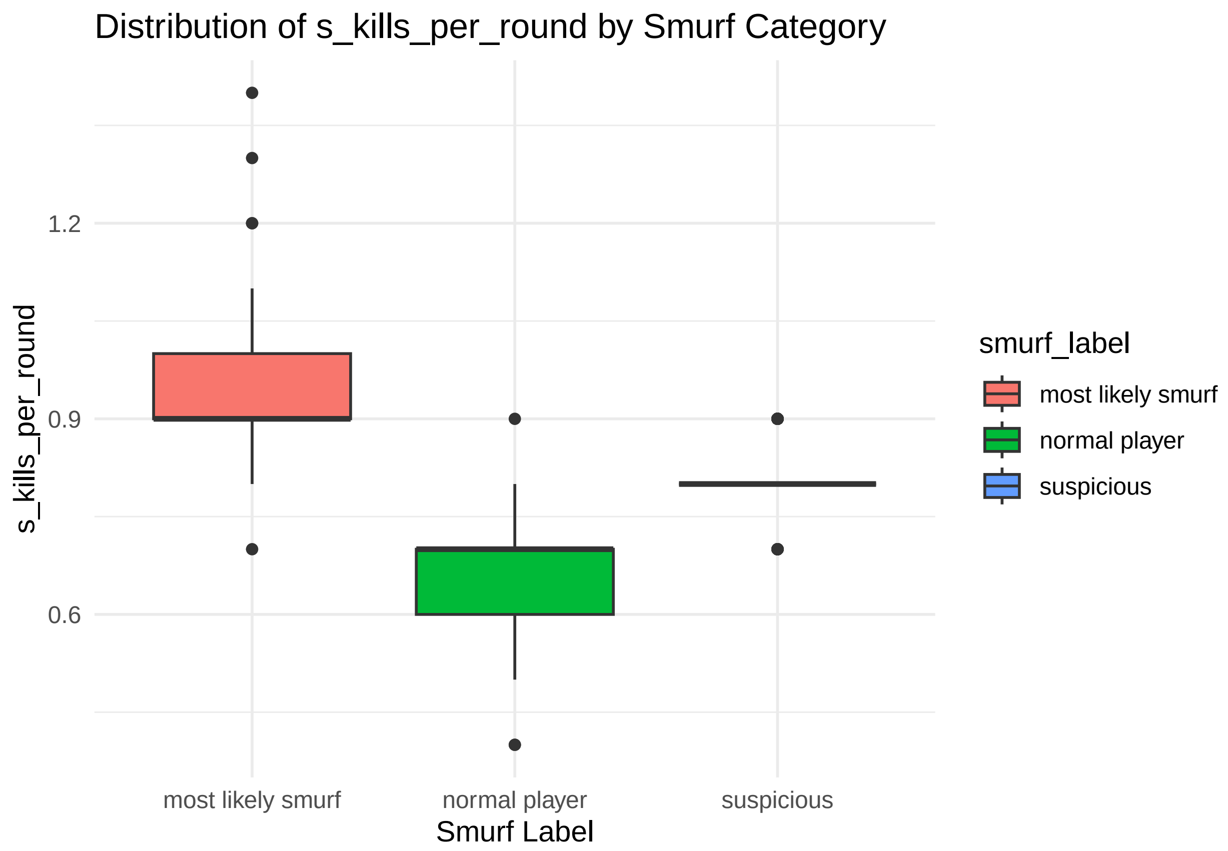
```
## Press [enter] to continue to next plot
```



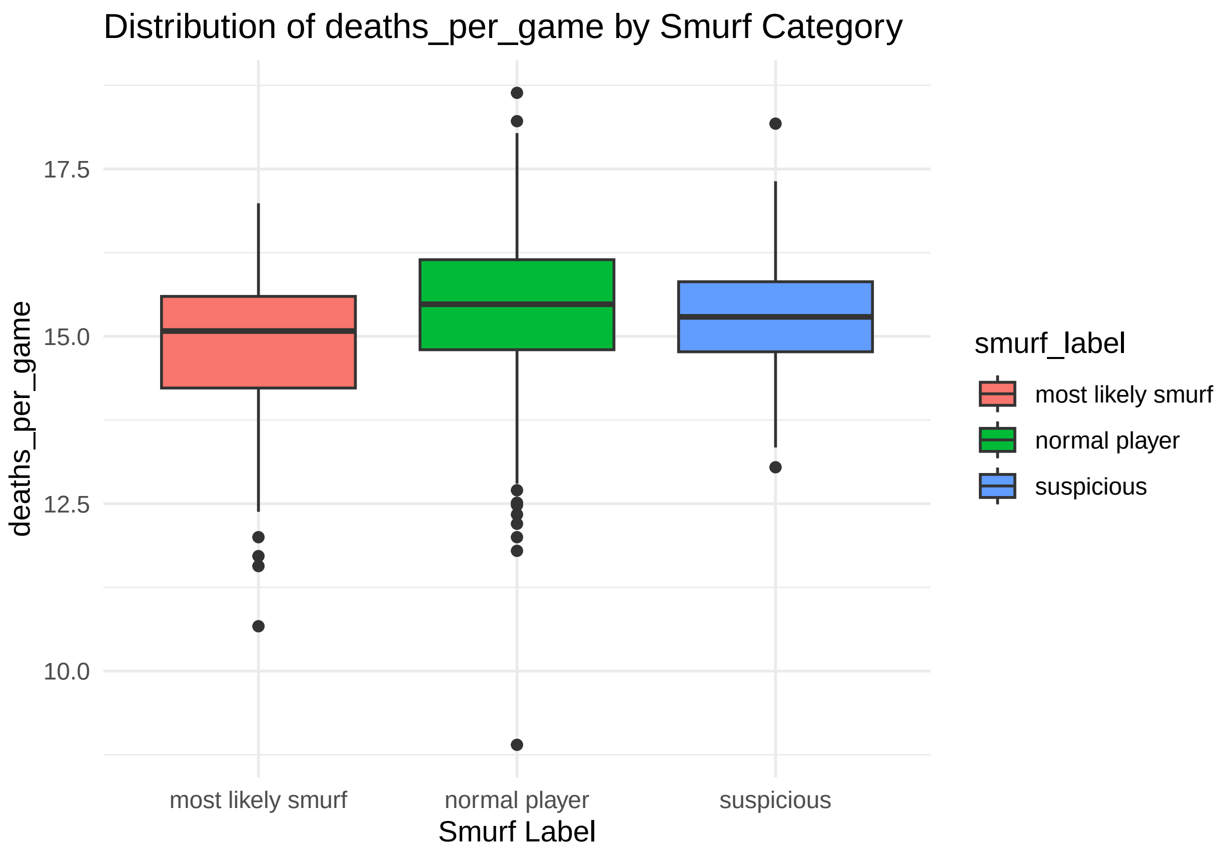
```
## Press [enter] to continue to next plot
```



```
## Press [enter] to continue to next plot
```

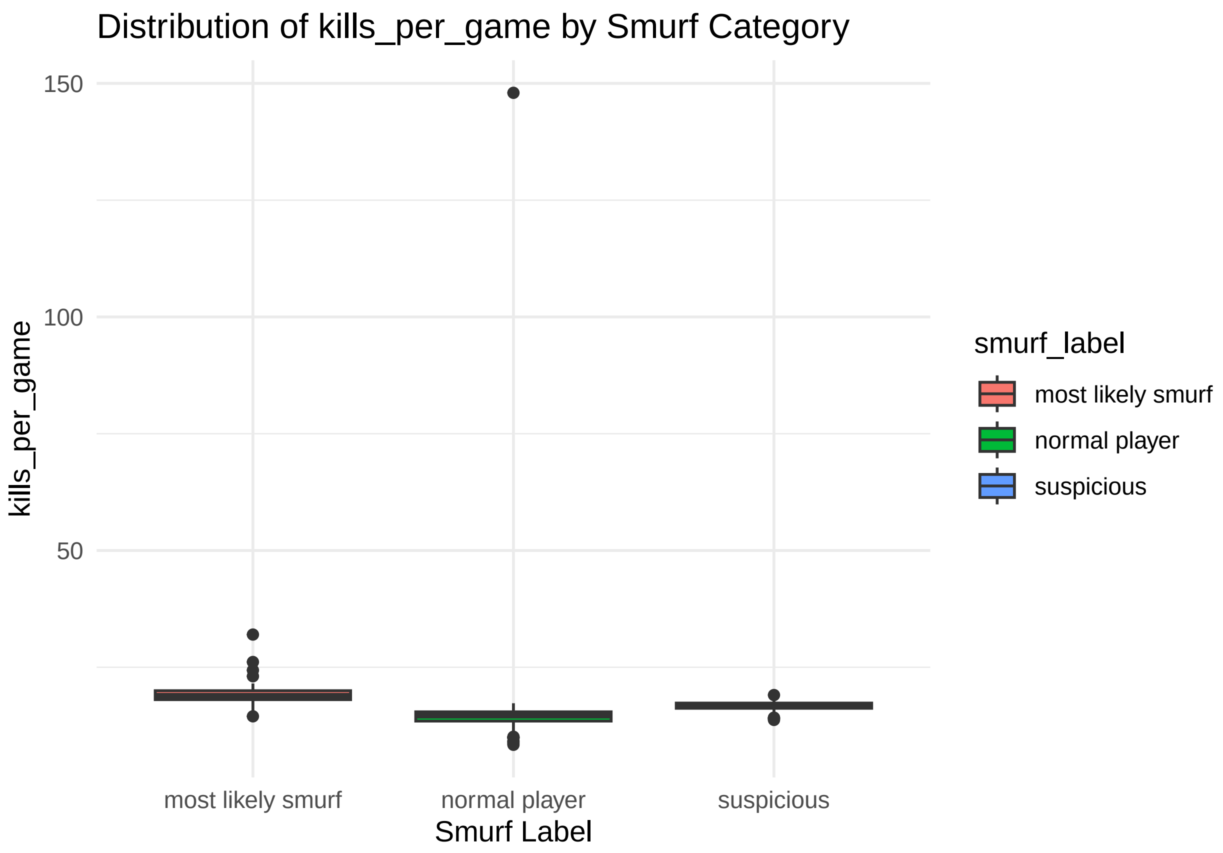


```
## Press [enter] to continue to next plot

## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



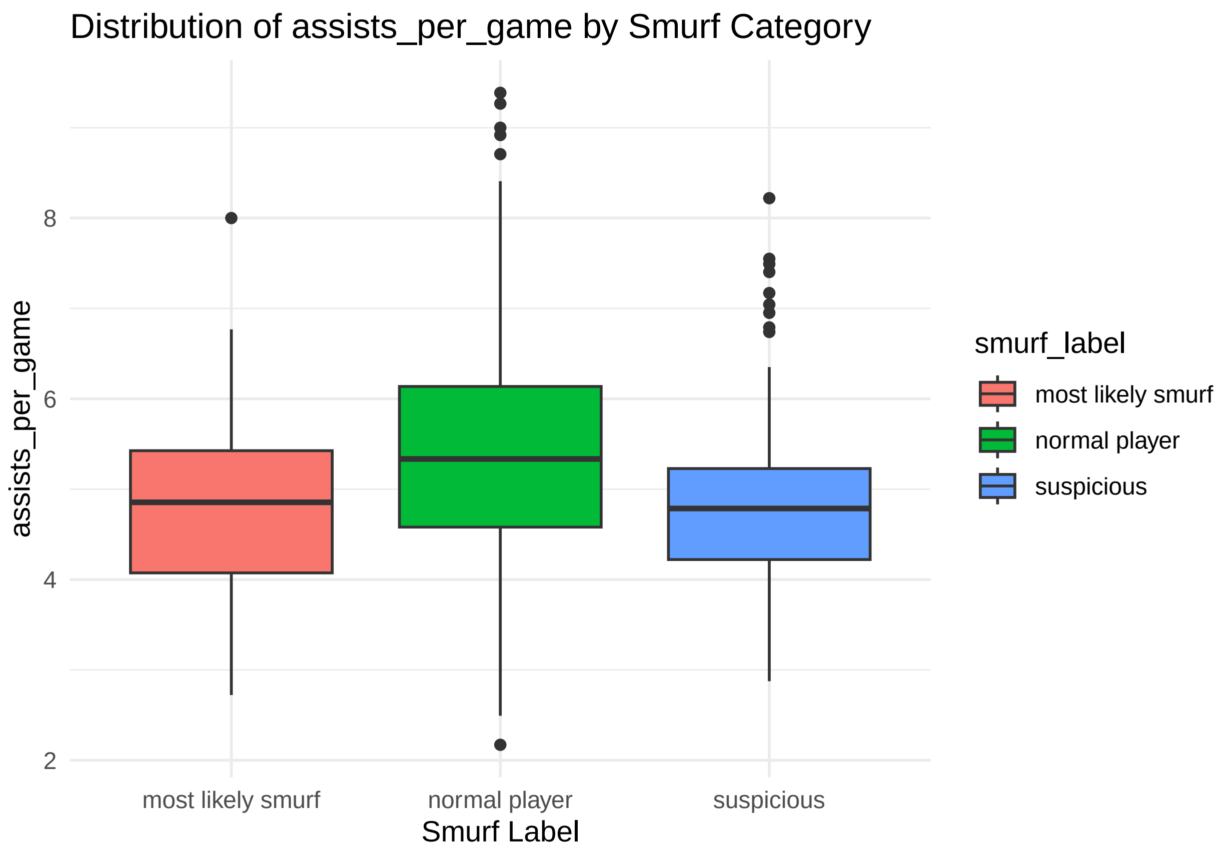
```
## Press [enter] to continue to next plot
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

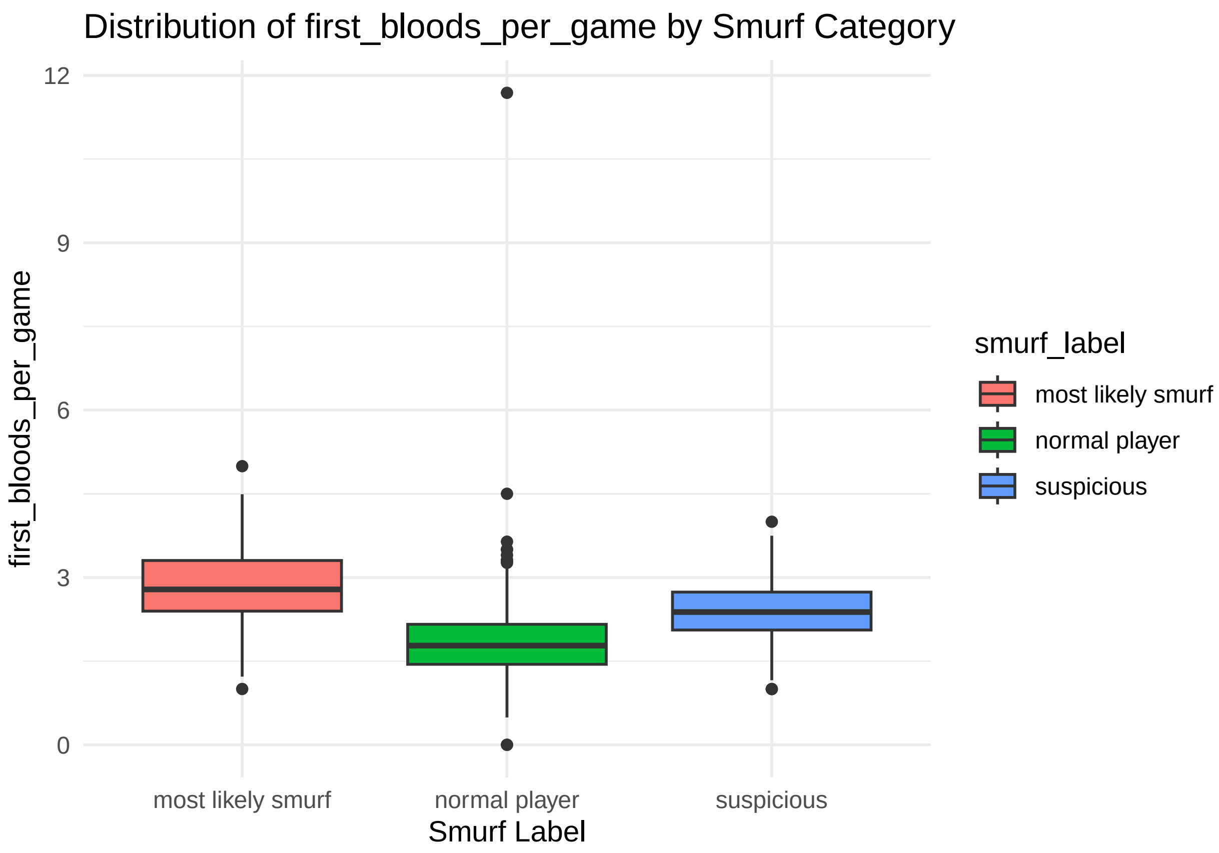


```
## Press [enter] to continue to next plot
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

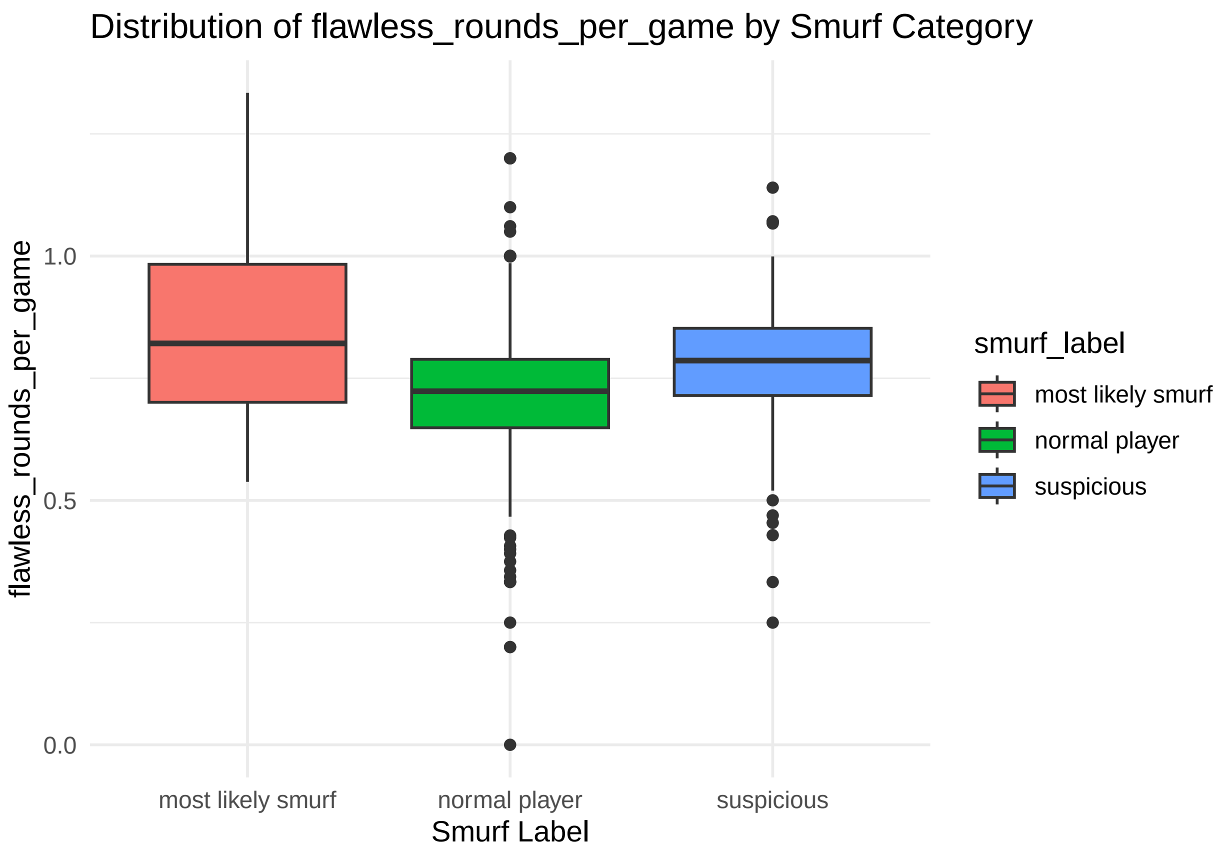


```
## Press [enter] to continue to next plot

## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

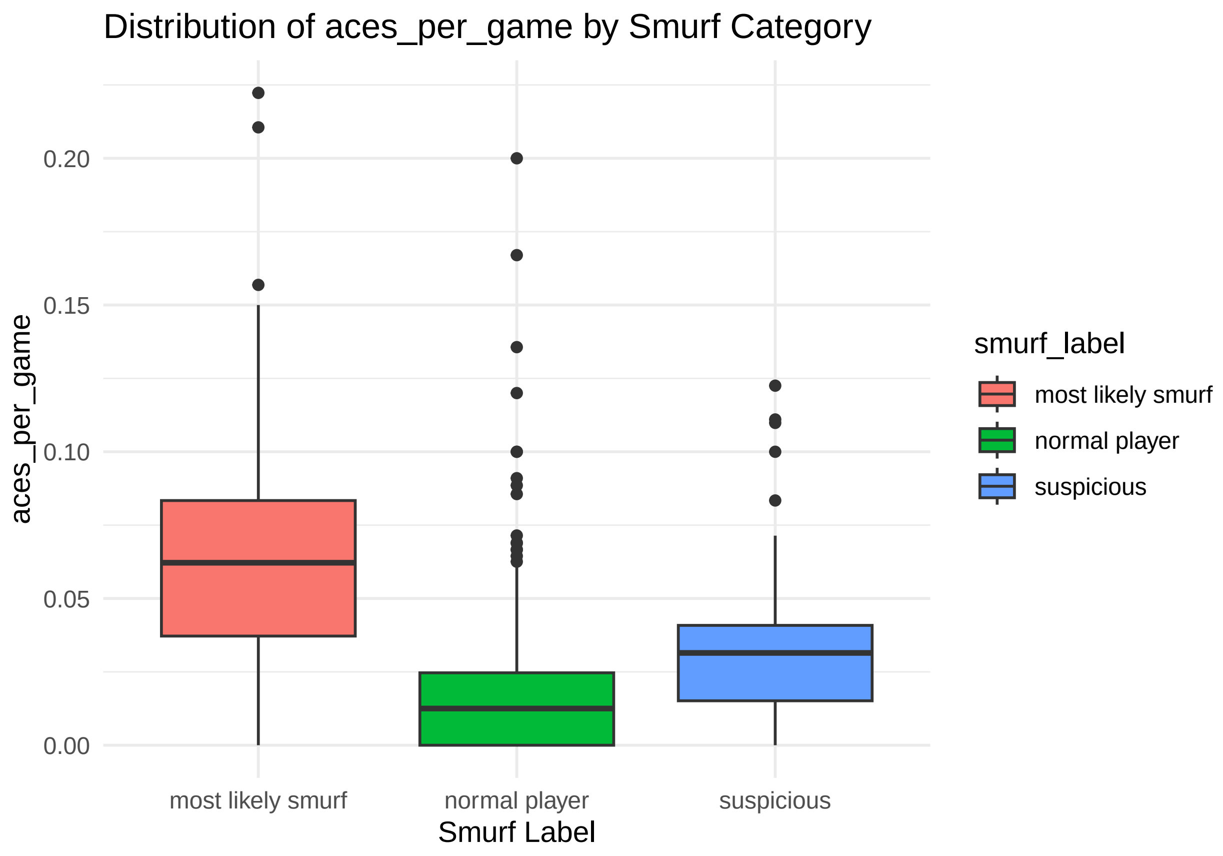


```
## Press [enter] to continue to next plot
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



```
## Press [enter] to continue to next plot

## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



```
## Press [enter] to continue to next plot
```

```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)

# Load the data
df <- read_csv("valorant_players_processedMay15,2025+morepreprocessing.csv")
```

```
## Rows: 547 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): puuid, user, tag
## dbl (17): hs_percent, leg_percent, s_damage_per_round, s_kd_ratio, s_win_per...
## lgl  (3): smurf_label_most likely smurf, smurf_label_normal player, smurf_la...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

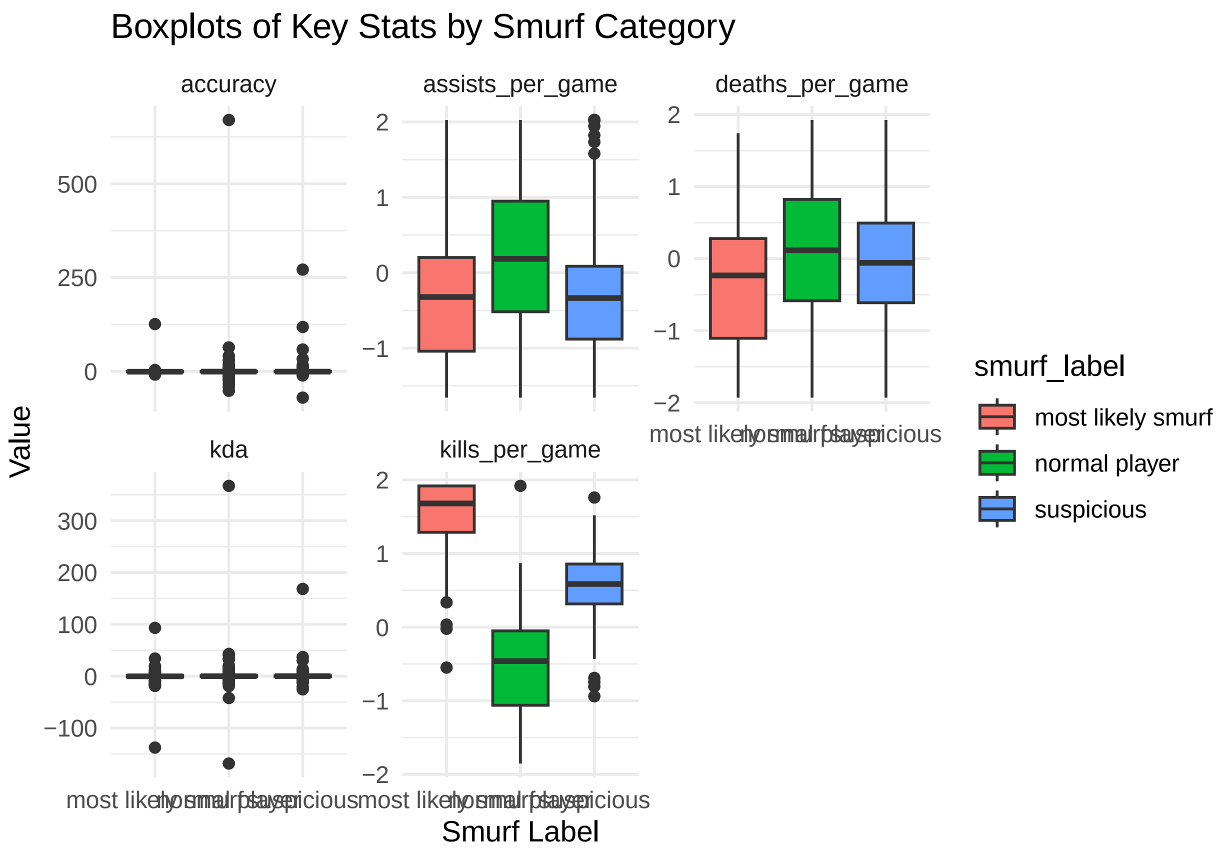
```
df <- df %>%
  mutate(smurf_label = case_when(
    `smurf_label_most likely smurf` == 1 ~ "most likely smurf",
    `smurf_label_suspicious` == 1 ~ "suspicious",
```

```r
    `smurf_label_normal player` == 1 ~ "normal player",
    TRUE ~ "unknown"
  ))
```

```r
# Summary of stats
key_stats <- c("kills_per_game", "deaths_per_game", "assists_per_game", "kda", "accuracy")

summary_stats <- df %>%
  select(all_of(key_stats)) %>%
  summary()

print(summary_stats)
```

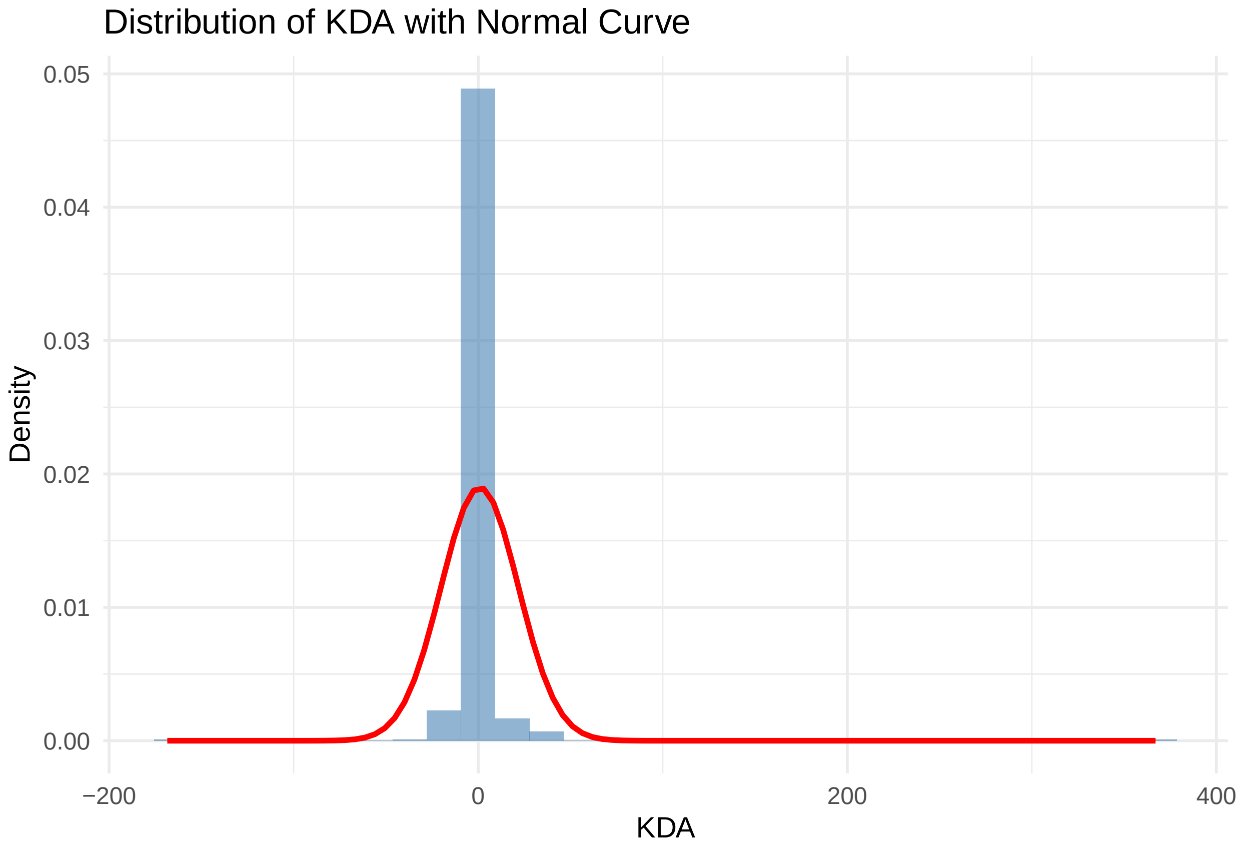
```
##  kills_per_game      deaths_per_game     assists_per_game
##  Min.   :-1.851653   Min.   :-1.930062   Min.   :-1.656516
##  1st Qu.:-0.687186   1st Qu.:-0.621382   1st Qu.:-0.761355
##  Median :-0.022950   Median :-0.008923   Median :-0.086627
##  Mean   :-0.001468   Mean   : 0.009246   Mean   :-0.002785
##  3rd Qu.: 0.628662   3rd Qu.: 0.670612   3rd Qu.: 0.639692
##  Max.   : 1.917576   Max.   : 1.924725   Max.   : 2.026356
##       kda                 accuracy
##  Min.   :-168.4494   Min.   :-70.2772
##  1st Qu.:  -1.1223   1st Qu.: -1.7009
##  Median :  -0.0219   Median : -1.0385
##  Mean   :   0.8495   Mean   :  1.1313
##  3rd Qu.:   1.2198   3rd Qu.: -0.2907
##  Max.   : 367.0384   Max.   :669.9229
```

```r
df_long <- df %>%
  pivot_longer(cols = all_of(key_stats), names_to = "statistic", values_to = "value")

ggplot(df_long, aes(x = smurf_label, y = value, fill = smurf_label)) +
  geom_boxplot() +
  facet_wrap(~ statistic, scales = "free_y") +
  theme_minimal() +
  labs(title = "Boxplots of Key Stats by Smurf Category", x = "Smurf Label", y = "Value")
```

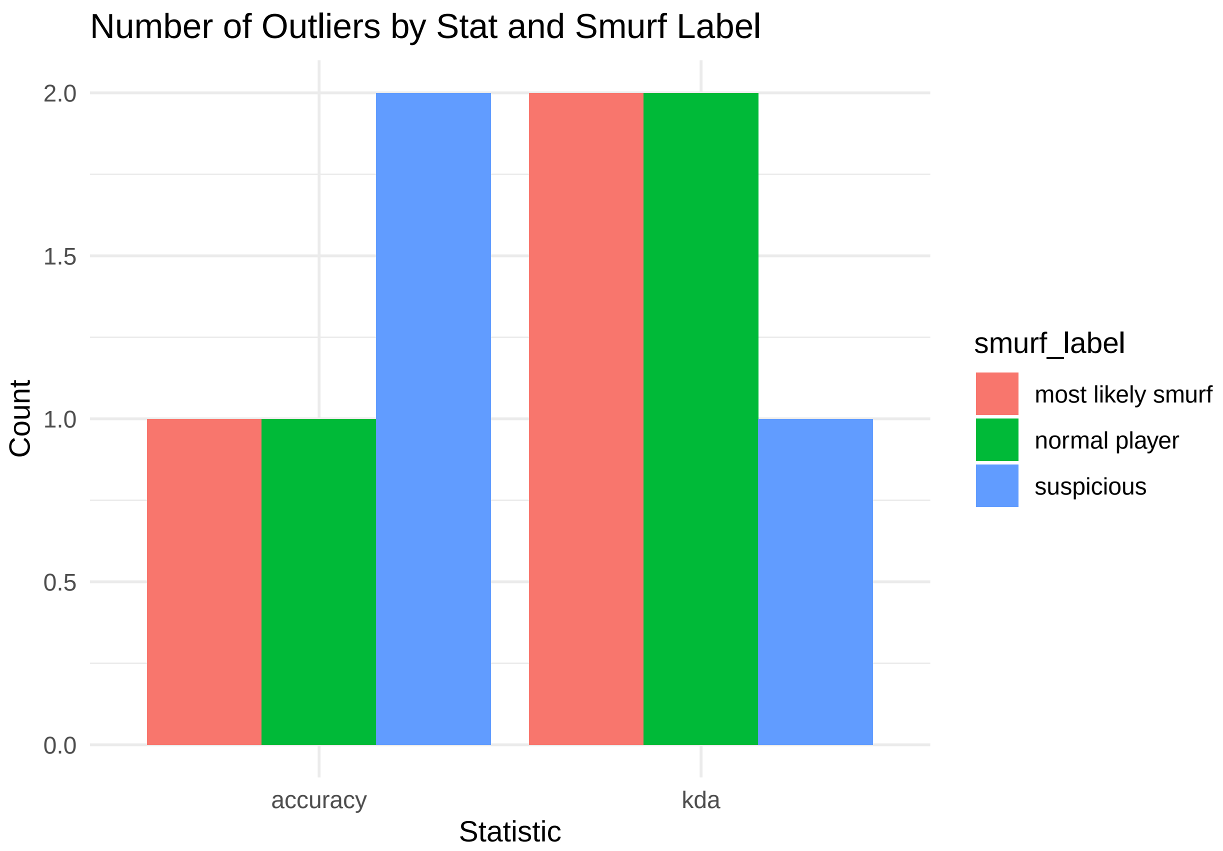


```r
ggplot(df, aes(x = kda)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", alpha = 0.6) +
  stat_function(fun = dnorm, args = list(mean = mean(df$kda, na.rm = TRUE),
                                         sd = sd(df$kda, na.rm = TRUE)),
                color = "red", size = 1) +
  theme_minimal() +
  labs(title = "Distribution of KDA with Normal Curve", x = "KDA", y = "Density")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

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



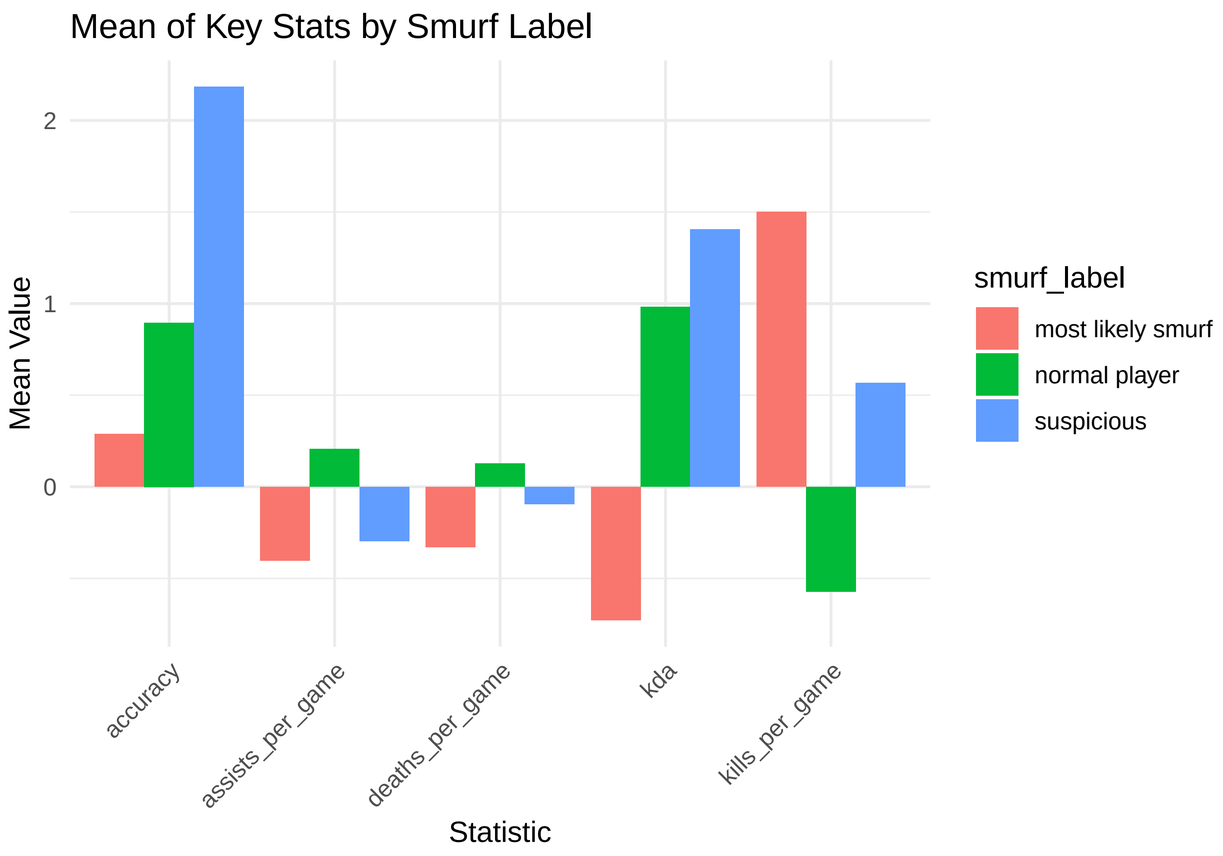
```
# Define outliers (Z-score > 3 or < -3)
df_outliers <- df_long %>%
  group_by(statistic) %>%
  mutate(z = (value - mean(value, na.rm = TRUE)) / sd(value, na.rm = TRUE)) %>%
  filter(abs(z) > 3)

ggplot(df_outliers, aes(x = statistic, fill = smurf_label)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Number of Outliers by Stat and Smurf Label", x = "Statistic", y = "Count")
```



```
df_summary_grouped <- df %>%
  group_by(smurf_label) %>%
  summarise(across(all_of(key_stats), mean, na.rm = TRUE)) %>%
  pivot_longer(-smurf_label, names_to = "stat", values_to = "mean_value")
```

```
## Warning: There was 1 warning in ‘summarise()‘.
## i In argument: ‘across(all_of(key_stats), mean, na.rm = TRUE)‘.
## i In group 1: ‘smurf_label = "most likely smurf"‘.
## Caused by warning:
## ! The ‘...‘ argument of ‘across()‘ is deprecated as of dplyr 1.1.0.
## Supply arguments directly to ‘.fns‘ through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
ggplot(df_summary_grouped, aes(x = stat, y = mean_value, fill = smurf_label)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Mean of Key Stats by Smurf Label", x = "Statistic", y = "Mean Value") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

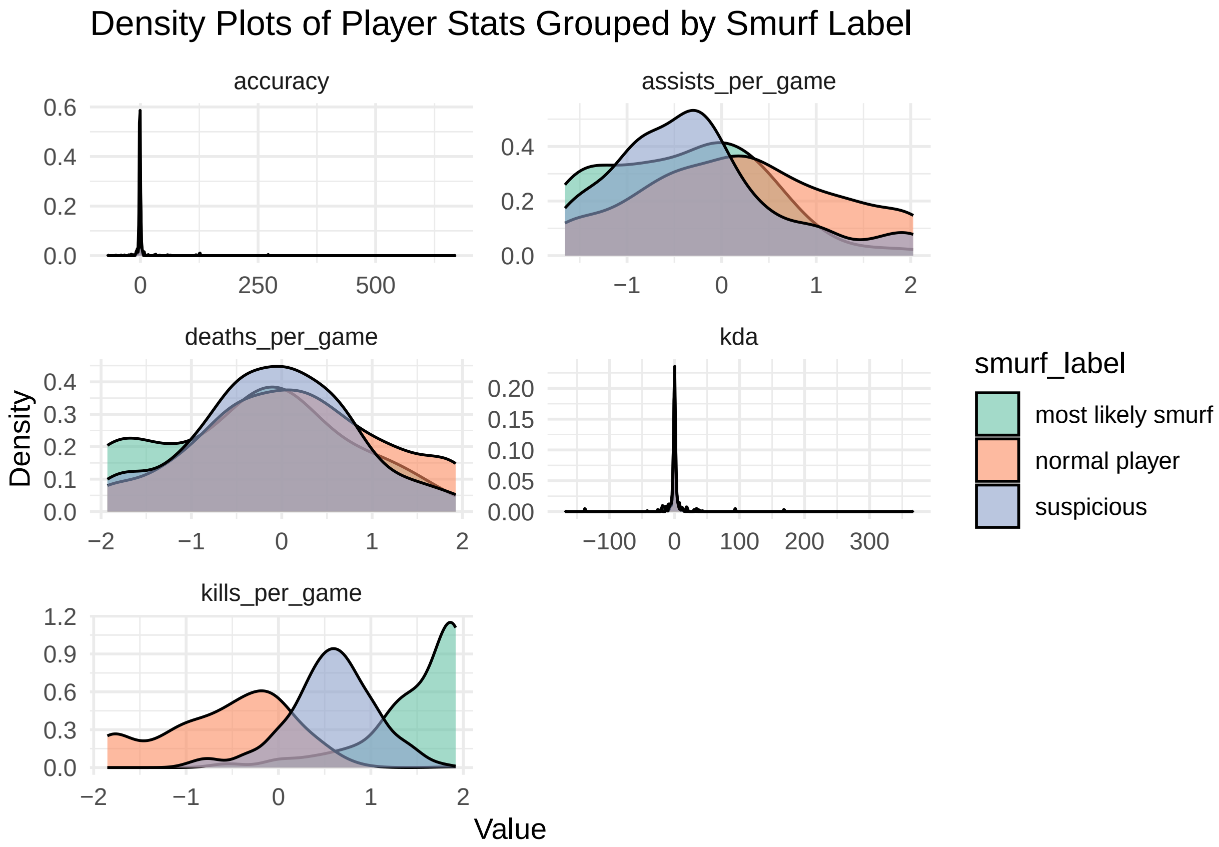


```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)

df <- read_csv("valorant_players_processedMay15,2025+morepreprocessing.csv")
```

```
## Rows: 547 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): puuid, user, tag
## dbl (17): hs_percent, leg_percent, s_damage_per_round, s_kd_ratio, s_win_per...
## lgl  (3): smurf_label_most likely smurf, smurf_label_normal player, smurf_la...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
df <- df %>%
  mutate(smurf_label = case_when(
    `smurf_label_most likely smurf` == 1 ~ "most likely smurf",
    `smurf_label_suspicious` == 1 ~ "suspicious",
    `smurf_label_normal player` == 1 ~ "normal player",
    TRUE ~ "unknown"
  ))
```

```r
key_stats <- c("kills_per_game", "deaths_per_game", "assists_per_game", "kda", "accuracy")

df_long <- df %>%
  pivot_longer(cols = all_of(key_stats), names_to = "statistic", values_to = "value")

ggplot(df_long, aes(x = value, fill = smurf_label)) +
  geom_density(alpha = 0.6) +
  facet_wrap(~ statistic, scales = "free", ncol = 2) +
  theme_minimal() +
  labs(title = "Density Plots of Player Stats Grouped by Smurf Label",
       x = "Value", y = "Density") +
  scale_fill_brewer(palette = "Set2")
```


Density Plots of Player Stats Grouped by Smurf Label
accuracy
assists_per_game
deaths_per_game
kda
kills_per_game
Density
Value
smurf_label
most likely smurf
normal player
suspicious