Visualizing 2D Categorical and Continuous Variables

2023-06-09

```
ohtani_batted_balls <- read_csv("https://shorturl.at/mnwL1")
head(ohtani_batted_balls)
```

```
## # A tibble: 6 x 7
##   pitch_type batted_ball_type  hit_x hit_y exit_velocity launch_angle outcome
##   <chr>      <chr>             <dbl> <dbl>         <dbl>        <dbl> <chr>
## 1 FC         line_drive        89.7  144.          113.            20 home_run
## 2 CH         fly_ball           3.35  83.9          83.9           55 field_out
## 3 CH         fly_ball         -65.6  126.          102.            38 field_out
## 4 CU         ground_ball       39.2   50.4          82.5            8 field_out
## 5 FC         fly_ball         -37.6  138.          101.            23 field_out
## 6 KC         popup            -51.9   41.6          84             65 field_out
```

```
ohtani_batted_balls <- ohtani_batted_balls %>%
  filter(pitch_type != "null") %>%
  mutate(pitch_type = fct_recode(pitch_type,
"Changeup" = "CH", "Breaking ball" = "CU", "Fastball" = "FC","Fastball" = "FF","Fastball" = "FS","Breaki

table(ohtani_batted_balls$pitch_type)
```

```
##
##      Changeup Breaking ball      Fastball
##            62           110           182
```

Chi-Squared Distribution

Question: Are all pitch types equally likely to occur?

To answer this, we can preform a chi-squared test!

Hypotheses:

- $H_0 : p_1 = p_2 = ... = p_k$
- H_a : at least two of p_a for i = 1, 2,..,k are not equal to one another.

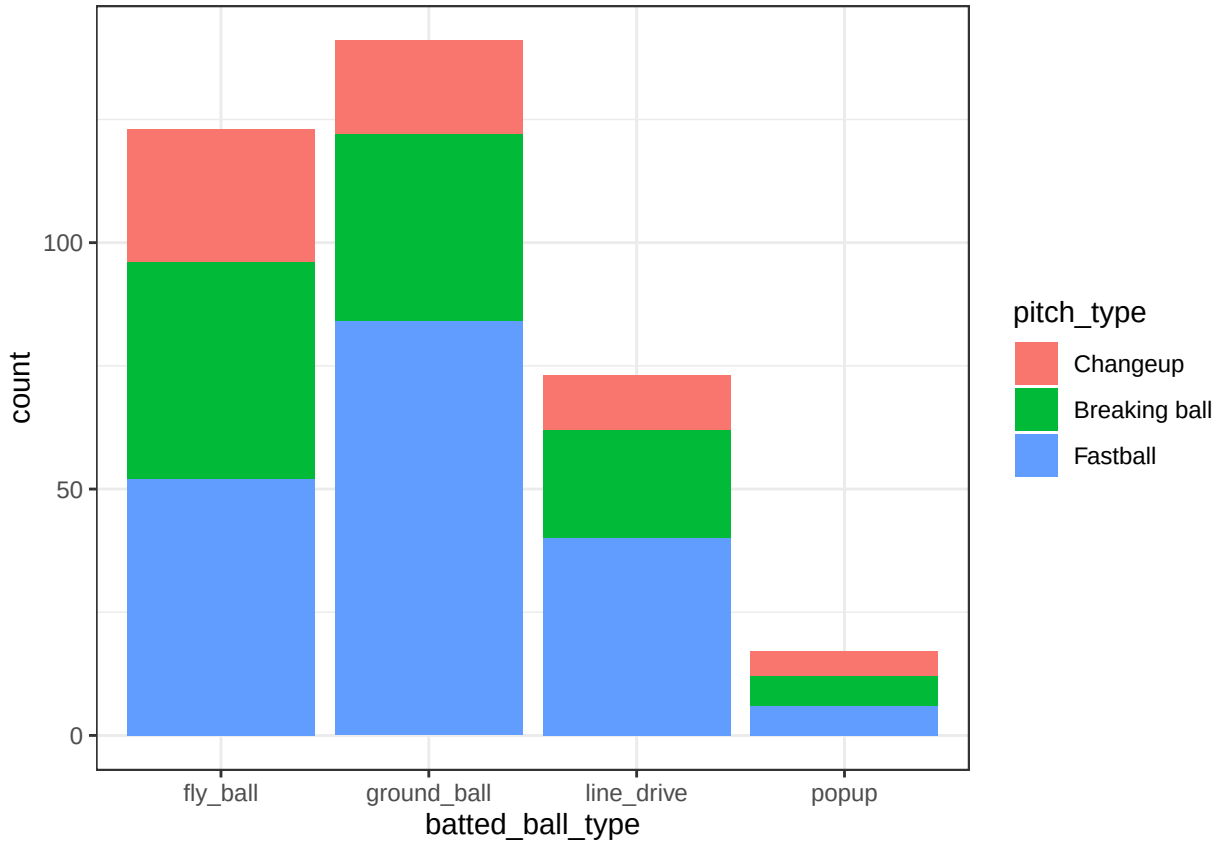
```
chisq.test(table(ohtani_batted_balls$pitch_type))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(ohtani_batted_balls$pitch_type)
## X-squared = 61.831, df = 2, p-value = 3.747e-14
```

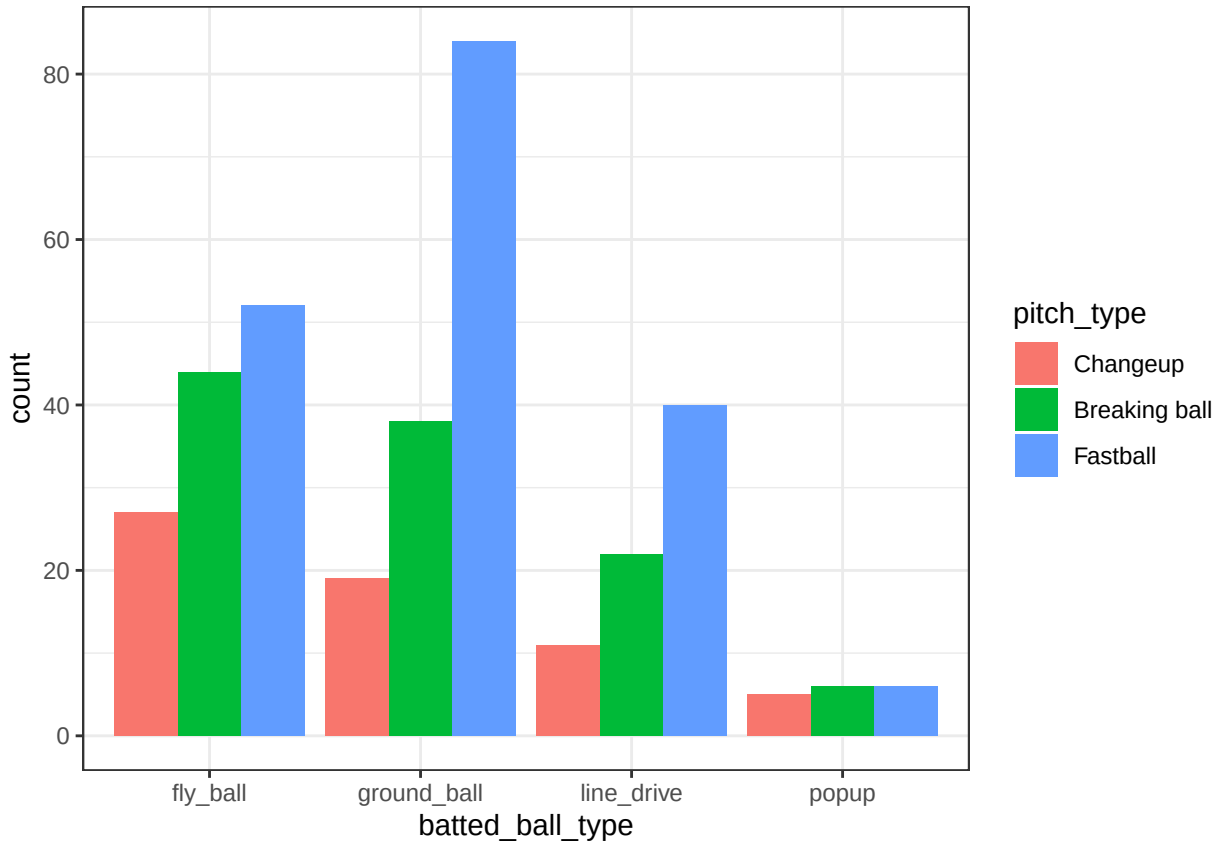
Conclusion: At a significance level of 0.05, we reject the null hypothesis in favor of there being very strong evidence (p-value approximately 0) that all pitch types are *not* equally likely to occur.

2D Categorical Visualization

```
ohtani_batted_balls %>%
ggplot(aes(x = batted_ball_type,
fill = pitch_type)) +
  geom_bar() +
  theme_bw()
```




```
ohtani_batted_balls %>%
ggplot(aes(x = batted_ball_type,
fill = pitch_type)) +
geom_bar(position = "dodge") +
  theme_bw()
```




```
table("Pitch type" = ohtani_batted_balls$pitch_type,
"Batted ball type" = ohtani_batted_balls$batted_ball_type)
```

```
##                Batted ball type
## Pitch type      fly_ball ground_ball line_drive popup
##   Changeup            27          19         11     5
##   Breaking ball       44          38         22     6
##   Fastball            52          84         40     6
```

```
proportions(table(ohtani_batted_balls$pitch_type, ohtani_batted_balls$batted_ball_type))
```

```
##
##                   fly_ball ground_ball line_drive      popup
##   Changeup      0.07627119  0.05367232 0.03107345 0.01412429
##   Breaking ball 0.12429379  0.10734463 0.06214689 0.01694915
##   Fastball      0.14689266  0.23728814 0.11299435 0.01694915
```

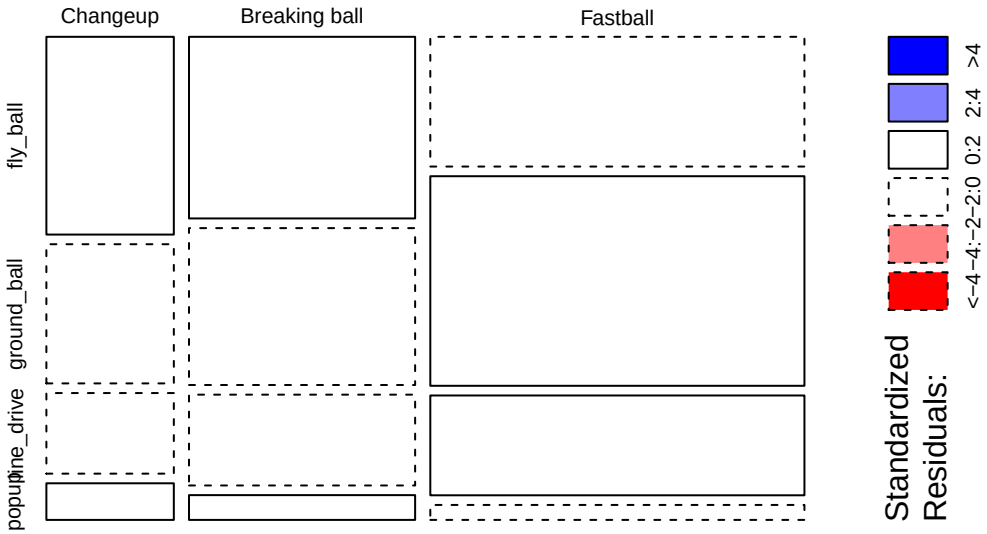
```
## joint probabilities table via dplyr

library(gt)
ohtani_batted_balls %>%
group_by(batted_ball_type, pitch_type) %>%
summarize(joint_prob = n() / nrow(ohtani_batted_balls)) %>%
pivot_wider(names_from = batted_ball_type, values_from = joint_prob,
values_fill = 0) %>%
gt()
```

pitch_type	fly_ball	ground_ball	line_drive	popup

Changeup	0.07627119	0.05367232	0.03107345	0.01412429
Breaking ball	0.12429379	0.10734463	0.06214689	0.01694915
Fastball	0.14689266	0.23728814	0.11299435	0.01694915

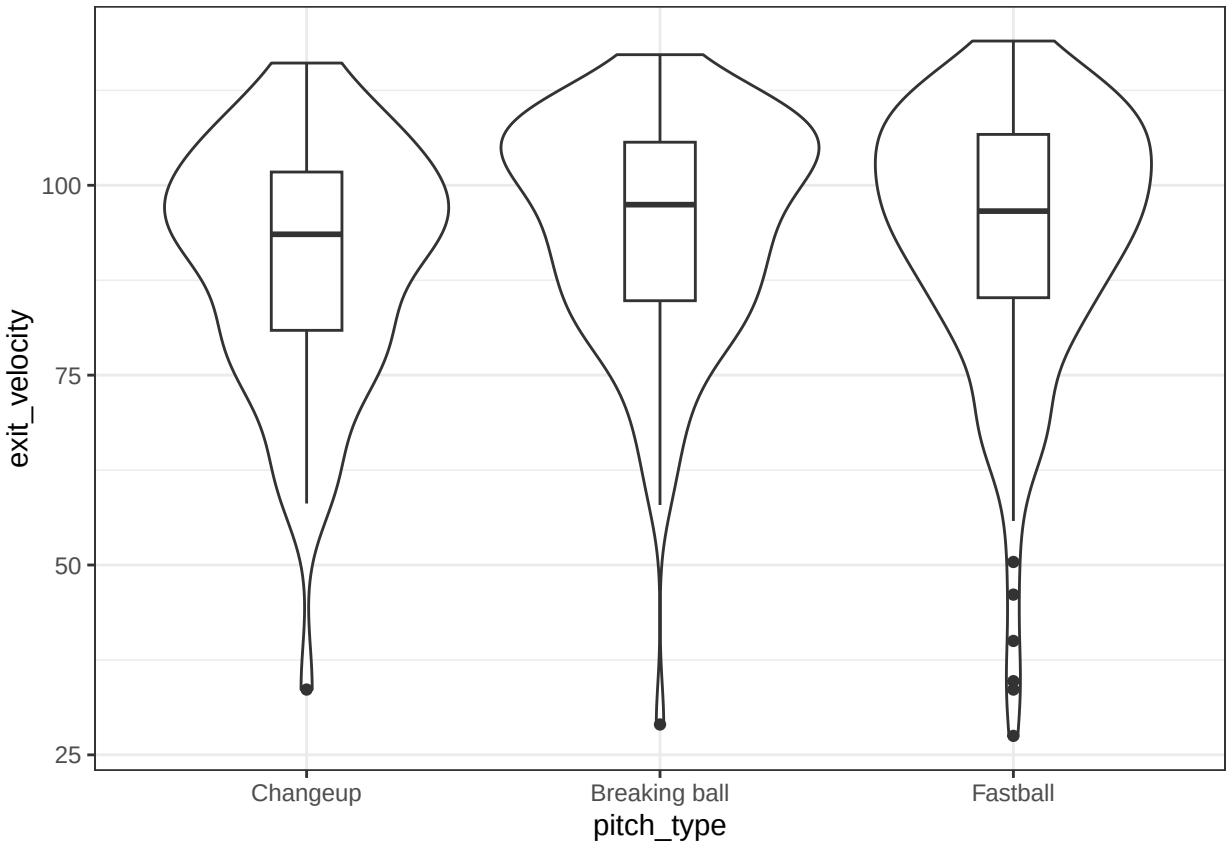
Visualizing Independence between 2 Categorical Variables

```
mosaicplot(table(ohtani_batted_balls$pitch_type, ohtani_batted_balls$batted_ball_type),
           shade = TRUE, #adds pearson residuals to mosiac plot
           # rij approx 0 means observed counts are close to expected counts
           # if abs(rij) > 2 means "signficant" at alpha = 0.05
           main = "Relationship between batted ball and pitch type")
```

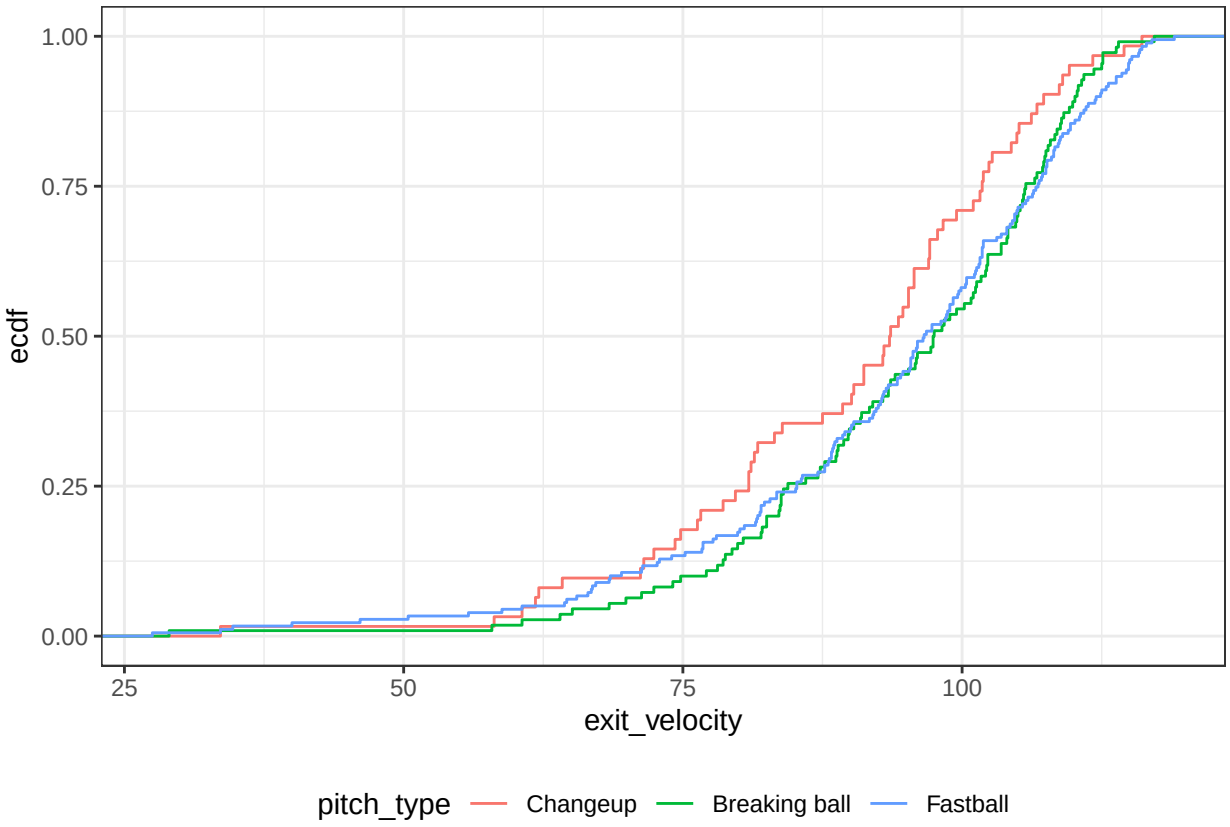


Continuous by Categorical

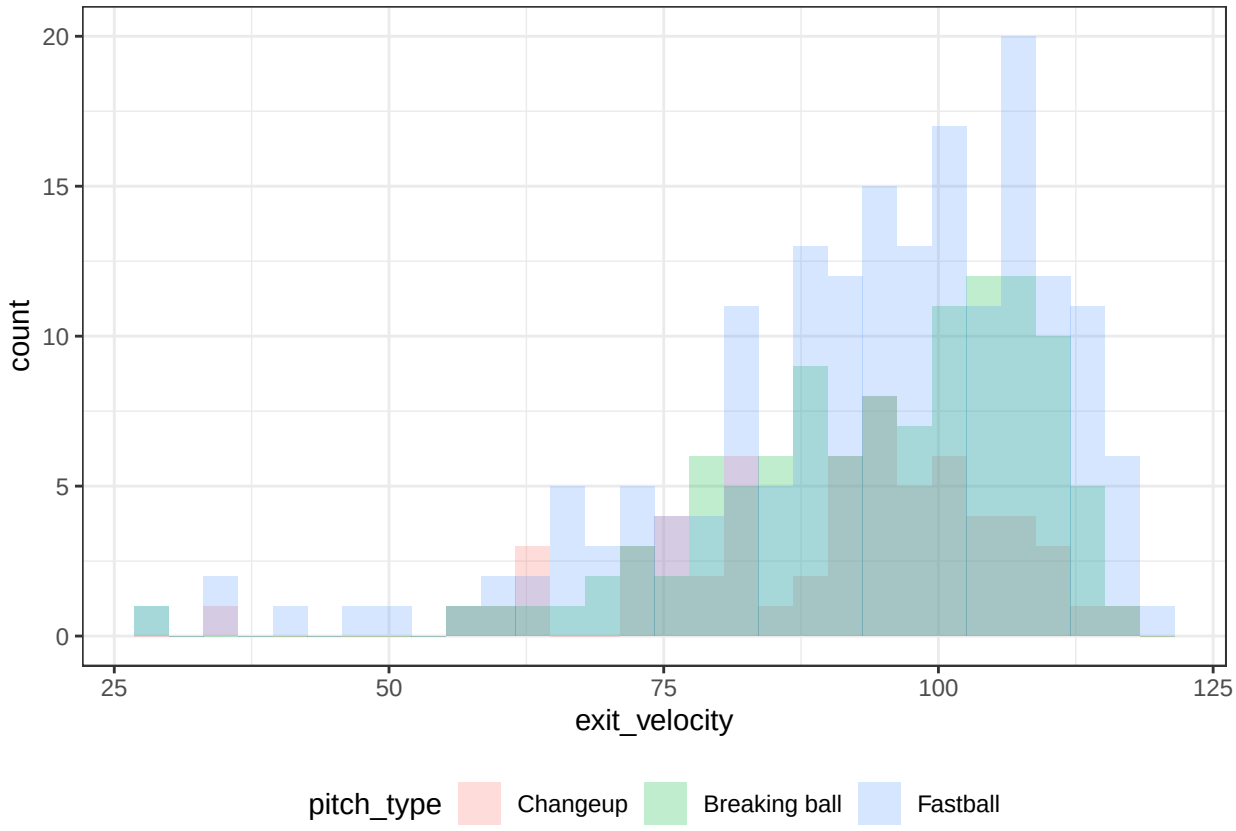
```
ohtani_batted_balls %>%
ggplot(aes(x = pitch_type,
y = exit_velocity)) +
geom_violin() +
geom_boxplot(width = .2)+
  theme_bw()
```



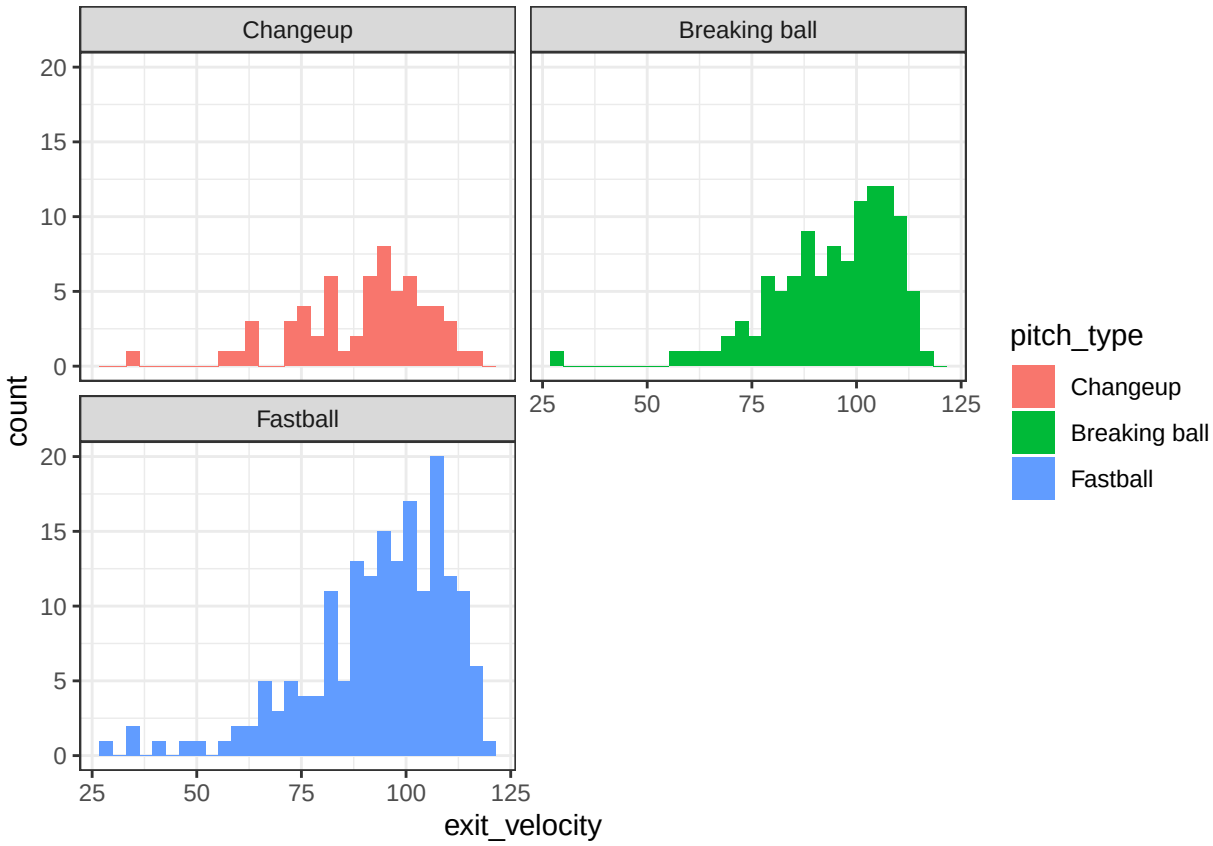

```
ohtani_batted_balls %>%
ggplot(aes(x = exit_velocity,
color = pitch_type)) +
stat_ecdf() +
theme_bw() +
theme(legend.position = "bottom")
```



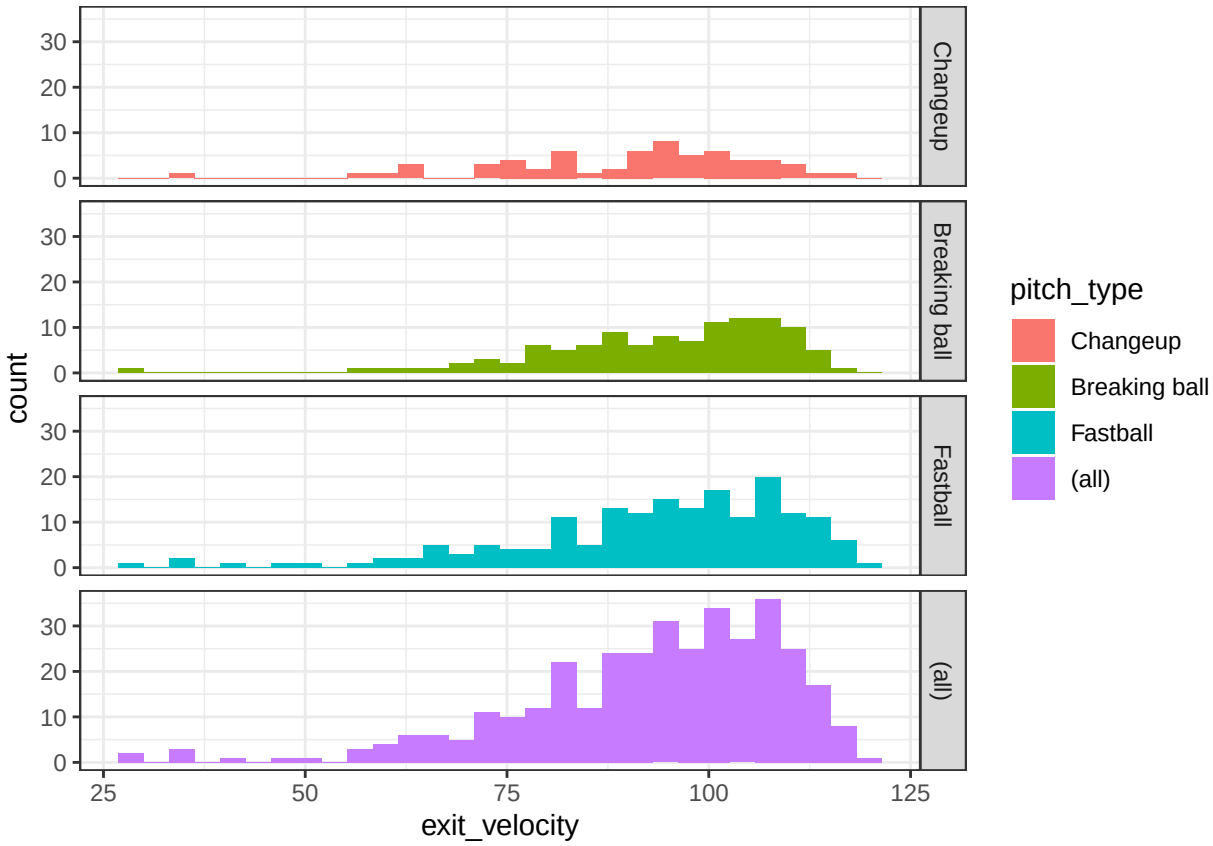

```
ohtani_batted_balls %>%
ggplot(aes(x = exit_velocity,
fill = pitch_type)) +
geom_histogram(alpha = .25, position = "identity") +
theme_bw() +
  theme(legend.position = "bottom")
```




```
ohtani_batted_balls %>%
ggplot(aes(x = exit_velocity)) +
geom_histogram(aes(fill = pitch_type)) +
theme_bw() +
facet_wrap(~ pitch_type, ncol = 2)
```

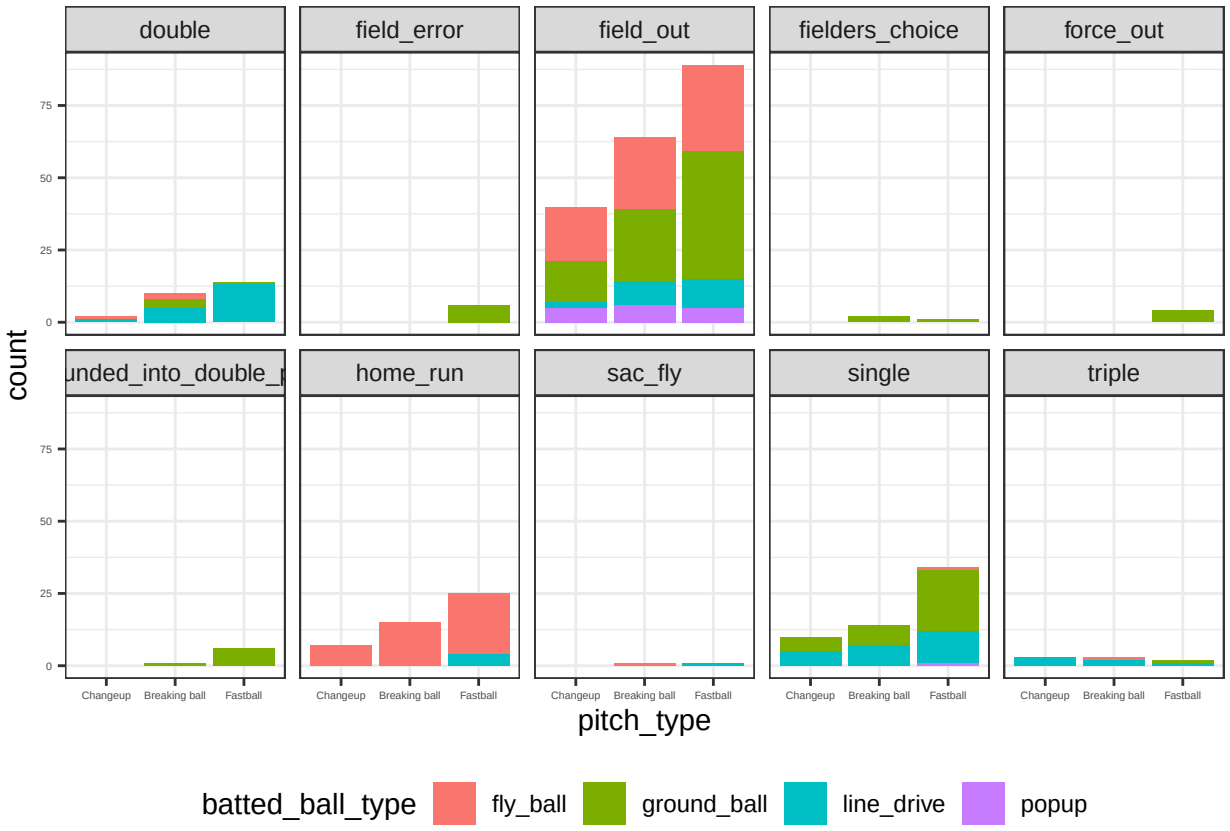



```
ohtani_batted_balls %>%
ggplot(aes(x = exit_velocity)) +
geom_histogram(aes(fill = pitch_type)) +
theme_bw() +
facet_grid(pitch_type ~., margins = TRUE)
```



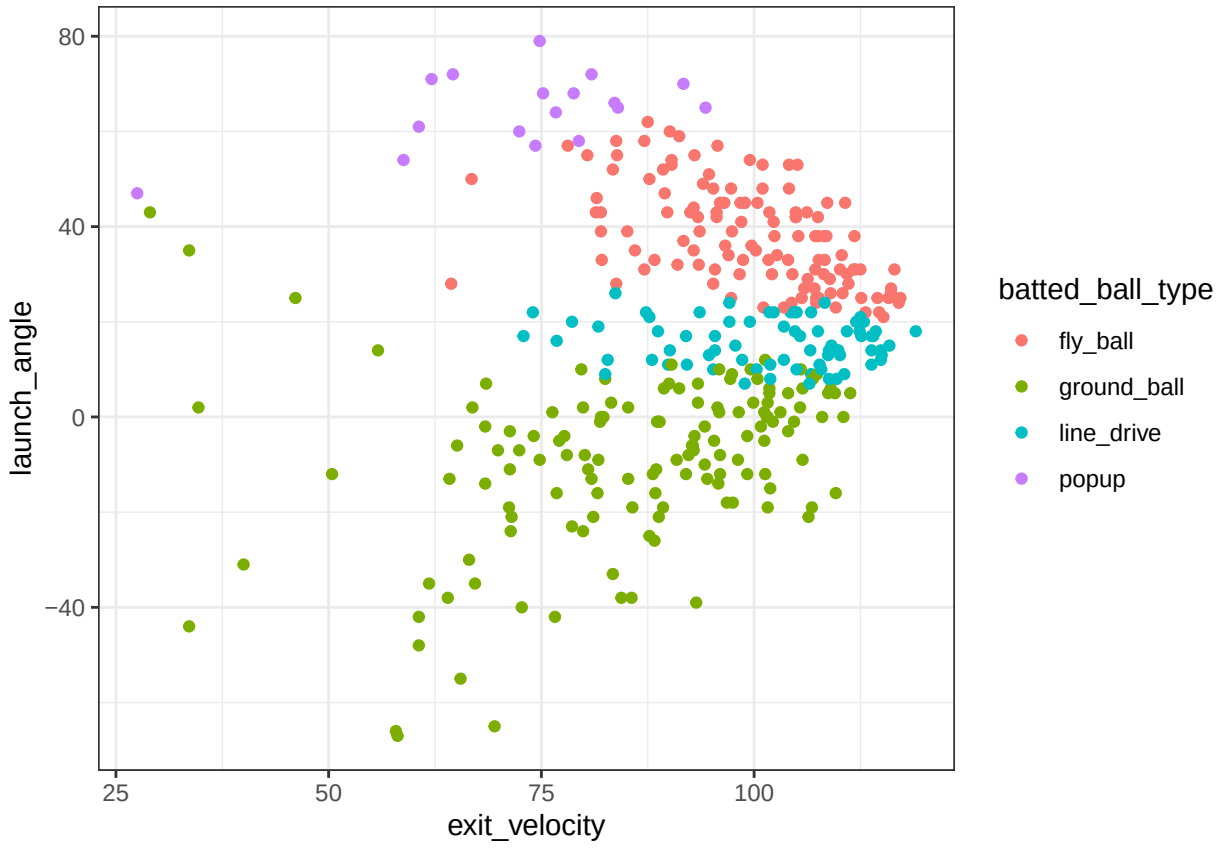
Facets make it easy to move beyond 2D

```
ohtani_batted_balls %>%
ggplot(aes(x = pitch_type,
fill = batted_ball_type)) +
geom_bar() + theme_bw() +
facet_wrap(~ outcome, ncol = 5) +
theme(legend.position = "bottom", axis.text = element_text(size = 4))
```



2D Continuous Relationships

```
ohtani_batted_balls %>%
ggplot(aes(x = exit_velocity,
y = launch_angle)) +
geom_point(aes(color = batted_ball_type)) +
theme_bw()
```


launch_angle
80
40
0
−40
25
50
75
100
exit_velocity
batted_ball_type
fly_ball
ground_ball
line_drive
popup