pokemon dataset

sarah, mars, juniper

Dataset

We are planning to use the Complete Pokemon Dataset that has information on different Pokemon up to Gen 7. The link where we got the dataset is included below. Dataset Link

Research Question

How do different Pokemon's base stats influence capture rate?

```
library(ggplot2)
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(mltools)
library(data.table)
```

```
Attaching package: 'data.table'

The following objects are masked from 'package:dplyr':

    between, first, last
```

```r
library(car)
```

```
Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

The professor recommended doing this so the empty string " " will be converted to a NA value so we can then use the is.na() function.

```r
pokemon <- read.csv("/Users/sarah/Desktop/SDS291/FinalProject/pokemon.csv", na.strings = c
```

Making the dual_type column that says if the Pokemon is a dual type based on if there is a second type.

```r
pokemon <- mutate(pokemon, dual_type = !is.na(type2))
```

Removing legendaries and making pokemone2 dataset

```r
pokemon2 <- filter(pokemon, is_legendary == 0)
```

Making sure that these variables are the right data type (specifically making sure capture rate is numeric)

```
pokemon2$'capture_rate' = as.numeric(pokemon2$'capture_rate')
```

```
Warning: NAs introduced by coercion
```

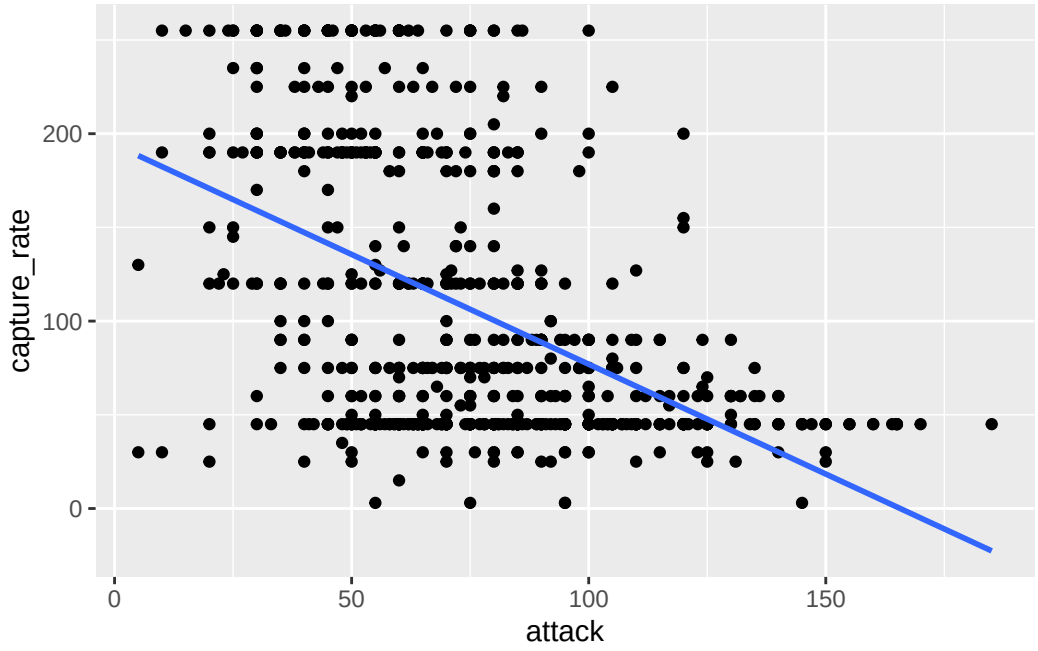
```
pokemon2$'type1' = as.factor(pokemon2$'type1')
pokemon2$'type2' = as.factor(pokemon2$'type2')
```

attack, hp, speed, defense visualizations

```
  ggplot(data = pokemon2, mapping = aes(x = attack, y = capture_rate)) +
    geom_point() +
    geom_smooth(method = lm, se = FALSE, formula = y~x)
```

```
Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

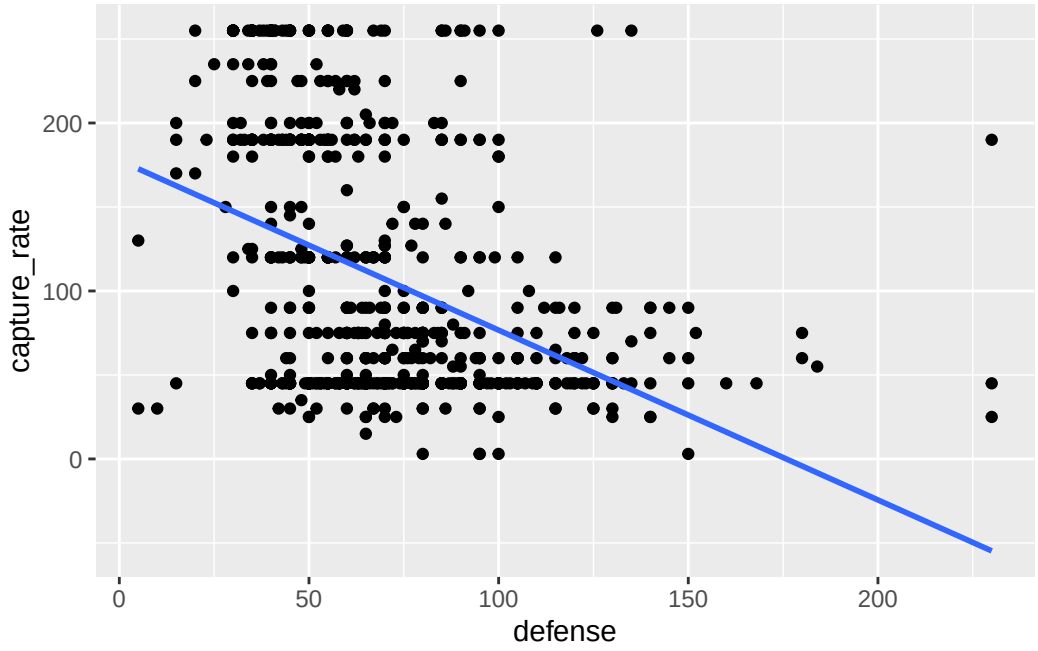
```
Warning: Removed 1 rows containing missing values (geom_point).
```




```
ggplot(data = pokemon2, mapping = aes(x = defense, y = capture_rate)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE, formula = y~x)
```

```
Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

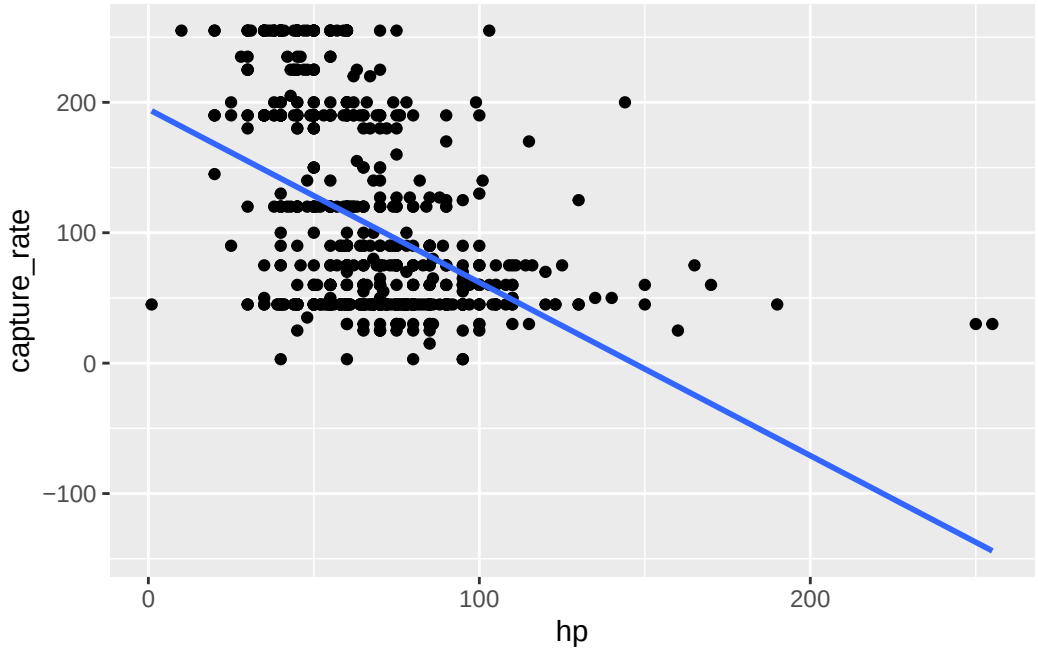
```
Warning: Removed 1 rows containing missing values (geom_point).
```




```
ggplot(data = pokemon2, mapping = aes(x = hp, y = capture_rate)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE, formula = y~x)
```

```
Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

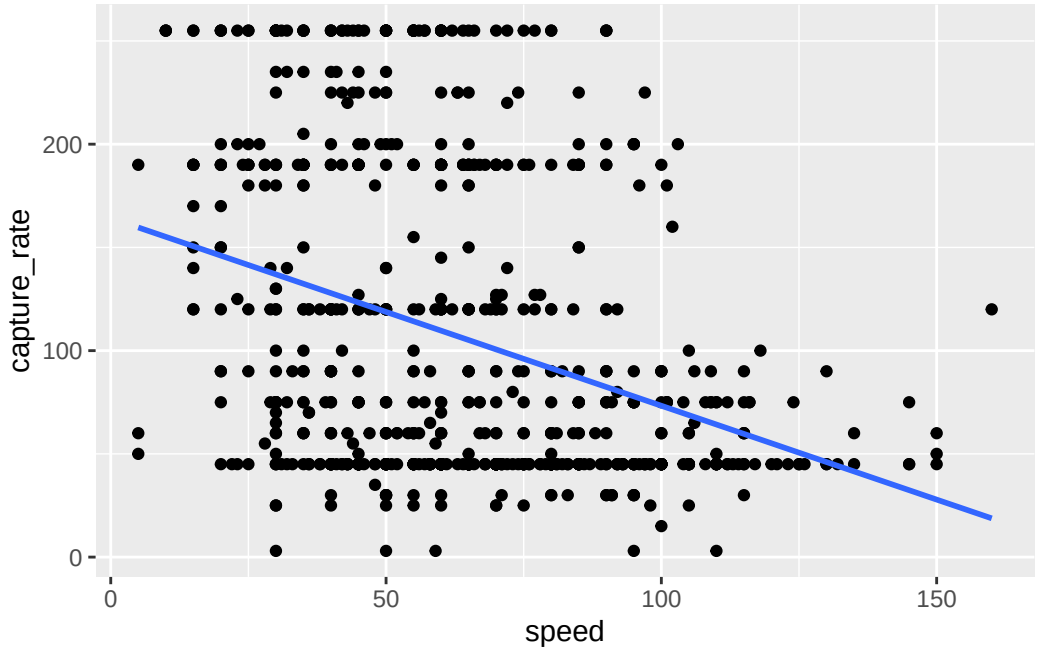
```
Warning: Removed 1 rows containing missing values (geom_point).
```




```
ggplot(data = pokemon2, mapping = aes(x = speed, y = capture_rate)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE, formula = y~x)
```

```
Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
Warning: Removed 1 rows containing missing values (geom_point).
```



Using pokemon2

```
mixed_model1 <- lm(capture_rate ~ dual_type + defense * speed , data = pokemon2)
```

```
summary(mixed_model1)
```

```
Call:
lm(formula = capture_rate ~ dual_type + defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-158.768  -44.858   -9.102   46.216  194.230

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   221.978429  13.955434  15.906  < 2e-16 ***
dual_typeTRUE  -0.278823   4.672326  -0.060   0.9524
defense        -0.761061   0.173289  -4.392 1.29e-05 ***
speed          -0.558421   0.229006  -2.438   0.0150 *
```

```
defense:speed  -0.005936   0.002966  -2.002   0.0457 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.03 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3049,    Adjusted R-squared:  0.301
F-statistic: 79.49 on 4 and 725 DF,  p-value: < 2.2e-16
```

```r
mixed_model2 <- lm(capture_rate ~ attack + defense * speed , data = pokemon2)
```

```r
summary(mixed_model2)
```

```

Call:
lm(formula = capture_rate ~ attack + defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-180.543  -42.316   -5.124   42.023  171.297

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   246.988423  13.914695  17.750  < 2e-16 ***
attack         -0.629191   0.087973  -7.152 2.09e-12 ***
defense        -0.636909   0.167645  -3.799 0.000157 ***
speed          -0.589761   0.221117  -2.667 0.007820 **
defense:speed  -0.002507   0.002905  -0.863 0.388440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.95 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3507,    Adjusted R-squared:  0.3471
F-statistic: 97.89 on 4 and 725 DF,  p-value: < 2.2e-16
```

```r
mixed_model3 <- lm(capture_rate ~ hp + defense * speed , data = pokemon2)
```

```r
summary(mixed_model3)
```

```
Call:
lm(formula = capture_rate ~ hp + defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-171.943  -35.934   -2.573   38.260  223.751

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.920e+02  1.417e+01  20.606  < 2e-16 ***
hp            -1.022e+00  8.852e-02 -11.545  < 2e-16 ***
defense       -8.693e-01  1.588e-01  -5.473 6.09e-08 ***
speed         -8.485e-01  2.117e-01  -4.007 6.77e-05 ***
defense:speed -5.687e-04  2.764e-03  -0.206    0.837
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.01 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4128,    Adjusted R-squared:  0.4096
F-statistic: 127.4 on 4 and 725 DF,  p-value: < 2.2e-16
```

```r
mixed_model4 <- lm(capture_rate~defense * speed , data = pokemon2)
```

```r
summary(mixed_model4)
```

```
Call:
lm(formula = capture_rate ~ defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-158.685  -44.746   -9.216   46.331  194.411

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   221.931622  13.923808  15.939  < 2e-16 ***
defense        -0.762040   0.172392  -4.420 1.14e-05 ***
speed          -0.559078   0.228584  -2.446   0.0147 *
defense:speed  -0.005932   0.002963  -2.002   0.0456 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.99 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3049,    Adjusted R-squared:  0.302
F-statistic: 106.1 on 3 and 726 DF,  p-value: < 2.2e-16
```

First Reduced: defense * speed Full: defense * speed + attack

```
anova(mixed_model4, mixed_model2)
```

```
Analysis of Variance Table

Model 1: capture_rate ~ defense * speed
Model 2: capture_rate ~ attack + defense * speed
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    726 2789829
2    725 2605963  1    183866 51.153 2.093e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is below 0.05 so attack adds to model

Second Reduced: defense * speed Full: defense * speed + dual_type

```
anova(mixed_model4, mixed_model1)
```

```
Analysis of Variance Table

Model 1: capture_rate ~ defense * speed
Model 2: capture_rate ~ dual_type + defense * speed
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    726 2789829
2    725 2789816  1    13.703 0.0036 0.9524
```

p-value is above 0.05 so dual_type doesn't add to model

Second Reduced: defense * speed Full: defense * speed + hp

```
anova(mixed_model4, mixed_model3)
```

```
Analysis of Variance Table

Model 1: capture_rate ~ defense * speed
Model 2: capture_rate ~ hp + defense * speed
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    726 2789829
2    725 2356607  1    433222 133.28 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is below 0.05 so hp does add to model.

adjusted r-squared and defense:speed p-value

base

```
summary(mixed_model4)
```

```
Call:
lm(formula = capture_rate ~ defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-158.685  -44.746   -9.216   46.331  194.411

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   221.931622  13.923808  15.939  < 2e-16 ***
defense        -0.762040   0.172392  -4.420 1.14e-05 ***
speed          -0.559078   0.228584  -2.446   0.0147 *
defense:speed  -0.005932   0.002963  -2.002   0.0456 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.99 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3049,    Adjusted R-squared:  0.302
F-statistic: 106.1 on 3 and 726 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.302 F-statistic: 106.1 on 3 and 726 DF, p-value: < 2.2e-16 p-value for defense:speed: 0.0456

hp

```
summary(mixed_model3)
```

```
Call:
lm(formula = capture_rate ~ hp + defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-171.943  -35.934   -2.573   38.260  223.751

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.920e+02  1.417e+01  20.606  < 2e-16 ***
hp            -1.022e+00  8.852e-02 -11.545  < 2e-16 ***
defense       -8.693e-01  1.588e-01  -5.473 6.09e-08 ***
speed         -8.485e-01  2.117e-01  -4.007 6.77e-05 ***
defense:speed -5.687e-04  2.764e-03  -0.206    0.837
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.01 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4128,    Adjusted R-squared:  0.4096
F-statistic: 127.4 on 4 and 725 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.4096 F-statistic: 127.4 on 4 and 725 DF, p-value: < 2.2e-16 p-value for defense:speed: 0.837

attack

```
summary(mixed_model2)
```

```
Call:
lm(formula = capture_rate ~ attack + defense * speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-180.543  -42.316   -5.124   42.023  171.297

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   246.988423  13.914695  17.750  < 2e-16 ***
attack         -0.629191   0.087973  -7.152 2.09e-12 ***
defense        -0.636909   0.167645  -3.799 0.000157 ***
speed          -0.589761   0.221117  -2.667 0.007820 **
defense:speed  -0.002507   0.002905  -0.863 0.388440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.95 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3507,    Adjusted R-squared:  0.3471
F-statistic: 97.89 on 4 and 725 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.3471 F-statistic: 97.89 on 4 and 725 DF, p-value: $<$ 2.2e-16 p-value for defense:speed: 0.388440

when p-value for defense:speed is above 0.05 so we're changing it

When checking the p-values on the interactive variable, it starts to lose its importance in certain models so we will be changing the interaction term to be parallel/additive.

explanatory variables: hp + defense + speed

```r
add_model3 <- lm(capture_rate ~ hp + defense + speed, data = pokemon2)
```

```r
summary(add_model3)
```

```
Call:
lm(formula = capture_rate ~ hp + defense + speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-172.348  -35.872   -2.588   38.100  223.862

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 294.36239    8.59605   34.24   <2e-16 ***
hp           -1.02500    0.08720  -11.75   <2e-16 ***
defense      -0.89839    0.07213  -12.46   <2e-16 ***
speed        -0.88905    0.07685  -11.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.98 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4128,    Adjusted R-squared:  0.4103
F-statistic: 170.1 on 3 and 726 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.4103

explanatory variables: attack + defense + speed

```r
add_model2 <- lm(capture_rate ~ attack + defense + speed, data = pokemon2)
```

```r
summary(add_model2)
```

```
Call:
lm(formula = capture_rate ~ attack + defense + speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-181.489  -43.022   -5.192   41.736  170.941

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 256.78786    8.04045  31.937  < 2e-16 ***
```

```
attack       -0.64170    0.08675  -7.397 3.86e-13 ***
defense      -0.76146    0.08527  -8.930  < 2e-16 ***
speed        -0.76566    0.08570  -8.934  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.94 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:   0.35, Adjusted R-squared:  0.3473
F-statistic: 130.3 on 3 and 726 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.3473

Nested F-test for hp + defense + speed vs hp + defense * speed

```r
anova(add_model3, mixed_model3)
```

```
Analysis of Variance Table

Model 1: capture_rate ~ hp + defense + speed
Model 2: capture_rate ~ hp + defense * speed
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    726 2356745
2    725 2356607  1    137.58 0.0423 0.8371
```

p-value: 0.8371 Is above 0.05 so the interaction between speed and defense is not needed

testing out interaction between hp and defense for fun

explanatory variables: hp + defense + speed

```r
hp_model3 <- lm(capture_rate ~ hp * defense + speed, data=pokemon2)
```

```r
summary(hp_model3)
```

```
Call:
lm(formula = capture_rate ~ hp * defense + speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
-173.102  -36.213   -2.474   38.236  221.544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.982e+02  1.287e+01  23.176  < 2e-16 ***
hp          -1.082e+00  1.649e-01  -6.558 1.03e-10 ***
defense     -9.645e-01  1.789e-01  -5.392 9.42e-08 ***
speed       -8.922e-01  7.728e-02 -11.544  < 2e-16 ***
hp:defense   9.832e-04  2.434e-03   0.404    0.686
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.01 on 725 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4129,    Adjusted R-squared:  0.4097
F-statistic: 127.5 on 4 and 725 DF,  p-value: < 2.2e-16
```

hp:defense has a p-value of 0.686 so its above 0.05 and not needed.

Best model

The best model is add_model3. We use capture_rate as the outcome variable, and hp, defense, and speed as explanatory variables in an additive model.

```r
add_model3 <- lm(capture_rate ~ hp + defense + speed, data = pokemon2)
```

```r
summary(add_model3)
```

```
Call:
lm(formula = capture_rate ~ hp + defense + speed, data = pokemon2)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-172.348  -35.872   -2.588   38.100  223.862

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 294.36239    8.59605   34.24   <2e-16 ***
hp           -1.02500    0.08720  -11.75   <2e-16 ***
defense      -0.89839    0.07213  -12.46   <2e-16 ***
speed        -0.88905    0.07685  -11.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.98 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4128,    Adjusted R-squared:  0.4103
F-statistic: 170.1 on 3 and 726 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0.4103