Pokemon - Dragon - CP - Analysis

Alan Cheun

2024-02-12

# Load necessary libraries  
library(readr)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ purrr 1.0.2  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(corrplot)

## corrplot 0.92 loaded

Pokemon <- read\_csv("Pokemon.csv")

## Rows: 800 Columns: 13  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): Name, Type 1, Type 2  
## dbl (9): #, Total, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed, Generation  
## lgl (1): Legendary  
##   
## 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.

#View(Pokemon)  
summary(Pokemon)

## # Name Type 1 Type 2   
## Min. : 1.0 Length:800 Length:800 Length:800   
## 1st Qu.:184.8 Class :character Class :character Class :character   
## Median :364.5 Mode :character Mode :character Mode :character   
## Mean :362.8   
## 3rd Qu.:539.2   
## Max. :721.0   
## Total HP Attack Defense   
## Min. :180.0 Min. : 1.00 Min. : 5 Min. : 5.00   
## 1st Qu.:330.0 1st Qu.: 50.00 1st Qu.: 55 1st Qu.: 50.00   
## Median :450.0 Median : 65.00 Median : 75 Median : 70.00   
## Mean :435.1 Mean : 69.26 Mean : 79 Mean : 73.84   
## 3rd Qu.:515.0 3rd Qu.: 80.00 3rd Qu.:100 3rd Qu.: 90.00   
## Max. :780.0 Max. :255.00 Max. :190 Max. :230.00   
## Sp. Atk Sp. Def Speed Generation   
## Min. : 10.00 Min. : 20.0 Min. : 5.00 Min. :1.000   
## 1st Qu.: 49.75 1st Qu.: 50.0 1st Qu.: 45.00 1st Qu.:2.000   
## Median : 65.00 Median : 70.0 Median : 65.00 Median :3.000   
## Mean : 72.82 Mean : 71.9 Mean : 68.28 Mean :3.324   
## 3rd Qu.: 95.00 3rd Qu.: 90.0 3rd Qu.: 90.00 3rd Qu.:5.000   
## Max. :194.00 Max. :230.0 Max. :180.00 Max. :6.000   
## Legendary   
## Mode :logical   
## FALSE:735   
## TRUE :65   
##   
##   
##

str(Pokemon)

## spc\_tbl\_ [800 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ # : num [1:800] 1 2 3 3 4 5 6 6 6 7 ...  
## $ Name : chr [1:800] "Bulbasaur" "Ivysaur" "Venusaur" "VenusaurMega Venusaur" ...  
## $ Type 1 : chr [1:800] "Grass" "Grass" "Grass" "Grass" ...  
## $ Type 2 : chr [1:800] "Poison" "Poison" "Poison" "Poison" ...  
## $ Total : num [1:800] 318 405 525 625 309 405 534 634 634 314 ...  
## $ HP : num [1:800] 45 60 80 80 39 58 78 78 78 44 ...  
## $ Attack : num [1:800] 49 62 82 100 52 64 84 130 104 48 ...  
## $ Defense : num [1:800] 49 63 83 123 43 58 78 111 78 65 ...  
## $ Sp. Atk : num [1:800] 65 80 100 122 60 80 109 130 159 50 ...  
## $ Sp. Def : num [1:800] 65 80 100 120 50 65 85 85 115 64 ...  
## $ Speed : num [1:800] 45 60 80 80 65 80 100 100 100 43 ...  
## $ Generation: num [1:800] 1 1 1 1 1 1 1 1 1 1 ...  
## $ Legendary : logi [1:800] FALSE FALSE FALSE FALSE FALSE FALSE ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `#` = col\_double(),  
## .. Name = col\_character(),  
## .. `Type 1` = col\_character(),  
## .. `Type 2` = col\_character(),  
## .. Total = col\_double(),  
## .. HP = col\_double(),  
## .. Attack = col\_double(),  
## .. Defense = col\_double(),  
## .. `Sp. Atk` = col\_double(),  
## .. `Sp. Def` = col\_double(),  
## .. Speed = col\_double(),  
## .. Generation = col\_double(),  
## .. Legendary = col\_logical()  
## .. )  
## - attr(\*, "problems")=<externalptr>

# Task III

## Introduction

In the realm of Pokémon battles, understanding Pokémon types is paramount. Pokémon types encompass various elements or specializations that a Pokémon possesses, ranging from nature-based categories like fire, water, and ground to domain-based classifications such as fighting, ghost, and psychic. These typings play a crucial role, offering distinct advantages or disadvantages in battles against other Pokémon.

Considered the foundational principle of Pokémon battling, understanding Pokémon typings is indispensable.

The objective of this project is to conduct a regression analysis to explore the relationship between Combat Power (CP), an overall measure of the effectiveness of a Pokémon in the Pokémon game, and selected Typing from the Pokémon dataset. Combat Power is a crucial metric in Pokémon, representing the battle strength of a Pokémon. This analysis aims to identify significant factors influencing CP and provide insights into how Total stats and explore if Typing contribute to the Combat Power of Pokémon.

We will explore Dragon, considered high in rank and of formidable power, not just in Pokémon lore, across many mythologies and legends.  
We will explore Bug, considered often weak and insignificant, this will provide a sound opposite to the Dragon typing.

Finally explore Water, this is the most prevalent typing across the first 6 generations of Pokémon .

## Dataset Selection:

For this analysis, we have chosen the Pokémon dataset, which contains information about various Pokémon species, their attributes, and Combat Power. The dataset includes both categorical and numeric variables, making it suitable for exploring relationships through regression analysis. Specifically, we are interested in understanding how the total stats (Total) and whether a Pokémon typing impacts its Combat Power.

### Data Wrangling

#CP function  
Combat\_Power <- function(attack, special\_atk, defense, special\_def, hp) {  
 total\_a <- attack + special\_atk  
 total\_d <- defense + special\_def  
 numerator <- floor(total\_a \* (total\_d^0.5) \* (hp^0.5) \* (0.667934^2))  
 return(floor(max(10, numerator / 10)))  
}  
#Add Combat Power to dataset and isTyping indicator  
Pokemon <- Pokemon %>%  
 mutate(isDragon = ifelse(`Type 1` == "Dragon" | `Type 2` == "Dragon", 1, 0)) %>%  
 mutate(isDragon = as.integer(ifelse(is.na(isDragon), 0, isDragon))) %>%  
 rowwise() %>%  
 mutate(isBug = ifelse(`Type 1` == "Bug" | `Type 2` == "Bug", 1, 0)) %>%  
 mutate(isBug = as.integer(ifelse(is.na(isBug), 0, isBug))) %>%  
 rowwise() %>%  
 mutate(isWater = ifelse(`Type 1` == "Water" | `Type 2` == "Water", 1, 0)) %>%  
 mutate(isWater = as.integer(ifelse(is.na(isWater), 0, isWater))) %>%  
 rowwise() %>%  
 mutate(CP = Combat\_Power(Attack, `Sp. Atk`, Defense, `Sp. Def`, HP))  
  
# Convert isDragon to a factor with levels "0" and "1"  
Pokemon$isDragon <- as.factor(Pokemon$isDragon)  
Pokemon$isBug <- as.factor(Pokemon$isBug)  
Pokemon$isWater <- as.factor(Pokemon$isWater)  
  
# Print the first few rows to check if CP values vary  
head(Pokemon)

## # A tibble: 6 × 17  
## # Rowwise:   
## `#` Name `Type 1` `Type 2` Total HP Attack Defense `Sp. Atk` `Sp. Def`  
## <dbl> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 Bulbas… Grass Poison 318 45 49 49 65 65  
## 2 2 Ivysaur Grass Poison 405 60 62 63 80 80  
## 3 3 Venusa… Grass Poison 525 80 82 83 100 100  
## 4 3 Venusa… Grass Poison 625 80 100 123 122 120  
## 5 4 Charma… Fire <NA> 309 39 52 43 60 50  
## 6 5 Charme… Fire <NA> 405 58 64 58 80 65  
## # ℹ 7 more variables: Speed <dbl>, Generation <dbl>, Legendary <lgl>,  
## # isDragon <fct>, isBug <fct>, isWater <fct>, CP <dbl>

We are to add dummy variables that will act as a Categorical variable in the regression model analysis.

**isDragon** is whether Type 1 or Type 2 is equal to Dragon, then set to 1 for True, 0 for False.

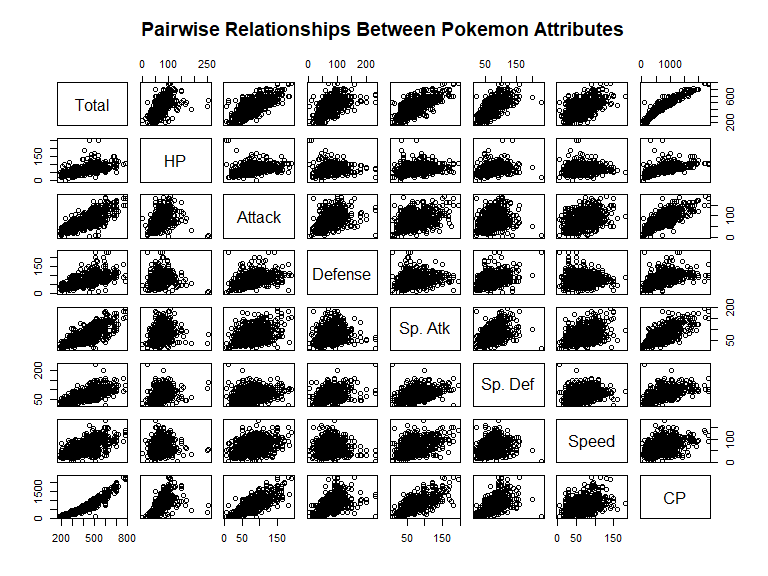
**isBug** is whether Type 1 or Type 2 is equal to Bug, then set to 1 for True, 0 for False.

**isWater** is whether Type 1 or Type 2 is equal to Water, then set to 1 for True, 0 for False.

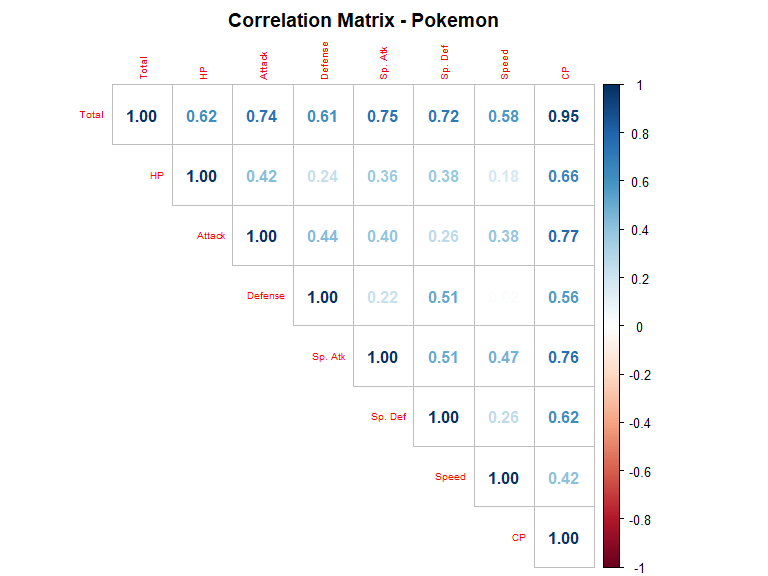
## Regression Model:

### Assumptions and Validation for Regression Model

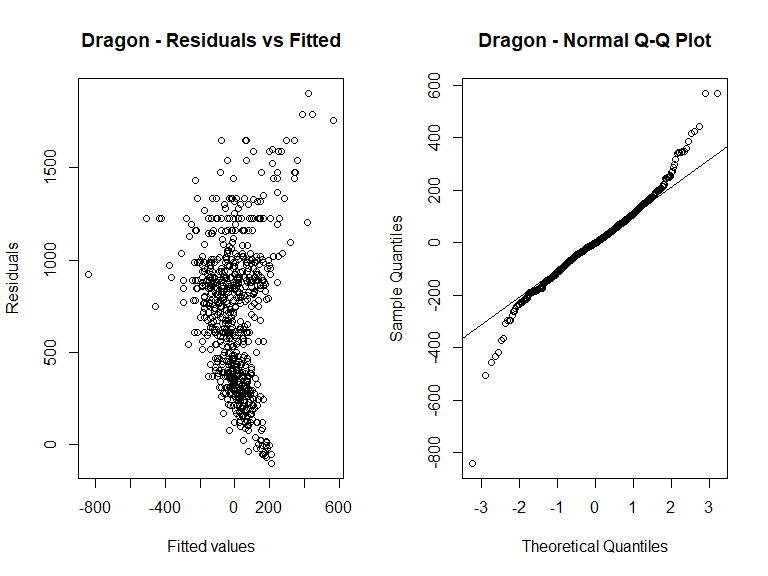
selected\_columns <- c("Total", "HP", "Attack", "Defense", "Sp. Atk", "Sp. Def", "Speed", "CP")  
  
# Create a scatterplot matrix with linear regression lines  
pairs(Pokemon[selected\_columns],main="Pairwise Relationships Between Pokemon Attributes")



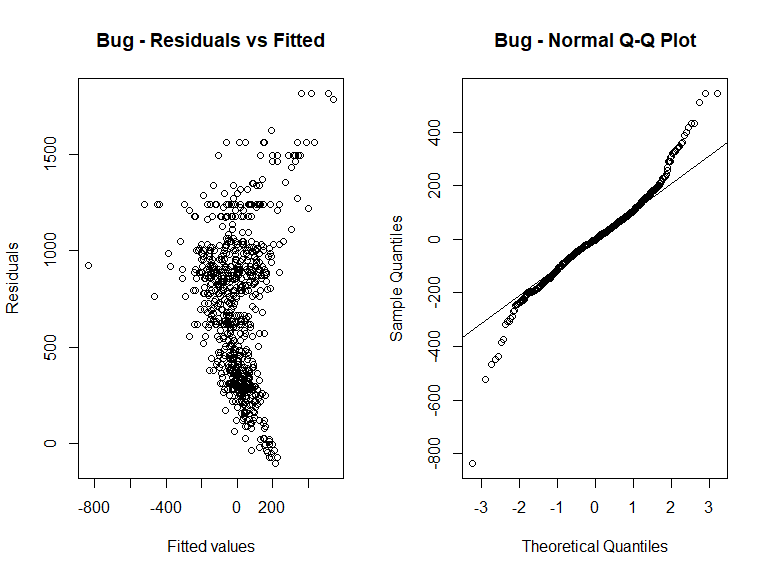
#Correlation   
#cor(Pokemon[selected\_columns])  
correlation\_matrix <- cor(Pokemon[selected\_columns])  
# Create a compact correlation plot (heatmap)  
corrplot(correlation\_matrix, method = "number", type = "upper", tl.cex = 0.6, main = "Correlation Matrix - Pokemon",mar = c(0, 0, 2, 2))



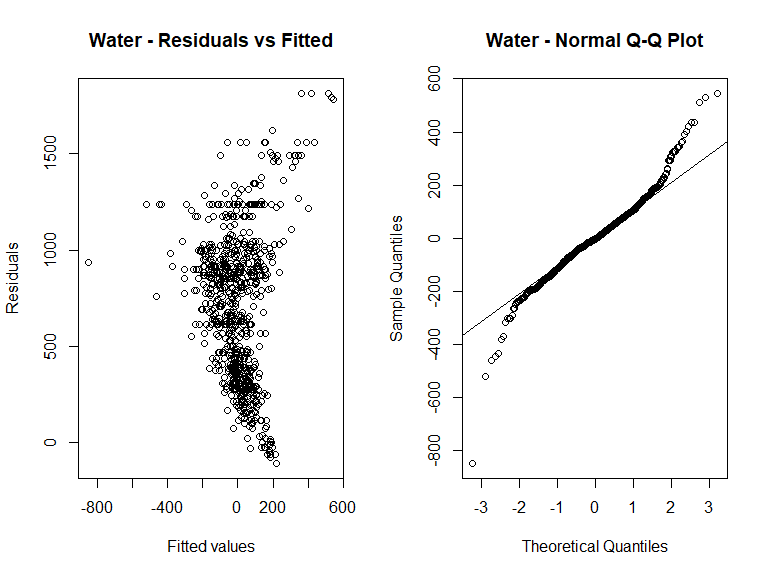
# Perform regression analysis  
cp\_dragon\_model <- lm(CP ~ Total + isDragon, data = Pokemon)  
cp\_bug\_model <- lm(CP ~ Total + isBug, data = Pokemon)  
cp\_water\_model <- lm(CP ~ Total + isWater, data = Pokemon)  
  
  
par(mfrow = c(1, 2))   
  
# Print regression summary for Dragon  
plot(cp\_dragon\_model$residuals, cp\_dragon\_model$fitted.values, main = "Dragon - Residuals vs Fitted", xlab = "Fitted values", ylab = "Residuals")  
# Plot Q-Q Residuals for Dragon model  
qqnorm(cp\_dragon\_model$residuals,main = "Dragon - Normal Q-Q Plot")  
qqline(cp\_dragon\_model$residuals)



# Print regression summary for Bug  
plot(cp\_bug\_model$residuals, cp\_bug\_model$fitted.values, main = "Bug - Residuals vs Fitted", xlab = "Fitted values", ylab = "Residuals")  
# Plot Q-Q Residuals for Bug model  
qqnorm(cp\_bug\_model$residuals,main = "Bug - Normal Q-Q Plot")  
qqline(cp\_bug\_model$residuals)



# Print regression summary for Water  
plot(cp\_water\_model$residuals, cp\_water\_model$fitted.values, main = "Water - Residuals vs Fitted", xlab = "Fitted values", ylab = "Residuals")  
# Plot Q-Q Residuals for Water model  
qqnorm(cp\_water\_model$residuals,main = "Water - Normal Q-Q Plot")  
qqline(cp\_water\_model$residuals)



In the scatterplots from “*Scatterplots for Pairwise Relationships Between* Pokémon *Attributes*”, there is a pattern apparent between Total and CP, while all other pairwise plots do not have a clear pattern identified. This provides us enough evidence that the Total attribute of the Pokémon should be included in a linear regression formula for predicting CP.

As we look further into In the Residuals vs Fitted plot into the focused Typings, the scatter plot does not indicate any sign of a pattern, thus residuals are independent, there is equal variances.  
This is a positive sign that the requirements for linear regression model is met.  
  
In the Q-Q Residuals plot, the points are falling close along the line, thus indicating normal distribution of the residuals for us to use certain statistical methods and tests.

### Estimate the Model

# Print regression summary for Water  
cat("LM - CP - Dragon Model")

## LM - CP - Dragon Model

summary(cp\_dragon\_model)

##   
## Call:  
## lm(formula = CP ~ Total + isDragon, data = Pokemon)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -839.61 -70.08 -2.21 71.39 567.69   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -666.06293 17.17480 -38.781 < 2e-16 \*\*\*  
## Total 3.14984 0.03863 81.545 < 2e-16 \*\*\*  
## isDragon1 111.96653 19.13106 5.853 7.07e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 127.5 on 797 degrees of freedom  
## Multiple R-squared: 0.9014, Adjusted R-squared: 0.9012   
## F-statistic: 3644 on 2 and 797 DF, p-value: < 2.2e-16

cat("LM - CP - Bug Model\n")

## LM - CP - Bug Model

summary(cp\_bug\_model)

##   
## Call:  
## lm(formula = CP ~ Total + isBug, data = Pokemon)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -835.74 -71.60 -2.62 69.08 543.71   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -677.91357 17.75633 -38.179 <2e-16 \*\*\*  
## Total 3.19637 0.03879 82.411 <2e-16 \*\*\*  
## isBug1 -15.51616 16.24821 -0.955 0.34   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.1 on 797 degrees of freedom  
## Multiple R-squared: 0.8973, Adjusted R-squared: 0.897   
## F-statistic: 3482 on 2 and 797 DF, p-value: < 2.2e-16

cat("LM - CP - Water Model")

## LM - CP - Water Model

summary(cp\_water\_model)

##   
## Call:  
## lm(formula = CP ~ Total + isWater, data = Pokemon)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -848.11 -71.40 -2.85 70.50 545.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -684.24405 17.47329 -39.159 <2e-16 \*\*\*  
## Total 3.20268 0.03837 83.461 <2e-16 \*\*\*  
## isWater1 13.88566 12.62929 1.099 0.272   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.1 on 797 degrees of freedom  
## Multiple R-squared: 0.8973, Adjusted R-squared: 0.8971   
## F-statistic: 3483 on 2 and 797 DF, p-value: < 2.2e-16

We have employed a linear regression model to investigate the associations between CP and the chosen predictors. The regression equation is given by:

#### Dragon

The p-values for the intercept, Total, and isDragon1 are all reported as ‘< 2e-16’, which means they are very close to zero. This suggests strong evidence to reject the null hypothesis that the corresponding coefficients are zero.

The Multiple R-squared value is 0.9014, indicating that approximately 90.14% of the variability in the response variable (CP) is explained by the linear regression model.

#### Bug

The p-values for the intercept and Total are both reported as ‘< 2e-16’, suggesting strong evidence to reject the null hypothesis that the corresponding coefficients are zero.

The p-value for isBug1 is 0.34, which is greater than 0.05, indicating that the *coefficient for isBug1 is not statistically significant*.

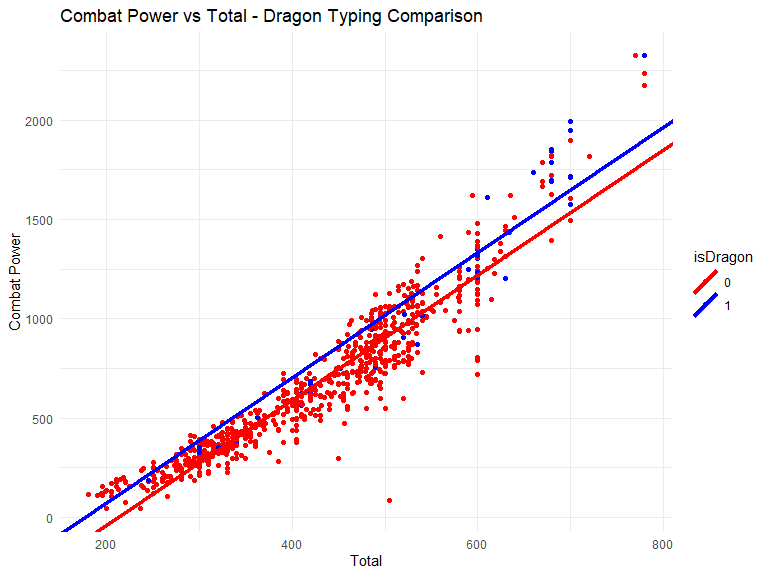
#### Water

The p-values for the intercept and Total are both reported as ‘< 2e-16’, suggesting strong evidence to reject the null hypothesis that the corresponding coefficients are zero.

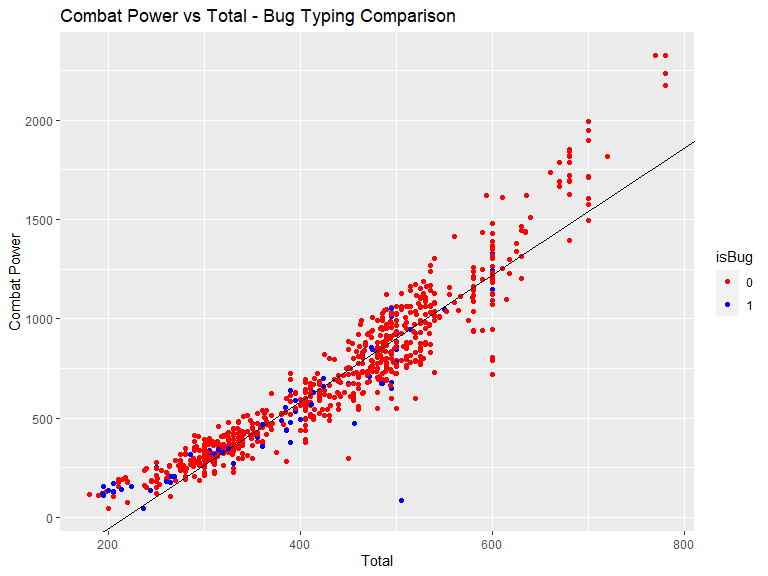
The p-value for isWater1 is 0.272, which is greater than 0.05, indicating that the coefficient for *isWater1 is not statistically significant*.

cb\_palette <- c("red", "blue")  
  
ggplot(Pokemon, aes(Total, CP, col = isDragon)) +  
 geom\_point() +  
 geom\_abline(aes(slope = 3.14984, intercept = -666.06293, col = "0"), size = 1.5) +  
 geom\_abline(aes(slope = 3.14984, intercept = -666.06293 + 111.96653, col = "1"), size = 1.5) +  
 labs(title = "Combat Power vs Total - Dragon Typing Comparison",  
 x = "Total",  
 y = "Combat Power") +  
 scale\_color\_manual(values = cb\_palette) +  
 theme\_minimal() # You can customize the theme as needed

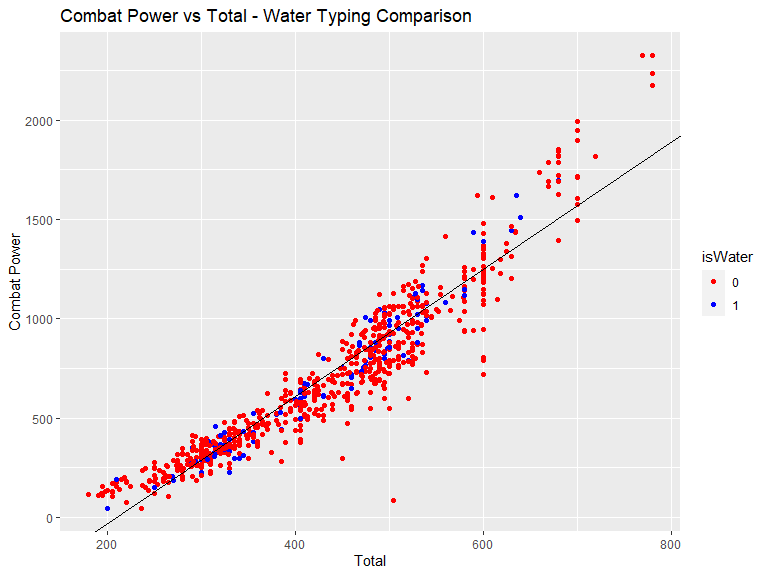
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



ggplot(Pokemon, aes(Total, CP, col = isBug)) +  
 geom\_point() +  
 geom\_abline(aes(slope = 3.19637, intercept = -677.91357), col = "0") +  
 geom\_abline(aes(slope = 3.19637, intercept = -677.91357 + -15.51616), col = "1") +  
 labs(title = "Combat Power vs Total - Bug Typing Comparison",  
 x = "Total",  
 y = "Combat Power") +  
 scale\_color\_manual(values = cb\_palette)



ggplot(Pokemon, aes(Total, CP, col = isWater)) +  
 geom\_point() +  
 geom\_abline(aes(slope = 3.20268, intercept = -684.2440), col = "0") +  
 geom\_abline(aes(slope = 3.20268, intercept = -684.2440 + 13.88566), col = "1") +  
 labs(title = "Combat Power vs Total - Water Typing Comparison",  
 x = "Total",  
 y = "Combat Power") +  
 scale\_color\_manual(values = cb\_palette)



In summary, the p-values for the Total in all models are well below 5% significance level and thus strongly support with statistical evidence that the coefficient of Total to CP is not zero.

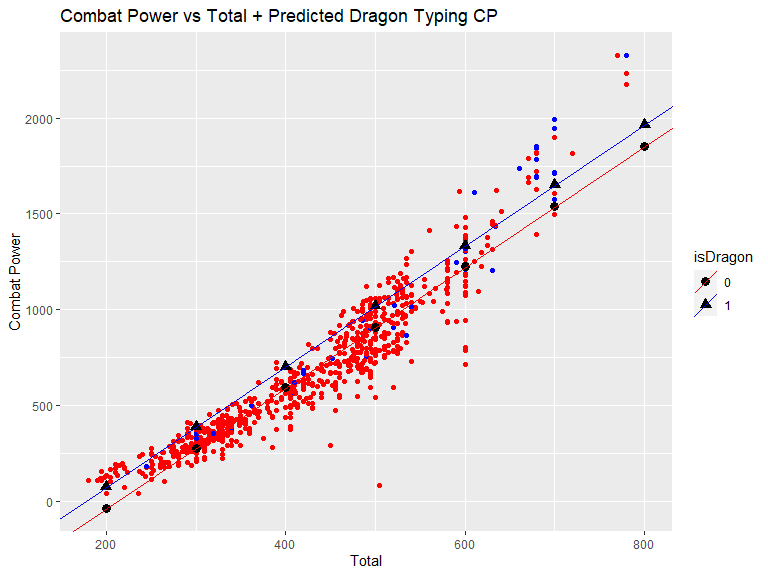
The Multiple R-squared values is around ~0.90, indicating that approximately 90% of the variability in the response variable (CP) is explained by the linear regression model. This is consistent with expectations given the strong coefficient correlation between Total and Combat Power.

Analyzing the three regression models, the Dragon typing had a p-value that strongly suggests there is a non-zero relationship between a Pokémon being of a Dragon typing and its impact to its CP. While the other Pokémon typings used in comparison show no statistical evidence for a relationship between a Bug or Water typing to its CP. This lack of correlation for Bug and Water can be visualized by the lack of any gap between the regression lines between its true or false status in Pokémon typing.

#Dragon  
cat("Dragon CP Prediction")

## Dragon CP Prediction

new\_CP\_Dragon\_Pokemon <-tibble(isDragon = c("0","1","0","1","0","1","0","1","0","1","0","1","0","1"),Total = c(200,200,300,300, 400,400,500,500,600,600,700,700,800,800))  
  
new\_CP\_Dragon\_Pokemon$Predicted\_CP <- predict(cp\_dragon\_model, newdata = new\_CP\_Dragon\_Pokemon)  
  
#Bug  
#cat("Bug CP Prediction")  
#predict\_CP\_Bug\_Pokemon <-tibble(isBug = c("0","1","0","1","0","1","0","1"),Total = c(200,200, 400,400, 600,600,800,800))  
#predict(cp\_bug\_model,predict\_CP\_Bug\_Pokemon)  
#Water  
#cat("Water CP Prediction")  
#predict\_CP\_Water\_Pokemon <-tibble(isWater = c("0","1","0","1","0","1","0","1"),Total = c(200,200, 400,400, 600,600,800,800))  
#predict(cp\_water\_model,predict\_CP\_Water\_Pokemon)  
# Plotting the ggplot with both observed and predicted values  
# Convert isDragon to factor for better plotting  
new\_CP\_Dragon\_Pokemon$isDragon <- factor(new\_CP\_Dragon\_Pokemon$isDragon)  
  
# Plotting the ggplot with both observed and predicted values  
ggplot() +  
 geom\_point(data = Pokemon, aes(x = Total, y = CP, col = isDragon)) +  
 geom\_point(data = new\_CP\_Dragon\_Pokemon, aes(x = Total, y = Predicted\_CP, shape = isDragon), col = "black", size = 3) +  
 geom\_abline(aes(slope = 3.14984, intercept = -666.06293, col = "0")) +  
 geom\_abline(aes(slope = 3.14984, intercept = -666.06293 + 111.96653, col = "1")) +  
 labs(title = "Combat Power vs Total + Predicted Dragon Typing CP ",  
 x = "Total",  
 y = "Combat Power") +  
 scale\_color\_manual(values = cb\_palette)



## Conclusion:

This regression analysis contributes to our understanding of the factors influencing Combat Power in Pokémon. The highly significant Total and isDragon variables indicate their importance in predicting Combat Power and have a possibility to be deployed. The regression models in utilizing Bug and Water typing are rejected to be deployed as it failed to reject the null hypothesis that its coefficient is significant.