H3 - Wins vs Losses (ESports significance)

```
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(tidyr)
aoe2 <- read.csv("../Data/aoe2_leaderboard_sample.csv")
```

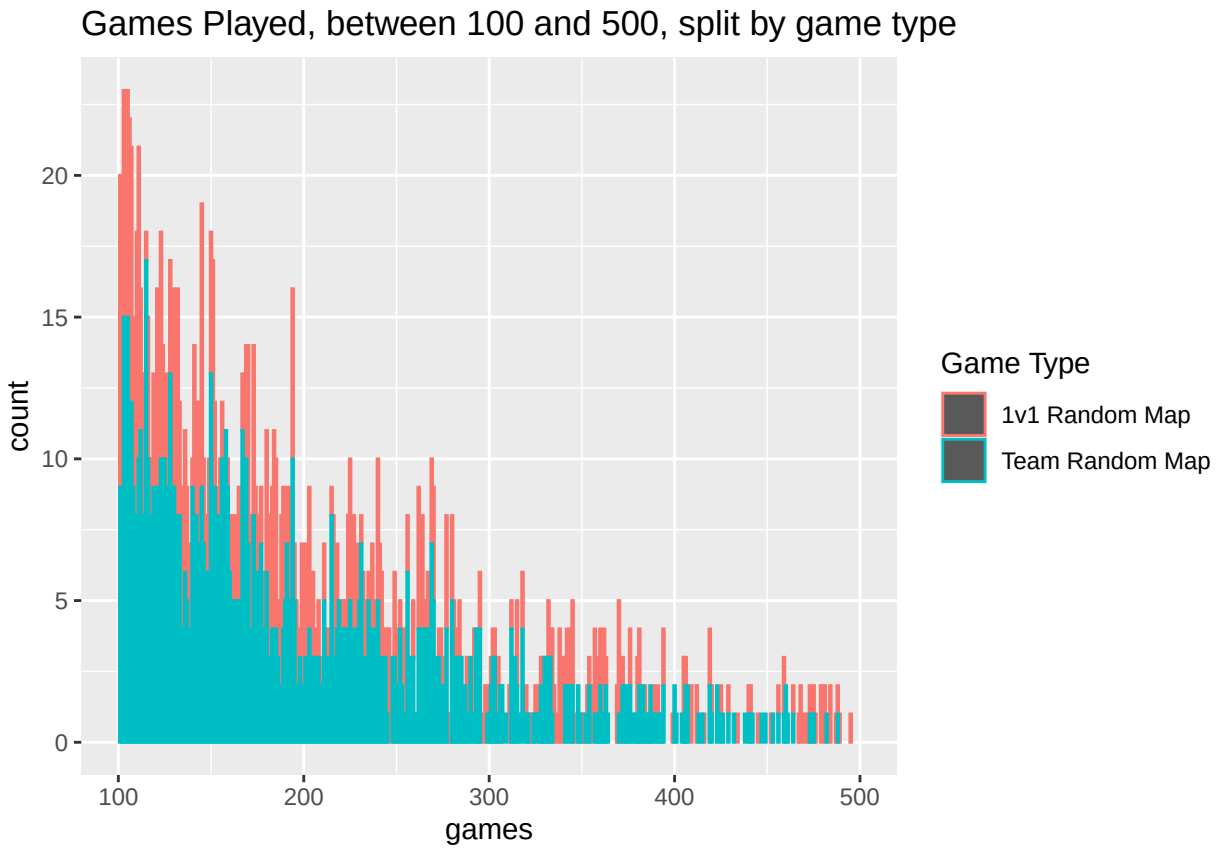
After exploring the popularity by country, we saw that Denmark, who played more average games, had more average wins and losses which makes sense because ESports, unlike other sports, do not have a set amount of games when looking at leaderboards. Therefore, let's see what we can see from this.

We first show the inconsistency of games played, even for just the average player that may play anywhere in between 100 and 500 games.

```
ggplot(aoe2, aes(games, color = game_type)) +
  geom_bar() +
  xlim(100, 500) +
  ggtitle("Games Played, between 100 and 500, split by game type") +
  labs(color = "Game Type")
```

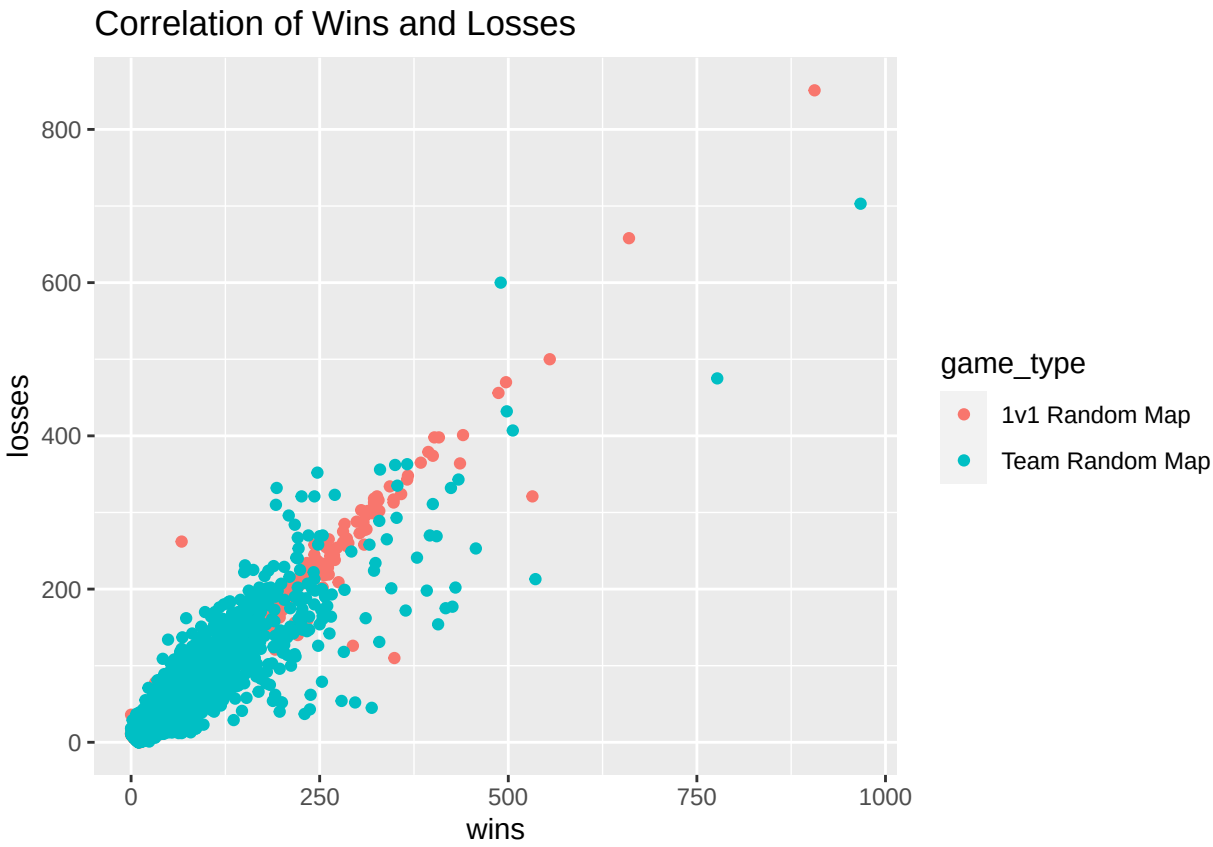
```
## Warning: Removed 4691 rows containing non-finite values (stat_count).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Basic Correlation Graph:

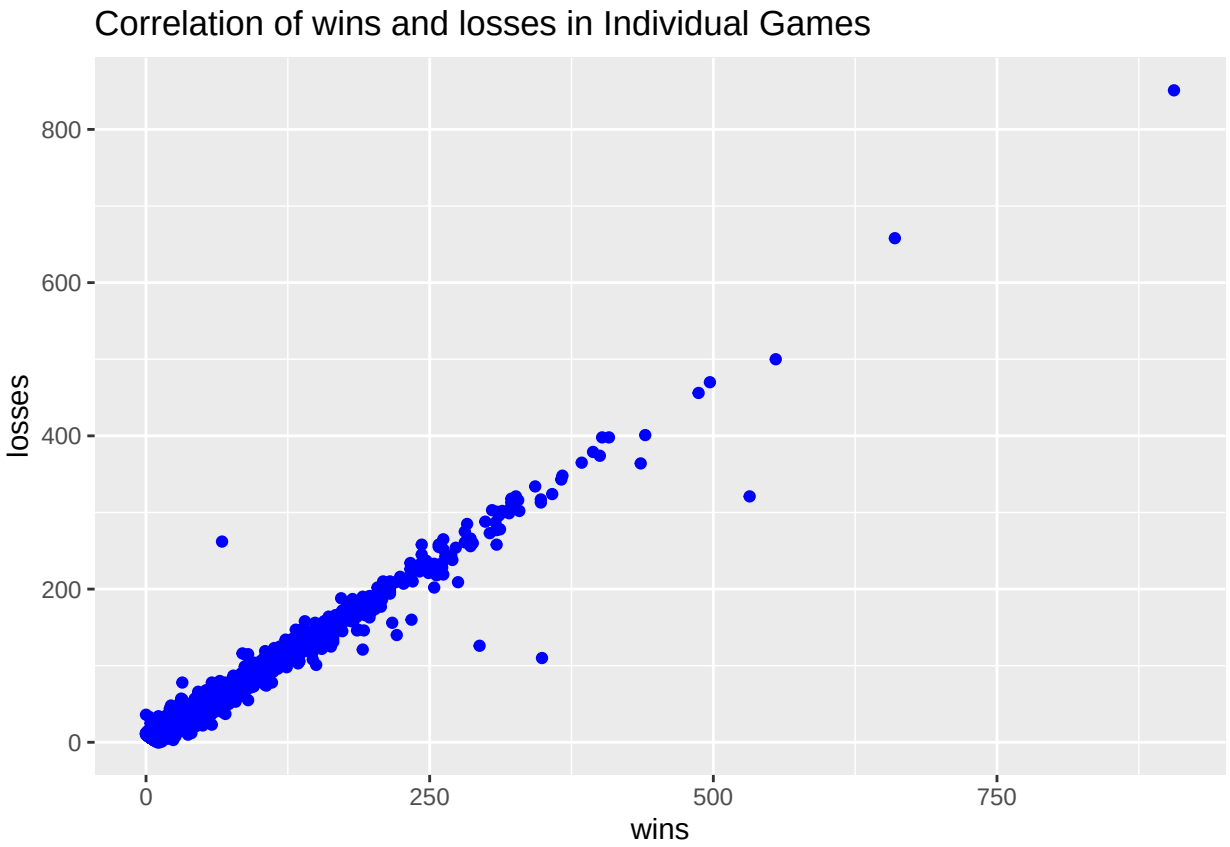
```
ggplot(aoe2, aes(wins, losses, color = game_type)) +
  geom_point() +
  ggtitle("Correlation of Wins and Losses")
```



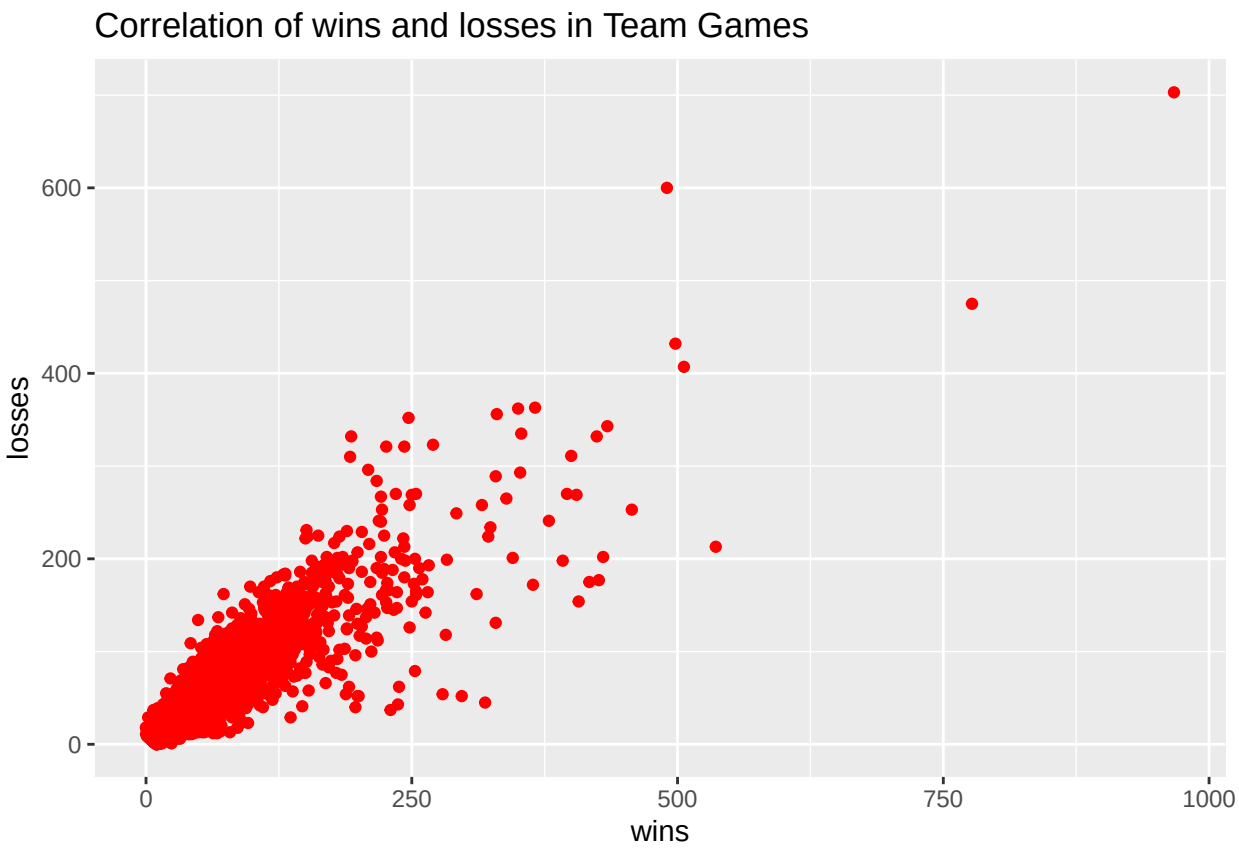
It generally seems like this positive correlation is true for both types of game, but let's verify:

```
individual = aoe2[aoe2$game_type == "1v1 Random Map", ]
team = aoe2[aoe2$game_type == "Team Random Map", ]

ggplot(individual) +
  geom_point(aes(wins, losses), color = "blue") +
  ggtitle("Correlation of wins and losses in Individual Games")
```




```
ggplot(team) +
  geom_point(aes(wins, losses), color = "red") +
  ggtitle("Correlation of wins and losses in Team Games")
```



It seems to be more or less true, but for team games, the grouping from about 50-150 seems to visually make the correlation weaker so lets see if the numbers back that up.

```
cor(individual$wins, individual$losses)
```

```
## [1] 0.9854718
```

```
cor(team$wins, team$losses)
```

```
## [1] 0.8999533
```

They in fact do, by about 8.5%.

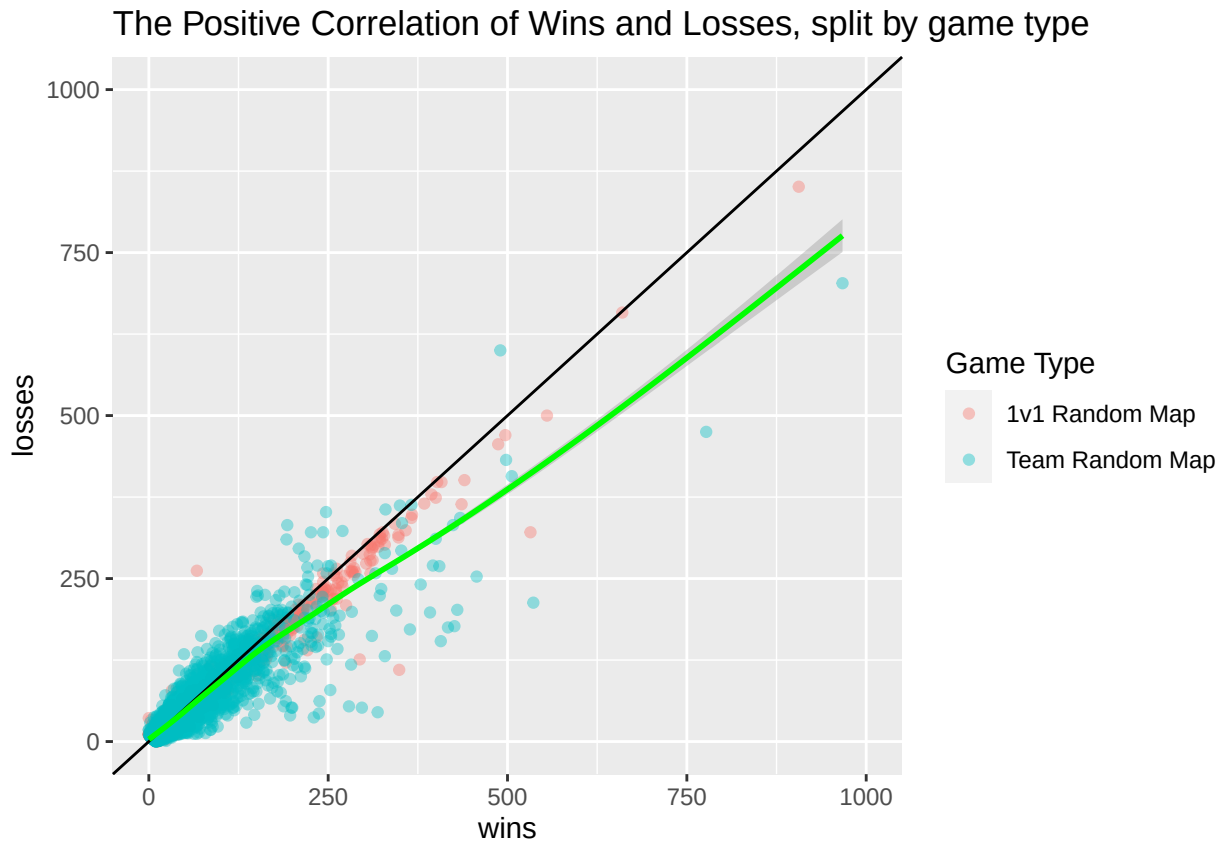
So now knowing this positive correlation is true, let's visualize a trend to see how the relationship between wins and losses changes as we get higher up on the leaderboard. In other words, does percentage of wins grow as wins get higher and if so, by how much?

In the below graphs, the green line is the best fit line for the data while the black line is the 45 degreee reference line, which in this case means the number of wins = number of losses.

```
ggplot(aoe2) +
  geom_point(aes(wins, losses, color = game_type), alpha = 0.4) +
  ggtitle("The Positive Correlation of Wins and Losses, split by game type") +
  xlim(0, 1000) +
  ylim(0, 1000) +
  geom_abline(intercept = 0, slope = 1) +
```

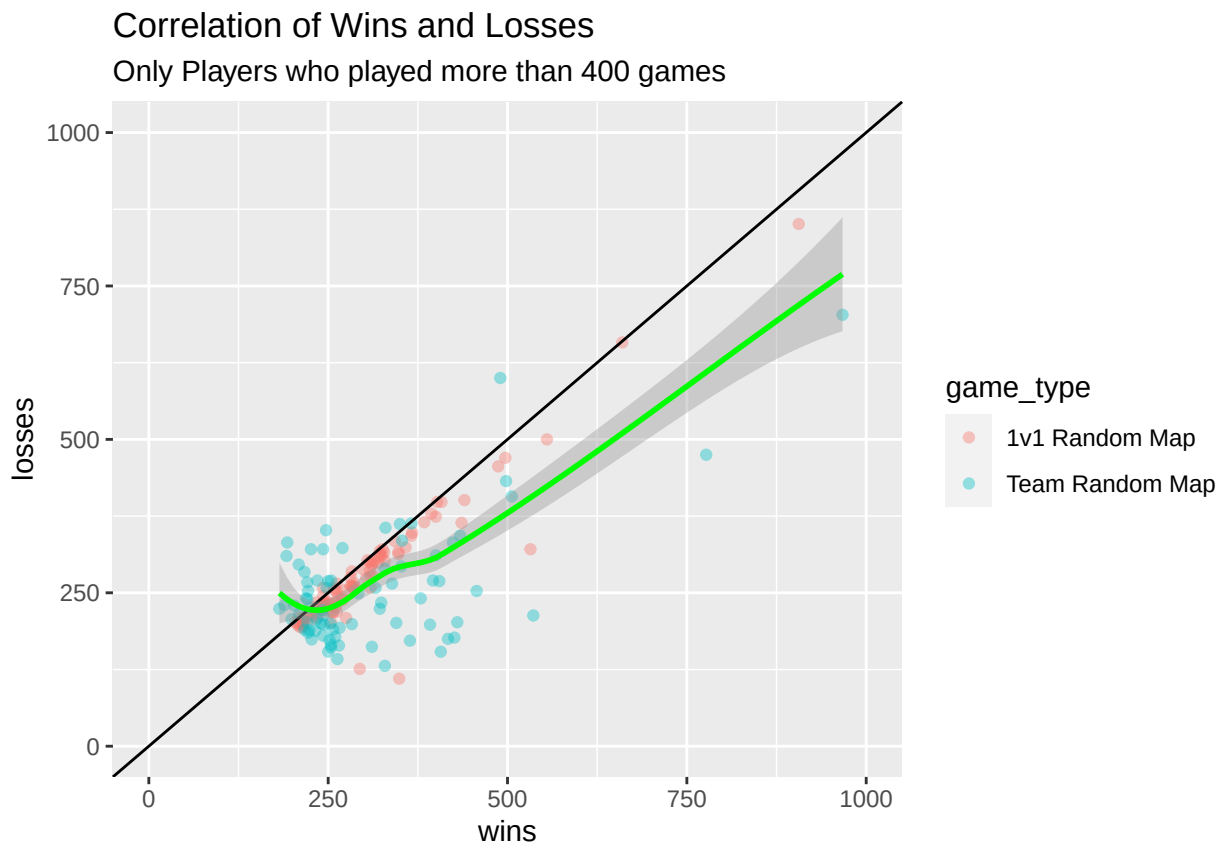
```
  geom_smooth(aes(x = wins, y = losses), color = "green") +
  labs(color = "Game Type")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



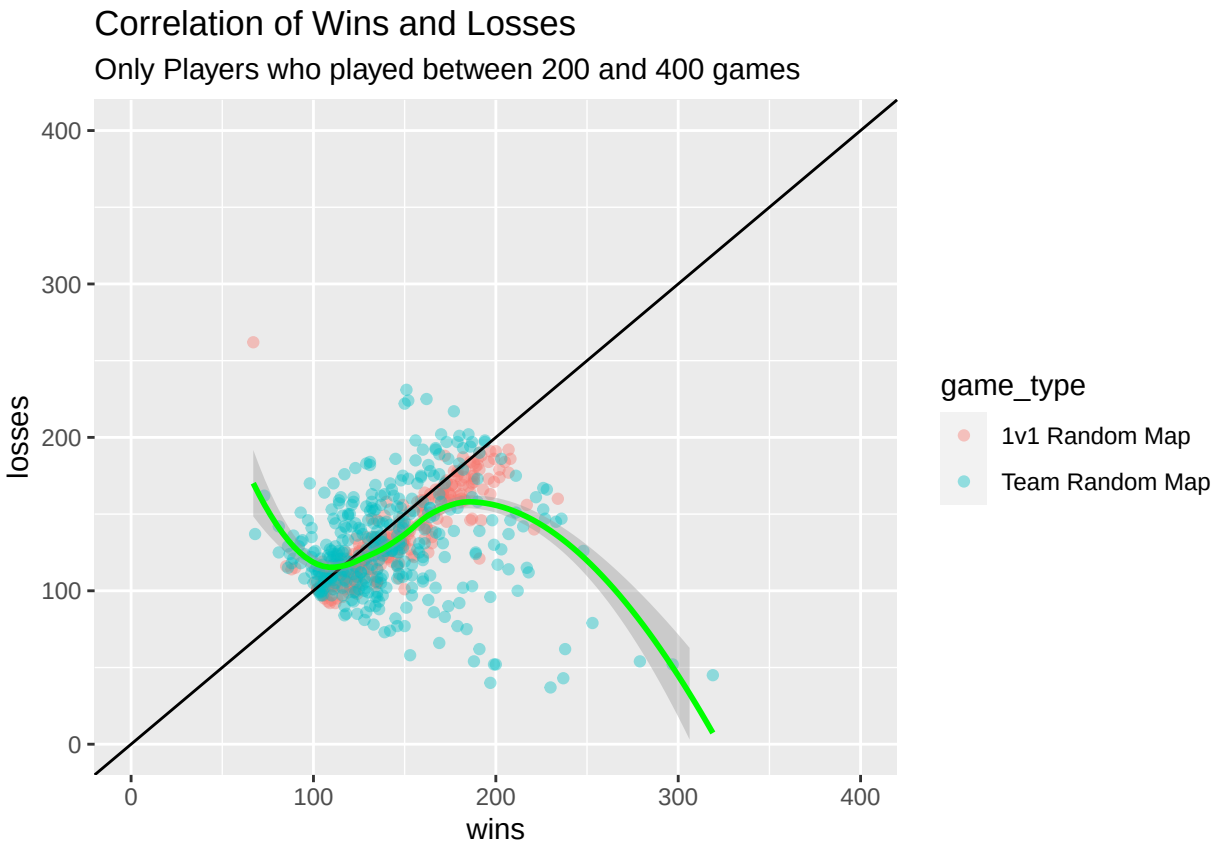

```
ggplot(aoe2[aoe2$games > 400, ]) +
  geom_point(aes(wins, losses, color = game_type), alpha = 0.4) +
  ggtitle("Correlation of Wins and Losses",
          subtitle = "Only Players who played more than 400 games") +
  xlim(0, 1000) +
  ylim(0, 1000) +
  geom_abline(intercept = 0, slope = 1) +
  geom_smooth(aes(x = wins, y = losses), color = "green")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggplot(aoe2[aoe2$games < 400 & aoe2$games > 200, ]) +
  geom_point(aes(wins, losses, color = game_type), alpha = 0.4) +
  ggtitle("Correlation of Wins and Losses",
          subtitle = "Only Players who played between 200 and 400 games") +
  xlim(0, 400) +
  ylim(0, 400) +
  geom_abline(intercept = 0, slope = 1) +
  geom_smooth(aes(x = wins, y = losses), color = "green")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



We see with the first graph that on average, as wins rise, the percentage of wins increases from the 50% average to around the 75% mark for the best players. There are not many players at this top range which arguably makes the data not 100% reliable, and that variance is shown in the 2nd graph.

With the the third graph however, we see a weird trend with players who have played in between the 200-400 range where their win percentage get increasingly larger when they hit that 250 win mark. All of these dots are team random map players, inspiring us to do the cluster analysis for the clusters of the best players, and what defines "the best player," as these players, as they are few, do not bring down the whole graph.

Overall, we proved our hypothesis correct and more, as wins and losses are definetly correlated, and its very positive.