Project Final report

I. Question of Interest

My question of interest was *« Is the average amount of pass of the winning team in a soccer game is different between Ligue 1 (French 1st soccer division) and Bundesliga (Germany 1st soccer division) ? ».*

I choosed this question because, as an European citizen, we are almost all fan of soccer, and among the 5 biggest European league (England, Spain, Germany, France and Italy), we always talk who is the « farmers league » (the worst league). So I wanted to compare something between Germany and France soccer league, as they are seen as not as good as the others. I choosed to see iof there was a difference on the number of pass of the winning team. Having in mind that if a team win with a low amount of pass, they play a less good soccer game than a team with a higher amount of pass.

II. Data Collection Methods

To have a good sample I took 30 individuals for each group, so 60 in total.

To select the soccer games, I used the website https://www.lequipe.fr/, a french website about sport that I know well. You can find every match of either Ligue 1 or Bundesliga (our 2 groups) and have access to games statistics, such as number of pass which is our variable of interest.

To randomly select a game, I used the website https://random.org which provide a randomizer (within a minimum and maximum).

I used games from 2021/2022 and current season (2022/2023). To select each individuals I followed these steps :

- 1st randomizer between 1 and 2 :
    - 1 = 2021/2022 season
    - 2 = 2022/2023 season
- 2nd randomizer : Selecting a random playday (between 1 and 38 for Ligue 1 and between 1 and 34 for Bundesliga in the case of 2021/2022 season, else between 1 and previous playday)
    - 1 = First playday
    - 18 = 18th playday
- 3rd randomizer : Selecting a random match in the playday (between 1 and 10 for Ligue 1 and between 1 and 9 for Bundesliga)
    - 1 = 1st match showed in lequipes' website (first match played on the playday)
    - 6 = 6th match showed in lequipes' website

We have to change range of the randomizer between Ligue 1 and Bundesliga because there is 20 teams in Ligue 1 and 18 teams in Bundesliga. In the case of selecting a game which is a draw, we randomly selected another one (by going again through all the steps).

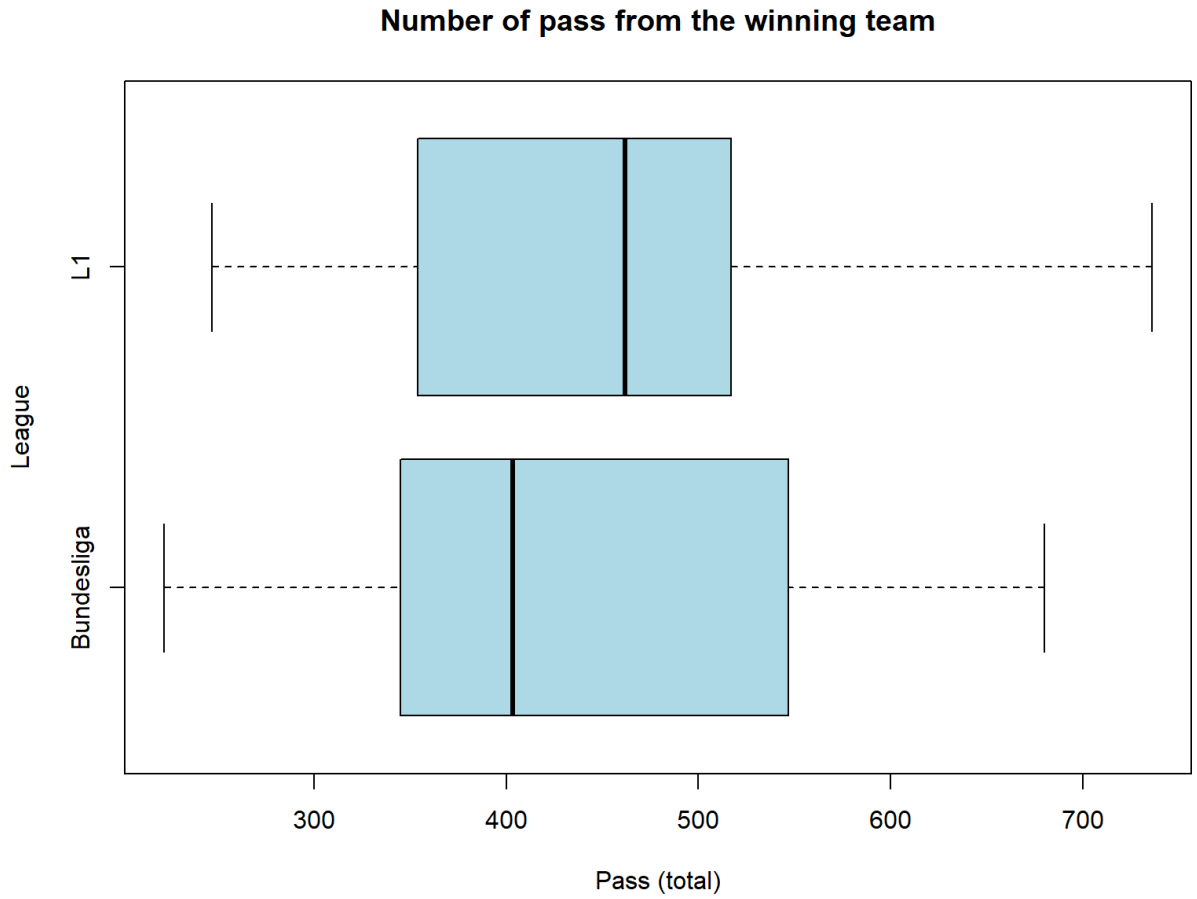
The only problem to happened during this data collection, is that I sampled some games 2 times (because of the randomizer) but I don't think I was supposed to do otherwise, maybe increase the population so I could have less « chance » of sampling 2 times the same game.

III. Your hypotheses

My null hypothesis is that the number of pass from winning teams coming from France 1st soccer) division (Ligue 1 is equal to the number of pass from winning teams coming from Germany 1st division (Bundesliga).

My alternative hypothesis is that there is difference between the number of pass from winning teams coming from France 1st soccer division and the number of pass from winning teams coming from Germany 1st soccer division.

IV. Exploratory analysis



League	Sample size	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
Ligue 1	30	247	357.50	462	451.6	513.25	736	119.4255
Bundesliga	30	222	346.25	403.5	434.1667	545.25	680	136.6143

Using the graphical display and summary statistics, we could procede to use the t-methods to obtain the p-value and construct a confidence interval because the variable of interest is quantitative, the sample was taken randomly and the data seems to be normally distributed (the data are not heavily skewed, only slightly). We have 30 individuals in each which is enough to perform the inference test.

I feel that I will end up failing to reject the null hypothesis because the box and whiskers plot overlap a lot and the mean of both sample is really close. So I don't believe that there is a true difference between the 2 populations.

V. Results of your hypothesis test

With RStudio, I calculated the p-value from the two-sample t-test : 0.6008.
RStudio used a 56.982 degree of freedom.

We got this result from the following line of code : *t.test(pass ~ league, data = Data, mu = 0, alternative = "two.sided", conf.level = 0.95)*

We fail to reject the null hypothesis at with a p-value of 0.6008 at a level of 95% confidence. We have close to no evidence that the difference of number of pass from the winning team between Bundesliga games and Ligue 1 games is different than 0.

VI. Confidence Interval for the difference in population means

With RStudio, I did calculate the 95% confidence interval for the difference in population means, and I ended up with this result :

95% confidence interval (µBundesliga - µLigue1): (-83.77346 pass; 48.90679 pass)

We are 95% confident that the difference of number of pass from the winning team between Bundesliga games and Ligue 1 games is between -83.77346 pass and 48.90679 pass.

VII. Final discussion

Before collecting the data, I thought that there would be a real difference between the 2 populations but it seems I was wrong. I thought Bundesliga would have more pass (from the winning team) because they're supposed to play better, and have a better control of the game (which mean that they would do more pass).

So I was quite surprised (and also a bit disappointed) about the result, which doesn't show much.

As most europeans thinks that France 1st soccer division is the worst league of the 5 big leagues, I also thought that would be the case. With this study I know tend to think that the level between leagues is maybe closer than what we think.