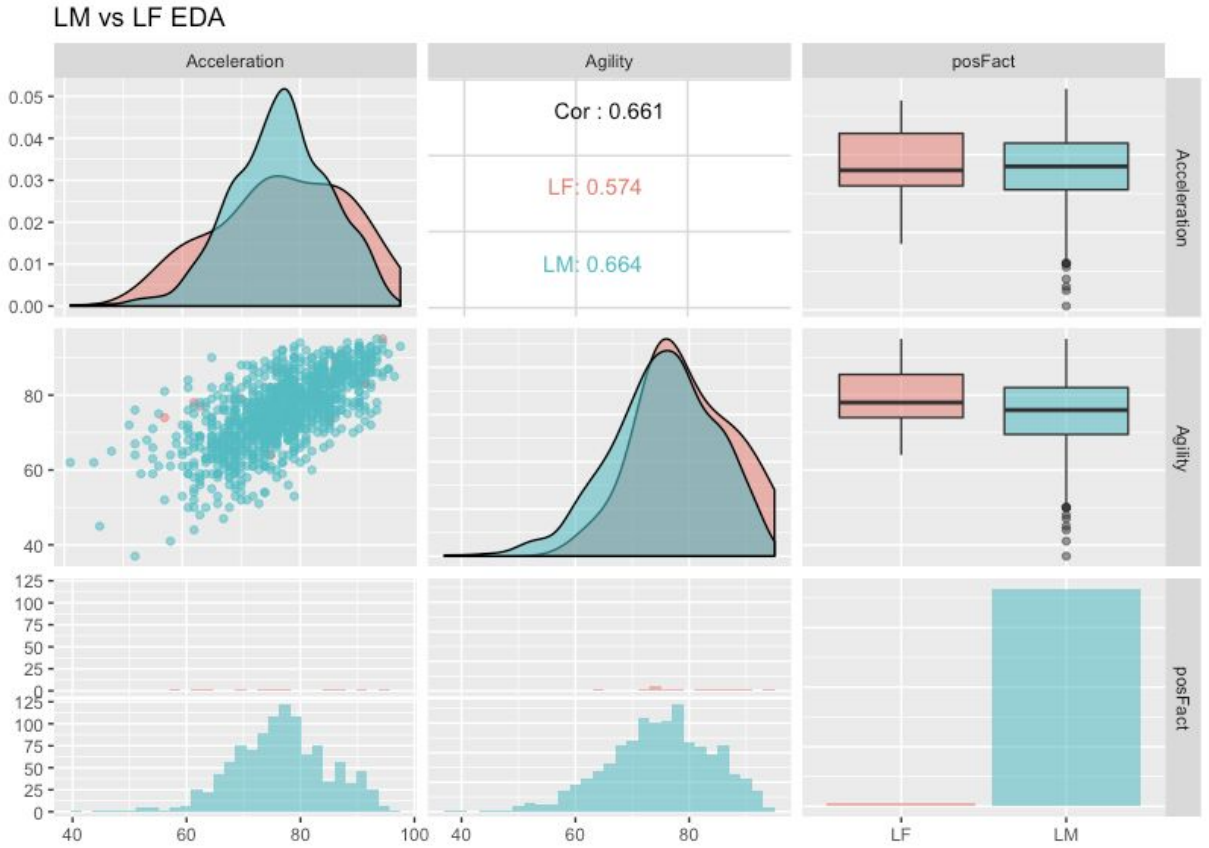
Unit3: For Live Session

DS6306
Garrity

PART 1A

Use Ggally and ggpairs() and the dataset to plot the categorical variable Position (LM and LF), versus the continuous variables Acceleration and Agility. What relationships do you see? Comment on these.

Acceleration and Agility ratings appear to be normally distributed for both positions, although there is a touch of left skew. The LM position has outliers to the low side for both ratings, however, the sample size is large so this is not terribly concerning. Other than the outliers, the variances appear to be similar for both positions and skills. Acceleration and Agility appear to be positively correlated. There is a large discrepancy in sample size between the two positions. A visual inspection suggests that assumptions for a t-test have been met.



PART 1B

Your client would like to formally test if the mean agility rating of left midfielders is different than that of the left forwards. Perform a 6 - step t-test to test for the difference in these means.

STEP 1: $\mathbf{H_0}$: $\mu_{LM} = \mu_{LF}$ | $\mathbf{H_a}$: $\mu_{LM} \neq \mu_{LF}$

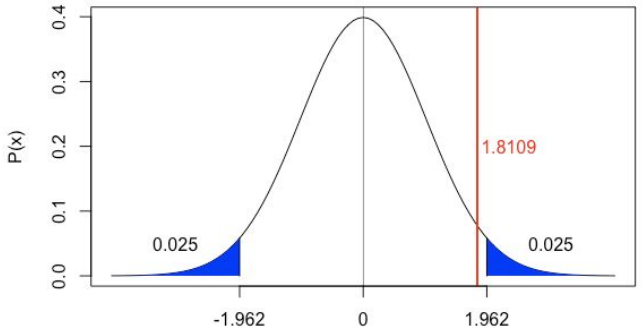
STEP 2: Draw and shade
---------------------------->

STEP 3: $t = 1.81$

STEP 4: $p = 0.070$

STEP 5: Fail to reject H_0

STEP 6: There is insufficient statistical evidence ($p = 0.07$ to suggest that the mean agility rating of left midfielders (LM) is different than that of left forwards (LF). We are 95% confident that the true mean difference in agility rating between LM and LF is within the interval (-0.36, 9.1). Note that the sample size for LF was much smaller than the sample size for LM (15 vs. 1095) and the p-value was fairly close to our cutoff of 0.05. Further investigation into this question should start with collecting additional samples of LF agility ratings.



t-test:

```
t.test(q1_fifa$Agility ~ q1_fifa$posFact, alternative
= "two.sided", paired = FALSE, var.equal = TRUE,
conf.level = 0.95)
```

Test results:

```
t = 1.8109, df = 1108, p-value = 0.07043
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -0.3636412  9.0741435
sample estimates:
mean in group LF mean in group LM
        79.73333         75.37808
```

PART 1C

Assumptions used for the t-test.

For a two-sample t-test we assume:

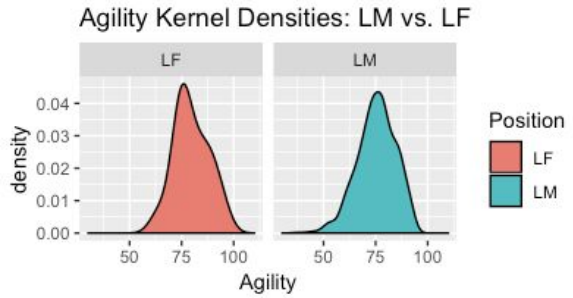
1) **Both samples are normally distributed.** Visual inspection of the three plots on the right demonstrate that normality is a reasonable assumption. There is left skew and outliers in the LF agility distribution, but we are dealing with a large sample size (n >> 30), so we can invoke the CLT to say that this assumption is met.
2) **Samples are independent.** I assume that the samples are all independent although this needs to be verified.
3) **Equal population variances.** We don't know the population variances, so we estimate them from the samples. Visual inspection of the plots on the right indicates that the variances are likely equal, and a f-test confirms this:

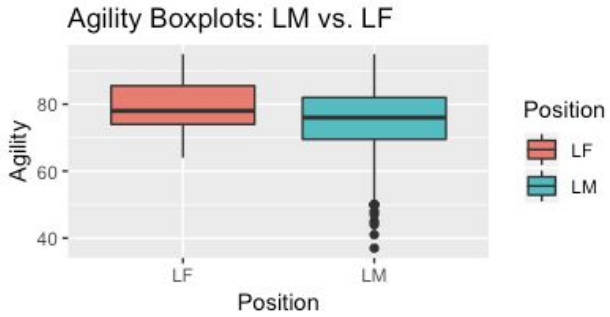
F-test:

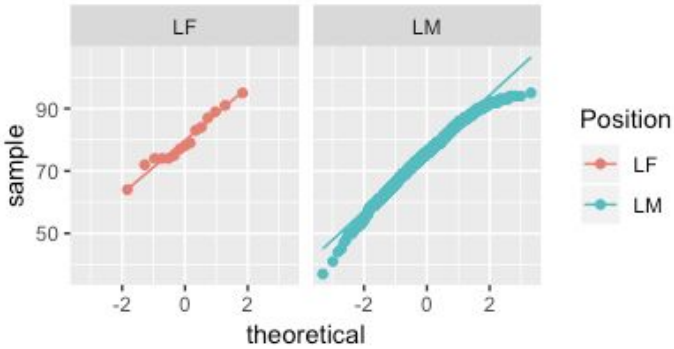
```
var.test(q1_fifa$Agility ~ q1_fifa$posFa, alternative =
"two.sided")
```

Test results:

```
F = 0.80497, num df = 14, denom df = 1094,p-value = 0.6711
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.4287003 2.0079723
sample estimates:
ratio of variances
         0.8049705
```







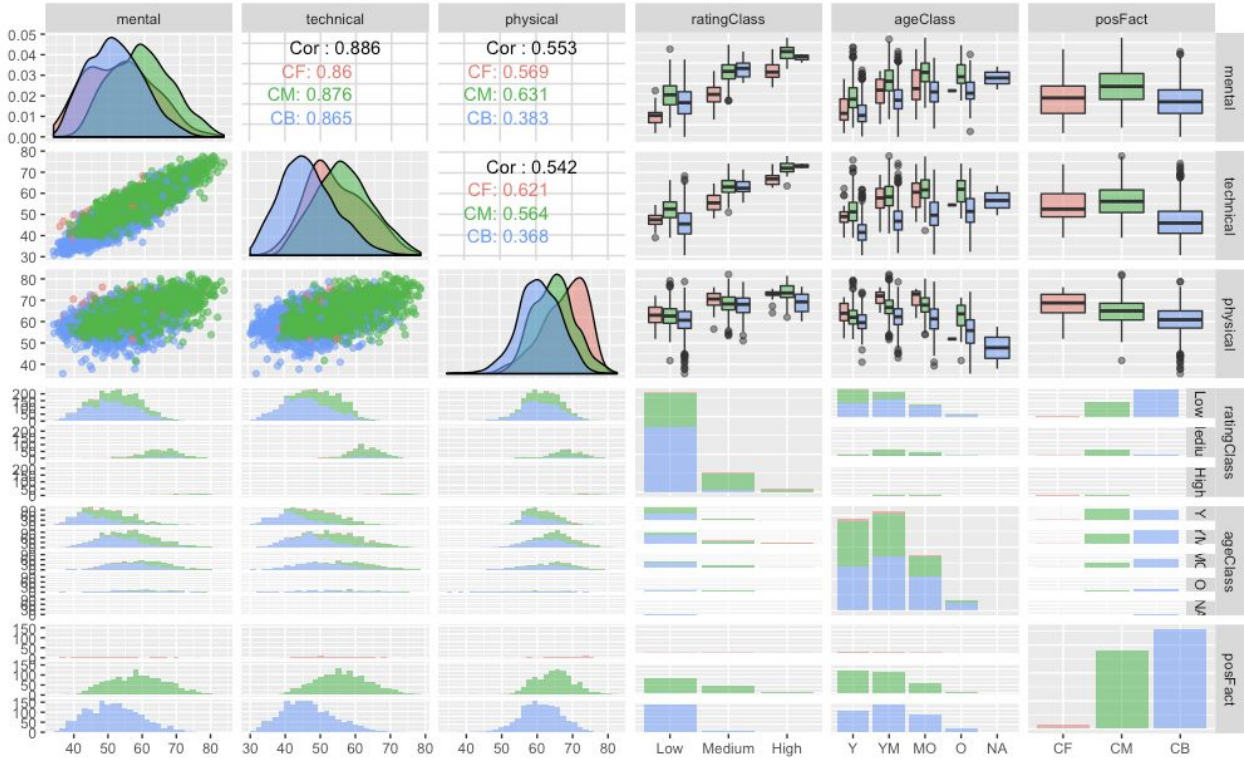
PART 2A

Mental, Technical, and Physical Ratings of CF, CM, and CB playing positions.

The motivating question is how center field positions (back, middle, and forward) compare three main skill areas: mental, physical, and technical. Mental, technical, and physical skill classes are defined here as the mean rating of attributes belonging to each skill class as defined by: https://www.fifauteam.com/fifa-19-attributes-guide/.

Mean ratings for each skill category were compared across the CF, CM, and CB positions as well as age class (young, mid-young, mid-old, old) and rating class (low, medium, and high).

CM positions tend to have the highest mental and technical ratings. CF positions appear to have the highest level of physical skill. CB positions tend to be rated lowest across all skill classes.

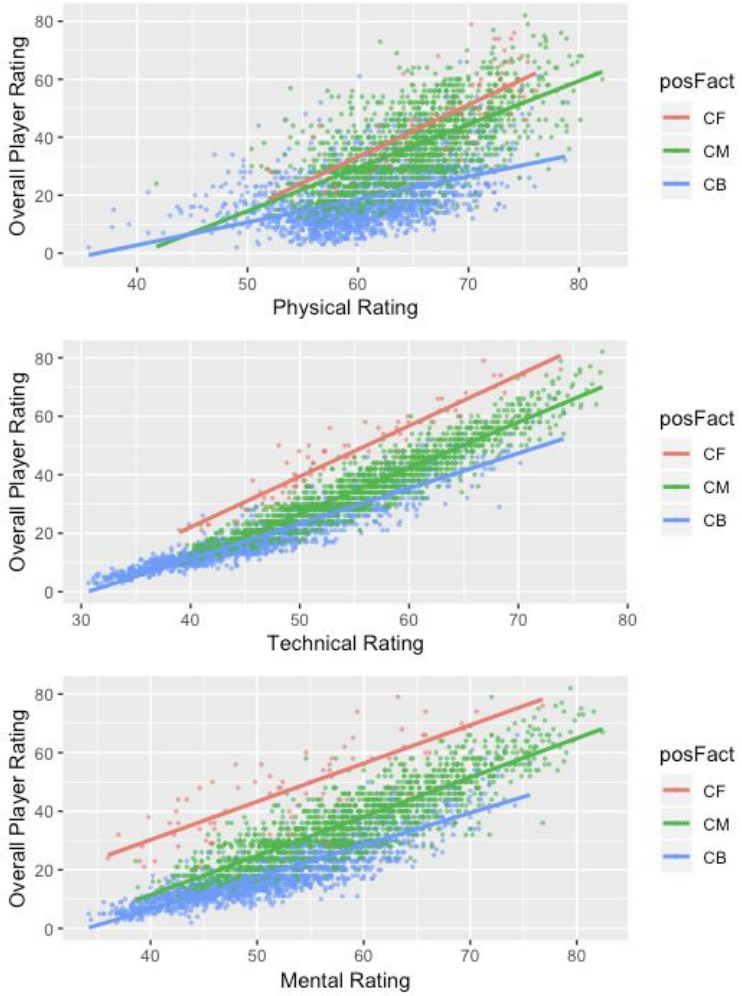


PART 2B

Impact of Mental, Technical, and Physical on Overall Player Rating.

I wanted to understand how each of the skill classes (mental, technical, and physical) was related to a player's overall rating for CB, CM, and CB positions.

- Higher ratings in any of the three skill classes has a positive impact on a player's overall rating.
- The physical skill class has less of an impact on CB overall rating compared to CF and CM positions.
- The relationship between overall player rating and technical or mental ratings was similar (similar slopes). There is a consistent offset (bias) in the fitted linear models among the positions, with CF tending to have the highest overall ratings and CB tending to have the lowest overall ratings for any given technical or mental rating.



Takeaways & Questions

EDA can be a lot of fun.

“Sapply” seems like a very powerful tool that will prevent having to write for loops. I wouldn’t mind spending some time becoming more comfortable with how and where to use it.

I'm confused by order vs. arrange when it comes to plotting. When we created the data frame with "arrange" the printed output was correctly arranged, but the plot still showed them out of order. It wasn't until we invoked "order" that things worked as expected.

Specifically why didn’t this work,

```
fifa %>% filter(!is.na(BallControl)) %>% group_by(Position) %>% summarize(meanBC = mean(BallControl),
count = n()) %>% arrange(meanBC) %>% ggplot(aes(x = Position, y = meanBC)) + geom_col()
```

...but, this did?

```
fifa_BC = fifa %>% filter(!is.na(BallControl)) %>% group_by(Position) %>% summarize(meanBC =
mean(BallControl), count = n()) %>% arrange(meanBC) %>% print(n=28)

fifa_BC$Position = factor(fifa_BC$Position, level =fifa_BC$Position[order(fifa_BC$meanBC)])

fifa_BC %>% ggplot(aes(x = Position, y = meanBC)) + geom_col()
```