Log- Baltimore Ravens Jersey Popularity vs. Twitter Followers

By Molly Ann Williams

https://github.com/MollyAnnWilliams

10/8/22

- Copied "table" from website (ESPN Baltimore Ravens Roster)

- Pasted into Excel

- Formatted text (changed all to Arial, size 12)

- Added column for length (LEN) of Name text string

- Added column for separating text from name (column 1, Name, had the Name combined with the Number)

- Ran code for LEN in Length column, pulled down to paste formula in the rest of the column

- To make it easier to separate the Name from Number, to be able to pull down and paste the MID formula to the cells below, I added a filter to be able to Sort by Length of Name

- Pulled formulas down for rest of column

- Hid the Length column. Also Hid Age, HT, WT, EXP, and COLLEGE

- Encountered a problem: Because I used MID for the Number column, I could not remove that number from the Name column(format- Firstname Lastname#). When I attempted to it removed the output from the Number column. I researched different methods to try to "Lock" the values into the column and have them not change when I change the source cell, but ran into a roadblock. Asked around for help.

- DISCOVERED using LEFT(A3,LEN(A3)-2) pulled just the text portion of the Name column! Will have to do each individually which is a pain but I can have a separate column for just names now.

- Finished inputting the player numbers for Def and ST. Renamed Just\_Name to Name

- Had trouble with the sort by Number. Three of the numbers I had an issue with the MID function and had to input them manually. This caused them to show up at the beginning of the sorted list (A-Z). Also there were several duplicates (#47, #59). I don't expect to find their jerseys for sale in the store though.

- Decided to switch first name with last name and add a comma in the Names column, to create a new column Last\_First. Combined the RIGHT, LEN, SEARCH and LEFT functions //=RIGHT(B5,LEN(B5)-SEARCH(" ",B5))&", "&LEFT(B5,SEARCH(" ",B5)-1)// to do so. Filled in the column. Renamed Names column to First\_Last, renamed Last\_First to Names

- Decided to create a new sheet with only the pertinent columns shown; leaving behind AGE, HT, WT, EXP, and COLLEGE. Named Roster\_Updated

- I ran into a coding issue trying to do the Roster\_Updated sheet. Will just leave columns hidden for the purposes of this excercise.

- Realizing I shouldn't have deleted the Offense sheet. Just to keep all raw data preserved.

\*

- Decided to combine the Roster spreadsheet tabs (Offense, Defense, ST) into one sheet (will combine this spreadsheet with the second spreadsheet which will have names, jersey numbers, positions, and Twitter followers)

- Combined all three sheets (Offense, Defense, and ST) into one Sheet. Cleaned and formatted, hid several columns. Formatted Cells in POS column, one color for each position (conditional formatting). Renamed Offense sheet to Roster.

- Added a new column for Side (Offense, Defense, ST). Could sort by that/have conditional formatting when the time comes.

- Will combine this spreadsheet with the jersey and Twitter spreadsheet

- Decided to combine the Roster spreadsheet tabs (Offense, Defense, ST) into one sheet (will combine this spreadsheet with the second spreadsheet which will have names, jersey numbers, positions, and Twitter followers).

10/9/22

- Went to jersey website. Wrote down Top Sellers in order by name (Top 20).

OBSERVATION- The usual suspects are in front (Jackson, Tucker etc), then a lot of retired players, and then the newest players.

- In Roster\_Jersey\_Twitter spreadsheet, added columns for Jersey\_Popularity and Twitter\_Followers.

- Inputted jersey information into Roster\_J\_T sheet for the top 20 players.

- Discovered that Rashad Bateman's old jersey (12) is more popular on the website than his current jersey (7).

- Froze panes so that header row and list of names were locked into place.

- Hid all players that were not in the Top 20.

- Found a list of Ravens players' Twitter handles. Visited each profile to see how many followers. Inputted their totals in the Twitter\_Followers column.

- Wanted to count how many Offense and how many Defense in ODST column, applied a COUNTIF but it included all the hidden rows. Researched how to exclude hidden rows on Google but the resolutions I found weren't applicable to my spreadsheet concerns. Manually counted: OFFENSE: 12, DEFENSE: 6, ST: 2

OBSERVATIONS: Morgan Moses does not have a Twitter account that I could find. Predictably, Lamar Jackson's jerseys were the most popular and he had the most Twitter followers as well. Surprisingly Kenyan Drake was number 20 on the jersey popularity list but 4th on the Twitter followers list. Meanwhile, Mark Andrews was number 2 on the jersey popularity list but 8th on the Twitter followers list. Simliar is Kyle Hamilton, who is number 6 on the jersey popularity list but 13th in Twitter followers.

- Copied and pasted Roster\_J\_T into a new sheet, titled Roster\_J\_T\_Trim. Deleted all unnecessary rows.

- Created a line chart showing relationship between JP and TF. Realized the chart is going the wrong way (highest followers on the left side). Will rename "Jersey Popularity vs. Twitter Followers V.1" and recreate in the correct direction. When I resorted the Twitter\_Followers column it automatically updated the chart.

- OBSERVATION: Lamar Jackson's Twitter following is so huge (1M) that his dot is way above the others'. This makes it very hard to read the numbers for the other players. They are all in pretty much the same spot on the chart. How to fix this? Maybe I could do a bar chart instead. Maybe if I remove Lamar and Morgan Moses the line chart will be easier to read. Not sure if I'm exactly allowed to do that but I could just make a note explaining why I did. I'm going to hide those two rows and see if that works. Otherwise I'll have to create a new sheet/table with just the 18 players I want to include.

- I changed the chart to something a little easier to read, and left out Jackson and Moses. However there are no labels for the players' names. I need to find a way to make a chart that includes the players' names, and is a little clearer in the lower numbers.

- I tried making a PivotChart but it is still confusing. I may be going about this all wrong. Tried a bar chart for just the Jersey popularity but it didn't include all the players. I have a problem. I might need to consult an online community.

- I tried drawing out what I wanted my chart to look like, on a piece of paper. I think if I could make the y axis the Twitter followers numbers and then put each player down on the x axis, I could do a stacked bar with... wait, that wouldn't work either. Basically I'm showing thousands of followers on one side, with numbers 1-18 on the other. There's no way I'll be able to show a visible correlation on that scale. Maybe I could make the followers numbers into a decimal (10.8 thousand instead of 10,800). I converted the numbers into decimals and created a new chart. It's a little better, however...

- I just realized, the Jersey rank column is from 1-20 (18 now). It represents a ranking, meaning #1 is "highest", but as a number, it's lowest. So it would show a reverse of what I'm trying to show, higher Twitter numbers, lower number (higher ranking). How to solve this? I could create a new column, Rank in Reverse? And assign each rank # with its opposite number (18 = 1, 17 = 2 etc). Then I could use only that column combined with the twitter column and name column. HOWEVER even when I do that I'm going to need to do a chart of just the rankings, like a close up of that bottom area. I could put them on top of each other. There must be a better way to solve this problem but I'm not thinking of it. I think it's a bit past me. At least Excel is not making it easy for me. Maybe Tableau or R will.

- I created a new chart showing Jersey Popularity: Most to Least Popular, using the inverse ranking column. It shows Mark Andrews as the most popular and Kenyan Drake the least. I made a line chart and then a bar chart, the bar chart matches the combo chart with Twitter followers. Except it's actually the opposite since I made it with the reverse rankings. Not gonna be perfect.

OBSERVATION: I'm trying to change the color theme to purple, for the Ravens, but I'm having trouble doing that. I will have to play with that for my final visualizations in Tableau or R. Right now it's orange and blue which would be fine if I were writing about the Broncos or Morgan State. Also, I need to remember, if I want to change anything about the data on the Roster\_J\_T\_Trim sheet, I need to copy and Paste Values into a new Sheet. Otherwise anything I do to it will adjust the charts.

- I just realized I will not be able to upload my data/make edits on BigQuery because I only have a Sandbox account. I think I'll have to pay for BQ in order to upload my own data. I'm gonna find out how much a month will cost.

- I researched pricing for BQ and it's way too much, even for a month. I realized maybe I hadn't tried the correct way to upload data, so I tried again. I was able to create a new project and upload a csv file of the Ravens\_Roster. Then I uploaded the jersey/twitter csv file.

10/10/22

- I realized while researching code for SQL Aliases and Joins that my two columns don't have the Name column's name in common. I deleted the csvs from BQ, renamed the name columns so they both say Names, and saved as csvs. Then I tried uploading them into BQ but I can't figure out how to create a database to upload them into. Figured this out.

- I realized I am signed in to BQ under my other Google email, mawbaltimore@gmail.com. Not sure if this will cause problems later.

- Now I am having trouble with the Aliases. I'm seeing the Alias'd columns just fine. However renaming the tables has not worked so far. They're still in long form and then have the Alias tacked on at the end. I've decided to move on to the JOIN steps. No wait- they didn't work on the columns! I wear last time I looked they were Aliased. Nope, they're not staying like that. As soon as I attempt another query and go back to check the dataset, the columns have gone back to their original (long) names.

WAIT

- Maybe I should remove the unnecessary rows from the Roster spreadsheet, as well as Length\_of\_Name? And number? Do I need the number? Before I upload that one to BQ I'll do that edit.

- I resaved the CSV files so that only the values and columns/rows I needed were in there, to avoid confusion. There was a little confusion between the names of the files but I think I sorted it out.

- I am at a standstill with the JOIN clause. I'm not sure what to do. I had thought that the hierarchy of stuff under the Explorer tab was like, Project name, Database name, table name. Which would make my project the ravenslongnumber thing, and the database is the long form name of my spreadsheet? Maybe the database name is the ravenslongnumber? The RR and RJT2 are the tables? Why can't I reference them in my queries? I'm very frustrated and I'm not sure what's wrong, why I can't figure this out, or even figure out a way to ask for help online.

- I've had an idea. Maybe the reason what I'm doing isn't sticking (the Aliases) is because I'm creating a new tab for each query. Maybe I should just create a new query right below the first one? That's probably how I was supposed to do it. It might just reset every time I open a new tab to make a new query. I might just start over, create a new Project. Then, before I upload the CSV files to that project (Create Dataset), I'll just bloody rename the columns and files. Forget Aliases. Would have been nice to show I can do it but I want to finish the project in time and I'd rather show that I tried at least. So I'll just wind up doing the JOIN.

- I started a new Project. Then I created a Dataset, and called it Ravens\_Tables. I uploaded tables, named Ravens\_Roster\_CSV and Roster\_J\_T\_CSV. Better names though not as short as I'd like. Now the trick is to see if I can join those two tables. I've written down the list of columns I want included. Maybe I'll start by doing a \* instead of listing the tables, but I'll have them join at Names.

- I am still having a problem with JOIN, I cannot get it to work at all. I feel like the syntax is right, it just seems confusing cuz when I put a table's name in the FROM clause, it says something about needing a dataset/database or something, but when I go and actually put the Ravens\_Tables there, it says basically the same thing. Even when I let it autofill and it puts the whole dang name starting with the project name, that doesn't work. I even found other websites with tips to look for help.

- I found a fix! I pasted the query into the SQL\_Plan note and saved it on BQ. It did return repeat columns though so I think I'll join RR into RJTC and leave out the columns I already listed. I guess that will work? If not maybe there's a remove duplicates for columns?

- I filtered through several steps, renaming tables and columns along the way, and wound up only with unique columns in a table I renamed table4.

- Uploaded the CSV files into BQ. Hit a snafu- I thought I uploaded the files into my project but they weren't available in the drop down arrow under the name of my project. I'm going to look up how to upload online. Now I'm trying to figure out if I have a dataBASE loaded, and how to upload it if not. On the SQL help page it's telling me to CREATE DATABASE but SQL is not recognizing that. Is there no space?

- Did an Inner Join in BQ to combine the two spreadsheets. I think the Inner Join will just bring all the columns together. I'll save this as a new spreadsheet and Save As the original ones as raw data.

10/11/22

- Opened up RStudio. I'll use my written plan for what visualizations to create. By the end of the day I've have several vizzes and a presentation.

- Installed "tidyverse" package. Attempted to install "ggplot2" but it's part of that package. Attempted to install "util" but it said it couldn't be installed. I wanted to use the "read.csv" package. I'm gonna look to see if it's part of the "tidyverse" package already. It's not, though .txt and .xls files' packages are. Attempted to install "read.csv" but it said it can't be installed in this version of RStudio. They provided a link to a source that might help me find the right package so I went to it. The link is unhelpful.

- I looked to see if I have a link saved for installing "read.csv". In the meantime I uploaded Ravens\_Joined\_Columns to the File library. Attempted to assign "Ravens1" to Ravens\_Joined\_Columns, but I forgot the quotation marks. Added them in, assigned the value "Ravens1" to the RJC table.

- Attemped to read.csv the Ravens1 file- read.csv(Ravens1). It says No such file. Attempted to do the same- read.csv(data/"Ravens1"). It says "non-numeric argument to binary operator". Attempted with- read.csv(data/Ravens1). Got the same error message.

- Looked through my resources for help but found none. I Googled "RStudio read.csv" and found a resource. Attempted- read.csv(file = 'data/Ravens1.csv'). Got the error message In file(file, "rt"): cannot open file 'data/Ravens1.csv': No such file or directory. Attempted the same except with 'Ravens\_Joined\_Columns.csv'. Same error messaage.

- SUCCESS: read.csv("Ravens\_Joined\_Columns.csv").

- Unfortunately, the file has extra columns. I found the SUBSET function and applied it as such- df = subset(Ravens\_Joined\_Columns, select = -c(Names\_1,Off\_Def\_ST\_1, Number\_1) ). Got the error message- object 'Ravens\_Joined\_Columns' not found. Attempted the same but with .csv at the end of the file name. Same error message. Added quotations around filename. Error message- argument "subset" is missing, with no default.

- Maybe I need to assign each column a number and cull those number-named columns out. Tried df = subset(Ravens1, select = -c(4,5,6)). Error message- argument "subset" is missing, with no default.

- Googled that error message. Found one suggestion, (name of dataset)[c(4:6),]. Returned error message- incorrect number of dimensions. Took out final comma. Result was [1] NA NA NA. No idea what this means. I am going to undo it, whatever I did. Tried subset(Ravens\_Joined\_Columns.csv(4:6)) Error message- could not find function "Ravens\_Joined\_Columns.csv". I think it thinks I'm trying to make RJC into a command/function. I'll add a comma and the c thing. Still didn't work, telling me that that object not found. Added df = to before subset. Still saying object (filename) not found, with RJC or Ravens1. There are no quotations used in any of these sample codes online.

- Found another source, tried- keeps <- c("Names","Off\_Def\_ST","Number","Jers\_Pop\_A\_V","Twit\_Foll\_By\_Thous")

> df = Ravens1[keeps] No error message but in the Values section it added df chr [1:5] NA NA NA NA NA and keeps chr [1:5] "Names" "off\_Def\_ST" (the other columns I said I wanted to keep. I'm gonna try running the read.csv code again and see if the unwanted columns have disappeared. No, they have not. It looks exactly the same. I think I'll just start over and remove the columns in Excel, then re-upload. Then I'll rename RJC to Ravens2 and move on to vizzes. Unless I should also do some sorting and filtering? I think I'll just sort within Excel, after removing the columns.

- Saved As RJC...\_v2. Deleted extra columns. Applied a filter. Sorted by Name. A-Z. Uploaded to RStudio. Realized I accidentally edited a different spreadsheet instead of RJC. I edited the Ravens\_Roster\_CSV file. Now it's edited and is the same as Ravens\_Joined\_Columns\_CSV\_v2. So I lost some raw data. I Saved As again and named this copy the filename above. I uploaded it up to RStudio. Assigned that file as Ravens2. Now I will read it and see what it looks like.

- For some reason the read.csv command I've been using that has worked is now not working. It's telling me No such file or directory. I am getting very frustrated.

- I clicked on the File RJCCSV2 in the File section. It had the option "Import File". I went through those steps and now it is imported and I feel a little stupid. Whatever, it seems to have worked. Now to apply some vizzes.

- Attempted ggplot with this code- ggplot(data = Ravens2, mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) + geom\_point() . Received this error message: Error in `fortify()`:! `data` must be a data frame, or other object coercible by `fortify()`, not a character vector. Run `rlang::last\_error()` to see where the error occurred. I will run rlang. Not much help, so I googled the error message. Will try taking away data = and just put the name of the dataframe. I got the same error.

- Tried ggplot(Ravens\_Joined\_Columns\_CSV\_v2, mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) I got a plot... with no data points. Found an article, attempted this: ggplot(Ravens\_Joined\_Columns\_CSV\_V2, mapping = aes(x=Names, y=Jers\_Pop\_A\_V)) + geom\_line() object RJCetc not found. Tried with a lower case v. Here's the code and error message:

ggplot(Ravens\_Joined\_Columns\_CSV\_v2, mapping = aes(x=Names, y=Jers\_Pop\_A\_V)) + geom\_line()

geom\_path: Each group consists of only one observation. Do you need

to adjust the group aesthetic?

I have no idea what this means. I'll try Googling it. I might need to rewatch the video about aes/geoms. Google didn't help me. I'm going to try removing the mapping = part.

ggplot(Ravens\_Joined\_Columns\_CSV\_v2, aes(x=Names, y=Jers\_Pop\_A\_V)) + geom\_line() Same error message.

- Started to watch refresher video, it referenced ggplot2 cheat sheet. I can't find that in my links but I found a site that suggested this as code: ggplot(data, aes(x=distance, y= dep\_delay, size=air\_time)) +

geom\_point() Got the error message: Warning message Using size for a discrete variable is not advised.

I'll remove the size part. Gave me a plot with just one dot in the center. No names listed, no numbers on the side. I screenshotted it and saved it as "FailedPlot1.jpg". On my computer.

- Watched refresher video. "Getting Started with ggplot()". Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = "Names", y = "Jers\_Pop\_A\_V")). Same kind of plot, with one point. Tried to replace point with line, no points now. Found RStudio Community, found a potential answer. Tried removing quotation marks: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) I got a plot!!! Doesn’t look great. I'll need to 45 degree angle the Names at the bottom, and I think I should make a bar chart to see how it looks. Error- stat\_count() can only have an x or y aesthetic. I knew that was going to happen. I'll see how it works if I use only the x aesthetic. As expected, the bars fill up the plot. If I can only control one axis how do I get the amounts into the plot? Maybe I'll try a scatterplot again but add in a size factor. It worked (size = twit foll...). The plot looks horrible though. There's really no logic to any of it. I guess it really doesn't prove my hypothesis completely, but I could make a note about how Twitter popularity doesn't necessarily equate to jersey popularity, but that other factors influence popularity on Twitter, like ease of use by player, or strength of content.

- I'm gonna change the color to purple. Wonder if I can do dark purple. I got it purple but it's more like pink. Also, I'm noticing that the Twit Foll by name size chart only shows three categories of size- 100,200,300. Wonder how I can add to that.

- I found a list of R colors. There's one called darkorchid4, that looks good. Wonder if I have to dl a package for all these or if they're base colors? I'll try it out. It may have changed the Names color? The dots are still pink. Maybe I'll try to adjust the angle of the x axis and come back to the color. There's also a mediumpurple4. Or a purple4. Trying theme(axis.text.x = element\_text(angle = 90, hjust = 1)

Running just that didn't work. Running this ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "darkorchid4", theme(axis.text.x = element\_text(angle = 90, hjust = 1)))) Got me this error message- "Error in `[.data.frame`(df, setdiff(names(df), names(transformed))) :

undefined columns selected

In addition: Warning message:

Ignoring unknown aesthetics: "

- Tried this: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "darkorchid4") + theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))) Got this message: Error in `check\_required\_aesthetics()`:

! geom\_point requires the following missing aesthetics: x and y

Run `rlang::last\_error()` to see where the error occurred.

- Going back to the next video to get ideas. Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4") This put the color variable outside the specifics about each axis. Got a good plot! Still need to angle the names.

- Found the video on angling etc. It's annotate. Tried this code:

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4") + annotate("text", x = 220, y = 3500, label = "Jersey Popularity and Twitter Followers", fontface = "bold", size = 4.5, angle = 25) It gave me a plot where the title is angled and huge, cramming all the data points and x axis names into the lefthand corner. I'm gonna adjust the size, angle, and remove the text and x and y parts.

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4") + annotate("text", label = "Jersey Popularity and Twitter Followers", fontface = "bold") This took away the label completely, but fixed the rest of it back to normal. Still need to figure out how to angle the x axis.

- Adding this to the plot code: theme(axis.title = element\_text( angle = 90, color="red", size=15, face=3)) Took away fontface, changed angle to 45, took away color designation. As I expected, it angled the TITLE and not the x axis labels. So I need to figure out a way to search for that.

- To save time and typing I saved the core ggplot code as plot1.

- From what I"m reading, it's not possible to angle the labels on the x axis unless you're making them perpendicular. I guess I could try that but I swear they did that in a video. Tried adding this to the theme section: theme(axis.text.x = element\_text(angle=90, hjust=1) Tried ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4"), theme(axis.text.x = element\_text(angle=90, hjust=1) + annotate("text", label = "Jersey Popularity and Twitter Followers") + theme(axis.text.x = element\_text(angle=90, hjust=1). Error message unexpected ',' Changed the comma in front of theme to a + as such:

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4") + theme(axis.text.x = element\_text(angle=90, hjust=1) + annotate("text", label = "Jersey Popularity and Twitter Followers")

It's just making a plus sign over and over again.

WAIT

- Maybe I should order the sheet in order of ranking, but still keep the size aspect? It might make a more understandable looking plot. I think I can do that in R... but I may just do it in Excel.

- Found a potential fix online. Will try this: writers\_df[order(Age.As.Writer),] with my own inputs. For some reason I created a tibble:

Ravens\_Joined\_Columns\_CSV\_v2[order("Jers\_Pop\_A\_V"),]

# A tibble: 1 × 5

Names Off\_Def\_ST Number Jers\_Pop\_A\_V Twit\_Foll\_By\_Thous

<chr> <chr> <dbl> <dbl> <dbl>

1 Andrews, Mark Offense 89 19 93

Where are the rest of the rows? I can't make this work. I'm going to start over with a new dataset, order it properly in Excel (maybe Google Sheets), then reupload, Alias, and run the table on it. I named it Ravens\_Joined\_Columns\_CSV\_v3. I downloaded it from Google Sheets.

- Will try ggtitle("Jersey Popularity vs. Twitter Followers") I got this: $title

[1] "Jersey Popularity vs. Twitter Followers"

attr(,"class")

[1] "labels"

No title appeared and I got this weird message.

- Tried this ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4", title = waiver(), subtitle = waiver())

Warning message; Ignoring unknown parameters: title, subtitle. I'm going to try adding them in the aes parentheses.

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, title = waiver(), subtitle = waiver()), color = "darkorchid4")

Same warning message.

- Trying ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3+ geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) + labs(title = 'Title')

Just puts more + signs at the end.

Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) + labs(title = 'Title')

This returned a plot, need to add the color in there but the Title was there. I figured out the correct syntax at least. Next step is to turn the dots purple, add the size option, rename the title, and perpendicularize the Names.

-Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = "Twit\_Foll\_By\_Thous")), color = "darkorchid4" + labs(title = 'Title') Error message said something about an extra comma

- Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = "Twit\_Foll\_By\_Thous")) + color = "darkorchid4" + labs(title = 'Title') Error in "darkorchid4" + labs(title = "Title") :

non-numeric argument to binary operator. Again, no idea what that means.

- Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = "Twit\_Foll\_By\_Thous")) + color = "darkorchid4" + labs(title = "Title") - changing the 'Title' to "Title"

Same error message.

- Tried switching things around: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = "Twit\_Foll\_By\_Thous")) + labs(title = "Title") + color = "darkorchid4") Unexpected ) I think it needs the parentheses before darkorchid

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = "Twit\_Foll\_By\_Thous")) + labs(title = "Title") + color = ("darkorchid4") Error= could not find function "+<-"

- An article I'm reading now has color INSIDE the aes parentheses. Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "darkorchid4")) This created a plot with pink dots. I think I need to do the color inside the aes function as the column it represents, and the color outside the function as the color of the dots. That would be:

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "Twit\_Foll\_By\_Thous") + color = "darkorchid4") Unexpected =

- I'm thinking maybe I don't need those quotation marks around color = twitter... Let's take them out.

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v2) + geom\_point (mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = Twit\_Foll\_By\_Thous) + color = "darkorchid4") Same error message.

- OK I did some thinking. I'm going to try to find sample code/example of ggplot2 code fragments. Hopefully it will become easier to figure out where I should be putting these pieces within the code. Found a cheatsheet.

- Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, color = "darkorchid4", size = Twit\_Foll\_By\_Thous) + labs(title = "Title"); several error messages. I'm going to just start from scratch with the code, make the most basic plot with no color or size, just the coordinates.

- ggplot(data + Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V)

That just prompted another + chain. I'm just going to research that error message/error. Internet says the code isn't close properly. So I guess I'll add another close parenthesis? I got this error message: Error in FUN(left, right) : non-numeric argument to binary operator. Internet says it's getting confused by the +? I'm going try using a comma instead. Now it doesn't like the comma. I think I'll put the + back in but also add .csv to the file.

WAIT

I put a + instead of a = after data. Let's do this again:

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3,csv) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V))

Now it's complaining about the .csv. I'll take that out. Finally got a plot

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V)) That's the base code. I'm gonna name it plotBase

- I'm going to use that Base to adjust color and size. I think the color inside the aes is about mapping, so I will skip that and add the color after the parentheses. I think I'll put size inside the aes code.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous))

- That worked! Now I'll add the color, outside the aes. If that works I'll name it plotBase2.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous) + color = "darkorchid4") Error message: unexpected equal sign . Gonna try adding another )

- Tried: ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous) + color = "darkorchid4")) Same error message. OK I'm taking all that color info out. I'll try adding it inside the aes brackets.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "darkorchid4")) This returned a plot. However the dots are pink again. I'm going to see if there's a different purple I can use. I'll try purple4.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous, color = "purple4"))

This showed the exact same result. The internet suggests I take color out of the aes section, add a comma and then the color, then one parentheses.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4")

THIS WORKED! I'm gonna save it as plotbase2

- Now the plot I have I'm going to save as a jpg, and try to find out a) how to change the x axis title, y axis title, legend title, and to b) perpendicularize the labels along the x axis.

ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jers\_Pop\_A\_V, size = Twit\_Foll\_By\_Thous), color = "darkorchid4") + scale\_x\_discrete(labels = c("Control", "Treatment 1", "Treatment 2")) I'm not going try this. I think I get the idea, but I don't understand the combine function in this setting, and why there are more than one "label". What I'm trying to change is the title of the legend, not the factor names in the legend. I think the way to do this is to assign a different name to that column, and then rerun the code including the updated column name, which will then show up on the top of the legend.

- Twitter Followers <- Ravens\_Joined\_Columns\_CSV\_v3.Twit\_Foll\_By\_Thous

That doesn't look right, and it wasn't: Error: unexpected symbol in "Twitter Followers".

- Tried: colnames(Ravens\_Joined\_Columns\_CSV\_v3) <- c(Names,Off\_Def\_ST,Number,Jersey\_Popularity,Twitter Followers)

I forgot the underscore in Twitter...

- colnames(Ravens\_Joined\_Columns\_CSV\_v3) <- c(Names,Off\_Def\_ST,Number,Jersey\_Popularity,Twitter\_Followers)

Error: object 'Names' not found

No idea. Maybe if I add quotations?

- colnames(Ravens\_Joined\_Columns\_CSV\_v3) <- c("Names","Off\_Def\_ST","Number","Jersey\_Popularity","Twitter\_Followers")

It WORKED! In the dataframe. The plot didn't change. I'll run that plotbase2

- I'm not sure how to run the code again. I tried View(plotbase2) but it just showed up in the dataframe section. Then I just typed in plotbase2. Nothing happened.

- ggplot(data = Ravens\_Joined\_Columns\_CSV\_v3) + geom\_point(mapping = aes(x = Names, y = Jersey\_Popularity, size = Twitter\_Followers), color = "darkorchid4")

This WORKED! It updated the column names. I guess I'll save this as plotbase3.

- OK so I've changed the color of the points. I've updated the labels of the x and y axis. Now I need to add a title and subtitle, and caption, and then make the names readable.

- I learned that to "run" plotbase3 again hit print(plotbase3). That reruns the code and shows the updated plot.

- Titles: plotbase3 + ggtitle("Jersey Popularity vs Twitter Followers")

That worked. It now has a title.

- I don't think I'll do a subtitle, maybe just a caption: Jersey Popularity is ranked from 1-20, with 20 being the most popular. Twitter Followers is scaled by the thousands. Data by Molly Williams.

- plotbase3 + ggtitle("Jersey Popularity vs Twitter Followers"), labs(caption = "Jersey Popularity is ranked from 1-20, with 20 being the most popular. Twitter Followers is scaled by the thousands. Data by Molly Williams.")

Unexpected comma. I will instead use a plus sign.

- plotbase3 + ggtitle("Jersey Popularity vs Twitter Followers") + labs(caption = "Jersey Popularity is ranked from 1-20, with 20 being the most popular. Twitter Followers is scaled by the thousands. Data by Molly Williams.")

- That's rather crowded but it worked. First concern, making the labels readable. More important, I want the order to be from least to most popular jersey. Maybe I'll just attempt that in Google Sheets? I will save this image and say it's the best I could do with my current knowledge.

- I need to find out how to save a jpg as a pdf, I need to go into Excel and try to produce a better chart, one I can use in Tableau. I can't forget about Lamar and Moses though. And I'll need more than one chart. I could make an interactive table on the Tableau, of the Names JersPop and Twit Followers, so people can see things ranked the way they want. Maybe I'll do a pie chart as well, showing how it tallies up for O/D/ST. I should add something to my hypothesis where I guess that Offensive players will be more popular in both categories. I'll do pie charts, one for Jers and one for Twit. Tomorrow is supposed to be only Tableau. Maybe in Excel I could figure out a way to make each plot point into a football, with different sizes by Twitter. Hopefully tonight I can research the things in this paragraph.

- Turning jpg into pdf was insanely easy. I knew how to do it already. It's all messed up, I didn't even notice that the caption was cut off because it's too long. I'll explain that as well. I guess I'll take the info from the caption and stick it somewhere else. Like I'll put a \* by the title and write the caption out below it somehow. How am I going to put in there that I left off Lamar and Moses? In the presentation it shouldn't be so hard, but on the Tableau? Maybe after I put my hypothesis (I think that goes on the Tableau?) I could write something like, "You may notice that #1 and #20 are missing from the table/chart. Lamar Jackson was the #1 most popular in jersey interest as well as Twitter following. His data point of Twitter followers (amounting to $1M +) skewed the charts so badly that you could no longer read the rest of the results. Morgan Moses is #20 in jersey interest, but has no Twitter account, so I did not include him in the equation." That's a long caption, I'll figure something out. I think one slide of the OOPS- I forgot about the R presentation. I'm going look that up really quick. Shit. That's impossible for me right now. I guess I'll just make a ppt presentation on my own, it will be good to show I can use that program, I suppose.

- Started with Excel. Doing conditional formatting, color coded the Off/Def/ST with shades of purple. Probably not good, as some people are not great with contrast, but I like how it looks. I was going try conditional formatting with the jersey or twitter columns, but why? This looks pretty good.

- Made a pivot table and chart of Twitter followers by Side. Turned it purple and ordered it ascending by number of followers.

- Made another pivot table and a pie chart of the same data. Attempted to change the color scheme but when I changed one slice it changed every slice to that color. I looked up how to change the color scheme and did so. Now it's shades of purple.

- I want to try using little footballs as data points, to see how it looks. I hid the unnecessary columns and created a scatter plot.

- I need to write a note somewhere explaining that, since the number scale of jersey popularity was between 1-20 (technically 2-18), and the number scale of Twitter followers included numbers up the hundreds of thousands, it would be very hard to tell the difference between plot points/notice any change between them. Therefore I adjusted the Twitter numbers to being by the thousand (i.e. 345,000 became 345, 59,600 became 59.6 etc) and increased the chart value of the jersey popularity numbers by multiplying them by 100. All numbers listed are therefore adjusted for scale.

- I made a combo chart, with the bars marking the Jersey pop by 100s, and the line marking the twitter followers by thousands. I changed the colors to purple and yellow and titled it. I'll figure out the caption tomorrow. I'll also attempt to make a scatterplot with little footballs tomorrow. I'll also do a pie chart showing Jersey popularity by side. Maybe I'll even find a pie chart that includes both data.

10/12/22

- I'm feeling sick. I can't find the spreadsheet I was working in all last night. With the charts, and the main page of joined columns. I can't find it in Excel, File Explorer, Google Sheets, or the Recycle Bin. If I have to start that all over... It's gone, I've done everything to find it. There's a csv file I was working on last night that I can open, but it only has one sheet in it. The one I made had multiple sheets with charts. I think I'm going to have to start from scratch.

- What I have to redo:

- Reformat table, including conditional formatting and filter/sort, hide unnecessary columns, add column to \* by 10

- pie chart of twitter popularity by side, jersey pop by side. make purple shades

- line/bar chart of jersey pop vs twit pop, make purple and yellow

- make chart of Lamar’s data point, with picture and title/caption

- Created pie chart of Twitter Followers by the Thousands by side. Changed colors to purple shades.

- Deleted unnecessary columns from main table (Names\_1, Off\_Def\_ST\_1, Number\_1)

- Created line/bar chart showing relationship between jersey popularity ranking and twitter followers.

- Created pie chart of Jersey Popularity by Hundreds by side.

- Created fake chart for Lamar, inserted stats and his picture. Saved chart as a screenshot and named it lamar2.jpg.

- I've decided not to do the footballs as data points idea. Maybe if I hadn't had to redo all my work this AM, but I don't have the heart. I'm going to upload this file to Google Drive just in case.

- I'm going to have to do all of this again. Apparently when you have a csv file it can't save multiple sheets, so it will only save the sheet you had open when you hit Save. It also doesn't save charts or images or anything. I went back to the original joined columns sheet and saved it as an xls file so this doesn't happen again. I'm going to have to do all the calculations and charts again. Also I won't have as much raw data because it's been deleted. I had no idea. How was I supposed to know that? And there were no warning messages. At least I saved that Lamar graphic.

- In Ravens\_Joined\_Columns\_XLS\_v4, created a Cleaned\_Filtered\_Sorted sheet to save only the columns I need for the project. From that sheet I created another sheet with just the data needed for the Jersey pie chart.

WAIT

The Jersey pie chart doesn't make any sense. The number assigned to each player refers to their rank out of 20, with 1 being the most popular jersey. Therefore adding up these numbers (which have also been multiplied by 1,000 to make the other charts easier to read) doesn't make sense. I don't think I'm going to do any charts for that data, besides the main chart with the correlation.

WAIT

Maybe my whole method is off. When I look at the chart of the main correlation, I realize it's like I'm trying to compare apples to oranges, between jersey rank and twitter followers. The numbers don't really match up. Maybe it's OK though. I just want to see it visually, the ups and downs and where there are outliers, and where my hypothesis is proven. I think I'm going to be OK. I just have to do a lot of explaining in captions and notes and whatnot.

OBSERVATIONS: Looking at the main correlation chart, I can say that, for the most part, there is no real correlation between jersey popularity and number of twitter followers. If there were, I would expect the yellow line to start near the bottom left corner and rise gradually to the other side. Instead, the player with the lowest rank of jersey popularity (Kenyan Drake #17) has the second highest Twitter following. However, Justin Tucker has the most Twitter followers, and his rank in jersey popularity is in second place. In that case the two values are in similar spaces. The top 4 most popular jerseys belong to players with higher than average Twitter followers, however the most popular jersey that I chose to include in my project belongs to Mark Andrews (#89) who has followers in the median range of those of the other players. Therefore I conclude that my original hypothesis is mostly incorrect.

I have potential explanations for this result. I noticed when ranking the jersey popularity of these players that a majority of the top 20 were rookies or in their second year with the team. This could explain their jersey popularity in the online store: I've noticed that, if a player is doing well at the beginning of their first/second season, people will be interested in purchasing their jerseys, as a sort of investment. "I should buy rookie Kyle Hamilton's jersey because it's going to be worth a lot more soon if he keeps doing well. I'll buy it now before their raise their prices due to increased popularity.

My hypothesis that the jerseys of players from the Offense would be the most popular was true.

- My recommendation for the sales team running the promotion is to promote the top 5 players by Twitter rankings. Since jersey sales can be driven by a player's novelty as well as other reasons, the best indicator of a player's overall popularity can be reflected in the strength of their social media presence. After all, players that fans like and find interesting are usually the ones with the most followers (no data to support this). These top 5 players would be: Justin Tucker (#9), Kenyan Drake (#17), Marlon Humphrey (#44), J.K. Dobbins (#27), and Demarcus Robinson (#10).

- I'm working on the PowerPoint presentation. I want to include a slide with an early table and early chart. Unfortunately I ruined that opportunity by saving my early work as CSV. For the table I think I'll use Roster\_Jersey\_Twitter2\_CSV.csv. Finished the presentation. I think it turned out pretty well.

- I finished the presentation, adding animation and revising my notes. I need to figure out how to print the notes only. It would be nice to know.

10/13/22

- I created a new file in Tableau. I uploaded the excel file (after saving it as an .xslx) Ravens\_Joined\_Columns\_XLS\_v4.xlsx. I assigned the columns and rows in order to make the Twitter Followers grouped by unit. I made a bar chart showing each player, what unit they were in, and their Twitter followers, assigning a purple color scale. Then I moved on to the pie chart for Twitter followers by unit. The chart was easy, however there is no legend. I researched how to add a legend and apparently you need to have one in the source data worksheet. I'm researching how to do that now. OK there is a legend with the pie chart I'm trying to recreate. Why isn't it showing it as an option? Then I had a data connection error. I decided not to do the pie chart anyway, it basically tells the same story as my first sheet.

- I created a Jersey Popularity vs. Twitter Followers sheet, with double bar charts for each player. I can't figure out how to combine the bars so they're stacked.

- I created a new sheet just to show twitter popularity by itself, to make something that looks different. I changed Off/Def/ST to Unit. I will make that adjustment everywhere else as well, including the Powerpoint.

- Created more sheets with charts, including a bubble chart that also will not show a legend. I'll just add a text box at the bottom later. More importantly, when I try to add that bubble chart onto a new Dashboard page, it won't display the actual chart, just the legend and title. I've done tons of searching online to figure out what's going on and I can't figure it out. I may have to screenshot that and add it in as an image.

- I realized my bubble chart is not very good, though I was able to get it on the dashboard/story. The color gradient doesn't mean anything, it's just a different shade of purple for each player. I just like it because it's pretty. I essentially said that on the dashboard.

- I made a new dashboard for the Twitter Followers by Unit/Player. I added an image. I want to add a Twitter quote, something funny or weird, or inspiring, by one of the players. Twitter is making me open an account to see these Tweets.

- I added a Tweet and tweaked the other Dashboards in my Story. I think I finished. The layout is a little weird: if it looks okay in the editor window, it's going to look different in the Presentation view as well as on the website. Not sure what to do about that. It's not a great Story, in fact it's basically another PowerPoint presentation but it's the best I can do with my current knowledge and capabilities.

- Running into a problem. I had assumed due to a very cursory introduction to Kaggle that it was a receptacle for data files, like I could have my own profile that I could upload files to (Excel, Powerpoint etc). That's not what it is. It's coding using Python, as far as I can understand. Where am I supposed to upload my work so that future employers can have access to it?

- I think the website I was thinking of was GitHub. You can upload files there to a free repository. My name is MollyAnnWilliams (duh). I went ahead and uploaded the Powerpoint presentation and the finalized last Excel file. Ravens\_Joined\_Columns\_XLS\_v4. I guess I have to provide a link to my tableau profile/project.

- I found the link for my Tableau profile and put it in my GitHub profile. Once I clean up and finish this Note I will upload that as well.

- I included the websites I used as resources on a dashboard in my Story. I also noted that I create all the spreadsheets etc. used to enact this project.

- Edited Log for brevity and clarity. I didn’t do a very good job on the brevity part. I consider this project to be complete. Thank you for reading.

WEBSITES:

https://www.espn.com/nfl/team/roster/\_/name/bal/baltimore-ravens

https://www.pro-football-reference.com/friv/nfl-player-twitter.htm

https://shop.baltimoreravens.com/baltimore-ravens-jerseys/t-14482411+d-25991993+z-85-2674144931

https://www.pro-football-reference.com/teams/rav/

\*

UPCOMING

- Will go through and neaten up formatting in my original Excel files before uploading my whole case study to Kaggle

- Will make official comments in my Log and other Notes that involved code.

\*