UFC-Fight Historical Data from 1993 to 2019
Exploratory Analysis

Sheridan Payne, spayne@bellarmine.edu

I. INTRODUCTION

I chose a dataset from Kaggle, title "UFC-Fight historical data from 1993 to 2019", created by Rajeev Warrier. This dataset consists of a list of evert UFC fight in the history of its organization. Each row represents a fight and details of the fighters, referees, location, dates, statistics on the fight & fighters techniques, and the descriptions of each fighters (e.g., height and weight). I chose this dataset for three reasons: (1) This dataset was one of the only datasets I found that met the 25 variable requirement, had diversity between variable types (i.e., categorical and continuous data), and was not missing a large amount of data; (2) I don't have much knowledge on professional fighting nor UFC, so I thought I could gain knowledge on professional fighting through analyzing the dataset; (3) I wanted to challenge myself with such a large dataset.
Reference / Link to Dataset: https://www.kaggle.com/rajeevw/ufcdata (data.csv)

II. DATA SET DESCRIPTION

The original dataset contains more than 5000 rows and 140 columns total. Hence, I truncated the original dataset to a smaller dataset, using all of the original rows, but only 25 of the columns. Majority of the columns are averages on specific fighting techniques that are not interesting enough to analyze nor are they clear as to what they represent. Most of the continuous variables I decided to use are separated by each fighter (color coded by red and blue) for each fight. Another interesting detail about the dataset that I like is how diverse the variables' data types are (i.e., there is float, int, boolean, datetime, etc.).

The variables listed in **Table 1** are the variables from the original dataset that are analyzed. The variables represent the following:

- *B_fighter* : name of blue fighter of match
- *R_fighter* : name of red fighter of match
- *Referee* : name of referee of match
- *date* : date of each match (formatted month / day / year)
- *location* : location of each match
- *Winner* : color of winning fighter of match or 'draw' if the match was a draw
- *title_bout* : Boolean value of whether the match was a title match or not
- *weight_class* : weight class of match
- *no_of_rounds* : number of rounds in match
- *B_avg_opp_BODY_landed* : number of significant strikes to the body of the blue fighter made by the red fighter
- *R_avg_opp_BODY_landed* : number of significant strikes to the body of the red fighter made by the blue fighter
- *B_avg_opp_HEAD_landed* : number of significant strikes to the head of the blue fighter made by the red fighter
- *R_avg_opp_HEAD_landed* : number of significant strikes to the head of the red fighter made by the blue fighter
- *B_total_rounds_fought* : average of total rounds fought by the blue fighter
- *R_total_rounds_fought* : average of total rounds fought by the red fighter
- *B_Stance* : stance of the blue fighter
- *R_Stance* : stance of the red fighter
- *B_Height_cms* : height of the blue fighter in centimeters
- *R_Height_cms* : height of the red fighter in centimeters
- *B_Reach_cms* : reach (arm span) of the blue fighter in centimeters
- *R_Reach_cms* : reach (arm span) of the red fighter in centimeters
- *B_Weight_lbs* : weight of the blue fighter in pounds

- *R_Weight_lbs* : weight of the red fighter in pounds
- *B_age* : age of the blue fighter
- *R_age* : age of the red fighter

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
R_fighter	nominal / object	0%
B_fighter	nominal / object	0%
Referee	nominal / object	0.427112%
date	ordinal / datetime64[ns]	0%
location	nominal / object	0%
Winner	nominal / object	0%
title_bout	nominal / bool	0%
weight_class	nominal / object	0%
no_of_rounds	interval / int64	0%
B_avg_opp_BODY_landed	ratio / float64	23.491179%
R_avg_opp_BODY_landed	ratio / float64	12.070566%
B_avg_opp_HEAD_landed	ratio / float64	23.491179%
R_avg_opp_HEAD_landed	ratio / float64	12.070566%
B_total_rounds_fought	interval / float64	0%
R_total_rounds_fought	interval / float64	0%
B_Stance	nominal / object	2.952646%
R_Stance	nominal / object	2.488394%
R_Height_cms	ratio / float64	0.074280%
B_Height_cms	ratio / float64	0.148561%
R_Reach_cms	ratio / float64	5.868152%
B_Reach_cms	ratio / float64	12.367688%
B_Weight_lbs	ratio / float64	0.111421%
R_Weight_lbs	ratio / float64	0.055710%
B_age	interval / float64	3.194058%
R_age	interval / float64	1.188487%

III. Data Set Summary Statistics

In **Table 2**, a summary of the statistics of the continuous variables is given. The table includes the count, average (mean), standard deviation (Std), minimum, maximum, and the 25th, 50th, and 75th quartiles.

Table 2: Summary Statistics for UFC_fighter_data

Variable Name	*Count*	*Mean*	*Std*	*Min*	*25th*	*50th*	*75th*	*Max*
no_of_rounds	*5144*	*3.119362*	*0.631457*	*1*	*3*	*3*	*3*	*5*
B_avg_opp_BODY_landed	*3879*	*5.639172*	*4.747421*	*0*	*2.333333*	*4.6*	*7.714286*	*48*
R_avg_opp_BODY_landed	*4494*	*5.498694*	*4.242793*	*0*	*2.5*	*4.8*	*7.5*	*41*
B_avg_opp_HEAD_landed	*3879*	*17.36783*	*12.70291*	*0*	*8.857143*	*15*	*23*	*126*
R_avg_opp_HEAD_landed	*4494*	*16.85842*	*11.67093*	*0*	*9*	*14.83333*	*22.47765*	*132*
B_total_rounds_fought	*5144*	*8.920879*	*11.26934*	*0*	*1*	*5*	*13*	*75*
R_total_rounds_fought	*5144*	*12.85342*	*13.36935*	*0*	*3*	*9*	*19*	*80*
R_Height_cms	*5140*	*179.2741*	*8.638978*	*152.4*	*172.72*	*180.34*	*185.42*	*210.82*
B_Height_cms	*5136*	*179.2386*	*8.515039*	*152.4*	*172.72*	*180.34*	*185.42*	*210.82*
R_Reach_cms	*4828*	*183.6644*	*10.30437*	*152.4*	*177.8*	*182.88*	*190.5*	*213.36*

B_Reach_cms	*4478*	*183.2861*	*10.14918*	*152.4*	*177.8*	*182.88*	*190.5*	*213.36*
B_Weight_lbs	*5138*	*172.1104*	*36.84702*	*115*	*145*	*170*	*185*	*770*
R_Weight_lbs	*5141*	*172.0759*	*35.16407*	*115*	*145*	*170*	*185*	*345*
B_age	*4972*	*29.17196*	*4.078538*	*18*	*26*	*29*	*32*	*51*
R_age	*5080*	*29.44232*	*4.141927*	*19*	*26*	*29*	*32*	*47*

In **Tables 3A-3K**, each table displays the categories of each categorical variable, its frequencies for each category, and the proportion of each category with respect to the other categories. Due to the enormous number of categories in the dataset for some of the categorical variables, **Tables 3A-3E** are truncated by displaying the sample of the table. The full tables can be seen in the original python code.

Table 3A: Proportions for Blue Fighters (*B_fighter*)

Blue Fighter	*Frequency*	*Proportion (%)*
Mackens Semerzier	*1*	*0.01944*
Casey Kenney	*1*	*0.01944*
Juan Adams	*2*	*0.03888*
Samy Schiavo	*2*	*0.03888*
Zak Ottow	*7*	*0.136081*
Brian Ortega	*6*	*0.116641*
Alexander Hernandez	*1*	*0.01944*
Tim Lajcik	*2*	*0.03888*
Brad Tavares	*7*	*0.136081*
Josh Shockley	*2*	*0.03888*

Table 3B: Proportions for Red Fighters (*R_fighter*)

Red Fighter	*Frequency*	*Proportion (%)*
Phillip Miller	*2*	*0.03888*
Leandro Silva	*3*	*0.05832*
Kyle Noke	*8*	*0.155521*
Tyron Woodley	*8*	*0.155521*
Tim Means	*9*	*0.174961*
Jesse Ronson	*1*	*0.01944*
Bobby Green	*5*	*0.097201*
Zak Ottow	*1*	*0.01944*
Paulo Thiago	*10*	*0.194401*
Rafael Dos Anjos	*14*	*0.272162*

Table 3C: Proportions for Referees (*Referee*)

Referee	*Frequency*	*Proportion (%)*

Taimak Guarriello	*1*	*0.019527*
Bobby Rehman	*7*	*0.136692*
Marcos Rosales	*6*	*0.117165*
Chris Tognoni	*89*	*1.737942*
Cecil Peoples	*3*	*0.058582*
Will Fisher	*1*	*0.019527*
Mike Bell	*2*	*0.039055*
Jim Perdios	*5*	*0.097637*
Nick Gamst	*3*	*0.058582*
Elvis Bello	*1*	*0.019527*

Table 3D: Proportions for Dates (*date*)

Date	*Frequency*	*Proportion (%)*
(2003, 6)	*8*	*0.155521*
(2009, 4)	*23*	*0.447123*
(2011, 10)	*32*	*0.622084*
(2016, 5)	*37*	*0.719285*
(2017, 10)	*34*	*0.660964*
(2007, 6)	*26*	*0.505443*
(2005, 2)	*9*	*0.174961*
(2010, 6)	*21*	*0.408243*
(2007, 4)	*28*	*0.544323*
(2006, 9)	*9*	*0.174961*

Table 3E: Proportions for Location (*location*)

Location	*Frequency*	*Proportion (%)*
Krakow, Poland	*12*	*0.233281*
Prague, Czech Republic	*13*	*0.252722*
Liverpool, England, United Kingdom	*11*	*0.213841*
Seoul, South Korea	*11*	*0.213841*
Oberhausen, North Rhine-Westphalia, Germany	*10*	*0.194401*
Norfolk, Virginia, USA	*13*	*0.252722*
Orlando, Florida, USA	*38*	*0.738725*
Jaragua do Sul, Santa Catarina, Brazil	*25*	*0.486003*
Beijing, China	*12*	*0.233281*
Las Vegas, Nevada, USA	*1216*	*23.63919*

Table 3F: Proportions for Winning Fighter (*Winner*)

Winner	*Frequency*	*Proportion*

		(%)
Red	3470	67.45723
Blue	1591	30.92924
Draw	83	1.61353

Table 3G: Proportions for Title Fight (*title_bout*)

Title Fight	Frequency	Proportion (%)
FALSE	4809	93.48756
TRUE	335	6.512442

Table 3H: Proportions for Weight Class (*weight_class*)

Weight Class	Frequency	Proportion (%)
Lightweight	989	19.22628
Welterweight	969	18.83748
Middleweight	725	14.09409
Heavyweight	507	9.856143
Light Heavyweight	502	9.758942
Featherweight	442	8.592535
Bantamweight	379	7.367807
Flyweight	187	3.635303
Women's Strawweight	143	2.779938
Women's Bantamweight	111	2.157854
Open Weight	92	1.788491
Women's Flyweight	50	0.972006
Catch Weight	38	0.738725
Women's Featherweight	10	0.194401

Table 3I: Proportions for Number of Rounds (*no_of_rounds*)

Number of Rounds	Frequency	Proportion (%)
3	4523	87.92768
5	423	8.223173
2	98	1.905132
1	78	1.51633
4	22	0.427683

Table 3J: Proportions for Stance of Blue Fighter (*B_Stance*)

Blue's Stance	Frequency	Proportion (%)
Orthodox	3829	76.81043

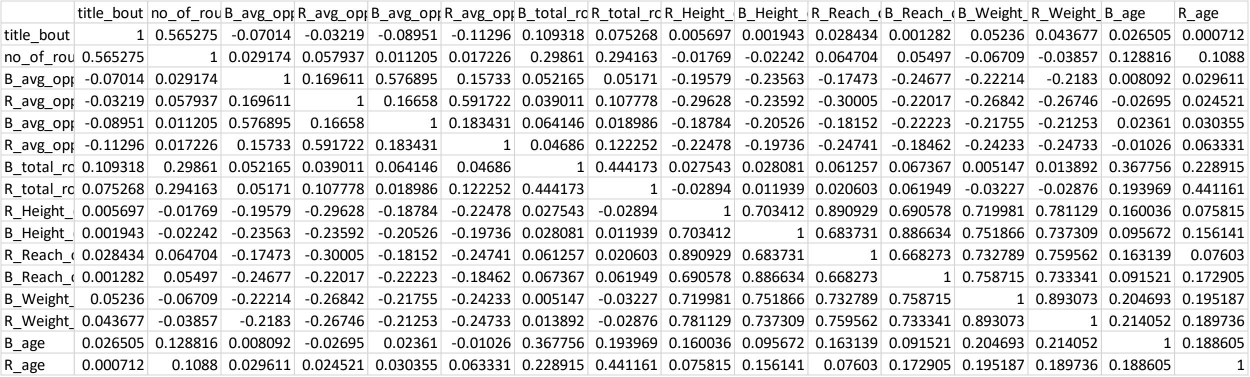
Southpaw	*975*	*19.55868*
Switch	*168*	*3.37011*
Open Stance	*9*	*0.180542*
Sideways	*4*	*0.080241*

Table 3K: Proportions for Stance of Red Fighters (*R_Stance*)

Red's Stance	*Frequency*	*Proportion (%)*
Orthodox	*3807*	*75.98802*
Southpaw	*1036*	*20.67864*
Switch	*150*	*2.994012*
Open Stance	*15*	*0.299401*
Sideways	*2*	*0.03992*

Using the continuous variables, a correlation matrix is created shown in **Table 4**. Although the table is small and difficult to read for the whole table to fit on the page, the values that are positive imply there is a positive correlation between the two variables. Similarly, a negative value implies there is a negative correlation between the two variables. Lastly, a value closer to 0 has a weak correlation between the two variables, while a value closer to 1 has a strong correlation between the two variables. A visual representation of **Table 4** is shown in **Figure 1**.

Table 4: Correlation Table/Tables



	title_bout	no_of_rou	B_avg_opp	R_avg_opp	B_avg_opp	R_avg_opp	B_total_ro	R_total_ro	R_Height_	B_Height_	R_Reach_	B_Reach_	B_Weight_	R_Weight_	B_age	R_age
title_bout	1	0.565275	-0.07014	-0.03219	-0.08951	-0.11296	0.109318	0.075268	0.005697	0.001943	0.028434	0.001282	0.05236	0.043677	0.026505	0.000712
no_of_rou	0.565275	1	0.029174	0.057937	0.011205	0.017226	0.29861	0.294163	-0.01769	-0.02242	0.064704	0.05497	-0.06709	-0.03857	0.128816	0.1088
B_avg_opp	-0.07014	0.029174	1	0.169611	0.576895	0.15733	0.052165	0.05171	-0.19579	-0.23563	-0.17473	-0.24677	-0.22214	-0.2183	0.008092	0.029611
R_avg_opp	-0.03219	0.057937	0.169611	1	0.16658	0.591722	0.039011	0.107778	-0.29628	-0.23592	-0.30005	-0.22017	-0.26842	-0.26746	-0.02695	0.024521
B_avg_opp	-0.08951	0.011205	0.576895	0.16658	1	0.183431	0.064146	0.018986	-0.18784	-0.20526	-0.18152	-0.22223	-0.21755	-0.21253	0.02361	0.030355
R_avg_opp	-0.11296	0.017226	0.15733	0.591722	0.183431	1	0.04686	0.122252	-0.22478	-0.19736	-0.24741	-0.18462	-0.24233	-0.24733	-0.01026	0.063331
B_total_ro	0.109318	0.29861	0.052165	0.039011	0.064146	0.04686	1	0.444173	0.027543	0.028081	0.061257	0.067367	0.005147	0.013892	0.367756	0.228915
R_total_ro	0.075268	0.294163	0.05171	0.107778	0.018986	0.122252	0.444173	1	-0.02894	0.011939	0.020603	0.061949	-0.03227	-0.02876	0.193969	0.441161
R_Height_	0.005697	-0.01769	-0.19579	-0.29628	-0.18784	-0.22478	0.027543	-0.02894	1	0.703412	0.890929	0.690578	0.719981	0.781129	0.160036	0.075815
B_Height_	0.001943	-0.02242	-0.23563	-0.23592	-0.20526	-0.19736	0.028081	0.011939	0.703412	1	0.683731	0.886634	0.751866	0.737309	0.095672	0.156141
R_Reach_	0.028434	0.064704	-0.17473	-0.30005	-0.18152	-0.24741	0.061257	0.020603	0.890929	0.683731	1	0.668273	0.732789	0.759562	0.163139	0.07603
B_Reach_	0.001282	0.05497	-0.24677	-0.22017	-0.22223	-0.18462	0.067367	0.061949	0.690578	0.886634	0.668273	1	0.758715	0.733341	0.091521	0.172905
B_Weight_	0.05236	-0.06709	-0.22214	-0.26842	-0.21755	-0.24233	0.005147	-0.03227	0.719981	0.751866	0.732789	0.758715	1	0.893073	0.204693	0.195187
R_Weight_	0.043677	-0.03857	-0.2183	-0.26746	-0.21253	-0.24733	0.013892	-0.02876	0.781129	0.737309	0.759562	0.733341	0.893073	1	0.214052	0.189736
B_age	0.026505	0.128816	0.008092	-0.02695	0.02361	-0.01026	0.367756	0.193969	0.160036	0.095672	0.163139	0.091521	0.204693	0.214052	1	0.188605
R_age	0.000712	0.1088	0.029611	0.024521	0.030355	0.063331	0.228915	0.441161	0.075815	0.156141	0.07603	0.172905	0.195187	0.189736	0.188605	1

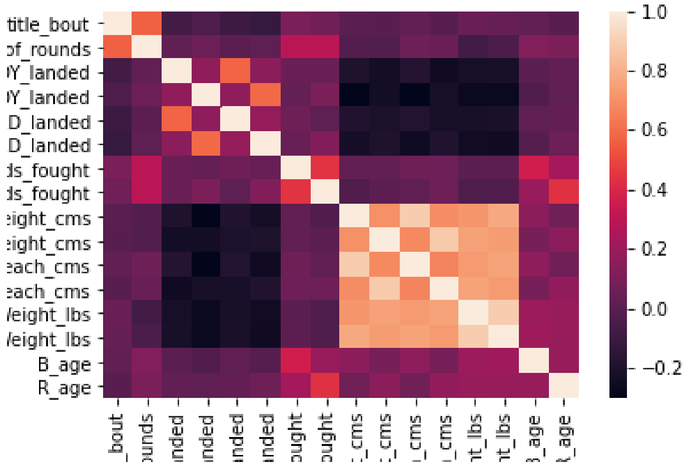
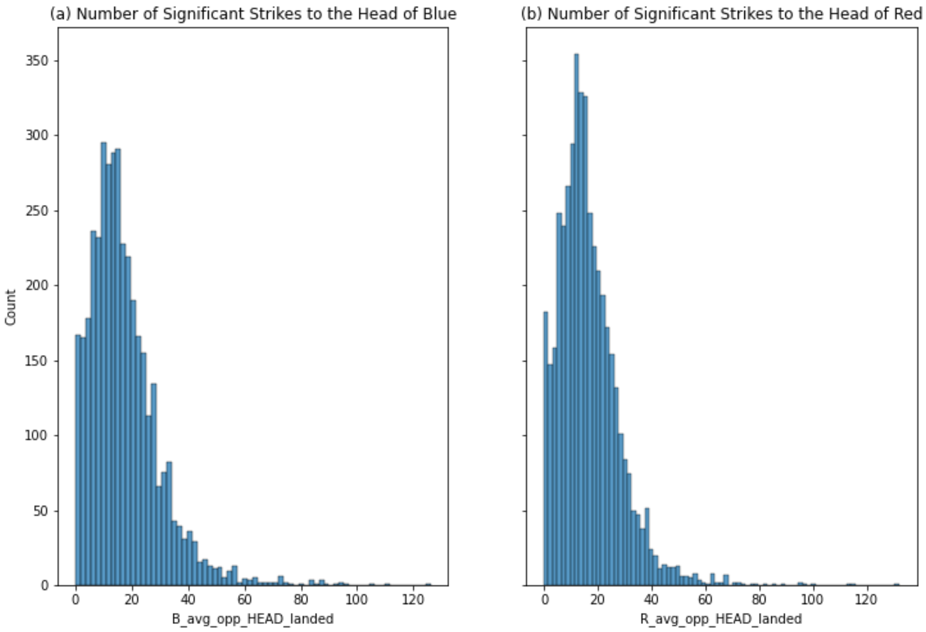


Figure 1 : Heatmap of the correlation matrix in Table 4

IV. DATA SET GRAPHICAL EXPLORATION

After retrieving the summary statistics of the continuous variables and having an idea of what types of categories each categorical value has, I developed various visualizations of the data to further investigate trends / correlations. This was generated through distribution charts, scatter plots, pairwise plots, bar charts, and other plots, such as line graphs and waffle charts.

When analyzing the distributions of head injuries by each fighter shown in **Figure 2**, what I found interesting was that the red fighters experienced more head injuries of about 20 on average, while the blue fighters experienced more diverse amounts of head injuries on average (see **Figure 2(a)**).


(a) Number of Significant Strikes to the Head of Blue

(b) Number of Significant Strikes to the Head of Red

Figure 2 : Histograms of the (a) number of significant strikes to the blue fighter's head on average; (b) number of significant strikes to the red fighter's head on average

When analyzing the distributions of body injuries by each fighter shown in **Figure 3**, the results had opposite conclusions of **Figure 2**; the red fighters experienced more diverse amounts of damage to the body on average, while the blue fighters had a peak average of body damage ranging from 0 to around 5.

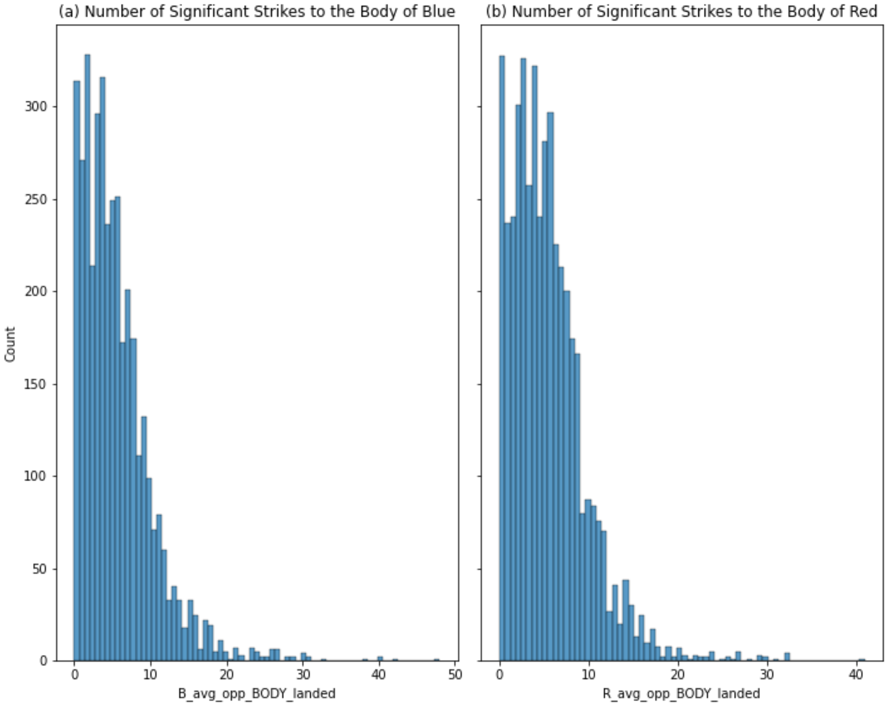


Figure 3 : Histograms of the (a) number of significant strikes to the blue fighter's body on average; (b) number of significant strikes to the red fighter's body on average

When analyzing the distributions of the ages of the fighters, I was not shocked by the results. I expected the majority of UFC fighters to be in the upper 20s to lower 30s, which is what **Figure 4** illustrates.

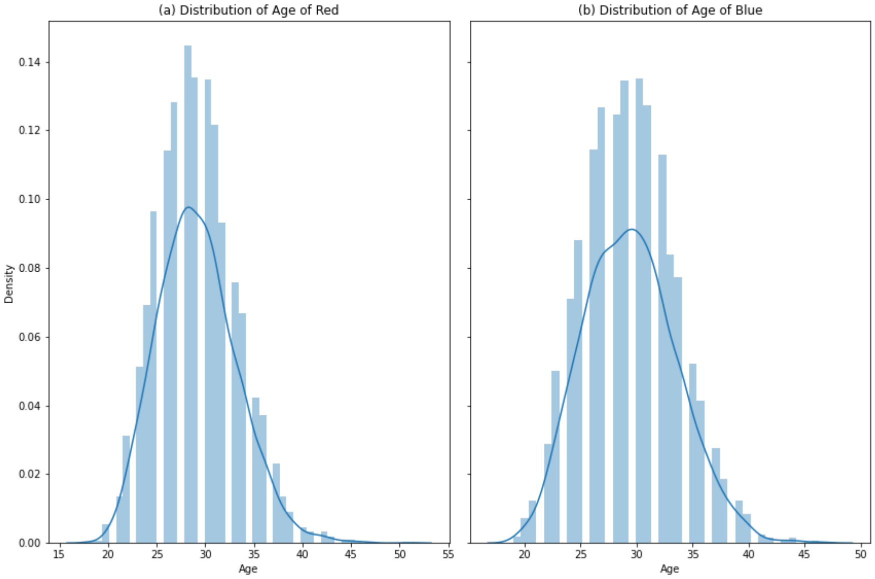


Figure 4 : Distribution plots of the (a) red fighter's age; (b) blue fighter's age

The scatter plots in pairwise plots are used for continuous data, similar to the variables shown in **Table 4**. I wanted to determine the relationship between the fighters' heights and reaches (arm spans), so I created a scatter plot for each type of fighter. Shown in **Figure 5**, the height and reach of the UFC fighters are strongly and positively correlated, which is what I expected.

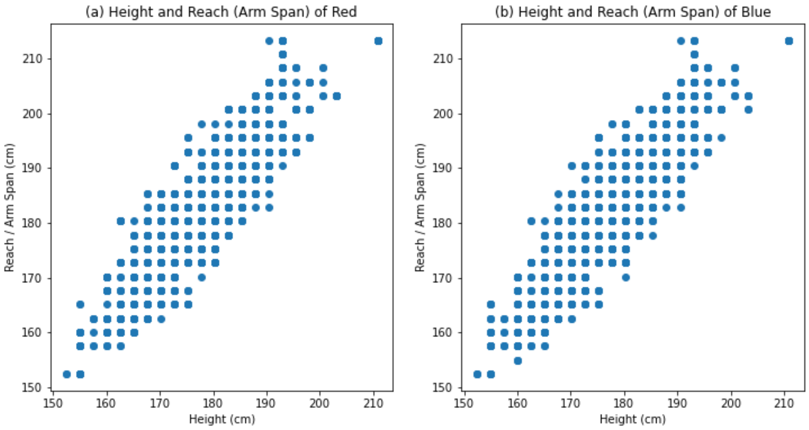


Figure 5 : Scatter plots of the (a) height and reach (arm span) in centimeters of the red fighter; (b) height and reach (arm span) in centimeters of the blue fighter

To get a general understanding of the relations / correlations between the continuous variables, pairwise plots were created (see **Figure 6** and **Figure 7**). **Figure 6** illustrates the relationships between the continuous variables that represent the fight statistics. For example, the variables used in **Figures 2** and **3** are considered fight statistics. What I found interesting about **Figure 6** was that the scatter plots between the total rounds fought and the average head and body damage received by the fighter were negatively correlated. In other words, the rounds fought by a fighter means they suffered less damage on average to the head and body. I do not consider this conclusion shocking, but I do consider it interesting as I did not think about it until the figure was created.

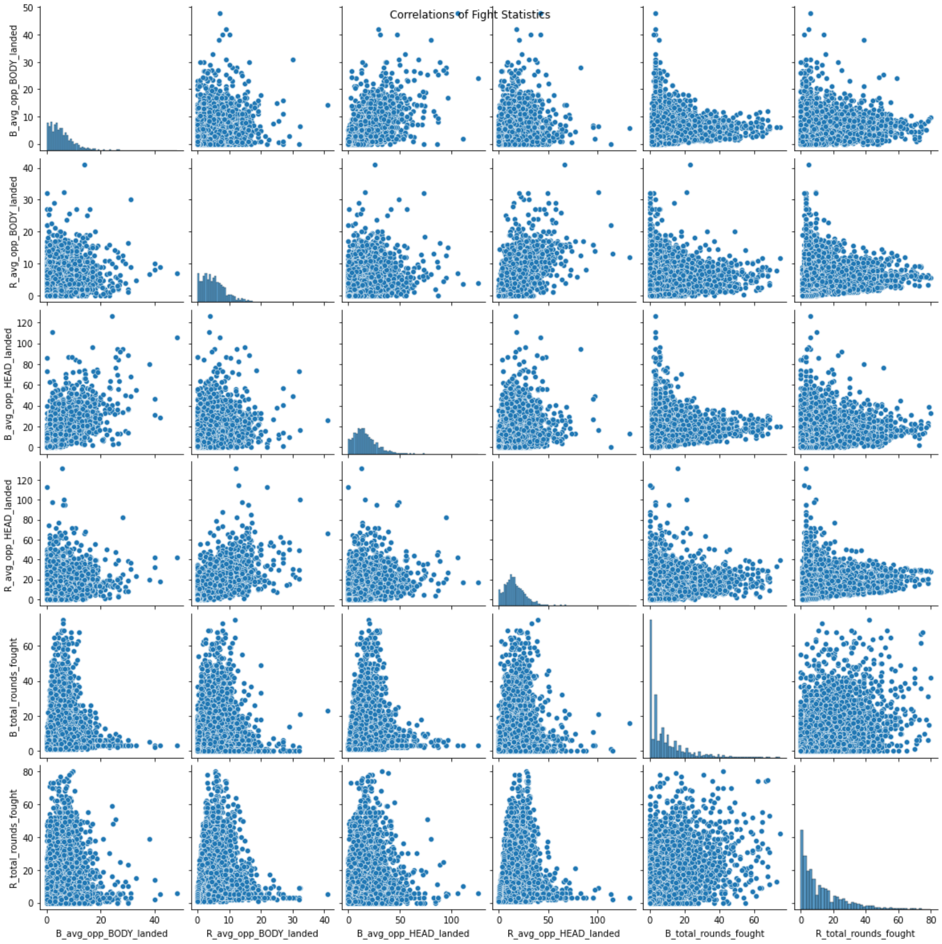


Figure 6 : Pairwise plot of the fight statistics variables (*B_avg_opp_BODY_landed*, *R_avg_opp_BODY_landed*, *B_avg_opp_HEAD_landed*, *R_avg_opp_HEAD _landed*, *B_total_rounds_fought*, *R_total_rounds_fought*) using scatter plots and distribution plots

Similar to **Figure 6**, I created a pairwise plot for the continuous variables that represent the fighter statistics. For example, the age variable used in **Figure 4** is considered fighter statistics. Many assumptions can be made from **Figure 7** since there are several different relationships between the fighter statistic variables shown. Some can easily be described as having a strong, positive correlation, while some do not have an easy description about them (e.g., plots shown in rows 1-4 and column 7-8 in **Figure 7**). These would be described as having weak correlations due to the unclarity of their relationship between the two variables.

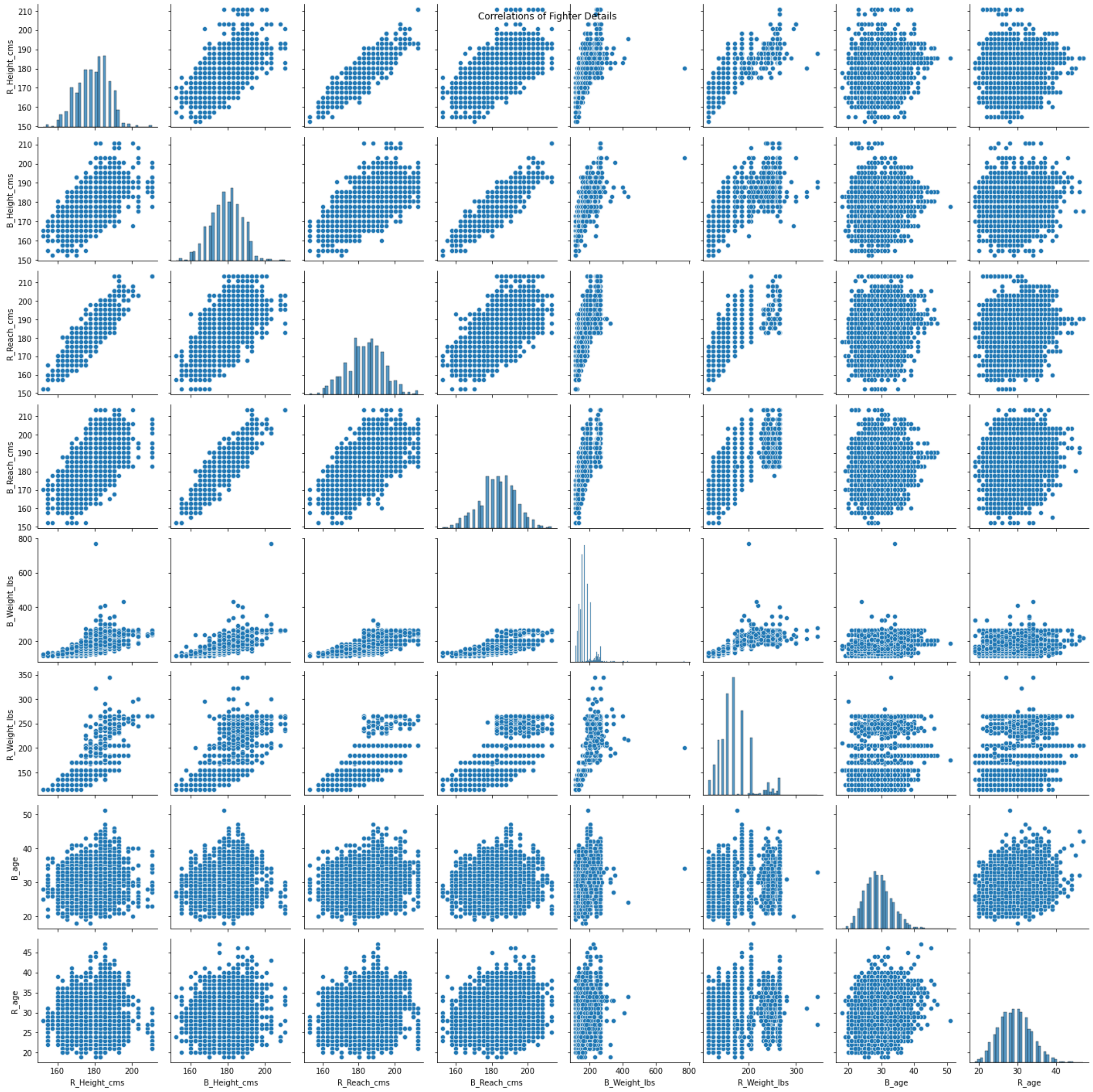


Figure 7 : Pairwise plot of the fighter statistics variables (*B_Height_cms*, *R_Height_cms*, *B_Reach_cms*, *R_Reach _cms*, *B_Weight_lbs*, *R_Weight_lbs*, *B_age*, *R_age*) using scatter plots and distribution plots

After analyzing loads of continuous data, I analyzed the categorical values variables to learn more about UFC fights and make specific assumptions. In **Figure 8**, the number of fights classified by the weight class is shown through a bar graph. What I did not know prior to investigating this data set is that UFC matches were split up by weight classes, nor did I know that there were weight classes specific to gender (e.g., women's featherweight). The specifics of the weight classes can be found here: https://www.abcboxing.com/unified-weight-classes-mma/

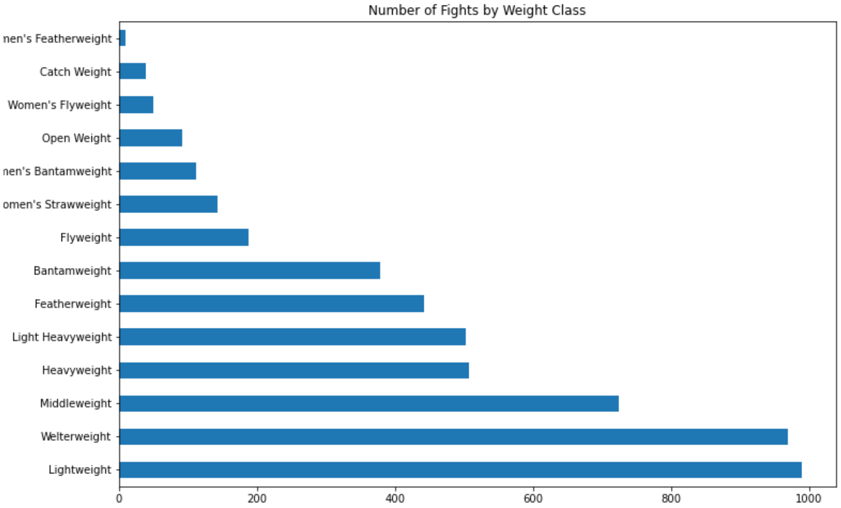


Figure 8 : Bar chart of the number of fights grouped by weight class

I wanted to determine whether there was a spike in UFC fights between 1993 to 2019 (the range of dates in the original dataset), so I created a line graph shown in **Figure 10(a)** to answer my curiosity. However, I wanted to show why a bar graph would seem ideal to determine and answer this question, but not ideal for a large dataset such as this. **Figure 9** illustrates the number of fights in each month and year from 1993 to 2019. This graph is impractical for this data set due to its size.

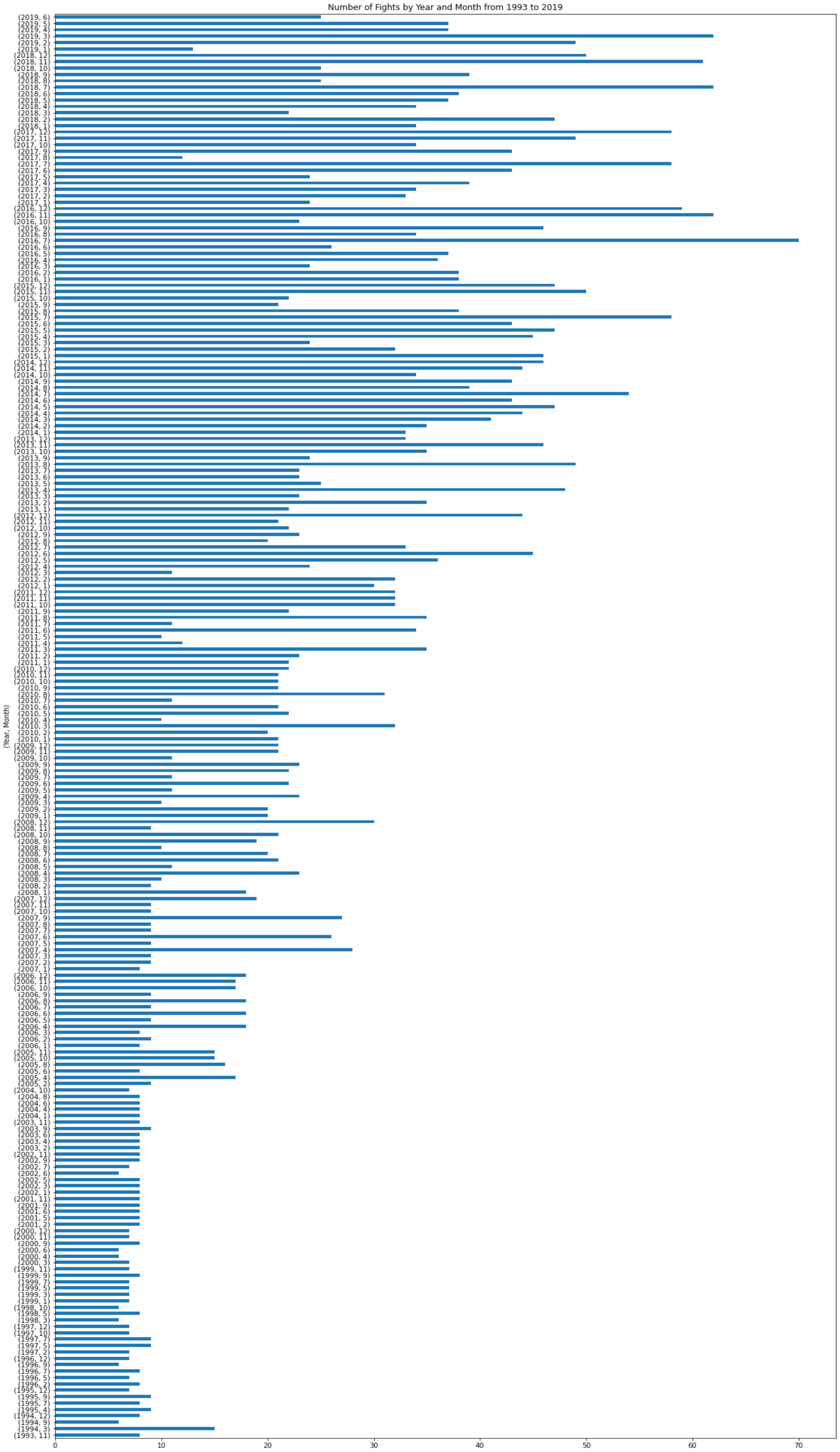


Figure 9 : Bar chart of the number of fights grouped by year and month from 1993 to 2019

As previously stated, I developed a line graph to show the spikes in UFC fights starting in 1993. **Figure 10** has three different line graphs: **10(a)** shows total number of fights between 1993 and 2019, **10(b)** shows the number of fights by month, and **10(c)** shows the number of fights by year. Since **Figure 10** is hard to analyze, I used **Figures 10(b)** and **10(c)** to make conclusions about **Figure 10(a).** The most popular month for UFC fights was November and December. There was also a spike in UFC fights from late 2004 to roughly 2012. Using those assumptions, a trend similar to what I just stated can be seen in **Figure 10(a)**.

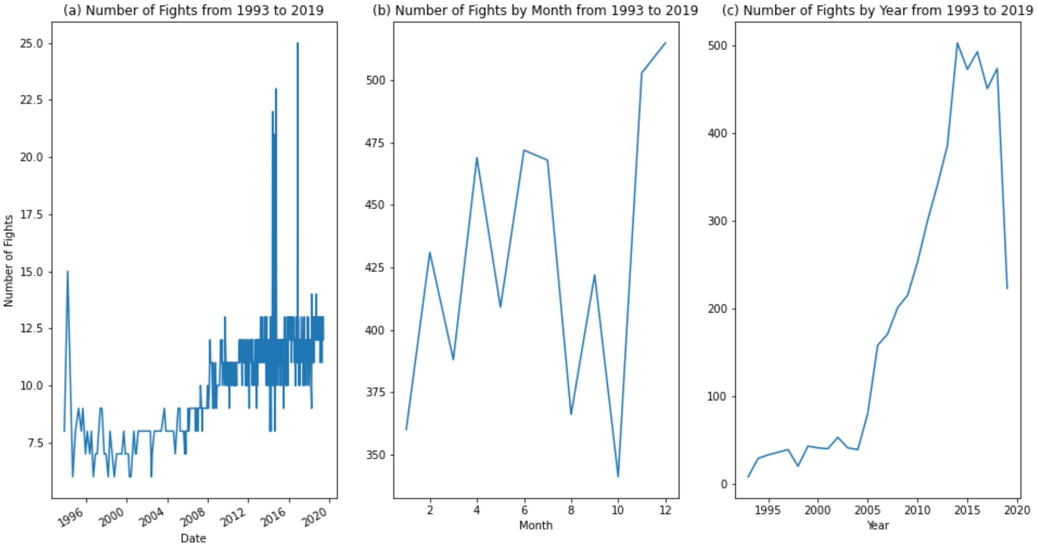


Figure 10 : Line graphs of the (a) number of fights from 1993 to 2019;
(b) number of fights by month from 1993 to 2019;
(c) number of fights by year from 1993 to 2019

For my last plot, I wanted to try something not discussed in previous classes: a waffle chart. A waffle chart is similar to a pie chart, as it shows the count or percentage of each category for a variable in relation to the other categories as a whole. I used the weight class variable once again since there were many categories but not as many as the *B_fighter* or *R_fighter* variables have. The top three largest weight classes shown in **Figure 11** are lightweight, welterweight, and middleweight, which corresponds to the results shown in **Figure 8**. Although not many conclusions can be made from **Figure 11**, I thought it was a plot type that not many people have seen, so I decided to investigate it through the pywaffle package in Python.

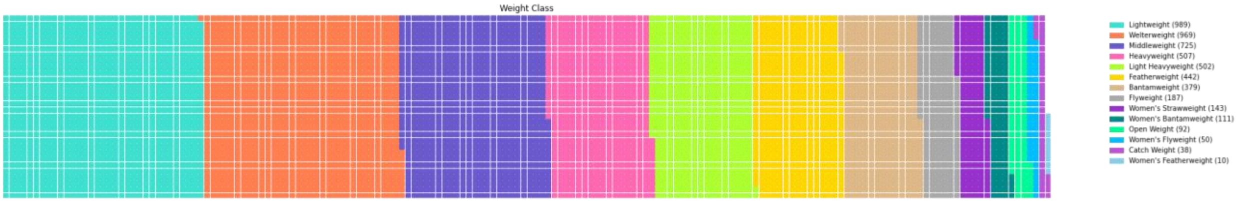


Figure 11 : Waffle Chart of the *weight_class* variable

V. SUMMARY OF FINDINGS

Despite my lack of knowledge on UFC fights, I was able to learn about specific trends based on the fights and the fighters themselves. I discovered that you have see matches are split up by weight classes of the fighters and relationships between the fight statistics variables and fighter variables. Specifically, I found that the less amount of strikes made to the fighters heads embodies on average implies more rounds were fought during a match. Another piece of information about UFC fighting I did not know as was that there was a spike in UFC matches from late 2004 to about mid 2012.