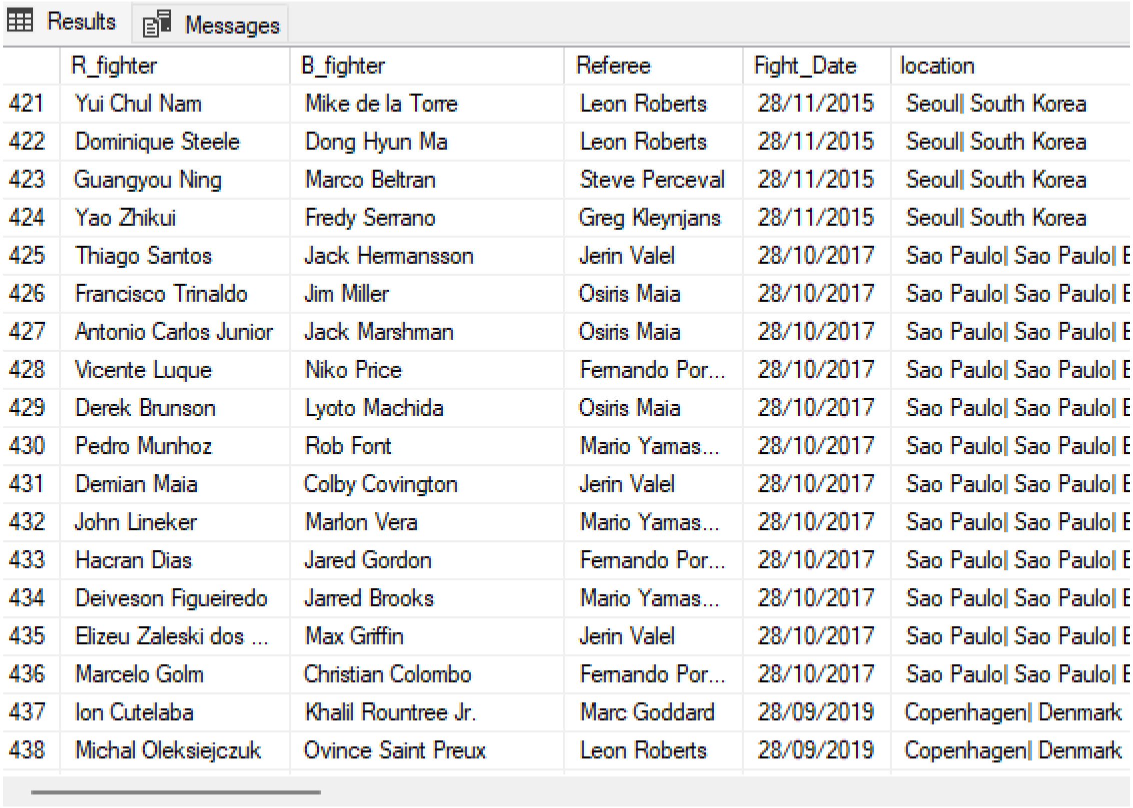
*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.

UFC data cleaning

First problem: I was asked to find out the most wins from a fighter in the last decade. My first thought was to check the 'Date' of the fights, as this would play a key component in the query that is executed

```
select * from UFC
order by fight_date DESC;
```

The following was shown, which demonstrates that the was a problem with the column with fight dates.



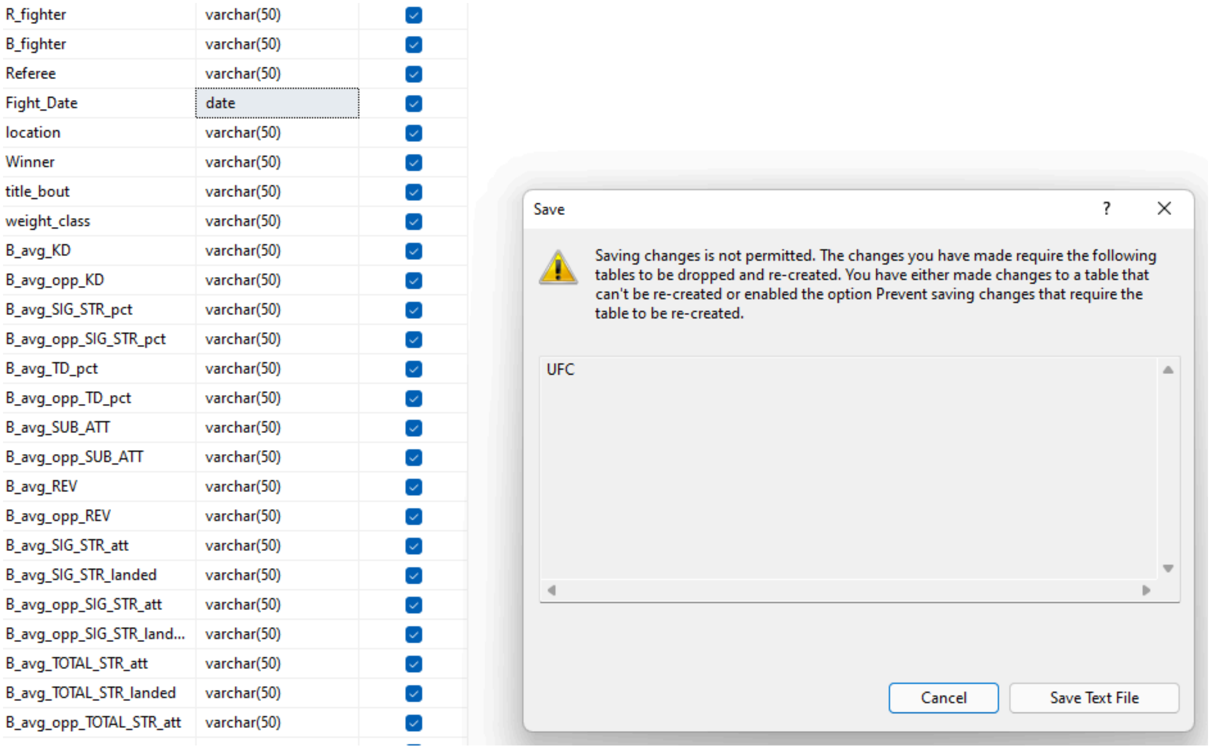
	R_fighter	B_fighter	Referee	Fight_Date	location
421	Yui Chul Nam	Mike de la Torre	Leon Roberts	28/11/2015	Seoul| South Korea
422	Dominique Steele	Dong Hyun Ma	Leon Roberts	28/11/2015	Seoul| South Korea
423	Guangyou Ning	Marco Beltran	Steve Perceval	28/11/2015	Seoul| South Korea
424	Yao Zhikui	Fredy Serrano	Greg Kleynjans	28/11/2015	Seoul| South Korea
425	Thiago Santos	Jack Hermansson	Jerin Valel	28/10/2017	Sao Paulo| Sao Paulo| E
426	Francisco Trinaldo	Jim Miller	Osiris Maia	28/10/2017	Sao Paulo| Sao Paulo| E
427	Antonio Carlos Junior	Jack Marshman	Osiris Maia	28/10/2017	Sao Paulo| Sao Paulo| E
428	Vicente Luque	Niko Price	Fernando Por...	28/10/2017	Sao Paulo| Sao Paulo| E
429	Derek Brunson	Lyoto Machida	Osiris Maia	28/10/2017	Sao Paulo| Sao Paulo| E
430	Pedro Munhoz	Rob Font	Mario Yamas...	28/10/2017	Sao Paulo| Sao Paulo| E
431	Demian Maia	Colby Covington	Jerin Valel	28/10/2017	Sao Paulo| Sao Paulo| E
432	John Lineker	Marlon Vera	Mario Yamas...	28/10/2017	Sao Paulo| Sao Paulo| E
433	Hacran Dias	Jared Gordon	Fernando Por...	28/10/2017	Sao Paulo| Sao Paulo| E
434	Deiveson Figueiredo	Jarred Brooks	Mario Yamas...	28/10/2017	Sao Paulo| Sao Paulo| E
435	Elizeu Zaleski dos ...	Max Griffin	Jerin Valel	28/10/2017	Sao Paulo| Sao Paulo| E
436	Marcelo Golm	Christian Colombo	Fernando Por...	28/10/2017	Sao Paulo| Sao Paulo| E
437	Ion Cutelaba	Khalil Rountree Jr.	Marc Goddard	28/09/2019	Copenhagen| Denmark
438	Michal Oleksiejczuk	Ovince Saint Preux	Leon Roberts	28/09/2019	Copenhagen| Denmark

We can see from the snapshot there was a problem with dates being ordered incorrectly.

My next thought was to check the table properties. The fight_date column had been set up with a VARCHAR data type by the uploader. This was the cause of the problem.

Second problem: Trying to rectify this by simply changing it from properties, I encountered a further problem (as seen in the next image):

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.



R_fighter	varchar(50)	☑
B_fighter	varchar(50)	☑
Referee	varchar(50)	☑
Fight_Date	date	☑
location	varchar(50)	☑
Winner	varchar(50)	☑
title_bout	varchar(50)	☑
weight_class	varchar(50)	☑
B_avg_KD	varchar(50)	☑
B_avg_opp_KD	varchar(50)	☑
B_avg_SIG_STR_pct	varchar(50)	☑
B_avg_opp_SIG_STR_pct	varchar(50)	☑
B_avg_TD_pct	varchar(50)	☑
B_avg_opp_TD_pct	varchar(50)	☑
B_avg_SUB_ATT	varchar(50)	☑
B_avg_opp_SUB_ATT	varchar(50)	☑
B_avg_REV	varchar(50)	☑
B_avg_opp_REV	varchar(50)	☑
B_avg_SIG_STR_att	varchar(50)	☑
B_avg_SIG_STR_landed	varchar(50)	☑
B_avg_opp_SIG_STR_att	varchar(50)	☑
B_avg_opp_SIG_STR_land...	varchar(50)	☑
B_avg_TOTAL_STR_att	varchar(50)	☑
B_avg_TOTAL_STR_landed	varchar(50)	☑
B_avg_opp_TOTAL_STR_att	varchar(50)	☑

Saving changes is not permitted. The changes you have made require the following tables to be dropped and re-created. You have either made changes to a table that can't be re-created or enabled the option Prevent saving changes that require the table to be re-created.

UFC

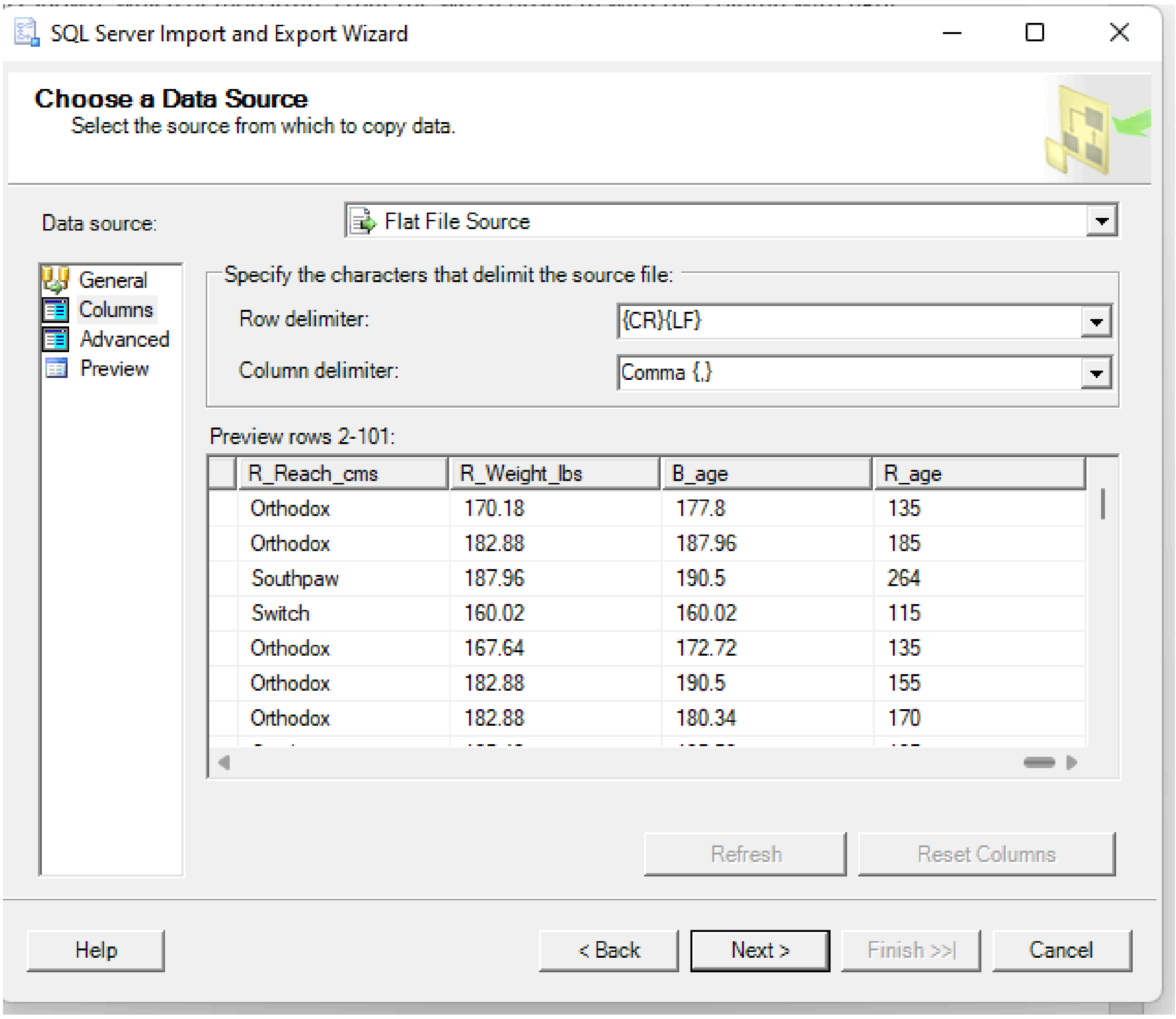
I could not simply make a change to the column data type.

Third problem: I then tried to use the ***CAST*** function to cast the fight_date column as date. This failed.

Lastly, I was required to re-import the data and this time checking that the columns had been set with the correct data types.

Simple fix, however, as I had not uploaded the original file the first time around. I checked the preview of the columns. Can you spot any issues in the image below?

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.



SQL Server Import and Export Wizard

Choose a Data Source

Select the source from which to copy data.

Data source: Flat File Source

Specify the characters that delimit the source file:

Row delimiter: {CR}{LF}

Column delimiter: Comma {,}

Preview rows 2-101:

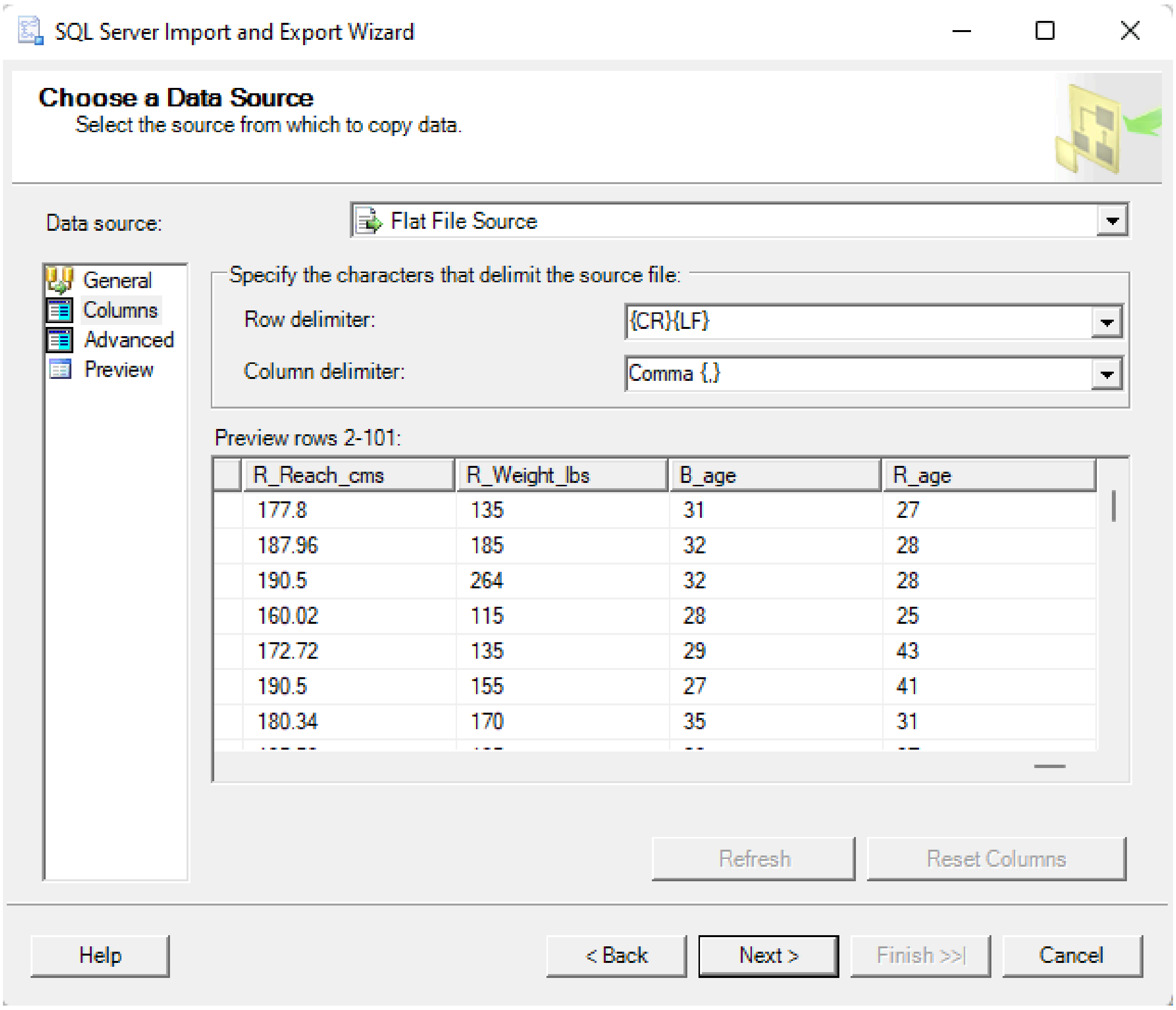
R_Reach_cms	R_Weight_lbs	B_age	R_age
Orthodox	170.18	177.8	135
Orthodox	182.88	187.96	185
Southpaw	187.96	190.5	264
Switch	160.02	160.02	115
Orthodox	167.64	172.72	135
Orthodox	182.88	190.5	155
Orthodox	182.88	180.34	170

Answer and fourth problem: you can see that the Age column has someone down as 177.8 and the rest are also incorrect. This was caused by the delimiter.

I then went through all the rows in the csv and discovered that the column named ‘location’ in the csv file had all data entered with ”city, state,country’’ i.e. separated with commas and this was causing the column values to spilt, hence the ages of fighter were different.

Solution: After replacing all commas with vertical bar lines (|) this rectified the issue. As can be seen in the image below:

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.



Data source: Flat File Source

Specify the characters that delimit the source file:

Row delimiter: {CR}{LF}

Column delimiter: Comma {,}

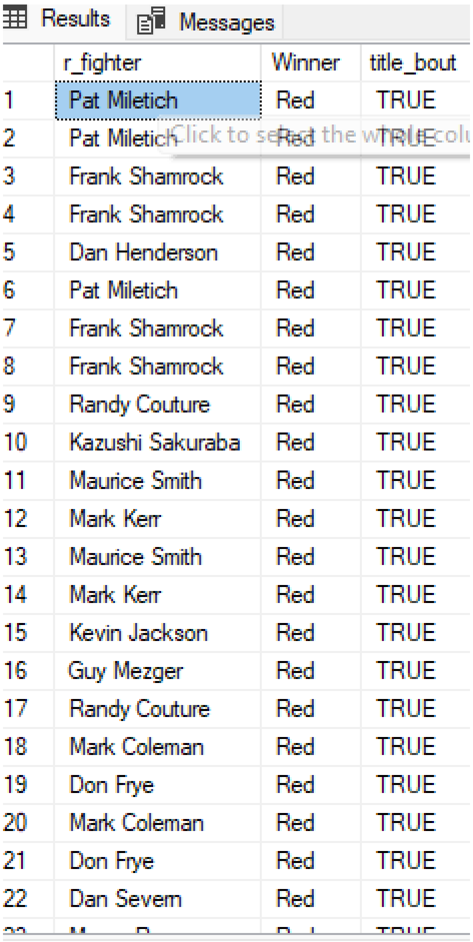
Preview rows 2-101:

R_Reach_cms	R_Weight_lbs	B_age	R_age
177.8	135	31	27
187.96	185	32	28
190.5	264	32	28
160.02	115	28	25
172.72	135	29	43
190.5	155	27	41
180.34	170	35	31

As it can be seen the Age is now correct and all other columns were checked for consistency.

QED.

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.

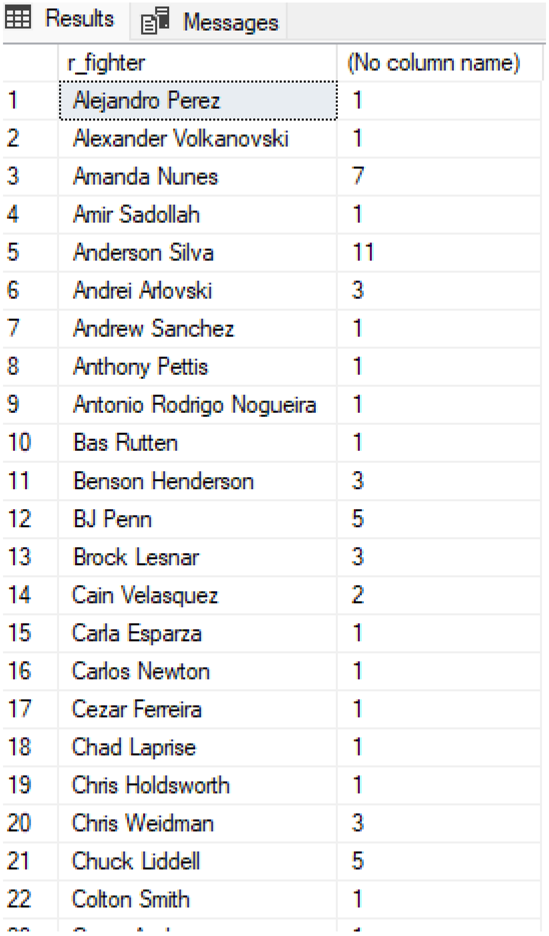


	r_fighter	Winner	title_bout
1	Pat Miletich	Red	TRUE
2	Pat Miletich	Red	TRUE
3	Frank Shamrock	Red	TRUE
4	Frank Shamrock	Red	TRUE
5	Dan Henderson	Red	TRUE
6	Pat Miletich	Red	TRUE
7	Frank Shamrock	Red	TRUE
8	Frank Shamrock	Red	TRUE
9	Randy Couture	Red	TRUE
10	Kazushi Sakuraba	Red	TRUE
11	Maurice Smith	Red	TRUE
12	Mark Kerr	Red	TRUE
13	Maurice Smith	Red	TRUE
14	Mark Kerr	Red	TRUE
15	Kevin Jackson	Red	TRUE
16	Guy Mezger	Red	TRUE
17	Randy Couture	Red	TRUE
18	Mark Coleman	Red	TRUE
19	Don Frye	Red	TRUE
20	Mark Coleman	Red	TRUE
21	Don Frye	Red	TRUE
22	Dan Severn	Red	TRUE

Looked good, so onto the fourth step - to group r_fighter, but I needed this more to count the number of times the R_fighters name shows, so that would determine how many times a title defense fight had been won.

```
select r_fighter, count(r_fighter)
from ufc
where title_bout = 'true' AND winner='red'
group by R_fighter;
```

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.

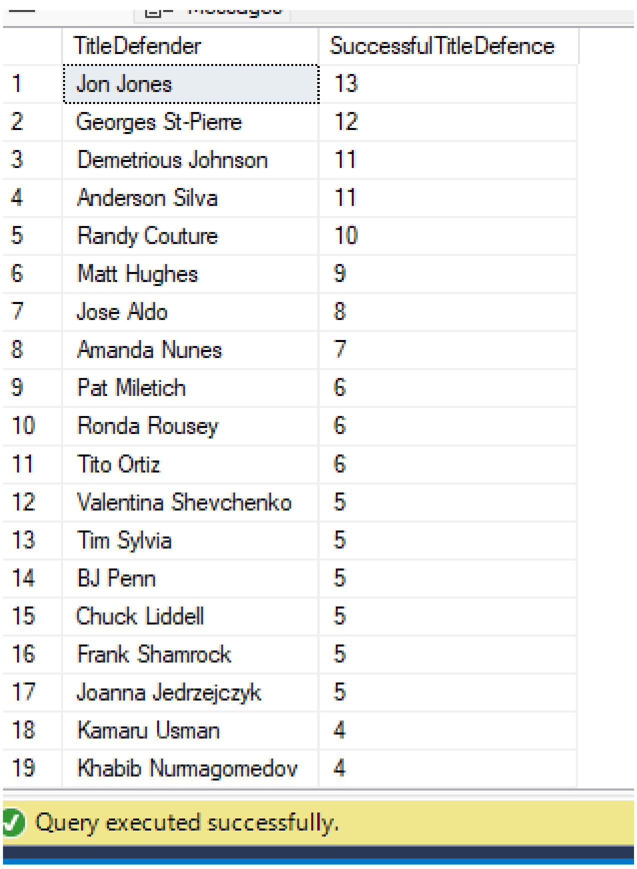


	r_fighter	(No column name)
1	Alejandro Perez	1
2	Alexander Volkanovski	1
3	Amanda Nunes	7
4	Amir Sadollah	1
5	Anderson Silva	11
6	Andrei Arlovski	3
7	Andrew Sanchez	1
8	Anthony Pettis	1
9	Antonio Rodrigo Nogueira	1
10	Bas Rutten	1
11	Benson Henderson	3
12	BJ Penn	5
13	Brock Lesnar	3
14	Cain Velasquez	2
15	Carla Esparza	1
16	Carlos Newton	1
17	Cezar Ferreira	1
18	Chad Laprise	1
19	Chris Holdsworth	1
20	Chris Weidman	3
21	Chuck Liddell	5
22	Colton Smith	1

Fifth, this is not in order. I wanted the fighter with the most defences to show i.e. in desc order. But the count column has no name, so an alias had to be assigned.

```
select r_fighter As TitleDefender, count(r_fighter) As SuccessfulTitleDefence
from ufc
where title_bout = 'true' AND winner='red'
group by R_fighter
order by SuccessfulTitleDefence DESC;
```

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.



	TitleDefender	SuccessfulTitleDefence
1	Jon Jones	13
2	Georges St-Pierre	12
3	Demetrious Johnson	11
4	Anderson Silva	11
5	Randy Couture	10
6	Matt Hughes	9
7	Jose Aldo	8
8	Amanda Nunes	7
9	Pat Miletich	6
10	Ronda Rousey	6
11	Tito Ortiz	6
12	Valentina Shevchenko	5
13	Tim Sylvia	5
14	BJ Penn	5
15	Chuck Liddell	5
16	Frank Shamrock	5
17	Joanna Jedrzejczyk	5
18	Kamaru Usman	4
19	Khabib Nurnagomedov	4

✓ Query executed successfully.

Now this looked good.

Some further scrutiny.

1. If the fighter had lost the champions title and then regained it, would they be included. Yes. As this would have meant they are in the Red corner and would have been making a title defence.

Limitations:

1. If there is a vacant title, the R_fighter would be the highest ranked fighter in that division so it does not necessarily mean it is a title defence fight. Though one could argue, that the title is there's as they are the next in line to take over and so should be included as a title defender.

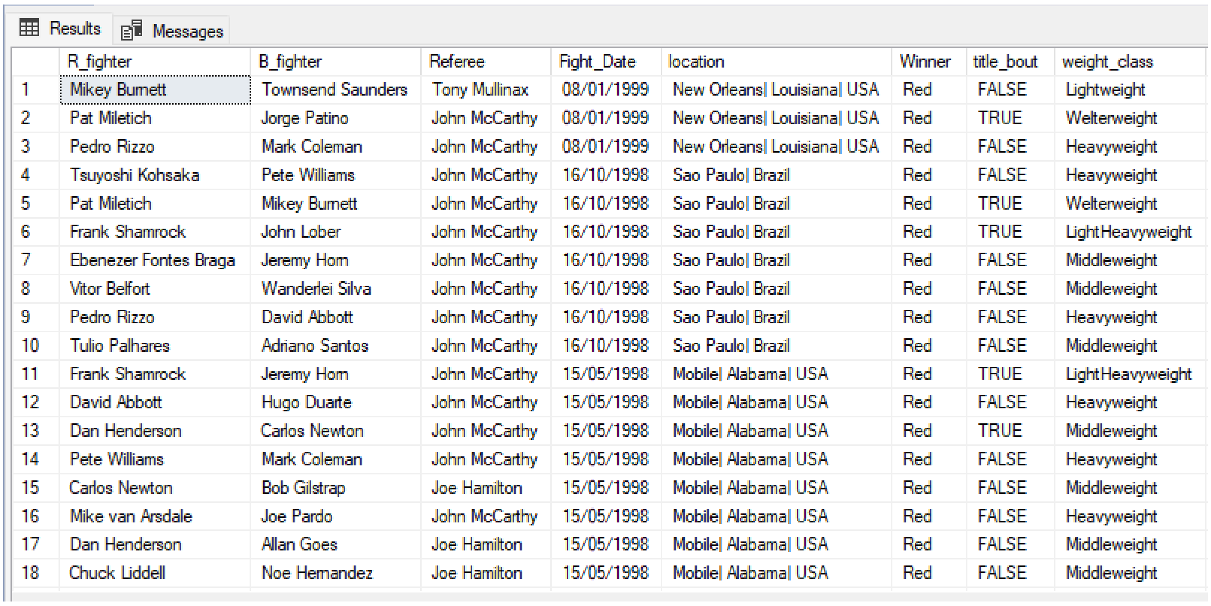
*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.

UPDATE 18th April

Second data project

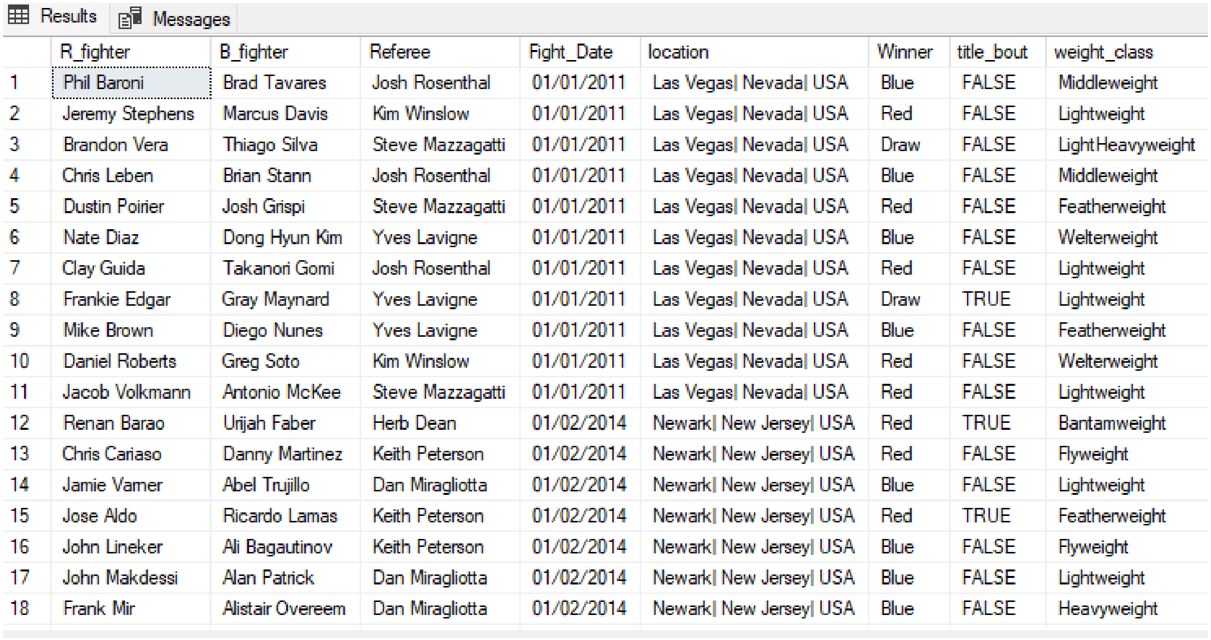
Cleaning the data – something that should have been done at the beginning.

1. As can be seen previously, there is a ‘date’ column named ‘date’ though SQL seemed to order it and understood what it needed to do, then ordered the date column. But this could cause confusion later on so the column name was changed to ‘Fight_Date’ to avoid the overlap of the function name and the column name.



	R_fighter	B_fighter	Referee	Fight_Date	location	Winner	title_bout	weight_class
1	Mikey Burnett	Townsend Saunders	Tony Mullinax	08/01/1999	New Orleans| Louisiana| USA	Red	FALSE	Lightweight
2	Pat Miletich	Jorge Patino	John McCarthy	08/01/1999	New Orleans| Louisiana| USA	Red	TRUE	Welterweight
3	Pedro Rizzo	Mark Coleman	John McCarthy	08/01/1999	New Orleans| Louisiana| USA	Red	FALSE	Heavyweight
4	Tsuyoshi Kohsaka	Pete Williams	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	FALSE	Heavyweight
5	Pat Miletich	Mikey Burnett	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	TRUE	Welterweight
6	Frank Shamrock	John Lober	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	TRUE	LightHeavyweight
7	Ebenezer Fontes Braga	Jeremy Horn	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	FALSE	Middleweight
8	Vitor Belfort	Wanderlei Silva	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	FALSE	Middleweight
9	Pedro Rizzo	David Abbott	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	FALSE	Heavyweight
10	Tulio Palhares	Adriano Santos	John McCarthy	16/10/1998	Sao Paulo| Brazil	Red	FALSE	Middleweight
11	Frank Shamrock	Jeremy Horn	John McCarthy	15/05/1998	Mobile| Alabama| USA	Red	TRUE	LightHeavyweight
12	David Abbott	Hugo Duarte	John McCarthy	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Heavyweight
13	Dan Henderson	Carlos Newton	John McCarthy	15/05/1998	Mobile| Alabama| USA	Red	TRUE	Middleweight
14	Pete Williams	Mark Coleman	John McCarthy	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Heavyweight
15	Carlos Newton	Bob Gilstrap	Joe Hamilton	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Middleweight
16	Mike van Arsdale	Joe Pardo	John McCarthy	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Heavyweight
17	Dan Henderson	Allan Goes	Joe Hamilton	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Middleweight
18	Chuck Liddell	Noe Hernandez	Joe Hamilton	15/05/1998	Mobile| Alabama| USA	Red	FALSE	Middleweight

2. However, it ordered it in a peculiar way, that I thought long and hard and then realised the column setting may be incorrect. It turned out the column was set up with VARCHAR.



	R_fighter	B_fighter	Referee	Fight_Date	location	Winner	title_bout	weight_class
1	Phil Baroni	Brad Tavares	Josh Rosenthal	01/01/2011	Las Vegas| Nevada| USA	Blue	FALSE	Middleweight
2	Jeremy Stephens	Marcus Davis	Kim Winslow	01/01/2011	Las Vegas| Nevada| USA	Red	FALSE	Lightweight
3	Brandon Vera	Thiago Silva	Steve Mazzagatti	01/01/2011	Las Vegas| Nevada| USA	Draw	FALSE	LightHeavyweight
4	Chris Leben	Brian Stann	Josh Rosenthal	01/01/2011	Las Vegas| Nevada| USA	Blue	FALSE	Middleweight
5	Dustin Poirier	Josh Grispi	Steve Mazzagatti	01/01/2011	Las Vegas| Nevada| USA	Red	FALSE	Featherweight
6	Nate Diaz	Dong Hyun Kim	Yves Lavigne	01/01/2011	Las Vegas| Nevada| USA	Blue	FALSE	Welterweight
7	Clay Guida	Takanori Gomi	Josh Rosenthal	01/01/2011	Las Vegas| Nevada| USA	Red	FALSE	Lightweight
8	Frankie Edgar	Gray Maynard	Yves Lavigne	01/01/2011	Las Vegas| Nevada| USA	Draw	TRUE	Lightweight
9	Mike Brown	Diego Nunes	Yves Lavigne	01/01/2011	Las Vegas| Nevada| USA	Blue	FALSE	Featherweight
10	Daniel Roberts	Greg Soto	Kim Winslow	01/01/2011	Las Vegas| Nevada| USA	Red	FALSE	Welterweight
11	Jacob Volkmann	Antonio McKee	Steve Mazzagatti	01/01/2011	Las Vegas| Nevada| USA	Red	FALSE	Lightweight
12	Renan Barao	Urijah Faber	Herb Dean	01/02/2014	Newark| New Jersey| USA	Red	TRUE	Bantamweight
13	Chris Cariaso	Danny Martinez	Keith Peterson	01/02/2014	Newark| New Jersey| USA	Red	FALSE	Flyweight
14	Jamie Varner	Abel Trujillo	Dan Miragliotta	01/02/2014	Newark| New Jersey| USA	Blue	FALSE	Lightweight
15	Jose Aldo	Ricardo Lamas	Keith Peterson	01/02/2014	Newark| New Jersey| USA	Red	TRUE	Featherweight
16	John Lineker	Ali Bagautinov	Keith Peterson	01/02/2014	Newark| New Jersey| USA	Blue	FALSE	Flyweight
17	John Makdessi	Alan Patrick	Dan Miragliotta	01/02/2014	Newark| New Jersey| USA	Blue	FALSE	Lightweight
18	Frank Mir	Alistair Overeem	Dan Miragliotta	01/02/2014	Newark| New Jersey| USA	Blue	FALSE	Heavyweight

*Disclaimer, the following data I am working with is available online for anyone to access and so this is not any confidential data I am working with.

3. As the file had been uploaded with ALL columns as VARCHAR, changing the setting on the original file was not possible.