Colin Mullaney

DS 4100

12/15/16

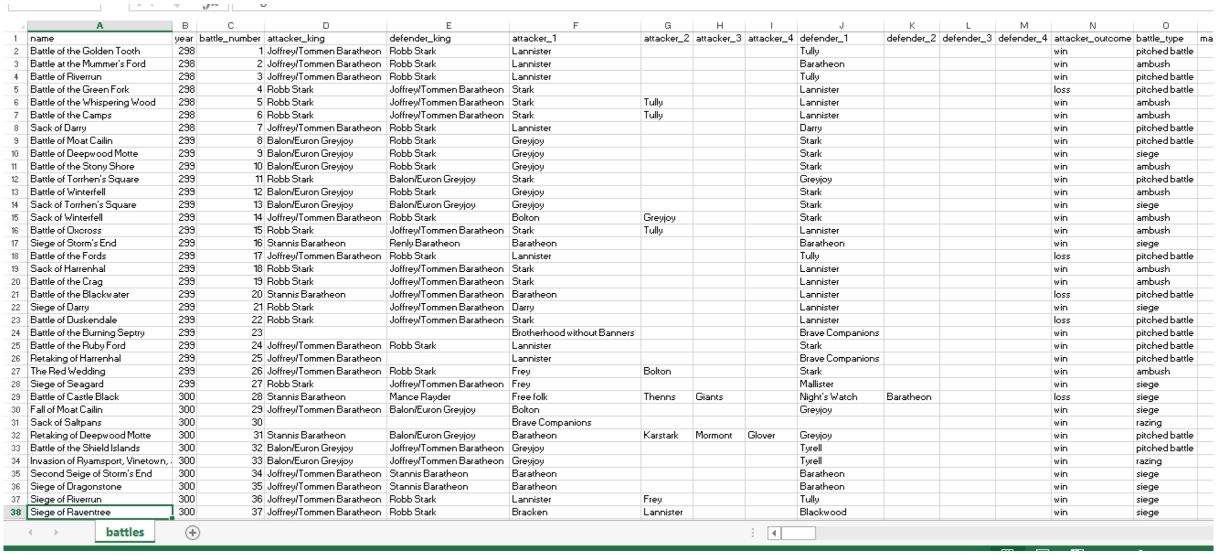
DS 4100 Term Project: Journal

For this project, I attempted to apply everything I had learned in class about the Data Science pipeline to conduct a start-to-finish analysis of an interesting data set of my choice. The two data sets I ended up using contained information about characters and battles in the popular book series, A Song of Ice and Fire (now a television series, "Game of Thrones"). After obtaining these data sets, I stored them in a MySQL database, and later cleaned them in R. Using Excel, I made multiple visualizations to get a better idea for the data. Finally, I constructed Naïve Bayes and Random Forest prediction models in R, to attempt to predict the outcome of a battle and the likelihood of a certain character dying in the series.

This project really helped to solidify the lessons and concepts I learned in class, while allowing me to choose my own domain for an analysis and manage the flow of the project. Below, I have described my experiences with each of the steps of the Data Science pipeline:

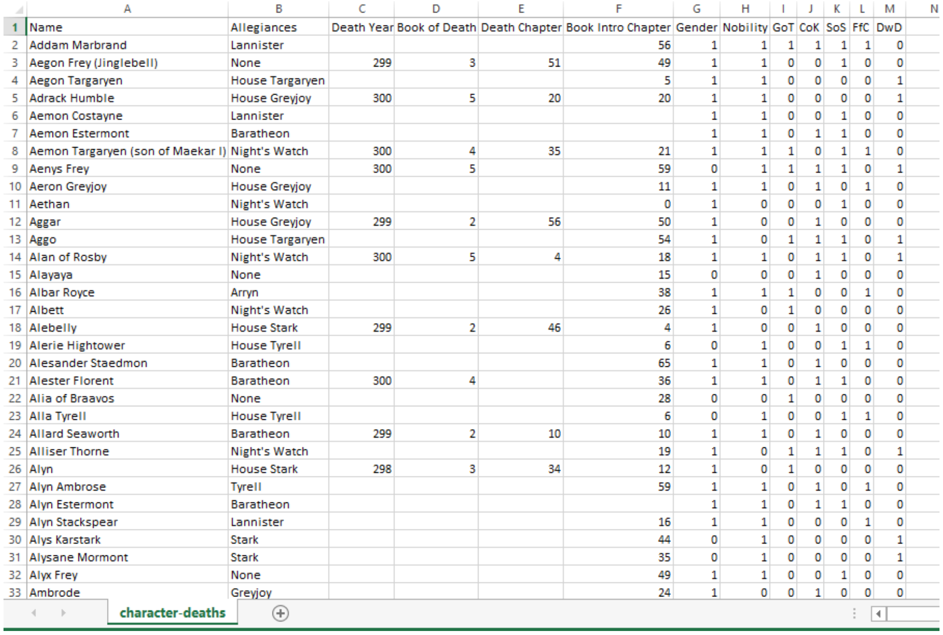
The Data:

I retrieved my two data sets from Kaggle.com as CSV files. The first data set, battles.csv, had information on 38 different battles that were mentioned in the book series. Each row contained the name of the battle, the year, the region, the attackers, the defenders, army sizes (for both attackers and defenders), battle type, and more. Below is a screen shot of the data set, which I initially opened in Excel:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	name	year	battle_number	attacker_king	defender_king	attacker_1	attacker_2	attacker_3	attacker_4	defender_1	defender_2	defender_3	defender_4	attacker_outcome	battle_type	ma
2	Battle of the Golden Tooth	298	1	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Tully				win	pitched battle	
3	Battle at the Mummer's Ford	298	2	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Baratheon				win	ambush	
4	Battle of Riverrun	298	3	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Tully				win	pitched battle	
5	Battle of the Green Fork	298	4	Robb Stark	Joffrey/Tommen Baratheon	Stark				Lannister				loss	pitched battle	
6	Battle of the Whispering Wood	298	5	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully			Lannister				win	ambush	
7	Battle of the Camps	298	6	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully			Lannister				win	ambush	
8	Sack of Darry	298	7	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Darry				win	pitched battle	
9	Battle of Moat Cailin	299	8	Balon/Euron Greyjoy	Robb Stark	Greyjoy				Stark				win	pitched battle	
10	Battle of Deepwood Motte	299	9	Balon/Euron Greyjoy	Robb Stark	Greyjoy				Stark				win	siege	
11	Battle of the Stony Shore	299	10	Balon/Euron Greyjoy	Robb Stark	Greyjoy				Stark				win	ambush	
12	Battle of Torrhen's Square	299	11	Robb Stark	Balon/Euron Greyjoy	Stark				Greyjoy				win	pitched battle	
13	Battle of Winterfell	299	12	Balon/Euron Greyjoy	Robb Stark	Greyjoy				Stark				win	ambush	
14	Sack of Torrhen's Square	299	13	Balon/Euron Greyjoy	Balon/Euron Greyjoy	Greyjoy				Stark				win	siege	
15	Sack of Winterfell	299	14	Joffrey/Tommen Baratheon	Robb Stark	Bolton	Greyjoy			Stark				win	ambush	
16	Battle of Oxcross	299	15	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully			Lannister				win	ambush	
17	Siege of Storm's End	299	16	Stannis Baratheon	Renly Baratheon	Baratheon				Baratheon				win	siege	
18	Battle of the Fords	299	17	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Tully				loss	pitched battle	
19	Sack of Harrenhal	299	18	Robb Stark	Joffrey/Tommen Baratheon	Stark				Lannister				win	ambush	
20	Battle of the Crag	299	19	Robb Stark	Joffrey/Tommen Baratheon	Stark				Lannister				win	ambush	
21	Battle of the Blackwater	299	20	Stannis Baratheon	Joffrey/Tommen Baratheon	Baratheon				Lannister				loss	pitched battle	
22	Siege of Darry	299	21	Robb Stark	Joffrey/Tommen Baratheon	Darry				Lannister				win	siege	
23	Battle of Duskendale	299	22	Robb Stark	Joffrey/Tommen Baratheon	Stark				Lannister				loss	pitched battle	
24	Battle of the Burning Septry	299	23			Brotherhood without Banners				Brave Companions				win	pitched battle	
25	Battle of the Ruby Ford	299	24	Joffrey/Tommen Baratheon	Robb Stark	Lannister				Stark				win	pitched battle	
26	Retaking of Harrenhal	299	25	Joffrey/Tommen Baratheon		Lannister				Brave Companions				win	pitched battle	
27	The Red Wedding	299	26	Joffrey/Tommen Baratheon	Robb Stark	Frey	Bolton			Stark				win	ambush	
28	Siege of Seagard	299	27	Robb Stark	Joffrey/Tommen Baratheon	Frey				Mallister				win	siege	
29	Battle of Castle Black	300	28	Stannis Baratheon	Mance Rayder	Free folk	Thenns	Giants		Night's Watch	Baratheon			loss	siege	
30	Fall of Moat Cailin	300	29	Joffrey/Tommen Baratheon	Balon/Euron Greyjoy	Bolton				Greyjoy				win	siege	
31	Sack of Saltpans	300	30			Brave Companions								win	razing	
32	Retaking of Deepwood Motte	300	31	Stannis Baratheon	Balon/Euron Greyjoy	Baratheon	Karstark	Mormont	Glover	Greyjoy				win	pitched battle	
33	Battle of the Shield Islands	300	32	Balon/Euron Greyjoy	Joffrey/Tommen Baratheon	Greyjoy				Tyrell				win	pitched battle	
34	Invasion of Ryamsport, Vinetown,	300	33	Balon/Euron Greyjoy	Joffrey/Tommen Baratheon	Greyjoy				Tyrell				win	razing	
35	Second Seige of Storm's End	300	34	Joffrey/Tommen Baratheon	Stannis Baratheon	Baratheon				Baratheon				win	siege	
36	Siege of Dragonstone	300	35	Joffrey/Tommen Baratheon	Stannis Baratheon	Baratheon				Baratheon				win	siege	
37	Siege of Riverrun	300	36	Joffrey/Tommen Baratheon	Robb Stark	Lannister	Frey			Tully				win	siege	
38	Siege of Raventree	300	37	Joffrey/Tommen Baratheon	Robb Stark	Bracken	Lannister			Blackwood				win	siege	

The next data set, character-deaths.csv, had information for 917 characters that were mentioned in the series. For each character, the data set contained a name, an allegiance, a year of death (if applicable), gender and nobility status, as well as information about which of the books they appeared in. Below is a screenshot of the data, opened in Excel:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Allegiances	Death Year	Book of Death	Death Chapter	Book Intro Chapter	Gender	Nobility	GoT	CoK	SoS	FfC	DwD
2	Addam Marbrand	Lannister				56	1	1	1	1	1	1	0
3	Aegon Frey (Jinglebell)	None	299	3	51	49	1	1	0	0	1	0	0
4	Aegon Targaryen	House Targaryen				5	1	1	0	0	0	0	1
5	Adrack Humble	House Greyjoy	300	5	20	20	1	1	0	0	0	0	1
6	Aemon Costayne	Lannister					1	1	0	0	1	0	0
7	Aemon Estermont	Baratheon					1	1	0	1	1	0	0
8	Aemon Targaryen (son of Maekar I)	Night's Watch	300	4	35	21	1	1	1	0	1	1	0
9	Aenys Frey	None	300	5		59	0	1	1	1	1	0	1
10	Aeron Greyjoy	House Greyjoy				11	1	1	0	1	0	1	0
11	Aethan	Night's Watch				0	1	0	0	0	1	0	0
12	Aggar	House Greyjoy	299	2	56	50	1	0	0	1	0	0	0
13	Aggo	House Targaryen				54	1	0	1	1	1	0	1
14	Alan of Rosby	Night's Watch	300	5	4	18	1	1	0	1	1	0	1
15	Alayaya	None				15	0	0	0	1	0	0	0
16	Albar Royce	Arryn				38	1	1	1	0	0	1	0
17	Albett	Night's Watch				26	1	0	1	0	0	0	0
18	Alebelly	House Stark	299	2	46	4	1	0	0	1	0	0	0
19	Alerie Hightower	House Tyrell				6	0	1	0	0	1	1	0
20	Alesander Staedmon	Baratheon				65	1	1	0	1	0	0	0
21	Alester Florent	Baratheon	300	4		36	1	1	0	1	1	0	0
22	Alia of Braavos	None				28	0	0	1	0	0	0	0
23	Alla Tyrell	House Tyrell				6	0	1	0	0	1	1	0
24	Allard Seaworth	Baratheon	299	2	10	10	1	1	0	1	0	0	0
25	Alliser Thorne	Night's Watch				19	1	0	1	1	1	0	1
26	Alyn	House Stark	298	3	34	12	1	0	1	0	0	0	0
27	Alyn Ambrose	Tyrell				59	1	1	0	1	0	1	0
28	Alyn Estermont	Baratheon					1	1	0	1	1	0	0
29	Alyn Stackspear	Lannister				16	1	1	0	0	0	1	0
30	Alys Karstark	Stark				44	0	1	0	0	0	0	1
31	Alysane Mormont	Stark				35	0	1	0	0	0	0	1
32	Alyx Frey	None				49	1	1	0	0	1	0	0
33	Ambrode	Greyjoy				24	1	0	0	1	0	0	0

After retrieving this data, I tried to assess its quality to ensure that it could be used for analysis. The data set containing the information on various characters was very complete, the only missing values being the fields regarding the characters' death. Having some background knowledge of the book series, I quickly looked through to find some of the more popular characters and validate their information was correct. Everything seemed to be in order, so I trust that this data set was of high enough quality to conduct an analysis. One problem with the data (that I cannot validate) is that the missing values for death year, death chapter etc, could be missing because the character did not die, or could be missing because the information regarding that character's death was simply not recorded. Since I considered any character without a death year to be alive, this could impact the validity of my analysis. However, I have no way around this, so I continued with my analysis.

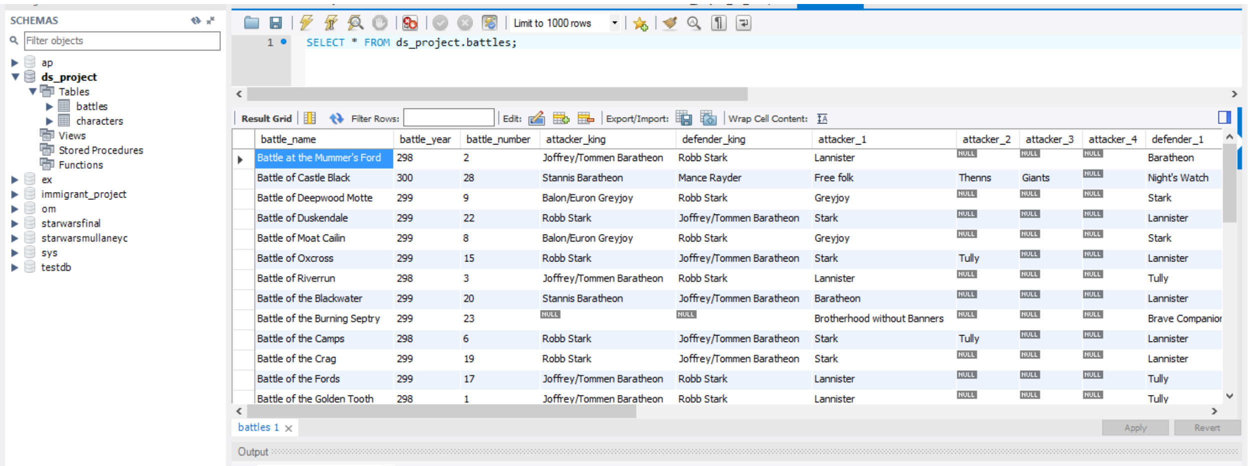
The battles data set was of significantly worse quality. Many of the rows had missing values, most predominantly in the attacker and defender army size columns. Of the 38 data points that were included, only 15 of them were complete, with absolutely no missing values. I chose to continue on with the analysis of this data set, for the sake of the learning experience, however I would not have used this data in an analysis for a company.

Data Storage:

I stored both of these data sets in MySQL, so that I could later load them into R. Normally I would not have bothered storing the battles data set in a database, because it only contained 38 rows, but I wanted to practice my data storage skills.

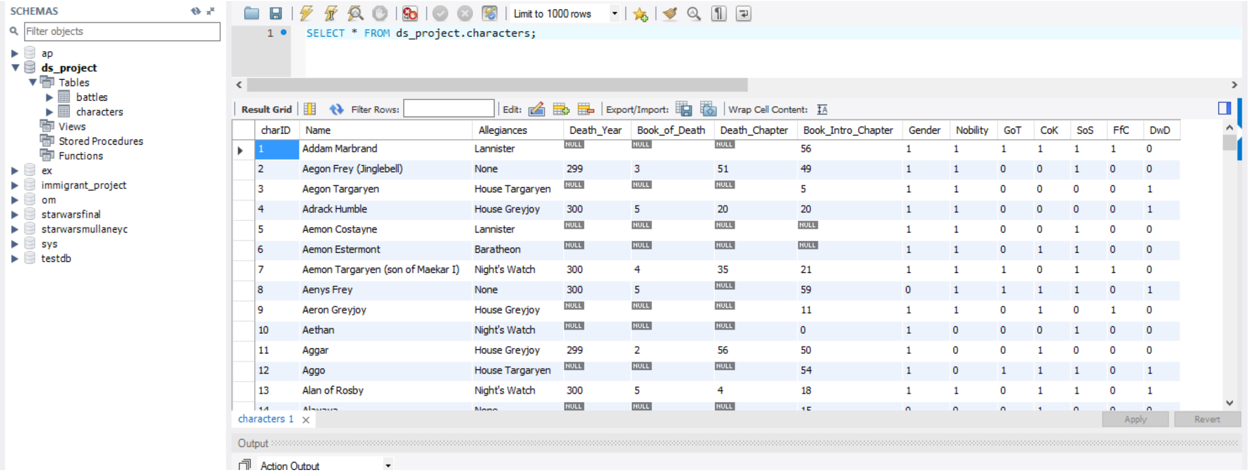
I used the MySQL workbench to load the data in through the CSV files, and automatically create the two different tables: battles and characters. I originally planned on using SQLite, because of its relative speed, but I later decided to use MySQL. I encountered troubles while trying to load the CSVs into SQLite, so I switched to MySQL, which had a much more user-friendly interface that allowed me to successfully load the CSV data. Also, if this were a larger scale project, SQLite would not be as efficient a choice, so it was better to get experience with MySQL. I chose to load the uncleaned data into the MySQL database first, and then clean it in R later on.

Below are screenshots of both data tables in MySQL:



```
1 ● SELECT * FROM ds_project.battles;
```

battle_name	battle_year	battle_number	attacker_king	defender_king	attacker_1	attacker_2	attacker_3	attacker_4	defender_1
Battle at the Mummer's Ford	298	2	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NULL	NULL	NULL	Baratheon
Battle of Castle Black	300	28	Stannis Baratheon	Mance Rayder	Free folk	Thenns	Giants	NULL	Night's Watch
Battle of Deepwood Motte	299	9	Balon/Euron Greyjoy	Robb Stark	Greyjoy	NULL	NULL	NULL	Stark
Battle of Duskendale	299	22	Robb Stark	Joffrey/Tommen Baratheon	Stark	NULL	NULL	NULL	Lannister
Battle of Moat Cailin	299	8	Balon/Euron Greyjoy	Robb Stark	Greyjoy	NULL	NULL	NULL	Stark
Battle of Oxcross	299	15	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully	NULL	NULL	Lannister
Battle of Riverrun	298	3	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NULL	NULL	NULL	Tully
Battle of the Blackwater	299	20	Stannis Baratheon	Joffrey/Tommen Baratheon	Baratheon	NULL	NULL	NULL	Lannister
Battle of the Burning Septry	299	23	NULL	NULL	Brotherhood without Banners	NULL	NULL	NULL	Brave Companior
Battle of the Camps	298	6	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully	NULL	NULL	Lannister
Battle of the Crag	299	19	Robb Stark	Joffrey/Tommen Baratheon	Stark	NULL	NULL	NULL	Lannister
Battle of the Fords	299	17	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NULL	NULL	NULL	Tully
Battle of the Golden Tooth	298	1	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NULL	NULL	NULL	Tully



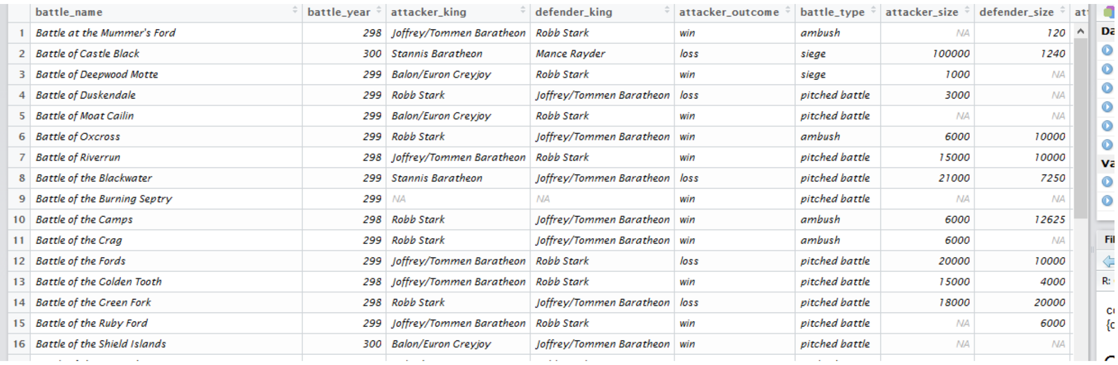
```
1 ● SELECT * FROM ds_project.characters;
```

charID	Name	Allegiances	Death_Year	Book_of_Death	Death_Chapter	Book_Intro_Chapter	Gender	Nobility	GoT	CoK	SoS	FfC	DwD
1	Addam Marbrand	Lannister	NULL	NULL	NULL	56	1	1	1	1	1	1	0
2	Aegon Frey (Jinglebell)	None	299	3	51	49	1	1	0	0	1	0	0
3	Aegon Targaryen	House Targaryen	NULL	NULL	NULL	5	1	1	0	0	0	0	1
4	Adrack Humble	House Greyjoy	300	5	20	20	1	1	0	0	0	0	1
5	Aemon Costayne	Lannister	NULL	NULL	NULL	NULL	1	1	0	0	1	0	0
6	Aemon Estermont	Baratheon	NULL	NULL	NULL	NULL	1	1	0	1	1	0	0
7	Aemon Targaryen (son of Maekar I)	Night's Watch	300	4	35	21	1	1	1	0	1	1	0
8	Aenys Frey	None	300	5	NULL	59	0	1	1	1	1	0	1
9	Aeron Greyjoy	House Greyjoy	NULL	NULL	NULL	11	1	1	0	1	0	1	0
10	Aethan	Night's Watch	NULL	NULL	NULL	0	1	0	0	0	1	0	0
11	Aggar	House Greyjoy	299	2	56	50	1	0	0	1	0	0	0
12	Aggo	House Targaryen	NULL	NULL	NULL	54	1	0	1	1	1	0	1
13	Alan of Rosby	Night's Watch	300	5	4	18	1	1	0	1	1	0	1
14	Alayaya	None	NULL	NULL	NULL	15	0	0	0	1	0	0	0

Retrieving the data from the MySQL database in R was relatively simple. I used the RMySQL package to connect to the database, and execute queries to grab the data. While all I did was select everything from the two tables, I could have easily filtered the data or executed more complex queries if they had been applicable to my project.

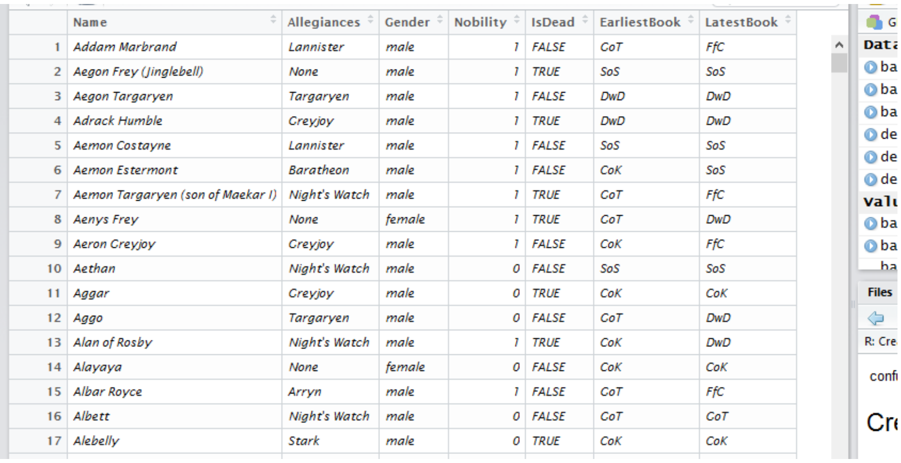
Data Cleaning:

After I loaded the data from my database into R, I began by cleaning both data frames. As expected, this step was the most time-consuming throughout the entire process. I started with the battles data set. To begin, I removed the 'Battle Number', 'Major Death', 'Major Capture', 'Location', and 'Note' columns, because they would not be used for my analysis. I also removed any rows that had missing values for the 'Attacker Outcome' column, as this was going to be my dependent variable for my analysis, and rows that were missing values would not be helpful. I did not remove rows that were missing other values, however. The Naïve Bayes model that I created has a way of handling NA values, so I decided to leave them as is, rather than removing them or guessing what value to replace them with. Next, I removed the 4 'attackers' columns and 4 'defenders' columns, and replaced them with one 'Attackers' and 'Defenders' column. These two columns contained the first attacker and first defender, respectively, from each row. This was a difficult choice for me because I was effectively losing information whenever I deleted attackers or defenders (because I only kept the first attacker or defender listed). I made this choice because of my method of prediction, and I will discuss it further in my final section about how I would change this analysis in the future. Below is a screenshot of the cleaned data in R:



	battle_name	battle_year	attacker_king	defender_king	attacker_outcome	battle_type	attacker_size	defender_size
1	Battle at the Mummer's Ford	298	Joffrey/Tommen Baratheon	Robb Stark	win	ambush	NA	120
2	Battle of Castle Black	300	Stannis Baratheon	Mance Rayder	loss	siege	100000	1240
3	Battle of Deepwood Motte	299	Balon/Euron Greyjoy	Robb Stark	win	siege	1000	NA
4	Battle of Duskendale	299	Robb Stark	Joffrey/Tommen Baratheon	loss	pitched battle	3000	NA
5	Battle of Moat Cailin	299	Balon/Euron Greyjoy	Robb Stark	win	pitched battle	NA	NA
6	Battle of Oxcross	299	Robb Stark	Joffrey/Tommen Baratheon	win	ambush	6000	10000
7	Battle of Riverrun	298	Joffrey/Tommen Baratheon	Robb Stark	win	pitched battle	15000	10000
8	Battle of the Blackwater	299	Stannis Baratheon	Joffrey/Tommen Baratheon	loss	pitched battle	21000	7250
9	Battle of the Burning Septry	299	NA	NA	win	pitched battle	NA	NA
10	Battle of the Camps	298	Robb Stark	Joffrey/Tommen Baratheon	win	ambush	6000	12625
11	Battle of the Crag	299	Robb Stark	Joffrey/Tommen Baratheon	win	ambush	6000	NA
12	Battle of the Fords	299	Joffrey/Tommen Baratheon	Robb Stark	loss	pitched battle	20000	10000
13	Battle of the Golden Tooth	298	Joffrey/Tommen Baratheon	Robb Stark	win	pitched battle	15000	4000
14	Battle of the Green Fork	298	Robb Stark	Joffrey/Tommen Baratheon	loss	pitched battle	18000	20000
15	Battle of the Ruby Ford	299	Joffrey/Tommen Baratheon	Robb Stark	win	pitched battle	NA	6000
16	Battle of the Shield Islands	300	Balon/Euron Greyjoy	Joffrey/Tommen Baratheon	win	pitched battle	NA	NA

For the characters data set, I removed "House" from all of the 'Allegiances' to make them consistent. Some rows in the data set, for example, had "House Stark", while others simply had "Stark". Since these two entries should be the same, I removed the "House" to ensure they were treated as the same allegiance. Next, I removed the three columns 'Death Year', 'Book of Death', and 'Death Chapter', and replaced them with a new column 'IsDead', that contained true for every row that had a value for 'Death Year' (rows with missing values would contain false). The three columns that I removed were not relevant for my analysis, so I removed them to reduce clutter in the data frame. Then, I changed the entries in the 'Gender' column from 1 and 0 to male and female, respectively. After that, I created two functions that would find the earliest book and latest book that each character appeared in. The data set that I found had a column for each of the 5 books in the series, and contained a 1 if the character appeared in that book. I reduced these 5 columns to 2 columns, 'EarliestBook' and 'LatestBook'. Below is a screenshot of the cleaned data frame in R:

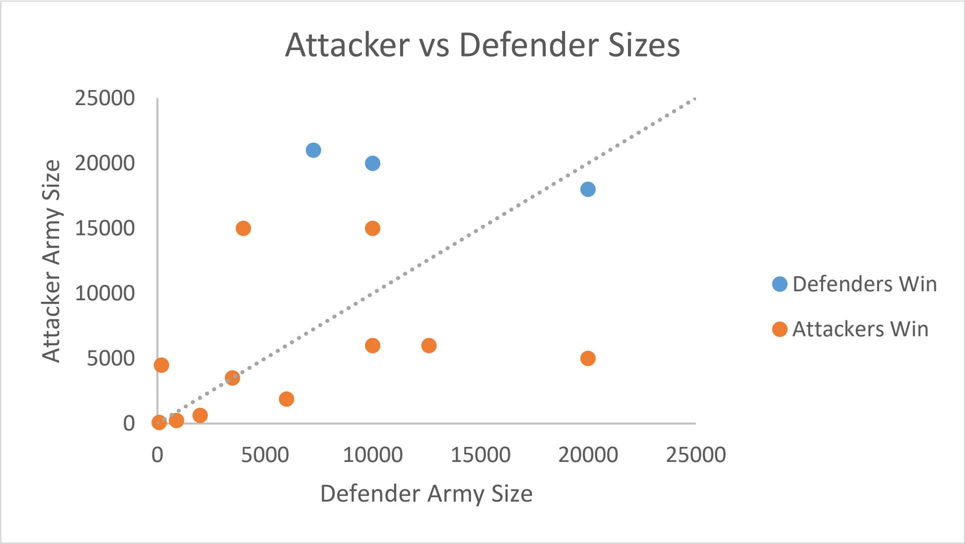


	Name	Allegiances	Gender	Nobility	IsDead	EarliestBook	LatestBook
1	Addam Marbrand	Lannister	male	1	FALSE	CoT	FfC
2	Aegon Frey (Jinglebell)	None	male	1	TRUE	SoS	SoS
3	Aegon Targaryen	Targaryen	male	1	FALSE	DwD	DwD
4	Adrack Humble	Greyjoy	male	1	TRUE	DwD	DwD
5	Aemon Costayne	Lannister	male	1	FALSE	SoS	SoS
6	Aemon Estermont	Baratheon	male	1	FALSE	CoK	SoS
7	Aemon Targaryen (son of Maekar I)	Night's Watch	male	1	TRUE	CoT	FfC
8	Aenys Frey	None	female	1	TRUE	CoT	DwD
9	Aeron Greyjoy	Greyjoy	male	1	FALSE	CoK	FfC
10	Aethan	Night's Watch	male	0	FALSE	SoS	SoS
11	Aggar	Greyjoy	male	0	TRUE	CoK	CoK
12	Aggo	Targaryen	male	0	FALSE	CoT	DwD
13	Alan of Rosby	Night's Watch	male	1	TRUE	CoK	DwD
14	Alayaya	None	female	0	FALSE	CoK	CoK
15	Albar Royce	Arryn	male	1	FALSE	CoT	FfC
16	Albett	Night's Watch	male	0	FALSE	CoT	CoT
17	Alebelly	Stark	male	0	TRUE	CoK	CoK

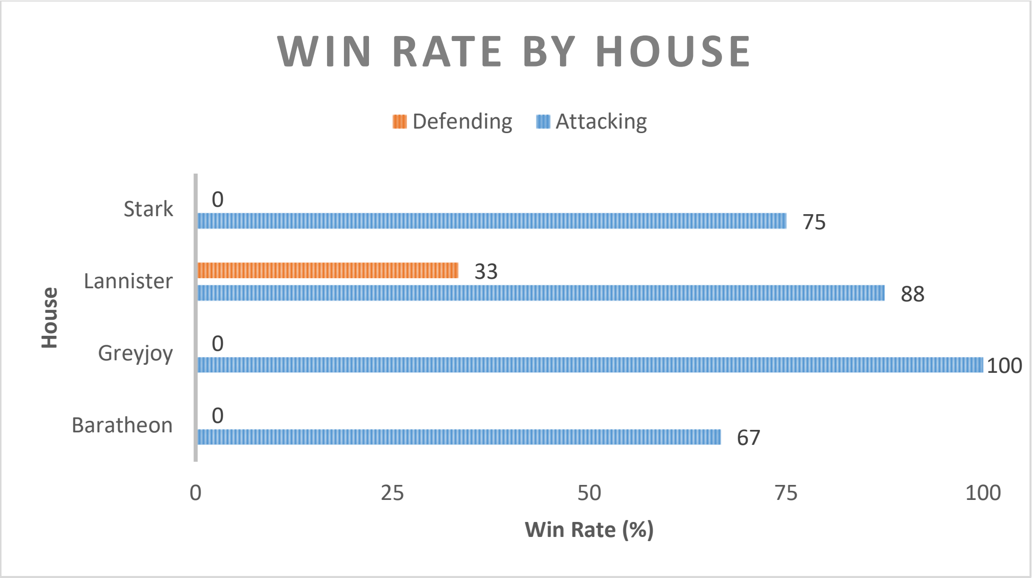
Visualizations:

In this section I will show, and explain, all of the visualizations that I created for this project. I used Excel to create each of these graphs. Since my data was largely categorical, and the results were Booleans (not continuous values), there was not a large variety of graphs that I could create. As such, I mostly relied on bar charts, since they let me easily compare different categories. Predictions from Naïve Bayes and Random Forest do not lend themselves as easily to visualizations (unlike linear regression which is easier to graph).

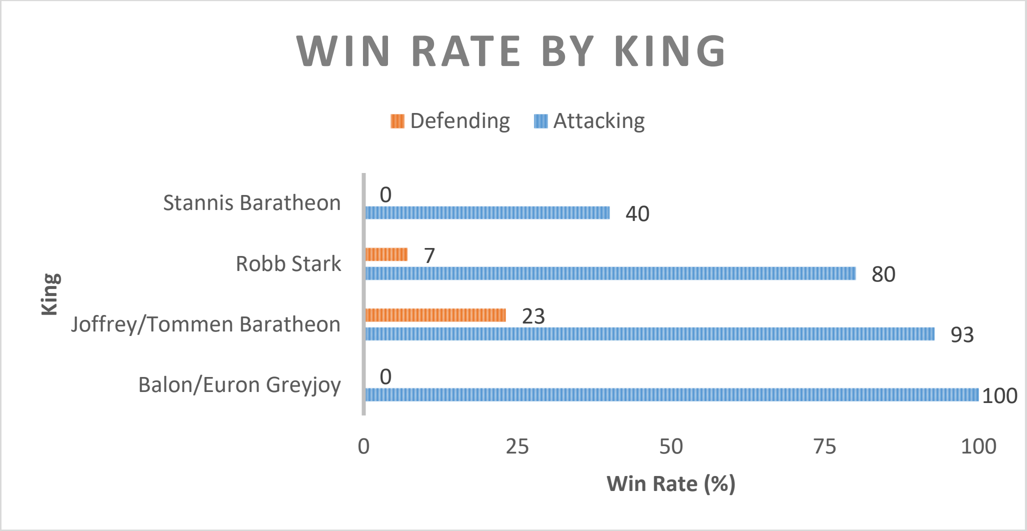
Battle Data:



I created this graph to see if there was a noticeable relationship between the army sizes and the outcome of the battle. I expected that whichever side had more troops would win, however this was not always the case. The 3 battles in which defenders won did not appear to have any trend, so I would conclude that army size alone is not enough to predict the winner of a battle. I excluded one data point that was an extreme outlier. The attacking army size was 100,000 and the defending army size was 1,240. The battle resulted in the defenders winning, which is unusual, given the massive difference in army sizes.

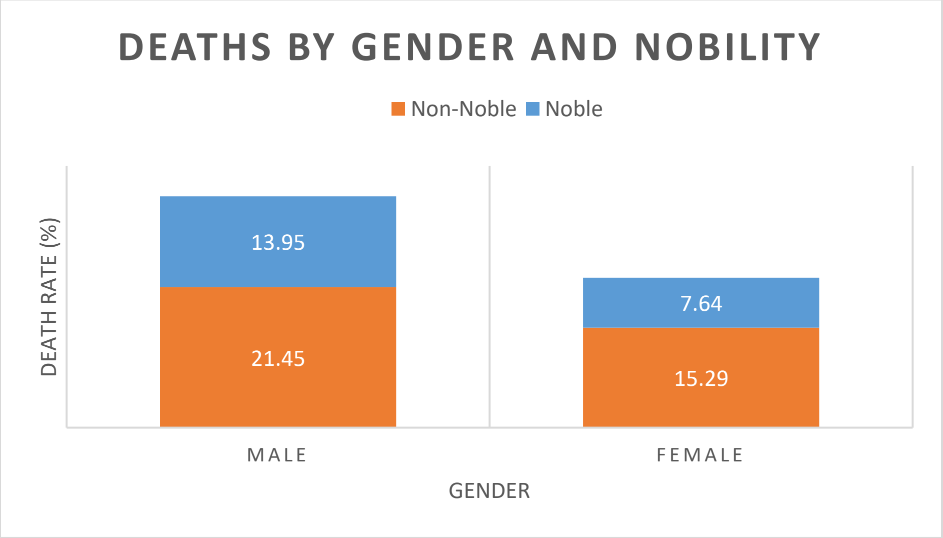


For this graph, I displayed the battle win rates for 4 major houses. I only chose Houses that had participated in at least 10 battles (whether attacking of defending). Many smaller Houses only participated in 1 or 2 battles, which would not be worth including. This graph shows that all of the Houses won more when attacking. Three of the four Houses lost every battle in which they were defenders.

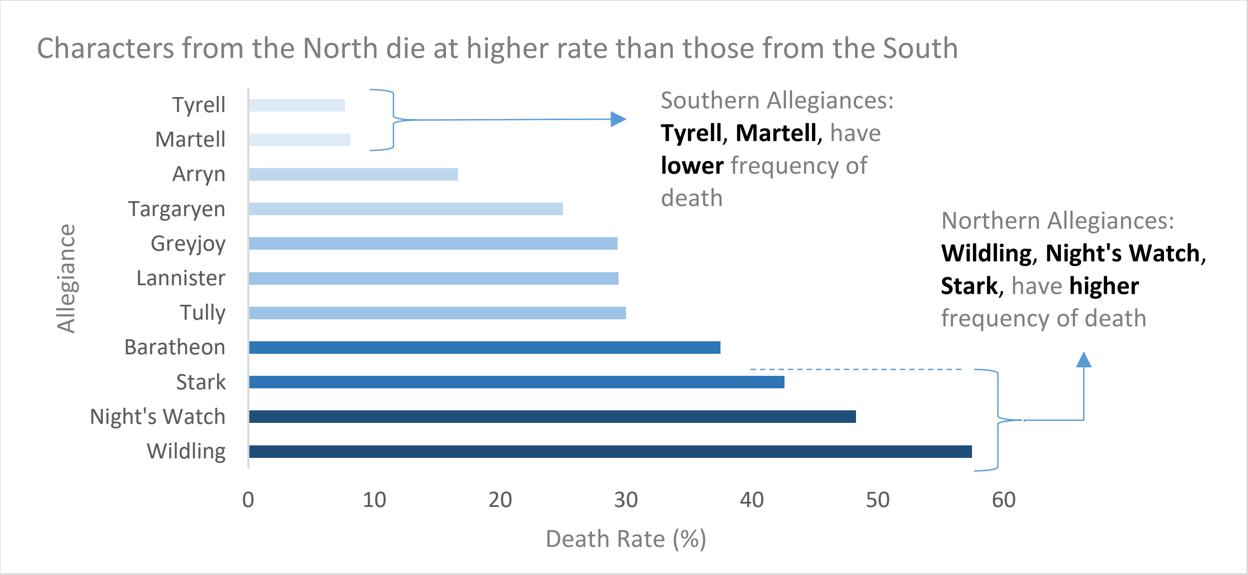


This graph is very similar to the previous one, as it shows win rates for different Kings. As with the previous graph, it shows that Kings had much more success when attacking. It also shows a fairly large difference between Stannis' win rate, compared to the three other Kings.

Character Data:



For this graph, I wanted to analyze death rates by both gender and nobility. Overall this graph shows that males died more often than females, and those who were nobles did not die as often as those who were not. Both of these conclusions make sense, as men were soldiers and more likely to die in battle, while nobles usually had guards and did not do their own fighting, so they were less likely to die.



This was my explanatory visualization for death rate among different groups/allegiances. When constructing this graph, I noticed that the groups with the highest death rates were all from the North, while the groups with the lowest death rates were from the South. I would not have discovered this trend if I did not have sufficient background information on the book series. This trend makes sense, as most of the fighting in the book series is done in the North, especially defending The Wall.

Predictive Models:

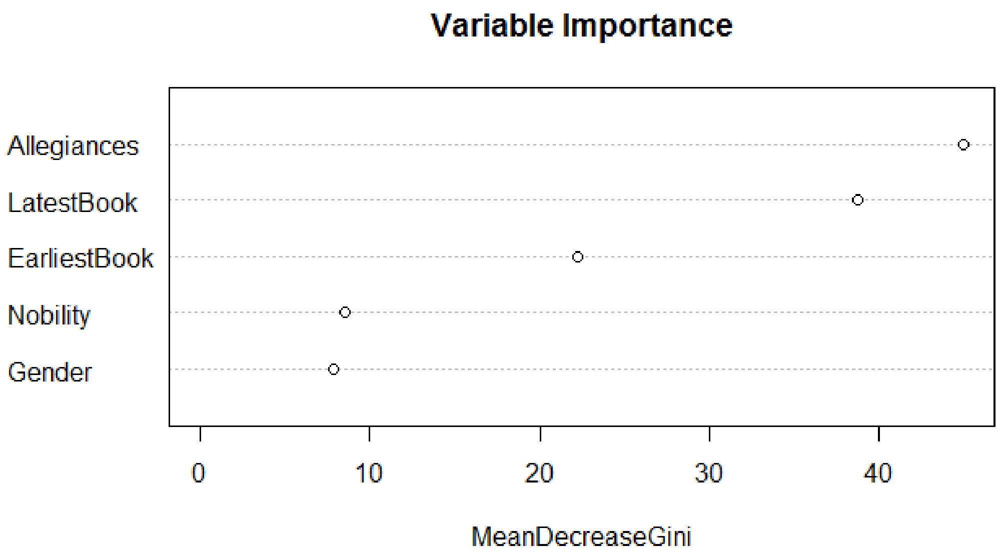
For this project, I created 2 different types of predictive models using R. I chose to create a Naïve Bayes model, using the 'e1071' package, and a Random Forest model, using the 'randomForest' package. We did not cover these predictive models in class, but I was able to do enough research of classification models to understand how they function. I chose both of these models because they are very useful in predicting the likelihood of a Boolean value, especially with a lot of categorical predictor variables. I also researched logistic regression, however it works better with more numerical or continuous variables, while Naïve Bayes and Random Forest work better with categorical variables.

For the battles data set, I created a Naïve Bayes inference model. The dependent variable that I tried to predict was the 'Attacker Outcome' column: basically, who would win this battle? The predictor variables that I used were the King, House, and army size of both the attackers and defenders, the battle type, the region, and whether or not the battle took place during summer. I chose not to create a Random Forest model because my battles data had a large number of missing values, and the Random Forest model cannot account for these. On the other hand, it is easy to simply omit missing values from the calculation of a Naïve Bayes model, without having to completely remove the data point.

I split my battles data into two subsets: a training data set with 30 entries and a test data set with 7 entries. I created these subsets by randomly selecting 30 of the 37 indices. I used the training data set to create the Naïve Bayes model, and then tested it with the test data set. Since the training set would change every time I ran the script, my accuracy was not always consistent. However, the Naïve Bayes model would most commonly correctly predict either 6 or 7 of the test data points. Given the extremely small data set size, I was fairly happy with the results. I think the accuracy of the model would be much better if I had a larger data set size.

For the characters data set, I created both a Naïve Bayes model and a Random Forest model. For both models, the dependent variable I tried to predict was the 'IsDead' variable: basically, will this character die? The predictor variables I used were allegiance, gender, nobility, and earliest and latest book appearance.

Like with the battles data set, I split the character data set into a training set of 700 rows, and a test set of 217 rows, determined through a random sample generator. I created both models using the training set, and then made predictions on the test data set. In general, the Naïve Bayes model had about a 66% accuracy, while the Random Forest model had about 75% accuracy. Neither of these are very high numbers, but since they are both moderately above 50%, I am satisfied. While I had plenty of data points in this data set, I think more information would be needed to correctly predict if a character would die. Some things, like the character's occupation (Knight, Squire, Sellsword, Cook, Page, etc), could heavily impact the character's likelihood of dying. For example, if a character was a knight or sellsword, they would be more likely to die than a cook. Below is a "variable importance" plot for the random forest model (created using the varImpPlot function). It shows that the 'Allegiance' and 'Latest Book' variables were more important than gender or nobility.



Potential Improvements:

After completing this project, there are definitely some things I would change, if I were to complete it again. The first thing would be to try to select a larger and more complete data set than the battles data set that I used. I found the domain of the data very interesting, but realistically the data set was too small and contained too many missing values to gather any impactful value. Another thing I would have wanted to improve would be how I handled the attackers and defenders in the battles data set. For every battle there was the potential to have up to 4 attackers and 4 defenders. I was not sure how to handle lists of values in a Naïve Bayes inference model, so that is why I reduced the attacker and defender fields to single values. This resulted in a loss of information that could have been avoided if I had done more research about Naïve Bayes models.