INTRODUCTION

The HBO series, an adaptation of George R. R. Martin’s bestselling books, has had an undeniable cultural impact. It has produced memes, merch, and (thousands of) articles.

If you’re among the very few who are not familiar, “Game of Thrones” is about a scramble for power in the fantasy world of Westeros (and Essos). At the center of this struggle is the Iron Throne and, with it, the right to rule the seven kingdoms.

The show features a famously long list of colorful characters, and it has a reputation for killing them off when you become attached and least expect it.

That’s why many of these data science projects attempt to predict when our favorite characters will bite the dust.

I have used various tools and packages of text mining to analyze this. Main steps of this project are as follows.

- Importing the data set.
- Refining the data with tidy text.
- Plotting the top characters with highest number of dialog.
- Counting the number of characters and finding out the top 10 characters.
- Plotting the percentage of the dialog spoken by top characters.
- Also we use many Python libraries likes: Pandas, Numpy, Matplotlib, Seaborn, Cufflinks, Sklearn, Counter, Scipy, and Plotly.

Tools:

1. Jupyter Notebooks, where all the code will live
2. Textstat, another Python package for analysing text
3. Textstat, another Python package for analysing text

The project ran the texts through Python with NLTK, and it used Seaborn for visualization and Networkx for network metrics and graphs.

As these projects demonstrate, data science isn’t limited to generating insights to help businesses increase their bottom line. Sometimes, it’s just wicked fun. If you are a data enthusiast or an aspiring data scientist, you can try such entertaining projects for fun and skill sharpening. And in this case, it’s certainly added a level of enjoyment and diversion to one of the biggest fandoms of our time.

DATA

This is not the sort of data model we’ve been featuring on the Vertabelo blog. Most of those are about running various types of enterprises or solving real-world business problems. This data model will focus on storing and modifying relationships and events in the story. So, let’s start with describing what this database will cover.

1. What should this data model cover?

In each series (movie, book, etc.) there are many plot changes. Characters are born, they change alliances, they travel, they fight other characters, they die, etc. To track all this action, we’ll need to store:

- Locations
- Events
- Characters
- Timeline
- All relationships between all these categories

Of course, we can’t store the whole series. But that is not the idea. I want to be able to store the backbone of the story in a structured format. That backbone should contain all we need to recreate the storyline from the start.

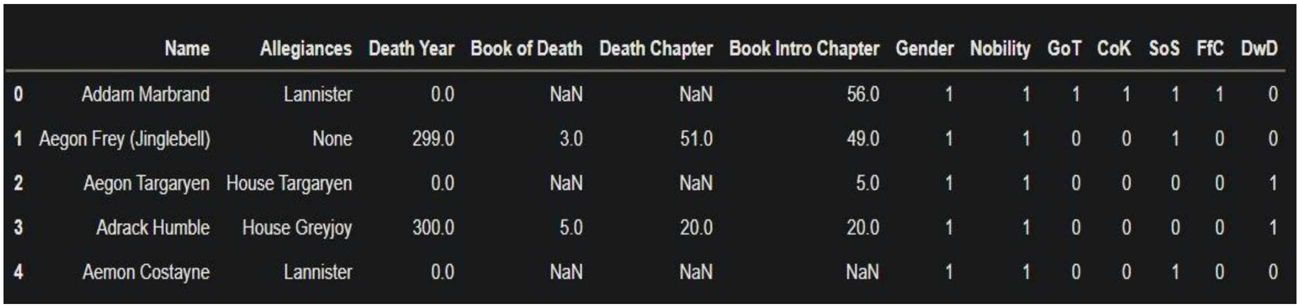
2. How is the data model limited?

Obviously, we won’t use this model for running a business or for project or process management. On the other hand, we could use something like it when storing relations between events, persons, and locations. After all, some elements of GoT are based on real-life events. And if we can store its plot, we should be able to store the “plot” for re-creating the sequence of real-life situations as well.

Marketers could use something like this to track what people do and buy with the idea of sending them personalized offers for products and services.

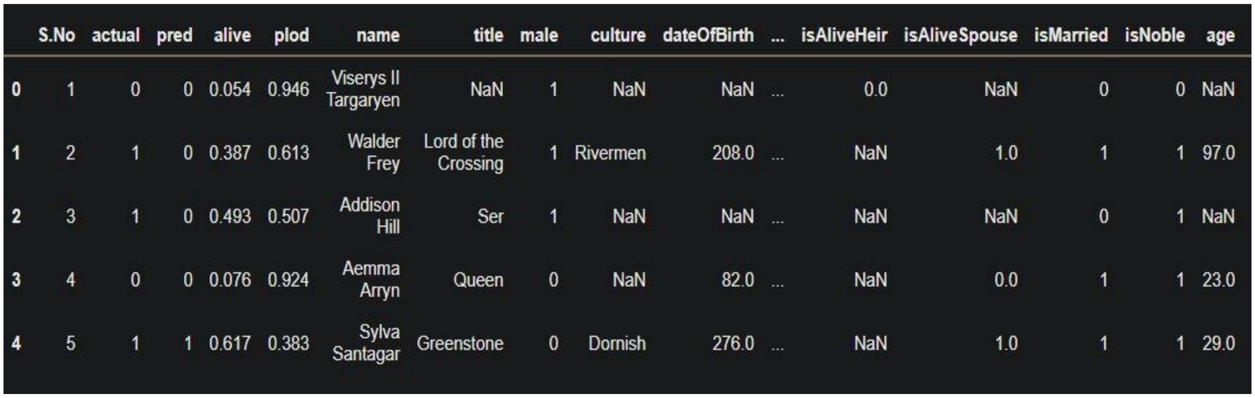
In these datasets we use head() function so only five raw are shown.

a) This csv contains information about all of the Character-deaths in game of thrones.



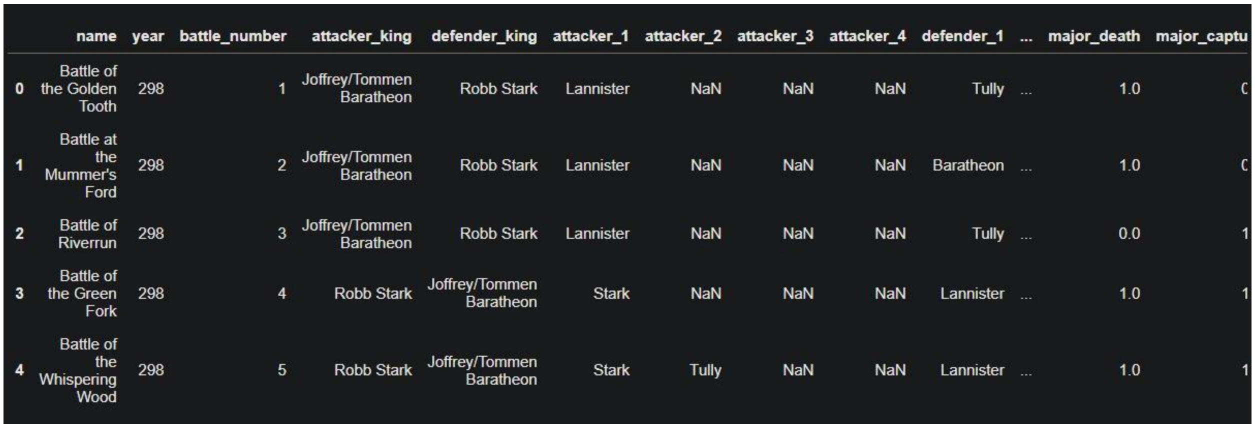
	Name	Allegiances	Death Year	Book of Death	Death Chapter	Book Intro Chapter	Gender	Nobility	GoT	CoK	SoS	FfC	DwD
0	Addam Marbrand	Lannister	0.0	NaN	NaN	56.0	1	1	1	1	1	1	0
1	Aegon Frey (Jinglebell)	None	299.0	3.0	51.0	49.0	1	1	0	0	1	0	0
2	Aegon Targaryen	House Targaryen	0.0	NaN	NaN	5.0	1	1	0	0	0	0	1
3	Adrack Humble	House Greyjoy	300.0	5.0	20.0	20.0	1	1	0	0	0	0	1
4	Aemon Costayne	Lannister	0.0	NaN	NaN	NaN	1	1	0	0	1	0	0

b) This csv contains information about all of the Character-predictions in game of thrones.



	S.No	actual	pred	alive	plod	name	title	male	culture	dateOfBirth	...	isAliveHeir	isAliveSpouse	isMarried	isNoble	age
0	1	0	0	0.054	0.946	Viserys II Targaryen	NaN	1	NaN	NaN	...	0.0	NaN	0	0	NaN
1	2	1	0	0.387	0.613	Walder Frey	Lord of the Crossing	1	Rivermen	208.0	...	NaN	1.0	1	1	97.0
2	3	1	0	0.493	0.507	Addison Hill	Ser	1	NaN	NaN	...	NaN	NaN	0	1	NaN
3	4	0	0	0.076	0.924	Aemma Arryn	Queen	0	NaN	82.0	...	NaN	0.0	1	1	23.0
4	5	1	1	0.617	0.383	Sylva Santagar	Greenstone	0	Dornish	276.0	...	NaN	1.0	1	1	29.0

c) This csv contains information about all of the Battles in game of thrones.



	name	year	battle_number	attacker_king	defender_king	attacker_1	attacker_2	attacker_3	attacker_4	defender_1	...	major_death	major_captu
0	Battle of the Golden Tooth	298	1	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NaN	NaN	NaN	Tully	...	1.0	C
1	Battle at the Mummer's Ford	298	2	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NaN	NaN	NaN	Baratheon	...	1.0	C
2	Battle of Riverrun	298	3	Joffrey/Tommen Baratheon	Robb Stark	Lannister	NaN	NaN	NaN	Tully	...	0.0	1
3	Battle of the Green Fork	298	4	Robb Stark	Joffrey/Tommen Baratheon	Stark	NaN	NaN	NaN	Lannister	...	1.0	1
4	Battle of the Whispering Wood	298	5	Robb Stark	Joffrey/Tommen Baratheon	Stark	Tully	NaN	NaN	Lannister	...	1.0	1

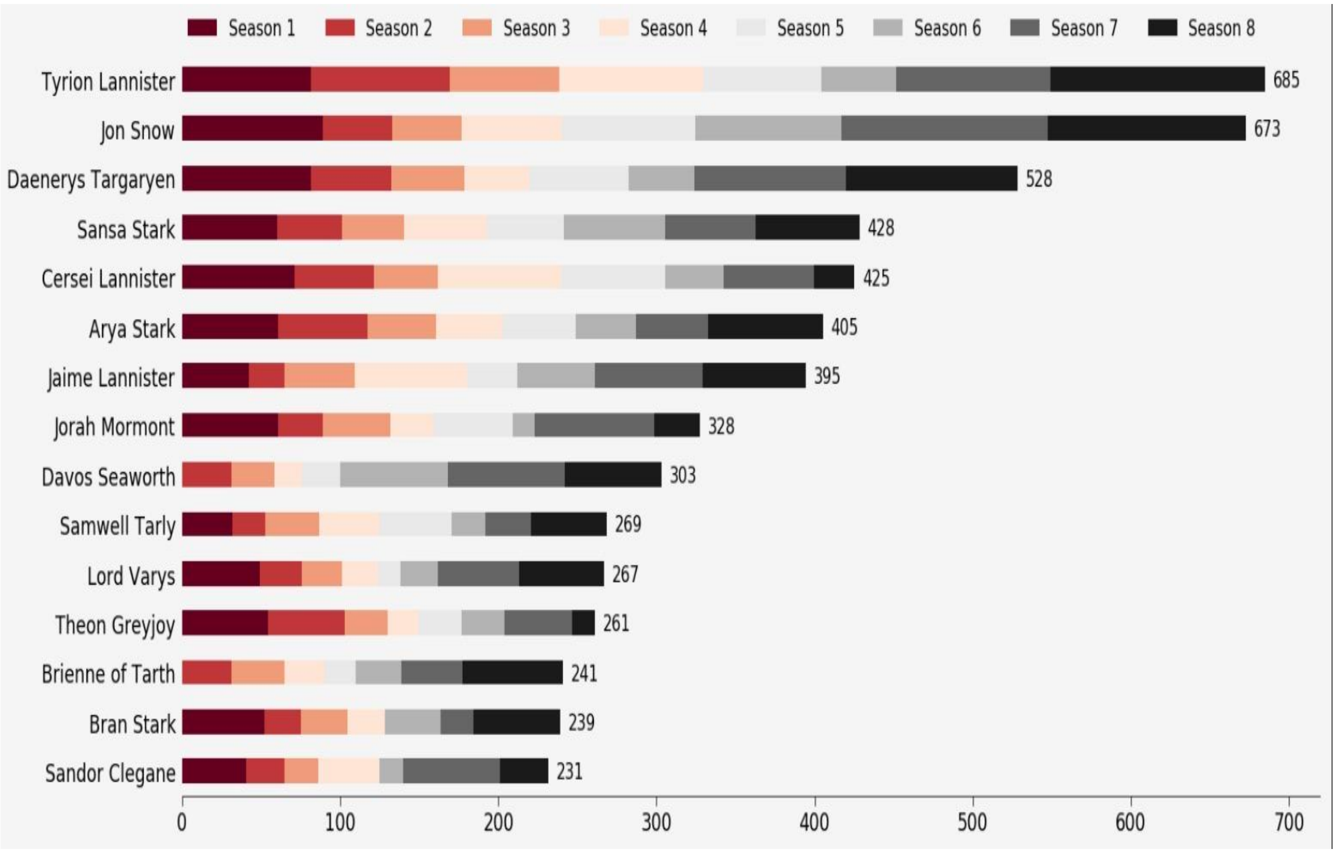
APPROACH

Without looking into the fire of the Lord of Light, I'll use good old bar charts to visualize the screen time of the characters and their houses. Here are the goals of this extremely scientific study:

- Finding out the screen time of the top 100 characters
- Screen time of the top 10 characters in each season
- Screen time of the nine primary houses in each season

Given below are the visualizations and results of the analysis:

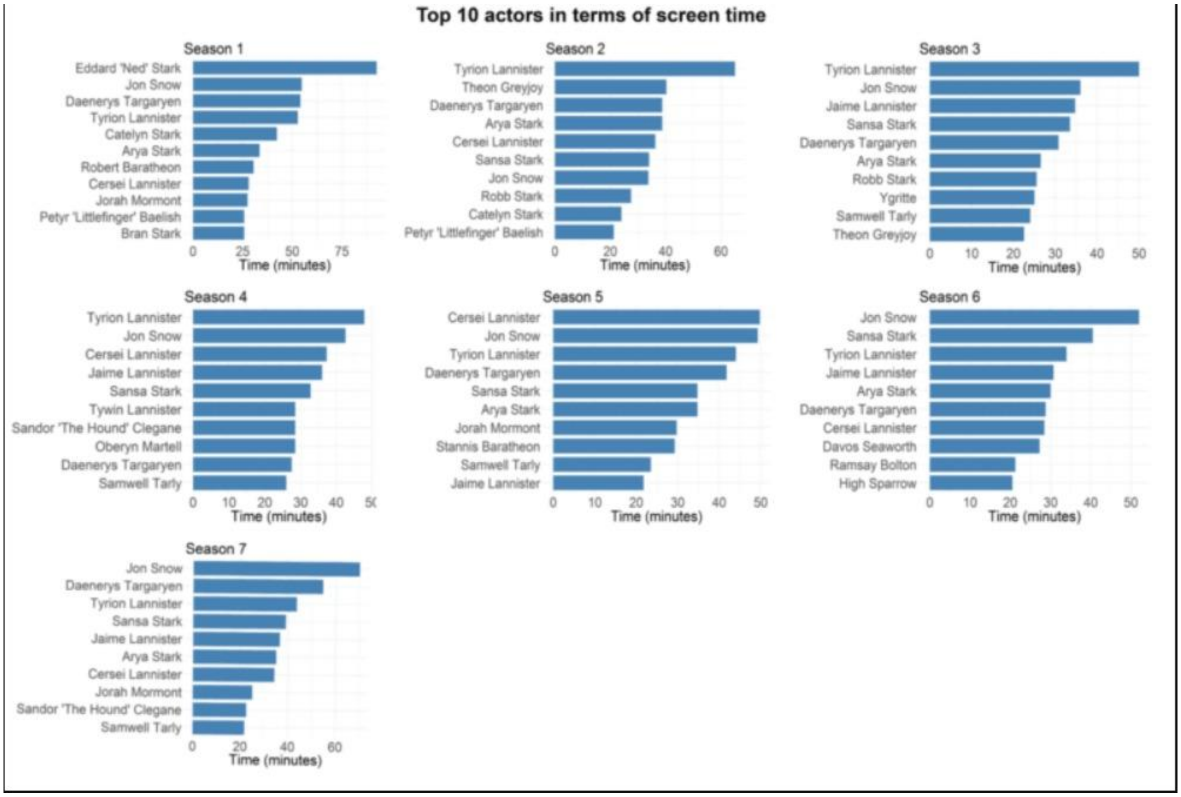
Screen time of the top main characters:



It looks like that Jon Snow and Tyrion Lannister are the two most important characters of the epic fantasy series, being almost tied in screen time at the top. After the delightful duo follows the titanic trio — Daenerys Targaryen, Sansa Stark and Cersei Lannister — with their impressive screen time, which shouldn't really come as a surprise. Awesome characters and great actresses. The average screen time of the top main characters stands at 7 minutes and 45 seconds, but our favorite five go way beyond that.

Top 10 characters in terms of screen time, by season:

Here you can see a visualization of the screen time of the top 10 character by season. Jon Snow certainly seems to be a strong contender for the main character of GoT as he’s always in first or second place — except for season 2 where he doesn’t even come close. Tyrion and, surprisingly, Theon Greyjoy are the biggest characters of that season.

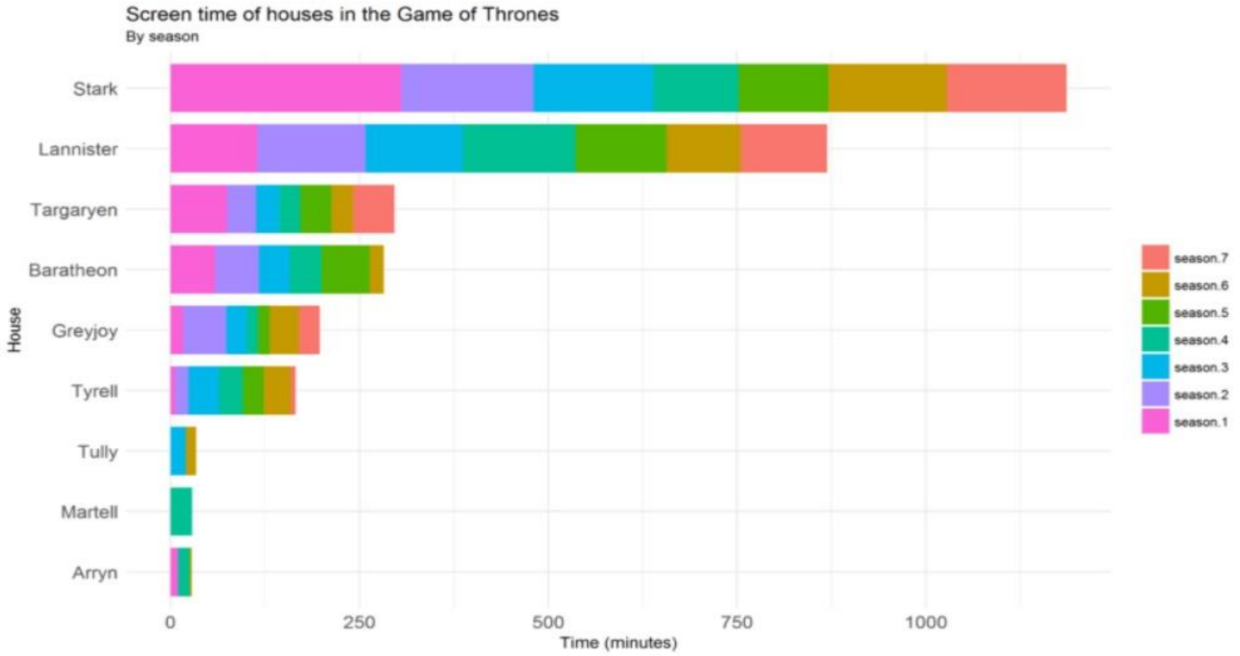


However, Jon’s character has only grown since then and has evolved to become one the most vital characters in the last two seasons. Another major character, Daenerys Targaryen (Mother of Dragons, Queen of blahblah, etc.), also started out slow as she didn’t even feature in the top 10 in season 4.

Screen time of the nine primary houses:

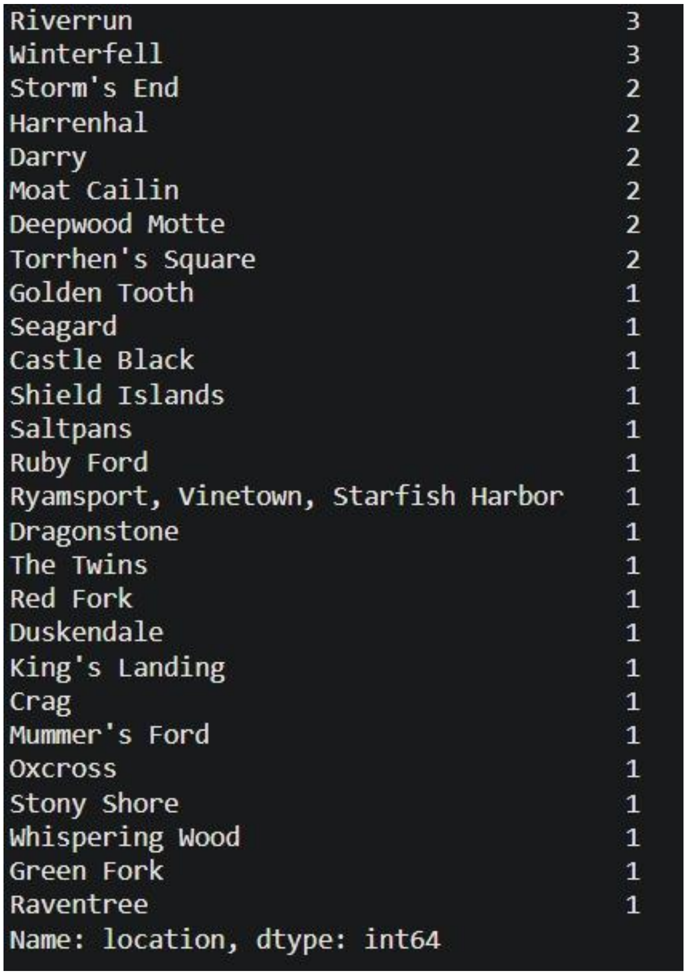
Here you can see the nine most important houses of Westeros. Unsurprisingly Stark and Lannister leave all the other houses in the dust. However, it should be noted that Jon Snow *SPOILER SPOILER SPOILER* was categorized as a Stark till the last season, where he's counted as a Targaryen.

This visualization shows that Starks, Lannisters and Targaryens have received the highest amount of screen time while Arryn and Martell have the least. The average screen time for these houses remains close to 12 minutes.



House of the Dragon is upon us, and it's packed with dynastic conflict, beards, armour and very silly wigs, just liked we hoped it would. It's also brought one important thing to mind: that we've forgotten where everywhere is in Westeros. King's Landing looks familiar, if a bit different, but where was Dragonstone again? What are the Stepstones? They seem to be important to those angry old guys and young women, so we'd better try to refresh our memories. That's where this handy series of maps comes in. Now you can remind yourself where Dragonstone is, relative to, say, King's Landing. You can see where the Iron Islands are! If you want to look up Winterfell, here it is: actual maps of Westeros and Essos, on which we've marked all the major locations of Game of Thrones and House of the Dragon. We've left off the continent of Sothoryos off because nothing happens there and no-one's really sure how big it is. (And you can forget the mysterious fourth continent, Ulthos, because till now you'd never even heard of it.)

Now we will calculate the count value of the locations where bottles have occurred……….



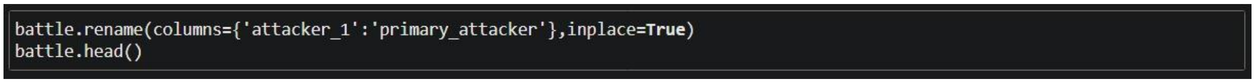
```
Riverrun                                3
Winterfell                              3
Storm's End                             2
Harrenhal                               2
Darry                                   2
Moat Cailin                             2
Deepwood Motte                          2
Torrhen's Square                        2
Golden Tooth                            1
Seagard                                 1
Castle Black                            1
Shield Islands                          1
Saltpans                                1
Ruby Ford                               1
Ryamsport, Vinetown, Starfish Harbor    1
Dragonstone                             1
The Twins                               1
Red Fork                                1
Duskendale                              1
King's Landing                          1
Crag                                    1
Mummer's Ford                           1
Oxcross                                 1
Stony Shore                             1
Whispering Wood                         1
Green Fork                              1
Raventree                               1
Name: location, dtype: int64
```

METHODS

➢ Data Pre-processing Steps

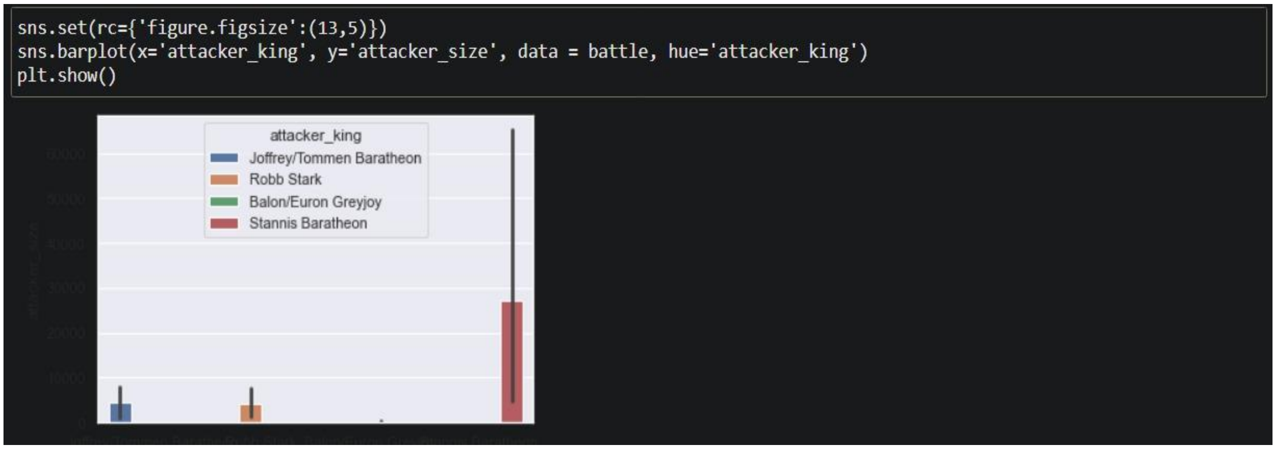
Step 1: Feature Engineering

Features considered- Title and Test (Title feature used for handling missing text feature values).



```
battle.rename(columns={'attacker_1':'primary_attacker'},inplace=True)
battle.head()
```

Step 2: Data distribution based on output class



```
sns.set(rc={'figure.figsize':(13,5)})
sns.barplot(x='attacker_king', y='attacker_size', data = battle, hue='attacker_king')
plt.show()
```

Step 3:

a. **Stemming:** Stemming is a method of text standardization(or word standardization) in the Natural Language Processing area that is used to prepare text, words and documents for further processing Stemming is a method of reducing word inflexion to its root forms such as mapping a group of words to the same stem even though the stem itself is not a valid word in the language. We use Snowball stemmer for this purpose. This algorithm is also known as thePorter2 stemming algorithm.

b. **Stop words:** Stop Words are words used to be used in search queries that do not contain important meaning. These words are typically filtered out of search queries because they return a vast quantity of unnecessary information.

c. Removing numerical and special characters and converting the words to lower case.

Step 4:

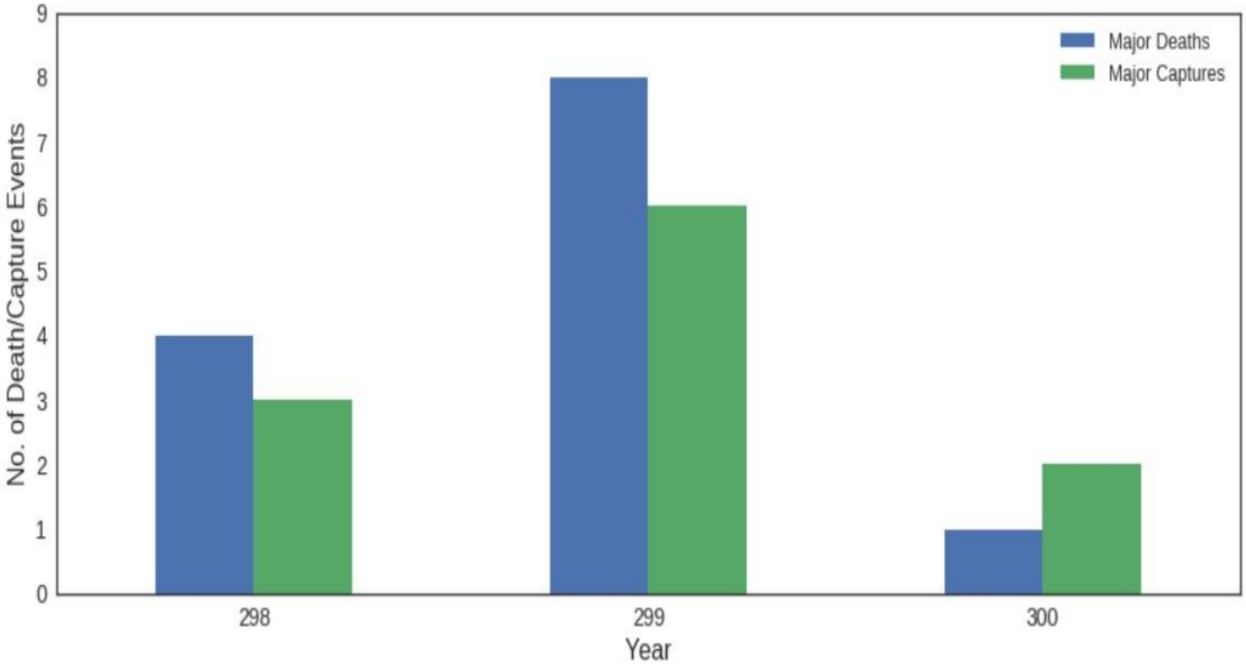
CountVectorizer: Convert a collection of text documents to a matrix of token counts Tf idf Transform: Transform a count matrix to a normalized tf or tf-idf representationTF-IDF: Vector representation of Text. TF-IDF is an abbreviation for Term Frequency-Inverse Document Frequency and is a very common algorithm to transform text into a meaningful representation of numbers.

Step 5:

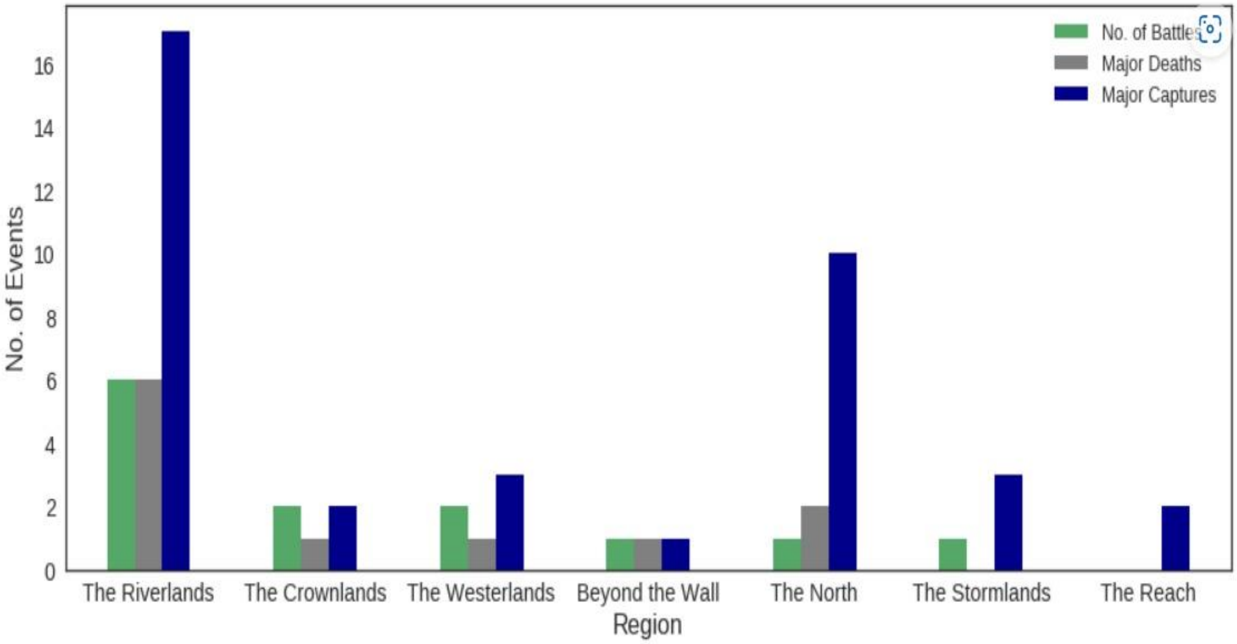
Visualizations: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In this project we will create visuals, multiple plots, and charts with the help of Python. So now we'll see some visuals.

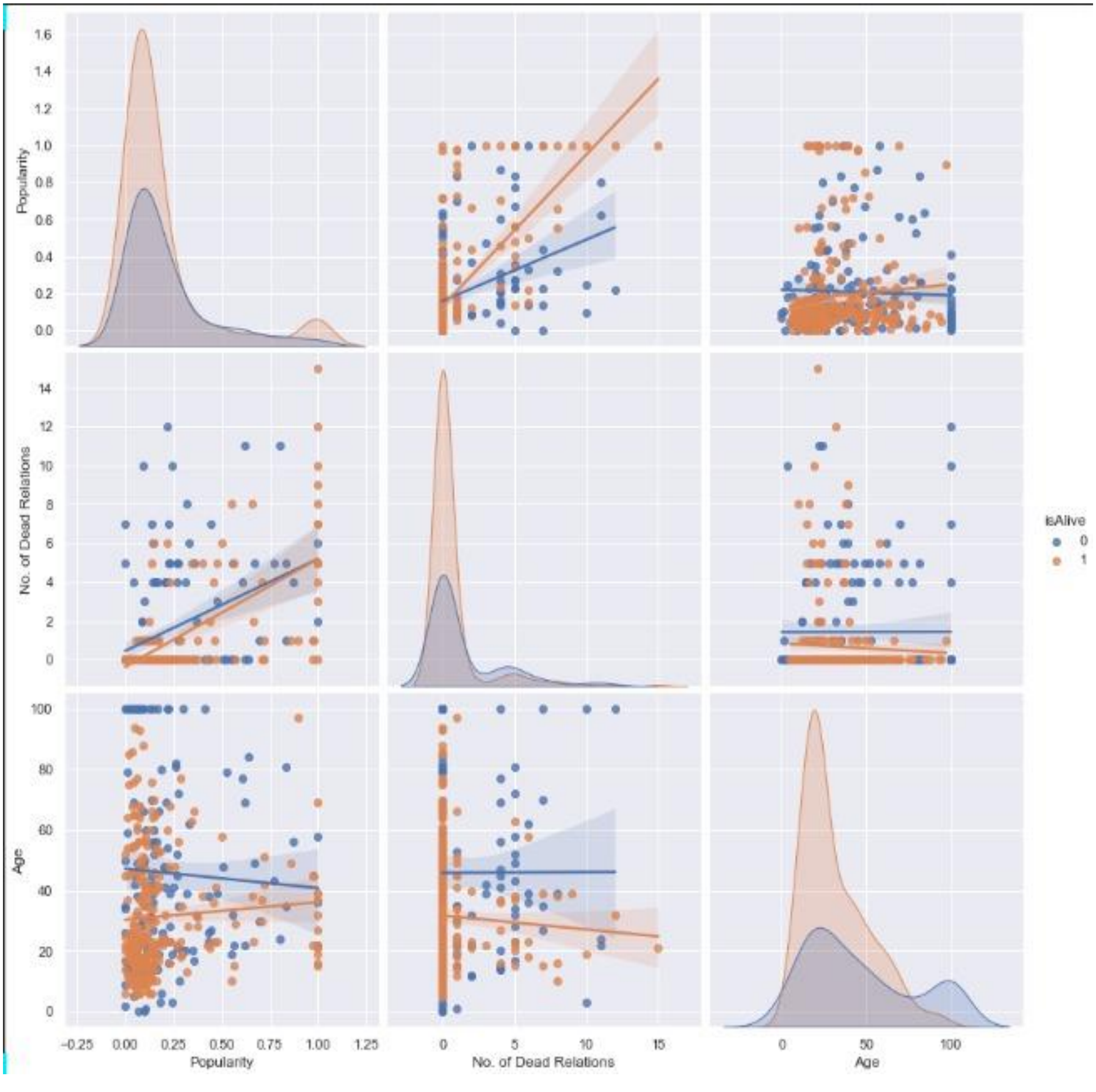
- **Major death/capture events by year –**



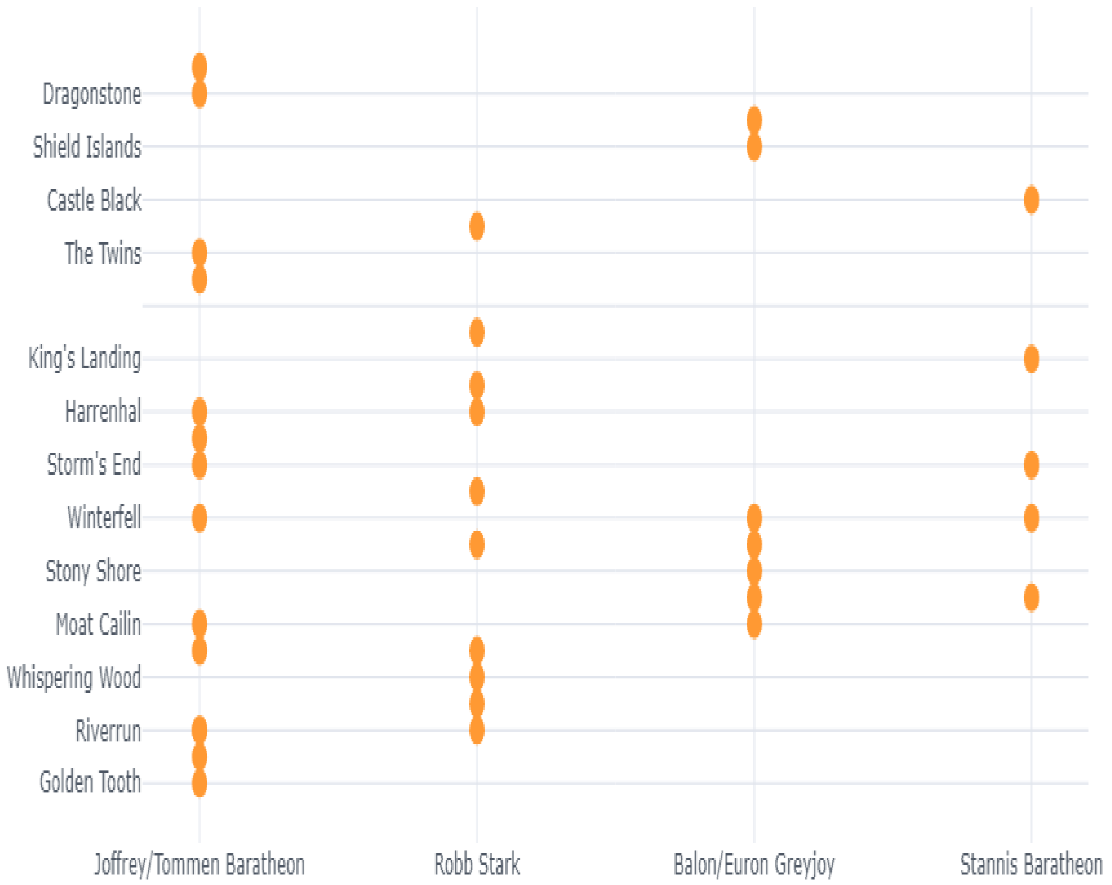
- **How many major death/capture events occur in each region?**



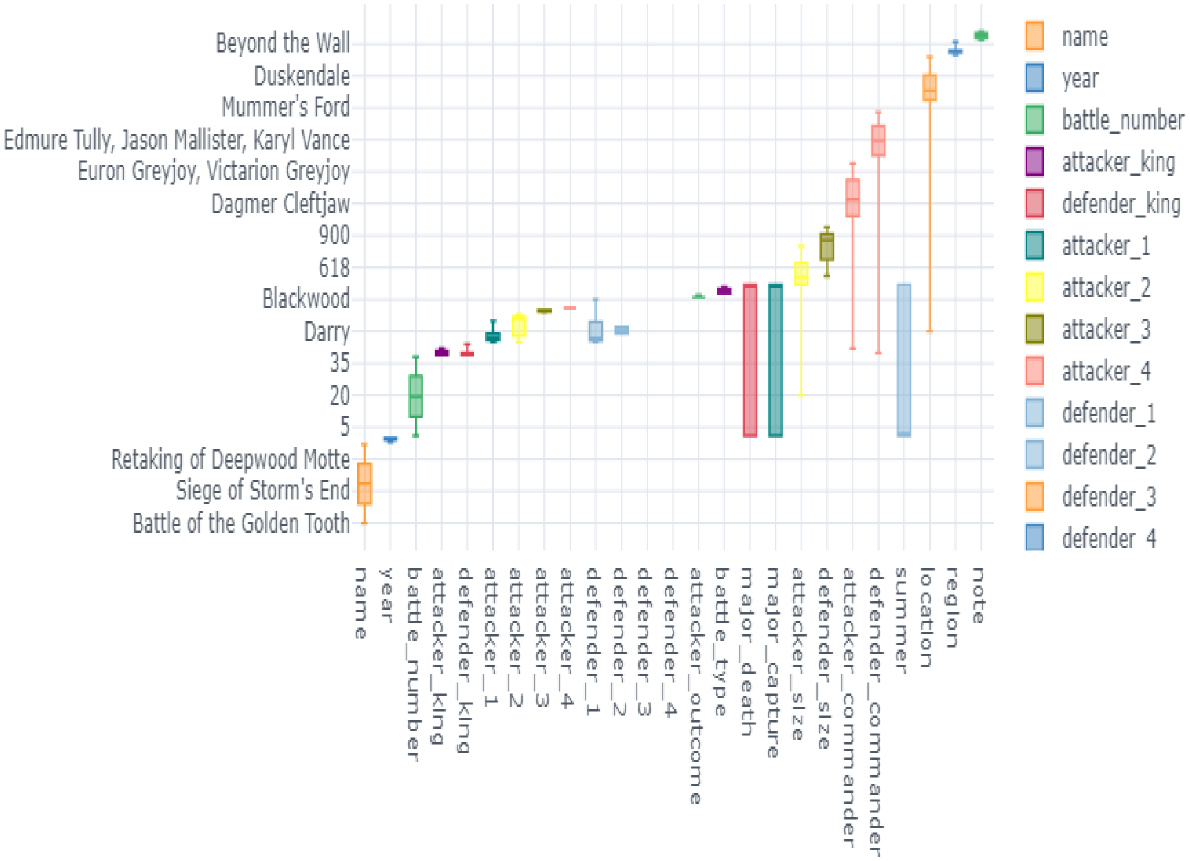
- **What is the relationship between popularity, number of dead relations, and age?**



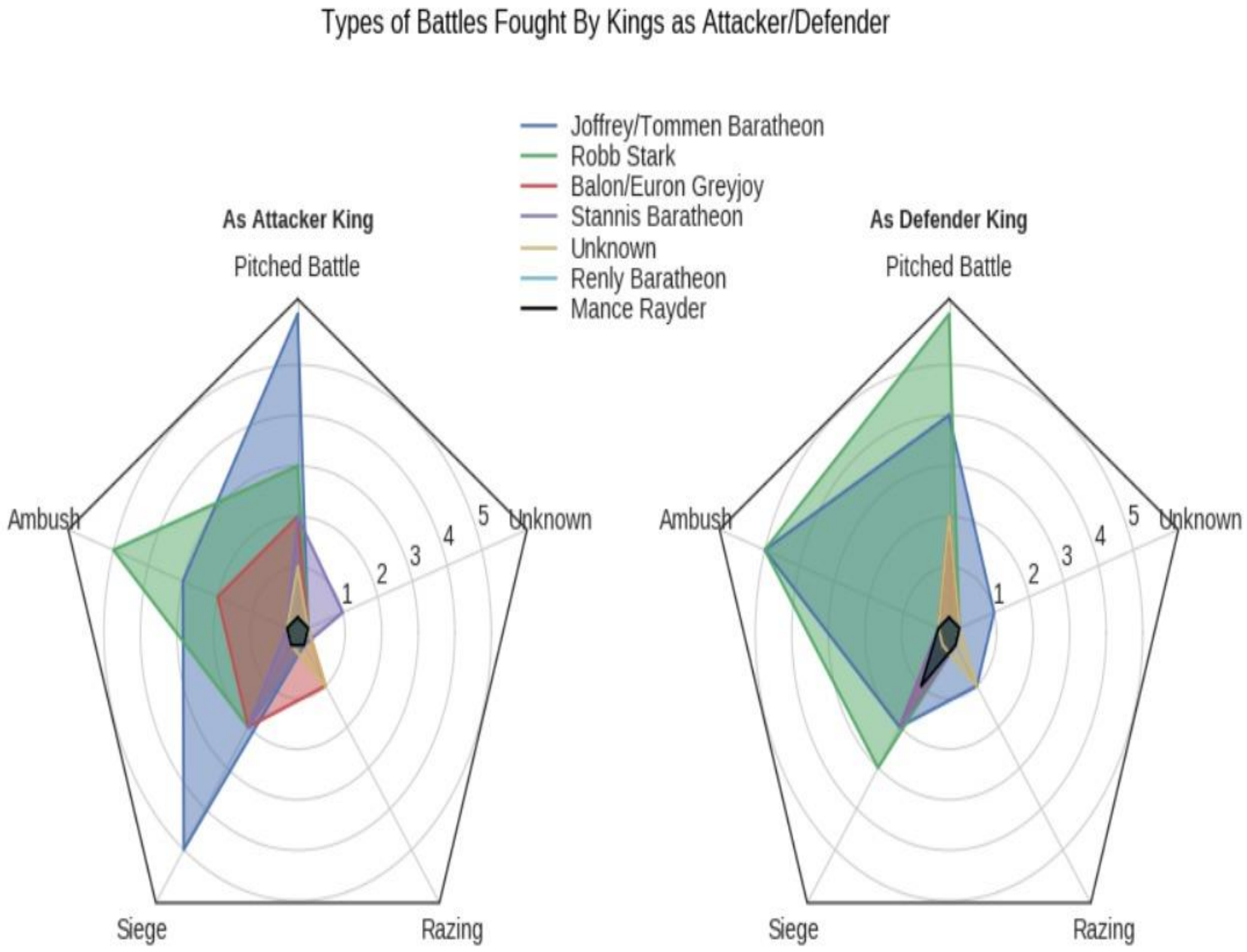
- **Which battle took place at which location?**



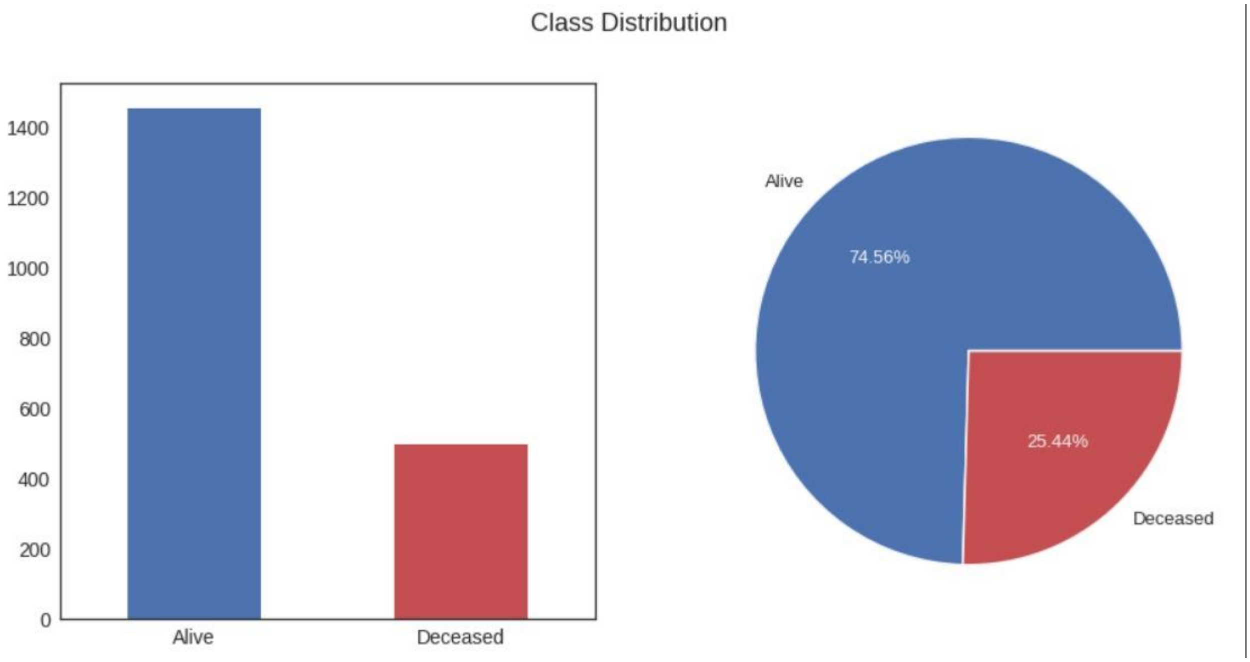
- **How many defender kings died during the battle.**



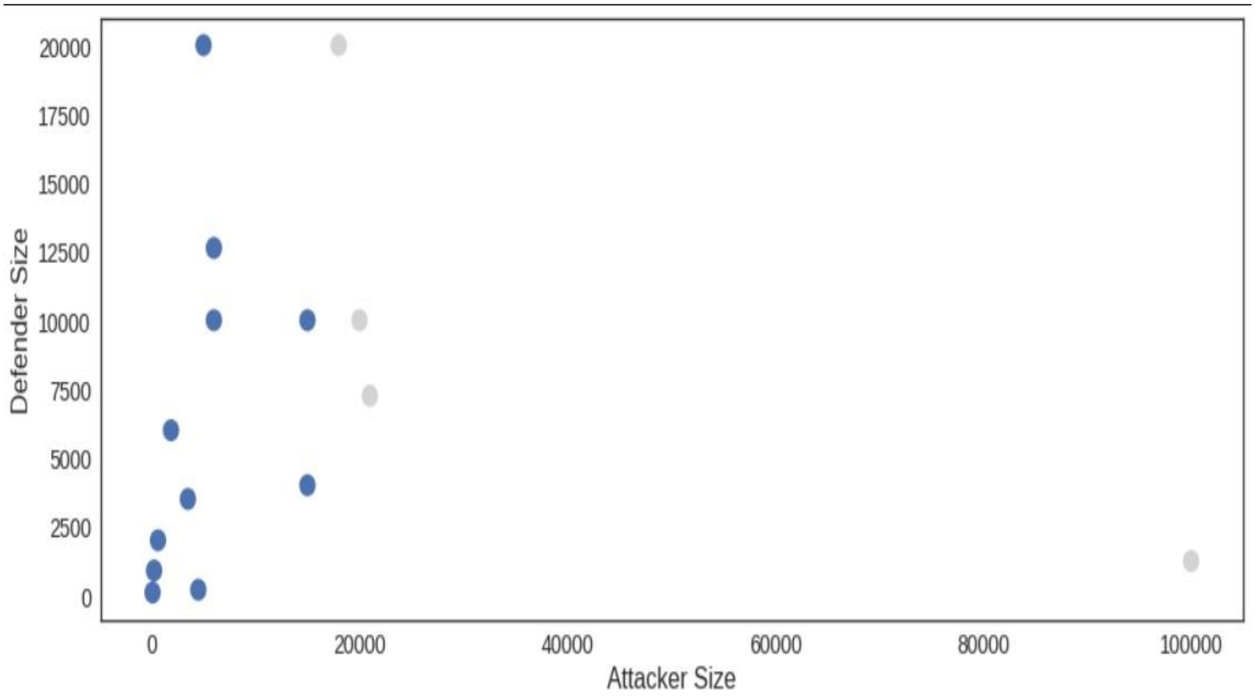
- **How often did kings fight different types of battles?**



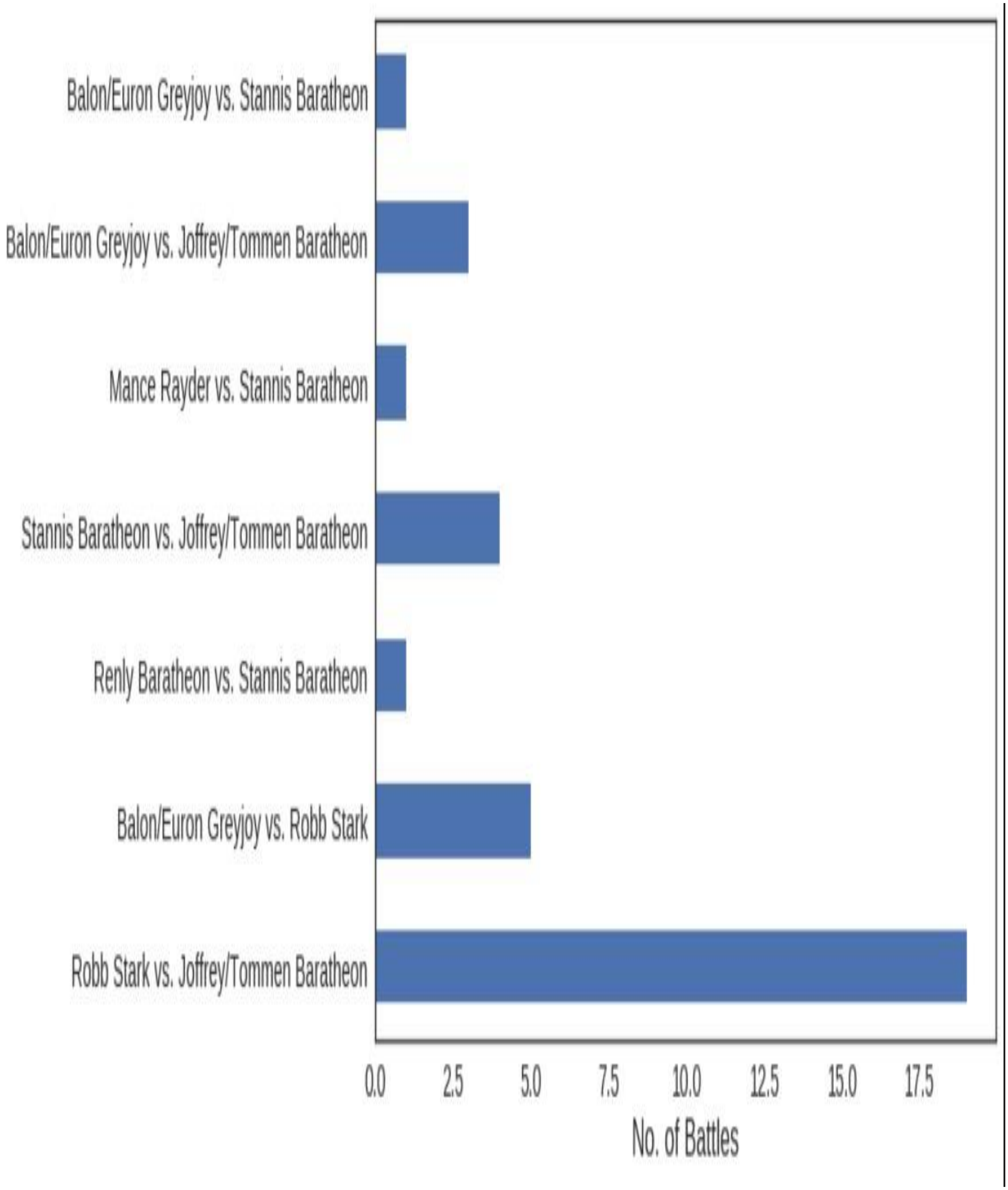
- **Is there a class imbalance**



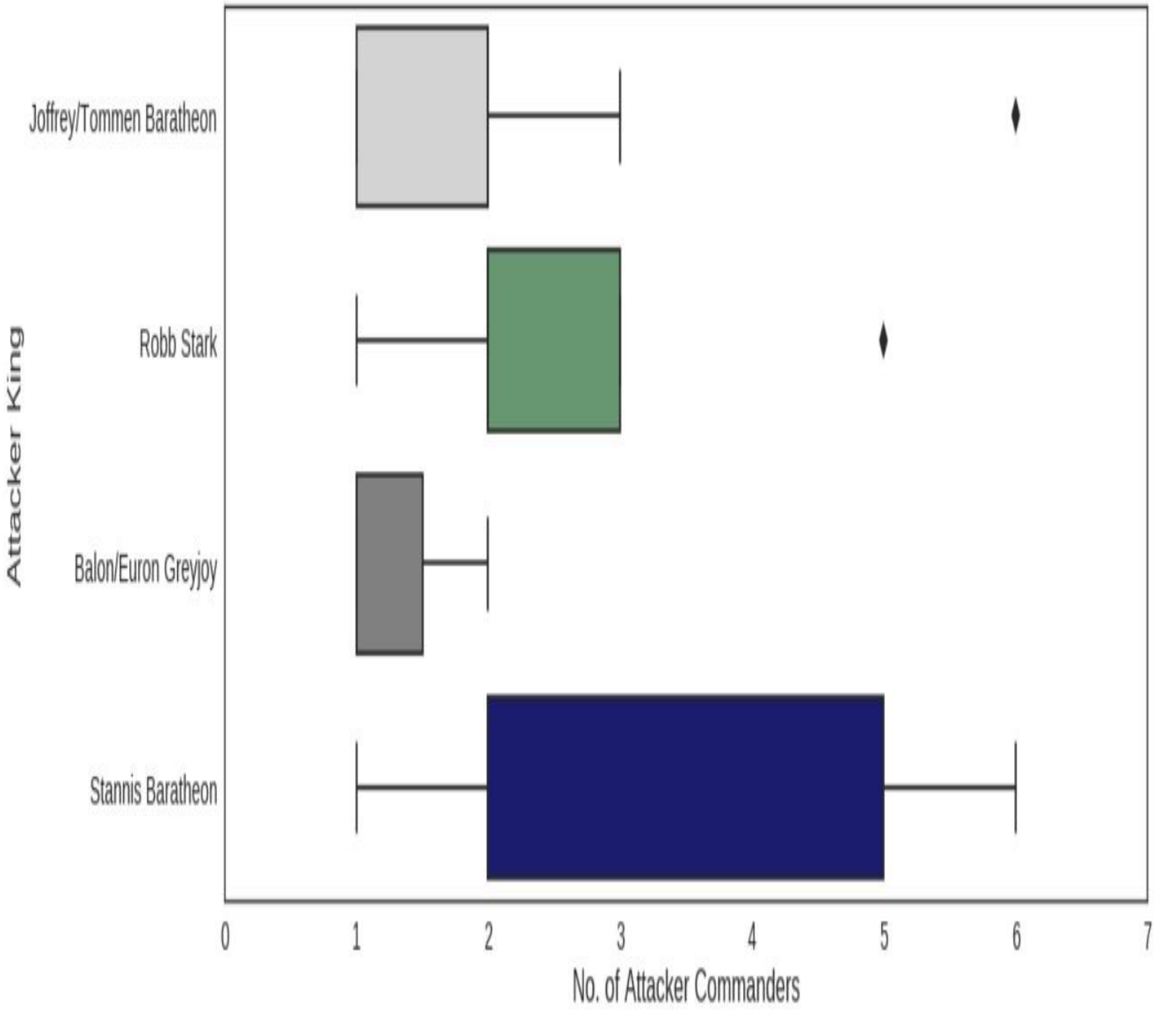
- **Impact of army size on outcome**



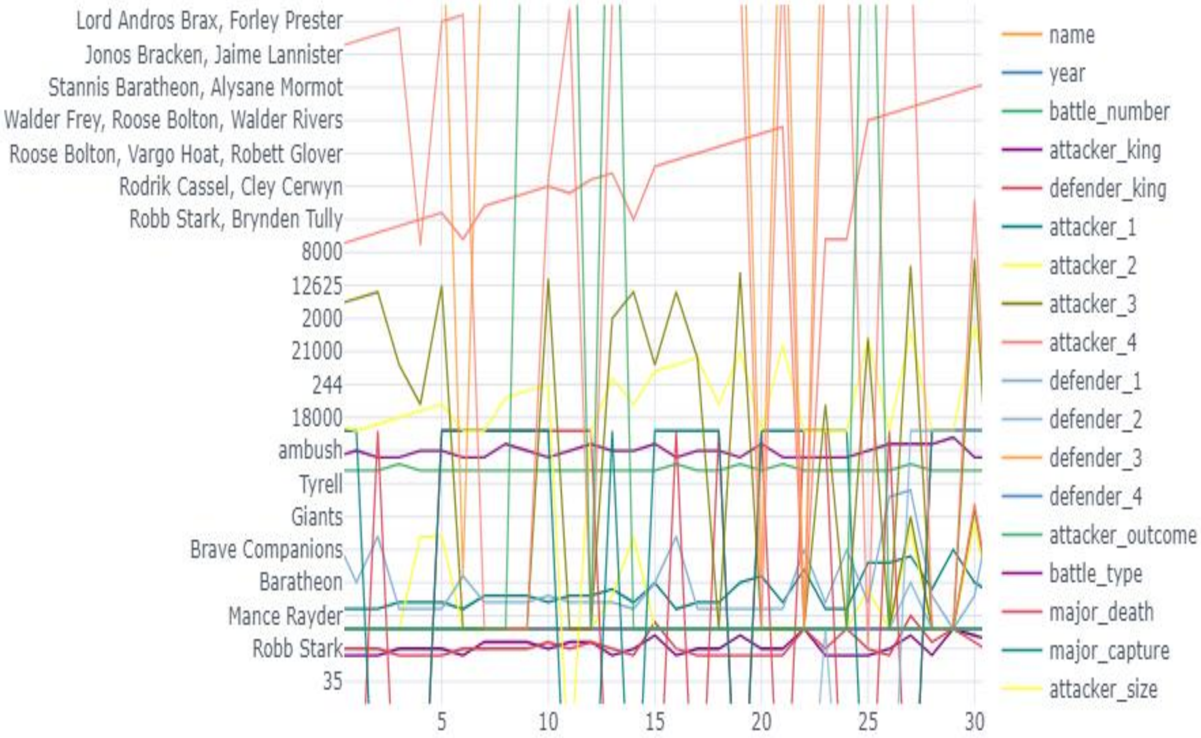
- **Which pairs fought the most battles?**



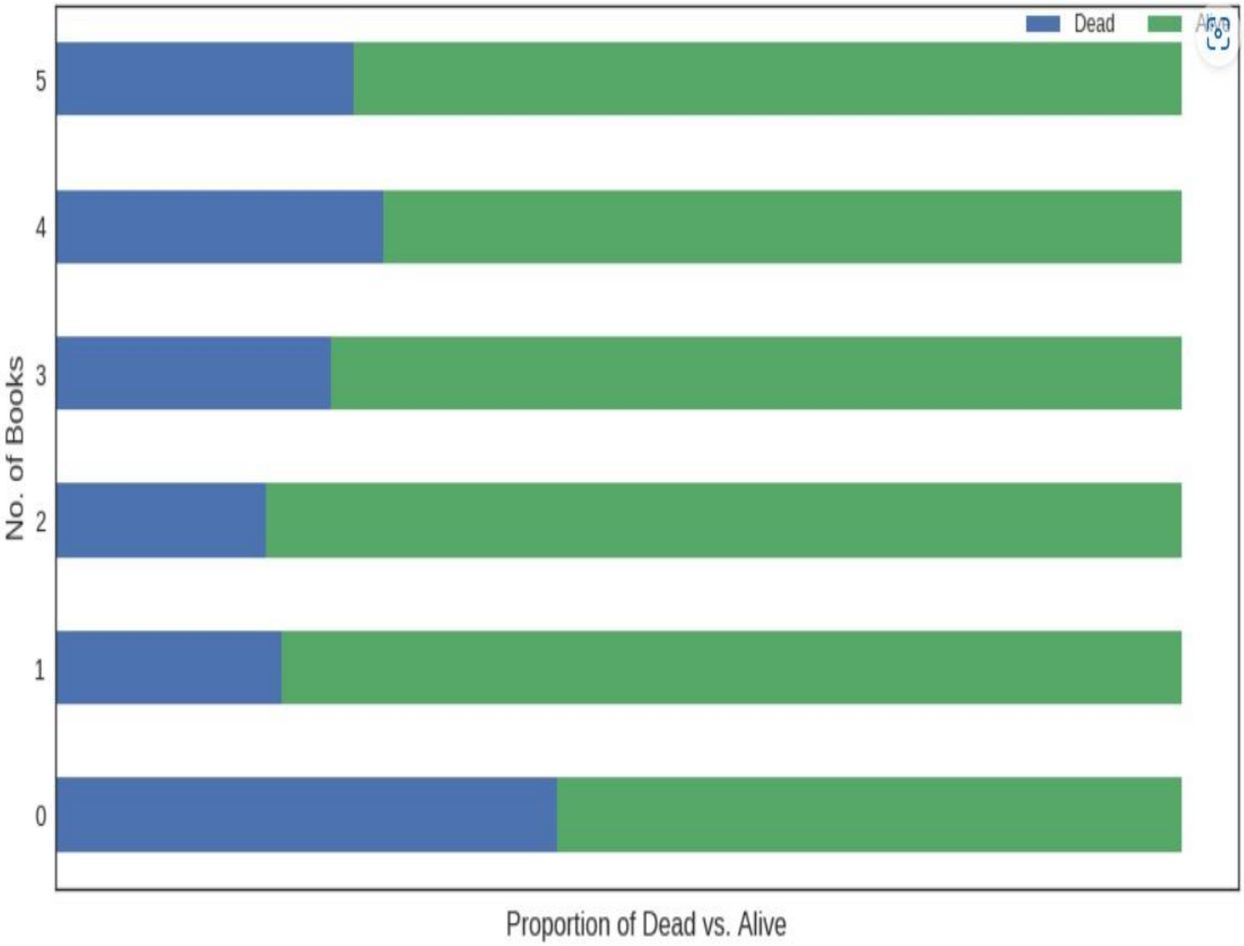
- **How many commanders did armies of different kings have?**



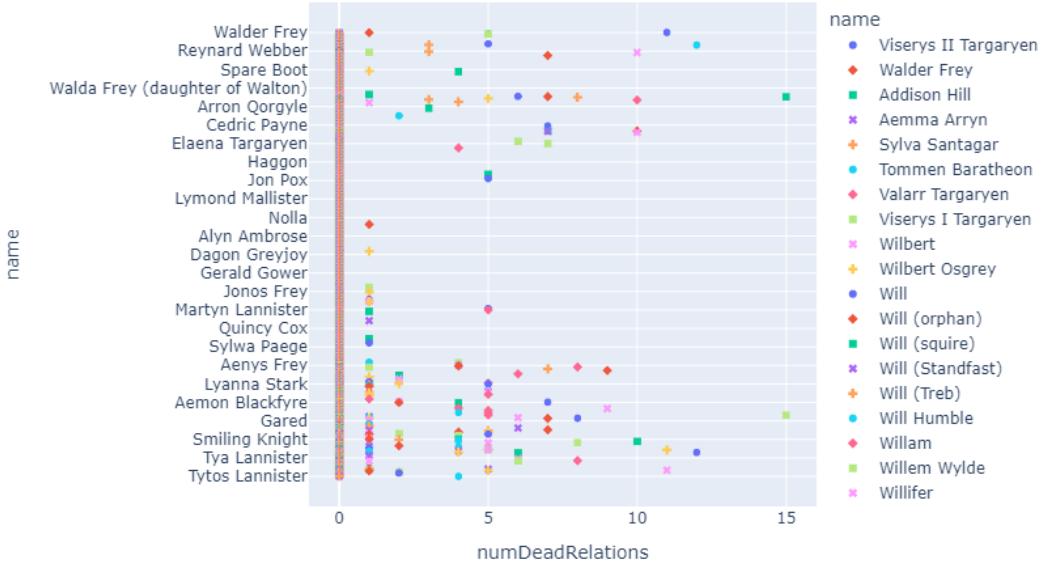
- **All battle data analysis**



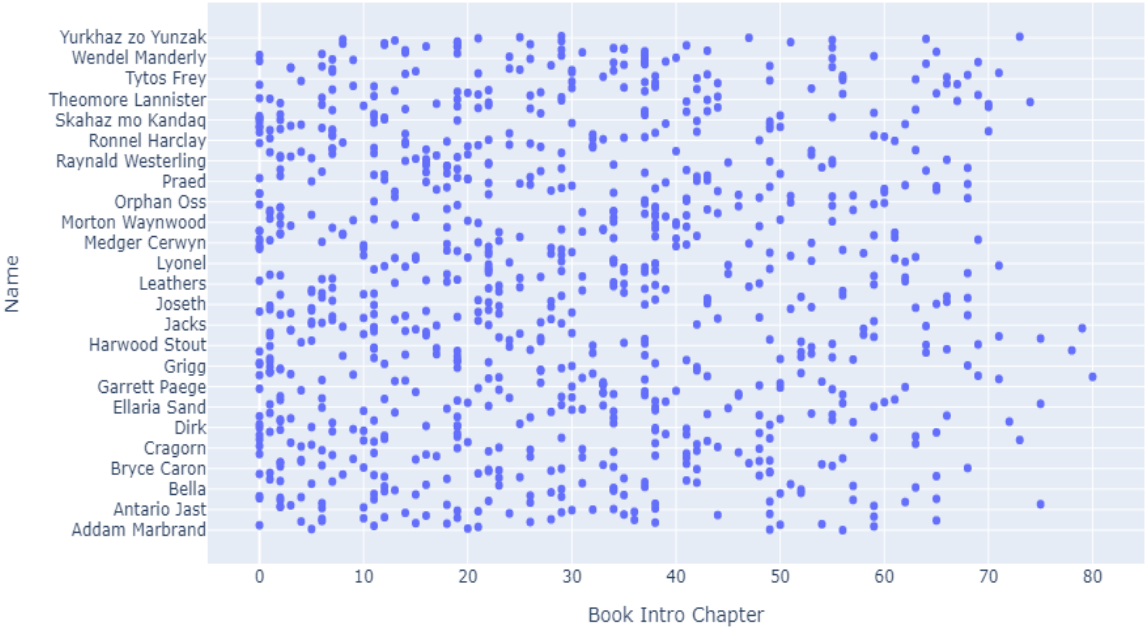
- **How does appearing in more books relate to survival?**



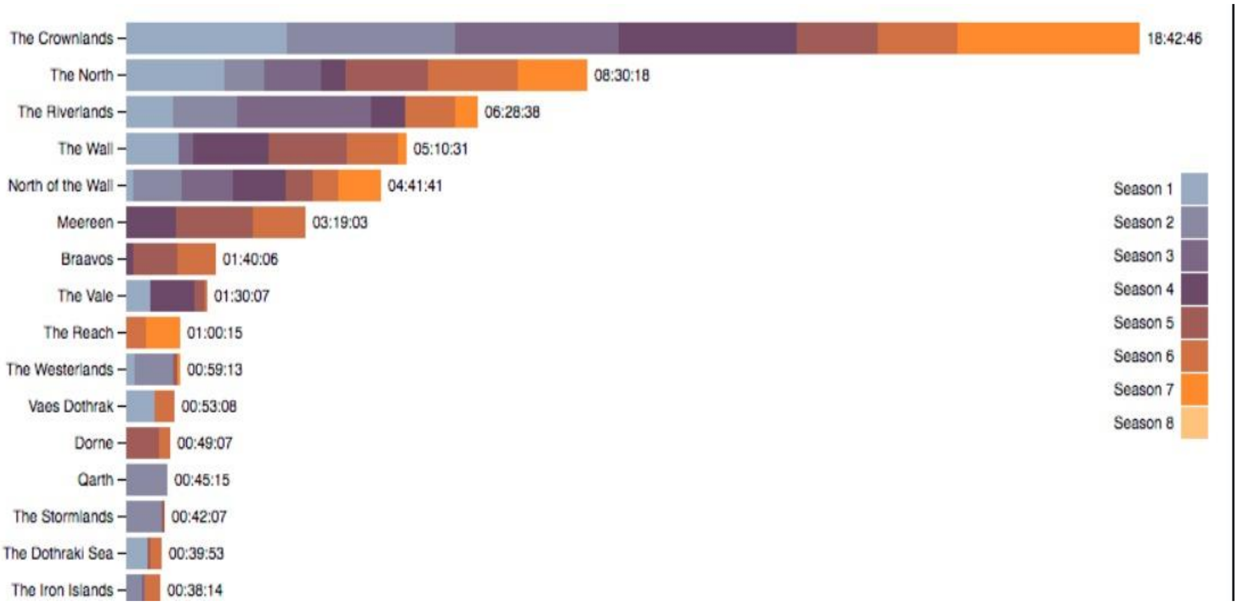
- **Number of death relation has to be analyzed by name.**



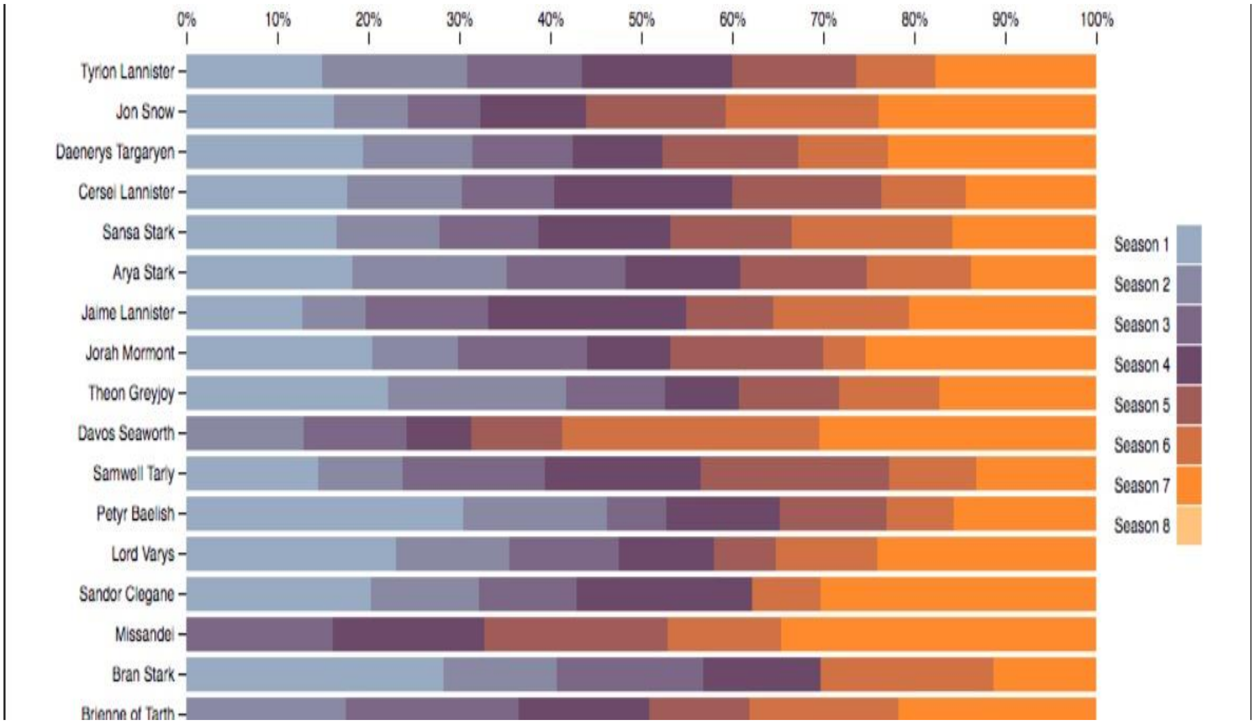
- **whose name appears in the book how many times**



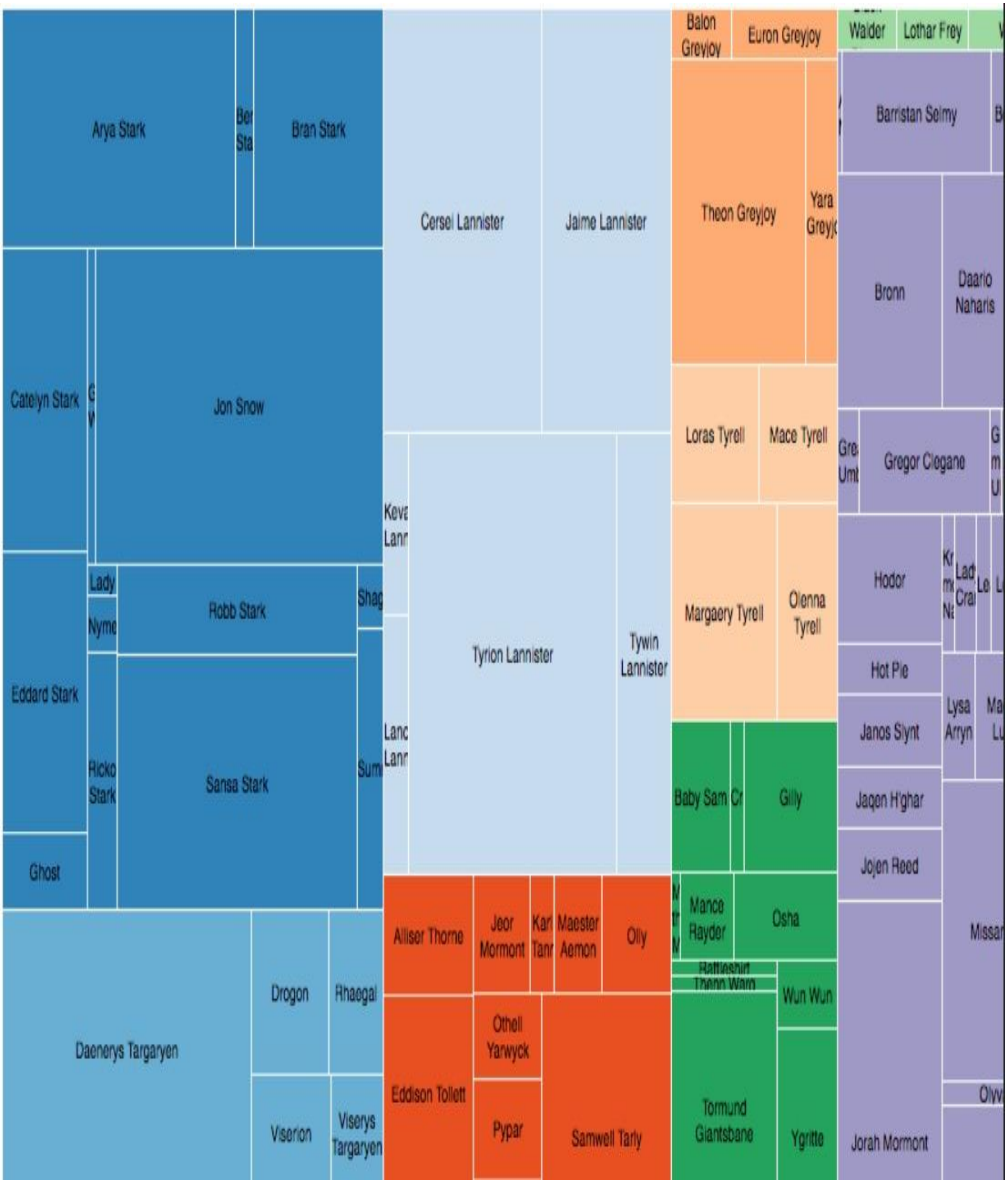
- **How long is spent in each location?**



- **What percentage of characters' time is in each season?**

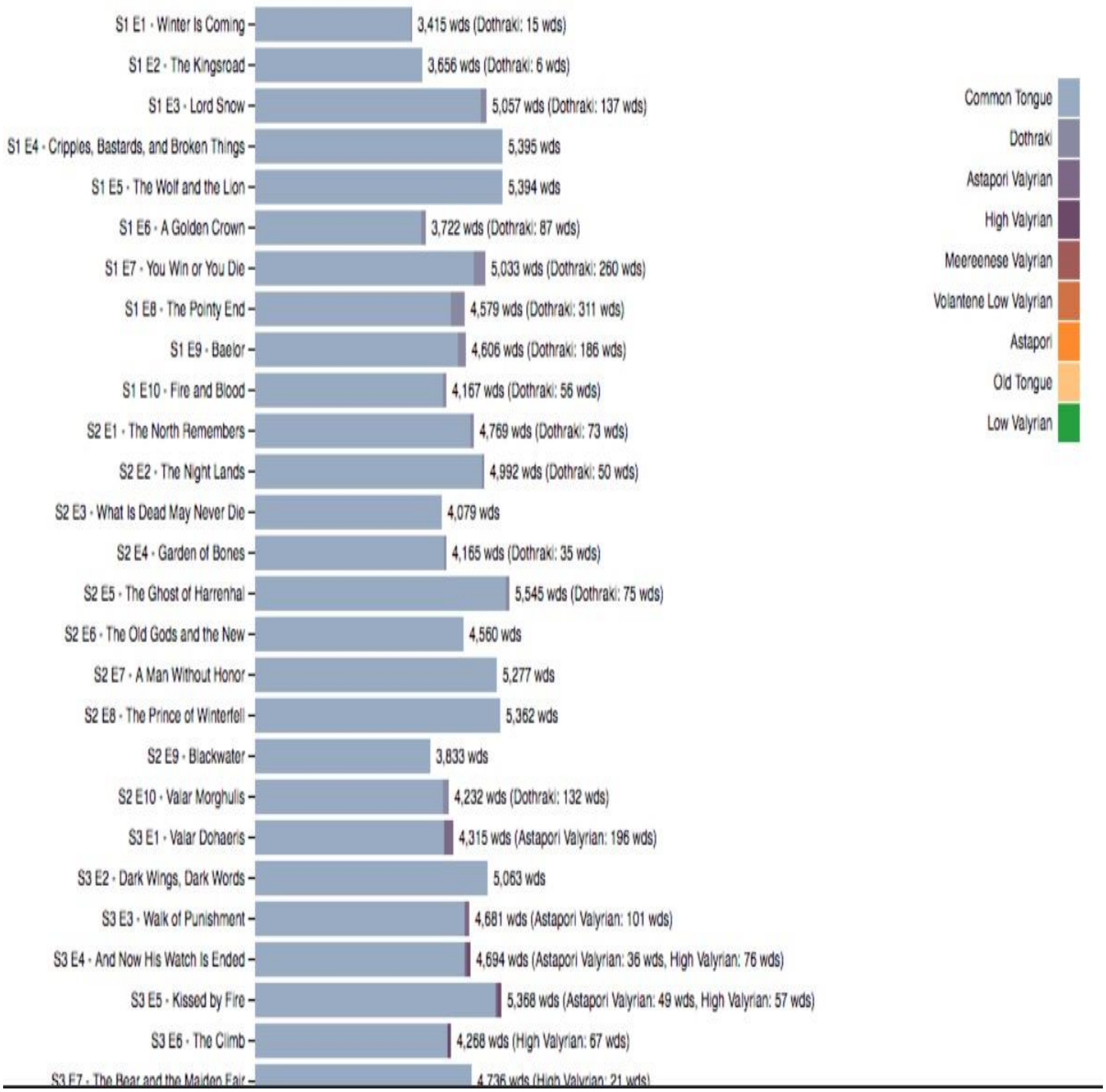


- **How long are characters on screen relative to all time?**

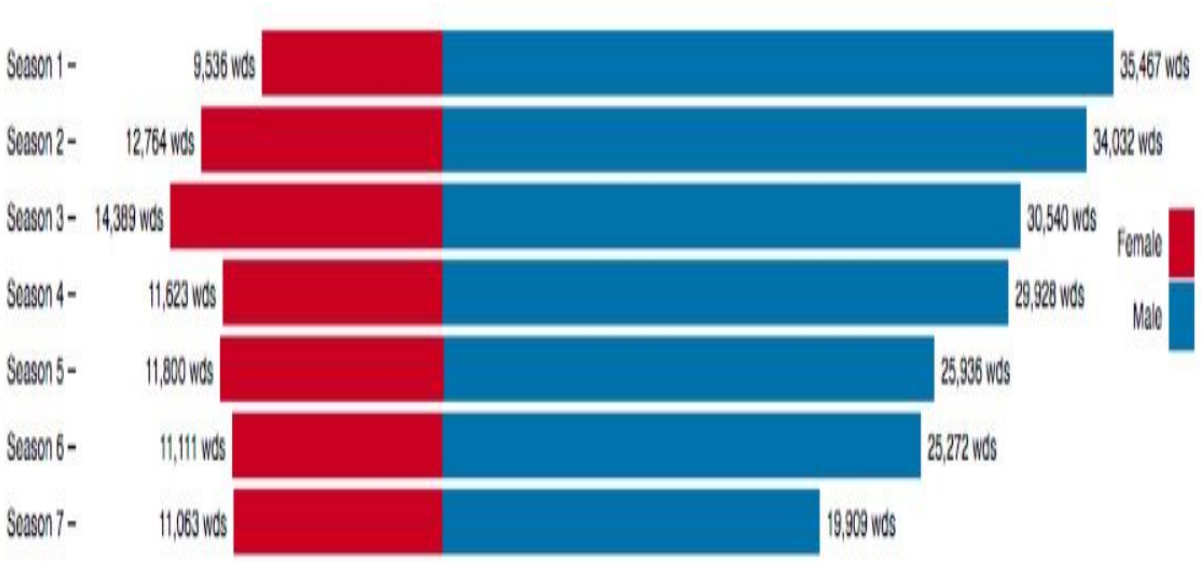


23

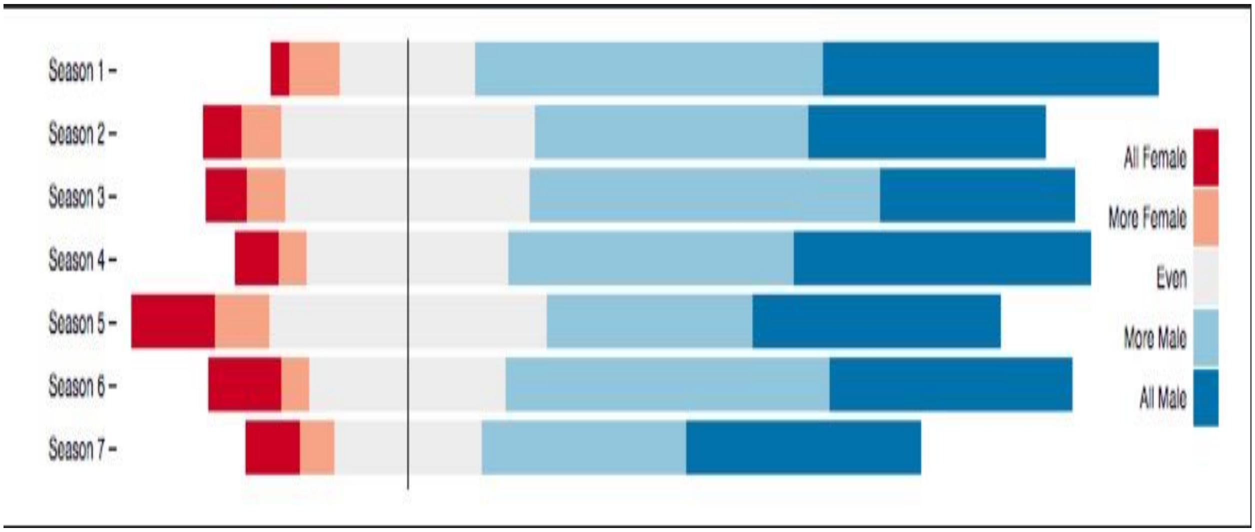
- **What languages are spoken in each episode?**



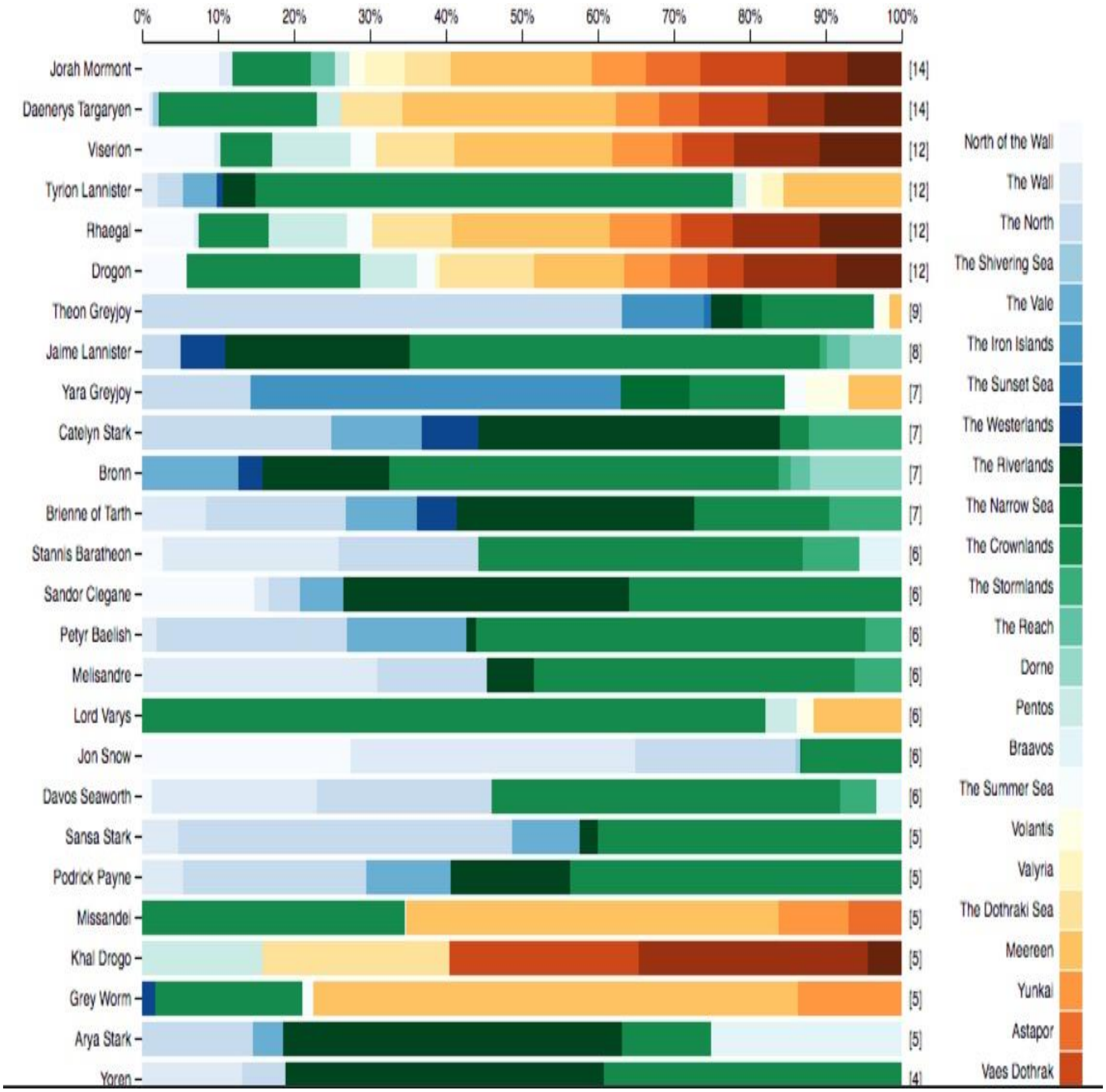
- **What's the gender balance of words spoken per season?**



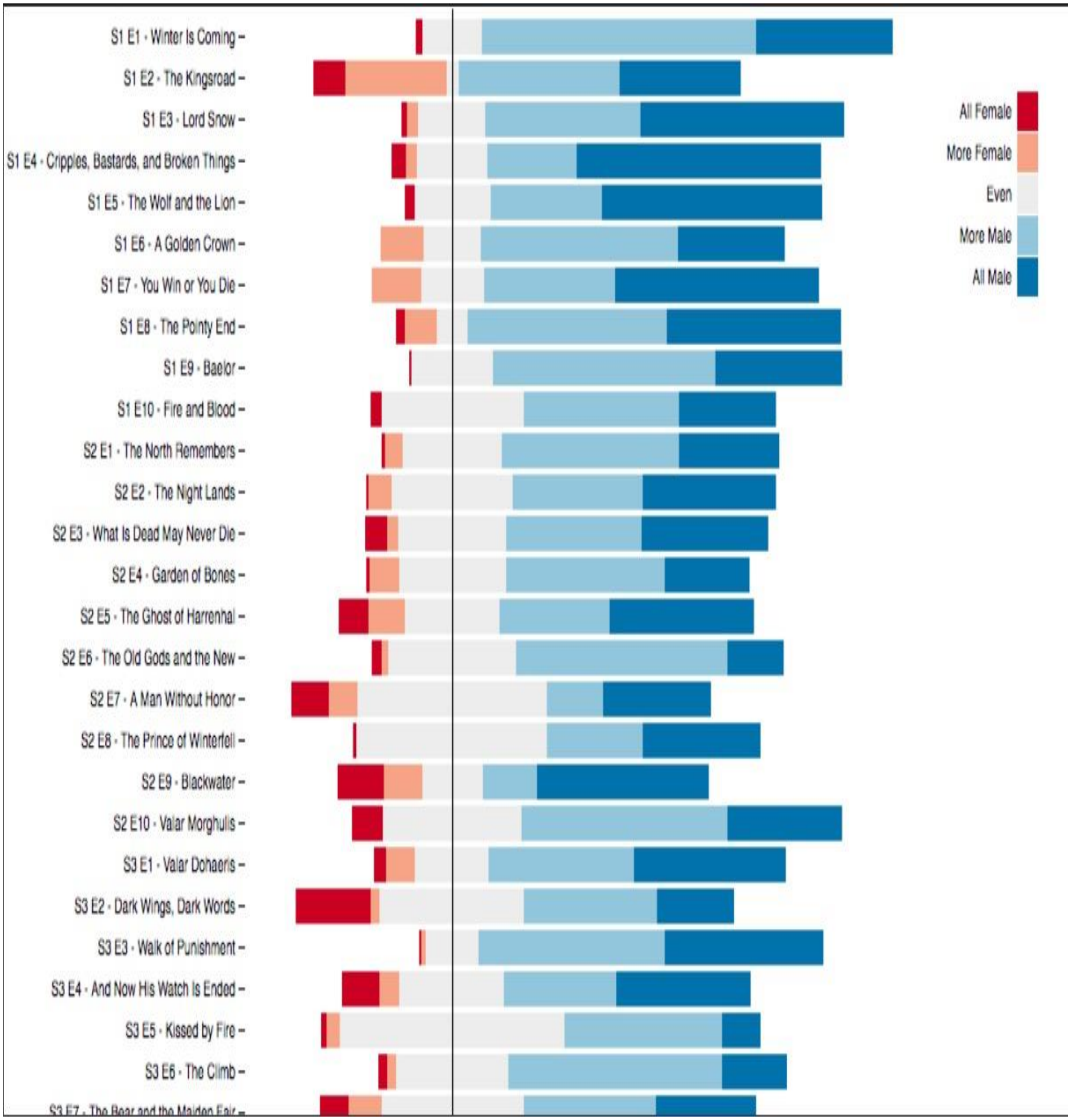
- **What's the gender balance of screen time per season?**



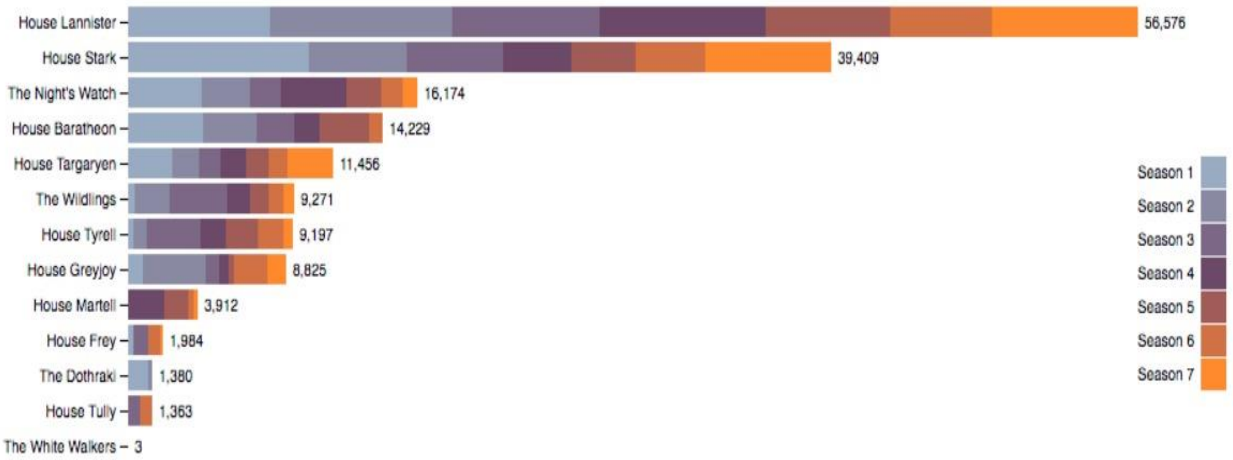
- **What percentage of time do characters spend in various locations?**



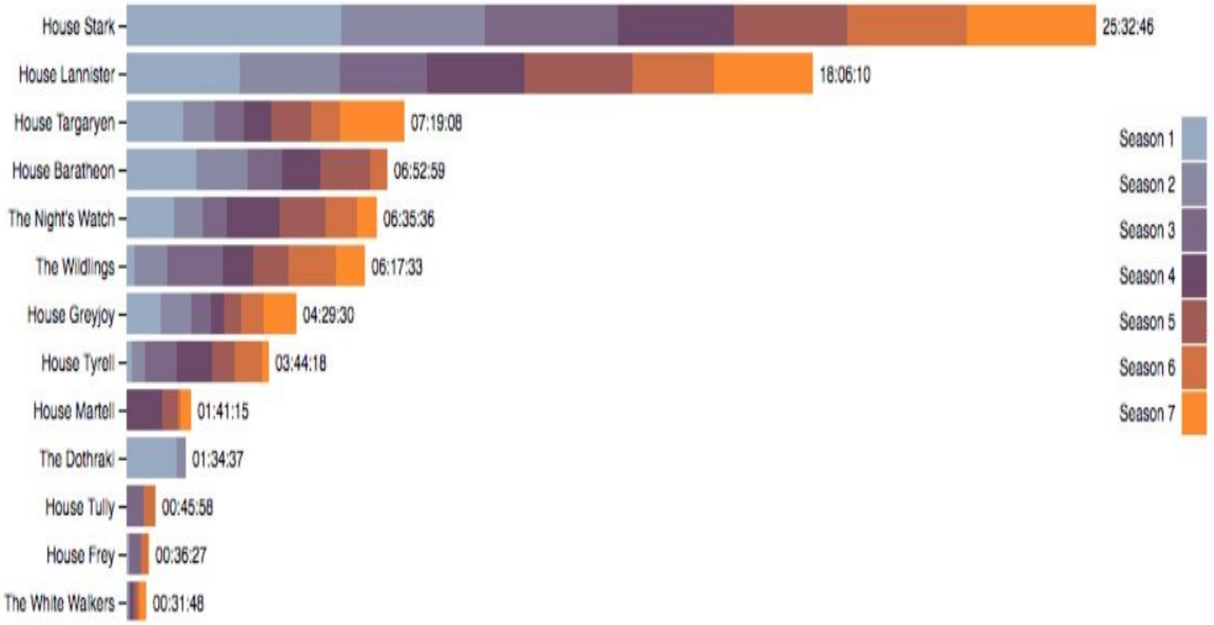
- **What's the gender balance of screen time per episode?**



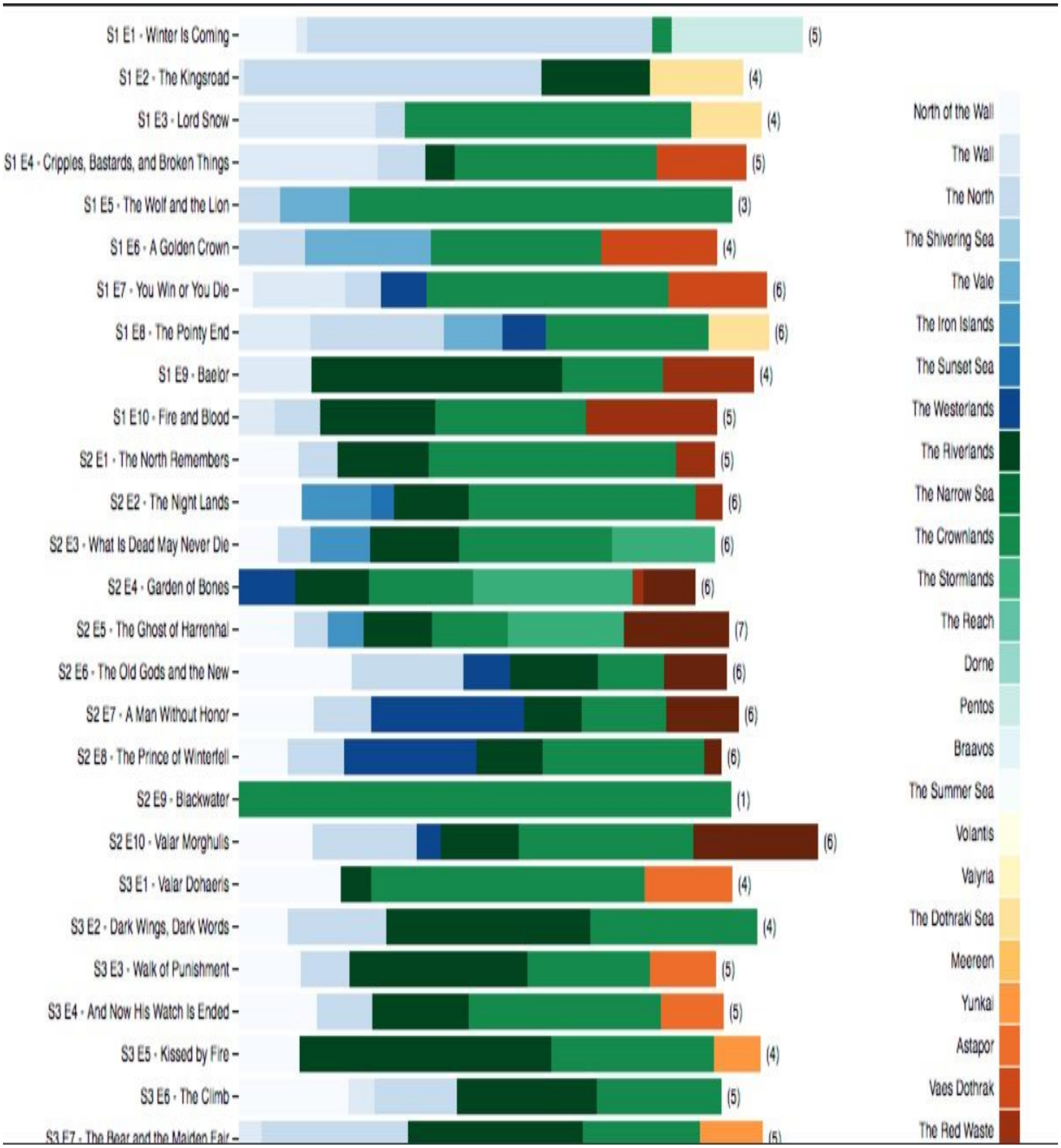
- **How many words do characters in each House speak?**



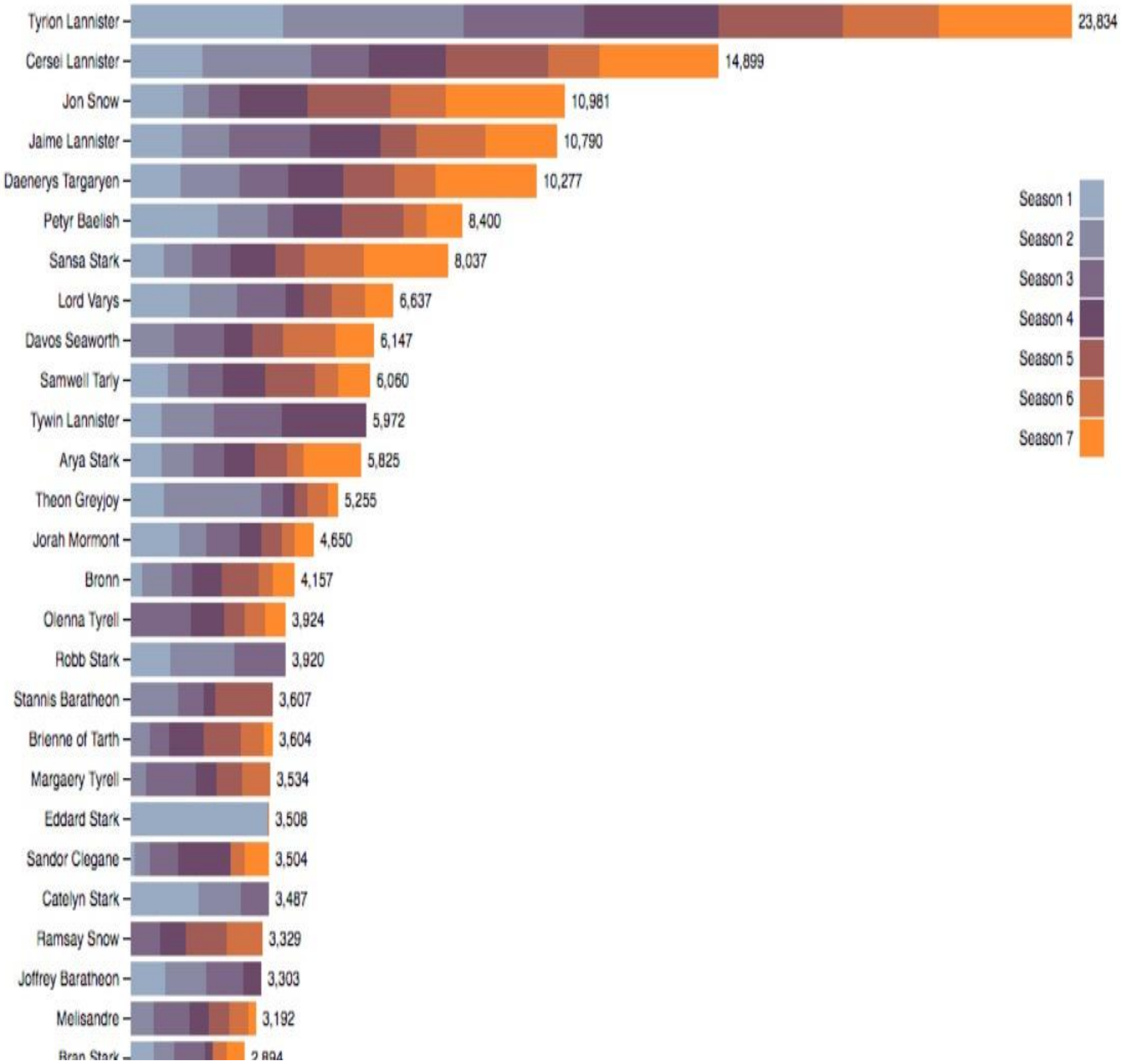
- **How long are Houses on screen?**



- **How many locations are in each episode?**

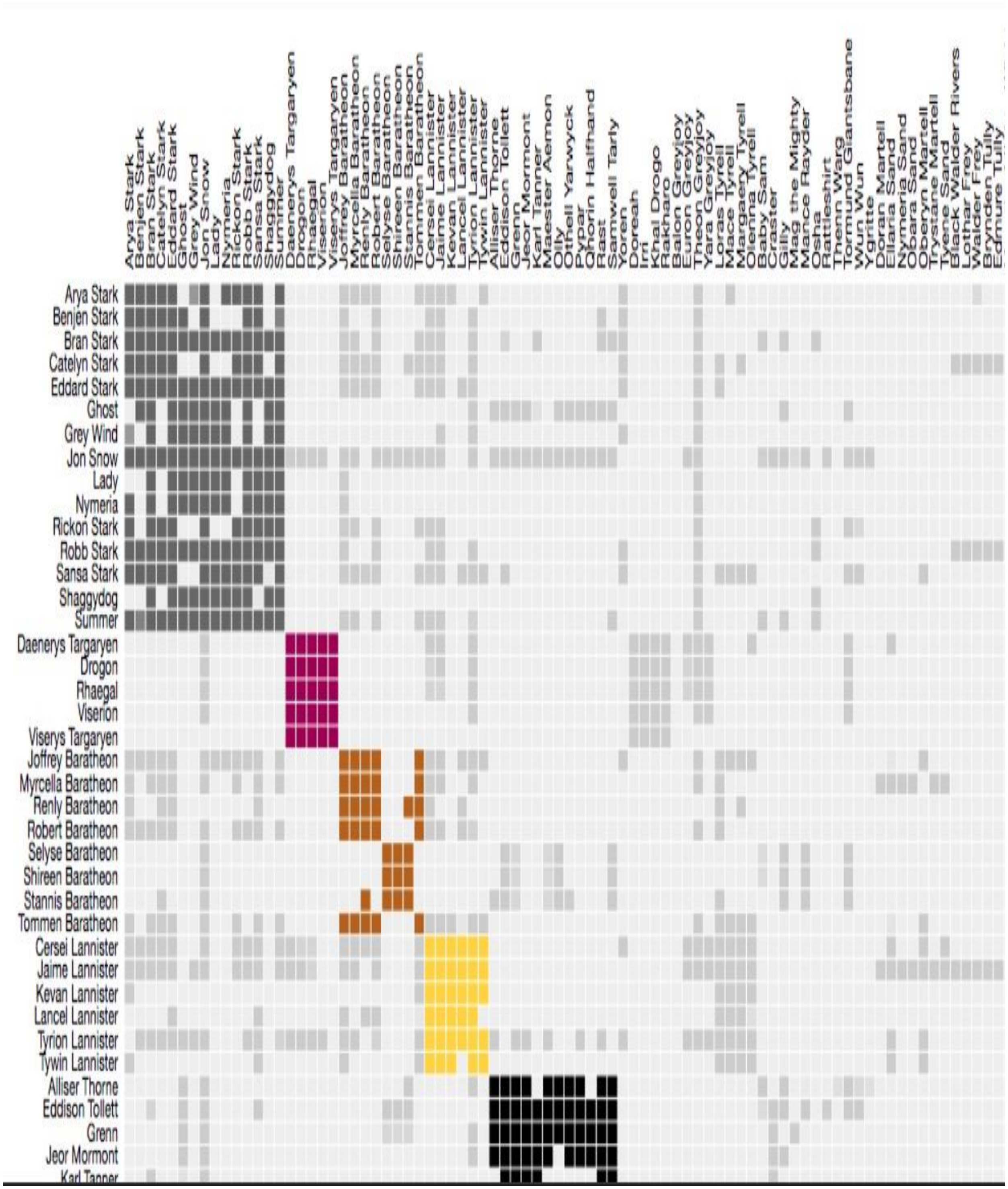


- **How many words do characters speak?**

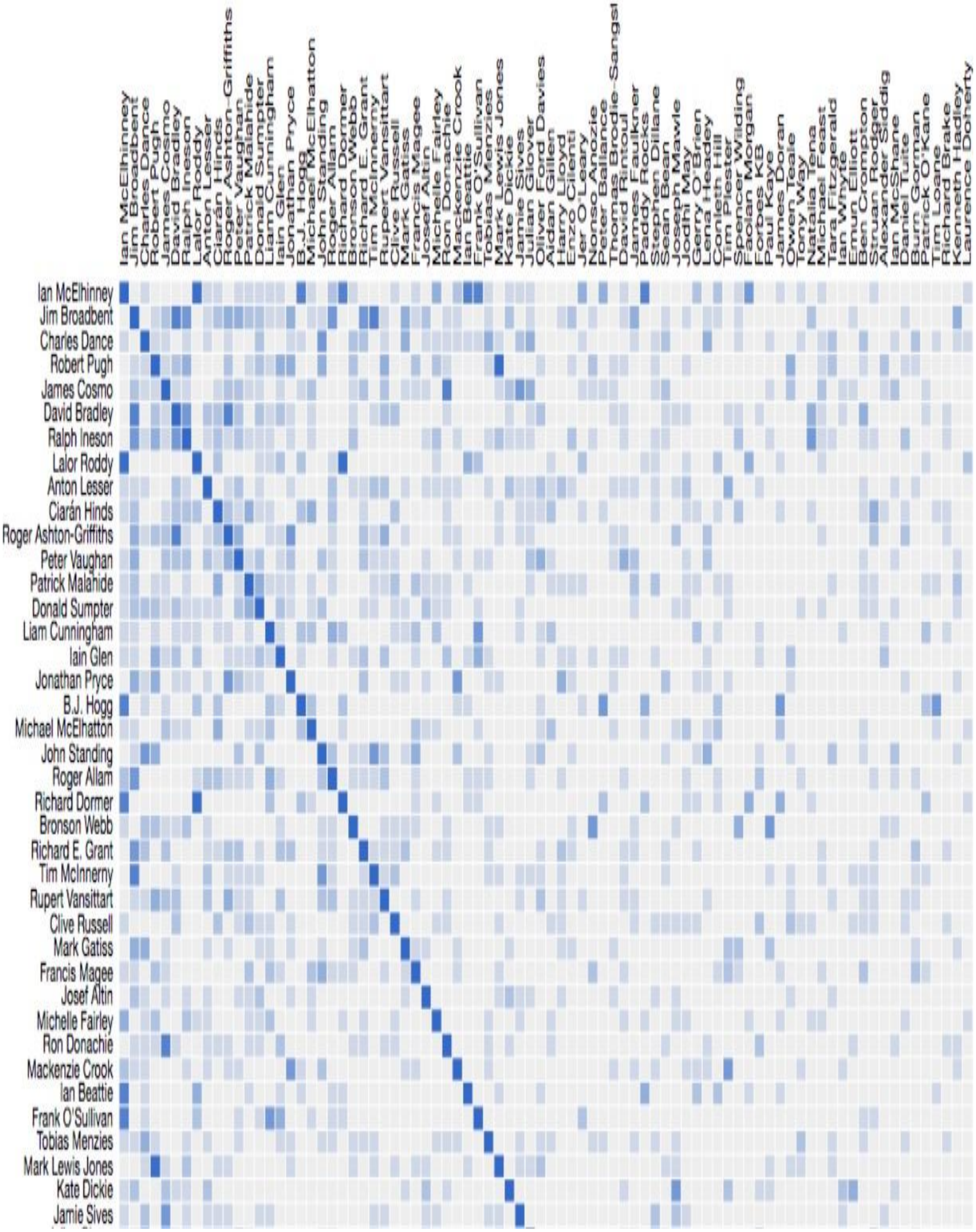


Co-Occurrence

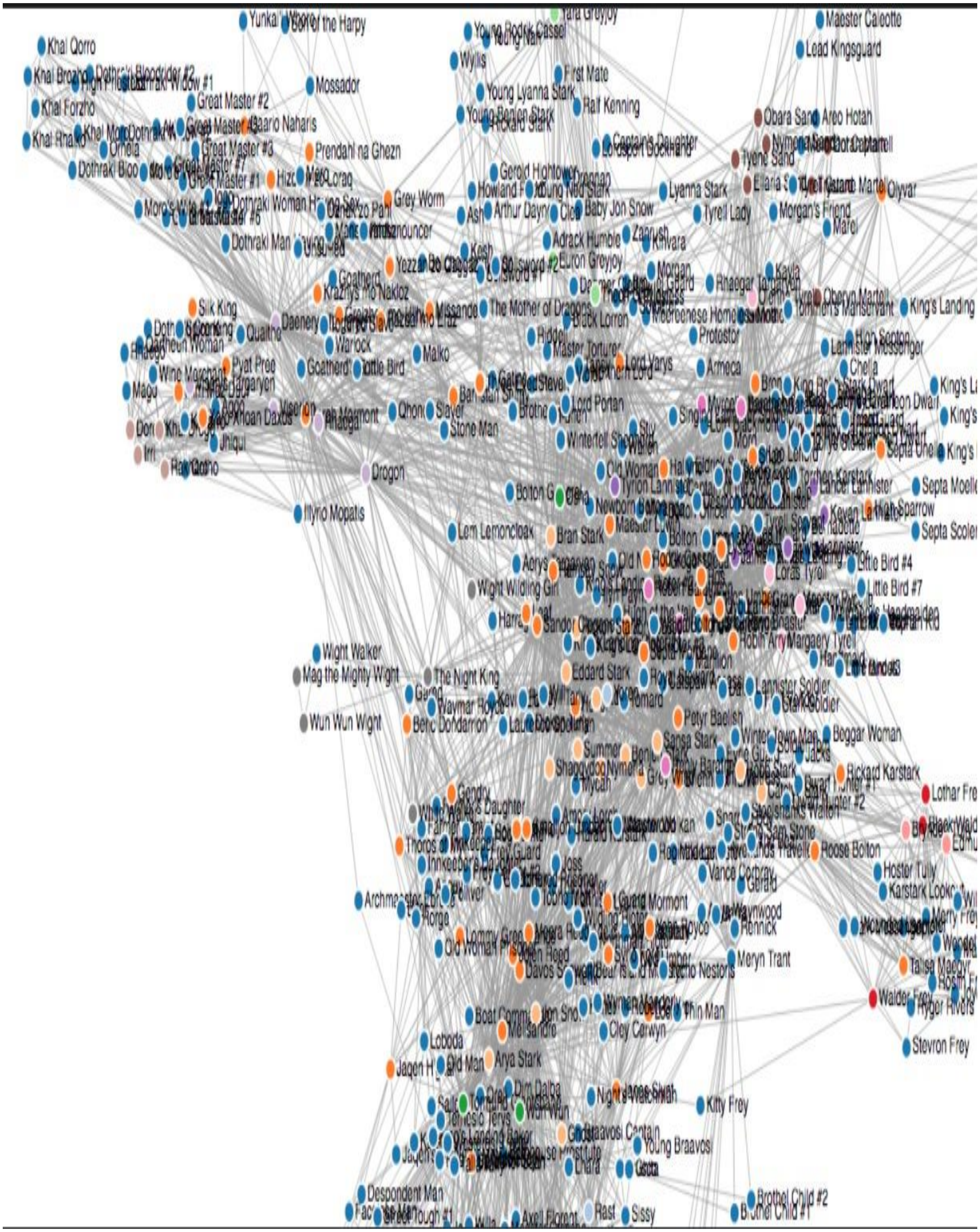
- **Which characters are on screen together?**



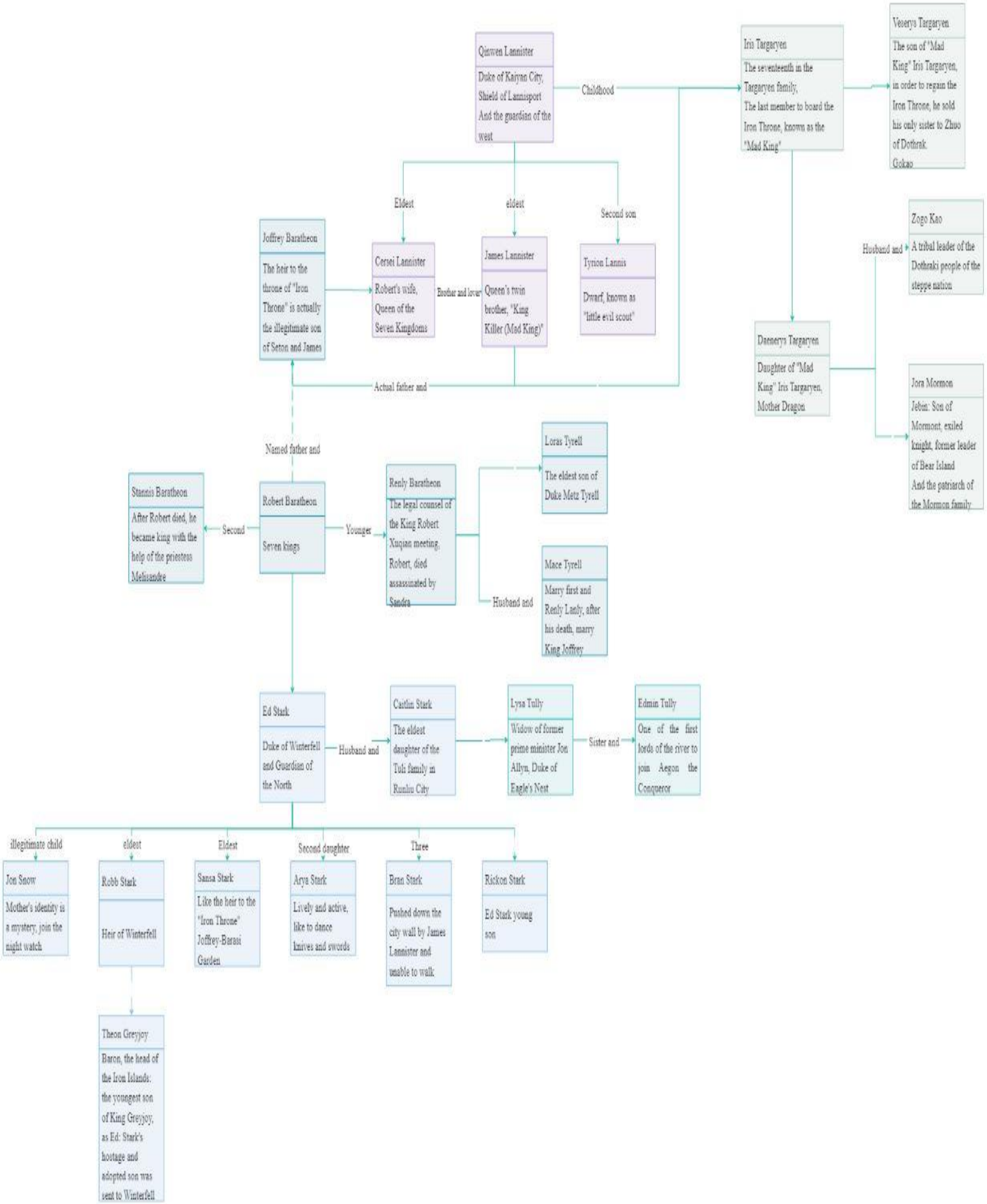
- **Which actors have been in movies with other *Game of Thrones* actors?**



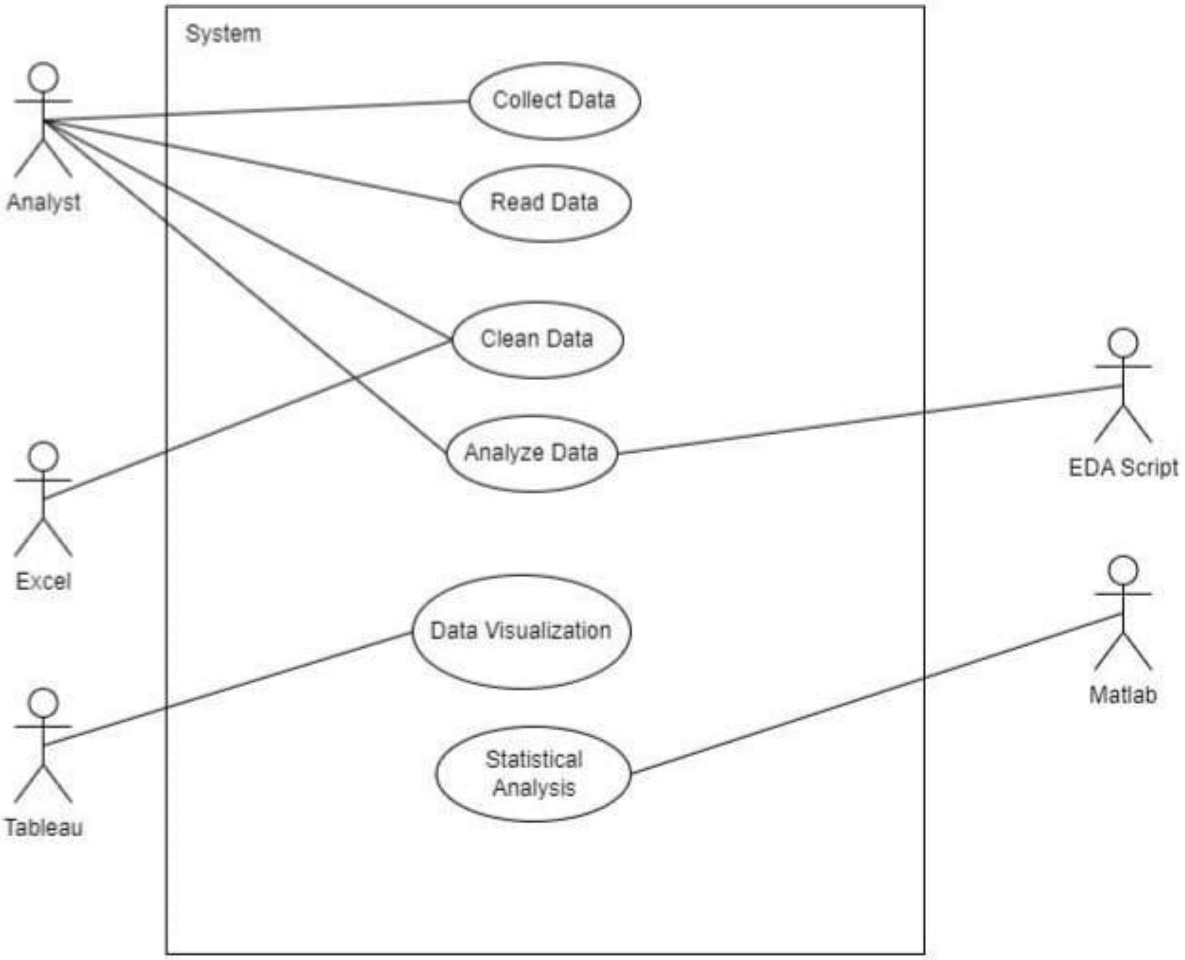
- **... or as a force-directed network?**



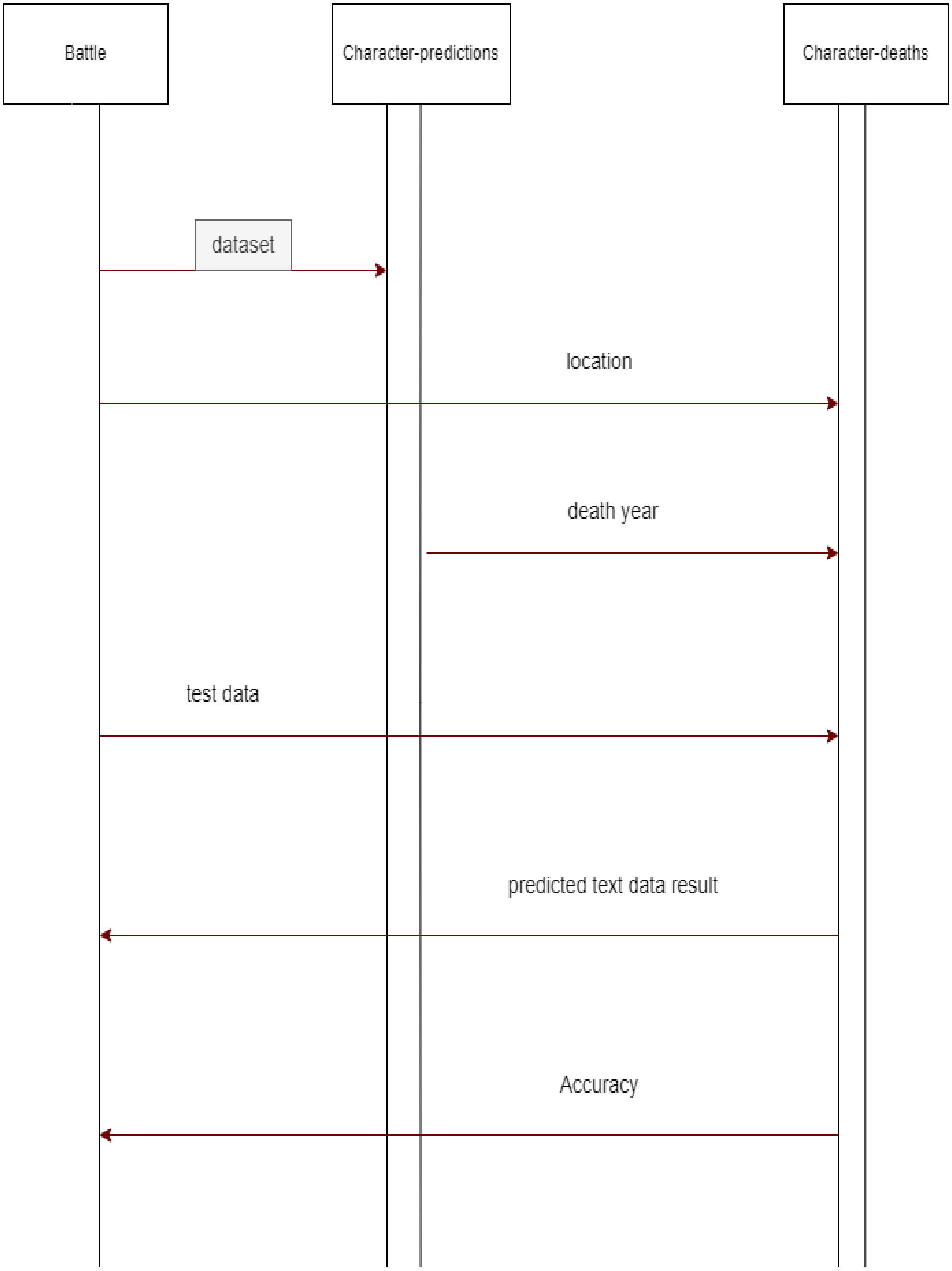
- **Relationship Diagram:**


Qinwen Lannister
Duke of Kaiyan City, Shield of Lannisport And the guardian of the west
Childhood
Iris Targaryen
The seventeenth in the Targaryen family, The last member to board the Iron Throne, known as the "Mad King"
Veserys Targaryen
The son of "Mad King" Iris Targaryen, in order to regain the Iron Throne, he sold his only sister to Zhuo of Dothrak. Gokao
Eldest
eldest
Second son
Joffrey Baratheon
The heir to the throne of "Iron Throne" is actually the illegitimate son of Seton and James
Cersei Lannister
Robert's wife, Queen of the Seven Kingdoms
Brother and lover
James Lannister
Queen's twin brother, "King Killer (Mad King)"
Tyrion Lannis
Dwarf, known as "little evil scout"
Husband and
Zogo Kao
A tribal leader of the Dothraki people of the steppe nation
Daenerys Targaryen
Daughter of "Mad King" Iris Targaryen, Mother Dragon
Jora Mormon
Jebin: Son of Mormont, exiled knight, former leader of Bear Island And the patriarch of the Mormon family
Actual father and
Named father and
Loras Tyrell
The eldest son of Duke Metz Tyrell
Stannis Baratheon
After Robert died, he became king with the help of the priestess Melisandre
Second
Robert Baratheon
Seven kings
Younger
Renly Baratheon
The legal counsel of the King Robert Xuqian meeting, Robert, died assassinated by Sandra
Mace Tyrell
Marry first and Renly Lanly, after his death, marry King Joffrey
Husband and
Ed Stark
Duke of Winterfell and Guardian of the North
Husband and
Caitlin Stark
The eldest daughter of the Tuli family in Runliu City
Lysa Tully
Widow of former prime minister Jon Allyn, Duke of Eagle's Nest
Sister and
Edmin Tully
One of the first lords of the river to join Aegon the Conqueror
illegitimate child
eldest
Eldest
Second daughter
Three
Jon Snow
Mother's identity is a mystery, join the night watch
Robb Stark
Heir of Winterfell
Sansa Stark
Like the heir to the "Iron Throne" Joffrey-Barasi Garden
Arya Stark
Lively and active, like to dance knives and swords
Bran Stark
Pushed down the city wall by James Lannister and unable to walk
Rickon Stark
Ed Stark young son
Theon Greyjoy
Baron, the head of the Iron Islands: the youngest son of King Greyjoy, as Ed: Stark's hostage and adopted son was sent to Winterfell

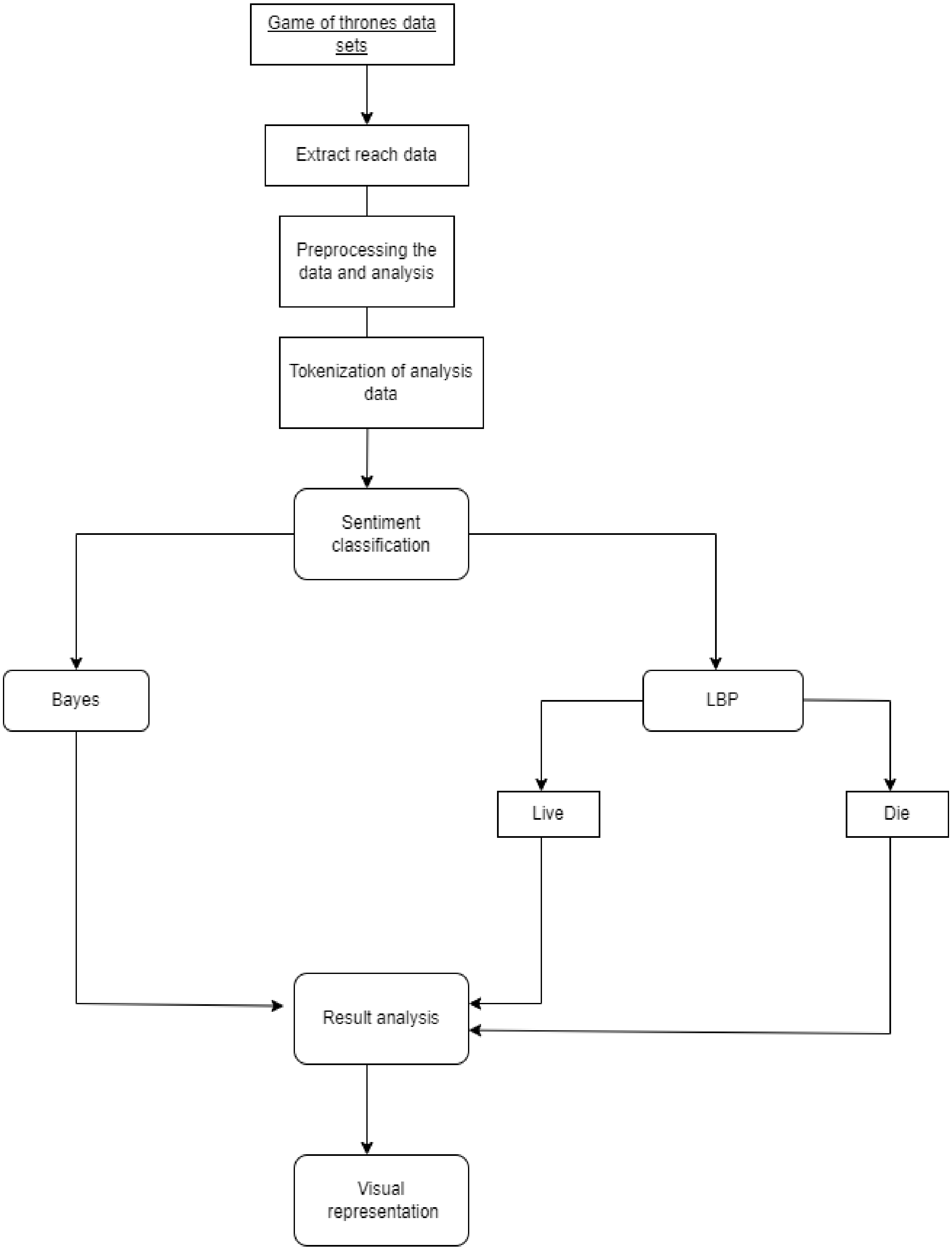
- **Use Case Diagram**



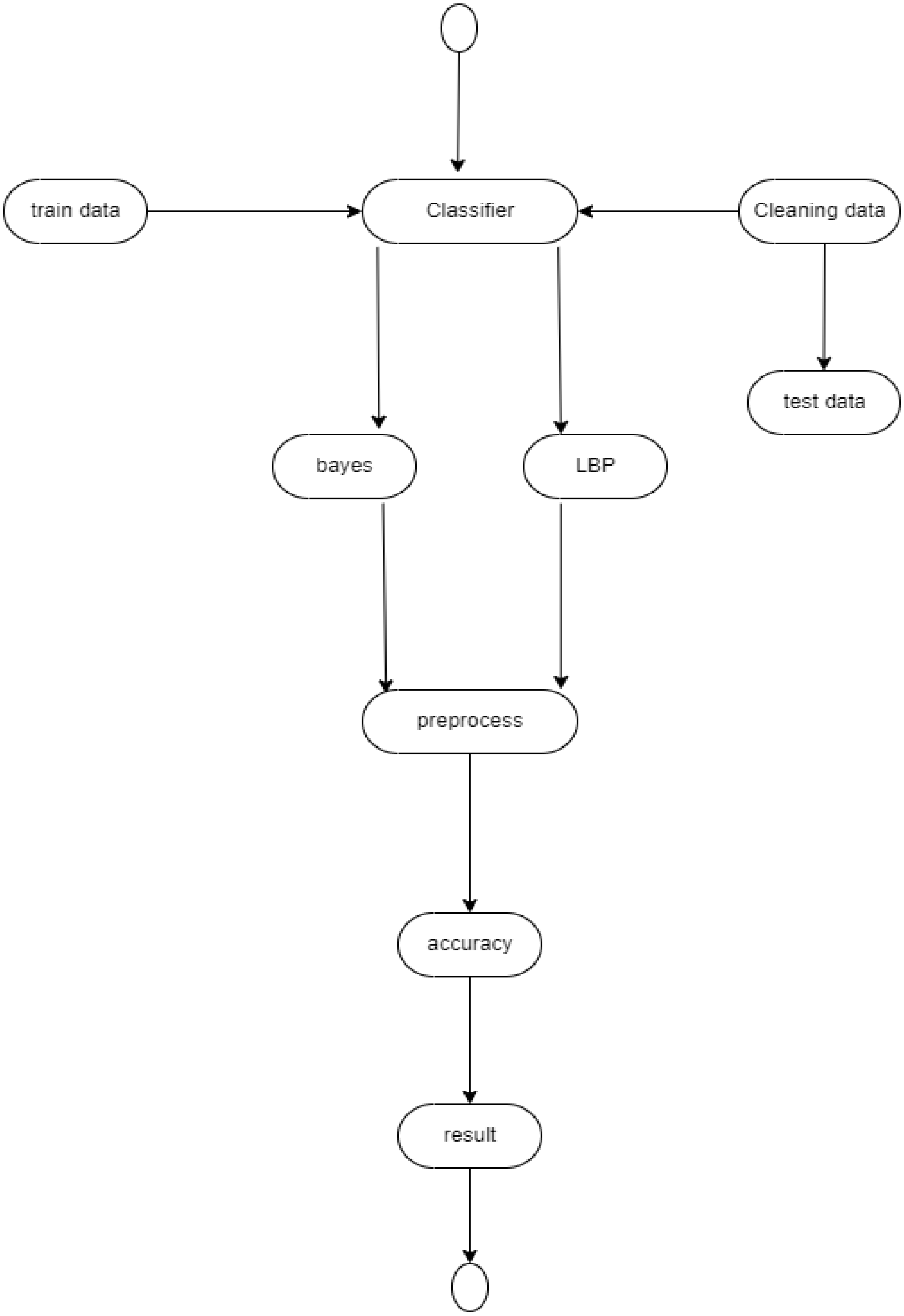
- **Sequence Diagram**



- **Flow chart Diagram**

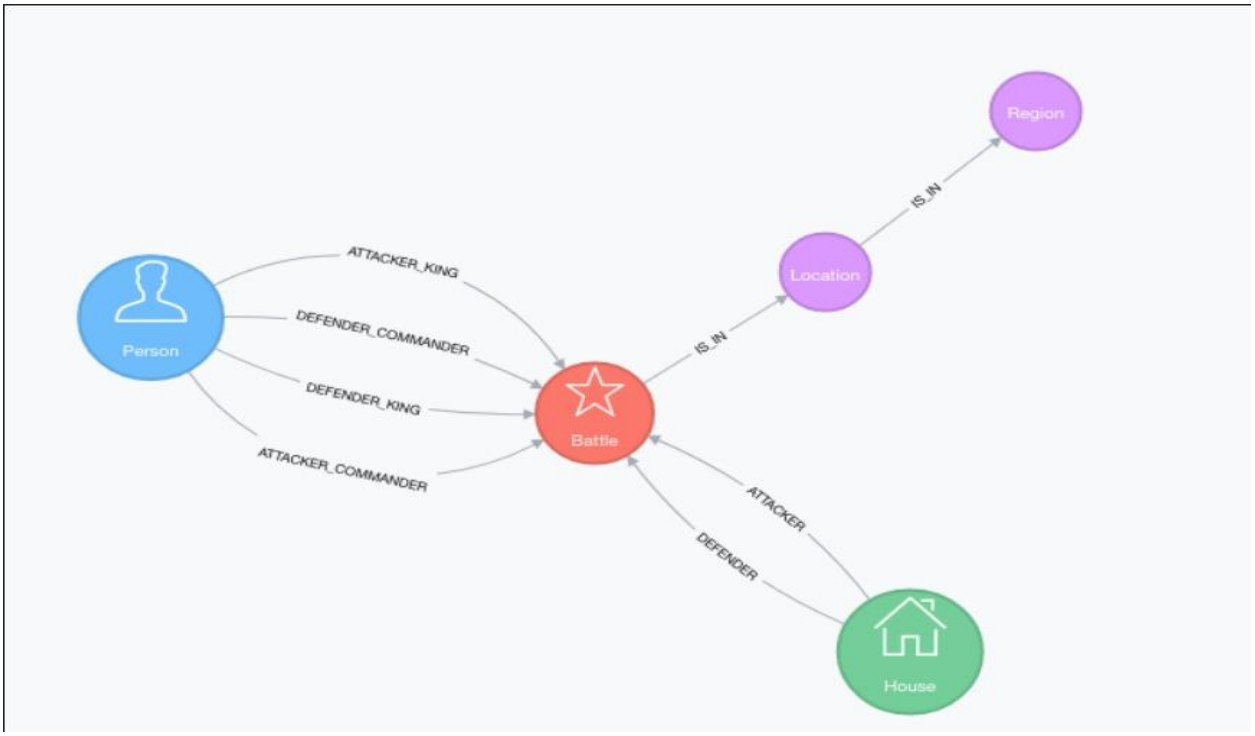


- **ACTIVITY DIAGRAM**

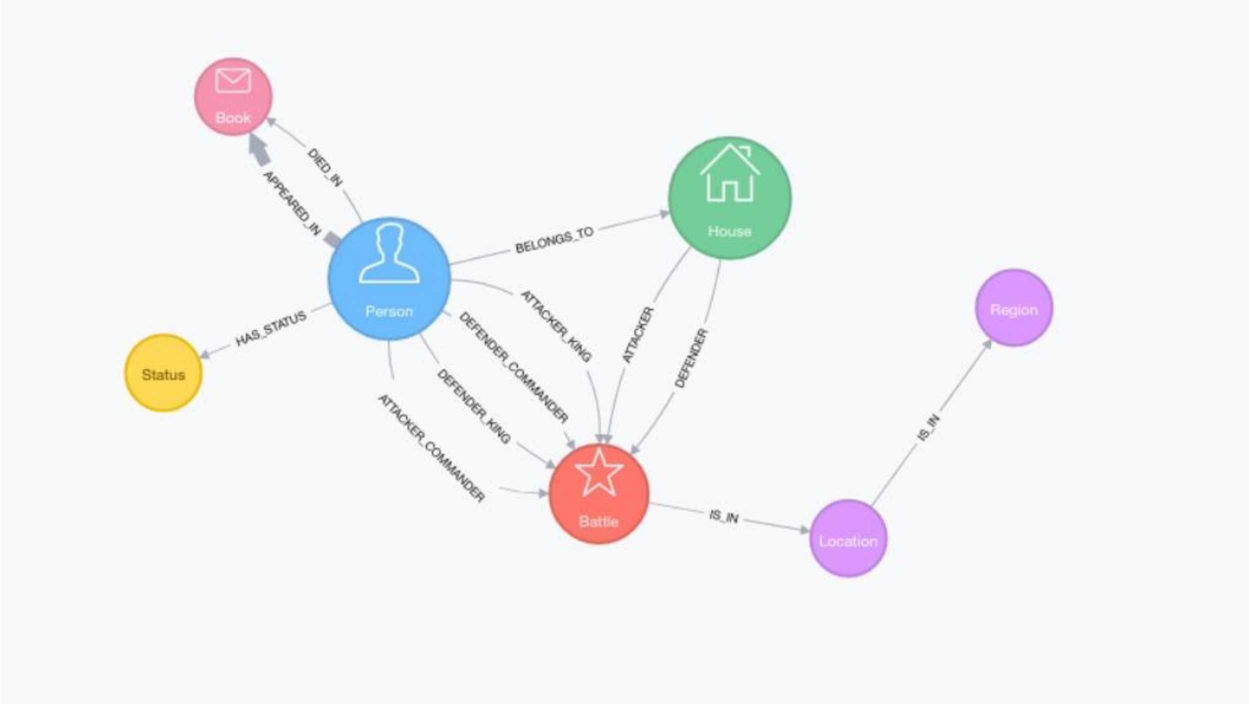


- **Game of thrones connection diagram**

1.



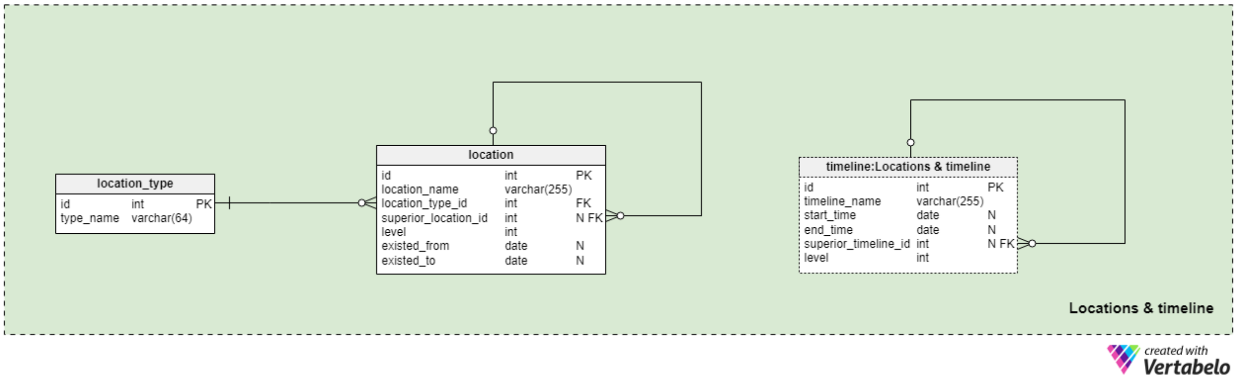
2.



- **Database Diagram**

This is not the sort of data model we've been featuring on the Vertabelo blog. Most of those are about running various types of enterprises or solving real-world business problems. This data model will focus on storing and modifying relationships and events in the story. So, let's start with describing what this database will cover.

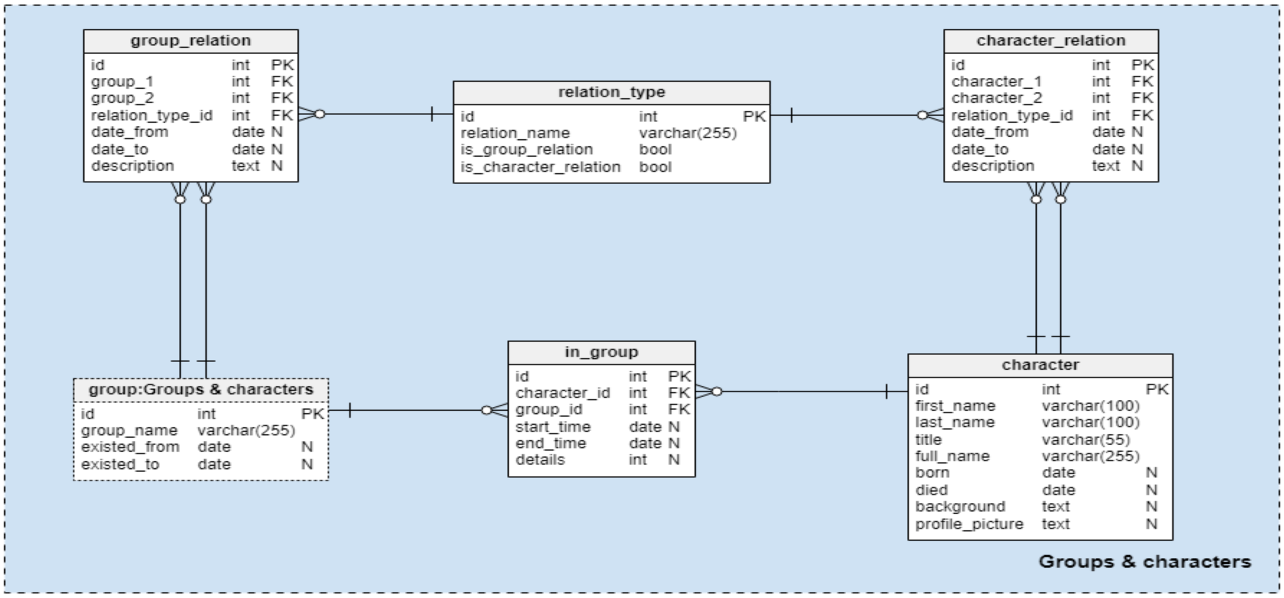
Section 1: Locations and Timeline



The first subject area we'll describe is the Locations & timeline subject area. The tables here are prerequisites for the tables in other subject areas.

The location table stores all locations in the series. For each location, we'll store its details and the first location that is "superior" to it. For a city (King's Landing), we will also store its region (The Crownlands) and its continent (Westeros). We'll have one record for each of these three and define the relation to the "superior" location.

Section 2: Groups and Characters



group_relation		
id	int	PK
group_1	int	FK
group_2	int	FK
relation_type_id	int	FK
date_from	date	N
date_to	date	N
description	text	N

relation_type		
id	int	PK
relation_name	varchar(255)	
is_group_relation	bool	
is_character_relation	bool	

character_relation		
id	int	PK
character_1	int	FK
character_2	int	FK
relation_type_id	int	FK
date_from	date	N
date_to	date	N
description	text	N

group:Groups & characters		
id	int	PK
group_name	varchar(255)	
existed_from	date	N
existed_to	date	N

in_group		
id	int	PK
character_id	int	FK
group_id	int	FK
start_time	date	N
end_time	date	N
details	int	N

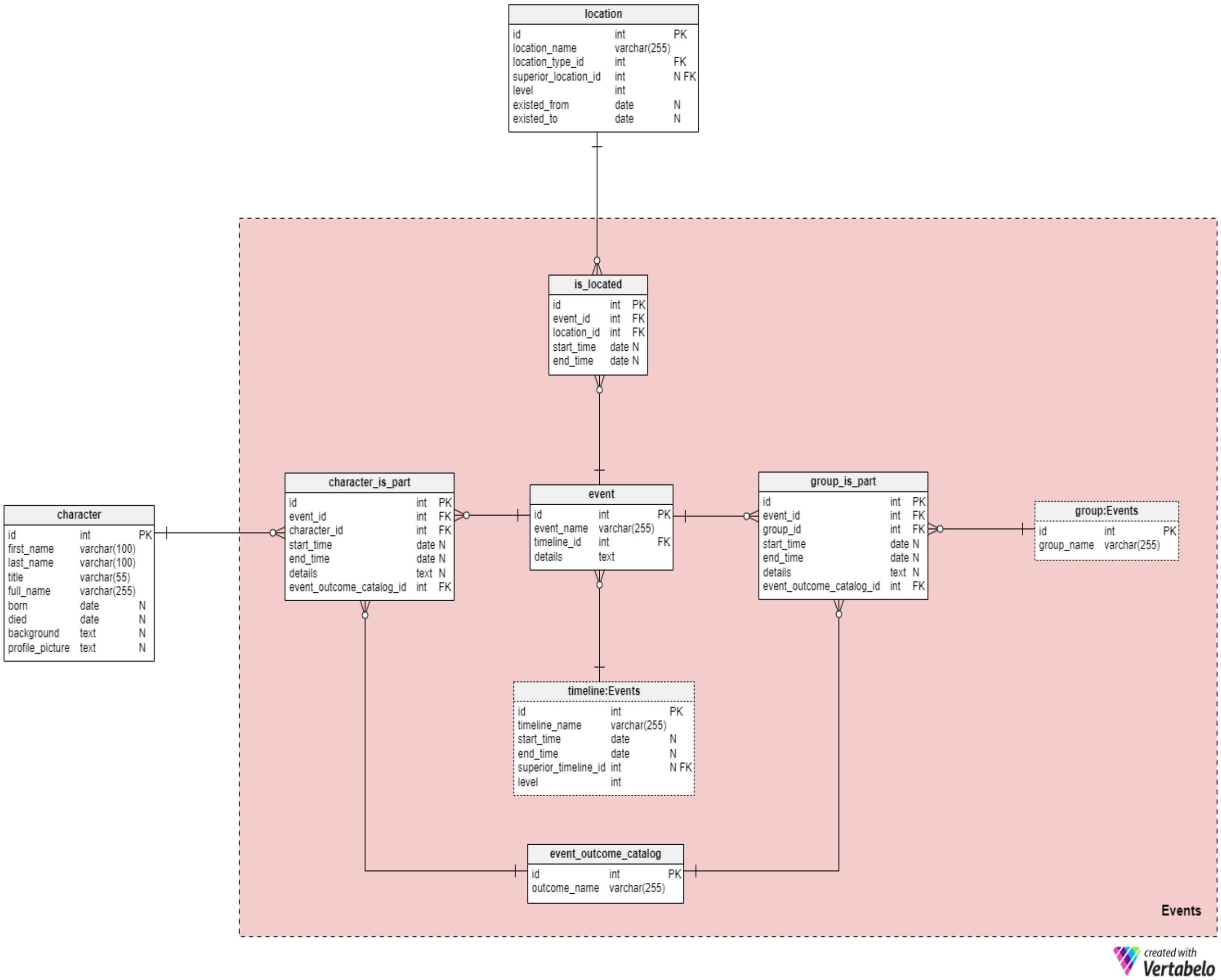
character		
id	int	PK
first_name	varchar(100)	
last_name	varchar(100)	
title	varchar(55)	
full_name	varchar(255)	
born	date	N
died	date	N
background	text	N
profile_picture	text	N

created with Vertabelo

The Groups & characters subject area stores all groups and organizations as well as the characters that belong to them. Characters could belong to several groups at once. For example, Daenerys Targaryen belongs to the House Targaryen and at the same time is part of the armies she leads.

We'll start with the character table. We'll store details about the characters that appear in the series and/or are important to the plot. We *could* store every single character, but there are many "nameless" characters that are not important to the story. For *Game of Thrones*, the most important characters include Tyrion Lannister, Cersei Lannister, Jaime Lannister, Daenerys Targaryen, Bran Stark, Arya Stark, Sansa Stark, Ned Stark, Petyr Baelish, Samwell Tarly, Robert Baratheon, and – of course – Jon Snow.

Section 3: Events



Events are the most important part of any story. They keep us reading or watching. The whole plot can change, either slightly or dramatically, based on its events.

We'll use five more tables to store event details. The remaining two tables – the group table and the timeline table – are copies, used here to avoid overlapping relations.

We'll start with the central table in this subject area, the event table. This is where we'll store events like battles, meetings, etc. An example of a GoT event is The Red Wedding.

Discussion

We can improve the model in the following ways:

a. **Adding more data** : Using huge volumes of data to train the models will help improve performance, leading to more accurate models.

b. Enhancing the quality of data by collecting news articles published in various domains as the vocabulary varies with domain. The news articles collected from social media platform will contain improper words like “awsm, fyn, baaad” etc are not used in news articles by news agencies.

c. **Using an Exhaustive Stop word List:** Apart from language stop words, there are some other supporting words as well which are of lesser importance than any other terms. These includes: Location stop words Country names, Cities names etc. Time stop words and Numerical stop words.

d. **Eliminating features with extremely low frequency:** There are words that rarely occur in many news articles and these words usually do not play much role in the text classification. Removing features for the words that rarely occur in news dataset can result in improving in the performance for different models.

e. **Use Complex Features:** N-grams and part of speech tags. Joining multiple words together to form a single feature and using these features along with single words for features can help in improving the model accuracy. Combination of N words into a single feature are known N-grams.

f. Using domain specific knowledge and human insight can help in choosing better featuresfor the prediction task under consideration.

Conclusion

In this project, we analyzed the data of Game of Thrones with the help of Python, for which we used Jupyter Notebook. We discussed a lot while doing this project, this project was like a challenge for us. When we started doing this project, many questions came in our mind, which we had to work very hard to find. But we got to learn a lot from this project as there were many visuals in this project.

Churn is another major consideration and one that has taken on new meaning during the streaming era. Warner Media plans to launch a new streaming service in the fourth quarter that will have to contend without new "Game of Thrones" episodes.

While many older consumers keep their pay TV package to avoid the hassle of switching, many younger viewers do not share that mindset, Brooks said. "After they've seen the end of a season or the end of a favorite series, they are ready to switch over to another service.

Reference

Game of Thrones Data sets:

https://www.kaggle.com/datasets/mylesoneill/game-of-thrones

Visualization:

https://jeffreylancaster.com/game-of-thrones/

Documentation:

https://blogs.sap.com/2014/12/11/data-geek-iii-analyzing-games-of-thrones-data-for-the-got-challenge/

Visualization:

https://github.com/Vishal-Sharma-26/EDA-on-Game-of-Thrones-Data

Diagrams:

https://app.diagrams.net/

Database Diagram & Documentation:

https://vertabelo.com/blog/a-song-of-ice-and-databases-a-game-of-thrones-data-model/