INTRODUCTION

This is an analysis of what viewers think about "House of the Dragon" series which is a spinoff of "Game of thrones" (GoT), arguably the biggest tv show ever created.

The show is set some 172 years before the birth of Daenerys Targaryen who is one of the main characters of GoT. It depicts the reign of house Targaryen and what ultimately led to the collapse of their all-powerful house.
The show was premiered on August 21, 2022.

A notable observation on social media, particularly twitter is that during or in the early aftermath of each episode's broadcast, the show is always among the top trending topics. However, trying to gauge how people feel about it from one's timeline will be skewed as tweets may be displayed according to individual preferences.
An objective analysis of this kind is therefore necessary to more accurately reflect how people feel about the series.

The analysis seeks to reveal insights like most talked about cast, most predominant hashtags and others.

DATA ANALYSIS PROCESS

The analysis will follow the six (6) main steps of data analysis namely:

1. **Ask** (*questions to make data-driven decisions)*
2. **Prepare** (*data for exploration)*
3. **Process** (*data from dirty to clean)*
4. **Analyze** (*data to answer questions)*
5. **Share** (*data through the Art of Visualization)*
6. **Act** (*on revealed insights to improve products or processes)*

Ask (*questions to make data-driven decisions)*

Some of the questions to guide analysis may include;

1. What character is talked about the most?
2. What are the popular hashtags used for commenting about the show?

It will also seek to provide a insights on some of the viewership numbers as well as key impressions like retweets and likes.

Business Task (Deliverable)

Identify what people (on twitter) think about the first season of "house of dragons"

Stakeholders:

Primary: Fans of "house of dragon" series
Secondary: 1. HBO product & marketing analytics team
2. Twitter marketing & analytics team

Prepare (*data for exploration)*

Python was used for the cleaning process.

Data for the analysis process was sourced from twitter using ***snscrape*** python library. The data contains 10 CSV files organized in long formats. This was done on 24th November, 2022.

An analysis of the biases and credibility of the data was done following the ***ROCCC*** method.

1. *Is the data **R**eliable?*
2. *Is the data **O**riginal?*

The data was sourced directly from twitter using snscrape. Its reliability and originality can be regarded as very high.

3. *Is the data **C**omprehensive?*

To more accurately reflect the perception of the show, it was important to consider tweets about each of the ten (10) episodes. There are over five hundred million (500,000) tweets sent out daily and so to use the showing period of the series, (August – October, 2022) as a condition for gathering tweets would skew the entire dataset to the last day of premiere.
To ensure the data is much more comprehensive, ten thousand (10,000) tweets were collected on the day of showing of each of the ten (10) episodes and stored in corresponding CSV files.

4. *Is the data **C**urrent?*
5. *Is the data **C**ited?*

The data is current and cited.

Process (*data from dirty to clean)*

The ten (10) CSV files containing the data are combined into one (1) CSV file and read into a data frame using python pandas library for cleaning and subsequent analysis.

Inspection of data revealed too many inconsistencies with the location column. Very few cells had actual countries listed and so there would be a lot of bias in any insights derived. The column was therefore dropped.

The Date column had time stamps and so were split to show just the date.

All hashtags were identified using the regex library and isolated into a list. Counter and collections python libraries were then used to determine all individual hashtags and their usage count. It was then read into a separate data frame.

Two (2) lists of possible spellings of cast and dragon names were created.

```
#define list of cast and dragons on the show with possible mispellings
Cast = ["rhaenyra", "rhanyra", "rhenyra", "raenyra", "renyra", "rhenera" "ranyra", "raenys", "renys", "ranys",
        "rhanys", "rhenys", "rhaenys", "daemon", "damon", "alicent", "alisent", "aliscent", "viserys", "vicerys"
        "viscerys", "aemond", "amond", "criston", "otto", "aegon", "agon", "corlys", "laenor", "lenor"
        "lanor", "larys", "vaemond", "vemond", "vamond"
        ]

Dragons = ["syrax", "cyrax", "sirax", "cirax" "caraxes", "karaxes", "seasmoke", "sea smoke", "sea-smoke", "Seasmoke",
           "arrax", "arax", "vhagar", "vhegar", "vagar", "vegar", "vhaegar" "meleys", "maelys", "malys"
           "vermax", "arrax", "vermithor"
          ]
```

Tweets were then split into words (tokens) using nltk.tokenize libraries and iterated against the cast and dragon lists. Returned cast and dragon names were isolated into a list for further processing.

To ensure consistency of extracted names, cast and dragon dictionaries were created with correct name spellings as keys and possible misspellings as values.

```
#dictionary of mispelled Cast names and their correct spelling
Cast_names_dict = {
                            "Rhaenyra": ["rhaenyra", "rhanyra", "rhenyra", "raenyra", "renyra", "rhenera", "ranyra"],
                            "Rhaenys": ["raenys", "renys", "ranys", "rhanys", "rhenys", "rhaenys"],
                            "Daemon": ["daemon", "damon"],
                            "Alicent": ["alicent", "alisent", "aliscent"],
                            "Viserys": ["viserys", "vicerys", "viscerys"],
                            "Aemond": ["aemond", "amond"],
                            "Aegon": ["aegon", "agon"],
                            "Laenor": ["laenor", "lenor", "lanor"],
                            "Vaemond": ["vaemond", "vemond", "vamond"],
                            "Otto": ["otto"],
                            "Larys": ["larys"],
                            "Criston": ["criston"],
                            "Corlys": ["corlys"],
                    }

#dictionary of mispelled dragon names and their correct spelling
Dragons_names_dict = {
                            "Syrax": ["syrax", "cyrax", "sirax", "cirax"],
                            "Caraxes": ["caraxes", "karaxes"],
                            "Seasmoke": ["seasmoke", "sea smoke", "sea-smoke", "seasmoke"],
                            "Arrax": ["arrax", "arax"],
                            "Vhagar": ["vhagar", "vhegar", "vagar", "vegar", "vhaegar"],
                            "Meleys": ["meleys", "maelys", "malys"],
                            "Vermithor": ["vermithor", "vemithor"],
                            "Vermax": ["vermax"],
                    }
```

Extracted names (both correct and misspelled) were iterated against values in the dictionary and any identified match would return the key (i.e., correct spelling of the name).
Returned list of correctly spelled cast and dragon names were then processed using counter and collection pandas libraries to determine all individual casts/dragons and their suage counts. They were then read into separate data frames.

A similar process was also used on the user column to reflect insights on accounts used to tweet.

Analyze (*data to answer questions)*

Python and Tableau were used in the analysis phase.

Using pandas library in python, the newly created data frames of Cast, Dragons, Users and Hashtags were all sorted in descending order to reflect the most talked about cast and dragon, the account with the most tweets as well the hashtag mostly used to tweet.

"Count" function in tableau was used to determine the total number of accounts used to tweet.

Analysis of all impressions, including retweets (RTs) and likes were also done using "*count*" functions in tableau.

Share (*data through the Art of Visualization)*

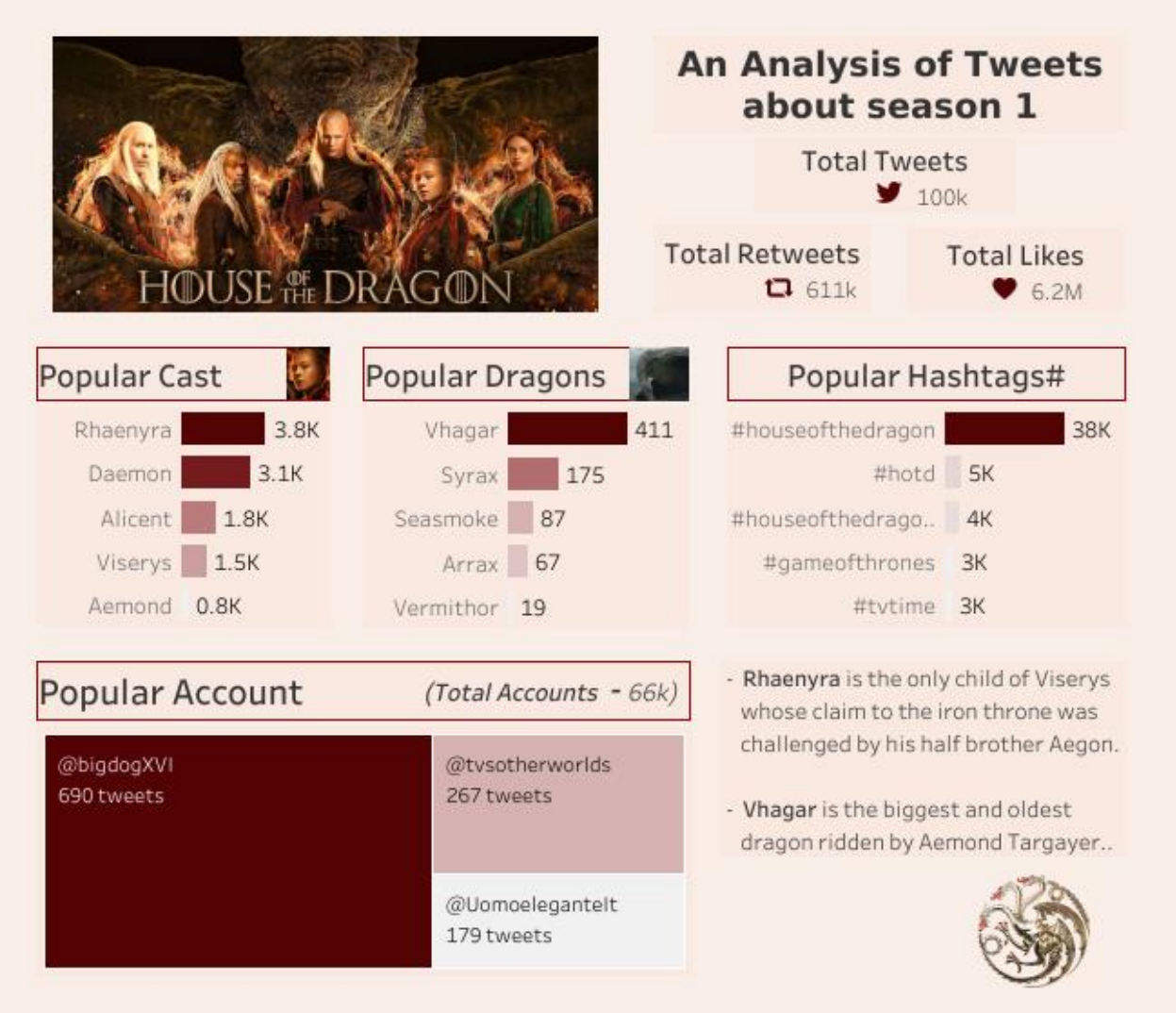
Final data was exported into "csv" files and subsequently exported to Tableau to be used to communicate insights via visualizations.

Analysis of the data revealed that Rhaenyra Targaryen was the most talked about cast in season 1 of the show with 3.8k mentions. Immediately behind her is her uncle/husband, Daemon Targaryen with also with 3.1k mentions. They appeared in all but one (1) of the ten (10) episodes of the show in very memorable scenes and so it is no wonder they are the two most talked about casts.

House Targaryen are dragon riders. At the height of their power, they had ten (10) adult dragons and nothing could stand against them in that time. The presence of dragons in the show is one aspect that viewers really love.

Analysis revealed that Vhagar is the most popular dragon of season 1 with four hundred and eleven (411) mentions.
Vhagar is the largest and oldest of all the dragons. She was ridden by Lady Laena Velaryon and then Aemond Targaryen after she died.

A lot of hashtags were used to tweet about the show. The most used however is the official hashtag of the show; "#houseofthedragon" with 38k mentions.
It is worth noting that 4th most used hashtag was "#gameofthrones" which is the original series from which House of the Dragon was spined off of. It garnered 3k mentions.



There were also 66k individual accounts involved which represents total viewers per 100k tweets.

Act (*on revealed insights to improve products or processes)*

Insights from this analysis can be used to guide production of season 2 considering how viewers feel about the show.

It could also help the marketing teams know how to promote the show to viewers.

Below is a link to the to the full project which has been posted on my GitHub profile.