With Great Power comes Great Visualisation (GPGV)
Svetlana Churina
A0228582X

The task for this event is to 'create visualization for your favourite fictional/real character'. After looking at the provided examples (like this one all-the-harry-potter-spells-when-they-were-used/ ) I remembered one of the greatest tv comedy show that have been existed - "The Office". I decided to analyse main character of this show - 'Michael Scott".

For the next visualisations I have used few datasets, that I get from  kaggle : the-office-lines , the_office_episodes.csv the_office_imdb. First dataset providing spoken text lines during the show for all characters. There are 5 attributes :

- index - Continuous
- Character - Discrete
- line - Nominal
- season - Ordinal
- episode_number - Ordinal

While working with this dataset I have found, that some episodes have been missing, but amount of the lost episodes is small enough for continue working with this dataset.

Second dataset give us more general information, such as title of the episode, director, writers, original air date. The attributes for this dataset are:

- season - Ordinal
- episode number in season - Ordinal
- episode number in series - Ordinal
- title of the episode - Nominal
- director - Discrete
- writers - Discrete
- original air date - Continuous
- production code - Discrete
- US viewers on original air date - Continuous

As for the third dataset, it's pretty much the same as the second one - this dataset providing general information. The only attributes, that we are looking for is Average IMDb rating.

For connection of all of those dataset I joined using two attributes - 'season' and 'episode_number'.

Because data, that have been provided can't be directly used for the questions,that I want to answer,I had to preprocessed it. Because I wanted to see the appearance of the famous Michael Scott line 'That's what she said' during the show, I had to store in which episodes are this line appeared and who said that with use of first dataset. As a result we have dataset that provide to us every appearance of this line during the show with next attributes:

- Season
- Episode
- Character.

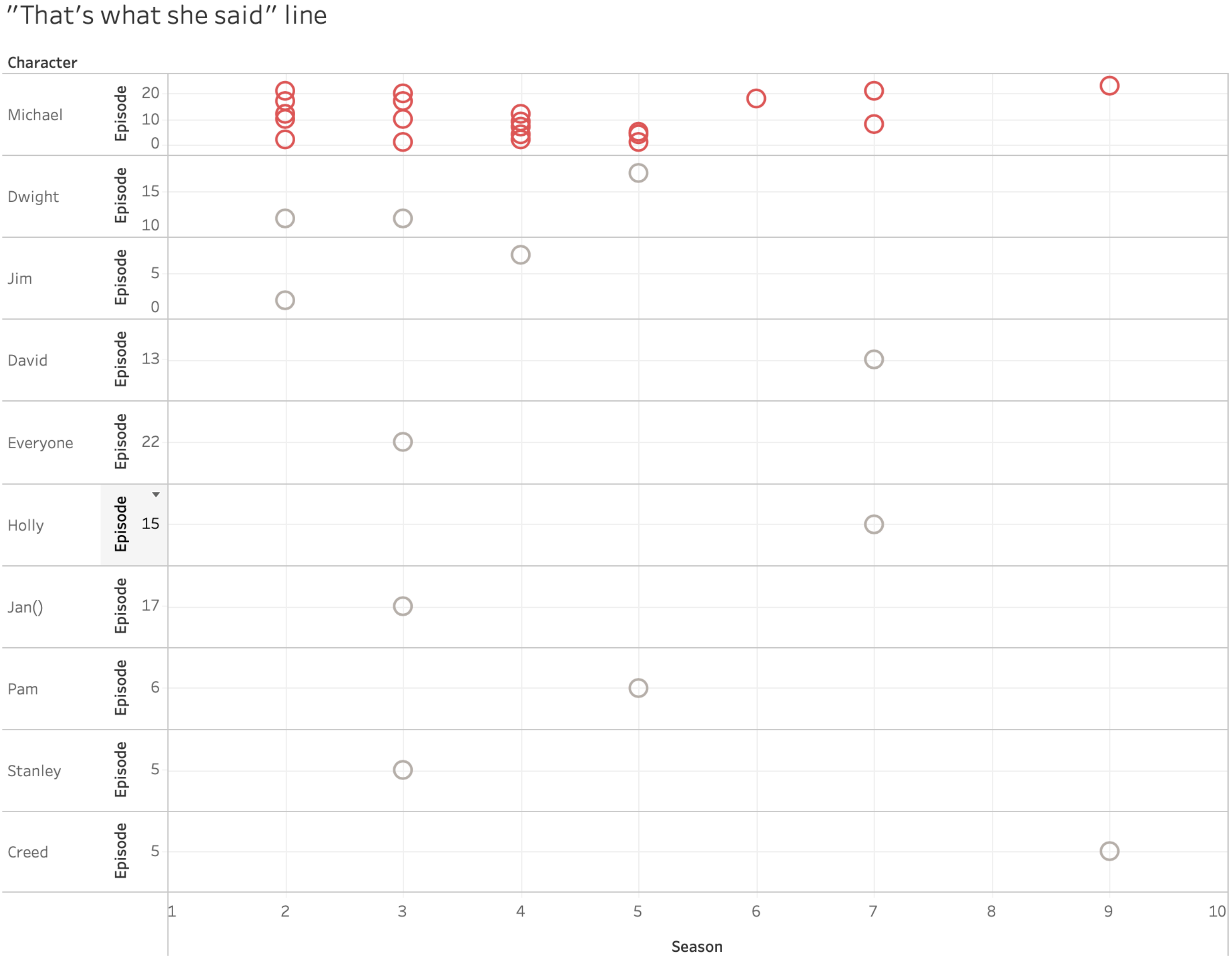
Also using first dataset, I have counted amount of lines from Michael Scott during the show. As a result I had dataset with next attributes:

- Season
- Episode
- Michael count (count of Michaels line in this episode).

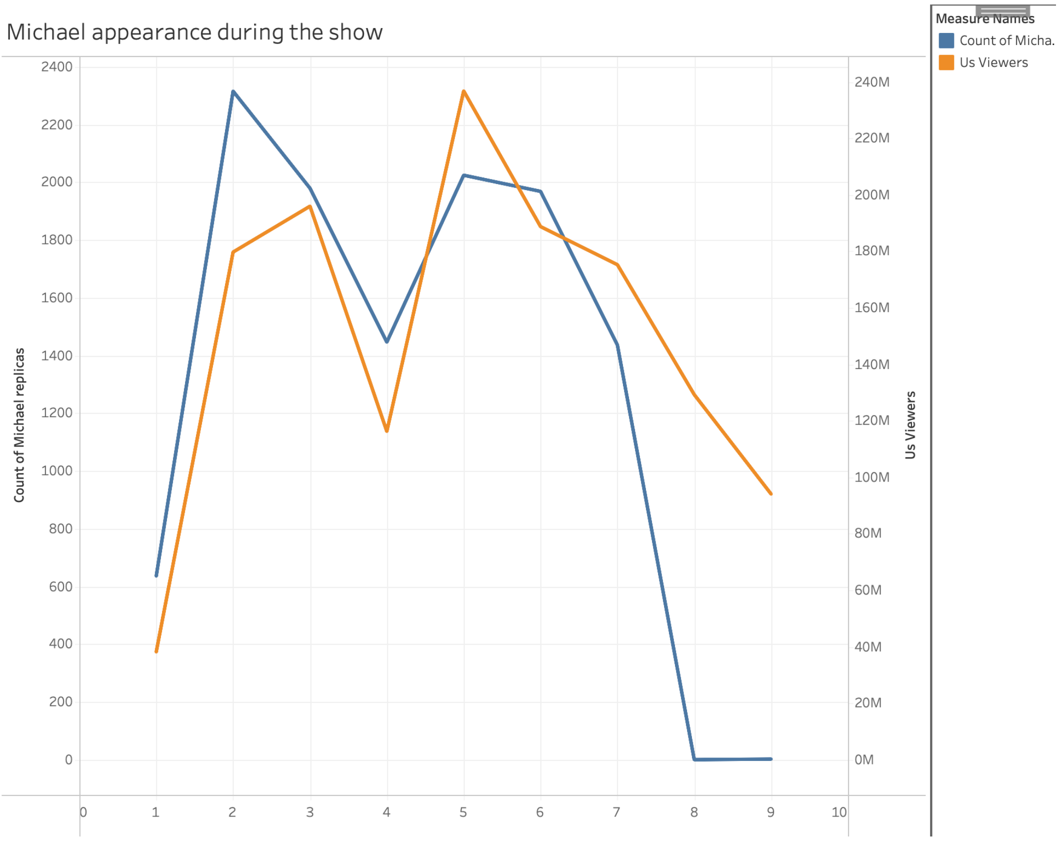
All of these pre-processing have been done with python.

As for my first visualisation "'that's what she said' line" I used 'season' as column and 'episode' and 'character' as rows. To separate Michael from other characters, I've made a set with this character and used it as a marks. The visualisation is a Gantt one.
For the second visualisation - "Michael appearance during the show" I used 'Season' as a column, and sum of Michael lines for the seasons with sum of US viewers as a rows. Also I'm marking median IMBd rating for the each season. The plot type is dual lines.
For the last visualisation I've made text visualisations using "character" attribute from the 'the-office-lines' dataset.

Questions, that can be pursued with first visualisation : 'In which episodes does 'that's what she said' line appears?' 'When was the first time Michael used that joke?" "Does this joke have been used by other people?"
Here is visualisation itself.
It was interesting to find out some unusual usage of this line : for example you don't expect it from "Creed". Also as we can see with this data visualisation, this line haven't been used that much, and started to appear only since season 2, while during the tv show it feels like this line is using very often.

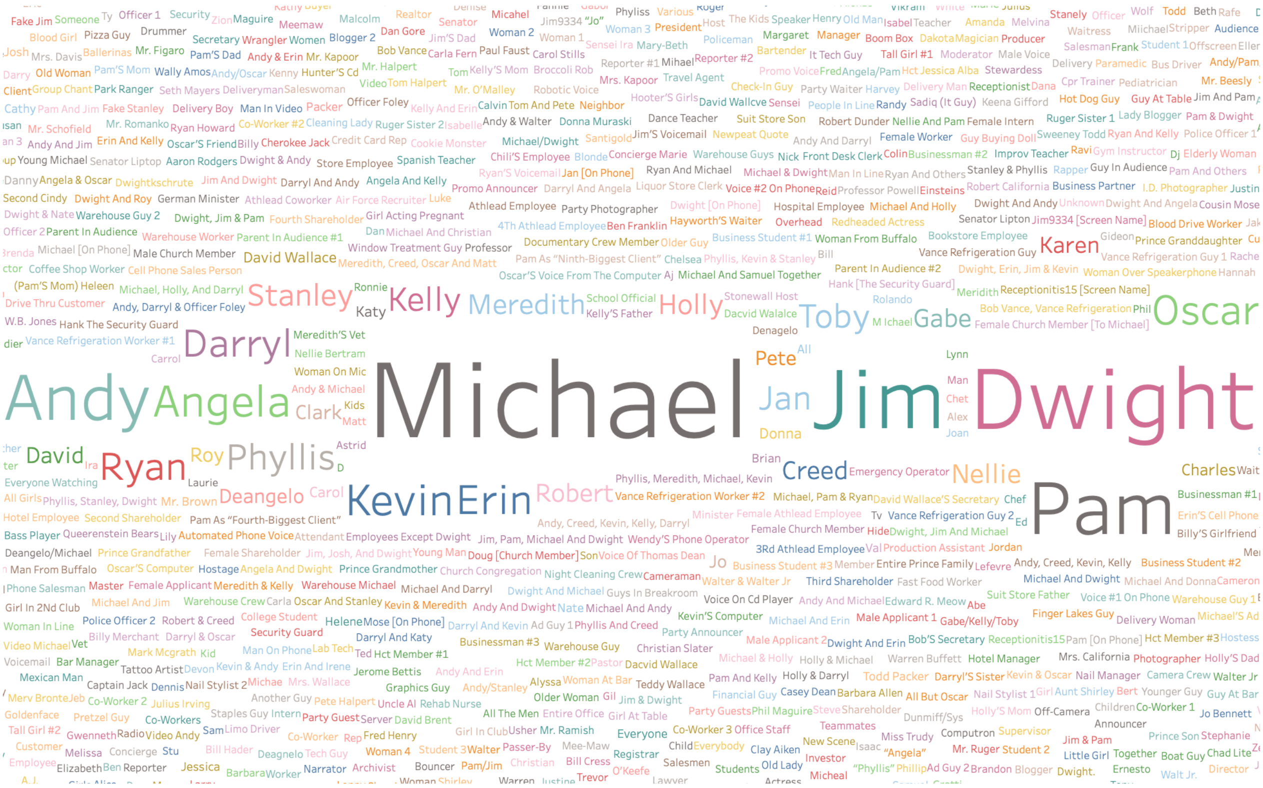


For the second data visualisation we can ask next question – 'When character Michael Scott disappeared in the show?" "Does disappearance of his character affected amount of viewers and IMDb rating"?
Most of the fans of this tv show saying, that after season 7, when Michael disappeared, show become not interesting and quality reduced a lot. With this plot we can see, that this is actually true. Disappearance of this character leads dramatic decrease of the viewers as well as IMBd rating also became lower, compare to other seasons. Of course metric of "Michael's lines" is not the one that we should use for judging reducing the quality of the show, there are a lot of other factors that can influence it, but from my point of view – Michael Scott appearance during the show is very indicative metric for popularity of this show.



With last visualisation we can ask "Who is/are the main character in this show?"
With this data visualisation we can see crystal clear, that Michael Scott is one of the main characters – he had the biggest amount of lines during the show.

Characters lines count



Link to the visualisations: https://prod-apnortheast-a.online.tableau.com/#/site/svetachurina/workbooks/508816?:origin=card_share_link