**Brief Description**

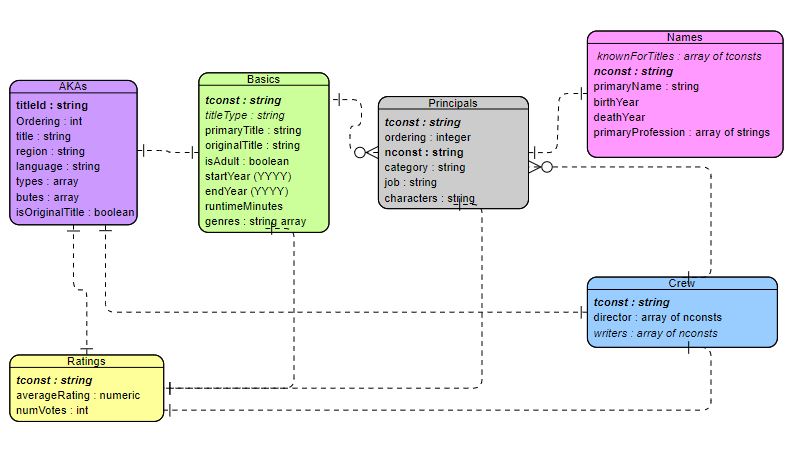
IMDb is the most popular online database of information related to films, television programs, etc. It includes a wide variety of data such as ratings, fan and critical reviews aswell as cast and production crew information.

Fig. 2 Structure of our Dataframes and how they relate to each other.

**Early Inspection and EDA**

In this project we decided to focus only on movies (title types ‘tvMovies’ and ‘movie’) , therefore we removed instances from other types, reducing our dataset by roughly 90% as we can see in the **Fig. 1**. ?\*leaving us with 296 734 observations from the starting 7 millions. \*? The datasets ‘title\_basics’ and ‘title\_ratings’ were merged by the ‘tconst’ column - an alphanumeric unique identifier. Inspecting the ‘runtimeMinutes’ we observed that some observations had 30 minutes or less of duration (68 cases) corresponding the durantion of a short movie meaning that they were wrongly classified as movies or tvMovies, we removed them.

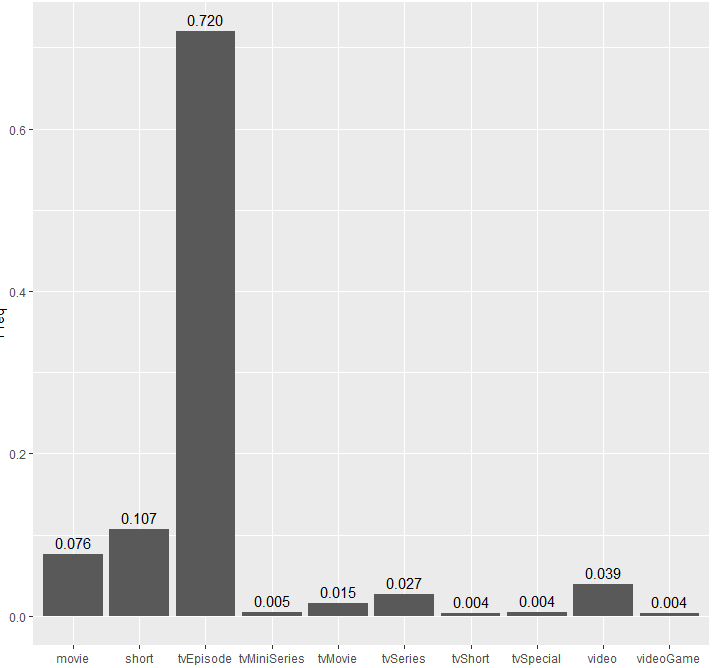
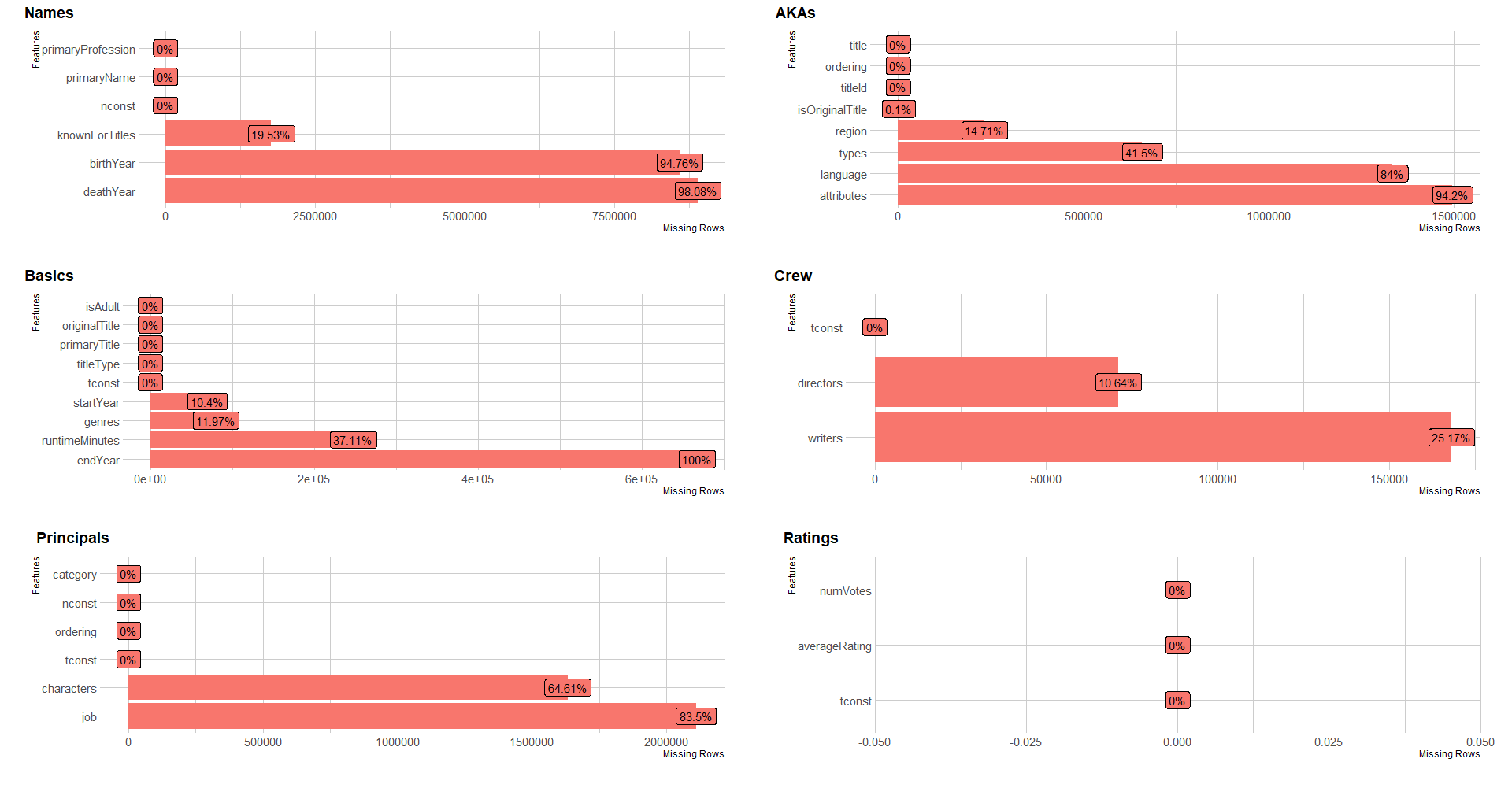


Fig.1 – Barplot of relative frequencie of title types.

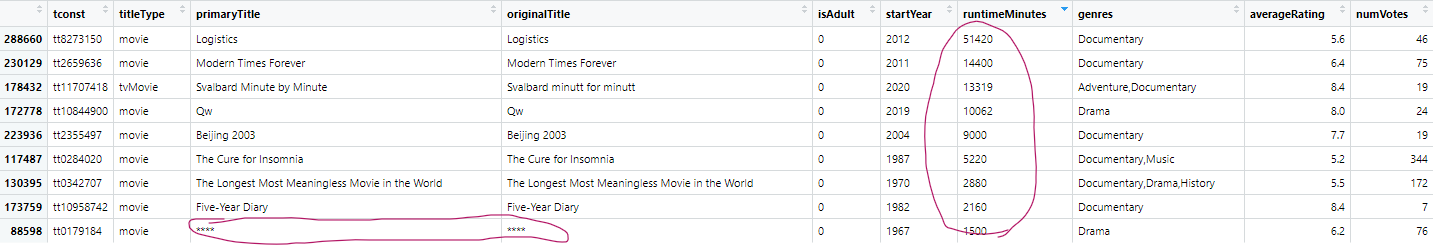
**Missing Values**

Originally, missing values were representend as “\N”, which was replaced by “NA” – the missing value indicator - so that R could correctly interpret these cells. Aftewards, the frequency of these missing values was plotted – **Fig.2**. The amount of missing values is especially high in AKAs, Names and Principals dataframes, which can lead to future complications when working with this data. **On Basics dataframe** the column endYear is used for television series and other types of titles that have a starting date and a end date, which explain the NA’s for a dataframe only with movies, this column is irrelevant for this project and it was removed. On the Names dataframe, both the column birthYear and deathYear have a very high amount of missing values, so they are irrelevant. On AKAs Dataframe it can be said the same of column language and attributes. On Principals Dataframe, column Job also has a very big percentage of missing values so it gives us little information.

**Fig 2.** - Relative frequencie of missing values in each dataframe.

**Wrong Entrances**

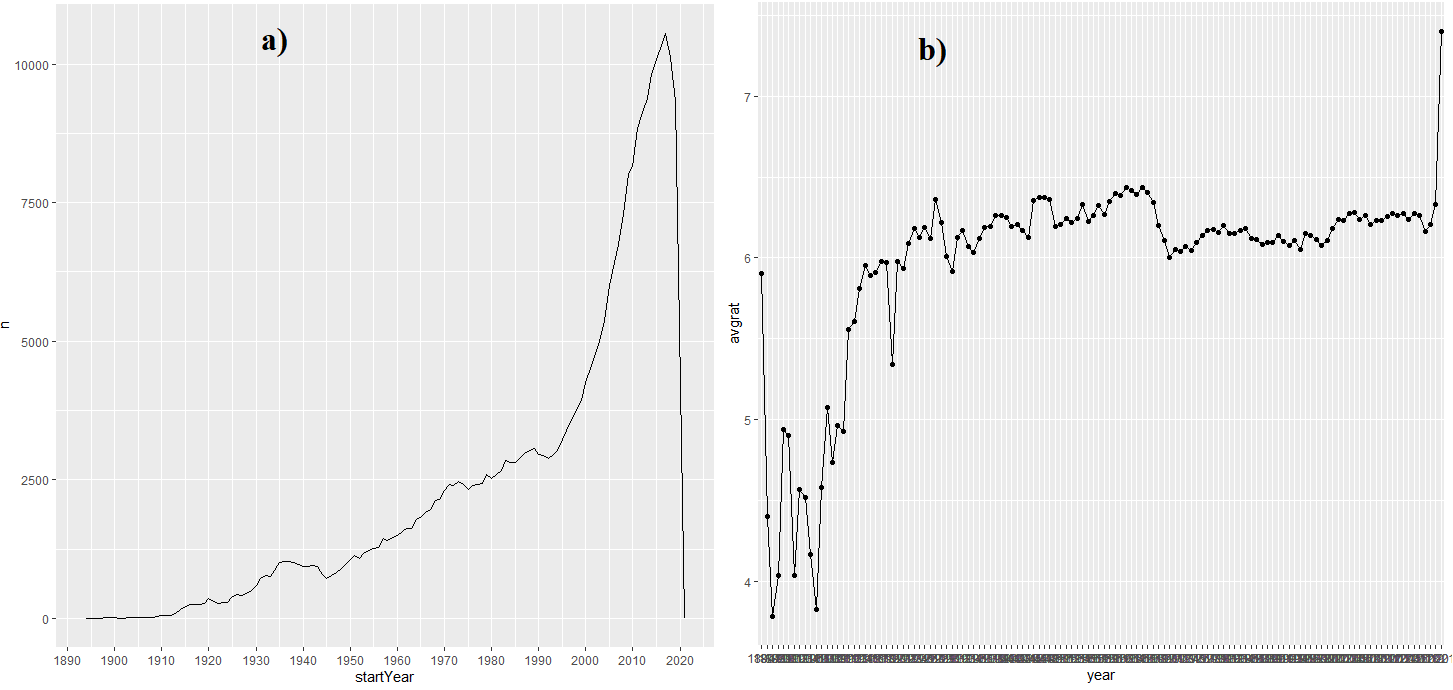
****While exploring the data, we came across wrong entrances that were removed, examples are shown on **Fig. 4**. During the project we are likely to find more of these observations.

****

## **Hypotheses**

**Are the movies getting better?**

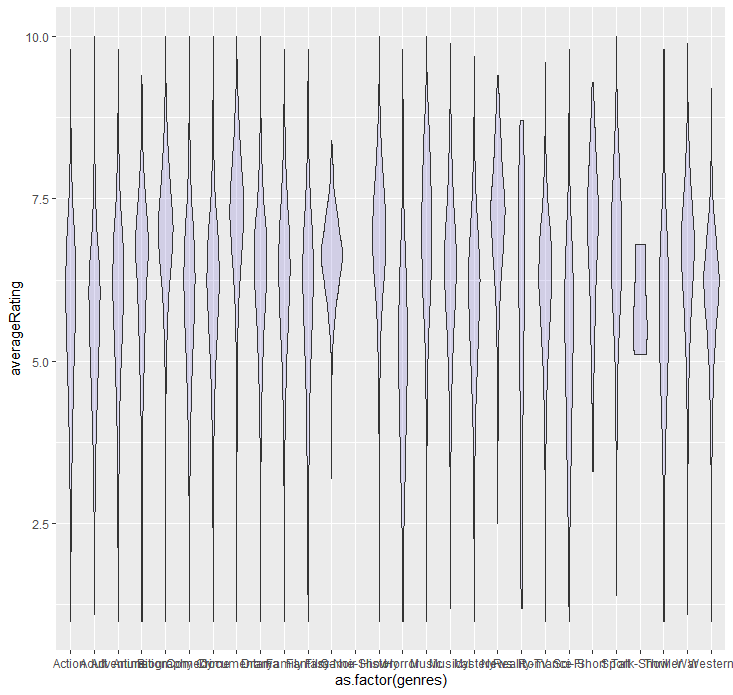
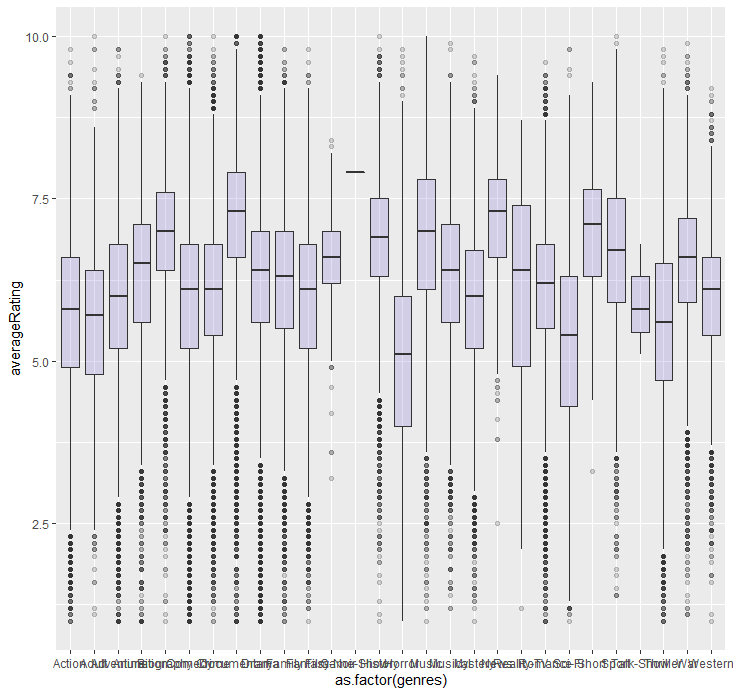
Although this question may look like a simple, straighfoward question, it’s actually quite complex and it may impose some challenges. To understand if the movies are getting better throughout the years we have to look at different factors as number of rating votes, average rating and number of films through years. We did some early graphs relating the average ratings and number of movies within the time line of our data **Fig 4.** . To have a clear visualization and answer this question we will need to clear the data from outliers and maybe define a a minimum rating vote that will work as a threshold in order to avoid bias. ?\*The feature ‘isAdult’ which includes porn movies can be a problem to.\*? Other transformations and statistical analysis will be necessary to have a final answer.



**Fig 4. –** a) Number of movies trough years, b) Averaging rating per year.

**Are the genres related to the ratings?**

Genres are a really interesting feature and we want to know more about it and how ratings can be affected. To have a early idea of what to expect we did a box plot and a violin plot to confirm the viability of the previous one **Fig 4**. We noticed that there was one genre ‘Game-Show’ with only one case that must be removed. Looking at the figure we can’t take certain elations, we must first aim to a clean and tidy dataset. New hypothises may arise while exploring more factors, for example we can look for interactions between genres. Some possible questions**:** How do different genres of movies rate through out the years. What genre of movies were more popular then and what genre is more popular now?



**Actor/Director**

Is there any combination of actor/diretor that achieves better ratings than the others? Was this consistent?

**Região/Rating**

**Filmes traduzidos em mais línguas melhor sucedidos, depende do genero?**

**Diretores/ Produtores que trabalham com mais homens e mulheres?**

**Generos de filmes que tendem para um genero do que pra outro nos seus papeis principais e se isso se relaciona com os ratings?**