At this project, I work with a dataset from Kaggle containing employee's age, gender, complaints-report per employee and Genre which is a general description of the type of complaint (e.g communication issues).

Firstly, I provide an exploratory data analysis to detect patterns and trends across the dataset. Then as a main task I work with the hypothesis that the Genre as information it is useful on the one hand, on the other hand though I believe is kind of abstract and vague. Hence I apply topic modeling technique such the LDA (Latent Dirichlet Allocation) which is the most common topic modeling technique to retrieve the topics of all complaints. To do that I first tokenize words of the corpus, remove the stopwords (words which does not add much meaning to a sentence such as "the", "an", etc.), remove whitespaces, convert to lowercases etc.. Then I tune a parameter just to optimize the number of topics in LDA. Finally, I use the visualization LDA library from genism to depict the produced topics and by relying a paper's idea [H.Gonen et al. (2021)] I try to analyze further the produced topics by constructing a word embedding space and calculate the centroid of each topic and fetch the 10 nearest neighbors of the topic-centroid.

The uploaded pngs depicted the centroids and the 10 nearest neighbors of each.

I believe from the final depictions we have a more accurate idea about the complaints of these employees, hence we have more fine-grained topics than the initials-Genre.

(I also work with kmeans in a small part of the given code)