To get started with the analysis, some basic statistics and visualizations have been created to get some insights to the data. We saw that few characters have a great part in terms of how often they speak. In addition, we could see that some words (like thou) and similar words "like" … are dominating Shakespeare plays.

In the second step, it was important to get some feeling for the genre, plays and roles. Therefore, a PCA and LSA analysis have been done based on the plays and its genre. The main questions we tried to answer were:

- Are the plays written differently in terms words they are using?
- Can one clearly differentiate between the four genres?
- Have all plays been written by Shakespeare or are there plays which are completely different from the others?
- How different are the plays as seen in terms of Networks?
- Using Clustering, trying to cluster the plays across all genres and trying to find what similarity does the plays and genres have with each other?

Through this approach one could see that Comedies and Tragedies are quite next to each other (which was indeed quite surprising) but Poetry and History are clearly different in terms of the words the genre are using. Therefore, one could differentiate between Genre and plays. In addition, to try to answer the question if all plays have been written by Shakespeare, we could observe in the LSA that "The Merry of Wives of Windsor" is kind of an outlier. Therefore, one can assume that this play was not written by Shakespeare.

Second step was to go deeper and do a role analysis and try to answer following questions:

- Can one differentiate between female and male?
- Can one identify characters which differ significantly from the others?

Doing a simple PCA, one could see, that Male and Female are not quite differentiable.

In addition, we wanted to see, whether the characters are different in terms of the words they use. One could see, that there are 10-15 characters, which are all men, powerful and influential characters in the plays.

Furthermore, in the next section we wanted to see, whether one could do classification and recognize speeches and tell which genre they are. A LDA showed 66% precision which is quite good. However, for the characters a precision of only 38% have been reached, which is obviously more difficult. However, considering the amount of plays (37) this is quite good.

Let us turn our attention to Network Analysis. For the initial analysis, we have plotted "Number of total Actors/Players" against each "Play" and "Total number of lines being spoken in each play" vs "Play". Some interesting things that came forward were that the play "Hamlet" had approximately 38 actors which was compared to 50 percent of other plays but the total number of lines spoken by the "Hamlet" actors was highest amongst them all, a total of 4200 lines. We also calculated the individual number of lines spoken by all the actors across all the plays and plotted which brought good insights.

A Graph of the all the 36 plays and a total of 971 actors/players as nodes along with 1328 edges between them calculated with an average degree of 2.7353 is plotted. Upon calculation of "Degree Centrality" and "Page Rank", the character "Messenger" is the most influential amongst all of the characters and appears in a total of 22 connected nodes to it. Similarly, the play "Richard III" is the most influential amongst all of them with a degree centrality of 0.073196 and Page Rank score of 0.023528. It is connected to a total of 71 nodes/neighbors.

Next we turn our focus on Clustering using K-means. Firstly, all the plays belonging to different genres like Comedy, Poetry, Tragedy and History were plotted individually as well as all combined on the basis of a dissimilarity score from which one obvious result came out that Poetry is not at all similar to the other genres and hence was present at the far right side of the plot.

There were certain outliers that came forward which were Cleopatra and Macbeth for Tragedy, Sonnets and VenusAndAdonis for Poetry, HenryV for History, and Cymbeline for Comedy.

After this, we did K-means clustering with K=20 at first on a particular poem "VenusAndAdonis" to see the words collected in 20 different clusters. Further, using all data with K=10, again k-means clustering was done and we could see the plots of various genres but to analyze the genres efficiently, a Dendrogram was plotted and the result displayed. As a whole, it has two primary clusters where one cluster has Poetry genre and the other one has all of the other genres combined which actually makes sense as the poetry is written in a very different way than other genres.

Furthermore, the comedy and tragedy genres are overlapping and are near to each other for most of them while history genre is little away from and in its own cluster.