**PLAY-BY-PLAY, PLAYER-BY-PLAYER – COLLEGE BASKETBALL ANALYTICS**

**Exploratory Data Analysis:**

The dataset contains 61,061 records and spans across 66 columns. During our analysis, we found that the original column names were somewhat complex and likely required specialized knowledge to interpret. To address this, we renamed these columns for clarity and ease of understanding. Our exploratory data analysis involved the creation of temporary dataframes, which helped us to delve into specific aspects of the data. We adhered to standard data cleaning practices, including identifying and handling null values, exploring unique values, and engaging in feature engineering. These steps were essential to uncover and address any intricacies in the dataset that could potentially complicate further processing.

**Pre-processing:**

In our assessment, we first focused on cleaning the 'Pick\_drafted' and 'year\_of\_college' columns. This involved checking for unique values in the 'Pick\_drafted' column for drafted players and examining the 'year\_of\_college' column for unique and null values. We decided to drop rows with specific values ('None', '0', '57.1', '42.9') in the 'year\_of\_college' column to ensure data consistency. Additionally, rows with null values in the 'year\_of\_college' column were dropped, ensuring that the dataset was cleaned and ready for further analysis in the context of college basketball player data.

Next, we addressed irregularities in the 'num' column. We dropped the 'ht' column as irrelevant and then examined the 'num' column to identify irregular entries such as 'None', '--', '23B', '4A', and '31/24'. These irregular values were replaced with NaN to standardize the data. Furthermore, missing jersey numbers were assigned NaN values based on the available numbers from 1 to 100 per team, ensuring a more uniform and accurate dataset for further analysis. We also managed null values in various columns. This involved replacing null values in the 'Recruit\_Rank' column with 0 and in the 'Assist / Turnover ratio' column with the median. Rows with null values in columns ranging from 'Defensive\_rating' to 'Points' were dropped. Additionally, we checked for null values in another subset of columns from 'Rim\_shots\_made' to 'Dunks %' without taking any specific action. These operations ensured that the dataset was cleaned and ready for further analysis, contributing to a more comprehensive understanding of the college basketball player data.

Finally, we further refined and standardized the dataset. We examined three columns ('Role/Position', 'conference', 'Type') to identify unique values and dropped the 'Unnamed: 65' column as it could not be interpreted. Missing values in 'Rim\_shots %', 'Midrange\_shots %', and 'Dunks %' columns were filled with 0. The 'Type' and 'num' columns were dropped as they did not provide meaningful insights. Unique player values in the 'player\_name' and 'Player\_ID\_(recurring for each year)' columns were identified, and a cleaned DataFrame was created by subsetting performance and player detail columns. The dataset was then standardized using StandardScaler from sklearn.preprocessing, ensuring consistency and accuracy for further analysis in the context of college basketball player data.

**Analysis Plan:**

Our objectives are to find the link between performance metrics and whether a player is drafted. This will tell us if we can search for certain performance metrics that signal whether a player is drafted, telling us about the types of players that make it professional or the type of metrics that are prioritized by scouts.

To achieve this, we used the K-Means Clustering algorithm, PCA, and t-SNE methods to find out our optimal number of clusters, ideal features, and to explore a dataset that already has many features.

**Preliminary Results:**

Moving ahead with our analysis, we performed PCA and found the top 14 components from our PCA model which explains 90% of the variance in our data, through which we created a dataframe for the threshold to compare with regular variables dataframe.

Secondly, the accuracy scores show that the Logistic Regression model with the top 14 PCA components performed slightly less effectively than the model using all variables. However, the confusion matrix revealed only one misclassification in the PCA-based regression. This minor performance drop is expected due to some information loss in PCA, and it doesn't necessarily imply inferiority of the model.

Lastly, when evaluating how well our clusters can predict draft status, we found that PCA-based hierarchical clustering was marginally more effective than KMeans clustering using the top 3 original variables. This suggests that there is not a clear separation between what makes a player get drafted to the NBA or not when simply clustering based on all of our variables in the dataset.

**t-SNE:**

In order to conduct t-SNE analysis, we had to standardize the numerical features while leaving our outcome variable in the dataframe. We then defined our categorical variables and used the standardized dataframe to run our t-SNE analysis. Through t-SNE, we were able to find two distinct clusters visually. Because our target variable “Pick\_drafted\_binary” has two outcomes, setting “n\_components” to 2 made the most sense so we can observe the two different clusters on a 2-dimensional setting. Our scatter plot divided into two separate clusters, one in which the outcome variable was drafted and the other was not drafted. The sizes of the clusters make sense because the amount of players drafted is much less than players not drafted. The separation of the clusters based on the target variable outcome indicates a good predictive performance of our t-SNE model.

**Sentiment Analysis:**

From our sentiment analysis, we can conclude that the textual descriptions of the players drafted into the NBA do not have strong correlation with what number they were picked in the draft. Furthermore, we can see that it is very difficult to predict the grade and position of the player based on the textual analysis that we have in our dataset. However, this makes sense realistically because there is often much more that goes into why a player is drafted before others and these textual descriptions of the players are also just predictions of how NBA analysts think a college player will do in the NBA.

The actual teams that are picking the college players have individual reasons for choosing them and often it depends on what kind of player they need (maybe more defensive or a certain position), which means the best prospects are not *always* chosen first or early in the draft. There are also many intangible aspects of players that cannot be measured in their stats (such as player potential) and although the textual analysis aims to look at these other intangibles, NBA analysts do not make perfect predictions which makes it hard to create a comparable sentiment between players based on these texts.

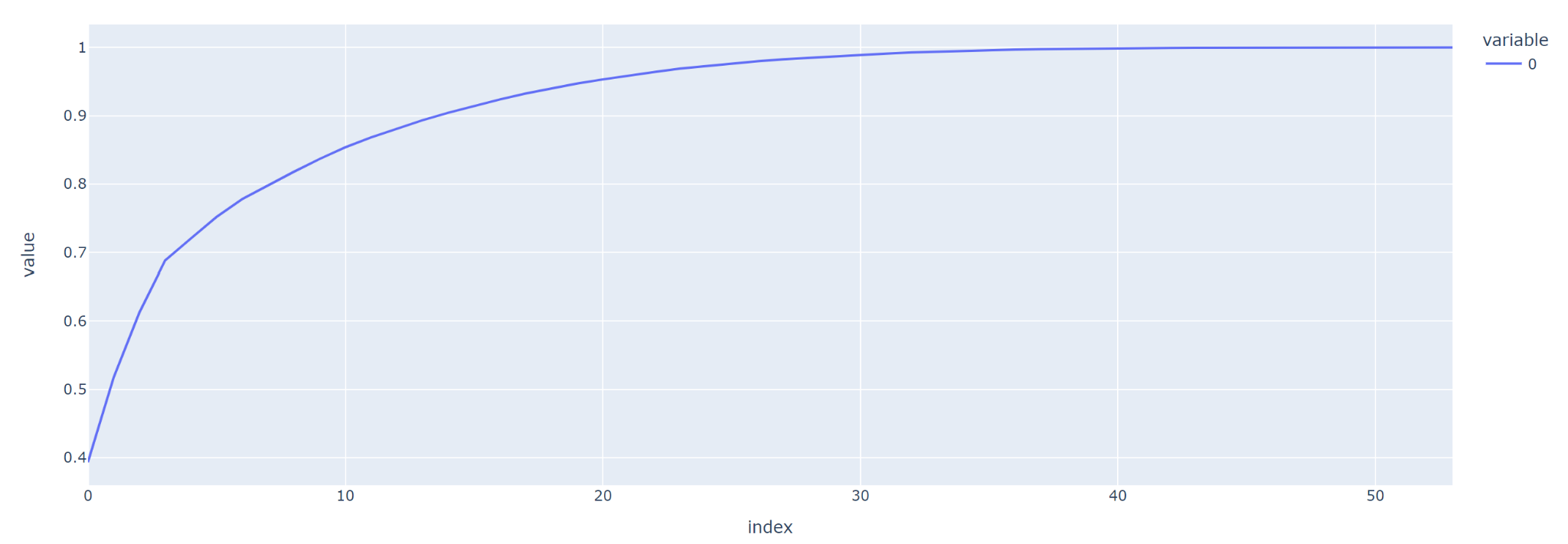
In addition, when examining the texts, we notice that there are many comparisons to past NBA stars which means that our model would need to understand the history of basketball and how these players that were mentioned played/succeeded in the NBA to be able to take this sentiment into account. We were unable to do this in our analysis, but this is something to look at if we were to continue with next steps.

**Conclusion:**

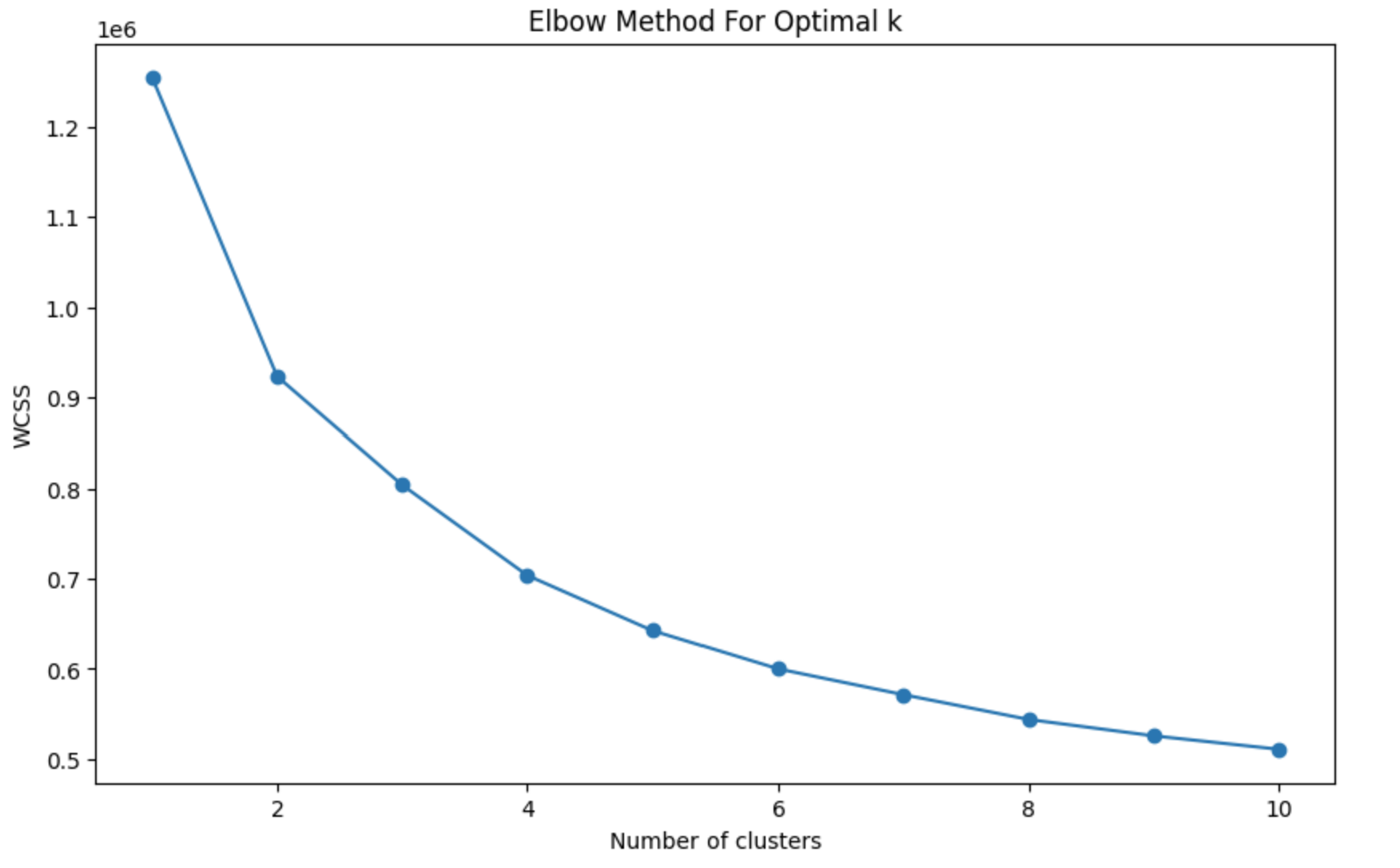
* NCAA data has no clear performance indicators for whether a player gets drafted
* The average drafted player stands out in offense vs the average NCAA player
* Sentiment can only be tracked with multiple years of data
* Dataset can be used for tracking general performance trends in college ball, but not used as a benchmark for whether they go pro

**Appendix:**

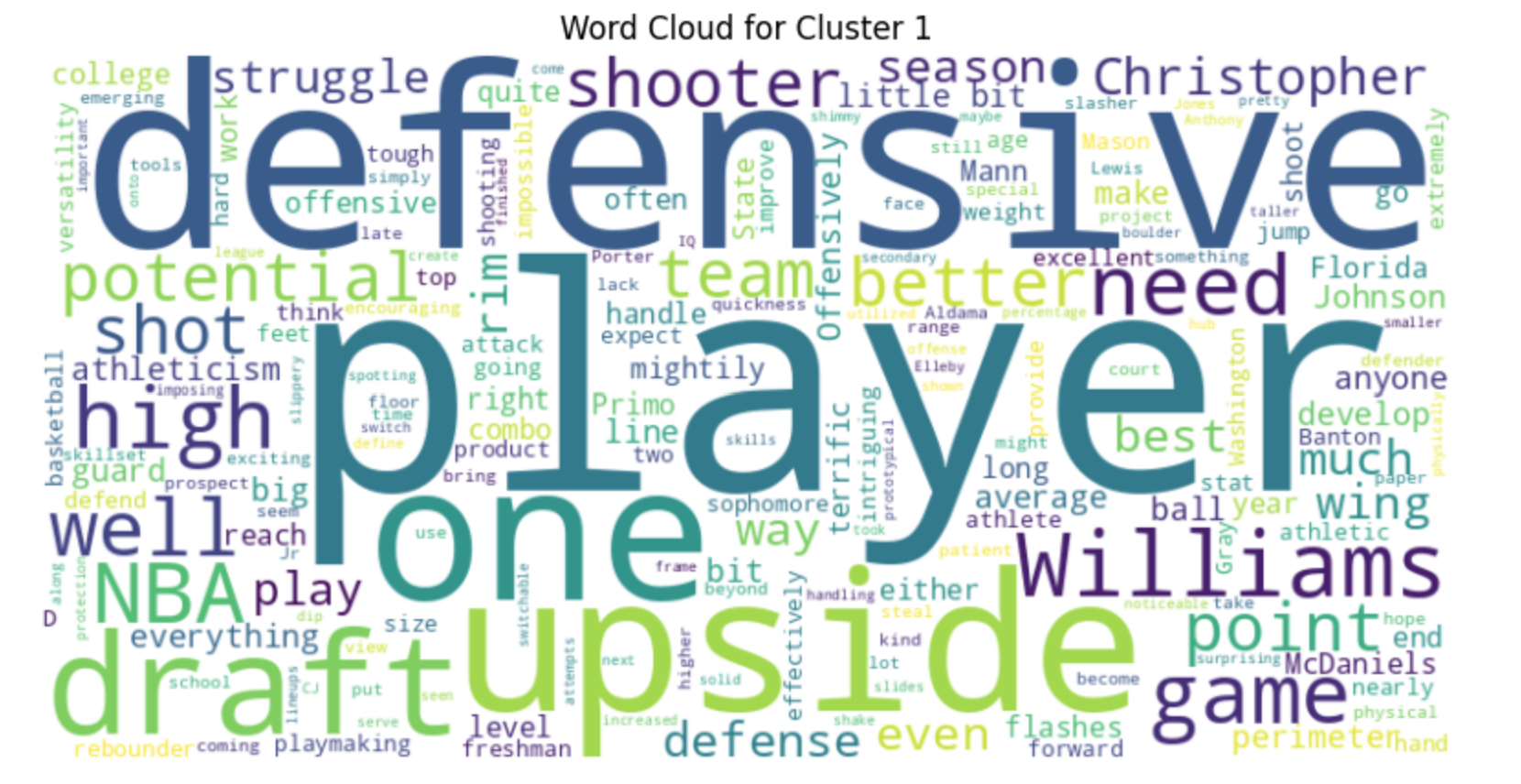
**Scree plot for PCA model:**

****

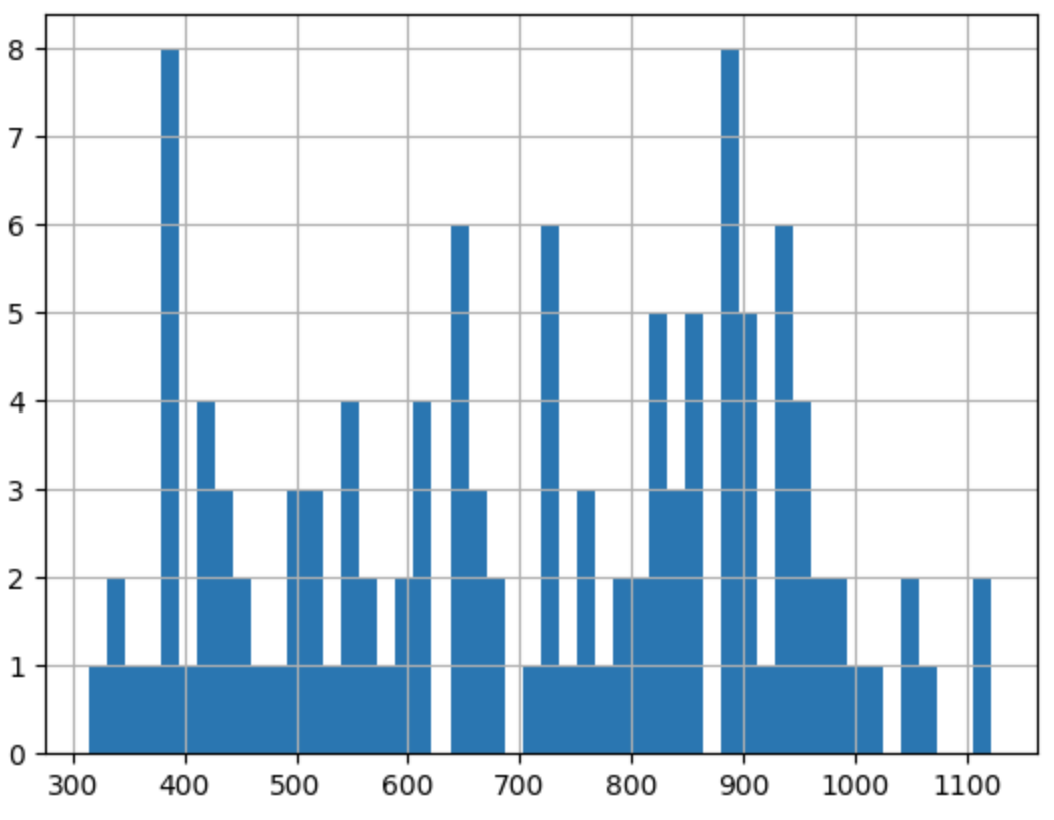
**Elbow Method For Optimal k for Clustering:**



**WordCloud of our clusters to see which words appear the most frequently in our clusters:**



**Histogram of the Instant Analysis of the players:**

****

**Link of Dataset:**

**[1] College basketball dataset:**

**[a]** [**https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset**](https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset)

**[b]**[**https://github.com/zackbatch/BA820-Project/blob/main/CollegeBasketballPlayers2009-2021.csv**](https://github.com/zackbatch/BA820-Project/blob/main/CollegeBasketballPlayers2009-2021.csv)

**[2]NCA website:**

[**https://github.com/zackbatch/BA820-Project/blob/main/Instant%20Analysis%20Draft%3B%202020-2021.csv**](https://github.com/zackbatch/BA820-Project/blob/main/Instant%20Analysis%20Draft%3B%202020-2021.csv)

**Contribution:**

| **Name** | **Coding** | **Theory** |
| --- | --- | --- |
| Varun Kaza | Exploratory Data Analysis | Report & Analysis in colab |
| Jaishankar Govindaraj | Preprocessing | Report & Presentation |
| Priyanka Chaudhari | Clustering | Report & Analysis in colab |
| Zachary Batchelder | PCA | Report & Analysis in colab |
| Audrey Sellers | t-SNE | Report , Presentation & Analysis in colab |