Restaurant Prediction System using R analysis

| Asst.Prof : Sang-Won Bae  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [sbae4@stevens.edu](mailto:sbae4@stevens.edu) |  | TA: Harsh Jain  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [hjain6@stevens.edu](mailto:hjain6@stevens.edu) |
| --- | --- | --- |

| Neha Patade  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [npatade@stevens.edu](mailto:npatade@stevens.edu) | Monil Shah  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [mshah83@stevens.edu](mailto:npatade@stevens.edu) | Sushant Ghorpade  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [sghorpad@stevens.edu](mailto:sghorpad@stevens.edu) |
| --- | --- | --- |

| Swaraj Kate  Stevens Institute of Technology  Hoboken, NJ 07030, USA  [skate@stevens.edu](mailto:skate@stevens.edu) | Niraj Chauhan  Stevens Institute of Technology Hoboken, NJ, USA  [nchauha4@stevens.edu](mailto:nchauha4@stevens.edu) | Sushant Ghaytadak  Stevens Institute of Technology Hoboken, NJ, USA  [sghaytad@stevens.edu](mailto:sghaytad@stevens.edu) |
| --- | --- | --- |

# ABSTRACT

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

*CHI 2020, April 25–30, 2020, Honolulu, HI, USA.*

© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-6708-0/20/04...$15.00.

DOI: <https://doi.org/10.1145/3313831.XXXXXXX>

\*update the above block and DOI per your rightsreview confirmation (provided after acceptance)

The goal of this project is to analyze and provide prediction results to an India based client who wants to invest in the food and beverage industry. Investing in a restaurant is a successful business as it provides maximum profits. Our client wants to open a restaurant in Bangalore city, and we will help him by providing analysis of different restaurants in Bangalore which are located at different locations by using R analysis. During R analysis, we have performed data manipulation, sentiment analysis, wordcloud and network analysis to provide basic graphs and proper visualization of graphs. We have also provided an interactive dashboard to enhance the graphs provided to the client. With the help of this visualization, we will help our client to set up his restaurant in a location which will provide maximum profits.

## Author Keywords

Restaurants; Barplot; Treemap; Scatterplot; Boxplot; WordCloud; Sentiment Analysis; Network Analysis; Interactive Dashboard.

# Background

Bangalore is the home to India’s IT hub with some of the best international and domestic technology. With the increase in companies related to the IT sector, a greater number of IT employers start to reside in this city. Due to this reason, restaurant owners have gained maximum profits by opening multiple restaurants. Hence one of our client wants to invest in the food and beverage industry by opening a restaurant in a location which will provide him maximum profit. We are going to help him by predicting and analyzing the performance of restaurants in Bangalore city through graphs and interactive dashboards.

# Objective

Our objective is to predict and take decisions based on data analysis and visualization for a client who wants to invest in the food and drink industry.

# Methodology

For this project, we would be analyzing the restaurant dataset which we acquired from Kaggle. This dataset consisted of some data which was not required. Hence, we had to first clean the dataset. Initially, we have imported the dataset in our environment and went through the summary and the structure of our datasets. The summary will give the basic mathematical information on each column and inform us about the missing value. Structure defines the structure of each variable like character, integer, numeric, date, etc. Here, we have filled the missing values with the mean. For factor type data, we have replaced the missing values with respective values. Data cannot be analyzed if it’s not cleaned properly as it may give faulty results. We have used dplyr and tidyse library for data cleaning and manipulation.

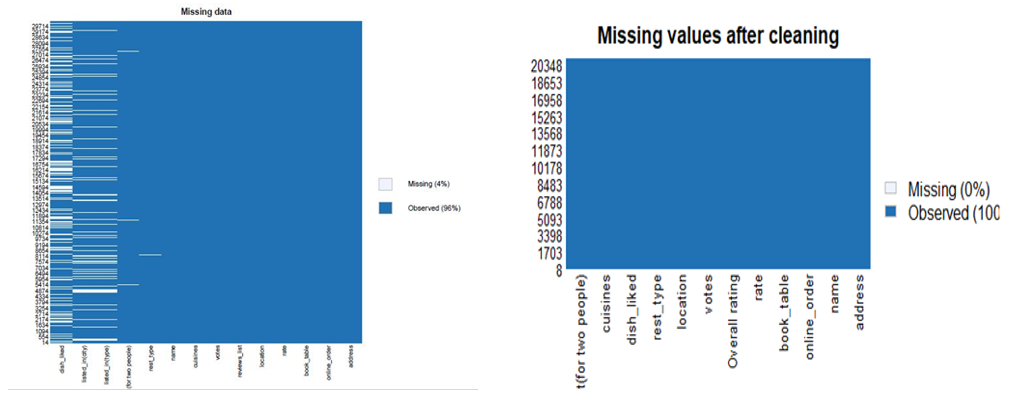


Figure 1:Missing values plot

After data cleaning and manipulation, we made use of libraries such as plotly and ggplot, to plot bar graphs for the count of online orders, top 10 best restaurants, top 10 locations for restaurants. After plotting the basic graphs, we made use of libraries such as dplyr, tm, and wordcloud to plot wordcloud of liked dishes from restaurants and individual wordcloud for the top 10 restaurants. We created the corpus and made us of tm library for text cleaning and analysis. During the text cleaning, we made sure that we removed all the punctuations, numbers, stopwords (and, or, the, was, for etc), and whitespace. We then plotted sentiment analysis for the reviews of the top 10 restaurants by using the function get\_nrc\_sentiments(). After sentiment analysis, we performed network analysis by using Gephi software to plot the neural network. For network analysis, we used the column reviews to see the cluster of words coming together. By using the Google Application Programming Interface (API), we extracted the latitude and longitude from all the addresses of the top 10 restaurants. This helped us to plot the geographical locations of the top 10 restaurants. After plotting all the necessary plots and graphs, we created an interactive dashboard using FlexDashboard which consists of all the interactive graphs and plots essential for the analysis and visualization. We created this dashboard to provide an easy way for our client to understand the graphs and plots.

# Results and Analysis:

Initially, we plotted a bar graph for the count of online orders. Figure 2 depicts the same

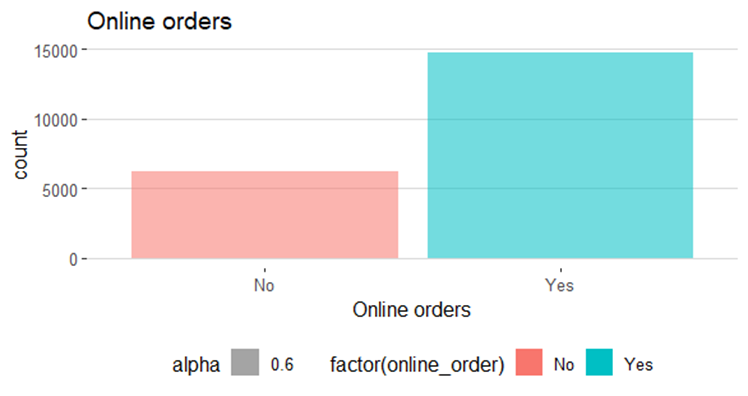


Figure 2: Bar plot for count of online orders

This plot helps us to understand that people mostly prefer to order food online rather than going out to restaurants. Hence, we can ask our client to focus more on online orders rather than decorating the interiors for dine-in.

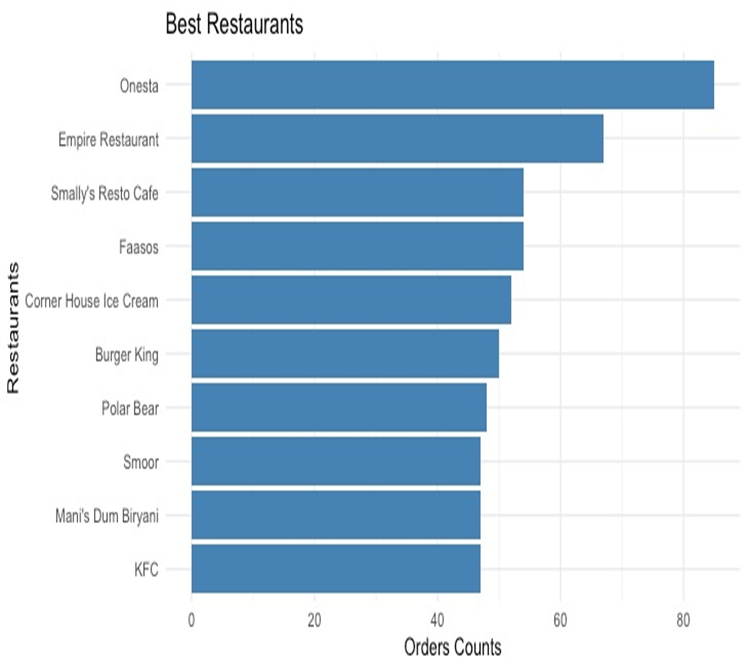


Figure : Bar plot for top 10 restaurants in Bangalore

Figure 3 depicts the bar graph for top 10 restaurants in Bangalore. This plot was extracted based on order count. This will help our client to understand the competition which he will face when he will set his own restaurant.

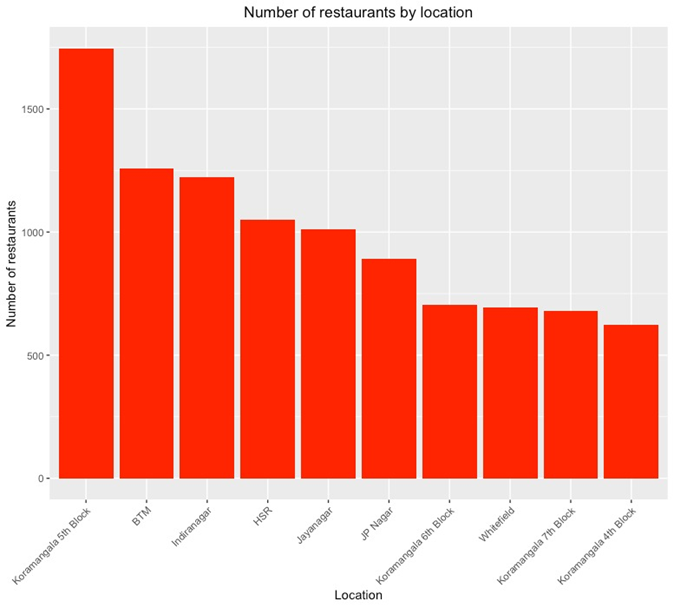


Figure : Bar plot for top 10 locations for restaurants in Bangalore

Figure 4 depicts the bar graph for top 10 locations in Bangalore. This plot could recommend our client to select a specific location, depending on the competition and the area in which our restaurant could acquire maximum profit and have more customers.

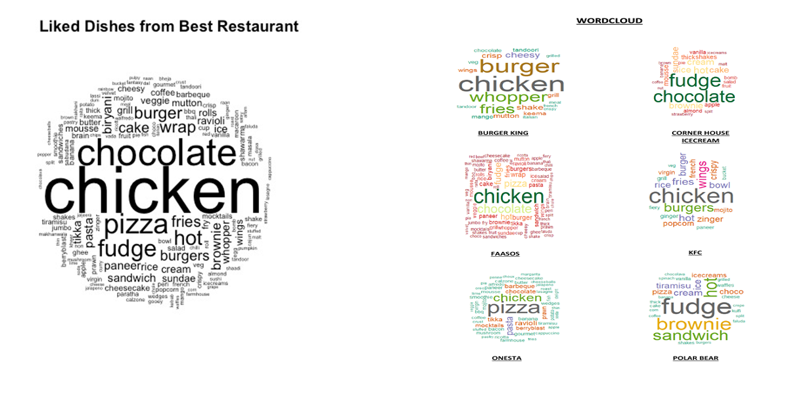


Figure : WordCloud for liked dishes from top 10 restaurants

Figure 5 provides two wordclouds, first one depicts liked dishes from the top 10 best restaurants while the second one depicts the individual wordcloud for top 10 restaurants. As seen from the wordclouds, chicken, chocolate, burger, pizza etc are the most like dishes. Based on this wordcloud plot, we can recommend our client to put these dishes on the restaurant’s menu list.

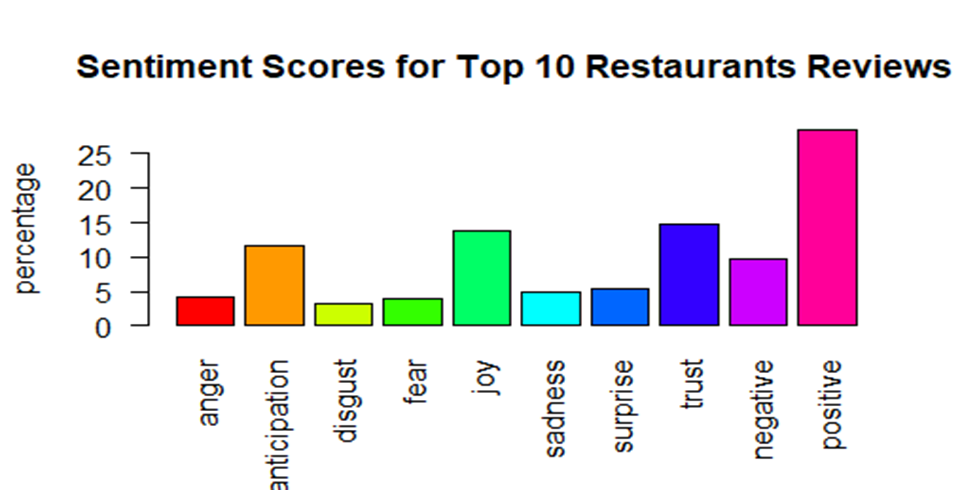


Figure : Sentiment analysis of reviews of top 10 restaurants

Figure 6 provides the sentiment analysis for the top 10 restaurants. Sentiment analysis is acquired from the customer reviews for the top 10 restaurants. Using this sentiment analysis, we can successfully quantify the sentiments of each customer. This can help our client decide which type of restaurant should they pursue, and which kind of restaurant is likely to be successful.



Figure : Network Analysis plot

Figure 7 depicts the network analysis which is used to plot the neural network. We have made use of the Gephi software to plot network analysis. To plot the network analysis, we have made use of the column reviews to see the cluster of words coming together. By implementing this, we can understand that there are 3 main hot topics between people which are Food, Place and Good.

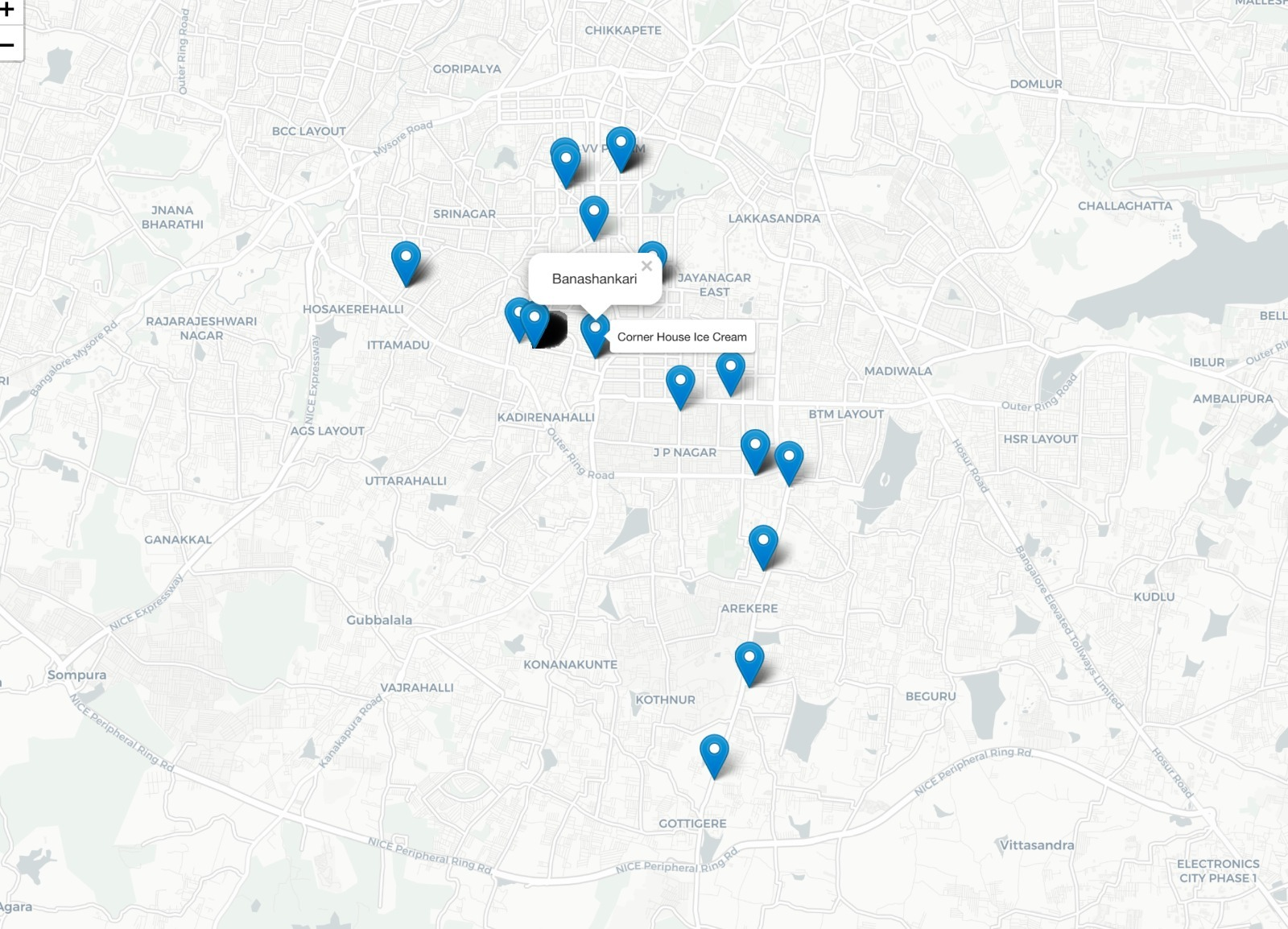


Figure : Geographical plot of the top 10 restaurants.

Figure 8 depicts geographical locations of the top 10 restaurants. To plot these locations, we have made use of the Google Application Programming Interfere (API) which helped us to extract the latitude and longitude from the addresses of the top 10 restaurants. As we hover on the marker, it provides the location name and as we press on the marker, we get the restaurant names for the same.

# Machine Learning:

With the help of machine learning algorithms, we have predicted the location as our target variable with respect to the rest of the columns. We have performed this machine learning on python. For which we have used various libraries such as sklearn and matplotlib to plot our findings.

To perform machine learning we have used two data sets which is textual data and numeric data.

Pre-processing:

We have performed one hot encoding to convert our data to numeric to get tangible results with our data set. We have plotted the correlation after converting the data to numeric.

A screenshot of a video game

Description automatically generated

Figure 9: Correlation plot

As test data we have included various columns such as rates, votes, location, votes etc. and using the test data we have the run our train data to calculate the accuracy of prediction. After having done the preprocessing we will then run a few algorithms to predict the data.

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| 1. Decision Tree | 74% |
| 1. Random Forrest | 80.01% |

We see that the best accuracy was given by Random Forest and we have plotted the predicted vs actual graph.

A close up of a mans face

Description automatically generated

**Figure 10: Prediction plot of Actual vs Predicted.**

Pre-processing:

For textual data preprocessing we have cleaned the data of various punctuation and stop words and the we have run our models on it.

To perform machine learning we have utilized various columns such as dishes liked cuisines, locations, book table, online order etc. We have also performed machine learning on various textual data. We’ve successfully running SVM and XG BOOST algorithm we see that the best accuracy is given by SVM model.

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| 1. Random Forest | 81% |
| 1. XG Boost | 84.03% |
| 1. SVM | 92% |

# Limitations:

1) Initially when we found our dataset, it was very big in size and very unclean. It took maximum of our disk space. Hence, we had to clean the data and make it compact.

2) We tried to plot network analysis on R, but we were unable to plot it. The nodes and edges matrix were so huge, that RStudio could not handle these matrices. Hence, we had to make use of Gephi software.

3) We did not have the latitude and longitude specifications for the address we had in the dataset. We had to make Google developer account to get developer API so that we could extract latitude and longitude for the address of our dataset.

4) Our dataset had lot of columns having characters as data points, while implementing machine learning model we need to have all the values as numeric. Hence, we had to use One Hot encoding to provide all unique numeric values to the strings (datapoints)

# Discussion:

From the results we have seen that some of the top locations for restaurants and the number of restaurants in each location. For lesser competition while starting a restaurant the investor can select a location with suitable number of restaurants at that location. Also, to have more footfall at that location we can included some of the top dishes liked in that area. For better visualization we have created interactive visualizations to better explain our story and help understand our findings.

There have been many studies and analysis performed on the data set but not many have performed a one hot encoding as well as textual machine learning on the data set. Also, we have performed sentiment analysis on the dataset to validate the ratings of the top 10 restaurants.

# Conclusion:

From the above visualizations we can infer the top 10 best locations, type of restaurant, which dishes, and cuisines should an investor select while setting up the restaurants. With the help of visualization, we have seen the top 10 restaurants in Bangalore area had a closer look at their cuisines and for capturing the attention of people/customers investor should include some of the dished which are liked the most by these top restaurants. If we consider our analysis, we can help set up a restaurant based on the amount to invest and which location to invest in.

# Acknowledgement:

It is a genuine pleasure to express our deep sense of appreciation and gratitude to our mentor and guide Prof. Sang Won Bae and to our teaching assistant Harsh Jain. Our professor helped us to provide us with significant feedback on our dataset which helped us to improve our project and our presentation as well. Our teaching assistant came up with some significant suggestions which helped us in improving our project. We would also like to thank School of System and Enterprises for providing immense support throughout our research and project.

# References:

Initially when we found our dataset on Kaggle.com

1. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press, Bloomington, IN.
2. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D Dissertation. Massachusetts Institute of Technology (MIT), Cambridge, MA.
3. Langdon Winner. 1999. Do artifacts have politics? In The Social Shaping of Technology (2nd. ed.), Donald MacKenzie and Judy Wajcman (Eds.). Open University Press, Buckingham, UK, 28-40.
4. Interactive Dashboards in R - YouTube Link <https://www.youtube.com/watch?v=_a4S4tq62OE&t=504s>
5. Interactive Dashboard using FlexDashboard - <https://rmarkdown.rstudio.com/flexdashboard/index.html>
6. ValueBox- <https://www.rdocumentation.org/packages/shinydashboard/versions/0.7.1/topics/valueBox>