

**DEPARTMENT: Computer Science**

**Module Code: Big Data**

**Academic Year: 2024**

**Submission Deadline: 20/04/2024**

**Actual Submission Date: 18/04/2024**

|  |
| --- |
| **Student Registration Number** |
| ARY23030 |

**Declaration**: *I have completed and submitted this work by myself without assistance from or communication with another person either external or fellow student or any AI type of content generator. I understand that not working on my own will be considered grounds for unfair means and will result in a fail mark for this work and might invoke disciplinary actions. This piece of assessment will be continuously checked for its academic integrity until my graduation and the mark will be revised if it is found to breach the unfair means policy. It is at the instructor’s discretion to conduct an oral examination which will result in the award of the final grade for that particular piece of work.*

**Abstract**

This research analyzes big data concepts and studies the result of big data analysis through different techniques. It shows the concept of cleaning the data and manipulating it. The aim of the study is to identify meaningful information from movie dataset through statistical techniques, understand the causes and effect of some features. This project also aims at time series data to see changes through time in different features. These sorts of analyses are carried out with the help of pyspark as it deals with large amount of data and it also used visualization tools such as Tableau which will help to find important trends in the data. You can find the visualizations in this website of mine - <https://public.tableau.com/app/profile/davidispiryan> called BigDataProject.Top of Form

Contents

[**1. Introduction** 4](#_Toc164096058)

[**2. Data Preprocessing** 4](#_Toc164096059)

[**3. Insight Analysis** 5](#_Toc164096060)

[**4. Storytelling for Dashboards** 7](#_Toc164096061)

[**5. ML models** 8](#_Toc164096062)

[**6. A\B Testing** 9](#_Toc164096063)

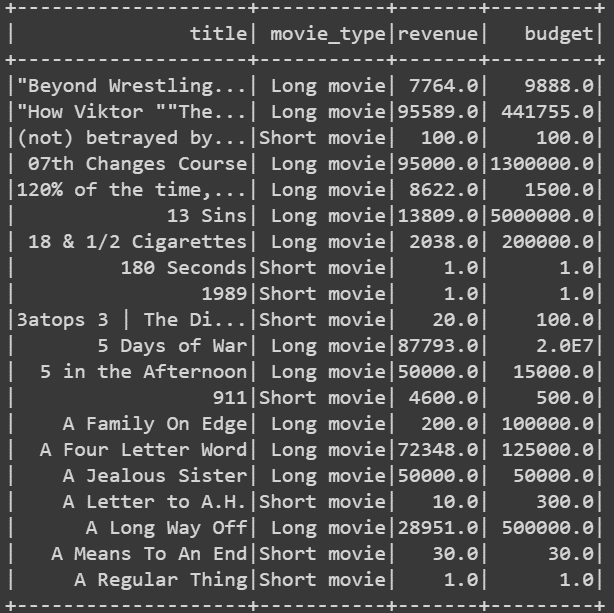
[**7. Conclusion** 10](#_Toc164096064)

# **1. Introduction**

The purpose of the project is to work with pyspark and Tabelau to determine meaningful information from the dataset which is in movie industry domain. The data itself is not so clean, hence the data undergoes a process of cleaning, manipulating and preprocessing including handling with missing and wrong values. Visualizations show trends in production companies, countries and audience engagement. Tableau offers interactive dashboards and storytelling for the analysis in time series and geographical trends including some impacts of covid pandemic in the industry. Machine learning models predict the profits of movies and do classification again predicting it with F1 score being the critical metric.

# **2. Data Preprocessing**

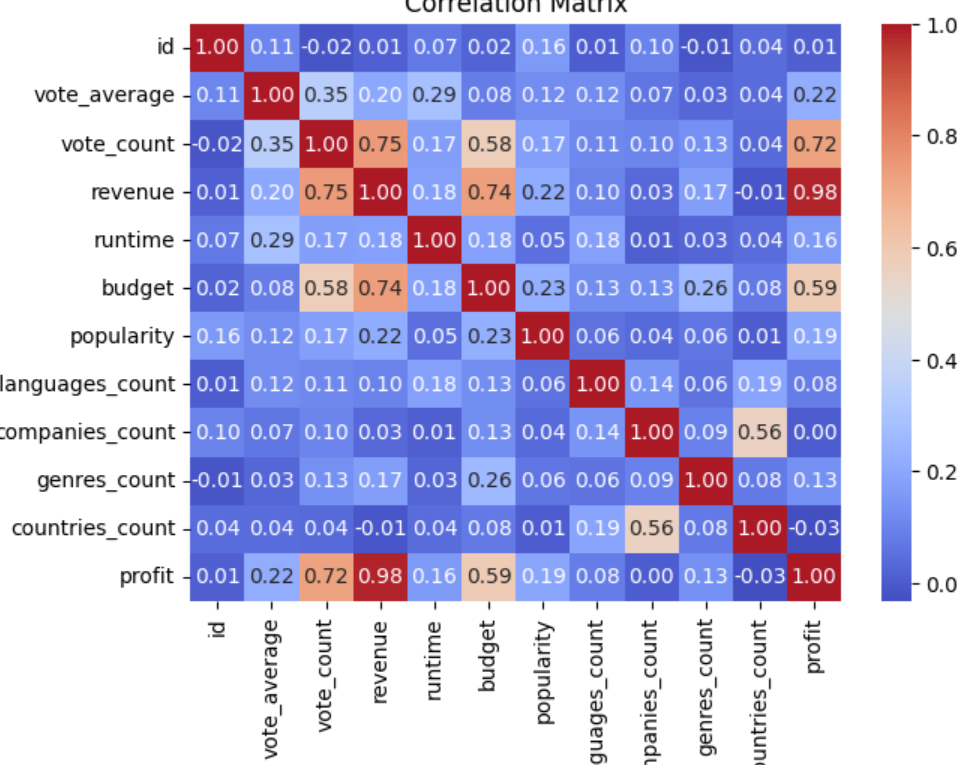
To start with, it is a must to have a scheme that we can use for the dataset so all features have a correct data type. After, it is important to check in the dataset if there are missing values or any blanks. As there are lots of missing values and a lot of blanks in each of features, firstly blanks are replaced with Nulls and then they are all removed. If only missing values of budget and revenue are removed that data will have 10,000 observations, and as this report is mostly about the profit and revenue, we can remove all of the data because the observations will almost be the same. It is necessary to check if some features have negative values, because if they do, they shall be removed as feature - budget, popularity cannot have negative numbers. Some features that have 0 values shall also be removed as that means that data is unknown for that values. After dealing with unnecessary data, some features like genre, production\_companies, production\_countries and spoken\_languages are transformed into new columns with the first value in a cell they had. This is done to understand the main genre, company, country and language. Also, the columns that had more than one value in a cell are transformed into such columns that can tell count of how many genres does the movie have or how many companies were producing on the movie. Regarding the outliers, it is obvious that there are some very strong outliers but it is not removed as there are some movies that have very huge popularity and so they have much revenue and therefore they had more budget. At the end only distinct values of titles and id variables are kept, and the duplicated ones are removed. In order to understand the data, it is important to see the descriptive statistics and whether there are some type errors or misleading data. As shown in Fig1 there are lots of budget and revenues that are nonsense, such as having a value 1. Of course, the data may not be 100 percent correct and that is why all the values for budget and revenue that are less than 100,000 shall also be removed to avoid possible unclear data. It is removed 100,000 as film budget shall at least be couple of times more than this number but until 100.000 there are a lot of uncertain data. After removing those values, the data won’t be so skewed. Also, there is a column created to understand profits and losses of each movie [1], [2].



*Fig.1: Filtered data having budget or revenue less 100,000*

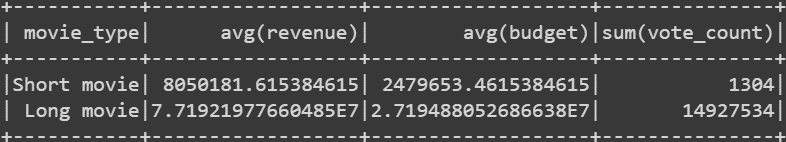
# **3. Insight Analysis**

To analyze important insights from the data, first it is important to identify the correlations between variables. As shown in Fig2 revenue with vote\_count and revenue with budget have the highest positive correlations. After, descriptive statistics of these 3 variables are calculated to get a better insight in the code.



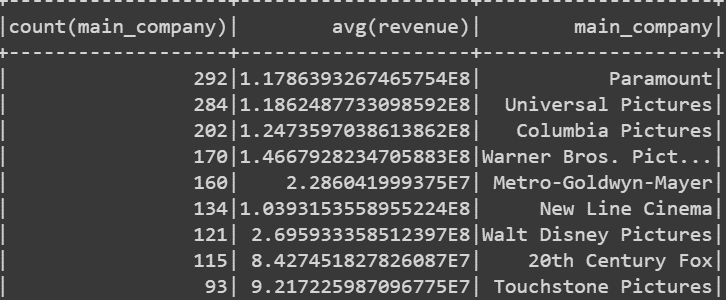
*Fig.2: Correlation matrix*

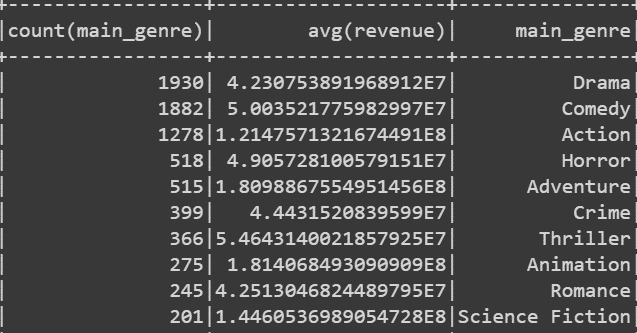
As the movie length is spitted into 2 types, it is necessary to find the revenue and budget for each of types. And as it shows, average revenue for long movies is 77,000,000 almost having 10 times more avg revenue than short movies which is 8,000,000. Average budget for long movies is also almost 10 times more than the budget of short movies which makes sense. Also, very few people vote for short movies which partly means that few people watch short movies – (Short movie < 45 minutes).



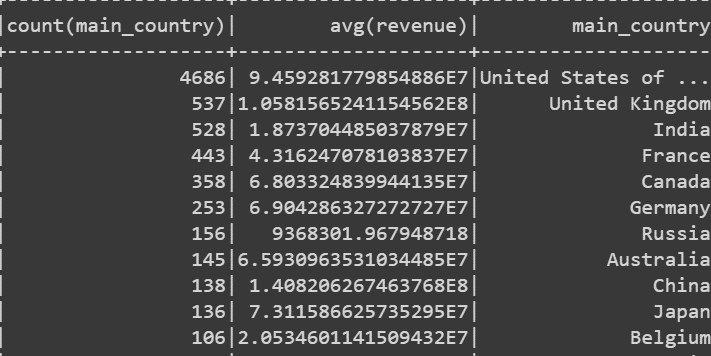
*Fig.3: Number of votes for each movie types and average of revenue & budget*

For movie market it is important to understand the best company that produces movies, how many movies were created by that company, country. Also, identifying the revenue or profit is necessary to know which genre had the highest revenue from all companies or countries. It is clear that Paramount pictures produced the greatest number of movies and the revenue is one of tops as well. Also, drama genre has the greatest number of movies having a great average revenue. Finally, it is obvious that USA produced most of the movies with a huge average revenue. All this is shown in Tableau in map visualizations. In order not to put each table here, you can find it in code that the highest revenue the company made was Syncopy, and the highest average revenue came from animations. The original language in almost 79 percent of movies was English. Based on these tables, people mostly like to watch drama, comedy and action movies as they produced more movies way more than other genres. Additionally, it is important to know what movies had the greatest profit, after making the movie, which is shown below. As there are more than one companies or genres included in each movie it important to mention that entire profit from movies was from those movies who had 3 genres. It also shows that the most profit came from 2 and 3 companies working together [3], [4].

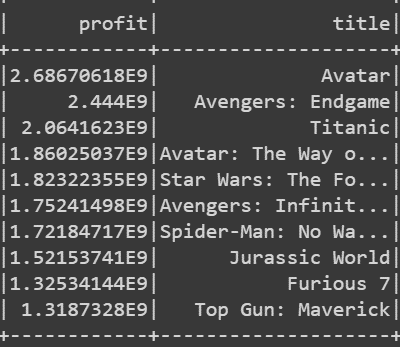


*Fig.4: Number of movies each company produced with its average revenue*

*Fig.5: Number of movies each genre has with its average revenue*



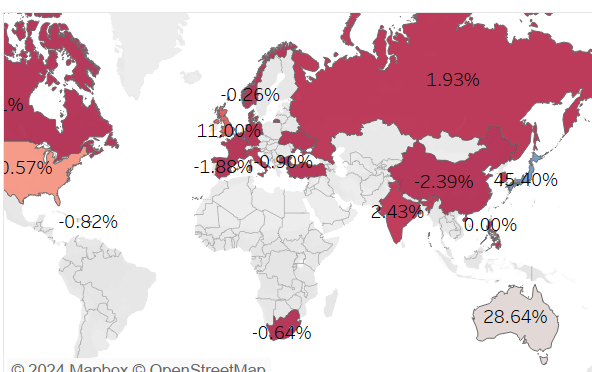
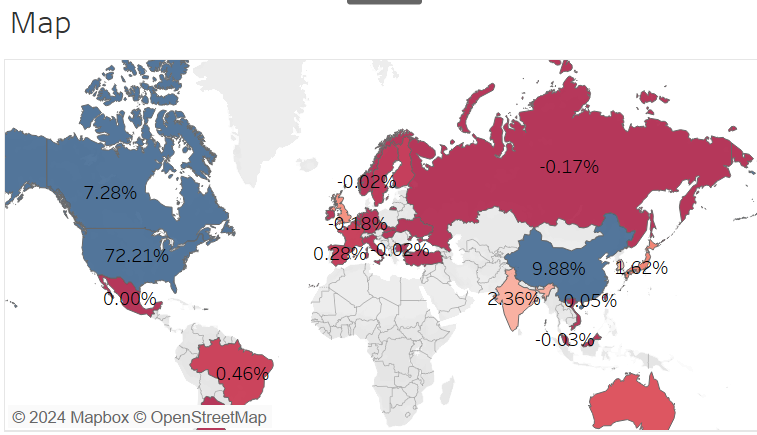
*Fig.6: Number of movies each country has produced with its average revenue*



*Fig.7: Top 10 movies with the highest profits*

# **4. Storytelling for Dashboards**

The section is provided with the Tableau file with all visualizations. It differs from other libraries from python as it has a variety of types for visualizing the data. As it is seen, there is a map and time series visualization which later was built into an interactive board that by clicking on a specific year, the data shown from all boards shows the exact data for that year. The same way a country can be chosen and it will show all information that a specific country has. In the map visualization it is clear that most of movies were made in the USA as it has the greatest number of profits from total of the world. As for the secondary large countries with largest profits, these are England, Canada, China, France and Germany having 20 % out of profits from total. It is also normal to see that in 2019 overall movie profits went down to almost 10 times because of the coronavirus and therefore less movies were released. And even until nowadays people do not make so many movies because of the pandemic and its affects which can also be seen in the tableau StatTable. But fortunately, Japan and Australia had increased overall the world profit from 0.81% to 28.6% and Japan from 1.62% to 45.5% during 2019 – 2020 as it is shown in figure 8. As a conclusion, it can be said that people do not make so many movies as they used to before 2020 and nowadays overall movies’ profits from all over the world have decreased by two times. In the dashboards it is specified red as being low profit and blue as high. Also, it clear that movie production rate started to increase very fast from 1970s and each decade it produced almost 2 times more movies according to the graph [5], [6].

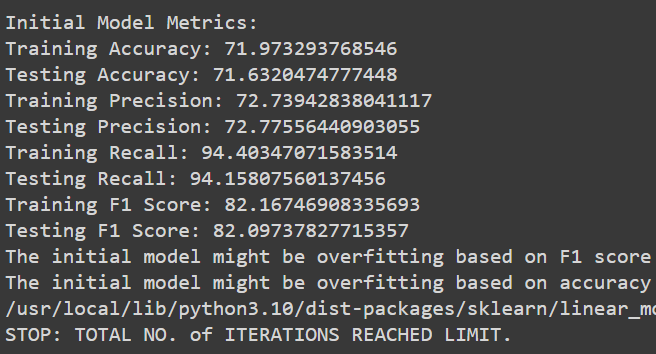
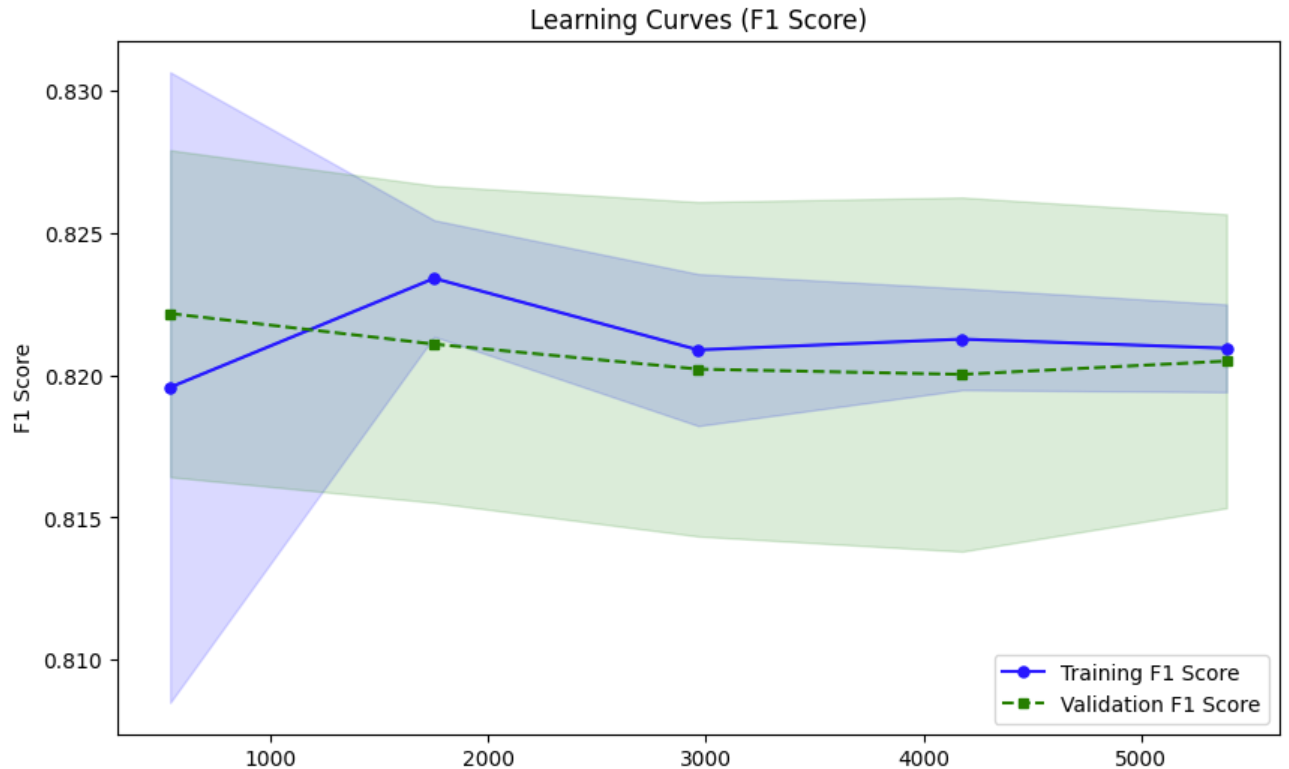


*Fig.8: Difference in overall profits in maps between 2019 and 2020.*

# **5. ML models**

In this section classification is used to understand how the model was working on profit or loss predictor. Precision shows how often model is correct when predicting the class, it shows the accuracy of positive predictions, where Recall identifies all positive instances out of all actual positive instances in the dataset. Meanwhile, F1 score is the harmonic mean of these 2 measures. For accuracy, it shows how often the model is predicted correctly, however it is not such a significant measure to use when the data is unbalanced. For this case, precision and recall and both quite high which is either overfitting or very good model. Most cases have different score of precision and recall as one of them goes up the other goes down. It will be shown later in this report [3].

It is important to understand if the model might be overfitting. In this case precision and recall are both quite high and have almost the same score, so it must be checked before moving on. By comparing the performance metrics between the training and test sets will help to assess whether the trained model is overfitting or generalizing well to unseen data.

*Fig.8-9: Checking overfitting and checking difference after regularization.*

Then it must be checked whether the regularization applied to the model effectively reduced overfitting and improved performance on unseen data. If the difference is too small than most probably regularization did not help and therefore it is not overfitting. A scalar value indicating the classifier's prediction accuracy on the test dataset is returned by the F1 score. Better performance is indicated by a greater F1 score, whereas the classifier may require more optimization if the score is lower. As the recall and precision scores are similar to each other and are quite high some would say the model is overfitting and some regularization methods should be used to better define the model. As it is seen in the figures, even with regularization method the model did not change dramatically. Therefore, it can be said that the model is very well fine-tuned and it is not overfitting [7]. As a result, it is trained and evaluated a logistic regression model with the F1 score as the performance metric. The dataset was divided into training and testing sets after undergoing normal scaling preprocessing. The model was evaluated using five folds of cross-validation, which produced the mean cross-validation F1 score. Following that, the model was evaluated on both training and testing sets and fitted to the training data to produce training and testing F1 scores. For every class, precision, recall, F1 score, and support were listed in a classification report that was created. To see how well the model performed in relation to the quantity of training examples, learning curves were plotted. In order to determine whether overfitting occurred, the training and testing F1 scores were compared, and learning curves were carefully examined for any indications of overfitting. If overfitting was found, regularization strategies, feature selection, alternate algorithms, or more data gathering might be taken into consideration in order to improve model generalization [8]. Also, linear regression is used to understand the impact on profit with different independent variables. The goal is to understand which variable have the greatest impact on the profit having normally distributed residuals.

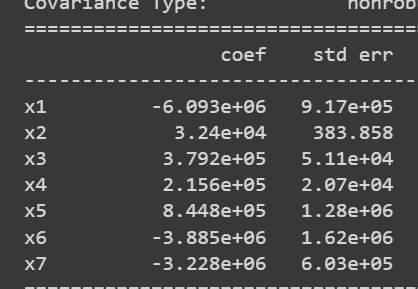


Fig.10: Impact on the model of independent variables.

x1 – x7 ('vote\_average', 'vote\_count', 'runtime', 'popularity', 'spoken\_languages\_count', 'countries\_count', 'companies\_count')

If the purpose is to fine tune the model features that have negative impact on the model like x1, x6 and x7 can be removed and the R square will be higher. After conducting regression assumption on the model, it shows that the models’ residuals are not normally distributed and the plot that shows residuals vs fitted values can approve the statement as well. But the highest impact the model has is from average vote which also has the highest std error.

# **6. A\B Testing**

After conducting T test in 2020 and 2019, it has become clear that there is a significant difference in profit in these two years. Also, comparing with the same year (2019) but having two different groups such as Russia and the USA, it can be said that there is again a huge difference in profit in these countries during COVID. The USA produced far more movies than Russia and hence did more profit. Conducting A\B testing in 2023 for different genres, one being the least popular and the other the most, it is clear that even in this case there is no significant difference in profit. Also, there is no significant difference in Marvel Studios and the 20th Century Fox companies in profit even with 5 years of difference, having 2018 compared with 2023 with different companies in each group.

# **7. Conclusion**

In summary, the purpose of the project was to find valuable information using pyspark and tableau. The dataset needed a lot of data preprocessing to have a clear dataset. It was cleaned, manipulated and transformed. Some descriptive statistics were used to understand the overall dataset. Visualizations show analysis of the countries, production firms, genres, and durations of the movies, as well as connections between the features. Tableau dashboards for storytelling included interactive visualizations that showed patterns in film output and profits across the world and years, as well as the COVID-19 pandemic's effects on the film business. A\B testing conducted shows a better understanding of difference in 2 or more groups with a specific filter. These processes make better understand the market trends in film industry through predictive modeling and insightful information.

**References**

[1] https://medium.com/bazaar-tech/apache-spark-data-cleaning-using-pyspark-for-beginners-eeeced351ebf

[2] https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[3] https://spark.apache.org/sql/

[4] Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, TensorFlow, and Keras. Packt Publishing Ltd.

[5] Tableau Software. (n.d.). Tableau: Business Intelligence and Analytics Software. Retrieved from <https://www.tableau.com/>

[6] PySpark Documentation. (n.d.). Apache Spark™ - Unified Analytics Engine for Big Data. Retrieved from https://spark.apache.org/docs/latest/api/python/index.html

[7] https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/

[8] https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-build.htm