Aspect Analysis and Visualization of Indian Unicorns

Sneha Suman
Department of Electronics and communication (AI) Indira Gandhi Delhi Technical University for Women
Kashmere Gate, New Delhi
snehasuman810@gmail.com

Kashish Saini
Department of Electronics and communication (AI) Indira Gandhi Delhi Technical University for Women
Kashmere Gate, New Delhi
sainikashish280304@gmail.com

Nonita Sharma
Department of Information Technology
Indira Gandhi Delhi Technical University for Women Kashmere Gate, New Delhi
nsnonita@gmail.com

Manik Rakhra
Assistant Profesor, Department of Computer Science and Engineering,
Lovely Professional University
Phagwara, India
rakhramanik786@gmail.com

***Abstract*— Unicorn companies have gained recognition recently. With the campaign of Startup India movement led by Prime Minister of India, India has become the hub for unicorns. The trend of unicorns flourishing in India is quite prominent and can be seen from the statistical analysis. Unicorn companies are those startups which are valued over $1 billion. The authors in this research work focusses on identifying the aspect analysis of the Unicorn numbers in India. The objective of this manuscript is to analyze the 103 unicorns that are present and successfully running in India. Further, python programming language is used in many fields, one of them is data analysis. This project was executed with the help of open-source packages and libraries such as Pandas, NumPy, Plotly express, Seaborn and Matplotlib. Plotting and analysis is designed and implemented using Google colab. Methodology used is fundamental and easy to understand. This analysis will be helpful in understanding more about the existing unicorns and about the recent bloom of unicorn startups that happened in the year 2021.**

Keywords— Dataset, Indian Unicorns, data analysis, python, packages, libraries, Pandas, NumPy, Seaborn, Matplotlib.

I. Introduction

Unicorn companies get their name for being rare, as a unicorn. These kinds of companies are tech-driven, highly innovative, privately owned and development-oriented. In the beginning of the year 2022, there were 1000 unicorns. From these 519 unicorns were a new addition, from this 44 were created in India. As of June 2022, there are 1150 unicorns over the world, 47 countries have at least one unicorn, U.S being at the top with 612 unicorns, followed by China, India with 174 and 65 unicorns respectively. Out of these total numbers, 44 unicorns were made in India in the year 2021 and 19 unicorns were born in the year 2022.

As of July, 2022 India is home for 105 unicorns, out of which numbers, 44 unicorns were made in the year 2021 and 19 unicorns were born in the year 2022. In this project, authors analyzed the dataset of 103 unicorns. Due to the recent pandemic situation all over the world, India experienced a bloom in number of unicorns. Digitalization of commerce and Smartphone's being prevalent during the worldwide pandemic also played an essential role in this. Many investors funded these companies which was one of major contributor. And by analysis of this data a lot of things can be concluded about their profit, losses, valuation and many more.

Writers opted python programming language and its packages/libraries for this work. Visualization helps in understand a dataset more profoundly. Any individual can get figure out any information easily with a visualization that would have taken a long time if done by looking at numbers in dataset. Hence, making it an effortless way of analysing data. Graphs and plot can show relations between any two variable and their distribution efficiently.

Author's objective with this data set was to analyse it for its profits and losses in the years 2020 and 2021, Operational Revenue, expenses, valuation, total fundings, head count (employees that work there), the year they were founded in, the year in which they attained the statues for being a unicorn start-up, and their age. For these analysis graphs are plotted and analysed.

The dataset was taken from Kaggle and imported using the package, Pandas on Google Colab file. After importing it, general data was obtained like its information, shape, number of columns and rows, datatype, non-null count, etc. The data was not clean and contained missing values. Those null values were handled by dropping them. For verifying this, sum of all null values is found, column wise, which comes out to be 0 again implying absence of null values Pre-processing of data was done at this stage. Statistical analysis is done by finding mean, minimum and maximum values for the individual columns. Visualization is done by plotting univariate and multivariate graphs with the help of Seaborn, Ploty Express and Matplotlib.

II. Proposed Methodology

Following steps are followed in order to perform the aspect analysis. The detailed visual representation in the form of a flow chart is given in Fig.1. Various steps are elaborated as under:

A. Data Collection:

Authors collected a dataset about Indian unicorns from Kaggle and imported it in a Google colab file by using crucial package called Pandas. Seaborn, matplotlib are used for visualization. These all packages and libraries are imported in the initial programs for smooth execution of the code. The authors uploaded the csv file in the Colab file and copied the path for importing dataset. Followings are the packages and libraries used in this analysis[1]

- NumPy: is a library for mathematical operations and used for arrays. It's a free open-source project.

- Plotly express: for visualising a variety of data along with making it detailed by changing the colour, size and other parameters.
- Pandas: is a package used for analysing and importing data.
- Matplotlib: library used for plotting and has its numeric extension as NumPy
- Seaborn: is a library used for making statistical graphics. It's based on Matplotlib and integrable with data frames of Pandas.

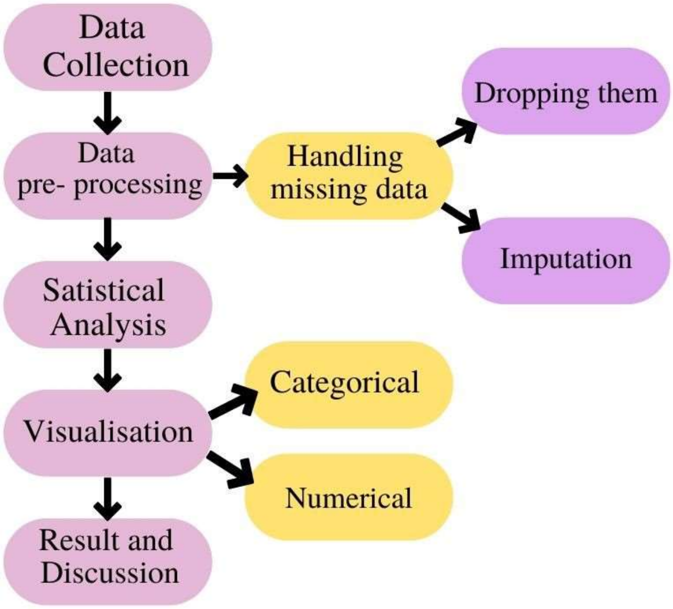

Figure 1 Flow chart representation of methodology

B. Data Preprocessing:

The data obtained is raw in its form and need to be pre-processed before any further analysis. All the columns are dropped which are not required for the actual analysis, which includes all the columns that do not have numerical values. As non-numeric value cannot be analysed. Inbuilt Functions are used for getting generic information about the dataset including number of rows and columns, data types, shape of data frame and extracting initial and last rows of data set.

Authors handled the null values present in the data by dropping them. The inbuilt function is applied for dropping all the rows that had a greater number of null values. For cross checking this handling, the sum of all the null values that are present in the data set is found by using an inbuilt function, again implying the absence of null values in the dataset. The data is now, clean and free of null values and hence further steps are implemented on it for analysis.

After removal of the null values, shape of the dataset changed, resulting in the following changes: -

- Number of rows = 60
- Number of columns = 13
- 6 columns had the data type: Object
- 6 columns had the data type: float64
- Only one column had the data type: Int64

C. Handling Missing Data:

Missing Data can occur when no information is provided for one or more items or for a whole unit. We use isnull() function to check if there are any missing data or not. After finding some missing data in the dataset, by using dropna() function we drop the values and continue working on our dataset. For again verifying the analysis, sum of all the null values was founded, which comes out to be zero again giving us a surety. After this, authors proceeded for further analysis.

D. Statistical Analysis:

In this analysis, descriptive statistics is used by the authors. It is used for summarising the data, and describe the basic features of a dataset. It cannot be used for drawing conclusions but can help in getting the idea of the data and familiarising well with it. It's used for providing quantitative data in a manageable form[2].

For this analysis, head (initial rows of the dataset), tail (last rows of dataset), mean, minimum and maximum values are found for all the columns in the data set using inbuilt functions. Mean in this analysis, is the usual arithmetic mean of the data. Maximum and minimum values are also found. Names of the columns are changed for the sake of ease in writing the names in the program.

E. Visualization:

Visualisation is the process of understanding a data by giving it a visual context. It helps us to understand a data far better. We can find trends, conclusions and correlation by this process. It makes analysing, observing and understanding the data effortless. The authors used packages and libraries. Boxplot, bar graph, count plots, multivariate graphs and joint graphs were plotted by using the above-mentioned packages. Normal, descriptive statistics is used for obtaining useful data that are mean, maximum and minimum value. They can't be used for drawing conclusions but provide some useful insights[3].

The Bellow graph in Fig.2 shows the valuation for each unicorn. Where the x-axis represents the name of unicorns and y-axis represents their valuation in $ Billion. Zomato holds the value for the highest valuation of $ 8.6 billion. Acko, Apna, CoinDCX, Lead and Mamaearth, have the value of $1.1 billion, which is the lowest in this graph. Amagi, BlackBuck, Blinkit, Darwinbox, Licious, Open Banking and Oxyzo have the valuation $1 billion[4].

	Mean	Maximum	Minimum
Unicorns	-	UpGrad	Acko
Valuation ($ Bn)	-	<1	1
FY21 P/L (Cr)	-	55.7	-101
FY20 P/L (Cr)	-	9.6	-100
FY21 Op Revenue (Cr)	1268.817167	12595.0	0.86
FY21 Expenses (Cr)	1747.457833	13257.0	15.51
FY21 EBITDA Margin	-	82.99%	-1.17%
FY21 Exp/Op Revenue	8.007500	331.4	0.73
Total Funding ($ Mn)	-	878	1,000
Head Count	3623.983333	18459.0	113.0
Founded In	2013.083333	2020	1998
Unicorn In	2020.100000	2022.0	2014.0
Unicorn Age	7.016667	22.0	2.0

Table 1 Analysis of maximum, minimum and mean values.

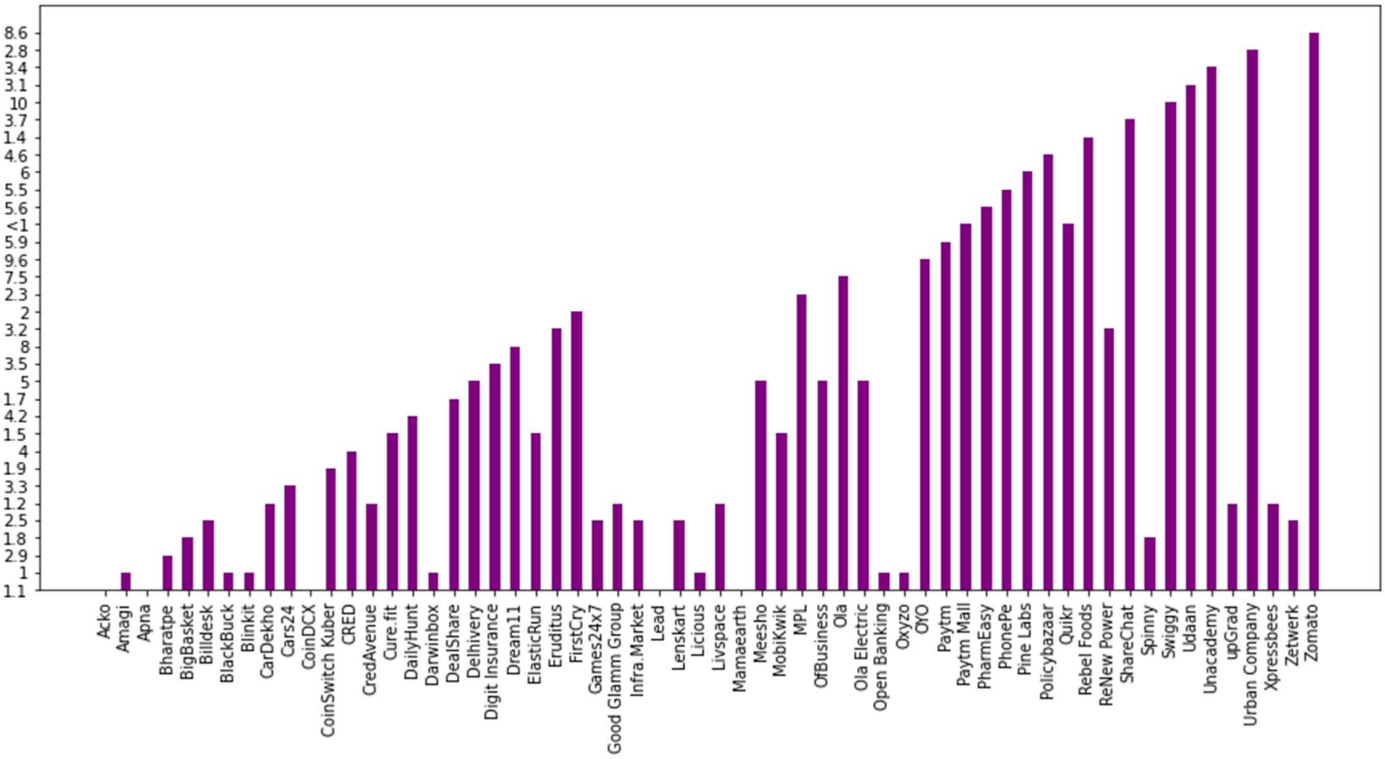

Figure 2 The graph above is used for analyzing the valuation of the unicorns in billion dollars.

Maximum, minimum and interquartile range are for operational revenue for the year 2020, expenses for the year 2021, expressional or operational revenue for the year 2021, head count, the year they were founded in, the year they attained the status of unicorn and their age. There are 7 outliers in the Head Count, which implies that these 7 companies hired the greatest number of employees that was out of the normal range. There are 5 outliers in the expenses of the year 2021 which means that these 5 companies had the most expenses that was out of the distribution. There are 4 outliers in the Operational Revenue implying these companies had the highest Operational Revenue which was out of the distribution.

Fig.3 shows the count of unicorns that were formed in particular year. 2014 and 2017 have the least number of unicorns formed whereas the year 2021 hold the largest value for number of unicorns formed. In the years 2019 and 2020, equivalent number of unicorns were formed.

The correlation between the profit and loss of the companies on the y-axis with their valuation on the x- axis. It is essential to mention here that valuation is in billion dollars and profit/loss is in million dollars. Here it is noticeable that they have a negative correlation.

In the figure 4, there is no such clear correlation between these two, but it can show that its density is mostly in lower area, with less expenses and by the companies that were founded in later years. Where the x-axis represents the year in which the unicorns were founded in, and y-axis represents the expenses for the year 2021. Histograms show these kind of results, in different manner. KDE (Kernel Density Estimation) in histograms shows smooth distribution and shape of data. Estimated regression line shows a negative a correlation between the two variable.

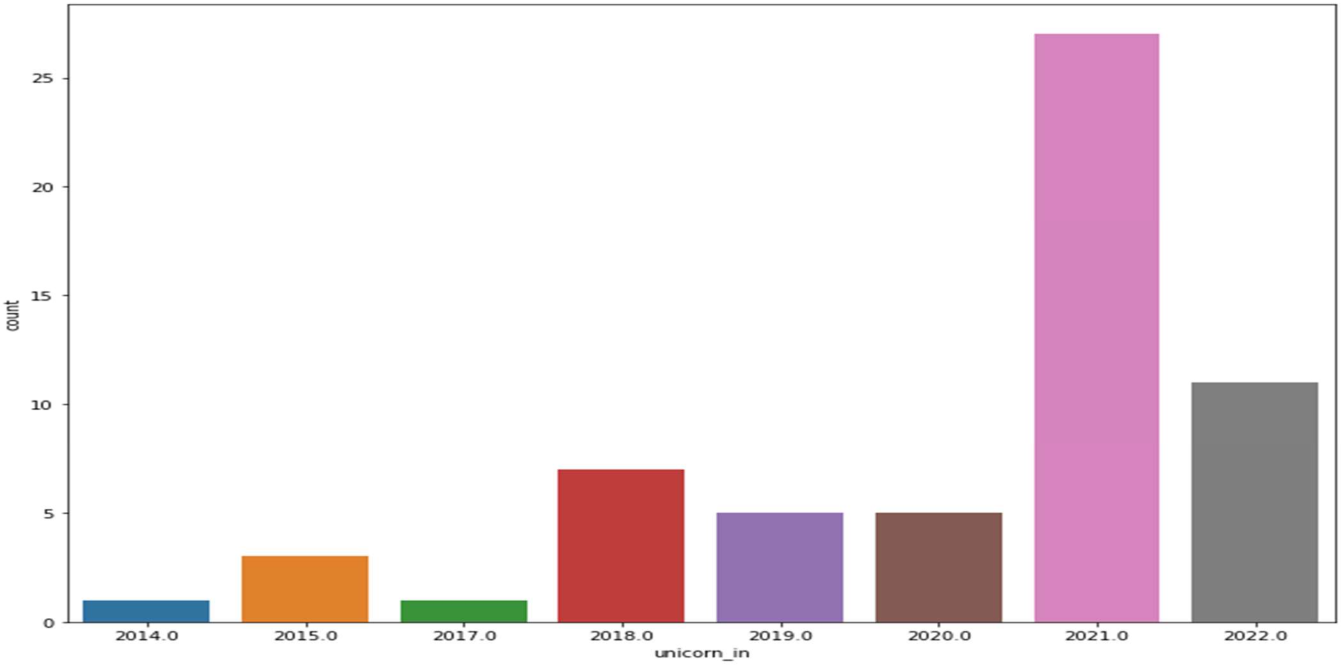

Figure 3 Categorical analysis for the year a company attained the status of unicorn

Figure 5 depicts seven different graphs which gives us meaningful insights for instance, in the 1st graph representing Operational Revenue of Unicorns in the year 2020 (where the number of Unicorns is on y-axis and Operational Revenue in Crores is on x-axis) more than 40 companies have their operational revenue as 1000 Cr in the year 2020, and less than 10 companies have the highest operational revenue. In the 2nd graph, representing the expenses for the year 2021 (where the number of unicorns is on the y-axis and expenses are on the x-axis) more than 35 companies have their expenses between the range 0 to 2000Cr, and approximately 11 companies have the highest expenses. In the 3rd graph, representing the Exp/Op Revenue for the year 2021 (where the number of unicorns is on the y- axis and Exp/Op Revenue are on the x-axis) approximately 59 companies have their Exp/Op Revenue between the range 0 to 35, and approximately 1 company has the highest Exp/Op Revenue. In the 4th graph, representing the.

Head Count (where the number of unicorns is on the y-axis and head count is on the x-axis) approximately 29 companies have the head count between the range 0-2500, and only less than 5 companies hire the most workforce. In the 5th graph, representing the year Unicorns were founded in (where the number of unicorns is on the y-axis and the year is on the x-axis) it is evident that most of the companies were created in the later years around the mid-2010s. In the 6th graph, representing the year the companies attained the status of being Unicorn (where the number of unicorns is on the y-axis and the year is on the x-axis) most of them were in the range 2018 – 2022, where 2021 has the peak.

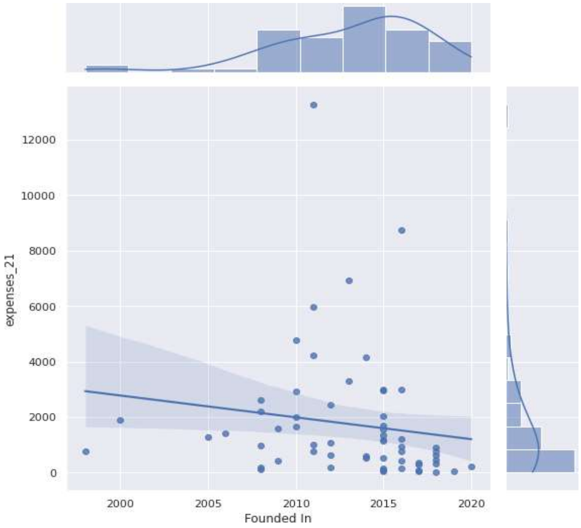

Figure 4 Analyzing expenses for the year 2021 with the year they were founded in

In the 7th graph, representing age of the unicorns (where the number of unicorns is on the y-axis and their age is on the x-axis) it can be seen that most of the companies are newly made and are of the ages 2.5 - 12 yrs., a few companies are 15 years old.

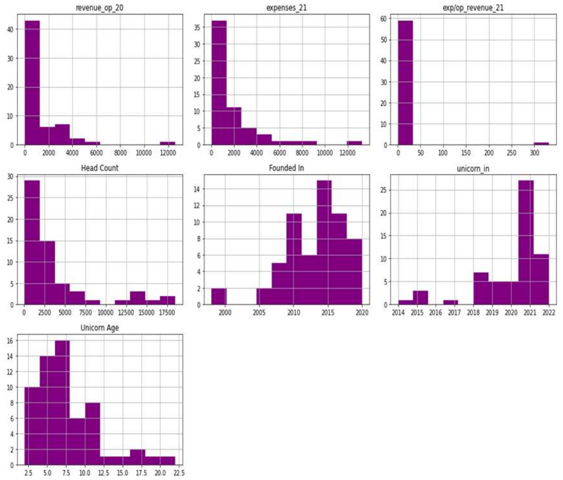

Figure 5 This analysis tells us about some essential

This figure 6 shows 36 graphs for analysis of relationship between two single variables and even distribution of the single variable. It gives us insights about data such as Operational Revenue for the year 2020 and expenses of the year 2021 are positively correlated .In the 23rd graph from start, where the year the unicorn was founded in is on the x-axis and Head count is on the y-axis, tells us that it's the few newly created companies that employees the most workforce[10]. There are 2 graphs that implies that the companies that attained the status of unicorns, were newly created and attained the status in the year 2021

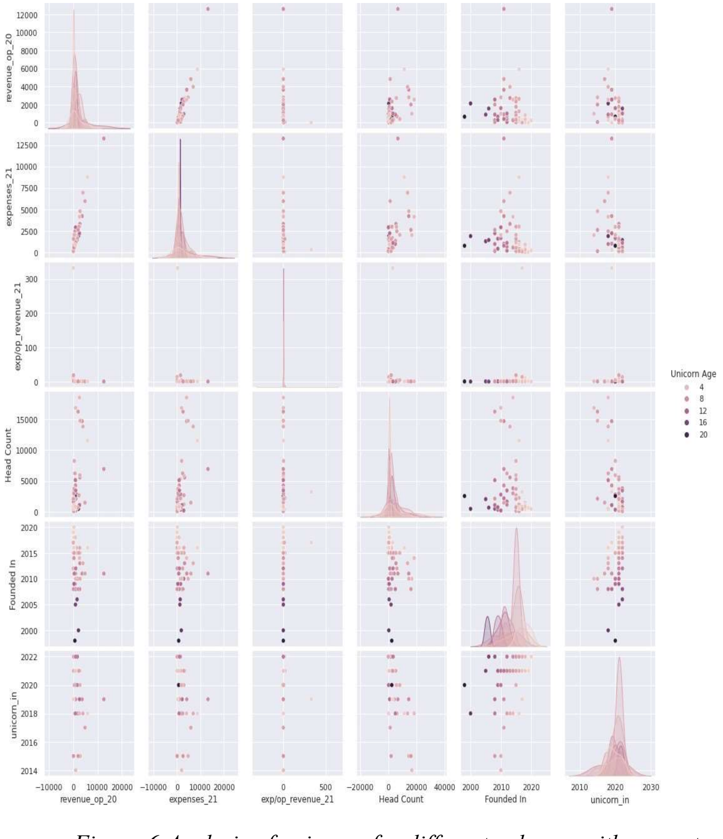

Figure 6 Analysis of unicorns for different columns with respect to their ages

III. Result And Discussions

The authors analysed that which company holds the maximum and minimum valuation (in $Billion). Zomato holds the value for the highest valuation of $ 8.6 billion. Acko, Apna, CoinDCX, Lead and Mamaearth, have the value of $1.1 billion, which is the lowest in this graph. Amagi, BlackBuck, Blinkit, Darwinbox, Licious, Open Banking and Oxyzo have the valuation $1 billion.

Analysis was also done for getting year in which maximum or minimum unicorns were formed. The year 2021 holds the record for formation of highest number of unicorns that is 44 and in the years 2014, 2017 lowest number of unicorns were formed. The reason for this can be pandemic in the years 2019 and 2020, which resulted in prevalence of smartphone and digitalization of commerce. It made funding for the investor easier and hence valuation increased [9].

In the analysis using boxplot it was found that there are 7 outliers in the Head Count, implying 7 companies hired the greatest number of employees that was out of the normal range. There are 5 outliers in the expenses of the year 2021 which means that these 5 companies had the most expenses that was out of the distribution. There are 4 outliers in the Operational Revenue implying these companies had the highest Operational Revenue which was out of the distribution.

Writers found that in the scatter plot analysis for the profit/loss and valuation, that they have a negative correlation represented through dots in the graph.

In the joint plot analysis, it is found that there is no such clear correlational between expenses (for year 2021) and valuation, it shows that its density is mostly in lower area, with less expenses and by the companies that were founded in later years. Histograms shows the same kind of results, in different manner. KDE (Kernel Density Estimation) in histograms shows smooth distribution and shape of data. Estimated regression line shows a negative a correlation between the two variables[7].

For the multiple histograms, multiple results can be found. In the 1st graph (Operational Revenue, 2020 v/s number of unicorns) more than 40 companies have their operational revenue as 1000 Cr in the year 2020, and less than 10 companies have the highest operational revenue. In the 2nd graph (expenses v/s number of unicorns) more than 35 companies have their expenses between the range 0 to 2000Cr, and approximately 11 companies have the highest expenses. In the 3rdgraph (Exp/Op revenue v/s number of unicorns) approximately 59 companies have their Exp/Op Revenue between the range 0 to 35, and approximately 1 company has the highest Exp/Op Revenue. In the 4th graph (head count v/s number of unicorns) approximately 29 companies have the head count between the range 0-2500, and only less than 5 companies hire the most workforce. In the 5th graph (the year Unicorns were founded in v/s number of unicorns) most of the companies were created in the later years around the mid-2010s. In the 6th graph (the year the companies attained the status of being Unicorn v/s number of unicorns) most of them were in the range 2018 – 2022, where 2021 has the peak. In the 7th graph (age v/s number of unicorns) it can be seen that most of the companies are newly made and are of the ages 2.5 - 12 yrs., a few companies are 15 years old [8].

The last analysis was done using pair plot that gives us insights about data such as Operational Revenue for the year 2020 and expenses of the year 2021 are positively correlated. It's the few newly created companies that employees the most workforce. The companies that attained the status of unicorns, were newly created and attained the status in the year 2021 implying situation for making a new start-ups has improved and enhancing.

IV. Conclusions

In this project, analysis of a dataset by using simple methodology and python packages and libraries is successfully executed. This gives us many useful insights that can help us understand more about these unicorns, their profit, losses, expenses, and many other essential parameters. It also assisted in getting to know the importance of digitalization that happened due to pandemic situation, which promoted the attracted the investors to give these companies funding's. A large growth can be seen in the EdTech, Online Shopping and many such innovative start-up's. Concluding that the recent years have provided a good environment for starting more such companies and innovative ideas, which would be continued in the coming years, that is a good sign for the economics as well as the general public as these are companies that helps in overcoming the problem of unemployment by hiring skilled workforce.

V. Refrences

[1] Singh, Shiwangi & Chauhan, Akshay & Dhir, Sanjay. (2019). Analyzing the startup ecosystem of India: a Twitter analytics perspective. Journal of Advances in Management Research. 17. 10.1108/JAMR-08-2019-0164.

[2] Pérez-Morón, James, Lina Marrugo-Salas, and Veronica Tordecilla-Acevedo. "ARE ASIAN UNICORNS IMMUNE TO COVID-19? A "LIVE" ASSESSMENT." In 13th Annual Conference of the Euromed Academy of Business: Business Theory and Practice across Industries and Markets, pp. 895-906. 2020.

[3] A. Singh, D. P. Kumar, K. Shivaprasad, M. Mohit and A. Wadhawan, "Vehicle Detection And Accident Prediction In Sand/Dust Storms," 2021 International Conference on Computing Sciences (ICCS), 2021, pp. 107-111, doi: 10.1109/ICCS54944.2021.00029.

[4] T. Fadhaeel, P. C. H, A. Al Ahdal, M. Rakhra and D. Singh, "Design and development an Agriculture robot for Seed sowing, Water spray and Fertigation," 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022, pp. 148-153, doi: 10.1109/CISES54857.2022.9844341.

[5] Rakhra, Manik, Partho Deb, Omdev Dahiya, Sahil Sonu Chandel, Brinderjit Bhutta, Sumit Badotra, Sunny Kumar, Aman Shaukat, and Dalwinder Singh. "An Analytical Study of the Types of Implements used by Farmers in Mechanized Agriculture." In 2022 International Mobile and Embedded Technology Conference (MECON), pp. 683-687. IEEE, 2022.

[6] Shruti, R. Soumya, M. Rakhra, N. S. Chintagunti, B. Singh and D. Singh, "Modern Data Mining Approach to Handle Multivariate Data and to Implement Best Saving Services for Potential Investor," 2022 International Mobile and Embedded Technology Conference (MECON), 2022, pp. 611-616, doi: 10.1109/MECON53876.2022.9752101.

[7] M. N. Gowda, D. Singh and M. Rakhra, "Machine Learning – Based Diagnosis of Covid-19 using Clinical Data," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), 2022, pp. 910-916, doi: 10.1109/ICIEM54221.2022.9853083.

[8] M. Rakhra and R. Singh, "Economic and Social Survey on Renting and Hiring of Agricultural Equipment of Farmers in Punjab," 2021 9th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir. ICRITO 2021, pp. 1–5, 2021, doi: 10.1109/ICRITO51393.2021.9596343.

[9] Chhabra, Abhishek, Tenzin Woeden, Dalwinder Singh, Manik Rakhra, Omdev Dahiya, and Aditya Gupta. "Image Steganalysis with Image decoder using LSB and MSB Technique." In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 900-905. IEEE, 2022.

[10] Takkar, Sakshi, Anuj Kakran, Veerpal Kaur, Manik Rakhra, Manish Sharma, Pargin Bangotra, and Neha Verma. "Recognition of Image-Based Plant Leaf Diseases Using Deep Learning Classification Models." *Nature Environment & Pollution Technology* 20 (2021).