

Course Completion Project

A REPORT ON

“EDA-LIST OF UNICORN STARTUP COMPANIES ANALYSIS”

A Project Report Submitted to the Loosies Technologies In partial fulfilment of the requirement for the award of the Course/Internship in

**MASTER OF DATA ANALYTICS**

year 2024-2025

SUBMITTED BY

**VIJAY BEDAGE**

UNDER THE GUIDANCE OF

Prof. Abhisheka Janhavi

Data science at

Loosies Technologies Belagavi

CERTIFICATE

Certified that this project report entitled "**Loosies Technologies Belagavi**" which is being submitted by **Mr. VIJAY BEDAGE** has successfully completed the course by demonstrating theoretical and practical understanding of **Data Analytics using Python** during the year 2024-25.

Guide

Prof. Abhisheka Janhavi

Data science at

Loosies Technologies Belagavi

Acknowledgment

The successful completion of this project on

**“Data Analytics using Python”** would not have been possible without the support and guidance, I would like to express my sincere gratitude to my supervisor, **Prof. Abhisheka Janhavi**, for her exceptional guidance, support, and the opportunity to gain hands-on exposure to real-world applications in Data Analytics using Python. Her advice and mentorship have been invaluable in the successful execution of this study.

Also, I am deeply grateful to **Mr. Manoj DN** sir for his valuable insights and consistent encouragement throughout the project. His suggestions and feedback at every stage have been instrumental in enhancing the quality of this work.

Finally, I am profoundly thankful to my parents, the faculty members, and my friends for their constant encouragement and assistance, which inspired me to strive for excellence in this project.

|  |  |  |
| --- | --- | --- |
| CHAPTER | CONTENTS | PAGE NO. |
| Introduction | 1.1. Definition of Data Science  1.2. Importance of Data Science  1.3. Objectives of the Study | **5 - 14** |
| Exploratory Data Analysis | 2.1. Definition and Significance  2.2. Types of EDA  2.3. Tools and Techniques | 15-17 |
| Methodology | 3.1. Dataset Description  3.2. Steps in Data Pre-processing  3.3. Exploratory Techniques Employed | 18-20 |
| Results and Analysis | 4.1. Country-Wise Distribution of Unicorns  4.2. Industry-Wise Insights  4.3. Valuation Analysis  4.4. Regional Comparisons | 21 |
| Visualization | 5.1. Line Charts and Bar Graphs  5.2. Pie Charts and Exploded Views  5.3. Top 10 Unicorns Analysis | 21-29 |
| Discussion | 6.1. Key Observations  6.2. Challenges Encountered  6.3. Limitations of the Study | 30-32 |
| Conclusion and Recommendations | 7.1. Summary of Findings  7.2. Recommendations for Stakeholders | 33 |
| Power BI Dashbord AND  References | 8.0. Dashbord  8.1. References | 34 |

index

DATA SCIENCE

Data science is the study of data to extract meaningful insights for business. It is a

multidisciplinary approach that combines principles and practices from the fields

of mathematics, statistics, artificial intelligence, and computer engineering to

analyse large amounts of data. This analysis helps data scientists to ask and

answer questions like what happened, why it happened, what wil happen, and

what can be done with the results.

IMPORTANCE OF DATA SCIENCE

Data science is important because it combines tools, methods, and technology to

generate meaning from data. Modern organizations are inundated with data;

there is a proliferation of devices that can automatically collect and store

information. Online systems and payment portals capture more data in the fields

of e-commerce, medicine, finance, and every other aspect of human life. We have

text, audio, video, and image data available in vast quantities.

ADVANTAGE OF DATA SCIENCE

In today's world, data is being generated at an alarming rate. Every second, lots of

data is generated; be it from the users of Facebook or any other social networking

site, or from the calls that one makes, or the data which is being generated from

different organizations.

Multiple Job Options :- Being in demand, it has given rise to a large number of

career opportunities in its various fields. Some of them are Data Scientist, Data

Analyst, Research Analyst, Business Analyst, Analytics Manager, Big Data Engineer,

etc.

**• Business benefits:-** Data Science helps organizations knowing how and when

their products sell best and that's why the products are delivered always to the

right place and right time. Faster and better decisions are taken by the

organization to improve efficiency and earn higher profits.

**Highly Paid jobs & career opportunities:-** As Data Scientist continues being the

sexiest job and the salaries for this position are also grand. According to a Dice

Salary Survey, the annual average salary of a Data Scientist $106,000 per year**.**

DATA ANALYTIC (DA)

It is the process of examining data sets to find trends and draw conclusions

about the information they contain. Increasingly, data analytics is done with the

help of specialized systems and software. Data analytics technologies and

techniques are widely used in commercial industries to enable organizations to

make more informed business decisions. Scientists and researchers also use data

analytics tools to verify or disprove scientific models, theories and hypotheses.

IMPORTANCE OF DATA ANALYTIC

Data analytics initiatives can help businesses increase revenue, improve

operational efficiency, optimize marketing campaigns and bolster customer

satisfaction efforts across multiple industries. It can also help organizations do the

following:

1. Personalize customer experiences. By going beyond traditional data

methods, data analytics connects insights with actions, enabling businesses to

create personalized customer experiences and develop related digital products.

2. Predict future trends. By using predictive analysis technologies, businesses can

create future-focused products and respond quickly to emerging market trends,

thereby gaining a competitive advantage over business rivals. Depending on the

application, the data that's analyzed can consist of either historical records or

new information that has been processed In addition, it can come from a mix of

internal systems and external data sources.

3.Reduce operational costs. By optimizing processes and resource allocation,

data analytics can help reduce unnecessary expenses and identify cost-saving

opportunities within the organization.

4. Provide risk management. Data analytics lets organizations identify and

mitigate risks by detecting anomalies, fraud and potential compliance issues.

5.Improve security. Companies use data analytics methods, such as parsing,

analyzing and visualizing audit logs, to look at past security breaches and find

the underlying vulnerabilities. Data analytics can also be integrated

with monitoring and alerting to quickly notify security professionals in the

event of a breach attempt.

6. Measure performance. Data analytics provide organizations with metrics and

key performance indicators to track progress, monitor performance and evaluate

the success of business initiatives. This helps businesses respond

promptly to changing market conditions and other operational challenges.

WHAT IS EXPLORATORY DATA ANALYSIS (EDA)?

• Exploratory data analysis (EDA) is used by data scientists to analyse and

investigate data sets and summarize their main characteristics, often

employing data visualization methods.

• EDA helps determine how best to manipulate data sources to get the

answers you need, making it easier for data scientists to discover patterns,

spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or

hypothesis testing task and provides a provides a better understanding of data set

variables and the relationships between them. It can also help determine if the

statistical techniques you are considering for data analysis are appropriate.

Originally developed by American mathematician John Tukey in the 1970s, EDA

techniques continue to be a widely used method in the data discovery process

today.

WHY IS EXPLORATORY DATA ANALYSIS IMPORTANT IN DATA SCIENCE?

The main purpose of EDA is to help look at data before making any

assumptions. It can help identify obvious errors, as well as better

understand patterns within the data, detect outliers or anomalous events,

find interesting relations among the variables.

• Data scientists can use exploratory analysis to ensure the results they

produce are valid and applicable to any desired business outcomes and

goals. EDA also helps stakeholders by confirming they are asking the right

questions. EDA can help answer questions about standard deviations,

categorical variables, and confidence intervals. Once EDA is complete

TYPES OF EXPLORATORY DATA ANALYSIS

There are four primary types of EDA:

• Univariate non-graphical. This is simplest form of data analysis, where the

data being analysed consists of just one variable. Since it's a single variable,

it doesn't

• deal with causes or relationships. The main purpose of univariate analysis is

to describe the data and find patterns that exist within it.

• Univariate graphical. Non-graphical methods don't provide a full picture of

the data. Graphical methods are therefore required. Common types of

univariate graphics include:

• Stem-and-leaf plots, which show all data values and the shape of the

distribution.

• Histograms, a bar plot in which each bar represents the frequency

(count) or proportion (count/total count) of cases for a range of

values.

• Box plots, which graphically depict the five-number summary of

minimum, first quartile, median, third quartile, and maximum.

• Multivariate nongraphical: Multivariate data arises from more than one

variable. Multivariate non-graphical EDA techniques generally show the

relationship between two or more variables of the data through cross-

tabulation or statistics.

• Multivariate graphical: Multivariate data uses graphics to display

relationships between two or more sets of data. The most used graphic is a

grouped bar plot or bar chart with each group representing one level of one

of the variables and each bar within a group representing the levels of the

other variable.**EXPLORATORY DATA ANALYSIS IN PYTHON**

Exploratory data analysis (EDA) is a critical initial step in the data science

workflow. It involves using Python libraries to inspect, summarize, and

visualize data to uncover trends, patterns, and relationships. Here's a

breakdown of the key steps in performing EDA with Python:

**1. Importing Libraries:**

• **pandas (pd):** For data manipulation and analysis.

• **NumPy (np):** For numerical computations.

• **Matplotlib.pyplot (pit):** For basic plotting functionalities.

• **Seaborn (sns):** A built-on top of Matplotlib, providing high-level

visualization.

• **Statistics(stats):** to understand the central tendency, position measure

of dispersion and covariance, correlations, association and causation.

**2. Loading the Data**:

Use pd.read\_csv() for C S files, similar functions exist for other data formats

(e.g., xIsx, json).

**3. Initial Inspection:**

• Get an overview of the data using df.head), tail), and info().

• Check data types with df.dtypes.

**4.Data Cleaning:**

• Identify and handle missing values using methods like df.isnull).sum).

• Find and address duplicates with df.duplicated).sum).

**5. Univariate Analysis:**

• Analyze single variables at a time.

• Use descriptive statistics with df.describe() for numerical data.

• Create histograms, box plots, and density plots to visualize distributions.

**6.Bivariate Analysis:**

• Explore relationships between two variables.

• Create scatter plots to identify trends and potential correlations.

**7. Visualization:**

* Pie charts for energy source distribution
* Bar graphs for sector-wise usage
* Heatmaps for regional adoption
* Line charts for historical trends

**DATA PRE-PROCESSING**

In data-driven processes, we prioritize refining our raw data through the

crucial stages of EDA (Exploratory Data Analysis). Both data pre-processing

and feature engineering play pivotal roles in this endeavour. EDA involves a

comprehensive range of activities, including data integration, analysis,

cleaning, transformation, and dimension reduction.

Data pre-processing involves cleaning and preparing raw data to facilitate

feature engineering. Meanwhile, feature engineering entails employing

various techniques to manipulate the data. This may include adding or

removing relevant features, handling missing data, encoding variables, and

dealing with categorical variables, among other tasks.

**EDA USING PYHTON**

**\* Data Collection**

Data Collection is an essential part of exploratory data analysis. It refers to the

process of finding and loading data into our system. Good, reliable data can be

found on various public sites private organizations. Some reliable data collection

The data depicted below represents the supermarket sales that is available on

Kaggle. It contains information invoice ID,Branch, City, Customer type, Gender,

Prout line, Unit price, Quantity, Tax 5%, Total, Date, Time, Payment, COBs,Gross

margin percentage, Gross income, Rating.

**Step 1:** Import Python Libraries

Import al libraries which are required for our analysis, such as Data Loading,

Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

**import pandas a s pd**

**importnumpyas np**

**importmatplotlib. pyplotasplt**

**import seaborn assns.**

**#to ignore warnings**

**import warnings**

**warnings.filterwarnings(ignore')**

**Step 2:** Reading Dataset

The Pandas library offers a wide range of possibilities for loading data into the

pandas DataFrame from files like JSON, csv, xIsx, sal, •pickle, html, .txt,

images etc.

Most of the data are available in a tabular format of CSV files. It is trendy and

easy to access. Using the **read \_csv()** function, data can be converted to a

pandas DataFrame.

In this article, the data to predict Used supermarket Sales is being used as an

example. In this dataset, we are trying to analyze the Market sales.

On identifying the factors influencing the Supermarket. We have stored the

data in the DataFrame data.

**df = pd.read\_csv("List Of Unicorn Startup Companies.csv")**

**Step 3: Data Cleaning**

Data Cleaning refers to the process of removing unwanted variables and values

from your dataset and getting rid of any irregularities in it. Such anomalies can

disproportionately skew the data and hence adversely affect the results. Some

steps that can be done to clean data are:

• Removing missing values, outliers, and unnecessary rows/ columns.

• Re-indexing and reformatting our data.

Before we make any inferences, we listen to our data by examining all

variables in the data.

The main goal of data understanding is to gain general insights about

the data, which covers the number of rows and columns, values in the

data, data types, and Missing values in the dataset.

• **shape - shape** will display the number of observations(rows) and

features(columns) in the dataset

There are 1001 observations and 16 variables in our dataset

**• head()** will display the top 5 observations of the dataset

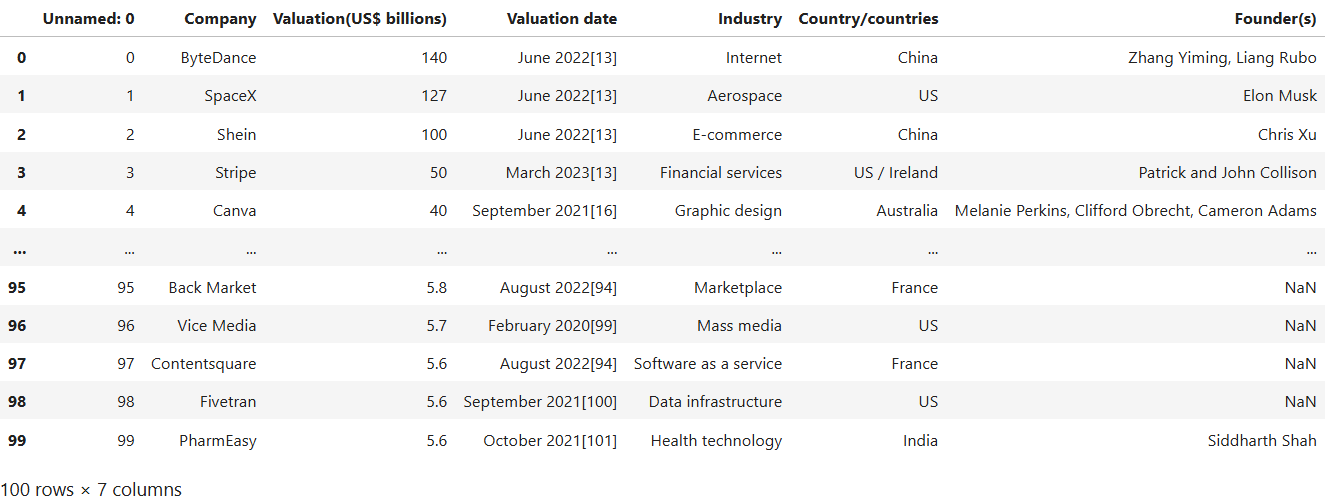
1. **Load and Explore the Dataset:**
   * Libraries such as Pandas, NumPy, Matplotlib, and Seaborn are imported.
   * The dataset is loaded into a DataFrame, and basic data structures are inspected.

# Load the dataset

df = pd.read\_csv('List\_of\_Unicorn \_Startup\_Companies.csv')

# Preview the first few rows

df.head(100)



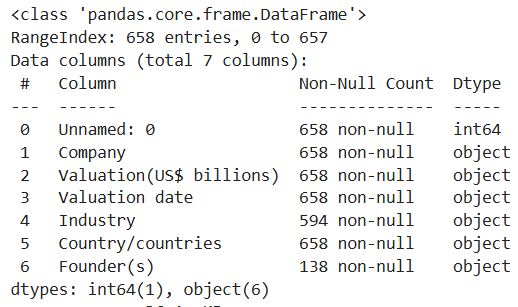
The df.head(100) output will display the first 100 rows of the dataset, including columns such as:

* **Company**: Name of the company
* **Valuation(US$ billions)**: The company's valuation in billions of US dollars
* **Valuation date**: Date when the valuation was recorded
* **Industry**: The industry in which the company operates
* **Country/countries**: The country or countries where the company is based
* **Founder(s)**: The founders of the company

This preview helps in understanding the data distribution and identifying any necessary data cleaning or preprocessing steps before further analysis.

1. **Data Cleaning:**

**df.info()**



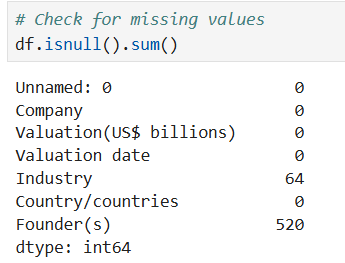
* **Total Entries**: 658, indexed from 0 to 657.
* **Columns**:
  1. **Unnamed: 0**: 658 non-null entries of type int64
  2. **Company**: 658 non-null entries of type object
  3. **Valuation (US$ billions)**: 658 non-null entries of type object
  4. **Valuation date**: 658 non-null entries of type object
  5. **Industry**: 594 non-null entries of type object
  6. **Country/countries**: 658 non-null entries of type object
  7. **Founder(s)**: 138 non-null entries of type object

### Observations:

* **Data Completeness**:
  + The columns "Company," "Valuation (US$ billions)," "Valuation date," and "Country/countries" are complete with 658 non-null entries.
  + The "Industry" column has 594 non-null entries, indicating some missing data.
  + The "Founder(s)" column has the least completeness with only 138 non-null entries.
* **Data Types**:
  + Most of the columns are of type object, except for "Unnamed: 0," which is of type int64.

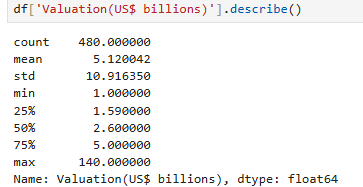
***Missing values are handled, and data is filtered based on conditions.:***

A Python code snippet and its output, revealing missing values in a DataFrame, with 64 missing in 'Industry' and 520 in 'Founder(s)'.



***Basic Descriptive Statistics:***

The output of the describe() method in Python for a DataFrame column named 'Valuation(US$ billions)', providing summary statistics such as count, mean, standard deviation, minimum, median, and maximum values. The column contains 480 entries with a mean valuation of $5.12 billion and a maximum valuation of $140 billion.



1. **Descriptive Statistics:**

**Top 10 highest-valued companies**

****

**Key Insights:**

**1. \*\*Top Valuation\*\*:**

**- \*\*ByteDance\*\* leads with the highest valuation of 140.0 billion dollars, showcasing its significant market presence and financial strength.**

**2. \*\*Diverse Industries\*\*:**

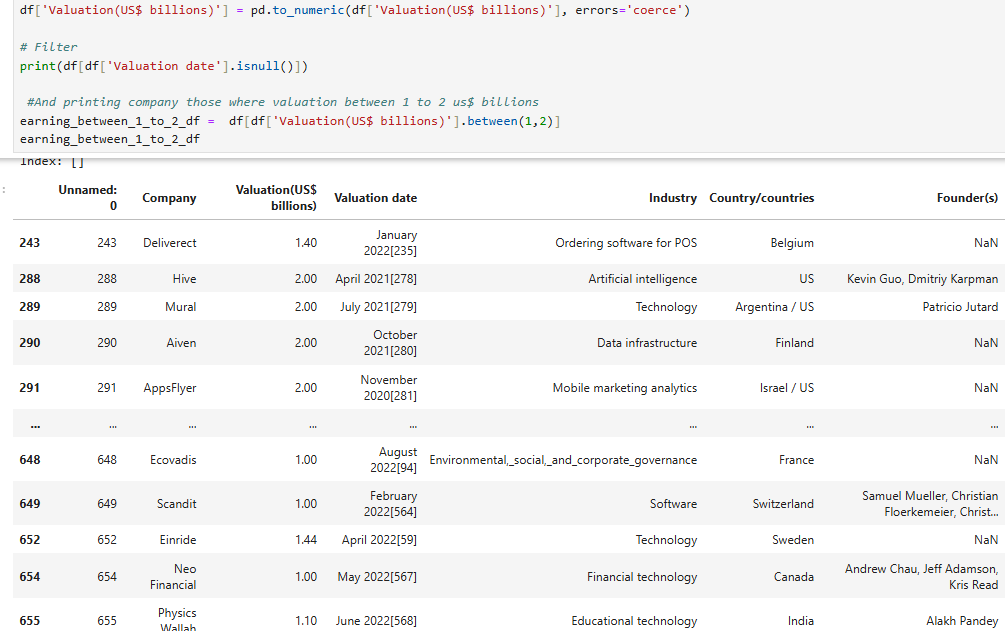
**- The companies represent a range of industries including internet services, aerospace, e-commerce, financial technology, and software development.**

**3. \*\*Global Presence\*\*:**

**- These companies are based in various countries, indicating the global nature of high-value private companies.**

**This table highlights the most valuable private companies globally, reflecting their substantial market valuations.**

***Convert 'Valuation(US$ billions)' to numeric, handling errors Filtering perticular column And printing company those were valuation between 1 to 2 US$ in Billions***



The code snippet in the image performs data manipulation and filtering operations on a DataFrame. It includes the following steps:

1. **Convert 'Valuation(US$ billions)' to Numeric**:
   * df['Valuation(US$ billions)'] = pd.to\_numeric(df['Valuation(US$ billions)'], errors='coerce')
   * This line converts the values in the 'Valuation(US$ billions)' column to numeric values, handling errors by setting invalid parsing as NaN.
2. **Filter for Missing Valuation Dates**:
   * print(df[df['Valuation date'].isnull()])
   * This line prints rows where the 'Valuation date' column is missing.
3. **Filter and Print Companies with Valuations Between 1 and 2 Billion US$**:
   * earning\_between\_1\_to\_2\_df = df[df['Valuation(US$ billions)'].between(1, 2)]
   * earning\_between\_1\_to\_2\_df
   * These lines filter the DataFrame for companies with valuations between 1 and 2 billion US dollars and store the result in earning\_between\_1\_to\_2\_df.

### Table Details:

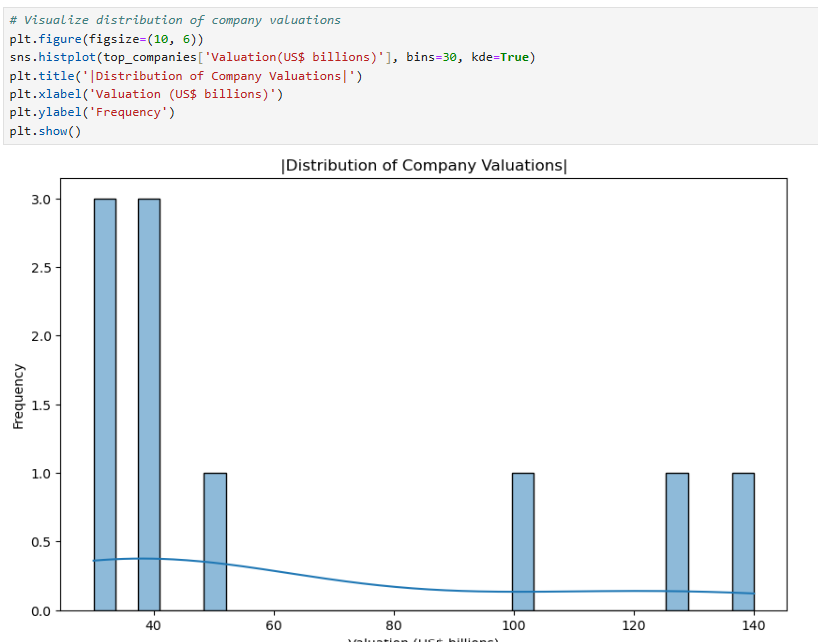
The table lists various companies along with their valuations, valuation dates, industries, countries, and founders. Here are the key points from the table:

1. **Columns**:
   * **Unnamed: 0**: Row index
   * **Company**: Name of the company
   * **Valuation(US$ billions)**: Valuation in billion US dollars
   * **Valuation date**: Date of the valuation
   * **Industry**: Industry the company operates in
   * **Country/countries**: Countries where the company is based
   * **Founder(s)**: Founders of the company
2. **Example Entries**:
   * **Deliverect**: Valuation of $1.40 billion, Valuation date: January 2022, Industry: Ordering software for POS, Country: Belgium, Founders: Not listed
   * **Hive**: Valuation of $2.00 billion, Valuation date: April 2021, Industry: Artificial intelligence, Country: US, Founders: Kevin Guo, Dmitriy Karpman
   * **Mural**: Valuation of $2.00 billion, Valuation date: July 2021, Industry: Technology, Countries: Argentina/US, Founder: Patricio Jutard
   * **Aiven**: Valuation of $2.00 billion, Valuation date: October 2021, Industry: Data infrastructure, Country: Finland, Founders: Not listed
   * **AppsFlyer**: Valuation of $2.00 billion, Valuation date: 2020, Industry: Mobile marketing analytics, Countries: Israel/US, Founders: Not listed
   * **Ecovadis**: Valuation of $1.00 billion, Valuation date: August 2022, Industry: Environmental, social, and corporate governance, Country: France, Founders: Not listed
   * **Scandit**: Valuation of $1.00 billion, Valuation date: February 2022, Industry: Software, Country: Switzerland, Founders: Samuel Mueller, Christian Floerkemeier
   * **Einride**: Valuation of $1.00 billion, Valuation date: April 2022, Industry: Technology, Country: Sweden, Founders: Not listed
   * **Neo Financial**: Valuation of $1.00 billion, Valuation date: 2022, Industry: Financial technology, Country: Canada, Founders: Andrew Chau, Jeff Adamson, Kris Read
   * **Physics Wallah**: Valuation of $1.10 billion, Valuation date: June 2022, Industry: Educational technology, Country: India, Founder: Alakh Pandey

This table and code provide a comprehensive overview of various companies' valuations, helping to identify trends and key players in different industries.

1. **Visualization:**

**1)Visualize the number of companies by country.**

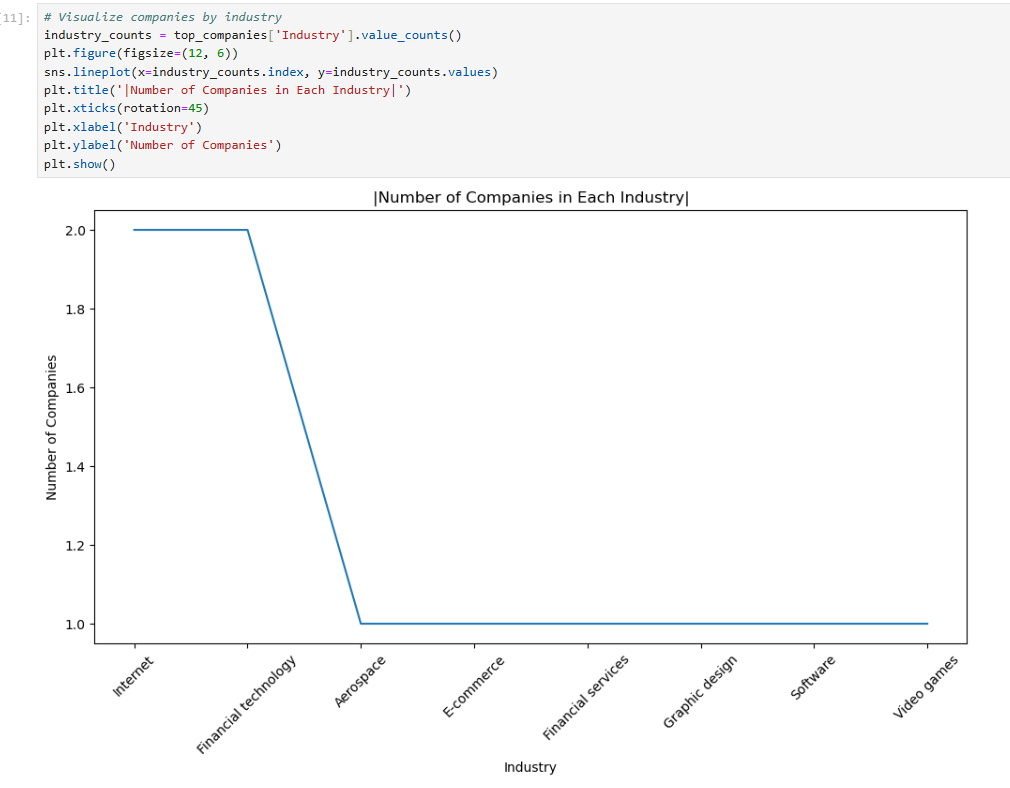


The image shows a Python code snippet and the resulting histogram plot. The code uses the matplotlib and seaborn libraries to visualize the distribution of company valuations in US$ billions. The histogram has 30 bins and includes a kernel density estimate (KDE) line.

### Plot Insights:

* **X-Axis (Valuation in US$ billions)**:Shows the range of company valuations from 0 to 140 billion dollars.
* **Y-Axis (Frequency):**Indicates how many companies fall into each valuation bin.
* **Histogram**:Bars show the frequency of companies within each valuation range.
* **KDE Line**:The smooth line overlaid on the histogram represents the estimated probability density function of the data, helping to visualize the distribution trend more clearly.Plot Title and Labels:
* **Title**:"Distribution of Company Valuations”
* **X-Axis Label**:"Valuation (US$ billions)"
* **Y-Axis Label**: "Frequency"

**2)Visualize the number of companies in Each Industry using line chart**

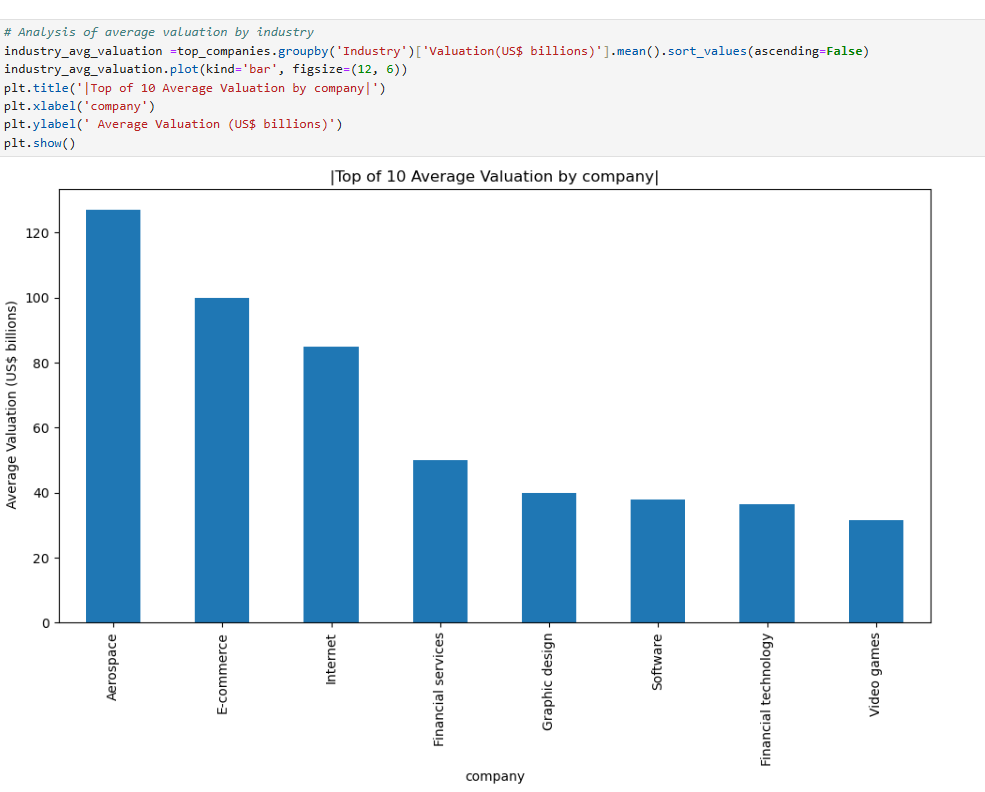


1. **Histogram Plot**:
   * The code creates a histogram to visualize the distribution of company valuations, using 30 bins and including a kernel density estimate (KDE) line to show the distribution's shape more smoothly.
2. **Figure Size**:
   * The plot's figure size is set to 10 by 6 inches for better readability.
3. **Plot Title and Labels**:
   * The plot is titled "Distribution of Company Valuations."
   * The x-axis is labeled "Valuation (US$ billions)."
   * The y-axis is labeled "Frequency."
4. **Plot Display**:
   * The plot is displayed using the plt.show() function.

### Plot Insights:

* **X-Axis (Valuation in US$ billions)**:
  + This axis shows the range of company valuations, from 0 to 140 billion dollars.
* **Y-Axis (Frequency)**:
  + This axis indicates how many companies fall into each valuation bin.
* **Histogram**:
  + The bars show the frequency of companies within each valuation range.
* **KDE Line**:The smooth line overlaid on the histogram represents the estimated probability density function of the data.

**3)Visualize the Top 10 Average Valuation by Company using bar chart**



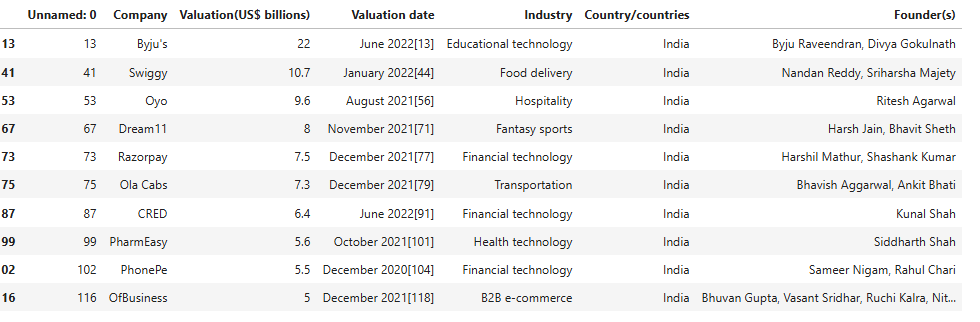
The image shows a bar chart titled **"Top of 10 Average Valuation by company,"** which presents the average valuation (in US$ billions) of companies across different industries.

Here are the key insights:

1. **Industries on the X-Axis**:
   * The industries listed are Aerospace, E-commerce, Internet, Financial Services, Graphic Design, Software, Financial Technology, and Video Games.
2. **Valuation Range on the Y-Axis**:
   * The y-axis represents the average valuation in US$ billions, ranging from 0 to 140.
3. **Top Industry**:
   * The Aerospace industry has the highest average valuation, suggesting significant financial value and growth in this sector.
4. **High-Valuation Sectors**:
   * E-commerce and Internet industries follow closely, indicating their strong market positions.
5. **Low-Valuation Sector**:
   * The Video Games industry has the lowest average valuation among the listed industries.

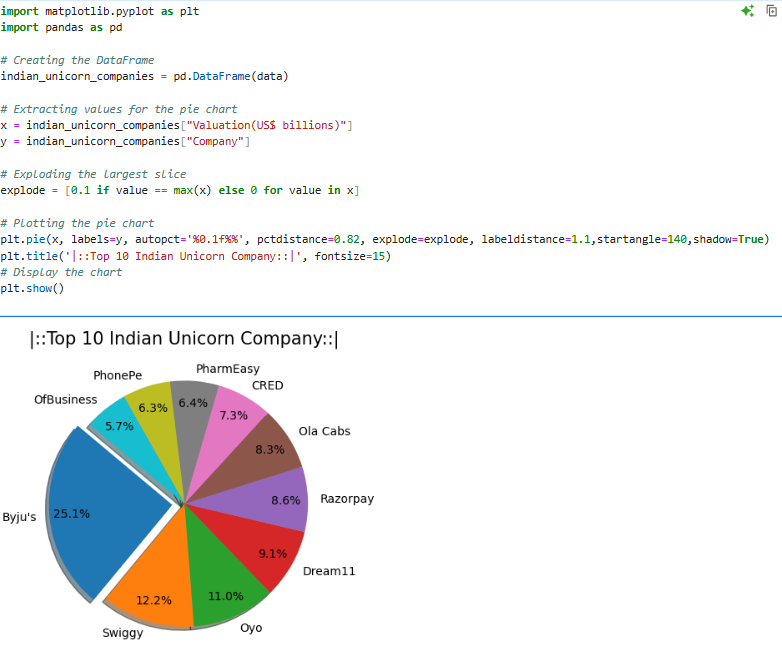
**4)Grouping perticular item from perticular column :-Grouping “India” from “Country/countries “**





1. **Company Names**:
   * **Byju's**: Valuation of $22 billion, Valuation date: June 22, 2022.
   * **Swiggy**: Valuation of $10.7 billion, Valuation date: January 2022.
   * **Oyo**: Valuation of $9.6 billion, Valuation date: August 2021.
   * **Dream11**: Valuation of $8 billion, Valuation date: November 2021.
   * **Razorpay**: Valuation of $7.5 billion, Valuation date: December 2021.
   * **Ola Cabs**: Valuation of $7.3 billion, Valuation date: December 2021.
   * **CRED**: Valuation of $6.4 billion, Valuation date: June 2022.
   * **PharmEasy**: Valuation of $5.6 billion, Valuation date: October 2021.
   * **PhonePe**: Valuation of $5.5 billion, Valuation date: December 2021.
   * **OfBusiness**: Valuation of $5 billion, Valuation date: December 2021.
2. **Industries**:
   * The companies span various industries such as Educational technology, Food delivery, Hospitality, Fantasy sports, Financial technology, Transportation, Health technology, and B2B e-commerce.
3. **Founders**:
   * **Byju Raveendran, Divya Gokulnath** (Byju's)
   * **Nandan Reddy, Sriharsha Majety** (Swiggy)
   * **Ritesh Agarwal** (Oyo)
   * **Harsh Jain, Bhavit Sheth** (Dream11)
   * **Harshil Mathur, Shashank Kumar** (Razorpay)
   * **Bhavish Aggarwal, Ankit Bhati** (Ola Cabs)
   * **Kunal Shah** (CRED)
   * **Siddharth Shah** (PharmEasy)
   * **Sameer Nigam, Rahul Chari** (PhonePe)
   * **Bhuvan Gupta, Vasant Sridhar, Ruchi Kalra, Nitin Jain** (OfBusiness)

**5)Visualize pie chart of Top 10 Indian Unicorn company Exploding Top one company**



### Pie Chart Insights:

The pie chart shows the following companies and their respective proportions based on valuation:

* Byju’s: 12.0%
* Swiggy: 11.0%
* Oyo: 9.1%
* Dream11: 8.6%
* Razorpay: 8.3%
* Ola Cabs: 7.3%
* CRED: 6.4%
* PharmEasy: 6.3%
* PhonePe: 5.7%
* OfBusiness: 4.6%

### Visual Description:

The chart is visually appealing and provides a clear comparison of the valuations of the top 10 Indian unicorn companies, highlighting their market presence and financial strength.

6)**Grouping perticular item from perticular column :- Grouping “China” from “Country/countries “**

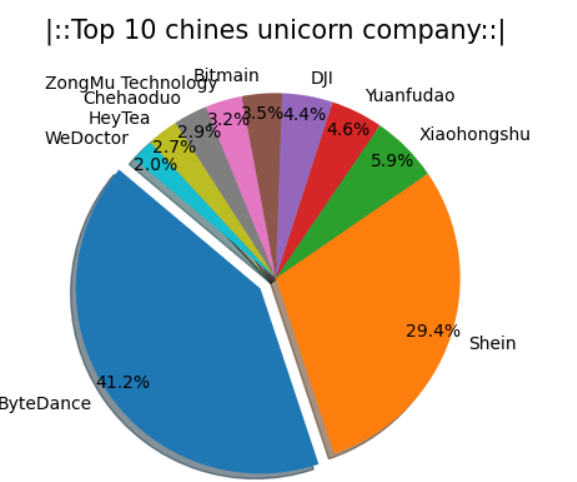
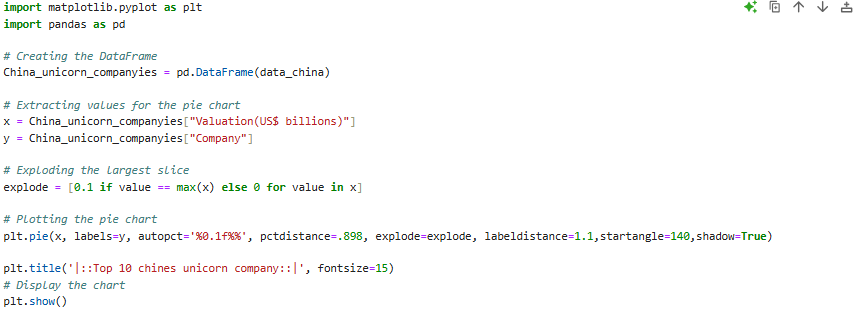




### Pie Chart Insights:

1. **Companies**:
   * **ByteDance**: Valuation of $140.0 billion, Valuation date: June 2021.
   * **Shein**: Valuation of $100.0 billion, Valuation date: June 2021.
   * **Xiaohongshu**: Valuation of $20.0 billion, Valuation date: June 2022.
   * **Yuanfudao**: Valuation of $15.5 billion, Valuation date: October 2020.
   * **DJI**: Valuation of $15.0 billion, Valuation date: September 2016.
   * **Bitmain**: Valuation of $12.0 billion, Valuation date: June 2018.
   * **ZongMu Technology**: Valuation of $11.0 billion, Valuation date: June 2021.
   * **Chehaoduo**: Valuation of $9.5 billion, Valuation date: July 2021.
   * **HeyTea**: Valuation of $9.3 billion, Valuation date: July 2021.
   * **WeDoctor**: Valuation of $6.8 billion, Valuation date: February 2021.
2. **Industries**:
   * The companies span various industries such as Internet, E-commerce, Educational technology, Technology, Cryptocurrency, Self-driving cars, Marketplace, Retail, and Healthcare.
3. **Countries**:
   * All listed companies are based in China.
4. **Founders**:
   * **ByteDance**: Zhang Yiming, Liang Rubo
   * **Shein**: Chris Xu
   * **Xiaohongshu**: Miranda Qu Fang, Charlwin Mao Wenchao
   * **DJI**: Frank Wang (Wang Tao)
   * This table highlights the impressive valuations of leading Chinese companies across various industries, showcasing their significant market presence and financial strength.

**7)Visualize pie chart of Top 10 China’s Unicorn company And Exploding Top one company**



### Pie Chart Insights:

The image is a pie chart titled **"Top 10 Chinese Unicorn Companies."** The chart visually represents the market share percentages of the top 10 Chinese unicorn companies. Here are the companies and their respective market shares:

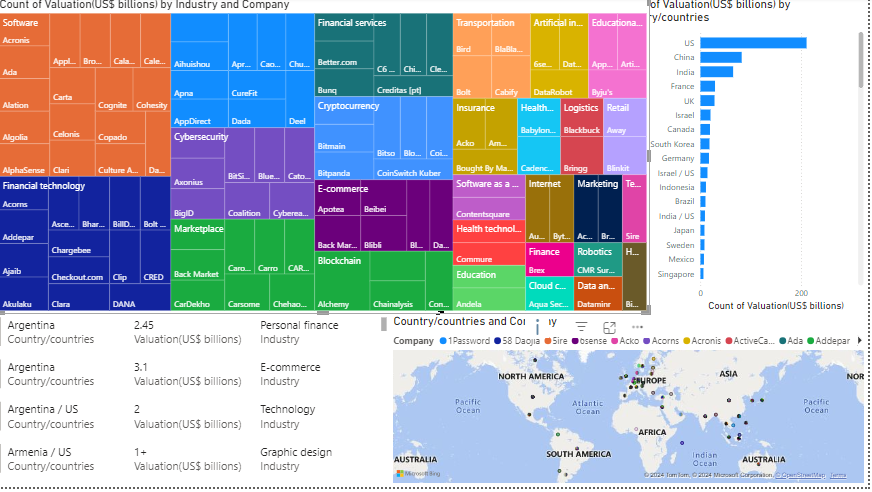
* **ByteDance**: 41.2%
* **Shein**: 29.4%
* **Xiaohongshu**: 5.9%
* **Yuanfudao**: 4.6%
* **DJI**: 4.4%
* **Bitmain**: 3.2%
* **ZongMu Technology**: 2.9%
* **Chehaoduo**: 2.9%
* **HeyTea**: 2.7%
* **WeDoctor**: 2.0%

### Key Insights:

1. **Dominant Player**:
   * ByteDance holds the largest market share at 41.2%, indicating its significant impact and valuation in the market.
2. **Major Competitor**:
   * Shein follows with a 29.4% market share, showcasing its strong market position.
3. **Other Notable Companies**:
   * Xiaohongshu, Yuanfudao, DJI, Bitmain, ZongMu Technology, Chehaoduo, HeyTea, and WeDoctor make up the remaining market shares, each contributing a smaller yet important part to the unicorn landscape in China

This pie chart provides a clear and concise visual comparison of the market shares of the top 10 Chinese unicorn companies, highlighting the leaders and their competitive standings.

POWER BI DASH BOARD:



*The image is a dashboard that visualizes data related to the valuation of companies by industry and country. It contains several key elements:*

*1. \*\*Treemap Chart\*\*:*

*- Displays the count of valuation (in US$ billions) by industry and company.*

*- Different industries are represented by various colors.*

*- Includes industries like Software, Financial services, Cryptocurrency, E-commerce, and more.*

*2. \*\*Bar Chart\*\*:*

*- Shows the count of valuation (in US$ billions) by countries.*

*- Lists countries such as China, the United States, India, and the United Kingdom, with China and the United States having the highest valuations.*

*3. \*\*Map\*\*:*

*- Depicts the geographical distribution of companies across different countries.*

*- Companies are represented by differently colored dots, each color corresponding to a different company.*

*4. \*\*Text Information\*\*:*

*- Provides details at the bottom left of the image, listing several countries, their valuations, and industries.*

*- Example entries:*

*- \*\*Argentina\*\*: 2.45 billion, Personal finance industry*

*- \*\*Argentina\*\*: 3.1 billion, E-commerce industry*

*- \*\*Argentina/US\*\*: 1 billion, Technology industry*

*- \*\*Armenia/US\*\*: 1+ billion, Graphic design industry*

*This dashboard offers a comprehensive visual representation of the top industries and countries in terms of company valuations,*

**conclusion**

*This project provided an extensive analysis of Unicorn Startup Companies using Exploratory Data Analysis (EDA) methodologies, highlighting the importance and application of data science in deriving actionable insights.*

**Exploratory Data Analysis in Data Science:**

EDA emerged as an indispensable tool for understanding and visualizing complex datasets. It facilitated the identification of trends, patterns, and anomalies, ensuring accuracy in deriving insights.

**Dataset Utilization:**

The dataset on Unicorn Companies, sourced from reliable platforms such as Kaggle, served as a foundation for analysis. Key variables like valuation, industry, and geographic distribution were explored, enabling comprehensive sectoral and regional insights.

**Data Preparation:**

Initial steps involved cleaning and pre-processing, including handling missing values, duplicates, and outliers. Techniques like re-indexing and transformation optimized the data for further analysis.

**Key Insights and Visualizations:**

Country Analysis: Countries such as the USA, China, and India dominated the Unicorn landscape. Detailed pie charts and bar graphs effectively illustrated these distributions.

Industry Insights: Key sectors, including technology, finance, and healthcare, were highlighted using line charts and bar graphs.

Valuation Analysis: The top 10 Unicorns by valuation were identified, and their statistics were visualized to discern market leaders.

Regional Comparisons: Comparative pie charts were generated for regions like India, China, and the USA, showcasing their top-performing Unicorns.

Impact of EDA in Decision-Making:

The project underscored EDA’s utility in validating business hypotheses and enabling data-driven strategies. Insights from the analysis can guide investment decisions, regional focus, and sectoral prioritization.

Technological Stack and Tools:

Python libraries like Pandas, NumPy, Matplotlib, and Seaborn proved critical in executing data manipulation and visualization tasks, demonstrating their relevance in modern data analytics.

**Challenges and Limitations:**

Incomplete datasets and reliance on public sources sometimes limited the scope of insights.

Visualizations were subject to interpretational biases, necessitating careful scrutiny.

Future Prospects:

The findings advocate for ongoing monitoring of the Unicorn landscape, integrating advanced analytics like machine learning to predict future trends.

References used for the project

* <https://www.edureka.co/>
* <https://leosias.com/>
* <https://copilot.microsoft.com/>
* <https://www.kaggle.com/>
* <https://openai.com/index/chatgpt/>
* https://www.microsoft.com/en-us/power-platform/products/power-bi