Indian Startup Funding Analysis

Objective:-

To analyze the Indian startup ecosystem by cleaning and exploring investment data, identifying key trends and patterns through descriptive statistics and clustering, and developing KPI-driven dashboards to provide actionable insights for investors, policymakers, and entrepreneurs.

Data Source:-

The dataset used for this project was obtained from Kaggle – Indian Startup Funding Dataset. It contains funding details of Indian startups, including funding amount, investors, funding type, and location.

Data Cleaning:-

The Data cleaning steps involved:-

i) Standardization of column names
ii) Removing unnecessary columns (Sr no)
iii) Checking for duplicate rows
iv) Fixing data types
v) Handling missing values
vi) Detecting and Handling Outliers.
vii) Handling inconsistencies and Inaccuracies
viii) Unit conversion

Exploratory Data Analysis (EDA):- (Univariate)

i) Univariate analysis of Categorical columns:-

1. **startup_name**:-

- This column is a list of all different startups that are being funded.
- There are 2437 different startups, indicating that investments are significantly diverse, and not concentrated to just a few companies. There are over 97% of the startups grouped into the 'other' category, which tells us that those startups

were rarely funded by investors (once mostly). It may indicate that those startups are at an early stage.

- Swiggy and Ola cabs are the 2 startups which have been funded the most followed by Paytm. yet their total contribution in the entire funding is just 2.8%, suggesting that these startups haven't dominated in the funding process even though they were funded the most times. This again tells us that investors haven't focussed only on dominant categories.
- as there is no dominance, the distribution is quite balanced, with almost every startup receiving an equal share in the funding process
- Because the cardinality is too high, one has to take care while encoding or grouping, as encoding (ohe) will lead to sparse data, and incorrect grouping may produce garbage results.
- there are no missing values.

2. **industry_vertical**:-

- This feature represents the sector in which the startup is operating.
- There are 785 unique sectors, showing how diversified each startup is and is not just belonging to major sectors like healthcare, finance, and ecommerce. but, there are over 29% of sectors grouped into 'other' categories showing that those are rarely funded. This indicates that investors have a bleak interest investing in those sectors or it may be that those sectors are risky for investments.
- The consumer internet sector shows strong domination with over 36% of funding done just within this sector followed by technology and ecommerce. This suggests that the ongoing investment trends are mainly in consumer internet followed by technology maybe because its demand is rising and hence letting the investors benefit the most.
- This imbalance due to majority investments in consumer internet makes the distribution skewed.
- one modelling, one has to take care of these overly dominant sectors, as they might even dominate in the modelling giving biased results. Due to high cardinality, one may need to group the similar categories into one in order to reduce the high dimensionality.
- there are no missing values

3. **subvertical**:-

- it represents the sub-sector or subcategory of each startup, coming under a particular sector.
- There are 1815 different subverticals found with 65% of them being rare and grouped under 'other' category. This may indicate that those subverticals belong to those industry verticals which are also found out to be rarely funded.

- Among the top 10 most frequently funded subverticals, we can find out that most belong to consumer internet and ecommerce, which are also most frequently industry verticals. But, they totally still contribute under 5% which aligns with the indication that the funding is diversified. There is no clear domination of a particular subvertical
- Due to no domination, the distribution of funding is balanced.
- due to high cardinality, while modelling, one needs to take care when encoding, which might increase the dimensionality too much. Hence, efficient grouping of subverticals is required.
- 30% of values are missing. Due to this, considering this column while modelling may not be a useful step. As we already have industry verticals, neglecting subverticals while modelling may not give a loss.

4. **city_location**:-

- This column represents different cities where the startups are located.
- The startups span over 98 different locations. This time, only 5% of the locations are grouped under the 'other' category, indicating that in those locations, funding of startups has been done rarely. It may indicate that those locations aren't a hub for startups and this might be a reason investors show a little interest when investing in such locations.
- 33% of all the startups which are situated in Bangalore got funded, followed by NCR region (29%) and Mumbai (18%). This suggests that the funding trends are aligning with the locations which are hubs of startups and have too many startups already. This may also indicate a linear relationship between locations dense in startups to funding done. These 3 locations in total contribute 80% funding, which shows clear domination.
- The domination of the 3 locations makes the distribution skewed and imbalanced, with rest locations getting funded much lesser times comparatively.
- cardinality is comparatively low but still significant hence consider grouping rarely funded locations before encoding to reduce dimensionality
- there are no missing values.

5. **investment_type**:-

- This feature represents the type of investment done
- There are 39 unique types of investments. 1.2% of it is grouped under ‘other’, suggesting how rarely those investment methods have been used.
- Seed funding and private equity investment methods are used the most. The seed funding all in all contributing 45% of all the fundings suggests that many startups are at an early stage and hence are being funded to raise the capital. It also suggests that 45% of all investments are too risky. On the other hand,

private equity investment being 45% too, suggests that many of the startups are also at a later stage and 45% of the investors have gone for a safer and profitable investment. These 2 together contribute 90% of all the investments, hence being totally dominant.
- due to these 2 investment_types being used too frequently, it makes the distribution skewed and imbalanced.
- Cardinality is comparatively lower but still > 15. Hence, grouping must be considered for investment_types that are rarely preferred before encoding.
- There are 0 missing values.

6. **investors_name**:-

- This feature is the temporary copy of the original feature, where each investor is individually presented.
- There are 3220 unique investors, suggesting that many of them, irrespective of being a big name or not, took part in the funding. But, due to 91% of them being categorized as 'other', it indicates that many of them just funded once mostly.
- Out of the frequent investors, the one which has funded the most haven't disclosed themselves (1.4%). The top 10 most frequent investors are all the private investing firms suggesting investments from individuals or low level investments like that of families are classified under 'other' as those have happened rarely.
- Then again, there is no clear domination of any investors, making the distribution balanced.
- As the cardinality is very high, one must consider grouping the less commonly appearing investors before encoding in order to reduce dimensionality. Due to cardinality this high, one may opt for frequency encoding instead of one hot encoding. Also, as the name of investors is not of much use while predicting the target variable, one may even exclude it.
- there are no missing values.

ii) Univariate analysis of Numerical columns:-

1. **amount_in_cr:-**

- This feature represents the amount the investors invested in each startup. The currency is INR and its unit is in crores (cr).
- The distribution is highly positively skewed (27.18), suggesting that most of the investments done are lower in amounts comparatively, and very few of them are extremely huge investments. 25% of investments are very low, lying between 0.14cr to 8.5cr. From the rest, 50% of investments lie between 8.5cr to 156.65cr.

This suggests that the majority of investments (75% of them) are very small compared to its extreme outliers.

- 4.8% of the amounts are outliers, suggesting an alignment with such a high skewness, where 95% are lying within the 0.14-378cr range. all the outliers, though extreme, are valid. no negative amounts found. This suggests that some of the startups are attracting very huge investments, indicating that they might possibly be later stage startups and hence such huge investments are done (not risky but safe).
- there are no missing values
- with a heavy skewness of 27.8, consider transformations like log, box-cox, yeo johnson, and go for binning if required by business or if the transformation doesn't solve the problem of skewness. After transformation is applied, the model may not perform well and get biased towards huge values. The values should be normalized/scaled before modelling.

iii) Univariate analysis of Datetime columns:-

1. **date**:-

- This feature represents the date in which each startup got funded.
- The data spans 2096 days of funding starting from 2015 Jan to 2020 Oct.
- Funding was the highest in 2016. there was a growth in funding from 2015 to 2016, where both the years show the highest contribution in funding. It suggests that the market was at its best position and many startups might have been at its early stage, hence requiring seed funding. Funding shows a decline from 2016 onwards. It may indicate that the early stage startups were funded mostly just once that too in 2015/2016 and then most of them have never been funded since. 2020 saw the least funding aligning with the covid-19 recession and market position in lockdown.
- Funding has been the highest in Dec followed by Aug and Apr. However, the graph is balanced and shows roughly equivalent funding done in all months and no month clearly dominating. This shows that funding is not seasonal.
- Startups have been actively funded the whole year.
- Funding was highest on Sundays followed by Monday and Tuesday. However, the graph here is also roughly balanced and no clear domination is seen.
- 1739 dates are missing. I have left it as NaT, because on imputing mode, the variability decreases to a great extent due to repetitive imputation of the same value for 57% of the date, causing false domination of a particular date, day, month, and year. Also, imputing mode by grouping startups is also avoided because still variability has been reduced to a great extent on doing so. Due to

dates having 57% missing values, any type of imputation can be risky and give false results.

Exploratory Data Analysis (EDA):- (Bivariate)

Note :- amount_in_cr is set to be the **'Target'** column.

i) Bivariate analysis of startup_name vs amount_in_cr:-

- **Leaders** - Paytm has the highest average funding value, being far ahead of others. It is followed by ola cabs which also has a substantially higher average funding done. This could indicate that these startups have a strong market position or are requiring high capital for growth. It may also indicate that these startups may be at a later stage.
- **Laggards** - Healthifyme and holachef have the least average fundings. it could suggest that these startups aren't in a good market position. Or, they might be at an early stage hence not being able to gather huge capital funds.
- **Volatility and outliers** - for startups like Paytm and ola cabs, where median is much lower than mean, it suggests that there might have happened a few extremely large fundings in these companies, thereby pulling up the averages.
- **Stability** - lenskart and urbanclaps are 2 startups having their medians close to means, indicating stable funding. This indicates that these startups have steady fundings. Due to stability, investing in them might be safe as it is easier to forecast.
- **Risk profile** - In startups with high variability, like Paytm and ola cabs, there is seen high growth but also volatility, suggesting they are high risk high reward type of startups.

ii) Bivariate analysis of industry_vertical vs amount_in_cr:-

- **Leaders** - Fintech shows total domination with a very high and unmatchable average funding happening in this sector. It aligns with Paytm also showing domination amongst all startups. This indicates that this sector is heavily funded and that there might be a majority of later stage startups falling under this sector. It could also suggest that there is a rise of fintech in the market place and hence many top investors may have shown deep interest in this sector.
- **Laggards** - food and beverage and edtech have seen least average funding. Though the number of times they are funded is greater than that of fintech which was funded just 10 times, the far lesser avg suggests that majority fundings are

minor. It could also indicate that startups falling under this sector are either small scale or early stage startups.

- **Volatility and outliers** - fintech, ecommerce, and finance are those sectors which have shown mean much higher than median, indicating that few investments would have been very large compared to others, pushing the mean up.
- **Stability** - None of the sectors have shown median being close to mean, indicating that none would have experienced consistent funding rounds.
- **Risk Profile** - all the sectors possess variability but out of all, fintech shows the highest variability, indicating that investing in it has higher risk as well as higher rewards due to strong fluctuations. It also suggests that forecasting for any sector could not be an easy task.

iii) Bivariate analysis of subvertical vs amount_in_cr:-

- **Leaders** - Food delivery with avg funding of 213cr is the leading subvertical. with a greater funding frequency and a greater avg, it indicates that this subvertical has attracted many investors who have funded high amounts. It could also suggest that startups under this subvertical are later stage startups. The average is not too high, suggesting no clear domination of subverticals.
- **Laggards** - learning and micro lending show least avg funding amount, which could suggest that startups under these subverticals are either small scale or early stage. Or, it may also indicate a niche market focus.
- **Volatility and Outliers** - Education and subverticals classified under missing show significant difference in mean and median suggesting majority investments being small and few large deals.
- **Stability** - pharmacy and payment gateway with roughly equivalent mean and median, show consistent funding performance and hence being easier to forecast.
- **Risk Profile** - Education's high variability shows that it is a high risk high reward kind of subvertical. Missing is a bag of all startups which have unspecified subverticals. Hence evaluating risk here is not possible.

iv) Bivariate analysis of city_location vs amount_in_cr:-

- **Leaders** - bangalore has seen greatest avg funding followed by ncr. This aligns with these 2 regions of India being a hub for startups. It could suggest that as these 2 regions might be dense with startups, the funding there will be higher than other regions for that reason. However, the avg fundings of these 2 regions aren't dominating excessively.

- **Laggards** - Ahmedabad and Hyderabad has seen least avg funding. It could be possible because many of the startups could be either small scale or could be in an early stage.
- **Volatility and Outliers** - The leaders, Bangalore and ncr, and Mumbai, pune have significant positive differences between mean and median. This indicates that there are many small fundings and a very few huge fundings done in this region, pushing the mean upwards.
- **Stability** - Jaipur and Ahmedabad have seen minor stability in fundings (having their close to median comparatively). This makes the funding forecasting a bit easier for these regions.
- **Risk Profile** - investing in high variability regions like bangalore, ncr, Mumbai and pune can give high risks as well as high rewards. On the contrary, regions which show least variability, like Jaipur and Ahmedabad, have lower investment risks due to lower fluctuations.

v) Bivariate analysis of investment_type vs amount_in_cr:-

- **Leaders** - Series b investment has shown dominance out of all the investment types. highest avg funding is through series b, with a max funding of 33150cr. it could suggest that later stage startups want huge capitals. It could also suggest that the investors investing to moderately later stages with such high investments are okay with taking moderate risks. It is followed by series D, the investment done at a late stage which is usually one of the highest because at that stage, investment risk is low and startup has market dominance.
- **Laggards** - seed funding, pre series a, debt funding, and seed angel funding have seen least mean values, which also aligns with them being used at a very early stage of startup, where investments are risky and hence huge amounts aren't invested outright.
- **Volatility and Outliers** - the dominators, series b and series d, have significant positive differences in their mean and median values. Series B has a median of just 135cr and mean of 2039cr. It has a max investment value of 33kcr. The max investment so high could be the only reason why it is observed to be a dominator pushing the average so high. series d has a mean of 1049cr and median of 552cr. It has the highest median which aligns with it being one of the investment types in which high amounts are invested in late stage startups. In series d too, most of the investments are revolving around the median, with few high values pushing the average upwards.
- **Stability** - series c and debt funding are comparatively stable, indicating that many investors, when using this type of investment, opt for taking risk within the limits and not going off limits.

- **Risk Profile** - series b showed max investment of 33k cr, much higher than the median of 135cr, suggesting that once an investor took a very high risk investing that huge amount in a moderately late stage startup. It could also be an entry error where a wrong value could have been entered under series b. It needs to be further investigated. Series D and seed round are also used for high risk high reward kinds of investment.

vi) Bivariate analysis of date vs amount_in_cr:-

- 2020 showed the highest mean funding done followed by 2019, indicating market boom of startups in these 2 years. 2016 saw least funding on average but with highest funding frequency, suggesting that majority startups could be in early stage during that time. The overall growth from 2015 to 2020 is seen to be positive but the funding frequency decreased over the years, suggesting that most of the investors could have done seed funding and a few of them preferred to fund in large amounts possibly through series b,c,d or private equity. 2020 has just 5 funding and became the year with topmost avg funding. This indicates that most of the 5 fundings are large comparatively thereby pushing the avg upwards.
- November has seen the highest average funding followed by January. But, these spikes aren't consistent over the years indicating no such seasonality.
- years 2020 and 2019, which have highest average funding, also possess outliers pushing their means upwards. Such a huge difference between mean and median suggests that the claim of 2019 and 2020 having market boom could be false because many of the fundings might be much smaller comparatively with few large funding rounds. November showed significant positive difference between mean and median, suggesting that it led due to few extremely large funding rounds which took place in this month, with majority rounds being small.
- 2016 showed minor steadiness comparatively suggesting presence of less dominant outliers. dec and may showed steady fundings comparatively with lesser fluctuations.
- Highest variability is seen in 2017 and then 2019. Nov and January show highest variability amongst all. This indicates that investing in these months and these years could be risky but also give higher rewards.

SQL KPIs and Dashboarding:-

1) Total funding raised (₹ Thousand Crores)
    - The cumulative of all fundings raised between 2015 and 2020.
    - Helps understand the overall scale of investment in India

2) Total number of deals closed
    - Total number of fundings done between 2015 and 2020
    - It is an indication of the strength of investor engagement and market activity
3) Average deal size (₹ Crores)
    - The average funding amount
    - It is an indication of whether the investors are making large investments (which could suggest that they are mostly funding later stage startups) or low value investments (which could suggest that they are mostly seed funding to early stage startups)

4) YoY% growth over the years
    - Growth rate of cumulative investments done each year
    - Gives an indication whether the funding ecosystem is expanding or contracting. A positive growth indicates investor confidence while negative growth rate indicates reduced investor appetite maybe due to market slowdown.

5) Top 10 funding stages by total funding
    - Top 10 most used investment methods
    - Investors can see which funding stage is attracting most capital. Startups can also identify which stage attracts most capital so that they know when to exactly raise the high funding.

6) Top 5 sectors by CAGR of total funding
    - CAGR is the steady growth % for each year, irrespective of ups and downs. For eg:- if the CAGR of a sector is 37%, it means that sector is seeing 37% growth in total funding all the years between 2015-2020.
    - Indicates which sectors attract high funding consistently. It also indicates which sector could be a fastest growing funding opportunity for the investor

7) Top 5 locations by total funding
    - Locations in India which attracted most funded
    - Indicates which locations the investors need to keep in mind. For people who want to start a startup, this data can also indicate locations with a market boom.

8) Investor concentration: Top contributors to total funding (%)

- % share in total funding by the top contributors
- It indicates either risk or stability for a startup. When concentration is high, it means that just a few investors are controlling a large chunk of total funding. It is risky for a startup ecosystem of a country because if all these investors pull back, the ecosystem might collapse due to it being excessively dependent on those few investors for funding.

9) Startup funding velocity by sector (average days between rounds)
   - how quickly startups in a given sector are able to raise their next round of funding
   - If the average days between rounds is low, it suggests that startups in that sector are able to raise funding frequently, and if average days are high, it means a slow growth of that sector due to the funding raising situation being tough.

10) Startup funding diversity index by sector
   - How funding is distributed across different investors on average in each sector
   - If index value is high, it means that the average number of investors investing in a startup of a sector is high and that funding is more distributed indicating lower risk in that sector.

Median Funding Amount by Cluster:-

Cluster	Median Funding Amount (Cr)
0	55.25
1	27,200.00
2	11,900.00
3	4,250.00
4	1,177.25

- Cluster 1 represents startups receiving the largest median funding, suggesting they may be later stage startups or well established startups.
- Cluster 4 represents startups with relatively low funding needs, possibly indicating early-stage or low-capital sectors.

- Variations among different clusters are significant indicating that different startups have different funding requirements, belong to different industries and attract different investors.

Key Conclusions:-

- Due to a few large deals, the Average funding of many startups is getting pushed upwards to a great extent, making it look like the startup is receiving high funding which in reality is not true.
- Most of the funding is going to startups in Bangalore, Delhi-NCR, and Mumbai — other cities get far less.
- Sectors like consumer internet and fintech are getting the most money and attention.
- Many deals are at the very early stage (seed funding) or very late stage (private equity).
- Over half of the funding records have no date, which makes it hard to track changes over time like year over year growth which may produce inaccurate results.
- Grouping startups into “funding clusters” shows clear ranges — some get small amounts, some medium, and a few get huge investments.

Recommendations:-

- Always check median funding along with the average so that very large deals don’t mislead the results.
- Check unusually large funding values (like one Series B round showing ₹33,150 crore) – confirm if they’re correct or typing errors.
- Try to fill in missing dates from other sources; if not possible, mark them clearly so they don’t affect time-based charts.
- Look more closely at other cities (apart from the 3 metro cities) and less-funded sectors – there may be good opportunities there.
- Clearly show the number of deals in each category to give context for averages.