INDIAN STARTUP FUNDING ECOSYSTEM ANALYSIS

1. Introduction/Objective

Funding plays a pivotal role in providing the necessary resources and support for startups to thrive. The project focuses on analyzing the funding received by start-ups in India from 2018 to 2021. The main objective is to gain insights into the Indian start-up ecosystem and propose the best course of action for our team's venture. By analyzing the data on funding amounts, start-up details, and investor information, we aim to unearth prevailing patterns and gain insights about the opportunities in India's start-up ecosystem to inform decision-making. To help deduce various insights the following guiding questions were used:

1.1. Research Questions

i. What is the overall trend in funding received by start-ups in India from 2018 to 2021?

ii. Which industries or sectors have received the highest funding during this period?

iii. What is the distribution of startups across the cities in India?

iv. What is the average funding amount received by start-ups in India during this period?

v. Is there a correlation between the funding amount and the number of investors involved in funding rounds?

To help draw conclusions based on the insights deduced, the following hypothesis were tested:

1.2. Hypothesis

i. **Hypothesis 1:**

Null: The funding received by start-ups in India has not demonstrated consistent upward trajectory over the years.

Alternate: The funding received by start-ups in India has demonstrated a consistent upward trajectory over the years.

ii. **Hypothesis 2:**

Null: There are no significant disparities in funding received by all sectors of the Indian startups.

Alternate: The technology sectors receive higher funding compared to other industries.

iii. **Hypothesis 3:**

Null: Situating a startup in a particular city does not influence funding.

Alternate: Situating a startup in a particular city significantly affects funding.

iv. **Hypothesis 4:**

Null: There are no significant disparities in funding received among different stages of the Indian startups.

Alternate: During different stages the startups received different funding amounts.

2. Project Structure

The Cross Industry Standard Process for Data Mining (CRISP-DM) framework was fully adopted in conducting the data analysis process on this project. The sequential steps followed were:

i. **Business understanding**. This stage involved setting the project objective.

ii. **Data understanding**. This stage verifying the quality of our data (is it clean?) and later exploring and describing our data.

iii. **Data preparation**. This stage entailed cleaning, integrating and reformatting our data.

iv. **Data Visualization and storytelling**.

v. **Deployment.**

2.1. Data understanding

The data used in this project was a combination of 4 different datasets. The dataset encompasses information on the Indian startup ecosystem for the years 2018, 2019, 2020, and 2021. Each dataset contains crucial details about funding stages, companies, founding years, funding amounts, investors, sectors, company information, founders, and headquarters locations.

2.2. Data preparation

During this stage, the datasets were cleaned separately, merged thereafter to form an aggregate dataset used in the analysis. The data preparation stage entailed:

i. Detecting and dealing with missing values and duplicates.

ii. Splitting Location and Industry Columns: The location and industry columns contained multiple values separated by commas. Only the first value was selected as the primary sector.

iii. Currency Conversion for 2018 Amounts: In the 2018 dataset, the amounts column contained a mix of Indian Rupees (INR) and US Dollars (USD). To standardize the amounts, the Indian Rupee was converted to US Dollars (USD).

iv. Removing commas and currency signs from the Amount column in all datasets. This allowed the amounts to be properly recognized as numeric values.

v. Correcting misplaced/erroneous values especially in the 2021 dataset.

vi. Dropping an extra column named "*column10*" in the 2020 dataset: This was to align the columns as this column was not present in other datasets.

vii. Adding "*Funding Year*" column to every dataset which would enable analyzing the

viii. Funding trends over time.

ix. Renaming Columns and Concatenation.

x. Data type conversions (e.g., numeric data mistakenly encoded as strings)

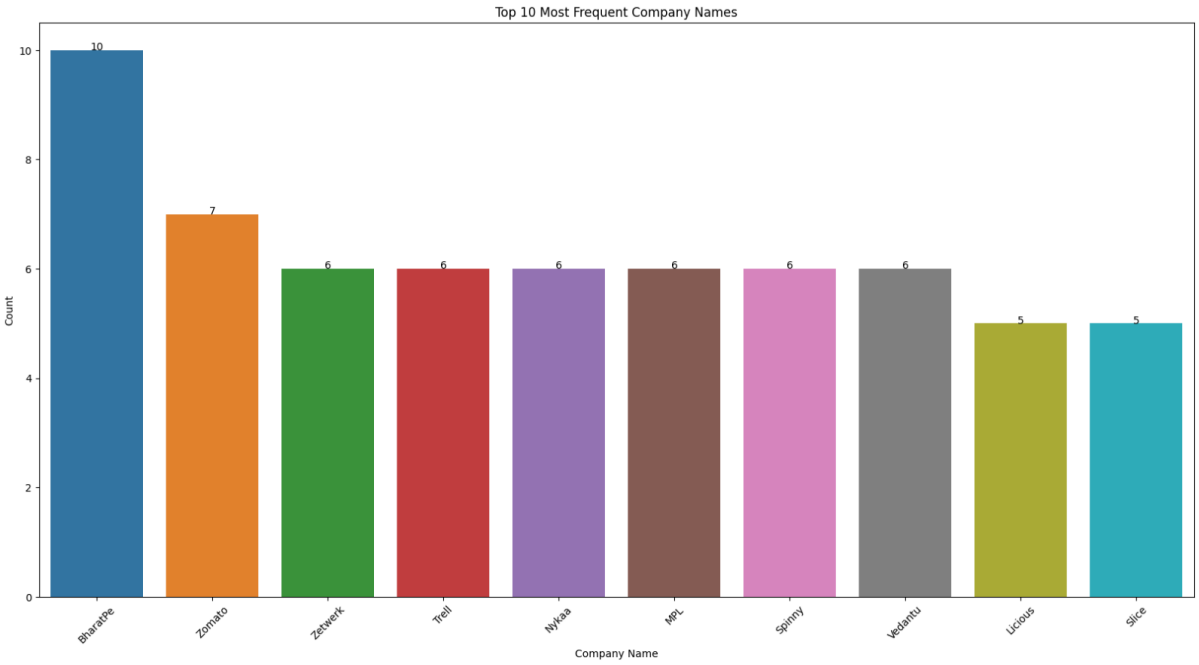
2.2.1. Exploratory data analysis of the merged dataset

This stage involved conducting an initial exploration of the merged dataset to gain insights and identify patterns or trends, generating summary statistics, visualizations, and descriptive analysis to understand the distribution, relationships, and characteristics of the data.

Univariate Analysis

Univariate analysis is a component of exploratory data analysis (EDA) that focuses on examining and interpreting individual variables in isolation. By scrutinizing a variable independently, researchers can uncover valuable information about its properties without considering the influence of other variables. In this each column of the merged dataset was closely examined.

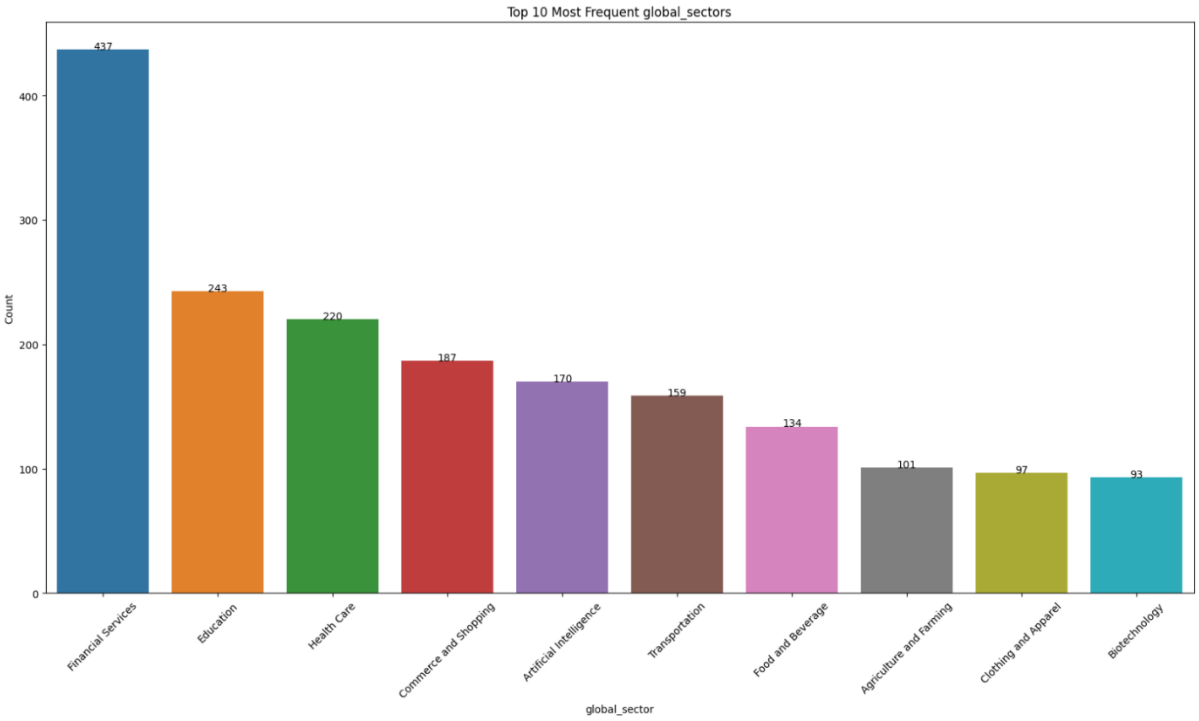
i. **The company column:**



ii. ***Sector* column:**

The recategorized sector column has 2839 rows with 48 unique sectors. '**Financial Services**' occurred most at 361 times.

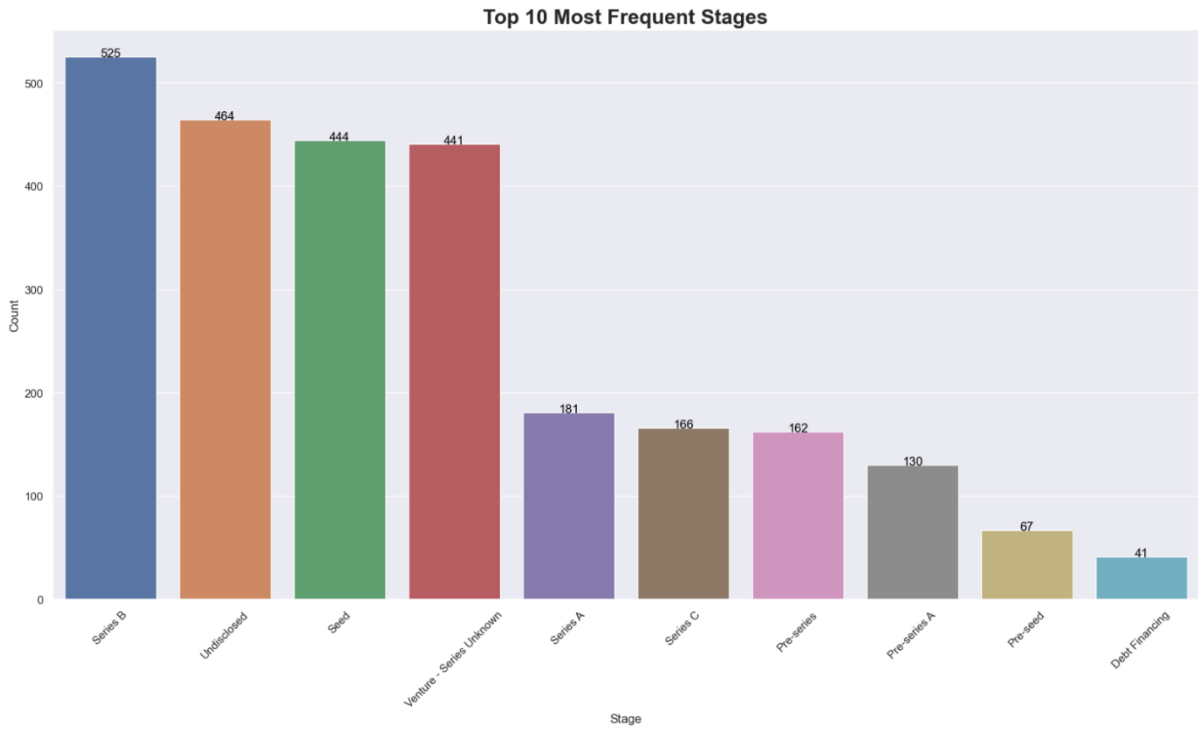
This indicates that, the financial services sector is the most dominant in the Indian start-up ecosystem for the period under review.



iii. ***Stage* column:**

The Series B funding stage was the most common funding stage at which Indian start-ups obtained funding. It dominated with 525 funding deals from 2018 to 2021.

Let us get the visual impressions about these stages and their distribution per counts.



i. The diagram above depicts the fact that most funded stage in the Indian start-up funding ecosystem is the **Series B stage**. 361 successful funding was obtained at this stage throughout the period.

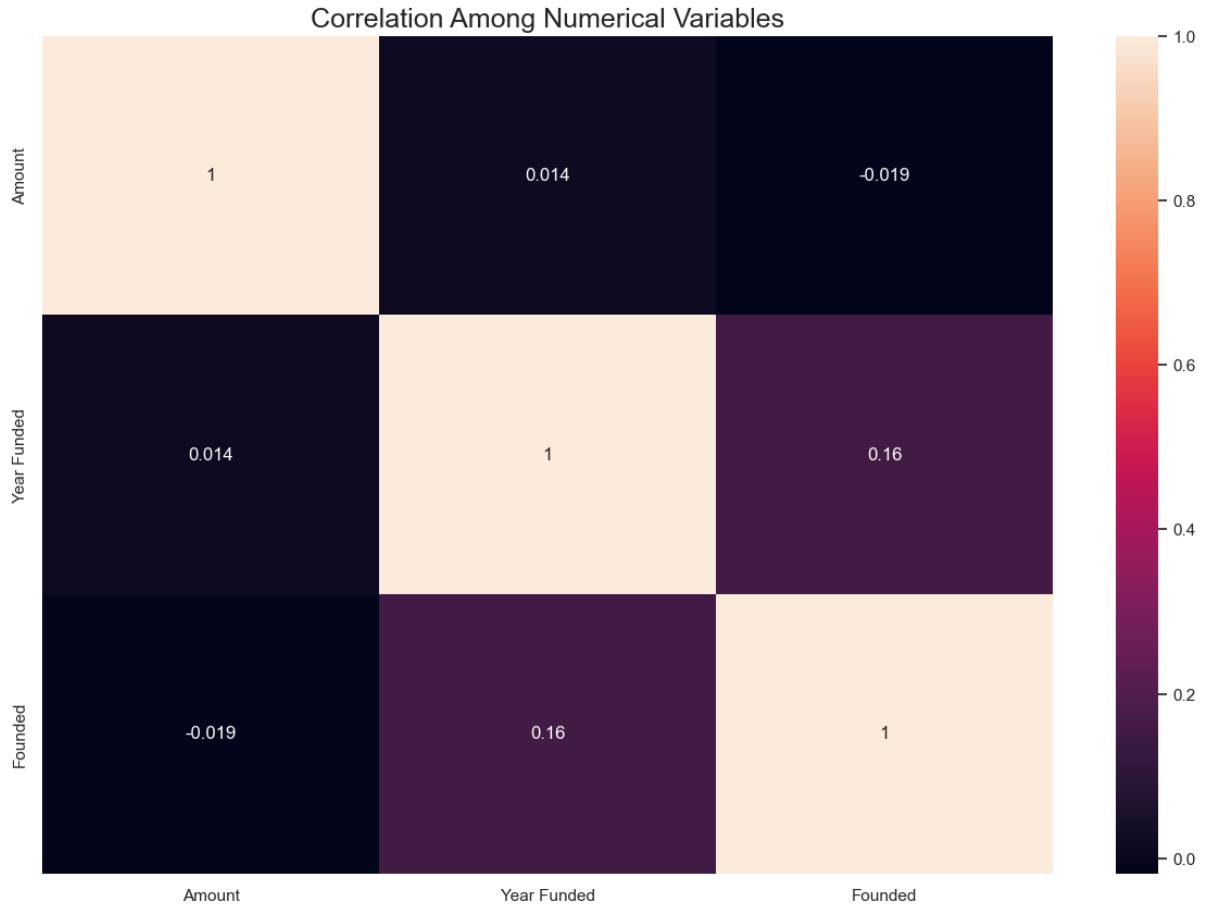
ii. However, the number of start-ups in the said ecosystem whose funding stages were not disclosed are just 61 short of that of the top funded stage.

iii. Also, the third most stage at which funding was obtained among these said start-ups is the **Seed** stage. Funding was obtained 444 times at this stage from 2018 through to 2021.

iv. Meanwhile at the venture equity stage whose series are unknown, 441 funding deals were obtained then a drastic drop in number of deals occurred with Series A, Series C, Pre-series, Pre-seed and the last for the top ten have been at the debt financing stage.

Multivariate analysis

To examine the correlation between the various variables in the data, a correlation heatmap was used. This form of exploratory data analysis seeks to find relationships among numeric variables.



i. From our correlation heatmap, there seems to be a very weak positive relationship between the year funded and funding amount received.

ii. For founding years of companies and funding obtained, there exists a weak negative relationship.

iii. This implies that the amount funded does not depend on the year in which it was received. Also, it does not rely on the age of such companies.

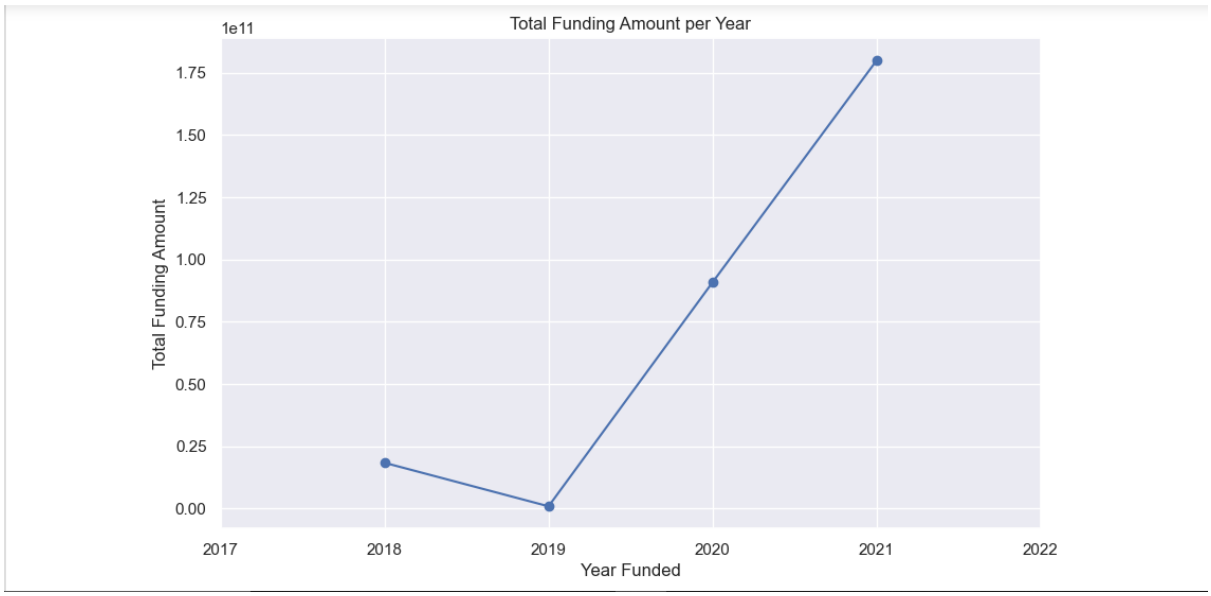
3. Data Visualization and storytelling.

This involves using compelling visuals and charts to present the data in an understandable and meaningful way. To establish a story told by the data on the overall funding of various startups in India, visualizations of various factors that affect the funding were displayed in a dashboard using PowerBI and later published to Power BI service. To help visualize and draw insights the research questions were used as a guide:

3.1. Through Research Questions

3.1.1. What is the overall trend in funding received by startups in India from 2018-2021?

A line graph of Total Funding Amount against the year of funding was plotted. To address this question, we used the 'groupby' function in python to group the data according to 'Year Funded' and aggregate 'Amount' by sum.

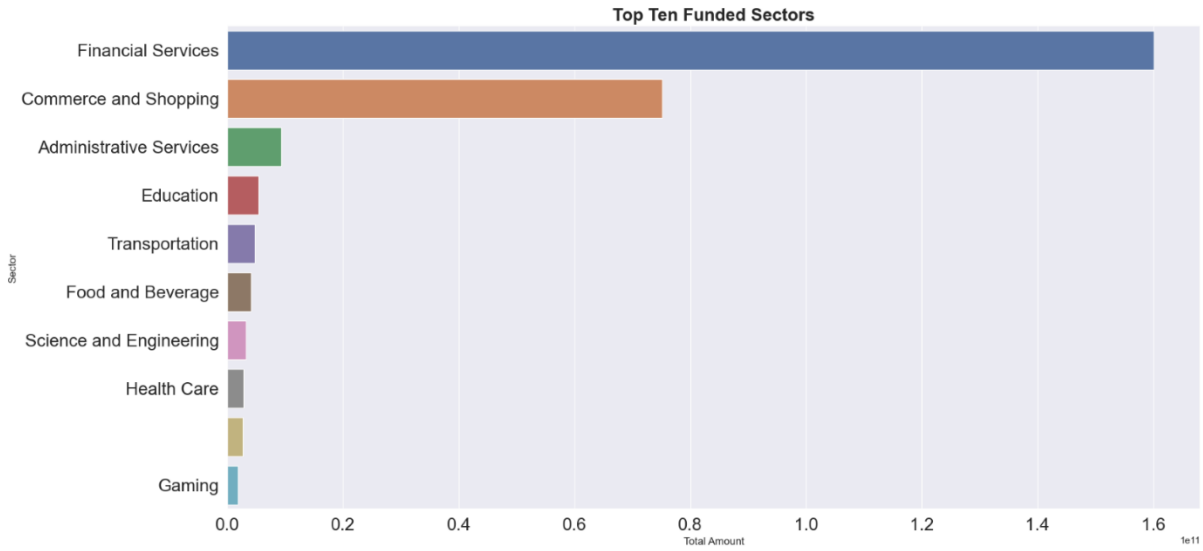


Insights Drawn

For the period 2018 -2019, there was a downward trend in the amount of money funded into startups. From 2019, the money funded has steadily increased.

3.1.2. Which industries or sectors have received the highest funding during this period?

A bar plot of Sector against the total funding amount was plotted. The data was grouped by the ***Sector*** column sum of ***Funding Amount*** calculated.

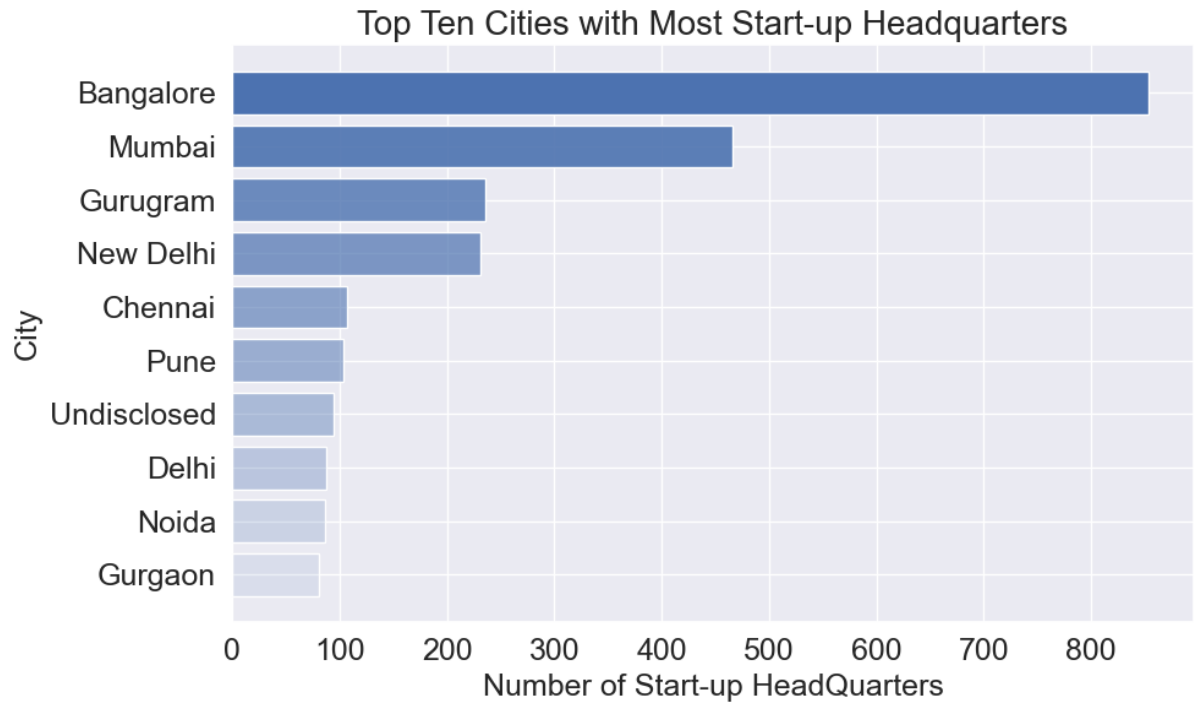


Insights Drawn

From the graph above, ***Financial Services*** is the highest funded sector, which is then followed by ***Commerce and Shopping*** while the ***Gaming*** sector was the least funded.

3.1.3. What is the Distribution of Start-ups Across the Cities in India?

This question answered the total amount of startups across the cities funded across the time period 2018-2021. A bar plot of Cities against the start-up count per city was plotted

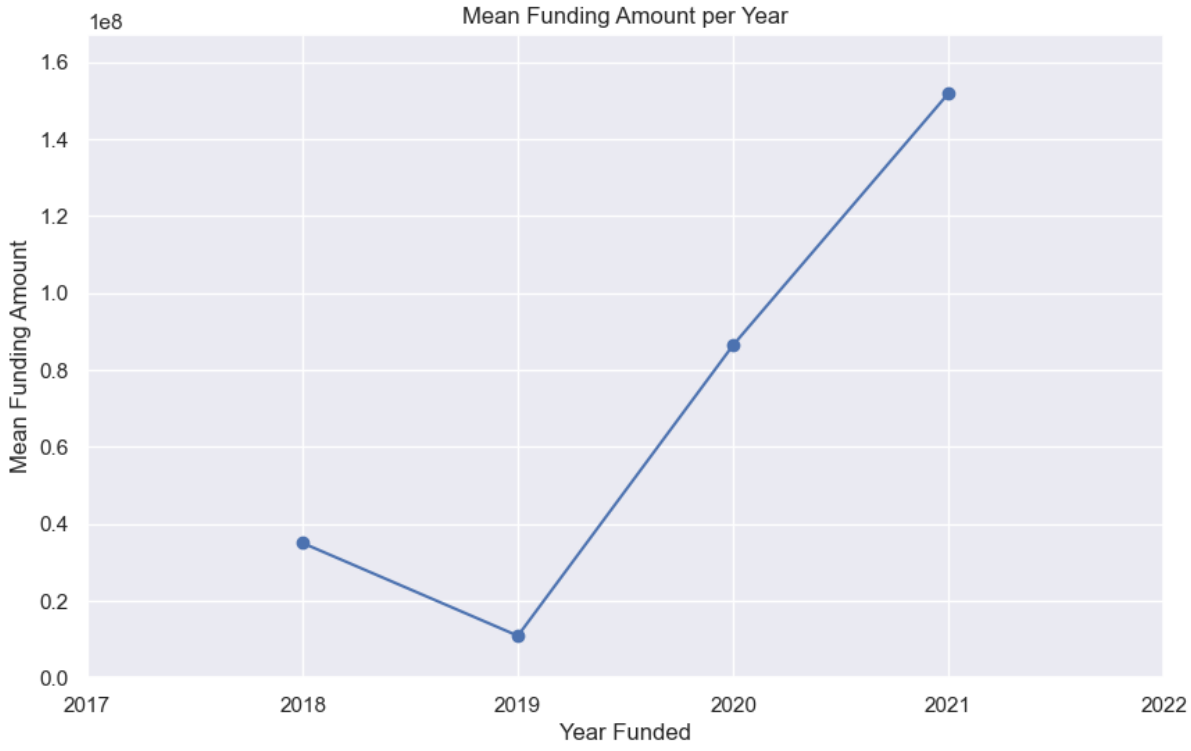


Insights Drawn

Most funded startups were in Bangalore. This can be explained by most startups having their headquarters located in ***Bengaluru (Bangalore),*** which is in the Indian State of Karnataka followed by ***Mumbai*** then ***Gurugram*** and ***New Delhi***.

3.1.4. What is the average funding amount received by start-ups in India during this period?

To illustrate the trend of average funding received by start-ups over the period of 2018 to 2021, we grouped the data by 'Year Funded' and aggregated the average (mean) of 'Amount'.

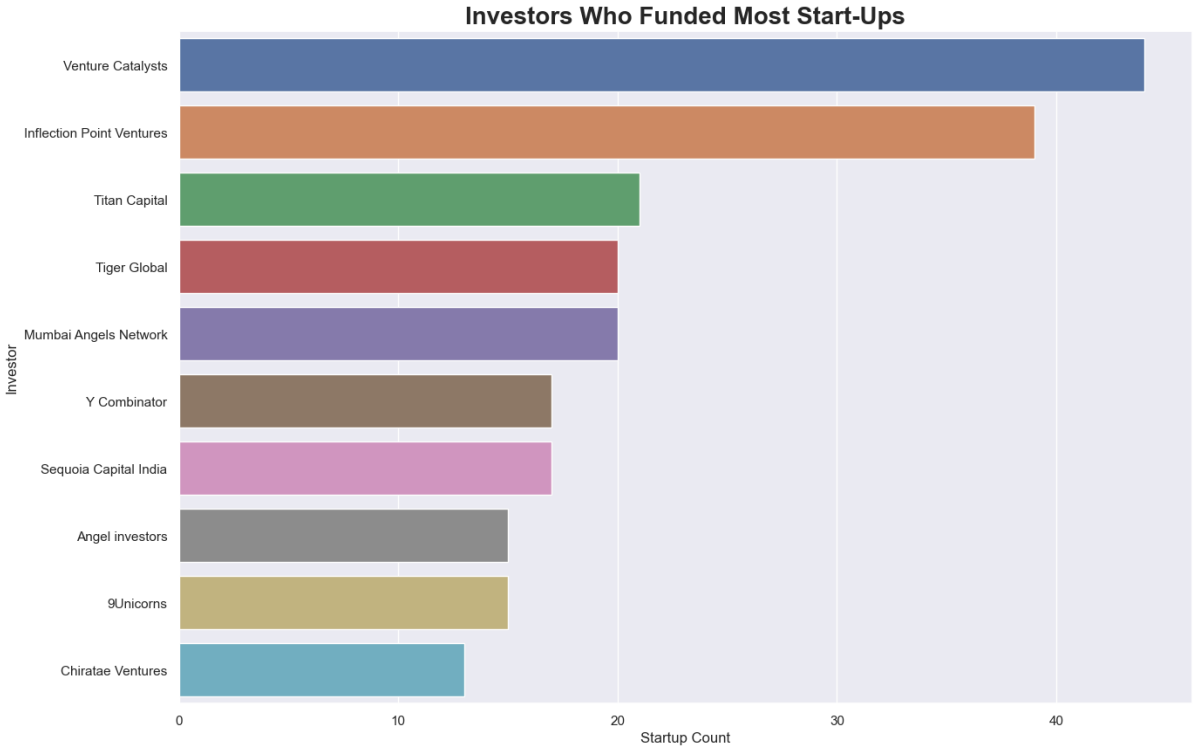


Insight Drawn

Following the same pattern exhibited by total funding, there was a downward trajectory in the average money funded in startups 2018-2019.There was a sharp increase in the average amount invested in startups from 2019-2020.

3.1.5. Which Investors Funded Most Start-Ups Over the Period?

A bar plot of investor against the startup count was plotted. The data was grouped by the Investor and a total count of startup calculated.



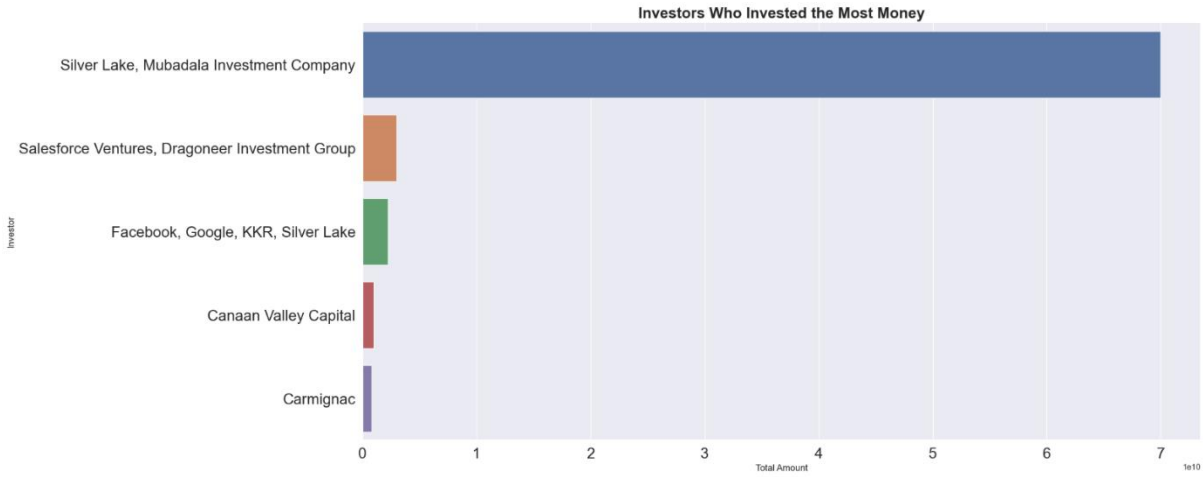
Insight Drawn

i. For most startups, their investors were either Unknown or Undisclosed. We assume it makes sense to expurgate such information from our analysis since it will not be of much benefit to our stakeholders.

ii. The leading investor that had invested in most startups was therefore Venture Catalysts, which was closely followed by Inflection Point Ventures. Thereafter came Titan Capital, Tiger Global and Mumbai Angels Network.

3.1.6. Which Investors Funded the Most Money on Average Over the Period?

A bar plot of investor against the total amount funded was plotted. The data was grouped by the Investor and sum of funds invested calculated.

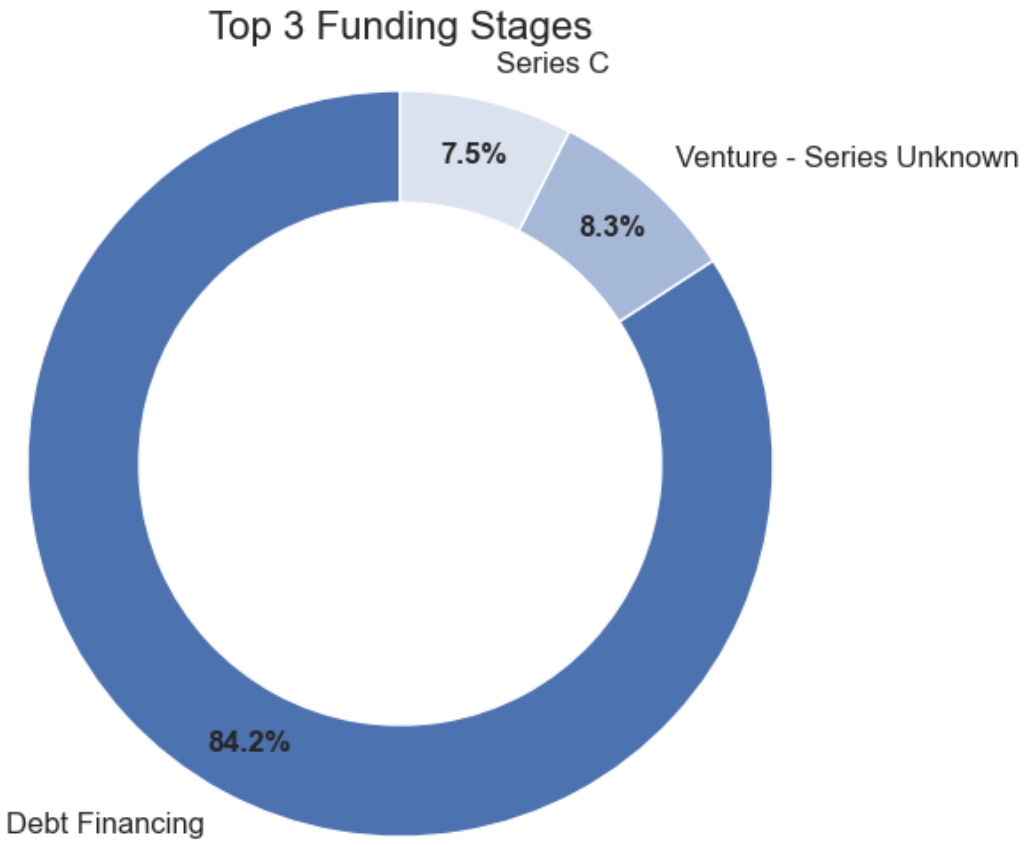


Insight Drawn

Silver Lake and Mubadala Investment Company are the leading investors who funded the highest average amount into startups.

3.1.7. At What Stages were Most Start-ups Funded?

A doughnut chart displaying the top three stages when the companies were funded is plotted. The data was grouped by the Stage and a total count of startup calculated.

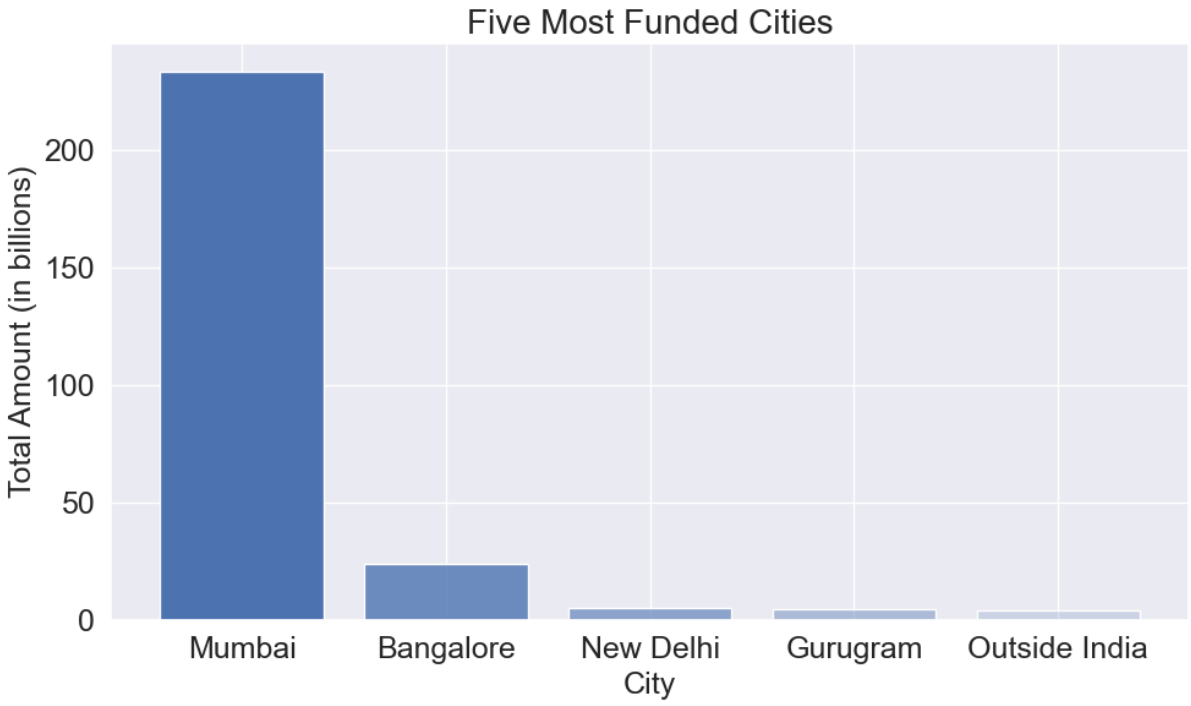


Insight Drawn

Dept financing recorded the highest funding in the Indian start-up ecosystem with a whopping ***84.2 percent*** share among three top stages of funding over the period of 2018 to 2021. This was followed by venture equity whose series was not known then Series C. These were just 8.3% and 7.5% respectively.

3.1.8. Which 5 Cities Recorded Most Funding?

A bar plot of Cities against the total amount funded was plotted. The data was grouped by the Cities and a sum of funds invested calculated.



Insight Drawn

According to the analysis, the most highly funded start-ups are in Mumbai as it received over 233 billion dollars over the period. The next was Bangalore, which had a little over 24 billion dollars. New Delhi then came with about 5 billion dollars with just minute gaps from Gurugram and those located outside India.

3.2. Hypothesis Testing

ANOVA is used to test the hypothesis at a significance level of 0.05 against the p-value. If the p-value is less than the alpha level of significance, we will reject the null hypothesis otherwise we don't reject. To choose between the non-parametric and parametric ANOVA, the distribution of the amount column is investigated. If it is normal, we use the parametric one-way ANOVA otherwise we use the non-parametric ANOVA known as Kruskal Wallis.

Shapiro Wilk Test

To test the distribution of the data the ***Amounts*** column is tested for normality using the Shapiro wilk test and the test below obtained.

```python
# Group the data by 'Year Funded'
grouped_data = df.groupby('Year Funded')['Amount']

# Perform Shapiro-Wilk test for each group
for year, group in grouped_data:
    statistic, p_value = stats.shapiro(group)
    print(f"Year: {year}")
    print("Shapiro-Wilk Test Results:")
    print("Statistic:", statistic)
    print("P-value:", p_value)
    if p_value < 0.05:
        print("The data does not follow a normal distribution.")
    else:
        print("The data follows a normal distribution.")
    print("-" * 30)
```

The results:

```
Year: 2018
Shapiro-Wilk Test Results:
Statistic: 0.16452401876449585
P-value: 6.291830104818429e-43
The data does not follow a normal distribution.
------------------------------
Year: 2019
Shapiro-Wilk Test Results:
Statistic: 0.7388180494308472
P-value: 1.5155211807726943e-10
The data does not follow a normal distribution.
------------------------------
Year: 2020
Shapiro-Wilk Test Results:
Statistic: 0.015582144260406494
P-value: 0.0
The data does not follow a normal distribution.
------------------------------
Year: 2021
Shapiro-Wilk Test Results:
Statistic: 0.012447237968444824
P-value: 0.0
The data does not follow a normal distribution.
------------------------------
```

Insights Drawn:

The distribution is not normal hence nonparametric ANOVA is used to test the hypothesis.

3.2.1. Hypothesis one:

To test Hypothesis 1, we will analyse the year-by-year funding amounts and calculate the average growth rate of funding.

Conclusion: There are significant disparities in funding received during different years.

To further analyze the differences, we use the post hoc Dunn test to establish the differences in years.

```python
def group_comparisons_by_pvalues(posthoc_result, alpha=0.05):

    is_significant = []  # p < alpha
    not_significant = []  # p > or = alpha

    for group1 in posthoc_result.index:
        for group2 in posthoc_result.columns:
            p_value = posthoc_result.loc[group1, group2]
            if p_value < alpha :
                is_significant.append((group1, group2))
            else:
                not_significant.append((group1, group2))

    return is_significant, not_significant

# Group the comparisons based on p-values
is_significant, not_significant = group_comparisons_by_pvalues(dunn_result, alpha=0.05)

# Print the results
print("Significant Differences:")
print(is_significant)


print("\nNot Significant Differences:")
print(not_significant)
```

```
Significant Differences:
[('2018', '2020'), ('2019', '2020'), ('2020', '2018'), ('2020', '2019'), ('2020', '2021'), ('2021', '2020')]

Not Significant Differences:
[('2018', '2018'), ('2018', '2019'), ('2018', '2021'), ('2019', '2018'), ('2019', '2019'), ('2019', '2021'), ('2020', '2020'), ('2021', '2018'), ('2021', '2019'), ('2021', '2021')]
```

3.2.2. Hypothesis 2:

Categorize start-ups based on industry and compare the funding amounts received by each sector.

```python
## print each amount in each sector as a list and get the amount per sector as an array

amounts_per_sector=df.groupby("global_sector")['Amount'].apply(list)
amount_sector = np.array(amounts_per_sector)
#amount_sector
```

```python
# Perform one-way ANOVA
test_statistic, p_value = stats.kruskal(*amount_sector)
print(f"P_value: {p_value}")


# Interpret the results
if p_value < alpha:
    print("Reject the null hypothesis concluding that there are significant disparities in funding received in different sectors.")
else:
    print("Fail to reject the null hypothesis concluding there is no significant evidence of disparities in funding among the sectors.")
```

```
P_value: 1.3299622453460806e-14
Reject the null hypothesis concluding that there are significant disparities in funding received in different sectors.
```

Conclusion: There are significant disparities in funding received in different sectors.

3.2.3. Hypothesis 3:

Examine the distribution of start-ups across cities and deduce which cities harbor most highly funded start-ups.

```
# Perform one-way ANOVA
test_statistic, p_value = stats.kruskal(*amount_city)
print(f"P_value: {p_value}")


# Interpret the results
if p_value < alpha:
    print("Reject the null hypothesis concluding that there are significant disparities in funding received in cities.")
else:
    print("Fail to reject the null hypothesis concluding there is no significant evidence of disparities in funding among the cities.")
```

```
P_value: 3.8317504844268007e-07
Reject the null hypothesis concluding that there are significant disparities in funding received in cities.
```

Conclusion: There are significant disparities in funding received in cities.

Investigating the differences among the cities

```
def group_comparisons_by_pvalues(posthoc_result, alpha=0.05):

    is_significant = []  # p < alpha
    not_significant = []  # p > or = alpha

    for group1 in posthoc_result.index:
        for group2 in posthoc_result.columns:
            p_value = posthoc_result.loc[group1, group2]
            if p_value < alpha :
                is_significant.append((group1, group2))
            else:
                not_significant.append((group1, group2))

    return is_significant, not_significant

# Group the comparisons based on p-values
is_significant, not_significant = group_comparisons_by_pvalues(dunn_result, alpha=0.05)

# Print the results
print("Significant Differences:")
print(is_significant)


print("\nNot Significant Differences:")
print(not_significant)
```

```
Significant Differences:
[('Bangalore', 'Hyderabad'), ('Hyderabad', 'Bangalore'), ('Hyderabad', 'Outside India'), ('Outside India', 'Hyderabad')]

Not Significant Differences:
[('Ahmadabad', 'Ahmadabad'), ('Ahmadabad', 'Ahmedabad'), ('Ahmadabad', 'Alleppey'), ('Ahmadabad', 'Alwar'), ('Ahmadabad', 'Ambernath'), ('Ahmadabad', 'Anand'), ('Ahmadabad', 'Andheri'),
```

Conclusion: There are significant differences among these pairs of cities ('Bangalore', 'Hyderabad'), ('Hyderabad', 'Bangalore'), ('Hyderabad', 'Outside India'), ('Outside India', 'Hyderabad').

3.2.4. Hypothesis 4

Categorize start-ups based on stages and compare the funding amounts received by each stage.

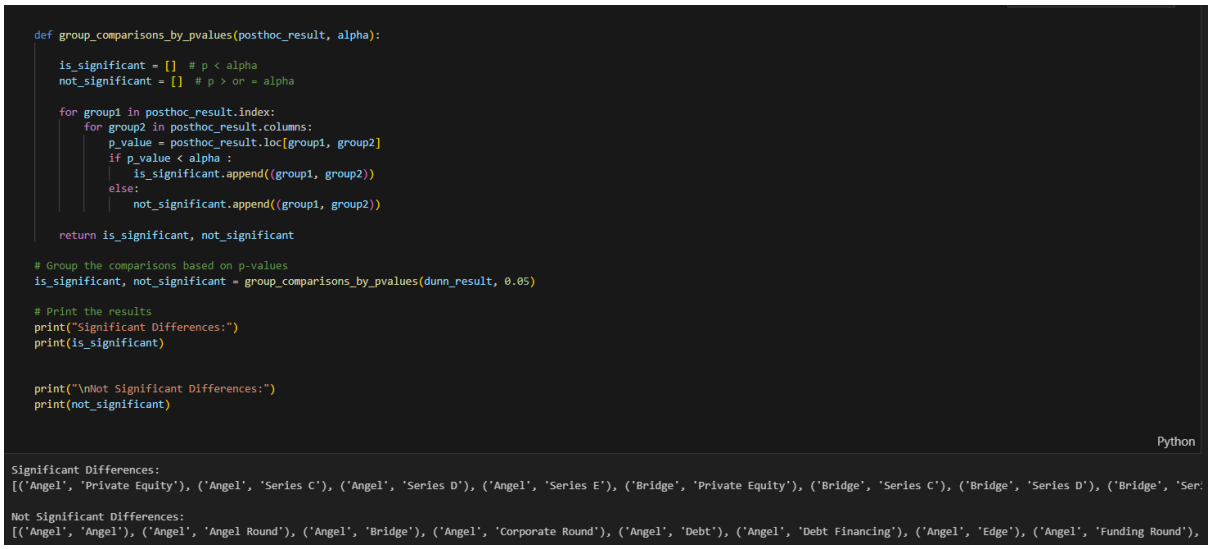
```
# Perform one-way ANOVA
test_statistic, p_value = stats.kruskal(*amount_stage)
print(f"P_value: {p_value}")


# Interpret the results
if p_value < alpha:
    print("Reject the null hypothesis concluding that there are significant disparities in funding received during different stages.")
else:
    print("Fail to reject the null hypothesis concluding there is no significant evidence of disparities in funding among the stages.")
```

```
P_value: 1.6442006660313595e-120
Reject the null hypothesis concluding that there are significant disparities in funding received during different stages.
```

Conclusion: There are significant disparities in funding received during different stages.

Identifying the differences:



```
def group_comparisons_by_pvalues(posthoc_result, alpha):

    is_significant = []  # p < alpha
    not_significant = []  # p > or = alpha

    for group1 in posthoc_result.index:
        for group2 in posthoc_result.columns:
            p_value = posthoc_result.loc[group1, group2]
            if p_value < alpha :
                is_significant.append((group1, group2))
            else:
                not_significant.append((group1, group2))

    return is_significant, not_significant

# Group the comparisons based on p-values
is_significant, not_significant = group_comparisons_by_pvalues(dunn_result, 0.05)

# Print the results
print("Significant Differences:")
print(is_significant)


print("\nNot Significant Differences:")
print(not_significant)
```

Python

```
Significant Differences:
[('Angel', 'Private Equity'), ('Angel', 'Series C'), ('Angel', 'Series D'), ('Angel', 'Series E'), ('Bridge', 'Private Equity'), ('Bridge', 'Series C'), ('Bridge', 'Series D'), ('Bridge', 'Ser

Not Significant Differences:
[('Angel', 'Angel'), ('Angel', 'Angel Round'), ('Angel', 'Bridge'), ('Angel', 'Corporate Round'), ('Angel', 'Debt'), ('Angel', 'Debt Financing'), ('Angel', 'Edge'), ('Angel', 'Funding Round'),
```

4. Overall Conclusions.

i. Funding dropped from 2018 to 2019 by over 17 billion dollars before rising steadily through to 2021. We therefore fail to reject the null hypothesis which states that the funding received by start-ups in India has not demonstrated consistent upward trajectory over the years.

ii. There are significant differences on the funding amounts received in through out the years. This can be proven by the post hoc Dunn test that illustrates the significant differences among the following pair of years ('2018', '2020'), ('2019', '2020'), ('2020', '2018'), ('2020', '2019'), ('2020', '2021'), ('2021', '2020').

iii. There are significant disparities in funding received during various sectors. Comparing the other sectors to the technology sectors, there was no significant evidence of a difference in funding amounts between the Tech sector and other sectors.

iv. There are significant disparities in funding received during various stages. The debt financing stage received the highest funding over the period with 150.7 billion. Venture capital with unknown series, however, came second with 14.9 billion and Series C with 13.4 billion.

5. Recommendations

i. Situation analysis for a venture into the Indian ecosystem regarding funding of start-ups is particularly important as it is highly likely to receive funding as a start-up when headquartered at Mumbai, Bangalore, New Delhi and Gurugram respectively.

ii. Alternative financing sources must be implored to support a start-up since most of the start-ups failed to receive highly significant funding at their initial stages. The earliest highly funded was at the Series C stage when it is expected to be a market fit.

iii. Considering the sector to venture into as far as funding for a start-up in India is concerned is vital as it is more likely to receive funding when at the financial services, commerce and shopping, administrative and education sectors, respectively.