The primary goal of this report is to analyse the factors that influence startup funding and devise strategies to make my startup more attractive to investors. I will achieve this by analysing funding data from the past five years (2017- 2021) to uncover trends and patterns that affect investment decisions. This includes assessing the impact of location on funding and identifying sectors that attract significant investment. I used Multiple Imputation by Chained Equations (MICE) to address missing values, creating multiple complete datasets to capture uncertainty in the missing data. By analysing each imputed dataset and pooling the results, I obtained a robust average scenario from the imputations, providing a solid foundation for strategic decision-making.

PART 1: PREPARING DATA

The dataset comprises 84 monthly datasets across 7 folders, each corresponding to a different year. I created a function to process each file, standardise column names and compile the DataFrames into a single dataset for easier analysis. The dataset had issues with repeated category names due to formatting errors, like 'PrivateEquity' versus 'Private Equity'. I implemented functions to standardize these names, reducing category variations and enhancing data consistency. The 'IndustryVertical' column may still contain a few entries that are unformatted, representing the same category. Initially, there were 1,253 category counts, but after standardisation and consolidation, we reduced them to 239, which is a manageable number that will not significantly impact our later analysis. The dataset featured high percentage of missing values across column, about 37% on average, especially in the 'InvestmentStage', 'founding_year' and 'Amount(inUSD)' columns. Removing these missing values would result in substantial data loss and compromise the robustness of our analysis. Given the variety of imputation methods available, it is crucial to understand the nature of the missing data. There are three main types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR occurs when the probability of a data point being missing is independent of any observed or unobserved data, whereas MAR is related to the observed data but is independent of the missing data itself. Lastly, MNAR occurs when the missingness is related to the value of the data itself, either observed or unobserved, requiring more sophisticated methods for estimation due to the complexity of the missingness mechanism (Donders *et al.*, 2006). Understanding these distinctions is critical to ensures the validity of the imputation results and helps avoid drawing incorrect conclusions based on biased or inappropriate imputations. I analysed the dataset's missing data patterns using graphical methods to determine if the data is MAR. I visualised the missing data with a matrix display showing the nullity matrix and a heatmap to explore relationships between variable pairs in terms of missingness. The analysis showed strong positive correlations between certain variables, like 'City' and 'IndustryVertical', and 'InvestmentType' and 'Amount(inUSD)', suggesting a systematic cause for the missingness and indicating that the data may be MAR, where missingness is related to observed variables. Given this indication, I have chosen to employ MICE. It is a robust imputation technique that can effectively address MAR values, as it operates under the assumption that, given the variables used in the imputation procedure, the missing data are MAR (Mongin *et al.*, 2019). In the MICE procedure, missing data are imputed using a series of regression models, each tailored to the type of variable being imputed—linear for continuous and logistic for binary variables. The process starts with simple placeholder imputations, such as mean values, which are iteratively refined using regression estimates that incorporate all available data, including previously imputed values. This iterative process is repeated for each variable until the imputation model stabilizes, resulting in a complete dataset with all missing values filled by statistically informed predictions (Zhang, 2016).

PART 2

Our analysis reveals that from 2017 to 2021, a total of 11,751 startups emerged, experiencing a major decline of -4.33% in new startups during 2018 and 2019. However, the sector rebounded with a growth of 3.26% in 2010 and 2021. During the same period, funding trends for Indian startups began robustly in 2017, with a steady increase through 2018, peaking at just above $2.25 billion. A significant downturn occurred in 2020, continuing into 2021, where funding levels dropped to approximately $0.5 billion. This downturn can be attributed to various factors, including economic downturns and global events like the COVID-19 pandemic, which had extensive economic impacts. Investments were predominantly led by the technology and internet sector, followed closely by the financial technology and platforms sector, each securing over a billion dollars. E-commerce and online marketplaces also attracted substantial funding, alongside specialized tech and other services. Regarding regional dynamics, Bengaluru emerged as the leading city in terms of startup investments, receiving over $18 billion. It was followed by Mumbai, Gurugram, New Delhi and Noida as the top Indian cities by funding across key industries. Notable funding figures include Roposo nearing $4 billion, Faasos with about $2.5 billion and Manch and Cashfree, each surpassing $1 billion. Startups typically received the most funding during their early or seed rounds, averaging about $700 million. Later stages, such as Series G, also attracted substantial funding, averaging near $200 million. Funding for Series B and H rounds was about $150 million each, while Series D rounds were slightly lower, at approximately $100 million. This pattern underscores that both early potential and more mature, expansion-focused stages attract larger investments. Annually, the startups receiving the most funding were Legalwiz.in in 2017 with $31 million, Nykaa in 2018 with $32 million, Innovapptive in 2019 with $30 million, Oyo in 2020 with $31 million, and Paytm with $15 million. While Bengaluru remains a major hub for startup funding, other cities like Gurugram, Mumbai and New Delhi also see substantial investments, particularly in prominent firms. However, smaller cities like Ahmedabad, Pune, Chennai and Hyderabad and Delhi, though less frequent in funding events, still host significant investments, highlighting how a company's location can influence investment levels. To explore how a company's location might impact the funding it receives, we structured our analysis around two primary hypotheses:

Null Hypothesis (H0): The amount of funding received by startups in India is not associated with their geographic location.

Alternative Hypothesis (H1): The amount of funding received by startups in India is associated with their geographic location.

Comparing Bengaluru, the city with the highest funding activity, against a collective of other significant cities including Gurugram, Mumbai and Ahmedabad, our findings indicate significant differences in the funding amounts between these cities. Consequently, we reject the null hypothesis, affirming that the funding a startup receives is indeed influenced by its city of operation. Thus, to make my startup attractive to investors, I will concentrate on entering high-growth sectors like technology and fintech, and ensure my product meets a significant market need. Develop and validate a minimum viable product to demonstrate potential and reduce investment risks. I will consider the benefits of being located in major startup hubs such as Bengaluru or Mumbai and plan how my startup can scale up operations and expand its customer base efficiently. Build a strong and diverse team to showcase leadership and expertise, crucial for investor confidence. I will tailor my pitch to align with the expectations of different funding stages, from seed to later rounds. These strategies will help increase my startup's appeal and align with successful trends in the startup ecosystem.

REFERENCES

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T. and Moons, K.G. (2006) ‘A gentle introduction to imputation of missing values’, *Journal of clinical epidemiology*, 59(10), pp.1087-1091

Mongin, D., Lauper, K., Turesson, C., Hetland, M.L., Kristianslund, E.K., Kvien, T.K., Santos, M.J., Pavelka, K., Iannone, F., Finckh, A. and Courvoisier, D.S. (2019) “Imputing missing data of function and disease activity in rheumatoid arthritis registers: what is the best technique?”, RMD open, 5(2), p.e000994

Zhang, Z. (2016) “Multiple imputation with multivariate imputation by chained equation (MICE) package”, *Annals of translational* medicine, 4(2)