Indian Startup Funding

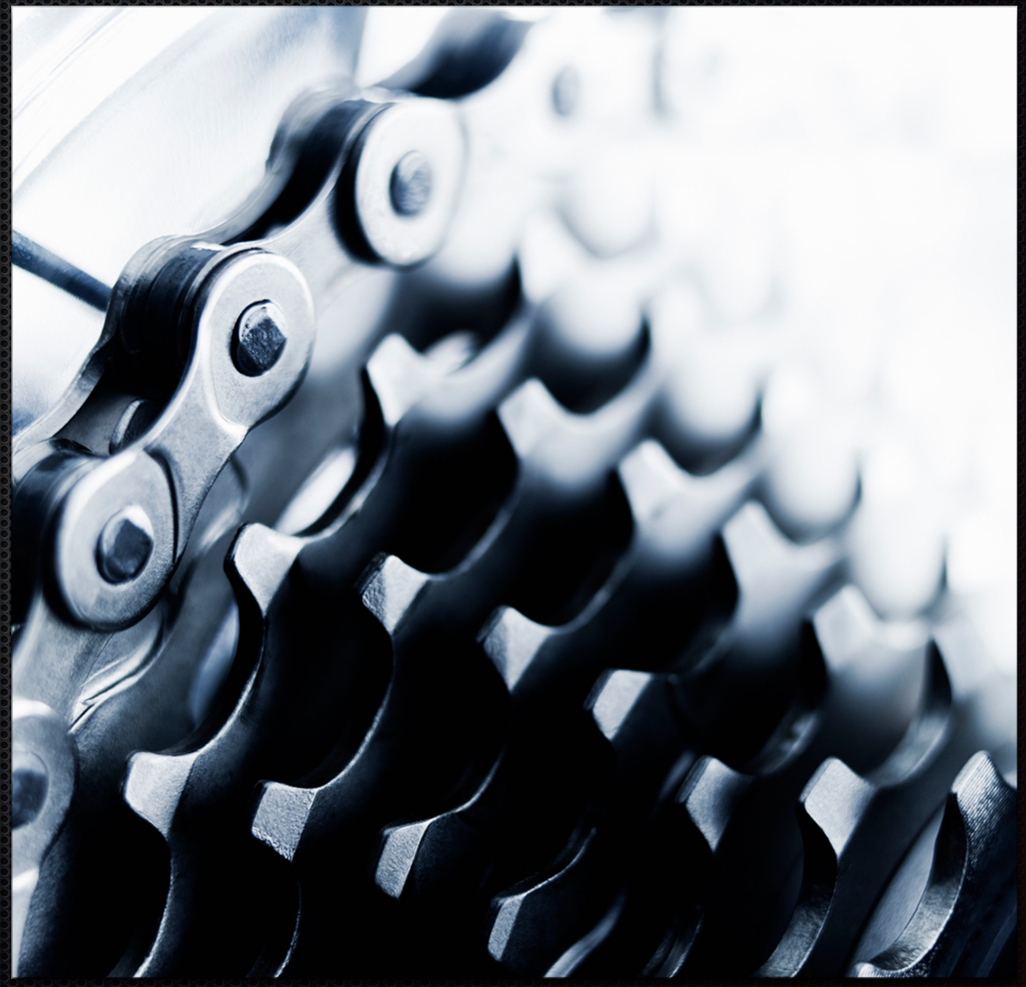
Predicting “How much funds does startup generally get in India?”

Group 2 - Hackathon (05-06 Jan’19)

Dhawal (on  Duty), Ramesh, Sonik

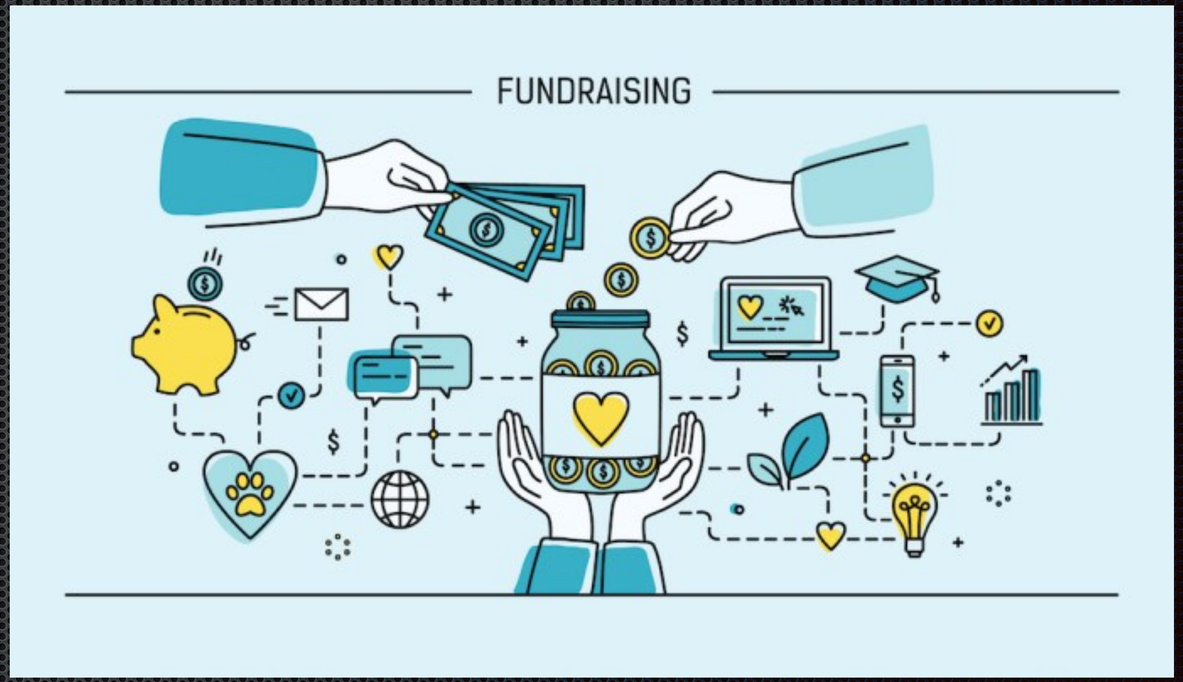
Agenda

- **Problem Statement**
- Data Source & Features
- EDA & Feature Engineering
  - Answering Few Questions
- Machine Learning
- Inference



Background & Problem Statement

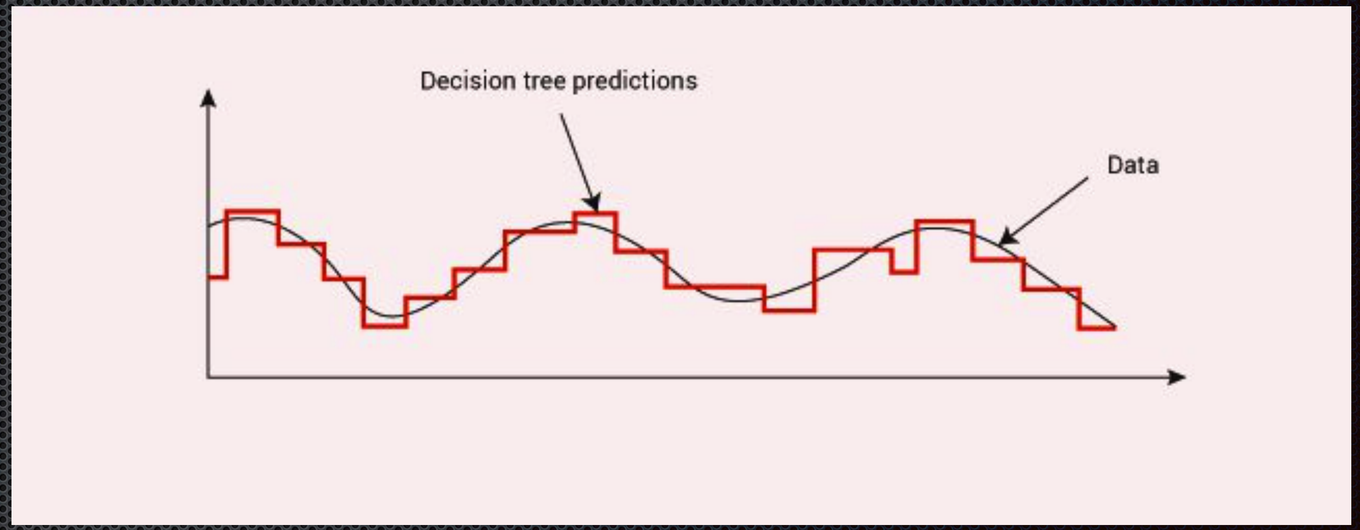
- Indian startup scene is rapidly increasing with 800+ startups in a year
- 5500 Startups currently in India
- Prime requirement of startups : Acquire capital to stabilise the business over the initial 2-3 years



Problem Statement : How much funds are allotted to Indian startups based on variables like timing, industry, location and investment type?

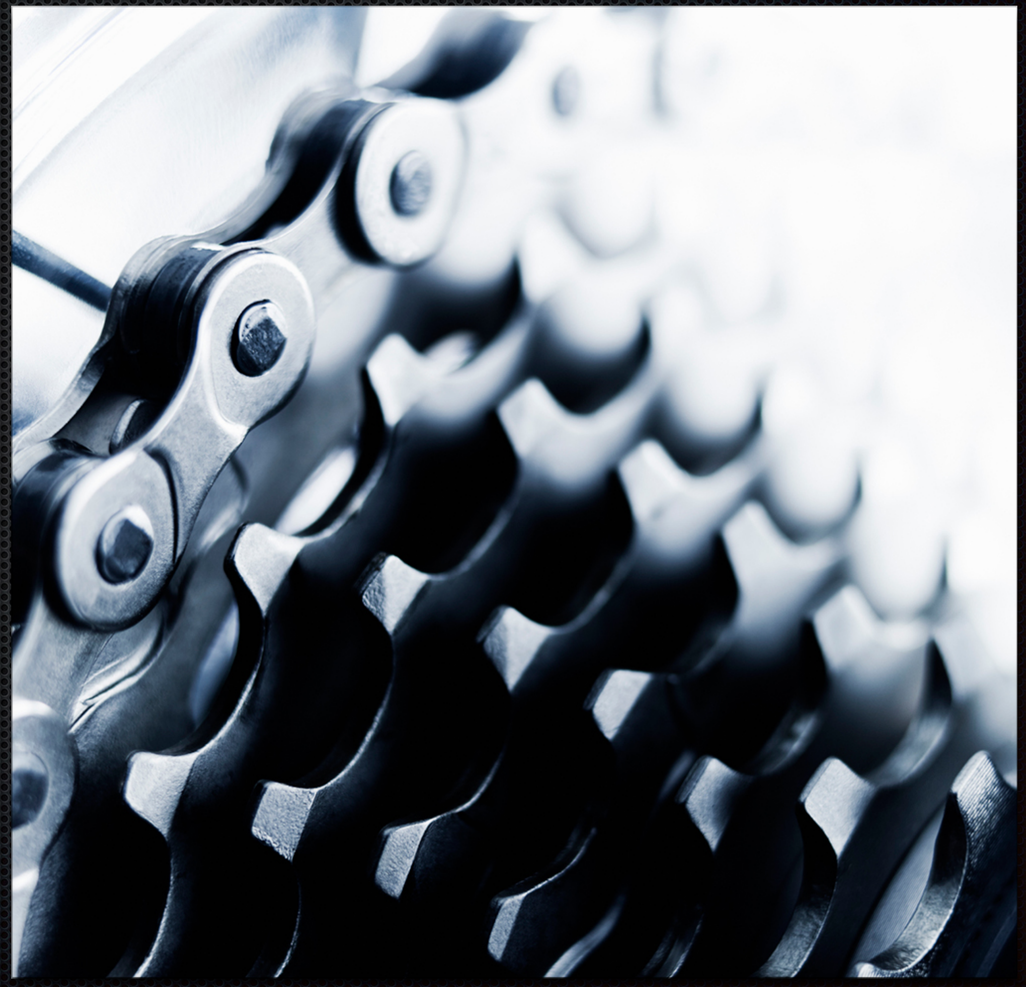
Problem Analysis

- Regression Problem
- Categorical Variables Inputs



Agenda

- Problem Statement
- **Data Source & Features**
- EDA & Feature Engineering
  - Answering Few Questions
- Machine Learning
- Inference



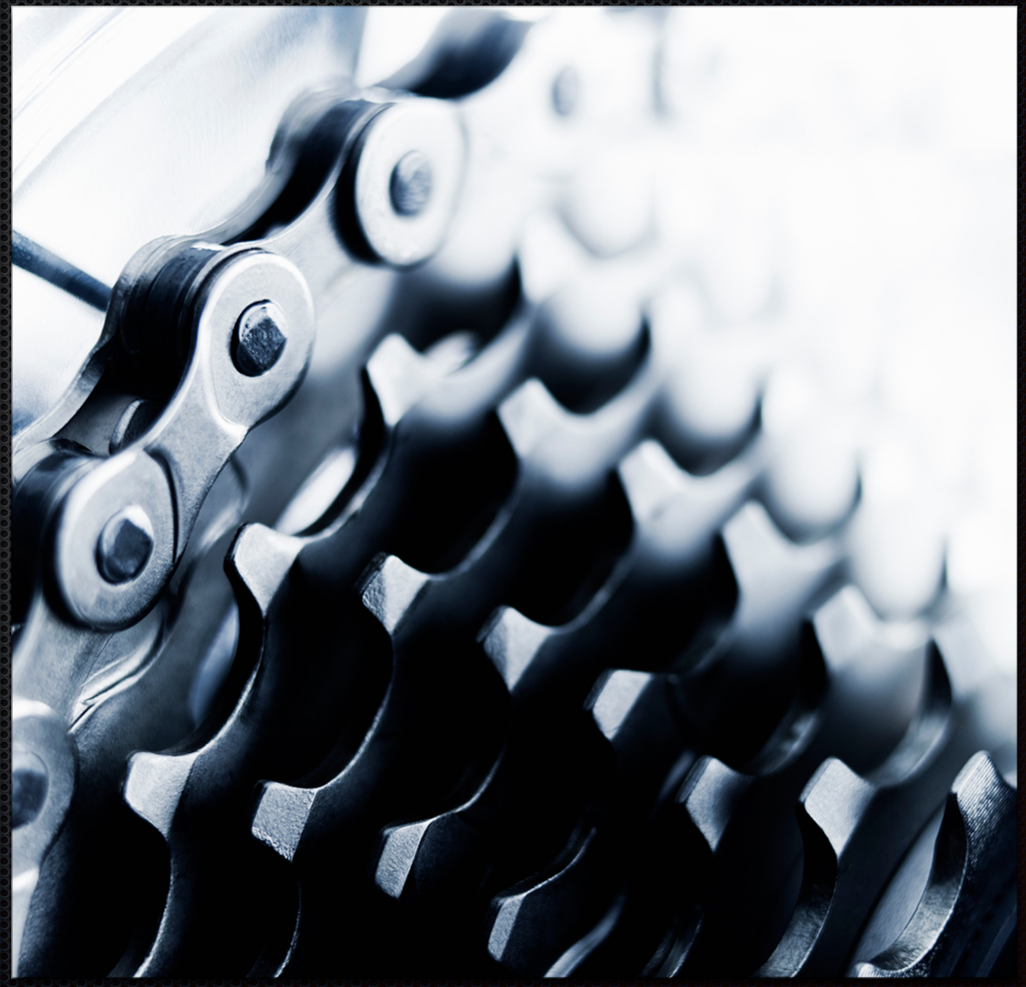
Data Source & Features

- **Important Features - All Categorical**
  - Date (Probably Time of the Year)
  - Industry
  - Location
  - Investor
  - Type of Investment
- **Target - Numerical**
  - Amount of Funding
- 2372 Data points
  - 846 (Missing $ Value of Funds)
  - 1536 (Value of Funds Present)
  - 1390 (With Industry Vertical Present)

```
RangeIndex: 2372 entries, 0 to 2371
Data columns (total 11 columns):
SNo                2372 non-null int64
Date               2372 non-null object
StartupName        2372 non-null object
IndustryVertical   2201 non-null object
SubVertical        1436 non-null object
CityLocation       2193 non-null object
InvestorsName      2364 non-null object
InvestmentType     2371 non-null object
AmountInUSD        1525 non-null object
Remarks            419 non-null object
```

Agenda

- Problem Statement
- Data Source & Features
- **Feature Engineering & EDA**
  - **Answering Few Questions**
- Machine Learning
- Inference



Feature Engineering & EDA (1/3)

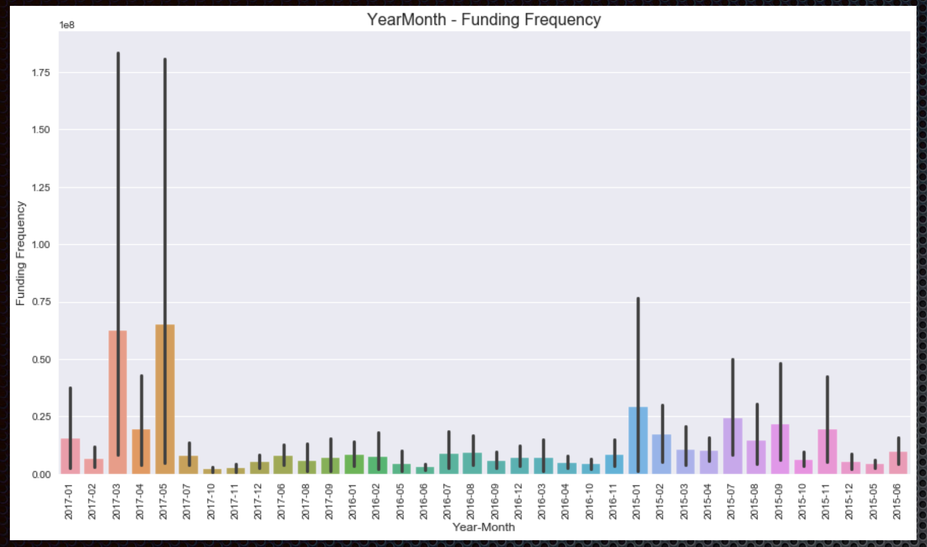
- % Missing Features
- Fixing Data
- Date Conversion to Month-Year
- Amount in USD: object to float
- Dropping empty rows, Not fit for imputation
- Dropping non-relevant features like serial number
- Drop Outliers (2%)

	Total	Percent
Remarks	1953	82.335582
SubVertical	936	39.460371
AmountInUSD	847	35.708263
CityLocation	179	7.546374
IndustryVertical	171	7.209106
InvestorsName	8	0.337268
InvestmentType	1	0.042159
StartupName	0	0.000000
Date	0	0.000000
SNo	0	0.000000

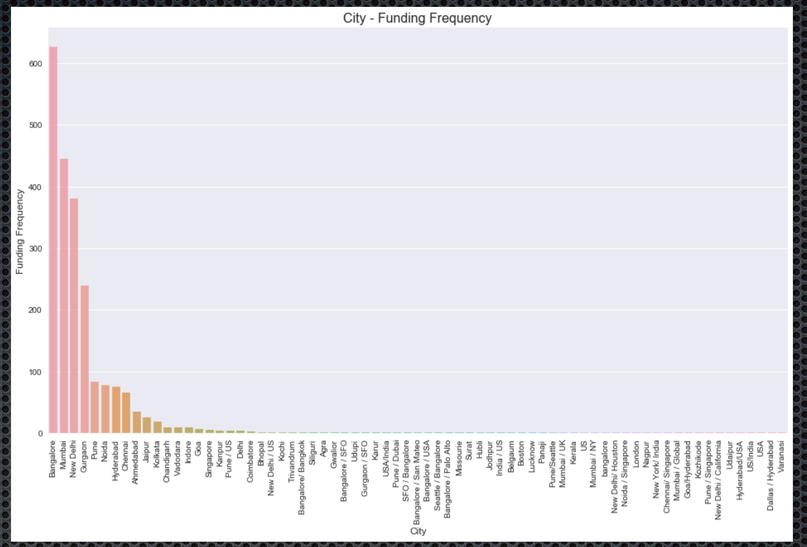
Fixing Data

In [60]:
```
fund['StartupName'] = fund['StartupName'].replace({"Flipkart.com":"Flipkart"})
fund['IndustryVertical']=fund['IndustryVertical'].replace({"ECommerce":"eCommerce"})
fund['IndustryVertical']=fund['IndustryVertical'].replace({"ecommerce":"eCommerce"})
fund['IndustryVertical']=fund['IndustryVertical'].replace({"Ecommerce":"eCommerce"})
fund['InvestmentType']=fund['InvestmentType'].replace({"Crowd funding":"Crowd Funding"})
fund['InvestmentType']=fund['InvestmentType'].replace({"SeedFunding":"Seed Funding"})
fund['InvestmentType']=fund['InvestmentType'].replace({"PrivateEquity":"Private Equity"})
```

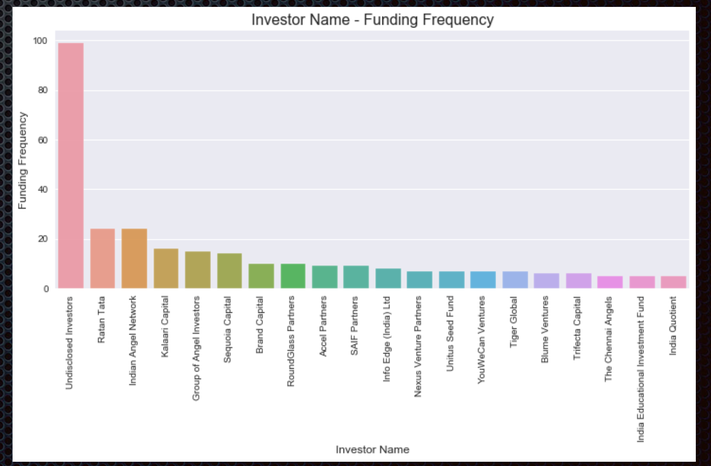
Feature Engineering & EDA (2/3)

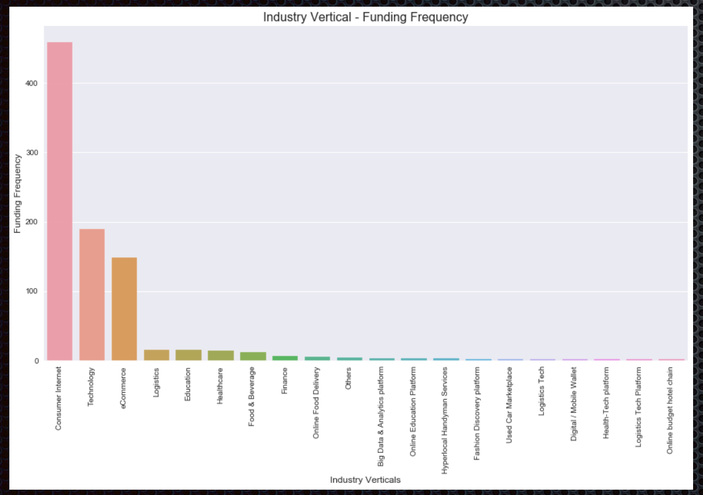


March '17 and May '17 - Flipkart and Paytm had received funding

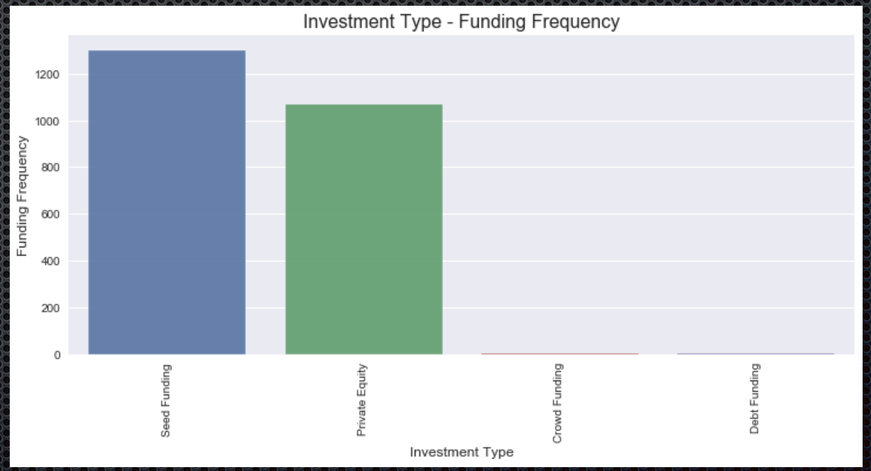


Bangalore	627 (T/E)	26.4%
Mumbai	446	19%
New Delhi	381	16%
Tier II/III	Varanasi	Indore
Karur	Nagpur	Belgaum
Siliguri	Kozikode	





Consumer Internet 20% > Technology 8% > eCommerce 6.3%

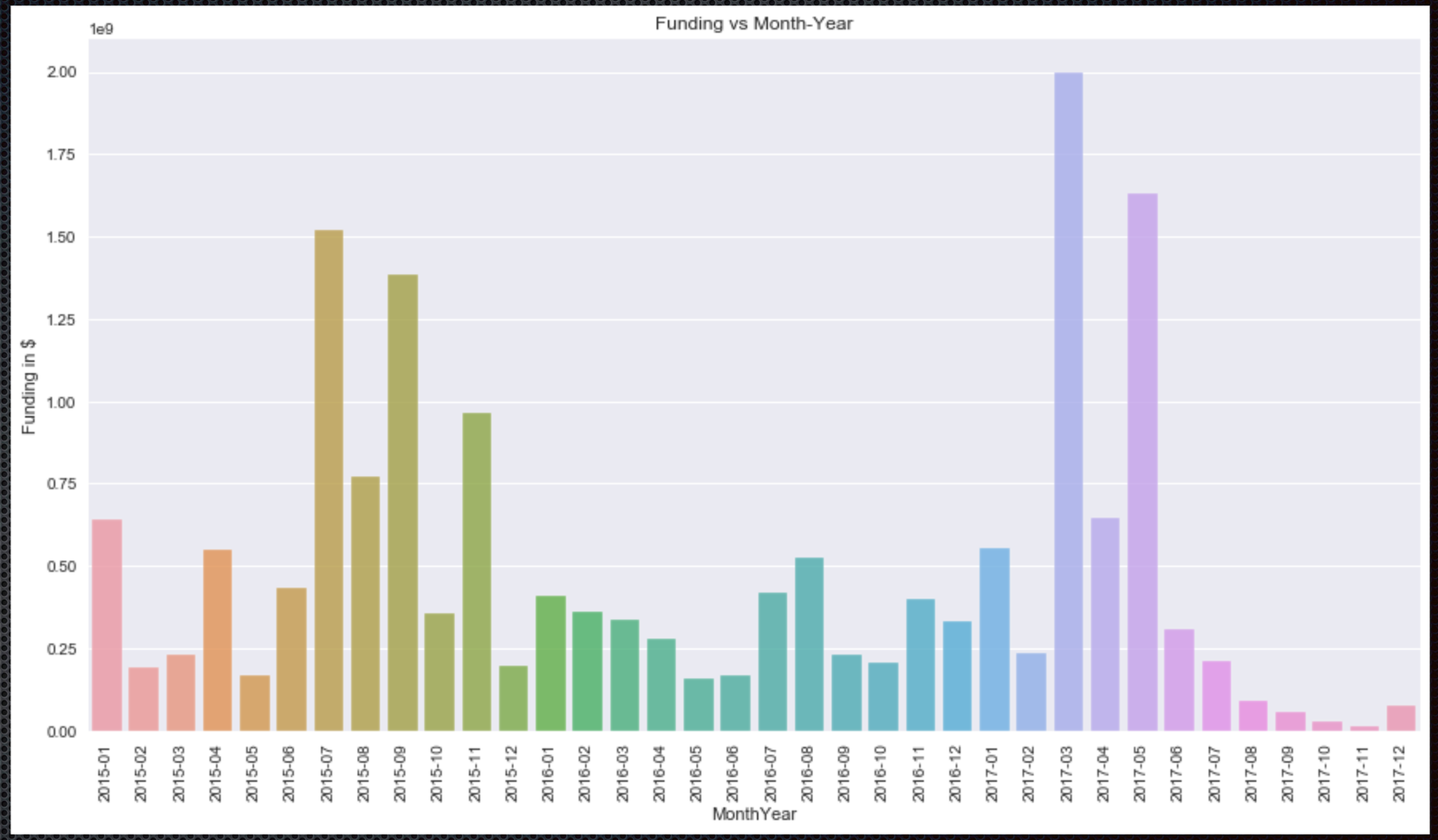


Seed Funding 54% PE 42%

Major Investors: Ratan Tata, Indian Angel Network, Kalaari Capital, Sequoia Capital

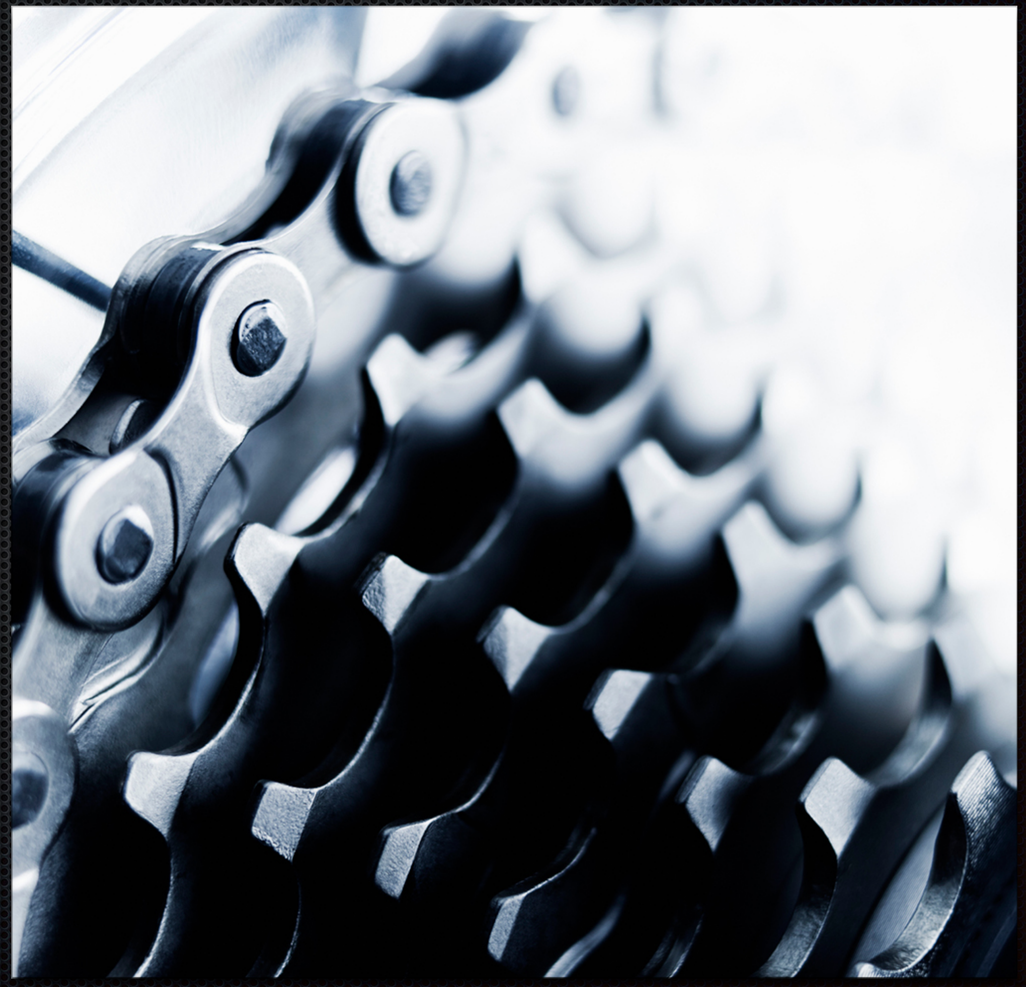
Feature Engineering & EDA (3/3)

Mean Investment	**$ 12.3 Mn**
Median Investment	**$ 1 Mn**
Maximum Single Investment	**$ 1.4 Bn**
Minimum Single Investment	**$ 18000**



Agenda

- Problem Statement
- Data Source & Features
- Feature Engineering & EDA
  - Answering Few Questions
- **Machine Learning**
- Inference



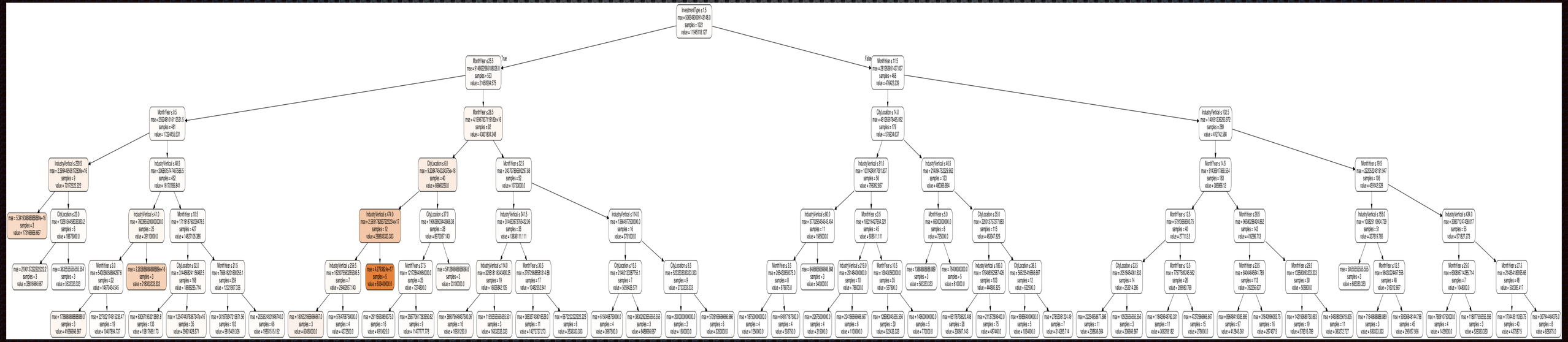
Machine Learning (1/3) - Linear Reg

Predictors	StartupName, InvestmentType, Final Remarks, IndustryVertical, InvestorsName, CityLocation, MonthYear	
Method	One Hot Encoding for categorical predictors	Dropped SubVertical (major missing values)
LR Accuracy Train/RMSE	100%	0
LR Accuracy Test/RMSE	**32%**	**$ 41,610,847**
Lasso	Depletes performance (24% Accuracy)	We go for Ridge
Ridge Accuracy Train/RMSE	89%	$ 24,397,428
Ridge Accuracy Test/ RMSE	**39%**	**$ 39,515,354**

Prevented overfitting with alpha = 0.8

Machine Learning (2/3) - Decision Tree

Predictors	IndustryVertical, InvestmentType, CityLocation, MonthYear		
Method	Label Encoding for Categorical Vars	GridSearchCV (criteria = ‘mse’, ‘mae’)	Depth 1-10, Features : Log, Sort Min wt of Leaf = 0.2%
R2 Score Train/RMSE	40%	**$ 55,297,893**	
R2 Score Test/RMSE	**19%**	**$ 49,421,051**	MSE, Depth=6, Wt_leaf=0.25%



Machine Learning (3/3) - Others

Random Forest	IndustryVertical, InvestmentType, CityLocation, MonthYear	Train Score = 49% Test Score = -8.2%
XGBoost	GridSearchCV, Label Encoding	Test Score = -8.5%

Agenda

- Problem Statement
- Data Source & Features
- Feature Engineering & EDA
  - Answering Few Questions
- Machine Learning
- **Inference**

Inference

- Machine Learning Models need more data to train (2015-2017)
- Amount of Funding is a Regression problem. Of 2372 Rows - 846 had missing amounts. **Workable Rows = 1390** (if we remove NaN - Zero Rows are left)
- **Linear Regression** with **Ridge** regularisation seems to best fit the limited data set
- More Cylical data required beyond 36 months
- Improve test accuracy (and the gap between training & CV score)