PROJECT REPORT

INTRODUCTION TO DATA MANAGEMEMT

(Project Semester August-December 2018)

STARTUP COMPANY ANALYSIS

Submitted by

Sowmith Kola

Registration No: -11603449

Section: -RKEM47

Course Code: -INT217

Under the Guidance of

Avinash Kaur

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Sowmith bearing Registration no. 11603449 has completed INT217 project titled, **“Data Analysis on various start-up companies and its financials”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science Engineering
Lovely Professional University
Phagwara, Punjab.

DECLARATION

I, Sowmith student of P132H: B.Tech. (Computer Science & Engineering) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 17-11-2018

Signature

Sowmith

11603449

ACKNOWLEDGEMENT

I have taken efforts in this course. However, it would not have been possible without the kind support and help of many individuals and University. I would like to extend my sincere thanks to all of them.

I am highly indebted to Ms. Avinash Kaur ma'am for her guidance and constant supervision as well as for providing necessary information for the course & also for her support in completing this course.

I would like to express my special gratitude and thanks to my faculty for giving me such attention and time.

My thanks and appreciations also go to my colleague and people who have willingly helped me out with their abilities.

TABLE OF CONTENTS

Introduction

A start-up or start-up is started by individual founders or entrepreneurs to search for a repeatable and scalable business model. More specifically, a start-up is a newly emerged business venture that aims to develop a viable business model to meet a marketplace need or problem. Founders design start-ups to effectively develop and validate a scalable business model. Hence, the concepts of start-ups and entrepreneurship are similar. However, entrepreneurship refers all new businesses, including self-employment and businesses that never intend to grow big or become registered, while start-ups refer to new businesses that intend to grow beyond the solo founder, have employees, and intend to grow large. Start-ups face high uncertainty and do have high rates of failure, but the minority that go on to be successful companies have to potential to become large and influential.

The Indian government has introduced over 50+ start-up schemes in past few years. Each start-up scheme is missioned towards boosting the Indian start-up ecosystem.

Consider this. Close to 4,400 technology start-ups exist in India and the number is expected to reach over 12,000 by 2020. India is also at third place behind US and Britain in terms of the number of start-ups. Furthermore, in line with its global counterparts, India has its own billion-dollar club to boast about. This includes start-ups like Flipkart, Snapdeal, Ola, InMobi, Hike, MuSigma, Paytm, Zomato, and Quikr. With the next $100 Mn funding raise, fintech startup MobiKwik too looks to join the unicorn club.

Inspiration

My dream of establishing a start-up   Inspired me to do this project.

Scope of analysis

My dataset contains a main dataset which consists of various details about start-ups like name, year of establishment, city, type of investment, investor name, company turnover etc.

I have gathered entire data from Kaggle and some references from official site of Start-up India. And the analysis is all about the following

1.Analysing the companies region using pivot tables.

2.Analysing the type of companies using slicers.

3.Analysing the year of establishments and representing them pictorially.

4.Showing the individual company details using LOOKUP Functions.

5.Protecting the Sensitive data.

EXISTING SYSTEM

The Existing system is given as follows:

System Name: STARTUP COMPANY ANALYSIS -Dashboard

Source of the System: https://www.kaggle.com/manasgarg/Startupdetails

DRAWBACKS OR LIMITATIONS OF EXISTING SYSTEM:

The existing data doesn’t provide the following techniques that are required for the analysis

- The optimal solutions and the way to control it is not represented.
- There are only limited number of data sets and charts in the system
- MATHEMATICAL CALCULATIONS: The existing system does not have logical conditioning and mathematical calculations such as average, maximum, minimum.
- SORTING: Sorting allows the apps to be sorted by category or cost.

SOURCE OF DATASET

The dataset is taken from Kaggle. Kaggle is a community of data scientists and data enthusiasts. This platform enables you to learn from and mentor each other on your personal, academic, and professional data science journeys.

Kaggle is an online community of data scientists and machine learners, owned by Google, Inc. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

The community spans 194 countries. It is the largest and most diverse data community in the world, ranging from those just starting out too many of the world's best-known researchers.

ETL PROCESS

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. The ETL process became a popular concept in the 1970s. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Since the data extraction takes time, it is common to execute the three phases in parallel. While the data is being extracted, another transformation process executes while processing the data already received and prepares it for loading while the data loading begins without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The separate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

Extract

The first part of an ETL process involves extracting the data from the source system(s). In many cases, this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. Most data-warehousing projects combine data from different source systems. Each separate system may also use a different data organization and/or format. Common data-source formats include relational databases, XML, JSON and flat files, but may also include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even formats fetched from outside sources by means such as web spidering or screen-scraping. The streaming of the extracted data source and loading on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing.

An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in each domain (such as a pattern/default or list of values). If the data fails, the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records.

Transform

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all; such data is known as "direct move" or "pass through" data.

An important function of transformation is the cleaning of data, which aims to pass only "proper" data to the target. The challenge when different systems interact is in the relevant systems' interfacing and communicating. Character sets that may be available in one system may not be so in others.

In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the server or data warehouse:

Selecting only certain columns to load: (or selecting null columns not to load). For example, if the source data has three columns (aka "attributes"), roll no, age, and salary, then the selection may take only roll no and salary. Or, the selection mechanism may ignore all those records where salary is not present (salary = null).

Translating coded values: (e.g., if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")

Encoding free-form values: (e.g., mapping "Male" to "M")

Deriving a new calculated value: (e.g., sale amount = qty * unit price)

Sorting or ordering the data based on a list of columns to improve search performance

Joining data from multiple sources (e.g., lookup, merge) and deduplicating the data

Aggregating (for example, rollup — summarizing multiple rows of data — total sales for each store, and for each region, etc.)

Generating surrogate-key values

Transposing or pivoting (turning multiple columns into multiple rows or vice versa)

Splitting a column into multiple columns (e.g., converting a comma-separated list, specified as a string in one column, into individual values in different columns)

Disaggregating repeating columns

Looking up and validating the relevant data from tables or referential files

Applying any form of data validation; failed validation may result in a full rejection of the data, partial rejection, or no rejection at all, and thus none, some, or all of the data is handed over to the next step depending on the rule design and exception handling; many of the above transformations may result in exceptions, e.g., when a code translation parses an unknown code in the extracted data

LOAD

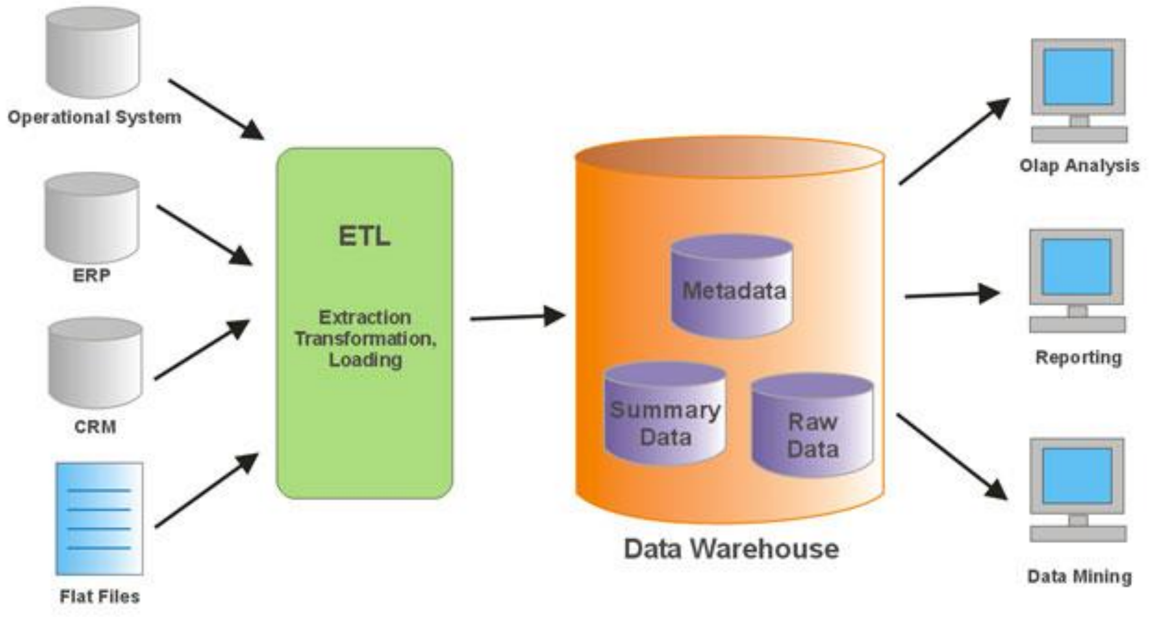
The load phase loads the data into the end target, which may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one-year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

For example, a financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all these data elements and consolidate them into a uniform presentation, such as for storing in a database or data warehouse.

Another way that companies use ETL is to move information to another application permanently. For instance, the new application might use another database vendor and most likely a very different database schema. ETL can be used to transform the data into a format suitable for the new application to use.

An example would be an Expense and Cost Recovery System (ECRS) such as used by accountancies, consultancies, and legal firms. The data usually ends up in the time and billing system, although some businesses may also utilize the raw data for employee productivity reports to Human Resources (personnel dept.) or equipment usage reports to Facilities Management.



In our scenario, dataset is Mobileappstore.csv, so during the ETL process the data is extracted from this dataset, transformed to eliminate irrelevant data mentioned in the scope of analysis section and loaded into the excel where the required data resides. From this analysis reporting can be done.

ANALYSIS ON EACH DATA SET:

Analysis 1:

Introduction

The first objective is all about Analysing the company's region using pivot tables.

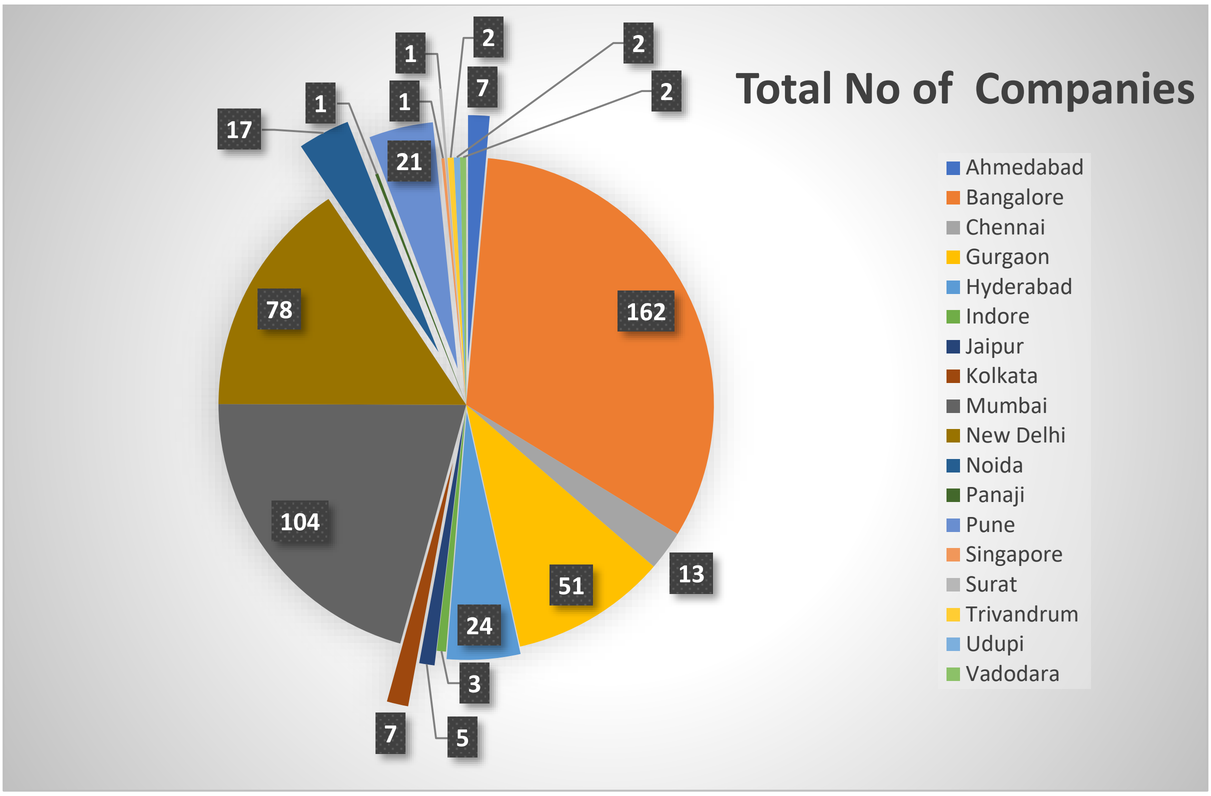
Specifications

Datasets contains city names, and no of companies in that city.

Data analysis:

City	No of Start-ups
Ahmedabad	7
Bangalore	162
Chennai	13
Gurgaon	51
Hyderabad	24
Indore	3

Visualisation



Analysis 2:

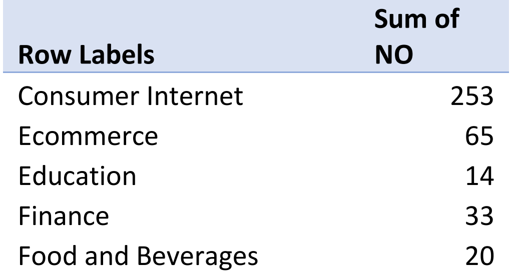
Introduction

The second objective is all about Analysing the type of companies using slicers.

Specifications

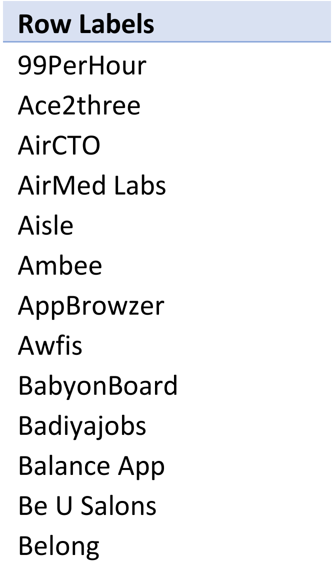
This dataset contains type of company in which genre they have established their company and no of companies present in that genre.

Data analysis:

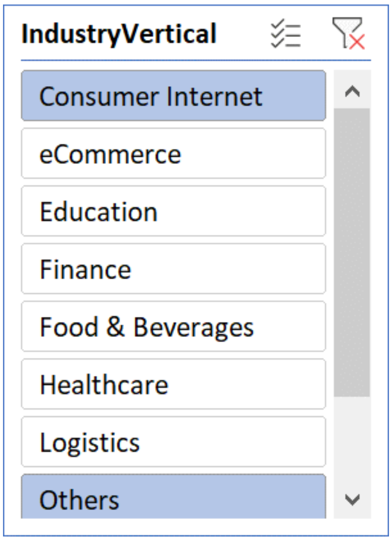


Row Labels	Sum of NO
Consumer Internet	253
Ecommerce	65
Education	14
Finance	33
Food and Beverages	20

Visualisation



Row Labels
99PerHour
Ace2three
AirCTO
AirMed Labs
Aisle
Ambee
AppBrowzer
Awfis
BabyonBoard
Badiyajobs
Balance App
Be U Salons
Belong



IndustryVertical

Consumer Internet

eCommerce

Education

Finance

Food & Beverages

Healthcare

Logistics

Others

Analysis 3:

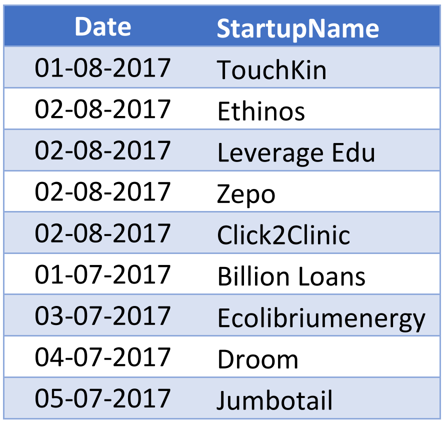
Introduction

The third objective is all about Analysing the year of establishments and representing them pictorially.

Specifications

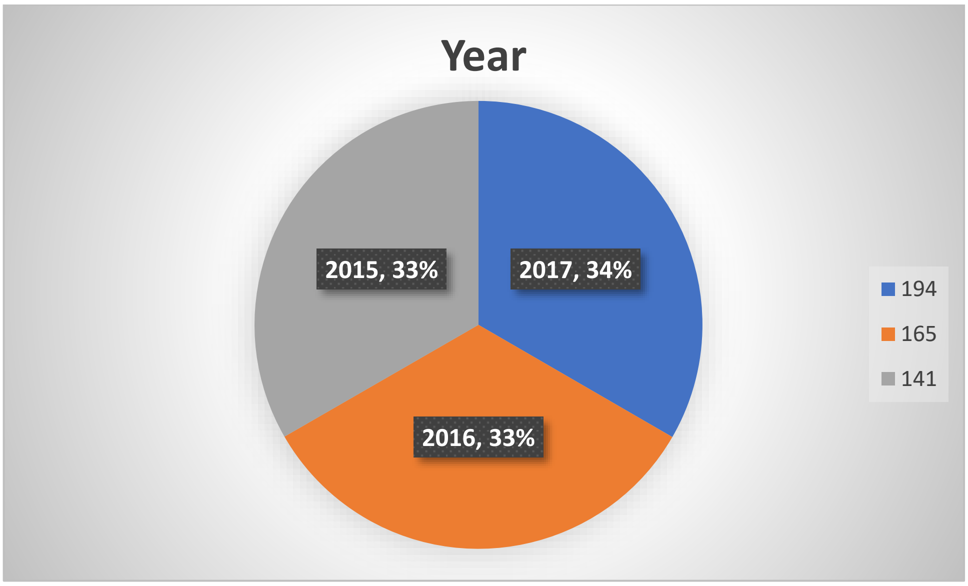
The data set contains years and the number of companies established.

Data analysis



Date	StartupName
01-08-2017	TouchKin
02-08-2017	Ethinos
02-08-2017	Leverage Edu
02-08-2017	Zepo
02-08-2017	Click2Clinic
01-07-2017	Billion Loans
03-07-2017	Ecolibriumenergy
04-07-2017	Droom
05-07-2017	Jumbotail

Visualisation



Analysis 4:

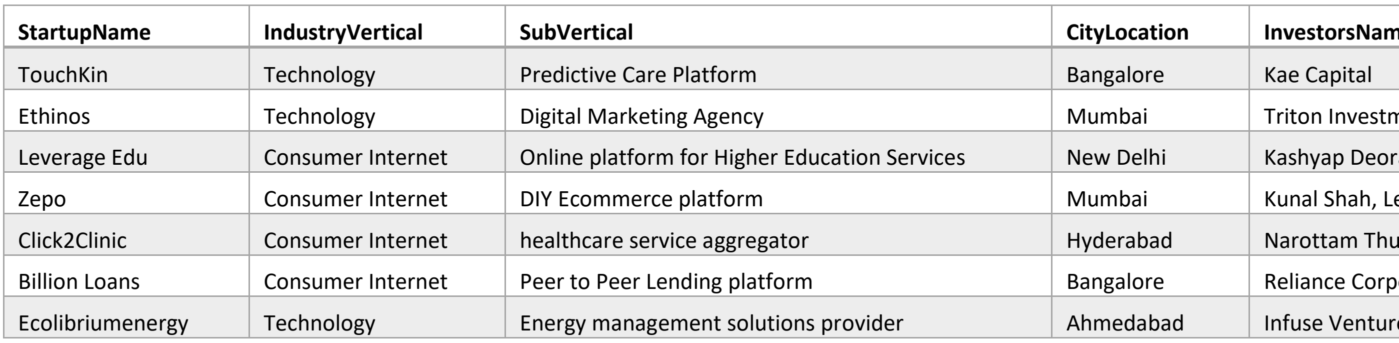
Introduction

The fourth objective is all about Showing the individual company details using LOOKUP Functions..

Specifications

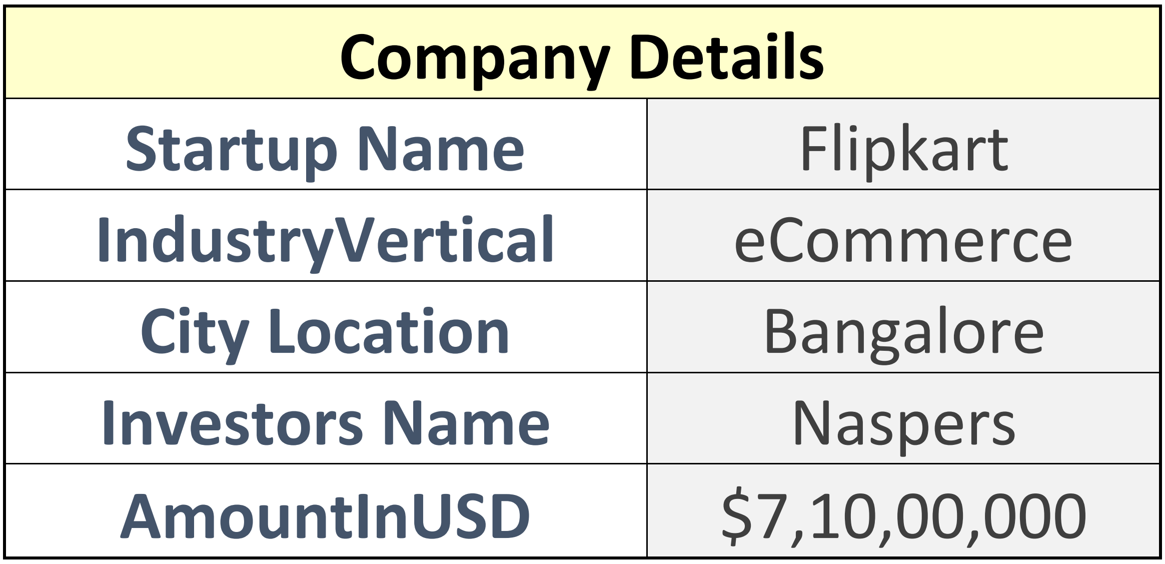
The data set is entering a company name.

Data analysis:



StartupName	IndustryVertical	SubVertical	CityLocation	InvestorsNam
TouchKin	Technology	Predictive Care Platform	Bangalore	Kae Capital
Ethinos	Technology	Digital Marketing Agency	Mumbai	Triton Investm
Leverage Edu	Consumer Internet	Online platform for Higher Education Services	New Delhi	Kashyap Deor
Zepo	Consumer Internet	DIY Ecommerce platform	Mumbai	Kunal Shah, Le
Click2Clinic	Consumer Internet	healthcare service aggregator	Hyderabad	Narottam Thu
Billion Loans	Consumer Internet	Peer to Peer Lending platform	Bangalore	Reliance Corp
Ecolibriumenergy	Technology	Energy management solutions provider	Ahmedabad	Infuse Ventur

Visualisation:



Company Details	
Startup Name	Flipkart
IndustryVertical	eCommerce
City Location	Bangalore
Investors Name	Naspers
AmountInUSD	$7,10,00,000

Analysis 5:

Introduction

The fifth objective is all about protecting the Sensitive data.

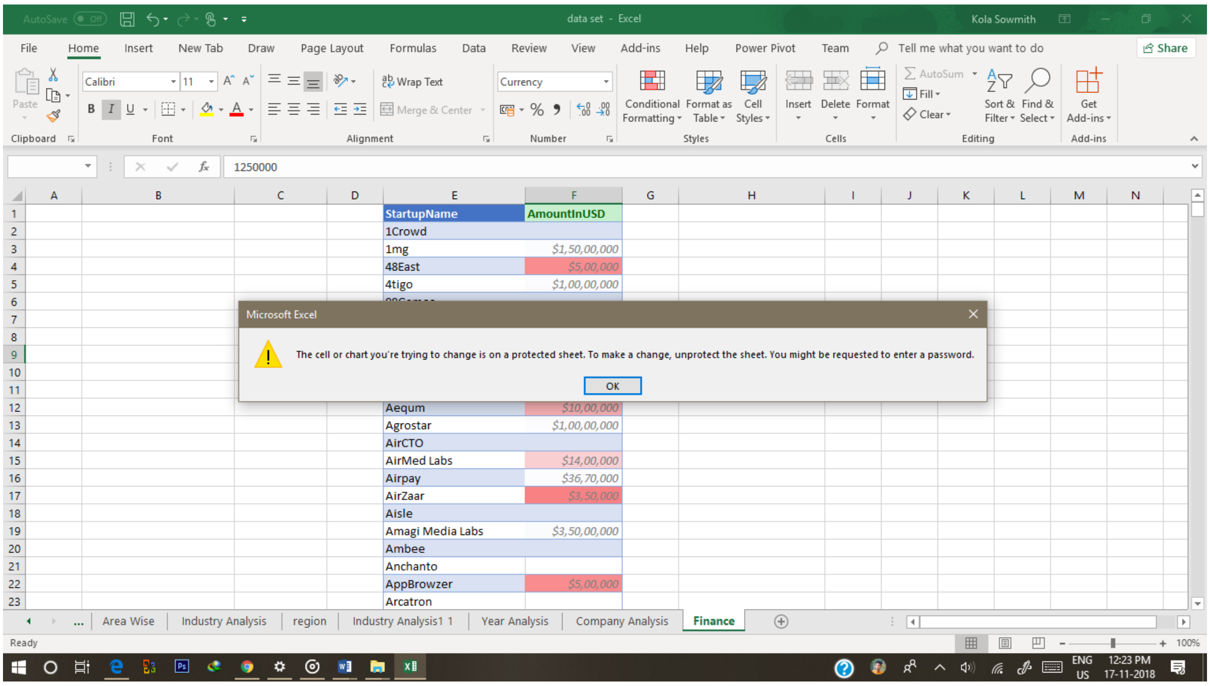
Specifications

The dataset is all about turnover of the companies since the finance details of a company is sensitive, we will be protecting the sheet with a password.

Data analysis

StartupName	AmountInUSD
1Crowd	
1mg	$1,50,00,000
48East	$5,00,000
4tigo	$1,00,00,000
99Games	
99PerHour	$3,00,000
ABI Health	$15,00,000
Absentia VR	$12,50,000
Ace Turtle	$50,00,000
Ace2three	$7,37,00,000
Aequm	$10,00,000

Visualisation



1250000

	E	F
1	StartupName	AmountInUSD
2	1Crowd	
3	1mg	$1,50,00,000
4	48East	$5,00,000
5	4tigo	$1,00,00,000
12	Aequm	$10,00,000
13	Agrostar	$1,00,00,000
14	AirCTO	
15	AirMed Labs	$14,00,000
16	Airpay	$36,70,000
17	AirZaar	$3,50,000
18	Aisle	
19	Amagi Media Labs	$3,50,00,000
20	Ambee	
21	Anchanto	
22	AppBrowzer	$5,00,000
23	Arcatron	

Microsoft Excel

The cell or chart you're trying to change is on a protected sheet. To make a change, unprotect the sheet. You might be requested to enter a password.

OK

LIST OF ANALYSIS WITH RESULTS

- Number of Start-up region.
- Companies in different genres.
- Number of companies established in a year.
- Searching for Company Details.
- Data protection

Future Scope

As the world is evolving very fast and world of youngsters is evolving to and STARTUP is the best example for that.it has changed the era of intelligence completely with their desperate ideas and business to a greater extent.

The analysis on the statistics of start-up in India leads to the better knowledge of different ideas to develop and for filtering the idea pool.

REFERENCES

- The general information is gathered from the offline survey in my university.
- The data sets and analysis have been gathered from the online sites.
- The introduction parts and the general description about the data sets have been collected from the Wikipedia.

BIBILOGRAPHY

https://www.kaggle.com/manasgarg/Startupdetails

https://inc42.com/startup-101/startup-scheme-indian-government-startups/