SOEN 6611 - SOFTWARE MEASUREMENT: THEORY AND PRACTICE

Project Report on Task 5

WINTER 2022

Course Instructor: Dr. Olga Ormandjieva

TEAM 5	
Student #	**Name**
40163582	Mohammod Suhel Firdus
40157109	Vivekananda Reddy Gottam
40184906	Saswati Chowdhury
40156971	Milesh Kotadia

Soen6611 Step 5 Team Report: suggested outline

Cover page: indicate the Step and list the names of your teammates.
Main content:

1 Describe the data set of your team

2 Describe the base measures data's collection procedure, trace it to your Step 4 plan for data collection.

3 Attach the collected data values (excel file, etc.)

4 For each of the 3 V's indicators:

  4.1 Attach the values of the corresponding derived measure(s), where applicable (excel file, etc.)

  4.2 Draw the graph of the indicator generated from the values of the derived measure(s)

5 Write your conclusions: would your data be used for a machine learning algorithm? Justify your answer.

1. Data Set Description:

1.1 Dataset name: Most Popular Superhero TV Shows

1.2 Source:
https://www.kaggle.com/datasets/anoopkumarraut/most-popular-superhero-tv-shows

1.3 Context:

Superheroes and superhero Tv-Shows are favorites among all of us. Essentially, the shows are a rehash of a superhero franchise with everything that it entails. A story is told, emotions are felt, and there's action and heroism displayed. Surely, most of us enjoyed them. Our goal is to delve into the data and see what the numbers say about these shows.

1.4 Content: Data Dictionary

Features	Define
show_title	name of the show
imdb_rating	ratings on IMDB.com
release_year	year shows were aired till the year it ended
runtime	show every episode runtime in minutes
genre	genre/category of the show
parental_guideline	parental advisory for the show
imdb_votes	number of votes on IMDb
synopsis	brief of the show

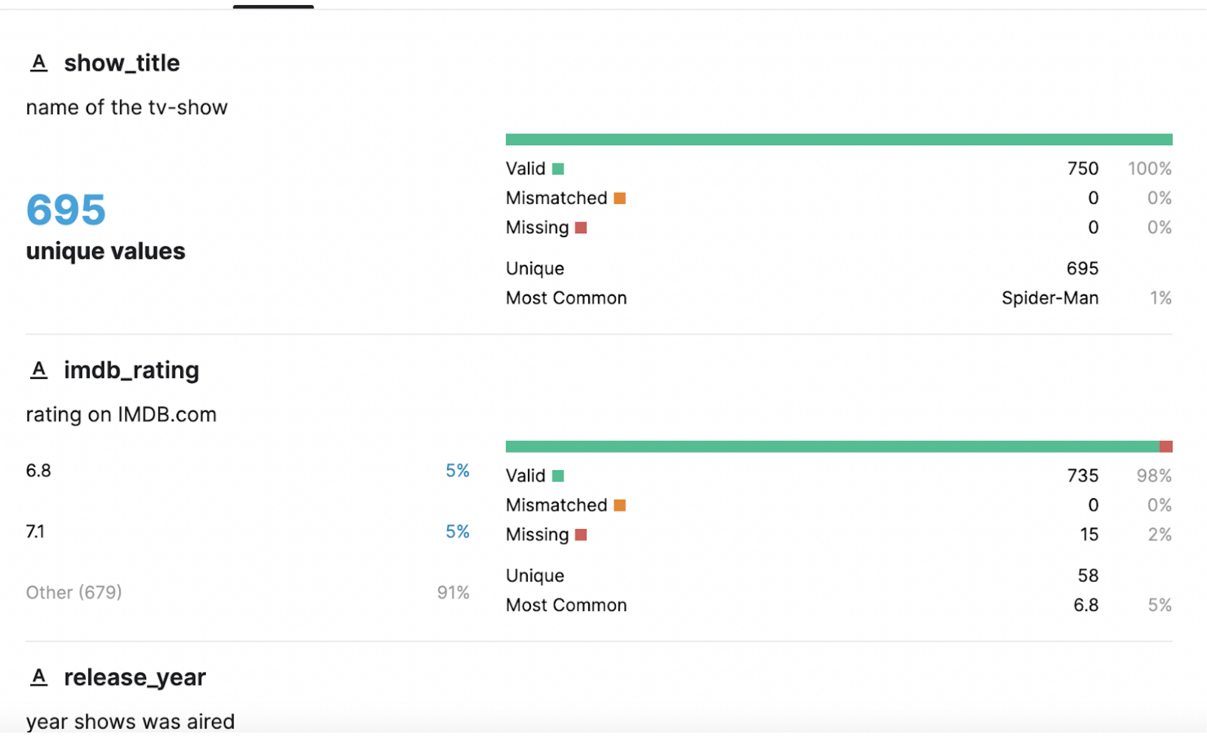
Table 1: Data dictionary of the dataset

1.5 Acknowledgements: IMDB.com, an open-source platform for movies and TV shows ratings, provided the data for this article.

1.6 Size of Dataset: 164.41 kB

1.7 Structure of data: 8 columns and 750 rows (text_format - String)

1.8 No of records: 750, No of Unique records: 740



show_title

name of the tv-show

695
unique values

Valid 🟩	750	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	695	
Most Common	Spider-Man	1%

imdb_rating

rating on IMDB.com

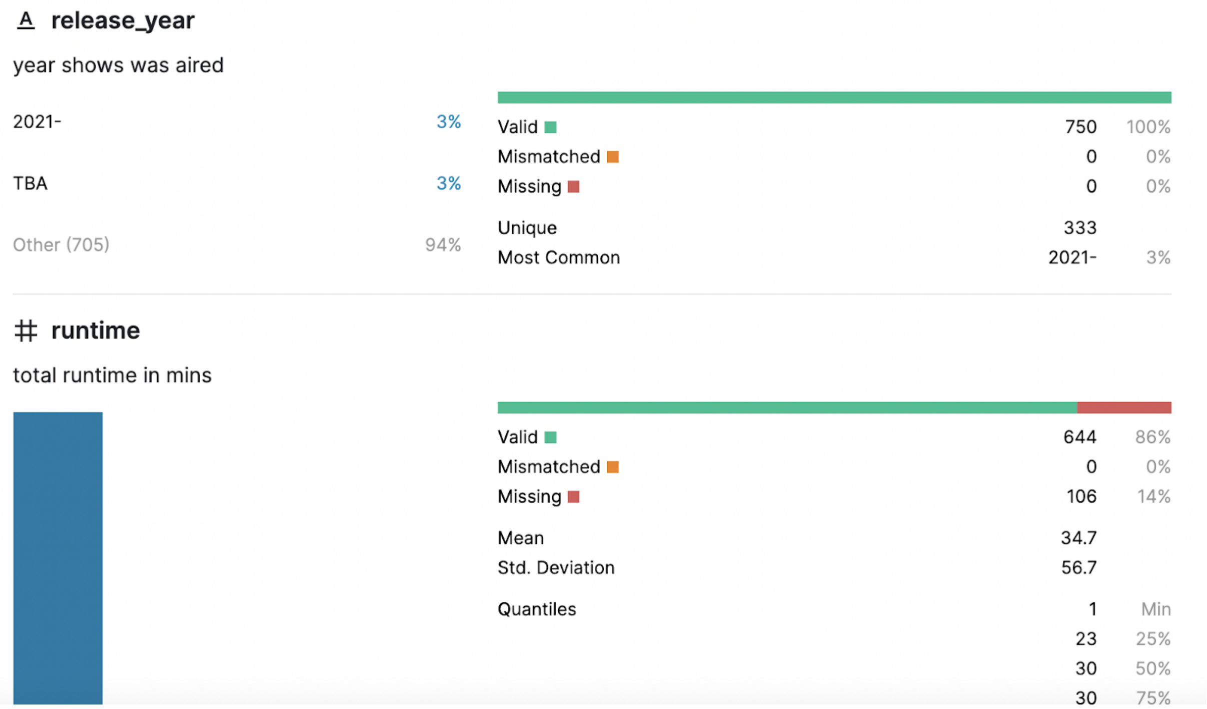
6.8	5%
7.1	5%
Other (679)	91%

Valid 🟩	735	98%
Mismatched 🟧	0	0%
Missing 🟥	15	2%
Unique	58	
Most Common	6.8	5%

release_year

year shows was aired

Figure 1.a: Synopsis of the dataset : Source: screenshot from Kaggle. Source: Kaggle



release_year

year shows was aired

2021-	3%
TBA	3%
Other (705)	94%

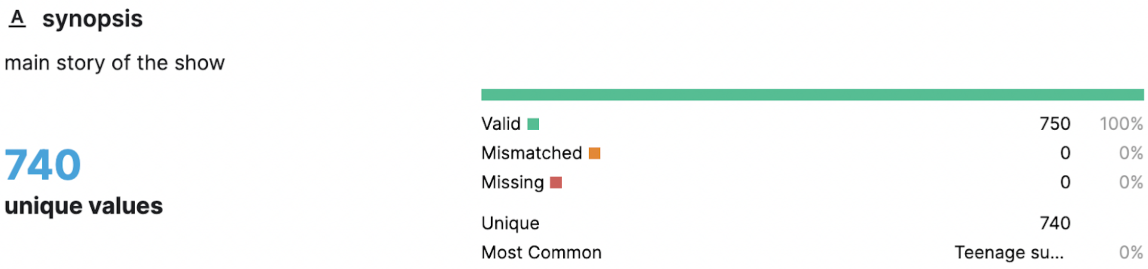
Valid 🟩	750	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	333	
Most Common	2021-	3%

runtime

total runtime in mins

Valid 🟩	644	86%
Mismatched 🟧	0	0%
Missing 🟥	106	14%
Mean	34.7	
Std. Deviation	56.7	
Quantiles	1	Min
	23	25%
	30	50%
	30	75%

Figure 1.b: Synopsis of the dataset : Source: screenshot from Kaggle. Source: Kaggle



synopsis

main story of the show

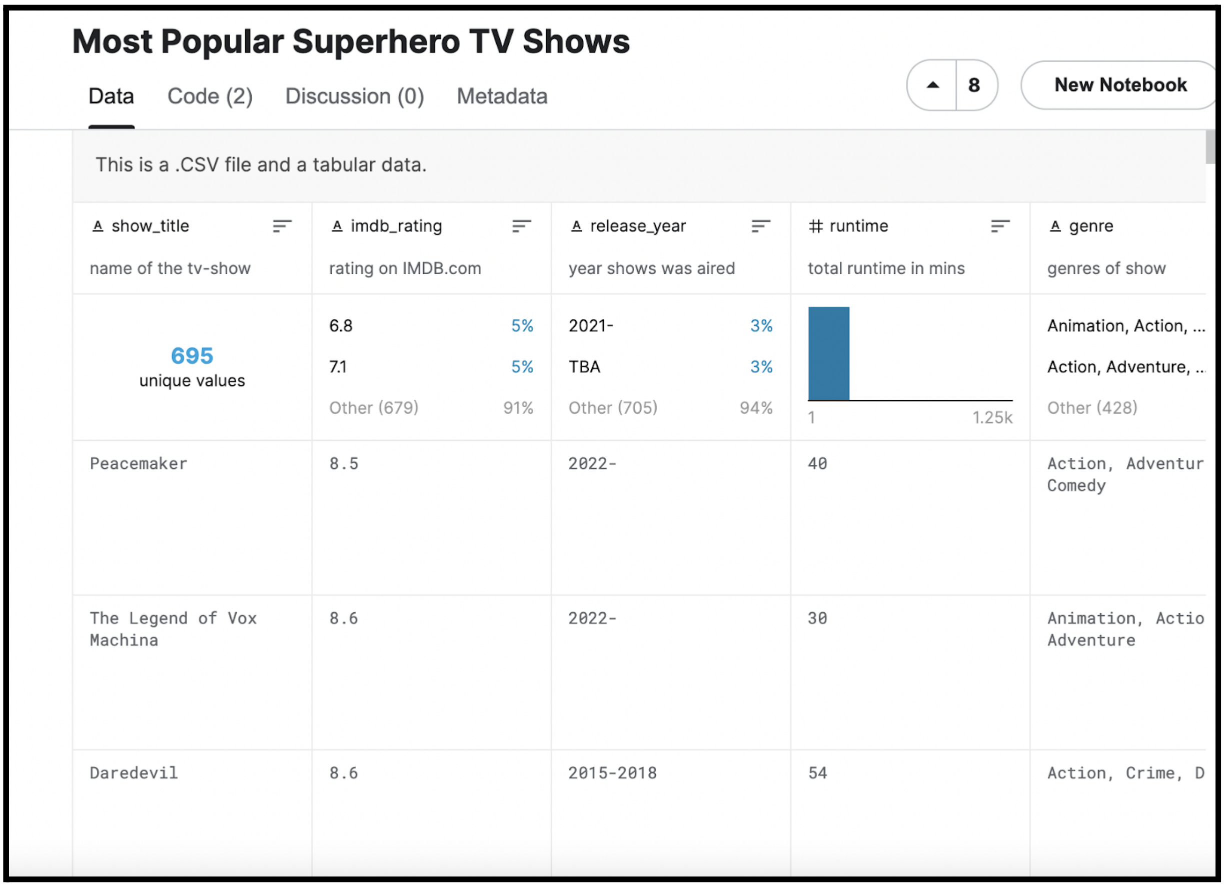
740
unique values

Valid 🟩	750	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	740	
Most Common	Teenage su...	0%

Figure 1.c: Synopsis of the dataset : Source: screenshot from Kaggle. Source: Kaggle

1.9 Details:

A sample details screenshot of our CSV file is given below:



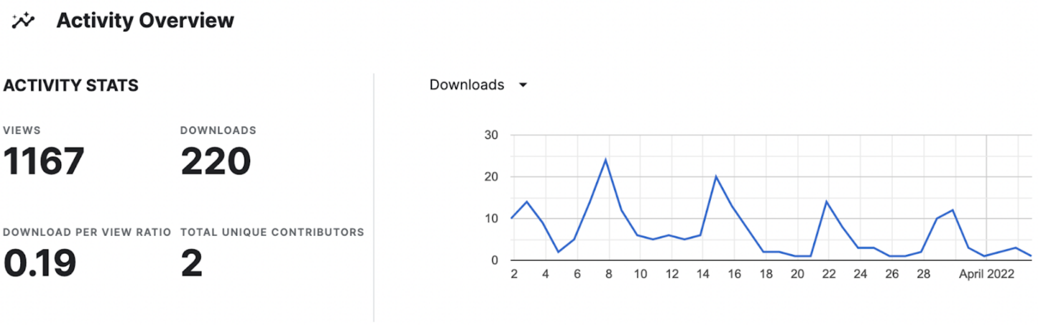
Most Popular Superhero TV Shows

This is a .CSV file and a tabular data.

show_title	imdb_rating	release_year	runtime	genre
name of the tv-show	rating on IMDB.com	year shows was aired	total runtime in mins	genres of show
695 unique values	6.8 5% 7.1 5% Other (679) 91%	2021- 3% TBA 3% Other (705) 94%	1 1.25k	Animation, Action, ... Action, Adventure, .. Other (428)
Peacemaker	8.5	2022-	40	Action, Adventur Comedy
The Legend of Vox Machina	8.6	2022-	30	Animation, Actio Adventure
Daredevil	8.6	2015-2018	54	Action, Crime, D

Figure 2: Summary of the dataset : Source: screenshot from Kaggle. Source: Kaggle

1.10 Activity Overview:


Activity Overview

ACTIVITY STATS

VIEWS 1167

DOWNLOADS 220

DOWNLOAD PER VIEW RATIO 0.19

TOTAL UNIQUE CONTRIBUTORS 2

Figure 3: Activity on the dataset : Source: screenshot from Kaggle. Source: Kaggle

1.11 Data Extraction:

Google Collab:

https://colab.research.google.com/drive/1NdWP_SZJlGmiGuCO54YRnihZCG0JAIQt?usp=sharing

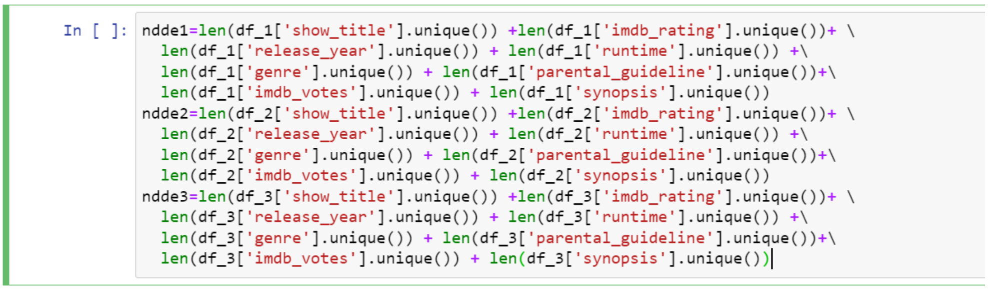
We had to divide the dataset into 3 subsets to assume we collected data on **three distinct times T1,T2, T3**

2. Base measures data's collection procedure

2.1 Collection procedure to measure Ndde:

The team of data scientists first measure the length of the dataset and then calculate the number of distinct data elements (Ndde) present in the dataset in about .5 days (4 person hours) at the beginning of every time frame T1, T2, T3 by using Google collab. It will be used to evaluate the value of big data volume.

To calculate the Ndde, we wrote a python code and used in-built methods to count the unique elements present in the dataframes.



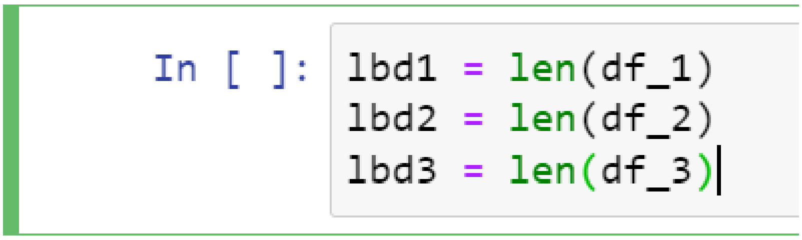
```
In [ ]: ndde1=len(df_1['show_title'].unique()) +len(df_1['imdb_rating'].unique())+ \
          len(df_1['release_year'].unique()) + len(df_1['runtime'].unique()) +\
          len(df_1['genre'].unique()) + len(df_1['parental_guideline'].unique())+\
          len(df_1['imdb_votes'].unique()) + len(df_1['synopsis'].unique())
        ndde2=len(df_2['show_title'].unique()) +len(df_2['imdb_rating'].unique())+ \
          len(df_2['release_year'].unique()) + len(df_2['runtime'].unique()) +\
          len(df_2['genre'].unique()) + len(df_2['parental_guideline'].unique())+\
          len(df_2['imdb_votes'].unique()) + len(df_2['synopsis'].unique())
        ndde3=len(df_3['show_title'].unique()) +len(df_3['imdb_rating'].unique())+ \
          len(df_3['release_year'].unique()) + len(df_3['runtime'].unique()) +\
          len(df_3['genre'].unique()) + len(df_3['parental_guideline'].unique())+\
          len(df_3['imdb_votes'].unique()) + len(df_3['synopsis'].unique())
```

Figure 4: Measuring Ndde using python

2.2 Collection procedure to measure Lbd:

Having calculated the Ndde, the development team goes on to measure the length of the dataset, which consists of the total number of records. The variance present in the dataset will be analyzed and compared in different time frames in this way. It took the same amount of time as Ndde, i.e. .5 days or about 4 person-hours by using Google collab.

For calculating Lbd, we used the in-built function for counting the number of rows in the dataset.



```
In [ ]: lbd1 = len(df_1)
        lbd2 = len(df_2)
        lbd3 = len(df_3)
```

Figure 5: Measuring Lbd using python

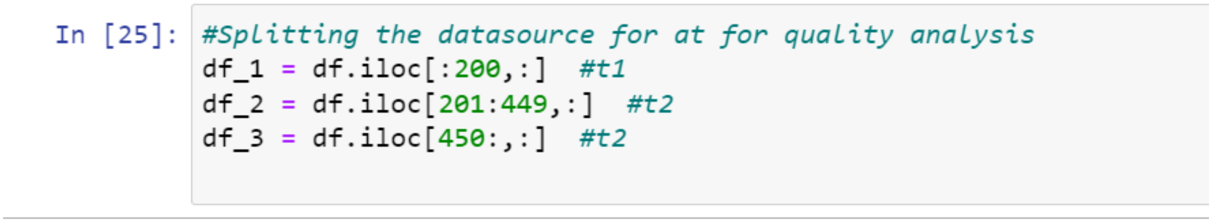
2.3 Collection procedure to measure Nds:

Currently, the team manually counts all datasets at the beginning of each time frame T1, T2, and T3 over the course of four person-hours (2.25 days). This will display the count of the various datasets that exist to evaluate the variety of big data.

The value of **Nds is calculated manually** which states the number of datasets that are present in different time frames.

2.4 Collection procedure to measure T (Time):

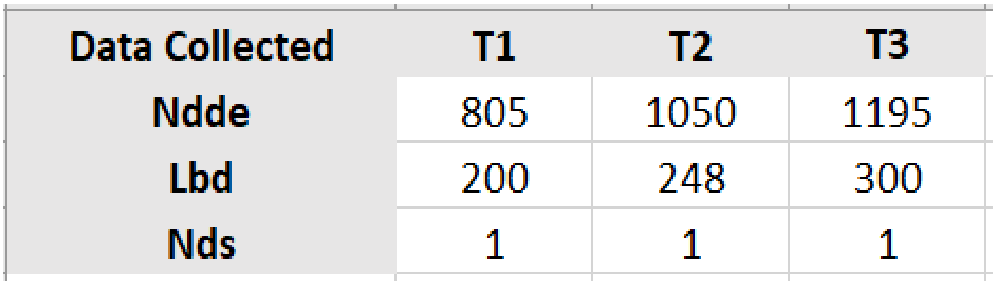
In order to analyze big data velocity, the developers track the time period (T). Specifically, we use it to measure how fast big data is growing as well as how fast it is processed. The dataset is split for three different timeframes - T1, T2, T3.



```
In [25]: #Splitting the datasource for at for quality analysis
         df_1 = df.iloc[:200,:]  #t1
         df_2 = df.iloc[201:449,:]  #t2
         df_3 = df.iloc[450:,:]  #t2
```

Figure 6: Splitting the dataset at T1, T2, T3

3. Attach the collected data values:



Data Collected	T1	T2	T3
Ndde	805	1050	1195
Lbd	200	248	300
Nds	1	1	1

Figure 7: Collected values of base measures

4. 3 V's indicators:

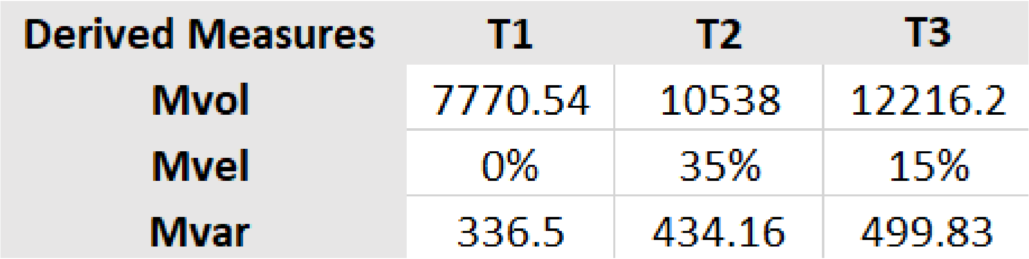
4.1 Attach the values of the corresponding derived measures

The values measured in *section 2* of all base measures were used to calculate the three derived measures by using the following formulas:

Derived Measures	Formula	Base Measures Used
Mvol	$Mvol\ (MDS) = Ndde\ (MDS) * \log_2 ((Ndde\ (NDS)))$	***Ndde*** *from section* **2.1** ***Nds*** *from section* **2.3**
Mvel	$Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100$	***T*** *(Time) from section* **2.4**
Mvar	$Mvar\ (MDS) = Ndde\ (DE) * W_{Ndde} + Lbd\ (MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}$	***Ndde*** *from section* **2.1** ***Lbd*** *from section* **2.2** ***Nds*** *from section* **2.3**

Table 2: Formulas for calculating derived measures

The values of derived measures in three different time frames:



Derived Measures	T1	T2	T3
Mvol	7770.54	10538	12216.2
Mvel	0%	35%	15%
Mvar	336.5	434.16	499.83

Figure 8: Calculated derived measures

4.2 Graph of the indicators:

Volume (Mvol):

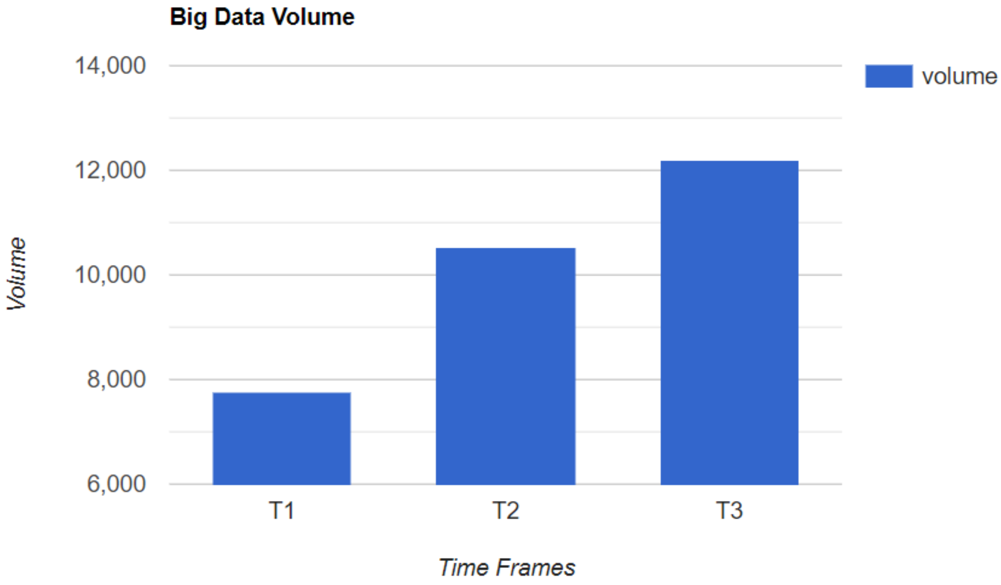


Figure 9: Big Data Volume against time frame

Velocity (Mvel):

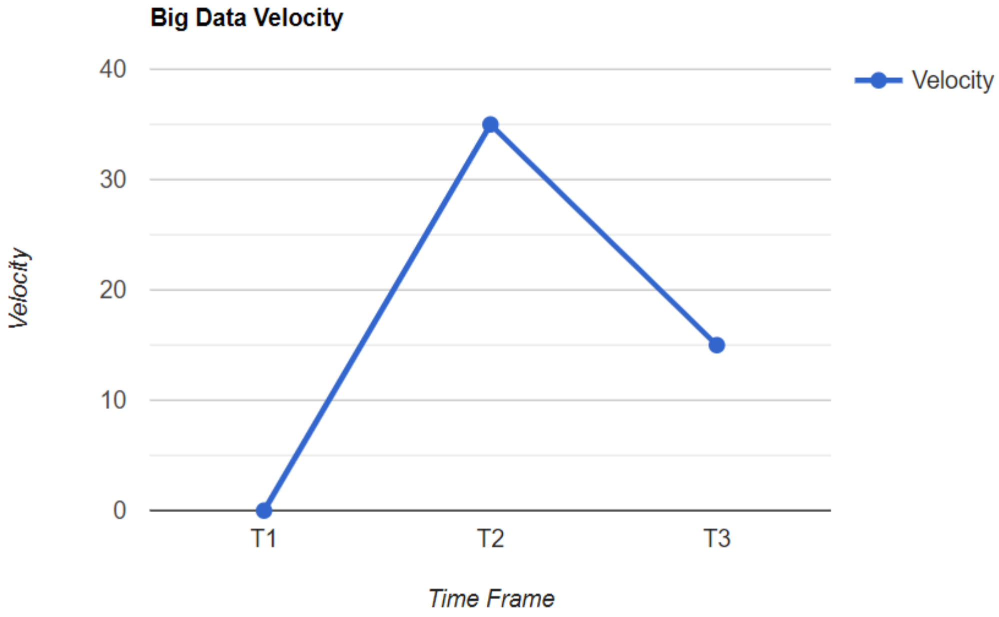


Figure 10: Big Data Velocity against time frame

Variety (Mvar):

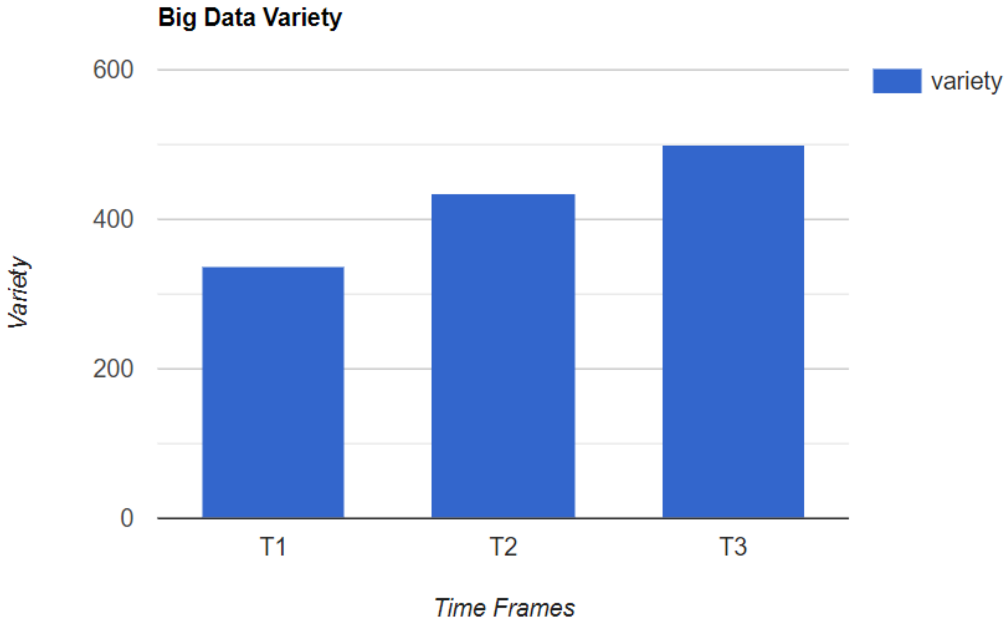


Figure 11: Big Data Variety against time frame

5. Conclusions:

From this indicator graph, we can see that there are clear trends in terms of the volume, velocity, and variety of datasets over the 3 time frames (T1, T2, and T3).

From T1 to T3, the volume of big data has been observed to increase gradually over time. It shows that there has been a substantial amount of data added to the dataset for the T2 and a considerable amount for the T3 timeframe.

As stated in previous steps, increasing or decreasing the value of varieties does not make them better or worse. It only gives information about the amount of Ndde, Lbd and Nds present in the dataset. It is evident from the indicator graph that there has been no significant change over the three timeframes (T1, T2, and T3). However, each time frame has seen a slight increase over the previous one.

Across the various time frames, the velocity measure varied the most. In the dataset at time T2, we observed a significant increase in velocity, whereas at time T3, the rate of change decreased. This is not surprising since the volume changed from T1 to T2 so much more than it did from T2 to T3.

Even though the dataset is not big enough, we used this dataset for prototyping and found the dataset suitable for exercising measurement activities. We could find that the dataset has structured data suitable for processing using a machine learning algorithm. The result of the machine learning algorithm usually requires more data for processing for accurate prediction, but considering more data can be added to the dataset in the future, and based on a visual analysis of three major quality characteristics, we can conclude that the data is suitable for machine learning algorithms.

Annexure:

Datasource:	https://www.kaggle.com/datasets/anoopkumarraut/most-popular-superhero-tv-shows
Data Analysis notebook:	*https://colab.research.google.com/drive/1NdWP_SZJlGmiGuCO54YRnihZCG0JAIQt?usp=sharing*