CONCORDIA UNIVERSITY

GINA CODY SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

SOEN 6611
SOFTWARE MEASUREMENT: THEORY AND PRACTICE

Source: SEI Implementing Goal-Driven Measurement course material (adapted)

Declaration:
We, the members of the team, have read and understood the Fairness Protocol and the Communal Work Protocol, and agree to abide by the policies therein, without any exception, under any circumstances, whatsoever.

Step 5 Team Report

FALL 2022

Submitted To: Dr. Olga Ormandjieva

Team No. 7	
#Student ID	**Name**
40200496	SIDDHARTHA NANDA
40201807	RAJAT KUMAR
40208194	BARIQ ISHTIAQ MOHAMMED
40218245	VIKYATH SRINIVASULU

TABLE OF CONTENTS

1. Dataset Description

1.1 Datasets Name

The two datasets we chose to perform our Data extraction are:

a) Netflix Movies and TV Shows
b) Netflix data with IMDB scores added

1.2 Source

Data Set 1 - https://www.kaggle.com/datasets/shivamb/netflix-shows
Data Set 2 - https://www.kaggle.com/datasets/sarahjeeeze/imdbfile

1.3 Context:

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc. One data set has ratings of iMbd and other one has of Mpaa.

1.4 Content: Data Dictionary

Both datasets have same Column Names which are as follows -

Features	Define
show_id	Ids of the show/movie
type	Tells if the record is of a Movie or TV Show
title	Title of the record
director	Gives the name of the director
cast	Tells the actors involved
country	Country that it is shot in
date_added	Body of the comment
release_year	Date and time of creation
rating	IMBD rating / MPAA rating (different datasets)
duration	Duration of the record
listed_in	Tells which category it is listed in
description	Describes the shoe/movie briefly

Table 1: Data Dictionary of the dataset

1.5 **Size of Dataset:**

a) Dataset 1 – 1 MB
b) Dataset 2 – 999 kB

1.6 **Acknowledgments:**
Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies on an internet-connected device. Depending on your plan, you can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.

1.7 **Structure of Data:**

Dataset 1 – 12 columns and 6234 rows (text_format - String and numerical)
Dataset 2 – 12 columns and 8807 rows (text_format - String and numerical)

Excel file after merging the datasets - merged data.xlsx

1.8 No of Records:
Dataset 1- 6234, No of unique records: 6234
Dataset 2- 8807, No of unique records: 8807

Files can be found –
1) NETFLIX DATASET\NETFLIX TVANDMOVIESHOWS\netflix_titles.csv
2) NETFLIX DATASET\NETFLIX TVANDMOVIESHOWS_With_rating\mycsvfile.csv

1.9 Details:
A sample details of screenshot of our CSV files is given below:

Dataset 1 -

netflix_titles.csv (3.4 MB)

Detail Compact Column 10 of 12 columns

About this file

All TV Shows and Movies meta data on Netflix. Updated every month.

show_id	type	title	director	cast	countr
Unique ID for every Movie / Tv Show	Identifier - A Movie or TV Show	Title of the Movie / Tv Show	Director of the Movie	Actors involved in the movie / show	Country v / show wa
8807 unique values	Movie 70% TV Show 30%	8807 unique values	[null] 30% Rajiv Chilaka 0% Other (6154) 70%	[null] 9% David Attenborough 0% Other (7963) 90%	United St India Other (5C
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United S
s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Gree...	South A
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddin...	
s4	TV Show	Jailbirds New Orleans			
s5	TV Show	Kota Factory		Mayur More, Jitendra Kumar, Ranjan Raj	India

Dataset 2 –

Netflix data with IMDB scores added

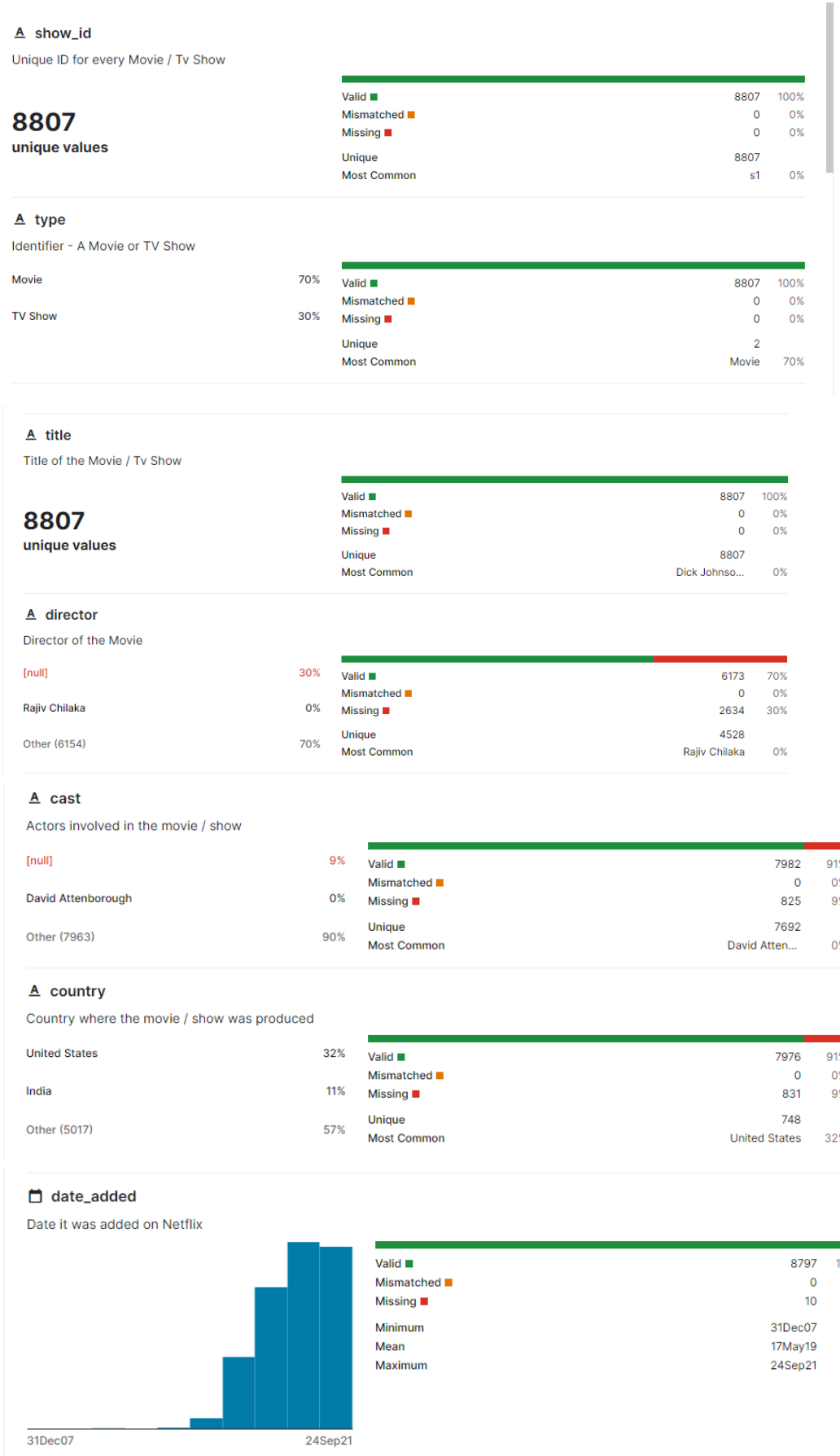
Data Code (2) Discussion (0) 8 New Notebook

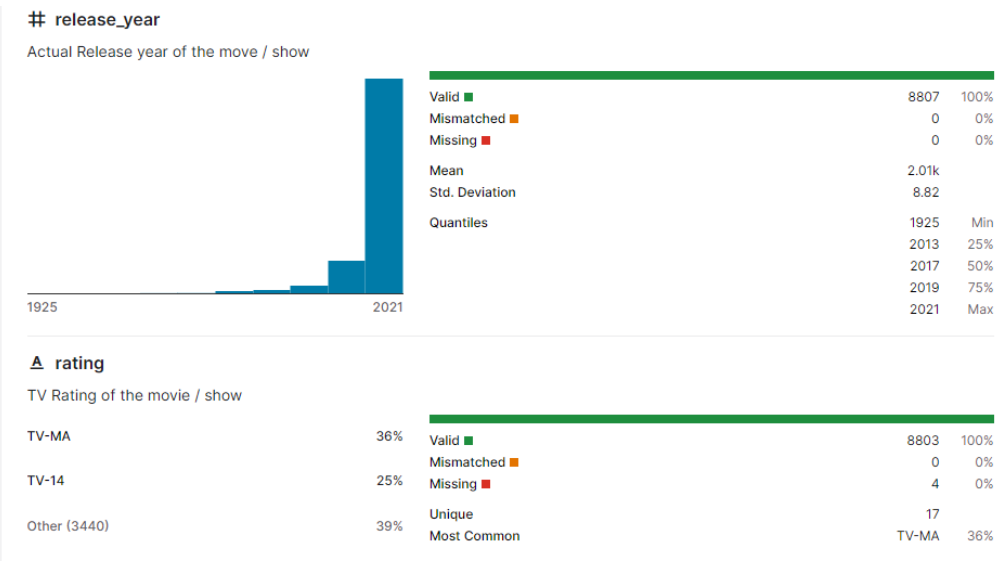
mycsvfile.csv (2.4 MB)

Detail Compact Column 10 of 12 columns

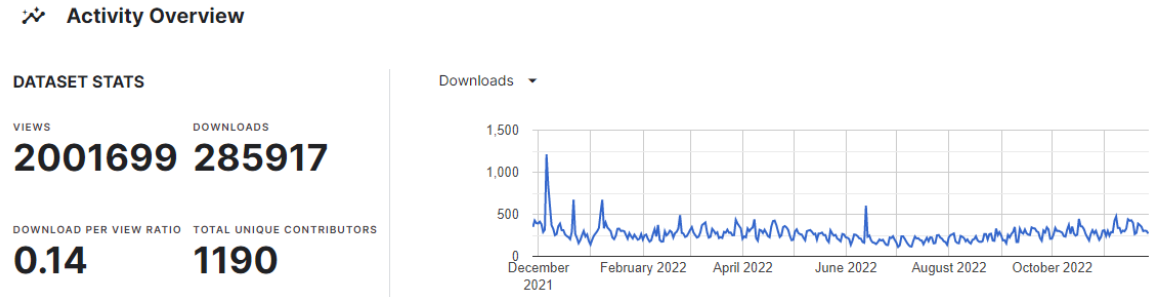
show_id	type	title	director	cast	countr
248k 81.2m	Movie 68% TV Show 32%	6172 unique values	[null] 32% Raúl Campos, Jan S... 0% Other (4247) 68%	bam 9% David Attenborough 0% Other (5646) 91%	United St India Other (34
81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole Howard, Jennifer Cameron, Jonathan Holmes, Lee Tockar...	United S India, S China
80117401	Movie	Jandino: Whatever it Takes		Jandino Asporaat	United K
70234439	TV Show	Transformers Prime		Peter Cullen, Sumalee Montano, Frank Welker, Jeffrey Combs, Kevin Michael Richardson, Tania Gunadi, ...	United S
80058654	TV Show	Transformers: Robots in Disguise		Will Friedle, Darren Criss, Constance Zimmer, Khary Payton, Mitchell Whitfield, Stuart Allan, Ted Mc...	United S
80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins, Keith Powers, Alicia Sanz, Jake Borelli, Kid Ink, Yo...	United S

1.10 Columns and Activity Overview:

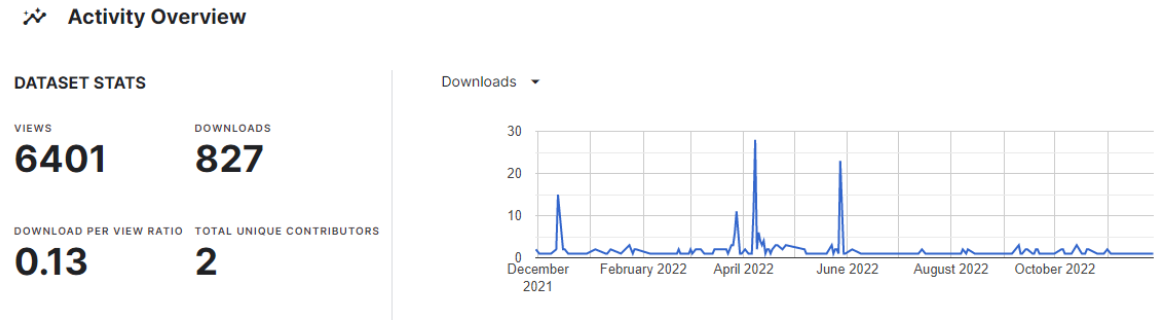




Dataset 1



Dataset 2



1.11 Data Extraction:

Google Collab link : (https://colab.research.google.com/drive/1Em36eUICt5B2BW79qaB3FmBIUkfQz__y?usp=sharing)

We have combined two datasets on the basis of common column and then further divided the data set into 3 subsets based on time frames (T1 , T2, T3) assuming that it was collected and recorded in real time.

2. Data Collection (Base Measures)

2.1 Collection Procedure for Nds:

Number of datasets (Nds) is calculated by the data scientists **manually**. The datasets for the project were two. This measure will be used to calculate various derived measures which in-turn will give us Variety.

2.2 Collection Procedure for Lbd:

Total number of records in the big data (LBD). As we had two data sets for the project but when combined to one, we divided it into time frames and calculated the number of records at the start of each frame

```
#LBD
lbd_1 = len(new_df_1)
lbd_2 = len(new_df_2)
lbd_3 =len(new_df_3)
```

Figure : Measuring lbd using python

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

```
#Ndde - unique records in T1
ndde_1=len(new_df_1['index'].unique()) +len(new_df_1['type'].unique())+ \
  len(new_df_1['title'].unique()) + len(new_df_1['director'].unique()) +\
  len(new_df_1['cast'].unique()) + len(new_df_1['country'].unique())+\
  len(new_df_1['date_added'].unique()) + len(new_df_1['release_year'].unique())+\
   len(new_df_1['rating_x'].unique()) + len(new_df_1['duration'].unique())+\
  len(new_df_1['listed_in'].unique()) + len(new_df_1['description'].unique())+\
   len(new_df_1['rating_y'].unique())
#Ndde - unique records in T2
ndde_2=len(new_df_2['index'].unique()) +len(new_df_2['type'].unique())+ \
  len(new_df_2['title'].unique()) + len(new_df_2['director'].unique()) +\
  len(new_df_2['cast'].unique()) + len(new_df_2['country'].unique())+\
  len(new_df_2['date_added'].unique()) + len(new_df_2['release_year'].unique())+\
   len(new_df_2['rating_x'].unique()) + len(new_df_2['duration'].unique())+\
  len(new_df_2['listed_in'].unique()) + len(new_df_2['description'].unique())+\
   len(new_df_2['rating_y'].unique())
#Ndde - unique records in T3
ndde_3=len(new_df_3['index'].unique()) +len(new_df_3['type'].unique())+ \
  len(new_df_3['title'].unique()) + len(new_df_3['director'].unique()) +\
  len(new_df_3['cast'].unique()) + len(new_df_3['country'].unique())+\
  len(new_df_3['date_added'].unique()) + len(new_df_3['release_year'].unique())+\
   len(new_df_3['rating_x'].unique()) + len(new_df_3['duration'].unique())+\
  len(new_df_3['listed_in'].unique()) + len(new_df_3['description'].unique())+\
   len(new_df_3['rating_y'].unique())
print("Ndde_1-",ndde_1)
print("Ndde_2-",ndde_2)
print("Ndde_3-",ndde_3)
```

Figure : Measuring Ndde using python

2.4 Collection procedure to measure T (Time):

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.

Before pre processing

T1 file – new_df_1 data.xlsx

T2 file – new_df_2 data.xlsx

T3 file - new_df_3 data.xlsx

After pre processing

T1 file – after_pre_process_df1 data.xlsx

T2 file – after_pre_process_df2 data.xlsx

T3 file - after_pre_process_df3 data.xlsx

3. Collected Data Values

3.1 The data values collected across three-time frames for the corresponding measure are:

a) Before pre-processing

Data Collected/ (Before Pre-processing)	T1	T2	T3
Ndde	7076	7427	8816
Nds	2	2	2
Lbd	1306	1305	1605
Volume	90492.9709	95500.5538	115541.6894
Velocity	0	5.5336	20.985
Variety	2794.666	2911.3333	3474.33333

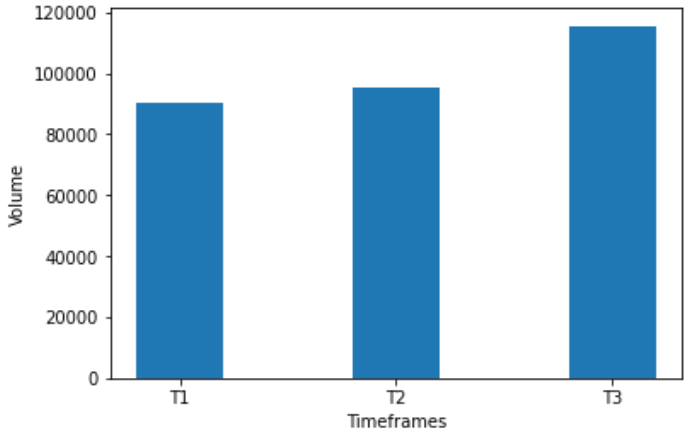
Table 2: Data Values before pre processing

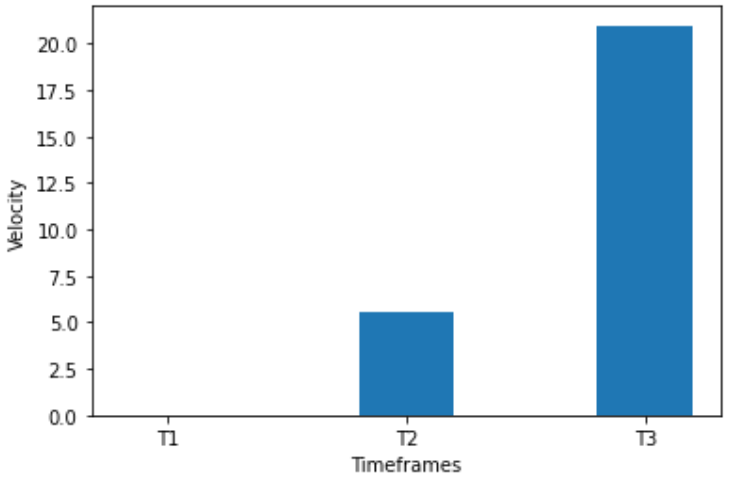
b) After pre-processing

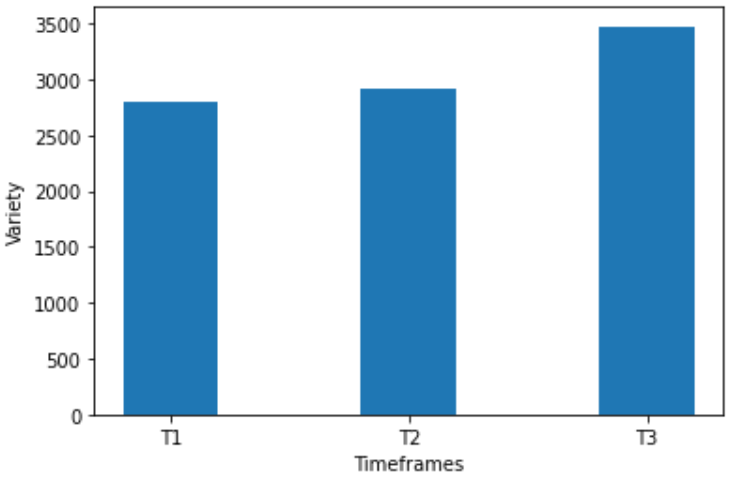
Data Collected/ (After Pre-processing)	T1	T2	T3
Ndde	1146	1553	2029
Nds	2	2	2
Lbd	206	281	376
Volume	11646.1004	16463.1078	22291.7163
Velocity	0	35.404	41.361
Variety	451.333	611.9999	802.3333

Table 3: Data Values after pre processing

The plots that can be mapped on the axes for the three V's are as follows for the base measures before pre-processing of data and timeframe division:


120000
100000
80000
60000
40000
20000
0
Volume
T1
T2
T3
Timeframes


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Velocity
T1
T2
T3
Timeframes


3500
3000
2500
2000
1500
1000
500
0
Variety
T1
T2
T3
Timeframes

4. Indicator Values (3V's)

4.1 Values of the corresponding derived measures

The base measures values calculated in previous sections were used to derive the 3V's by using the following formulas:

Derived Measures	Formulas	Base measures used
MVol	$Mvol\ (MDS) = Ndde\ (MDS) * log_2\ ((Ndde\ (NDS)))$	Ndde Nds
MVel	$Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100$	T (Time)
MVar	$Mvar\ (MDS) = Ndde\ (DE) * W_{Ndde} + Lbd\ (MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}$	Lbd Ndde Nds

The derived values of each derived measures in three different time frame:

Measures (Before pre-processing)	T1	T2	T3
Mvol	90492.9709	95500.5538	115541.6894
Mvel	0	5.5336	20.9853
Mvar	2794.6666	2911.3333	3474.3333

Table 4: 3Vs Values of derived measures before pre-processing

Measures (After pre-processing)	T1	T2	T3
Mvol	11646.1004	16463.1078	22291.7163
Mvel	0	35.404	41.36
Mvar	451.3333	611.999	802.3333

Table 5: Final Values of derived measures after pre processing

	Average Value	
Measures	**Before**	**After**
Mvol	100511.7380	16800.3081
Mvel	8.8396	25.4533
Mvar	3060.1110666	621.8884

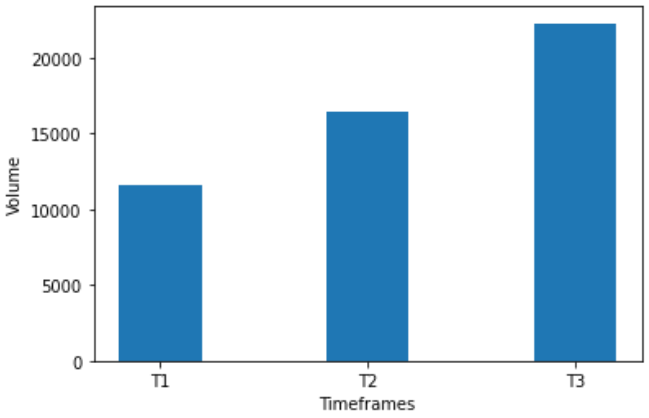
Table 6: Average values of derived measures

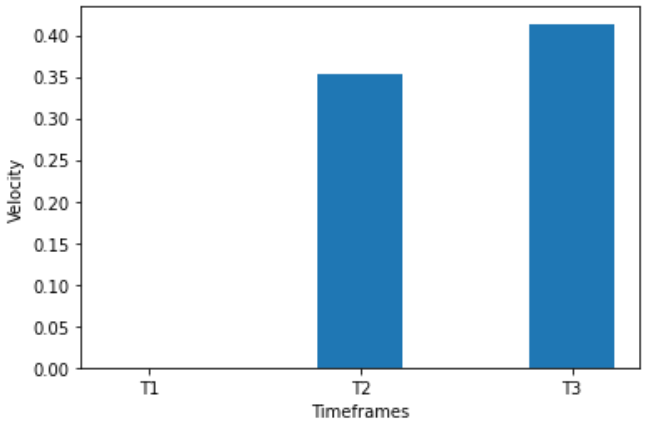
Final Values at the end of the process:

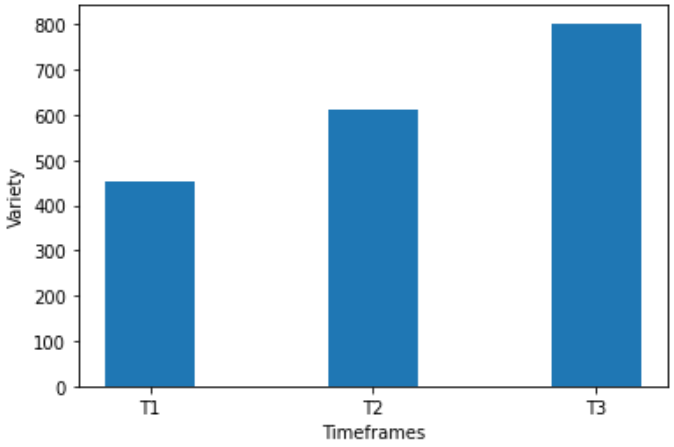
Mvol -> 16800.3081
Mvel -> 25.4533
Mvar -> 621.8884

3V's Data Quality Value (W_Vol*Vol + W_Vel*Vel + W_Var*Var) = 5815.675

Final Graphs after – Pre processing







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 gradual change over the three timeframes (T1, T2, and T3). However, each time frame has seen a slight increase over the previous one.

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.

6. Additional links

Data source:	Data Set 1 - https://www.kaggle.com/datasets/shivamb/netflix-shows Data Set 2 - https://www.kaggle.com/datasets/sarahjeeeze/imdbfile
Data Analysis notebook:	https://colab.research.google.com/drive/1Em36eUICt5B2BW79qaB3FmBIUkfQz__y?usp=sharing