Stakeholder Update.

predicting movie success using Movies MARK®

Stakeholder Updates:

- Link to Repository
- Link to ReadMe.md file for project overview
- Link to description of machine learning model
- Link to database overview
- Link to dashboard blueprint
- See below for presentation blueprint

What movie do you want to watch next?

Note: this yellow box format indicates working notes for the team to serve as reminders

predicting movie success using Movies MARK®

Everyone loves movies.

Streaming has reshaped cinema and the COVID-19 pandemic has left many of us wondering “what should we watch next?”

Using data-wrangling, programming, and machine learning skills, we plan to answer:

What makes movies successful?

Our Team

Of budding Data Scientists collaborated virtually across Zoom and Slack (as "the_clever_crew") to bring you this fine work.

- **Maggie Allen**
  Presentation
- **Andrew Malony**
  GitHub + Graphs
- **Rose Baumann**
  Database
- **Kathy Morrissey**
  Machine Learning Model

But first we must ask
How do we define Movie Success?

suc·cess

/səkˈses/

noun

1. the accomplishment of an aim or purpose.
   "there is a thin line between success and failure"
2. **ARCHAIC**
   the good or bad outcome of an undertaking.
   "the good or ill success of their maritime enterprises"

suc·cess

/səkˈses/

noun sadfasdfa

1. Popularity (Proprietary ratings, User Ratings)
2. Estimated Profitability (Revenue-Budget)
3. Awards

Meet Ellen.

She is the owner of a new start up streaming service, Serenity Streaming. She's looking to use AI and Machine Learning to help connect users with their favorite movie they have never even heard of.

Right now she's still working out of her home office and realizes despite a ton of data, she needs a proof of concept machine learning model to get investment interest.

Meet Sam.

Sam is a newly promoted executive at ABC Movie Productions and is interested in determining the right mix of movie genre, Director's talents, and A-list actors are going to be the recipe for the next blockbuster.

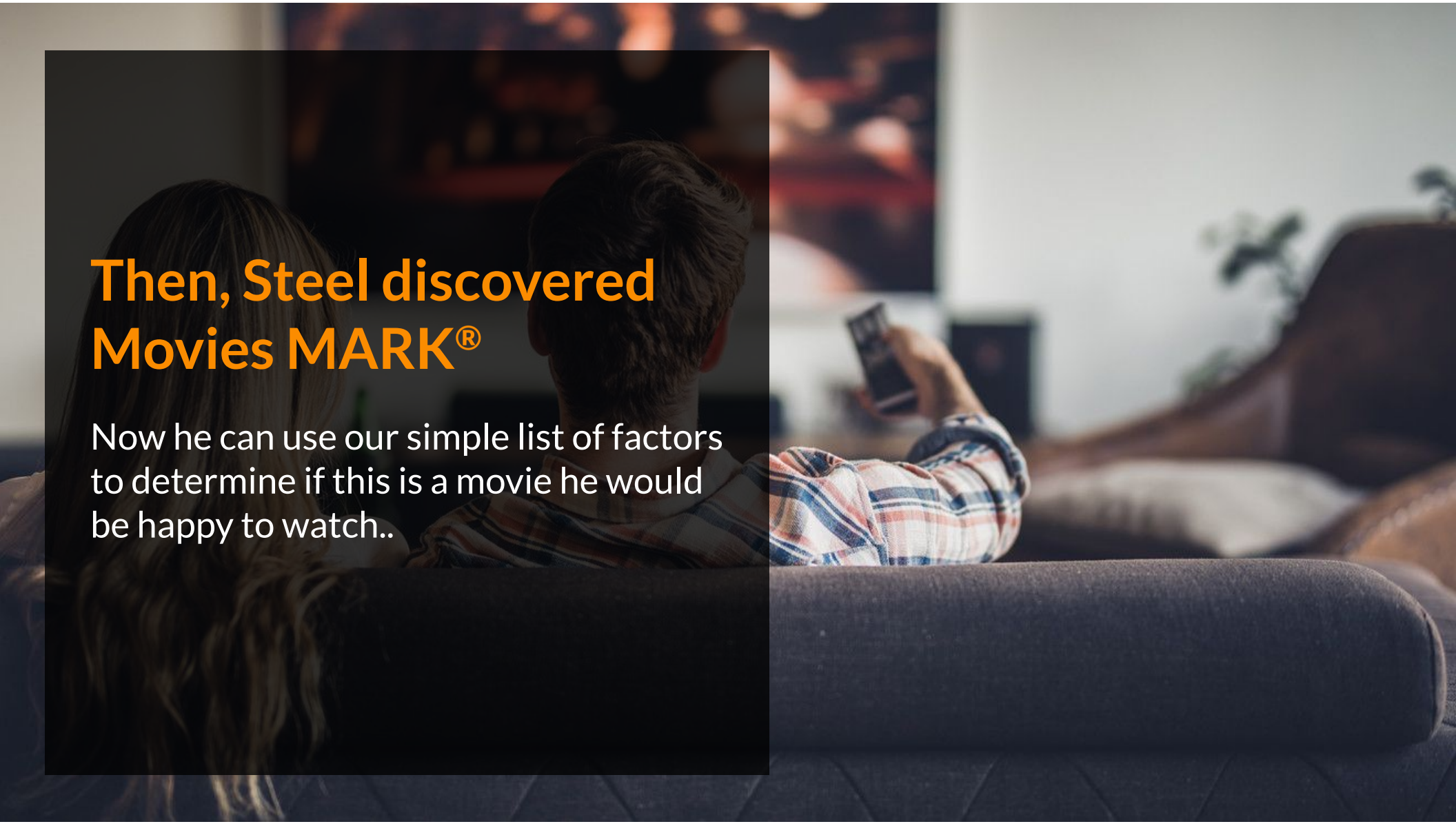
Before him, the boomers were sitting in rooms making all the calls but he thinks data science can flip the script.

Meet Steel.

He recently had a baby so he has no time to watch a bunch of bad movies. When he finally has a free evening, he wants the first movie he streams to be one he's happy to talk about with his new baby boy.

Let's see what he should look for in his next popcorn night's entertainment...


Whatever our
solution will
suggest, we should
summarize here.


Then, Steel discovered Movies MARK®
Now he can use our simple list of factors to determine if this is a movie he would be happy to watch..

Our Data

- Internet Movies DataBase (IMDb)
- The Movies DataBase (TMDB)
- Kaggle Movies DataSet

1970+

Source: https://www.imdb.com
https://www.themoviedb.org/?language=en-US
https://www.kaggle.com/rounakbanik/the-movies-dataset

Data Clean Up

Datatypes..

tbd

TBD

tbd

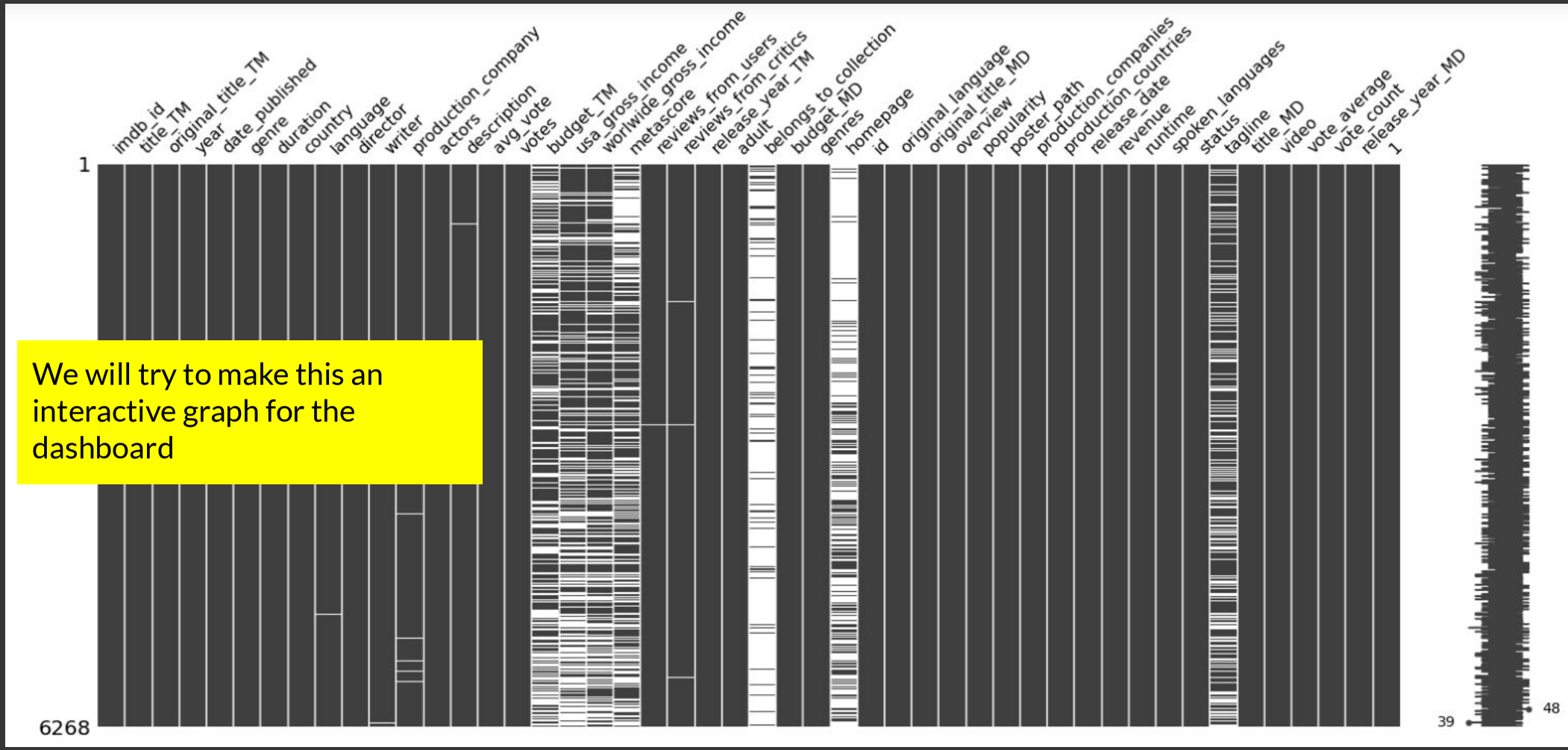
TBD

TBD

`country` has constant value "USA"	Constant
`imdb_id` has a high cardinality: 28511 distinct values	High cardinality
`title` has a high cardinality: 27678 distinct values	High cardinality
`original_title` has a high cardinality: 27056 distinct values	
`year` has a high cardinality: 111 distinct values	
`date_published` has a high cardinality: 13734 distinct values	High cardinality
`genre` has a high cardinality: 874 distinct values	High cardinality
`language` has a high cardinality: 650 distinct values	High cardinality
`director` has a high cardinality: 12463 distinct values	High cardinality
`writer` has a high cardinality: 23560 distinct values	High cardinality
`production_company` has a high cardinality: 11479 distinct values	High cardinality
`actors` has a high cardinality: 28469 distinct values	
`description` has a high cardinality: 28407 distinct values	
`budget` has a high cardinality: 1511 distinct values	
`usa_gross_income` has a high cardinality: 7333 distinct values	High cardinality
`worlwide_gross_income` has a high cardinality: 7643 distinct values	High cardinality

Overview of Values in Each Variable
(cardinality is a measure of set size)

Reformat into a excel table and add columns to say what we did with the data - Directors - binned, actors ignored, writers dropped, etc.



Missing Data Fields by Variable (Column)

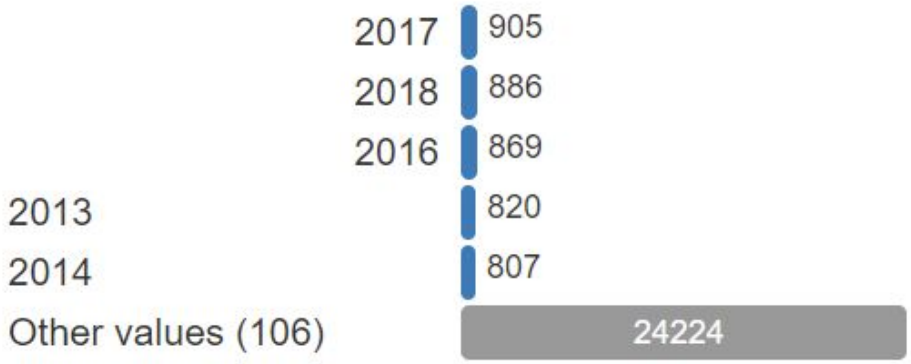

2019
2006
2001
2015
2002
2007
2003
2008
2016
Year: 2019
Reviews From Users: 250,446
2018
2010
1997
1994
1995
1996
2014
1999
2000
1993
1988
1987
1992
2004
2011
2020
1990
1980
2017
1986
1972
1985
1979
2005
2013
2009
1998
1983
1974
1991
1976
2 nulls

year

Categorical

HIGH CARDINALITY

Distinct	111
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	222.9 KiB

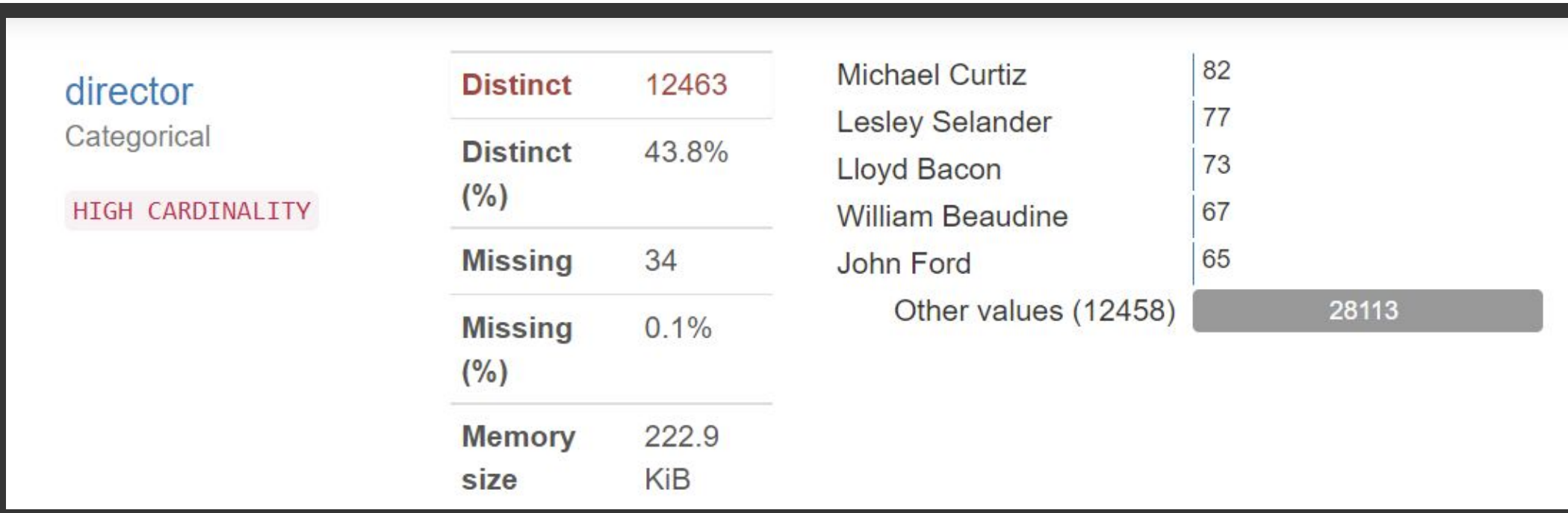


Observations in Year Variable

Common Values

Value	Count	Frequency (%)
2017	905	3.2%
2018	886	3.1%
2016	869	3.0%
2013	820	2.9%
2014	807	2.8%
2015	800	2.8%
2012	738	2.6%
2019	700	2.5%
2009	656	2.3%
2011	652	2.3%
Other values (101)	20678	72.5%

Choose which shows this better from prior slide



Observations in Director Variable

Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1.1
5-th percentile	3.1
Q1	4.8
median	5.8
Q3	6.5
95-th percentile	7.3
Maximum	9.7
Range	8.6
Interquartile range (IQR)	1.7

Descriptive statistics

Standard deviation	1.284809426
Coefficient of variation (CV)	0.2312437167
Kurtosis	0.05842961363
Mean	5.556083617
Median Absolute Deviation (MAD)	0.8
Skewness	-0.6237699936
Sum	158409.5
Variance	1.650735261
Monotonicity	Not monotonic

Average Vote Summary Statistics

Minimum 5 values | Maximum 5 values

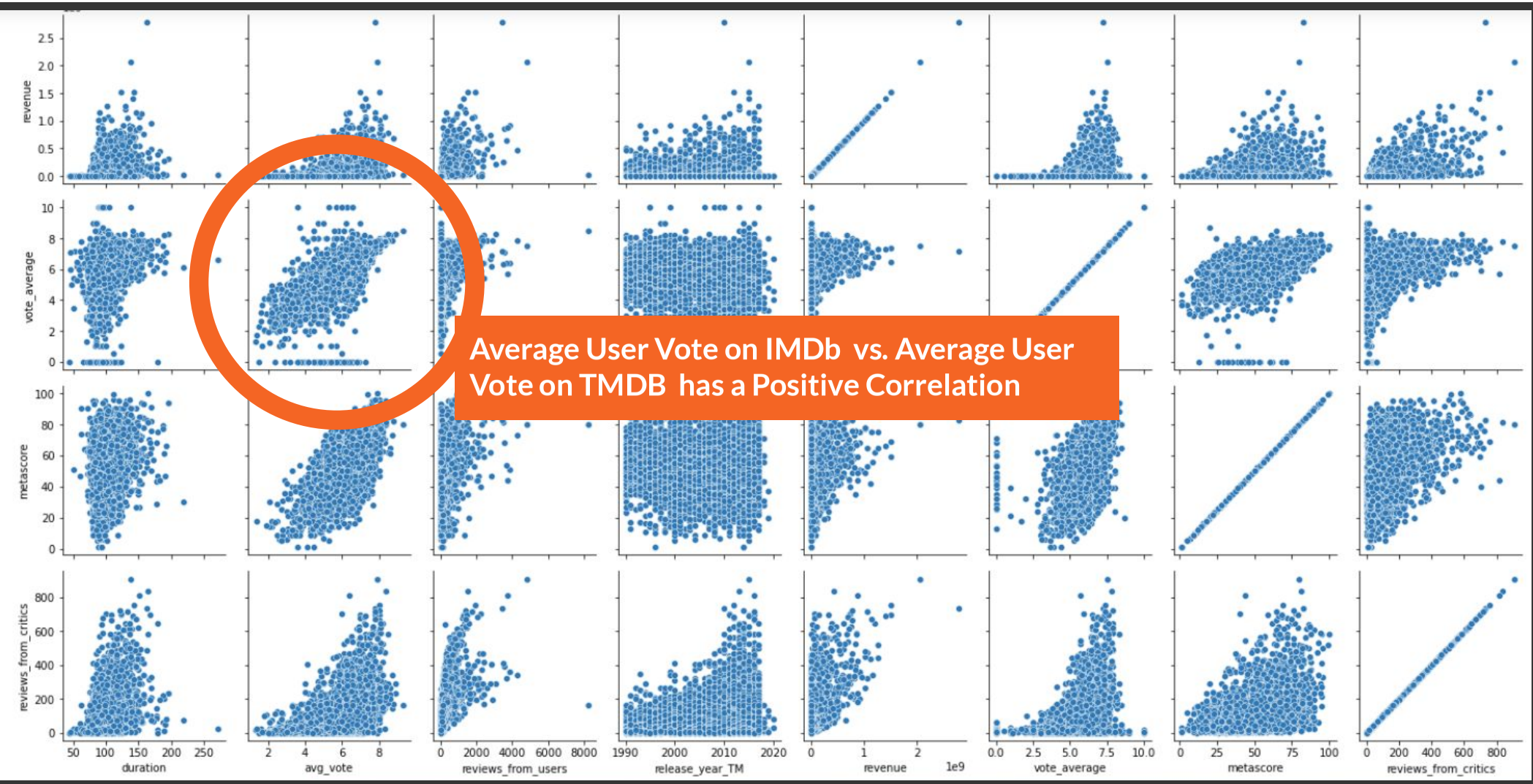
Value	Count	Frequency (%)
1.1	4	< 0.1%
1.2	8	< 0.1%
1.3	8	< 0.1%
1.4	8	< 0.1%
1.5	17	0.1%

Average Vote Minimums and Maximums

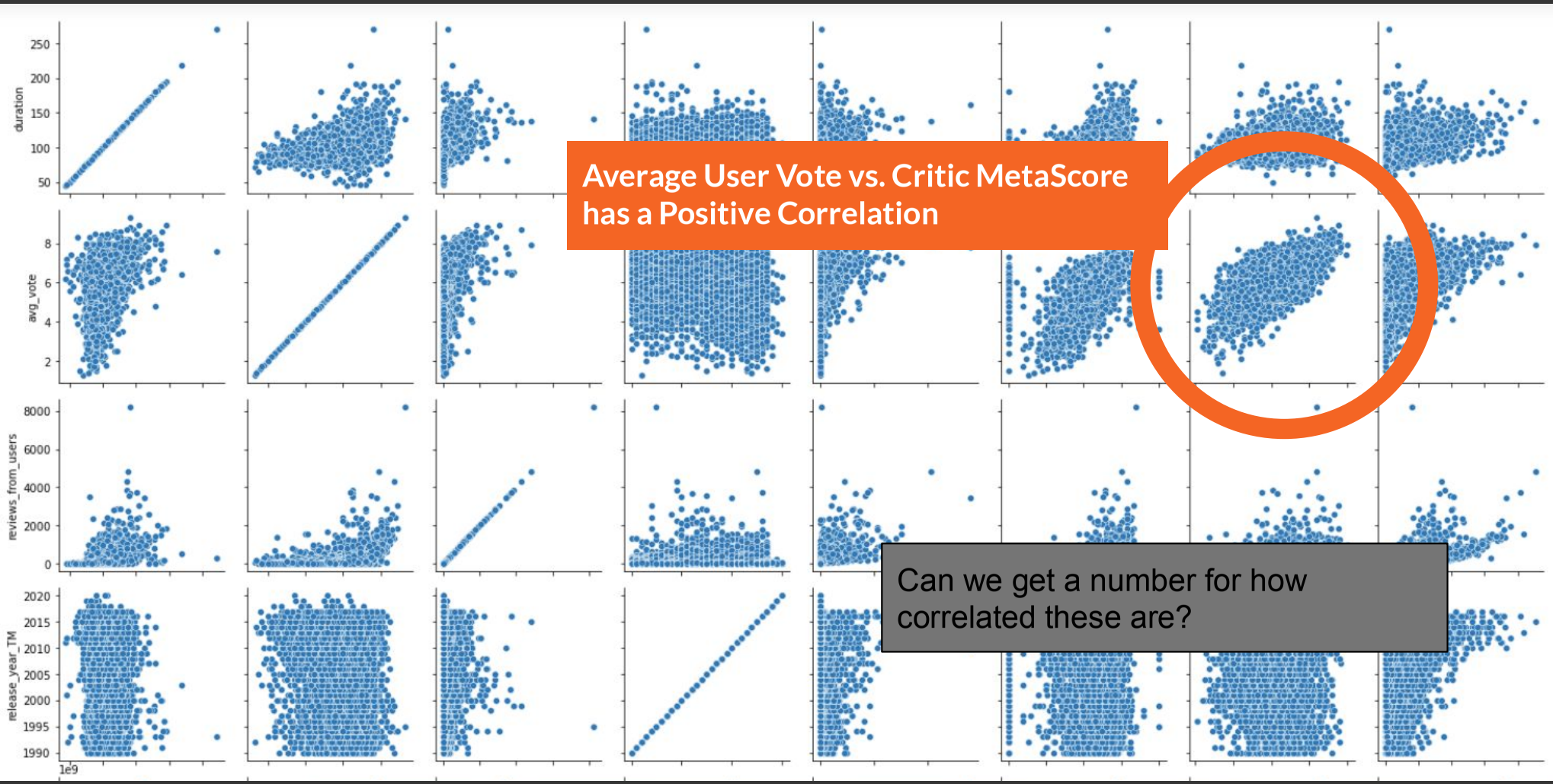
Our Findings

Summary Statement...

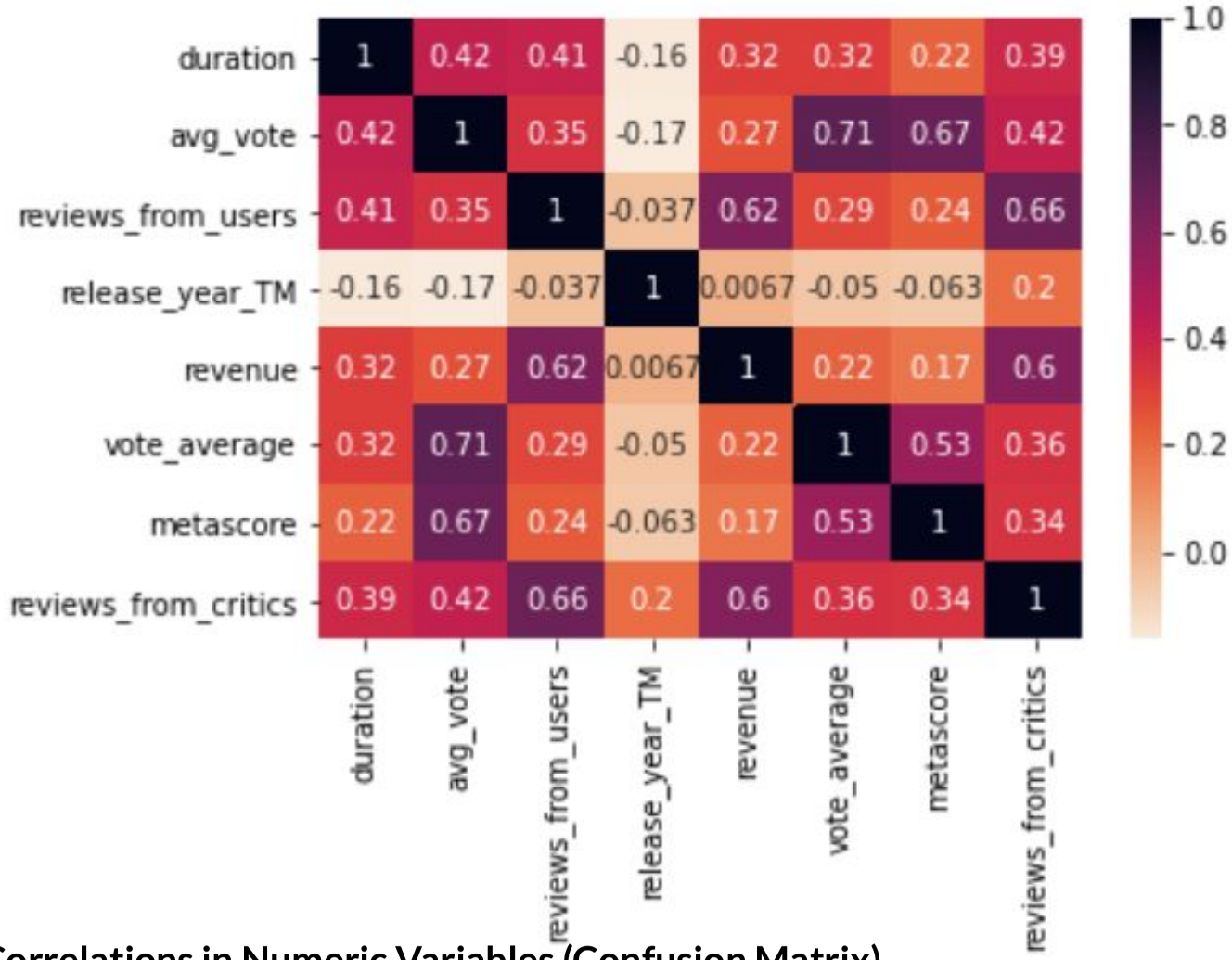
tbd



Relationships of Numerical Variables in Movies Data



Relationships of Numerical Variables in Movies Data

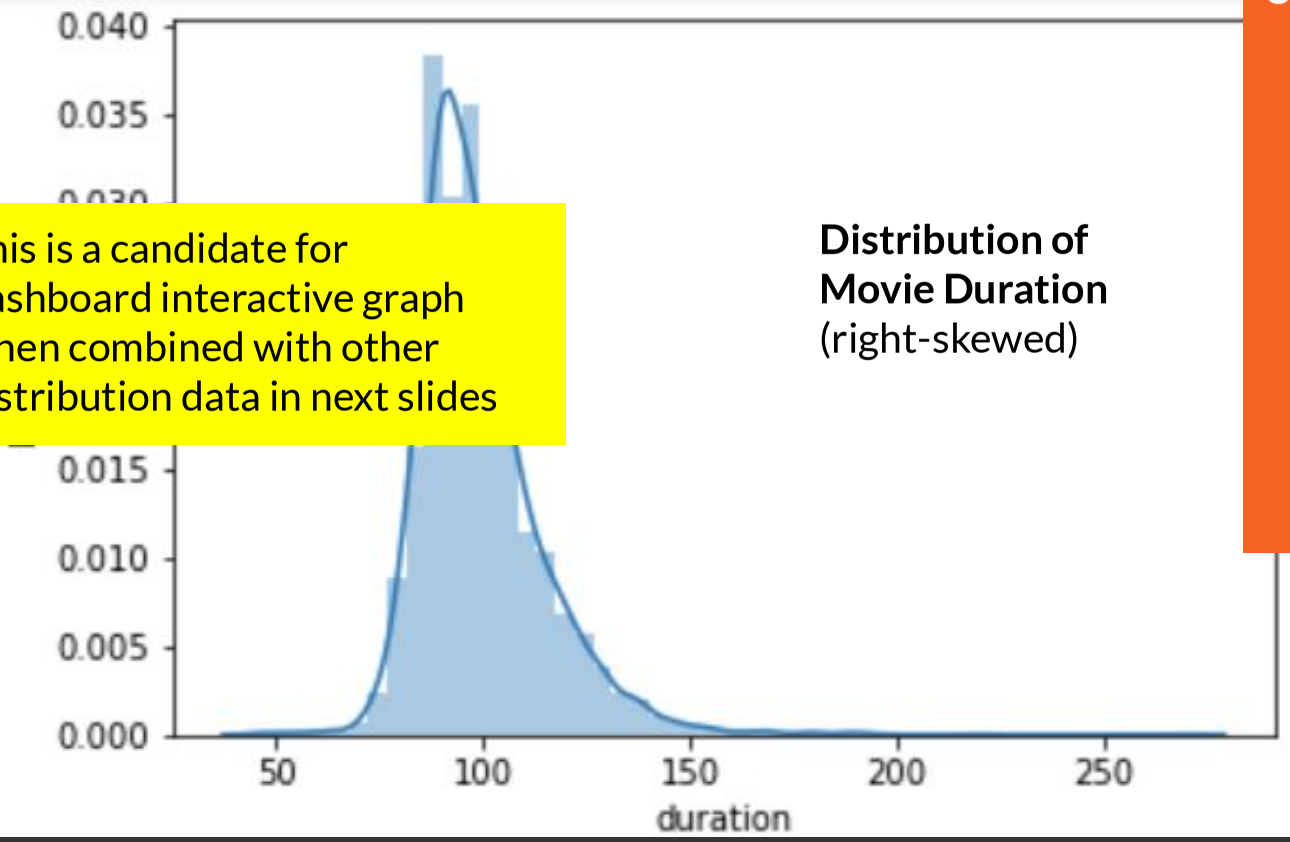


Correlations in Numeric Variables (Confusion Matrix)

CONCLUSION:

- Some positive correlation between revenue and user reviews
- Duration and Release Year has minimal correlations to other variables
- Note: Dropping revenue variable because data is incomplete
-

```
1 for column in new_movies_df.columns:
2     plt.figure()                  # <==========
3     sb.distplot(new_movies_df[column])
```

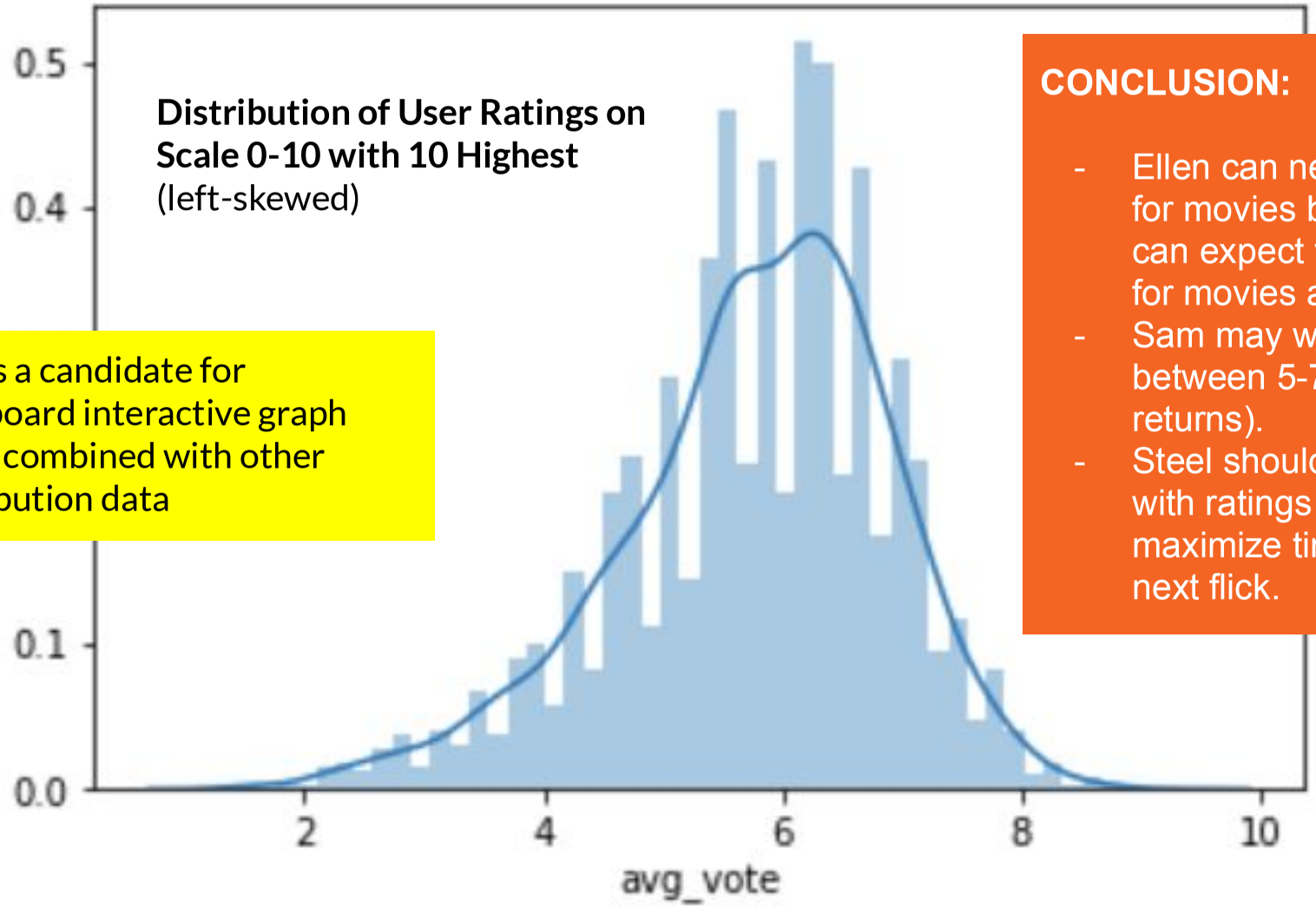


This is a candidate for dashboard interactive graph when combined with other distribution data in next slides

Distribution of Movie Duration
(right-skewed)

CONCLUSION:

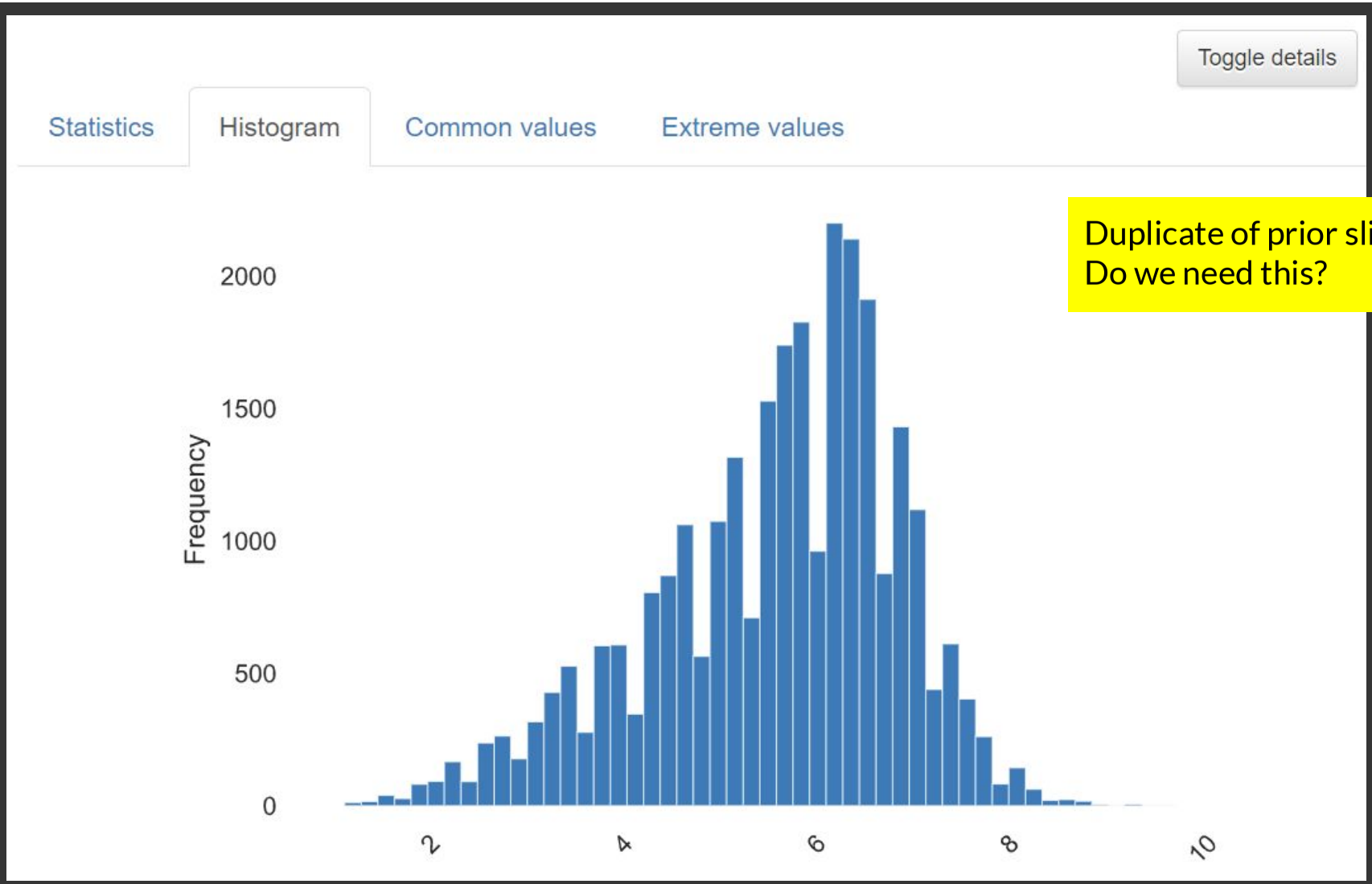
- Most movies are an average length of just over 90 minutes (mean = 93.05 minutes)
- Mode is 90 minutes with a standard deviation of 18.58 minutes… (SEE SPEAKER NOTES FOR MORE MEASURES OF CENTRALITY)



This is a candidate for dashboard interactive graph when combined with other distribution data

CONCLUSION:

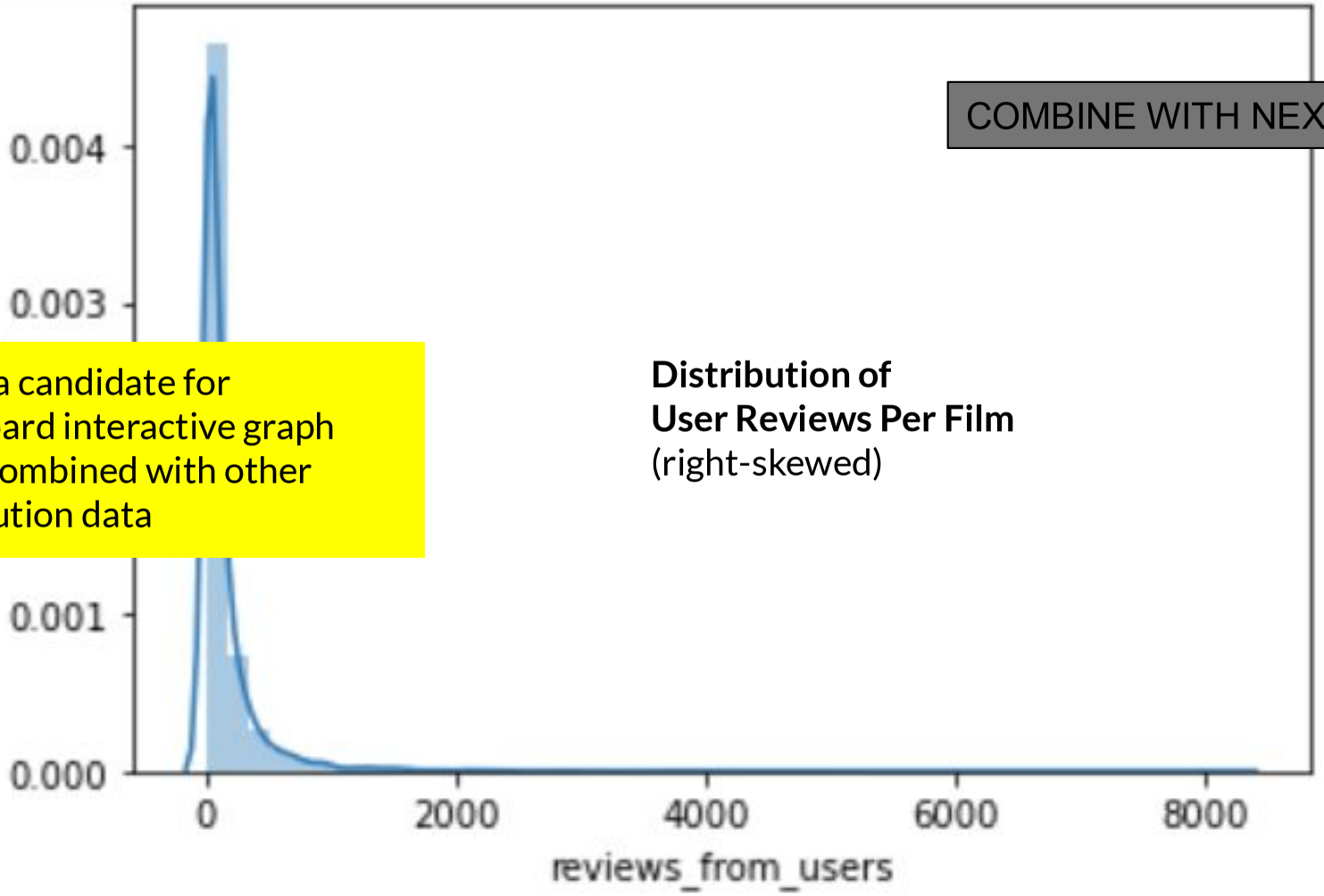
- Ellen can negotiate better deals for movies between 3-5 and can expect to pay a premium for movies above 8.
- Sam may want to target ratings between 5-7 (7+ is diminishing returns).
- Steel should look for movies with ratings between 7-8 to maximize time looking for his next flick.



Duplicate of prior slide-
Do we need this?

Distribution of Average Votes

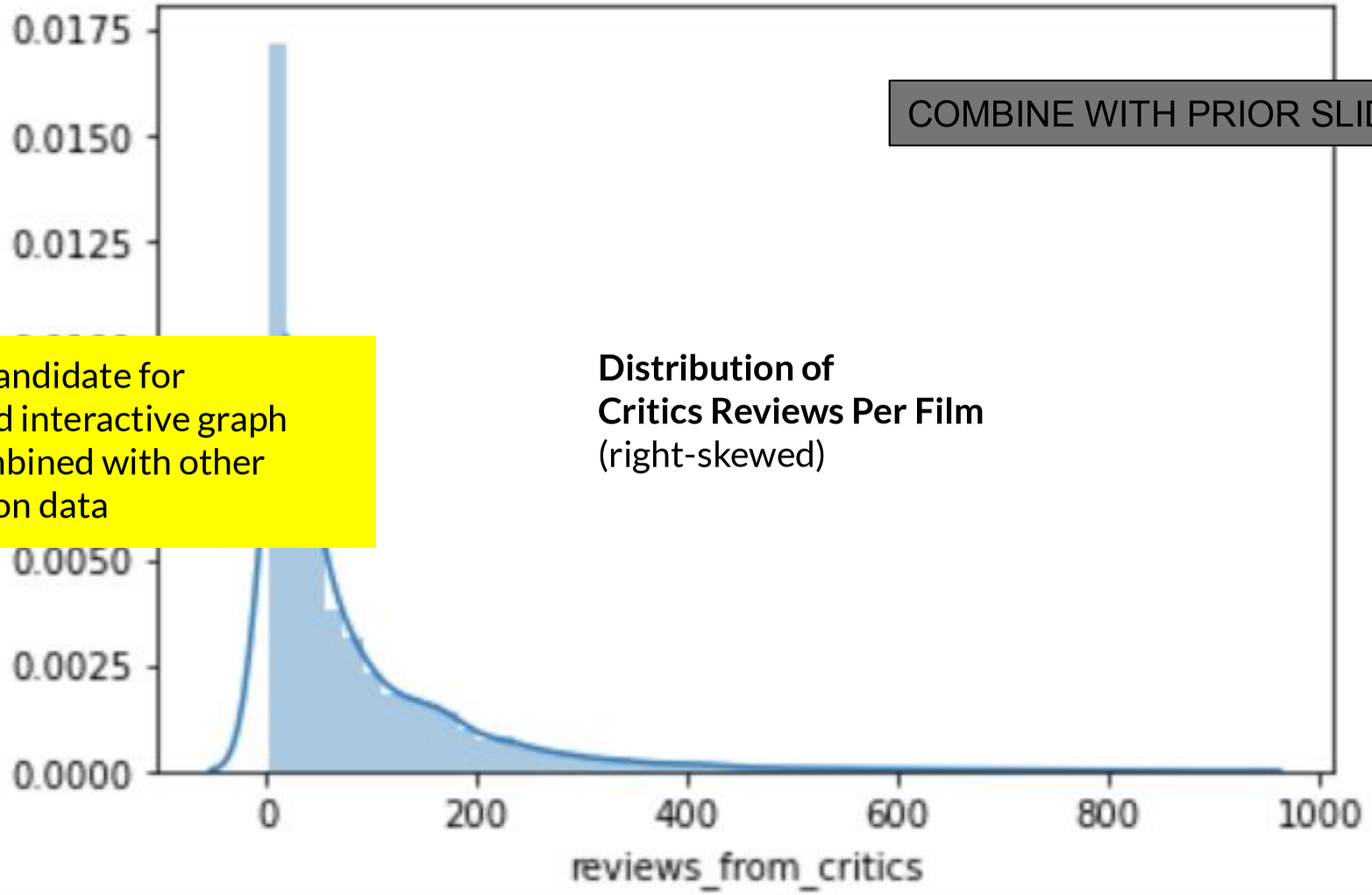
COMBINE WITH NEXT SLIDE



This is a candidate for dashboard interactive graph when combined with other distribution data

Distribution of
User Reviews Per Film
(right-skewed)

COMBINE WITH PRIOR SLIDE



This is a candidate for dashboard interactive graph when combined with other distribution data

Distribution of
Critics Reviews Per Film
(right-skewed)

Placeholder - How many critic reviews vs user reviews? Can we compare across datasets (IMDb v TMDB)

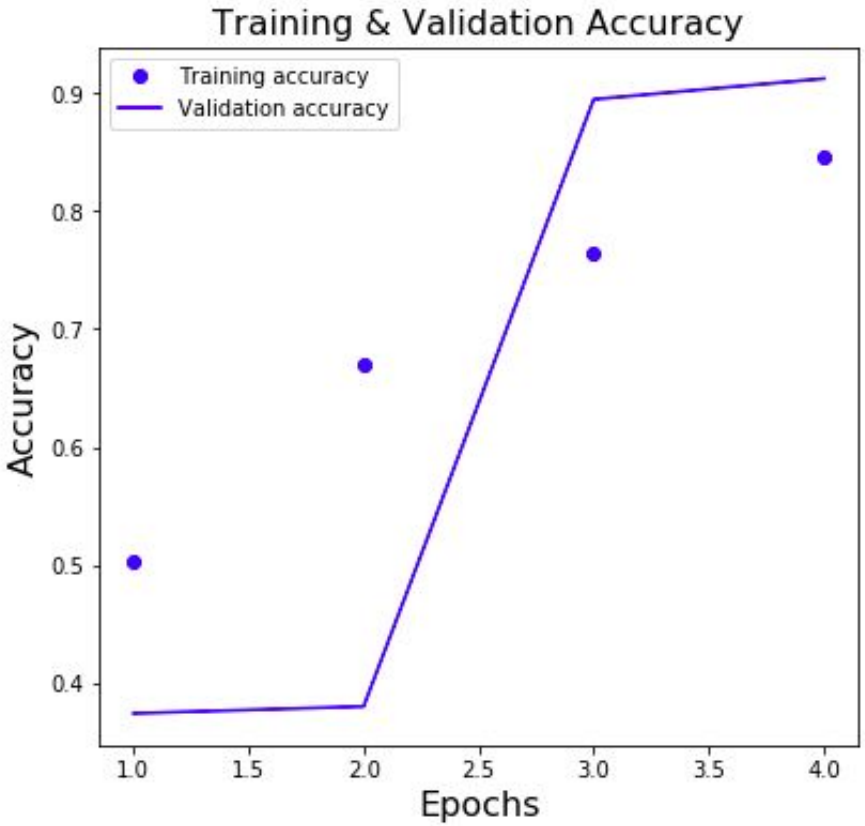
This is a candidate for dashboard interactive graph

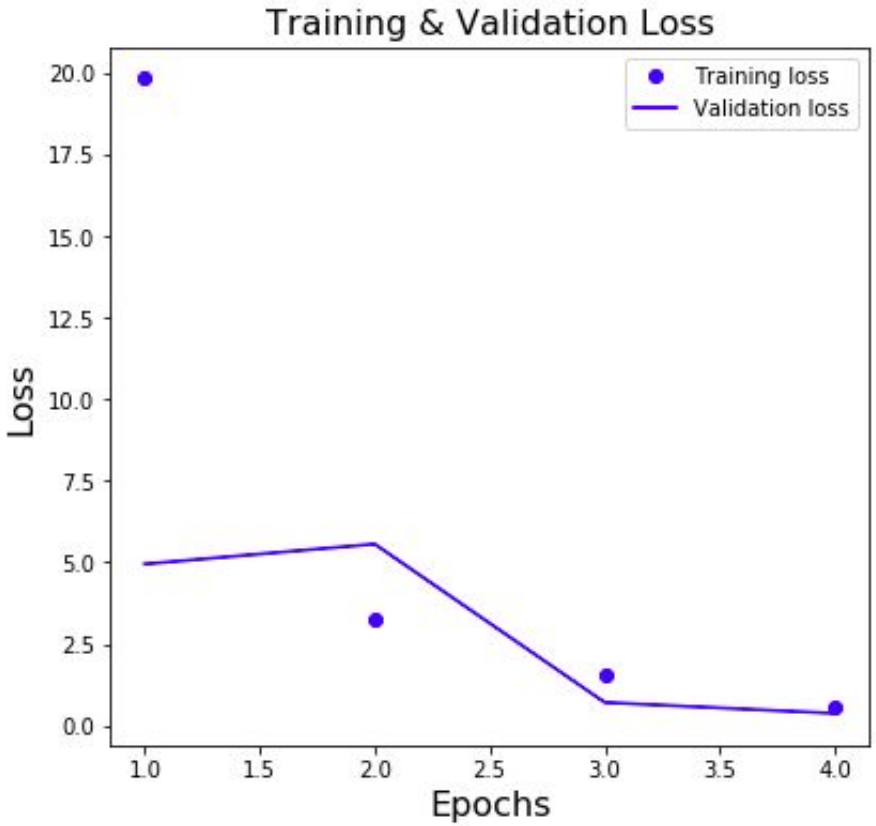
Placeholder - review counts by year

Placeholder for Machine Learning Model Analysis

Graphs from Machine Learning

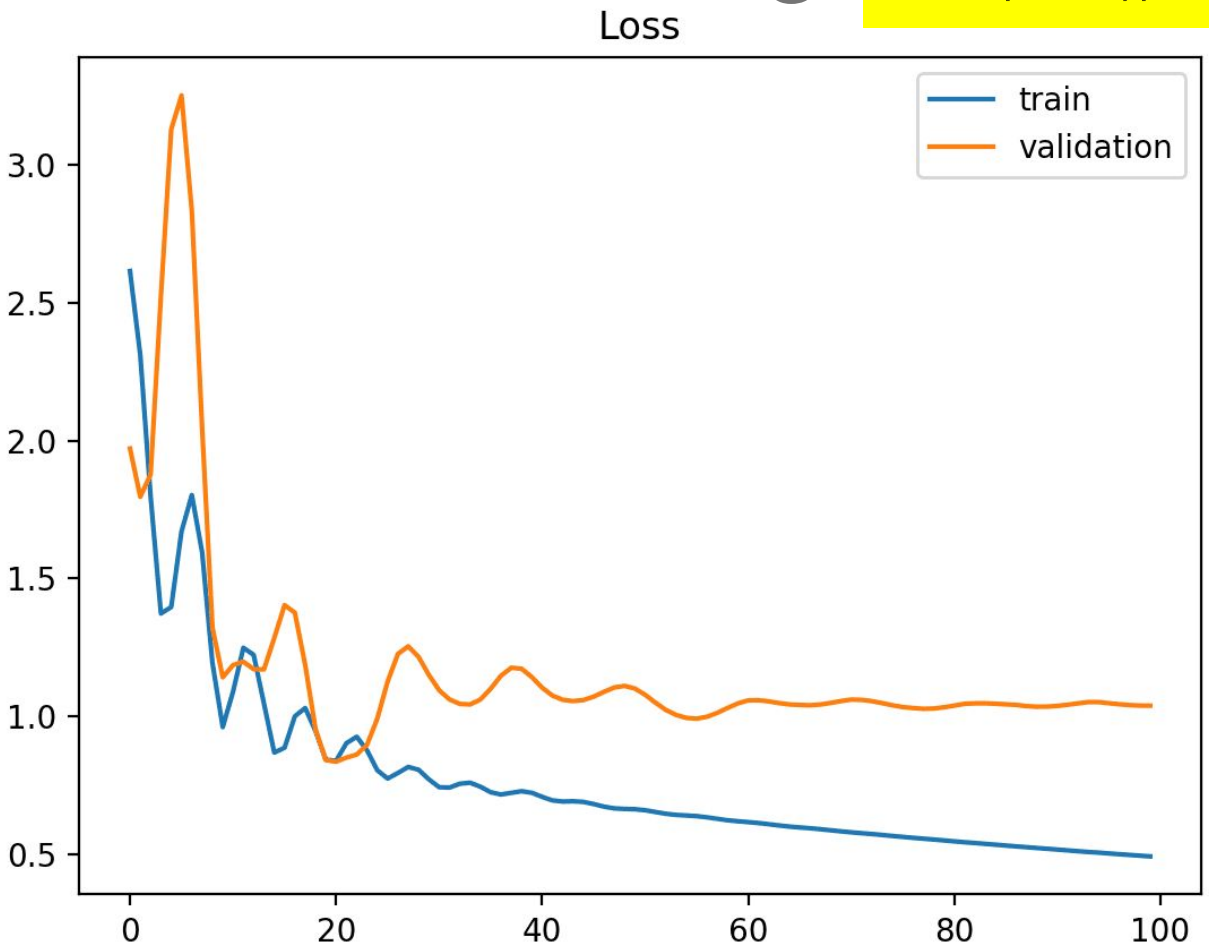
Placeholder - this is temp graphs from our machine learning model prototype





Graphs from Machine Learning

Placeholder - this is temp graphs from our machine learning model prototype



—

Clever saying to finish us off on a note with a touch of humor..

Some sort of summary bullet of interest.

Thank you.

APPENDIX

—

- Link to Repository
- Link to ReadMe.md file

Link to Profile Report

Pretty Graphs