Creating a new movie studio: Exploratory data analysis of movie data.

Saif Kasmani · Follow

11 min read · May 8, 2020

6

Project: Microsoft sees all the big companies creating original video content, and they want to get in on the fun. They have decided to create a new movie studio, but the problem is they don't know anything about creating movies. They have hired you to help them better understand the movie industry.

Our Goal: Your team is charged with doing data analysis and creating a presentation that explores what type of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the CEO can use when deciding what type of films they should be creating.

Data we worked with: some movie-related data from:
* Box Office Mojo
* IMDB
* Rotten Tomatoes
* TheMovieDB.org

We were given the following tables:

> *bom.movie_gross.csv*
> *name.basics.csv*
> *rt.movie_info.tsv*
> *rt.reviews.tsv*
> *title.akas.csv*
> *title.basics.csv*
> *title.crew.csv*
> *title.principals.csv*
> *title.ratings.csv*
> *tmdb.movies.csv*
> *tn.movie_budgets.csv*

All the above files can be found here :

https://github.com/learn-co-students/dsc-mod-1-project-v2-1-onl01-dtsc-ft-041320/tree/master/zippedData

Data Cleaning: First thing we had to look at was the necessary data cleaning to be done for the tables we would use from the given tables. We created some functions which would help us achieve DRY style of coding. The functions we used to clean/ change datatypes of certain columns are below:

```
import pandas as pd
import string

def convert_dollars(dollar):
 return int(dollar.replace("$", "").replace(",", ""))

def to_date(df,col):
 df[col] = pd.to_datetime(df[col])
 return df

def remove_punctuations(text):
 for punctuation in string.punctuation:
 text = text.lower().replace(punctuation, "")
 return text
```

In the Data Cleaning notebook we prepared our data to be exported for EDA. Removed unwanted columns and converted respective objects to date-time where needed. We had a goal to make one primary df with all the data we needed and formatted it as the desired data types needed for plotting and analysis.

Our final table structure is given below:

id	release_date	movie	production_budget	domestic_gross	genres	tconst	runtime	popularity	year_released	release_day	release_month	domestic_gross_in_mill	production_budget_in_mill	domestic_net_in_mill	Return_on_Investment
1	2009-12-18	Avatar	425000000	760507625	Horror	tt1775309	93.0	26.526	2009	4	12	760.507625	425.0	335.507625	78.942971
3	2019-06-07	Dark Phoenix	350000000	42762350	Action,Adventure,Sci-Fi	tt6565702	113.0	NaN	2019	4	6	42.762350	350.0	-307.237650	-87.782186
4	2015-05-01	Avengers: Age of Ultron	330600000	459005868	Action,Adventure,Sci-Fi	tt2395427	141.0	44.383	2015	4	5	459.005868	330.6	128.405868	38.840250
5	2017-12-15	Star Wars Ep. VIII: The Last Jedi	317000000	620181382	None	None	NaN	NaN	2017	4	12	620.181382	317.0	303.181382	95.640815
6	2015-12-18	Star Wars Ep. VII: The Force Awakens	306000000	936662225	None	None	NaN	NaN	2015	4	12	936.662225	306.0	630.662225	206.098766

The data has 4383 records and 16 columns.

Here we begin to add rows that will be valuable information to analyze later on. The production_budget and domestic_gross numbers are quite large so it may be easier to digest these numbers in terms of millions of dollars. We also create a column with the year released so that we may begin to do some time series analysis later on.

I have also added some columns that produce profitability metrics such as profit_in_millions and return_on_investment(%). These are important because we want to advise our client to make smart decisions with their money. No one is in the business of losing money.


```
tn_movie_budgets[‘domestic_gross_in_mill’] =
tn_movie_budgets[‘domestic_gross’] / 10**6

tn_movie_budgets['production_budget_in_mill'] =
tn_movie_budgets['production_budget'] / 10**6

tn_movie_budgets['domestic_net_in_mill'] =
tn_movie_budgets['domestic_gross_in_mill'] -
tn_movie_budgets['production_budget_in_mill']

tn_movie_budgets['Return_on_Investment'] =
(tn_movie_budgets['domestic_net_in_mill'] /
tn_movie_budgets['production_budget_in_mill']).apply(lambda x: x*100)

tn_movie_budgets['release_day_num'] =
tn_movie_budgets['release_date'].apply(lambda x:x.day)

tn_movie_budgets['release_month_num'] =
tn_movie_budgets['release_date'].apply(lambda x:x.month)

tn_movie_budgets['release_day'] =
tn_movie_budgets['release_date'].dt.day_name()

tn_movie_budgets['release_month'] =
tn_movie_budgets['release_date'].dt.month_name()
```

Merging columns to DataFrame from other tables:

We defined a function to merge genres from title_basics table to tn_movie_budgets table:

```
def get_genres(movie_title):
 try:
 genres =
title_basics.loc[title_basics['primary_title']==movie_title,
'genres'].values[0]
 except:
 genres = None
 return genres
```

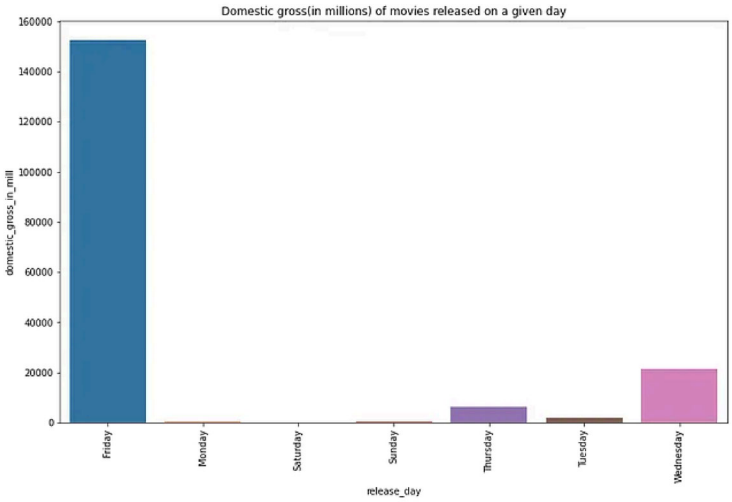
Similarly we merged runtime_minutes from title_basics and popularity from tmdb_movies to our combined DataFrame.

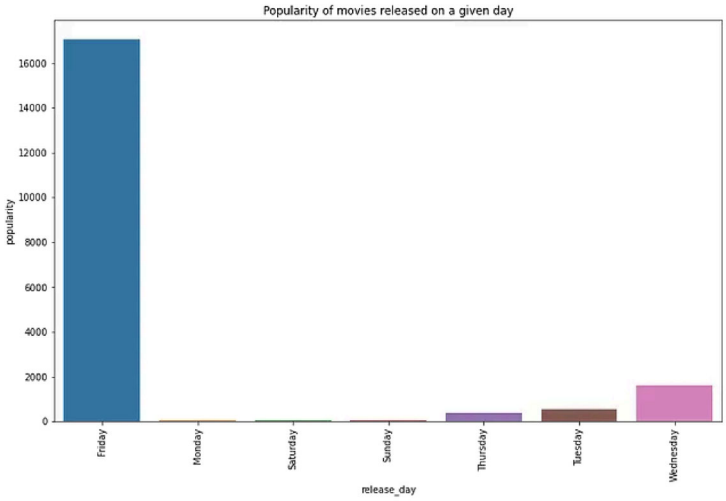
```
tn_movie_budgets =
tn_movie_budgets.loc[(tn_movie_budgets['release_date'] >= '2000-01-
01') & (tn_movie_budgets['year_released'] < 2020)]
```

New technology is coming out everyday. With the creation of netflix in 1997 and the increased internet speed and bandwith the way we consume film media has changed dramatically. While historical data is very important, for this business case we have decided to only go back to the year 2000. We have also decided to remove any films that were not released as of January 1, 2020 as it may not have complete data, or the film might not have been released yet.

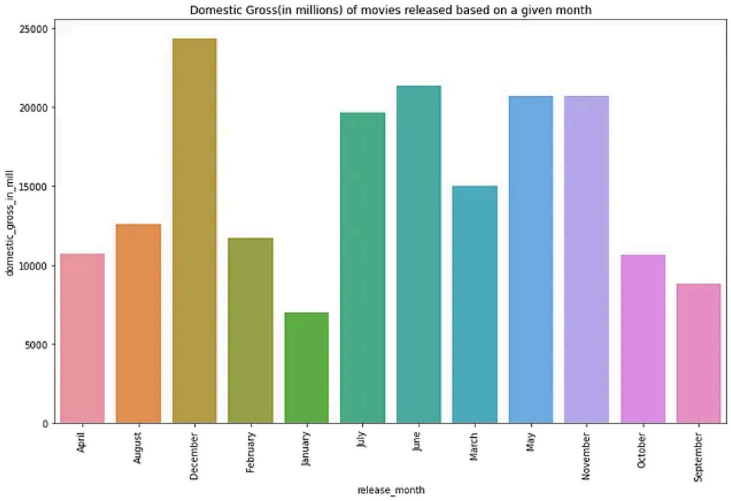
Our EDA analysis with questions:

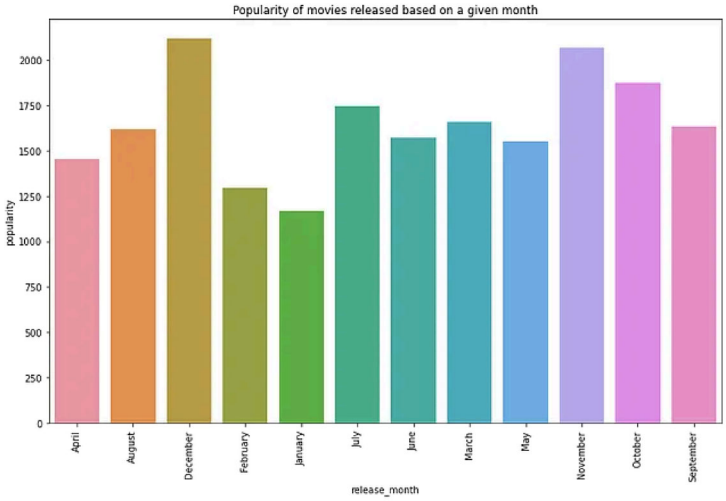
Q1) What is the best day/month to release movies vs popularity/domestic gross?





Here we did the sum of Domestic gross and Popularity vs Release Day

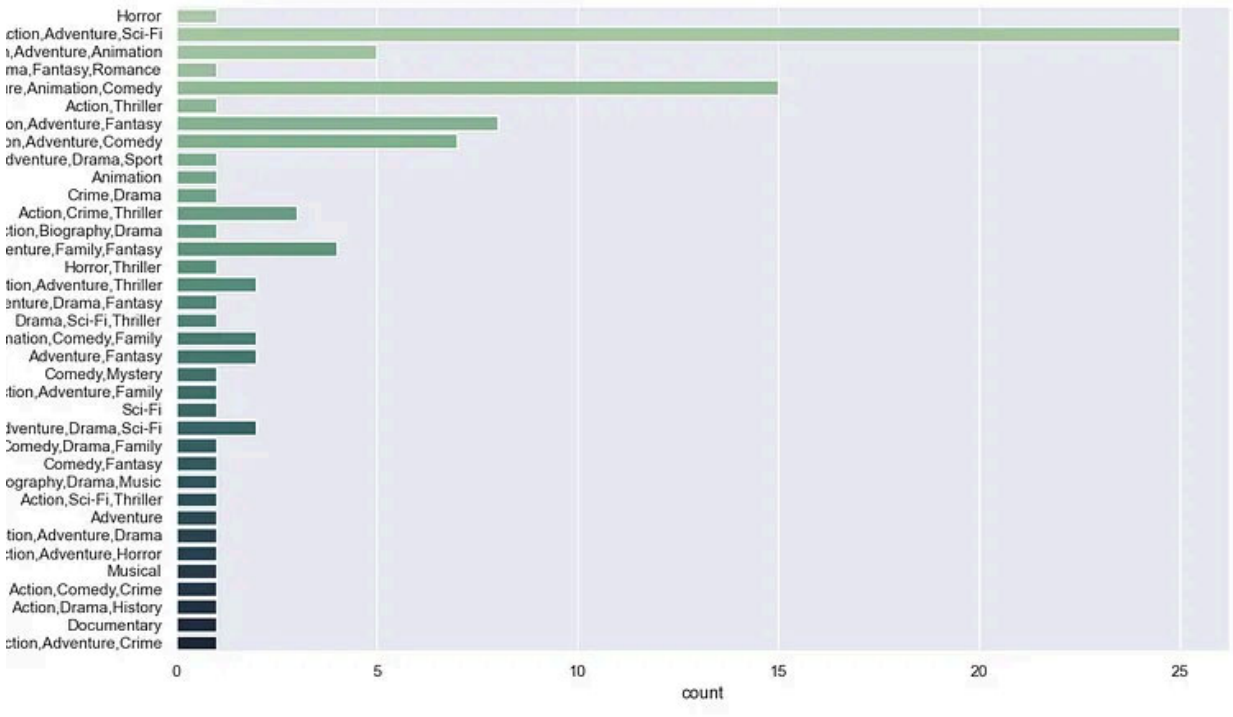




Sum of Domestic gross in millions and sum of Popularity vs Release Month

Conclusion: Friday is the best day to release a movie, in terms of both popularity and also domestic gross. **December** is the best month to release a movie, in terms of both popularity and domestic gross.

Q2) What is the most successful genre?



We determined that a bar chart was the best way to show our findings. These are the counts of movies in a genre.

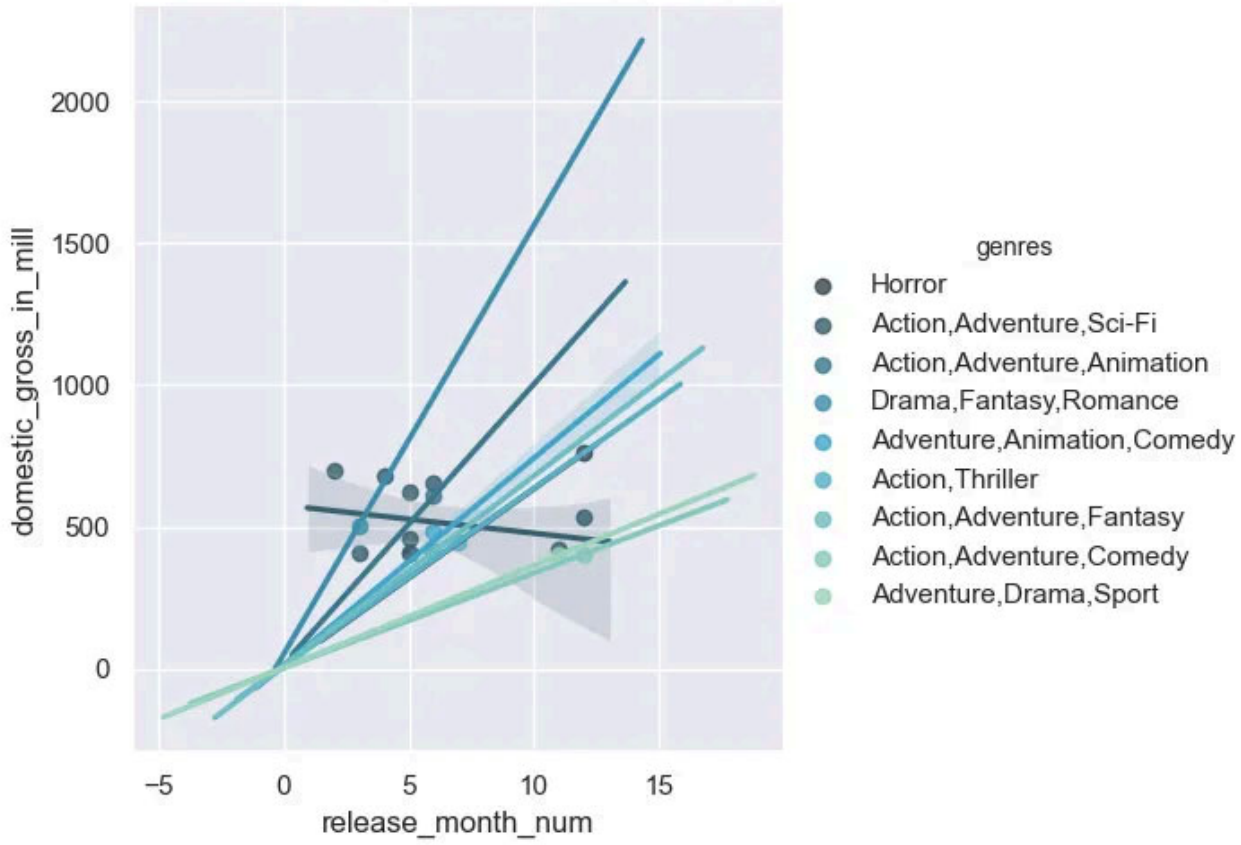
The next sub question to determine is if there is any correlation between the success of a particular genre being released at a specific time of the year. We extracted the top movies based on box office sales and compared that to release months and genre to see if there were any relationships. We decided to work with a smaller sampling size of 20 to get a better understanding of exactly how profitable the top movies of all time were. Did certain movies perform better being released in certain months?

A visualization was needed to actually see the correlation because it was hard to see our findings from the above dataset. We plotted our findings in a regression plot. This visual gave us a lot of interesting information.

1. 60% of the top movies of all time fell in the ‘Action, Adventure, Sci-Fi’ genre category.

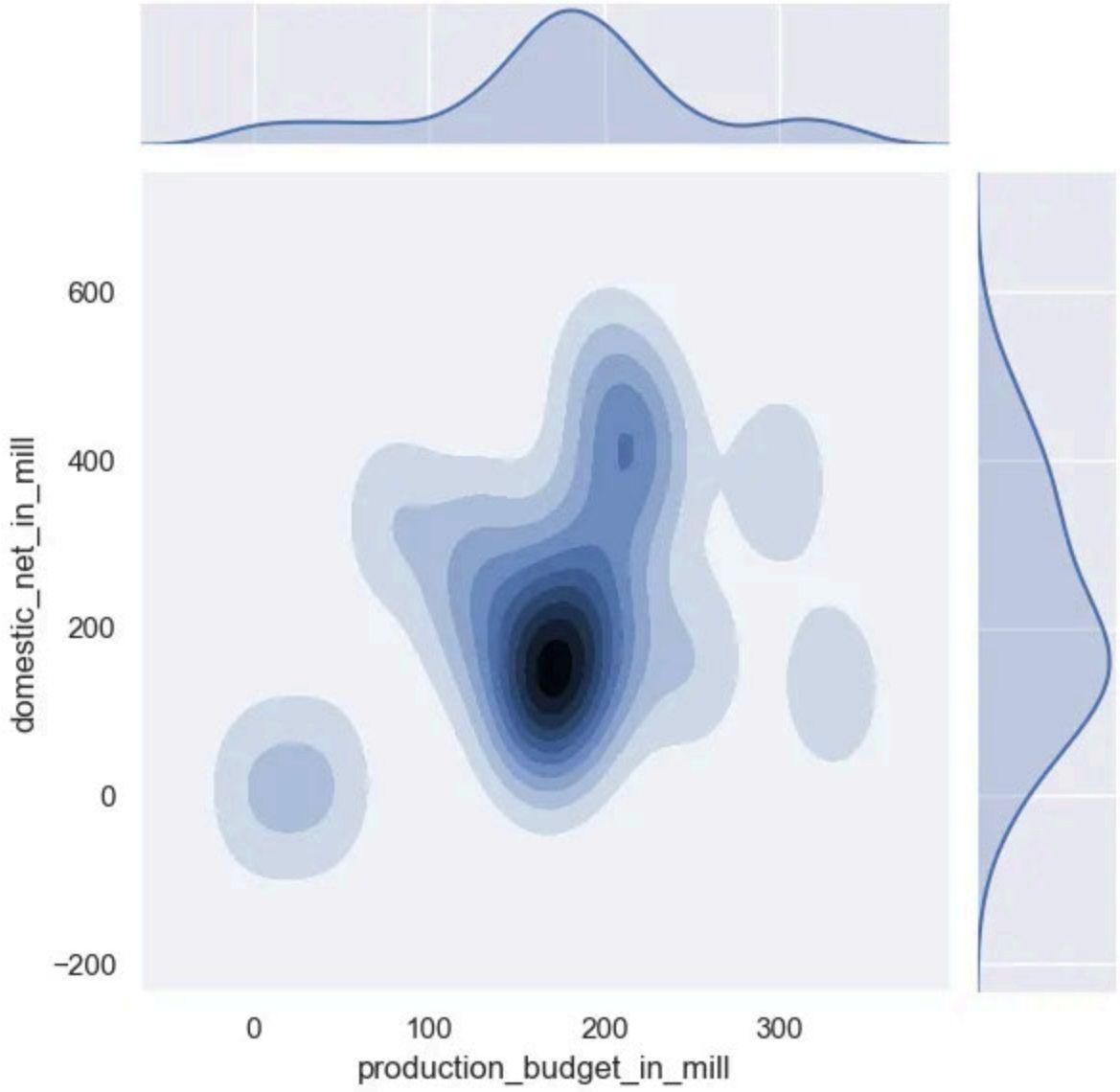
2. The majority of the ‘Action, Adventure, Sci-Fi’ movies were released in late Spring, Early-mid summer. The other popular seasons were Spring break and the holidays. But there was one other finding that was interesting as well.

3. One outlier ‘Action, Adventure, Sci-Fi’ movie which was the second highest grossing movie of all time and the most profitable movie of all time was released in February. That movie was ‘Black Panther’. It was released during Black History Month because it was a cultural Marvel movie. So this data shows that if you want to create an ‘Action, Adventure, Sci-Fi’ cultural movie, it will have better success being released during that culture’s heritage month.



The last question that we wanted to answer about genre was how costly is the overall opportunity cost of the most successful genre to produce, and if the

profits are worth it.



We then create a graph to visually display our findings to see the correlation of production budget vs. domestic net gross in the 'Action, Adventure, Sci-Fi' genre. We used a hexbin marginal plot to show that correlation. This shows that the most successful 'Action, Adventure, Sci-Fi' movies had a production budget of around 200 million dollars, and a domestic net gross between 200–500 million dollars.

Conclusion: The final recommendation to Microsoft pertaining to what genre would be the most profitable for them to make movies in would be 'Action, Adventure, Sci-Fi'. We concluded this finding from the following analysis discoveries:

1. Out of the top 100 domestic gross movies over the past 30 years, the genre ‘Action, Adventure, Sci-Fi’ made up the largest successful genre group in that data sample.

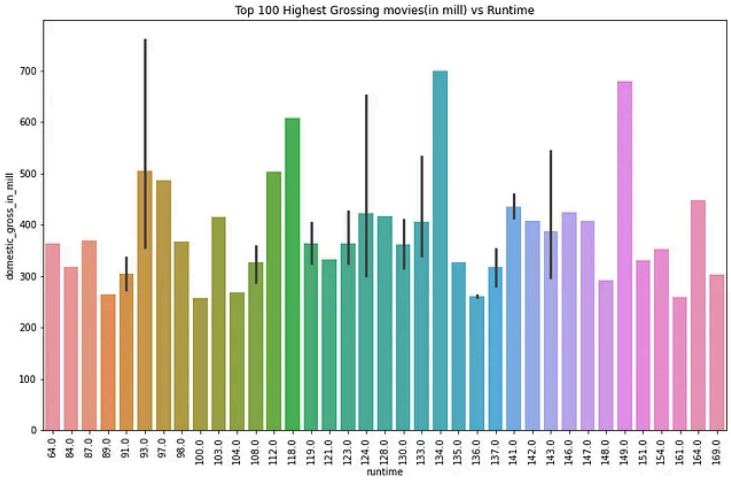
Our findings showed that releasing ‘Action, Adventure, Sci-Fi’ movies in late Spring/early-mid Summer, Spring Break week, during the holidays, and if it is a cultural movie, released during that culture’s Heritage month, all proved to be the most profitable times of the year to release that genre.

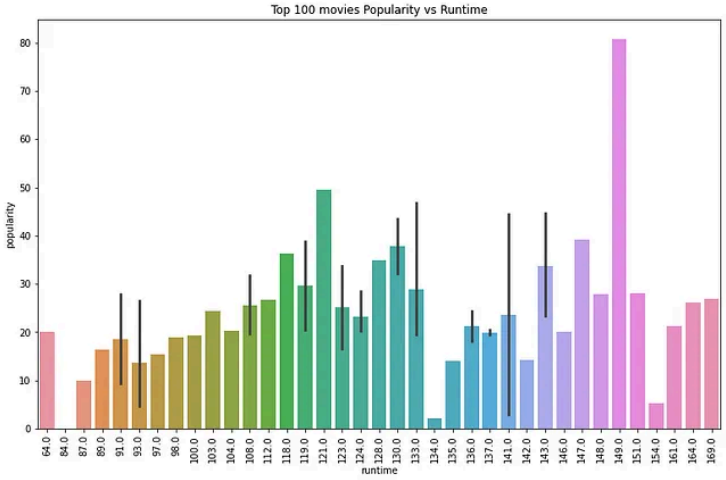
1. Sticking to a production budget of 200 million dollars while producing an ‘Action, Adventure, Sci-Fi’ movie has proven to be the key ingredient to high net profitability that can be forecasted to be between 200–500 million dollars.

Q3) Is there a relationship of run time of movies vs domestic gross and production budget?

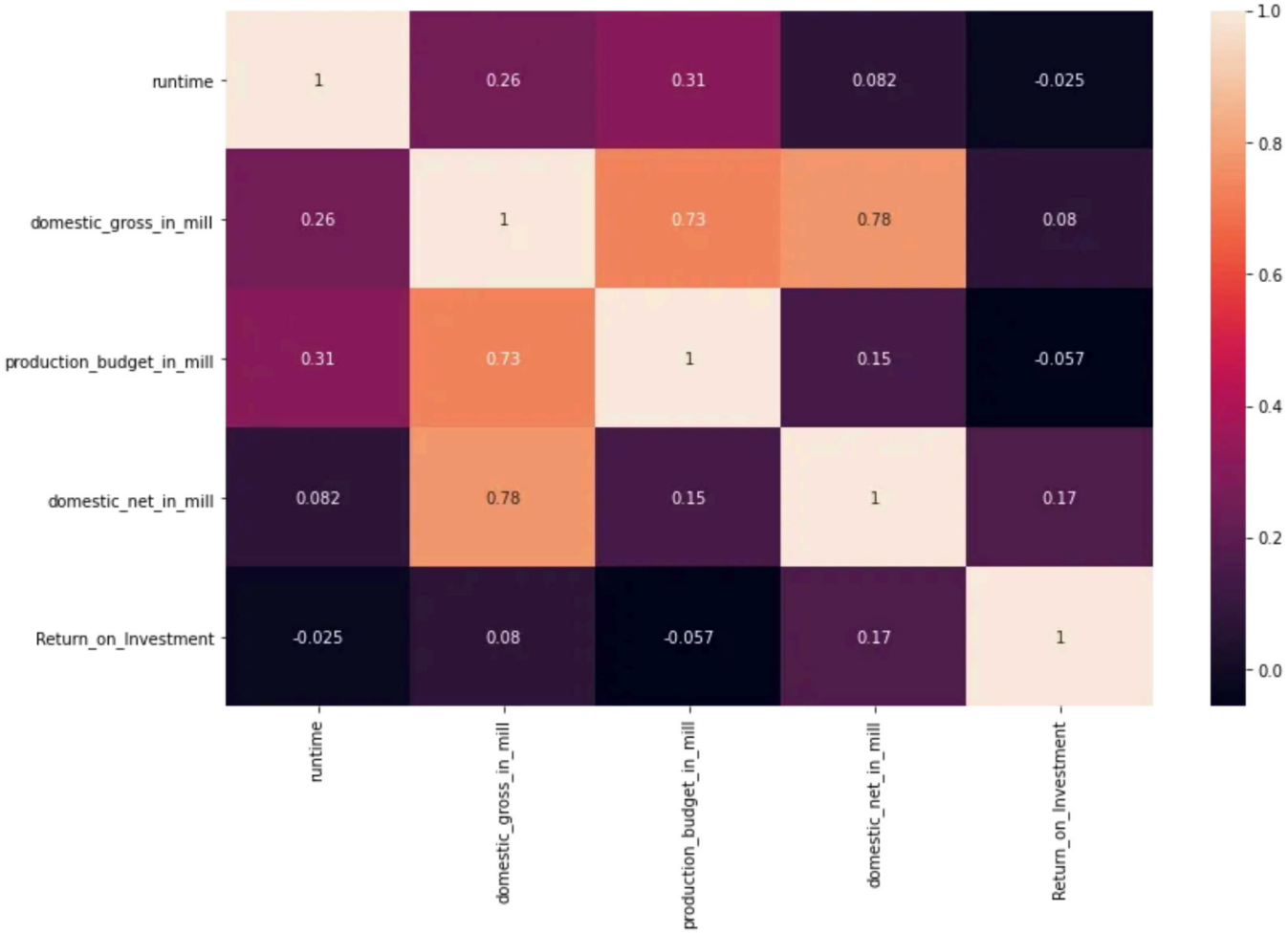
Here we create two dataframes: i) 100 Highest grossing movies and ii) 100 Lowest grossing movies.

```
top100 = runtime_df.nlargest(100,'domestic_gross')
bot100 = bot100.nsmallest(100,'domestic_gross')
```





Top 100 highest grossing movies and Popularity of those Top 100 movies vs Runtime of movies in minutes.



Examinining a relationship between runtime and production budget using a heatmap.

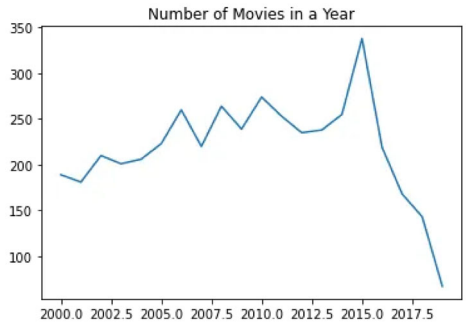
Conclusion: It seems that the **highest grossing** movies average to be around **123 minutes** while the **lowest grossing movies** average around the **95 minute** mark. The **most popular movies** in the **top 100 movies** have a runtime of **149 minutes**. If we take into consideration the most popular movies like Titanic, Avatar, All Marvel movies, this is what we would expect. As per our numbers, people normally like longer movies. Also as we can see in the heatmap that the correlation coefficient of runtime to production budget is positively correlated and is **0.31** which is moderately strong.

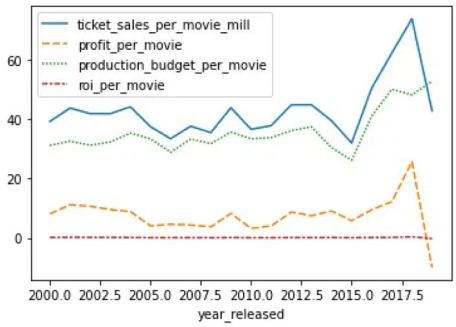
Q4) Can the film industry be a consistent profit center?

Here we will evaluate questions such as:

1. Are movies making more or less profit since 2000?
2. Are movies getting more expensive to make since 2000?
3. Does spending more money on production increase your chances of being profitable?

The intent of these questions are to provide an insight if the movie industry is thriving or failing. We want our clients to make the smartest decisions. We are looking to see if “an ounce of prevention equals a pound of cure”. If we can inform our clients that entering the movie business will not only be a waste of time, but also a waste of resources not beginning down that path is the smartest choice to make.





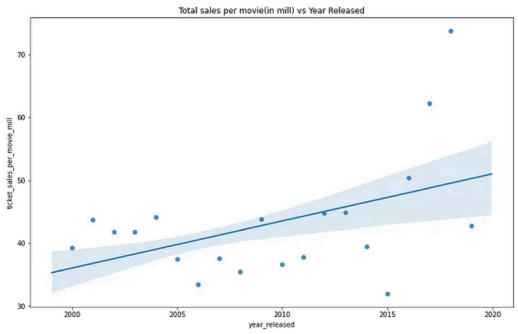
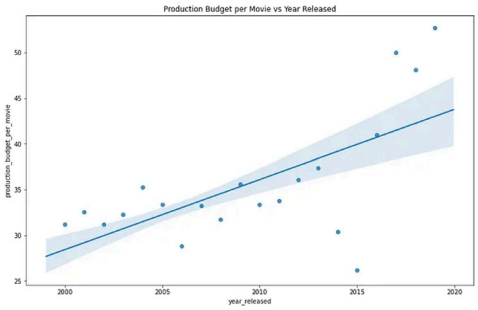
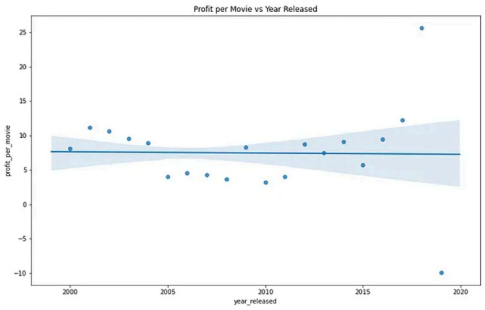


Fig 1, 2 ,3





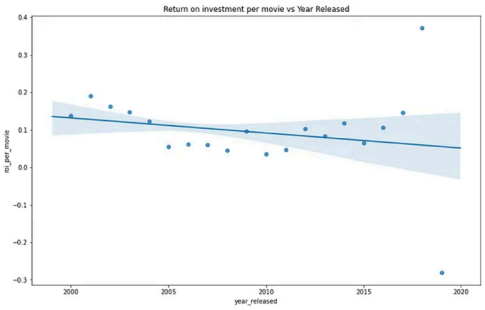


Fig 4, 5, 6

Figure 1 illustrates the number of movies that hit the movie theaters that year. As you can see the number of movies made increased rapidly from 2000–2015 and has since been in a steep downtrend. This can be for many reasons. The first could be imperfect data — perhaps this isn't all the movies that hit box office. The second — and more probable reason — is that movies aren't hitting theaters anymore and going straight to in-demand services like Netflix.

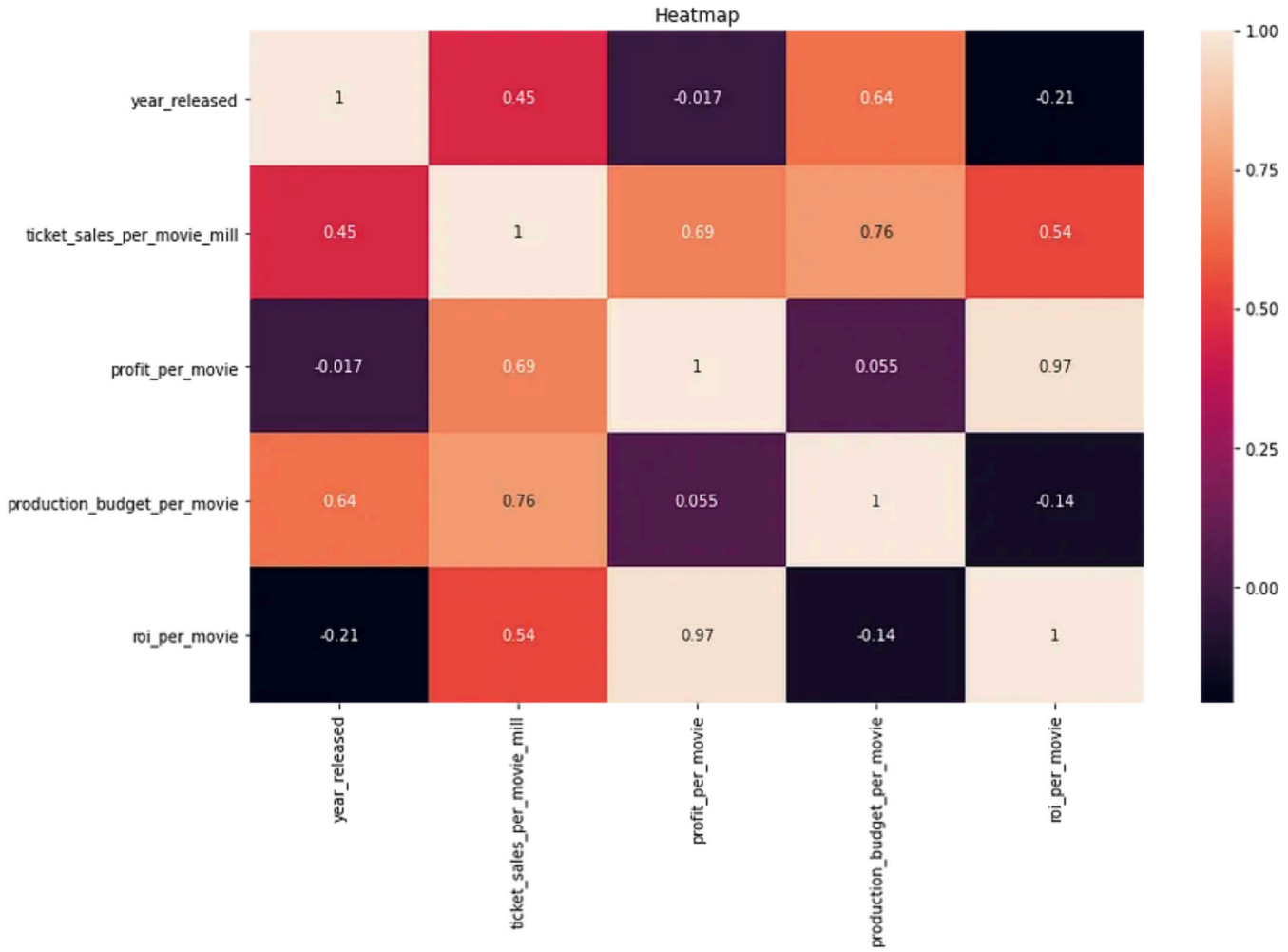
Figure 2 shows the normalized to number of movies data over the desired years. You can see in this figure how all of the variables interact with each other. The ticket sales for example shot up from 2015 to 2018 and then has falled off steeply. While ticket sales have increased the profit per movie takes a very hard drop and actually goes negative around 2019. This is interesting because you can see the production budget spike during this time and ticket sales to fall — which is the direct inverse of what you want to see happen.

Figure 3 shows a positive sloped line from 2000 to 2019 which means that over this time period ticket sales have been increasing. This is good if you are in the movie industry as you want to see an appetite for the product and steady growth.

Figure 4 shows the production budget over time and another positively sloped line. While not on the same exact scale, it appears that the production budget line of best fit has a much steeper slope that the line in figure 2.0. This means production budgets are increasing faster than the growth in ticket sales — not a good sign. This can also be seen in Figure 2 where the green line (production budget) actually surpasses the blue line (ticket sales).

Figure 5 shows the profit per movie over time. This line of best fit is essentially flat. This means that over time the movie industry as a whole is not making any more money per movie today than it was 20 years ago. If there is one conclusion that can be drawn from this is predictability is often a good thing as it means you will have accurate information to make informed business decisions.

Figure 6 shows a decline in ROI per movie over time. This is not good. This means that it is getting harder every year to risk investing money in making the movies.



Correlation Matrix — heatmap

Interesting Finds:

ROI per movie & Year Released = -0.21

This was also seen in the regression plot. This means that as the years increase the ROI per movie is decreasing at about a 1:5 ratio.

Production Budget & ROI = -0.14

It was not unexpected to see that the production budget to ROI relationship is negative, but it was a shock to see just how little the negative relationship is. This is saying that there is a weak negative relationship. My intuition was that obviously the higher the production budget the lower the ROI because

the production budget is a huge cost. However, it seems that the higher the production budget might also generate more ticket sales.

Production Budget & Profit = 0.05

Very similar to the production budget & ROI, this was expected to have a negative relationship. Instead we found that production budget and profit have 0 correlation.

1. Are ticket sales growing since 2000?

Figure 2 answers this question pretty plainly looking at the blue line representing ticket sales. Ticket sales shows an unencumbered picture of the demand for the movies created that year. From 2000 to 2015 ticket sales were not growing until a crazy spike going around 2018. Ticket sales after 2019 have declined dramatically after an impressive 2018 back to the same levels they were at the entire 2 decades we evaluated. *There is no measurable growth in demand from 2000 to 2019.*

2. Are movies making more or less profit since 2000?

This question is easily answerable looking at Figure 2, Figure 4, and Figure 7. Looking at Figure 4 we see the relationship of profit per movie over time with a line of best fit plotted showing the trend. The slope is flat to barely negative. Figure 7 shows the correlation coefficient between profit per movie and the year it was released is -0.17. This also shows there is a negligible negative correlation to the year it was released and the profit of the movie. *This shows that movies are not becoming more profitable over time.*

3. Are movies getting more expensive to make since 2000?

Similarly to question 2 lets look at Figure 3 and Figure 7. Figure 3 shows production budgets per movie over time. The line of best fit is showing a steep ascent which means that productions budgets have been growing as

the years passed. Figure 7 shows the correlation coefficient between these two variables is 0.64. This is proof of a strong positive correlation of these values as the years go on the production budgets also increase. ***This shows that movies are becoming increasingly expensive to make over time.***

4. Does a larger production budget increase your chances of producing a profitable movie?

Figure 7 measures the correlation between multiple variables including production budget and profit per movie and ROI. The values for production budget vs profit is 0.055 and production budget vs ROI is -0.14. This means that while production budget does have a little effect on ticket sales the increased cost in the budget is greater therefore hurting your return metrics. ***This shows that a larger production budget has no change to your profit and actually will hurt your return metrics.***

Summary

In summary, on this data I would not recommend entering the movie industry as an inexperienced content creator. The majority of movies are not doing well and there is an extremely wide range in possible outcomes. That being said, I believe this data is leaving out a major part of the revenue stream for movies in the on demand market. If we had data on the income generated from on demand services such as Netflix it may shed a much more positive light on becoming a content creator.

Data Science | Exploratory Data Analysis | Movies

Written by Saif Kasmani

3 Followers

Follow

More from Saif Kasmani

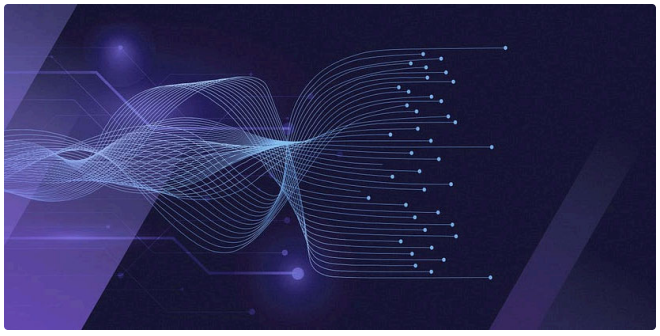
Characteristics	Matplotlib	Seaborn
Use Cases	Matplotlib plots various graphs using Pandas and Numpy	Seaborn is the extended version of Matplotlib which uses Matplotlib along with Numpy and Pandas for plotting graphs
Complexity of Syntax	It uses comparatively complex and lengthy syntax.	It uses comparatively simple syntax which is easier to learn and understand.
Multiple figures	Matplotlib has multiple figures can be opened	Seaborn automates the creation of multiple figures which sometimes leads to out of memory issues
Flexibility	Matplotlib is highly customizable and powerful.	Seaborn avoids a ton of boilerplate by providing default themes which are commonly used.

Saif Kasmani

Matplotlib vs Seaborn library

Data Visualization tools are of great importance in the analytics industry as they...

2 min read · Dec 9, 2020

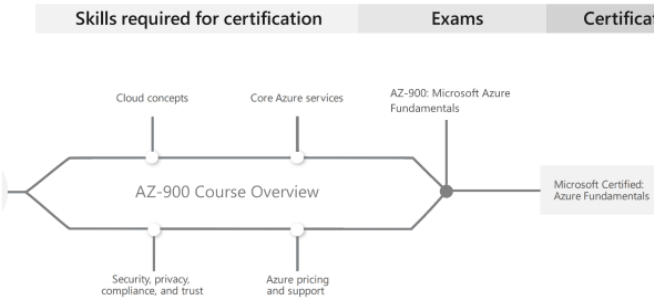


Saif Kasmani

Time Series Analysis: Time-Series Forecasting Machine learning...

While all the numerous advanced tools and techniques are employed for data analysis...

7 min read · Nov 26, 2020



Saif Kasmani

Telecom Churn Analysis

Project: Predicting churn for a telecom company so it can can effectively focus a...

7 min read · Jul 15, 2020

Saif Kasmani

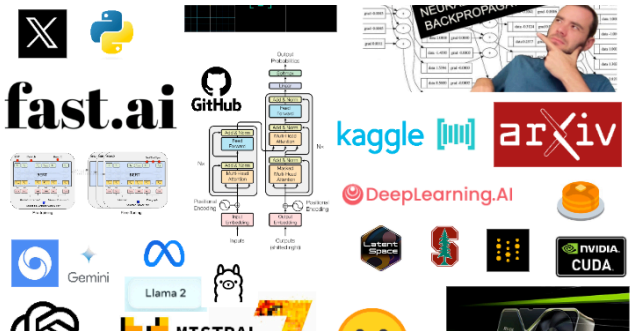
Azure Data Fundamentals

Summary of the basics of Azure Data Fundamentals and what to expect in the...

2 min read · Dec 20, 2020

See all from Saif Kasmani

Recommended from Medium



Benedict Neo in bitgrit Data Science Publication

Roadmap to Learn AI in 2024

A free curriculum for hackers and programmers to learn AI

11 min read · 2 days ago

5.7K 71

Data Scian by Imad Adrees

Best Portfolio Projects for Data Science

“How can I showcase my data skills to the world?” you may be asking. Fear not, for the...

5 min read · Sep 19, 2023

1.2K 8

Lists

Predictive Modeling w/ Python

20 stories · 998 saves

Practical Guides to Machine Learning

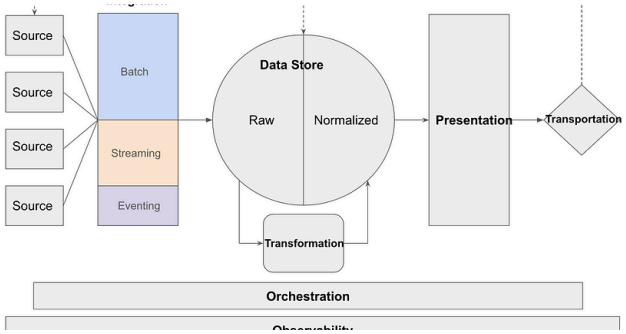
10 stories · 1189 saves

Coding & Development

11 stories · 501 saves

ChatGPT prompts

47 stories · 1258 saves



Nathan Rosidi

Data Science in 2024—What Has Changed

What has changed in the data science landscape, and what are the challenges of th...

4 min read · Jan 29, 2024

917 16

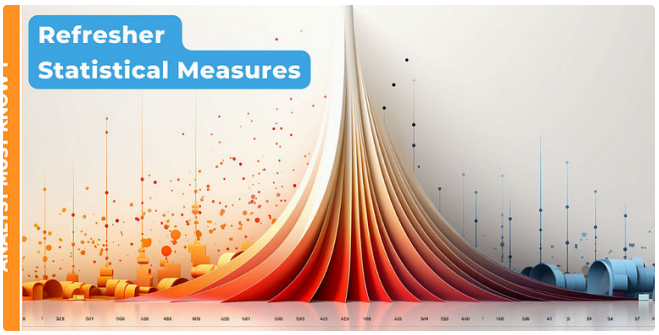
Dave Melillo in Towards Data Science

Building a Data Platform in 2024

How to build a modern, scalable data platform to power your analytics and data science...

9 min read · Feb 6, 2024

2K 31



Prof. Frenzel

Statistical Measures Every Analyst Must Know—Part1

Measures of Central Tendency, Variability, Quartiles, Z-Scores, and as always:...

11 min read · Feb 4, 2024

769 5

Vinay Kumar Moluguri

10 Data Analysis Projects to Land Your First Analytics Job

Looking to break into data analytics but need projects to showcase skills to employers?...

· 4 min read · Oct 1, 2023

200 3

See more recommendations