Data Analytics Portfolio

GERALD YUEN
2022

About Me

My name is Gerald and I graduated from Cal State Fullertion, California. My goal is to become a data analyst in Los Angeles area. I value hard work, fairness, and honesty. I have been in the service industry for over 15 years dealing with customers and colleagues which makes me a good problem solver and a great team player.

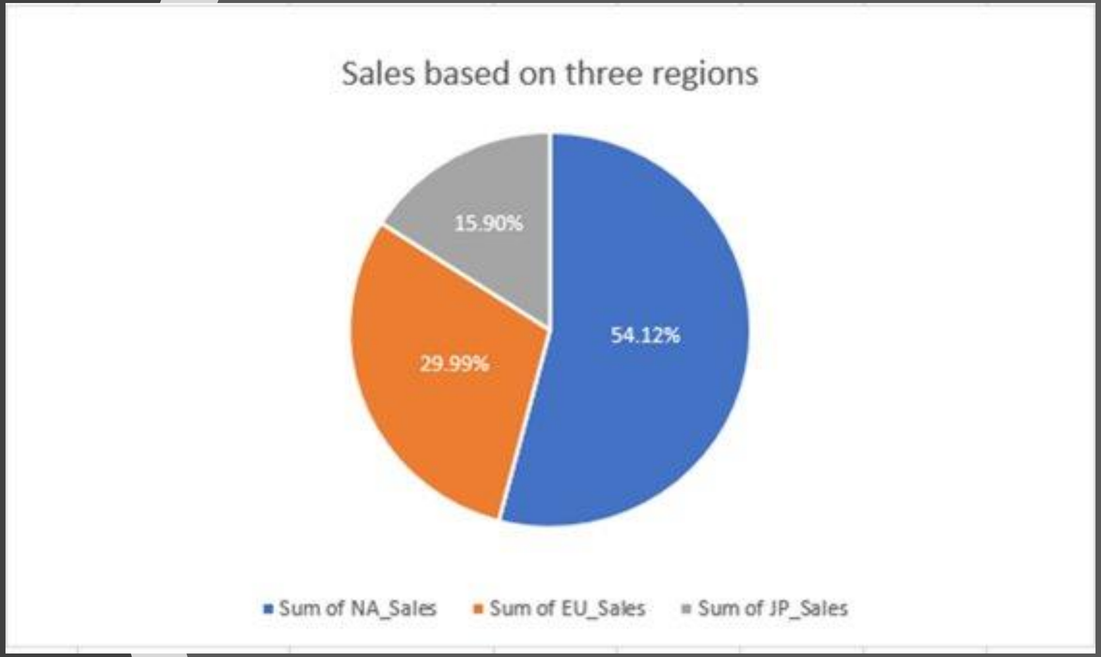
Table of Contents

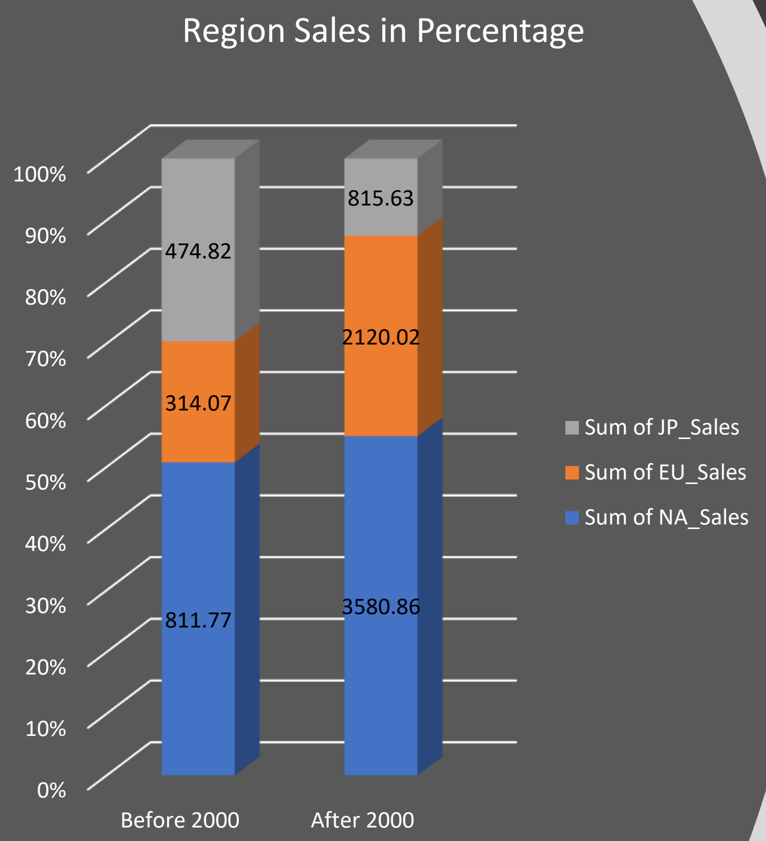
1. GAMECO --- Excel
2. Influenza Season Analysis ----------------- Tableau
3. Rockbuster --------------------------------------- SQL
4. Instacart --------------------------------------- Python
5. Pig E. Bank ------ Data Ethics and Data Analytics
6. IMDb ------------------------- End of course Project

1) GAMECO

A video game company wants to use data to inform the development of new games. This scenario involve answer some questions regarding for which region is more popular and the sales from 1980 to 2016. The objective is to find out what new games to sell next year.

GAMECO

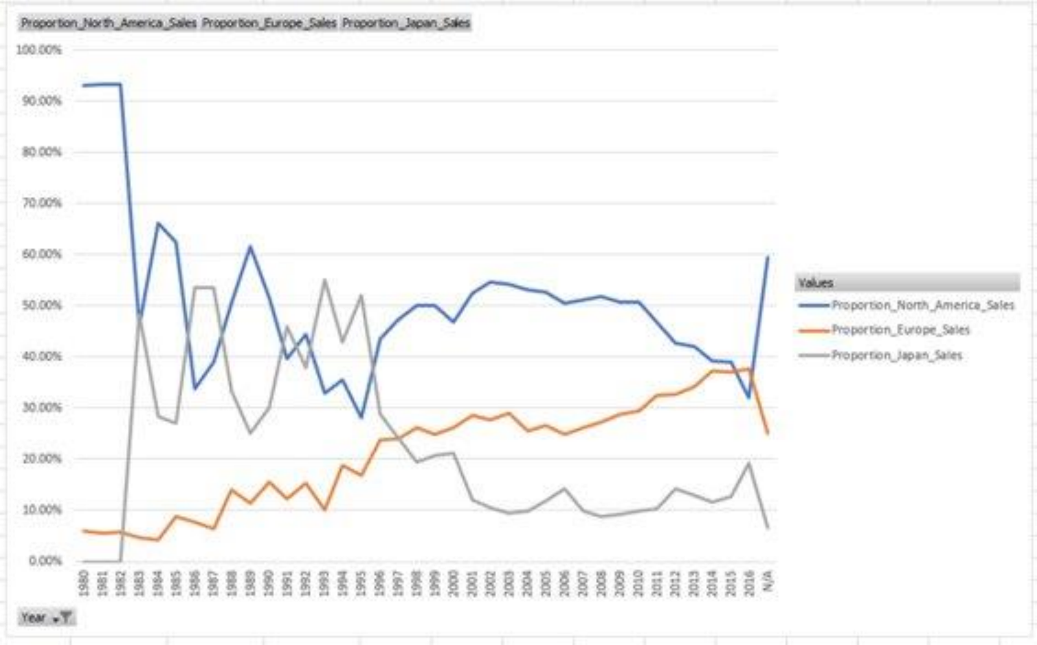




I use Excel to analyze the project based on the three regions, North America Sales, Europe Sales, and Japan Sales. North America sales is the best of the three regions and Europe has sufficiently increased sales after 2000.

GAMECO

From the line chart, it shows North America sales is the highest, but Europe sales is the only one that increases over the year.



GAMECO
Result

- We need to allocate more resources to North America region to increase sales and investigate Europe region to see what makes Europe sales increase.

- Excel presentation
- Excel clean dataset

- Skills in Excel

- Data cleaning
- Sorting
- Pivot tables
- Graphs

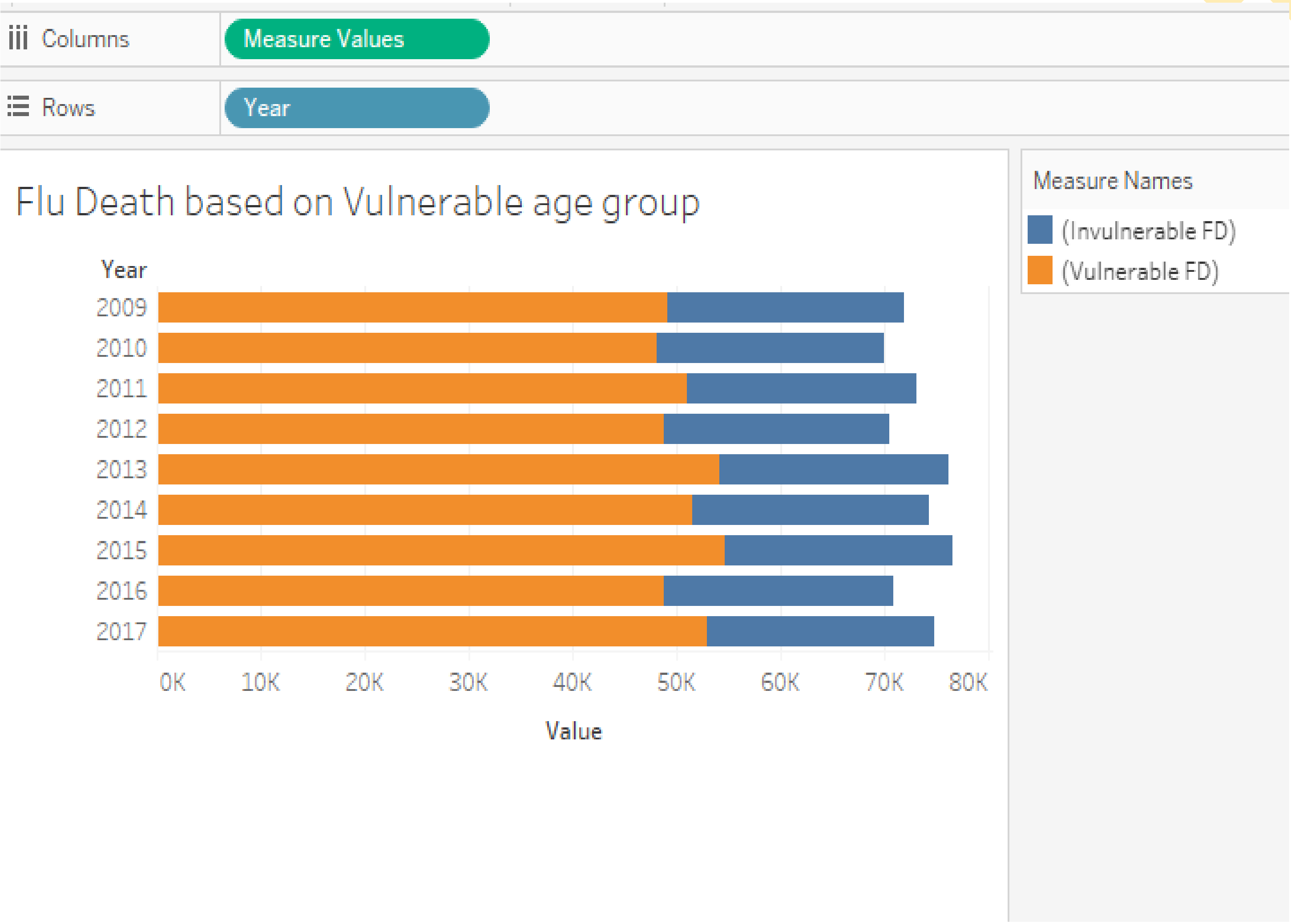
2) Influenza Season Analysis

A US Influenza data provided by the CDC to prepare the upcoming Influenza season to place medical staff in the states that need more assistance.

Influenza Season Analysis

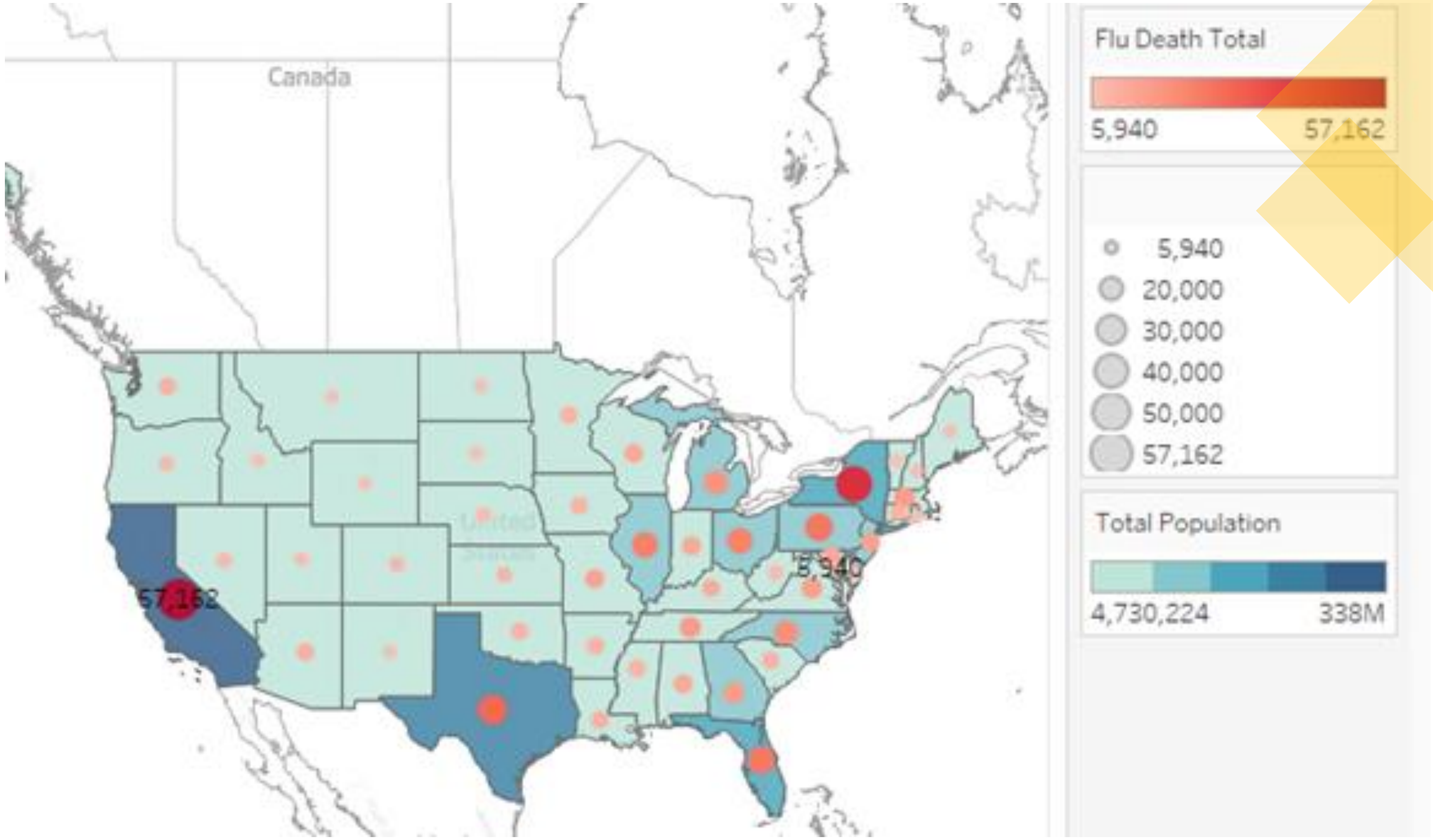
- I cleaned up the data and combined it so they can be spilt between 55+ years old (vulnerable age) and 54 and under (invulnerable age). The result shows there are more flu death in the vulnerable age group than the invulnerable age group.

- Data Integration

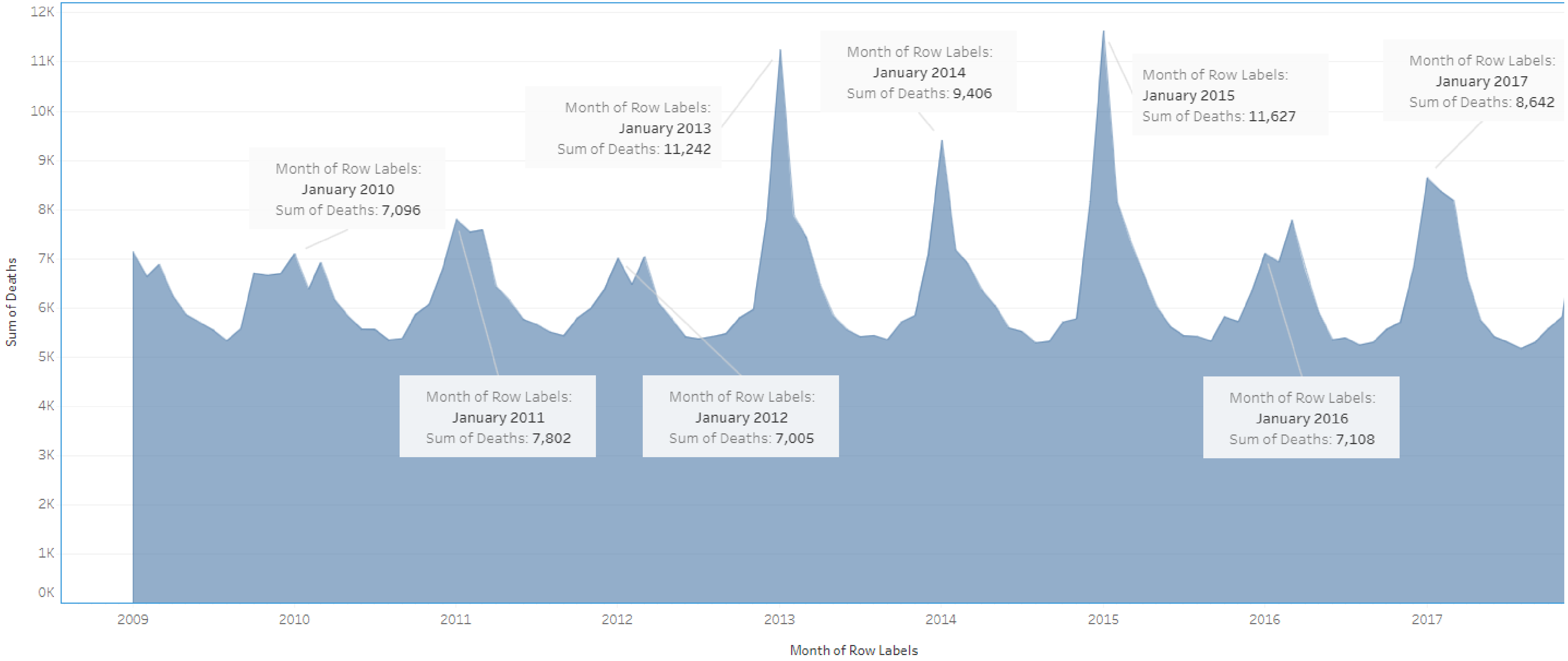


Influenza Season Analysis

I created a map with Tableau about the population in each state while showing the how many flu death in each state. We are also filter the map based on the years. The map shows California with the highest death total and also with the highest population.



Influenza Season Analysis



- The line graph shows the highest death rate are mostly in January, with the highest in 2015.

Influenza Season Analysis Result

- Five of the highest Flu death States are California, New York, Texas, and Pennsylvania. We need to pay extra attention to these states and allocate more resources.
- Peak of the Flu season is in Januarys so more medical staff should be scheduled.

Skills in Tableau

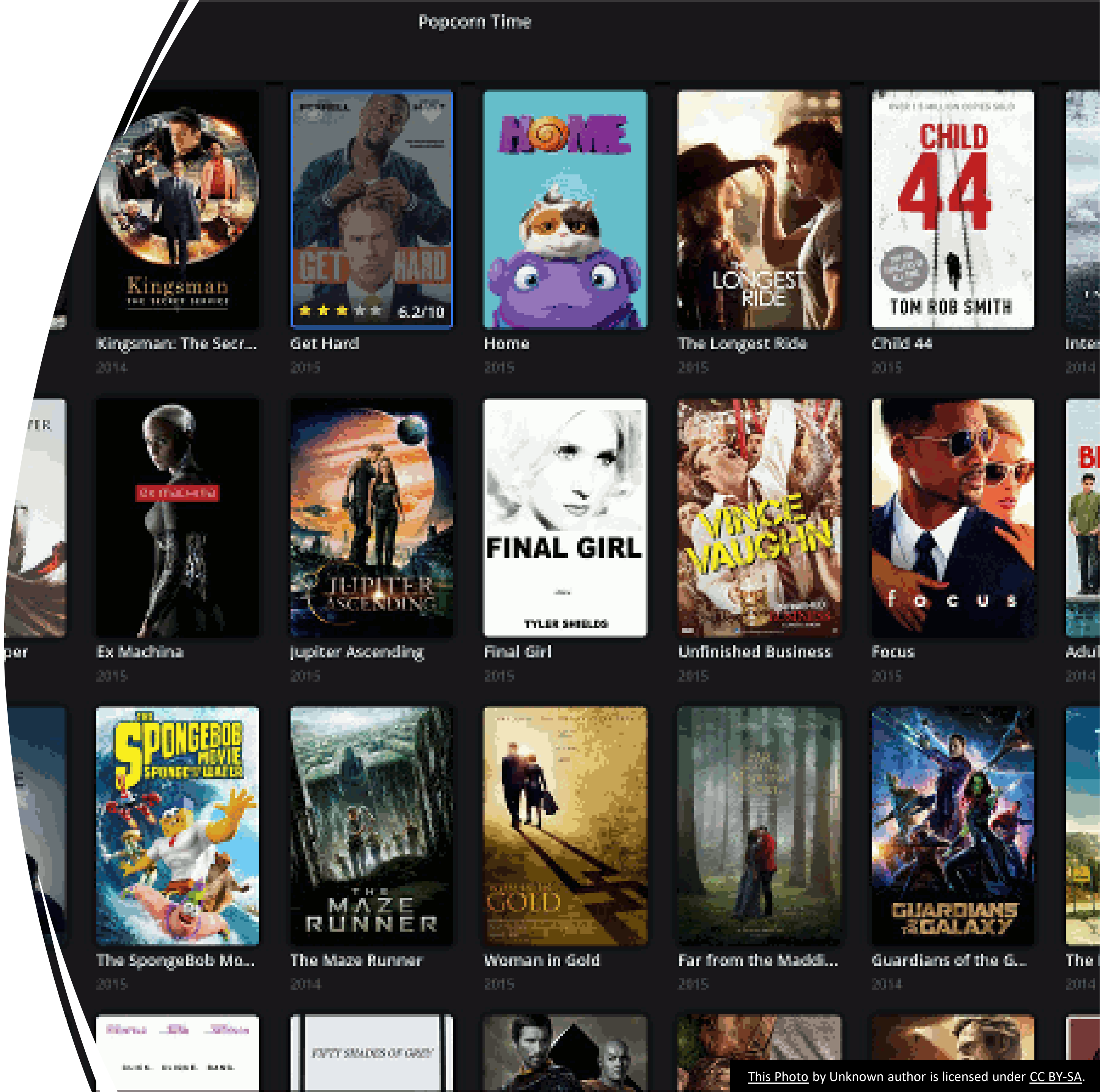
- Data cleaning
- Bar/pie/treemaps
- Dual maps with symbol map and regular map
- Linking the maps together for more interactive presentation.

Tableau presentation

3) Rockbuster Stealth LLC

Rockbuster Stealth LLC is a movie rental company that have stores worldwide and is planning to sell online video rental service.

The company wants to know which country the customers are from and which movies are profitable. They also want to know if the sales figures vary between geographic regions.



Rockbuster Stealth LLC

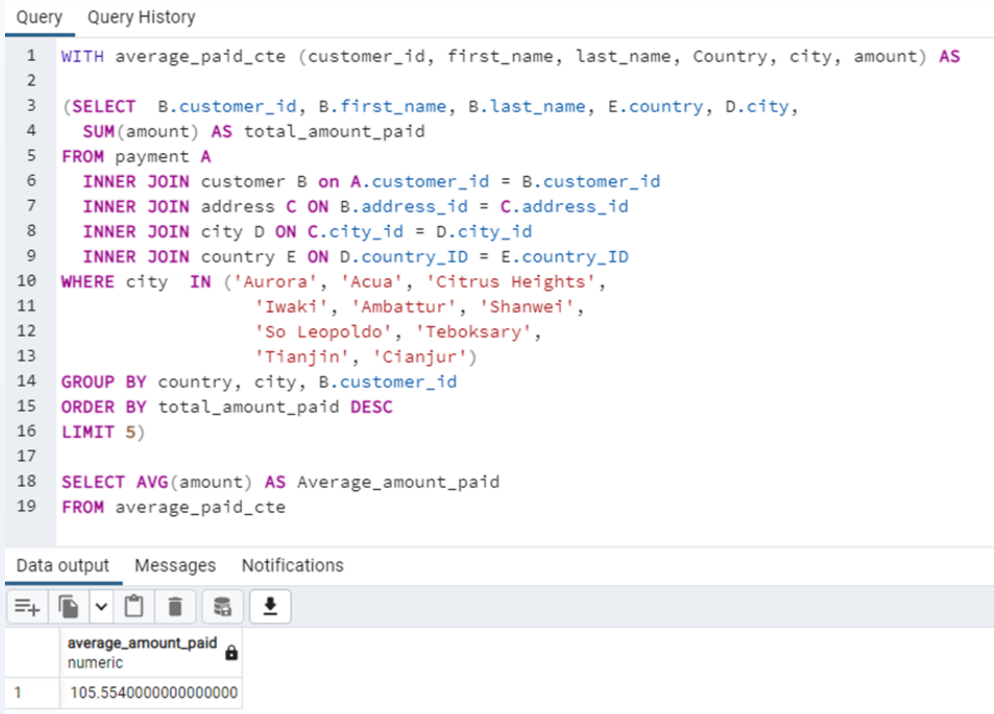
With the SQL software, I can sort and put tables together to gather data more efficiently .

The country with the most revenue is India with $6034.78 and 1422 customers.

Country	Customer count	Total Revenue
India	1422	$6034.78
China	1297	$5251.03
United States	869	$3685.31
Japan	749	$3122.51
Mexico	718	$2984.82
Brazil	681	$2919.19
Russian Federation	638	$2765.62
Philippines	530	$2219.70
Turkey	351	$1498.49
Indonesia	331	$1352.69

Rockbuster Stealth LLC

With the SQL software, I can sort and put tables together to gather data more efficiently .
I put five different tables together to find out the average amount paid for each customer is $105.56.

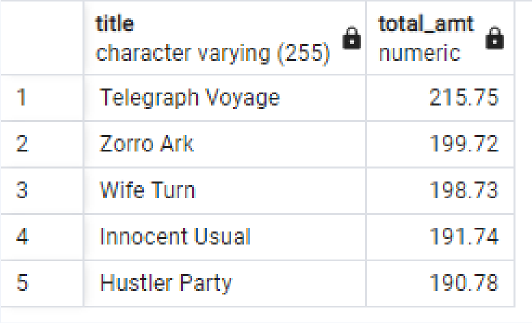


```sql
WITH average_paid_cte (customer_id, first_name, last_name, Country, city, amount) AS

(SELECT  B.customer_id, B.first_name, B.last_name, E.country, D.city,
  SUM(amount) AS total_amount_paid
FROM payment A
  INNER JOIN customer B on A.customer_id = B.customer_id
  INNER JOIN address C ON B.address_id = C.address_id
  INNER JOIN city D ON C.city_id = D.city_id
  INNER JOIN country E ON D.country_ID = E.country_ID
WHERE city  IN ('Aurora', 'Acua', 'Citrus Heights',
                 'Iwaki', 'Ambattur', 'Shanwei',
                 'So Leopoldo', 'Teboksary',
                 'Tianjin', 'Cianjur')
GROUP BY country, city, B.customer_id
ORDER BY total_amount_paid DESC
LIMIT 5)

SELECT AVG(amount) AS Average_amount_paid
FROM average_paid_cte
```

```
   average_amount_paid
   numeric
1  105.5540000000000000
```

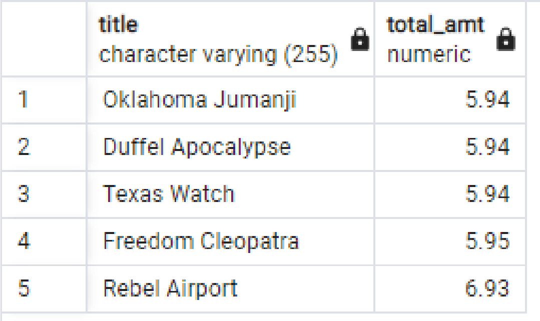
Rockbuster Stealth LLC

Five top movies



	title character varying (255)	total_amt numeric
1	Telegraph Voyage	215.75
2	Zorro Ark	199.72
3	Wife Turn	198.73
4	Innocent Usual	191.74
5	Hustler Party	190.78

Five bottom movies



	title character varying (255)	total_amt numeric
1	Oklahoma Jumanji	5.94
2	Duffel Apocalypse	5.94
3	Texas Watch	5.94
4	Freedom Cleopatra	5.95
5	Rebel Airport	6.93

The top movie with the most revenue is “Telegraph Voyage” for $215.75, and the least revenue gain movie are “Oklahoma Jumanji”, “Duffel Apocalypse”, and “Texas Watch”, which are all $5.94.

Rockbuster Stealth LLC Result

I suggest to retire the least profitable movies and invest in movies that are more popular. We can also advertise in the top ten regions to generate more profit. We can evaluate the process by checking the average amount paid by customers.

Data Dictionary
Tableau graphs
Powerpoint Presentation
Github

Skills

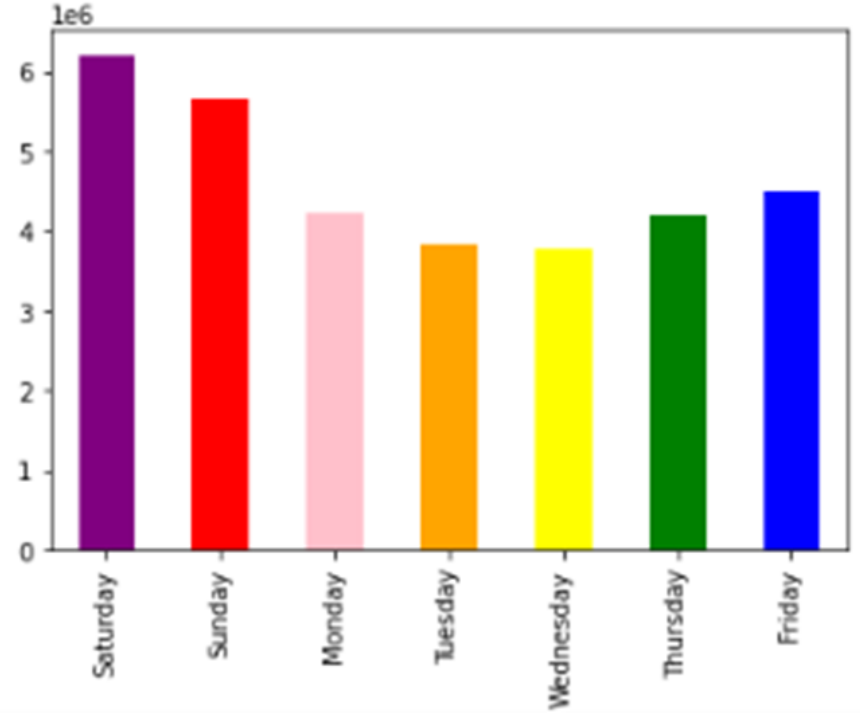
- SQL
  - Filtering data
  - Joining tables of data
  - Common Table Expressions
- Tableau
  - Maps and graphs connected as filters

4) Instacart

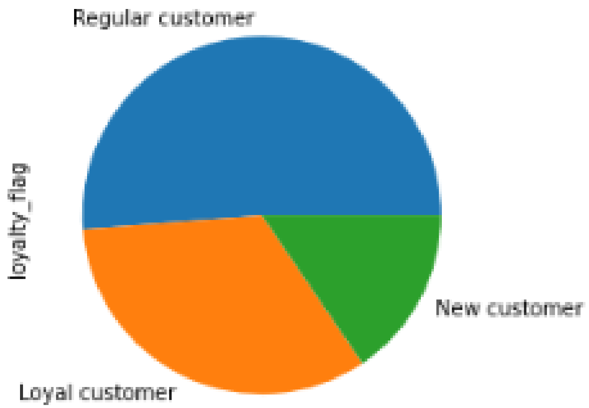
- Instacart is an online grocery store that operate through an app. The company wants to use the initial data to gather more information of sales patterns and have better marketing strategy.

Instacart

With Python, I can compute using a huge data base and find out most customers buy products on Sundays.

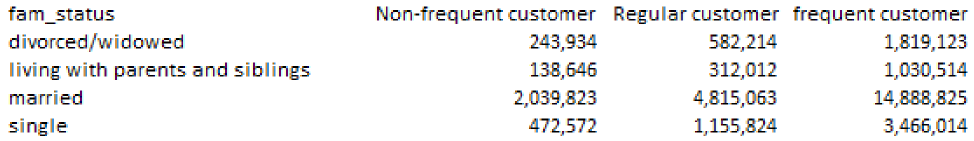


Instacart

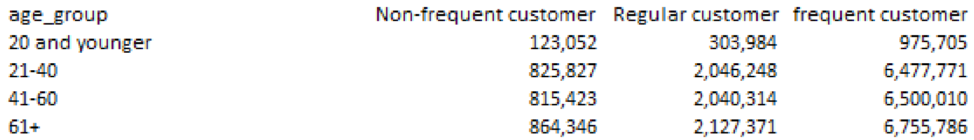


Over 50% of the customers are regular customers, and after that are loyal customers and the rest are new customers.

I also group data together and find out most frequent customer are married couples who are over 21 years old.



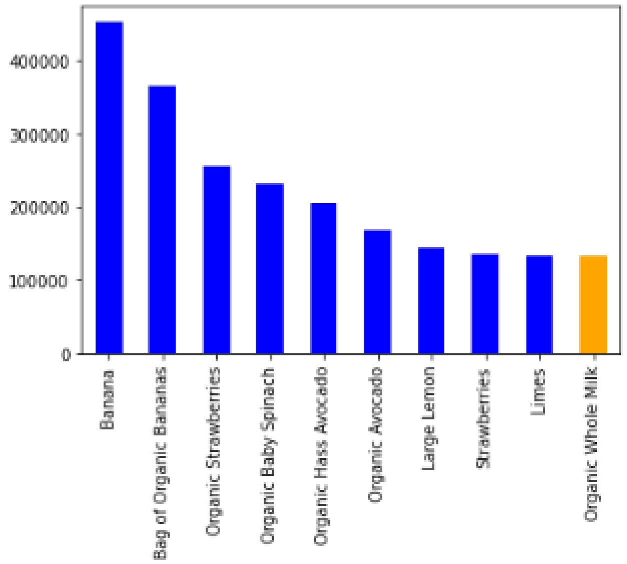
fam_status	Non-frequent customer	Regular customer	frequent customer
divorced/widowed	243,934	582,214	1,819,123
living with parents and siblings	138,646	312,012	1,030,514
married	2,039,823	4,815,063	14,888,825
single	472,572	1,155,824	3,466,014

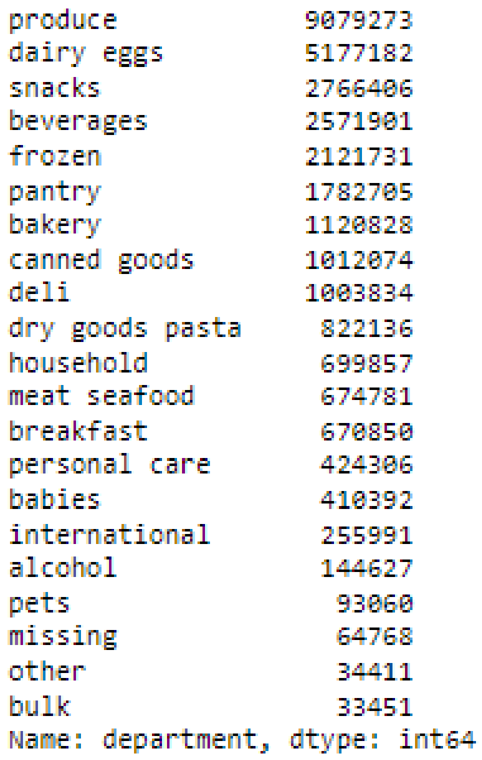


age_group	Non-frequent customer	Regular customer	frequent customer
20 and younger	123,052	303,984	975,705
21-40	825,827	2,046,248	6,477,771
41-60	815,423	2,040,314	6,500,010
61+	864,346	2,127,371	6,755,786

Instacart

I sort out the Product key and found out the most product sold are in the “produce” section. Top 9 items are in the “produce” section and the 10th item is in the “dairy eggs” section.





```
produce            9079273
dairy eggs         5177182
snacks             2766406
beverages          2571901
frozen             2121731
pantry             1782705
bakery             1120828
canned goods       1012074
deli               1003834
dry goods pasta     822136
household           699857
meat seafood        674781
breakfast           670850
personal care       424306
babies              410392
international       255991
alcohol             144627
pets                 93060
missing              64768
other                34411
bulk                 33451
Name: department, dtype: int64
```

Instacart – Result

- We can target married couples who are over 21 years old for marketing and focus on the "produce" section and negotiate for better pricing.
- We can have more incentive program for regular customers to buy more to boost them up to loyal customers.

- Github
- Excel presentation

Skill

- Python
  - Data wrangling
  - Combine/export data
  - Grouping/aggregating data
  - Data visualization
- Excel
  - Copy and paste to excel with df.to_clipboard()

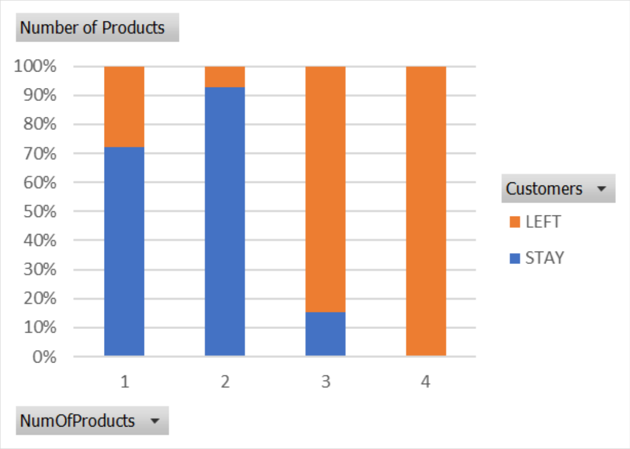
5) Pig E. Bank

Pig E. Bank is a global financial service company and I'm acted as a junior data analyst. This project will be using data ethic and applied Analytics.

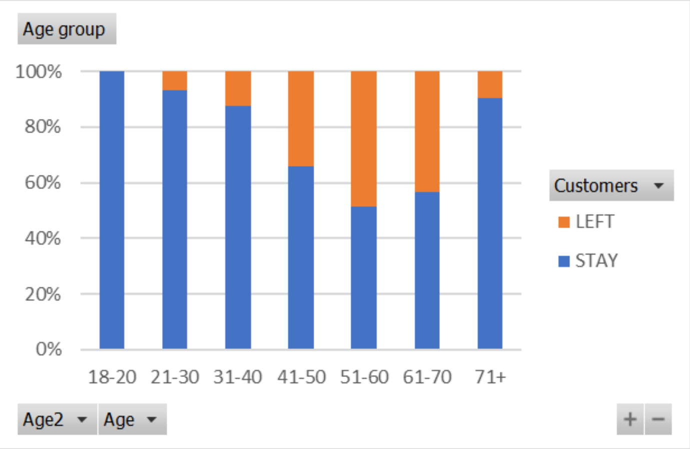
The company is trying to figure out why customers leave. Also, the bank is setting up an anti-money laundering plan.

In part of the decision tree, I set up charts to identity the reason why customers left and some of the reason are high number of products the customer have and the age rage between 51-70.

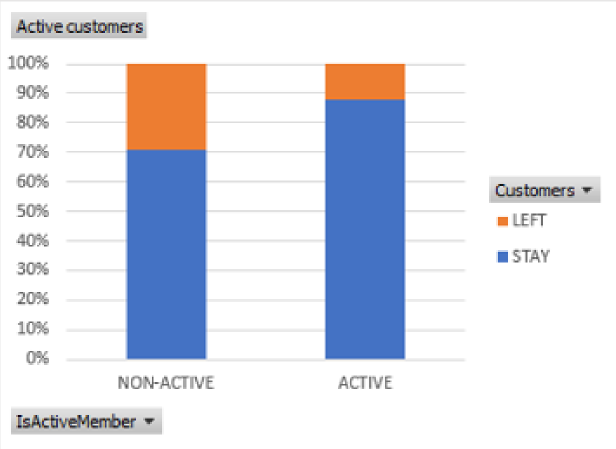
Pig E. Bank



The graph shows customers with a high number of products tends to leave the bank.

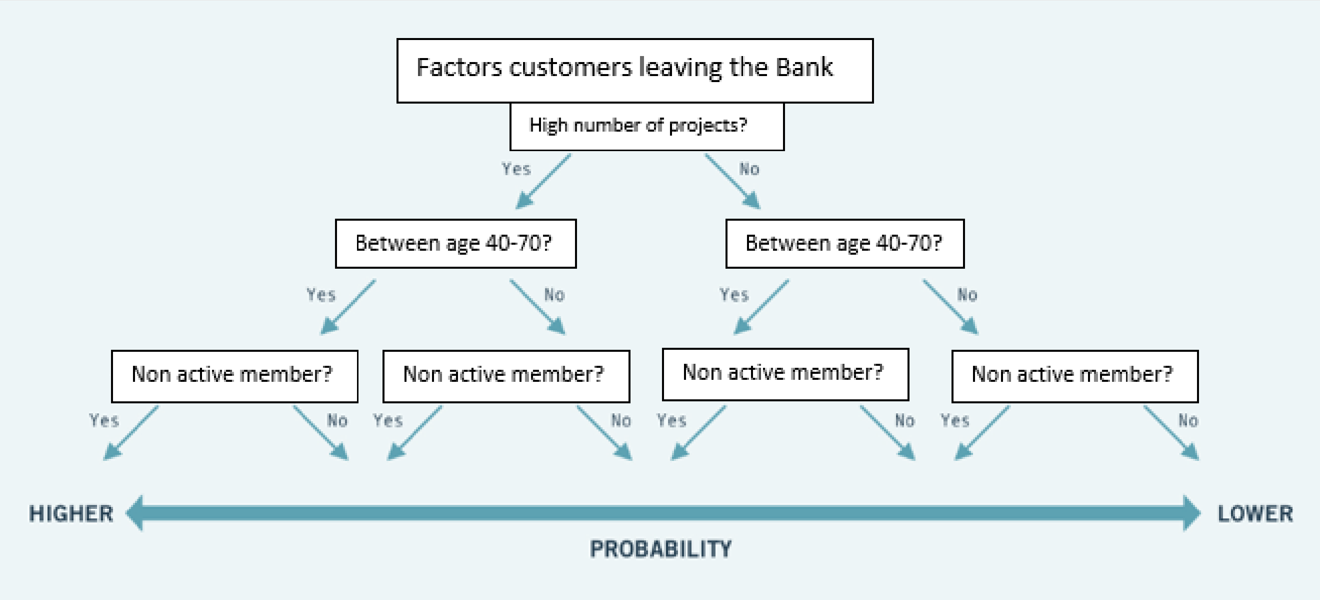


In the age group 51-70, more than 40% of the age group leave the bank.



Non-active customers are more likely to leave the bank.

Pig E. Bank

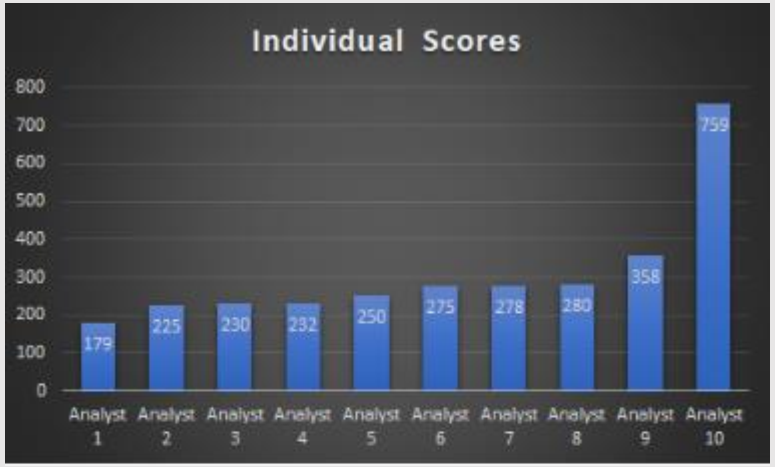


With the bar charts earlier, I set up the decision tree to visuals the probability of the factors customer leaving the bank. A high number of projects who are between 40 to 70 years old non-active customers has the highest probability to leave the bank and a low number of projects who are not between 40-70 active customers has the lowest probability to leave the bank.

Pig E. Bank

The bank is setting up an anti-money laundering plan by having analysts outsource from another department to check data to see if the transactions are fraudulent.

The analysts resulted in measurement bias since the analysts don't have the same experience and training, they produce different result. As seen from the graph one of the analyst has a very high score because of different viewpoints of the transactions.



Pig E. Bank
Result

Based on the decision tree, we can focus on turning customers to low probability to leave the bank such as less profiles and have promotion to make the customers active again.

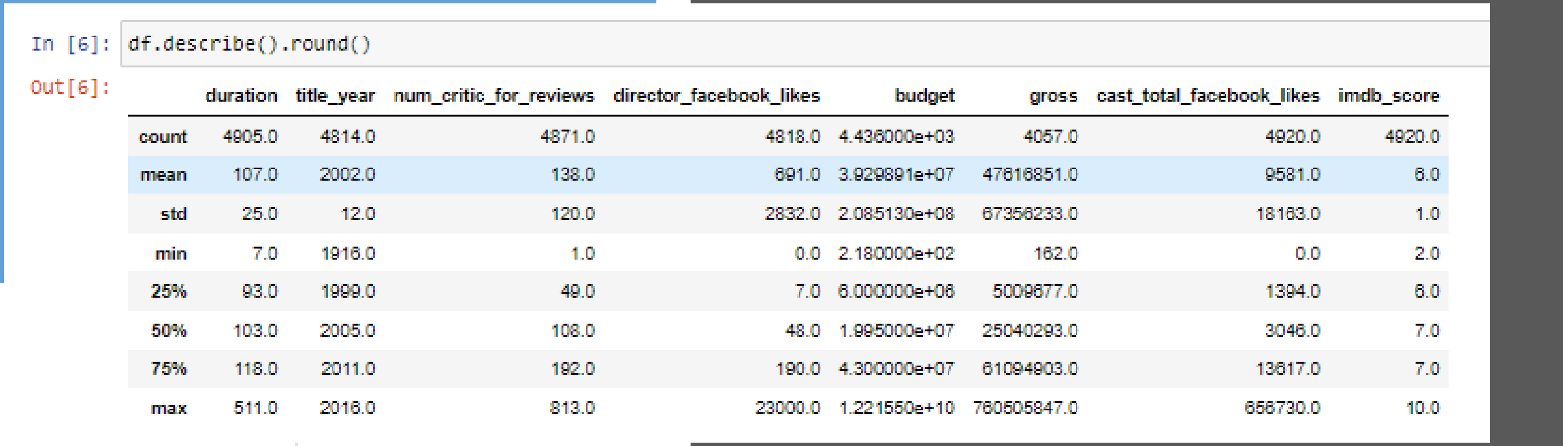
For the money laundering scenario, we can have a set rule of how to detect fraudulent transactions and set up training for analysts to avoid bias.

- Skill
  - Excel chart/pivot table
  - Data ethics
  - Decision tree

6) IMDb

- IMDb is an online database for movies, tv shows and more. It includes a lot of information about the movies and also let viewers rate the movie.
- I use the dataset which includes 5000 movies worldwide between 1916 to 2016. I analysis the data to see if the IMDb score affects the movie data.

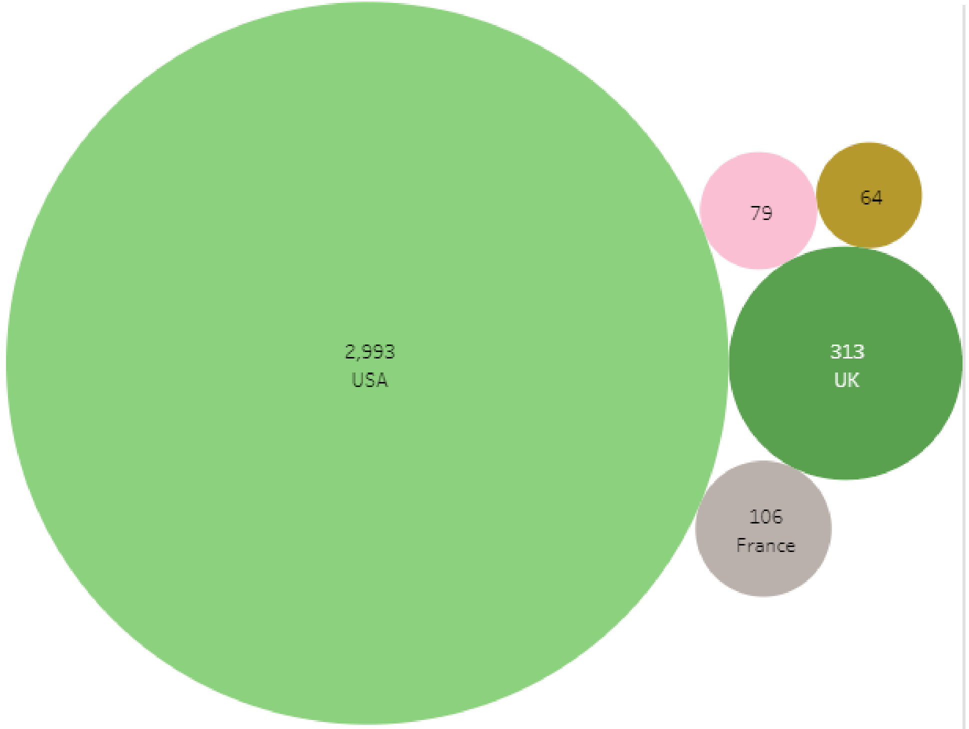
IMDb



In [6]:
```
df.describe().round()
```

Out[6]:

	duration	title_year	num_critic_for_reviews	director_facebook_likes	budget	gross	cast_total_facebook_likes	imdb_score
count	4905.0	4814.0	4871.0	4818.0	4.436000e+03	4057.0	4920.0	4920.0
mean	107.0	2002.0	138.0	691.0	3.929891e+07	47616851.0	9581.0	6.0
std	25.0	12.0	120.0	2832.0	2.085130e+08	67356233.0	18163.0	1.0
min	7.0	1916.0	1.0	0.0	2.180000e+02	162.0	0.0	2.0
25%	93.0	1999.0	49.0	7.0	6.000000e+06	5009677.0	1394.0	6.0
50%	103.0	2005.0	108.0	48.0	1.995000e+07	25040293.0	3046.0	7.0
75%	118.0	2011.0	192.0	190.0	4.300000e+07	61094903.0	13617.0	7.0
max	511.0	2016.0	813.0	23000.0	1.221550e+10	760505847.0	656730.0	10.0



Over 60% of the movies in the dataset are made in USA. The reason can be because majority of the movies are made in USA and IMDb is a website in English and therefore less foreign movies are shown.

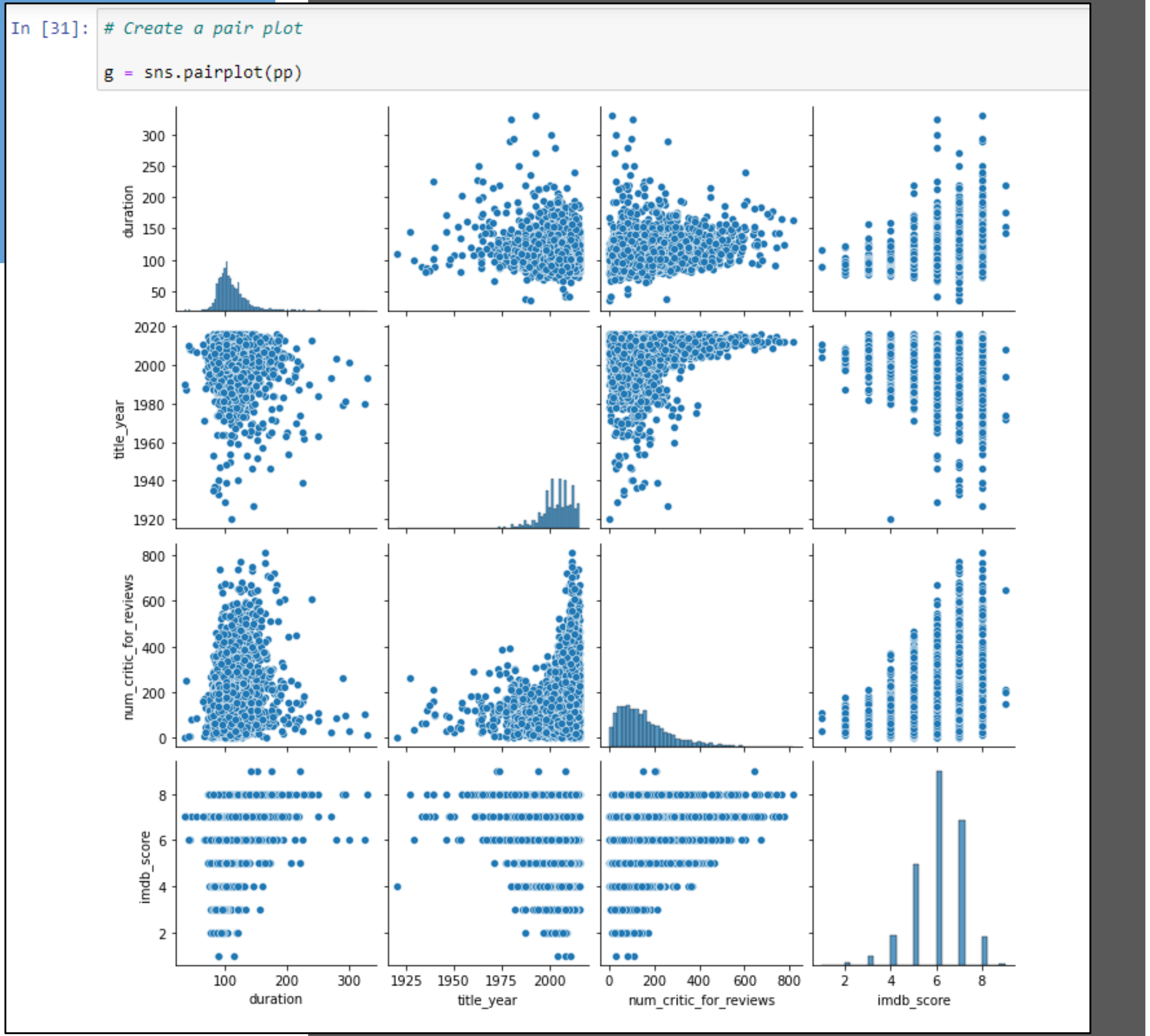
The average IMDb score is 6. It means people think the average score of movies are 6 out of 10.

IMDb

This is a pair plot prepared from Python. It compares different variables in graphs at the same time.

It shows there are less low IMDb score and not many high critic reviews with movies before the 1070s.

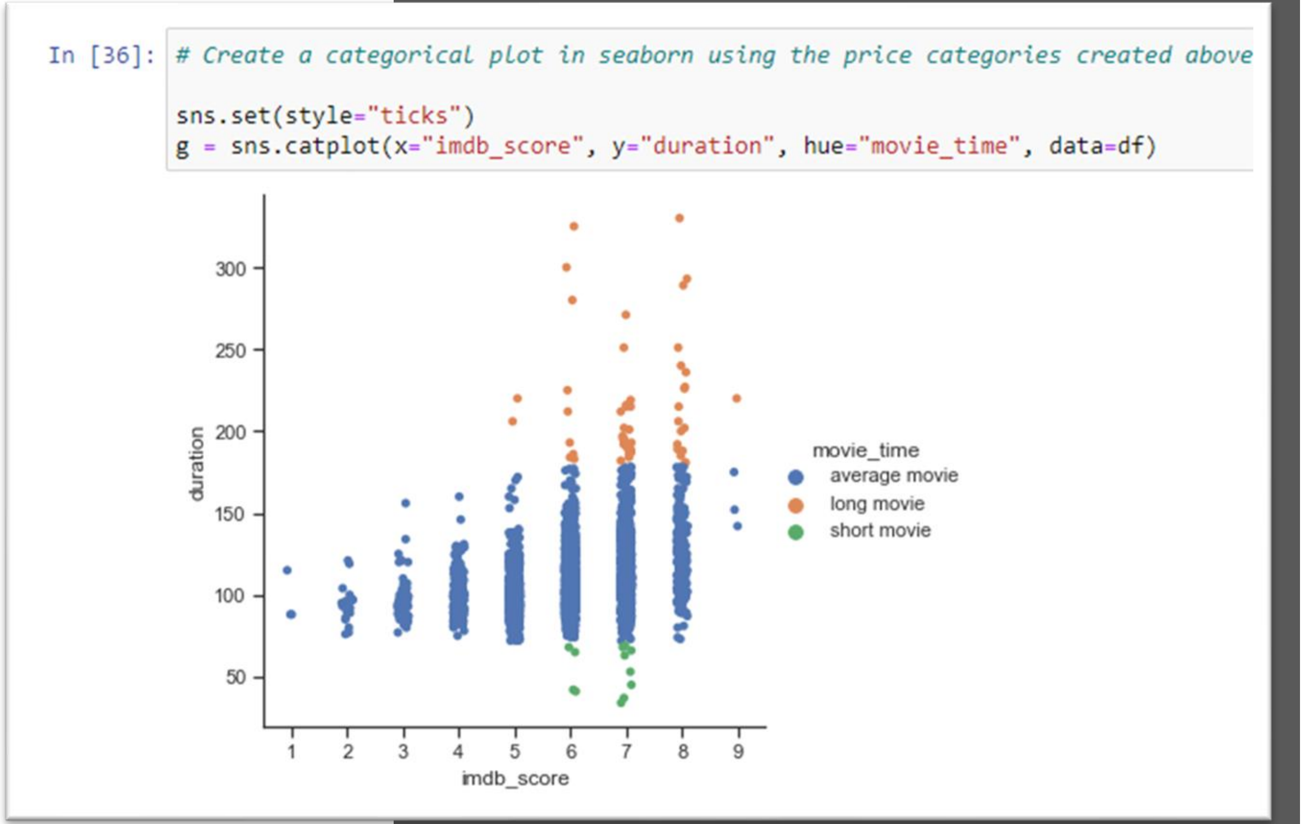
IMDb score shows average is 6 and movies with high number critic reviews have high IMDb score.



```
In [31]: # Create a pair plot

         g = sns.pairplot(pp)
```

IMDb

- The categorical plot shows long movies have an IMDb score of 5 and over and short movies have an IMDb score of 6 and 7. To find a “good” movie that has an IMDb score of 5 and over, it should be either a long movie or a short movie.



```
In [36]: # Create a categorical plot in seaborn using the price categories created above

         sns.set(style="ticks")
         g = sns.catplot(x="imdb_score", y="duration", hue="movie_time", data=df)
```

IMDb Result:

- More than half of the movies in the IMDb dataset are made in United States.
- The average score of IMDb is 6 with majority of the rating between 3 and 9.
- There are high critic reviews of movies with less cast total Facebook likes, and there are low critic reviews of movies with high cast total Facebook likes. we can look into other factors and see how they correlated .

Suggestions:

- We can add more foreign movies to the dataset.
- The genre section has multiple theme in each movie cause confusion in the report. cleaning up to one genre per movie and we can find out more patterns about the genre of movies.
- Put in more details of which State in the United State flim the movies for a more detail report.

[Kaggle dataset]

[Tableau presentation]

Skill:

-Excel

sorting

data cleaning

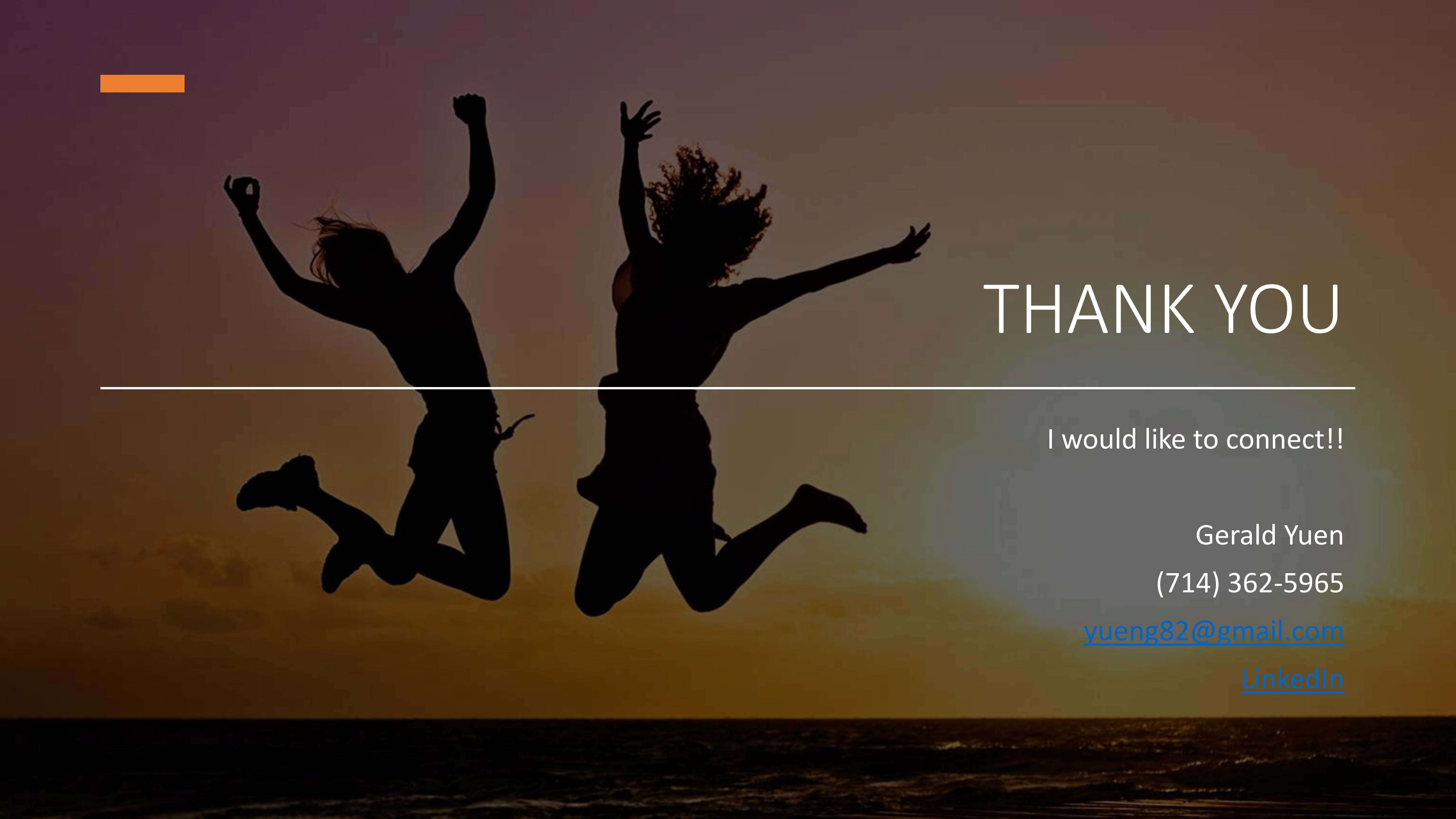
-Python

Pair plot

linear regression

Categorical plot

-Tableau



THANK YOU

I would like to connect!!

Gerald Yuen

(714) 362-5965

yueng82@gmail.com

LinkedIn