KRISHNENDU BARMAN

DATA SCIENCE

PORTFOLIO

2023

PROFESSIONAL BACKGROUND

Currently, I am pursuing my M.Tech. Degree in Environmental Engineering from IIT Bombay with a CGPA of 8.65, have several skills, including Machine Learning, Data Analysis, and Python.

I have taken training from Trainity with Eight real-world Projects, and also worked on some data sets as a self-project.

As a fresher, I am very excited to work on real-world problems. I want to see closely how the corporate world deals with problems. Being a fresher, I am very adaptive to learning a new skills. I always been a team guy who loves to face new problems and get to the solution through critical thinking.

TABLE OF CONTENTS

Data Analytic Process

In this project, the main USP was to learn how a data analytics or data science project to be staged. What are the common questions, and what approach should I take to reach the answer? And at the end of the process, I have to make the right decision according to the requirements.

The Problem

How to stage a data analytics problem in the projects.

Findings

There are some common steps to take -

- Plan
- Prepare
- Process
- Analyze
- Share
- Act

Instagram User Analytics

User analysis is how we track how users engage and interact with our digital product (software or mobile application) to derive business insights for marketing, product & development teams. Teams across the business then use these insights to launch a new marketing campaign, decide on features to build for an app, track the app's success by measuring user engagement and improve the experience while helping the business grow.

The Problem

The problem, Using the dataset of Instagram, is to provide insights on the questions asked by for future steps.

Approach

•The 1st approach to see in the data tables and get an outline what are the given information I have.
•Find out and make a rough roadmap for every problem I have to solve.
•Then try out the concepts of SQL from aggregator to join.

Findings 1

Top 5 old Instagram users

id	username	created_at
80	Darby_Herzog	06-05-2016 00:14
67	Emilio_Bernier52	06-05-2016 13:04
63	Elenor88	08-05-2016 01:30
95	Nicole71	09-05-2016 17:30
38	Jordyn.Jacobson2	14-05-2016 07:56

Theses are the IDs and their names who are most old users of Instagram.

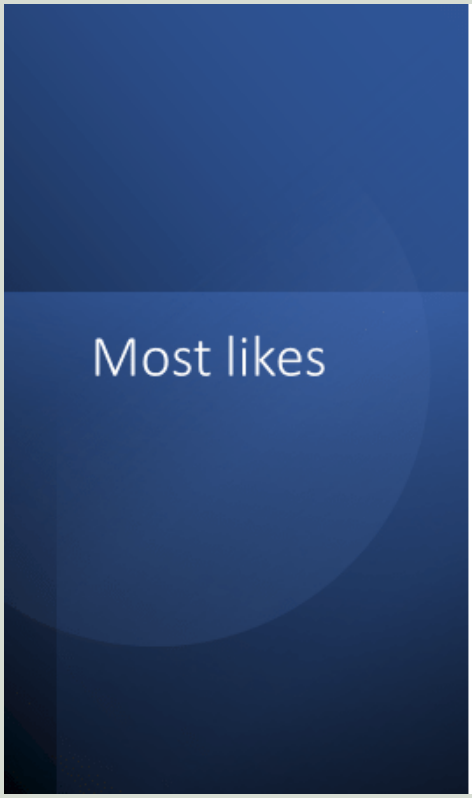
Findings 2

IDs with no photo posting

id	username	photo_id
5	Aniya_Hackett	NULL
7	Kassandra_Homenick	NULL
14	Jaclyn81	NULL
21	Rocio33	NULL
24	Maxwell.Halvorson	NULL
25	Tierra.Trantow	NULL
34	Pearl7	NULL
36	Dillie_Ledner37	NULL
41	Mckenna17	NULL
45	David.Osinski47	NULL
49	Morgan.Kassulke	NULL
53	Linnea59	NULL
54	Duane60	NULL
57	Julien_Schmidt	NULL
66	Mike.Auer39	NULL
68	Franco_Keebler64	NULL
71	Nia_Haag	NULL
74	Hulda.Macejkovic	NULL
75	Leslie67	NULL
76	Janelle.Nikolaus81	NULL
80	Darby_Herzog	NULL
81	Esther.Zulauf61	NULL
83	Bartholome.Bernhard	NULL
89	Jessyca_West	NULL
90	Esmeralda.Mraz57	NULL
91	Bethany20	NULL

Not all Instagram users post photos on thee media, Theses are some users who never post a on Instagram.

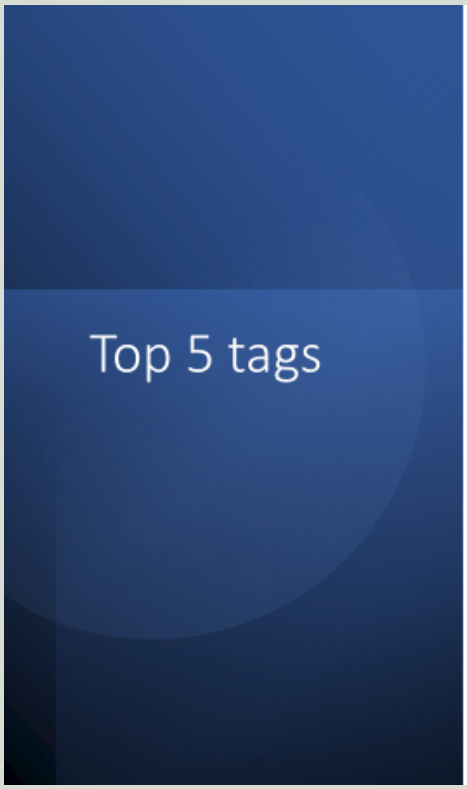
Findings 3



id	username	photo_id	like_count	posting_date
52	Zack_Kemmer93	145	48	26-11-2022 17:11
65	Adelle96	182	43	26-11-2022 17:11
46	Malinda_Streich	127	43	26-11-2022 17:11

Some photos got lots of likes from other users, Instagrame can promote those photos and also photos from those users to engage more users on Instagram. And here is the list of the top 3 photos and the users whoo posted it.

Findings 4



tag_id	count	tag_name
21	59	smile
20	42	beach
17	39	party
13	38	fun
5	24	food

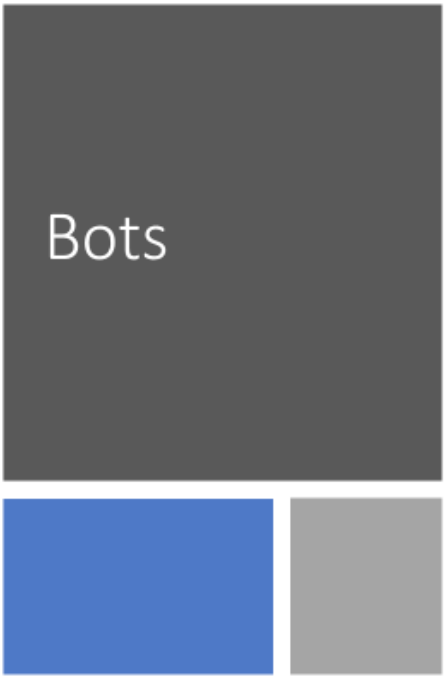
There are some trending tags under the photos for a period of time, the tags help more reach through out the media. Here are the tags which got most like on this social media.

Findings 5

DayofWeek	count
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

The user registration (new) on Instagram can vary depending on the no. of days of the week, but here for this data I can't find any significant difference in new registration in a different day of week on instagram

Findings 6



id	username	no_of_likes
5	Aniya_Hackett	257
14	Jaclyn81	257
21	Rocio33	257
24	Maxwell.Halvorson	257
36	Ollie_Ledner37	257
41	Mckenna17	257
54	Duane60	257
57	Julien_Schmidt	257
66	Mike.Auer39	257
71	Nia_Haag	257
75	Leslie67	257
76	Janelle.Nikolaus81	257
91	Bethany20	257

There are some bots on every social media who likes each and every photo which is not likely for a normal user. Here I have found out the bots from the dataset.

Analysis and conclusion

All the findings are made using SQL. By analyzing the whole data, many insights and essential information are found, from the business point of view, which are very important for the upcoming marketing campaign, more user engagement, and also get an estimation of how users are actively singing Instagram.

Operation Analytics and Investigating Metric Spike

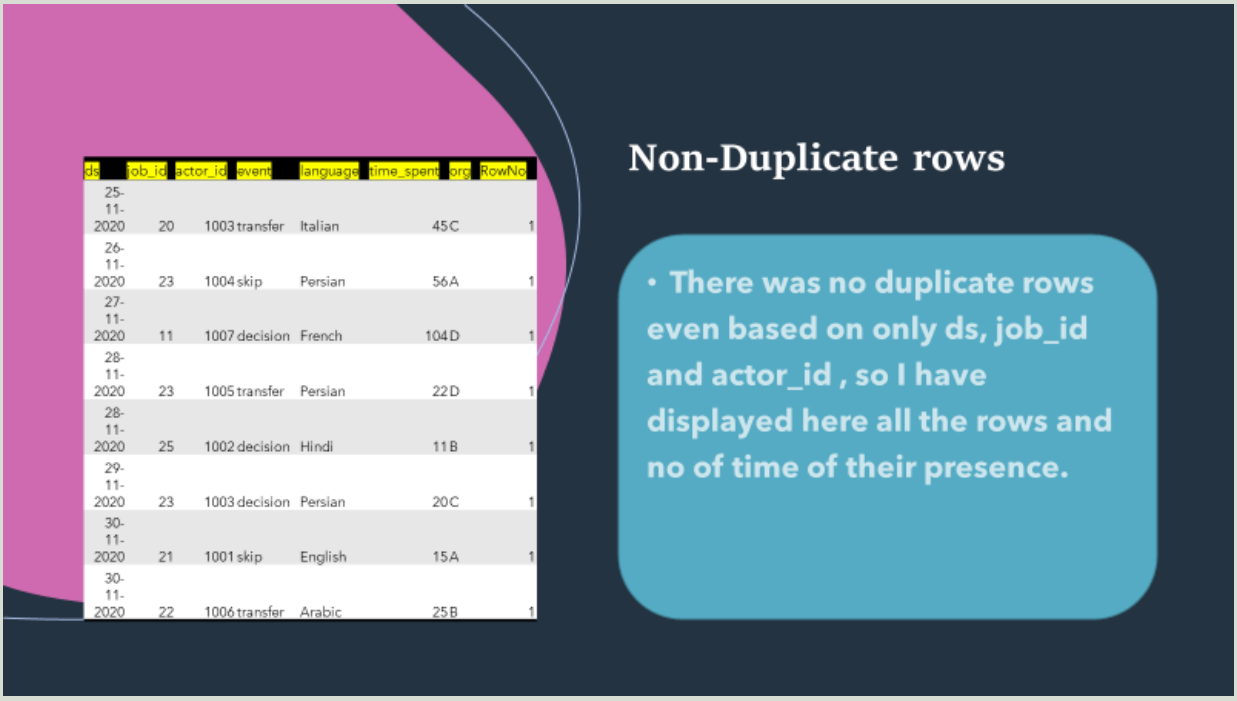
In this project I had to deal with some specific questions, which were asked based on the Operations Metric dataset. This project helped closely with the ops team, support team, marketing team, to derive insights out of the data they collect.

Findings 1

language	perc_jobs
English	12.5
Arabic	12.5
Hindi	12.5
French	12.5
Italian	12.5
Persian	37.5

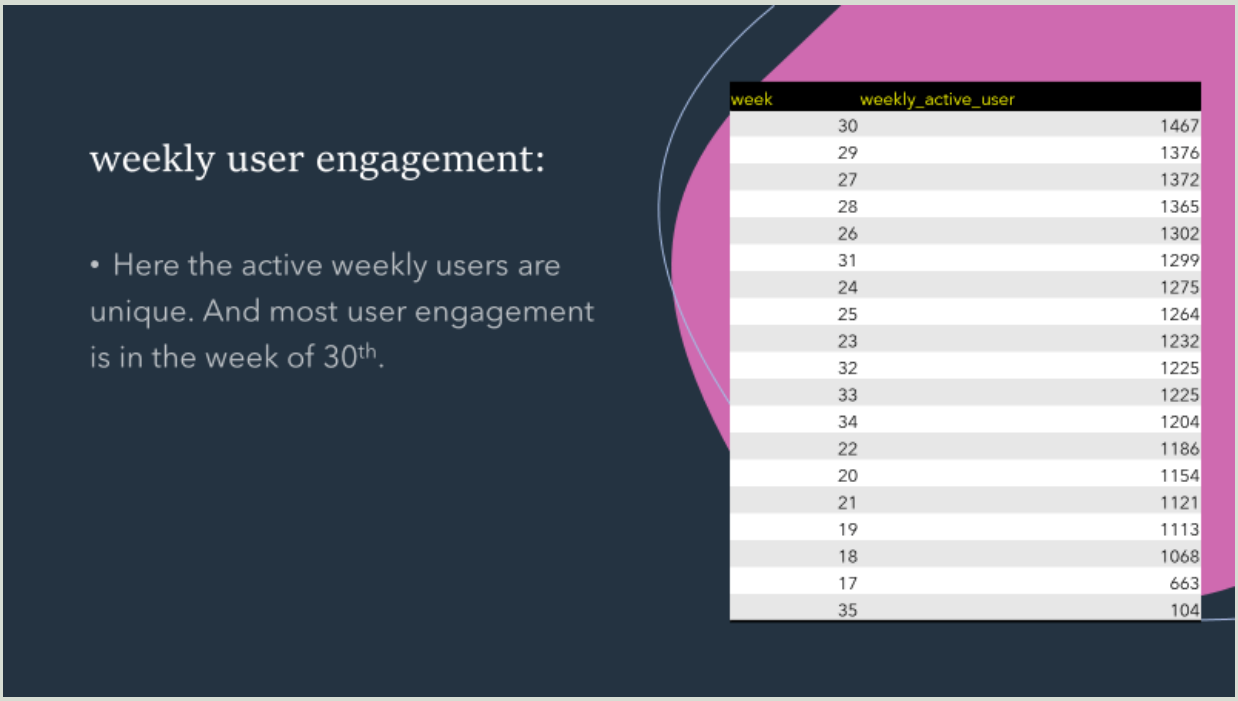
Here, the Persian language job is searched most compared to other languages, So the company should emphasis on the Persian language a little bit more than other languages.

Findings 2



No duplicate rows were found in the data set , every row was unique.

Findings 3



Findings 4

The user growth in monthly basis for both no. of total user and active user.

Analysis and conclusion

Much info from the data set, like weekly engagement, email engagement, and also the average no of jobs reviewed, were found with the help of SQL queries, this information helps the company which type of job they should emphasize and also helps for tracking the company growth.

Hiring Process Analytics

In this project, the hiring process of an MNC is discussed. The statistical information of this company like no of men and women, no of people in different departments and also their salaries and many more insights.

Findings 1

How many Males and Females are there?

Male	Female	Do not know	Blank
4085	2675	393	15

This results are obtained from the statistical analysis using COUNTIF function keeping the range as D coloumn.

Findings 2

Average Salary

Count of the employees	7168
Total salary	3.6E+08
Average salary	49976.1

This result is obtained using the COUNT , SUM and Divided by the no. of employees .
We can also use the AVG function on the salary to find this.

Findings 3

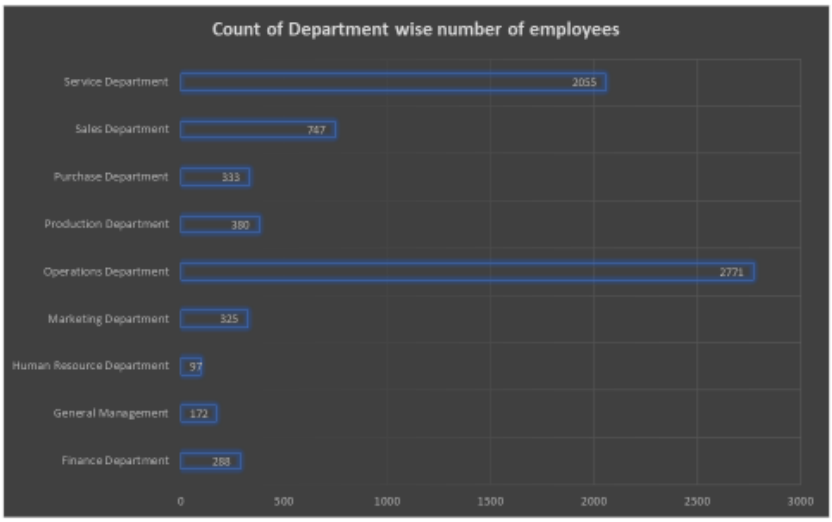
Highest and Lowest salary

Highest salary	400000
Lowest Salary	100

This is obtained by using MAX and MIN function on Salary offered

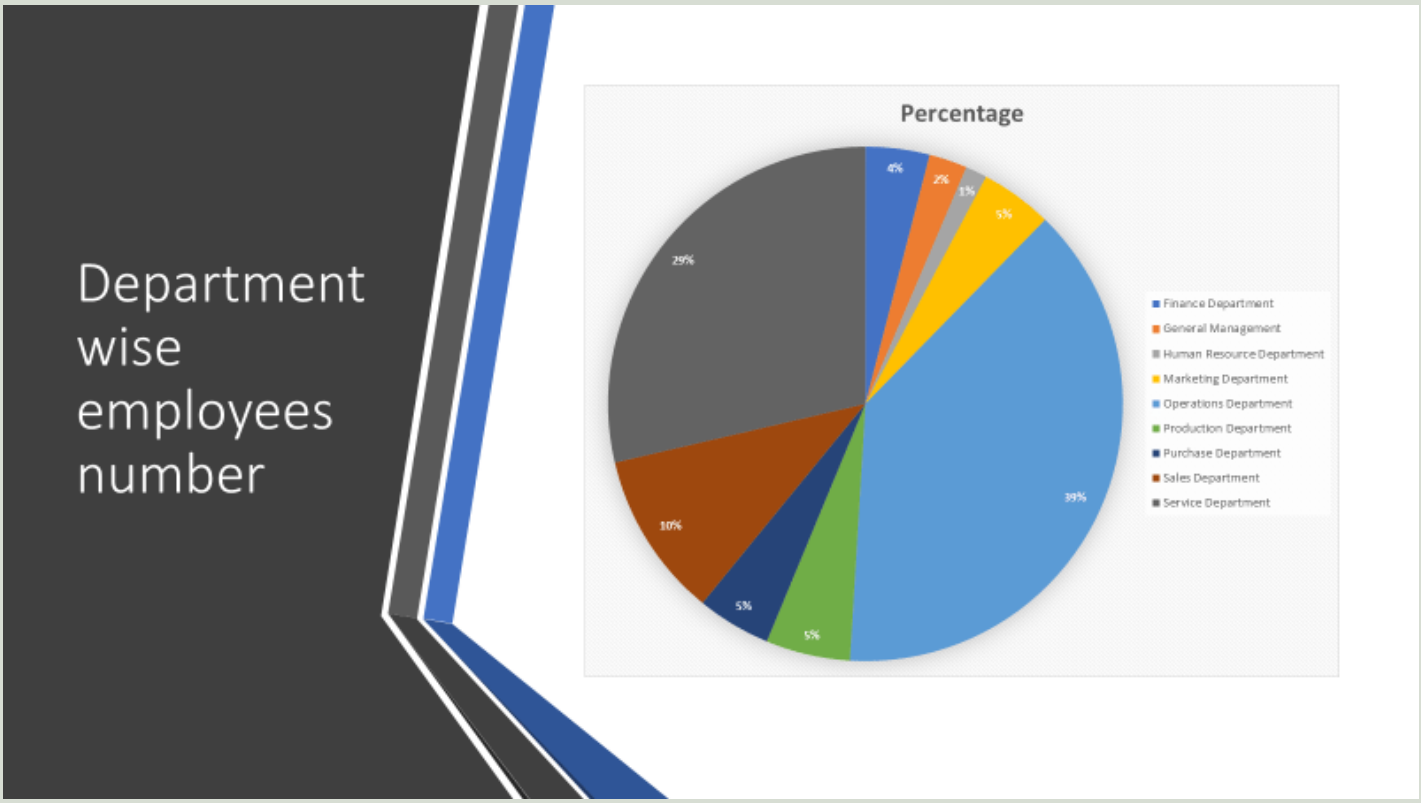
Findings 4

Department wise employees number

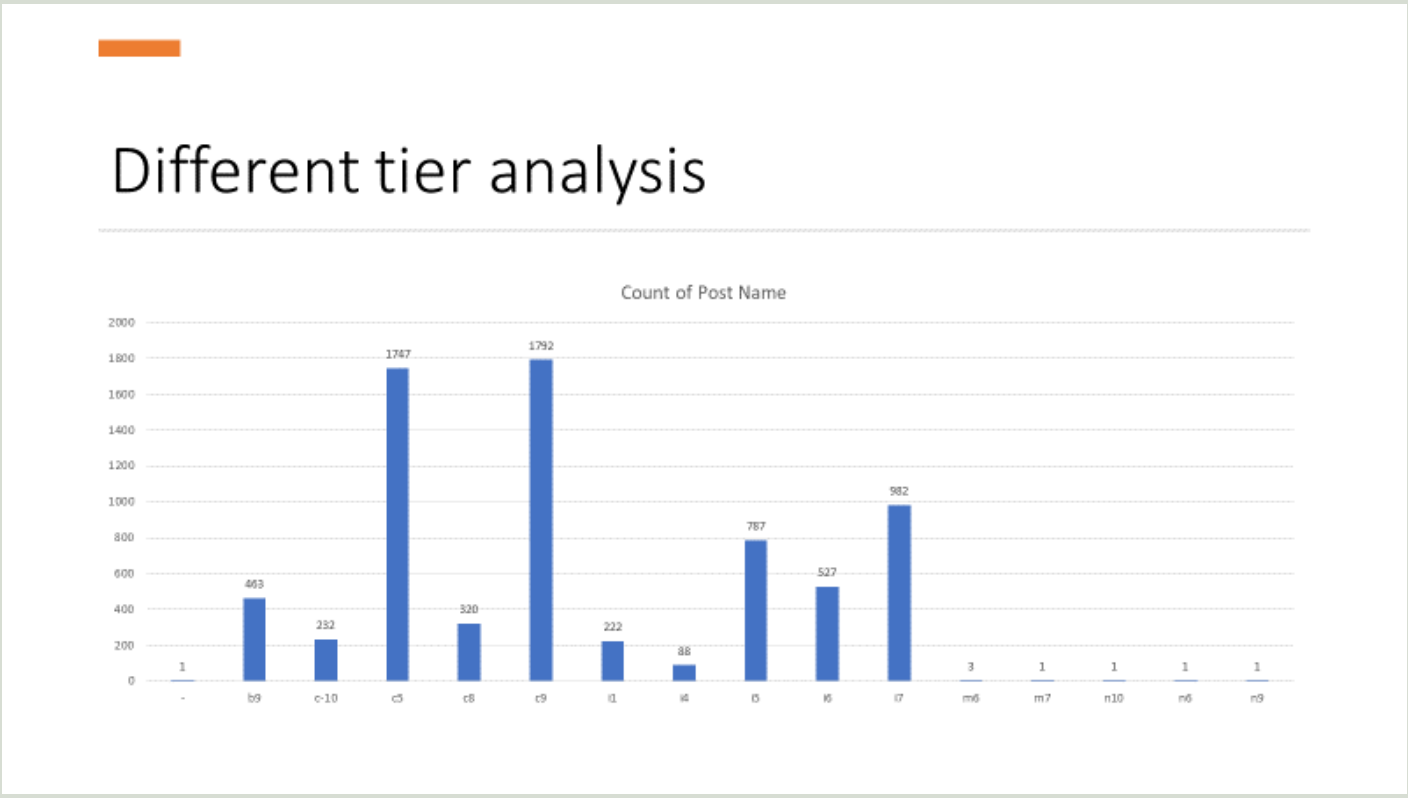


This analysis is done using pivot chart analysis

Findings 5

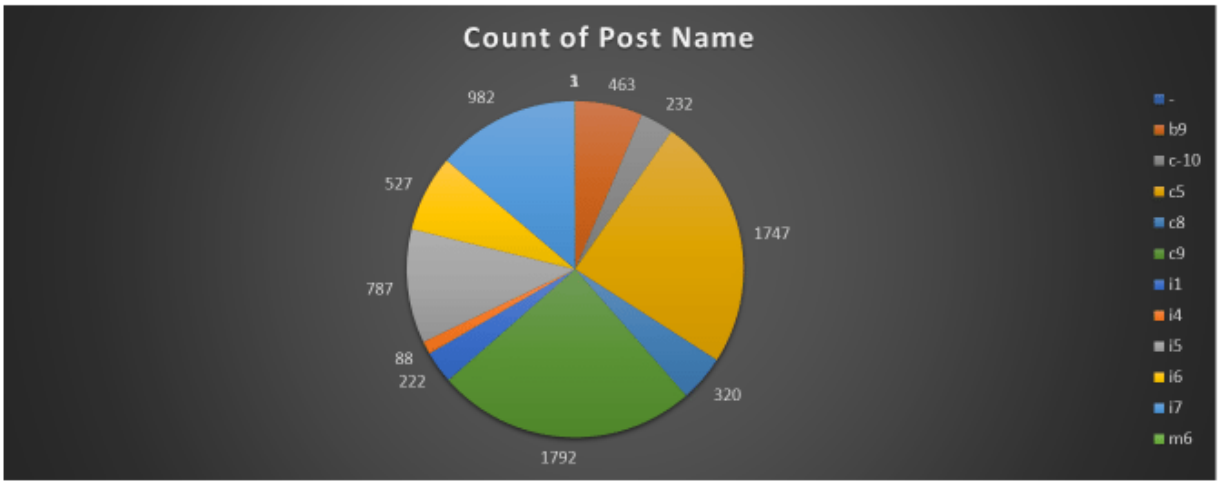


Findings 6



Findings 7

Different tier analysis



Both the bar and a pie chart of the different tiers of the employments.

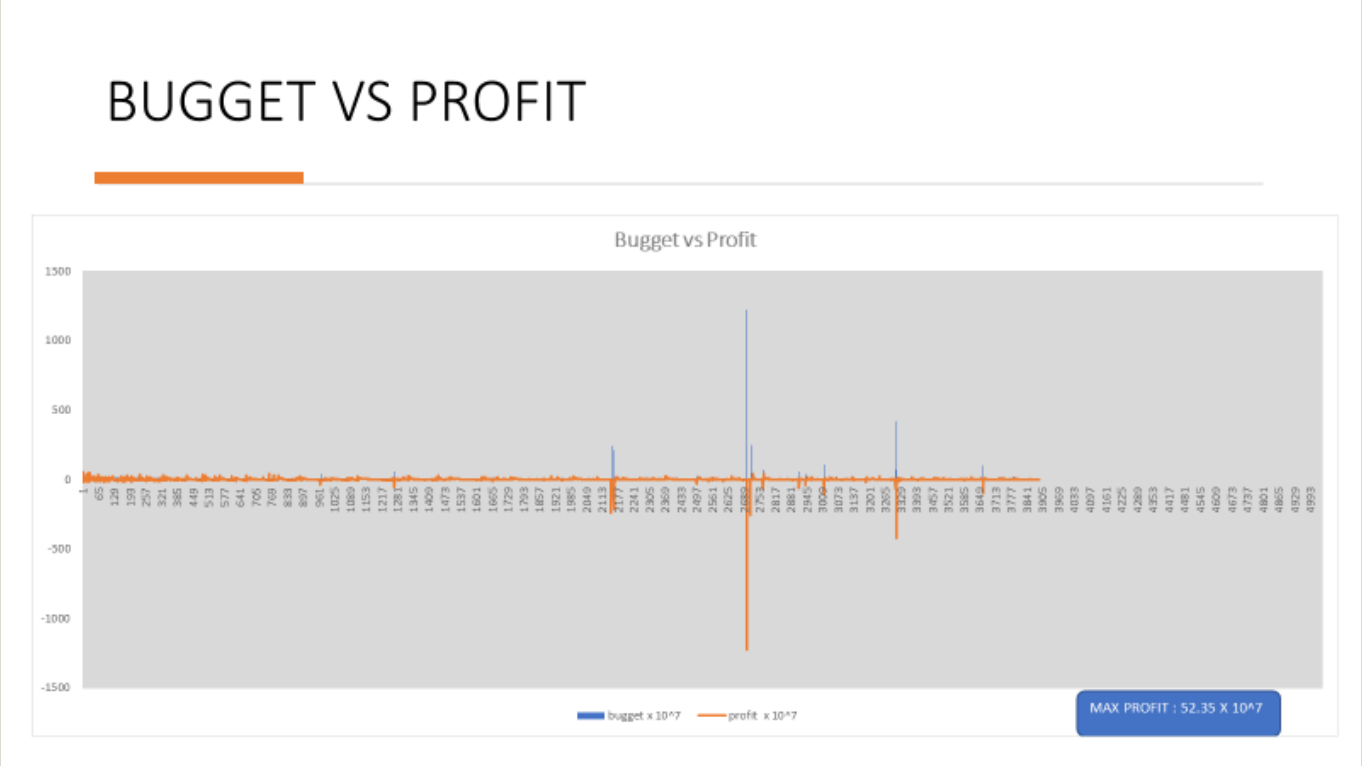
Analysis and conclusion

This type of analysis is necessary for a company to track its employment information. That may help the MNC how to manage the workload of every employee, whom to give promotions to in the future time or manage gender equality in the company.

IMDB Movie Analysis

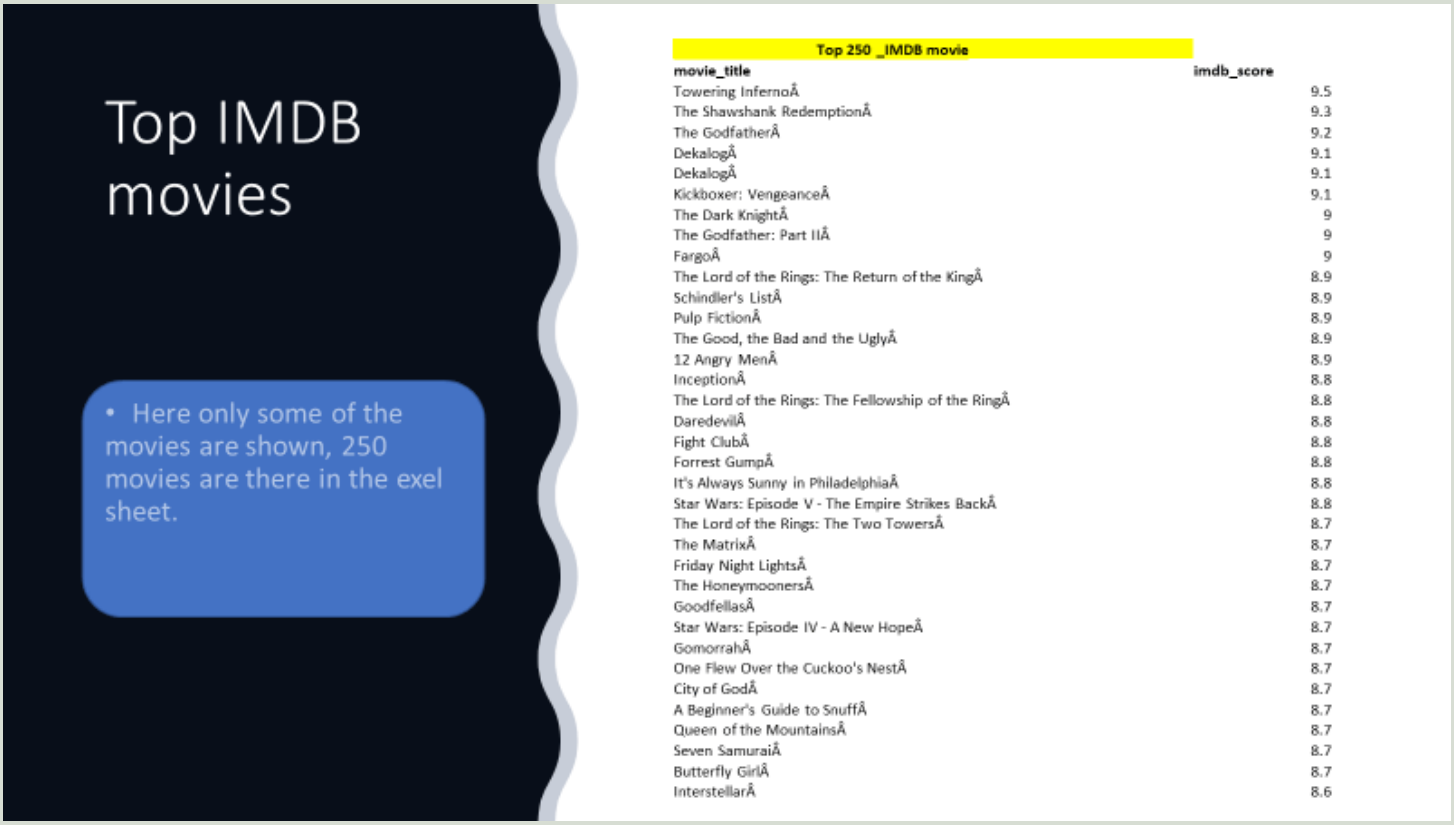
This dataset has various columns of different IMDB Movies. Where the number of columns is very large and many information like movie name their leading actor name, date, number of critics' review,s and many more, I have tried to find out some of its insights from the data set.

Findings 1



There are a vast number of movie information in the data set. So it was easy situation to visualize all the budgets and profits for an inappropriate graph. But the problem was finding the most profitable movie on the list. It was found using this double-sided bar graph where the upper side shows the budget of the movies and the lower one shows the profit from the movies.

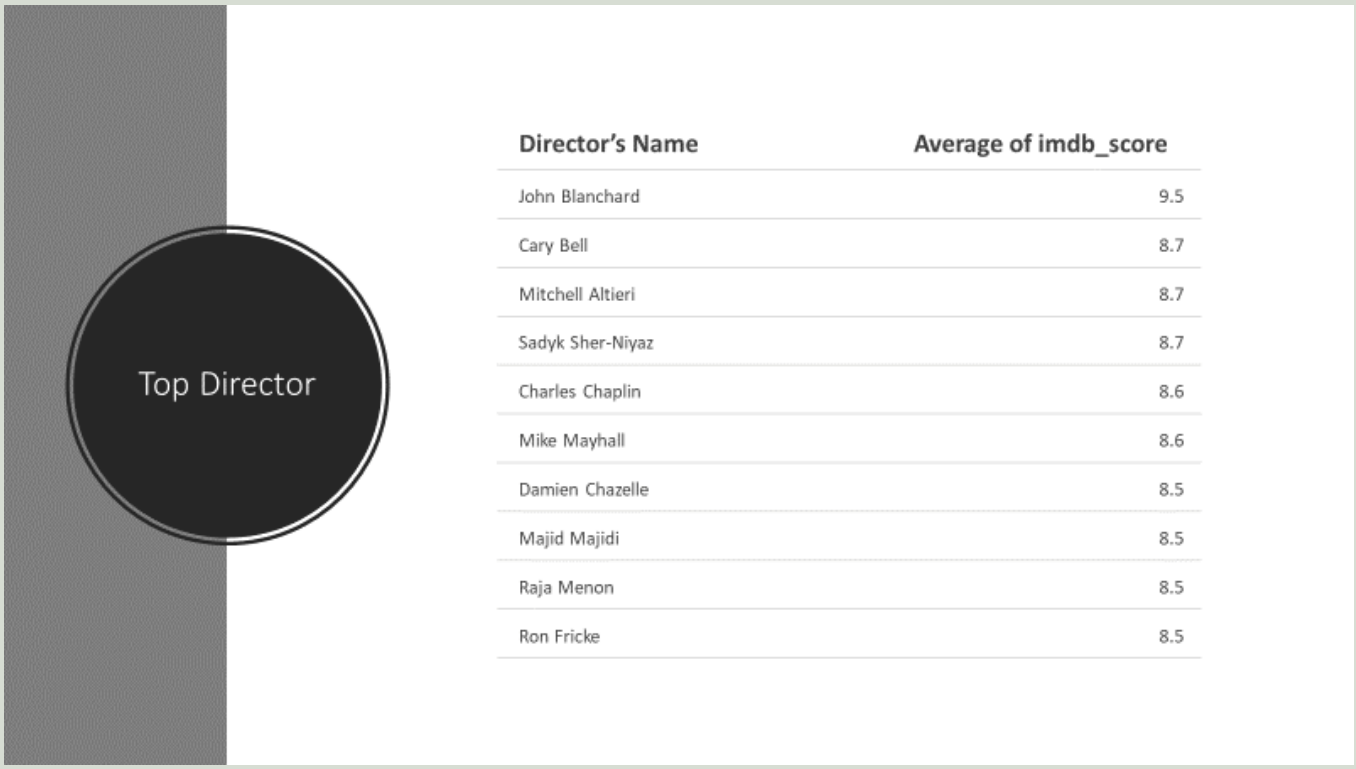
Findings 2



Top 250 _IMDB movie	
movie_title	imdb_score
Towering InfernoÂ	9.5
The Shawshank RedemptionÂ	9.3
The GodfatherÂ	9.2
DekalogÂ	9.1
DekalogÂ	9.1
Kickboxer: VengeanceÂ	9.1
The Dark KnightÂ	9
The Godfather: Part IIÂ	9
FargoÂ	9
The Lord of the Rings: The Return of the KingÂ	8.9
Schindler's ListÂ	8.9
Pulp FictionÂ	8.9
The Good, the Bad and the UglyÂ	8.9
12 Angry MenÂ	8.9
InceptionÂ	8.8
The Lord of the Rings: The Fellowship of the RingÂ	8.8
DaredevilÂ	8.8
Fight ClubÂ	8.8
Forrest GumpÂ	8.8
It's Always Sunny in PhiladelphiaÂ	8.8
Star Wars: Episode V - The Empire Strikes BackÂ	8.8
The Lord of the Rings: The Two TowersÂ	8.7
The MatrixÂ	8.7
Friday Night LightsÂ	8.7
The HoneymoonersÂ	8.7
GoodfellasÂ	8.7
Star Wars: Episode IV - A New HopeÂ	8.7
GomorrahÂ	8.7
One Flew Over the Cuckoo's NestÂ	8.7
City of GodÂ	8.7
A Beginner's Guide to SnuffÂ	8.7
Queen of the MountainsÂ	8.7
Seven SamuraiÂ	8.7
Butterfly GirlÂ	8.7
InterstellarÂ	8.6

In the problem statement, the list of the top 250 movies was asked whose rating was the top. It is not possible to show all the movie's names in this slide, but some of hem are shown here.

Findings 3



Director's Name	Average of imdb_score
John Blanchard	9.5
Cary Bell	8.7
Mitchell Altieri	8.7
Sadyk Sher-Niyaz	8.7
Charles Chaplin	8.6
Mike Mayhall	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Raja Menon	8.5
Ron Fricke	8.5

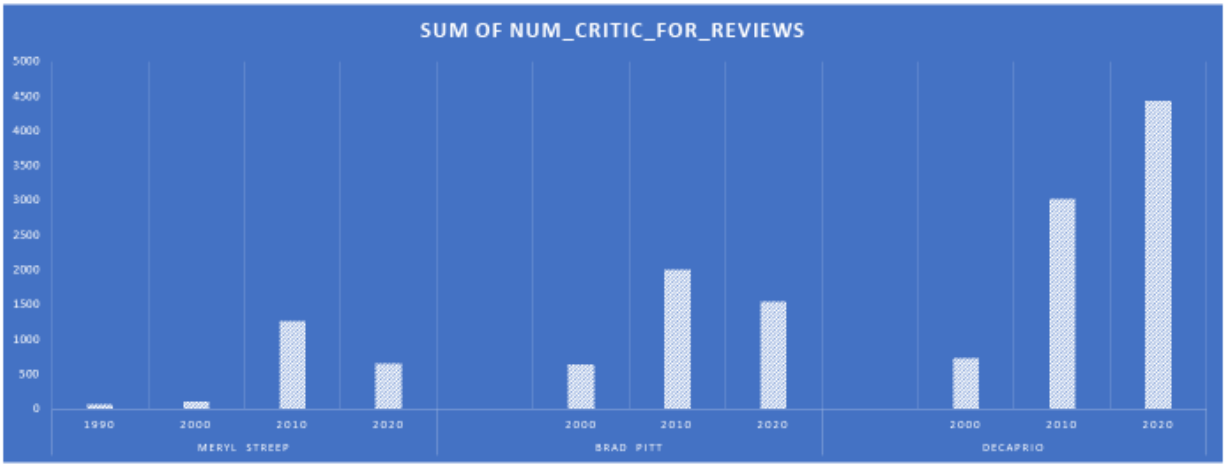
Findings 4

Top Genre

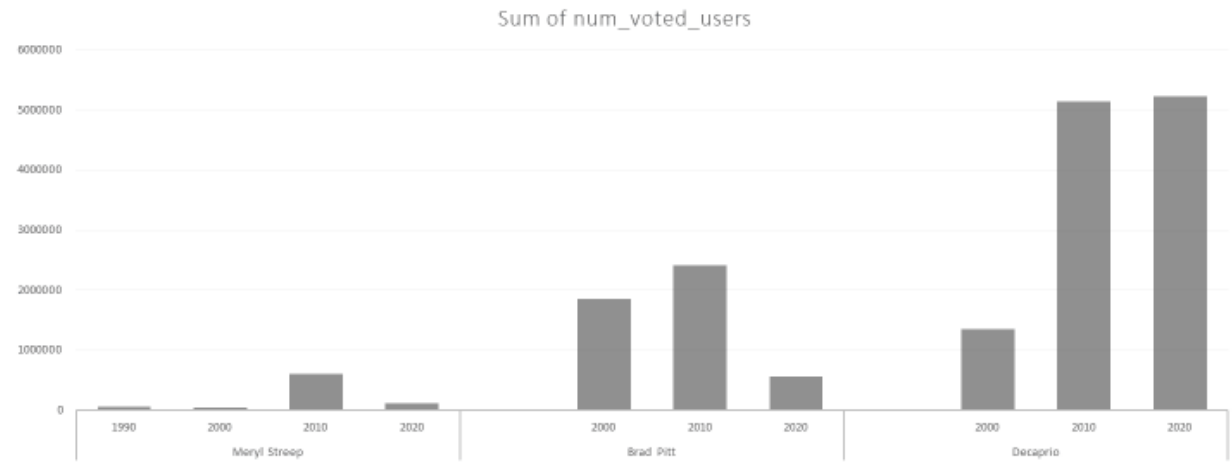
Genre of the movies	Average of imdb_score
Action|Adventure|Crime|Drama|Sci-Fi|Thriller	8.80
Action|Adventure|Biography|Drama|History	8.60
Action|Drama|History|Thriller|War	8.50
Adventure|Animation|Drama|Family|Musical	8.50
Crime|Drama|Fantasy|Mystery	8.50
Action|Adventure|Drama|Fantasy|War	8.40
Action|Animation|Crime|Sci-Fi|Thriller	8.40
Adventure|Drama|Thriller|War	8.40
Comedy|Drama|History|Romance	8.40
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.30

Findings 5

No of critic reviews decade wise



No of user reviews decade wise



The cinema industry evolves with time, both in critic review numbers and also in user review number.

Analysis and conclusion

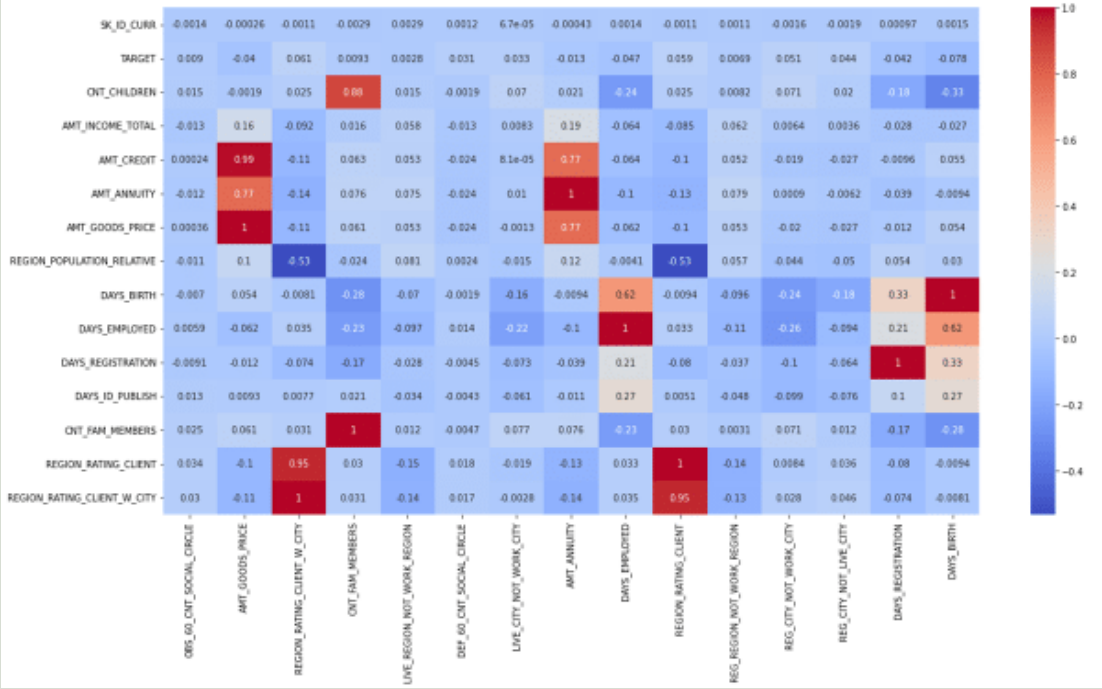
IMDB analysis of the movies can give the movie producers as directors which Genres are being hit in the recent time, which actors are able to collect more profit from the market. Also in the case of actors, they can be sure which directors they should work with for making a good film both critically acclaimed and also in terms of box office collection.

Bank Loan Case Study

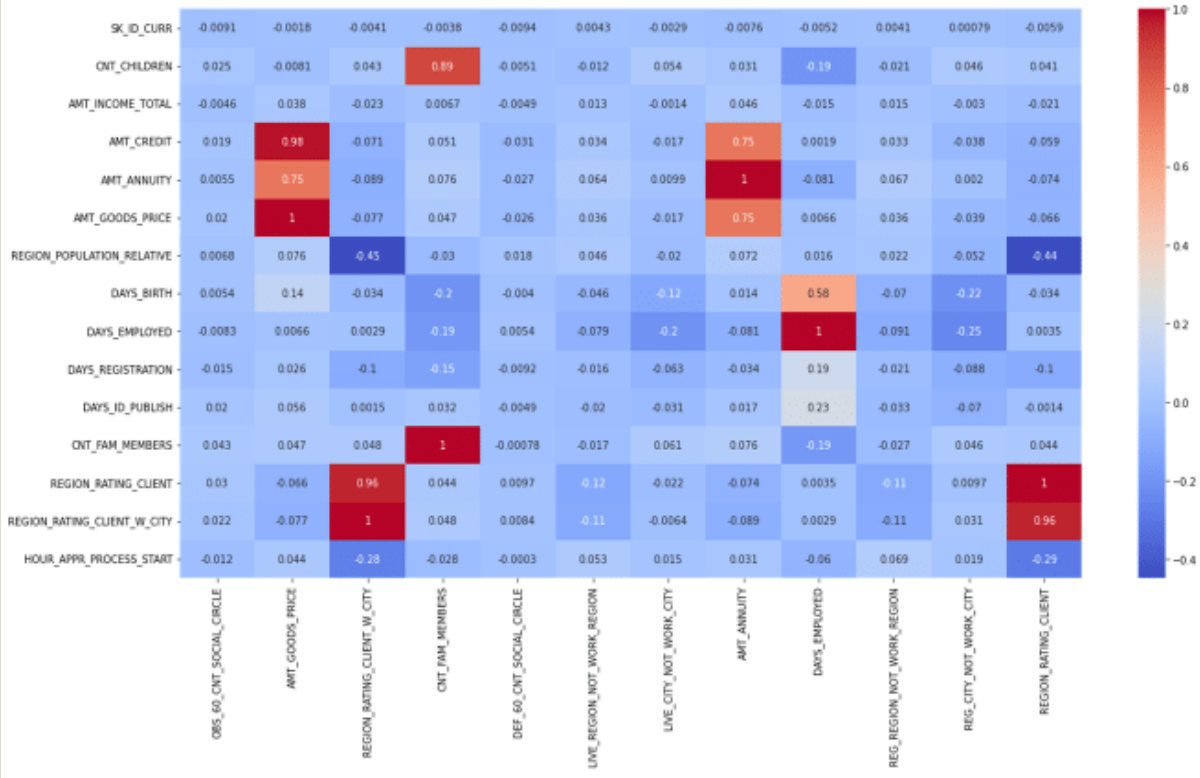
•The data set is huge, So to visualize the data set and to get insights from the dataset, I have used python (jupyter notebook) for the EDA of this data. Python libraries like pandas, NumPy, seaborn and klib, etc. Many insights from the data set, which factors are causing more to affect the target columns that is whoo are repayes and who are defaulters.

Findings 1

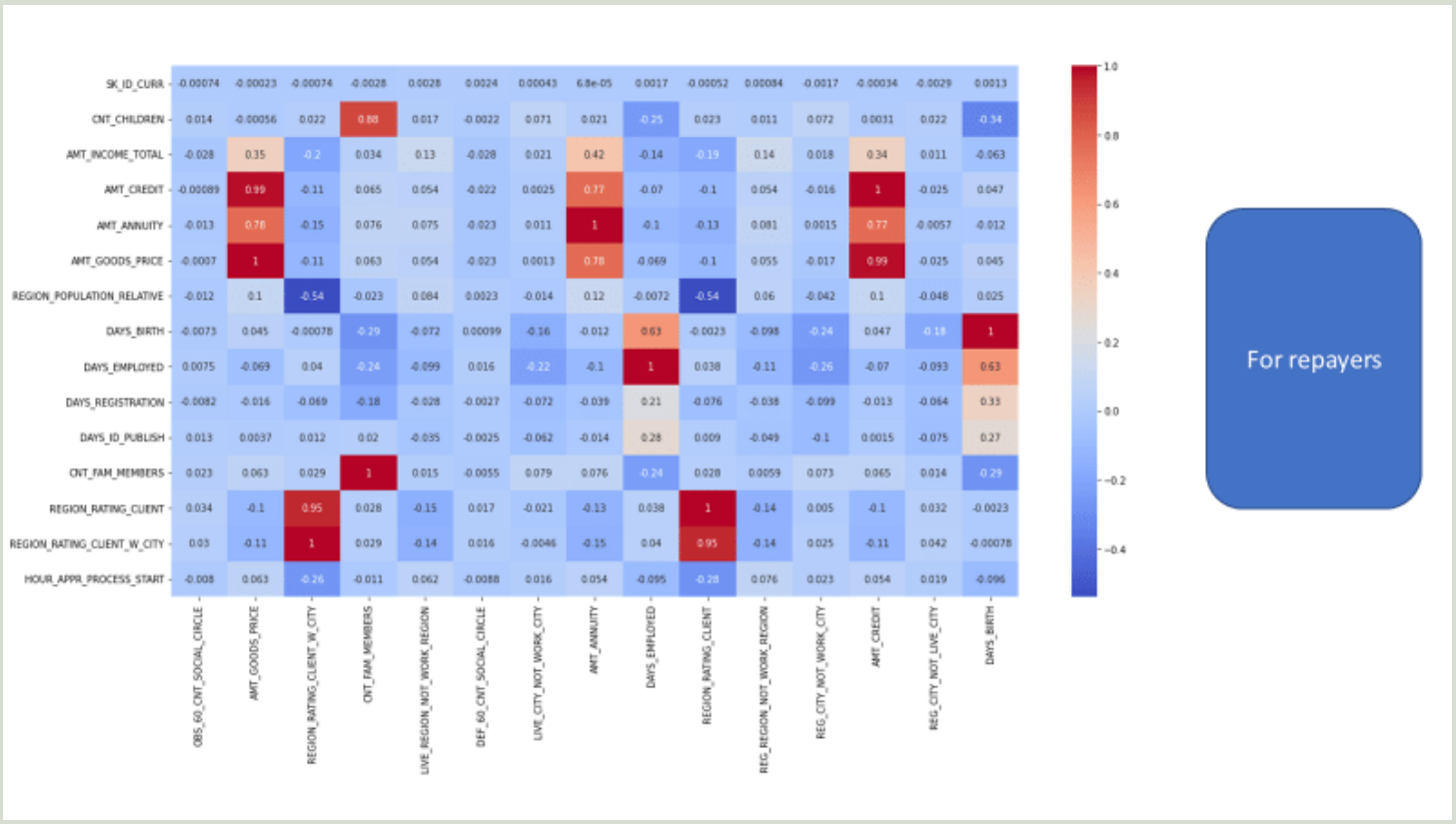
Most corelated numerical columns



For whole Dataset



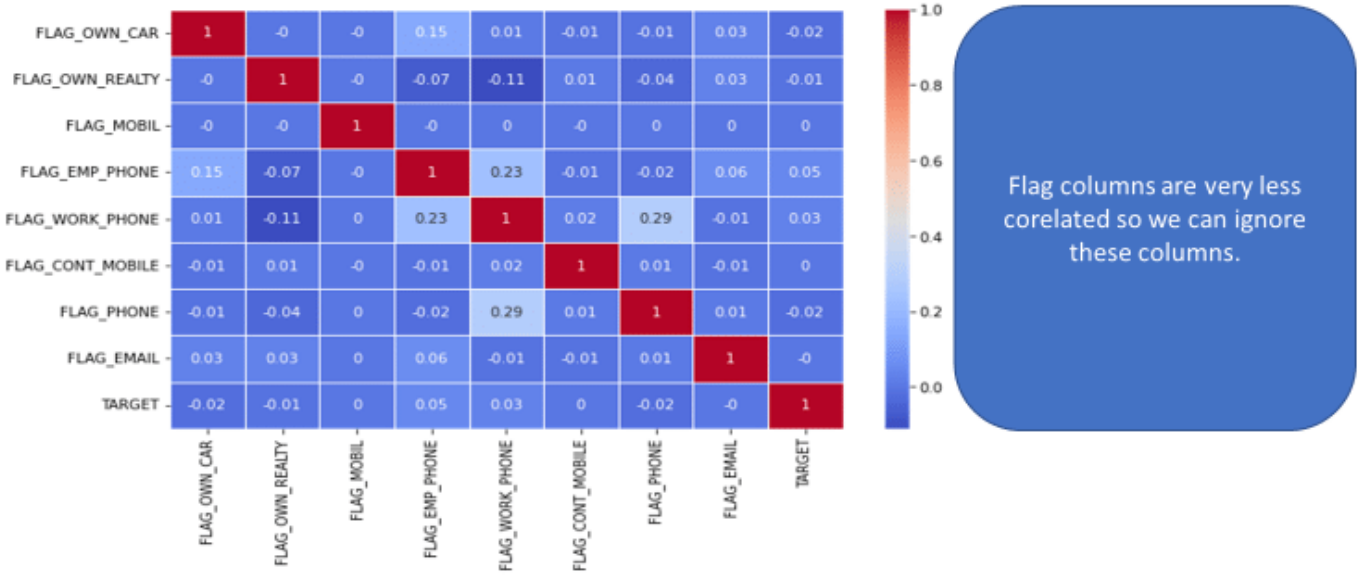
For defaulters



These are the dataset columns most correlated in the data set among the 47 columns. For repayers, defaulters and overall also.

Findings 2

Flag corelation



Flag columns are very less corelated so we can ignore these columns.

More columns ignored

EXT_SOURCE_2 and EXT_SOURCE_3 are also ignored.

Findings 3

Deal with missing values

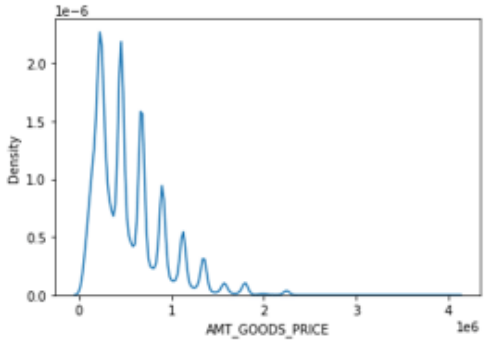
- DAYS_LAST_PHONE_CHANGE 0.000325
- CNT_FAM_MEMBERS 0.000650
- AMT_ANNUITY 0.003902
- AMT_GOODS_PRICE 0.090403
- DEF_60_CNT_SOCIAL_CIRCLE 0.332021
- OBS_60_CNT_SOCIAL_CIRCLE 0.332021
- DEF_30_CNT_SOCIAL_CIRCLE 0.332021
- OBS_30_CNT_SOCIAL_CIRCLE 0.332021
- NAME_TYPE_SUITE 0.420148
- AMT_REQ_CREDIT_BUREAU_QRT 13.501631
- AMT_REQ_CREDIT_BUREAU_HOUR 13.501631
- AMT_REQ_CREDIT_BUREAU_DAY 13.501631
- AMT_REQ_CREDIT_BUREAU_WEEK 13.501631
- AMT_REQ_CREDIT_BUREAU_MON 13.501631
- AMT_REQ_CREDIT_BUREAU_YEAR 13.501631
- OCCUPATION_TYPE 31.345545

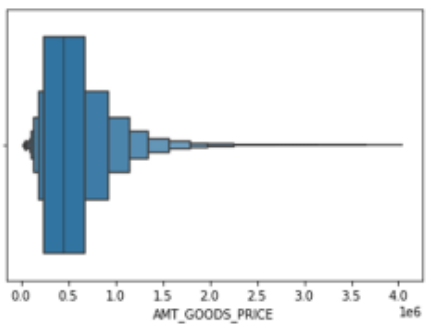
Percentage of missing values of different columns, others has no missing values

The numerical columns are filled with mean,median or mode according the statistic of that particular columns. The categorical columns are filled with most_frequent values. And all the columns are filled.

Findings 4

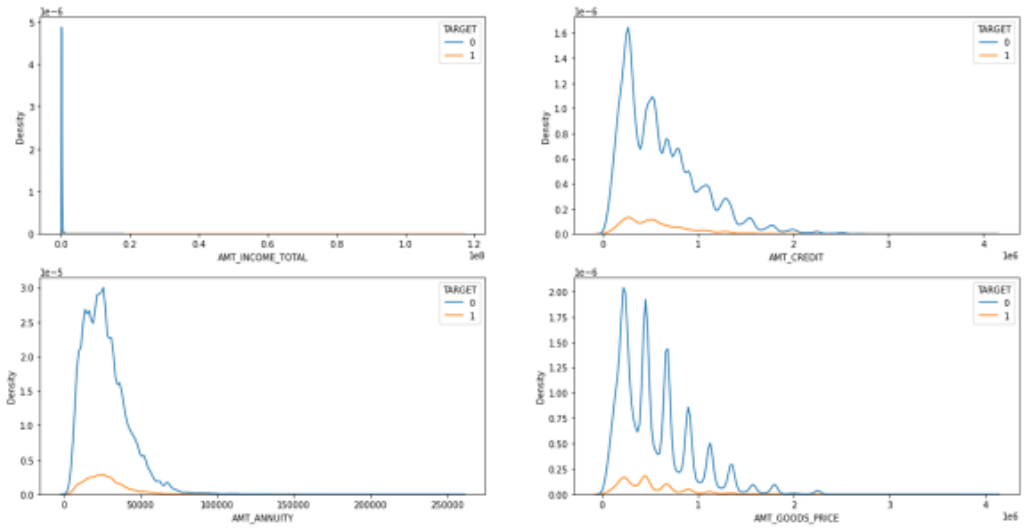
Density plots for different columns





Findings 5

Density plots for different columns



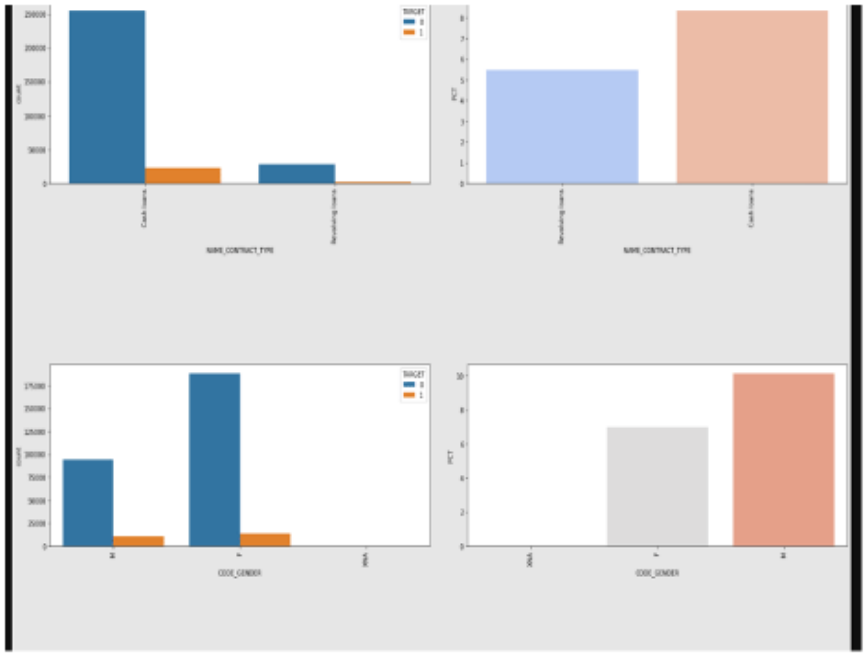
Insights from previous univariate analysis

- 1. AMT_INCOME_TOTAL : Moset of the repayers are in the range in 1 milloin of income.
- 2. AMT_GOODS_PRICE : The price of the goods are mostly 0 - 1 million
- 3. AMT_CREDIT : And so the amount credit is also mostly 0 - 1 million.

4. AMT_ANNUITY : Most of the customurs are paying annuity between 0 - 50000.

Findings 6

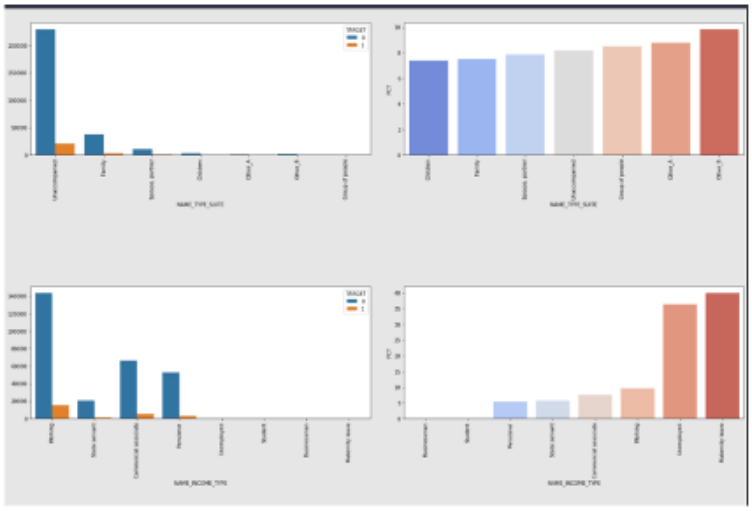
Insights from the different columns



LOAN_TYPE :-
1. The number of cash loan taker is higher than the Revolving loan taker. And also the nomber of defaulter is high in case of cash loan taker

CODE_GENDER :-
2. Female loan takers are higher and also the percentage of female [8%] defaulter is less than male[10%] loan takers.

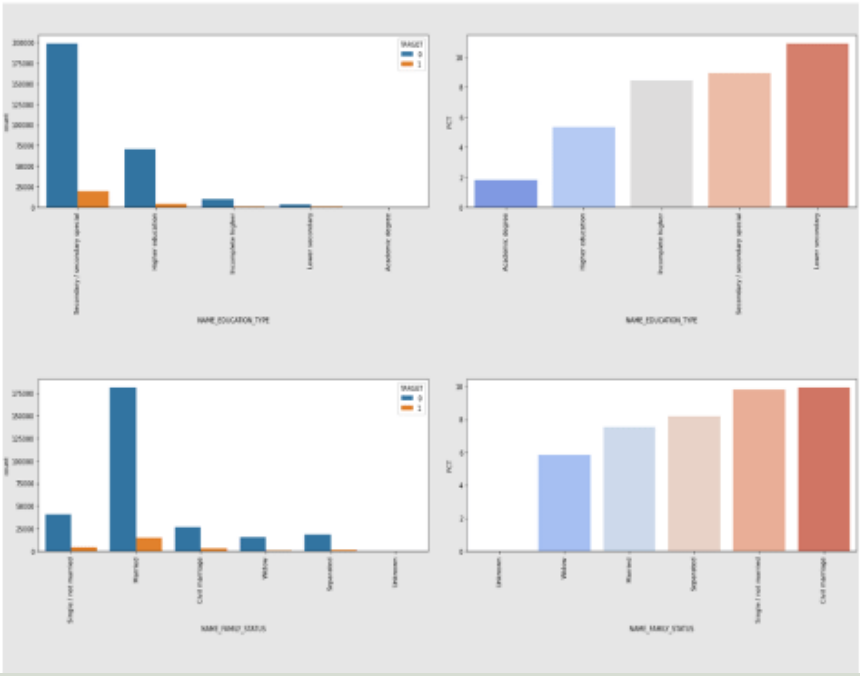
Insights from the different columns



NAME_TYPE_SUIT :-
3. MOst of the time unacompanied people are taking loans and the default rate is 8.5% which is still ok.

NAME_INCOME_TYPE :-
4. Most of the loans are given to the 'working' profesionals followed by 'commercial associate and 'pentioners' and also their defaulter rate is also low.

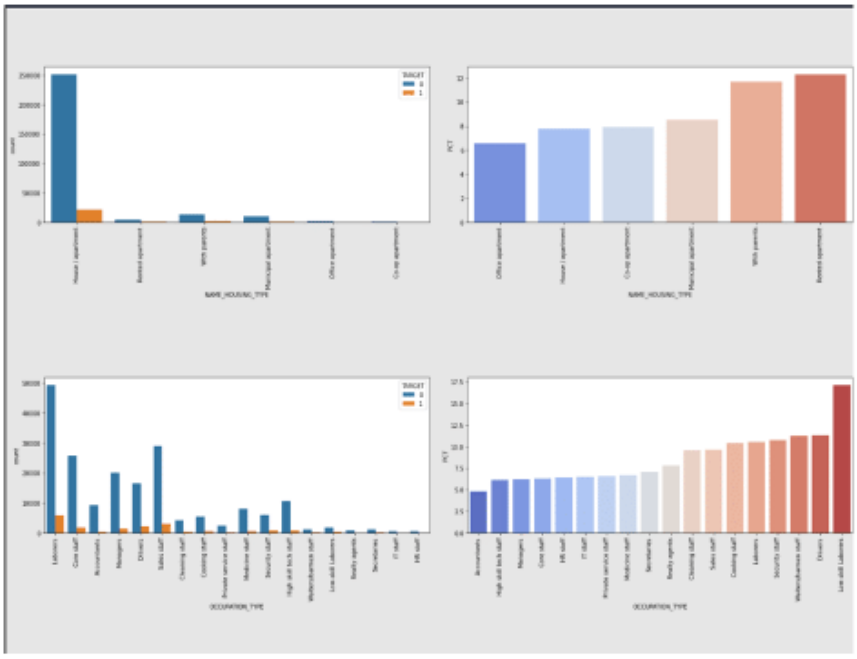
Insights from the different columns



NAME_EDUCATION_TYPE :-
5. Most of the loans are given for the secondary and higher eduction. The default rate for higher education [5%] is less than secondary education loan [9%]. So higher EDUCATIONIS SAFER TO give loans.

NAME_FAMILY_STATUS :-
6. Most of the loans are given to the married people and also their defaulter rate is less than 8% , which is quite acceptable.

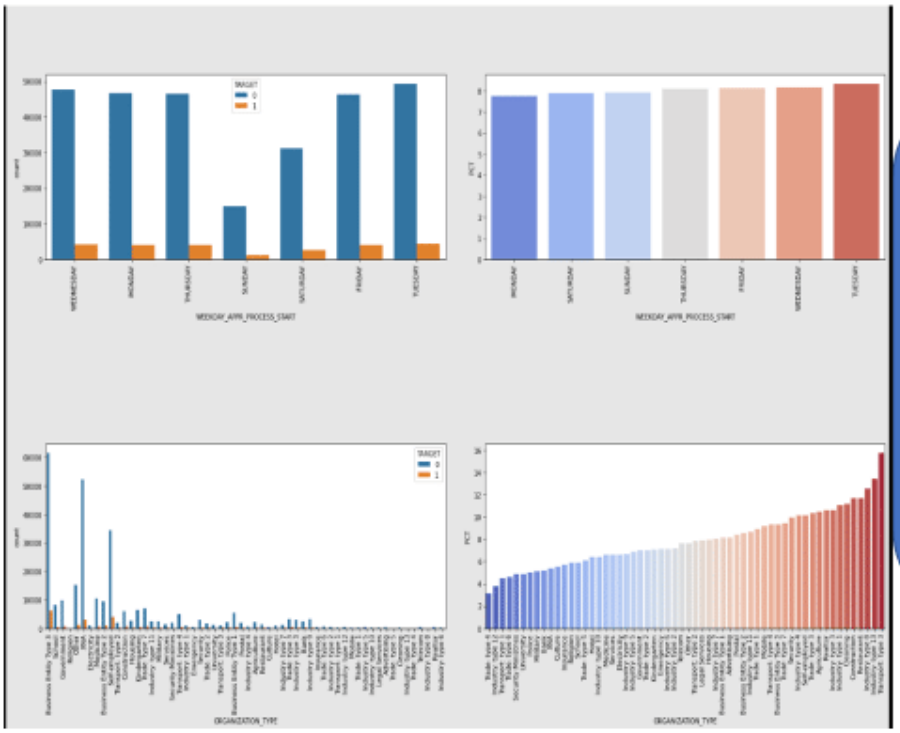
Insights from the different columns



NAME_HOUSING_TYPE :-
7. People havbing own house are safer to give the loan with the defaulter rate of 8%.

OCCUPATION_TYPE :-
8. Low-skilled laboureres and drivers are highest defaulter. Accountants are less defaulters.Number of Labours, core-stuff, manegers are high and also their default rate arein safer side.

Insights from the different columns



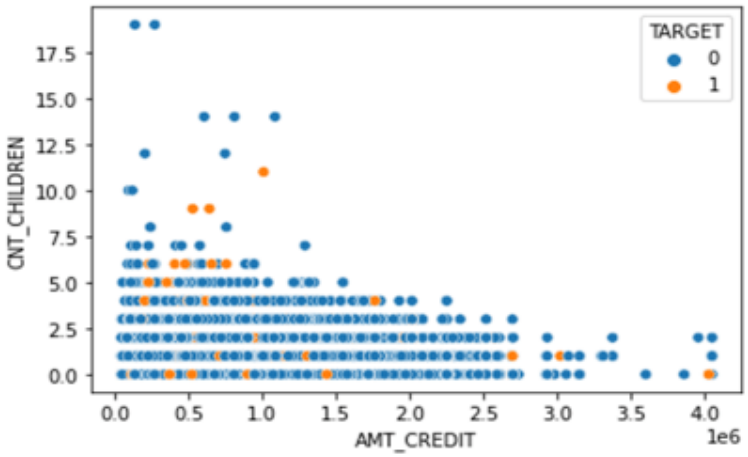
WEEKDAY_APPRVAL_PROCESS_START :-
9. This data will be make not that much sence so we are ignoring this column.

ORGANISTAION_TYPE :-
10. Highest defaulter - Transport type -3 [nuber is very less] XNA are the safest as their number is high and default rate is also less. Bussiness entity 3 , self employed and other also in good number and comperatively safe to give loan.

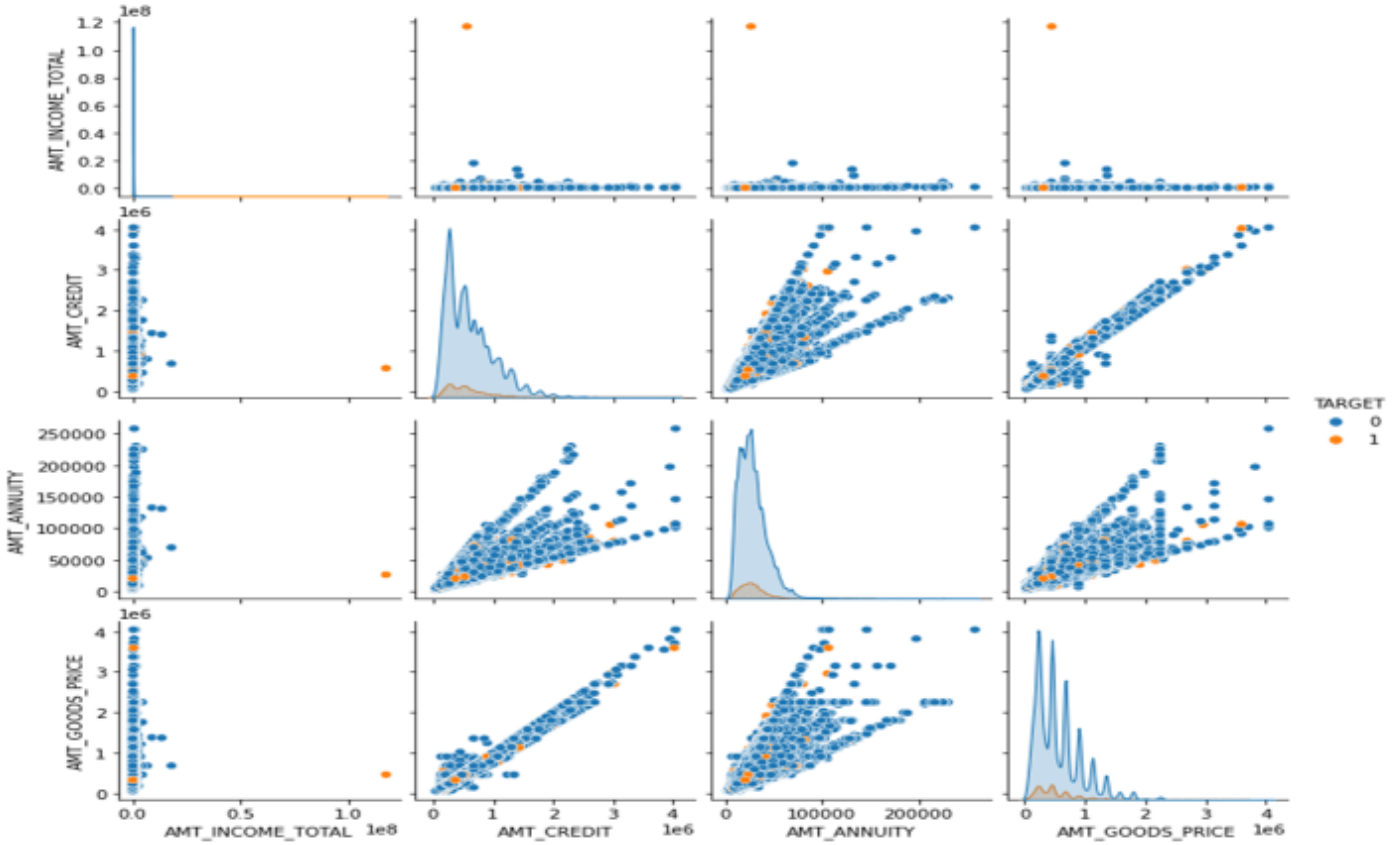
Depending on these analyses company can take the decision that whom to give the loan and whom to not..

Findings 7

Scatter plots between columns (bivariate analysis)



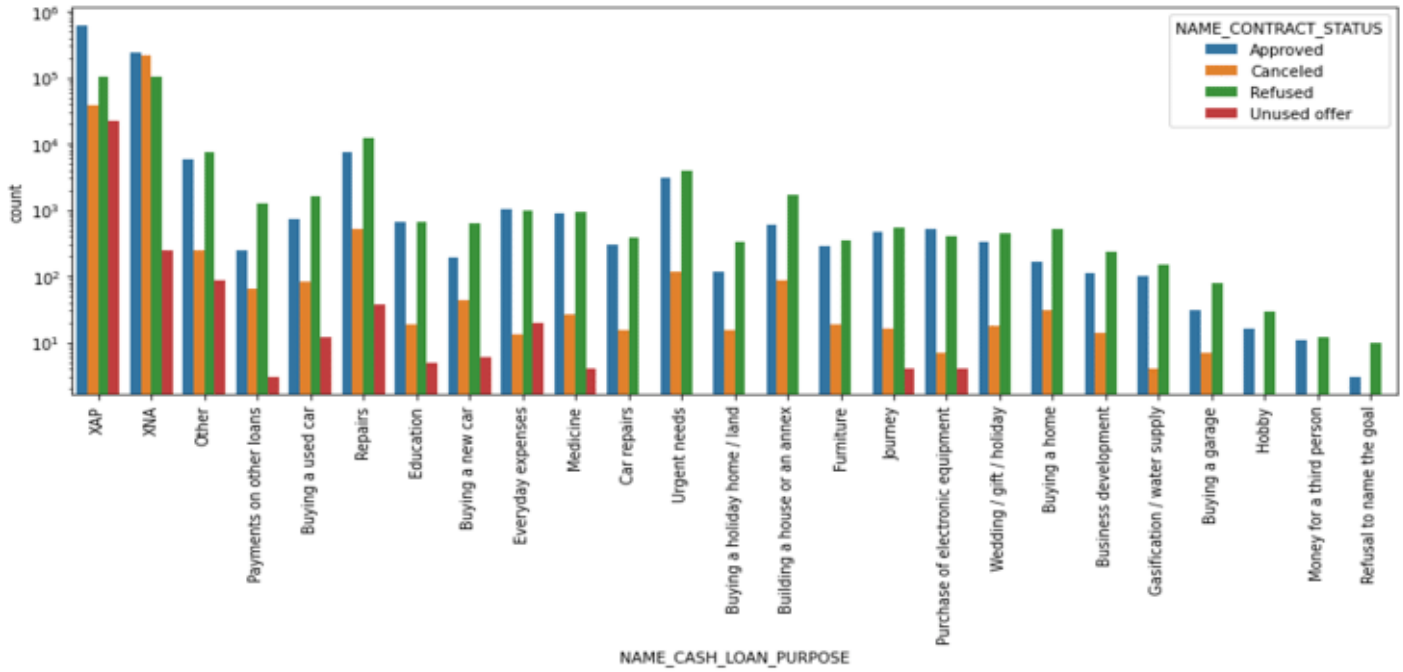
This bi-variate analysis is showing the amount of loans given to the customers according to thee number of their children.



Insights from bi variate analysis

- 1. AMT_CREDIT and AMT_GOODS_PRICE are nearly co-related. If the amount credit increases the number
- of defaulter decreases.
- 2. The people with less annual income (1 milion) are more likely to take loans , out of which
- who are given loans less than 1.5 million are likely to be defaulter.
- 3. Most number of loans are given to the people who have a annuity amount of 100k or less than that.

From the previous application data



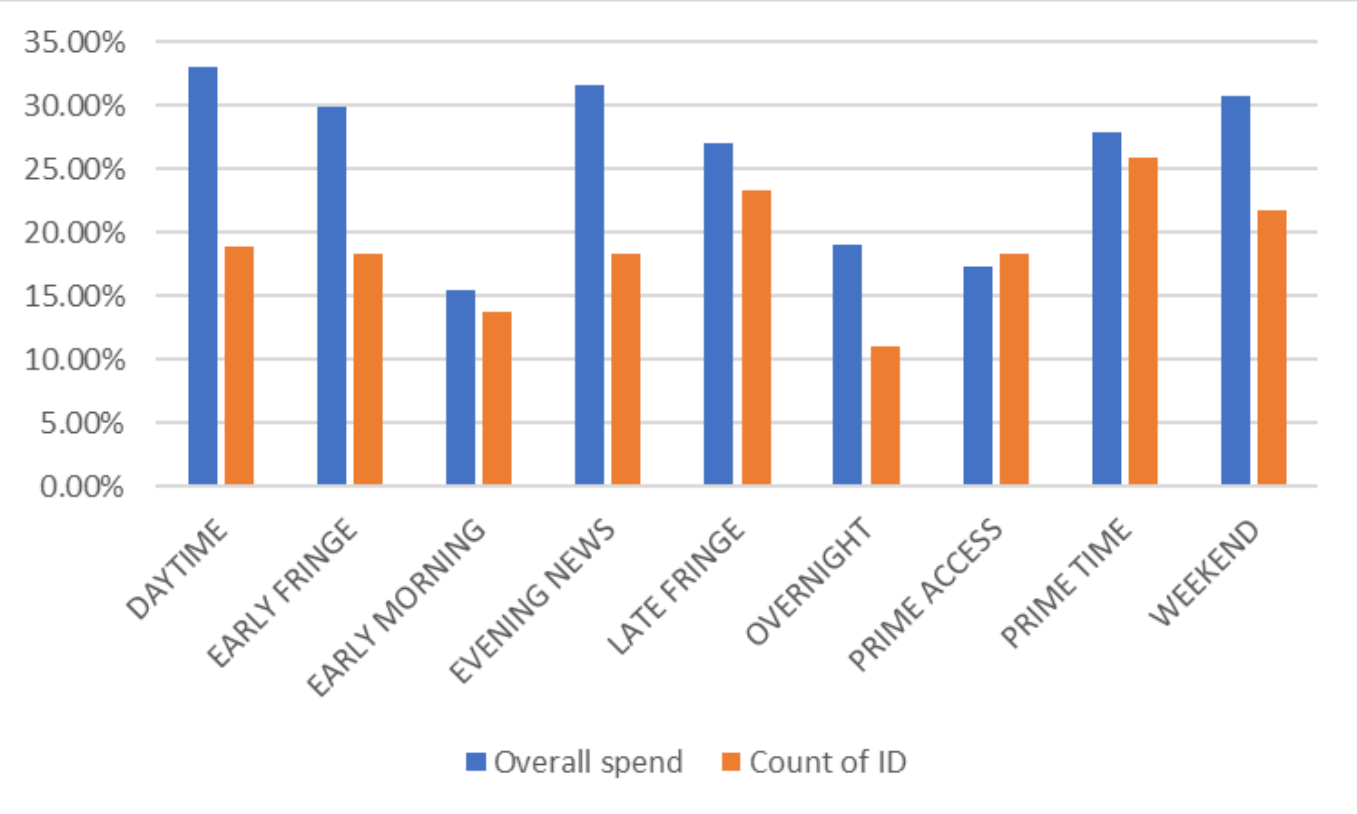
Analysis and conclusion

•This dataset was huge to analyze, but much information was extracted from the dataset. There are many characteristics about the defaulters and repayers, who is safer to give loans, and who is not safer to give loans. And many more insights are given. For further analysis like Machine Learning or any other campaign started by the company like to begin a loan at a low-interest rate, which can be offered to those customers who are no defaulters. And there is a high probability that they will repay the money to the bank.

XYZ Ads Airing Report Analysis

Dataset has different TV Airing Brands, their product, and their category. The dataset includes the network through which Ads are airing, types of networks like Cable/ Broadcast, and the show name on which Ads aired. You can also see the data of Dayparts, Time zone, and the time & date at which Ads got aired. IT also includes other data like Pod Position (the lesser the valuable), the duration for which Ads aired on screen, Equivalent sales &, and the total amount spent on the Ads aired.

Findings 1



There are not that many graphical analyses I did; most of the analyses were in numbers. Only this one was the compare between different day time of overall money spent and also the count of ads were shown.
Other analysis is given in the conclusion sections.

Analysis and conclusion

This analysis was most of the number driven, which was found using excel. Some of them are

1. Pod Position :

Pod position is the number of ad positions for a particular advertisement slot.

Max spend In Pod Position = 1st position => $ 324025029 So the most expensive Pod Position is 1st pod position, So the companies are paying more money for the 1st pod.

And the spending for the 2nd, 3rd, and 4th positions is nearly about same.

2. If we see the quarter-wise money spent we can see, the money spend in the 3rd quarter is less than the others, on the other hand, the money spent in the 1st quarter is the highest.

Maybe because Christmas and new year's celebrations affect the sale of cars.

3. a) Most companies target 'Daytime' and 'Prime Tiem' to run their ads, except Toyota and Honda Cars. Maruti Suzuki is showing most of the ads on tv over the days. They are more relying on 'Prime Time' even more than 'Day Time.' 'Weekend' has a little jump in ads but not more than 'DayTime' for any of the companies.

b) "Maruti Suzuki" and "Mahindra" mostly acquire the 30-sec duration advertisement, where the number of 15 and 20 second durations ads are significantly less compared to 30
second. On the other hand, other companies have more 15-second duration advertisements.

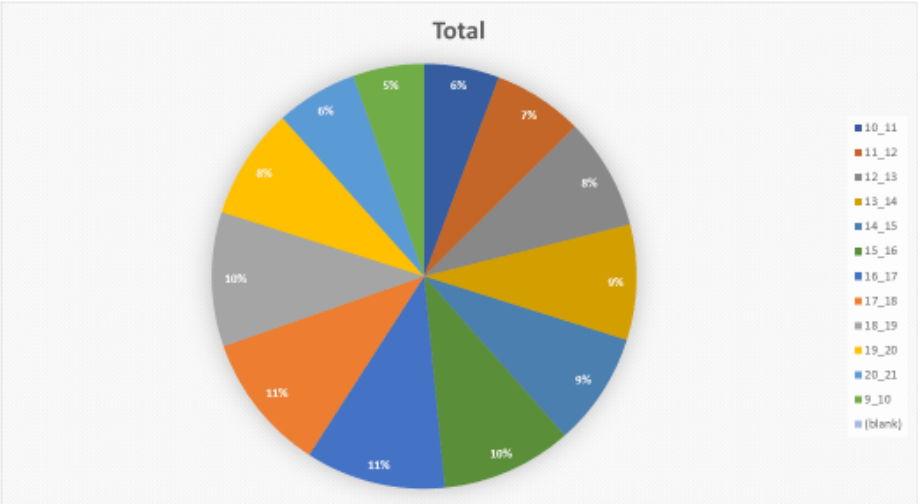
4. Seeing this data, we can say the best time to get the most add in a reasonable spent in 'PRIME TIME' for Mahindra and Mahindra.

ABC Call Volume Trend Analysis

dataset of a Customer Experience (CX) Inbound calling team for 23 days. Data includes Agent_Name, Agent_ID, Queue_Time [duration for which customer have to wait before they get connected to an agent], Time [time at which call was made by customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred). And I have to find some nights and also find some answers to some of the questions asked by the managing team.

Findings 1

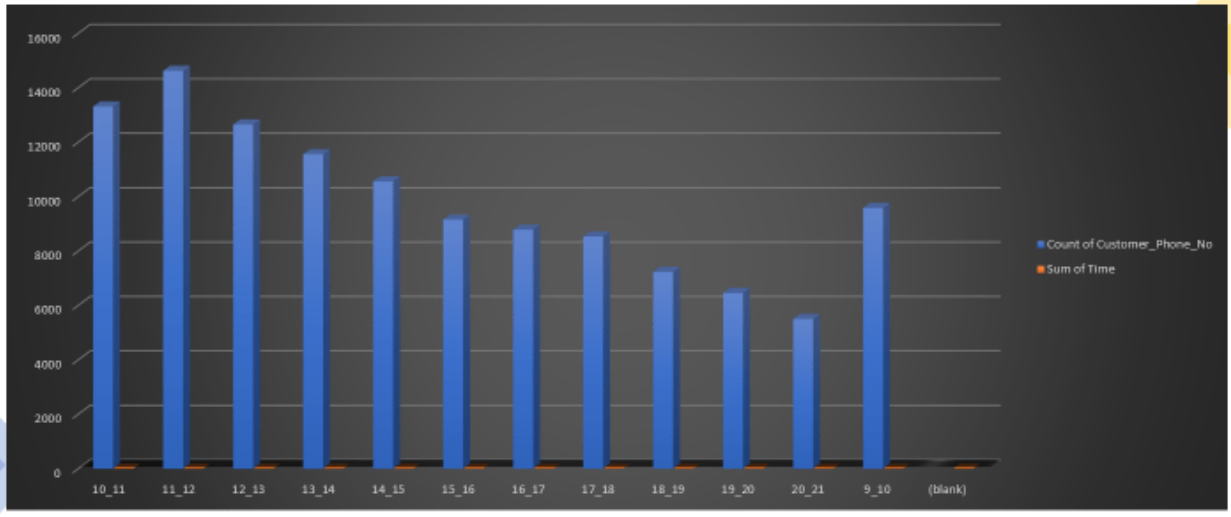
Average call time duration



The time durations are given given in the legend and all the pivot charts are done in exel

Findings 2

Total volume/ number of incoming calls



Analysis and conclusion

In this project, the main problem was managing the call service team so that they can manage to attend to most of the calls 90%. Also, there was no night shift, I had to analyze and give they an employee management plan so that 90% of calls can be attended to along with the constraint that the working time of the employees should not be over time.

Appendix

Link for trainity projects:

https://drive.google.com/drive/folders/1UQicHYwpQpyLXdzo9RGOnj1cImoHUE_A?usp=sharing

Some other works on kaggle:

Krishnendu Barman | Notebooks Novice | Kaggle