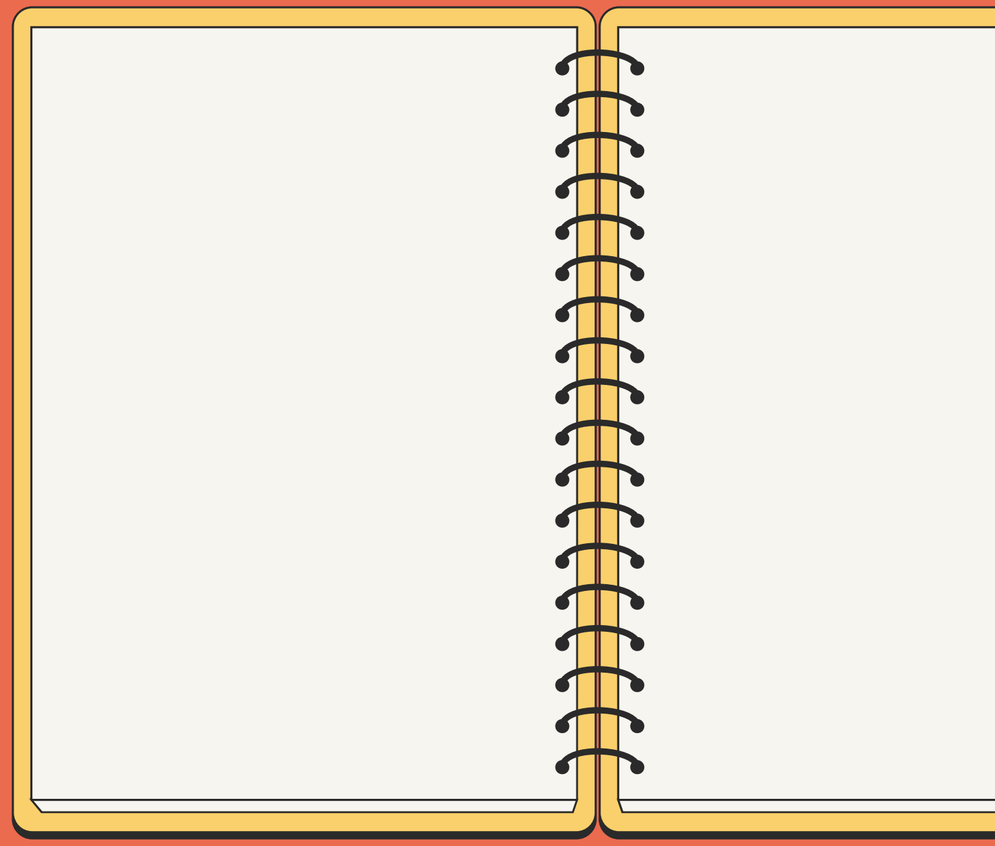
TRAINITY

DATA ANALYTICS PORTFOLIO



AMIT BHAT

PROFESSIONAL BACKGROUND

A BCom graduate in Commerce with a silver medal in my Masters in Globalization and Labour (School of Management & Labour Studies) from Tata Institue of Sciences. I developed an early interest in data and research while pursuing my master's which encouraged me to work on Quantitative Analysis for my Final Year dissertation.

Building on to what I leaned from my dissertation, I subsequently took role in an Analytics company (Ohm Analytics) working as a Senior Research Analyst, at present, where my work primarily involves collection, cleaning and analyzing the data.

Table of Contents

MODULE 1 - PROJECT

REAL LIFE SCANERIO OF DATA ANALYSIS FOR FORMING DECISIONS

PLAN

- Defining the problem statement

In a city like Pune, Public transport is too poor which in turn leads to high cost of travel if one opts for cab aggregators but even that is an issue considering how frequently the ride is cancelled and in some cases hard to find a ride if one is living in outskirts of the city. Taking a view point of a working professional living in outskirts of Pune city, one would mainly face three critical issues in this scenario.

1. High travel cost
2. Wastage of time and energy
3. Frequently being late at office which also adds to mental pressure

In this case, one needs to ask themself questions like does one needs a personal vehicle and if yes by what level it can solve the three problems listed above thereby a decision can bring a satisfactory result. Will a two wheeler or a four wheeler be the best buy to satisfy daily needs.

PREPARE

- Next step would be to decide the financing option to buy a personal two wheeler or four wheeler vehicle. One can ponder over few points like –

1. Taking a full or partial loan.
2. Amount of savings available.
3. If invested amount can be withdrawn to cover the deficit of estimate price of vehicle.
4. Opportunity cost of either paying full money by cash or selling investment.

PROCESS

- Next decision would be to decide whether a two wheeler or a four wheeler proves to be a good fit considering the cost, usage etc.
- If, as per the cost, a two wheeler looks like a probable options then further data gathering would be around two wheelers.
- Data could be gathered around different segments of bike, mileage, performance,amount of interest if opted for loan etc.

ANALYZE

- Analysis would be done based on several parameters under different segments

ADVENTURE
- Mileage
- Cost of bike
- Type of usage
- Level of comfort
- Color, and other specifications
- Spare parts availability

CRUISDER
- Mileage
- Cost of bike
- Type of usage
- Level of comfort
- Color, and other specifications
- Spare parts availability

SPORTS
- Mileage
- Cost of bike
- Type of usage
- Level of comfort
- Color, and other specifications
- Spare parts availability

Cont.

- Someone with heavy daily usage in city would prefer a bike with good mileage and might try to avoid a cruiser bike because of discomfort in maneuvering the bike in city traffic.
- Someone who likes to travel long distance would prefer a cruiser or adventure bike.
- Someone with daily city usage and also occasionally weekend long rides would prefer adventure segment and ignore sports bike because of seating discomfort.

Cont.

- Additionally, one should check what are the current bike in that segment and will there be any upcoming new models.
- If any upgraded models are planned, are the current models expected to continue in production by the company and if not will spare parts be available after the production stops.
- Which segment of bike will better suit one in terms of looks, color, and specifications of bikes etc.

SHARE

- After analyzing one's own needs, it is of prime essence to effectively communicate the specific requirements to the showroom advisor which will ensure that the products best suited are presented and provided for test rides.
- Providing additional details like if someone is naïve in riding, if one is an aggressive rider is also important because it is possible that only bikes with moderate power would be suggested then to avoid any accident that can help with high performance powered bikes.
- One should try to visit different showrooms to compare the similar products and identify price variation if any.

ACT

- The end process would be one finally buying the product considering all the options one had and deciding upon the form of payment.

MODULE 2- PROJECT

INSTAGRAM USER ANALYTICS

PROJECT DECSRIPTION

The following assignment is about analyzing the INSTAGRAM users and providing actionable insights to the "MARKETING TEAM" on areas related to launching advertisement campaigns, and user activity. The project also encapsulates metrics on user engagement to let "INVESTORS" know that if the platform is still relevant and growing in the market and user being active or not.

TECH STACK USED

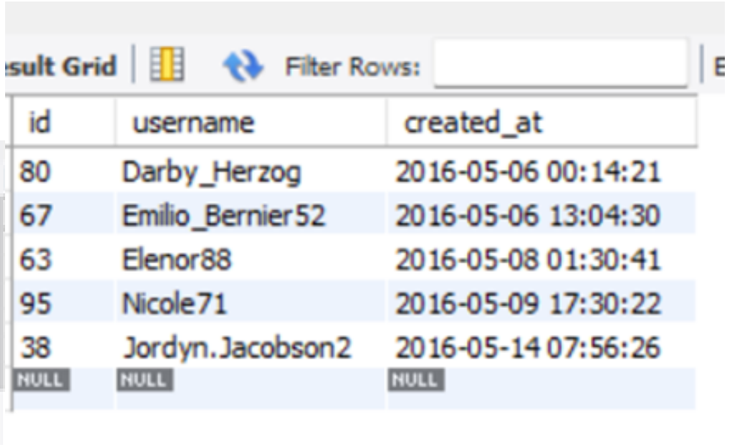
For this project, I have used platform **MYSQL Workbench** and the query language is **MYSQL**.

A) Marketing

- **Rewarding Most Loyal Users:**

```
SELECT *
FROM Users
ORDER BY Created_at ASC
LIMIT 5;
```

The 5 oldest users on the platform are mentioned in the image.



id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26
NULL	NULL	NULL

Cont.

- **Remind Inactive Users to Start Posting**

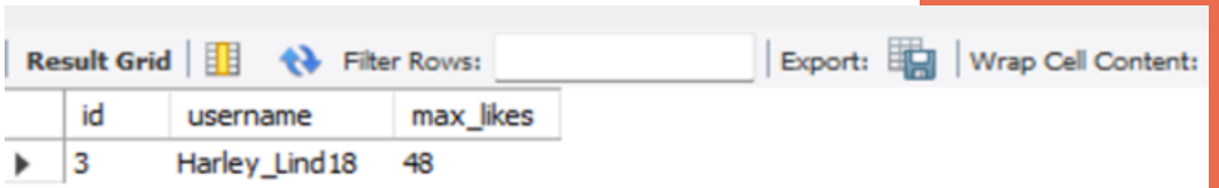
```
SELECT id, username
FROM users
LEFT JOIN photos
ON users.id = photos.user_id
WHERE photos.user_id IS NULL;
```

A total of 26 users are such who have been inactive on the platform.

Cont.

- **Declaring Contest Winner**
- SELECT user.id, user.username, COUNT(photo_likes.photo_id) as max_likes
- FROM Likes photo_likes
- INNER JOIN users user
- ON photo_likes.user_id = user.id
- GROUP BY photo_likes.photo_id
- ORDER BY max_likes DESC
- LIMIT 1;

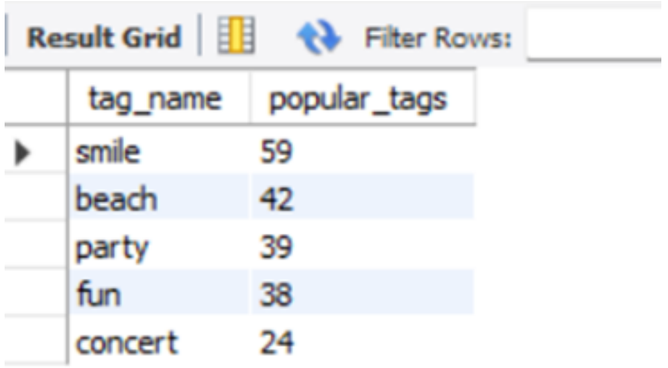
User Harley_lind18 having user id 3
Has got the most number of likes
i.e. 48.



Cont.

- **Hashtag Researching**
- SELECT tag.tag_name, COUNT(tags_used.tag_id) as popular_tags
- FROM photo_tags tags_used
- INNER JOIN tags tag
- ON tags_used.tag_id = tag.id
- GROUP BY tags_used.tag_id
- ORDER BY popular_tags DESC
- LIMIT 5;

The mostly commonly used hashtags in order from highest to lowest are shown in the image.



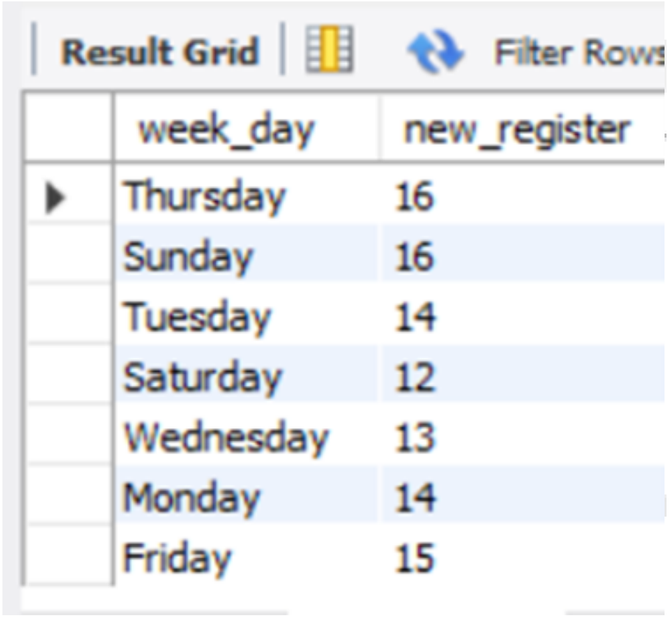
tag_name	popular_tags
smile	59
beach	42
party	39
fun	38
concert	24

Cont.

- **Launch AD Campaign**

- SELECT week_day, COUNT(week_day) as new_register
- FROM (SELECT dayname(created_at) as week_day FROM users) as weekday
- GROUP BY week_day;

To launch an ad campaign, the most Preferred days should be Thursday And Sunday since these have most number of users registrations in the week.



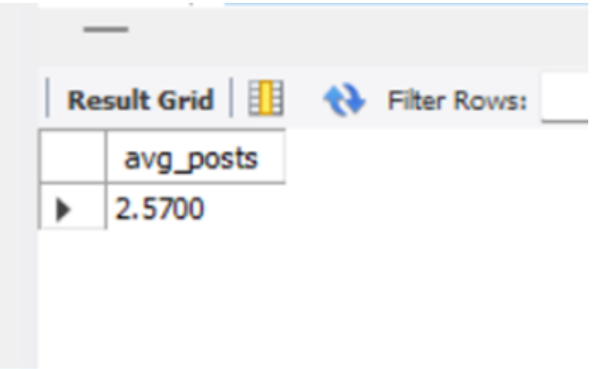
week_day	new_register
Thursday	16
Sunday	16
Tuesday	14
Saturday	12
Wednesday	13
Monday	14
Friday	15

(B)Investor Metrics

- Count of post by users
- SELECT user_id, COUNT(user_id)
- FROM Photos
- GROUP BY user_id;

- AVG photos posted
- SELECT ((SELECT COUNT(pictures_posted.id)
- FROM photos pictures_posted)/(SELECT COUNT(total_users.id)
- FROM users total_users)) AS avg_posts;

Average number of photos posted per user in the Platform us 2.57 .

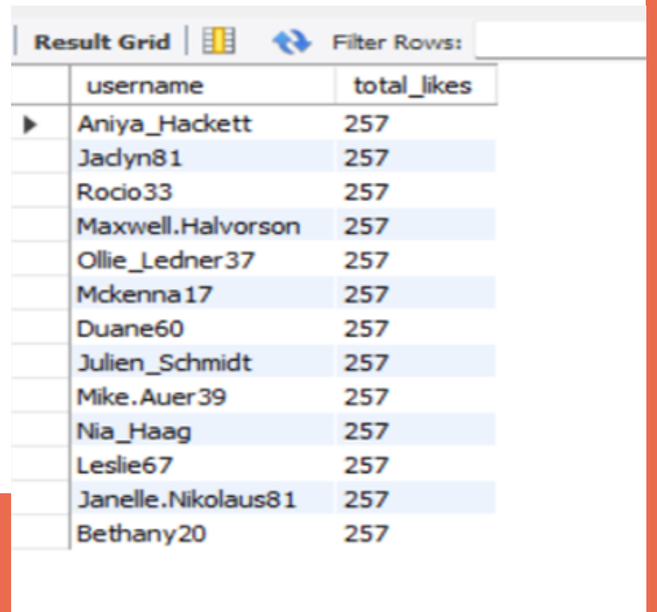


Cont.

- **Bots & Fake Accounts**

```
SELECT username, COUNT(likes.user_id) as total_likes
FROM users
INNER JOIN Likes
ON id = user_id
GROUP BY likes.user_id
HAVING total_likes = (SELECT COUNT(photos.id) FROM photos);
```

TOTAL OF 13 users are such who have liked Every single photo posted on the platform and are Termed as bots.



Result Grid | Filter Rows:

username	total_likes
Aniya_Hackett	257
Jadyn81	257
Rocio33	257
Maxwell.Halvorson	257
Ollie_Ledner37	257
Mckenna17	257
Duane60	257
Julien_Schmidt	257
Mike.Auer39	257
Nia_Haag	257
Leslie67	257
Janelle.Nikolaus81	257
Bethany20	257

DRIVE LINK FOR PPT & SQL FILE

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1KD5DYBPOOSFDUYCLSTM2U1A_OSD3S6JJ?USP=SHARING

MODULE 3- PROJECT

Operation Analytics and Investigating Metric Spike

PROJECT DECSRIPTION

The following assignment is about Operations analytics and metric spike thereby analyzing the JOB DATA for Microsoft and providing insights on areas related number of jobs reviewed, and other events happening around it. The project also encapsulates analyzing the users activity at various stages of engagement.

TECH STACK USED

For this project, I have used platform **MYSQL Workbench** and the query language is **MYSQL.**

Job Data Analysis

- **Number of jobs reviewed:**

- SELECT
  ds, job_id,
  (COUNT(Job_id)/SUM(time_spent))/3600
- FROM job_data
- GROUP BY ds;

cont.

- **Throughput:**

- SELECT

  SUM(event_persecond) OVER(ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) as moving_average
- FROM (SELECT ds, COUNT(event)/SUM(time_spent) as event_persecond FROM job_data GROUP BY ds ORDER BY ds) through_put;

Cont.

- **Percentage share of each language**

- SELECT Language, (COUNT(language)/total_languages)*100 as percent_language
- FROM job_data
- CROSS JOIN
  (SELECT COUNT(DISTINCT language) as Total_languages
- FROM job_data) b
- GROUP BY Language;

Cont.

- **Duplicate rows:**

- SELECT *
- FROM (SELECT*, ROW_NUMBER() OVER(PARTITION BY job_id ORDER BY event) as dup_job FROM job_data) job_part
- WHERE dup_job >1;

Investigating metric spike

- **User Engagement:**

- SELECT EXTRACT(WEEK FROM occurred_at) as weekly_data, COUNT(User_id)
- FROM events
- WHERE event_type = 'engagement'
- GROUP BY weekly_data ;

Cont.

- **User Growth:**

- SELECT Year, Month, SUM(Monthly_user_growth) OVER(ORDER BY Year ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) as Cum_Growth_overtime
- FROM (SELECT COUNT(user_id) as Monthly_user_growth, EXTRACT(month FROM created_at) as Month, EXTRACT(YEAR FROM created_at) as Year FROM users GROUP BY Month, Year) as Monthly_breakup;

Cont.

- **Weekly Retention**

- SELECT COUNT(User_id) active_weekly_usercount, EXTRACT(Year FROM created_at) as yearly, EXTRACT(Week FROM created_at) as weekly
- FROM users
- WHERE State = 'active'
- GROUP BY yearly, weekly;

Cont.

- **Weekly Engagement:**

- SELECT COUNT(User_id) weekly_volumne, device, EXTRACT(Year FROM occurred_at) as yearly, EXTRACT(Week FROM occurred_at) as weekly
- FROM events
- WHERE event_type = 'engagement'
- GROUP BY device;

Cont.

- **Email Engagement:**

- SELECT EXTRACT(Week FROM occurred_at) as Week,
- COUNT(CASE WHEN action = 'sent_weekly_digest' THEN 1 ELSE NULL END) AS weekly_emails_sent,
- COUNT(CASE WHEN action = 'email_open' THEN 1 ELSE NULL END) AS weekly_emails_open,
- COUNT(CASE WHEN action = 'sent_reengagement_email' THEN 1 ELSE NULL END) AS weekly_reengagement_email,
- COUNT(CASE WHEN action = 'email_clickthroughs' THEN 1 ELSE NULL END) AS weekly_emails_click
- FROM email_events
- GROUP BY Week;

DRIVE LINK FOR PPT & SQL FILE

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1EWS1IG7HJZNS8X16LXYQ_YCMZDUHAE9-?USP=SHARING

MODULE 4- PROJECT

HIRING PROCESS ANALYTICS

PROJECT DECSRIPTION

The following assignment is about hiring process analytics which is a crucial element of every organization. It analysis various matric related to salary, male female participation, Departmental employees being hired etc.

TECH STACK USED

For this project, I have used MS EXCEL for performing all the analysis using functions like pivot tables, averages etc.

How many males and females are Hired ?

- A total of 1856 females and 2563 male candidates we hired

Row Labels	Count of application_id
-	10
Don’t want to say	268
Female	1856
Male	2563
Grand Total	**4697**

What is the average salary offered in this company ?

- Average salary breakdown offered by the company is

Row Labels	Average of Offered Salary
b9	49666.76458
c-10	51134.62069
c5	50213.50372
c8	50701.4625
c9	50201.18583
i1	49943.93694
i4	48877.84091
i5	49391.92503
i6	48839.24858
i7	50065.36086
m6	34521.33333
m7	41402
n10	26990
n6	44700
n9	46219

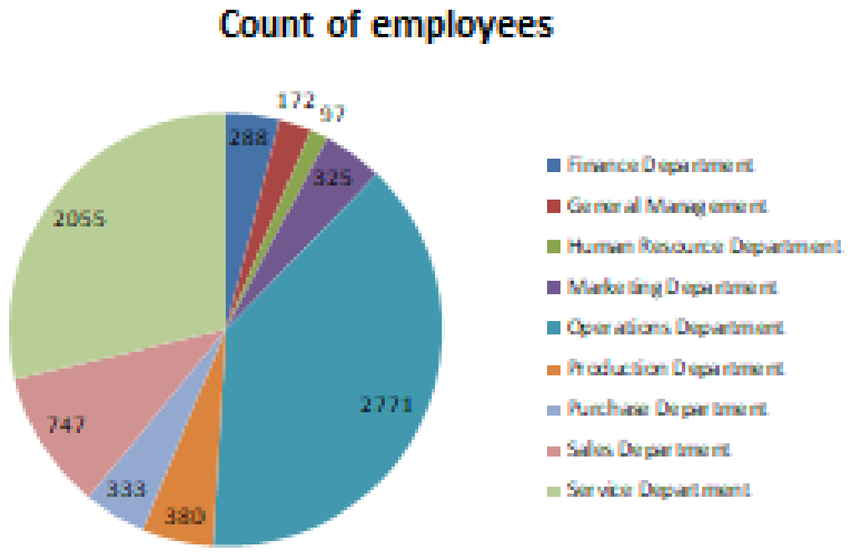
Draw the class intervals for salary in the company ?

- Class Interval for the salary is 399900

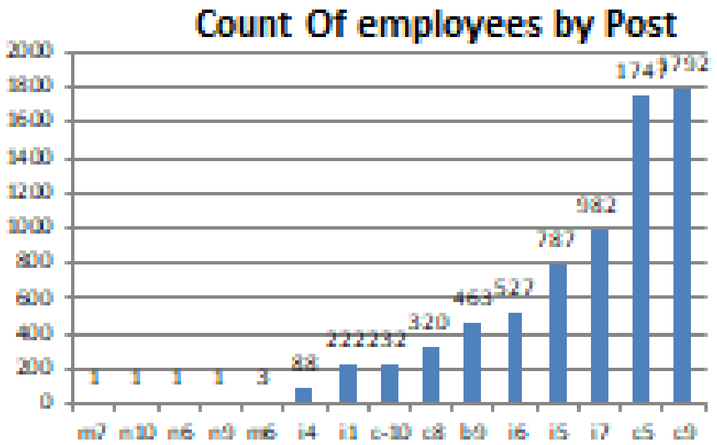
Maximum Salary	Minimum Salary	Class Interval
400000	100	399900

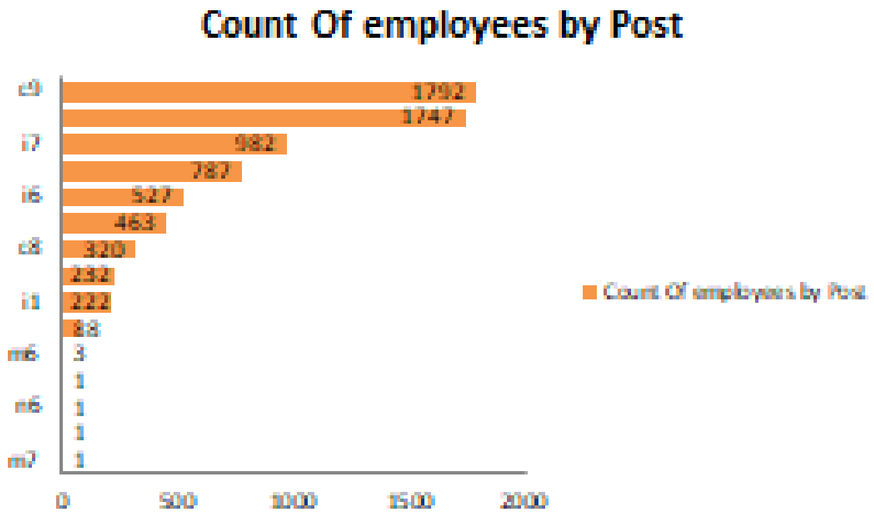
Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?

- Majority of people are working in Finance department.



Represent different post tiers using chart/graph?





DRIVE LINK FOR PPT & EXCEL SOLVED FILES

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1FWAWMWPQVBPNWBOODB0ZSRIOA4ROPTZQ?USP=SHARING

MODULE 5- PROJECT

IMDB Movie Analysis

PROJECT DECSRIPTION

The following assignment is based on analysis the imdb movie data to figure out what is working in the industry, what users like the movie, who are some fo the popular actor and some popular movies in the movie world.

TECH STACK USED

For this project, I have used MS EXCEL for performing all the analysis using functions like pivot tables, averages, sum, Vlookup count etc.

Cleaning the data

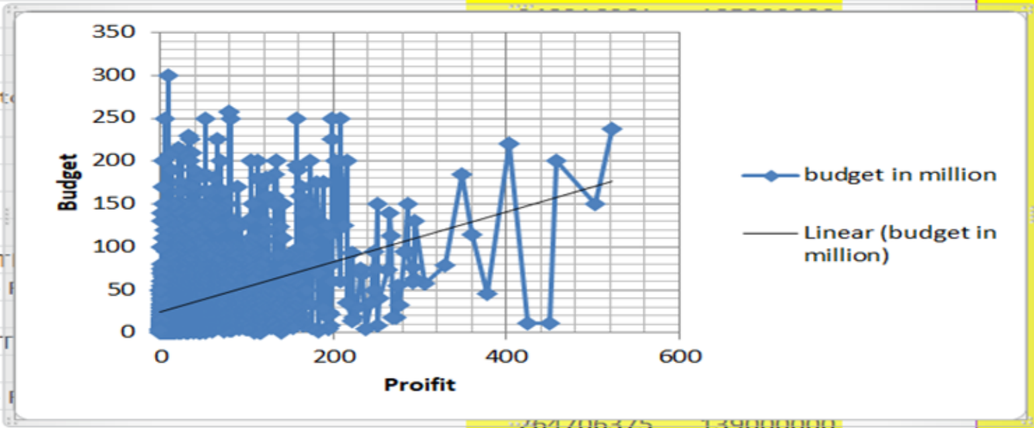
- Data cleaning method like deleting the unnecessary columns, deleting the blank cells and duplicates as well if any.

Note: Data cleaning for this project was majorly done for every questions separately in order to keep the maximum data available for every question.

Movies with highest profit

- Movie with the highest profit is Avatar

AE	AF	AG	AH	AI	AJ
Movie_Ttile	Profit	budget		Profit in million	budget in million
AvatarÂ	523505847	237000000		523.505847	237
Jurassic WorldÂ	502177271	150000000		502.177271	150
TitanicÂ	458672302	200000000		458.672302	200
Star Wars: Episode IV - A New HopeÂ	449935665	11000000		449.935665	11
E.T. the Extra-TerrestrialÂ	424449459	10500000		424.449459	10.5
The AvengersÂ	403279547	220000000		403.279547	220
The AvengersÂ	403279547	220000000		403.279547	220
The Lion KingÂ	377783777	45000000		377.783777	45
Star Wars: Episode I - The Phantom MenaceÂ	359544677	115000000		359.544677	115
The Dark KnightÂ				348.316061	185
The Hunger GamesÂ				329.999255	78
DeadpoolÂ				305.024263	58
The Hunger Games: Cat				294.645577	130
Jurassic ParkÂ				293.784	63
Despicable Me 2Â				292.049635	76
American SniperÂ				291.323553	58.8
Finding NemoÂ				286.83887	94
Shrek 2Â				286.471036	150
The Lord of the Rings: T				283.019252	94
Star Wars: Episode VI -				276.625409	32.5
Forrest GumpÂ				274.691196	55
Star Wars: Episode V - T				272.158751	18
Home AloneÂ				267.761243	18
Star Wars: Episode III -				267.262555	113
Spider-ManÂ	264706375	139000000		264.706375	139
MinionsÂ	262029560	74000000		262.02956	74
The Sixth SenseÂ	253501675	40000000		253.501675	40
JawsÂ	252000000	8000000		252	8
FrozenÂ	250736600	150000000		250.7366	150



Top 250

- List of 250 imdb movies num_voted_users is greater than 25,000 with rank given

	movie_title	language	imdb_score	IMDb_Top_250_Rank	
	The Shawshank RedemptionÂ	English	9.3	1	
	The GodfatherÂ	English	9.2	2	
	The Dark KnightÂ	English	9	3	
	The Godfather: Part IIÂ	English	9	3	
	FargoÂ	English	9	3	
	The Lord of the Rings: The Return of the KingÂ	English	8.9	6	
	Schindler's ListÂ	English	8.9	6	
	Pulp FictionÂ	English	8.9	6	
0	The Good, the Bad and the UglyÂ	Italian	8.9	6	
1	12 Angry MenÂ	English	8.9	6	
2	InceptionÂ	English	8.8	11	
3	The Lord of the Rings: The Fellowship of the RingÂ	English	8.8	11	
4	DaredevilÂ	English	8.8	11	
5	Fight ClubÂ	English	8.8	11	
5	Forrest GumpÂ	English	8.8	11	
7	It's Always Sunny in PhiladelphiaÂ	English	8.8	11	
3	Star Wars: Episode V - The Empire Strikes BackÂ	English	8.8	11	
9	The Lord of the Rings: The Two TowersÂ	English	8.7	18	
0	The MatrixÂ	English	8.7	18	
1	Friday Night LightsÂ	English	8.7	18	
2	GoodfellasÂ	English	8.7	18	
3	Star Wars: Episode IV - A New HopeÂ	English	8.7	18	
4	One Flew Over the Cuckoo's NestÂ	English	8.7	18	
5	City of GodÂ	Portuguese	8.7	18	
5	Seven SamuraiÂ	Japanese	8.7	18	
7	InterstellarÂ	English	8.6	26	
3	HannibalÂ	English	8.6	26	
9	Saving Private RyanÂ	English	8.6	26	
0	LutherÂ	English	8.6	26	
1	Se7enÂ	English	8.6	26	

Top_Foreign_Lang_Film

Top_Foreign_Lang_Film	language	imdb_score	IMDb_Top_250
The Good, the Bad and the UglyÂ	Italian	8.9	6
City of GodÂ	Portuguese	8.7	18
Seven SamuraiÂ	Japanese	8.7	18
Spirited AwayÂ	Japanese	8.6	26
AirliftÂ	Hindi	8.5	40
The Lives of OthersÂ	German	8.5	40
Children of HeavenÂ	Persian	8.5	40
AmÃ©lieÂ	French	8.4	62
Baahubali: The BeginningÂ	Telugu	8.4	62
Princess MononokeÂ	Japanese	8.4	62
Das BootÂ	German	8.4	62
Rang De BasantiÂ	Hindi	8.4	62
OldboyÂ	Korean	8.4	62
A SeparationÂ	Persian	8.4	62
MetropolisÂ	German	8.3	86
DownfallÂ	German	8.3	86
The HuntÂ	Danish	8.3	86
Howl's Moving CastleÂ	Japanese	8.2	117
Pan's LabyrinthÂ	Spanish	8.2	117
IncendiesÂ	French	8.2	117
The Secret in Their EyesÂ	Spanish	8.2	117
Lage Raho Munna BhaiÂ	Hindi	8.2	117
SolarisÂ	Russian	8.1	147
The Sea InsideÂ	Spanish	8.1	147
Tae Guk Gi: The Brotherhood of WarÂ	Korean	8.1	147
AkiraÂ	Japanese	8.1	147
Elite SquadÂ	Portuguese	8.1	147
Amores PerrosÂ	Spanish	8.1	147
The CelebrationÂ	Danish	8.1	147
The ReturnÂ	Russian	8	203
The Diving Bell and the ButterflyÂ	French	8	203
My Name Is KhanÂ	Hindi	8	203
PersepolisÂ	French	8	203

Best Directors

- After cleaning the data, pivot table was used to group the directors and to calculate mean of score, average function in pivot was used

director_name	mean_imdb_score	Rank
John Blanchard	9.5	1
Cary Bell	8.7	2
Mitchell Altieri	8.7	3
Sadyk Sher-Niyaz	8.7	4
Charles Chaplin	8.6	5
Mike Mayhall	8.6	6
Damien Chazelle	8.5	7
Majid Majidi	8.5	8
Raja Menon	8.5	9
Ron Fricke	8.5	10

Popular Genres

- Post cleaning, pivot was used to group the genres and then to find popularity I assumed that out of all the column, number of Facebook likes will provide more accurate result of the popularity for the movies and their gneres.

G	H	I
genres	**Mean_movie_facebook_likes**	
Adventure|Drama|Thriller|Western	190000	
Adventure|Comedy|Crime|Drama	149000	
Adventure|Drama|Sci-Fi	128466	
Adventure|Animation|Comedy|Drama|Family|Fantasy	118000	
Crime|Drama|Mystery|Thriller|Western	114000	
Action|Biography|Drama|History|Thriller|War	112000	
Adventure|Comedy|Drama|Fantasy|Musical	90000	
Action|Adventure|Family|Fantasy|Romance	89000	
Biography|Drama|Thriller|War	82500	
Biography|Crime|Drama|History|Music	76000	
Action|Horror|Romance	73000	
Adventure|Drama|Sci-Fi|Thriller	71750	
Action|Crime|Drama|Sci-Fi|Thriller	71000	
Adventure|Comedy|Drama|Fantasy|Romance	70000	
Action|Biography|Drama|Thriller|War	58000	
Action|Adventure|Comedy|Sci-Fi	55400	
Action|Adventure|Crime|Drama|Sci-Fi|Thriller	55000	
Adventure|Drama|Fantasy	49614	
Biography|Comedy|Crime|Drama	46000	
Biography|Drama|History|Thriller	44500	
Popularity I assumed based on facebook likes and not imdb rating		

Charts-Popular Actor

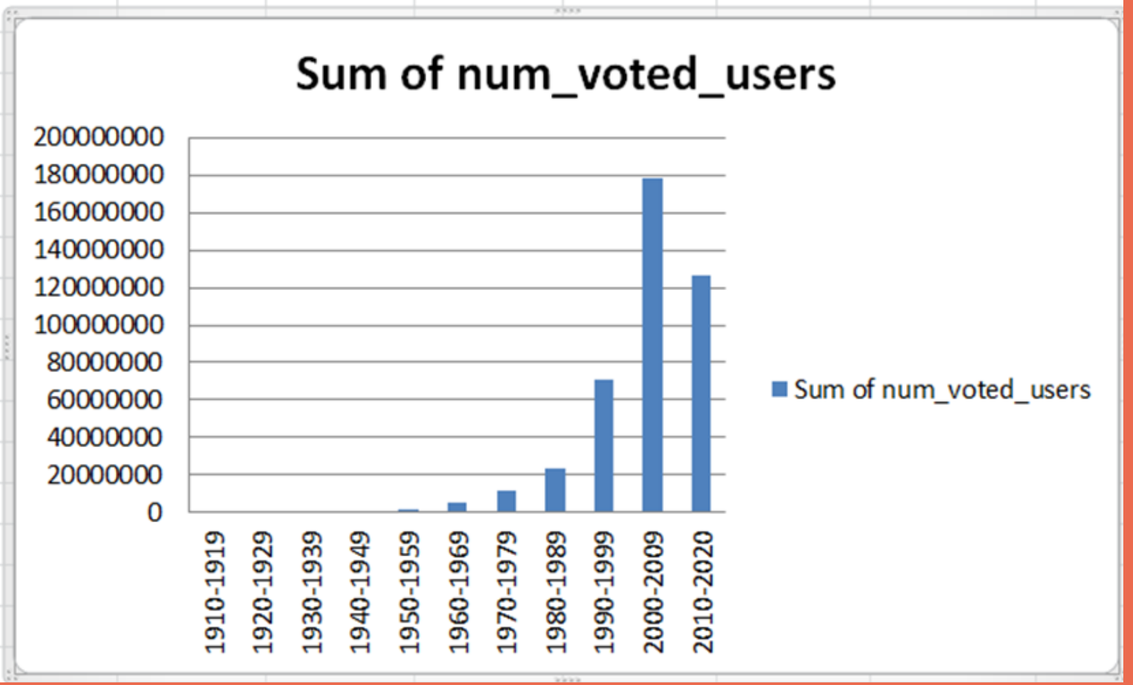
- Leonardo DiCaprio appears to be the actor with highest mean for number of critics for reviews and number of user for reviews.
- Again, for calculation this, after cleaning the data, for these three actors pivot was used to group the actors and then average function used in value section of pivot.

Actor	Average of num_critic_for_reviews	Average of num_user_for_reviews
Brad Pitt	231.9444444	702.4444444
Leonardo DiCaprio	330.1904762	914.4761905
Meryl Streep	163.1538462	257.3076923

Charts-Decade user vote

- Number of voted users seems to rise after every decade and was at its peak in 20s (2000 till 2009) post which it saw a decline.
- This also was calculated using group function in pivot setting a value of 10 for years to group it as decade and then using sum function to calculate the voted users.

Decade	Sum of num_voted_users
1910-1919	10718
1920-1929	128672
1930-1939	984397
1940-1949	1211888
1950-1959	1638504
1960-1969	5153052
1970-1979	11312705
1980-1989	23176169
1990-1999	70633270
2000-2009	178354686
2010-2020	126202706



DRIVE LINK FOR PPT & EXCEL SOLVED FILES

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1S-OOXGGA0BEFGBFLLU-PUJ6TEDHTGZXS?USP=SHARING

MODULE 6- PROJECT

Bank Loan Case Study

- **PROJECT DECSRIPTION**

The following assignment is based on bank loan case study data which tries to figure under various conations an applicant is likely to default out and what are the indicators under which an applicant would not default. This will ensure that base don variables explained, only right candidates are approved of loan so there is no loss of business.

- **TECH STACK USED**
- For this project, I have used MS EXCEL for performing all the analysis using functions like pivot tables, averages, sum, Vlookup, Count, Mean, Median etc.

- **Note that, it is not possible to explain every step in the pdf hence I have included the workbooks worked upon which has all the formula and different sheets for different analysis that was performed.**

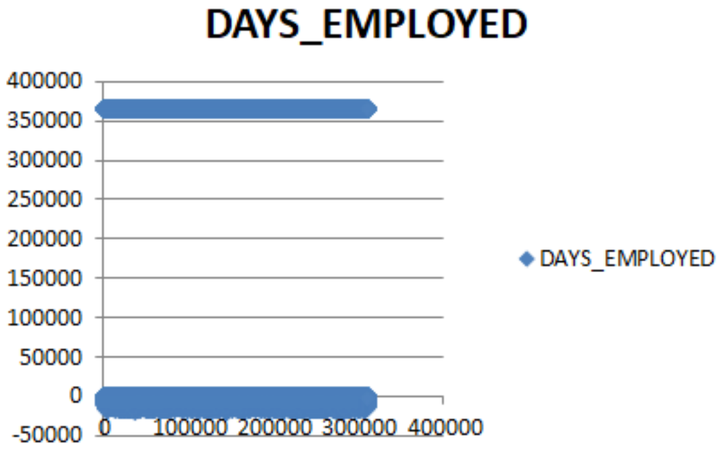
Identify the missing data

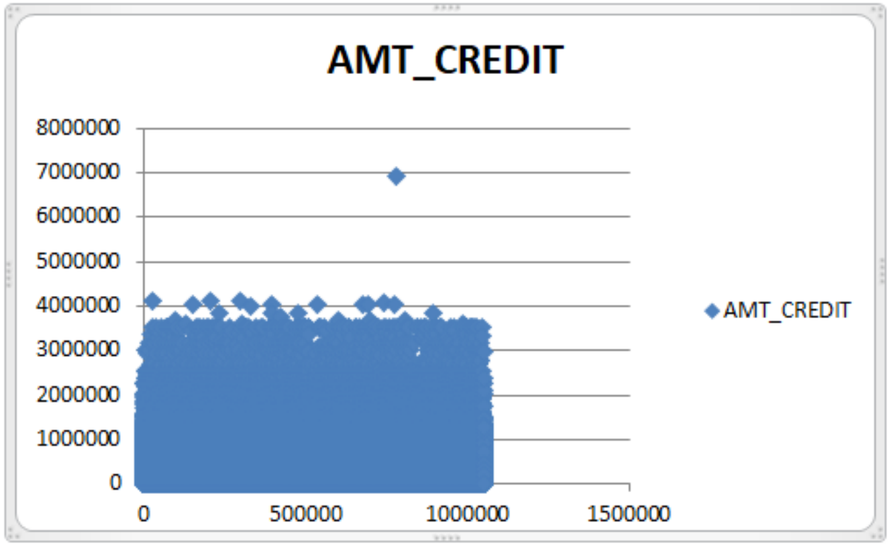
- Percentage of missing data was identified using the **count function** in excel and the dividing that with the total count of data points in that column. For example - **=COUNTBLANK(B5:B1048576).**
- Once missing values identified, column with high percentage of missing value i.e more than 40% is deleted, in this exercise I have sorted the data **(Using sort function for columns)** so the higher missing percentage data column are on the right side of the table and not deleted them.
- Column with missing value of less than 40% was replaced with mean or median in continuous variable. IN some cases, data is skewed as proven by outliers, this Median was used to fill the missing data points.

Identify if there are outliers in the dataset

- Creating scatter plot for identifying various outliers

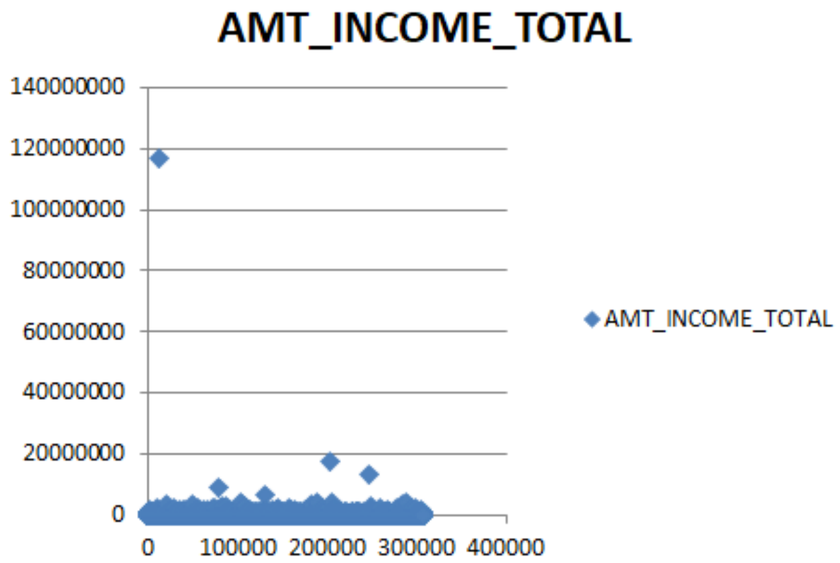
Value of 350000 is clearly an outlier in the column days_employed



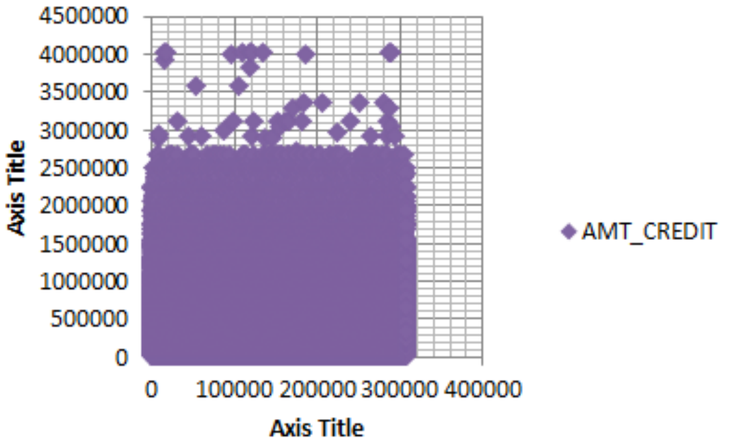


Value of 7000000 is clearly an outlier in the column AMT_CREDIT.

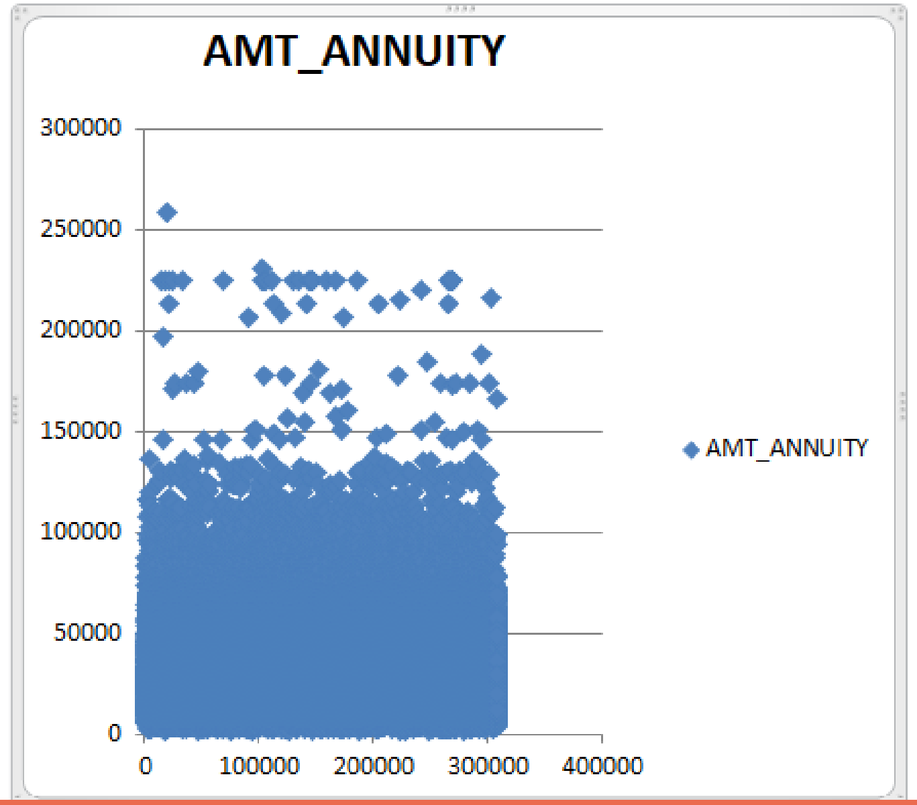
Median could be used to fill the blanks in this

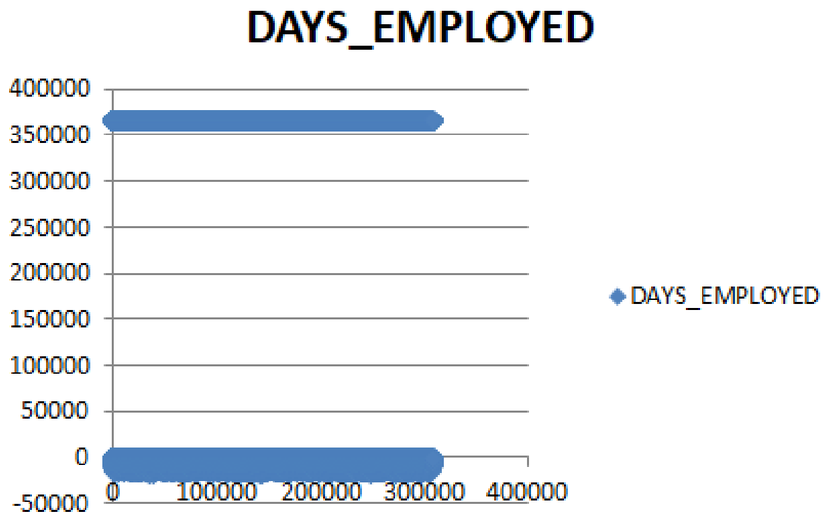


Income above 4 lakh is an outlier in the column

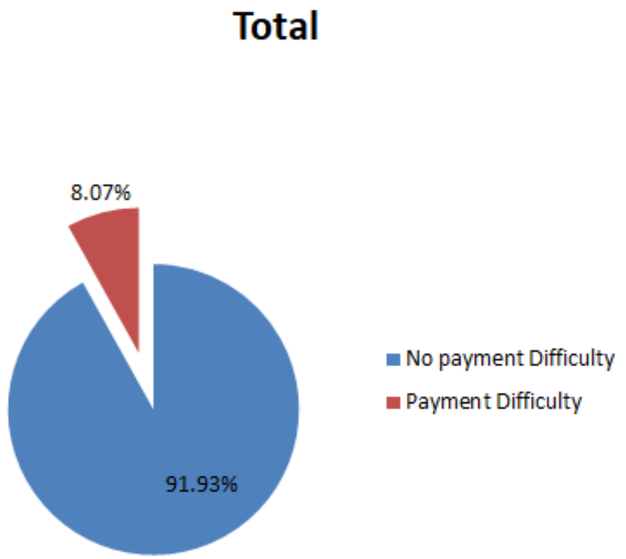


Anything over 25 lakh is an outlier which make data right skewed in the column amount of credit asked for

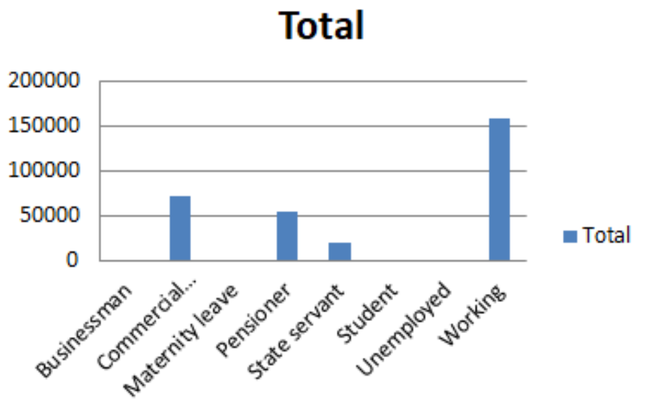

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
0
100000
200000
300000
400000
AMT_ANNUITY


DAYS_EMPLOYED
400000
350000
300000
250000
200000
150000
100000
50000
0
-50000
0
100000
200000
300000
400000
DAYS_EMPLOYED

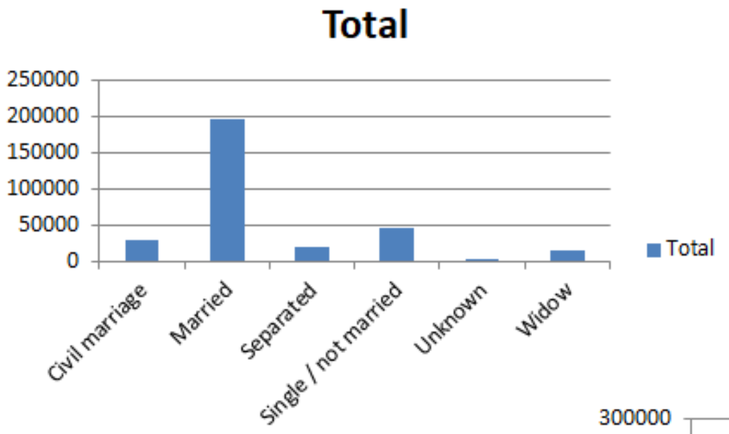
Data imbalance



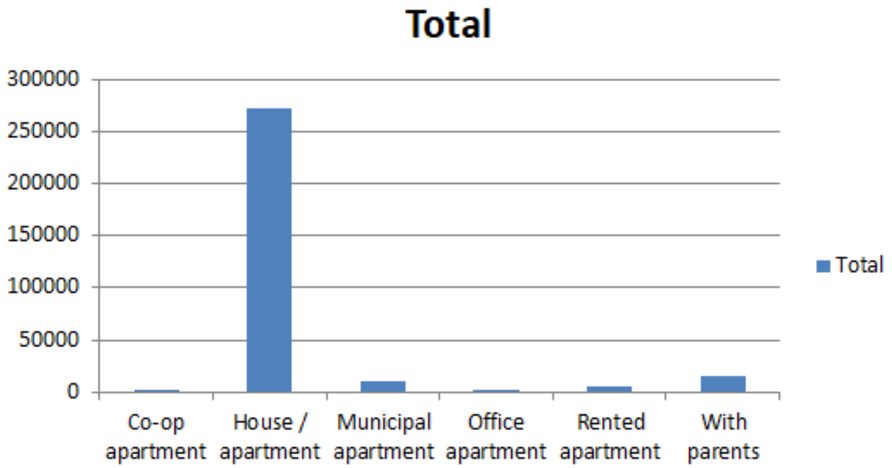
There is a high percentage of people with having no payment difficulty



Data sample is highly loaded with working class people.

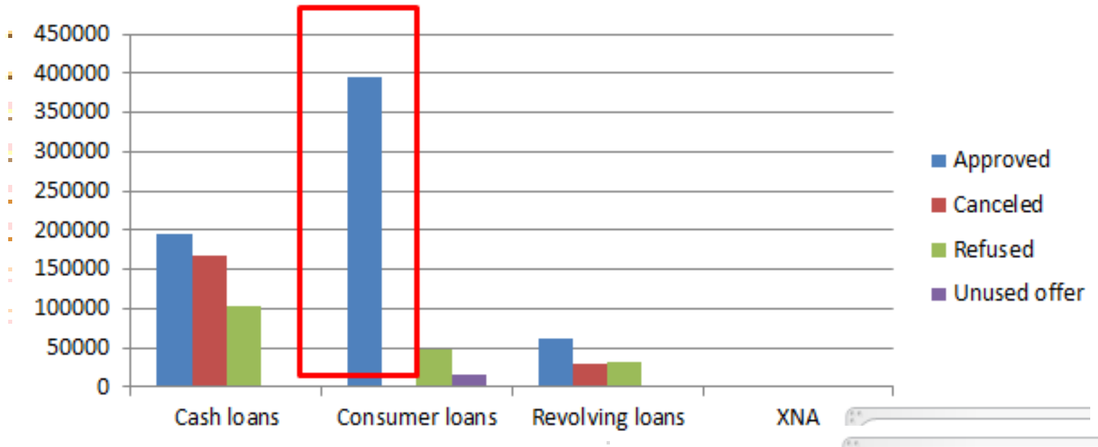


Results conclude that there are close t0 2lakh records of people being married and low percentage of separated or widow



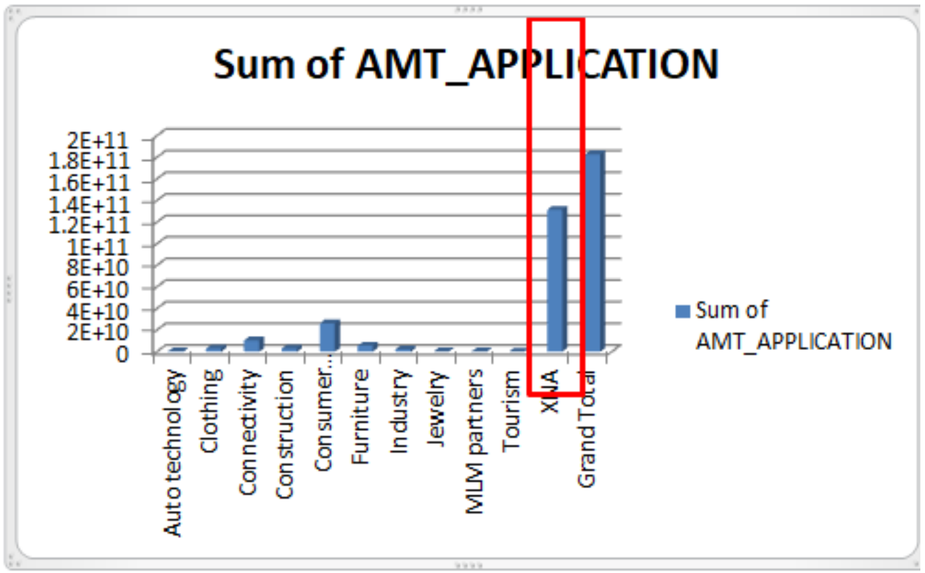
Person with having their own house have a major presence in our dataset

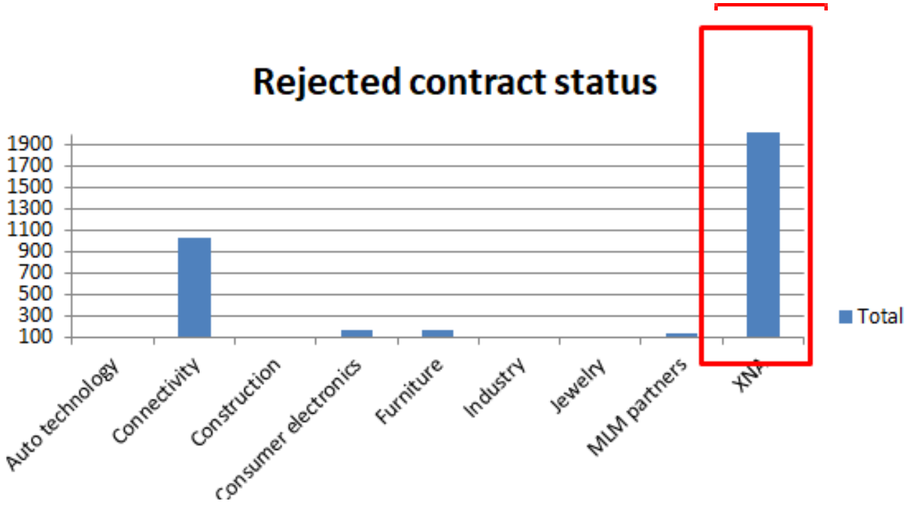
Univariate, segmented univariate, bivariate analysis,



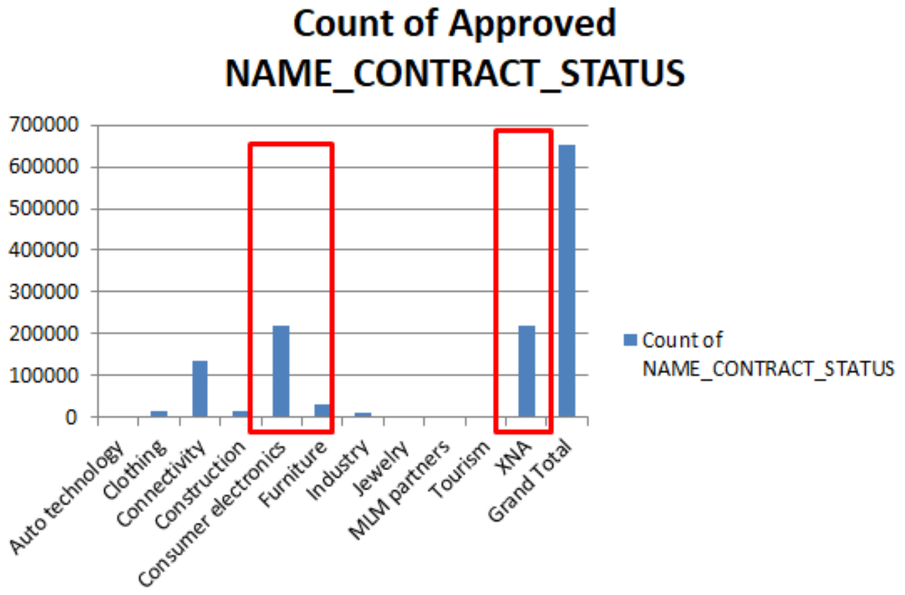
We can infer from the graph that there are high chance of loan being approved if is a consumer loan.

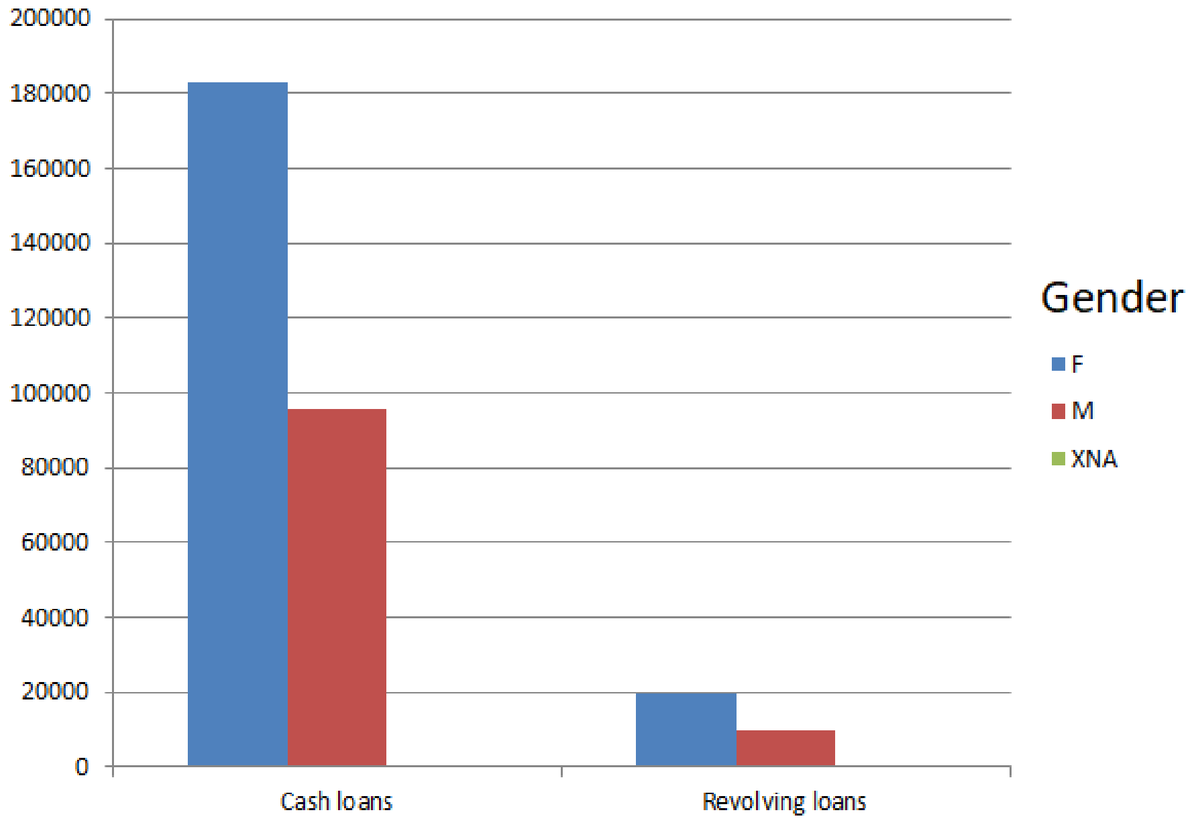
For various previous application higher amount if loans were asked fro XNA and consumer loans





Someone asking loan for XNA and connectivity areas have higher chances of being rejected





We can infer from this graph that female have more number of females have asked for cash or revolving loans compared to men applicants

Top 10 **correlation**

- With payment default

	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_2	OBS_30_CNT_SOCIAL_CIRCLE	60_CNT_SOCIAL_CI	EXT_SOURCE_3
AMT_INCOME_TOTAL	1.00												
AMT_CREDIT	0.04	1.00											
REGION_POPULATION_F	0.01	0.07	1.00										
DAYS_BIRTH	0.00	-0.14	-0.05	1.00									
DAYS_EMPLOYED	-0.01	0.00	0.02	-0.58	1.00								
DAYS_REGISTRATION	0.00	-0.03	-0.06	0.29	-0.19	1.00							
DAYS_ID_PUBLISH	0.00	-0.05	-0.02	0.25	-0.23	0.10	1.00						
AMT_ANNUITY	0.05	0.75	0.07	-0.01	-0.08	0.03	-0.02	1.00					
AMT_GOODS_PRICE	0.04	0.98	0.08	-0.14	0.00	-0.03	-0.06	0.75	1.00				
EXT_SOURCE_2	0.01	0.12	0.17	-0.11	0.00	-0.07	-0.06	0.12	0.13	1.00			
OBS_30_CNT_SOCIAL_CI	0.00	0.02	0.01	-0.01	-0.01	0.01	-0.02	0.00	0.02	0.02	1.00		
OBS_60_CNT_SOCIAL_CI	0.00	0.02	0.01	-0.01	-0.01	0.01	-0.02	0.01	0.02	0.02	1.00	1.00	
EXT_SOURCE_3	-0.02	0.08	-0.01	-0.17	0.09	-0.09	-0.13	0.04	0.08	0.08	0.01	-0.01	1.00

With people having payment difficulty, there is a higher and positive correlation in OBS-60_CNT_SOCIAL and OBS_3-_CNT SOCIAL, followed by Amount of good price to amount annuity etc.

- Without payment default

	AMT_INCOME_TOTA	AMT_CREDIT	REGION_POPULATION_REL	DAYS_BIRTH	DAYS_EMPLO	DAYS_REGISTRATI	DAYS_ID_PL	AMT_ANNUITY	GOODS_	T_SOURCE	OBS_30_CNT_SO	DEF_30_CNT_SOCIA	OBS_60_CNT_SC	NT_SOCIA	T_SOURCE_3
AMT_INCOME_TOTAL	1.00														
AMT_CREDIT	0.34	1.00													
REGION_POPULATION_RELATI	0.17	0.10	1.00												
DAYS_BIRTH	0.06	-0.05	-0.03	1.00											
DAYS_EMPLOYED	-0.14	-0.07	-0.01	-0.62	1.00										
DAYS_REGISTRATION	0.06	0.01	-0.05	0.33	-0.21	1.00									
DAYS_ID_PUBLISH	0.02	0.00	0.00	0.27	-0.27	0.10	1.00								
AMT_ANNUITY	0.42	0.77	0.12	0.01	-0.11	0.04	0.01	1.00							
AMT_GOODS_PRICE	0.35	0.99	0.10	-0.04	-0.07	0.02	0.00	0.78	1.00						
EXT_SOURCE_2	0.14	0.13	0.20	-0.08	-0.03	-0.05	-0.04	0.13	0.14	1.00					
OBS_30_CNT_SOCIAL_CIRCLE	-0.03	0.00	-0.01	0.01	0.01	0.01	-0.01	-0.01	0.00	-0.02	1.00				
DEF_30_CNT_SOCIAL_CIRCLE	-0.03	-0.02	0.01	0.00	0.02	0.00	0.00	-0.02	-0.02	-0.03	0.33	1.00			
OBS_60_CNT_SOCIAL_CIRCLE	-0.03	0.00	-0.01	0.01	0.01	0.01	-0.01	-0.01	0.00	-0.02	1.00	0.33	1.00		
DEF_60_CNT_SOCIAL_CIRCLE	-0.03	-0.02	0.00	0.00	0.02	0.00	0.00	-0.02	-0.02	-0.03	0.25	0.86	0.25	1.00	
EXT_SOURCE_3	-0.07	0.04	-0.01	-0.20	0.11	-0.10	-0.12	0.03	0.04	0.08	0.00	-0.03	0.00	-0.03	1.00

With people not having payment difficulty, there is a higher and positive correlation in OBS-60_CNT_SOCIAL and OBS_3-_CNT SOCIAL, followed by amount of goods price to amount credit etc.

DRIVE LINK FOR PPT & EXCEL SOLVED FILES

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1ZVGKZD6RIDGKUCEJPZCMHJ9-1ODDMO-Z?USP=SHARING

MODULE 7- PROJECT

XYZ Ads Airing Report Analysis

PROJECT DECSRIPTION

The following assignment is based on advertisement data of 6 automobiles companies. The data provides insights on different marketing strategies adopted by companies. How their different products preform based in their EQ units.

TECH STACK USED

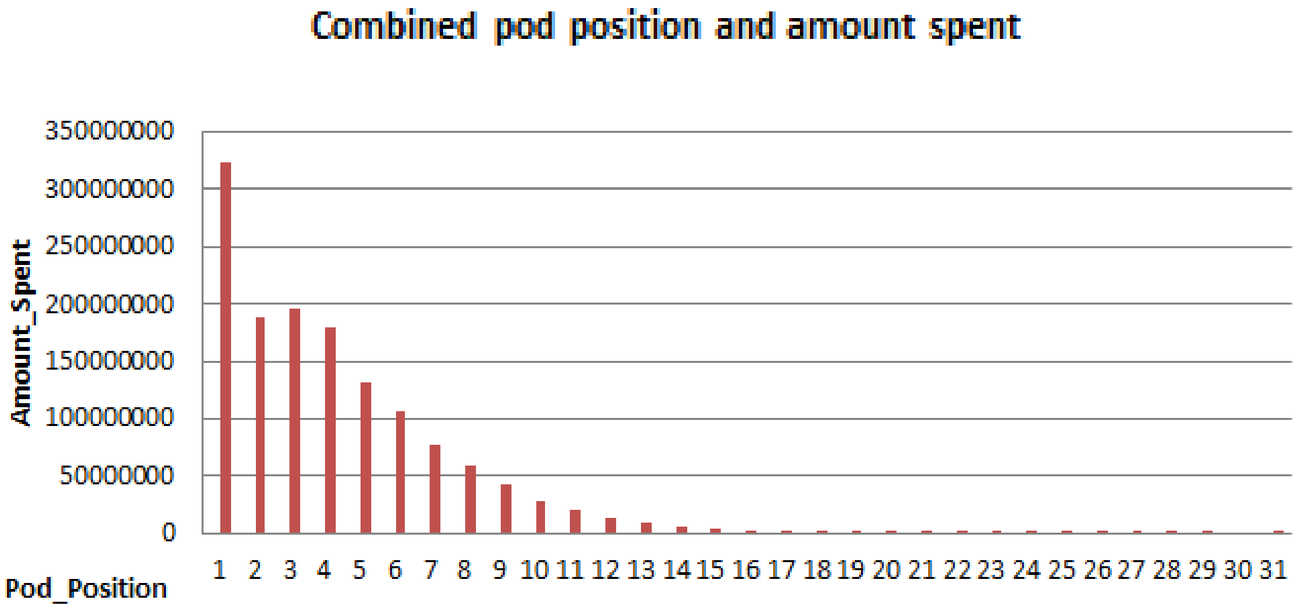
For this project, I have used MS EXCEL for performing all the analysis using functions like pivot tables, averages, sum, various mathematical operators, Count, etc.

Note that, it is not possible to explain every step in the pdf hence I have included the workbooks worked upon which has all the formula and different sheets for different analysis that was performed.

What is Pod Position? Does the Pod position number affect the amount spent on Ads for a specific period of time by a company?

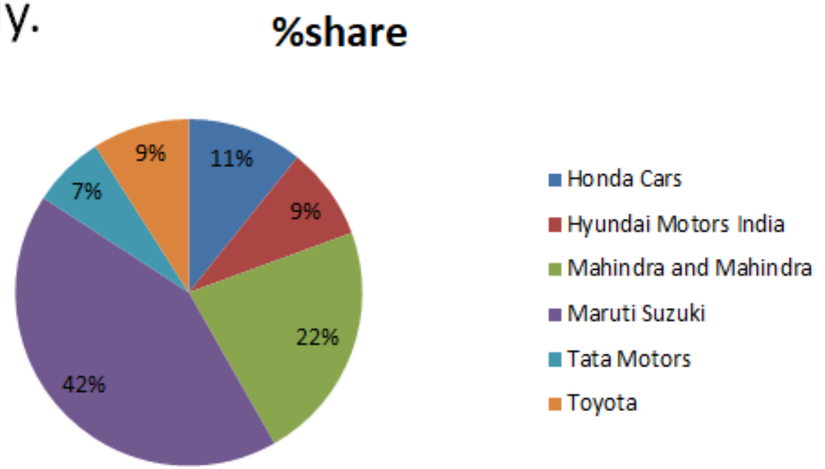
- Pod Position refers to the series in which ads are run, where the idea is the ads which is first in the pod gets the higher attention of the viewers.
- So, in conclusion lower the pod position higher the value since it have the chance of getting high viewership.
- For example : Honda spends the highest total amount of **Rs:7454554** on Pod position 1 out of their total spending, similarly Hyundai spend **Rs:41981693**

- If we look at the combined pod position and the amount spend by all the companies, it is evident that in order to get the lower pod position companies need to spend more amount of money.

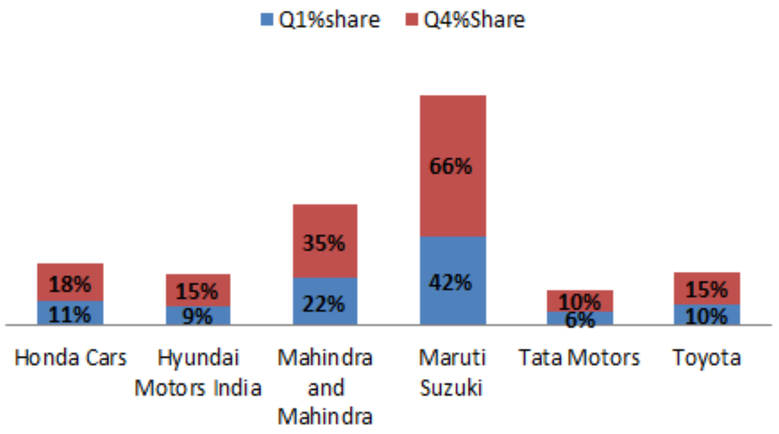


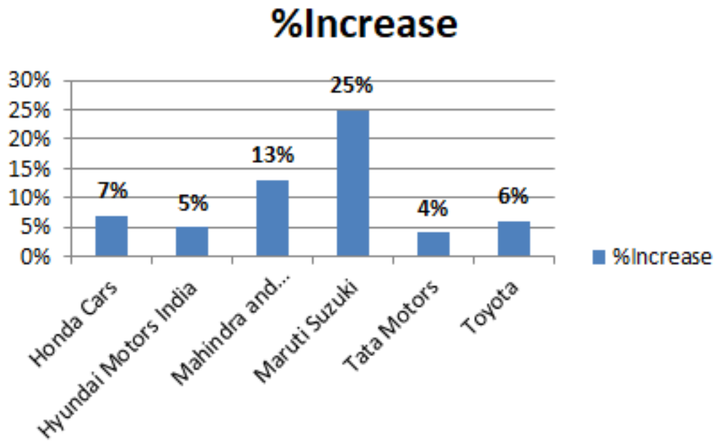
What is the share of various brands in TV airings and how has it changed from Q1 to Q4 in 2021?

- To find the market share of each brand, we have used the criteria that the amount of time an ad is run compared to the total amount of time all the ads are run will give the market share of ads being run for that particular brand
- From the below pie chart it is clear that, Maruti has the highest percentage share in the market, followed by Mahindra, Honda respectively.



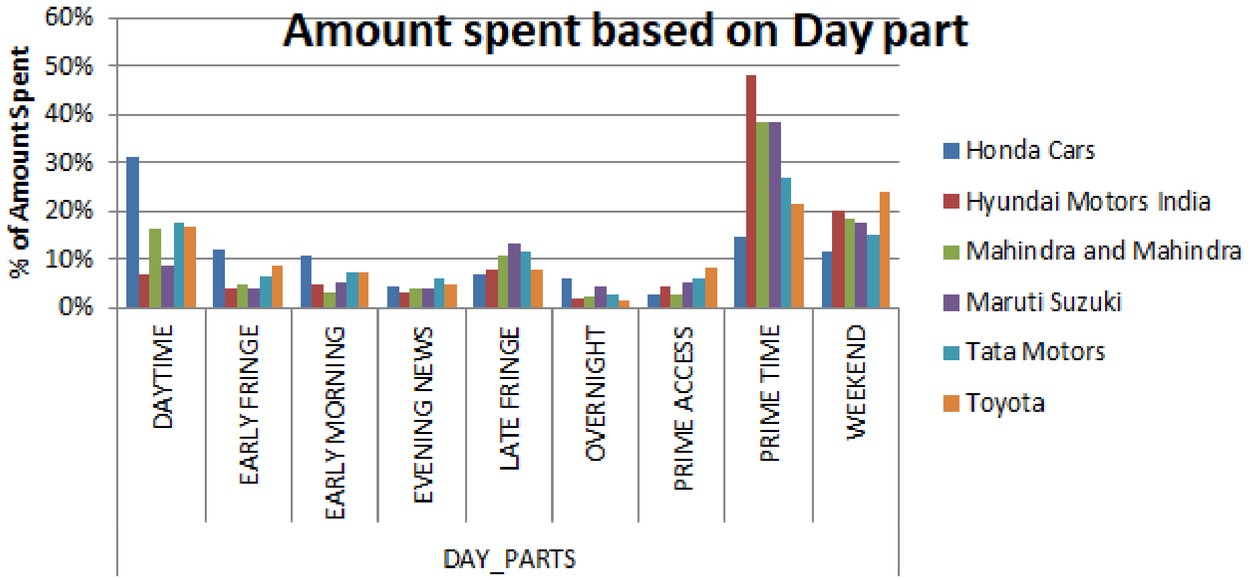
- There has been an increase in percentage share of market for all the 6 automobile brands. The highest jump is from Market Suzuki of 25% growth from 42% to 66%, followed by Mahindra by 13% growth and so on



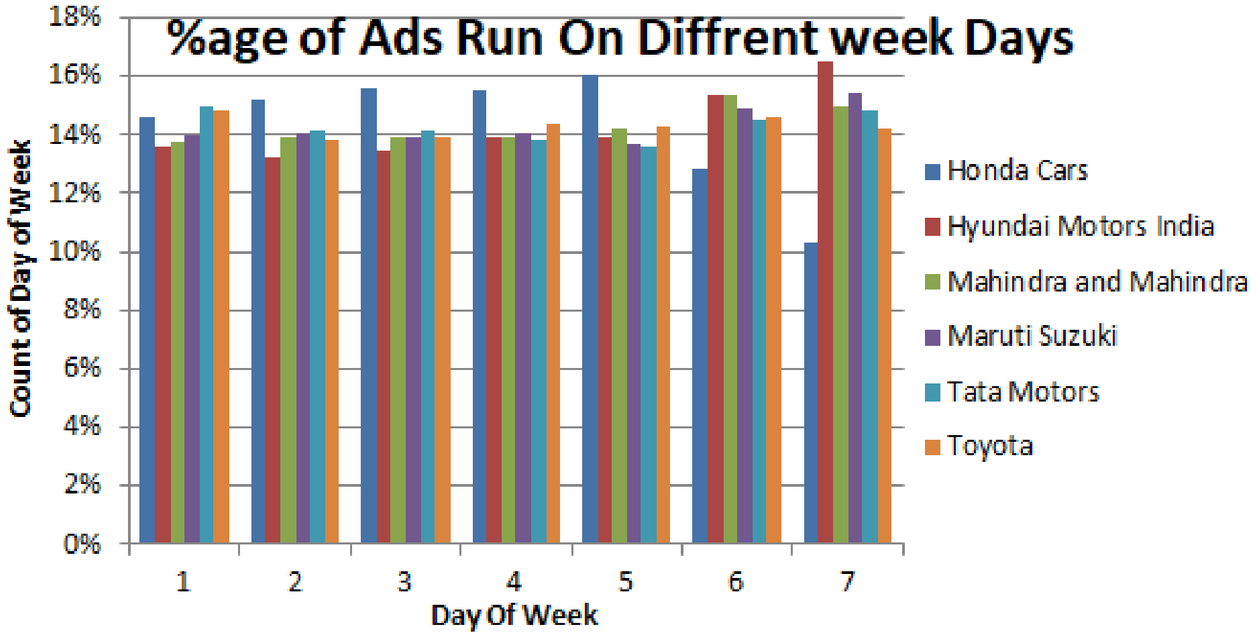


Conduct a competitive analysis for the brands and define advertisement strategy of different brands and how it differs across the brands.

- Looking at the amount being spend on various day parts, it is clear that Honda cars spend most of their money on Daytime ads(**31%**), whereas companies like Hyundai(**48%**), Mahindra (**38%**) and Maruti (**38%**)likes to spend most of Prime Time ads, and Toyota (**24%**) spends highest on weekends.



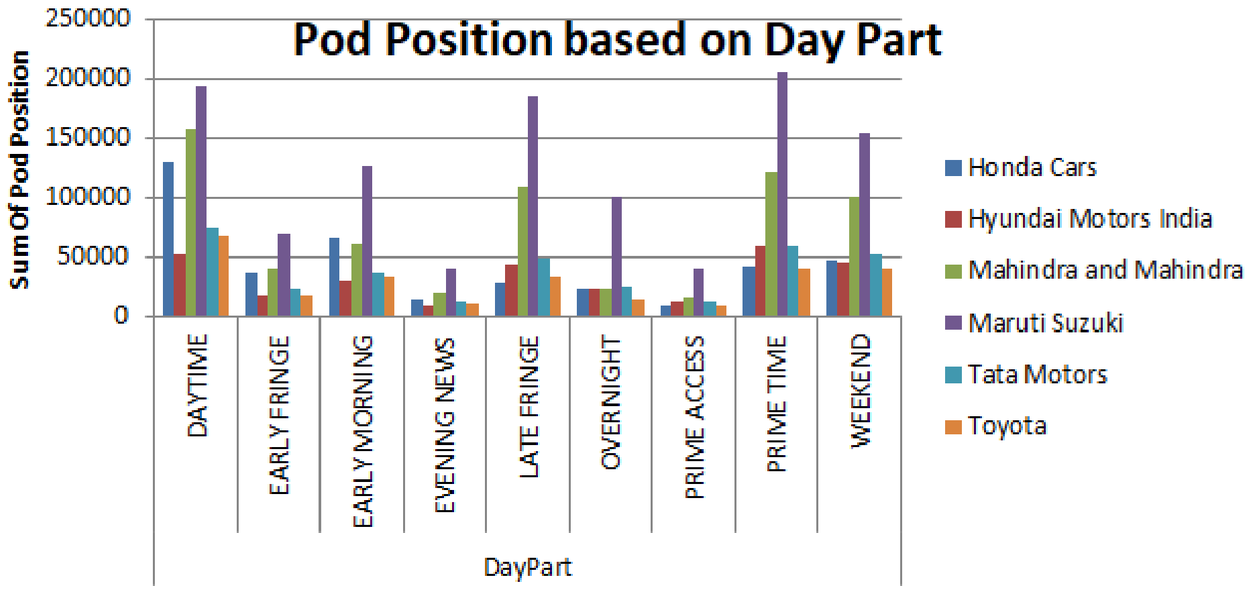
- If we consider the number of days being run on a particular day of week, there is not evident variants among companies. Number of ads are evenly spread across all days by all companies.



- Since we know that, lower the pod position, the more valuable it is. Taking that in consideration, we can see something quite evidently.

1. Hyundai has sum of pod position lower in Evening, prime access and overnight fringe, where they spend lowest amount based on day parts as we have seen in first chart of comparative analysis, which means that if they are not spending more on shows whey they get better attention line prime time, then thy try to get more valuable pod position on those days.

2. Same goes for Maruti as they spend lower amount on Evening Fringe, but the sum of pod position is lower as well on this day which means better pod positions



Mahindra and Mahindra wants to run a digital ad campaign to complement its existing TV ads in Q1 of 2022. Based on the data from 2021, suggest a media plan to the CMO of Mahindra and Mahindra. Which audience should they target?

- To find the target audience, there could be various strategies based on different permanents like conversion rate, Pod position, Geography etc.
- First out of all, if we look at cable ads, and filter the pod position only till 10, there is no EQ units data for Northern India, which means there is a huge untapped markey by all 6 companies and Mahindra should definitely target this.

Pod Position	(Multiple Items)			
EQ Units	(All)			
Average of EQ Units		Network Type		
Brand	TimeZone	broadcast	cable	Grand Total
Honda Cars	Central India	0.79	0.75	0.78
	Northeast India	0.77	0.83	0.82
	Northern India	0.70		0.70
	Southern India	0.86	0.94	0.93
Honda Cars Total		0.77	0.83	0.82
Hyundai Motors India	Central India	0.79	0.95	0.89
	Northeast India	0.83	0.79	0.79
	Northern India	0.81		0.81
	Southern India	0.78	0.80	0.80
Hyundai Motors India Total		0.83	0.79	0.79
Mahindra and Mahindra	Central India	0.99	1.00	1.00
	Northeast India	1.01	0.98	0.99
	Northern India	1.00		1.00
	Southern India	0.99	1.00	1.00
Mahindra and Mahindra Total		1.01	0.98	0.99
Maruti Suzuki	Central India	0.99	0.99	0.99
	Northeast India	0.97	0.99	0.99
	Northern India	1.00		1.00
	Southern India	0.98	0.99	0.99
Maruti Suzuki Total		0.97	0.99	0.99
Tata Motors	Central India	0.72	0.54	0.62
	Northeast India	0.58	0.54	0.55
	Northern India	0.84		0.84
	Southern India	0.50	0.59	0.59
Tata Motors Total		0.59	0.54	0.55
Toyota	Central India	0.87	2.86	2.74
	Northeast India	0.94	0.87	0.88
	Northern India	0.93		0.93
	Southern India	0.83	0.77	0.77
Toyota Total		0.94	0.90	0.90
Grand Total		0.89	0.89	0.89

untapped cable markrt in nothern india

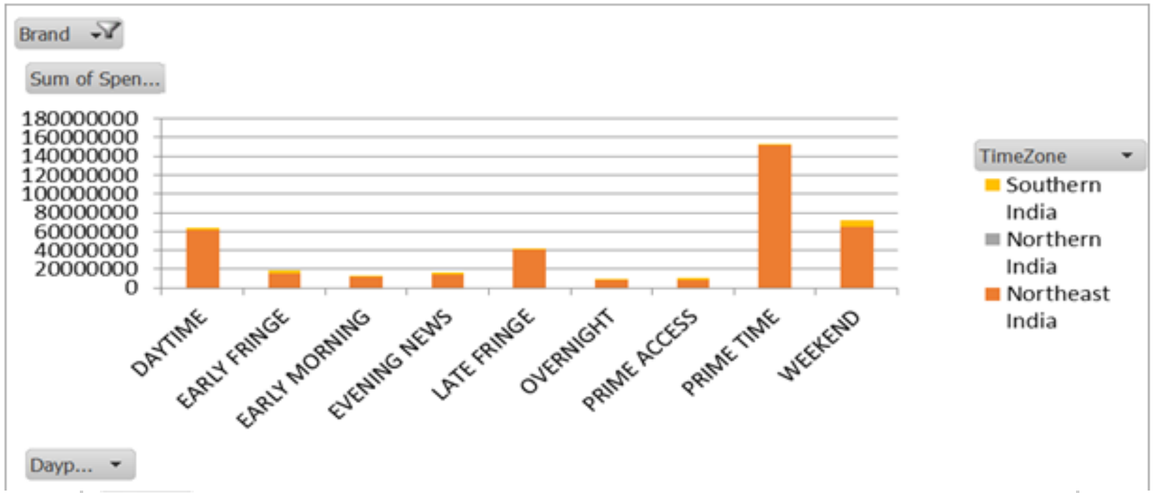
- Mahindra and Mahindra spends low amount of money On Northern India and Central India timezone, however their Average Equivalent Sales unit from central India is close to others.

Brand	Mahindra and Mahindra

	Column Labels				
	Central India	Northeast India	Northern India	Southern India	Grand Total
Sum of Spend ($)	951460	375103729	191725	21058741	397305655

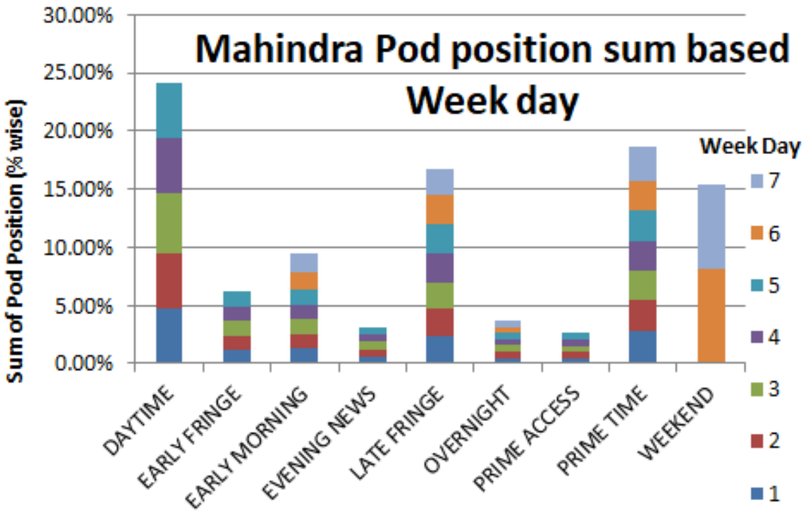
Pod Position	(Multiple Items)			
EQ Units	(All)			
Average of EQ Units		Network Type		
Brand	TimeZone	broadcast	cable	Grand Total
Honda Cars	Central India	0.79	0.75	0.78
	Northeast India	0.77	0.83	0.82
	Northern India	0.70		0.70
	Southern India	0.86	0.94	0.93
Honda Cars Total		0.77	0.83	0.82
Hyundai Motors India	Central India	0.79	0.95	0.89
	Northeast India	0.83	0.79	0.79
	Northern India	0.81		0.81
	Southern India	0.78	0.80	0.80
Hyundai Motors India Total		0.83	0.79	0.79
Mahindra and Mahindra	Central India	0.99	1.00	1.00
	Northeast India	1.01	0.98	0.99
	Northern India	1.00		1.00
	Southern India	0.99	1.00	1.00
Mahindra and Mahindra Total		1.01	0.98	0.99

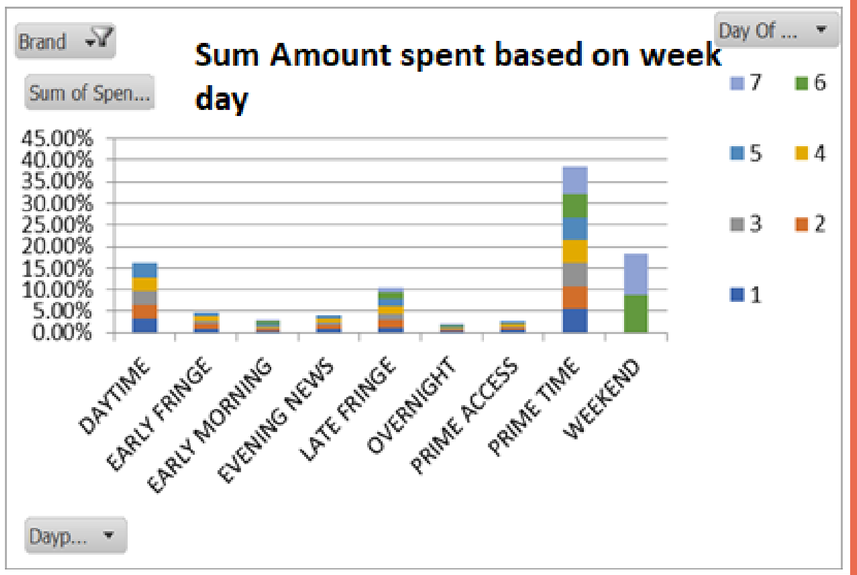
- Mahindra Spends most of their money in Northeast India for all of their products but no ads of Scorpio and XUV in central and Northern India which results in low product visibility



Brand	Mahindra and Mahindra				
Sum of Spend ($)	**Column Labels**				
Row Labels	**Central India**	**Northeast India**	**Northern India**	**Southern India**	**Grand Total**
Mahindra New Thar	951460	371098161	191725	20976563	393217909
Mahindra Scorpio		2268849		9380	2278229
Mahindra XUV 700		1736719		72798	1809517
Grand Total	**951460**	**375103729**	**191725**	**21058741**	**397305655**

- One things noted across many charts is that Mahindra's prime focus has majorly on prime time day as evident in the below chart as well. They invest some of the higher amount in Prime stand but their sum of pod position is also quite high in prime time which means less valuable pod position.





Additional Insights

- There is a stark difference in amount spend by companies son their various products. For example, Honda spends a meager amount on Honda city, Maruti Spends extremely low on Ignis and swift etc., which leads to lower product visibility

Brand	Product	Sum of Spend ($)
Honda Cars	Honda City	59051
	Honda Civiz	34777159
	Honda Jazz	13422130
Honda Cars Total		**48258340**
Hyundai Motors India	Hyundai I20	180808756
Hyundai Motors India Total		**180808756**
Mahindra and Mahindra	Mahindra New Thar	393217909
	Mahindra Scorpio	2278229
	Mahindra XUV 700	1809517
Mahindra and Mahindra Total		**397305655**
Maruti Suzuki	Maruti Suzuki Baleno	332492531
	Maruti Suzuki Celerio	104075
	Maruti Suzuki Ciaz	199405767
	Maruti Suzuki Ertiga	4038393
	Maruti Suzuki Ignis	35283
	Maruti Suzuki Swift	15347
	Maruti Suzuki WagonR	22555076
Maruti Suzuki Total		**558646472**
Tata Motors	Tata Nexon	68498763
	Tata Safari	24022082
	Tata Tiago	2269382
Tata Motors Total		**94790227**
Toyota	Toyota Etios	43551141
	Toyota Fortuner	159554
	Toyota Innova	68942417
Toyota Total		**112653112**
(blank)	(blank)	
(blank) Total		
Grand Total		**1392462562**

DRIVE LINK FOR PPT & EXCEL SOLVED FILES

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1-ZZZY4DWDJ3F-GGASKCIDQQI6GM9NGJZ?USP=SHARING

MODULE 8- PROJECT

ABC Call Volume Trend Analysis

PROJECT DECSRIPTION

The following project is based on call center volume trends. It provides data on the volume of calls received on regular basis, what is the average time spend on answering the calls. It also provides the breakup of the calls received based on abandoned, transferred and answered. These data points would help in determine the actual manpower requirements based on volume trends.

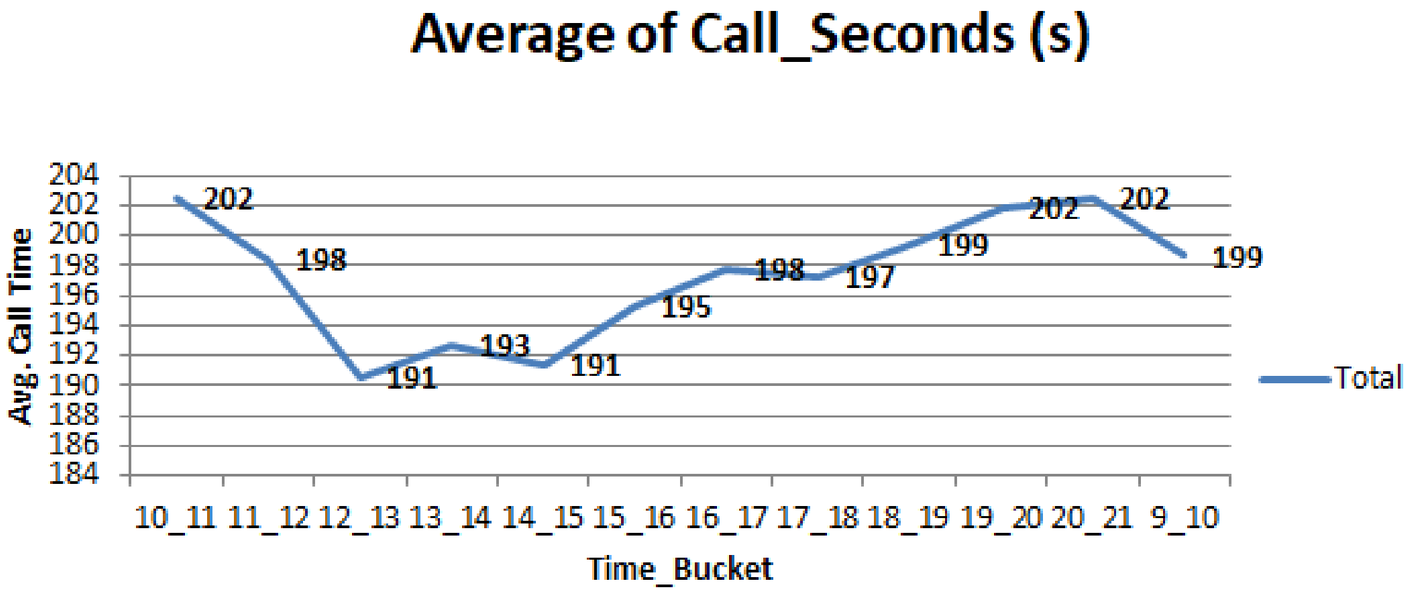
TECH STACK USED

For this project, I have used MS EXCEL for performing all the analysis using functions like pivot tables, averages, sum, various mathematical operators, Count, etc.

Note that, it is not possible to explain every step in the pdf hence I have included the workbooks worked upon which has all the formula and different sheets for different analysis that was performed.

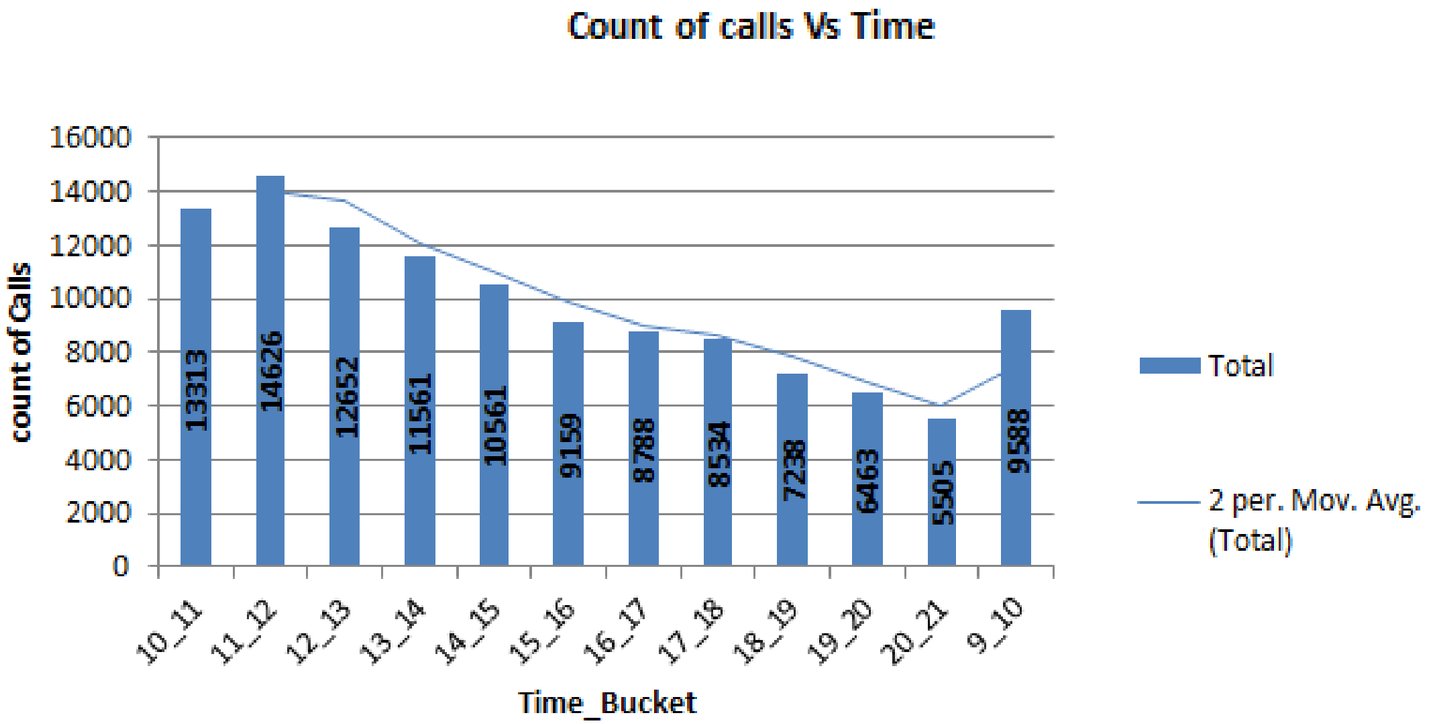
Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

- The aggregate avg. time for all calls duration is 196 seconds.
- The avg. time lowers downs as the day progress but picks up again from around middle of the day and then reduce a bit in the end



Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time].

- The volume of calls coming in decreases as we go into day but it pick a bit in the last bucket of the day



As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. You have to calculate minimum number of agents required in each time bucket so that at least 90 calls should be answered out of 100.)

- Since we know, the tolerance rate is 10%, we first try to ascertain total no of calls received in every bucket and what was the average call duration for that bucket. This will help us to determine the total seconds required to attend per bucket in order to achieve 90% success.

Count of Agent_Name	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
10_11	6911	6368	34	13313
11_12	6028	8560	38	14626
12_13	3073	9432	147	12652
13_14	2617	8829	115	11561
14_15	2475	7974	112	10561
15_16	1214	7760	185	9159
16_17	747	7852	189	8788
17_18	783	7601	150	8534
18_19	933	6200	105	7238
19_20	1848	4578	37	6463
20_21	2625	2870	10	5505
9_10	5149	4428	11	9588
Grand Total	**34403**	**82452**	**1133**	**117988**

Row Labels	Average of Call_Seconds (s)
10_11	202.4989852
11_12	198.4062028
12_13	190.5965442
13_14	192.6287465
14_15	191.4335224
15_16	195.3163048
16_17	197.6802628
17_18	197.2947003
18_19	199.3934893
19_20	201.8658458
20_21	202.447067
9_10	198.6478271
Grand Total	196.4807137

- Calculation for total working seconds
- In this, we try to figure out what would be the total seconds in a month an employee would be answering on calls.
- Since actual working seconds is 16200 in day, we will multiply it with total working day which would come out as 22 after reducing Sundays and leaves.

Total days of data provided	23		
Avg. Month days	**30**		
Total working hours/ day	9		
Actual Working Hours/day	7.5		
Hours spend on call with users/day (60% of 7.5)	**4.5**	**16200**	**Actual callings time in second / per day (Total hours per day multiply by total 3600**
Hours spent in 22 days		**356400**	**Out of 30 days, Assuming month starts with Monday and a typical 4 Sunday month we will minus 4 sundays and 4 unplanned levaes then remains 22 days**

- Finally, we know the total actual working seconds in a month which is 356400 seconds which we need to divide by from total working seconds required to achieve 90% calls for every bucket. And this will provide us the, manpower requirements which would be finally rounded to next digit, which is 59 employees.

Row Labels	Average of Call_Seconds (s)
10_11	202.4989852
11_12	198.4062028
12_13	190.5965442
13_14	192.6287465
14_15	191.4335224
15_16	195.3163048
16_17	197.6802628
17_18	197.2947003
18_19	199.3934893
19_20	201.8658458
20_21	202.447067
9_10	198.6478271
Grand Total	196.4807137

Count of Agent_Name	Column Labels										
Row Labels	abandon	answered	transfer	Grand Total	% Of call abondend	10% of total	rounded (Calls that can be abonded)	Count of answered (To achive 90%)	Total seconds required for 90% calls	Total Manpower Required	
10_11	6911	6368	34	13313	51.91166529	1331.3	1331	11982	2426342.84	6.807920427	
11_12	6028	8560	38	14626	41.21427595	1462.6	1463	13163	2611620.848	7.327780156	
12_13	3073	9432	147	12652	24.28865002	1265.2	1265	11387	2170322.849	6.089570282	
13_14	2617	8829	115	11561	22.63645013	1156.1	1156	10405	2004302.108	5.623743287	
14_15	2475	7974	112	10561	23.43528075	1056.1	1056	9505	1819575.631	5.105431063	
15_16	1214	7760	185	9159	13.25472213	915.9	916	8243	1609992.3	4.517374579	
16_17	747	7852	189	8788	8.500227583	878.8	879	7909	1563453.199	4.386793487	
17_18	783	7601	150	8534	9.175064448	853.4	853	7681	1515420.593	4.252021866	
18_19	933	6200	105	7238	12.89030119	723.8	724	6514	1298849.189	3.644357994	
19_20	1848	4578	37	6463	28.59353242	646.3	646	5817	1174253.625	3.294763257	
20_21	2625	2870	10	5505	47.68392371	550.5	551	4954	1002922.77	2.814036953	
9_10	5149	4428	11	9588	53.70254485	958.8	959	8629	1714132.1	4.809573793	
Grand Total	34403	82452	1133	117988	29.15804997	11798.8	11799	106189	20864090.51	58.54121915	59

Since we know avg. call time per buckt, we will multiply that with total additonal calls that need to be answeered

pend 16200 seconds per day in a j

Let's say customers also call this ABC insurance company in night but didn't get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am] and distribution of those 30 calls are as follows:

Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

- For this we know the basic calculation from the previous question like, actual working seconds in a month, avg. answering seconds for calls etc.
- Since, we will get 30% of the calls that we get in daytime in the night as well. We will calculate the 30% of 117988 (days calls volume) which comes out as count 35396.4.

		Row Labels	Total_calls		
	AM	10_11	13313		
	AM	11_12	14626		
	PM	12_13	12652		
	PM	13_14	11561		
	PM	14_15	10561		
	PM	15_16	9159		
Day Job	PM	16_17	8788		
	PM	17_18	8534		
	PM	18_19	7238		
	PM	19_20	6463		
	PM	20_21	5505		
	AM	9_10	9588		
		Grand Total	**117988**	**30% of 117988**	**35396.4**

- After figuring out the total calls receive during night, we will calculate the calls received per bucket in night as per the breakup given which will be eventually be multiplied by avg. of call duration which we calculated in the last question as well.
- In the end we will divide the total seconds with actual working seconds that we calculated earlier. And finally we will get the additional manpower requirement in the night as 18 employees.

				% of calls per day calls	Total calls during night per bucket	90%calls answered	Avg sec per bucket	Total Secs	Manpower		
Night Job	PM	21_22	3	10%	3539.64	3185.676	198.6478	632827.6	1.775610592	356400	(Since we know
	PM	22_23	3	10%	3539.64	3185.676	202.499	645096.2	1.810034111		manpower sec
	PM	23_24	2	7%	2359.76	2123.784	198.4062	421371.9	1.182300558		worked from last
	PM	24_1	2	7%	2359.76	2123.784	190.5965	404785.9	1.135762882		questiion
	AM	1_2	1	3%	1179.88	1061.892	192.6287	204550.9	0.573936377		
	AM	2_3	1	3%	1179.88	1061.892	191.4335	203281.7	0.570375213		
	AM	3_4	1	3%	1179.88	1061.892	195.3163	207404.8	0.581943944		
	AM	4_5	1	3%	1179.88	1061.892	197.6803	209915.1	0.588987345		
	AM	5_6	3	10%	3539.64	3185.676	197.2947	628517	1.763515689		
	AM	6_7	4	13%	4719.52	4247.568	199.3935	846937.4	2.376367577		
	AM	7_8	4	13%	4719.52	4247.568	201.8658	857438.9	2.405833072		
	AM	8_9	5	17%	5899.4	5309.46	202.4471	1074885	3.015950068		
		Total	30		0	31856.76		0	17.78061743	18	Total Additonal manpower for night

DRIVE LINK FOR PPT & EXCEL SOLVED FILES

HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1M7O_UXQVJYCKXSARKJ1WW0YTZZIHJBFK?USP=SHARING

LEARNINGS

Throughout these projects, I have gained immense Knowlege primarily about the tools and also about different industries as well some of which I would try to encapsulate here.

- From the Data Analytics Process project, I got to know about the analytical process and how to break down any situation or question analytical into various steps and how to act on the.
- Working through the Instagram User Analytics, I got the opportunity to hone my tech skills in SQL by writing basic to average queries. I learned a little bit of extension of basic ststaments like SELECT, GROUPS etc

- From the projects on Operations Analytics, I was able to learn some above average SQL queries like Window function as well as writing the basic sql queries in a different way based on the questions being asked.
- By completing rest of the project, on technical front I basically learned about excel function and how to effective use different case statement.

On the process front, I immensely leaned about the EDA and what things to look for before starting to analyze the data. More specifically from the Bank Loan case study and IMBD data analysis, I got to know how to take care of blank data points, when to use median or mode to fill the blanks and in what cases we should replace the value appropriately and what percentage of blank data we can tolerate based on industry standards.

In addition, I become aware of various business term like Pod positions in marketing, how analytics can help generate revenue via increase in sales as well as how a budget management issue possibly turns out to be an ineffective research issue in the business.