DATA ANALYSIS PORTFOLIO

By Ravi Kumar

ABOUT ME

Hi, I am Ravi Kumar. I have a technical background and hold a bachelor's degree (B. Tech) in Electrical Engineering from KIIT University, Bhubaneswar – Odisha. I have around 4 years of experience working as a QA Analyst in several IT Companies. Having great interest in Data Analytics, I plan to transition my career from a QA Analyst to Data Analyst or Business Analyst roles.

Further, I will list below the data analytics projects I have worked upon and also the technology stack used to perform the required analysis.

TABLE OF CONTENTS

Module -1 Project: Data Analytics Process

This was the very first project we were given. This involved about the real-life applications of data analytics. In our day-to-day life activities, we use data analytics everyday without even realizing it.

Our task was to give examples of real-life scenarios where we use data analytics and also write down the different data analytics processes. Data Analytics process contains several steps i.e., PLAN, PREPARE, PROCESS, ANALYZE, SHARE, ACT. I gave two examples – first was travel planning and second was house renovation or interior designing.

I used MS Office Word to write down the scenarios.

Module -2 Project: Instagram User Analytics

The second project was Instagram user Analytics. This project was based on the queries raised by the product team of Instagram on various issues. They raised various queries and I used different SQL functions to solve and then used MS Office PowerPoint to prepare a detailed report on them.

The following queries were raised by the marketing team: -

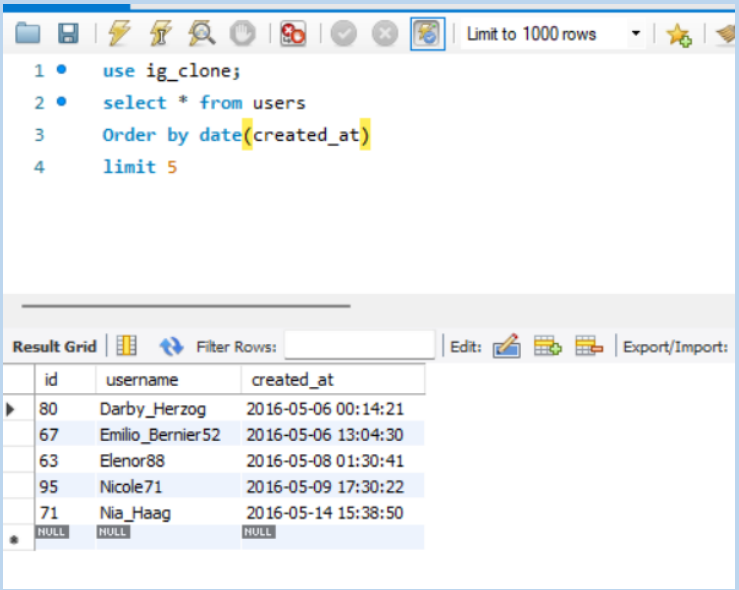
1. People who have been using platform for the longest time.
2. Who are the inactive users.
3. Which user has got the most likes on a photo.
4. Which are the most popular hashtags on Instagram.
5. Which is the best day to launch Ads.

The following queries were raised by the investors: -

1. How active are users on Instagram
2. Is Instagram crowded with fake and dummy accounts.

I first began with importing the data to SQL database. Then closely went through the data set by analyzing each column and type of values they store. After that I checked for any duplicate or null values or if any data cleaning is required. Then I began with the analysis and found out the answers.

1. People who have been using platform for the longest time.



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
71	Nia_Haag	2016-05-14 15:38:50
NULL	NULL	NULL

2. Who are the inactive users.

```
use ig_clone;
select *
from users
left join photos
on users.id = photos.user_id
where user_id is NULL
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	username	created_at	id	image_url	user_id	created_dat
5	Aniya_Hackett	2016-12-07 01:04:39	NULL	NULL	NULL	NULL
7	Kasandra_Homenick	2016-12-12 06:50:08	NULL	NULL	NULL	NULL
14	Jaclyn81	2017-02-06 23:29:16	NULL	NULL	NULL	NULL
21	Rocio33	2017-01-23 11:51:15	NULL	NULL	NULL	NULL
24	Maxwell.Halvorson	2017-04-18 02:32:44	NULL	NULL	NULL	NULL
25	Tierra.Trantow	2016-10-03 12:49:21	NULL	NULL	NULL	NULL
34	Pearl7	2016-07-08 21:42:01	NULL	NULL	NULL	NULL
36	Ollie_Ledner37	2016-08-04 15:42:20	NULL	NULL	NULL	NULL
41	Mckenna17	2016-07-17 17:25:45	NULL	NULL	NULL	NULL
45	David.Osinski47	2017-02-05 21:23:37	NULL	NULL	NULL	NULL
49	Morgan.Kassulke	2016-10-30 12:42:31	NULL	NULL	NULL	NULL
53	Linnea59	2017-02-07 07:49:34	NULL	NULL	NULL	NULL
54	Duane60	2016-12-21 04:43:38	NULL	NULL	NULL	NULL
57	Julien_Schmidt	2017-02-02 23:12:48	NULL	NULL	NULL	NULL

3. Which user has got the most likes on a photo.

```
use ig_clone;
select users.username, photos.id,photos.image_url,count(*) as total_likes
from likes
join photos on photos.id=likes.photo_id
join users on users.id=likes.photo_id
group by photos.id
order by total_likes desc
limit 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

username	id	image_url	total_likes
Kaley9	30	http://kenny.com	41
Jayson65	61	https://dejon.name	41
Zack_Kemmer93	52	https://hershel.com	41
Tomas.Beatty93	97	https://carolanne.com	40
Alexandro35	13	https://fred.com	40
Javonte83	100	https://brook.com	39
Ressie_Stanton46	62	https://rigoberto.net	39
Seth46	44	http://golden.org	39
Mike.Auer39	66	http://lionel.net	39
Harley_Lind18	3	http://vicky.biz	38

Result 19

4. Which is the best day to launch Ads.

```sql
use ig_clone;
SELECT
    DAYNAME(created_at) AS day, COUNT(*) AS total
FROM
    users
GROUP BY day
ORDER BY total DESC
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

day	total
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

The following queries were raised by the investors: -

1. How active are users on Instagram

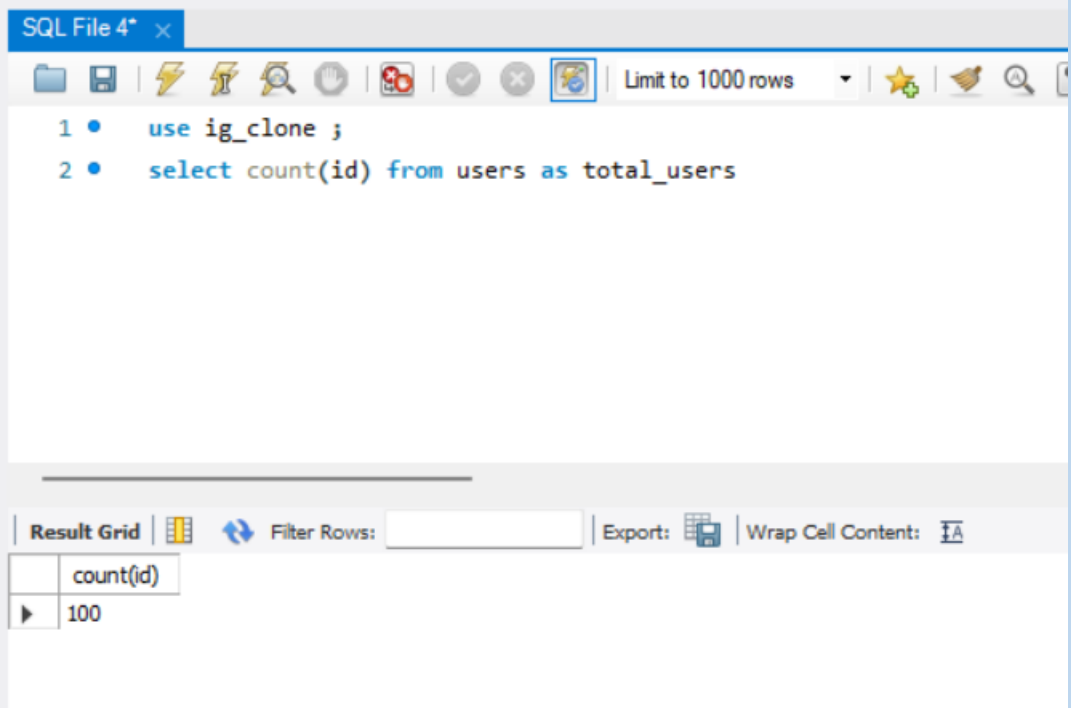
SQL File 4* | Limit to 1000 rows

```sql
use ig_clone ;
SELECT((SELECT COUNT(*)FROM photos)/(SELECT COUNT(*) FROM users)) as avg_post_time;
```

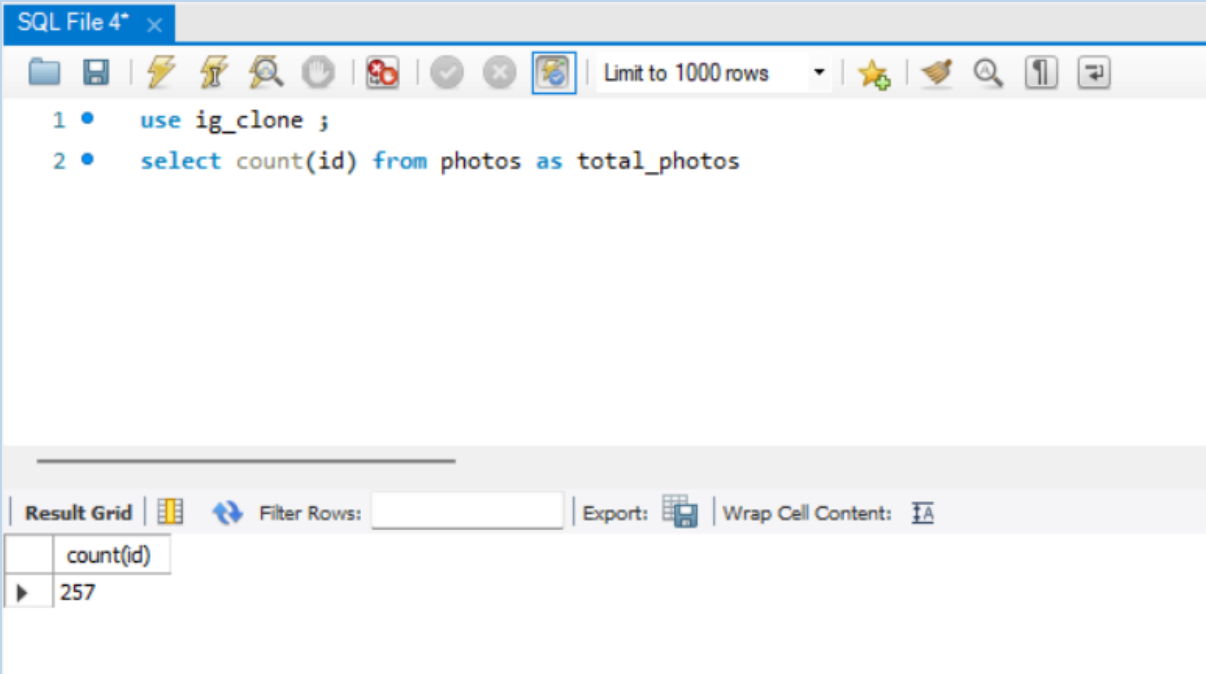
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

avg_post_time
2.5700

Total no of users on Instagram



Total no of photos



Module -3 Project: Operation & Metric Analytics

The third project was Operation Analytics and investigating metric spikes. Operation Analytics is the analysis done for the complete end to end operations of a company. This helps the company to understand in which areas they can improve. On the other hand, metric spike is done to understand certain trends like daily increase or dip in engagement, sales figures etc.

I first imported the data to SQL database and used certain SQL functions to get desired answers and then prepared a presentation in MS Office PowerPoint with relevant explanation.

There were two different case studies with different data sets for each one – Job Data and Investigating metric spike.

A brief description of both case studies and output required form both of them are given below: -

CASE STUDY 01 (JOB DATA)

1. Amount of jobs reviewed over time.
2. No. of events happening per second.
3. Share of each language for different contents.
4. Rows that have same values present in them.

CASE STUDY 02 (INVESTIGATING METRIC SPIKE)

5. To measure the activeness of a user.
6. Amount of users growing over time for a product.
7. Users getting retained weekly after signing up for a product.
8. To measure the activeness of a user weekly.
9. Users engaging with the email service.

I went through the data set closely, analyzed every column and checked for any duplicates or if any data cleaning is required. Then I began with finding solutions for above mentioned queries. Solutions to them are as below –

CASE STUDY 01 (JOB DATA)

1. Amount of jobs reviewed over time.

```
use opr_n_ma ;
SELECT ((COUNT(job_id))/sum(time_spent)/3600) as job_reviewed
 FROM job_data
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_reviewed
0.00000746

2. No. of events happening per second.

```
select ds , job_nums, total_time,
sum(job_nums) over (order by ds rows between 6 preceding and current row)/sum(total_time) as 7day_rolling_average
from
(
select ds ,count(job_id) as job_nums,sum(time_spent) as total_time
from job_data where ds >='2020-11-01'  and ds <='2020-11-30'
group by ds
) a
group by
ds
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ds	job_nums	total_time	7day_rolling_average
2020-11-25	1	45	0.0222
2020-11-26	1	56	0.0357
2020-11-27	1	104	0.0288
2020-11-28	2	33	0.1515
2020-11-29	1	20	0.3000
2020-11-30	2	40	0.2000

3. Share of each language for different contents.

```sql
-- select * from job_data
select language, lang_count,
sum(lang_count)/(sum(lang_count)over(order by language rows between unbounded preceding and unbounded following )) * 100.0 as perc_language
from
(
select language, count(language) as lang_count
from job_data
group by language
) a
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

language	lang_count	perc_language
Arabic	1	12.50000
English	1	12.50000
French	1	12.50000
Hindi	1	12.50000
Italian	1	12.50000
Persian	3	37.50000

4. Rows that have same values present in them.

```sql
select job_id, count(job_id) as duplicate_count
from job_data
group by job_id
having duplicate_count > 1
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_id	duplicate_count
23	3

CASE STUDY 02 (INVESTIGATING METRIC SPIKE)

5. To measure the activeness of a user.

```sql
use opr_n_ma ;

SELECT
    extract(year from occurred_at) as year,
    extract(week from occurred_at) as weeknum,
    count(distinct user_id) as user_enagagement
FROM
    events
GROUP BY
    year, weeknum
order by
    year, weeknum
```

year	weeknum	user_enagagement
2014	17	663
2014	18	1068
2014	19	1113
2014	20	1154
2014	21	1121

6. Amount of users growing over time for a product.

```sql
use opr_n_ma ;

select
    year,
    weeknum,
    new_active_user,
    sum(new_active_user)over(order by year,weeknum rows between unbounded preceding and current row) as cum_active_user
FROM
(
SELECT
    extract(year from activated_at) as year,
    extract(week from activated_at) as weeknum,
    count(distinct user_id) as new_active_user
 FROM
    users
where
    state = 'active'
 group by
    year,
    weeknum
) a
```

year	weeknum	new_active_user	cum_active_user
2013	0	23	23
2013	1	30	53
2013	2	48	101
2013	3	36	137
2013	4	30	167

7. Users getting retained weekly after signing up for a product.

```sql
SELECT user_id, count(user_id),
sum(case when retention_week = 1 then 1 else 0 end) as week_1,
sum(case when retention_week = 2 then 1 else 0 end) as week_2,
sum(case when retention_week = 3 then 1 else 0 end) as week_3,
sum(case when retention_week = 4 then 1 else 0 end) as week_4,
sum(case when retention_week = 5 then 1 else 0 end) as week_5,
sum(case when retention_week = 6 then 1 else 0 end) as week_6,
sum(case when retention_week = 7 then 1 else 0 end) as week_7,
sum(case when retention_week = 8 then 1 else 0 end) as week_8,
sum(case when retention_week = 9 then 1 else 0 end) as week_9
from
(
SELECT
    a.user_id,
    a.signup_week,
    b.engagement_week,
    b.engagement_week - a.signup_week as retention_week
FROM
```

Result Grid

user_id	count(user_id)	week_1	week_2	week_3	week_4	week_5	week_6	week_7	week_8	week_9
11920	1	0	0	0	0	0	0	0	0	0
11924	1	0	0	0	0	0	0	0	0	0
11926	8	1	1	1	1	1	1	1	0	0
11928	8	0	0	0	0	0	0	0	0	1
11929	1	0	0	0	0	0	0	0	0	0
11931	6	1	1	1	1	0	0	0	0	0
11933	6	1	1	1	1	1	0	0	0	0
11936	3	0	0	1	0	0	0	0	0	0

Result 12

8. To measure the activeness of a user weekly.

SQL File 4* events SQL File 6*

```sql
select
    extract(year from occurred_at) as year,
    extract(week from occurred_at) as week,
    device,
    count(distinct user_id ) as user_count
FROM
    events
where
    event_type = 'engagement'
GROUP BY
    year,week,device
order by
    year,week,device
```

Result Grid

year	week	device	user_count
2014	17	acer aspire desktop	9
2014	17	acer aspire notebook	20
2014	17	amazon fire phone	4
2014	17	asus chromebook	21
2014	17	dell inspiron desktop	18

Result 2

9. Users engaging with the email service.

```sql
SELECT
100.0*SUM(case when email_cat = 'email_open' then 1 else 0 END)/SUM(case when email_cat = 'email_sent' then 1 else 0 END) as email_open_rate,
100.0*SUM(case when email_cat = 'email_clicked' then 1 else 0 END)/SUM(case when email_cat = 'email_sent' then 1 else 0 END) as email_clicked_rate
FROM
(
select *,
case
when action IN ('sent_weekly_digest','sent_reengagement_email')
    THEN 'email_sent'
when action IN ('email_open')
    THEN 'email_open'
when action IN ('email_clickthrough')
    THEN 'email_clicked'
END as email_cat
FROM email_events
) a
```

email_open_rate	email_clicked_rate
33.58339	14.78989

After thorough analysis, we found out the answers for every question asked.

Module -4 Project: Hiring Process Analytics

The fourth project I did was hiring process analytics. It is the most important function of a company. Companies get to know about number of hirings, number of resignations, number of rejections, interviews, types of jobs, vacancies etc.

A brief description of the case study and output required is given below: -

1. How many males and females are Hired?
2. What is the average salary offered in this company?
3. Draw the class intervals for salary in the company?
4. Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department?
5. Represent different post tiers using chart/graph?

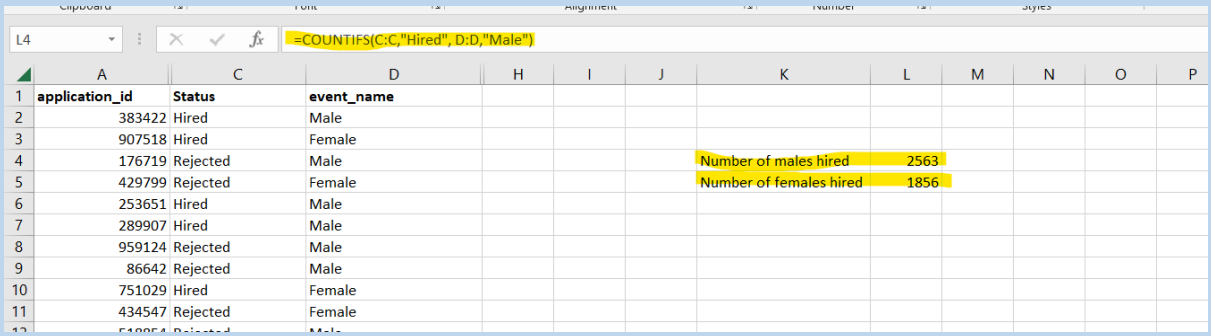
Before beginning with the analysis, I first went through the data, tried to understand every column, then checked for any duplicate data or if any data cleaning is required. After verifying all above mentioned processes I began with getting solutions for above mentioned problem statements.

The data was present in excel so performed the analysis in MS Excel and prepared a detailed report in MS PowerPoint. In order to make the report more presentable, used Pivot Tables to draw graphs. The answers are as follows -

1. How many males and females are Hired?

   The formula used in excel was as follows: -

   - COUNTIFS (C:C,"Hired", D:D,"Male")
   - COUNTIFS (C:C,"Hired", D:D, “Female")



L4 =COUNTIFS(C:C,"Hired", D:D,"Male")

	A	C	D	K	L
1	application_id	Status	event_name		
2	383422	Hired	Male		
3	907518	Hired	Female		
4	176719	Rejected	Male	Number of males hired	2563
5	429799	Rejected	Female	Number of females hired	1856
6	253651	Hired	Male		
7	289907	Hired	Male		
8	959124	Rejected	Male		
9	86642	Rejected	Male		
10	751029	Hired	Female		
11	434547	Rejected	Female		

2. What is the average salary offered in this company?

L4 =AVERAGE(G:G)

	A	C	G	K	L
1	application_id	Status	Offered Salary		
2	383422	Hired	56553		
3	907518	Hired	22075		
4	176719	Rejected	70069	Average Salary offered in this company	49983.03
5	429799	Rejected	3207		
6	253651	Hired	29668		
7	289907	Hired	85914		
8	959124	Rejected	69904		
9	86642	Rejected	11758		
10	751029	Hired	15156		

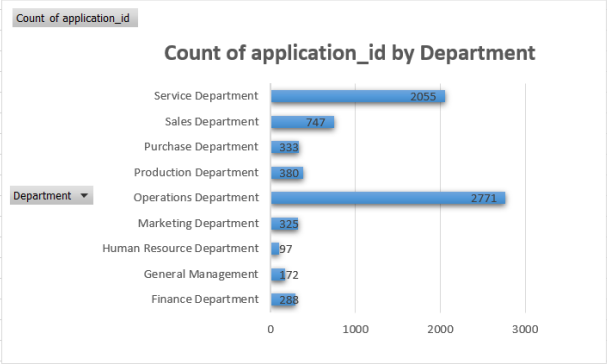
3. Draw the class intervals for salary in the company?

L5 =MAX(G:G)-MIN(G:G)

	A	G	K	L
1	application_id	Offered Salary		
2	383422	56553		
3	907518	22075	Maximum Salary	400000
4	176719	70069	Minimum Salary	100
5	429799	3207	Class Interval	399900
6	253651	29668		
7	289907	85914		
8	959124	69904		
9	86642	11758		
10	751029	15156		
11	434547	49515		
12	518854	26990		

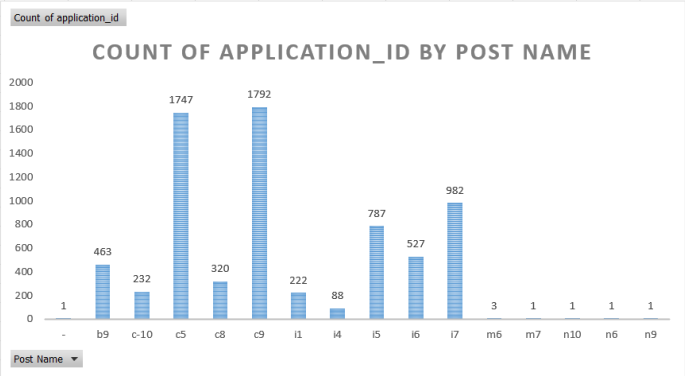
4. Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department?

Department	Count of application_id
Finance Department	288
General Management	172
Human Resource Department	97
Marketing Department	325
Operations Department	2771
Production Department	380
Purchase Department	333
Sales Department	747
Service Department	2055



5. Represent different post tiers using chart/graph?

Post Name	Count of application_id
-	1
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1



Module -5 Project: IMDB Movie Analysis

This is the fifth project I did. This was based on the movie rating platform IMDB. IMDB collects large amount of data of each movie like actors, directors, producers' names, public rating, budget, earnings, stores them in their databases and then analyses them based on their requirements. This project gives a brief idea about the movie analysis done by IMDB to give ratings based on different criteria.

Output required is given below: -

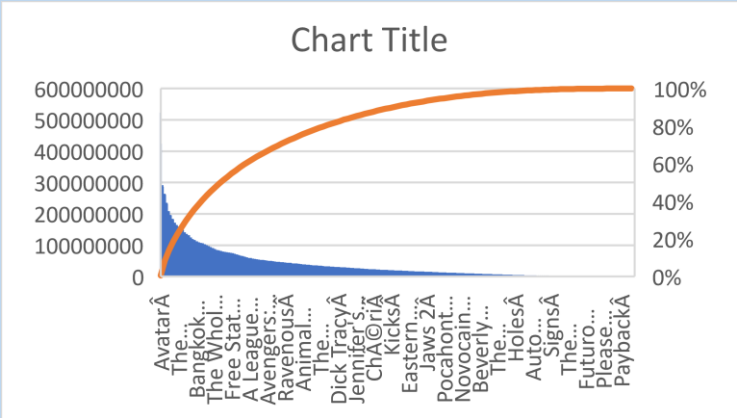
1. Clean the data
2. Find the movies with the highest profit?
3. Find IMDB Top 250
4. Find the best directors
5. Find popular genres
6. Find the critic-favorite and audience-favorite actors

The data was given in excel format, so I first tried to understand the data and variables stored in each column, then went on with data cleaning and finally beginning with the analysis. I used MS Excel to study and analyze data and then used MS PowerPoint to prepare a detailed report for the same.

The insights drawn from the data sets are as follows –

1. Clean the data
2. Find the movies with the highest profit?

   Ans – Movie with the highest profit is Avatar.



3. Find IMDB Top 250

D2 | The Shawshank RedemptionÂ

	A	B	C	D	E	F	G	H	I	J
1	**movie_title**	**imdb_score**	**num_voted_users (applied filter > 25000)**	**IMDB_top_250 (num_voted_users>25000)**						
2	The Shawshank Redem	9.3	1689764	The Shawshank RedemptionÂ						
3	The GodfatherÂ	9.2	1155770	The GodfatherÂ						
4	The Dark KnightÂ	9	1676169	The Dark KnightÂ						
5	The Godfather: Part IIÁ	9	790926	The Godfather: Part IIÂ						
6	FargoÂ	9	170055	FargoÂ						
7	The Lord of the Rings:	8.9	1215718	The Lord of the Rings: The Return of the KingÂ						
8	Schindler's ListÂ	8.9	865020	Schindler's ListÂ						
9	Pulp FictionÂ	8.9	1324680	Pulp FictionÂ						
10	The Good, the Bad and	8.9	503509	The Good, the Bad and the UglyÂ						
11	12 Angry MenÂ	8.9	447785	12 Angry MenÂ						
12	InceptionÂ	8.8	1468200	InceptionÂ						
13	The Lord of the Rings:	8.8	1238746	The Lord of the Rings: The Fellowship of the RingÂ						
14	DaredevilÂ	8.8	213483	DaredevilÂ						

4. Find the best directors

The top 10 directors are as below: -

director_name	**imdb_score**		**top 10 directors**
John Blanchard Average	9.5		John Blanchard Average
John Blanchard	9.5		Frank Darabont Average
Frank Darabont Average	9.3		Francis Ford Coppola Average
Frank Darabont	9.3		John Stockwell Average
Francis Ford Coppola Average	9.2		Christopher Nolan Average
Francis Ford Coppola	9.2		Francis Ford Coppola Average
John Stockwell Average	9.1		Peter Jackson Average
John Stockwell	9.1		Steven Spielberg Average
Christopher Nolan Average	9		Quentin Tarantino Average
Christopher Nolan	9		Sergio Leone Average
Francis Ford Coppola Average	9		
Francis Ford Coppola	9		
Peter Jackson Average	8.9		

5. Find popular genres

genres	**imdb_score**		**top 10 genres**
Comedy	9.5		Comedy Average
Comedy Average	9.5		Crime|Drama Average
Crime|Drama	9.3		Drama Average
Crime|Drama Average	9.3		Action Average
Crime|Drama	9.2		Action|Crime|Drama|Thriller Average
Crime|Drama Average	9.2		Crime|Drama|Thriller Average
Drama	9.1		Biography|Drama|History Average
Drama Average	9.1		Western Average
Drama	9.1		Action|Adventure|Sci-Fi|Thriller Average
Drama Average	9.1		Action|Adventure|Drama|Fantasy Average
Action	9.1		
Action Average	9.1		
Action|Crime|Drama|Thriller	9		
Action|Crime|Drama|Thriller Average	9		

6. Find the critic-favorite and audience-favorite actors

	B	C	D	E	F	G	H	I	J
1	**Meryl Streep**	**Leonardo DiCaprio**	**Brad Pitt**	**Combined column (appended)**				**help column**	**actor_1_name**
2	It's ComplicatedÂ	TitanicÂ	The Curious Case of Benjamin ButtonÂ	It's ComplicatedÂ				CCH Pounder:1	CCH Pounder
3	The River WildÂ	The Great GatsbyÂ	TroyÂ	The River WildÂ				Johnny Depp:1	Johnny Depp
4	Julie & JuliaÂ	InceptionÂ	Ocean's TwelveÂ	Julie & JuliaÂ				Christoph Waltz:1	Christoph Waltz
5	The Devil Wears PradaÂ	The RevenantÂ	Mr. & Mrs. SmithÂ	The Devil Wears PradaÂ				Tom Hardy:1	Tom Hardy
6	Lions for LambsÂ	The AviatorÂ	Spy GameÂ	Lions for LambsÂ				Doug Walker:1	Doug Walker
7	Out of AfricaÂ	Django UnchainedÂ	Ocean's ElevenÂ	Out of AfricaÂ				Daryl Sabara:1	Daryl Sabara
8	Hope SpringsÂ	Blood DiamondÂ	FuryÂ	Hope SpringsÂ				J.K. Simmons:1	J.K. Simmons
9	One True ThingÂ	The Wolf of Wall StreetÂ	Seven Years in TibetÂ	One True ThingÂ				Brad Garrett:1	Brad Garrett
10	Florence Foster JenkinsÂ	Gangs of New YorkÂ	Fight ClubÂ	Florence Foster JenkinsÂ				Chris Hemsworth:1	Chris Hemsworth
11	The HoursÂ	The DepartedÂ	Sinbad: Legend of the Seven SeasÂ	The HoursÂ				Alan Rickman:1	Alan Rickman
12	The Iron LadyÂ	Shutter IslandÂ	Interview with the Vampire: The Vampire ChroniclesÂ	The Iron LadyÂ				Henry Cavill:1	Henry Cavill
13	A Prairie Home CompanionÂ	Body of LiesÂ	The Tree of LifeÂ	A Prairie Home CompanionÂ				Kevin Spacey:1	Kevin Spacey
14	JuliaÂ	Catch Me If You CanÂ	The Assassination of Jesse James by the Coward Robert FordÂ	JuliaÂ				Giancarlo Giannini:1	Giancarlo Giannini
15		The BeachÂ	BabelÂ	TitanicÂ				Johnny Depp:2	Johnny Depp
16		Revolutionary RoadÂ	By the SeaÂ	The Great GatsbyÂ				Johnny Depp:3	Johnny Depp
17		The Man in the Iron MaskÂ	Killing Them SoftlyÂ	InceptionÂ				Henry Cavill:2	Henry Cavill
18		J. EdgarÂ	True RomanceÂ	The RevenantÂ				Peter Dinklage:1	Peter Dinklage
19		The Quick and the DeadÂ	Johnny SuedeÂ	The AviatorÂ				Chris Hemsworth:2	Chris Hemsworth
20		Marvin's RoomÂ		Django UnchainedÂ				Johnny Depp:4	Johnny Depp
21		Romeo + JulietÂ		Blood DiamondÂ				Will Smith:1	Will Smith
22		The Great GatsbyÂ		The Wolf of Wall StreetÂ				Aidan Turner:1	Aidan Turner

This completes our analysis as all questions have now been answered.

Module -6 Project: Bank Loan Case Study

This project is about a case study relayed to a Bank Loan. We have to carry out an EDA (Exploratory Data Analysis). Based on our analysis, we will get the solution for required questions.

I first analyzed the data. While analyzing, I found out that data had a lot of missing values. So, my first task was to get the missing values by performing mean, median and mode functions as required. So, I began by cleaning the data and then finding the outliers so as to make the data standardized. To perform the analysis, I used MS Excel 2019 for analysis and used MS Word to prepare the report.

So, let's begin with analysis……

1. **Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly**

First, we imported the data to excel.

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family

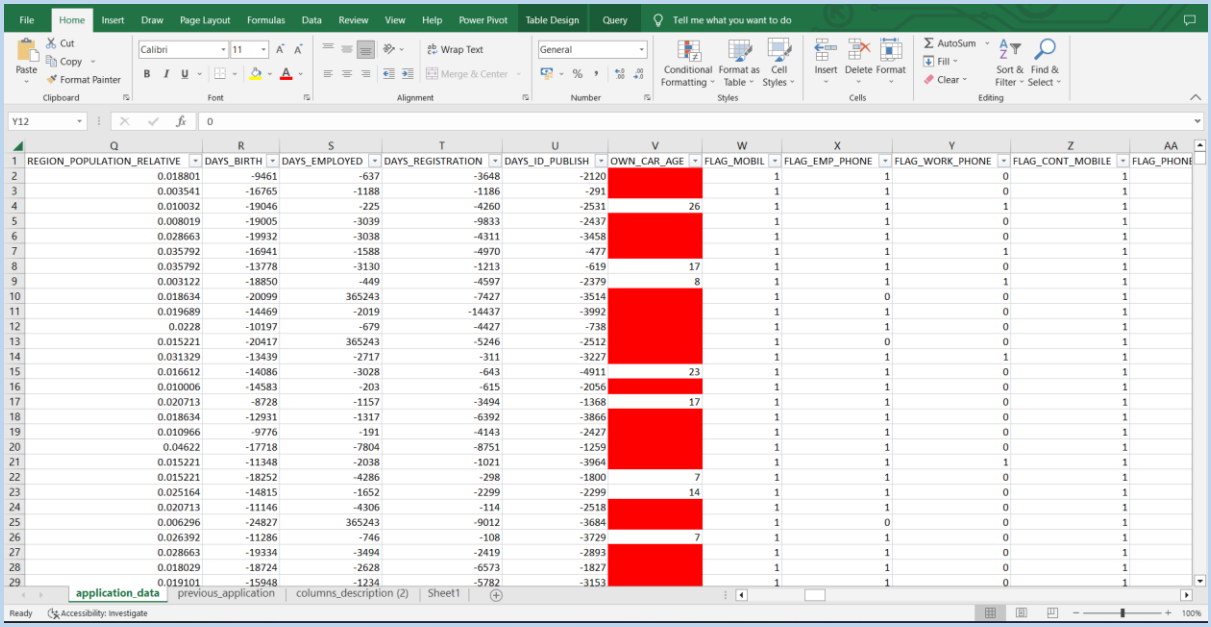
application_data | columns_description | previous_application | Sheet1

Column1	Table	Row	Description	Special
1	application_data	SK_ID_CURR	ID of loan in our sample	
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at	
5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	
6	application_data	CODE_GENDER	Gender of the client	
7	application_data	FLAG_OWN_CAR	Flag if the client owns a car	
8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat	
9	application_data	CNT_CHILDREN	Number of children the client has	
10	application_data	AMT_INCOME_TOTAL	Income of the client	
11	application_data	AMT_CREDIT	Credit amount of the loan	
12	application_data	AMT_ANNUITY	Loan annuity	
13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given	
14	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan	
15	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)	
16	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved	
17	application_data	NAME_FAMILY_STATUS	Family status of the client	
18	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)	
19	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more po	normalized
20	application_data	DAYS_BIRTH	Client's age in days at the time of application	time only relative to the application
21	application_data	DAYS_EMPLOYED	How many days before the application the person started current employment	time only relative to the application
22	application_data	DAYS_REGISTRATION	How many days before the application did client change his registration	time only relative to the application
23	application_data	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applie	time only relative to the application
24	application_data	OWN_CAR_AGE	Age of client's car	
25	application_data	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)	
26	application_data	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)	
27	application_data	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)	
28	application_data	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)	
29	application_data	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)	
30	application_data	FLAG_EMAIL	Did client provide email (1=YES, 0=NO)	

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	11
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	11
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	7
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	9
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8
2315218	175704	Cash loans		0	0			TUESDAY	11
1656711	296299	Cash loans		0	0			MONDAY	7
2367563	342292	Cash loans		0	0			MONDAY	15
2579447	334349	Cash loans		0	0			SATURDAY	15
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	7
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	16
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	15
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	16
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	7
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	11
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	6
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4
1172842	302212	Cash loans		0	0			TUESDAY	9
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	5
1555330	199353	Cash loans		0	0			SATURDAY	6
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8
2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESDAY	11
1676258	433469	Cash loans	22242.825	247500	268083	0	247500	THURSDAY	14

Then, in order to clean data I, highlighted the blank cells first .

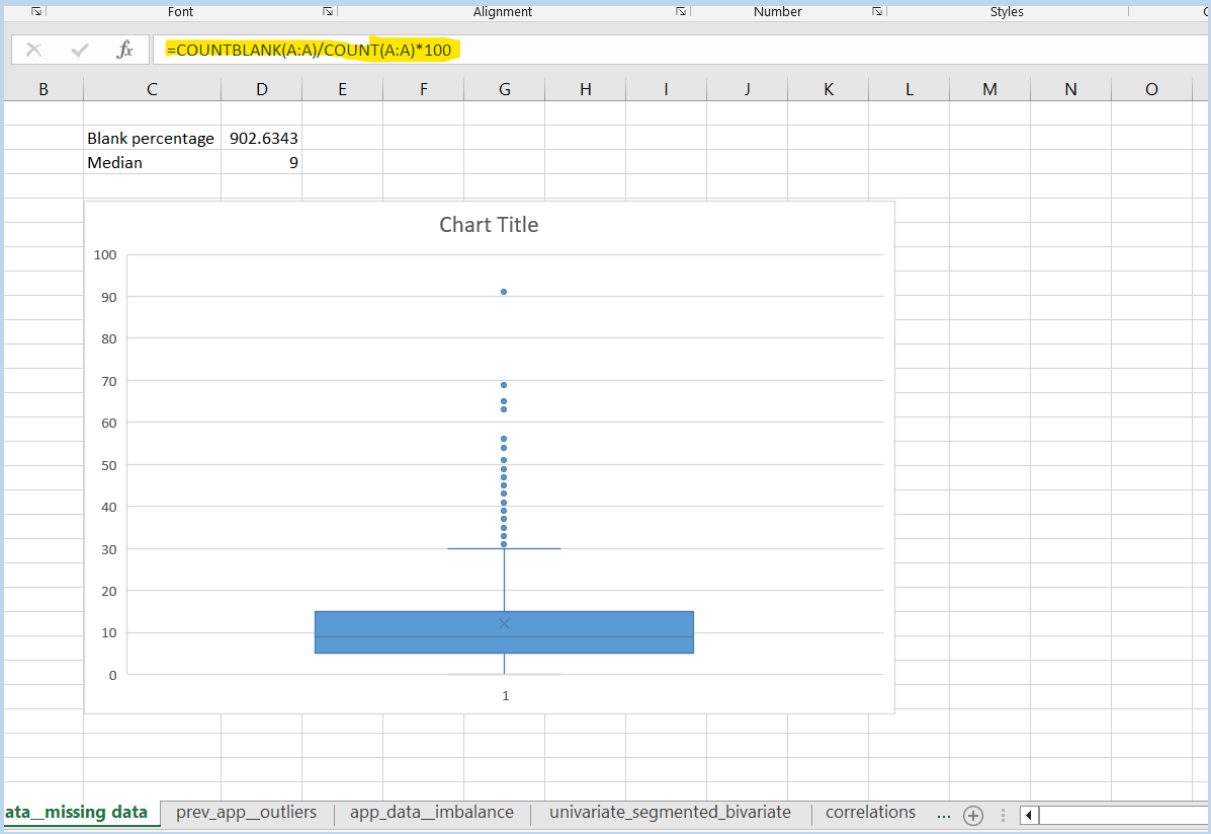
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	11
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	11
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	7
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	9
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8
2315218	175704	Cash loans		0	0			TUESDAY	11
1656711	296299	Cash loans		0	0			MONDAY	7
2367563	342292	Cash loans		0	0			MONDAY	15
2579447	334349	Cash loans		0	0			SATURDAY	15
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	7
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	16
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	15
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	16
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	7
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	11
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	6
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4
1172842	302212	Cash loans		0	0			TUESDAY	9
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	5
1555330	199353	Cash loans		0	0			SATURDAY	6
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8
2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESDAY	11
1676258	433469	Cash loans	22242.825	247500	268083	0	247500	THURSDAY	14



REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	OWN_CAR_AGE	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE
0.018801	-9461	-637	-3648	-2120		1	1	0	1	
0.003541	-16765	-1188	-1186	-291		1	1	0	1	
0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	
0.008019	-19005	-3039	-9833	-2437		1	1	0	1	
0.028663	-19932	-3038	-4311	-3458		1	1	0	1	
0.035792	-16941	-1588	-4970	-477		1	1	1	1	
0.035792	-13778	-3130	-1213	-619	17	1	1	0	1	
0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	
0.018634	-20099	365243	-7427	-3514		1	0	0	1	
0.019689	-14469	-2019	-14437	-3992		1	1	0	1	
0.0228	-10197	-679	-4427	-738		1	1	0	1	
0.015221	-20417	365243	-5246	-2512		1	0	0	1	
0.031329	-13439	-2717	-311	-3227		1	1	1	1	
0.016612	-14086	-3028	-643	-4911	23	1	1	0	1	
0.010006	-14583	-203	-615	-2056		1	1	0	1	
0.020713	-8728	-1157	-3494	-1368	17	1	1	0	1	
0.018634	-12931	-1317	-6392	-3866		1	1	0	1	
0.010966	-9776	-191	-4143	-2427		1	1	0	1	
0.04622	-17718	-7804	-8751	-1259		1	1	0	1	
0.015221	-11348	-2038	-1021	-3964		1	1	1	1	
0.015221	-18252	-4286	-298	-1800	7	1	1	0	1	
0.025164	-14815	-1652	-2299	-2299	14	1	1	0	1	
0.020713	-11146	-4306	-114	-2518		1	1	0	1	
0.006296	-24827	365243	-9012	-3684		1	0	0	1	
0.026392	-11286	-746	-108	-3729	7	1	1	0	1	
0.028663	-19334	-3494	-2419	-2893		1	1	0	1	
0.018029	-18724	-2628	-6573	-1827		1	1	0	1	
0.019101	-15948	-1234	-5782	-3153		1	1	0	1	

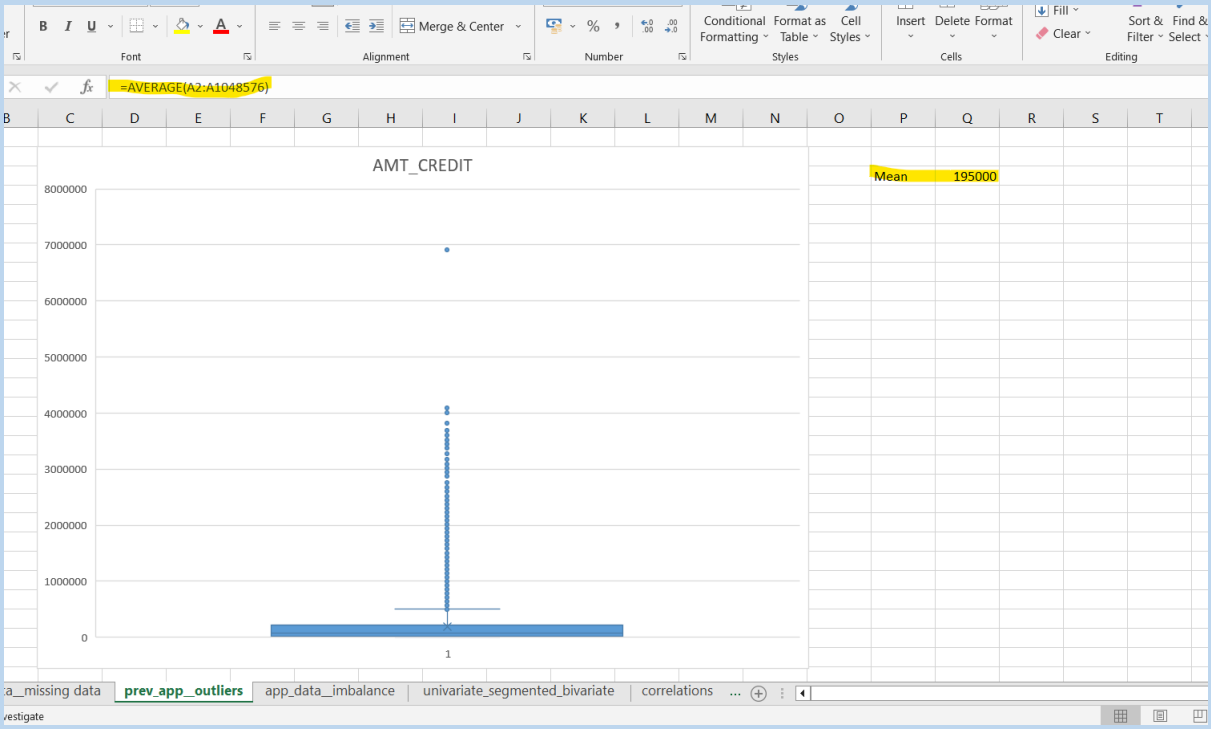
2. Identify **the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**

I found out the blank percentage and median of the column and filled the empty spaces there. (This is just for one table. Actual cleaning and filling of data is shown in excel file attached for other columns).

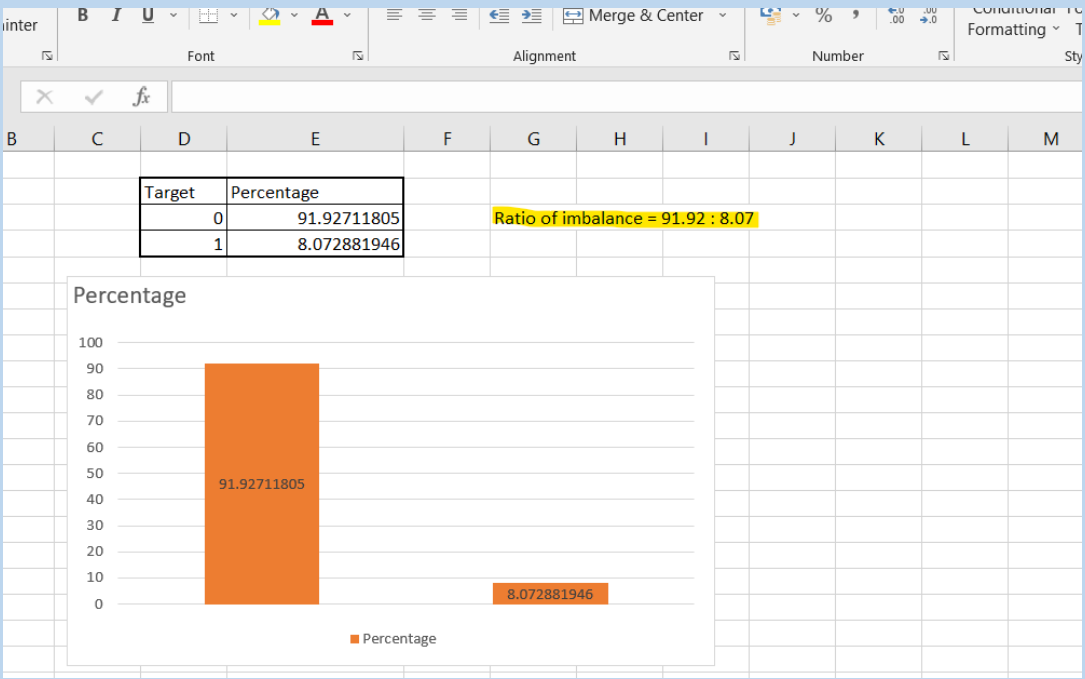


3. **Identify if there are** outliers **in the dataset. Also, mention why do you think it is an outlier.**

Foe Numerical columns, I found out the outliers and chose the value for the upper whisker as shown below. The credit amount value above 195000 is considered to be an upper whisker.



4. **Identify if there is data imbalance in the data. Find the ratio of data imbalance. The ratio of imbalance for Target Table came out to be 91.92:8.07.**

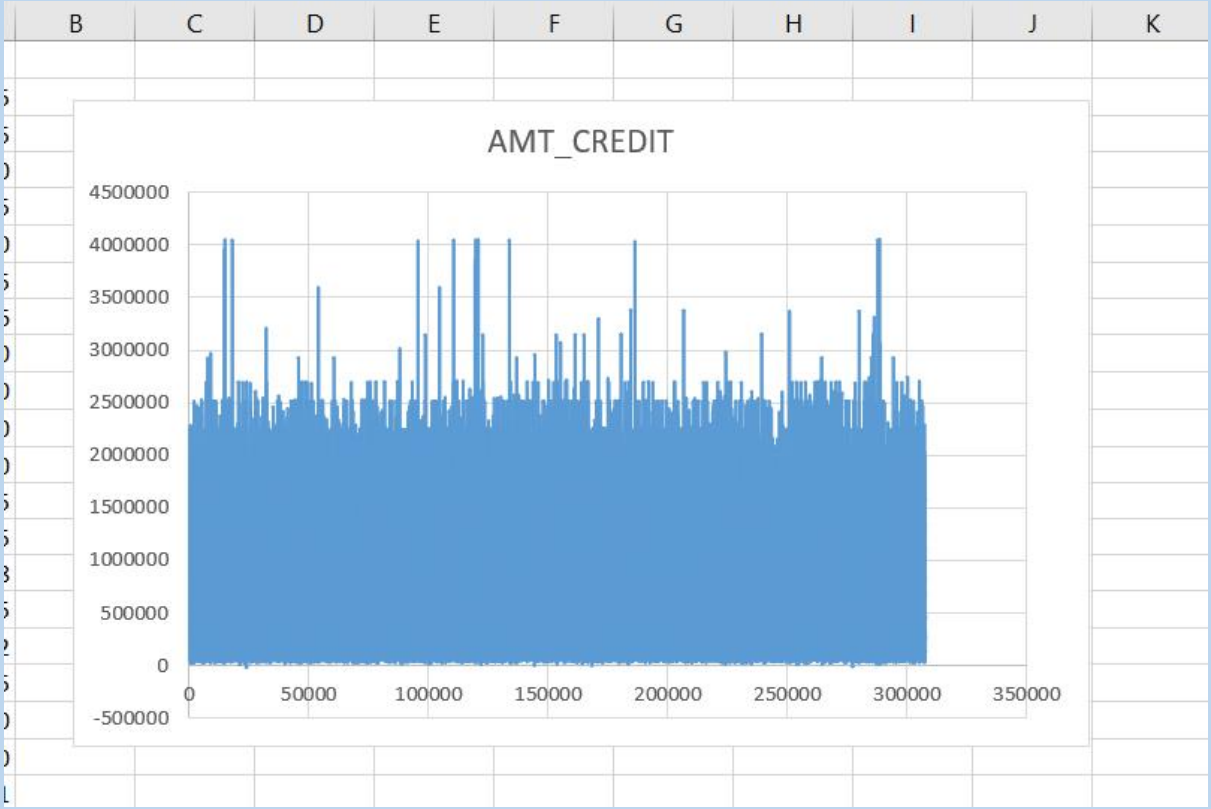


5. **Explain the** results of univariate, segmented univariate, bivariate analysis, etc. **in business terms.**

The **results of univariate, segmented univariate, bivariate analysis are as follows –**

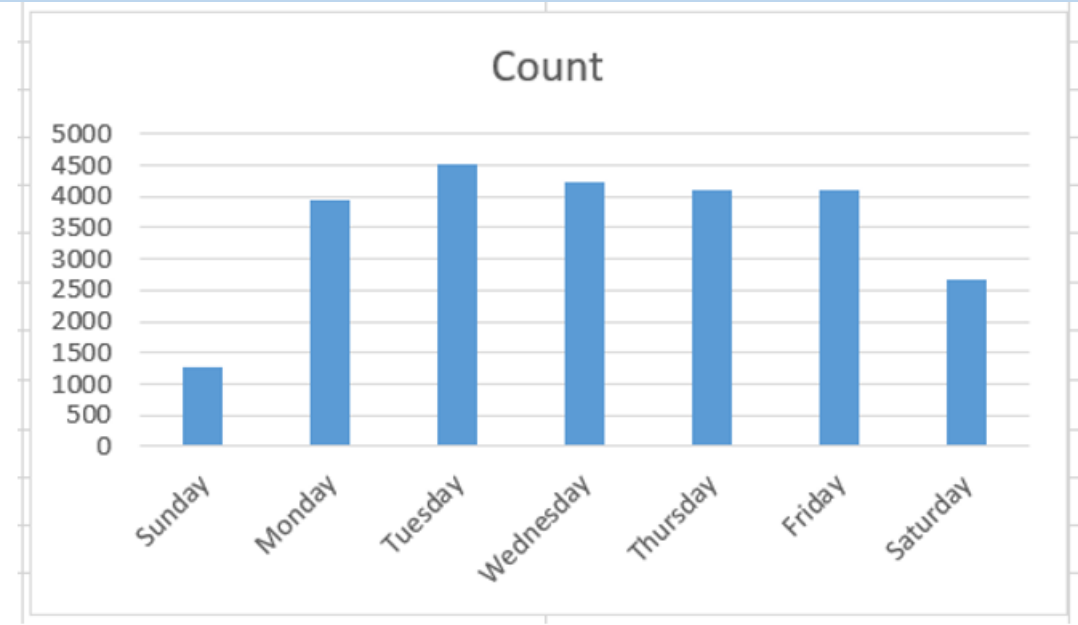
To perform the analysis, I first divided the data into two sets i.e. Target - 0 and Target – 1

AMT_CREDIT

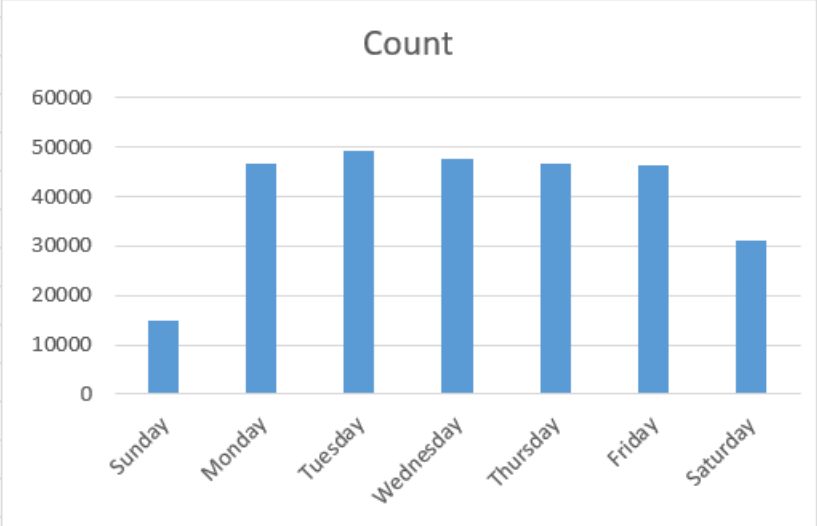


WEEKDAY_APPR_PROCESS_START

Target - 1	
WEEKDAY_APPR_PROCESS_START	**Count**
Sunday	1283
Monday	3934
Tuesday	4501
Wednesday	4238
Thursday	4098
Friday	4101
Saturday	2670



Target - 0	
WEEKDAY_APPR_PROCESS_START	**Count**
Sunday	14898
Monday	46780
Tuesday	49400
Wednesday	47696
Thursday	46493
Friday	46237
Saturday	31182



INSIGHTS – We can conclude that application starting process is less on Saturday and Sunday.

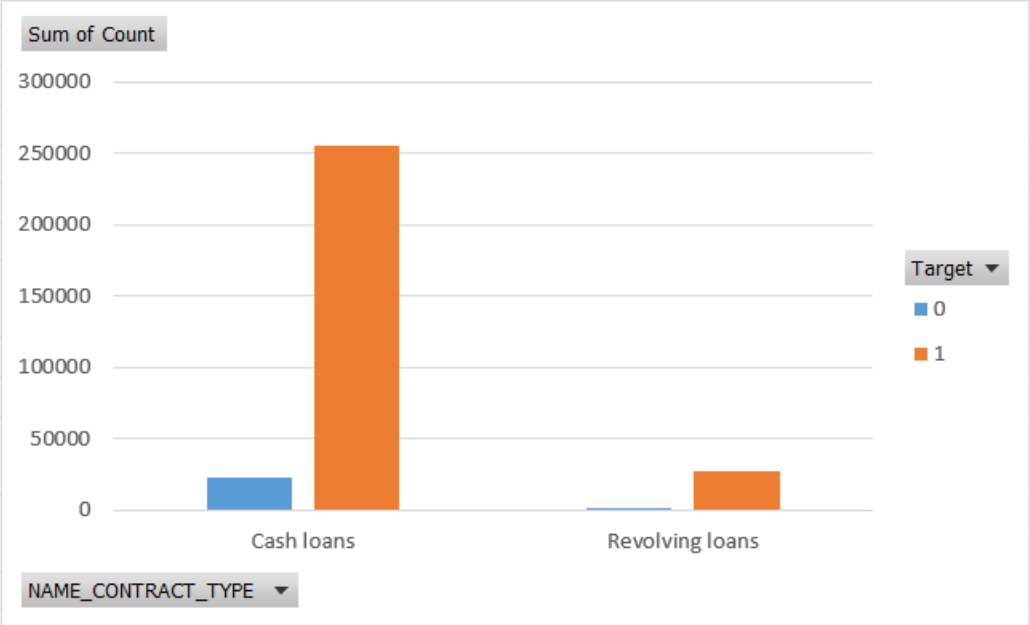
NAME_CONTRACT_TYPE

Target - 1	
NAME_CONTRACT_TYPE	**Count**
Cash loans	23221
Revolving loans	1604

Target - 0	
NAME_CONTRACT_TYPE	**Count**
Cash loans	255011
Revolving loans	27675

NAME_CONTRACT_TYPE	Count	Target
Cash loans	255011	1
Revolving loans	27675	1
Cash loans	23221	0
Revolving loans	1604	0

Sum of Count	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Cash loans	23221	255011	278232
Revolving loans	1604	27675	29279
Grand Total	**24825**	**282686**	**307511**



INSIGHTS – We can conclude that people prefer cash type loans more than other. People take more cash loans.

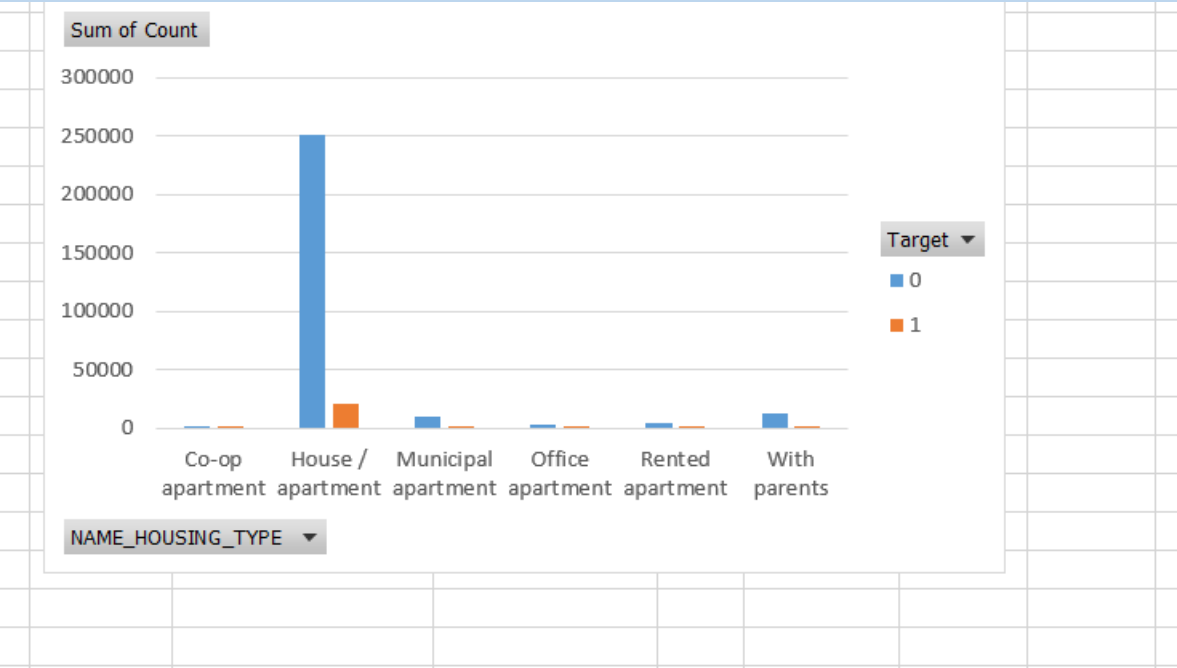
NAME_HOUSING_TYPE

Target - 1	
NAME_HOUSING_TYI	Count
House / apartment	21272
Co-op apartment	89
Municipal apartment	955
Office apartment	172
Rented apartment	601
With parents	1736

Target - 0	
NAME_HOUSING_TYI	Count
House / apartment	251596
Co-op apartment	1033
Municipal apartment	10228
Office apartment	2445
Rented apartment	4280
With parents	13104

NAME_HOUSING_TYI	Count	Target
House / apartment	21272	1
Co-op apartment	89	1
Municipal apartment	955	1
Office apartment	172	1
Rented apartment	601	1
With parents	1736	1
House / apartment	251596	0
Co-op apartment	1033	0
Municipal apartment	10228	0
Office apartment	2445	0
Rented apartment	4280	0
With parents	13104	0

Sum of Count	Column Labels		
Row Labels	0	1	Grand Total
Co-op apartment	1033	89	1122
House / apartment	251596	21272	272868
Municipal apartment	10228	955	11183
Office apartment	2445	172	2617
Rented apartment	4280	601	4881
With parents	13104	1736	14840
Grand Total	**282686**	**24825**	**307511**



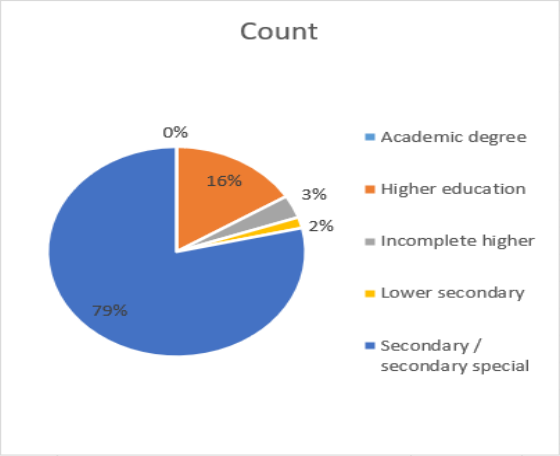
INSIGHTS – We can conclude that people living in houses fall in both the category of default loans and non-default loans.

6. Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable).

To find the correlation, we again divide the data into two sets based on Targets and consider Target – 1 as defaulters.

NAME_EDUCATION_TYPE

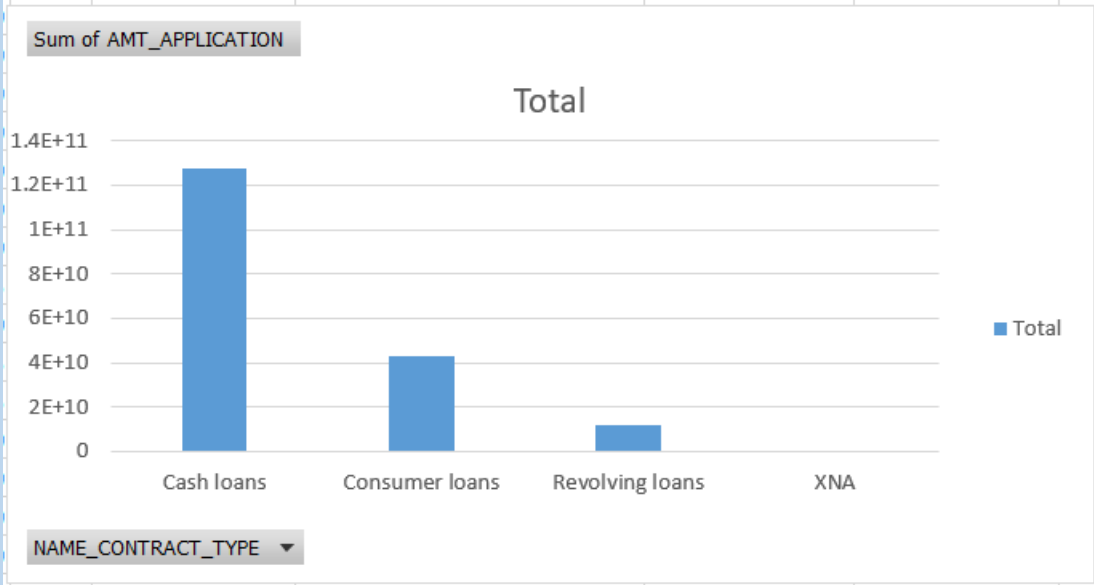
NAME_EDUCATION_TYPE	Count
Academic degree	3
Higher education	4009
Incomplete higher	872
Lower secondary	417
Secondary / secondary special	19524



INSIGHTS – We can find that people with education type as Secondary/Secondary Special are more likely to default and people with education type Academic degree default the least.

NAME_CONTRACT_TYPE and AMT_APPLICATION

Row Labels	Sum of AMT_APPLICATION
Cash loans	1.27622E+11
Consumer loans	43263132837
Revolving loans	11849616540
XNA	0
Grand Total	**1.82735E+11**



INSIGHTS – If we sum the total amount for loan in applications, we find that that people mostly take cash loans.

AMT_APPLICATION and AMT_CREDIT

AMT_APPLICATION	AMT_CREDIT		
17145	17145		Correlation Coefficient
607500	679671		0.975777217
112500	136444.5		
450000	470790		
337500	404055		
315000	340573.5		
0	0		
0	0		

INSIGHTS – We find that the correlation coefficient is 0.9758 using excel formula =CORREAL.

AMT_INCOME_TOTAL and AMT_ANNUITY

W	X	Y	Z
AMT_INCOME_TOTAL	AMT_ANNUITY		
202500	24700.5		Correlation Coefficient
270000	35698.5		0.191657428
67500	6750		
135000	29686.5		
121500	21865.5		
99000	27517.5		

INSIGHTS – We find that the correlation coefficient is 0.19166 using excel formula =CORREAL.

CONCLUSION – From the above analysis, we can find out what kind of people and can repay loan, what kinds of loan people prefer to take, people taking loans come from what background, what is their source of income, for what type of people, the loan applications are refused and based on which conditions.

RESULTS: -

1. People with academic degree have less defaults.
2. People prefer cash loans more than any other type.
3. People with secondary/secondary special as education type have more chances of defaulting loans.
4. People who have less than 5 years of employment have high default rate.
5. Focused variable for application file – Target.
6. Focused variable for Previous application file – NAME_CONTRACT_STATUS.
7. Important fields to consider for loan repayment are –
8. NAME_EDUCATION_TYPE
9. AMT_INCOME_TOTAL
10. DAYS_EMPLOYED
11. AMT_CREDIT
12. People with lower total income are more likely to default.
13. People with high Credit amount are less likely to default.

Module -7 Project: XYZ Ads Airing Report

This project is based on TV Ads airing report analysis. XYZ is an ads airing company. In this project we are provided with dataset having different TV Airing Brands, their product, their category. Dataset includes the network through which Ads are airing, types of networks like Cable/ Broadcast and the show name also on which Ads got aired. We can also see the data of Dayparts, Time zone and the time & date at which Ads got aired.

Here we have to analyse the brands and their advertisement strategies and most favourable brands and which brand has the highest share.

In this dataset, I first went through the data set to understand the details of the different variables and columns. I checked for any null values, missing or blank cells, duplicate data or if any data cleaning is required. After checking all these fields, I went up to perform the data analysis and answer the required questions. I used MS Excel 2019 to perform the analysis (used Pivot tables, different formulas) and MS Word 2019 to prepare a detailed report.

The questions asked along with their solutions are as follows -

Q1.) What is Pod Position? Does the Pod position number affect the amount spent on Ads for a specific period of time by a company? (Explain in Details with examples from the dataset provided)

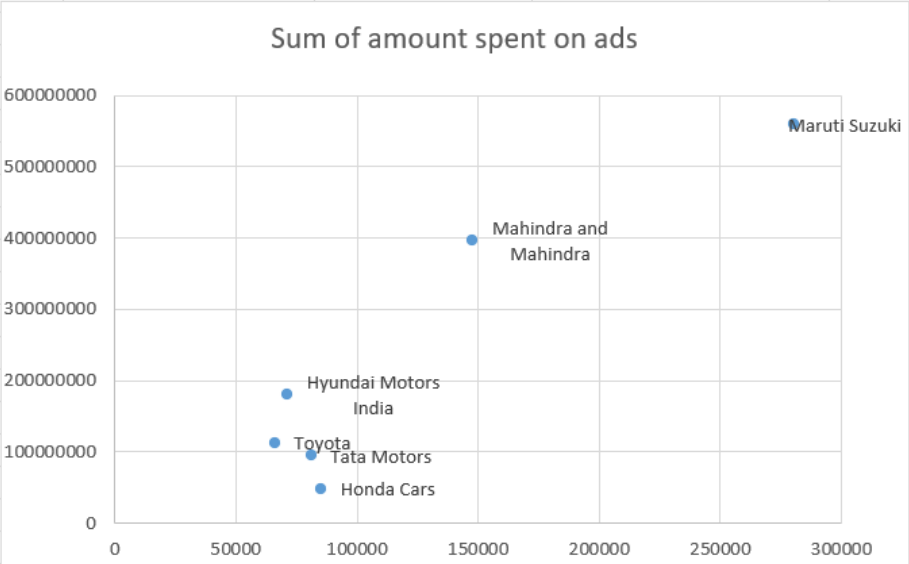
Ans - Ad Pods" or "Podding" is a term used to refer to multiple ads that are placed together and then played back one after the other (back-to-back playback) in a single ad break.

Few important things to know about ad pod: -

- Individual ads in an ad pod have "sequence numbers" and these numbers determine the order in which the player should playback the ads.
- If an ad cannot be played, then the player moves on to the next ad based on the sequence number.
- The player must attempt to play all the ads in the pod, unless, the ads cannot be played back or they do not fit into the ad slot. For example, if three 30-second ads are returned for a 60-second ad slot.
- If an ad cannot be played, the player can playback the next ad in the sequence or playback a stand-alone/non-sequenced ad

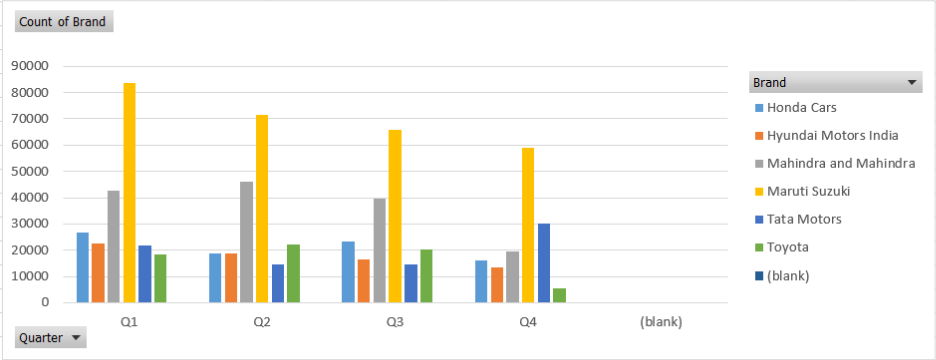
Yes, the Pod position number affects the amount spent on Ads for a specific period of time by the company.

Company	Count of Ads	Sum of amount spent on ads
Honda Cars	85265	48258340
Hyundai Motors India	71296	180808756
Mahindra and Mahindra	147890	397305655
Maruti Suzuki	280274	558646472
Tata Motors	81041	94790227
Toyota	66018	112653112



Q2.) What is the share of various brands in TV airings and how has it changed from Q1 to Q4 in 2021?

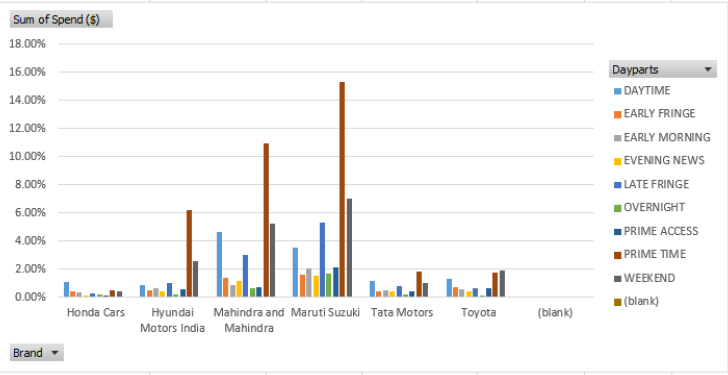
Count of Brand	Column Labels						
Row Labels	Honda Cars	Hyundai Motors India	Mahindra and Mahindra	Maruti Suzuki	Tata Motors	Toyota	(blank)	Grand Total
Q1	26839	22600	42522	83648	21836	18251		215696
Q2	18751	18887	46084	71632	14633	21981		191968
Q3	23450	16543	39788	65951	14499	20225		180456
Q4	16225	13266	19496	59043	30073	5561		143664
(blank)								
Grand Total	**85265**	**71296**	**147890**	**280274**	**81041**	**66018**		**731784**



- Maruti Suzuki has the highest share in almost all quarters followed by Mahindra and Mahindra.
- Toyota has the lowest share in TV Airings.

Q.3.) Conduct a competitive analysis for the brands and define advertisement strategy of different brands and how it differs across the brands.

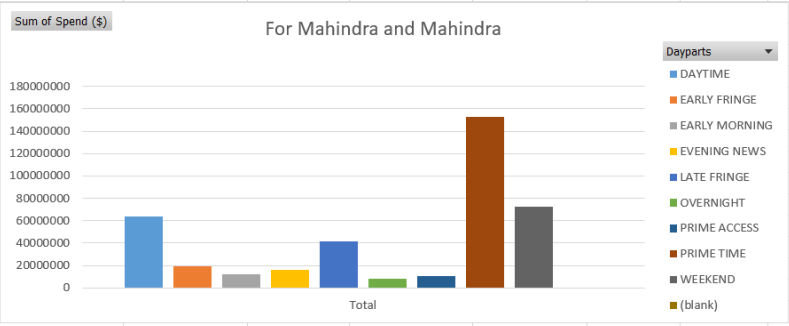
Sum of Spend ($)	Column Labels										
Row Labels	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	(blank)	Grand Total
Honda Cars	1.08%	0.41%	0.37%	0.15%	0.25%	0.20%	0.10%	0.50%	0.39%	0.00%	3.47%
Hyundai Motors India	0.89%	0.51%	0.63%	0.39%	0.98%	0.23%	0.55%	6.23%	2.58%	0.00%	12.98%
Mahindra and Mahindra	4.61%	1.38%	0.87%	1.15%	3.00%	0.62%	0.74%	10.97%	5.20%	0.00%	28.53%
Maruti Suzuki	3.50%	1.63%	2.08%	1.49%	5.32%	1.70%	2.08%	15.34%	6.98%	0.00%	40.12%
Tata Motors	1.19%	0.44%	0.51%	0.42%	0.80%	0.18%	0.42%	1.84%	1.01%	0.00%	6.81%
Toyota	1.33%	0.70%	0.60%	0.39%	0.64%	0.12%	0.64%	1.73%	1.93%	0.00%	8.09%
(blank)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Grand Total	**12.59%**	**5.08%**	**5.05%**	**3.99%**	**10.98%**	**3.05%**	**4.54%**	**36.62%**	**18.11%**	**0.00%**	**100.00%**



- Most of the brands spend most in daytime and least in overnight.
- Honda Cars spends relatively more in daytime than in other parts of day.
- Mahindra and Mahindra spend most in daytime.
- Maruti Suzuki spends most in primetime.

Q.4.) Mahindra and Mahindra want 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?

	Column Labels										
	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	(blank)	Grand Total
Sum of Spend ($)	64154402	19204408	12119383	16018235	41781609	8597788	10299276	152713257	72417297		397305655



Company gave most ads during the primetime.

However, they should also increase ads during other times of day as there is audience throughout the day.

Module -8 Project: ABC Call Volume Trend

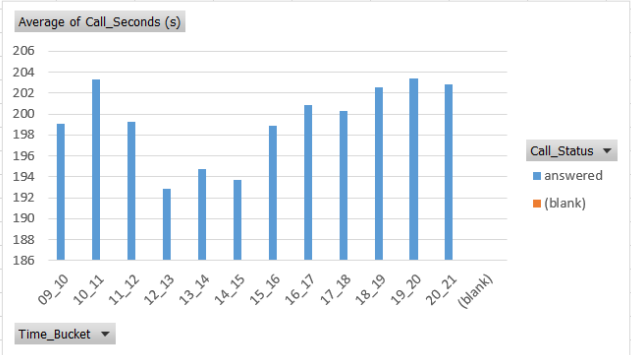
This project is based on call volume report analysis of a call center. ABC is a call center which has a separate team for voice process. In this project we are provided with dataset having details of agents, call duration, time duration, details on calls answered, abandoned and transferred. Data set also contains customer phone number, queue time, IVR time, date and time of call.

Here we have to analyse the rate of call which went unanswered and how many more agents are required to answer the call in both day and night shifts.

In this dataset, I first went through the data set to understand the details of the different variables and columns. I checked for any null values, missing or blank cells, duplicate data or if any data cleaning is required. After checking all these fields, I went up to perform the data analysis and answer the required questions. To perform the analysis, I used MS Excel 2019 for analysis and used MS Word to prepare the report.

Q1.) Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

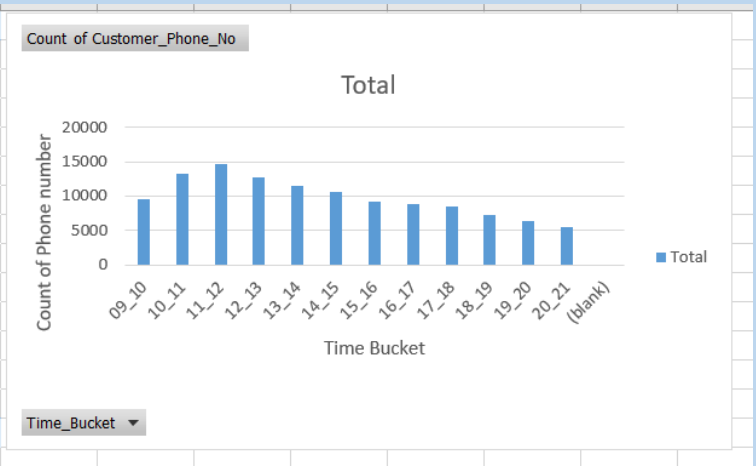
Average of Call_Seconds (s)	Column Labels		
Row Labels	**answered**	**(blank)**	**Grand Total**
09_10	199.0691057		199.0691057
10_11	203.3310302		203.3310302
11_12	199.2550234		199.2550234
12_13	192.8887829		192.8887829
13_14	194.7401744		194.7401744
14_15	193.6770755		193.6770755
15_16	198.8889175		198.8889175
16_17	200.8681864		200.8681864
17_18	200.2487831		200.2487831
18_19	202.5509677		202.5509677
19_20	203.4060725		203.4060725
20_21	202.845993		202.845993
(blank)			
Grand Total	**198.6227745**		**198.6227745**



- Average of call time duration answered by agents is 198.62 seconds.
- Average of call duration is highest between 10 to 11 am and 7 to 8 pm.

Q2.) Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]. You can select time in a bucket form (i.e. 1-2, 2-3, …..)

Row Labels	Count of Customer_Phone_No
09_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
(blank)	
Grand Total	**117988**



- The number of calls increases from 9 am to 12 noon and then decreases.

Q3.) 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.)

Description	unit	
Total no of working days per week	days	6
Total unplanned leaves per month	days	4
Total working hours	hrs	9
Time spent in lunch and snacks	hrs	1.5
Actual working hours	perc	60%
Avg time agent is occupied (60% of 7.5 hrs) daily	hrs	4.5
Total no. of days in a month	days	30

Time taken on average to answer calls (sec)	199.314176
Time requirement to answer 90% of the calls (hrs)	255.6204308
Total working person required per day	57

Day Calls	Avg answered calls	total calls	total calls in %	No. of agents required
09_10	199.0691057	9588	8.13%	5
10_11	203.3310302	13313	11.28%	6
11_12	199.2550234	14626	12.40%	7
12_13	192.8887829	12652	10.72%	6
13_14	194.7401744	11561	9.80%	6
14_15	193.6770755	10561	8.95%	5
15_16	198.8889175	9159	7.76%	4
16_17	200.8681864	8788	7.45%	4
17_18	200.2487831	8534	7.23%	4
18_19	202.5509677	7238	6.13%	3
19_20	203.4060725	6463	5.48%	3
20_21	202.845993	5505	4.67%	3
Average	**199.314176**		100.00%	57

Count of Duration(hh:r	Call Status			
Days	**abandon**	**answered**	**transfer**	**Grand Total**
01-Jan	**684**	**3883**	**77**	**4644**
02-Jan	**356**	**2935**	**60**	**3351**
03-Jan	**599**	**4079**	**111**	**4789**
04-Jan	**595**	**4404**	**114**	**5113**
05-Jan	**536**	**4140**	**114**	**4790**
06-Jan	**991**	**3875**	**85**	**4951**
07-Jan	**1319**	**3587**	**42**	**4948**
08-Jan	**1103**	**3519**	**50**	**4672**
09-Jan	**962**	**2628**	**62**	**3652**
10-Jan	**1212**	**3699**	**72**	**4983**
11-Jan	**856**	**3695**	**86**	**4637**
12-Jan	**1299**	**3297**	**47**	**4643**
13-Jan	**738**	**3326**	**59**	**4123**
14-Jan	**291**	**2832**	**32**	**3155**
15-Jan	**304**	**2730**	**24**	**3058**
16-Jan	**1191**	**3910**	**41**	**5142**
17-Jan	**16636**	**5706**	**5**	**22347**
18-Jan	**1738**	**4024**	**12**	**5774**
19-Jan	**974**	**3717**	**12**	**4703**
20-Jan	**833**	**3485**	**4**	**4322**
21-Jan	**566**	**3104**	**5**	**3675**
22-Jan	**239**	**3045**	**7**	**3291**
23-Jan	**381**	**2832**	**12**	**3225**
Grand Total	**34403**	**82452**	**1133**	**117988**
	1496	3585	49	5130
	29.16%	69.88%	0.96%	
	30.00%	70%	1%	

- Total agents required to answer 90% of calls per day is 57.
- The amount of answered calls are 70%, abandon are 30% and transferred are 1% approximately.

Q4.) 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:

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

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

Call volume daily (9 AM - 9pm)	5130
If we provide support in night, (9 PM - 9 AM)	1539
Additional hours required	76.68613
Additional HC	17
Total HC	74
Note - all values are taken from previous sheet	
4.5 - avg time agent is occupied - taken from prev sheet	

Nights Call	Calls Distribution	Time Distribution	Agents Required
21_22	3	10%	2
22_23	3	10%	2
23_24	2	7%	1
00_01	2	7%	1
01_02	1	3%	1
2_3	1	3%	1
3_4	1	3%	1
4_5	1	3%	1
5_6	3	10%	2
6_7	4	13%	2
7_8	4	13%	2
8_9	5	17%	3
	30	100%	17

- First calculated the Time Distribution by dividing each calls distribution by total calls i.e. 30.
- Total agents required to answer 90% of calls at night is 17.

CONCLUSION

After performing all the projects, I came to know about the real-world application of this data. I came to know how these big product companies use data driven insights to get best results in minimum time. I came to know about the various applications and importance of tools like SQL, MS Excel, MS Word, MS PowerPoint, use of formulas, logics, commands, Pivot Tables, graphs etc. I also came to understand how meaningful results can be drawn form such huge amount of data if they are properly sorted and cleaned. They can give the most accurate insights about the operations of a company and what improvements can be made in order to grow.

THANK YOU