Data Analysis Portfolio

By-

Shubhanshu Pratap Singh

Professional Background

I am a B.Tech Graduate in Computer Science & System Engineering. I completed my Graduation in 2021 from Kalinga Institute of Industrial Technology, Bhubaneswar with an overall CGPA of 8.21.

While in college, I interned with Highradius Technologies, Bhubaneswar. During my internship also, I had a part in project where we applied Data Analysis and Machine Learning Algorithms while building and deploying an AI Enabled Fintech B2B Cloud Application.

I have also worked with Accenture as an Associate Software Engineer. I was a part of Release and Change Management team. I worked for a diamond client of retail and e-commerce domain requiring my knowledge of computer science. The project demanded a hands-on over Azure DevOps tool for release management domain. I also gathered knowledge for the change management domain

But, as a data enthusiast, I have been looking to transition into the field of Analytics. I believe I am a perfect fit in any company as a Data Analyst/Business Analyst, based upon my current skills. I am eagerly looking for an opportunity to prove my skills while also learn more from the real-life work, deliver good results and to support the organization that I work for.

Table of Contents

Data Analytics Process

Project Description-

The task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process. Prepare a PPT/PDF on a real-life scenario explaining it with the above process (Plan, Prepare, Process, Analyze, Share, Act) and submit it as part of this task.

Outcome-

I took a real-life example of Selecting a cinema hall for a movie and explained how we put data analysis to work for it. The process has been explained with the help of the process: Plan, Prepare, Process, Analyse, Share, Act

Link to the report-

https://docs.google.com/presentation/d/1sqhqWb9fosQN9wAfGblKPgp6fJOi06z2/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true

Instagram User Analytics

Project Description-

- In this project we are going to perform user analysis on the user database of Instagram to gain insights and key metrics by analyzing how users are engaging with the application.
- The insights gained will help the management and marketing team in deciding the future updates and features for the application to enhance user experience and increase user interaction, as well as, help in launching Ad campaigns for the business to grow and benefit our investors and the company.

The Problem-

A) Marketing: The marketing team wants to launch some campaigns, and they need your help with the following

1. Rewarding Most Loyal Users: People who have been using the platform for the longest time.
   Your Task: Find the 5 oldest users of the Instagram from the database provided
2. Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo.
   Your Task: Find the users who have never posted a single photo on Instagram
3. Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.
   Your Task: Identify the winner of the contest and provide their details to the team
4. Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.
   Your Task: Identify and suggest the top 5 most commonly used hashtags on the platform
5. Launch AD Campaign: The team wants to know which day would be the best day to launch ADs.
   Your Task: What day of the week do most users register on? Provide insights on when to schedule an ad campaign

B) Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds

1. User Engagement: Are users still as active and post on Instagram or they are making fewer posts
   Your Task: Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users
2. Bots & Fake Accounts: The investors want to know if the platform is crowded with fake and dummy accounts
   Your Task: Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).

Analysis-

To solve the problem MySQL Workbench 8.0.32 was used.

1. Rewarding most loyal users-

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

2. Remind Inactive users to start posting-

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

username
Aniya_Hackett
Kasandra_Homenick
Jaclyn81
Rocio33
Maxwell.Halvorson
Tierra.Trantow
Pearl7
Ollie_Ledner37
Mckenna17
David.Osinski47
Morgan.Kassulke
Linnea59
Duane60
Julien_Schmidt
Mike.Auer39

Result 1 | Read Only

3. Declaring Contest Winner –

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

username	id	image_url	likes
Kaley9	30	http://kenny.com	41

Result 7 | Read Only

4. Hashtag researching-

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

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

Result 6 × Read Only

5. Launch AD Campaign-

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

Day_of_week	Users_registered
Thursday	16

Result 7 × Read Only

Output

6. User engagement-

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

Posts_per_User
2.5700

Result 8 × Read Only

7. Bots & Fake Accounts-

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

username	Total_likes_given
Aniya_Hackett	257
Jaclyn81	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

Result Grid | Form Editor | Field Types | Query Stats

Result 4 | Read Only

Link to the report-
https://docs.google.com/presentation/d/1dlsDbfPR53n59ROWABQJboopZliifCL2/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true

Operation Analytics and Investigating Metric Spike

Project Description-

The project focuses on operation analytics and investigating metric spike. We will analyze the different datasets provided to us and will draw insights to answer the various questions raised by the departments.
Since, operation analytics is a complete end to end analysis of the operations of an organization, by performing it we will try to find the areas to improve upon. The insights derived by us will help the ops team, support team, marketing team and others to plan their course of action ahead. The analysis will also help in predicting the overall growth and decline of the organization's fortune.
We will also be investigating metric spike to answer questions related to engagement of the users with the services of our organization in a specified time interval.

The Problem-

I. Case Study 1 (Job Data)

A. Number of jobs reviewed: Amount of jobs reviewed over time.
Your task: Calculate the number of jobs reviewed per hour per day for November 2020?
B. Throughput: It is the no. of events happening per second.
Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?
C. Percentage share of each language: Share of each language for different contents.
Your task: Calculate the percentage share of each language in the last 30 days?
D. Duplicate rows: Rows that have the same value present in them.
Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

II. Case Study 2 (Investigating metric spike)

A. User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.
Your task: Calculate the weekly user engagement?
B. User Growth: Amount of users growing over time for a product.
Your task: Calculate the user growth for product?
C. Weekly Retention: Users getting retained weekly after signing-up for a product.
Your task: Calculate the weekly retention of users-sign up cohort?

D. Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.
   Your task: Calculate the weekly engagement per device?
E. Email Engagement: Users engaging with the email service.
   Your task: Calculate the email engagement metrics?

Analysis-

To solve the problem MySQL Workbench 8.0.32 was used.

1. Number of jobs reviewed

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

ds	Throughput
2020-11-30	144.0000
2020-11-29	180.0000
2020-11-28	218.1818
2020-11-27	34.6154
2020-11-25	80.0000

Result 11 | Read Only

2. Throughput

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

ds	7d_Rolling_Avg
2020-11-25	0.0222
2020-11-27	0.0134
2020-11-28	0.0220
2020-11-29	0.0248
2020-11-30	0.0264

3. Percentage share of each language

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

language	Percentage_share
Persian	33.3333
Arabic	16.6667
Hindi	16.6667
French	16.6667
Italian	16.6667

4. User Engagement

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

Week_no	Weekly_Active_Users
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104

Result 4

5. User Growth

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

Month_no	Users_registered	Perc_growth_in_user
1	712	NULL
2	685	-3.7921
3	765	11.6788
4	907	18.5621
5	993	9.4818
6	1086	9.3656
7	1281	17.9558
8	1347	5.1522
9	330	-75.5011
10	390	18.1818
11	399	2.3077
12	486	21.8045

Result 1

6. Weekly Retention

Week No	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16	Week 17	Week 18
17	740	472	324	251	205	187	167	146	145	145	136	131	132	143	116	91	82	77	5
18	788	362	261	203	168	147	144	127	113	122	106	118	127	110	97	85	67	4	0
19	601	284	173	153	114	95	91	81	95	82	68	65	63	42	51	49	2	0	0
20	555	223	165	121	91	72	63	67	63	65	67	41	40	33	40	0	0	0	0
21	495	187	131	91	74	63	75	72	58	48	45	39	35	28	2	0	0	0	0
22	521	224	150	107	87	73	63	60	55	48	41	39	31	1	0	0	0	0	0
23	542	219	138	101	90	79	69	61	54	47	35	30	0	0	0	0	0	0	0
24	535	205	143	102	81	63	65	61	38	39	29	0	0	0	0	0	0	0	0
25	500	218	139	101	75	63	50	46	38	35	2	0	0	0	0	0	0	0	0
26	495	181	114	83	73	55	47	43	29	0	0	0	0	0	0	0	0	0	0
27	493	199	121	106	68	53	40	36	1	0	0	0	0	0	0	0	0	0	0
28	486	194	114	69	46	30	28	3	0	0	0	0	0	0	0	0	0	0	0
29	501	186	102	65	47	40	1	0	0	0	0	0	0	0	0	0	0	0	0
30	533	202	121	78	53	3	0	0	0	0	0	0	0	0	0	0	0	0	0
31	430	145	76	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	496	188	94	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	499	202	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	518	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

7. Weekly Engagement

Week_No	Dell_Inspiron_Notebook	iPhone_5	iPhone4S	Windows_Surface	Macbook_air	iPhone_5S	Macbook_Pro	Kindle_Fire	iPad_Mini	Nexus_7	Nexus_5	Samsung_Galaxy_S4	Lenovo_Thinkpad	Samsumg_Galaxy_Tablet	Acer_Aspire_Notebook	Asus_Chrom
17	46	65	21	10	54	42	143	6	19	18	40	52	86	8	20	21
18	77	113	46	10	121	73	252	27	30	30	73	82	153	11	33	42
19	83	115	44	16	112	79	266	21	36	41	87	91	178	6	41	27
20	84	125	55	21	119	79	256	23	32	32	103	93	173	9	40	41
21	80	137	45	17	110	74	247	30	23	29	91	84	167	6	47	38
22	92	125	45	15	145	71	251	21	34	45	96	105	176	10	41	52
23	103	152	53	14	124	79	266	25	33	36	88	99	176	14	43	49
24	99	142	53	22	152	79	255	25	39	49	87	101	165	11	40	43
25	105	137	40	22	121	78	275	24	30	51	89	99	197	12	47	38
26	89	152	50	21	134	94	269	26	43	46	87	112	192	12	35	49
27	89	163	67	33	142	83	302	25	35	40	84	116	202	15	49	52
28	103	151	61	33	148	93	295	31	35	39	85	122	220	9	49	50
29	113	144	60	28	148	90	295	37	34	45	77	123	209	13	53	49
30	127	152	65	19	159	103	322	25	35	62	84	103	206	9	60	56
31	113	135	56	19	147	71	321	14	27	38	69	100	207	8	55	56
32	104	119	34	10	125	67	307	12	30	25	67	82	179	6	55	62
33	110	110	35	15	133	65	312	14	28	30	70	80	191	12	46	49
34	105	101	50	18	136	70	292	13	25	33	70	90	193	14	63	47
35	9	2	6	3	10	3	17	3	2	2	4	6	16	0	3	6

8. Email Engagement

Week_no	Weekly_Digest_emails	Reengagement_emails	Opened_emails	Clickthrough_emails
18	2602	157	912	430
19	2665	173	972	477
20	2733	191	1004	507
21	2822	164	1014	443
22	2911	192	987	488
23	3003	197	1075	538
24	3105	226	1155	554
25	3207	196	1096	530
26	3302	219	1165	556
27	3399	213	1228	621
28	3499	213	1250	599
29	3592	213	1219	590
30	3706	231	1383	630
31	3793	222	1351	445
32	3897	200	1337	418
33	4012	264	1432	490
34	4111	261	1528	490
17	908	73	310	166
35	0	48	41	38

Link to the report-
https://docs.google.com/presentation/d/1-arwe9rl8K633BLAh4HUi6uZgrvyWXlx/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true

Hiring Process Analytics

Project Description

- Hiring process is an essential and integral function of a company. By analyzing their previous data records, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyze before hiring freshers or any other individual.
- The project is on Hiring Process Analytics where we will be performing Exploratory Data Analysis (EDA) on the dataset provided by the company. The dataset contains data related to previous hiring done by the company.

Software used for doing the overall Analysis: Microsoft Excel

The Problem-

A. Hiring: Process of intaking of people into an organization for different kinds of positions.
   Your task: How many males and females are Hired ?
B. Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.
   Your task: What is the average salary offered in this company ?
C. Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.
   Your task: Draw the class intervals for salary in the company ?
D. Charts and Plots: This is one of the most important part of analysis to visualize the data.
   Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?
E. Charts: Use different charts and graphs to perform the task representing the data.
   Your task: Represent different post tiers using chart/graph?

Analysis-

1. Hiring

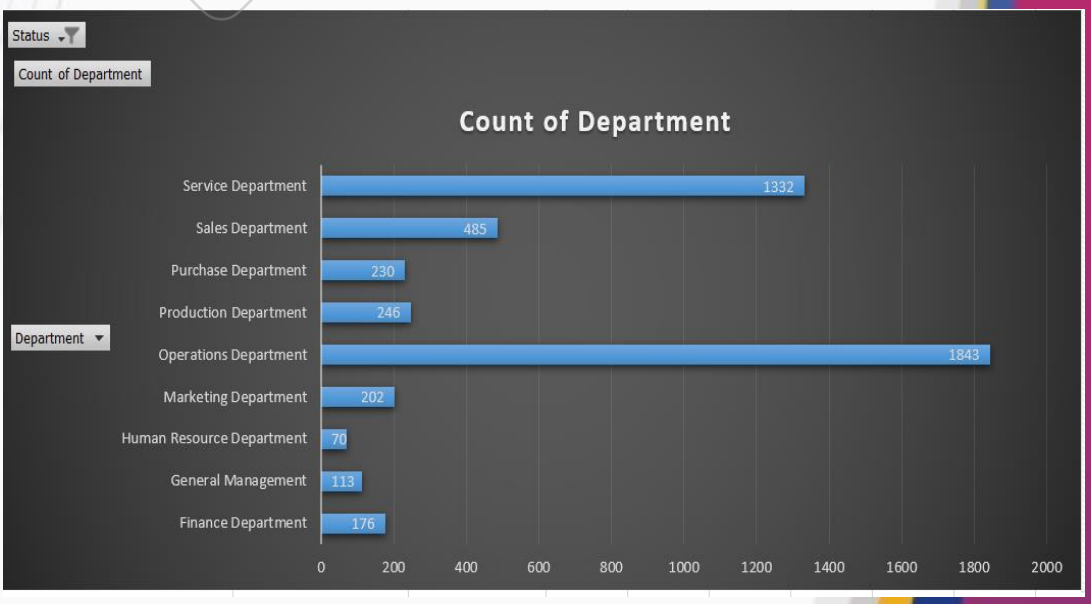
Total_Males_hired	Total_Females_hired
2563	1856

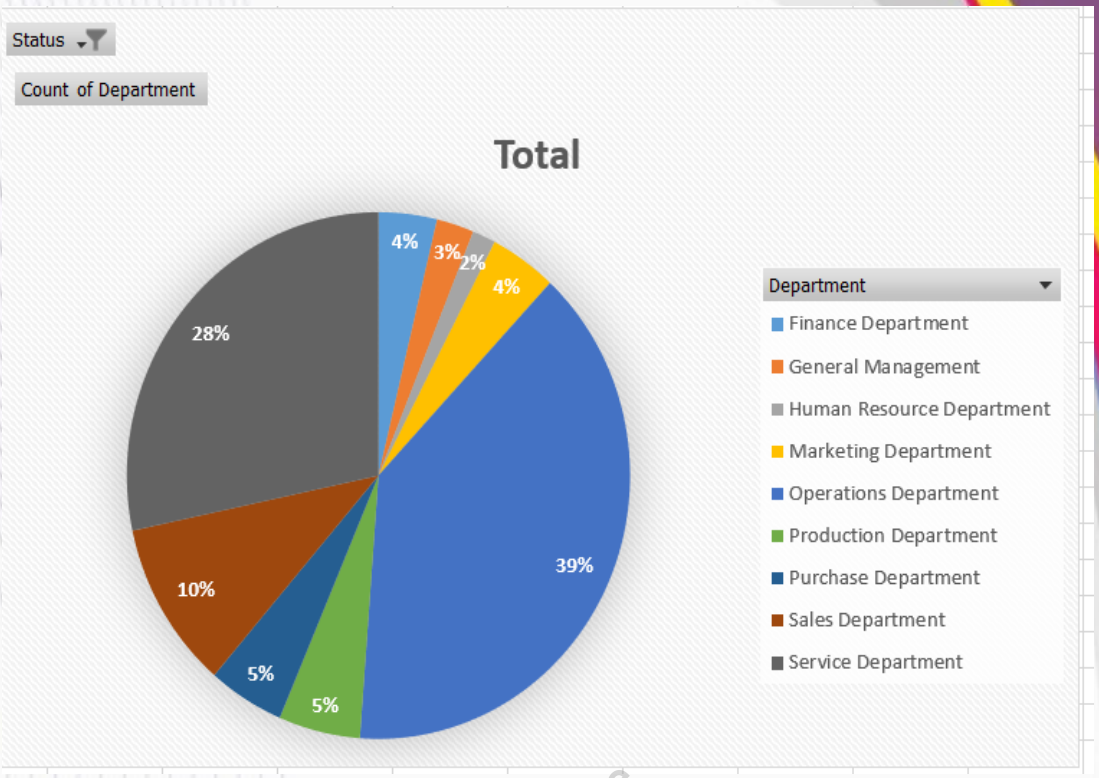
Average_Salary_Offered
49983.02902

2. Average Salary
3. Class Intervals

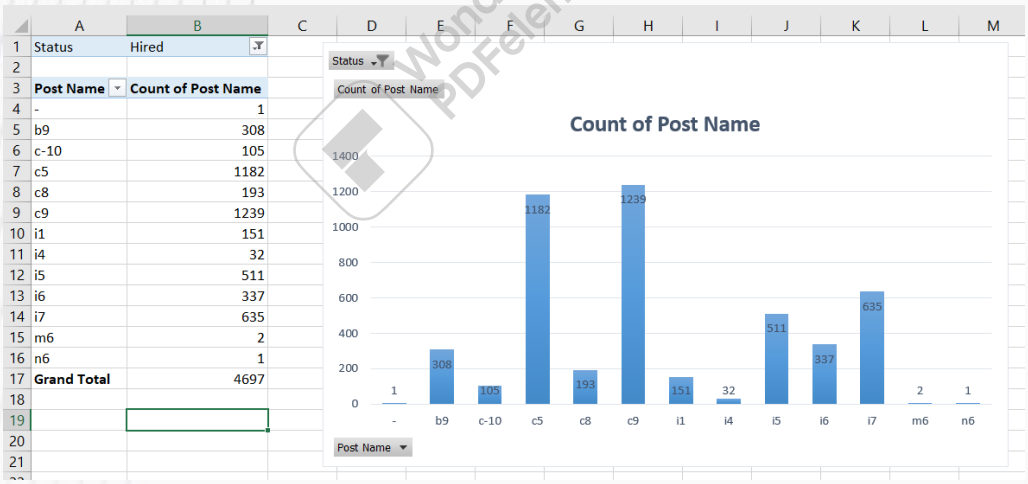
Status	Hired
Class Interval of Salary	**Number of Employees**
100-10099	444
10100-20099	487
20100-30099	457
30100-40099	488
40100-50099	523
50100-60099	496
60100-70099	450
70100-80099	479
80100-90099	462
90100-100099	408
190100-200099	1
290100-300099	1
390100-400099	1
Grand Total	**4697**

4. Charts and Plots





5. Charts-We can observe that most number of hired employees are in the post tier c9, that is 1239 employees.



Status	Hired
Post Name	**Count of Post Name**
-	1
b9	308
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
n6	1
Grand Total	4697

Link to the report-

https://docs.google.com/presentation/d/1D3XAjTmO0HWPkoa40DQYVYGjZWkiRQYS/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true

IMDB Movie Analysis

Project Description-

For the project, we have been provided with a dataset having various columns of different IMDB Movies. The data in these columns pertains Movie's name, Director's name, Actors, Budget, Gross, Language, Colour, Genre, etc.
Since, we are required to frame the problems for this task, we will need to define some problem we want to shed some light on.
Therefore, we will try to solve these problems by performing our Analysis on the data and visualizing it where necessary.

The problem-

A. Cleaning the data:This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
B. Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
C. Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
D. Best Directors: TGroup the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
E. Popular Genres: Perform this step using the knowledge gained while performing previous steps.
F. Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

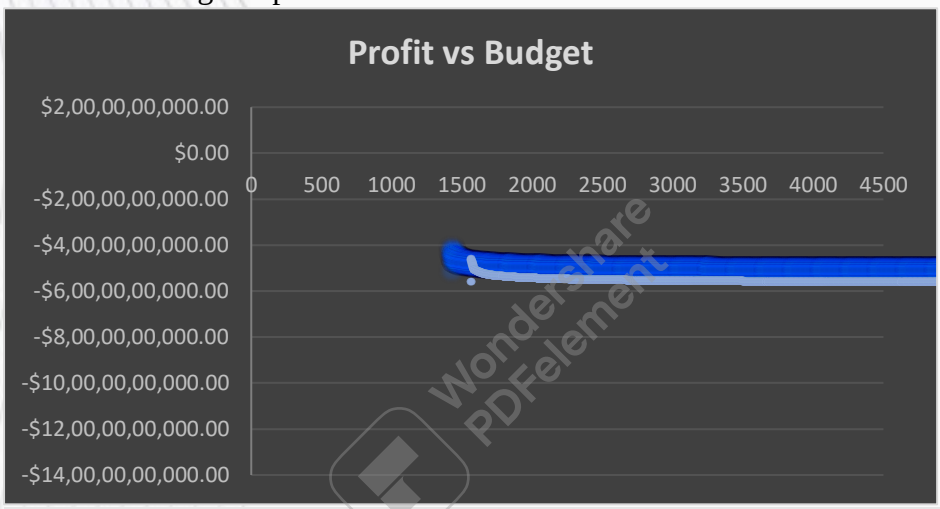
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Analysis -

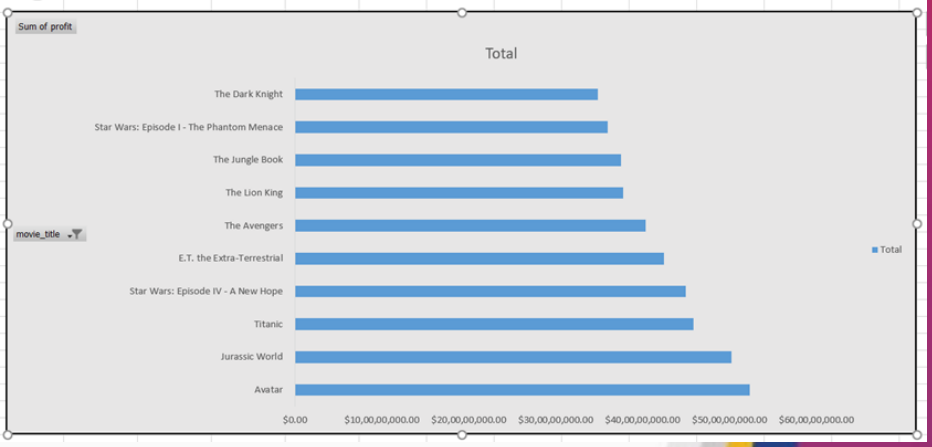
1. Cleaning the data

	A	B	C	D	E	F	G	H	I	J	K	L
1	director_nam	num_critic_for_revie	gross	genres	actor_1_name	movie_title	num_voted_use	num_user_for_revie	langu	budget	title_year	imdb_sco
2	Àmile Gaudreaul	67	6239558	Comedy|Drama	Paul Sorvino	Mambo Italiano	5548	67	English	5000000	2003	6.7
3	Àlex de la Iglesia	71	3607	Crime|Mystery|Thriller	Jim Carter	The Oxford Murders	22753	94	English	10000000	2008	6.1
4	Aaron Schneider	160	9176553	Drama|Mystery	Bill Murray	Get Low	19147	97	English	7500000	2009	7.1
5	Aaron Seltzer	99	48546578	Comedy|Romance	Alyson Hannigan	Date Movie	50415	613	English	20000000	2006	2.7
6	Abel Ferrara	48	1227324	Crime|Drama	Isabella Rossellini	The Funeral	6921	48	English	12500000	1996	6.6
7	Adam Carolla	14	105943	Comedy	Jay Mohr	Road Hard	1351	11	English	1500000	2015	6.1
8	Adam Goldberg	22	2580	Drama|Mystery	Judy Greer	I Love Your Work	1618	40	English	1650000	2003	5.4
9	Adam Marcus	112	15935068	Fantasy|Horror|Thriller	Kane Hodder	Jason Goes to Hell: The Final Friday	19331	317	English	2500000	1993	4.3
10	Adam McKay	265	119219978	Action|Comedy|Crime	Dwayne Johnson	The Other Guys	189806	316	English	100000000	2010	6.7
11	Adam McKay	164	148213377	Action|Comedy|Sport	Will Ferrell	Talladega Nights: The Ballad of Ricky	130776	437	English	73000000	2006	6.6
12	Adam McKay	173	100468793	Comedy	Will Ferrell	Step Brothers	212499	277	English	65000000	2008	6.9
13	Adam McKay	272	2175312	Comedy	Harrison Ford	Anchorman 2: The Legend Continue	131227	346	English	50000000	2013	6.3
14	Adam McKay	426	70235322	Biography|Comedy|Drama|History	Ryan Gosling	The Big Short	182983	374	English	28000000	2015	7.8
15	Adam McKay	181	84136909	Comedy	Darcy Donavan	Anchorman: The Legend of Ron Burg	267921	577	English	26000000	2004	7.2
16	Adam Rapp	50	101228	Comedy|Drama	Zooey Deschanel	Winter Passing	7228	53	English	3500000	2005	6.4
17	Adam Rifkin	42	4193025	Comedy|Music	Natasha Lyonne	Detroit Rock City	30682	194	English	15000000	1999	6.8
18	Adam Shankman	144	109993847	Comedy|Family|Fantasy|Romance	Adam Sandler	Bedtime Stories	72326	116	English	80000000	2008	6.1
19	Adam Shankman	219	118823091	Comedy|Drama|Family|Music|Musical|Romance	Jerry Stiller	Hairspray	98693	401	English	75000000	2007	6.7
20	Adam Shankman	360	38509342	Comedy|Drama|Musical|Romance	James Martin Kelly	Rock of Ages	61995	356	English	75000000	2012	5.9
21	Adam Shankman	77	82569532	Adventure|Comedy|Family	Taylor Lautner	Cheaper by the Dozen 2	42737	128	English	60000000	2005	5.4
22	Adam Shankman	125	113006880	Action|Comedy|Drama|Family|Thriller	Vin Diesel	The Pacifier	66308	225	English	56000000	2005	5.5
23	Adam Shankman	108	60400856	Comedy|Romance	Matthew McConaughey	The Wedding Planner	61317	229	English	35000000	2001	5.2
24	Adam Shankman	121	132541238	Comedy	Angus T. Jones	Bringing Down the House	30058	236	English	33000000	2003	5.5
25	Adam Shankman	80	41227069	Drama|Romance	Lauren German	A Walk to Remember	162701	962	English	11000000	2002	7.4
26	Adrian Lyne	152	52752475	Drama|Thriller	Olivier Martinez	Unfaithful	63067	533	English	50000000	2002	6.7
27	Adrian Lyne	38	6734844	Drama|Romance	David Margulies	9½ Weeks	29591	118	English	17000000	1986	5.9
28	Adrian Lyne	94	156645693	Drama|Romance|Thriller	Fred Gwynne	Fatal Attraction	55101	201	English	14000000	1987	6.9

2. Movies with highest profit



Row Labels	Sum of profit
Avatar	$52,35,05,847.00
Jurassic World	$50,21,77,271.00
Titanic	$45,86,72,302.00
Star Wars: Episode IV - A New Hope	$44,99,35,665.00
E.T. the Extra-Terrestrial	$42,44,49,459.00
The Avengers	$40,32,79,547.00
The Lion King	$37,77,83,777.00
The Jungle Book	$37,52,90,282.00
Star Wars: Episode I - The Phantom Menace	$35,95,44,677.00
The Dark Knight	$34,83,16,061.00
Grand Total	**$4,22,29,54,888.00**



3. Top 250

	A	B	C	D	E
1	Rank	IMDb_Top_250	imdb_score	num_voted_users	language
2	1	The Shawshank Redemption	9.3	1689764	English
3	2	The Godfather	9.2	1155770	English
4	3	The Dark Knight	9	1676169	English
5	3	The Godfather: Part II	9	790926	English
6	5	The Lord of the Rings: The Return of the King	8.9	1215718	English
7	5	Pulp Fiction	8.9	1324680	English
8	5	The Good, the Bad and the Ugly	8.9	503509	Italian
9	5	Schindler's List	8.9	965020	English
10	9	Inception	8.8	1468200	English
11	9	Fight Club	8.8	1347461	English
12	9	Star Wars: Episode V - The Empire Strikes Back	8.8	837759	English
13	9	The Lord of the Rings: The Fellowship of the Ring	8.8	1238746	English
14	9	Forrest Gump	8.8	1251222	English
15	14	Seven Samurai	8.7	229012	Japanese
16	14	City of God	8.7	533200	Portuguese
17	14	Star Wars: Episode IV - A New Hope	8.7	911097	English
18	14	The Matrix	8.7	1217752	English
19	14	Goodfellas	8.7	728685	English
20	14	One Flew Over the Cuckoo's Nest	8.7	680041	English
21	14	The Lord of the Rings: The Two Towers	8.7	1100446	English
22	21	The Usual Suspects	8.6	740918	English
23	21	Modern Times	8.6	143086	English
24	21	Interstellar	8.6	928227	English
25	21	Se7en	8.6	1023511	English
26	21	Spirited Away	8.6	417971	Japanese
27	21	The Silence of the Lambs	8.6	887467	English
28	21	Saving Private Ryan	8.6	881236	English
29	21	American History X	8.6	782437	English
30	29	Psycho	8.5	422432	English
31	29	The Dark Knight Rises	8.5	1144337	English
32	29	The Prestige	8.5	844052	English
33	29	Memento	8.5	845580	English
34	29	Whiplash	8.5	399138	English
35	29	The Lives of Others	8.5	259379	German
36	29	Apocalypse Now	8.5	450676	English
37	29	The Green Mile	8.5	782610	English
38	29	Terminator 2: Judgment Day	8.5	744891	English
39	29	Children of Heaven	8.5	27882	Persian
40	29	The Departed	8.5	873649	English
41	29	Django Unchained	8.5	955174	English
42	29	Gladiator	8.5	982637	English
43	29	Alien	8.5	563827	English
44	29	Back to the Future	8.5	732212	English
45	29	The Lion King	8.5	644348	English
46	29	The Pianist	8.5	497946	English
47	29	Samsara	8.5	22457	None
48	29	Raiders of the Lost Ark	8.5	661017	English
49	48	WALL·E	8.4	718837	English
50	48	A Separation	8.4	151812	Persian
51	48	Oldboy	8.4	356181	Korean
52	48	Requiem for a Dream	8.4	573541	English
53	48	Lawrence of Arabia	8.4	192775	English
54	48	Princess Mononoke	8.4	221552	Japanese
55	48	Aliens	8.4	488537	English
56	48	AmÃ©lie	8.4	534262	French
57	48	The Other Dream Team	8.4	3086	English
58	48	Braveheart	8.4	736638	English
59	48	Reservoir Dogs	8.4	664719	English
60	48	Star Wars: Episode VI - Return of the Jedi	8.4	681857	English
61	48	Baahubali: The Beginning	8.4	62756	Telugu
62	48	American Beauty	8.4	822500	English
63	48	Once Upon a Time in America	8.4	221000	English
64	48	Das Boot	8.4	168203	German
65	64	Some Like It Hot	8.3	175196	English
66	64	Scarface	8.3	537442	English
67	64	No End in Sight	8.3	7314	English
68	64	Batman Begins	8.3	980946	English
69	64	Unforgiven	8.3	277505	English
70	64	L.A. Confidential	8.3	414219	English
71	64	Metropolis	8.3	111841	German
72	64	The Sting	8.3	175607	English
73	64	Good Will Hunting	8.3	604904	English
74	64	Snatch	8.3	600996	English
75	64	Toy Story	8.3	623757	English
76	64	Toy Story 3	8.3	544884	English
77	64	Room	8.3	161288	English
78	64	Raging Bull	8.3	235133	English
79	64	Eternal Sunshine of the Spotless Mind	8.3	666937	English
80	64	Amadeus	8.3	270790	English
81	64	Downfall	8.3	249354	German
82	64	Up	8.3	665575	English
83	64	Inside Out	8.3	345198	English
84	64	Inglourious Basterds	8.3	885175	English
85	64	2001: A Space Odyssey	8.3	427357	English
86	64	Hoop Dreams	8.3	18980	English
87	64	Indiana Jones and the Last Crusade	8.3	515306	English
88	64	Monty Python and the Holy Grail	8.3	382240	English
89	64	The Hunt	8.3	170155	Danish
90	89	Finding Nemo	8.2	692482	English
91	89	Captain America: Civil War	8.2	272670	English
92	89	Gran Torino	8.2	561773	English
93	89	Trainspotting	8.2	469561	English
94	89	The Bridge on the River Kwai	8.2	149444	English
95	89	How to Train Your Dragon	8.2	485430	English
96	89	Incendies	8.2	80429	French
97	89	On the Waterfront	8.2	100890	English
98	89	Warrior	8.2	332276	English
99	89	Pan's Labyrinth	8.2	467234	Spanish
100	89	Lock, Stock and Two Smoking Barrels	8.2	414976	English
101	89	Howl's Moving Castle	8.2	214091	Japanese
102	89	V for Vendetta	8.2	791783	English

movies which were of foreign origin in the list of Top 250 movies

F	G	H
language	(Multiple Item	
Row Labels	Rating	
The Good, the Bad and the Ugly	8.9	
Seven Samurai	8.7	
City of God	8.7	
Spirited Away	8.6	
Samsara	8.5	
Children of Heaven	8.5	
The Lives of Others	8.5	
AmÃ©lie	8.4	
A Separation	8.4	
Baahubali: The Beginning	8.4	
Princess Mononoke	8.4	
Oldboy	8.4	
Das Boot	8.4	
The Hunt	8.3	
Downfall	8.3	
Metropolis	8.3	
Howl's Moving Castle	8.2	
The Act of Killing	8.2	
Incendies	8.2	
Pan's Labyrinth	8.2	
The Secret in Their Eyes	8.2	
The Sea Inside	8.1	
Akira	8.1	
Amores Perros	8.1	
The Celebration	8.1	
Elite Squad	8.1	
Tae Guk Gi: The Brotherhood of \	8.1	
A Fistful of Dollars	8	
Persepolis	8	
Central Station	8	
Waltz with Bashir	8	
My Name Is Khan	8	
Hero	7.9	
Amour	7.9	
Crouching Tiger, Hidden Dragon	7.9	
The Chorus	7.9	
The Second Mother	7.9	
Letters from Iwo Jima	7.9	
Veer-Zaara	7.9	
Ernest & Celestine	7.9	
4 Months, 3 Weeks and 2 Days	7.9	
Nine Queens	7.9	

4. Best Directors

Director	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4

5. Popular Genres

Genres	genre1	genre2	genre3	genre4	genre5	genre6	genre7	genre8	Total
Drama	690	913	293	42	2	0	0	0	1940
Comedy	1028	291	164	19	0	0	0	0	1502
Thriller	3	140	447	387	121	12	2	2	1114
Action	962	0	0	0	0	0	0	0	962
Romance	3	307	372	147	35	8	5	1	878
Adventure	375	412	0	0	0	0	0	0	787
Crime	255	342	101	13	1	0	0	0	712
Fantasy	37	131	186	90	56	13	1	0	514
Sci-Fi	8	90	229	101	43	19	7	0	497
Family	3	137	152	125	30	3	0	0	450
Horror	160	143	73	11	4	0	0	0	391
Mystery	23	176	113	55	10	4	1	0	382
Biography	207	33	3	0	0	0	0	0	243
Animation	46	125	28	0	0	0	0	0	199
Musical	2	18	29	25	18	9	2	0	103
Documentary	40	17	7	0	0	0	0	0	64
Western	3	8	18	12	14	3	0	0	58

6. Charts

Actor Name	Meryl_Streep	Leonardo_DiCaprio	Brad_Pitt
Movie Done	It'S Complicated	Titanic	The Curious Case Of Benjamin Button
	The River Wild	The Great Gatsby	Troy
	Julie & Julia	Inception	Ocean'S Twelve
	The Devil Wears Prada	The Revenant	Mr. & Mrs. Smith
	Lions For Lambs	The Aviator	Spy Game
	Out Of Africa	Django Unchained	Ocean'S Eleven
	Hope Springs	Blood Diamond	Fury
	One True Thing	The Wolf Of Wall Street	Seven Years In Tibet
	The Hours	Gangs Of New York	Fight Club
	The Iron Lady	The Departed	Sinbad: Legend Of The Seven Seas
	A Prairie Home Companion	Shutter Island	Interview With The Vampire: The Vampire Chronicles
		Body Of Lies	The Tree Of Life
		Catch Me If You Can	The Assassination Of Jesse James By The Coward Robert Ford
		The Beach	Babel
		Revolutionary Road	By The Sea
		The Man In The Iron Mask	Killing Them Softly
		J. Edgar	True Romance
		The Quick And The Dead	
		Marvin'S Room	
		Romeo + Juliet	
		The Great Gatsby	

7. movies which were the favorites of the critics and favorites of the users.

Critic Favourite	Albert Finney
Audience Favourite	Heather Donahue

Actor Name	Average of num_critic_for_reviews	Average of num_userr_for_review
Cch Pounder	260	966
Johnny Depp	250	580
Christoph Waltz	374	412
Tom Hardy	341	744
Daryl Sabara	374	451
J.K. Simmons	238	504
Brad Garrett	324	387
Chris Hemsworth	412	702
Alan Rickman	141	388
Henry Cavill	362	1067
Kevin Spacey	178	563
Giancarlo Giannini	403	1243
Peter Dinklage	154	198
Will Smith	222	563
Aidan Turner	381	737
Emma Stone	304	489
Mark Addy	234	581
Christopher Lee	155	1237

Link to the report-

https://docs.google.com/presentation/d/1gDJxS08QTQmC_AzWa1Zrd4jwu4ktJohw/edit?usp=drive_link&ouid=109923652620940529642&rtpof=true&sd=true

Bank Loan Case Study

Project Description-

We are provided with two datasets containing details about a bank loan for our final project. Due to people's inadequate or non-existent credit history, lending institutions struggle to give out loans. As a result, some customers take advantage of the situation by defaulting.
It aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

The problem-

- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

Analysis-

1. Identification and Handling of Missing data
   a. In ‘application_data.csv’

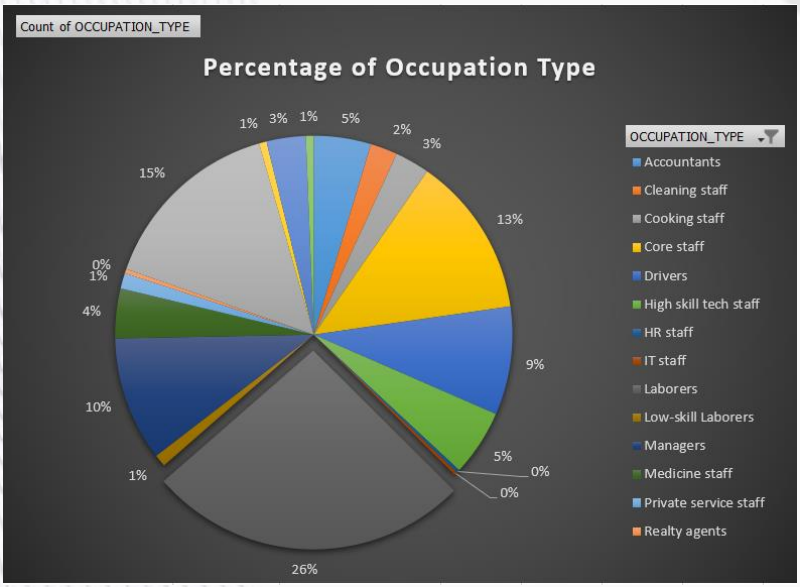
Columns	Missing Value cells count	Non missing value cells count	Percentage blank cells	Column dropped or not?		
COMMONAREA_AVG	214865	92647	69.87%	Yes, dropped due to high percentage of missing values	No.of columns dropped due to high % of missing values	49
COMMONAREA_MODE	214865	92647	69.87%	Yes, dropped due to high percentage of missing values	No of columns dropped as they are not needed	37
COMMONAREA_MEDI	214865	92647	69.87%	Yes, dropped due to high percentage of missing values	No. of columns to be used for analysis	36
NONLIVINGAPARTMENTS_AVG	213514	93998	69.43%	Yes, dropped due to high percentage of missing values	**Total**	**122**
NONLIVINGAPARTMENTS_MODE	213514	93998	69.43%	Yes, dropped due to high percentage of missing values		
NONLIVINGAPARTMENTS_MEDI	213514	93998	69.43%	Yes, dropped due to high percentage of missing values		
FONDKAPREMONT_MODE	210295	97217	68.39%	Yes, dropped due to high percentage of missing values		
LIVINGAPARTMENTS_AVG	210199	97313	68.35%	Yes, dropped due to high percentage of missing values		
LIVINGAPARTMENTS_MODE	210199	97313	68.35%	Yes, dropped due to high percentage of missing values		
LIVINGAPARTMENTS_MEDI	210199	97313	68.35%	Yes, dropped due to high percentage of missing values		
FLOORSMIN_AVG	208642	98870	67.85%	Yes, dropped due to high percentage of missing values		
FLOORSMIN_MODE	208642	98870	67.85%	Yes, dropped due to high percentage of missing values		
FLOORSMIN_MEDI	208642	98870	67.85%	Yes, dropped due to high percentage of missing values		
YEARS_BUILD_AVG	204488	103024	66.50%	Yes, dropped due to high percentage of missing values		
YEARS_BUILD_MODE	204488	103024	66.50%	Yes, dropped due to high percentage of missing values		
YEARS_BUILD_MEDI	204488	103024	66.50%	Yes, dropped due to high percentage of missing values		
OWN_CAR_AGE	202929	104583	65.99%	Yes, dropped due to high percentage of missing values		
LANDAREA_AVG	182590	124922	59.38%	Yes, dropped due to high percentage of missing values		
LANDAREA_MODE	182590	124922	59.38%	Yes, dropped due to high percentage of missing values		
LANDAREA_MEDI	182590	124922	59.38%	Yes, dropped due to high percentage of missing values		
BASEMENTAREA_AVG	179943	127569	58.52%	Yes, dropped due to high percentage of missing values		
BASEMENTAREA_MODE	179943	127569	58.52%	Yes, dropped due to high percentage of missing values		
BASEMENTAREA_MEDI	179943	127569	58.52%	Yes, dropped due to high percentage of missing values		
EXT_SOURCE_1	173378	134134	56.38%	Yes, dropped due to high percentage of missing values		
NONLIVINGAREA_AVG	169682	137830	55.18%	Yes, dropped due to high percentage of missing values		
NONLIVINGAREA_MODE	169682	137830	55.18%	Yes, dropped due to high percentage of missing values		
NONLIVINGAREA_MEDI	169682	137830	55.18%	Yes, dropped due to high percentage of missing values		
ELEVATORS_AVG	163891	143621	53.30%	Yes, dropped due to high percentage of missing values		
ELEVATORS_MODE	163891	143621	53.30%	Yes, dropped due to high percentage of missing values		
ELEVATORS_MEDI	163891	143621	53.30%	Yes, dropped due to high percentage of missing values		
WALLSMATERIAL_MODE	156341	151171	50.84%	Yes, dropped due to high percentage of missing values		

   b. In ‘previous_application.csv’

Columns	Missing Value cells count	Non missing value cells count	Percentage blank cells	Column dropped or no		
RATE_INTEREST_PRIMARY	1043561	5015	99.52%	Yes, dropped	No.of columns dropped due to high % of missing values	11
RATE_INTEREST_PRIVILEGED	1043561	5015	99.52%	Yes, dropped	No of columns dropped as they are not needed	11
AMT_DOWN_PAYMENT	895844	152732	85.43%	Yes, dropped	No. of columns to be used for analysis	15
RATE_DOWN_PAYMENT	895844	152732	85.43%	Yes, dropped	**Total**	**37**
DAYS_FIRST_DRAWING	673065	375511	64.19%	Yes, dropped		
DAYS_FIRST_DUE	673065	375511	64.19%	Yes, dropped		
DAYS_LAST_DUE_1ST_VERSION	673065	375511	64.19%	Yes, dropped		
DAYS_LAST_DUE	673065	375511	64.19%	Yes, dropped		
DAYS_TERMINATION	673065	375511	64.19%	Yes, dropped		
NFLAG_INSURED_ON_APPROVAL	673065	375511	64.19%	Yes, dropped		
NAME_TYPE_SUITE	616506	432070	58.79%	Yes, dropped		
AMT_GOODS_PRICE	385515	663061	36.77%	No		
AMT_ANNUITY	372232	676344	35.50%	No		
CNT_PAYMENT	372230	676346	35.50%	No		
PRODUCT_COMBINATION	346	1048230	0.03%	No		
AMT_CREDIT	1	1048575	0.00%	No		
SK_ID_PREV	0	1048576	0.00%	No		
SK_ID_CURR	0	1048576	0.00%	No		
NAME_CONTRACT_TYPE	0	1048576	0.00%	No		
AMT_APPLICATION	0	1048576	0.00%	No		
WEEKDAY_APPR_PROCESS_START	0	1048576	0.00%	No		
HOUR_APPR_PROCESS_START	0	1048576	0.00%	Yes, not needed		
FLAG_LAST_APPL_PER_CONTRACT	0	1048576	0.00%	Yes, not needed		
NFLAG_LAST_APPL_IN_DAY	0	1048576	0.00%	Yes, not needed		
NAME_CASH_LOAN_PURPOSE	0	1048576	0.00%	Yes, not needed		

2. Identification of Outliers-
   a. In ‘application_data.csv’

	AMT_ANNUITY	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE
Mean	27108.57391	168797.9193	599025.9997	538396.2075
min	1615.5	25650	45000	40500
q1	16524	112500	270000	238500
median	24903	147150	513531	450000
q3	34596	202500	808650	679500
max	258025.5	117000000	4050000	4050000
IQR	18072	90000	538650	441000
Upper Bound	61704	337500	1616625	1341000
Lower Bound	-10584	-22500	-537975	-423000

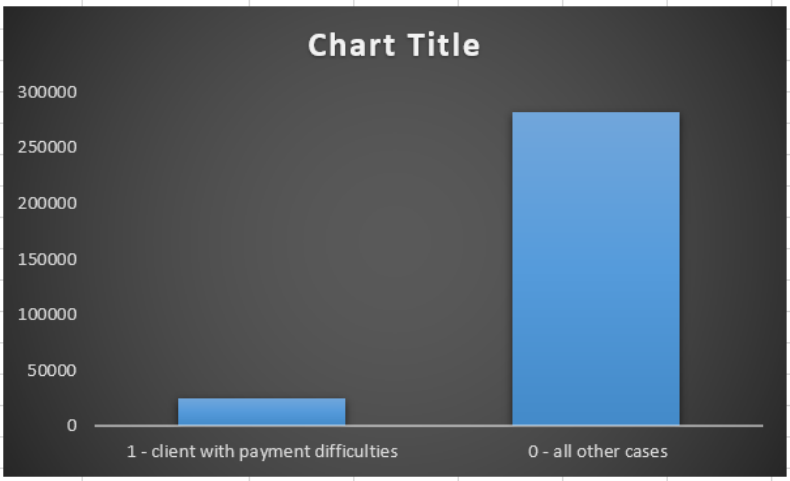


b. In ‘previous_application.csv’

	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_PAYMENT	AMT_CREDIT	AMT_APPLICATION
Q1	9000	90000	6	0	0
Median	16583.535	225000	12	101880	71946
Q3	28502.82	450000	36	337500	270000
IQR	19502.82	360000	30	337500	270000
Max	418058.145	6905160	84	6905160	6905160
Min	0	0	0	0	0
Lower Bounc	-20254.23	-450000	-39	-506250	-405000
Upper Bounc	57757.05	990000	81	843750	675000

3. Finding the Data Imbalance.

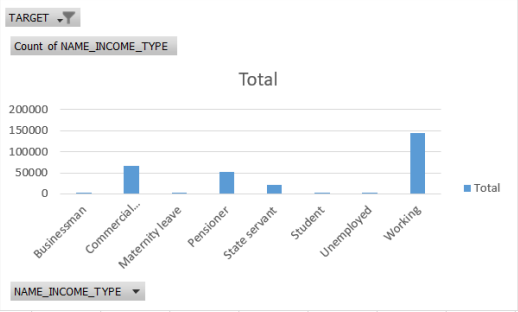
	A	B	C	D	E	F	G
1							
2		Count of Target		Ratio of Data Imbalance			
3	1 - client with payment difficulties	24825		8.07%			
4	0 - all other cases	282686		91.93%			
5	Total Clients	307511					
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							



4. Univariate Analyses
   a. In ‘application_data.csv’ –
      i. Univariate Analysis on Clients income type

TARGET	0
Client Type	**Count of NAME_INCOME_TYPE**
Businessman	10
Commercial associate	66257
Maternity leave	3
Pensioner	52380
State servant	20454
Student	18
Unemployed	14
Working	143550
Grand Total	**282686**

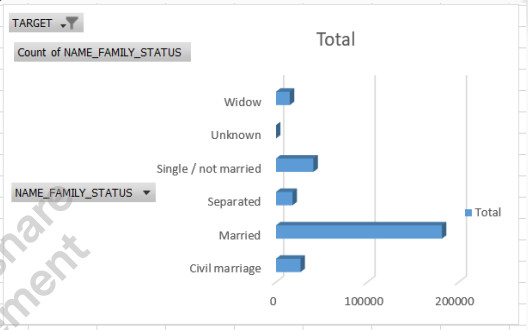
Defaulter & Non Defaulter included in All
1- Defaulters
0- Non-defaulters



      ii. Univariate Analysis on Family status of the client

TARGET	0
Row Labels	**Count of NAME_FAMILY_STATUS**
Civil marriage	26814
Married	181582
Separated	18150
Single / not married	40987
Unknown	2
Widow	15151
Grand Total	**282686**

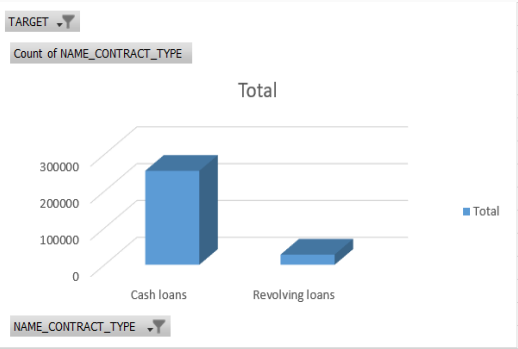
Defaulter & Non Defaulter included in All
1- Defaulters
0- Non-defaulters



      iii. Univariate Analysis on Clients type of loan

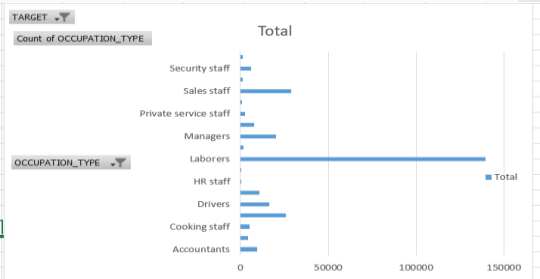
TARGET	0
Type of Loan	**Count of NAME_CONTRACT_TYPE**
Cash loans	255011
Revolving loans	27675
Grand Total	**282686**

Defaulter & Non Defaulter included in All
1- Defaulters
0- Non-defaulters



      iv. Univariate Analysis on kind of occupation does the client have

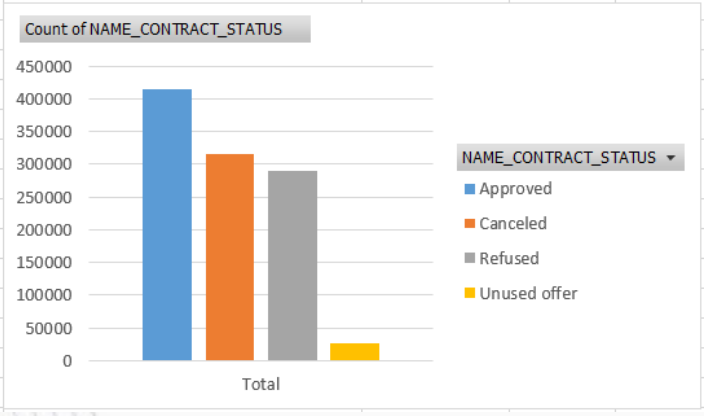
TARGET	0
Occupation	**Count of OCCUPATION_TYPE**
Accountants	9339
Cleaning staff	4206
Cooking staff	5325
Core staff	25832
Drivers	16496
High skill tech staff	10679
HR staff	527
IT staff	492
Laborers	139461
Low-skill Laborers	1734
Managers	20043
Medicine staff	7965
Private service staff	2477
Realty agents	692
Sales staff	29010
Secretaries	1213
Security staff	5999
Waiters/barmen staff	1196
Grand Total	**282686**



b. In ‘previous_application.csv’

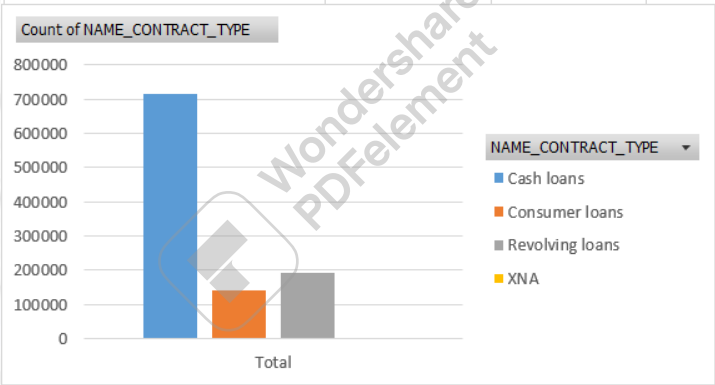
i. Univariate Analysis on contract status of client’s previous application.

	Column Labe			
	Approved	Canceled	Refused	Unused offer
Count of NAME_CONTRACT_STATUS	415142	316319	290678	26436



ii. Univariate Analysis on Contract product type of previous application

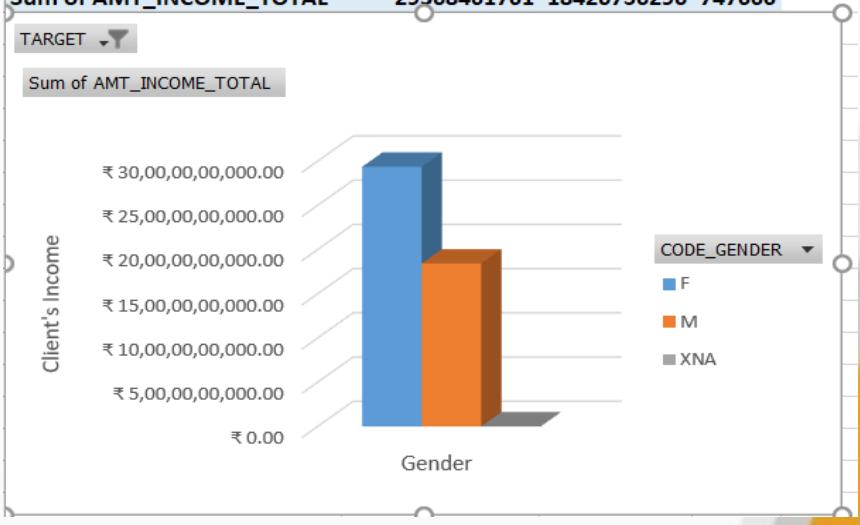
	Column Labels			
	Cash loans	Consumer loans	Revolving loans	XNA
Count of NAME_CONTRACT_TYPE	713940	141141	193148	346



5. Bivariate Analyses

a. Bivariate analysis of Income of the client on the basis of their gender

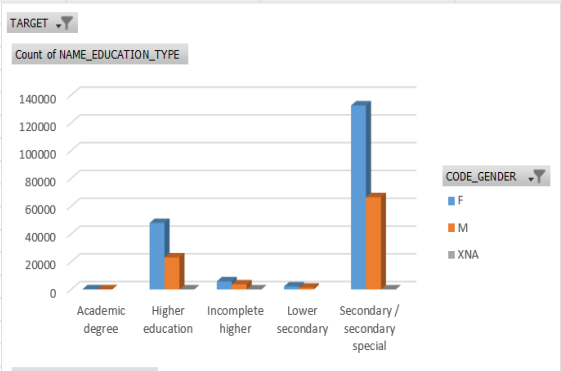
TARGET	0		
	Column Labels		
	F	M	XNA
Sum of AMT_INCOME_TOTAL	29368401701	18426756296	747000



b. Bivariate Analysis of Client's highest level of education achieved on the basis of gender

TARGET 0

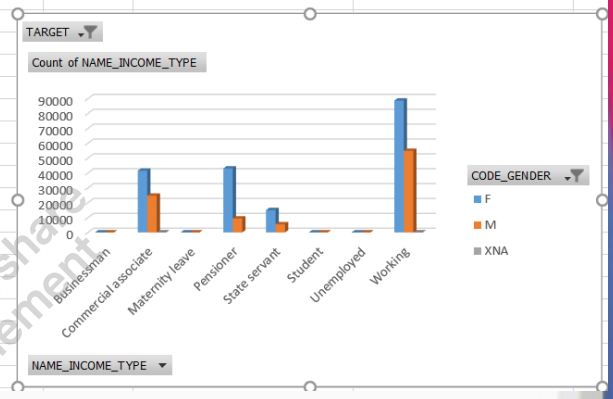
Count of NAME_EDUCATION_T	Column Labels		
Highest level of Education	F	M	XNA
Academic degree	97	64	
Higher education	47851	23002	1
Incomplete higher	5779	3625	1
Lower secondary	2097	1302	
Secondary / secondary special	132454	66411	2
Grand Total	**188278**	**94404**	**4**



c. Bivariate Analysis of client's income type on the basis of their gender

TARGET 0

Count of NAME_INCOME_TYPE	Column Labels		
Row Labels	F	M	XNA
Businessman	3	7	
Commercial associate	41551	24705	1
Maternity leave	2	1	
Pensioner	43018	9362	
State servant	15009	5445	
Student	7	11	
Unemployed	11	3	
Working	88677	54870	3
Grand Total	**188278**	**94404**	**4**



6. Top 10 Correlations
   a. Correlations of our Defaulter Clients

Top 10 Correlations		
AMT_GOODS_PRICE	AMT_CREDIT	0.983102519
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994436
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885176
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.77853974
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699196
AMT_ANNUITY	AMT_CREDIT	0.752194735
DAYS_EMPLOYED	DAYS_BIRTH	0.582185148
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497936543
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052287
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337181024

b. Correlations of our Non- Defaulter Clients

Top 10 Correlations		
AMT_GOODS_PRICE	AMT_CREDIT	0.987250457
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861361
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859331835
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381135
AMT_GOODS_PRICE	AMT_ANNUITY	0.776674087
AMT_ANNUITY	AMT_CREDIT	0.771296945
DAYS_EMPLOYED	DAYS_BIRTH	0.626113878
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446100857
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.43551371
AMT_ANNUITY	AMT_INCOME_TOTAL	0.418948086

Link to the report-

https://docs.google.com/presentation/d/13wz14dyNiYpTpUNZccm5TTvXuLPvLeyb/edit?usp=drive_link&ouid=109923652620940529642&rtpof=true&sd=true

Analyzing the Impact of Car Features on Price and Profitability

Project Description-

The automotive industry is evolving rapidly on multiple dimension, with the need of fuel-efficient automobile right now more than ever, technological innovation, adoption of environmentally sustainable practices and structural shifts with changing consumer preferences.

In recent times, there has been an increase in the demand of electric and hybrid vehicles as an alternative to gasoline powered vehicles due to their carbon emission being less compared to gasoline powered vehicle.

Therefore, on our client's demand to learn how they can optimize pricing and product development decisions to increase their profitability as well as meet consumers' demand, we are going to perform our analysis on the Impact of Car Features on Price and Profitability.

The dataset that will be used for our analysis contains information on various car models and their specifications, and is titled "Car Features and MSRP". It was collected and made available on Kaggle by Cooper Union, a private college located in New York City.

The dataset contains information on over 11,000 car models and their specifications, including details on the car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP).

The Problem-

- How does the popularity of a car model vary across different market categories?
- What is the relationship between a car's engine power and its price?
- Which car features are most important in determining a car's price?
- How does the average price of a car vary across different manufacturers?
- What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

The client has requested these questions given below:

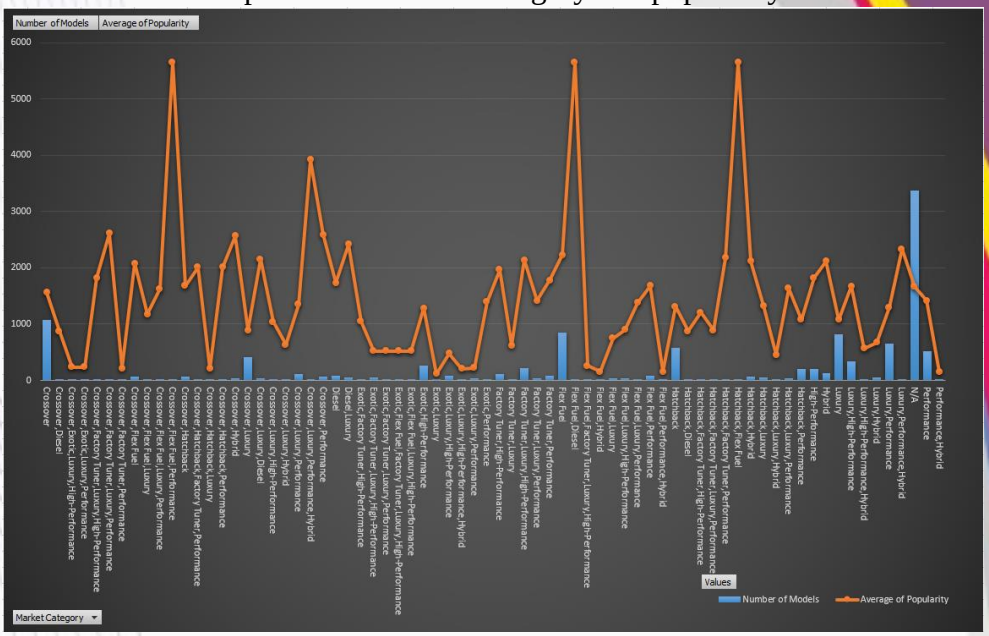
- How does the distribution of car prices vary by brand and body style?
- Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?
- How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?
- How does the fuel efficiency of cars vary across different body styles and model years?
- How does the car's horsepower, MPG, and price vary across different Brands?

Analysis-

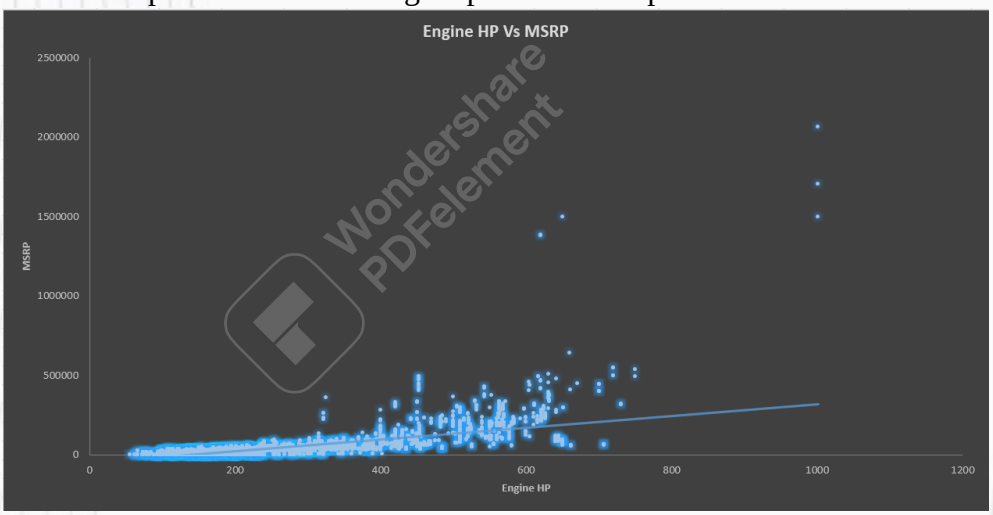
1. Popularity of a car model varies across different market categories

Category	Number of Models	Average of Popularity
Crossover	1075	1556.168372
Crossover,Diesel	7	873
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.461538
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Factory Tuner,Performance	4	210
Crossover,Flex Fuel	64	2073.75
Crossover,Flex Fuel,Luxury	10	1173.2
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover,Flex Fuel,Performance	6	5657
Crossover,Hatchback	72	1675.694444
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Luxury	7	204
Crossover,Hatchback,Performance	6	2009
Crossover,Hybrid	42	2563.380952
Crossover,Luxury	406	889.2142857
Crossover,Luxury,Diesel	34	2149.411765
Crossover,Luxury,High-Performance	9	1037.222222
Crossover,Luxury,Hybrid	24	630.9166667
Crossover,Luxury,Performance	112	1349.089286
Crossover,Luxury,Performance,Hybrid	2	3916
Crossover,Performance	69	2585.956522
Diesel	84	1730.904762
Diesel,Luxury	47	2416.106383
Exotic,Factory Tuner,High-Performance	21	1046.380952
Exotic,Factory Tuner,Luxury,High-Performance	51	523.0196078
Exotic,Factory Tuner,Luxury,Performance	3	520
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performan	13	520
Exotic,Flex Fuel,Luxury,High-Performance	11	520
Exotic,High-Performance	254	1280.047244
Exotic,Luxury	12	112.6666667
Exotic,Luxury,High-Performance	77	473.025974
Exotic,Luxury,High-Performance,Hybrid	1	204
Exotic,Luxury,Performance	36	217.0277778
Exotic,Performance	10	1391
Factory Tuner,High-Performance	104	1966.442308
Factory Tuner,Luxury	2	617
Factory Tuner,Luxury,High-Performance	215	2133.367442
Factory Tuner,Luxury,Performance	31	1413.419355
Factory Tuner,Performance	84	1774.047619
Flex Fuel	855	2225.71345
Flex Fuel,Diesel	16	5657
Flex Fuel,Factory Tuner,Luxury,High-Performance	1	258
Flex Fuel,Hybrid	2	155
Flex Fuel,Luxury	39	746.5384615
Flex Fuel,Luxury,High-Performance	32	898.3125
Flex Fuel,Luxury,Performance	28	1380.071429
Flex Fuel,Performance	87	1680.471264
Flex Fuel,Performance,Hybrid	2	155
Hatchback	574	1308.65331
Hatchback,Diesel	14	873
Hatchback,Factory Tuner,High-Performance	13	1205.153846
Hatchback,Factory Tuner,Luxury,Performance	9	886.8888889
Hatchback,Factory Tuner,Performance	21	2173.714286
Hatchback,Flex Fuel	7	5657
Hatchback,Hybrid	64	2111.15625
Hatchback,Luxury	45	1323.133333
Hatchback,Luxury,Hybrid	3	454
Hatchback,Luxury,Performance	36	1632.25
Hatchback,Performance	198	1073.661616
High-Performance	198	1823.378788
Hybrid	121	2116.586777
Luxury	819	1079.214896
Luxury,High-Performance	334	1668.017964
Luxury,High-Performance,Hybrid	12	568.8333333
Luxury,Hybrid	52	673.6346154
Luxury,Performance	659	1293.062215
Luxury,Performance,Hybrid	11	2333.181818
N/A	3376	1664.832938
Performance	520	1415.209615
Performance,Hybrid	1	155

Relationship between market category and popularity



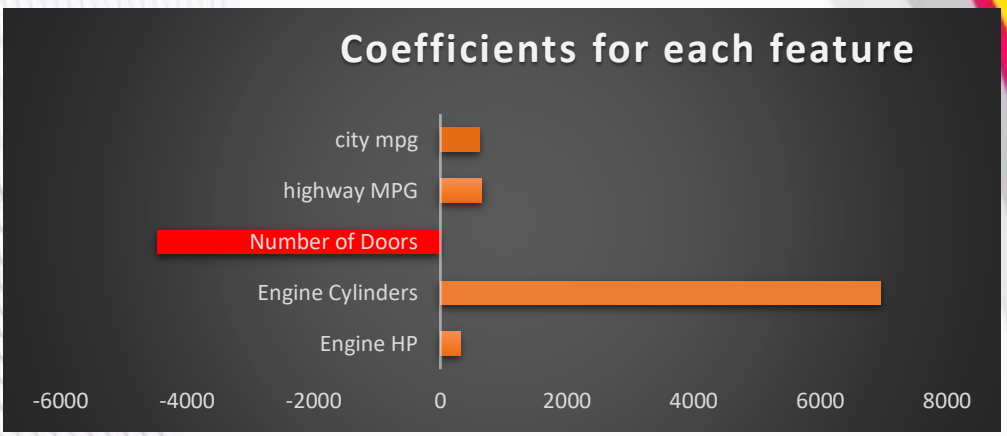
2. Relationship between a car's engine power and its price



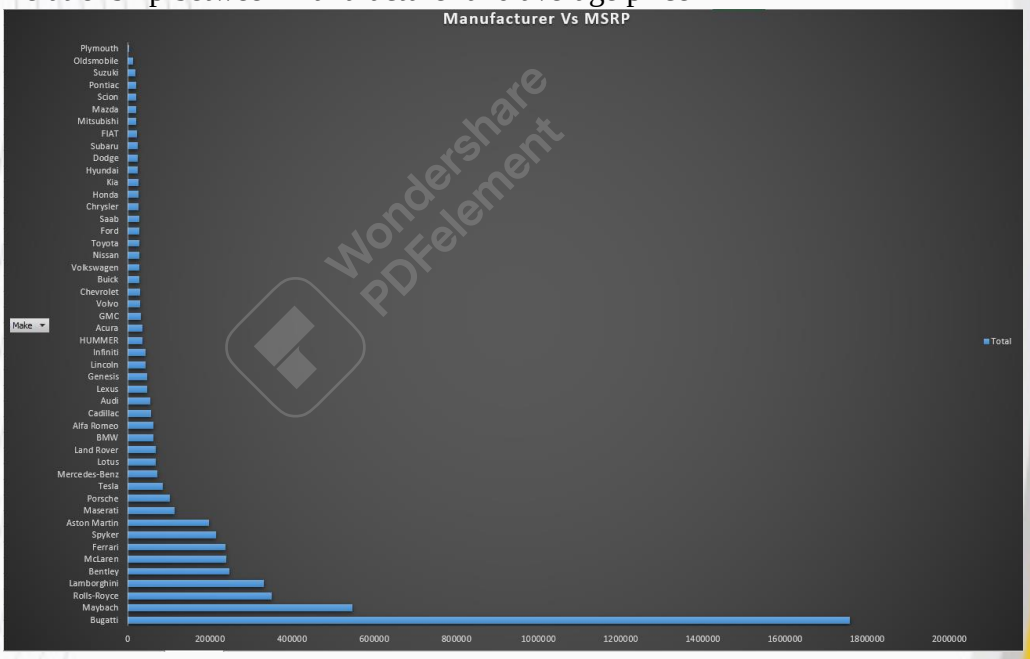
3. car features are most important while determining a car's price

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-91036.00786	3592.216139	-25.34257526	7.9782E-138	-98077.38354	-83994.63218	-98077.38354	-83994.63218
Engine HP	313.919297	6.310889994	49.74247647	0	301.5488422	326.2897518	301.5488422	326.2897518
Engine Cylinders	6963.906079	455.0931785	15.30215439	2.50075E-52	6071.843375	7855.968782	6071.843375	7855.968782
Number of Doors	-4480.869402	496.9421369	-9.016883594	2.24953E-19	-5454.963427	-3506.775377	-5454.963427	-3506.775377
highway MPG	656.1929999	107.2506949	6.118310007	9.77342E-10	445.9627672	866.4232327	445.9627672	866.4232327
city mpg	613.3660972	101.5817676	6.038151447	1.60836E-09	414.2479594	812.484235	414.2479594	812.484235

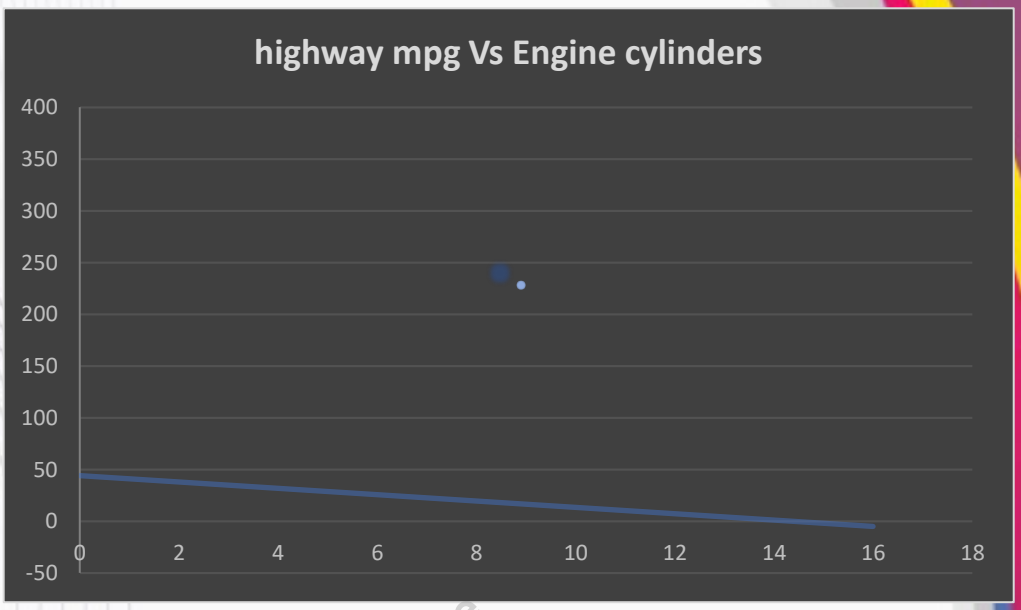
chart showing the coefficient values for each variable to visualize their relative importance



4. Relationship between manufacturer and average price

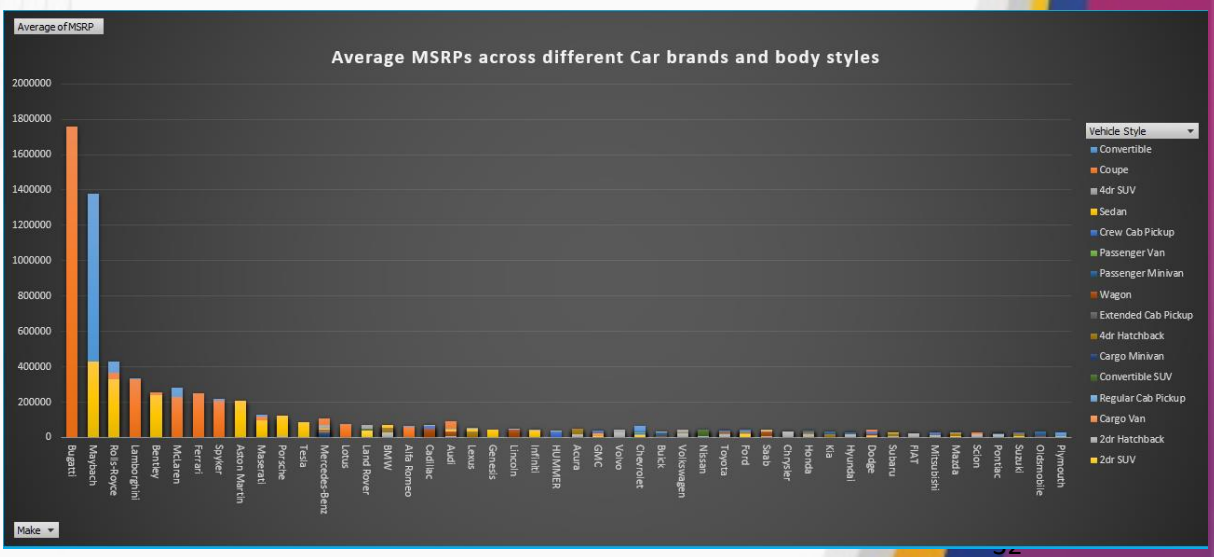


5. relationship between fuel efficiency and the number of cylinders in a car's engine

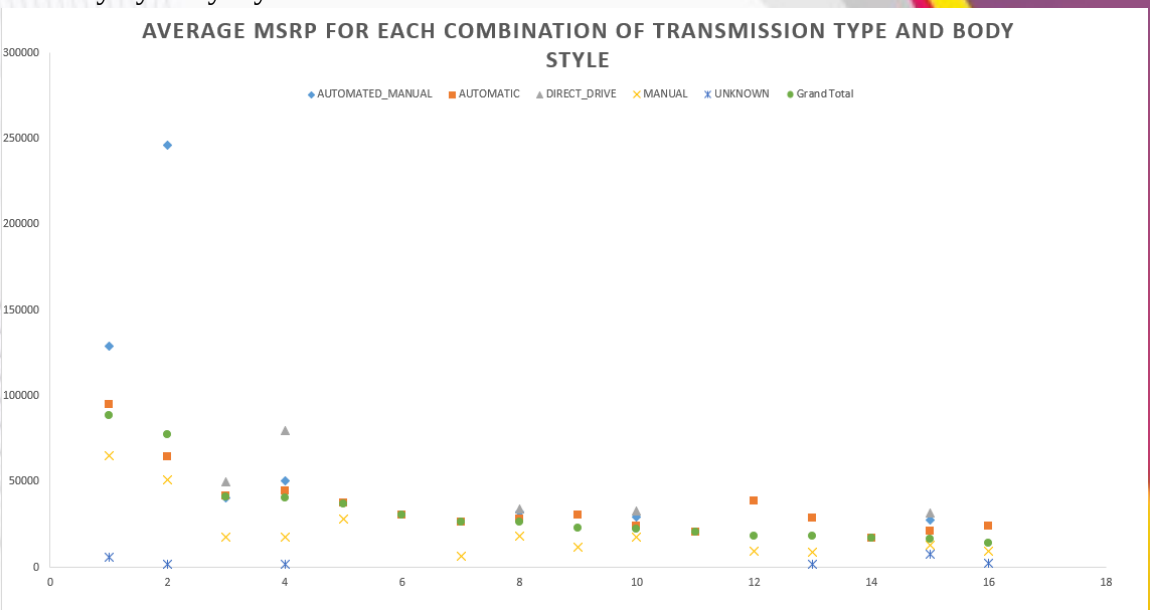


6. Distribution of car prices vary by brand and body style

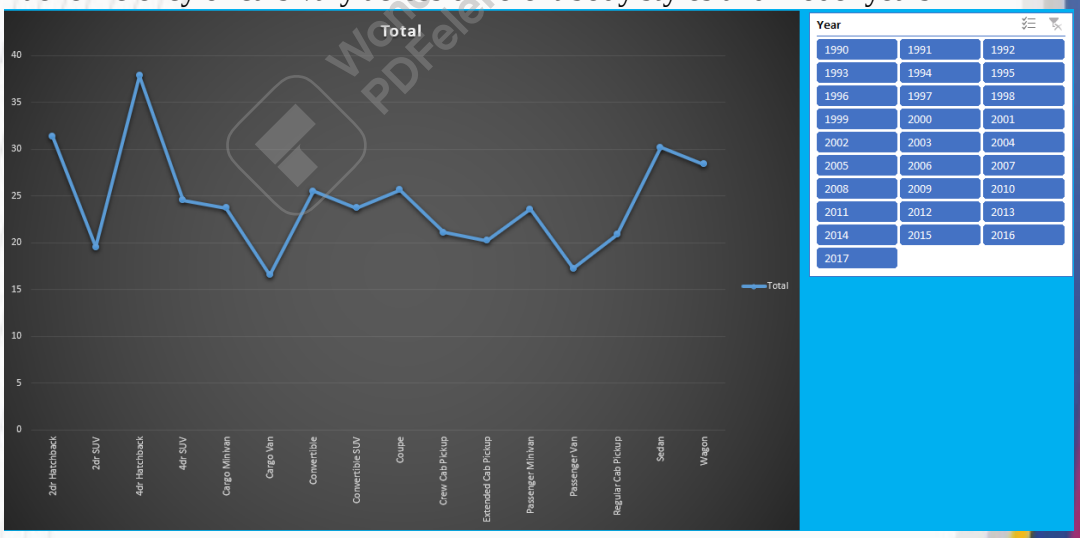
7. highest and lowest average MSRPs, and how does this vary by body style



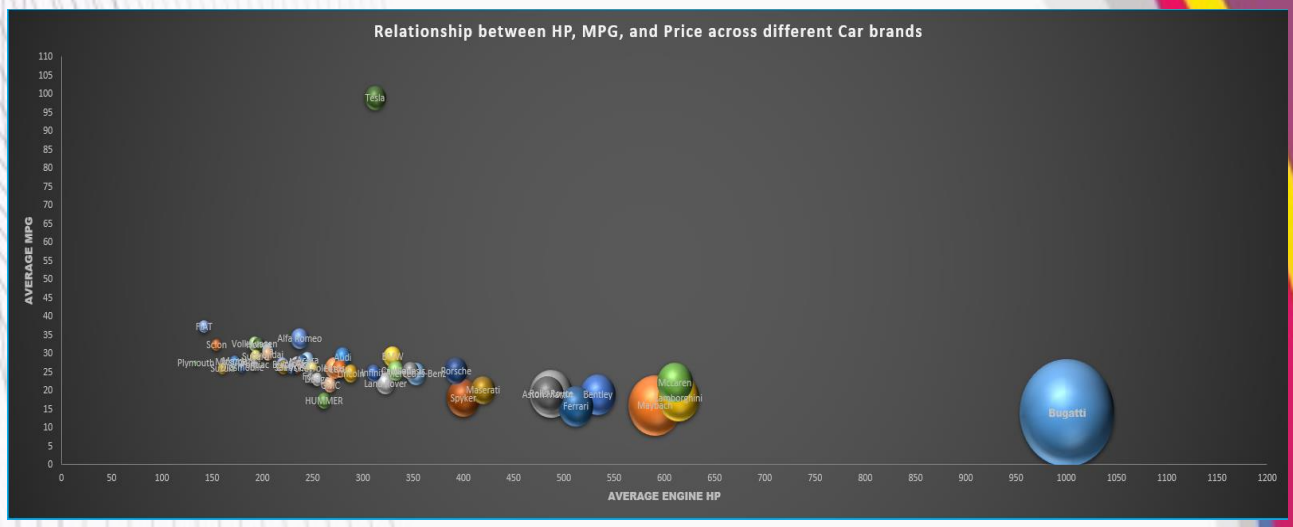
8. Different feature such as transmission type affect the MSRP, and how does this vary by body style



9. Fuel efficiency of cars vary across different body styles and model years



10. Car's horsepower, MPG, and price vary across different Brands



Link to the report-

https://docs.google.com/presentation/d/1SSxkv8t6Wu-SLCuALLKmKm2__Fb6vPZG/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true

ABC Call Volume Trend Analysis

Project Description-

- A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. Typically, these teams fulfil various roles and responsibilities such as: Customer experience programs (CX programs), Digital customer experience, Design and processes, Internal communications, Voice of the customer (VoC), User experiences, Customer experience management, Journey mapping, Nurturing customer interactions, Customer success, Customer support, Handling customer data, Learning about the customer journey.
- In a Customer Experience team, there is a huge employment opportunity for Customer service representatives A.k.a. call center agents, and customer service agents. Some of their roles include Email support, Inbound support, Outbound support, and social media support.
- Inbound customer support is defined as the call center which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of existing customers or prospective customers for our business which are attended by customer care representatives. Inbound customer service is the methodology of attracting, engaging, and delighting our customers to turn them into our business' loyal advocates. By solving our customers' problems and helping them achieve success using our product or service, we can delight our customers and turn them into a growth engine for our business.

The Problem-

a. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).
b. 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, ….)
c. 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.)
d. 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

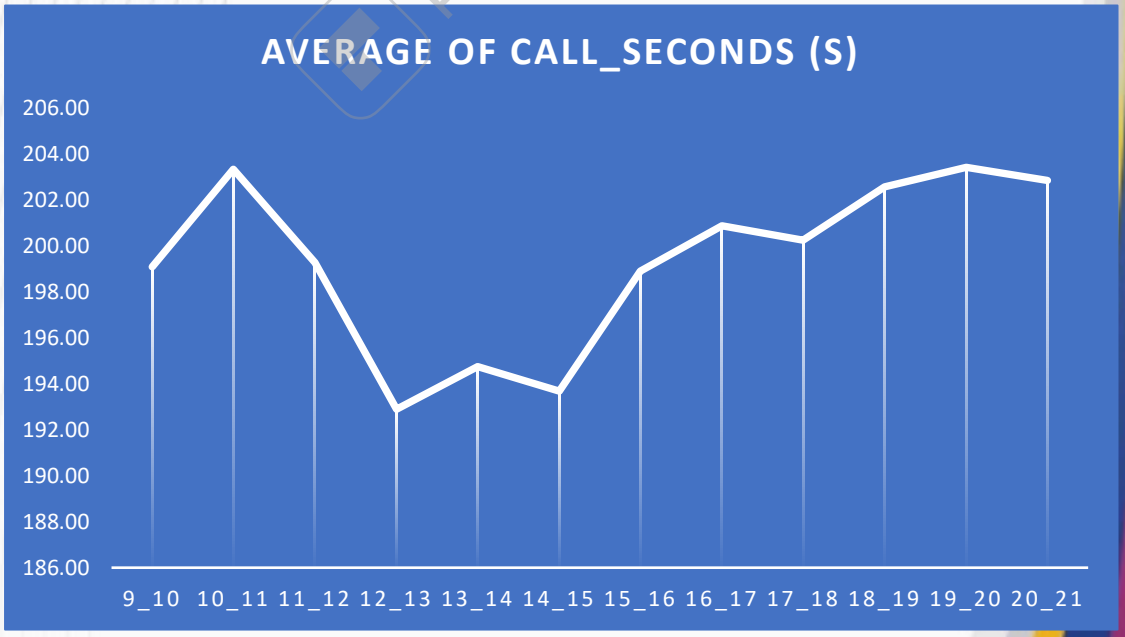
Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

Assumption: An agent work for 6 days a week; On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into lunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e 60% of 7.5 Hrs) on call with customers/ users. Total days in a month is 30 days.

Analysis-

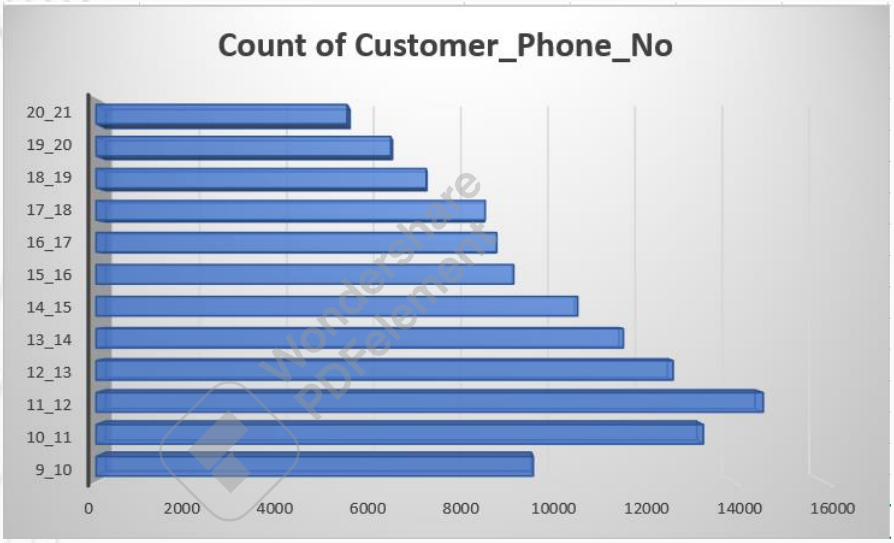
1. Average time duration for all incoming calls received by agents in each Time bucket

Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
10_11	203.33
11_12	199.26
12_13	192.89
13_14	194.74
14_15	193.68
15_16	198.89
16_17	200.87
17_18	200.25
18_19	202.55
19_20	203.41
20_21	202.85
9_10	199.07
Grand Total	**198.62**



2. The total volume/ number of calls coming in via charts/ graphs

Time-Bucket	Count of Customer_Phone_No	Percentage of Calls per Bucket
9_10	9588	8.13%
10_11	13313	11.28%
11_12	14626	12.40%
12_13	12652	10.72%
13_14	11561	9.80%
14_15	10561	8.95%
15_16	9159	7.76%
16_17	8788	7.45%
17_18	8534	7.23%
18_19	7238	6.13%
19_20	6463	5.48%
20_21	5505	4.67%
Grand Total	**117988**	**100.00%**



3. Manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.

Time-Bucket	Count of Customer_Phone_No	Percentage of Calls per Bucket	TotalHours needed to answer the % calls per bucket	Agents required
9_10	9588	8.13%	20.72	5
10_11	13313	11.28%	28.77	6
11_12	14626	12.40%	31.61	7
12_13	12652	10.72%	27.34	6
13_14	11561	9.80%	24.99	6
14_15	10561	8.95%	22.82	5
15_16	9159	7.76%	19.79	4
16_17	8788	7.45%	18.99	4
17_18	8534	7.23%	18.44	4
18_19	7238	6.13%	15.64	3
19_20	6463	5.48%	13.97	3
20_21	5505	4.67%	11.90	3
			255	**57**

4. manpower plan for the night time, that is, from 9 Pm to 9 Am, Abandon rate assumption would be same 10%

Time Bucket	Calls Received	Percentage of calls	Actual No. of Calls per bucket	Total hours required	Agents required per bucket
21_22	3	10.00%	154	8	2
22_23	3	10.00%	154	8	2
23_24	2	6.67%	103	5	1
00_01	2	6.67%	103	5	1
01_02	1	3.33%	51	3	1
02_03	1	3.33%	51	3	1
03_04	1	3.33%	51	3	1
04_05	1	3.33%	51	3	1
05_06	3	10.00%	154	8	2
06_07	4	13.33%	205	10	2
07_08	4	13.33%	205	10	2
08_09	5	16.67%	257	13	3
Total	30	100.00%	1539	76.42	17

Link to the report-

https://docs.google.com/presentation/d/119uPjnzMhqmyO5f5q8wth1BTUsS6sY_y/edit?usp=sharing&ouid=109923652620940529642&rtpof=true&sd=true