Data Analysis

Portfolio



Presented by:

Ruchita Parma

Table Contents

Table of Contents ----------------------------- 2

Data Analysis Process---------------------

Description------------------------------------- 7

Project step ---------------------------------------7-8

Instagram user analytics -----------

Description------------------------------------9

Design -------------------------------------------- 10

Findings ---------------------------------------11-12

Data Analysis -----------------------------------13

Conclusion --------------------------------------14

Operation Analytics and Investigating Metric Spike--------------------------------------------------

Description-------------------------------------15

Design -------------------------------------------- 16

Findings ---------------------------------------17-24

Conclusion --------------------------------------25

Hiring Process Analytics-----------

Descriotion----------------------------------26

Design -------------------------------------------- 27

Findings ---------------------------------------28-32

Data Analysis -----------------------------------33

Conclusion --------------------------------------34

IMDB Movie Analysis--------------------

Description-------------------------------35

Design -----------------------------------------36

Findings ---------------------------------------37-40

Data Analysis -----------------------------------41

Conclusion --------------------------------------42

Bank Loan Case Study-----------

Description------------------------------------43

Design -------------------------------------------- 44

Missing Values Analysis----------------------------45

Data Imbalance----------------------------------46

Outlier--------------------------------------47-48

Finding----------------------------------49-51

Analysis ----------------------------------------52

Conclusion----------------------------------53

Impact of Car Features---------------------

Description----------------------------------54-55

Design----------------------------------------56

Problem-------------------------------------57

Findings ---------------------------------------58-68

Conclusion --------------------------------------69

ABC Call Volume Trend ---------------------

Description----------------------------------------

Design --------------------------------------------

Findings ---------------------------------------

Data Analysis -----------------------------------

Conclusion --------------------------------------



Data analysis process



Description

We use Data Analytics in everyday life without even knowing it. For e.g. Going to a market to purchase car.

**Plan:** I first decide my budget than which things I need before going to purchase. Its new car or old car

**Prepare:** Next I need to check how much I am willing to spend and how to get that money.

**Process:** Then I need to check how much I want from the data. Like if I am going to buy new car what do I want – Tata motors /Volvo /Honda/ Kia/ Hyundai Motors etc.

**Analysis:** I obviously won't buy car which are old year manufacture, Also you need to check does the car have all future like company Name, manufacture Year, Mileage, fuel type, Engine, Seats, Model, Power, Price etc. That I want and, will it is in my budget. Also take a test drive

**Share:** Now i communicate my idea to the dealer to find the best suitable fit for me.

**Act:** Then i finally buy it!



Instagram user analytics

Description

User analysis is the process by which we track how users engage and interact with our digital Product (software or mobile application) in an attempt to derive business insights for marketing, Product & development teams. These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.

Design

Get the information from given description about dataset and understand the problem.

Go through the row data and understand the Attribute, variable that are given.

I have used My SQL to extract the required data from the given database using the Join function, sub-queries, Aggregation, where condition, Group by, Distinct and other functions required.

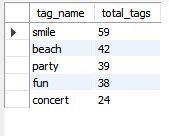
Execute the queries and if there are any errors occurred then modified it without any errors.

Done the assessment like that.

FINDING 1

Identify and suggest the top 5 most commonly used hash tags on the platform

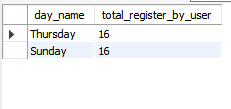
select tags.tag\_name, count(\*) as total\_tags from photo\_tags join tags on photo\_tags.tag\_id=tags.id group by tag\_id order by total\_tags desc limit 5;



FINDING 2

What day of the week do most users register on? Provide insights on when to schedule an ad campaign.

SELECT DAYNAME(created\_at) AS day\_name,count(\*) as total\_register\_by\_user FROM users GROUP BY day\_name ORDER BY total\_register\_by\_user DESC LIMIT 2;



FINDING 3

Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users

SELECT (SELECT COUNT(\*) FROM photos) / (SELECT COUNT(\*) FROM users) AS Avg ;



ANALYSIS

In this analysis I find out who have been using the platform for the longest time. How to determine user who gets the most likes on a single photo. I also analysis

Witch week do most users register on? Provide insights on when to schedule an ad campaign

CONCLUSION

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to create the task using that data. I infer about this project is the dataset provided was very limited and small in respect of Rows and columns, But still, it was a very good experience working on such kind of project. this project help me how to work real time data.

Operation & Metric Analytics



Description

Analytics is the analysis done for the complete end to end operations of an organization. with the help of this, the company then finds the areas on which it must improve upon. This kind of analysis is further used to predict the overall growth.

Design

Get the information from given description about dataset and understand the problem.

Go through the row data and understand the Attribute, variable that are given.

I have used My SQL to extract the required data from the given database using the Join function, sub-queries, Aggregation, where condition, Group by, Distinct and other functions required.

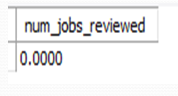
Execute the queries and if there are any errors occurred then modified it without any errors.

Done the assessment like that

Finding

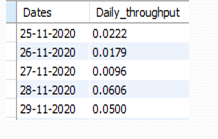
1. Calculate the number of jobs reviewed per hour per day for November 2020?

select count(distinct job\_id)/(30\*24) as num\_jobs\_reviewed from job\_data where ds between '2020-11-01' and '2020-11-30';



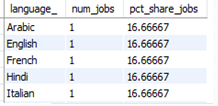
1. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?

select ds as Dates, count(event\_)/sum(time\_spent) as Daily\_throughput from job\_data group by ds order by ds;



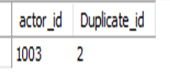
1. Calculate the percentage share of each language in the last 30 days?

select language\_, num\_jobs,100.0\* num\_jobs/total\_jobs as pct\_share\_jobs from(select language\_, count(distinct job\_id) as num\_jobsfrom job\_datagroup by language\_)a cross join(select count(distinct job\_id) as total\_jobs from job\_data)b;



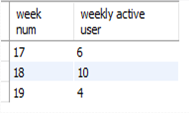
1. Let’s say you see some duplicate rows in the data. How will you display duplicates from the table?

SELECT actor\_id, COUNT(actor\_id) as Duplicate\_id FROM job\_data GROUP BY actor\_id HAVING COUNT(actor\_id) > 1



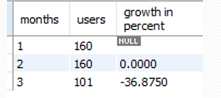
1. Calculate the weekly user engagement?

select extract(week from occurred\_at) as "week num",count(distinct user\_id) as "weekly active user" from events where event\_type="engagement" group by 1;



1. Calculate the user growth for product?

select months,users,((users/lag(Users) over (order by Months)-1)\*100) as "growth in percent“ from(select extract(MONTH from created\_at) as Months,count(activated\_at) as Usersfrom users where activated\_at not in ("")group by 1order by 1)a;

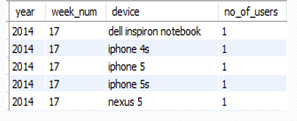


1. Calculate the weekly retention of users-sign up cohort?

select count(user\_id), sum(case when retention\_week = 1 then 1 else 0 end) as per\_week\_retention from(select a.user\_id, a.sign\_up\_week, b.engagement\_week, b.engagement\_week - a.sign\_up\_week as retention\_weekfrom((select distinct user\_id, extract(week from occurred\_at) as sign\_up\_weekfrom eventswhere event\_type = 'signup\_flow'and event\_name = 'complete\_signup'and extract(week from occurred\_at)=18)aleft join(select distinct user\_id, extract(week from occurred\_at) as engagement\_weekfrom eventswhere event\_type = 'engagement')bon a.user\_id = b.user\_id)group by user\_idorder by user\_id;

1. Calculate the weekly engagement per device?

Select extract(year from occurred\_at) as year,extract(week from occurred\_at) as week\_num,device,count(distinct user\_id) as no\_of\_users from events where event\_type = 'engagement'group by 1,2,3 order by 1,2,3;



1. Calculate the email engagement metrics?

select100.0 \* sum(case when email\_cat = 'email\_opened' then 1 else 0 end) /sum(case when email\_cat = 'email\_sent' then 1 else 0 end)as email\_opening\_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\_clicking\_ratefrom(select \*,case when action in ('sent\_weekly\_digest', 'sent\_reengagement\_email') then 'email\_sent' when action in ('email\_open') then 'email\_opened' when action in ('email\_clickthrough') then 'email\_clicked'end as email\_catfrom email\_events)a;



Conclusion

In this project, I learned how to apply advanced SQL concepts like Windows Functions, etc. I understood how the real-world industry works.

It helped me in improve my SQL concepts.

The challenge here is that the data in case study 2 is very huge for that I load the data but the main thing I learn a lot, how to work on large dataset.

HIRING PROCESS ANALYSIS



Description

Provided dataset of the company. Where the details about people who registered for a particular post in a department of this company as an analyst I have use my knowledge in statistics and use different formulas in excel and draw necessary conclusions about the company.

Attributes: application\_id, Interview Taken,

Status event\_name, Department,

Post Name, Offered Salary

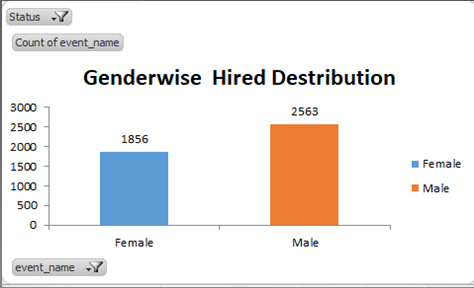
Design

Firstly i upload the dataset and try to understand the data and its pattern

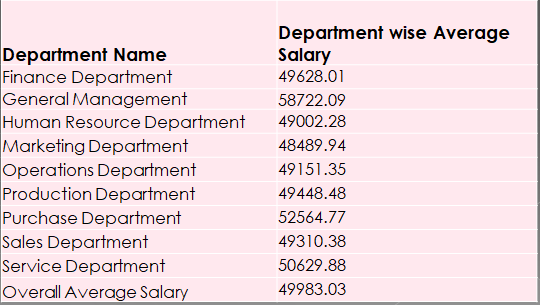
I performing analysis on Microsoft Excel using various in-built formula of Excel and we will go step by step in finding out the answers to the questions asked related to the data set. Create table use MIN, MAX, AVG, COUNTIF and other statistic formulas. Draw the given task and save my answer sheet

Finding

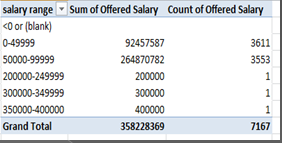
1. . How many males and females are hired?

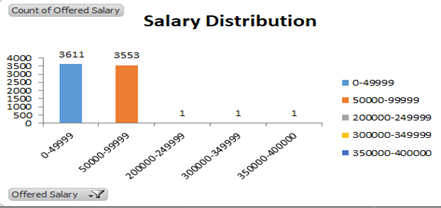


1. .What is the average salary offered in this company?

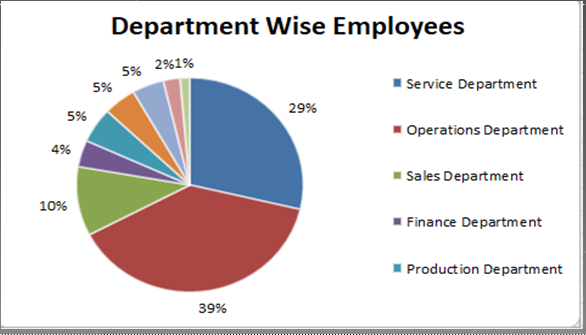


1. .Draw the class intervals for salary in the company?

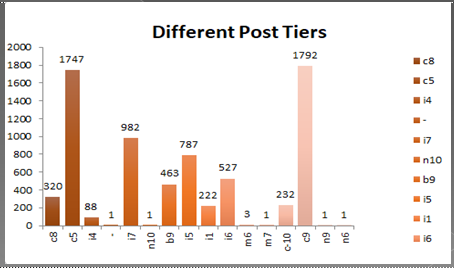




1. .Draw Pie Chart / Bar Graph ( or any other graph ) to show proportion of people working different department ?



1. .Represent different post tiers using chart/graph?

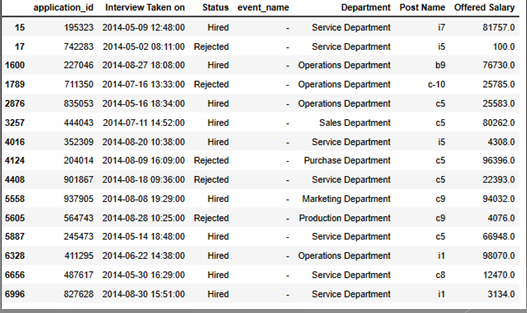


EDA

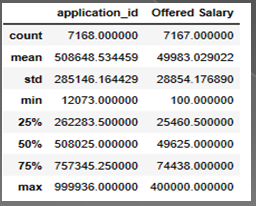
In the EDA I analysis the null values, find the outlier, Remove and correlation.



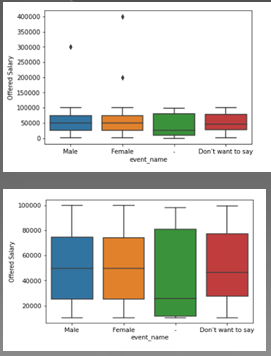
Null values Analysis.



Statistical analysis and information about data



Outliers find and Remove.



Analysis

Males are 54.57% of all the employees,39.51% are Females and 5.92% don’t want to

say

• General Management Department has the highest average salary of 60,810.20 and

Marketing the lowest of 47,843.40

• Most of the employees are in the salary class interval of 40,001-50,000

• Operations Department has the most employees total of 1843 and Human Resources

the lowest of 70 Employees

• Most of the employees are in the post tier of c9 a total of 1239 Employee

Males are 54.57% of all the employees,39.51% are Females and 5.92% don’t want to

say

• General Management Department has the highest average salary of 60,810.20 and

Marketing the lowest of 47,843.40

• Most of the employees are in the salary class interval of 40,001-50,000

• Operations Department has the most employees total of 1843 and Human Resources

the lowest of 70 Employees

• Most of the employees are in the post tier of c9 a total of 1239 Employee

Males are 54.57% of all the employees,39.51% are Females and 5.92% don’t want to

say

• General Management Department has the highest average salary of 60,810.20 and

Marketing the lowest of 47,843.40

• Most of the employees are in the salary class interval of 40,001-50,000

• Operations Department has the most employees total of 1843 and Human Resources

the lowest of 70 Employees

• Most of the employees are in the post tier of c9 a total of 1239 Employee

Males are 54.57% of all the employees,39.51% are Females and 5.92% don’t want to

say

• General Management Department has the highest average salary of 60,810.20 and

Marketing the lowest of 47,843.40

• Most of the employees are in the salary class interval of 40,001-50,000

• Operations Department has the most employees total of 1843 and Human Resources

the lowest of 70 Employees

• Most of the employees are in the post tier of c9 a total of 1239 Employees

Males are 54.57% of all the employees,39.51% are Females and 5.92% don’t want to say

General Management Department has the highest average salary of 60,810.20 and

Marketing the lowest of 47,843.40

Most of the employees are in the salary class interval of 40,001-50,000

Operations Department has the most employees total of 1843 and

Human Resources the lowest of 70 Employees

Conclusion

This project has helped me improve my excel skill.

This project has leaded me to understand the analysis required in the company’s hiring process.

This project has understood the concept of exploratory data analysis.

##### IMDB Movie Analysis



Description

This project aims to carry out the in depth analysis of IMBD movie data set this project is focused on carry out analysis which will help bring insights relevant to IMDB. The project will answer to the important questions like

Clean data.

Movies with highest profit.

Top movies with best IMDB rating score.

Audience favorite actors.

Critic favorite actors, User rating over decades.

DESIGN

Import the dataset in excels.

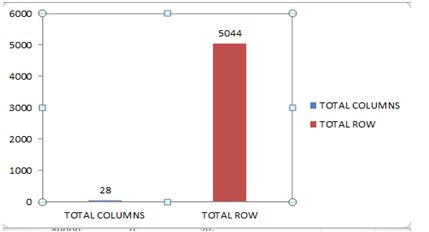
Clean the data like remove null values and duplicate.

Use pivot, sum, average count if and filter and give the answer of asked quotations.

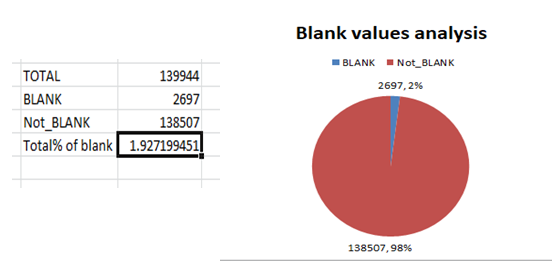
Create charts and bar

FINDINGS

**Before clean the data we have 28 columns and 5044 rows**



**Null values and not null values after remove the duplicate values.**



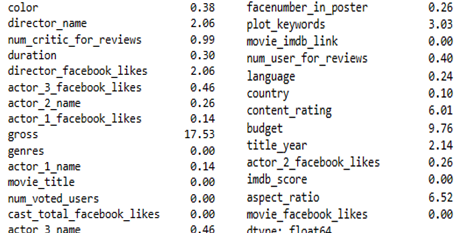
**Fill and remove null values**

Fill the null value by “mean” for numerical columns

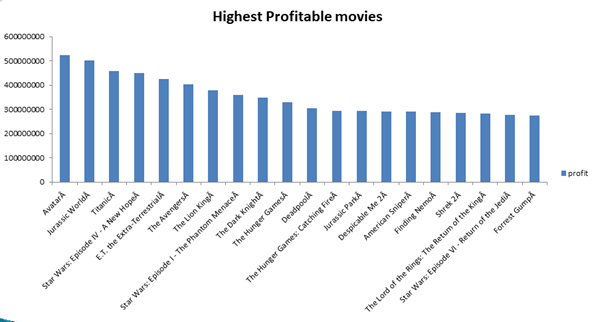
Fill the null values by “most frequently” values in categorical columns

Some null values drop witch is less than 2%(show below the list of null values in percent)

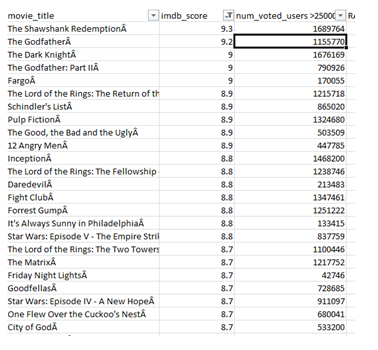
**Null values in columns % wise**



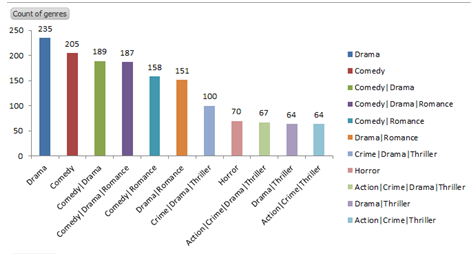
1. **Find the movies with the highest profit?**



**Top 250 MOVIES**



**Find popular genres**



**Most popular genres is “DRAMA”**

**ANALYSIS**

Individually inspecting budget and profit, both are slightly skewed to right, which means most of the movies have positive profit and also budgets are high

We can observe that there are a couple of movies that have very high budgets but still having negative profits

We can see a slightly positive trend between ‘budget’ and ‘profit’ i.e., as budget increases, profit also increases.

CONCLUSION

Analysis of the movie dataset shows that majority of the movies have runtime between 90 and 120 minutes. The project successfully brings out the required insights. All the questions were answers by carry out analysis using python and excel. The insights drawn were important and answer the required questions. Top genres are drama, action, comedy, adventure and thriller. Although we couldn't establish acceptable correlation between rating and revenue, we were able to establish moderate positive correlation between rating and vote.

Bank Loan Analysis



DESCRIPTION

This project aims at analyzing the risk appetite of banks. When the company receives a loan application, the company must decide for loan approval based on the applicant’s profile**.**

The data given contains the information about the loan application at the time of applying for the loan.

Design

Firstly i upload the dataset and try to understand the data and its pattern

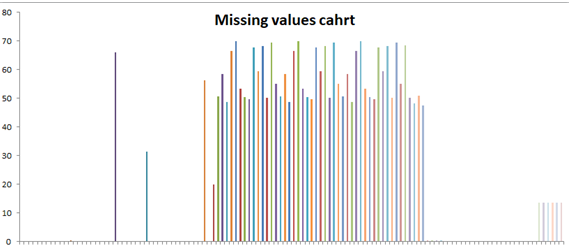
I performing analysis on Microsoft Excel using various in-built formula of Excel and we will go step by step in finding out the answers to

The questions asked related to the data set.

Create table use MIN, MAX, AVG, COUNTIF and other statistic formulas.

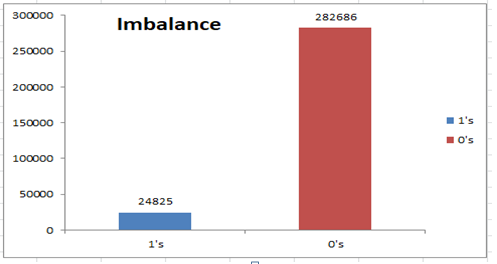
Draw the given task and save my answer sheet

Identify the missing data columns wise missing values in %



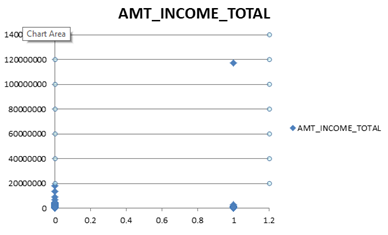
* In above graph I find the null values % columns wise.
* I can drop the columns that have 40% and above null values
* I can fill the null value by MEAN, MEDIAN, MODE.

Data imbalance



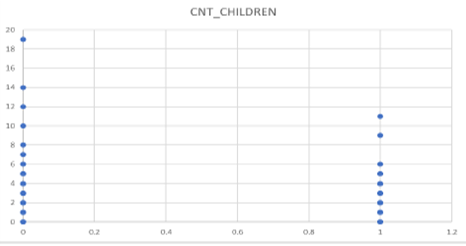
Shows the ratio of total applicants with payment difficulties (1) to the total applicants with installments being paid on time (0)

Identify if there are **outliers** in the dataset

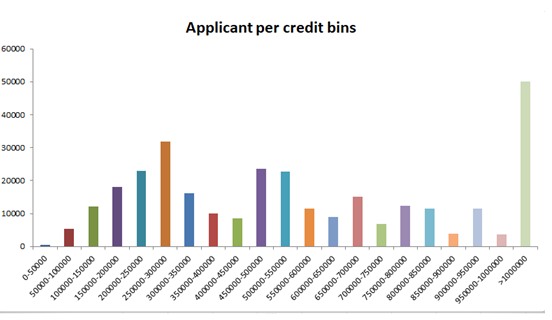


In the income variable majority of income in lac but few income in CORER so this corer outlier

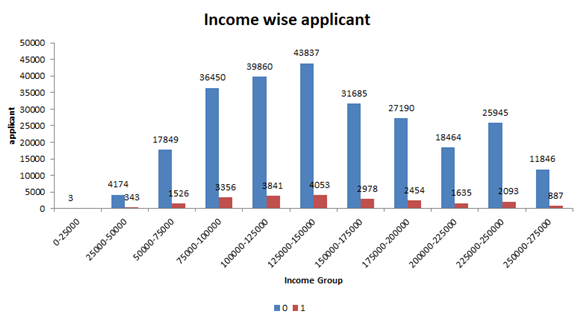
In CNT\_CHILDREN columns there are outliers for the target column 0 and as well as 1. The XY Plotter for 0 shows 19 children which is highly unusual these days. The XY plotter for 1 shows more than 7 children.



In univariate analysis AMT\_CREDIT grouped in different credit bins. Majority of the applicants were offered loans in the credit range of 10 Lac and Above.

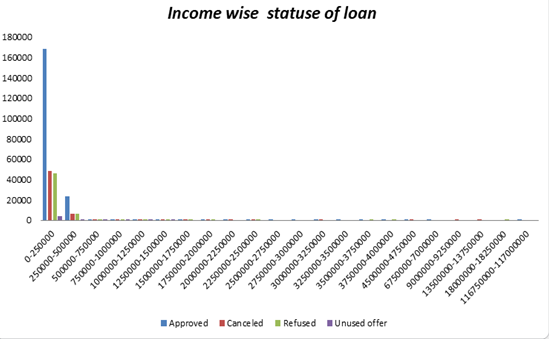


SEGMENTED UNIVARIAITE ANALYSIS



* The above graph is an example of univariate segmented analysis which depicts simply the count of segmented applicants (0 & 1) for the variable AMT\_TOTAL\_INCOME grouped in different income bins.
* Also, maximum applicants (0,1) draw an income between 1.25 Lacs to 1.5 Lacs but there are applicants which are having payment difficulties despite belonging to the same income range.

BIVARIAITE ANALYSIS



In the above graph 0-250000 have more approve more cancelled and more refused and unused offers from other income band.

Analysis

Applicants drawing higher income were offered higher loan amount by the bank.

Majority of applicants drawn an income range between 1.25 Lacs – 1.5 Lacs, also the defaults drawn income between the same range. Majority of applicants were offered loans in the credit range of 9 Lacs and above.

Conclusion

This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain.

Analysing the Impact of Car Features on Price and Profitability



Description

The automobile industry is one of the most important drivers of economic growth of a country.

The growth of this sector will support & helped to carve a unique path among the manufacturing sectors.

The automobiles produced in the country uniquely cater to the demands of low- and middle-income groups of population.

This chapter analyzes the roles of Brands, car models, Engine quality, type, cylinders & its power, Market category and other enabling factors in the expansion of the automobile and automotive components.

To meet the future needs of customers including the electrical vehicles and stay ahead of competition, manufacturers are now catching up on up gradation, digitization, and automation.

The chapter also analyzes Company's in light of these developments.

Design

Firstly i upload the dataset and try to understand the data and its pattern

I performing analysis on Microsoft Excel using various in-built formula of Excel and we will go step by step in finding out the answers to

The questions asked related to the data set.

Create table use pivot, MIN, MAX, AVG, COUNTIF and other statistic formulas.

Draw the given task and save my answer sheet

Problem

In talk in basic business problem, Company need to analyze customer and market demands so as they meet the expectations for its growth.

Below were the main questions and scenario that we need to handle and implement its solutions.

1. Affordability

2. Comfort

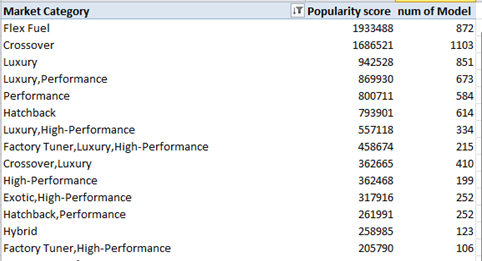
3. Efficiency

4. Market trends for customer choice for Engine and its transmission

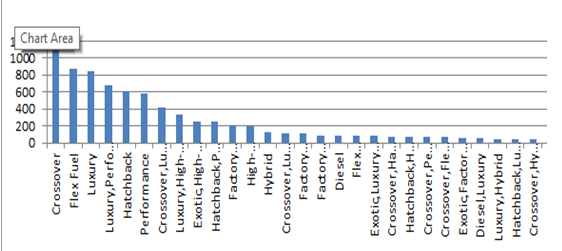
5. Technology (Automatic or Manual)

Finding 1

1. Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores

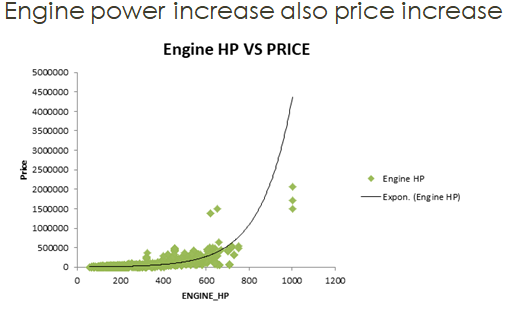


**B**. Relationship between market category and popularity



Finding 2

**Task 2:**  Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trend line to the chart to visualize the relationship between these variable

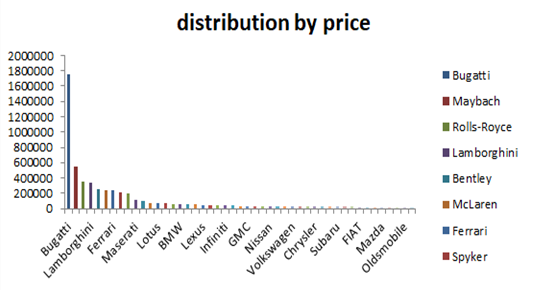


Finding

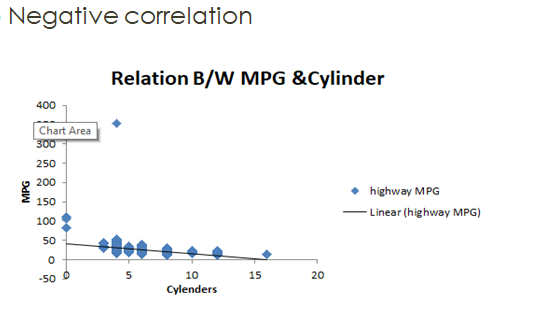
**Task 4.A:** Create a pivot table that shows the average price of cars for each manufacturer



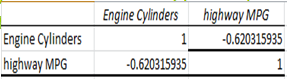
**Task 4.B:** Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.

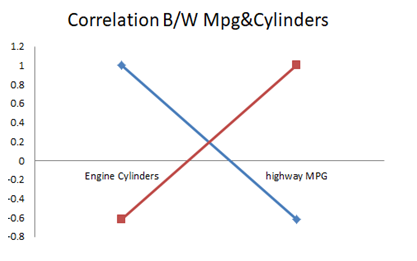


**Task 5.A:** Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.

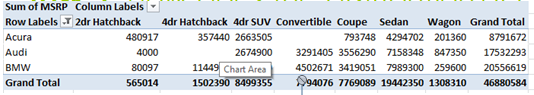


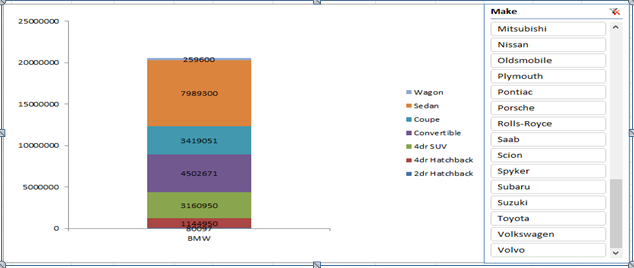
**Task 5.B:** Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.



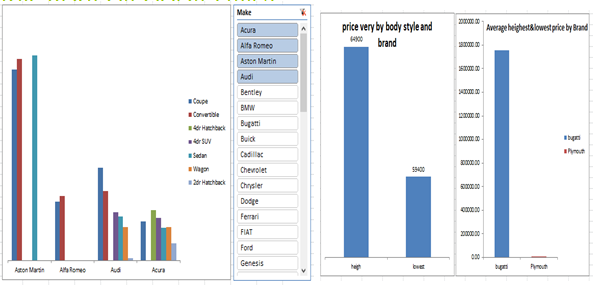


**Task 1:** How does the distribution of car prices vary by brand and body style?

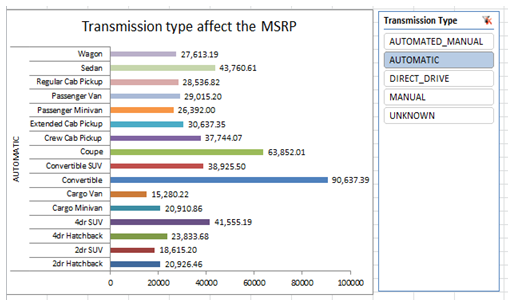




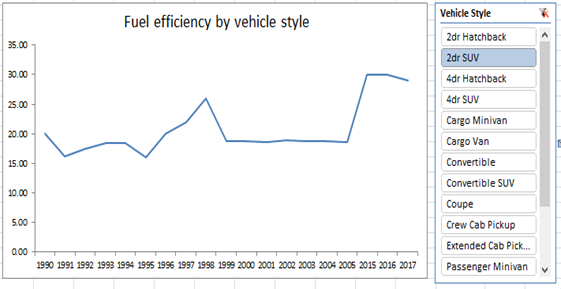
**Task 2:** Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?



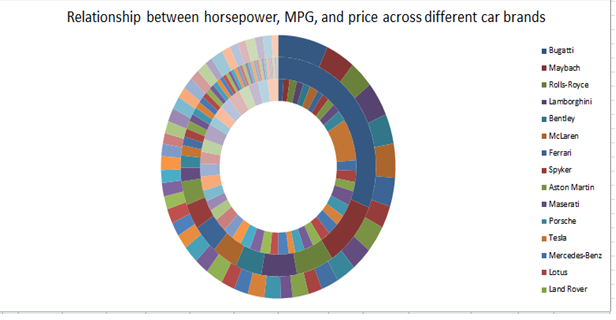
**Task 3:** How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?



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?



Conclusion

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is pre-processing of the data. In this research clean data to avoid unnecessary noise for the data set. Data cleaning is one of the processes that are important for analyse the data and create a use full visualization.

ABC Call Volume Trend

Description

I have a dataset of a Customer Experience (CX) Inbound calling team for 23 days. Dataset contain multipl columns likCustomer\_Phone\_No, Date\_&\_Time, Time, Time\_Bucket,Duration(hh:mm:ss),Call\_Seconds(s),Call\_Status,Wrapped

In the given project we have to calculate the average call duration and total number of call per hour

We had to propose a manpower plan required during each time bucket and reduce the abandon rate to 10%

We also have to propose a manpower plan when calls are received across the 24 hours

Design

Import the dataset and understand the dataset.

Data cleaning.

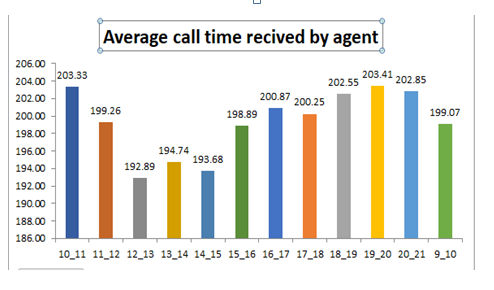
Understanding the required insight.

Create a pivot tables or using formulas and functions to find the required insight.

Drown Charts for better Visualization.

Finding

1. Calculate the average call time duration for all incoming calls received by agents(in eachTime\_Bucket)

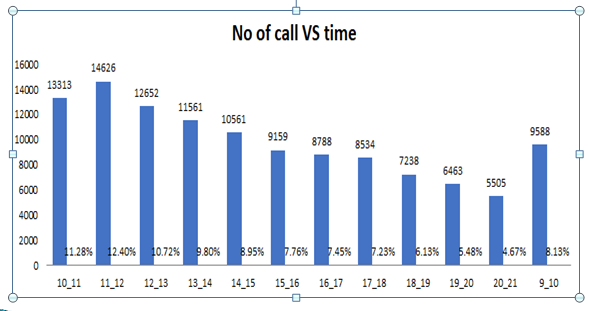


In the above graph we can easily analyse call answered by time\_bucket.

Total average call time is 198.62 seconds for all incoming call received by agent.

Highest received call by agent between 7pm - 8pm and followed by 10am-11am, 6pm-7pm.

1. 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, …..)



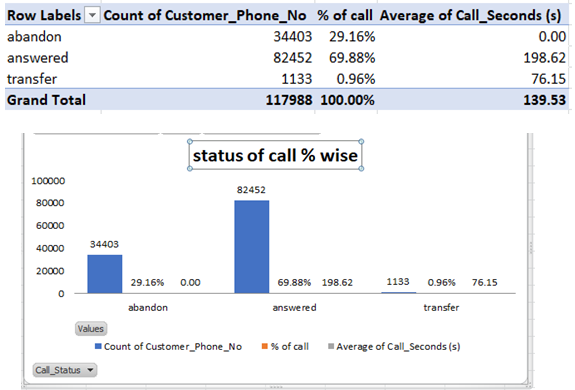
Create a pie vote table put bucket\_time in rows and customer phone no in value column and count the customer phone no.

Maximum customer call in between 11am to12pm

Lowest call in 8pm-9pm

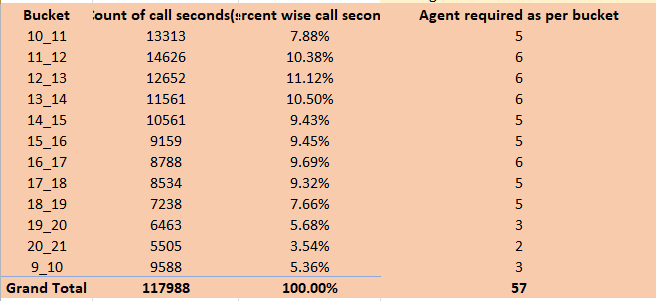
Drown column bar for better visualisation.

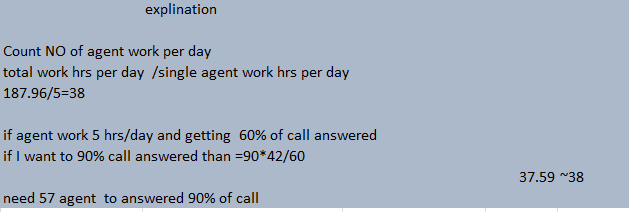
1. As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. You have to calculate minimum number of agents required in each time bucket so that at least 90 calls should be answered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

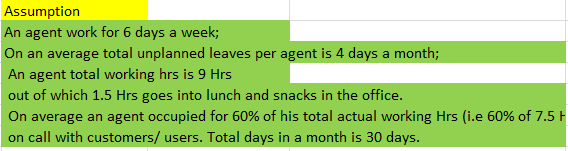












Conclusion

I have learned how to analyse the call centre data and report to the management. I have also learned to use pivot table and pivot charts more effectively