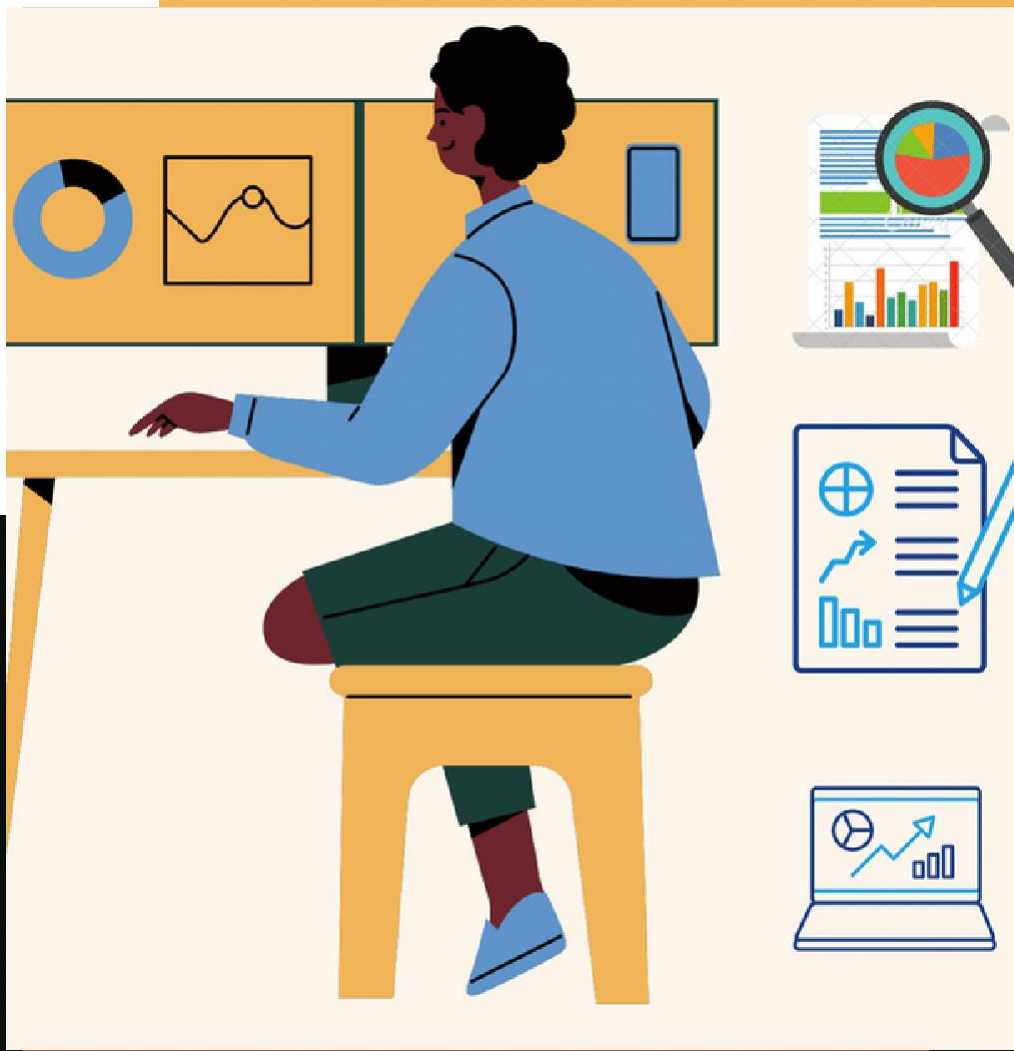
DATA ANALYTICS

PORTFOLIO PROJECT

PREPARED BY APARAJITA SINGH



CONTACT:

LinkedIn

Email Id: aparajitasingh7068@gmail.com

PROFESSIONAL BACKGROUND

EDUCATION

- I am pursuing my bachelors of technology degree in Electronics and Communication Engineering from Feroze Gandhi Institute Of Engineering and Technology Raebareli in with 8.05 SGPA(till now). I will be graduating in 2024 .
- I have done Data Analytics Trainee certificate course from Trainity.

HARD SKILLS

- Structured Query Language
- (SQL) Advance Microsoft Excel
- Data Visualization (Tableau)
- Statistics

SOFT SKILLS

Storytelling
Communication Skills
Critical Thinking
Teamwork
Research
Presentation and Report

OTHER SKILLS

- Predictive Analytics
- Risk Analytics
- Behavioral Analytics

ABOUT ME

Thorough and meticulous Data Analyst passionate about helping businesses succeed. Looking forward to work in an esteemed company where skills, knowledge, talents matter more than the degree. Through development of portfolio projects and internship, I learned the importance of having an iterative, hypothesis oriented approach to analysis. I am eager to leverage that approach at your company as a data analyst.

Certifications

Data Analytics Virtual Internship:
InternshipCertificate.pdf

Data Analytics Live Training:
TrainingCertificate.pdf

Mini certificate:
MiniCertificate.pdf

TABLE OF CONTENTS

Data Analytics Process

Description: We use Data Analytics in everyday life.
For example: Searching in Web Searching Engines like Google.

6 Steps Process:

Plan:
First we decide to buy laptop from the Amazon. For buying cell laptop we register our number on Amazon and make a customer profile by using our phone number and email id. Then we go and search for laptop.

Prepare:
Next we need to check which website will give us the correct information in the simplest way.

Process:
Then we need to check how much we want from the data. Like if we want – difference between the different price ranges is the best choice according to the different models / etc.

Analyze:
We then analyze the different trendy models and generations of laptops according to our need.

Share:
Now we search it in the Google. And Google gives us the best results at the top-most by using the process of data analytics / data science.

Act:
Then we finally click it and get the necessary information which we need. We, then use that information in our project.

Project Link:- DATA ANALYTICS PROCESS (1).pdf

Instagram User Analytics

Description: This project is about how the users engage and interact with Instagram. We will analyse these users 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.

Findings:

- **Rewarding Most Loyal Users:** The 5 oldest users of the Instagram
- Remind Inactive Users to Start Posting: The users who have never posted a single photo on Instagram
- **Declaring Contest Winner:** The user who gets the most likes on a single photo
- **Hashtag Researching:** The top 5 most commonly used hashtags
- **Launch AD Campaign:** Day of the week do most users register on
- **User Engagement:** Average user posts on Instagram
- **Bots & Fake Accounts:** Users (bots) who have liked every single photo on the site

Approach: We are working with the product team of Instagram and the product manager has asked us to provide insights on the questions asked by the management team. We use SQL to derive different insights from the dataset provided by the management team. First, we run the necessary commands for creating the database to work on. Then, we performed analysis to generate valuable insights for the company.

Insights:

- There are total of 100 users using Instagram clone.
- Around 26% of the users are inactive in Instagram. We can remind the inactive users by sending them promotional emails to post their 1st photo.
- The most liked photo in Instagram is posted by Zack_Kemmer93, which is liked by 48% of the users. The team can start the contest for the most liked photos. This will make the users to post more such good posts.
- The most used hashtag is "smile". Around 59% of the users use the "smile" hashtags. If a partner brand use the "smile" hashtag, it will be able to reach the most users in the platform.
- The best days to launch ads are Sunday and Thursday. As the most users register on Instagram on Sunday and Thursday.
- 13% of Instagram IDs are fake and dummy accounts.

Project Link: INSTAGRAM USERS ANALYTICS.pdf

Operation Analytics & Investigating Metric Spikes

Description: Operation Analytics is the analysis done for the complete end to end operations of a company.With the help of this, the company then finds the areas on which it must improve upon.
Being one of the most important parts of a company, this kind of analysis is further used to understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst we must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that it's very important to investigate metric spike.

Findings:

- **Number of jobs reviewed**
- **Throughput**
- **Percentage share of each language**
- **Duplicate rows**
- **User Engagement**
- **User Growth**
- **Weekly Retention**
- **Weekly Engagement**
- **Email Engagement**

Approach: I am working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which I must derive certain insights out of it and answer the questions asked by different departments.
Firstly, I spent some time on understanding the data/table given. I cleared the questions which was in my mind and what are the things to consider while reviewing the data. I use SQL to derive different insights from the dataset provided by the management team. I first created a database "operation_analytics" and then the tables using the structure and links provided by the team. Then, we performed analysis to generate valuable insights for the company.

Insights:

1. Case Study 1 (Job Data):

- The number of distinct jobs reviewed per hour per day for November 2020 is 83%.
- We used the 7-day rolling average of throughput as it gives the average for all the days right from day 1 to day 7 whereas, daily metric gives the average for only that particular day itself.
- The percentage share of Persian language is the most (37.5%).
- There are two duplicate rows if we partition the data by job_id. But if we look the overall columns, all the rows are unique.

2. Case Study 2 (Investigating metric spike):

- The weekly user engagement increased from week 18th to week 31st and then started declining from then onward. This means that some of the users do not find much quality in the product/service in the last of the weeks.
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014.
- The overall count of weekly engagement per device used is the most for MacBook users and iPhone users.
- The email opening rate is around 34% and email clicking rate is around 15%. The users are engaging with the email service which is good for the company to expand.

Project Link: operation analytics and investigation metric spike.pdf

Hiring Process Analytics

Description: Hiring process is the fundamental and the most important function of a company. Here, 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 analyse before hiring freshers or any other individual. Being a Data Analyst, our job is to go through these trends and draw insights out of it for hiring department to work upon.

Findings:

- **Hiring:** How many males and females are Hired ?
- **Average Salary:** What is the average salary offered in this company ?
- **Class Intervals:** Draw the class intervals for salary in the company ?
- **Charts and Plots:** Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?
- **Charts:** Represent different post tiers using chart/graph?

Approach: I am working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked me to answer certain questions making sense out of that data.

We will use EDA to generate different insights and to answer the questions asked by the company.

The dataset given by the company contains the details about people who registered for a particular post in a department of this company. I used MS Excel to analyze the data with different tables and columns.

Insights:

- The rejection rate of male applicant is 6% higher than the female applicant.
- The average salary paid in this company is 50K.
- Most of the employers are in the Operation Department and then in the Human Resource Department.
- The applicant is most likely to get hired if he/she is applying for the HR Department as the rejection rate here is the least.
- There are only 3 candidates in the company who are paid more than 100K.

Project Link: hiring process analytics (1).pdf

IMDB Movie Analysis

Description: The dataset provided by the company contains various columns of different IMDB Movies. We are required to Frame the problem. For this task, we will need to define a problem we want to shed some light on.

We can do this by asking the following 'What?' :

- What do we see happening?
- What is our hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Findings: The things that we find out through the project are:

- movies with the highest profit
- top movies as per imdb rating
- top directors
- most popular genres
- top foreign language films

Approach: Firstly, I cleaned the data. Then, we used Five 'Whys' approach to determine its root cause by repeatedly asking the question "Why". While asking Why is easy, what we're interested in is the answer. Each time we answer why the next time gets more difficult as we must think deeper behind the reasons for this. As we ask why, we may find that we have multiple answers for the same question.

Insights:

- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is 'Avatar' followed by 'Jurassic World' and 'Titanic' and so on.
- The Shawshank Redemption is the top-most movie with the highest IMDB rating.
- The Good, the Bad and the Ugly (Italian) is the top-most foreign language movie.
- Charles Chaplin is the top-most director followed by Tony Kaye.
- The most popular genres is Drama followed by Comedy.
- 'Leonardo DiCaprio' is the critic-favorite as well as the audience-favorite actor.
- The most users voted in the decade 2000s and the least in the decade 1940s.

Project Link: IMDB ANALYTICS.pdf

EXCEL LINK : IMDB_Movies (1).xlsx

Bank Loan Case Study

Description: This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding: The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Suppose we work for a consumer finance company which specialises in lending various types of loans to urban customers. We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Findings:

- Our aim is to identify the 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.
- The driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- Presenting the overall approach of the data analysis, cleaning the dataset, finding outliers, data imbalance, univariate, segmented univariate, bivariate analysis, etc.
- The top 10 correlation for the Client with payment difficulties and all other cases (Target variable).

Approach:

- Used the COUNTA function to count the total rows in each column.
- Secondly, found the percentage of null values in each column using the formula – total row counts for each column / total row counts.
- Further, removed all the columns having null value percentages of more than 30%. For columns having less than 30% null value percentages, have done mean, median, and mode imputations for the missing values for columns having null value percentages less than 30%.
- Also found the outliers using the interquartile range method considering relevant columns.
- After going through each column description, have kept only relevant columns to bring out the insights.
- The columns having days were converted into years by simply dividing the days by 365.

Insights:

1. Non-Default:

- Academic degree has less defaults.
- Student and Businessmen have no defaults.
- Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%.
- People above age of 50 have low probability of defaulting.
- Applicant with Income more than 700,000 are less likely to default.
- People with zero to two children tend to repay the loans.

2. Default:

- Men are at relatively higher default rate.
- Clients who are either at Maternity leave OR Unemployed default a lot.
- Not approving the loan of young people who are in age group of 20-40 as they have higher probability of defaulting.
- When the credit amount goes beyond 3M, there is an increase in defaulters.
- People who have less than 5 years of employment have high default rate.

Project Link:

BANK LOAN CASE STUDY 1.pdf

IMPACT OF CAR FEATURES on price and profitability

Description:

- The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

- The given tasks below based on the business problem would require advanced Excel skills and knowledge of data analysis techniques such as regression analysis, pivot tables, sensitivity analysis, optimization, and time series analysis.

Findings:

Tasks: Analysis

- 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?

Building the Dashboard:

- 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?

Approach:

- This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis , descriptive statistics and visualization.

- The dataset contain 11,914 records along 16 dimensions. We remove duplicate rows. There are 715 duplicate rows. Find some blank rows then I apply some statistical function to fill the blank rows.

Insights:

- The higher car price has the maximum body style that indicates price plays an important role when it's come to body style, we can see that in the dashboard clearly.
- That also represents more variety for consumer demand since the value of the car required many features & style to give variety to people .
- When did the average for brand & body style with price its show that highest & lowest price does not vary by body style its remain the same no. of body style in both the category.
- Automatic transmission type has the highest no. of body style & Direct_drive is the lowest amount of car style which clearly tell us that consumers have more demand in Automatic cars.
- Average MPG & Engine does not have much different when we compare with brand they all vary between 14 to 18 MPG & 300 to 400 HP except Bugatti that has 14 to 200.

Project Link:

Impacr of car features.pdf

ABC Call Volume Trend Analysis

Description: The attached dataset is of Inbound calls of an ABC company from the insurance category consists of a Customer Experience (CX) Inbound calling team for 23 days. Data includes Agent_Name, Agent_ID, Queue_Time [duration for which customer have to wait before they get connected to an agent], Time [time at which call was made by customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred).

Findings:

- The average call time duration for all incoming calls received by agents (in each Time_Bucket).
- The total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time].
- Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.
- Propose a manpower plan required during each time bucket in a day[9 pm to 9 am]. Maximum Abandon rate assumption would be same 10%.

Approach:

- We used pivot table and pivot charts to get the valuable insights of the data.
- We assumed 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.
- We also assumed total days in a month is 28 days for easy calculation.

Insights:

- The customers call the least in the evening. So, the company can reduce the number of agents at that time for answering the calls.
- The company can hire 17 customer support agents for the night shift work.
- The company can shift some of the day workers for the night shift.
- The employees who are working 9 am to 9 pm. The manager can change some of the workers shift from 5 am to 2 pm and some workers from 2 pm to 11 pm to get the most calls answered.
- The company can make the employers divide into 3 parts too, so that the agents are always available 24/7.

Project Link: ABC Call volume analysis.pdf

Conclusion

Learnings: The things I learned from the projects are:

- Data Analysis 6 Steps Processes.
- How to use Advance SQL Concepts in the real world business case.
- How to use Advance Excel Concepts in the real business case scenario.
- How to analyze the huge datasets in Excel, etc.
- How to visualize the data to gain the valuable insights.
- Concepts of Operation Analysis & Investigating Metric Spikes.
- HR Analytics
- Predictive Analytics
- Risk Analytics
- Behavioral Analytics
- Business scenario of impact of car features.
- Customer Experience Team and Inbound Customer Support
- Google to get the concepts and answers whenever get stuck.

Portfolio Project Link: portfolio projects

Thank You