DATA ANALYSIS PORTFOLIO

BY – MOHAMMED MAAZ

PROFESSIONAL BACKGROUND

I have pursued B.E in Electronics and Communication engineering from MVJ College of Engineering in Bangalore.

During my academics I have done multiple projects like Biped Robot, Mini CNC Plotter using DVD Writers and Real time Arduino based Cardiac Disease Detection using Neural Network.

I have done Internships in Southwestern Railways (SWR) and Defence Research and Development Organization (DRDO).

I graduated from engineering in 2023 with overall 8.3 CGPA.

I developed interest in Data Analytics and did research on it and enrolled in Trainity Data Analytics course.

I developed skills on MS-Excel, SQL, Power BI, Tableau, Python, Statistics etc through this course. And now I’m looking forward to pursue my career in Data analytics field and showcase my skills and learn more.

Table of contents

Data analytics process

We have faced data analytics in real life always.

Like ex:- Planning a vacation

Data analytics process Steps:

Plan: I'm planning a vacation. I need to determine the destination, date of travel, travel budget, accommodation etc.

Prepare: I research on different travel destinations, travel dates, travel costs, accommodations and gather information.

Process: I process various factors such as weather conditions, tourist attractions, hotel reviews, travel options. I will make few choices based on preferences and constraints.

Analyze: I analyze the collected data to make decisions. For example, I might consider historical weather data to choose the best time to travel for my preferred destination, see reviews of hotels to make best accommodation choice, and see travel videos for travel experiences etc.

Share: I'll share my vacation plans with people who I have planned the vacation. I will see online reviews of people who have been to that tourist destination and consult with people who I have planned the vacation.

Act: Finally, after all the above 5 points, I'll select my tourist destination, book my flights, accommodations based on analysis and recommendations. During vacation I might use present data to make on spot decisions.

Project link - Click here

Instagram User Analytics

Description:

The project is about how the users engage and interact with Instagram. We will analyze 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.

Tasks:

A) Marketing Analysis:

Loyal User Reward: Identify the five oldest users on Instagram from the provided database.

Inactive User Engagement: Identify users who have never posted a single photo on Instagram.

Contest Winner Declaration: Determine the user with the most likes on a single photo and provide their details to the team.

Hashtag Research: Identify and suggest the top five most commonly used hashtags on the platform.

Ad Campaign Launch: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

B) Investor Metrics:

User Engagement: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.

Bots & Fake Accounts: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

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 perform analysis to generate valuable insights for the company.

Insights: By running SQL queries,

- We found out the five oldest users on Instagram. By this the management can give the Loyal User Reward.
- We found out that there are 26 users who have never posted a photo. The management should understand user behaviour and analyze the profiles of these users and find out reasons for not posting.
- We found that Zack_Kemmer93 with user id 52 has got highest number of likes i.e. , 48 on its photo with photo_id as 145. The management shall recognize the user publicly such as a dedicated post on yourofficial Instagram account, give them the declared and promised award.
- We found that the most commonly used #hastags are smile, beach, party, fun, concert. The management shall communicate the above hashtags to the partner but shall also give disclaimer that this is not a static data and can however change in near future as Instagram is highly dynamic so the partner should be careful and also to keep tracking changes if any, in future.
- We found that Thursday is the day when most of the users were registered.
- We found that out that there are 13 users who have liked all the 257 photos which any normal user would not be able to do and are therefore fake.

Project link - Click here

Operation & Metric Analytics

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:

- Jobs reviewed over time
- Throughput analysis
- Language share analysis
- Duplicate rows detection
- 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 - Click here

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:

 Males are 138% of the females employees.

 I found out that the average salary of general management is highest. Average salary is around 58,722.

 Most of the employers are in the Operation Department and then in the Service Department

 Most employees were employed on c9 post

 There are only 3 candidates in the company who are paid more than 100K.

Project link - Click here

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:

 distribution of movie genres and their impact on the IMDB score

 distribution of movie durations and its impact on the IMDB score.

 distribution of movies based on their language.

 Influence of directors on movie ratings.

 relationship between movie budgets and their financial success.

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:

 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.

 We found that movie with duration 100-120 minutes has maximum number of movies.

 The language in which majority of movies are made is English i.e., 3606. This amounts to 95% of the movies under analysis. The maximum std deviation is of Italian movies.

 The maximum number of movies were made by “steven Spielberg” I.e., 25 movies.

Project Link - Click here

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. The project is about to carry out an exploratory data analysis (EDA) for a bank. Here we are going to consider the different types of EDA (univariate, bivariate and multivariate analysis) in order to carry out the analysis and find the solution to problem statement.

Findings:

- Identify Missing Data and Dealing with it Appropriately.
- Identify Outliers in the Dataset
- Determining if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- Perform univariate analysis, segmented univariate analysis, and bivariate analysis using Excel functions.
- Identify Top Correlations for Different Scenarios.

Approach:

My first approach was to analyse the dataset, clean the dataset finding the blanks and missing values, imputing the missing values with the appropriate Method (mean, median, mode). Then i tried to find the outliers in the dataset, there are some anomalies such as negative values which need either to be deleted or standardized. After all these I used pivot tables and basic charts to visualise the data. Moreover, insights were drawn based on my understandings.

Insights:

 Customers who requested more than Rs.350,000 will probably be turned down.

 The majority of the loans requested by new customers were accepted, which has them ecstatic.

 A large percentage of loans requested through Credit and Cash firms are rejected.

 Almost 80% of the loans were approved, and the remaining 20% were consistently denied.

 Consumer loans have the highest approval rate and almost no cancellations. For the first Selling place area group, a number of loans were cancelled.

 The rate of rejection is higher for walk-in loans.

 Loans given to clients of MLM partners are probably going to be cancelled.

 Customers are more likely to have their loan cancelled if they apply for another loan within ten months of their prior loan.

Project Link - Click here

Impact of Car Features

Description: The project is focused on analyzing the impact of various car features on consumer demand and pricing in the automotive industry. By examining a dataset containing information on car models and their specifications, the project aims to identify which features are most popular among consumers and which are most influential in determining a car's price. Using data analysis techniques such as regression analysis, the project seeks to provide insights into how car manufacturers can optimize pricing and product development decisions to maximize profitability while meeting consumer demand. The project may also explore how changes in consumer preferences and trends in the industry, such as the shift towards electric and hybrid vehicles, are affecting the impact of car features on demand and pricing. Overall, the project aims to provide valuable insights for car manufacturers to improve their competitiveness in the market and increase profitability over time.

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

 Creating pivot tables that shows the number of car models in each market category and their corresponding popularity scores.

 Create a combo chart that visualizes the relationship between market category and popularity.

 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.

 Creating dashboards about various analysis.

Approach: Firstly, I cleaned the data. Understanding the data, features it contains, meaning behind it & what impact does it have on the target Handling the irregularities in the data such as null/missing values Analyzing the data, various relations between the features to derive conclusions Visualize the Analysis and create a Dashboard.

Insights:

 People in US like engine with higher horse power. As the engine power increases, so does the price.

 Engine Horse Power and Number of Cylinder a car has are the major contributors deciding the price

followed by efficiency of the car.

 Exotic cars hold up their value and aspiration over the years.

 Electric cars tend to lose their value over the years.

 The efficiency of the car is inversely related to the Engine Horse Power and the Number of Cylinders. As the Power and No. of Cylinders increase, efficiency of the car decreases.

 US market is different from Indian market, Indian market is more price conscious. In India, people usually look for efficient cars and do not mind compromising on Engine HP or how big the engine is. But in US, people like big powerful cars.

Project link - Click here

ABC Call Volume Trend

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 9Hrs out of which 1.5 Hrs goes into lunch and snacks in the office.

 On average an agent occupied 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 average call duration is more or less same for all the time buckets

 The number of calls received by the company increased during the peak time which is between 9 am to 1 pm but then as the day passed the calls reduced and by the end of the day the it were only about 5 % of total calls.

 The customers call the least in the evening. So, the company can reduce the number of agents at that time for answering the calls.

 We found that especially during peak time which is around 9am to 2 pm the number of agents required are substantially high as compared to currently employed workforce. The management needs to urgently employee more workforce as the current abandoned rate of calls is around 30% which is very high and in some time slots its up to 50%.

 In the last task I calculated the estimated calls in each time bucket that the company can receive during the night. I simply calculated the calls prorate basis based on the estimation provided to me.

This way I calculated the agents that the company is required during the night to answer the calls.

Project link - Click here

Conclusion

Trainity Data Analytics course has helped me learn a lot about Data Analytics. I have gained knowledge about Data Analytics and developed skills like MS-Excel, SQL, Power BI, Tableau, Statistics, Python etc. I have done projects related to real life scenarios using these skills and solved problems and developed insights. I'm looking forward to work in Data Analytics field and utilize the skills I have learned and gain experience.

Thank You