Data Analyst Portfolio Report

About me

Hi, I'm MD Ariful Islam! I have an analytical background in computer science and currently, I am on track to completing my Level 3 degree in Department For Education (DfE) Data Essential Skills Bootcamp from Cambridge Spark. I have developed a strong foundation in computer science and a passion for using data to uncover meaningful insights. I am excited to bring my technical and analytical skills to the field of data analysis as an entry-level data analyst.

I have been working for more than ten years with experience in IT, Telecommunication, Logistics, Retail, Healthcare and the Food Manufacturing industry. I have worked for more than 10 companies which gives me the ability to translate complex questions into understandable, customer service, management and leadership skills for the industry.

In my free time, I enjoy exploring new data analysis tools and techniques, and I am always looking for opportunities to expand my knowledge and skills. Whether working on a team or independently, I am driven by the thrill of discovering new insights and the satisfaction of using data to solve complex problems.

Education

Cambridge Spark, UK **Sep 2023 - Nov 2023**

DfE - Level 3 Bootcamp in Data Essential Skills

Oxford Home Study Center, UK **Jun 2021 - Dec 2021**

Level 7 Diploma in Global Project Management

University of Wales, UK. **Oct 2010 - Mar 2012**

Master's Degree in Information Technology Management

Darul Ihsan Univerity, Bangladesh. **Aug 2002 – Dec 2006**

Bachelor's Degree in Computer Science and Engineering

Project Work

Hangman Game Project (GitHub)

Using Python basic for making hangman game. Two players are playing the hangman game. I have used the if else condition in the function for choosing the guess and selecting hangman after three times loss.

Computer Vision Rock Paper Scissors Game Project (GitHub)

Using Python basic for making Rock Paper Scissors game. Here I have used TensorFlow and OpenCV in Python to take input from webcam and play games using webcam.

Multinational Retail Data Centralisation (GitHub)

In this project, I have implemented good practices of data extraction, cleaning and querying to subsequently assist in making business decisions for example real world environment. For this project, I have collected data from multiple sources such as AWS RDS databases, PDF conversions, API extractions, and S3 buckets containing CSV files and JSON files. For cleaning data I have used Pandas data frame in Python. Finally, I have used PostgreSQL for querying data.

Professional Certification

Data science and Machine learning Foundation(CFI)

Python Essentials (AiCore)

Python Programming (AiCore)

Overview of Learning

1 What is data and data type?

In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. Store data as 1 and 0 in the computer.

1.1 Data life cycle

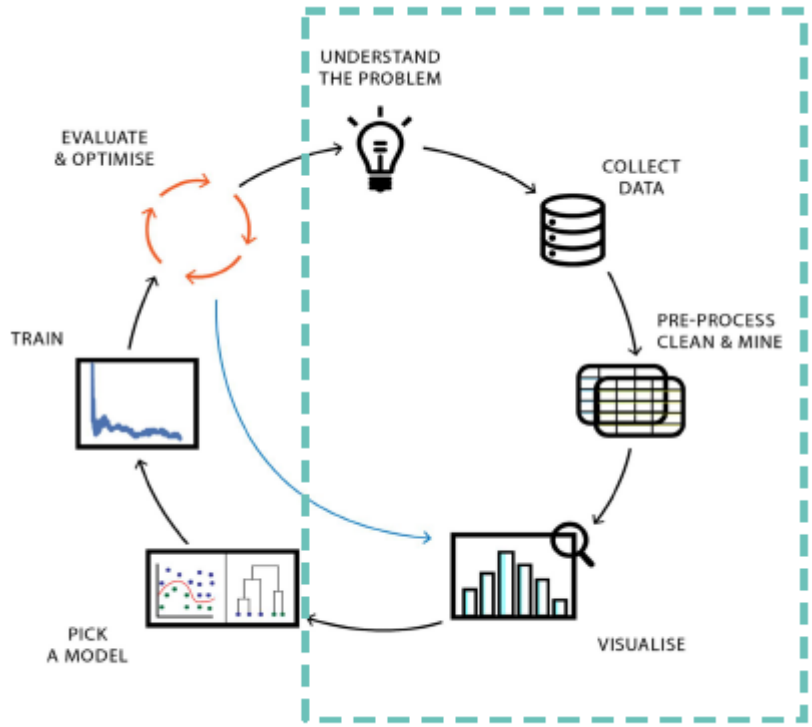


Figure : Data life cycle

In this data essential skills program, we have learned to understand the problem, collect, pre-process clean and mine and visualisation. The rest of the three steps are to pick a model, train, and evaluate and optimise for more professionals like Data Scientists.

1.1.1 Understand the Problem

Data is only useful if it can help to solve a problem. The first phase of the data lifecycle involves getting to grips with what the issues are around the problem. This can help to frame and focus on what data will be useful and relevant to solving this problem so you can then dive into the analysis to try and come up with a solution.

1.1.2 Collect Data

Once we have clarity on the problem(s) then we are trying to solve, this will make it much easier to know what types and sources of data will be most relevant. When understand the problem properly for the project then we will start data collection.

1.1.3 Pre-Process Clean and Mine

Once we have our relevant data that will likely help solve my problem, it will probably need editing before it gets analysed First, I will need to use a range of techniques to

get my data into shape and ready for analysis by cleaning and processing the data.

1.1.4 Data visualization in communication

Using data visualization to communicate business insights effectively involves translating data into visual representations that are easy to understand and analyze. Data visualization is a powerful tool that will help any audience grasp complex information quickly and make informed business decisions.

2. Data Sourcing & Integration

There are two sources primary data source and secondary data source. Primary data is generated by the company itself through questionnaires, surveys, interviews, and so on. Secondary data is generated by someone else and then can be extracted and used for specific purposes.

2.1 Data Integration

Data integration is the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process and includes steps such as cleansing, ETL(extract, transform and load,) mapping, and transformation. Data integration ultimately enables analytics tools to produce effective, actionable business intelligence.

3. Numbers for Data Analysis

Know the different types of data you might come across as a data analyst such as qualitative, quantitative, statistical analysis Mean, Median, and Average. Discrete and continuous data.

4. Microsoft Excel Function

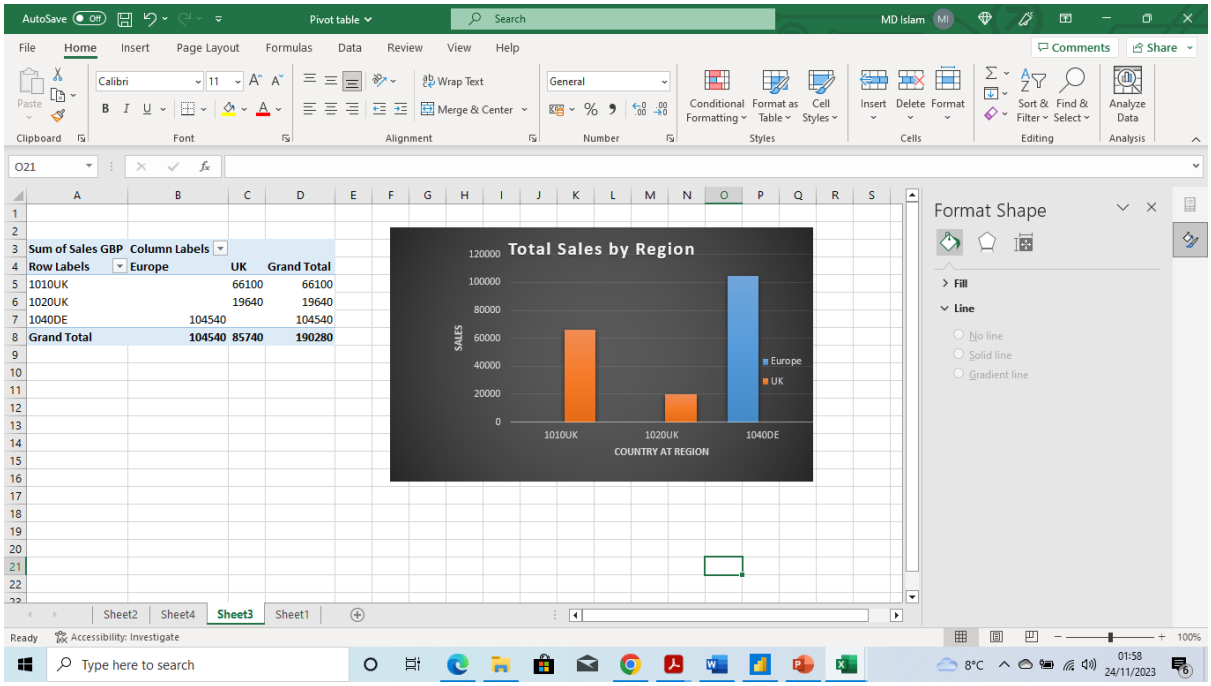
During this course, I have learned basic functions such as SUM, MEAN, MEDIAN, AVERAGE, MODE, LARGE, SMALL, MAX, MIN, and COUNTA and advanced functions such as VLOOKUP, HLOOKUP, LOOKUP, SUMIF, SUMIFS, COUNTIF,

CONCATENATE, TRIM, and conditional function AND, OR

5. Pivot table and power Query

As long as we can connect to the data, whether it be locally in the same workbook or remotely in other locations, we can build Pivot Table reports that rearrange the raw data and change it into meaningful information.

5.1 Creating Pivot Chart



5. 2 Power Query Editor

Power Query is a data loading transformation and preparation tool tool which we can find in Excel. Power Query Editor is a powerful tool for ETL (Extract, Transform, Load) processes in Power BI. It allows you to import data from various sources, clean and transform that data

to suit your needs, and then load the processed data into Power BI for further analysis and visualization.

6. Graph and chart

The effectiveness of the visualisation depends on what we choose to use. I have learned in this program such as histograms, bar charts, pie charts, cluster column charts, area charts, line charts, mapping doughnut charts and many more. I have also learned data visualisation such as MS Excell, MS Power BI, and Tableau.

7. SQL

7.1 What is SQL?

SQL is a very neat way to describe the logical layout of data. Structured query language (SQL) commands are specific keywords or SQL statements that developers use to manipulate the data stored in a relational database.

It is a language dedicated to making it easy to insert, update, and read from your database.

- Standard language for querying data.

- Supported by many modern distributed processing engines.

- Executed on the server.

7.2 Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with data both easy and intuitive. It is a fundamental high-level building block for doing practical, real-world data analysis in Python**.**

7.3 SQL vs. Pandas

- SQL is executed on server.

- Pandas is executed on client.

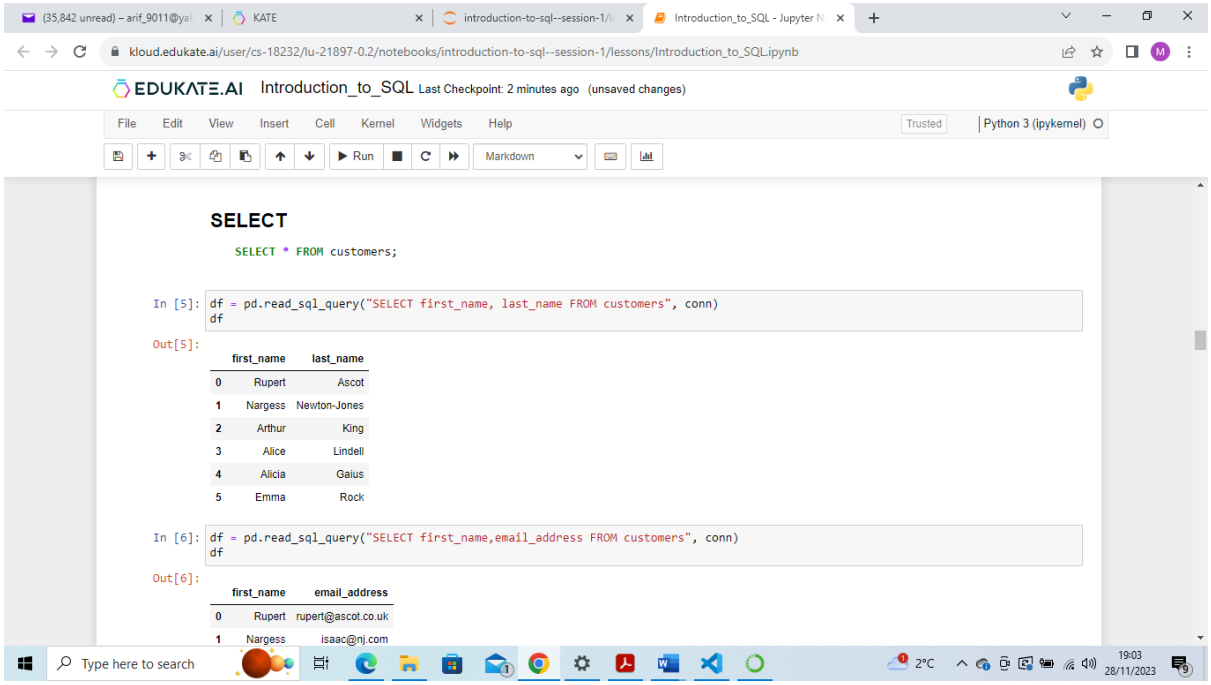
- SQL is decoupled from the implementation.

7.4 SQL Operator

SELECT

SELECT * FROM customers;

Screenshot

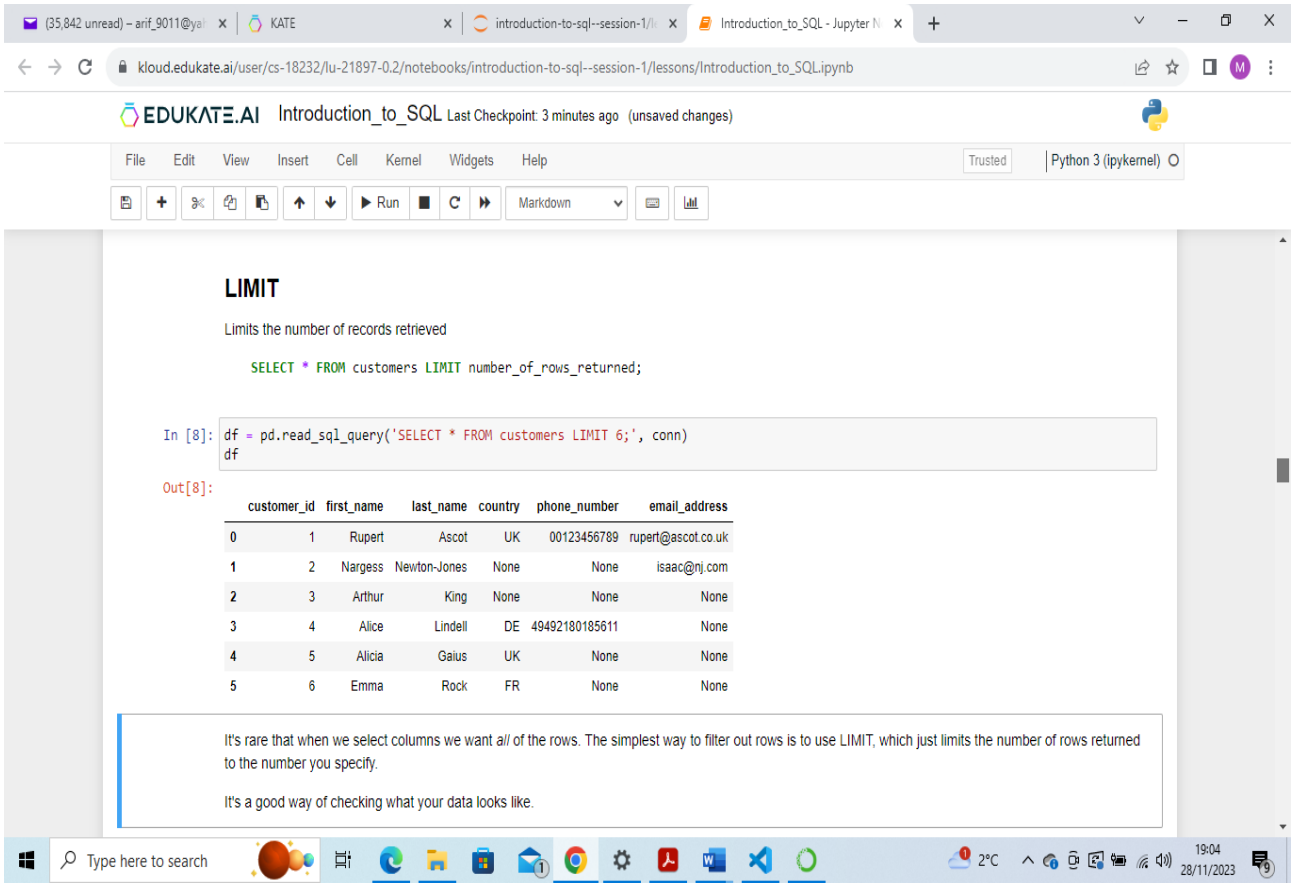


LIMIT

Limits the number of records retrieved

SELECT * FROM customers **LIMIT** number_of_rows_returned;

Screenshot

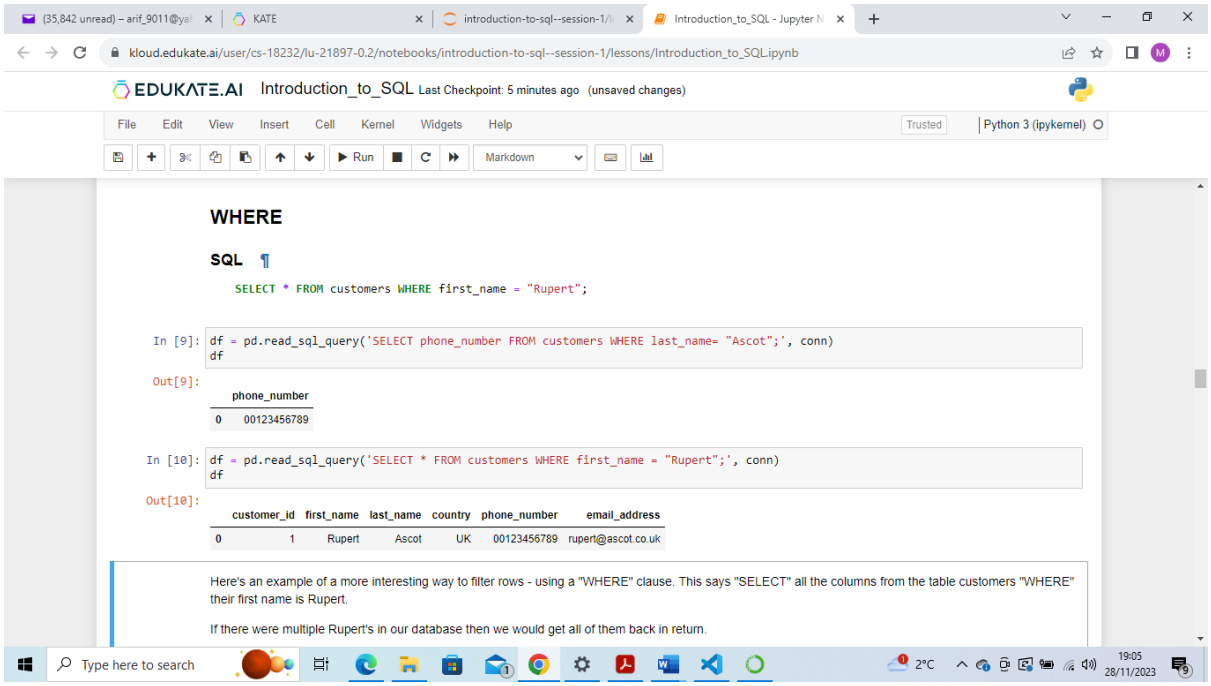


WHERE

SQL

SELECT * FROM customers **WHERE** first_name = "Rupert";

Screenshot

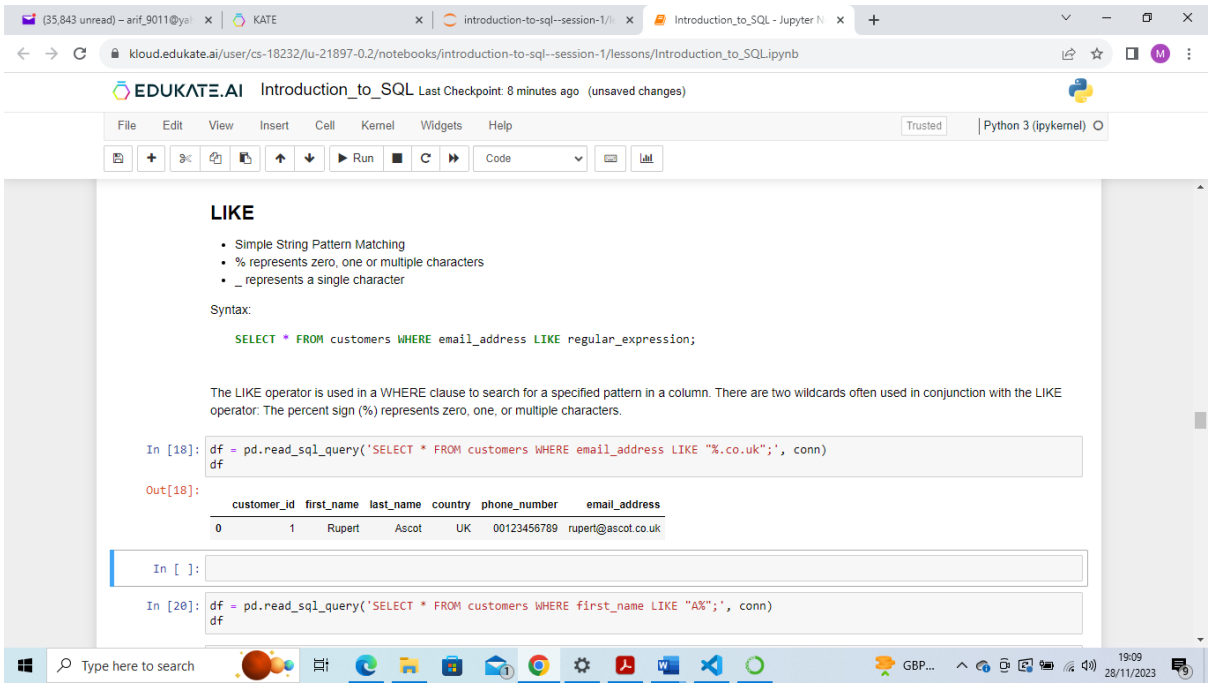


LIKE

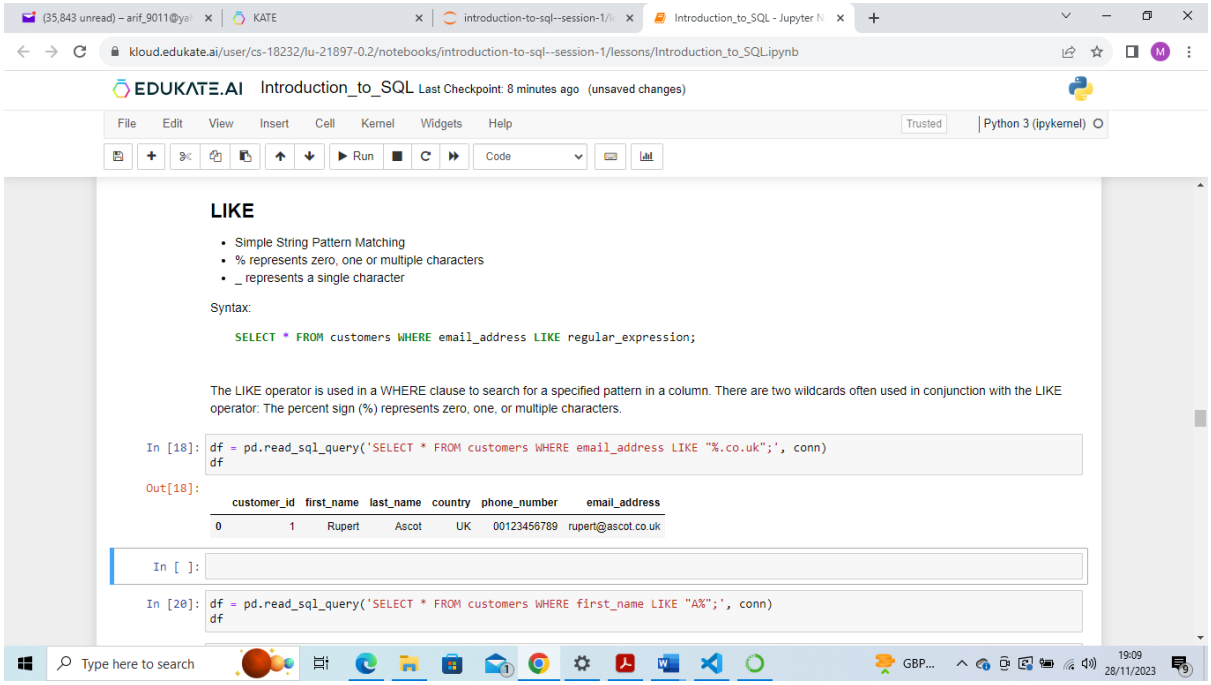
- Simple String Pattern Matching
- % represents zero, one or multiple characters
- _ represents a single character

Syntax:

SELECT * FROM customers **WHERE** email_address **LIKE** regular_expression;



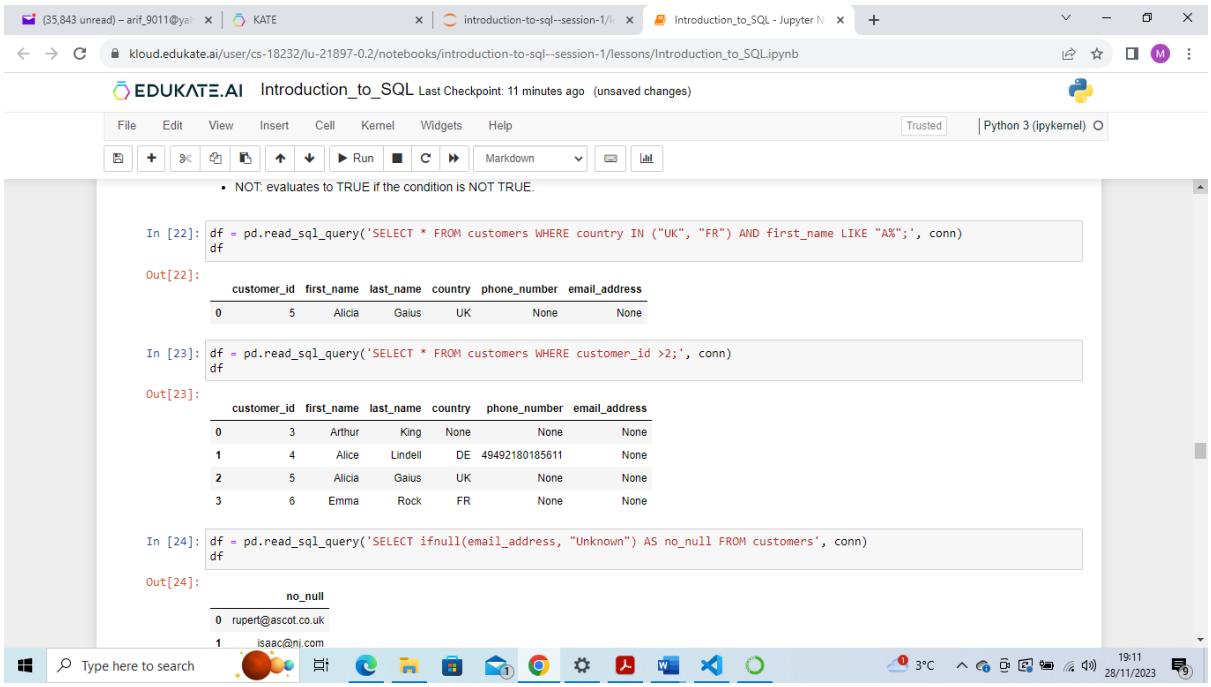
Screenshot



AND, OR, NOT

- AND: evaluates to TRUE if all conditions are TRUE.
- OR: evaluates to TRUE if any of the conditions are TRUE.
- NOT: evaluates to TRUE if the condition is NOT TRUE.

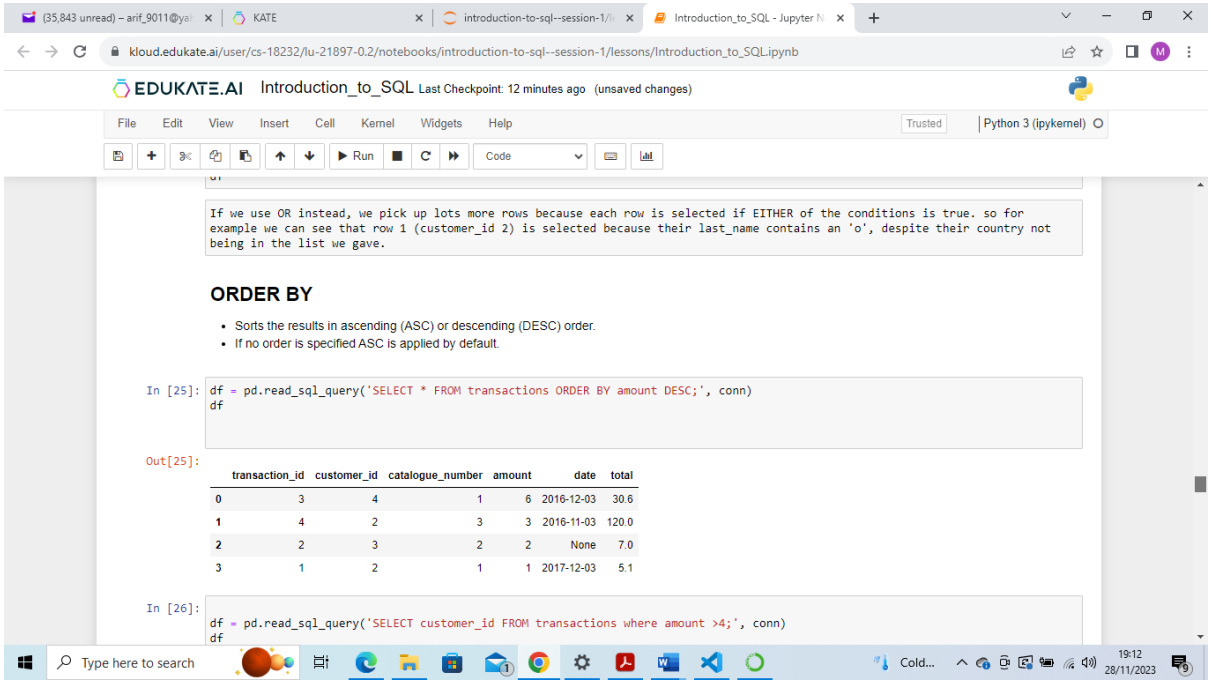
Screenshot



ORDER BY

- Sorts the results in ascending (ASC) or descending (DESC) order.
- If no order is specified ASC is applied by default.

Screenshot



Data Portfolio Project

IBM HR Analytics Employee Attrition and Performance

Introduction

HR Data Analytics portfolio project that deals with HR KPIs like Performance tracking, and attrition rate. The goal of this report is to analyse employee attrition. Attrition is described as the gradual loss of employees over time. Attrition is a major issue for all organizations, where it can lead to implications in staffing, employee morale, project costs, loss of experience, and a general hindrance to organizational growth.

We will examine the most important factors that influence attrition within an organization. We will consider if these factors are within the control of the organization and what actions can used to mitigate or combat attrition. We will also analyse current trends in HR and how these apply to our analysis, and finally, based on our results, we will conclude with insights and recommendations.

Project Object

My client is the IBM human resources director. He is trying to figure out the roots of employee attrition and improve the performance of the company. For that, he focuses on defining the parameters which cause the employee attrition rate via a proactive approach and tries to overcome that/those with the project's outcome. This study also helps to find out which employee are likely to leave organization.

Problem statement

Employee capital is one of the greatest assets an organization can possess. Companies can spend as much as 70% of total business costs on employees. These costs include salaries, training, recruitment, and skill investments. Furthermore, recruiting and keeping top talent is important to the growth and long-term viability of any company. Often employees hold key characteristics that are instrumental in moving the company forward. Knowing this, when employees decide to quit or leave a company, it can be a serious issue. With each employee, the company loses its direct investment along with all the knowledge and experience that the employee would have inherently provided. In the field of Human Resources, and HR, when

employees decide to quit, this is referred to as employee attrition, and this is the focus of our analysis.

Overview of Data

The core of our analysis is derived through a publicly available online dataset, providing variables akin to what a typical Human Resources personnel would have available. This dataset is known on Kaggle as the IBM HR Analytics Employee Attrition & Performance dataset. It is comprised of over one thousand and four hundred observations and thirty-five features—features and variables are used interchangeably in this report to represent the columns for which observations pertain.

Within the dataset, we have a mix of numeric and categorical datatypes. The initial dataset contains twenty-six numeric variables and it contains nine categorical data types. Examples of numeric data types include employee age, their monthly income, and the years that they've been working with the company. Examples of available categorical variables include education, gender, job role, or the department that the employee works with. It is important to verify the construction of the dataset and its variable types. This is because the make-up of the data directly influences the nature of what is possible in the analysis process.

Data Collection and Cleaning:

Data Collection

I use IBM HR Analytics Employee Attrition & Performance data from Kaggle, which was created by IBM data scientists. The dataset is on the open-source website and can be reached from this link. It has 1470 rows x 35 columns and contains numeric and categorical data types in columns. I loaded the dataset from this link in xlsx format and transformed it into Microsoft Power BI after importing the necessary libraries.

Data Cleaning

I began by meticulously cleaning and preparing the HR dataset in Excel. This involved handling missing values, standardizing data formats, and ensuring data integrity. According to checking spelling, I did not find any spelling mistakes with the data sets. I have searched for missing values in every column of the dataset, all rows look like having 1470 non-null entries. However, missing or null values can be found in Excel with True or False by using the ISBLANK function. For this reason, I have checked both missing values and duplicate

values in the dataset. Luckily, it was okay to continue to the next step. Because I did not have any missing or null and duplicate values.

This research also gave me a general impression of unique and top values for each attribute rate in addition to their frequencies in the dataset. I made double-checks on some of the features to make sure that everything was good to go. Those results were also okay.

I inspected the useless features to drop in the dataset. "YearsWithCurrManager", "StockOptionLevel", and " Standard Hours " for each observation and that did not impact or change anything in the data. For that reason, I have deleted those three useless columns.

Another important aspect is to make sure that all the data is in the correct format. This means ensuring that dates are in the correct date format, numbers are formatted as numbers, and text is formatted as text. So, I have corrected the formatting data and data type.

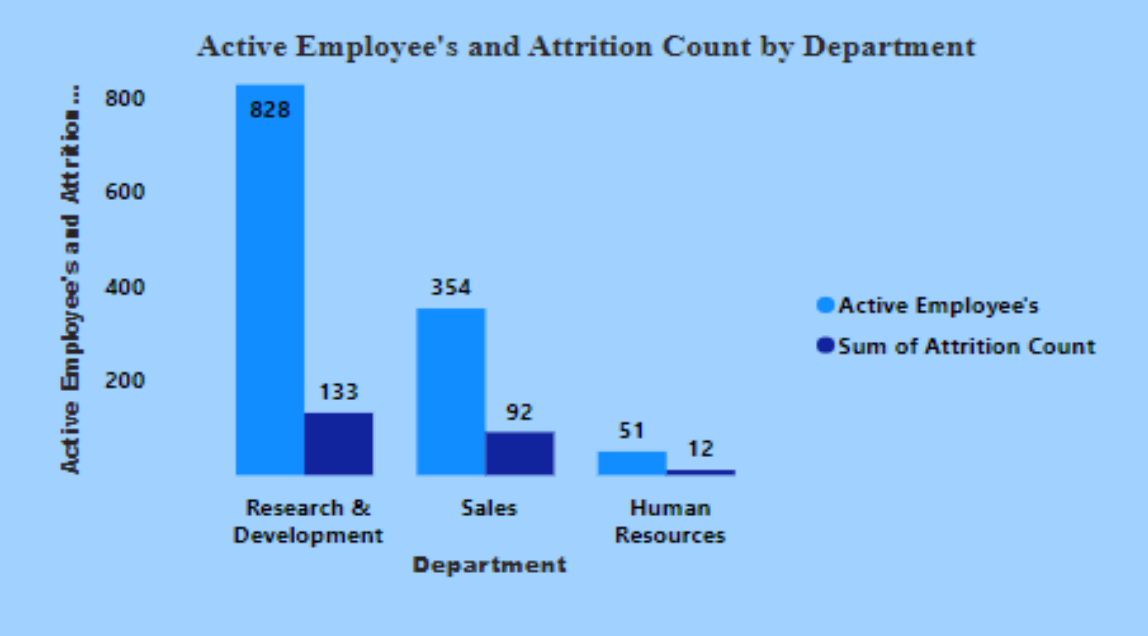
To be able to use it effectively in the further steps, I reassigned the response variable (Attrition) which had "Yes" and "No" values previously. I have created conditional columns "Attrition Count" in the power query editor and were assigned 1 for Yes and 0 for No respectively. After that, I moved the response variable to the last column place and changed the text format to number format.

I have created a conditional column age group with the help of the reference column of age.

groups like – Under 18, 19-30, 31-40,41-50, 50 above in power query editor.

Finding and Analysis

According to the clustered column graph presentation in Figure 1, active employees and the attrition count in the department of research and development are 828 and 133 respectively, the sales department 354 and 92 respectively, and finally human resources department 51 and 12 respectively. So, active employees and attrition are research and development are higher than in the sales and human resources department.



Fig**ure**1: Active Employee's and Attrition count by By Department

According to the pie chart presentation in Figure 2, attrition rates by gender, male and female are 87 and 150 respectively which is 36.71% male and 63.29% female. So, during this analysis, the attrition rate of females is higher than in males.

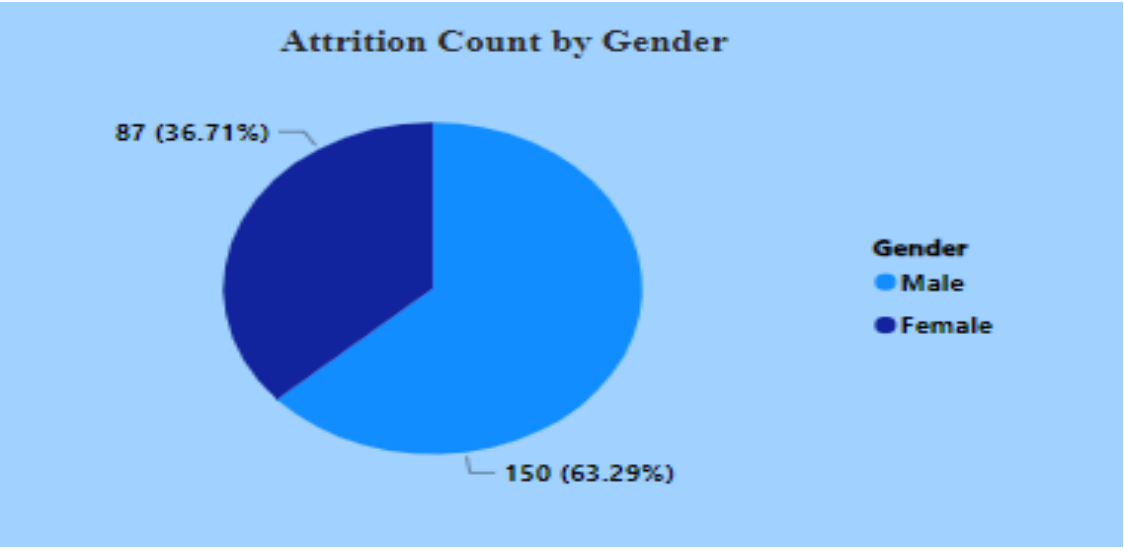


Fig2: Attrition count by Gender

According to the Clusterd bar chart presentation in Figure 3, the Sales Executive position has the highest score 1019. Then Research Scientist has the second highest score 925 and Laboratory Technician score is 819. Then manufacturing director is 462, Healthcare Representation is 413, Management is 326, Sales Representation is 26, Research Director and Human Resource is 163. The performance rating of sales executives is the highest and human resources is the lowest. So, I found human resource staff needs to improve their performance according to Figure 3.

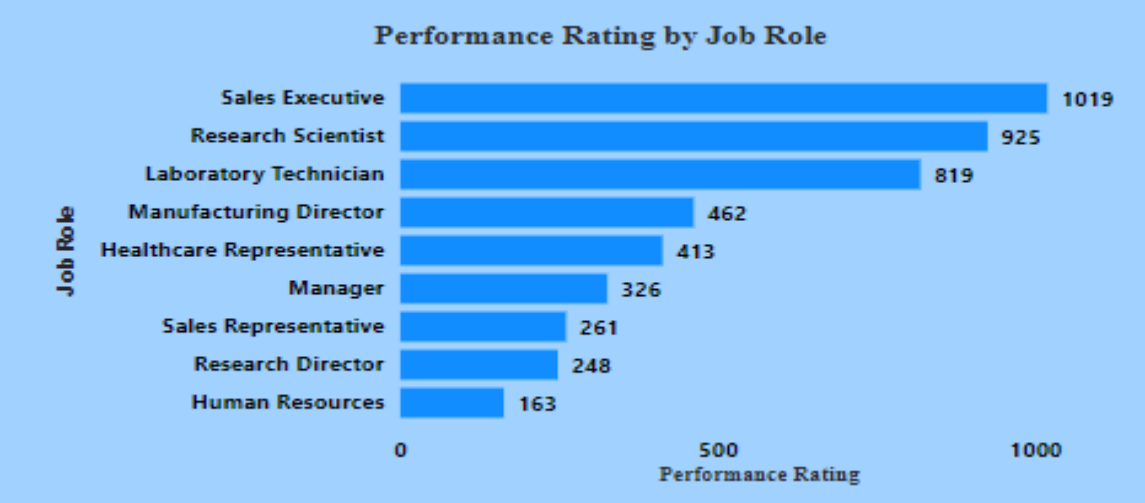


Figure 3: Performance Rating count by Job Role

According to the clustered column chart presentation in Figure 4, the average monthly income of married males is 7.2k and females 6.5k, divorced males 6.8k and females 6.8k, and single males 6.0k and females 5.8k. So, according to figure 4, I can say that married male average monthly earning is higher than female monthly earning. But with divorced male and female average monthly earning is equal. The single male average monthly earnings is slightly higher than a single female.

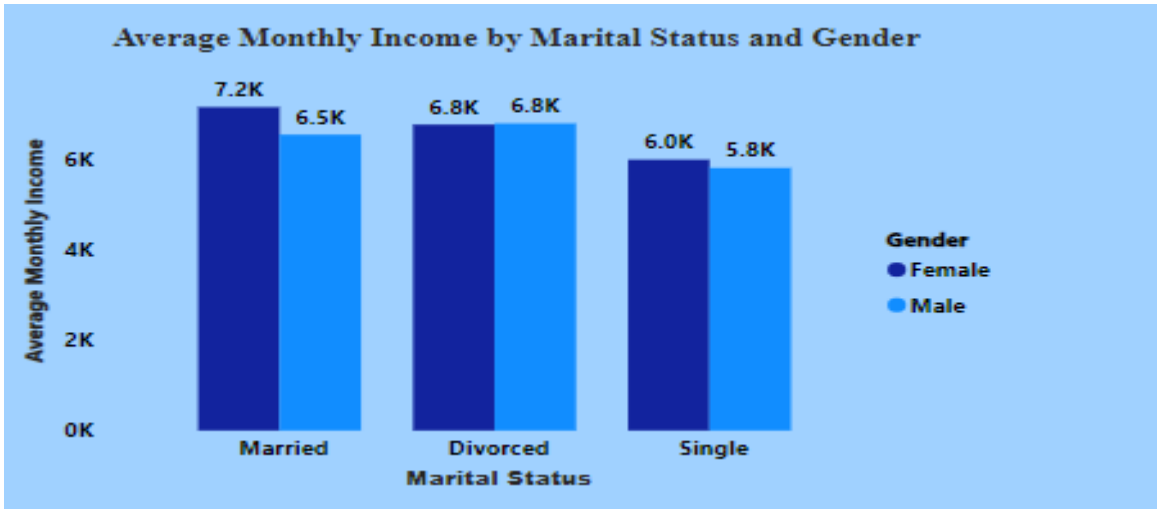


Figure 4: Average Monthly Income by Marital Status and Gender

According to the clustered column chart presentation in Figure 5, the age group under 18, is 50%, the age group 19-30, is 25.40%, the age group 31-40, is 13.73%, the age group 50 above is 12.59%, and the age group 41-50, is 10.56%. We can see in Figure 5 age group under 18 has the highest attrition rate in the company which means 50% overall attrition rate from the age under 18. The age group 19-30 are the second-highest attrition rate in the company.

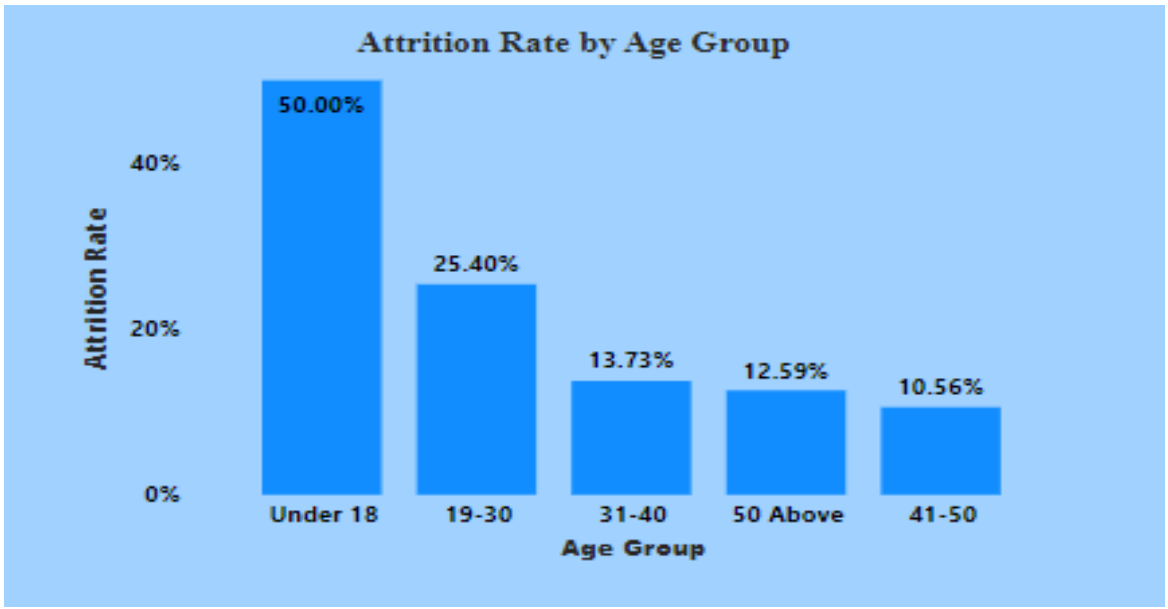


Figure 5: Attrition Count by Age Group

According to the line and clustered column chart in Figure 6, attrition count by business travel, there are high number of employees who frequently, followed by those employees who travel frequently and non - travellers. The highest number of Attrition rate is of the employee who travel frequently, while the lowest being the non -traveller one. So, in Figure 6, those travel frequently regarding their job role, have the highest attrition rate**.**

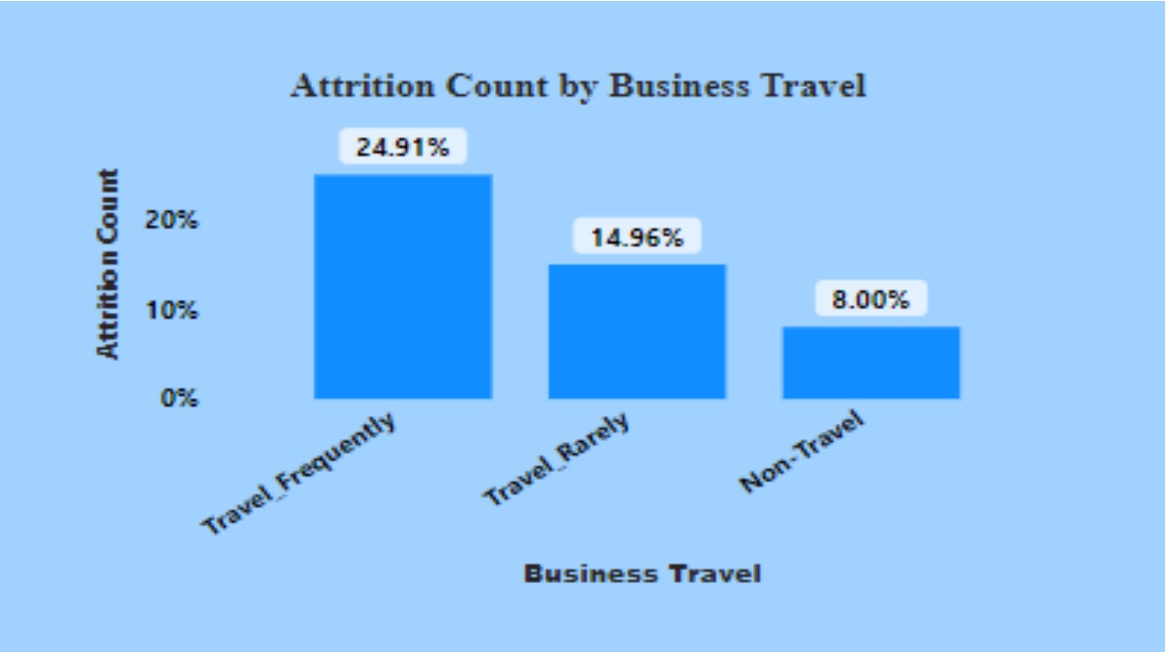


Figure 6: Attrition Count Business Travel

According to Figure 7, attrition count by education field shows in figure 7, life science 89, medical 63, marketing 35, technical degree 32, human resource 7 and others 11. So, life science is the highest attrition count. And medical is the second hight attrition count. Human resources is the lowest attrition count. In figure 6, Life Science education filed staff are leaving frequently than any other education filed.

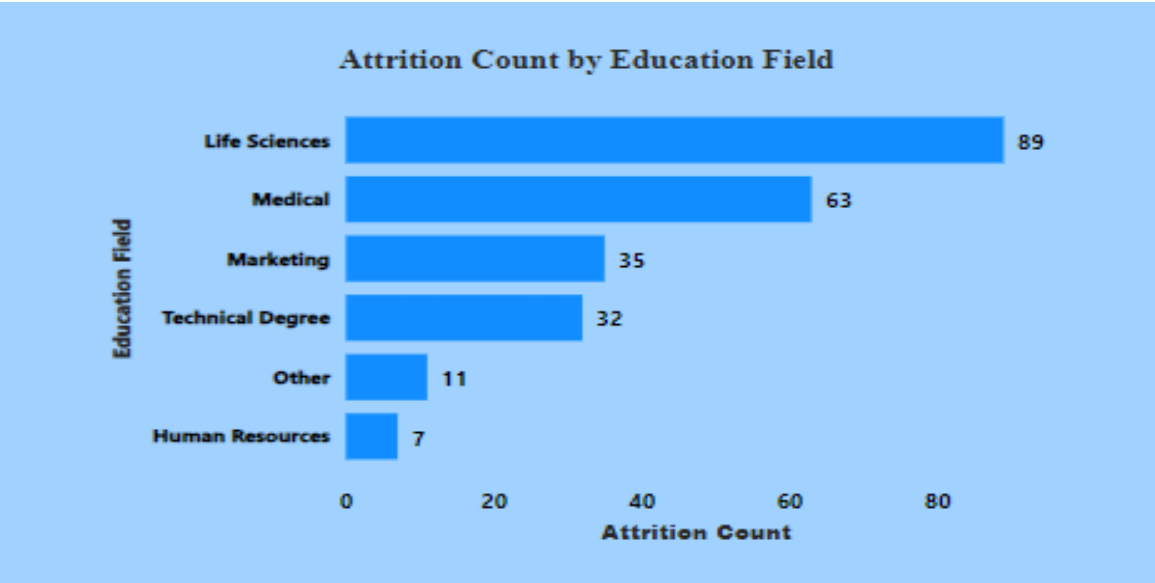


Figure 7: Attrition Count By education field

Dashboard Analysis

Project Description: This is an interactive HR analytics dashboard created using Power BI. The dashboard leverages data from an HR dataset provided by IBM on Kaggle. The dashboard focuses on exploring popular HR metrics and key performance indicators (KPIs), shedding light on various aspects such as attrition rates, job satisfaction ratings, and demographic insights.

Dashboard Highlights

1. **Attrition Analysis:** Explore attrition count and attrition rate across gender, education qualification, and department to gain insights into the factors contributing to employee turnover.

2. **Job Satisfaction:** Analyse job satisfaction ratings across different professions and age groups, providing a comprehensive overview of employee satisfaction levels.

Dashboard Features:

- Interactive visuals for exploring and analysing HR data.
- Utilization of DAX (Data Analysis Expressions) to create custom measures and KPIs.
- Card visuals displaying key metrics such as total employees, attrition count, attrition rate, active employees, female count, male count and average age.
- Detailed breakdowns using slicers to filter data based on gender, marital status.

Dashboard Preview

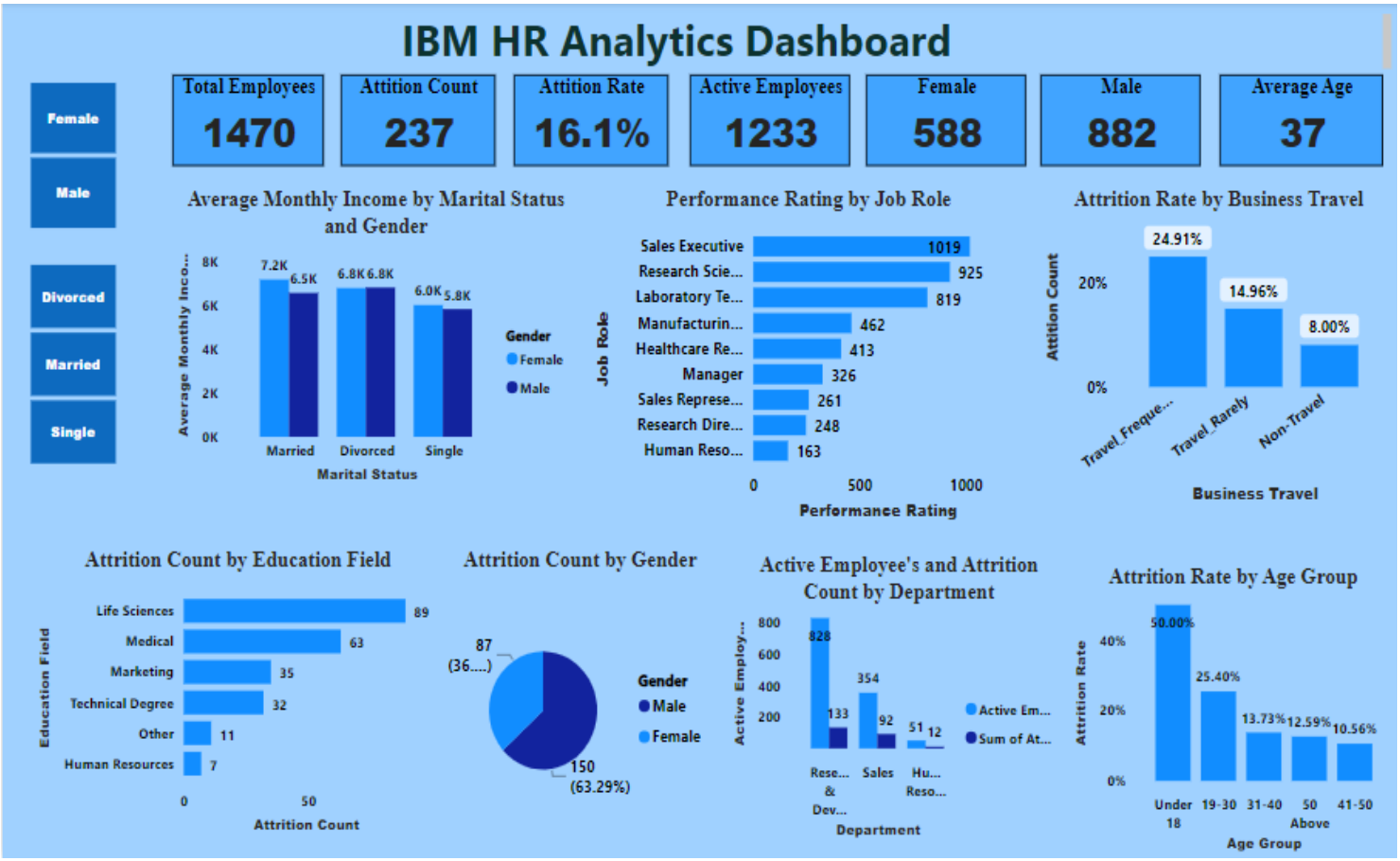


Figure: IBM HR Analytics Dashboard

Conclusion

This Power BI portfolio demonstrates my expertise in leveraging Power BI and related technologies to create interactive and insightful dashboards. Each project showcases different aspects of data analysis and visualization, highlighting my skills in data modelling, report creation, and data-driven decision-making. Feel free to explore each project to get a better understanding of my capabilities in Power BI in the future.