Data Analytics - SQL Project

PLEASE READ IT THOROUGHLY

For this project, you will be utilising the skills you have gained over the last few months to complete a Data Analytics Project, which mirrors the steps you would take in a data analyst or data science role.

The assignment sections are based on techniques and methods we have covered in the SQL, Data Visualisation and Excel sessions.

This project will need to be uploaded to your Github profile, so you can share your Github link on your social media channels. Your github repository should include your SQL code, your data visualisation, a write up/powerpoint of what you have done and any excel calculations you have done.

1. Choosing a data set

Choose **ONE** dataset from the following datasets listed below, or YOU can choose your own and use this one dataset **throughout** your project.
Not all the columns will be relevant, so please pick out the columns you think will be best suited to your project needs.
A zip folder containing each dataset will be sent to you as well. Each zip folder should contain one or two datasets, depending on what dataset you pick. This is mainly for the purposes of the SQL task.

Below is a description of each dataset:

Healthcare:

- **Life Expectancy WHO(World Health Organisation) Dataset**: this dataset will focus on immunisation factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.
  For more info see:
  https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Politics: This contains two datasets - will require using joins

- **2016 EU Referendum in the United Kingdom**: A referendum was held on the 23 June 2016 to decide whether the United Kingdom should remain a member of the European Union or leave. Approximately 52%, or more than 17 million people, voted to leave the EU. The referendum turnout was 72.2%, with more than 33.5 million votes cast. The Electoral Commission published the results of the EU referendum by

district and region after the vote. The Office of National Statistics provided the population demographics by district from the 2011 United Kingdom Census. https://www.kaggle.com/datasets/electoralcommission/brexit-results

Marketing:

- **Customer Personality**: is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers. Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyse which customer segment is most likely to buy the product and then market the product only on that particular segment. For more info on the columns: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Finance:

- **Credit Card Customers**: A manager at the bank is disturbed with more and more customers leaving their credit card services. This dataset consists of 10,000 customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. There are nearly 18 features. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Retail: This contains two datasets - will require using joins

- **Customer Transactions**: With the retail market getting more and more competitive by the day, there has never been anything more important than the ability for optimising service business processes when trying to satisfy the expectations of customers. Channelizing and managing data with the aim of working in favour of the customer as well as generating profits is very significant for survival. Ideally, a retailer's customer data reflects the company's success in reaching and nurturing its customers. A Retail store is required to analyse the day-to-day transactions and keep a track of its customers spread across various locations along with their purchases/returns across various categories. https://www.kaggle.com/datasets/darpan25bajaj/retail-case-study-data?select=prod_cat_info.csv

Education:

- **Covid19 and Education**: COVID-19 and its impact on education, social life and mental health of students: A Survey Link to the paper In this study, a cross-sectional survey is conducted with a sample size of 1182 students of different age groups from different educational institutions in Delhi National Capital Region (NCR). https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students

2. Excel for Data Manipulation

Choose 2 or 3 techniques

This section focuses on utilising the skills gained from Excel to manipulate your data.
You can use techniques/Formulas such as:

- SUM
- VLOOKUP
- XLOOKUP
- HLOOKUP
- IF
- COUNTIF
- MAX/MIN

3. SQL For Data Analytics

The data analysis section is based on the content we covered in the SQL sessions.
Your project **MUST** contain all of the following techniques:

- Aggregations of your chosen dataset, using SUM, COUNTS, AVG
- GROUP BY
- Using WHERE to filter your data
- CASE WHEN to add additional columns

If appropriate:

- Sub queries
- Joins
- Window Functions

It is very important in a data investigation to write your observations and assumptions down in the SQL editor so we can follow your thoughts.

4. Dashboard

Based on what you have discovered in your data analysis in sections 2&3, start building out your data visualisation. You can use either PowerBI or Tableau for this section.
Remember all your visualisation best practises.

- Look out for any trends in your data.
- Are there any correlations
- Can you see any changes overtime

5. Python - Optional

If you feel confident and comfortable - then you are more than welcome to add any Python work to your project.

6. Add your project to your Github Profile

Guidelines on how to add to create a Github profile were covered in the last session.

Your Github profile must contain the following:

- Your SQL file with your queries and comments (Upload as .sql)
- Your Dashboard (PDF)
- The datasets used
- Your excel workbook - (CSV File)
- Powerpoint as a PDF

Additional Guidance

Healthcare - Life Expectancy WHO (World Health Organisation) Dataset

- How does infant deaths and adult mortality impact life expectancy?
- Focus on one country and see how the life expectancy has changed across the years?
- What impact does alcohol, BMI impact life expectancy
- In 2015, what were the top 10 countries with highest life expectancy vs lowest
- Of those top 10 countries and bottom countries, what was the impact of issues like polio, measles, HIV_AIDs etc?

Dataset Description		
Status: Developing or developed country	Life Expectancy: life expectancy in age	Adult Mortality: Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant Deaths: Number of Infant Deaths per 1000 population	Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis: immunisation coverage among 1-year-olds (%)	Measles: number of reported cases per 1000 population	BMI: Average Body Mass Index of entire population
Under Five Deaths: Number of under-five deaths per 1000 population	Polio:Polio (Pol3) immunisation coverage among 1-year-olds (%)	Total Expenditure: General government expenditure on health as a percentage of total government expenditure (%)
diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	HIV_AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)	GDP: Gross Domestic Product per capita (in USD)
Population: Population of the country	Thinness_1_to_19year: Prevalence of thinness among children and adolescents for Age	Thinness_5_to_9years: Prevalence of thinness among children for Age 5 to 9(%)

	10 to 19 (%)	
Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Schooling: Number of years of Schooling(years)	

Politics: 2016 EU Referendum in the United Kingdom

- What is the average leave percentage & average remain percentage?
- How does the votes differ across Area
- How many residents are there in each Area? What % are old v young?
- Sum the under 30s age range and the over 30s age range & do some analysis on the voting preferences
- How do these voting preferences by age differ by region too?
- What area had the max/min remain/leave percentage?

Columns you may not know
Referendum:
Electorate: all the people in a country or area who are entitled to vote in an election
Rejected Ballots = No official mark + multiple marks + writing or mark + unmarked or void
Census is just population data for each age bracket

Marketing: Customer Personality

- What is the average age of people who respond to a marketing campaign?
- How does the response rate differ across demographics?
- How does the response differ across the number of purchases?
- Average number of store purchases vs websites
- What's the impact of this on response?

Dataset

People
ID: Customer's unique identifier
Year_Birth: Customer's birth year
Education: Customer's education level
Marital_Status: Customer's marital status
Income: Customer's yearly household income
Kidhome: Number of children in customer's household
Teenhome: Number of teenagers in customer's household
Dt_Customer: Date of customer's enrollment with the company
Recency: Number of days since customer's last purchase
Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products
MntWines: Amount spent on wine in last 2 years
MntFruits: Amount spent on fruits in last 2 years

MntMeatProducts: Amount spent on meat in last 2 years
MntFishProducts: Amount spent on fish in last 2 years
MntSweetProducts: Amount spent on sweets in last 2 years
MntGoldProds: Amount spent on gold in last 2 years

Promotion
NumDealsPurchases: Number of purchases made with a discount
AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place
NumWebPurchases: Number of purchases made through the company's website
NumCatalogPurchases: Number of purchases made using a catalogue
NumStorePurchases: Number of purchases made directly in stores
NumWebVisitsMonth: Number of visits to company's website in the last month

Finance - Credit Card Customers:

- How many customers have churned (attrition flag = 1)
- How does churn (attrition) look across the different demographics?
- What's the average utilisation?

CLIENTNUM: Client number. Unique identifier for the customer holding the account
Attrition_Flag: Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age: Demographic variable - Customer's Age in Years
Gender: Demographic variable - M=Male, F=Female
Dependent_count: Demographic variable - Number of dependents
Education_Level: Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status: Demographic variable - Married, Single, Divorced, Unknown
Income_Category: Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category: Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book: Period of relationship with bank
Total_Relationship_Count: Total no. of products held by the customer
Months_Inactive_12_mon: No. of months inactive in the last 12 months
Contacts_Count_12_mon: No. of Contacts in the last 12 months
Credit_Limit:Credit Limit on the Credit Card
Total_Revolving_Bal: Total Revolving Balance on the Credit Card
Avg_Open_To_Buy: Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1: Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt: Total Transaction Amount (Last 12 months)

Total_Trans_Ct: Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1: Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio: Average Card Utilization Ratio

Retail: Customer Transactions

- Number of transactions across gender
- Number of items in Personal appliances - total sales, total tax and qty
- Top selling products in clothing - women sub category
- products with most returns
- highest transaction date

Education: Covid19 and Education

- Avg time spent on online class
- Categorise the age groups and look at the impact on time spent on online classes
- link between health issues and time spent on online classes
- link between time spent on social media vs sleep vs fitness
- No of good, excellent responses

The actual research paper:
https://www.researchgate.net/publication/347935769_COVID-19_and_its_impact_on_education_social_life_and_mental_health_of_students_A_Survey