Capstone Project

Graded Project – 90 marks

Part I – Excel (45 Marks)

Problem Statement:

The case is about a bank which has a growing customer base. Majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 7% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

The department wants to analyse and visualize data that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

The file Bank.csv include various information (data) collected by the bank on the customer like their demographics, information (age, bank balance, etc.), the customer's relationship with the bank, the customer response to the last personal loan campaign (poutcome) etc

Domain: Marketing analytics

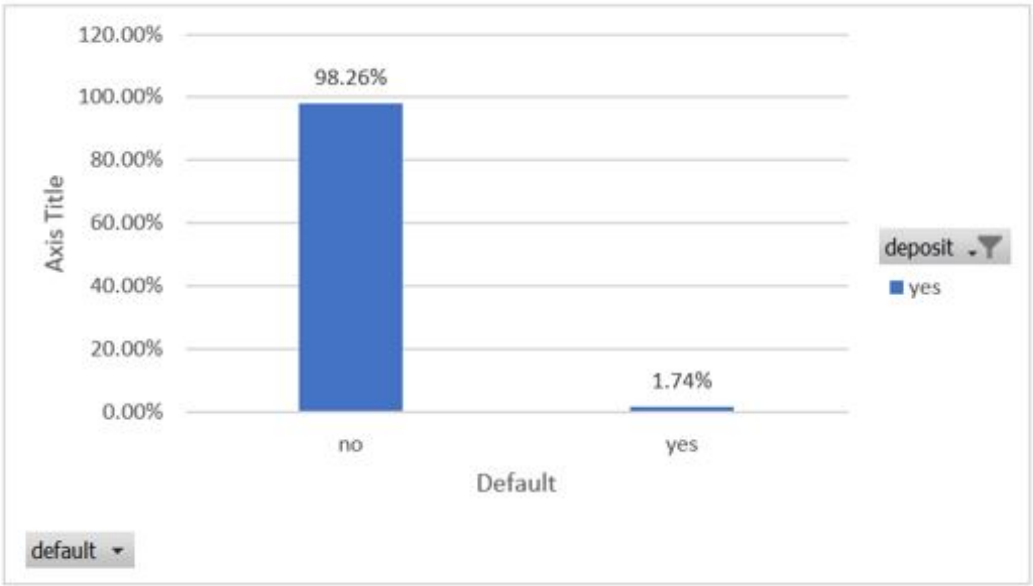
Data Description: The data set contains the following fields:

- age: Age of the customer
- Job: Type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital: Marital status (categorical: 'divorced','married','single'; note: 'divorced' means divorced or widowed)
- Education: Level of education attained. (Categorical: "primary", "secondary", "tertiary", "unknown")
- default: Has credit in default? (categorical: 'no','yes')
- balance: amount of money held in the bank account at the given moment.
- Housing: Has housing loan? (categorical: 'no','yes','unknown')
- Contact: Contact communication type (categorical: 'cellular','telephone')
- Day: Last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- Month: Last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Duration: Last contact duration, in seconds (numeric)
- Campaign: Number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Previous: Number of contacts performed before this campaign and for this client (numeric)
- Poutcome: Outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
- Loan Deposit: Has the customer subscribed to the loan deposit or not?

Tasks/ Questions to be Answered:

You are free to use excel functions, pivot tables, pivot charts, slicers and charts to answer the questions below. As part of this problem, you will use only basic visual metrics to gather powerful insights on data. You will be also able to appreciate why visualization is the key to any leader who wants to derive business value out of the data

1. Understanding Data and Missing Value treatment (6 points)
    a. Provide a basic data statistics
    b. Observe if there are missing values in the data and report the no. of missing values in the dataset. (Hint: Report the number of missing values variable vice)
    c. Replace the missing data with suitable values for Numerical and Categorical Columns (Hint: Use median values for numerical variables and Modal values for categorical variables)
2. Understanding the customer (6 points)
    a. What is the mean age of the customer base?
    b. What percentage of the customers in the given data set are less than 45 years old?
    c. What is the minimum and maximum account balance of the customer base given in the dataset?
3. Use visual analysis to gain insights on customers (6 points)
    a. What percentage of customers have made a loan deposit with the bank?
    b. Under the “job” column, what are the different categories
    c. Under the “job” column, which category of customers are more prone to subscribe to bank’s services?
    d. In the “job” column, what category of people have taken the highest number of home loans? (Hint: Express figures in both numbers and percentages.)
4. Use visual analysis to gain insights on transitions (6 points)
    a. Refer to the chart given identify what should the title of y-axis be?
    b. Plot the same using the data
    c. What conclusion can you draw from the given chart?



5. Correlating the variables for effective decision making (6 points)
    a. What mode of contact has been most efficient in getting more people to make loan deposits?
    b. What age group of people made the greatest number of loan deposits?

   c. Within the above-mentioned age group, which job category has the highest percentage of loan deposits?
6. Hypothesis Testing: (10 points)
   a. A marketing specialist says “customers who are married are a better target audience for a loan deposit when compared to customers who are not married”. With the help of visual analysis can you prove or disprove this hypothesis?
      i. Pivot table: 2 points
      ii. Visualization: 2 points
      iii. Inference - 1 points
      iv. Use any hypothesis test (Optional)
   b. A marketing executive claims “Blue collar customers with secondary level of education are more likely to make a loan deposit when compared to other category of jobs.
      i. Pivot table: 2 points
      ii. Visualization: 2 points
      iii. Inference - 1 points
      iv. Use any hypothesis test (Optional)
7. Business report, Inferences and observations for all the Tasks (5 Marks)
   a. Each step needs to be executed clearly separately in separate sub sheet of excel before making a final dashboard
   b. Please note the inferences, Observations and Conclusion should be available both in the workbook and the report clearly.
   c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
   d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
   e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis

Part II – SQL (45 Marks)

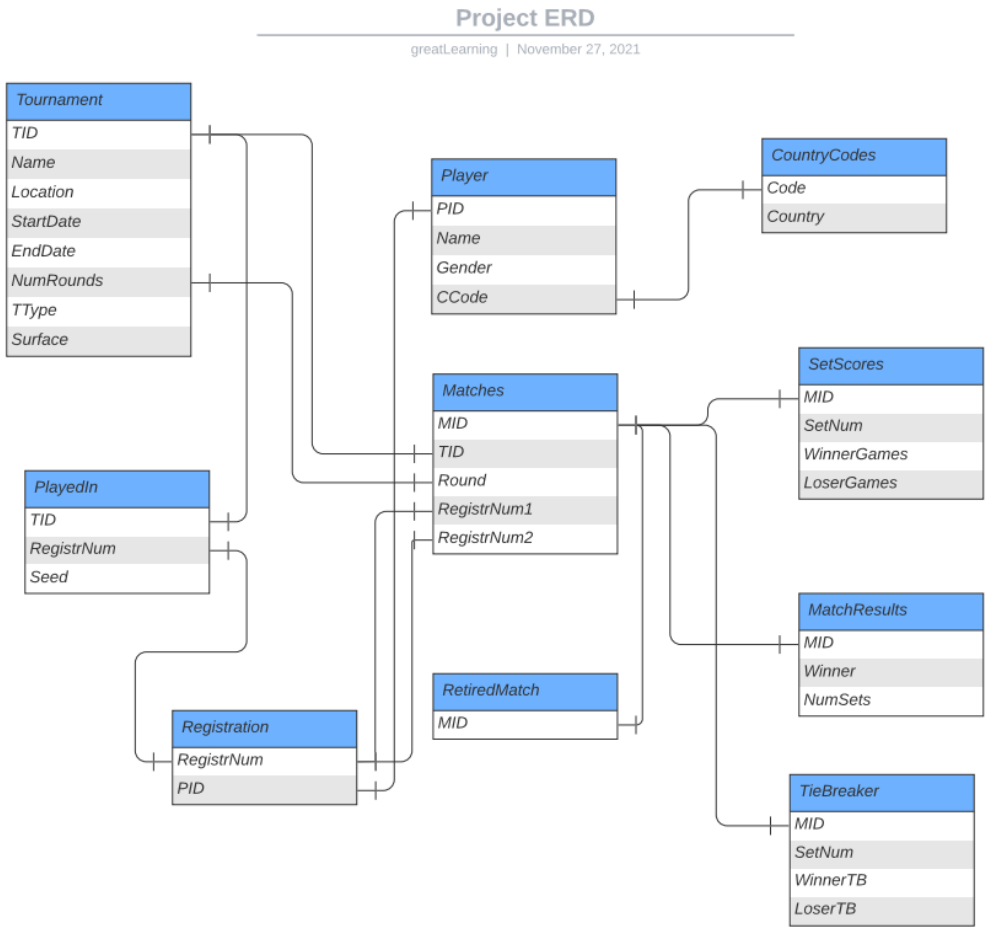
Problem Statement:

The database you will be writing queries about contains information concerning the Association of Tennis Professionals. It contains information about players, matches and tournaments concerning the ATP. If you are unfamiliar with tennis terminology (especially concerning scoring), a quick description can be found on Wikipedia at here. This assignment will enable constructing SQL queries for the supplied data based and preparing data for future analysis.

Domain: Sports

Data Description:

The ERD diagram of the Database is as given below



Several assumptions about the database are included below:

- The source of the data used for this database only contained information about the ATP (men's tennis) and not the WTA (women's tennis). All players involved are male; we have no information pertaining to women's singles or doubles, or mixed doubles.
- While it should be clear that Player's PID and Tournament's TID field should be unique, this is also true for Match's MID field; no IDs are re-used throughout different tournaments. Thus, a

Match's MID field uniquely represents it across all Matches known to exist. This avoids having to compare Tournament IDs as well as Match IDs for several of the queries.

- When a player participates in a tournament, he first has to register. For the purposes of our database, this includes receiving a registration number for the tournament. Because the ATP manages all of the tournaments in our database, we assume that a registration number is unique throughout all of the tournaments. This means that given a only a registration number, we can determine the tournament it corresponds to.
- A Player's ID (Player.PID) is only used in our database for registration in tournaments. All other locations the Player is referred to by the RegistrNum (notably in the Winner attribute of MatchResults).
- Two players participating in doubles tennis as a team share the same registration number. Note that several players register for both singles and doubles tennis in the same tournament, players in this category get two separate registration numbers.
- Both a player registered for singles play or two people registered for doubles play under the same number (i.e., are playing with each other) can both be referred to as a team: a team of one or a team of two.
- In most tournaments, the top fraction of the players registering are seeded, to help spread out the players expected to perform well (this makes the later matches more interesting!). Seed values are given on a per-tournament basis. Players who do not receive a seed in a particular tournament have a NULL value instead.
- The Tournament relation stores the number of rounds the tournament has as an attribute, which is an integer. The Round attribute of the Match relation is similar. A value of 1 corresponds to the first round, 2 to the second, and so on. Match.Round does not say "Quarterfinals," "Semifinals," or "Finals." You will have to figure out how to compute what rounds correspond to these. Don't try to hard-code values in (i.e. if you look at the database and realize that the US open singles had a total of 7 rounds, so the Semifinals is round. Rather, use a nested query to help.
- Each entity in the Set relation contains information about a particular set of some match. The attributes involving the "Winner" refer to the winner of the winner of the match, not necessarily the set (similarly for the "Loser" attributes). In cases where the set was determined by a tie-breaker, the TieBreaker relation has WinnerTB and LoserTB, which contain the points won in the tie-breaker game. Again, the "Winner" column of the TieBreaker refers to the winner of the match, not necessarily the set.
- Sometimes players will need to retire from a match, which counts as a forfeit. If this happens, we keep records for all completed sets of the match. For the set in which the player retired, the WinnerGames and LoserGames attributes hold the number of games won and lost. The RetiredMatch relation holds all Match IDs where the loser retired.
- Many of the queries will require some date manipulation. Specifically, you may have to extract the year a particular tournament was held. Assume that no tournaments wrap around years (i.e. late December-early January), and only concern yourself with the year of the start date.

Tasks/Questions to be Answered:

Please Note: You are required to answer the following questions in SQL and paste screenshots of the query along with the output in case of SQL statements in the submission document

1. To answer the following question, include results from both doubles and singles matches. List the names of all players who have ever been assigned a seed for any tournament (either doubles or singles). (6 Points)
2. There are 2 parts of the question (8 Points)
    a. List all tournaments having more than 5 rounds.
    b. Print the name of the tournament, the tournament type, the start and end dates, and the number of rounds. (Hint: Use joins)
3. There are 2 parts of the question (8 Points)
    a. For all tournaments in the database, list the name, tournament type, surface type, and the number of rounds it has.
    b. Sort the results in descending order by the number of rounds.

4. List the names, tournament types, and lengths (in days) of all tournaments that were longer than one week. (6 points)
5. To answer the following question, include results from both doubles and singles matches. List the names of all players from the following countries, who played against “Tommy Haas”. (6 points)
    i. Russian,
    ii. Chilean
    iii. US
6. There are 2 parts the question (only include results only for singles tennis) (8 Points)
    a. List the names of all players who have lost to “Roger Federer” in the finals of any tournament,
    b. List the name of the tournament they lost in also in the same table
7. Business Report & detailed Explanation of Inferences for all the Tasks(3 Marks)
    a. You need to submit all the answers to the questions in a sequential manner
    b. Please note the outputs, inferences, Observations and Conclusion should be available both in the SQL workbook and business report.
    c. The Business report should include a detailed explanation of the approach used/Query used, all outputs of the queries, insights/inferences based on outputs etc. The Level of detail and explanation in business report should be deeper with regards to why this query was used for that question
    d. Your report should contain both Queries & output along with inferences