Introduction

Title: Empowering Insights: Using Data Analysis to Create Targeted Insurance Products and Marketing Strategies

Objective: To employ SQL and R programming languages to analyze a dataset, identify significant associations among consumer characteristics, various insurance types, and the effectiveness of diverse marketing techniques in the competitive insurance market.

Table of Contents

1. Introduction and Background

1.1 Background and Objectives of the Problem

In the highly competitive insurance market, managing all data is difficult. Our consultant must consolidate data types from numerous files in the organization’s repository. The firm struggles to find significant data due to its widespread distribution. The primary objective of our study is to employ SQL and R programming languages to analyze a dataset and identify significant associations among consumer characteristics, various insurance types, and the effectiveness of diverse marketing techniques.

2. Database Development

2.1 Microsoft Access

The data analysis of the provided dataset was conducted using SQL, with Microsoft Access being employed as the software tool. Microsoft Access is chosen due to its user-friendly interface, which facilitates the efficient creation of databases. Additionally, it facilitates seamless integration of data management across multiple platforms and offers a cost-effective solution.

2.2 Steps Taken to Create Database

- Created a new database using the create function on Microsoft Access.
- Imported all four Excel sheets as tables to our database using the external data import function.

2.3 Description of Given Data and Creation of Analytical Base Table (ABT)

Customer Data:

- **File Name**: Data_1_Customer.xlsx
- **Variables**: CustomerID, Title, GivenName, MiddleInitial, Surname, CardType, Occupation, Gender, Age, Location, ComChannel, MotorID, HealthID, TravelID

Motor Policies Data:

- **File Name**: Data_2_Motor_Policies.xlsx
- **Variables**: MotorID, PolicyStart, PolicyEnd, MotorType, veh_value, Exposure, Clm, Numclaims, v_body, v_age, LastClaimDate

Health Policies Data:

- **File Name**: Data_3_Health_Policies.xlsx
- **Variables**: HealthID, PolicyStart, PolicyEnd, HealthType, HealthDependentsAdults, DependentsKids

Travel Policies Data:

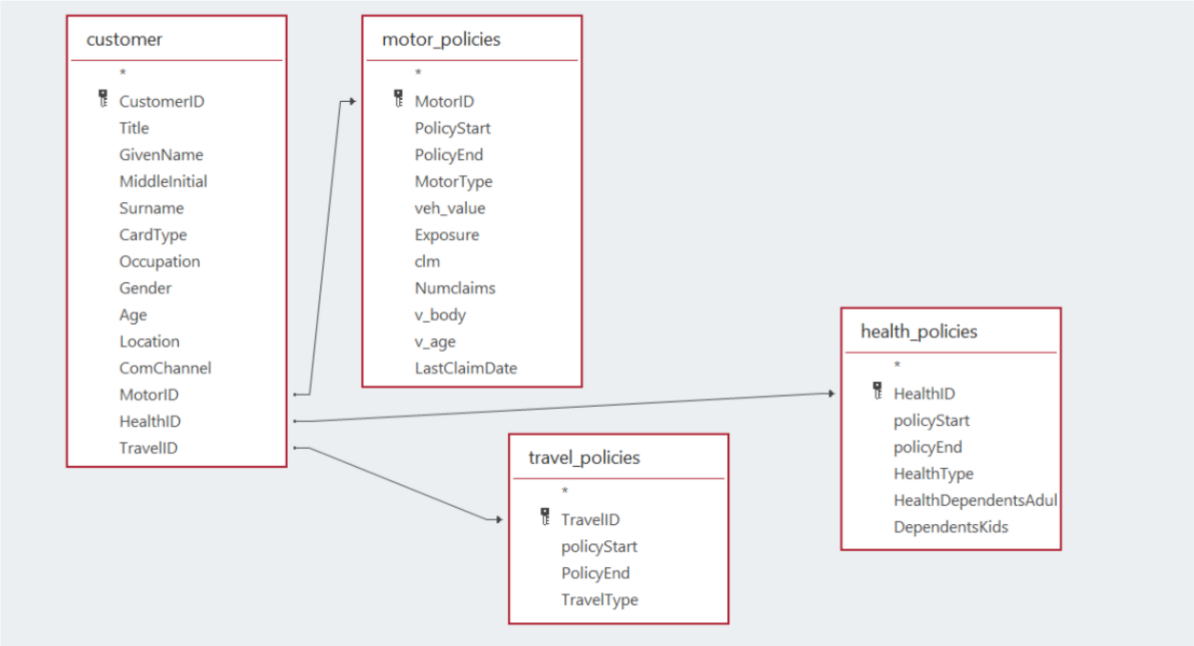
- **File Name**: Data_4_Travel_Policies.xlsx
- **Variables**: TravelID, PolicyStart, PolicyEnd, TravelType

Analytical Base Table (ABT):

- Consolidated data from the four tables using SQL queries.
- Ensured data integrity and offered significant insights into the connections among the data.

2.4 Database Structure

- The primary key of the customer table is “CustomerID,” while the foreign keys are “HealthID,” “MotorID,” and “TravelID.”
- The following figure constitutes the database’s structure:



2.5 Limitations of the Approach and Technologies Used

- Scalability concerns when dealing with extensive datasets.
- Complexity of handling unstructured data.
- Learning curve associated with executing sophisticated queries.
- Possible difficulties in achieving optimal speed in concurrent contexts.

2.6 Overcoming Limitations for the Business

- NoSQL databases or SQL partitioning could be beneficial for scalability.
- Integration of SQL and NoSQL databases and SQL’s JSON/XML capabilities for managing different data formats.

- Incorporation of R into data analytic workflows for improved statistical analysis, data visualization, and machine learning.

3. Data Quality Report (R)

3.1 Steps to Create ABT Table in R

- Imported necessary libraries in R: tidyverse, readxl, dplyr.
- Imported raw data into R and assigned to variables: data1, data2, data3, data4.
- Constructed the ABT using left_join function from dplyr.

3.2 Identification of Data Quality Issues

- Utilized functions: count(), summary(), str(), is.na(), and plotting various box plots to identify data quality issues.

3.3 Implication of Data Quality Issues

- Inconsistent categorical values and inaccurate numerical entries affect analysis and decision-making.

3.4 Approaches Used to Address Data Quality

- Used mutate(), filter(), and replace() functions from R to address data quality issues.
- Created a new variable ABT_clean to reflect the changes.

3.5 Business Strategies for Preventing Data Quality Issues

- Standardized gathering techniques, regular audits, comprehensive employee training.
- Automated validation tools, explicit governance rules, and frequent data maintenance.

4. Insights

4.1 SQL Functions Used for Insights

- **SUM()**: Determines the total number of policies and counts the number of renewed policies.
- **IIF()**: Conditional function to count renewed policies.
- **DATEDIFF()**: Calculates the time interval between PolicyStart and PolicyEnd.
- **COUNT()**: Determines the overall count of policies.
- **GROUP BY**: Aggregates data for analysis based on particular categories.

4.2 Insights Obtained

1. **Retention Rate**:
   - Health insurance plans have higher client retention compared to automobile and travel insurance plans.
2. **Gender, Age, and Different Policies**:
   - Uniform pattern in all age groups and both genders, suggesting low transactions in the travel insurance area.
3. **Location, Gender, and Policy Count**:
   - Urban residents exhibit a higher propensity to embrace policy compared to rural residents.
4. **ComChannel, Location, and Customer Count**:
   - Urban clientele prefer digital communication methods, while rural clients respond faster to phone interactions.

5. **Age Group and Total Claims**:
   - Higher insurance claims in the 56-60 age bracket due to health issues, with lower claims in the retirement age group (60-85).
6. **Age Group, Count of Customer, and AvgClaimPercent**:

- Insurance claims peak in the 56-60 age bracket, with lower claims in the 18-25 age group due to better health and less risky activities.