Vehicle Insurance Claim Fraud database

Aditya Satpute

[asatpute@ucdavis.edu](mailto:asatpute@ucdavis.edu)

Table of contents

|  |  |
| --- | --- |
| **Topic** | **Page no.** |
| Procedure for database creation | 2 |
| Challenges faced during data import | 3 |
| Data dictionary | 4 |
| Business Questions | 5 |

**Procedure for database creation**

The original dataset had 15420 unique rows with 33 columns, which would not have been very useful for analysis. The approach used for breaking the dataset into logical tables was based on

1. Common sense
2. and grouping based on similarity of attributes.

The following steps were under taken after creation of the initial ER diagram:

*Data preparation:*

1. Create new sheet for each table
2. Cut and paste columns in accordance with the ER diagram
3. Understand if the tables could exist as standalone tables or would need a composite primary key
4. Analyze for unique values in each column of the tables. For example, the Vehicle table has a column named category, which lists out the different categories of the vehicle i.e sport, utility etc.
5. If a table could exist as a unique standalone table, the number of categories existing for each row was checked. If it was finite (approx. 20 max categories) and made sense to make combinations with the other columns, the categories were noted.
6. Different combinations of the categories for each column were made programmatically using python.
7. These were used to populate the table.
8. A synthetic/surrogate primary key was provided to these kinds of tables. Example: VehicleID, InvestigationID, AgentID etc.
9. Note: Because the original dataset had 15420 unique rows, with a unique policy number that uniquely identified each customer, the customer table was cut down to 12000 rows to add some complexity to the analyses and a surrogate primary key was added.
10. The above steps would help with the relationship, one customer can have one or many policies.
11. For tables where no grouping was possible, they were left alone with 15420 records.
12. Dates were calculated from the following columns that were provided: WeekOf Month, DayOfWeek, Month, Year which were disjoin in the dataset.
13. Foreign key columns were populated accordingly in excel using VLookup from the different sheets.
14. Data for each table was stored in separate excel sheets

*Data Import:*

1. A schema named VehicleInsuranceFraud was created in MySQL.
2. Using table import wizard, all the tables were created by importing the spreadsheet for each table.

*Table setup and relationships in MySQL:*

1. Primary keys were set by using alter table
2. Null and non-null values were set using alter table.
3. Foreign keys were set according to ERD.
4. Using reverse engineer in MySQL, the relationships were setup logically, following with forward engineer command was implemented.

In the course of this process, an additional table was formed. The ERD was adjusted accordingly.

*The following drive folder has all the table creation commands, ERD created in MySQL backend and the excel sheets for the tables:*

[Click here to go to drive](https://drive.google.com/drive/folders/1vC7sUK2XN_hufIL7FVsEYBPtn9r7hRu4?usp=sharing)

**Challenges faced during data import**

Not many challenges were faced when importing the data using the table import wizard. A few issues that were addressed are as follows:

1. Null value rows were not inserted.

Remedy: Recheck the tables for null values and fix them

1. Data type changes when importing.

Remedy: When importing using import wizard, date is by default set to string. This was changed to datetime format before proceeding to import. Even after changing format, timestamps were added to the date column, which was cleaned up by changing the column type to include only date.

1. AgentID column type change.

Remedy: The AgentID column is an Alpha-numeric column. Hence, it was imported as string. String data types are not allowed as primary key. Hence the when setting AgentID as primary ket, its type was changed to char(10). Similar changes were applied to AgentID in claim table, where it is present as a foreign key.

**Data dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Description** | **Type of variable** | **Sample values** |
| Month | Month of Accident | Categorical | Jan, Feb, Mar, Apr,…,Dec |
| WeekOfMonth | Week number when accident happened | Categorical | 1,2,3,4,5 |
| DayOfWeek | Day of Accident | Categorical | Sunday, Monday, …, Saturday |
| Make | Vehicle brand | Categorical | Toyota, Ford, Mercedes etc. |
| AccidentArea | Area type where accident occurred | Categorical | Urban, Rural |
| DayOfWeekClaimed | Day of Claim generated | Categorical | Sunday, Monday, …, Saturday |
| MonthClaimed | Month of Claim generated | Categorical | Jan, Feb, Mar, Apr,…,Dec |
| WeekOfMonthClaimed | Week number when Claim was genrated | Categorical | 1,2,3,4,5 |
| Sex | Sex | Categorical | Male, Female |
| MaritalStatus | Marital status | Categorical | Married, Single |
| Age | Age of policy holder | Continuous | 25,27,43,.. |
| Fault | Who was at fault when the accident occurred | Categorical | Policy Holder, Third Party |
| PolicyType | Type of policy | Categorical | Sports- Liability, Sedan- Liability… |
| VehicleCategory | Category of vehicle | Categorical | Sport, Utility, Sedan |
| VehiclePrice | Price category range of the vehicle | Categorical | 20000 to 29000, less than 20000,… |
| FraudFound\_P | Was fraud found when claim was investigated. 0- yes, 1- no | Categorical | 0,1 |
| PolicyNumber | Unique policy number | Continuous |  |
| RepNumber | Representative number | Continuous |  |
| Deductible | Deductible for the insurance | Continuous | 300, 400, 500,… |
| DriverRating | Policy holder driver rating | Ordinal | 1,2,3,4,5 |
| Days\_Policy\_Accident | How many days since accident has happened. Given in range of days | Categorical | 1 to 7, 8 to 15,... |
| Days\_Policy\_Claim | Days since claim was requested. Given in range of days | Categorical | 1 to 7, 8 to 15,... |
| PastNumberOfClaims | How many claims has the policy holder generated in the past | Categorical | none, 1, 2 to 4, more than 4 |
| PoliceReportFiled | Was police report filed | Categorical | Yes, No |
| WitnessPresent | Was a witness presnt during the accident | Categorical | Yes, No |
| AgentType | Type of agent | Categorical | External, Internal |
| NumberOfSuppliments | Number of supplimental policies for vehicles | Categorical | Nonw, 1 to 2, … |
| AddressChange\_Claim | Has address changed since claim was made | Categorical | no change, 1 year, .. |
| NumberOfCars | Number of cars insured | Categorical | 1 vehicle, 2 vehicles, .. |
| Year | Year when the accident happened and claim made | Continuous |  |
| BasePolicy | Type of base policy | Categorical | Collision, Liability |

**Business Questions**

1. What are the customer demographics in terms of gender and number of cars owned. Is there a relationship between them?
2. What are the attributes of customers whose claims were found to be fraudulent?
3. Which are the top 5 most popular car makes and models enrolled with our firm for an insurance?
4. Are cars in certain price ranges more represented in our database than the others, i.e are expensive cars more likely to be insured with our than low range ones? Does this vary by car maker?
5. Does marital status determine the type of car and insurance owned?
6. Who are our top 2 and bottom 2 performing agents in terms of the number of policies sold?
7. Is there any seasonality in accidents? Are there certain months when more accidents occur?
8. Can we attribute fraudulent claims to certain vehicle price categories and deductibles?
9. Are sports cars more likely to be involved in accidents?
10. Is a claim more likely to be a fraudulent in the absence of a witness and if police report wasn’t filed?
11. Which are the top 3 policy types and their area of accident in terms or percentages?