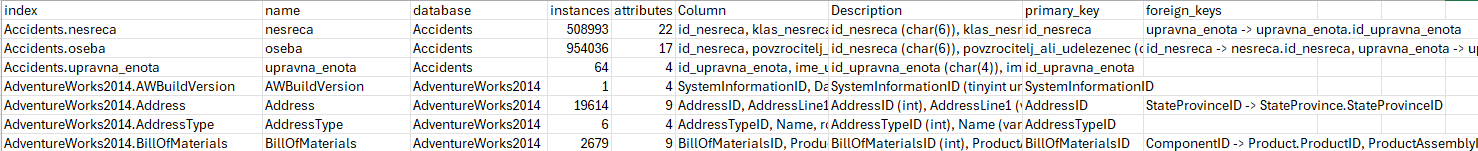
Appendix 2 - Explanation for selection of 50 database tables from the Prague Relational Learning Repository

These are the steps taken to obtain 50 database tables from the Prague Relational Learning Repository.

1. The paper that described the Repository **[Motl & Schulte (2024)]** had this table 1:

|  |  |
| --- | --- |
|  | In it we can see, in the first line, important information that allow us to choose the Databases.  There are 49 Databases described here, some with numbers that point to research papers associated.  There are the #Relations, that show the number of tables in each database.  There are the #Instances, which show the count of rows in the target table in each database.  There is the Size in MB including indexes.  There is the Type, Real or Synth (synthetically generated).  There are many different Domains.  And there are two possible kinds of Tasks, Regr(ession) or Class(ification). |

1. The next step was the download of all information currently available at the Prague Relational Learning Repository. Over 2000 tables.
2. We brought the following information, similar to what was brought earlier from all the 622 datasets of the UCI Machine Learning Repository:



* 1. The index, with the content of the database followed by a dot and the table
  2. The name of the table
  3. The name of the database
  4. The number of Instances/records in this table
  5. The number of attributes/columns in this table
  6. The columns in this table
  7. The description, where the columns contain the format associated
  8. The primary\_key(s) in this table
  9. The foreign\_key(s) in this table

1. From the Table 1 we checked if they existed in the current repository. Unfortunately, some databases do not exist anymore or have different names currently.
2. These are the 3 databases not found currently: PTC, Thrombosis and VOC.
3. Analyzing the content of all different Domains it was discovered that the Kinship Domain is not suitable for our research, because they all their tables only contain few columns with information about names of relatives, such as brother, father, etc.:

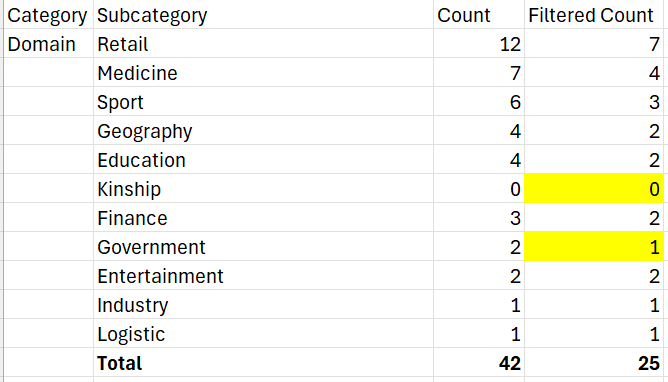
|  |  |  |
| --- | --- | --- |
| Dunur.brother | Dunur.husband | Elti.target |

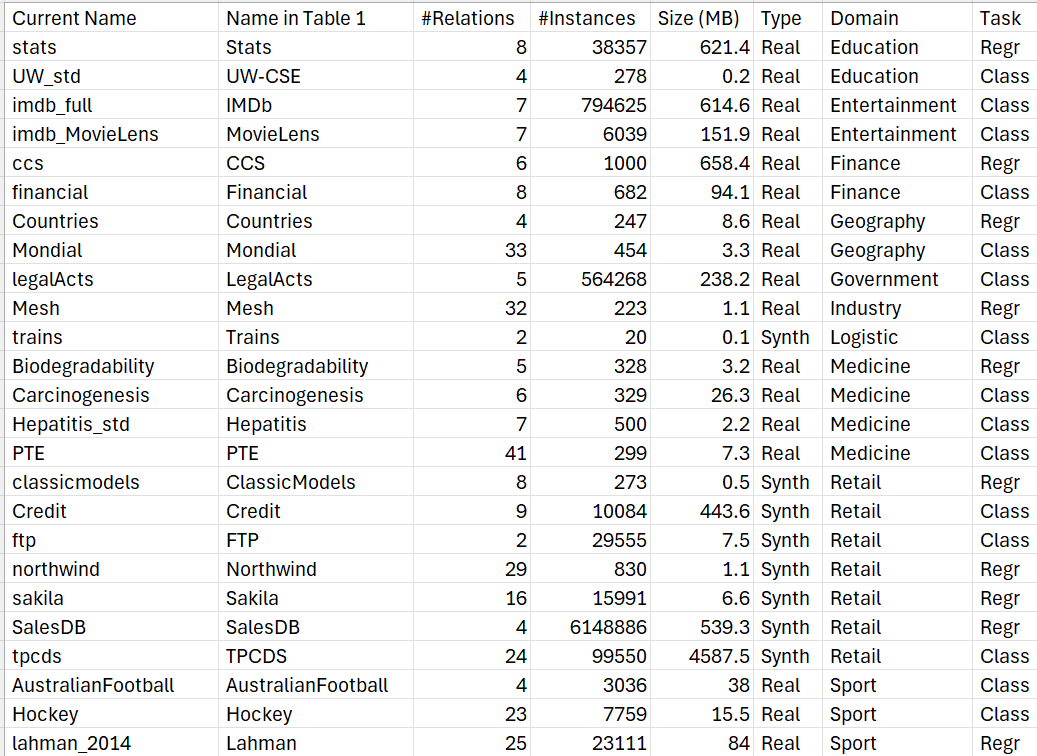
1. So, the initial total of 49 databases became 49 – 3 – 4 (Kinship) = 42 databases.
2. See below the list of the 49 databases, with the Current Name and the Name in Table 1:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Current Name | Name in Table 1 | #Relations | #Instances | Size(MB) | Type | Domain | Task |
| Accidents | Accidents | 4 | 495760 | 210 | Real | Government | Class |
| AdventureWorks2014 | AdventureWorks | 71 | 30669 | 234.6 | Synth | Retail | Regr |
| AustralianFootball | AustralianFootball | 4 | 3036 | 38 | Real | Sport | Class |
| Biodegradability | Biodegradability | 5 | 328 | 3.2 | Real | Medicine | Regr |
| Carcinogenesis | Carcinogenesis | 6 | 329 | 26.3 | Real | Medicine | Class |
| ccs | CCS | 6 | 1000 | 658.4 | Real | Finance | Regr |
| classicmodels | ClassicModels | 8 | 273 | 0.5 | Synth | Retail | Regr |
| Countries | Countries | 4 | 247 | 8.6 | Real | Geography | Regr |
| Credit | Credit | 9 | 10084 | 443.6 | Synth | Retail | Class |
| cs | CS | 8 | 100 | 0.3 | Synth | Finance | Class |
| Dunur | Dunur | 20 | 276 | 0.8 | Real | Kinship | Class |
| Elti | Elti | 14 | 1081 | 0.7 | Real | Kinship | Class |
| employee | Employee | 7 | 2838426 | 344.6 | Synth | Retail | Regr |
| financial | Financial | 8 | 682 | 94.1 | Real | Finance | Class |
| ftp | FTP | 2 | 29555 | 7.5 | Synth | Retail | Class |
| genes | Genes | 3 | 862 | 1.9 | Real | Medicine | Class |
| Hepatitis\_std | Hepatitis | 7 | 500 | 2.2 | Real | Medicine | Class |
| Hockey | Hockey | 23 | 7759 | 15.5 | Real | Sport | Class |
| imdb\_full | IMDb | 7 | 794625 | 614.6 | Real | Entertainment | Class |
| imdb\_MovieLens | MovieLens | 7 | 6039 | 151.9 | Real | Entertainment | Class |
| lahman\_2014 | Lahman | 25 | 23111 | 84 | Real | Sport | Regr |
| legalActs | LegalActs | 5 | 564268 | 238.2 | Real | Government | Class |
| Mesh | Mesh | 32 | 223 | 1.1 | Real | Industry | Regr |
| Mondial | Mondial | 33 | 454 | 3.3 | Real | Geography | Class |
| MooneyFamily | MooneyFamily | 72 | 92 | 3.3 | Synth | Kinship | Class |
| mutagenesis | Mutagenesis | 3 | 188 | 0.9 | Real | Medicine | Class |
| nations | Nations | 3 | 14 | 2.1 | Real | Geography | Class |
| NBA | NBA | 4 | 30 | 0.3 | Real | Sport | Class |
| NCAA | NCAA | 10 | 268 | 40.6 | Real | Sport | Class |
| northwind | Northwind | 29 | 830 | 1.1 | Synth | Retail | Regr |
| Pima | Pima | 14 | 768 | 0.8 | Real | Medicine | Class |
| PremierLeague | PremiereLeague | 4 | 363 | 11.3 | Real | Sport | Class |
| Database not found | PTC | 4 | 343 | 7.8 | Real | Medicine | Class |
| PTE | PTE | 41 | 299 | 7.3 | Real | Medicine | Class |
| pubs | Pubs | 11 | 18 | 0.4 | Synth | Retail | Regr |
| sakila | Sakila | 16 | 15991 | 6.6 | Synth | Retail | Regr |
| SalesDB | SalesDB | 4 | 6148886 | 539.3 | Synth | Retail | Regr |
| Same\_Gen | SameGen | 7 | 1081 | 0.3 | Real | Kinship | Class |
| stats | Stats | 8 | 38357 | 621.4 | Real | Education | Regr |
| Student\_loan | StudentLoan | 13 | 1000 | 0.9 | Real | Education | Class |
| Database not found | Thrombosis | 3 | 806 | 1.9 | Real | Medicine | Class |
| tpcc | TPCC | 9 | 28433 | 174.1 | Synth | Retail | Class |
| tpcds | TPCDS | 24 | 99550 | 4587.5 | Synth | Retail | Class |
| tpch | TPCH | 8 | 148255 | 1925.1 | Synth | Retail | Regr |
| trains | Trains | 2 | 20 | 0.1 | Synth | Logistic | Class |
| university | University | 5 | 38 | 0.3 | Synth | Education | Class |
| UW\_std | UW-CSE | 4 | 278 | 0.2 | Real | Education | Class |
| Database not found | VOC | 8 | 8215 | 2.7 | Real | Logistic | Class |
| world | World | 3 | 239 | 0.8 | Real | Geography | Class |

1. An analysis was made, and this was the distribution of the databases: A screenshot of a data sheet

   Description automatically generated
2. The next step was filtering to 25 databases. It was found out that the Accidents' Database tables had Czech words in titles, not English words, which means that our code could not check the words in them. So, 1 more database was cancelled. It is from the Government Domain. A proportion was made regarding the Count of Domain databases:



1. This was the list of the 25 databases chosen, by alphabetical order of Domain:
2. This list of 25 databases contained 331 tables from which 50 were chosen.
3. The final definition was based in analysis of the content, to choose interesting tables from the maximum amount of the 25 databases.
4. Some databases did not contain interesting information for attribute analysis, such as Carcinogenesis and Mesh, and were discarded.
5. Another decision was made regarding the number of Foreign Keys. In a future step of our research, we will analyse the relationship between different tables, and the fact that there are tables with many FKs will enable that, and this was an important decision taken.
6. This is the final definition of the 50 tables:

A close-up of a document

Description automatically generated 

This final definition is in file 50\_Database\_Tables\_from\_Prague\_Repository.xlsx [24]

This is the analysis of the 50 final tables:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Current Name | Number | | sakila | 10 | | Mondial | 5 | | northwind | 5 | | lahman\_2014 | 4 | | imdb\_full | 3 | | imdb\_MovieLens | 3 | | AustralianFootball | 2 | | PTE | 2 | | tpcds | 2 | | Biodegradability | 1 | | ccs | 1 | | classicmodels | 1 | | Countries | 1 | | Credit | 1 | | financial | 1 | | ftp | 1 | | Hepatitis\_std | 1 | | Hockey | 1 | | legalActs | 1 | | SalesDB | 1 | | stats | 1 | | trains | 1 | | UW\_std | 1 | | Carcinogenesis | 0 | | Mesh | 0 | | **Total** | **50** |   **Database Distribution Analysis** | Total number of databases: 23  **Area Distribution Analysis**   * Retail: 21 tables * Sport: 7 tables * Entertainment: 6 tables * Geography: 6 tables * Medicine: 4 tables * Education: 2 tables * Finance: 2 tables * Government: 1 table * Logistic: 1 table   Total number of areas: 9  **Table Size Distribution Analysis**   * Small (0-1,000 instances): 23 tables * Medium (1,001-100,000 instances): 20 tables * Large (100,001+ instances): 7 tables   **Notable large tables:**   1. Credit.charge: 1,600,000 instances 2. financial.trans: 1,056,320 instances 3. SalesDB.Sales: 6,715,221 instances 4. tpcds.store\_sales: 2,880,404 instances |

**Primary Key Analysis**

Total tables: 50 Tables

Tables with defined Primary Keys: 47 Tables (39 use single-column primary keys and 8 use composite keys).

3 Tables without defined Primary Keys:

1. Hockey.Master
2. northwind.Invoices
3. northwind.Order Details Extended

**Composite Key Analysis**

Total tables with composite keys: 8

Tables with composite keys:

1. imdb\_full.movies2actors (movieid, actorid)
2. imdb\_MovieLens.u2base (userid, movieid)
3. lahman\_2014.halloffame (hofID, yearID)
4. lahman\_2014.salaries (yearID, teamID, lgID, playerID)
5. lahman\_2014.teams (yearID, lgID, teamID)
6. Mondial.city (Name, Province)
7. Mondial.province (Name, Country)
8. tpcds.store\_sales (ss\_item\_sk, ss\_ticket\_number)

* The lahman\_2014 database uses composite keys most frequently, with 3 out of its 4 tables having composite keys.
* The Mondial database uses composite keys for geographic entities, likely to handle cases where names might be repeated across different regions.
* The imdb databases use composite keys for relationship tables, which is a common practice in many-to-many relationships.

**Foreign Keys analysis**

Total tables: 50 Tables

Tables with defined Foreign Keys: 26 Tables

|  |  |
| --- | --- |
|  | Observe that the databases with more tables chosen contain FKs, such as sakila, Mondial, and northwind, allowing future FK analysis: |