DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Project Description

CS-322 Introduction to Database Systems
Spring 2021

Table of Contents

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Introduction

In this project you will get a set of files collected by a real-world entity, based on which you need to i) design the database schema and implement the relational schema, ii) parse, clean, and load the data into a DBMS, iii) write queries, and, finally, iv) evaluate and optimize queries with index structures/query plan analysis in order to analyze the performance impact on generated query plans and discuss about the query optimizer decisions on querying the given dataset.

Therefore, the goal is to guide you through the design process from getting the unstructured raw data that needs organization, abstract reasoning about the entities and relations that exist, parsing and preparing the data for loading using the programming tools of your choice, to the point where this data is ready to be queried using a relational DBMS. This project simulates a business use-case and synthesizes your programming and analysis skills in a practical task with a concrete end goal, along with practicing and implementing the theoretical principles acquired in this class.

IMPORTANT: Read the whole document before starting doing any work.

Project summary

The dataset contains a subset of data from the California Highway Patrol collected from the Statewide Integrated Traffic Records System (SWITRS), covering traffic collisions in the state of California in 2018. These reports have been collected by the Highway Patrol, recorded electronically for archival and preservation based on the forms filed by officers. This data has an immense value for urban planners and scientists that may want to analyze risks and how to improve the traffic situation and which factors or locations may be prevalent. For example, data scientists may want to obtain an answer if during some holidays there are more or fewer collisions, or which weather conditions may significantly increase the risk of an accident. Insurance companies could find significant value in this data to analyze and optimize their policies, and would want to use this data too. While the reports have been digitized and provided in CSV files, these files are not ready to be queried – some databases may support using the CSV files directly or using spreadsheet software, but this process is error-prone, cumbersome, and does not scale to data that providers have promised to span over more years, which would yield gigabytes of data over the past decades. You are given the data subset for 2018 to test and implement the proof of concept.

Suppose you have been hired to take these CSV files and produce a database that will enable analysts to get answers in a very fast way – they have also provided you with a list of their queries that they would like to receive answers very quickly. The TAs are your quality control managers, and are there to provide feedback and make sure your final deliverable is up to standards and that the tentative clients are happy with the performance of their database, as the analysts always want their answers immediately and with the minimal delay and latency, whatever the size of the data may be, and make it future-proof and scalable by using a relational DBMS.

The project is done in teams of 3 people. The project is separated into 3 milestones, which follow the material taught in the lectures. We have synchronized each milestone with the material of the lectures for your convenience.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

The first milestone requires you to analyze the dataset and extract the ER (Entity-Relationship) model, translating it to relational schema, as well as getting acquainted with a DBMS.

The second milestone requires you to express a set of queries on top of the loaded database. The goal of this part is to familiarize with data loading and the task of data processing and eventual cleaning (meaning transforming data if needed). You will also get to apply your SQL skills, and get a first intuition about how query performance is directly dependent on i) the way you formulate a query and ii) the logical and physical design of your database.

Finally, in the third part of the project you will express a set of more sophisticated SQL queries, which you will also analyze to come up with a detailed description of the execution and propose an improvement. You will analyze the queries and their respective query plans in order to optimize the execution, either with building appropriate index structures or rewriting the queries to make the execution more efficient (or both), and discuss about the decisions that query optimizer took, such as if it even considered the newly created index structure.

For each of these milestones the you should prepare a document following the provided template which describes the completed work. The grading will be done based on the final report as well on a presentation and short discussion and Q&A with the TAs. The final report should contain material about all the work done for the 3 milestones combined into one document, where the project deliverable template is included.

The reports for the first two milestones are optional, while the final (3rd) deliverable is mandatory and graded. However, only the teams that submit the intermediate reports will get feedback on their progress, and we strongly advise you to do it so you can benefit from our feedback – the goal is to guide you through this process and help break down this project into manageable chunks.

IMPORTANT: Only the teams that deliver the intermediate milestone deliverables within the deadline will receive feedback! You do not need to use the exact template document (.docx or .odt), however your deliverable must include all the elements of the provided report. You are free to add sections or elements if needed.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Deliverable 1: Create the ER model, Design & Create the Schema

Deadline (to get feedback): 29/03/2021

You have received the following files:

- collisions2018.csv
- parties2018.csv
- victims2018.csv

The goal of this deliverable is to design an ER model and a corresponding relational schema, and create the database tables in a database system (using SQL DDL). The organization of the data in files and the given description ***DOES NOT IMPLY*** an ER model or a relational schema (e.g. that these are the only 3 entities of the ER model). The provided CSV files are simply the way it is the most convenient to collect the data, however you need to reason about what are the entities, relations, and how can you logically organize the data into self-sufficient actors in the business use-case. You need to discuss about necessary constraints (key, foreign key constraints, nullable values, and others), and have an understanding why certain design choices remove repeated or redundant data points/attributes. This material is covered in the first weeks of the course, and will allow you to start on time to analyze and provide a first version of your model.

In the 1st deliverable you should:

1. Create an ER model for the provided data. For your ease, you may provide a relational translation of the ER model, that will help you with the next point.
2. Design the database and the constraints needed to maintain the database consistent.
3. Provide the SQL DDL commands to create the tables and the constraints in a relational database system.
4. Describe their work in the form of a report which should contain an ER diagram, SQL DDL code for table creation, description of the data constraints, and justification of the design choices (in a few paragraphs). The report should be submitted as a single PDF file (**one PDF document per group**).

Important Note: Before designing the ER model, understand the data and read carefully the notes given in the form of **FAQ** at the end of the project description, as well as the detailed data description. If you need any clarifications, ask the TAs during the project session or office hours, or ask on Moodle forum.

Tip: Analyze the data and keep in mind that a column in CSV file does not always map to a column in entity/table. Remember that the column values have to be atomic (not a list) in relational model (1st Normal Form). Some data columns may become separate tables for this reason. Feel free to group some values/attributes into a separate entity/table if they seem to **repeat** or appear to be logically a separate entity (explain your assumptions over the data and design decision). For example, a frequent design choice is a *star schema* – where attribute groups become entities/tables called *dimension* tables, while *fact* tables refer to them via foreign keys. Think first of the entities and relations based on attributes and their meaning, from the high-level perspective.

Points breakdown for elements of this deliverable: 18 points for the ER model, 4 points for DDL + constraints.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Deliverable 2: Import the Data. First part of SQL queries

Deadline (to get feedback): 03/05/2021

In this phase, you have to import the provided raw CSV data into the database. You will need to process and split or reorganize the data into different CSV files that represent your entities with all the necessary attributes, and eventually new keys and foreign keys, that are ready to be imported to the DBMS. You should know how to insert/delete/update data via SQL DML commands, as well as to execute exploratory queries over the data.

You have to implement the following queries in SQL:

1. List the year and the number of collisions per year. Suppose there are more years than just 2018.
2. Find the most popular *vehicle make* in the database. Also list the number of vehicles of that particular make.
3. Find the fraction of total collisions that happened under *dark* lighting conditions.
4. Find the number of collisions that have occurred under snowy weather conditions.
5. Compute the number of collisions per day of the week, and find the day that witnessed the highest number of collisions. List the day along with the number of collisions.
6. List all weather types and their corresponding number of collisions in descending order of the number of collisions.
7. Find the number of at-fault collision parties with *financial responsibility* and *loose material* road conditions.
8. Find the median victim age and the most common victim seating position.
9. What is the fraction of all participants that have been victims of collisions while using a belt?
10. Compute and the fraction of the collisions happening for each hour of the day (for example, *x% at 13*, where 13 means period from 13:00 to 13:59). Display the ratio as percentage for all the hours of the day.

In summary, in the 2nd deliverable you should:

1. Parse the given data and import them in the created database as described in your 1st deliverable.
2. Implement (using SQL) the queries described above.
    a. Provide the SQL code as well as the first 20 rows (when applicable) of the result for each query.
    b. Note: Consider the use of indexes to accelerate long-running queries, useful for the next part.
3. Extend the project report from the first deliverable with the description of the work done for the second deliverable and an explanation for the design choices. Include any changes to the design covered in the first deliverable, with justification of the changes. The report should be submitted as a single PDF file.

Points breakdown for elements of this deliverable: 15 points for the queries, 8 points for data cleaning and loading to the DBMS.

IMPORTANT: You can use any RDBMS, you are not limited to Oracle. In that case you must set up the database locally on your device(s), and if needed we will ask you to run certain queries to make sure you have a working database with at least one team member, in order to avoid plagiarism or reports without functioning DBMS.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Deliverable 3: Interesting and insightful SQL queries

Deadline (you must submit the report to get graded): 31/05/2021

A series of more interesting queries should be implemented with SQL. In addition, the performance of **any 5 queries** should be optimized and analyzed in depth by using indexes and evaluated based on the produced query plans and their cost – compare the cost and plans before and after the optimization to justify the difference.

The queries to be implemented are:

1. For the drivers of age groups: underage (less and equal to 18 years), young I [19, 21], young II [22,24], adult [24,60], elder I [61,64], elder II [65 and over), find the ratio of cases where the driver was the party at fault. Show this ratio as percentage and display it for every age group – if you were an insurance company, based on the results would you change your policies?
2. Find the top-5 *vehicle types* based on the number of collisions on roads with holes. List both the *vehicle type* and their corresponding number of collisions.
3. Find the top-10 *vehicle makes* based on the number of victims who suffered either a *severe injury* or were *killed*. List both the *vehicle make* and their corresponding number of victims.
4. Compute the *safety index* of each seating position as the fraction of total incidents where the victim suffered *no injury*. The position with the highest *safety index* is the safest, while the one with the lowest is the most unsafe. List the most safe and unsafe *victim seating position* along with its *safety index*.
5. How many vehicle types have participated in at least 10 collisions in at least half of the cities?
6. For each of the top-3 most populated cities, show the *city location*, *population*, and the *bottom-10 collisions* in terms of average victim age (show collision id and average victim age).
7. Find all collisions that satisfy the following: the collision was of type *pedestrian* and all victims were above 100 years old. For each of the qualifying collisions, show the *collision id* and the *age* of the eldest collision victim.
8. Find the vehicles that have participated in at least 10 collisions. Show the *vehicle id* and *number of collisions* the vehicle has participated in, sorted according to number of collisions (descending order). What do you observe?
9. Find the top-10 (with respect to *number of collisions*) cities. For each of these cities, show the city location and number of collisions.
10. Are there more accidents around dawn, dusk, during the day, or during the night? In case lighting information is not available, assume the following: the dawn is between 06:00 and 07:59, and dusk between 18:00 and 19:59 in the period September 1 - March 31; and dawn between 04:00 and 06:00, and dusk between 20:00 and 21:59 in the period April 1 - August 31. The remaining corresponding times are night and day. Display the number of accidents, and to which group it belongs, and make your conclusion based on absolute number of accidents in the given 4 periods.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

In total, in the 3rd deliverable you should:

1. Accommodate all above queries by giving the corresponding SQL code.
2. Select 5 queries from Deliverable 3, and accelerate them by using indexes. Explain the necessities of indexes based on the queries and the query plans that you can find from the system (you are free to select any 5 queries you like from the queries of this deliverable).
3. After the introduced optimizations, report the runtime of all queries in (milli)seconds and explain the distribution of the cost (based again on the plans) for the 5 queries selected in part 2, as well as the discussion based on the cost of the query plan – and how this plan has changed and why.
4. Present the results of the queries.
6. Complete the project report written for the previous deliverables by adding description of the queries, explanation for the design choices, analysis of the chosen queries, as well as the changes compared to the work described in the previous deliverables. The report should be submitted as a single PDF file.

Points breakdown for elements of this deliverable: 40 points for the queries, 15 points for optimization.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

California Traffic Collision data description

In this section, we present the dataset, the meaning of the columns/attributes, as well as the eventual mappings from a code to a human-readable value. Read carefully the data description, the FAQ and if in doubt ask the TAs for clarification. The data is stored in three CSV (comma separated values) files.

parties2018.csv

This file contains information about the groups of people (parties) that were involved in the accident. The mappings may be already applied, or partially applied due to varying data collection mechanism.

Column Name	Description
at_fault	Indicates whether the party was at fault in the collision
case_id	The unique identifier of the collision report
cellphone_use	B - Cell Phone in Use C - Cell Phone Not in Use D - No Cell Phone/Unknown Blank or - - Not Stated
financial_responsibility	N - No Proof of Insurance Obtained Y - Yes, Proof of Insurance Obtained O - Not Applicable (used for parked cars, bicyclists, pedestrians, and party type others) E - Used if the officer is called away from the scene of the collision prior to obtaining the insurance information Blank or - - not stated
hazardous_materials	A - Hazardous Materials Blank or - - Not Stated
id	The identifier of the party
movement_preceding_collision	A - Stopped B - Proceeding Straight C - Ran Off Road D - Making Right Turn E - Making Left Turn F - Making U-Turn G - Backing H - Slowing/Stopping I - Passing Other Vehicle J - Changing Lanes K - Parking Maneuver L - Entering Traffic M - Other Unsafe Turning N - Crossed Into Opposing Lane

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	O - Parked P - Merging Q - Traveling Wrong Way R - Other Blank - Not Stated
other_associated_factor_1	A - Violation E - Vision Obscurements F - Inattention G - Stop and Go Traffic H - Entering/Leaving Ramp I - Previous Collision J - Unfamiliar With Road K - Defective Vehicle Equipment L - Uninvolved Vehicle M - Other N - None Apparent O - Runaway Vehicle P - Inattention, Cell Phone Q - Inattention, Electronic Equip. R - Inattention, Radio/CD S - Inattention, Smoking T - Inattention, Eating U - Inattention, Children V - Inattention, Animal W - Inattention, Personal Hygiene X - Inattention, Reading Y - Inattention, Other Blank or - - Not Stated
other_associated_factor_2	The same as other_associated_factor_1
party_age	The age of the party at the time of the collision
party_drug_physical	E - Under Drug Influence F - Impairment - Physical H - Not Applicable I - Sleepy/Fatigued Blank or - - Not Stated
party_number	The number associated to the party in the particular case
party_safety_equipment_1	A - None in Vehicle B - Unknown C - Lap Belt Used D - Lap Belt Not Used E - Shoulder Harness Used F - Shoulder Harness Not Used

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	G - Lap/Shoulder Harness Used H - Lap/Shoulder Harness Not Used J - Passive Restraint Used K - Passive Restraint Not Used L - Air Bag Deployed M - Air Bag Not Deployed N - Other P - Not Required Q - Child Restraint in Vehicle Used R - Child Restraint in Vehicle Not Used S - Child Restraint in Vehicle, Use Unknown T - Child Restraint in Vehicle, Improper Use U - No Child Restraint in Vehicle V - Driver, Motorcycle Helmet Not Used W - Driver, Motorcycle Helmet Used X - Passenger, Motorcycle Helmet Not Used Y - Passenger, Motorcycle Helmet Used Blank or - - Not Stated
party_safety_equipment_2	The same as party_safety_equipment_1
party_sex	M - Male F - Female Blank or - - Not Stated
party_sobriety	A - Had Not Been Drinking B - Had Been Drinking, Under Influence C - Had Been Drinking, Not Under Influence D - Had Been Drinking, Impairment Unknown G - Impairment Unknown H - Not Applicable Blank or - - Not Stated
party_type	1 - Driver (including Hit and Run) 2 - Pedestrian 3 - Parked Vehicle 4 - Bicyclist 5 - Other Blank or - - Not Stated
school_bus_related	E - School Bus Related Blank or - - Not Stated
statewide_vehicle_type	A - Passenger Car/Station Wagon B - Passenger Car with Trailer C - Motorcycle/Scooter D - Pickup or Panel Truck E - Pickup or Panel Truck with Trailer

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	F - Truck or Truck Tractor G - Truck or Truck Tractor with Trailer H - Schoolbus I - Other Bus J - Emergency Vehicle K - Highway Construction Equipment L - Bicycle M - Other Vehicle N - Pedestrian O - Moped Blank or - - Not Stated
vehicle_make	The full name of the vehicle make of the party involved
vehicle_year	The model year of the vehicle of the party involved

Tip: As this is a CSV, meaning that the values are split by commas (“,”), and new line (“\n”) separates different rows, make sure to properly handle the **Text** field, as it may contain some of these values between its quotes!

Tip: When there are more than 2 values in the possible value set (for example, something that is natural to be represented as a Boolean or an atomic value such as an integer or double), due to 1st Normal Form and value constraints it is common for some of these attributes to become dimensions that are separate entities. For example, this enables you to subsequently add new categories of attributes, or for example have a localization (display the text in different language, simply by knowing which key corresponds to the text in a given language).

Tip: You are free to add synthetic keys to your data (for example, add a column which represents the key), if needed.

Tip: When you see entities which may appear multiple times for example (_1, _2), what kind of relation cardinality may this be? Try to generalize this, what if there are _3, and subsequent requirement for more attributes for these attributes, such as lists of such items.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

collisions2018.csv

This file contains information about the collision, where it happened, and the vehicles involved. The mappings may be already applied, or partially applied due to varying data collection mechanism.

Column Name	Description
case_id	The identifier of the collision report
collision_date	The date of the collision
collision_severity	1 - Fatal 2 - Injury (Severe) 3 - Injury (Other Visible) 4 - Injury (Complaint of Pain) 0 - PDO (Property Damage Only)
collision_time	The time when the collision occurred (24-hour time)
county_city_location	The location code of where the collision occurred
hit_and_run	F - Felony M - Misdemeanor N - Not Hit and Run
jurisdiction	Four digits assigned by the DOJ (Department of Justice)
lighting	A - Daylight B - Dusk - Dawn C - Dark - Street Lights D - Dark - No Street Lights E - Dark - Street Lights Not Functioning Blank or - - Not Stated
location_type	H - Highway I - Intersection R - Ramp (or Collector) Blank or - - Not State Highway
officer_id	The ID of the officer attending the collision
pcf_violation	Primary collision factor numerical code
pcf_violation_category	01 - Driving or Bicycling Under the Influence of Alcohol or Drug 02 - Impeding Traffic 03 - Unsafe Speed 04 - Following Too Closely 05 - Wrong Side of Road 06 - Improper Passing 07 - Unsafe Lane Change 08 - Improper Turning 09 - Automobile Right of Way 10 - Pedestrian Right of Way 11 - Pedestrian Violation

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	12 - Traffic Signals and Signs 13 - Hazardous Parking 14 - Lights 15 - Brakes 16 - Other Equipment 17 - Other Hazardous Violation 18 - Other Than Driver (or Pedestrian) 19, 20, 21 - Unsafe Starting or Backing 22 - Other Improper Driving 23 - Pedestrian or "Other" Under the Influence of Alcohol or Drug 24 - Fell Asleep 00 - Unknown Blank or - - Not Stated
pcf_violation_subsection	Blank if no subsection
population	1 - Incorporated (less than 2500) 2 - Incorporated (2500 - 10000) 3 - Incorporated (10000 - 25000) 4 - Incorporated (25000 - 50000) 5 - Incorporated (50000 - 100000) 6 - Incorporated (100000 - 250000) 7 - Incorporated (over 250000) 9 - Unincorporated (Rural) 0 - University (Private Property) Blank or - - Not Stated
primary_collision_factor (PCF)	A - (Vehicle) Code Violation B - Other Improper Driving C - Other Than Driver D - Unknown E - Fell Asleep Blank or - - Not Stated
process_date	The date the collision case was processed
ramp_intersection	1 - Ramp Exit, Last 50 Feet 2 - Mid-Ramp 3 - Ramp Entry, First 50 Feet 4 - Not State Highway, Ramp-related, Within 100 Feet 5 - Intersection 6 - Not State Highway, Intersection-related, Within 250 Feet 7 - Highway 8 - Not State Highway Blank or - - Not Stated
road_condition_1	A - Holes, Deep Ruts B - Loose Material on Roadway

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	C - Obstruction on Roadway D - Construction or Repair Zone E - Reduced Roadway Width F - Flooded G - Other H - No Unusual Condition Blank or - - Not Stated
road_condition_2	The same as road_condition_1
road_surface	A - Dry B - Wet C - Snowy or Icy D - Slippery (Muddy, Oily, etc.) Blank or - - Not Stated
tow_away	Y - Yes N - No
type_of_collision	A - Head-On B - Sideswipe C - Rear End D - Broadside E - Hit Object F - Overturned G - Vehicle/Pedestrian H - Other Blank or - - Not Stated
weather_1	A - Clear B - Cloudy C - Raining D - Snowing E - Fog F - Other G - Wind Blank or - - Not Stated
weather_2	The same as weather_1

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

victims2018.csv

This file contains information about the specific people involved in the collisions (victims) and their injuries. The mappings may be already applied, or partially applied due to varying data collection mechanism.

Column Name	Description
case_id	The identifier of the collision report
id	The victim ID
party_number	The number associated to the party in the particular case
victim_age	0 – 125 998 – Not Stated 999 – Pregnancy
victim_degree_of_injury	1 - Killed 2 - Severe Injury 3 - Other Visible Injury 4 - Complaint of Pain 5 – Suspected Serious Injury 6 – Suspected Minor Injury 7 – Possible Injury 0 - No Injury
victim_ejected	0 - Not Ejected 1 - Fully Ejected 2 - Partially Ejected 3 - Unknown Blank or - - Not Stated
victim_role	1 - Driver 2 - Passenger (includes non-operator on bicycle or any victim on/in parked vehicle or multiple victims on/in non-motor vehicle) 3 - Pedestrian 4 - Bicyclist 5 - Other (single victim on/in non-motor vehicle; e.g. ridden animal, horse-drawn carriage, train, or building) 6 - Non-Injured Party
victim_safety_equipment_1	A - None in Vehicle B - Unknown C - Lap Belt Used D - Lap Belt Not Used E - Shoulder Harness Used F - Shoulder Harness Not Used G - Lap/Shoulder Harness Used H - Lap/Shoulder Harness Not Used J - Passive Restraint Used

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

	K - Passive Restraint Not Used L - Air Bag Deployed M - Air Bag Not Deployed N - Other P - Not Required Q - Child Restraint in Vehicle Used R - Child Restraint in Vehicle Not Used S - Child Restraint in Vehicle, Use Unknown T - Child Restraint in Vehicle, Improper Use U - No Child Restraint in Vehicle V - Driver, Motorcycle Helmet Not Used W - Driver, Motorcycle Helmet Used X - Passenger, Motorcycle Helmet Not Used Y - Passenger, Motorcycle Helmet Used Blank or - - Not Stated
victim_safety_equipment_2	The same as victim_safety_equipment_1
victim_seating_position	1 - Driver 2 thru 6 - Passengers 7 - Station Wagon Rear 8 - Rear Occupant of Truck or Van 9 - Position Unknown 0 - Other Occupants A thru Z - Bus Occupants Blank or - - Not Stated
victim_sex	M - Male F - Female Blank or - - Not Stated

You can find the dataset here (.zip file, ~191MB):
https://drive.switch.ch/index.php/s/VKsJDwreSk6QITN

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Frequently Asked Questions

How does one browse the data?

The dataset size is substantial, so it is hard to open most files using a notepad or text editor.
Applications such as Notepad++ and Sublime Text do a better job, but may still have issues with bigger files.
We thus also propose using Unix commands such as:

1. head: prints the first 50 lines of the file
2. less: allows backward movement in the file as well as forward movement
3. vi text editor: this editor does not open the whole file but only the part that is displayed

A useful and recommended method is to browse the data using scripting languages such as Python, where you can use Pandas library to load the CSV as a DataFrame, and explore parts of data via the functions of the library. This way it is also useful to explore the data for future data cleaning, transformation, and loading to DBMS, and the library also provides a method to explore basic statistics and features of the data.

Which is the format of the given data?

The given data is CSV files (Comma Separated Values) which are values separated with comma (,). Each column represents a specific attribute. Usually in CSV files the name of the attribute is given in the first line of the file.

Why are the datasets "dirty"?

Real-world data is almost always dirty; missing values are commonplace; users abuse DBMS datatypes and store values based on their arbitrary, ad-hoc rules. We consider data cleaning to be a part of your project. Regarding how to perform data cleaning, there is more than one correct solution. Some possible ways are the following:

- Use your favorite scripting language, or a typical program that handles data inconsistencies. For example, Python, Java, and Scala all feature CSV parsers which you can use to read and transform the data.
    - One proposed way is to use libraries such as Pandas for Python for fast data manipulation, then you can use the same library to save the DataFrame structure to CSV suitable for DBMS loading.
- Use Unix commands such as sed, grep, awk.
- Load the data in a DBMS and then use DBMS functions to transform them based on your requirements.

Which database system should I use?

You are free to use any DBMS you want. Typical open-source examples are MySQL and PostgreSQL. We will also grant you access to an Oracle installation located on a server at EPFL, which you can use. We will do our best to troubleshoot any issues with the Oracle server and help with issues related to your own installations. To access the Oracle database with the group accounts we provided to you (we will communicate this separately), you need to connect via EPFL network, therefore you need to use EPFL VPN. You can use any system/frontend that allows JDBC connection and file upload to the database backend (Oracle SQLDeveloper, JetBrains DataGrip, …).

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Which character encoding should I set?

All files use UTF-8 encoding. Take care of initializing your database using the correct encoding before creating tables or loading the data.

What should I do if it takes too long to load the data?

The two most common reasons for a slow data loading process are the following:

1. Defining too many indexes/foreign key relations in your tables can delay loading significantly. **We therefore propose that you first create simple tables with only primary key properties, or without any constraints specified at all**. Once data is loaded, add the more complex table relations and indexes.
2. If you are using the database system provided by us, make sure that you are connected to the EPFL network via VPN, it may take a long time to upload the data files, thus leading to longer loading times.

An additional scenario is that you have set up your own database server (e.g., PostgreSQL or MySQL), and that the default resources allocated to it are very few. In that case, some useful links are the following:

- https://dev.mysql.com/doc/refman/5.5/en/innodb-buffer-pool.html
- http://www.rathishkumar.in/2017/01/how-to-allocate-innodb-buffer-pool-size-in-mysql.html
- https://wiki.postgresql.org/wiki/Tuning_Your_PostgreSQL_Server

What should I pay attention to?

1. **There is no intermediate grading**
    a. We still urge you to complete the milestones on time, so that you will not be overwhelmed at the end of the semester.
    b. The parts of the project are created in a way so that you will use the things you learn in the course and the exercise session and have hands-on experience.
    c. Every deliverable should include the text from the previous one too, explicitly updated to reflect the changes you made to address the feedback we gave you.
2. **Collaboration**
    a. We want you to collaborate
    b. We DO NOT GO INTO how you will split the work -> As long as you do equal parts of the work
    c. Writing the queries can (and should!) be done by everyone!
        i. You can solve the queries in multiple ways to find the optimal one (and help the optimizer)
3. The only important deadline on which you are graded is the last one **BUT** if you want feedback make sure to send us the milestone deliverables!

Can I discard some data?

Dropping some rows is acceptable, in absence of non-nullable data or data that cannot be inferred. Under no circumstances, however, should you drop a significant chunk of the data or columns/attributes. Whenever you drop some data, you should include the description of the dropped data and the reason for doing so.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

How long should the deliverables be?

There is no strict page limit, as long as the deliverables report on the points we requested and are informative.

How should I choose my team?

Putting teams together is entirely up to you. Our advice is that every team member should be exposed equally to every task of the project. While, for example, it might appear tempting to a good data analyst to focus on the data cleaning and loading and quickly finish her assigned task, she will then be disadvantaged in the course midterm and final, because her SQL and query optimization experience will be limited.

What should I do if one of my teammates does not work?

We advise that you address the issue early on, before you encounter high load due to a deadline. We cannot be more lenient to such teams as a whole for fairness. During the final project presentation, however, it becomes obvious whether a team member did not place equal effort; this student will get a lower grade. Please inform us in case there is a conflict or if your teammate decides to withdraw from the course, then we can try to address this issue.

When can I ask questions about the project?

The weekly project session is the intended place for questions. Otherwise, please use the Moodle forum for questions that are of interest to your colleagues. Finally, every TA has specified office hours that you can use for further clarifications.