 **🧪 Lab**  | Traffic Collisions in California

**INTRODUCTION:** Data is often stored across multiple tables to keep the storage requirements compact, and to organize different types of data. Knowing how to use a join is a vital skill when working with data, since bringing tables together can open the door to additional insights that are cumbersome or impossible looking at just one table at a time.

In this lab, you’ll use your proficiency with joins to help a reporter in California use data to support an article they’re writing on the causes of motor vehicle accidents. In particular, they want some information about how many accidents are caused by the influence of alcohol, or due to inattention (such as using a cell phone to text or talk to others), and when these types of accidents tend to occur.

.

PROMPT: To help the reporters out, you will be making use of data regarding traffic accidents in the state of California released by the California Highway Patrol. Certain insights can be found by looking at data on the incident level, while other insights are possible by looking deeper at the parties involved in an incident. But to make insights across those two levels, we need a join to be able to relate the unique information contained in each table.

**—** Data Set **Description**

Data for this lab comes from the California Highway Patrol’s Statewide Integrated Traffic Records System (SWITRS). The SWITRS data we’ve provided (switrs) consists of two tables from the 2019 data collection: switrs\_collisions and switrs\_parties. The tables are related hierarchically. At the top level, there is a unique row and identifier for each incident in the collisions table. Then, in the lower level, each collision is between one or more parties, which include vehicles, pedestrians, etc.

The original collisions table has 469 664 rows and 76 columns, but we’ll be focusing on only the following four columns in this project:

* **case\_id** - unique identifier for each collision
* **collision\_time** - time of day when collision occurred, in 24 hour format
* **day\_of\_week** - day of week when collision occurred. Note that numbering starts at 1 = Monday and ends at 7 = Sunday (instead of 0 = Sunday)
* **party\_count** - number of parties involved in the collision

The original parties table has 940 216 rows and 33 columns, with the following five columns of interest:

* **case\_id** - associated with a collision with matching case\_id, may not be unique
* **party\_number** - numbering of parties involved, always starts from 1 for each collision
* **at\_fault** - Y/N indicating whether party was at fault for collision
* **party\_sobriety** - encodings for whether or not the party had been drinking
* **oaf\_1**, **oaf\_2** - encodings for other associated factors

Most of the features in the dataset are coded in some way for efficient data storage, which can make working with highly detailed data like this tricky. This includes the party\_sobriety, oaf\_1, and oaf\_2 columns you’ll be investigating in the lab. Don’t sweat that point, though: the instructions will explain the encoding values relevant to the tasks.

**— Task 1:** How frequently does alcohol use or lack of attention feature in accidents?

To start, we should run some queries on the switrs\_parties table to understand how fault, alcohol use, and inattention are attributed to accidents.

1. Write a query that answers the following question: According to this dataset, how many people are at fault for a collision?

(paste your query below 👇)

|  |
| --- |
| select count(\*)  from switrs\_parties  where at\_fault='Y' |

(write your **answer** below 👇)

|  |
| --- |
| 438491 |

1. The party\_sobriety field takes on a value of 'B' when the party is known to have been drinking, and under the influence of alcohol. Modify your query from part A to answer the following question: How many parties were found at fault while under the influence of alcohol?

(paste your query below 👇)

|  |
| --- |
| select count(\*)  from switrs\_parties  where at\_fault='Y'  and party\_sobriety='B' |

(write your **answer** below 👇)

|  |
| --- |
| 33512 |

1. The **oaf\_1** or **oaf\_2** feature takes on a value of 'F' if inattention was a factor in the collision. Modify your query to answer the following question: How many parties were found at fault while lack of attention was a factor in the collision?

(paste your query below 👇)

|  |
| --- |
| select count(\*)  from switrs\_parties  where at\_fault='Y'  and (oaf\_1='F' or oaf\_2='F') |

(write your **answer** below 👇)

|  |
| --- |
| 18311 |

**— Task 2:** When do accidents occur by day of the week?

Now that we have a way to identify whether or not a collision can be attributed to alcohol or inattention, let’s add in the switrs\_collisions table to answer the journalist’s question of whether or not there are differences between the two accident sources.

1. Let’s start with the switrs\_collisions table on its own. Write a query that returns the number of collisions, grouped by day of the week. Which days have the highest number of collisions, and which days have the least number? Sunday is indicated by a 7 instead of a 0.

(paste your query below 👇)

|  |
| --- |
| SELECT  day\_of\_week,  count(case\_id) as number\_of\_collisions  from switrs\_collisions  group by 1  order by 2 DESC |

(write your **answer** below 👇)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 5 | 75654 | | 4 | 70298 | | 3 | 69919 | | 2 | 69189 | | 1 | 65983 | | 6 | 63462 | | 7 | 55159 | |  |  | |

1. The switrs\_collisions table and switrs\_parties tables share values in the **case\_id** column. Write a new query that inner joins the two tables on that column, returning the number of rows. How many rows are in the combined output table, and why?

(paste your query below 👇)

|  |
| --- |
| SELECT count(\*)  from switrs\_collisions sc  join switrs\_parties sp  on sc.case\_id = sp.case\_id |

(write your **answer** below 👇)

|  |
| --- |
| 940216 |

1. Combine the queries from parts A and B to return the number of collisions grouped by the day of the week. Add a condition for the involved parties so that we only count accidents where the party was found to be at fault AND under the influence of alcohol. Which days have the highest number of collisions, and which days have the smallest number?

(paste your query below 👇)

|  |
| --- |
| SELECT  day\_of\_week,  count(sc.case\_id) as number\_of\_collisions  from switrs\_collisions sc  join switrs\_parties sp  on sc.case\_id = sp.case\_id  where at\_fault = 'Y' AND party\_sobriety = 'B'  group by 1  order by 2 desc |

(write your **answer** below 👇)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 7 | 7603 | | 6 | 7523 | | 5 | 5097 | | 4 | 3552 | | 1 | 3478 | | 3 | 3189 | | 2 | 3070 | |

1. Modify your query to look at the number of accidents by the day of the week where the party was found to be at fault AND inattention was a factor. Which days have the highest number of collisions, and which days have the smallest number?

(paste your query below 👇)

|  |
| --- |
| SELECT  day\_of\_week,  count(sc.case\_id) as number\_of\_collisions  from switrs\_collisions sc  join switrs\_parties sp  on sc.case\_id = sp.case\_id  where at\_fault = 'Y' AND (oaf\_1 = 'F' or oaf\_2 = 'F')  group by 1  order by 2 desc |

(write your **answer** below 👇)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 5 | 3030 | | 4 | 2762 | | 3 | 2760 | | 2 | 2754 | | 1 | 2672 | | 6 | 2273 | | 7 | 2060 | |

**— LevelUp**

Simply because an accident was such that inattention was a factor does not necessarily mean that a cell phone was the source of the driver’s distraction. In the parties table, there is a column called **sp\_info\_2.** This feature takes on a value of B, 1, or 2 if a cell phone was known to be in use at the time of the accident.

1. If you’re interested in digging deeper, you might want to try seeing what proportion of accidents were caused by cell phone distraction, and if they differ from other ‘inattention’ accidents.

Keep in mind that the **sp\_info\_2** column is a string data type, so you’ll need to treat the '1', and '2' codes appropriately!

(paste your query below 👇)

|  |
| --- |
| SELECT  day\_of\_week,  count(sc.case\_id) as number\_of\_collisions  from switrs\_collisions sc  join switrs\_parties sp  on sc.case\_id = sp.case\_id  where sp\_info\_2 in ('1','2','1.0','2.0','B')  group by 1  order by 2 desc |

(write your **answer** below 👇)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (write   |  |  | | --- | --- | | 5 | 1915 | | 4 | 1871 | | 2 | 1761 | | 3 | 1737 | | 6 | 1651 | | 1 | 1622 | | 7 | 1453 |   your answer here) |

**— Investigation (Open-Ended)**

Don't let these prompts limit what you would like to examine in the dataset! What else can you find? Are there any interesting trends, correlations, or insights you can uncover? Any visualizations you can create to better illustrate your findings? Dive in and explore the data to its fullest!

(paste your query below 👇)

|  |
| --- |
| (write your query here) |

(write your **insights** below 👇)

|  |
| --- |
| (write your answer here) |

---------------------------------------------------------------------

(paste your query below 👇)

|  |
| --- |
| (write your query here) |

(write your **insights** below 👇)

|  |
| --- |
| (write your answer here) |

---------------------------------------------------------------------

(Share any visualizations below 👇)

|  |
| --- |
|  |

|  |
| --- |
|  |