Milestone 6 | 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 Milestone, 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.

HOW IT WORKS: Follow the prompts in the questions below to investigate your data. Post your answers in the provided boxes: the **yellow boxes** for the queries you write, **purple boxes** for visualizations and **blue boxes** for text-based answers. When you're done, export your document as a pdf file and submit it on the Milestone page – see instructions for creating a PDF at the end of the Milestone.

RESOURCES: If you need hints on the Milestone or are feeling stuck, there are multiple ways of getting help. Attend Drop-In Hours to work on these problems with your peers, or reach out to the HelpHub if you have questions. Good luck!

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 Milestone 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: `collisions` and `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 Milestone:

- **`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 Milestone. Don't sweat that point, though: the instructions will explain the encoding values relevant to the tasks.

If you're curious to explore the data further on your own, or want to see what other parts of the dataset that aren't available are like, you can find a comprehensive description of the data in full here, on the SWITRS information page.

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

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

A. 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
  at_fault,
  COUNT(case_id) AS n_collisions
FROM switrs.parties
GROUP BY at_fault
```

Note: You could also just do a COUNT on the remaining data after filtering `WHERE at_fault = 'Y'` to just get the requested count. One advantage of the query listed here is that you can directly compare at-fault parties to not-at-fault parties.

(write your **answer** below 👇)

There are 438 491 parties listed as at-fault in the table. Interestingly, this is fewer than the number of collisions, so there must be collisions where no party is at fault.

B. 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
  at_fault,
  COUNT(case_id) AS n_collisions
FROM switrs.parties
WHERE party_sobriety = 'B'
GROUP BY at_fault
```

(write your **answer** below 👇)

There are 33 512 parties listed as at-fault in the table while also listed as under the influence of alcohol. This is the majority of all parties listed as under the influence (only 1547 not at fault).

C. 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
  at_fault,
  COUNT(case_id) AS n_collisions
FROM switrs.parties
WHERE
  oaf_1 = 'F'
  OR oaf_2 = 'F'
GROUP BY at_fault
```

(write your **answer** below 👇)

There are 18 311 parties listed as at-fault, where inattention was a factor. (There are 1411 parties where inattention was a factor, but the party was not found to be at fault.)

– **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 `collisions` table to answer the journalist's question of whether or not there are differences between the two accident sources.

A. Let's start with the `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? Note: Day of week is encoded slightly differently than what comes out of the `date_part` function: Sunday is indicated by a 7 instead of a 0.

(paste your query below 👇)

```
SELECT
  day_of_week,
  COUNT(case_id) AS n_collisions
FROM switrs.collisions
GROUP BY day_of_week
ORDER BY day_of_week ASC
```

Note: Sorting is not required, but is useful for the analysis.

(write your **answer** below 👇)

Accidents are higher during the week than on the weekends, with their highest count on Fridays. Saturdays see the second-fewest collisions, and Sundays the least.

B. The `collisions` table and `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(*) AS n_rows
FROM switrs.collisions AS c
INNER JOIN switrs.parties AS p
  ON c.case_id = p.case_id
```

(write your **answer** below 👇)

There are 940 216 rows in the joined table, equal to the number of rows in the parties table alone. This makes sense since there is a guaranteed relationship between collisions and parties, where every collision involves one or more parties. We have no risk of losing rows to an inner join.

C. 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
  c.day_of_week,
  COUNT(c.case_id) AS n_collisions
FROM switrs.collisions AS c
INNER JOIN switrs.parties As p
  ON c.case_id = p.case_id
WHERE
  p.at_fault = 'Y'
  AND p.party_sobriety = 'B'
GROUP BY c.day_of_week
ORDER BY c.day_of_week ASC
```

(write your **answer** below 👇)

In the overall data, there were more collisions on weekdays than weekends; the opposite is true for collisions where the at-fault party was under the influence of alcohol. Now, Saturday and Sunday have the highest number of accidents, while Tuesday sees the smallest number of accidents.

D. 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
  c.day_of_week,
  COUNT(c.case_id) AS n_collisions
FROM switrs.collisions as c
INNER JOIN switrs.parties as p
  ON c.case_id = p.case_id
WHERE
  p.at_fault = 'Y'
  AND (
    p.oaf_1 = 'F'
    OR p.oaf_2 = 'F'
```

```
    )
GROUP BY c.day_of_week
ORDER BY c.day_of_week ASC
```

(write your **answer** below 👇)

The accidents by day of the week distribution is fairly similar to the overall distribution. The highest number of collisions occurred on Fridays, with the lowest number of collisions on Saturdays and Sundays.

— **Task 3:** When do accidents occur by the time of day?

A data analyst colleague of yours has taken interest in your project with the journalist and has pitched in their own contribution by providing you a summary of the dataset with five features:

- **`alcohol_involved`** - TRUE/FALSE whether or not the party at fault was under the influence of alcohol
- **`inattention_involved`** - TRUE/FALSE whether or not inattention was a factor for the party at fault
- **`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)
- **`hour_of_day`** -hour of day when collision occurred, in 24 hour format (0–2300). Values of 2500 indicate an unknown time of day.
- **`n_collisions`** - number of collisions matching the conditions of the first four columns

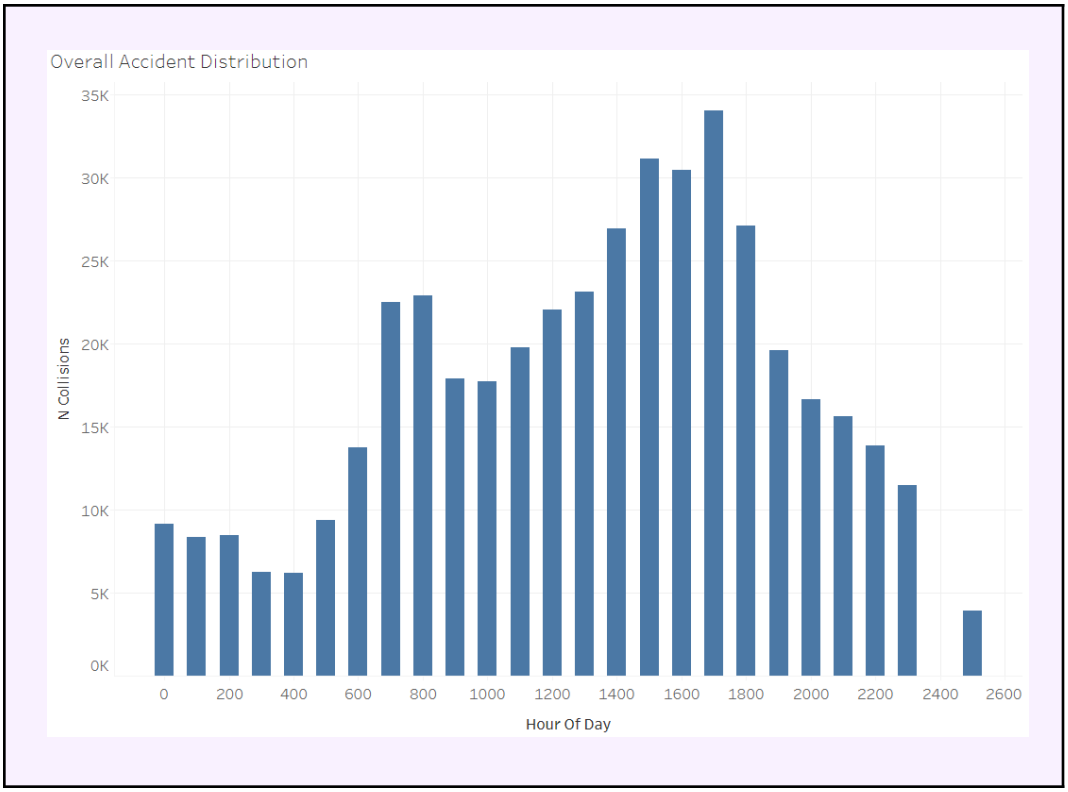
Let's use this new data summary to look at how accident patterns change based on the time of day. Since the data has already been queried, we'll do this visually within Tableau! Click this link **to navigate to the workbook you'll use to complete the remainder of this Milestone.** Once you've published your Tableau Workbook in the folder named Upload Workbooks Here, paste the Share Link in the box below.

SHARE LINK

Continue to post your answers in the provided boxes: purple boxes for your visualizations, and blue boxes for text-based answers.

A. On Sheet 1, create a bar chart of the number of collisions by the hour of day. Describe the pattern in the data. Are there times of day where more accidents occur? Does this fit in with your expectations?

HINT: Drag the `Hour Of Day` pill to the Columns and the `N Collisions` pill to the Rows.

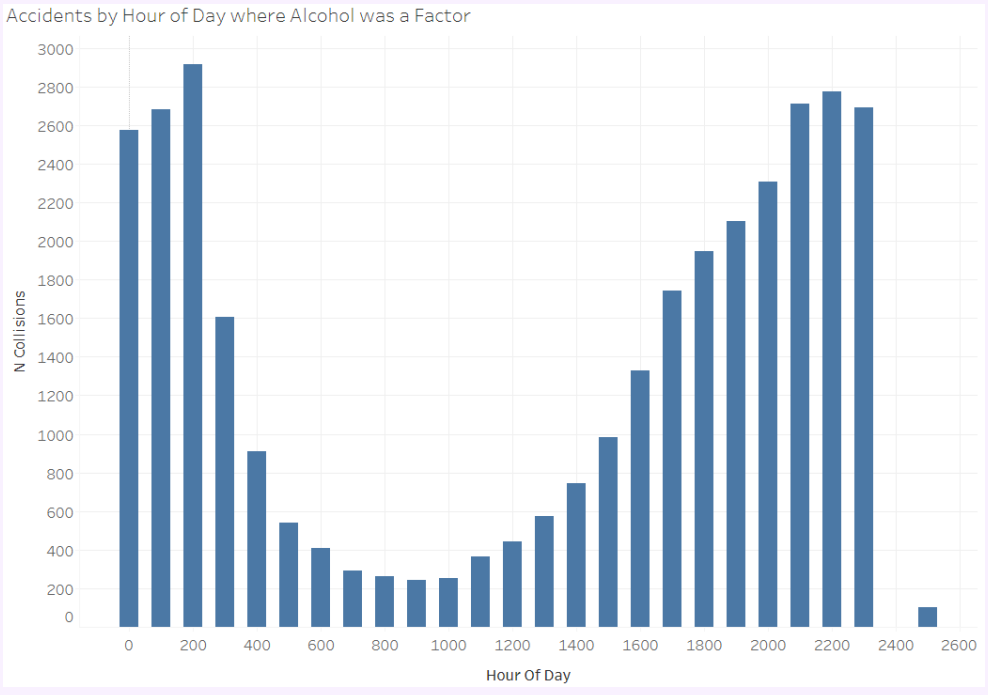


(write your **answer** below 👇)

The distribution of collisions by hour of day is somewhat bimodal, with a weak peak corresponding to the morning rush hours at 7am and 8am, and a stronger peak corresponding to the evening rush hours from 3pm to 5pm.

Notice that there are a number of collisions listed in the 2500 hour, which does not correspond with an actual hour of day.

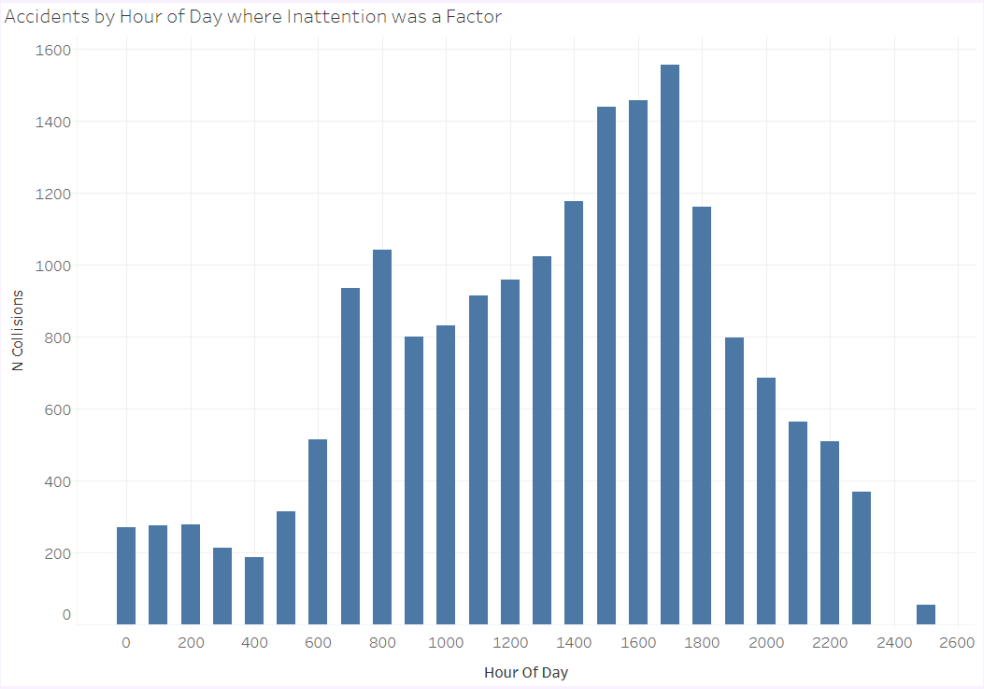
B. Copy the chart into a new sheet and add a filter so that the bar chart only shows accidents where the party at fault was found to be under the influence of alcohol. How does this distribution of accidents by time of day compare to the overall distribution?



(write your **answer** below 👇)

The distribution of collisions where alcohol was an at-fault feature is very different from the overall distribution. Both rush hour humps are absent, and there is only a single peak in the evening hours, peaking from 9pm to 2am.

C. Copy the chart into one more sheet, but now change the filter to only look at accidents where inattention was a factor from the party-at-fault. How does this distribution compare to the overall distribution?



(write your **answer** below 👇)

Like with the analysis of accidents by the day of the week, the distribution of collisions where inattention was a factor with an at-fault party are not particularly different from the distribution of collisions as a whole. The peaks in incidents by hour of day are a little sharper, but still focused around the morning and

afternoon rush hour periods and the time in between those peaks. In general, we can infer that accidents caused by inattention can occur any time any other normal accident might occur!

— Level Up

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. 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
  at_fault,
  COUNT(case_id) AS n_collisions
FROM switrs.parties
WHERE sp_info_2 IN ('B', '1', '2')
GROUP BY at_fault
```

(write your **answer** below 👇)

We found in Task 1 that 18,311/438,491 = 42% of at-fault collisions listed inattention as a factor.

From this query, we can see that there are 7885 parties listed as at-fault where a cell phone was in use. That's roughly 2% of

at-fault collisions, significantly less than the proportion listing inattention as a factor.

— Submission

Great work completing this Milestone! To submit your completed Milestone, you will need to download / export this document as a PDF and then upload it to the Milestone submission page. You can find the option to download as a PDF from the File menu in the upper-left corner of the Google Doc interface.