✓ **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.

SQL App: Here's that link to our specialized SQL app, where you'll write your SQL queries and interact with the data.

– 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.

– **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?

```
SELECT
  COUNT(at_fault) AS total_at_fault
FROM
  switrs.parties
WHERE
  at_fault = 'Y';
```

According to this dataset, there are 438,491 people at fault for a collision.

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?

```
SELECT
  COUNT(at_fault) AS total_at_fault
FROM
  switrs.parties
WHERE
```

```
  at_fault = 'Y'
  AND party_sobriety = 'B';
```

33,512 people were at fault of a collision while under the influence in this dataset.

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?

```
SELECT
  COUNT(at_fault) AS total_at_fault
FROM
  switrs.parties
WHERE
  at_fault = 'Y'
  AND (
    oaf_1 = 'F'
    OR oaf_2 = 'F'
  );
```

There were 18,311 people where the parties were at fault while lack of attention was a factor in the collision.

– **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.

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

Accidents are highest on Fridays, and lowest on Sundays.

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?

```
SELECT
  COUNT(*) as n_collisions
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties AS b ON a.case_id = b.case_id;
```

The query returned an output of 940216 rows of data when both tables were joined by the case_id.

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?

```
WITH car_accident_table AS (
  SELECT
    *
  FROM
    switrs.collisions AS a
    INNER JOIN switrs.parties AS b ON a.case_id = b.case_id
)
SELECT
  day_of_week,
  COUNT(at_fault) AS total_at_fault
FROM
  car_accident_table
WHERE
  at_fault = 'Y'
  AND party_sobriety = 'B'
GROUP BY
  day_of_week;
```

Saturday and Sunday appear to have the highest number of collisions meeting the criteria of both at fault and under the influence. Sunday accounted for 7,603 collisions and Saturday accounted for 7,523 collisions. On the other hand, Tuesday and Wednesday had the least number of collisions for at fault and under the influence. Tuesday accounted for 3,070 collisions and Wednesday accounted for 3,189 collisions.

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?

```
WITH car_accident_table AS (
  SELECT
    *
  FROM
    switrs.collisions AS a
    INNER JOIN switrs.parties AS b ON a.case_id = b.case_id
)
SELECT
  day_of_week,
  COUNT(at_fault) AS total_at_fault
FROM
  car_accident_table
WHERE
  at_fault = 'Y'
  AND (
    oaf_1 = 'F'
    OR oaf_2 = 'F'
  )
GROUP BY
  day_of_week;
```

Thursday and Friday had the highest number of collisions meeting the criteria of at fault and inattention being a contributing factor. Thursday accounted for 2,762 collisions and Friday accounted for

3,030 collisions meeting our criteria. On the other hand, Saturday and Sunday held the least number of collisions for being at fault and inattention was a contributing factor. Saturday accounted for 2,273 collisions and Sunday accounted for 2,060 collisions.

– **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

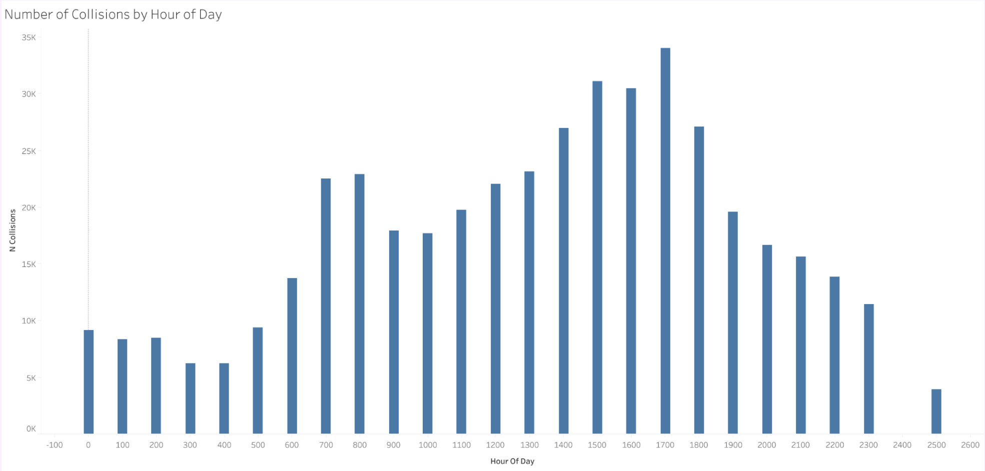
Let's use this new data summary to look at how accident patterns change based on the time of day. Since the data has already been queried, we'll do this visually within Tableau! Click this link **to navigate to the workbook you'll use to complete the remainder of this Milestone.** Once you've published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

https://prod-useast-b.online.tableau.com/#/site/globaltech/workbooks/1237675?:origin=card_share_link

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

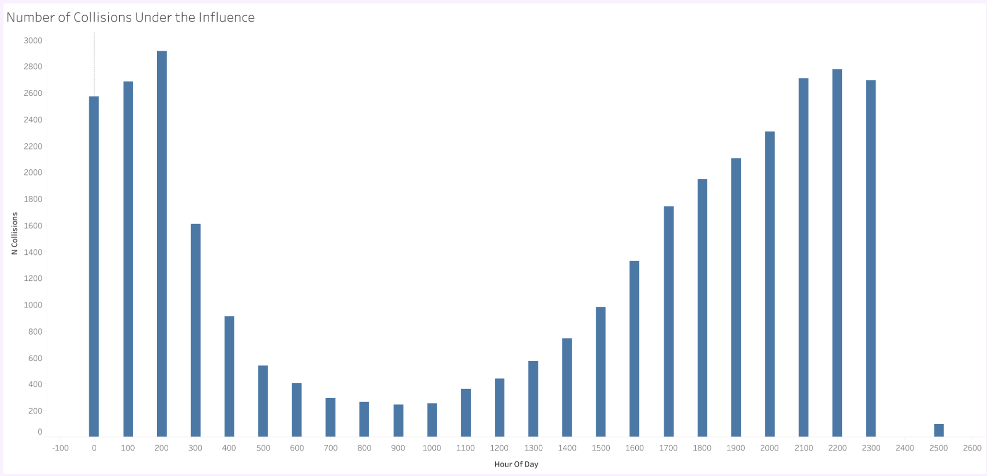
A. On Sheet 1, create a bar chart of the number of collisions by the hour of day. Describe the pattern in the data. Are there times of day where more accidents occur? Does this fit in with your expectations?

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



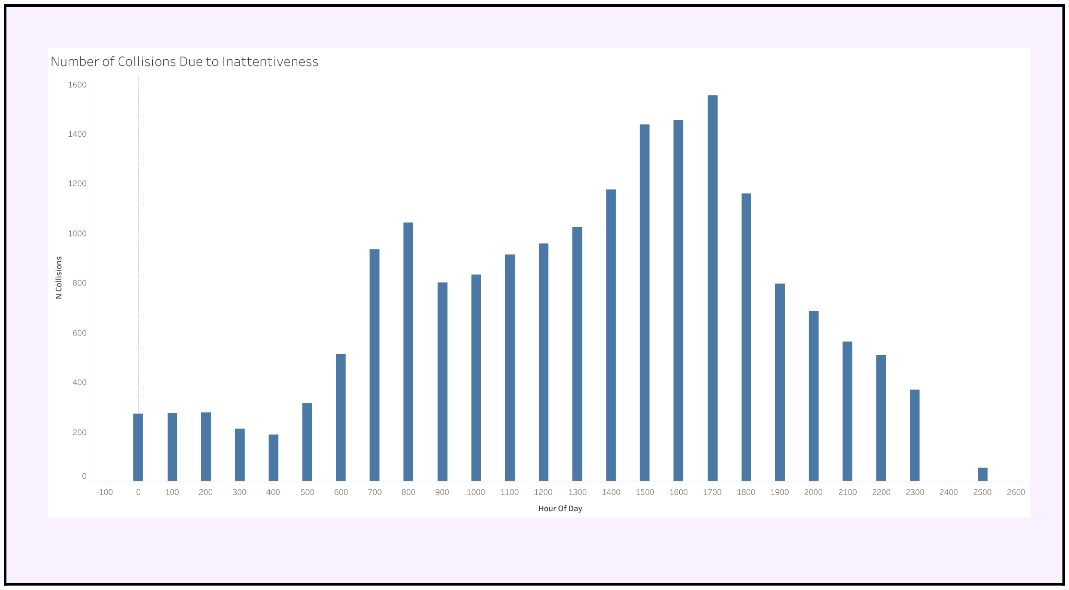
At 1400 and 1700 hours or 3 pm and 5 pm exists a peak time when a high number of collisions occur. Another peak is present at 7 am and 8 am. This fits with my expectations because these are the times when traffic rush occurs.

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?



This second graph looks like an inverse of our previous chart. There is an abundance of collisions due to alcohol-related cases from midnight to 2 am. From 9 pm to 11 pm there is another peak of collisions related to alcohol influence. Finally, from 4 am to 4 pm, exists a range where the number of collisions concerning alcohol influence is at its lowest numbers.

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?



The distribution of this chart closely resembles the first chart that we created. However, it does not have a strong connection to our second chart.

— 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.

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!

```
SELECT
  COUNT(at_fault) AS total_at_fault
FROM
  switrs.parties
WHERE
  at_fault = 'Y'
  AND sp_info_2 IN ('B', '1', '2');
```

People at fault with cell phones being a contributing factor for the collision accounted for 7,885 cases. On the other hand, people at fault and inattentiveness being a contributing factor accounted for 18,311 collisions. This 10,426 difference can indicate that inattentiveness causes more collisions than phone usage while driving.

– 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.