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.

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 and answer the following question: How many parties are cited as being at fault for a collision?

```
SELECT COUNT(party_number) AS
num_of_parties_that_are_cited_for_being_at_fault_for_a_collis
on
FROM switrs.parties
WHERE at_fault = 'Y'
```

438491 are the number of parties that are cited as being 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(party_number) AS
num_of_parties_at_fault_for_a_collison_under_the_influence_
FROM switrs.parties
WHERE at_fault = 'Y'
AND party_sobriety = 'B'
```

33512 are the number of parties that were found at fault for a collision while under the influence.

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(party_number)
AS num_of_parties_at_fault_for_a_collison_due_to_inattention
FROM switrs.parties
WHERE at_fault = 'Y'
AND (oaf_1 = 'F' OR oaf_2 = 'F')
```

18311 are the number of parties that were found at fault for a collision due to their lack of attention.

— **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(day_of_week) AS number_of_collisions
FROM switrs.collisions
GROUP BY day_of_week
ORDER BY number_of_collisions DESC
```

The days of the week that have the highest number of collisions are day 5 and 4, while day 6 and 7 have 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?

```
SELECT
COUNT (collisions.case_id) as
combined_amount_of_rows_in_case_id
```

```
FROM switrs.collisions
INNER JOIN switrs.parties
ON collisions.case_id = parties.case_id
```

940216 are the number of rows that are combined in the output table that shares values in the case_id.
The reason that there are 940216 rows is because I am combining the case.id that shares values from the switrs.collision database and the switrs.parties database.

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?

```
SELECT
collisions.day_of_week,
COUNT(collisions.case_id) AS
num_of_collisons_where_the_party_was_at_fault_and_innebriated
FROM
switrs.collisions
INNER JOIN
switrs.parties ON collisions.case_id = parties.case_id
WHERE
parties.at_fault = 'Y'
AND parties.party_sobriety = 'B'
GROUP BY
collisions.day_of_week
ORDER BY
```

```
num_of_collisons_where_the_party_was_at_fault_and_innebriated
DESC
```

The days of the week that have the highest number of collisions where the party was at fault and under the influence are days 7 and 6, while days 3 and 2 have the least.

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?

```
SELECT
collisions.day_of_week,
COUNT(collisions.case_id) AS
num_of_collisons_where_the_party_was_at_fault_and_inatentive
FROM
switrs.collisions
INNER JOIN
switrs.parties ON collisions.case_id = parties.case_id
WHERE
parties.at_fault = 'Y'
AND (parties.oaf_1 = 'F' OR parties.oaf_2 = 'F')
GROUP BY
collisions.day_of_week
ORDER BY
num_of_collisons_where_the_party_was_at_fault_and_inatentive
DESC
```

The days of the week that have the highest number of collisions where the party was at fault and inattentive are days 5 and 4, while days 6 and 7 have the least.

— **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:

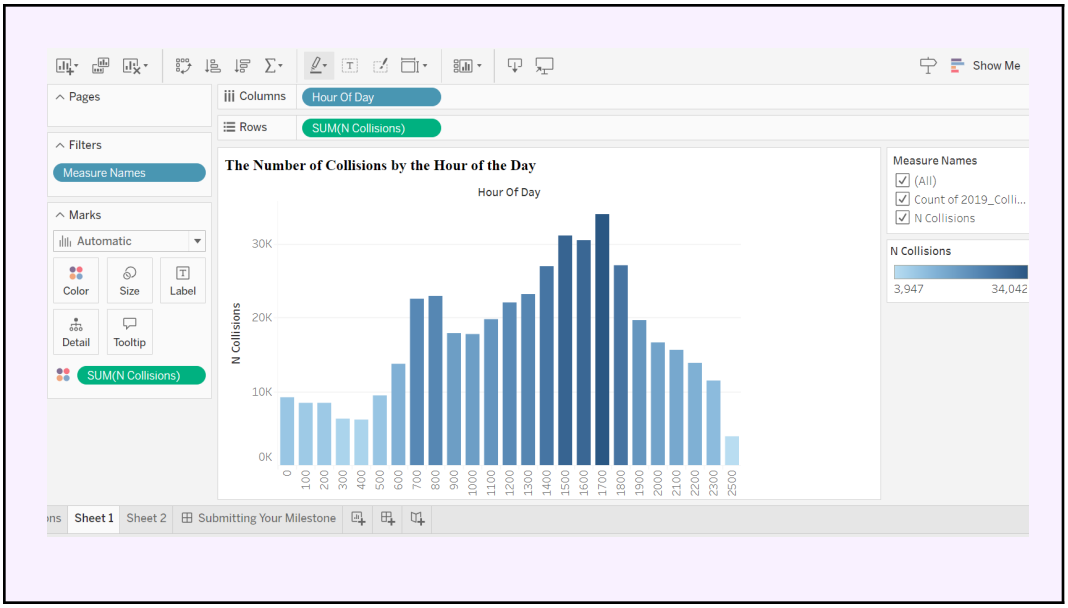
- **`alcohol_involved`** - TRUE/FALSE whether or not the party at fault was under the influence of alcohol
- **`inattention_involved`** - TRUE/FALSE whether or not inattention was a factor for the party at fault
- **`day_of_week`** - day of week when collision occurred. Note that numbering starts at 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/t/globaltech/views/milestone6a/SubmittingYourMilestone

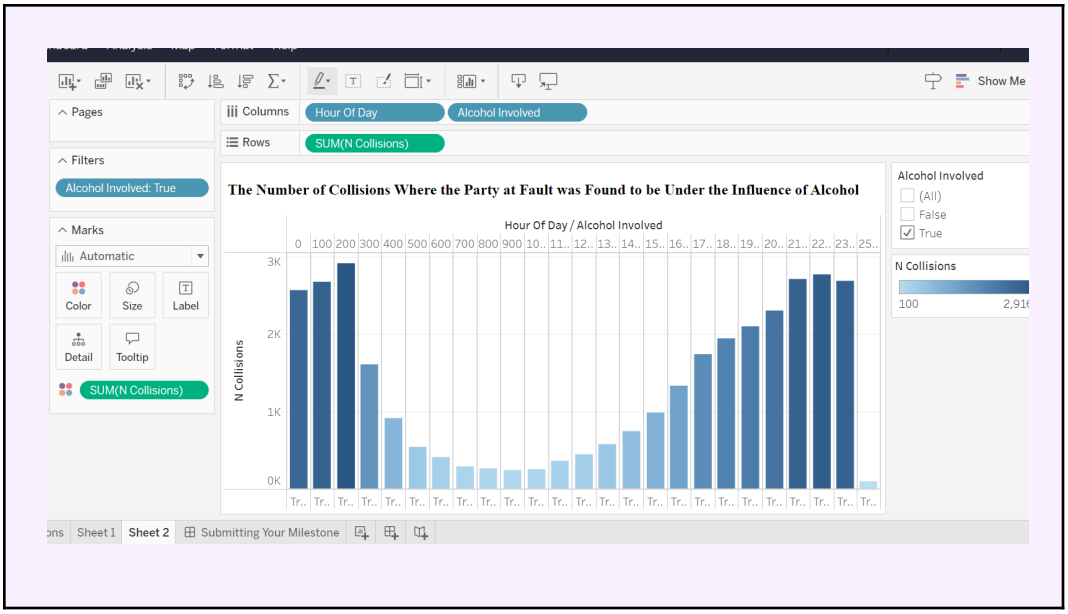
Continue to post your answers in the provided boxes: purple boxes for your visualizations, and blue boxes for text-based answers.

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?



The number of accidents increases progressively as the day goes by until it reaches a peak at hour 17 and then begins to decrease for the rest of the day.
This data fits within my expectations because at the day progresses accidents are more likely to occur, since people are out and about commuting to work and the highest amount of accidents are during rush hour.

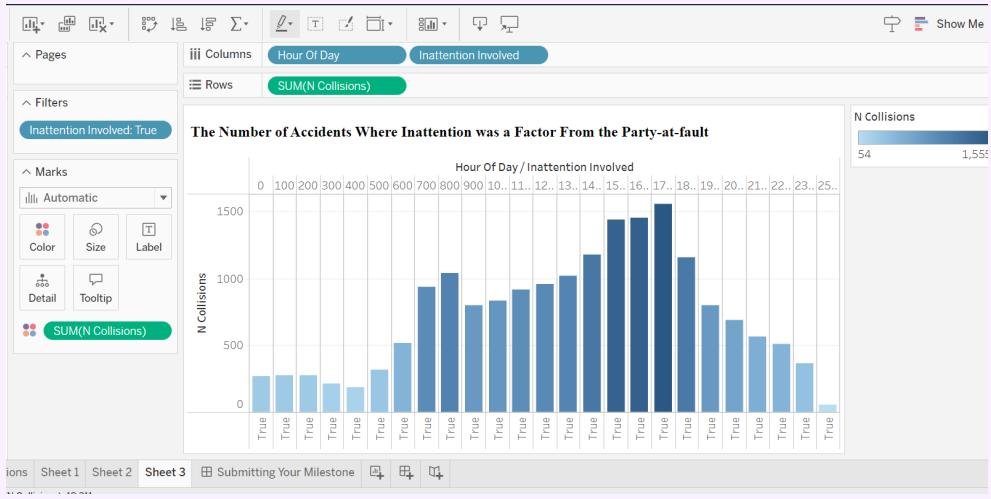
B. Copy the chart into a new sheet and add a filter so that the bar chart only shows accidents where the party at fault was found to be under the influence of alcohol. How does this distribution of accidents by time of day compare to the overall distribution?



Unlike the first bar chart that showed the amount of collisions by the hour of the day increased progressively, this bar shows the opposite. The amount of accidents where the party was under the influence of alcohol starts extremely high at the beginning of the day until the accidents decrease sharply during the day, until it suddenly increases later into the evening.

This is also within expectations because I would suspect that the most amount of accidents that occur, are during time periods where a person is mostly likely to be drinking, such as the early morning and later day/evening, which is backed up by the data.

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?



Lastly, This bar chart closely resembles the 1st bar chart. The amount of accidents increased progressively as the day went on, spiking during rush hours such as hour 7 and 17.
This leads me to believe that the majority of accidents during the day are due to people being inattentive while commuting to and back from work, while the majority of accidents during the night is caused by alcohol.

— 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!

```
SELECT
CASE
WHEN sp_info_2 IN ('1', '2') THEN 'Phone in Use'
ELSE 'Phone Not in Use'
END AS phone_usage,
COUNT(*) AS number_of_accidents
FROM switrs.parties
WHERE oaf_1 = 'F' OR oaf_2 = 'F'
GROUP BY phone_usage
```

I broke the amount of people that have gotten into an accident due to inattentiveness into two groups, people that were and were not using their phone.
17231 people that got into an accident due to inattentiveness did not use their phone, while 2491 people did.
This makes me wonder what were the other reasons for inattentiveness that people had when that crashed.

— 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.