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?

```
--INNER JOIN is not necessary for TASK 1.A
SELECT
  b.at_fault,
  COUNT(*) as n_of_faults -- Counts the # of "Faults"
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
GROUP BY
  b.at_fault
```

The number of parties shown to be at "fault" for creating a collision is 438,491. The number of parties in a collision whose "fault" it wasn't is 501,725. Totally, out of the nearly 1M collisions, it was the "fault" of 46.63%.

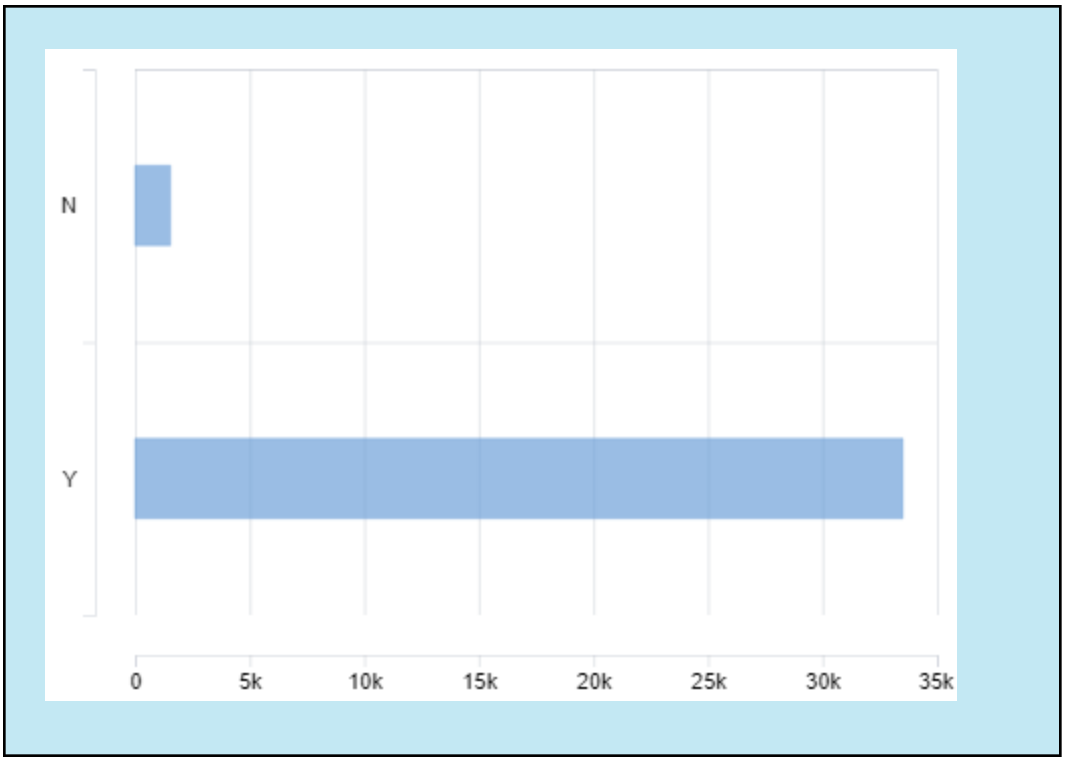
B. The party_sobriety field takes on a value of 'B' when the party is known to have been drinking, and under the influence of alcohol. Modify your query

from part A to answer the following question: How many parties were found at fault while under the influence of alcohol?

```
--INNER JOIN is not necessary for TASK 1.B
SELECT
  b.at_fault,
  b.party_sobriety,
  COUNT(*) as n_of_faults -- Counts the # of "Faults"
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
WHERE b.party_sobriety = 'B'
GROUP BY
  b.at_fault,
  b.party_sobriety
```

The data shows that 33,512 collisions were the “fault” of party members who were under the influence of alcohol. This number is 2164.66% higher than party members who were in a collision and whose fault it was AND were NOT under the influence of alcohol.

Visually, this is what the data looks like:



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?

```
--INNER JOIN is not necessary for TASK 1.C
SELECT
  b.at_fault,
  COUNT(*) as n_of_faults -- Counts the # of "Faults"
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
WHERE oaf_1 = 'F' OR oaf_2 = 'F'
GROUP BY
  b.at_fault
```

Inattention is another large contributor in the # of “faults” for collisions made by a party. 18311 parties were “at fault” for the collision they were in due to or in relation with inattention.

Visually, here is a bar graph because I prefer visual aids to help illustrate data:

The Data is significantly greater with Y’s then N’s. With Y’s being Yes to Inattention and N’s being No for Inattention.

– **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.

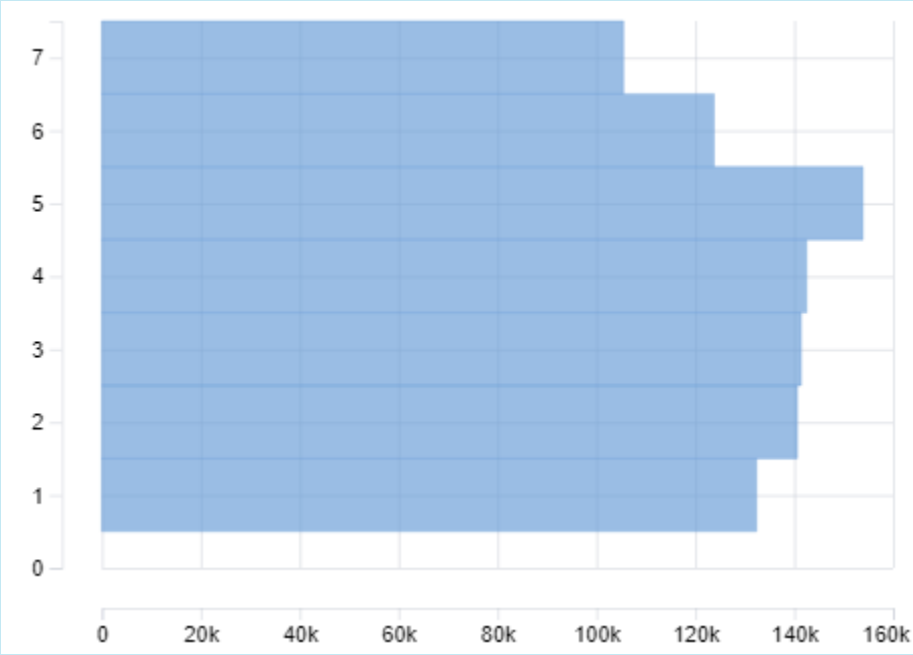
```
SELECT
  a.day_of_week,
  COUNT(*) as n_collisions
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
GROUP BY
  a.day_of_week
ORDER BY n_of_faults DESC
```

Friday has the highest # of collisions at 153,911 and Sunday is the lowest being 105,508. Surprisingly the data shows that Saturday is a less dangerous day to go driving than a Thursday or Wednesday. This data is surprising because my first assumption would be that the Weekend has the highest amount of accidents.

However, after more critical thought, in my opinion this makes sense after thinking about the type of person who would drink and drive. While nothing is wrong with drinking on the weekdays, in my experience, it's more typical for people to drink socially on weekends instead of socially on weekdays.

Where I grew up, heavy drinking on weekdays is a symptom of irresponsibility. Thus, it makes sense that higher accidents happen on Weekdays. Here is a graph to show the total visual

data:



Sunday = 7 Monday = 0 and so forth.

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_total
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
```

940,216 collision cases have been documented in this combined table. I do not understand the “Why” portion of this question.

Why are there 940,216 cases? The answer is because that's how many cases were collected. I'm confused by that question.

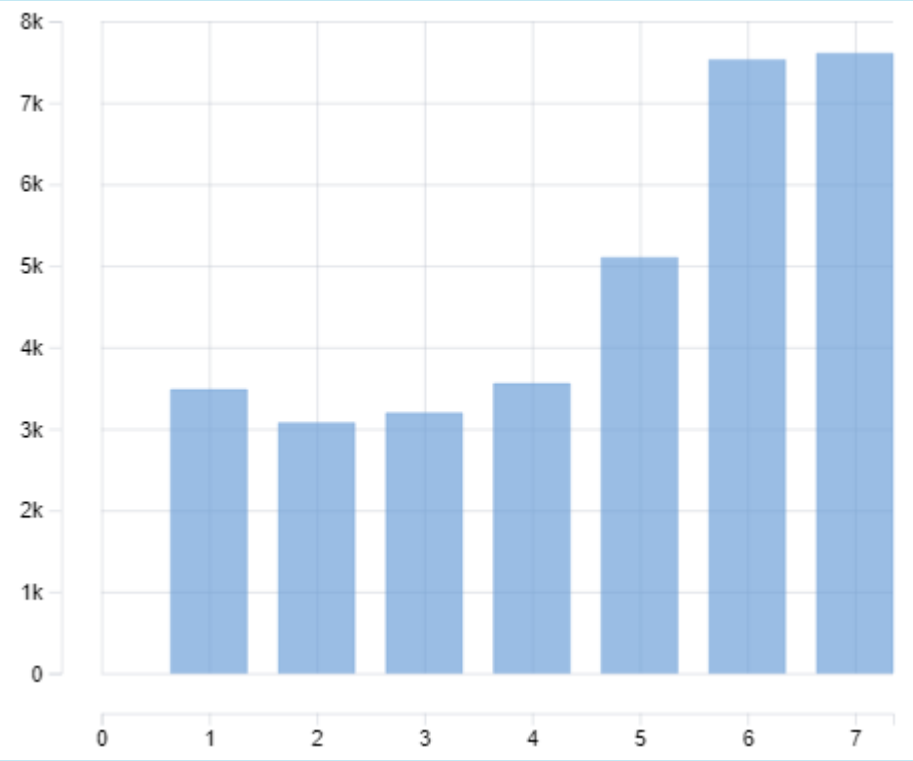
C. Combine the queries from parts A and B to return the number of collisions grouped by the day of the week. Add a condition for the involved parties so that we only count accidents where the party was found to be at fault AND under the influence of alcohol. Which days have the highest number of collisions,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

```
SELECT
  b.at_fault,
  a.day_of_week,
  COUNT(*) as n_collisions --Counts the # of "Faults"
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
WHERE
  party_sobriety = 'B'
  AND b.at_fault = 'Y'
GROUP BY
  b.at_fault,
  a.day_of_week
ORDER BY
  n_collisions DESC
```

Sunday had the highest number of collisions at 7603 that fit the parameters of being under the influence and whose fault it was. Tuesday had the lowest at 3070 collisions.

Sunday has a 247% higher likelihood of collisions happening compared to its lowest counterpart Tuesday.

Here is a graph to provide further visual detail:



.

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?

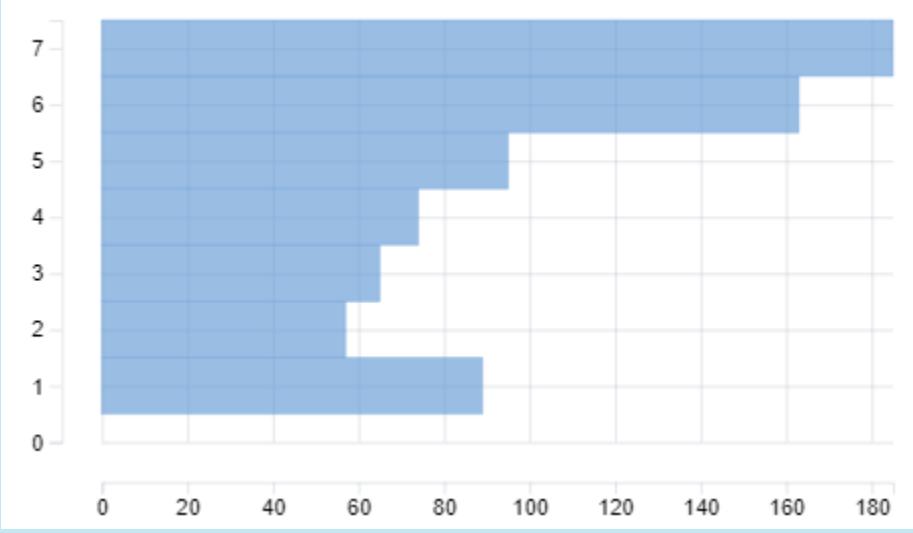
```
SELECT
  b.at_fault,
  a.day_of_week,
  COUNT(*) as n_collisions --Counts the # of "Faults"
FROM
  switrs.collisions AS a
  INNER JOIN switrs.parties as B ON a.case_id = b.case_id
WHERE
  (
```

```
    oaf_1 = 'F'
    OR oaf_2 = 'F'
  )
  AND party_sobriety = 'B'
  AND b.at_fault = 'Y'
GROUP BY
  b.at_fault,
  a.day_of_week
ORDER BY
  n_collisions DESC
```

This data is extremely surprising and it just strengthens the claim I've heard that data can be misleading!

The highest number of collisions that fit the given parameters: it was the party's "fault", it was under the influence of alcohol, and it was combined with inattention, was Sunday at 185, and Saturday coming in second at 163.

Here is a graph to visually see the data.



— **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:

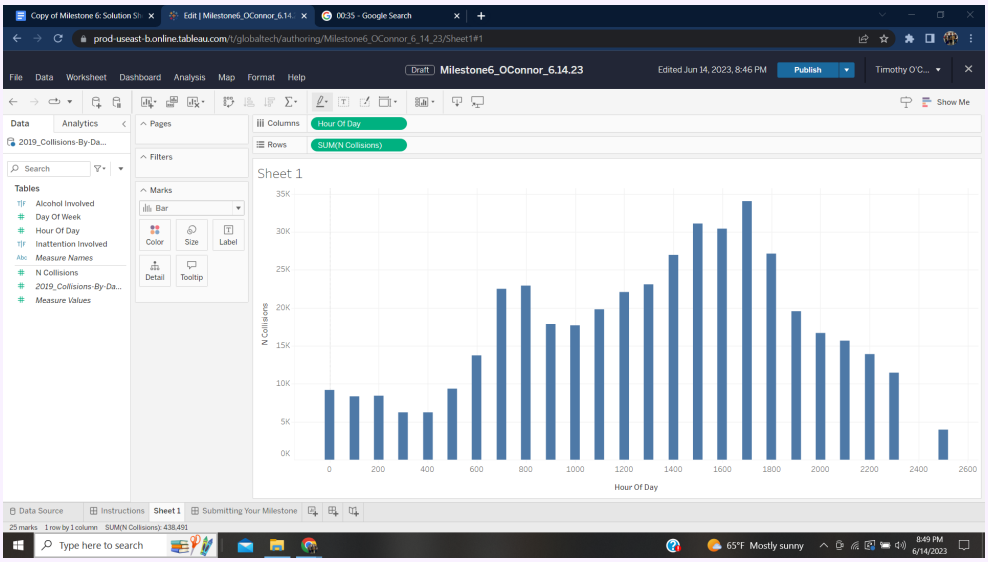
- **`alcohol_involved`** - TRUE/FALSE whether or not the party at fault was under the influence of alcohol
- **`inattention_involved`** - TRUE/FALSE whether or not inattention was a factor for the party at fault
- **`day_of_week`** - day of week when collision occurred. Note that numbering starts at 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/560988?: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?



My expectations were that the majority of people would be drinking and driving during what I would call “Gremlin Hours” or more specifically, 1am-3am.

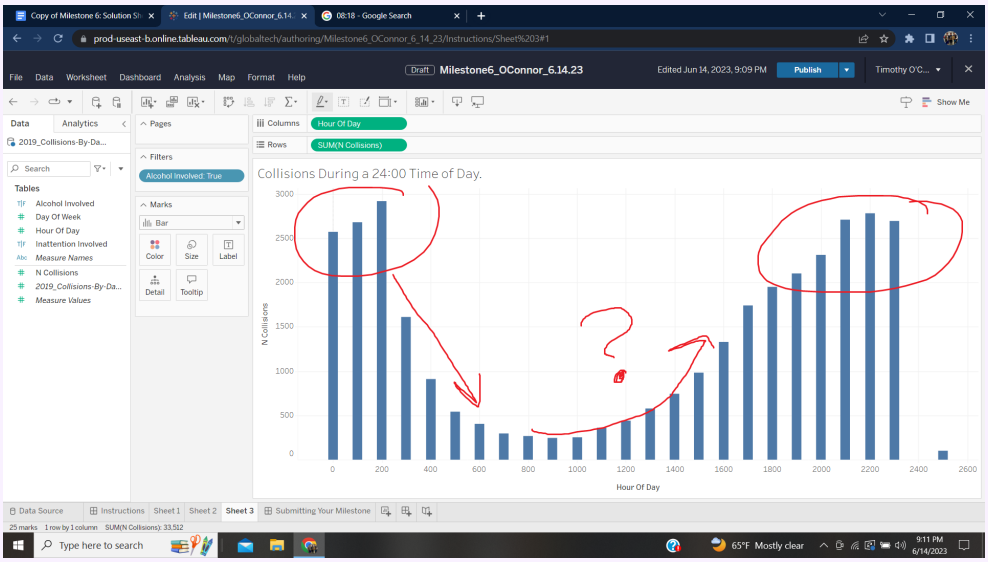
There is a dramatic spike during the hours of 7am and 8am which is confusing and I am worried about those people who are drinking that early to begin with.

Secondly, there is a steady growth in the amount of collisions throughout the day with a peak at 5pm which does not match my expectations either.

The pattern I notice is that from 9am to 7pm, there is a steady and increasing growth rate of collisions. After 8pm, there is a dramatic decline in the number of collisions.

This is reassuring to me as someone who likes to drive late at night. Although this data is from California and I live in NJ so this data is only reassuring if I were to fly out to Cali.

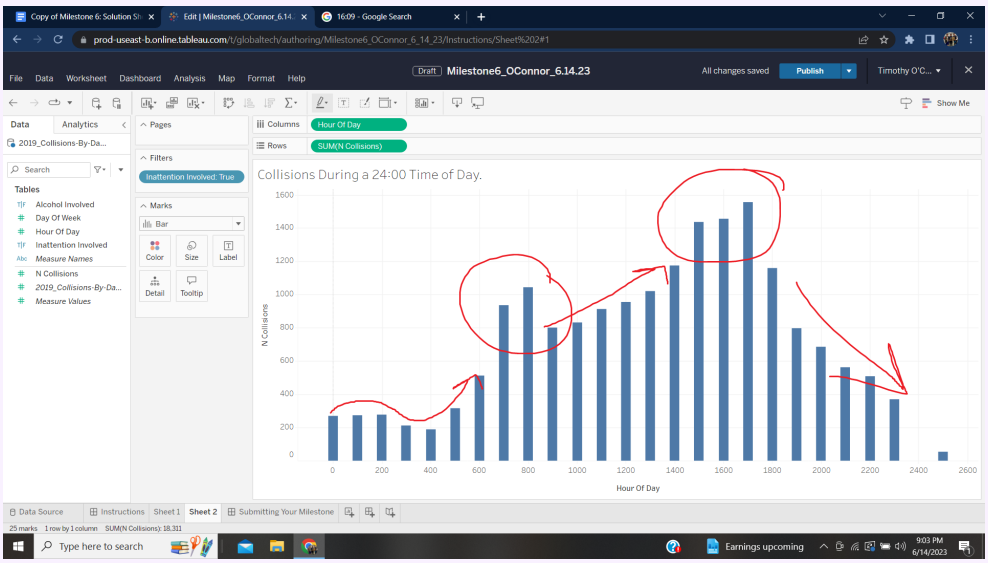
B. Copy the chart into a new sheet and add a filter so that the bar chart only shows accidents where the party at fault was found to be under the influence of alcohol. How does this distribution of accidents by time of day compare to the overall distribution?



The highest volume of collisions happens during 5pm to 2am. Peaks appear at 12am-2am and 8pm-11pm. The questionable and more saddening detail I gathered from this graph is that there are accidents occurring in the late morning and afternoon.

Alcoholism hinders the lives of so many people and it's sad to see data that shows people allowing themselves to get into accidents.

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?

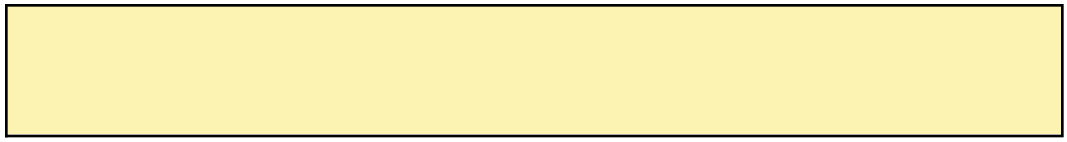


The peaks agree with my expectations. There are a lot of collisions during the hours of 7am-8am when people are driving to work, and a steady incline to its peak at 5pm-7pm when people are driving back home.

– 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!



— 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.