Traffic Collisions in California

Data Set **Description**

The original collisions table has 469 664 rows and 76 columns, but I'll only be focusing on the following four columns:

- **`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

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

To start analyzing the parties table, I need to understand how fault is attributed in accidents. Specifically, I'm looking for how many people are marked as "at fault" for collisions. To do this, I'll run a SQL query that counts the total number of people marked with a "Y" (for "Yes") in the `at_fault` column. This will give me the number of

individuals who were determined to be at fault for collisions based on the available data.

```
SELECT
  COUNT(*) AS total_fault
FROM
  switrs.parties
WHERE
  at_fault = 'Y'
```

There are a total of **438,491** collisions where the person is indeed at fault.

To further refine my analysis, I want to know how many parties were found at fault while under the influence of alcohol. In the dataset, the `party_sobriety` field is encoded with a 'B' when the individual involved had been drinking. I'll modify my previous query to not only check for those marked as "at fault" (`at_fault = 'Y'`), but also filter for those whose `party_sobriety` field is equal to 'B', indicating they were under the influence of alcohol.

```
SELECT
  COUNT(*) AS total_fault
FROM
  switrs.parties
WHERE
  at_fault = 'Y'
  AND party_sobriety = 'B'
```

There were **33,512** cases where the party who was found at fault was also under the influence of alcohol.

Now, I want to analyze how many parties were found at fault where inattention played a role in the collision. In the dataset, the `oaf_1` or `oaf_2` fields take on a value of 'F' when inattention was a contributing factor. I'll modify the previous query to include a condition that checks whether either `oaf_1` or `oaf_2` is equal to 'F', in addition to the `at_fault` condition.

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

There were **18,311** cases where the party who was found at fault was also guilty of having a lack of attention.

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

To understand how collisions vary by day of the week, I'll start by querying the `collisions` table and grouping the results by the `day_of_week` column. This will allow me to see how many collisions occurred on each day, with the day of the week encoded as 1 for Monday through 7 for Sunday.

```
SELECT
  day_of_week,
  COUNT(*) as num_collisions
FROM
  switrs.collisions
GROUP BY
  day_of_week
ORDER BY
  day_of_week
```

The day with the **most collisions** is **Friday** (5) with a total of **75,654** collisions. The day with the **least amount** of collisions is **Sunday** (7) with a total of **55,159** collisions.

Next, I'm going to join the `collisions` and `parties` tables using the `case_id` column, which they both share. By performing an inner join, I'll get only the rows where there's a match between the two tables. This will allow me to see all the parties involved in each collision. I'll then count the number of rows in the combined output to understand how many records the join creates.

```
SELECT
  COUNT(*) AS n_collisions
FROM
  switrs.collisions AS c
INNER JOIN switrs.parties AS p
ON c.case_id = p.case_id
```

There are 940,216 rows in the joined table. Which is actually equal to the number of rows in the parties table.

To dive deeper into the data, I'll combine the queries from earlier parts A and B to look at the number of collisions grouped by the day of the week. This time, I'll add a condition to ensure that only accidents where the party was both found to be at fault and under the influence of alcohol are counted. By doing this, I'll be able to see how these types of collisions vary throughout the week and identify which days have the most and fewest such incidents.

```
SELECT
  c.day_of_week,
  COUNT(*) AS total_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
```

The two days with the highest number of collisions are Saturday and Sunday at 7,523 and 7,603 respectively. The two days with the least amount of collisions are Tuesday and Wednesday at 3,070 and 3,189 respectively.

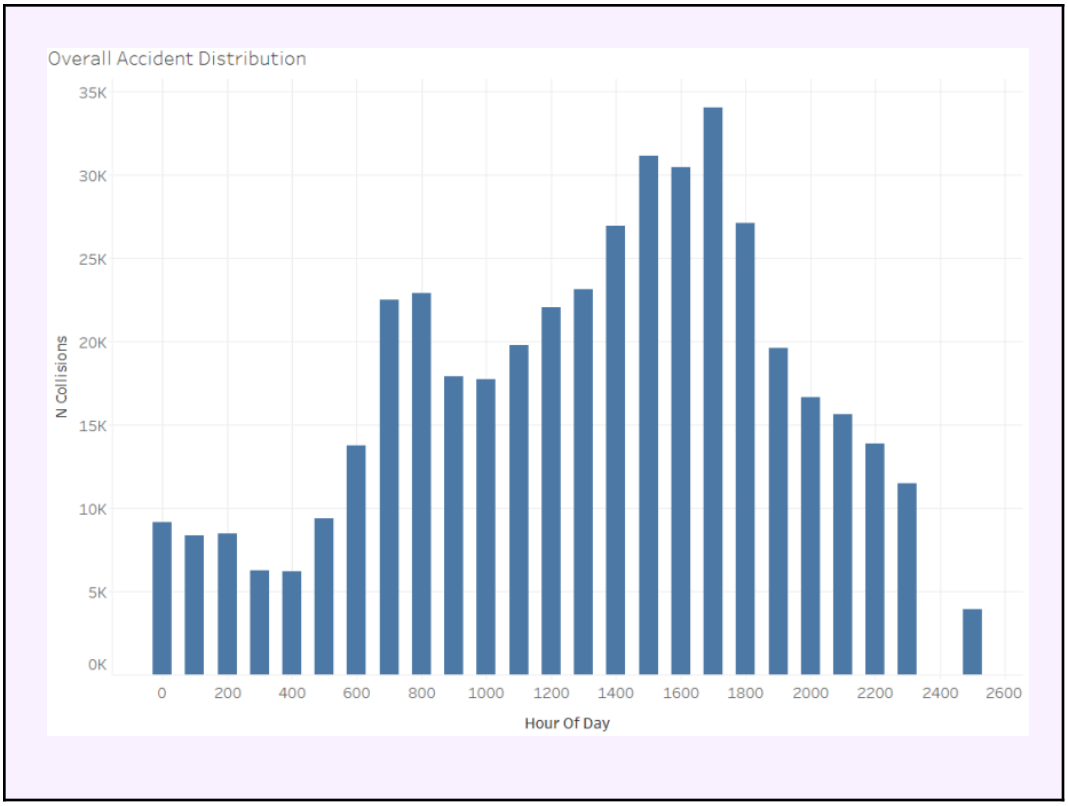
Next, I'll modify my query to focus on the number of accidents by the day of the week where the party was found to be at fault and inattention was a contributing factor. This will help identify how often inattention plays a role in accidents on different days of the week, giving us insight into when such collisions are most and least frequent.

```
SELECT
  c.day_of_week,
  COUNT(*) AS total_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'
    OR (
      p.oaf_1 = 'F'
      OR p.oaf_2 = 'F'
    )
  )
GROUP BY
  c.day_of_week
ORDER BY
  c.day_of_week ASC;
```

The two days with the highest number of collisions are Saturday and Sunday at 9,633 and 9,478 respectively. The two days with the least amount of collisions are Tuesday and Wednesday at 5,767 and 5,884 respectively.

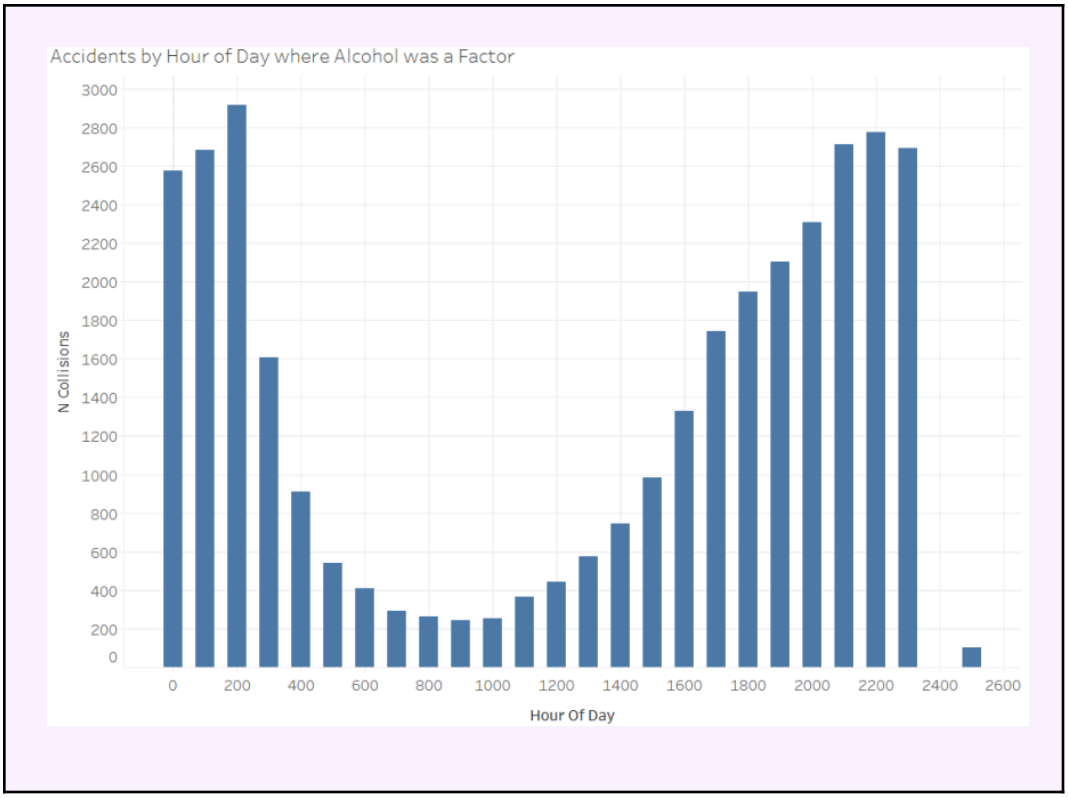
Task 3: When do accidents occur by the time of day?

Now I am going to create a bar chart of the number of collisions by the hour of day. Creating a visualization will allow for an easier understanding of the data.



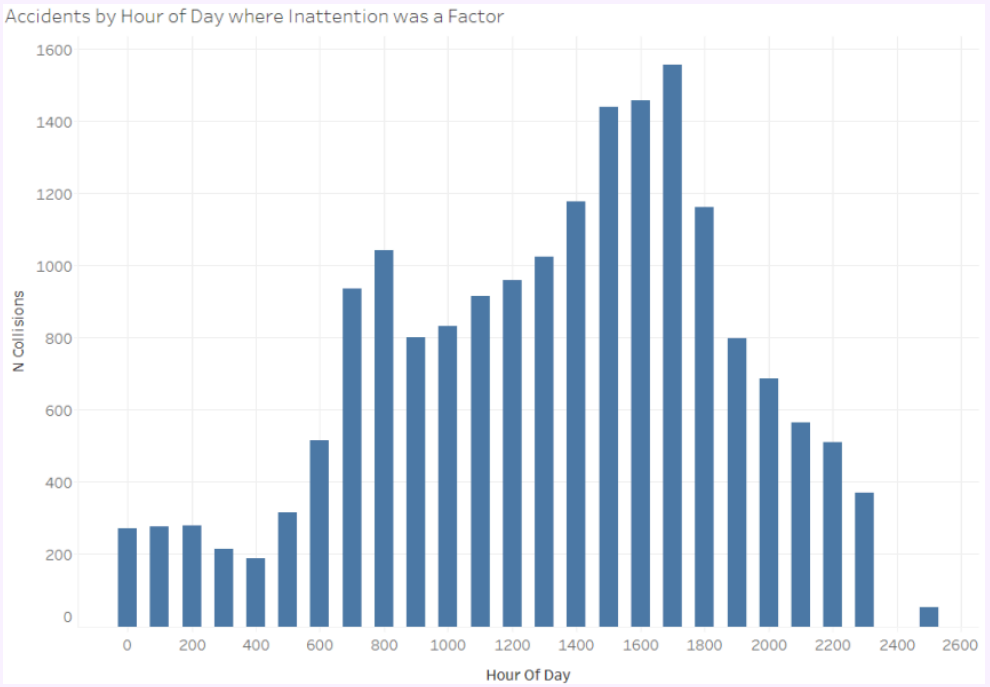
There are certain times in the day where accidents happen more frequently and quite honestly that makes sense. It happens more during the morning and during the noon-evening. These times are usually when people are driving to work and/or back home.

Now I am going to 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.



The graph shows an even bigger gap and is much more distinctive with the times. It shows that early morning and later in the day to late into the night is when the most accidents that are influenced by alcohol occur.

After seeing that, I am now going to change the filter to only look at accidents where inattention was a factor from the party-at-fault.



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!

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.

```
SELECT
```

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

We originally found 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