

## **MIS 6309.503 - Business Data Warehousing - F24**

### **Final Project**

#### **Members**

Anuja Thakar AXT230040 anujaabhay.thakar@utdallas.edu

Siddhi Javalkar SXJ230040 siddhinarendra.javalkar@utdallas.edu

**Business Vertical:** Sales Department

#### **Business Questions:**

1. Which vehicle type has highest no. of violations?
2. How many accidents occurred causing property damage?
3. What are the top 5 search reasons?
4. Which is the highest violation type?
5. Which are the Top 5 states with highest no. of arrests?
6. How many accidents occurred involving alcohol?
7. How many accidents occurred due to not wearing a seat belt?
8. Which are the Top 5 driver states with highest no. of arrests?

**Data:** Traffic Violations

**Data Source:** <https://catalog.data.gov/dataset/traffic-violations>

#### **ELT Process:**

1. Extract:  
Export .csv file from data source data.texas.gov  
Analyze file to create import .csv files
2. Transform:  
Create tables and attributes in Postgres SQL (Create queries)  
Data cleaning: Updated inconsistencies for duplicates, missing data, and unknown states  
(Update queries) using Chatgpt  
Data joining and Data normalization is done for tables for easy information access and standardizing date format
3. Load  
Import .csv files of data for each table

**Schema:** STAR

**Fact Table:** violations\_fact (violations registered)

#### **Dimension Tables:**

agency\_dim (captured by agency)

sub\_agency\_dim (captured by sub-agency)

arrest\_type\_dim (type of arrest made)

state\_dim (state in which arrest was made)

driver\_state\_dim (state to which driver belongs)

dl\_state\_dim (state which the driver's license belong to)

search\_type\_dim (type of search conducted by officer)

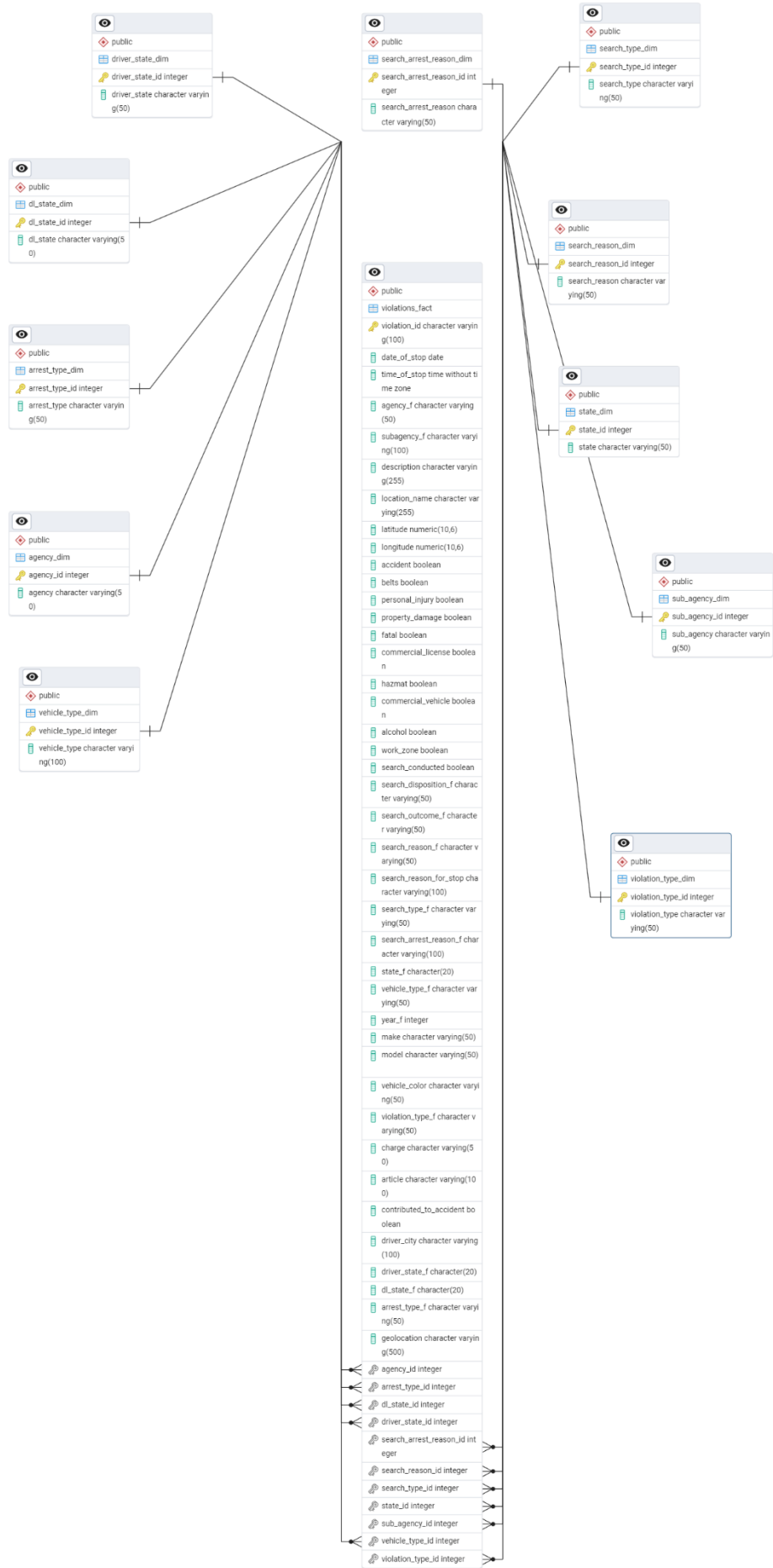
search\_reason\_dim (reason due to which search was conducted)

search\_arrest\_reason\_dim (arrested due to)

vehicle\_type\_dim (type of vehicle which the driver was driving)

violation\_type\_dim (type of violation)

**ERD:**



## Data Cleaning:

FTS Dictionaries

FTS Parsers

FTS Templates

Foreign Tables

Functions

Materialized Views

Operators

Procedures

Sequences

Tables (13)

agency\_dim

arrest\_type\_dim

dl\_state\_dim

driver\_state\_dim

search\_arrest\_reason\_dim

```
118
119 -----Data cleaning using update queries for the fact table-----
120 ----- Description column: handling invalid and missing values-----
121
122 ✓ update violations_fact
123     set description = 'Unknown'
124     where (description is null or description ~ '^[0-9]+$');
125
126
127
```

Data Output Messages Notifications

UPDATE 8

Query returned successfully in 1 secs 868 msec.

utions

les

s

s

/views

n

\_dim

m

> dim

```
103 -----Data cleaning using update queries for the fact table-----
104 ----- Description column: handling invalid and missing values-----
105
106 ✓ update violations_fact
107     set description = 'Unknown'
108     where (description is null or description ~ '^[0-9]+$');
109
110     select * from violations_fact;
111
112 ----- search_conducted column: handle missing values-----
113
114 ✓ update violations_fact
115     set search_conducted = 'No'
116     where search_conducted is null;
117
```

Data Output Messages Notifications

UPDATE 392494

Query returned successfully in 24 secs 6 msec.

tes

les

Views

im

be\_dim

dim

ite\_dim

rrest\_reason\_dim

utcome\_dim

ason\_dim

pe\_dim

i

```
117
118
119 -----search_disposition_f column: handle missing values-----
120
121 ✓ update violations_fact
122     set search_disposition_f = 'Nothing'
123     where search_disposition_f is null;
124
```

Data Output Messages Notifications

UPDATE 788879

Query returned successfully in 36 secs 409 msec.

Tables  
Foreign Tables  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (12)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim  
search\_arrest\_reason\_dim  
search\_outcome\_dim

```
125  
126 --- search_outcome_f column: handle missing values-----  
127  
128 v update violations_fact  
129 set search_outcome_f = 'Unknown'  
130 where search_outcome_f is null;
```

Data Output Messages Notifications

UPDATE 369197

Query returned successfully in 8 secs 264 msec.

FTS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (12)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim

```
130 where search_outcome_f is null;  
131  
132  
133 ---- search_reason_f column: handle missing values-----  
134  
135 v update violations_fact  
136 set search_reason_f = 'Unknown'  
137 where search_reason_f is null;
```

Data Output Messages Notifications

UPDATE 788878

Query returned successfully in 19 secs 888 msec.

FTS Templates  
Foreign Tables  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (12)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim  
search\_arrest\_reason\_dim

```
138  
139  
140 ---- search_reason_for_stop column: handle missing values----  
141  
142 v update violations_fact  
143 set search_reason_for_stop = 'Unknown'  
144 where search_reason_for_stop is null;
```

Data Output Messages Notifications

UPDATE 352693

Query returned successfully in 14 secs 709 msec.

Tables  
Foreign Tables  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (12)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim  
search\_arrest\_reason\_dim

```
146  
147 --- search_type_f column: handle missing values-----  
148  
149 v update violations_fact  
150 set search_type_f = 'Unknown'  
151 where search_type_f is null;
```

Data Output Messages Notifications

UPDATE 788878

Query returned successfully in 23 secs 424 msec.

les  
d Views

Jim  
pe\_dim  
dim  
ate\_dim  
rrest\_reason\_dim

```
152  
153  
154 --- search_reason_f column: handle missing values-----  
155  
156 update violations_fact  
157 set search_arrest_reason_f = 'Other'  
158 where search_arrest_reason_f is null;
```

Data Output Messages Notifications

UPDATE 794887

Query returned successfully in 14 secs 655 msec.

f Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (12)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim  
search\_arrest\_reason\_dim

```
160  
161 ---- state_f column: handle missing and invalid values----  
162  
163 update violations_fact  
164 set state_f = 'Unknown'  
165 where (state_f is null or state_f = 'XX');
```

Data Output Messages Notifications

UPDATE 7882

Query returned successfully in 1 secs 685 msec.

oreign Tables  
unctions  
aterialized Views  
erators  
ocedures  
equences  
bles (13)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim

```
183 ---- year_f column: handle invalid and null values----  
184  
185 update violations_fact  
186 set year_f = 00  
187 where (year_f is null or year_f = 0);
```

Data Output Messages Notifications

UPDATE 4743

Query returned successfully in 3 secs 714 msec.

FIS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (13)  
agency\_dim  
arrest\_type\_dim  
dl\_state\_dim  
driver\_state\_dim  
search\_arrest\_reason\_dim

```
189  
190 ----model column: handle invalid and null values-----  
191 update violations_fact  
192 set model = 'Unknown'  
193 where (model is null or model = '0');  
194
```

Data Output Messages Notifications

UPDATE 42

Query returned successfully in 3 secs 381 msec.

FTS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (13)  
  > agency\_dim  
  > arrest\_type\_dim  
  > dl\_state\_dim  
  > driver\_state\_dim  
  > search\_arrest\_reason\_dim

```
195  
196 --- article column: handle missing values----  
197  
198 v update violations_fact  
199   set article = 'Unknown'  
200   where article is null;
```

Data Output Messages Notifications

UPDATE 2312

Query returned successfully in 3 secs 734 msec.

Parsers  
Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (13)  
  > agency\_dim  
  > arrest\_type\_dim  
  > dl\_state\_dim

```
201  
202 --- driver_city: handle invalid and missing values----  
203  
204 v update violations_fact  
205   set driver_city = 'Unknown'  
206   where (driver_city is null or driver_city = '0')  
207
```

Data Output Messages Notifications

UPDATE 72

Query returned successfully in 2 secs 14 msec.

FTS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (13)  
  > agency\_dim  
  > arrest\_type\_dim  
  > dl\_state\_dim

```
208  
209 --- driver_state_f column: handling invalid and null values----  
210  
211 update violations_fact  
212   set driver_state_f = 'Unknown'  
213   where (driver_state_f is null or driver_state_f = 'XX');
```

Data Output Messages Notifications

UPDATE 421

Query returned successfully in 1 secs 887 msec.

FTS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (13)  
  > agency\_dim  
  > arrest\_type\_dim  
  > dl\_state\_dim  
  > driver\_state\_dim  
  > search\_arrest\_reason\_dim

```
215  
216 ---- dl_state_f column: handling invalid and null values-----  
217  
218 v update violations_fact  
219   set dl_state_f = 'Unknown'  
220   where (dl_state_f is null or dl_state_f = 'XX');
```

Data Output Messages Notifications

UPDATE 20095

Query returned successfully in 4 secs 91 msec.

stors  
rdures  
ences  
s(13)  
ency\_dim  
est\_type\_dim  
state\_dim  
ver\_state\_dim  
arch\_arrest\_reason\_dim  
arch\_outcome\_dim  
arch\_reason\_dim  
arch\_type\_dim  
ite\_dim  
b\_agency\_dim

222  
223  
224  
225  
226  
227  
228  
229  
230

-----Update invalid states in violations\_fact table-----  
  
update violations\_fact  
set state\_f = 'UNKNOWN',  
dl\_state\_f = 'UNKNOWN',  
driver\_state\_f = 'UNKNOWN'  
where state\_f in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US')  
or dl\_state\_f in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US')  
or driver\_state\_f in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');

Data Output Messages Notifications

UPDATE 1238  
  
Query returned successfully in 8 secs 345 msec.

232  
233  
234  
235  
236  
237  
238

-----Update invalid states for dimension tables-----  
  
update state\_dim  
set state = 'Unknown'  
where state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');

Data Output Messages Notifications

UPDATE 12  
  
Query returned successfully in 86 msec.

232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243

-----Update invalid states for dimension tables-----  
  
update state\_dim  
set state = 'Unknown'  
where state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');  
  
update dl\_state\_dim  
set dl\_state = 'Unknown'  
where dl\_state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');

Data Output Messages Notifications

UPDATE 12  
  
Query returned successfully in 73 msec.

231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247

-----Update invalid states for dimension tables-----  
  
update state\_dim  
set state = 'Unknown'  
where state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');  
  
update dl\_state\_dim  
set dl\_state = 'Unknown'  
where dl\_state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');  
  
update driver\_state\_dim  
set driver\_state = 'Unknown'  
where driver\_state in ('AB', 'BC', 'MH', 'NB', 'NF', 'NU', 'NV', 'ON', 'PE', 'PQ','QC','US');

Data Output Messages Notifications

UPDATE 12  
  
Query returned successfully in 86 msec.



## Data Analysis:

1.

```
246
247
248 ----- Find the vehicle type with the highest violations-----
249
250 select
251     vtd.vehicle_type,
252     count(*) as total_violations
253 from
254     violations_fact vf
255 join vehicle_type_dim vtd on vf.vehicle_type_id = vtd.vehicle_type_id
256 group by
257     vtd.vehicle_type
258 order by
259     total_violations desc
260 limit 1;
261
```

Data Output Messages Notifications

	vehicle_type character varying (100)	totalViolations bigint
1	08 - Recreational Vehicle	30191

2.

Views  
Data Wrappers  
ges  
tions  
as (1)  
ic  
ggregates  
ollations  
omains  
FS Configurations  
FS Dictionaries  
FS Parsers  
FS Templates  
oreign Tables  
unctions  
aterialized Views  
entences

```
258 -----Number of accidents caused due to alcohol-----
259
260 select
261     count(*) as total_accidents_with_alcohol
262 from
263     violations_fact
264 where
265     alcohol = true
266     and accident = true;
267
```

Data Output Messages Notifications

	total_accidents_with_alcohol bigint
1	160

3.

guages  
ilications  
emas (1)  
ublic  
Aggregates  
ollations  
Domains  
FS Configurations  
FS Dictionaries  
FS Parsers  
FS Templates  
Foreign Tables  
Functions  
aterialized Views  
Operators

```
270 -----Number of accidents caused due to not wearing seat belt-----
271
272 select
273     count(*) as total_accidents_without_belts
274 from
275     violations_fact
276 where
277     belts = FALSE and accident = TRUE;
278
```

Data Output Messages Notifications

	total_accidents_without_belts bigint
1	40176

4.

276  
277 -----Number of accidents that caused property damage-----  
278 **select**  
279     **count**(\*) **as** total\_accidents\_with\_property\_damage  
280 **from**  
281     violations\_fact  
282 **where**  
283     property\_damage = true  
284     **and** accident = true;  
285

Data Output   Messages   Notifications

total\_accidents\_with\_property\_damage  
bigint

1	15939
---	-------

5.

283  
284 -----Top 5 States with highest number of arrests-----  
285  
286 **select**  
287     sd.state,  
288     **count**(\*) **as** total\_arrests  
289 **from**  
290     violations\_fact vf  
291 **join** state\_dim sd **on** vf.state\_id = sd.state\_id  
292 **where**  
293     vf.arrest\_type\_f **is not null**  
294 **group by**  
295     sd.state  
296 **order by**  
297     total\_arrests **desc**  
298 **limit** 5;  
299

Data Output   Messages   Notifications

	state character varying (50)	total_arrests bigint
1	Unknown	162035
2	IA	12859
3	KS	12784
4	TN	12774
5	NY	12771

6.



sub\_agency\_dim  
vehicle\_type\_dim  
violation\_type\_dim  
violations\_fact  
gger Functions  
es  
ws  
ctions

s  
iggers  
ins  
Data Wrappers  
es  
ions  
s (1)  
>  
gregates  
lations  
mains  
S Configurations  
S Dictionaries  
S Parsers

```
335  
336 -----Find the top 5 search reasons-----  
337  
338 select  
339     sr.search_reason,  
340     count(*) as total_searches  
341 from  
342     violations_fact vf  
343 join search_reason_dim sr on vf.search_reason_id = sr.search_reason_id  
344 where  
345     vf.search_conducted = TRUE  
346 group by  
347     sr.search_reason  
348 order by  
349     total_searches desc  
350 limit 5;
```

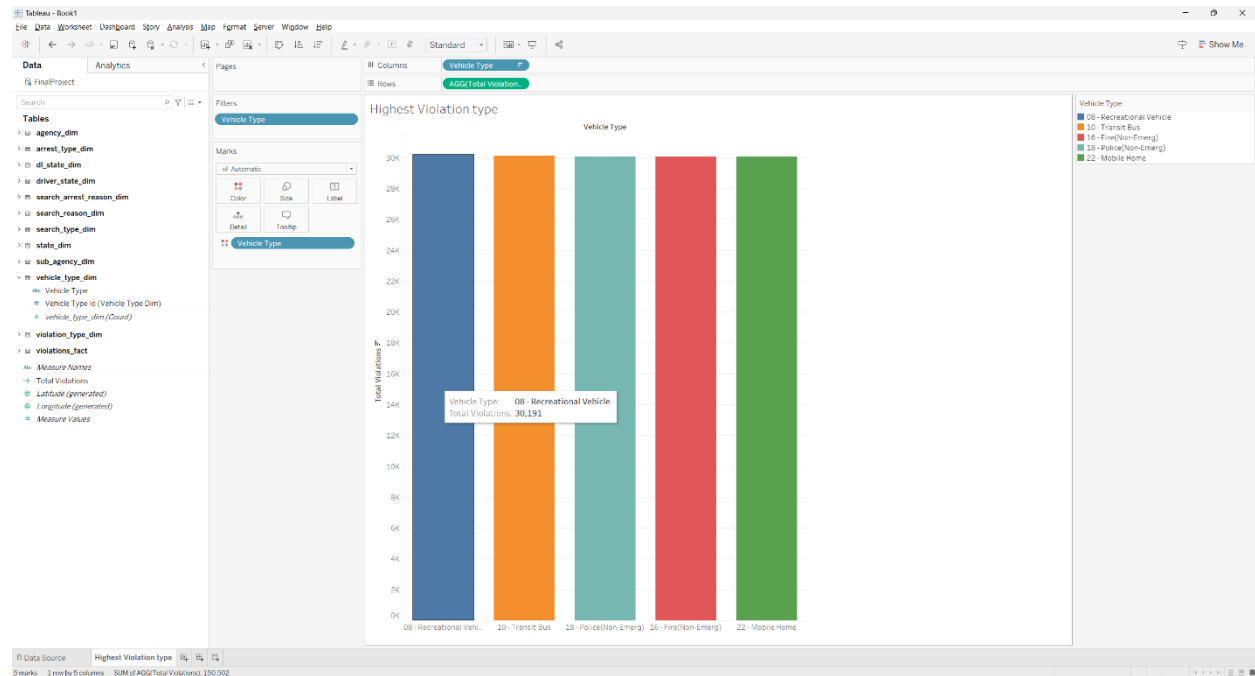
Data Output Messages Notifications

	search_reason character varying (50)	total_searches bigint
1	Exigent Circumstances	6448
2	Probable Cause	6415
3	Probable Cause for CDS	6387
4	K-9	6331
5	Arrest/Tow	6327

## REPORTS:

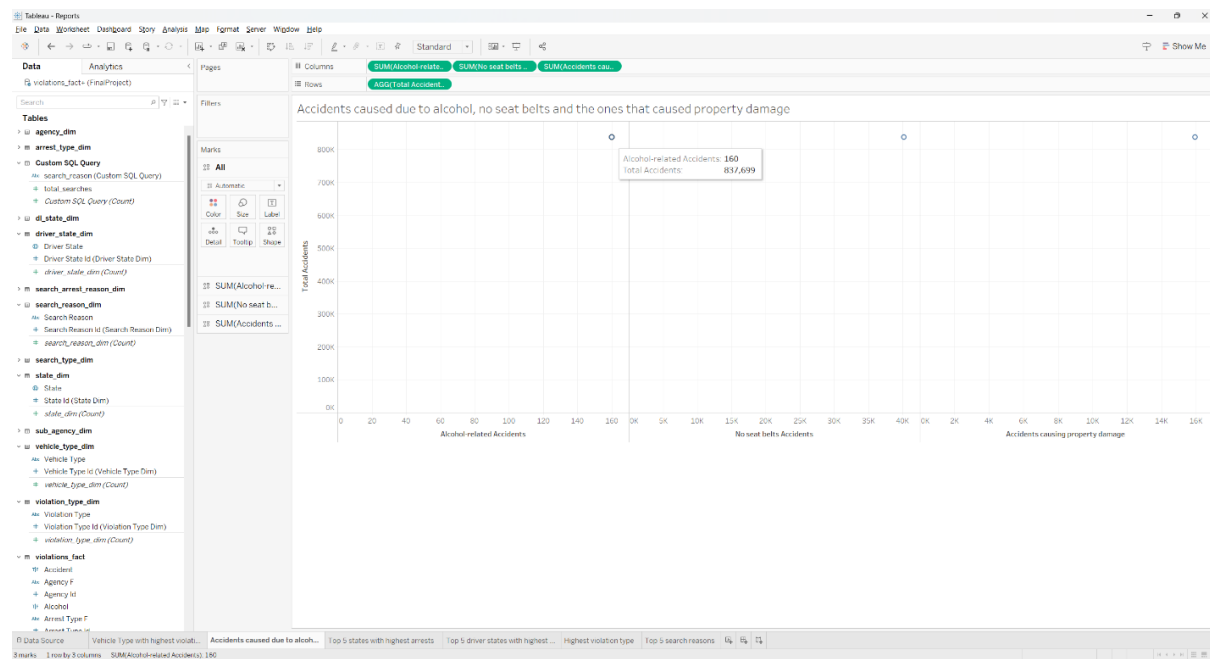
Question1: Which vehicle type has highest no. of violations?

Answer: 30,191 violations by Recreational vehicles



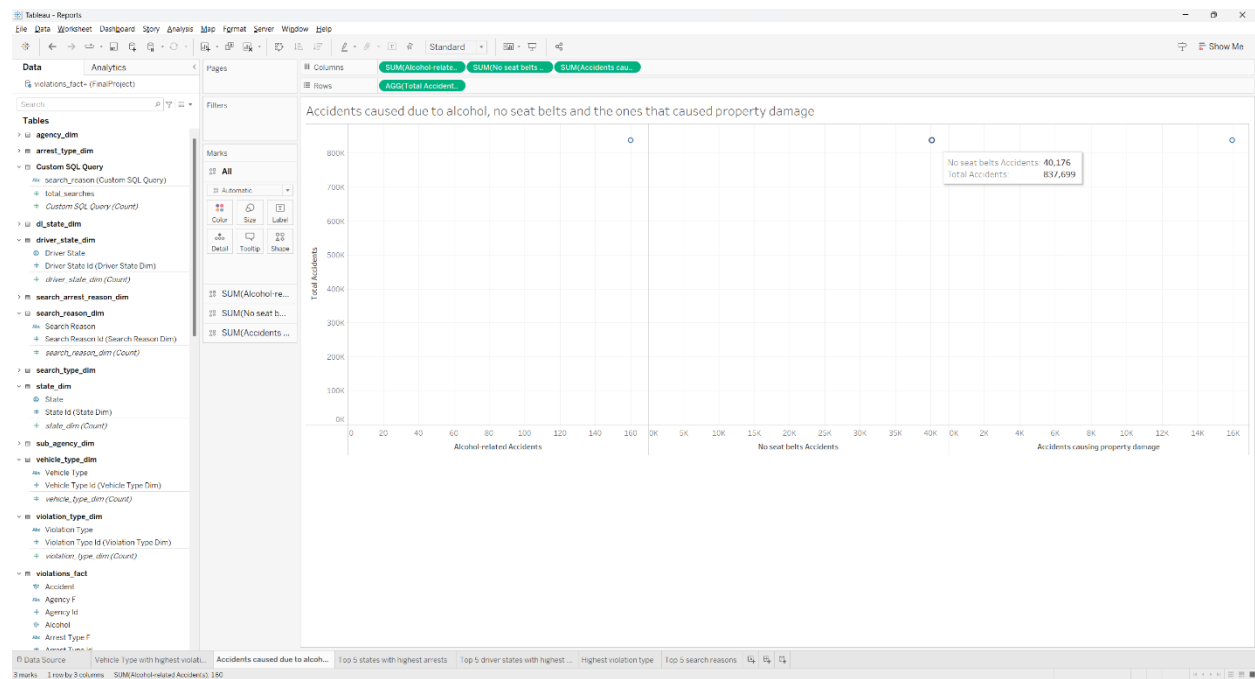
Question 2: How many accidents occurred involving alcohol?

Answer: 160 Accidents



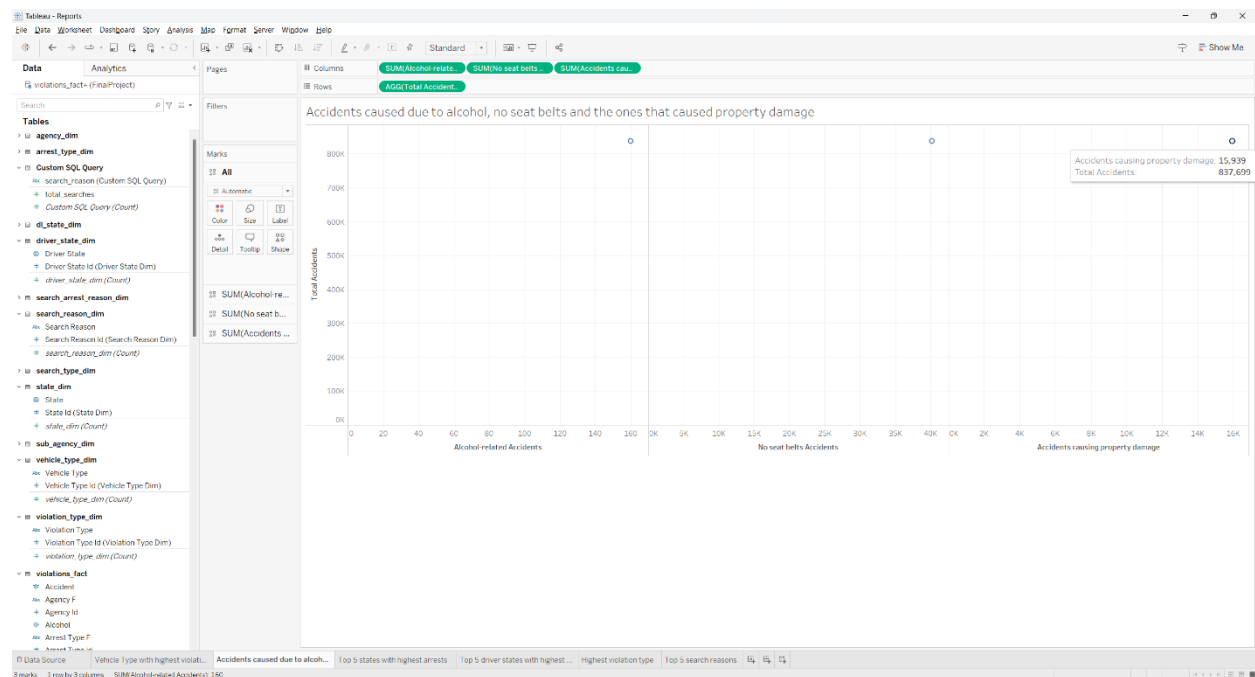
Questions 3: How many accidents occurred due to not wearing a seat belt?

Answer: 40,176 accidents



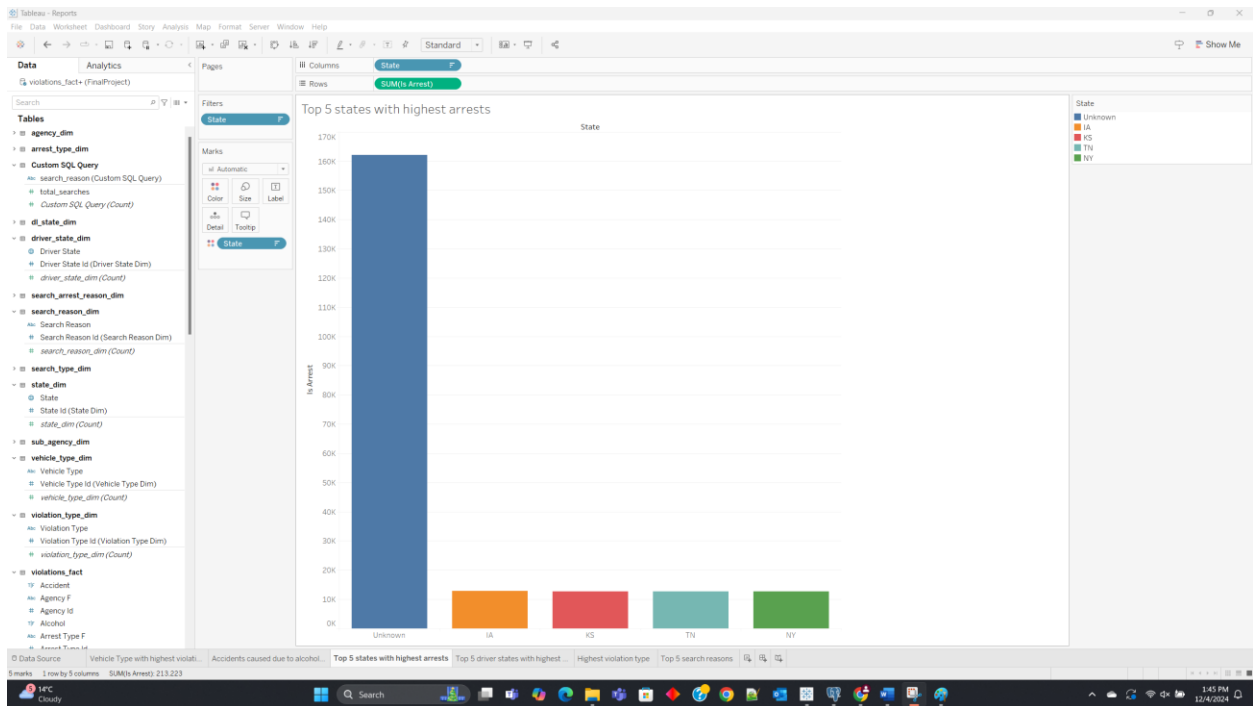
Questions 4: How many accidents occurred causing property damage?

Answer: 15,939 accidents



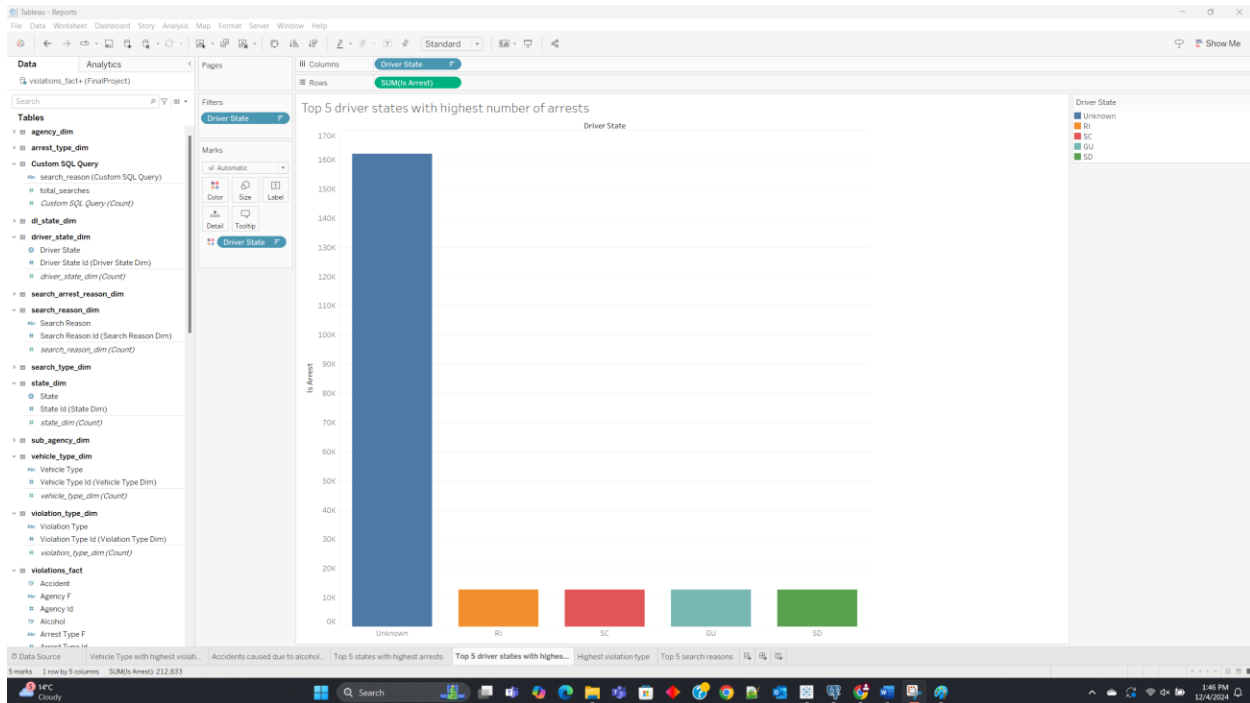
Question 5: Which are the Top 5 states with highest no. of arrests?

Answer: Iowa, Kansas, Tennessee and New York



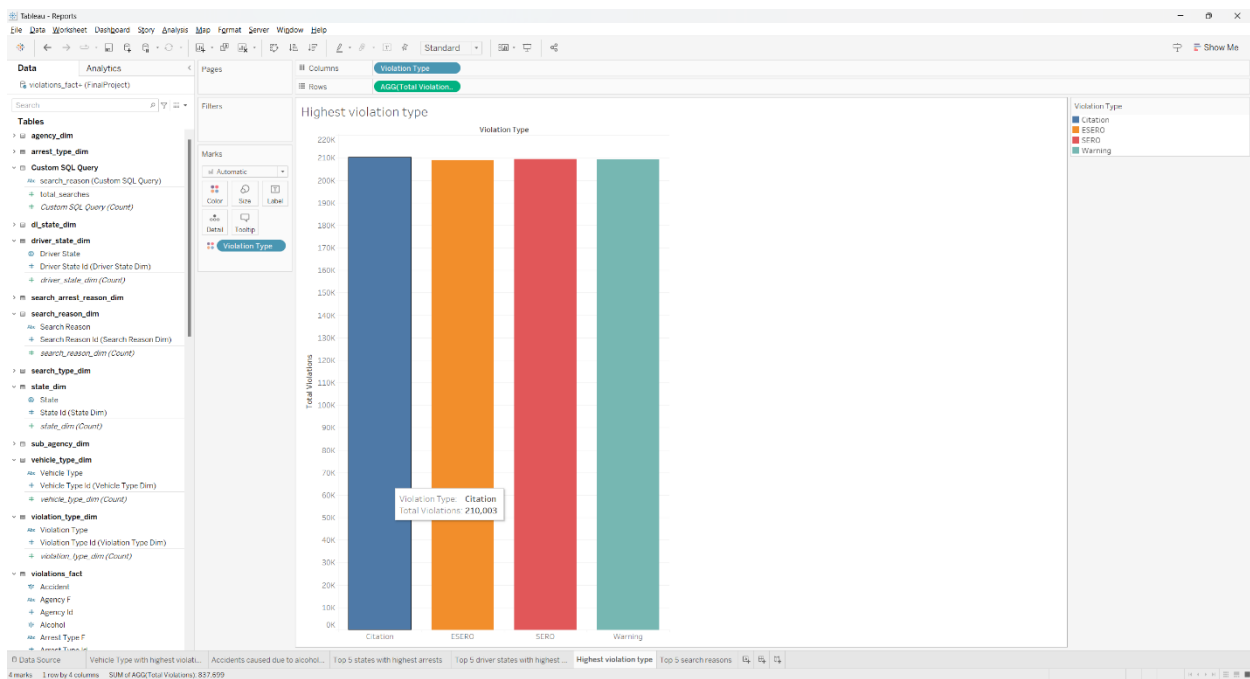
Question 6: Which are the Top 5 driver states with highest no. of arrests?

Answer: Rhode Island, South Carolina, Guam and South Dakota



Question 7. Which is the highest violation type?

Answer: Citation (210,003)



Question 8. What are the top 5 search reasons?



Answer: Exigent circumstances, Probable cause, Probable cause for CDS, K-9, Arrest/Tow

