

VEHICLE SALES

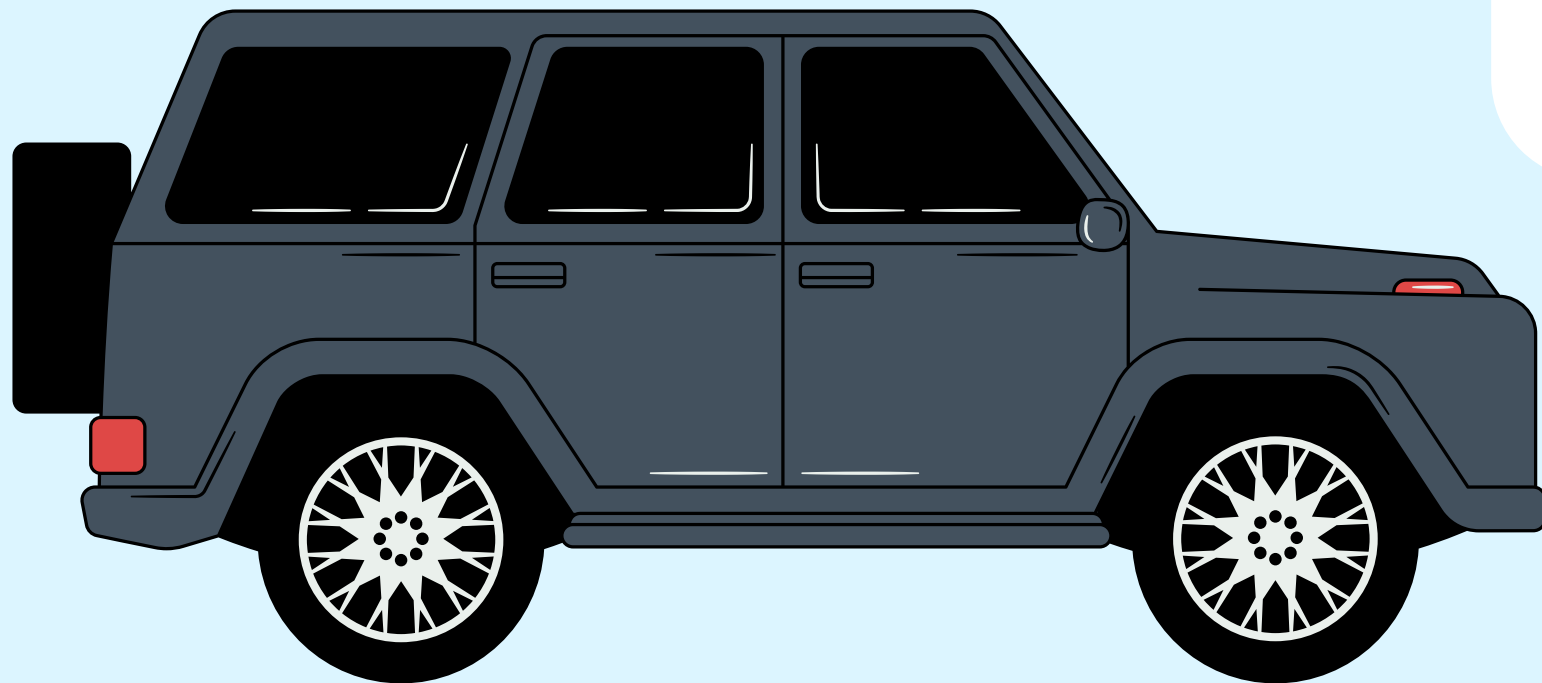
By Juan Granillo



PROBLEM STATEMENT

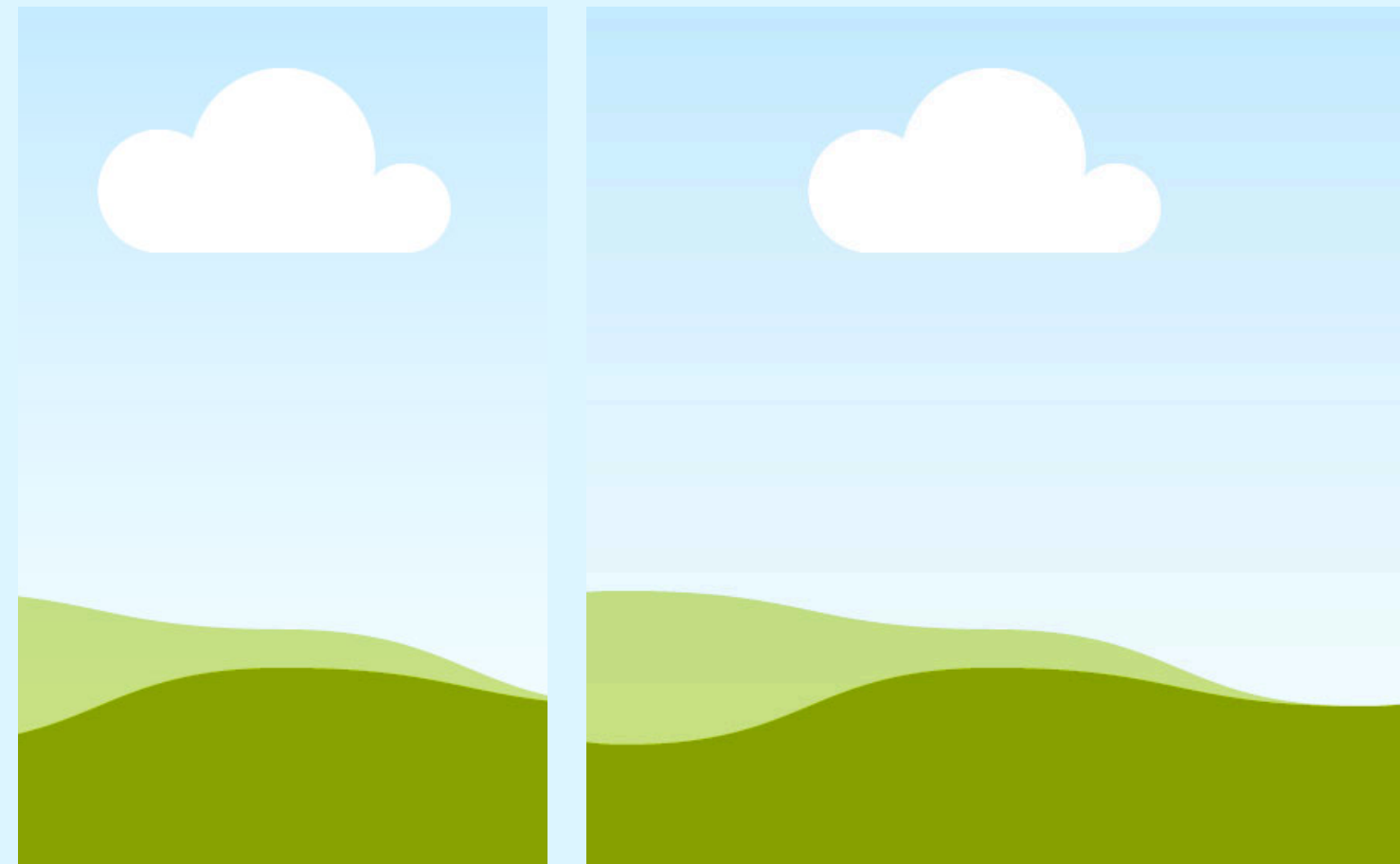
Understanding vehicle sales trends is critical for optimizing inventory, marketing strategies, and dealership management

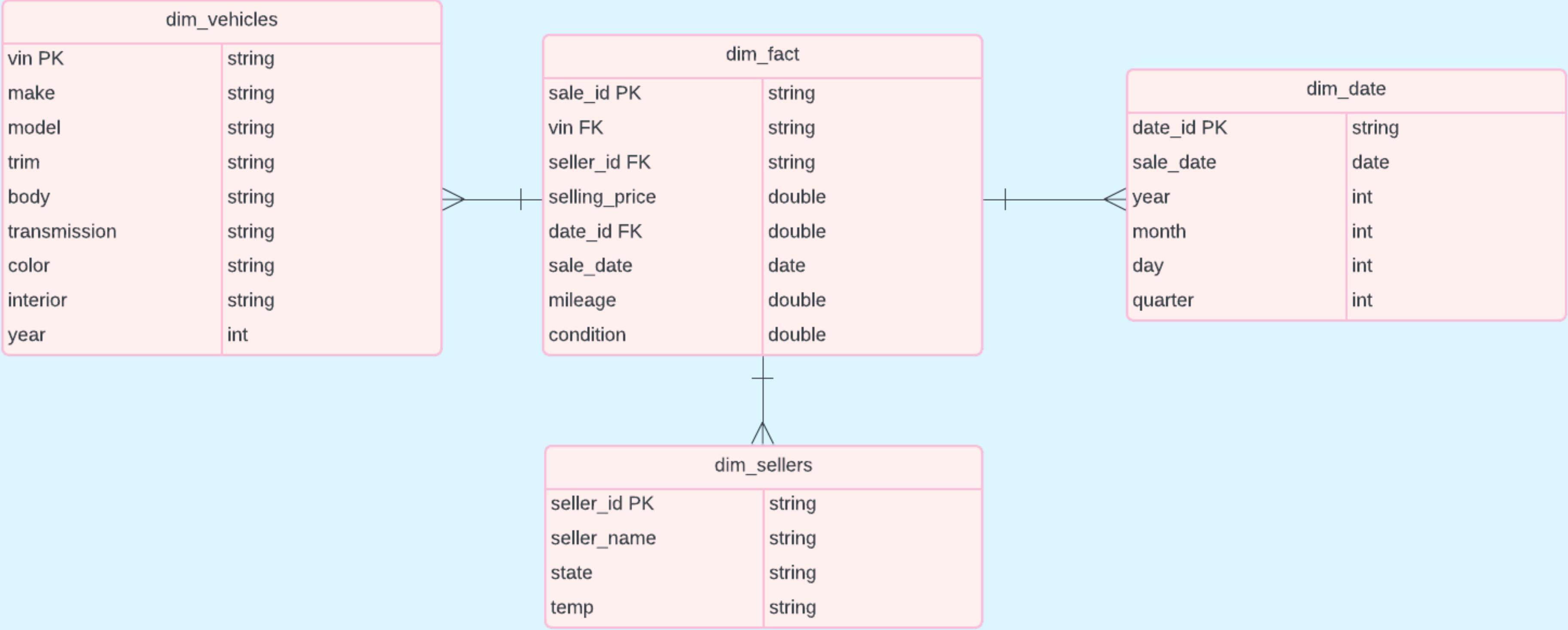
The challenge is to analyze these factors together and gain insights into how climate and vehicle characteristics influences vehicle sales patterns in different states

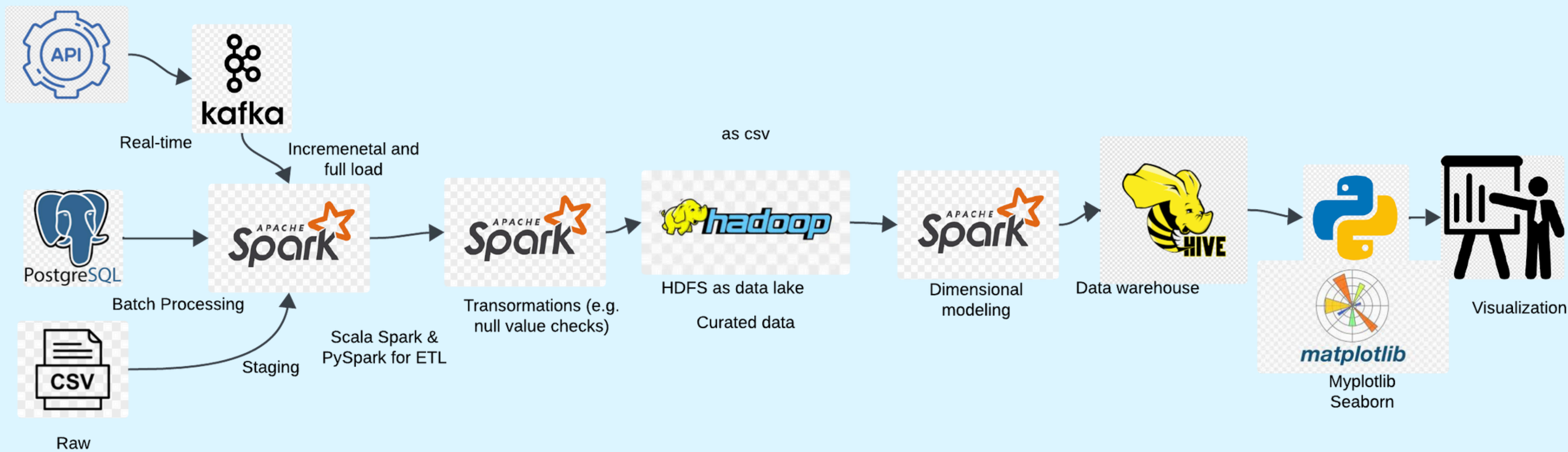


BUSINESS USE CASE

- Objective: To analyze how different climates and vehicle characteristics affect vehicle sales in different states.
- Outcome: Providing dealerships and manufacturers with data-driven insights to optimize inventory, sales strategies, and marketing campaigns based on state-specific temperatures
- Value: Improved decision-making for stock management, targeted promotions, and efficient resource allocation based on climate trends.







CDC

```
testdb=> select * from timestamp_table;
 id |          created_at
----+-----
  1 | 2004-05-01 12:00:00
 24 | 2024-10-16 03:15:04.141387
 25 | 2024-10-19 21:41:24.797246
(3 rows)
```

Increment load if vehicle sales >
than timestamp table

Incremental load

```
[ec2-user@ip-172-31-14-3 ~]$ hdfs dfs -ls /tmp/vehicle_sales_curated
Found 3 items
-rw-r--r--  3 ec2-user hadoop 0 2024-10-19 21:48 /tmp/vehicle_sales_curated/_SUCCESS
-rw-r--r--  3 jenkins hadoop 389 2024-10-19 23:29 /tmp/vehicle_sales_curated/part-00000-3d9d388b-7ad7-486c-ae59-6e879ae99355-c000.csv
-rw-r--r--  3 ec2-user hadoop 598807 2024-10-16 03:26 /tmp/vehicle_sales_curated/part-00000-7eaca917-75ac-4668-8491-ec68bb73cd1b-c000.csv
[ec2-user@ip-172-31-14-3 ~]$ hdfs dfs -ls /tmp/vehicle_sales_curat^C
```

Recent entries in new file

```
[ec2-user@ip-172-31-14-3 ~]$ hdfs dfs -ls /tmp/vehicle_sales_curated
[ec2-user@ip-172-31-14-3 ~]$ hdfs dfs -cat /tmp/vehicle_sales_curated/part-00000-3d9d388b-7ad7-486c-ae59-6e879ae99355-c000.csv
2022,toyota,camry,xse,sedan,automatic,4t1b11hxxnu123456,ca,82.0,15000,blue,black,toyota dealership,30000,32000,2024-01-12 10:45:00
2022,acura,tlx,a-spec,sedan,automatic,5j8tc1h58nl017239,ca,85.0,31500,black,red,acura of pasadena,29000,30500,2024-02-21 14:45:00
2021,ford,f-150,lariat,truck,automatic,1ftfw1e54mke12345,ny,75.0,25000,red,gray,ford dealership,40000,42000,2024-01-15 09:15:00
```



Original Data

year	make	model	trim	body	transmission	vin	stat
e	condition	odometer	color	interior	seller	mmr	sellingprice
	saledate		created_at				
2002					automatic	137fa90362e197965	nc
	25	79808	white	gray	performance auto center inc	47000	36000
Mon Feb 23 2015	01:30:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
2001	HUMMER	11		Wagon	SUV	137za84341e193591	ca
	21	65612	blue	gray	auto city sales/leasing	43400	45750
Thu May 28 2015	05:00:00	GMT-0700	(PDT)	2024-10-16 03:15:04.141387			
2000					automatic	137za8435ye187468	ca
	23	84028	silver	tan	aaero sweet company	40300	42000
Thu Feb 27 2015	05:30:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	Sedan	19uua5640xa034760	ga
	19	233154	black	tan	enterprise vehicle exchange / tra / rental / tulsa	1075	1050
Wed Feb 11 2015	16:00:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	Sedan	19uua5640xa053244	fl
	23	173310	blue	beige	coggin nissan	1325	2100
Thu Feb 05 2015	08:20:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	Sedan	19uua5640xa053826	oh
		113958	red	black	bargain wheels llc	1825	1600
Thu Dec 18 2014	10:05:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	Sedan	19uua5641xa040227	va
	2	130002	green	tan	purple heart	1975	2800
Thu Mar 05 2015	03:50:00	GMT-0800	(PST)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	sedan	19uua5641xa053415	pa
	19	153673	green	beige	r hollenshead auto sales inc	1500	1450
Fri Jun 05 2015	02:00:00	GMT-0700	(PDT)	2024-10-16 03:15:04.141387			
1999	Acura	TL		3.2	sedan	19uua5642xa002702	nj
	35	121985	gray	gray	wayne auto sales inc	1725	2000

--More--

Curated Data after ETL

```
0
2012,honda,civic,ex,sedan,automatic,19xfb2f89ce023740,wi,48.0,41961,silver,gray,international honda,11750,12500,2015-02-18 02:00:00
2006,chrysler,town and country,base,minivan,automatic,1a4gp45r06b731009,wi,19.0,123768,red,gray,dt credit corporation,2200,1500,2015-01-07 10:00:00
2006,chrysler,town and country,base,minivan,automatic,1a4gp45r16b551750,wi,19.0,140003,gold,gray,credit acceptance corp/vr,so uthfield,2025,1650,2015-02-18 02:00:00
2009,chrysler,aspen,limited,suv,automatic,1a8hw58p89f701582,wi,21.0,83435,black,gray,bredemann lexus in glenview,11350,12900,2015-02-25 02:00:00
[ec2-user@ip-172-31-14-3 ~]$
[ec2-user@ip-172-31-14-3 ~]$
```



- Null values, duplicates, outliers, invalid values dropped
- To lowercase
- Dates normalized

Tests report on curated data

Column: interior, Null Count: 0

Column: seller, Null Count: 0

Column: mmr, Null Count: 0

Column: sellingprice, Null Count: 0

Column: saledate, Null Count: 0

2. Duplicate Records Check:

Duplicate Record Count: 0

3. Negative Values Check (in numerical columns):

Column: year, Negative Value Count: 0

Column: condition, Negative Value Count: 0

Column: odometer, Negative Value Count: 0

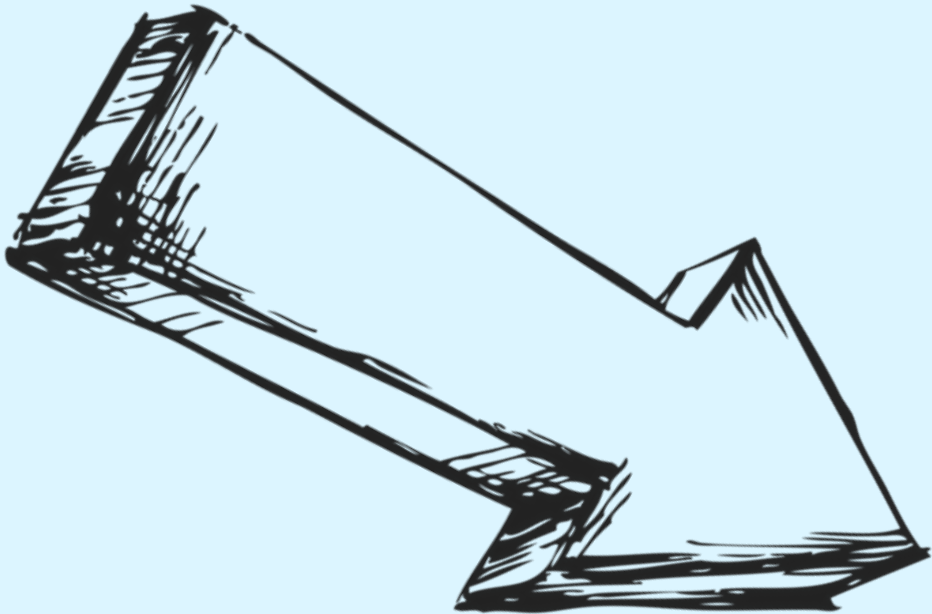
Column: mmr, Negative Value Count: 0

Column: sellingprice, Negative Value Count: 0

Real-time data via Kafka consumer to hdfs

```
Listening to partition 0 of topic avg_state_temp...
Received message: AL, 64.0
Received message: AK, 26.6
Received message: AZ, 72.3
Received message: AR, 61.7
Received message: CA, 59.4
Received message: CO, 45.6
Received message: CT, 51.9
Received message: DE, 55.5
Received message: FL, 70.7
Received message: GA, 64.5
Received message: HI, 70.0
```

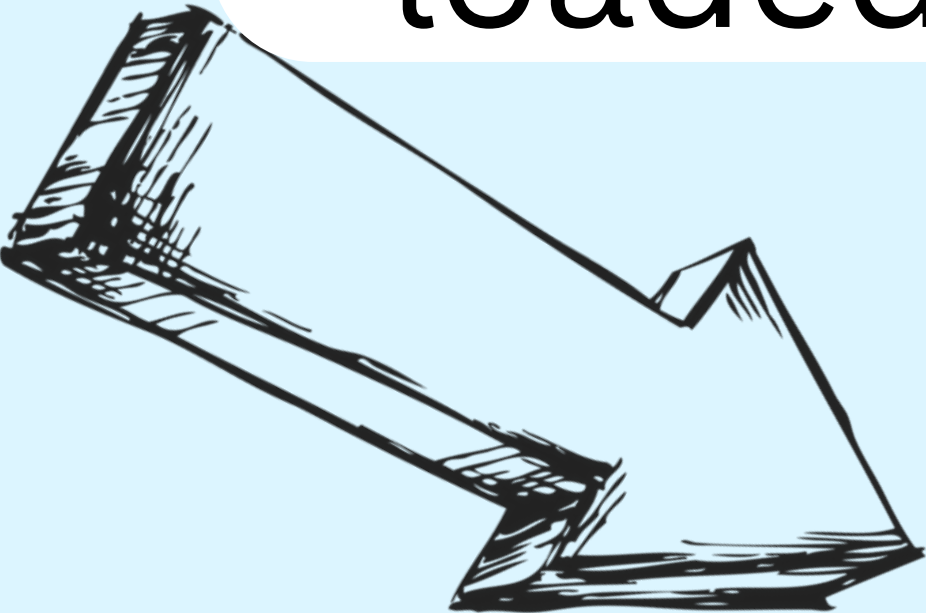
Topic: avg_state_temp



```
Found 101 items
-rw-r--r--  3 ec2-user hadoop      0 2024-10-18 01:02 /tmp/vehicle_temp_fin/_SUCCESS
-rw-r--r--  3 ec2-user hadoop      0 2024-10-18 01:01 /tmp/vehicle_temp_fin/part-00000-0dfcb9b8-9ede-4526-92e9-cbc2feb510
0e-c000.csv
-rw-r--r--  3 ec2-user hadoop      0 2024-10-18 01:02 /tmp/vehicle_temp_fin/part-00000-1db3957c-3b09-484a-8e7f-431e909c2e
16-c000.csv
-rw-r--r--  3 ec2-user hadoop      0 2024-10-18 01:01 /tmp/vehicle_temp_fin/part-00000-21e5d38f-6149-4adf-b01b-1516755fe0
45-c000.csv
-rw-r--r--  3 ec2-user hadoop      0 2024-10-18 01:01 /tmp/vehicle_temp_fin/part-00000-23c09766-946d-475e-aaf0-3a68f649cf
85-c000.csv
```

Dimensional models created with spark + loaded onto Hive

```
2015,1,7,1,2015-01-07T10:00:00.000Z,VWcT
2015,2,11,1,2015-02-11T02:00:00.000Z,c0bb
2015,2,25,1,2015-02-25T02:00:00.000Z,ftho
2015,2,25,1,2015-02-25T02:00:00.000Z,jWJe
2014,12,31,4,2014-12-31T10:00:00.000Z,LRgY
2015,2,25,1,2015-02-25T02:00:00.000Z,TPNa
2015,1,14,1,2015-01-14T02:00:00.000Z,B01F
2015,1,21,1,2015-01-21T02:00:00.000Z,04xq
2015,1,28,1,2015-01-28T02:00:00.000Z,GzbX
2015,1,21,1,2015-01-21T02:00:00.000Z,aVWc
2015,1,21,1,2015-01-21T02:00:00.000Z,cJqx
2015,2,18,1,2015-02-18T02:00:00.000Z,6Jnv
2015,1,7,1,2015-01-07T10:00:00.000Z,JTt4
2015,2,18,1,2015-02-18T02:00:00.000Z,5iNW
2015,2,25,1,2015-02-25T02:00:00.000Z,BW8p
```



dim_date	
date_id PK	string
sale_date	date
year	int
month	int
day	int
quarter	int

dim_date_vehicles.year	dim_date_vehicles.month	dim_date_vehicles.day	dim_date_vehicles.quarter	dim_date_vehicles.saledate	dim_date_vehicles.date_id
2015	1	28	1		2015-01-28
2015	1	28	1	6Trs	2015-01-28
2015	3	3	1	zafm	2015-03-03
2015	1	28	1	fdFV	2015-01-28
2015	3	4	1	piY4	2015-03-04

- Queries using PyHive + Hue
- Visualization using Matplotlib and SeaBorn

```
def plot_most_popular_body_per_state(df):  
  
    # Plot using Seaborn  
    plt.figure(figsize=(12, 6))  
    sns.barplot(x='state', y='body_count', hue='body', data=df, dodge=False)  
  
    # Add labels and title  
    plt.title('Most Popular Vehicle Body Type by State', fontsize=16)  
    plt.xlabel('State', fontsize=12)  
    plt.ylabel('Count of Most Popular Body Type', fontsize=12)  
  
    # Rotate x labels for better readability  
    plt.xticks(rotation=90)
```


Tech used

Python

Python used to created visuals using myplotlib and seaborn

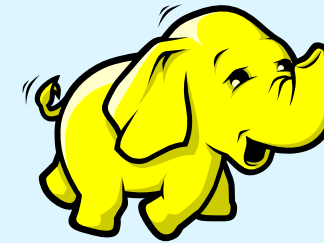


PySpark/Spark Scala + kafka



Spark scala used for batch processing and transforamations
Pyspark and kafka used for real-time data consumer

Hadoop



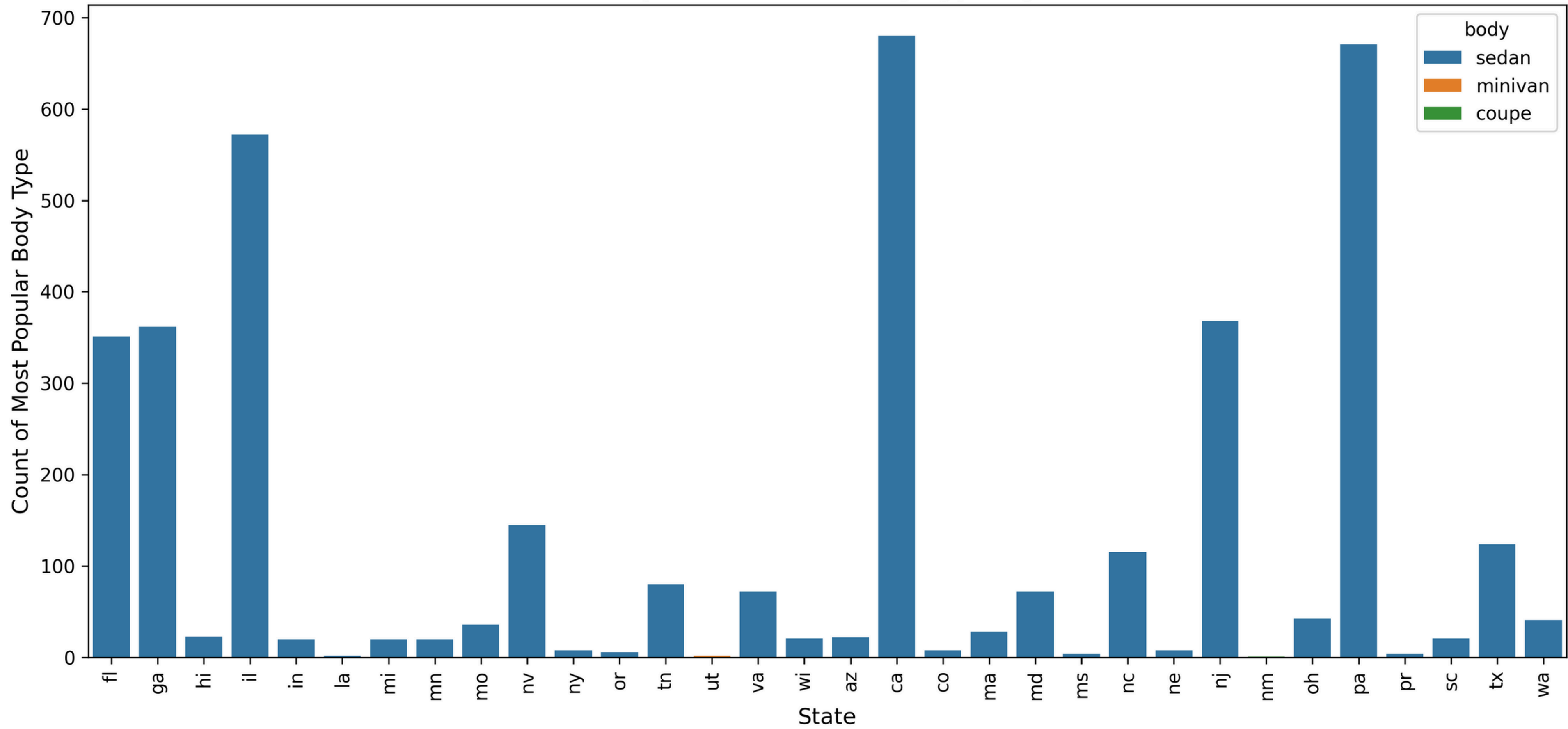
HDFS used as data lake for curated data and dimensional model tables

Hive

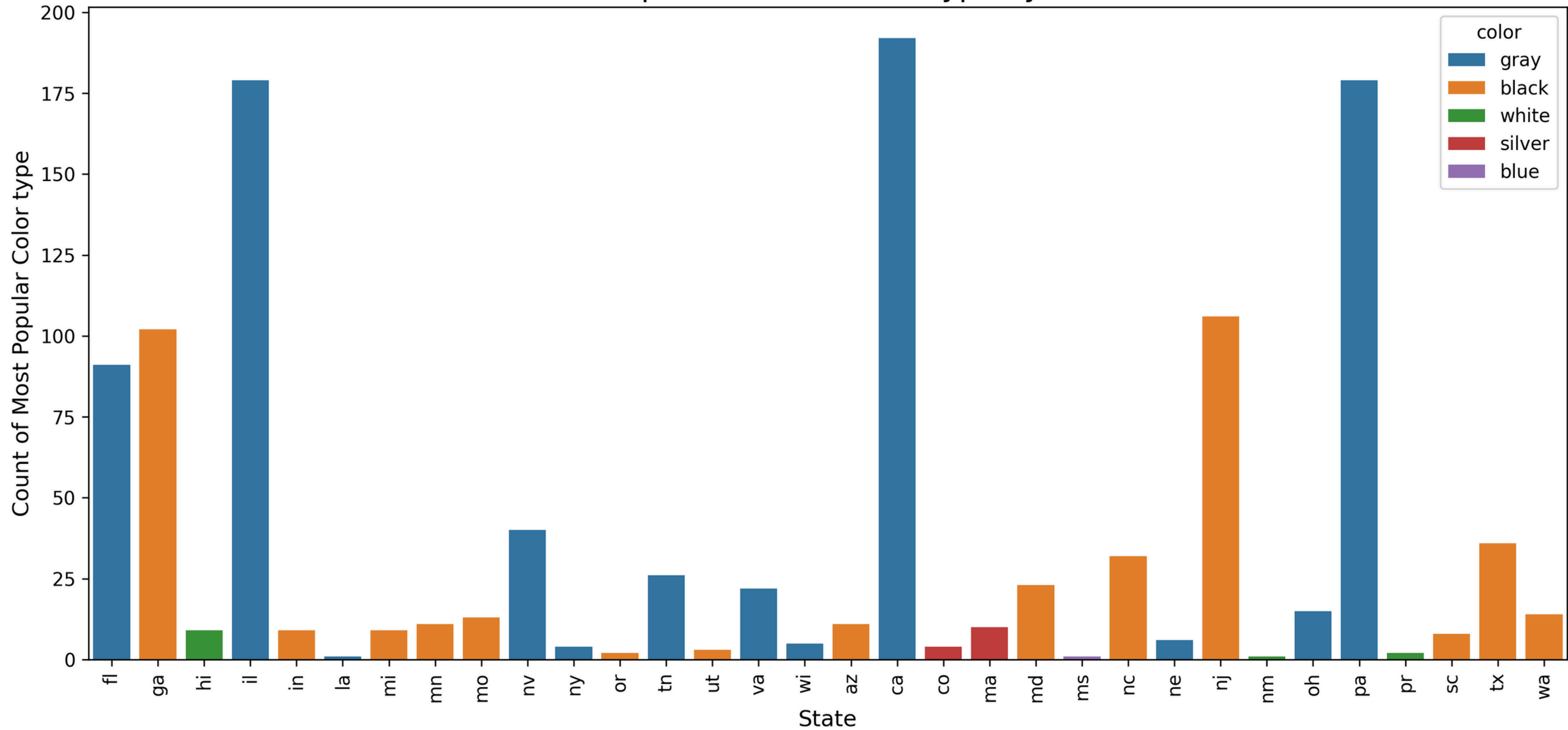


Hive used as a data warehouse to query from using python and Hue

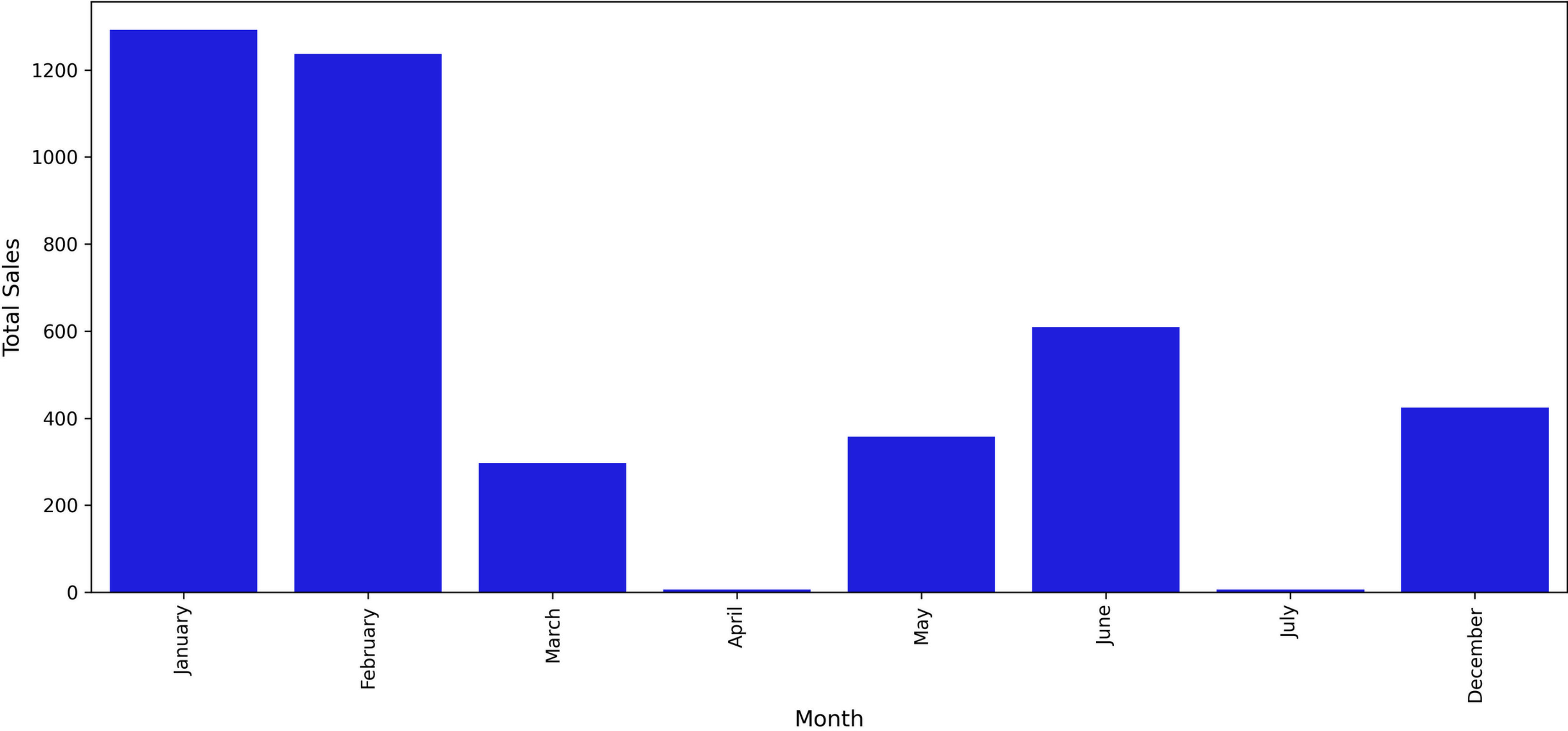
Most Popular Vehicle Body Type by State



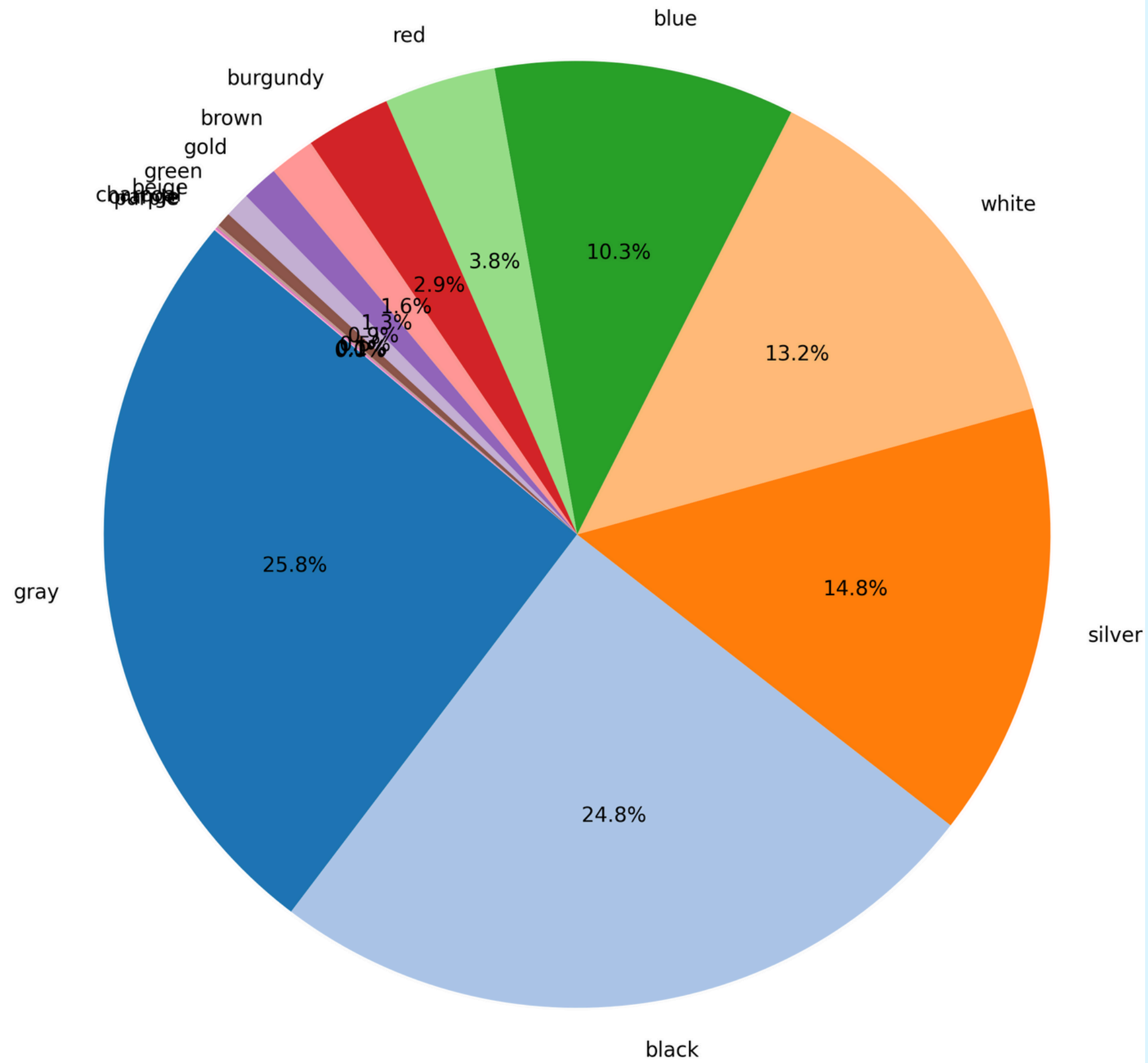
Most Popular Vehicle Color Type by State



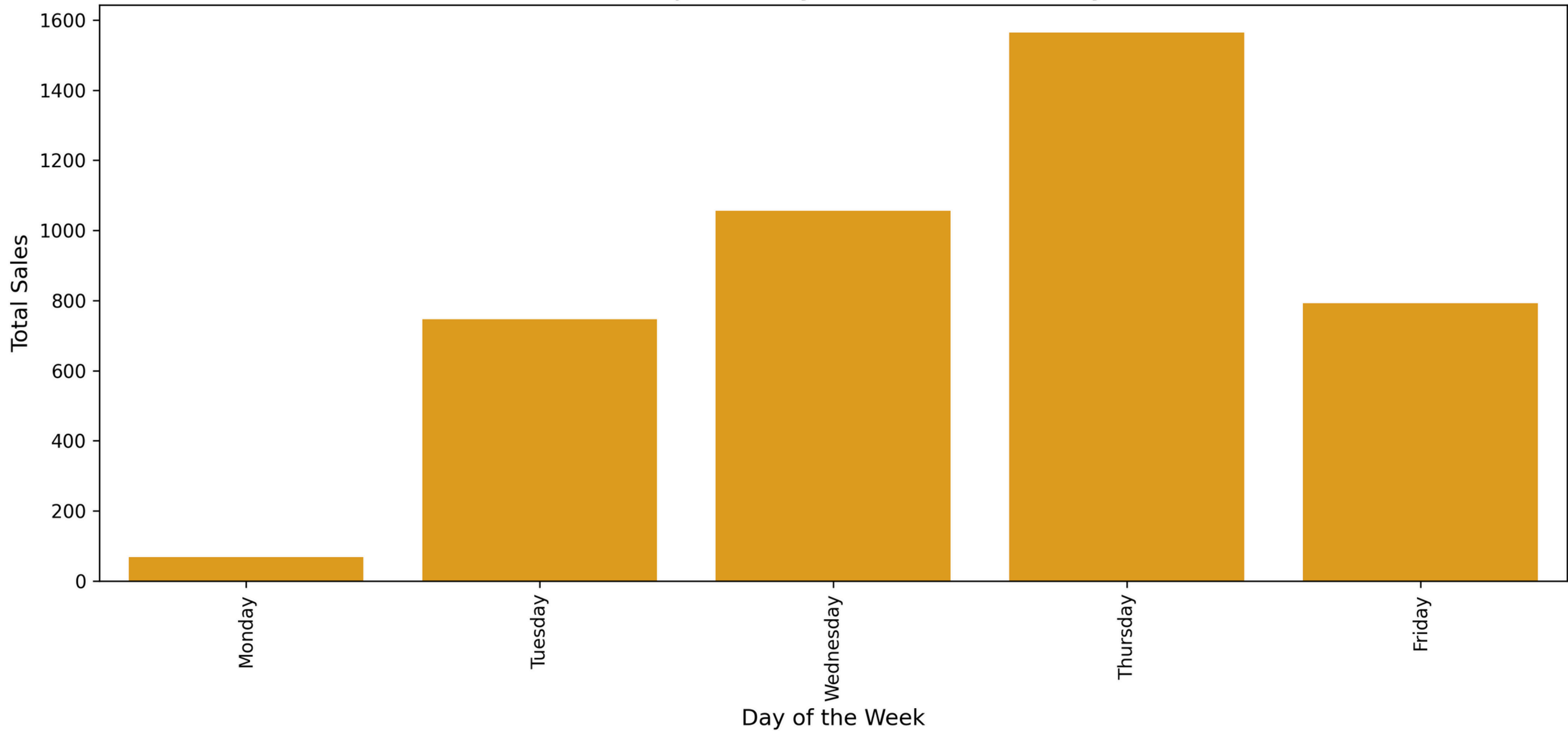
Most Popular Month to Buy



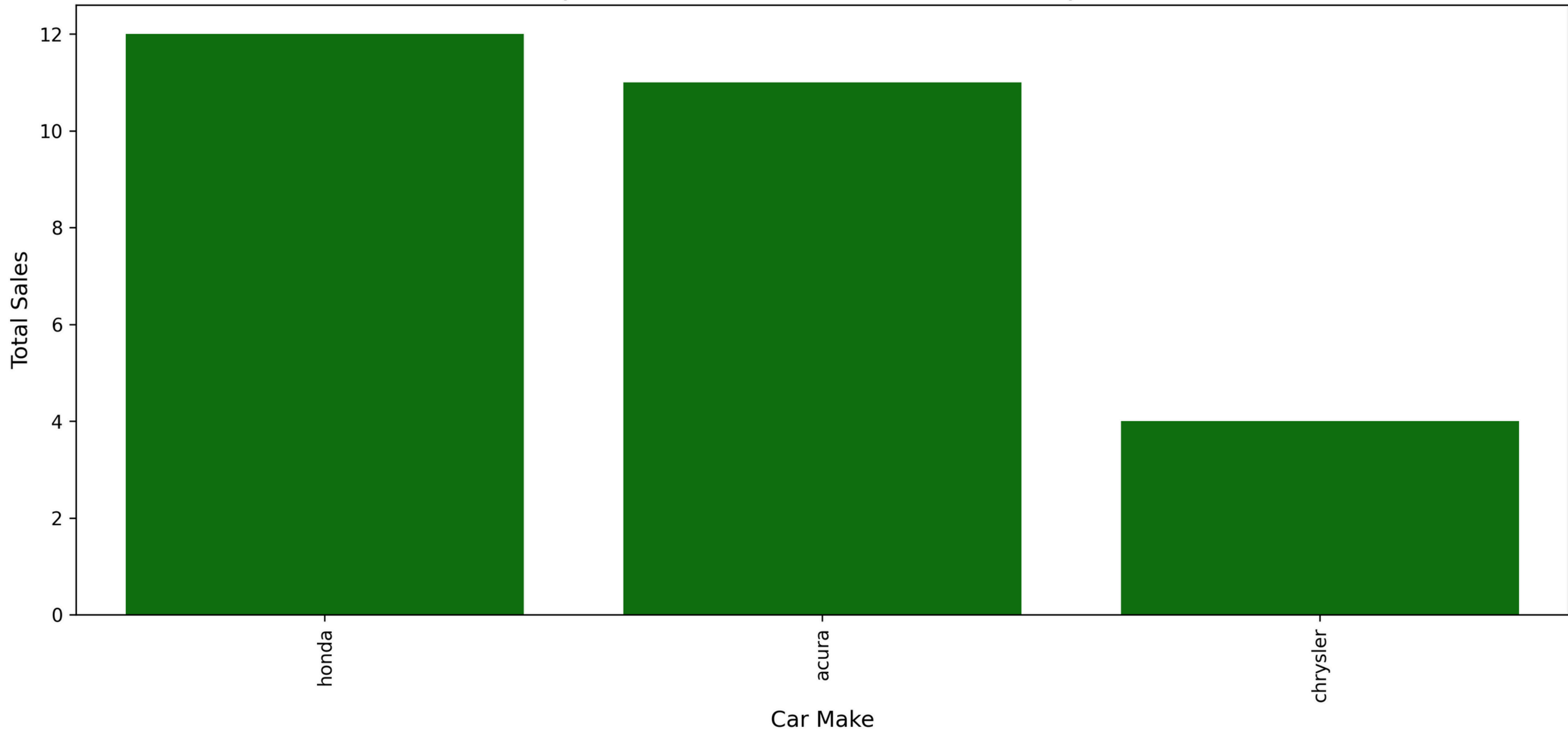
Most Popular Vehicle Colors in the US



Most Popular Day of the Week to Buy

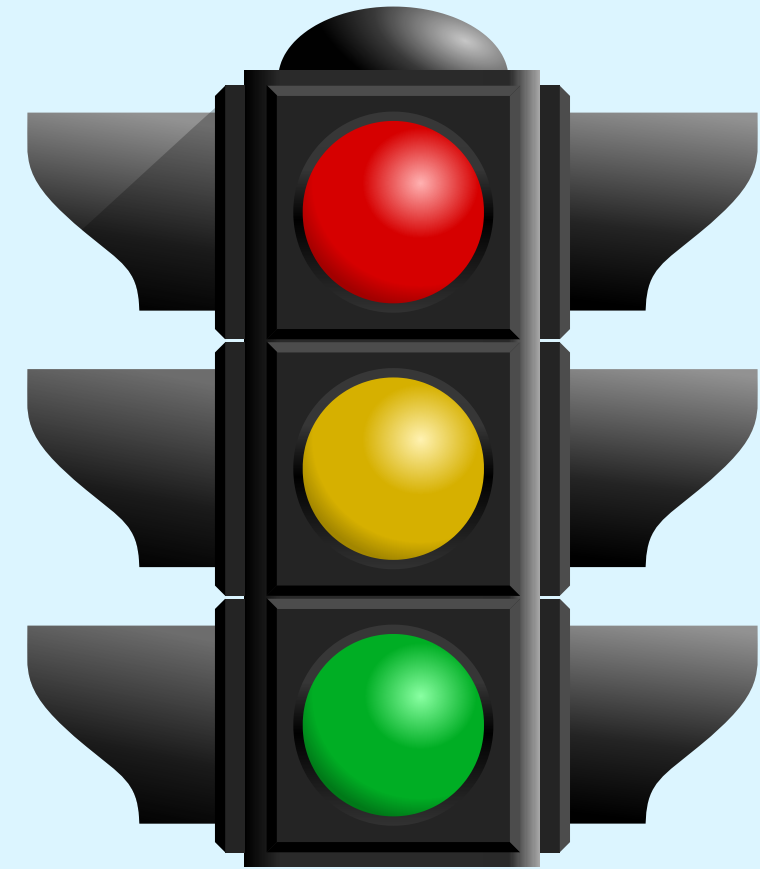


Most Popular Car Make in States with Temp > 72°F

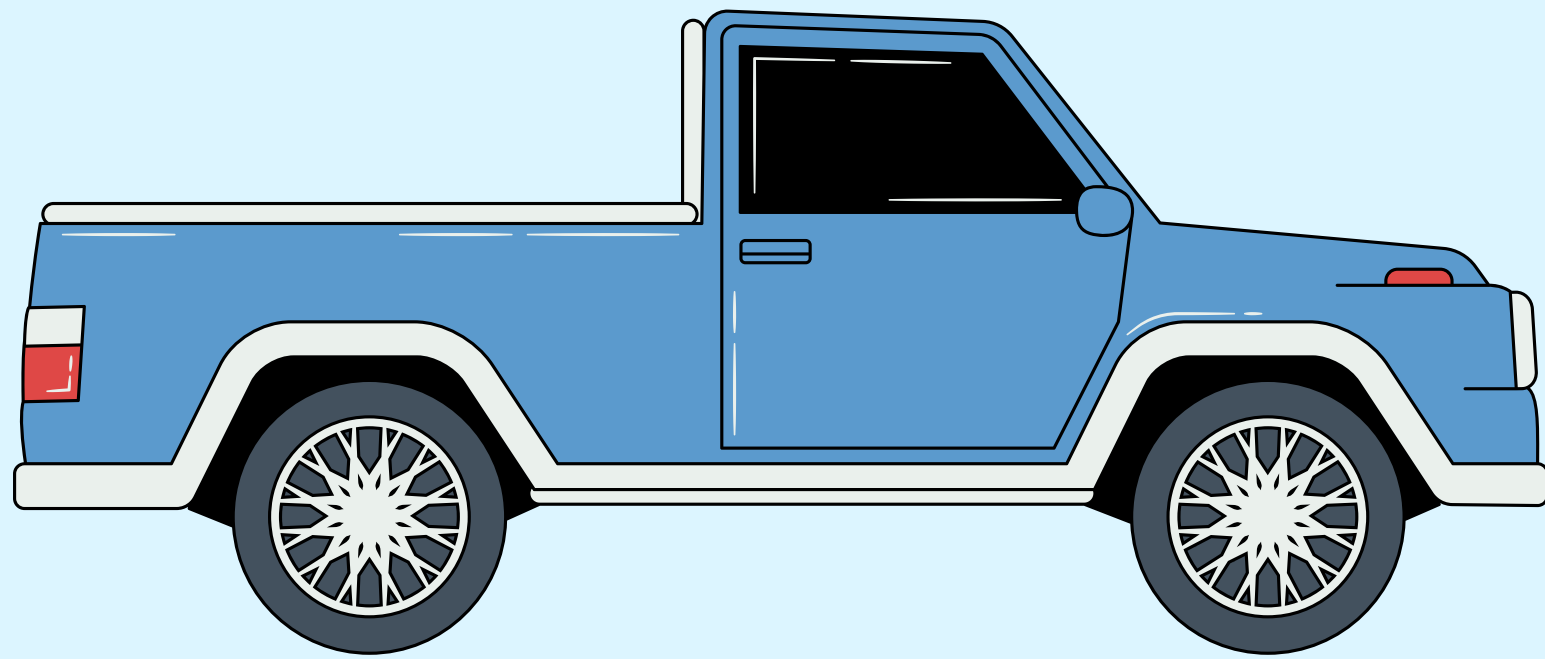


Challenges faced

- Creating queries to get desired data
- Transforming raw data to get desired curated data
- Creating dimensional models
- Debugging



Future Enhancements



- Gather additional weather statistics such rainfall per year.. etc.
- Do ELT instead of ETL
- Implement JSON in Kafka
- Use an interactive dashboard to display visuals
- Use additional datasets

Q&A

Thank you

