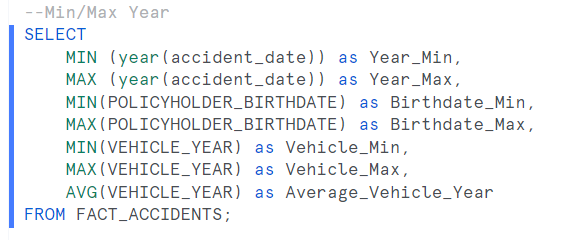
**CAPSTONE PROJECT – JDUSCHAK: Answers and Documentation to Questions 1 – 3:**

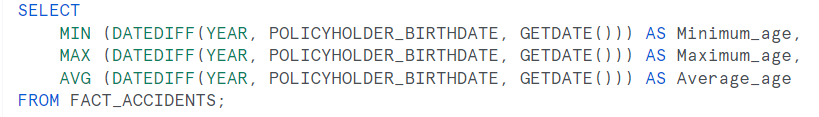
**And other “scratch” work incl. custom SQL for Tableau**

**Link to Snowflake code:** <https://app.snowflake.com/bsyjopy/fpb03241/w4IAJYFFrsEt#query>

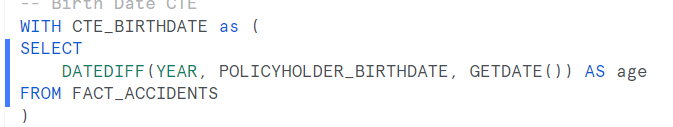
1. **DATA EXPLORATION:**

* Noticed that the initial ERD was a very denormalized schema with repetitive keys and IDs and started wrapping my head around where the redundancies lie.
* ETL was converting this to a more normalized “Star-Schema.”
* Noticed anomalous data in the that there were some irregularities in the birth days in the policyholder data table (ages ranging from 0-116 yrs old) so I surfaced this concern, and we were able to get a new upload of accurate data for Monday. New Max Age appears to be 60 which is also curious.
* Performed Basic Min/Min/Average/Count functions to get insight on data range.
  + Example:

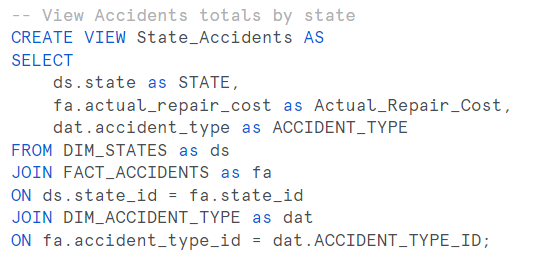




* Created Date Diff to get an age out of Birth Date field:
  + Saved as CTE



* Create view of Accident total repair by state



1. **ETL:**

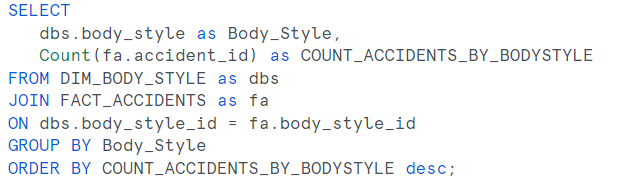
* See attached for ETL script – saved in case of reload needed.



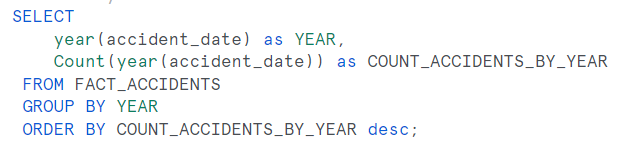
* I did have a challenge with the Fact\_Accidents table as there were multiple joins so I wrote it in the exact order as the Create Table so I would not miss anything.
* DIM tables were more straightforward.
  + “Vehicle\_Use” was misnamed “Vehicles\_Use” in the ERD, so encountered a compilation error on that one when I initially ran.

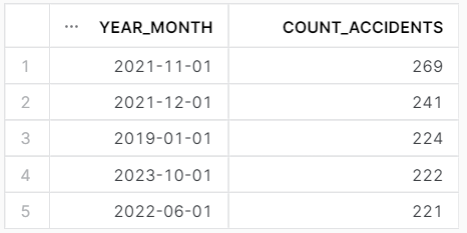
1. **SQL ANALYSIS:**

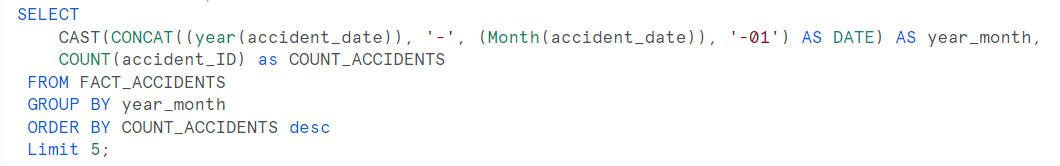
* Which body style of vehicle has the most claims? **4-Door Sedan (3736)**



* Which year has the most claims? **2019 (2,324)**

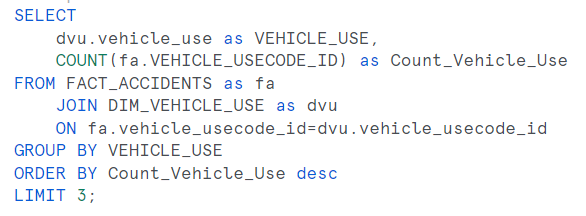


* What are the Top 5 Year & Months? **Nov 2021 (269), Dec 2021 (241), Jan 2019 (224), Oct 2023 (222), Jun 2022   
  (221)**
  + 

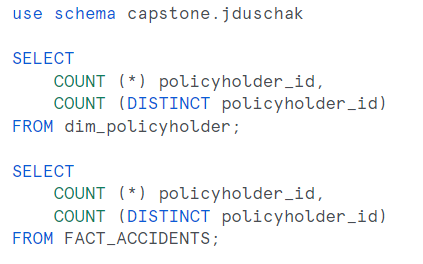


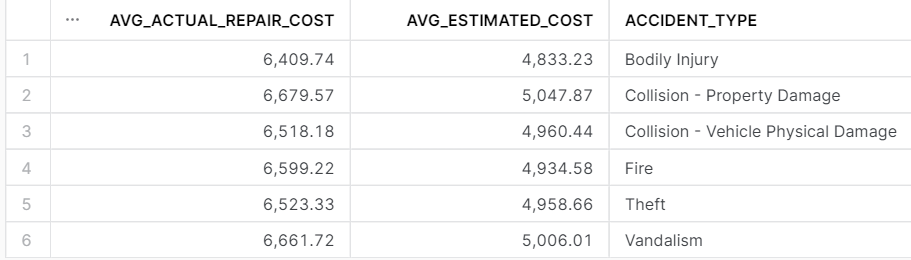
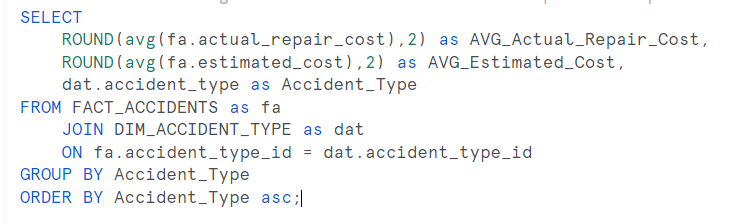
* Top 3 most common Vehicle Use?

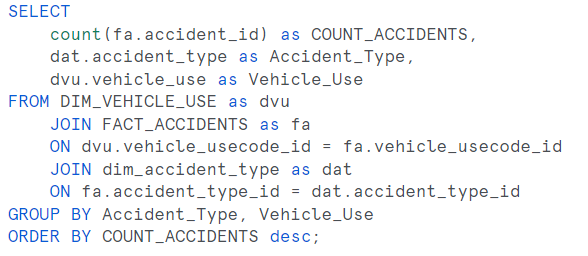




* How many policyholders have more than one claim in the dataset? **Total records 9907, unique 3922**



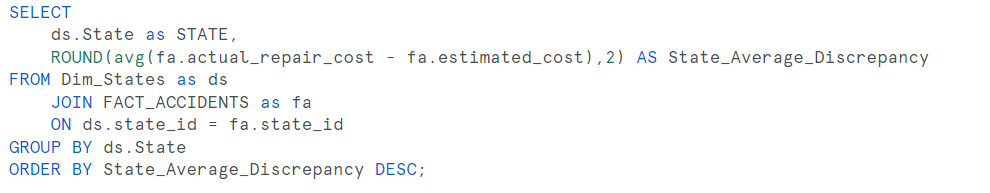
* What's the average estimated cost and actual repair cost per accident type?
  + 
  + 
* Are there any correlations between vehicle use (e.g., work, pleasure) and accident type?
  + Business — No Youthful Operator was highest among all accident types with Work/School — Over 15 Miles coming in second and Work/School — Under 15 Miles coming in third.
    - Vandalism seems to be doubled up (roughly twice the amount) when compared to other claim types. Is this because it is listed twice in Accident type ID? Do we have to eliminate one of these?
    - Collision is split into Vehicle damage and other PD. Does this matter?
  + Solved by writing a query, exporting and analyzing further via pivot table in Excel.
    - **QUERY**:



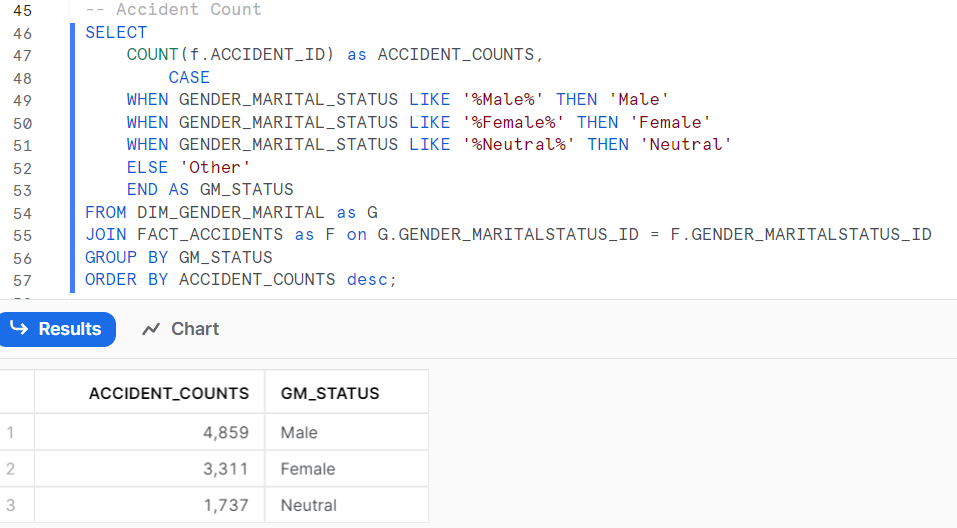
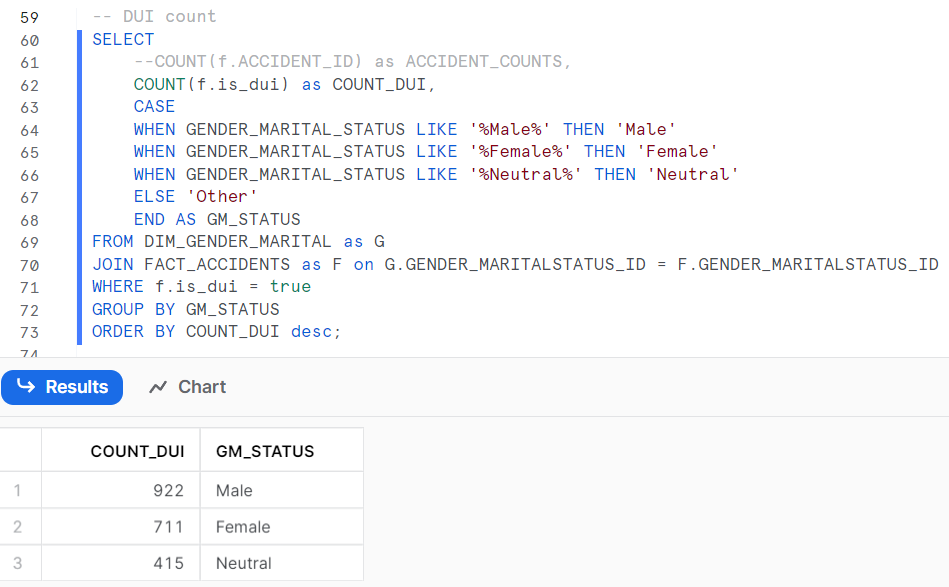
* + - **PIVOT**:



* Which state has the highest discrepancy between estimated and actual repair costs? **WV with average $1,909.58**



1. DEEP DIVE ANALYSIS

* **Claims Time Analysis:** 
  + **yrs 2019-2023**
* **Policyholder Analysis:** 
  + **Min age = 18, max age = 60, average age = 35**
  + **Gender/Marital Analysis – simplified status and did accident count and DUI count:**
    - **Acc count**
  + 
    - **DUI Count**
  + 
* **Vehicle Analysis:** 
  + **Not all Body styles are used**
* **Statewide Analysis:**
  + **CA has largest dollar amount paid in Claim settlements**
  + **WV has largest discrepancy between claim estimate and final cost**
  + **DC has highest DUI count**

1. INSIGHT COMMUNICATION (need x5 – assuming this is based on the above)
   1. Gender breakdown – Accident frequency and DUI
   2. Highest DUI states w/age of driver
   3. Nov 2021 Deep-Dive on claim
   4. Claim type vs. Body Style
   5. Repair discrepancies by state (avg and sum)
2. DATA STORYTELLING AND VISUALIZATION - TABLEAU
3. PRESENTATION:

* Powerpoint: 5 Slides for Insights
* Tableau dashboard requirements:
  + Total number of claims by month for the entire timeframe. xx
  + Distribution of claims by accident type.
  + Highlight the states with the highest number of DUI-related claims xx have data in custom sql
  + Time-series plots showcasing claims trends over time. xx
  + Heat maps to show geographical distribution and concentrations of claims xx

1. RECCOMENDATIONS
   1. Initial dataset had a few anomalies and required some cleaning, but after some discussion with my stakeholders and SMEs, I was able to resolve.
   2. Younger drivers are prone to DUI. Perhaps look into some sort of education program for younger drivers – especially around DUI. Propose a driver education credit.
   3. DC and West Virginia have the highest discrepancies in estimates vs. payout.
   4. DC appears to be an outlier on many different factors. Perhaps we need to consider the drivers for this, and given it is such a small geographical area, perhaps consider a moratorium on writing new business in this area.

**CTEs/Cleanup queries I wrote: Used in Tableau as Custom SQL**

