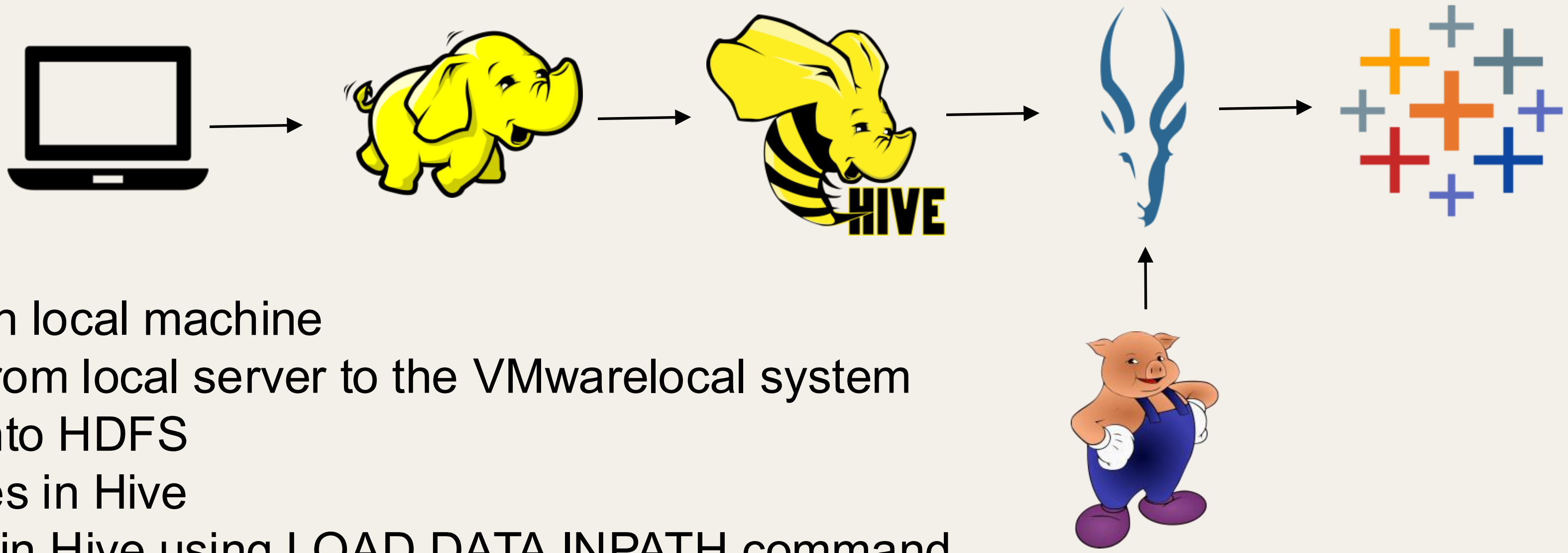
Big Data Project – ANT Trucks

Team 7:
Mehreen Ahmed
Prajakta Bhavsar
Jiapeng Cao
Wan-Chuan Chang
Mahip Singh Charan
Sundeep Chowdary

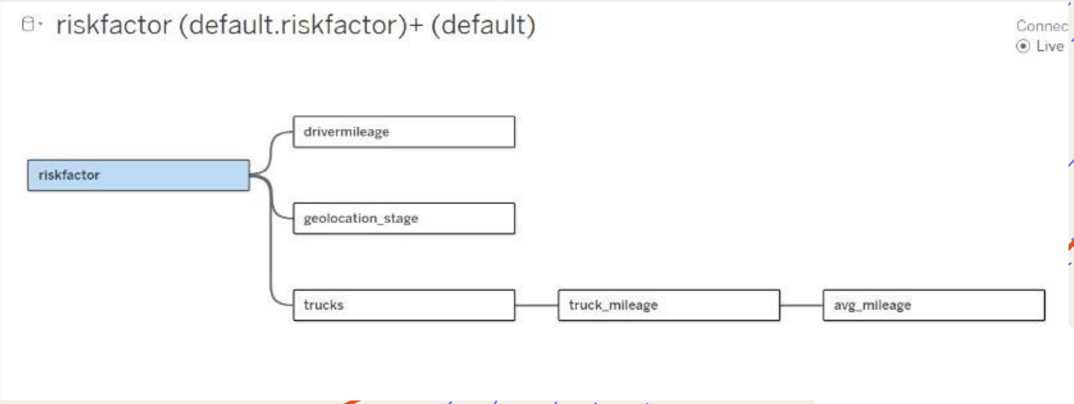
Problem Statement and Objectives

- Large commercial truck accidents continue to be a major cause of injuries and fatalities across the United States. With thousands of miles driven across various geographic and regulatory regions, identifying and mitigating driver-based risk is a critical challenge for fleet management.
- Our objective: To identify truck models and drivers that are at risk of compromising road safety and operational compliance.
- We will be using:
  - Hadoop, Hive, and Impala to process, load, and query large-scale geographic, truck, and driver data
  - Tableau to visualize and recommend data-driven decisions through dashboards and analytics reporting

Analysis Workflow Diagram



1. Save files on local machine
2. Move files from local server to the VMwarelocal system
3. Copy files into HDFS
4. Create tables in Hive
5. Import data in Hive using LOAD DATA INPATH command
6. Create and upload Pig script to populate Riskfactor
7. Import data in Tableau using Impala ODBC driver
8. Create and design schema in Tableau
9. Visualize findings


riskfactor (default.riskfactor)+ (default)

Connec

○ Live

riskfactor

drivermileage

geolocation_stage

trucks

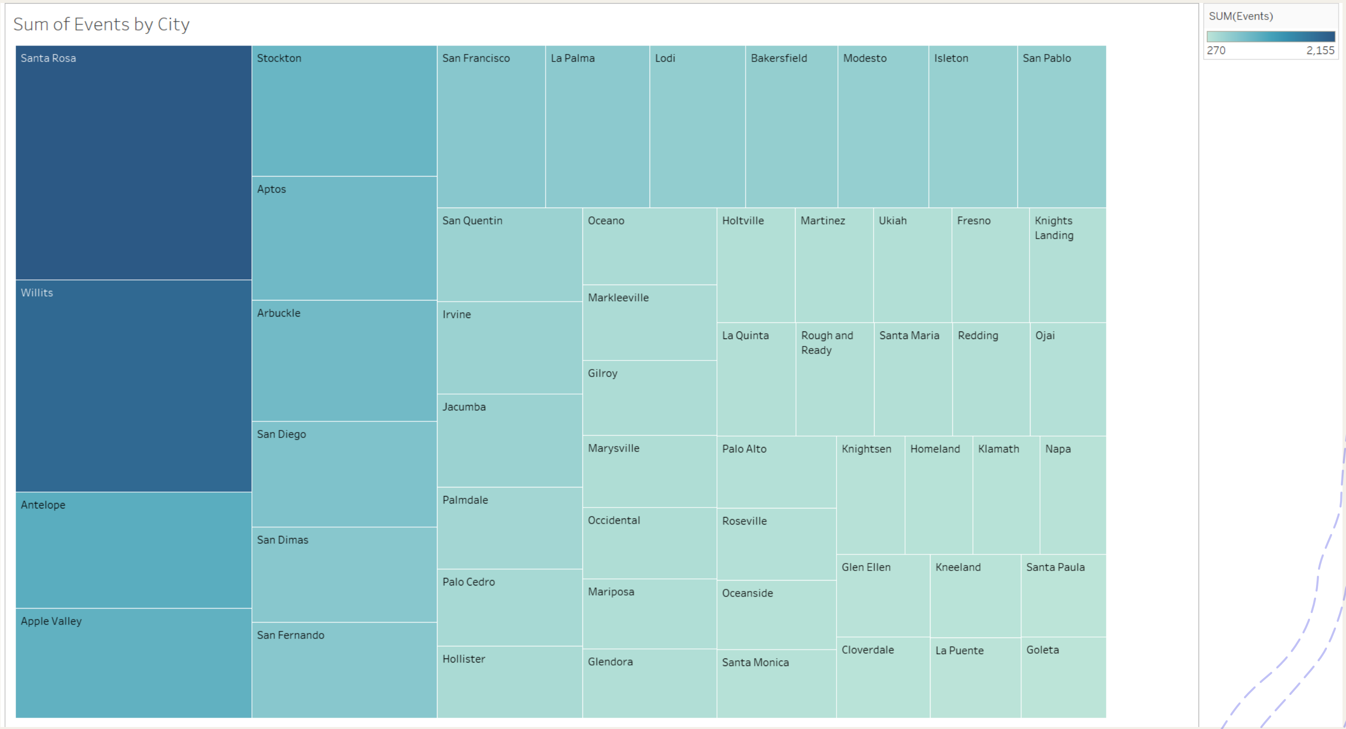
truck_mileage

avg_mileage

Total number of events per city in a Tree Map

Key Insights:

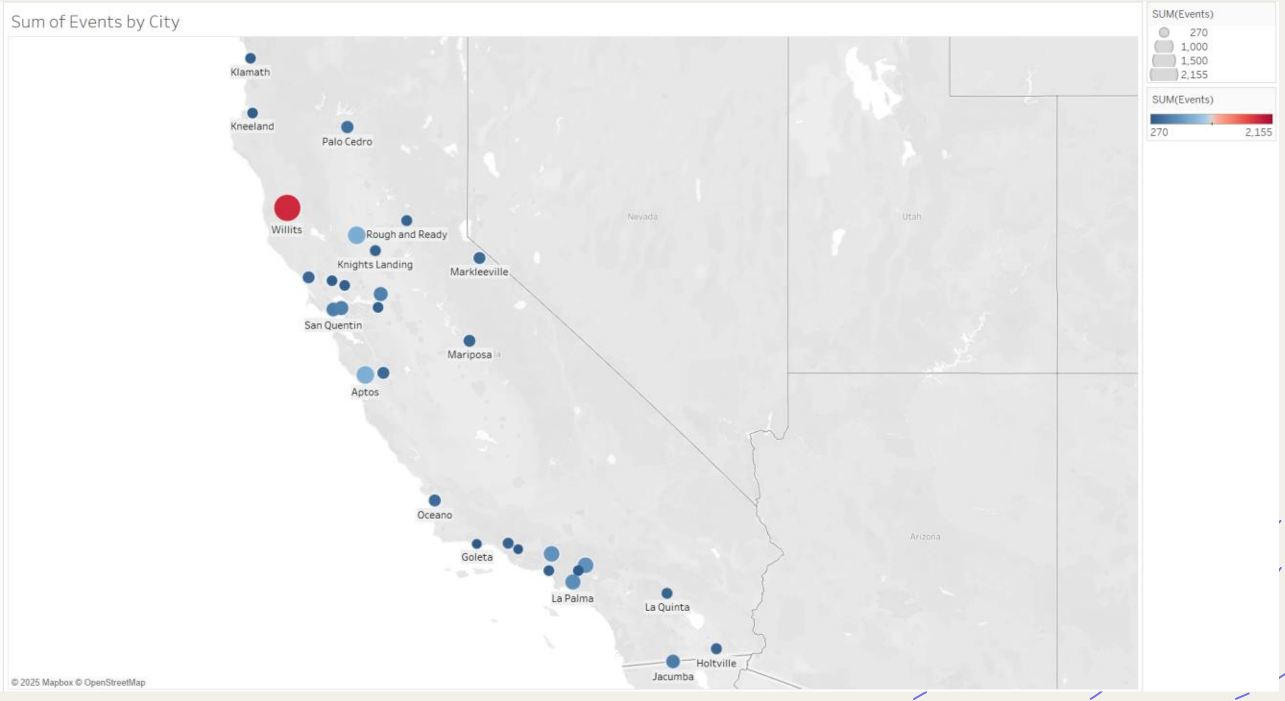
- Santa Rosa, Willits, and Stockton dominate the tree-map, signifying high event counts.
- This view reinforces the insights from the map, showing how certain cities are responsible for a disproportionately large share of incidents.
- Cities like Santa Monica, Cloverdale, and La Puente show up with very minimal contributions.



Total number of events per city in a Filled Map

Key Insights:

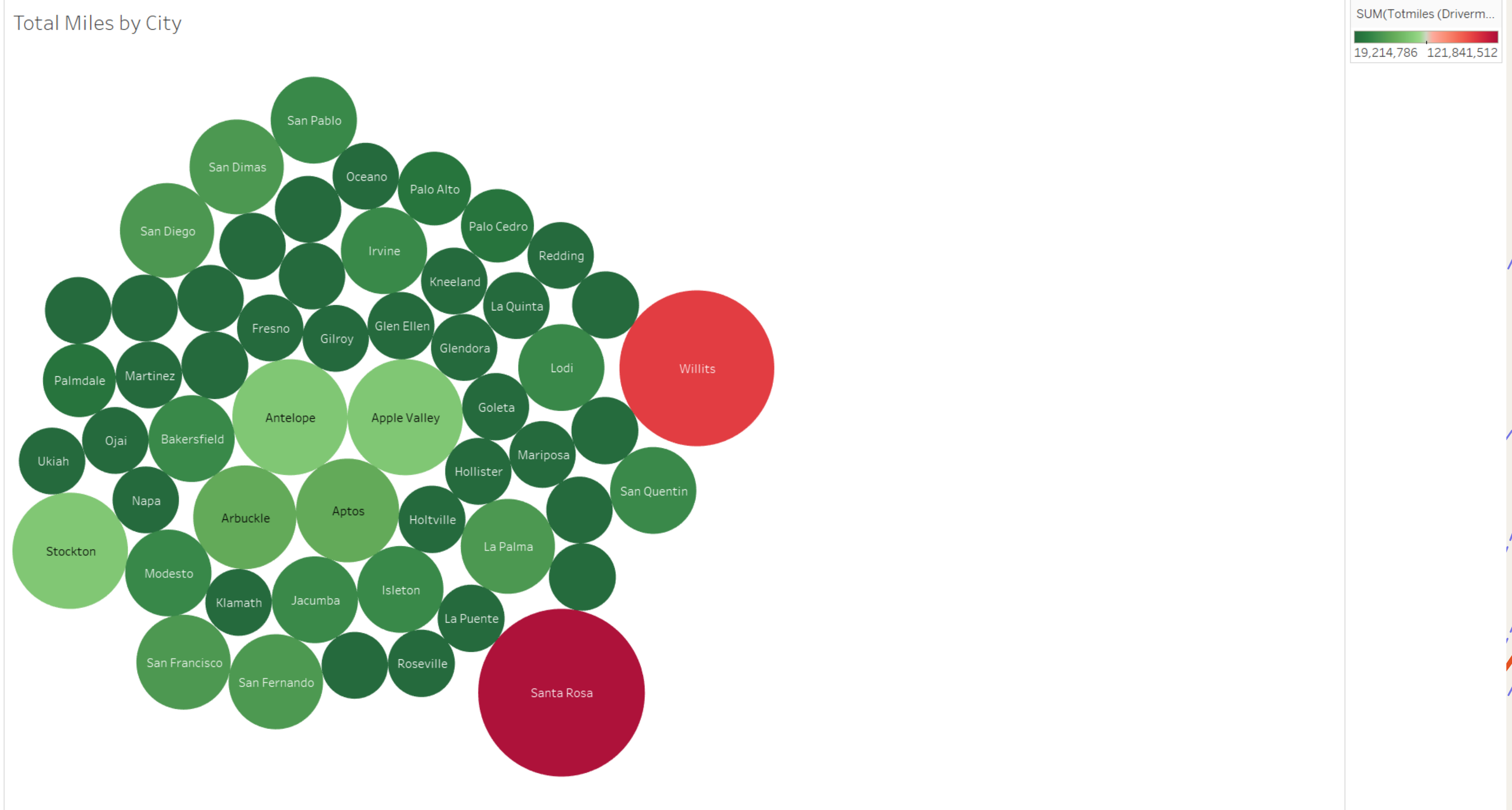
- Willits shows the highest number of events (2,155), highlighted in dark red.
- Other cities such as Aptos, San Quentin, and Rough and Ready show moderate levels of events.
- There’s a geographical clustering in California, with events spread throughout both northern and southern regions.



Total mileage driven per city in a packed Bubble Chart

Key Insights:

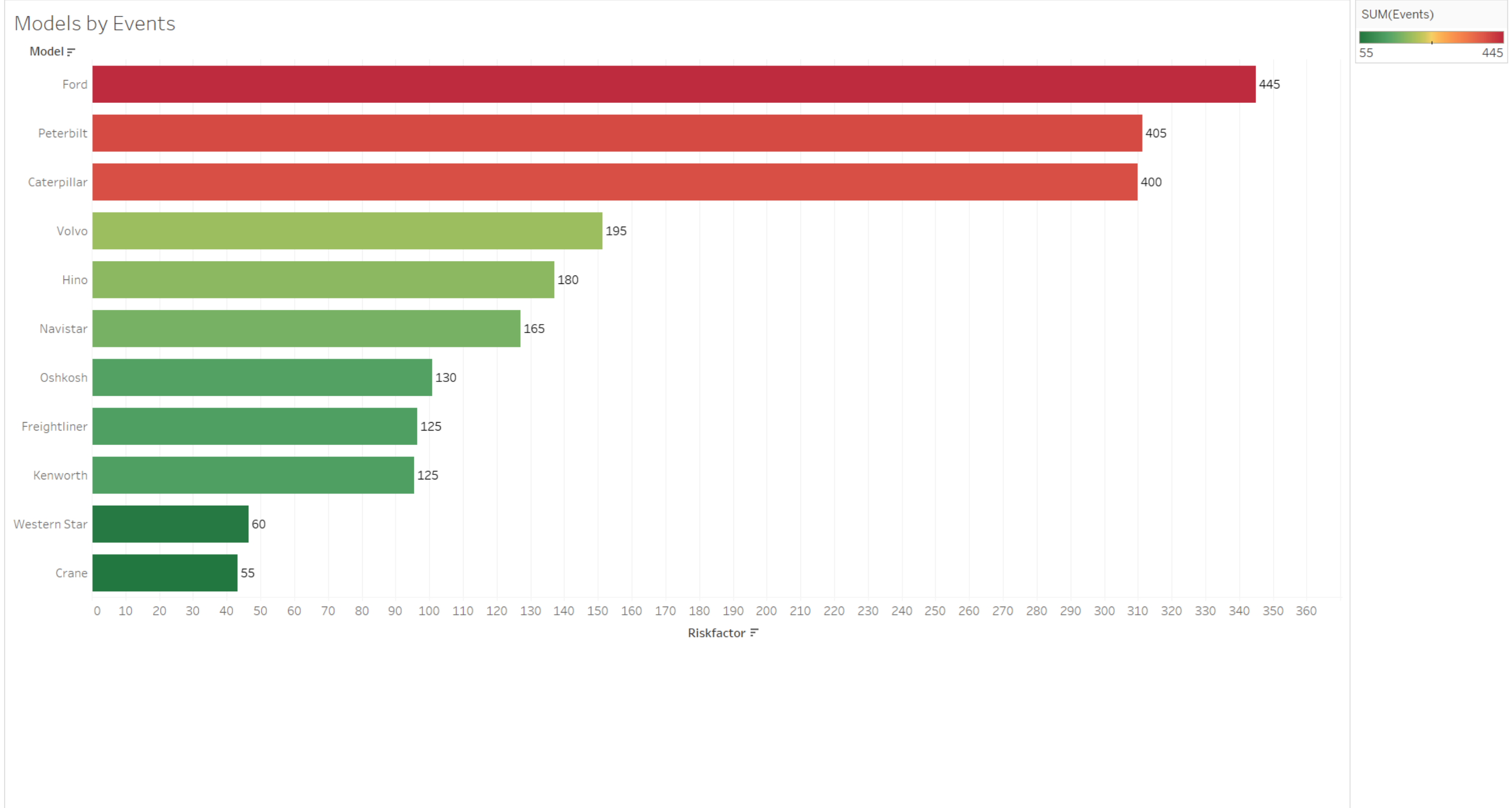
- Santa Rosa and Willits stand out with very high mileage (highlighted in red), indicating heavy vehicle usage.
- Most other cities fall into a mid-tier range of mileage (shown in varying greens).
- Cities like Antelope, Apple Valley, and Stockton also show relatively high mileage but are not in the highest category.



Number of reported events by vehicle model in a Bar Chart

Key Insights:

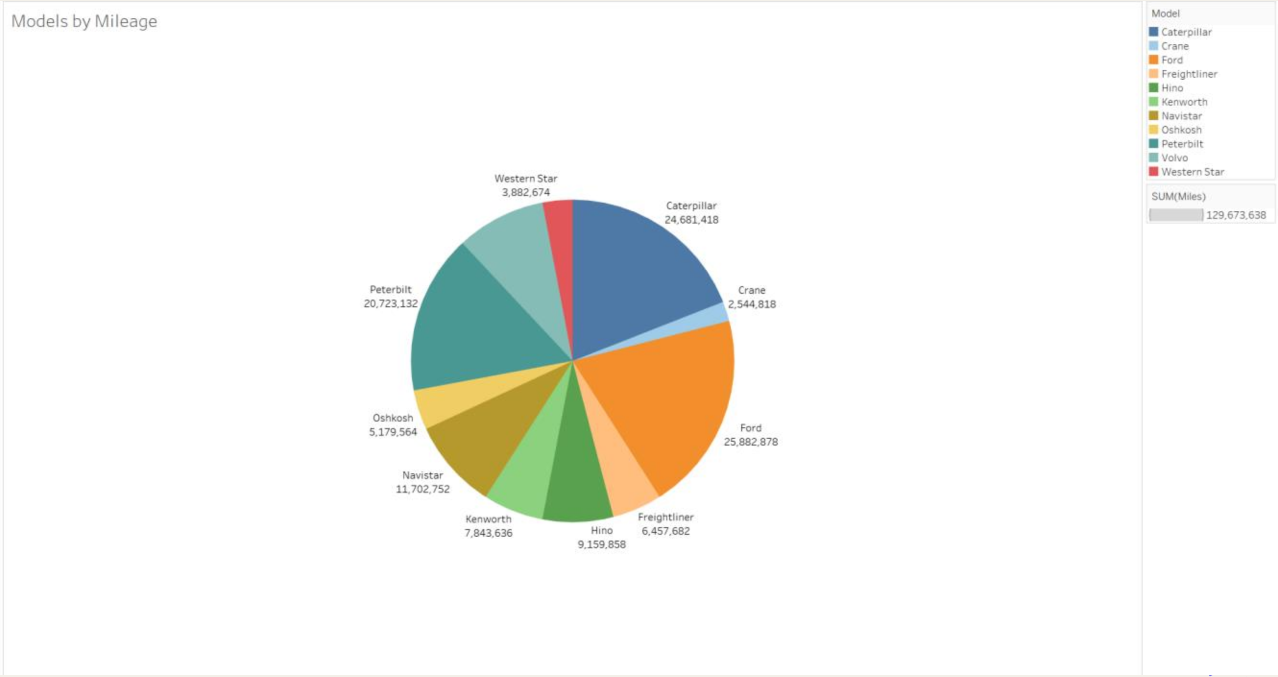
- Ford (445 events), Peterbilt (405), and Caterpillar (400) again top the list, but this time in terms of reported events—correlating with high mileage.
- Crane and Western Star report the least number of events (55–60).
- Some lower-mileage models like Hino (180 events) and Volvo (195 events) still report a high number of events, suggesting possible maintenance issues unrelated to total distance traveled.



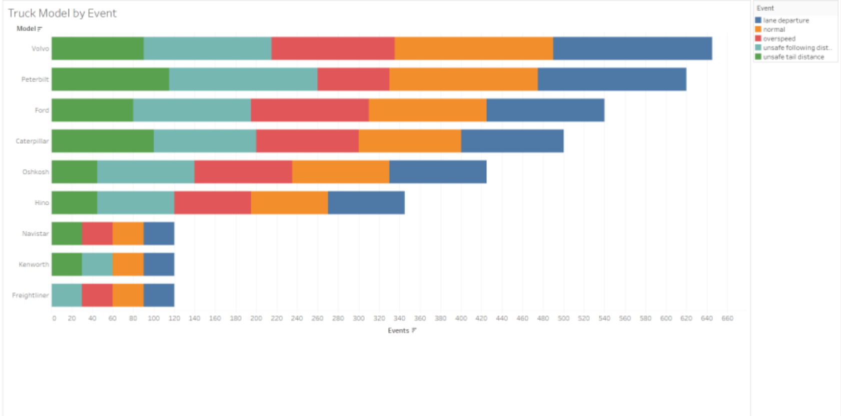
Total mileage covered per vehicle model in a Pie Chart

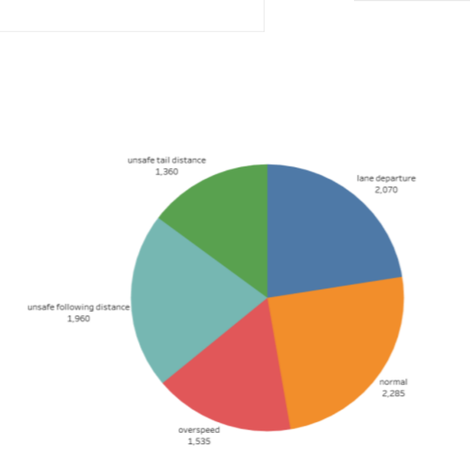
Key Insights:

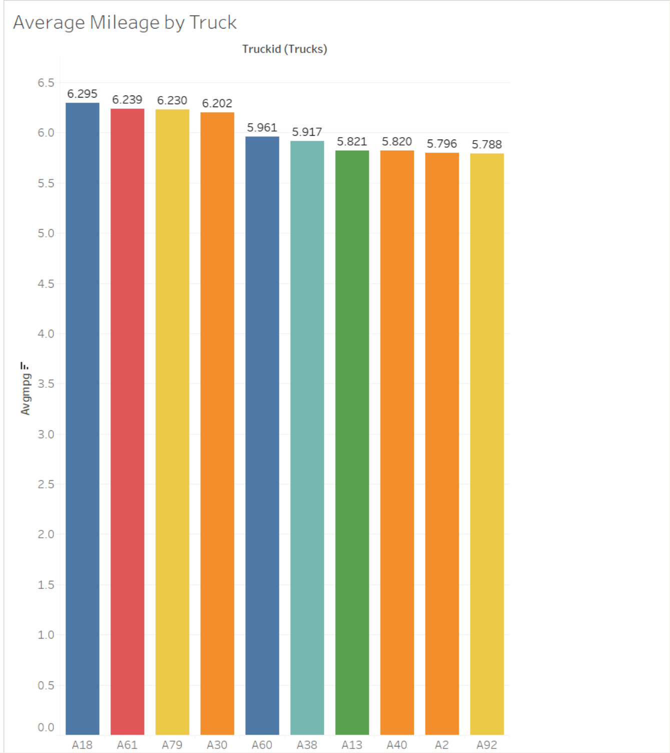
- Ford leads with the highest mileage: 25,882,878 miles, followed by Caterpillar (24.6M) and Peterbilt (20.7M).
- Crane has the lowest mileage at 2.5M.
- These three brands (Ford, Caterpillar, Peterbilt) together account for over 50% of the total 129.7M miles logged.
- Mileage is heavily concentrated among a few models, indicating a skewed usage pattern.

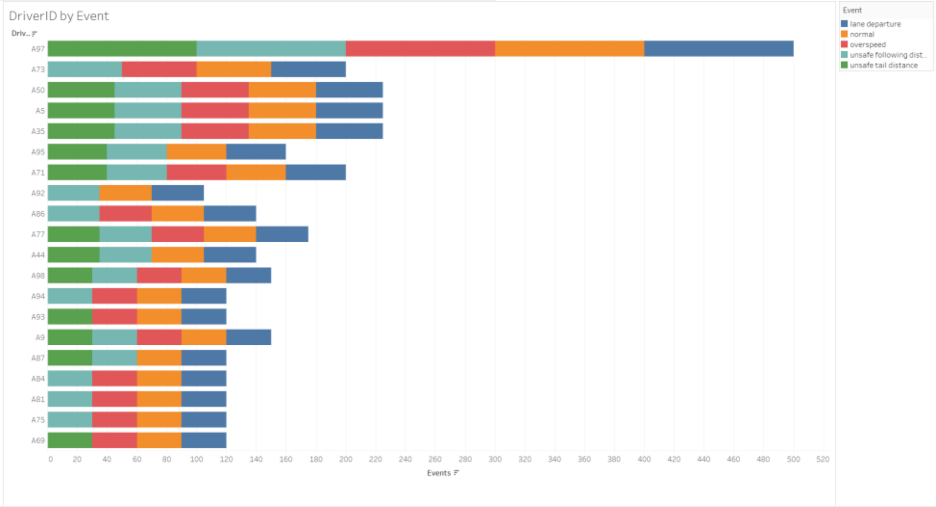


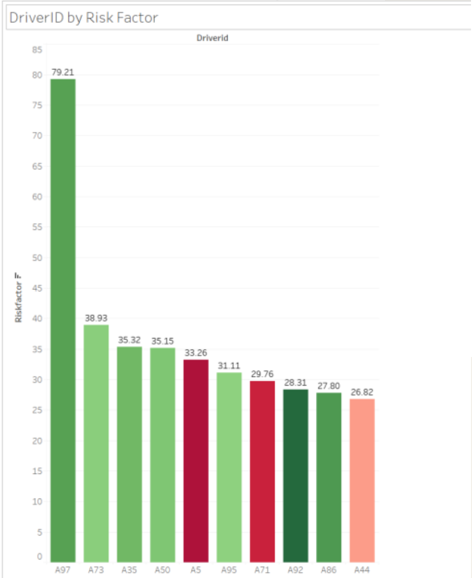
Graphs sorted for the top 10 entries in the charts shown previously











Conclusions we have drawn:

- Driver A97 is the highest risk contributor (Risk Factor: 79.21); several others exceed the threshold (>35).
- Lane departure and unsafe following distance are top risk events.
- Volvo, Peterbilt, and Ford trucks report the most event types.
- Santa Rosa and Willits are cities with the highest incident counts.

1. High Mileage Correlates with High Event Counts

- Ford, Caterpillar, and Peterbilt are the most driven models and also have the highest number of events.
- While high usage justifies higher events, the event-to-mile ratio needs closer monitoring to assess true reliability.

2. Certain Event Types Dominate

- The majority of incidents are “Normal” events, followed by Lane Departure and Unsafe Following Distance.
- These are behavior-related, indicating a need for driver coaching rather than mechanical fixes.

3. City-Level Risk Concentration

- Santa Rosa and Willits are high-risk zones—logging the most miles and the most incidents.
- Urban planning, traffic congestion, or driver familiarity may be influencing this.

4. Driver Performance Is Uneven

- Driver A97 shows disproportionately high risk (score: 79.21) and event count—possibly due to unsafe habits.
- A few other drivers like A73 and A50 also show higher-than-average incident profiles.

5. Vehicle Efficiency Varies Slightly

- Caterpillar and Freightliner trucks demonstrate better fuel efficiency (~6.2–6.3 mpg) than others.
- Minor MPG differences compound over time—small improvements here mean significant fuel savings.

Our Suggestions:

- Retrain or monitor high-risk drivers (A97, A73, A50).
- Prioritize safer models/routes for high-mileage tasks.
- Use behavior-based alerts to reduce overspeed and lane deviation events.
- Reroute away from high-incident cities to improve safety.

1. Implement a Driver Risk Monitoring Program

- Flag high-risk drivers (e.g., A97) for immediate retraining or driver coaching.
- Create a monthly risk score dashboard by integrating event type severity and mileage.

2. Shift Routing Away from Risky Cities

- Reduce deployments in cities like Willits and Santa Rosa unless necessary.
- Consider route optimization algorithms to divert traffic from historically risky zones.

3. Model-Specific Maintenance Review

- Ford and Caterpillar trucks should undergo frequent checks to validate if issues are mechanical or behavioral.
- Examine event type distribution per model to distinguish between wear-and-tear and user-based causes.

4. Use MPG Data to Assign Long Routes

- Assign high-mileage routes to trucks like Caterpillar A18 and Freightliner A61 for better fuel economy.
- Maintain a fleet efficiency leaderboard to encourage proper maintenance practices.

5. Behavior-Based Event Reduction Strategy

- Target overspeed, tailgating, and lane departure events with:
- Real-time driver alerts
- Monthly behavioral feedback reports
- Incentives for clean driving logs

THANK YOU!

Any questions?