# USED CARS

Data Warehousing Project Presentation

University of Applied Sciences Ulm

Jossin Antony, Louis Reichert, Alperen Sarac, Jan Schrade, David Todorov

# Business Understanding

- As the inflation and interest rates rise over the years, buying and owning a car costs more than before. Average American spends over 700 $ a month on car payments alone, excluding the insurance, maintenance etc..
- And 15% of the people pays over a 1000$ monthly in 2023.

# Business Understanding

- Since most cars are depreciating assets over time, we are taking a look at the used market, which already took the first depreciation hit going from ‘New’ to ‘Used’
- In this regard, the dataset we worked on, shows us data about the US market. Due to its significance in the industry and present oportunities for both buyers and sellers.

# Business Understanding

- For our analysis, we are utilizing a dataset, sourced from the platform Kaggle, which offers a diverse collection of information.
- Our goal by working on this dataset, is to gain insights into the patterns, trends and dynamics of the American used car market.

# Data Understanding

- The initial phase of out scientific inquiry involved a detailed examination of the information provided in the dataset.
- As with most of the datasets, there were missing/incomplete data.
- We worked out way around some of these by filling in missing fields by values like ‘Unknown’ in order to have a complete dataset.

# Data Understanding

- While observing the data, we categorized multiple columns to gain deeper insight.
- We divided the fuel type column in 5 categories.
- As well as performing categorizations on mileage, year, transmission modes, accidents, one owner status and driver rating columns.
- With this, we were able to explore relationships, patterns and trends within the dataset.

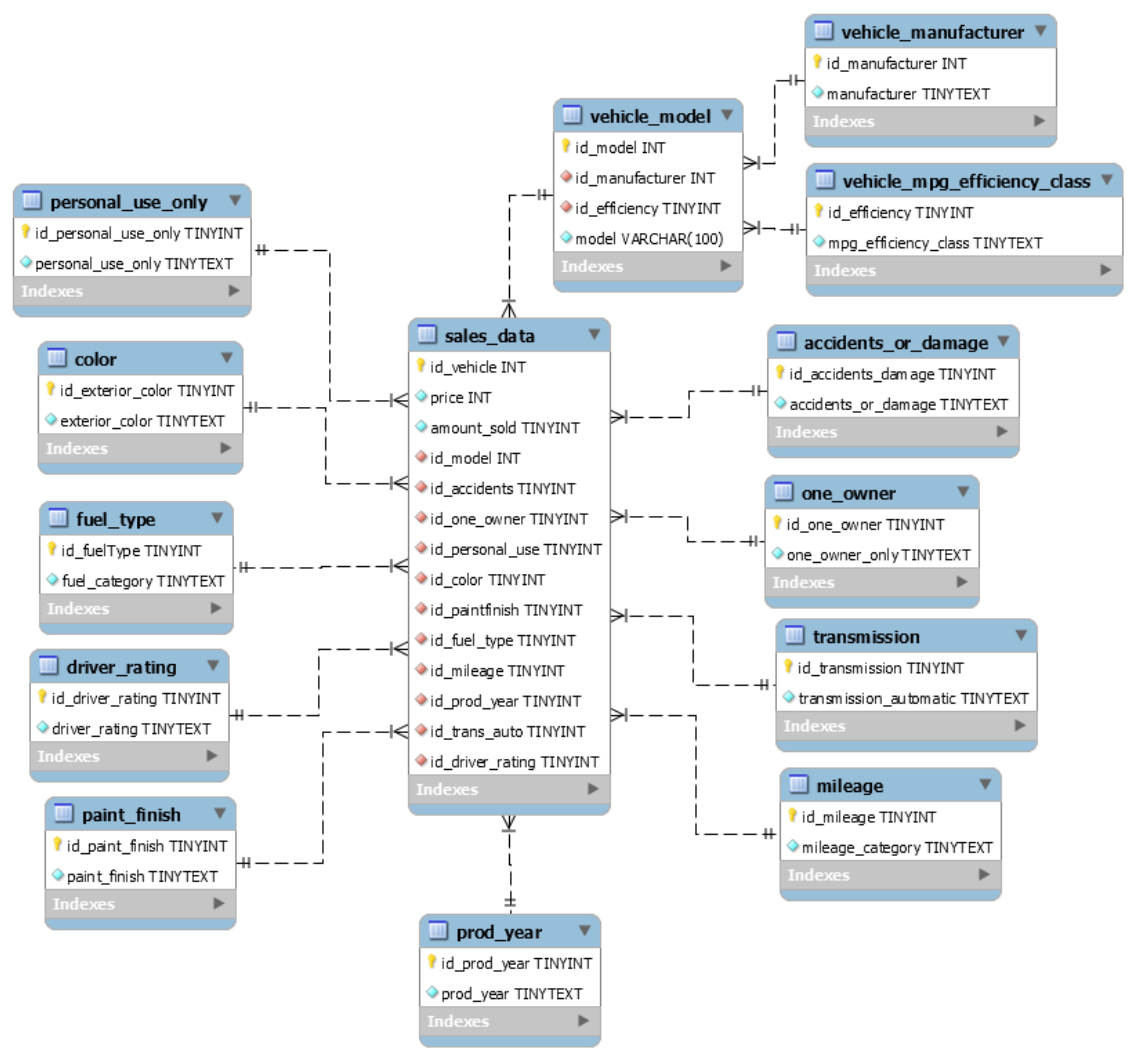
# Data Preparation

As mentioned, the categorization was needed in order to properly assess the data set.

An example would be colors. Since many companies have different names for internal or external colors etc., we simplified them by putting them in readable categories.

We excluded some data that wasn't either believable or outright didn't make sense, such as used cars with build year of 2024.

This categorization process enhanced the interpretability and usability of the dataset, facilitating more nuanced explo- rations of the relationships between different variables.




# Data Modeling

- Data was loaded into a temporary staging warehouse for cleaning, transformation, and integration before being loaded into the Core Data Warehouse (CDWH).
- CDWH utilized a snowflake schema consisting of a central fact table (price and amount) connected to multiple dimension tables (e.g., fuel type, mileage) to organize the data hierarchically.
- Categorizations were used in dimension tables such as model, accidents and damages, one owner, personal use only, color, paint finish, mileage, product year, transmission, and driver rating.

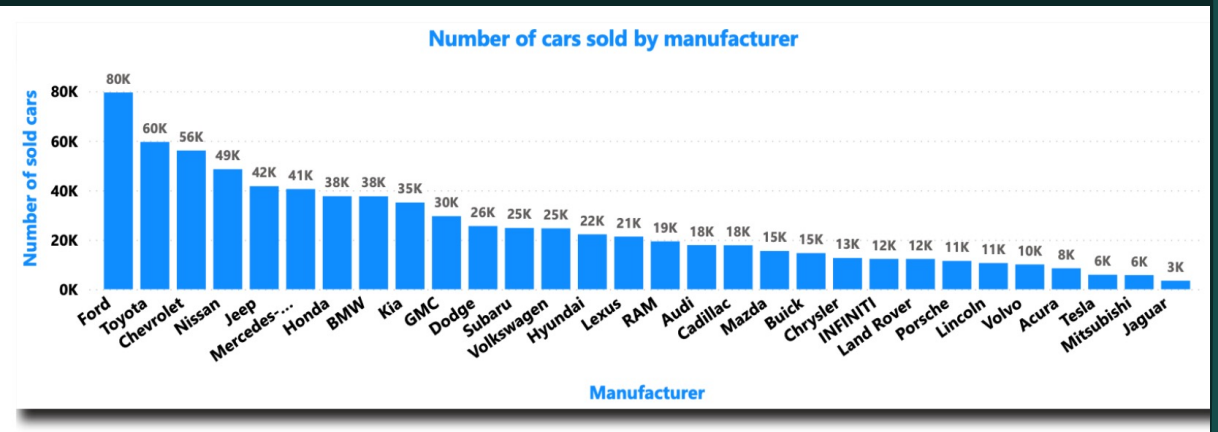
# Data Modeling

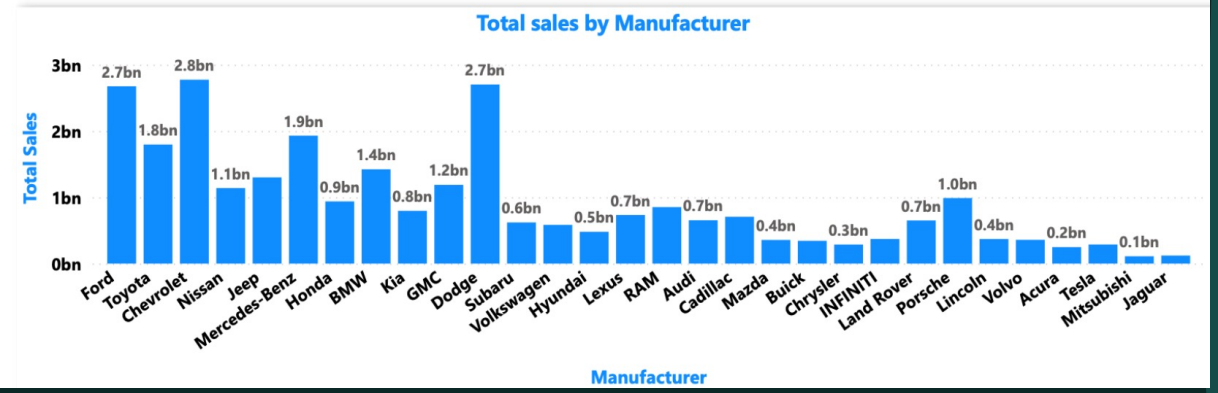
Staging warehouse and snowflake schema played a crucial role in the success of the CDWH project, enabling data access and analysis for valuable insights.

Indexes were added to necessary tables in the ETL database, and a new dimension was created among the thirteen dimensions with unique categorizations.

An issue with mapping miles per gallon in a one-to-one relationship was encountered during the creation of intermediary databases and tables but was successfully resolved.

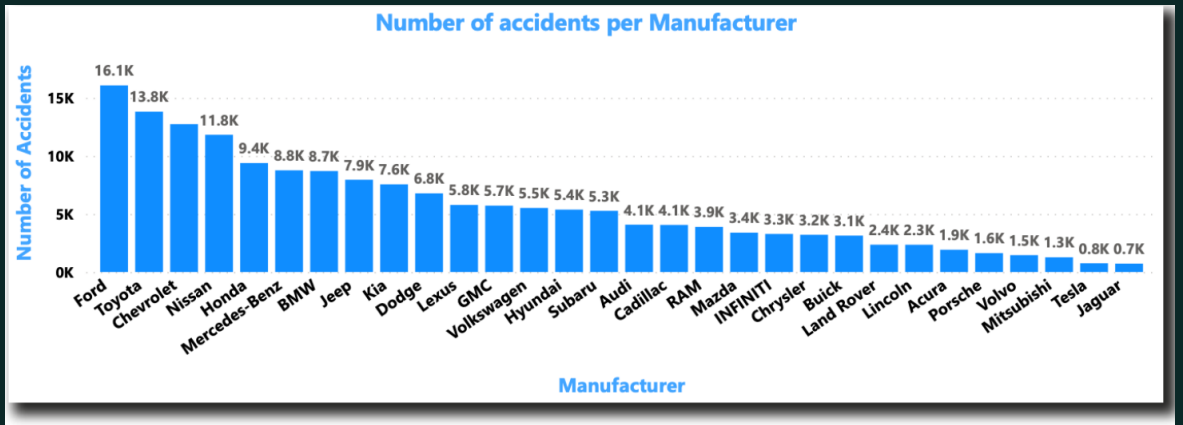
Creating the data warehouse required attention to detail and a thorough understanding of the data being modelled.

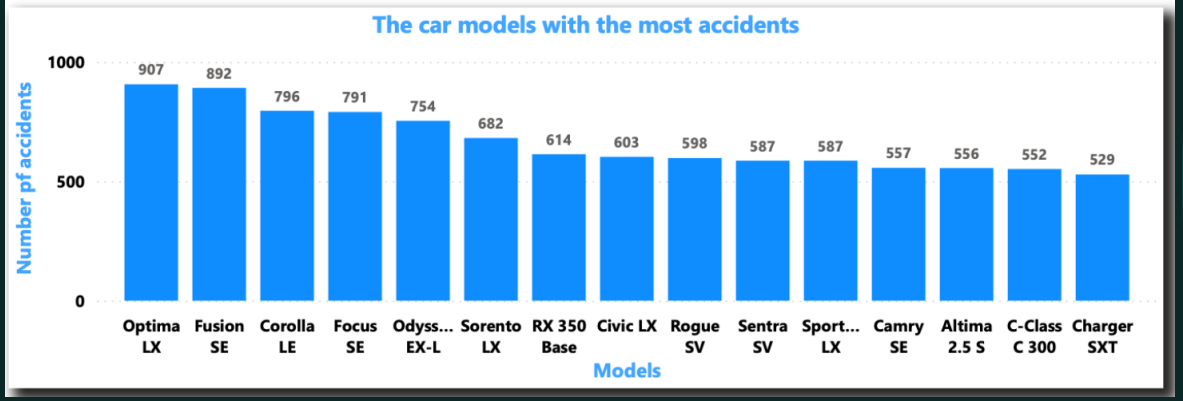



# Evaluation

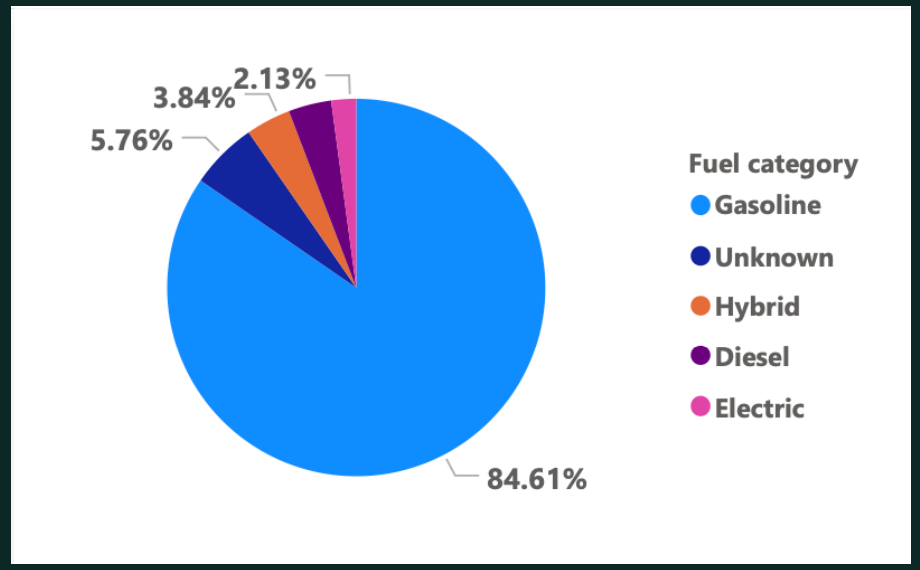
- Car sales vary significantly among different manufacturers, with Ford emerging as the leading manufacturer, selling approximately 80,000 cars.
- Ford's sales surpass Toyota by a difference of 20,000 cars, while Jaguar has the lowest sales figures, selling around 3,000 cars.
- On average, each manufacturer has sold approximately 25,000 cars.
- These findings indicate substantial variations in sales performance among manufacturers in the automotive industry.
- Despite Ford cars being involved in the highest number of accidents, they also exhibit the highest resale volume compared to other manufacturers.



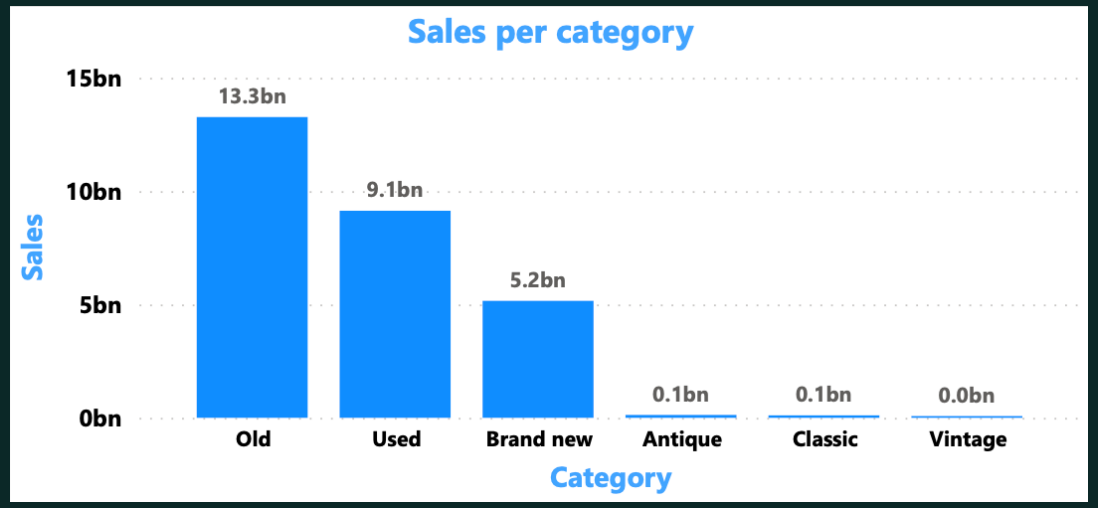

# Evaluation

- The graph demonstrates a similar trend to the previous findings, suggesting a correlation between the number of cars sold and the frequency of accidents.
- The proportion between the number of cars sold by each manufacturer and their accident involvement ranges from approximately one-fourth to one-fifth
- This relationship becomes evident when considering Ford as an example. With a sales volume of around 80,000 cars, Ford experienced approximately 16,000 accidents, resulting in the aforementioned proportion.




# Evaluation

- The figure reveals that gasoline is the predominant fuel type for used cars, representing approximately 85% of all vehicles.
- This finding suggests a continued reliance on conventional gasoline-powered vehicles, indicating that a significant portion of cars in use are older models.
- Electric or hybrid cars account for only around 6% of the total used car market, indicating a relatively low adoption of more environmentally friendly fuel options.
- A notable observation is the relatively low prevalence of diesel consumption in the American market compared to some other countries.
- This indicates that diesel-powered vehicles are not as common or popular among American buyers.
- As a result, prospective buyers seeking a used car in America are highly likely to encounter vehicles fuelled by gasoline.




# Evaluation

- This figure provides insights into the sales distribution across different car categories.
- Owners who have recently purchased their vehicles are more inclined to retain them for a longer duration, possibly due to the novelty factor.
- This is reflected in higher sales volumes of older cars, indicating that owners are more likely to part ways with them if the car had only one owner.
- Reselling old cars generates a substantial revenue of $13.3 billion, while used cars contribute approximately $9.1 billion in sales.
- The graph trend suggests a growing attachment and longer ownership, with owners selling their cars when they significantly age or lack the latest technological advancements.
- Cars older than 23 years are particularly rare in the resale market, potentially due to sentimental value or owners' inclination for car preservation or collection as a hobby.

# Deployment

The completion of the data warehouse project enables the provision of valuable data to a specific company or website specializing in car sales.

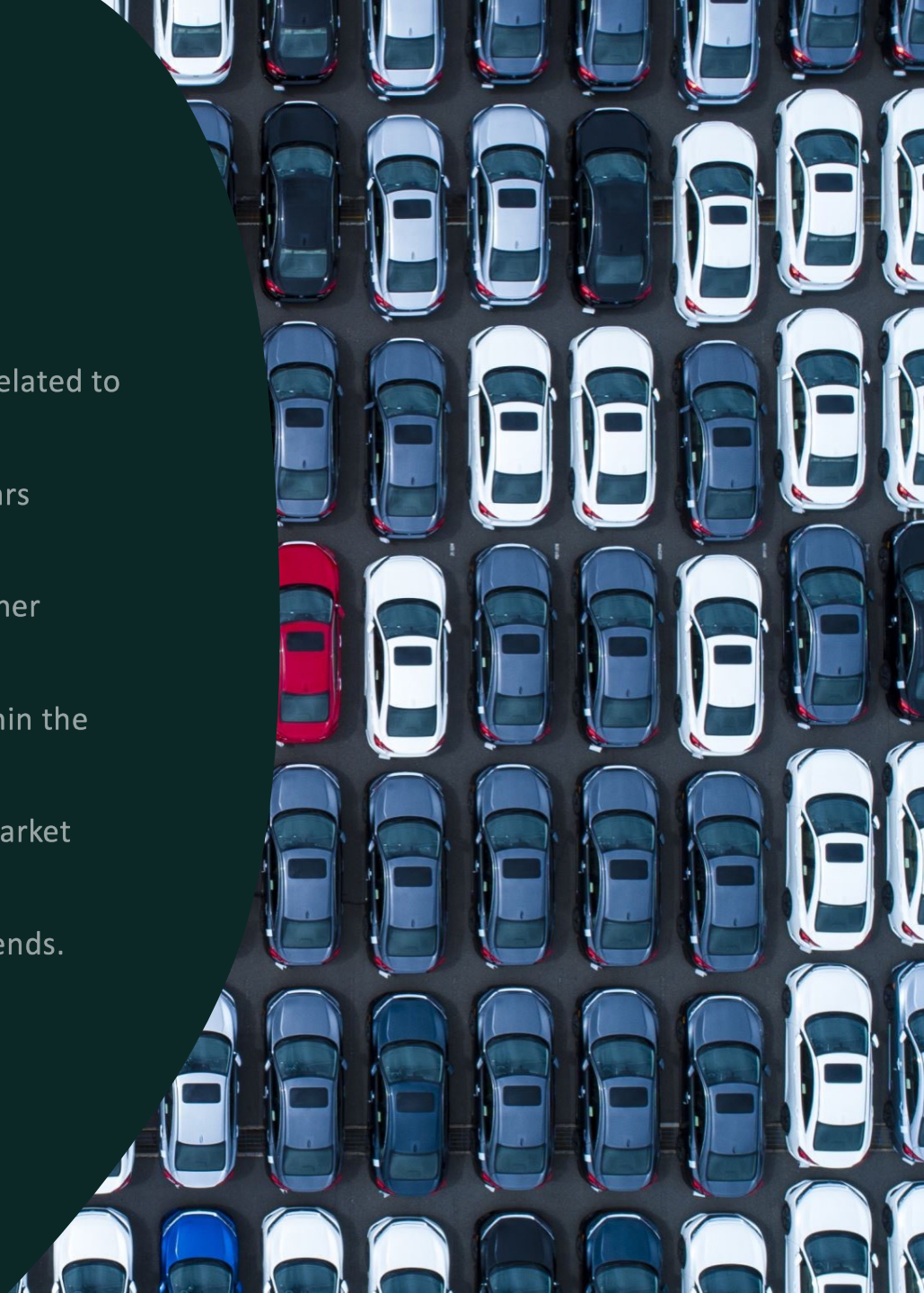
The data warehouse includes an overview of used cars and their individual attributes, facilitating quick and efficient searches for users.

Users can easily find the best car that matches their requirements, streamlining the car buying process.

The data warehouse benefits both car sellers and buyers, serving as a valuable resource for anyone involved in buying or selling a car.

# Conclusion

- The comprehensive analysis of the data warehouse reveals insights about various aspects related to the purchase of used cars.
- A correlation is observed between a car's age and its probability of being sold, except for cars exceeding 23 years, which often possess classic or vintage status.
- Ford demonstrates a dominant position in the used car resale market, aligning with the higher incidence of accidents involving Ford vehicles.
- Gasoline emerges as the primary fuel type among car owners, indicating its preference within the industry.
- These findings provide valuable insights into the dynamics of the used car market, Ford's market presence, accident trends, and fuel preferences.
- The analysis contributes to a deeper understanding of consumer behaviour and industry trends.



# References


- (1) Used Cars Dataset. kaggle. (https://www.kaggle.co m/datasets/used-cars-dataset)

Thank You for your attention !

Questions ?