Samantha

- The automotive industry has faced a shortage in the Semiconductor Integrate Chips globally
- Semicoductor IC is a critical component for controlling several electronic devices in the vehicle
- Car sales industry is making up for the shortage by raising their APR and prices
- Increasing demand of used cars which is making the prices of used cars higher as well
- Give an example and talk about how buying a car works
- Limit our studies and findings for US market only
- Dataset collected from Kaggle
- Analyzed different manufacturers and years
- Price is our target variable
- Go over questions

Ryoichi

- Data included more than one hundred thousand used car price data with other features such as id, year, manufacturer, condition, odometer, number of cylinders, and state.
- Eliminated used cars older than the year 2000 and used cars that drove more than 200,000 miles as we believed nobody wants to buy such cars.
- Removed used cars that are valued more than 100000 for the prediction purpose.

- As feature engineering, categorized the state input into four new categories including northeast, Midwest, south, and west to simplify the information.
- For the number of cylinders, remove the word cylinders to make it a numerical value.
- Get rid of those rows that include null values to finalize the dataset we are going to use for analysis.
- The tools we used in the project:

1. Python as programming language for data processing and machine learning
2. Postgres PgAdmin to store the database
3. Tableau as well as Python library for visualization

Reviewed the relationship between mileage based on odometer and the price.

The more the car runs, the more the car becomes worn away, which will make the used car value lower.

This hypothesis is correct, even though we had some irregular prices.

No longer contains cars that ran more than 200,000 miles as a result of data processing.

Matthew

- Go over dashboard
- Explain interactive element
- Go over charts from dashboard

Shahla

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application. It is used by the Train Using Auto ML tool and classifies or regresses the data using true or false answers to certain questions. The resulting

structure, when visualized, is in the form of a tree with different types of nodes—root, internal, and leaf.

one of the important Advantage of Decision tree is that it can be used for both regression and classification problem also it requires less effort for data preparation during pre-processing

Disadvantage: A small change in the data can cause a large change in the structure of the decision tree causing instability.

Machine learning models are the mathematical engines of Artificial Intelligence, expressions of algorithms that find patterns and make predictions faster than a human can.

There are two types of Machine Learning:

- Supervised Machine Learning: It is an ML technique where models are trained on labeled data i.e., output variable is provided in these types of problems. Here, the models find the mapping function to map input variables with the output variable or the labels.
- Regression and Classification problems are a part of Supervised Machine Learning.
- Unsupervised Machine Learning: It is the technique where models are not provided with the labeled data and they have to find the patterns and structure in the data to know about the data.
- Clustering and Association algorithms are a part of Unsupervised ML. For our project we are using Supervised Machine learning.

Linear Regression may be one of the most commonly used models in the real world. It is a linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). Linear regression is used in everything from biological, behavioral, environmental and social sciences to business.

Advantage: Linear Regression is easier to implement, interpret and very efficient to train.
Disadvantage: Linear Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.

Results:
The r-squared result from the linear regression model is R^2 train: 0.757, test: 0.754 meaning almost 76% of the training and 75% of the testing data observed variation can be explained by the model's inputs.
Recommendations/Final Thoughts
After evaluating both models, we find the Decision Tree Regressor model better fit for our dataset as its r square value is higher than in the Linear Regression Model.