THINK!

Application :

GOAL:

- Build, train, test, and deploy a machine learning regression model to predict used car prices based on their features

PRACTICAL REAL-WORLD APPLICATION:

- This project can be effectively used by car dealerships to predict used car prices and understand key factors that contribute to used car prices.

DATA:

- **INPUTS:**
  - Make, Model, Type, Origin, Drivetrain, Invoice, EngineSize, Cylinders, Horsepower, MPG_City, MPG_Highway, Weight, Wheelbase, and Length

- **OUTPUT:**
  - MSRP (Price)

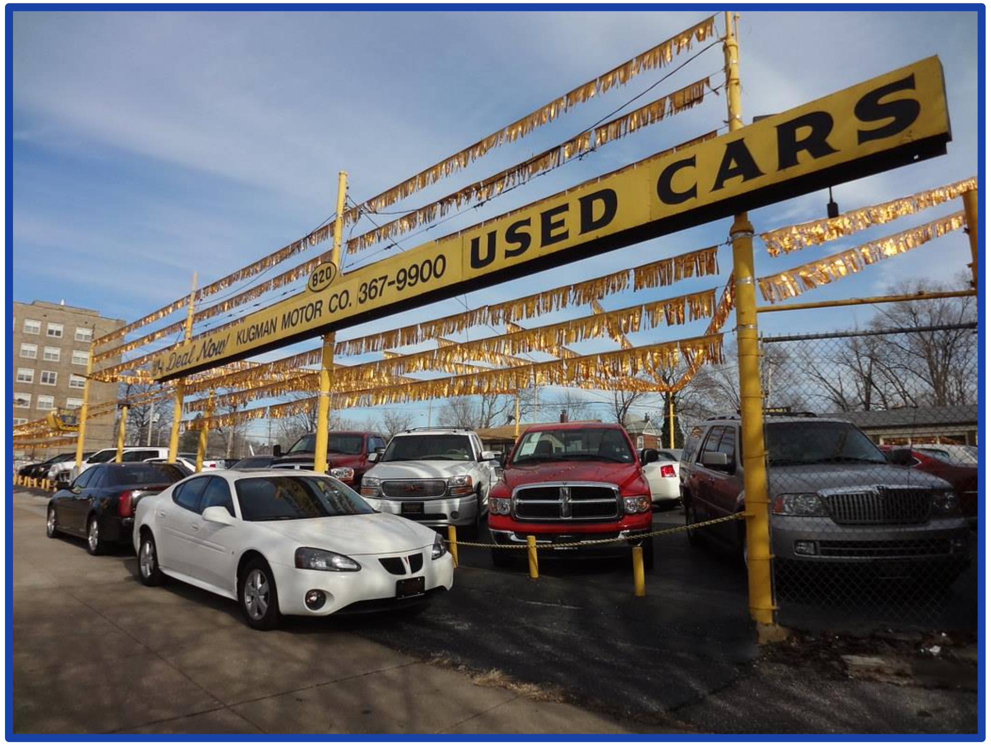
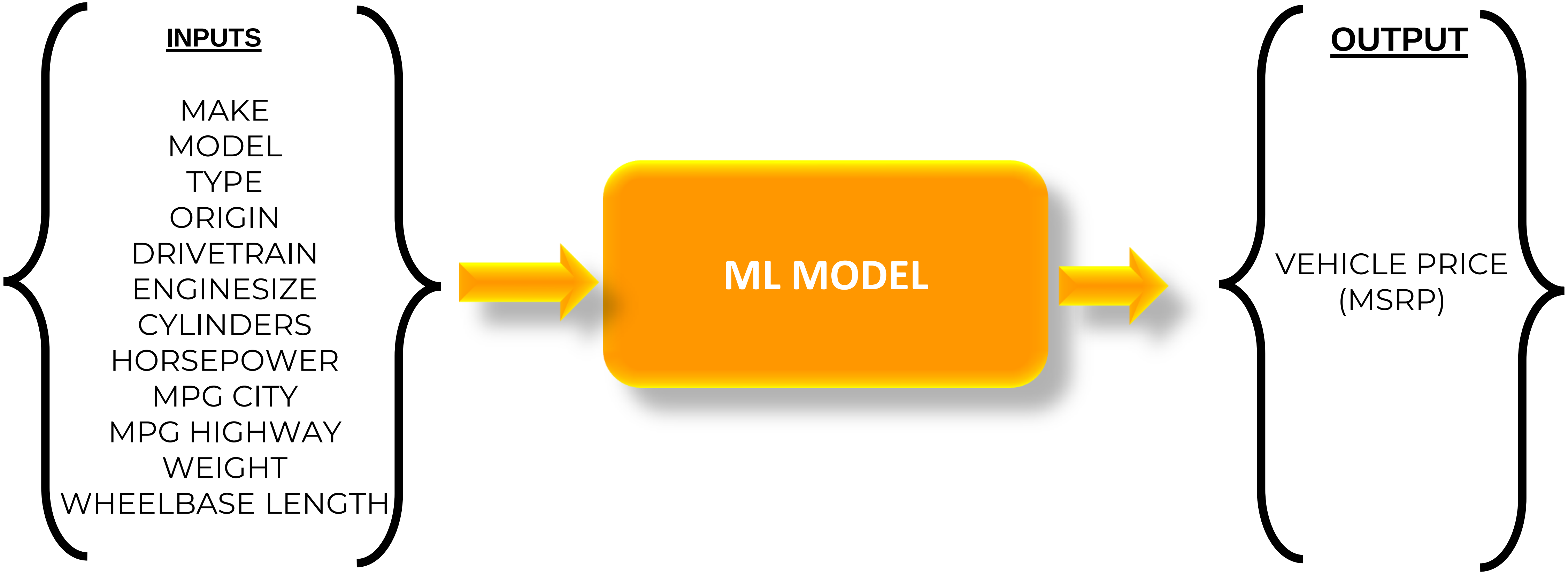


Image Source: https://www.flickr.com/photos/pasa/6757993805
Dataset Source: https://www.kaggle.com/ljanjughazyan/cars1

INPUTS AND OUTPUTS



INPUTS

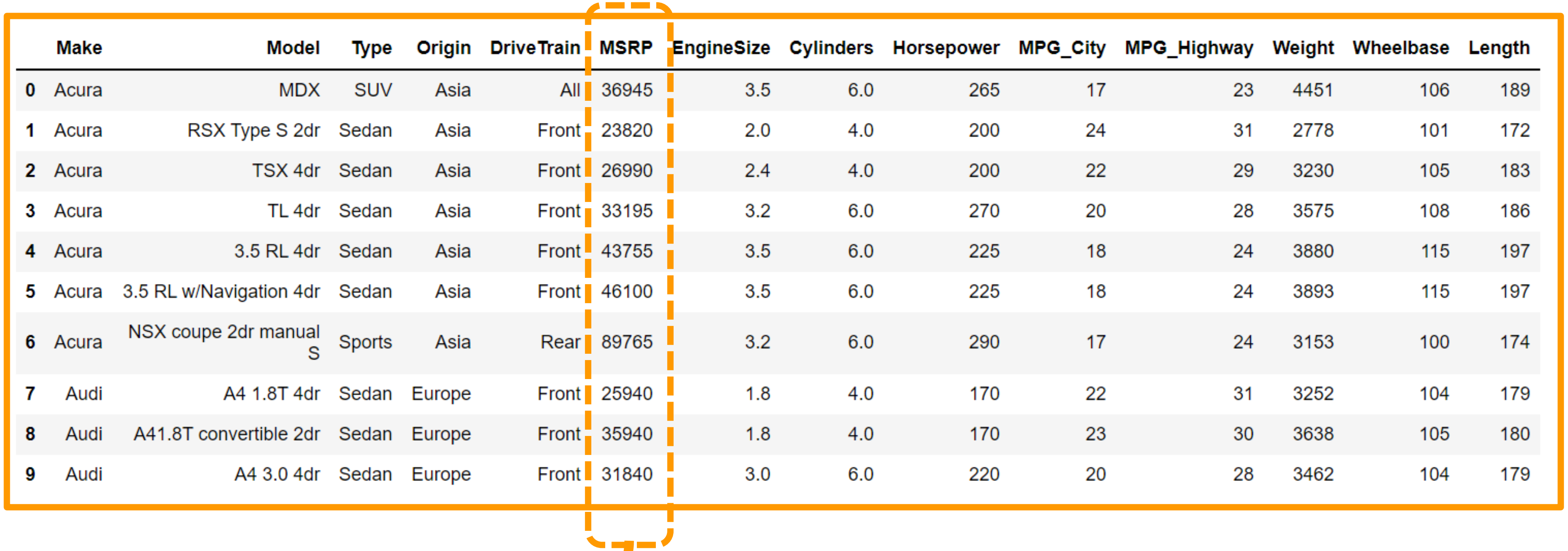
MAKE
MODEL
TYPE
ORIGIN
DRIVETRAIN
ENGINESIZE
CYLINDERS
HORSEPOWER
MPG CITY
MPG HIGHWAY
WEIGHT
WHEELBASE LENGTH

ML MODEL

OUTPUT

VEHICLE PRICE
(MSRP)

DATA OVERVIEW

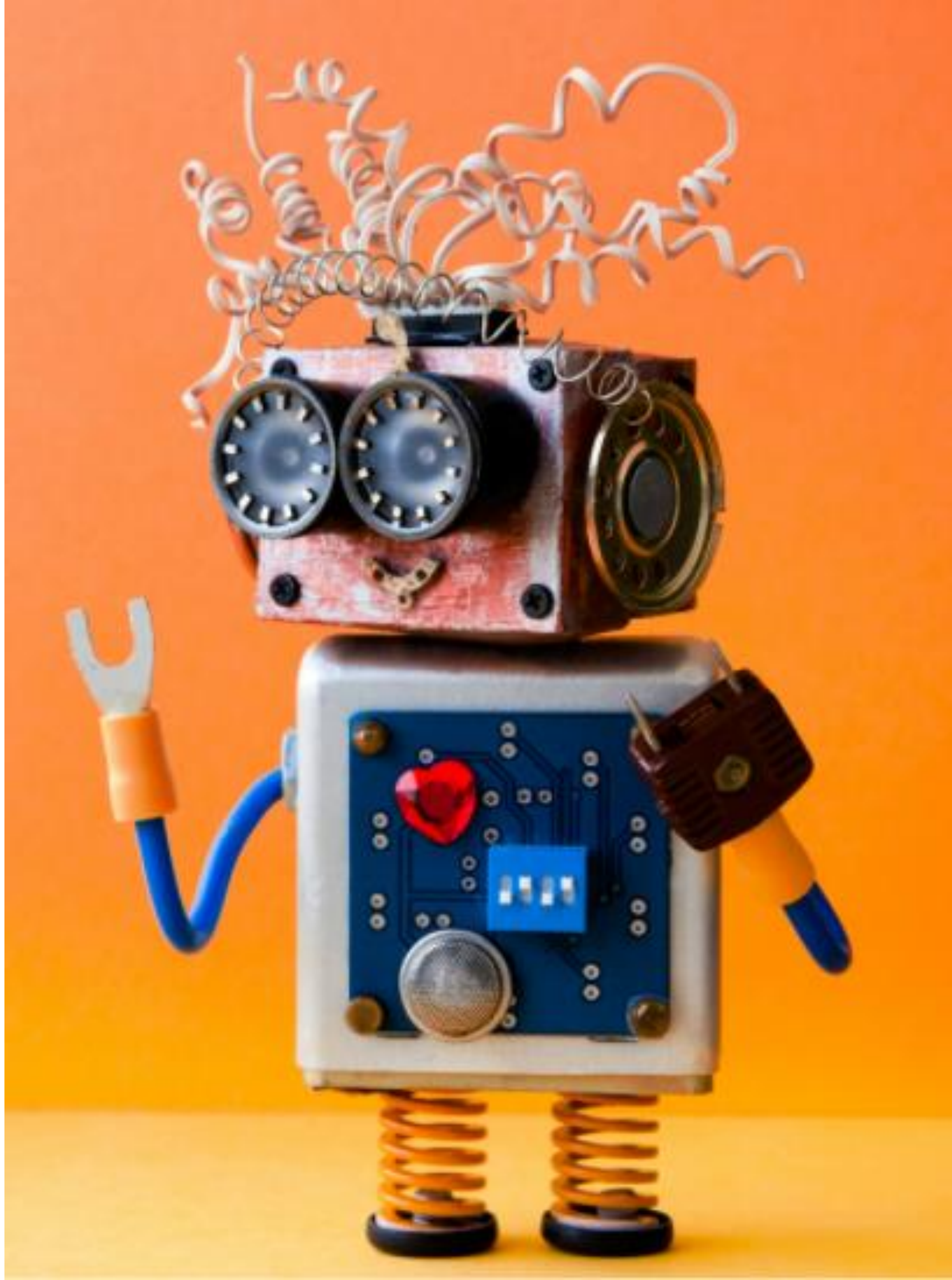


	Make	Model	Type	Origin	DriveTrain	MSRP	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
0	Acura	MDX	SUV	Asia	All	36945	3.5	6.0	265	17	23	4451	106	189
1	Acura	RSX Type S 2dr	Sedan	Asia	Front	23820	2.0	4.0	200	24	31	2778	101	172
2	Acura	TSX 4dr	Sedan	Asia	Front	26990	2.4	4.0	200	22	29	3230	105	183
3	Acura	TL 4dr	Sedan	Asia	Front	33195	3.2	6.0	270	20	28	3575	108	186
4	Acura	3.5 RL 4dr	Sedan	Asia	Front	43755	3.5	6.0	225	18	24	3880	115	197
5	Acura	3.5 RL w/Navigation 4dr	Sedan	Asia	Front	46100	3.5	6.0	225	18	24	3893	115	197
6	Acura	NSX coupe 2dr manual S	Sports	Asia	Rear	89765	3.2	6.0	290	17	24	3153	100	174
7	Audi	A4 1.8T 4dr	Sedan	Europe	Front	25940	1.8	4.0	170	22	31	3252	104	179
8	Audi	A41.8T convertible 2dr	Sedan	Europe	Front	35940	1.8	4.0	170	23	30	3638	105	180
9	Audi	A4 3.0 4dr	Sedan	Europe	Front	31840	3.0	6.0	220	20	28	3462	104	179

MODEL OUTPUT: MSRP

MANUFACTURER'S SUGGESTED RETAIL PRICE

SUCCESS STORIES



SUCCESS STORIES

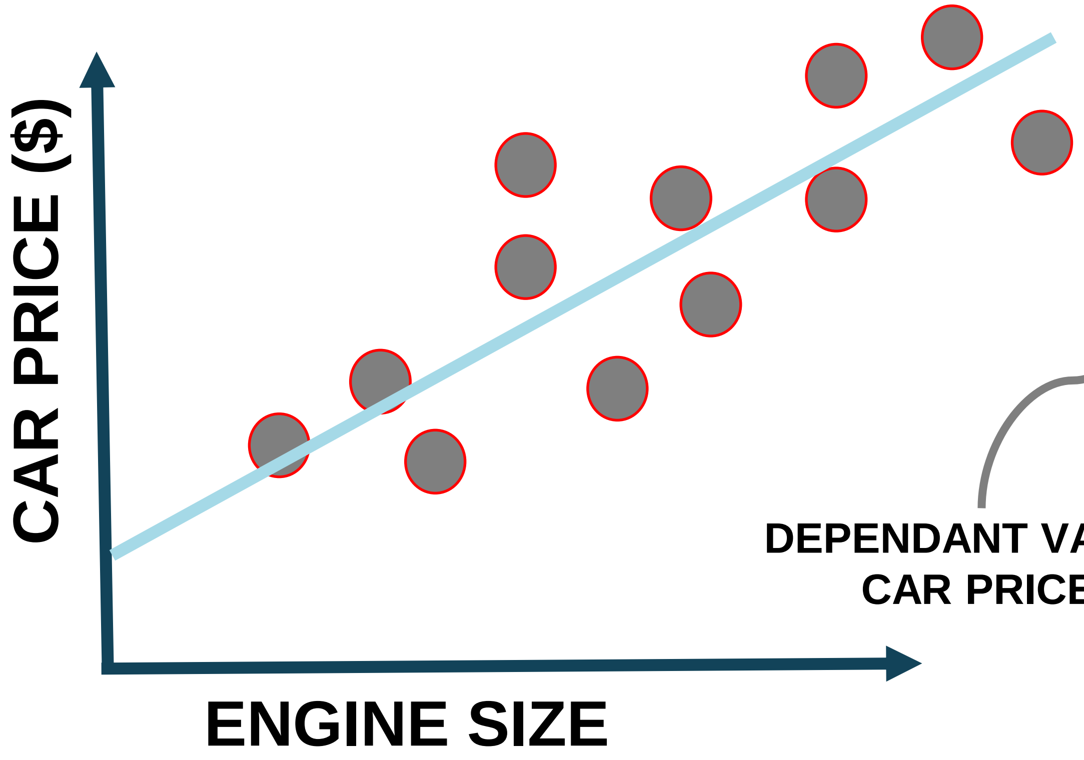
- Price prediction of products and services is critical for any company to maximize revenues and reduce costs.
- Fareboom.com is an innovative tool that leverages machine learning to predict flight prices. The tool has been developed by AltexSoft.
- The fare forecast feature has been developed to help users make better purchasing decisions.
- The tool can guide customers to select the best time to purchase a flight.
- The tool is built on a self-learning machine-learning algorithm that can predict future price movements while taking into account historical data, airline deals, demand, and seasonal effects.
- Great case studies: https://www.altexsoft.com/case-studies/
- Fare price prediction tool: https://www.altexsoft.com/case-studies/travel/altexsoft-creates-unique-data-science-and-analytics-based-fare-predictor-tool-to-forecast-price-movements/

Source: https://www.altexsoft.com/blog/datascience/data-science-and-ai-in-the-travel-industry-9-real-life-use-cases/

MULTIPLE LINEAR REGRESSION

RECALL SIMPLE LINEAR REGRESSION?

- Goal is to obtain a relationship (model) between two variables only such as age and insurance cost for example.



MODEL! (GOAL)

$$y = \boxed{b} + \boxed{m} * x$$

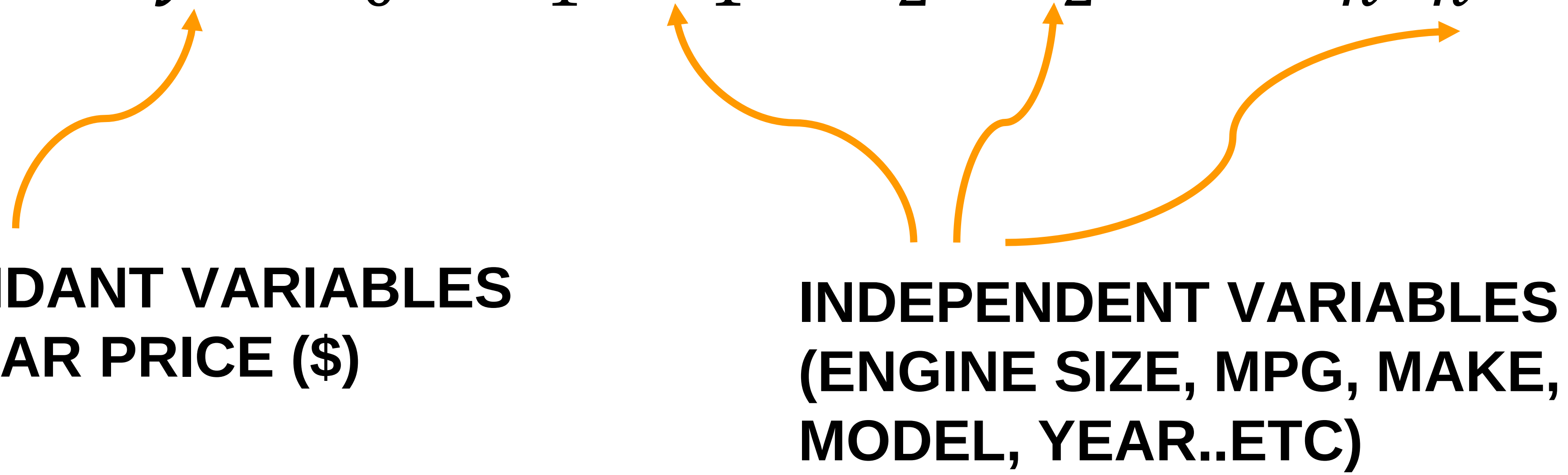
DEPENDANT VARIABLE
CAR PRICE ($)

INDEPENDENT VARIABLE
ENGINE SIZE

MULTIPLE LINEAR REGRESSION: INTUITION

- Multiple Linear Regression: examines relationship between more than two variables.
- Recall that Simple Linear regression is a statistical model that examines linear relationship between two variables only.
- Each independent variable has its own corresponding coefficient.

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + .. + b_n x_n$$



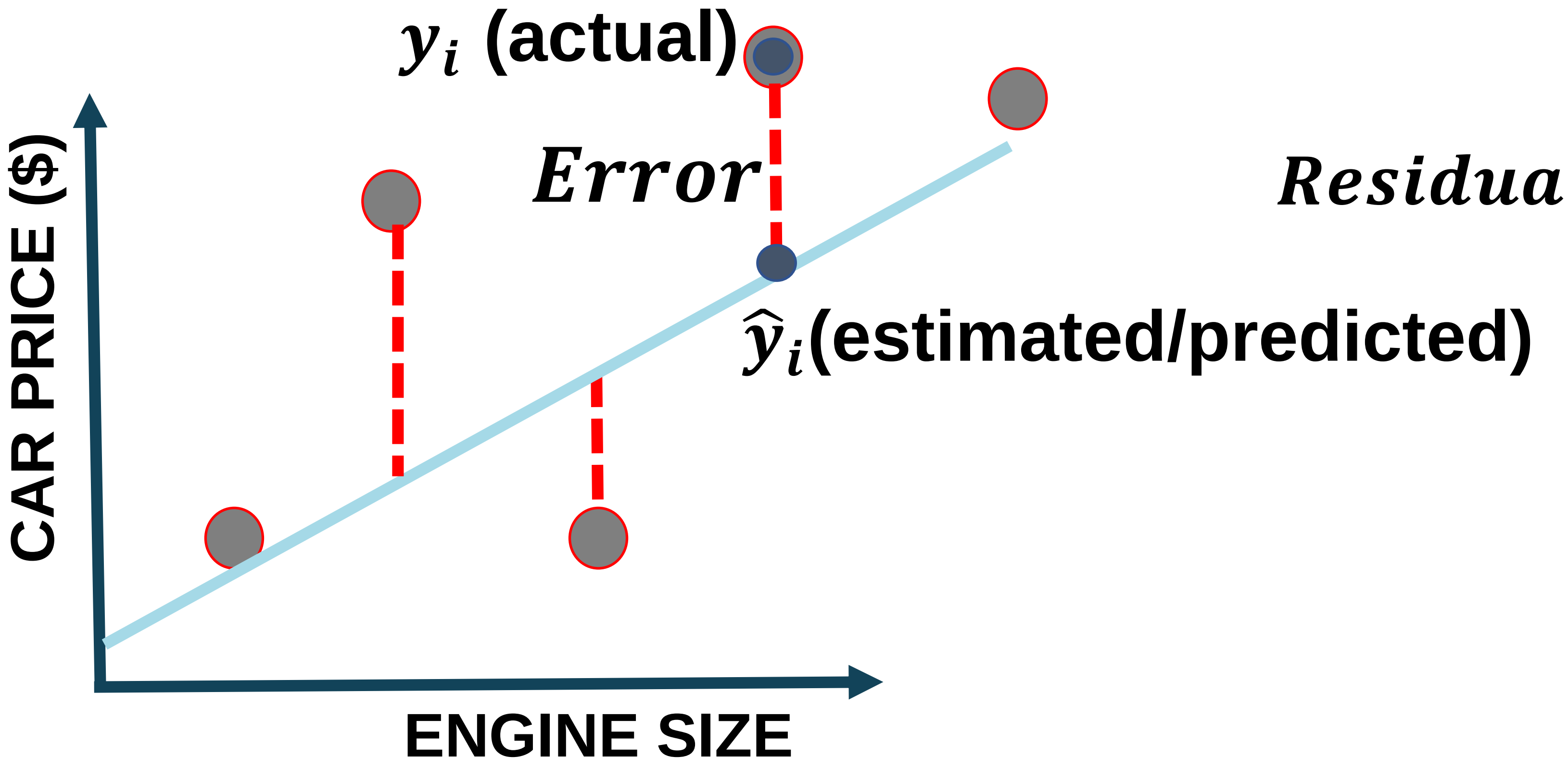
DEPENDANT VARIABLES
CAR PRICE ($)

INDEPENDENT VARIABLES
(ENGINE SIZE, MPG, MAKE, MODEL, YEAR..ETC)

REGRESSION METRICS AND KPIs

REGRESSION METRICS: HOW TO ASSESS MODEL PERFORMANCE?

- After model fitting, we would like to assess the performance of the model by comparing model predictions to actual (True) data



$$Residuals\ (Error) = \hat{y}_i - \mathbf{y}_i$$

REGRESSION METRICS: MEAN ABSOLUTE ERROR (MAE)

- Mean Absolute Error (MAE) is obtained by calculating the absolute difference between the model predictions and the true (actual) values
- MAE is a measure of the **average magnitude of error** generated by the regression model
- The mean absolute error (MAE) is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- MAE is calculated by following these steps:
  1. Calculate the residual of every data point
  2. Calculate the absolute value (to get rid of the sign)
  3. Calculate the average of all residuals
- If MAE is zero, this indicates that the model predictions are perfect.

REGRESSION METRICS: MEAN SQUARE ERROR (MSE)

- Mean Square Error (MSE) is very similar to the Mean Absolute Error (MAE) but instead of using absolute values, squares of the difference between the model predictions and the training dataset (true values) is being calculated.
- MSE values are generally **larger** compared to the MAE since the **residuals are being squared**.
- In case of data outliers, MSE will become much larger compared to MAE.
- In MSE, error increases in a **quadratic fashion** while the error increases in **proportional fashion in MAE.**
- In MSE, since the error is being squared, prediction error is being heavily penalized.
- The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- MSE is calculated by following these steps:
  1. Calculate the residual for every data point
  2. Calculate the squared value of the residuals
  3. Calculate the average of results from step #2

REGRESSION METRICS: ROOT MEAN SQUARE ERROR (RMSE)

- Root Mean Square Error (RMSE) represents the **standard deviation of the residuals** (i.e.: differences between the model predictions and the true values (training data)).
- RMSE can be **easily interpreted** compared to MSE because **RMSE units match the units of the output**.
- The RMSE is calculated as follows:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}$$

- RMSE is calculated by following these steps:
  1. Calculate the residual for every data point
  2. Calculate the squared value of the residuals
  3. Calculate the average of the squared residuals
  4. Obtain the square root of the result

REGRESSION METRICS: MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

- Mean Absolute Percentage Error (MAPE) is the equivalent to MAE but provides the error in a percentage form and therefore overcomes MAE limitations.
- Issues with MAE: Since MAE values can range from 0 to infinity which makes it difficult to interpret the result as compared to the training data.
- MAPE might exhibit some limitations if the data point value is zero (since there is division operation involved)
- The MAPE is calculated as follows:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} |(y_i - \hat{y}_i)/y_i|$$

REGRESSION METRICS: MEAN PERCENTAGE ERROR (MPE)

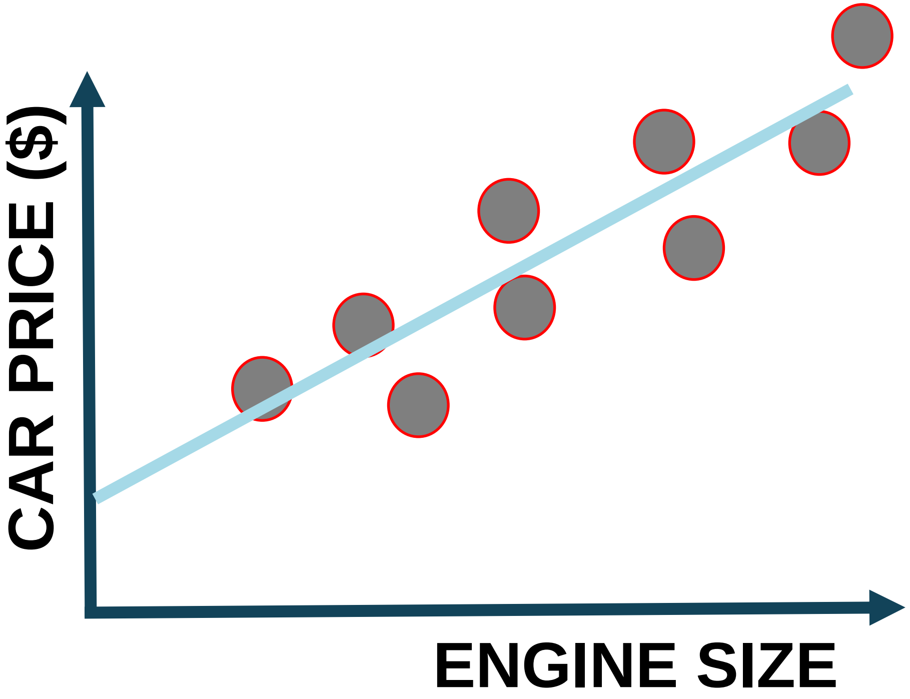
- MPE is similar to MAPE but without the absolute operation
- MPE is useful to provide an insight of how many positive errors as compared to negative ones
- The MPE is calculated as follows:

$$MPE = \frac{100\%}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)/y_i$$

REGRESSION METRICS AND KPIs

REGRESSION METRICS: R SQUARE (R^2)-COEFFICIENT OF DETERMINATION

- R-square or the coefficient of determination represents the proportion of variance (of y) that has been explained by the independent variables in the model.
- If $R^2 = 80$, this means that 80% of the increase in the car price is due to increase in the engine size.



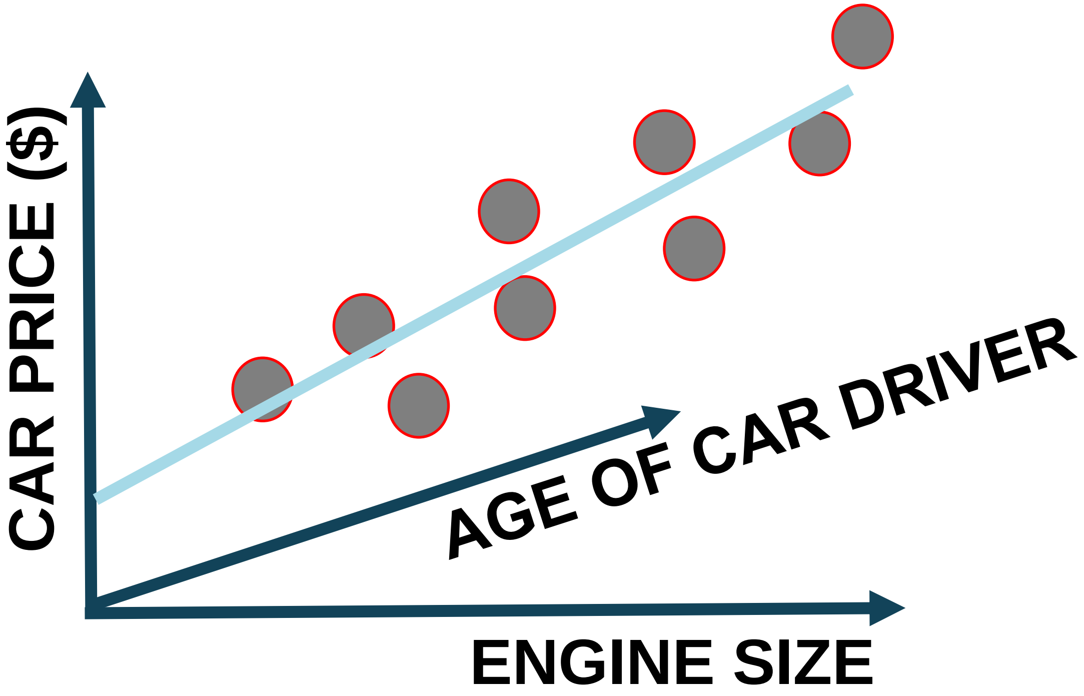
REGRESSION METRICS: R SQUARE (R^2)-COEFFICIENT OF DETERMINATION

- R-square represents the proportion of variance of the dependant variable (y) that has been explained by the independent variables.
- R-square provides an insight of **goodness of fit**.
- It gives a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance.
- Maximum R^2 value is 1
- A constant model that always predicts the expected value of y, disregarding the input features, will have an R^2 score of 0.0.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

REGRESSION METRICS: ADJUSTED R SQUARE (R^2)

- If $R^2 = 80$, this means that 80% of the increase in the car's price is due to increase in engine size.
- Let's add another 'useless' independent variable, let's say "age of the car driver" to the Z-axis.
- Now R^2 increases and becomes: $R^2 = 85\%$



REGRESSION METRICS: ADJUSTED R SQUARE (R^2)

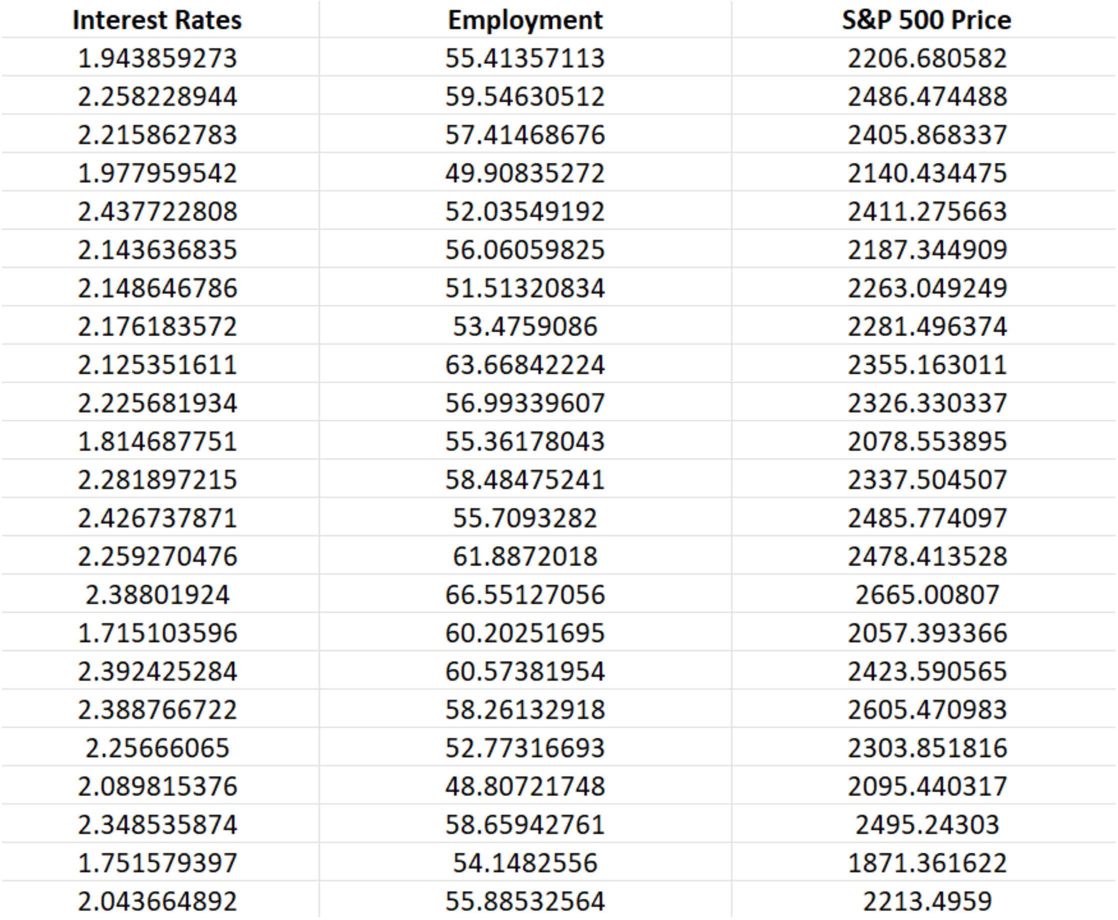
- One limitation of R^2 is that it increases by adding independent variables to the model which is misleading since some added variables might be useless with minimal significance.
- Adjusted R^2 overcomes this issue by **adding a penalty** if we make an attempt to add independent variable that does not improve the model.
- Adjusted R^2 is a modified version of the R^2 and takes into account the **number of predictors in the model**.
- If useless predictors are added to the model, Adjusted R^2 will decrease
- If useful predictors are added to the model, Adjusted R^2 will increase
- K is the number of independent variables and n is the number of samples

$$R^2_{adjusted} = 1 - [\frac{(1 - R^2)(n - 1)}{n - k - 1}]$$

PROJECT

PROJECT

- We would like to predict the S&P500 Price using interest rate and employment.
  - Independent variable X: Interest Rate and Employment
  - Dependent variable Y: S&P 500 Price



Interest Rates	Employment	S&P 500 Price
1.943859273	55.41357113	2206.680582
2.258228944	59.54630512	2486.474488
2.215862783	57.41468676	2405.868337
1.977959542	49.90835272	2140.434475
2.437722808	52.03549192	2411.275663
2.143636835	56.06059825	2187.344909
2.148646786	51.51320834	2263.049249
2.176183572	53.4759086	2281.496374
2.125351611	63.66842224	2355.163011
2.225681934	56.99339607	2326.330337
1.814687751	55.36178043	2078.553895
2.281897215	58.48475241	2337.504507
2.426737871	55.7093282	2485.774097
2.259270476	61.8872018	2478.413528
2.38801924	66.55127056	2665.00807
1.715103596	60.20251695	2057.393366
2.392425284	60.57381954	2423.590565
2.388766722	58.26132918	2605.470983
2.25666065	52.77316693	2303.851816
2.089815376	48.80721748	2095.440317
2.348535874	58.65942761	2495.24303
1.751579397	54.1482556	1871.361622
2.043664892	55.88532564	2213.4959

PROJECT

Using the Jupyter notebook “*Multiple Linear Regression with SKLearn - Project Skeleton*”, perform the following:

1. Load the “*S&P500_Stock_Data.csv*” dataset
2. Perform data visualization and basic exploratory data analysis
3. Split the data into 80% for training and 20% for testing
4. Train a machine linear regression model in Scikit-Learn
5. Assess trained model performance
6. Visualize the results in 3D