DigitalSkola

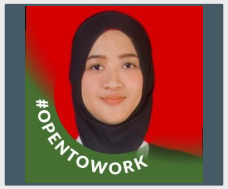
OTOMOTIC DATASET

By **VarieTeam**

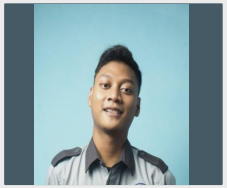
Kelas B
Selasa, 30 November 2021

PROFIL

Mentor

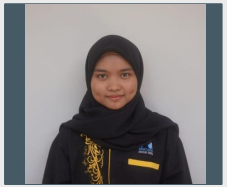


Lathifa Millati Saifullah



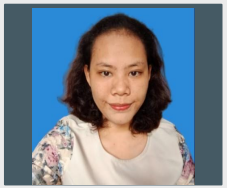
Ahmad Maulana
Sistem Informasi
Universitas Singaperbangsa Karawang.

Dwika Ananda Agustina
Sistem Informasi
Universitas Negeri Semarang.



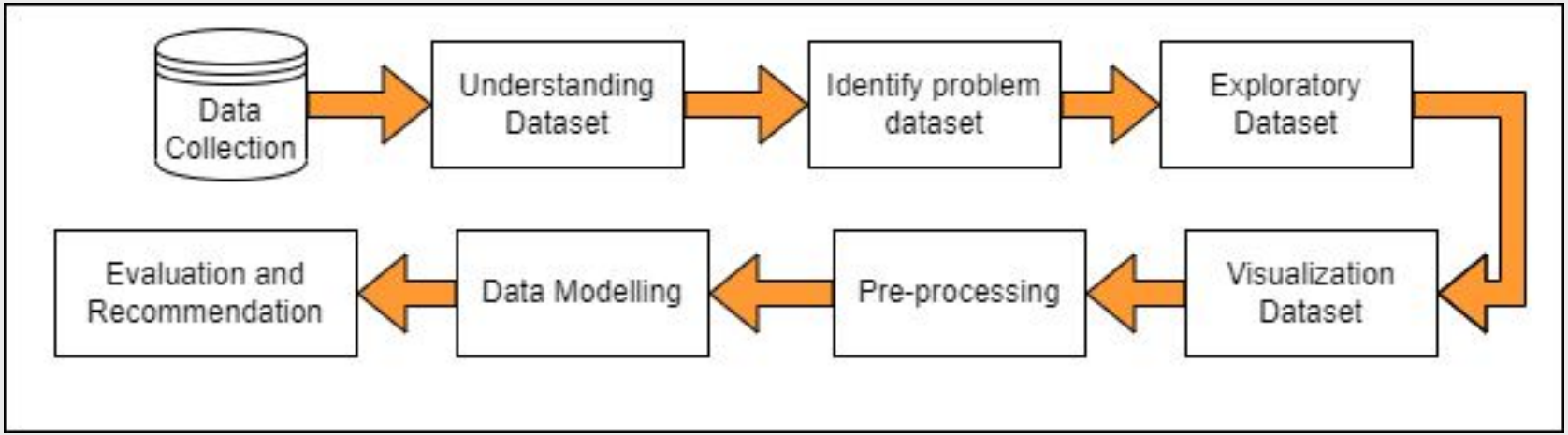
Linda Kushernawati
Informatika
Universitas Muhammadiyah Sidoarjo.

Novendy Farhanudin
Pendidikan Teknik Elektro
Universitas Pendidikan Indonesia.



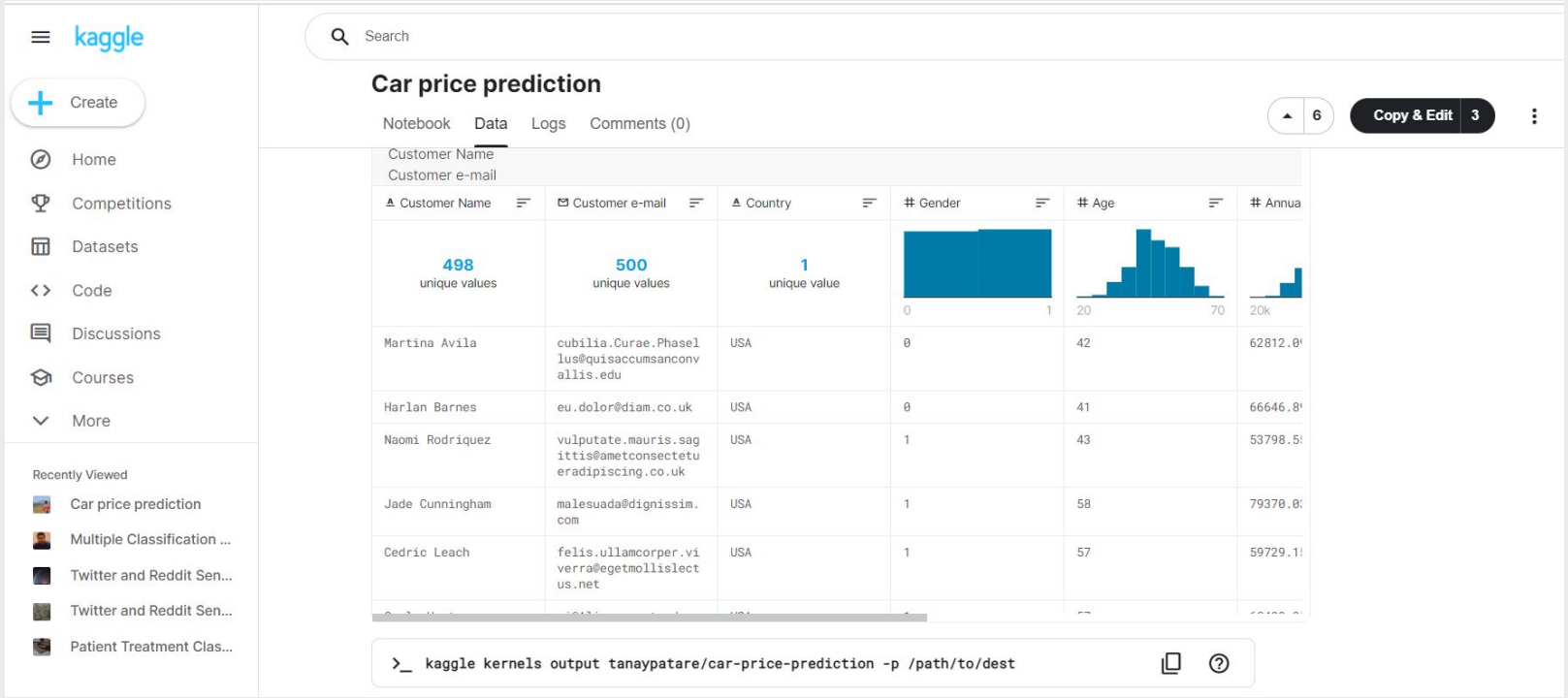
Restina Silalahi
Matematika
Universitas Negeri Medan.

STAGE IN DATASET & MODELLING



I. DATA COLLECTION

Dataset Car Price Prediction collect from **kaggle.com**, with **9 variable** and **500 record**.



II. UNDERSTANDING DATASET

BACKGROUND & PURPOSE

Along with the high activity and business, the car has become a basic need. The high level of public interest in cars makes this business increase, this is indicated by the number of car showrooms. Not to mention that showrooms are very competitive to compete in order to continue to exist in the car business.

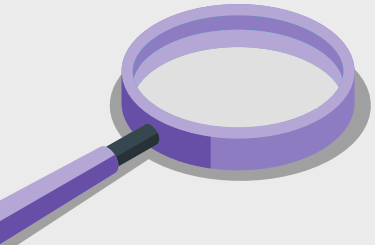
BACKGROUND & PURPOSE

One of the problems faced by all showrooms is to determine the price quickly and accurately so that the showroom can sell its merchandise and immediately get revenue. The current condition of car price predictions is still popular (Pandey, Rastogi, & Singh, 2020). Various showrooms compete with each other for prices to get customers.

The research objective is to get the best MAE score from machine learning models for predicting future car prices.

WHAT THE DATASET ABOUT

The **otomotic** dataset contains several columns represents customer data of car purchase.



DESCRIBE COLUMNS

Customer Name : Name of customers.

Customer e-mail : E-mail of customers.

Country : Country of customers.

Gender : Gender of customers. 0 means women, 1 means man.

Age : Age of customers.

DESCRIBE COLUMNS

Annual Salary : Annual salary of customers.

Credit Card Debt : Credit Card Debt of customers.

Net Worth : Liabilities of customers. Could also means

Assets (what you own) - Liabilities (what you owe).

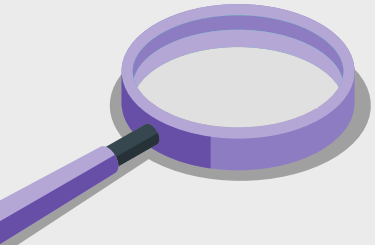
Car Purchase Amount : The price of amount car to purchase.

DATA TYPES

Columns	DataFrame	Statistics
Customer Name	object	nominal
Customer e-mail	object	nominal
Country	object	nominal
Gender	int64	nominal
Age	int64	continue
Annual Salary	float64	discrete
Credit Card Debt	float64	discrete
Net Worth	float64	discrete
Car Purchase Amount	float64	discrete

NULL VALUE

Customer Name	0
Customer e-mail	0
Country	0
Gender	0
Age	0
Annual Salary	0
Credit Card Debt	0
Net Worth	0
Car Purchase Amount	0



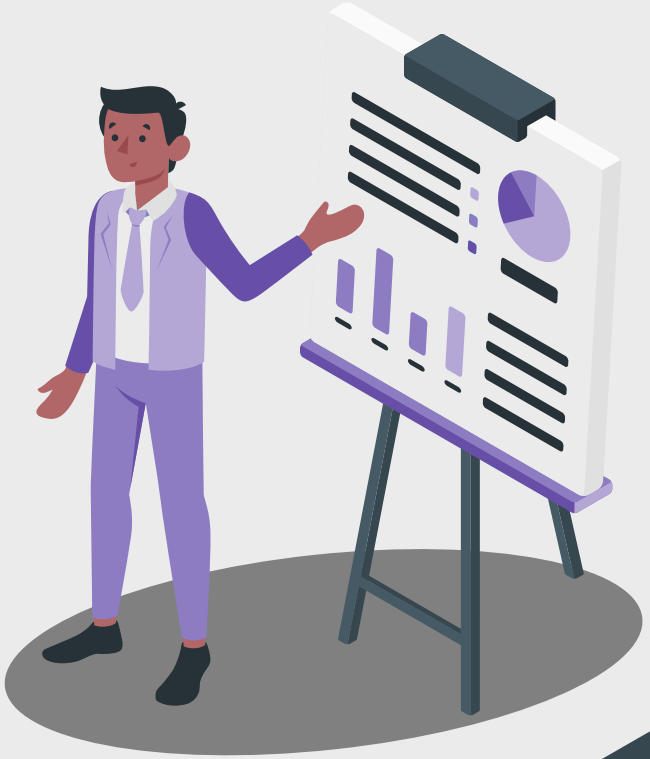
DATAFRAME DESCRIBE

DigitalSkola

```
df.describe()
```

	Gender	Age	Annual Salary	Credit Card Debt	Net Worth	Car Purchase Amount
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	0.506000	46.224000	62127.239608	9607.645049	431475.713625	44209.799218
std	0.500465	7.990339	11703.378228	3489.187973	173536.756340	10773.178744
min	0.000000	20.000000	20000.000000	100.000000	20000.000000	9000.000000
25%	0.000000	41.000000	54391.977195	7397.515792	299824.195900	37629.896040
50%	1.000000	46.000000	62915.497035	9655.035568	426750.120650	43997.783390
75%	1.000000	52.000000	70117.862005	11798.867487	557324.478725	51254.709517
max	1.000000	70.000000	100000.000000	20000.000000	1000000.000000	80000.000000

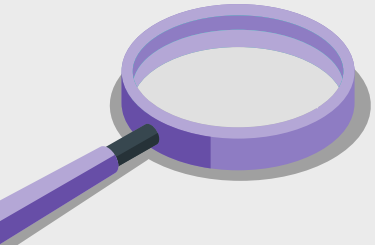
III. IDENTIFY REQUIRED ACTIVITIES



CHOOSE LABEL

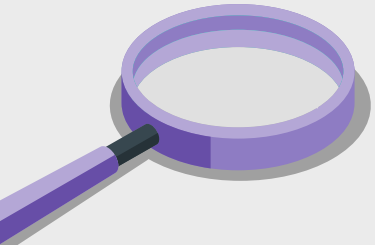
The dataset could be use to predict the price of a car someone would buy.

Hence, the **Car Purchase Amount** column should be used as the label.



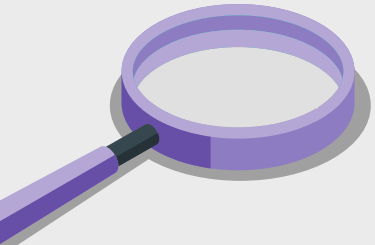
INDEPENDENT VARIABLE

- Gender
- Age
- Annual Salary
- Credit Card Debt
- Net Worth



TRAIN/TEST DATA

The dataset has 500 rows.
80% of it would be used as the **train** data.
And the rest, **20%** would be used as the **testing** data.



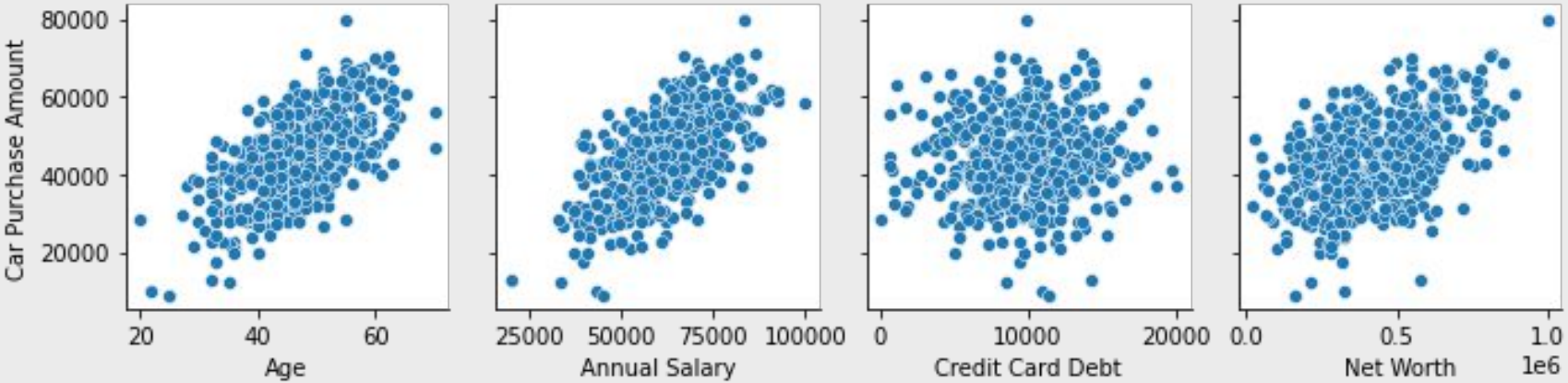
DigitalSkola

IV. EXPLORATORY DATA ANALYSIS & VISUALIZATION DATASET

DigitalSkola

VARIABLE CORRELATIONS

VARIABLE CORRELATIONS



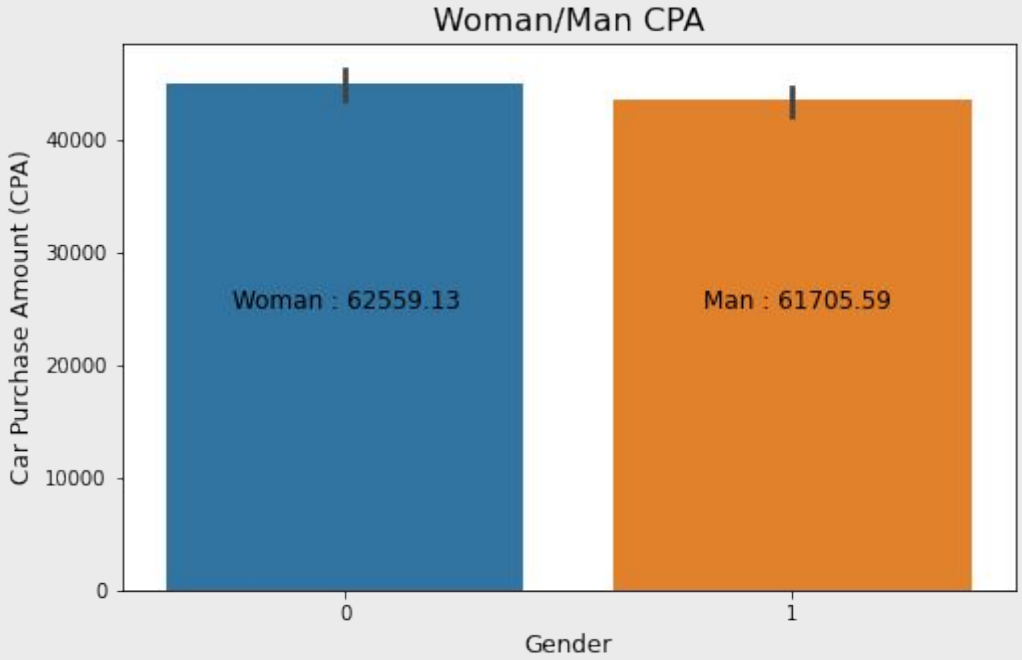
DigitalSkola

GENDER VS ANNUAL SALARY

Woman has *slightly* higher **Annual Salary** than Man

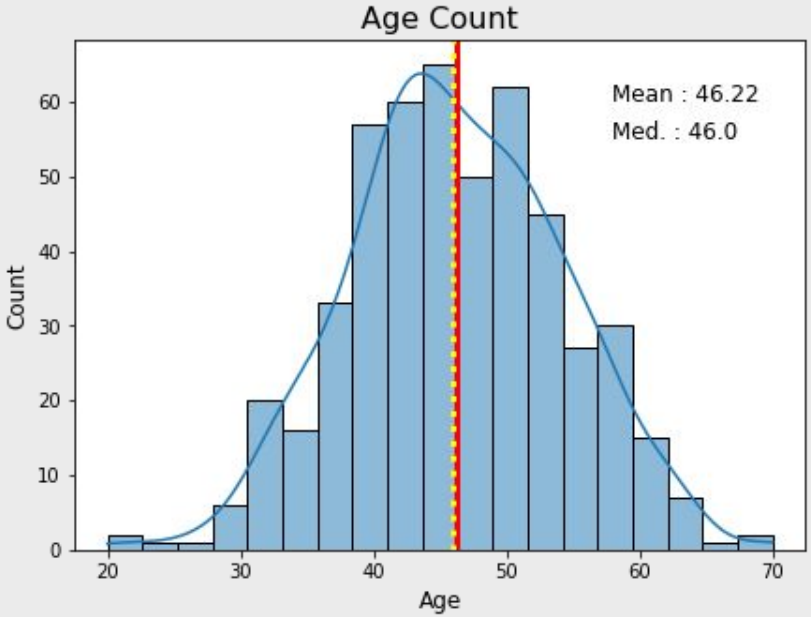
DigitalSkola

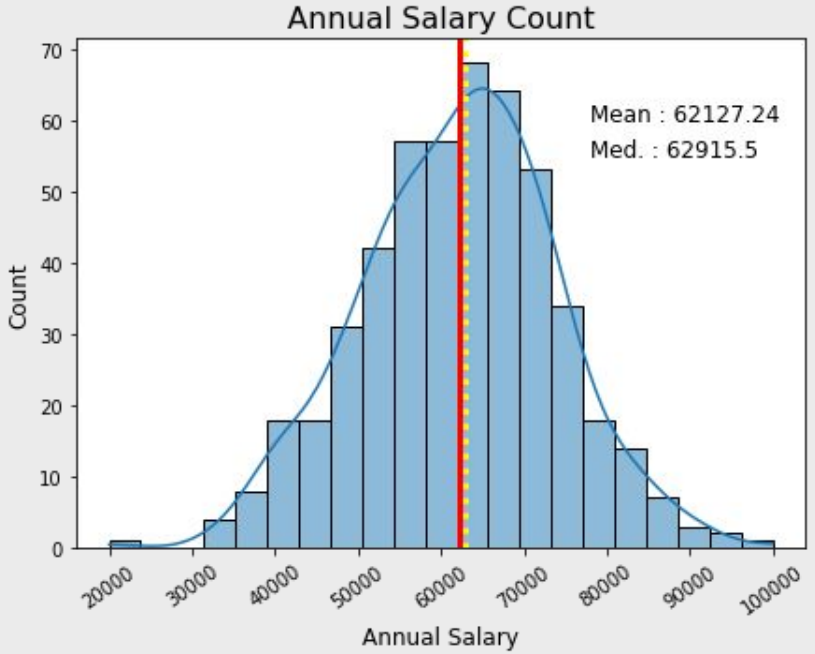
GENDER VS CAR PURCHASE AMOUNT



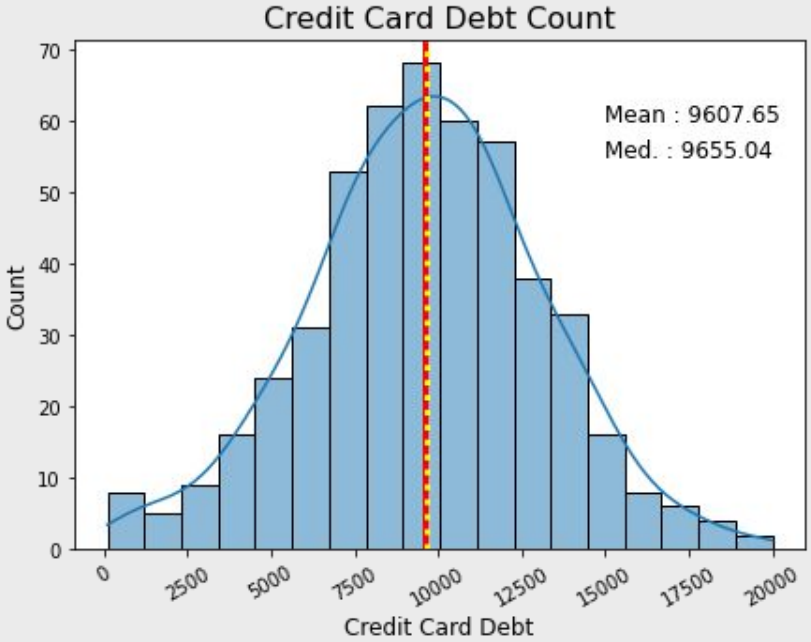
Woman has *slightly* higher **Car Purchase Amount** than Man

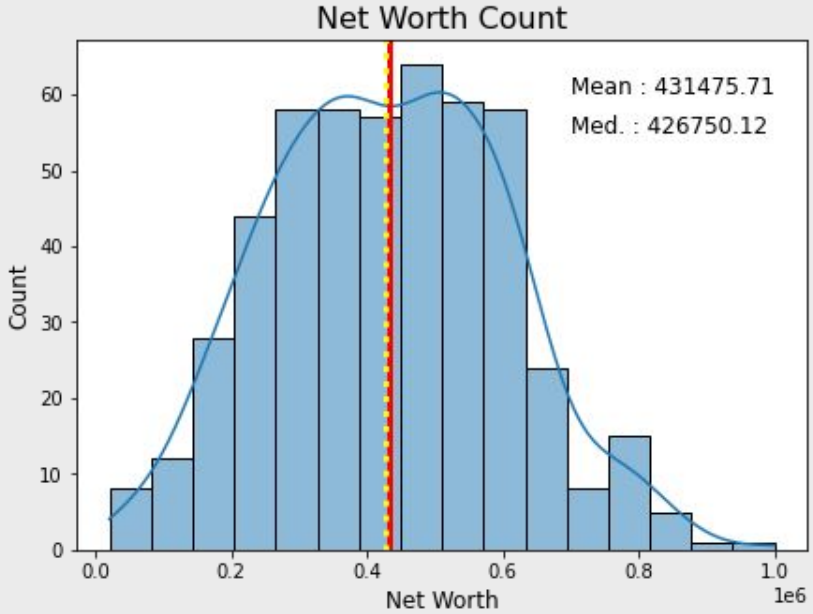
DATA DISTRIBUTION



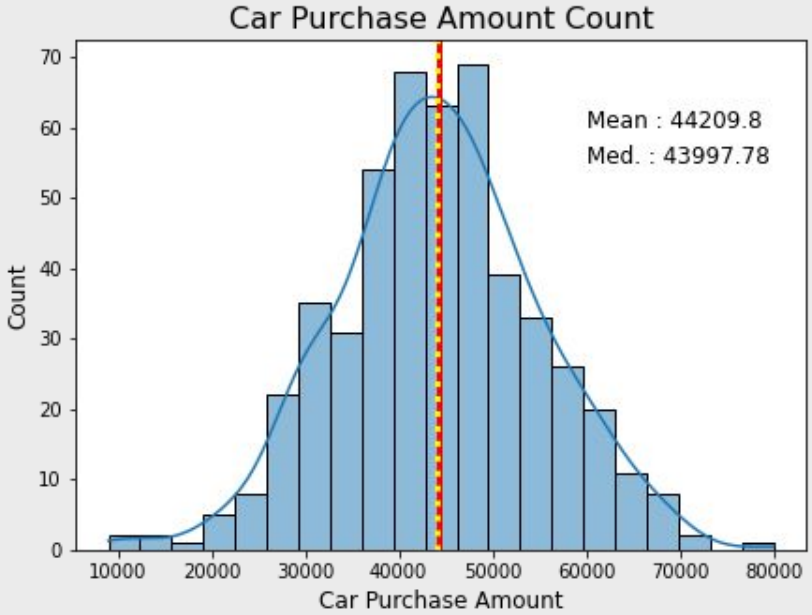


DATA DISTRIBUTION

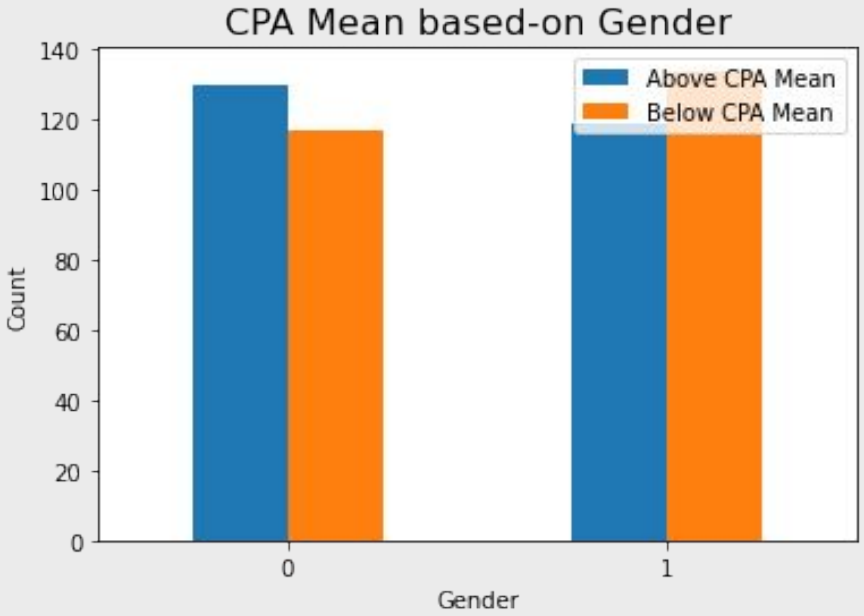




DATA DISTRIBUTION



CAR PURCHASE AMOUNT (ABOVE AND BELOW AVG)



DigitalSkola

V. DATA PRE-PROCESSING

Advanced Label Encoding: Gender

Drop Non-predictable Features

Train Test Split

```
X = df.drop(['Car Purchase Amount'], axis=1)
y = df['Car Purchase Amount']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

VI. DATA MODELLING

LINEAR REGRESSION

```
Train    : 209.8162618000001
Test     : 195.9006304999999
```

LASSO

```
Train    : 209.7764118286068
Test     : 195.62556049352838
```

ELASTIC NET

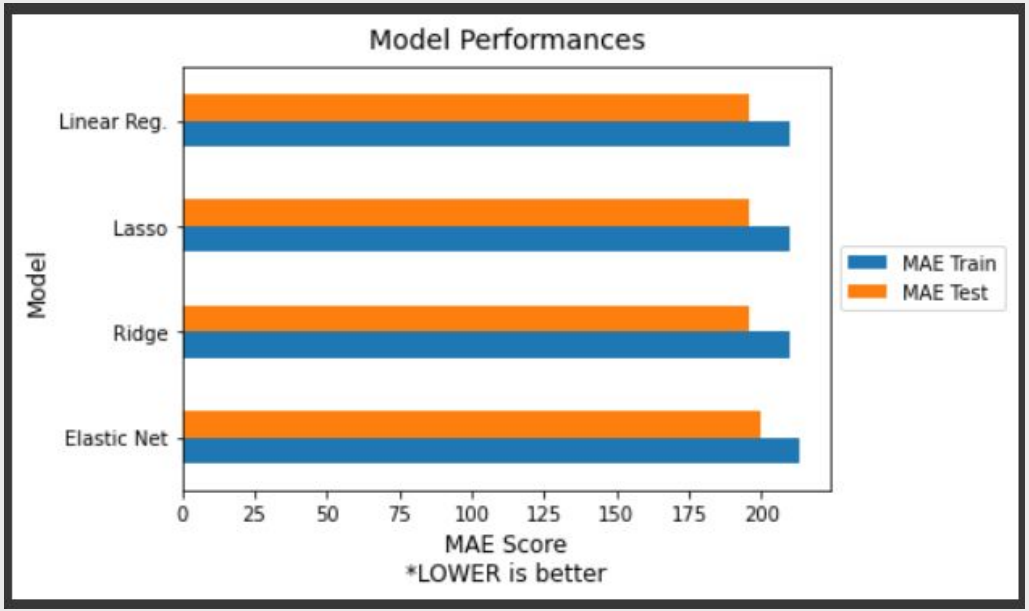
```
Train    : 213.4144494138161
Test     : 199.4939631737842
```

RIDGE

```
Train    : 209.77823246645187
Test     : 195.6294163352639
```

MODEL COMPARISON

	MAE Train	MAE Test
Elastic Net	213.414448	199.493962
Ridge	209.778232	195.629416
Lasso	209.776412	195.625560
Linear Reg.	209.762705	195.520630



VII. EVALUATION AND RECOMMENDATION

Get the best MAE score from machine learning models for predicting future car prices.

Recommendation model:
Linear Regression and **Lasso Regression**.

DigitalSkola

THANKS

Any further questions?

Don't be hesitate to reach us!