

DA Capstone (AutoScout) Intro_new_DE

Pear Deck Session

Training Clarusway

Pear Deck - March 10, 2023 at 10:47AM

Part 1 - Summary

Use this space to summarize your thoughts on the lesson

Part 2 - Responses

Slide 1



Use this space to take notes:

Slide 2

▶ Car Price Prediction EDA

CLARUSWAY®
WAY TO REINVENT YOURSELF



Use this space to take notes:

Slide 3

▶ Table of Contents

- ▶ Aim & Goals
- ▶ Big Picture
- ▶ Description
- ▶ What is expected of you?
- ▶ Need to Study
- ▶ Assumptions
- ▶ Hints



3

Use this space to take notes:

Slide 4

Your Response

You Chose

Slide 4

I've started working on the project and examined the dataset?

True

False

Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

Your Response

- **False**

Other Choices

- True

Use this space to take notes:

Slide 5

► Aim

- ▶ To get the dataset ready to provide an appropriate input to ML model predicting car prices by applying Exploratory Data Analysis (EDA) process.



Goals

- To ensure that all our students complete all projects.
- To increase soft skill abilities within the scope of project management (self-study, group work, time planning, task sharing, etc.).

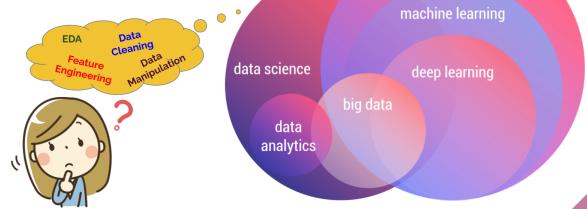


Use this space to take notes:

Slide 6

► Big Picture

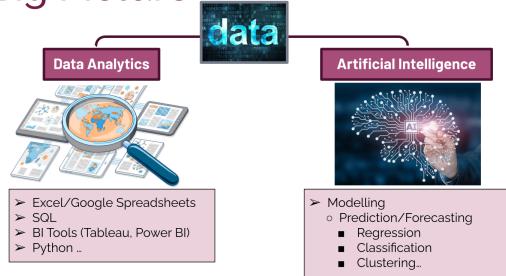
- ▶ Where am I?
- ▶ Why will I learn these?



Use this space to take notes:

Slide 7

► Big Picture



Use this space to take notes:

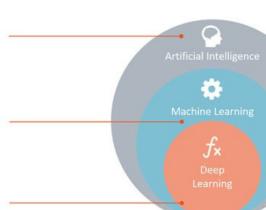
Slide 8

► Big Picture

Artificial Intelligence
Any technique which enables computers to mimic human behavior.

Machine Learning
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning
Subset of ML which make the computation of multi-layer neural networks feasible.

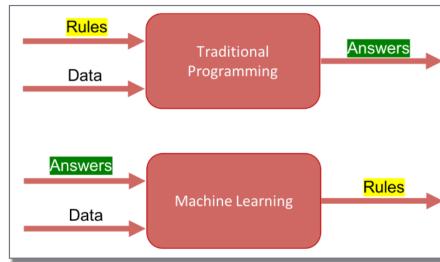


8

Use this space to take notes:

Slide 9

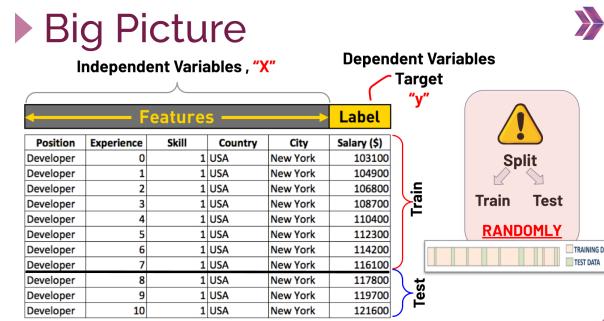
► Big Picture



9

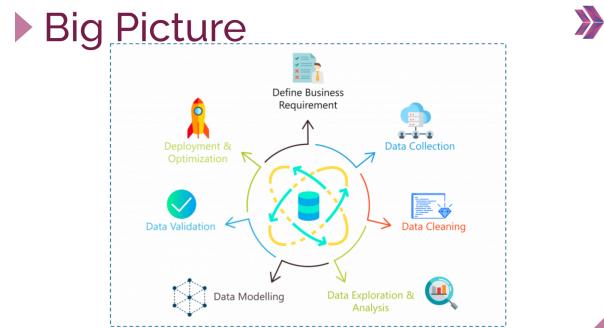
Use this space to take notes:

Slide 10



Use this space to take notes:

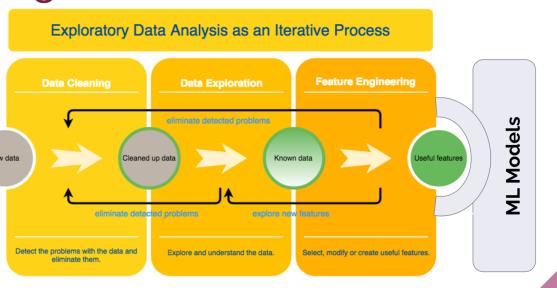
Slide 11



Use this space to take notes:

Slide 12

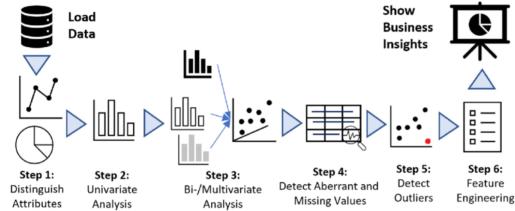
► Big Picture



Use this space to take notes:

Slide 13

► Big Picture



13

Use this space to take notes:

Slide 14

► Description ➤

- ▶ A ``.json`` file containing a dataset consisting of **29480 rows and 58 columns** is provided.
- ▶ This dataset, scraped from the online car trading company in 2022, contains many features of **various number of car models**.
- ▶ The features (variables) of this dataset are **too messy and distorted**.

14

Use this space to take notes:

Slide 15

► What is expected of you? ➤

- ▶ Read the ``.json`` file and assign the dataset into a **“DataFrame”** using “pandas”.
- ▶ Implement all aspects of the **“EDA process”** to the dataset.
 - Fix corrupted **data formats**
 - Handle with **missing values and outliers**
 - **Domain knowledge** (automobiles) is important
 - Always use the **internet** to do the research that you need (Domain Knowledge)
 - Think carefully to decide whether a data is **outliers or not**
 - **Drop the columns/rows** you determined unnecessary as a result of your analysis
 - Use **visualization tools** while doing all these processes

15

Use this space to take notes:

Slide 16

► What is expected of you? ➤



df_head(1), T

	0	1
url	https://www.autoscout24.com/offers/a1-a1-1.html	https://www.autoscout24.com/offers/a1-a1-1.html
make_model	Audi A1	Audi A1
body_type	Sedan	Sedan
price	14000	14000
km	95.07 km	Price negotiable
reg_date	05/2018	05/2018
perm_mile	2 previous owners	Vehicle history
VIN	N/A	N/A
hp	95 kW	95 kW
Type	1.4 liter, Diesel (Petrol/Electric) FWD	1.4 liter, Gasoline
Precious_Duration	1 day	1 day
Next_Inspection	(06/2021), valid p. 02/2021 (control)	N/A
Inspection_new	(06/2021), valid p. 02/2021 (control)	N/A
Fuel_Service	[n, n, n] (control)	N/A
Non-existing_Vehicle	[n, n]	N/A
Adult	2	1
Male	Infant	Infant
Model	[n, A1, N1]	N/A
Other_Make	14000	N/A
First_Registrat	[> 2010, n]	[> 2017, n]
Body_Color	[n, Black, n]	[n, Red, n]

df_head(1), T

	0	1
make_model	Audi A1	Audi A1
body_type	Sedan	Sedan
price	14000	14000
km	95.07 km	Price negotiable
reg_date	05/2018	05/2018
perm_mile	2 previous owners	Vehicle history
VIN	N/A	N/A
Type	1.4 liter, Diesel (Petrol/Electric) FWD	1.4 liter, Gasoline
Conflicting_Contract	0	0
Entertainment_Media	0	0
Exterior	0	0
Safety_Security	0	0
Interior	0	0
Powertrain	0	0
Drivetrain	0	0
Weight_kg	1220.000	1220.000
Drive_Axle	Front	Front
Front_Axle	Front	Front
Ctr_Fuel_economy	16.000	16.000

16

Use this space to take notes:

Slide 17

► What is expected of you? ➤



17

Use this space to take notes:

Slide 18

► What is expected of you? ➤

The diagram illustrates the transformation of a dataset. On the left, a table labeled `df_head[1:T]` shows the initial state of the data. On the right, a table labeled `df_final_head[1:T]` shows the final state after processing. An arrow points from the left table to the right table.

	0	1	2	3
make	Audi	Audi	Audi	Audi
body_type	Sedan	Sedan	Sedan	Sedan
price	52776	14010	14010	14010
age	10.000	10.000	10.000	10.000
mpg_diesel	0.0000	0.0000	0.0000	0.0000
km	50513.000	35500.000	35500.000	35500.000
Type	Used	Used	Used	Used
Fuel	Diesel	Diesel	Diesel	Diesel
Power	130.000	130.000	130.000	130.000
kmes	7.000	7.000	7.000	7.000
Cooling_Condition	No conditioning	Automatic	Automatic	Automatic
Bluetooth_Headset	No	Bluetooth Headset available On board	Bluetooth Headset available On board	Bluetooth Headset available On board
Environment_Media	No	Bluetooth Headset available On board	Bluetooth Headset available On board	Bluetooth Headset available On board
Carina	Alloy wheels	Catalytic Converter	Catalytic Converter	Catalytic Converter
Safety_Security	ABS Central door lock	Daytime running light D.	Daytime running light D.	Daytime running light D.
Phantom_Driver	Age	2.000	2.000	2.000
Insurance_new	1	1	1	1
Insulation_new	1	1	1	1
Paint_Type	Metallic	Metallic	Metallic	Metallic
Upkeepkeep_Note	Club	Club	Club	Club
Wt_kg	1.800	2.000	2.000	2.000
Nr_of_Seats	4.000	4.000	4.000	4.000
Gearing_Type	Automatic	Automatic	Automatic	Automatic
Dimensions_m	4.530.000	4.530.000	4.530.000	4.530.000
Weight_kg	1220.000	1220.000	1220.000	1220.000
Drive_chasis	Front	Front	Front	Front
Color_hexcode	3.000	5.000	5.000	5.000
CO2_Emissions	99.000	129.000	129.000	129.000

	0	1	2	3
price	52776.000	14010.000	14010.000	14010.000
km	50513.000	35500.000	35500.000	35500.000
Gears	7.000	7.000	7.000	7.000
age	10.000	10.000	10.000	10.000
Previous_Owner	1.000	1.000	1.000	1.000
hp_MHP	60.000	141.000	85.000	86.000
Inspector_new	1.000	0.000	0.000	1.000
Dimensions_m	4.530.000	4.530.000	4.530.000	4.530.000
Weight_kg	1220.000	1220.000	1220.000	1220.000
Carina	No	No	No	No
Bluetooth_Headset	No	No	No	No
Environment_Media	No	No	No	No
Carina	No	No	No	No
ABS_Central_door_lock	No	No	No	No
Daytime_running_light	No	No	No	No
Upkeepkeep_Note	Club	Club	Club	Club
Wt_kg	1.800	2.000	2.000	2.000
Nr_of_Seats	4.000	4.000	4.000	4.000
Gearing_Type	Automatic	Automatic	Automatic	Automatic
Dimensions_m	4.530.000	4.530.000	4.530.000	4.530.000
Weight_kg	1220.000	1220.000	1220.000	1220.000
Drive_chasis	Front	Front	Front	Front
Color_hexcode	3.000	5.000	5.000	5.000
CO2_Emissions	99.000	129.000	129.000	129.000

Use this space to take notes:

Slide 19

► Need to Study ➤

- str.method
- contains()
- extract()
- get_dummies()
- add_prefix()
- sample()
- to_numeric()
- isin()
- apply()
- replace()
- split()
- join()
- regex
- def
- lambda

Use this space to take notes:

19

Slide 20

► Assumptions



- ▶ Assume the year you are currently in is 2022

20

Use this space to take notes:

Slide 21

► Hints



- ▶ Domain Knowledge is one of the most important things to evaluate your data.
- ▶ You have to evaluate each column by target label.

21

Use this space to take notes:

Slide 22

► Hints

- ▶ Check the **column names**.
(You can change the column names to something more useful.)
 - ▶ Check the percentage of **null values** for each column.
(You can drop columns having more than %... null value.)
 - ▶ Check the **value_counts** of each column, evaluate them and **take notes** about what you'll do.
(drop, similarity between columns, how to clean, define the pattern etc.)

A small, stylized icon consisting of two chevrons pointing to the right, colored in a gradient of blue and red.

2

Use this space to take notes:

Slide 23

► Hints

- ▶ How to exclude each value of columns from list.

3

Use this space to take notes:

Slide 24

Hints

- You can create functions to clean values

```
df["Fuel"].value_counts(dropna=False)
Diesel (Particulate Filter)      4315
Super 95                         4100
Gasoline                          3175
Diesel                            2984
Regular                           503
Super E10 95                      402
Super 95 (Particulate Filter)    268

benzine = ["Gasoline", "Super 95", "Regular", "Super E10 95", "Super Plus 98", "Super Plus E10 98", "Others"]
lpg = ["LPG", "Liquid petroleum gas", "CNG", "Biogas", "Domestic gas H"]
def fueltype(x):
    if x in benzine:
        return "Benzine"
    elif x in lpg:
        return "LPG/CNG"
    else:
        return x
df["Fuel"] = df.Fuel.apply(fueltype)
```



Use this space to take notes:

Slide 25

Hints

- Relatively hard-to-handle columns

- Consumption

To create separate columns, define the patterns for each consumption type.
Then evaluate which one is enough to ML Model.

```
Nah:
[[3.9 1/100 km (comb)], [4.1 1/100 km (city)], [3.7 1/100 km (country)]]           1986
[[4.2 1/100 km (comb)], [5 1/100 km (city)], [3.7 1/100 km (country)]]             304
[[4.4 1/100 km (comb)], [6.8 1/100 km (city)], [4.5 1/100 km (country)]]            276
[[3.8 1/100 km (comb)], [4.3 1/100 km (city)], [3.5 1/100 km (country)]]            257
[[3.6 1/100 km (comb)], [], [4.4 1/100 km (country)]]                                ...
[[\n, 4.8 1/100 km (comb), \n, 5.6 1/100 km (city), \n, 4.3 1/100 km (country), \n]       1
[[7.6 1/100 km (comb)], [], []]                                                       1
[[5.6 1/100 km (comb)], [7.6 1/100 km (city)], [4.4 1/100 km (country)]]            1
[\n, 4.7 1/100 km (comb), \n, \n, \n]                                                 1
Name: Consumption, Length: 882, dtype: int64
```



Use this space to take notes:

Slide 26

► Hints

- ▶ Relatively hard-to-handle columns

- Comfort_Convenience
- Entertainment_Media
- Extras
- Safety_Security

How can missing values in
these columns be filled?

NaN	1374
[Bluetooth, Hands-free equipment, On-board computer, Radio, USB]	1282
[Bluetooth, Hands-free equipment, MP3, On-board computer, Radio, USB]	982
[Bluetooth, CD player, Hands-free equipment, MP3, On-board computer, Radio, USB]	783
[On-board computer, Radio]	487

```
df["Entertainment_Media"] = [", ".join(item) if type(item) == list else item for item in df["Entertainment_Media"]]
```

28

Use this space to take notes:

Slide 27

► Hints

- ▶ How to examine columns to fill missing values

```
df.groupby("age").km.describe()
```

```
df.groupby(["make_model", "age"]).km.describe()
```

```
df.groupby(["make_model", "body_type", "age"]).price.describe()
```

27

Use this space to take notes:

Slide 28

► Hints

- ▶ How to fill missing values by groups

Example-1

```
#Step-1  
#df["body_type"].fillna(df["body_type"].mode()[0])  
  
#Step-2  
#df.loc[df["make_model"]=="Audi A1", "body_type"].fillna(df[df["make_model"]=="Audi A1"]["body_type"].mode()[0])  
  
#Step-3  
for group in list(df["make_model"].unique()):  
    cond = df["make_model"]==group  
    mode = list(df[cond]["body_type"].mode())  
    if len(mode) > 1:  
        df.loc[cond, "body_type"] = df.loc[cond, "body_type"].fillna(df[cond]["body_type"].mode()[0])  
    else:  
        df.loc[cond, "body_type"] = df.loc[cond, "body_type"].fillna(df["body_type"].mode()[0])
```

You can generalize this loop to create your own function in your analysis

28

Use this space to take notes:

Slide 29

► Hints

- ▶ How to fill missing values by groups

Example-2

```
#Step-1  
#df["Previous_Owners"].fillna(method="ffill")  
  
#Step-2  
#df.loc[df["age"]==0, "Previous_Owners"].fillna(method="ffill")  
  
#Step-3  
for group in list(df["age"].unique()):  
    cond = df["age"]==group  
    df.loc[cond, "Previous_Owners"] = df.loc[cond, "Previous_Owners"].fillna(method="ffill").fillna(method="bfill")  
    df["Previous_Owners"] = df["Previous_Owners"].fillna(method="ffill").fillna(method="bfill")
```

You can generalize this loop to create your own function

29

Use this space to take notes:

Slide 30

► Hints

- ▶ How to fill missing values by groups

Example-3

```
# Step-1
# df[“Paint_Type”].fillna(method=“ffill”)

# Step-2
# df.loc[df[“make_model”]==“Audi A1”, “Paint_Type”].fillna(method=“ffill”)

# Step-3
# for group in list(df[“make_model”].unique()):
#     cond = df[“make_model”]==group
#     df.loc[cond, “Paint_Type”] = df.loc[cond, “Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)
#     df[“Paint_Type”] = df[“Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)

# Step-4
for group1 in df[“make_model”].unique():
    for group2 in list(df[“body_type”].unique()):
        cond1 = df[“make_model”]==group1
        cond2 = df[“body_type”]==group2
        df.loc[cond1, “Paint_Type”] = df.loc[cond1, “Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)
        df.loc[cond2, “Paint_Type”] = df.loc[cond2, “Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)

for group1 in list(df[“make_model”].unique()):
    cond1 = df[“make_model”]==group1
    df.loc[cond1, “Paint_Type”] = df.loc[cond1, “Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)
df[“Paint_Type”] = df[“Paint_Type”].fillna(method=“ffill”).fillna(method=“bfill”)
```

30

Use this space to take notes:

Slide 31

► Hints

- ▶ Dummy Operation

(The get_dummies function has 2 different uses)

pd.get_dummies(df)

need to be
research

df[“col_name”].str.get_dummies(sep = “ ”)

31

Use this space to take notes:

Slide 32

► Hints



- ▶ pd.factorize()
- ▶ count()
- ▶ map()
- ▶ cat.codes
- ▶ LabelEncoder()
- ▶ OneHotEncoder()

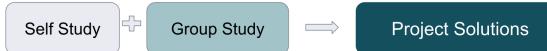
Use this space to take notes:

32

Slide 33

► Capstone Project Period

Capstone Project Period Duration
11 March - 16 March 2023



Project Solution Sessions

16 -17 March 2023

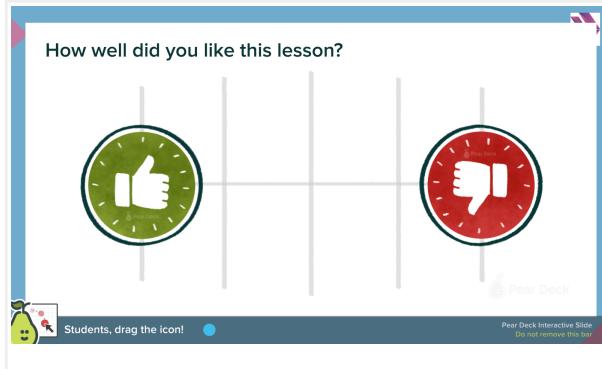
33

Use this space to take notes:

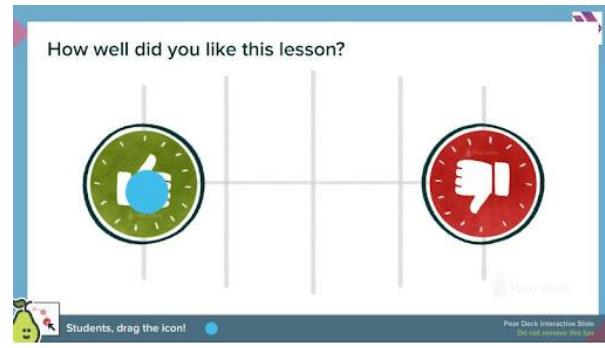
Slide 34

Your Response

Slide 34



Your Response



Use this space to take notes:

Slide 35

THANKS!

Any questions?

You can find us at:

- ▶ #questions-answers@Slack



Use this space to take notes:

CLARUSWAY®
www.clarusway.com

35