

# Data Engineering For Car Price Predictor Application

การทำวิศวกรรมข้อมูลเพื่อสร้างเว็บแอปพลิเคชันที่คำนวณ  
ราคารถมือสอง

จัดทำโดย  
นาย ธโนดม โซติบารุ่งพงศ์  
อาจารย์ที่ปรึกษาโครงการ ผศ.ดร.นันทวุฒิ วงศ์อังกู

# Data Engineering For Car Price Predictor Application

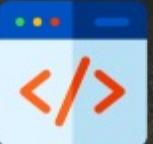


BACKGROUND  
STORY

Machine  
Learning



DATAPIPELINE



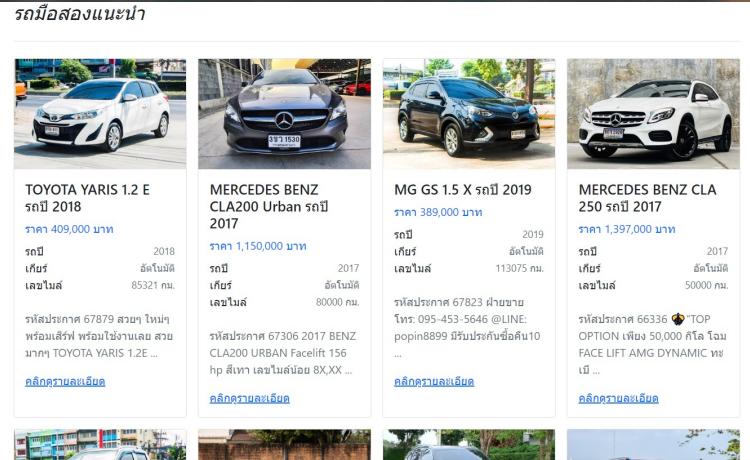
Application



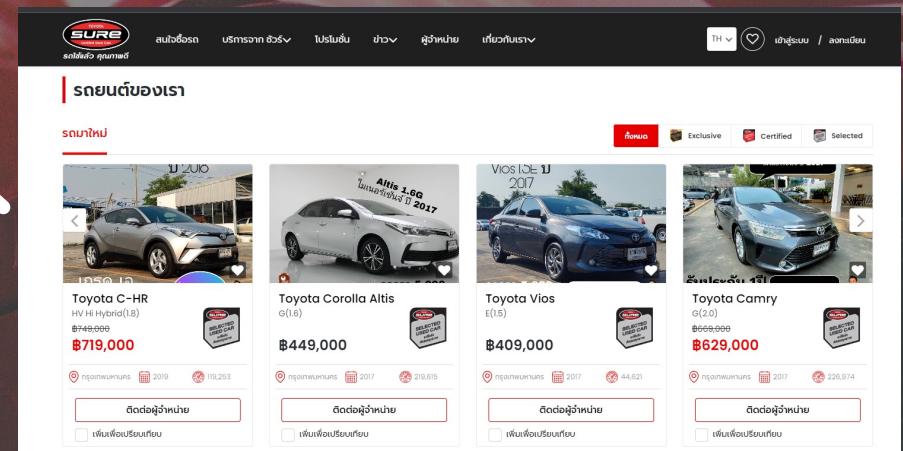
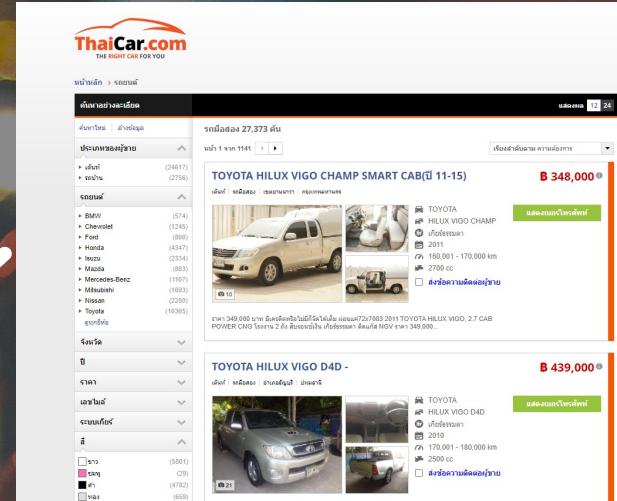
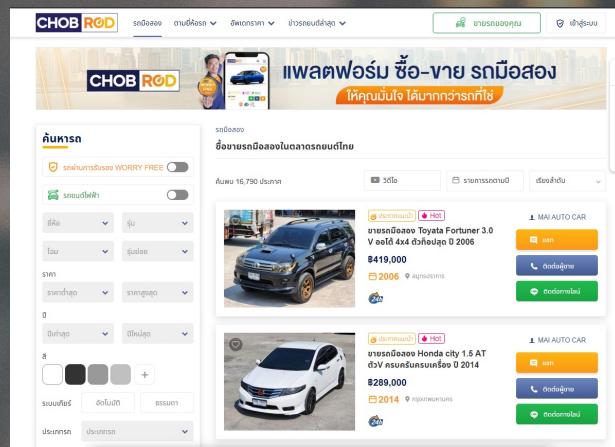
# BACKGROUND STORY



# BACKGROUND STORY

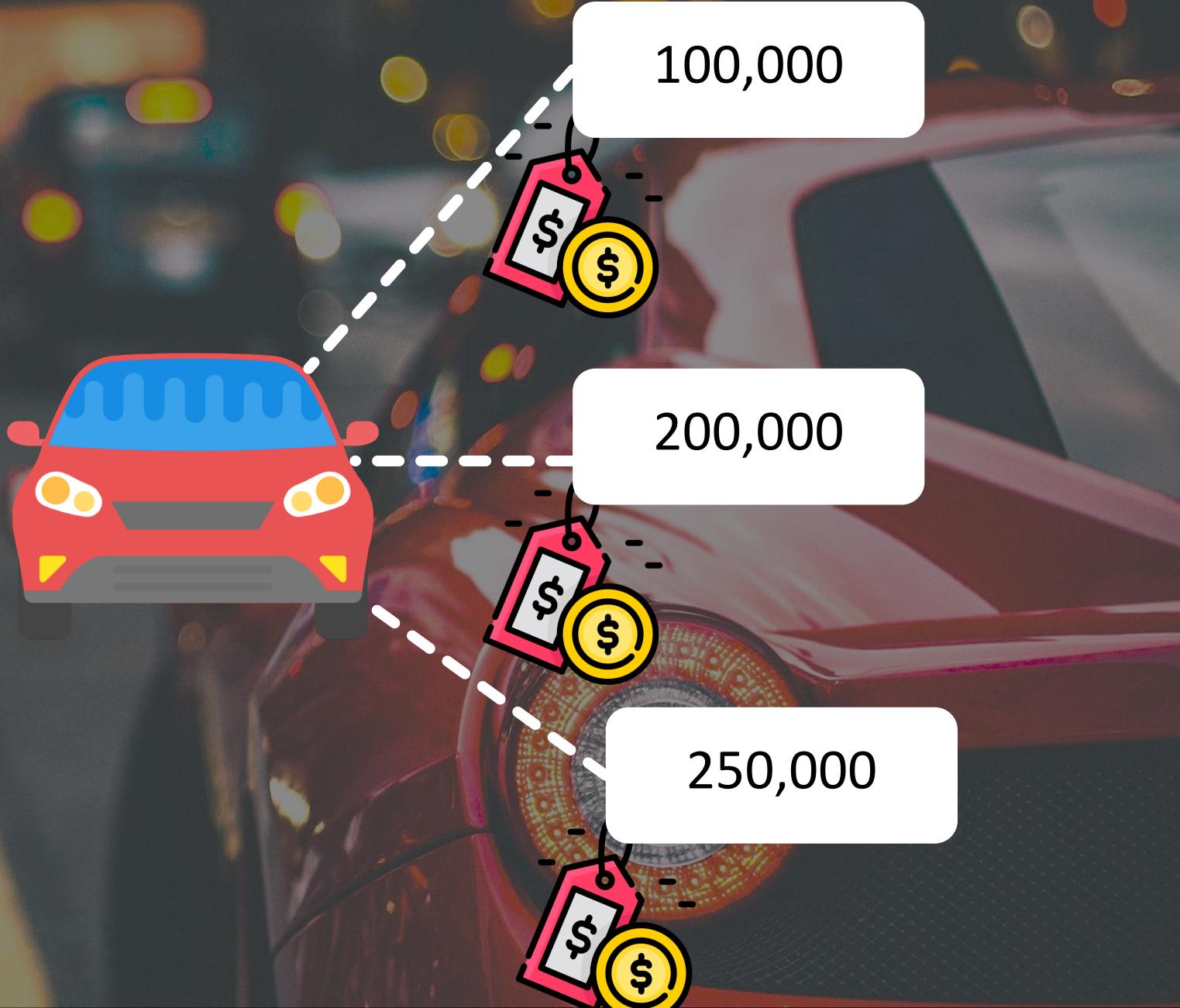


ขายเท่าไหร่ดีนะ ?  ซื้อเท่าไหร่ดีนะ ?





## BACKGROUND STORY

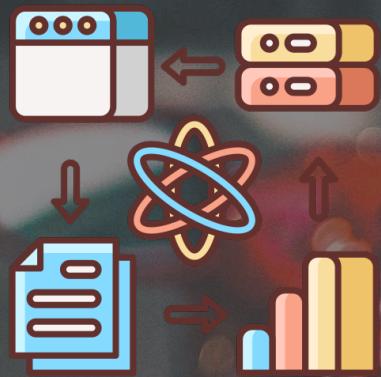




## BACKGROUND STORY



DATAPIPELINE



MACHINE LEARNING



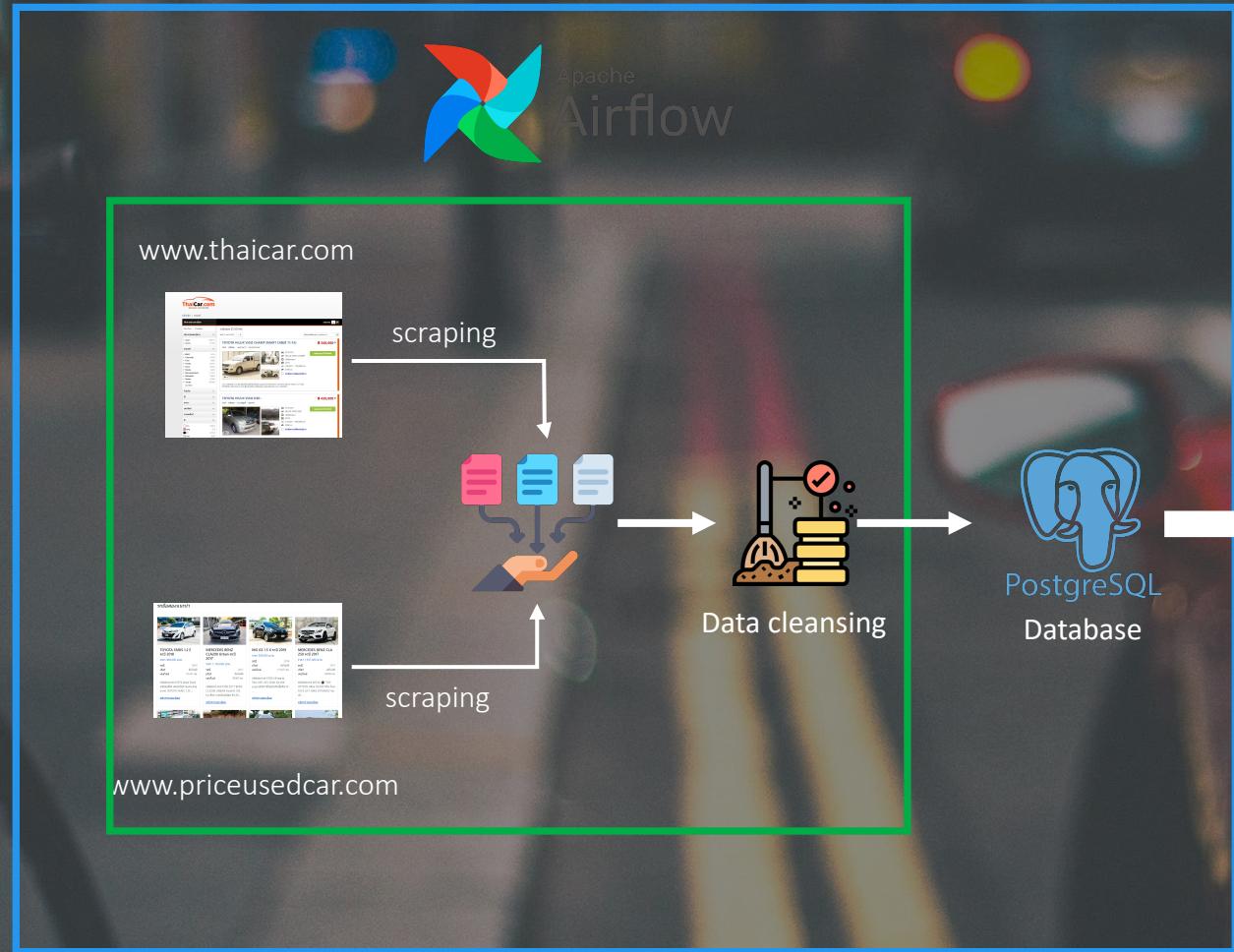
APPLICATION

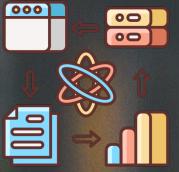


DATAPIPELINE



Docker Container

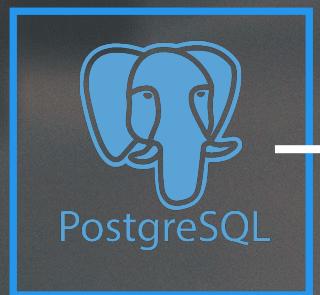




## MACHINE LEARNING



Docker Container



PostgreSQL



Train data

linear regression

random forest regression

xgboost regression



Best model from  
model evaluation



# APPLICATION

Program & API





**DATAPIPELINE**

# DATAPIELINE



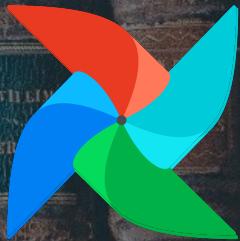
Apache  
Airflow

Apache Airflow คืออะไร?

Airflow คือ เครื่องมือหนึ่งในการสร้าง Data Pipeline โดยเป็น Open Source Platform ตัวหนึ่งที่ผู้ใช้สามารถเขียนโปรแกรมที่จะมาควบคุมการไหลของ Workflow และสามารถคอนเฟิร์มการไหลได้ ซึ่งถูกพัฒนาโดย Airbnb และเริ่มใช้งานมาตั้งแต่ปี 2015 ซึ่งนอกจากสร้าง Data Pipelines ได้แล้วมันยังสามารถนำไปสร้าง หรือพัฒนา ETL (Extract-Transform-Load), Machine Learning และ Predictive ได้อีกด้วย

โดย Airflow เข้ามาช่วยจัดการ Task ต่างๆ อีกทั้งยังรองรับแบบ Hybrid & Multi-Cloud โดยเกิดขึ้นมาจาก Data Pipeline ที่มีจำนวนมากซึ่งทำให้ควบคุมได้ยาก โดย Airflow มีการเขียน Workflow เป็น DAG (Directed Acyclic Graph) กราฟที่มีหัวลูกลศรทิศทางเดียว โดยไม่สามารถกลับมาที่จุดเดิมได้ ซึ่ง DAG ประกอบไปด้วยหลายๆ Task ที่เชื่อมต่อกันและในแต่ละ Task นั้น ก็มีความสามารถที่แตกต่างกัน

# DATAPIPELINE



Apache  
Airflow

Data collection

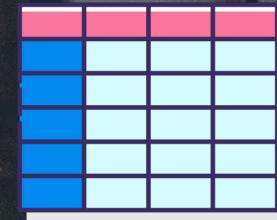


Data cleansing



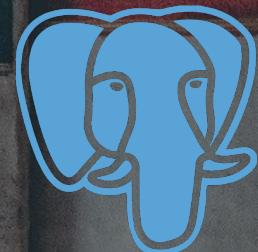
Extract

Cleaned Data



Transform

Database



Load

PostgreSQL

ในกระบวนการสร้าง Data pipeline หรือกระบวนการ ETL นี้ มีการใช้ Apache Airflow เข้ามาช่วยในการเขียน pipeline โดยข้อดีของการใช้ airflow ในโปรเจคนี้ คือ ใช้ python ใน การเขียน , มี interface เօ้าไว้เพื่อดู task การทำงานในแต่ละส่วนของ pipeline ที่เราเขียน เป็นต้น



Apache Airflow

DAGs Datasets Security Browse Admin Docs 21:28 UTC AA

**DAG: collect\_data\_project\_dags** collect data from site

running Schedule: @daily Next Run: 2023-02-20, 00:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

2023-02-20T21:26:31Z Runs 25 Run manual\_2023-02-20T21:26:30.116094+00:00 Layout Left > Right Update Find Task...

PythonOperator deferred failed queued removed restarting running scheduled shutdown skipped success up\_for\_reschedule up\_for\_retry upstream\_failed upstream\_failed no\_status

Auto-refresh C

```
graph LR; A[get_carprice_today] --> B[print1]; B --> C[print2]; C --> D[merge_csv]; D --> E[clean_data]; E --> F[to_db]; G[get_carprice_today1] --> B;
```



# Data collection

หน้าหลัก / รถมือสองทั้งหมด

## เช็คราคารถมือสองทั้งหมด

ราคา ปีรถ ลักษณะ

9491 รายการ หน้า 1 จาก 633 เรียงลำดับ

**NISSAN ALMERA 1.2 E รถปี 2016**  
ราคา 245,000 บาท  
รถปี: 2016  
เกียร์: อัตโนมัติ  
เลขไมล์: 124000 กม.  
รหัสประจำค 68531 Nissan Almera 1.2E ปี 2016 มือเดียวประวัติดีศูนย์ เดิมบัง  
น้อโลไม่ ...  
[คลิกดูรายละเอียด](#)

แนะนำรถ SUV มือสอง ราคาคุ้ม  
Review ลูกค้ารับรถมากกว่า 3,000 คัน หรือลูกค้าไม่มีรถมา茱ราเพิบกิจการไปรับ  
ลูก้าสิ่งที่  
[สมชายค้าร์ชีนเดอร์ อรัญ](#)

เงินใช้ที่ใช้คุกเก้ เพื่อเพิ่มประสิทธิภาพและประสบการณ์ที่ดีในการใช้งานเว็บไซต์ คุณสามารถอ่านรายละเอียดเพิ่มเติมได้ที่ [นโยบายคุคกี้](#) [ยอมรับ](#)

หน้าหลัก > รถมือสอง

**ThaiCar.com**  
THE RIGHT CAR FOR YOU

รถมือสอง 27,373 คัน  
หน้า 1 จาก 1141

ดันพารอตมือสอง

ค้นหาใน | ล้างค่าลูบ

ประเภทของผู้ขาย

- เด่นๆ (24617)
- รถบ้าน (2756)

รอบปี

- BMW (574)
- Chevrolet (1245)
- Ford (800)
- Honda (4347)
- Isuzu (2334)
- Mazda (883)
- Mercedes-Benz (1107)
- Mitsubishi (1693)
- Nissan (2280)
- Toyota (10365)

จังหวัด

ปี

ราคา

เชื่อมโยง

**TOYOTA HILUX VIGO CHAMP SMART CAB(ปี 11-15)**  
เด่นๆ | รถมือสอง เช็คสภาพ กรุณาตรวจสอบ  
ราคา 349,000 บาท มีเครื่องหัวใจใหม่ที่ดีไกด์เพิ่ม ส่องแฉค์72x7083 2011 TOYOTA HILUX VIGO. 2.7 CAB POWER CNG...  
  
**TOYOTA HILUX VIGO D4D -**  
เด่นๆ | รถมือสอง เชื่อมโยงที่ดี | ปี 2011  
ราคา 439,000 บาท



# Data collection

HTML

The screenshot shows a search result for cars. It includes a header with filters like 'รถเก๋ง', 'รถตู้', etc. Below are two main entries:

- NISSAN ALMERA 1.2 E 2016**  
วันที่ 24 กันยายน 2016  
ราคา 244,000 บาท  
ดูรายละเอียด
- TOYOTA HILUX VIGO D4D 2015**  
วันที่ 24 กันยายน 2016  
ราคา 349,000 บาท  
ดูรายละเอียด

ชื่อรถ, ราคา, ปี, เลข. ไมล์

HTML

The screenshot shows a search result for cars. It includes a header with filters like 'รถเก๋ง', 'รถตู้', etc. Below are two main entries:

- NISSAN ALMERA 1.2 E 2016**  
วันที่ 24 กันยายน 2016  
ราคา 244,000 บาท  
ดูรายละเอียด
- TOYOTA HILUX VIGO D4D 2015**  
วันที่ 24 กันยายน 2016  
ราคา 349,000 บาท  
ดูรายละเอียด

ชื่อรถ, ราคา, ปี, เลข. ไมล์



Data Frame



CSV



DAGS



# Data cleansing

- ลบข้อมูลที่ราการถเป็น 0 และข้อมูลที่มี null

123 index	RBC name	123 price	RBC year	RBC brand	123 km_drive
9,232	MERCEDES-BENZ C-CLASS C200 Kom W2C	0	2009	MERCEDES-BEI	[NULL]
10,139	HONDA Accord โฉมปี 13-15	0	2014	HONDA	[NULL]
9,374	TOYOTA WISH -	0	2008	TOYOTA	[NULL]
9,376	HONDA Accord โฉมปี 08-11	0	2008	HONDA	[NULL]
9,351	MAZDA 3 โฉมปี 05-10	0	2007	MAZDA	[NULL]
9,377	HONDA Civic โฉมปี 05-12	0	2009	HONDA	[NULL]
9,357	MAZDA BT-50 PRO FREE STYLE CAB (12-15)	0	2012	MAZDA	[NULL]
9,358	KIA JUMBO PICUP โฉมปี 09-13	0	2008	KIA	[NULL]
9,359	TOYOTA ALPHARD โฉมปี 08-14	0	2008	TOYOTA	[NULL]
9,378	HONDA Jazz โฉมปี 08-13	0	2011	HONDA	[NULL]
9,344	TOYOTA YARIS โฉมปี 06-12	0	2008	TOYOTA	[NULL]
9,368	HONDA Accord โฉมปี 08-11	0	2011	HONDA	[NULL]
9,360	MITSUBISHI SPACE WAGON โฉมปี 04-10	0	2010	MITSUBISHI	[NULL]

- ลบข้อมูลที่เลขไม่ถูกเป็น 0

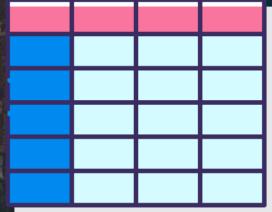
	RBC name	123 price	RBC year	RBC brand	123 km_drive
1,980	MITSUBISHI TRITON 2.4 CNG	248,000	2012	MITSUBISHI	0
1,979	ISUZU MU-7 3.0 Primo	348,000	2008	ISUZU	0
5,893	TOYOTA CAMRY 2.4 V	288,000	2007	TOYOTA	0
3,648	TOYOTA HILUX REVO 2.4 J Plus	438,000	2018	TOYOTA	0
4,765	TOYOTA AVANZA 1.5 S	278,000	2013	TOYOTA	0
4,767	HONDA ACCORD 2.4 EL	378,000	2011	HONDA	0
3,637	FORD ESCAPE 3.0 XLT 4WD	148,000	2003	FORD	0
505	TOYOTA HILUX REVO 2.4 E	570,000	2010	TOYOTA	0

- ลบข้อมูลที่ปีไม่ใช่ปีและปรับให้เป็น int

123 index	RBC name	123 price	RBC year	RBC brand	123 km_drive	
1	2,918	MITSUBISHI PAJERO SPORT 2.4	1,269,000	[NULL]	MITSUBISHI	68
2	2,970	TOYOTA HILUX REVO 2.4	639,000	[NULL]	TOYOTA	8
3	2,971	TOYOTA HILUX REVO 2.4	639,000	[NULL]	TOYOTA	7
4	2,702	TOYOTA HILUX REVO 2.4	639,000	[NULL]	TOYOTA	50
5	2,969	TOYOTA	599,000	REVO	TOYOTA	8
6	2,959	[NULL]	1,589,000	FORTUNER	TOYOTA	0
7	2,953	[NULL]	879,000	ALTIS	TOYOTA	8
8	2,952	[NULL]	869,000	ALTIS	TOYOTA	8

- เพิ่ม column (car\_old) เพื่อเปลี่ยนข้อมูลปีรถ ให้เป็น  
ความเก่าของรถ เพื่อให้เป็นข้อมูลเชิงปริมาณ

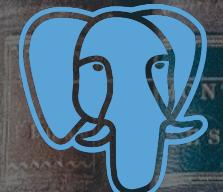
123 year	RBC brand	123 km_drive	123 car_old
10	1,980	MITSUBISHI	100,000
10	1,981	TOYOTA	250,000
10	1,981	TOYOTA	253,000
10	1,981	TOYOTA	110,000
10	1,989	NISSAN	360,790
10	1,989	MERCEDES-BEI	95,001
10	1,990	HONDA	130,000
10	1,991	HONDA	180,000
10	1,991	HONDA	180,000
10	1,991	HONDA	180,000



# Data

	A	B	C	D	E	F	G
1		name	price	year	brand	km_drive	car_old
2	0	TOYOTA FO	639000	2012	TOYOTA	233133	11
3	1	MITSUBISHI	459000	2012	MITSUBISHI	138388	11
4	2	HONDA ACC	368000	2010	HONDA	201901	13
5	3	FORD FIEST.	159000	2010	FORD	134716	13
6	4	HONDA ACC	559000	2013	HONDA	157479	10
7	5	NISSAN X-TI	509000	2015	NISSAN	150156	8
8	6	ISUZU D-MAX	669000	2018	ISUZU	78658	5
9	7	TOYOTA VIC	299000	2013	TOYOTA	93767	10
10	8	ISUZU D-MAX	499000	2014	ISUZU	91229	9
11	9	MITSUBISHI	480323	2017	MITSUBISHI	72746	6
12	10	HONDA CITY	439000	2018	HONDA	38301	5
13	11	CHEVROLET	299000	2012	CHEVROLET	619741	11
14	12	TOYOTA VIC	399000	2018	TOYOTA	54598	5
15	13	TOYOTA HIL	489000	2012	TOYOTA	212817	11
16	14	NISSAN NP300	369000	2017	NISSAN	61118	6
17	15	TOYOTA CAI	189000	2003	TOYOTA	260324	20
18	16	TOYOTA FO	659000	2011	TOYOTA	215527	12
19	17	MITSUBISHI	889000	2017	MITSUBISHI	116836	6
20	18	ISUZU D-MAX	390000	2013	ISUZU	180479	10
21	19	TOYOTA HIL	438000	2014	TOYOTA	93000	9
22	20	TOYOTA HIL	458000	2014	TOYOTA	177000	9
23	21	HONDA CIVI	938000	2019	HONDA	48000	4
24	22	MERCEDES I	598000	2012	MERCEDES	183000	11
25	23	MITSUBISHI	288000	2012	MITSUBISHI	180000	11
26	24	TOYOTA FO	738000	2013	TOYOTA	82000	10
27	25	TOYOTA HIL	598000	2018	TOYOTA	107000	5
28	26	TOYOTA WIT	248000	2006	TOYOTA	196000	17
29	27	ISUZU D-MAX	388000	2018	ISUZU	143000	5
30	28	CHEVROLET	358000	2017	CHEVROLET	87000	6

```
▶ [4] car.info()
✓ 0.0s
...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8085 entries, 0 to 8084
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   name        8085 non-null   object 
 1   price       8085 non-null   int64  
 2   year        8085 non-null   int64  
 3   brand       8085 non-null   object 
 4   km_drive    8085 non-null   float64
 5   car_old     8085 non-null   int64  
dtypes: float64(1), int64(3), object(2)
memory usage: 379.1+ KB
```



PostgreSQL

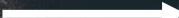
# Load data to database



Docker Container



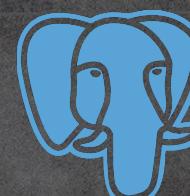
DAGS



Data Frame



Connect,  
Load to database



PostgreSQL



# Load data to database

DBeaver 22.3.3 - car\_price

Database Navigator X Projects

postgres - localhost:5432

Databases

- airflow
- postgres

project1

- Schemas
  - public
- Tables
  - car\_price (1.1M)
  - Views
  - Materialized Views
  - Indexes
  - Functions
  - Sequences
  - Data types
  - Aggregate functions

Project - General X

Name: DataSource

Bookmarks

Diagrams

Scripts

car\_price

Properties Data ER Diagram

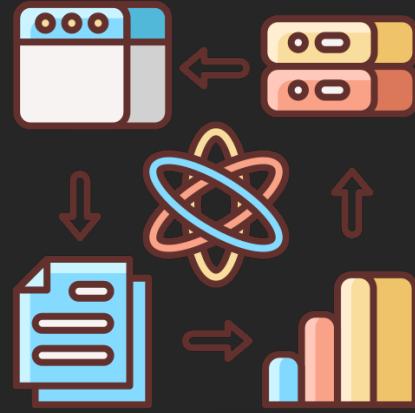
Enter a SQL expression to filter results (use Ctrl+Space)

index	name	price	year	brand	km_drive	car_old
0	TOYOTA FORTUNER 3.0 V	639,000	2,012	TOYOTA	233,133	11
1	MITSUBISHI PAJERO SPOR	459,000	2,012	MITSUBISHI	138,388	11
2	HONDA ACCORD 2.0 EL	368,000	2,010	HONDA	201,901	13
3	FORD FIESTA 1.6 Sport	159,000	2,010	FORD	134,716	13
4	HONDA ACCORD 2.0 EL NA	559,000	2,013	HONDA	157,479	10
5	NISSAN X-TRAIL 2.0 V 4WC	509,000	2,015	NISSAN	150,156	8
6	ISUZU D-MAX 1.9 Hi-Lande	669,000	2,018	ISUZU	78,658	5
7	TOYOTA VIOS 1.5 TRD	299,000	2,013	TOYOTA	93,767	10
8	ISUZU D-MAX 3.0 Hi-Lande	499,000	2,014	ISUZU	91,229	9
9	MITSUBISHI TRITON 2.4 GL	480,323	2,017	MITSUBISHI	72,746	6
10	HONDA CITY 1.5 S i-VTEC	439,000	2,018	HONDA	38,301	5
11	CHEVROLET CAPTIVA 2.0 L	299,000	2,012	CHEVROLET	619,741	11
12	TOYOTA VIOS 1.5 G	399,000	2,018	TOYOTA	54,598	5
13	TOYOTA HILUX VIGO 3.0 G	489,000	2,012	TOYOTA	212,817	11
14	NISSAN NP 300 NAVARA 2.	369,000	2,017	NISSAN	61,118	6
15	TOYOTA CAMRY 2.4 Q	189,000	2,003	TOYOTA	260,324	20
16	TOYOTA FORTUNER 3.0 TR	659,000	2,011	TOYOTA	215,527	12
17	MITSUBISHI PAJERO SPOR	889,000	2,017	MITSUBISHI	116,836	6
18	ISUZU D-MAX 2.5 L	390,000	2,013	ISUZU	180,479	10
19	TOYOTA HILUX VIGO 2.5 D	438,000	2,014	TOYOTA	93,000	9
20	TOYOTA HILUX VIGO 3.0 G	458,000	2,014	TOYOTA	177,000	9
21	HONDA CIVIC 1.5 Turbo RS	938,000	2,019	HONDA	48,000	4
22	MERCEDES BENZ E200	598,000	2,012	MERCEDES	183,000	11
23	MITSUBISHI LANCER EX 1.8	288,000	2,012	MITSUBISHI	180,000	11
24	TOYOTA FORTUNER 3.0 V	738,000	2,013	TOYOTA	82,000	10

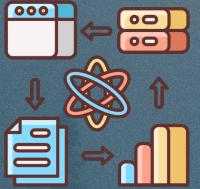
Refresh Save Cancel Export data 200 200+

200 row(s) fetched - 5ms (2ms fetch), on 2023-02-21 at 05:02:30

ICT en TH



# Machine Learning

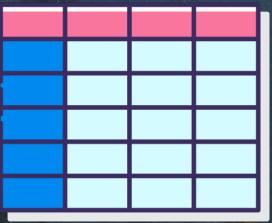


# Machine Learning

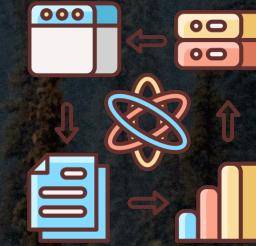


PostgreSQL

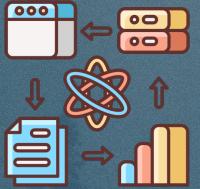
→  
Pull data



→  
Training



→  
Best model



# Machine Learning



PostgreSQL



Year,car\_old,price

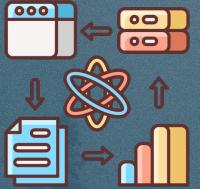
Brand,Name

OneHot Encoder

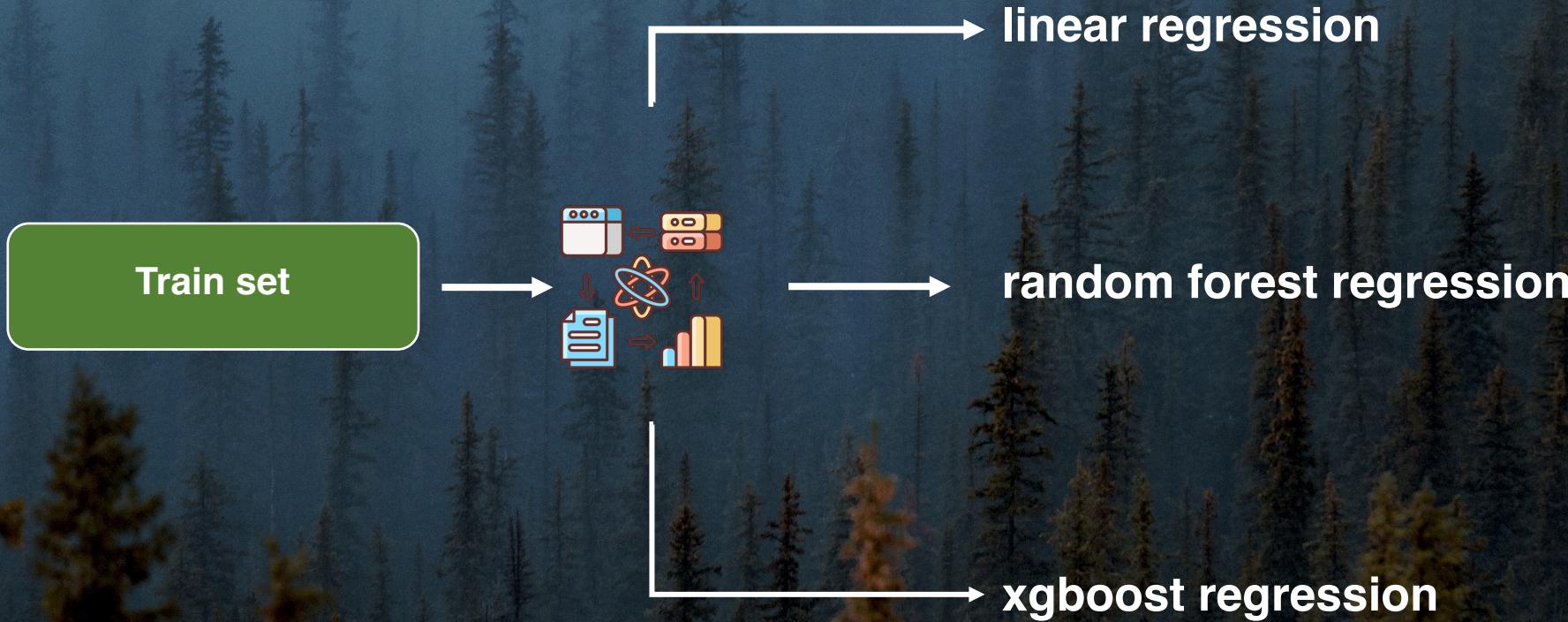


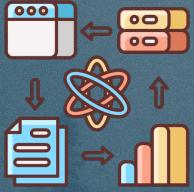
DataFrame





# Machine Learning

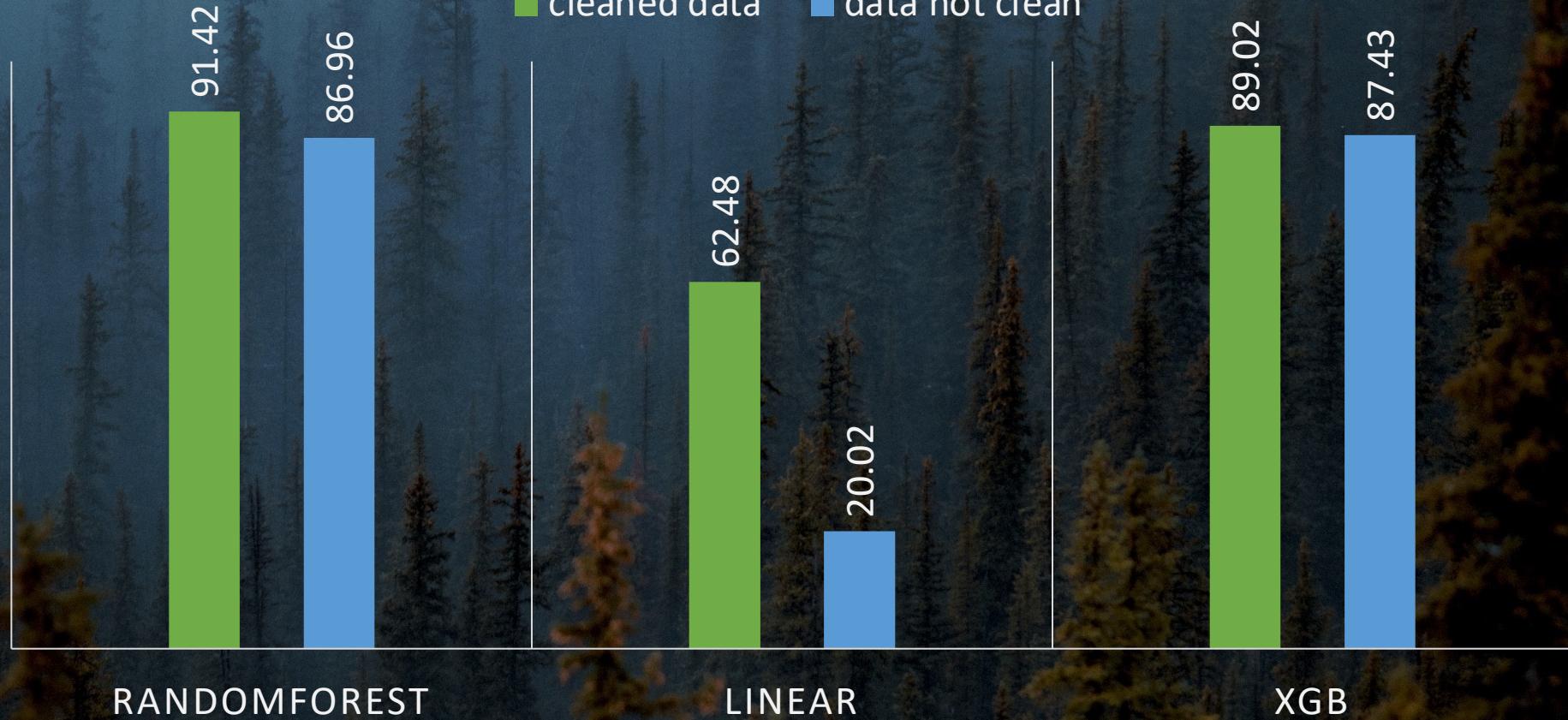


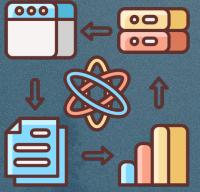


# Model score by r2score

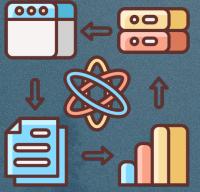
## MODEL WITH R2SCORE

cleaned data    data not clean

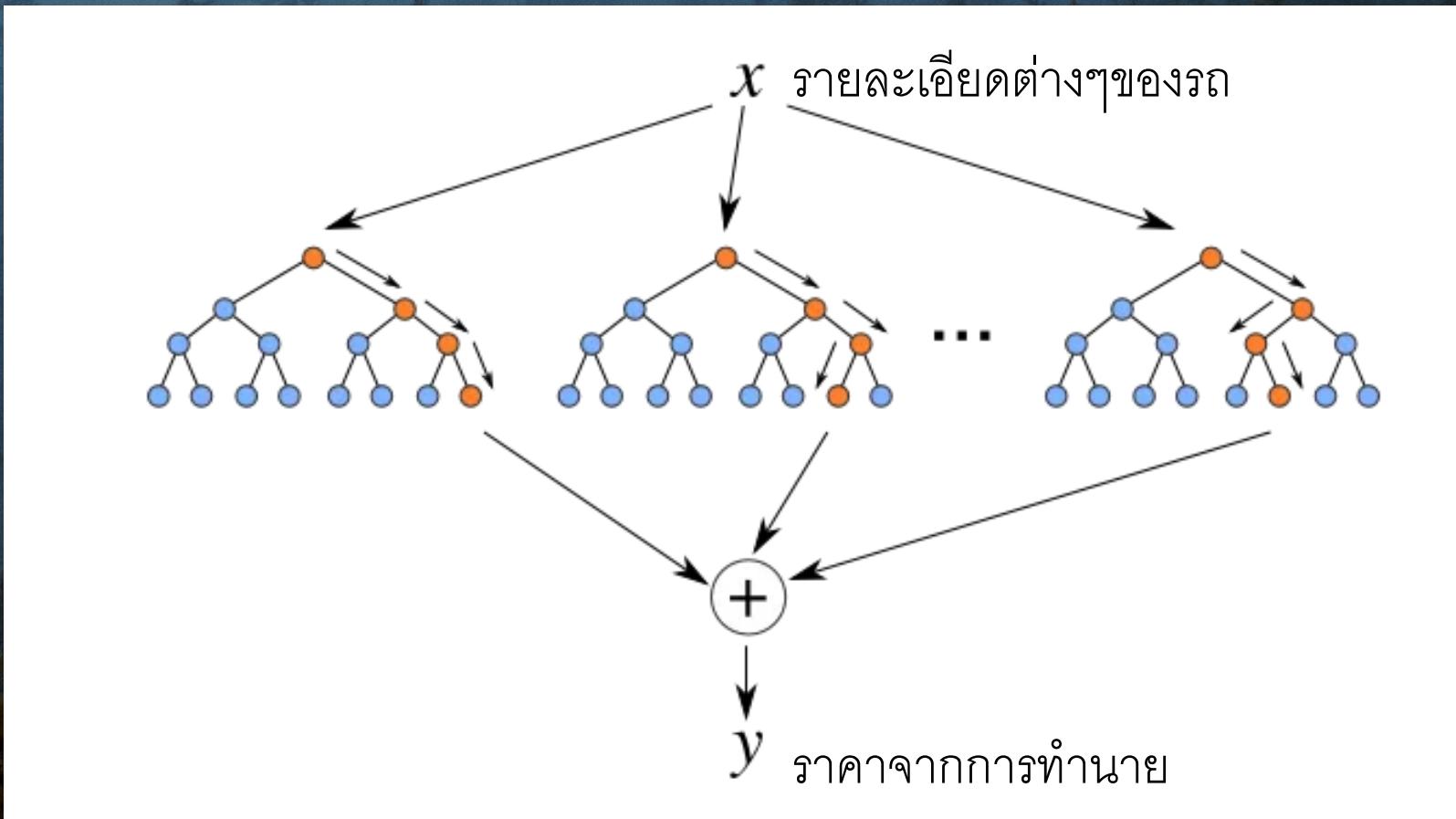




# Random Forest Regression Model



# Random Forest Regression





# Application



# Application



Prediction Model



Flask  
web development,  
one drop at a time



# Application

Used Car Price Predictor

Select Brand of Your Car:

NISSAN

Select Name:

NISSAN ALMERA 1.2 E

Select Year of Purchase:

2016

Kilometers travelled:

124000

Predict Price!

Prediction: ₩281290.0



# Used Car Price Predictor

Select Brand of Your Car:

NISSAN

Select Name:

NISSAN ALMERA 1.2 E

Select Year of Purchase:

2016

Kilometers travelled:

124000

Predict Price!

Prediction: ₧281290.0



## NISSAN ALMERA 1.2 E รถปี 2016

ราคา 245,000 บาท

รถปี: 2016

เกียร์: อัตโนมัติ

เลขไมล์: 124000 กม.

รหัสประกาศ 68531 Nissan Almera 1.2E ปี 2016 มือเดียว  
บางน้อดใหม่ ...

[คลิกดูรายละเอียด](#)

NISSAN ALMERA โฉมปี 14-16

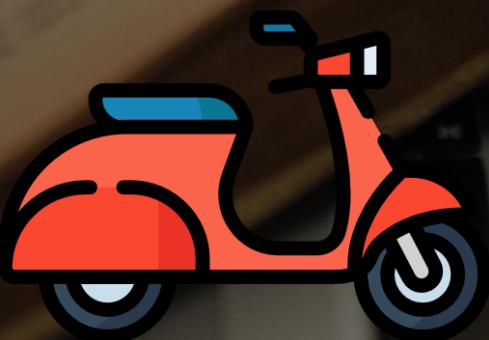
เดือน | รถมือสอง | เช็คเนวี่ | กรุงเทพมหานคร

฿ 349,000

แสดงเบอร์โทรศัพท์

NISSAN  
ALMERA  
เกียร์อัตโนมัติ  
2014  
1200 cc  
 ส่งข้อความติดต่อผู้ขาย

แอร์, วิทยุ, CD, ABS, AIRBAG, พ.พาวเวอร์, กระจกไฟฟ้า, กระจกมองข้างปรับไฟฟ้า, ช. ล็อก, กุญแจรีโมท



# ปัญหาที่พบ

- การเก็บข้อมูลจากการ **scraping** เป็นการเก็บข้อมูลจากการอ่านข้อมูลที่แสดงผลบนหน้าเว็บไซต์แล้วแยกข้อมูลต่างๆ เมื่อมีข้อมูลที่มีลักษณะผิดปกติต้องทำการแก้ไขกระบวนการทำความสะอาดข้อมูลใน **data pipeline** ทำให้ต้องใช้เวลาค่อนค้างนานในขั้นตอน **data cleansing**
- หากมีการอปเดทแก้ไขรูปแบบการแสดงผลเว็บไซต์ต้นทางของข้อมูลต้องออกแบบกระบวนการดึงข้อมูล (**extract**) ใหม่
- เนื่องมาจากรถในการนำมายเป็นรถมือสอง ต้องมีการประเมินสภาพรถจริง จึงจะตัดสินราคาที่ว่างชายได้ แต่ในเว็บไซต์ต้นทางของข้อมูลไม่ได้แสดงรายละเอียดที่ลึกและครบถ้วนในรถแต่ละคัน ราคานี้ได้จากการทำนายจึงยังไม่เหมาะสมในการนำไปอ้างอิงในการขายจริง ตัวแอปเจิงเหมาะสมในการดูราคาร่วง หรือดูแนวโน้มของราคารถรุ่นที่เราต้องการจะขายหรือซื้อ

A close-up photograph of an open book and a laptop keyboard. The book is open to a page with dense text, and the laptop keyboard is visible in the foreground, slightly blurred. The lighting is warm and focused on the book's pages.

# THANK YOU