รายงานหัวข้อโปรเจค Car Sale Advertisements

จัดทำโดย

นาย กิตติภณ สุรุ่งเรืองสกุล 61070278

นำเสนอ

รศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนนึงของวิชา 06026117 FUNDAMENTALS OF DATA SCIENCE

ปีการศึกษา 2562 ภาคเรียนที่ 2

Business understanding

ธุรกิจเกี่ยวกับการขายรถมือสองในประเทศยูเครน ซึ่งมีรถหลากหลายยี่ห้อที่นำมาขาย รถที่นำมาขายก็มีทั้งรถที่จดทะเบียนในประเทศยูเครนและไม่ได้จดทะเบียนและมีขายตั้งแต่รถที่ผลิตในปี 1953 ไปจนถึงรถรุ่นใหม่ ๆ ซึ่งในโครงงานนี้จะศึกษาว่า รถยี่ห้อใดที่นิยมนำมาขายมือสองเพื่อนำมาใช้ศึกษาหาเหตุผลว่าทำไมรถเหล่านี้ถึงถูกนำมาขายต่อเช่น นำมาขายมือสองเพราะเป็นรถยี่ห้อดังถึงจะผ่านการใช้มาแล้วก็ยังขายได้ราคาดี นำมาขายมือสองเพราะเป็นรถที่หายากหรือเลิกผลิตแล้วเมื่อนำมาขายสามารถทำเงินได้สูงเช่น Volkswagen van เป็นต้น และศึกษาความสัมพันธ์ของตัวแปร mileage กับ price โดยศึกษาว่าเมื่อ mileage มีค่าน้อยจะส่งผลให้ price สูงขึ้น

Data understanding

ข้อมูลได้มาจาก Kaggle (https://www.kaggle.com/antfarol/car-sale-advertisements) ซึ่งเป็นข้อมูลทุติยภูมิมีคนรวบรวมข้อมูลไว้แล้ว ข้อมูลที่ได้มาเป็นไฟล์ประเภท Comma-separated values (csv)

	car	price	body	mileage	engV	engType	registration	year	model	drive
count	9576	9576.000000	9576	9576.000000	9142.000000	9576	9576	9576.000000	9576	9065
unique	87	NaN	6	NaN	NaN	4	2	NaN	888	3
top	Volkswagen	NaN	sedan	NaN	NaN	Petrol	yes	NaN	E-Class	front
freq	936	NaN	3646	NaN	NaN	4379	9015	NaN	199	5188
mean	NaN	15633.317316	NaN	138.862364	2.646344	NaN	NaN	2006.605994	NaN	NaN
std	NaN	24106.523436	NaN	98.629754	5.927699	NaN	NaN	7.067924	NaN	NaN
min	NaN	0.000000	NaN	0.000000	0.100000	NaN	NaN	1953.000000	NaN	NaN
25%	NaN	4999.000000	NaN	70.000000	1.600000	NaN	NaN	2004.000000	NaN	NaN
50%	NaN	9200.000000	NaN	128.000000	2.000000	NaN	NaN	2008.000000	NaN	NaN
75%	NaN	16700.000000	NaN	194.000000	2.500000	NaN	NaN	2012.000000	NaN	NaN
max	NaN	547800.000000	NaN	999.000000	99.990000	NaN	NaN	2016.000000	NaN	NaN

ชุดข้อมูลมีทั้งหมด 9576 แถวและมี 10 ตัวแปรได้แก่

car: ยี่ห้อรถ

- เป็นประเภทตัวอักษร

price: ราคาขาย (USD)

- เป็นประเภทตัวเลขทศนิยม

body: ประเภทตัวถังรถ

- เป็นประเภทตัวอักษร

mileage: ระยะทางที่วิ่งไปแล้ว (‘000 กม.)

- เป็นประเภทตัวเลขทศนิยม

engV: ปริมาตรรอบเครื่องยนต์ (‘000 ลบ.ม.)

- เป็นประเภทตัวเลขทศนิยม

engType: ประเภทของเชื้อเพลิง (“Other” ในกรณีนี้คือค่าว่าง)

- เป็นประเภทตัวอักษร

registration: เป็นรถที่จดทะเบียนในยูเครนหรือไม่

- เป็นประเภทบูลีน
- Yes: จดทะเบียนในยูเครน No: ไม่จดทะเบียนในยูเครน

year: ปีที่ผลิต

- เป็นประเภทตัวเลข

model: ชื่อรุ่นเฉพาะ

- เป็นประเภทตัวอักษร

drive: ประเภทไดร์ฟ

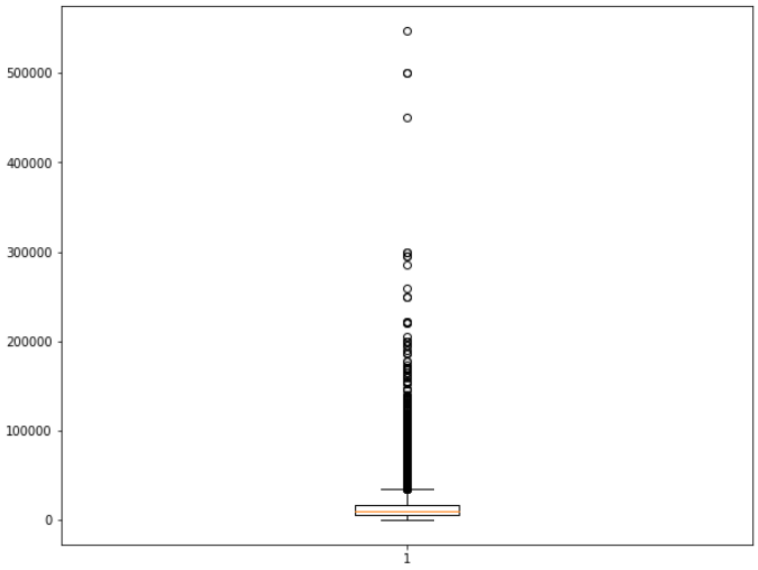
- เป็นประเภทตัวอักษร

ทำการตรวจสอบว่าตัวแปรใดที่มีค่าว่างอยู่บ้าง โดยถ้าตัวแปรไหนมีค่า True คือตัวแปรนั้นมีค่าว่าง ดังนั้นตัวแปรที่มีค่าว่างคือ engV drive และตัวแปร engType ค่า “Other” คือค่าว่าง

```
False    9576
Name: car, dtype: int64

False    9576
Name: price, dtype: int64

False    9576
Name: body, dtype: int64

False    9576
Name: mileage, dtype: int64

False    9142
True      434
Name: engV, dtype: int64
```

```
False    9576
Name: engType, dtype: int64

False    9576
Name: registration, dtype: int64

False    9576
Name: year, dtype: int64

False    9576
Name: model, dtype: int64

False    9065
True      511
Name: drive, dtype: int64
```

ทำการพล็อต box plot เพื่อตรวจสอบค่า outlier ของตัวแปร price



Data preparation

1. แก้ไขค่าว่างของตัวแปร drive

```
n = 0
for i in car['drive'].isnull():
  if i == True:
    check = car['model'].loc[n]
    new = np.where(car['model'] == check)
    new = car.loc[new[0]]
    new_drive = new['drive'].mode()

    if type(new_drive.any()) == str:
      car.loc[car.model == check, 'drive'] = car.loc[car.model == check, 'drive'].fillna(new_drive.loc[0])
    elif type(new_drive.any()) == bool:
      car.loc[car.model == check, 'drive'] = car.loc[car.model == check, 'drive'].fillna(car['drive'].mode().loc[0])
  n += 1
```

ผู้จัดทำคิดว่าตัวแปร model และ drive มีความเกี่ยวข้องกัน เช่น model “E-Class” ค่าส่วนใหญ่ในตัวแปร drive จะเท่ากับ “rear”

```
x = car[car.model == 'E-Class']
x['drive'].value_counts()
```

```
rear     165
full       9
front      5
Name: drive, dtype: int64
```

จึงจะทำการดูที่ตัวแปร model เช่น ใน model “E-Class” ส่วนใหญ่จะมีค่าในตัวแปร drive เท่ากับ “rear” จึงทำการแทนค่าว่างในตัวแปร drive เมื่อ model เท่ากับ “E-Class” ด้วยค่า mode เพราะว่าเป็นค่าที่มีความถี่มากที่สุดหรือซ้ำกันมากที่สุด ซึ่งแปลว่าค่าว่างมีโอกาสที่จะเป็นค่า mode มากที่สุด

```
x = car[car.model == 'CL 550']
print(x['drive'].mode(), '\n')
print(car['drive'].value_counts(), '\n')
print(car['drive'].mode())
```

```
Series([], dtype: object)

front    5105
full     1647
rear     1256
Name: drive, dtype: int64

0    front
dtype: object
```

แต่ถ้า model นั้นเมื่อหาค่า mode ของตัวแปร drive แล้วไม่มี จะทำการแทนที่ค่าว่างด้วยค่า mode ของ drive ทั้งหมด เช่น model “CL 550” ไม่มีค่า mode ในตัวแปร drive จึงทำการแทนค่าว่างโดยค่า mode จาก drive ทั้งหมดคือ “front”

ซึ่งวิธีนี้ใช้หลักการเดียวกันกับ Bayes' theorem โดยจะแทนที่ค่าว่างด้วยค่าที่มีโอกาสจะเกิดขึ้นมากที่สุด และ ค่าในส่วนที่ไม่ได้เป็นค่าว่างเปรียบเสมือนค่าที่ไว้สร้างโมเดลและค่าว่างเปรียบเสมือนเหมือนข้อมูลใหม่ที่เข้ามาและต้องการทำนาย

2. แก้ไขค่าว่างของตัวแปร engV

```
engV_median = car['engV'].median()
car = car.fillna(value={'engV' : engV_median})
```

ทำการแทนที่ค่าว่างของตัวแปร engV ด้วยค่า median เพื่อป้องกันการรบกวนถ้าหากมีค่า Outlier

3. แก้ไขค่าว่างของตัวแปร engType

```
car = car[car.engType != 'Other']
```

ทำการลบแถวที่มีค่า engType เท่ากับ Other เพราะเป็นค่าว่าง

```
print(len(car), '\n')
print(len(car[car.engType == 'Other']), '\n')
print(len(car[car.engType != 'Other']))
```

```
9576

462

9114
```

ใช้วิธีการลบแถวเพราะว่าเมื่อทำการลบออกไปแล้วข้อมูลที่เหลือยังมีจำนวนมากและจำนวนแถวที่ลบออกไปเป็นเพียง 462/9576 = 4.82% จากทั้งหมด

4. ทำการกำจัดค่า outlier ของ price และค่าผิดปกติเช่น ราคาเท่ากับ 0

```
car = car[car.price != 0];
```

ทำการลบแถวที่มีค่า price เท่ากับ 0 เพราะเป็นค่าผิดปกติและไม่มีจริง

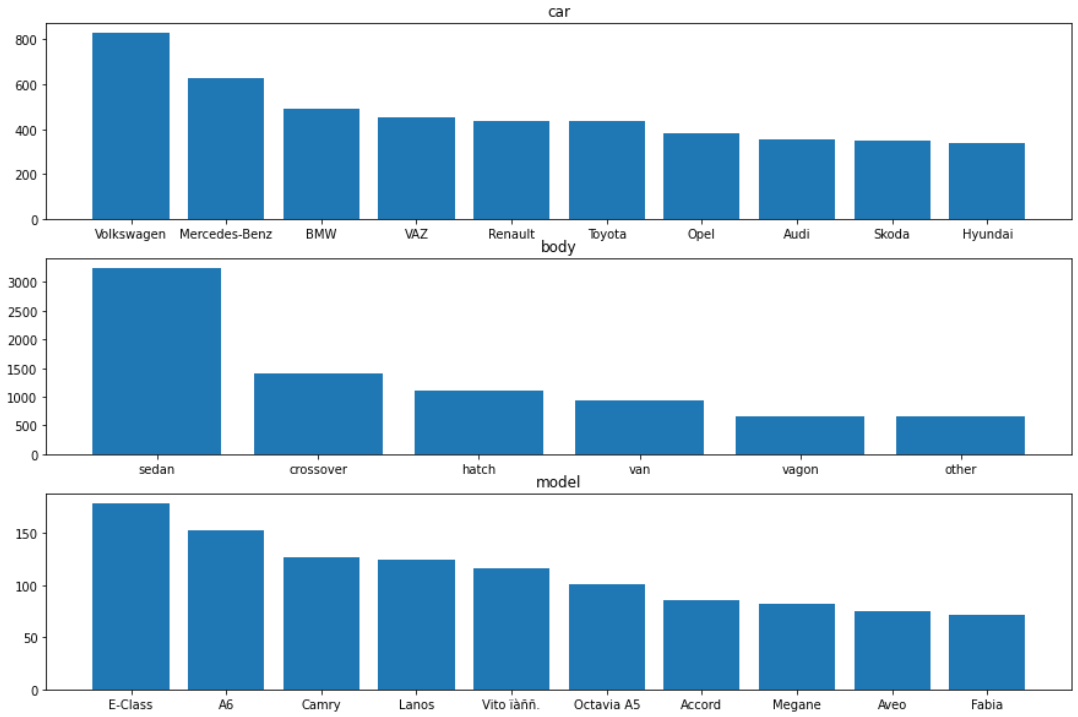
```
q1 = np.quantile(car['price'], 0.25) #quantile 1 ของ price
q3 = np.quantile(car['price'], 0.75) #quantile 3 ของ price
iqr = q3 - q1 #IQR ของ price
lower = q1 - (1.5 * iqr) #lower 1.5*IQR whisker
higher = q3 + (1.5 * iqr) #higher 1.5*IQR whisker
# print(lower, '-', higher)
car = car[car.price >= lower]#ลบข้อมูลที่มีค่าน้อยกว่า lower
car = car[car.price <= higher]#ลบข้อมูลที่มีค่ามากกว่า higher
```

ใช้ np.quantile() ในการหา quantile1และ3 เพื่อนำไปหาค่า IQR จากนั้นนำค่า IQR ไปหาขอบทั้งสองข้างจึงได้เป็น lower และ higher คือค่าที่น้อยที่สุดและค่าที่มากที่สุดที่เป็นไปได้ตามลำดับ แล้วจึงนำค่าทั้งสองไปตั้งเงื่อนไขโดยจะลบแถวที่มีค่า price น้อยกว่า lower และลบแถวที่มีค่า price มากกว่า higher

Modeling

1. ทำการพล็อตกราฟ Bar chart

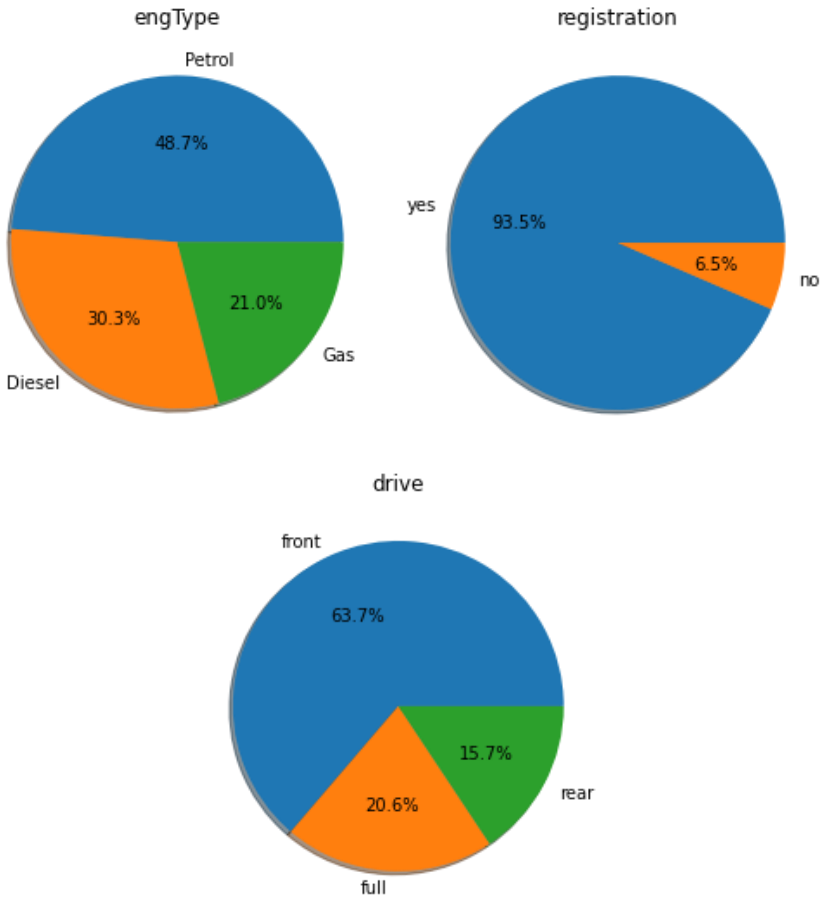
```
bar_plot = ['car', 'body', 'model']
fig = plt.rcParams["figure.figsize"] = (15,10)

n = 1
for i in bar_plot:
  plt.subplot(len(bar_plot),1,n)
  x = car[i].value_counts().head(10).keys();
  y = car[i].value_counts().head(10)
  plt.bar(x, y)
  # print(x, '\n', y, '\n')
  plt.title(i)
  n += 1
```



ทำการพล็อตกราฟเพื่อดูว่า ยี่ห้อ ประเภทตัวถังรถและชื่อรุ่นเฉพาะใดที่มีการนำมาขายมือสองมากที่สุด โดยจะแสดงเฉพาะ 10 อันดับแรกแต่ถ้ามีไม่ถึง 10 อันดับก็จะแสดงเท่าที่มี ที่เลือกใช้ Bar chart เพราะว่าข้อมูลมีหลายตัวแปรและเป็นข้อมูลประเภทตัวอักษร โดยการแสดงค่าจะแสดงจากมากไปน้อยและแกน Y คือจำนวนที่มีการนำมาขายมือสอง ซึ่งถ้าลูกค้าท่านใดต้องการหารถยี่ห้อนี้ ประเภทตัวถังรถและชื่อรุ่นเฉพาะเหล่านี้ก็สามารถมาหาได้ที่นี่เพราะมีการนำมาขายมือสองจำนวนมาก

2. ทำการพล็อตกราฟ Pie chart

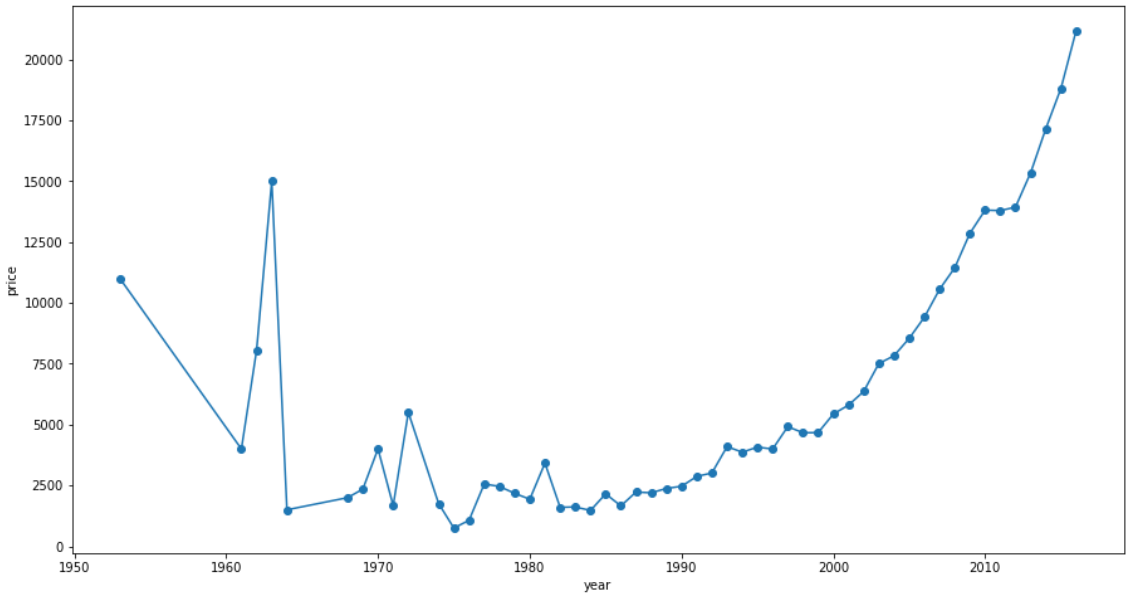
```
pie_plot = ['engType', 'registration', 'drive']
fig = plt.rcParams["figure.figsize"] = (15,15)

n = 1
for i in pie_plot:
  plt.subplot(len(pie_plot),1,n)
  # print(car[i].value_counts())
  plt.pie(car[i].value_counts(), labels = car[i].value_counts().keys(), autopct='%1.1f%%', shadow=True)
  plt.title(i)
  n += 1
```



ทำการพล็อตกราฟเพื่อดูว่า ประเภทของเชื้อเพลิง เป็นรถที่จดทะเบียนในยูเครนหรือไม่และประเภทไดรฟ์ใดที่มีการนำมาขายมือสองมากที่สุด ที่เลือกใช้ Pie chart เพราะว่าเมื่อนำมาพล็อตด้วยกราฟนี้จะทำให้มองเห็นได้ชัดเจนว่าสัดส่วนข้อมูลเป็นอย่างไร ข้อมูลมีจำนวนตัวแปรน้อยและเป็นข้อมูลประเภทตัวอักษร โดยค่าที่แสดงคือจำนวนที่มีการนำมาขายมือสอง

3. ทำการพล็อตกราฟ Line chart

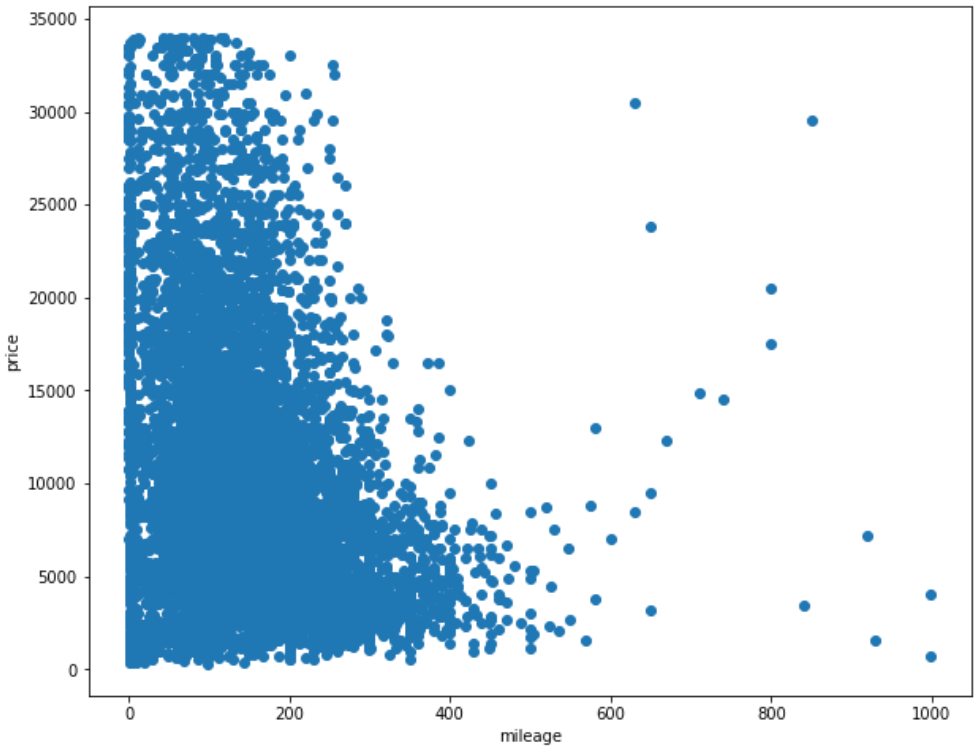
```
fig = plt.rcParams["figure.figsize"] = (15,8)
x = np.sort(car['year'].drop_duplicates()) # เอาปีที่ซ้ำกันออกและเรียงลำดับจากน้อยไปมาก
y = car[['year', 'price']].groupby(['year']).mean() # ทำการ group by ด้วย year และหาค่า mean ของ price ในแต่ละปี
plt.plot(x, y, marker='o') # ทำการมาร์คจุด o ในแต่ละปี
plt.xlabel('year')
plt.ylabel('price')
# จะเห็นได้ว่าราคาเฉลี่ยมีแนวโน้มสูงขึ้นเรื่อย ๆ ตั้งแต่ประมาณปี 2000
```



ทำการพล็อตด้วย Line chart เพราะต้องการศึกษาแนวโน้มของราคารถที่เปลี่ยนแปลงในแต่ละปีที่ถูกผลิต ซึ่งจะเห็นได้ว่าแนวโน้มของราคาขายเฉลี่ยสูงขึ้นเรื่อย ๆ ตั้งแต่รถที่ผลิตปี 2000 เป็นต้นมา สามารถตีความได้ว่ารถที่ผลิตออกมาในปีหลัง ๆ เมื่อนำมาขายมือสองจะมีราคาที่สูงกว่ารถในปีเก่า ๆ แต่ถ้ารถที่ผลิตในปีเก่า ๆ นั้นหายากหรือเป็นที่ต้องการก็จะมีราคาสูงเช่นกันเช่น รถที่ผลิตประมาณปี 1963 มีราคาที่ค่อนข้างสูงเมื่อเทียบกับรถที่ผลิตในปีหลัง ๆ สรุปได้ว่ารถรุ่นใหม่ที่ออกมาเมื่อถูกนำมาขายเป็นรถมือสองจะได้ราคาที่สูงกว่ารถรุ่นที่ผลิตในปีเก่า

4. Linear Regression

```
fig = plt.rcParams["figure.figsize"] = (10,8)
plt.scatter(car['mileage'], car['price'])
plt.ylabel('price')
plt.xlabel('mileage');
```



ทำการพล็อตด้วย Scatter plot เพื่อดูความสัมพันธ์ของตัวแปร mileage และ price เพื่อศึกษาว่าถ้า mileage มีค่าน้อยจะส่งผลให้ price มีค่าสูงขึ้นหรือไม่ แต่เมื่อลองทำการพล็อตกราฟจะเห็นว่าทั้งสองตัวแปรไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น จึงไม่สามารถทำ Linear Regression ได้

Evaluation

การเตรียมข้อมูลถือว่ามีข้อผิดพลาดเล็กน้อยตรงส่วนของตัวแปร engType ที่ทำการแก้ไขค่าว่างด้วยวิธีการลบออกซึ่งการลบออกอาจทำให้ข้อมูลที่สำคัญหายไป อาจจะต้องใช้วิธีอื่นในการแก้ไขค่าว่างตรงส่วนนั้นแทน ในส่วนของขั้นตอน Modeling ที่ทำการพล็อตกราฟเพื่อดูข้อมูลถือว่าเลือกใช้กราฟได้เหมาะสม แต่ควรจะใช้กราฟที่หลากหลายมากกว่านี้เช่น Histogram เพื่อดูความถี่ของราคารถว่ารถส่วนใหญ่ที่นำมาขายอยู่ในช่วงราคาใดเพื่อที่เวลาคนขายนำรถมาขายจะได้ตั้งราคาได้อย่างเหมาะสมและในส่วนของการทำ Linear regression ได้ทำการทดสอบกับหลาย ๆ ตัวแปรแล้วแต่ก็ไม่มีความสัมพันธ์กัน ซึ่งอาจจะมีปัจจัยอื่น ๆ ที่ผู้จัดทำไม่ทราบจึงควรศึกษาเพิ่มเติมเพื่อปรับปรุงในโครงงานต่อไป

สรุปได้ว่ายี่ห้อรถ ประเภทตัวถังรถ ชื่อรุ่นเฉพาะ ประเภทของเชื้อเพลิง เป็นรถที่จดทะเบียนในยูเครนหรือไม่และประเภทไดรฟ์ที่มีการขายมือสองมากที่สุดคือ Volkswagen sedan E-Class Petrol จดทะเบียนในยูเครนและfront ตามลำดับ ซึ่งถ้าต้องการซื้อรถที่มีส่วนประกอบเหล่านี้ก็สามารถหาซื้อได้ที่นี่เพราะสามารถหาซื้อได้ง่ายและมีขายเป็นจำนวนมาก และรถรุ่นใหม่ที่ออกมาเมื่อถูกนำมาขายเป็นรถมือสองจะได้ราคาที่สูงกว่ารถรุ่นที่ผลิตในปีเก่า

อ้างอิง

https://www.kaggle.com/antfarol/car-sale-advertisements (แหล่งข้อมูล)

https://matplotlib.org/ (ศึกษาการพล็อตกราฟ)