DSI205

LEAST-SQUARES PROBLEM PROJECT

- PREPARE DATASET
- MODEL BUILDING
- RESULTS
- CONCLUSION

MEMBERS

- 6524651038 ปุณณวิช ศิลปเสริฐ
- 6524651061 จีรัฏฐ์ โนดไธสง
- 6524651210 ชัญญานุช เจริญพนารัตน์
- 6524651244 ณิชพน รักยาบัณฑิต

DATA DESCRIPTION

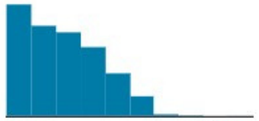
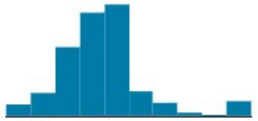
Description

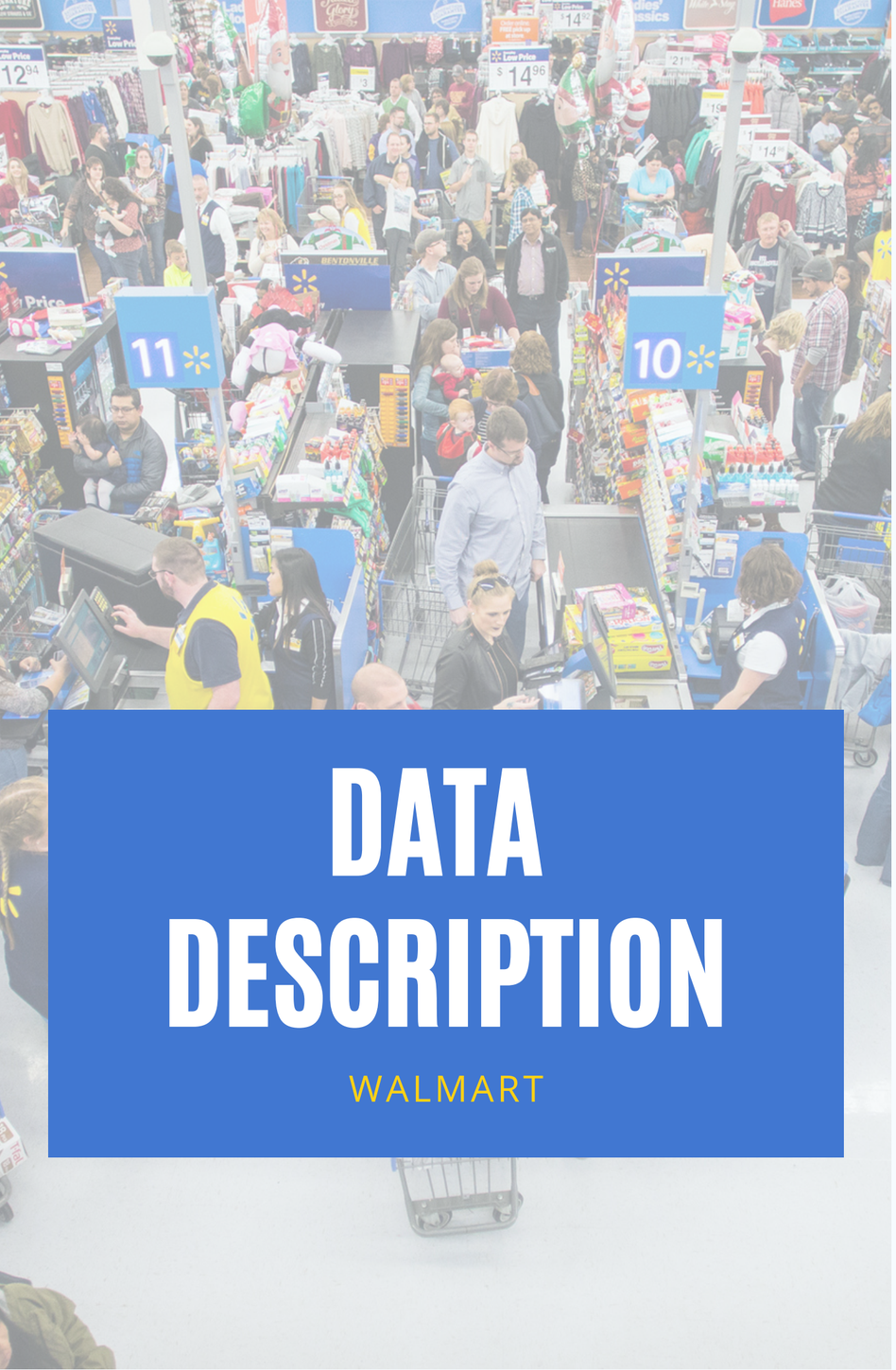
Walmart เป็นหนึ่งในร้านค้าปลีกชั้นนำในสหรัฐอเมริกา ซึ่งมีข้อมูลการขายสำหรับร้าน Walmart อยู่ 45 สาขา และมีข้อมูลที่ครอบคลุมยอดขายตั้งแต่วันที่ 5 กุมภาพันธ์ 2010 ถึงวันที่ 1 พฤศจิกายน 2012

Objective

เข้าใจชุดข้อมูล clean ข้อมูล และสร้างโมเดล Regression เพื่อทำนายยอดขายโดยมีตัวแปรอิสระหลายตัวแปร ประเมินผลโมเดลและเปรียบเทียบคะแนนที่เกี่ยวข้อง เช่น R2, RMSE, เป็นต้น

Walmart.csv

# Store	Date	# Weekly_Sales	# Holiday_Flag	# Temperature	# Fuel_Price	# CPI	# Unemployment
The store number	The week of sales	Sales for the given store	Whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week	Temperature on the day of sale	Cost of fuel in the region	Prevailing consumer price index	Prevailing unemployment rate
 1 45	143 unique values	 210k 3.82m	 0 1	 -2.06 100	 2.47 4.47	 126 227	 3.88 14.3
1	05-02-2010	1643690.9	0	42.31	2.572	211.0963582	8.106
1	12-02-2010	1641957.44	1	38.51	2.548	211.2421698	8.106
1	19-02-2010	1611968.17	0	39.93	2.514	211.2891429	8.106
1	26-02-2010	1409727.59	0	46.63	2.561	211.3196429	8.106
1	05-03-2010	1554806.68	0	46.5	2.625	211.3501429	8.106
1	12-03-2010	1439541.59	0	57.79	2.667	211.3806429	8.106
1	19-03-2010	1472515.79	0	54.58	2.72	211.215635	8.106
1	26-03-2010	1404429.92	0	51.45	2.732	211.0180424	8.106
1	02-04-2010	1594968.28	0	62.27	2.719	210.8204499	7.808
1	09-04-2010	1545418.53	0	65.86	2.77	210.6228574	7.808
1	16-04-2010	1466058.28	0	66.32	2.808	210.4887	7.808



DATA DESCRIPTION

WALMART

คอลัมน์ทั้งหมดของชุดข้อมูล

- **Store** - the store number
- **Date** - the week of sales
- **Weekly_Sales** - sales for the given store
- **Holiday_Flag** - whether the week is a special holiday week

  1 – Holiday week

  0 – Non-holiday week
- **Temperature** - Temperature on the day of sale
- **Fuel_Price** - Cost of fuel in the region
- **CPI** - Prevailing consumer price index
- **Unemployment** - Prevailing unemployment rate

OVERVIEWS

Prepare Dataset

Model Building

Results

Conclusion

PREPARE DATASET

kaggle

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106
5	1	12-03-2010	1439541.59	0	57.79	2.667	211.380643	8.106
6	1	19-03-2010	1472515.79	0	54.58	2.720	211.215635	8.106
7	1	26-03-2010	1404429.92	0	51.45	2.732	211.018042	8.106
8	1	02-04-2010	1594968.28	0	62.27	2.719	210.820450	7.808
9	1	09-04-2010	1545418.53	0	65.86	2.770	210.622857	7.808

CHECK NUMBER OF RECORD

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106
5	1	12-03-2010	1439541.59	0	57.79	2.667	211.380643	8.106
6	1	19-03-2010	1472515.79	0	54.58	2.720	211.215635	8.106
7	1	26-03-2010	1404429.92	0	51.45	2.732	211.018042	8.106
8	1	02-04-2010	1594968.28	0	62.27	2.719	210.820450	7.808
9	1	09-04-2010	1545418.53	0	65.86	2.770	210.622857	7.808

(6438 row x 8 columns)

DATA CLEANING

ดู meta data ของ df

```
[ ]  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Store         6435 non-null   int64  
 1   Date          6435 non-null   object 
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64  
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
```

เช็ค Missing value กับ Duplicate

Check duplicate

```
[ ]  df[df.duplicated()]
```

Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment

Check missing values

```
[ ]  df.isnull().mean()
```

```
Store           0.0
Date            0.0
Weekly_Sales    0.0
Holiday_Flag    0.0
Temperature     0.0
Fuel_Price      0.0
CPI             0.0
Unemployment    0.0
dtype: float64
```

FIXING OUTLIER

สร้างฟังก์ชั่นเพื่อคำนวนและนับจำนวน **outlier** ใน **column** โดยเมื่อตรวจสอบลึกไปในแต่ละคอลัมน์ **Unemployment , Holiday_Flag** และ ให้แสดงผลในทางสถิติ ยกเว้นคอลัมน์ **Weekly_Sales** และ **Temperature**

```
[ ]  count_outliers(df).sort_values('outlier_count', ascending=False)
```

	outlier_count	outlier_percent
Unemployment	481.0	7.474747
Holiday_Flag	450.0	6.993007
Weekly_Sales	34.0	0.528361
Temperature	3.0	0.046620

FEATURE ENGINEERING

```
[ ]  df['Employment'] = 100 - df['Unemployment']
     df.sample(3)
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Employment
2436	18	2010-12-03	1138800.32	0	42.39	2.805	131.784000	9.202	90.798
3361	24	2011-06-24	1304850.67	0	68.88	3.964	135.265267	8.212	91.788
233	2	2011-10-28	1769296.25	0	65.87	3.372	217.325182	7.441	92.559

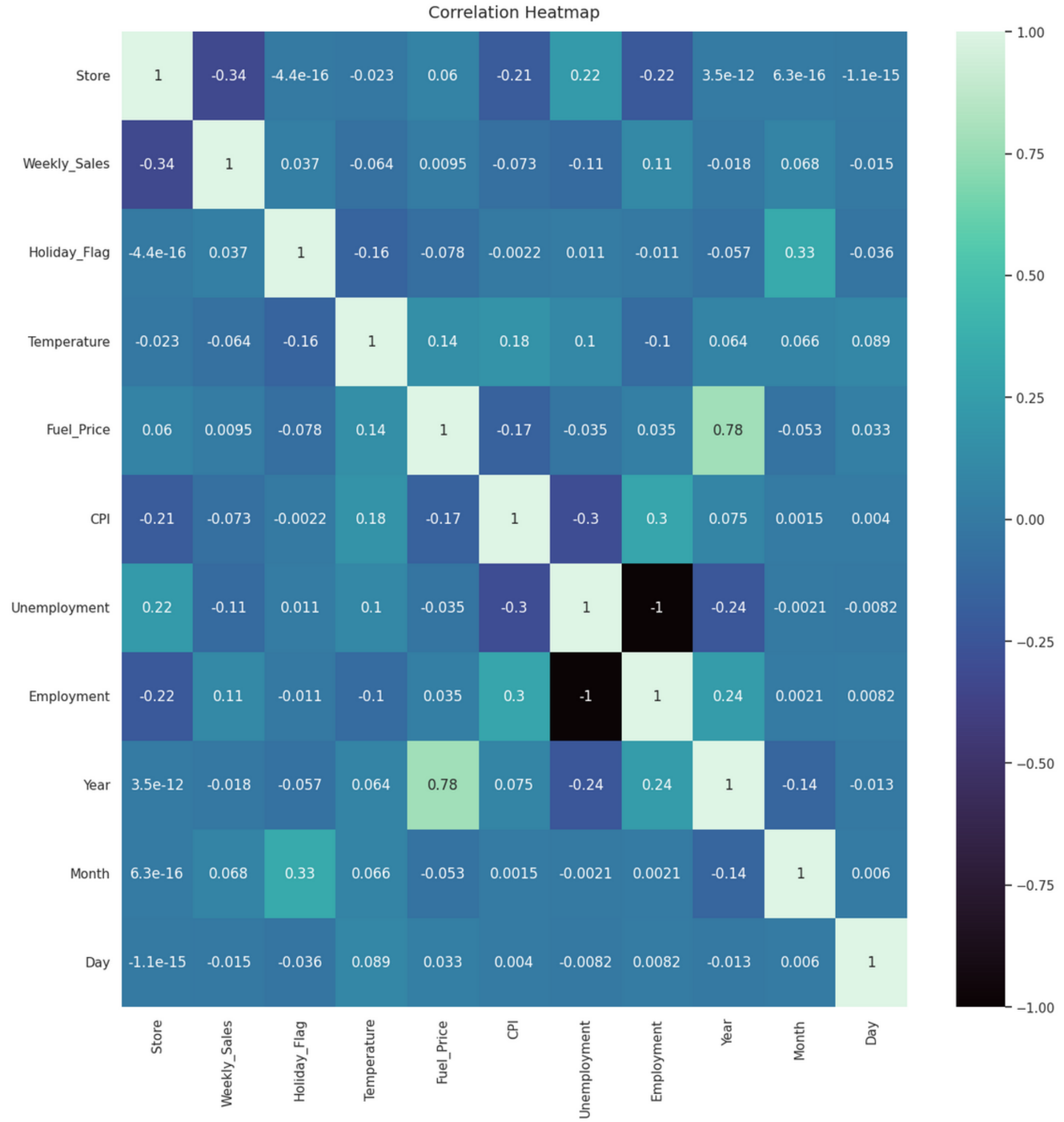
```
[ ]  df['Year'] = df['Date'].dt.year
     df['Month'] = df['Date'].dt.month
     df['Day'] = df['Date'].dt.day
     df.head(3)
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Employment	Year	Month	Day
0	1	2010-05-02	1643690.90	0	42.31	2.572	211.096358	8.106	91.894	2010	5	2
1	1	2010-12-02	1641957.44	1	38.51	2.548	211.242170	8.106	91.894	2010	12	2
2	1	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106	91.894	2010	2	19

Correlation Heatmap

หาความสัมพันธ์(correlation) **ระหว่างคอลัมน์ต่างๆ กับคอลัมน์** Weekly_Sales

- **ความสัมพันธ์ที่มากที่สุดคือ** Weekly_Sales **กับตัวแปร** Employment **มีค่า** correlation coefficient **เท่ากับ** 0.11



```
df_copy = df.copy()
df_copy.head(3)
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Employment	Year	Month	Day
0	1	2010-05-02	1643690.90	0	42.31	2.572	211.096358	8.106	91.894	2010	5	2
1	1	2010-12-02	1641957.44	1	38.51	2.548	211.242170	8.106	91.894	2010	12	2
2	1	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106	91.894	2010	2	19

	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Employment	Year	Month	Day
0	1	1643690.90	0	42.31	2.572	211.096358	91.894	2010	5	2
1	1	1641957.44	1	38.51	2.548	211.242170	91.894	2010	12	2
2	1	1611968.17	0	39.93	2.514	211.289143	91.894	2010	2	19

SEPERATE X AND y

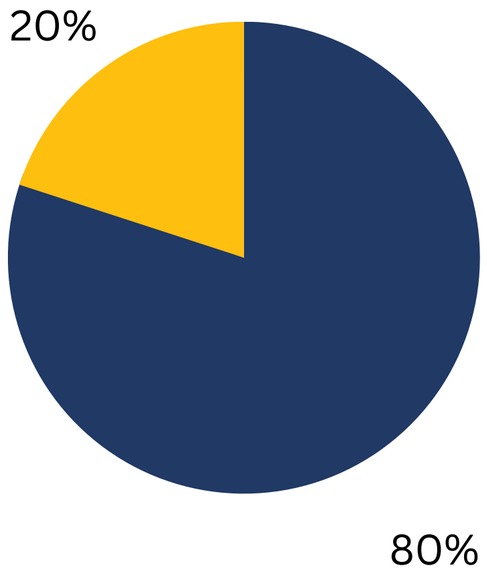
```
X = df_copy.drop('Weekly_Sales', axis=1)
y = df_copy['Weekly_Sales']
```

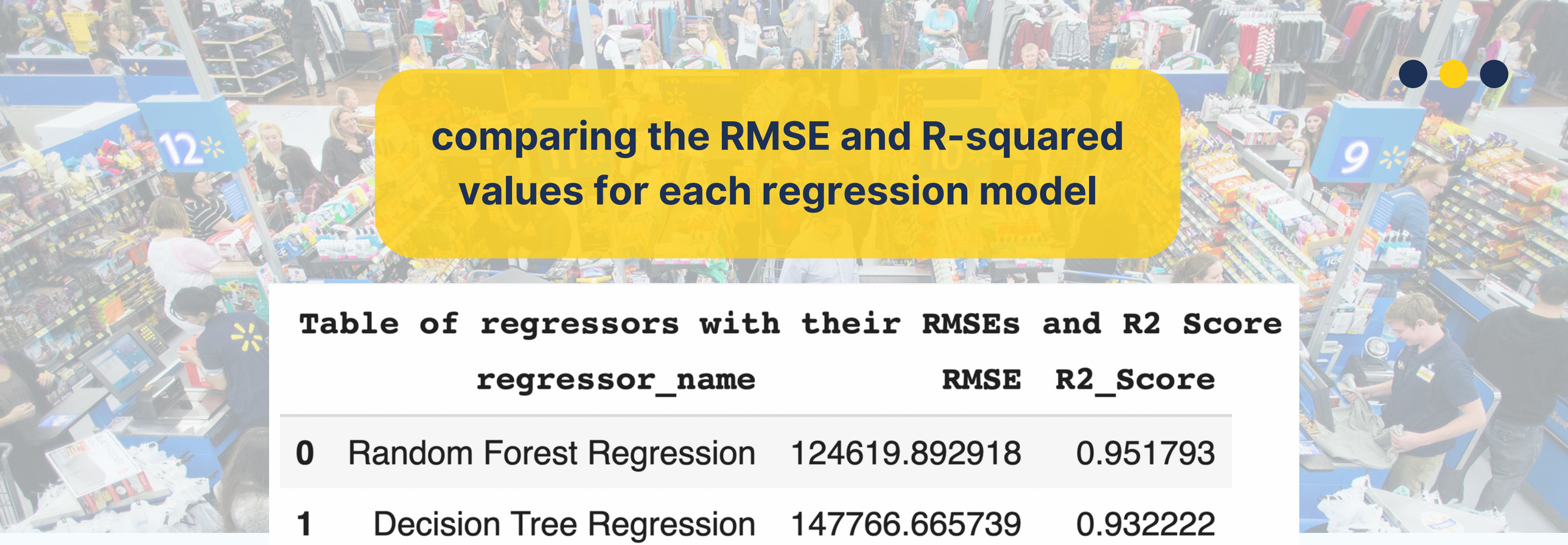
SPLIT TRAIN AND TEST DATA

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
print(f'Train data: {X_train.shape}')
print(f'Train target: {y_train.shape}')
print(f'Test data: {X_test.shape}')
print(f'Test target: {y_test.shape}')
```

```
Train data: (5148, 9)
Train target: (5148,)
Test data: (1287, 9)
Test target: (1287,)
```





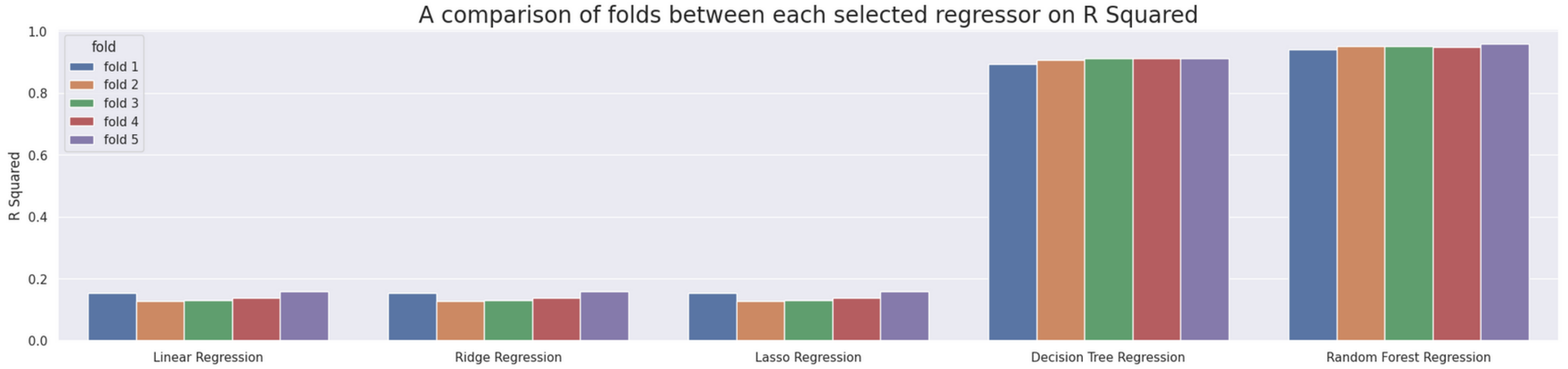
comparing the RMSE and R-squared values for each regression model

Table of regressors with their RMSEs and R2 Score

	regressor_name	RMSE	R2_Score
0	Random Forest Regression	124619.892918	0.951793
1	Decision Tree Regression	147766.665739	0.932222
2	Linear Regression	522417.695958	0.152828
3	Lasso Regression	522418.001490	0.152827
4	Ridge Regression	522420.773473	0.152818

Cross-Validation

วิธีการนี้จะแบ่งข้อมูลออกเป็นกลุ่ม ๆ แล้วทำซ้ำตามจำนวนที่กำหนด



HYPERPARAMETER TUNING AND VISUALIZING MODEL

Multiple Linear Regression

Multiple Linear Regression Formula:

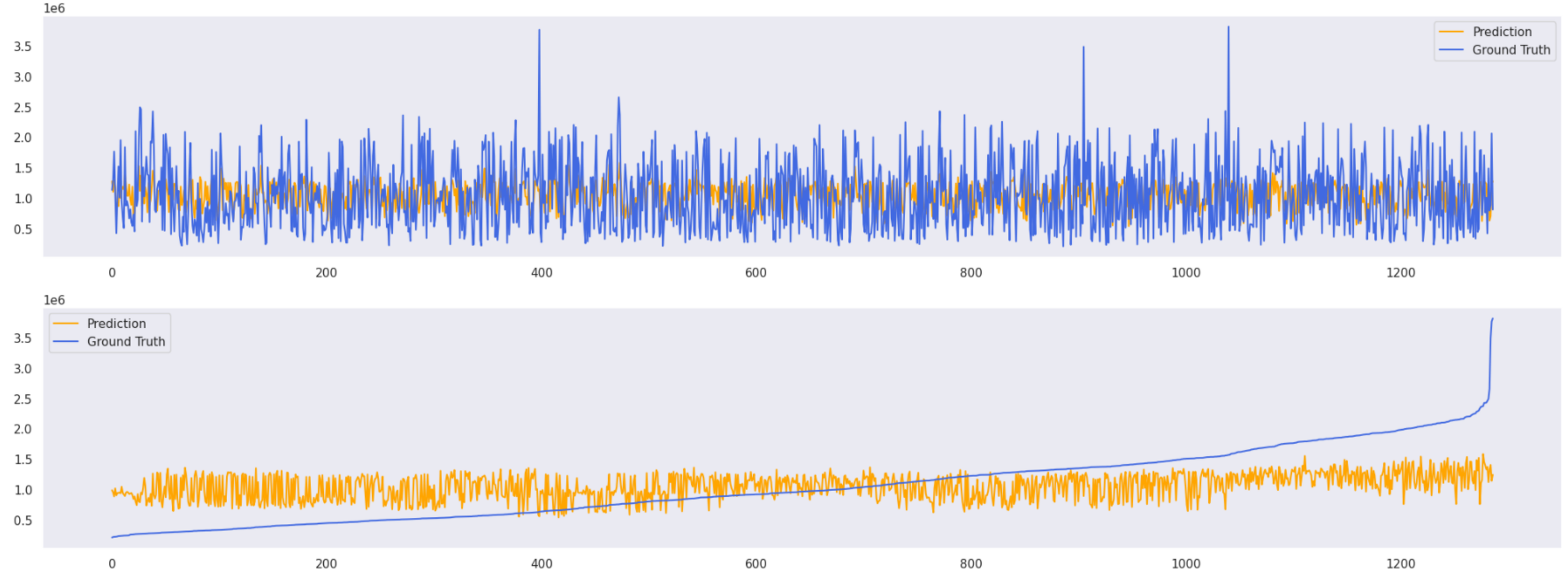
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

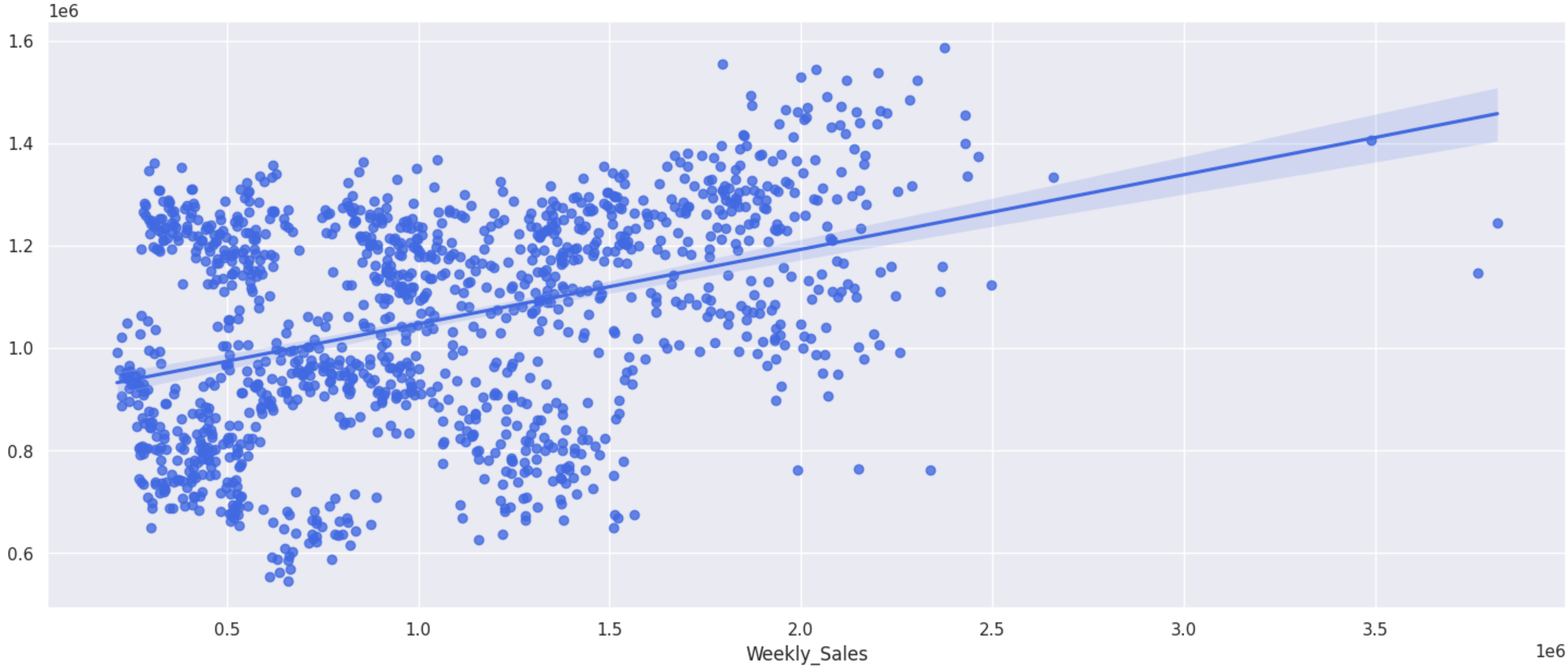
Multiple Linear Regression Formula of our Dataset

Weekly_Sales = 73901024.50215305 - 15076.402159991327(Store) + 25071.866287670422(Holiday_Flag) - 1105.137079939652(Temperature) + 53045.089734490095(Fuel_Price) - 2147.397289377168(CPI) + 26267.934578226337(Employment) - 37155.85088079869(Year) + 11054.534727010883(Month) - 1323.578355437419 8(Day)

Multiple Linear Regression actual vs predicted values

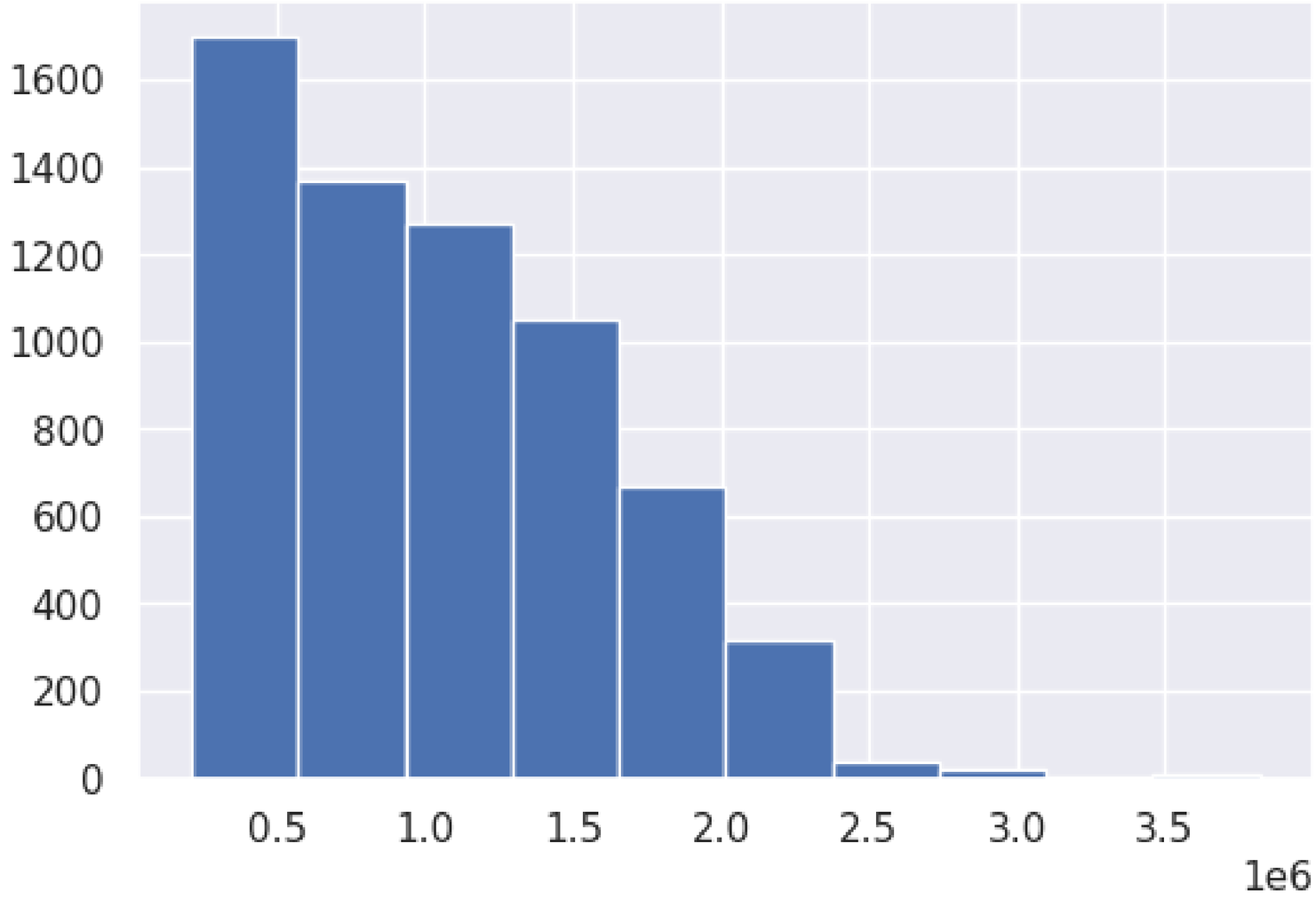


พล็อต Scatter plot พร้อมเส้น Regression line ของ multiple linear regresion



พล็อต Histogram plot ของตัวแปรเป้าหมาย Weekly_Sales

ดูการกระจายของข้อมูล เพื่อกำหนดค่าทางสถิติ



คำนวณค่า Percentage Error จากค่า RMSE

Percentage Error = (RMSE / Median of target variable) * 100

Multiple Linear Regressor มีค่า percentage error เท่ากับ 54.38%

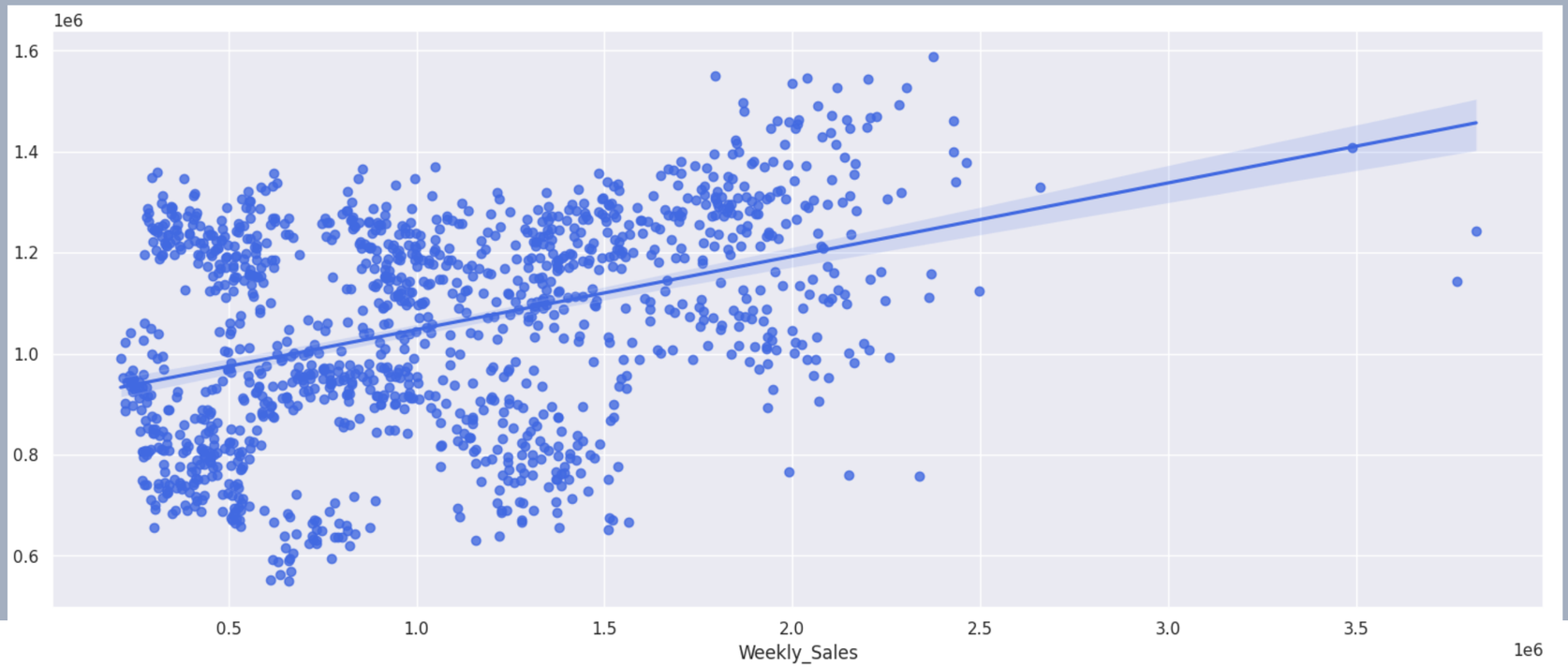
hyperparameter ของ Ridge คือค่า alpha

COST FUNCTION

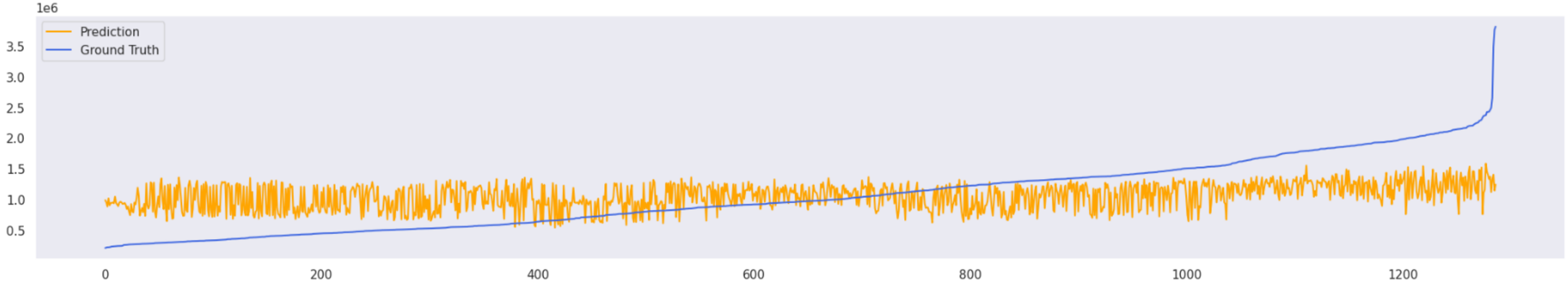
$$\sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

	alpha	RMSE
0	0.001	522417.699044
1	0.010	522417.726821
2	0.100	522418.004505
3	1.000	522420.773473
4	10.000	522447.693299
5	100.000	522655.251411

ค่า RMSE สำหรับแต่ละค่า ALPHA ใน RIDGE REGRESSION ไม่แตกต่างกันมากนัก ดังนั้นเราจะใช้ค่า ALPHA เดิมที่เป็นค่า DEFAULT ต่อไป (ALPHA=1.0)



SCATTER PLOT พร้อมเส้น REGRESSION LINE ของโมเดล



กราฟเปรียบเทียบค่า ACTUAL กับค่า PREDICTED ของโมเดล

PERCENTAGE ERROR เท่ากับ 54.38%

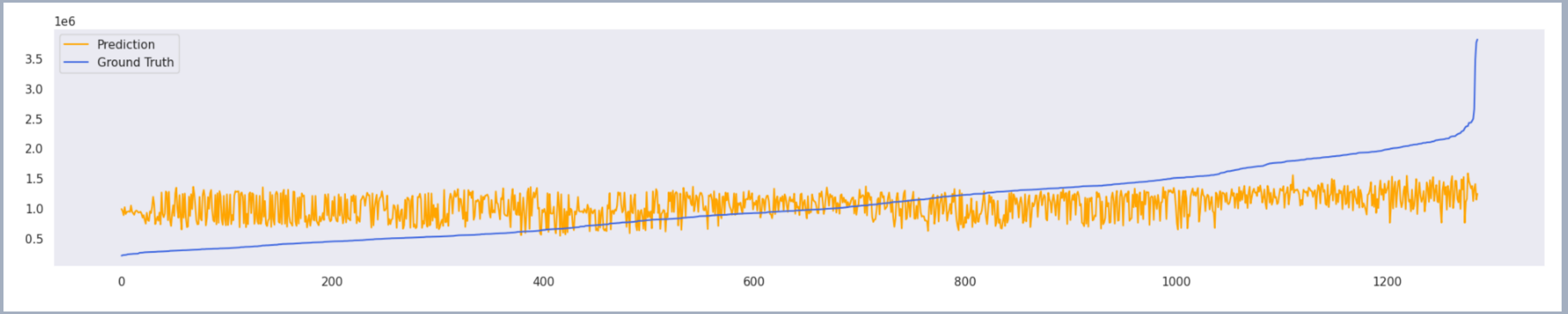
hyperparameter ของ Lasso คือค่า alpha

COST FUNCTION

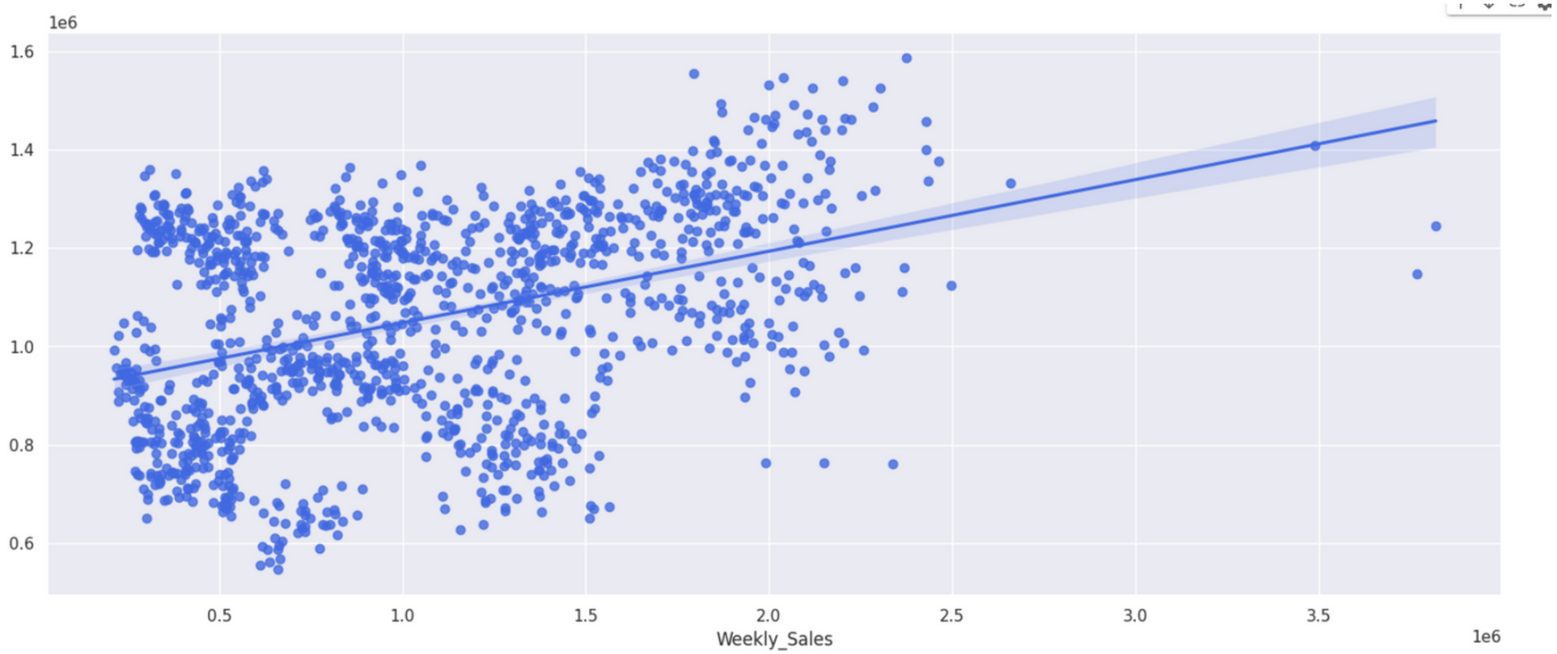
$$\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

	alpha	RMSE
0	0.001	522417.696263
1	0.010	522417.699007
2	0.100	522417.726482
3	1.000	522418.001490
4	10.000	522420.757654
5	100.000	522448.897885

RMSE สำหรับแต่ละค่า ALPHA ใน LASSO REGRESSION ไม่แตกต่างกันมากนัก ดังนั้นเราจะใช้ค่า ALPHA เดิมที่เป็นค่า DEFAULT ต่อไป (ALPHA=1.0)



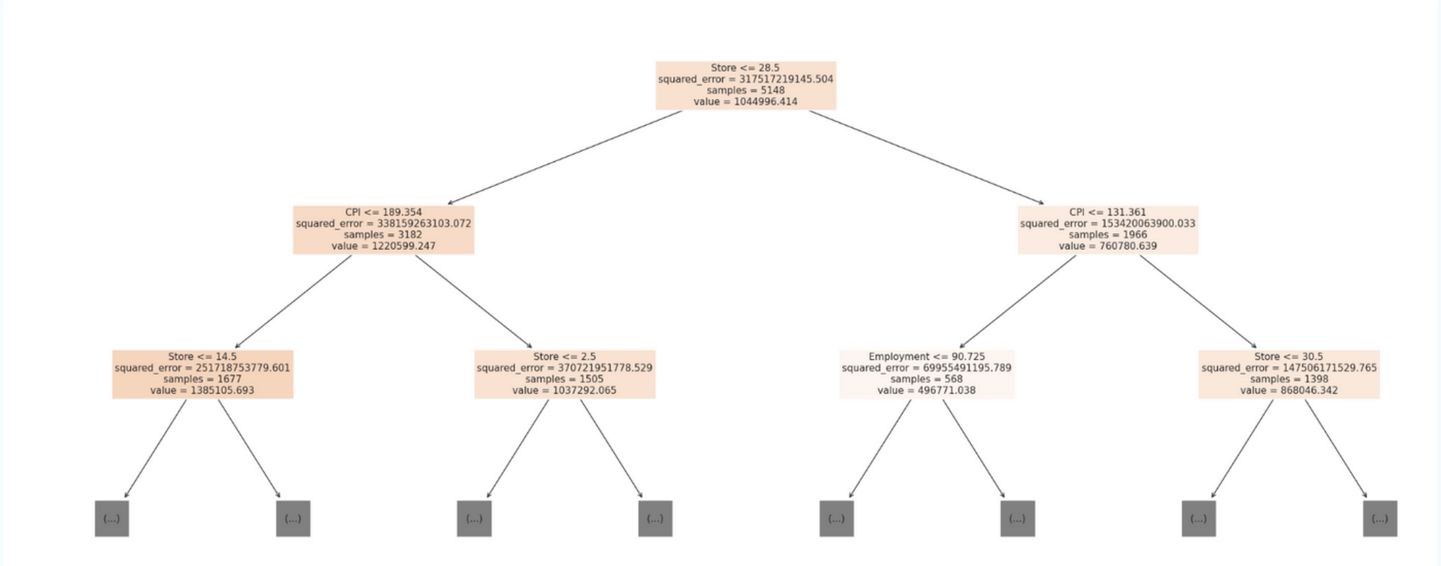
กราฟเปรียบเทียบค่า ACTUAL กับค่า PREDICTED ของโมเดล



SCATTER PLOT พร้อมเส้น REGRESSION LINE ของโมเดล

PERCENTAGE ERROR
เท่ากับ 54.38%

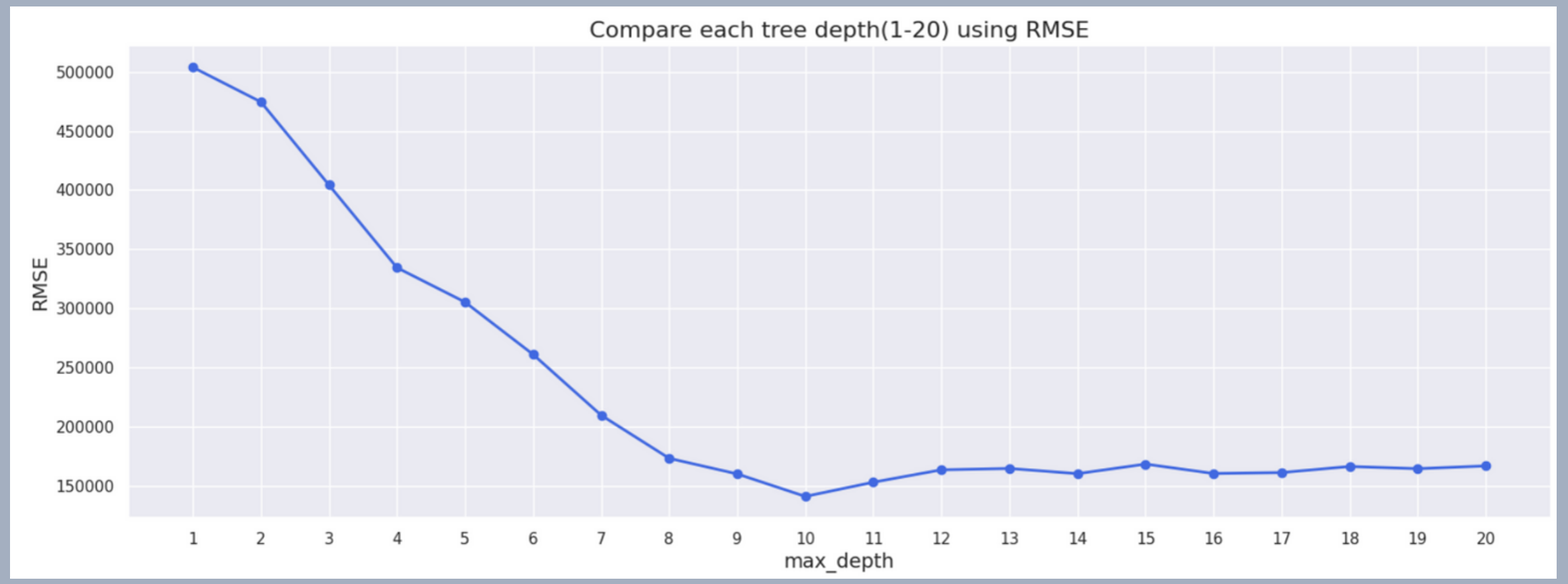
hyperparameter ของ decision tree คือค่า (max-depth)



TREE ของ DECISION_TREE_REGRESSOR ที่ (MAX_DEPTH = 2)

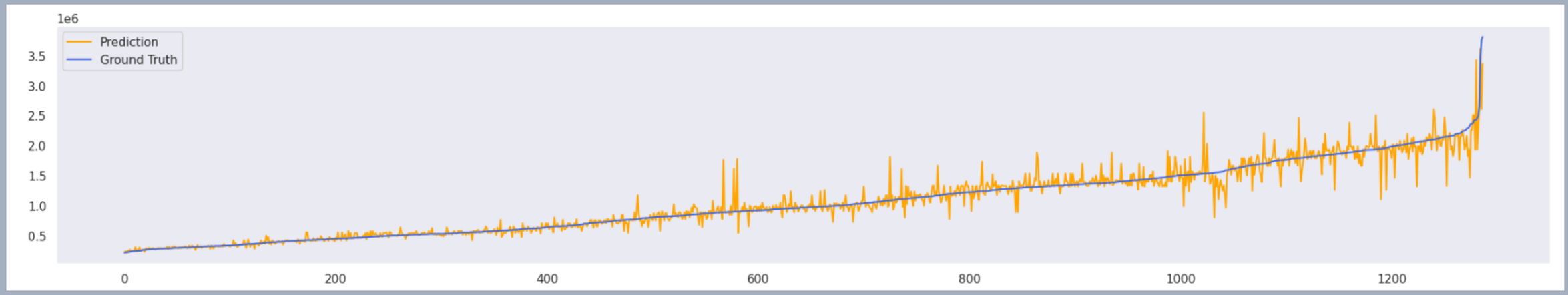
	Variable	Importance
0	Store	0.656773
4	CPI	0.142185
5	Employment	0.130524
8	Day	0.021777
7	Month	0.015958
3	Fuel_Price	0.015345
2	Temperature	0.012598
1	Holiday_Flag	0.004744
6	Year	0.000096

ตรวจสอบความสำคัญของแต่ละตัวแปร(FEATURE)

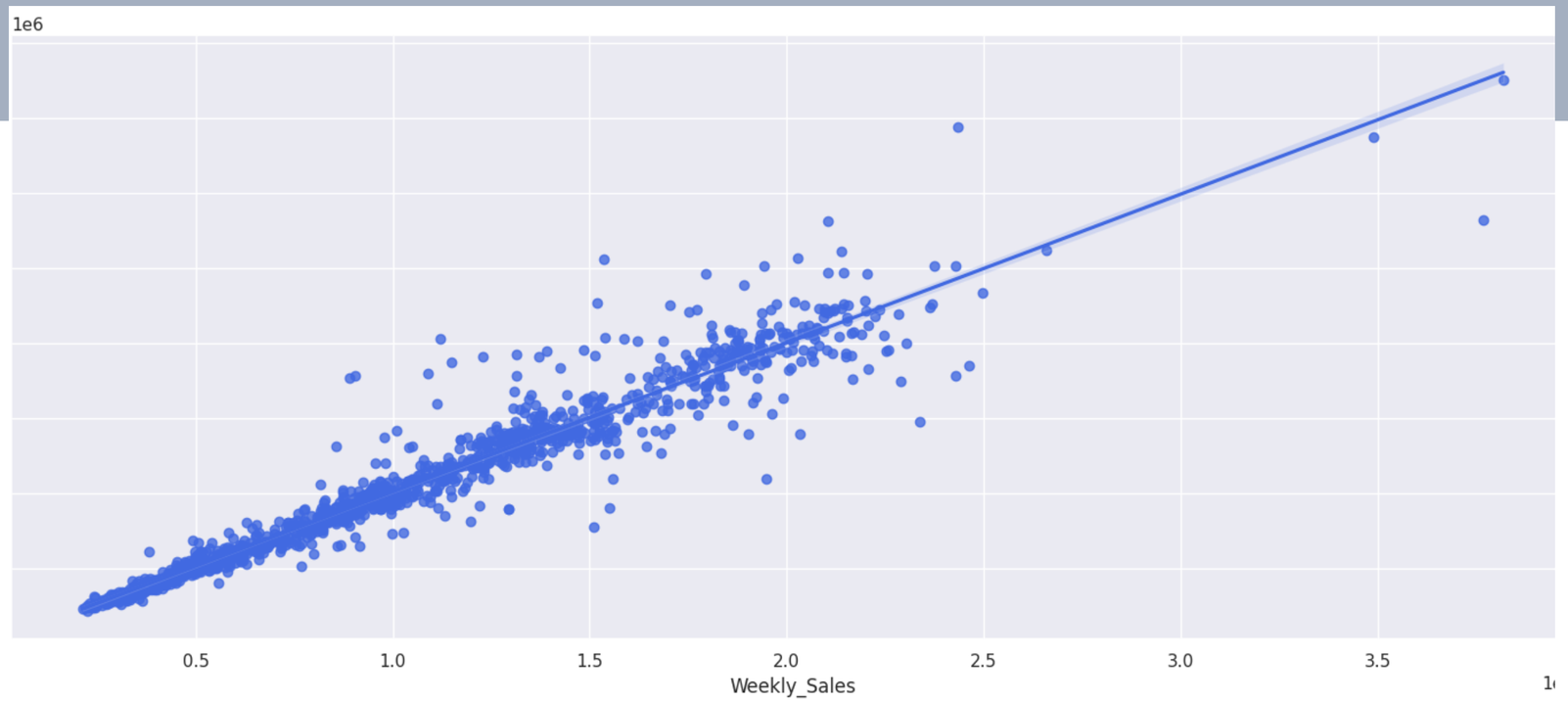


เปรียบเทียบค่า RMSE ของแต่ละ MAX_DEPTH

- **ค่า RMSE ที่ต่ำที่สุดและตอนโมเดลเริ่มเรียนรู้จากชุดข้อมูล TRAINING แต่ก่อนที่จะเกิดการ OVERFITTING เราจะเลือก TREE DEPTH ที่เท่ากับ 10**



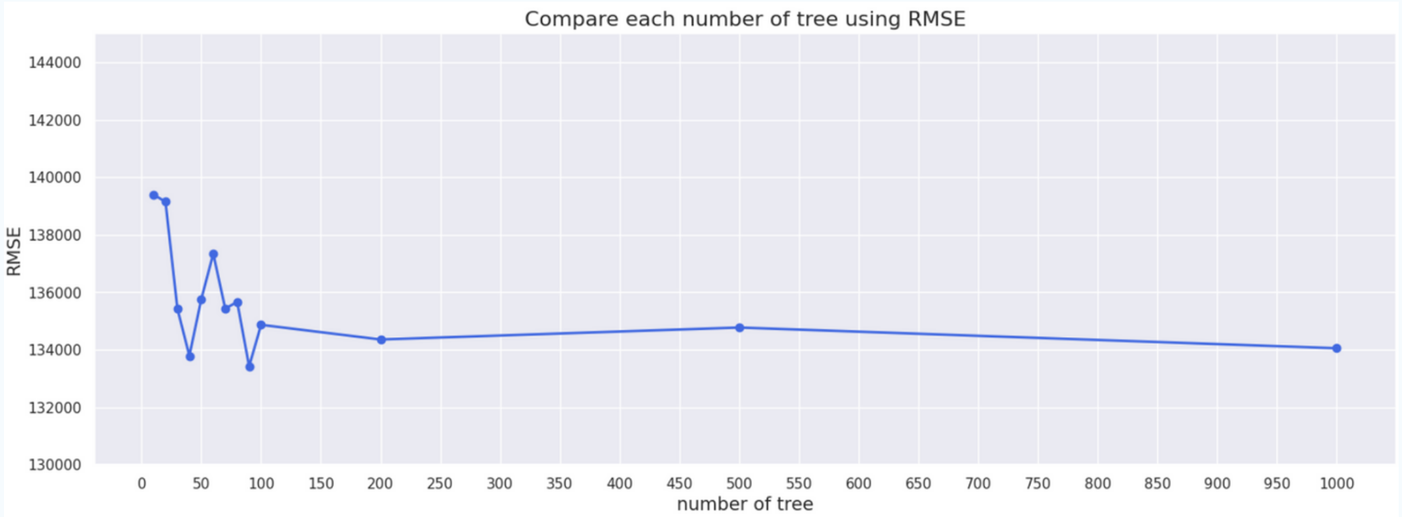
กราฟเปรียบเทียบค่า ACTUAL กับค่า PREDICTED ของโมเดล



SCATTER PLOT พร้อมเส้น REGRESSION LINE ของโมเดล

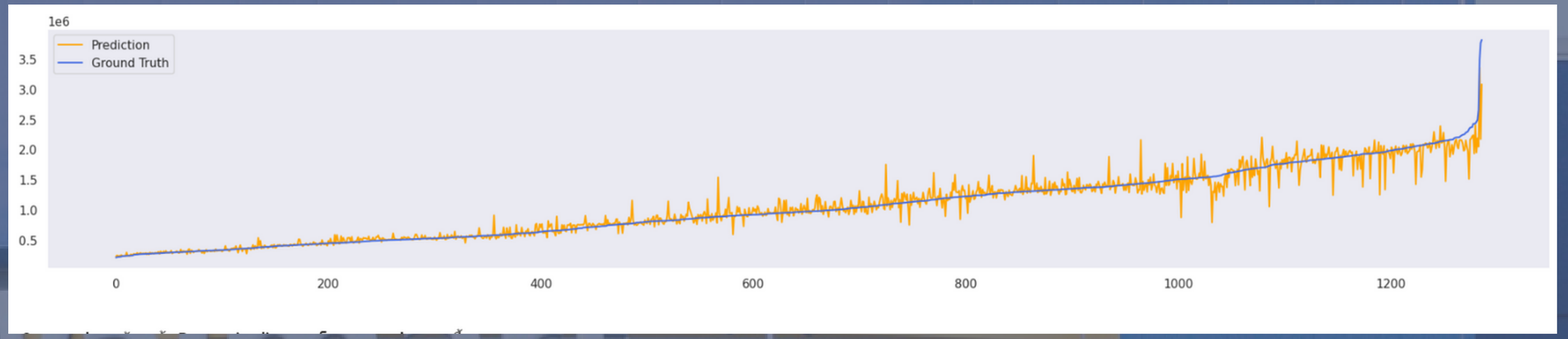
PERCENTAGE ERROR
เท่ากับ 15.43%

hyperparameter ของ random forest คือค่า n -estimate

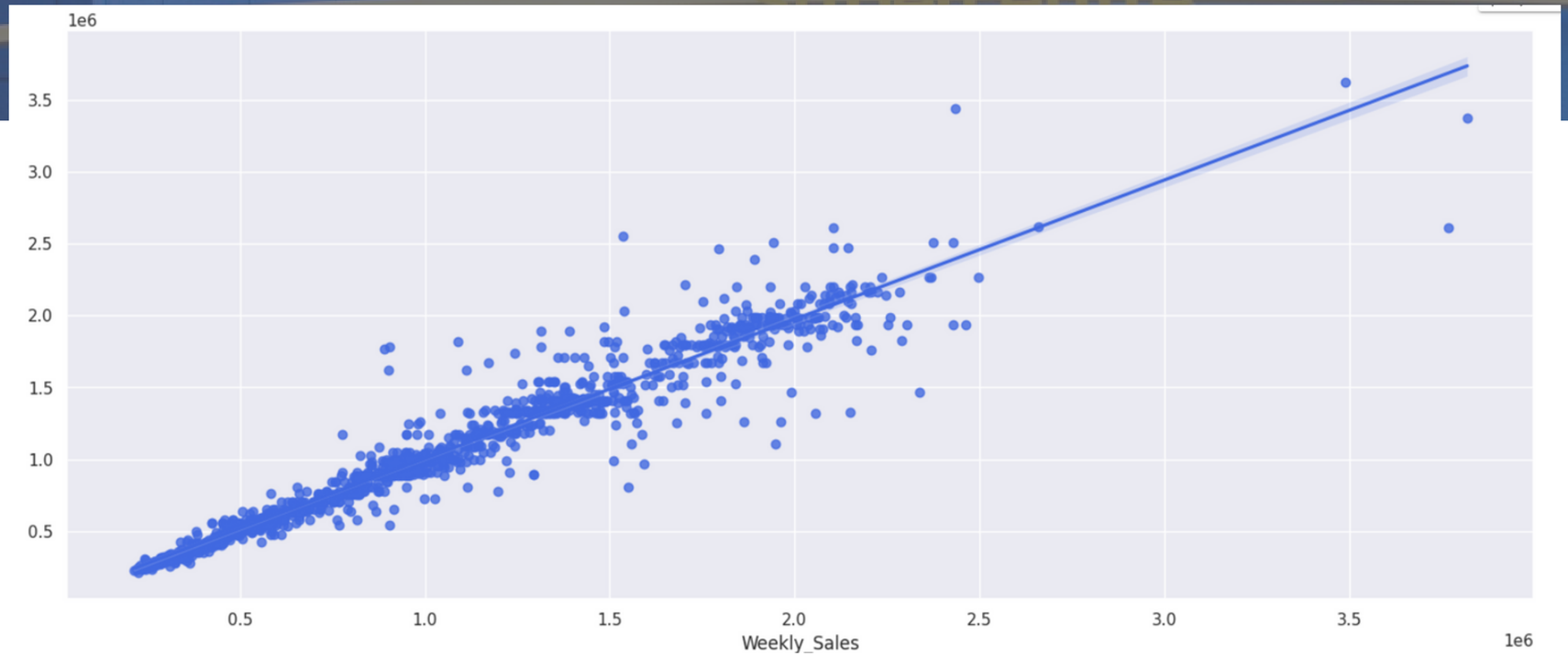


เปรียบเทียบค่า RMSE ของแต่ละ N - ESTIMATE

	number_of_trees	RMSE
0	90	133423.425640
1	40	133790.230612
2	1000	134053.723197
3	200	134357.488247
4	500	134773.484425
5	100	134867.708749
6	30	135426.629817
7	70	135431.985147
8	80	135665.325313
9	50	135764.981716
10	60	137340.970439
11	20	139147.846581
12	10	139399.067231

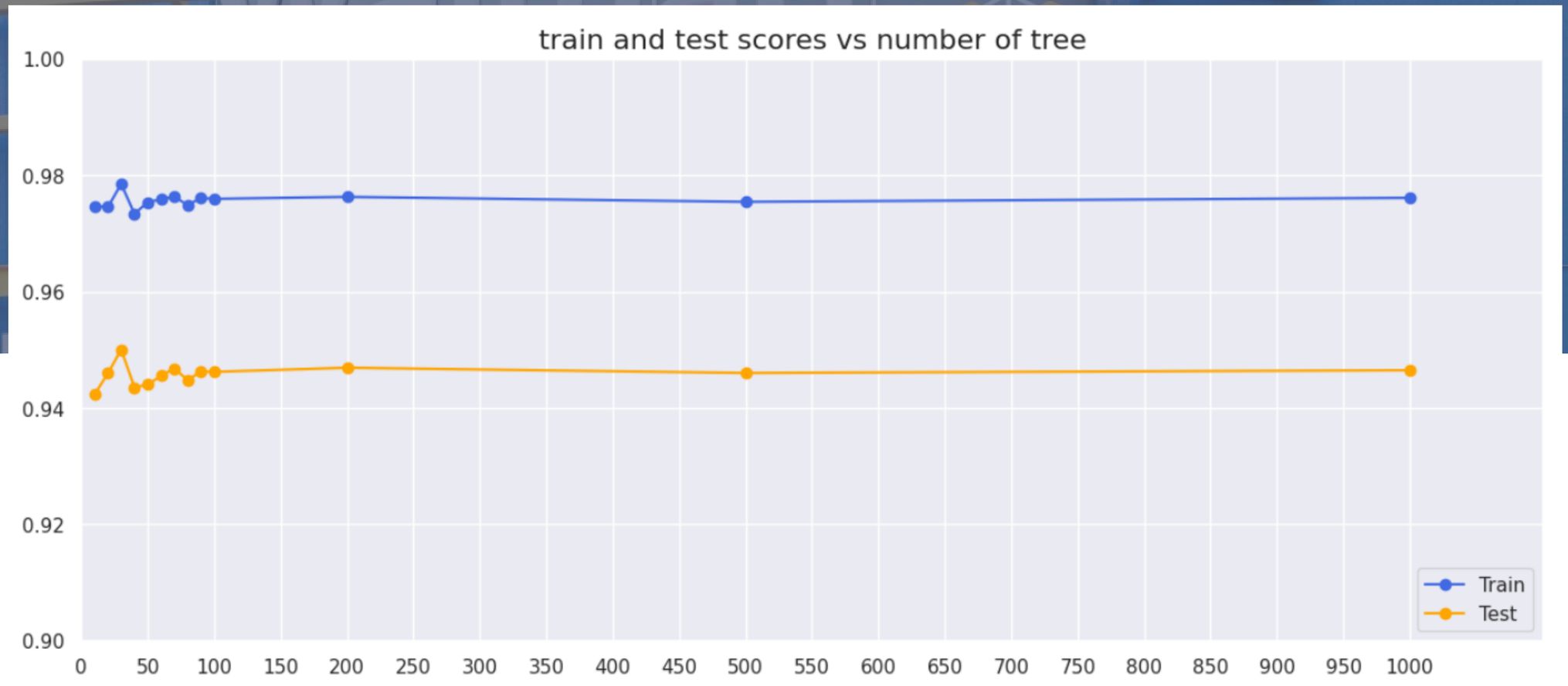


กราฟเปรียบเทียบค่า ACTUAL กับค่า PREDICTED ของโมเดล



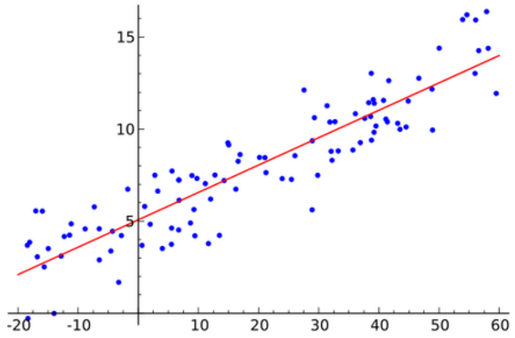
PERCENTAGE ERROR เท่ากับ 15.01%

SCATTER PLOT พร้อมเส้น REGRESSION LINE ของโมเดล

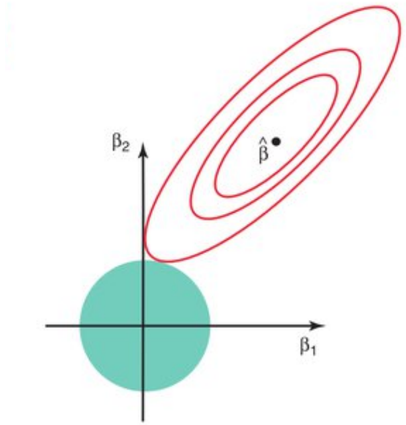


ผลการทดสอบแต่ละครั้งก็แตกต่างกันไป โดยก็อาจนำไปสู่การ OVERFITTING
ดังนั้นเราจะต้องทำการตรวจสอบว่าโมเดลเริ่มมีความ OVERFIT ที่จำนวนต้นไม้เท่าใด

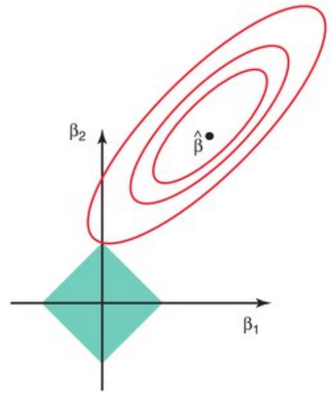
Conclusion



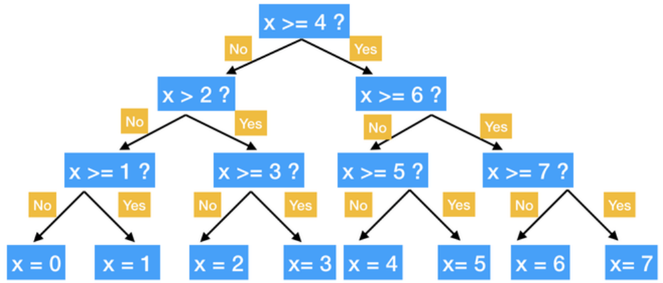
Multiple Linear Regression



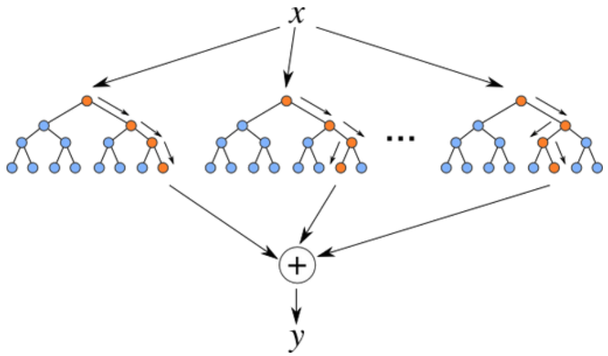
Ridge Regression



Lasso Regression

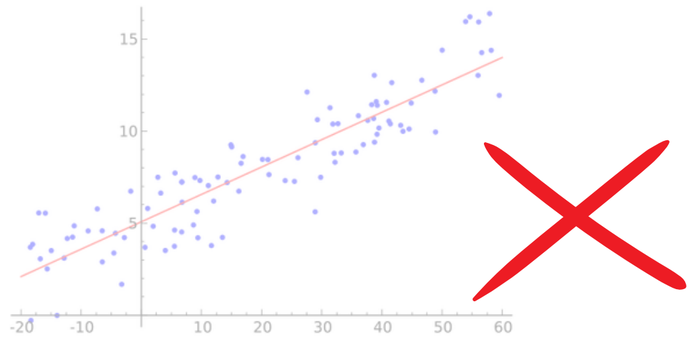


Decision Tree Regression

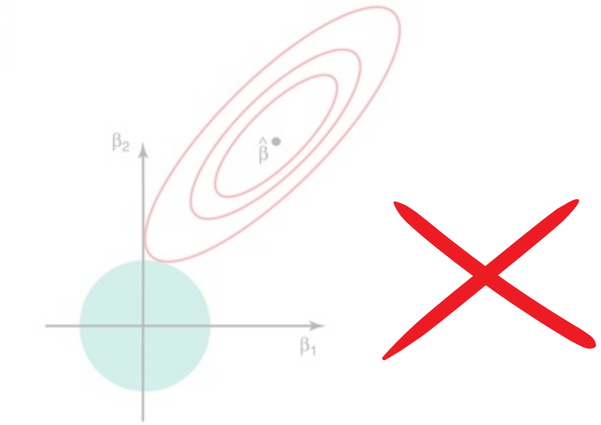


Random Forest Regression

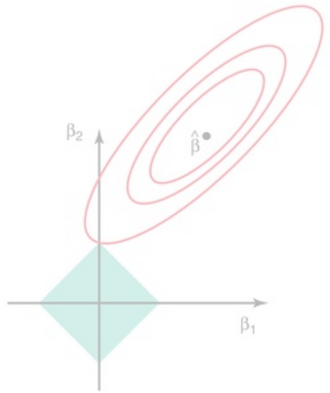
Conclusion



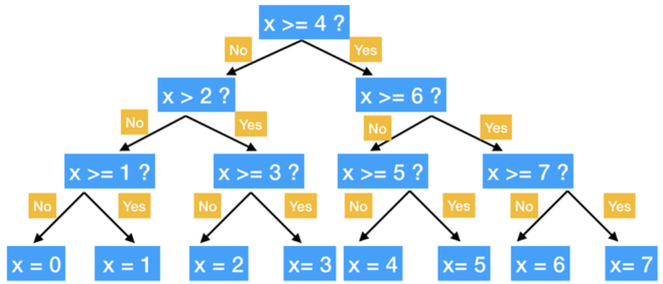
Multiple Linear Regression



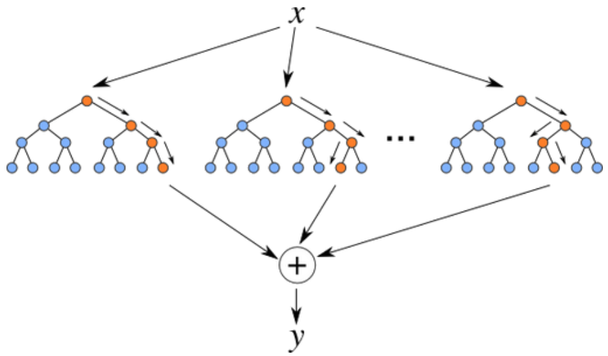
Ridge Regression



Lasso Regression

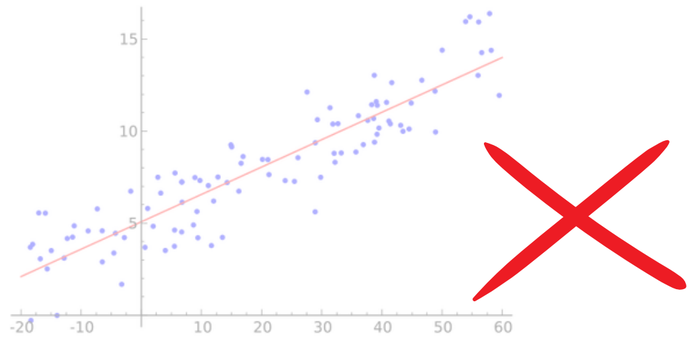


Decision Tree Regression

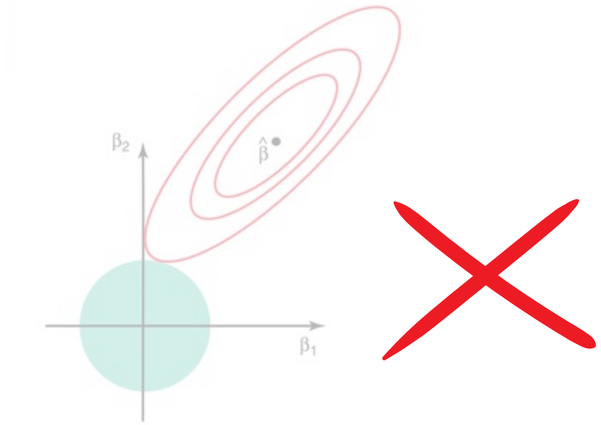


Random Forest Regression

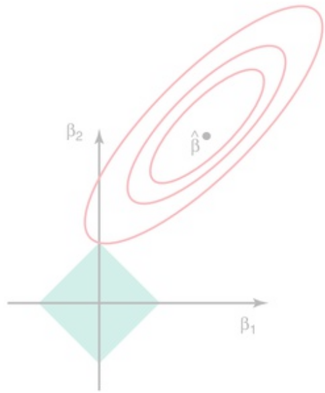
Conclusion



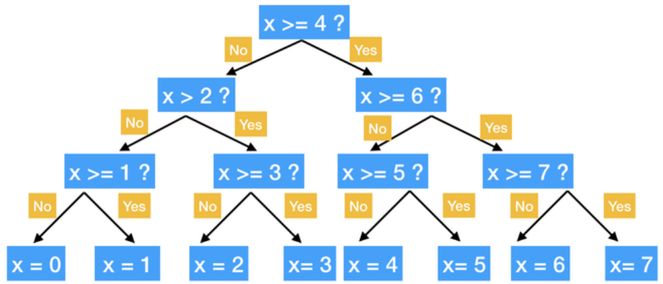
Multiple Linear Regression



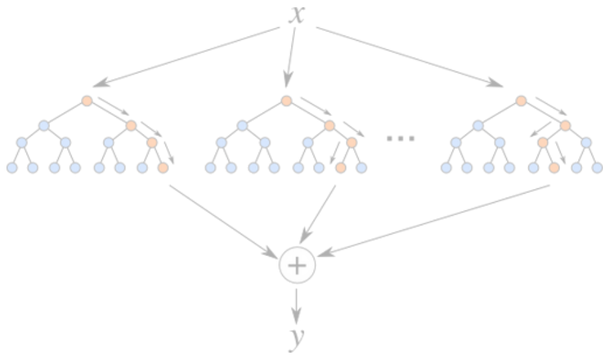
Ridge Regression



Lasso Regression

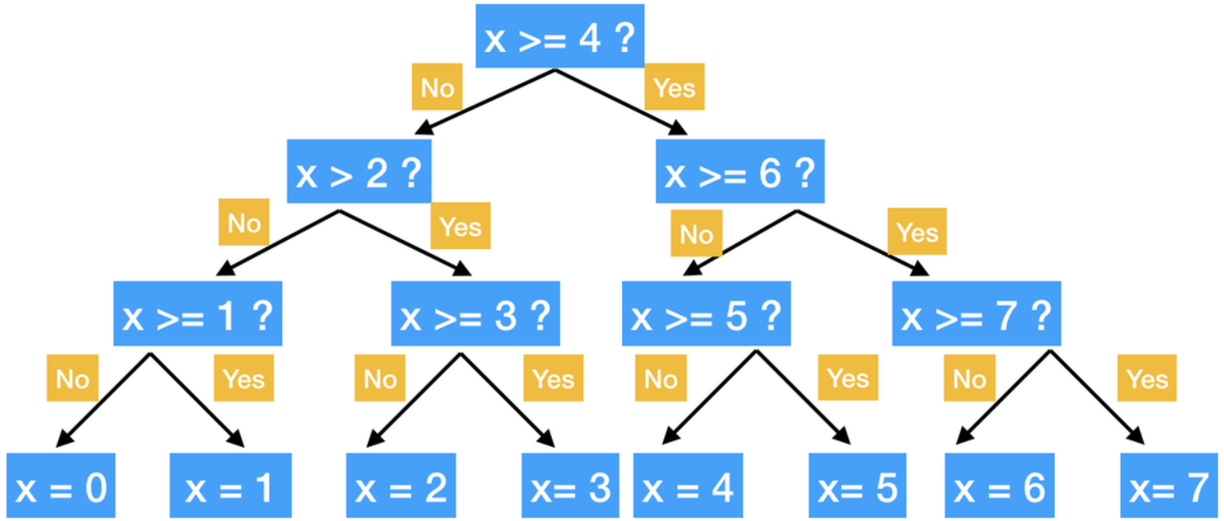


Decision Tree Regression



Random Forest Regression

Conclusion



Decision Tree Regression

Thanks For Listening

DSI205: Least-Squares Problem Project