Regression Project

Nicola Szwaja
Piotr Droś

Goal of the project

1

- Analyse the data

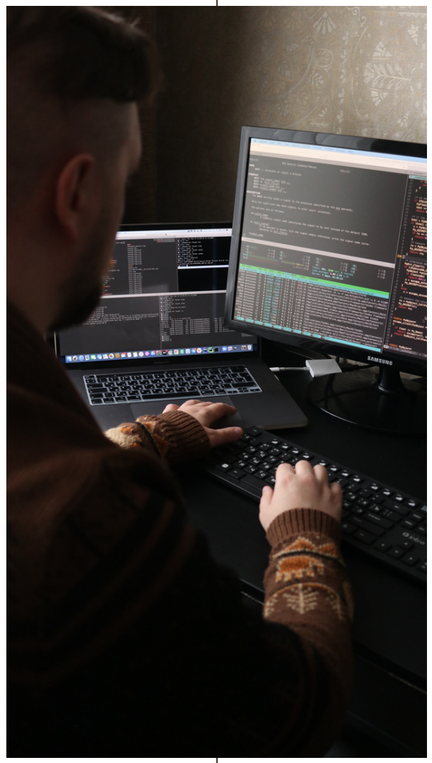
2

- Pick what the model will predict
- process the data

3

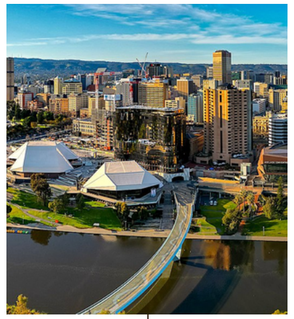
- Create models
- Pick the best one

Dataset overview



Data consists of **positions** and **absorbed power** outputs of wave energy converters (WECs) in four locations.

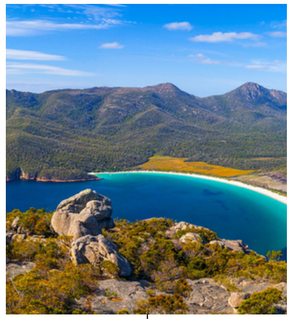
WEC’s Locations



Adelaide

Sydney

Perth



Tasmania

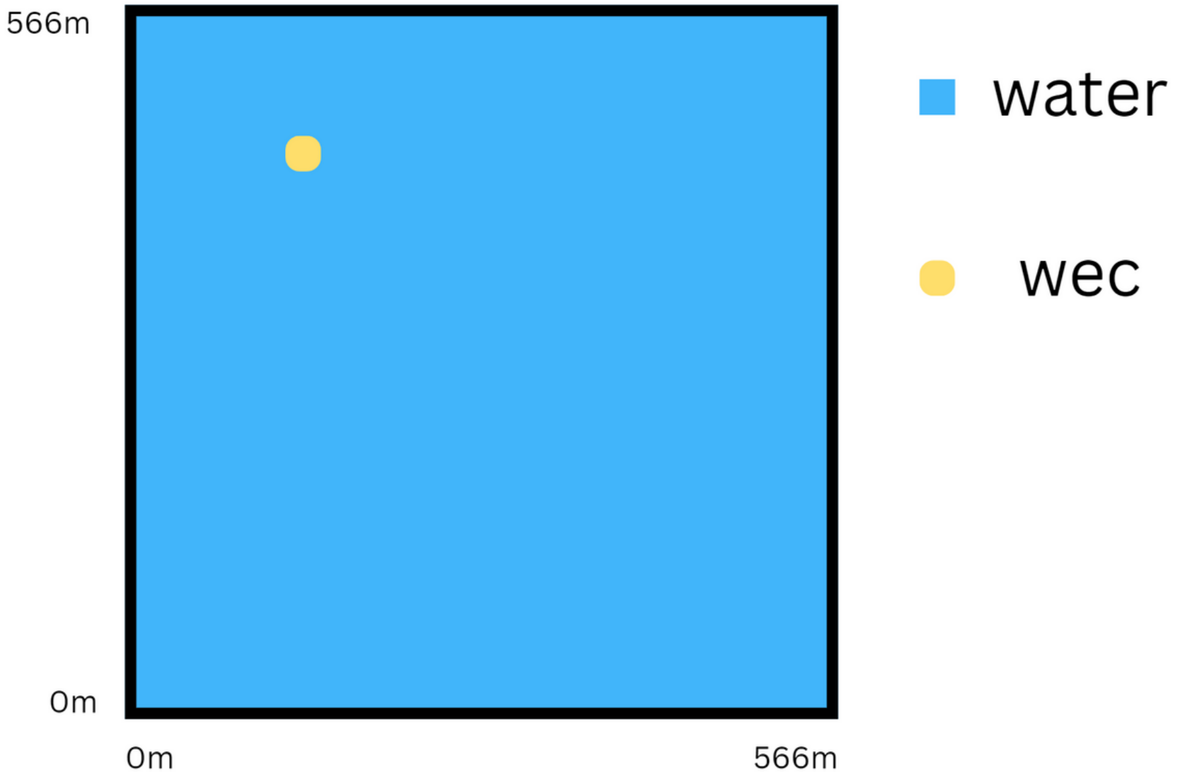
Example dataset

	0	1	2	3	4	5	6	7	8	9	...
0	316.5855	223.9277	182.3434	551.5497	7.8641	243.1339	361.0877	115.9284	78.6087	468.3799	...
1	530.3136	68.7031	31.5983	175.2582	516.1441	63.4652	67.0954	369.4486	14.0930	375.4462	...
2	27.3967	399.0488	565.6854	394.0466	120.2245	558.1293	546.4520	27.3256	314.1051	235.9476	...
3	346.1526	59.6375	226.2742	280.9095	402.2161	218.7181	207.0407	339.5676	0.0000	0.0000	...
4	317.9144	551.8542	335.4745	40.0240	316.6285	365.6434	416.3060	562.1028	211.3577	143.1255	...
...	...	...	...	...	...	...	...	...	...	...	...
71994	3.8797	404.4992	234.2780	376.1119	102.9226	36.2912	514.6270	46.8826	445.1958	349.5316	...
71995	46.1547	487.8102	219.0245	86.9150	0.0000	220.0181	245.4543	566.0000	186.7975	387.0030	...
71996	46.1547	487.8102	219.0245	91.3298	0.0000	170.9280	245.4543	566.0000	159.1193	354.2496	...
71997	46.1547	487.8102	219.0245	86.9150	0.0000	170.9280	270.4428	566.0000	186.7975	383.7056	...
71998	46.1547	487.8102	310.7891	86.9150	0.0000	170.9280	245.4543	566.0000	186.7975	433.9646	...

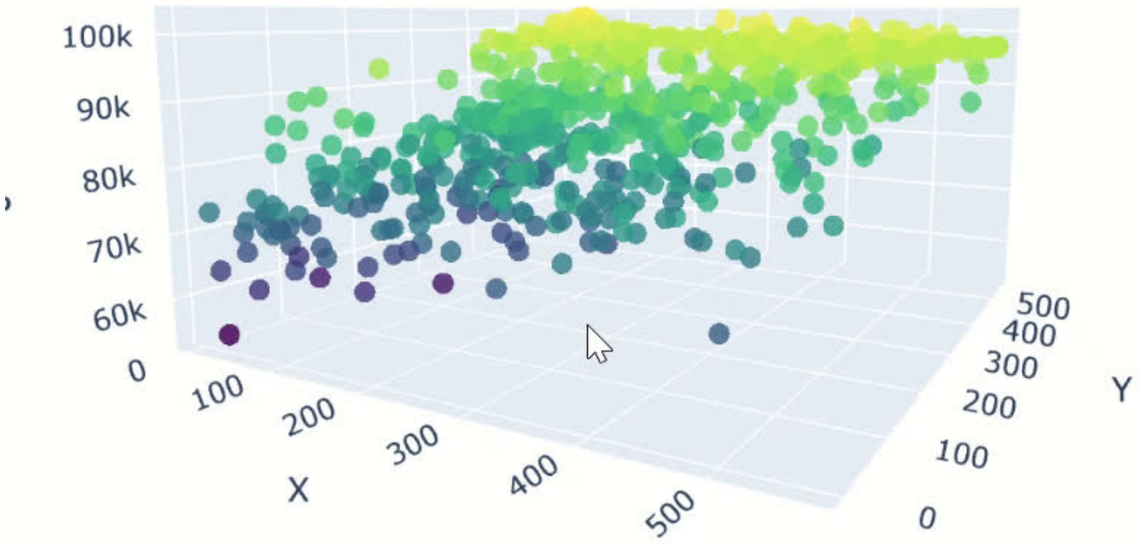
Column naming

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	P8	P9	P10	P11	P12	P13	P14	P15	P16	power_all
0	316.5855	223.9277	182.3434	551.5497	7.8641	243.1339	361.0877	115.9284	78.6087	468.3799	...	82322.0277	98069.1011	86578.6330	93016.4133	63145.1829	98353.1952	80225.1390	98447.2846	97570.2225	1370374.145

Understanding data



- 16 WECs
- coordinates range from 0 - 566m



Visualising data

Determine dataset prediction goals

1) Predicting power from single coordinates

```python
coordinates_wec1 = [1, 4]

predicted_power = model.predict(coordinates_wec1 )
```

2) Predicting total farm power from 16 pairs of coordinates

```python
coordinates = [1, 2, 3, 4, 5, 6, ... 28, 29, 30, 31, 32]

predicted_power_all = model.predict(coordinates)
```

3) Predicting power from single coordinates

```python
coordinates = [1, 4]

wec_1_adel_model.predict(coordinates)
wec_2_adel_model.predict(coordinates)
```

Determine dataset prediction goals

1) Predicting power from single coordinates

```python
coordinates_wec1 = [1, 4]

predicted_power = model.predict(coordinates_wec1 )
```

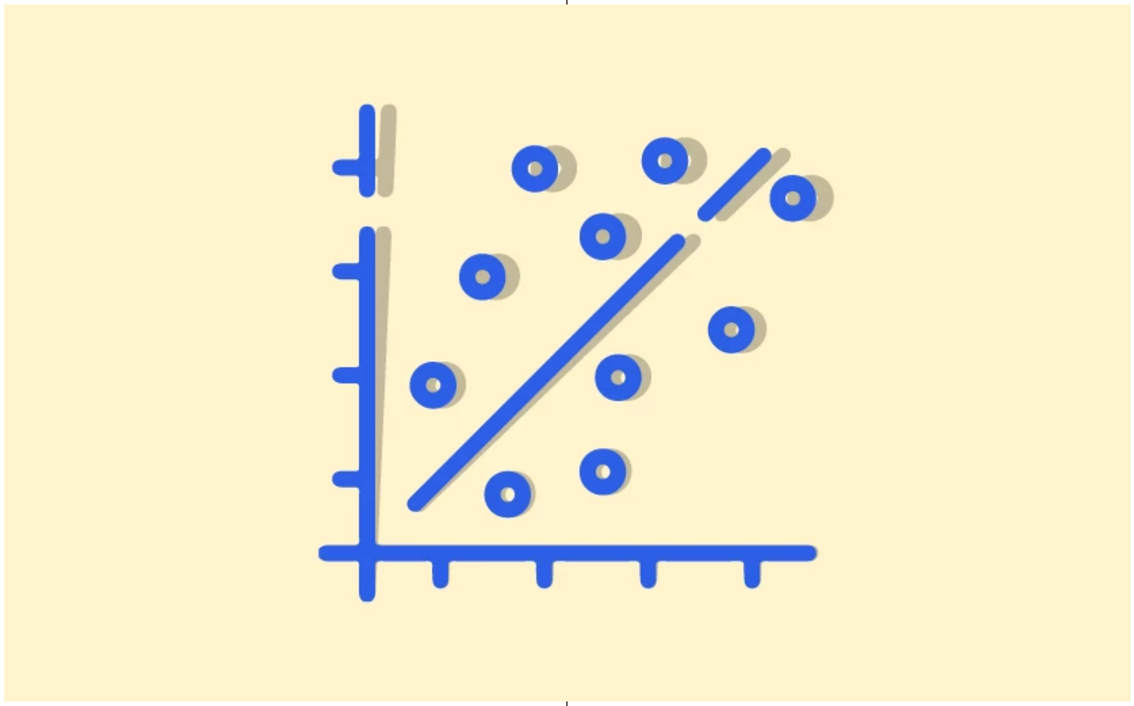
2) Predicting total farm power from 16 pairs of coordinates

```python
coordinates = [1, 2, 3, 4, 5, 6, ... 28, 29, 30, 31, 32]

predicted_power_all = model.predict(coordinates)
```

Predicting power from single coordinates

```python
coordinates = [1, 4]

wec_1_adel_model.predict(coordinates)
wec_2_adel_model.predict(coordinates)
```

Correlation



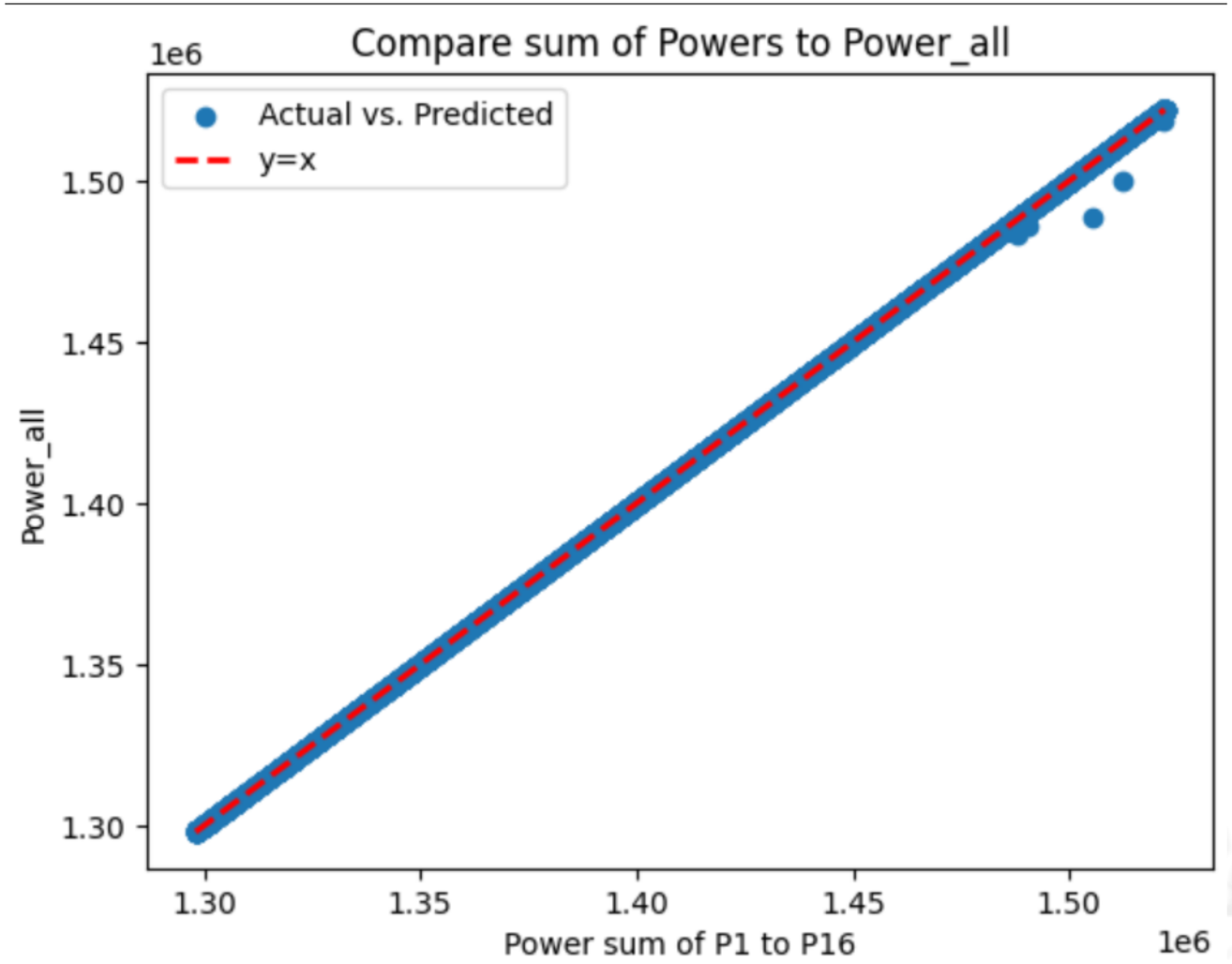
What is correlation?

statistical concept that measures the degree to which two variables change together.

Why do we do that?

- provides valuable insights for decision-making
- If two variables are strongly correlated, the value of one variable may provide information about the likely value of the other.

Correlation



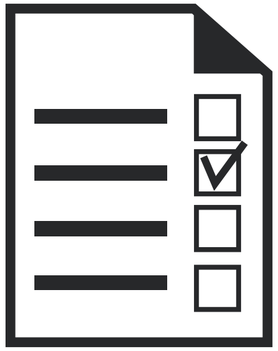
Data preprocessing

Outliers

What are outliers?

Data points that significantly differ from the rest of the observations in a dataset.

What might they cause?

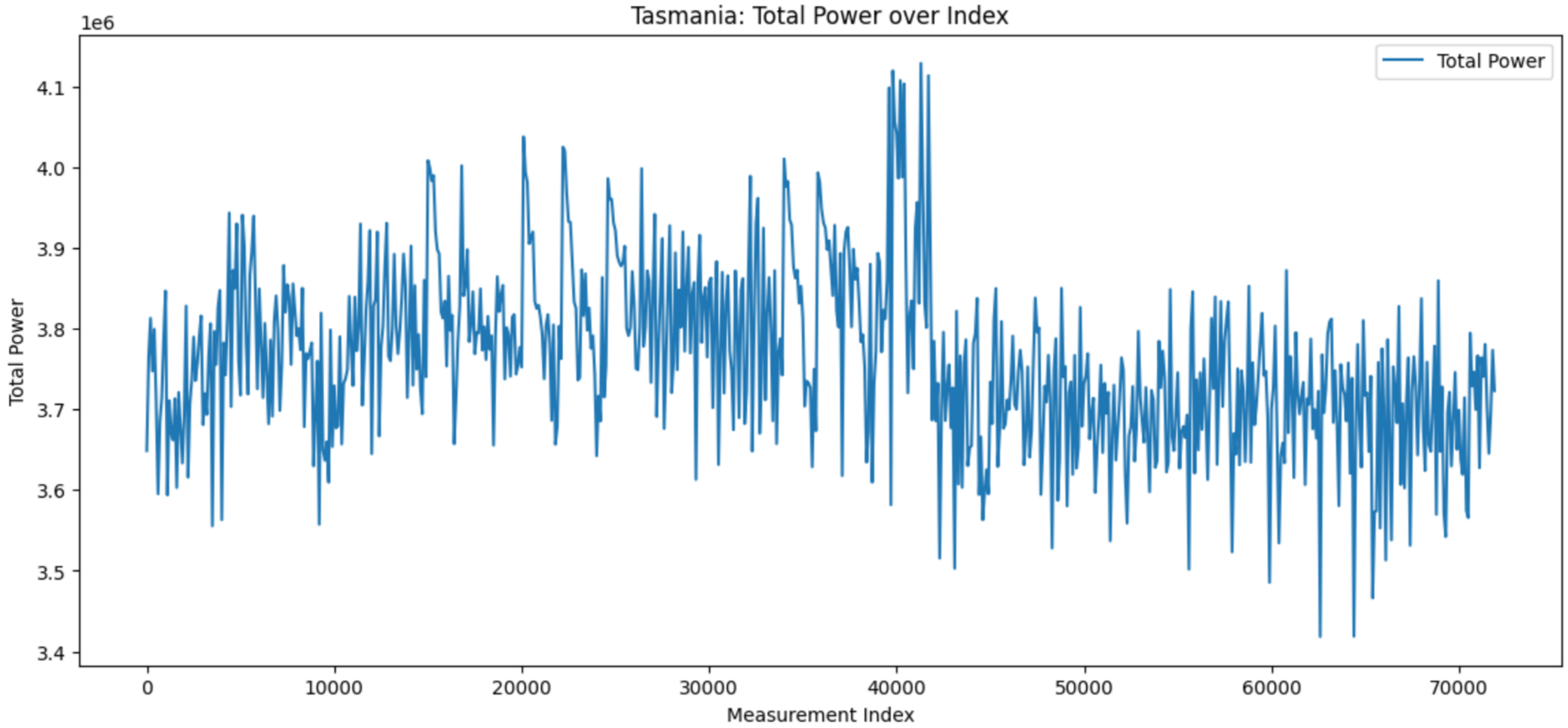


Strongly influence statistical measures such as the **mean** and **standard deviation**

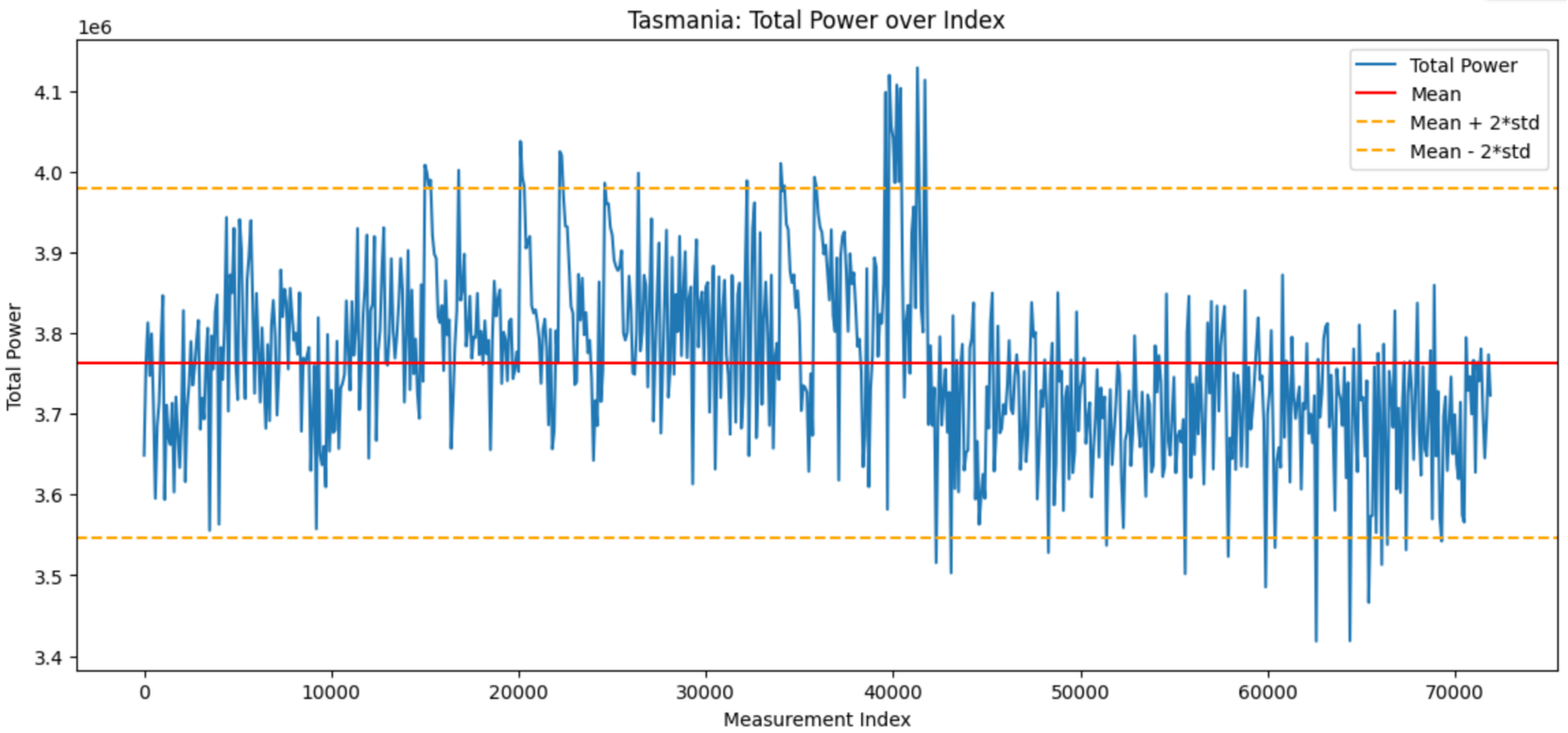
How can we handle them?

- transforming the data
- excluding them from analysis

Outliers: plot

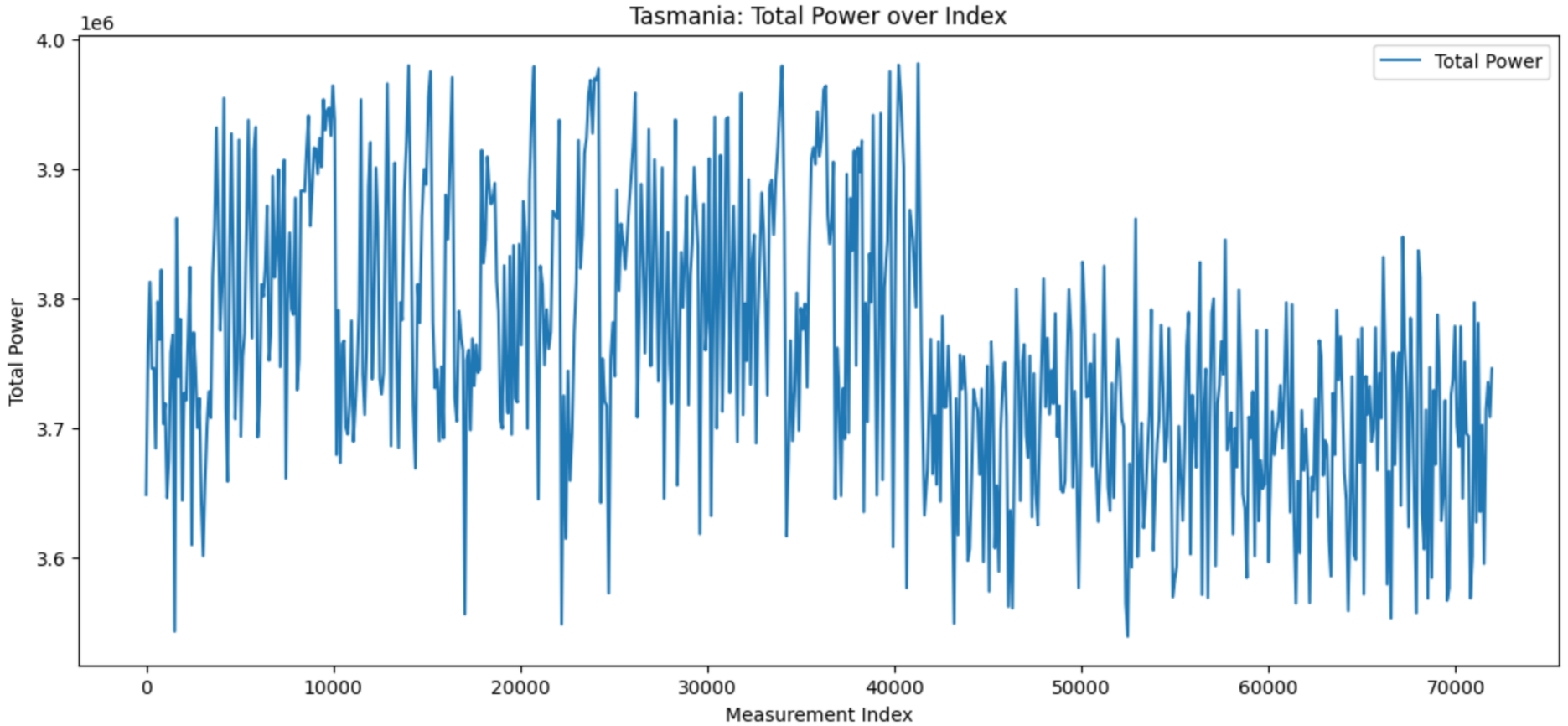


Outliers: restricted



...

Outliers: removed

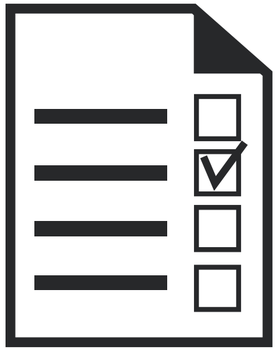


Missing Values

What are Missing Values?

Missing values are defined as the values or data that are not stored for some variable/s in the given dataset.

What might they cause?



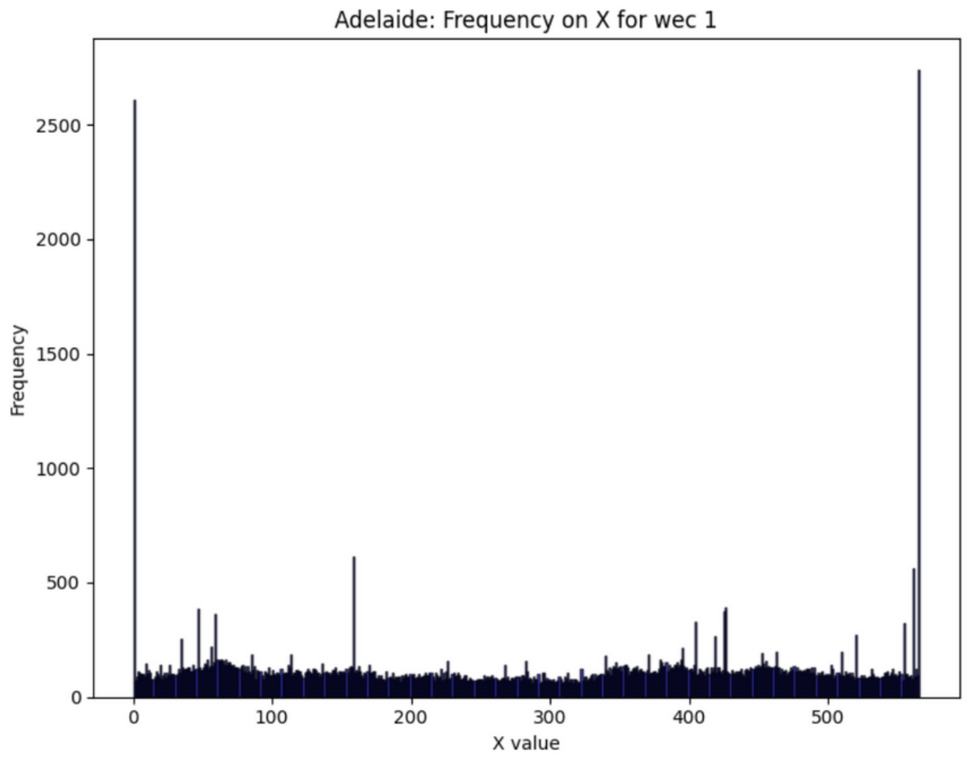
- reduce the accuracy of the model
- machine learning algorithms fail if the dataset contains missing values
- worse prediction precision

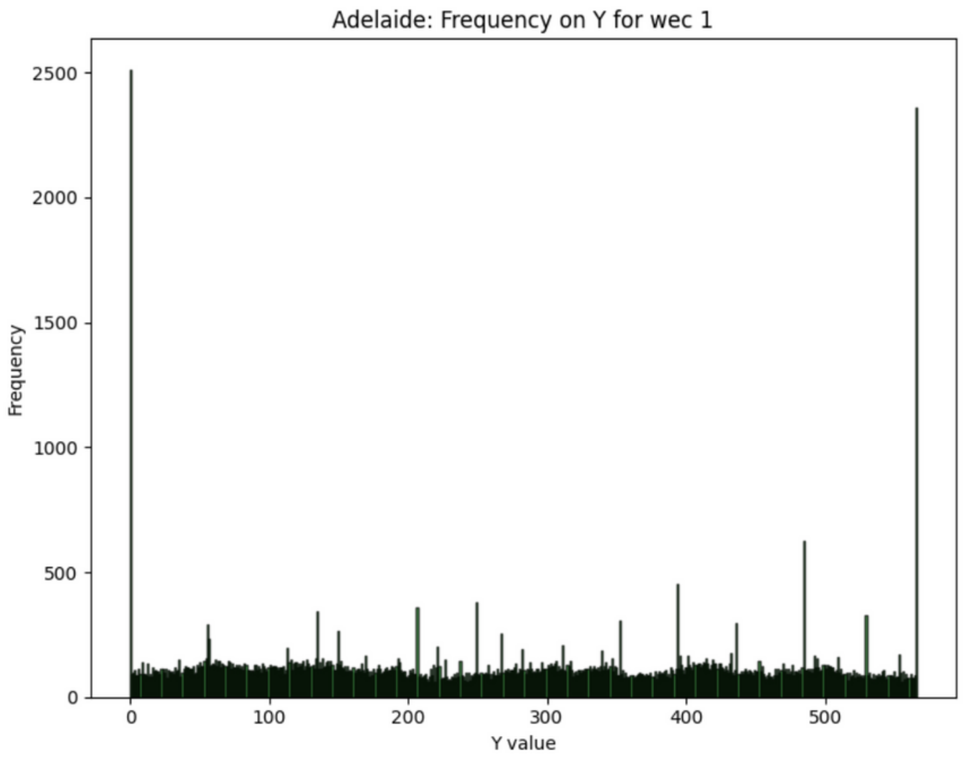
How can we handle them?

- deleting the Missing Values
- imputing the Missing Values

No missing values found in the dataset!

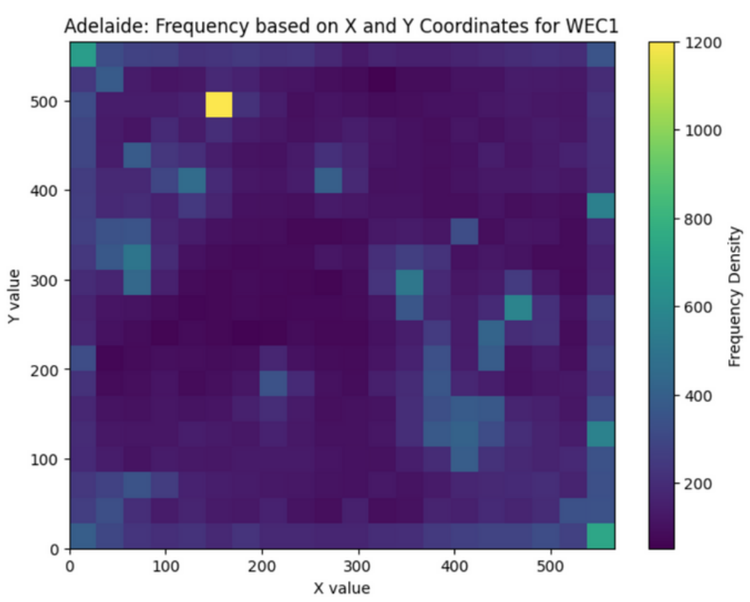
Extreme Input Values

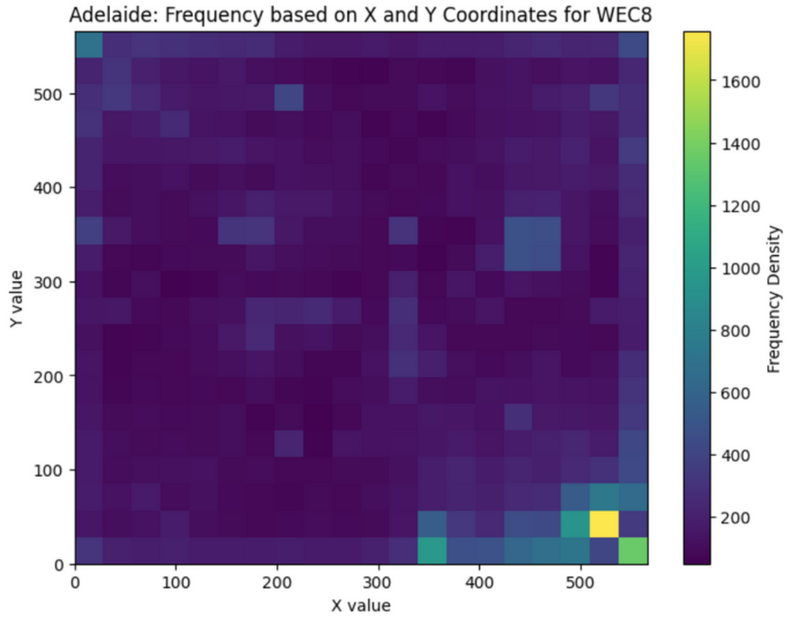


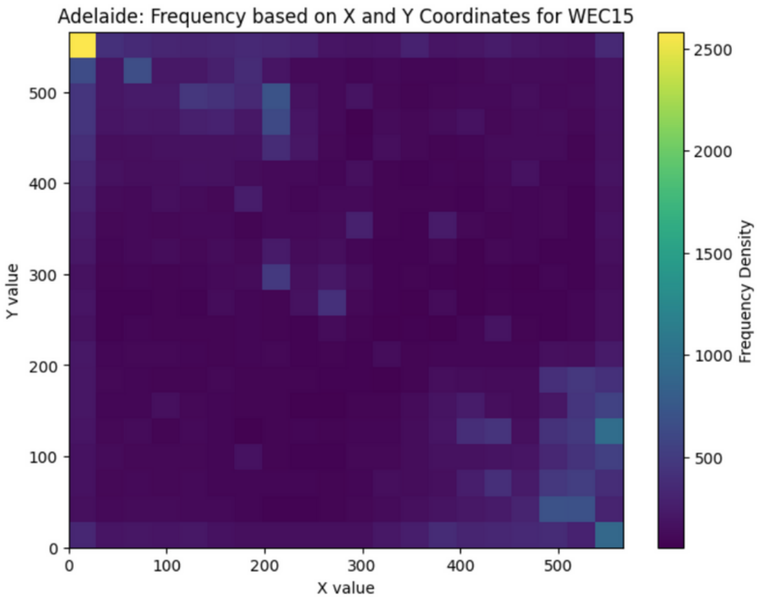


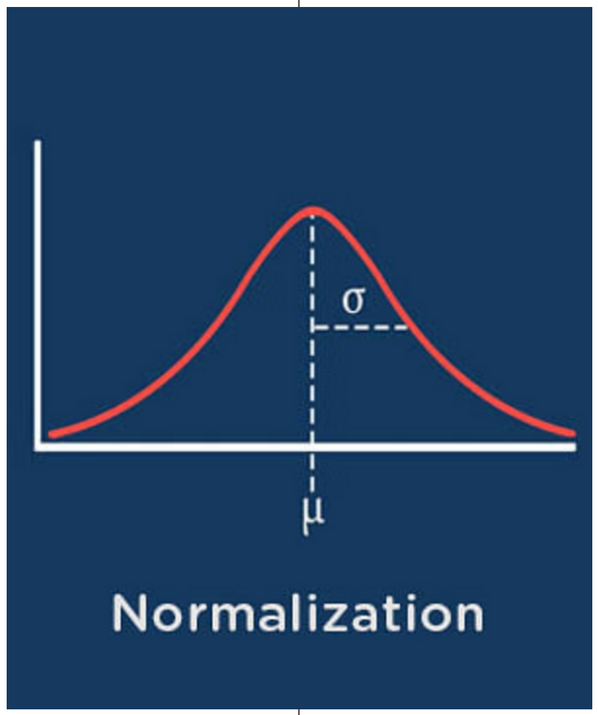
...

Extreme Input Values









Normalization

What is normalization?

Adjusting values measured on different scales to a notionally common scale

Why do we do that?

- ensures **optimal performance** of machine learning algorithms
- the **convergence speed** during model training is **enhanced** - learning processes
- **consistent interpretation** and comparison of features, allowing for a **clearer understanding** of their contributions to the overall model

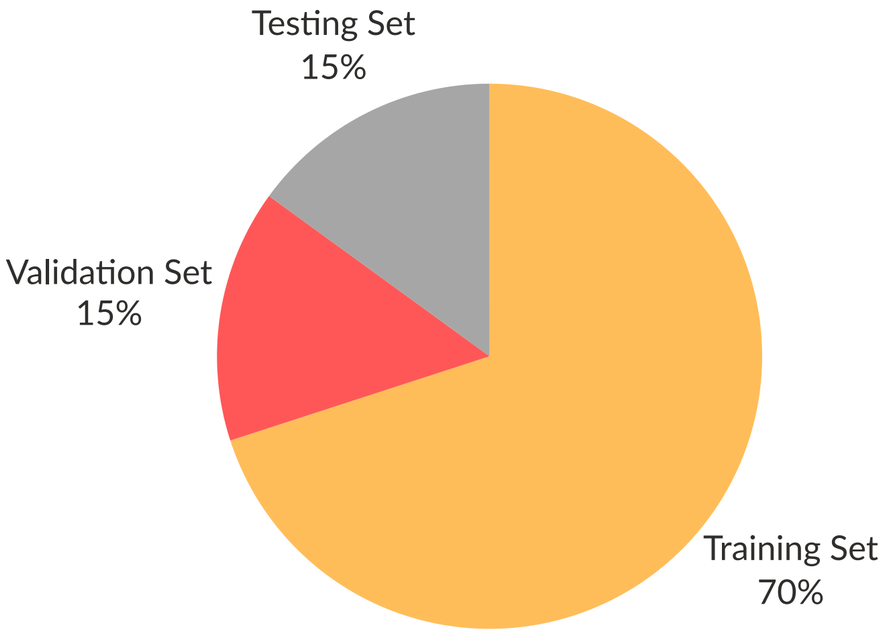
Normalized data

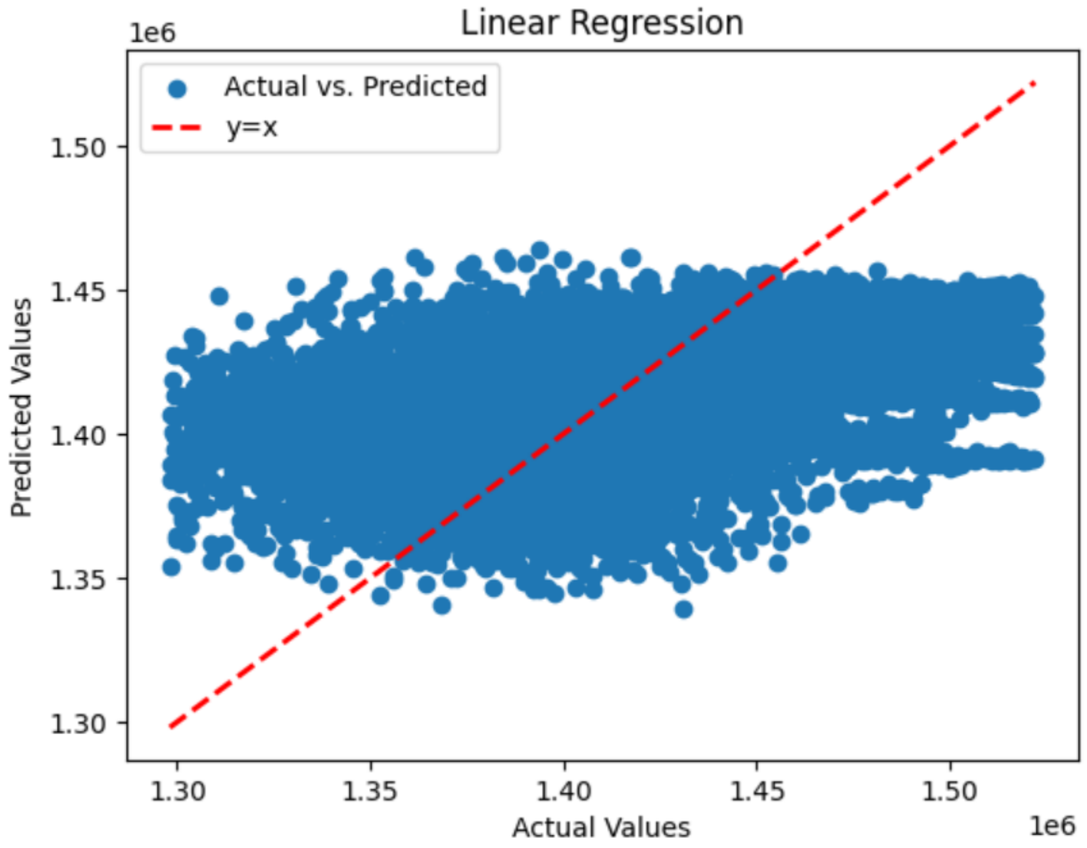
ower_all_normalized
0.322739
0.191277
0.240844
0.440520
0.257261
...
0.755442
0.805866
0.818474
0.871022
0.872522

Models

Data Split

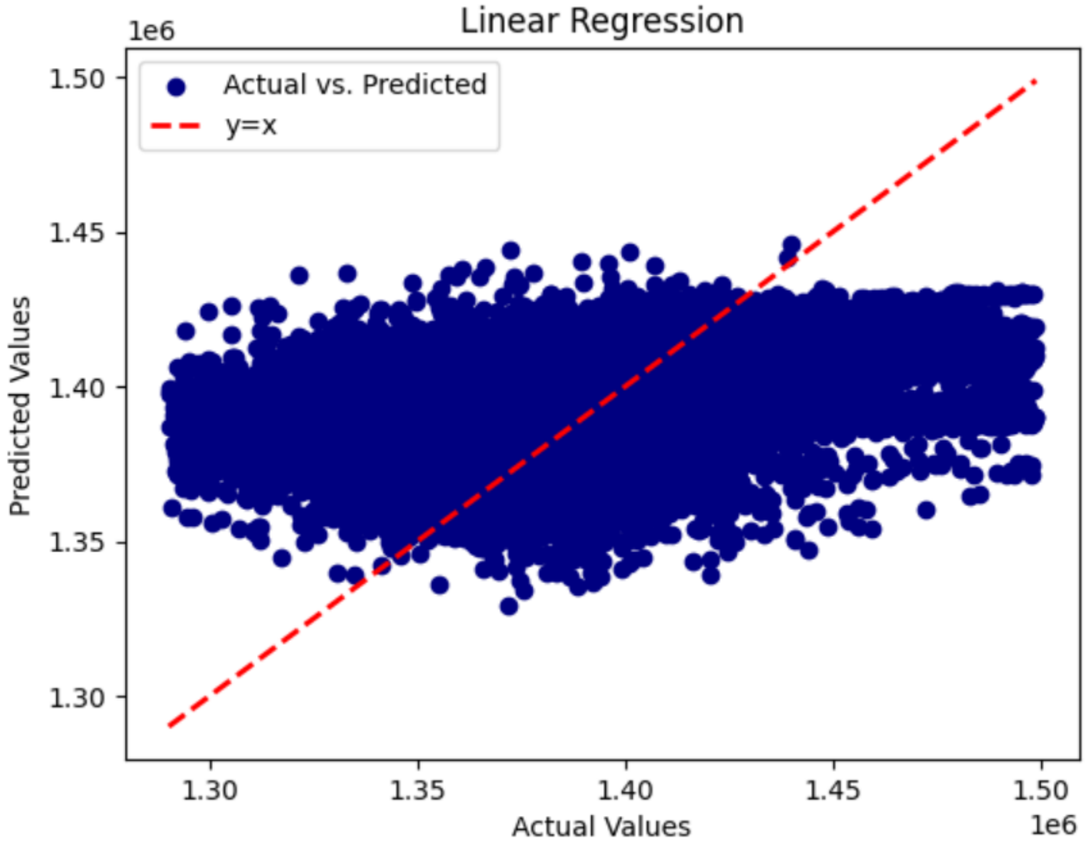
Training and Validation Sets





Linear Regression

Adelaide



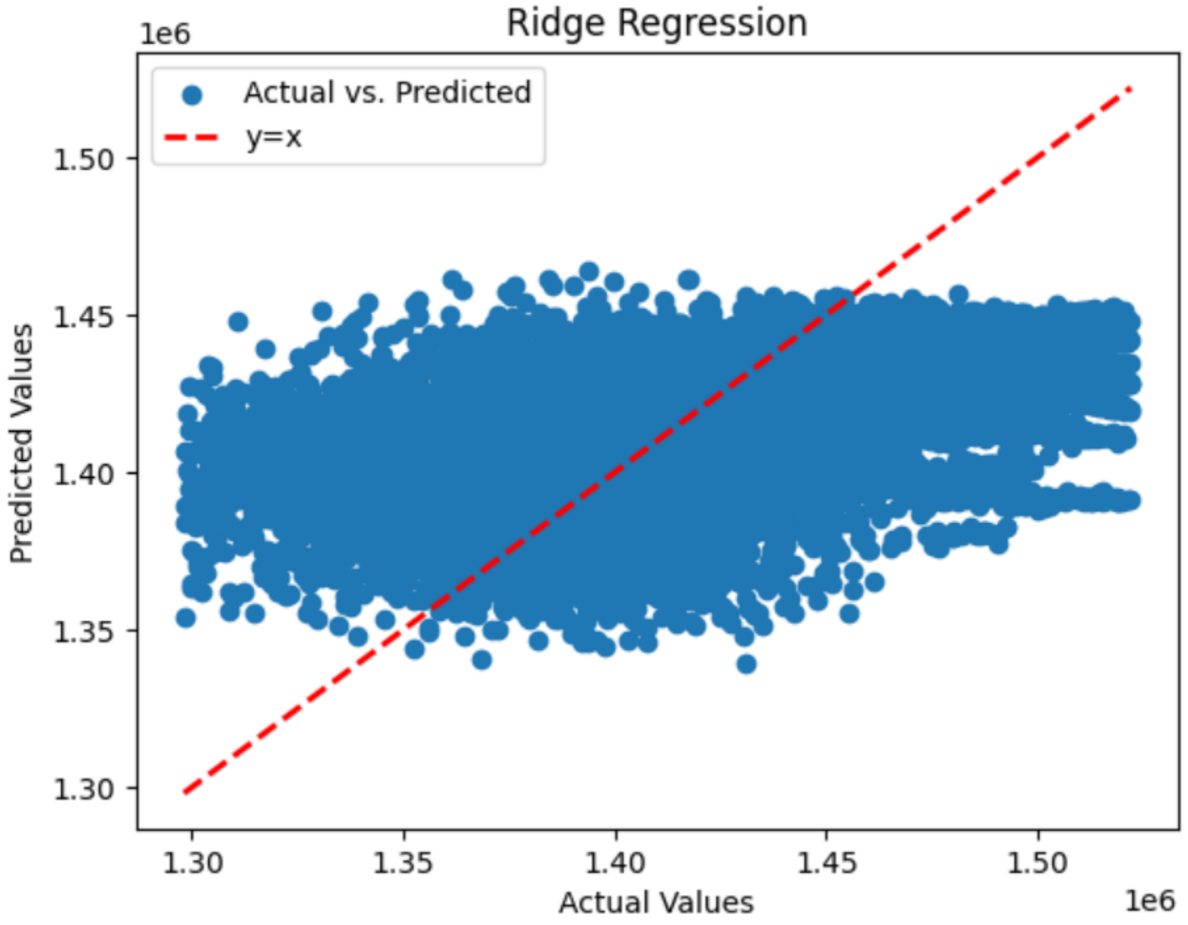
Linear Regression

Perth

...

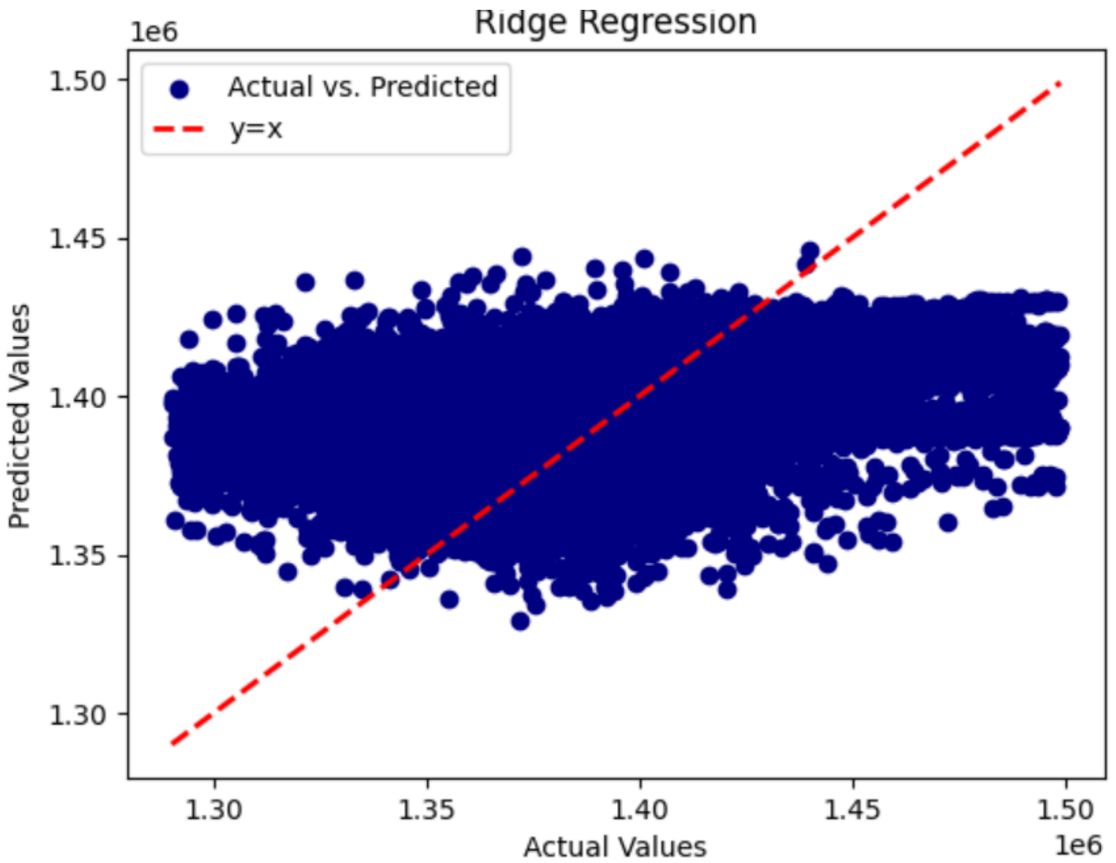
Ridge Regression

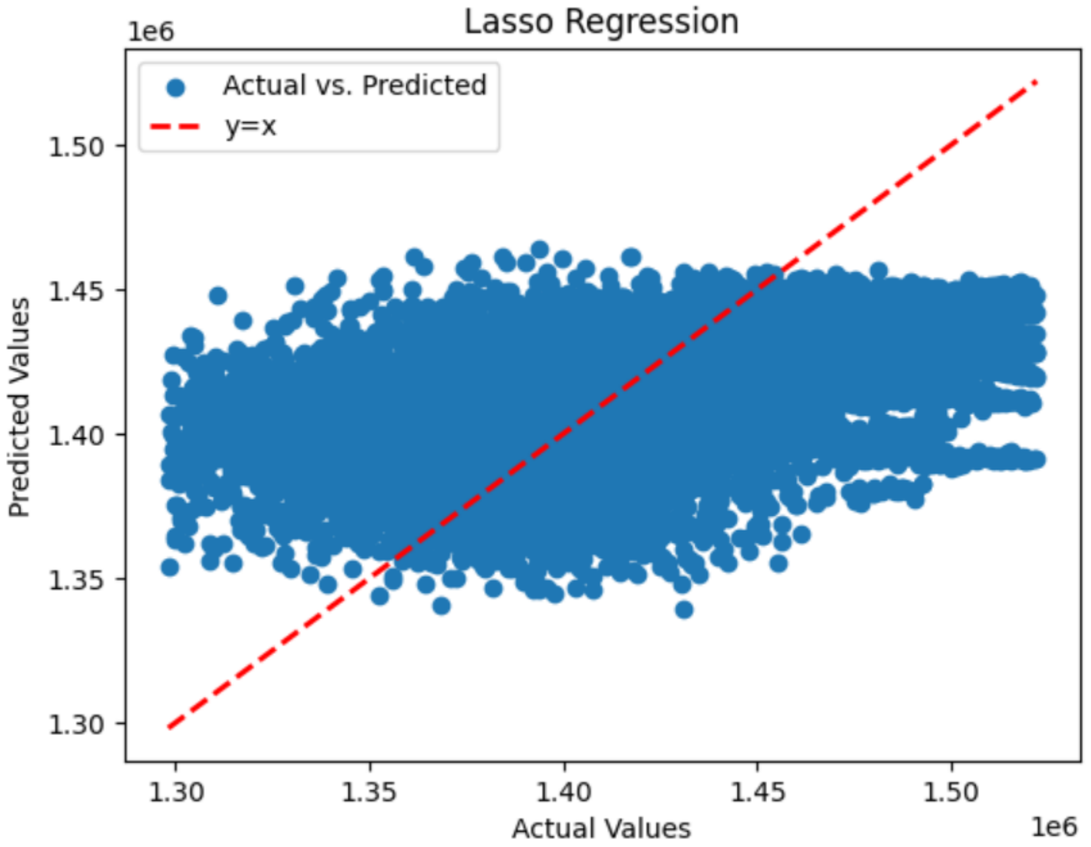
Adelaide



Ridge Regression

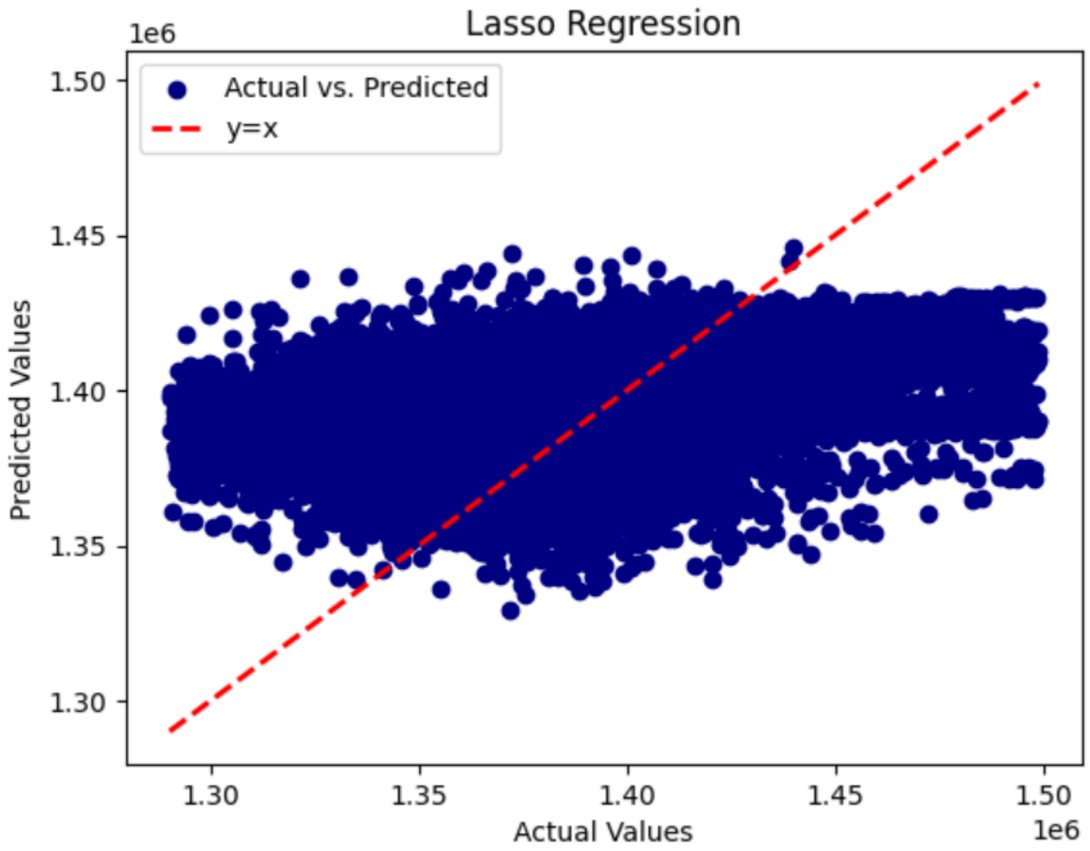
Perth





Lasso Regression

Adelaide

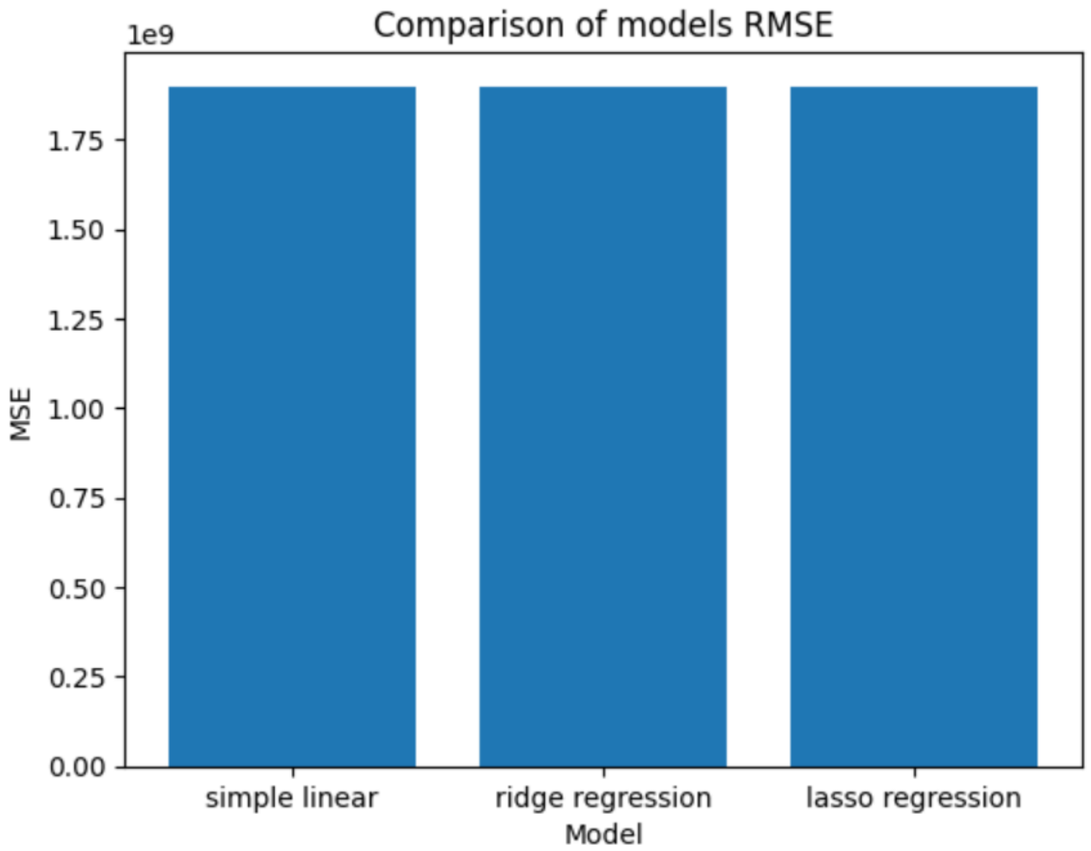


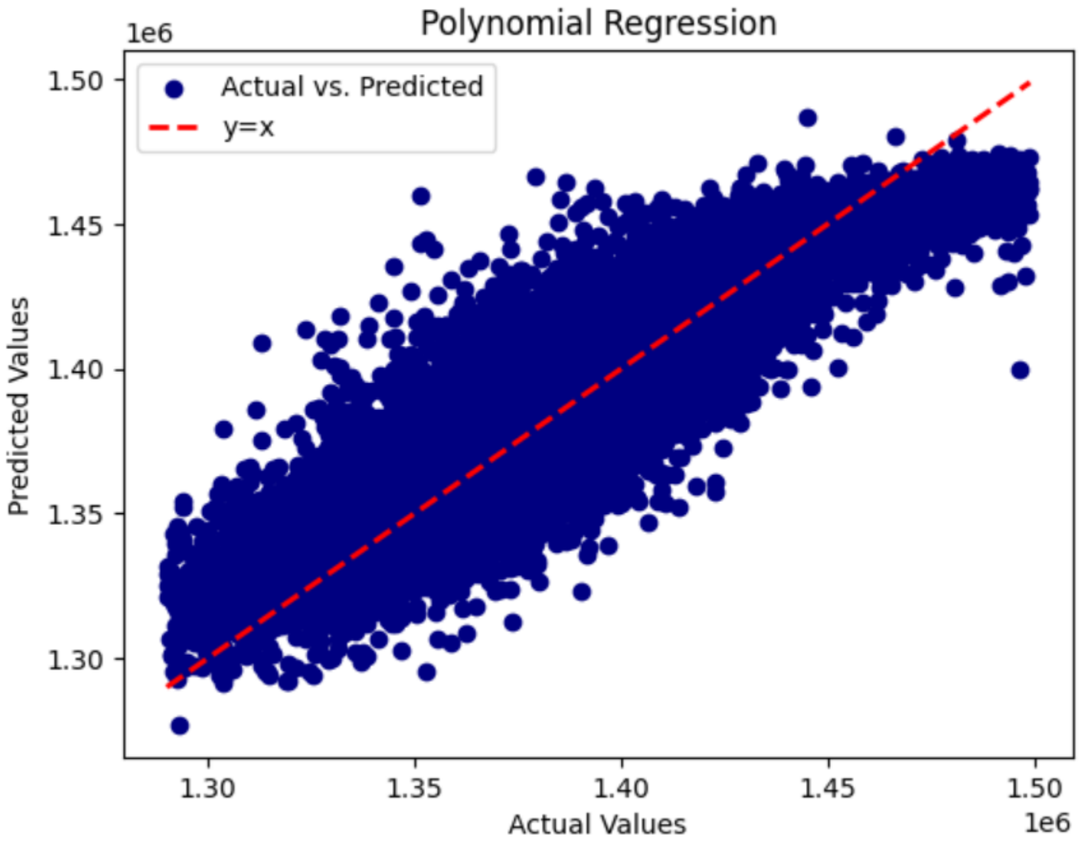
Lasso Regression

Perth

Comparasion

Perth

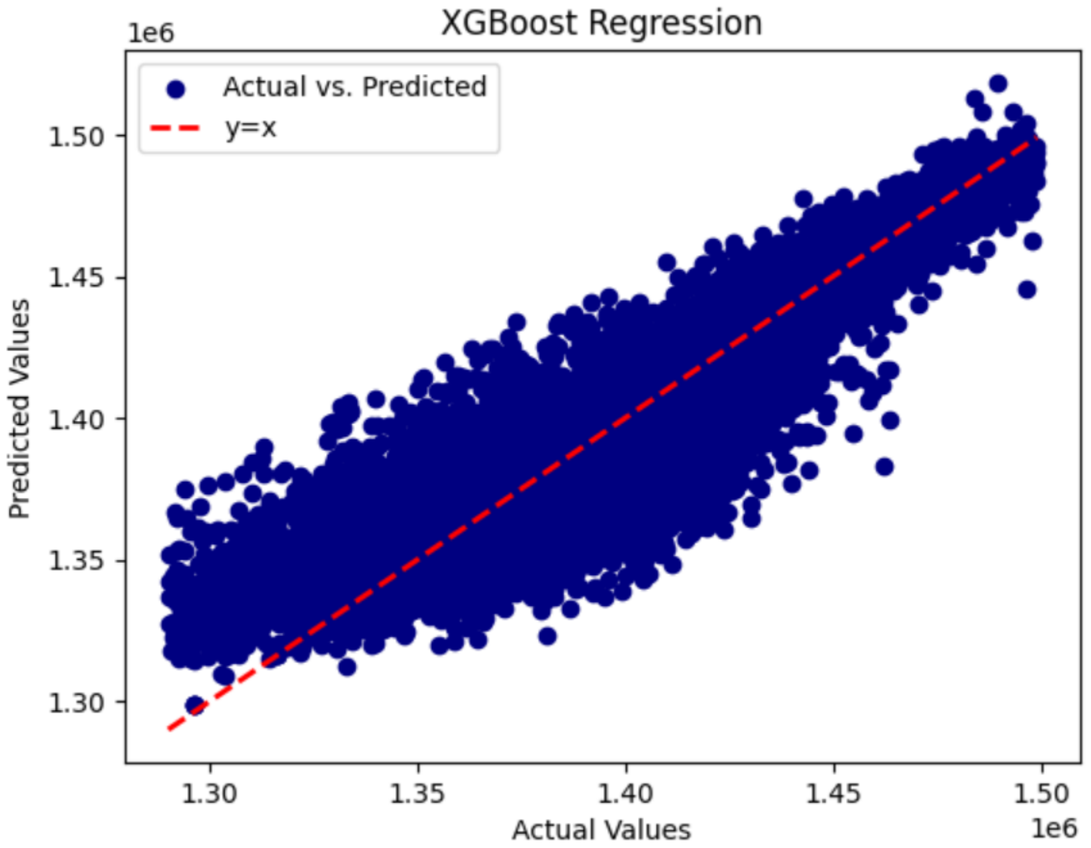


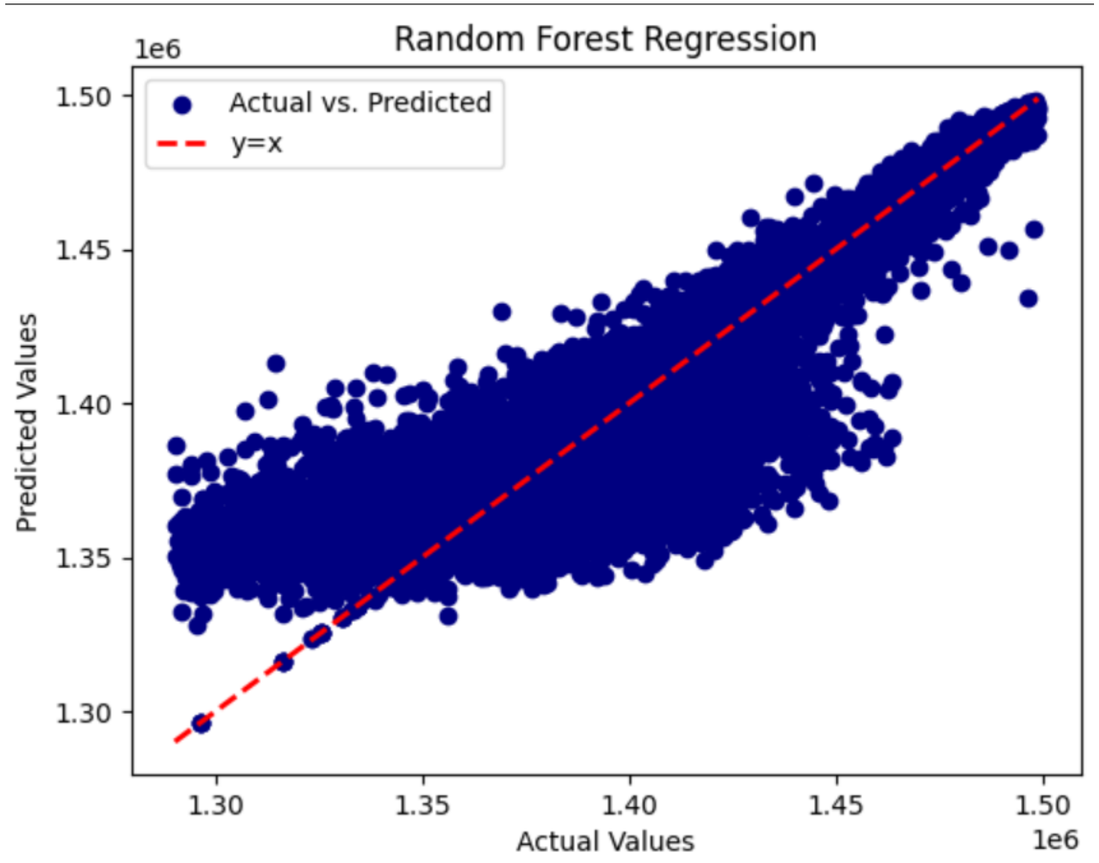


Polynomal Regression

degree: 2

XGBoost


XGBoost Regression
1e6
Actual vs. Predicted
y=x
1.50
1.45
1.40
1.35
1.30
Predicted Values
1.30
1.35
1.40
1.45
1.50
1e6
Actual Values

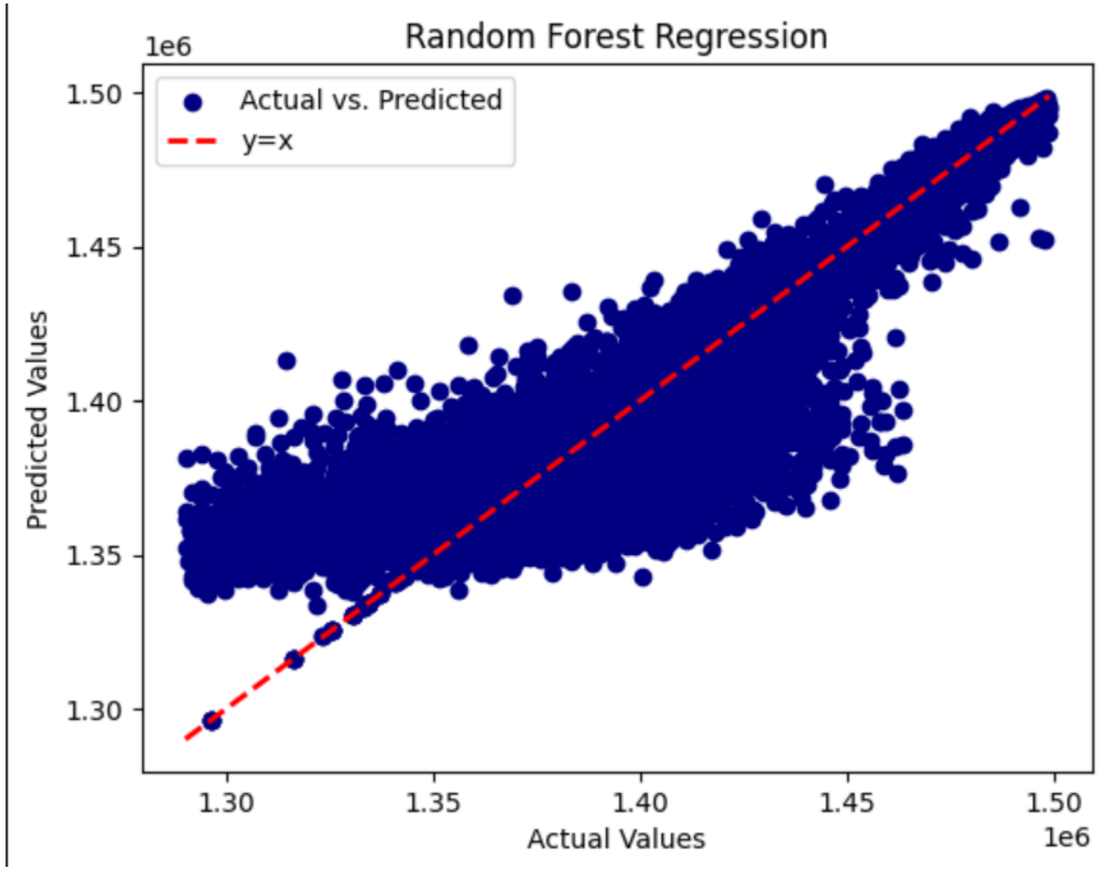


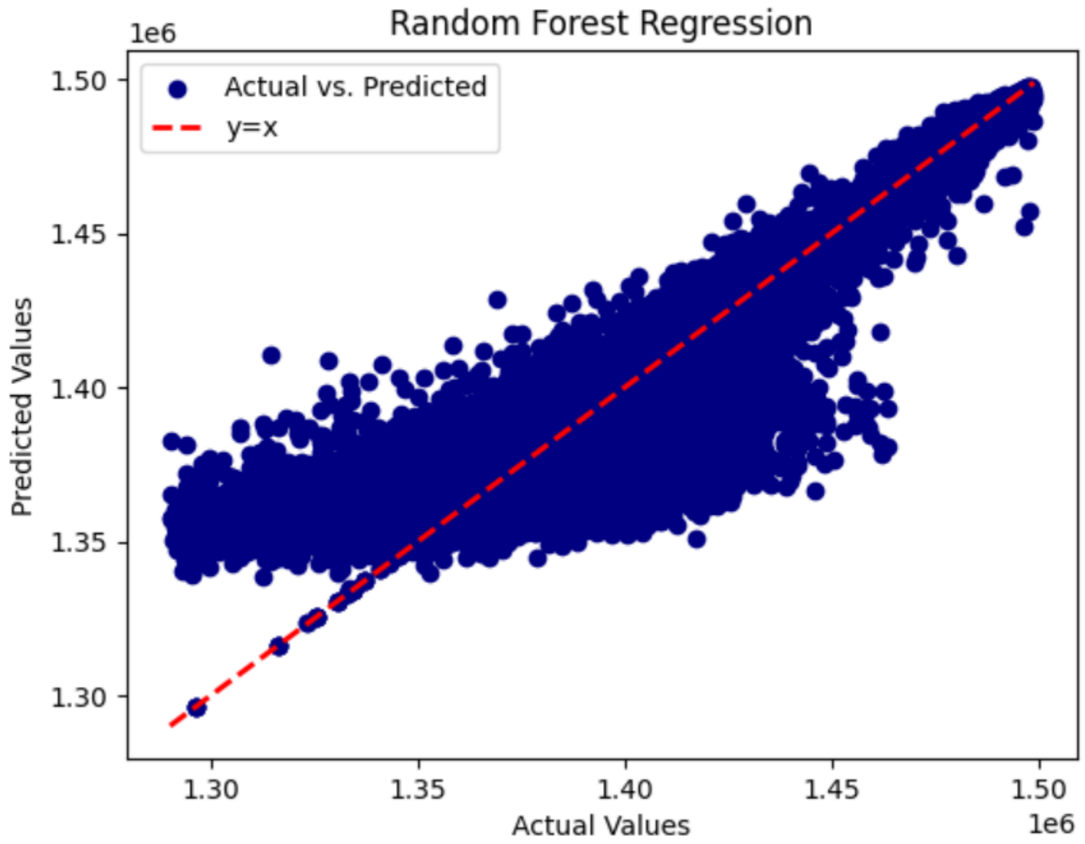
Random Forest

trees: 20

Random Forest

trees: 50



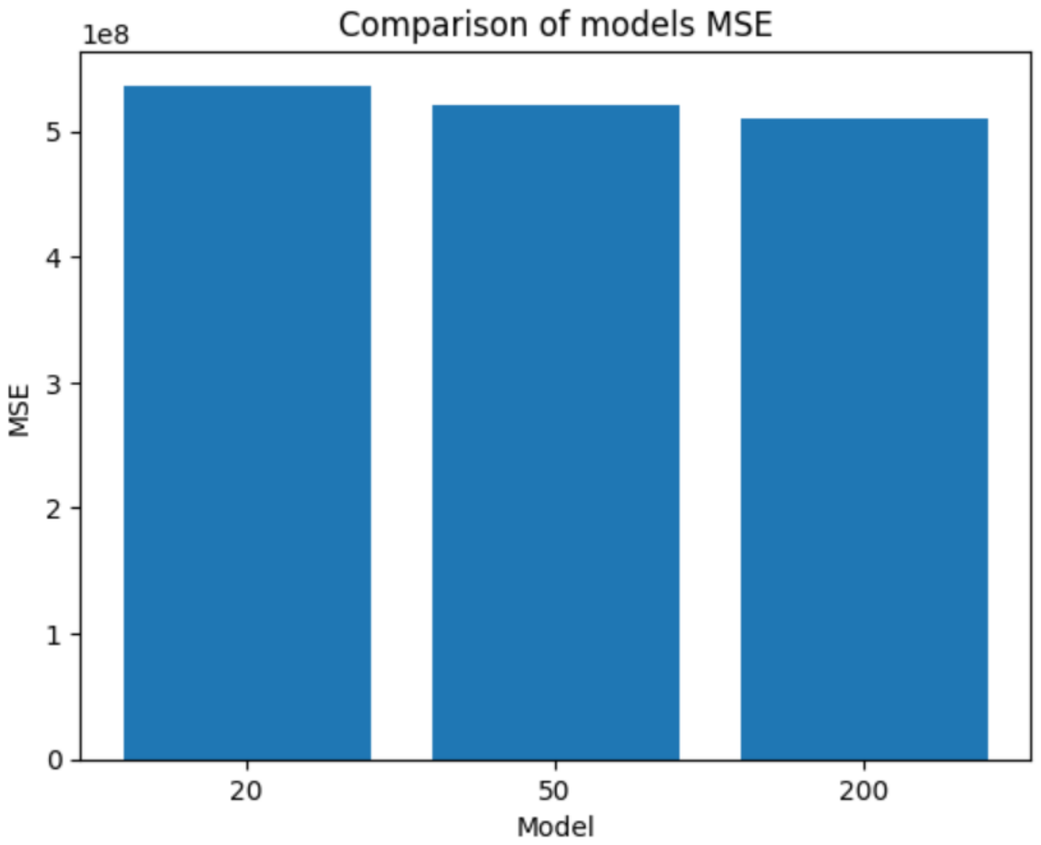


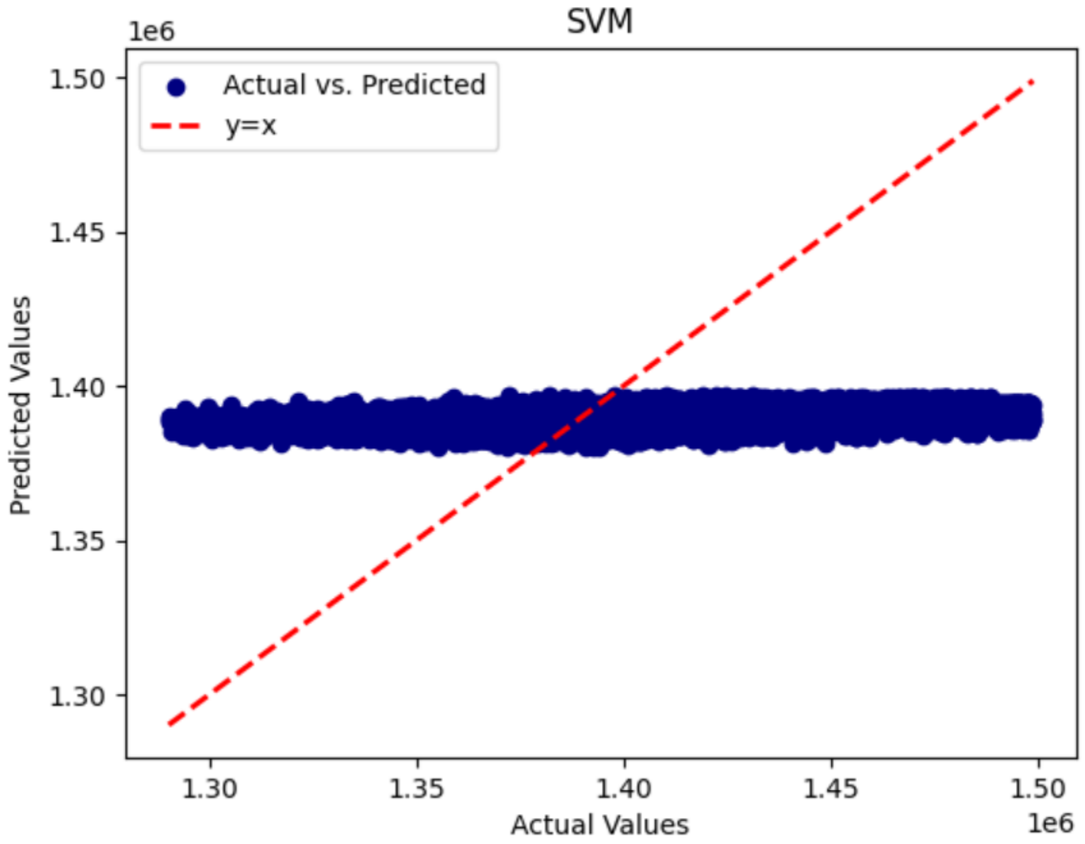
Random Forest

trees: 200

Random Forest Comparasion

Perth



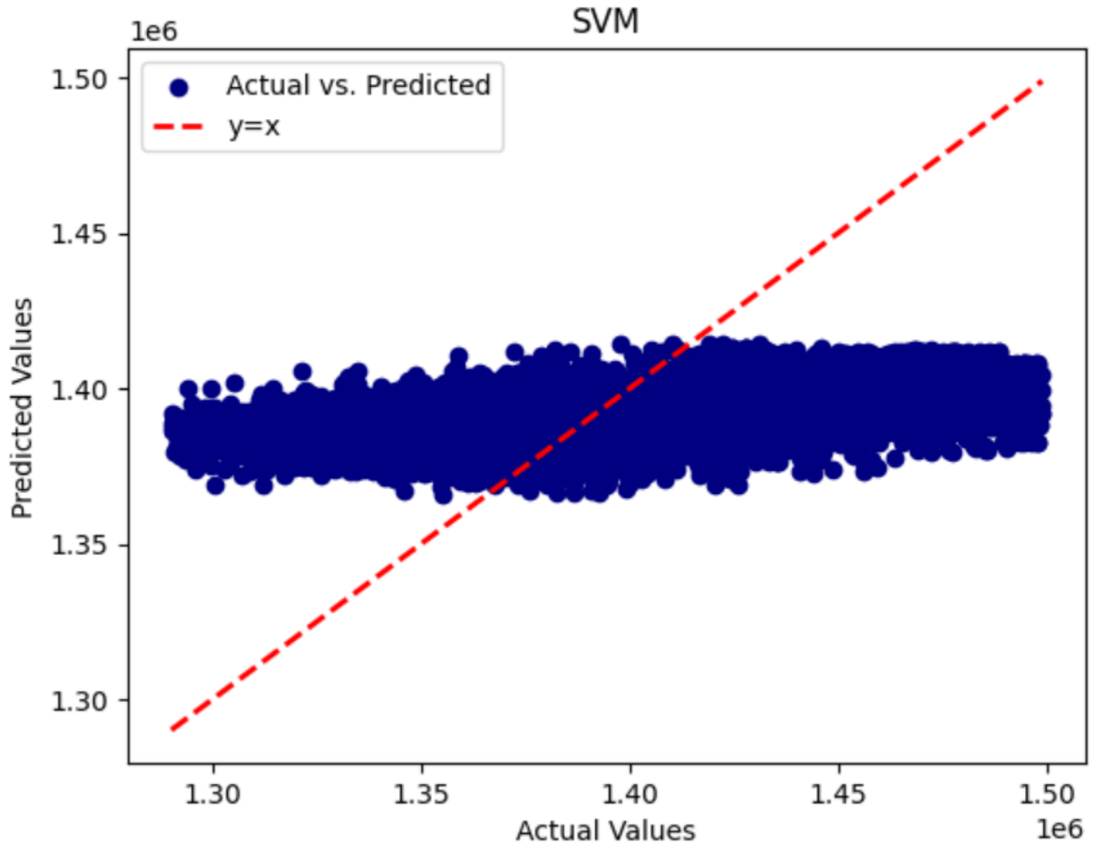


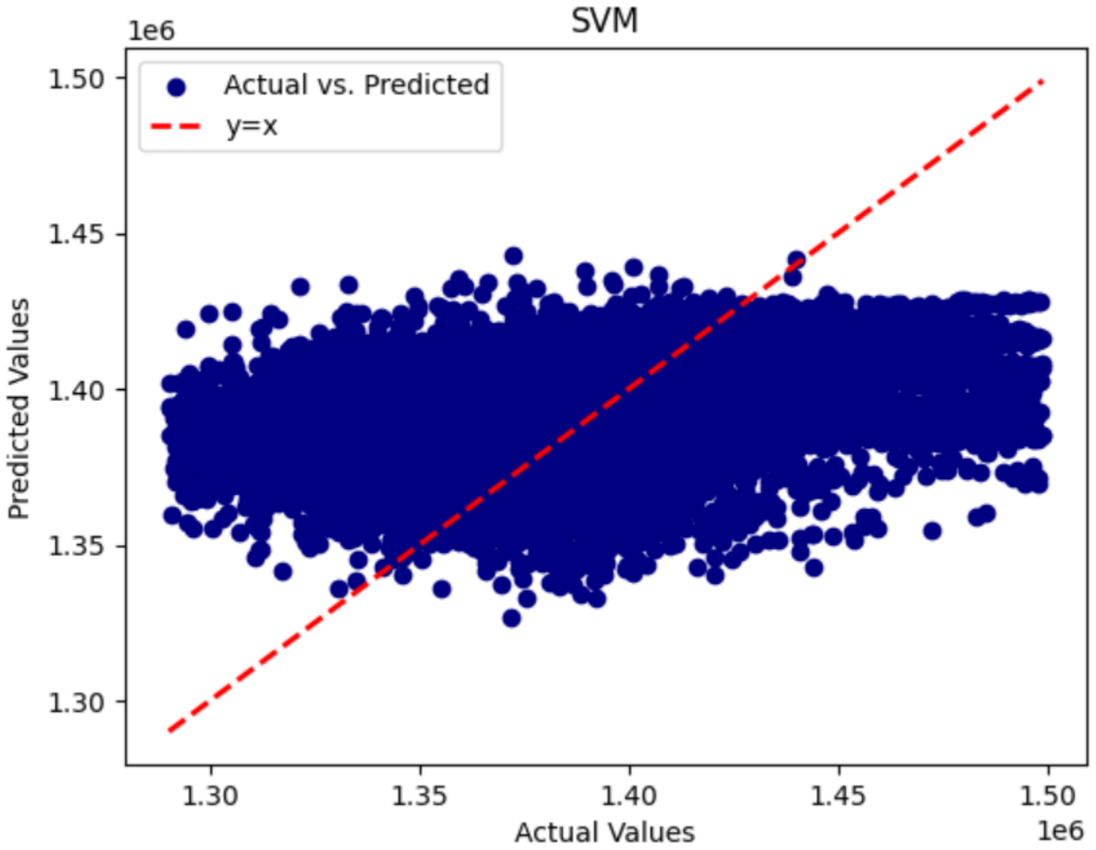
SVM

poly
degree: 2

SVM

poly
degree: 3



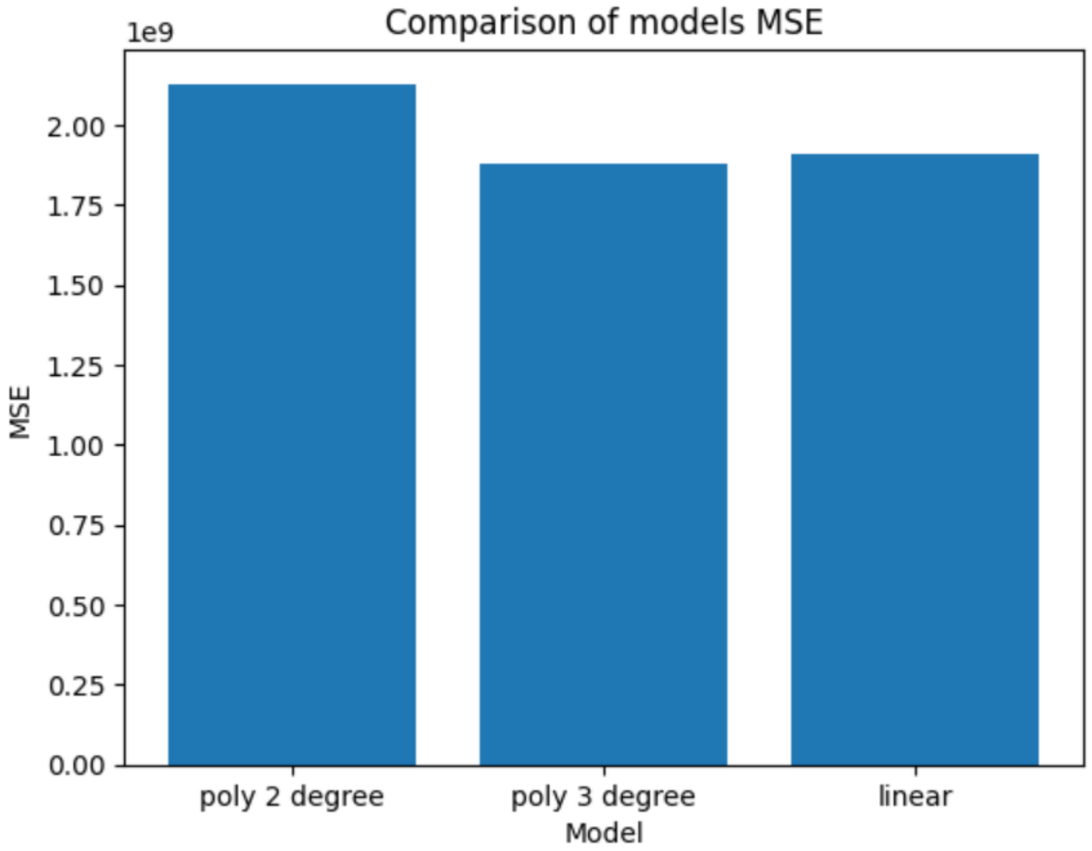

SVM
1e6
Actual vs. Predicted
y=x
1.50
1.45
1.40
1.35
1.30
Predicted Values
1.30
1.35
1.40
1.45
1.50
1e6
Actual Values

SVM

linear

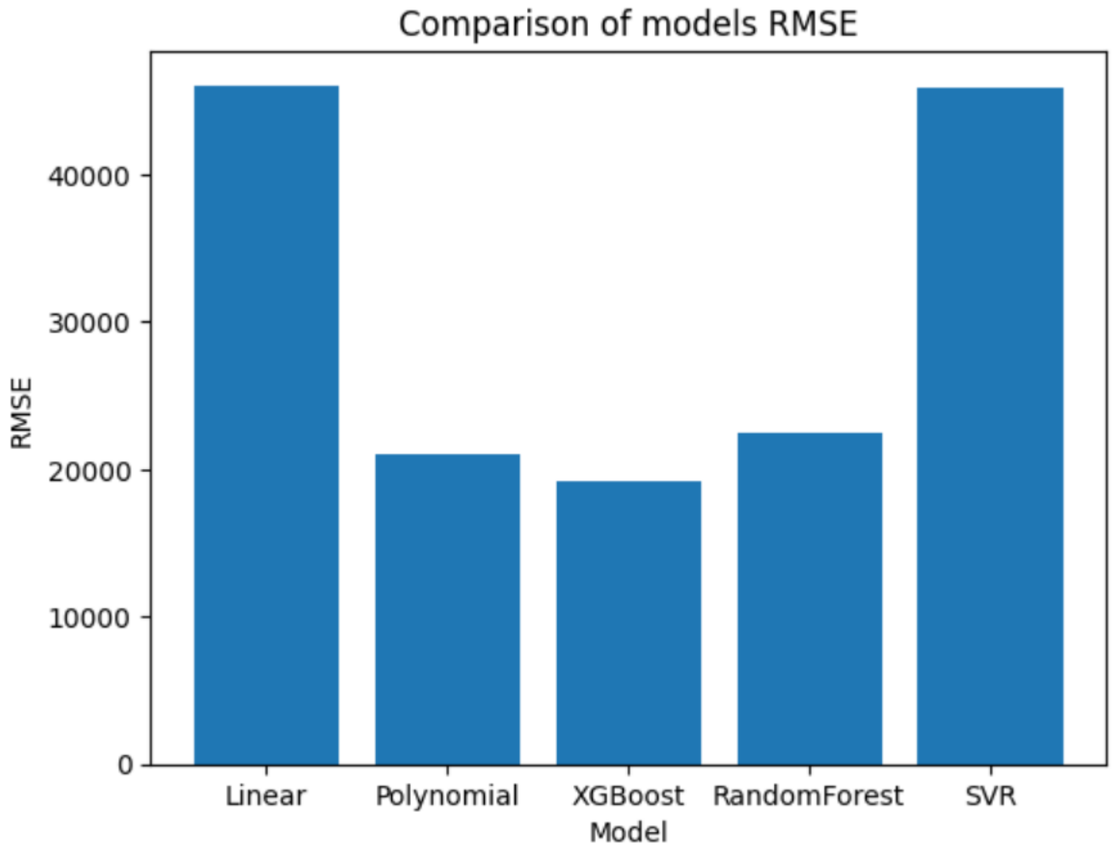
SVM Comparasion

Perth



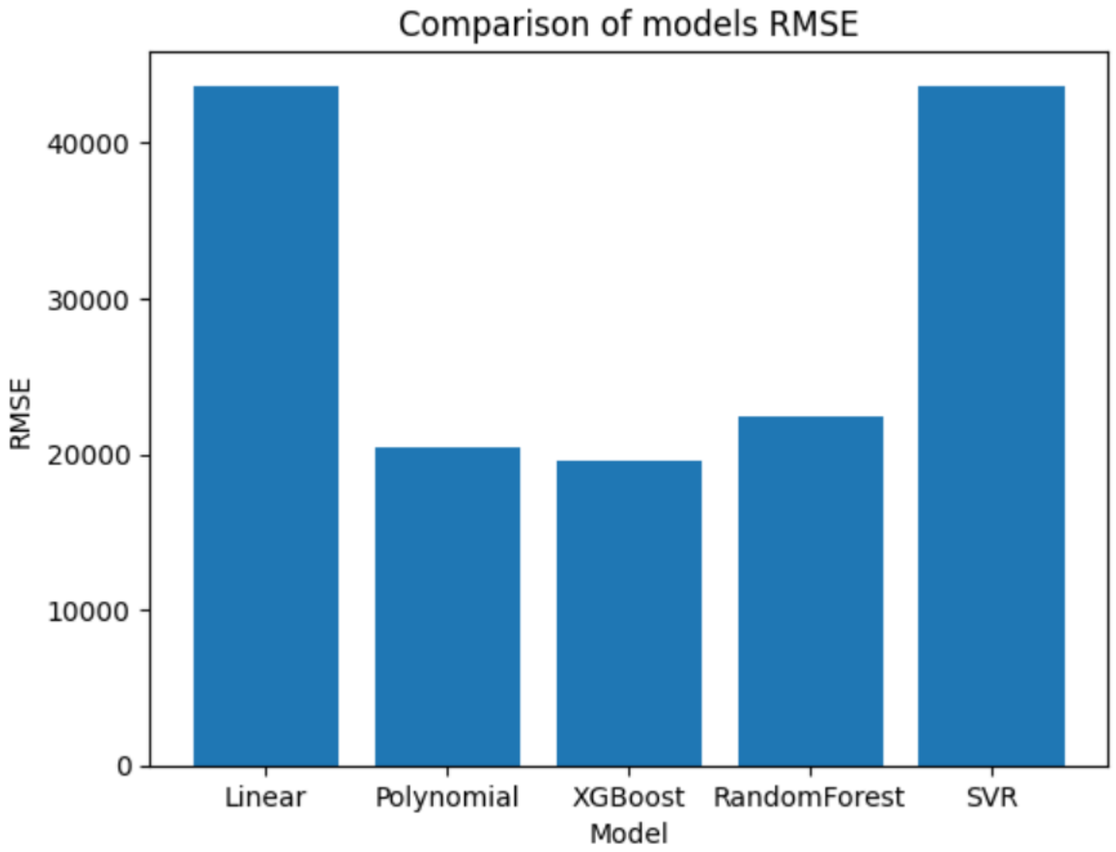
Testing models' performence

RMSE Comparasion



Adelaide

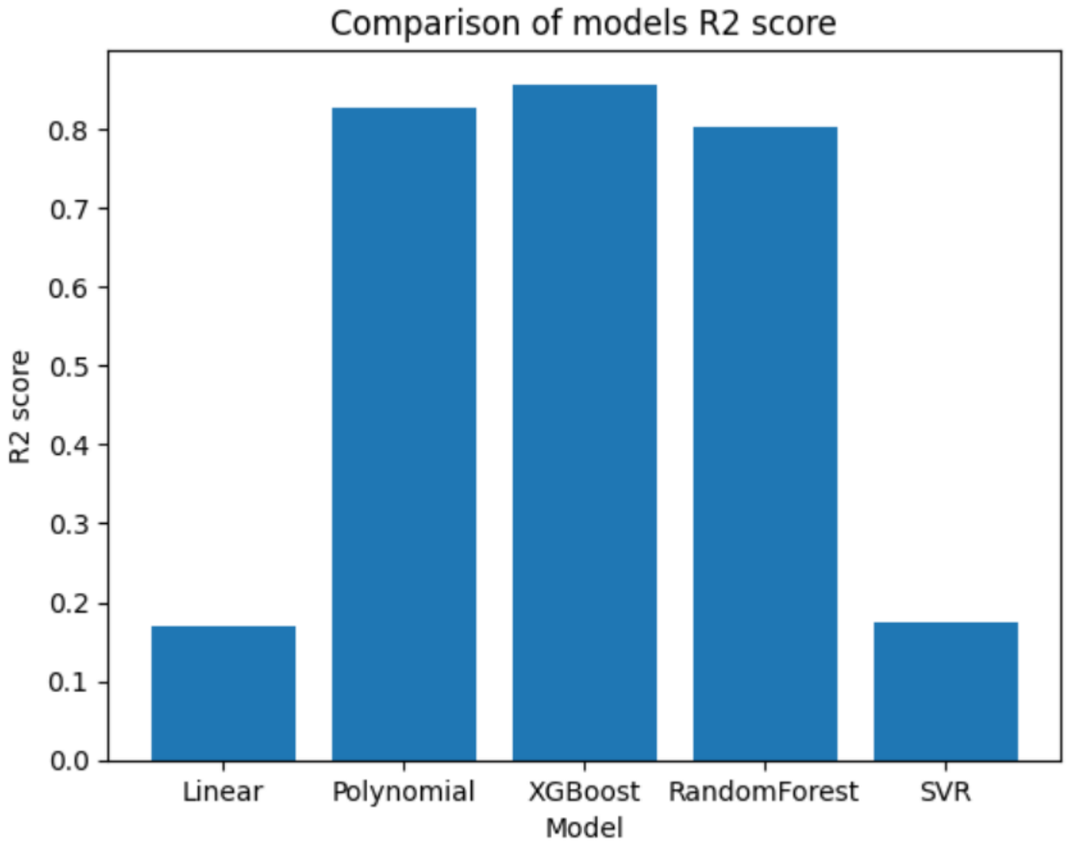
RMSE Comparasion



Perth

...

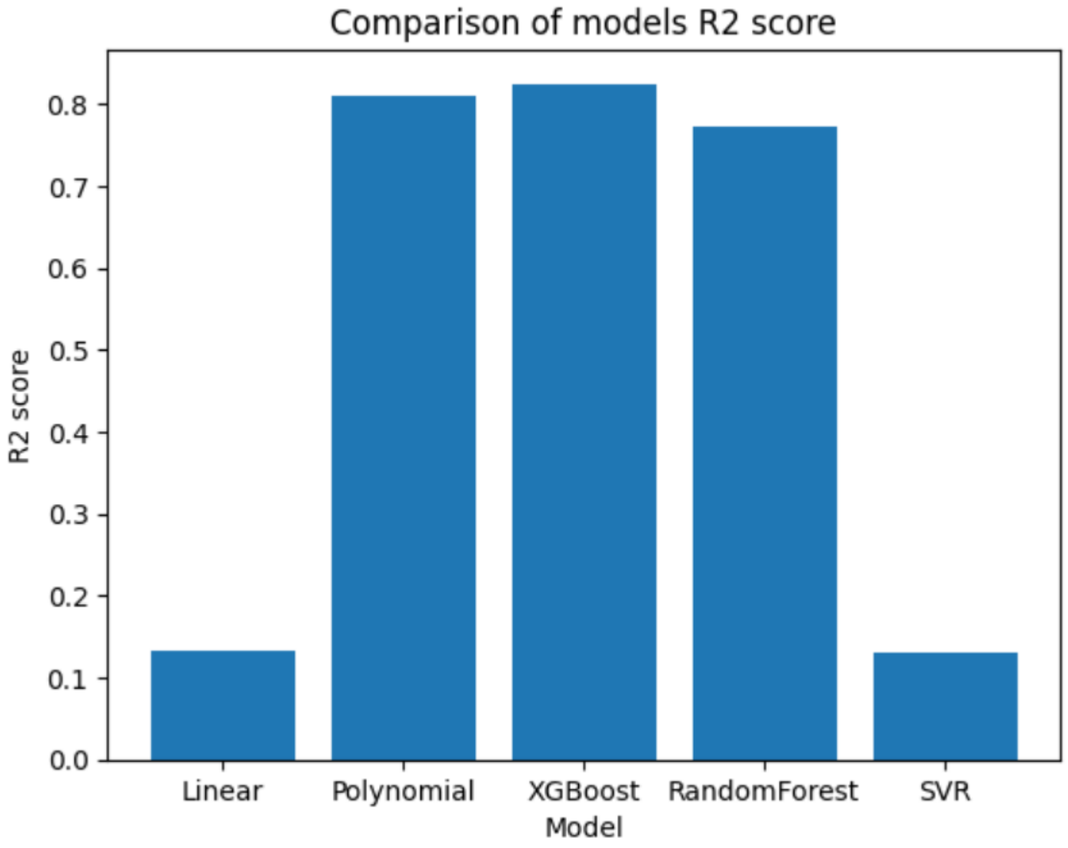
R2 Comparasion



Adelaide

...

R2 Comparasion



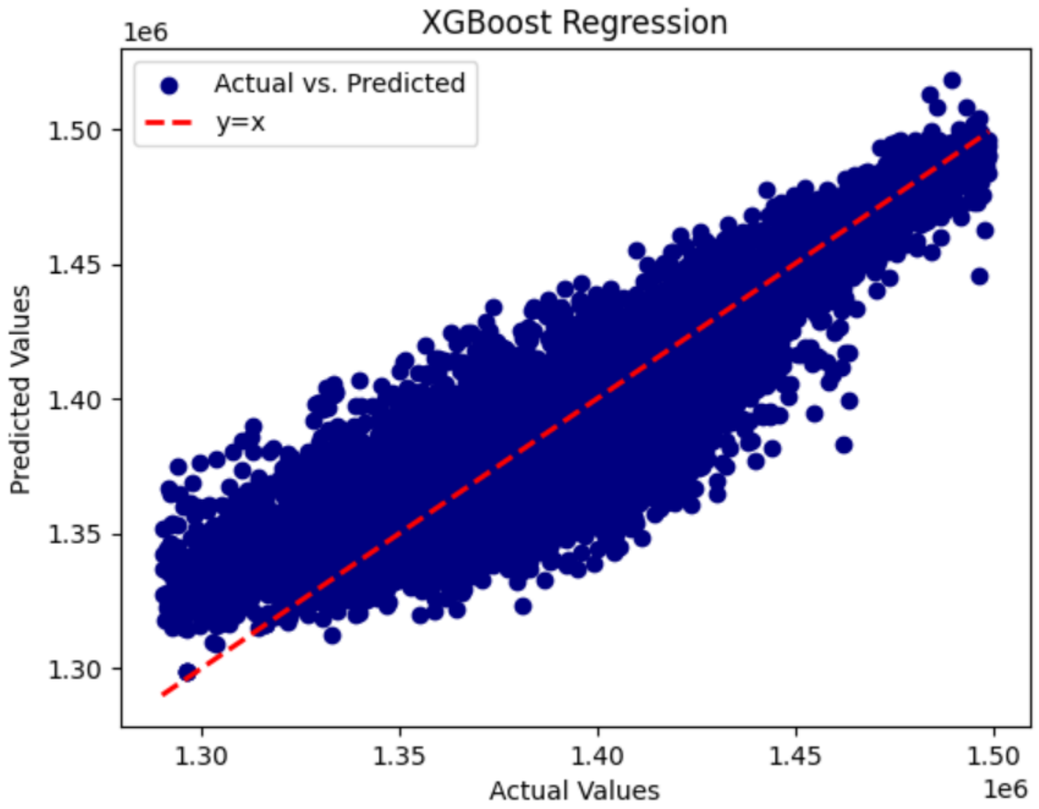
Perth

...

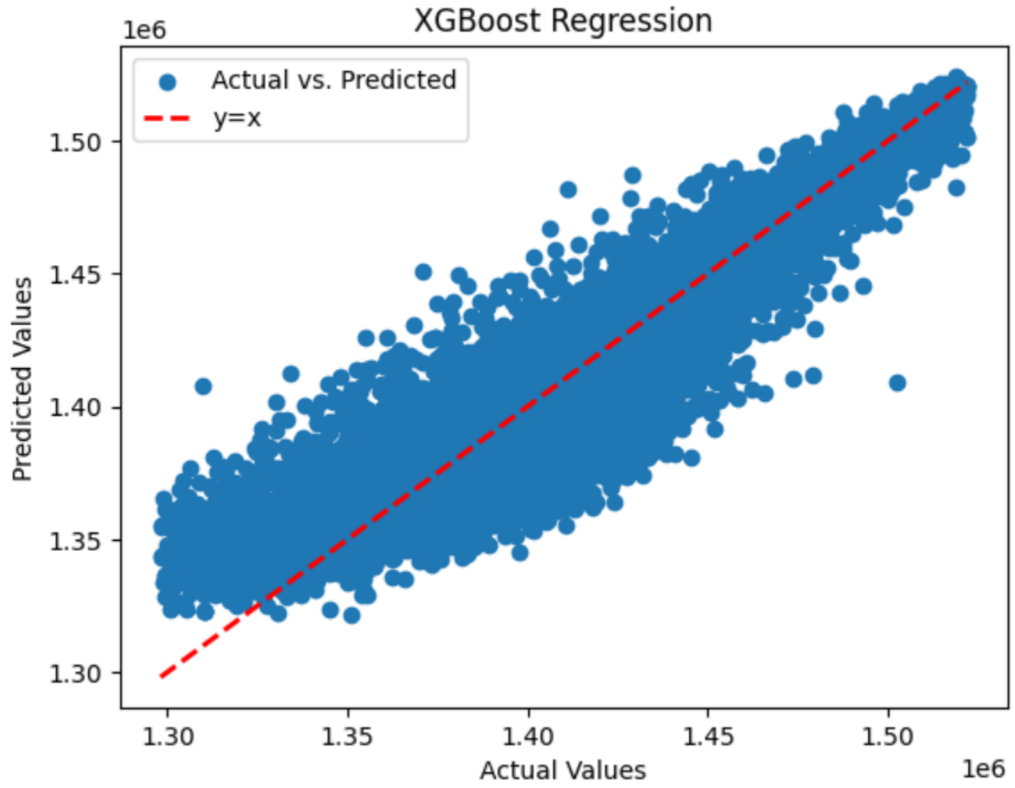
WINNER!

XGBoost

XGBoost

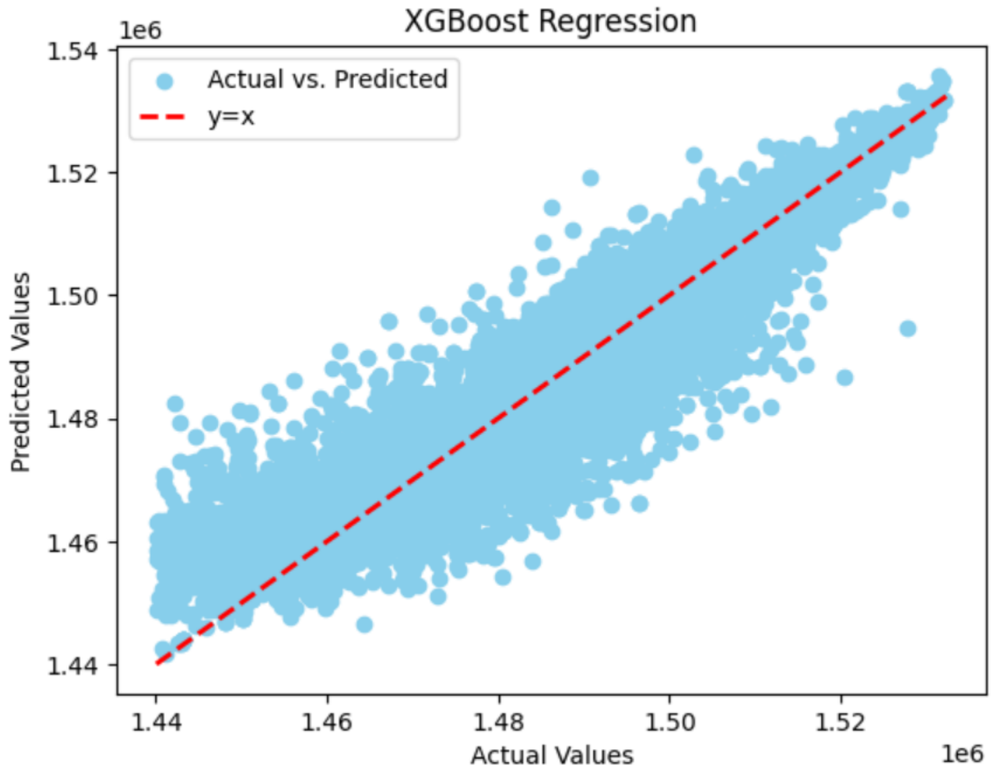


Perth

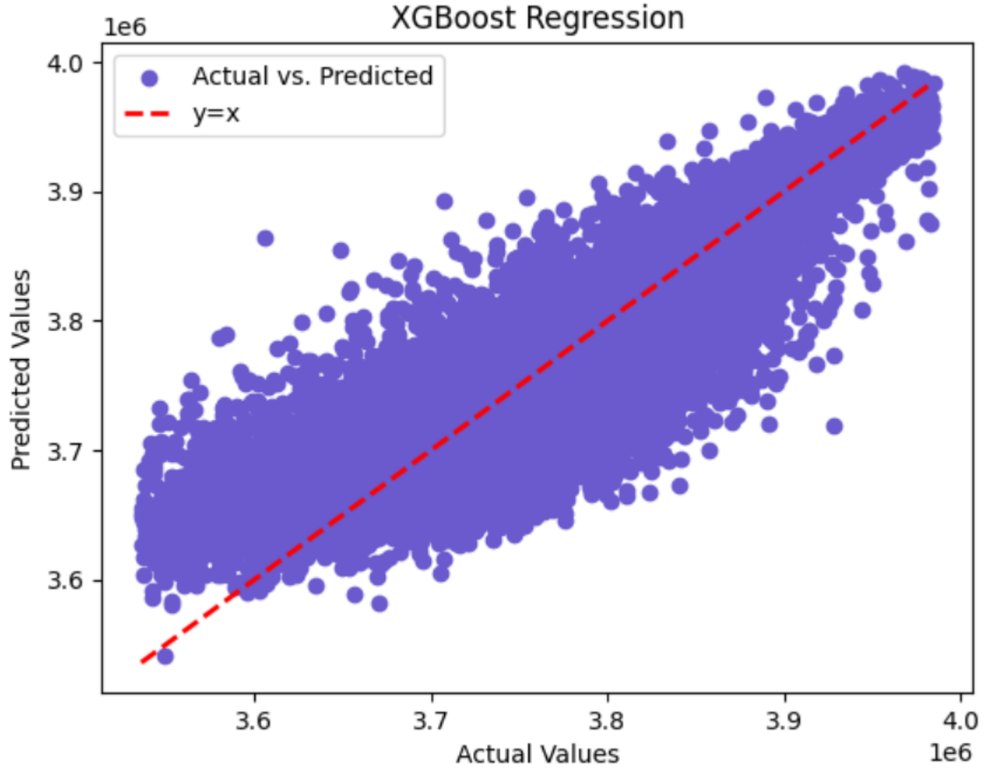


Adelaide

XGBoost



Sydney



Tasmania

RMSE
of the **final model**

(non-normalised data)

Sydney	7409.66
Tasmania	229350.58
Adelaide	75350.18
Perth	85165.58

RMSE
of the **final model**

(normalised data)

Sydney	0.2217364519244979 2
Tasmania	0.1779274769848658 6
Adelaide	0.167537921497995 5
Perth	0.0938330791674778 2

...

Final RMSE

Normalized	Non-nomalized
99 319	**0.165259**

Thank you

Nicola Szwaja
Piotr Droś