ANALYSIS AND PREDICTION USING THE WEC DATASET

Jakub Dziurka, Paweł Hermansdorfer

PROJECT OBJECTIVE

Main Goals:

- Analyze the Wave Energy Converter (WEC) dataset.
- Build a regression model to predict a real-valued target (e.g., power output).

Scope of Work:

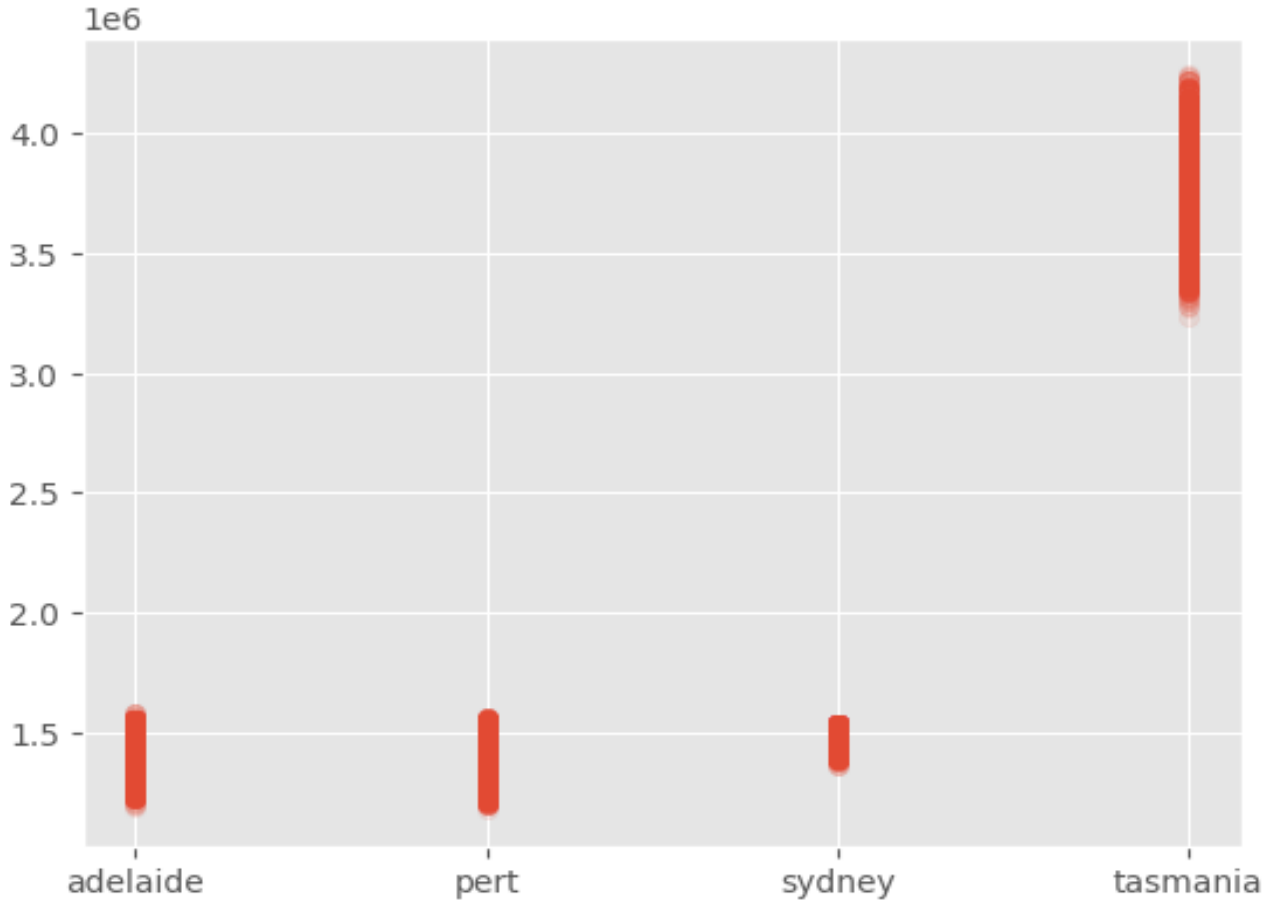
- Explore and clean the dataset.
- Train and compare machine learning models.
- Evaluate the best model and draw insights.

UNDERSTANDING THE DATASET

The dataset represents data collected from four regions: Adelaide, Perth, Sydney, and Tasmania.

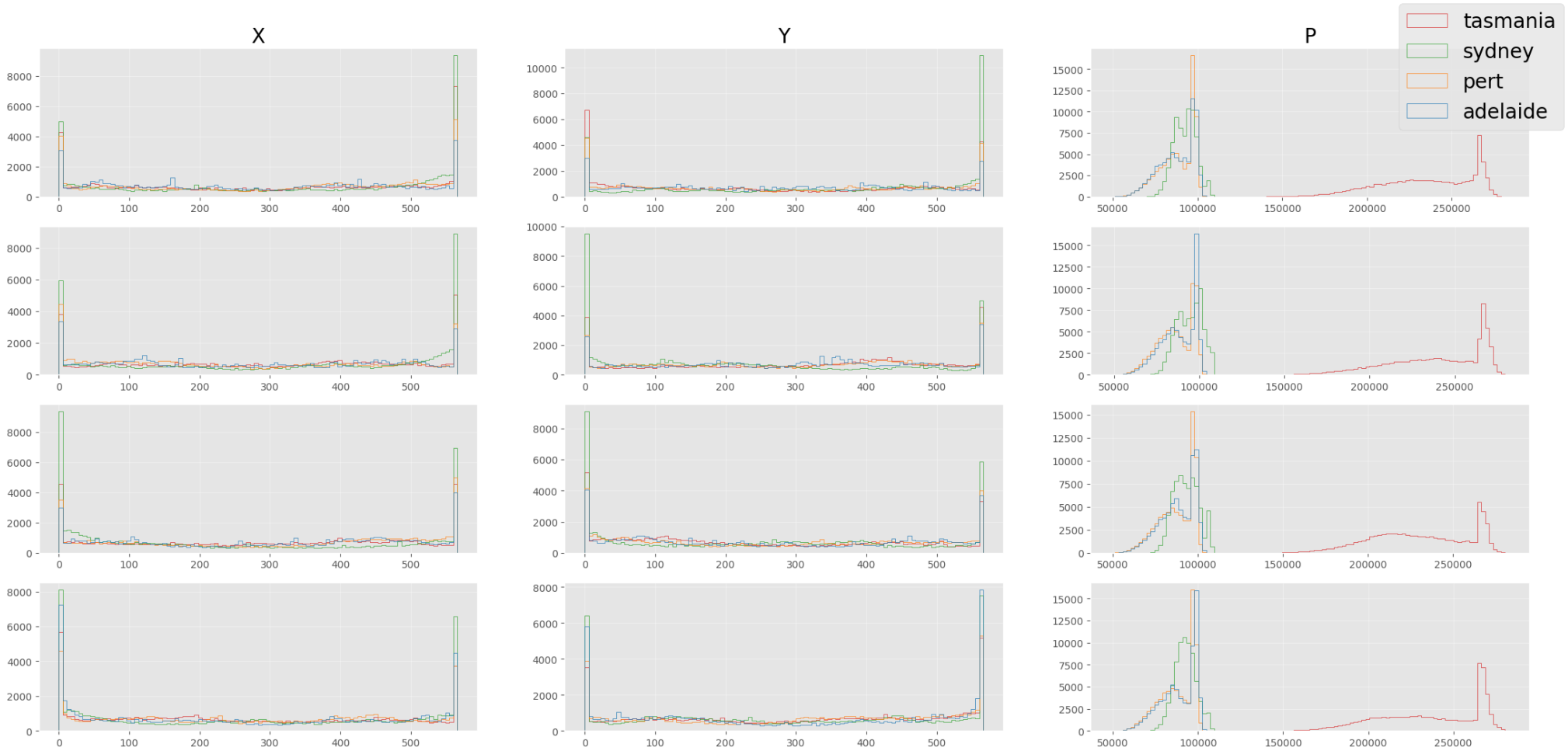
It contains information on:

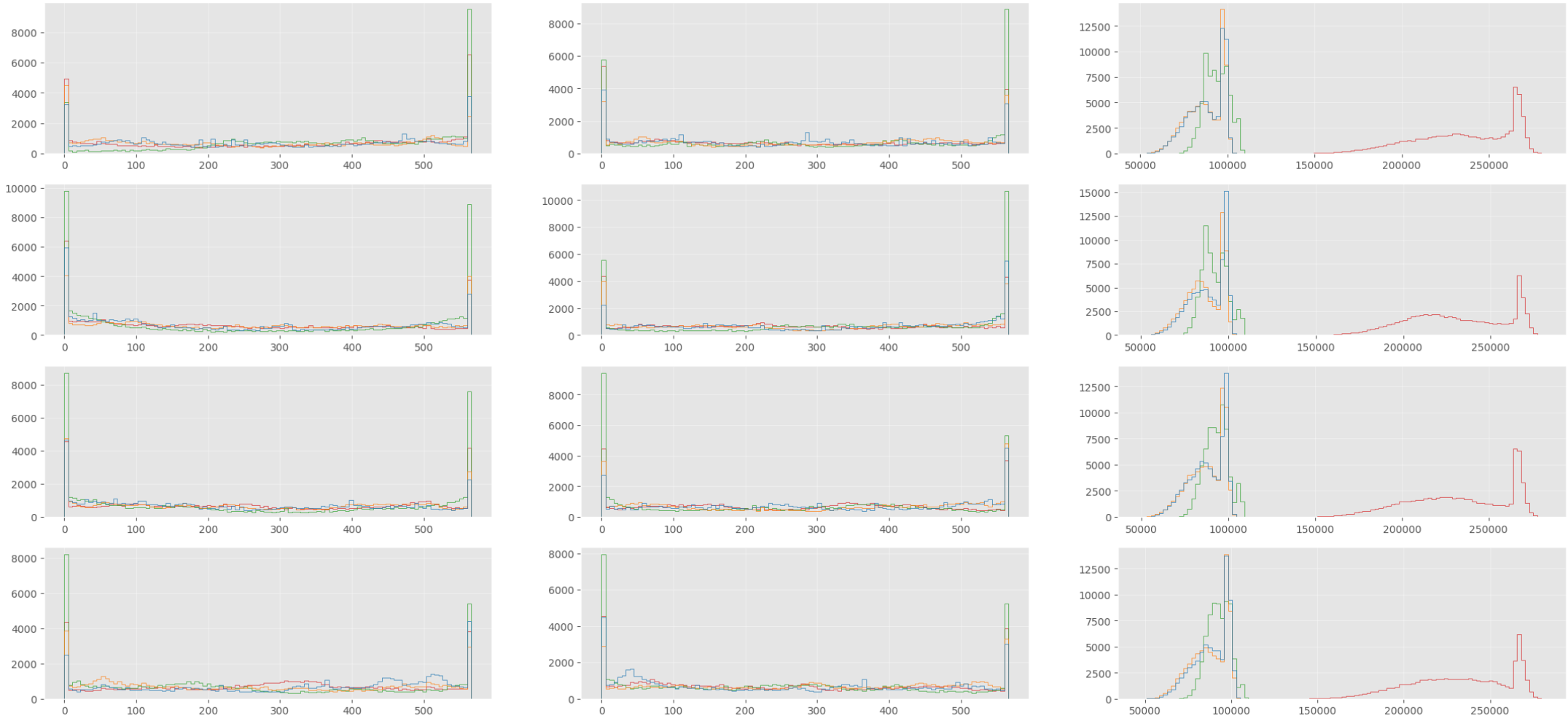
- Positions of 16 Wave Energy Converters (WECs): X1, X2, ..., X16 and Y1, Y2, ..., Y16 (coordinates in meters).
- Absorbed power of each WEC: P1, P2, ..., P16.
- Total power output of the farm: power_output.

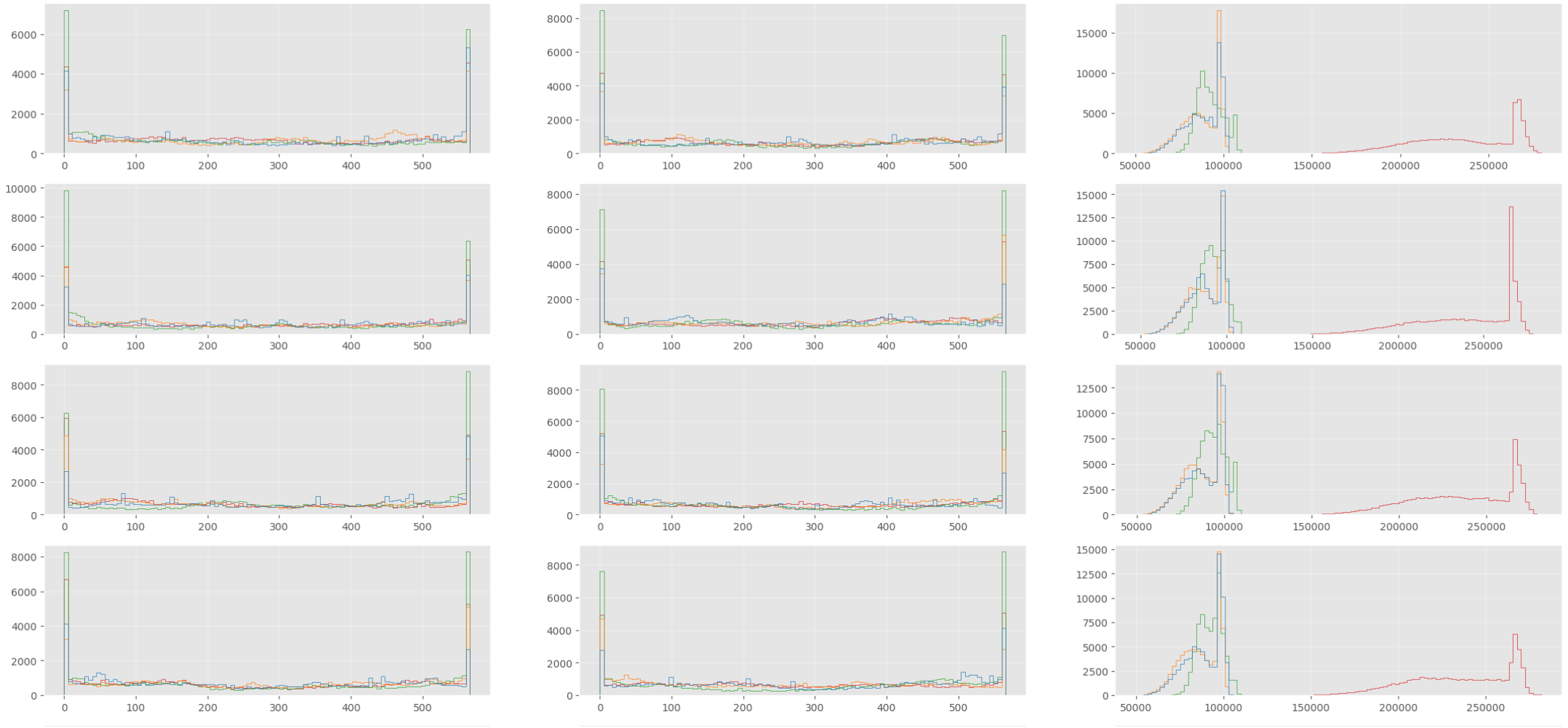


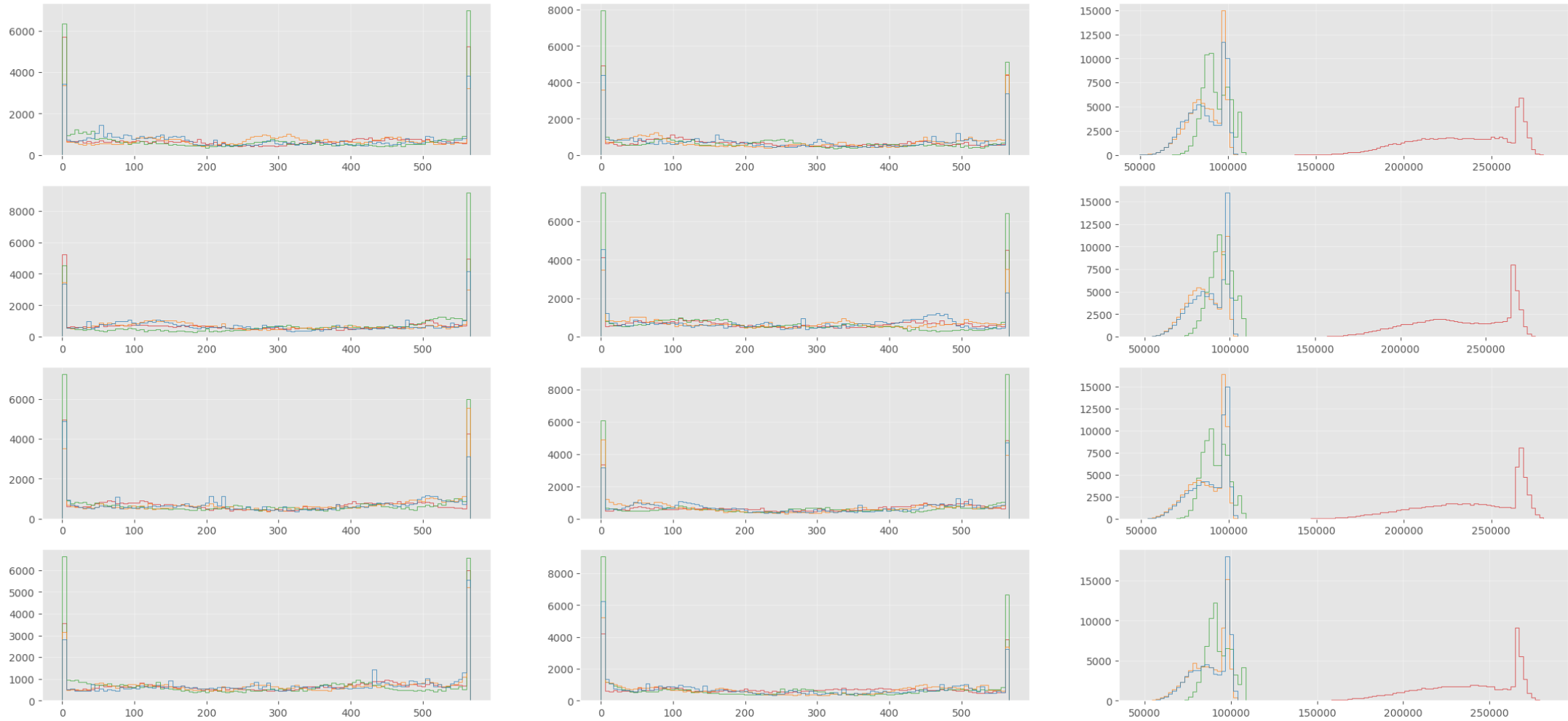
A scatter plot of region vs. power_output shows distribution across regions

Histogram plots for X, Y, and P columns visualize their distributions across regions




0 2000 4000 6000 8000
0 100 200 300 400 500
0 2000 4000 6000 8000
0 100 200 300 400 500
0 2500 5000 7500 10000 12500
50000 100000 150000 200000 250000
0 2000 4000 6000 8000 10000
0 100 200 300 400 500
0 2000 4000 6000 8000 10000
0 100 200 300 400 500
0 2500 5000 7500 10000 12500 15000
50000 100000 150000 200000 250000
0 2000 4000 6000 8000
0 100 200 300 400 500
0 2000 4000 6000 8000
0 100 200 300 400 500
0 2500 5000 7500 10000 12500
50000 100000 150000 200000 250000
0 2000 4000 6000 8000
0 100 200 300 400 500
0 2000 4000 6000 8000
0 100 200 300 400 500
0 2500 5000 7500 10000 12500
50000 100000 150000 200000 250000


0 2000 4000 6000
0 2000 4000 6000 8000
0 5000 10000 15000
0 2000 4000 6000 8000 10000
0 2000 4000 6000 8000
0 2500 5000 7500 10000 12500 15000
0 2000 4000 6000 8000
0 2000 4000 6000 8000
0 2500 5000 7500 10000 12500
0 2000 4000 6000 8000
0 2000 4000 6000 8000
0 2500 5000 7500 10000 12500 15000
0 100 200 300 400 500
50000 100000 150000 200000 250000


0 100 200 300 400 500
50000 100000 150000 200000 250000

The histograms illustrate the distribution of variables X, Y, and P from the Wave Energy Converters dataset for four locations: Tasmania, Sydney, Perth, and Adelaide:

- **X and Y Variables**: Show a similar distribution, with most values concentrated in lower ranges and a significant spike at the higher end.
- **P Variable**: Displays a distinct peak around 100,000–125,000, with differences in shape between locations, suggesting variability in energy output or environmental conditions.
- **Location Differences**: Each location exhibits unique patterns, with Tasmania and Sydney showing more pronounced peaks, reflecting local influences on wave energy conversion.

A correlation value close to 1 indicates a very strong positive linear relationship between Power Input and Power Output. A high correlation can indicate the need to exclude one of the variables to avoid multicollinearity. Both variables likely provide similar information to the model.

1.000000	0.941346	0.937452	0.939979	0.938531	0.935371	0.937399	0.939096	0.936759	0.944144	0.939321	0.940137	0.936412	0.939312	0.942305	0.938874	0.972082
0.941346	1.000000	0.938508	0.939181	0.939783	0.935239	0.937427	0.938581	0.939949	0.943601	0.941717	0.939070	0.940534	0.938686	0.942425	0.943372	0.973021
0.937452	0.938508	1.000000	0.933547	0.935110	0.933677	0.933089	0.936273	0.927833	0.934302	0.935846	0.932086	0.931423	0.930861	0.936619	0.934798	0.967106
0.939979	0.939181	0.933547	1.000000	0.941199	0.931059	0.934780	0.935827	0.935282	0.938915	0.936624	0.934123	0.933968	0.933825	0.941403	0.938603	0.969664
0.938531	0.939783	0.935110	0.941199	1.000000	0.930869	0.940260	0.938592	0.936275	0.940425	0.936117	0.936959	0.937295	0.938290	0.940014	0.938823	0.970955
0.935371	0.935239	0.933677	0.931059	0.930869	1.000000	0.926720	0.935706	0.931400	0.932784	0.933799	0.933759	0.932129	0.933623	0.934483	0.934642	0.966068
0.937399	0.937427	0.933089	0.934780	0.940260	0.926720	1.000000	0.933220	0.932074	0.938942	0.932605	0.933803	0.936293	0.933604	0.936680	0.934674	0.967850
0.939096	0.938581	0.936273	0.935827	0.938592	0.935706	0.933220	1.000000	0.933548	0.939732	0.935915	0.934095	0.932629	0.934609	0.939318	0.937313	0.969303
0.936759	0.939949	0.927833	0.935282	0.936275	0.931400	0.932074	0.933548	1.000000	0.938481	0.934642	0.938786	0.938078	0.939445	0.938139	0.939971	0.969166
0.944144	0.943601	0.934302	0.938915	0.940425	0.932784	0.938942	0.939732	0.938481	1.000000	0.940501	0.938984	0.937365	0.940810	0.943078	0.940094	0.972577
0.939321	0.941717	0.935846	0.936624	0.936117	0.933799	0.932605	0.935915	0.934642	0.940501	1.000000	0.935027	0.933904	0.937006	0.939835	0.938878	0.969852
0.940137	0.939070	0.932086	0.934123	0.936959	0.933759	0.933803	0.934095	0.938786	0.938984	0.935027	1.000000	0.938881	0.938168	0.940838	0.936777	0.969841
0.936412	0.940534	0.931423	0.933968	0.937295	0.932129	0.936293	0.932629	0.938078	0.937365	0.933904	0.938881	1.000000	0.933536	0.936810	0.938166	0.968910
0.939312	0.938686	0.930861	0.933825	0.938290	0.933623	0.933604	0.934609	0.939445	0.940810	0.937006	0.938168	0.933536	1.000000	0.936824	0.938926	0.969596
0.942305	0.942425	0.936619	0.941403	0.940014	0.934483	0.936680	0.939318	0.938139	0.943078	0.939835	0.940838	0.936810	0.936824	1.000000	0.939953	0.972311
0.938874	0.943372	0.934798	0.938603	0.938823	0.934642	0.934674	0.937313	0.939971	0.940094	0.938878	0.936777	0.938166	0.938926	0.939953	1.000000	0.971320
0.972082	0.973021	0.967106	0.969664	0.970955	0.966068	0.967850	0.969303	0.969166	0.972577	0.969852	0.969841	0.968910	0.969596	0.972311	0.971320	1.000000

Correlation matrix for power input and power output values

CALCULATING MEAN DISTANCE BETWEEN POINTS

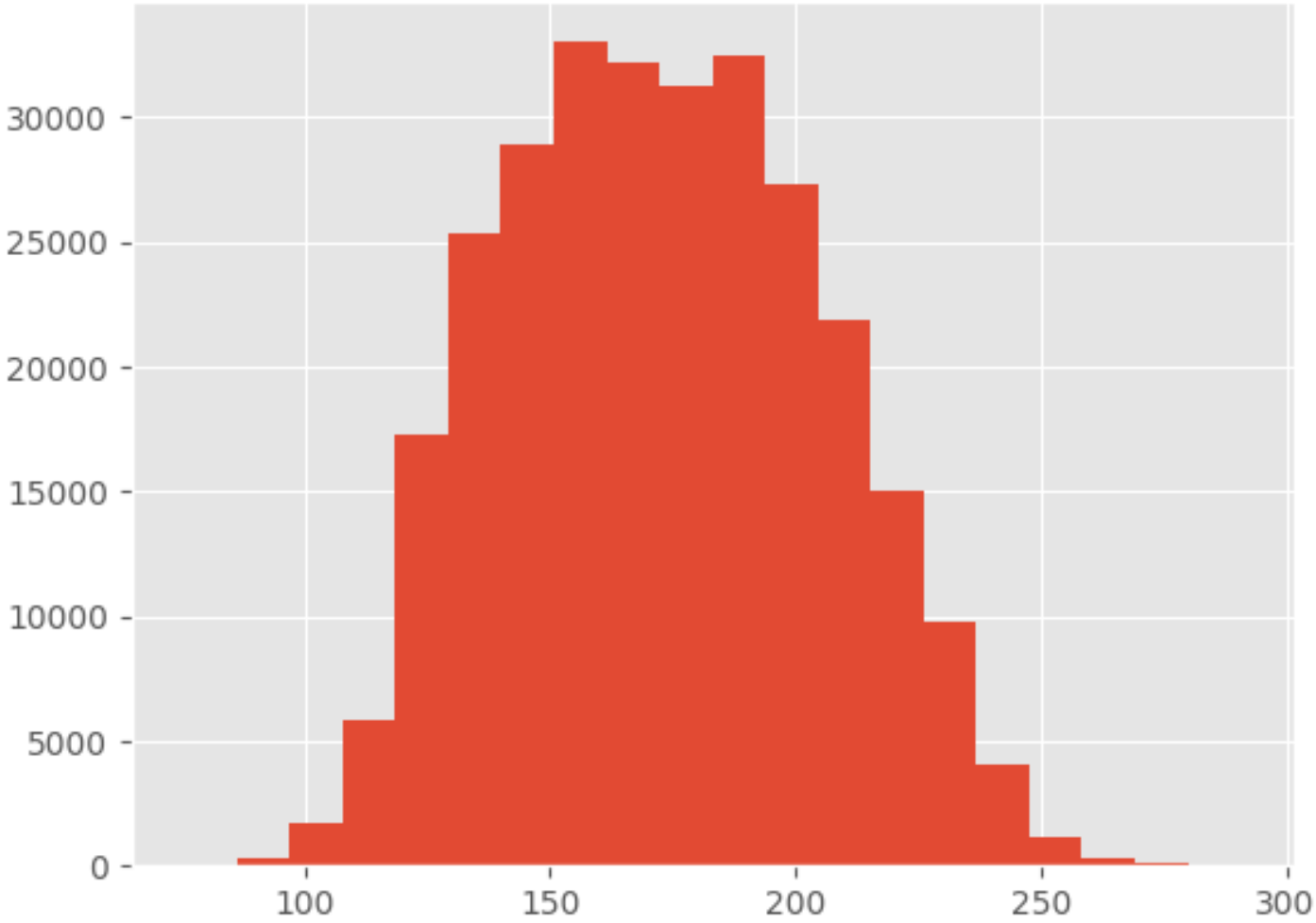
Key Observations:

- **Range:** The mean_dist ranges from approximately 75 to 291, showing a wide variety of point layouts.
- **Central Tendency:** The mean and median are close, suggesting a relatively symmetric distribution.
- **Spread:** A standard deviation of ~32 indicates moderate variability in point dispersion.

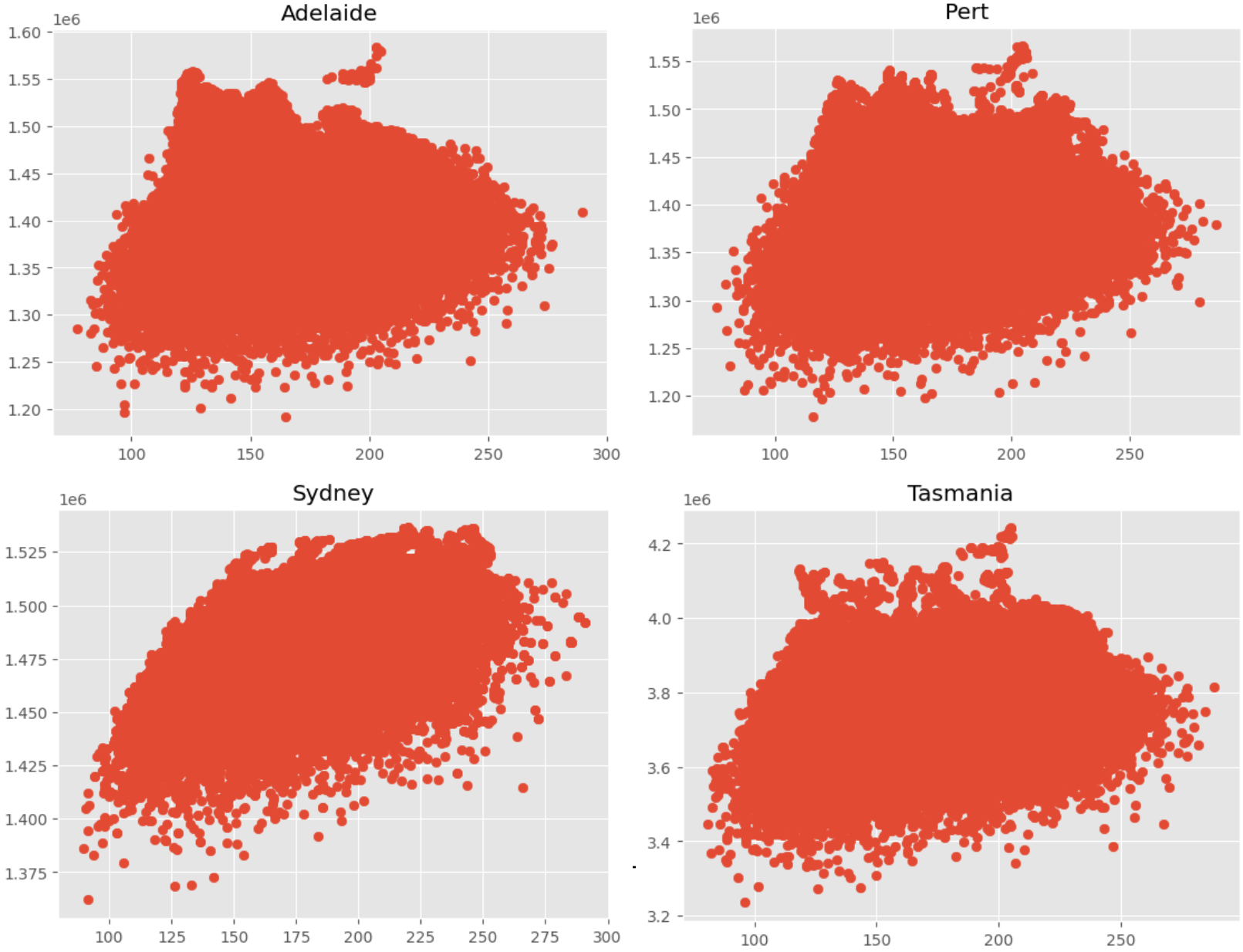
Statistic	Value
Count	287999.00
Mean	172.96
Std	32.10
Min	75.35
25%	148.01
50% (Median)	172.03
75%	196.51
Max	290.69

DISTRIBUTION OF THE MEAN_DIST VALUES

The histogram of the mean_dist values shows a distribution with a peak around 150-175, indicating that most data points have moderate distances between points. The distribution is roughly symmetric with a slight right skew, suggesting a normal-like distribution with fewer instances of larger distances.



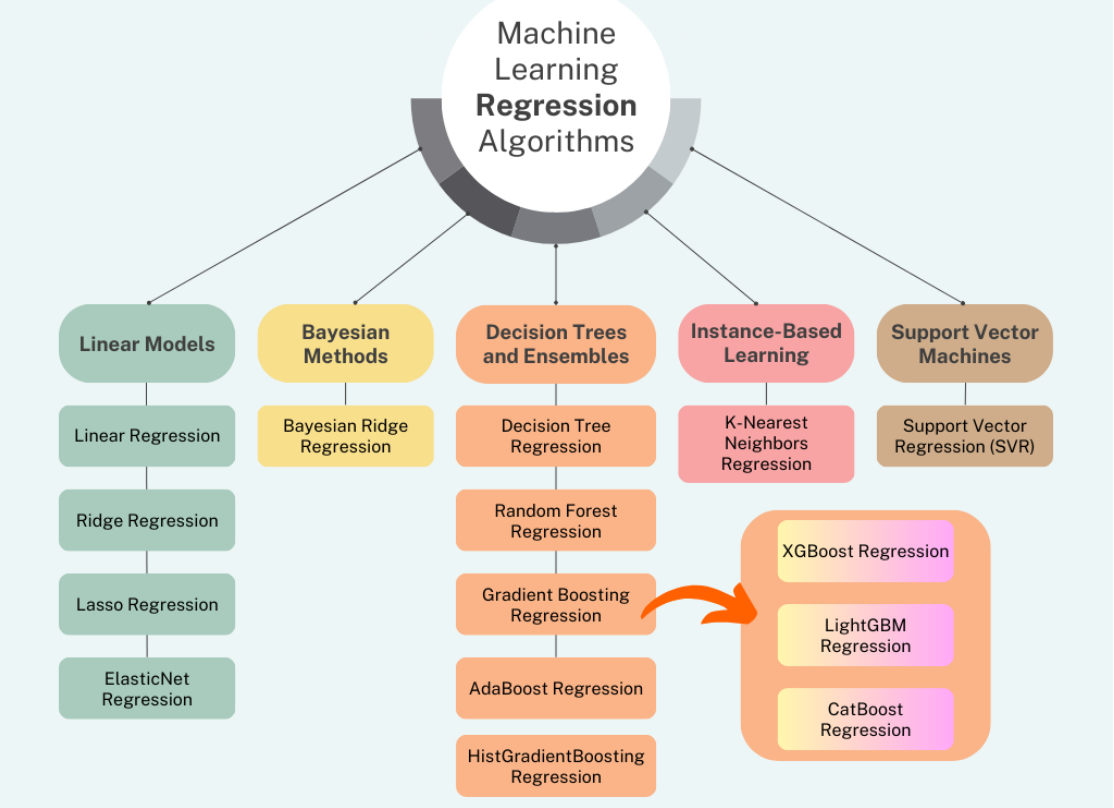
RELATIONSHIP BETWEEN MEAN DISTANCE AND POWER_OUTPUT



Observations:

- Potential Pattern in Mean Distance and Power Output:
  - These scatter plots suggest a dense cluster of power_output values between 1.2M and 1.55M (3.2M to 4.2M in Tasmania), with mean_dist mostly in the range of 100 to 250.
  - Although the direct relationship appears non-linear and scattered, there may be underlying patterns worth exploring further.
- Hypothesis for Modeling:
  - The variability in mean_dist could provide meaningful signals, either on its own or through interactions with other features, that may improve the predictive power of models.
  - Its role in explaining power_output in different regions will be tested in models.

MODELS AND METRICS



The following models that are used:

- Linear Regression
- Bayesian Ridge Regression
- Random Forest Regressoin
- XGBoost Regression
- Light GBM Regression
- K-Nearest Neighbors Regression

ADELAIDE

Default inputs

LinearRegression:

Mean error: 51029.5 Mean error % 3.62

BayesianRidge:

Mean error: 51029.3 Mean error % 3.62

KNeighborsRegressor:

Mean error: 25783.9 Mean error % 1.83

RandomForestRegressor:

Mean error: 23701.3 Mean error % 1.68

XGBRegressor:

Mean error: 19933.3 Mean error % 1.41

LGBM:

Mean error: 21177.9 Mean error % 1.50

Extended inputs with mean dist

LinearRegression:

Mean error: 50878.2 Mean error % 3.61

BayesianRidge:

Mean error: 50879.3 Mean error % 3.61

KNeighborsRegressor:

Mean error: 25722.7 Mean error % 1.82

RandomForestRegressor:

Mean error: 23799.9 Mean error % 1.69

XGBRegressor:

Mean error: 19956.6 Mean error % 1.42

LGBM:

Mean error: 21407.0 Mean error % 1.52

PERT

Default inputs

LinearRegression:

Mean error: 48198.2 Mean error % 3.46

BayesianRidge:

Mean error: 48197.4 Mean error % 3.46

KNeighborsRegressor:

Mean error: 25506.0 Mean error % 1.83

RandomForestRegressor:

Mean error: 24320.4 Mean error % 1.74

XGBRegressor:

Mean error: 20423.4 Mean error % 1.46

LGBM:

Mean error: 21145.2 Mean error % 1.52

Extended inputs with mean dist

LinearRegression:

Mean error: 47840.8 Mean error % 3.43

BayesianRidge:

Mean error: 47840.8 Mean error % 3.43

KNeighborsRegressor:

Mean error: 25453.7 Mean error % 1.83

RandomForestRegressor:

Mean error: 24194.8 Mean error % 1.74

XGBRegressor:

Mean error: 20449.2 Mean error % 1.47

LGBM:

Mean error: 21502.8 Mean error % 1.54

SYDNEY

Default inputs

LinearRegression:

Mean error: 21594.3 Mean error % 1.45

BayesianRidge:

Mean error: 21593.7 Mean error % 1.45

KNeighborsRegressor:

Mean error: 14012.9 Mean error % 0.94

RandomForestRegressor:

Mean error: 8933.1 Mean error % 0.60

XGBRegressor:

Mean error: 8378.5 Mean error % 0.56

LGBM:

Mean error: 8960.4 Mean error % 0.60

Extended inputs with mean dist

LinearRegression:

Mean error: 18117.4 Mean error % 1.22

BayesianRidge:

Mean error: 18117.4 Mean error % 1.22

KNeighborsRegressor:

Mean error: 13949.8 Mean error % 0.94

RandomForestRegressor:

Mean error: 8464.6 Mean error % 0.57

XGBRegressor:

Mean error: 8088.6 Mean error % 0.54

LGBM:

Mean error: 8772.6 Mean error % 0.59

TASMANIA

Default inputs

LinearRegression:

Mean error: 103493.2 Mean error % 2.75

BayesianRidge:

Mean error: 103493.0 Mean error % 2.75

KNeighborsRegressor:

Mean error: 68337.8 Mean error % 1.82

RandomForestRegressor:

Mean error: 60661.3 Mean error % 1.61

XGBRegressor:

Mean error: 52988.7 Mean error % 1.41

LGBM:

Mean error: 55300.7 Mean error % 1.47

Extended inputs with mean dist

LinearRegression:

Mean error: 101862.9 Mean error % 2.71

BayesianRidge:

Mean error: 101861.3 Mean error % 2.71

KNeighborsRegressor:

Mean error: 68241.3 Mean error % 1.81

RandomForestRegressor:

Mean error: 61438.1 Mean error % 1.63

XGBRegressor:

Mean error: 53010.2 Mean error % 1.41

LGBM:

Mean error: 55543.8 Mean error % 1.48

ADELAIDE

Impact of parameters in XGBRegressor

Adelaide best model:

model = XGBRegressor(n_estimators=500, booster='gbtree', max_depth=7, learning_rate=0.3)

Mean error: 19478.9 Mean error % 1.38

```
n_estimators: 10                Mean error: 28227.4    Mean error % 2.00
n_estimators: 50                Mean error: 21518.6    Mean error % 1.53
n_estimators: 100               Mean error: 19933.3    Mean error % 1.41
n_estimators: 200               Mean error: 19312.5    Mean error % 1.37
n_estimators: 500               Mean error: 19157.0    Mean error % 1.36
n_estimators: 1000              Mean error: 19146.1    Mean error % 1.36

booster: gbtree          Mean error: 19933.3    Mean error % 1.41
booster: gblinear        Mean error: 51028.5    Mean error % 3.62
booster: dart            Mean error: 19933.3    Mean error % 1.41

max_depth: 0          Mean error: 25228.6    Mean error % 1.79
max_depth: 1          Mean error: 41060.1    Mean error % 2.91
max_depth: 2          Mean error: 27263.5    Mean error % 1.93
max_depth: 3          Mean error: 23023.8    Mean error % 1.63
max_depth: 4          Mean error: 21397.8    Mean error % 1.52
max_depth: 5          Mean error: 20379.9    Mean error % 1.45
max_depth: 6          Mean error: 19933.3    Mean error % 1.41
max_depth: 7          Mean error: 19775.4    Mean error % 1.40
max_depth: 8          Mean error: 20189.9    Mean error % 1.43
max_depth: 9          Mean error: 20509.8    Mean error % 1.45

learning_rate: 0.05       Mean error: 25314.2    Mean error % 1.80
learning_rate: 0.1        Mean error: 22115.6    Mean error % 1.57
learning_rate: 0.2        Mean error: 20050.5    Mean error % 1.42
learning_rate: 0.3        Mean error: 19933.3    Mean error % 1.41
learning_rate: 0.4        Mean error: 20508.1    Mean error % 1.45
learning_rate: 0.5        Mean error: 21148.7    Mean error % 1.50
```

PERT

Impact of parameters in XGBRegressor

Pert best model:

model = XGBRegressor(n_estimators=1000, booster='gbtree', max_depth=7, learning_rate=0.3)

Mean error: 20039.8 Mean error % 1.44

```
n_estimators: 10                Mean error: 28258.5    Mean error % 2.03
n_estimators: 50                Mean error: 21815.6    Mean error % 1.56
n_estimators: 100               Mean error: 20423.4    Mean error % 1.46
n_estimators: 200               Mean error: 19906.1    Mean error % 1.43
n_estimators: 500               Mean error: 19891.3    Mean error % 1.43
n_estimators: 1000              Mean error: 19860.0    Mean error % 1.42

booster: gbtree          Mean error: 20423.4    Mean error % 1.46
booster: gblinear        Mean error: 48191.8    Mean error % 3.46
booster: dart            Mean error: 20423.4    Mean error % 1.46

max_depth: 0          Mean error: 25758.9    Mean error % 1.85
max_depth: 1          Mean error: 41477.3    Mean error % 2.97
max_depth: 2          Mean error: 27678.4    Mean error % 1.98
max_depth: 3          Mean error: 23056.2    Mean error % 1.65
max_depth: 4          Mean error: 21604.2    Mean error % 1.55
max_depth: 5          Mean error: 20616.6    Mean error % 1.48
max_depth: 6          Mean error: 20423.4    Mean error % 1.46
max_depth: 7          Mean error: 20446.7    Mean error % 1.47
max_depth: 8          Mean error: 20649.1    Mean error % 1.48
max_depth: 9          Mean error: 21096.6    Mean error % 1.51

learning_rate: 0.05       Mean error: 25333.7    Mean error % 1.82
learning_rate: 0.1        Mean error: 22092.5    Mean error % 1.58
learning_rate: 0.2        Mean error: 20392.7    Mean error % 1.46
learning_rate: 0.3        Mean error: 20423.4    Mean error % 1.46
learning_rate: 0.4        Mean error: 21043.9    Mean error % 1.51
learning_rate: 0.5        Mean error: 22015.0    Mean error % 1.58
```

SYDNEY

Impact of parameters in XGBRegressor

Sydney best model:

model = XGBRegressor(n_estimators=1000, booster='gbtree', max_depth=9, learning_rate=0.3)

Mean error: 20959.4 Mean error % 1.50

```
n_estimators: 10                Mean error: 12730.7    Mean error % 0.86
n_estimators: 50                Mean error: 9350.3     Mean error % 0.63
n_estimators: 100               Mean error: 8378.5     Mean error % 0.56
n_estimators: 200               Mean error: 7544.0     Mean error % 0.51
n_estimators: 500               Mean error: 6800.5     Mean error % 0.46
n_estimators: 1000              Mean error: 6569.9     Mean error % 0.44

booster: gbtree          Mean error: 8378.5     Mean error % 0.56
booster: gblinear        Mean error: 21597.7    Mean error % 1.45
booster: dart            Mean error: 8378.5     Mean error % 0.56

max_depth: 0          Mean error: 8928.2     Mean error % 0.60
max_depth: 1          Mean error: 13644.0    Mean error % 0.92
max_depth: 2          Mean error: 11071.3    Mean error % 0.74
max_depth: 3          Mean error: 10264.0    Mean error % 0.69
max_depth: 4          Mean error: 9632.1     Mean error % 0.65
max_depth: 5          Mean error: 8948.7     Mean error % 0.60
max_depth: 6          Mean error: 8378.5     Mean error % 0.56
max_depth: 7          Mean error: 7857.3     Mean error % 0.53
max_depth: 8          Mean error: 7602.9     Mean error % 0.51
max_depth: 9          Mean error: 7415.7     Mean error % 0.50

learning_rate: 0.05       Mean error: 11017.9    Mean error % 0.74
learning_rate: 0.1        Mean error: 9625.0     Mean error % 0.65
learning_rate: 0.2        Mean error: 8559.7     Mean error % 0.58
learning_rate: 0.3        Mean error: 8378.5     Mean error % 0.56
learning_rate: 0.4        Mean error: 8410.1     Mean error % 0.57
learning_rate: 0.5        Mean error: 8519.6     Mean error % 0.57
```

TASMANIA

Impact of parameters in XGBRegressor

Tasmania best model:

model = XGBRegressor(n_estimators=500, booster='gbtree', max_depth=7, learning_rate=0.3)

Mean error: 51611.5 Mean error % 1.37

```
n_estimators: 10                Mean error: 71946.8    Mean error % 1.91
n_estimators: 50                Mean error: 56837.4    Mean error % 1.51
n_estimators: 100               Mean error: 52988.7    Mean error % 1.41
n_estimators: 200               Mean error: 51171.0    Mean error % 1.36
n_estimators: 500               Mean error: 50638.6    Mean error % 1.35
n_estimators: 1000              Mean error: 50631.0    Mean error % 1.35

booster: gbtree          Mean error: 52988.7    Mean error % 1.41
booster: gblinear        Mean error: 103546.2   Mean error % 2.75
booster: dart            Mean error: 52988.7    Mean error % 1.41

max_depth: 0          Mean error: 64339.1    Mean error % 1.71
max_depth: 1          Mean error: 92920.6    Mean error % 2.47
max_depth: 2          Mean error: 70094.6    Mean error % 1.86
max_depth: 3          Mean error: 60781.1    Mean error % 1.62
max_depth: 4          Mean error: 56344.6    Mean error % 1.50
max_depth: 5          Mean error: 54290.0    Mean error % 1.44
max_depth: 6          Mean error: 52988.7    Mean error % 1.41
max_depth: 7          Mean error: 52753.7    Mean error % 1.40
max_depth: 8          Mean error: 53275.8    Mean error % 1.42
max_depth: 9          Mean error: 54684.5    Mean error % 1.45

learning_rate: 0.05       Mean error: 64917.9    Mean error % 1.73
learning_rate: 0.1        Mean error: 58264.6    Mean error % 1.55
learning_rate: 0.2        Mean error: 53454.1    Mean error % 1.42
learning_rate: 0.3        Mean error: 52988.7    Mean error % 1.41
learning_rate: 0.4        Mean error: 54045.5    Mean error % 1.44
learning_rate: 0.5        Mean error: 55973.1    Mean error % 1.49
```