ML Checklist: Getting Started

- What is the objective in business terms?
- Understand how your solution will be used
- Are there current solutions/workarounds?
- What categorization? (supervised/unsupervised, etc)
- How will performance be measured?
- Does the performance measure match the business objective?
- What's the minimum acceptable performance?
- Any reuse possible?
- Is human expertise available?
- What would be the manual solution?
- Are there any assumptions?
- Document! Document! Document!

Selecting Performance Measures

- How will you measure "how good" your model is performing?

Confution Matrix

N=300	Predicted: No	Predicted: Yes	
Actual: No	TN=140	FP=15	155
Actual: Yes	FN=100	TP=45	145
	240	60	

Common Regression Measures

RMSE: Room Mean Squared Error

- Most commonly used $RMSE = \sqrt{\frac{\sum_{i=i}^{N}(Predicted_i - Actual_i)^2}{N}}$

MAE: Mean Absolute Error

- Preferred when many outliers $MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - X|$

R^2: R-squared

- Also called coefficient of determination Division of the $\text{Sum of Square Regression} = \sum(y' - \bar{y'})^2$ and $\text{Sum of Square Total} = \sum(y - \bar{y})^2$ where the Sum of Square Regression of made up of $1 - SSE \div SST = 1 - \frac{\sum(y-\bar{y'})^2}{SST}$

Summar of Common Measures

Acronym	Full Name	Residual Operation	Robust to Outliers
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes
MPE	Mean Percentage Error	N/A	Yes

Check and Validate All Your Assumptions

Module 2.2 Assignment

Using the data for the California Census, answer the following questions

1. What are the attributes for each district?

In [ ]:
```
# package imports
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
```

In [ ]:
```
# upload the data
df_cal_cens = pd.read_csv('housing.csv')
df_cal_cens.head()
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

In [ ]:

```
# get summary stats on each numerical feature
df_cal_cens.describe()
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_va
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000

In [ ]:

```
# determine unique values of the categorical features
print(f"Attributes for each district:\n{df_cal_cens['ocean_proximity'].unique()}")
```

```
Attributes for each district:
['NEAR BAY' '<1H OCEAN' 'INLAND' 'NEAR OCEAN' 'ISLAND']
```

2. What attributes are confusing to you?

The most confusing attributes to are the difference between the "<1H OCEAN" and "NEAR OCEAN". I would assume that NEAR OCEAN would mean >1H from the ocean but would most likely have to plot on a map chart to understand that information more.

In [ ]:

```

```

3. Without graphing tools, what observations can you make about the data?

After reviewing the below information we can point out some key observations

1. There are only 5 ISLAND homes and the largest count attribute is <1H OCEAN at 9,136 observations

2. ISLAND homes have the greatest median age with INLAND homes having the lowest median age
3. ISLAND homes have the lowest number of total rooms and total bedrooms compared to the other attributes
4. ISLAND has the lowest number of households (unsure what this feature means)
5. ISLAND has the lowest median income but also has the lowest std of median income compared to the other homes
6. INLAND has the lowest median house value

In [ ]:
```
# without graphing tools, we can look at the summary stats for each attribute/district
sum_stats_dict = {}
for val in df_cal_cens['ocean_proximity'].unique():
    # print(f"summary stats for {val}\n{df_cal_cens[df_cal_cens['ocean_proximity'] == val].describe()}\n")
    sum_stats_dict[val] = df_cal_cens[df_cal_cens['ocean_proximity'] == val].describe()
```

In [ ]:
```
sum_stats_dict['NEAR BAY']
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	2290.000000	2290.000000	2290.000000	2290.000000	2270.000000	2290.000000	2290.000000	2290.000000	2290.000000
mean	-122.260694	37.801057	37.730131	2493.589520	514.182819	1230.317467	488.616157	4.172885	259212.311790
std	0.147004	0.185434	13.070385	1830.817022	367.887605	885.899035	350.598369	2.017427	122818.537064
min	-122.590000	37.350000	2.000000	8.000000	1.000000	8.000000	1.000000	0.499900	22500.000000
25%	-122.410000	37.730000	29.000000	1431.250000	289.000000	718.250000	275.000000	2.834750	162500.000000
50%	-122.250000	37.790000	39.000000	2083.000000	423.000000	1033.500000	406.000000	3.818650	233800.000000
75%	-122.140000	37.907500	52.000000	3029.750000	628.750000	1495.000000	599.250000	5.054425	345700.000000
max	-122.010000	38.340000	52.000000	18634.000000	3226.000000	8276.000000	3589.000000	15.000100	500001.000000

In [ ]:
```
sum_stats_dict['<1H OCEAN']
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	9136.000000	9136.000000	9136.000000	9136.000000	9034.000000	9136.000000	9136.000000	9136.000000	9136.000000
mean	-118.847766	34.560577	29.279225	2628.343586	546.539185	1520.290499	517.744965	4.230682	240084.285464
std	1.588888	1.467127	11.644453	2160.463696	427.911417	1185.848357	392.280718	2.001223	106124.292213
min	-124.140000	32.610000	2.000000	11.000000	5.000000	3.000000	4.000000	0.499900	17500.000000
25%	-118.500000	33.860000	20.000000	1464.000000	303.000000	857.750000	293.000000	2.864900	164100.000000
50%	-118.275000	34.030000	30.000000	2108.000000	438.000000	1247.000000	421.000000	3.875000	214850.000000
75%	-118.000000	34.220000	37.000000	3141.000000	652.000000	1848.000000	617.000000	5.180500	289100.000000
max	-116.620000	41.880000	52.000000	37937.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

In [ ]:
```
sum_stats_dict['NEAR OCEAN']
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	2658.000000	2658.000000	2658.000000	2658.000000	2628.000000	2658.000000	2658.000000	2658.000000	2658.000000
mean	-119.332555	34.738439	29.347254	2583.700903	538.615677	1354.008653	501.244545	4.005785	249433.977427
std	2.327307	2.275386	11.840371	1990.724760	376.320045	1005.563166	344.445256	2.010558	122477.145927
min	-124.350000	32.540000	2.000000	15.000000	3.000000	8.000000	3.000000	0.536000	22500.000000
25%	-122.020000	32.780000	20.000000	1505.000000	313.000000	778.500000	299.000000	2.630525	150000.000000
50%	-118.260000	33.790000	29.000000	2195.000000	464.000000	1136.500000	429.000000	3.647050	229450.000000
75%	-117.182500	36.980000	37.000000	3109.000000	666.000000	1628.000000	614.000000	4.837400	322750.000000
max	-116.970000	41.950000	52.000000	30405.000000	4585.000000	12873.000000	4176.000000	15.000100	500001.000000

In [ ]:
```
sum_stats_dict['ISLAND']
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.000000
mean	-118.354000	33.358000	42.400000	1574.600000	420.400000	668.000000	276.600000	2.74442	380440.000000
std	0.070569	0.040866	13.164346	707.545264	169.320111	301.691067	113.200265	0.44418	80559.561816
min	-118.480000	33.330000	27.000000	716.000000	214.000000	341.000000	160.000000	2.15790	287500.000000
25%	-118.330000	33.340000	29.000000	996.000000	264.000000	422.000000	173.000000	2.60420	300000.000000
50%	-118.320000	33.340000	52.000000	1675.000000	512.000000	733.000000	288.000000	2.73610	414700.000000
75%	-118.320000	33.350000	52.000000	2127.000000	521.000000	744.000000	331.000000	2.83330	450000.000000
max	-118.320000	33.430000	52.000000	2359.000000	591.000000	1100.000000	431.000000	3.39060	450000.000000

In [ ]:
```
sum_stats_dict['INLAND']
```

Out[ ]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	6551.00000	6551.000000	6551.000000	6551.000000	6496.000000	6551.000000	6551.000000	6551.000000	6551.000000
mean	-119.73299	36.731829	24.271867	2717.742787	533.881619	1391.046252	477.447565	3.208996	124805.392001
std	1.90095	2.116073	12.018020	2385.831111	446.117778	1168.670126	392.252095	1.437465	70007.908494
min	-123.73000	32.640000	1.000000	2.000000	2.000000	5.000000	2.000000	0.499900	14999.000000
25%	-121.35000	34.180000	15.000000	1404.000000	282.000000	722.000000	254.000000	2.188950	77500.000000
50%	-120.00000	36.970000	23.000000	2131.000000	423.000000	1124.000000	385.000000	2.987700	108500.000000
75%	-117.84000	38.550000	33.000000	3216.000000	636.000000	1687.000000	578.000000	3.961500	148950.000000
max	-114.31000	41.950000	52.000000	39320.000000	6210.000000	16305.000000	5358.000000	15.000100	500001.000000