CALIFORNIA HOUSING PRICE

PRESENTATION PROJECT DSI204

Colab : https://colab.research.google.com/drive/1qvNiy4SOI9t8zN1k_XGmLHuNHM5zrugT?usp=sharing

HOUSING.CSV

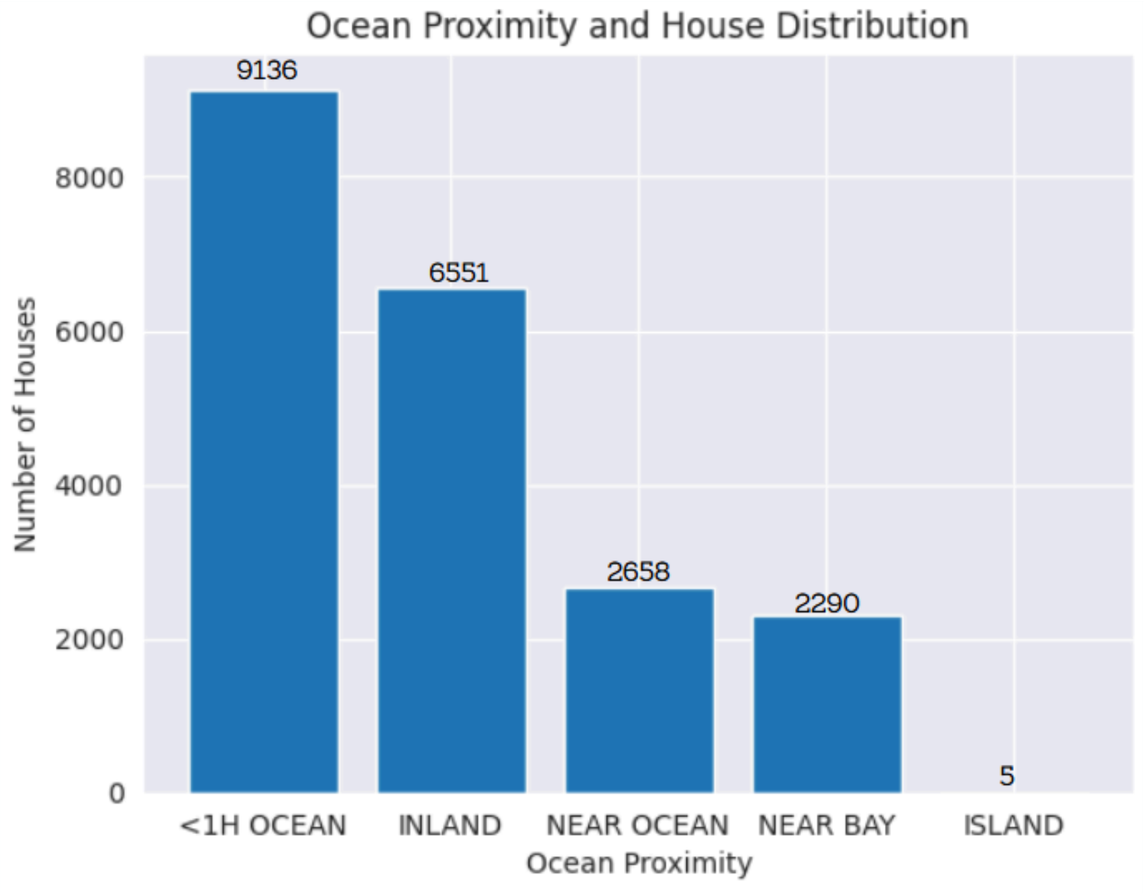
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	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
...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	INLAND
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	INLAND
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	INLAND
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	INLAND
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	INLAND

20640 rows × 10 columns

20640 rows × 10 columns

แหล่งอ้างอิง : https://www.kaggle.com/datasets/camnugent/california-housing-prices

OCEAN PROXIMITY COLUMN

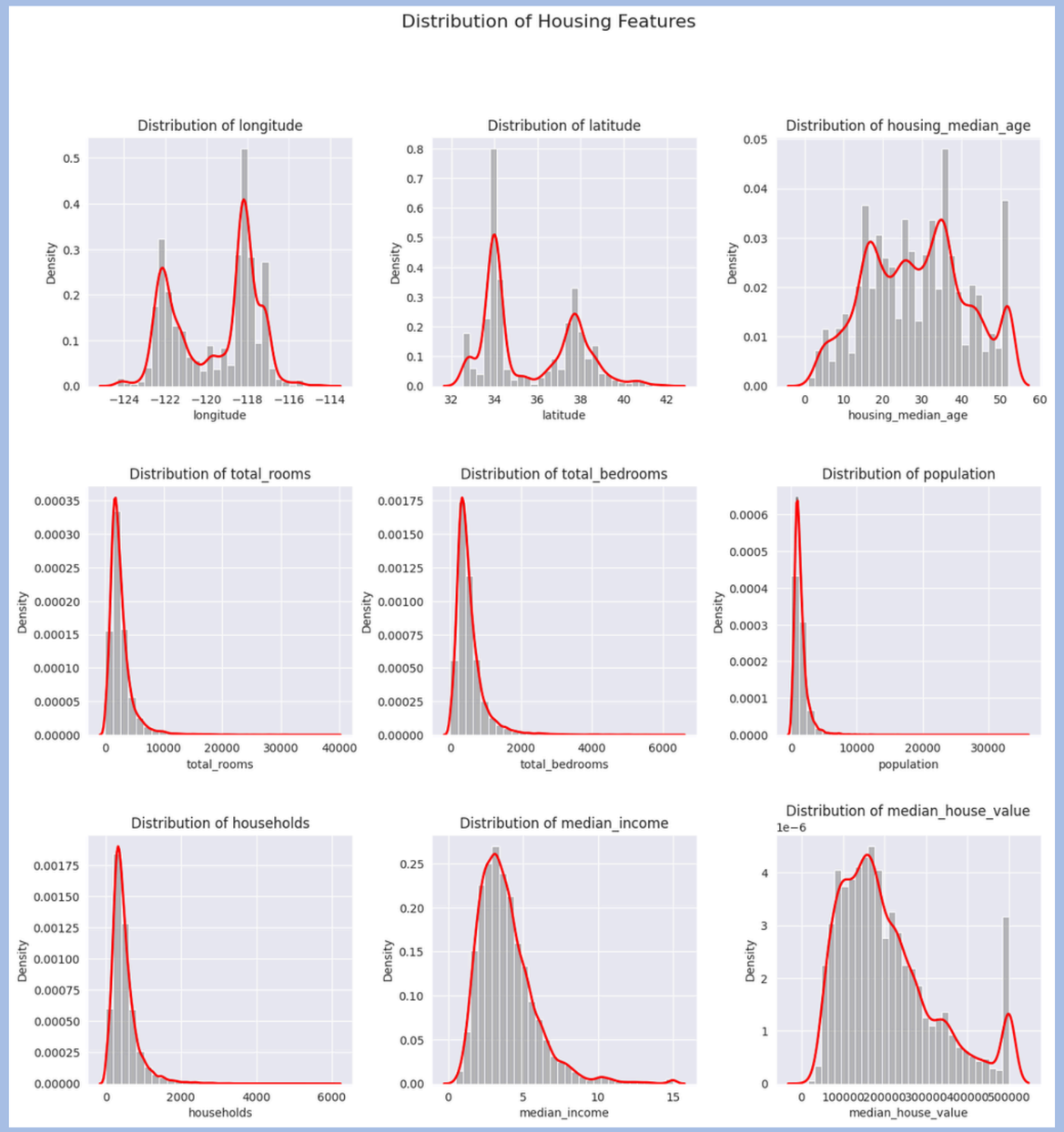


ocean_proximity	count
<1H OCEAN	9136
INLAND	6551
NEAR OCEAN	2658
NEAR BAY	2290
ISLAND	5

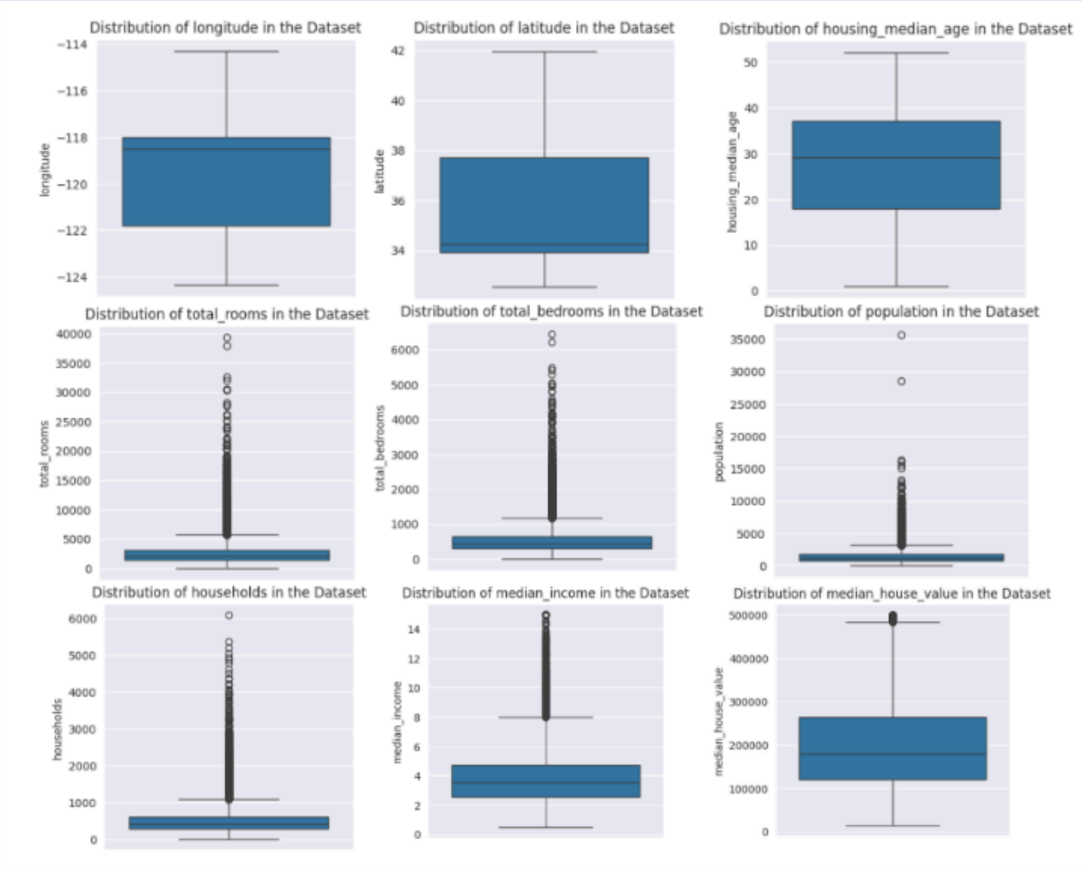
dtype: int64

DATA DESTRIBUTION

กราฟไม่ค่อยมีความ normal distribution



เมื่อเราใช้สถิติเชิงพรรณนาในการวัดข้อมูล



TOO MUCH OUTLIERS

ดังนั้น เราจึง ทำการตัด **outliers**

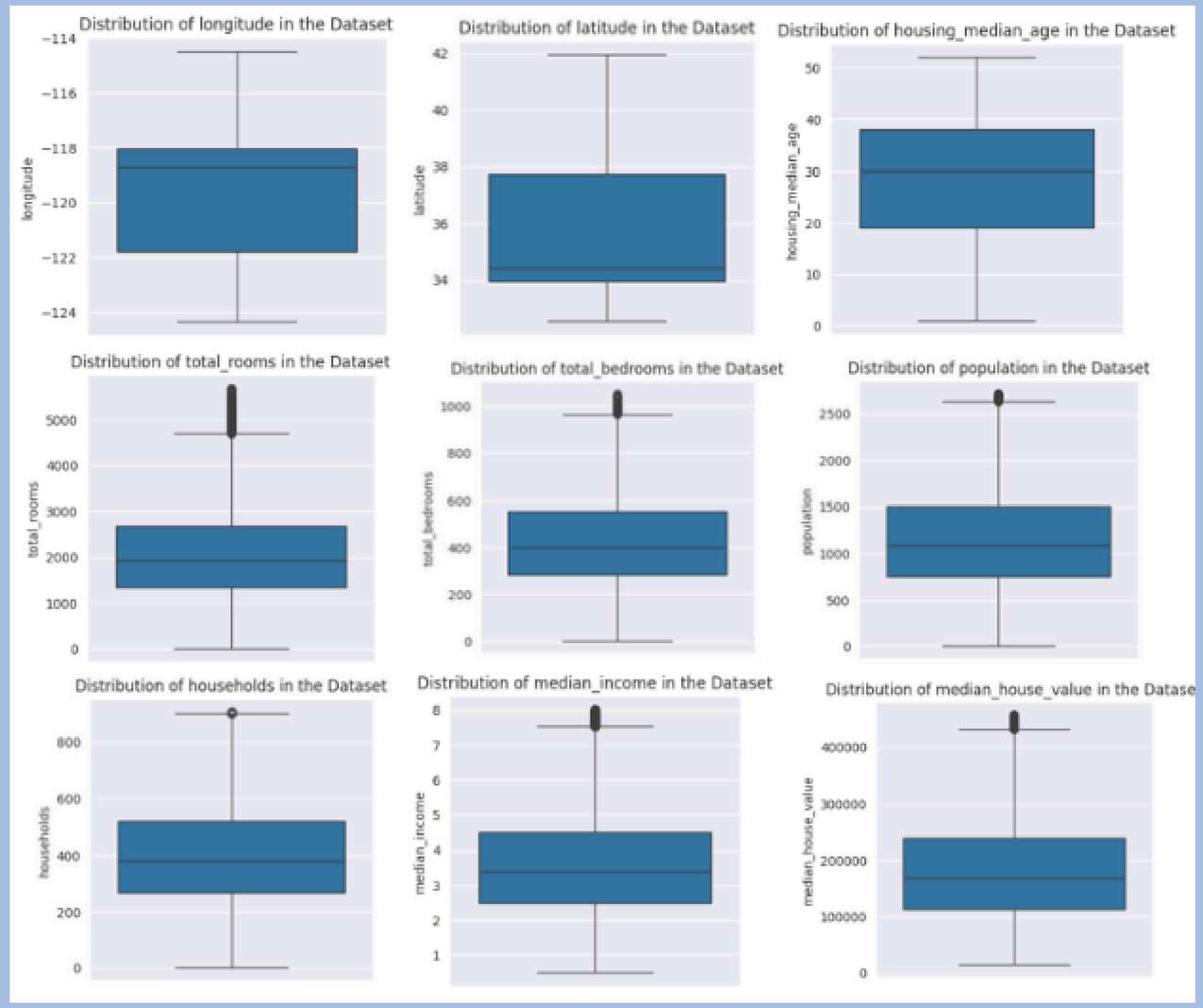
ด้วยสูตร

$$IQR = Q_3 - Q_1$$

$$Outlier = (Q1 - 1.5IQR) \cup (Q3 + 1.5IQR)$$

แหล่งอ้างอิง : https://datarockie.com/blog/excel-detect-outliers/

ผลลัพธ์หลังจากการตัด **outliers**



Mean

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
Before	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
After	-119.615230	35.716213	29.656622	2071.818296	426.704634	1148.766517	399.271749	3.577796	185334.899611

STD.

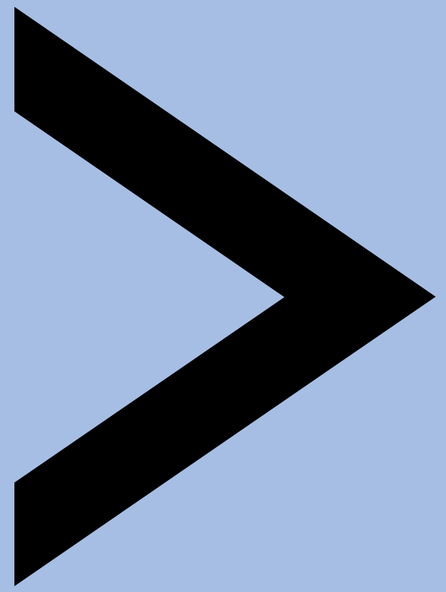
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
Before	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
After	2.005515	2.173725	12.218930	1024.143588	200.473017	548.697330	185.945188	1.446608	92032.935991

ตารางเปรียบเทียบค่า **Mean** และ **STD.** **ก่อนตัด**และ**หลังตัด** outliers

MEAN & STD.

Before

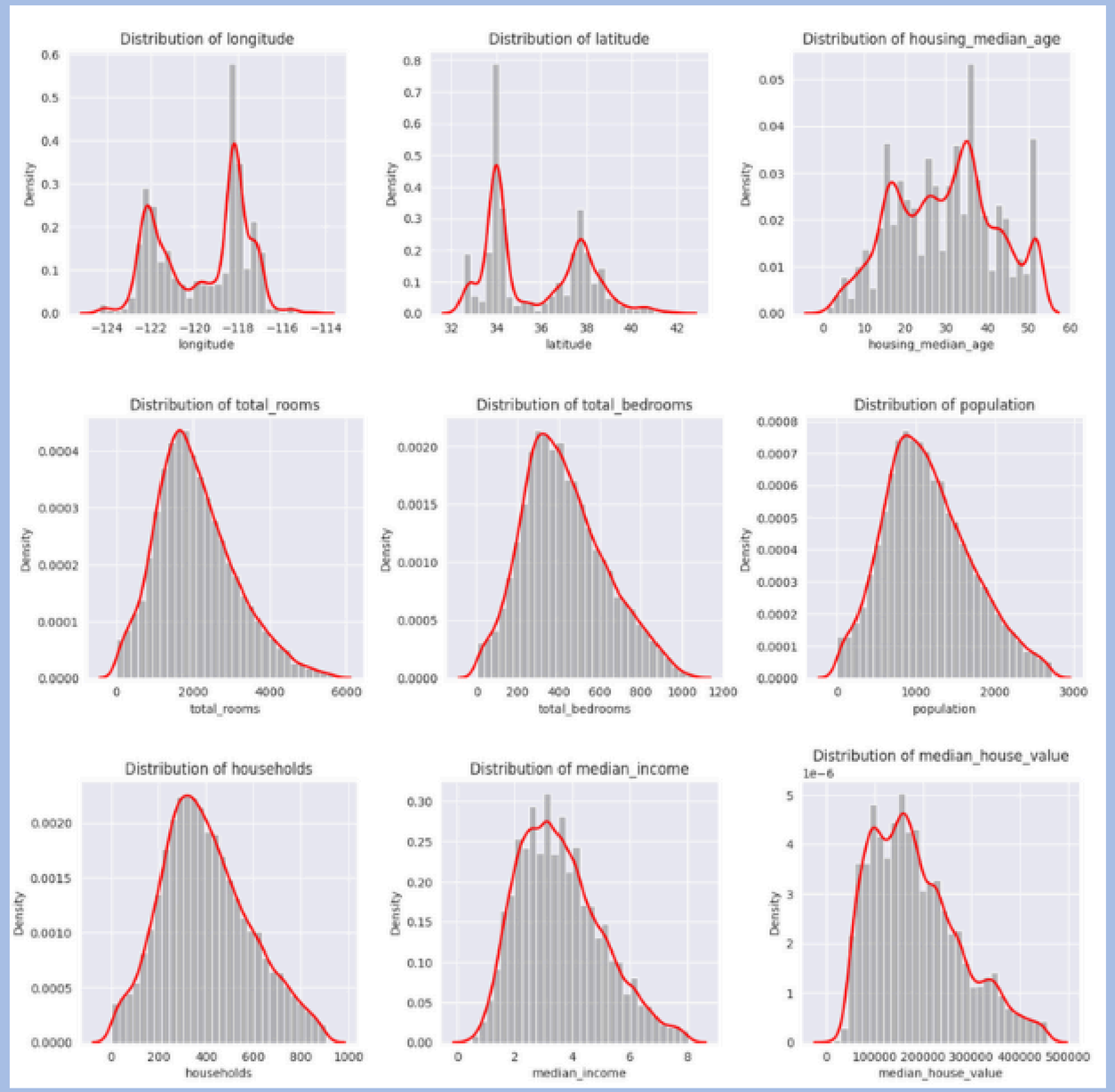
ตัด Outliers



After

ตัด Outliers

เปรียบเทียบค่า **Mean** และ **STD.** **ก่อนตัด**และ**หลังตัด** outliers



NEW

DATA DESTRIBUTION

เห็นได้ว่า ข้อมูลมีความ normal มากขึ้น

OCEAN PROXIMITY COLUMN

	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
2	0	0	0	1	0
3	0	0	0	1	0
4	0	0	0	1	0
5	0	0	0	1	0
6	0	0	0	1	0

ทำ One-Hot Encoding

แหล่งอ้างอิง : https://datarockie.com/blog/excel-detect-outliers/

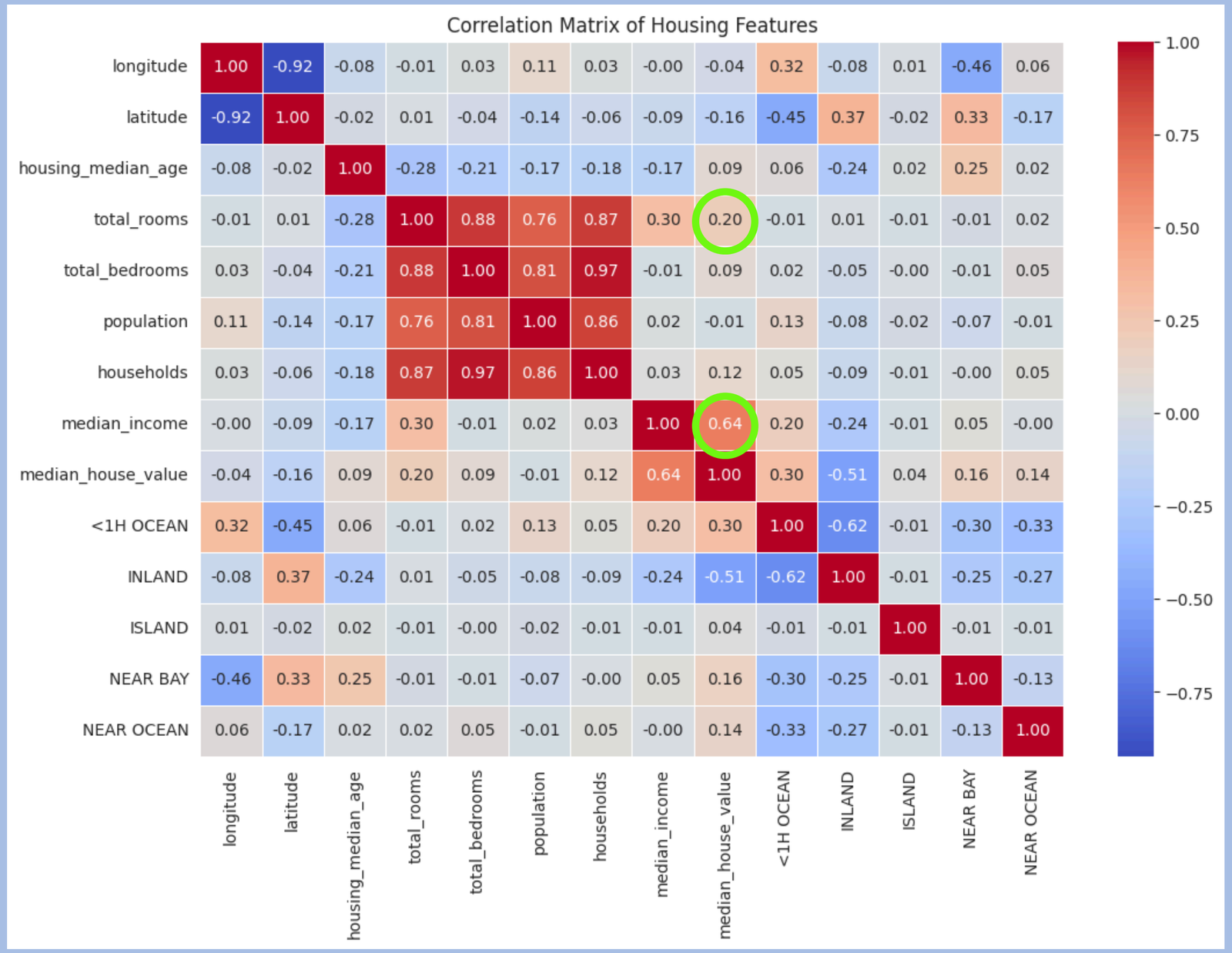
HOUSING.CSV (NEW)

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	0	0	0	1	0
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	0	0	0	1	0
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	0	0	0	1	0
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	0	0	0	1	0
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	3.6591	299200.0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	0	1	0	0	0
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	0	1	0	0	0
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	0	1	0	0	0
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	0	1	0	0	0
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	0	1	0	0	0

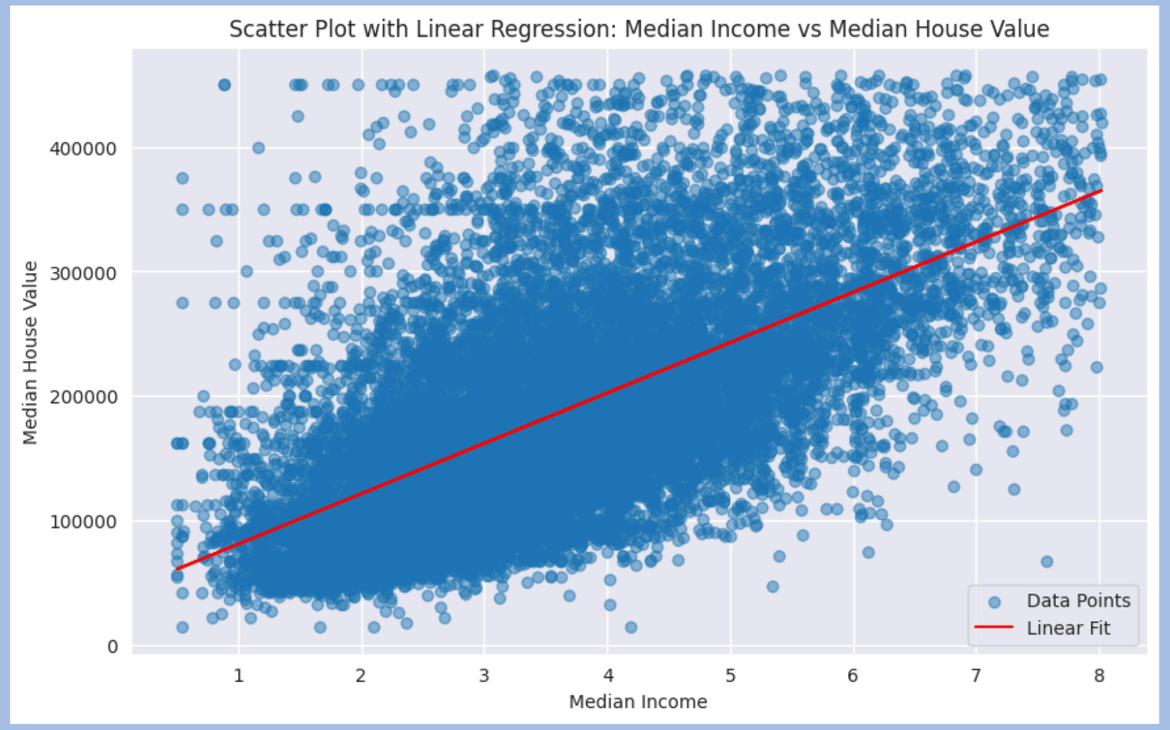
16725 rows × 14 columns

16725 rows × 14 columns

จำนวน **rows ลดลง** จำนวน **columns เพิ่มขึ้น**



INFERENTIAL STATISTICS




```
Two-Sample t-test:
data: median_house_value and total_rooms
t = 10.687, df = 58, p-value = 2.493e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 138405.9 202200.0
sample estimates:
 mean of x = 170306.667
 mean of y = 3.712
```

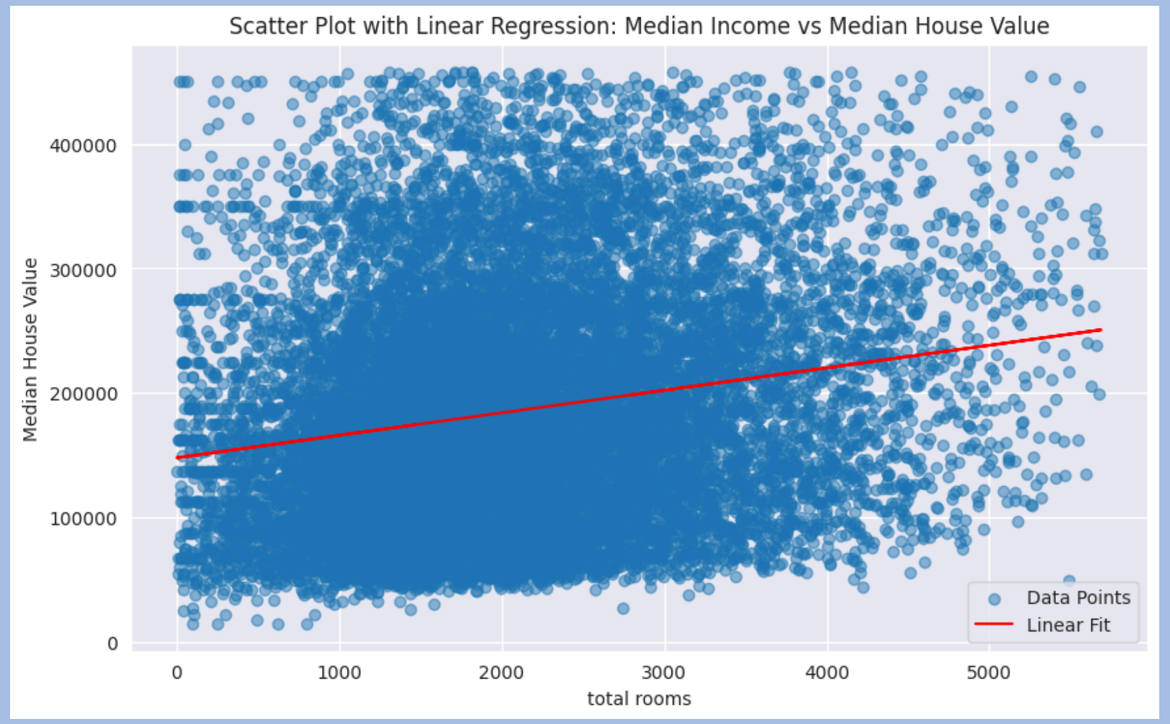
สถิติเชิงอนุมาน

Inferential Statistics

$H_0: \beta_1 = 0$ (การเปลี่ยนแปลงของคอลัมน์ median_income ไม่มีผลต่อ median_house_value)

$H_1: \beta_1 \neq 0$ (การเปลี่ยนแปลงของคอลัมน์ median_income มีผลต่อ median_house_value)

ทดสอบสมมติฐานและ**ดูความสัมพันธ์**ระหว่าง **Median_incomec** และ **Median_house_value**




```
Two-Sample t-test:
data: median_house_value and total_rooms
t = 10.567, df = 58, p-value = 3.862e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 136497.8 200296.4
sample estimates:
 mean of x = 170306.667
 mean of y = 1909.567
```

สถิติเชิงอนุมาน

Inferential Statistics

$H_0: \beta_1 = 0$ (การเปลี่ยนแปลงของคอลัมน์ total_rooms ไม่มีผลต่อ median_house_value)

$H_1: \beta_1 \neq 0$ (การเปลี่ยนแปลงของคอลัมน์ total_rooms มีผลต่อ median_house_value)

ทดสอบสมมติฐานและ**ดูความสัมพันธ์**ระหว่าง **Total_rooms** และ **Median_house_value**

CONCLUSION

JANE & PANDA & LEENA & BONUS, REAL ESTATE AGENTS

MEMBERS

กมลชนก	บุตะกะ	6624650120
ณัฐชยา	ฤทธิรุณ	6624650237
ธมนวรรณ	ธีร์รัชชัยกุล	6624650260
ประภัสสร	พัศดร	6624650328


THANK YOU
Q&A