«به نام خدا»

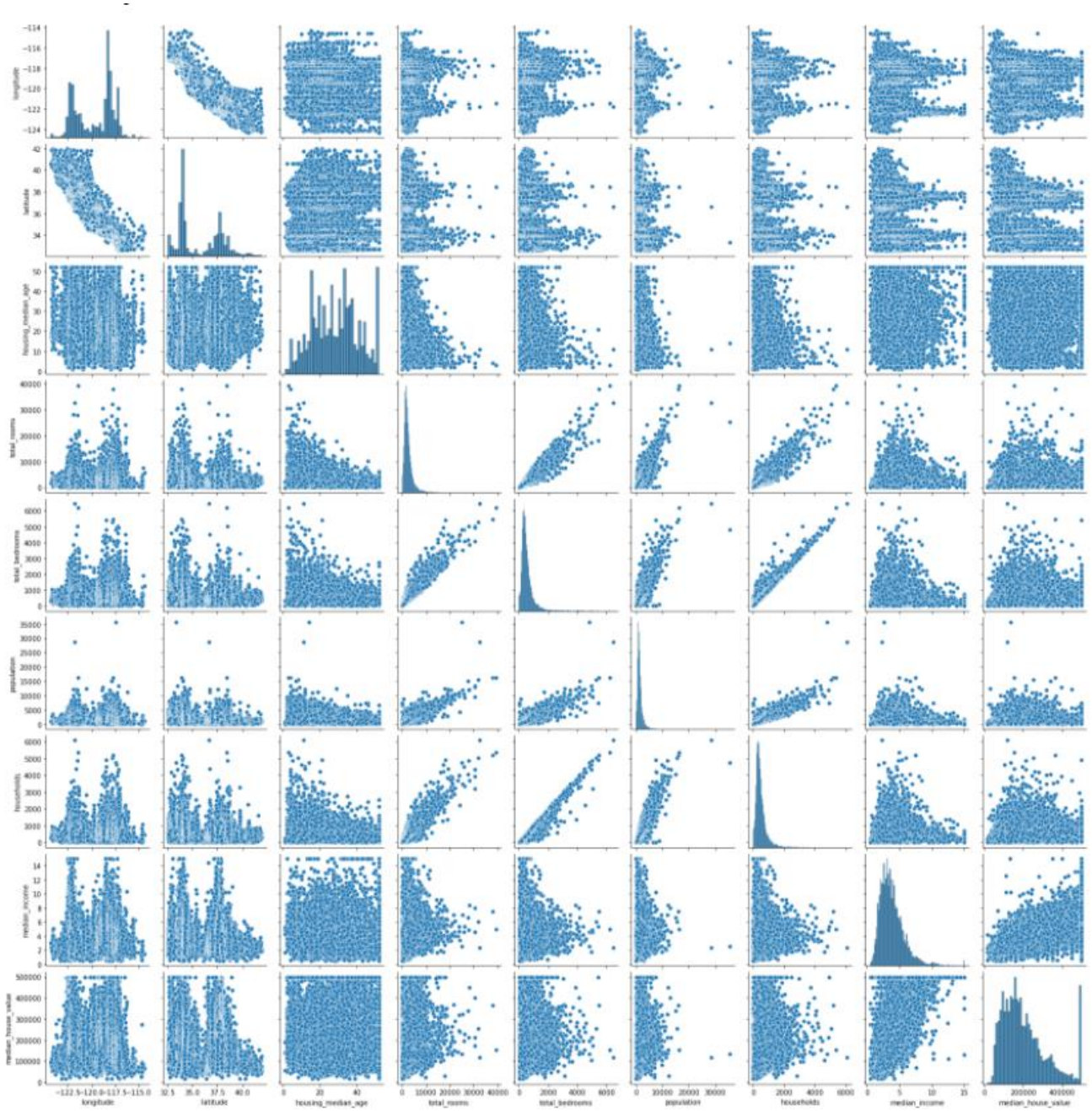
تکلیف سوم – سوال دوم – مرضیه علیدادی – 9631983

(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند.)

.2

a) pairPlot به صـورت پیش فرض، فقط سـتون های عددی را plot می کند. در ادامه می توانیم با اسـتفاده از رنگ کردن، ستون های categorical را هم در این plot در نظر بگیریم.

Pairplot پیش فرض برای این دیتاست، بدین شکل است:



همانطور که مشخص است، نمودار pairPlot دو شکل را برپا ما نمایش می دهد: histogram و scatter plot .

همانطور که می دانیم، histogram در یک نمودار دوبعدی، طریقه ی توزیع یک متغیر واحد را نشـــان می دهد. و scatter plot رابطه ی بین دو متغیر را نشان می دهد.

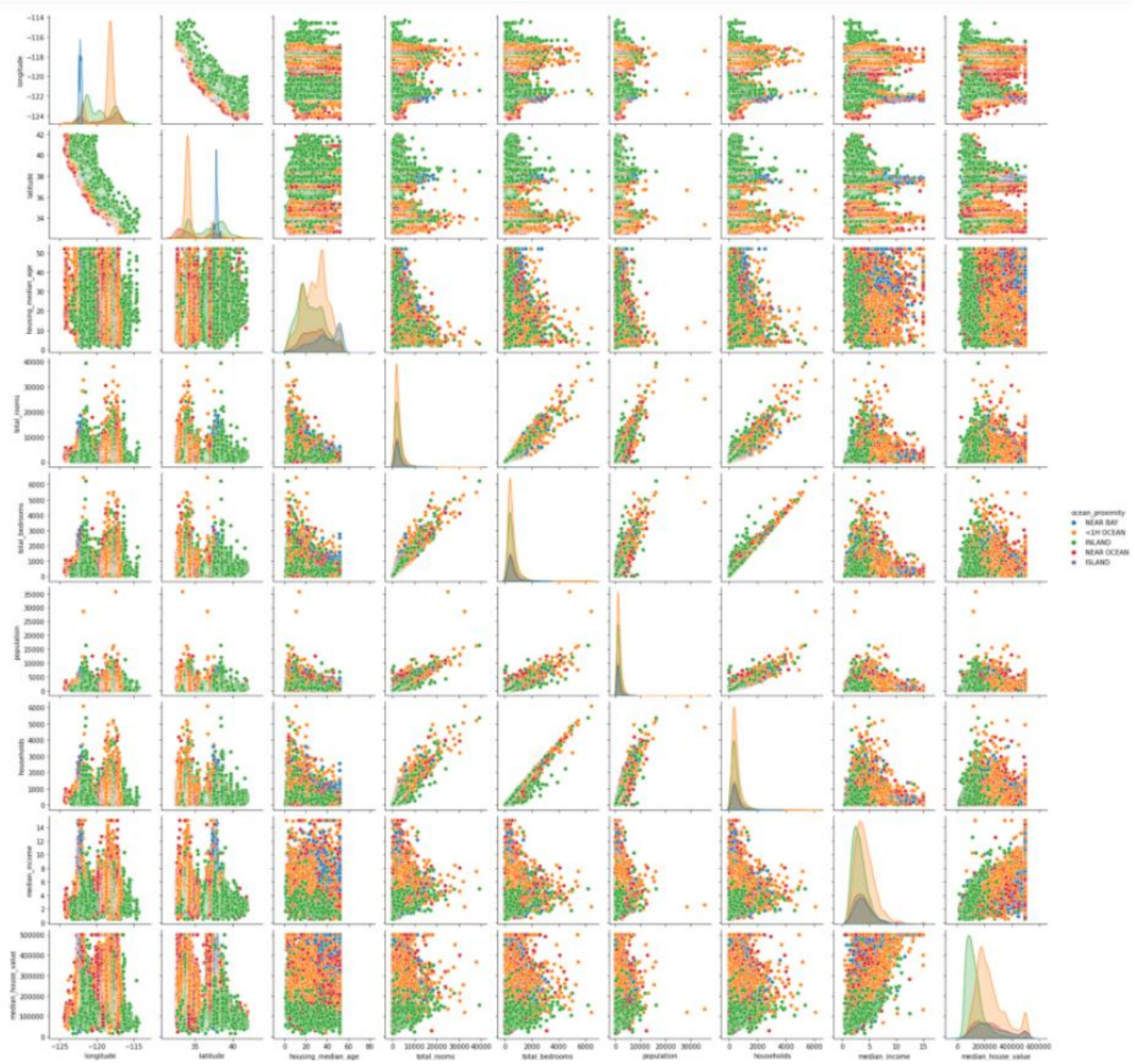
برای مثـال، نمودار دوم از ســـمـت چـپ، در ردیف اول را در نظر می‌گیرم: این نمودار، نمودار scatter plot بین longitude و latitude را نشـان می دهد. این نمودار نشـان می دهد که معمولاً هر چقدر طول جغرافیایی افزایش پیدا می کند، عرض جغرافیایی کاهش می یابد. که این مسئله، در نقشه ی کشیده شده در سوال قبل نیز همخوانی دارد.

به همین ترتیب بقیه ی scatter plot ها نیز تفسیر می شوند.

همچنین برای مثال، می‌توان از نمودار های histogram مربوط به total_bedrooms و total_rooms ، دریافت که این دو متغیر دارای کجی راست هستند.

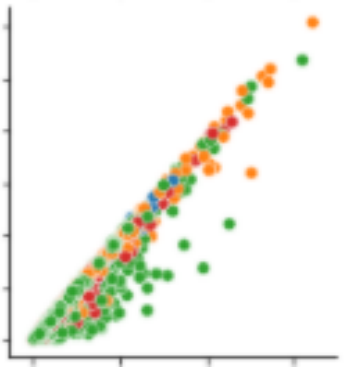
(نمودار های قرار گرفته در طول و عرضِ برابر در این pairPlot ، همان نمودار histogram مربوط به متغیر نظیر آن طول و عرض است. بقیه ی نمودار ها، نمودار های scatter plot میان دو متغیر قرار گرفته در طول و عرض هستند.)

همانطور که قبلا گفتم، با رنگ آمیزی نمودار ها، می توان متغیر categorical را هم در آن ها دخیل کرد. که در این دیتاست، ocean_proximity از این دسته است:



همانطور که مشخص است، همان نمودار های رسم شده در شکل قبل، اینجا با در نظر گرفتن category های مختلف شرکت کننده در متغیر ocean_proximity ، رنگ آمیزی شده اند. که بدین ترتیب، می توان تاثیر دسته های مختلف این متغیر را روی توزیع بقیه ی متغیر ها بررسی کرد. و بدین ترتیب، میتوان رابطه ی بین این دسته ها با بقیه ی متغیر ها را نیز کشف کرد.

b) همانطور که از نمودار pairPlot نظیر میان دو متغیر households و total_bedrooms مشخص است، رابطه ی این دو نزدیک به خط x=y است و نشان می دهد بیشترین ارتباط را با یکدیگر دارند:



به علاوه، رابطه ی میان متغیر ها، با استفاده از قابلیت corr با متد pearson نیز بررسی می شود. که اینجا نزدیکترین عدد به یک، 0.979 است که مربوط به رابطه ی میان همین دو متغیر است:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.000000	-0.924664	-0.108197	0.044568	0.069608	0.099773	0.055310	-0.015176	-0.045967
latitude	-0.924664	1.000000	0.011173	-0.036100	-0.066983	-0.108785	-0.071035	-0.079809	-0.144160
housing_median_age	-0.108197	0.011173	1.000000	-0.361262	-0.320451	-0.296244	-0.302916	-0.119034	0.105623
total_rooms	0.044568	-0.036100	-0.361262	1.000000	0.930380	0.857126	0.918484	0.198050	0.134153
total_bedrooms	0.069608	-0.066983	-0.320451	0.930380	1.000000	0.877747	0.979728	-0.007723	0.049686
population	0.099773	-0.108785	-0.296244	0.857126	0.877747	1.000000	0.907222	0.004834	-0.024650
households	0.055310	-0.071035	-0.302916	0.918484	0.979728	0.907222	1.000000	0.013033	0.065843
median_income	-0.015176	-0.079809	-0.119034	0.198050	-0.007723	0.004834	0.013033	1.000000	0.688075
median_house_value	-0.045967	-0.144160	0.105623	0.134153	0.049686	-0.024650	0.065843	0.688075	1.000000

بنابراین، اینجا با استفاده از هر دوی این روش ها اثبات شد که دو متغیر households و total_bedrooms بیشترین correlation را با یکدیگر دارند.

c) pearsonr(x, y) ، ضریب همبستگی pearson و p-value را برای تست عدم همبستگی محاسبه می کند.

ضریب همبستگی pearson رابطه ی خطی بین دو مجموعه داده را اندازه گیری می کند. به طور دقیق تر، همبستگی pearson نیازمند توزیع نرمال در هر مجموعه داده است. مانند هر ضریب همبستگی دیگریریا، عددی که نتیجه می‌دهد، در فاصله ی 1- تا 1 است. که 0 نشان دهنده ی عدم وجود هیچ همبستگی است. 1 و 1- نشان دهنده ی وجود رابطه ی خطی دقیق است. همبستگی مثبت به این معناست که با افزایش x ، y هم افزایش می یابد. و همبستگی منفی به این معناست که با افزایش x ، y کاهش می یابد.

p-value تقریباً احتمال وجود یک سیستم غیر همبسته برای تولید مجموعه های داده را نشان می دهد، که حداقل به اندازه آنچه از این مجموعه داده ها محاسبه می شود، همبستگی pearson دارند.

ضریب همبستگی، با این فرمول محاسبه می شود:

$$r_{pb} = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2(y - m_y)^2}}$$

```
r = \frac{\sum (x - m_x) (y - m_y)}
        {\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}
```

که m_x در آن میانگین وکتور x است. و m_y در آن میانگین وکتور y است.

d) خروجی بدین شکل شد:

ضریب همبستگی pearson : 0.688 (همانند عدد محاسبه شده در بخش b) و p-value : 0.0

e) -0.3571622692099669

f) 0.688075207958548 (همان عدد نظیر به دست آمده در بخش های b و d)

g) در این نمودار، عددی که با استفاده از corr برای نمایش همبستگی میان متغیر ها محاسبه می شود، به یک طیف رنگی نظیر شده است. که در کنار نمودار، عدد متناظر هر رنگ نمایش داده شده است. مثلا رنگ های نزدیک به سفید، نزدیک به عدد 1 هستند. و رنگ های نزدیک به سیاه، نزدیک به عدد 1- هستند. و هرچه رنگ ها به میانه ی این طیف نزدیک تر باشند، به عدد 0 نزدیک هستند.