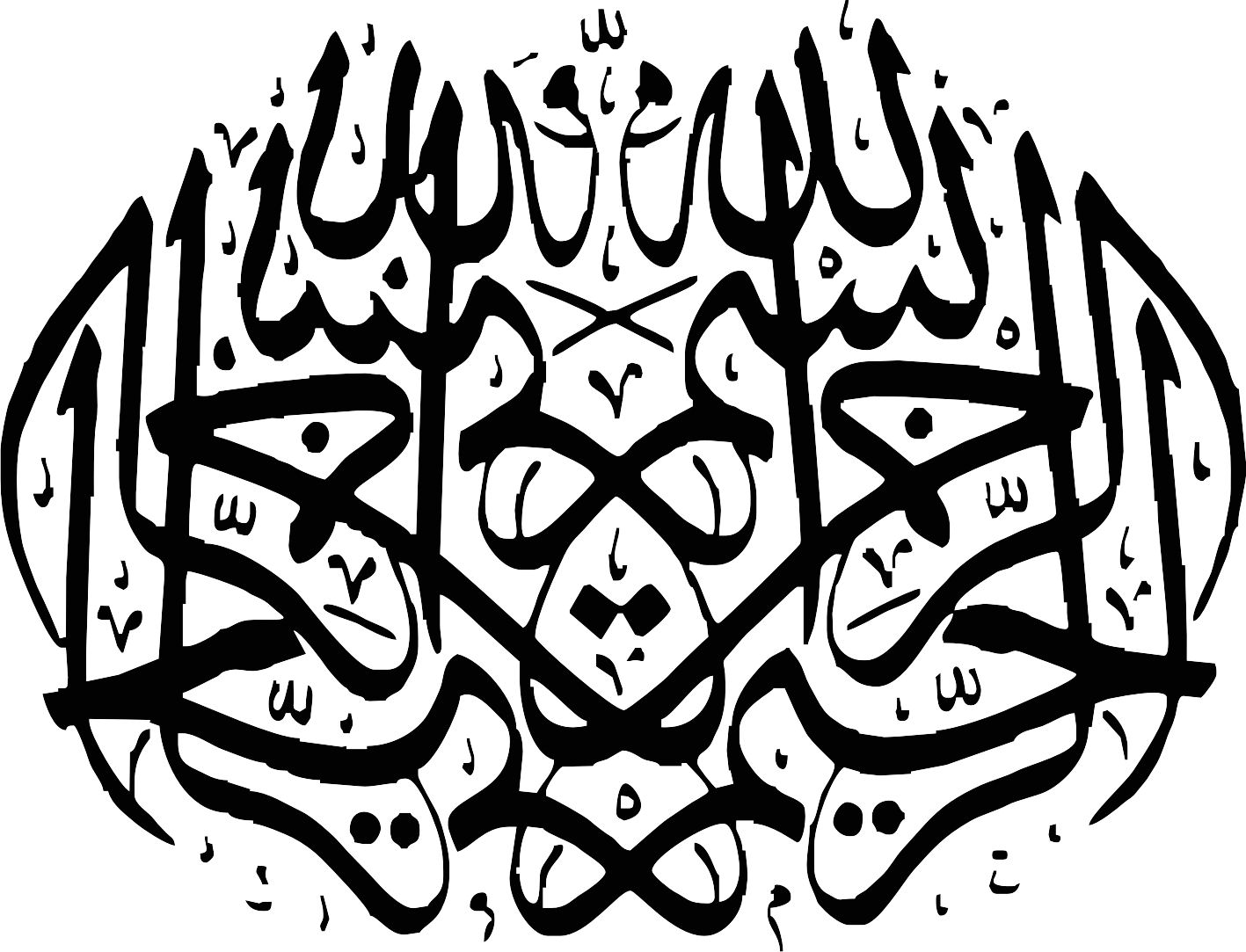


**گزارش تکلیف دوم درس یادگیری ماشین کاربردی**

**استاد درس: دکتر ناظرفرد**

تهیه کننده: سید نیما محمودیان

شماره دانشجویی: 402125005



فهرست مطالب

[1-پاسخ سوال اول 1](#_Toc163208856)

[1-1-لود کردن دیتاست، نمایش ده ردیف به صورت تصادفی، و بررسی کلی داده‌ها 1](#_Toc163208857)

[2-1-مدیریت مقادیر گمشده 5](#_Toc163208858)

[3-1-مصور سازی داده‌ها 8](#_Toc163208859)

[4-1-مدیریت داده‌های پرت 14](#_Toc163208860)

[5-1-مهندسی ویژگی‌ها 16](#_Toc163208861)

[2-پیش‌پردازش تصویر 18](#_Toc163208862)

[1-2- gray scale کردن عکس‌ها 20](#_Toc163208863)

[2-2-تنظیم روشنایی و کنتراست تصاویر 21](#_Toc163208864)

[3-2-نرمال سازی تصاویر 23](#_Toc163208865)

[3-پیش‌پردازش متن 25](#_Toc163208866)

[1-3-خواندن داده‌ها از پیکره همشهری 25](#_Toc163208867)

[2-3-پیش‌پردازش متن‌ها 29](#_Toc163208868)

[3-3-TF-IDF 32](#_Toc163208869)

[4-3-مصورسازی داده‌ها 33](#_Toc163208870)

[4-پیوست 40](#_Toc163208871)

فهرست شکل‌ها

[شکل 1: وارد کردن کتابخانه‌های مورد نیاز 1](#_Toc163166210)

[شکل 2: حذف ستون‌های اضافه 1](#_Toc163166211)

[شکل 3: خواندن دیتاست و نمایش ده سطر تصادفی 2](#_Toc163166212)

[شکل 4: بررسی مجموع مقادیر ناموجود در هر ستون 2](#_Toc163166213)

[شکل 5: اطلاعات نمایش داده شده توسط متد info() 3](#_Toc163166214)

[شکل 6: تعداد مقادیر موجود در هر یک از متغیرهای دسته‌ای 3](#_Toc163166215)

[شکل 7: دسته‌بندی مجدد ستون nationality 4](#_Toc163166216)

[شکل 8: دسته‌بندی مجدد ستون edu 5](#_Toc163166217)

[شکل 9: جایگذاری ؟ با np.nan 6](#_Toc163166218)

[شکل 10: train test split 6](#_Toc163166219)

[شکل 11: پایپ‌لاین پیش پردازش داده‌های دسته‌ای 6](#_Toc163166220)

[شکل 12: نحوه تعریف کلاس ColumnTransformer() 7](#_Toc163166221)

[شکل 13: تعریف مدل و پایپ‌لاین اصلی 7](#_Toc163166222)

[شکل 14: شمای کلی پایپ‌لاین 8](#_Toc163166223)

[شکل 15:معیارسنجی مدل 8](#_Toc163166224)

[شکل 16: حذف سطرهای دارای مقادیر خالی 8](#_Toc163166225)

[شکل 17: کد رسم نمودار ستونی 9](#_Toc163166226)

[شکل 18: کد مربوط به نمودار جعبه‌ای 9](#_Toc163166227)

[شکل 19: نمودارهای ستونی رسم شده از داده‌های دسته‌ای 10](#_Toc163166228)

[شکل 20: نمودار جعبه‌ای از داده‌های عددی 11](#_Toc163166229)

[شکل 21: کد مربوط به هیستوگرام 11](#_Toc163166230)

[شکل 22: هیستوگرام‌های رسم شده 11](#_Toc163166231)

[شکل 23: نحوه انتخاب ستون‌های عددی و محاسبه ماتریس همبستگی 12](#_Toc163166232)

[شکل 24: هیت‌مپ رسم شده 12](#_Toc163166233)

[شکل 25: نحوه محاسبه همبستگی با ستون هدف 13](#_Toc163166234)

[شکل 26: نمودار میله‌ای همبستگی با ستون هدف 13](#_Toc163166235)

[شکل 27: نحوه محاسبه هیت‌مپ برای داده‌های دسته‌ای 14](#_Toc163166236)

[شکل 28: هیت‌مپ ایجاد شده از ستون‌های دسته‌ای 15](#_Toc163166237)

[شکل 29: نحوه حذف سطرهایی که داده پرت دارند 15](#_Toc163166238)

[شکل 30: one hot encoding 16](#_Toc163166239)

[شکل 31: انتقال ستون هدف به انتهای دیتافریم 17](#_Toc163166240)

[شکل 32: پایپ‌لاین ارزیابی روش PCA 17](#_Toc163166241)

[شکل 33: ایجاد کلاس و پایپ‌لاین برای mutual information 18](#_Toc163166242)

[شکل 34: شمای کلی پایپ‌لاین ایجاد شده برای mutual information 18](#_Toc163166243)

[شکل 35: وارد کردن کتابخانه‌ها و تولید سه عدد تصادفی 19](#_Toc163166244)

[شکل 36: لود کردن تصاویر و ذخیره آنها در متغیرها 19](#_Toc163166245)

[شکل 37: نمایش ابعاد عکس‌ها 19](#_Toc163166246)

[شکل 38: تبدیل عکس‌ها به rgb و نمایش آنها 20](#_Toc163166247)

[شکل 39: gary scale کردن عکس‌ها 22](#_Toc163166248)

[شکل 40: تابع تنظیم کنتراست و روشنایی 23](#_Toc163166249)

[شکل 41: تنظیم روشنایی و کنتراست شکل اول 24](#_Toc163166250)

[شکل 42: تابع نرمال کننده عکس 25](#_Toc163166251)

[شکل 43: نرمال کردن عکس با استفاده از تابع 25](#_Toc163166252)

[شکل 44: کتابخانه‌های استفاده شده در سوال سوم 26](#_Toc163166253)

[شکل 45: جداسازی رکوردها و قرار دادن آنها در چهار فیلد 26](#_Toc163166254)

[شکل 46: تابع تقسیم کننده فایل به یازده بخش 27](#_Toc163166255)

[شکل 47: تمیز کردن متن‌ها و ذخیره آنها در اکسل 28](#_Toc163166256)

[شکل 48: تبدیل فایل‌های اکسل به یک دیتافریم 28](#_Toc163166257)

[شکل 49: پیدا کردن نحوه انکودینگ متن 29](#_Toc163166258)

[شکل 50: حذف و جایگذینی کاراکترها و type casting 29](#_Toc163166259)

[شکل 51: حذف علائم نگارشی 30](#_Toc163166260)

[شکل 52: نحوه حذف اعداد 30](#_Toc163166261)

[شکل 53: توکنایز کردن متن 31](#_Toc163166262)

[شکل 54: حذف stop wordsها 31](#_Toc163166263)

[شکل 55: نمایش 5 توکنی که بیشتری استفاده را داشته‌اند 31](#_Toc163166264)

[شکل 56: تبدیل توکن‌ها به متن و نرمال سازی متن 32](#_Toc163166265)

[شکل 57: اجرای TF-IDF 32](#_Toc163166266)

[شکل 58: نمایش 5 کلمه مهم هر متن 33](#_Toc163166267)

[شکل 59: نمایش 5 کلمه مهم در همه متون 33](#_Toc163166268)

[شکل 60: فرایند ایجاد wordcloud 34](#_Toc163166269)

[شکل 61: ابر کلمه تولید شده 35](#_Toc163166270)

[شکل 62: انجام تحلیل عواطف با استفاده از کتابخانه polyglot 36](#_Toc163166271)

[شکل 63: نمودار سری زمانی 37](#_Toc163166272)

[شکل 64: نمودار میله‌ای تعداد متن‌های مثبت، خنثی و منفی 37](#_Toc163166273)

[شکل 65: نمودار سهم هر دسته از قطبیت‌ها از کل متون 38](#_Toc163166274)

[شکل 66: هسیتوگرام میزان قطبیت 38](#_Toc163166275)

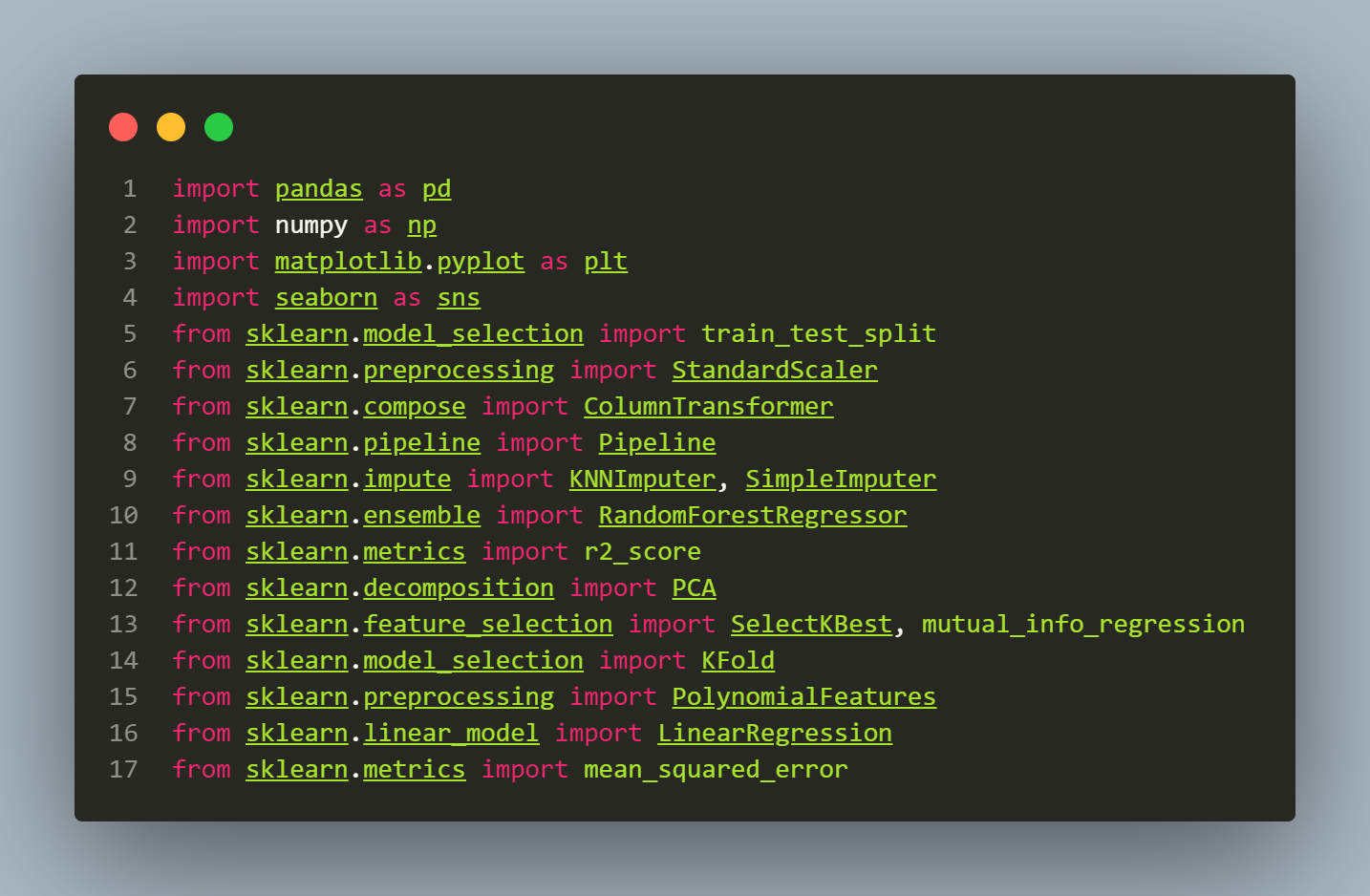
[شکل 67: نمودار تعداد اخبار مثبت، خنثی و منفی در طول زمان 39](#_Toc163166276)

[شکل 68: نمودار میله‌ای تعداد متون در بیست دسته اول موضوعات 40](#_Toc163166277)

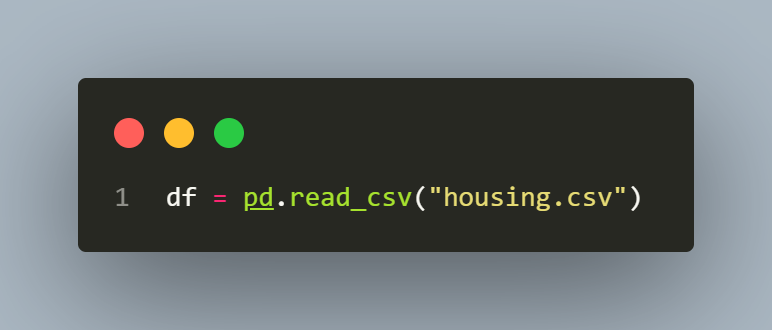
# 1-پاسخ سوال اول

## 1-1-لود کردن دیتاست، نمایش ده ردیف به صورت تصادفی، و بررسی کلی داده‌ها

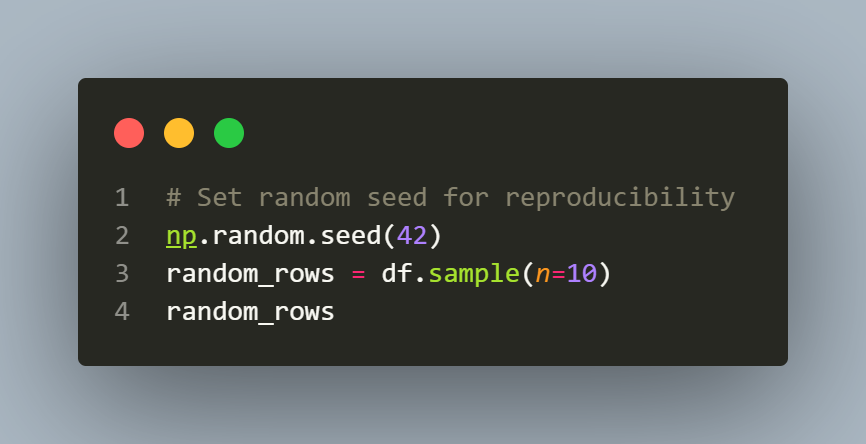
توجه شود که هرجا احتیاجی به ایپمورت کردن یک کتابخانه بود، برای حفظ نظم، ایمپورت در اولین بلوک کد انجام شده است شکل یک بلوک مربوط به وارد کردن کتابخانه‌ها را نمایش می‌دهد. برای لود کردن دیتاست، و نمایش ده ردیف رندوم از کتابخانه pandas استفاده می‌کنیم. با استفاده از کتابخانه numpy یک random seed ایجاد می‌کنیم. Random seed به مظور بازتولید نتایج مشابه تعریف می‌شود. شکل (2)، خواندن دیتاست و شکل (3) نشان دادن ده ردیف رندوم را نمایش می‌دهد.



شکل 1: وارد کردن کتابخانه‌های مورد نیاز

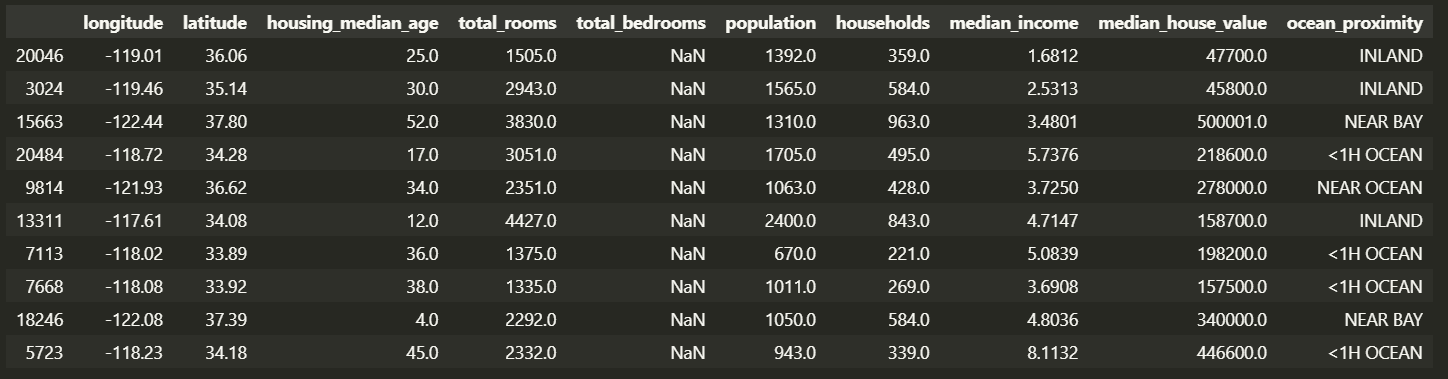


شکل 2: خواندن دیتاست



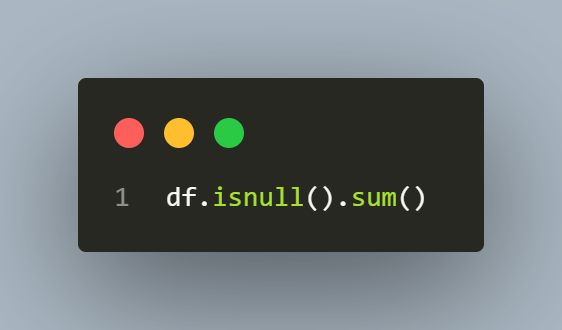
شکل 3: کد مربوط به نمایش ده سطر تصادفی

همچنین شکل (4) ده سطر تصادفی نمایش داده شده را نشان می‌دهد.



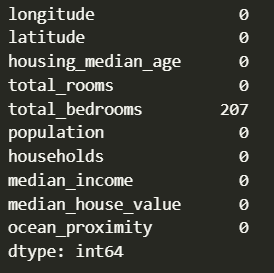
شکل 4: ده سطر تصادفی نمایش داده شده

در ادامه با استفاده از متد info() نگاهی کلی به ستون‌های باقی‌مانده می‌اندازیم. سپس با استفاده از method chaining نمایش داده شده در شکل (5) بررسی می‌کنیم که در هر ستون در مجموع چه تعداد مقادیر تعریف نشده وجود دارد.



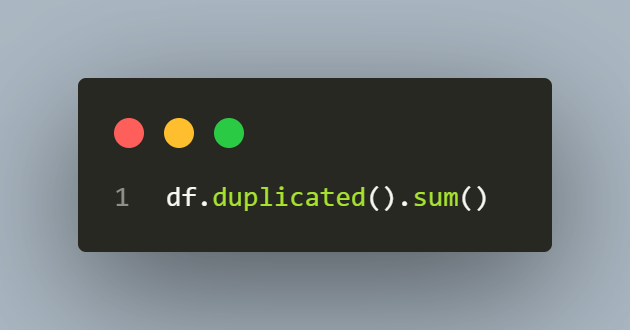
شکل 5: کد مربوط به نشان دادن تعداد مقادیر گمشده

با توجه به شکل (6) مشاهده می‌شود که ستون total\_bedrooms دارای مقادیر خالی است که در ادامه به مدیریت آنها خواهیم پرداخت.

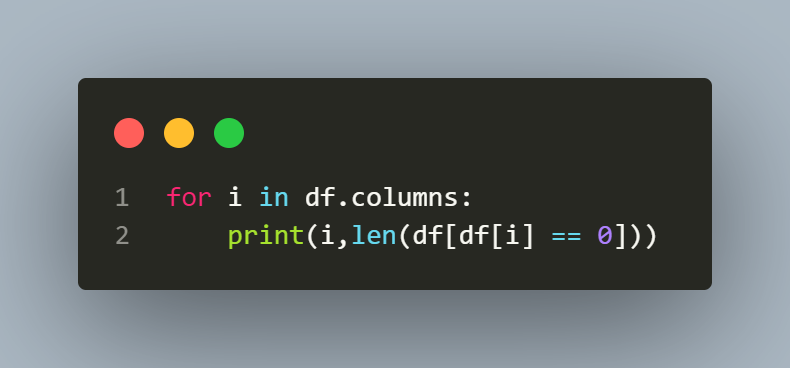


شکل 6: تعداد مقادیر خالی در هر سطر

در ادامه سه مورد را بررسی می‌کنیم: اول اینکه چه تعداد سطر تکراری داریم. شکل (7) نحوه انجام این کار را نمایش می‌دهد. سپس بررسی می‌کنیم که آیا در ستون‌هایی که مقدار صفر معنایی ندارد، رکوردی با مقدار صفر داریم یا خیر. شکل (8) نحوه انجام این کار را نمایش می‌دهد.



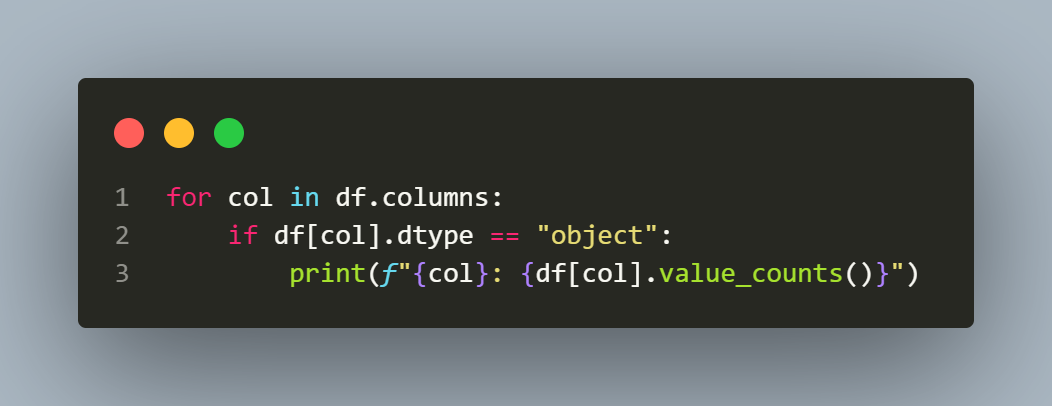
شکل 7: کد بررسی تعداد سطرهای تکراری



شکل 8: کد بررسی تعداد صفرها در هر ستون

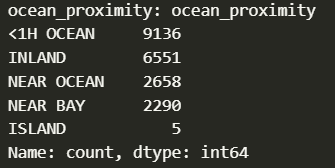
با بررسی خروجی این دو کد، متوجه می‌شویم که هیچ سطر تکراری در دیتافریم وجود ندارد. همچنین در هیچ یک از ستون‌ها مقدار صفر نداریم.

در قدم بعدی بررسی می‌کنیم که در سطرهایی که نوع داده‌ی دسته‌ای دارد، چه کلاس‌هایی موجود داریم و جمعیت هر دسته چقدر است.



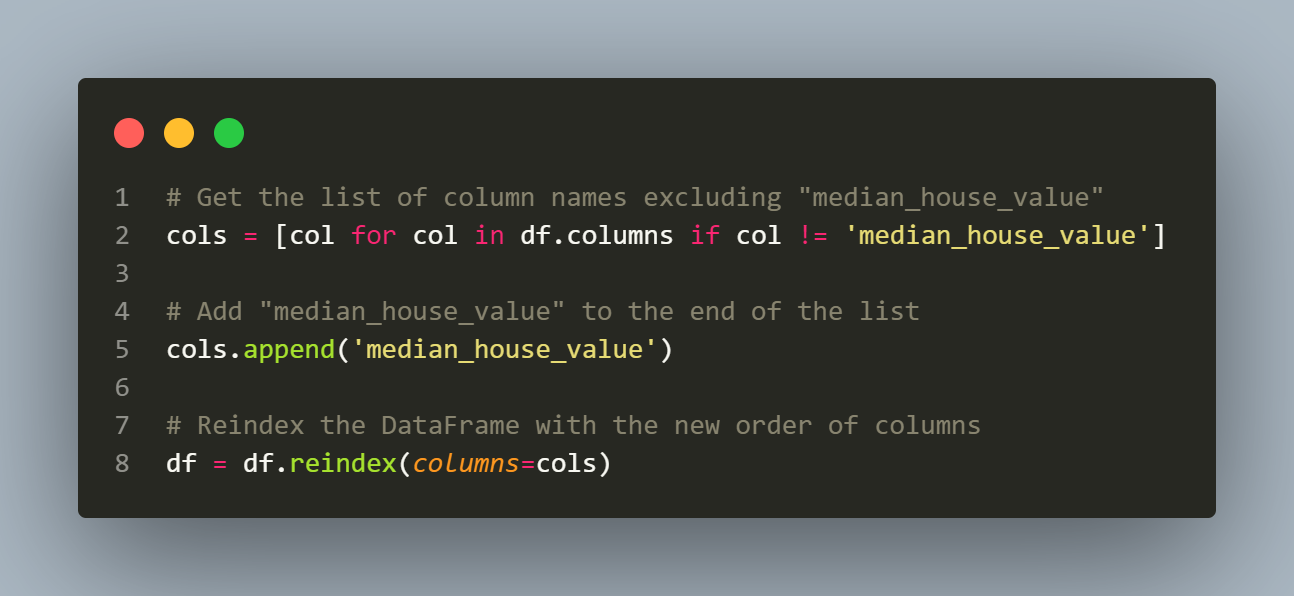
شکل 9: بررسی تعداد و جمعیت موجود در سطرهای گسسته

شکل (10) خروجی این کد را نمایش می‌دهد. مشاهده می‌شود که دسته‌ای با نام غیر معمول در ستون ocean\_proximity وجود ندارد.



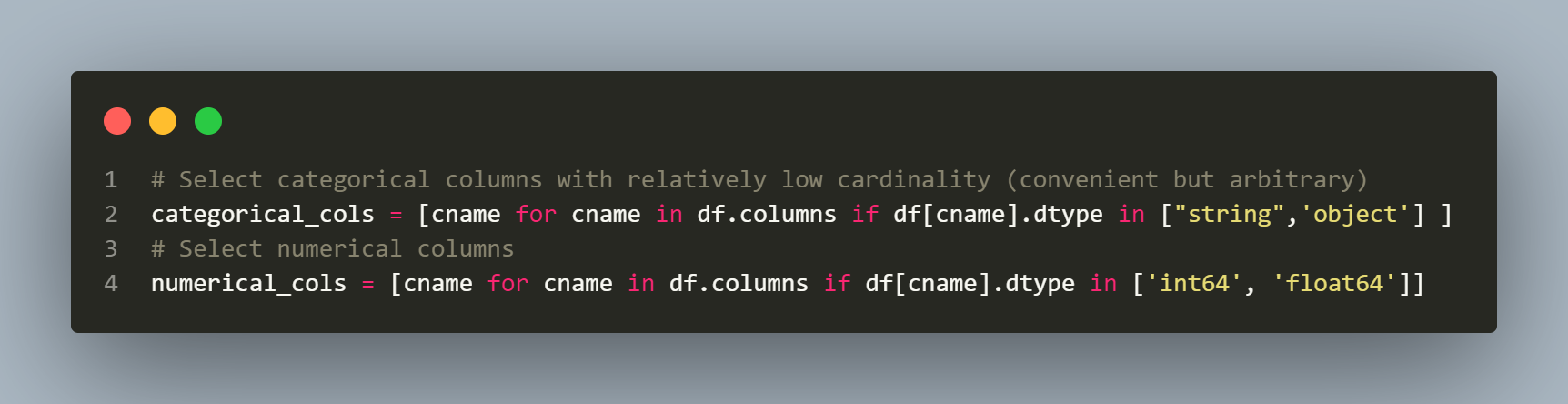
شکل 10: خروجی کد شکل 9

حال برای حفظ قرارداد، سطر هدف را به سطر آخر دیتافریم انتقال می‌دهیم. شکل (11) کدی که این کار را بر عهده دارد نمایش می‌دهد.



شکل 11: انتقال سطر هدف به سطر آخر

حال یک تقسیم‌بندی انجام می‌دهیم که در ادامه‌ی کار مفید واقع می‌شود. با استفاده از List comprehension نام ستون‌هایی که داده‌های عددی دارند را در یک لیست و نام ستون‌هایی که دارای مقادیر دسته‌ای هستند را در یک لیست دیگر قرار می‌دهیم. شکل (12) نحوه انجام این کار را نمایش می‌دهد.



شکل 12: جداسازی داده‌های عددی و دسته‌ای در دو لیست جدا

## 2-1-مصور سازی داده‌ها

### 1-2-1-رسم و تفسیر هیستوگرام

شکل 13 نحوه رسم هیستوگرام را نمایش می‌دهد. num\_rows و num\_cols ابعاد شبکه فرعی را مشخص می کنند. در این مورد، یک شبکه سه در سه ایجاد می کند، بنابراین می توان تا 9 نمودار فرعی را در خود جای داد. تابع plt.subplots برای تولید یک شکل و مجموعه ای از نمودارهای فرعی (محور) استفاده می شود. (20,10) figsize اندازه کل شکل (20 اینچ عرض و 10 اینچ ارتفاع) را مشخص می کند.

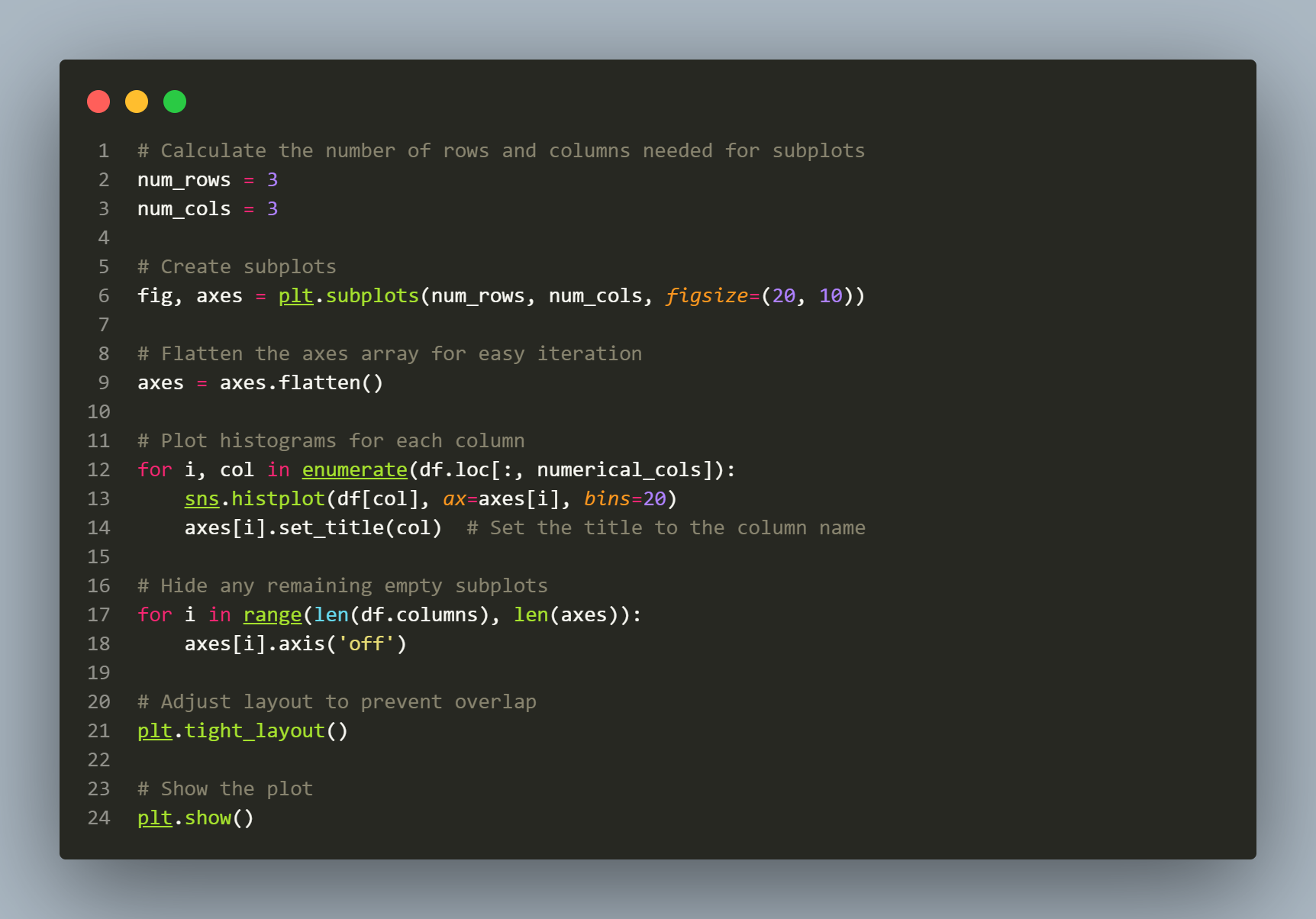
سپس حلقه روی هر ستون مشخص شده در لیست numerical\_cols از df تکرار می شود. برای هر ستون:

sns.histplot یک هیستوگرام از ستون ایجاد می کند و آن را بر روی نمودار فرعی مربوطه ترسیم می کند bins=20 مشخص می کند که هر هیستوگرام باید 20 میله داشته باشد. سپس عنوان هر نمودار بالای آن مشخص می‌شود.

سپس لوپ بعدی هر فضای خالی که در آن نمودار وجود ندارد را حذف می‌کند.

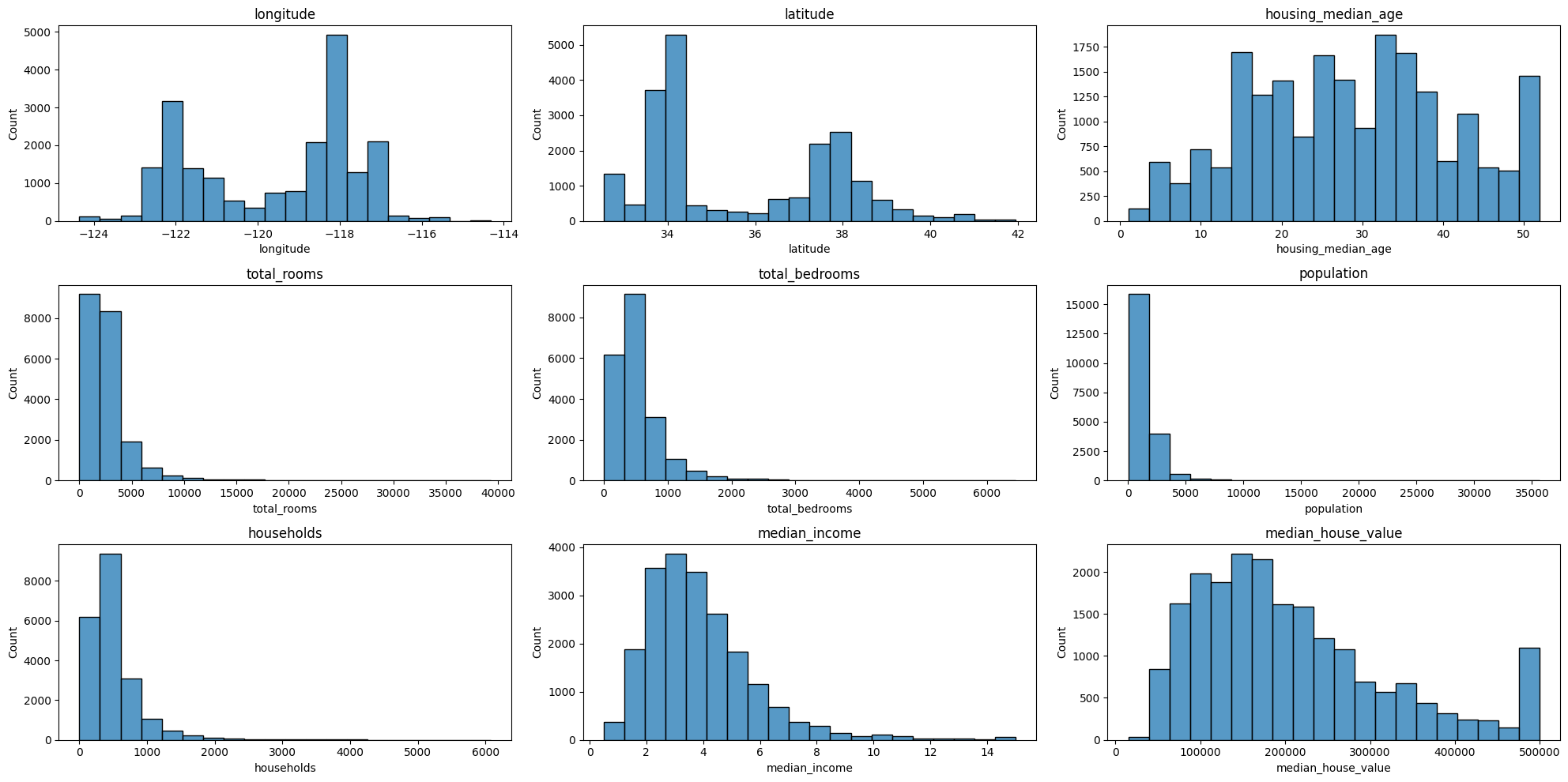
()plt.tight\_layout، نمودارهای فرعی و برچسب ها را طوری تنظیم می کند که به خوبی در ناحیه شکل قرار گیرند و روی هم قرار نگیرند.

plt.show () شکل کامل را با تمام نمودارهای فرعی نمایش می دهد.



شکل 13: کد مربوط به رسم هیستوگرام

همچنین شکل (14) هیستوگرام‌های رسم شده را نمایش می‌دهد.



شکل 14: هیستوگرام‌های رسم شده

**تفسیر:**

**Longitude:**

توزیع چند وجهی است، که نشان دهنده چندین قله است که در آن داده های مسکن به صورت طولی متمرکز شده اند.

این نشان می دهد که خوشه های متعددی از خانه ها در طول های جغرافیایی مختلف، احتمالاً مربوط به شهرها یا مناطق مختلف است.

Latitude:

مشابه طول جغرافیایی، توزیع چند وجهی با تمرکز داده های مسکن در عرض های جغرافیایی خاص است. این خوشه ها را در مناطق جغرافیایی خاصی نشان می دهد، که وقتی با طول جغرافیایی ترکیب می شوند، می توانند مناطق جغرافیایی خاصی را مشخص کنند.