تمرین اول درس یادگیری ماشین (آشنایی با کتابخانه های پایه)

برای انجام تمرینات درس یادگیری ماشین شما نیازمند آشنایی با مهارت های زیر هستید:

- بارگذاری و دستکاری داده ها(به وسیله pandas و numpy)
- رسم نمودارها و تصویر پردازی(به وسیله matplotlib)
- محاسبات بر روی داده های ماتریسی n بعدی(به وسیله numpy)

هدف از این تمرین آشنایی با کتابخانه های بالا و استفاده از متدهای پر کاربرد در حل مسائل یادگیری ماشین می‌باشد.

خلاصه تمرین

در این تمرین در ابتدا دیتاست frcities که شامل اطلاعات مهمی از شهرهای فرانسه است را بارگذاری میکنیم.سپس موقعیت شهرها را با توجه به ستون های lat و lng (طول و عرض جغرافیایی) رسم میکنیم. با استفاده از کتابخانه های numpy و pandas تغییراتی بر روی داده ها انجام داده و در نهایت میزان همبستگی میان ویژگی ها را محاسبه میکنیم.نظر به آنکه توانایی لازم برای حل مسئله را داشته باشید پیشنهاد می‌شود حتما داکیومنت کتابخانه ها را مطالعه فرمایید.

بخش اول

۱) دیتاست frcities.csv را با استفاده از کتابخانه pandas در محیط پایتون بارگذاری کنید.

برای راهنمایی به این لینک مراجعه کنید

*فایل مجموعه داده همران تمرین آپلود شده است

پس از بارگذاری دیتاست با استفاده از دستور shape تعداد نمونه ها و ستون ها را بررسی کنید.

۲) در محاسبات آینده وجود سطر یا ستونی با مقدار null یا duplicate مشکل ایجاد میکند. همانطور که در خلاصه ای از دیتاست زیر مشاهده می‌کنیم در ستون population مقادیر null وجود دارد. این سطرها را حذف کنید.

	city	lat	lng	iso2	density	population	ranking
0	Saint-Oblas	45.5674	5.0447	FR	129.2	NaN	4
1	Louresse	47.2394	-0.3136	FR	33.8	872.0	3
2	Olmet	45.7100	3.6614	FR	10.4	161.0	3
3	Olmet	44.9542	2.6108	FR	71.0	NaN	4
4	Gottenhouse	48.7208	7.3611	FR	305.6	382.0	3
...	...	...	...	...	...	...	...
59059	La Rochette	45.2609	6.2881	FR	55.5	NaN	4
59060	La Rochette	45.3056	3.4747	FR	14.4	NaN	4
59061	Saint-Eutrope	45.4181	0.1114	FR	62.9	168.0	3
59062	Saint-Eutrope	44.4535	0.5204	FR	34.2	NaN	4
59063	Taillis	48.1889	-1.2389	FR	81.2	996.0	3

توضیح دهید که دستور زیر چه عملی را انجام میدهد و آرگومان های آن‌را شرح دهید:

```
dataset = dataset.drop_duplicates(subset=['city'],keep='last')
```

بخش دوم : تصویرپردازی داده ها

۳) در این قسمت با استفاده از کتابخانه matplotlib داده های موجود در دیتاست را بر اساس موقعیت جغرافیایی(lat و lng) رسم کنید برای آشنایی با این کتابخانه به لینک مراجعه کنید.
- محور افقی را lat در نظر بگیرید.
- از Scatter استفاده کنید.

۴) مشخصات جغرافیایی که داده شده بر اساس مایل آمده است. ابتدا مشخصات را به کیلومتر تبدیل کرده و به دیتاست بعنوان ۲ ستون جدید lat_km و lng_km اضافه کنید (راهنما: از متد apply() استفاده کنید).
- هر واحد lat معادل ۸۷ کیلومتر و هر واحد lng معادل ۱۱۰ کیلومتر در نظر گرفته شود.
- با مقادیر بدست آمده مجددا داده ها را رسم کنید.

بخش سوم :محاسبات

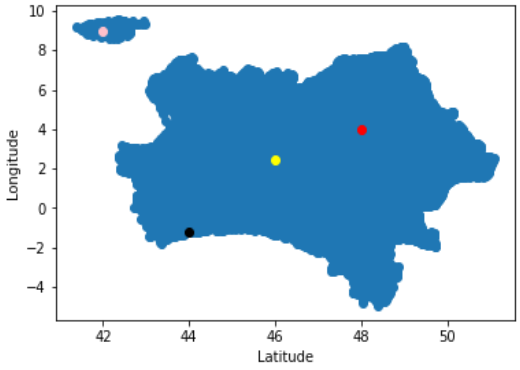
در این قسمت با توجه به جدول زیر نقطه مناسب را انتخاب کرده و نقاط در شعاع ۱۵۰ کیلومتری مرکز مختصاتی را جدا کنید.سپس با استفاده از ستون های population و density میزان تراکم جمعیت را برای داده ها محاسبه کنید:

۵) برای راحتی کار بهتر است داده ها را به فرم آرایه numpy تبدیل کنید. همچنین مرکز مختصاتی را هم به کیلومتر تبدیل کنید.

۶) تابعی برای پیدا کردن شهرها در شعاع ۱۵۰ کیلومتری بنویسید.

۷) ذخیره داده های بدست آمده بصورت دیتافریم.

۸) محاسبه تراکم جمعیت برای داده هایی که در مرحله قبل بدست آوردیم (در دیتاست بعنوان pop/dens اضافه شود).



longitude	latitude	شماره دانشجویی
4	48	۰–۲
-1.2	44	۳–۵
2.44	46	۶–۸
9	42	۹

معادله دایره:

$$(x-h)^2+(y-k)^2=r^2$$

x = x coordinate of circle point
h = x coordinate of center point
y = y coordinate of circle point
k = y coordinate of center point
r = radius of the circle

بخش چهارم

در این بخش به بررسی ضریب همبستگی میان ویژگی ها(ستون ها) میپردازیم. برای این منظور ۳ نوع همبستگی pearson, spearman,Kendal tau را از کتابخانه scipy در نظر میگیریم.میزان ضریب همبستگی میان population و density و همینطور pop/dens و lat را محاسبه کرده و تحلیل خود را بیان کنید. برای راهنمایی بیشتر به این لینک مراجعه کنید.