Amirkabir University of Technology
(Tehran Polytechnic)

Electrical Engineering Department

Data Mining – Dr.Amir Mazlaghani

Report Exp 1

نام دانشجو: علی بابالو – ۹۸۳۱۳۲۲

ایمیل : alibabaloo@aut.ac.ir

برای تمرین عملی اول از کتابخانه های pandasa, numpy, matplotlib, math استفاده می‌کنیم. برای خواندن دیتاست از دستور read_csv استفاده می‌کنیم و دیتافریم را نمایش می‌دهیم. برای اینکه ستون های خواسته شده را حذف کنیم از دستور drop استفاده کرده و نام ستون هارا مشخص می‌کنیم. برای عوض کردن نام ستون ها نیز از دستور rename استفاده کرده و نام ستون و نام اصلاح شده آن را بعنوان یک دیکشنری به تابع می‌دهیم. سپس در دیتافریم شهرهایی که جمعیت آنها از یک میلیون کمترند را حذف می‌کنیم.

با دستور drop_duplicates می‌توان سطر های تکراری را حذف کرد و برای حذف سطرهایی که بیش از یک آیتم NaN دارند بازهم از دستور drop استفاده می‌کنیم اما ترشهولد آن را یکی کمتر از تعداد ستون ها در نظر می‌گیریم که یعنی سطر هایی که صفر یا یک داده از دست رفته دارند ولید هستند و آنها را نگه دار و باقی که بیش از ۲ داده از دست رفته دارند را دراپ کن.

برای پر کردن داده های از دست رفته از groupby استفاده می‌کنیم که بتوان یک عملیات را روی دیتاقریم انجام داد سپس پس از مشخص کردن محل انجام اوپریشن خود عملیان را مشخص می‌کنیم که ان عبارت است از ترنسفورم کردن داده هایی که Nan هستند با مقدار میانگین آن ستون که برای آن از یم تابع لامبدا استفاده می‌کنیم.

برای محاسبه فاصله هاورسین یک تابع به نام haversine می‌نویسیم که مقادیر عرض و طول جغرافیایی تهران و شهرهای دیگر را بعنوان ورودی بگیرد، سپس آنها را با استفاده از تابع map به رادیان تبدیل می‌کنیم.سپس با توجه به رابطه داده شده مقدار فاصله را حساب کرده و آن را ریترن می‌کنیم. سپس برای اضافه کردن آن مقادیر به دیتافریم یک ستون جدید به نام Distance_from_tehran ایجاد می‌کنیم که برای محاسبه مقادیر آن از تابع اپلای استفاده کرده و مقادیر هر سطر به همراه مفادیر مربوط به تهران را با استفاده از یک لامبدا فانکشن به تابع Haversine می‌دهیم.

برای سورت کردن از تابع sort_values استفاده می‌کنیم و در ارگومان تابع مشخص می‌کنیم که دیتا فریم را بر اساس کدام ستون سورت کند (در اینجا city و lat).

برای سیو کردن دیتافریم نهایی به فایل csv از دستور to_csv استفاده می‌کنیم که محل و نام ذخیره شده فایل را بعنوان ورودی با تابع می‌دهیم تا آن را ذخیره کند.

برای قسمت پلات کردن داده ها ابتدا دیتافریم را بر اساس Distance_from_tehran سورت می‌کنیم و ۱۰ داده اول آن را بر می‌داریم و سپس bar chart را برحسب شهر و فاصله آن از تهران نمایش می‌دهیم. برای جمعیت آن شهر ها نیز همین کار را تکرار می‌کنیم و نمودار را بر اساس جمعیت و نام شهر ها نمایش می‌دهیم. (چون خود تهران نیز در این نمودار ها ظاهر می‌شود – چون فاصله تهران از خودش ۰ است پس نزدیک ترین محسوب می‌شود – یک نمودار دیگر نیز از شهر های سورت شده ۱ تا ۱۱ کشیده شده است که ۱۰ شهر نزدیک به تهران اند اگر خود تهران را در نظر نگیریم)

برای قسمت آخر نیز از دستور plt.scatter استفاده می‌کنیم که یک اسکتر پلات برحسب طول و عرض جغرافیایی نمایش دهد که شکا خروجی شبیه نقشه جهان است.

پایان