

**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**