



تمارين درس داده کاوی

## تمرین ۲: انبار داده

استاد: دکتر حسین رحمانی  
فروردین ۱۴۰۰

## راهنمای تمرین

- مهلت ارسال تمرین تا ساعت ۲۳:۵۹ تاریخ ۱۴۰۰/۰۲/۰۴ است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر ۲۵ درصد از نمره تمرین کسر خواهد شد.
- پاسخ به سوالات این تمرین باید در قالب یک گزارش با فرمت PDF ارائه شود.
- به همراه فایل گزارش تمرین، فایل کدهای اجراشده نیز پیوست شود.
- تمامی فایل‌های این تمرین (گزارش و کدها) در قالب یک فایل فشرده (rar یا zip) با نام‌گذاری زیر ارسال شود.

**StudentNumber\_FirstName\_LastName\_HW2.zip**

- رعایت نکات نگارشی در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این تمرین حتما باید از زبان برنامه‌نویسی پایتون استفاده شود.

در این مجموعه داده قصد داریم بر روی نمونه ۲ درصدی از داده های رفاهی مردم ایران کار کنیم. این مجموعه داده از لینک زیر قابل دسترس است.

<http://mashghema.ir/500000FamilySample-990402.rar>

**باتوجه به مجموعه داده رفاه ایرانیان به هریک از سوالات زیر پاسخ دهید:**

۱- برای هر یک از موارد زیر یک Cube ایجاد نمایید:

- میانگین درآمد افراد هر شهر در ماه
- تعداد اعضای دارای بیماری خاص استان تهران به تفکیک شهر و وضعیت بهزیستی
- تعداد پروانه های صنفی افراد شهرنشین در استان مازندران به تفکیک شهر و سال تولد
- مقدار واریزی افراد به تفکیک سال، استان، جنسیت و وضعیت شهر نشینی
- میانگین قیمت ماشین های هر خانواده به تفکیک شهر و تعداد اعضای خانواده

۲- به نظر شما چه Cube های دیگری می توان با استفاده از مجموعه داده رفاه ایرانیان ساخت که برای مراحل بعدی داده کاوی موثر باشد. ۵ مورد Cube دلخواه دیگر با ذکر دلیل ایجاد نمایید.

\* برای هر کدام از سوالات فوق مراحل مختلف کار خود را توضیح دهید. استفاده از کتابخانه pandas برای انجام این تمرین توصیه می شود.

**امتیازی:** انجام پیش پردازش و مستندسازی مراحل مختلف پیش پردازش قبل از پاسخ به سوالات فوق، دارای نمره اضافه است.

**موفق باشید**