

درک و توصیف دادهها- قسمت دوم

به پیوست مجموعه دادههای زیر برای حل تمرین ایفاد گردید.

1. A Clinical Breast Cancer Dataset

2. The Salaries of year 2015

۱. برای هر یک از مجموعه دادههای فوق، معیارهای تعیین شده مرکزیت داده و شاخصهای آماری مورد نیاز را محاسبه کنید و در صورت نیاز تفسیری از نتایج بدست آمده، ارائه دهید. ویژگیهای هر مجموعه داده به شرح زیر است:

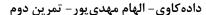
Dataset#1

- Complete TCGA ID
- Gender
- Age at Initial Pathologic Diagnosis: Mean, Mode, Median
- ER Status
- PR Status
- HER2 Final Status
- Tumor : Mode
- Tumor--T1 Coded
- Node: Mode
- Node-Coded
- Metastasis
- Metastasis-Coded
- AJCC Stage: Mode
- Vital Status

Dataset#2

- Employee Name
- Job Title: Mode
- Base Pay: Mean, Mode, Median
- Overtime Pay: Mean, Mode, Median, a horizontal boxplot of the Overtime Pay
- Other Pay: Mean, Mode, Median
- Benefits: Mean, Mode, Median, a horizontal boxplot of the Benefits
- Total Pay: Mean, Mode, Median
- Total Pay & Benefits
- Year

۲. برای مجموعه داده شماره ۱، Boxplot از ویژگیهای Age و Tumor و Boxplot از ویژگیهای Age و Age و Stage
۲. برای مجموعه داده شماره ۱، Boxplot از ویژگیهای Stage و Stage





- ۳. برای مجموعه داده شماره ۲، چند Boxplot از ویژگیهای Job Title و سایر ویژگیها رسم کنید. انتخاب نوع شغل و قیژگیهای دیگر اختیاری است (مثلا میتوانید Boxplot حقوق پایه افرادی که در عنوان شغل آنها عبارت ویژگیهای دیگر اختیاری است را با ASSOCIATE TAX AUDITORها مقایسه کنید).
- ۴. برای مجموعه داده شماره ۱ و ۲ بصورت جداگانه تمامی مراحل محاسبه Dissimilarity بین دو نمونه را بصورت دستی بنویسید.

مهلت تحویل: ۱۴۰۲/۰۲/۲۰

پاسخ تمرین را در قالب یک فایل PDF بصورت یک گزارش فنی (Technical Report) آماده سازید.

می توانید از زبان پایتون برای رسم نمودارها و پاسخ به سوالات استفاده نمایید.

اگر تمرین شامل برنامهنویسی است ارائه سورس کد برنامه در فایل پیوست الزامی است.

تمامی فایلهای ارسالی به فرمت rar. و نام فایلها به فرمت زیر باشد که # شماره تمرین شماست:

DM_HW#_YourStudentNumber_Family.rar