داده کاوی ـ ترم پاییز ـ سال ۹۸
دانشگاه صنعتی خواجه نصیرالدین طوسی ـ دانشکده ریاضی

تمرین سری اول

هدف از این تکلیف، کار با داده واقعی و تمرین با نمایش داده‌ها و رسم نمودار است. برای این منظور مجموعه داده adult را از لینک زیر دانلود کنید.

```
https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
```

هر خط از فایل اطلاعات مربوط به یک شخص ساکن ایالات متحده را نشان می‌دهد که شامل اطلاعاتی مثل سن، میزان تحصیلات، جنسیت، شغل و غیره می‌باشد. در زیر یک نمونه از اطلاعات یک رکورد را نشان می‌دهد.

```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical,
Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
```

در این میان فیلد اول سن فرد را نشان می‌دهد و فیل پنجم عددی است که تعداد سالی که فرد مورد نظر تحصیل کرده را نشان می‌دهد. همچنین در انتهای هر خط مشخص شده که شخص مورد نظر حقوق سالیانه بالای پنجاه هزار دلار دارد >50k و یا کمتر از پنجاه هزار $<=50$k

توصیفات مربوط به کلیه فیلدها را در لینک زیر ببینید.

```
https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names
```

با توجه به داده های فایل adult می‌خواهیم نمودارهای زیر را رسم کنیم.

۱. می‌خواهیم اطلاعات مربوطه به پنج دسته از افراد را جدا کنیم. دسته اول، کسانی که تعداد سالهای تحصیلشان کمتر مساوی 5، دسته دوم بیشتر از 5 و کمتر و مساوی 7، دسته سوم بیشتر از 7 و کمتر مساوی 10، دسته چهارم بیشتر از 10 و کمتر مساوی 14 و دسته آخر کسانی که تعداد سالهای تحصیلشان بیشتر از 14 بوده است. نموداری ستونی (bar) را رسم کنید که برای هر دسته تعداد افراد در آن دسته را بصورت ستونی نشان دهد.

۲. نموداری از نوع boxplot را رسم کنید بطوریکه برای هر پنج دسته بالا، ماکزیمم، مینیمم، Q_1, Q_2, Q_3 فیلد سن را نشان دهد.

۳. در تمرین بالا برای هر دسته میانگین و انحراف معیار دسته مورد نظر را نیز محاسبه کنید.

۴. صد رکورد به تصادف انتخاب کرده و نموداری نقطه‌ای رسم کنید بطوری که محور x ها تعداد سالهای تحصیلات و محور y ها سن فرد را نشان دهد. اگر فرد مورد نظر حقوق سالیانه بالای پنجاه هزار داشته باشد، نقطه را با رنگ آبی و در غیر این صورت با رنگ قرمز نشان دهد.