به نام او که جان را فکرت آموخت

**تمرین دوم درس وب معنایی**

نفیسه عامری1

1 دانشجو، دانشکده مهندسی، دانشگاه فردوسی، مشهد

**صورت سوال**

1. 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 مجموعه‌داده به شرح زیر است.
2. برای مجموعه داده شماره 1 Boxplot از ویژگی‌های Age و Tumor ، و Boxplot از ویژگی‌های Age و AJCC Stage را با استفاده از پایتون رسم نمایید.
3. برای مجموعه داده شماره 2، چند Boxplot از ویژگی‌های Job Title و سایر ویژگی‌ها رسم کنید. انتخاب نوع شغل و ویژگی‌های دیگر اختیاری است (مثلا می‌توانید Boxplot حقوق پایه افرادی که در عنوان شغل آن‌ها عبارت ASSOCIATE TAX AUDITOR است را با PHYSICIAN AND SURGEON ها مقایسه کنید).
4. برای مجموعه داده شماره 1 و 2 بصورت جداگانه تمامی مراحل محاسبه Dissimilarity بین دو نمونه را بصورت دستی بنویسید.

# شرح تکنیکال

مجموعه داده مورد استفاده در این پروژه دارای 49000 رکورد و یک برچسب دوجمله‌ای است که حقوق کمتر یا بیشتر از پنجاه هزار دلار آمریکا را نشان می‌دهد. جدول زیر توضیحات برای قسمت اول و دوم تمرین را نوشتم.

برای قسمت پنجم تمرین:

سن: سن به طور بالقوه می‌تواند با وضعیت تاهل، تحصیلات و شغل ارتباط داشته باشد. به عنوان مثال، افراد مسن ممکن است بیشتر متاهل باشند و سطح تحصیلات بالاتر و تجربه بیشتری در شغل خود داشته باشند.

طبقه کارگر: طبقه کار می‌تواند با شغل و درآمد مرتبط باشد. به عنوان مثال، افرادی که در بخش خصوصی کار می‌کنند ممکن است نسبت به افرادی که برای دولت کار می‌کنند، مشاغل پردرآمدتری داشته باشند.

Fnlwgt: fnlwgt وزن نهایی است که تعداد واحدهای موجود در مجموعه داده اصلی توسط هر رکورد نشان داده می‌شود. بعید است که هیچ ارتباطی با سایر ویژگی‌های مجموعه داده داشته باشد.

تحصیلات: تحصیلات می‌تواند با شغل و درآمد مرتبط باشد. به عنوان مثال، افرادی با سطوح تحصیلات بالاتر ممکن است مشاغلی داشته باشند که به مهارت‌های تخصصی بیشتری نیاز دارند و حقوق بیشتری دریافت می‌کنند.

Education-num: Education-num نمایش عددی آموزش است. بعید است که هیچ ارتباطی با سایر ویژگی‌های مجموعه داده داشته باشد.

وضعیت تاهل: وضعیت تاهل می‌تواند با سن و شغل مرتبط باشد. به عنوان مثال، افراد مسن ممکن است بیشتر متاهل باشند، و افراد در مشاغل خاص ممکن است بیشتر از دیگران متاهل باشند.

شغل: شغل می‌تواند با تحصیلات و درآمد مرتبط باشد. به عنوان مثال، افرادی که در پست‌های مدیریتی هستند ممکن است سطح تحصیلات بالاتری داشته باشند و حقوق بیشتری نسبت به افرادی که در موقعیت‌های کارگری قرار دارند، دریافت کنند.

رابطه: رابطه می‌تواند با وضعیت تاهل و سن همبستگی داشته باشد. به عنوان مثال، افرادی که متاهل هستند ممکن است همسر یا فرزندانی در این زمینه داشته باشند.

نژاد: بعید است که نژاد هیچ ارتباطی با سایر ویژگی‌های مجموعه داده داشته باشد.

جنسیت: جنسیت می‌تواند با شغل و درآمد مرتبط باشد. برای مثال، مردان ممکن است بیشتر از زنان مشاغل پردرآمد در صنایع خاص داشته باشند.

سود سرمایه: سود سرمایه می‌تواند با درآمد مرتبط باشد. افراد با درآمد بالاتر ممکن است فرصت‌های سرمایه گذاری بیشتری داشته باشند که منجر به افزایش سرمایه می‌شود.

زیان سرمایه: زیان سرمایه نیز می‌تواند با درآمد مرتبط باشد. افراد با درآمد بالاتر ممکن است دارایی‌های بیشتری برای از دست دادن داشته باشند که منجر به زیان سرمایه بیشتر می‌شود.

ساعت در هفته: ساعت‌ها در هفته می‌تواند با شغل و درآمد مرتبط باشد. به عنوان مثال، افرادی که در مشاغلی که ساعتی پرداخت می‌کنند ممکن است ساعت‌های طولانی‌تری برای کسب درآمد بیشتر کار کنند.

کشور بومی: کشور بومی هیچ ارتباطی با سایر ویژگی‌های مجموعه داده بعید است داشته باشد.

درآمد: درآمد متغیر هدف است و مهمترین ویژگی است که باید در هنگام تجزیه و تحلیل این مجموعه داده در نظر گرفت. می‌توان آن را با بسیاری از ویژگی‌های دیگر در مجموعه داده، مانند تحصیلات، شغل، و ساعت‌ها در هفته مرتبط کرد.

با توجه به پراکندگی داده‌های هر ویژگی، ما باید داده‌ها را بیشتر تجزیه و تحلیل کنیم تا محدوده، میانگین، میانه، حالت، انحراف استاندارد و سایر معیارهای آماری برای هر ویژگی تعیین شود. بدون این اطلاعات، تعیین پراکندگی داده ها دشوار است.

# شرح نتایج

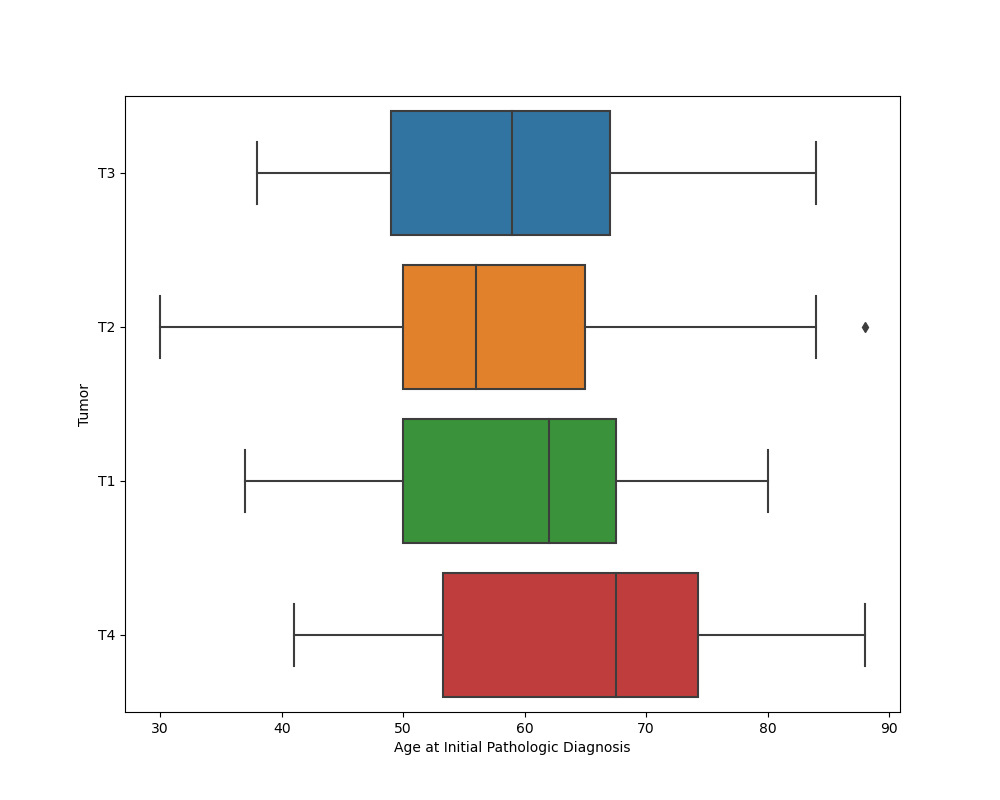
خروجی نتایج مسئله به شرح زیر است.

A screenshot of a computer

Description automatically generated

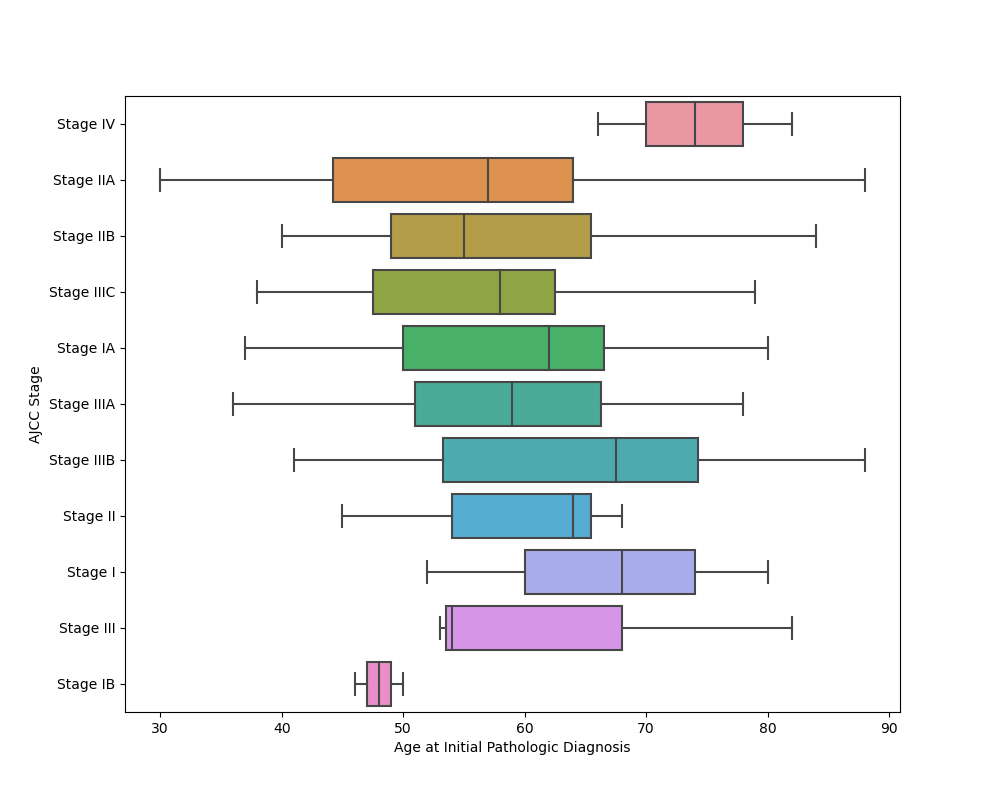
شکل ‏2‑1- خروجی کد برنامه برای مجموعه داده اول

همانطور که در تصویر بالا مشاهده می‌کنید، آمار کلی از جنسیت افراد، سن آن‌ها و درآمدشان در تصویر بالا آمده است؛ اما در ادامه برای تجزیه و تحلیل بهتر دیتاست، چند نمودار هیستوگرام با استفاده از زبان پایتون نوشته ام.



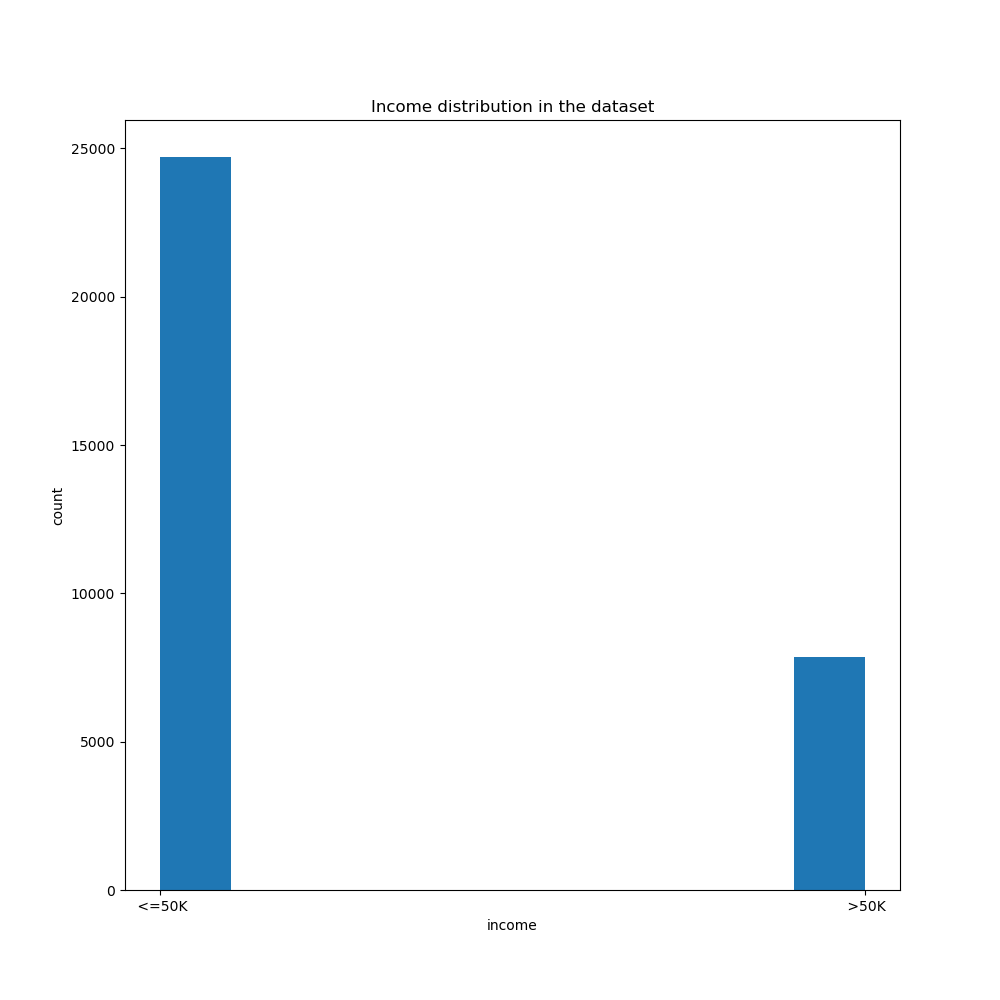
شکل ‏2‑2: نمودار هیستوگرام سن افراد

همانطور که در تصویر بالا مشاهده می‌کنید، میانگین سن افراد 38 سال است و اکثر افراد این دیتاست جوان و زیر 50 سال سن دارند.



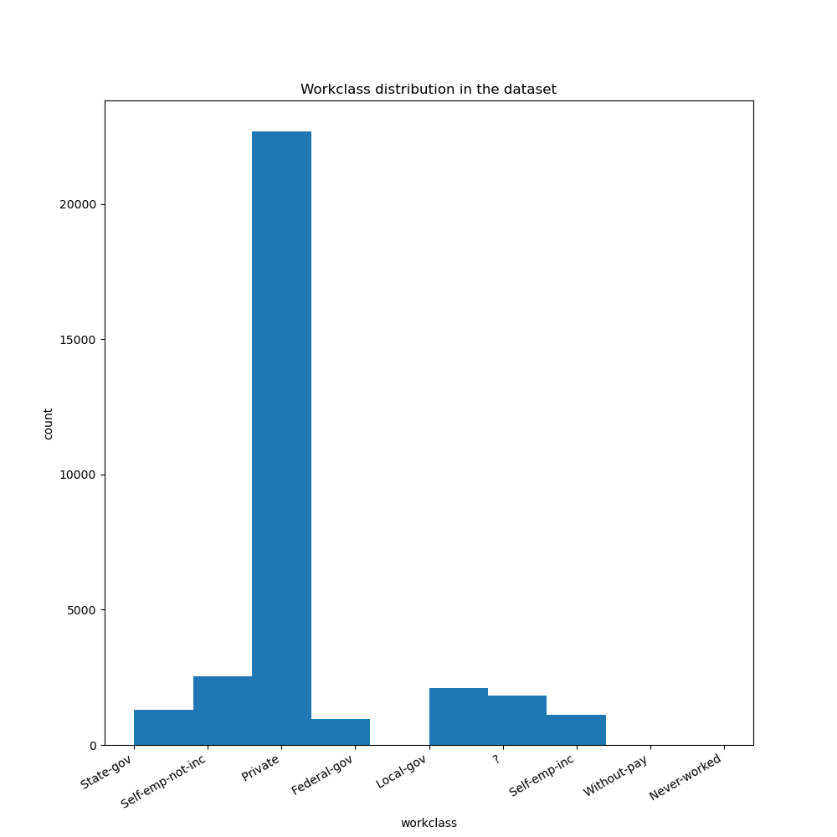
شکل ‏2‑3: نمودار هیستوگرام تحصیلات افراد

همانطور که در تصویر بالا مشاهده می‌کنید، تحصیلات بیشتر افراد کارشناسی است و بعد از آن تحصیلات کالج.



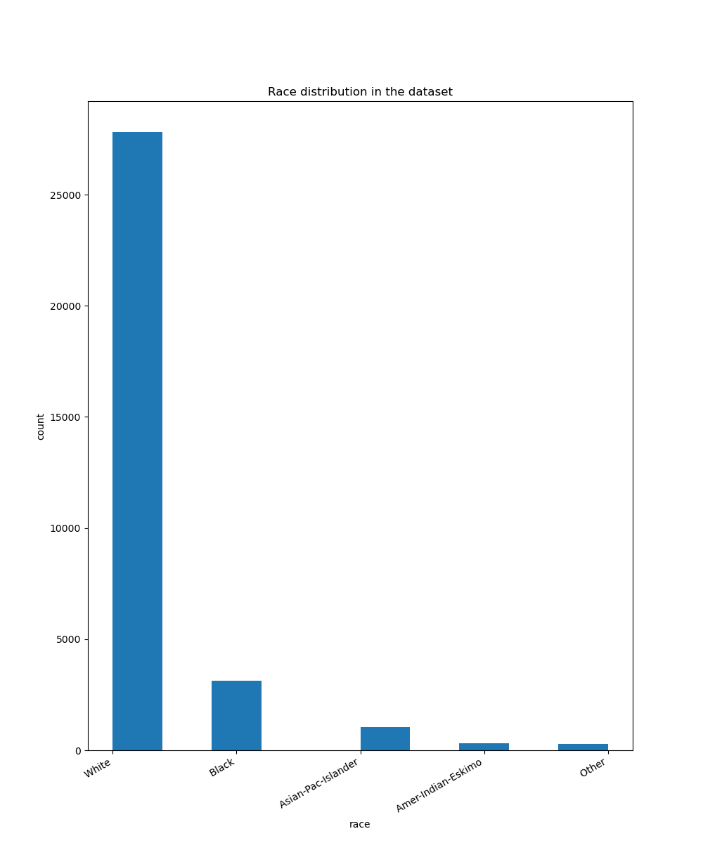
شکل ‏2‑4: نمودار هیستوگرام درآمد افراد

همانطور که در تصویر بالا مشاهده می‌کنید، تقریبا 70 درصد افراد جامعه حقوق زیر 50 هزار دلار در سال دریافت می‌کنند.



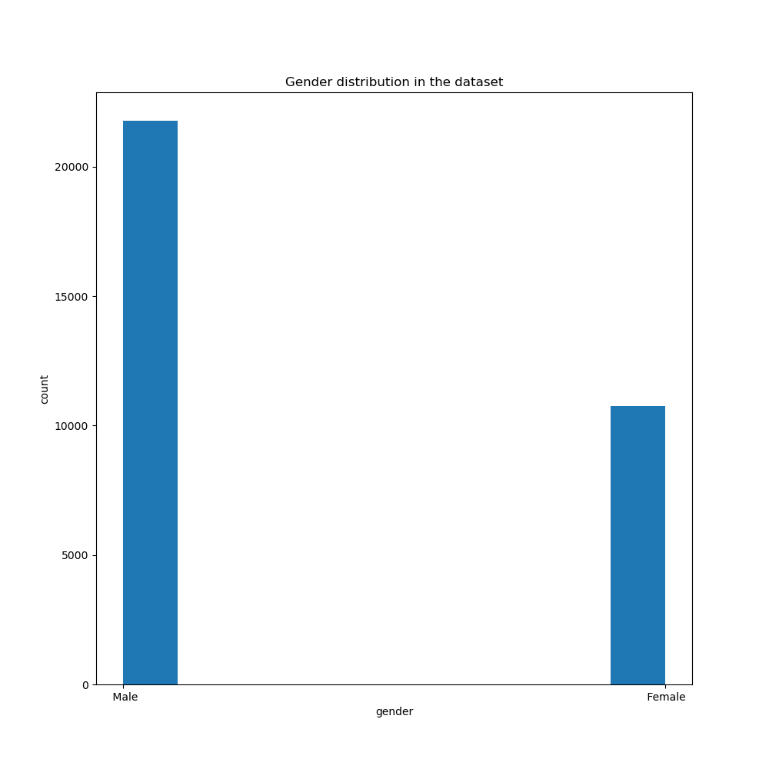
شکل ‏2‑5: نمودار هیستوگرام کلاس کاری افراد

همانطور که در تصویر بالا مشاهده می‌کنید، اکثر افراد این دیتاست شغل خصوصی خود را دارند یا در شرکت‌های خصوصی مشغول به کارند.



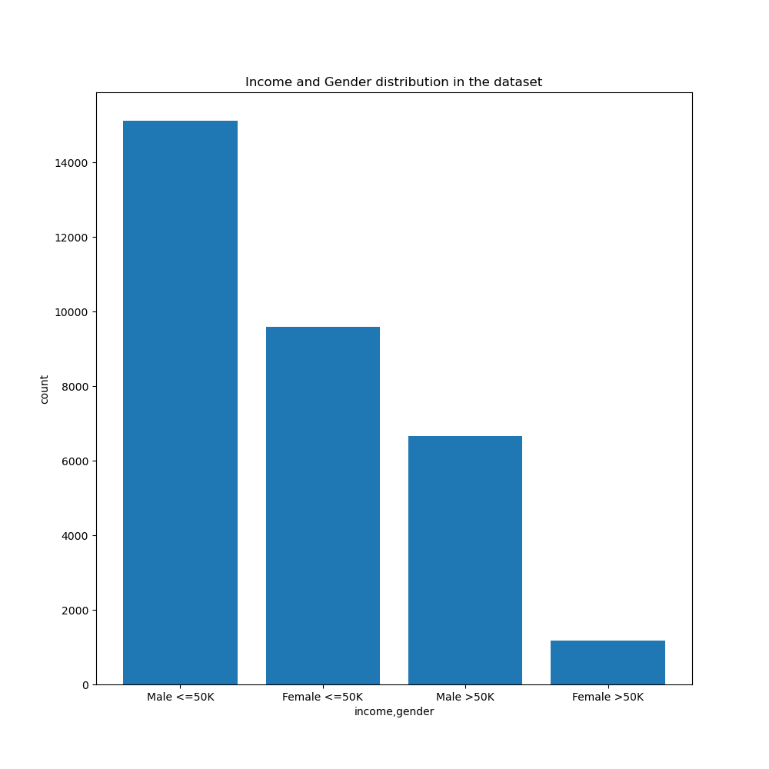
شکل ‏2‑6: نمودار هیستوگرام نژاد افراد

همانطور که در تصویر بالا مشاهده می‌کنید، نژاد اکثر افراد این دیتاست سفید است.



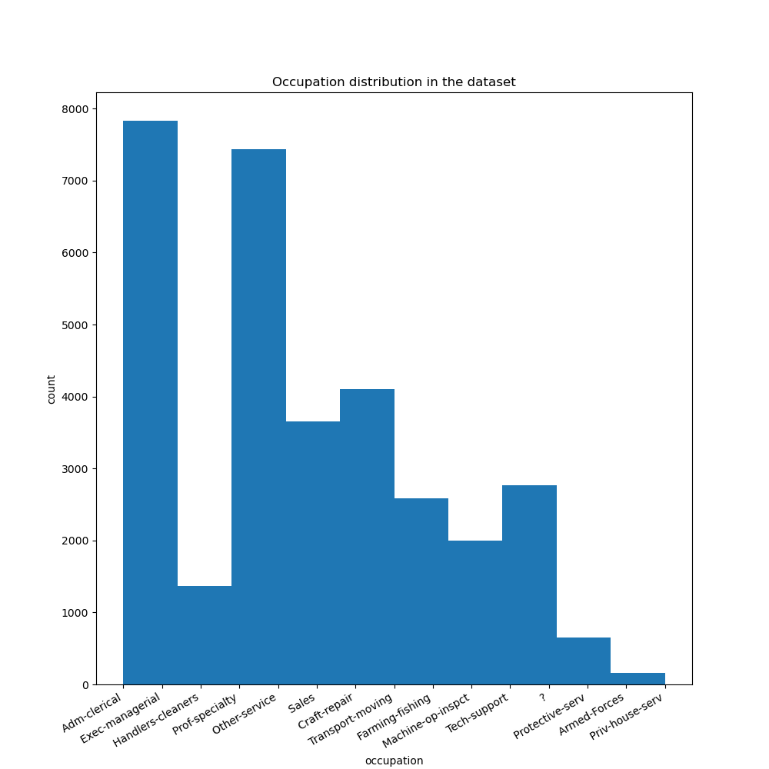
شکل ‏2‑7: نمودار هیستوگرام جنسیت افراد

همانطور که در تصویر بالا مشاهده می‌کنید، تعداد مردها در این دیتاست دوبرابر زن‌ها است.



شکل ‏2‑8: نمودار هیستوگرام درآمد افراد بر اساس جنسیت آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، تقریبا 60 درصد درآمد مردان کمتر از 50 هزار دلار است. در حالی که تقریبا 90 درصد درآمد زنان در این دیتاست کمتر از 50 هزار دلار است. این نشان می‌دهد که 40 درصد مردان شغل‌های پر درآمد دارند در حالی که 10 درصد زنان شغل‌های پر درآمد دارند.



شکل ‏2‑9: نمودار هیستوگرام نوع شغل افراد

همانطور که در تصویر بالا مشاهده می‌کنید، اکثر افراد این دیتاست در شغل‌های Prof-specialty و Adm-clerical هستند.

**پیوست**

**کد برنامه برای مجموعه داده اول**

import statistics

import seaborn as sns

import pandas as pd

from matplotlib import pyplot as plt

dataset1 = pd.read\_csv('dataset1.csv')

print('============== Age ===============')

age = dataset1['Age at Initial Pathologic Diagnosis']

mean\_age = age.mean()

print('Median:', age.median())

print('Mean:', mean\_age)

print('Mode:', statistics.mode(age))

plt.figure(figsize=(10, 8))

plt.hist(age, bins=20)

plt.axvline(mean\_age, color='red', linestyle='dashed', linewidth=2)

plt.xlabel('age')

plt.ylabel('count')

plt.title('Age distribution in the dataset')

plt.savefig('age.png')

print('============== Tumor ===============')

tumor = dataset1['Tumor']

tumor\_mode = statistics.mode(tumor)

print('Mode:', tumor\_mode)

plt.figure(figsize=(10, 8))

plt.hist(tumor)

plt.axvline(tumor\_mode, color='red', linestyle='dashed', linewidth=2)

plt.xlabel('tumor')

plt.ylabel('count')

plt.title('Tumor distribution in the dataset')

plt.savefig('tumor.png')

print('============== Node ===============')

node = dataset1['Node']

node\_mode = statistics.mode(node)

print('Mode:', node\_mode)

plt.figure(figsize=(10, 8))

plt.hist(node)

plt.axvline(node\_mode, color='red', linestyle='dashed', linewidth=2)

plt.xlabel('node')

plt.ylabel('count')

plt.title('Node distribution in the dataset')

plt.savefig('node.png')

print('============== AJCC Stage ===============')

AJCCStage = dataset1['AJCC Stage']

AJCCStage\_mode = statistics.mode(AJCCStage)

print('Mode:', AJCCStage\_mode)

plt.figure(figsize=(10, 8))

plt.hist(AJCCStage)

plt.axvline(AJCCStage\_mode, color='red', linestyle='dashed', linewidth=2)

plt.xlabel('AJCCStage')

plt.ylabel('count')

plt.title('AJCCStage distribution in the dataset')

plt.savefig('AJCCStage.png')

# ============ boxplot ==============

plt.figure(figsize=(10, 8))

sns.boxplot(x='Age at Initial Pathologic Diagnosis', y='Tumor', data=dataset1)

plt.savefig('Boxplot Age & Tumor.png')

plt.figure(figsize=(10, 8))

sns.boxplot(x='Age at Initial Pathologic Diagnosis', y='AJCC Stage', data=dataset1)

plt.savefig('Boxplot Age & AJCC Stage.png')

**کد برنامه برای مجموعه داده دوم**