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.

# شرح نتایج

خروجی نتایج مسئله به شرح زیر است.

A screenshot of a computer

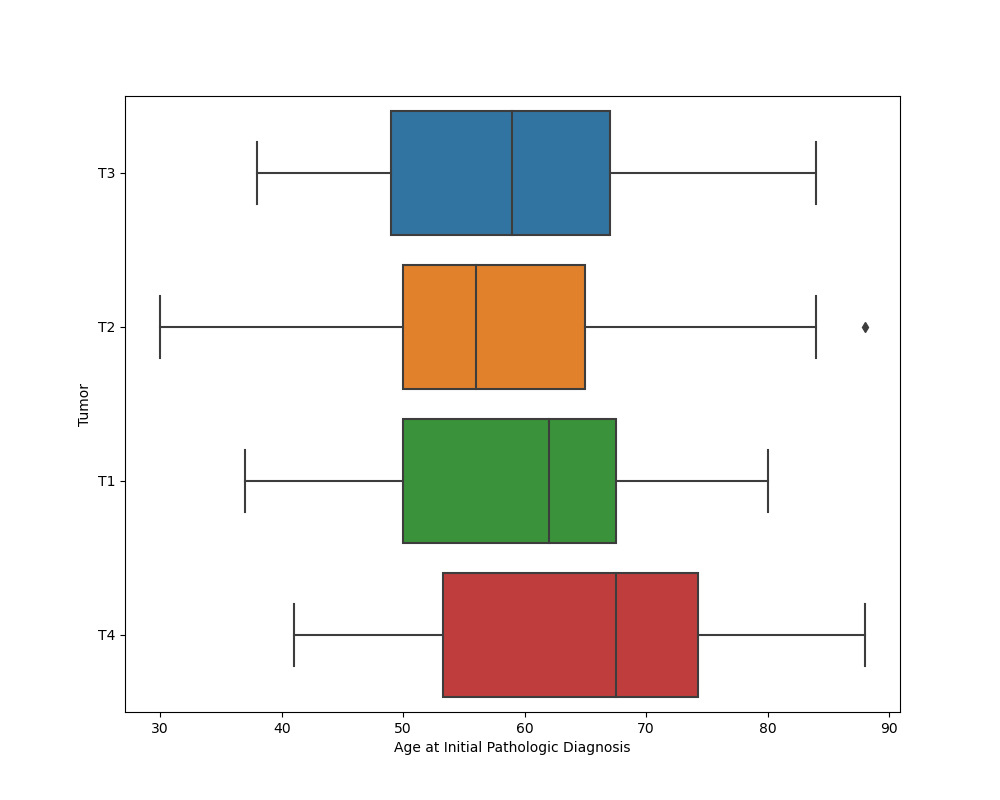
Description automatically generated

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

همانطور که در تصویر بالا مشاهده می‌کنید، شاخص‌های آماری مد میانه و میانگین از سن افراد، نوع تومور آن‌ها و نودشان و همچنین از AJCC Stage را مشاهده می‌کنید. در ادامه برای تجزیه و تحلیل بهتر دیتاست، چند نمودار هیستوگرام با استفاده از زبان پایتون نوشته ام.

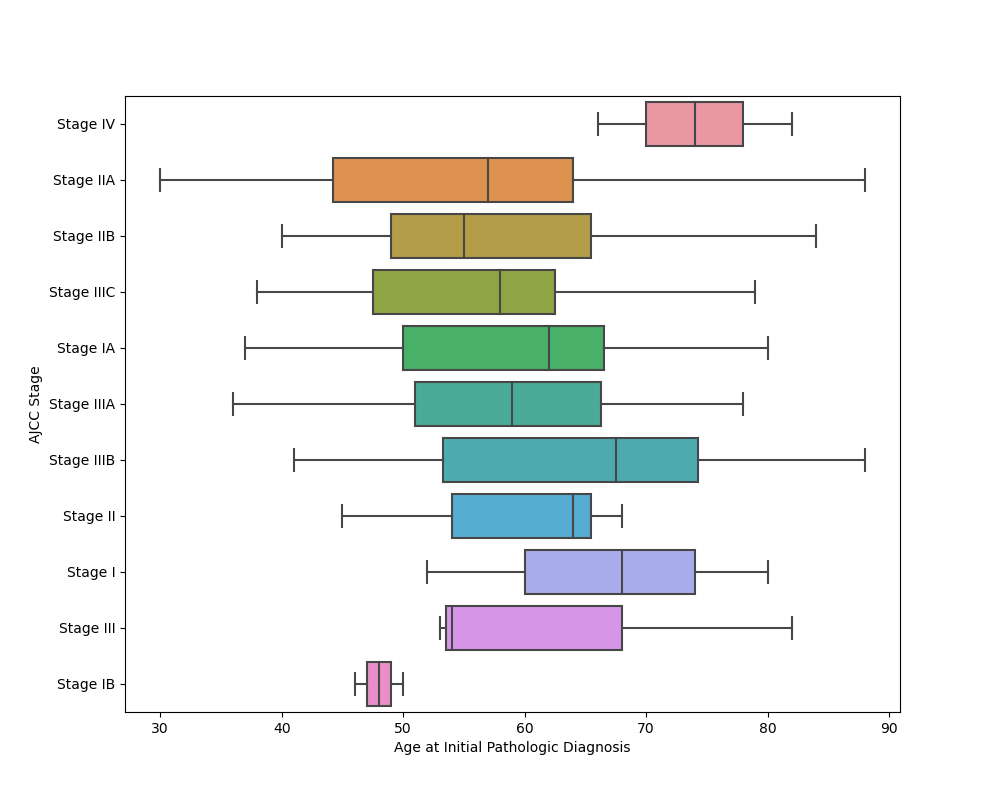
|  |  |
| --- | --- |
|  |  |

شکل ‏2‑2: نمودارهای هیستوگرام از سن و نوع تومور و نود و AJCC Stage



شکل ‏2‑3: نمودار باکس پلات از سن افراد و نوع تومور آن‌ها

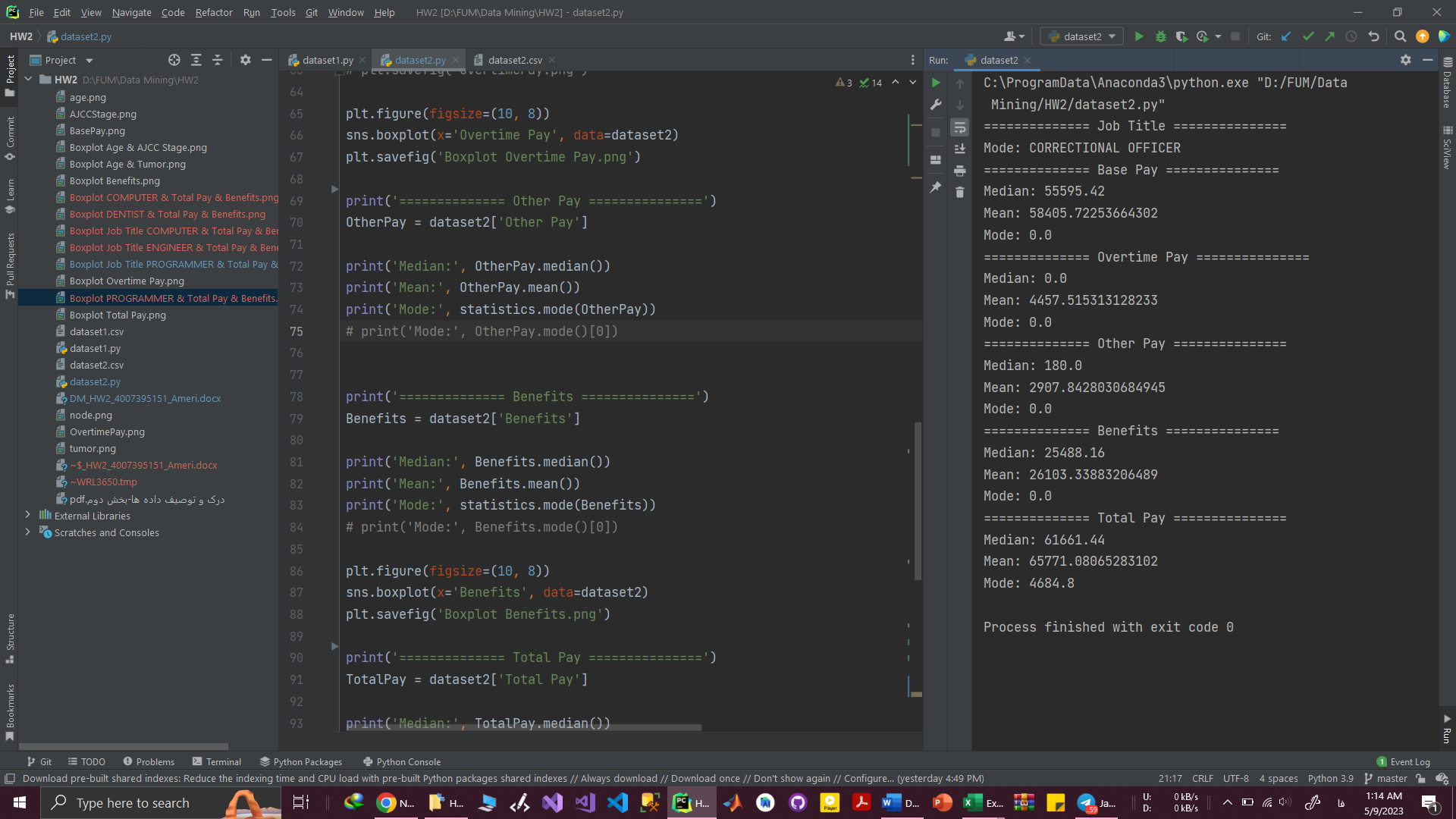
همانطور که در تصویر بالا مشاهده می‌کنید، بیشتر افرادی که تومور دارند سنی بین 50 تا 75 دارند؛ همچنین تومور نوع T4 در بین افراد با سن بالاتر شایع‌تر است. بیشترین تعداد افرادی که داده‌های آن‌ها در این نمودار نمایش داده شده است، در سنین 60 تا 70 سال هستند و اندازه تومورشان در حدود 2 تا 3 سانتیمتر است. همچنین در نمودار مشاهده می‌شود که افرادی که سنشان بین 30 تا 50 سال است، در مقایسه با سنین بالاتر، تومورهایی با اندازه کوچک‌تر دارند. در کل، توزیع داده‌ها در این نمودار نسبتاً یکنواخت است و به‌طور معمول مشاهده نمی‌شود که تعداد زیادی داده‌ها در کمینه یا بیشینه باشند.



شکل ‏2‑4: نمودار باکس پلات از میزان گسترش تومور بر اساس سن افراد

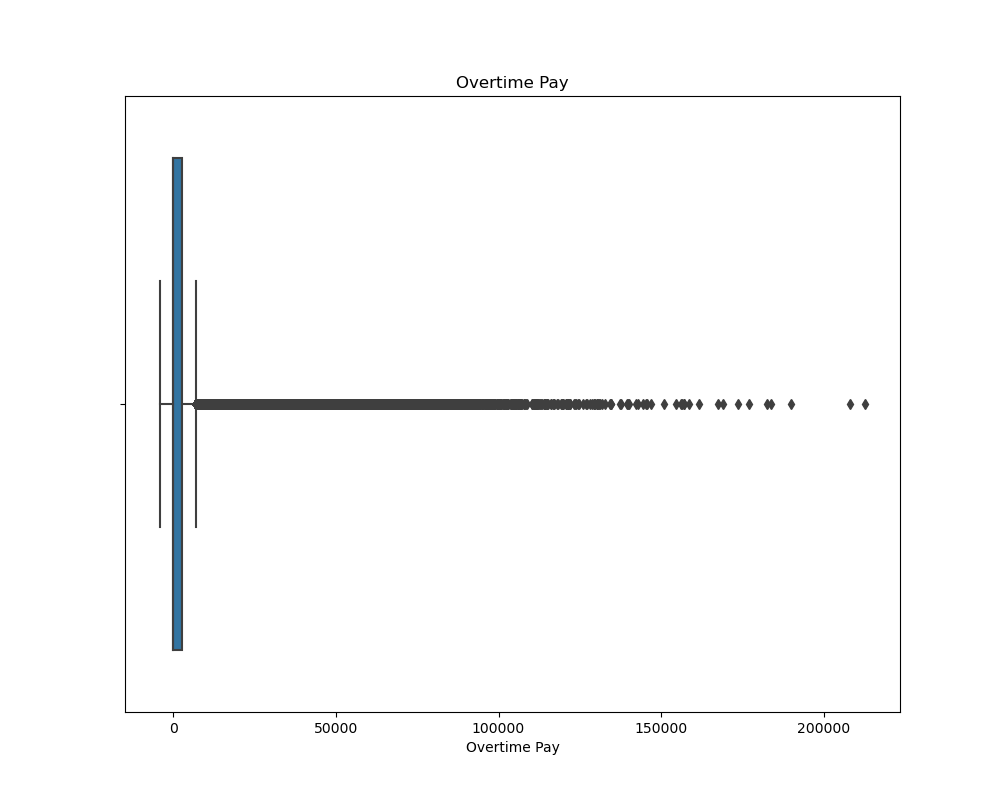
همانطور که در تصویر بالا مشاهده می‌کنید، به عنوان مثال افراد بین 70 تا 80 سال در مرحله چهارم از میزان گسترش تومور هستند. سن متوسط در تشخیص اولیه پاتولوژیک حدود 60 سال است. دامنه سنی بسیار گسترده است، از 30 تا 88. محدوده بین چارکی (IQR) تقریباً بین 48 تا 67 است، به این معنی که 50٪ از داده‌ها در این محدوده قرار می‌گیرند. چندین نقطه پرت وجود دارد که با نقاط منفرد خارج از سبیل‌ها نشان داده شده است، که نشان می‌دهد برخی از بیماران در سنین بسیار جوان‌تر یا بزرگ‌تر از اکثر بیماران در مجموعه‌داده تشخیص داده شده‌اند. توزیع تقریباً متقارن به نظر می‌رسد، با کادر در مرکز میانه. گسترش داده‌ها نسبتاً گسترده به نظر می‌رسد، با کادری که از حدود 43 تا 74 گسترش یافته است.

برای مجموعه داده دوم داریم:



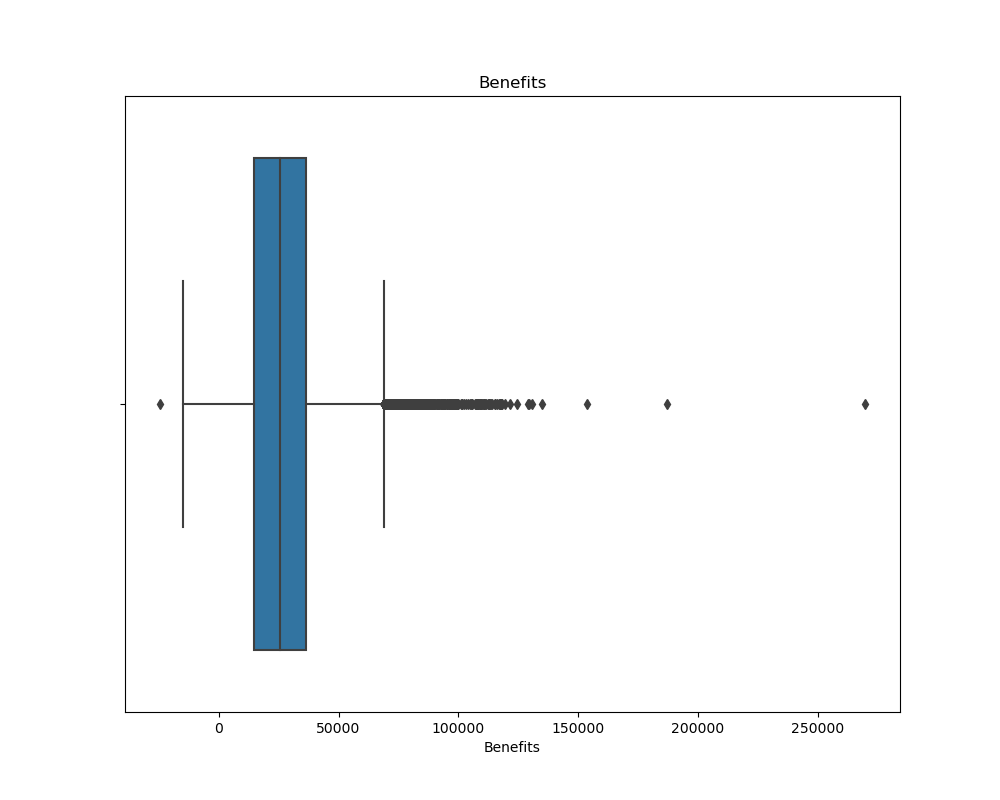
شکل ‏2‑5: خروجی کد برنامه برای مجموعه داده دوم

همانطور که در تصویر بالا مشاهده می‌کنید، شاخص‌های آماری مد میانه و میانگین از شغل افراد، حقوق پایه آن‌ها و حقوق اضافه کاری و مزیا و در مجموع کل درآمد آن‌ها را مشاهده می‌کنید.



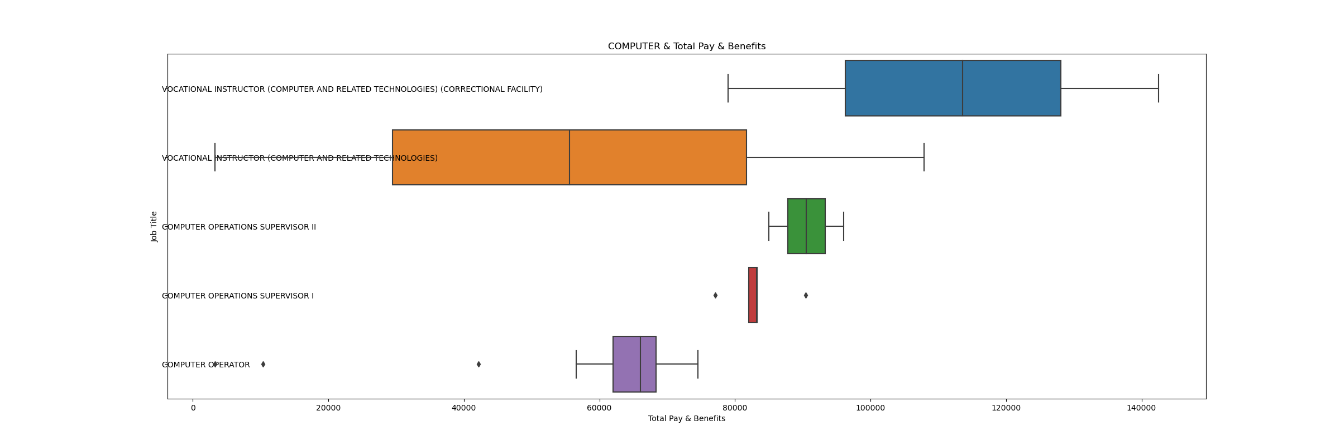
شکل ‏2‑6: نمودار باکس پلات حقوق اضافه کاری

همانطور که در تصویر بالا مشاهده می‌کنید، در اکثر شغل‌ها حقوق اضافه کاری صفر هست؛ اما در چند مورد خاص پرداخت اضافه کاری به 200000 هم رسیده است.



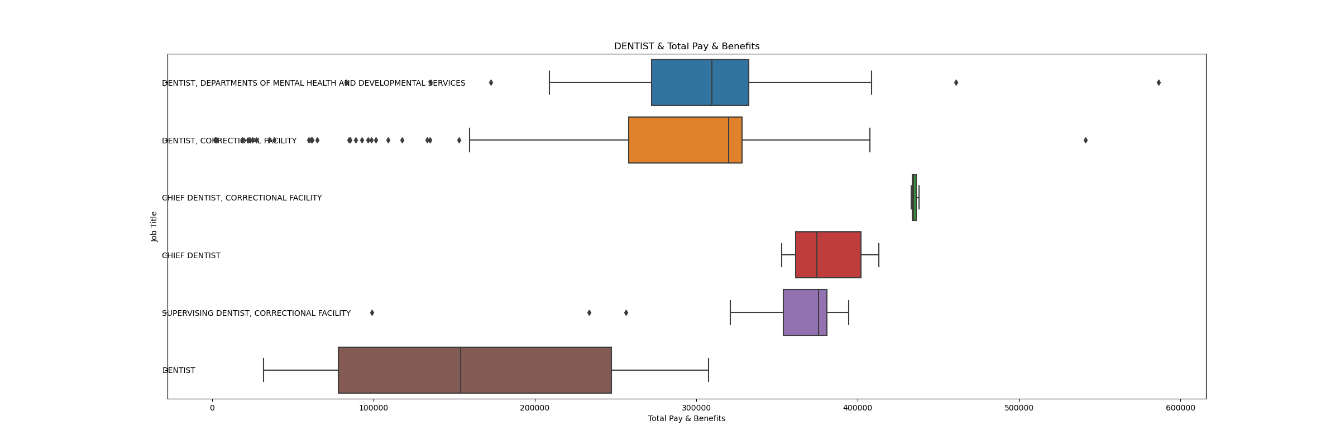
شکل ‏2‑7: نمودار باکس پلات مزایا

همانطور که در تصویر بالا مشاهده می‌کنید، مزایا چیزی حدود صفر تا 70000 است و در موارد خاص به 200000 هم رسیده است.



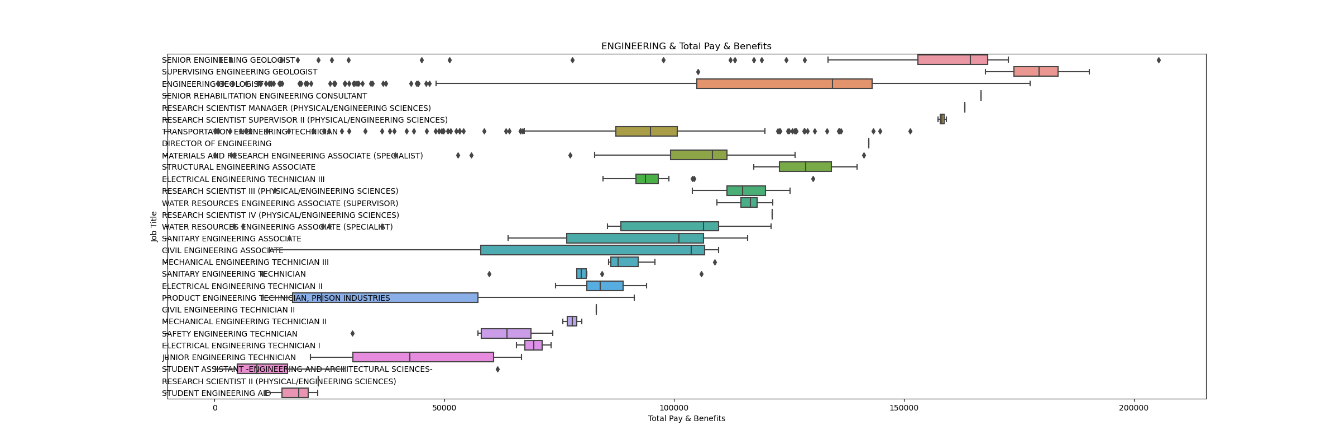
شکل ‏2‑8: نمودار هیستوگرام درآمد افراد بر اساس جنسیت آن‌ها

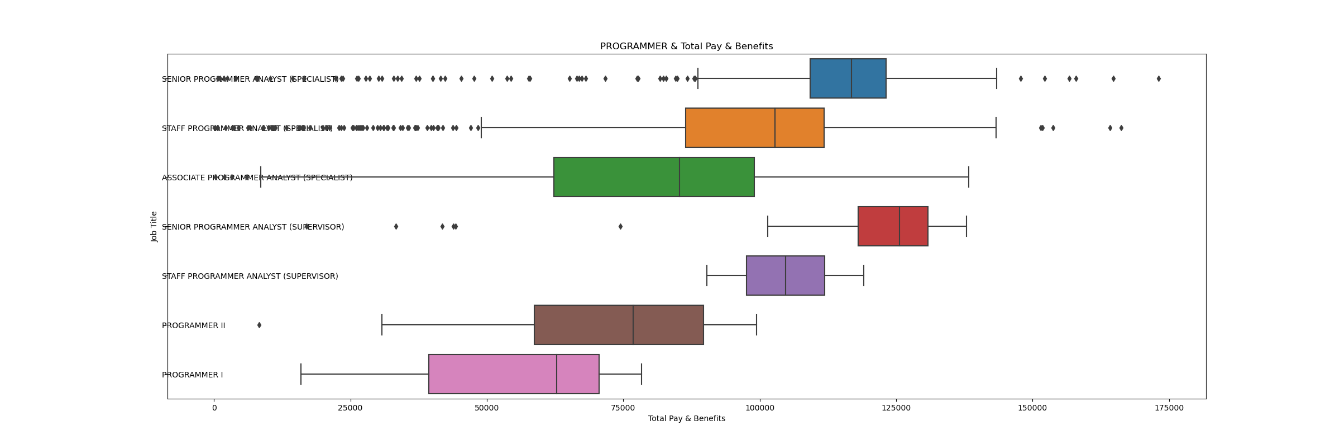
همانطور که در تصویر بالا مشاهده می‌کنید، تقریبا 60 درصد درآمد مردان کمتر از 50 هزار دلار است. در حالی که تقریبا 90 درصد درآمد زنان در این دیتاست کمتر از 50 هزار دلار است. این نشان می‌دهد که 40 درصد مردان شغل‌های پر درآمد دارند در حالی که 10 درصد زنان شغل‌های پر درآمد دارند.

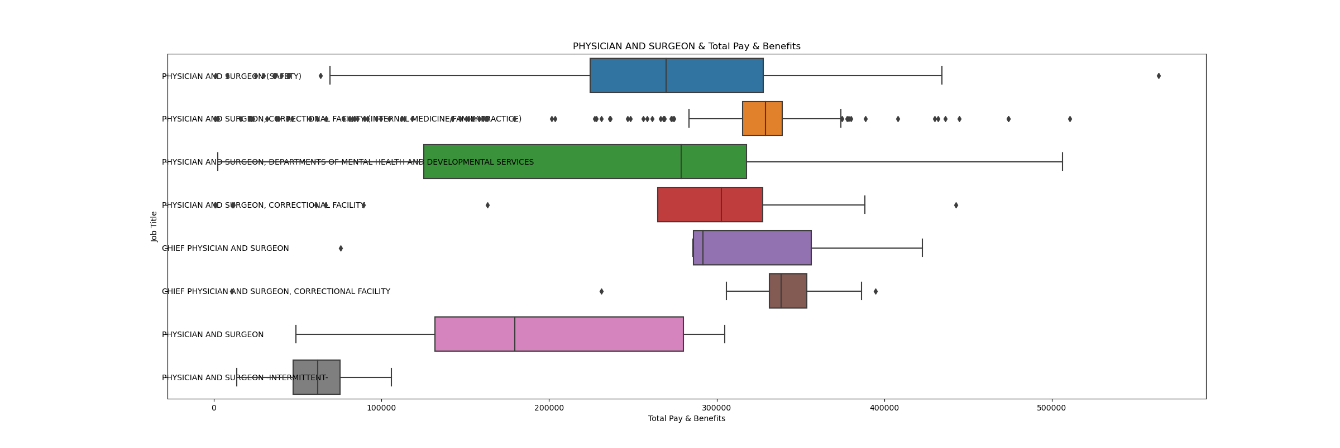


شکل ‏2‑9: نمودار هیستوگرام نوع شغل افراد

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**

import statistics

import seaborn as sns

import pandas as pd

from matplotlib import pyplot as plt

def boxplot(data, name):

    plt.figure(figsize=(24, 8))

    sns.boxplot(x='Total Pay & Benefits', y='Job Title', data=data)

    # plt.yticks(rotation=45, ha="right")

    plt.yticks(ha="left")

    plt.title(name)

    plt.savefig('Boxplot ' + name + '.png')

dataset2 = pd.read\_csv('dataset2.csv')

print('============== Job Title ===============')

JobTitle = dataset2['Job Title']

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

print('============== Base Pay ===============')

BasePay = dataset2['Base Pay']

mean\_BasePay = BasePay.mean()

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

print('Mean:', mean\_BasePay)

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

print('============== Overtime Pay ===============')

OvertimePay = dataset2['Overtime Pay']

mean\_OvertimePay = OvertimePay.mean()

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

print('Mean:', mean\_OvertimePay)

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

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

sns.boxplot(x='Overtime Pay', data=dataset2)

plt.title('Overtime Pay')

plt.savefig('Boxplot Overtime Pay.png')

print('============== Other Pay ===============')

OtherPay = dataset2['Other Pay']

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

print('Mean:', OtherPay.mean())

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

print('============== Benefits ===============')

Benefits = dataset2['Benefits']

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

print('Mean:', Benefits.mean())

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

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

sns.boxplot(x='Benefits', data=dataset2)

plt.title('Benefits')

plt.savefig('Boxplot Benefits.png')

print('============== Total Pay ===============')

TotalPay = dataset2['Total Pay']

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

print('Mean:', TotalPay.mean())

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

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

sns.boxplot(x='Total Pay', data=dataset2)

plt.title('Total Pay')

plt.savefig('Boxplot Total Pay.png')

print('============== Total Pay & Benefits ===============')

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

sns.boxplot(x='Total Pay & Benefits', data=dataset2)

plt.title('Total Pay & Benefits')

plt.savefig('Boxplot Total Pay & Benefits.png')

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

PROGRAMMER = dataset2[JobTitle.str.contains('PROGRAMMER')]

data = pd.concat([PROGRAMMER])

boxplot(data, 'PROGRAMMER & Total Pay & Benefits')

COMPUTER = dataset2[JobTitle.str.contains('COMPUTER')]

data = pd.concat([COMPUTER])

boxplot(data, 'COMPUTER & Total Pay & Benefits')

ENGINEERING = dataset2[JobTitle.str.contains('ENGINEERING')]

data = pd.concat([ENGINEERING])

boxplot(data, 'ENGINEERING & Total Pay & Benefits')

# RESEARCHER = dataset2[JobTitle.str.contains('RESEARCHER')]

# data = pd.concat([RESEARCHER])

# boxplot(data, 'RESEARCHER & Total Pay & Benefits')

PHYSICIANANDSURGEON = dataset2[JobTitle.str.contains('PHYSICIAN AND SURGEON')]

data = pd.concat([PHYSICIANANDSURGEON])

boxplot(data, 'PHYSICIAN AND SURGEON & Total Pay & Benefits')

DENTIST = dataset2[JobTitle.str.contains('DENTIST')]

data = pd.concat([DENTIST])

boxplot(data, 'DENTIST & Total Pay & Benefits')