به نام او که جان را فکرت آموخت

تمرین دوم درس وب معنایی

نفیسه عامری[۱]

[۱] دانشجو، دانشکده مهندسی، دانشگاه فردوسی، مشهد

صورت سوال

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 مجموعه‌داده به شرح زیر است.

۲) برای مجموعه داده شماره ۱ Boxplot از ویژگی‌های Age و Tumor ، و Boxplot از ویژگی‌های Age و AJCC Stage را با استفاده از پایتون رسم نمایید.

۳) برای مجموعه داده شماره ۲، 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).

۴) برای مجموعه داده شماره ۱ و ۲ بصورت جداگانه تمامی مراحل محاسبه Dissimilarity بین دو نمونه را بصورت دستی بنویسید.

فصل ۱– شرح تکنیکال

در این تمرین از دو مجموعه داده استفاده شده است. در فصل بعد به توصیف و تفسیر نتایج می‌پردازیم.

در ادامه، برای قسمت چهارم تمرین محاسبه عدم شباهت روی دو مجموعه‌داده را مشاهده می‌کنید.

$$d(i,j) = \frac{p-m}{p}$$

که در آن m تعداد موارد منطبق است (یعنی تعداد مشخصه‌هایی که i و j در یک حالت هستند) و p تعداد کل ویژگی‌هایی است که اشیاء را توصیف می‌کنند. وزن‌ها را می‌توان برای افزایش اثر m یا اختصاص وزن بیشتر به منطبق‌ها در ویژگی‌هایی که تعداد حالت‌های بیشتری دارند، اختصاص داد.

طبق این تعریف برای مجموعه داده اول، برای نمونه اول و دوم داریم:

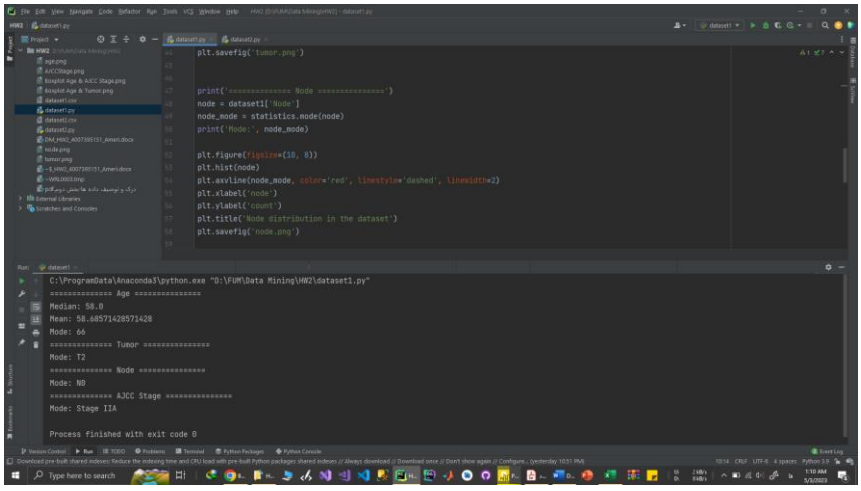
$$d(A0T2, A0CM) = \frac{30-12}{30} = 0.6$$

و برای مجموعه داده دوم، برای نمونه اول و دوم داریم:

$$d(Theodore\ H\ Eliopoulos, Vernon\ L\ Steiner) = \frac{12-4}{12} = 0.66$$

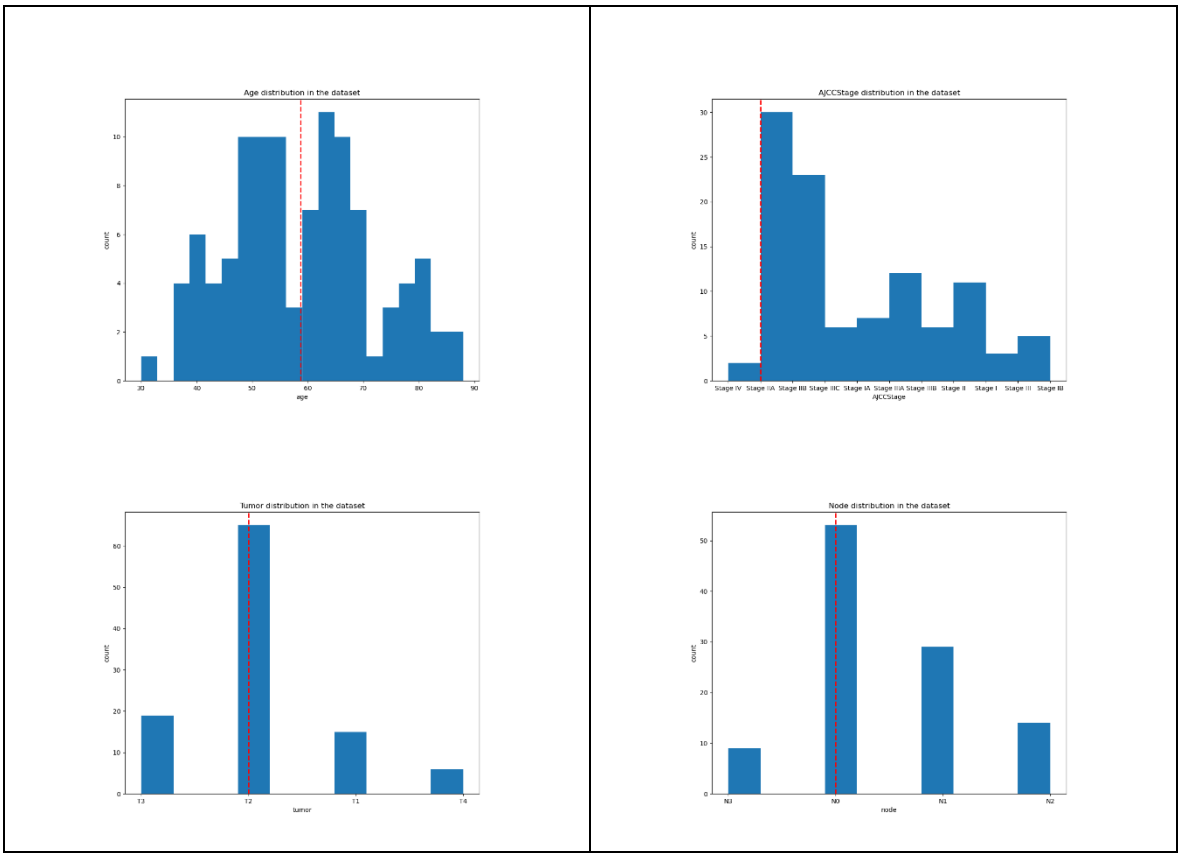
فصل ۲–شرح نتایج

خروجی نتایج مسئله به شرح زیر است.

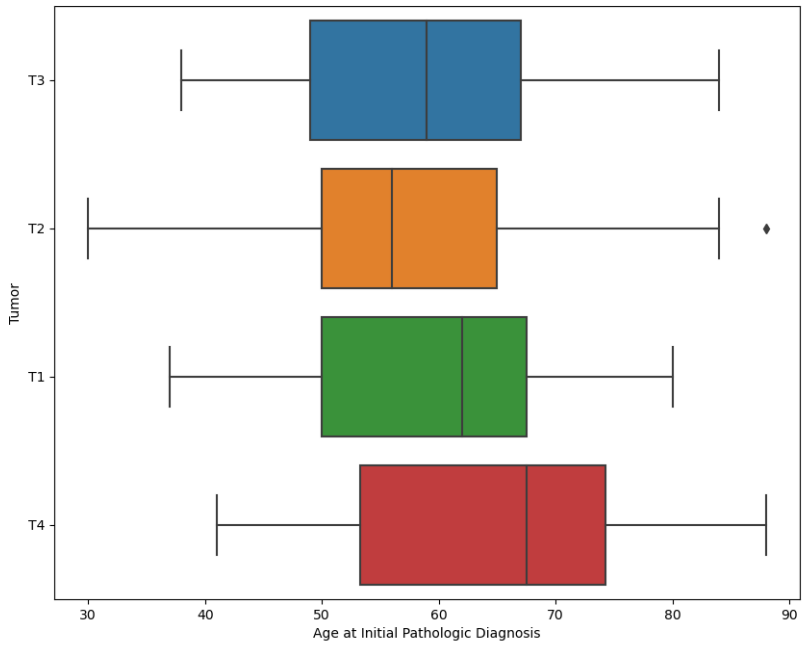


شکل ۱-۲- خروجی کد برنامه برای مجموعه داده اول

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.

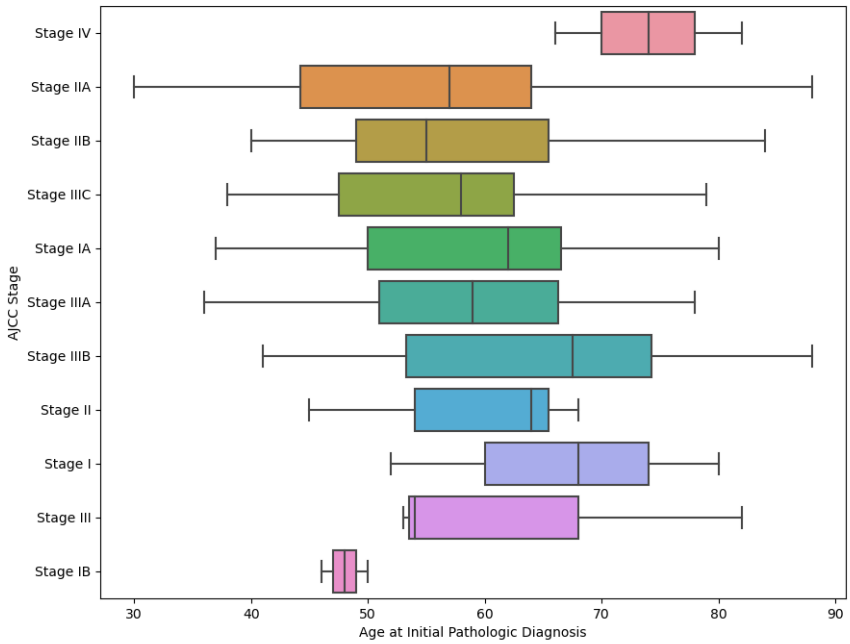


شکل ۲-۲: نمودارهای هیستوگرام از سن و نوع تومور و نود و AJCC Stage



شکل ۳-۲: نمودار باکس پلات از سن افراد و نوع تومور آن‌ها

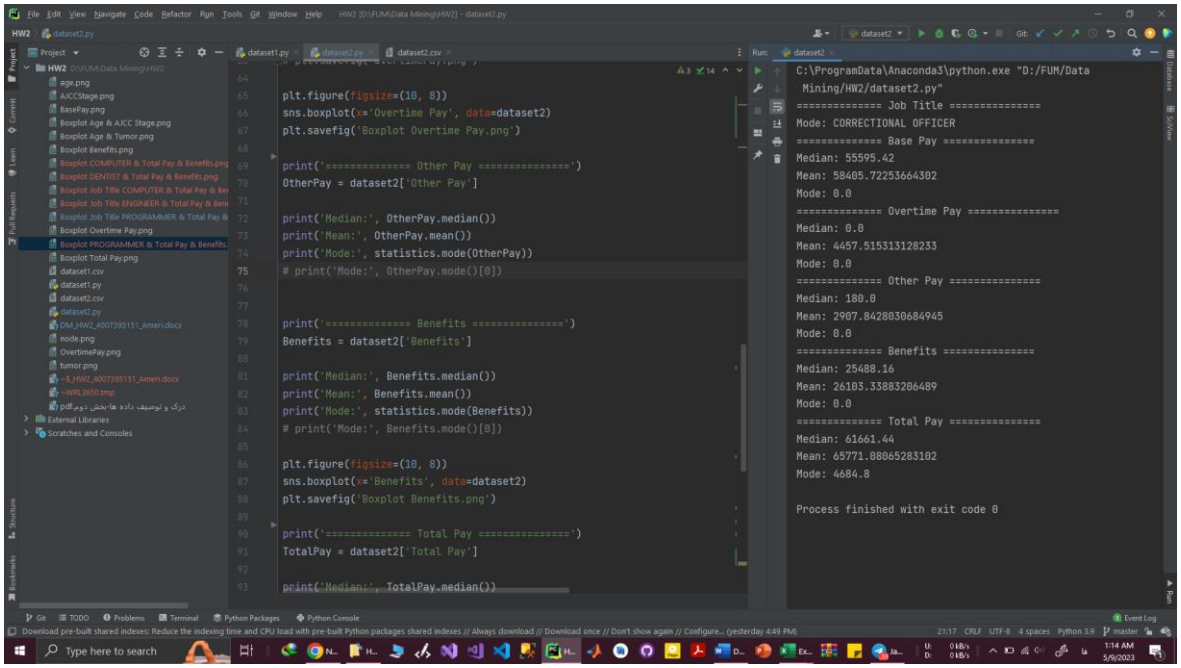
همانطور که در تصویر بالا مشاهده می‌کنید، بیشتر افرادی که تومور دارند سنی بین ۵۰ تا ۷۵ دارند؛ همچنین تومور نوع T4 در بین افراد با سن بالاتر شایع‌تر است. بیشترین تعداد افرادی که داده‌های آن‌ها در این نمودار نمایش داده شده است، در سنین ۶۰ تا ۷۰ سال هستند و اندازه تومورشان در حدود ۲ تا ۳ سانتیمتر است. همچنین در نمودار مشاهده می‌شود که افرادی که سنشان بین ۳۰ تا ۵۰ 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 ۴-۲: نمودار باکس پلات از میزان گسترش تومور بر اساس سن افراد

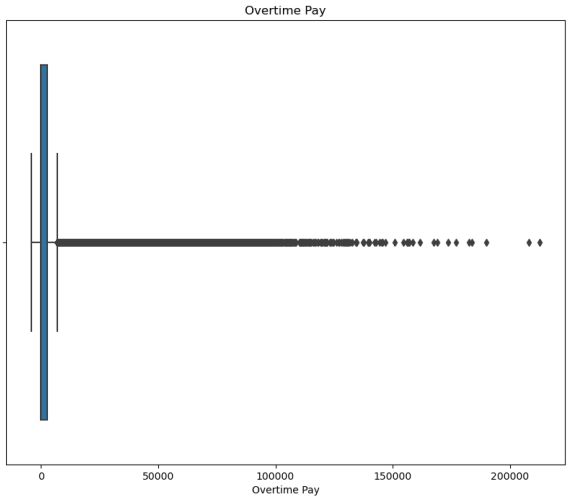
همانطور که در تصویر بالا مشاهده می‌کنید، به عنوان مثال افراد بین ۷۰ تا ۸۰ سال در مرحله چهارم از میزان گسترش تومور هستند. سن متوسط در تشخیص اولیه پاتولوژیک حدود ۶۰ سال است. دامنه سنی بسیار گسترده است، از ۳۰ تا ۸۸. محدوده بین چارکی (IQR) تقریباً بین ۴۸ تا ۶۷ است، به این معنی که ۵۰٪ از داده‌ها در این محدوده 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 ۴۳ تا ۷۴ گسترش یافته است.

برای مجموعه داده دوم داریم:



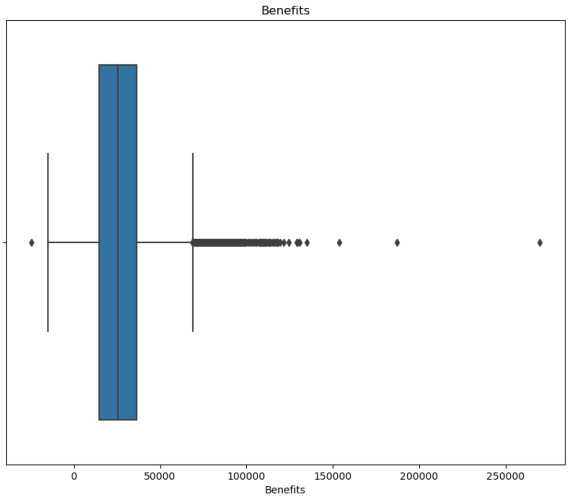
شکل ۵-۲: خروجی کد برنامه برای مجموعه داده دوم

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.



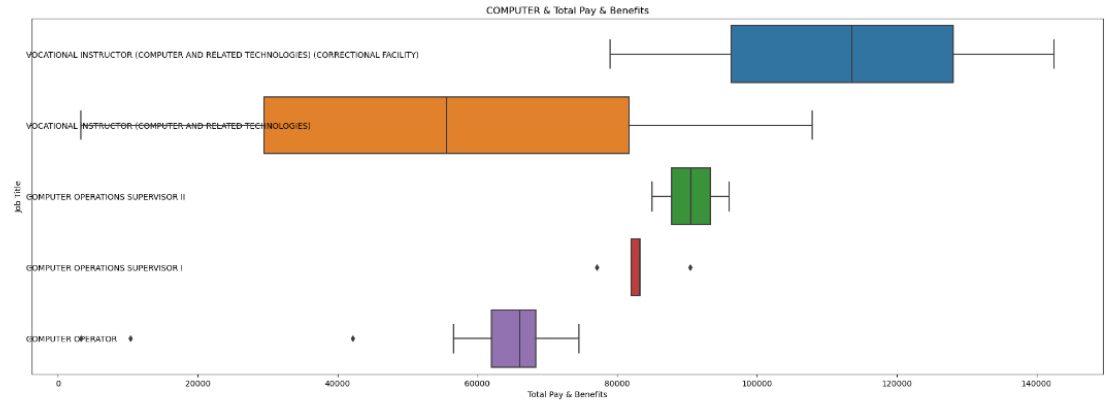
شکل ۶-۲: نمودار باکس پلات حقوق اضافه کاری

همانطور که در تصویر بالا مشاهده می‌کنید، در اکثر شغل‌ها حقوق اضافه کاری صفر هست؛ اما در چند مورد خاص پرداخت اضافه کاری به ۲۰۰۰۰۰ هم رسیده است.



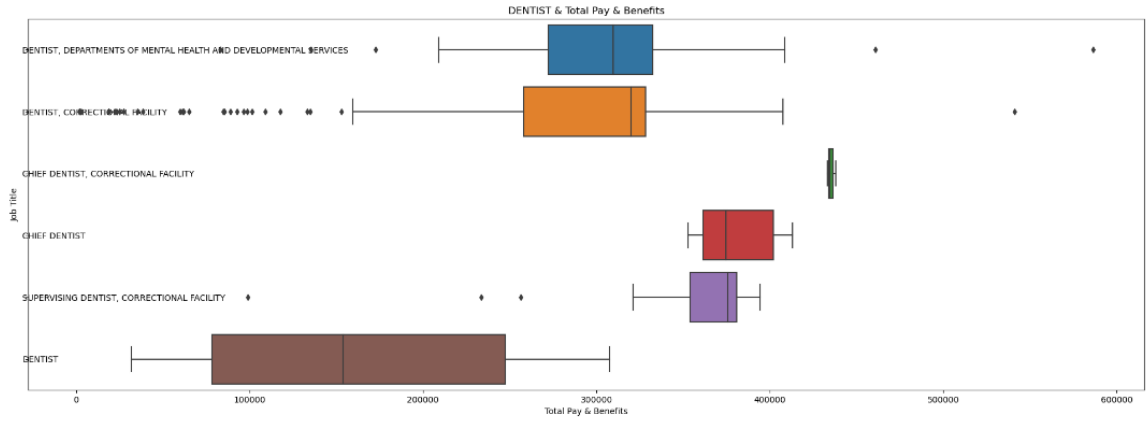
شکل ۷-۲: نمودار باکس پلات مزایا

همانطور که در تصویر بالا مشاهده می‌کنید، مزایا چیزی حدود صفر تا ۷۰۰۰۰ است و در موارد خاص به ۲۰۰۰۰۰ هم رسیده است.



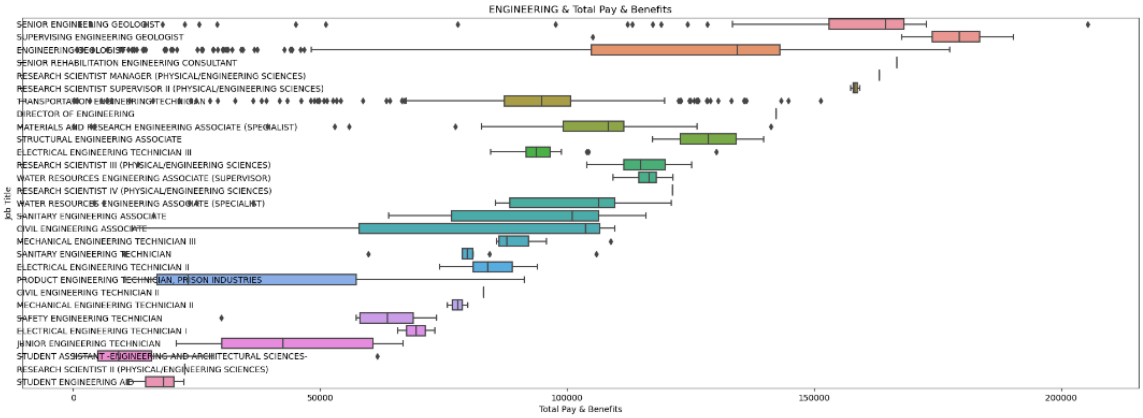
شکل ۸-۲: نمودار باکس پلات از نوع شغل کامپیوتری و درآمد و مزیا آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، انواع شغل‌های کامپیوتری درامدی بین ۳۰۰۰ تا ۱۴۰۰۰۰ دلار دارند. و همانطور که می‌بینید بالاترین درآمد در بین شغل‌های کامپیوتری مربی حرفه‌ای (کامپیوتر و فناوری‌های مرتبط) (تأسیسات اصلاحی) است. و کمترین متعلق به اوپراتور کامپیوتر است.



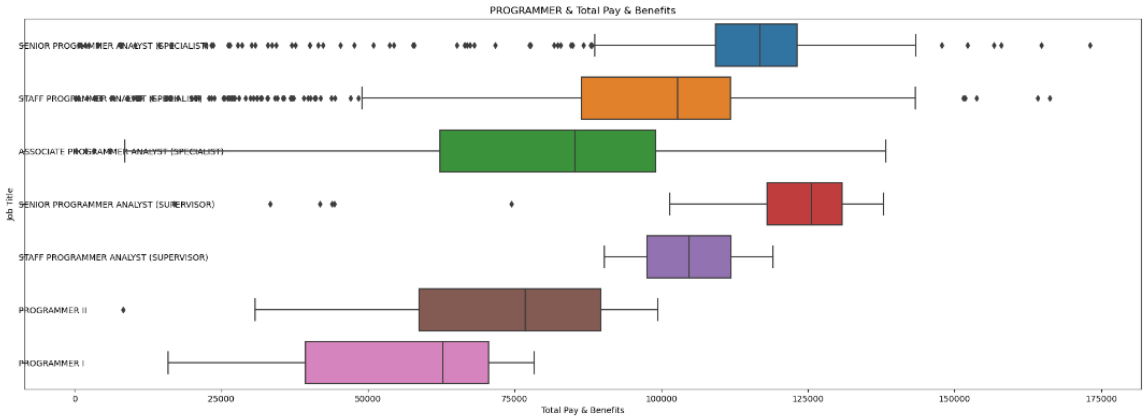
شکل ۹-۲: نمودار باکس پلات از نوع شغل دندان پزشکی و درآمد و مزیا آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، انواع شغل‌های دندان پزشکی درامدی بین ۹۰۰۰۰ تا ۴۵۰۰۰۰ دلار دارند. و همانطور که می‌بینید بالاترین درآمد در بین شغل‌های دندان پزشکی، دندانپزشک، بخش سلامت روان و خدمات رشد است. و کمترین متعلق به دندانپزشک است.



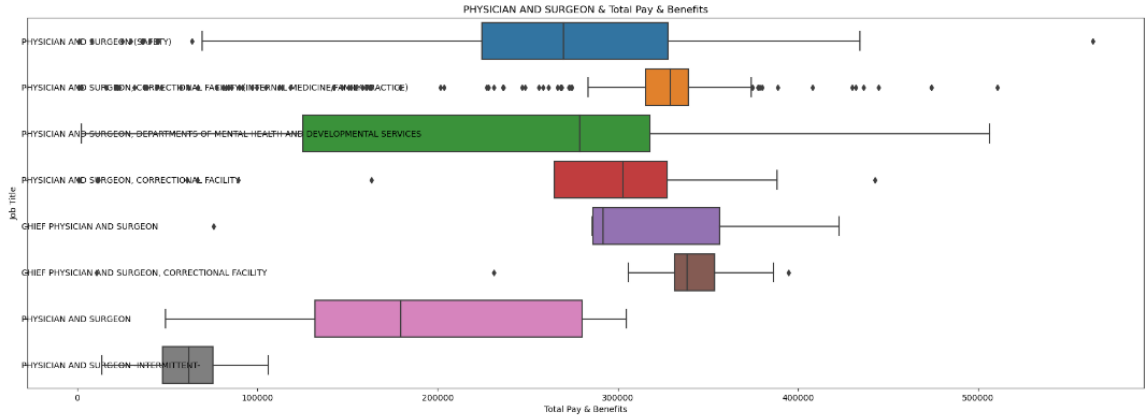
شکل ۱۰-۲: نمودار باکس پلات از نوع شغل مهندسی و درآمد و مزیا آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، انواع شغل‌های مهندسی درامدی بین صفر تا ۲۰۰۰۰۰ دلار دارند. و همانطور که می‌بینید بالاترین درآمد در بین شغل‌های مهندسی، کارشناس ارشد زمین شناس مهندسی است. و کمترین متعلق به کمک مهندسی دانشجویی است.



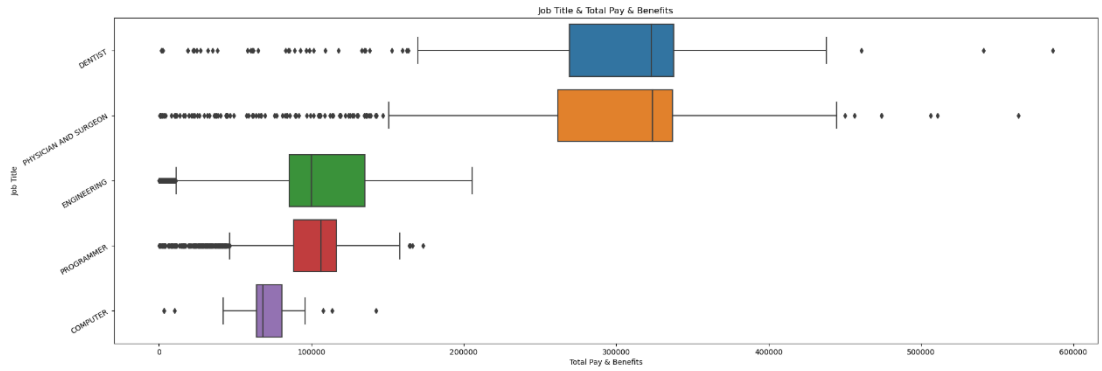
شکل ۱۱-۲: نمودار باکس پلات از نوع شغل برنامه نویسی و درآمد و مزیا آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، انواع شغل‌های برنامه نویسی درامدی بین ۳۰۰۰۰ تا ۱۵۰۰۰۰ دلار دارند. و همانطور که می‌بینید بالاترین درآمد در بین شغل‌های برنامه نویسی، برنامه نویس ارشد تحلیلگر (متخصص) است. و کمترین متعلق به برنامه نویس سطح یک است.



شکل ۱۲-۲: نمودار باکس پلات از نوع شغل پزشک و جراح و درآمد و مزیا آن‌ها

همانطور که در تصویر بالا مشاهده می‌کنید، انواع شغل‌های پزشکی و جراحی درامدی بین ۵۰۰۰۰ تا ۴۰۰۰۰۰ دلار دارند. و همانطور که می‌بینید بالاترین درآمد در بین شغل‌های پزشکی و جراحی، پزشک و جراح (ایمنی) است. و کمترین متعلق به پزشک و جراح -متناوب- است.



شکل ۱۳-۲: نمودار باکس پلات از انواع شغل و درآمد و مزیا آن‌ها

در نمودار بالا همه شغل‌هایی که قبل‌تر بررسی کردیم را به صورت دسته بندی شده با یکدیگر مقایسه می‌کنیم. همانطور که در تصویر بالا مشاهده می‌کنید، شغل‌های دندان پزشک، جراح و پزشک درآمدی حدود ۳۰۰۰۰۰ دلار دارند. و شغل‌های مهندسی و برنامه نویسی و کامپیوتری درآمدی بین ۷۰۰۰۰ دلار تا ۱۵۰۰۰۰ دلار دارند.

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=30, 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')
```

```
dataset2.loc[JobTitle.str.contains('PROGRAMMER'), 'Job Title'] = 'PROGRAMMER'
dataset2.loc[JobTitle.str.contains('DENTIST'), 'Job Title'] = 'DENTIST'
dataset2.loc[JobTitle.str.contains('COMPUTER'), 'Job Title'] = 'COMPUTER'
dataset2.loc[JobTitle.str.contains('ENGINEERING'), 'Job Title'] = 'ENGINEERING'
dataset2.loc[JobTitle.str.contains('PHYSICIAN AND SURGEON'), 'Job Title'] = 'PHYSICIAN
AND SURGEON'


PROGRAMMER = dataset2[JobTitle.str.contains('PROGRAMMER')]
COMPUTER = dataset2[JobTitle.str.contains('COMPUTER')]
ENGINEERING = dataset2[JobTitle.str.contains('ENGINEERING')]
PHYSICIANANDSURGEON = dataset2[JobTitle.str.contains('PHYSICIAN AND SURGEON')]
DENTIST = dataset2[JobTitle.str.contains('DENTIST')]
data = pd.concat([DENTIST, PHYSICIANANDSURGEON, ENGINEERING, PROGRAMMER, COMPUTER])

boxplot(data, 'Job Title & Total Pay & Benefits')
```