دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

# مبانی داده کاوی

## پاسخنامه تمرین شماره ۴

بهار ۱۴۰۳

# فهرست مطالب

# ۱ پاسخ سوالات

## ۱.۱ پاسخ سوال ۱

(آ) :

- میانگین:

۲۰ + ۱۸ + ۲۷ + ۲۴ + ۲۲ + ۱۹ + ۱۶ + ۱۷ + ۲۰ + ۲۳ + ۱۸ + ۱۵ + ۲۱ = ۲۶۰

۲۶ ÷ ۱۳ = ۲۰

- میانه:ارزش عددی واقع شده در وسط یک مجموعه داده

[۱۵, ۱۶, ۱۷, ۱۸, ۱۸, ۱۹, ۲۰, ۲۰, ۲۱, ۲۲, ۲۳, ۲۴, ۲۷]

که با ۲۰ برابر است.

- مد: عدد یا اعداد دارای بیشترین تکرار در مجموعه که ۱۸ و ۲۰ است.
- واریانس:

$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

که $\bar{x}$ میانگین نمونه است و $x_i$ مقدار هر نمونه می باشد و n نیز تعداد نمونه هاست. حاصل در نهایت برابر با ۱۰٫۶۲ است.

پاسخ گستردهٔ یکی از دانشجویان به عنوان نمونه آمده است:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$= \frac{1}{13}((15-20)^2 + (16-20)^2 + (17-20)^2 + (18-20)^2 + (18-20)^2$$
$$+ (19-20)^2 + (20-20)^2 + (20-20)^2 + (21-20)^2 + (22-20)^2$$
$$+ (23-20)^2 + (24-20)^2 + (27-20)^2) = \frac{138}{13} \approx 10.62$$

- چارک ها

مشاهده ای از مجموعه داده های مورد بررسی است که یک چهارم داده ها (یعنی ۲۵ درصد مشاهدات ) از آن کوچکتر و سه چهارم داده ها (یعنی ۷۵ درصد مشاهدات ) از آن بزرگتر می باشد.

روش بدست آوردن چارک اول: ابتدا میانه داده ها را بدست آورده سپس برای نیمه اول داده ها (از کوچکترین عدد تا میانه ) مجددا یکبار دیگر میانه را محاسبه می نماییم. این عدد که میانه نیمه اول داده ها است همان چارک اول می باشد.

چارک دوم چارک دوم همان میانه می باشد، داده ای که ۵۰ درصد (نیمی ) از مشاهدات از آن کوچکتر یا مساوی و ۵۰ درصد (نصف دیگر ) از آن بزرگتر می باشند.

چارک سوم مشاهده ای از مجموعه داده های مورد بررسی است که سه چهارم داده ها (یعنی ۷۵ درصد مشاهدات) از آن کوچکتر و یک چهارم داده ها (یعنی ۲۵ درصد مشاهدات) از آن بزرگتر می باشد.

روش بدست آوردن چارک سوم: ابتدا میانه داده ها را بدست آورده سپس برای نیمه دوم داده ها (از میانه تا بزرگترین عدد) مجددا یکبار دیگر میانه را محاسبه می نماییم. این عدد که میانه نیمه دوم داده ها است همان چارک سوم می باشد.

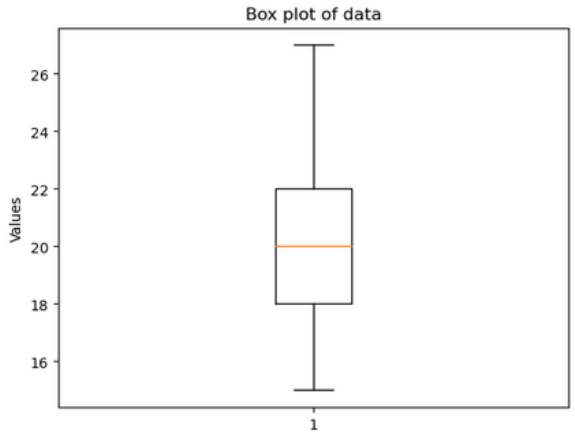
برای این سوال چارک اول برابر با ۱۷٫۵، چارک دوم برابر با ۲۰ و چارک سوم برابر با ۲۲٫۵ است.

(ب) استفاده از کتابخانه های مختلف در پایتون مجاز است و در ادامه یک نمونه برای مثال مطرح شده است:

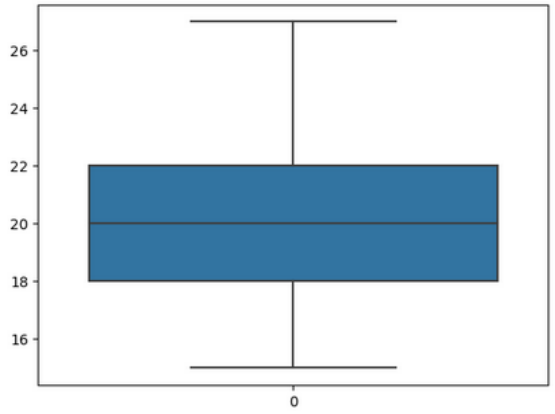
```
import numpy as np

data = [20, 18, 27, 24, 22, 19, 16, 17, 20, 23, 18, 15, 21]

# Mean
mean = np.mean(data)

# Median
median = np.median(data)

# Quartiles
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)

# Variance
variance = np.var(data)

print("Mean:", mean)
print("Median:", median)
print("Q1:", q1)
print("Q3:", q3)
print("Variance:", variance)
```

(ج) نمودار با دو کتابخانه رسم شده است:

```
In [1]: import matplotlib.pyplot as plt
data = [20, 18, 27, 24, 22, 19, 16, 17, 20, 23, 18, 15, 21]
plt.boxplot(data)
plt.show()
```




```
In [2]: import seaborn as sns
data = [20, 18, 27, 24, 22, 19, 16, 17, 20, 23, 18, 15, 21]
sns.boxplot(data=data)
plt.show()
```




(د) نمونه ای از محاسبه Z-Score که توسط یکی از دانشجویان به طور گسترده نوشته شده، درج شده است.

$$\rightarrow \sigma = \sqrt{10.62} = 3.26$$

| داده | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|
| مقدار Z | $\frac{15-20}{3.26} = -1.53$ | $\frac{16-20}{3.26} = -1.23$ | $\frac{17-20}{3.26} = -0.92$ | $\frac{18-20}{3.26} = -0.61$ | $\frac{19-20}{3.26} = -0.31$ | $\frac{20-20}{3.26} = 0$ |
| داده | 21 | 22 | 23 | 24 | 27 | |
| مقدار Z | $\frac{21-20}{3.26} = 0.31$ | $\frac{22-20}{3.26} = 0.61$ | $\frac{23-20}{3.26} = 0.92$ | $\frac{24-20}{3.26} = 1.23$ | $\frac{27-20}{3.26} = 2.15$ | |

فرمول محاسبه:

$$z = \frac{x-\mu}{\sigma}$$

نماد $\sigma$ برابر است با انحراف استاندارد مجموعه داده

نماد $\mu$ برابر است با میانگین مجموعه داده

و خود x هم در هر نمونه جایگذاری می شود.

(ه) با استفاده از کتابخانه scipy این بخش انجام شده است:

```
In [3]: from scipy.stats import zscore

        data = [20, 18, 27, 24, 22, 19, 16, 17, 20, 23, 18, 15, 21]

        z_scores = zscore(data)

        print("Z-Scores:", z_scores)
```

```
Z-Scores: [ 0.         -0.61384981  2.14847435  1.22769963  0.61384981 -0.30692491
 -1.22769963 -0.92077472  0.          0.92077472 -0.61384981 -1.53462454
  0.30692491]
```

## ۲.۱ پاسخ سوال ۲

| داده | رکورد | گراف | ترتیبی |
|---|---|---|---|
| آ | | | * |
| ب | * | | |
| ج | * | | |
| د | | * | |
| ه | | | * |
| و | | * | |
| ز | | * | |
| ح | | | * |
| ط | * | | |
| ی | | | * |
| ک | | * | |
| ل | * | | |
| م | | * | |
| ن | | * | |
| س | | | * |

## ۳.۱ پاسخ سوال ۳

به دلیل اشتباهی که اکثر دوستان در مورد توان در این سوال دچار شدن، نمره این سوال به صورت امتیازی حساب می‌شود و نمره این سوال برای کسانی که سوال را اشتباه پاسخ داده اند، روی سوالات دیگر پخش می شود.

پاسخ صحیح یکی از دانشجویان در ادامه قرار داده شده است.

$$Mahalanobis_i = [(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)]^{0.5}$$

$$x_i - \mu = \begin{bmatrix} 5 - 4.7 \\ 8 - 4.7 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 3.3 \end{bmatrix}$$

$$(x_i - \mu)^T = [0.3 \quad 3.3]$$

$$\Sigma^{-1} = \frac{1}{(110 \times 51) - (41 \times 14)} \begin{bmatrix} 51 & -41 \\ -14 & 110 \end{bmatrix} = \begin{bmatrix} 0.010 & -0.008 \\ -0.003 & 0.022 \end{bmatrix}$$

$$Mahalanobis_i = ([0.3 \quad 3.3] \begin{bmatrix} 0.010 & -0.008 \\ -0.003 & 0.022 \end{bmatrix} \begin{bmatrix} 0.3 \\ 3.3 \end{bmatrix})^{0.5} = ([-0.0069 \quad 0.0702] \begin{bmatrix} 0.3 \\ 3.3 \end{bmatrix})^{0.5}$$
$$= \sqrt{0.22959} = 0.479$$

برای فهم بهتر، کد پایتون برای حل این سوال نیز قرار داده شده است:

```
import numpy as np

# Data
X = 5
Y = 8
mean_vector = np.array([4, 7])
covariance_matrix = np.array([[10, 4], [4, 5]])

# Calculate (X - mean)
data_point = np.array([X, Y])
diff_mean = data_point - mean_vector

# Calculate inverse of covariance matrix
covariance_inv = np.linalg.inv(covariance_matrix)

# Calculate Mahalanobis Distance
mahalanobis_distance = np.sqrt(np.dot(np.dot(diff_mean,
 covariance_inv), diff_mean))

print("Mahalanobis Distance:", mahalanobis_distance)
```

## ۴.۱ پاسخ سوال ۴

مطابق نکته ای که در ویدیوی راهنمای تمرین سری ۲ برای این سوال بیان شد، بایستی مجموعهٔ داده شده به صورت برداری درآید تا بتوان مقادیر خواسته شده را برای آن محاسبه کرد.

مطابق فرمول ها برای هر کدام از موارد Jaccard و smc تعداد را محاسبه و سپس در فرمول قرار می دهیم:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

شکل ۱: رابطهٔ jaccard

$$\text{SMC} = \frac{\text{number of matching attributes}}{\text{total number of attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{01} + M_{10}}$$

شکل ۲: رابطهٔ smc

از میان پاسخ های صحیح دانشجویان یک مورد به عنوان نمونه قرار داده شده است:

مجموعه مرجع: {apple, banana, cherry, date, elderberry}

A = 11110 B = 01111

SMC = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) = (3 + 0) / 5 = 3/5 = 0.6$

Jaccard = $f_{11} / (f_{01} + f_{10} + f_{11}) = 3 / (1+1+3) = 3/5 = 0.6$

در اینجا [۱] و اینجا [۲] مثال های بیشتری مطرح شده است. مثال مشابهی نیز در اینجا [۳] بیان شده است.

## ۵.۱ پاسخ سوال ۵

از میان پاسخ های صحیح دانشجویان یک مورد به عنوان نمونه قرار داده شده است:

[1] https://www.youtube.com/watch?v=R0ORgpi25d0
[2] https://www.youtube.com/watch?v=kZE0ytXGqgw
[3] https://en.wikipedia.org/wiki/Talk%3AJaccard_index

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$H(x) = -(p_A \log_2 p_A + p_B \log_2 p_B + p_C \log_2 p_C)$$
$$= -(0.4 \log_2 0.4 + 0.3 \log_2 0.3 + 0.3 \log_2 0.3)$$

$$\log_2 0.4 \cong -1.322$$

$$\log_2 0.3 \cong -1.737$$

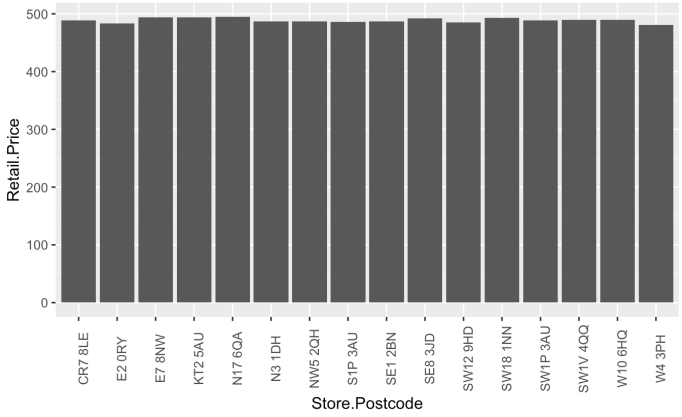
$$H(x) = -((0.4 * -1.322) + ((0.3 * -1.737) * 2)) \approx 1.571$$

## ۶.۱ پاسخ سوال ۶

۱. الف

کمترین میانگین فروش: W4 3PH

بالاترین میانگین فروش: N17 6QA

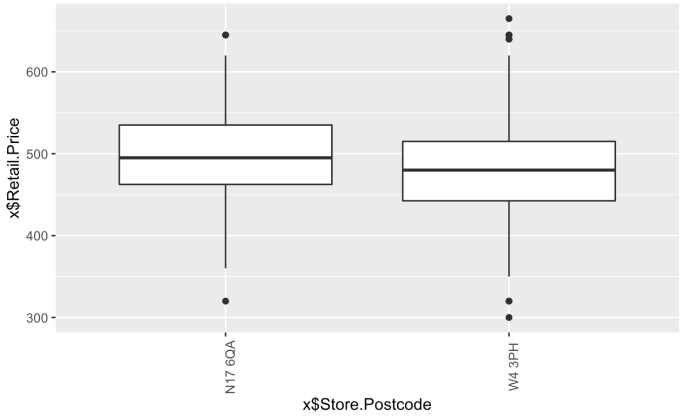



```python
# قسمت اول سوال ششم
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("LaptopSalesJanuary2008.csv")

# ایجاد نمودار میله ای برای نمایش میانگین قیمت خرده فروشی بر اساس فروشگاه
mean_prices = df.groupby('Store Postcode')['Retail Price'].mean()
mean_prices.plot(kind='bar', ylabel='Mean Price', title='Mean Retail Price by Shop')
max_store = mean_prices.index[mean_prices.argmax()]
min_store = mean_prices.index[mean_prices.argmin()]

print(f"فروشگاه با بالاترین میانگین: {max_store}")
print(f"فروشگاه با کمترین میانگین: {min_store}")
plt.show()
```

شکل ۳: نمونه کد یکی از دانشجویان

۲. ب

مطابق شکل زیر، میانگین فروش و همچنین چارک ۳ و ۱ فروشگاه فروش بالاتر (سمت چپ) قوی تر است. علاوه بر این، فروشگاه W43PH دارای چند علامت پرت اضافی است که نیاز به بررسی بیشتر دارد.




```python
# قسمت دوم سوال ششم
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("LaptopSalesJanuary2008.csv")

# رسم نمودار جعبه ای برای همه فروشگاه ها
plt.figure(figsize=(10, 6))
plt.boxplot(df.groupby('Store Postcode')['Retail Price'].apply(list), labels=df['Store Postcode'].unique())
plt.title('Comparison of Retail Price Distribution for All Shops')
plt.xlabel('Store Postcode')
plt.ylabel('Retail Price')
plt.xticks(rotation=45)
plt.show()

# یافتن فروشگاه با بیشترین و کمترین میانگین
max_shop = mean_prices.idxmax()
min_shop = mean_prices.idxmin()

# فیلتر کردن داده ها برای فروشگاه با بیشترین میانگین
max_shop_data = df[df['Store Postcode'] == max_shop]['Retail Price']

# فیلتر کردن داده ها برای فروشگاه با کمترین میانگین
min_shop_data = df[df['Store Postcode'] == min_shop]['Retail Price']

# رسم نمودار جعبه ای برای قیمت خرده فروشی در فروشگاه های بیشترین و کمترین میانگین
plt.figure(figsize=(10, 6))
plt.boxplot([max_shop_data, min_shop_data], labels=[max_shop, min_shop])
plt.title('Comparison of Retail Price Distribution between Shops with Highest and Lowest Mean')
plt.xlabel('Store Postcode')
plt.ylabel('Retail Price')
plt.show()
```

شکل ۴: نمونه کد یکی از دانشجویان

## ۷.۱ پاسخ سوال ۷

بخش اول این سوال، حالت پژوهشی داشته اما در بخش دوم انواع نمودار ها با کتابخانه های مختلف و متنوعی توسط دانشجویان رسم شده که می توانید فایل کد های دوستان خود را که در گروه تلگرام قرار داده شده مشاهده کنید.