پاسخ سوال دوم تمرین دوم هوش مصنوعی

گزارش مراحل تحلیل و پیش‌پردازش داده‌ها و همچنین ارائه مدلی برای خوشه‌بندی

الف) اضافه کردن کتابخانه‌های مورد نیاز و خواندن داده‌ها به کمک Pandas

نخست تمامی کتابخانه‌های لازم را وارد می‌کنیم. پس از آن باید با استفاده از تابع read_csv() داده‌ها را از فایل csv خواند. در نتیجه این کار فایل به فرمت DataFrame درخواهد آمد. با استفاده از تابع head() می‌توان چند سطر اول این مجموعه داده را مشاهده کرد.

	Favorite Color	Favorite Music Genre	Favorite Soft Drink	Gender
0	Cool	Rock	7UP/Sprite	F
1	Neutral	Hip hop	Coca Cola/Pepsi	F
2	Warm	Rock	Coca Cola/Pepsi	F
3	Warm	Folk/Traditional	Fanta	F
4	Cool	Rock	Coca Cola/Pepsi	F

ب) شناسایی داده‌های از دست رفته

به سادگی می‌توان با استفاده از متد isna() داده‌های از دست رفته را شناسایی کرد و به کمک sum() تعداد کل missing valueها را برای هر ستون از داده‌ها محاسبه کرد. طبق خروجی بدست آمده هیچ یک از ویژگی‌ها دارای مقادیر از دست رفته نیستند.

	missing value count
Favorite Color	0
Favorite Music Genre	0
Favorite Soft Drink	0
Gender	0

پ) تمیزسازی داده‌ها

در این مرحله باید داده‌های دسته‌بندی شده یا همان categorical را به مقادیر عددی تبدیل کنیم که برای اینکار از label encoder استفاده شده است. شمای نهایی داده‌ها به شکل زیر خواهد بود.

	Favorite Color	Favorite Music Genre	Favorite Soft Drink	Gender
0	0	6	0	0
1	1	2	1	0
2	2	6	1	0
3	2	1	2	0
4	0	6	1	0

ج) انجام EDA و Visualization برای بدست آوردن بینش از داده‌ها

ابتدا برای بدست آوردن اطلاعات آماری این مجموعه داده، می‌توان با استفاده از تابع describe() توصیفی آماری از متغیرهای این مجموعه داده را مشاهده کرد. البته برای درک بهتر نتایج این مرحله پیش از استفاده از label encoder اجرا شده است.

	Favorite Color	Favorite Music Genre	Favorite Soft Drink	Gender
count	66	66	66	66
unique	3	7	4	2
top	Cool	Rock	Coca Cola/Pepsi	F
freq	37	19	32	33

همچنین با استفاده از تابع info() و shape می‌توان تعداد سطرها و ستون‌های داده به همراه نوع آن‌ها را مشاهده کرد.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66 entries, 0 to 65
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Favorite Color       66 non-null     int32
 1   Favorite Music Genre  66 non-null     int32
 2   Favorite Soft Drink   66 non-null     int32
 3   Gender               66 non-null     int32
dtypes: int32(4)
memory usage: 1.2 KB
```

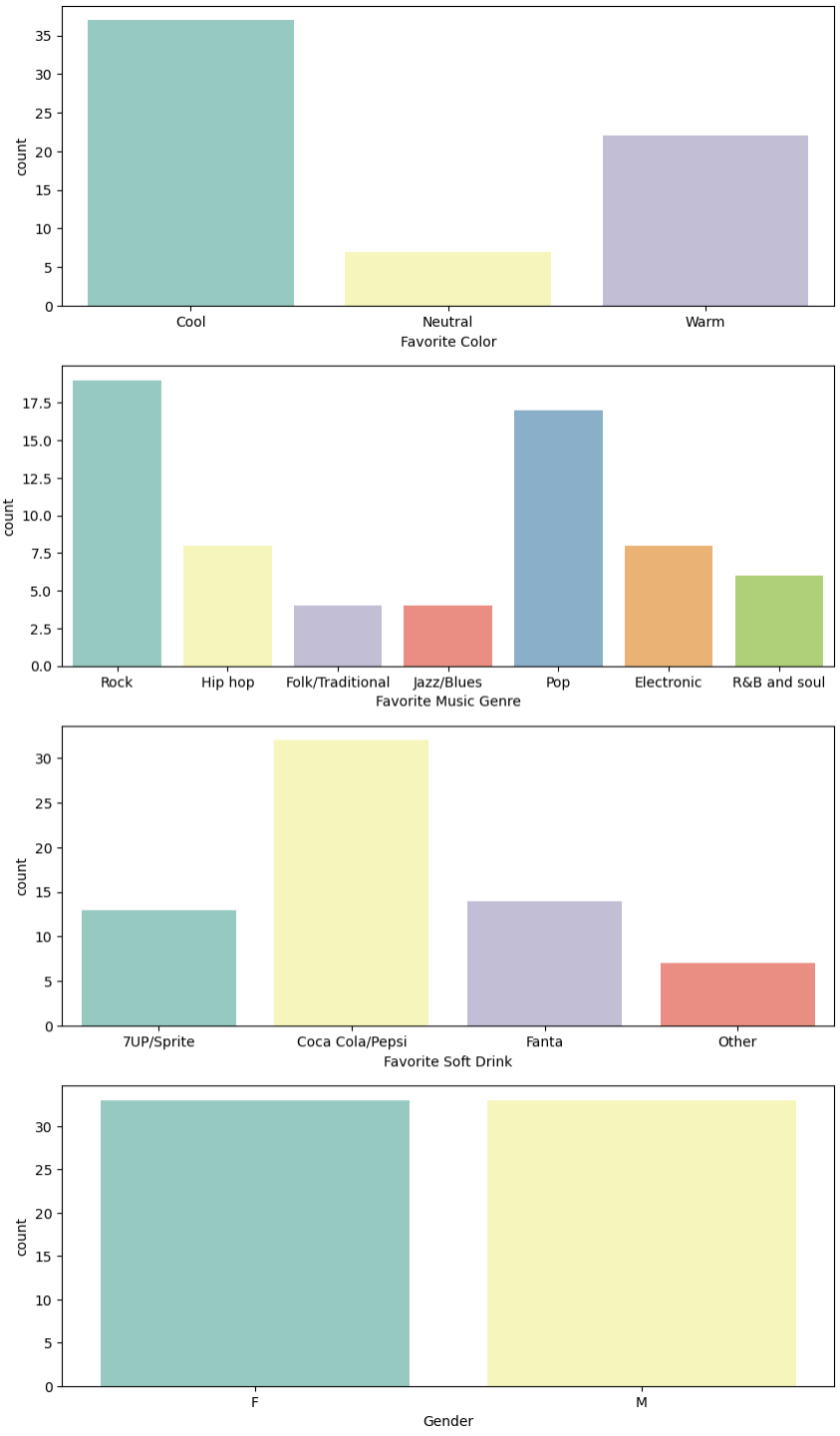
```
df.shape
```

```
(66, 4)
```

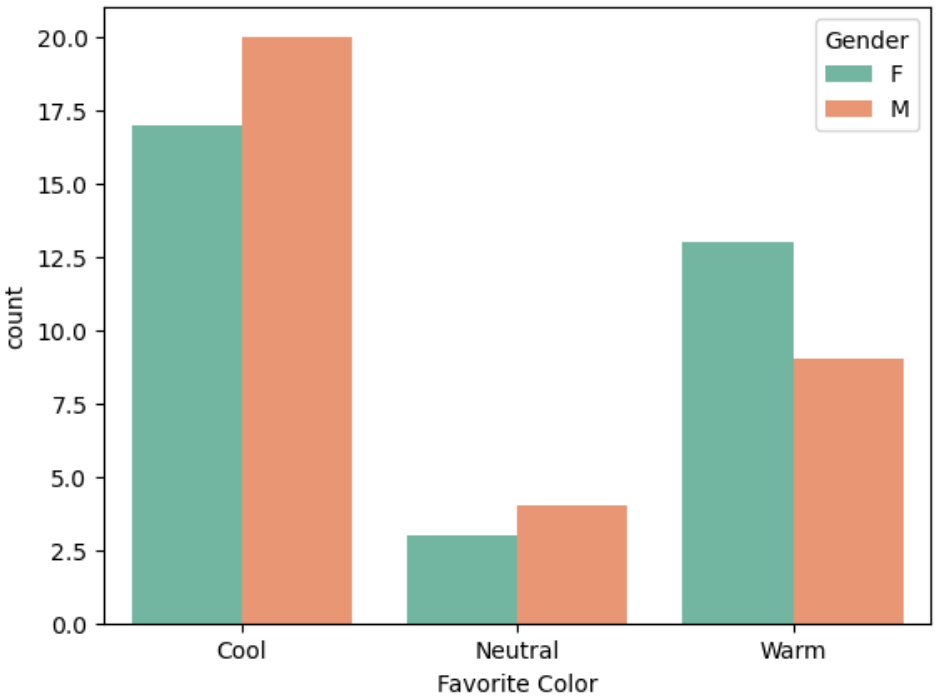
سپس به سراغ رسم نمودارها و data visualization می‌رویم.

نمودار توزیع

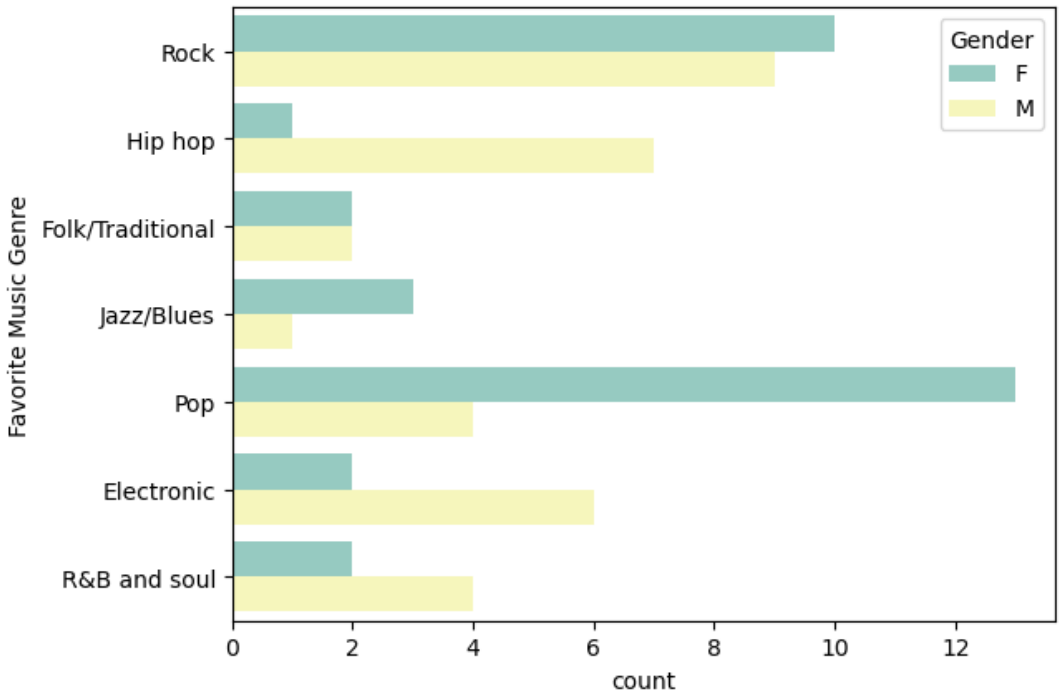
در ادامه با رسم نمودارهایی تعداد مشاهدات در هر دسته از ویژگی‌ها را تصویرسازی می‌کنیم تا دید بهتری نسبت به داده‌ها داشته باشیم.



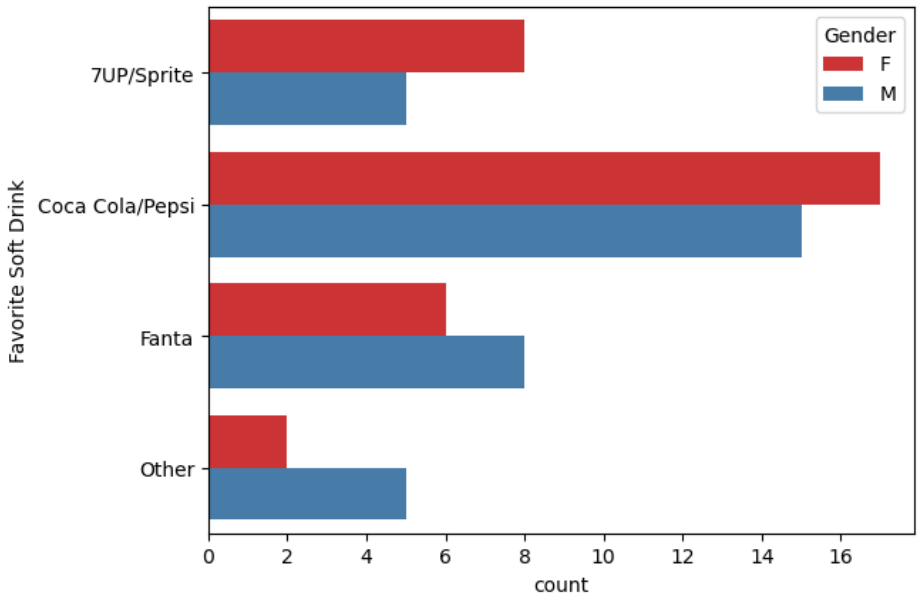
نمودار ۱: رنگ cool، موسیقی راک و پاپ و نوشیدنی coca cola/pepsi از دیگر مقادیر محبوبیت بالاتری دارند.



نمودار ۲: رنگ‌های سرد در بین مردان و رنگ‌های گرم بین زنان محبوب‌ترند



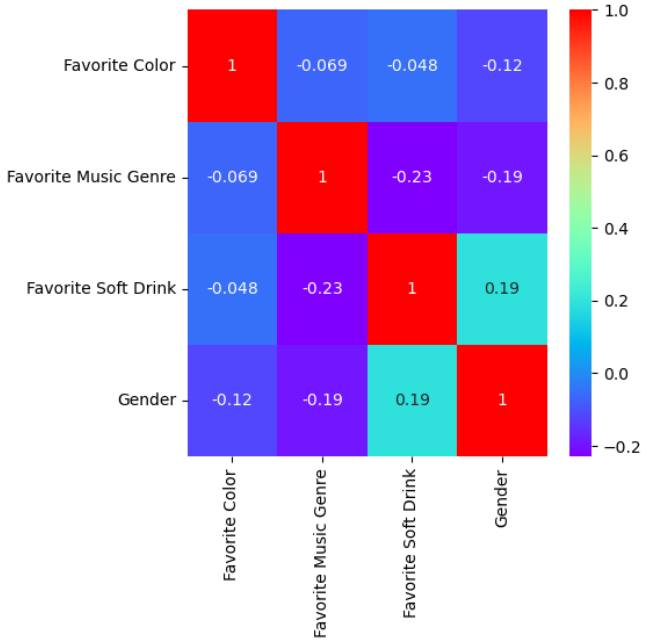
نمودار ۳: موسیقی پاپ، راک و جز بین زنان محبوب‌ترند



نمودار ۴: مردان نوشیدنی فانتا و زنان نوشیدنی‌های اسپرایت و کوکاکولا را ترجیح می‌دهند.

د) رسم ماتریس هبستگی

پیش‌تر داده‌های دسته‌بندی شده یا همان categorical را به داده‌های عددی تبدیل کردیم. حال با رسم نمودار هبستگی میزان همبستگی دوبه‌دوی متغیرها به یکدیگر را نمایش می‌دهیم.



نمودار ۵

ه) آماده‌سازی داده‌ها جهت شروع عملیات خوشه‌بندی

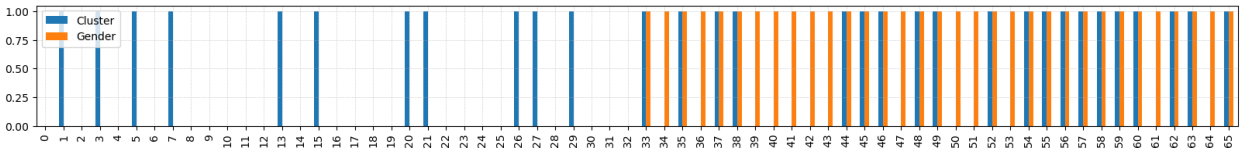
در این مرحله می‌خواهیم با استفاده از داده‌ها مدلی برای خوشه‌بندی ارائه کنیم. ابتدا ویژگی Gender را کنار می‌گذاریم. سپس مقیاس‌بندی ویژگی‌ها را نیز انجام می‌دهیم؛ چراکه برای خوشه‌بندی صحیح به آن نیاز داریم. برای اینکار از هر دو ماژول StandardScaler و MinMaxScale استفاده شد که در نهایت به دلیل نتایج بهتری که استفاده از StandardScaler نسبت به ماژول دیگر بدست آورد، این روش انتخاب شد.

و) خوشه‌بندی اولیه

ابتدا خوشه‌بندی را روی داده‌هایی که ویژگی جنسیت در آن‌ها وجود ندارد انجام می‌دهیم. سپس خوشه‌های تولید شده را باید با مقادیر ستون جنسیت مقایسه کنیم تا ببینیم خوشه‌بندی صورت گرفته مطابق طبقه‌بندی جنسیتی داده‌ها هست یا نه. نتیجه این مرحله نشان می‌دهد که از ۶۶ داده موجود، ۴۲ تای آن‌ها برچسبی مطابق طبقه‌بندی جنسیتی داده‌هاخورده‌اند.

```
Number of correct lables are 42 out of 66.
Accuracy score: 0.64
```

سپس نتیجه بدست آمده را روی نمودار می‌بریم. این نمودار در محور x خود ۶۶ داده موجود را نمایش می‌دهد و در محور y جنسیت را با اعداد ۰ و ۱ نشان می‌دهد. هر داده یک مقدار واقعی که نشان دهنده جنسیت صحیح آن و یک برچسب از خوشه‌بندی دارد که به ترتیب با رنگ‌های نارنجی و آبی ترسیم شده‌اند. این نمودار نشان می‌دهد از بین داده‌ها کدام یک از آن‌ها برچسبی مطابق طبقه‌بندی جنسیتی داده‌ها دریافت کرده‌اند.



نمودار ۶

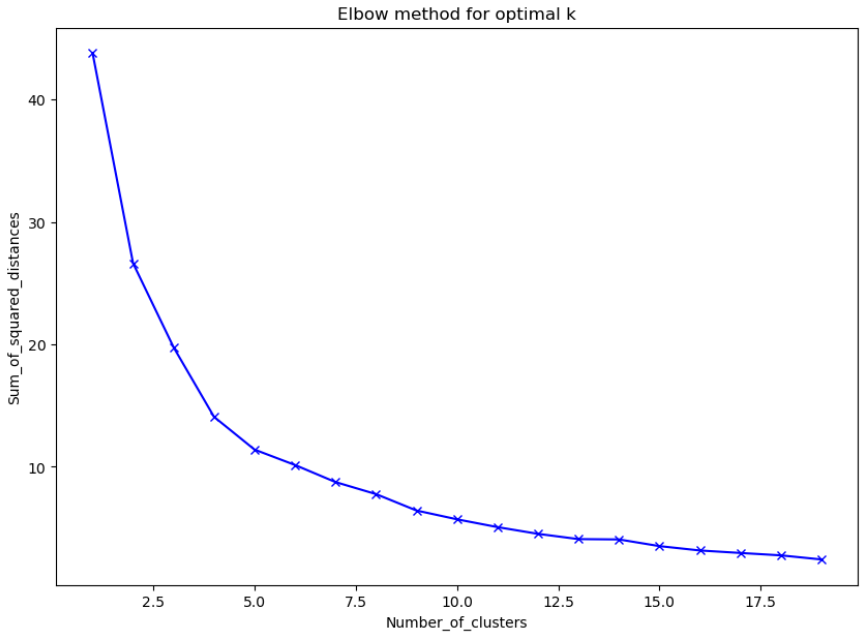
در ادامه تلاش شد تا به کمک نمودار scatter نمایی از خوشه‌بندی داده‌ها به همراه برچسب جنسیت آن‌ها نمایش داده شود که موفقیت‌آمیز نبود.

در نهایت مقادیر دو معیار silhouette score و davies bouldin score برای خوشه‌بندی انجام شده محاسبه شده است. این معیار که می‌تولند بین −۱ و ۱ باشد برای مقدار ۱ نشان می‌دهد که نقطه داده در خوشه‌ای که به آن تعلق دارد بسیار فشرده است و از سایر خوشه‌ها دور است که مطلوب است در مدل به مقداری نزدیک ۱ دست پیدا کنیم. بدترین مقدار −۱ است و در اینجا این معیار به ۱ نزدیک‌تر است تا به −۱؛ پس می‌توان آن را تا حدی خوب در نظر گرفت. معیار دوم یک معیار اعتبارسنجی است که اغلب به منظور ارزیابی تعداد بهینه خوشه‌ها برای استفاده استفاده می‌شود. این معیار به عنوان نسبتی بین پراکندگی خوشه و جدایی خوشه تعریف می‌شود و مقدار کمتر به این معنی است که خوشه بندی بهتر است.

```
Silhouette Coefficient:  0.30194349595273 58
Davies Bouldin Score:  1.411909925941235
```

ی) خوشه‌بندی نهایی

در نهایت روی کل ستون‌های داده از الگوریتم بازو استفاده می‌کنیم تا تعداد خوشه‌های مناسب را بیابیم.



طیق این نمودار به نظر می‌رسد مقدار ۲ برای این مجموعه داده مناسب است. همچنین با محاسبه دقت در آن می‌توان این فرضیه تا تایید کرد.

```
kmeans = KMeans(n_clusters=2)
kmeans.fit(X)
y_clusters = kmeans.fit_predict(X)
if y_clusters[0] != df.Gender[0]:
    for i, label in enumerate(y_clusters):
        if label == 1:
            y_clusters[i] = 0
        else:
            y_clusters[i] = 1
correct_labels = sum(y == y_clusters)
print('Accuracy score: {0:0.2f}'. format(correct_labels/float(y.size)))
```

```
Accuracy score: 1.00
```