Segmentation de client et recommandation de films

Présenté par Rebecca Leygonie et Nemanja Kostadinovic

Analyse exploratoire

Données clients

```
+----------+------+---+-----------------+----------------------+
|CustomerID|Gender|Age|Annual Income (k$)|Spending Score (1-100)|
+----------+------+---+-----------------+----------------------+
|         1|  Male| 19|               15|                    39|
|         2|  Male| 21|               15|                    81|
|         3|Female| 20|               16|                     6|
|         4|Female| 23|               16|                    77|
|         5|Female| 31|               17|                    40|
|         6|Female| 22|               17|                    76|
|         7|Female| 35|               18|                     6|
|         8|Female| 23|               18|                    94|
|         9|  Male| 64|               19|                     3|
|        10|Female| 30|               19|                    72|
+----------+------+---+-----------------+----------------------+
```

```
+-------+-----------------+
|summary|       CustomerID|
+-------+-----------------+
|  count|              200|
|   mean|            100.5|
| stddev|57.879184513951124|
|    min|                1|
|    max|              200|
+-------+-----------------+
```

Null rows of CustomerID 0

```
+-------+------+
|summary|Gender|
+-------+------+
|  count|   200|
|   mean|  null|
| stddev|  null|
|    min|Female|
|    max|  Male|
+-------+------+
```

Null rows of Gender 0

```
+-------+-----------------+
|summary|              Age|
+-------+-----------------+
|  count|              200|
|   mean|            38.85|
| stddev|13.969007331558888|
|    min|               18|
|    max|               70|
+-------+-----------------+
```

```
+-------+-----------------+
|summary|Annual Income (k$)|
+-------+-----------------+
|  count|              200|
|   mean|            60.56|
| stddev| 26.26472116527124|
|    min|               15|
|    max|              137|
+-------+-----------------+
```

Null rows of Annual Income (k$) 0

```
+-------+----------------------+
|summary|Spending Score (1-100)|
+-------+----------------------+
|  count|                   200|
|   mean|                  50.2|
| stddev|     25.823521668370173|
|    min|                     1|
|    max|                    99|
+-------+----------------------+
```
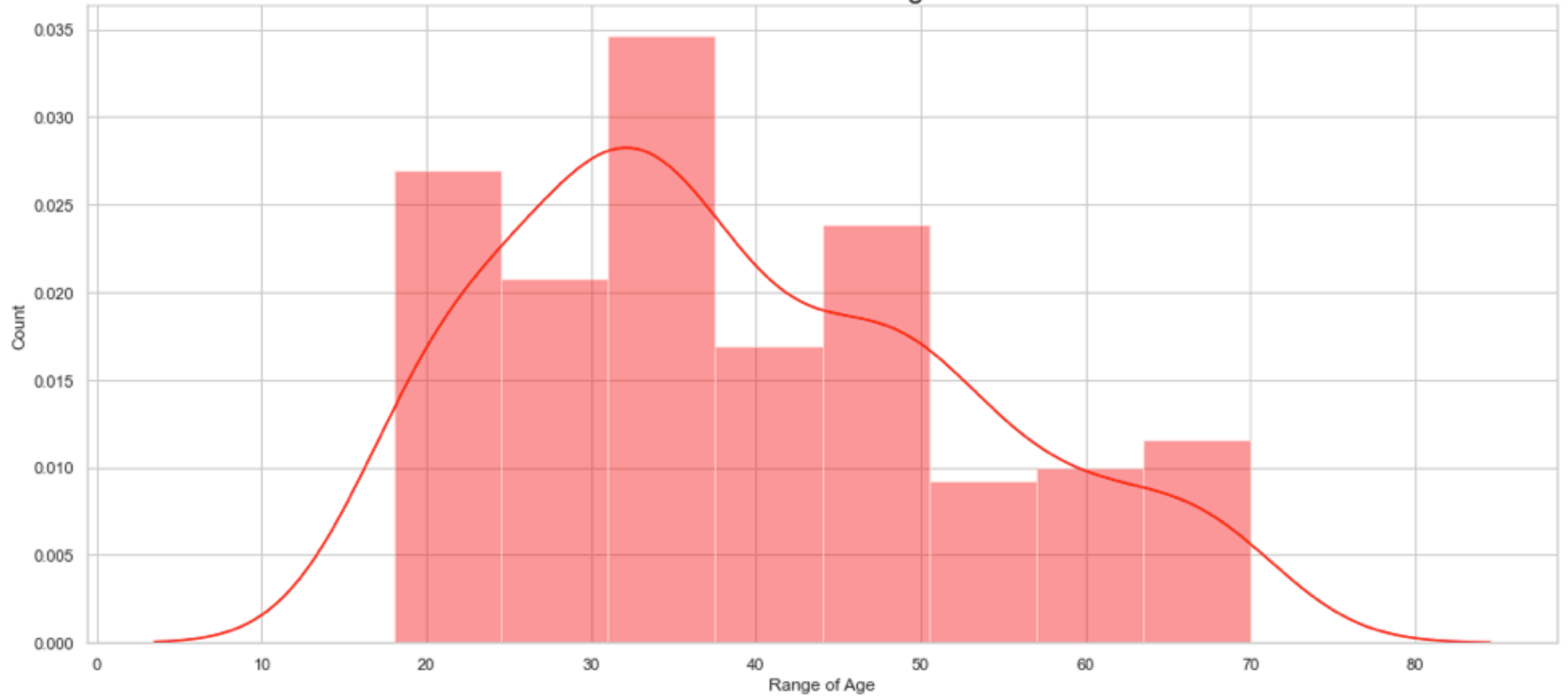
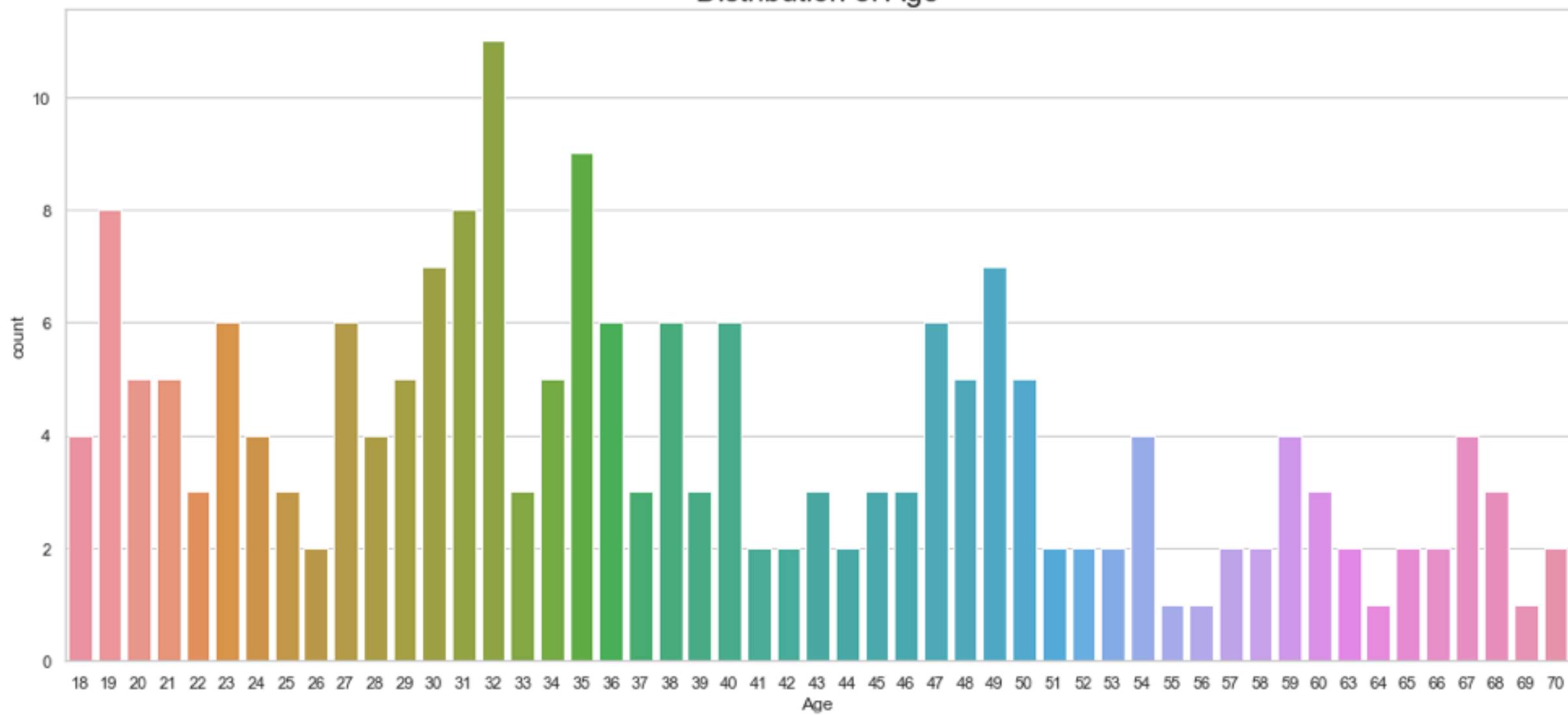Null rows of Spending Score (1-100) 0

Distribution of Annual Income

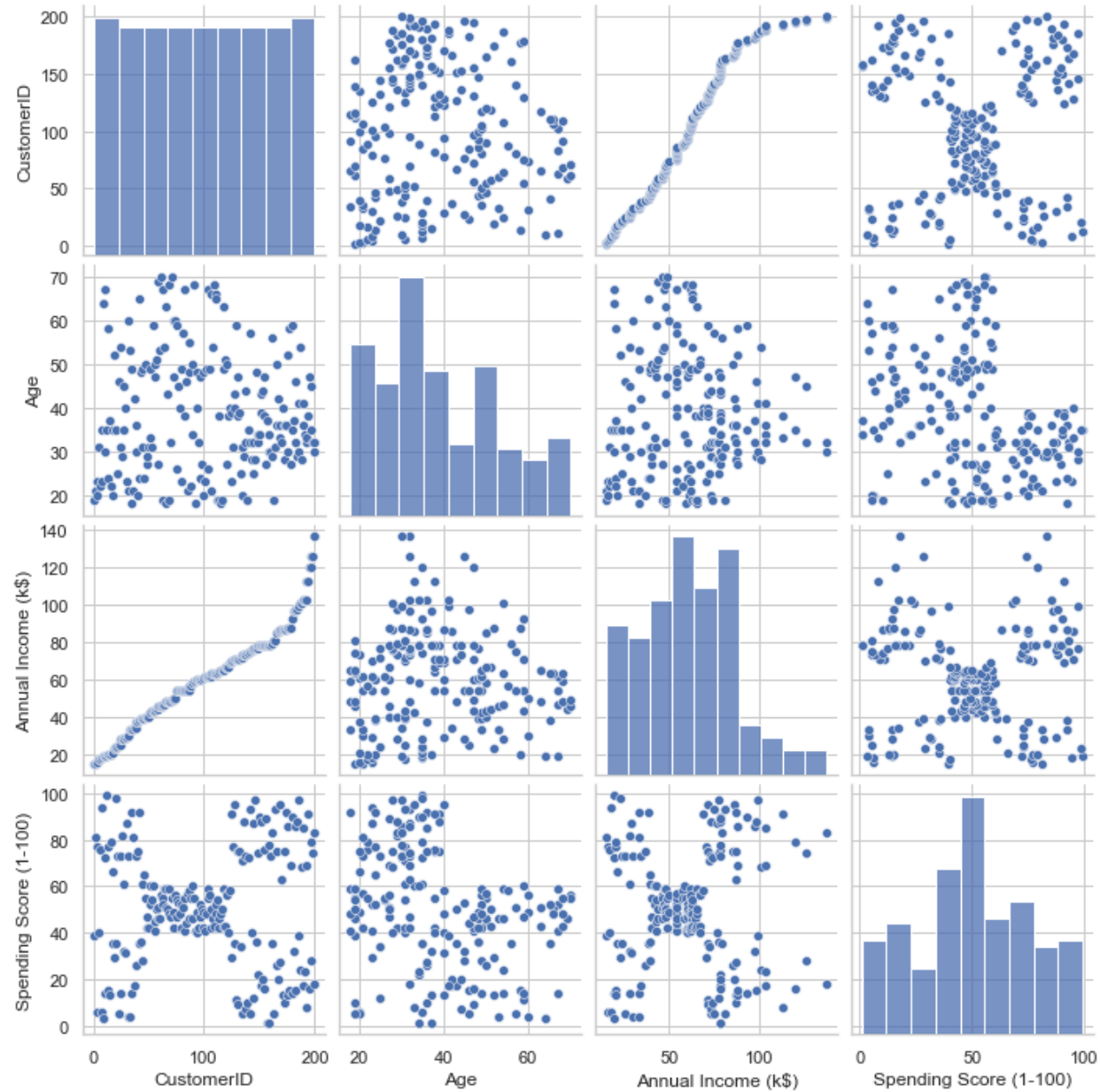Distribution of Annual Income
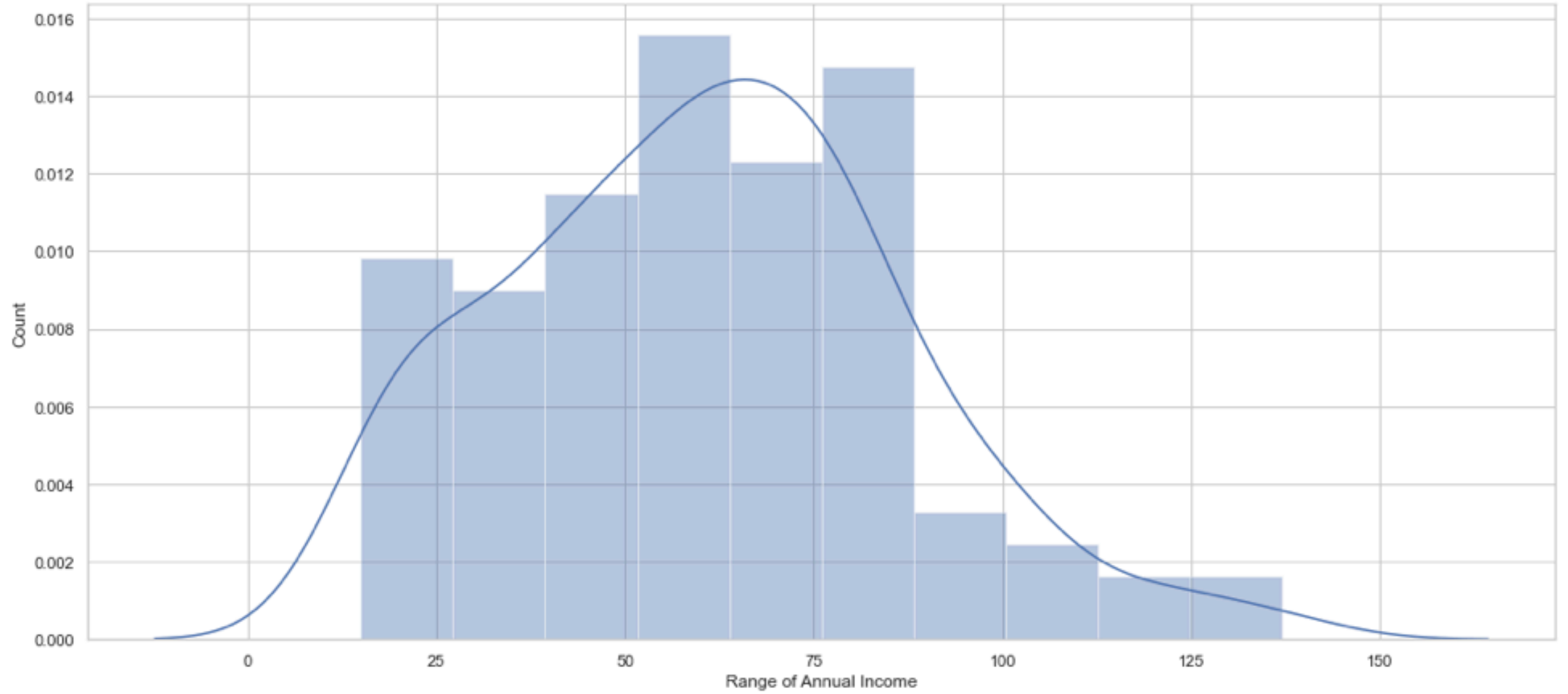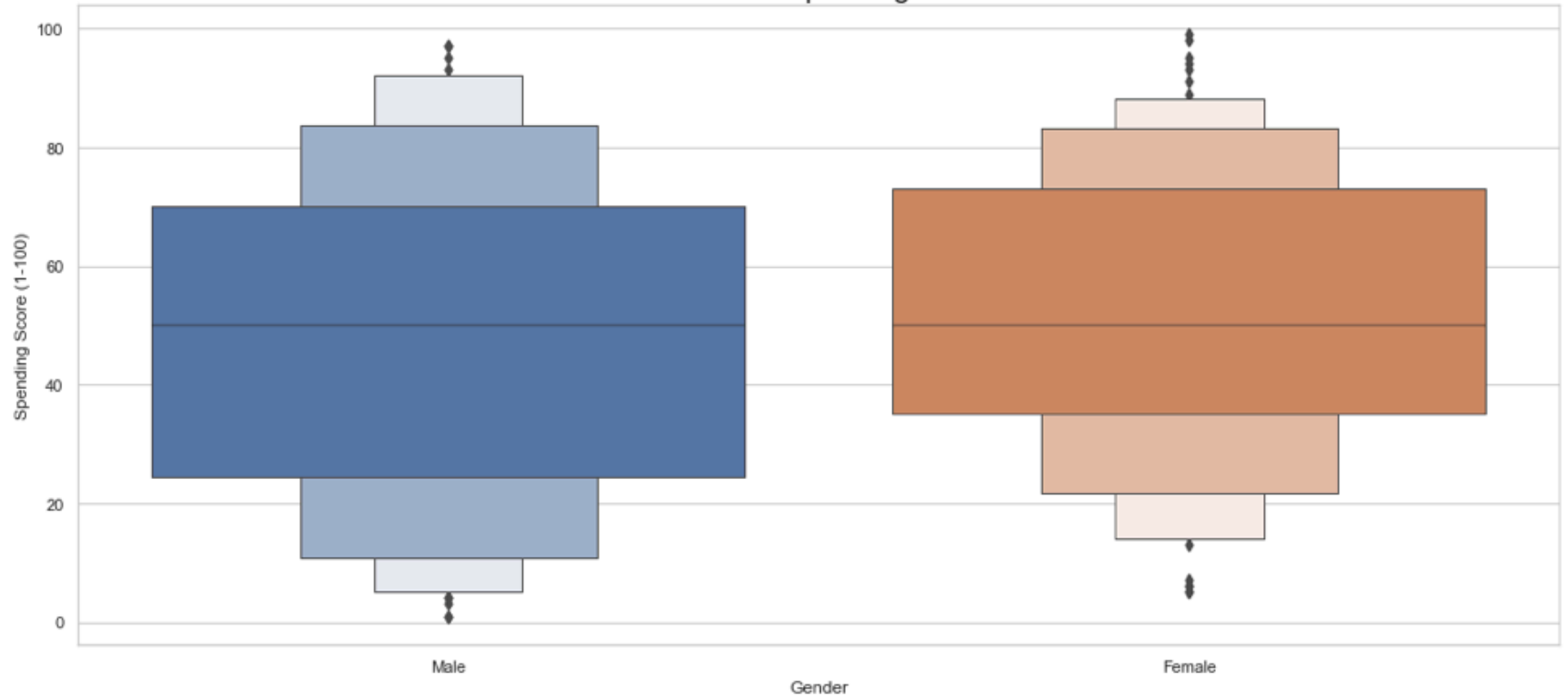
Distribution of Age
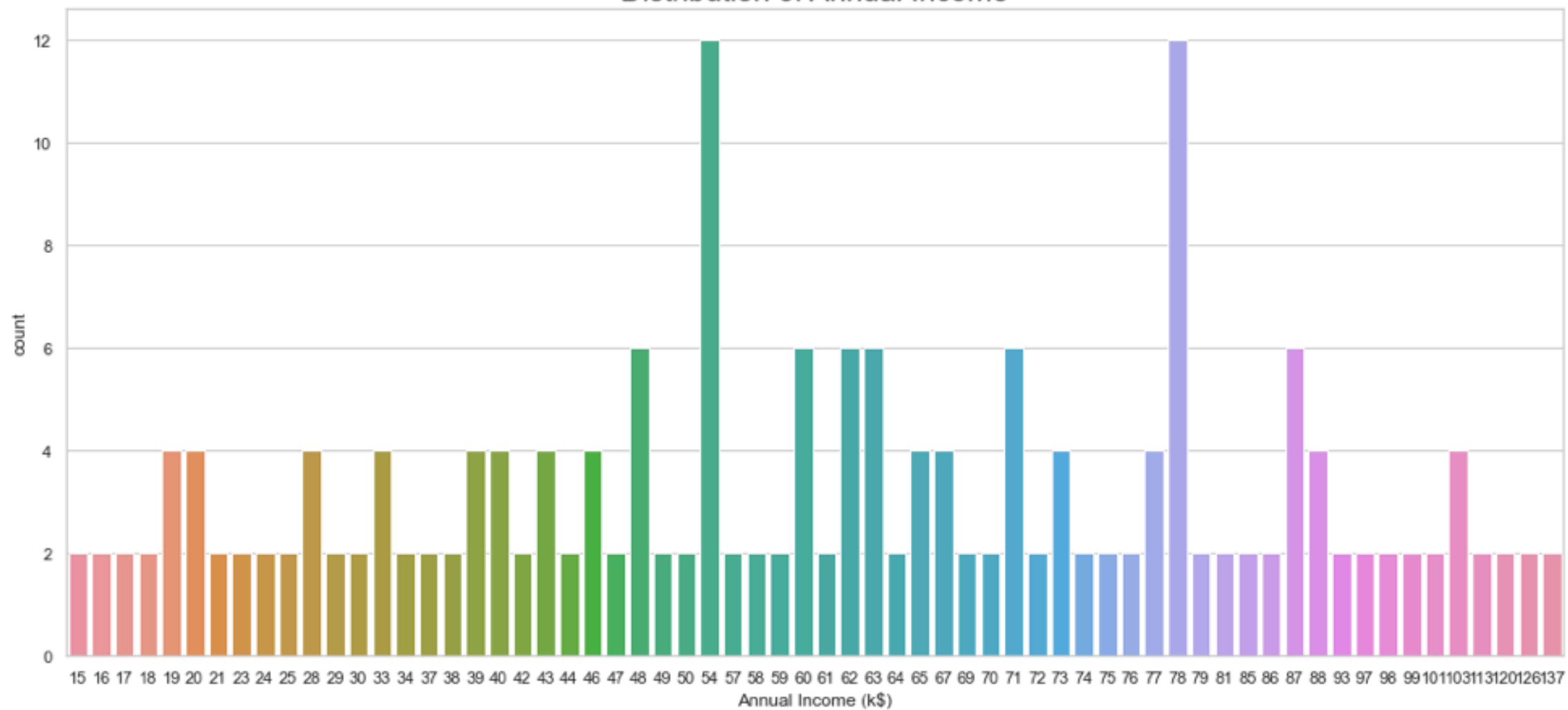
Distribution of Age

Distribution of Spending Score
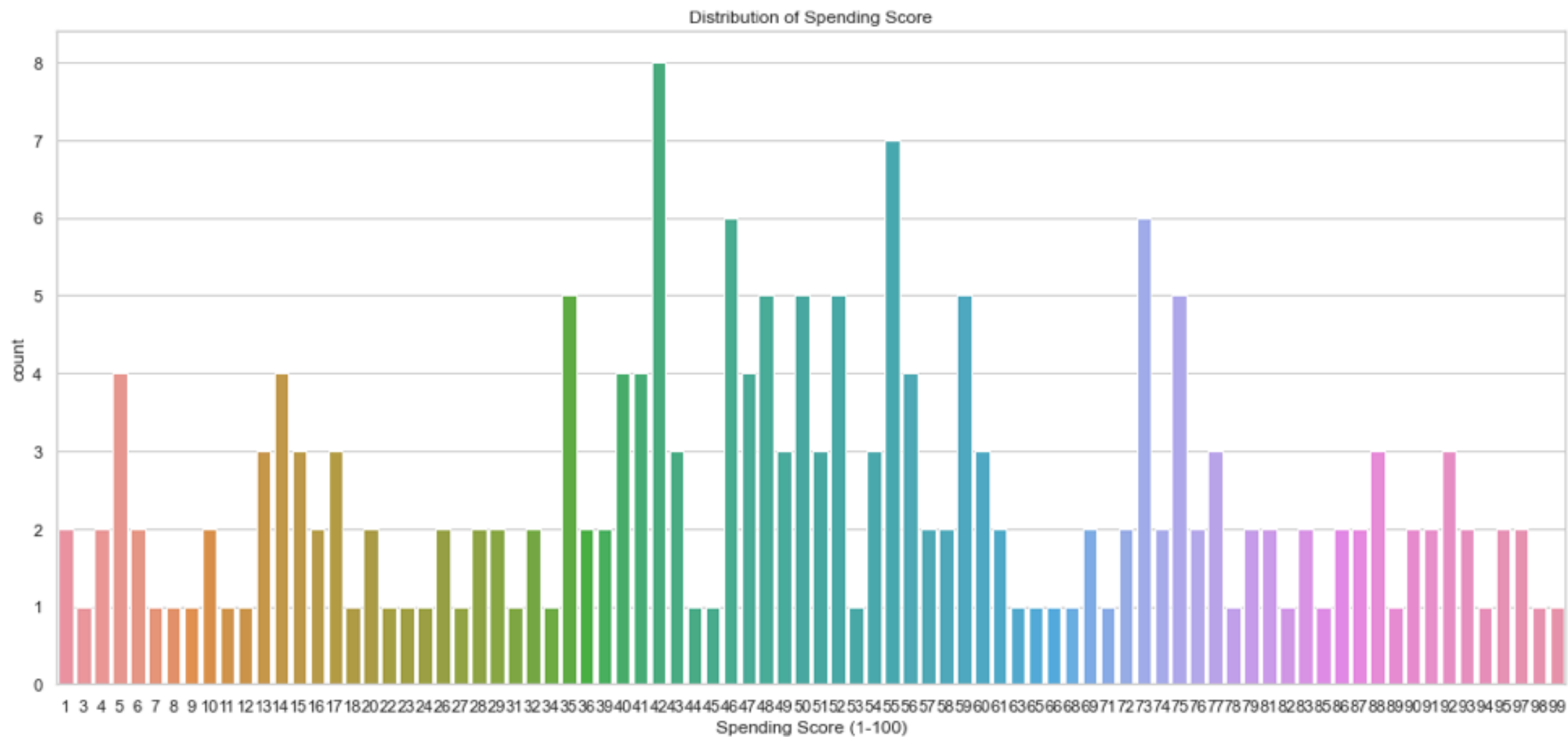
Corrélation entre variables

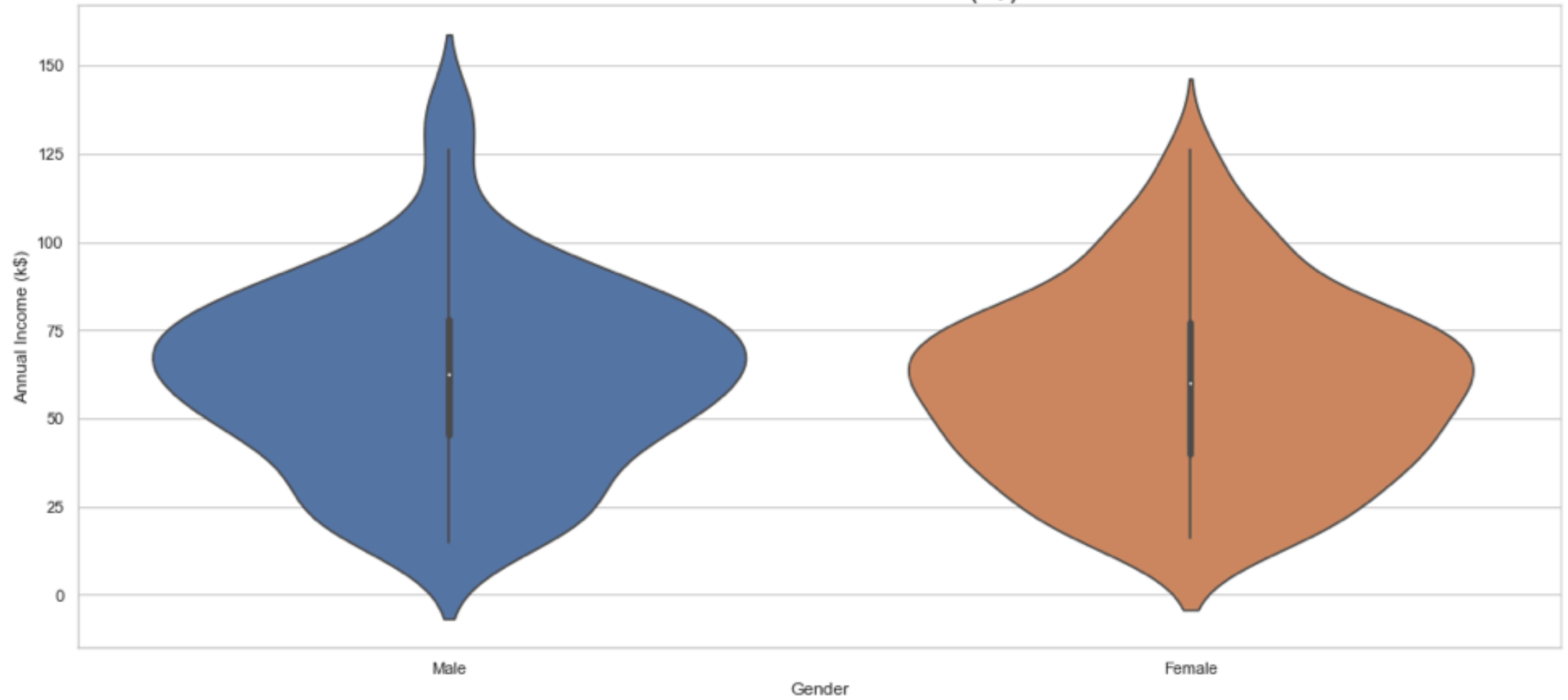Gender vs Spending Score

Gender vs Annual Income (k$)

Segmentation et classification de clients

Objectif :

- Repartir les clients dans différents groupes partageant des caractèristiques communes
  -> comparaison entre *Kmeans* et *Bisecting Kmeans* pour k variant de 2 à 10

- Attribuer un groupe trouvé par le clustering à un nouveau client
  -> comparaison entre *RandomForest* et *Decision Tree*

Clustering par *Kmeans* avec k=5 et silhouette = 0.62
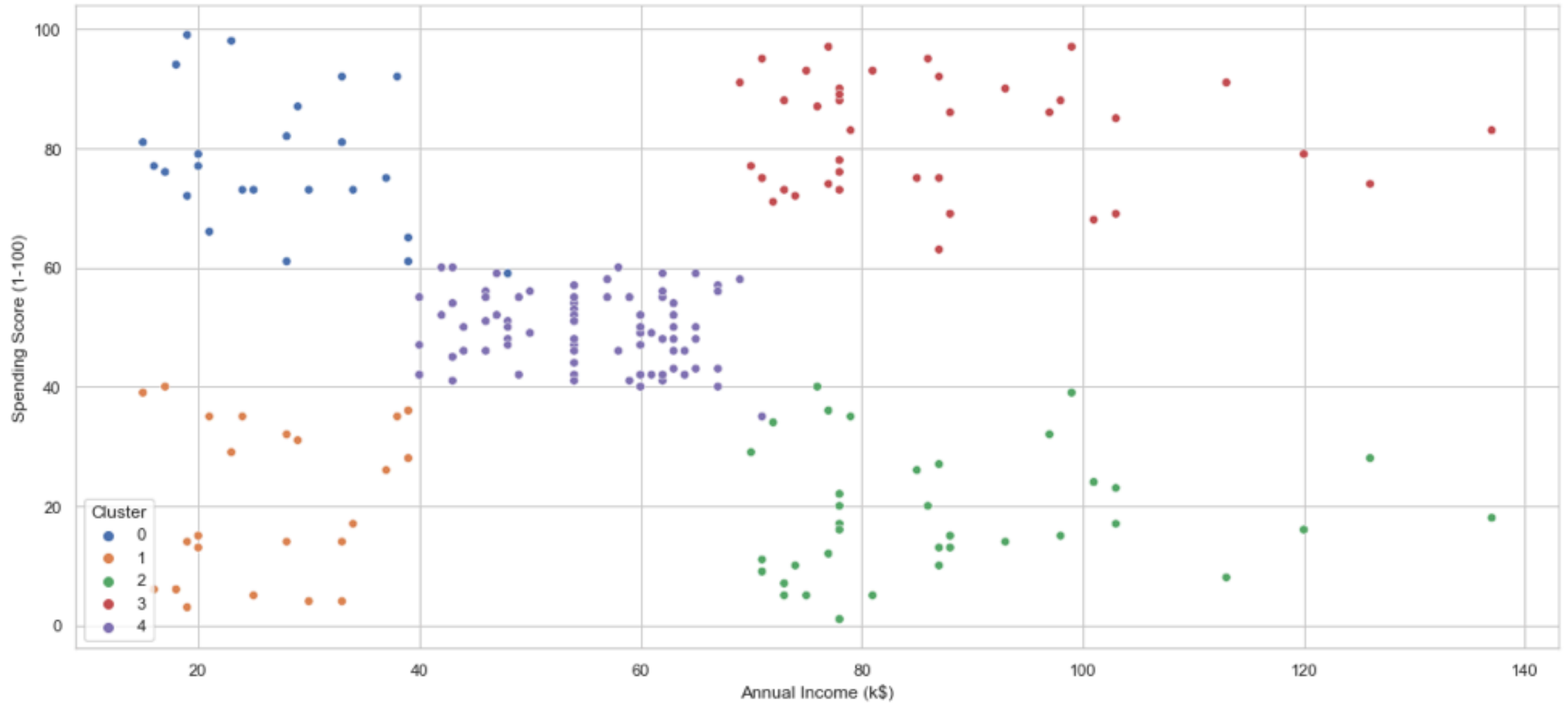
RandomForest sur les données test, accuracy = 1.0

```
+------------------+-------+--------------------+--------------------+----------+
|          features|Cluster|       rawPrediction|         probability|prediction|
+------------------+-------+--------------------+--------------------+----------+
|[0.0,18.0,65.0,48.0]|     4|[1.0,0.0,0.0,0.0,...|[0.1,0.0,0.0,0.0,...|       4.0|
|[0.0,21.0,62.0,42.0]|     4|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       4.0|
|[0.0,23.0,16.0,77.0]|     0|[10.0,0.0,0.0,0.0...|[1.0,0.0,0.0,0.0,...|       0.0|
|[0.0,23.0,18.0,94.0]|     0|[10.0,0.0,0.0,0.0...|[1.0,0.0,0.0,0.0,...|       0.0|
|[0.0,23.0,70.0,29.0]|     2|[0.0,0.0,10.0,0.0...|[0.0,0.0,1.0,0.0,...|       2.0|
|[0.0,25.0,72.0,34.0]|     2|[0.0,0.0,10.0,0.0...|[0.0,0.0,1.0,0.0,...|       2.0|
|[0.0,27.0,78.0,89.0]|     3|[0.75,0.0,0.0,9.2...|[0.075,0.0,0.0,0....|       3.0|
|[0.0,30.0,34.0,73.0]|     0|[7.52777777777777...|[0.75277777777777...|       0.0|
|[0.0,30.0,78.0,78.0]|     3|[0.75,0.0,0.0,9.2...|[0.075,0.0,0.0,0....|       3.0|
|[0.0,30.0,88.0,86.0]|     3|[0.75,0.0,0.0,9.2...|[0.075,0.0,0.0,0....|       3.0|
|[0.0,31.0,39.0,61.0]|     0|[4.77777777777777...|[0.47777777777777...|       4.0|
|[0.0,31.0,72.0,71.0]|     3|[0.0,0.0,0.0,10.0...|[0.0,0.0,0.0,1.0,...|       3.0|
|[0.0,32.0,60.0,42.0]|     4|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       4.0|
|[0.0,32.0,76.0,87.0]|     3|[0.0,0.0,0.0,10.0...|[0.0,0.0,0.0,1.0,...|       3.0|
|[0.0,34.0,78.0,22.0]|     2|[0.0,0.0,10.0,0.0...|[0.0,0.0,1.0,0.0,...|       2.0|
|[0.0,37.0,97.0,32.0]|     2|[0.0,0.0,10.0,0.0...|[0.0,0.0,1.0,0.0,...|       2.0|
|[0.0,38.0,78.0,76.0]|     3|[0.0,0.0,0.0,10.0...|[0.0,0.0,0.0,1.0,...|       3.0|
|[0.0,40.0,69.0,58.0]|     4|[0.0,0.0,0.0,2.0,...|[0.0,0.0,0.0,0.2,...|       4.0|
|[0.0,41.0,99.0,39.0]|     2|[0.0,0.0,6.0,1.0,...|[0.0,0.0,0.6,0.1,...|       2.0|
|[0.0,44.0,78.0,20.0]|     2|[0.0,0.0,10.0,0.0...|[0.0,0.0,1.0,0.0,...|       2.0|
+------------------+-------+--------------------+--------------------+----------+
only showing top 20 rows
```

Recommandation de films pour les clients du centre commercial

- Jointure des données d'avis et de description des films
- Recommandation basée sur l'historique d'avis du client et des autres utilisateurs
- Application d'un modèle de filtrage collaboratif : ALS

Recommandation de films pour l'utilisateur 3

```
+-------+--------------------------------------------------+----------------------------------------------------------------------------------------+
|movieId|recommendations                                   |title                                                                                   |
+-------+--------------------------------------------------+----------------------------------------------------------------------------------------+
|148    |[[83, 2.7878337], [169, 2.7877886], [168, 2.787404]] |The Secret Life of Words                                                               |
|471    |[[83, 3.4891572], [169, 3.4891005], [168, 3.4886198]]|Bandyta                                                                                |
|496    |[[83, 3.1753914], [169, 3.17534], [168, 3.1749024]]  |Borat: Cultural Learnings of America for Make Benefi
t Glorious Nation of Kazakhstan|
|833    |[[83, 2.6024714], [169, 2.6024292], [168, 2.6020706]]|Umberto D.                                                                             |
|1088   |[[83, 3.0998192], [169, 3.0997689], [168, 3.0993419]]|Whale Rider                                                                            |
|1580   |[[83, 3.4191318], [169, 3.4190764], [168, 3.4186056]]|Released                                                                               |
|1591   |[[83, 2.5174723], [169, 2.5174315], [168, 2.5170846]]|Nowhere in Africa                                                                      |
|1645   |[[83, 3.3626971], [169, 3.3626425], [168, 3.362179]] |A Time to Kill                                                                         |
|1829   |[[83, 2.9150498], [169, 2.9150026], [168, 2.9146008]]|The Story of Adele H                                                                   |
|1959   |[[83, 3.4865506], [169, 3.4864936], [168, 3.4860134]]|Swept from the Sea                                                                     |
|2122   |[[83, 2.5333629], [169, 2.5333216], [168, 2.5329726]]|The Whole Ten Yards                                                                    |
|2142   |[[83, 2.924748], [169, 2.9247005], [168, 2.9242976]] |Cop Land                                                                               |
|3175   |[[83, 3.4257376], [169, 3.425682], [168, 3.4252102]] |Barry Lyndon                                                                           |
|4519   |[[83, 3.1578095], [169, 3.1577582], [168, 3.1573234]]|Conversations with My Gardener                                                         |
|4935   |[[83, 3.2363183], [169, 3.236266], [168, 3.23582]]   |Howl's Moving Castle                                                                   |
|5156   |[[83, 2.7750869], [169, 2.7750423], [168, 2.7746596]]|Bicycle Thieves                                                                        |
|6466   |[[83, 3.4009705], [169, 3.4009151], [168, 3.4004467]]|Freddy vs. Jason                                                                       |
|6620   |[[83, 3.637123], [169, 3.637064], [168, 3.6365633]]  |[{'iso_3166_1': 'US', 'name': 'United States of Amer
ica'}]|
|7340   |[[83, 2.8415718], [169, 2.8415256], [168, 2.841134]] |If you've got a taste for terror... take Carrie to t
he prom.|
|8592   |[[83, 3.1870253], [169, 3.1869733], [168, 3.186535]] |Dick Tracy                                                                             |
+-------+--------------------------------------------------+----------------------------------------------------------------------------------------+
```

Conclusion