GE2324 Assignment1

LIU Hengche 57854329

Q1:

(a):

- Iteration 1 - clustering:
  - Cluster 0 color indexes: [5, 14]
  - Cluster 1 color indexes: [1, 3, 4, 7, 8, 9, 11, 12, 15, 16]
  - Cluster 2 color indexes: [2, 6, 10, 13]
- Iteration 1 - new centroids:
  - Cluster 0 center: (176.0, 176.5, 157.0)
  - Cluster 1 center: (67.0, 96.5, 101.7)
  - Cluster 2 center: (203.2, 214.8, 219.8)
- Iteration 2 - clustering:
  - Cluster 0 color indexes: [5, 8, 9, 12, 14]
  - Cluster 1 color indexes: [1, 3, 4, 7, 11, 15, 16]
  - Cluster 2 color indexes: [2, 6, 10, 13]
- Iteration 2 - new centroids:
  - Cluster 0 center: (147.2, 162.4, 158.6)
  - Cluster 1 center: (40.9, 72.3, 76.9)
  - Cluster 2 center: (203.2, 214.8, 219.8)
- Iteration 3 - clustering:
  - Cluster 0 color indexes: [5, 8, 9, 12, 14]
  - Cluster 1 color indexes: [1, 3, 4, 7, 11, 15, 16]
  - Cluster 2 color indexes: [2, 6, 10, 13]
- **Final clustering result**:
  - Cluster 0 ((147.2, 162.4, 158.6)): [5, 8, 9, 12, 14]
  - Cluster 1 ((40.9, 72.3, 76.9)): [1, 3, 4, 7, 11, 15, 16]
  - Cluster 2 ((203.2, 214.8, 219.8)): [2, 6, 10, 13]

(b):

- Iteration 1 - clustering:
  - Cluster 0 color indexes: [1, 4, 7, 11]
  - Cluster 1 color indexes: [2, 5, 6, 8, 10, 13, 14]
  - Cluster 2 color indexes: [3, 9, 12, 15, 16]
- Iteration 1 - new centroids:
  - Cluster 0 centroids: (24.8, 49.5, 46.0)
  - Cluster 1 centroids: (186.1, 197.4, 197.9)

    - Cluster 2 centroids: (86.6, 119.4, 128.2)
- Iteration 2 - clustering:
    - Cluster 0 color indexes: [1, 4, 7, 11]
    - Cluster 1 color indexes: [2, 5, 6, 8, 10, 13, 14]
    - Cluster 2 color indexes: [3, 9, 12, 15, 16]
- **Final clustering result**:
    - Cluster 0 ((24.8, 49.5, 46.0)): [1, 4, 7, 11]
    - Cluster 1 ((186.1, 197.4, 197.9)): [2, 5, 6, 8, 10, 13, 14]
    - Cluster 2 ((86.6, 119.4, 128.2)): [3, 9, 12, 15, 16]

No. The results are not the same. This is because K-means clustering converges to **local** optima, meaning the final clusters depend heavily on where the centroids start. Different initializations lead to different cluster boundaries.

Q2:

(a)

- Step 1: Means

```
X̄ = 50.00, Ȳ = 2.00
```

- Step 2: Individual Calculations

```
Index | X_i   | Y_i   | (X_i-X̄)   | (Y_i-Ȳ)    | Product
-----------------------------------------------------------------
1     | 50    | 1     |      0.00 |     -1.00 |     -0.00
2     | 30    | 2     |    -20.00 |      0.00 |     -0.00
3     | 80    | 3     |     30.00 |      1.00 |     30.00
4     | 20    | 1     |    -30.00 |     -1.00 |     30.00
5     | 60    | 2     |     10.00 |      0.00 |      0.00
6     | 40    | 1     |    -10.00 |     -1.00 |     10.00
7     | 70    | 3     |     20.00 |      1.00 |     20.00
8     | 10    | 1     |    -40.00 |     -1.00 |     40.00
9     | 90    | 4     |     40.00 |      2.00 |     80.00
```

- Step 3: Sum of Products

```
Covariance term: Σ(X_i-X̄)(Y_i-Ȳ) = 210.00
```

- Step 4: Standard Deviations

```
√(Σ(X_i-X̄)²) = √6000.00 = 77.46
√(Σ(Y_i-Ȳ)²) = √10.00 = 3.16
Denominator = 244.95
```

- Step 5: Final Calculation

Pearson r = 0.857

- Conclusion:

```
There is a strong positive correlation (r = 0.857) between
Distance Driven and Number of Previous Owners. This means:
• As distance driven increases, number of previous owners tends to INCREASE
• Vehicles with higher mileage generally have had MORE previous owners
```

(b)

```
============================================================
Analysis of Car Age vs Price
============================================================

Original Data:
        Car Age |           Price
-----------------------------------
              5 |              20
              3 |              25
              8 |              15
              2 |              30
              6 |              18
              4 |              22
              7 |              16
              1 |              35
              9 |              12

Ranks with Tie Handling:
   Car Age Rank |      Price Rank |  Difference (d) |              d²
-------------------------------------------------------------------------------
           5.00 |            5.00 |            0.00 |            0.00
           3.00 |            7.00 |           -4.00 |           16.00
           8.00 |            2.00 |            6.00 |           36.00
           2.00 |            8.00 |           -6.00 |           36.00
           6.00 |            4.00 |            2.00 |            4.00
           4.00 |            6.00 |           -2.00 |            4.00
           7.00 |            3.00 |            4.00 |           16.00
           1.00 |            9.00 |           -8.00 |           64.00
           9.00 |            1.00 |            8.00 |           64.00

Intermediate Calculations:
Sum of d² (Σd²) = 240.00
n(n² - 1) = 9*(9²-1) = 720
Spearman's ρ = 1 - (6*240.00)/720 = -1.000
```

```
============================================================
Analysis of Previous Owners vs Price
============================================================

Original Data:
Previous Owners |           Price
-----------------------------------
              1 |              20
              2 |              25
              3 |              15
              1 |              30
              2 |              18
```

```
               1 |              22
               3 |              16
               1 |              35
               4 |              12

Ranks with Tie Handling:
Previous Owners Rank |      Price Rank |  Difference (d) |              d²
------------------------------------------------------------------------------
           2.50 |            5.00 |           -2.50 |            6.25
           5.50 |            7.00 |           -1.50 |            2.25
           7.50 |            2.00 |            5.50 |           30.25
           2.50 |            8.00 |           -5.50 |           30.25
           5.50 |            4.00 |            1.50 |            2.25
           2.50 |            6.00 |           -3.50 |           12.25
           7.50 |            3.00 |            4.50 |           20.25
           2.50 |            9.00 |           -6.50 |           42.25
           9.00 |            1.00 |            8.00 |           64.00

Intermediate Calculations:
Sum of d² (Σd²) = 210.00
n(n² - 1) = 9*(9²-1) = 720
Spearman's ρ = 1 - (6*210.00)/720 = -0.750
```

FINAL CONCLUSION:

```
Car Age vs Price Correlation:       ρ = -1.000
Previous Owners vs Price Correlation: ρ = -0.750
```

Car Age is the stronger predictor of Price. There is a perfect negative correlation (ρ = -1.0), meaning car price decreases consistently as vehicles get older.

(c)

```
============================================================
PROBLEM 2: Kendall's Tau Calculation
============================================================

========================================
Calculating Kendall's Tau for Distance Driven vs Price
========================================

Pair-wise Analysis:
Pair (1,2): D
Pair (1,3): D
Pair (1,4): D
Pair (1,5): D
Pair (1,6): D
... Total 36 pairs analyzed

Count Summary:
Concordant pairs (C): 0
Discordant pairs (D): 36
Total pairs: 36.0
Kendall's Tau = (C-D)/Total = (0-36)/36.0 = -1.000
```

Final Kendall's Tau: **τ = -1.000**