Project 2: Big Data Management

"Advanced Hadoop"

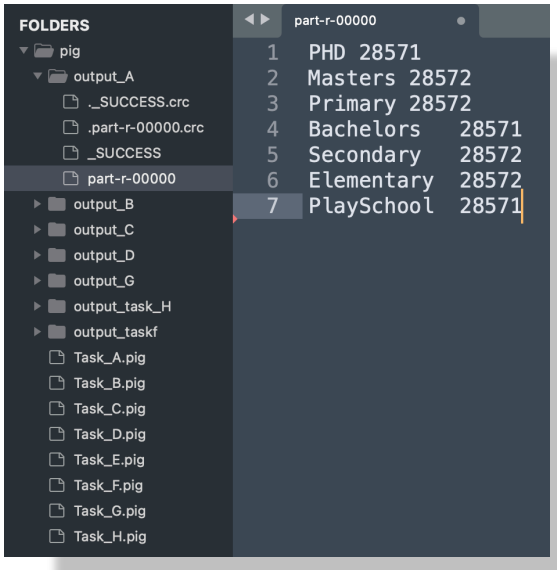
Ehu Shubham Shaw
Supreeth
Roee shraga
October 10, 2024

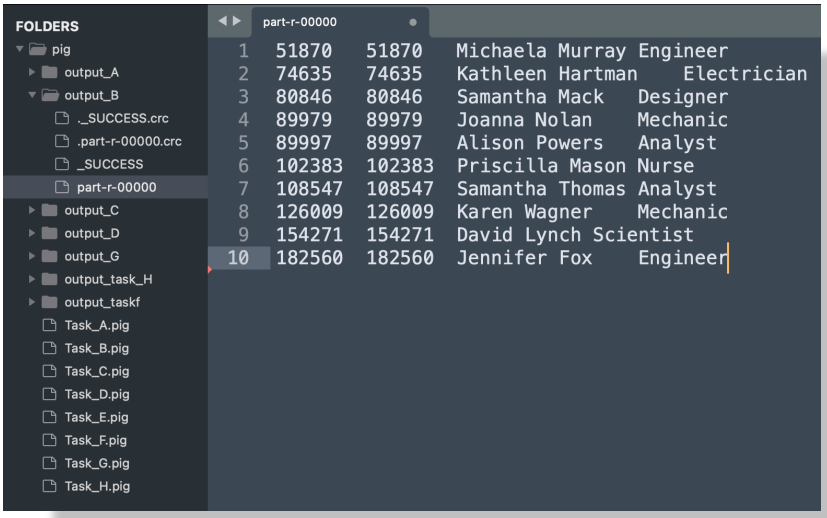
1- Revisiting LinkBook Network with Pig [16 Points]

The way I chose to implement the task is by using Pig latin. For Task A Pig Requires fewer lines of code and is easier to maintain compared to the custom Mapper and Reducer classes in Java. Wheres in MapReduce Offers more control over the logic and performance but requires more boilerplate code. Task B pig is easier for aggregations like top N. The TOP function simplifies this task, making it more intuitive. in MapReduce You need a custom reducer to track and sort page accesses, which adds complexity. in Task C Pig Uses Simple filtering operation with a straightforward query.
Where as MapReduce Requires writing a custom Mapper to filter the data, which is less intuitive than Pig's syntax. In Task D Pig Simplifies joins and counting. Pig's high-level syntax saves time and effort.where MapReduce Requires custom Mapper and Reducer for counting and handling relationships, which increases complexity. Task E pig supports counting and distinct operations in a straightforward way. On the other hand MapReduce Requires more work to implement distinct counting, needing additional logic.Task F Pig uses calculation of averages is easier and more abstracted. And in MapReduce Requires a custom reducer to calculate the average and compare each user's relationship count. Task G Pig Date comparison and filtering are easier.compared to MapReduce Requires custom logic to handle dates and filtering. Task H Joins and filtering based on relationships and accesses are handled easily in pig. And for in MapReduce Requires complex logic to join and filter the data.

Pig Implementation: To calculate the frequency of each education level for Task a, utilise GROUP and COUNT on the HighestEdu field. Utilize GROUP, COUNT, and TOP to determine the ten most frequently accessed pages for Task B. Use FILTER in Task c to pick persons whose HighestEdu matches yours. For Task d, count relationships and calculate the pleasure factor by using JOIN on LinkBookPage and Associates. To determine the total and unique accesses in Task e, utilise GROUP and COUNT on the AccessLog. Determine the average number of relationships for Task f and eliminate people with more relationships than the norm. To find people who haven't accessed LinkBook in ninety days, use FILTER on the AccessLog in Task g.Lastly, for Task h, find individuals who have relationships but have never visited their friends' LinkBook sites by using JOIN on Associates and AccessLog.

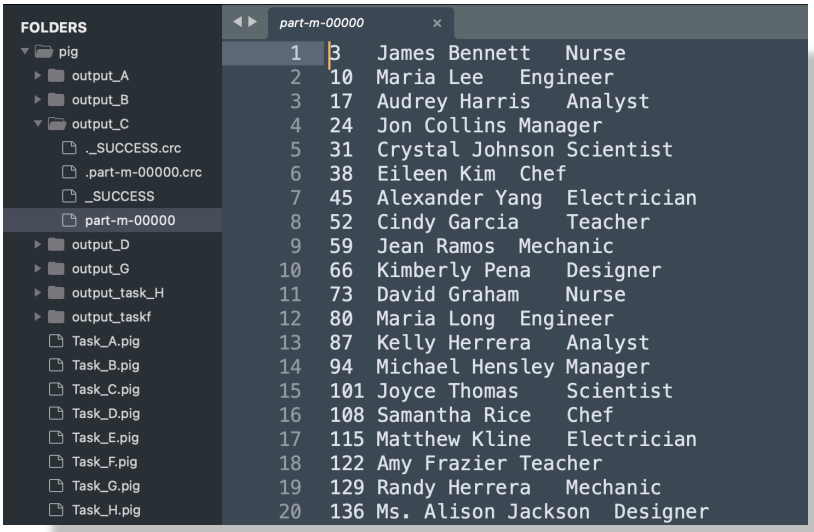
TASK A

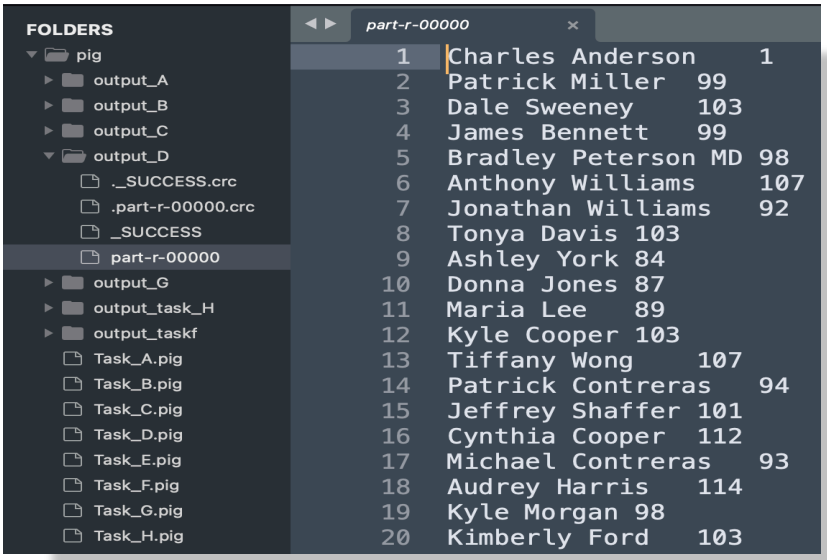




Task B

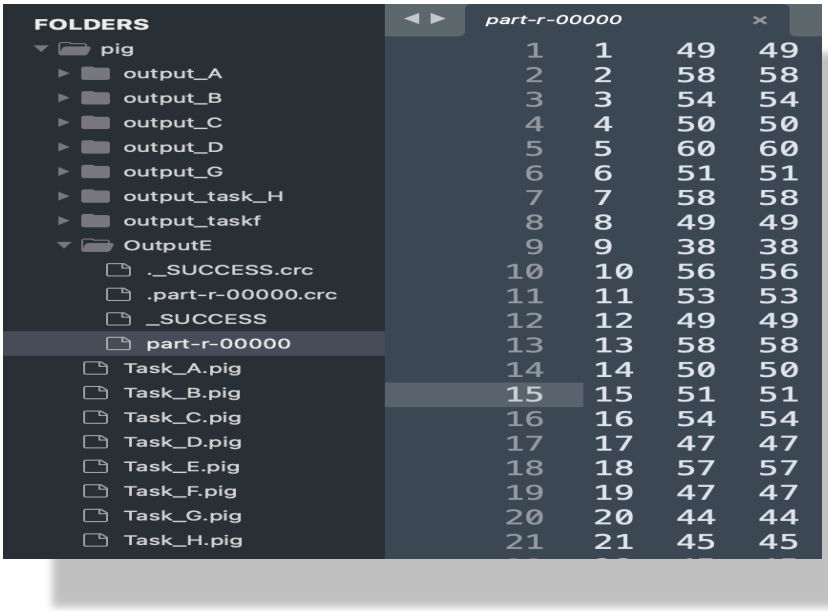
Task C



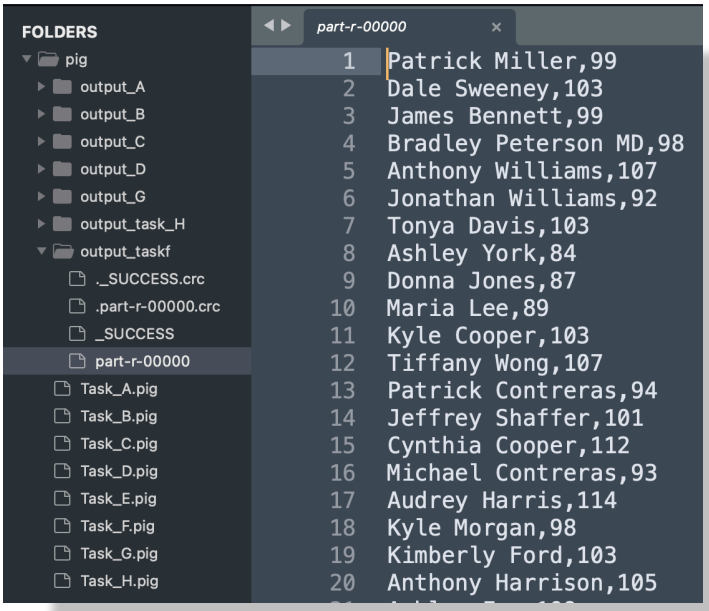


Task D

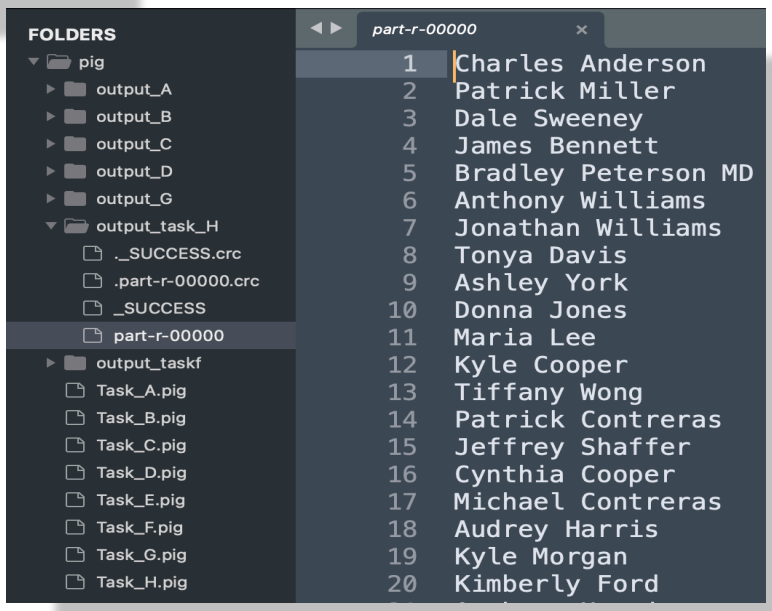
Task E

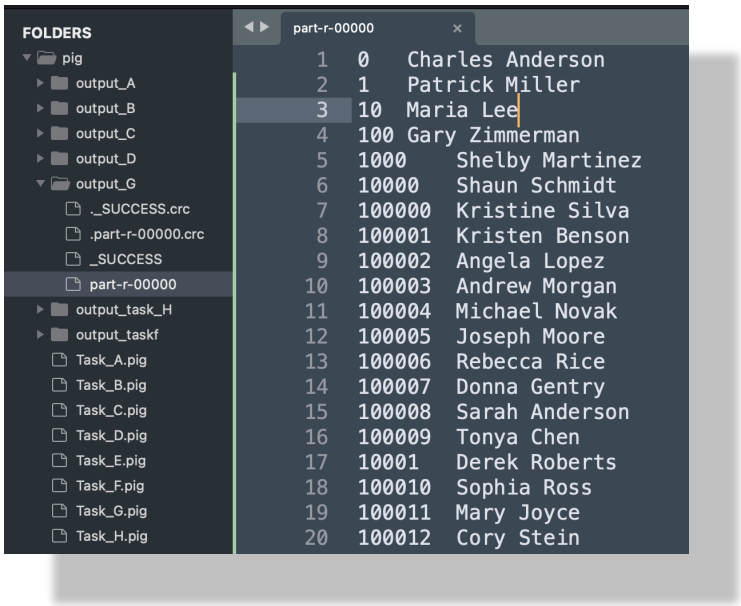


Task F



Task G





Task H

2- Clustering [44 Points]

For Task 1 (Single-Iteration K-Means, R=1), the solution runs only one iteration of the K-means algorithm, assigning each data point to the nearest centroid and updating the centroids once. I tested with different values of K (the number of clusters) to see how the clustering changed depending on the initial centroid pick. Because this task only requires one repetition, it is computationally fast. However, the quality of the clusters is strongly influenced by the centroids' initial placements.

In Task 2 (Multi-Iteration K-Means with K=10), the centroids are modified after each iteration using the points provided to them. Running studies with different K values revealed how the centroid positions change. However, if convergence comes quickly, the method may perform unneeded computations, rendering subsequent iterations redundant. Performance improves with decreasing K, but plateaus after a few repetitions.

In Task 3 (Multi-Iteration K-Means with Early Termination), the method employs early stopping when centroids converge below a certain threshold (epsilon = 0.001) before reaching the maximum number of iterations. I experimented with different K and R numbers to see how often the algorithm converged before reaching the maximum number of iterations. In one experiment, it converged at 36 iterations, significantly reducing computation time. This strategy improves performance when the dataset allows for rapid convergence.

For Task 4 (K-Means using Hadoop MapReduce Combiner), I employed a combiner to improve data shuffling and aggregation throughout the MapReduce process. The combiner conducts local aggregation at the mapper level before transferring data to the reducer, which reduces computation time. For example, when K=100 was used, the combiner reduced runtime by around one minute compared to running without it.

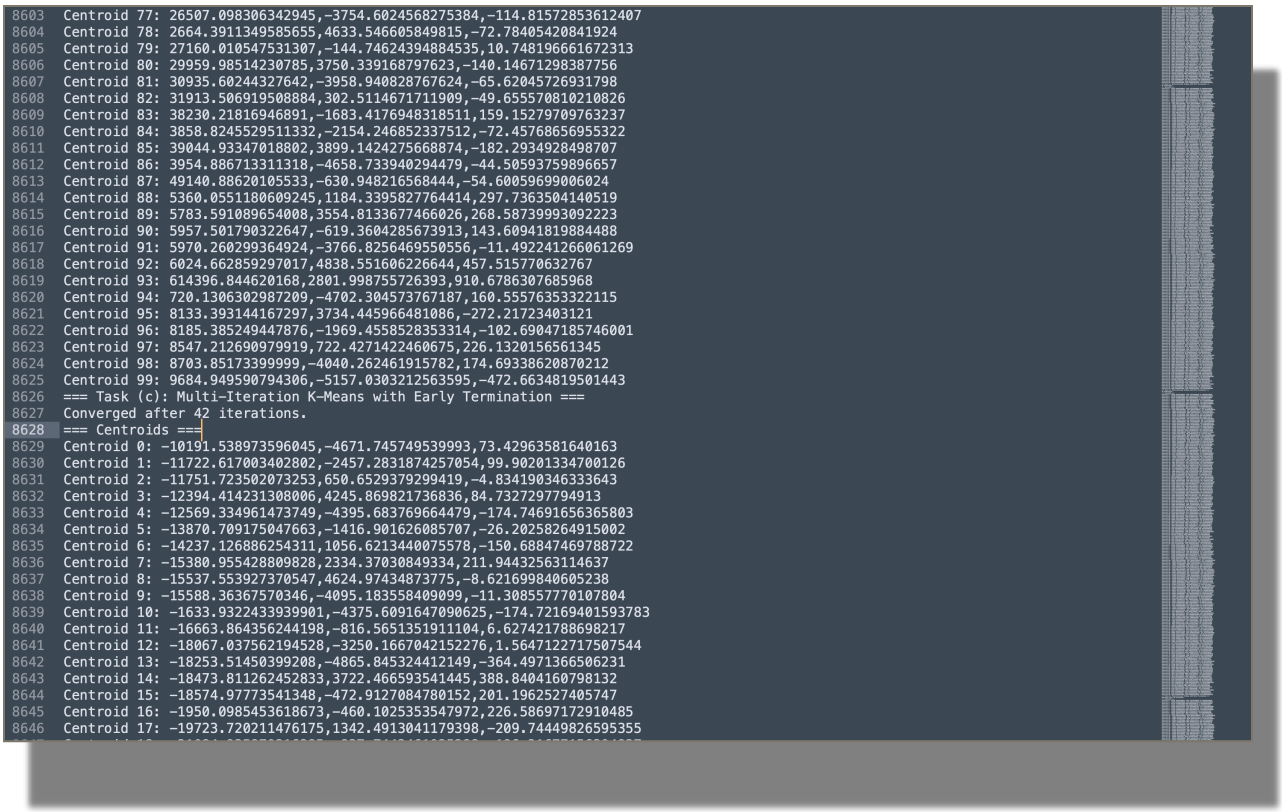
Task 5 (Clustered Points Evaluation) displayed the final centroid coordinates and the number of points allocated to each cluster. I tested the technique with K values of 5, 10, and 100 and noticed that the points allocated to clusters changed with each iteration. The distance between centroid locations rarely remained constant between iterations. As K grows, clusters get more granular and exact, but calculation time increases as well. More iterations (higher R) improve accuracy while increasing computing time, whereas early termination balances accuracy and efficiency by ending when the clusters stabilize.

Kaggle dataset **Python dataset**

```
=== Task (a): Single-Iteration K-Means (R=1) ===
=== Centroids ===
Centroid 0: -17740.987914720223,-32.16045677347778,10.576229271312487
Centroid 1: -2514.9741295953554,-151.52640521084564,-1.2005583594879 4
Centroid 2: -32860.765504206676,183.78441897614778,14.955900103370988
Centroid 3: 14623.378944979117,41.813621191371 29,1.7845258742 45664
Centroid 4: 34674.33962300714,69.99590980174459,-31.619689426619725
=== Task (b): Basic Multi-Iteration K-Means (R=10) ===
=== Centroids ===
Centroid 0: -21733.334005332043,-457.4049553091571,8.675713033463772
Centroid 1: -34760.48071888516,1003.3590712997714,17.120329849990842
Centroid 2: -8625.728433781891,713.4684178839218,-17.896191675183015
Centroid 3: 6722.9562649859,-590.9771873189216,6.054176348511625
Centroid 4: 28273.975990499515,8.893537741046963,-4.033804905436877
Centroid 0: -21137.929245250176,-107.39727011217039,4.075352195086014 5
Centroid 1: -34510.76108701358,298.1762445529446,32.68398574371951
Centroid 2: -8346.294307810853,18.454136694523257,-13.050949720941 2
Centroid 3: 29157.802434707457,37.81124654669398,-6.626105685498191
Centroid 4: 8091.183309784175,-83.78213383470552,1.5006189853865126
Centroid 0: -21137.929245250176,-107.39727011217039,4.075352195086024 5
Centroid 1: -34510.76108701358,298.1762445529446,32.68398574371951
Centroid 2: -8346.294307810853,18.454136694523257,-13.050949720941 2
Centroid 3: 29157.802434707457,37.81124654669398,-6.626105685498191
Centroid 4: 8091.183309784175,-83.78213383470552,1.5006189853865126
Centroid 0: -20648.623978892316,82.19865506172474,19.268672355664664
Centroid 1: -34262.79349647997,229.4702918721198,23.679907437262663
Centroid 2: -7661.999934065268,-172.50730567804996,-25.908397254461157
Centroid 3: 30188.474270009257,-35.288065860841584,-19.160693929027488
Centroid 4: 9094.852265659583,21.859141393981 41,12.971339977084417
Centroid 0: -20648.623978892316,82.19865506172474,19.268672355664664
Centroid 1: -34262.79349647997,229.4702918721198,23.679907437262663
Centroid 2: -7661.999934065268,-172.50730567804996,-25.908397254461157
Centroid 3: 30188.474270009257,-35.288065860841584,-19.160693929027498
Centroid 4: 9094.852265659583,21.859141393981 41,12.971339977084417
Centroid 0: -20220.396660990853,67.75878988442898,19.885438179788222
Centroid 1: -34012.42214299034,256.6156545878215,21.826060549754 15
Centroid 2: -6697.544985100782,-203.96799855495343,-31.984102753753078
Centroid 3: 31069.469338595O7,-8.763385998152163,-21.530402816243875
Centroid 4: 10328.335627574186,30.016463056782612,19.588432701533883
Centroid 0: -20220.396660990853,67.75878988442898,19.885438179788222
Centroid 1: -34012.42214299034,256.6156545878215,21.826060549754 15
Centroid 2: -6697.544985100782,-203.96799855495343,-31.984102753753078
Centroid 3: 31069.469338595O7,-8.763385998152163,-21.530402816243875
Centroid 4: 10328.335627574186,30.016463056782612,19.588432701533883
Centroid 0: -19613.2908907426,15.497313526234 31,0.16319498653062223
Centroid 1: -33846.736812932,335.44981137771794,26.408872339082855
Centroid 2: -5754.799602435291,-234.75787718867002,-10.027125511598850
Centroid 3: 11283.466633053948,67.64878711725233,16.910133000036534
Centroid 4: 32052.169148597644,-33.19399539766349,-30.368118428700 47
Centroid 0: -19613.2908907426,15.497313526234 31,0.16319498653062223
Centroid 1: -33846.736812932,335.44981137771794,26.408872339082855
Centroid 2: -5754.799602435291,-234.75787718867002,-10.027125511598857
Centroid 3: 11283.466633053948,67.64878711725233,16.910133000036534
Centroid 4: 32052.169148597644,-33.19399539766349,-30.368118428700 47
Centroid 0: -19188.395197301885,70.06321917521164,-4.602109066171814
Centroid 1: -33574.84778501426,234.491985289192,36.75734182417191
Centroid 2: -4870.102610436248,-187.11807395595153,-6.466754229921 65
Centroid 3: 12197.571695558243,29.931547636548892,17.884095568742943
Centroid 4: 32591.516200706625,-30.219928777554 19,-38.654455208787624
=== Task (c): Multi-Iteration K-Means with Early Termination ===
No Convergence.

=== Task (d): K-Means with Hadoop MapReduce with Combiner ===
=== Centroids ===
Centroid 0: -0.0034813649679084646,-0.14797973332989647,-0.098758507311974 55
Centroid 1: -33574.84778501426,234.491985289192,36.75734182417191
Centroid 2: -6697.544985100782,-203.96799855495343,-31.984102753753078
Centroid 3: 31069.469338595O7,-8.763385998152163,-21.530402816243875
Centroid 4: 8091.183309784175,-83.78213383470552,1.5006189853865126

=== Task (e-1): Final Output of cluster centers ===
=== Centroids ===
Centroid 0: -17740.987914720223,-32.16045677347778,10.576229271312487
Centroid 1: -2514.9741295953554,-151.52640521084564,-1.2005583594879 4
Centroid 2: -32860.765504206676,183.78441897614778,14.955900103370988
Centroid 3: 14623.378944979117,41.813621191371 29,1.7845258742 45664
Centroid 4: 34674.33962300714,69.99590980174459,-31.619689426619725
=== Task (e-2): Final Output of Clustered Points ===
Cluster Center: -17740.987914720223,-32.16045677347778,10.576229271312487 | Number of Points: 505 points
Cluster Center: -2514.9741295953554,-151.52640521084564,-1.2005583594879 4 | Number of Points: 544 points
Cluster Center: -32860.765504206676,183.78441897614778,14.955900103370988 | Number of Points: 282 points
Cluster Center: 14623.378944979117,41.813621191371 29,1.7845258742 45664 | Number of Points: 593 points
Cluster Center: 34674.33962300714,69.99590980174459,-31.619689426619725 | Number of Points: 315 points
```

```
=== Task (a): Single-Iteration K-Means (R=1) ===
=== Centroids ===
Centroid 0: 0.1917807828718958,4.026208655621477,-1.1535601067337677
Centroid 1: -1.1978864576413168,2.321546029684262,-0.14163034204442798
Centroid 2: -0.3787205051755565 5.334917978894541,0.3705515551916600 3
Centroid 3: 0.3892605340088976,2.455966824837509,0.7322969123511548
Centroid 4: 0.2569019707368709 3,0.3498508340120633 5,-0.4496599658265828
=== Task (b): Basic Multi-Iteration K-Means (R=10) ===
=== Centroids ===
Centroid 0: 0.1917807828718958,4.026208655621477,-1.1535601067337677
Centroid 1: -1.1978864576413168,2.321546029684262,-0.14163034204442798
Centroid 2: -0.3787205051755565 5.334917978894541,0.3705515551916600 3
Centroid 3: 0.3892605340088976,2.455966824837509,0.7322969123511548
Centroid 4: 0.2569019707368709 3,0.3498508340120633 5,-0.4496599658265828
Centroid 0: -0.2179378468731465 5,5.354131208465669,0.3452803646105714 4
Centroid 1: -1.1160437708096491,2.267209283869940 7,-0.1031082644405637 8
Centroid 2: 0.3065572459482709,3.832033159226905,-1.021504841129958 5
Centroid 3: 0.1468802580140346 2,0.1082854846170969 3,-0.1942984822958039
Centroid 4: 0.5162813023019023,2.486382152235047,0.6965635012330967
Centroid 0: -0.1732374661328455,5.377323254311311,0.3078241692394607
Centroid 1: -1.0527062762840398,2.258425583199141 8,-0.0590602536119013 9
Centroid 2: 0.0930202958091610 6,-0.0128551545527128 55,-0.1118511323168098 2
Centroid 3: 0.3205890639409507,3.674192991295235,-0.9945289391573808
Centroid 4: 0.6053942859916 13,2.458505507256412,0.6622065693242569
Centroid 0: -0.1628350042158237,5.361249252157499,0.2784734276547002 3
Centroid 1: -1.0268488062990832,2.259942035584955 6,-0.0165871882240493 93
Centroid 2: 0.0581620437956818 8,-0.0738056103835595,-0.0902910013162745 7
Centroid 3: 0.3143194442775185,3.565013990979596,-0.9814696913707054
Centroid 4: 0.6719870531712189,2.401652303325526,0.6578128173688119
Centroid 0: -0.1543375678650555 7,5.337383016338755,0.2523275094382583
Centroid 1: -1.0322933172390365,2.260755843093374 3,0.0489642859433884 1
Centroid 2: 0.0363531661766071 86,-0.0958565802756141 8,-0.0908211774174316
Centroid 3: 0.3070212706708343 5,3.464485416707879 6,-0.9794565351120299
Centroid 4: 0.7118324775505788,2.355123545235398,0.6669203287234202
Centroid 0: -0.1402240801559703 8,5.309623831059216,0.2346136352240503 5
Centroid 1: -1.0410235927962346,2.257626837579279 4,0.1150102414356743 13
Centroid 2: 0.0254128693382529 4574,-0.1118156653979480 9,-0.0868787513985001 14
Centroid 3: 0.2918907068030435,3.394767666150349,-0.9807700055882800
Centroid 4: 0.7417266912299031,2.318410265630850 3,0.6570337102674553
Centroid 0: -0.1195180374729809,5.288265069948517,0.2320650844220974 5
Centroid 1: -1.0435082984453548,2.261221330219018 7,0.1722277544004802 7
Centroid 2: 0.0160750073467258 14,-0.1298449481288908 3,-0.1002904927435612 1
Centroid 3: 0.2684415882422973 6,3.335666990747836 6,-0.9914535950043666
Centroid 4: 0.7595697988401049,2.285725757910807 2,0.6438650725628575
Centroid 0: -0.1010738654617437 8,5.271998901241229,0.2193357041191327 5
Centroid 1: -1.0525832661127565,2.272794293255506 3,0.2217468709397354
Centroid 2: 0.0086721162900201 56,-0.1374833429066947,-0.1014263180819548 6
Centroid 3: 0.2415913743407563 6,3.289460782781688 5,-1.0057480872370015
Centroid 4: 0.7795067433116066,2.269646507673990 2,0.6285509660713312
Centroid 0: -0.0861705790481337 9,5.255393701430878,0.2120570485800859 8
Centroid 1: -1.0552357235464869,2.288562676569293,0.2451918903103513
Centroid 2: -0.0034813649679084646,-0.14797973332989647,-0.098758507311974 55
Centroid 3: 0.2203612824268546 8,3.234855339163051,-1.0187682810835472
Centroid 4: 0.7998479654575688,2.255428837605369 3,0.6128930793399755
Centroid 0: -0.0129982786214945 8,-0.1561513027072947,-0.0961741603981018 1
Centroid 1: -0.0833438501086358 3,5.248264849468372 5,0.2063992527705031 2
Centroid 2: -1.0484986463125 95,2.301785046519184 3,0.2656865609739627
Centroid 3: 0.2162255133769112 3,3.200318267289530 3,-1.0278154882723047
Centroid 4: 0.8118666889354365,2.245556182535447 8,0.6157323317967864
=== Task (c): Multi-Iteration K-Means with Early Termination ===
No Convergence.
=== Task (d): Advanced K-Means with Hadoop MapReduce with Combiner ===
=== Centroids ===
Centroid 0: -0.0129982786214945 8,-0.1561513027072947,-0.0961741603981018 1
Centroid 1: -0.0833438501086358 3,5.248264849468372 5,0.2063992527705031 2
Centroid 2: -1.0484986463125 95,2.301785046519184 3,0.2656865609739627
Centroid 3: 0.2162255133769112 3,3.200318267289530 3,-1.0278154882723047
Centroid 4: 0.8118666889354365,2.245556182535447 8,0.6157323317967864

=== Task (e-1): Final Output of cluster centers ===
=== Centroids ===
Centroid 0: -0.0129982786214945 8,-0.1561513027072947,-0.0961741603981018 1
Centroid 1: -0.0833438501086358 3,5.248264849468372 5,0.2063992527705031 2
Centroid 2: -1.0484986463125 95,2.301785046519184 3,0.2656865609739627
Centroid 3: 0.2162255133769112 3,3.200318267289530 3,-1.0278154882723047
Centroid 4: 0.8118666889354365,2.245556182535447 8,0.6157323317967864
=== Task (e-2): Final Output of cluster centers with Points ===
Cluster Center: -0.0129982786214945 8,-0.1561513027072947,-0.0961741603981018 1 | Number of Points: 1121 points
Cluster Center: -0.0833438501086358 3,5.248264849468372 5,0.2063992527705031 2 | Number of Points: 984 points
Cluster Center: -1.0484986463125 95,2.301785046519184 3,0.2656865609739627 | Number of Points: 942 points
Cluster Center: 0.2162255133769112 3,3.200318267289530 3,-1.0278154882723047 | Number of Points: 920 points
Cluster Center: 0.8118666889354365,2.245556182535447 8,0.6157323317967864 | Number of Points: 1033 points
```

Task C (Convergence)



```
Centroid 77: 26507.098306342945,-3754.6024568275384,-114.81572853612407
Centroid 78: 2664.3911349585655,4633.546605849815,-72.78405420841924
Centroid 79: 27160.010547531307,-144.74624394884535,10.748196661672313
Centroid 80: 29959.98514230785,3750.339168797623,-140.14671298387756
Centroid 81: 30935.60244327642,-3958.940828767624,-65.62045726931798
Centroid 82: 31913.506919508884,302.5114671711909,-49.655570810210826
Centroid 83: 38230.21816946891,-1683.4176948318511,159.15279709763237
Centroid 84: 3858.8245529511332,-2154.246852337512,-72.45768656955322
Centroid 85: 39044.93347018802,3899.1424275668874,-182.2934928498707
Centroid 86: 3954.886713311318,-4658.733940294479,-44.50593759896657
Centroid 87: 49140.88620105533,-630.948216568444,-54.60059699006024
Centroid 88: 5360.0551180609245,1184.3395827764414,297.7363504140519
Centroid 89: 5783.591089654008,3554.8133677466026,268.58739993092223
Centroid 90: 5957.501790322647,-619.3604285673913,143.60941819884488
Centroid 91: 5970.260299364924,-3766.8256469550556,-11.492241266061269
Centroid 92: 6024.662859297017,4838.551606226644,457.515770632767
Centroid 93: 614396.844620168,2801.998755726393,910.6786307685968
Centroid 94: 720.1306302987209,-4702.304573767187,160.55576353322115
Centroid 95: 8133.393144167297,3734.445966481086,-272.21723403121
Centroid 96: 8185.385249447876,-1969.4558528353314,-102.69047185746001
Centroid 97: 8547.212690979919,722.4271422460675,226.1320156561345
Centroid 98: 8703.85173399999,-4040.262461915782,174.69338620870212
Centroid 99: 9684.949590794306,-5157.0303212563595,-472.6634819501443
=== Task (c): Multi-Iteration K-Means with Early Termination ===
Converged after 42 iterations.
=== Centroids ===
Centroid 0: -10191.538973596045,-4671.745749539993,73.42963581640163
Centroid 1: -11722.617003402802,-2457.2831874257054,99.90201334700126
Centroid 2: -11751.723502073253,650.6529379299419,-4.834190346333643
Centroid 3: -12394.414231308006,4245.869821736836,84.7327297794913
Centroid 4: -12569.334961473749,-4395.683769664479,-174.74691656365803
Centroid 5: -13870.709175047663,-1416.901626085707,-26.20258264915002
Centroid 6: -14237.126886254311,1056.6213440075578,-150.68847469788722
Centroid 7: -15380.955428809828,2334.37030849334,211.94432563114637
Centroid 8: -15537.553927370547,4624.97434873775,-8.621699840605238
Centroid 9: -15588.36367570346,-4095.183558409099,-11.308557770107804
Centroid 10: -1633.9322433939901,-4375.6091647090625,-174.72169401593783
Centroid 11: -16663.864356244198,-816.5651351911104,6.627421788698217
Centroid 12: -18067.037562194528,-3250.1667042152276,1.5647122717507544
Centroid 13: -18253.51450399208,-4865.845324412149,-364.4971366506231
Centroid 14: -18473.811262452833,3722.466593541445,388.8404160798132
Centroid 15: -18574.97773541348,-472.9127084780152,201.1962527405747
Centroid 16: -1950.0985453618673,-460.1025353547972,226.58697172910485
Centroid 17: -19723.720211476175,1342.4913041179398,-209.74449826595355
```

3- Get Creative! [40 Points]

a. BYOD (Bring Your Own Data):

- **Customer Personality Analysis** is a technique that allows businesses to acquire a better knowledge of their ideal customers. Businesses can tailor their products and services to the unique demands, behaviors, and concerns of each client segment by analyzing them. This targeted approach enables businesses to optimize marketing efforts by focussing on clients who are most likely to buy a product rather than selling to the complete customer base.

- For Pre-Processing the Dataset - We began using a Kaggle dataset containing patient data and 29 features. To make our Java code easier to process and build, we decreased the dataset's dimensionality to three characteristics, denoted by x, y, and z. This dimensionality reduction was done in Task 2 to simplify the dataset for computational efficiency.

- Data Cleaning and Standardization, Before lowering dimensionality, we cleaned the data. This stage required dealing with missing or incorrect values in the matrix's rows

and columns (r and c). After cleaning the dataset, we used standardization to guarantee that the features had a uniform scale, which is required for effective dimensionality reduction. Standardization rescales the data so that each feature has a mean of 0 and a standard deviation of 1, allowing the Principal Component Analysis (PCA) procedure to function more reliably.

- Lastly, Dimensionality Reduction Using PCA After cleaning and standardizing. The dataset, we reduced it to three dimensions using Principal Component Analysis (PCA). PCA is a method for converting data into a set of orthogonal components with the greatest variance. By condensing the dataset's information to three primary components, we were able to keep the most important parts. We used NumPy to reduce dimensionality. PCA created a dense matrix of updated data that we could easily input into our Java code for further processing. This stage streamlined the data, making it easier to handle and assess in future operations for our code.

b. Evaluate the output of the clustering algorithm "*Silhouette*".

K = 3 : The clustering technique performs well with a Silhouette score of 0.6186 (K = 3). This implies that the clusters are well-separated, with data points closer to their particular clusters than to others, implying that the K=3 solution is extremely effective for this dataset.

```
Updated Clustering Evaluation Report:
1. K-Means Clustering Results:
=== Centroids ===
Centroid 0: -21802.68390426056,28.37865096345726,11.36707411598497
Centroid 1: 27797.449153536334,44.017145706239766,-3.5132460876444624
Centroid 2: 3734.4599600024626,-63.86339219188164,-10.024505995564208

2. Silhouette Score Evaluation:
   Overall Silhouette Score: 0.6186280655201676
   Interpretation: [See interpretation above]
```

K = 5 : A Silhouette score of 0.476 indicates acceptable grouping. Though the score is lower than for K=3, it still suggests adequate cluster separation and cohesiveness, with a little decrease in performance as the number of clusters grows.
The average silhouette score of 0.5473 for K=3 and K=5 indicates balanced performance, but further optimization may improve results. Both K=3 and K=5 offer plausible solutions, with K=3 marginally exceeding. Finally, the Silhouette evaluation demonstrates how the method performs

```
Updated Clustering Evaluation Report:
1. K-Means Clustering Results:
=== Centroids ===
Centroid 0: -18833.225458581812,-982.6554890294938,16.612895922471644
Centroid 1: -28512.2279590318,2602.1598554772518,12.389644658679542
Centroid 2: -40261.364989388196,2484.992129075358,60.42818034064102
Centroid 3: -43247.327029960456,-2914.1440401112536,-121.86159084876817
Centroid 4: 12773.271936976354,9.914676986659785,-8.328557297933298

2. Silhouette Score Evaluation:
   Overall Silhouette Score: 0.4761188000180853
```

with various K values. The higher score for K=3 indicates better-defined clusters, but K=5 still performs pretty well.

K=7 : Total Silhouette Score: 0.4858 ,Interpretation: The clustering for K=7 lacks a strong or distinguishable structure. It suggests that the clusters may not be well-separated, with data points potentially being assigned to arbitrary clusters, implying that the algorithm or dataset requires further optimisation or changes.

```
=== Task (e): Silhouette Score Evaluation ===
Overall Silhouette Score: 0.4857700718891826
Silhouette Score Interpretation:
Weak structure found. The clustering may be arbitrary.

Updated Clustering Evaluation Report:
1. K-Means Clustering Results:
=== Centroids ===
Centroid 0: -12856.263837201734,328.46249210690564,-4.909393313138605
Centroid 1: -20506.928850287666,6.042531813107754,8.732081232098766
Centroid 2: -33016.19551870957,147.32488912040483,17.69195631424592
Centroid 3: -6355.963149896548,-695.3588878048778,-1.9881955558847955
Centroid 4: 14744.31559905834,-26.52936129732165,-5.356933516850448
Centroid 5: 2141.679289988163,120.51979055729242,28.399263803192554
Centroid 6: 33225.25814256535,45.73608158741366,-36.66985585672809

2. Silhouette Score Evaluation:
   Overall Silhouette Score: 0.4857700718891826
   Interpretation: [See interpretation above]
```

K=50 : Overall silhouette score: 0.4253. Interpretation: Clustering with K=50 has a weaker structure than prior evaluations, implying that the huge number of clusters may have resulted in overfitting or breaking data into numerous arbitrary or less meaningful groups. This result suggests that the clustering algorithm may not be well-suited to this value of K, and additional tweaking may be required.

```
Updated Clustering Evaluation Report:
1. K-Means Clustering Results:
=== Centroids ===
Centroid 0: -10706.278296834293,-118.16595110416331,18.02258017119816
Centroid 1: -11561.25652663303,-4095.382410092767,-432.8849354258641
Centroid 2: -12832.766978364254,4594.925726892419,3.4371348518488922
Centroid 3: -13386.432647361027,-2486.6194702649814,36.09976866113162
Centroid 4: -15028.414955594382,2265.6986467592055,77.26048613719486
Centroid 5: -16015.866204127122,-4800.664247424829,-264.12183133507324
Centroid 6: -1785.7830956584623,2809.1885031107363,-110.18146137546853
Centroid 7: -17535.66136976127,-3021.9428744777333,24.24768072712916
Centroid 8: -18248.37893270136,770.794754599433,49.650654146422646
Centroid 9: -20174.941101670738,4588.838172925966,179.1565049114489
Centroid 10: -20920.74981996778,-1262.89730109506,-52.269624474570186
Centroid 11: -21717.098288737423,2154.6713609435606,40.05568164992845
Centroid 12: -22913.523202861616,-4076.8094176921104,109.41209639282776
Centroid 13: -23528.308846226933,-1829.087259672445,-517.5183493689922
Centroid 14: -24952.14707727689,1387.8833223302177,63.448870943725446
Centroid 15: -27973.51692653096,-3516.043815176994,-8.777182637064989
Centroid 16: -28320.002628790426,3898.519108456147,-26.18314336844569
Centroid 17: -31545.36504114053,-1839.0116634227843,157.57744531998327
Centroid 18: -35559.70503906677,1216.4181889225242,-16.22376538215205
Centroid 19: -44350.99786288196,1277.6314768693528,-34.589342060235765
Centroid 20: -5373.616127427295,-2452.3167881994177,-15.319796744570233
Centroid 21: -6155.649774323215,4304.291824712866,68.99412861149978
Centroid 22: -8802.8954102653,2524.180138978442,-80.80251506929311
Centroid 23: -9250.348041000076,-4379.589040448941,166.63058376629934
Centroid 24: 9930.749631314118,-4236.239484743317,-84.81219095715522
Centroid 25: 11180.213090964167,-1705.4919992022667,-73.51358861385756
Centroid 26: 12388.402725555128,683.5517522224276,-60.83108730447821
Centroid 27: 12826.477273909248,4330.85903149043,129.35418292119687
Centroid 28: 14368.78640847944,3881.43649302593,-203.70954653566397
Centroid 29: 457.80135411870816,-1360.7601301067255,35.15063543255052
Centroid 30: 13875.438641809937,-4864.872445999399,115.82251856359495
Centroid 31: 14798.777915851335,-2441.5077565973193,30.998256244193335
Centroid 32: 16101.333061196869,449.709678401825,-88.63801707440889
Centroid 33: 17592.241367642397,3866.360203547392,41.29521841542299
Centroid 34: 18302.13440610193,-3241.6312919602046,267.1823419665626
Centroid 35: 20212.96912341134,8.271609988607828,-40.30801793056302
Centroid 36: 20979.391673827315,4323.199506709797,-216.67568620167367
Centroid 37: 21961.885652014895,-3925.112750248122,28.954863636090952
Centroid 38: 25260.11928165965,2534.9324014303866,-75.93075630738143
Centroid 39: 25855.51785114567,-2050.733337077978,50.02975032231433
Centroid 40: 2831.192151336269,-4200.452564123226,-73.5070254005828
Centroid 41: 29653.60143961574,1409.6293154732416,9.789436456203335
Centroid 42: 31824.97577507028,-3738.2311049198865,-64.24417725589957
Centroid 43: 3020.038513632848,3207.1182497728287,76.38907838068533
Centroid 44: 37188.69676606274,1816.1323887001945,5.927393196543938
Centroid 45: 114180.24895362444,144.669014862147,-204.40019381472968
Centroid 46: 6849.81354039599,3662.936202753248,75.29099362078107
Centroid 47: 7414.846390557808,-640.6270909953435,146.6128893808934
Centroid 48: 10065.246435551326,2462.8556708525693,-29.862473616317324
Centroid 49: 10096.204322739286,4441.151119724976,-55.46818460144751

2. Silhouette Score Evaluation:
   Overall Silhouette Score: 0.42533327961968476
```