Matt Adragna, Camilo Escobar, Jackson Lundberg

1-Spark RDDS

Step 1-

- We used a python script generateConcert.py in order to generate the data necessary. The python script includes optional parameters that allow for modification of the size of the data, percentage chance someone was infected and where it is saved. This is outlined in the readme in the project folder. The defaults are a size of 100MB and a .01 chance of someone being infected.

Step 2-

Query 1) For query 1, we mapped each infected person into 25x25 cells. If they were close to the edge of the cell (within 6 meters) we would add a dummy one of them to that adjacent cell as well.
I then joined the two RDDS on that cell as a key.
Following that -> filtered out every pair of infected,non infected that was not within 6 feet of one another and then returned it.

Query 2) Took the .distinct of a mapped query1 that only returned the infected -> this made it just infected who were within 6 feet with duplicates removed.

Query 3) Mapped people_some_infected to the same cells.Then grouped the cells by key so that it is in form (cell,[list of people in the cell]). Then flat mapped those to infected,sum(ofPeopleInfected) pairs. This is done by finding every infected person in the list and then iterating through the rest of the list to see if they had been in close contact with them (excluding the infected we are checking against).

2A)
Step 1:

- We used a python script generateTransactions.py in order to generate the data necessary. The python script includes optional parameters that allow for modification of the amount of customers, the amount of purchases and where it is saved. This is outlined in the readme in the project folder. The defaults are 50000 customers and 5000000 purchases

Step 2:

1) T1 was created by taking the purchases View that was created and running the SQL query `"SELECT * FROM purchases WHERE TransTotal <= 600"` over it
2) Part 2 was created by running the below query over t1

```
"SELECT TransNumItems, " +
"percentile_approx(TransTotal, 0.5) AS median," +
```

```
                "MIN(TransTotal) AS min," +
                "MAX(TransTotal) AS max " +
                "FROM T1 " +
                "GROUP BY TransNumItems"
                 "
```

T

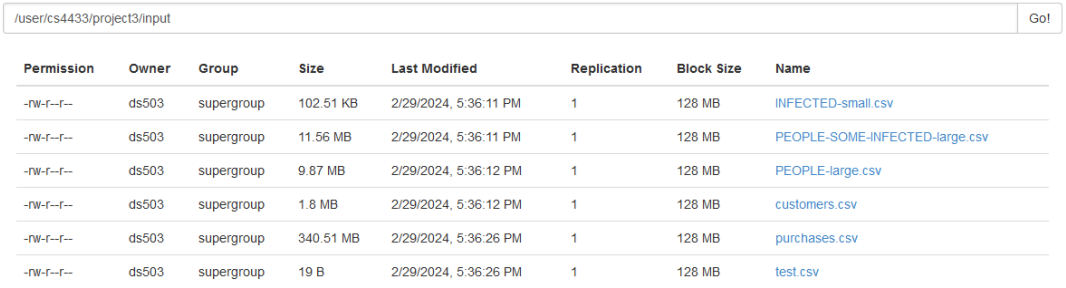
3) Part 3 was created by running the below query over t1

```
   " SELECT ID, Age, COUNT(TransNumItems) AS TotalItems, SUM(TransTotal)
AS TotalSpent " +
   "FROM T1 " +
   "JOIN customers ON T1.CustID = customers.ID " +
   "WHERE Age BETWEEN 18 AND 25 " +
   "GROUP BY ID, Age"
```

T

Part 1:

Inputted Data:

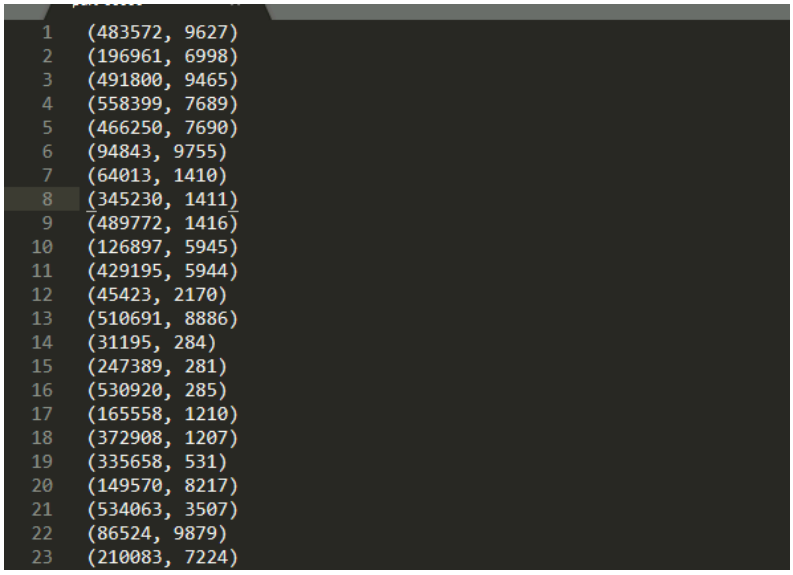


Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	ds503	supergroup	102.51 KB	2/29/2024, 5:36:11 PM	1	128 MB	INFECTED-small.csv
-rw-r--r--	ds503	supergroup	11.56 MB	2/29/2024, 5:36:11 PM	1	128 MB	PEOPLE-SOME-INFECTED-large.csv
-rw-r--r--	ds503	supergroup	9.87 MB	2/29/2024, 5:36:12 PM	1	128 MB	PEOPLE-large.csv
-rw-r--r--	ds503	supergroup	1.8 MB	2/29/2024, 5:36:12 PM	1	128 MB	customers.csv
-rw-r--r--	ds503	supergroup	340.51 MB	2/29/2024, 5:36:26 PM	1	128 MB	purchases.csv
-rw-r--r--	ds503	supergroup	19 B	2/29/2024, 5:36:26 PM	1	128 MB	test.csv

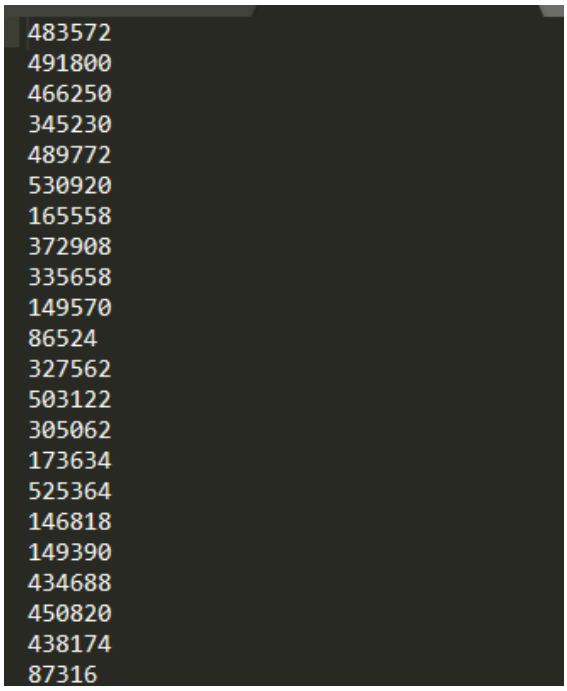
Outputted Data:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 2:44:42 PM	0	0 B	query1Results
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 2:44:44 PM	0	0 B	query2Results
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 2:44:53 PM	0	0 B	result3

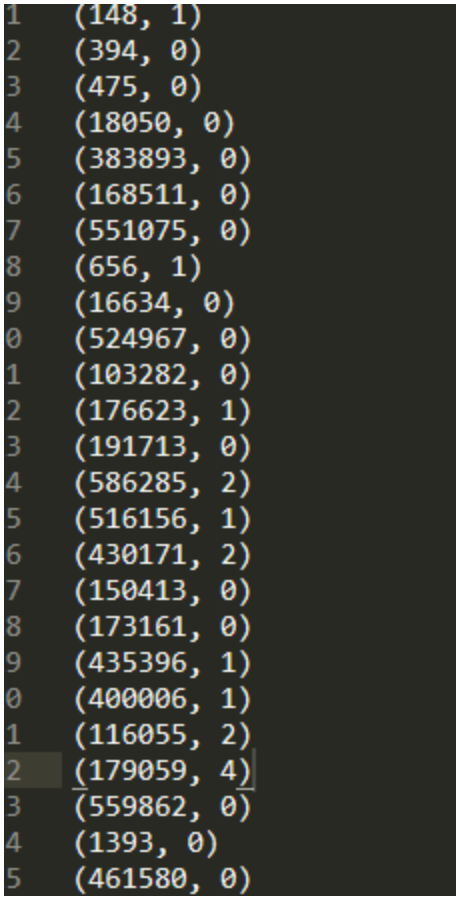
Portion of query 1 Results:



Portion of query 2 Results:

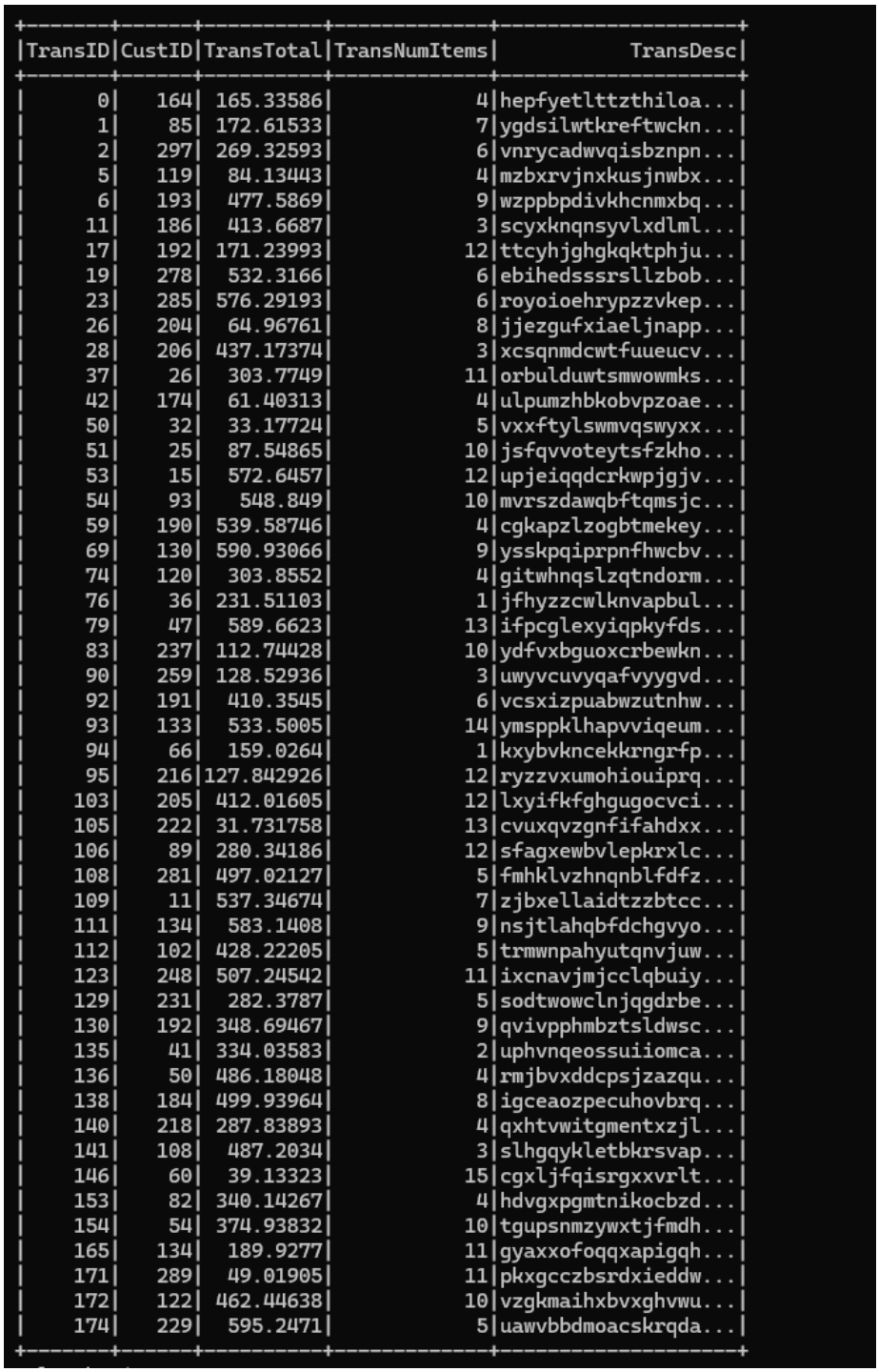


Portion of query 3 Results:


(148, 1)
(394, 0)
(475, 0)
(18050, 0)
(383893, 0)
(168511, 0)
(551075, 0)
(656, 1)
(16634, 0)
(524967, 0)
(103282, 0)
(176623, 1)
(191713, 0)
(586285, 2)
(516156, 1)
(430171, 2)
(150413, 0)
(173161, 0)
(435396, 1)
(400006, 1)
(116055, 2)
(179059, 4)
(559862, 0)
(1393, 0)
(461580, 0)

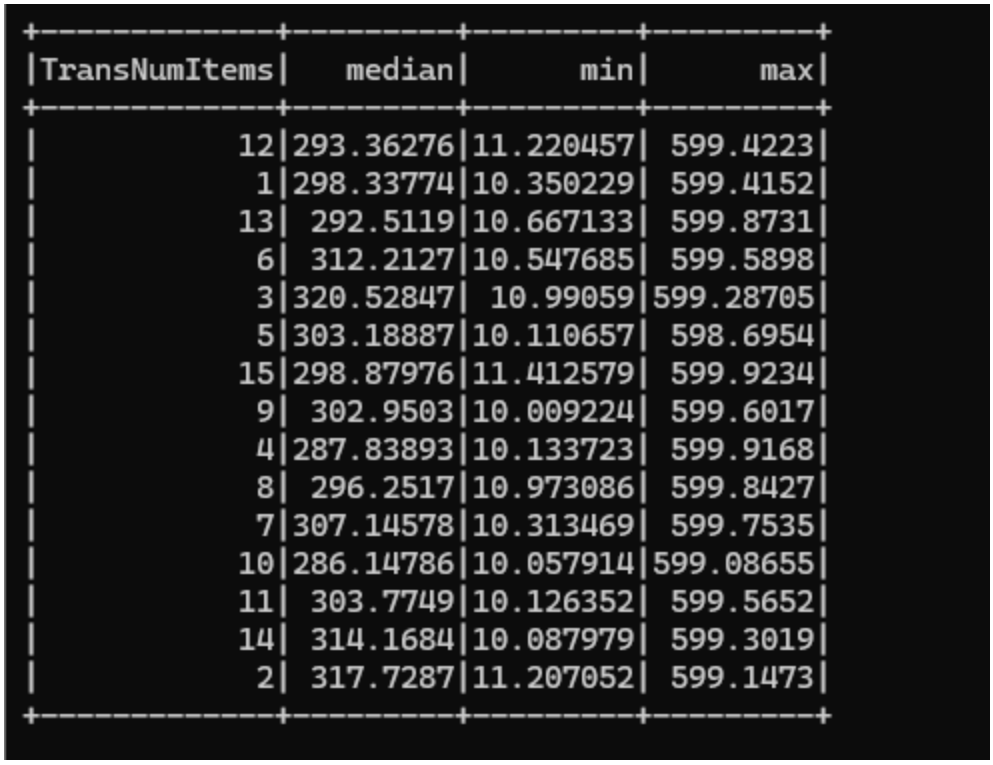
Part 2:

a)



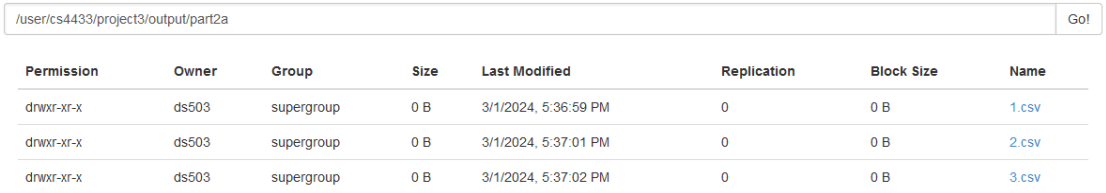
TransID	CustID	TransTotal	TransNumItems	TransDesc
0	164	165.33586	4	hepfyetlttzthiloa...
1	85	172.61533	7	ygdsilwtkreftwckn...
2	297	269.32593	6	vnrycadwvqisbznpn...
5	119	84.13443	4	mzbxrvjnxkusjnwbx...
6	193	477.5869	9	wzppbpdivkhcnmxbq...
11	186	413.6687	3	scyxknqnsyvlxdlml...
17	192	171.23993	12	ttcyhjghgkqktphju...
19	278	532.3166	6	ebihedsssrsllzbob...
23	285	576.29193	6	royoioehrypzzvkep...
26	204	64.96761	8	jjezgufxiaeljnapp...
28	206	437.17374	3	xcsqnmdcwtfuueucv...
37	26	303.7749	11	orbulduwtsmwowmks...
42	174	61.40313	4	ulpumzhbkobvpzoae...
50	32	33.17724	5	vxxftylswmvqswyxx...
51	25	87.54865	10	jsfqvvoteytsfzkho...
53	15	572.6457	12	upjeiqqdcrkwpjgjv...
54	93	548.849	10	mvrszdawqbftqmsjc...
59	190	539.58746	4	cgkapzlzogbtmekey...
69	130	590.93066	9	ysskpqiprpnfhwcbv...
74	120	303.8552	4	gitwhnqslzqtndorm...
76	36	231.51103	1	jfhyzzcwlknvapbul...
79	47	589.6623	13	ifpcglexyiqpkyfds...
83	237	112.74428	10	ydfvxbguoxcrbewkn...
90	259	128.52936	3	uwyvcuvyqafvyygvd...
92	191	410.3545	6	vcsxizpuabwzutnhw...
93	133	533.5005	14	ymsppklhapvviqeum...
94	66	159.0264	1	kxybvkncekkrngrfp...
95	216	127.842926	12	ryzzvxumohiouiprq...
103	205	412.01605	12	lxyifkfghgugocvci...
105	222	31.731758	13	cvuxqvzgnfifahdxx...
106	89	280.34186	12	sfagxewbvlepkrxlc...
108	281	497.02127	5	fmhklvzhnqnblfdfz...
109	11	537.34674	7	zjbxellaidtzzbtcc...
111	134	583.1408	9	nsjtlahqbfdchgvyo...
112	102	428.22205	5	trmwnpahyutqnvjuw...
123	248	507.24542	11	ixcnavjmjcclqbuiy...
129	231	282.3787	5	sodtwowclnjqgdrbe...
130	192	348.69467	9	qvivpphmbztsldwsc...
135	41	334.03583	2	uphvnqeossuiiomca...
136	50	486.18048	4	rmjbvxddcpsjzazqu...
138	184	499.93964	8	igceaozpecuhovbrq...
140	218	287.83893	4	qxhtvwitgmentxzjl...
141	108	487.2034	3	slhgqykletbkrsvap...
146	60	39.13323	15	cgxljfqisrgxxvrlt...
153	82	340.14267	4	hdvgxpgmtnikocbzd...
154	54	374.93832	10	tgupsnmzywxtjfmdh...
165	134	189.9277	11	gyaxxofoqqxapigqh...
171	289	49.01905	11	pkxgcczbsrdxieddw...
172	122	462.44638	10	vzgkmaihxbvxghvwu...
174	229	595.2471	5	uawvbbdmoacskrqda...

b)



TransNumItems	median	min	max
12	293.36276	11.220457	599.4223
1	298.33774	10.350229	599.4152
13	292.5119	10.667133	599.8731
6	312.2127	10.547685	599.5898
3	320.52847	10.99059	599.28705
5	303.18887	10.110657	598.6954
15	298.87976	11.412579	599.9234
9	302.9503	10.009224	599.6017
4	287.83893	10.133723	599.9168
8	296.2517	10.973086	599.8427
7	307.14578	10.313469	599.7535
10	286.14786	10.057914	599.08655
11	303.7749	10.126352	599.5652
14	314.1684	10.087979	599.3019
2	317.7287	11.207052	599.1473

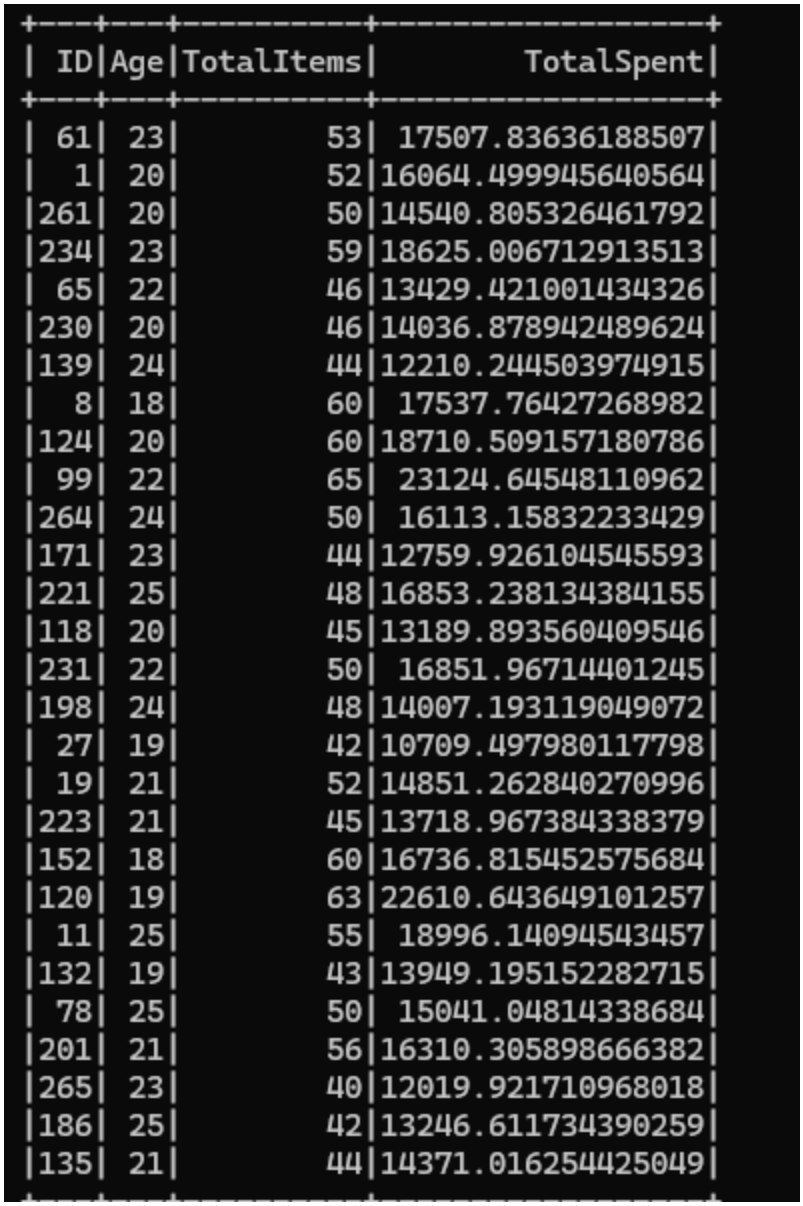
Outputted data:



/user/cs4433/project3/output/part2a Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 5:36:59 PM	0	0 B	1.csv
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 5:37:01 PM	0	0 B	2.csv
drwxr-xr-x	ds503	supergroup	0 B	3/1/2024, 5:37:02 PM	0	0 B	3.csv

c)



ID	Age	TotalItems	TotalSpent
61	23	53	17507.83636188507
1	20	52	16064.499945640564
261	20	50	14540.805326461792
234	23	59	18625.006712913513
65	22	46	13429.421001434326
230	20	46	14036.878942489624
139	24	44	12210.244503974915
8	18	60	17537.76427268982
124	20	60	18710.509157180786
99	22	65	23124.64548110962
264	24	50	16113.15832233429
171	23	44	12759.926104545593
221	25	48	16853.238134384155
118	20	45	13189.893560409546
231	22	50	16851.96714401245
198	24	48	14007.193119049072
27	19	42	10709.497980117798
19	21	52	14851.262840270996
223	21	45	13718.967384338379
152	18	60	16736.815452575684
120	19	63	22610.643649101257
11	25	55	18996.14094543457
132	19	43	13949.195152282715
78	25	50	15041.04814338684
201	21	56	16310.305898666382
265	23	40	12019.921710968018
186	25	42	13246.611734390259
135	21	44	14371.016254425049

Contribution Statement

Matt, Camilo, Jackson: All members of our team met and communicated over text to discuss and work on the project. We all helped each other set up and work on the tasks. We would reach out to each other over text messages to ask for help. We worked together for the most part, and any individual progress was explained to the rest of the group to catch us all up.

Resource Usage Statement

Matt:

- I used the resources on canvas such as the slides and helpful links. I consulted the pyspark documentation on https://spark.apache.org/docs/latest/api/python in order understand what I could do and the syntax for it.
- I also used ChatGPT in order to help with explaining example code, ‘helping’ diagnose bugs and provide syntax in context.

Jackson:

- I used the resources provided by the professor and TAs on canvas, such as discussion boards, helpful links, and the project resources to help me with this assignment. I used the Spark documentation to help me learn more about working with spark and understanding the syntax.
- https://spark.apache.org/docs/latest/api/python/index.html

Camilo:

- Used resources from the lectures and slides on Canvas to help with the assignment. I also watched YouTube videos and online documentation to assist in the query creation.
- Used ChatGPT to help format code and ask for method explanations