RBDA Project

Project - Rachit Jain - rj2219

Explanation

Dataset description:

1. The dataset that I used is NY citywide payroll dataset for a fiscal year [1]. This data was collected to study how the state's budget was being allocated.
2. The data has data points from fiscal year 2014 to 2020.
3. The data consists of the following columns:

Column Name	Description	Preprocessed	Kept/Dropped/Converted
Fiscal Year	Fiscal year for the data point	No	Kept
Payroll number	Agency number	No	Dropped
Agency Name	Name of the agency which was making payments	The agency names had commas which needed to be preprocessed so that they could be read in the `mapper`.	Kept
Last Name	Last name of the employee	No	Dropped
First Name	First name of the employee	No	Dropped
Mid Init	Middle initial of the employee	No	Dropped
Agency Start Date	Date on which employee began working	No	Dropped

Column Name	Description	Preprocessed	Kept/Dropped/Converted
Work Location Borough	Borough of employee's work location	Removed boroughs other than NYC boroughs - had boroughs like Washington DC, Albany, etc.	Kept
Title Description	Civil service title description of the employee	The agency names had commas which needed to be preprocessed so that they could be read in the `mapper`.	Dropped
Leave Status as of June 30th	Status of employment	No	Dropped
Base Salary	Base salary of the employee	Some salaries were not in `per annum` calculation. Converted those.	Converted
Pay Basis	When the employee is paid	Used this in calculation for getting total salary.	Converted
Regular Hours	Number of hours worked	No. Data issue	Dropped
Regular Gross Paid	Amount paid to the employee	No. Data issues	Dropped
OT Hours	Number of over time hours worked	No	Kept
Total OT Paid	Total over time paid	No	Converted
Total Other Pay	Salary in addition to base	No	Converted

4. The data set had a few issues where some columns did not have the correct data.
5. Total rows:
    a. Before profiling and cleaning: 3923291

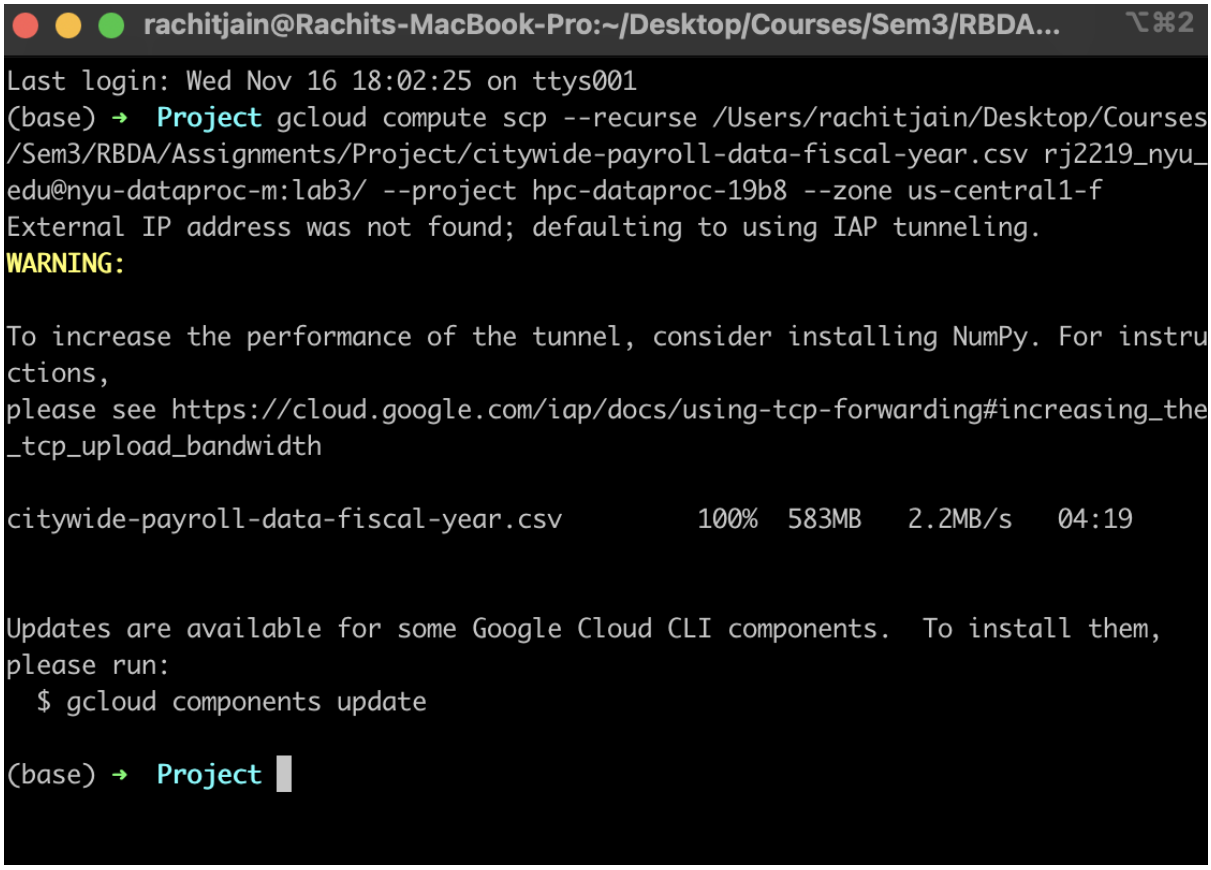
b. After profiling and cleaning: 2470563

Data Profiling and Cleaning:

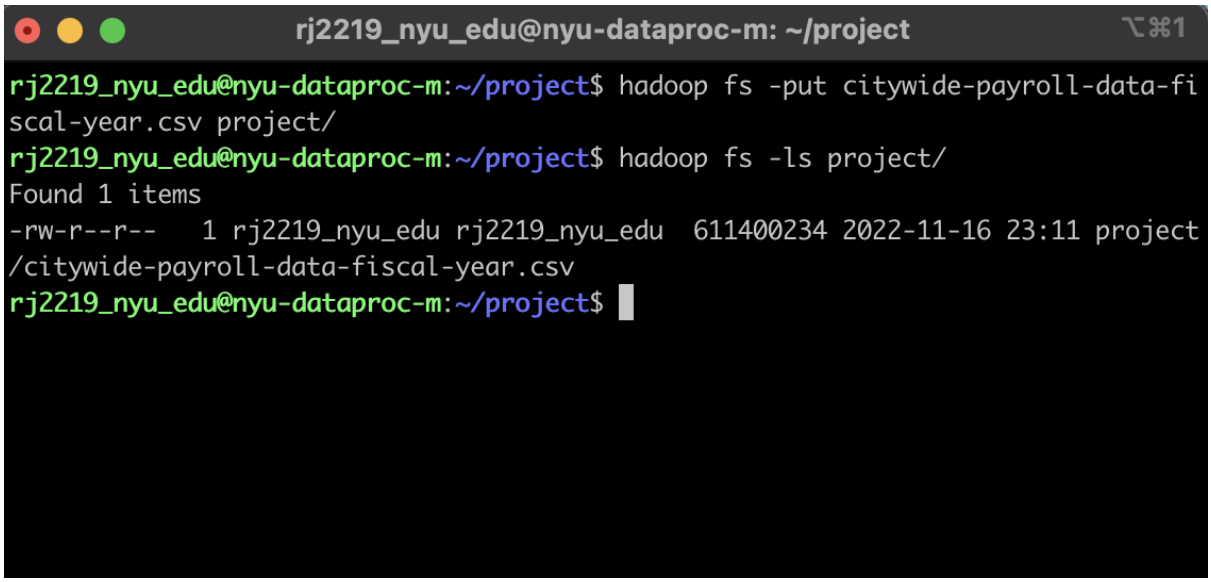
1. The code has been uploaded to `GitHub` in a private repository [2].
2. The overall idea for the data is to drop the columns that are not required and profile the data to get the overall `totalPay` of a person which we will later relate to the purchasing capacity of the borough.
3. I set the `mapred.textoutputformat.separator` parameter as `,` as I need to write the data in `.csv` format.
4. My input data is in the `.csv` format and output data is being written as `key -> processed_row`.
5. In my case, the `key` corresponds to the borough since all our analysis is borough based and the `processed_row` corresponds to the row of the csv after the columns are dropped and the row has been profiled and cleaned.
6. The mapper has a `set` of boroughs which are used for the processing - `"MANHATTAN", "BRONX", "BROOKLYN", "QUEENS", "STATEN ISLAND”`.
7. The first step of cleaning the data was reading the input file which had `,` in the descriptions. This was solved by checking if the text was in a `string` format indicating one column in the data and the comma was converted to a semi colon. (`, → ;`).
8. The above step made the parsing of the row based on a comma and we could easily divide the arguments.
9. If the borough was not an NYC borough, it was not included in the final HDFS cluster and the row was dropped.
10. I checked the `payBasis` and based on that, I converted each calculation into `per annum`.
11. Based on heuristics in [3], the following calculations were made:
    a. `per annum or prorated annual` → `base salary + total OT paid + other allowances`
    b. `per day` → `(base salary * number of days * working weeks) + total OT paid + other allowances`
    c. `per hour` → `(base salary * number of hours * number of days * working weeks) + total OT paid + other allowances`

12. The average working hours are `7` and the average number of working weeks are `49`.
13. Based on the above, the total pay of the person was calculated.
14. In the reducer, the output was going to be of the same format as that from the mapper but the only difference would be the I am dropping the rows which have a pay of less than `20000.`
15. Minimum wage in NYC is `$14.20` per hour by end of December 2022 [4].
16. Using this for the calculation:
    a. $14.20 * 7 * 5 * 49 = 24,453$
    b. But the minimum wage has been lesser in the previous years.
    c. Using this, I came up with the heuristic of removing the total salaries of `< 20,000` since that data might be incorrectly recorded.
17. The rows that were dropped in the reducer were used to form a statistic as to how many data points could be incorrect during the parsing of the data.
18. The dropped rows were aggregated and the final count of the `borough_year` were written to the HDFS.
19. This could give us more insights as to how the data was recorded and how many rows could finally be used for the processing.
20. The final outputs were written to the HDFS with the type `Text, Text`.

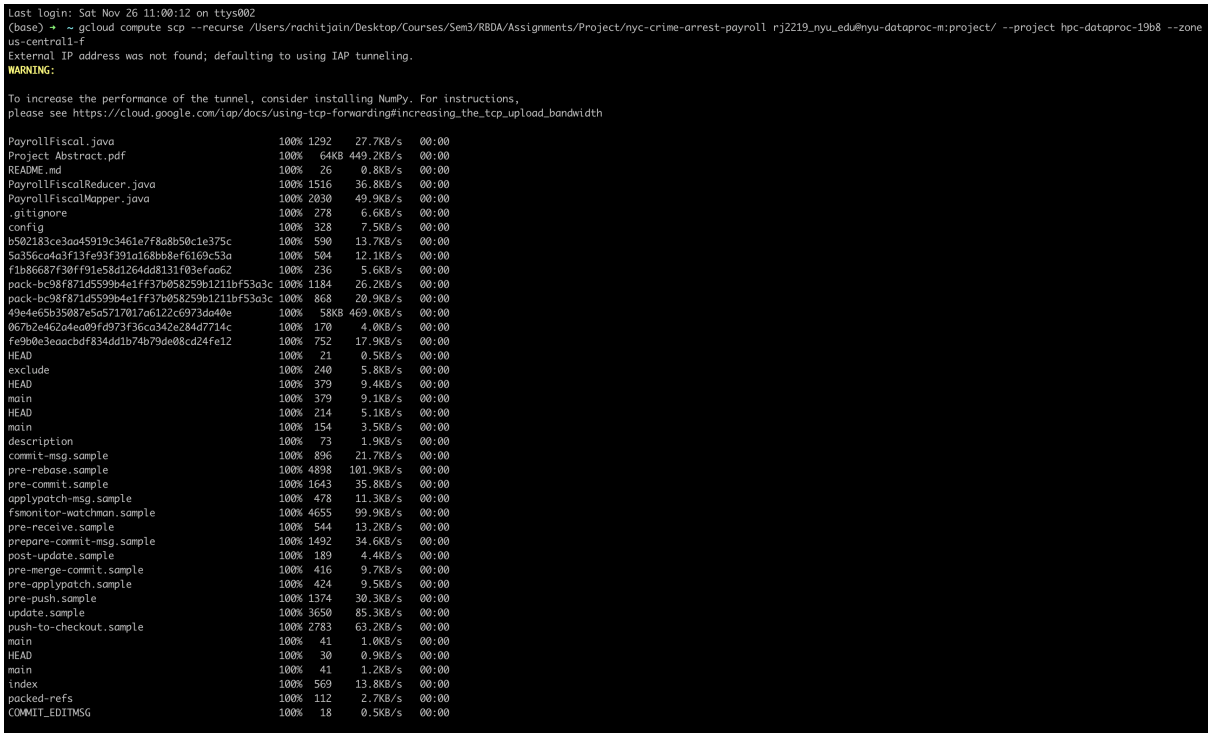
Screenshots



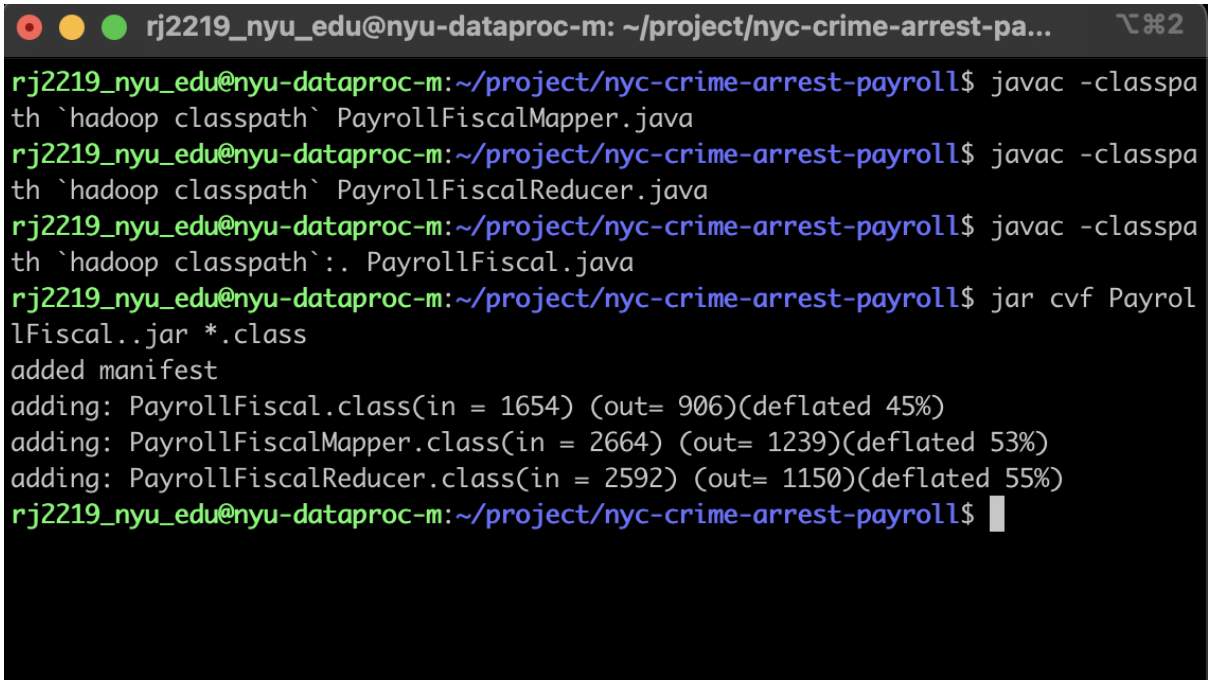
`scp` command for pushing data to the dataproc cluster



Showing that the dataset has been pushed onto the `hadoop` file system



Code files being pushed to the cluster. The extra files here are the files due to `git`.



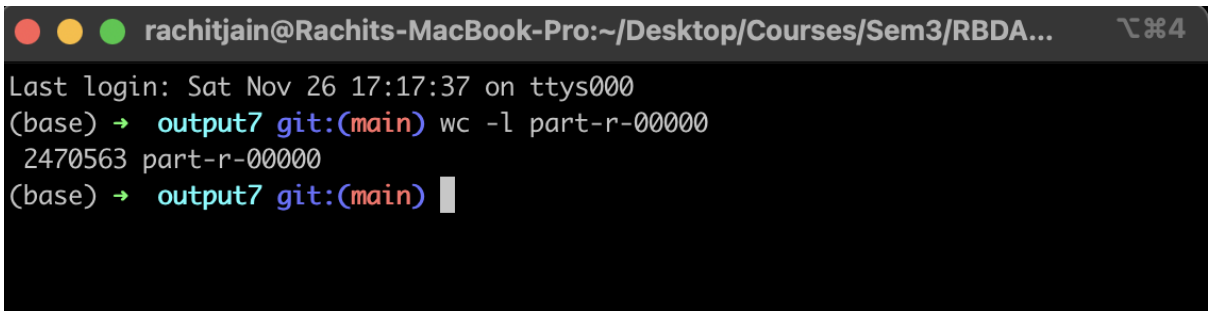
Setting `classpath` and creating the `.jar` on hadoop.

```
rj2219_nyu_edu@nyu-dataproc-m:~/project/nyc-crime-arrest-payroll$ hadoop jar PayrollFiscal.jar PayrollFiscal project/citywide-payroll-data-fiscal-year.csv project/output7
2022-11-27 19:37:53,010 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.39:8032
2022-11-27 19:37:53,181 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.39:10200
2022-11-27 19:37:53,485 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-11-27 19:37:53,502 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/rj2219_nyu_edu/.staging/job_1668611926937_1075
2022-11-27 19:37:53,748 INFO input.FileInputFormat: Total input files to process : 1
2022-11-27 19:37:53,871 INFO mapreduce.JobSubmitter: number of splits:5
2022-11-27 19:37:53,945 INFO Configuration.deprecation: mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2022-11-27 19:37:54,043 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1668611926937_1075
2022-11-27 19:37:54,044 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-27 19:37:54,223 INFO conf.Configuration: resource-types.xml not found
2022-11-27 19:37:54,223 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-27 19:37:54,354 INFO impl.YarnClientImpl: Submitted application application_1668611926937_1075
2022-11-27 19:37:54,389 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1668611926937_1075/
2022-11-27 19:37:54,390 INFO mapreduce.Job: Running job: job_1668611926937_1075
2022-11-27 19:38:02,546 INFO mapreduce.Job: Job job_1668611926937_1075 running in uber mode : false
2022-11-27 19:38:02,548 INFO mapreduce.Job:  map 0% reduce 0%
2022-11-27 19:38:11,657 INFO mapreduce.Job:  map 20% reduce 0%
2022-11-27 19:38:12,663 INFO mapreduce.Job:  map 40% reduce 0%
2022-11-27 19:38:13,668 INFO mapreduce.Job:  map 100% reduce 0%
2022-11-27 19:38:27,738 INFO mapreduce.Job:  map 100% reduce 100%
2022-11-27 19:38:28,750 INFO mapreduce.Job: Job job_1668611926937_1075 completed successfully
2022-11-27 19:38:28,838 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=204221036
                FILE: Number of bytes written=409919379
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=611417378
                HDFS: Number of bytes written=147370588
                HDFS: Number of read operations=20
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Rack-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=171216
                Total time spent by all reduces in occupied slots (ms)=43980
                Total time spent by all map tasks (ms)=42804
                Total time spent by all reduce tasks (ms)=10995
                Total vcore-milliseconds taken by all map tasks=42804
                Total vcore-milliseconds taken by all reduce tasks=10995
                Total megabyte-milliseconds taken by all map tasks=175325184
                Total megabyte-milliseconds taken by all reduce tasks=45035520
        Map-Reduce Framework
```

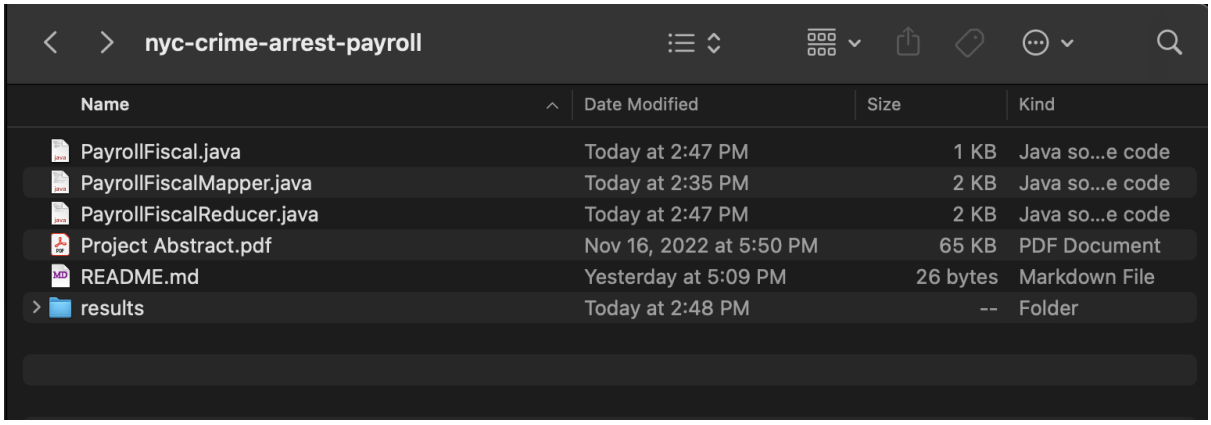
Screenshot showing running of the `mapreduce` task.

```
(base) ➜  nyc-crime-arrest-payroll git:(main) ✗ git status
On branch main
Your branch is up to date with 'origin/main'.

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
        modified:   .gitignore
        modified:   PayrollFiscal.java
        modified:   PayrollFiscalMapper.java
        modified:   PayrollFiscalReducer.java

no changes added to commit (use "git add" and/or "git commit -a")
(base) ➜  nyc-crime-arrest-payroll git:(main) ✗ git add .
(base) ➜  nyc-crime-arrest-payroll git:(main) ✗ git commit -m "Implemented Payro
llFiscal"
[main 434849b] Implemented PayrollFiscal
 4 files changed, 25 insertions(+), 22 deletions(-)
(base) ➜  nyc-crime-arrest-payroll git:(main) git push origin main
Enumerating objects: 11, done.
Counting objects: 100% (11/11), done.
Delta compression using up to 8 threads
Compressing objects: 100% (6/6), done.
Writing objects: 100% (6/6), 1001 bytes | 1001.00 KiB/s, done.
Total 6 (delta 4), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (4/4), completed with 4 local objects.
To https://github.com/rachitjain2706/nyc-crime-arrest-payroll.git
   f6067b2..434849b  main -> main
(base) ➜  nyc-crime-arrest-payroll git:(main)
```

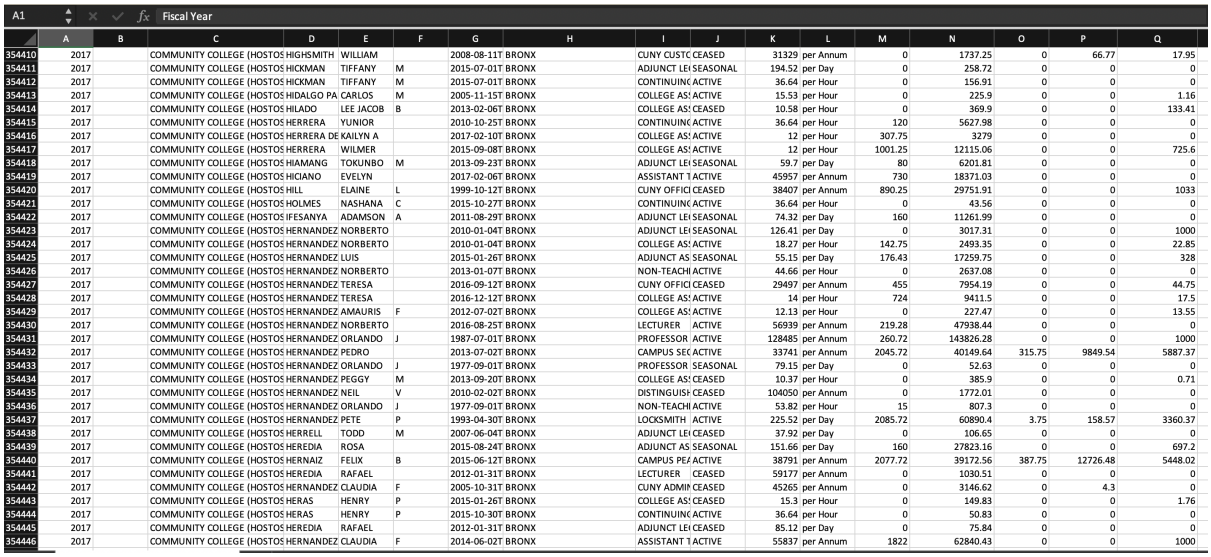
`git status` output shown and data pushed to git.



Output file after the mapreduce task ends. Shows the number of records kept after preprocessing.



Directory structure



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
354410	2017		COMMUNITY COLLEGE (HOSTOS	HIGHSMITH	WILLIAM		2008-08-11T	BRONX	CUNY CUSTC	CEASED	31329	per Annum	0	1737.25	0	66.77	17.95
354411	2017		COMMUNITY COLLEGE (HOSTOS	HICKMAN	TIFFANY	M	2015-07-01T	BRONX	ADJUNCT LEI	SEASONAL	194.52	per Day	0	258.72	0	0	0
354412	2017		COMMUNITY COLLEGE (HOSTOS	HICKMAN	TIFFANY	M	2015-07-01T	BRONX	CONTINUINC	ACTIVE	36.64	per Hour	0	156.91	0	0	0
354413	2017		COMMUNITY COLLEGE (HOSTOS	HIDALGO PA	CARLOS	M	2005-11-15T	BRONX	COLLEGE AS	ACTIVE	15.53	per Hour	0	225.9	0	0	1.16
354414	2017		COMMUNITY COLLEGE (HOSTOS	HILADO	LEE JACOB	B	2013-02-06T	BRONX	COLLEGE AS	CEASED	10.58	per Hour	0	369.9	0	0	133.41
354415	2017		COMMUNITY COLLEGE (HOSTOS	HERRERA	YUNIOR		2010-10-25T	BRONX	CONTINUINC	ACTIVE	36.64	per Hour	120	5627.98	0	0	0
354416	2017		COMMUNITY COLLEGE (HOSTOS	HERRERA DE	KAILYN A		2017-02-10T	BRONX	COLLEGE AS	ACTIVE	12	per Hour	307.75	3279	0	0	0
354417	2017		COMMUNITY COLLEGE (HOSTOS	HERRERA	WILMER		2015-09-08T	BRONX	COLLEGE AS	ACTIVE	12	per Hour	1001.25	12115.06	0	0	725.6
354418	2017		COMMUNITY COLLEGE (HOSTOS	HIAMANG	TOKUNBO	M	2013-09-23T	BRONX	ADJUNCT LEI	SEASONAL	59.7	per Day	80	6201.81	0	0	0
354419	2017		COMMUNITY COLLEGE (HOSTOS	HICIANO	EVELYN		2017-02-06T	BRONX	ASSISTANT T	ACTIVE	45957	per Annum	730	18371.03	0	0	0
354420	2017		COMMUNITY COLLEGE (HOSTOS	HILL	ELAINE	L	1999-10-12T	BRONX	CUNY OFFICI	CEASED	38407	per Annum	890.25	29751.91	0	0	1033
354421	2017		COMMUNITY COLLEGE (HOSTOS	HOLMES	NASHANA	C	2015-10-27T	BRONX	CONTINUINC	ACTIVE	36.64	per Hour	0	43.56	0	0	0
354422	2017		COMMUNITY COLLEGE (HOSTOS	IFESANYA	ADAMSON	A	2011-08-29T	BRONX	ADJUNCT LEI	SEASONAL	74.32	per Day	160	11261.99	0	0	0
354423	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	NORBERTO		2010-01-04T	BRONX	ADJUNCT LEI	SEASONAL	126.41	per Day	0	3017.31	0	0	1000
354424	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	NORBERTO		2010-01-04T	BRONX	COLLEGE AS	ACTIVE	18.27	per Hour	142.75	2493.35	0	0	22.85
354425	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	LUIS		2015-01-26T	BRONX	ADJUNCT AS	SEASONAL	55.15	per Day	176.43	17259.75	0	0	328
354426	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	NORBERTO		2013-01-07T	BRONX	NON-TEACHI	ACTIVE	44.66	per Hour	0	2637.08	0	0	0
354427	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	TERESA		2016-09-12T	BRONX	CUNY OFFICI	CEASED	29497	per Annum	455	7954.19	0	0	44.75
354428	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	TERESA		2016-12-12T	BRONX	COLLEGE AS	ACTIVE	14	per Hour	724	9411.5	0	0	17.5
354429	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	AMAURIS	F	2012-07-02T	BRONX	COLLEGE AS	ACTIVE	12.13	per Hour	0	227.47	0	0	13.55
354430	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	NORBERTO		2016-08-25T	BRONX	LECTURER	ACTIVE	56939	per Annum	219.28	47938.44	0	0	0
354431	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	ORLANDO	J	1987-07-01T	BRONX	PROFESSOR	ACTIVE	128485	per Annum	260.72	143826.28	0	0	1000
354432	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	PEDRO		2013-07-02T	BRONX	CAMPUS SEC	ACTIVE	33741	per Annum	2045.72	40149.64	315.75	9849.54	5887.37
354433	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	ORLANDO	J	1977-09-01T	BRONX	PROFESSOR	SEASONAL	79.15	per Day	0	52.63	0	0	0
354434	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	PEGGY	M	2013-09-20T	BRONX	COLLEGE AS	CEASED	10.37	per Hour	0	385.9	0	0	0.71
354435	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	NEIL	V	2010-02-02T	BRONX	DISTINGUISH	CEASED	104050	per Annum	0	1772.01	0	0	0
354436	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	ORLANDO	J	1977-09-01T	BRONX	NON-TEACHI	ACTIVE	53.82	per Hour	15	807.3	0	0	0
354437	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	PETE	P	1993-04-30T	BRONX	LOCKSMITH	ACTIVE	225.52	per Day	2085.72	60890.4	3.75	158.57	3360.37
354438	2017		COMMUNITY COLLEGE (HOSTOS	HERRELL	TODD	M	2007-06-04T	BRONX	ADJUNCT LEI	CEASED	37.92	per Day	0	106.65	0	0	0
354439	2017		COMMUNITY COLLEGE (HOSTOS	HEREDIA	ROSA		2015-08-24T	BRONX	ADJUNCT AS	SEASONAL	151.66	per Day	160	27823.16	0	0	697.2
354440	2017		COMMUNITY COLLEGE (HOSTOS	HERNAIZ	FELIX	B	2015-06-12T	BRONX	CAMPUS PEA	ACTIVE	38791	per Annum	2077.72	39172.56	387.75	12726.48	5448.02
354441	2017		COMMUNITY COLLEGE (HOSTOS	HEREDIA	RAFAEL		2012-01-31T	BRONX	LECTURER	CEASED	59177	per Annum	0	1030.51	0	0	0
354442	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	CLAUDIA	F	2005-10-31T	BRONX	CUNY ADMIN	CEASED	45265	per Annum	0	3146.62	0	4.3	0
354443	2017		COMMUNITY COLLEGE (HOSTOS	HERAS	HENRY	P	2015-01-26T	BRONX	COLLEGE AS	CEASED	15.3	per Hour	0	149.83	0	0	1.76
354444	2017		COMMUNITY COLLEGE (HOSTOS	HERAS	HENRY	P	2015-10-30T	BRONX	CONTINUINC	ACTIVE	36.64	per Hour	0	50.83	0	0	0
354445	2017		COMMUNITY COLLEGE (HOSTOS	HEREDIA	RAFAEL		2012-01-31T	BRONX	ADJUNCT LEI	CEASED	85.12	per Day	0	75.84	0	0	0
354446	2017		COMMUNITY COLLEGE (HOSTOS	HERNANDEZ	CLAUDIA	F	2014-06-02T	BRONX	ASSISTANT T	ACTIVE	55837	per Annum	1822	62840.43	0	0	1000

Dataset screenshot - 1

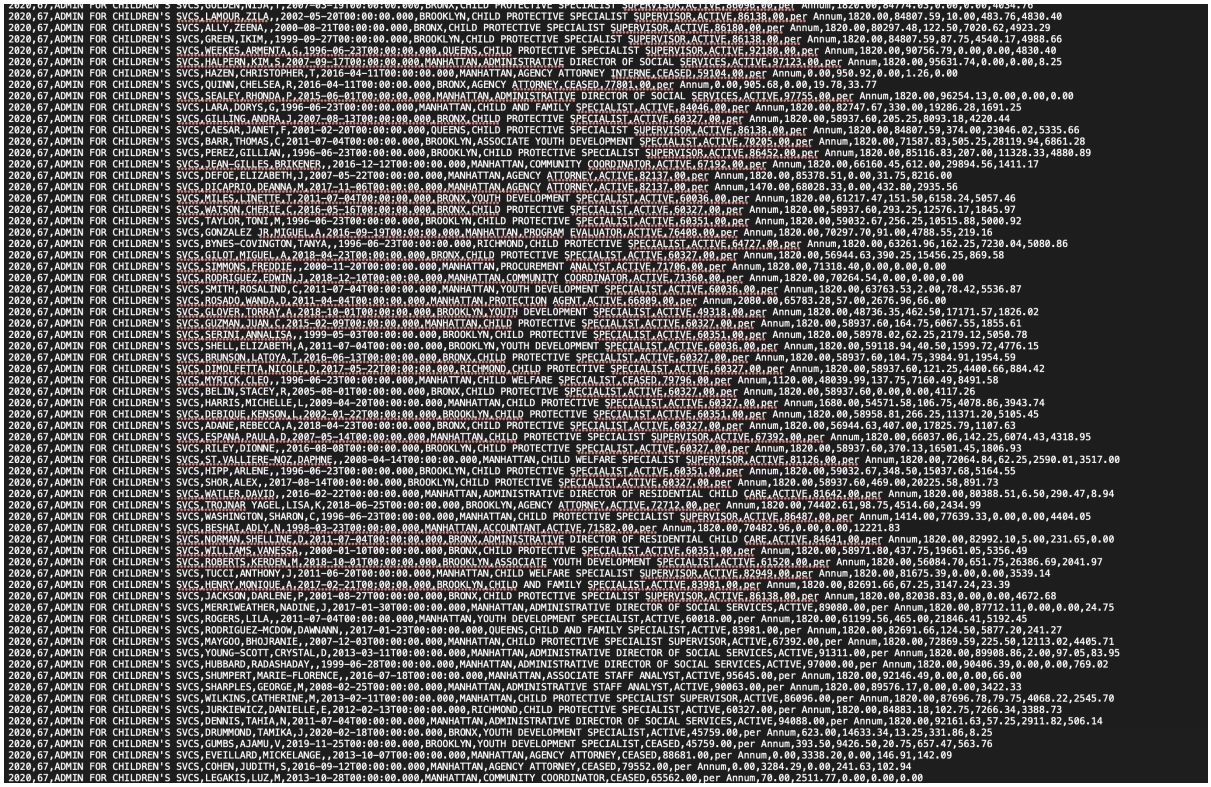
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Fiscal Year	Payroll Numl	Agency Name	Last Name	First Name	Mid Init	Agency Start	Work Location Borough	Title Descrip	Leave Status	Base Salary	Pay Basis	Regular Hou	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay
2	2020	67	ADMIN FOR CHILDREN'S SVCS	MAKHRINSK	IRINA		2011-10-03T	MANHATTAN	STAFF ANAL	ACTIVE	81509	per Annum	1820	79329.44	0	0	0
3	2020	67	ADMIN FOR CHILDREN'S SVCS	ACEVEDO	MIGUELINA	P	2006-02-27T	BRONX	CHILD PROTE	ACTIVE	86096	per Annum	1820	89599.83	142.25	9830.52	4501.92
4	2020	67	ADMIN FOR CHILDREN'S SVCS	ARGUETA	REGINA	A	2015-02-09T	BROOKLYN	CHILD PROTE	ACTIVE	60327	per Annum	1820	58937.6	336	14387.01	2027.07
5	2020	67	ADMIN FOR CHILDREN'S SVCS	SANTOS	WESLEY		2017-02-21T	MANHATTAN	PROGRAM E	ACTIVE	82987	per Annum	1820	79325.99	0	0	0
6	2020	67	ADMIN FOR CHILDREN'S SVCS	BURKE	KEVIN	C	2000-03-13T	MANHATTAN	ADMINISTRA	ACTIVE	90764	per Annum	1820	89370.34	0	0	795.64
7	2020	67	ADMIN FOR CHILDREN'S SVCS	SHAH	HETAL	M	2017-06-12T	BROOKLYN	AGENCY ATT	ACTIVE	91563	per Annum	1820	90972.86	147.25	9746.24	3196.65
8	2020	67	ADMIN FOR CHILDREN'S SVCS	DELROW	ANGELE		1997-09-22T	MANHATTAN	PRINCIPAL A	ACTIVE	68688	per Annum	1820	69826.26	15	927.94	1721.4
9	2020	67	ADMIN FOR CHILDREN'S SVCS	DAY	CATIANA		2006-02-27T	BROOKLYN	CHILD PROTE	ACTIVE	67392	per Annum	1820	66037.06	122.25	5172.59	4137.62
10	2020	67	ADMIN FOR CHILDREN'S SVCS	WILLIAMS	LINDA		1996-06-23T	QUEENS	CHILD WELF/	ACTIVE	72552	per Annum	1820	70352.8	6.75	277.81	4784.07
11	2020	67	ADMIN FOR CHILDREN'S SVCS	MARCIAL	INGRID		1996-06-23T	BRONX	CHILD PROTE	ACTIVE	60351	per Annum	1820	58990.37	351.25	15379.97	4913.41
12	2020	67	ADMIN FOR CHILDREN'S SVCS	ANGUS	DESMOND		2006-09-04T	MANHATTAN	ADMINISTRA	ACTIVE	85031	per Annum	1820	83725.26	0	0	803.89
13	2020	67	ADMIN FOR CHILDREN'S SVCS	HICKS	LILLIAN	L	1996-06-23T	BROOKLYN	CHILD PROTE	ACTIVE	86419	per Annum	1820	85084.25	0	0	5094.4
14	2020	67	ADMIN FOR CHILDREN'S SVCS	SPRINGER	ANDRE	R	1996-06-23T	MANHATTAN	CHILD PROTE	ACTIVE	92351	per Annum	1820	90933	0	0	4830.4
15	2020	67	ADMIN FOR CHILDREN'S SVCS	ABRAHAM	VAYOLA	K	2014-12-01T	MANHATTAN	AGENCY ATT	ACTIVE	91563	per Annum	1820	94887.78	74	4412.59	4583.81
16	2020	67	ADMIN FOR CHILDREN'S SVCS	MORALES-IS	JUANA		2011-07-04T	QUEENS	SUPERVISIN	ACTIVE	55853	per Annum	2080	55216.73	338.5	13875.27	2172.75
17	2020	67	ADMIN FOR CHILDREN'S SVCS	WHITE	QUANIQUA	L	2012-06-18T	BROOKLYN	YOUTH DEVE	ACTIVE	59995	per Annum	1750	58410.41	158	7351.05	6690.81
18	2020	67	ADMIN FOR CHILDREN'S SVCS	PINNICK	MONITE	N	2017-10-16T	MANHATTAN	CHILD PROTE	ACTIVE	60327	per Annum	1820	58929.41	289.5	12778.57	3626.93
19	2020	67	ADMIN FOR CHILDREN'S SVCS	ANTOINE	LUCY		2006-09-04T	MANHATTAN	ADMINISTRA	ACTIVE	80106	per Annum	1820	72346.34	0	0	2998.15
20	2020	67	ADMIN FOR CHILDREN'S SVCS	WOODS	ANNJEANETTE		2015-07-06T	BRONX	COMMUNITY	ACTIVE	65896	per Annum	1820	64876.12	0	0	1324.26
21	2020	67	ADMIN FOR CHILDREN'S SVCS	RODULFO	SHAVON	E	2017-02-13T	BRONX	COMMUNITY	ACTIVE	62215	per Annum	1820	61259.83	312	13804.06	313.5
22	2020	67	ADMIN FOR CHILDREN'S SVCS	SMITH-AGBA	JUNE	A	2005-08-01T	MANHATTAN	CHILD PROTE	ACTIVE	67392	per Annum	1820	62045.47	221	9472.76	4376.44
23	2020	67	ADMIN FOR CHILDREN'S SVCS	HERRERA	MANUEL		1996-06-23T	BRONX	CHILD PROTE	ACTIVE	64727	per Annum	1820	63261.96	258.25	11029.57	4968.41
24	2020	67	ADMIN FOR CHILDREN'S SVCS	OLIVER	DALE		1998-09-08T	MANHATTAN	ASSOCIATE S	ACTIVE	77083	per Annum	1820	76666.4	0	0	5437.45
25	2020	67	ADMIN FOR CHILDREN'S SVCS	VIDAL	SIGRY	E	2015-02-09T	BRONX	CHILD PROTE	ACTIVE	60327	per Annum	1820	58937.6	543.75	23504.42	2138.84
26	2020	67	ADMIN FOR CHILDREN'S SVCS	KINARD	TRAMAINE	T	2016-07-18T	MANHATTAN	MOTOR VEH	ACTIVE	49927	per Annum	2080	49103.96	1054.5	37436.55	1239.55
27	2020	67	ADMIN FOR CHILDREN'S SVCS	WALKER	SERENA		1996-06-23T	MANHATTAN	CHILD WELF/	ACTIVE	86452	per Annum	1820	85116.83	3	150	4844.15
28	2020	67	ADMIN FOR CHILDREN'S SVCS	SUAREZ	JOSEPHINE		2001-05-21T	MANHATTAN	DIRECTOR O	ACTIVE	95831	per Annum	1820	94375.02	0	0	0
29	2020	67	ADMIN FOR CHILDREN'S SVCS	HAMMONDS	MONIQUE	A	2017-02-13T	MANHATTAN	ADMINISTRA	ACTIVE	94191	per Annum	1820	92213.57	0	0	3455.34
30	2020	67	ADMIN FOR CHILDREN'S SVCS	LACEWELL	MARTIN		1997-01-06T	MANHATTAN	CONGREG C/	ACTIVE	63691	per Annum	1820	62705.3	540.25	25184.16	7884.17
31	2020	67	ADMIN FOR CHILDREN'S SVCS	DOUGLAS-AI	SHERILL	A	2004-01-26T	BROOKLYN	CHILD PROTE	ACTIVE	86096	per Annum	1820	84774.03	244	14085.24	5014.25
32	2020	67	ADMIN FOR CHILDREN'S SVCS	WILLIAMS	DAMITA	D	1996-06-23T	MANHATTAN	PRINCIPAL A	ACTIVE	65758	per Annum	1820	67042.68	0	0	4219.81
33	2020	67	ADMIN FOR CHILDREN'S SVCS	WILLIAMS-M	HELENA	C	2007-07-16T	MANHATTAN	CHILD PROTE	ACTIVE	86096	per Annum	1820	88842.54	136.75	8251.65	4227.38
34	2020	67	ADMIN FOR CHILDREN'S SVCS	SUBAIR	BASIRAT	F	2015-04-20T	MANHATTAN	STRATEGIC	ACTIVE	92700	per Annum	1820	91291.4	39.25	2173.67	117.13
35	2020	67	ADMIN FOR CHILDREN'S SVCS	SMITH	MAI	T	2018-06-25T	MANHATTAN	CONGREG C/	ACTIVE	42731	per Annum	1546	36027.04	1021.73	34038.5	2369.59
36	2020	67	ADMIN FOR CHILDREN'S SVCS	SINGH	LISA	B	2018-02-05T	QUEENS	CHILD PROTE	ACTIVE	60327	per Annum	1680	53434.94	472.5	20301.75	827.53
37	2020	67	ADMIN FOR CHILDREN'S SVCS	TAYLOR	JAMI	L	2006-07-24T	BRONX	CHILD PROTE	ACTIVE	60327	per Annum	1820	58937.6	279.5	11597.2	4795.12

Dataset screenshot - 2

Dataset before MapReduce

```
Fiscal Year,Payroll Number,Agency Name,Last Name,First Name,Mid Init,Agency Start Date,Work Location Borough,Title Description,Leave Status as of June 30,Base Salary,Pay Basis,Regular Hours,Regular Gross Paid,OT Hours,Total OT Paid,Total Other Pay
2020,67,ADMIN FOR CHILDREN'S SVCS,MAKHRINSKY,IRINA,,2011-10-03T00:00:00.000,MANHATTAN,STAFF ANALYST,ACTIVE,81509.00,per Annum,1820.00,79329.44,0.00,0.00,0.00
2020,67,ADMIN FOR CHILDREN'S SVCS,ACEVEDO,MIGUELINA,P,2006-02-27T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86096.00,per Annum,1820.00,89599.83,142.25,9830.52,4501.92
2020,67,ADMIN FOR CHILDREN'S SVCS,ARGUETA,REGINA,A,2015-02-09T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,336.00,14387.01,2027.07
2020,67,ADMIN FOR CHILDREN'S SVCS,SANTOS,WESLEY,,2017-02-21T00:00:00.000,MANHATTAN,PROGRAM EVALUATOR,ACTIVE,82987.00,per Annum,1820.00,79325.99,0.00,0.00,0.00
2020,67,ADMIN FOR CHILDREN'S SVCS,BURKE,KEVIN,C,2000-03-13T00:00:00.000,MANHATTAN,ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES,ACTIVE,90764.00,per Annum,1820.00,89370.34,0.00,0.00,795.64
2020,67,ADMIN FOR CHILDREN'S SVCS,SHAH,HETAL,M,2017-06-12T00:00:00.000,BROOKLYN,AGENCY ATTORNEY,ACTIVE,91563.00,per Annum,1820.00,90972.86,147.25,9746.24,3196.65
2020,67,ADMIN FOR CHILDREN'S SVCS,DELROW,ANGELE,,1997-09-22T00:00:00.000,MANHATTAN,PRINCIPAL ADMINISTRATIVE ASSOCIATE - NON SUPVR,ACTIVE,68688.00,per Annum,1820.00,69826.26,15.00,927.94,1721.40
2020,67,ADMIN FOR CHILDREN'S SVCS,DAY,CATIANA,,2006-02-27T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,67392.00,per Annum,1820.00,66037.06,122.25,5172.59,4137.62
2020,67,ADMIN FOR CHILDREN'S SVCS,WILLIAMS,LINDA,,1996-06-23T00:00:00.000,QUEENS,CHILD WELFARE SPECIALIST SUPERVISOR,ACTIVE,72552.00,per Annum,1820.00,70352.80,6.75,277.81,4784.07
2020,67,ADMIN FOR CHILDREN'S SVCS,MARCIAL,INGRID,,1996-06-23T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60351.00,per Annum,1820.00,58990.37,351.25,15379.97,4913.41
2020,67,ADMIN FOR CHILDREN'S SVCS,ANGUS,DESMOND,,2006-09-04T00:00:00.000,MANHATTAN,ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES,ACTIVE,85031.00,per Annum,1820.00,83725.26,0.00,0.00,803.89
2020,67,ADMIN FOR CHILDREN'S SVCS,HICKS,LILLIAN,L,1996-06-23T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86419.00,per Annum,1820.00,85084.25,0.00,0.00,5094.40
2020,67,ADMIN FOR CHILDREN'S SVCS,SPRINGER,ANDRE,R,1996-06-23T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,92351.00,per Annum,1820.00,90933.00,0.00,0.00,4830.40
2020,67,ADMIN FOR CHILDREN'S SVCS,ABRAHAM,VAYOLA,K,2014-12-01T00:00:00.000,MANHATTAN,AGENCY ATTORNEY,ACTIVE,91563.00,per Annum,1820.00,94887.78,74.00,4412.59,4583.81
2020,67,ADMIN FOR CHILDREN'S SVCS,MORALES-ISLEY,JUANA,,2011-07-04T00:00:00.000,QUEENS,SUPERVISING SPECIAL OFFICER,ACTIVE,55853.00,per Annum,2080.00,55216.73,338.50,13875.27,2172.75
2020,67,ADMIN FOR CHILDREN'S SVCS,WHITE,QUANIQUA,L,2012-06-18T00:00:00.000,BROOKLYN,YOUTH DEVELOPMENT SPECIALIST,ACTIVE,59995.00,per Annum,1750.00,58410.41,158.00,7351.05,6690.81
2020,67,ADMIN FOR CHILDREN'S SVCS,PINNICK,MONITE,N,2017-10-16T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58929.41,289.50,12778.57,3626.93
2020,67,ADMIN FOR CHILDREN'S SVCS,ANTOINE,LUCY,,2006-09-04T00:00:00.000,MANHATTAN,ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES,ACTIVE,80106.00,per Annum,1820.00,72346.34,0.00,0.00,2998.15
2020,67,ADMIN FOR CHILDREN'S SVCS,WOODS,ANNJEANETTE,,2015-07-06T00:00:00.000,BRONX,COMMUNITY COORDINATOR,ACTIVE,65896.00,per Annum,1820.00,64876.12,0.00,0.00,1324.26
2020,67,ADMIN FOR CHILDREN'S SVCS,RODULFO,SHAVON,E,2017-02-13T00:00:00.000,BRONX,COMMUNITY COORDINATOR,ACTIVE,62215.00,per Annum,1820.00,61259.83,312.00,13804.06,313.50
2020,67,ADMIN FOR CHILDREN'S SVCS,SMITH-AGBAYOR,JUNE,A,2005-08-01T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,67392.00,per Annum,1820.00,62045.47,221.00,9472.76,4376.44
2020,67,ADMIN FOR CHILDREN'S SVCS,HERRERA,MANUEL,,1996-06-23T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,64727.00,per Annum,1820.00,63261.96,258.25,11029.57,4968.41
2020,67,ADMIN FOR CHILDREN'S SVCS,OLIVER,DALE,,1998-09-08T00:00:00.000,MANHATTAN,ASSOCIATE STAFF ANALYST,ACTIVE,77083.00,per Annum,1820.00,76666.40,0.00,0.00,5437.45
2020,67,ADMIN FOR CHILDREN'S SVCS,VIDAL,SIGRY,E,2015-02-09T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,543.75,23504.42,2138.84
2020,67,ADMIN FOR CHILDREN'S SVCS,KINARD,TRAMAINE,T,2016-07-18T00:00:00.000,MANHATTAN,MOTOR VEHICLE OPERATOR,ACTIVE,49927.00,per Annum,2080.00,49103.96,1054.50,37436.55,1239.55
2020,67,ADMIN FOR CHILDREN'S SVCS,WALKER,SERENA,,1996-06-23T00:00:00.000,MANHATTAN,CHILD WELFARE SPECIALIST SUPERVISOR,ACTIVE,86452.00,per Annum,1820.00,85116.83,3.00,150.00,4844.15
2020,67,ADMIN FOR CHILDREN'S SVCS,SUAREZ,JOSEPHINE,,2001-05-21T00:00:00.000,MANHATTAN,DIRECTOR OF FIELD OPERATIONS,ACTIVE,95831.00,per Annum,1820.00,94375.02,0.00,0.00,0.00
2020,67,ADMIN FOR CHILDREN'S SVCS,HAMMONDS,MONIQUE,A,2017-02-13T00:00:00.000,MANHATTAN,ADMINISTRATIVE STAFF ANALYST,ACTIVE,94191.00,per Annum,1820.00,92213.57,0.00,0.00,3455.34
2020,67,ADMIN FOR CHILDREN'S SVCS,LACEWELL,MARTIN,,1997-01-06T00:00:00.000,MANHATTAN,CONGREG CARE SPEC - ACS,ACTIVE,63691.00,per Annum,1820.00,62705.30,540.25,25184.16,7884.17
2020,67,ADMIN FOR CHILDREN'S SVCS,DOUGLAS-ALEXIS,SHERILL,A,2004-01-26T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86096.00,per Annum,1820.00,84774.03,244.00,14085.24,5014.25
2020,67,ADMIN FOR CHILDREN'S SVCS,WILLIAMS,DAMITA,D,1996-06-23T00:00:00.000,MANHATTAN,PRINCIPAL ADMINISTRATIVE ASSOCIATE - NON SUPVR,ACTIVE,65758.00,per Annum,1820.00,67042.68,0.00,0.00,4219.81
2020,67,ADMIN FOR CHILDREN'S SVCS,WILLIAMS-MURPHY,HELENA,C,2007-07-16T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86096.00,per Annum,1820.00,88842.54,136.75,8251.65,4227.38
2020,67,ADMIN FOR CHILDREN'S SVCS,SUBAIR,BASIRAT,F,2015-04-20T00:00:00.000,MANHATTAN,STRATEGIC INITIATIVE SPECIALIST,ACTIVE,92700.00,per Annum,1820.00,91291.40,39.25,2173.67,117.13
2020,67,ADMIN FOR CHILDREN'S SVCS,SMITH,MAI,T,2018-06-25T00:00:00.000,MANHATTAN,CONGREG CARE SPEC - ACS,ACTIVE,42731.00,per Annum,1546.00,36027.04,1021.73,34038.50,2369.59
2020,67,ADMIN FOR CHILDREN'S SVCS,SINGH,LISA,B,2018-02-05T00:00:00.000,QUEENS,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1680.00,53434.94,472.50,20301.75,827.53
2020,67,ADMIN FOR CHILDREN'S SVCS,TAYLOR,JAMI,L,2006-07-24T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,279.50,11597.20,4795.12
2020,67,ADMIN FOR CHILDREN'S SVCS,WALTERS,JULTET,A,2004-09-20T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST,ON SEPARATION LEAVE,60327.00,per Annum,1820.00,58937.60,132.17,5976.98,4441.24
2020,67,ADMIN FOR CHILDREN'S SVCS,WILLIAMS,LEONIE,,2018-02-05T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1960.00,61980.10,169.25,6583.42,797.33
2020,67,ADMIN FOR CHILDREN'S SVCS,NELSON,TAMEIKA,L,2014-04-28T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,164.50,6046.62,2748.90
2020,67,ADMIN FOR CHILDREN'S SVCS,DANIELS,KRISTAL,A,2014-10-20T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,67392.00,per Annum,1820.00,59855.60,107.50,3655.33,2691.63
2020,67,ADMIN FOR CHILDREN'S SVCS,TWENOFU,EUNICE,U,1996-06-23T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST,CEASED,60351.00,per Annum,1575.00,50945.95,183.75,7609.03,5132.38
2020,67,ADMIN FOR CHILDREN'S SVCS,MARCUS,LENORE,E,2017-02-13T00:00:00.000,MANHATTAN,COMMUNITY COORDINATOR,ACTIVE,64539.00,per Annum,1820.00,63547.91,0.00,0.00,132.00
2020,67,ADMIN FOR CHILDREN'S SVCS,VARGAS,VIRGIN,M,2016-05-16T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,335.50,14122.49,2359.85
2020,67,ADMIN FOR CHILDREN'S SVCS,ERNEST,MARSHA,,2002-04-29T00:00:00.000,QUEENS,CHILD PROTECTIVE SPECIALIST,ACTIVE,60351.00,per Annum,1820.00,58958.81,271.50,11813.35,5109.79
2020,67,ADMIN FOR CHILDREN'S SVCS,HODGE,CELIA,C,2005-08-22T00:00:00.000,MANHATTAN,COMMUNITY COORDINATOR,ACTIVE,77915.00,per Annum,1820.00,76718.89,0.00,0.00,528.62
2020,67,ADMIN FOR CHILDREN'S SVCS,GALLARDO,CYNTHIA,L,2008-01-14T00:00:00.000,QUEENS,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,67392.00,per Annum,1820.00,66357.31,199.75,8765.57,4130.28
2020,67,ADMIN FOR CHILDREN'S SVCS,KHAN,BILAL,,2016-06-13T00:00:00.000,MANHATTAN,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,329.00,16983.83,4810.00
2020,67,ADMIN FOR CHILDREN'S SVCS,MOE,HEMSLEY,O,2001-04-30T00:00:00.000,MANHATTAN,ASSOCIATE STAFF ANALYST,ACTIVE,77083.00,per Annum,1820.00,76666.40,0.00,0.00,5437.45
2020,67,ADMIN FOR CHILDREN'S SVCS,SKEEN,MARSHA,M,2008-05-19T00:00:00.000,BROOKLYN,CHILD AND FAMILY SPECIALIST,ACTIVE,83981.00,per Annum,1820.00,82691.66,0.00,0.00,528.62
2020,67,ADMIN FOR CHILDREN'S SVCS,HARDY,TANISHA,M,2000-09-18T00:00:00.000,MANHATTAN,DIRECTOR OF FIELD OPERATIONS,ACTIVE,86485.00,per Annum,1820.00,83677.55,0.00,0.00,812.86
2020,67,ADMIN FOR CHILDREN'S SVCS,MINUS,SHARLENE,M,2007-01-07T00:00:00.000,MANHATTAN,CHILD WELFARE SPECIALIST SUPERVISOR,ACTIVE,84967.00,per Annum,1820.00,83654.42,0.00,0.00,2141.14
2020,67,ADMIN FOR CHILDREN'S SVCS,LEWIS,DOTTIA,,1998-04-13T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60351.00,per Annum,1820.00,58982.62,470.50,22247.85,6553.44
2020,67,ADMIN FOR CHILDREN'S SVCS,GOLDEN,NIJA,T,2007-03-19T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86096.00,per Annum,1820.00,84774.03,0.00,0.00,4034.76
2020,67,ADMIN FOR CHILDREN'S SVCS,LAMOUR,ZILA,,2002-05-20T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86138.00,per Annum,1820.00,84807.59,10.00,483.76,4830.40
2020,67,ADMIN FOR CHILDREN'S SVCS,ALLY,ZEENA,,2000-08-21T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86180.00,per Annum,1820.00,80297.48,122.50,7020.62,4923.29
2020,67,ADMIN FOR CHILDREN'S SVCS,GREEN,IKIM,,1999-09-27T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86138.00,per Annum,1820.00,84807.59,87.75,4540.17,4988.66
2020,67,ADMIN FOR CHILDREN'S SVCS,MEEKES,ARMENTA,G,1996-06-23T00:00:00.000,QUEENS,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,92180.00,per Annum,1820.00,90756.79,0.00,0.00,4830.40
2020,67,ADMIN FOR CHILDREN'S SVCS,SALPERN,KIM,S,2007-03-19T00:00:00.000,MANHATTAN,ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES,ACTIVE,97123.00,per Annum,1820.00,95621.74,0.00,0.00,8.25
2020,67,ADMIN FOR CHILDREN'S SVCS,HAZEN,CHRISTOPHER,T,2016-04-11T00:00:00.000,MANHATTAN,AGENCY ATTORNEY INTERNE,CEASED,59104.00,per Annum,0.00,950.92,0.00,1.26,0.00
2020,67,ADMIN FOR CHILDREN'S SVCS,QUINN,CHELSEA,R,2016-04-11T00:00:00.000,BRONX,AGENCY ATTORNEY,CEASED,77801.00,per Annum,0.00,905.68,0.00,19.78,33.77
2020,67,ADMIN FOR CHILDREN'S SVCS,SEALEY,RHONDA,P,2015-06-01T00:00:00.000,MANHATTAN,ADMINISTRATIVE DIRECTOR OF SOCIAL SERVICES,ACTIVE,97755.00,per Annum,1820.00,96254.13,0.00,0.00,0.00
2020,67,ADMIN FOR CHILDREN'S SVCS,LARA,DORYS,G,1996-06-23T00:00:00.000,MANHATTAN,CHILD AND FAMILY SPECIALIST,ACTIVE,84046.00,per Annum,1820.00,82747.67,330.00,19286.28,1691.25
2020,67,ADMIN FOR CHILDREN'S SVCS,GILLING,ANDRA,J,2007-08-13T00:00:00.000,BRONX,CHILD PROTECTIVE SPECIALIST,ACTIVE,60327.00,per Annum,1820.00,58937.60,205.25,8093.18,4220.44
2020,67,ADMIN FOR CHILDREN'S SVCS,CAESAR,JANET,F,2001-02-20T00:00:00.000,QUEENS,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86138.00,per Annum,1820.00,84807.59,374.00,23046.02,5335.66
2020,67,ADMIN FOR CHILDREN'S SVCS,BARR,THOMAS,C,2011-07-04T00:00:00.000,BROOKLYN,ASSOCIATE YOUTH DEVELOPMENT SPECIALIST,ACTIVE,70205.00,per Annum,1820.00,71587.83,505.25,28119.94,6861.28
2020,67,ADMIN FOR CHILDREN'S SVCS,PEREZ,GILLIAN,,1996-06-23T00:00:00.000,BROOKLYN,CHILD PROTECTIVE SPECIALIST SUPERVISOR,ACTIVE,86452.00,per Annum,1820.00,85116.83,207.00,11328.33,4880.89
2020,67,ADMIN FOR CHILDREN'S SVCS,JEAN-GILLES,BRIKENER,,2016-12-12T00:00:00.000,MANHATTAN,COMMUNITY COORDINATOR,ACTIVE,67192.00,per Annum,1820.00,66168.45,612.00,29894.56,1411.17
```

Raw data screenshot - 1



Raw data screenshot - 2

Dataset after MapReduce

```
BRONX,2015,POLICE DEPARTMENT,357.77,22147.74,108875.49
BRONX,2018,POLICE DEPARTMENT,59.13,2109.36,55327.06
BRONX,2016,POLICE DEPARTMENT,41.25,2817.61,61492.44
BRONX,2017,POLICE DEPARTMENT,242.25,22182.34,124892.76999999999
BRONX,2017,POLICE DEPARTMENT,88.17,3499.45,56625.009999999995
BRONX,2017,POLICE DEPARTMENT,195.08,18961.07,145875.09
BRONX,2016,POLICE DEPARTMENT,384.83,21720.32,114479.08
BRONX,2020,POLICE DEPARTMENT,240.67,8658.92,111157.17
BRONX,2019,POLICE DEPARTMENT,30.30,2278.18,108282.81999999999
BRONX,2015,POLICE DEPARTMENT,0,0.00,24342.350000000002
BRONX,2016,POLICE DEPARTMENT,549,50522.10,174474.04
BRONX,2017,POLICE DEPARTMENT,142.75,5432.70,55840.32
BRONX,2018,POLICE DEPARTMENT,41,2123.08,64925.6
BRONX,2015,POLICE DEPARTMENT,308,9279.39,49619.11
BRONX,2017,POLICE DEPARTMENT,20,805.93,56944.47
BRONX,2019,POLICE DEPARTMENT,42.33,3024.28,103764.65
BRONX,2020,POLICE DEPARTMENT,122.00,3109.95,53304.659999999996
BRONX,2017,POLICE DEPARTMENT,224.48,11478.42,70140.38
BRONX,2018,POLICE DEPARTMENT,29.5,2126.31,100009.95999999999
BRONX,2017,POLICE DEPARTMENT,95.45,5113.75,67561.41
BRONX,2016,POLICE DEPARTMENT,410,25508.29,126601.23000000001
BRONX,2017,POLICE DEPARTMENT,30,3901.00,110196.22
BRONX,2017,POLICE DEPARTMENT,201.83,25004.83,131555.32
BRONX,2016,POLICE DEPARTMENT,532.78,39022.50,136057.22
BRONX,2015,POLICE DEPARTMENT,400.18,23749.15,118730.98
BRONX,2019,POLICE DEPARTMENT,33.42,1346.43,68592.81
BRONX,2015,POLICE DEPARTMENT,347.67,20455.85,113260.70000000001
BRONX,2015,POLICE DEPARTMENT,565.25,21371.87,78423.33
BRONX,2020,POLICE DEPARTMENT,520.22,29395.57,139531.83000000002
BRONX,2019,POLICE DEPARTMENT,0.00,0.00,188989.5
BRONX,2018,POLICE DEPARTMENT,28.25,2140.12,110443.66
BRONX,2017,POLICE DEPARTMENT,293.75,7298.56,48346.42
BRONX,2020,POLICE DEPARTMENT,355.25,10910.68,54273.32
BRONX,2019,POLICE DEPARTMENT,237.42,13751.01,115215.64
BRONX,2019,POLICE DEPARTMENT,0.00,0.00,25938.0
BRONX,2019,POLICE DEPARTMENT,246.10,9949.98,67661.76
BRONX,2020,POLICE DEPARTMENT,614.92,26872.99,125298.71
BRONX,2016,POLICE DEPARTMENT,49.25,1265.73,42275.770000000004
BRONX,2018,POLICE DEPARTMENT,47.67,2145.93,70392.64
BRONX,2016,POLICE DEPARTMENT,72.68,1861.30,47035.82
BRONX,2018,POLICE DEPARTMENT,61.22,1983.19,48660.5
BRONX,2019,POLICE DEPARTMENT,213.75,15592.27,119816.04000000001
BRONX,2015,POLICE DEPARTMENT,345.68,21272.80,121778.52
BRONX,2016,POLICE DEPARTMENT,511.42,36732.95,140203.75
BRONX,2019,POLICE DEPARTMENT,464.25,15867.81,63988.49
BRONX,2019,POLICE DEPARTMENT,50.58,5271.81,186014.44
BRONX,2017,POLICE DEPARTMENT,0,0.00,44178.89
BRONX,2019,POLICE DEPARTMENT,8.42,279.57,44408.159999999996
BRONX,2017,POLICE DEPARTMENT,113.75,4556.84,66972.93
BRONX,2020,POLICE DEPARTMENT,268.25,15358.02,118756.59
BRONX,2019,POLICE DEPARTMENT,233.82,17715.25,119467.95999999999
BRONX,2015,POLICE DEPARTMENT,146.92,9582.60,100893.90000000001
BRONX,2020,POLICE DEPARTMENT,140.40,8781.16,108618.27
BRONX,2018,POLICE DEPARTMENT,27.57,2156.79,110671.79999999999
BRONX,2016,POLICE DEPARTMENT,379.38,27216.78,130983.34
BRONX,2017,POLICE DEPARTMENT,298.87,25334.74,129738.35
BRONX,2018,POLICE DEPARTMENT,60.17,2158.30,55708.36
BRONX,2016,POLICE DEPARTMENT,387.42,30279.88,151772.68
BRONX,2015,POLICE DEPARTMENT,647.15,43505.18,142977.75
BRONX,2020,POLICE DEPARTMENT,411.42,22875.56,139479.55
BRONX,2019,POLICE DEPARTMENT,66.75,5664.00,107116.33
BRONX,2015,POLICE DEPARTMENT,214.5,6114.69,56636.05
BRONX,2017,POLICE DEPARTMENT,58.8,4750.65,106889.15
BRONX,2020,POLICE DEPARTMENT,352.58,20591.34,125301.26999999999
BRONX,2018,POLICE DEPARTMENT,46.17,2168.74,74345.05
BRONX,2017,POLICE DEPARTMENT,478.97,42657.25,163718.01
BRONX,2020,POLICE DEPARTMENT,249.50,5775.95,52991.45
BRONX,2017,POLICE DEPARTMENT,561.17,49712.68,159316.88
BRONX,2020,POLICE DEPARTMENT,254.50,9436.53,61924.06
```

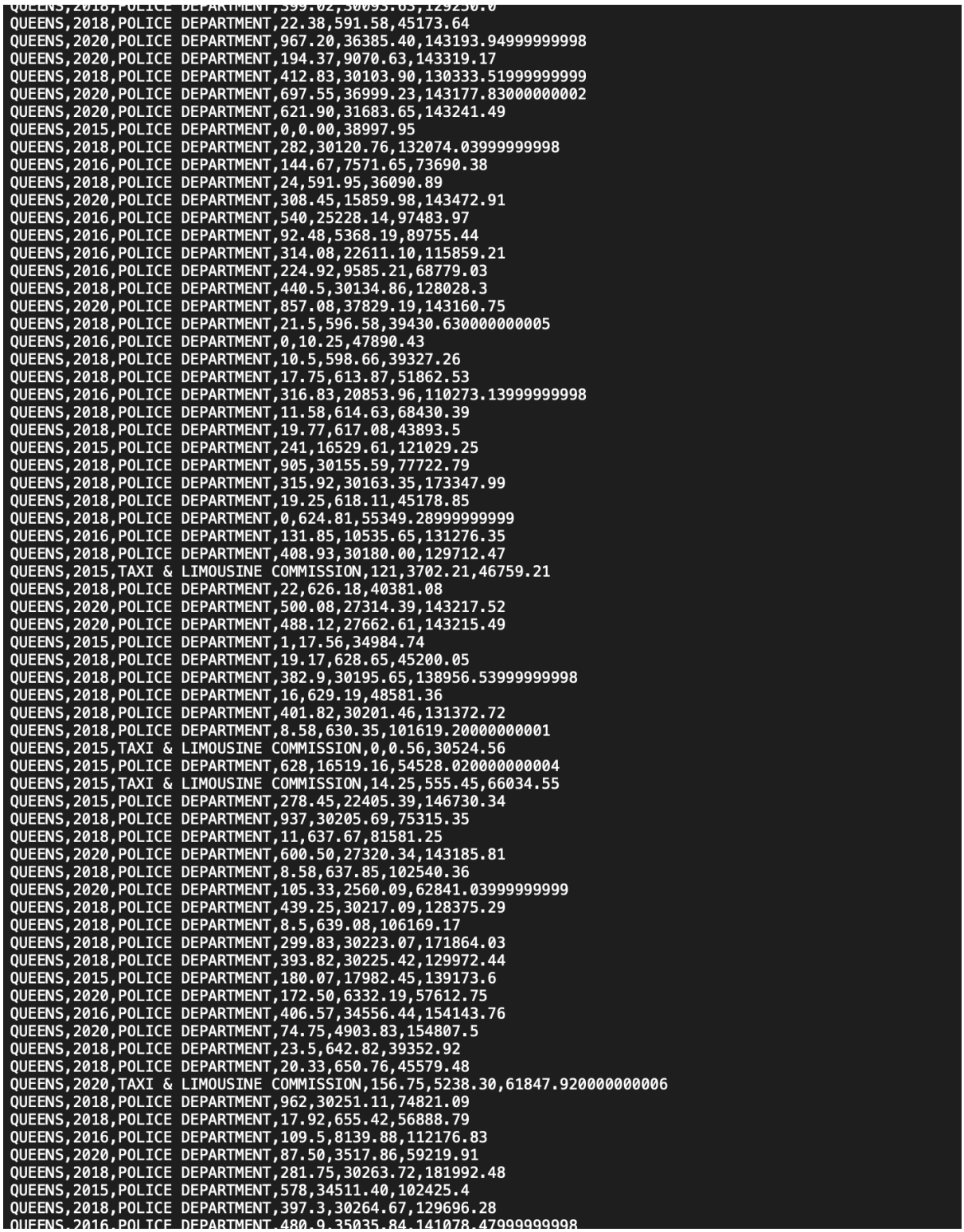
Preprocessed data screenshot - 1

```
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,71,2802.47,72512.23
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,1.25,40.41,58596.020000000004
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,112095.39
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,8,499.38,96697.26000000001
BROOKLYN,2017,DEPT OF ENVIRONMENT PROTECTION,404.25,26948.00,96879.46
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,198.5,7306.53,71015.27
BROOKLYN,2019,NYC HOUSING AUTHORITY,128.00,3864.66,58988.53999999999
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,122,2295.39,33165.12
BROOKLYN,2019,DEPT OF HEALTH/MENTAL HYGIENE,0.00,0.00,66064.56
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,536.75,15838.90,58988.740000000005
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,4.88,66653.68
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,19,406.43,40825.68
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,67,2470.36,50178.84
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,43176.41
BROOKLYN,2019,DEPT OF HEALTH/MENTAL HYGIENE,0.00,0.00,66681.88
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,59284.21
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,9,173.11,33152.560000000005
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,29.75,1776.75,98949.16
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,29.5,875.32,43432.58
BROOKLYN,2020,DEPT. OF HOMELESS SERVICES,5.75,191.09,62415.17999999999
BROOKLYN,2019,NYC HOUSING AUTHORITY,0.00,0.00,35483.26
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,42499.92
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,113.25,5739.79,68685.54000000001
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,8.5,276.17,62429.659999999996
BROOKLYN,2016,NYC HOUSING AUTHORITY,5,81.91,31106.91
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,50.61,42586.03
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,63345.0
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,175.5,5261.59,49973.829999999994
BROOKLYN,2019,NYC HOUSING AUTHORITY,243.25,6099.55,43151.880000000005
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,317.25,10417.00,55938.88
BROOKLYN,2019,NYC HOUSING AUTHORITY,116.75,3004.65,35979.5
BROOKLYN,2019,NYC HOUSING AUTHORITY,279.25,8997.43,55421.4
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,38.75,1446.17,53360.82
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,7,243.69,53061.92
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,23.5,850.65,57483.11
BROOKLYN,2019,DEPT OF HEALTH/MENTAL HYGIENE,0.00,0.00,66151.56
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,359.5,11071.48,57734.03
BROOKLYN,2019,NYC HOUSING AUTHORITY,222.75,5633.09,50256.14
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,48995.17
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,5,187.94,69148.04000000001
BROOKLYN,2017,DEPT OF ENVIRONMENT PROTECTION,512.5,34549.56,129859.55999999998
BROOKLYN,2019,NYC HOUSING AUTHORITY,124.00,3429.18,40452.53
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,342,24105.00,135979.38
BROOKLYN,2019,NYC HOUSING AUTHORITY,165.25,3925.06,41114.52
BROOKLYN,2019,DEPT OF HEALTH/MENTAL HYGIENE,0.00,1.88,66153.54
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,49070.17
BROOKLYN,2019,DEPT OF HEALTH/MENTAL HYGIENE,0.00,0.00,66183.79
BROOKLYN,2019,NYC HOUSING AUTHORITY,150.00,3705.46,40931.549999999996
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,248,11483.75,71835.85
BROOKLYN,2016,NYC HOUSING AUTHORITY,183.75,7857.76,64237.780000000006
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,22.75,1061.86,47319.44
BROOKLYN,2019,NYC HOUSING AUTHORITY,120.00,3199.77,40558.38
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,141,6973.24,83282.12000000001
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,858.75,34472.66,87898.72
BROOKLYN,2019,NYC HOUSING AUTHORITY,133.50,3679.37,40728.54
BROOKLYN,2019,NYC HOUSING AUTHORITY,181.50,4547.58,36622.72
BROOKLYN,2019,NYC HOUSING AUTHORITY,0.00,0.00,41408.58
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,40411.0
BROOKLYN,2020,DEPT. OF HOMELESS SERVICES,716.25,19235.10,61772.65
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,9.25,275.81,53127.78999999999
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,1,25.18,58309.83
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,7.93,56858.78
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,58.5,2039.04,54807.770000000004
BROOKLYN,2019,NYC HOUSING AUTHORITY,172.00,4154.48,35839.829999999994
BROOKLYN,2020,DEPT. OF HOMELESS SERVICES,0.00,0.00,62215.0
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,0,0.00,43639.770000000004
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,125.75,4717.61,61428.23
BROOKLYN,2017,HRA/DEPT OF SOCIAL SERVICES,399.5,15373.05,74903.91
BROOKLYN,2019,NYC HOUSING AUTHORITY,265.25,6833.54,44327.32
```

Preprocessed data screenshot - 2

```
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36823.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,677.99,28294.99
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33329.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22818.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,26474.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,35915.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,34540.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,26312.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27684.71
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,26312.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27196.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,135.70,32321.7
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,29352.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,1499.24,34828.24
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,32186.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,38966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22818.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36300.52
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22818.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27980.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36298.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,35915.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33329.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36823.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,26279.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27617.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,31661.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,2.68,26839.68
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22149.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,2.40,37968.4
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,26279.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27617.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22818.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,35915.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27684.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33329.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33329.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36298.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,27980.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33329.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,87.15,27775.050000000003
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.67,26312.67
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,22818.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,34247.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,33371.450000000004
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,37966.0
MANHATTAN,2015,DEPT OF ED PARA PROFESSIONALS,0,0.00,36298.0
```

Preprocessed data screenshot - 3



Preprocessed data screenshot - 4

Shell commands

Running commands:

```
# Make the project directory
hadoop fs -mkdir project

# Setting hadoop classpath and creating the jar
javac -classpath `hadoop classpath` PayrollFiscalMapper.java
javac -classpath `hadoop classpath` PayrollFiscalReducer.java
javac -classpath `hadoop classpath`:. PayrollFiscal.java
jar cvf PayrollFiscal.jar *.class

# Putting the data file onto hadoop
hadoop fs -put citywide-payroll-data-fiscal-year.csv project/

# Command to run the task
hadoop jar PayrollFiscal.jar PayrollFiscal project/citywide-payroll-data-fiscal-year.c
sv project/output7

# Getting the output to dataproc
hadoop fs -get project/output7
```

Other shell commands:

```
# Log in to Google cloud
gcloud compute ssh nyu-dataproc-m --project hpc-dataproc-19b8 --zone us-central1-f

# scp command to push data from local
gcloud compute scp --recurse /Users/rachitjain/Desktop/Courses/Sem3/RBDA/Assignments/P
roject/nyc-crime-arrest-payroll rj2219_nyu_edu@nyu-dataproc-m:project/ --project hpc-d
ataproc-19b8 --zone us-central1-f

# scp command to get the output to local
gcloud compute scp --recurse rj2219_nyu_edu@nyu-dataproc-m:project/nyc-crime-arrest-pa
yroll/output7/ /Users/rachitjain/Desktop/Courses/Sem3/RBDA/Assignments/Project/nyc-cri
me-arrest-payroll/results/ --project hpc-dataproc-19b8 --zone us-central1-f

# Extra commands
hadoop fs -ls project/
hadoop fs -ls project/output7
hadoop fs -cat project/output7/part-r-00000

# Count how many records left after preprocessing
wc -l part-r-00000
```

Team Members

[1] Jahnvi Arya (ja4158)
[2] Rachit Jain (rj2219)
[3] Shobhit Sinha (ss13881)

References

[1] York, C. of N. (2021, January 1). *NY citywide payroll data (fiscal year)*. Kaggle. Retrieved November 27, 2022, from https://www.kaggle.com/datasets/new-york-city/ny-citywide-payroll-data-fiscal-year

[2] https://github.com/rachitjain2706/nyc-crime-arrest-payroll

[3] https://www.indeed.com/career-advice/career-development/work-weeks-in-year

[4] https://www.ny.gov/new-york-states-minimum-wage/new-york-states-minimum-wage#:~:text=The Minimum Wage Act