2020-Summer-CSE-5331-002-Software Testing
Prof. Sharma Chakravarthy
Project 2 - Team 11
Jay Denton, ID - 1001559386

I. Overall Status (90% completed)

1. **Task 1**: Compute the average salary of males and females for each state for these 5 years. Compare and analyze them.

Codes, output files and analysis excel file are in the directory /CalcTask

Completeness: 100%

2. **Task 2**: Equiwidth Histogram on Age: use the buckets 0 to 9,10 to 19, …., 90 to 99 for each year of data given and compare them.

Codes, output files and analysis excel file are in the directory /EquiWidth

Completeness: 100%

3. **Performance Analysis**
I can run multiple producers, but I'm having a problem setting up Comet with multiple nodes. Thus this part incomplete.

Completeness: 40%

4. **Analysis**

The analysis has been applied to the both tasks. And will be in this report.

Completeness: 100%

II. Performance Measure

For each task sent on Comet, it can be shown with the following command:
```
$ squeue -u [user]
```
and the info of the task ran by the user can be shown.
With the command:
```
$ squeue -ij <job_list>
```
Can display the information for specified job(s).

In hadoop, the org.apache.hadoop.util.Timer.monotonicNowNanos can be used to track the task start and finish time, thus the performance can be recorded accordingly.

However, I had a hard time setting up Comet multiple M/P nodes, and didn't quite complete this part.

III. File Descriptions

Task 1: Compute the average salary of males and females for each state for these 5 years.

In the folder of /CalcTask, the "CalcAvg.java" implements the mapreduce for hadoop to run. The code is written in Java. The class CalcAvg holds everything. Within it, I have implements my custom org.apache.hadoop.io.Writable function. I extended org.apache.hadoop.mapreduce. Mapper, org.apache.hadoop.mapreduce.Reducer, and Partitioner.

The classes are separated in the /CalcTask/classes folder and and jar file is at /CalcTask.

This file calculates the average salary for different gender partitioned by the states. Regarding the null and zero value of wage.

The input pair for Mapper is <LongWritable, Text>, and then takes <Text gender, Custom Writable <Wage, Count, State> Tuple > as output.

The input pair for Reducer is <Text, Costumed Writable Tuple> and output pair is the same.

The Partitioner separate the reduce tasks by state value.

The final output files are located at /CalcTask/output-distr.

Task 2: Equiwidth Histogram on Age: use the buckets 0 to 9,10 to 19, …., 90 to 99 for each year.

In the folder of /EquiWidth, the "Histogram.java" implements the mapreduce for hadoop to run. The code is written in Java. The class Histogram holds everything. Within it, I have implements my custom Mapper, Reducer and Partitioner.

The classes are separated in the /EquiWidth/classes folder and and jar file is at /EquiWidth.

This file calculates the average salary for different gender partitioned by the equivalent bucket width regarding to ages.

The input pair for Mapper is <LongWritable, Text>, and then takes <Text year, Custom Writable People Count Tuple > as output.

The input pair for Reducer is <Text, Writable People Count Tuple> and output pair is the same.

The Partitioner separate the reduce tasks by age group value.

The final output files are located at /EquiWidth/output-distr.

IV. Division of Labor

Task 1: Jay.
Task 2: Jay.
Hadoop set up on Commet and PC: Jay.

V. Logical Errors and Solutions

1. Hadoop setup on Comet. The JAVA_HOME variable exported in the ~/.bashrc does not compile as expected when run the hadoop build and dist.run files.

Solution: Add the variable directly to the hadoop setup files.

2. Hadoop setup on PC. When I stop the hadoop by stop-all.sh, and try to back up by start-all.sh, the datanode does not back up, shown in jps.

Solution: Since the duplicate of datanode is set to 1, which meaning, there cannot exist a second datanode, thus have to delete /mydata/hdfs/datanode/current. Then datanode will restart.

3. When override the Mapper, Reducer and other functions, I tried to code the way I thought should work, but it wouldn't.

Solution: By running javac to compile the finished java file for mapreduce, there are always some notes shown as: "use -Xlint:deprecation for details". And I add the linker to the end of the compilation command, which gives the warning string.

Most of the time, it is because the data structure is not legit with the hadoop version of what I use, which is same with comet, Hadoop-2.6.0. Thus, I have to look up on the API documentation for that specific version, instead of other versions.

4. At first, the java file can be compiled with no error, and the hadoop can compile the file as well, but I'm getting no out put values or anything.

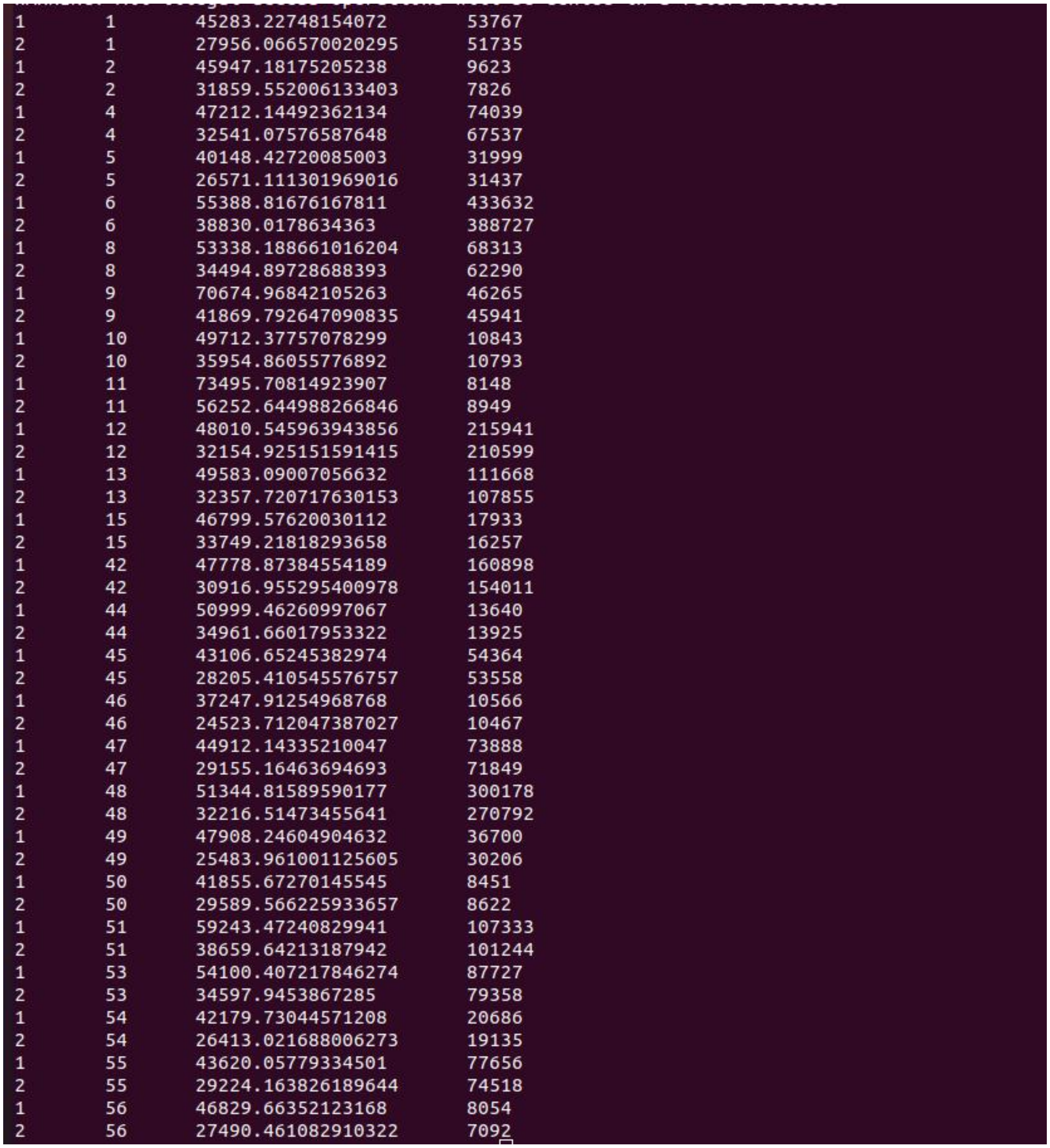
Solution: Firstly, because I forgot to parse the correct output values from Mapper class. Since there is no values to output, then the Reducer will have null or empty values as input, thus cannot generate final results. The solution is to context.write correct output pair.

Secondly, not only the input and output pairs have to be in the right format, but also at the main method, the format of job.setOutputKeyClass and job.setOutputValueClass has to be the correct classes, otherwise it still causes empty output. I figured that out.

VI. Analysis

Task 1:

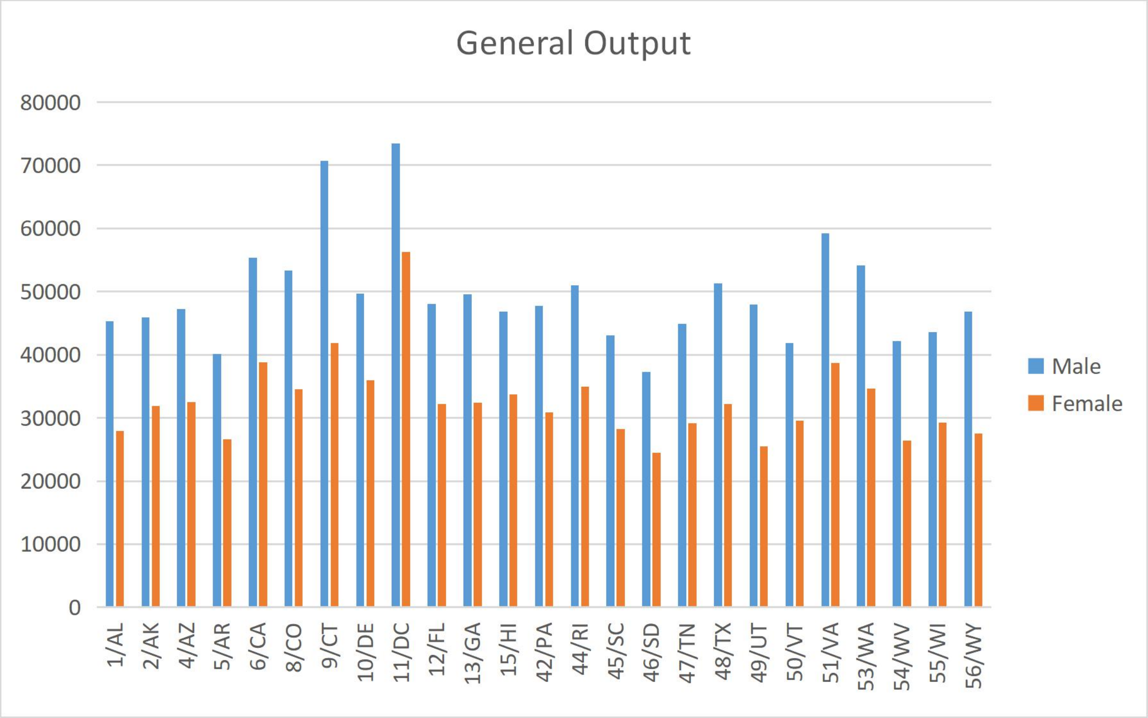
Here is the general output <int Gender, int State, double Average_Wage, long Count>



```
1       1       45283.22748154072       53767
2       1       27956.066570020295      51735
1       2       45947.18175205238       9623
2       2       31859.552006133403      7826
1       4       47212.14492362134       74039
2       4       32541.07576587648       67537
1       5       40148.42720085003       31999
2       5       26571.111301969016      31437
1       6       55388.81676167811       433632
2       6       38830.0178634363        388727
1       8       53338.188661016204      68313
2       8       34494.89728688393       62290
1       9       70674.96842105263       46265
2       9       41869.792647090835      45941
1       10      49712.37757078299       10843
2       10      35954.86055776892       10793
1       11      73495.70814923907       8148
2       11      56252.644988266846      8949
1       12      48010.545963943856      215941
2       12      32154.925151591415      210599
1       13      49583.09007056632       111668
2       13      32357.720717630153      107855
1       15      46799.57620030112       17933
2       15      33749.21818293658       16257
1       42      47778.87384554189       160898
2       42      30916.955295400978      154011
1       44      50999.46260997067       13640
2       44      34961.66017953322       13925
1       45      43106.65245382974       54364
2       45      28205.410545576757      53558
1       46      37247.91254968768       10566
2       46      24523.712047387027      10467
1       47      44912.14335210047       73888
2       47      29155.16463694693       71849
1       48      51344.81589590177       300178
2       48      32216.51473455641       270792
1       49      47908.24604904632       36700
2       49      25483.961001125605      30206
1       50      41855.67270145545       8451
2       50      29589.566225933657      8622
1       51      59243.47240829941       107333
2       51      38659.64213187942       101244
1       53      54100.407217846274      87727
2       53      34597.9453867285        79358
1       54      42179.73044571208       20686
2       54      26413.021688006273      19135
1       55      43620.05779334501       77656
2       55      29224.163826189644      74518
1       56      46829.66352123168       8054
2       56      27490.461082910322      7092
```

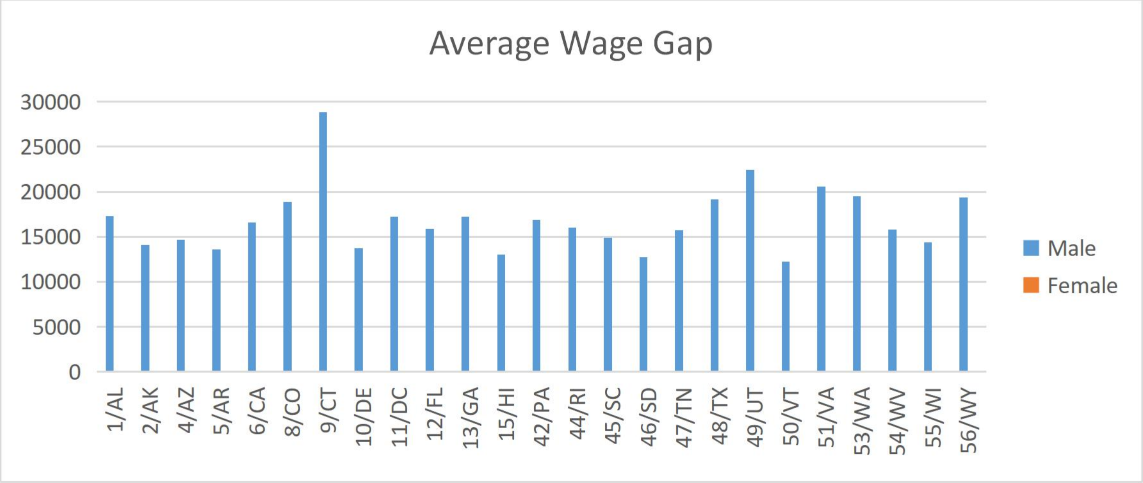
The problem here, is that there are 27 states missing from the csv file. Since the instruction file states that there should be 52 states in total. I only have 25 states available. Anyway, I will do the analysis according to the output where I can get.

Here is the General output of the different gender group average wage in 25 different states.



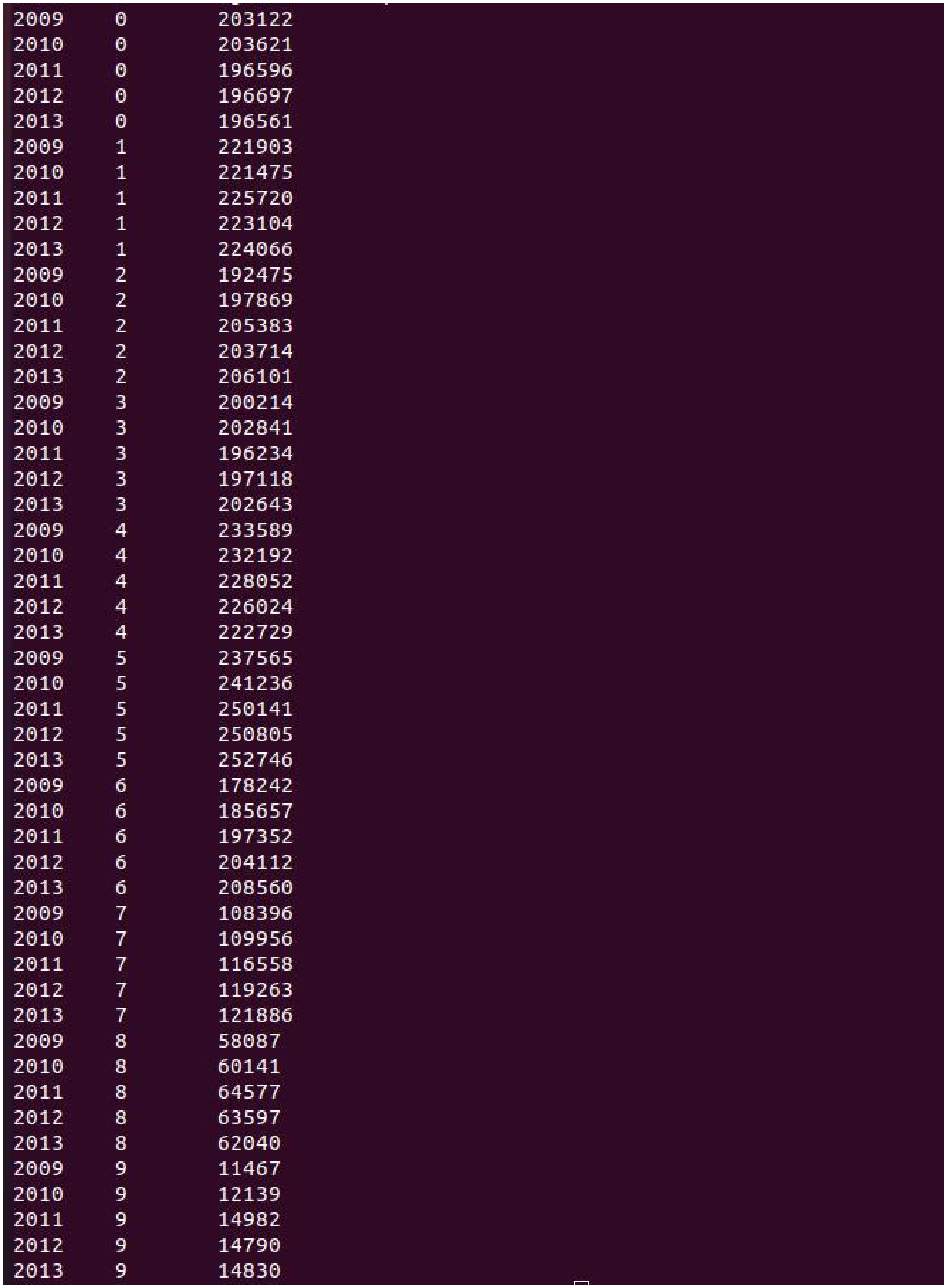
We can see that the highest average wage of male is in the 11/District of Columbia with value $73,495.70. The highest average wage of female is in also in the 11/District of Columbia, with the value $56,252.64. The lowest average wage of male is in the 46/South Dakota and with $37,247.91. And the female making lowest average in this state as well, with $24523.71. Luckily, Texas average wage is above median.

Sorry ladies, but seems like male is making more average wage in all the states. And the biggest gap is in 9/Conneticut with the difference $28,805.17, and the smallest gap is in 51/Virginia with the difference $12,266.11. By the meantime, Texas is the 6th highest among these 25 state.



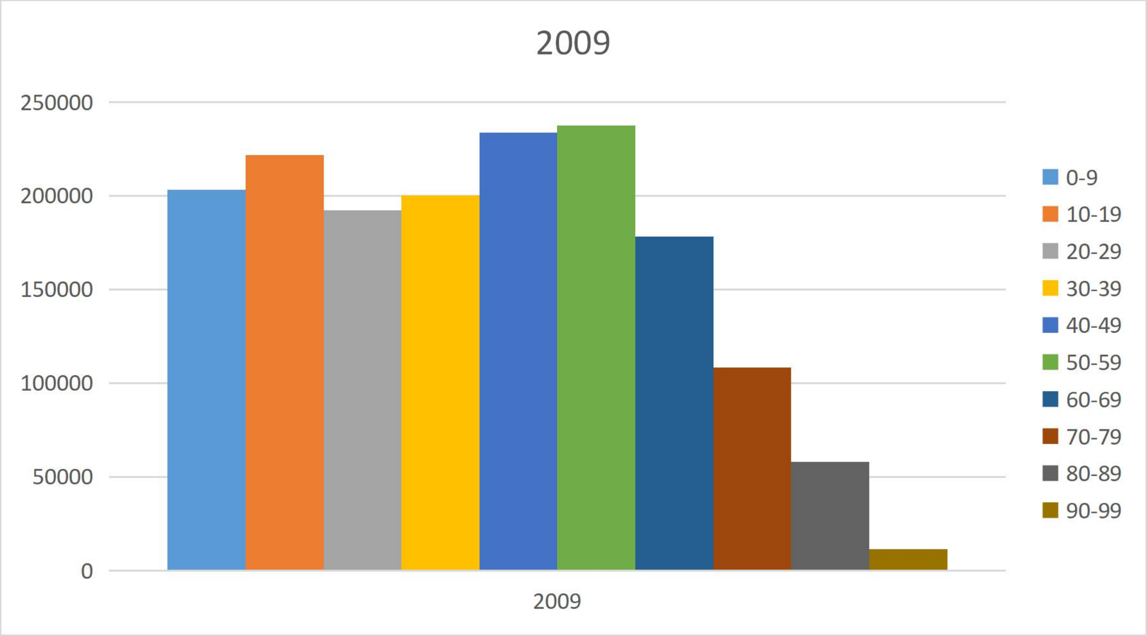
Task 2:

Here is the general output   <string Year, int Age_Group, long Count>

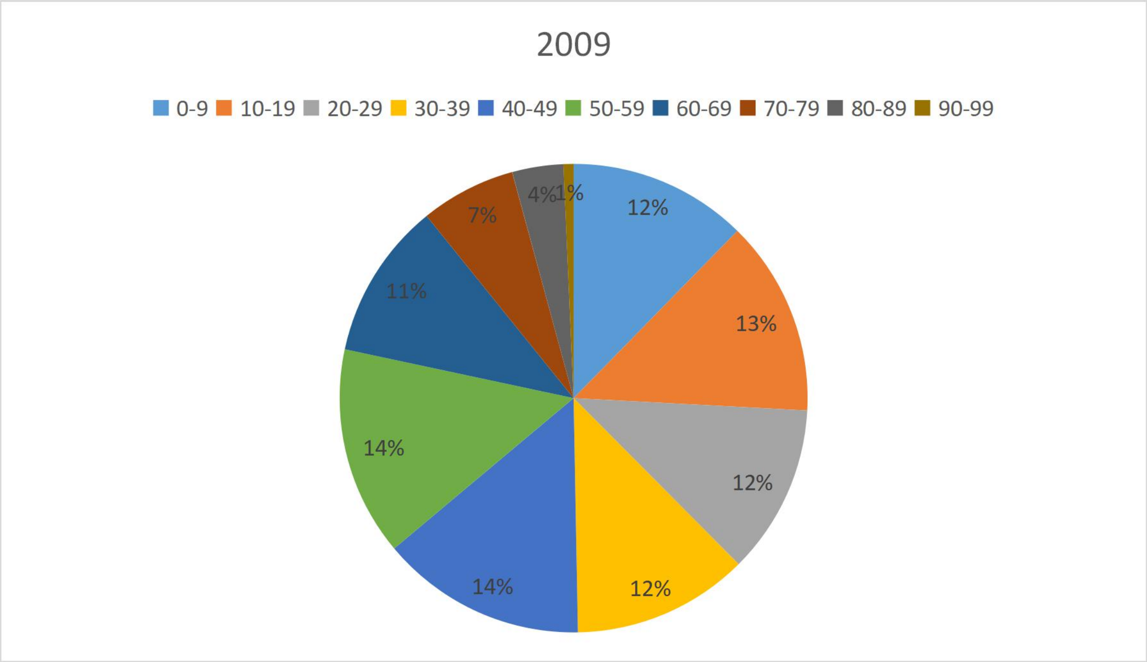


```
2009    0       203122
2010    0       203621
2011    0       196596
2012    0       196697
2013    0       196561
2009    1       221903
2010    1       221475
2011    1       225720
2012    1       223104
2013    1       224066
2009    2       192475
2010    2       197869
2011    2       205383
2012    2       203714
2013    2       206101
2009    3       200214
2010    3       202841
2011    3       196234
2012    3       197118
2013    3       202643
2009    4       233589
2010    4       232192
2011    4       228052
2012    4       226024
2013    4       222729
2009    5       237565
2010    5       241236
2011    5       250141
2012    5       250805
2013    5       252746
2009    6       178242
2010    6       185657
2011    6       197352
2012    6       204112
2013    6       208560
2009    7       108396
2010    7       109956
2011    7       116558
2012    7       119263
2013    7       121886
2009    8       58087
2010    8       60141
2011    8       64577
2012    8       63597
2013    8       62040
2009    9       11467
2010    9       12139
2011    9       14982
2012    9       14790
2013    9       14830
```

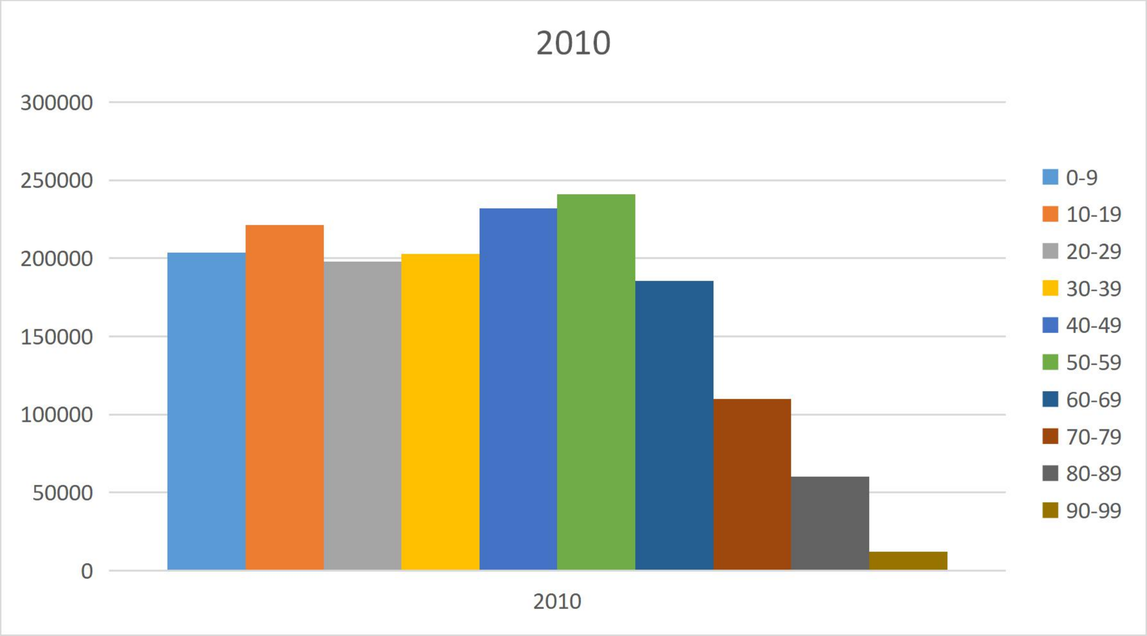
Here is the general histogram output of 2009:



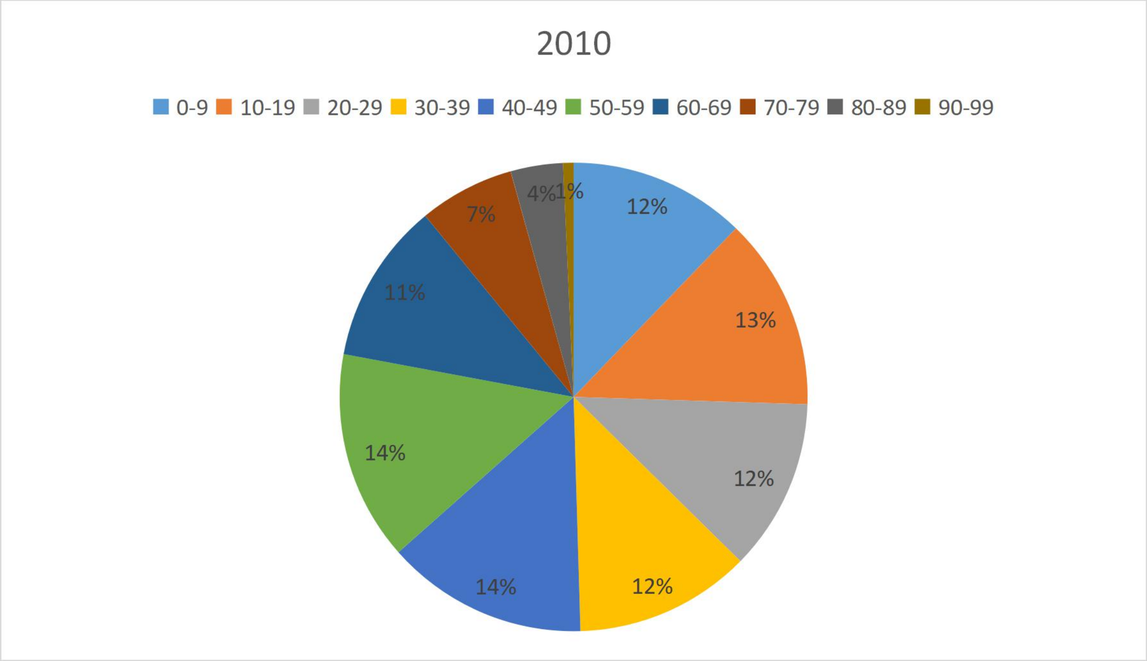
Let’s take a look with a different view of different age groups in percentages with pie chart:



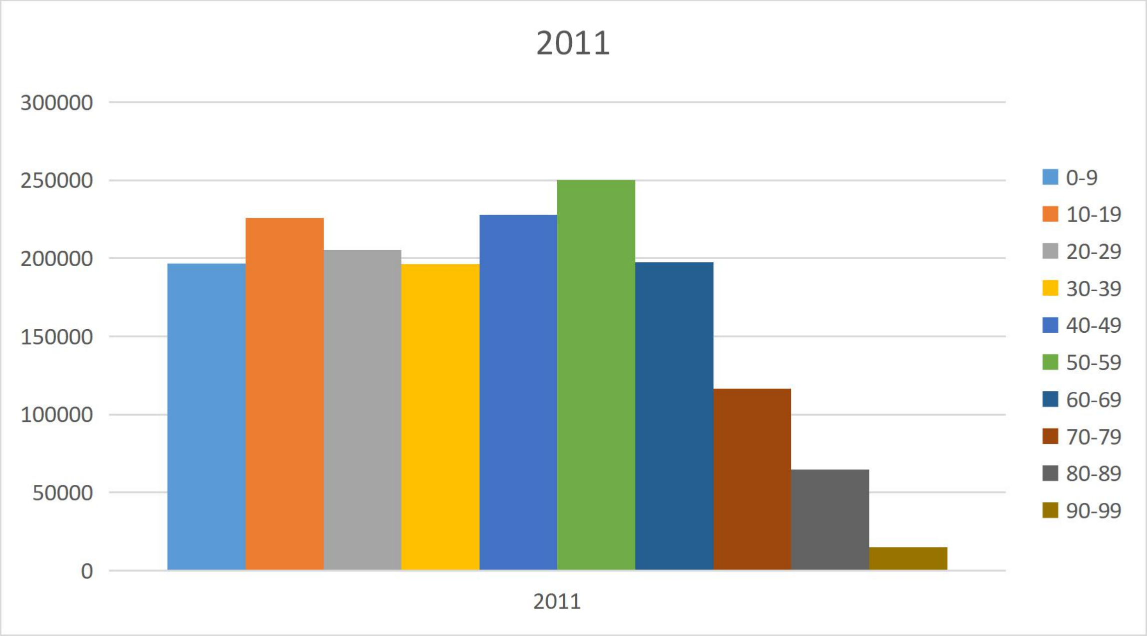
Here is the general histogram output of 2010:



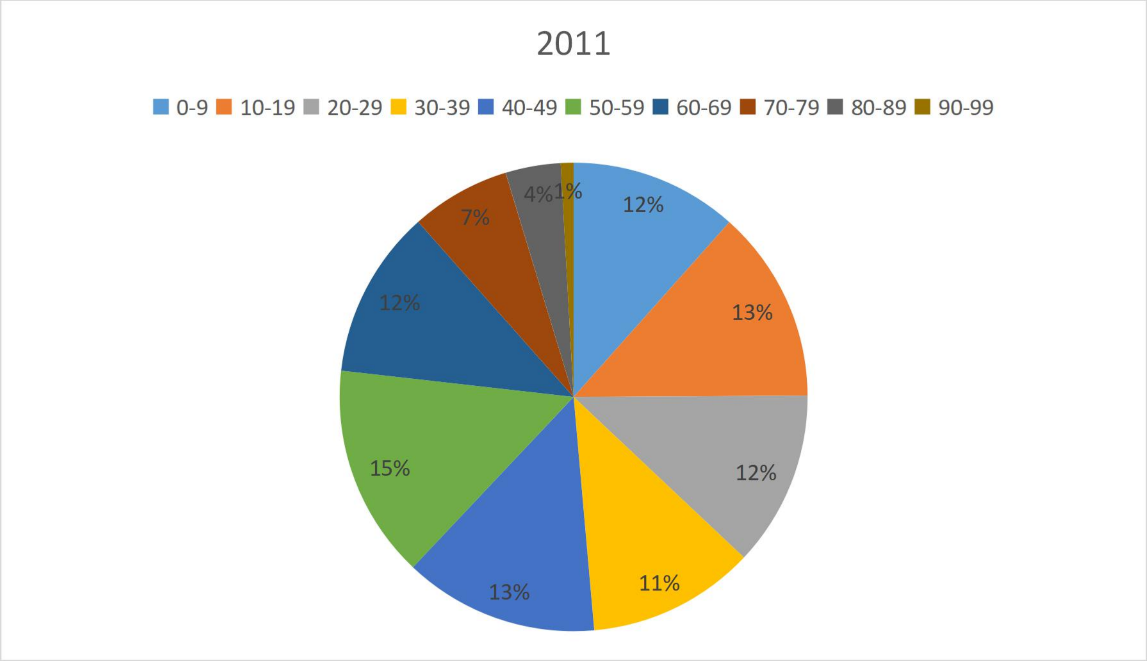
Let's take a look with a different view of different age groups in percentages with pie chart:



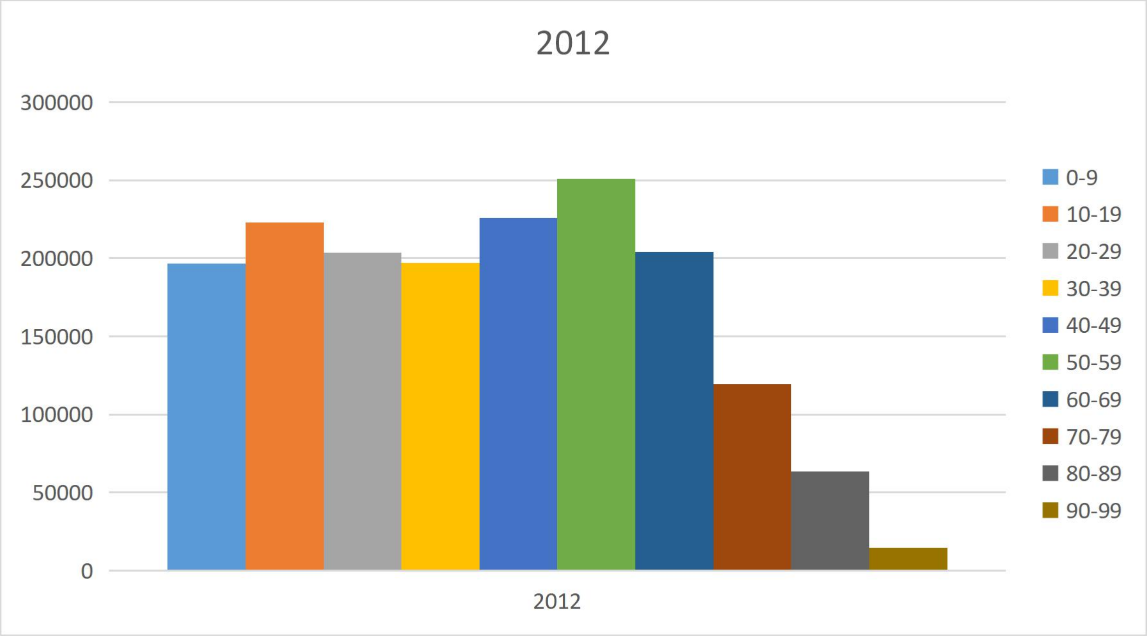
Here is the general histogram output of 2011:



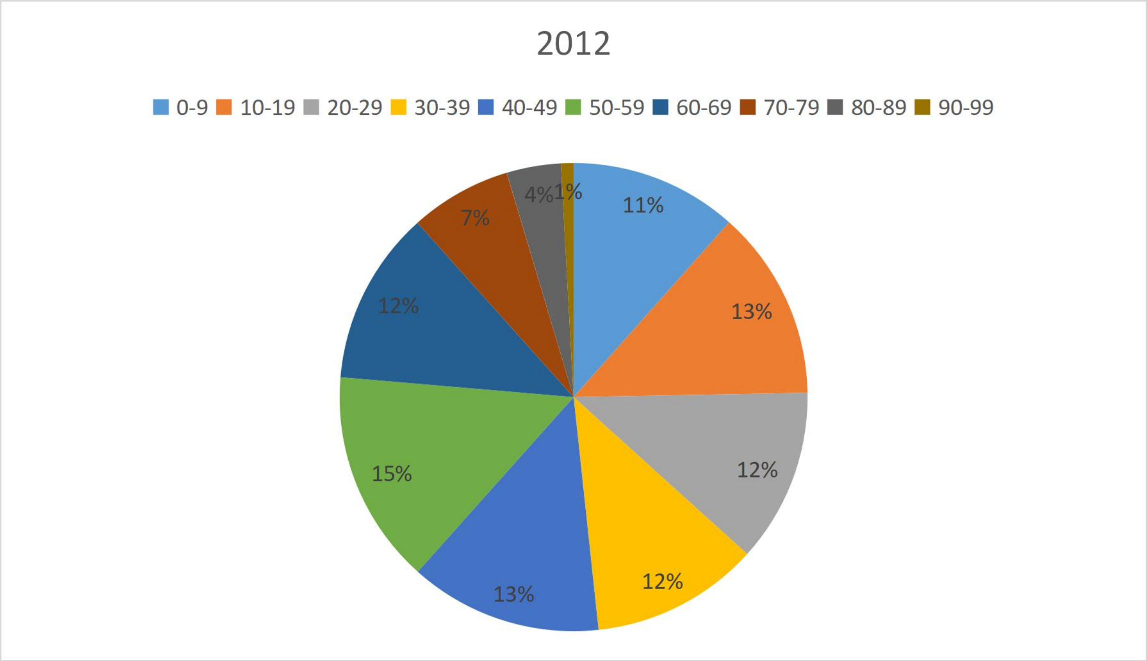
Let’s take a look with a different view of different age groups in percentages with pie chart:



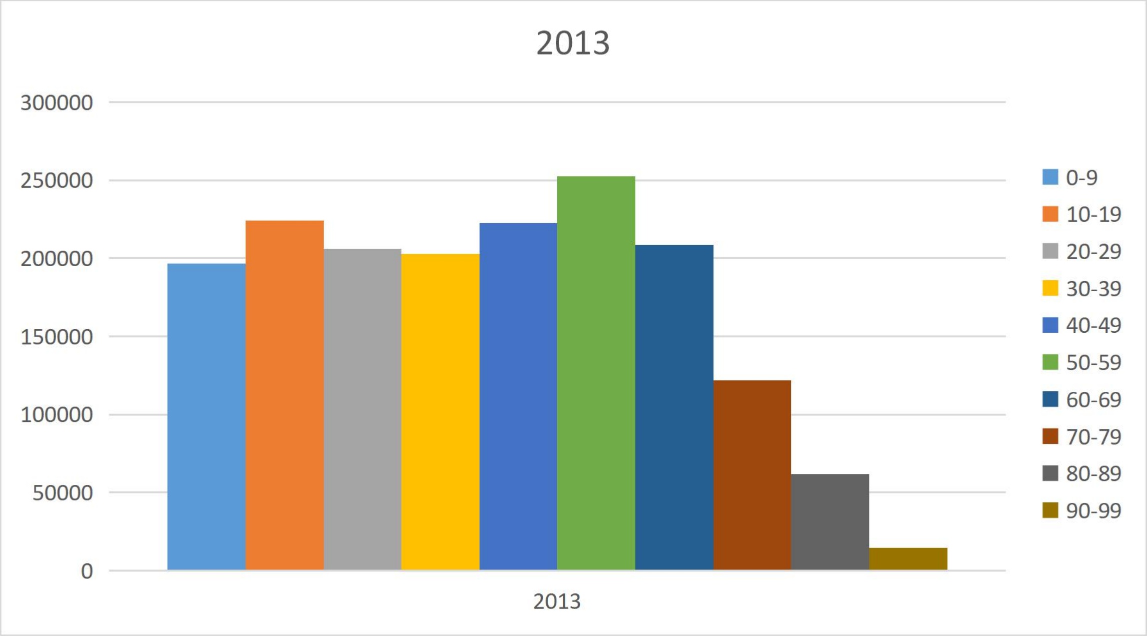
Here is the general histogram output of 2012:



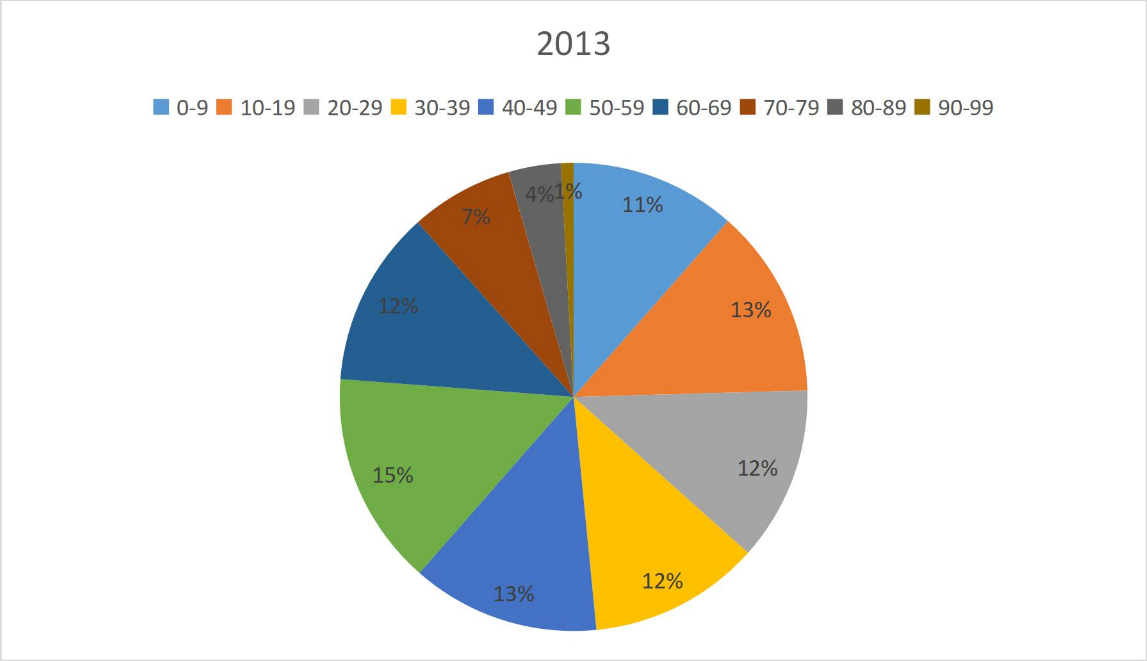
Let’s take a look with a different view of different age groups in percentages with pie chart:



Here is the general histogram output of 2013:



Let’s take a look with a different view of different age groups in percentages with pie chart:



From the above Equiwidth histograms, we can see that the distribution of different age groups in each five year is almost the same. With a large amount of 0-69 young and middle-age groups and a small amount of 70-99 years old group.

Largest percentage age group is 50-59 and 40-49 during 2009-2010 with both 14%. After 2011, the only largest age groups is 50-59 with 15%, and the 40-49 age group becomes the second age group with 13%.

The 70-79, 80-89, 90-99 age groups are very stable, they haven't changed in these 5 years, and they have percentage with 7%, 4%, and 1%. Also the teenager and young adult age groups, 10-19, 20-29 has also not changed in these 5 year, with percentage 13% and 12% each.