Data source: Appended in directly Project which has given access: FinalProjtable.csv

Text

Description automatically generated

(Data profiling directory)

Java file: CountRecsMapper ,CountRecsReducer ,CountRecs

Jar: test8

Aim to find the distinct records for each sectors (I deliberately doing this for later split of sectors)

Graphical user interface, text

Description automatically generatedText

Description automatically generated

Explore Mean Median Mode

Java: MeanModeMedian.java

The output form: Sector name—Mean, Median, Mode, aiming to explore the data distribution

Jar: meanmodemedian3.jar

Text

Description automatically generated Text

Description automatically generated

(Data cleaning directory)

Java: CleanMapper, CleanReducer, Clean

Jar: test15

Result: wipe out all the null value in the input for each sector, below is the cleaned file

Background pattern

Description automatically generated

Result of CountRecs Using Clean file

Jar: Testafter1.jar

Result: we can see a significantly decrement in records for each sector as shown below

Graphical user interface, text

Description automatically generatedText

Description automatically generated

The difference is because unnecessary or abnormal column is eliminated

Data ingesting directory

Distinct column values

Java: FindDistinct.java

Jar: distinctcol.jar

Result: the distinct value for each sector after cleaning the file

Text

Description automatically generatedText

Description automatically generated

Explain the data formatting and processing for the project(Key code)

Java: SectorDataDifference

Jar cleanprocess1.jar

Result: Noticing that this project need the difference between each year, so not only the further cleaning(we need to discard all the sectors with missing values or the comparison between each year will be invalid) is conducted, but also I output the value into the form of a double list keeping track of all the difference of data between 2 years. [2020-2019,2019-2018….., 1991-1990]. Choosing any technical related-job as the final data(software, computer designing…)Graphical user interface, text

Description automatically generatedA picture containing text

Description automatically generated

Finally, using the code in MLFinalcode.txt, I want to use regression to prove that the upper and down of tech employment isn’t a trend, but due to accidental event happening in the world. The RMSE’s tiny value validates it and the final data also shows a decline recently as opposed to the prospering development in the earlier time due to the covid 19 in Washington DC.

The Screenshot for the final code (processing the result of the previous procedural and use machine learning)Text

Description automatically generated