Report:

For this assignment we read in and parsed a large dataset in text format from the changes\_python.txt file, line by line and put it into a list: data. We took its length: 5255 lines.

We then scrubbed (cleaned) the data and placed it into container objects inside the Commit class. We also appended them into a list and my finding out its length we saw that there were 422 different sets of commit objects in our dataset.

We are required to perform an analysis of these 422 objects and come up with 3 interesting statistical pieces of information from this dataset with supporting evidence of interestingness.

With this in mind, first we conducted an analysis of the authors in our dataset.

While parsing the commits, we gathered each author name and kept tracking the number of commits they were responsible for, in a dictionary inside our Statistic class. Once done, we printed this information to the screen. That way we found out all the author names and the corresponding number of commits they were responsible for in our dataset.

We got the length of this author dictionary and that gave us the number of authors in our dataset: 10.

We proceeded to sort the authors from highest to lowest in terms of the number of commits they were responsible for. That gave us their overall ranking.

We found out the author with the highest number of commits: viacheslav.vdovenko with 191 commits. We also found the author with the lowest number of commits: murari.krishnan with 1.

We decided to find the average number of commits each author was responsible for per day. For this, once again, while parsing the commits, we gathered each date in our dataset and put it into a dictionary inside our Statistic class. At the end we got the length of this date dictionary and that gave us the number of days in our dataset: 76. We divided each author’s corresponding number of commits by this length to get the number of commits each author is responsible on average per day. We printed this information to the screen.

We found something else while performing this analysis of the authors in our dataset. While most authors names are either in id form: i.e. dred0001 or in firstname.surname form: viacheslav.vdovenko, there was one that was not: /OU=Domain Control Validated/CN=svn.synchronoss.net. This is a CRON job (scheduled task) that automatically runs and is responsible for 24 commits in our dataset. We have 6 author names in id form, 3 in fname,sname form and one CRON job. So id form is the preferred way.

Second we conducted an analysis of the days of the week in our dataset.

While parsing the commits, we gathered the day of the week and kept tracking the number of commits they had under and put this in a dictionary inside our Statistic class. Once done, we printed this information to the screen. That way we found out the days of the week these commits are being performed: Mon, Tue, Wed, Thu and Fri, and the number of commits each day of the week has under in our dataset. Therefore our commits are performed Mon to Fri i.e. weekdays not weekend.

We got the length of this day dictionary and that gave us the number of days of the week commits are performed in our dataset: 5, confirming the above information: commits are performed 5 days of the week, during weekdays not weekends.

We proceeded to sort the days of the week from highest to lowest in terms of the number of commits they had under. That gave us their overall ranking: 1. Thu with 118 commits, 2. Fri with 95, 3. Tue with 80, 4. Wed with 76, 5. Mon with 53. So most commits are performed towards the end of the working week i.e. Thu and Fri and the least commits are performed during the start of the working week i.e. Mon.

We found out the day of the week with the highest number of commits: Thu with 118 commits. We also found the day of the week with the lowest number of commits: Mon with 53. Mon has less than half the number of commits that Thu has, which is important to know, to see how it could be improved.

We decided to find the average number of commits per day of the week. For this, once again, we used the length of the date dictionary: 76. We divided each day of the week corresponding number of commits by this length to get the number of commits each day of the week had on average. We printed this information to the screen.

Third we conducted an analysis of the changes in our dataset.

While parsing the commits, we gathered each change unique identifier and kept tracking the number of changes they had under and put this in a dictionary inside our Statistic class. Once done, we printed this information to the screen. That way we found out all the types of changes: A, D, M, R and the corresponding number of changes they had under in our dataset.

We got the length of this change dictionary and that gave us the number of type of changes in our dataset: 4, confirming above information.

We proceeded to sort the changes from highest to lowest in terms of the number of changes they had under. That gave us their overall ranking.

We found out the most performed type of change: M with 1186 commits. We also found the least performed type of change: R with 2.

We decided to find the average number of times each change type happened per day. For this, once again, we used the length of the date dictionary: 76. We divided each change type corresponding number of changes by this length to get the number of times each change type happened on average per day. We printed this information to the screen.

Last we conducted an analysis of the times in our dataset.

While parsing the commits, we gathered the time and parse it again to get the hour, minutes and seconds. By checking if the hour was below or above 12 i.e. noon, we gathered data about the time of the day i.e. morning or afternoon each commit was performed and put this in a dictionary inside our Statistic class. Once done, we printed this information to the screen. That way we found out how many commits were performed in the morning and how many in the afternoon.

We found out that most commits are performed during the afternoon: afternoon had 284 commits while morning had 138 commits. From this we can see that there are more than double the number of commits being performed in the afternoon than the morning. This is something interesting to note but maybe not surprising as if they start working in the morning chances are they will tend to finish most of their work in the afternoon thus recording the commits at that particular time of the day.

We decided to find the number of commits performed each time of day on average. For this, once again, we used the length of the date dictionary: 76. We divided each time of day corresponding number of changes by this length to get the number of times commits happened on average on this time of day. We printed this information to the screen.

This concluded our basic analysis of the dataset.