# CA4 – Statistical Analysis in Python Pandas

422 objects transformed from a large dataset of >5000 lines of text were analysed in python using pandas and the following interesting statistical pieces of information were derived. This information is supported by graphs and plots produced using Pandas. Tables contain statistics generated in pandas.

Code: Data transformation – process\_changes\_with\_object.py

Statistical analysis in pandas – pandas.py

Tests: test\_simple\_with\_object.py

## Main Findings

#### Finding #1:

*Most revisions consist of one line of comment.*

#### Finding #2:

*The contribution of each author to the overall revision of this dataset is not directly correlated to the number of revisions made by each.*

#### Finding #3:

*The top 2 authors between them are responsible for 81.28% of all revisions of this dataset, but as seen in Figure 2, may not have over 80% influence on the dataset due to other authors entering more lines per revision than Thomas and Jimmy.*

#### Finding #4:

*The median of the author count is 8. With an upper quartile value of 25.5 and a max of 191 we can see that this dataset is skewed by the high counts contributed by 2 authors (Figure 3).*

#### Finding #5:

*As the mean and median are almost the same value (5 and 5.55 respectively) we may assume that this is an almost symmetrical relationship between date and revisions, with zero skewness. Revisions appear to be made on a routine basis.*

#### Finding #6:

*The most popular date for making revisions was 04/08/2015.*

#### Finding #7:

*Most of revisions are supervised, 95%, with only 5% unsupervised.*

## Data Preparation

The cleaned data in the changes.csv are analysed using Pandas version 0.20.3 in Python 2.7.13. Data are read from .csv into data frame format (df) and explored within this environment. On initial inspection the ‘author’ column has 422 rows and has a data type of object which contains mostly first names, with a potential outlier in the form of an automated entry – this will be investigated further later (see Table 1 and section 3).

Each column is explored using descriptive statistics:

* Author: Object data type with 10 unique entries. Highest frequency is ‘Thomas’ 191 revisions.
* Revision: Object data type with 422 unique entries, 1 for each of the 422 different commit objects. Can be used as a unique ID.
* Date: Object data type with 76 unique entries, the most frequent (04/08/2015) occurs 19 times.
* Time: Object data type with 419 unique entries.
* Number\_of\_lines: Float data type. A range of statistics returned for this is a numeric field:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| 422 | 1.315166 | 0.845672 | 1 | 1 | 1 | 1 | 7 |

6 columns contain various numbers of rows of comments – these data will not be used further in the statistical analyses and are dropped from the df. The revision column is also dropped.

## Summary Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **author** | **date** | **time** | **revision** |
| **count** | 422 | 422 | 422 | 422 |
| **unique** | 10 | 76 | 419 | 422 |
| **top** | Thomas | 04/08/2015 | 11:25:18 | R1493044 |
| **freq** | 191 | 19 | 2 | 1 |

The df shape is now 422 rows x 4 columns. Indexing the df starts it at 0 and steps it in units of 1 until a stop at 422.

## Numeric Plots

Various methods were used to produce plots showing the distribution of the number of lines column within the df (Figure 1). This is the only numeric column in the original df. Figure 1 shows the 422 rows, the revisions, on the X axis and the number of lines within each revision along the Y axis.

#### Finding #1:

*Most revisions consist of one line of comment.*

A close up of a logo

Description generated with high confidence

Figure 1: Lines of comment per revision.

## Categorical Plots

The 10 unique authors in the dataset were categorised by their revisions and the numbers of lines within each of these revisions (Figure 2). Thomas made the most revisions (see Table 1) but these were mostly consisting of one line of comment only; whereas Vincent made much less number of revisions (26, Table 1) but his revisions tended to contain more than one line of code (Figure 2).

#### Finding #2:

*The contribution of each author to the overall revision of this dataset is not directly correlated to the number of revisions made by each.*

A screenshot of a cell phone

Description generated with very high confidence

Figure 2: Author revisions by numer of lines.

## Statistical Analysis

### Author Analysis

A table summarising the authors by the number/count of revisions each made was produced – Thomas has made the most revisions, followed by Jimmy, then there is a large gap to third place, Vincent, with 26 revisions (Table 1, Figure 3).

#### Finding #3:

*The top 2 authors between them are responsible for 81.28% of all revisions of this dataset, but as seen in Figure 2, may not have over 80% influence on the dataset due to other authors entering more lines per revision than Thomas and Jimmy.*

Table 1: Revision count per author.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Count** | **% Total** | **Supervised** | **Unsupervised** |
| Thomas | 191 | 45.26 | ✓ |  |
| Jimmy | 152 | 36.02 | ✓ |  |
| Vincent | 26 | 6.16 | ✓ |  |
| /OU=Domain Control Validated/CN=svn.company.net | 24 | 5.69 |  | ✓ |
| Ajon002 | 9 | 2.13 | ✓ |  |
| Freddie | 7 | 1.66 | ✓ |  |
| Alan | 5 | 1.18 | ✓ |  |
| Nicky | 5 | 1.18 | ✓ |  |
| Dave | 2 | 0.47 | ✓ |  |
| Murari.krishnan | 1 | 0.24 | ✓ |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| 10 | 42.2 | 69.292937 | 1 | 5 | 8 | 25.5 | 191 |

A screenshot of a cell phone

Description generated with high confidence

Descriptive statistics for the author count column were generated (above).

#### Finding #4:

*The median of the author count is 8. With an upper quartile value of 25.5 and a max of 191 we can see that this dataset is skewed by the high counts contributed by 2 authors (Figure 3).*

A new data frame, df1, was created to calculate the percentage of all revisions each author was responsible for (Table 1). These percentages are plotted in Figure 4 and Figure 5 as a line graph and pie chart respectively. Figure 4 mirrors the bar chart in Figure 3, while Figure 5 clearly displays the large proportion of the revisions contributed by Thomas and Jimmy. Figure 5 is clearly a more appropriate method of displaying these percentages.

Figure 3: Number of revisions by author.

A picture containing racquetball

Description generated with high confidence

Figure 4: Percentage of all revisions contributed by each author.

A close up of a logo

Description generated with high confidence

Figure 5: Percentage of all revisions contributed by each author.

### Date Analysis

The date column was plotted by count to illustrate the temporal span and frequency of when the revisions were made (Figure 6). From this plot we can see that on most dates between 1 and 2.5 revisions were made.

A screenshot of a cell phone

Description generated with high confidence

Figure 6: Revision dates.

Descriptive statistics for the date column were generated. An average of 5 revisions were made per unique date, with the most revisions on one date being 19.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| 76 | 5.552632 | 3.889798 | 1 | 2.75 | 5 | 7.5 | 19 |

#### Finding #5:

*As the mean and median are almost the same value (5 and 5.55 respectively) we may assume that this is an almost symmetrical relationship between date and revisions, with zero skewness. Revisions appear to be made on a routine basis.*

A new data frame, df2, was created and sorted to display the dates when most revisions were entered. Table 2 shows the highest 10 counts which make up 30.33% of the total 76 dates when revisions were entered.

#### Finding #6:

*The most popular date for making revisions was 04/08/2015.*

Table 2: Dates when most revisions took place.

|  |  |  |
| --- | --- | --- |
| **Date** | **Count** | **% Total** |
| 04/08/2015 | 19 | 4.5 |
| 13/07/2015 | 14 | 3.32 |
| 29/10/2015 | 13 | 3.08 |
| 15/07/2015 | 13 | 3.08 |
| 20/11/2015 | 12 | 2.84 |
| 09/10/2015 | 12 | 2.84 |
| 12/11/2015 | 12 | 2.84 |
| 26/11/2015 | 11 | 2.61 |
| 22/10/2015 | 11 | 2.61 |
| 02/11/2015 | 11 | 2.61 |

Using df2 the percentage of all revisions occurring on each date was calculated (Table 2) and plotted (Figure 7 and Figure 8). The pie chart in Figure 8 supports the assumption that the data are almost symmetrical with almost zero skewness – clearly a pie chart is not a useful method of displaying this type of data distribution. The line chart in Figure 7 shows the same data as in Figure 8 but the small range in the percentage of revisions is clear here.

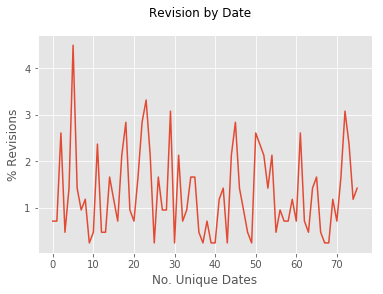


Figure 7: Percentage of total revisions made per date.

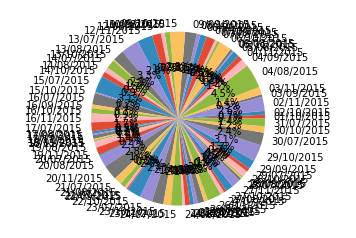


Figure 8: Percentage of total revisions made per date.

### Supervised V’s Unsupervised Analysis

As mentioned previously, most of the author column contains mostly first names, with a potential outlier in the form of an automated entry. This is likely the result of unsupervised revisions, as opposed to the supervised revisions made by the named authors (Table 1). The percentage of supervised and unsupervised revisions of the total revisions are derived by merging data frames (df (original data), and df1 (author summary)), and calculating the percentage of the total revisions into a new data frame, dfnew, for plotting (Table 3, Figure 9).

#### Finding #7:

*Most of revisions are supervised, 95%, with only 5% unsupervised.*

Table 3: Supervised v’s unsupervised revisions.

|  |  |
| --- | --- |
| Supervised % | Unsupervised % |
| 94.31 | 5.69 |

A screenshot of a cell phone

Description generated with very high confidence

Figure 9: Supervised v’s unsupervised revisions.