The face of DEATH

A study in PCA as based on the GRIM books

Hendrik A. Dreyer

Abstract

Can we put a face to death? And, can that face tell us something about the way we, as a people, die in general? The purpose of this study is to deploy the mechanisms of Principle Component Analysis (PCA) on a dataset, which contains the categorized death records of the Australian populace since 1907. This report, through meticulous dimensionality reduction of a complex dataset, seeks to cast new light on how we perceive our exodus from this realm, which we call “reality”. The dataset in this analysis was freely obtained from the Australian government (https://data.gov.au). The dataset was processed, interpreted, analysed and visualized utilizing the R analytical programming language (http://www.rstudio.com/). This report presents the interesting idea of visualizing death data in three dimensions. A subtle personification of death is created by visualising death data. In doing so, numerous observations come to light whereby we can gain a better understanding of related entities that are hidden in and between the death numbers, such as correlation between deaths in world war one and two as well as disassociation within deaths in the age groups 85+. As constant, certain and baffling death is in the light of the sobering visuals in this report, one unyielding question arises, “What should the landscape of death look like? Why? And on who’s authority?” Can a modern society define a utopian death landscape and is it achievable? This report does not aim to answer those questions but rather, to bring those questions into the light as viable and just questions. As the saying goes, “The only things certain in life are death and taxes.” In that case, let’s get a bit more acquainted with the former.[1].

Introduction

The Australian government has been maintaining a General Record of Incidence of Mortality (GRIM) database since 1907 (a.k.a “GRIM books”). The GRIM books (https://data.gov.au/dataset/grim-books) contains deaths data for specific causes of death in Australia from 1907 to 2016. Initial investigation of the GRIM books (https://data.gov.au/dataset/grim-books) let the author to believe that valuable insight can be gained by better understanding the variances between different categories of death data. We, as an Australian society, are particularly good at categorising and recording deaths incidences. Therefore, the context and substance that the GRIM books (https://data.gov.au/dataset/grim-books) offer seemed particularly favourable for data mining. Looking at death data inevitably leaves one with deeper questions, such as - Are there any correlations between death categories, ages, gender and/or time frames? Can we, as a society, better understand these correlations and/or differences. The objective of this report is to decimate the GRIM books (https://data.gov.au/dataset/grim-books) and to summarize and present core aspects of Australian death data.

Data

All of the data in the GRIM books originate from two sources, 1) tabulations of deaths data from 1907 to 1964 and 2) the AIHW National Mortality Database (https://www.aihw.gov.au/about-our-data/our-data-collections/national-mortality-database) as from 1964 onwards. Death records in the GRIM books (https://data.gov.au/dataset/grim-books) (Welfare 2018a) are categorized by the International Statistical Classification of Diseases and Related Health Problems (“ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification” 2018). The data for this project has been freely sourced from a dataset as provided by the Australian government, which is aptly called, GRIM

(https://data.gov.au/dataset/grim-books/resource/edcbc14c-ba7c-44ae-9d4f-2622ad3fafe0) (Welfare 2018b). The dataset is stored in a comma-separated-value (CSV) format and can therefore be consumed by utilising the read.csv() function in R.

```
df <- read.csv("grimdatagovau.csv", header = TRUE, sep = ",")
```

The GRIM (https://data.gov.au/dataset/488ef6d4-c763-4b24-b8fb-9c15b67ece19/resource/edcbc14c-ba7c-44ae-9d4f-2622ad3fafe0/download/grimdatagovau.csv) dataset, in its raw state, consists of 369600 records and has the following nine variables:

- **id**: Integer. Index of dataset
- **Grim code**: 55 factor levels. Each level is associated with a specific grim code, which correlate with an ICD-10 code (http://apps.who.int/classifications/icd10/browse/2010/en). Each grim code, in essence, bundles a set of related death causes and are as such used in the GRIM books
- **Cause of death**: ICD-10 code (http://apps.who.int/classifications/icd10/browse/2010/en)
- **Age group**: 18 factors levels: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-50, 51-54, 55-59, 60-64, 65-70, 71-74, 75-79, 80-84, 85+. Each record in the GRIM books summarizes the number of deaths for a specific age group in a specific year as associated to a specific ICD-10 code (http://apps.who.int/classifications/icd10/browse/2010/en)
- **Year**: Integer - The year in which the deaths occurred
- **Sex**: 4 Factors: Males, Females, Persons, Missing
- **Deaths**: Integer - The number of deaths that occurred
- **Rate**: Numerical (A calculated field)
- **Age standardized rate**: Numerical (A calculated field)

In its raw format, the GRIM books (https://data.gov.au/dataset/grim-books), is not suitable for PCA and the following data wrangling steps had to be applied:

- All rows of type Sex(Persons) or Age group(Missing or Total) were removed as these records were summarising the entries for the Males and Females rows.
- All rows marked with the grim code “GRIM0000” were also removed as they summarized all the other grim codes and are therefore duplicates.
- All records with empty death entries were also removed.
- Only the columns labelled, Grim, Year, Sex, Age group and Deaths were kept from the original dataset.

The new filtered data set contains 129366 records with 5 variables (Grim Code, Year, Sex, Age Group and Deaths). Since the observations for each GRIM category is scattered across multiple rows, the data set has to be transformed into columns, which contains the GRIM categories. For this transformation the function spread(), which is part of the tidyverse library, is used. An index counter is added to the newly formed dataset and any newly created NAs’ are replaced with zeros. NAs are implicitly created due to the transformation process. The newly created dataset has 3960 rows and 58 variables, which is suitable for consumption by the PCA process. The 58 variables in the newly transformed dataset consists of 54 GRIM categories (all entries of GRIM0000 was filtered out due to being duplicates), Year, Sex and Age group. Thus, the PCA process is fed with a scaled matrix of 54 dimensions.

Methods

The statistical analytical software package R (Version: 1.1.456) and its native RStudio work environment (“RStudio. RStudio” 2014) was used to analyse the dataset, generate the results and this report. The prcomp() function (“Prcomp Function R Documentation” 2018), which is part of the R statistical group of functions is used to perform the PCA on the newly transformed dataset. The scale parameter in the prcomp function is set to TRUE in order to scale the input dataset to have unit variance before the analysis takes place.

```
#Perform PCA on grim predictors
PCA <- prcomp(df_grims, scale = TRUE)
```

The prcomp function produces a prcomp object, which amongst other things contain a matrix x (PCA$x). This matrix contains the principal components and their associated weightings. To understand the proportion of variance each component exercises on the dataset, the PVE (Proportion of the Variance Explained) is calculated and listed as follows:

```
# Calc the Proportion of the Variance Explained
PVE <- (PCA$sdev^2)/sum(PCA$sdev^2)
```

Figure 1: Principle Component Variances

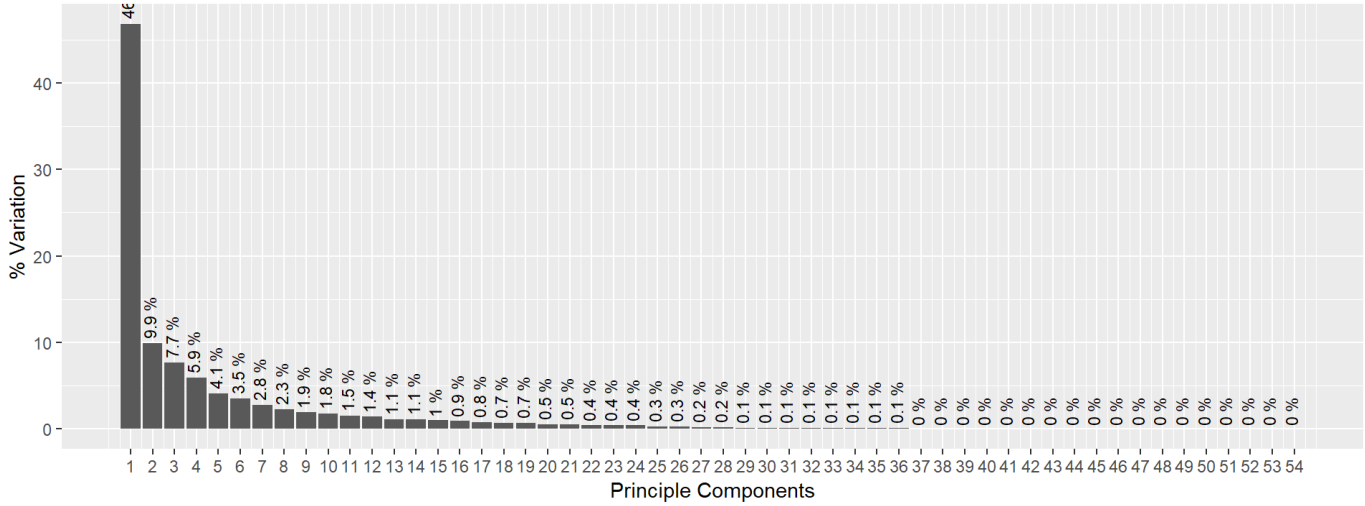


Figure 1 displays the PVE (scaled to %) values for each principal component in descending order. From Figure 1, it can be observed that the first two components account for 56.7% (46.8% + 9.9%) of the variance in the dataset. This warrants further investigation as to exactly how these two components, PCA1 and PCA2, contributes towards such a large portion of the variance. The scatterplot in Figure 2 illustrates the relationship between the values of PCA1 and PCA2. The x-axis in the Figure 2 illustrates what portion of the variance in the dataset PC1 accounts for and the y-axis illustrates what portion of the variance in the dataset PC2 accounts for.

In order to understand which GRIMs exhort the largest effect on where the points are plotted, the loading scores can be utilised as contained in the prcomp object labelled, “PCA$rotation”. For instance, let’s examine the loading scores for PCA1 since it accounts for 46.8% alone of the variance in the data. GRIMS that pushes the points to the right of the graph will have large positive values and GRIMS that pushes the points to the left will have large negative values. But, as can be viewed in the Figure 2, there are several points that are distributed far to the right. Hence, only the top 10 largest loadings will be evaluated.

```
#The loadings for PCA1 accounts for 46.8% of the variation in the data
grim_loading <- PCA$rotation[,1]

#Sort the loadings as decreasing
grim_scores_rank <-sort(grim_loading, decreasing = TRUE)
top_10_grims <- names(grim_scores_rank[1:10])

#Display top 10 loadings
PCA$rotation[top_10_grims,1]
```

```
##  GRIM0400  GRIM1100  GRIM1400  GRIM0403  GRIM0005  GRIM1300  GRIM0600
## 0.1948536 0.1915667 0.1896101 0.1893450 0.1889115 0.1876869 0.1873762
##  GRIM1407  GRIM1200  GRIM1000
## 0.1836022 0.1833417 0.1761204
```

By cross-referencing the GRIM codes in the ground truth dataset, the associated ICD codes can be found. Below are the top 10 GRIM codes listed with the highest loading scores and therefore has the largest effect (push to the right) on the points in the Figure 2:

- GRIM0400 (0.1948536) - All endocrine, nutritional and metabolic diseases (ICD-10 E00-E90)
- GRIM1100 (0.1915667) - All diseases of the digestive system (ICD-10 K00-K93)
- GRIM1400 (0.1896101) - All diseases of the genitourinary system (ICD-10 N00-N99)
- GRIM0403 (0.1893450) - Diabetes (ICD-10 E10-E14)
- GRIM0005 (0.1889115) - Chronic kidney disease (ICD-10 B52.0, D59.3, E10.2, E11.2, E12.2, …)
- GRIM1300 (0.1876869) - All diseases of the musculoskeletal system and connective tissue (ICD-10 M00-M99)
- GRIM0600 (0.1873762) - All diseases of the nervous system (ICD-10 G00-G99)
- GRIM1407 (0.1836022) - Kidney failure (ICD-10 N17-N19)
- GRIM1200 (0.1833417) - All diseases of the skin and subcutaneous tissue (ICD-10 L00-L99)
- GRIM1000 (0.1761204) - All diseases of the respiratory system (ICD-10 J00-J99)

To get a better idea of the structures and clusters presented in Figure 2, two different colours, for Males (blue) and Females (pink), are overlaid onto the plotted points as presented in Figure 3 below.

Figure 2: PC1/PC2

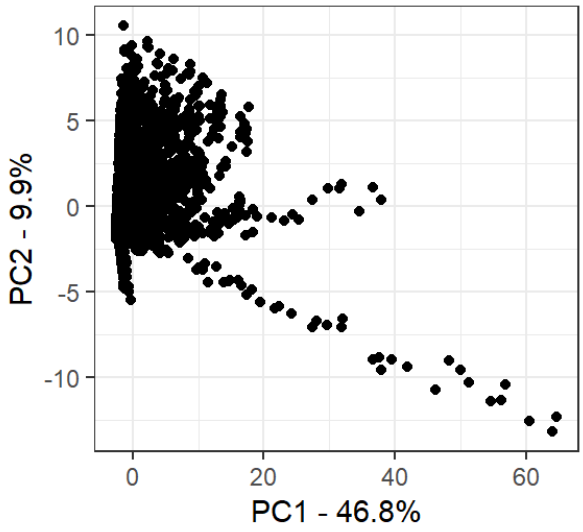
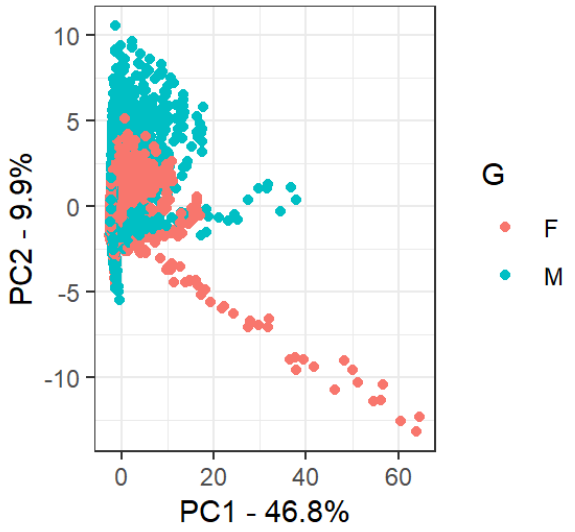
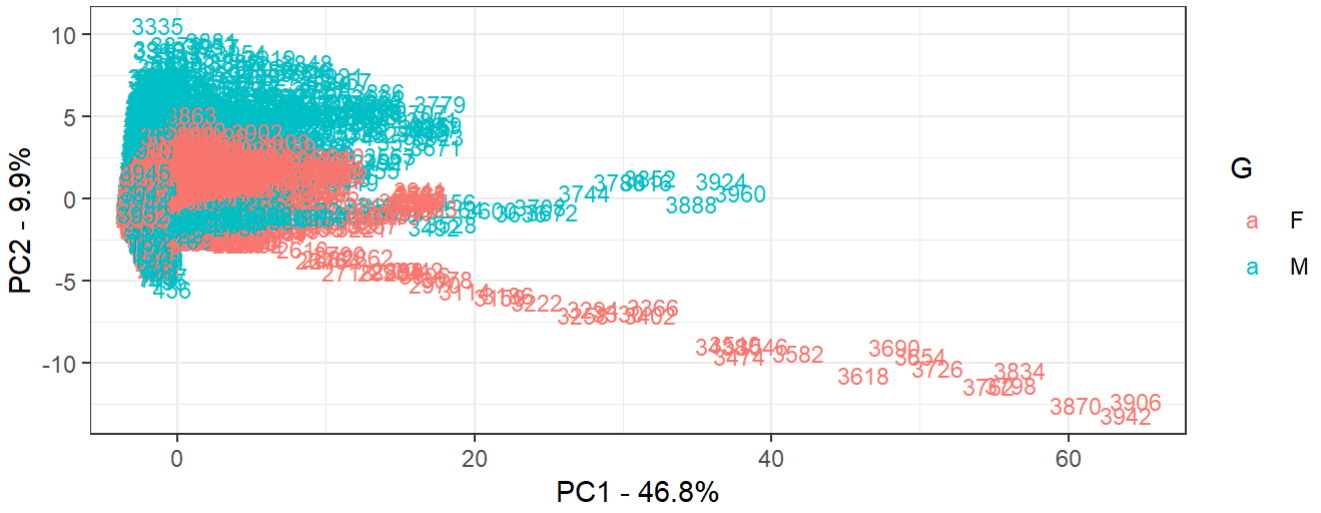


Figure 3: PC1/PC2 - Gender



Even more insight can be gained by overlaying the plotted points in Figure 3 with index numbers from the PCA tables, as presented in Figure 4 below. Each of the indexes can then, in turn, be cross referenced with GRIM codes and their associated ICD-10 codes. For instance, the label 3942 situated in the bottom right hand corner of Figure 4, when used to index the ground truth data reveals that, that point describes the huge variances of deaths for females in the age group 85+. Numerous GRIM codes in the indexed row contribute towards the disproportionate number of deaths for that socio-economic sub-category.

Figure 4: PC1/PC2 - Gender(Ind)



Figures 2, 3 and 4 reveals a few interesting aspects regarding the dataset. Already a few groupings can be viewed, and this leads to even further investigation and analysis by exploring a third dimension of the principle components as well as selective filter and display of the other categories such as, sex, age group and year. The year category especially offers some interesting insights into groupings and distributions of deaths over the last century, especially during World War 2 and the Spanish Flu in 1919. The 3D scatter plots as listed in APPENDIX A - D, presents a chilling visual display as the data can be rotated in 3D. The 3D plots reveal several clusters and vectors (aptly referred to as DVs - Death Vectors). The 3D plots in the Appendices were generated by utilising the scatter3d function, which is part of the car library. These plots can be rotated in 3D and zoomed, both positive and negative. By utilising these spatial features in the 3D plots, a full sense of PCA1, PCA2 and PCA3 can be appreciated. A full discussion and interpretation of these visualisation are presented in the Results and Discussions section.

Results and Discussions

NOTE: For the results and discussions, please refer to the 3D plot as listed in Appendix A - D.
First off, a comment must be made on the general 3D structure, which the top three PCAs express in all of the 3D plots. By rotating any of the plots (Figures 5 to 12) it can be observed that the overall structure is roughly the form of a baseball glove (catcher's mitten). Upon closer inspection it can be observed that the central part of the structure consists of a few smaller vectors of elongated, clustered points. From here on forth, these vectors will be referred to as DVs (Death Vectors). These smaller DVs are of interest and should be further studied and identified as they might present valuable information regarding the finer structures of the variances in the dataset. From the main structure three larger DVs can be observed. At first sight, these larger DVs seem like outliers. But, are they?

The plotted points in Figure 5 (Appendix A) has been coloured by a continuous colour scale for the period 1907 to 2016, which ranges from white to blue. From this colour scheme it can be observed that most of the recorded deaths took place towards the latter end of the 1907-2016 time period. This makes sense, as the Australian government became more meticulous and accurate in recording deaths data during the second half of the twentieth century.

In Figure 6, the data points are separated between Males (blue) and Females (pink). An number of interesting aspects arises from this colour scheme. The largest DV is represented by females and the smaller DV is represented by males. An ellipsoid is drawn around the centre cluster of the points and upon zooming in on this ellipsoid it can be observed that a larger portion of female deaths are encapsulated within the volume of the ellipsoid and that a vast number of male deaths fall far outside of the ellipsoid. These latter two observations indicate that there are a large portion of female deaths that contributes towards the variance associated with this DV. The same argument goes towards the smaller DV represented by male deaths. The male deaths that are scattered outside of the ellipsoid indicates that males, in general, die of numerously more

causes than females. This is probably due to the socio-economic structure of the previous century, wherein males represented the larger portion of the work force and therefore where more prone to succumb to work related injuries and also taking a more active role in warfare than females do in general. On the other hand, the female deaths inside the ellipsoid could most likely be described to deaths that are caused due to female physiological attributes and abilities such as, child birth.

Figure 7 illustrates, by way of a continuous colour scale (red to black), the various age categories. Red represents the younger side of the age scale (0-5) and black represents the older side of the age scale (85+). From this figure we learn that the two larger DVs for females and males are all elderly people. This indicates that a very large portion of deaths in our society is represented by the elderly. it is a known fact that Australians live longer today than ever before (Wade 2018). The two larger black DVs (oriented nearly perpendicular to each other in 3D space) also indicates that males and females die of different old-age related diseases. A third larger DV, coloured in red, represents deaths by children in the age category (0-5). This also makes sense, as children of that age would die of different diseases than the elderly and hence, the cluster is orientated differently in 3D space than the others.

Figure 8 presents the deaths of everyone from 1907-1917(red) and everyone from 2006-2016(black). It can be observed that a far larger and more nuanced spectrum of DVs represents deaths in the last decade than a hundred years ago. This makes sense in the light that we, today, are far better and more meticulous in categorising and recording deaths data.

Figure 9 and 10 gives a comparison between all the deaths of Australians during the second world war and during the Spanish Flu. It can be observed that the deaths during the Spanish Flu all lies on the same smaller DV. This makes sense, as most deaths would have been related to the flu. Another intrusting comparison is the size of both the clusters in both the figures. It is believed that up to 15000 Australians died of the Spanish Flu (“1918 Influenza Pandemic - Australia: Fatalities genealogy project. geni_family_tree” 2018) and that 23000 Australian were killed in WW2 (“Conflicts” 2018).

Figures 11 and 12 presents the deaths of elderly people (85+) in red and the rest in black for the years 1907-1917 and 2006-2016. It can be observed that a far larger portion of deaths for this age category presents the variance in the dataset over the last decade. We know we are living longer due to health and socio-economic advances. In order to view the red coloured points in Figure 11, the reader has to rotate the plot 180 degrees and zoom into the main section of the plot to see the correct data points.

Conclusions

We as a democratic and capitalistic society who is for ever driven by the need to monetize each aspect of our lives can take a hint or two from the visualizations in this report. Clearly, the visualisations indicate that the a dis-proportionate number of elderly people are dying daily. Any nation that strives to keep its social, moral and economic house in order should take the latter aspect into account. But, then, also to be fair, we should surely look to the needs of the elderly as well. The two main DVs (elderly males and females), as illustrated in the visualisations, spells out two monetized markets - healthcare and funeral. It also defines a fiscal weight for government - at least the healthcare market. On the other hand, the healthcare market is a huge job creation sector. So, like any normal government, they are trying to have it both ways - cheap/underfunded healthcare but high numbers of healthcare personnel. What is the solution to this, or is this just the way life is? Or, should we be investigating other exit strategies such as voluntary euthanasia? How would this impact the social, moral and economic aspects of our society? Let’s argue for a moment, that the two death vectors are an abnormal phenomenon in our society. And, let’s for a moment, imagine that those to vectors are shortened. Would the plot then display an acceptable variation pattern? Or, might we see the two death vectors as the outcome of a society that has at least, for now, achieved the state in which we can grow old and experience life longer? There is probably no straight yes or no answers to these questions. Instead, the landscape of death as illustrated in the many 3D visualisation in this report is a mere reflection of the ebbs and flows of what our society is going through as we flow down the river of time and change.

In the light of these findings, this report has achieved in what it set out in the first place. It managed to simplify a complex dataset with a high number of dimensions. It found a way to present the dataset and to explain various aspects of the presentation. As an end result, it managed to find several interesting and thought-provoking ideas and aspects hidden in the data.

APPENDICES

NOTE: Scatterplot can be zoomed in and out by way of the mouse wheel. It is advisable to do so, as the plots contain a high density of points.

A - Deaths: Continuous Scale % Center Sphere (Males/Females)

Fig. 5: years 1907(white) - 2016(blue)

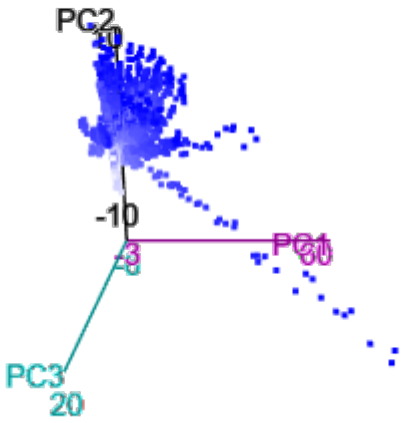
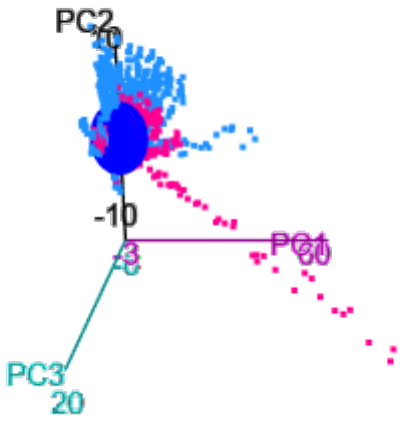


Fig. 6: Males (blue) and Females (pink)



B - Deaths: Continuous Scale (Age) & Then and Now

Fig. 7: 0-5(red) to 85+(black)

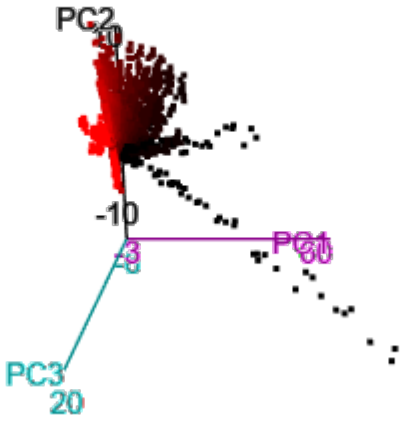
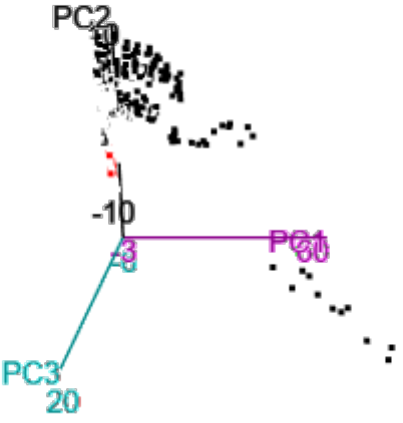


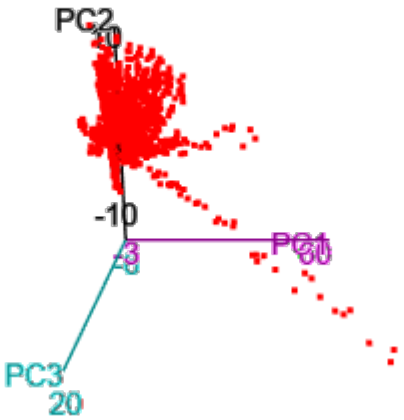
Fig. 8: Now:2016(black) and Then:1907(red)

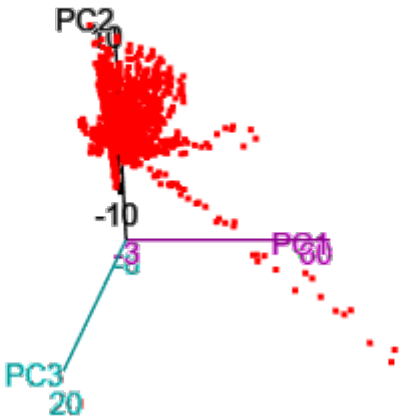


C - Deaths: WW2 & Spanish Flu

Fig. 9: WW2 (black) the rest(red)

panish Flu (black) the rest

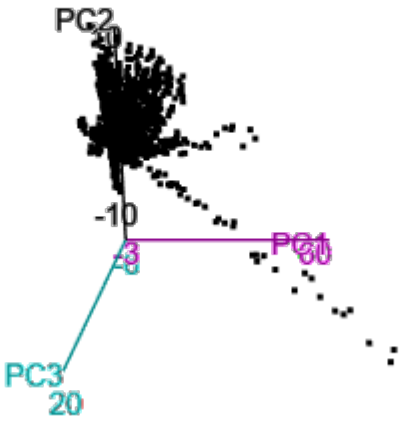


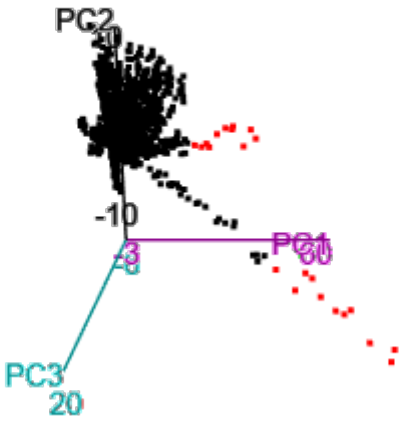


D - Deaths: Age 85+ (1907) & Age 85+ (2016)

Fig. 11: 85+ 1900(black the rest (red)

Fig. 12: 85+ 2016(black the rest (red)





References

"1918 Influenza Pandemic - Australia: Fatalities genealogy project. geni_family_tree." 2018. Accessed December 5. https://www.geni.com/projects/1918-Influenza-Pandemic-Australia-Fatalities/24220 (https://www.geni.com/projects/1918-Influenza-Pandemic-Australia-Fatalities/24220).

"Conflicts." 2018. Accessed December 5. http://www.naa.gov.au/collection/explore/defence/conflicts.aspx (http://www.naa.gov.au/collection/explore/defence/conflicts.aspx).

"ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification." 2018. July 26. https://www.cdc.gov/nchs/icd/icd10cm.htm (https://www.cdc.gov/nchs/icd/icd10cm.htm).

"Prcomp Function R Documentation." 2018. Accessed December 5. https://www.rdocumentation.org/packages/kazaam/versions/0.1-0/topics/prcomp (https://www.rdocumentation.org/packages/kazaam/versions/0.1-0/topics/prcomp).

“RStudio. RStudio.” 2014. April 4. https://www.rstudio.com/products/rstudio/ (https://www.rstudio.com/products/rstudio/).

Wade, Matt. 2018. “Trend for Australians to Live Longer Reshapes Economy. the Sydney Morning Herald.” August 11. https://www.smh.com.au/business/the-economy/trend-for-australians-to-live-longer-reshapes-economy-20180810-p4zwuv.html (https://www.smh.com.au/business/the-economy/trend-for-australians-to-live-longer-reshapes-economy-20180810-p4zwuv.html).

Welfare, Australian Institute of Health and. 2018a. “General Record of Incidence of Mortality (GRIM) Books.” Accessed December 4. https://data.gov.au/dataset/grim-books (https://data.gov.au/dataset/grim-books).

———. 2018b. “GRIM.” Accessed December 4. https://data.gov.au/dataset/grim-books/resource/edcbc14c-ba7c-44ae-9d4f-2622ad3fafe0 (https://data.gov.au/dataset/grim-books/resource/edcbc14c-ba7c-44ae-9d4f-2622ad3fafe0).

---

1. Footnote: It is not the author’s intent to cast a sombre light on the topic at hand but, instead to seek a better understanding of how the landscape of death is changing as we progress through time and the advances of medical technology and social changes. Although there is only one way to enter this world, there seems to be a plethora of ways to exit and it seems ever changing. As the data suggest we all can’t just peacefully slip away in our sleep one blissful night but must face the reality that there exists a real possibility to be a hit-and-run victim in a McDonalds drive-thru in Frankston (https://www.heraldsun.com.au/leader/news/record-number-of-cars-impounded-in-victoria/news-story/f04bf8bb723bb7f5f4d051382bbf24c4)↩