# Predicting the cause of death based on age and musical skill

## System Analysis and Decision Support Methods

Natalia Malinowska
272669
Applied Computer Science
Wroclaw University of Science and Technology

Wed. (even) 15:15

10th of June 2024

# Contents

# 1 Problem intro

The problem addressed in the research is the relationships between data on deceased musicians, including the causes of their deaths. The inspiration for these considerations was asking whether it is possible to effectively predict what a musician will die from, taking into account factors such as their age, place of residence, genre of music played, or musical skills.

The aim of this study is to analyze whether this data can be used to build a predictive model for the causes of death among musicians. By analyzing the acquired dataset, patterns and correlations are sought that may provide information on how these factors interact, and potentially influence the mortality of musicians. We will use various statistical methods to analyze the data, conduct experiments on them, and create a model.

The assumption is that there is some relationship between the deaths of musicians and the genre of music they engage in. For example, musicians performing hip hop or rock music are known for a more impulsive, risky lifestyle than jazz or classical musicians - therefore, such a relationship should emerge from our considerations.

# 2 Dataset acquisition and exploratory analysis

## 2.1 Dataset acquisition

Acquiring complete, reliable data, then cleaning, and preparing it for analysis, was likely the most challenging task in this project.

The initial dataset of deceased artists was obtained from the *rateyourmusic.com* website using web scraping method with the help of the *BeautifulSoup* Python library. However, there was a lack of data about artists' birth dates – necessary for calculating their age – as well as several parameters that could potentially be used for additional analysis, such as gender or music genre. To address this, an API provided by the *musicbrainz.org* website was utilized.

After obtaining two datasets, they were synthesized into one dataset by matching them based on the artists' names. Then, a multi-stage process of data cleaning and preparation was carried out. The age was calculated, the places of residence were shortened (to countries only), and the data was generalized – for example, nearly 200 causes of death that came in various forms were categorized into 11 categories, similarly done with music genres, and skills.

### 2.1.1 Individual stages of data acquisition and processing

a) Obtaining the first part of the data using a web scraping method

b) Obtaining the second part of the data using an API method

c) Data synthesis

d) Standardization of data

e) Removal of erroneous and empty data

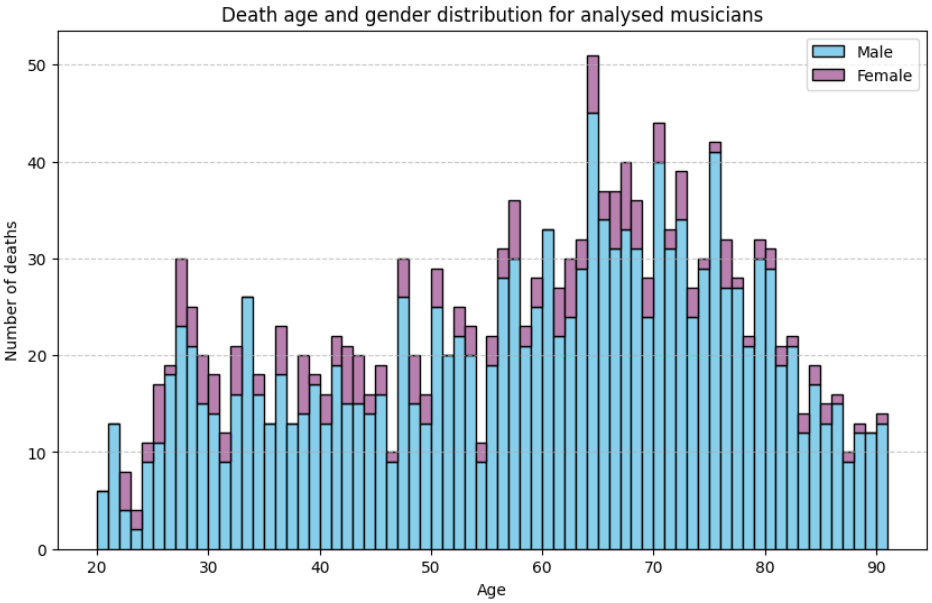
f) Generalization/categorization of data

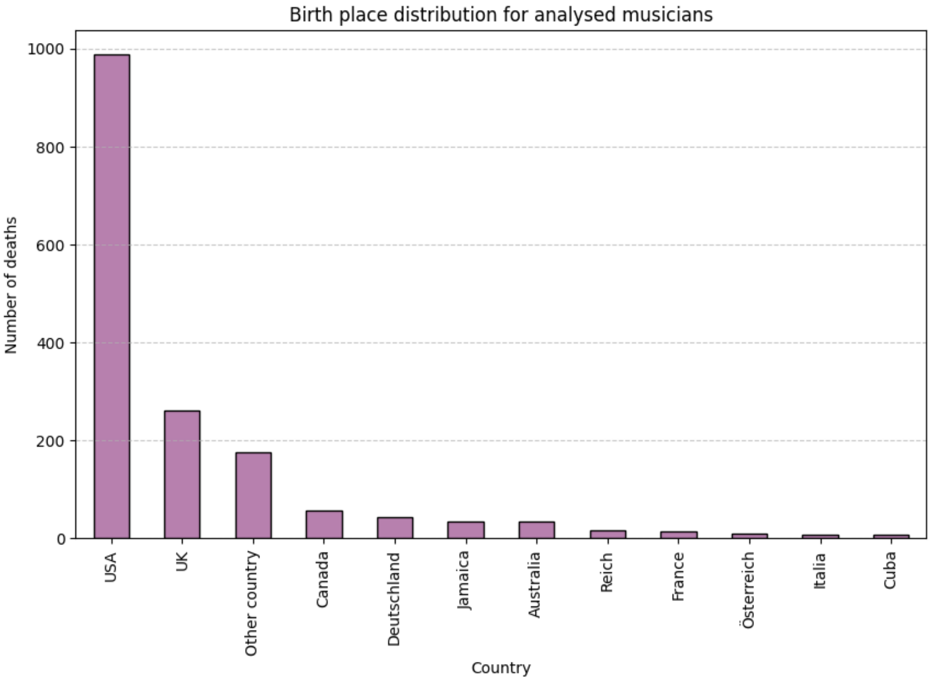
g) Repairing data formatting errors

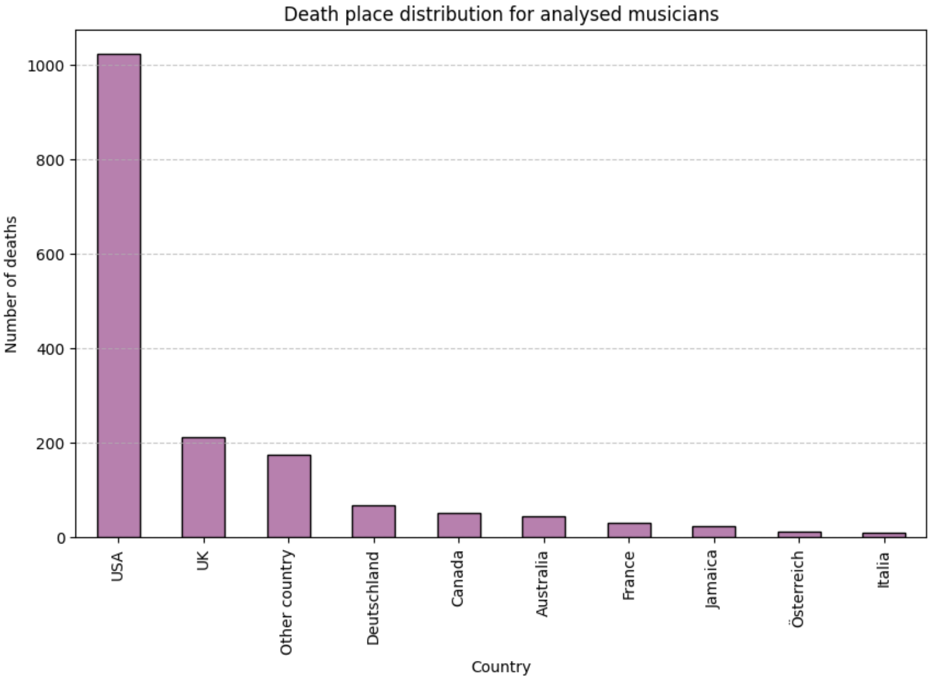
h) Generating separate datasets (non-atomic data in columns, e.g., ”piano, guitar, vocals” in the ”musical_skills” column, were split into atomic values to enable proper analysis)

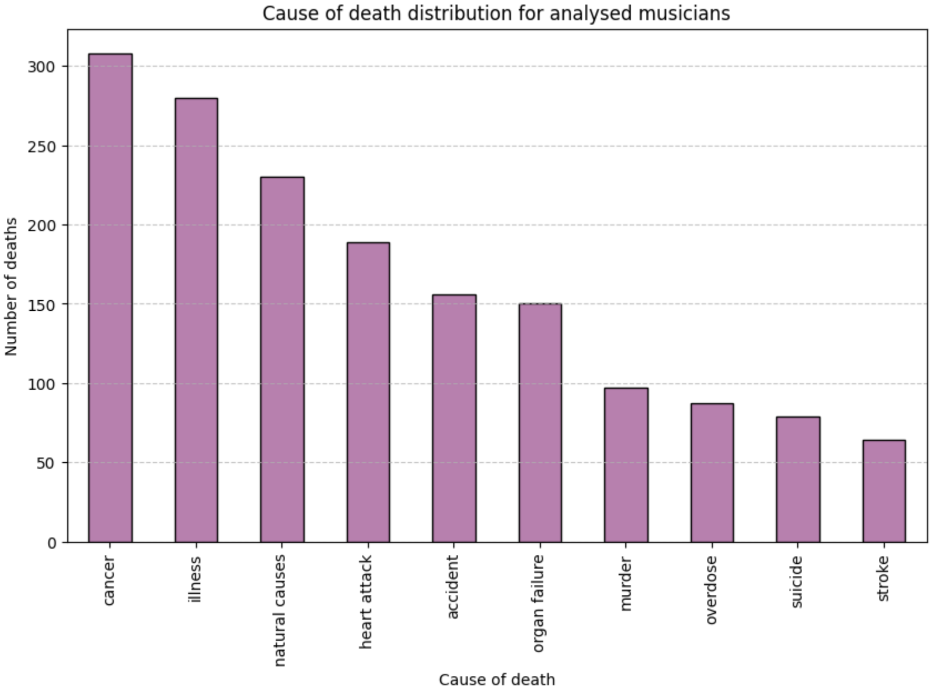
## 2.2 Exploratory analysis of the dataset

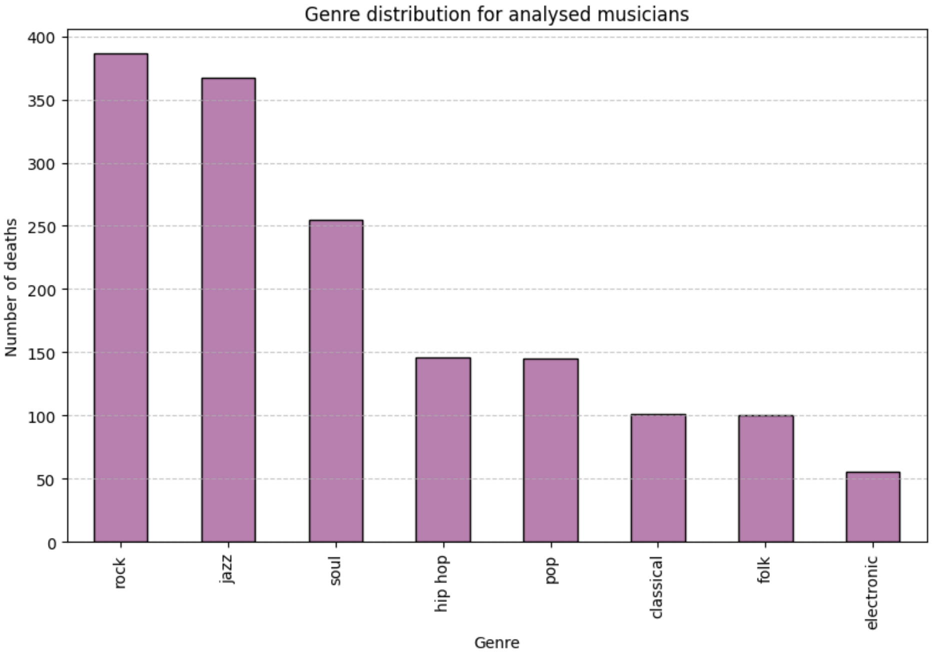
### 2.2.1 Analysis of the amount of deaths within a particular parameter

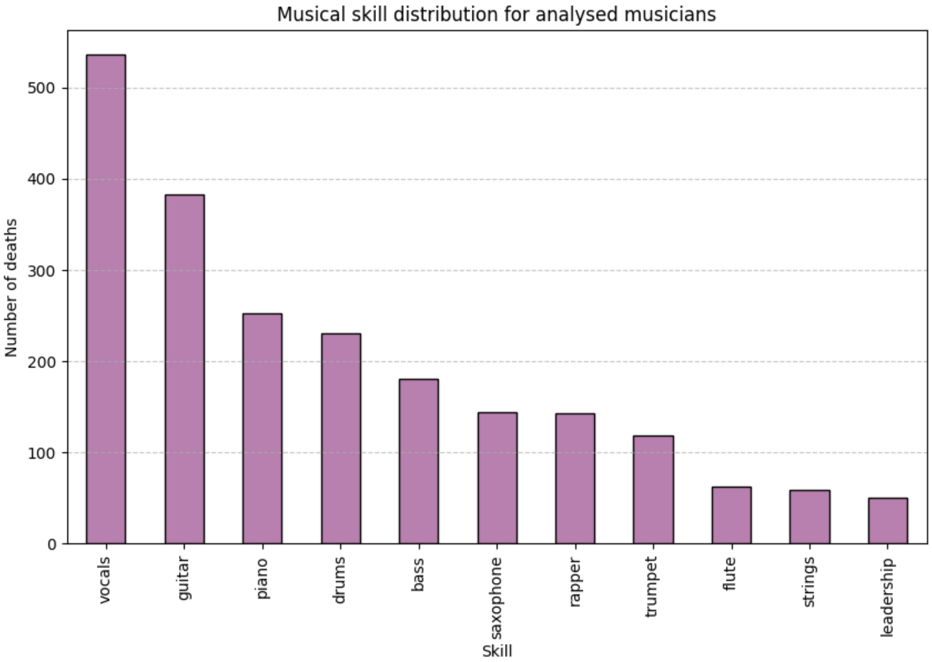




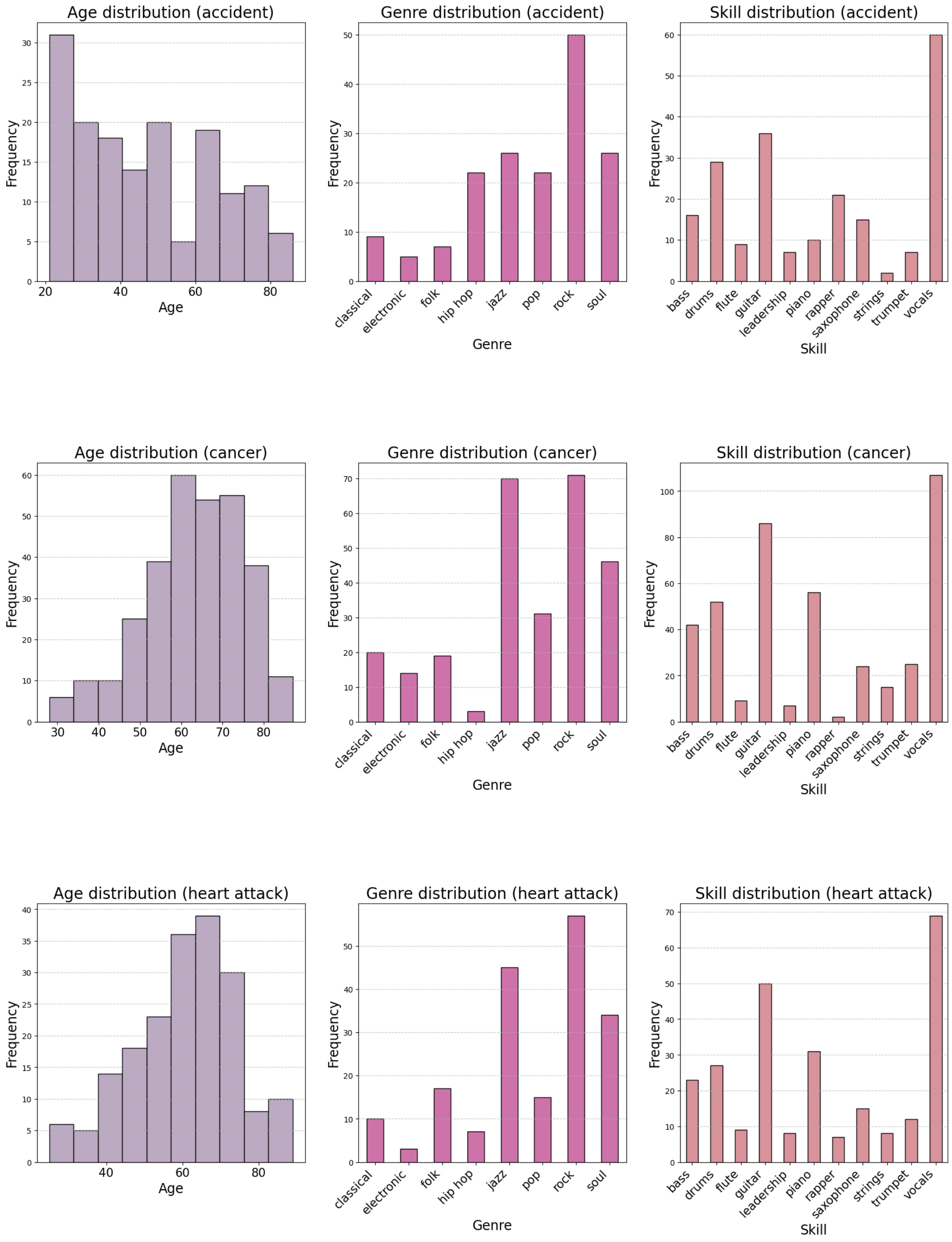

Death place distribution for analysed musicians
Number of deaths
0
200
400
600
800
1000
USA
UK
Other country
Deutschland
Canada
Australia
France
Jamaica
Österreich
Italia
Country

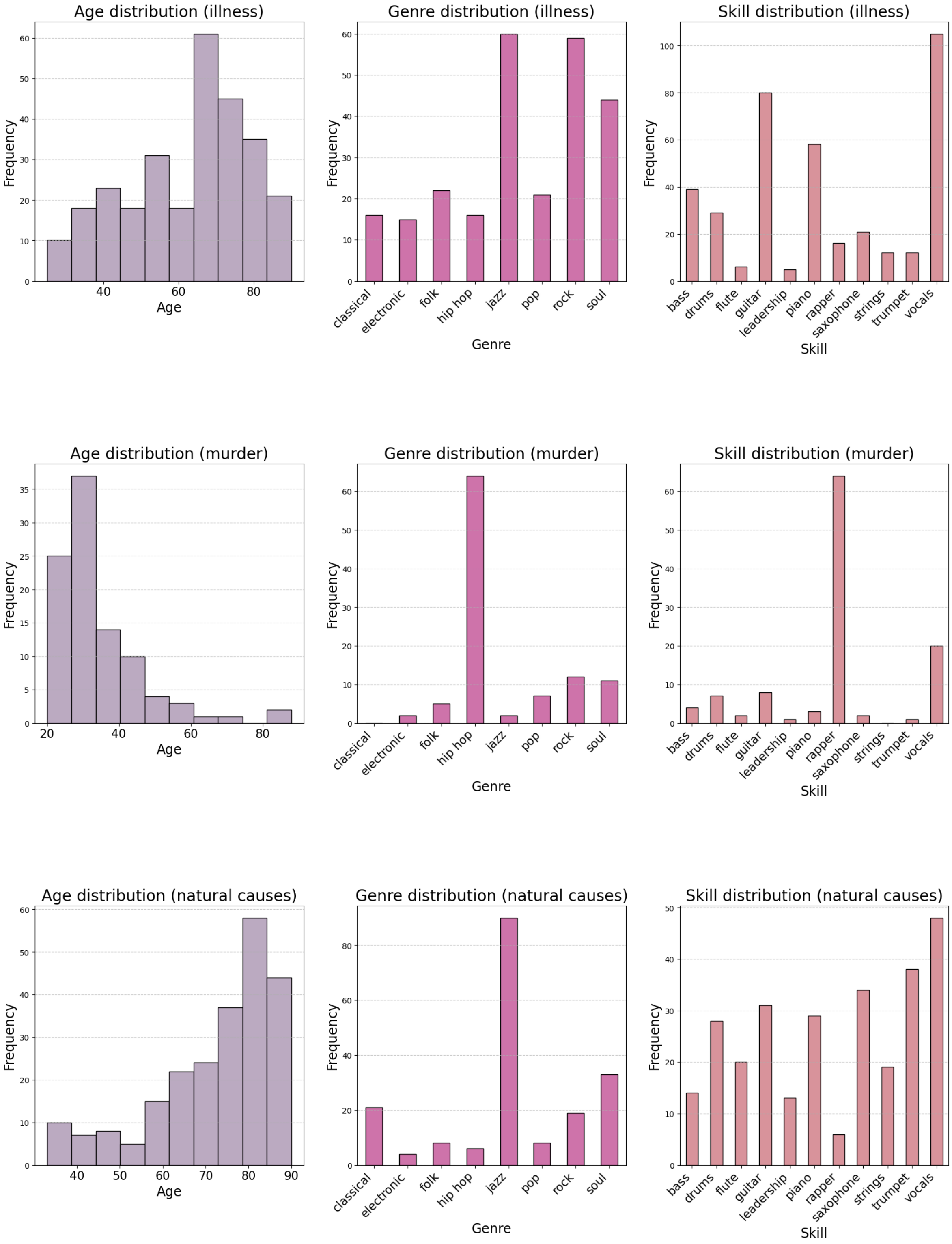

Cause of death distribution for analysed musicians
Number of deaths
0
50
100
150
200
250
300
cancer
illness
natural causes
heart attack
accident
organ failure
murder
overdose
suicide
stroke
Cause of death

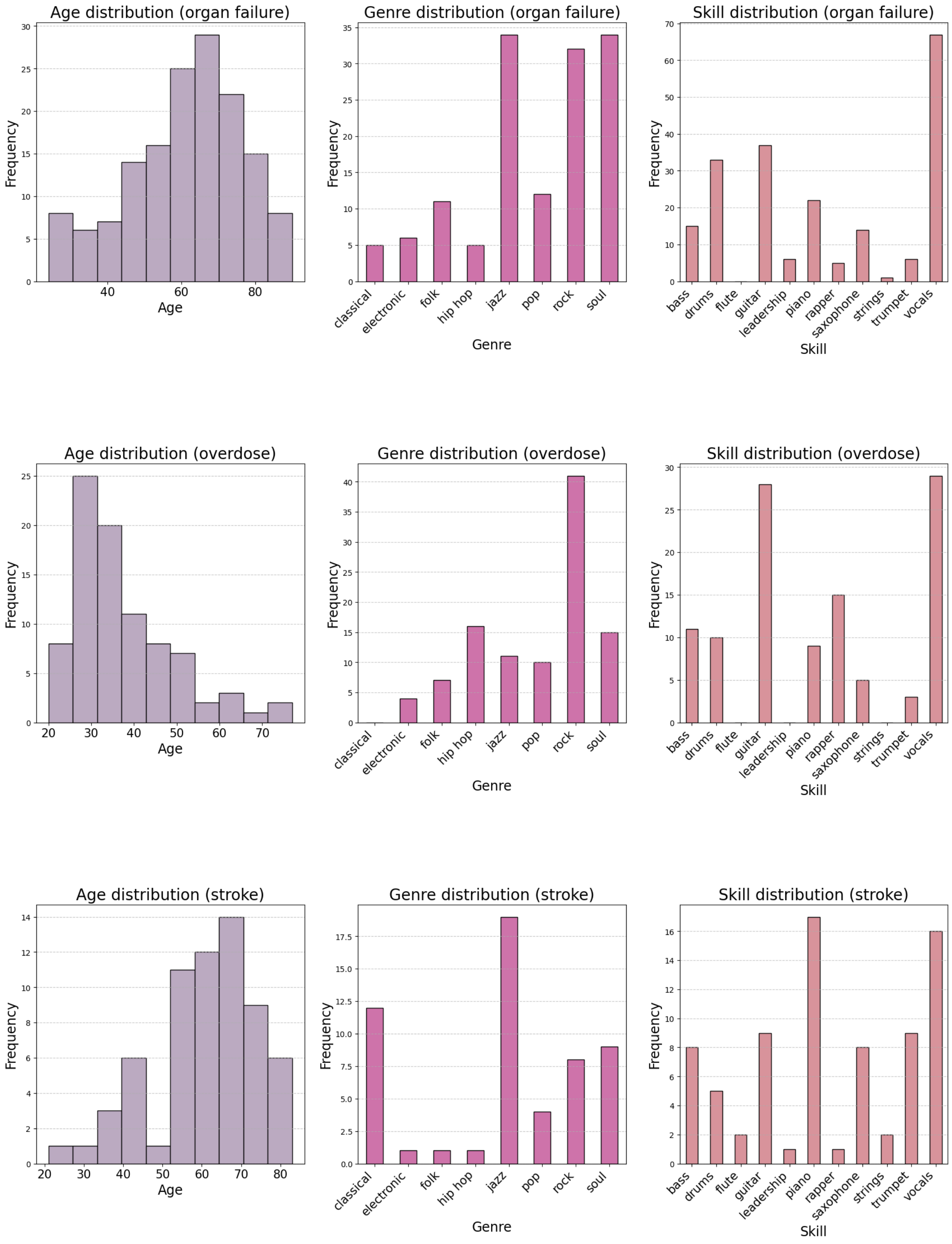

Genre distribution for analysed musicians
Number of deaths
0
50
100
150
200
250
300
350
400
rock
jazz
soul
hip hop
pop
classical
folk
electronic
Genre

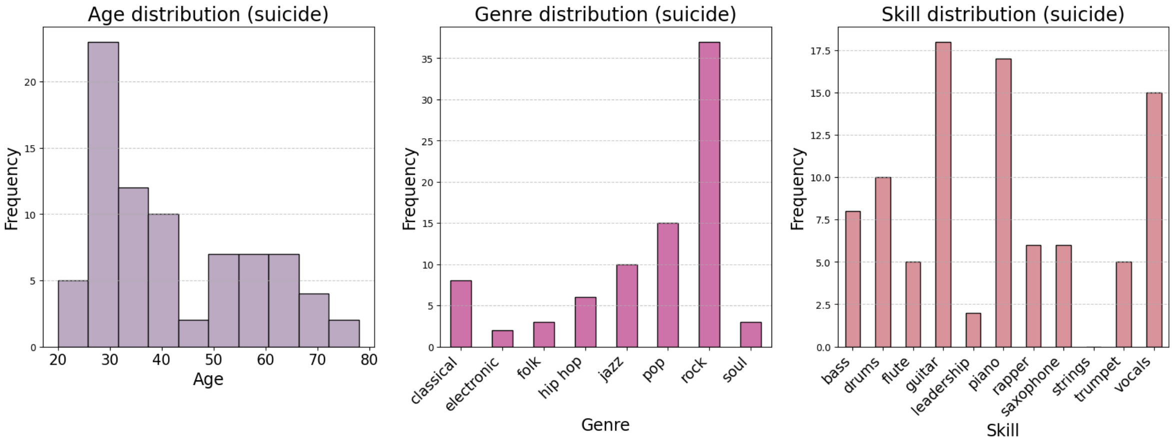

Musical skill distribution for analysed musicians
Number of deaths
0
100
200
300
400
500
vocals
guitar
piano
drums
bass
saxophone
rapper
trumpet
flute
strings
leadership
Skill

### 2.2.2 Individual cause of death analysis


Age distribution (accident)
Genre distribution (accident)
Skill distribution (accident)
Age distribution (cancer)
Genre distribution (cancer)
Skill distribution (cancer)
Age distribution (heart attack)
Genre distribution (heart attack)
Skill distribution (heart attack)


Age distribution (illness)
Frequency
0
10
20
30
40
50
60
40
60
80
Age
Genre distribution (illness)
Frequency
0
10
20
30
40
50
60
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Genre
Skill distribution (illness)
Frequency
0
20
40
60
80
100
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals
Skill
Age distribution (murder)
Frequency
0
5
10
15
20
25
30
35
20
40
60
80
Age
Genre distribution (murder)
Frequency
0
10
20
30
40
50
60
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Genre
Skill distribution (murder)
Frequency
0
10
20
30
40
50
60
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals
Skill
Age distribution (natural causes)
Frequency
0
10
20
30
40
50
60
40
50
60
70
80
90
Age
Genre distribution (natural causes)
Frequency
0
20
40
60
80
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Genre
Skill distribution (natural causes)
Frequency
0
10
20
30
40
50
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals
Skill


Age distribution (organ failure)
Frequency
Age
0
5
10
15
20
25
30
40
60
80
Genre distribution (organ failure)
Frequency
Genre
0
5
10
15
20
25
30
35
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Skill distribution (organ failure)
Frequency
Skill
0
10
20
30
40
50
60
70
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals
Age distribution (overdose)
Frequency
Age
0
5
10
15
20
25
20
30
40
50
60
70
Genre distribution (overdose)
Frequency
Genre
0
5
10
15
20
25
30
35
40
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Skill distribution (overdose)
Frequency
Skill
0
5
10
15
20
25
30
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals
Age distribution (stroke)
Frequency
Age
0
2
4
6
8
10
12
14
20
30
40
50
60
70
80
Genre distribution (stroke)
Frequency
Genre
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
classical
electronic
folk
hip hop
jazz
pop
rock
soul
Skill distribution (stroke)
Frequency
Skill
0
2
4
6
8
10
12
14
16
bass
drums
flute
guitar
leadership
piano
rapper
saxophone
strings
trumpet
vocals



## 2.3 Characteristics of the dataset and conclusions

Analyzing the data led to the decision to specify the following parameters:

- *age* - ranged 20-100
- *birth_place* - primarily USA and UK
- *death_place* - primarily USA and UK
- *cause_of_death* - 10 categories: cancer, illness, natural causes, heart attack, organ failure, accident, murder, overdose, suicide, stroke
- *musical_skills* - 11 categories: vocals, guitar, piano, drums, bass, saxophone, rapper, trumpet, flute, strings, leadership
- *genres* - 8 categories: rock, jazz, soul, pop, hip hop, folk, electronic, classical

After conducting the data exploration, the data on the gender and birth/death locations of the artists had to be discarded during the further analysis. Unfortunately, despite efforts, the acquired dataset predominantly consisted of males (approx. 80% of the data) and artists residing in the USA (approx. 70% of the data). Such an imbalance in the data would not lead to any meaningful results; consequently, these parameters were omitted from the final analysis.

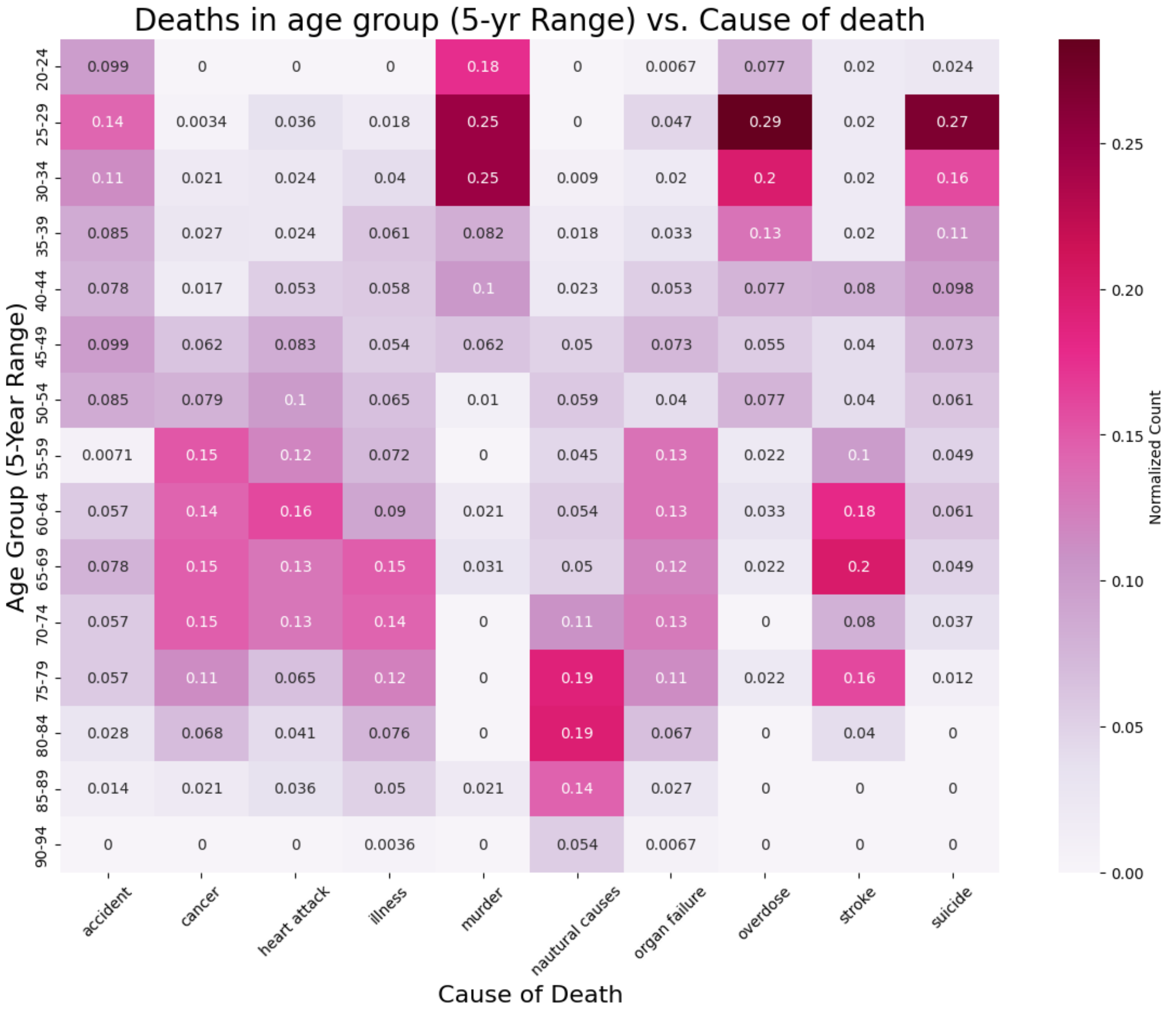
It is worth mentioning that due to the lack of data, not every musician is provided with a genre. It has led to creating a separate dataset, consisting of around 1300 records - it is smaller than the original dataset, consisting of around 2000 records after separating data (point h. mentioned in the data acquisition process).

# 3 Experiments

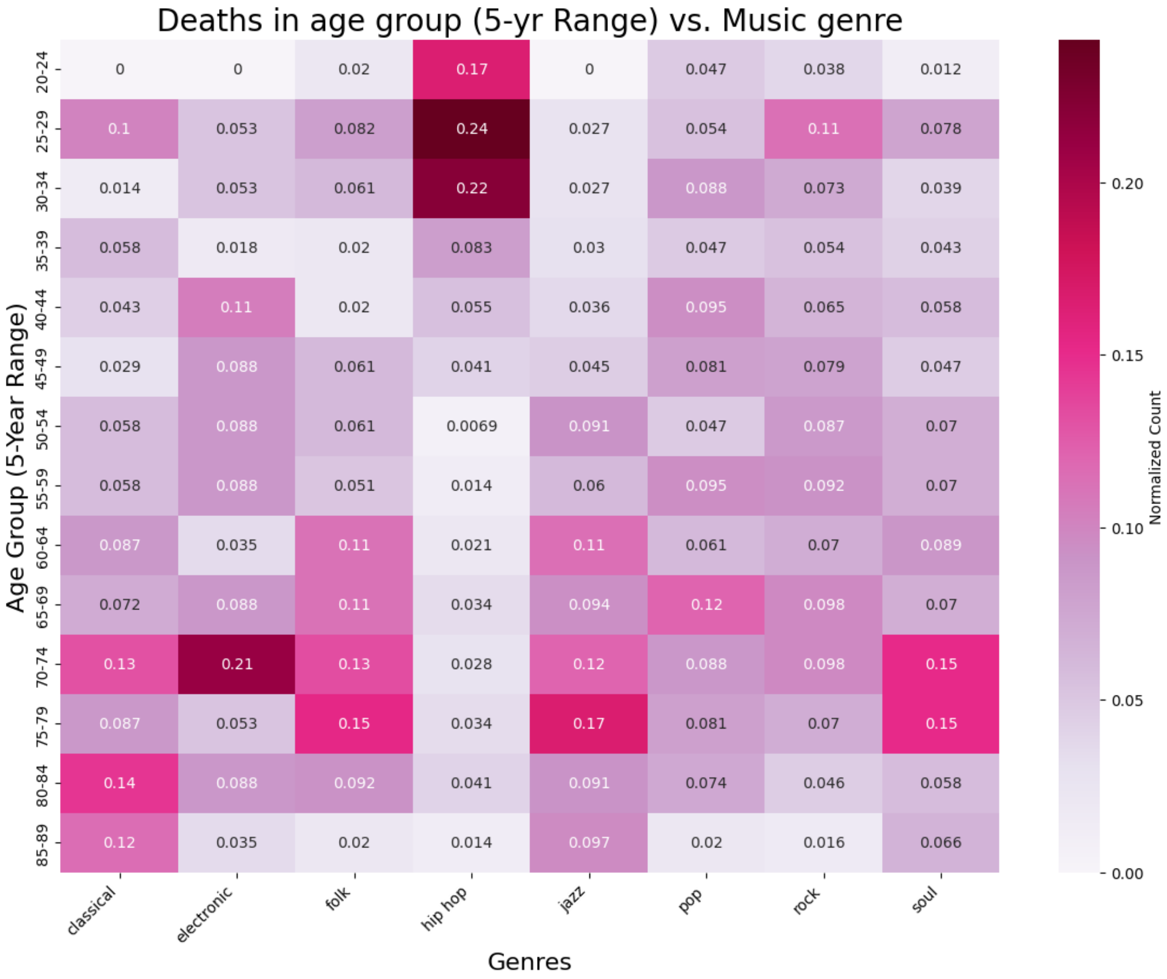
## 3.1 Correlation matrices

The following heatmaps showcase the correlation between the amount of deaths in certain age groups and chosen parameters.

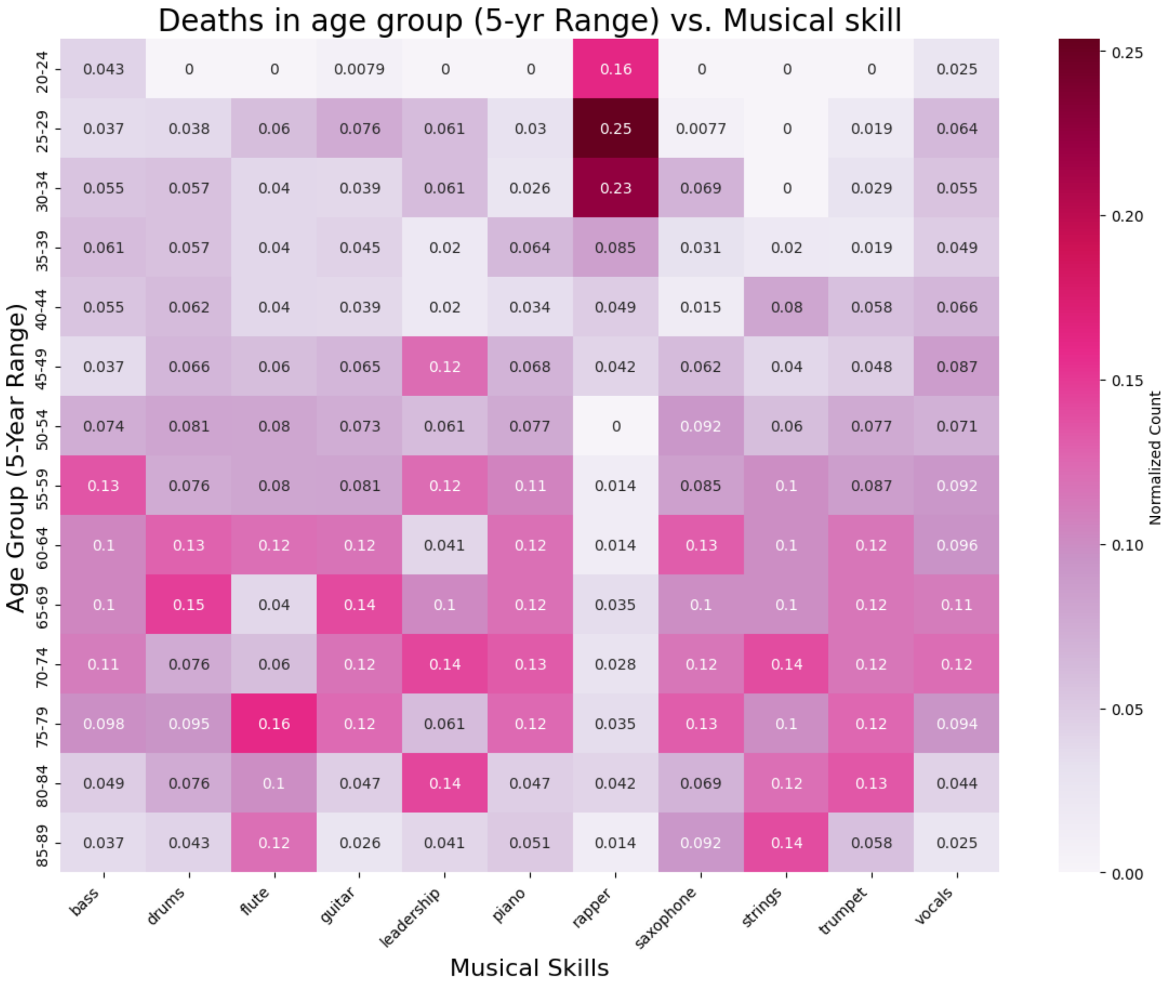
It is important to mention that the counts have been normalised to make up for the data imbalances, by dividing the total amount of the deaths by the given cause (as discovered in the exploratory analysis, there is - for example - much more data on certain death causes, leading to the impairment of the credibility of the results).



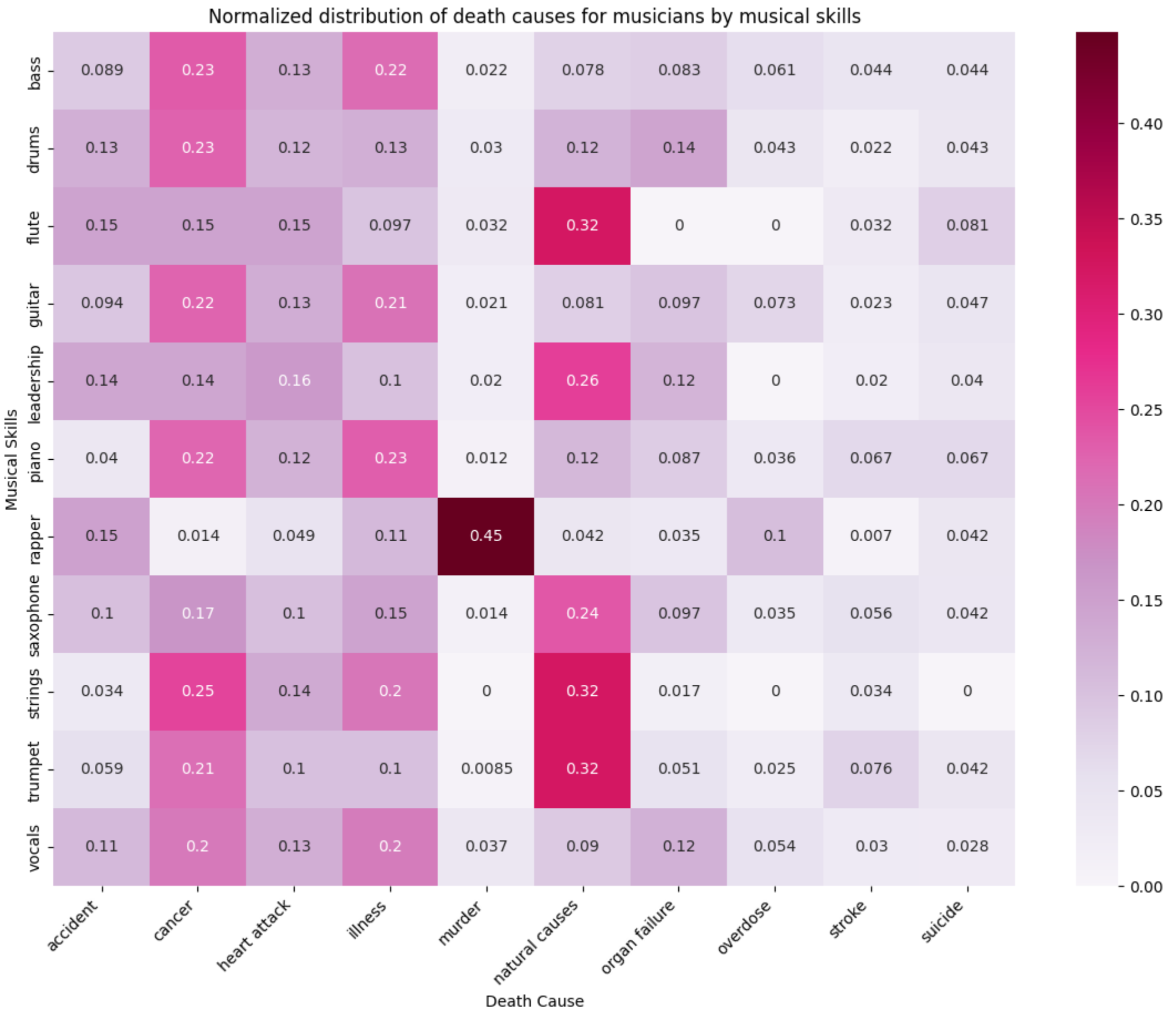
Despite the limitations of the dataset, we can clearly see some patterns forming already - for example, deaths by natural causes occurring primarily in older age groups, whereas deaths by overdose, suicide or murder tend to occur within younger age groups.



Although not too dependable, some correlations can be observed - for example, musicians performing ”calmer”, more sophisticated genres tend to live longer lives, whereas rappers or rock stars on average have shorter lifespans, most likely due to their unhealthy, more impulsive lifestyles.



In this heatmap we can see the clear dataset bias towards rappers who have died at a young age. Some correlations can be observed, but it seems obvious that more high-quality data is needed to draw better conclusions, more rooted in reality.



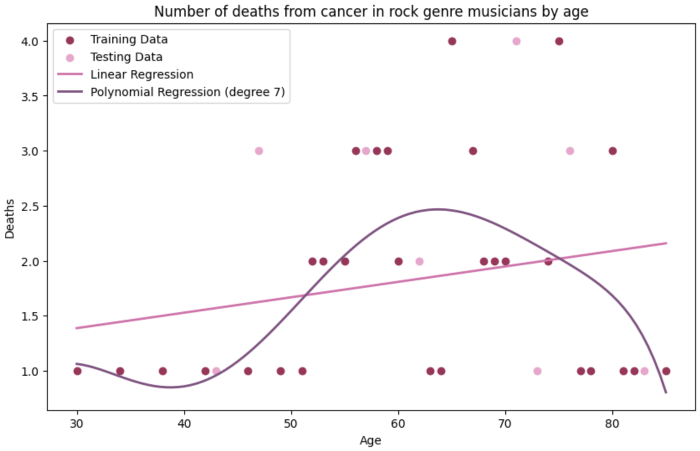
Here we see another clear example of dataset bias, particularly within the murder category.

## 3.2 Linear and polynomial regression

While trying to fit the regression models to the datasets, the train/test split has been taken with a test_size of 0.2 (the models have been trained on 80% of data and tested on the remaining 20% of them).

The first trial of fitting a regression model to the dataset ended up in a failure. The data was too scattered. As a result, the results of the adopted performance indicators - Root Mean Squared Error and Rˆ2 score - were extremely low and therefore unsatisfying.

The results of the original trial are as follows:

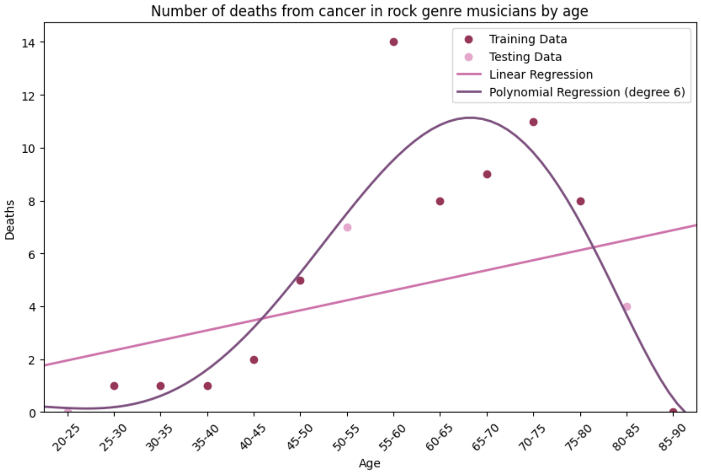


```
Linear Regression - Training Set:  RMSE = 0.9269, R^2 = 0.0467
Linear Regression - Testing Set:  RMSE = 1.1785, R^2 = -0.1697
Polynomial Regression - Training Set:  RMSE = 0.7817, R^2 = 0.322
Polynomial Regression - Testing Set:  RMSE = 1.0855, R^2 = 0.0078
```

Further explorations have led to grouping the data into ranges - taking a sum of deaths for every 5-year range. This has led to a much better fitting of the models, and therefore the performance indicators returned a much more satisfactory results, presented below.

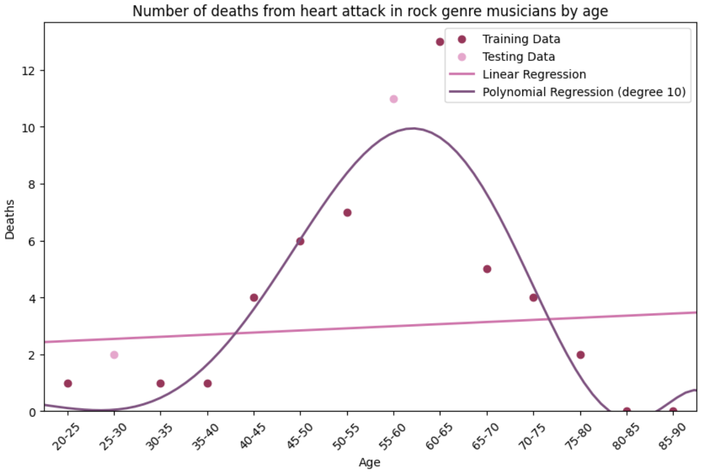
Only a few combinations of death causes, genres and skills were chosen for demonstration purposes.

### 3.2.1 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - cancer, genre - rock



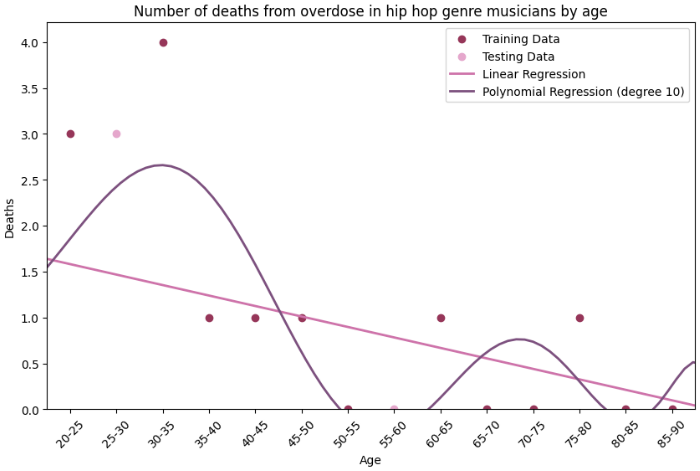
```
Linear Regression - Training Set:  RMSE = 4.1543, R^2 = 0.1859
Linear Regression - Testing Set:  RMSE = 2.1185, R^2 = 0.4834
Polynomial Regression - Training Set:  RMSE = 1.5802, R^2 = 0.8822
Polynomial Regression - Testing Set:  RMSE = 0.4669, R^2 = 0.9749
```

### 3.2.2 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - heart attack, genre - rock



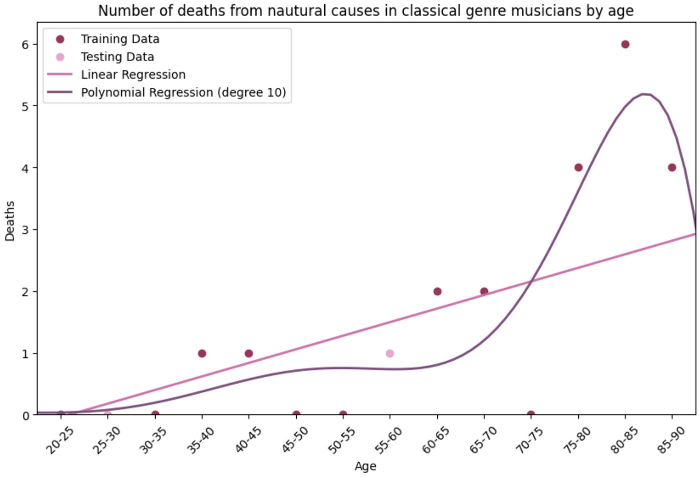
```
Linear Regression - Training Set:  RMSE = 3.5293, R^2 = 0.0111
Linear Regression - Testing Set:  RMSE = 4.3084, R^2 = 0.1027
Polynomial Regression - Training Set:  RMSE = 1.2461, R^2 = 0.8767
Polynomial Regression - Testing Set:  RMSE = 1.2231, R^2 = 0.9277
```

### 3.2.3 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - overdose, genre - hip hop



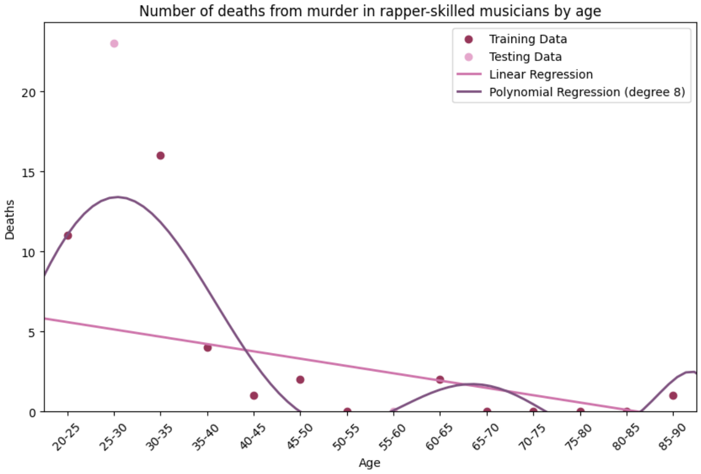
```
Linear Regression - Training Set:  RMSE = 0.9924, R^2 = 0.2513
Linear Regression - Testing Set:  RMSE = 1.6445, R^2 = -0.6026
Polynomial Regression - Training Set:  RMSE = 0.7541, R^2 = 0.5677
Polynomial Regression - Testing Set:  RMSE = 0.5984, R^2 = 0.7878
```

### 3.2.4 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - natural causes, genre - classical



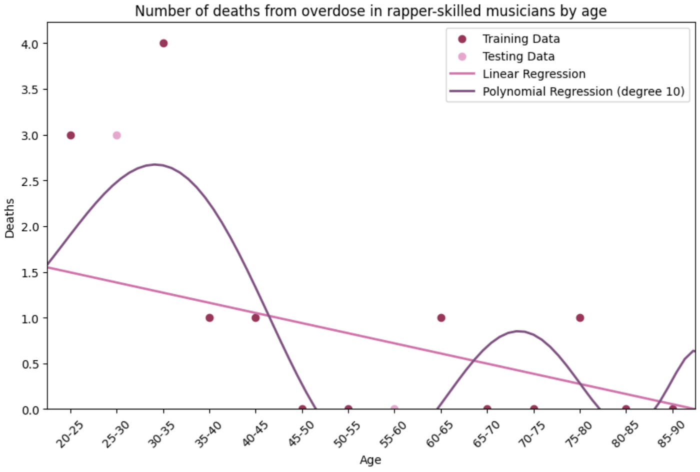
```
Linear Regression - Training Set:  RMSE = 1.4823, R^2 = 0.358
Linear Regression - Testing Set:  RMSE = 0.6359, R^2 = -1.1569
Polynomial Regression - Training Set:  RMSE = 0.8183, R^2 = 0.8043
Polynomial Regression - Testing Set:  RMSE = 0.1455, R^2 = 0.887
```

### 3.2.5 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - murder, skill - rapper



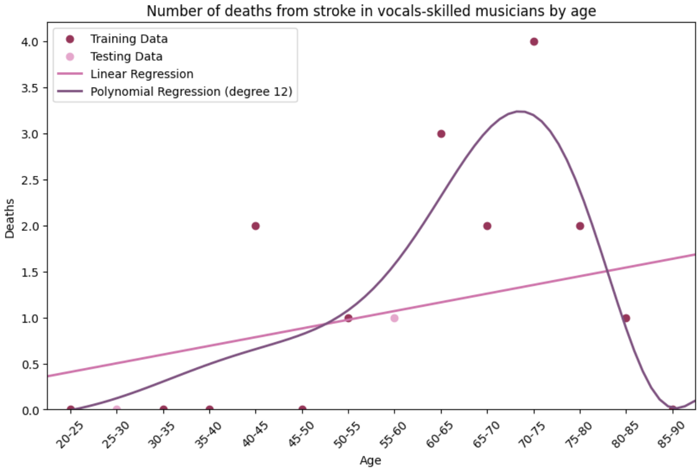
```
Linear Regression - Training Set:  RMSE = 3.8946, R^2 = 0.2587
Linear Regression - Testing Set:  RMSE = 10.3486, R^2 = -0.0797
Polynomial Regression - Training Set:  RMSE = 1.7932, R^2 = 0.8429
Polynomial Regression - Testing Set:  RMSE = 5.0576, R^2 = 0.7421
```

### 3.2.6 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - overdose, skill - rapper



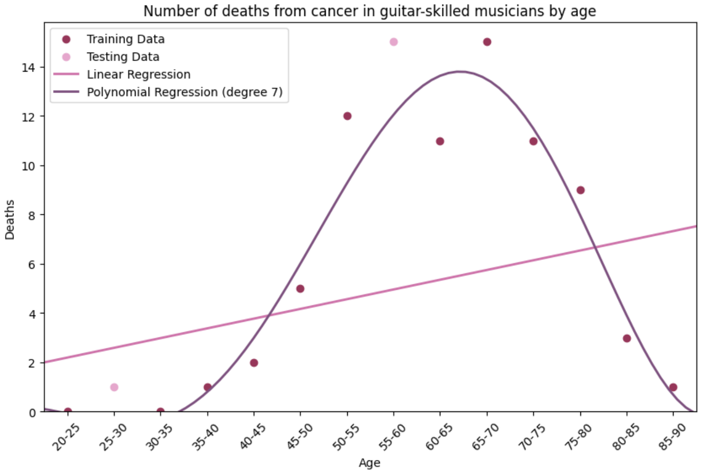
```
Linear Regression - Training Set:  RMSE = 1.0239, R^2 = 0.2292
Linear Regression - Testing Set:  RMSE = 1.5984, R^2 = -0.514
Polynomial Regression - Training Set:  RMSE = 0.7494, R^2 = 0.5871
Polynomial Regression - Testing Set:  RMSE = 0.6037, R^2 = 0.784
```

### 3.2.7 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - stroke, skill - vocals



```
Linear Regression - Training Set:  RMSE = 1.1718, R^2 = 0.1418
Linear Regression - Testing Set:  RMSE = 0.2614, R^2 = 0.6355
Polynomial Regression - Training Set:  RMSE = 0.5866, R^2 = 0.7849
Polynomial Regression - Testing Set:  RMSE = 0.3019, R^2 = 0.5138
```

### 3.2.8 Fitting a linear regression model and a polynomial regression of degree n to the data: death cause - cancer, skill - guitar



```
Linear Regression - Training Set:  RMSE = 4.8028, R^2 = 0.1464
Linear Regression - Testing Set:  RMSE = 5.1186, R^2 = 0.3531
Polynomial Regression - Training Set:  RMSE = 1.1845, R^2 = 0.9481
Polynomial Regression - Testing Set:  RMSE = 1.6843, R^2 = 0.93
```

# 4 Predictive model

## 4.1 Approach description

After the experiments have been done, there was an attempt at creating a predictive model for the chances of dying to various causes, based on age and skills.

After some research, the *SVC* algorithm was chosen to perform this task.

*Support Vector Classification (SVC)* is a supervised learning algorithm used for classification tasks. It belongs to the family of Support Vector Machines (SVMs), which are powerful algorithms used for both classification and regression tasks.

The data has been properly categorized, encoded and split into training and testing sets. Then, the model has been fitted to the data and the probabilities were calculated using the *predict_proba* method. The classification report, containing performance indicators and support (column that indicates the number of instances of each class in the testing dataset), has then been generated using *classification_report* method.

The *SVC* algorithm, *predict_proba* method and *classification_report* method were all taken from the scikit-learn library.

First constructed model predicts the probability of dying to two different death causes, based on age and skills provided by the user. Second constructed model predicts the probability of dying to all available death causes, also based on user input.

The same model can be constructed for predictions based on genres. However, for demonstration purposes, I have chosen the skills dataset, since it has more data to base the model on.

## 4.2 Sample results with performance evaluation scores

### 4.2.1 First model (two death causes)

```
MODEL REPORT:
User input: age=60, musical_skill='drums', death_causes=['cancer', 'overdose']

Probabilities for age 60 and musical skill 'drums':
Probability of dying from cancer: 0.9313
Probability of dying from overdose: 0.0687

Training Accuracy: 0.9065420560747663
Testing Accuracy: 0.897196261682243

Classification Report:
              precision    recall  f1-score   support

      cancer       0.91      0.96      0.93        82
    overdose       0.85      0.68      0.76        25

    accuracy                           0.90       107
   macro avg       0.88      0.82      0.85       107
weighted avg       0.89      0.90      0.89       107



==============================================================================
MODEL REPORT:
User input: age=25, musical_skill='rapper', death_causes=['murder', 'natural causes']

Probabilities for age 25 and musical skill 'rapper':
Probability of dying from murder: 0.9183
Probability of dying from natural causes: 0.0817

Training Accuracy: 0.9233226837060703
Testing Accuracy: 0.8987341772151899

Classification Report:
                precision    recall  f1-score   support

        murder       0.85      0.65      0.73        17
natural causes       0.91      0.97      0.94        62

      accuracy                           0.90        79
     macro avg       0.88      0.81      0.84        79
  weighted avg       0.90      0.90      0.89        79


==============================================================================

MODEL REPORT:
User input: age=40, musical_skill='strings', death_causes=['cancer', 'heart attack']

Probabilities for age 40 and musical skill 'strings':
Probability of dying from cancer: 0.6272
Probability of dying from heart attack: 0.3728

Training Accuracy: 0.6343692870201096
Testing Accuracy: 0.6131386861313869

Classification Report:
              precision    recall  f1-score   support

      cancer       0.61      0.96      0.75        81
heart attack       0.67      0.11      0.18        56

    accuracy                           0.61       137
   macro avg       0.64      0.54      0.47       137
weighted avg       0.63      0.61      0.52       137
```

### 4.2.2 Second model (all death causes)

```
MODEL REPORT:
User input: age=25, musical_skill='rapper'

Probabilities for age 25 and musical skill 'rapper':
Probability of dying from accident: 0.1637
Probability of dying from cancer: 0.015
Probability of dying from heart attack: 0.0565
Probability of dying from illness: 0.0537
Probability of dying from murder: 0.4348
Probability of dying from natural causes: 0.0212
Probability of dying from organ failure: 0.0434
Probability of dying from overdose: 0.1202
Probability of dying from stroke: 0.0163
Probability of dying from suicide: 0.0752

Training Accuracy: 0.27710144927536234
Testing Accuracy: 0.30324074074074076
Classification Report:
                precision    recall  f1-score   support

      accident       0.32      0.52      0.40        42
        cancer       0.26      0.86      0.40        79
  heart attack       1.00      0.00      0.00        51
       illness       0.00      0.00      0.00        88
        murder       0.41      0.50      0.45        24
natural causes       0.47      0.47      0.47        60
 organ failure       1.00      0.00      0.00        38
      overdose       1.00      0.00      0.00        24
        stroke       1.00      0.00      0.00         7
       suicide       0.17      0.05      0.08        19

      accuracy                           0.30       432
     macro avg       0.56      0.24      0.18       432
  weighted avg       0.45      0.30      0.20       432

==============================================================================

MODEL REPORT:
User input: age=70, musical_skill='guitar'

Probabilities for age 70 and musical skill 'guitar':
Probability of dying from accident: 0.0626
Probability of dying from cancer: 0.2549
Probability of dying from heart attack: 0.1498
Probability of dying from illness: 0.2174
Probability of dying from murder: 0.0214
Probability of dying from natural causes: 0.1129
Probability of dying from organ failure: 0.1093
Probability of dying from overdose: 0.0088
Probability of dying from stroke: 0.0452
Probability of dying from suicide: 0.0177

Training Accuracy: 0.27710144927536234
Testing Accuracy: 0.30324074074074076
Classification Report:
                precision    recall  f1-score   support

      accident       0.32      0.52      0.40        42
        cancer       0.26      0.86      0.40        79
  heart attack       1.00      0.00      0.00        51
       illness       0.00      0.00      0.00        88
        murder       0.41      0.50      0.45        24
...
      accuracy                           0.30       432
     macro avg       0.56      0.24      0.18       432
  weighted avg       0.45      0.30      0.20       432

==============================================================================

MODEL REPORT:
User input: age=45, musical_skill='drums'

Probabilities for age 45 and musical skill 'drums':
Probability of dying from accident: 0.2031
Probability of dying from cancer: 0.1767
Probability of dying from heart attack: 0.1116
Probability of dying from illness: 0.0975
Probability of dying from murder: 0.0282
Probability of dying from natural causes: 0.054
Probability of dying from organ failure: 0.134
Probability of dying from overdose: 0.0811
Probability of dying from stroke: 0.043
Probability of dying from suicide: 0.0707

Training Accuracy: 0.27710144927536234
Testing Accuracy: 0.30324074074074076
Classification Report:
                precision    recall  f1-score   support

      accident       0.32      0.52      0.40        42
        cancer       0.26      0.86      0.40        79
  heart attack       1.00      0.00      0.00        51
       illness       0.00      0.00      0.00        88
        murder       0.41      0.50      0.45        24
natural causes       0.47      0.47      0.47        60
 organ failure       1.00      0.00      0.00        38
      overdose       1.00      0.00      0.00        24
        stroke       1.00      0.00      0.00         7
       suicide       0.17      0.05      0.08        19

      accuracy                           0.30       432
```

```
   macro avg       0.56      0.24      0.18       432
weighted avg       0.45      0.30      0.20       432
```

## 5 Conclusions

The most obvious conclusion that can be drawn from the report is the importance of high-quality, varying data. The analysis seems to be very biased in all the areas that are lacking within the acquired dataset. However, with the right techniques, some dependencies and correlations could be clearly observed.

Regarding the linear and polynomial regression experiments, it was clear that the polynomial regression had the upper hand, resulting in much better performance indicator scores. However, in most cases, the degree of the polynomial had to be quite high (6-12). This might be the case due to the nature of the dataset - there isn't enough individual data and it is quite scattered across the graph.

Overall, the death age predictions during the experiments and death cause predictions made by the model have been somewhat of a success.

When it comes to the predictive models, it has been observed that - on average - the first model has much better performance indicator scores. However, upon observation, the second model returns probabilities that seem to be matching with the provided dataset and can be observed in reality.

Age and musical skills appear to influence the likelihood of death causes among musicians, with certain genres and skills potentially associated with higher risks.

The predictive model demonstrates promising performance in forecasting death causes based on available attributes, although further refinement and more data may be necessary to improve accuracy, particularly for less prevalent causes.

Overall, the analyses and predictions provide valuable insights into the factors contributing to mortality among musicians and underscore the importance of considering various demographic and occupational factors in health-related research and interventions.

## 6 References

- *https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python*
- *https://musicbrainz.org/doc/MusicBrainz_API*
- *https://rateyourmusic.com/list/montezuma/all-dead-all-dead-worlds-largest-list-of-deceased-musicians/*
- *https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/*