

Portfolio assignment 101&102

Gemaakt door Martijn Laffort 2159310

Portfolio assignment 101

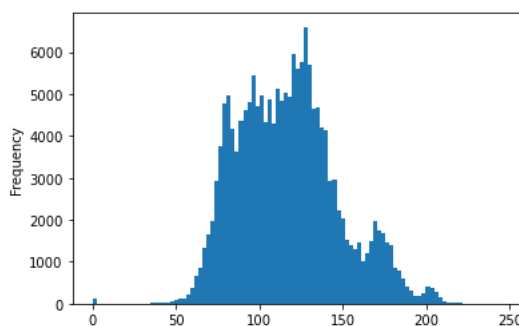
Als ik terug kijk naar portfolio assignment 1 dan moet ik zeggen dat ik niet heel veel zou veranderen. Wel zou ik een paar aspecten willen toevoegen die ik toen nog niet had benoemt. Een toevoeging die ik zeker moet benoemen is de correlatie tussen verschillende data, als je een dataset eenmaal aan het bekijken bent kan je met data science de correlatie vinden tussen verschillende kolommen. Ik had iets soortgelijks gezegd in de eerste opdracht alleen daar had ik het meer over verborgen patronen die je kan onderzoeken en dat is toch verschillend met de correlaties tussen twee of meer stukken data.

Verder ben ik best tevreden met mijn originele antwoord op de vraag wat is data science.

Portfolio assignment 102(101.2)

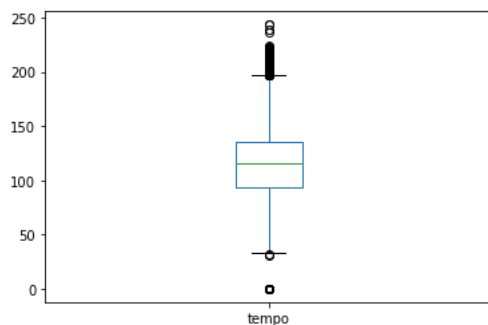
```
In [14]: spotify[column].plot(kind='hist', bins = 100)
```

```
Out[14]: <AxesSubplot:ylabel='Frequency'>
```



```
In [15]: spotify[column].plot(kind='box')
```

```
Out[15]: <AxesSubplot:>
```

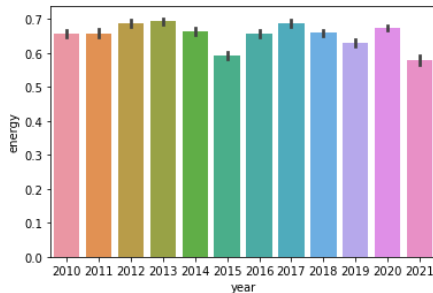


Hier kunnen we zien hoe de tempo van de verschillende liedjes over de afgelopen jaren is verdeelt. In de eerste afbeelding zien we een histogram met een normale verdeling(rechtscheve verdeling zou ik ook nog goedkeuren). Dit houdt in dat het tempo redelijk gelijk is verdeelt en dat zien we in de tweede afbeelding ook, we zien hier dat de

boxplot mooi verdeeld is maar dat er wel veel uitschieters zijn aan de bovenkant van de boxplot(tempo over 200)

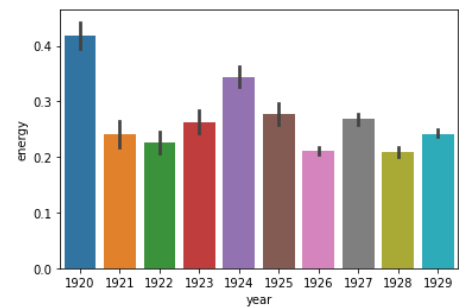
```
In [69]: sns.barplot(y="energy", x="year", data=current_decade)
```

```
Out[69]: <AxesSubplot:xlabel='year', ylabel='energy'>
```



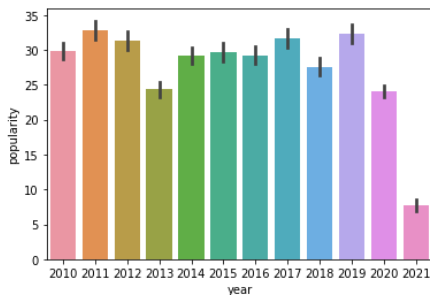
```
In [73]: sns.barplot(y="energy", x="year", data=first_decade)
```

```
Out[73]: <AxesSubplot:xlabel='year', ylabel='energy'>
```



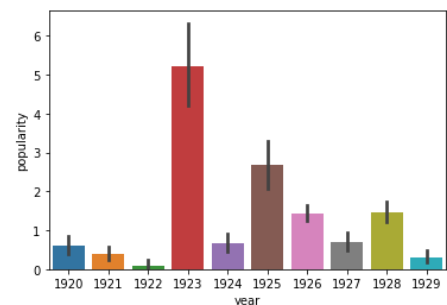
```
In [70]: sns.barplot(y="popularity", x="year", data=current_decade)
```

```
Out[70]: <AxesSubplot:xlabel='year', ylabel='popularity'>
```



```
In [74]: sns.barplot(y="popularity", x="year", data=first_decade)
```

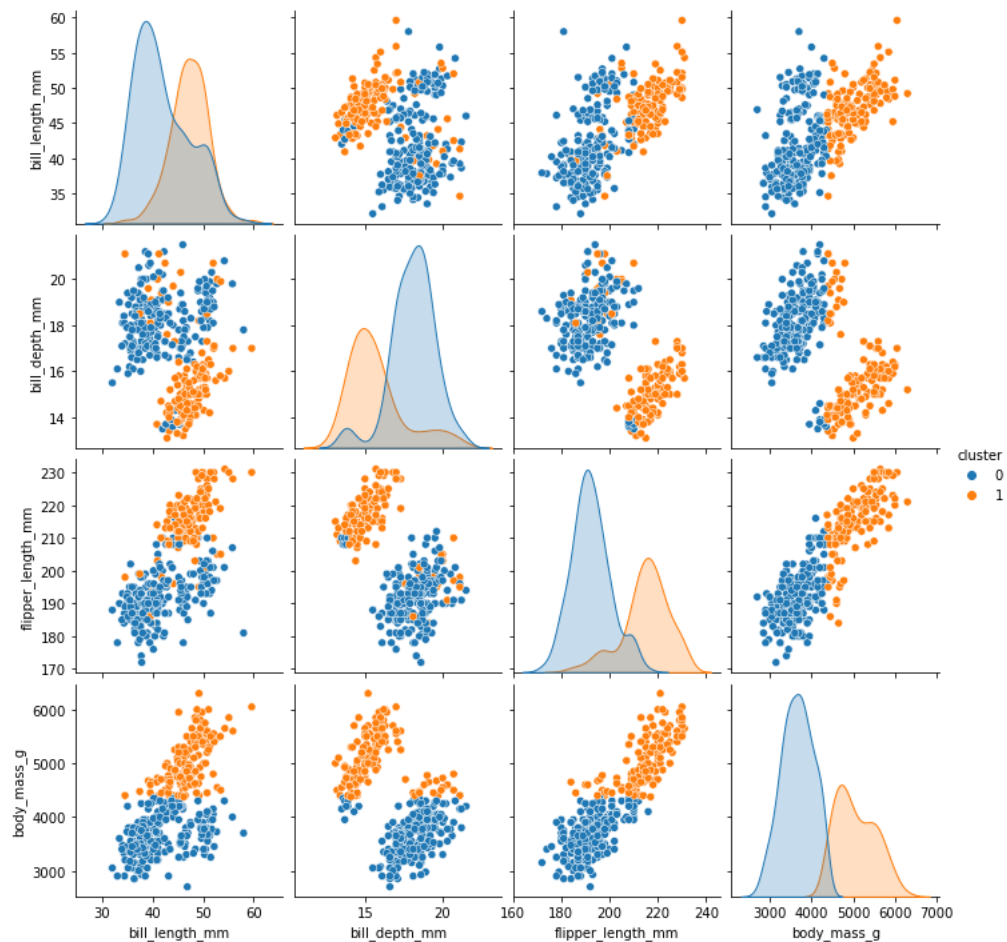
```
Out[74]: <AxesSubplot:xlabel='year', ylabel='popularity'>
```



Bij deze twee afbeeldingen zien we de energie en de populariteit van de liedjes op spotify. Bij de linkse afbeelding hebben we het over de laatste 10 jaar aan liedjes en bij de rechtse afbeelding hebben we het over de eerste 10 jaar aan liedjes. Bij deze barplots kunnen we goed het verschil zien tussen de twee verschillende decennium. We kunnen hier de conclusie trekken dat de liedjes van het afgelopen decennium meer energie hebben en populairder zijn de liedjes van 1920 tot 1930. Een data scientist moet bijvoorbeeld met deze gegevens beslissingen kunnen maken voor de organisatie. Dit onderdeel is weer een goed voorbeeld van BI, bij BI moet de organisaties betere beslissingen maken met de analyse van data.

```
In [10]: sns.pairplot(penguins, hue="cluster")
```

```
Out[10]: <seaborn.axisgrid.PairGrid at 0x195c86f1be0>
```



In deze afbeelding wordt er gebruik gemaakt van clustering, namelijk cluster 0(blauw) en 1(oranje). De clusters zijn ingedeeld op basis van de waarden van de penguins. Deze afbeelding laat zien wat de correlaties zijn tussen de waarden en tussen de clusters. In dit voorbeeld maken we gebruik van een pairplot wat het makkelijk maakt voor de data scientist om een goede analyse te kunnen doen.