Data Humanism

Authors

Nik Bear Brown

What is Data Humanism?

Humanism And Dataism

Humanism originated from an ideological tendency, which tends to care about human personality, emphasizes the maintenance of human dignity, advocates a tolerant secular culture, opposes violence and discrimination, pursues freedom, equality and self-worth, and eventually develops into a philosophy and a theoretical system of worldview.From the above description we can also know that Humanism has an important feature, it holds a view that each person is valuable in himself or herself.

For a long time, humanism was considered advanced and universal. But With the development of computer science and biology, another theory was born -- dataism. Dataism believes that the universe is made up of data streams, and the value of any phenomenon or entity lies in its contribution to data processing. It is the only brand-new value created by human after the creation of humanism. For dataism, the highest value is information streams. Therefore, freedom of information is the highest good.Dataism equates human experience with data patterns, undermine human authority and source of meaning.

Data Humanism

Humanism and Dataism can be said to have many opposites, especially in recognizing personal value. But recently, a new idea was born in the exploration of data-related technologies and means. That's what I'm going to talk about next, Data Humanism. Data humanism believes data is very important, but humans should not be abandoned. We should find a way to make data and humans closely connected. And this method is data visualization. People can better understand your purpose by transforming data into various visual images and feel the connection between this content and their life. The result of data visualization is a combination of art and data science. It is scientific and also takes into account the author's self-expression. It can be said to be a good combination of humanism and data. And also, Data humanism is a philosophy and approach to data analysis and decision-making that emphasizes the importance of understanding the human

context and impact of data. It stresses the need to consider the ethical and societal implications of data usage and ensure that data is used in a way that benefits individuals and society. Data humanism also emphasizes the importance of transparency and accountability in data usage and encourages the active participation of individuals and communities in the data decision-making process.

Data Humanism and Data Visualization

Most of the previous explanations of data humanism in this article stay in terms of history and central idea. These are abstract interpretations, but for data science, data humanism is a concrete way of working. In fact, in my mind, data humanism is actually a kind of advanced data visualization, which adds more humanistic considerations and artistic pursuits than traditional data visualization. So obviously, in order to understand data humanism, we need to understand data visualization first.Next, I will use some pictures and two examples to help you understand data visualization.

Data Visualization and examples

Consider the following matrix, you need to find all the numbers 9 from this matrix.

```
1 2 9 5 3 4 3 0 3 5 8
2 1 3 5 3 5 5 3 4 8 1
9 8 4 3 5 3 4 5 6 3 4
5 1 3 5 1 5 8 5 6 3 9
4 5 2 1 3 1 5 2 1 3 4
0 8 2 1 4 8 6 3 0 8 5
```

This is not a difficult task, but it still takes a lot of time.

So what if the matrix becomes like this

```
1 2 9 5 3 4 3 0 3 5 8
2 1 3 5 3 5 5 3 4 8 1
9 8 4 3 5 3 4 5 6 3 4
5 1 3 5 1 5 8 5 6 3 9
4 5 2 1 3 1 5 2 1 3 4
0 8 2 1 4 8 6 3 0 8 5
```

I marked all the nines with green squares so that you can complete the task in a second.

This is the power of vision, in this case, the matrix is equivalent to the raw data we got, and the green square marks are an easy way to visualize the key points. Of course, in general, we will not use this simple method but use various charts for data visualization.

First, let's take a look at the first data set we will use —— **Covid Deaths and Cases WorldWide**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import regex as re

df=pd.read_csv('../content/sample_data/covid_worldwide.csv', index_col=0)
df.describe()
```

	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Population
count	231	231	225	210	212	213	228
unique	231	231	213	210	181	212	228
top	USA	104,196,861	38	101,322,779	0	78,646	334,805,269
freq	1	1	3	1	8	2	1

The data in this dataset is string, this is not conducive to our analysis, so we convert all data to float type.

```
df['Total Deaths'] = df['Total Deaths'].astype(str).str.replace(',', '').astype(float)
df['Total Cases'] = df['Total Cases'].astype(str).str.replace(',', '').astype(float)
df['Total Recovered'] = df['Total Recovered'].astype(str).str.replace(',', '').astype(float)
df['Active Cases'] = df['Active Cases'].astype(str).str.replace(',', '').astype(float)
df['Total Test'] = df['Total Test'].astype(str).str.replace(',', '').astype(float)
df['Population'] = df['Population'].astype(str).str.replace(',', '').astype(float)
df.head(10)
```

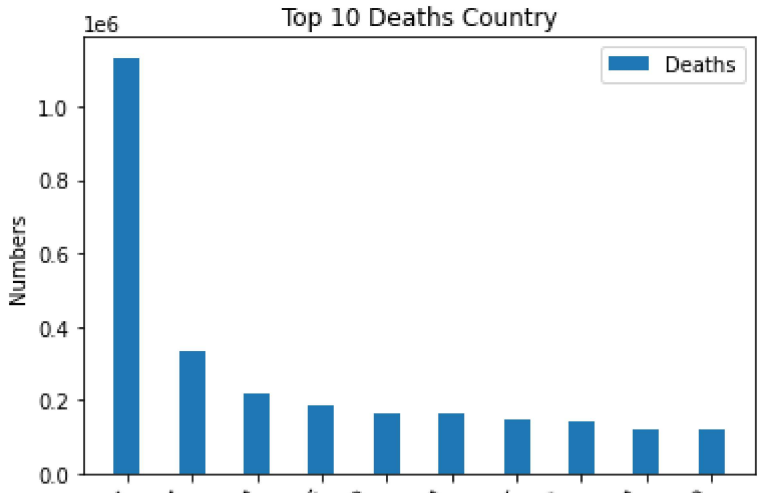
Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Populati
1	USA	104196861.0	1132935.0	101322779.0	1741147.0	1.159833e+09	3.348053e+
2	India	44682784.0	57.0	44150289.0	1755.0	9.152658e+08	1.406632e+
3	France	39524311.0	164233.0	39264546.0	95532.0	2.714902e+08	6.558452e+

Here, if we want to explore the number of deaths in the top ten death countries, we can use the **histogram**. Of course, before drawing, we also need to sort the required data.

```
topDeath = df.sort_values(by="Total Deaths",ascending=False)
topDeath = topDeath.head(10)
topDeath
```

Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Populatior
1	USA	104196861.0	1132935.0	101322779.0	1741147.0	1.159833e+09	334805269.(
19	Mexico	7368252.0	332198.0	6606633.0	429421.0	1.935620e+07	131562772.(
35	Peru	4481621.0	218931.0	4258688.0	4002.0	3.775460e+07	33684208.(
8	Italy	25453789.0	186833.0	25014986.0	251970.0	2.654782e+08	60262770.(
4	Germany	37779833.0	165711.0	37398100.0	216022.0	1.223324e+08	83883596.(
3	France	39524311.0	164233.0	39264546.0	95532.0	2.714902e+08	65584518.(
18	Iran	7564350.0	144749.0	7337549.0	82052.0	5.442078e+07	86022837.(
22	Colombia	6356309.0	142486.0	6179501.0	34322.0	3.695151e+07	51512762.(
21	Poland	6380225.0	118736.0	5335940.0	925549.0	3.811863e+07	37739785.(

```
plt.bar([x for x in range(10)], topDeath['Total Deaths'], label="Deaths", width=0.4)
plt.xticks([x for x in range(10)], topDeath['Country'], rotation=330)
plt.xlabel("Country")
plt.ylabel("Numbers")
plt.title("Top 10 Deaths Country")
plt.legend()
plt.show()
```



This is the simplest histogram, we can intuitively see the number of deaths in each country.

But sometimes, we want to know more. For example, we want to explore the relationship between the number of tests and the number of cases in the countries with top ten highest number of deaths.

In this case, we can use a **two-column histogram**

```
totalWidth=0.8
labelNums=2
barWidth=totalWidth/labelNums
seriesNums=10

plt.bar([x for x in range(seriesNums)], topDeath['Total Test'], label="Test", width=barWidth
plt.bar([x+barWidth for x in range(seriesNums)], topDeath['Total Cases'], label="Cases", wi

plt.xticks([x+barWidth/2*(labelNums-1) for x in range(seriesNums)], topDeath['Country'], ro
plt.xlabel("Country")
plt.ylabel("Numbers")
plt.title("Tests and Cases of Top 10 Deaths Country")
plt.legend()
plt.show()
```

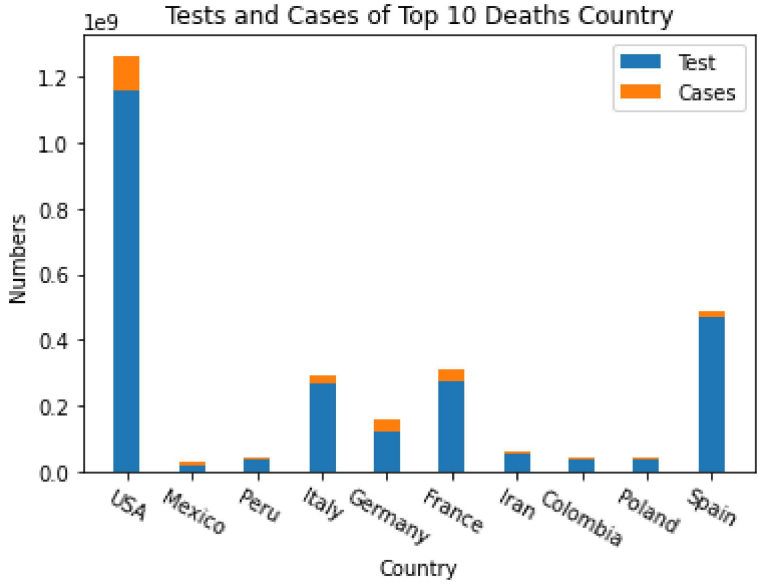
In the figure above, we compare the relationship between the number of tests and the number of cases in the countries with the top ten highest number of deaths.

In the similar way, we can also draw a histogram with three or more columns.

We also have another way to represent the same situation, which is to use a **stacked-histogram**

```
plt.bar([x for x in range(10)], topDeath['Total Test'], label="Test", width=0.4)
plt.bar([x for x in range(10)], topDeath['Total Cases'], label="Cases", width=0.4, bottom=tc

plt.xticks([x for x in range(10)], topDeath['Country'], rotation=330)
plt.xlabel("Country")
plt.ylabel("Numbers")
plt.title("Tests and Cases of Top 10 Deaths Country")
plt.legend()
plt.show()
```

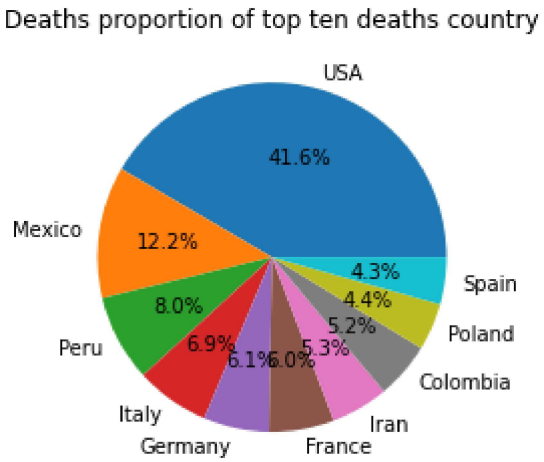


After looking at so many histograms, we can find that histograms have their limitations. The histogram does not have a clear overall concept. Although we can compare different parts of data, but if you want to know the proportion of each different part, you cannot draw conclusions intuitively.

And in this case you can use **pie chart**, let's take the sum of the deaths of the top ten deaths countries as 1 to see what the proportion of each country is.

```
plt.title('Deaths proportion of top ten deaths country')
plt.pie(topDeath['Total Deaths'],autopct='%1.1f%%', labels=topDeath['Country'])
```

```
plt.show()
```



In this example, we explore several basic ways of data visualization, using various methods, that can be used to compare the relationship between one or more aspects of different objects.

Next, let's move on to another example. We're going to illustrate how to use a graph to show the changing trend of a certain property of an object.

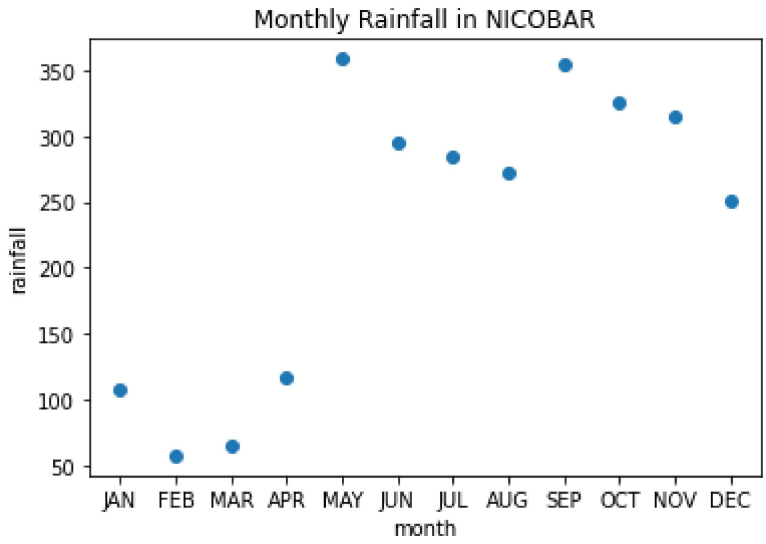
Dataset 2 -- **Rainfall in India**

```
df2=pd.read_csv('/content/sample_data/district wise rainfall normal.csv', index_col=0)
df2
```

In this data set, we know the monthly rainfall for each region of India, and these data have a clear time order.

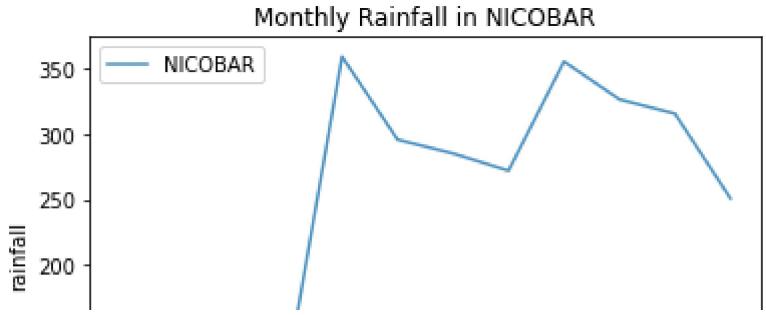
If we want to know the rainfall in the NICOBAR area during the year, we can use a **scatterplot**

```
nico=df2[['DISTRICT', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT',
plt.scatter(['JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT', 'NOV', 'D
plt.xlabel('month')
plt.ylabel('rainfall')
plt.title('Monthly Rainfall in NICOBAR')
plt.show()
```



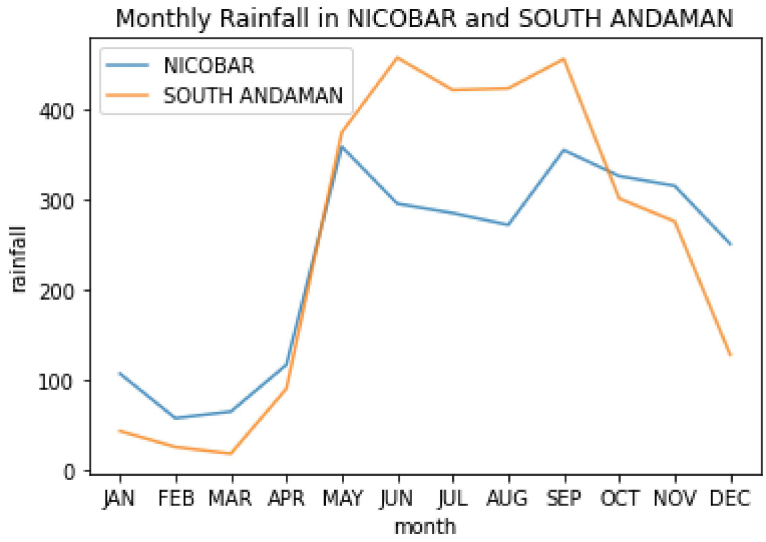
In the picture above, we can indeed vaguely find out the trend of rainfall changes, but it is not really intuitive. Scatter plots are more used in classification. In order to clearly know the trend of rainfall over time, we should use **line chart**.

```
plt.plot(['JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT', 'NOV', 'DEC'
plt.xlabel('month')
plt.ylabel('rainfall')
plt.title('Monthly Rainfall in NICOBAR')
plt.legend()
plt.show()
```



In this way, we can clearly see the trend of rainfall over time. Of course, we can also add data from more regions to compare the trends of them.


```
sa=df2[['DISTRICT', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT', 'N
plt.plot(['JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT', 'NOV', 'DEC'
plt.plot(['JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL' ,'AUG', 'SEP', 'OCT', 'NOV', 'DEC'
plt.xlabel('month')
plt.ylabel('rainfall')
plt.title('Monthly Rainfall in NICOBAR and SOUTH ANDAMAN')
plt.legend()
plt.show()
```



Apparently the rainy season in SOUTH ANDAMAN is more violent

Through the above two examples, we have learned a lot of basic charts for data visualization. After years of development, in fact, the current data visualization have more intuitive and diverse forms.

Advanced Data Visualizaion

↳ *4 cells hidden*

In the End, Data Humanism

In the latest data visualization attempts, data has become a variety of interactive, dynamic and concise elements.

data humanism wants to go a step further, to add some elements of humanism -- artistry.

Each week, people represented their activities in a unique and creative way. They decided to represent each week with a bouquet of flowers, one for each activity. This person assigned a different activity to each flower, including fitness, family time, food, and more. To track their progress, this person came up with a clever system: the more they engaged in an activity, the more petals they earned for the corresponding flower. Their goal was to win up to 5 petals for each flower by completing a week's worth of goals. This method allowed the person to visually see their progress and identify areas for improvement. At the end of each week, they had a beautiful bouquet of flowers representing their activities and accomplishments.

- It is a vision that wants their bouquet to evolve as they grow as individuals. They envisioned that the composition of the flowers would change as they built stronger habits and discovered new interests.
- During the second week of the experiment, the first digital family gathering took place. They found this activity to be very fulfilling and decided to make it a weekly habit. To represent this new activity, this person added a new flower, purple salvia, to their bouquet and began to track its development.
- This person is excited about the possibility of their bouquet growing and developing in this way and is looking forward to seeing the changes in their personal development.

Reference:

[1] Qlik-oss. (n.d.). Sn-mekko-chart - 梅科图. GitHub. https://github.com/qlik-oss/sn-mekko-chart

[2] Zhihu. (n.d.). 什么是梅克图？. Retrieved February 23, 2023, from https://www.zhihu.com/question/52240981

[3] Kaggle. (n.d.). Rainfall in India. Retrieved February 23, 2023, from https://www.kaggle.com/datasets/rajanand/rainfall-in-india?select=district+wise+rainfall+normal.csv

[4] Kaggle. (n.d.). Covid deaths and cases worldwide. Retrieved February 23, 2023, from https://www.kaggle.com/code/finnheaslop/beginner-covid-deaths-and-cases-work-in-prog/notebook

[5] Zhihu. (n.d.). 饼图与其他图形. Retrieved February 23, 2023, from https://zhuanlan.zhihu.com/p/345262150

[6] Zhihu. (n.d.). 什么是数据可视化？. Retrieved February 23, 2023, from https://zhuanlan.zhihu.com/p/162338503

[7] Khowala, D. (2018, June 7). Data Humanism: Visualizing data to connect people with numbers. Devika Khowala. https://devikakhowala.com/data-humanism

[8] Zhihu. (n.d.). 数据可视化. Retrieved February 23, 2023, from https://zhuanlan.zhihu.com/p/439354353

[9] Winkler, L. (2019, July 15). Data Visualization for Humans: How I Turned My Data Into Watercolour Art. Towards Data Science. https://towardsdatascience.com/data-visualization-for-humans-how-i-turned-my-data-into-watercolour-art-651d6acb16a3

License

AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.