data_cleaning

Samuel

2024-04-10

First we have a brief overview of the data set.

```
library(ggplot2)
substance_use <- read.csv("C:/Users/Samuel/Downloads/substance_use.csv")
dim <-dim(substance_use)
sprintf("There is %d row and %d column in the dataset",dim[1],dim[2])
```

```
## [1] "There is 15120 row and 10 column in the dataset"
```

```
head(substance_use)
```

```
##   measure                 location    sex      age                 cause
## 1  Deaths East Asia & Pacific - WB   Male 25 to 29 Alcohol use disorders
## 2  Deaths East Asia & Pacific - WB Female 25 to 29 Alcohol use disorders
## 3  Deaths East Asia & Pacific - WB   Male 30 to 34 Alcohol use disorders
## 4  Deaths East Asia & Pacific - WB Female 30 to 34 Alcohol use disorders
## 5  Deaths East Asia & Pacific - WB   Male 35 to 39 Alcohol use disorders
## 6  Deaths East Asia & Pacific - WB Female 35 to 39 Alcohol use disorders
##    metric year         val       upper       lower
## 1 Percent 1990 0.004355489 0.005574785 0.003579575
## 2 Percent 1990 0.002316023 0.002622133 0.002052042
## 3 Percent 1990 0.006539015 0.007974114 0.005392593
## 4 Percent 1990 0.002667792 0.002950154 0.002417720
## 5 Percent 1990 0.007597508 0.010585770 0.006359210
## 6 Percent 1990 0.002744876 0.003049935 0.002468063
```

Then, we explore the data whether there is null data in the dataset.

```
anyNA(substance_use)
```

```
## [1] FALSE
```

It shows that there is no null data within the data set.

In the following part, we text whether there is any duplication in data set.

```
count=0

for (i in duplicated(substance_use)){
```

```
  if (i){
    count=count+1
  }
}
sprintf("There is %d duplication in the dataset",count)
```

```
## [1] "There is 0 duplication in the dataset"
```

We make a box plot of two diseases prevalence and deaths from all the years, age, sex and location in the data set to explore the distribution relation between them.

```
ggplot(data=substance_use,aes(,x=cause,y=val,fill=measure))+
  geom_boxplot(outliers = FALSE)+
  ggtitle("boxplot of two diseases prevalence and deaths in the whole dataset")
```