Experimental Statistics for Engineers, Descriptive Statistics

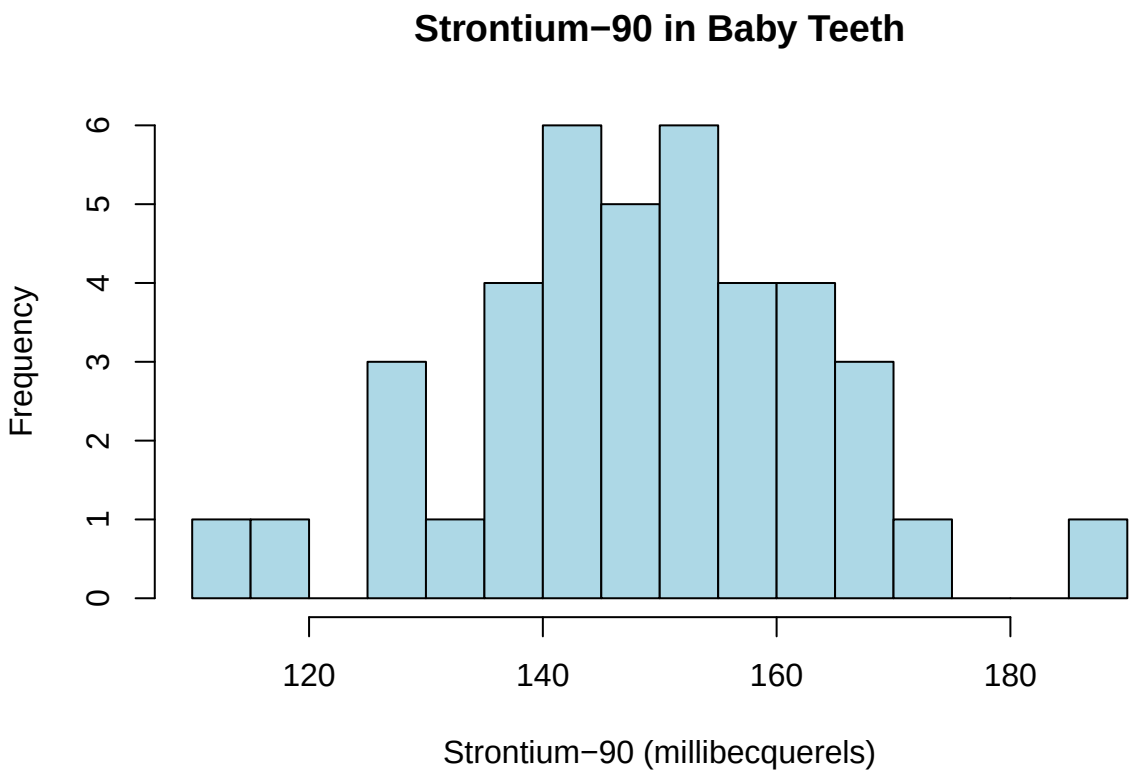
Kamala Dadashova

Problem 1: The amount of strontium-90 (in millibecquerels) of baby teeth obtained from Pennsylvania residents born after 1979 (based on data from "An Unexpected Rise in Strontium-90 in U.S. Deciduous Teeth in the 1990s" by Mangano, et., al., Science of the Total Environment) are listed below.

114, 116, 128, 129, 130, 133, 136, 137, 138, 140, 142, 142, 144, 145, 145, 145, 147, 149, 150, 150, 150, 151, 151, 151, 151, 152, 155, 156, 156, 158, 158, 161, 163, 163, 165, 166, 169, 170, 172, 188

a. Construct a histogram of the data.

```
strontium<-c(114,116, 128, 129, 130, 133, 136, 137, 138, 140, 142, 142, 144, 145, 145, 145, 147, 149,
             150, 150, 150, 151, 151, 151, 151, 152, 155, 156, 156, 158, 158, 161, 163, 163, 165, 166,
             169, 170, 172, 188)
hist(strontium,breaks= 15, main = "Strontium-90 in Baby Teeth",
       xlab="Strontium-90 (millibecquerels)",col="light blue")
```



b. Construct a box plot and five-number summary.

```
summary(strontium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   114.0   141.5   150.0   149.2   158.0   188.0
```

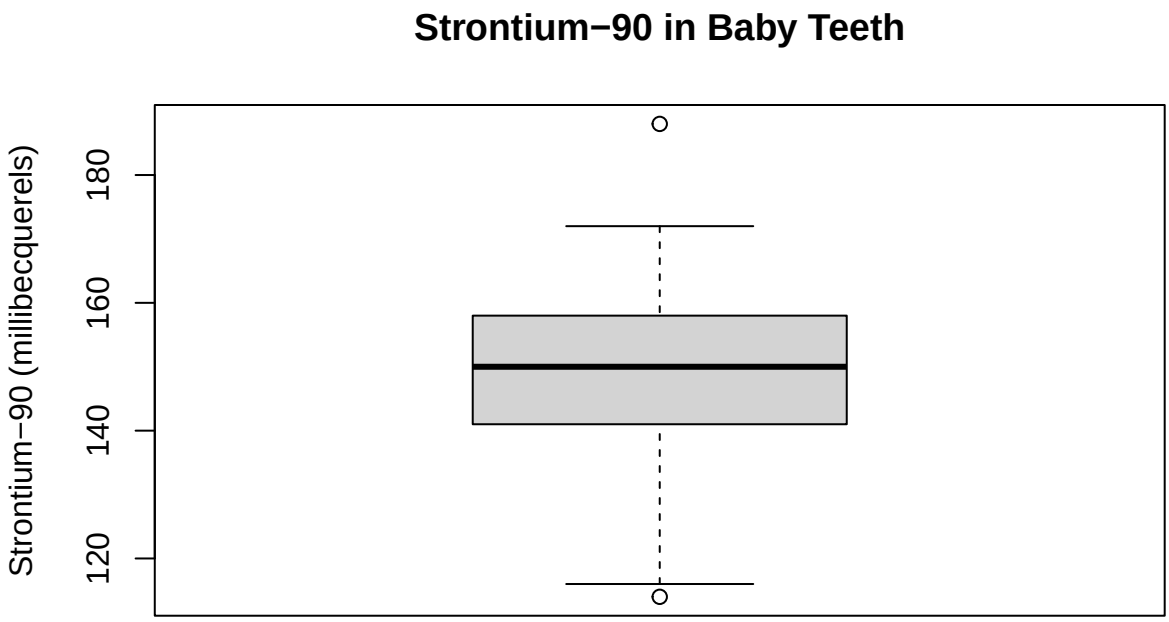
```
interquartile_range<-IQR(strontium)
Q1<-quantile(strontium)[2]
Q3<-quantile(strontium)[4]
upper_fence<-as.numeric(Q3)+1.5*interquartile_range
upper_fence
```

```
## [1] 182.75
```

```
lower_fence<-as.numeric(Q1)-1.5*interquartile_range
lower_fence
```

```
## [1] 116.75
```

```
boxplot(strontium, main="Strontium-90 in Baby Teeth",ylab="Strontium-90 (millibecquerels)")
```



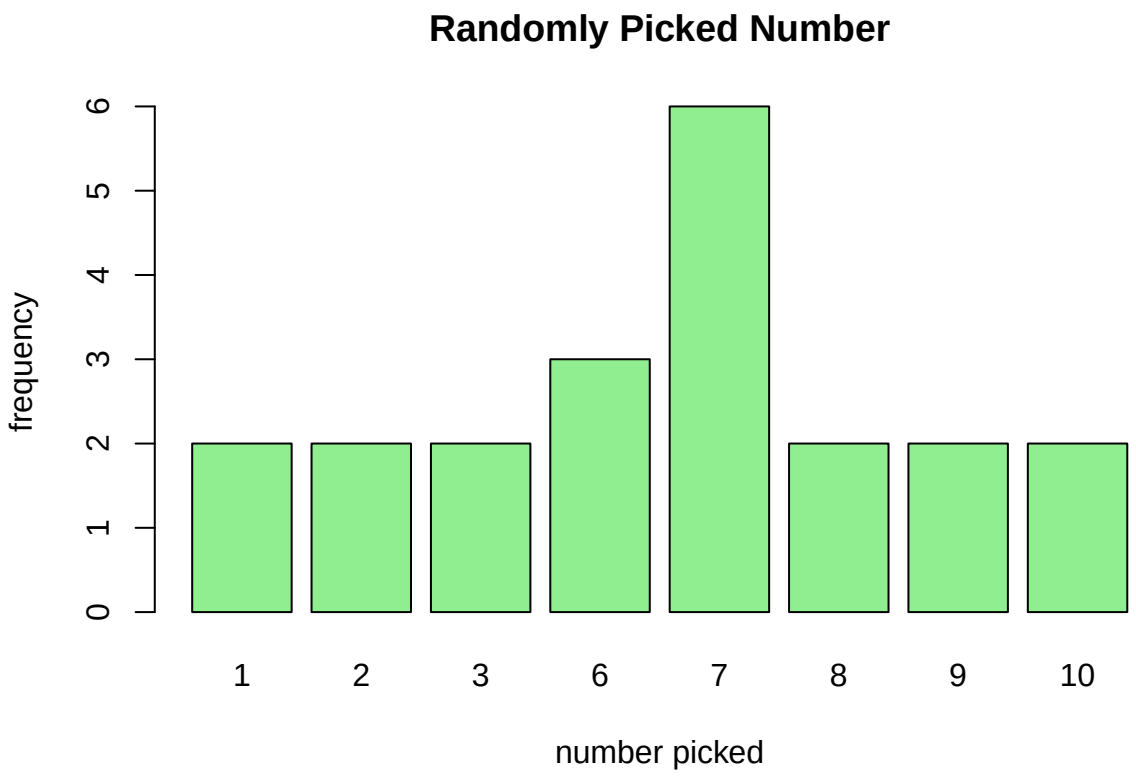
c. What does each plot above tell you about the data, outliers, shape?

By looking at histogram plot, one can observe that the data seems to be symmetrical. We can also confirm this by comparing mean and median which are 149.2 and 150, respectively. Their closeness also prove that histogram is symmetric. Secondly, we observe that there are two potential outliers that one of them is bigger than upper face (182.75) and another is less than lower face(116.75).

Problem 2: In a survey, the participants are asked to choose a "random integer" between 1 and 10. Their answers are: 9, 7, 7, 6, 7, 6, 7, 8, 10, 3, 2, 2, 10, 6, 1, 9, 7, 1 ,7, 8, 3.

a. Create a barchart of this data.

```
random_integer=c(9, 7, 7, 6, 7, 6, 7, 8, 10, 3, 2, 2, 10, 6, 1, 9, 7, 1, 7, 8, 3)
freq=table(random_integer)
barplot(freq,main = "Randomly Picked Number",xlab="number picked", ylab="frequency",col="light green")
```



b. Create a relative frequency table of this data.

```
rel_freq=freq/sum(freq)
rel_freq
```

```
## random_integer
##         1         2         3         6         7         8         9        10
## 0.0952381 0.0952381 0.0952381 0.1428571 0.2857143 0.0952381 0.0952381 0.0952381
```

c. Based on your barchart and relative frequency table, does it appear that asking participants to choose a "random" integer is a good mechanism for generating random numbers? Why or why not?

It is not a good mechanism to select random numbers because there are numbers like 4 and 5 even do not exist in selection. Also, there are numbers like 7 that has higher likelihood being selected comparing to numbers like 2 that has nearly 10% to be chosen.

Problem 3:

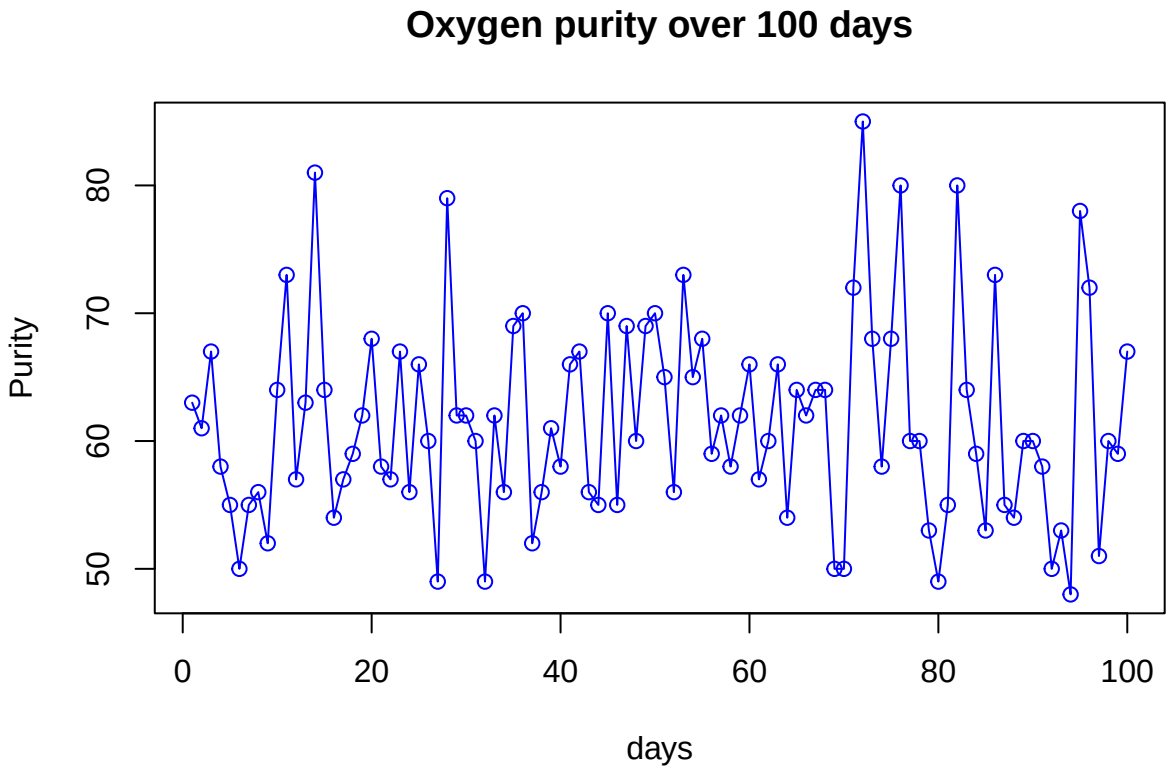
The data in oxygen.csv are 100 measured daily purities of oxygen delivered by a single supplier. The values are in hundredths of a percent purity above 99.00 percent (so 63 means 99.63percent). The supplier's lower specification limit (lowest acceptable value) is 99.50percent. (The data are taken from D. C. Jacobs (1990), Watch Out for Nonnormal Distributions, Chemical Engineering Progress.)

a. Make a time-sequence plot for the data. Are there any obvious time trends? What would be the practical usefulness of early detection of any such time trend?

We are not able to observe any obvious time trend from time-sequence plot. There exist a noticeable number of sudden upwards or downwards spikes. The area between 40ˆth day and 60ˆth day appears to be somewhat steady but we see instantaneous jumping ups and down for other times. Obtaining trends without having that much significant change in purities would be given as an example for early detection of any such time

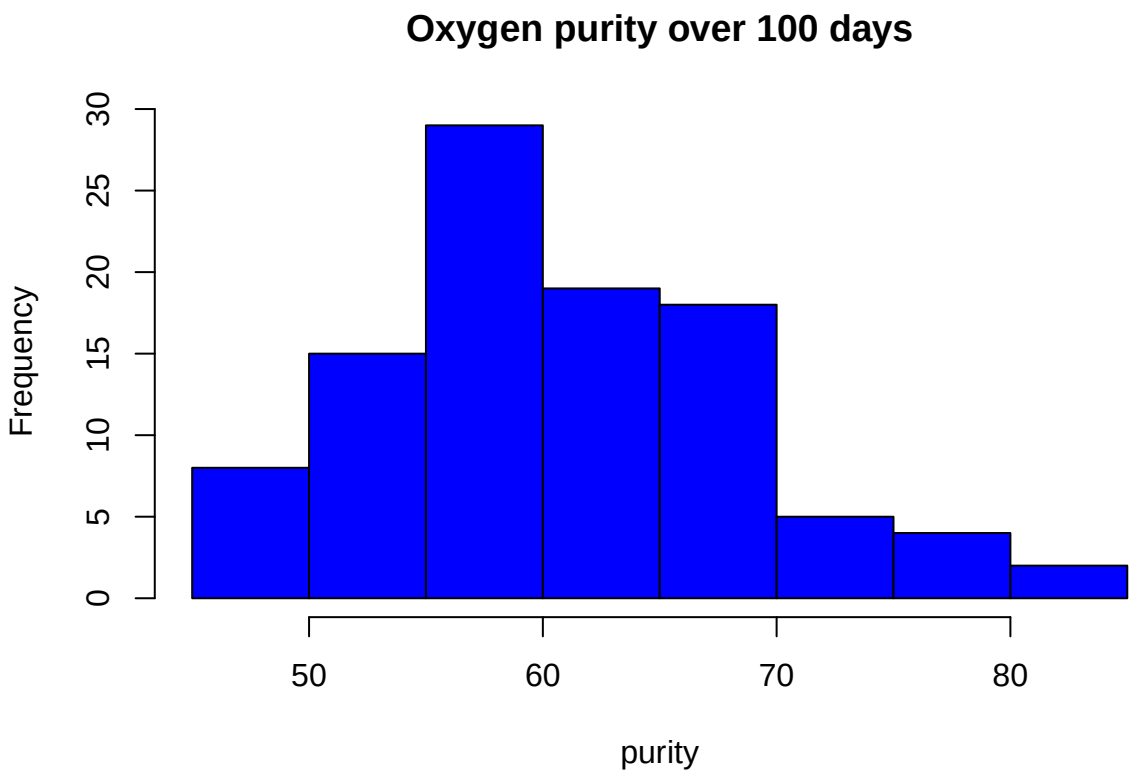
trend. If we were able to detect such trend earlier, we took precautions. For example, if oxygen concentrator filters oxygen at a lower purity level, then it means that the filters need to be cleaned.

```
ts.plot(oxygen$purity, type = "o",col="blue",
        main = "Oxygen purity over 100 days", xlab = "days", ylab = "Purity")
```



b. Ignore the time order of the data collection and represent these data with a histogram. Describe the "shape" of the data, including symmetry and outliers.

```
hist(oxygen$purity, col="blue",main = "Oxygen purity over 100 days", xlab = "purity")
```




```
interquartile_range<-IQR(oxygen$purity)
Q1_purity = quantile(oxygen$purity)[2]
Q3_purity = quantile(oxygen$purity)[4]
upper_fence<-as.numeric(Q3_purity)+1.5*interquartile_range
upper_fence
```

```
## [1] 81.625
```

```
lower_fence<-as.numeric(Q1_purity)-1.5*interquartile_range
lower_fence
```

```
## [1] 40.625
```

The histogram demonstrates that overall data is left skewed towards lower purity levels. Since lower face is 40.625 and there is no data less than that in our histogram that we can say outlier. However, we have some data bigger than 81.625. So, we conclude that we have a few outliers at 80% purity or higher.

c. Calculate the mean, standard deviation, and five-number summary. Which statistics would you choose to summarize this data set and why?

```
mean(oxygen$purity)
```

```
## [1] 61.52
```

```
sd(oxygen$purity)
```

```
## [1] 7.868207
```

```
summary(oxygen$purity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   56.00   60.00   61.52   66.25   85.00
```

Since we have skewed data, the median is a better measure of central tendency than the mean. So, five-number

summary would be more useful to summarize data rather than mean and standard deviation.

Problem 4: From either popular or professional literature, find a graphic that describes a single continuous variable. Turn in a copy of the graphic with your homework and include a complete citation. In a paragraph, describe the context (using who-what-where-when-why-how) for the data. For the graphic to be acceptable, you must be able to identify at least the experimental units ("who") and the variable ("what").

{Source: https://www.climate.gov/media/12885}

The graphic shows yearly surface temperature compared to the 20th-century average from 1880–2020. Blue bars indicate cooler-than-average years; red bars show warmer-than-average years. The experimental units is time (years) and the variable is difference in average surface temperature. The graph illustrates that Earth's temperature has increased by 0.08 Celsius per decade since 1880, and the rate of warming over the past 40 years is more than twice that: 0.18 Celsius per decade since 1981. Secondly, 2020 has been the second warmest year. Averaged across land and ocean, the surface temperature in 2020 was 0.98 Celsius warmer than the twentieth-century average of 13.9 Celsius and 21.19 Celsius warmer than 1880-1900. The ten warmest years has been recorded since 2005.

Problem 5: An experiment involves tossing a pair of dice, one green and one red, and recording the numbers that come up.

a. If x equals the outcome on the green die and y the outcome on the red die, describe the sample space S.

By denoting the number of green dice with x and number of red dice with y. We have

$$S = \{(x, y)\} = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5),$$
$$(2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5),$$
$$(4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$$

b. List the elements corresponding to the event A that the sum is greater than 8.

$$A = \{(3,6), (4,5), (4,6), (5,4), (5,5), (5,6), (6,3), (6,4), (6,5), (6,6)\}.$$

c.List the elements corresponding to the event B that a 2 occurs on either die.

$$B = \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (1,2), (3,2), (4,2), (5,2), (6,2)\}.$$

d.List the elements corresponding to the event C that a number greater than 4 comes up on the green die.

$$C = \{(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}.$$

e.List the elements corresponding to the event $A \cap C$.

$$A \cap C = \{(5,4), (5,5), (5,6), (6,3), (6,4), (6,5), (6,6)\}.$$

f.List the elements corresponding to the event $A \cap B$.

$$A \cap B = \emptyset$$

g.List the elements corresponding to the event $B \cap C$.

$$B \cap C = \{(5, 2), (6, 2)\}.$$

h.Construct a Venn diagram to illustrate the intersections and unions of the events A, B, and C.

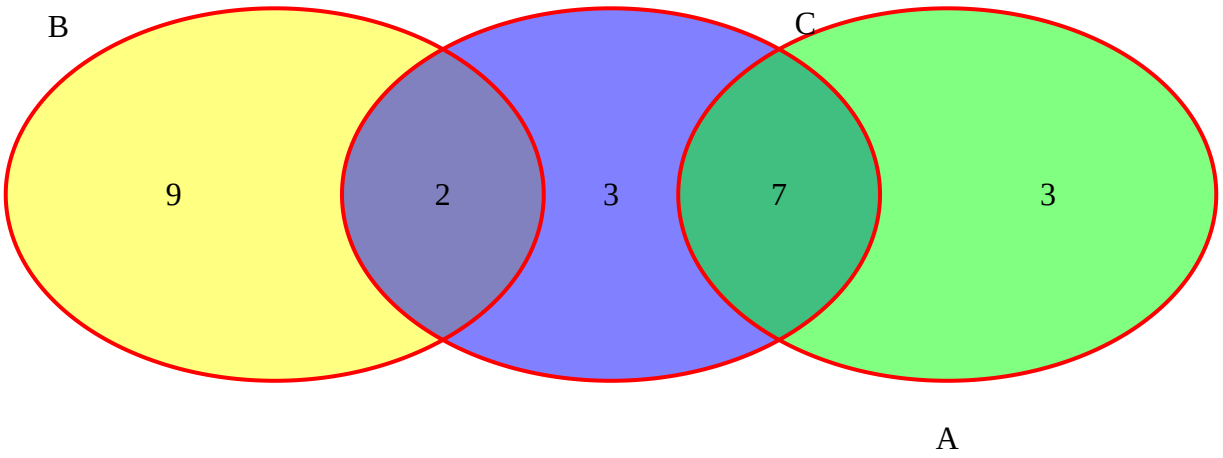
```
# load Venn diagram package
library("VennDiagram")
```

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```

```
# move to new plotting page
grid.newpage()

# create Venn diagram with three sets
draw.triple.venn(area1=10, area2=11, area3=12,
                 n12=0, n23=2, n13=7, n123=0,
                 category=c("A","B","C"),
                 col="Red",fill=c("Green","Yellow","Blue"))
```




```
## (polygon[GRID.polygon.1], polygon[GRID.polygon.2], polygon[GRID.polygon.3], polygon[GRID.polygon.4],
```