assignment_2

Stepan Kropachev

2025-01-02

The Task

Consider the data set pollutants_22.csv.

Using suitable statistical hypotheses tests, check if the three types of pollutants seem to have **significant differences in the means**, with respect to the location of the station (city center/suburbs) and with respect to the year (2000/2020).

Write a short report with the results and with your comments.

The results

First, let's test the data for normality using the Shapiro test.

```
##
##  Shapiro-Wilk normality test
##
## data:  data$PM10
## W = 0.99159, p-value = 0.6384
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Ozone
## W = 0.99337, p-value = 0.8122
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$NO2
## W = 0.9947, p-value = 0.9177
```

As we see, all the variables are normally distributed. And we can apply test for investigating whether the three types of pollutants seems to have significant difference in the means. For this purpose we can use:

- T-test to analyse the pollutants separately (but there's a high risk of Type 1 error);
- MANOVA and ANOVA for multivariate analysis;

Since the data is distibuted normally, let's apply the MANOVA test.

```
##             Df   Wilks approx F num Df den Df Pr(>F)
## center       1 0.95383  1.98442      3    123 0.1198
## year_group   1 0.98181  0.75953      3    123 0.5189
## Residuals  125
```

Here we see, that there's no significant multivariate effect of location on the combined pollutant levels. There's no significant multivariate effect of year on the combined pollutant levels, either.

Let's apply the ANOVA to investigate the variables individually.

```
##  Response PM10 :
##              Df Sum Sq Mean Sq F value Pr(>F)
## center        1      4    4.27  0.0026 0.9591
## year_group    1    555  555.32  0.3433 0.5590
## Residuals   125 202173 1617.38
##
##  Response Ozone :
##              Df Sum Sq Mean Sq F value Pr(>F)
## center        1    336  335.74  0.5827 0.4467
## year_group    1    191  191.26  0.3319 0.5656
## Residuals   125  72022  576.18
##
##  Response NO2 :
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## center        1   420.7  420.74  5.0431 0.02648 *
## year_group    1   169.8  169.83  2.0356 0.15615
## Residuals   125 10428.8   83.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of three pollutants (**PM10**, **Ozone**, and **NO2**) across different locations and years reveals that when considered together through MANOVA, there are no significant multivariate effects of either location or year.

Follow-up univariate ANOVAs show that only **NO2** exhibits a significant difference between city center and suburbs. Neither **PM10** nor **Ozone** show significant differences across locations or years. This suggests that while most pollutant levels remain consistent across space and time, **NO2** concentrations vary significantly depending on the location of measurement.