Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

Module 3 - Data Exploration and Analysis

Mohamad Lukman Aidid bin Mohd Yusoff

2023-10-30

Lesson 1: Testing assumptions for statistical analysis

1. Download the dataset "obis_red_list_filtered_1000.csv" if it's not already done. Explore the structure of the dataset using the str() function. If necessary, change the class of your variable using the as.factor() function.
2. You can create a graph of the temperature data distribution using the hist(), qqnorm() and qqline() functions. What can you see? What assumptions can you make about the normality of the data?
3. Test the normality of the temperature data using the shapiro.test() function. What is the p-value of the normality test? What can you conclude?
4. Test for homoscedasticity for temperature and country using the leveneTest() function in the car package. What is the p-value of the homoscedasticity test? What is your conclusion?
5. Use dplyr's mutate() function to create a new column that is the **logarithm** of the original temperature column. You can then re-run the normality and homoscedasticity tests. Does the transformation improve the temperature data?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(stats)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
```

```
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
setwd('C:/Users/Administrator/Desktop/R/')
```

Question 1:

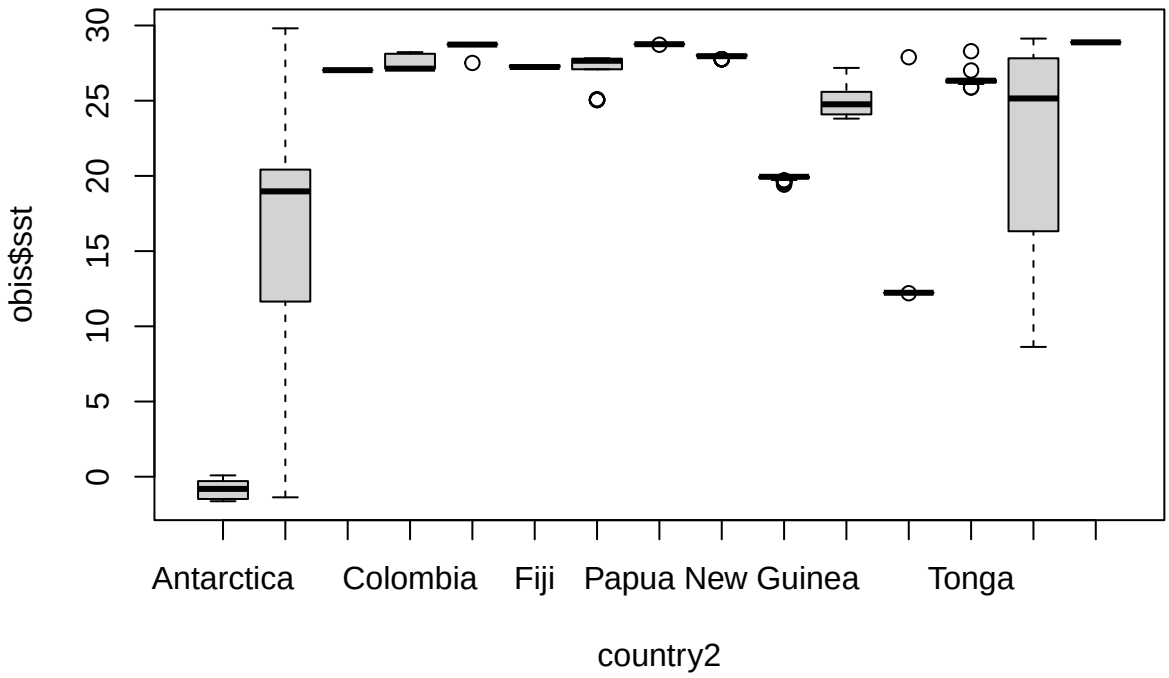
```
obis <- read.csv("C:/Users/Administrator/Desktop/R/obis_red_list_filtered_1000.csv")

head(obis)
```

```
##          scientificName date_year          family minimumDepthInMeters
## 1 Balaenoptera physalus      2003 Balaenopteridae                    0
## 2 Balaenoptera physalus      2003 Balaenopteridae                    0
## 3 Balaenoptera physalus      2003 Balaenopteridae                    0
## 4 Balaenoptera physalus      2003 Balaenopteridae                    0
## 5 Balaenoptera physalus      2003 Balaenopteridae                    0
## 6 Balaenoptera physalus      2002 Balaenopteridae                    0
##   shoredistance   sst   sss individualCount    country status
## 1        182964 -1.47 34.03               2 Antarctica     VU
## 2        135623 -1.58 34.01               2 Antarctica     VU
## 3        138638 -1.58 34.01               9 Antarctica     VU
## 4         77966 -1.57 34.06               4 Antarctica     VU
## 5        141441 -1.59 34.02               3 Antarctica     VU
## 6        -14124 -1.43 33.71               3 Antarctica     VU
```

```
str(obis)
```

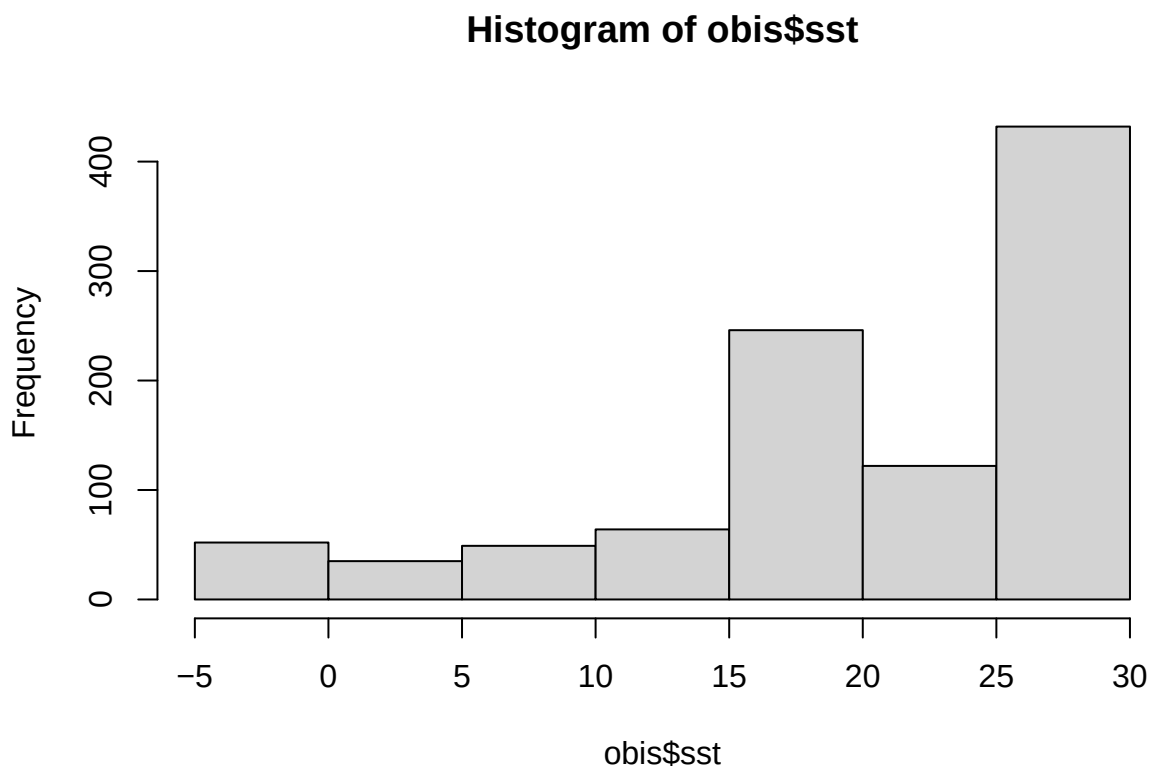
```
## 'data.frame':    1000 obs. of  10 variables:
##  $ scientificName      : chr  "Balaenoptera physalus" "Balaenoptera physalus" "Balaenoptera physalus
##  $ date_year           : int  2003 2003 2003 2003 2003 2002 2003 2003 2003 2020 ...
##  $ family              : chr  "Balaenopteridae" "Balaenopteridae" "Balaenopteridae" "Balaenopteridae
##  $ minimumDepthInMeters: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ shoredistance       : int  182964 135623 138638 77966 141441 -14124 727065 184171 144748 478287 .
##  $ sst                 : num  -1.47 -1.58 -1.58 -1.57 -1.59 -1.43 -0.51 -1.48 -1.55 0.35 ...
##  $ sss                 : num  34 34 34 34.1 34 ...
##  $ individualCount     : num  2 2 9 4 3 3 3 6 6 8 ...
##  $ country             : chr  "Antarctica" "Antarctica" "Antarctica" "Antarctica" ...
##  $ status              : chr  "VU" "VU" "VU" "VU" ...
```

```
country2 <- as.factor(obis$country)
plot(obis$sst~country2)
```

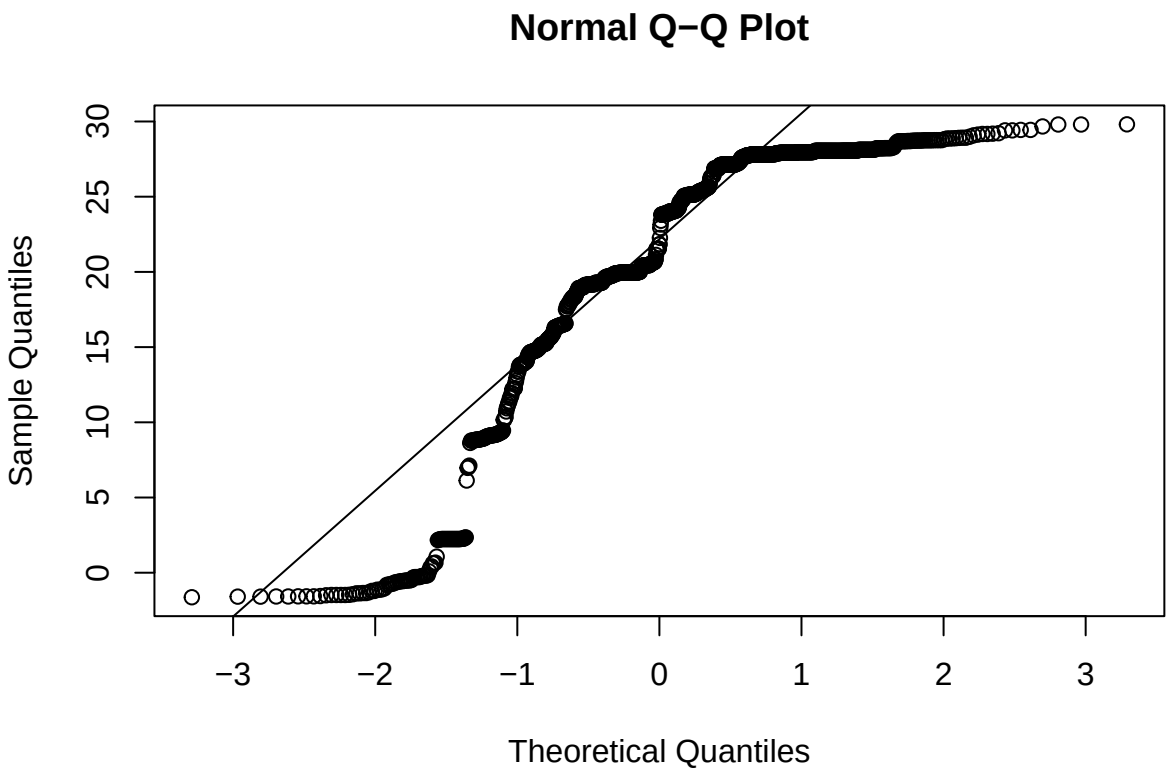

obis$sst
0
5
10
15
20
25
30
Antarctica
Colombia
Fiji
Papua New Guinea
Tonga
country2

Question 2:

```
hist(obis$sst)
```




```
qqnorm(obis$sst)
qqline(obis$sst)
```



From the histogram and Q-Q Plot, it is observed that the data for temperature is not normally distributed.

Question 3:

```
shapiro.test(obis$sst)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  obis$sst
## W = 0.84335, p-value < 2.2e-16
```

p-value is much less than 0.05, thus it is not normally distributed.

Question 4:

```
leveneTest(sst ~ country, obis)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group  14  22.372 < 2.2e-16 ***
##       985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

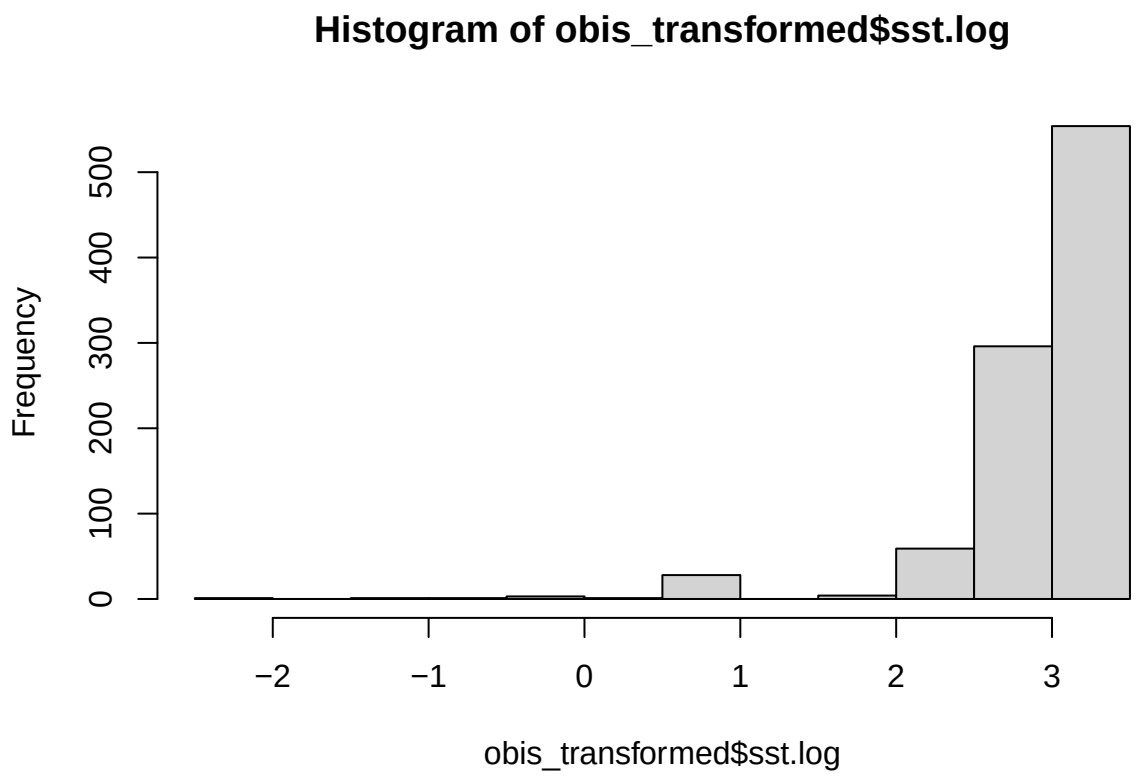
p-value for the Levene test is also much less than 0.05, thus it is not homoscedastic.

Question 5:

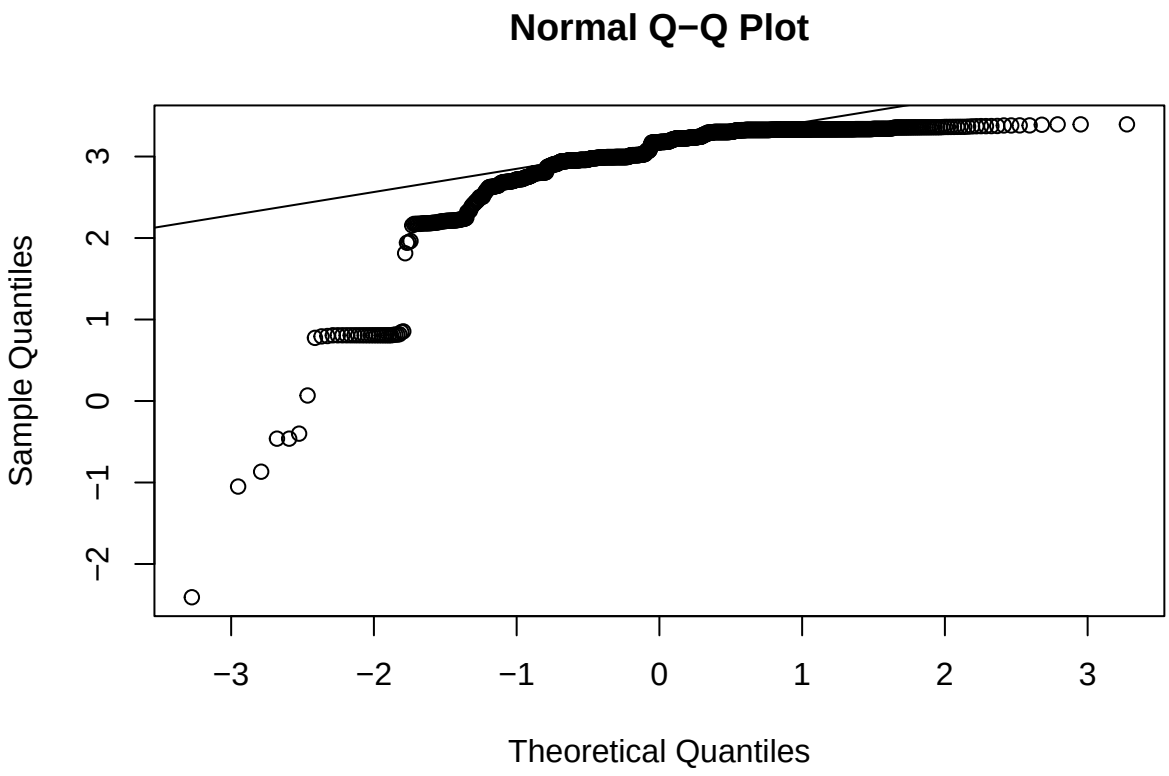
```
obis_transformed <- mutate(obis, sst.log = log(sst))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `sst.log = log(sst)`.
## Caused by warning in `log()`:
## ! NaNs produced
```

```
hist(obis_transformed$sst.log)
```




```
qqnorm(obis_transformed$sst.log)
qqline(obis_transformed$sst.log)
```




```
shapiro.test(obis_transformed$sst.log)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  obis_transformed$sst.log
## W = 0.60476, p-value < 2.2e-16
```

```
leveneTest(sst.log ~ country, obis_transformed)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group  14  11.145 < 2.2e-16 ***
##       933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For both the Shapiro-Wilk test and Levene test, the p-values are not improved when using the log transformation on the temperature data. Thus, the transformation does not help in improving the temperature data.