Exercise-004

2024-03-29

Εργασία 2

Να μελετήσετε το dataset diamonds του ggplot2 package.

Να εντοπίσετε δύο (2) ζευγάρια μεταβλητών που θεωρείτε ότι παρουσιάζουν ενδιαφέρον:

- Να μελετήσετε τη μεταξύ τους συσχέτιση (cor)
- Να δημιουργήσετε το αντίστοιχο διάγραμμα διασποράς (scatterplot)
- Να σχολιάσετε τα ευρήματά σας.

Η εργασία θα είναι μια σύντομη παρουσίαση σε google slides (+κώδικας).

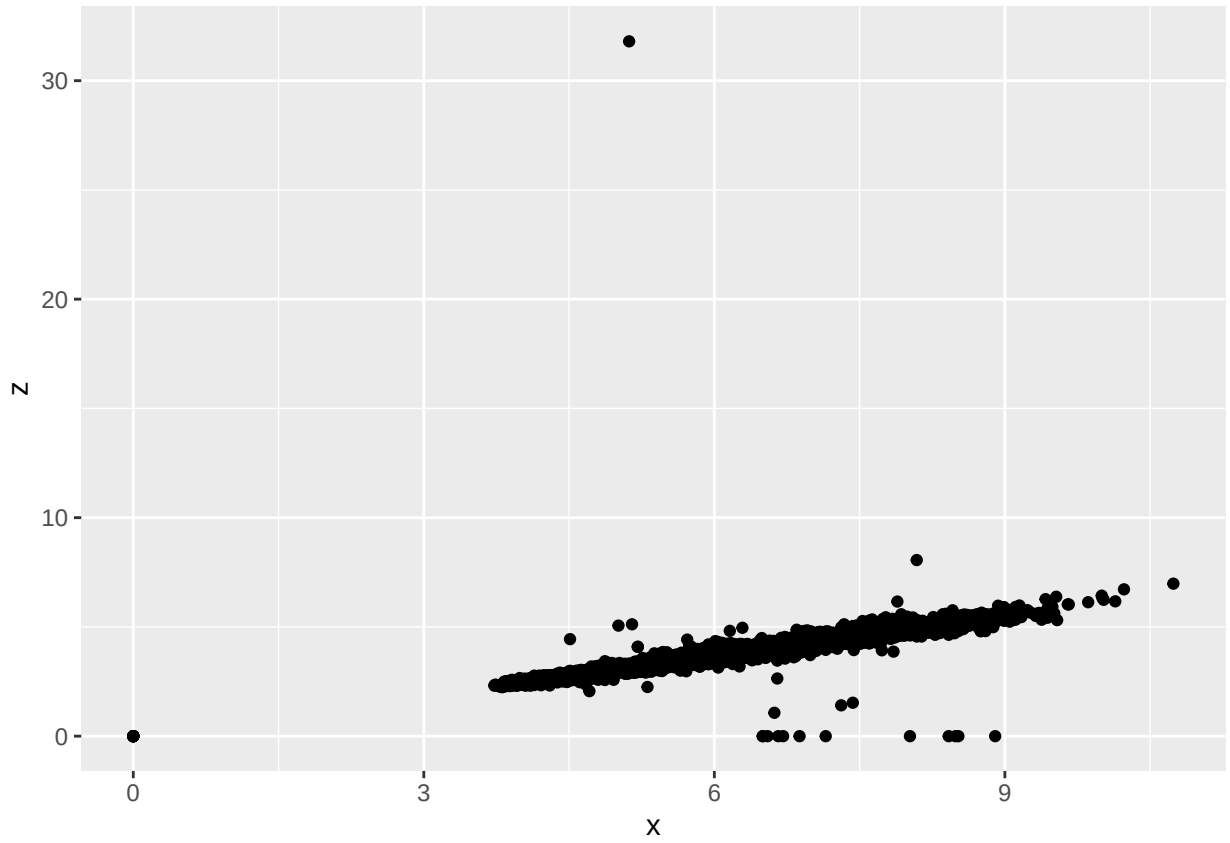
```
library(ggplot2)


# Compute pearson correlation and plot histograms for variables x and z

 print(cor(diamonds$x,diamonds$z,method="pearson"))
```

```
## [1] 0.9707718
```

```
 print(ggplot(diamonds,aes(x=x,y=z)) + geom_point())
```

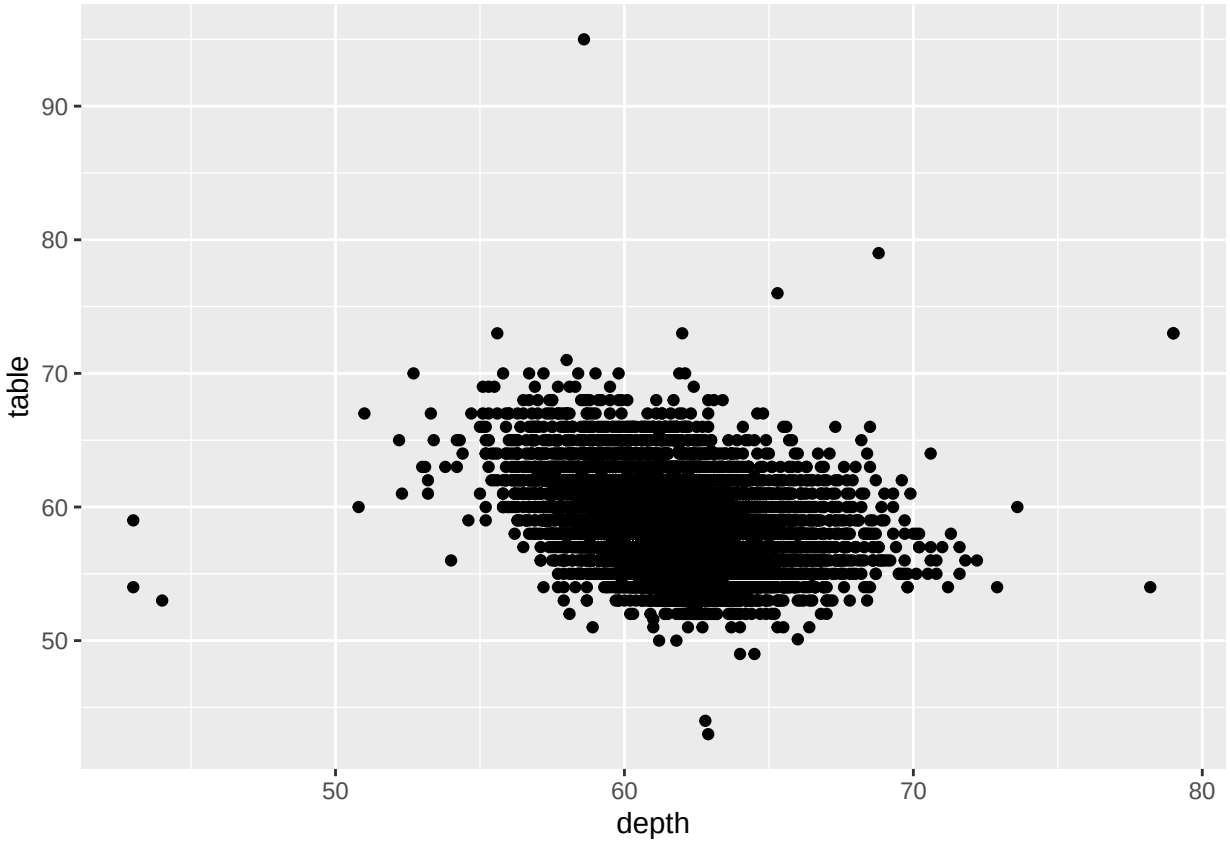



```
# Compute pearson correlation and plot histograms for variables depth and table

print(cor(diamonds$depth,diamonds$table,method="pearson"))
```

```
## [1] -0.2957785
```

```
print(ggplot(diamonds,aes(x=depth,y=table)) + geom_point())
```



Σχολιασμός

Στο διάγραμμα διασποράς των μεταβλητών x και z βλέπουμε ότι τα σημεία δεν είναι διεσπαρμένα στο επίπεδο, το οποίο σημαίνει ότι υπάρχει γραμμική σχέση μεταξύ αυτών. Στο διάγραμμα διασποράς των μεταβλητών depth και table παρατηρούμε ότι τα σημεία είναι διεσπαρμένα στο επίπεδο, το οποίο σημαίνει ότι υπάρχει καμπυλόγραμμη σχέση μεταξύ αυτών.

Εργασία 3

Να μελετήσετε το dataset airquality (base R). Το dataset περιέχει ημερήσιες μετρήσεις της ποιότητας του αέρα (air quality) στη Νέα Υόρκη για μια περίοδο 5 μηνών.

Να απαντήσετε στις παρακάτω ερωτήσεις:

- Ποιά είναι η μέση τιμή της θερμοκρασίας για τη δεδομένη περίοδο;
- Ποιά ημέρα ήταν η θερμότερη;
- Ποιά είχε τον πολύ αέρα;
- Ποιές ημέρες η θερμοκρασία ήταν μεγαλύτερη από 90 βαθμούς Fahrenheit?

Επιπλέον, με τα διαθέσιμα δεδομένα, να δημιουργήσετε ένα διάγραμμα διασποράς (scatterplot), ένα boxplot, ένα histgram κι ένα bar chart.

Να σχολιάσετε τα ευρήματά σας. Η εργασία θα είναι μια σύντομη παρουσίαση σε google slides στο επισυναπτόμενο αρχείο... εκεί να έχετε screenshots από τα διαγράμματα, editable κώδικα, τα αποτελέσματά σας, κι έναν σύντομο σχολιασμό.

```
# Mean Temperature
print(mean(airquality$Temp))
```

```
## [1] 77.88235
```

```
# Hottest Day
cat("Day: ",airquality$Day[which.max(airquality$Temp)],"Month: ",airquality$Month[which.max(airquality$
```

```
## Day:  28 Month:  8
```

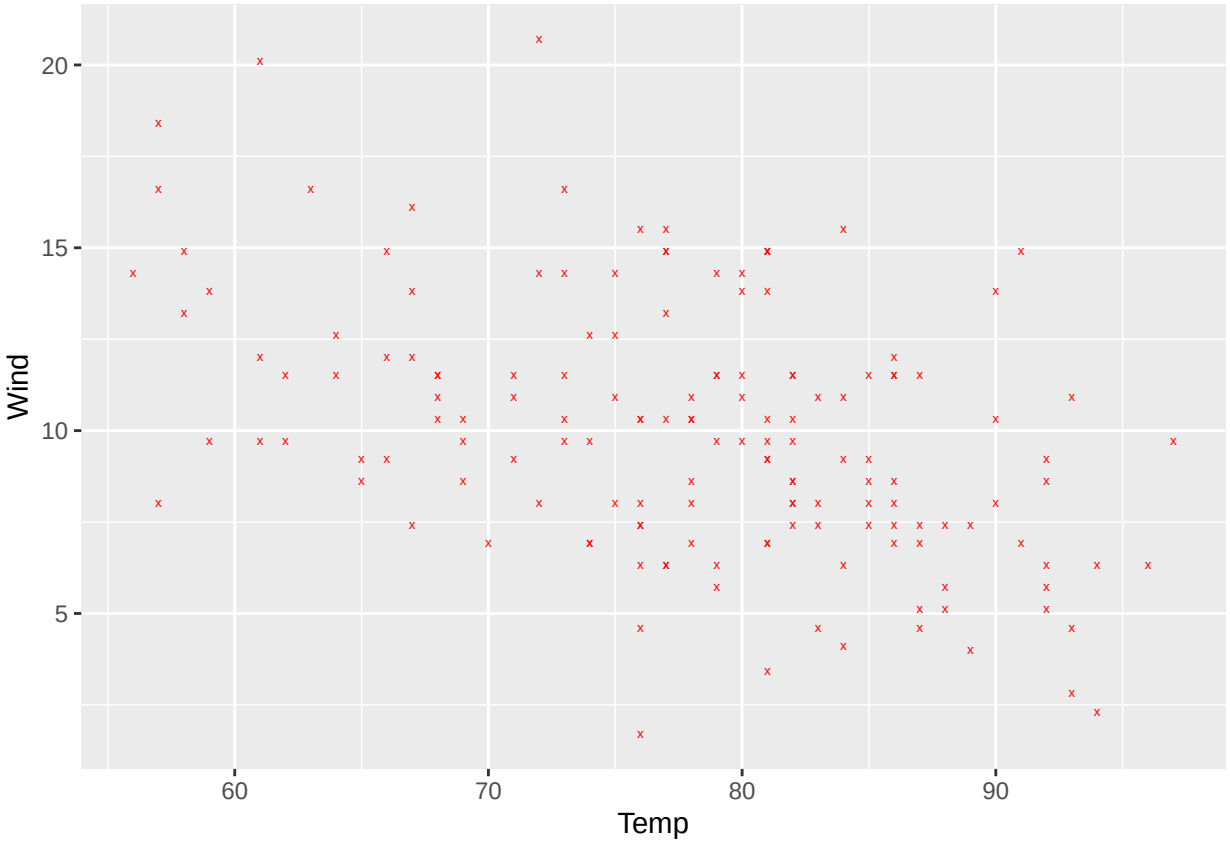
```
# Max wind speed day
cat("Day: ",airquality$Day[which.max(airquality$Wind)],"Month: ",airquality$Month[which.max(airquality$
```

```
## Day:  17 Month:  6
```

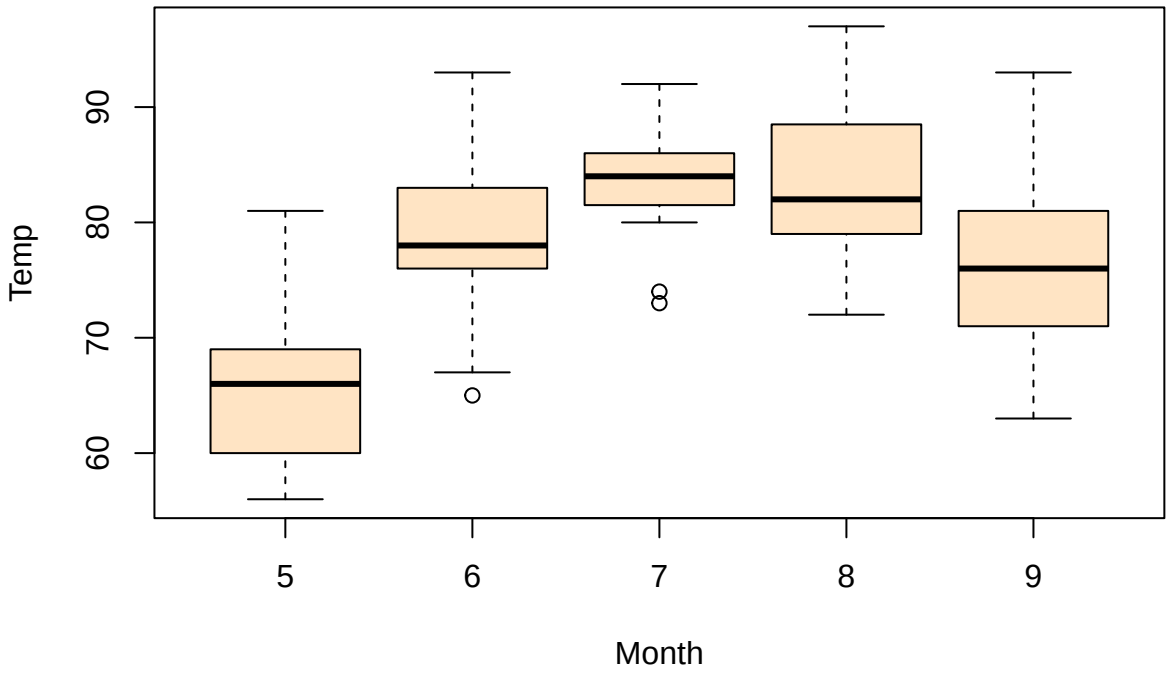
```
# Days with temperature greater than 90 Fahrenheit
Greater_Than_90 <- subset(airquality,Temp>90)
print(Greater_Than_90)
```

```
##     Ozone Solar.R Wind Temp Month Day
## 42     NA     259 10.9   93     6  11
## 43     NA     250  9.2   92     6  12
## 69     97     267  6.3   92     7   8
## 70     97     272  5.7   92     7   9
## 75     NA     291 14.9   91     7  14
## 102    NA     222  8.6   92     8  10
## 120    76     203  9.7   97     8  28
## 121   118     225  2.3   94     8  29
## 122    84     237  6.3   96     8  30
## 123    85     188  6.3   94     8  31
## 124    96     167  6.9   91     9   1
## 125    78     197  5.1   92     9   2
## 126    73     183  2.8   93     9   3
## 127    91     189  4.6   93     9   4
```

```
# Scatterplot
print(ggplot(airquality, aes(Temp,Wind))+geom_point(shape="x",color="red"))
```



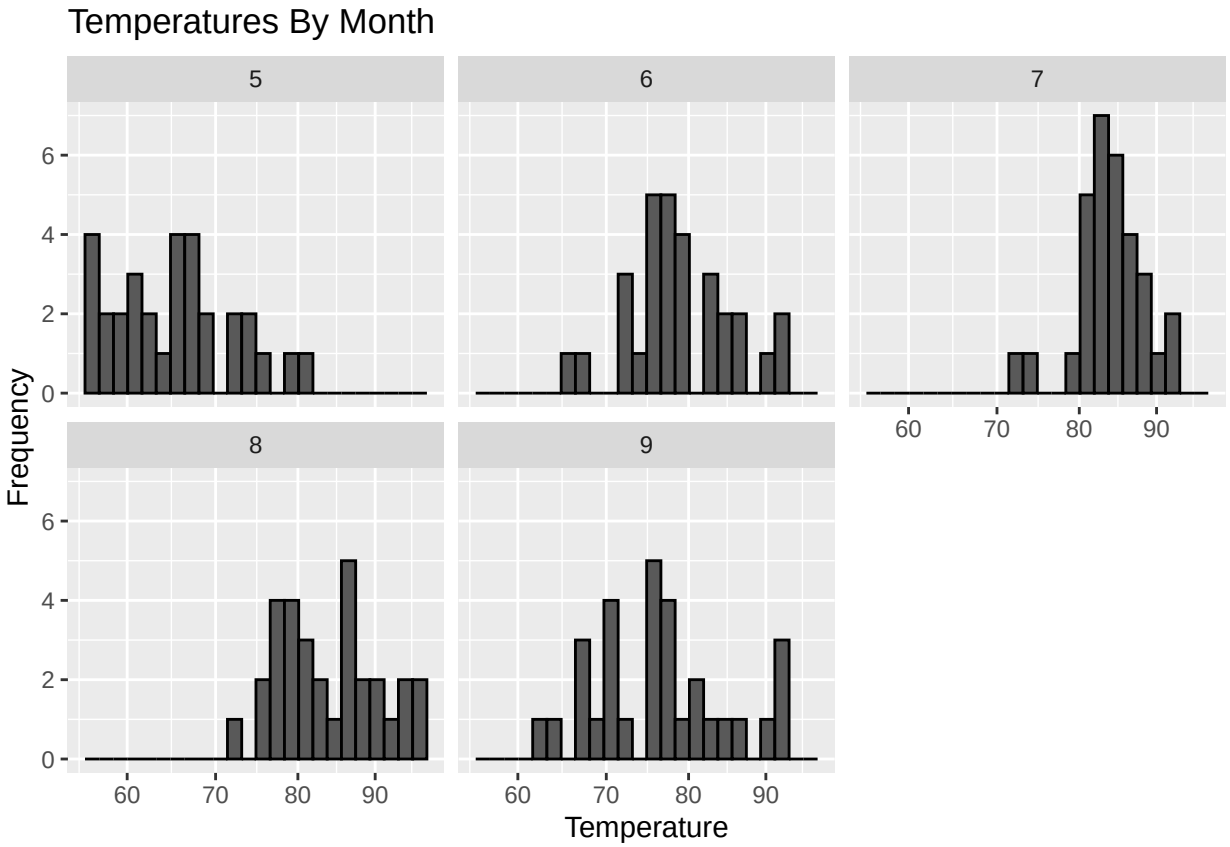

```
# Boxplot
print(with(airquality,boxplot(Temp ~ Month, col="bisque")))
```



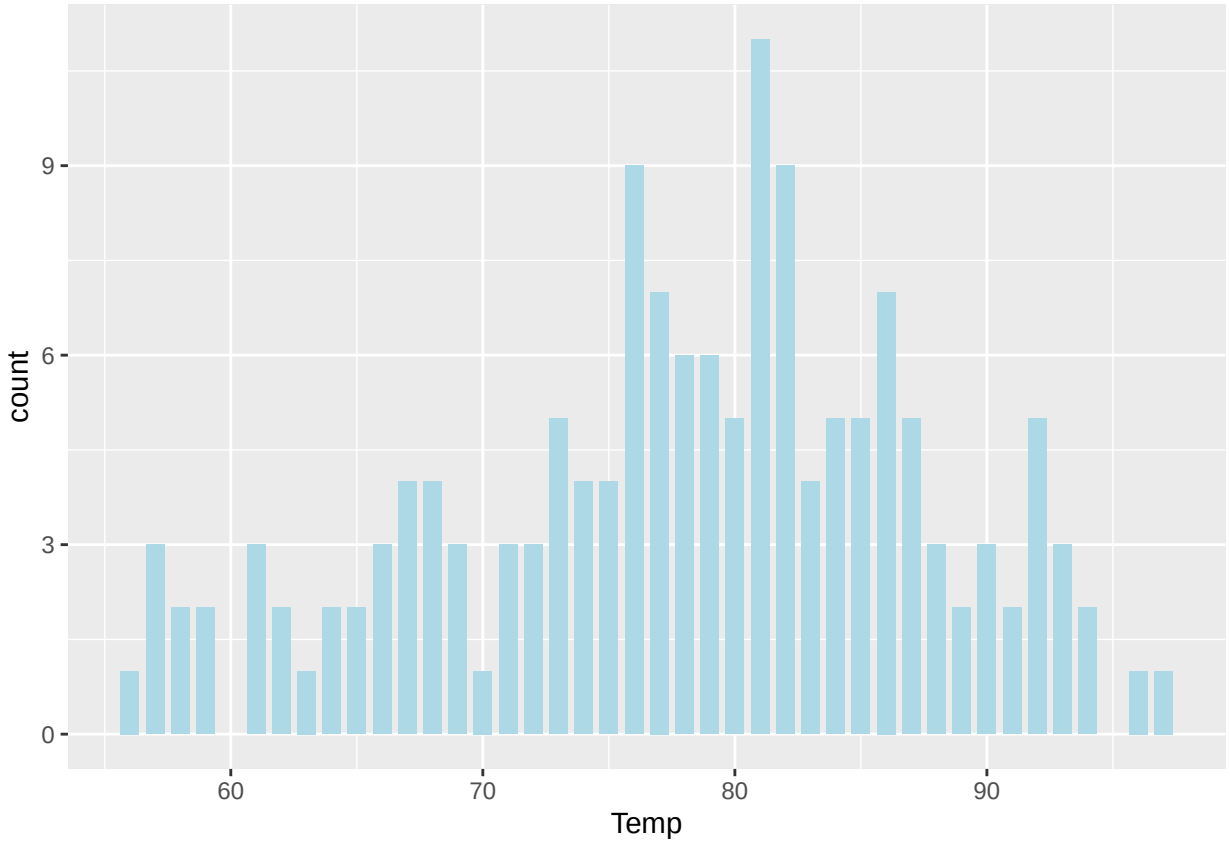

```
## $stats
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   56   67 80.0 72.0   63
## [2,]   60   76 81.5 79.0   71
## [3,]   66   78 84.0 82.0   76
## [4,]   69   83 86.0 88.5   81
## [5,]   81   93 92.0 97.0   93
##
## $n
## [1] 31 30 31 31 30
##
## $conf
##          [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 63.44601 75.98073 82.72301 79.30412 73.11533
## [2,] 68.55399 80.01927 85.27699 84.69588 78.88467
##
## $out
## [1] 65 73 74
##
## $group
## [1] 2 3 3
##
## $names
## [1] "5" "6" "7" "8" "9"
```

```r
# Histogram
print(ggplot(aes(x = Temp, fill = Day), data = airquality) + geom_histogram(binwidth = .1, color="black
        facet_wrap(~Month) + labs(title = "Temperatures By Month", x = "Temperature", y = "Frequency")
        scale_x_sqrt())
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
# Bar chart
print(ggplot(airquality, aes(Temp)) +
        geom_bar(fill="lightblue", position="dodge", width=0.75))
```



Σχολιασμός

Στο διάγραμμα διασποράς παρατηρούμε την σχέση μεταξύ των μεταβλητών Wind και Temp. Φαίνεται ότι τα σημεία είναι τυχαία διεσπαρμένα στο επίπεδο, το οποίο σημαίνει ότι υπάρχει καμπυλόγραμμη (και όχι γραμμική) σχέση μεταξύ των δύο μεταβλητών. Στο boxplot παρατηρούμε κάτι που επιβεβαιώνεται και από τα ιστογράμματα, ότι δηλαδή όσο περνάνε οι μήνες ανεβαίνουν οι θερμοκρασίες και στην συνέχεια “πέφτουν” με χαμηλότερο ρυθμό. Τέλος, στο bar chart παρατηρούμε την κατανομή των θερμοκρασιών για αυτό το διάστημα των 5 μηνών.