# 3. EDA - Eksploracyjna Analiza Danych

# 3.1 Importowanie potrzebnych bibliotek oraz pakietów

library(tidyverse)  
library(lubridate)  
library(corrplot)  
library(kableExtra)  
library(leaflet)  
library(scales)  
library(tmap)  
library(maps)

# 3.2 Importowanie danych oraz łączenie w jeden zbiór

dane\_1a <- read.csv("data/pipeline\_661\_austin.csv")  
dane\_1b <- read.csv("data/pipeline\_1642\_austin.csv")  
dane\_1c <- read.csv("data/pipeline\_2335\_austin.csv")  
dane\_2a <- read.csv("data/pipeline\_2358\_ny.csv")  
dane\_2b <- read.csv("data/pipeline\_4550\_ny.csv")  
dane\_2c <- read.csv("data/pipeline\_558\_ny.csv")  
dane\_3a <- read.csv("data/pipeline\_3687\_cali.csv")  
dane\_3b <- read.csv("data/pipeline\_6377\_cali.csv")  
dane\_3c <- read.csv("data/pipeline\_7062\_cali.csv")

Importowanie danych dla 3 różnych lokalizacji w każdym z 3 stanów: Texas, California, Nowy Jork. Wybrany zbiór jest reprezentowalny dla całego datasetu.

Głównym pakietem używanym przy tworzeniu eksploracyjnej analizy danych jest *tidyverse*. Jest to w zasadzie zbiór pakietów najbardziej potrzebnych przy analizie danych. Dwa główne komponenty tego pakietu to *dplyr* oraz *ggplot2*. W niniejszej analizie są one używane na każdym etapie tworzenia wykresów, najpierw tworzymy ograniczony zbiór danych **temp** za pomocą tzw. pipe’a %>% a następnie wykres za pomocą funkcji ggplot.

temp <- dane\_all %>% select(grid, state)   
  
ggplot(temp, aes(grid)) +  
 geom\_histogram(bins = 50, aes(y = ..density..), fill = "blue") +   
 geom\_density(alpha = 0.002, fill = "black") +   
 ggtitle("Rozkład zużycia prądu") +  
 geom\_vline(xintercept = round(mean(temp$grid), 2), size = 1, linetype = 2) + facet\_wrap(~ state, ncol =2)

# 3.3 Spojrzenie na strukturę danych:

dane\_all %>% str()

## 'data.frame': 261735 obs. of 15 variables:  
## $ X : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ dataid : int 661 661 661 661 661 661 661 661 661 661 ...  
## $ local\_15min : chr "2018-11-21 15:15:00" "2018-11-21 15:30:00" "2018-11-21 15:45:00" "2018-11-21 16:00:00" ...  
## $ grid : num 0.124 0.251 0.419 0.833 1.105 ...  
## $ city : chr "Austin" "Austin" "Austin" "Austin" ...  
## $ state : chr "Texas" "Texas" "Texas" "Texas" ...  
## $ station\_id : chr "KTXAUSTI1619" "KTXAUSTI1619" "KTXAUSTI1619" "KTXAUSTI1619" ...  
## $ latitude : num 30.3 30.3 30.3 30.3 30.3 ...  
## $ longitude : num -97.7 -97.7 -97.7 -97.7 -97.7 ...  
## $ temp\_avg : num 13 13 13 12 12 12 12 12 12 12 ...  
## $ wind\_speed\_avg: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ wind\_dir\_avg : num 346 346 346 292 292 292 292 277 277 277 ...  
## $ pressure\_max : num 1021 1021 1021 1021 1021 ...  
## $ humidity\_avg : num 64 64 64 67 67 67 67 72 72 72 ...  
## $ holiday : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...

## 3.3.1 Podsumowanie:

dane\_all %>% summary()

## X dataid local\_15min grid   
## Min. : 0 Min. : 558 Length:261735 Min. :-4.7940   
## 1st Qu.: 9349 1st Qu.:1642 Class :character 1st Qu.: 0.2430   
## Median : 19473 Median :2358 Mode :character Median : 0.4340   
## Mean : 29484 Mean :3408 Mean : 0.5851   
## 3rd Qu.: 38713 3rd Qu.:6377 3rd Qu.: 0.8060   
## Max. :104147 Max. :7062 Max. :10.4640   
##   
## city state station\_id latitude   
## Length:261735 Length:261735 Length:261735 Min. :30.27   
## Class :character Class :character Class :character 1st Qu.:30.27   
## Mode :character Mode :character Mode :character Median :32.72   
## Mean :33.71   
## 3rd Qu.:32.72   
## Max. :42.44   
##   
## longitude temp\_avg wind\_speed\_avg wind\_dir\_avg   
## Min. :-117.16 Min. :-33.0 Min. : 0.000 Min. : 0   
## 1st Qu.:-117.16 1st Qu.: 15.0 1st Qu.: 1.000 1st Qu.: 96   
## Median : -97.74 Median : 19.0 Median : 2.000 Median :219   
## Mean :-101.23 Mean : 19.6 Mean : 3.415 Mean :196   
## 3rd Qu.: -97.74 3rd Qu.: 24.0 3rd Qu.: 5.000 3rd Qu.:286   
## Max. : -76.39 Max. : 44.0 Max. :25.000 Max. :360   
## NA's :7104 NA's :7104 NA's :7104   
## pressure\_max humidity\_avg holiday   
## Min. : 937.7 Min. : 5.00 False:177944   
## 1st Qu.:1008.1 1st Qu.:56.00 True : 83791   
## Median :1013.5 Median :74.00   
## Mean :1010.1 Mean :70.01   
## 3rd Qu.:1016.6 3rd Qu.:87.00   
## Max. :1044.7 Max. :99.00   
## NA's :3856 NA's :7104

# 3.3.2 Wnioski:

Zbiór posiada 34 700 obserwacji dla każdego domu oraz 15 róznych zmiennych:

**X** - liczba porządkowa

**dataid** - id dla obserwacji, oznacza powiązanie dom - mieszkańcy

**local\_15min** - zmienna daty, rok - miesiąc - dzień godzina:minuty

**grid** - zużycie prądu mierzone co 15 minut

**city** - miasto

**state** - stan

**station\_id** - id stacji pogodowej, z której pobrano dane pogodowe w danym momencie czasu

**latitude** - szerokość geograficzna

**longitude** - długość geograficzna

**temp\_avg** - średnia wartość temperatury w stopniach C

**wind\_speed\_avg** - średnia siła wiatru

**wind\_dir\_avg** - średni kierunek wiatru

**pressure\_max** - maksymalna wartość ciśnienia

**humidity\_avg** - wilgotność

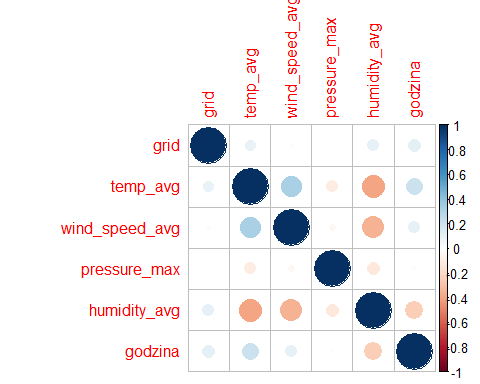
**holiday** - zmienna jakościowa oznaczająca czy dany dzień jest świętem

## 3.3.3 Spojrzenie na 6 pierwszych wierszy

dane\_all %>% head()

## X dataid local\_15min grid city state station\_id latitude  
## 1 0 661 2018-11-21 15:15:00 0.124 Austin Texas KTXAUSTI1619 30.27  
## 2 1 661 2018-11-21 15:30:00 0.251 Austin Texas KTXAUSTI1619 30.27  
## 3 2 661 2018-11-21 15:45:00 0.419 Austin Texas KTXAUSTI1619 30.27  
## 4 3 661 2018-11-21 16:00:00 0.833 Austin Texas KTXAUSTI1619 30.27  
## 5 4 661 2018-11-21 16:15:00 1.105 Austin Texas KTXAUSTI1619 30.27  
## 6 5 661 2018-11-21 16:30:00 0.869 Austin Texas KTXAUSTI1619 30.27  
## longitude temp\_avg wind\_speed\_avg wind\_dir\_avg pressure\_max humidity\_avg  
## 1 -97.74 13 0 346 1021.33 64  
## 2 -97.74 13 0 346 1021.33 64  
## 3 -97.74 13 0 346 1021.33 64  
## 4 -97.74 12 0 292 1021.33 67  
## 5 -97.74 12 0 292 1021.33 67  
## 6 -97.74 12 0 292 1021.33 67  
## holiday  
## 1 False  
## 2 False  
## 3 False  
## 4 False  
## 5 False  
## 6 False

# 3.4 Analiza korelacji zmiennych liczbowych



### 3.4.1 Wnioski:

* negatywna korelacja między temperaturą a ciśnieniem (< -0.3)
* negatywna korelacja między siłą wiatru a wilgotnością (< -0.3)
* pozytywna korelacja między prędkością wiatru a temperaturą (0.3)
* pozytywna korelacja między zmienną celu grid a wilgotnością (< 0.2)
* pozytywna korelacja między zmienną celu grid a godziną (< 0.2)

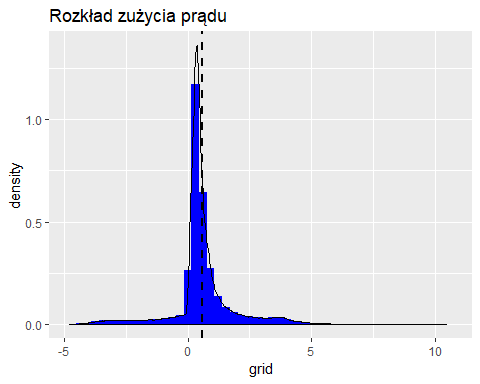
# 3.5 Analiza zmiennej ‘grid’ - zużycie prądu

Przy prezentacji rozkładów użyta została funkcja gęstości (density).

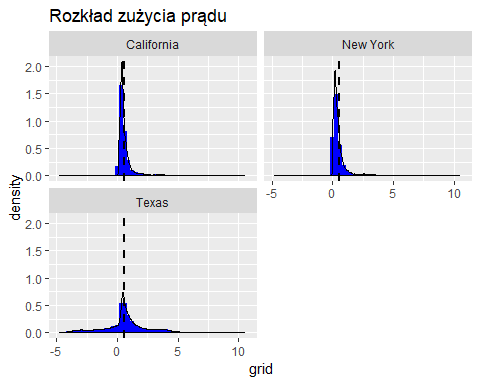
Podział wartości zużycia prądu na kwantyle:

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
 -4.794 0.028 0.185 0.286 0.361 0.434 0.532 0.688 0.971 1.835 10.464

## 3.5.1 Wykres rozkładu zmiennej zużycie prądu



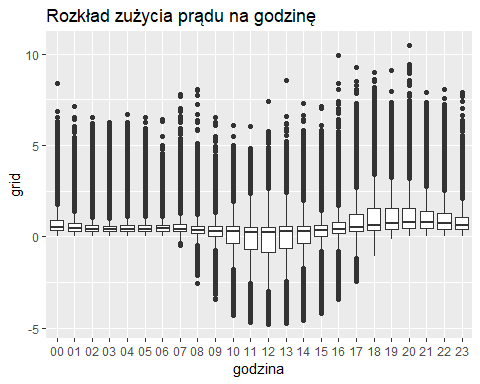
## 3.5.2 Wykres rozkładu zmiennej zużycie prądu z podziałem na stany



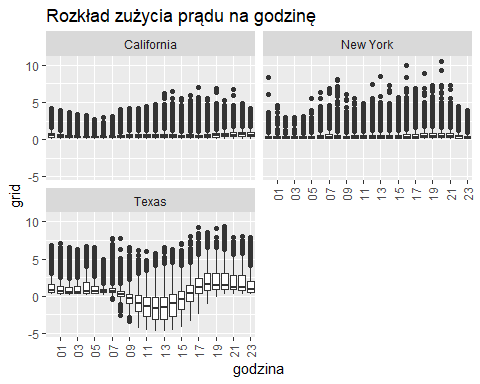
### 2.5.3 Wnioski:

* Analiza rozkładu zmiennej wskazuje na na duże zagęszczenie wartości wokół średniej. Zdecydowana większość wartości pozostaje w przedziale 0 - 1.
* Dodanie podziału na stan wyjaśnia zagęszczenie w przedziale 0-1, w stanach Kalifornia oraz Nowy Jork zużycie prądu przyjmuje tylko wartości dodatnie - wybrane id-domy nie posiadają własnych źródeł prądu. W wybranych domach w Teksiasie solary słoneczne produkują energię elektryczną.

## 3.5.4 Wykres pudełkowy prezentujący zużycie prądu w ciągu doby



## 3.5.5 Wykres pudełkowy prezentujący zużycie prądu w ciągu doby w podziale na stany



### 3.5.6 Wnioski:

* Analiza rozkładów wskazuje na duże podobieńtwo dla Kaliforni oraz Nowego Jorku, z niewiele większym rozmyciem danych dla drugiego stanu. W obu przypadkach jednak widać wyższe zużycie w ciągu dnia ze szczytem późnym popołudniem/wieczorem. W stanie Teksas ze względu na występowanie wartości ujemnych rozkład diametralnie się zmienia. Zauważyć można jednak silny wzrost w godzinach popołudniowych, gdy rośnie zużycie, a jednocześnie przez spadek promieniowania słonecznego produkcja energii maleje.

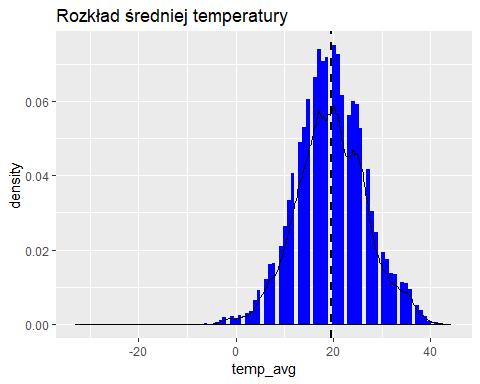
# 3.6 Analiza zmiennej ‘temp\_avg’ - średniej temperatury

Podział wartości średniej temperatury na kwantyle:

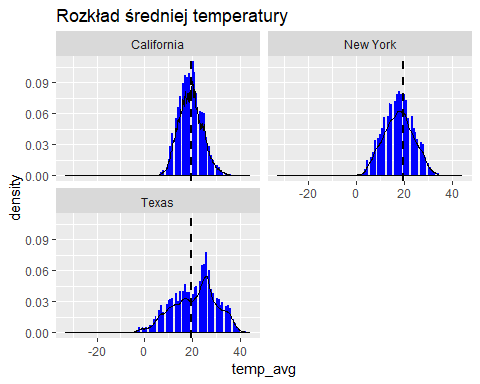
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## -33 11 14 16 18 19 21 23 26 29 44

## 

## 3.6.1 Wykres rozkładu średniej temperatury



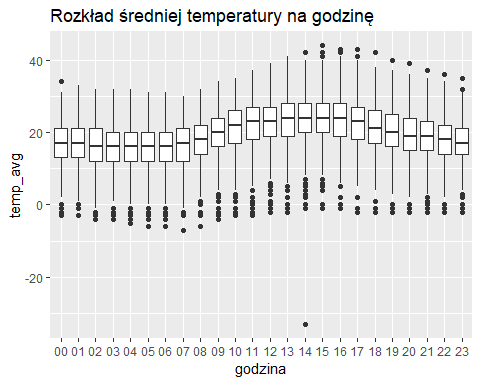
## 3.6.2 Wykres rozkładu średniej temperatury z podziałem na stany



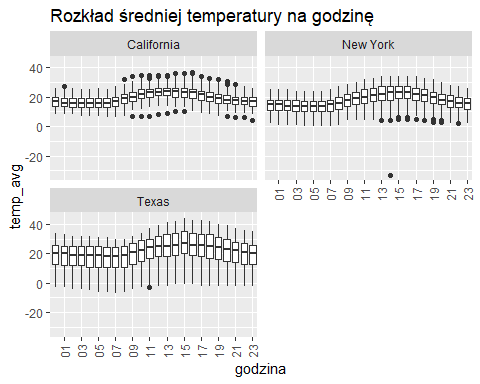
### 3.6.3 Wnioski:

* Analiza rozkładu zmiennej wskazuje na rozkład normalny. W podziale na stany widać różnice, jednak wartości rozkładają się po obu stronach średniej.
* Dodanie podziału na stany wyjaśnia pokazuje dodatkowo najmniejsze wahania temperatury w Kalifornii oraz największe w Teksasie.

## 3.6.4 Wykres pudełkowy prezentujący średnie temperatury w ciągu doby



## 3.6.5 Wykres pudełkowy prezentujący średnią temperaturę w ciągu doby w podziale na stany



### 3.6.6 Wnioski:

* Rozkład średniej temperatury wskazuje niewielkie zmiany dobowe. Podział na stany nie wnosi wiele do obserwacji poza różnicą w amplitudach temperatur, najniższą w Nowym Jorku i najwyższą w Teksasie. Ten ostatni stan charakteryzuje się również na najwyższą zmiennością dobową oraz najwyższymi temperaturami maksymalnymi.

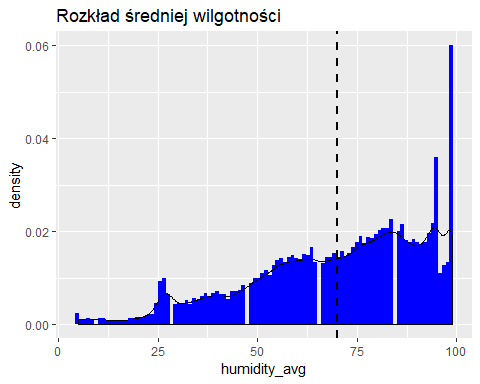
# 3.7 Analiza zmiennej ‘humidity\_avg’ - średniej wilgotności

Podział wartości średniej wilgotności na kwantyle:

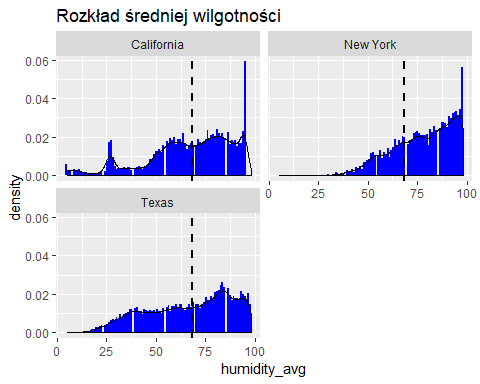
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## 5 38 52 60 67 74 80 85 90 95 99

## 

## 3.7.1 Wykres rozkładu średniej wilgotności



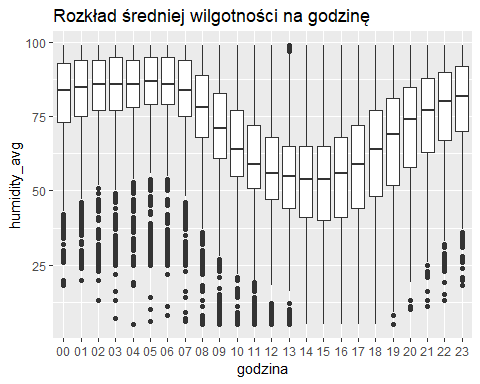
## 3.7.2 Wykres rozkładu średniej wilgotności z podziałem na stany



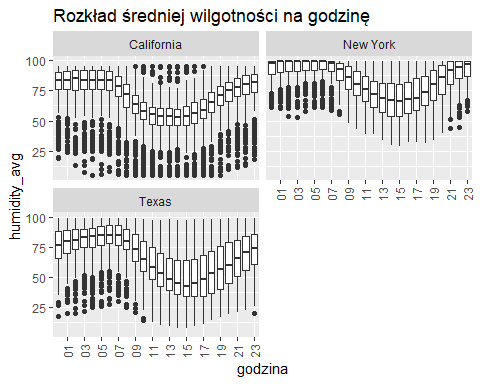
### 3.7.3 Wnioski:

* Analiza rozkładu zmiennej wskazuje na większą gęstość wartości wilgotności w górze przedziału.
* Dodanie podziału na stany wskazuje najwyższą wilgotność dla Nowego Jorku.

## 3.7.4 Wykres pudełkowy prezentujący średnią wilgotność w ciągu doby



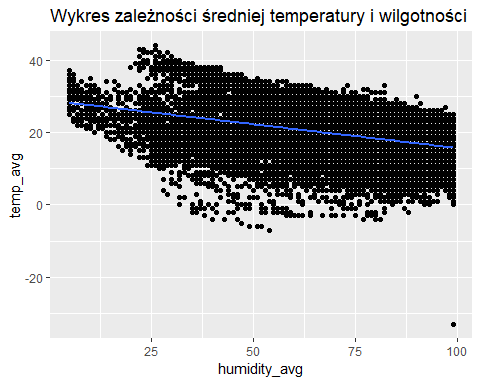
## 3.7.4 Wykres pudełkowy prezentujący średnią wilgotność w ciągu doby w podziale na stany



### 3.7.5 Wnioski:

* Rozkład średniej wilgotności wskazuje na silną zależność dobową.
* W przypadku indywidualnych stanów, Kalifornia wykazuje znacznie większe rozmycie wartości skrajnych, przy zachowaniu tendencji dobowej.

# 3.8 Wskazanie zależności między średnią temperaturą a wilgotnością



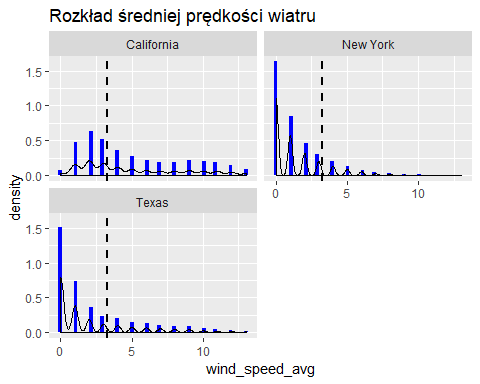
Zauważyć można pozytywną korelację. Jednak rozmycie na wykresie sugeruje iż korelacja nie jest silna.

# 3.9 Analiza zmiennej ‘wind\_speed\_avg’ - średniej prędkości wiatru

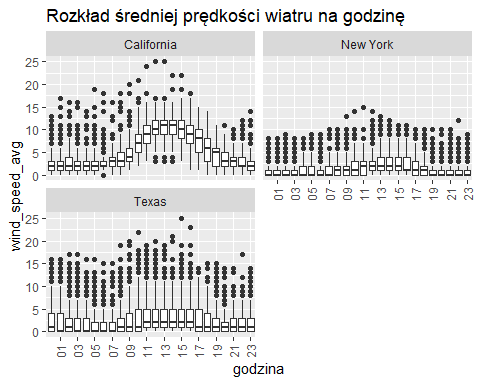
Podział wartości średniej wilgotności na kwantyle:

## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%   
## 0 0 0 1 1 2 3 4 6 9 25

## 3.9.1 Wykres rozkładu średniej prędkości wiatru z podziałem na stany



## 3.9.3 Wykres pudełkowy prezentujący średnią prędkość wiatru w ciągu doby w podziale na stany

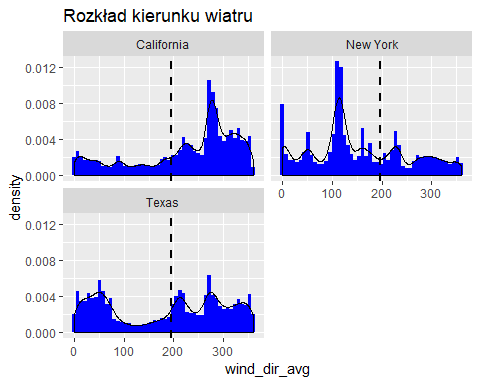


### 3.9.4 Wnioski:

* Rozkład charakteryzuje się dużą zmiennością godzinową jedynie dla Kalifornii. Średnia siła wiatru znacząco rośnie w godzinach około południowych.
* Analiza rozkładu zmiennej wskazuje na podobieństwo zjawiska w Nowym Jorku oraz Teksasie - większość obserwacji poniżej średniej, często występująca wartość 0. Może to również oznaczać brak danych.

# 3.10 Analiza zmiennej ‘wind\_dir\_avg’ - kierunku wiatru wyrażonego w stopniach 0-360

## 3.10.1 Wykres rozkładu kierunku wiatru

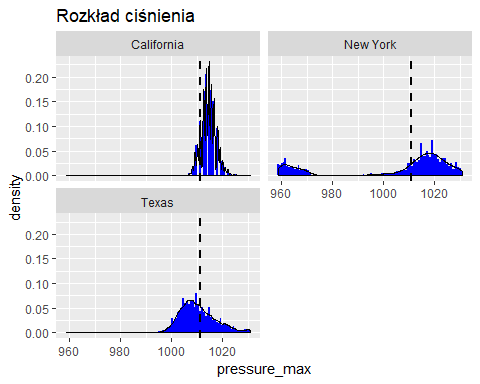


### 3.10.2 Wnioski:

* Każdy ze stanów charakteryzuje się innym rozkładem kierunku wiatru. Ma to ścisły związek z ich położeniem: ocean spokojny, ocean atlantycki, zatoka meksykańska.

# 3.11 Analiza zmiennej ‘pressure\_max’ - ciśnienia

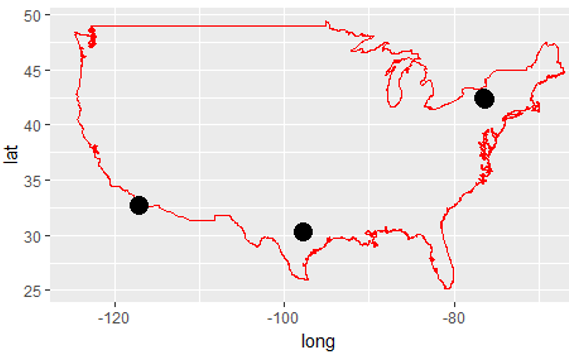
## 3.11.1 Wykres rozkładu ciśnienia z podziałem na stany



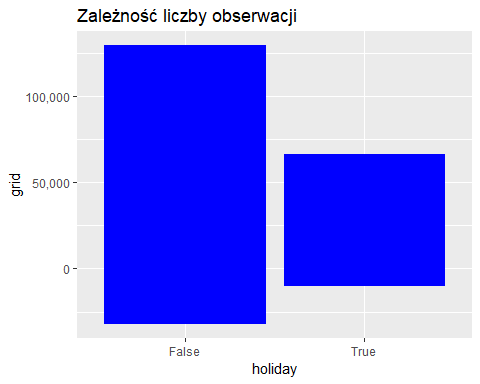
### 3.11.2 Wnioski:

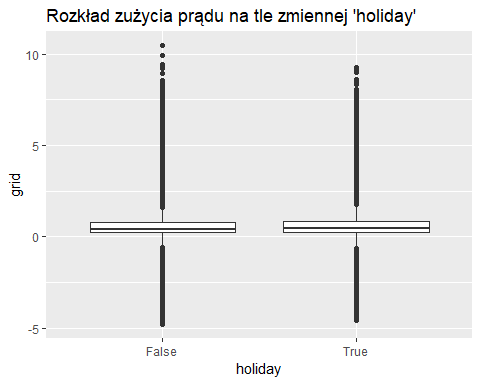
* Silne zróżnicowanie rozkładów ze względu na stan. Szczególnie silne zagęszczenie w przypadku Kalifornii, duże zróżnicowanie w przypadku Nowego Jorku.

## 3.11.3 Lokalizacje występujące w analizie, na podstawie długości i szerokości geograficznej.



## 3.11.4 Wykresy zależności zużycia prądu od zmiennej ‘holiday’, przedstawiające stosunek ilości obserwacji oraz rozkład

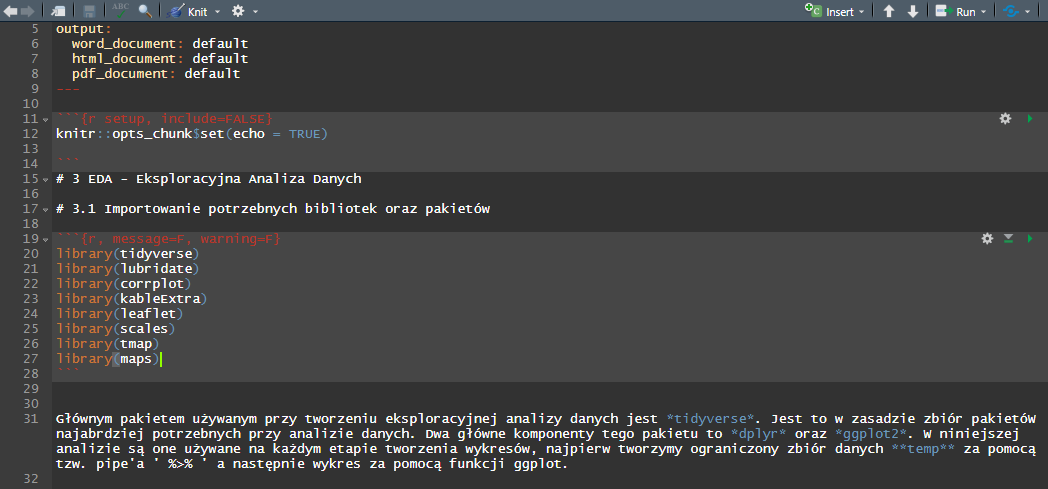




### 3.11.5 Wnioski:

* Porównanie ilościowe, wskazuje na oczywistą różnicę w liczbie obserwacji - więcej jest dni pracujących niż świątecznych.
* W przypadku rozkładu zaprezentowanego na wykresie pudełkowym brak istotnych różnic. Wartości pierwszego i trzeciego kwartyla są przybliżone, tak samo jak wartość mediany.

\*Eksploracyjna analiza danych - EDA, została w całości napisana w R Markdown i automatycznie przekonwertowana w dokument typu .docx. Zaletą tego rozwiązania jest możliwość powielania tych samych kroków analizy dla różnych zbiorów danych, bez konieczności tworzenia dokumentu od nowa.



**Rys. 3.1 Widok programu R Studio, w którym tworzony jest skrypt .rmd**

źródło: opracowanie własne