EDA projekt dyplomowy

Paweł Ciechanowicz

23 05 2020

## Importowanie potrzebnych pakietóW

library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------- tidyverse 1.2.1 --

## <U+221A> ggplot2 3.1.1 <U+221A> purrr 0.3.2   
## <U+221A> tibble 2.1.1 <U+221A> dplyr 0.8.0.1  
## <U+221A> tidyr 0.8.3 <U+221A> stringr 1.4.0   
## <U+221A> readr 1.3.1 <U+221A> forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(corrplot)

## corrplot 0.84 loaded

library(kableExtra)

##   
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':  
##   
## group\_rows

# Importowanie danych

Importowanie danych dla 3 różnych lokalizacji w każdym z 3 stanów: Texas, California, Nowy Jork.

Łączenie danych w jeden zbiór.

## Spojrzenie na strukturę danych:

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## 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

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## X dataid local\_15min grid city state station\_id  
## 1 0 2358 2019-06-24 14:45:00 0.161 Brooktondale New York KNYCAROL8  
## 2 1 2358 2019-06-24 14:30:00 0.060 Brooktondale New York KNYCAROL8  
## 3 2 2358 2019-06-24 14:15:00 0.032 Brooktondale New York KNYCAROL8  
## 4 3 2358 2019-06-24 14:00:00 0.032 Brooktondale New York KNYCAROL8  
## 5 4 2358 2019-06-24 15:45:00 0.100 Brooktondale New York KNYCAROL8  
## 6 5 2358 2019-06-24 15:30:00 0.032 Brooktondale New York KNYCAROL8  
## latitude longitude temp\_avg wind\_speed\_avg wind\_dir\_avg pressure\_max  
## 1 42.38 -76.39 23 1 93 956.65  
## 2 42.38 -76.39 23 1 93 956.65  
## 3 42.38 -76.39 23 1 93 956.65  
## 4 42.38 -76.39 23 1 93 956.65  
## 5 42.38 -76.39 24 2 51 955.98  
## 6 42.38 -76.39 24 2 51 955.98  
## humidity\_avg holiday  
## 1 74 False  
## 2 74 False  
## 3 74 False  
## 4 74 False  
## 5 71 False  
## 6 71 False

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## X dataid local\_15min grid city state station\_id  
## 1 0 3687 2014-01-01 00:00:00 1.717 San Diego California KCASANDI97  
## 2 1 3687 2014-01-01 00:15:00 1.905 San Diego California KCASANDI97  
## 3 2 3687 2014-01-01 00:30:00 1.923 San Diego California KCASANDI97  
## 4 3 3687 2014-01-01 00:45:00 1.535 San Diego California KCASANDI97  
## 5 4 3687 2014-01-01 01:00:00 1.010 San Diego California KCASANDI97  
## 6 5 3687 2014-01-01 01:15:00 0.957 San Diego California KCASANDI97  
## latitude longitude temp\_avg wind\_speed\_avg wind\_dir\_avg pressure\_max  
## 1 32.72 -117.16 9 2 4 1017.95  
## 2 32.72 -117.16 9 2 4 1017.95  
## 3 32.72 -117.16 9 2 4 1017.95  
## 4 32.72 -117.16 9 2 4 1017.95  
## 5 32.72 -117.16 9 2 14 1018.96  
## 6 32.72 -117.16 9 2 14 1018.96  
## humidity\_avg holiday  
## 1 95 True  
## 2 95 True  
## 3 95 True  
## 4 95 True  
## 5 93 True  
## 6 93 True

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

## Warning in bind\_rows\_(x, .id): binding factor and character vector,  
## coercing into character vector

## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector  
  
## Warning in bind\_rows\_(x, .id): binding character and factor vector,  
## coercing into character vector

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 ...

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

: Wnioski

Zbiór posiada 34 700 obserwacji dla jednego 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

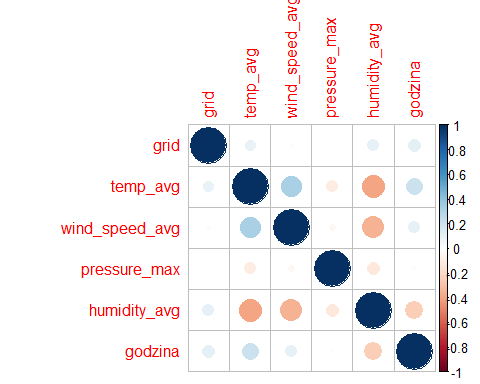
# 5 pierwszych wierszy

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

## Analiza korelacji zmiennych liczbowych

## grid temp\_avg wind\_speed\_avg pressure\_max  
## grid 1.00000000 0.09484063 -0.01915328 0.00782906  
## temp\_avg 0.09484063 1.00000000 0.32781810 -0.10704683  
## wind\_speed\_avg -0.01915328 0.32781810 1.00000000 -0.04011247  
## pressure\_max 0.00782906 -0.10704683 -0.04011247 1.00000000  
## humidity\_avg 0.10073517 -0.39276973 -0.34754626 -0.12910857  
## godzina 0.11663975 0.21619553 0.10440696 -0.01065025  
## humidity\_avg godzina  
## grid 0.1007352 0.11663975  
## temp\_avg -0.3927697 0.21619553  
## wind\_speed\_avg -0.3475463 0.10440696  
## pressure\_max -0.1291086 -0.01065025  
## humidity\_avg 1.0000000 -0.24867046  
## godzina -0.2486705 1.00000000

 # Wnioski:

* silna negatywna korelacja między temperaturą a ciśnieniem (< -0.3)
* silna 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)

## Analiza zmiennej ‘grid’ - zużycie prądu

Podział wartości zużycia prądu na kwantyle:

knitr::kable(Grid,caption = 'Rozkład zmiennej Grid')

Rozkład zmiennej Grid

Grid

0%

-4.794

10%

0.028

20%

0.185

30%

0.286

40%

0.361

50%

0.434

60%

0.532

70%

0.688

80%

0.971

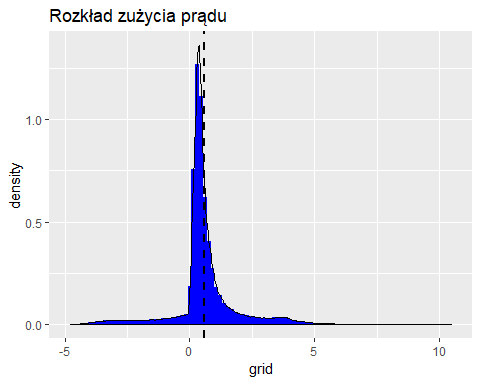
90%

1.835

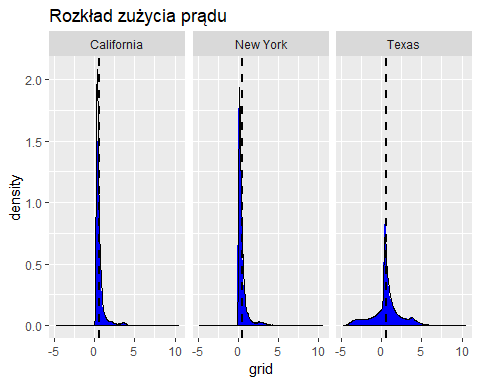
100%

10.464

# Wykres rozkładu zmiennej zużycie prądu



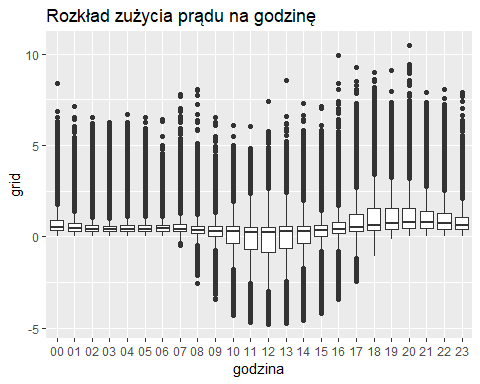
# Wykres rozkładu zmiennej zużycie prądu z podziałem na stany



# Wnioski:

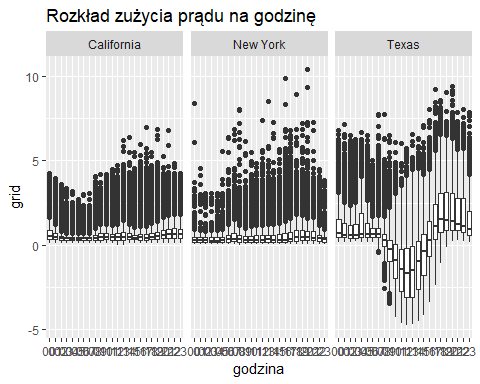
* Analiza rozkładu zmiennej wskazuje na na duże zagęszczenie wartości wokół średniej. Zdecydowana większość wartości pozostaje w przedziale 0 - 1.
* Dodanie podziału na stan wyjaśnia zagęszczenie w przedziale 0-1, w stanach Kalifornia oraz Nowy Jork zużycie prądu przyjmuje tylko wartości dodatnie - wybrane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

# Wykres pudełkowy prezentujący zużycie prądu w ciągu doby



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

ggplot(temp, aes(x = godzina, y = grid)) +  
 geom\_boxplot() +  
 ggtitle("Rozkład zużycia prądu na godzinę") + facet\_wrap(~ state)



# Wnioski:

* Analiza rozkłądów 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 szcytem 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.

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

Podział wartości średniej temperatury na kwantyle:

knitr::kable(Grid,caption = 'Rozkład zmiennej temp\_avg')

Rozkład zmiennej temp\_avg

Grid

0%

-33

10%

11

20%

14

30%

16

40%

18

50%

19

60%

21

70%

23

80%

26

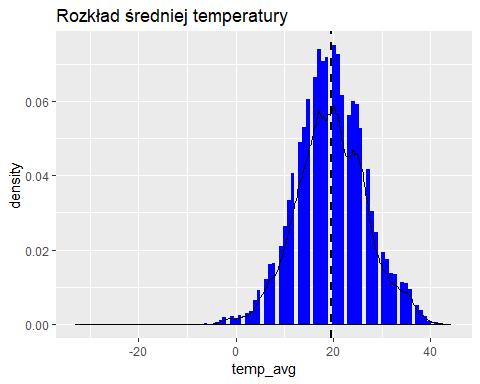
90%

29

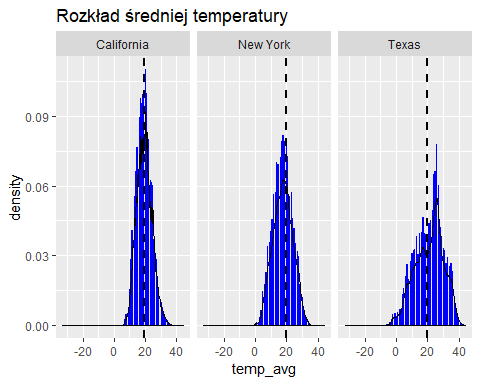
100%

44

# Wykres rozkładu średniej temperatury



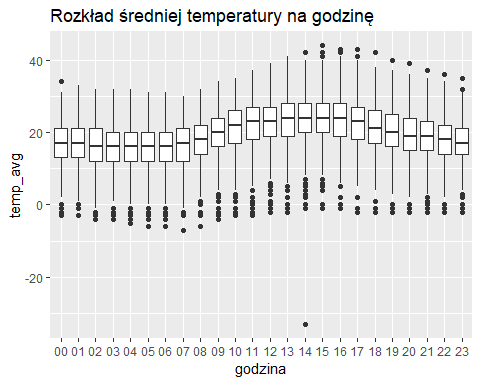
# Wykres rozkładu średniej temperatury z podziałem na stany



# Wnioski:

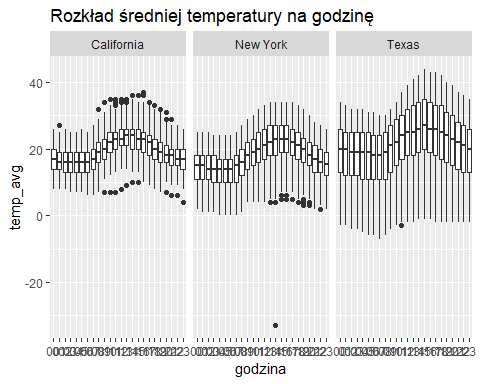
* Analiza rozkładu zmiennej wskazuje na rozkład normalny. W podziale na stany widać różnice, jednak warotści rozkładają się po obu stronach średniej.
* Dodanie podziału na stany wyjaśnia pokazuje dodatkowo najmniejsze wahania temperatury w Kaliforni oraz największe w Teksasie.

# Wykres pudełkowy prezentujący średnie temperatury w ciągu doby



# Wykres pudełkowy prezentujący średnią temperatyrę w ciągu doby w podziale na stany

ggplot(temp, aes(x = godzina, y = temp\_avg)) +  
 geom\_boxplot() +  
 ggtitle("Rozkład średniej temperatury na godzinę") + facet\_wrap(~ state)



# Wnioski:

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

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

Podział wartości średniej wilgotności na kwantyle:

knitr::kable(Grid,caption = 'Rozkład zmiennej humidity\_avg')

Rozkład zmiennej humidity\_avg

Grid

0%

5

10%

38

20%

52

30%

60

40%

67

50%

74

60%

80

70%

85

80%

90

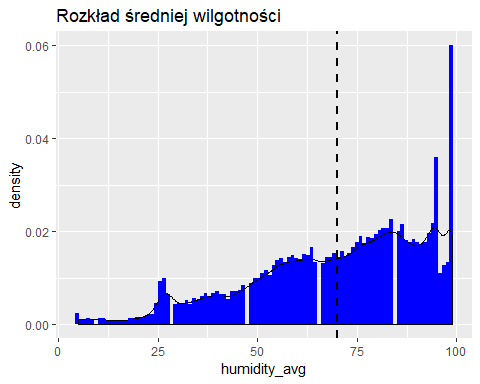
90%

95

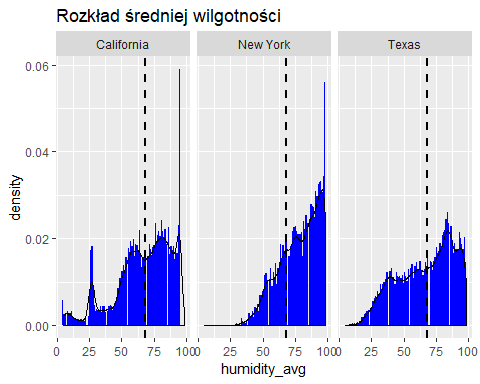
100%

99

# Wykres rozkładu średniej wilgotności



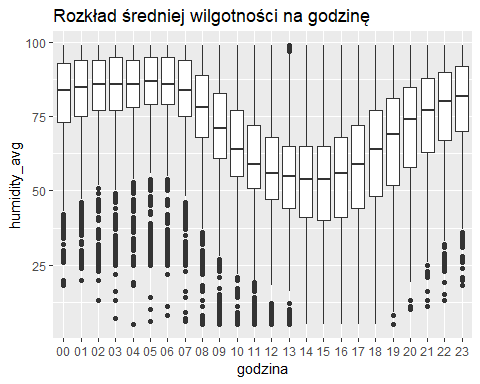
# Wykres rozkładu średniej wilgotności z podziałem na stany



# Wnioski:

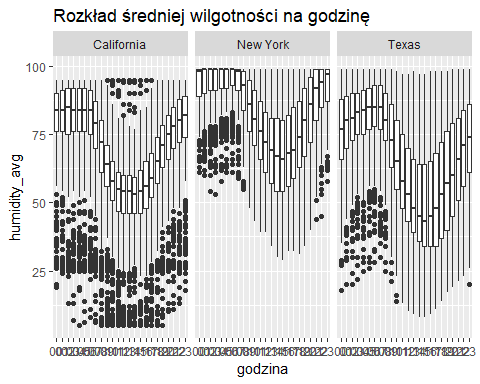
* Analiza rozkładu zmiennej wskazuje na większą gęstość wartości wilgotności w górze przedziału.
* Dodanie podziału na stany wskazuje najwyższą wilgotność dla Nowego Jorku.

# Wykres pudełkowy prezentujący średnią wilgotność w ciągu doby



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

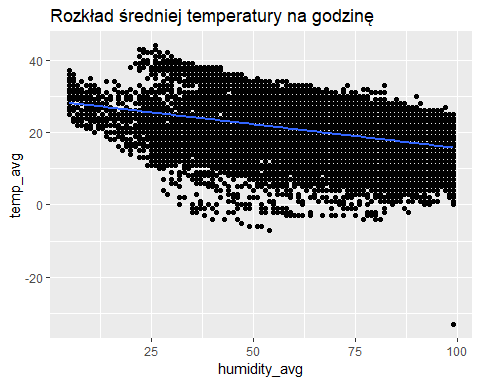
ggplot(temp, aes(x = godzina, y = humidity\_avg)) +  
 geom\_boxplot() +  
 ggtitle("Rozkład średniej wilgotności na godzinę") + facet\_wrap(~ state)



# Wnioski:

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

# Wyskazanie zależności między średnią temperaturą a wiglotnością



Zauważyć można negatywną korelację.

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

Podział wartości średniej wilgotności na kwantyle:

knitr::kable(Grid,caption = 'Rozkład zmiennej wind\_speed\_avg')

Rozkład zmiennej wind\_speed\_avg

Grid

0%

0

10%

0

20%

0

30%

1

40%

1

50%

2

60%

3

70%

4

80%

6

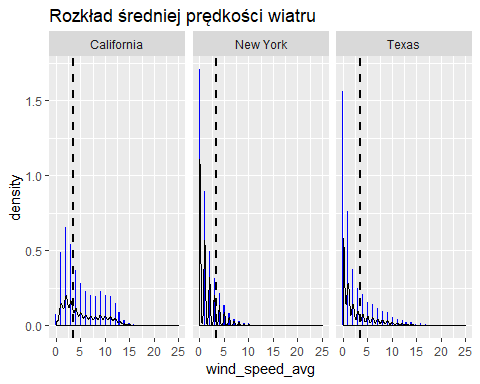
90%

9

100%

25

# Wykres rozkładu średniej prędkości wiatru z podziałem na stany

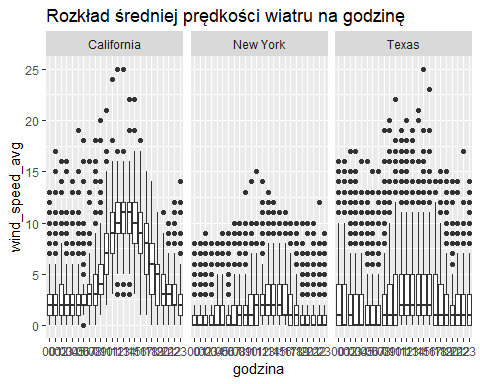


# Wnioski:

* 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.

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

temp <- dane\_all %>% select(wind\_speed\_avg, state) %>% mutate(godzina = format(as.POSIXct(strptime(dane\_all$local\_15min,"%Y-%m-%d %H:%M",tz="")) ,format = "%H")) %>% drop\_na()  
  
ggplot(temp, aes(x = godzina, y = wind\_speed\_avg)) +  
 geom\_boxplot() +  
 ggtitle("Rozkład średniej prędkości wiatru na godzinę") + facet\_wrap(~ state)

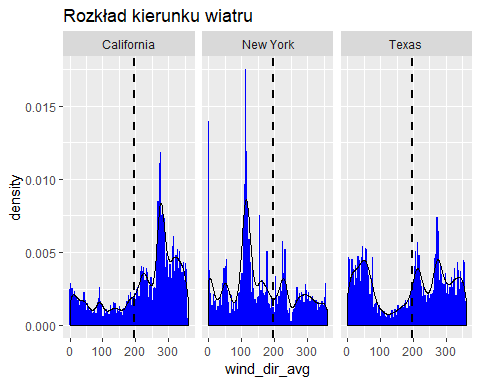


# Wnioski:

* Rozkład charakteryzuje się dużą zmiennością godzinową jedynie dla Kalifornii. Średnia siła wiatru znacząco rośnie w godzinach około południowych.

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

# Wykres rozkładu kierunku wiatru

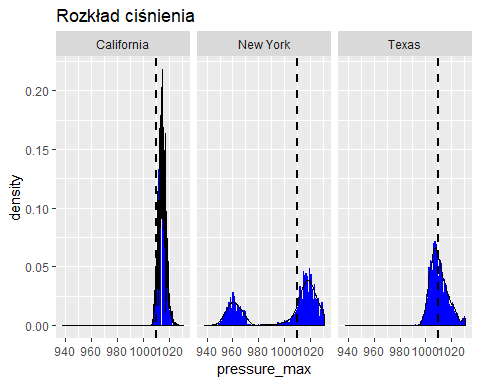


# Wnioski:

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

## Analiza zmiennej ‘pressure\_max’ - ciśnienia

# Wykres rozkładu ciśnienia z podziałem na stany

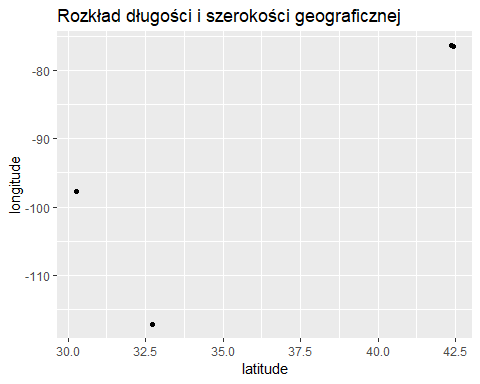


# Wnioski:

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

## Rozkład długości i szerokości geograficznej

temp <- dane\_all %>% select(latitude, longitude, state) %>% drop\_na()  
  
ggplot(temp, aes(x = latitude, y = longitude)) +   
 geom\_point() +  
 #geom\_histogram(bins = 100, aes(y = ..density..), fill = "blue") +   
 #geom\_density(alpha = 0.002, fill = "black") +   
 ggtitle("Rozkład długości i szerokości geograficznej")# + geom\_jitter(width = 0.5, height = 0.5, alpha = 0.5)



# geom\_vline(xintercept = round(mean(temp$wind\_dir\_avg), 2), size = 1, linetype = 2)# + facet\_wrap(~ state)

# Wykres zależności zużycia prądu od zmiennej ‘holiday’

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

