ANOVA Adam Matuszczyk MSAD 2018/2019

Celem pracy jest sprawdzenie czy sprzedaż konkretnego produktu zależy od zmiennych czasu oraz regionów w jakich odbywa się dystrybucja. Do analizy zastosowałem dane sprzedaży pewnego produktu leczniczego, który jest dostępny w całej Polsce w dystrybucji zamkniętej (bezpośredni importer-> hurtownia farmaceutyczna->apteka->pacjent).

Ze względu na czytelność analizy uprościłem dostępna dane do analizy regionów oraz czasu wg. klucza: **Region[**C_G (kujawsko-pomorskie, pomorskie); D_F(dolnośląskie, lubuskie); E_T(łódzkie, świętokrzyskie), K_R(małopolskie, podkarpackie); N_B_L(warmińsko-mazurskie, podlaskie, lubelskie), P_Z(wielkopolskie, zachodniopomorskie); S O(śląskie, opolskie); W(mazowieckie)**]**, **Czas[** zima(styczeń, luty, marzec); wiosna(kwiecień, maj, czerwiec), lato(lipiec, sierpień, wrzesień); jesień(listopad, październik, grudzień)**]**. W wypadku **Regionu** skróty literowe zastosowałem z kodów tablic rejestracyjnych.

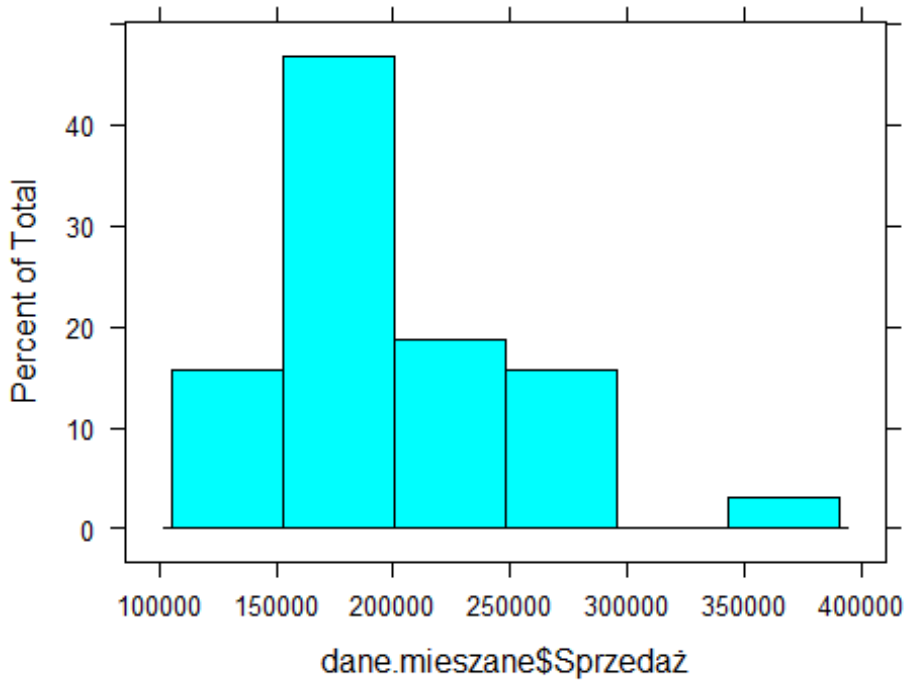
```
library(tidyverse)
library(lubridate)
library(reshape2)
library(graphics)
library(gplots)
library(nortest)
library(lattice)
library(skimr)

#https://docs.google.com/spreadsheets/d/e/2PACX-1vS16A2lcUKaVem5fW21_Jah4ZoSGMFmb9FFVPh5VMpqxuaNYV2qOT9lnPYMi-ZrLaLxztHrijGO4db1/pub?gid=1299943662&single=true&output=csv
moj_url <- "https://docs.google.com/spreadsheets/d/e/2PACX-1vTbOkma4SkS38S4vpMX0bGCung32QNSyHhmyrfOVn1-FuacQYV6ADwW1llm3lBqFOisTgZUPhM18q2M/pub?gid=1903822578&single=true&output=csv"
dane  <- read.delim(url(moj_url), header =TRUE, stringsAsFactors = FALSE, sep = ";")
colnames(dane) = c("region", "zima", "wiosna", "lato","jesień")

dane.mieszane <-  melt(dane, id= "region")
colnames(dane.mieszane) = c("Region", "Czas", "Sprzedaż")
dane.mieszane <- dane.mieszane %>% mutate_if(is.character, as.factor) %>% glimpse()

## Observations: 32
## Variables: 3
## $ Region   <fct> C_G, D_F, E_T, K_R, N_B_L, P_Z, S_O, W, C_G, D_F, E_T...
## $ Czas     <fct> zima, zima, zima, zima, zima, zima, zima, zima, wiosn...
## $ Sprzedaż <int> 245585, 118727, 166579, 209002, 165909, 197267, 29094...
```

```
head(dane.mieszane)
```

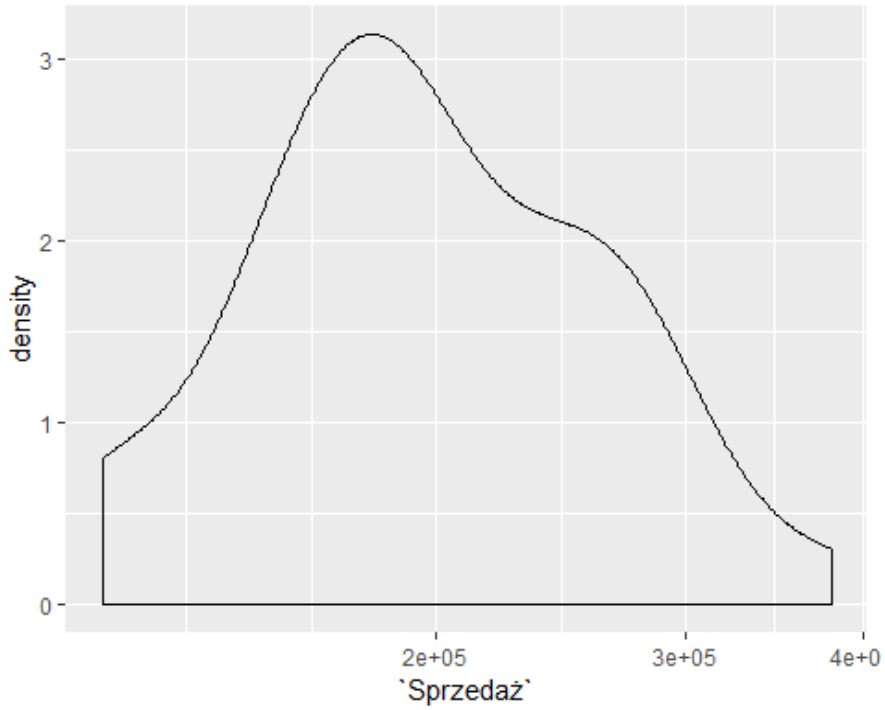
```
##   Region Czas Sprzedaż
## 1    C_G zima   245585
## 2    D_F zima   118727
## 3    E_T zima   166579
## 4    K_R zima   209002
## 5  N_B_L zima   165909
## 6    P_Z zima   197267
```

Po uporządkowaniu danych przedstawię kilka wykresów aby intuicyjnie ukierunkować się na kolejne kroki analizy.

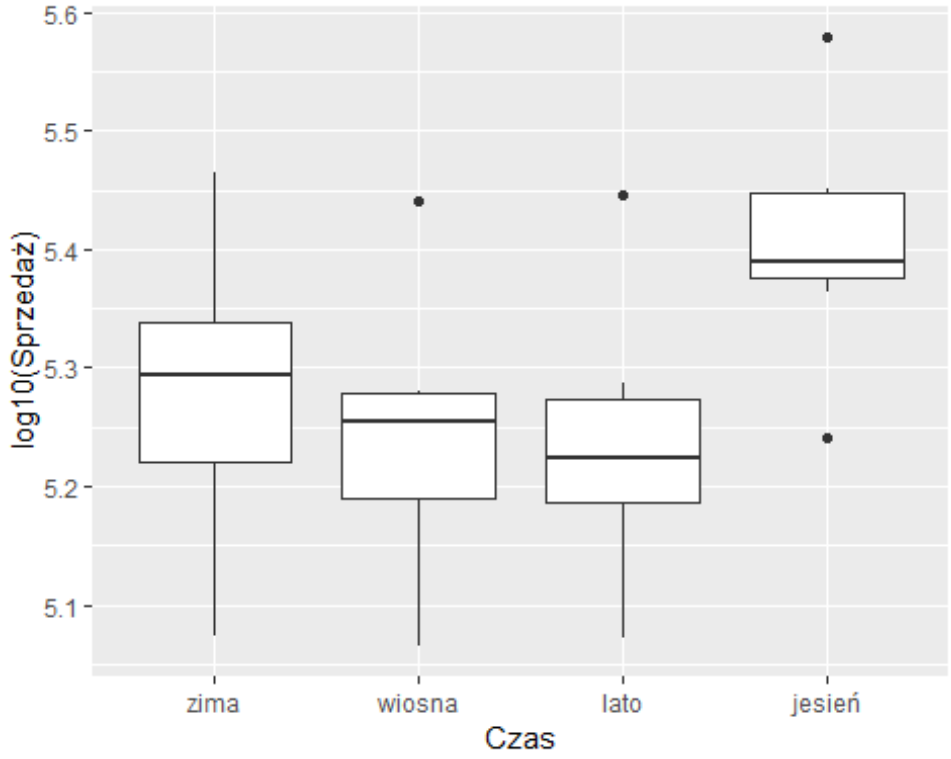
```
histogram(dane.mieszane$Sprzedaż)
```



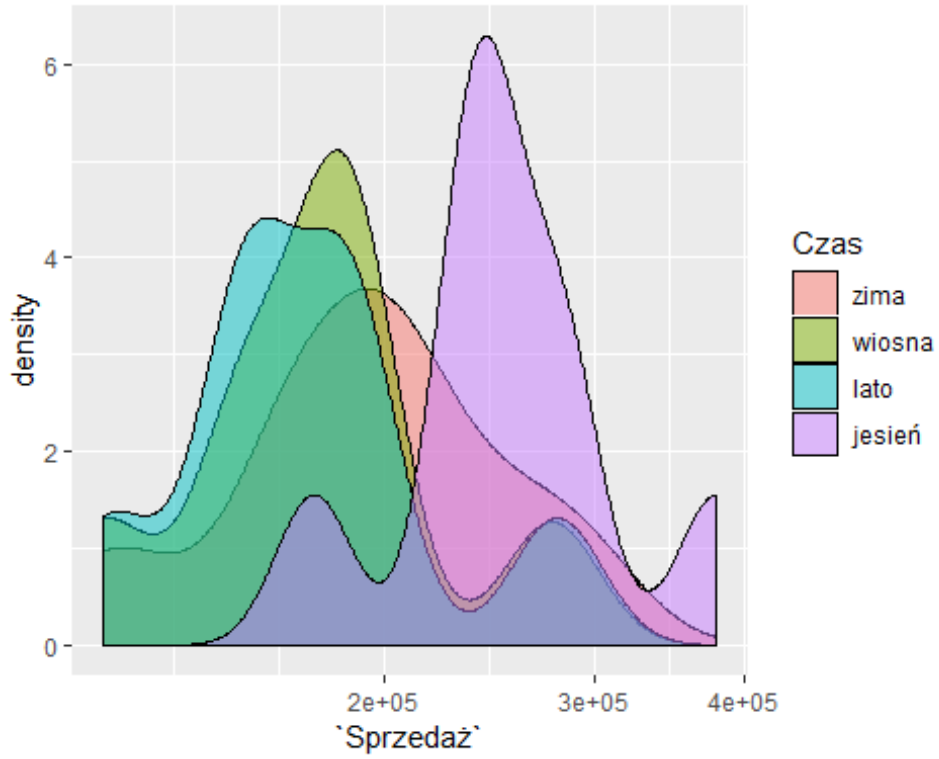
```
ggplot(dane.mieszane, aes(Sprzedaż)) + geom_density() + scale_x_log10()
```



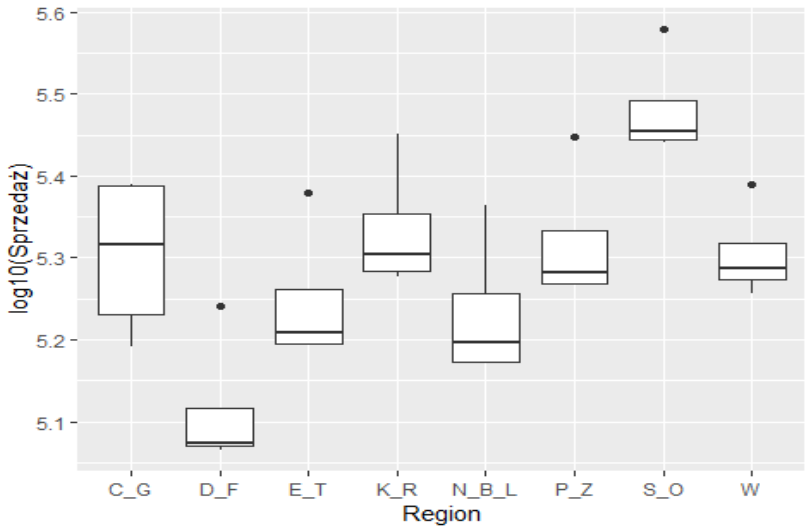

```
#zima(styczeń, luty, marzec); wiosna(kwiecień, maj, czerwiec)
#lato(lipiec, sierpień, wrzesień); jesień(listopad, październik, grudzień)
ggplot(dane.mieszane, aes(Czas, log10(Sprzedaż))) + geom_boxplot()
```



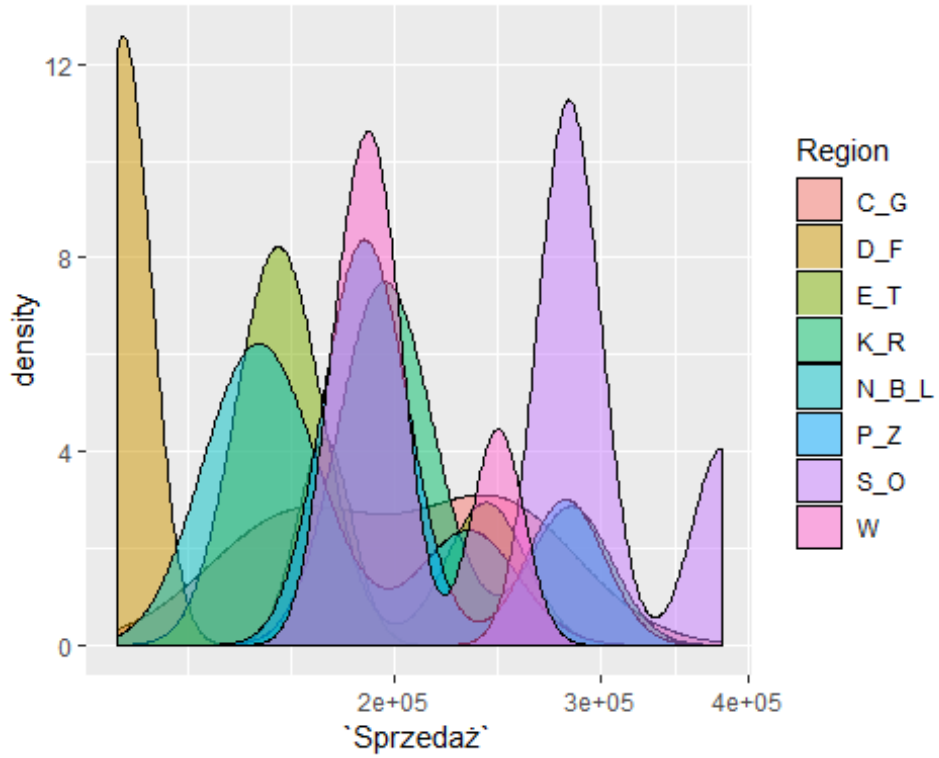
```
ggplot(dane.mieszane, aes(Sprzedaż, fill = Czas )) + geom_density(alpha = 0.5
) + scale_x_log10()
```




```
#C_G (kujawsko-pomorskie, pomorskie); D_F(dolnośląskie_lubuskie); E_T(łódzkie
, świętokrzyskie)
#K_R(małopolskie, podkarpackie); N_B_L(warmińsko-mazurskie, podlaskie, lubels
kie)
#P_Z(wielkopolskie_zachodniopomorskie); S_O(śląskie, opolskie); W(mazowieckie
)
ggplot(dane.mieszane, aes(Region,log10(Sprzedaż))) + geom_boxplot()
```



```
ggplot(dane.mieszane, aes(Sprzedaż, fill = Region)) + geom_density(alpha = 0.
5) + scale_x_log10()
```



Jak się należało spodziewać sprzedaż zależy od regionów i czasu, chociaż wizualnie widać, że będą problemy z odrzuceniem normalności rozkładu. Ciekawostką jest to, że w czasie lata liczba klientów aptek spada nawet o 50%. Polacy w wakacje zdrowieją co oczywiście widać w tej analizie natomiast ze względu na typ choroby raczej większość klientów naszego specyfiku jest w pozostałej grupie.

```
tapply(dane.mieszane$Sprzedaż, dane.mieszane$Region, shapiro.test)
```

```
## $C_G
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.82443, p-value = 0.1537
##
##
## $D_F
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.66396, p-value = 0.004058
##
##
```

```
## $E_T
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.71836, p-value = 0.01862
##
##
## $K_R
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.78476, p-value = 0.07759
##
##
## $N_B_L
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.76803, p-value = 0.05616
##
##
## $P_Z
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.71975, p-value = 0.01926
##
##
## $S_O
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.73649, p-value = 0.02863
##
##
## $W
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.83928, p-value = 0.1933
```

```r
tapply(dane.mieszane$Sprzedaż, dane.mieszane$Czas, shapiro.test)
```

```
## $zima
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96926, p-value = 0.8921
##
##
## $wiosna
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.89689, p-value = 0.2708
##
##
## $lato
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.87644, p-value = 0.1741
##
##
## $jesień
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.89777, p-value = 0.2759
```

```r
kruskal.test(dane.mieszane$Sprzedaż ~dane.mieszane$Czas)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dane.mieszane$Sprzedaż by dane.mieszane$Czas
## Kruskal-Wallis chi-squared = 9.4858, df = 3, p-value = 0.02348
```

```r
kruskal.test(dane.mieszane$Sprzedaż ~dane.mieszane$Region)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dane.mieszane$Sprzedaż by dane.mieszane$Region
## Kruskal-Wallis chi-squared = 18.972, df = 7, p-value = 0.008277
```

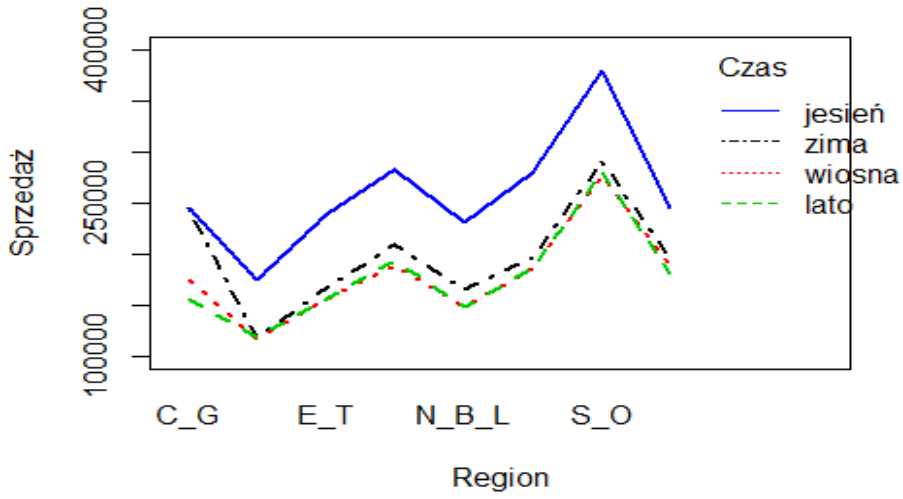
```
bartlett.test(dane.mieszane$Sprzedaż ~ dane.mieszane$Czas)

##
##  Bartlett test of homogeneity of variances
##
## data:  dane.mieszane$Sprzedaż by dane.mieszane$Czas
## Bartlett's K-squared = 0.46966, df = 3, p-value = 0.9255

bartlett.test(dane.mieszane$Sprzedaż ~ dane.mieszane$Region)

##
##  Bartlett test of homogeneity of variances
##
## data:  dane.mieszane$Sprzedaż by dane.mieszane$Region
## Bartlett's K-squared = 1.5051, df = 7, p-value = 0.9821
```

Jak widać z normalnością rozkładu jest różnie. W niektórych regionach p-wartość jest korzystna, ale w pozostałych niestety jednoznacznie wysokie wartości nie pozwalają na odrzucenie H0 normalności rozkładu. Sprzedaż w czasie ma rozkład normalny we wszystkich zakresach.

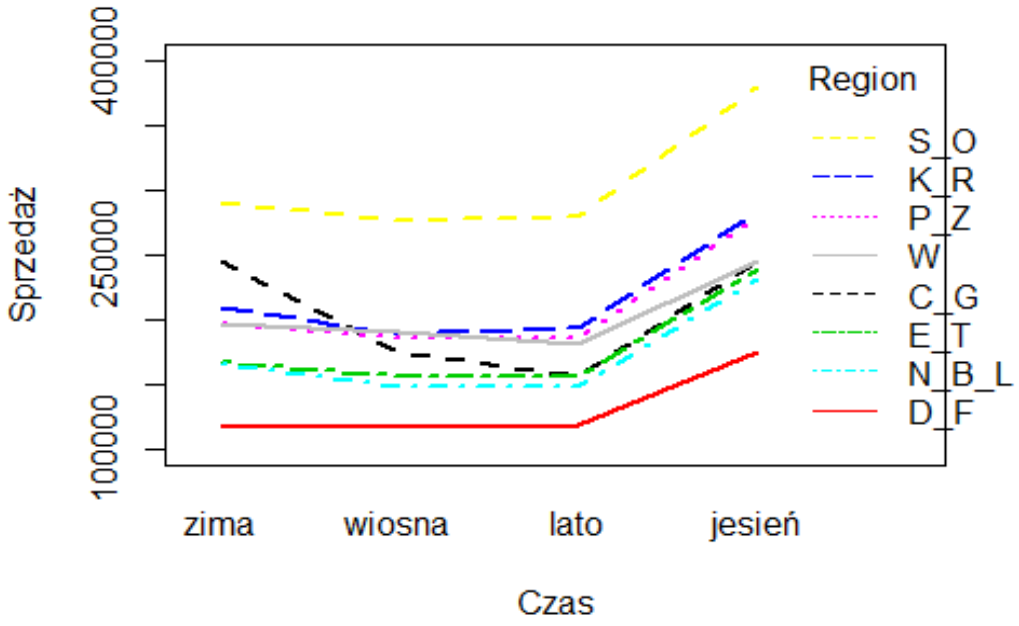
Wykonałem alternatywnie **test Kruskala- Wallisa** gdzie i wypadku **Czasu** i **Regionów** możemy odrzucić H0 o równości dystrybuant rozkładów w porównywanych populacjach.

W celu dalszych kroków wykonałem analizę interakcji między czynnikami.

```
interaction.plot(dane.mieszane$Region,dane.mieszane$Czas, dane.mieszane$Sprzedaż, ylim = c(100000,400000), ylab = "Sprzedaż", xlab = "Region", lwd = 2, trace.label = "Czas", col = 1 : 4)
```



```
interaction.plot(dane.mieszane$Czas,dane.mieszane$Region, dane.mieszane$Sprze
daż, ylim = c(100000,400000), ylab = "Sprzedaż", xlab = "Czas", lwd = 2,
trace.label = "Region", col = 1 : 8)
```



Jak widać obydwa czynniki mają wpływ na sprzedaż(**Czas**, **Region**), ale specjalnie między sobą nie wchodzą w interakcję. W zasadzie tylko linia „C_G" zachowuje się inaczej niż pozostałe, ale kształt jej świadczy o zgodności z pozostałymi liniami natomiast spadek sprzedaży w okresie „zima-wiosna" mógł być spowodowany zapasami, których dokonali klienci na zakończenie poprzedniego roku. Dane pochodzą jeszcze z okresu gdy publikacja listy leków refundowanych była zaskoczeniem dla lekarzy, firm i pacjentów, także każda plotka mogła zmienić obraz sprzedaży. W późniejszym czasie ustawa farmaceutyczna jednoznacznie określiła termin publikacji listy leków refundowanych.

Kolejnym krokiem jest zbudowanie modelu dwuczynnikowej analizy wariancji gdzie interakcje nie będą brane pod uwagę.

```
dwuczynnikowa <- aov(Sprzedaż ~ Czas + Region, data = dane.mieszane)
summary(dwuczynnikowa)
```

```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## Czas         3 3.543e+10 1.181e+10   53.42 5.31e-10 ***
## Region       7 6.984e+10 9.977e+09   45.13 3.00e-11 ***
## Residuals   21 4.642e+09 2.211e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na podstawie F wartości i odpowiadającym im wartościom prawdopodobieństwa stwierdziłem, że **Czas** i **Region** wpływają w sposób istotny na **Sprzedaż.**

Ostatnim krokiem jest wykonanie analizy *post hoc*.

```
post_dwuczynnikowa <- TukeyHSD(dwuczynnikowa, which = c("Czas", "Region"))
post_dwuczynnikowa
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sprzedaż ~ Czas + Region, data = dane.mieszane)
##
## $Czas
##                     diff       lwr        upr     p adj
## wiosna-zima   -19126.75 -39847.65   1594.147 0.0771881
## lato-zima     -21758.50 -42479.40  -1037.603 0.0373291
## jesień-zima    60724.88  40003.98  81445.772 0.0000003
## lato-wiosna    -2631.75 -23352.65  18089.147 0.9843583
## jesień-wiosna  79851.62  59130.73 100572.522 0.0000000
## jesień-lato    82483.37  61762.48 103204.272 0.0000000
##
## $Region
##                  diff         lwr        upr     p adj
## D_F-C_G    -73298.00 -108560.835 -38035.165 0.0000166
## E_T-C_G    -25493.25  -60756.085   9769.585 0.2800001
## K_R-C_G     13470.75  -21792.085  48733.585 0.8958578
## N_B_L-C_G  -31413.25  -66676.085   3849.585 0.1040108
## P_Z-C_G      6700.50  -28562.335  41963.335 0.9978028
## S_O-C_G    101165.25   65902.415 136428.085 0.0000001
## W-C_G       -1865.75  -37128.585  33397.085 0.9999996
## E_T-D_F     47804.75   12541.915  83067.585 0.0036247
## K_R-D_F     86768.75   51505.915 122031.585 0.0000012
## N_B_L-D_F   41884.75    6621.915  77147.585 0.0128698
## P_Z-D_F     79998.50   44735.665 115261.335 0.0000044
## S_O-D_F    174463.25  139200.415 209726.085 0.0000000
## W-D_F       71432.25   36169.415 106695.085 0.0000241
## K_R-E_T     38964.00    3701.165  74226.835 0.0236833
## N_B_L-E_T   -5920.00  -41182.835  29342.835 0.9989980
## P_Z-E_T     32193.75   -3069.085  67456.585 0.0900679
## S_O-E_T    126658.50   91395.665 161921.335 0.0000000
## W-E_T       23627.50  -11635.335  58890.335 0.3651883
## N_B_L-K_R  -44884.00  -80146.835  -9621.165 0.0067973
## P_Z-K_R     -6770.25  -42033.085  28492.585 0.9976559
## S_O-K_R     87694.50   52431.665 122957.335 0.0000010
## W-K_R      -15336.50  -50599.335  19926.335 0.8197568
## P_Z-N_B_L   38113.75    2850.915  73376.585 0.0282008
## S_O-N_B_L  132578.50   97315.665 167841.335 0.0000000
## W-N_B_L     29547.50   -5715.335  64810.335 0.1450899
## S_O-P_Z     94464.75   59201.915 129727.585 0.0000003
```

```
## W-P_Z       -8566.25  -43829.085  26696.585 0.9902588
## W-S_O     -103031.00 -138293.835 -67768.165 0.0000001
```

Podsumowując, analiza **„Czas”** jest zdominowana przez okres **„jesień”**. Czyli sprzedaż w miesiącach od października do grudnia ma największy wpływ na pozostałe cykle czasu. Inne kombinacje zostały uznane za nieistotne z jednym granicznym wyjątkiem "**lato-zima**”, który jednak na siłę „**jesień**” nie brałbym pod uwagę. Zresztą pewna część recept grudniowych jest realizowana w styczniu stąd siłą rozpędu zima jest trochę „uprzywilejowana”.

Bardziej skomplikowane są kombinacje w „**Region”,** ale można założyć, że sprzedaż „S_O” ma istotny wpływ na pozostałe regiony.