Sprawozdanie 1

Yana Negulescu

Spis treści:

1. Opis danych, cel analizy, pytania badawcze.
2. Wczytanie danych, nadanie odpowiednich etykiet, typów oraz wartości, obsługa braków danych
3. Analiza danych
4. Podsumowanie

1. Opis danych

W tym sprawozdaniu będę analizowała dane dotyczące choroby otyłością u ludzi, pobrałam je ze strony internetowej Kagle: https://www.kaggle.com/datasets/myroslavsobchyshyn/obesity-dataset/data.

Dane obejmują ocenę poziomu otyłości u osób z krajów Meksyku, Peru i Kolumbii, w wieku od 14 do 61 lat, o zróżnicowanych nawykach żywieniowych i kondycji fizycznej, dane zebrano za pomocą platformy internetowej wraz z ankietą, w której anonimowi użytkownicy odpowiadali na pytania, następnie informacje zostały przetworzone, uzyskując 17 atrybutów i 2111 rekordów.

- Uzyskane zmienne: Gender, Age, Height i Weight.
- Zmienne związane z nawykami żywieniowymi: występowanie choroby otyłością w rodzinie (family_history_with_overweight), częste spożywanie żywności wysokokalorycznej (FAVC), częstotliwość spożycia warzyw (FCVC), liczba posiłków głównych (NCP), spożywanie pokarmów między posiłkami (CAEC), dzienne spożycie wody (CH20) oraz spożycie alkoholu (CALC).
- Zmienne związane ze stanem fizycznym: monitorowanie spożycia kalorii (SCC), częstotliwość aktywności fizycznej (FAF), czas korzystania z urządzeń technologicznych (TUE) oraz używany środek transportu (MTRANS), ocena stanu ciała (NObesity)

Pytanie badawcze: Jak i w jakim stopniu pewne nawyki behawioralne wpływają na otyłość u ludzi.

2. Wczytanie danych, nadanie odpowiednich etykiet, typów oraz wartości, obsługa braków danych

Pierwsze sześć wierszów:

```
# A tibble: 6 x 17
  Gender   Age Height Weight family_history_with_overw~1 FAVC   FCVC   NCP CAEC
  <chr>  <dbl>  <dbl>  <dbl> <chr>                       <chr> <dbl> <dbl> <chr>
1 Female    21   1.62   64   yes                         no        2     3 Some~
2 Female    21   1.52   56   yes                         no        3     3 Some~
3 Male      23   1.8    77   yes                         no        2     3 Some~
4 Male      27   1.8    87   <NA>                        no        3     3 Some~
5 Male      22   1.78   89.8 <NA>                        no        2     1 Some~
6 Male      29   1.62   53   <NA>                        yes       2     3 Some~
# i abbreviated name: 1: family_history_with_overweight
# i 8 more variables: SMOKE <chr>, CH2O <dbl>, SCC <chr>, FAF <dbl>, TUE <dbl>,
#   CALC <chr>, MTRANS <chr>, NObeyesdad <chr>
```

Teraz przeanalizujemy jakie mamy zmienne:

- Gender (płec): zmienna kategoryczna. Przyjmuje wartości: female/male.
- Age (wiek): zmienna ciągła. Przyjmuje wartości: 14, ..., 61.
- Height (wzrost): zmienna ciągła. Przyjmuje wartości: 1.45, ..., 1.98.
- Weight (waga): zmienna ciągła. Przyjmuje wartości: 39, ..., 173..
- family_history_with_overweight (choroby w rodzinie): zmienna kategoryczna. Przyjmuje wartości: yes/now. Zamienimy to na 1/0 dla wygodności.
- FAVC (spożywanie żywności wysokokaloryczne): zmienna kategoryczna. Przyjmuje wartości: yes/no. Zamienimy to na 1/0 dla wygodniejszej analizy.
- FCVC (ilość spożywanych ważyw dziennie): zmienna ciągła. Przyjmuje wartości: 1, ..., 3.
- NCP (ilość posiłków głównych): zmienna ciągła. Przyjmuje wartości: 1, ..., 4.
- CAEC(spożywanie pokarmów pomiędzy posiłkami): zmienna kategoryczna. Przyjmuje wartości: no/Sometimes/Frequently/Always. Zamienimy to na 0/1/2/3 dla wygodności.
- SMOKE: zmienna kategoryczna. Przyjmuje wartości: yes/no. Zamienimy to na 1/0 dla wygodności.
- CH2O(spożywanie wody w litrach): zmienna ciągła. Przyjmuje wartości: 1, ..., 3.

- CALC(częstość spożuwania alkoholu): zmienna kategoryczna. Przyjmuje wartości: no/Sometimes/Frequently/Always. Zamienimy to na 0/1/2/3 dla wygodności.
- SCC (kontrola kalorii): zmienna kategoryczna. Przyjmuje wartości: yes/no. Zamienimy to na 1/0 dla wygodności.
- FAF (godziny aktywności fizycznej): zmienna ciągła. Przyjmuje wartości: 0, ..., 3.
- TUE (czas korzystania z urządzeń w tygodniu): zmienna ciągła. Przyjmuje wartości: 0, ..., 2.
- MTRANS: zmienna kategoryczna. Przyjmuje wartości: Walking/Bike/Public_Transportation/Motorbike/
- NObeyesdad (Poziom otyłości): zmienna kategoryczna. Definiujemy ją zgodnie z wskaźnikiem BMI (wskaźnik masy ciała), jaki wynosi $\frac{waga}{wzrost^2}$. Przyjmuje wartości:
  1. Insufficient_Weight: BMI < 18,49
  2. Normal_Weight: 18,5 < BMI < 24,99
  3. Overweight_Level_I: 25,0 < BMI < 27,49
  4. Overweight_Level_II: 27,5 < BMI < 29,99
  5. Obesity_Type_I: 30,0 < BMI < 34,99
  6. Obesity_Type_II: 35,0 < BMI < 39,99
  7. Obesity_Type_III: BMI > 40

  Dla wygodności zamienimy na -1/0/1/2/3/4/5.

Obsługa braków danych:

Mamy 73 braka danych. W stosunku do 2111 rekordów to nie zbyt dużo więc podejmowałam decyzje ich usunąć.

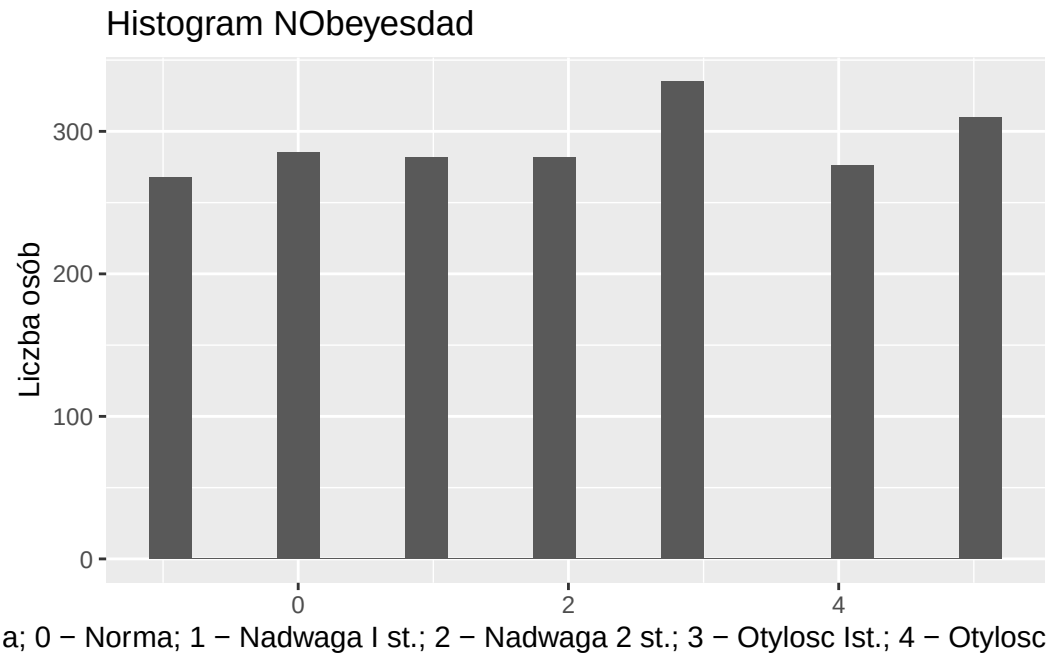
Teraz mamy 2038 rekordów oraz 17 atrybutów.

3. Analiza danych

Histogram danych `dane$NObeyesdad`

Zobaczymy histogram danych odpowiadujących za stan ciała:

```
Warning: Use of `dane$NObeyesdad` is discouraged.
i Use `NObeyesdad` instead.
```

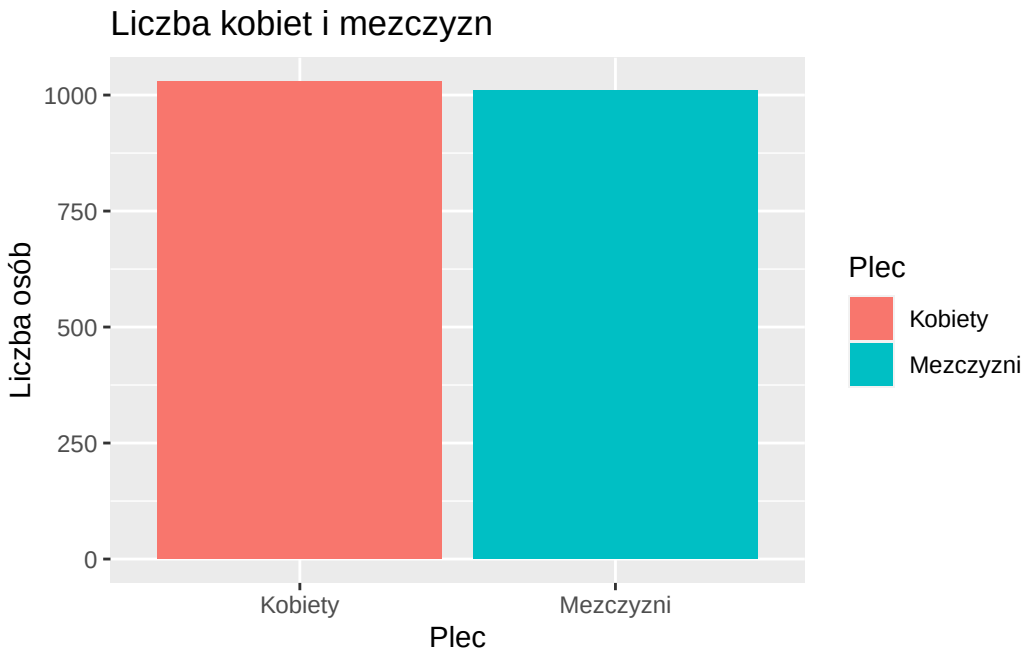


a; 0 − Norma; 1 − Nadwaga I st.; 2 − Nadwaga 2 st.; 3 − Otylosc Ist.; 4 − Otylosc I

Widzimy że w naszym datasecie przeważają dane dotyczące osób z otyłością pierwszego stopnia, jednak liczba wszystkich danych dotyczących innych stanów ciała jest mniej więcej na tym samym pożiomie.

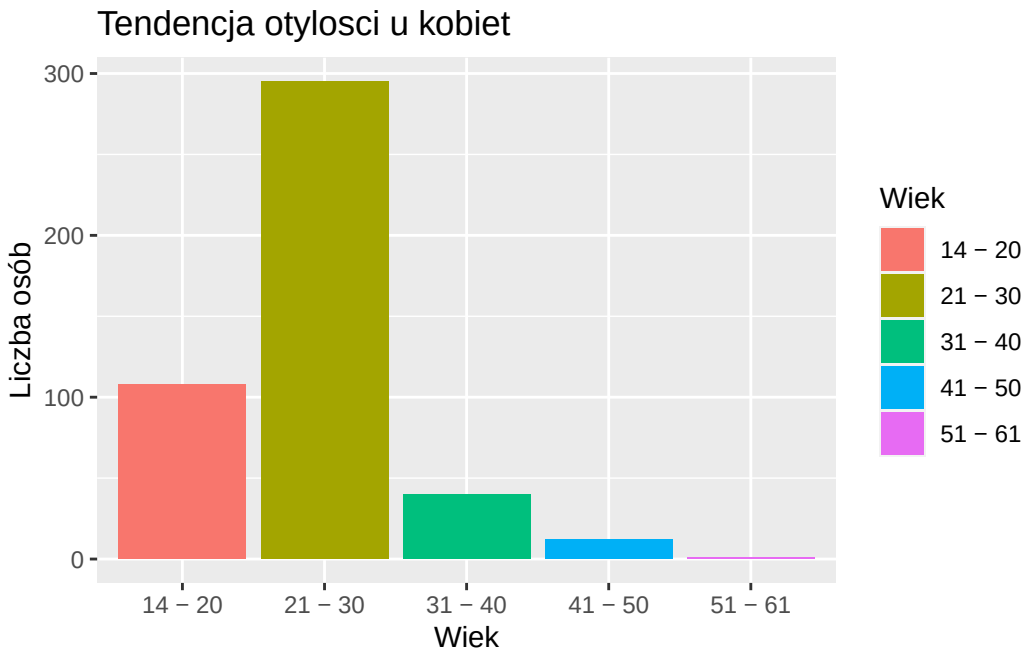
Mężczyzny i kobiety: ryzyko choroby na otyłość w zależności od wieku

Podzielimy nasze dane na kategorii mężczyżni i kobiety i sprawdżmy w jakim wieku jest najwiękrze ryzyko choroby otyłością spośród kobiet i mężczyżn:

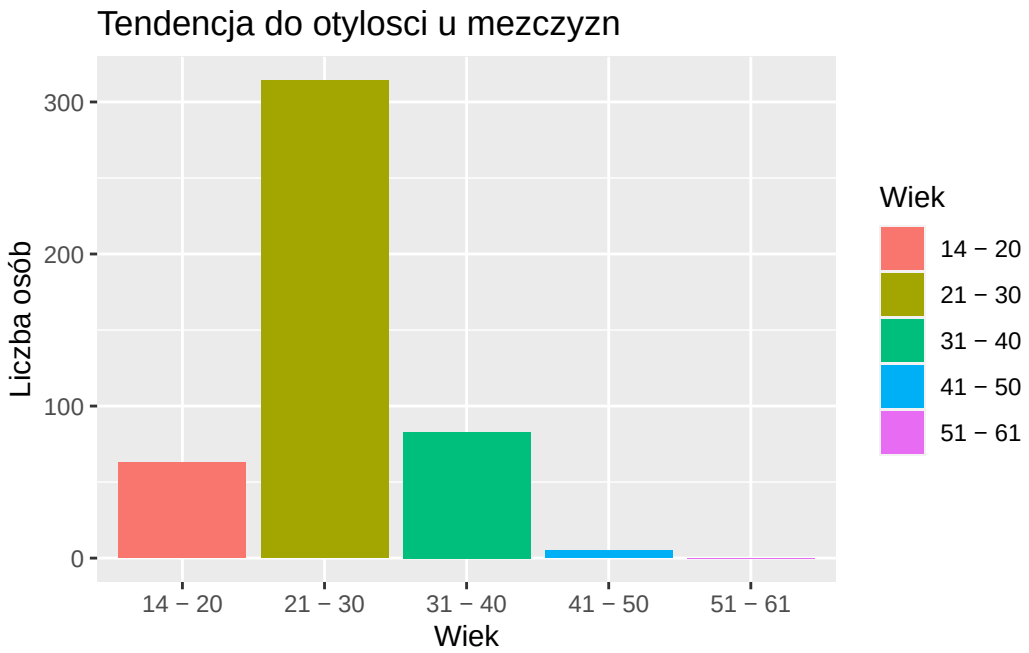


Odfiltrujemy kobiet ze stanem ciała uznawanym za akceptowalny dla zdrowia, to znaczy normal_weight, owerweight_level_1, overweight_level_2 (ostatnie dwa to stopni nadwagi, nie są chorobą) oraz kobiet chorujących na otyłość. To samo zrobimy dla kategorii mężczyżn.

Dalej odfiltrujemy wszystkich po wieku i narysujemy wykresy słupkowe.



To samo zrobimy dla kategorii mężczyżn:



Z wykresów możemy zauważyć, że najbardziej podatne na otyłość są osoby w wieku 21 - 30 lat, jednak nie wydaje się, aby to było prawdą. Przepuszczam, że skoro te dane sązostały wzięte

z wyników ankiet na stronie internetowej, nie są oni zbyt dokładne odnośnie wieku, ponieważ osoby starsze rzadziej korzystają z internetu.

Zależność zmiennej `NObeyesdad` od reszty zmiennych

R-squared wynosi 0.4312 , co oznacza że wszystkie zmienne wyjaśniają 43.12% zmiennej zależnej NObeyesdad.

Największy wpływ na zmienną `stan_ciała` mają zmienne `posilki`, `liczenie_kalorii` mają oni silną negatywną korelacje$-0.96890, -0.72495$ blisko -1, co oznacza że kiedu jedna z nich maleje zmienna `stan_ciała` rożnie i na odwrót. Też wielki wpływ mają zmienne `dużo_kalorii` i `warzywa` mają duży współczynnik dodatniej korelacji $0.60457, 0.80530$ blisko 1, to znaczy kiedy jedna z nich rośnie, zmienna stan_ciala też rośnie.

Zmienne `aktywnosc_fizyczna`, `tech_urzadzenia`, `alkohol` mają mniejszy wpływ na zmienną `stan_ciala` niż powyższe. Ich współczynniki korelacji: $-0.35769, -0.26595, 0.41770$

Zmienne `dania_glowne`, `palennie` i `woda` mają najmniejszy wpływ, ich współczynniki korelacji $0.04792, 0.19169, 0.13463$ są blisko 0.

Mężczyzny i kobiety: ryzyko choroby na otyłość w zależności od spożycia ważyw

Teraz spradźmy jaki wpływ ma spożycie warzyw na stan ciała u kobiet i meżczyżn: obliczymy podstawowe statystyki dla atrybuta FCVC(ilość sporzytych warzyw) każdej kategorii (ludzi z otyłością i bez).

Dla nas jest oczewiste że spożycie warzyw i owoców ma wpływ na zdrowość ciała, ale z naszysz obliczeń można wywnioskować że spożycie warzyw nie ma wielkiego wpływu, bo średnia dla kobiet z pierwszej kategorji wynosi 2,361 oraz odchylenie standardowe 0,5418936, co znaczy że większość wartości znajduje sie w przedziale [1,8191064; 2,9028936], oraz dla kobiet z drugiej kategorji średnia wynosi 2,755 oraz odchylenie standardowe 0,4543976, znajdują się oni w przedziale [2,3006024; 3,00]

Mamy podobne wyniki dla mężczyzn z kategorii 1 i 2, więc wnioskuę, że spożycie warzyw nie ma wpływu na wskażnik BMI.

Mężczyzny i kobiety: ryzyko choroby na otyłość w zależności od przypadków otyłości w rodzinie

Okazuje się że spośród kobiet z otyłością u 99% są członki rodziny chore na otyłość, podczas gdy u kobiet nie chorujących ten procent jest znacznie niższy - 73.69%

To samo możemy zobaczyć i u mężczyżn, gdzie u 98.92% meżczyn z otyłością w rodzinie występowały przypadki otyłości, odnośnie 73.11% u mężczyzn nie chorujących.

Oznacza to, że osoby z przypadkami choroby w rodzinie mają większe ryzyko na chorobę otyłością.

Logistyczna analiza regresji: wpływ użycia użądzeń technologcznych na zdrowie

Teraz skorzystamy z logistycznej analizy regresji do oszacowania wpłuwu użycia użądzeń technologcznych na zdrowie. Musimy skonstułować zależną zmienną binarną `NObeyesdad` (stan ciała) Dlatego odfiltrujemy dane na ludzi z otyłością i bez, oznaczmy ich 1 i 0, oraz skonstrulujemy model.

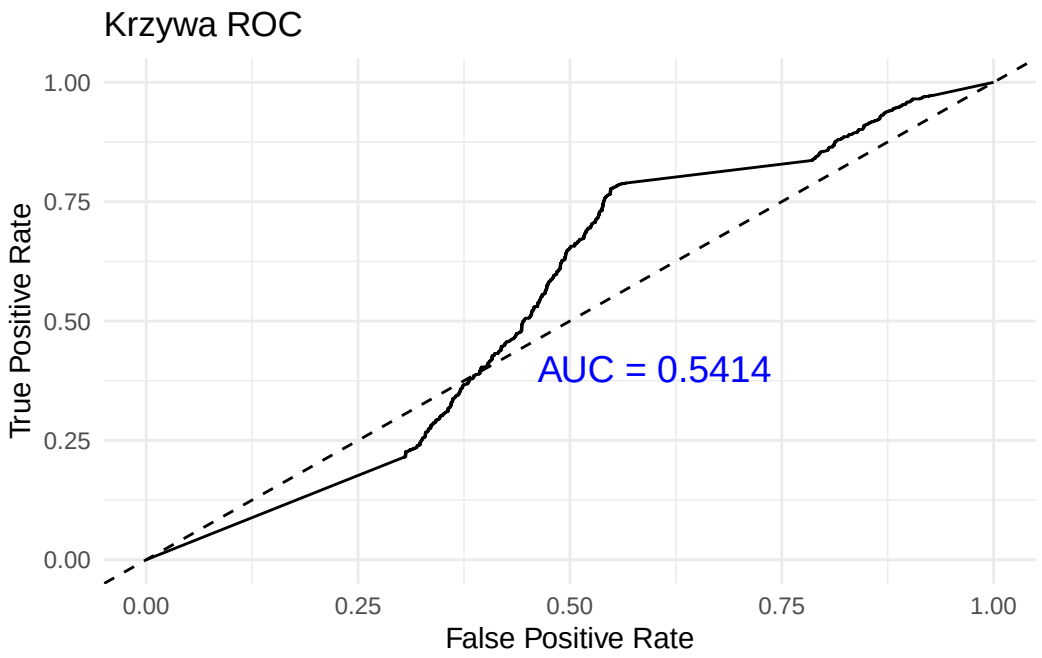
P-wartości, odpowiednio 0.908 dla wyrazu wolnego i $3.48e-05$ dla `newdata$TUE` (korzystanie z urządzeń technologicznych). Niska p-wartość wskazuje, że istnieje statystycznie istotna zależność między `newdata$TUE` a szansami na `newdata$NObeyesdad`.

Różnica między Null deviance a Residual deviance wskazuje na to, ile model tłumaczy. Zmniejszenie odchylenia resztowego wskazuje na pewne, ale nie znaczące, dopasowanie modelu.

Narysujemy krzywą ROC(Receiver Operating Characteristic), aby zobaczyć, jak dobrze `newdata$TUE` przewidują `newdata$NObeyesdad`, oraz policzymy współczynnik AUC(Area Under the Curve):

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```



Wartość AUC powyżej 0.5 oznacza, że model jest lepszy niż losowa klasyfikacja.

Wniosek: model wydaje się mieć pewną zdolność do wyjaśniania zmienności danych, ale wartości odchylenia wskazują na to, że może być jeszcze sporo niewyjaśnionej zmienności.

Logistyczna analiza regresji: wpływ aktywności fizycznej na zdrowie

Jeszcze raz skorzystamy z logistycznej analizy regresji do oszacowania wpłuwu aktywności fizycznej na zdrowie:

P-wartości dla współczynników są odpowiednio 0.0218 dla wyrazu wolnego i 8.25e-11 dla `newdata$FAF` (czas aktywności fizycznej). Niska p-wartość dla `newdata$FAF` (znacznie mniejsza niż 0.05) wskazuje, że istnieje statystycznie istotny związek między `newdata$FAF` a szansami na `newdata$NObeyesdad`.

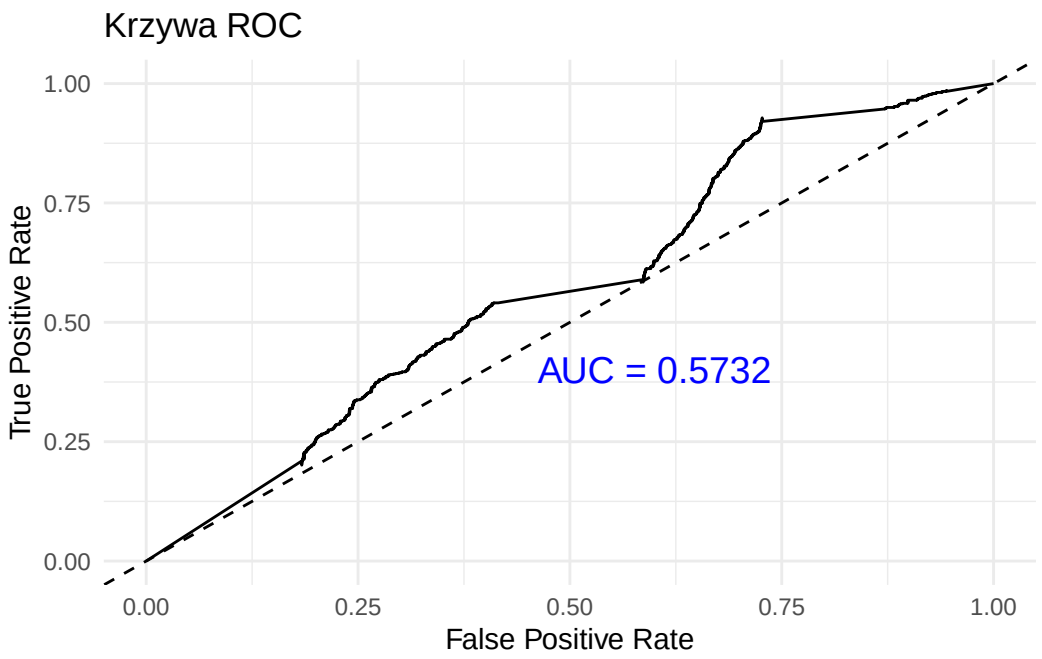
Null deviance: 2806.4 na 2037 stopniach swobody i Residual deviance: 2762.9 na 2036 stopniach swobody. Zmniejszenie odchylenia wskazuje na to, że model z `newdata$FAF` lepiej pasuje do danych niż model zawierający tylko wyraz wolny.

Wpływ `newdata$FAF` na `newdata$NObeyesdad` jest istotny statystycznie i negatywny, co znaczy przy zmniejszeniu aktywności fizycznej wzrasta szansa na chorobę otyłością.

Narysujemy krzywą ROC(Receiver Operating Characteristic), aby zobaczyć, jak dobrze `newdata$FAF` przewidują `newdata$NObeyesdad`, oraz policzymy współczynnik AUC(Area Under the Curve):

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```



AUC wynosi 0.5732 co wskazuje na to że zmienna `newdata$FAF` przewiduje ponad połowe danych.

Model ma pewną zdolność do wyjaśniania zmienności w danych, ale nadal istnieje znacząca niewyjaśniona zmienność (jak wskazuje odchylenie resztowe).

4. Podsumowanie

Przeanalizowaliśmy dane dotyczące choroby otyłością spośród kobiet i mężczyzn w różnym wieku. Wyjaśniliśmy, że prowadzenie siedzącego trybu życia, częste korzystanie z urządzeń technicznych (czyli przebywanie w pozycji statycznej), czeste spożywanie wysokokalorycznych potraw powoduje ryzyko na chorobe otyłością. Też wpływ ma przypadki otyłości w rodzinie. Z drugiej strony wyjaśniliśmy, że palenie i piecie wody nie wpływają tak bardzo na ryzyko otyłości.