Politechnika Krakowska
Wydział Informatyki i Telekomunikacji

Sprawozdanie z przedmiotu:

Projektowanie i Analiza Eksperymentów

Temat projektu:

Analiza wariancji i metoda składowych głównych

Wykonał: **Rafał Gęgotek**
Kierunek: Informatyka
Stopień studiów: II stopnia
Specjalizacja: Data Science
Rok akademicki: 2021/2022

Spis Treści

1. Cel Projektu

Celem projektu jest przeprowadzenie na indywidualnie znalezionym zbiorze danych jednokierunkowej analizy wariancji, porównań wielokrotnych a la Tukey, próby klasyfikacji, analizy dwukierunkowej wariancji z interakcjami i bez oraz wykonaniu metody PCA.

2. Zbiór danych

Zbiór danych dotyczyczy zestawienia rachunków klientów kard kredytowych. Próba zawiera kombinację danych z marca 2013 r. i danych historycznych obejmujących ostatnie 12 miesięcy przed wskazaną datą. Zestaw danych pochodzi ze strony https://leaps.analyttica.com/home, skąd został oczyszczony z danych brakujących i upubliczniony pod aresenm *https://kaggle.com/sakshigoyal7/credit-card-customers*.

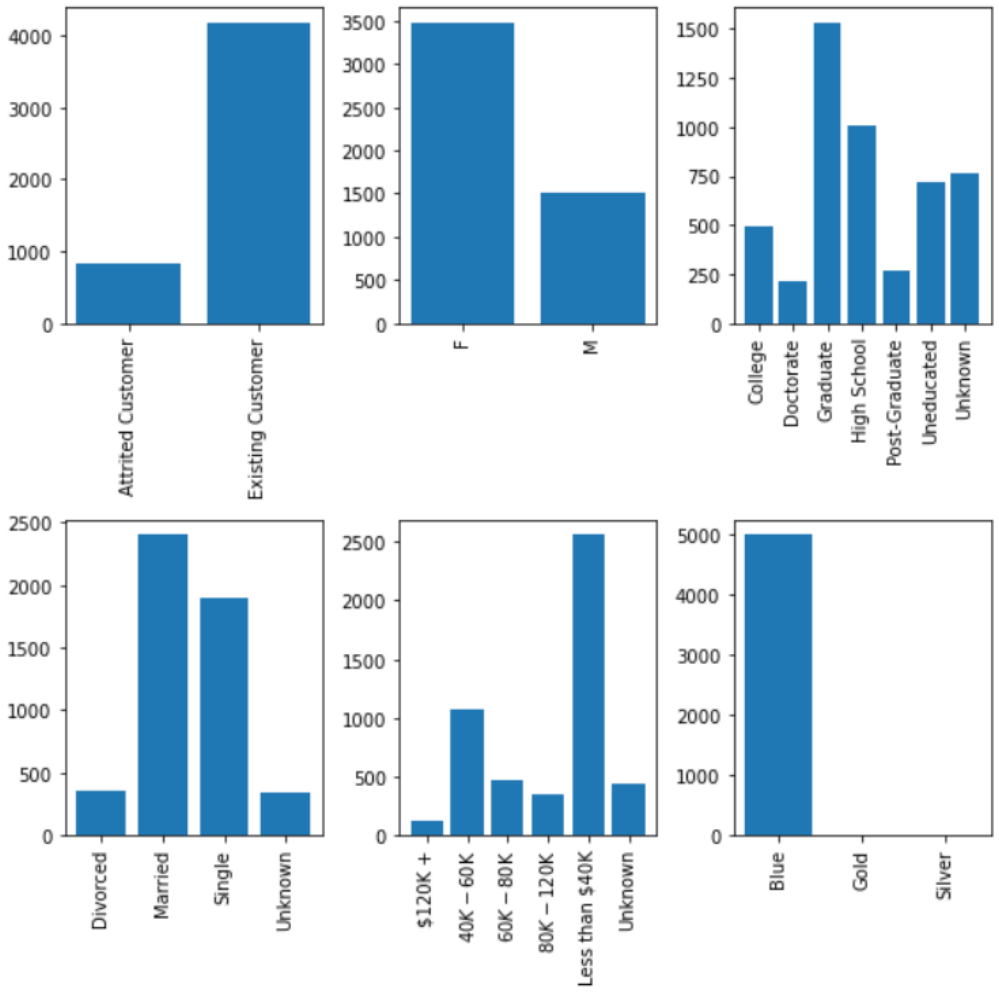
Zbiór liczy łącznie 10127 elementów, z których losowo wybrano 1000 i są opisane przez 21 cech:

Clientnum	unikalny identyfikator klienta posiadającego konto
Attrition_Flag	cecha wskazująca, czy klient dalej jest posiadaczem konta
Customer_Age	wiek klienta w latach
Gender	płeć (F-kobieta, M-mężczyzna)
Dependent_count	liczba osób na utrzymaniu
Education_Level	poziom edukacji klienta (high school, college graduate)
Marital_Status	stan cywilny (Married, Single, Unknown)
Income_Category	roczny dochód posiadacza rachunku, zmienna jakościowa (<40K, 40K - 60K,60K−80K, 80K−120K, > $120K, Unknown)
Card_Category	typ karty kredytowej (Blue, Silver, Gold, Platinum)
Months_on_book	okres relacji z bankiem
Total_Relationship_Count	liczba produktów posiadanych przez klienta
Months_Inactive_12_mon	liczba miesięcy nieaktywnych w ciągu roku
Contacts_Count_12_mon	liczba kontaktów w ciągu roku
Credit_Limit	limit na karcie kredytowej
Total_Revolving_Bal	saldo odnawialne na karcie kredytowej
Avg_Open_To_Buy	dostępne środki w ciągu miesiąca, średni z całego roku
Total_Amt_Chng_Q4_Q1	zmiany kwoty transakcji (IV kwartał do I kwartał)
Total_Trans_Amt	łączna kwota transakcji w roku
Total_Trans_Ct	całkowita liczba transakcji
Total_Ct_Chng_Q4_Q1	zmiany liczby transakcji (IV kwartał do I kwartał)
Avg_Utilization_Ratio	średni współczynnik wykorzystania karty

Poniżej fragment zbioru danych:

1	Clientnum	Attrition Flag	Customer Age	Gender	Dependent count	Education Level	Marital Status	Income Category	Card Category	Months on book	Total Relationship Count	Months Inactive 12 mon	Contacts Count 12 mon	Credit Limit	Total Revolving Bal	Avg Open To Buy	Total Amt Chng Q4 Q1	Total Trans Amt	Total Trans Ct	Total Ct Chng Q4 Q1	Avg Utilization Ratio
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4	1	0	3418	0	3418	2.594	1887	20	2.333	0
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3	4	1	3313	2517	796	1.405	1171	20	2.333	0.76
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5	1	0	4716	0	4716	2.175	816	28	2.5	0
5	713061558	Existing Customer	44	M	2	Graduate	Married	$40K - $60K	Blue	36	3	1	2	4010	1247	2763	1.376	1088	24	0.846	0.311
6	712396908	Existing Customer	57	F	2	Graduate	Married	Less than $40K	Blue	48	5	2	2	2436	680	1756	1.19	1570	29	0.611	0.279
7	714885258	Existing Customer	44	M	4	Unknown	Unknown	$80K - $120K	Blue	37	5	1	2	4234	972	3262	1.707	1348	27	1.7	0.23
8	806160108	Existing Customer	61	M	1	High School	Married	$40K - $60K	Blue	56	2	2	3	3193	2517	676	1.831	1336	30	1.143	0.788
9	784725333	Existing Customer	41	M	3	High School	Married	$40K - $60K	Blue	33	4	2	1	4470	680	3790	1.608	931	18	1.571	0.152
10	811604133	Existing Customer	47	F	4	Unknown	Single	Less than $40K	Blue	36	3	3	2	2492	1560	932	0.573	1126	23	0.353	0.626
11	806624208	Existing Customer	47	M	4	High School	Married	$40K - $60K	Blue	42	6	0	0	4785	1362	3423	0.739	1045	38	0.9	0.285
12	778348233	Existing Customer	53	M	3	Unknown	Married	$80K - $120K	Blue	33	3	2	3	2753	1811	942	0.977	1038	25	2.571	0.658
13	712991808	Existing Customer	53	M	2	Uneducated	Married	$60K - $80K	Blue	48	2	5	1	2451	1690	761	1.323	1596	26	1.6	0.69
14	788658483	Existing Customer	53	F	2	College	Married	Less than $40K	Blue	38	5	2	3	2650	1490	1160	1.75	1411	28	1	0.562
15	715318008	Existing Customer	55	F	1	College	Single	Less than $40K	Blue	36	4	2	1	3520	1914	1606	0.51	1407	43	0.483	0.544
16	713962233	Existing Customer	55	F	3	Graduate	Married	Less than $40K	Blue	36	6	2	3	3035	2298	737	1.724	1877	37	1.176	0.757
17	715190283	Existing Customer	57	F	1	Graduate	Unknown	$40K - $60K	Blue	49	3	3	2	3672	886	2786	1.32	1464	28	0.556	0.241
18	778493808	Existing Customer	49	M	3	High School	Married	$60K - $80K	Blue	37	5	2	1	3906	0	3906	1.214	1756	32	1	0
19	789172683	Existing Customer	56	M	2	Doctorate	Married	$60K - $80K	Blue	45	6	2	0	2283	1430	853	2.316	1741	27	0.588	0.626
20	738406533	Existing Customer	59	M	1	Doctorate	Married	$40K - $60K	Blue	52	3	2	2	2548	2020	528	2.357	1719	27	1.7	0.793
21	771490833	Existing Customer	52	M	1	College	Single	$80K - $120K	Blue	40	5	1	1	4745	1227	3518	0.624	1140	40	0.6	0.259
22	720756708	Existing Customer	52	F	3	Unknown	Married	Less than $40K	Blue	41	6	3	2	2622	1549	1073	1.321	1878	30	1.143	0.591
23	711525033	Existing Customer	66	F	0	High School	Married	Less than $40K	Blue	54	3	4	2	3171	2179	992	1.224	1946	38	1.923	0.687
24	717891558	Existing Customer	49	F	4	Graduate	Unknown	Less than $40K	Blue	36	6	4	2	3298	2200	1098	0.678	1052	32	0.6	0.667
25	716632758	Existing Customer	49	F	3	Graduate	Single	Less than $40K	Blue	36	2	2	0	2802	2363	439	0.75	1295	40	0.6	0.843
26	768563658	Existing Customer	56	M	2	Uneducated	Married	$40K - $60K	Blue	50	4	2	3	4458	1880	2578	1.107	1424	29	1.417	0.422
27	714091983	Existing Customer	42	M	2	High School	Single	$60K - $80K	Blue	34	4	4	3	3336	1753	1583	0.69	1168	27	1.25	0.525
28	787584108	Existing Customer	55	M	3	Unknown	Married	$80K - $120K	Blue	47	4	2	3	3436	2016	1420	0.901	1097	33	0.833	0.587
29	788730933	Existing Customer	44	F	2	Uneducated	Single	Less than $40K	Blue	20	6	3	3	2084	1468	616	1.004	1132	28	0.556	0.704
30	711314058	Existing Customer	49	M	2	Graduate	Married	$60K - $80K	Blue	32	2	2	2	1687	1107	580	1.715	1670	17	2.4	0.656
31	720096558	Existing Customer	55	F	2	Graduate	Married	Less than $40K	Blue	42	5	3	3	2216	1034	1182	0.758	1540	36	0.286	0.467

Tak natomiast prezentuję się rozstaw grup dla zmiennych typu jakościowego:



Poniżej zostały również przedstawione główne statystyki dla zmiennych typu ilościowego:

	count	mean	std	min	25%	50%	75%	max
Customer_Age	1000.0	46.333000	8.188355	26.0	41.00000	46.0000	52.00000	67.000
Dependent_count	1000.0	2.294000	1.317308	0.0	1.00000	2.0000	3.00000	5.000
Months_on_book	1000.0	35.884000	8.022400	13.0	31.00000	36.0000	40.00000	56.000
Total_Relationship_Count	1000.0	3.899000	1.520234	1.0	3.00000	4.0000	5.00000	6.000
Months_Inactive_12_mon	1000.0	2.393000	1.049596	0.0	2.00000	2.0000	3.00000	6.000
Contacts_Count_12_mon	1000.0	2.418000	1.094749	0.0	2.00000	2.0000	3.00000	6.000
Credit_Limit	1000.0	2930.112000	980.744061	1439.0	2169.75000	2744.5000	3468.75000	5282.000
Total_Revolving_Bal	1000.0	1224.064000	797.904941	0.0	697.50000	1354.5000	1794.00000	2517.000
Avg_Open_To_Buy	1000.0	1706.048000	1177.993877	14.0	786.75000	1392.0000	2384.25000	5267.000
Total_Amt_Chng_Q4_Q1	1000.0	0.763668	0.227866	0.0	0.63700	0.7410	0.85350	2.357
Total_Trans_Amt	1000.0	3750.410000	2065.553992	647.0	2272.00000	4022.5000	4639.50000	15867.000
Total_Trans_Ct	1000.0	63.371000	20.845188	12.0	45.00000	68.0000	80.00000	131.000
Total_Ct_Chng_Q4_Q1	1000.0	0.718963	0.245290	0.0	0.57700	0.7140	0.83300	3.000
Avg_Utilization_Ratio	1000.0	0.439850	0.292068	0.0	0.20775	0.4945	0.67525	0.992

3. Opis problemu

Rozpatrywanym problemem w ramach analizy wariancji będzie zbadanie czy istnieją statystycznie istotne różnice w grupach poziomu edukacyjnego i grupach stanu cywilnego w zależności do kwoty limitu na karcie kredytowej.

W drugiej części projekt przy użyciu metody PCA zostanie sprawdzone w jakim stopniu i ile głównych składowych zawierających zmienne ilościowe z zestawu danych objaśnia parametr poziomu edukacji.

4. Jednoczynnikowa analiza wariancji

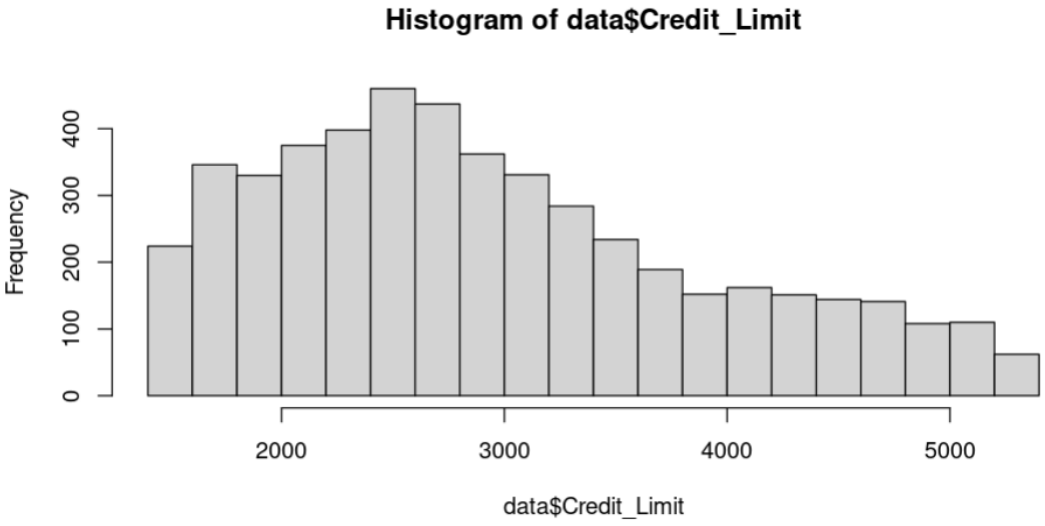
4.1 Sprawdzenie założeń do analizy wariancji

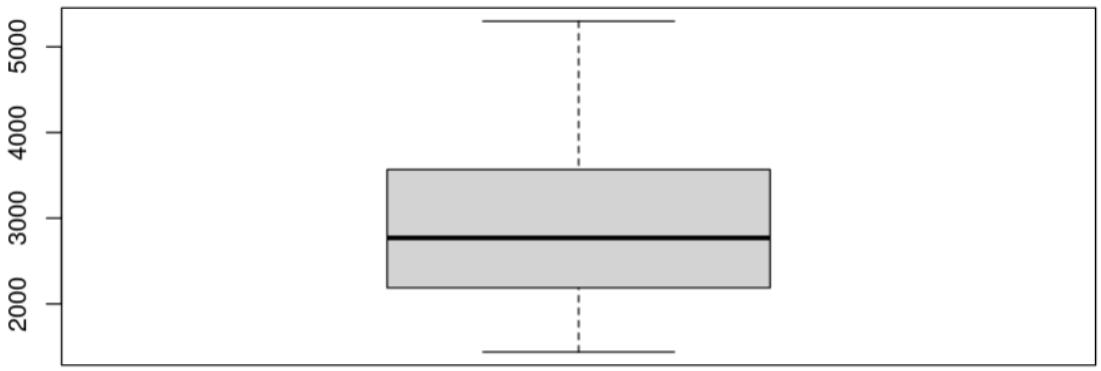
Aby przy testowaniu układu hipotez możliwe było posługiwanie się metodami analizy wariancji Anova muszą być spełnione poniższe założenia:

- zmienna zależna powinna być typu ilościowego
- Próby wybiera się losowo, a także niezależnie od siebie, z każdej z populacji.
- Każda z badanych populacji cechuje się rozkładem normalnym
- W analizowanych populacjach wariancje są takie same (homogeniczność)

4.1.1 Normalność rozkładu

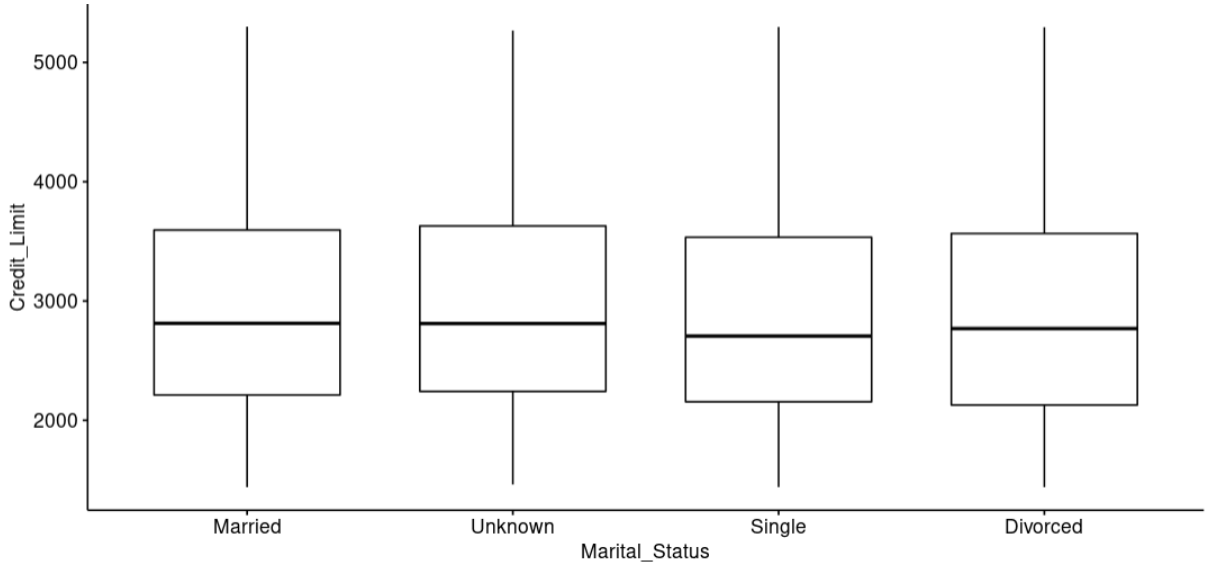
Poniżej zostały zamieszczone histogram ilustrujący rozkład danych dla parametru limitu na karcie kredytowe, oraz wykres boxplot, z którego można odczytać, iż nie występują dane odstające.



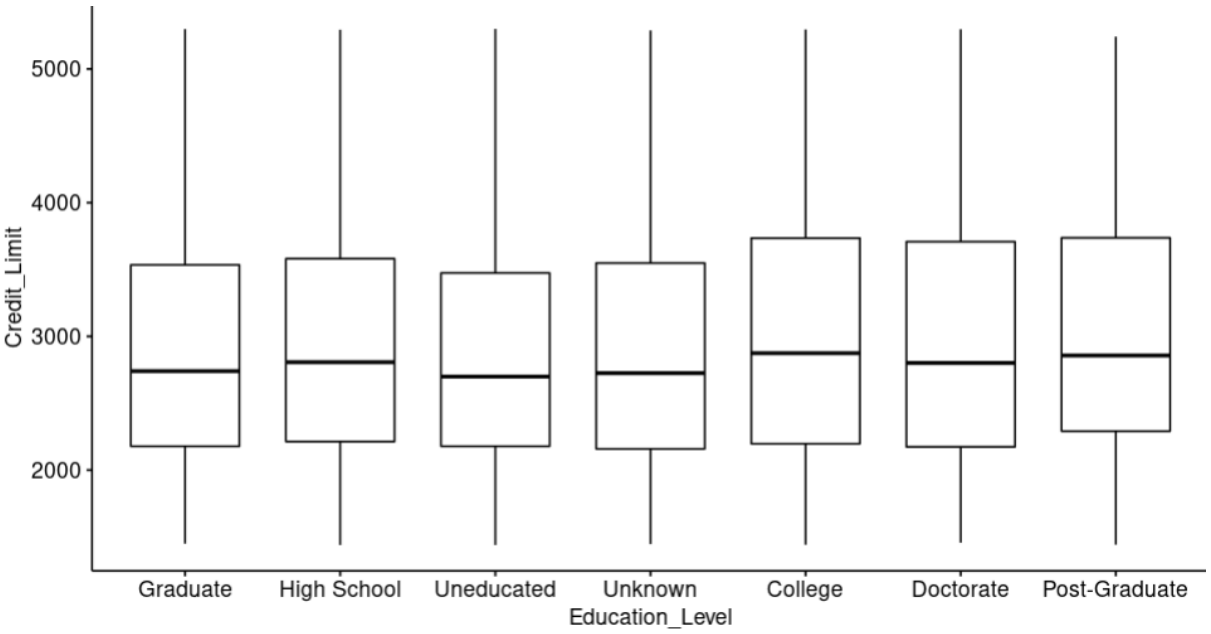


Kolejne wykresy pudełkowe prezentują zależności dla konkretnych podgrup w uwzględnianych czynnikach analizy wariancji względem limitu karty kredytowej:

- wykres Box Plot dla czynnika stanu cywilnego - brak danych odstających



- wykres Box Plot dla czynnika poziomu edukacji - brak danych odstających



Główną metodą służącą do sprawdzenie normalności rozkładu jest test Shapiro-Wilka, w ramach którego H0 wskazuje, iż próba pochodzi z rozkładu normalnego (warunek: p > 0.05).

- test Shapiro-Wilka dla parametru limitu karty kredytowej
  - odrzucenie hipotezy zerowej

```
> shapiro.test(data$Credit_Limit)

        Shapiro-Wilk normality test

data:  data$Credit_Limit
W = 0.95044, p-value < 2.2e-16
```

- test Shapiro-Wilka dla czynnika stanu cywilnego, dla każdej jego podgrupy względem zmiennej objaśnianej
  - odrzucenie hipotezy zerowej dla każdej grupy czynnika

```
> splitMarital_Status = split(data, data$Marital_Status)
> shapiro.test(splitMarital_Status$"Divorced"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitMarital_Status$Divorced$Credit_Limit
W = 0.95153, p-value = 2.152e-09

> shapiro.test(splitMarital_Status$"Married"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitMarital_Status$Married$Credit_Limit
W = 0.95502, p-value < 2.2e-16

> shapiro.test(splitMarital_Status$"Single"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitMarital_Status$Single$Credit_Limit
W = 0.94327, p-value < 2.2e-16

> shapiro.test(splitMarital_Status$"Unknown"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitMarital_Status$Unknown$Credit_Limit
W = 0.94828, p-value = 1.46e-09
```

- test Shapiro-Wilka dla dla czynnika poziomu edukacji, kolejna dla każdej jego podgrupy względem zmiennej objaśnianej
  - odrzucenie hipotezy zerowej dla każdej grupy czynnika

```
> shapiro.test(splitEducation_Level$"College"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$College$Credit_Limit
W = 0.9535, p-value = 2.119e-11

> shapiro.test(splitEducation_Level$"Doctorate"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$Doctorate$Credit_Limit
W = 0.94588, p-value = 3.554e-07

> shapiro.test(splitEducation_Level$"Graduate"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$Graduate$Credit_Limit
W = 0.94815, p-value < 2.2e-16

> shapiro.test(splitEducation_Level$"High School"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$"High School"$Credit_Limit
W = 0.95614, p-value < 2.2e-16

> shapiro.test(splitEducation_Level$"Post-Graduate"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$"Post-Graduate"$Credit_Limit
W = 0.94853, p-value = 5.104e-08

> shapiro.test(splitEducation_Level$"Uneducated"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$Uneducated$Credit_Limit
W = 0.94637, p-value = 1.756e-15

> shapiro.test(splitEducation_Level$"Unknown"$Credit_Limit)

        Shapiro-Wilk normality test

data:  splitEducation_Level$Unknown$Credit_Limit
W = 0.94366, p-value < 2.2e-16
```

Jak wykazały testy Shapiro założenia normalności rozkładu danych nie zostało spełnione. Dla każdego testu wartość **p** była zdecydowanie mniejsza niż 5%, co było podstawą do odrzucenia H0.
Pomimo iż założenie normalności, które jest wymogiem do wykonaniu istotnej analizy wariancji przy użyciu Anova, nie zostało spełnione, to kolejne kroki sprawdzenia założeń zostaną wykonane.

4.1.2 Sprawdzenie homogeniczności

W celu zbadania homogeniczności zostały wykorzystany Test Levene'a oraz Barletta. Test Levene'a jest alternatywą dla Bartlet'a, ale jest mniej wrażliwy na odstępstwa od normalności. Dla obu technik jeżeli p-value jest mniejsze niż 5% to możemy stwierdzić, że występują różnice między wariancjami w porównywanych grupach (H0).

- test Levene'a dla dla czynników poziomu edukacji i stanu cywilnego
  - brak podstaw do odrzucenia hipotezy zerowej dla obu czynnika

```
> leveneTest(data = data, Credit_Limit ~ Education_Level)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   6   0.774 0.5904
      993
> leveneTest(data = data, Credit_Limit ~ Marital_Status)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  0.4753 0.6995
      996
```

- test Bartlet'a dla dla czynników poziomu edukacji i stanu cywilnego
  - brak podstaw do odrzucenia hipotezy zerowej dla obu czynnika

```
> bartlett.test(Credit_Limit ~ Education_Level, data)

        Bartlett test of homogeneity of variances

data:  Credit_Limit by Education_Level
Bartlett's K-squared = 2.4707, df = 6, p-value = 0.8717

> bartlett.test(Credit_Limit ~ Marital_Status, data)

        Bartlett test of homogeneity of variances

data:  Credit_Limit by Marital_Status
Bartlett's K-squared = 0.80293, df = 3, p-value = 0.8488
```

Oba testy wykazały brak różnic między wariancjami w porównywanych grupacha, a więc można założyc homogeniczność wariancji.

4.1.3 Niezależność wariancji

Dane zostały wybrane niezależnie w sposób losowy, tak jak zostało przedstawione na poniższych poleceniach:

```
> data <- read.table("BankChurners.csv", header = TRUE, sep = ",")
> length(data$Credit_Limit)
[1] 10127
> data <- data[sample(nrow(data),1000),]
> length(data$Credit_Limit)
[1] 1000
```

4.1.4 Typ ilościowy

Zmienna zależna jest typu ilościowego co zostało już wcześniej pokazane na histogramie, dodatkowo poniżej zostały dodane krótkie podsumowanie statystyk tego parametru.

```
> summary(data$Credit_Limit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1439    2187    2770    2949    3566    5299
```

4.1.5 Wnioski do założeń analizy wariancji

Założenia analizy wariancji nie zostały spełnione, dlatego też wyniki przeprowadzenia analizy wariancji Anova będą nieistotne.
Mimo tego faktu w kolejnych etapach projektu w celu edukacyjnym zostaną zrealizowane czynności analizy przy pomocy metody Anova oraz testu Kruskala-Wallisa, który jest rekomendowanym rozwiązaniem analizy wariancji w przypadku braku normalności rozkładu.

4.2 Jednoczynnikowa analiza wariancji Anova

4.2.1 Dla czynnika poziomu edukacji

Przeprowadzona analiza **Anova** wskazała wartość p znacznie większą niż 5%. Dzięki temu można stwierdzić, że pomiędzy grupami nie występują statystycznie istotne różnice, a więc nie ma podstaw do odrzucenia hipotezy zerowej o równości grup dla czynnika poziomu edukacji - poziom edukacji nie wpływa na limit na karcie kredytowej.

```
> daneanova <- lm(Credit_Limit ~ Education_Level, data = data)
> anova(daneanova)
Analysis of Variance Table

Response: Credit_Limit
                 Df    Sum Sq Mean Sq F value Pr(>F)
Education_Level   6   7348528 1224755  1.2257 0.2904
Residuals       993 992243055  999238
```

Ze względu na potwierdzenie hipotezy zerowej, nie jest konieczne przeprowadzenie testów post-hoc.
Jednakże w celu przedstawienia wartości p dla wszystkich kombinacji grup wykonano **test Tukeya**. Dla większości kombinacji współczynnik p wynosi blisko 1, tak więc pomiędzy grupami nie ma statystycznie istotnych różnic.

```
> TukeyHSD(aov(daneanova))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = daneanova)

$Education_Level
                                diff       lwr      upr     p adj
Doctorate-College         -22.891282 -575.6218 529.8392 0.9999997
Graduate-College         -273.713034 -620.4288  73.0027 0.2295574
High School-College      -237.756589 -606.7779 131.2647 0.4784090
Post-Graduate-College    -179.220799 -699.5863 341.1447 0.9500513
Uneducated-College       -245.441626 -631.0498 140.1665 0.4938168
Unknown-College          -239.592199 -632.8369 153.6525 0.5483237
Graduate-Doctorate       -250.821752 -740.8760 239.2325 0.7376028
High School-Doctorate    -214.865308 -720.9464 291.2158 0.8723411
Post-Graduate-Doctorate  -156.329517 -781.4068 468.7478 0.9901211
Uneducated-Doctorate     -222.550344 -740.8505 295.7498 0.8661987
Unknown-Doctorate        -216.700917 -740.7074 307.3055 0.8857643
High School-Graduate       35.956444 -230.1617 302.0746 0.9996865
Post-Graduate-Graduate     94.492235 -358.7432 547.7277 0.9963205
Uneducated-Graduate        28.271408 -260.4080 316.9508 0.9999523
Unknown-Graduate           34.120835 -264.6826 332.9243 0.9998824
Post-Graduate-High School  58.535790 -411.9823 529.0539 0.9998060
Uneducated-High School     -7.685036 -322.8052 307.4351 1.0000000
Unknown-High School        -1.835609 -326.2558 322.5846 1.0000000
Uneducated-Post-Graduate  -66.220827 -549.8573 417.4156 0.9996614
Unknown-Post-Graduate     -60.371400 -550.1182 429.3754 0.9998161
Unknown-Uneducated          5.849427 -337.3202 349.0190 1.0000000
```

Ze względu na niespełnienie założeń analizy wariancji wykonano analizę nieparametrycznym odpowiednikiem ANOVA jakim jest test **test Krukskala Wallisa**.

```
> kruskal.test(Credit_Limit ~ Education_Level, data = data)

        Kruskal-Wallis rank sum test

data:  Credit_Limit by Education_Level
Kruskal-Wallis chi-squared = 7.632, df = 6, p-value = 0.2663
```

Wartość p-value wskazała wartość znacznie większą niż 5%, tak więc nie ma podstaw do odrzucenia hipotezy zerowej. Można stwierdzić że dla czynnika poziomu edukacji nie istnieją statystyczne różnice między grupami, czyli poziom edukacji nie wpływa na limit na karcie kredytowej.

4.2.2 Dla czynnika stanu cywilnego

Przeprowadzona analiza Anova wskazała wartość p znacznie mniejsze niż 5%. Dzięki temu można stwierdzić, że pomiędzy grupami występują statystycznie istotne różnice, a więc można odrzucenia hipotezy zerowej o równości grup dla czynnika stanu cywilnego - rodzaj stanu cywilnego ma wpływa na limit na karcie kredytowej (oczywiście tylko jeżeli spełnione są założenia analizy wariancji).

```
> daneanova <- lm(Credit_Limit ~ Marital_Status, data = data)
> anova(daneanova)
Analysis of Variance Table

Response: Credit_Limit
               Df    Sum Sq Mean Sq F value   Pr(>F)
Marital_Status   3  12380019 4126673  4.1634 0.006078 **
Residuals      996 987211563  991176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ze względu na odrzucenie hipotezy zerowej, adekwatnym jest przeprowadzenie testów post-hoc. W tym celu przedstawienia wartości p dla wszystkich kombinacji grup poprzez test Tukeya. Można zaobserwować, że tylko jedna z 6 kombinacji jest statystycznie znacząca, gdzie wartość p_adj jest mniejsza od 5%. Zależność ta występuje między grupami singli i osób w związku małżeńskim.

```
> TukeyHSD(aov(daneanova))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = daneanova)

$Marital_Status
                       diff        lwr       upr     p adj
Married-Divorced   65.21702 -241.15715 371.59120 0.9471493
Single-Divorced  -156.04588 -468.75440 156.66264 0.5732572
Unknown-Divorced  148.23672 -264.02763 560.50107 0.7913071
Single-Married   -221.26290 -398.91035 -43.61545 0.0075954
Unknown-Married    83.01970 -239.05776 405.09716 0.9108296
Unknown-Single    304.28260  -23.82619 632.39138 0.0803142
```

Ponieważ grupy nie są równoliczne wykonano również testy Dunnetta:

```
> DunnettTest(x=data$Credit_Limit, g=data$Marital_Status,control = "Married")

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$Married
                     diff    lwr.ci    upr.ci   pval
Divorced-Married  -65.21702 -349.2138 218.77973 0.9238
Single-Married   -221.26290 -385.9351 -56.59073 0.0041 **
Unknown-Married    83.01970 -215.5334 381.57278 0.8741

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> DunnettTest(x=data$Credit_Limit, g=data$Marital_Status,control = "Unknown")

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$Unknown
                    diff    lwr.ci    upr.ci   pval
Divorced-Unknown -148.2367 -513.8107 217.33731 0.5932
Married-Unknown   -83.0197 -368.6208 202.58141 0.7791
Single-Unknown   -304.2826 -595.2320 -13.33323 0.0384 *

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> DunnettTest(x=data$Credit_Limit, g=data$Marital_Status,control = "Divorced")

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$Divorced
                     diff    lwr.ci   upr.ci   pval
Married-Divorced   65.21702 -207.9844 338.4184 0.8630
Single-Divorced  -156.04588 -434.8958 122.8040 0.3738
Unknown-Divorced  148.23672 -219.3896 515.8630 0.6066

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> DunnettTest(x=data$Credit_Limit, g=data$Marital_Status,control = "Single")

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$Single
                   diff       lwr.ci   upr.ci   pval
Divorced-Single 156.0459 -133.3124327 445.4042 0.4694
Married-Single  221.2629   56.8805378 385.6453 0.0041 **
Unknown-Single  304.2826    0.6739739 607.8912 0.0494 *

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testy potwierdziły wcześniejsze wskazania, że istnieją kombinacje statystycznie istotne, a dokładniej istnieją dwie takie kombinacje jako Single-Unknown oraz Single-Married.

Ze względu na niespełnienie założeń analizy wariancji wykonano analizę nieparametrycznym odpowiednikiem ANOVA jakim jest test **Krukskala Wallisa**.

```
> kruskal.test(Credit_Limit ~ Marital_Status, data = data)

        Kruskal-Wallis rank sum test

data:  Credit_Limit by Marital_Status
Kruskal-Wallis chi-squared = 14.379, df = 3, p-value = 0.002432
```

Wartość p-value wskazała wartość mniejszą niż 5%, tak więc można odrzucić hipotezę zerową. Można stwierdzić że dla czynnika stanu cywilnego istnieją statystyczne różnice między grupami, czyli rodzaj stanu cywilnego ma wpływa na limit na karcie kredytowej.

5. Dwuczynnikowa analiza wariancji

5.1 Założenia do analizy wariancji

5.1.1 Niezależność wariancji, typ ilościowy

W związku że analizujemy te same parametry z zestawu danych co dla jednoczynnikowej analizy wariancji, to możemy potwierdzić spełnienie dwóch założeń jako o typie ilościowym zmiennej objaśniającej oraz o niezależności wariancji, co zostało wykazane we wcześniejszej części projektu (pkt 5.1.)

5.1.2 Homogeniczność wariancji

Test Levene'a wskazał wartość p znacznie większą niż niż 5%, dlatego też nie ma podstaw do odrzucenia hipotezy zerowej o braku różnic między wariancjami w porównywanych grupach.

```
> leveneTest(data = data, Credit_Limit ~ Marital_Status * Education_Level)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group  27  0.4493 0.9934
      972
```

5.1.2 Normalność rozkładu danych

Ponownie przy pomocy testu Shapiro sprawdzono normalność rozkładu dla każdej z kombinacji badanych grup. Wyniki testu pokazują, że dla blisko połowy możliwych kombinacji wartość p była mniejsza niż 5%, a więc należy odrzucić hipotezę zerową, a dane nie pochodzą z rozkładu normalnego.

```
> print(data %>%
+   group_by(Education_Level, Marital_Status) %>%
+   shapiro_test(Credit_Limit)
+ , n=40)
# A tibble: 28 × 5
   Education_Level Marital_Status variable     statistic          p
   <chr>           <chr>          <chr>            <dbl>      <dbl>
 1 College         Divorced       Credit_Limit     0.908 0.337
 2 College         Married        Credit_Limit     0.956 0.0470
 3 College         Single         Credit_Limit     0.951 0.131
 4 College         Unknown        Credit_Limit     0.928 0.496
 5 Doctorate       Divorced       Credit_Limit     0.877 0.316
 6 Doctorate       Married        Credit_Limit     0.897 0.0608
 7 Doctorate       Single         Credit_Limit     0.944 0.396
 8 Doctorate       Unknown        Credit_Limit     0.945 0.550
 9 Graduate        Divorced       Credit_Limit     0.898 0.125
10 Graduate        Married        Credit_Limit     0.951 0.0000502
11 Graduate        Single         Credit_Limit     0.925 0.00000916
12 Graduate        Unknown        Credit_Limit     0.949 0.232
13 High School     Divorced       Credit_Limit     0.902 0.0624
14 High School     Married        Credit_Limit     0.945 0.000711
15 High School     Single         Credit_Limit     0.914 0.0000700
16 High School     Unknown        Credit_Limit     0.872 0.0360
17 Post-Graduate   Divorced       Credit_Limit     0.969 0.662
18 Post-Graduate   Married        Credit_Limit     0.937 0.0755
19 Post-Graduate   Single         Credit_Limit     0.941 0.362
20 Post-Graduate   Unknown        Credit_Limit     0.994 0.853
21 Uneducated      Divorced       Credit_Limit     0.947 0.374
22 Uneducated      Married        Credit_Limit     0.948 0.00566
23 Uneducated      Single         Credit_Limit     0.921 0.00210
24 Uneducated      Unknown        Credit_Limit     0.885 0.178
25 Unknown         Divorced       Credit_Limit     0.918 0.299
26 Unknown         Married        Credit_Limit     0.948 0.00679
27 Unknown         Single         Credit_Limit     0.921 0.000541
28 Unknown         Unknown        Credit_Limit     0.846 0.0252
```

5.1.2 Wnioski do założeń

Założenia analizy wariancji nie zostały spełnione, dlatego też wyniki przeprowadzenia analizy wariancji Anova będą nieistotne.
Mimo tego faktu w kolejnych etapach projektu w celu edukacyjnym zostaną zrealizowane czynności analizy przy pomocy metody Anova oraz testu Friedmana, który jest rekomendowanym rozwiązaniem nieparametrycznej dwuczynnikowej analizy wariancji.

5.2 Analiza bez interakcji

Na podstawie testu Anova dla wariancji dwuczynnikowej bez interakcji możemy zaobserwować, żę dla obu czynników wartość p wskazuje na statystycznie istotne różnice między wariancjami, chociaż dla poziomu edukacji wartość ta jest graniczna na poziomie lekko powyżej 5%.

```
> res.aov3 <- aov(Credit_Limit ~ Education_Level + Marital_Status, data = data)
> summary(res.aov3)
                Df    Sum Sq Mean Sq F value Pr(>F)
Education_Level   6  11306594 1884432   2.029 0.0593 .
Marital_Status    3   9789539 3263180   3.513 0.0148 *
Residuals       990 919670426  928960
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ze względu na odrzucenie hipotezy zerowej, adekwatnym było przeprowadzenie testów post-hoc. W tym celu przedstawienia wartości p dla wszystkich kombinacji grup poprzez test Tukeya.
Podobnie jak w analizie jednoczynnikowej dla parametru statusu cywilnego tylko jedna kombinacja spełnia warunek p < 5%, jest to Single-Married.
Natomiast dla czynnika poziomu edukacji wszystkie wartości p przekraczają 5%.

```
> TukeyHSD(res.aov3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Credit_Limit ~ Education_Level + Marital_Status, data = data)

$Education_Level
                               diff       lwr       upr     p adj
Doctorate-College          -344.746535 -855.1090 165.61592 0.4179961
Graduate-College           -323.365522 -648.8521   2.12102 0.0528163
High School-College        -238.194154 -583.0036 106.61525 0.3894136
Post-Graduate-College       -86.429868 -585.6373 412.77759 0.9986998
Uneducated-College         -349.536675 -720.1897  21.11632 0.0794491
Unknown-College            -299.042187 -671.9212  73.83684 0.2127655
Graduate-Doctorate           21.381013 -432.3234 475.08546 0.9999994
High School-Doctorate       106.552381 -361.2080 574.31275 0.9940220
Post-Graduate-Doctorate     258.316667 -332.5423 849.17564 0.8558978
Uneducated-Doctorate         -4.790141 -491.9142 482.33395 1.0000000
Unknown-Doctorate            45.704348 -443.1157 534.52436 0.9999637
High School-Graduate         85.171368 -168.3467 338.68944 0.9556494
Post-Graduate-Graduate      236.935654 -204.1834 678.05466 0.6908625
Uneducated-Graduate         -26.171154 -313.8542 261.51189 0.9999691
Unknown-Graduate             24.323335 -266.2221 314.86879 0.9999811
Post-Graduate-High School   151.764286 -303.7991 607.32768 0.9574170
Uneducated-High School     -111.342522 -420.7186 198.03357 0.9385736
Unknown-High School         -60.848033 -372.8876 251.19154 0.9974613
Uneducated-Post-Graduate   -263.106808 -738.5310 212.31735 0.6595559
Unknown-Post-Graduate      -212.612319 -689.7740 264.54935 0.8443541
Unknown-Uneducated           50.494489 -289.8859 390.87488 0.9994622

$Marital_Status
                     diff        lwr       upr     p adj
Married-Divorced  176.35633 -128.35119 481.06385 0.4442673
Single-Divorced    -8.36651 -317.86886 301.13583 0.9998792
Unknown-Divorced  265.79509 -159.34794 690.93812 0.3740208
Single-Married   -184.72284 -354.56955 -14.87613 0.0268155
Unknown-Married    89.43876 -247.90795 426.78546 0.9038589
Unknown-Single    274.16160  -67.52222 615.84541 0.1654439
```

5.3 Analiza z interakcjami

Zastosowanie analizy wariancji dwuczynnikowej z interakcjami przyniosło pozytywny efekt, gdyż dla obu czynników wartość p miałą mniejsze wartości niż przy modelu bez interakcji. Ponadto okazała się, że najbardziej istotna statystycznie różnice między wariancjami są dla interakcji poziomu edukacji ze statusem cywilnym.

```
> res.aov4 <- aov(Credit_Limit ~ Education_Level * Marital_Status, data = data)
> summary(res.aov4)
                              Df    Sum Sq Mean Sq F value   Pr(>F)
Education_Level                6  11306594 1884432   2.080 0.053081 .
Marital_Status                 3   9789539 3263180   3.602 0.013141 *
Education_Level:Marital_Status 18  39174072 2176337   2.403 0.000894 ***
Residuals                     972 880496355  905860
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.4 Analiza na podstawie testu Friedmana

Ze względu na niespełnienie założeń dwuczynnikowej analizy wariancji wykonano analizę nieparametrycznym odpowiednikiem ANOVA jakim dla analizy dwuskładnikowej jest test **Friedmana**.

In [14]:
```
from scipy import stats

Marital_Status_class, Marital_Status = np.unique(data["Marital_Status"], return_inverse=True)
Education_Level_class, Education_Level = np.unique(data["Education_Level"], return_inverse=True)

stats.friedmanchisquare(data["Credit_Limit"], Education_Level, Marital_Status)
```

Out[14]:
```
FriedmanchisquareResult(statistic=13.351351351351344, pvalue=0.0012612201221243594)
```

Test wykonany z pomocą narzędzi w języku python wskazał wartość p znacznie mniejszą niż 5%. Dlatego też możemy odrzucić hipotezę zerową i stwierdzić że jest istotna statystycznie różnica limitu na karcie kredytowej pomiędzy grupami.

Przeprowadzony został równie test pairwise z metodą Bonferroniego, aby zbadać między którymi grupami zachodzą istotne różnice.
Dla zmiennej objaśnianej (limit na karcie kredytowej) istotna statystycznie różnica (wartość p mniejsza od 5%) została wykryta dla czynnika stanu cywilnego a dokładniej między grupą osób rozwiedzionych, a grupą singli i osób w związku małżeńskim. Dla czynnika poziomu edukacji nie ma istotnych statystycznie różnic.

```
> pairwise.wilcox.test(data$Credit_Limit, data$Marital_Statu, p.adj="bonferroni")

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  data$Credit_Limit and data$Marital_Statu

        Divorced Married Single
Married 0.022    -       -
Single  0.047    1.00    -
Unknown 1.00     1.00    1.00

P value adjustment method: bonferroni
> pairwise.wilcox.test(data$Credit_Limit, data$Education_Level, p.adj="bonferroni")

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  data$Credit_Limit and data$Education_Level

             College Doctorate Graduate High School Post-Graduate Uneducated
Doctorate    1       -         -        -           -             -
Graduate     1       1         -        -           -             -
High School  1       1         1        -           -             -
Post-Graduate 1      1         1        1           -             -
Uneducated   1       1         1        1           1             -
Unknown      1       1         1        1           1             1

P value adjustment method: bonferroni
```

6. Analiza głównych składowych - PCA

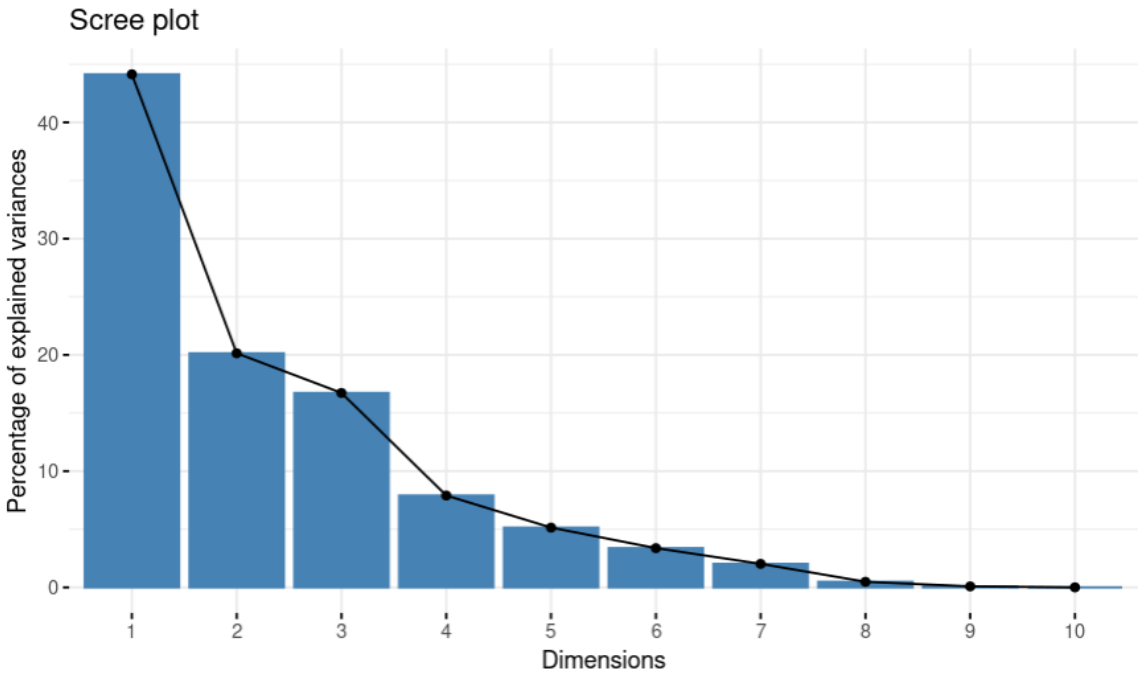
W ramach analizy głównych składowych zostało wybrane 10 zmiennych ilościowych, mianowicie:

- Customer_Age,
- Credit_Limit,
- AVG_Open_To_Ba
- Total_trans_Amt
- Total_Ct_Chng_Q4_Q1
- Months_On_Nook,
- Total_Revolving_Ba
- Total_Amt_Chng_Q4_Q1
- Total_Trans_Ct
- Avg_Utiliztation_Ratio

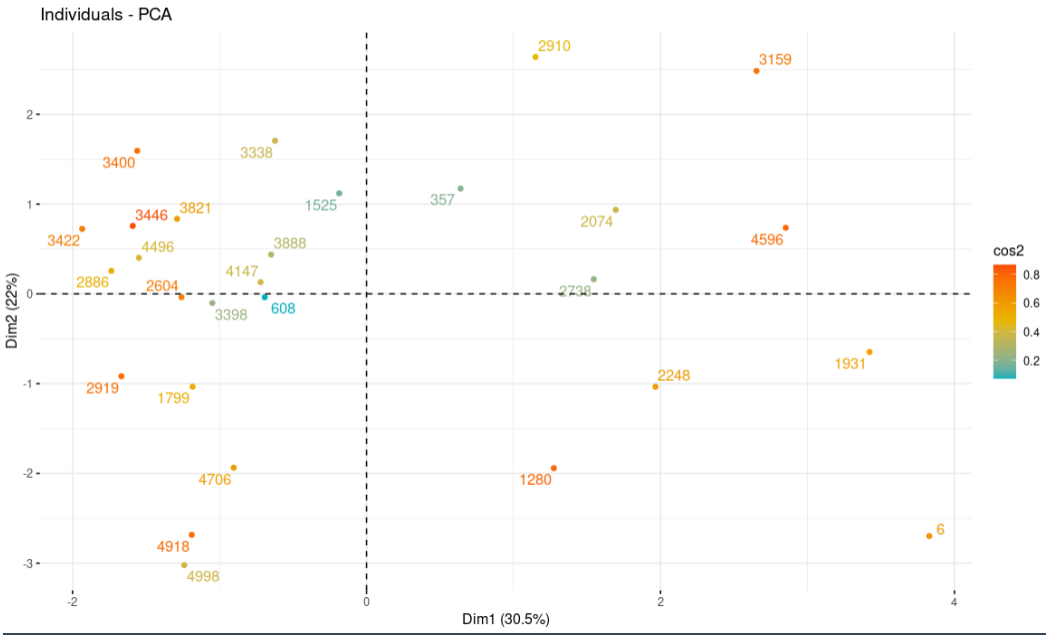
```
> head(data.active)
  Customer_Age Months_on_book Credit_Limit Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt Total_Trans_Ct Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
1           51            36         3418                   0            3418              2.594            1887            20              2.333                 0.000
2           40            34         3313                2517             796              1.405            1171            20              2.333                 0.760
3           40            21         4716                   0            4716              2.175             816            28              2.500                 0.000
4           44            36         4010                1247            2763              1.376            1088            24              0.846                 0.311
5           57            48         2436                 680            1756              1.190            1570            29              0.611                 0.279
6           44            37         4234                 972            3262              1.707            1348            27              1.700                 0.230
```

Zmienną objaśnianą, dla której będzie wykonywana analiza głównych składowych jest poziom edukacji (Education_Level).

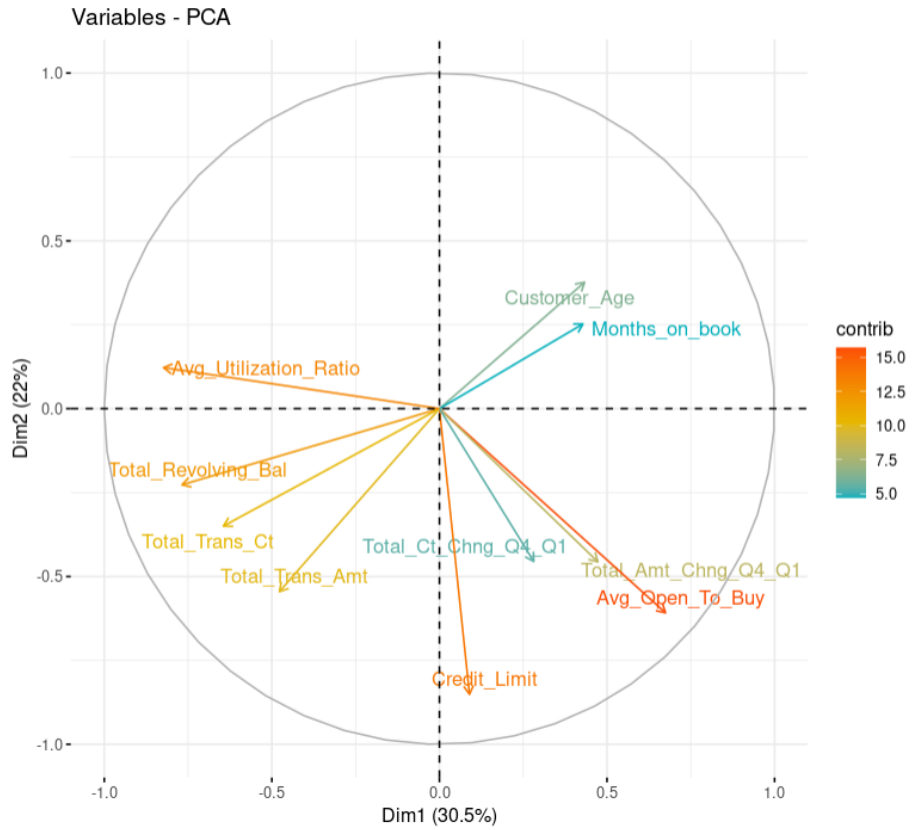
Do stworzenia modelu PCA została wykorzystana funkcji prcomp. Poniżej został zobrazowany procentowy udział zmiennych niezależnych w stopniu w jakim objaśniają zmienną zależną. Jak widać dwie główne składowe objaśniają w ok. 65% wariancję zbioru.



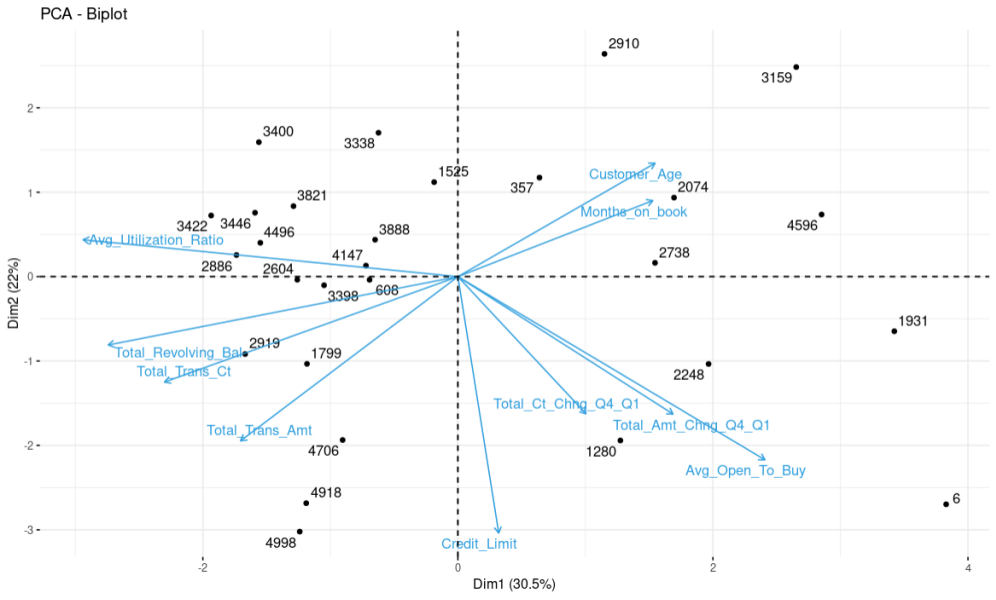
Kolejne ilustracje przedstawiają wykresy opisujące osobników (obserwacje), zmienne oraz wykres biplot łączący dwa poprzednie.



Z wykresu możemy odczytać znaczenia dwóch głównych składników dla danej obserwacji, tak że im dalej od początku układu współrzędnych znajduję się obserwacja tym mocniej wplywają na nią główne składowe.



Powyższy wykres zmiennych, prezentuje takie zależności jak skorelowania zmiennych. Dodatnio skorelowane wskazują tę samą stronę wykresu np Total_Trans_Ct i Total_Trans_Amt, natomiast ujmnie skorelowaną dla zmiennej Total_Trans_Ct jest zmienna Customer_age.
Podobnie jak na poprzednim wykresie im zmienna znajduje się dalej od początku układu współrzędnych tym mocniej wpływają na nią dwie główne składowe.



Ostatni wykres jest połączeniem dwóch poprzednio opisywanych.

7. Wnioski

Podsumowując:

- W ramach analizy wariancji jednoczynnikowej dla objaśnienia limitu na karcie kredytowej nie zostały spełnione założenia odnośnie normalności rozkładu, przez co przeprowadzona analiza Anova nie miała istotnych rezultatów.
  Wynik nieparametrycznego odpowiednika Anova (test Kruskala-wallisa) wykazała, że są istotne statystyczne różnice między grupami dla czynnika stanu cywilnego, a więc rodzaj stanu cywilnego ma wpływa na limit na karcie kredytowej. Przeciwnie sytuacja wyglądała dla czynnika poziomu edukacji nie było aby podstaw do odrzucenia hipotezy zerowej.
- W ramach analizy wariancji dwuczynnikowej dla tych samych zmiennych niezależnych i zmiennej zależnej nie zostały spełnione założenia odnośnie normalności rozkładu, przez co przeprowadzona analiza Anova nie miała istotnych rezultatów.
  Wynik nieparametrycznego odpowiednika Anova (test Friedmana) wykazał, że możemy odrzucić hipotezę zerową i stwierdzić, iż jest istotna statystycznie różnica limitu na karcie kredytowej pomiędzy grupami dla czynników poziomu edukacji i stanu cywilnego.
- W ramach analizy głównych składowych dla zmiennej objaśnianej poziomu edukacji i 10 zmiennych zależnych typu ilościowego wykazano, że dwie główne składowe objaśniają w ok 65% wariancję zbioru.