Проект по Статистика

Зорница Николаева Димитрова, фак.№ 71843

Тема:

„Pokemons Information”

Таблицата с данните е взета от:

https://www2.cs.arizona.edu/classes/cs120/fall17/ASSIGNMENTS/assg02/EXAMPLES-pokemon.html

използван файл:

„PokeInfo-250.csv.“

Зареждаме данните:

```
> rm(list = ls())
> library(fpp2)
> library(readxl)
> Pokemon_Info<-read.csv("E:/Downloads/PokeInfo-250.csv")
> View(Pokemon_Info)
```

Част от таблицата, с която работи проекта:

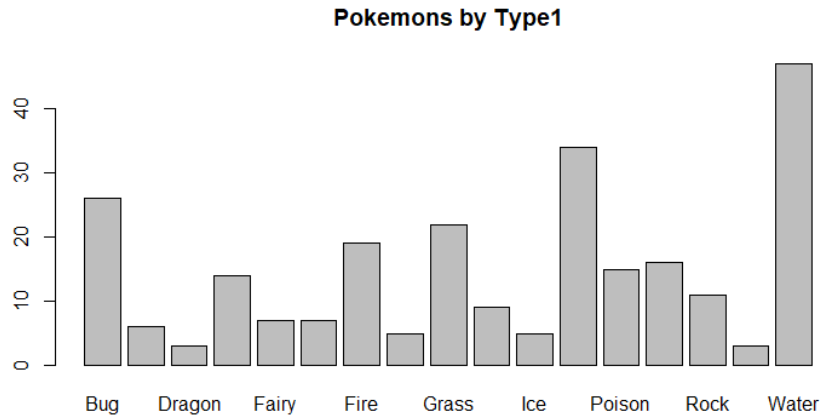
	X.	Name	Type1	Type2	Total	HP	Attack	Defense	SpAtk	SpDef	Speed	Generation	Legendary
1	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
2	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
3	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
4	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
5	4	Charmander	Fire		309	39	52	43	60	50	65	1	False
6	5	Charmeleon	Fire		405	58	64	58	80	65	80	1	False
7	6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	False
8	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	False
9	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	False
10	7	Squirtle	Water		314	44	48	65	50	64	43	1	False
11	8	Wartortle	Water		405	59	63	80	65	80	58	1	False
12	9	Blastoise	Water		530	79	83	100	85	105	78	1	False
13	9	BlastoiseMega Blastoise	Water		630	79	103	120	135	115	78	1	False
14	10	Caterpie	Bug		195	45	30	35	20	20	45	1	False
15	11	Metapod	Bug		205	50	20	55	25	25	30	1	False
16	12	Butterfree	Bug	Flying	395	60	45	50	90	80	70	1	False
17	13	Weedle	Bug	Poison	195	40	35	30	20	20	50	1	False
18	14	Kakuna	Bug	Poison	205	45	25	50	25	25	35	1	False
19	15	Beedrill	Bug	Poison	395	65	90	40	45	80	75	1	False
20	15	BeedrillMega Beedrill	Bug	Poison	495	65	150	40	15	80	145	1	False

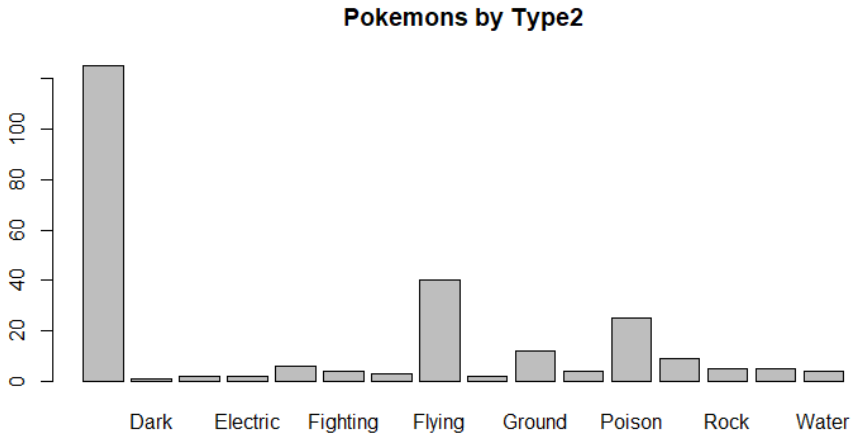
Обща информация върху количествените данни(Total, HP, Attack, Defense, SpeedAttack, SpeedDefense, Speed):

```
> summary(Total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  180.0   323.0   420.0   418.3   500.0   780.0
> summary(HP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   50.00   65.00   66.26   80.00  250.00
> summary(Attack)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   50.00   70.00   74.13   90.00  190.00
> summary(Defense)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   50.00   65.00   71.37   85.00  230.00
> summary(SpeedAttack)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.0    45.0    65.0    69.1    85.0   194.0
> summary(SpeedDefense)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   50.00   70.00   69.64   85.00  230.00
> summary(Speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   45.00   65.00   67.79   90.00  150.00
```

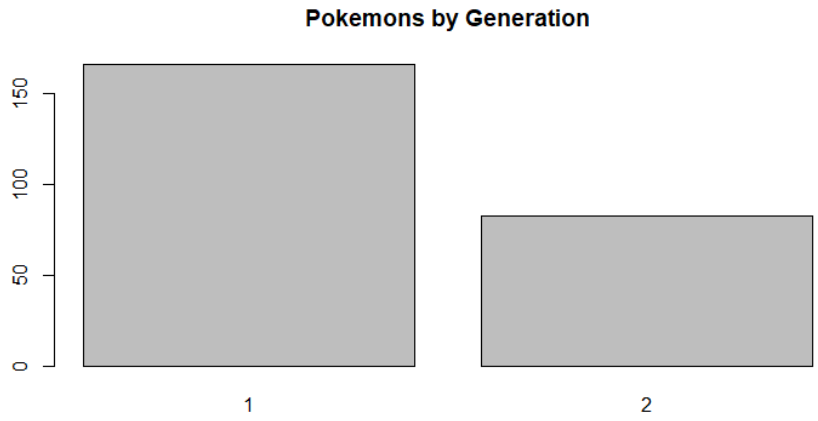
Графични представяния според един категориен признак(Type1, Type2, Generation, Legendary):

```
> barplot(table(Type1), main = "Pokemons by Type1")
> barplot(table(Type2), main = "Pokemons by Type2")
> barplot(table(Generation), main = "Pokemons by Generation")
> barplot(table(Legendary), main = "Pokemons by Legendary")
```

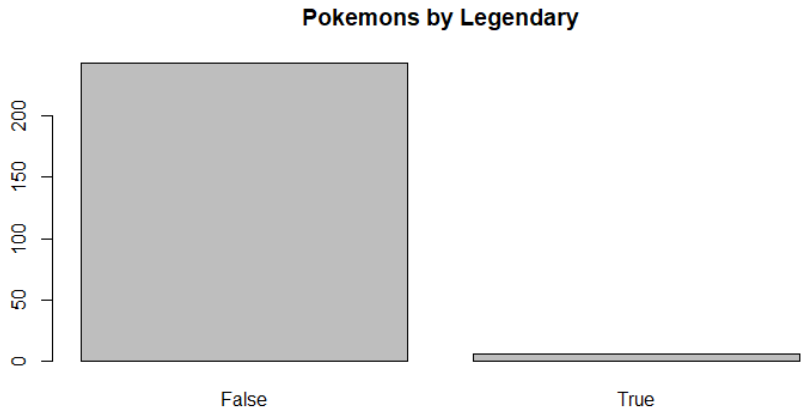




Забелязваме, че повечето покемони нямат Type2.



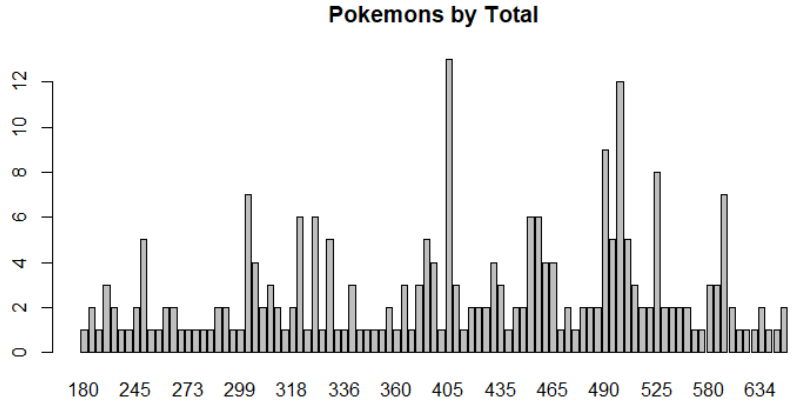
Забелязваме, че повечето покемони са от Generation 1.

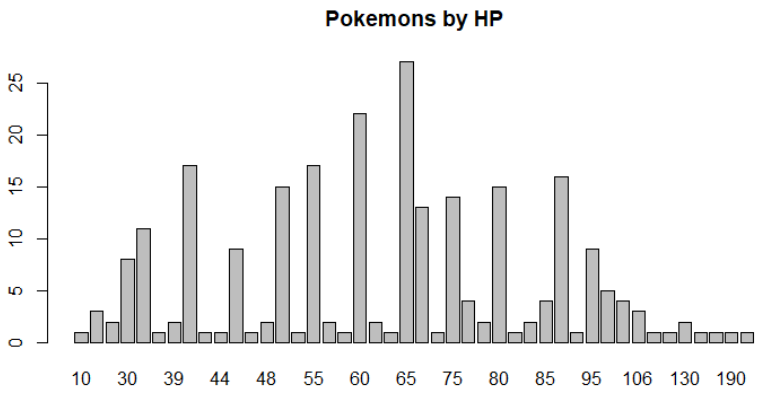


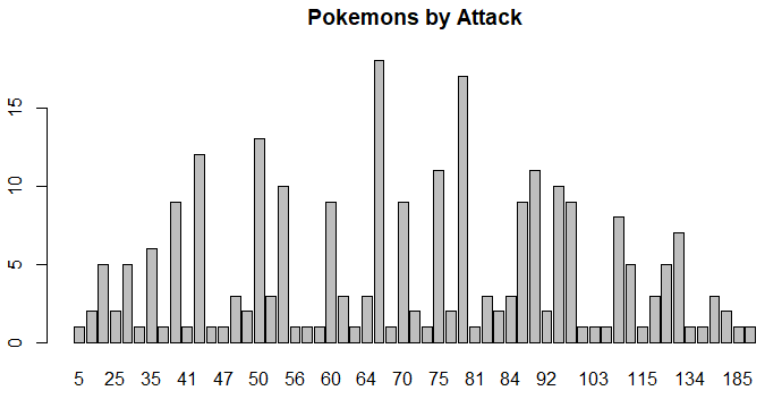
Забелязваме, че много малка част от покемоните са Legendary.

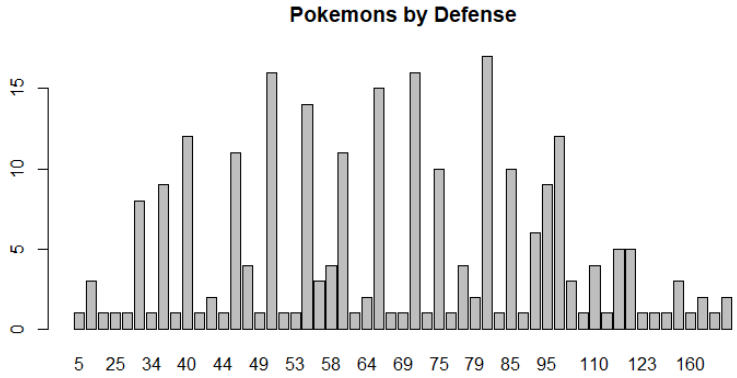
Графични представяния според един количествен признак(Total, HP, Attack, Defense, SpAtk, SpDef, Speed):

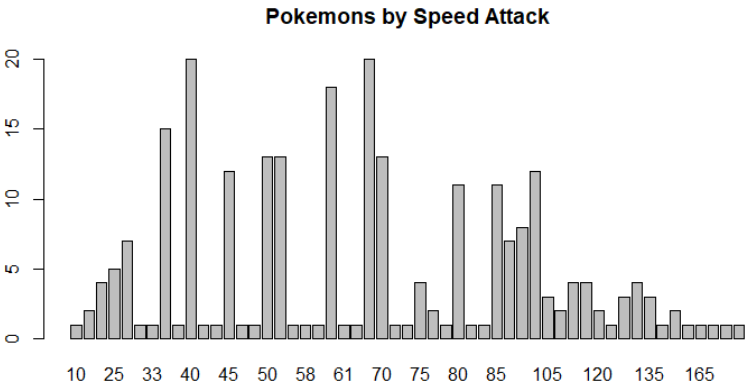
```
> barplot(table(Total), main = "Pokemons by Total")
> barplot(table(HP), main = "Pokemons by HP")
> barplot(table(Attack), main = "Pokemons by Attack")
> barplot(table(Defense), main = "Pokemons by Defense")
> barplot(table(SpeedAttack), main = "Pokemons by Speed Attack")
> barplot(table(SpeedDefense), main = "Pokemons by Speed Defense")
> barplot(table(Speed), main = "Pokemons by Speed")
```

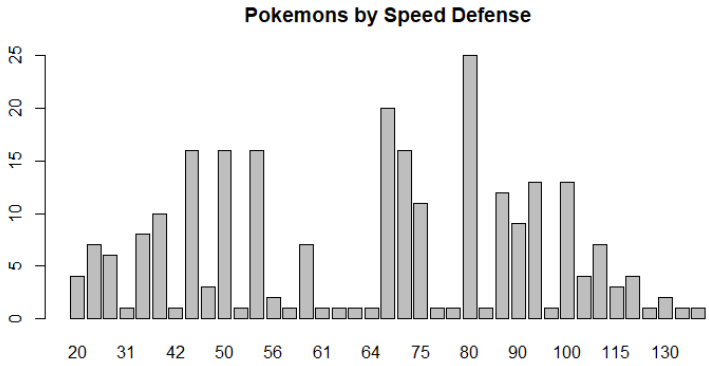


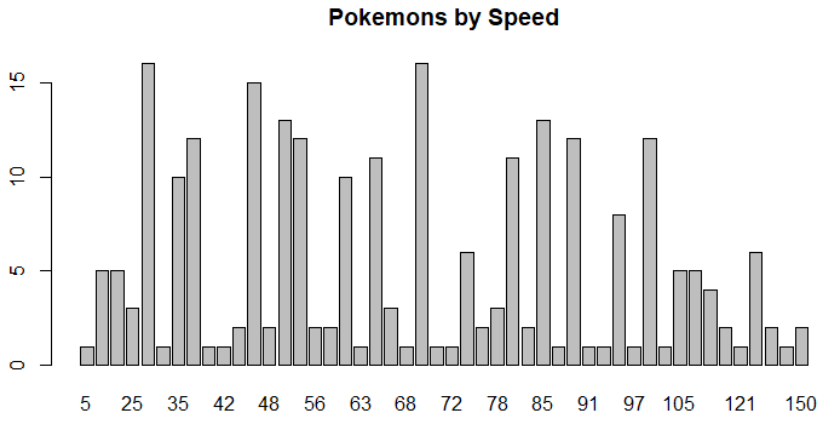



Pokemons by Attack
0
5
10
15
5 25 35 41 47 50 56 60 64 70 75 81 84 92 103 115 134 185


Pokemons by Defense
0
5
10
15
5 25 34 40 44 49 53 58 64 69 75 79 85 95 110 123 160

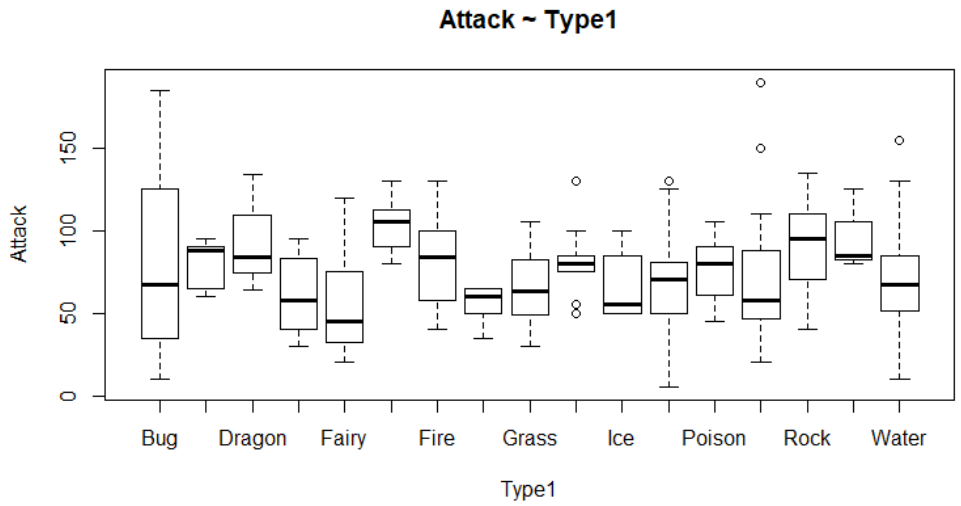

Pokemons by Speed Attack
0
5
10
15
20
10 25 33 40 45 50 58 61 70 75 80 85 105 120 135 165

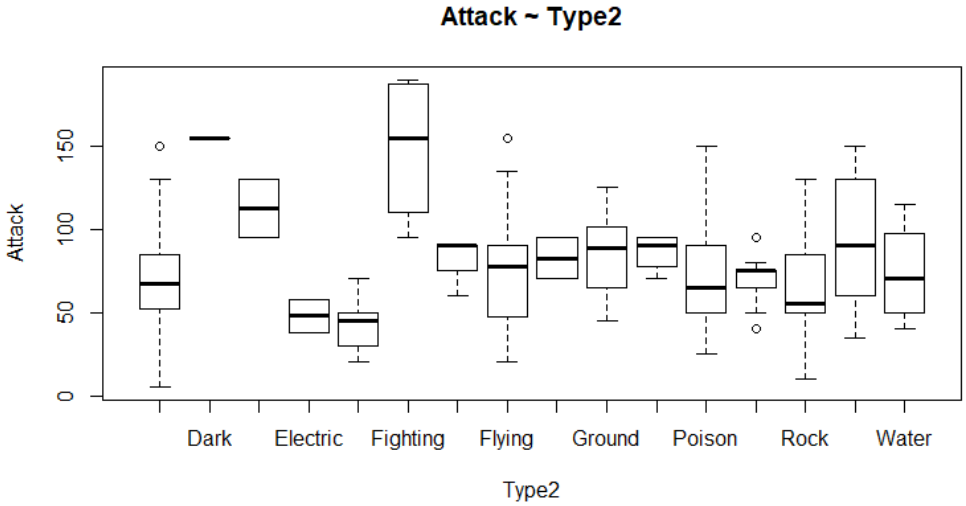

Pokemons by Speed Defense
0
5
10
15
20
25
20 31 42 50 56 61 64 75 80 90 100 115 130



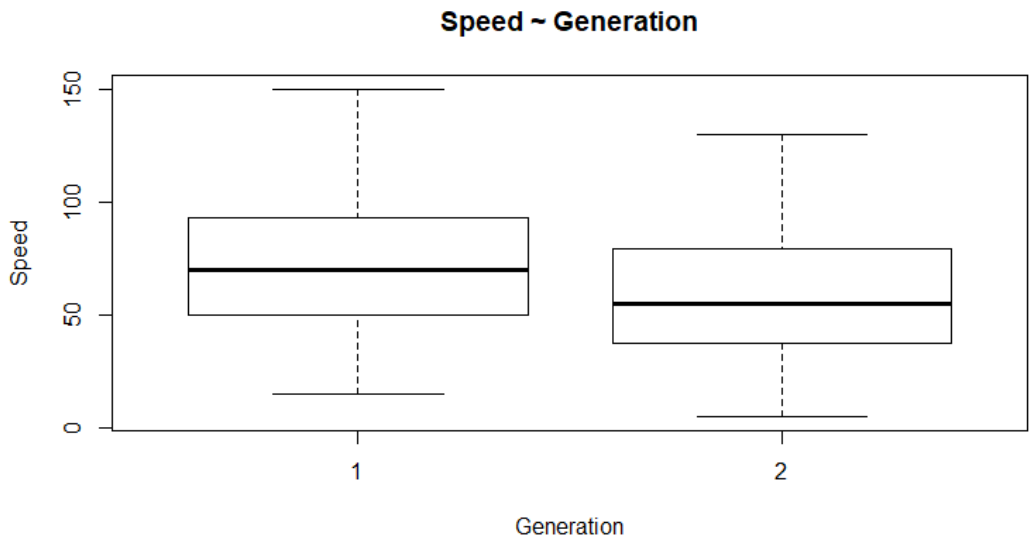
Графично представяне според един количествен и един качествен признак (Attack~Type1, Attack~Type2, Speed~Generation, Speed~Legendary):

```
> boxplot(Attack ~ Type1, main = "Attack ~ Type1")
> boxplot(Attack ~ Type2, main = "Attack ~ Type2")
> boxplot(Speed ~ Generation, main = "Speed ~ Generation")
> boxplot(Speed ~ Legendary, main = "Speed ~ Legendary")
```

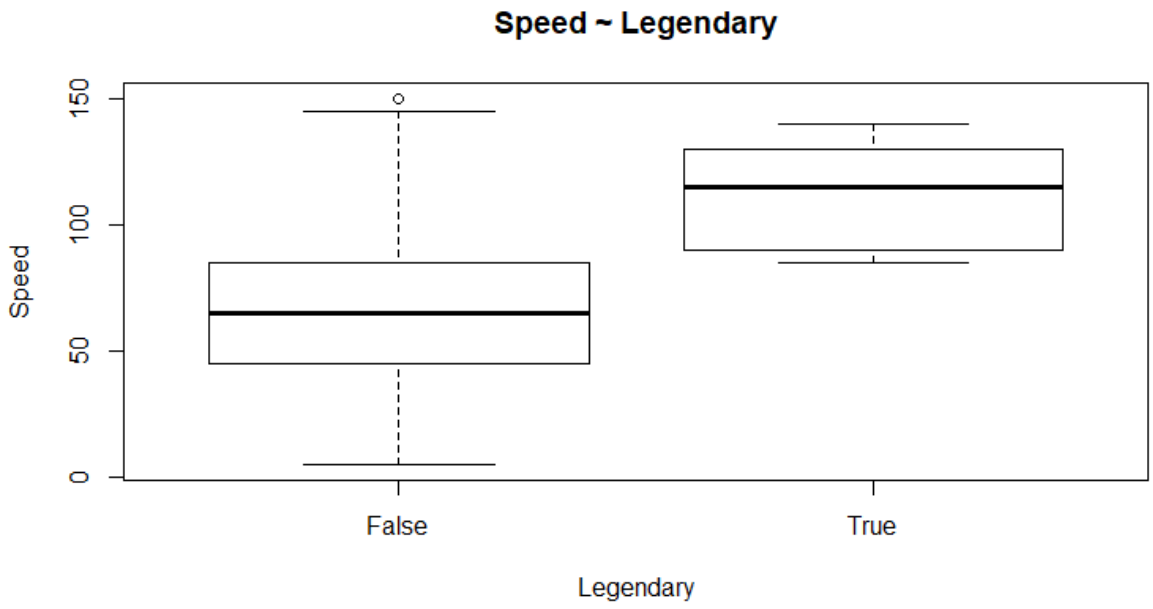




Забелязваме, че Attack при Type1 и Type2 не е средно разпределена. При някои покемони средната атака е по-ниска, при други-по-висока.



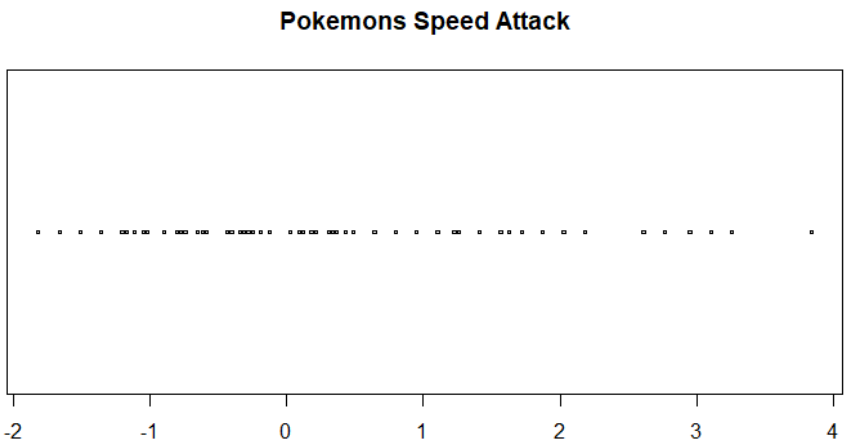
Speed при Generation1 и Generation2 са сравнително средно разпределени, покемоните от Generation1 са малко по-бързи.

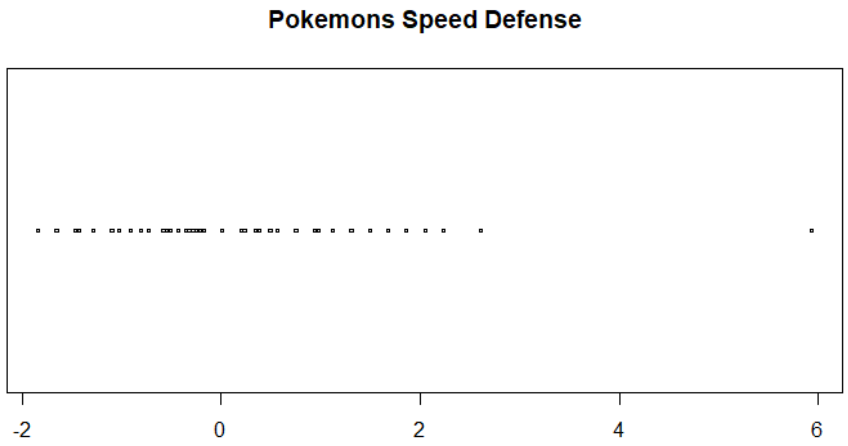


Забелязваме, че бързината не е средно разпределена, покемоните които са Legendary са по-бързи.

Лента на SpeedAttack и SpeedDefense:

```
stripchart(scale(SpeedAttack), cex = 0.3, main = "Pokemons Speed Attack")
stripchart(scale(SpeedDefense), cex = 0.3, main = "Pokemons Speed Defense")
```

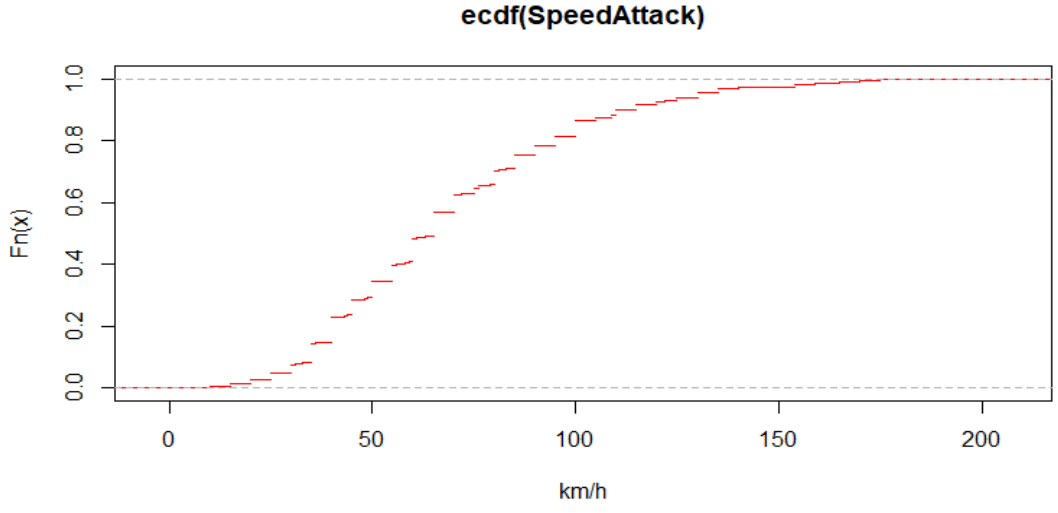


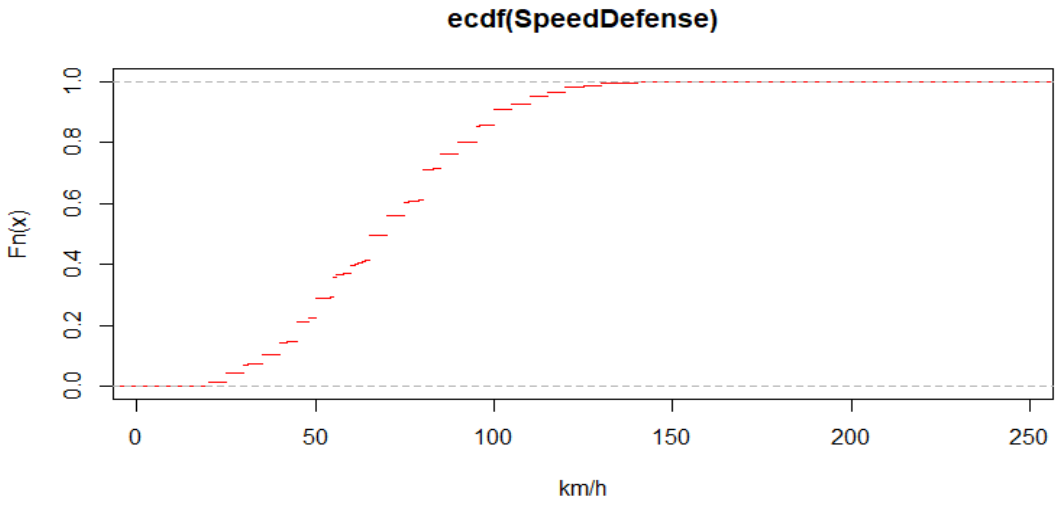


Забелязваме, че и при двете ленти имаме струпване в лявата част и единично наблюдение в дясно. Тези ленти ни показват, че SpeedAttack и SpeedDefense са в корелация.

Емпирични функции на разпределение на SpeedAttack и SpeedDefense:

```
par(mfrow = c(2,1))
plot(ecdf(SpeedAttack), verticals = FALSE, col = "RED", do.points = FALSE, lwd = 1, xlab = "km/h")
plot(ecdf(SpeedDefense), verticals = FALSE, col = "RED", do.points = FALSE, lwd = 1, xlab = "km/h")
```

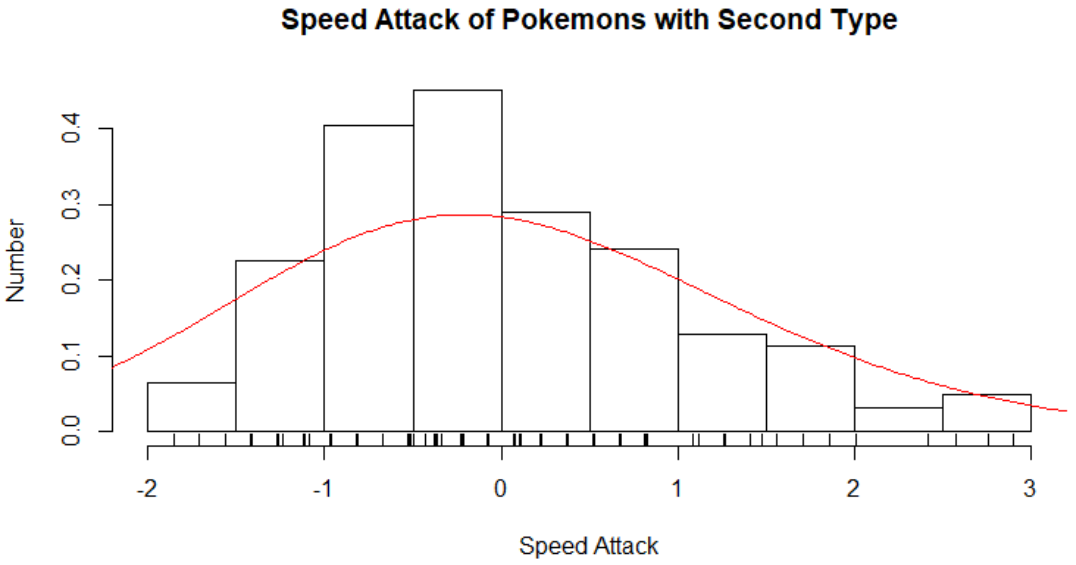


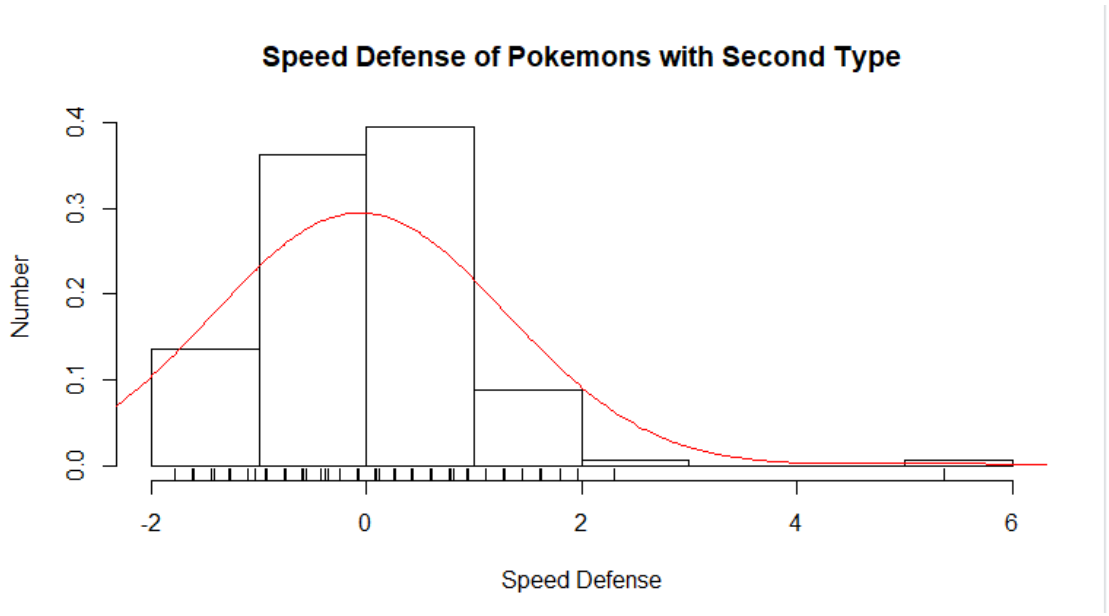


Хистограми на SpeedAttack и SpeedDefense на покемони, които имат Type2:

```
> type2pokemons = Pokemon_Info[(Type2 != ''),]
> hist(scale(type2pokemons$SpAt),
+      probability = TRUE,
+      main = "Speed Attack of Pokemons with Second Type",
+      xlab = "Speed Attack",
+      ylab = "Number")
> rug(jitter(scale(type2pokemons$SpAtk)))
> lines(density(scale(type2pokemons$SpAtk), bw = 1), col = "red")
```

```
> hist(scale(type2pokemons$SpDef),
+      probability = TRUE,
+      main = "Speed Defense of Pokemons with Second Type",
+      xlab = "Speed Defense",
+      ylab = "Number")
> rug(jitter(scale(type2pokemons$SpDef)))
> lines(density(scale(type2pokemons$SpDef), bw = 1), col = "red")
```



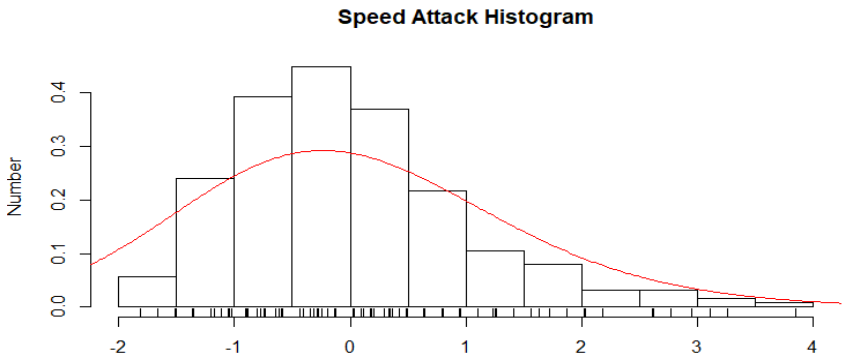


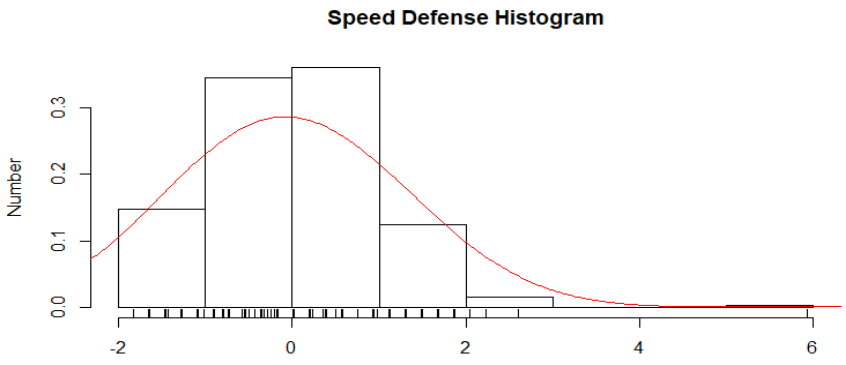
Къде точно се намират наблюденията отбелязваме с функцията rug, след това чертаем червена линия върху съществуващата хистограма.

Върхът при „Speed Defense of Pokemons with Second Type“ се намира наляво от средното. Налична е дясна положителна асиметрия, тъй като имаме по-голямо струпване на данни около по-малките значения на признака (SpeedDefense).

Хистограми на SpeedAttack и SpeedDefense:

```
hist(scale(SpeedAttack), probability = TRUE, right = FALSE, main = "Speed Attack Histogram", xlab = "", ylab = "Number")
rug(jitter(scale(SpeedAttack)))
lines(density(scale(SpeedAttack), bw = 1), col = "red")
hist(scale(SpeedDefense), probability = TRUE, right = FALSE, main = "Speed Defense Histogram", xlab = "", ylab = "Number")
rug(jitter(scale(SpeedDefense)))
lines(density(scale(SpeedDefense), bw = 1), col = "red")
```





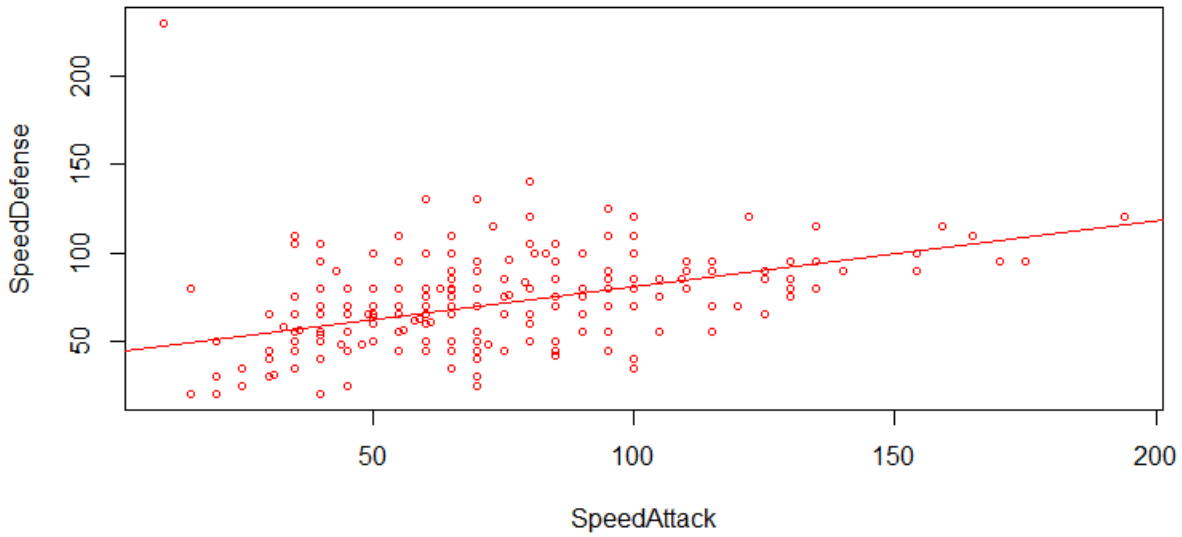
Сравняваме хистограмите за SpeedAttack и SpeedDefense. Графиките изглеждат доста подобно. Повече от 60% от наблюденията и за двата признака се намират вляво от средното. Останалите наблюдения се намират вдясно и са доста по-разпръснати. Линиите на плътност също изглеждат сравнително подобни. От тези наблюдения можем да твърдим, че има положителна корелация между признаците SpeedAttack и SpeedDefense.

```
> cor(SpeedAttack, SpeedDefense)
[1] 0.4477387
```

Стойността е 0.45, което попада в скалите на умерена и значителна корелация.

Построяваме корелационно поле и регресионна права през полученото поле

```
plot(SpeedAttack, SpeedDefense, col = "red", cex = 0.70)
abline(lm(SpeedDefense~SpeedAttack), col = "red")
```



Регресионната права минава близо до повечето точки, така че можем да кажем, че SpeedAttack и SpeedDefense са сравнително правопропорционални в нарастването си.

```
> result = lm(SpeedDefense~SpeedAttack)
> result$coefficients[1]
(Intercept)
   43.91148
> result$coefficients[2]
SpeedAttack
  0.3722928
```

Покемон със SpeedAttack 80km/h, на базата на линейния модел се очаква да е със SpeedDefense 73, 694904km/h:

80*0,3722928 + 43.91148 = 73, 694904

Корелации между различни двойки количествени признаци:

0 < R < 0,3 – слаба корелация

0,3 < R < 0,5 – умерена корелация

0,5 < R < 0,7 – значителна корелация

0,7 < R < 0,9 – висока корелация

0,9 < R < 1 – много висока корелация

Другата скала е:

0 – 0,2 – слаба корелация

0,2 – 0,4 – умерена корелация

0,4 – 0,6 – значителна корелация

0,6 – 0,8 – висока корелация

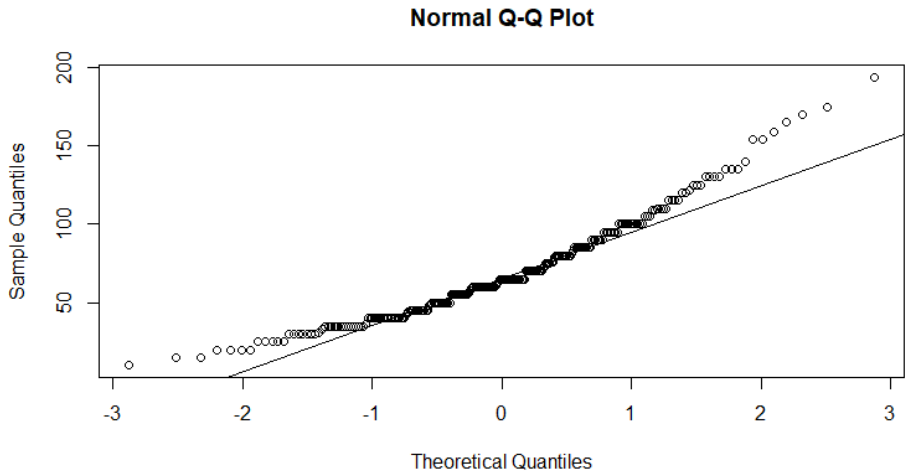
0,8 – 1 – много висока корелация

Скалите са взети от:
https://bg.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%B5%D0%BB%D0%B0%D1%86%D0%B8%D1%8F

```
> cor(Attack, SpeedAttack)
[1] 0.2346357
> cor(Defense, SpeedDefense)
[1] 0.3942657
> cor(Defense, Attack)
[1] 0.4398876
> cor(HP, Total)
[1] 0.5522511
```

Според изследваните данни първата двойка се намира в слаба корелация, втората и третата в умерени, а последната в значителна корелация.

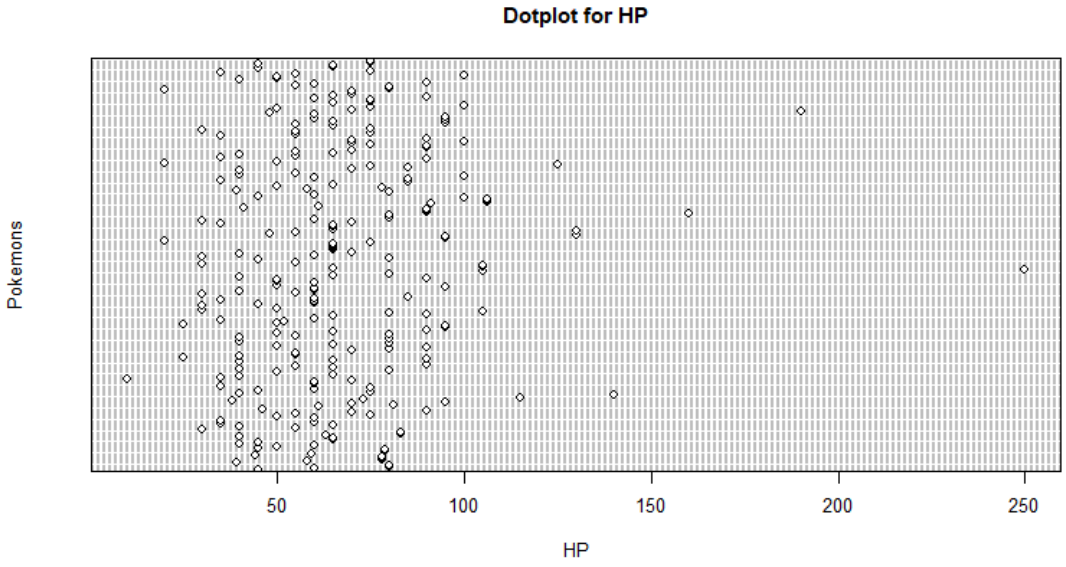
```
> mean(SpeedAttack)
[1] 69.10442
> qqnorm(SpeedAttack)
> qqline(SpeedAttack)
> shapiro.test(SpeedAttack)

        Shapiro-Wilk normality test

data:  SpeedAttack
W = 0.94316, p-value = 3.002e-08
```



Построяваме права и забелязваме, че повечето наблюдения се намират по нея или близо до нея и можем да предположим, че работим с нормално разпределени данни.

Shapiro–Wilk test-а отхвърля нулевата хипотеза за равномерно разпределение на входните данни. (p-value < 0.05)

Dotplot за HP:



Заключение:

-Много малка част от покемоните са Legendary, повечето нямат Type2 и са от Generation 1.

-Attack при Type1 и Type2 не е средно разпределена. При някои покемони средната атака е по-ниска, при други-по-висока.

-Speed при Generation1 и Generation2 са сравнително средно разпределени, покемоните от Generation1 са малко по-бързи. По условие на Legendary, бързината не е средно разпределена, покемоните, които са Legendary са по-бързи.

-От лентите на SpeedAttack и SpeedDefense забелязваме, че и при двете имаме струпване в лявата част и единично наблюдение в дясно. Тези ленти ни показват, че SpeedAttack и SpeedDefense са в корелация.

-Върхът при хистограмата „Speed Defense of Pokemons with Second Type“ се намира наляво от средното. Налична е дясна положителна асиметрия, тъй като имаме по-голямо струпване на данни около по-малките значения на признака (SpeedDefense).

-От хистограмите за SpeedAttack и SpeedDefense забелязваме, че графиките изглеждат доста подобно. Повече от 60% от наблюденията и за двата признака се намират вляво от средното. Останалите наблюдения се намират вдясно и са доста по-разпръснати. Линиите на плътност също изглеждат сравнително подобни. От тези наблюдения можем да твърдим, че има положителна корелация между признаците SpeedAttack и SpeedDefense. Изчисляваме стойността ѝ и виждаме, че тя е 0.45, което попада в скалите на умерена и значителна корелация.

-От корелационното поле и регресионна права през полученото поле забелязваме, че регресионната права минава близо до повечето точки, така че можем да кажем, че SpeedAttack и SpeedDefense са сравнително правопропорционални в нарастването си.

-Изчисляваме, че на базата на линейния модел покемон със SpeedAttack 80km/h, се очаква да е със SpeedDefense 73, 694904km/h.

-От Normal Q-Q Plot-а забелязваме, че повечето наблюдения се намират по правата или близо до нея, от което предполагаме, че работим с нормално разпределени данни, а Shapiro–Wilk test-а отхвърля нулевата хипотеза за равномерно разпределение на входните данни(SpeedAttack).(p-value < 0.05)