# Визуализация

# Деректерді өндіру және білімді ашу

# Кіріспе

Деректерді өндіру бұл деректерден пайдалы ақпаратты алу немесе ашу үшін қолайлы машиналық оқытуды немесе статистикалық әдістерді әзірлеу, анықтау, пайдалану туралы ғылым.

Негізгі мақсат деректерден маңызды ақпаратты табу.

Деректерді өндіру көп салалы, әр түрлі салалардан үлес алады

- информатика: жасанды интеллект, машиналық оқыту, дерек қоры
- статистика
- инженерлік, географиялық ақпараттық жүйелер және сауда

# Кіріспе

Деректерді өндіруде барлығына сәйкес келетін тәсіл жоқ.

Алдымен деректердің қасиетін түсіну және тапсырманы түсіну маңызды.

Содан кейін тапсырма үшін сәйкес деректерді іздеу әдісін анықтау маңызды.

Біз нені үйренеміз

- деректерді визуализациялау (осы лекция)
  - деректерді өндірудегі бірінші қадам деректердің қасиетін түсіну екенін есте сақта
  - бұны деректерді визуализациялау арқылы жүзеге асыруға болады
- классификация ағашы
- классификация үшін векторлық машинаны қолдау
- naive Bayes классификаторы
- регрессия әдістері
- өлшемді азайту әдістері: PCA және көпөлшемді масштабтау

# Кіріспе

Әртүрлі терминология әртүрлі пәндерде (информатика, статистика, машиналық оқыту, коммерция және инженерия) бірдей ұғымдар үшін қолданылады, мысалы

- деректерді өндіру ↔ аналитика, машиналық оқыту, үлкен деректерді талдау
- енгізу ↔ түсіндірмелі, болжаушы немесе тәуелсіз айнымалы, атрибут, белгілер
- шығару ↔ жауап, сынып немесе тәуелді айнымалы

# Жоспары

- Шолу
  - **Визуализация дегеніміз не?**
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- Техникалар

# Кіріспе

- Сурет мың сөзге, ал график мың санға тең.
- Біз графиктерді деректер туралы хабарларды жеткізу және қорытындылау үшін пайдаланамыз.
- Кестелер нақты сандарды оқу үшін тиімді.
- Мысалы оқырманды бір топ үшін 0,0356, ал екіншісі үшін 0,0750 орташа бағалау мән бермеуі мүмкін. Мәселе мынада, соңғысы екі есе дерлік жоғары. Мұндай ондаған салыстырулар мүмкін, оларды жақсы таңдалған графикалық дисплейде көруге болады.

# Кіріспе

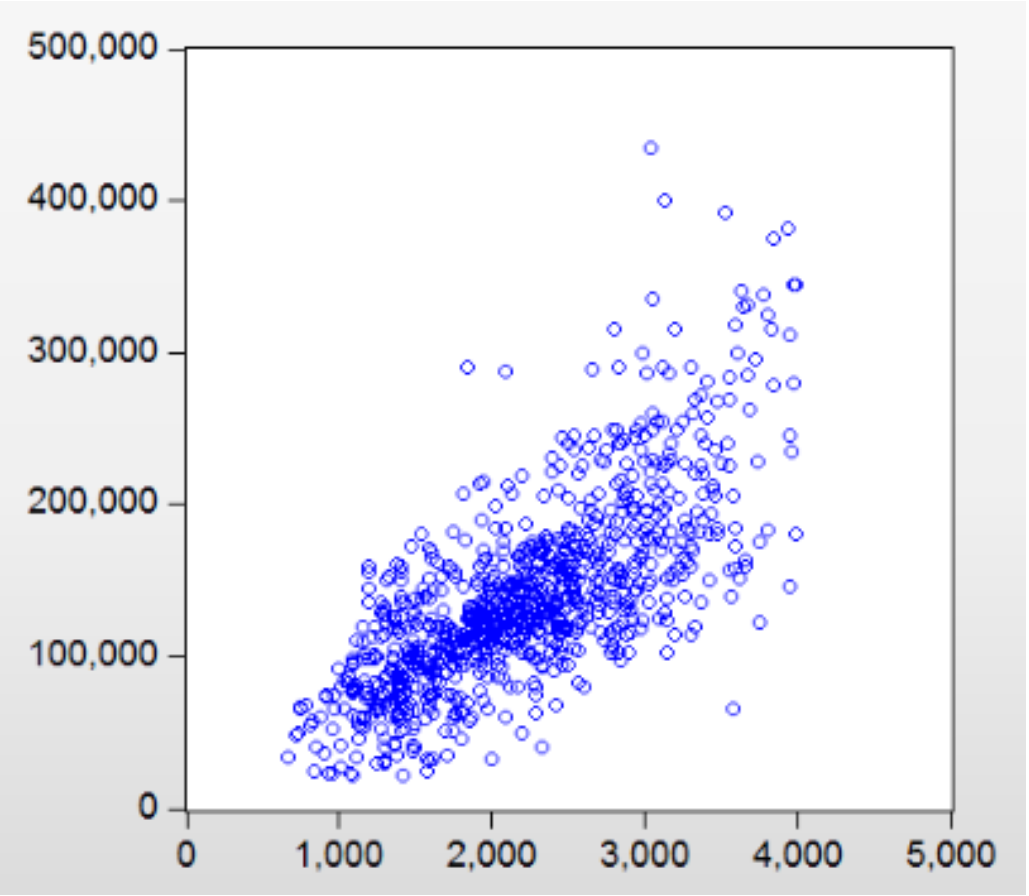
- График жақсы болады егер көрсететін деректер де жақсы болса
- Шығармашылық көп болса да, күмәнді деректерден жақсы график ала алмайсыз.
- Графиктерте қиын емес интерпретациялары бар жақсы бейнелер болуы керек.
- Қорытындылай алмайтын немесе түсіндіруді жеңілдетпейтін графикті кесте немесе сандар тізімін беруден де жаман.

# Визуализация дегеніміз не?

- Деректерді визуализациялау ақпаратты графикалық форматта көрсету.
- Визуализацияны қолданудың негізгі мотивациясы
  - адамдар әдетте көрнекі ақпараттың үлкен көлемін қабылдай алады
  - көрнекі ақпараттағы үлгілерді немесе ауытқуларды таба алады
- Сәтті визуализация деректерді (ақпаратты) визуалды пішімге түрлендіруді талап етеді, осылайша деректер сипаттамалары мен деректер элементтері немесе атрибуттары арасындағы қарым-қатынастар талдануы немесе хабарлануы мүмкін.
- Деректерді визуализациялаудағы негізгі мәселе қызығушылық қатынастарын оңай байқауға болатын әдісті немесе әдісті таңдау болып табылады.

# Қарапайым мысал

- Үй бағасы мен үй көлемі арасындағы оң байланыс.
- Үлкен үйдегі үй бағасының ауытқуы шағын үйге қарағанда әлдеқайда көп.
- x осі - үй көлемі
- y осі - үй бағасы



# Визуализация дегеніміз не?

Деректерді өндіру контекстінде нені визуализациялауға болады?

- Өңделмеген немесе түрлендірілген деректер
  - өңделмеген айнымалының лог-трансформациясы
- Айнымалылар арасындағы байланыстар
  - үй бағасы мен үй көлемі арасындағы қатынас
- Үлгілерден болжамды шығыс
  - 2021-2025 жылдардағы туристер санының болжамы
- Деректер мен үлгілер арасындағы сәйкессіздіктер ("қателер")

# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - **Неліктен визуализация?**
  - Қалай визуализациялау керек?
- Техникалар

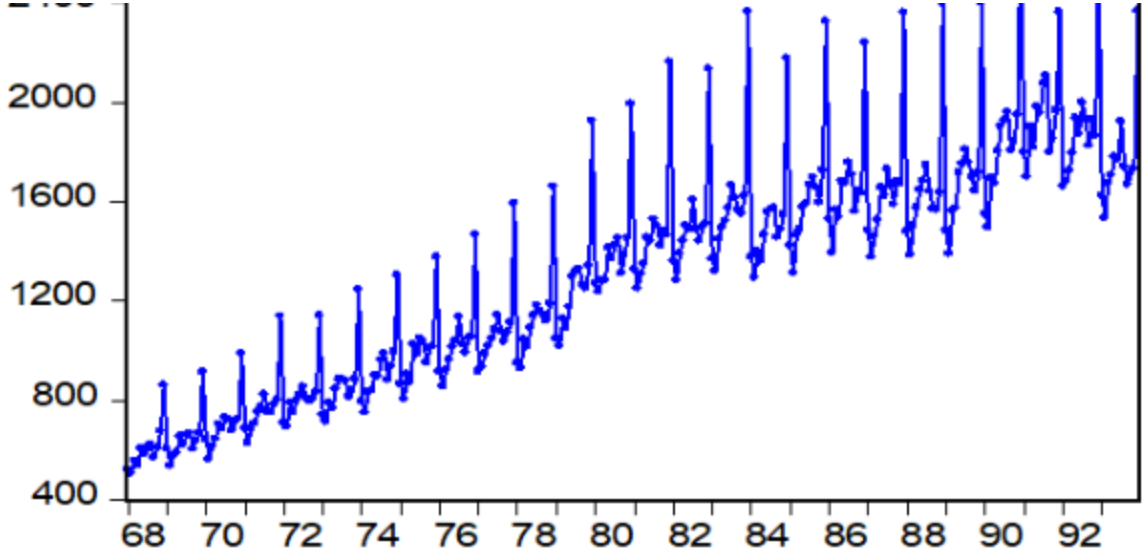
# Неліктен визуализация?

Деректерді өндіруде визуализация маңызды

- тазалықты талап ететін бастапқы деректегі ақауларды анықтайды
  - мысалы әртүрлі өлшем бірліктері, шектен тыс көрсеткіштер
- атрибуттарды таңдауға көмектеседі
- шығу айнымалымен байланысы жоқ кіріс айнымалы мәнін жояды
- модель құруға көмектеседі, өйткені маңызды үлгілерді көрсетеді
- нәтижелерді түсіндіруге көмектеседі
- нәтижелерді басқаларға хабарлайды

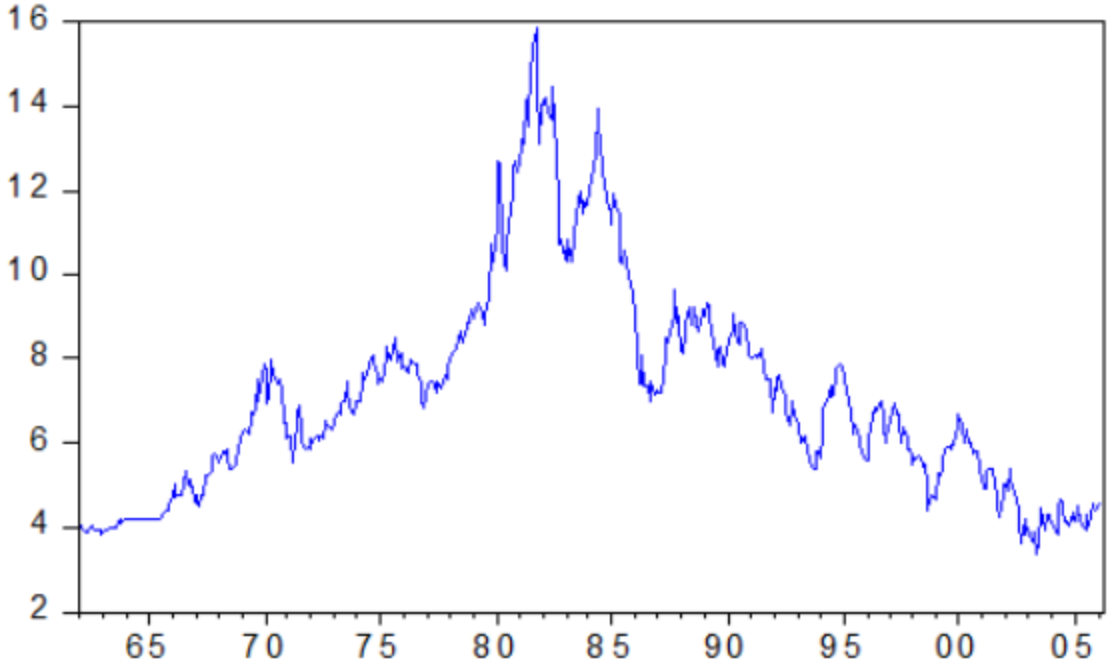
# Мысал, уақыт сериясының сюжеті

- тренд
  - орташа уақыт бойынша өсу
- маусымдық
  - қайталанатын үлгі
  - желтоқсанның ең жоғары
  - ақпанның ең төменгісі



# Визуализация, уақыт сериясының сюжеті

- тұрақты
  - кешегі мән бүгінгі мәнге өте жақын
- кездейсоқ тренд
  - ішкі үлгілер өте әртүрлі



# Математик Джон Тукидің дәйексөздері

- Визуализация жиі зұлымдық үшін қолданылады. елеусіз деректер өзгерістерін бұрмалау және оларды мағыналы ету. Менің досым болғың келсе олай істеме. Нәтижелерді анық және шынайы көрсетіңіз. Егер бірдеңе жұмыс істемесе, шолу нәтижелерін білуі керек.
- Сандық шамалар күтілетін мәндерге, графикалық қорытындылар күтпеген мәндерге бағытталған.

# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - **Қалай визуализациялау керек?**
- Техникалар

# Қалай визуализациялау керек?

Графикалық шеберліктің принциптері

- Деректердің мағынасын бұрмалауға жол берме
- Шағын кеңістікте көп ақпаратты көрсет
- Графиканы ауызша және статистикалық сипаттамалармен біріктір
- Бірнеше деректер жиынының қасиеттерін салыстыр

# Қалай визуализациялау керек?

Саналы араласусыз тез қабылданатын графикалық қасиеттердің үш категориясы

- Түс
- Пішін
  - пішіні
  - өлшемі
- Қозғалыс
  - анимация, айналдыру, масштабтау
  - кейбіреулері, бірақ тым көп емес

# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - Сандық айнымалылар
    - Бір
    - Екі (немесе одан да көп)
    - Көп
  - Категориялық айнымалылар
  - Екі

# Мысал деректер жинағы: Андерсонның ирис деректері

## Иристің үш түрінен 150 гүл бойынша өлшемдер (әрқайсысы 50-ден):

| Сепал.Ұзындығы | Сепал.Ені | Жапырақ.Ұзындығы | Жапырақ.Ені | Түрлер |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | сетоза |
| 5.8 | 4.0 | 1.2 | 0.2 | сетоза |
| 5.2 | 3.5 | 1.5 | 0.2 | сетоза |
| 4.5 | 2.3 | 1.3 | 0.3 | сетоза |
| 6.5 | 2.8 | 4.6 | 1.5 | версикс |
| 6.2 | 2.2 | 4.5 | 1.5 | версикс |
| 5.5 | 2.4 | 3.7 | 1.0 | версикс |
| 5.7 | 3.0 | 4.2 | 1.2 | версикс |
| 6.7 | 2.5 | 5.8 | 1.8 | вирджиника |
| 7.7 | 2.8 | 6.7 | 2.0 | вирджиника |
| 7.7 | 3.0 | 6.1 | 2.3 | вирджиника |
| 5.9 | 3.0 | 5.1 | 1.8 | вирджиника |

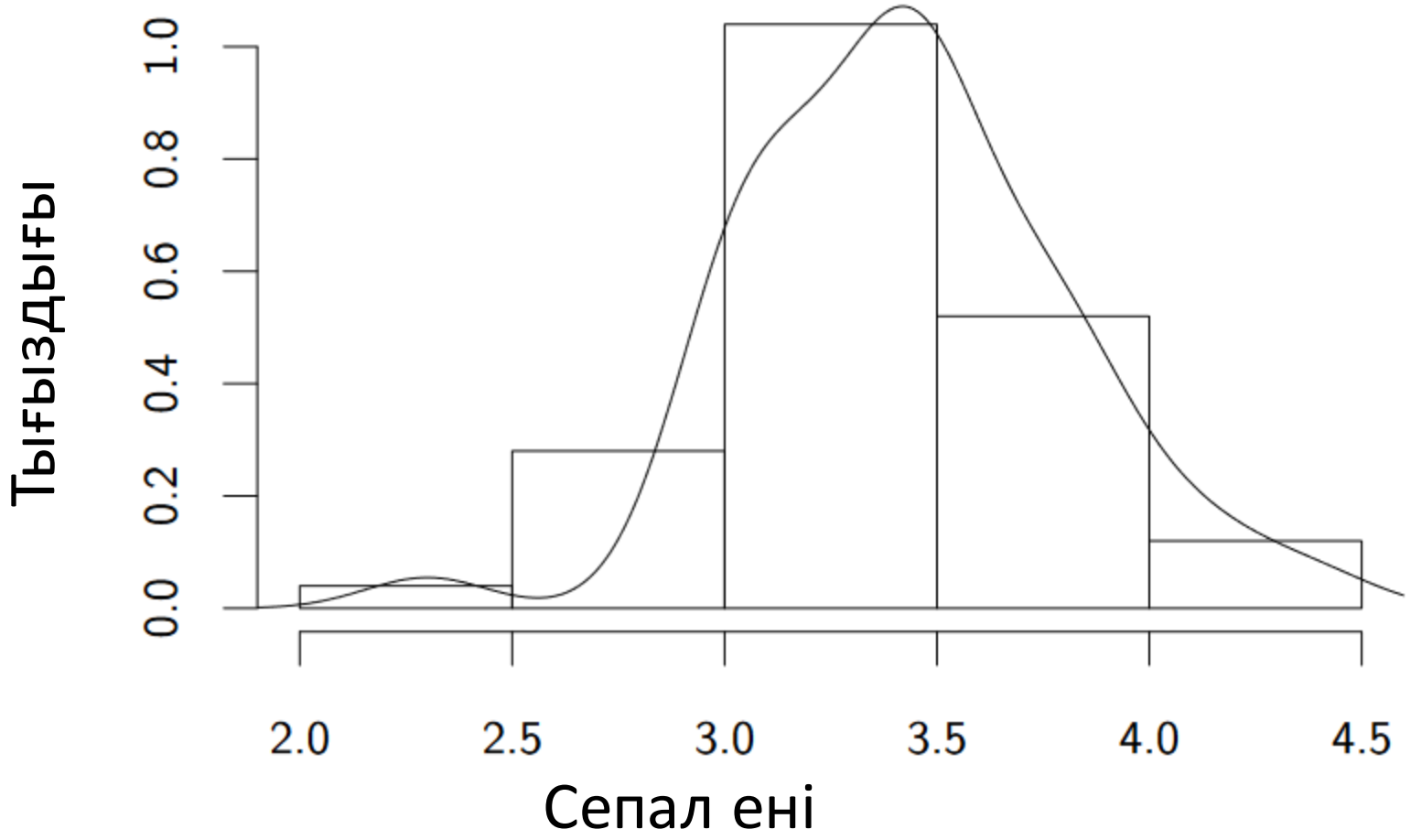
# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - **Сандық айнымалылар**
    - **Бір**
    - Екі (немесе одан да көп)
    - Көп
  - Категориялық айнымалылар
  - Екі

# Гистограмма

- Гистограмма деректердің қалай таратылатынын сипаттайды
- Гистограмма көлденең ось бойымен (әдетте тұрақты ені) әрбір "қалтадағы" бақылаулардың жиілігін (тік ось) көрсетеді.
- Тым көп қоқыс (бұдырлы учаске) мен тым аз (қызықты деталь көрінбейді) арасында ымыраға келу керек.
- Интерпретация: шектен тыс мәндерді (ерекше жоғары немесе төмен деректер мәндерін) анықтауға мүмкіндік береді және таралудың қиғаштығын (симметриясының жоқтығын) көрсетеді.
- R коды: гистограмма немесе пішінді көрсету үшін қисыққа сәйкес келетін тығыздық.

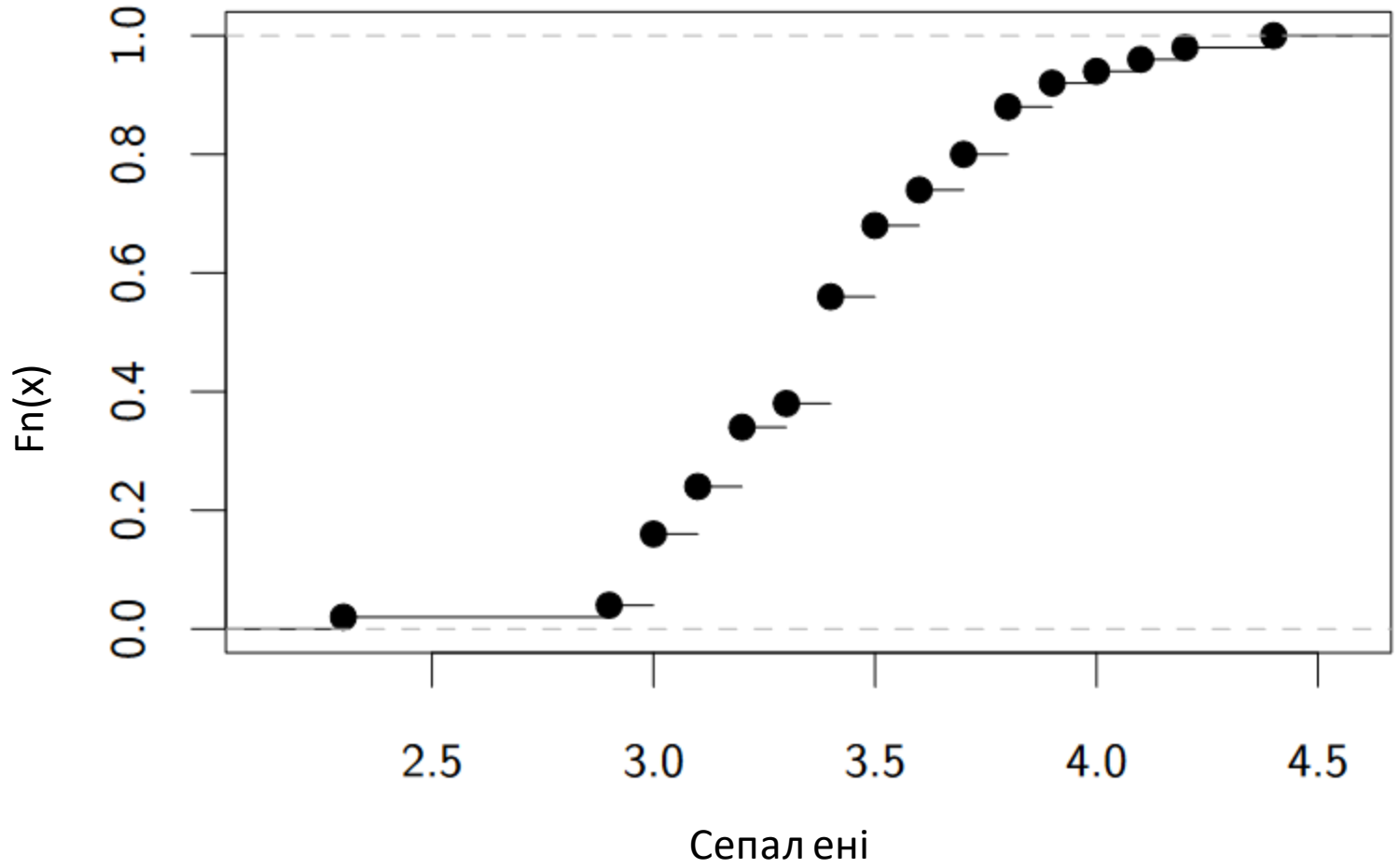
# Гистограмма (қабатталған тығыздық сәйкестігі)



# Эмпирикалық жинақтаушы таралу функциясы

- Эмпирикалық жинақтаушы үлестіру функциясының сызбасы деректер ауқымындағы x (көлденең ось) үшін деректер мәндерінің ≤ x үлесін көрсетеді.
- Бұл k байланыстырылған деректер мәндері бар жерлерде k/n-ге секіретін қадамдық функция пішімін алады.
- Сызба төмендемейді, деректер шоғырланған аймақтарда тік, әдетте S пішінді.
- R коды: ecdf(x)

# Эмпирикалық жинақтаушы таралу функциясы



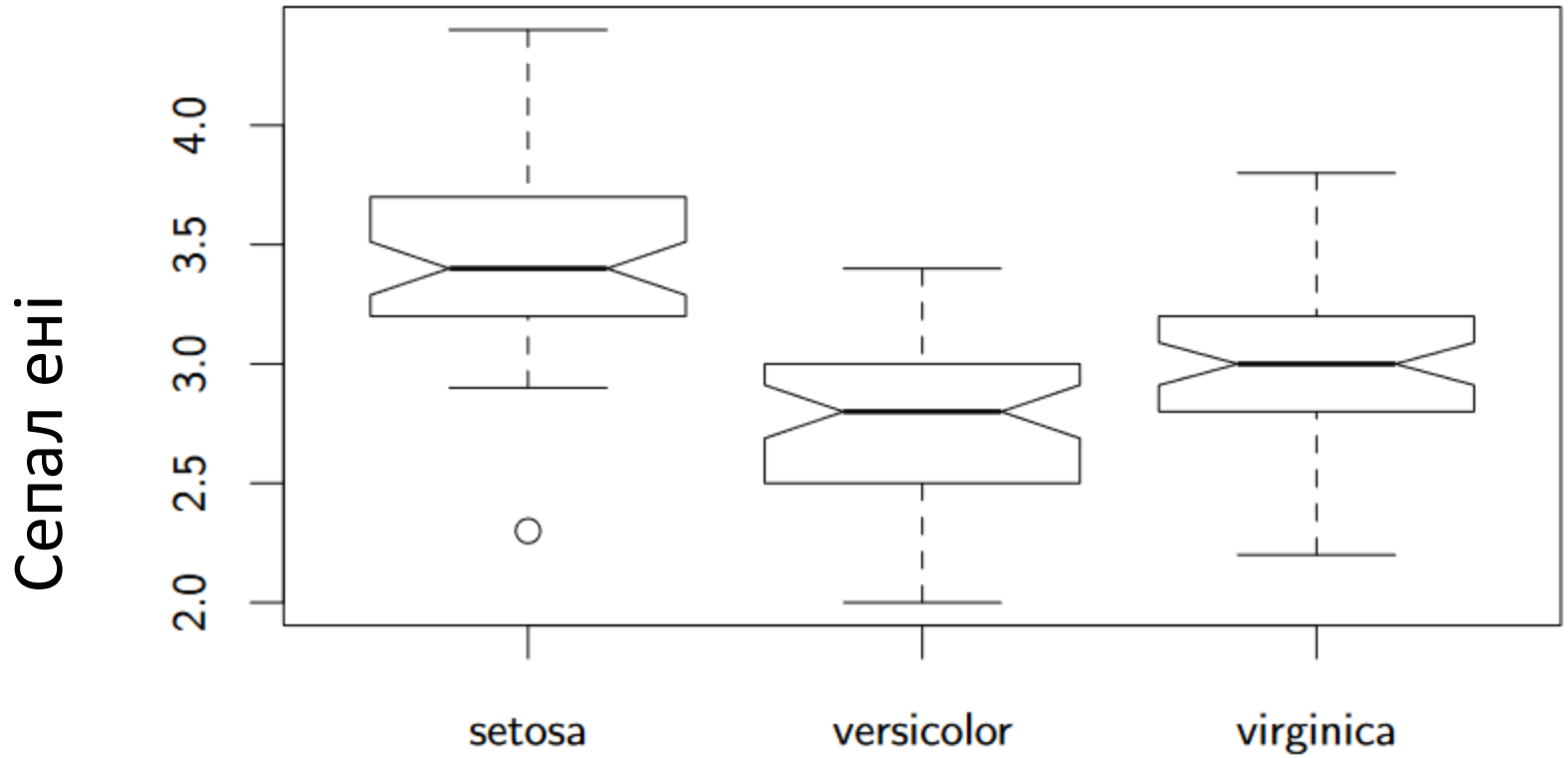
# Квартилдер

- Жалғыз сандық айнымалының төменгі немесе бірінші квартилі деректер мәндерінің 25%-ынан асады.
- Үшінші квартильдің жоғарғы бөлігі деректер мәндерінің 75%-дан асады.
- IQR квартиларалық диапазон квартилдер арасындағы айырмашылық болып табылады.
- R коды: quantile(x,c(0,25,0,75))

# Қорап сызбасы

- Қарапайым "қорап және мұрт" сызбасы төменгі және жоғарғы квартилдер арасындағы қорапты және экстремалды деректер мәндеріне дейін созылатын мұрттарды көрсетеді.
- Әдетте жоғарғы квартильден 1,5× IQR жоғары немесе төменгі квартилден төмен орналасқан шектен тыс мәндер бөлек нүктелер ретінде көрсетіледі - содан кейін мұрт тек ең шеткі емес мәндерге таралады.
- Параллель сызбалар әртүрлі топтарды салыстыру үшін пайдалы.
- R медиана үшін 95% сенімділік аралықтарын көрсету үшін ойықтар қоса алады

# Қорап сызбасы



# Жоспары

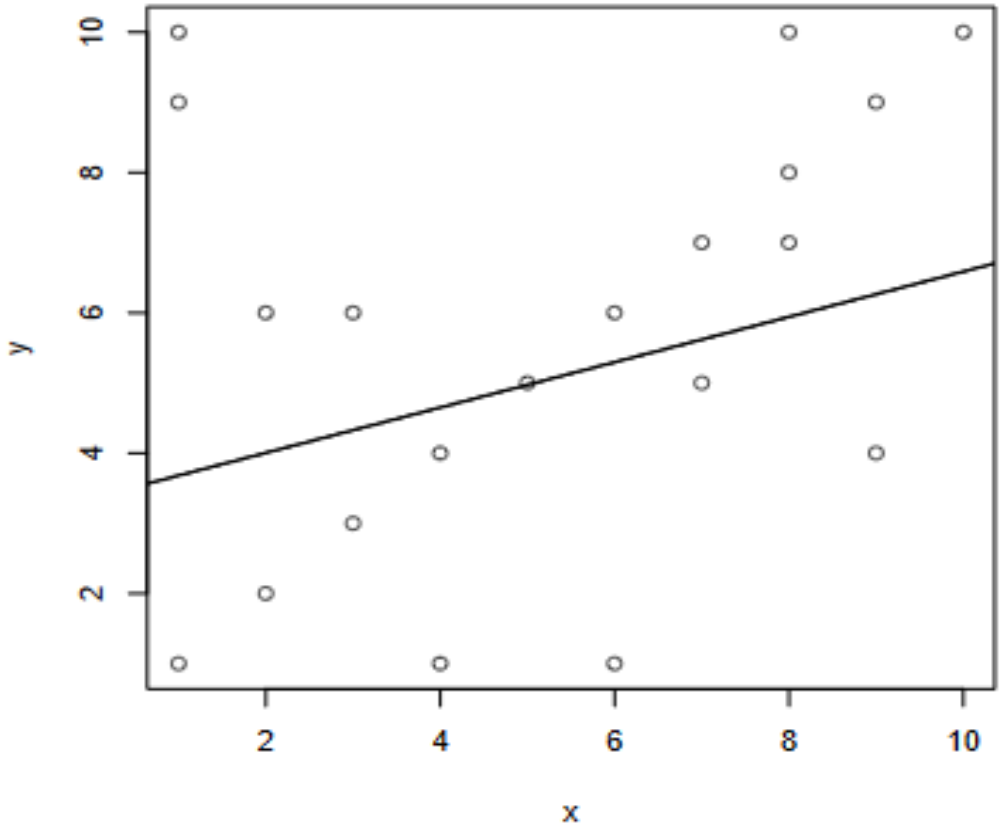
- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - **Сандық айнымалылар**
    - Бір
    - **Екі (немесе одан да көп)**
    - Көп
  - Категориялық айнымалылар
  - Екі

# Шашырау сызбасы

- Шашырау диаграммасы – x және y екі сандық айнымалының нүктелік графигі.
- яғни x y болжау үшін пайдаланылады.

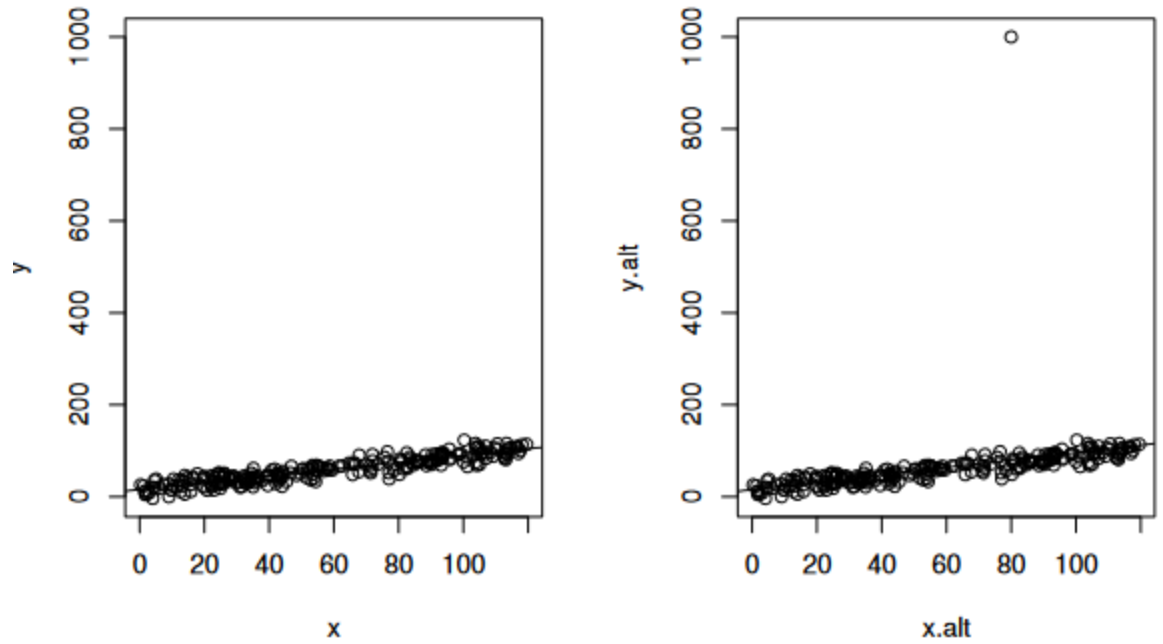
# Шашырау сызбасы

R code:
plot(x,y)
abline(lm(x y))



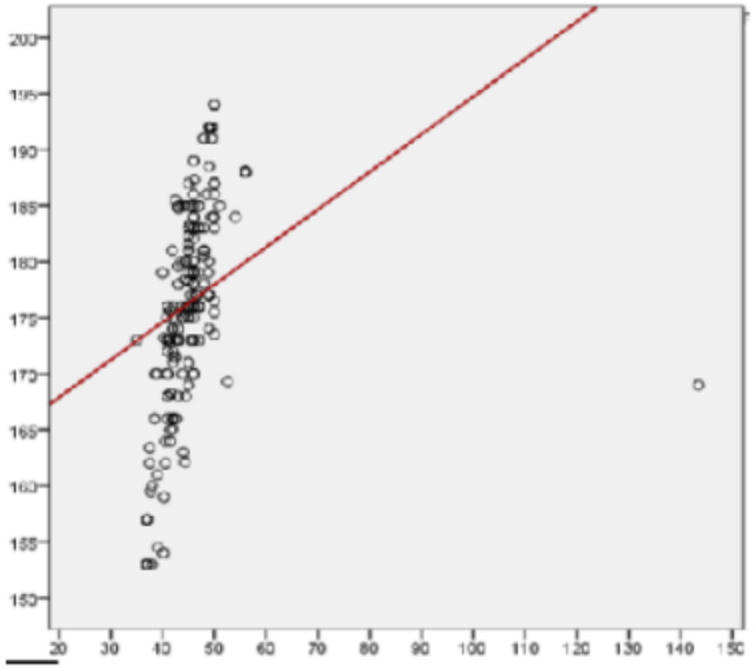
# Шектеу дегеніміз не?

- Шектеу - бұл деректердің негізгі трендінен айтарлықтай ерекшеленетін бақылау.



# Ықпал етуші нүкте дегеніміз не?

- Әрбір деректер нүктесі үлгі параметрлерінің бағалауларына әсер етеді; ықпалды бақылау мұны басқа нүктелерге қарағанда көбірек жасайды.



# Шашырау сызбасы

Шашырау сызбасы мынаны көрсетеді

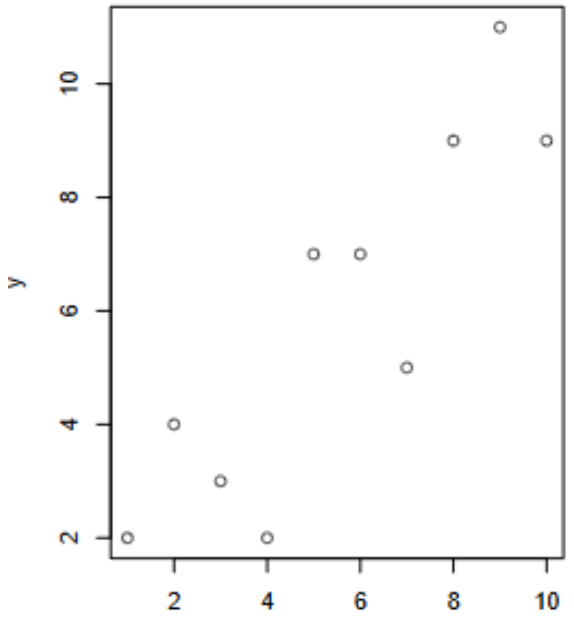
- оң/теріс және әлсіз/күшті ассоциация
- сызықтық немесе қисық қатынас
- шектен тыс мәндер
- ең қолайлы сызыққа әсер ететін ықпалды нүктелер
- кластерлер мен бос орындар

# Шашырау сызбасы

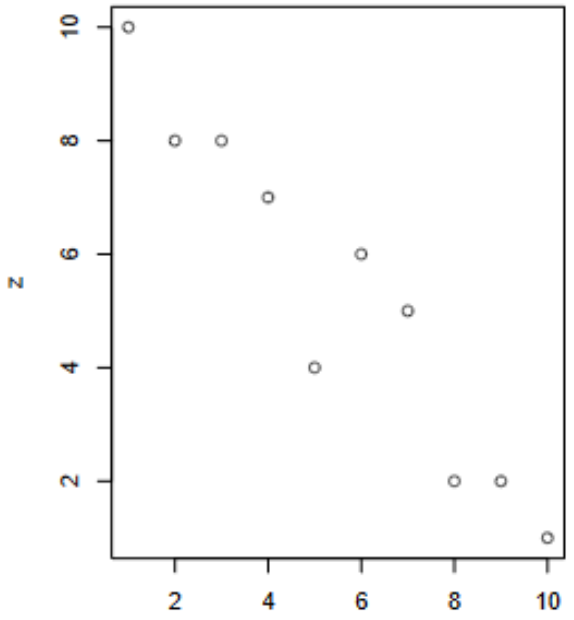
- Екі сандық айнымалылар арасындағы қатынасты шашыраңқы сызбасы ретінде көрсетуге болады.
- R коды: plot(x,y)
- Таңбаларды құру (R plot() қосымша аргументі: pch) және colours(аргумент: col)

# Ассоциация

- Бірінің жоғары мәндері екіншісінің жоғары мәндерін тудырғанда, айнымалылар *оң ассоциацияға* ие деп аталады
- Егер біріндегі жоғары мәндер екіншісінің төмен мәндерін тудырса, олар *теріс ассоциацияға* ие болады




оң ассоциация

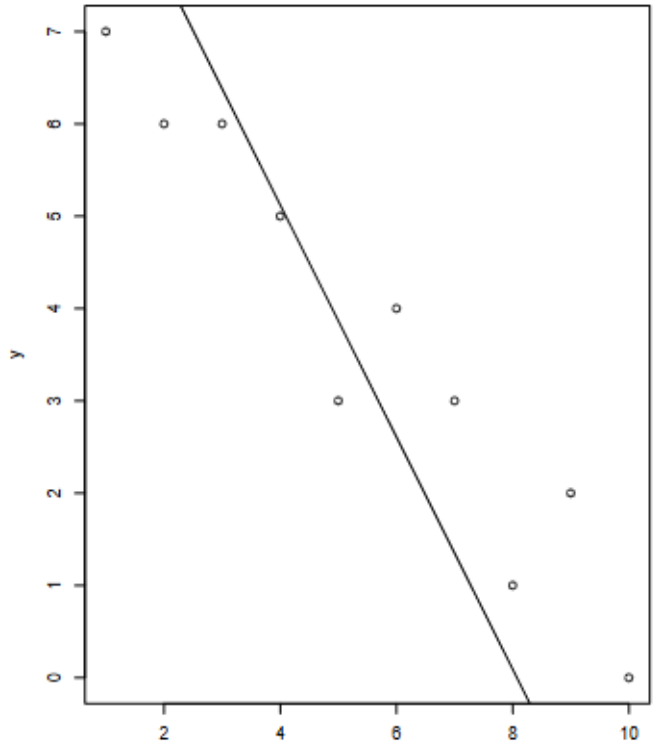



теріс ассоциация

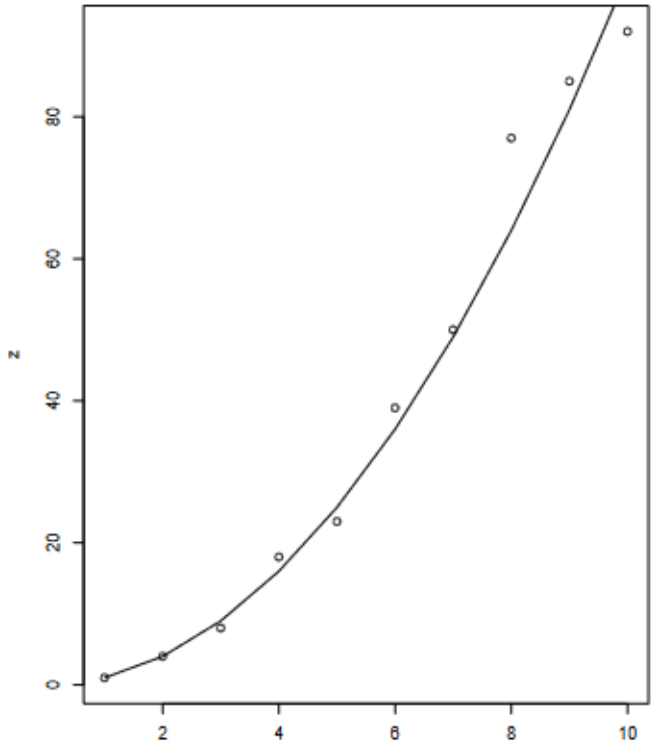
# Ең қолайлы сызық

- Егер түзу көптеген нүктелерге жақсы жақындаса, қатынас *сызықтық* деп аталады
- Егер қисық сызық жақсы жуықтау болса, ол *қисық* (сызық емес)
- Егер айнымалылар бір-бірімен байланыссыз болса, ең жақсы сәйкес келетін сызық *көлденең* болады (бірақ керісінше болуы міндетті емес)

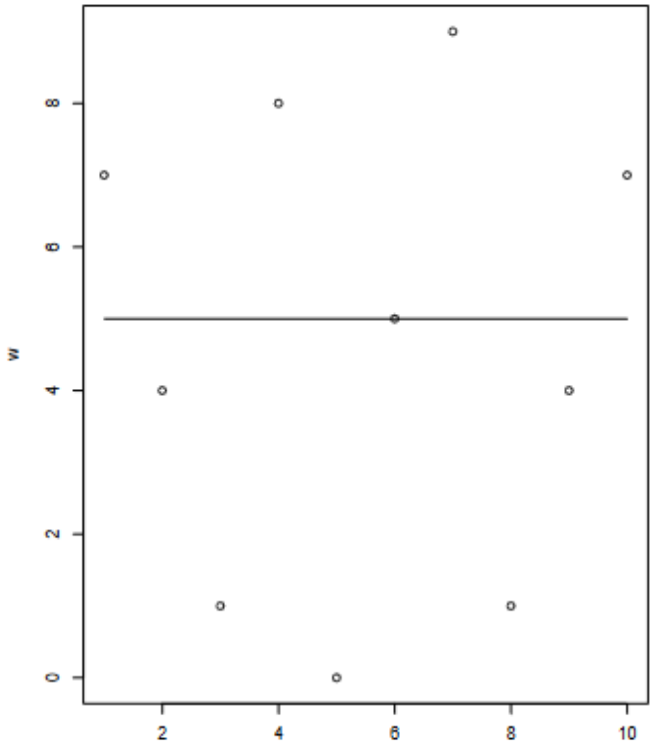
# Ең қолайлы сызық




сызықтық




қисық




көлденең

# Корреляция

Екі сандық айнымалылар арасындағы статистикалық корреляция түзу сызықты қатынастың күші мен бағытын көрсететін сан болып табылады.

Байланыстың **күші** нүктелердің түзу сызыққа жақындығымен анықталады

- **күшті** немесе **әлсіз** болуы мүмкін

**Бағыт оң** немесе **теріс** болуы мүмкін

- Екі айнымалы **оң корреляцияланады**, егер біреуінің ұлғаюы екіншісінің ұлғаюымен байланысты болса.
- Екі айнымалы **теріс корреляцияға** ие, егер біреуінің ұлғаюы екіншісінің төмендеуіне байланысты болса.

# Корреляция

**Корреляция коэффициенті** сызықтық байланыстың күші мен бағытын өлшейді.

R коды: cor(x,y)

- $-1 \leq r \leq 1$
- Белгі оң/теріс корреляцияны көрсетеді
- $r \approx 0$ байланыстың жоқтығын көрсетеді
- $r = \pm 1$ - тамаша сызықтық ассоциация

# Классикалық мысал: Анскомб квартеті

Әрқайсысында 11 бақылаудан тұратын төрт екі айнымалы (x, y) деректер жиыны.

| obs | X1 | Y1 | X2 | Y2 | X3 | Y3 | X4 | Y4 |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.5 |
| 9 | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

# Классикалық мысал: Анскомб квартеті

Әрқайсысында 11 бақылаудан тұратын төрт екі айнымалы (x, y) деректер жиыны, барлығы келесі қасиеттерге ие

$$\bar{x} = 9$$

$$s_x = 3.3166$$

$$\bar{y} = 7.5009$$

$$s_y = 2.0316$$

$$r = 0.8164$$

# Классикалық мысал: Анскомб квартеті

Әрқайсысында 11 бақылаудан тұратын төрт екі айнымалы (x, y) деректер жиыны, барлығы келесі қасиеттерге ие

Олар қаншалықты әртүрлі болуы мүмкін?

$$\bar{x} = 9$$

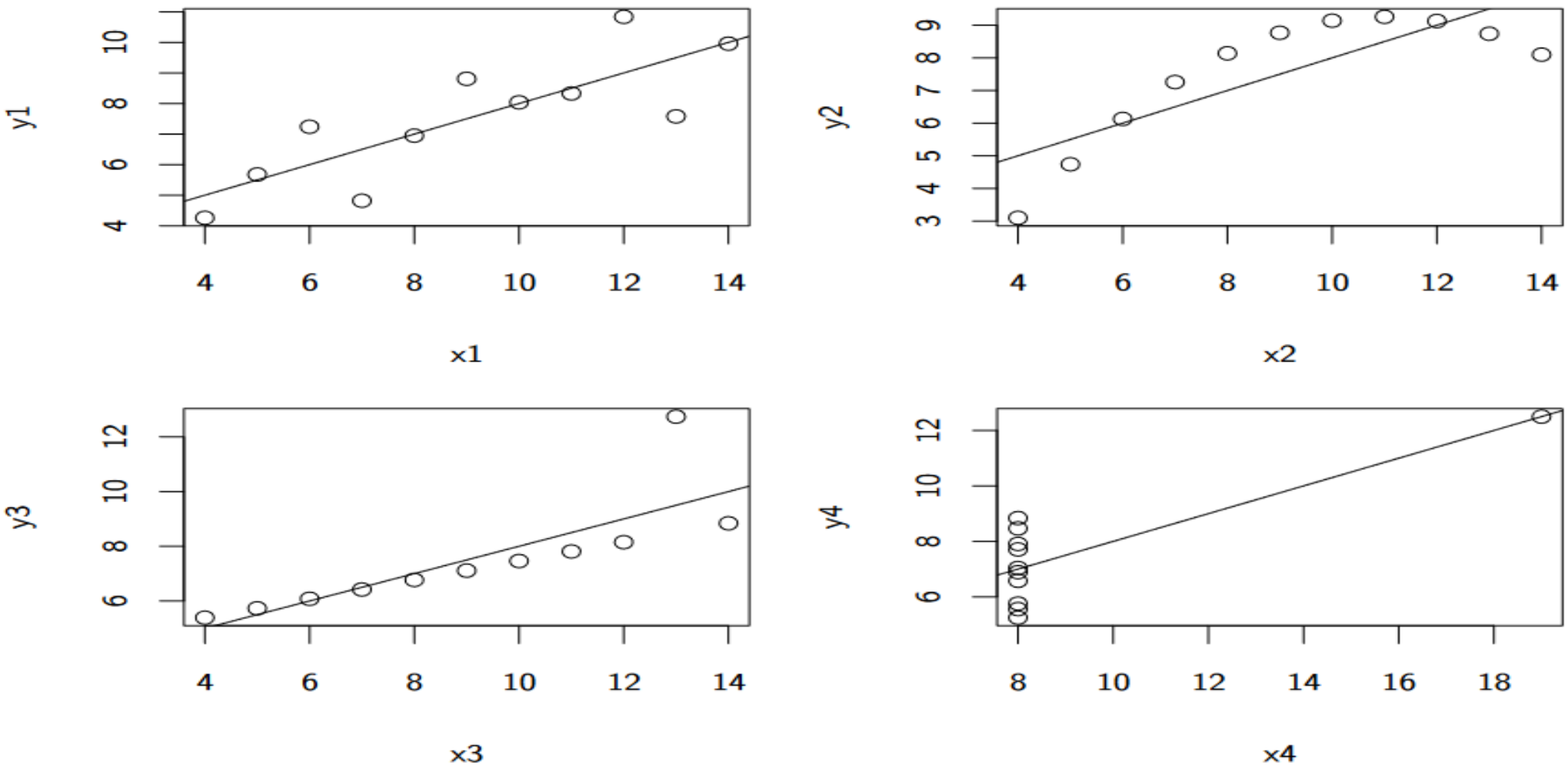
$$s_x = 3.3166$$

$$\bar{y} = 7.5009$$

$$s_y = 2.0316$$

$$r = 0.8164$$

# Өте әртүрлі


y1
4
6
8
10
4
6
8
10
12
14
x1
y2
3
4
5
6
7
8
9
4
6
8
10
12
14
x2
y3
6
8
10
12
4
6
8
10
12
14
x3
y4
6
8
10
12
8
10
12
14
16
18
x4

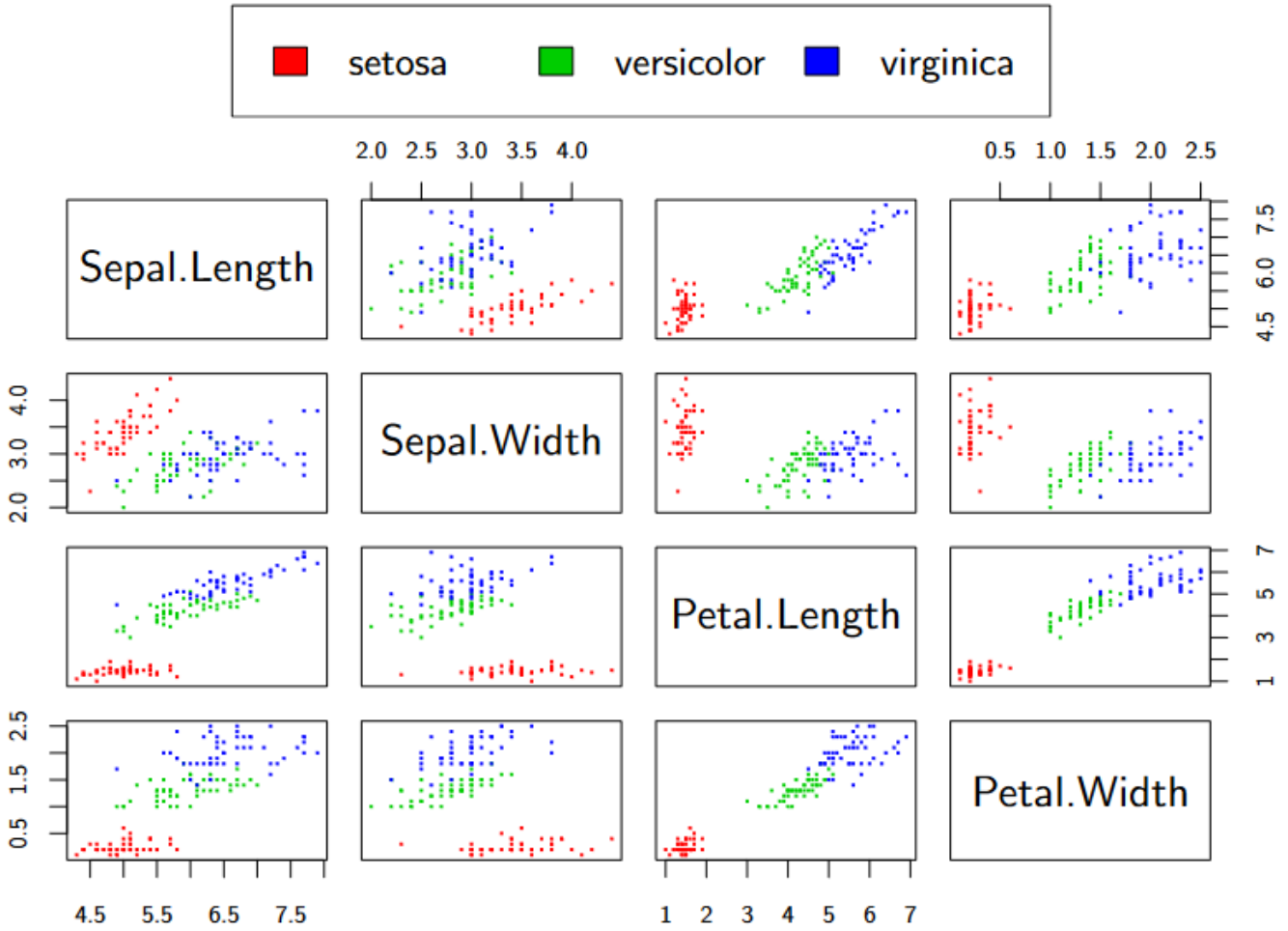
# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - **Сандық айнымалылар**
    - Бір
    - Екі (немесе одан да көп)
    - **Көп**
  - Категориялық айнымалылар
  - Екі

# Шашырау матрицасы

- Шашырау диаграммасы матрицасы – осьтері тураланған, әр атрибут бір-біріне (өзінен басқа) атрибутқа қарсы салынған шашыраңқы диаграммалар массиві
- R коды: pairs(x) (x бірнеше бағандары бар)

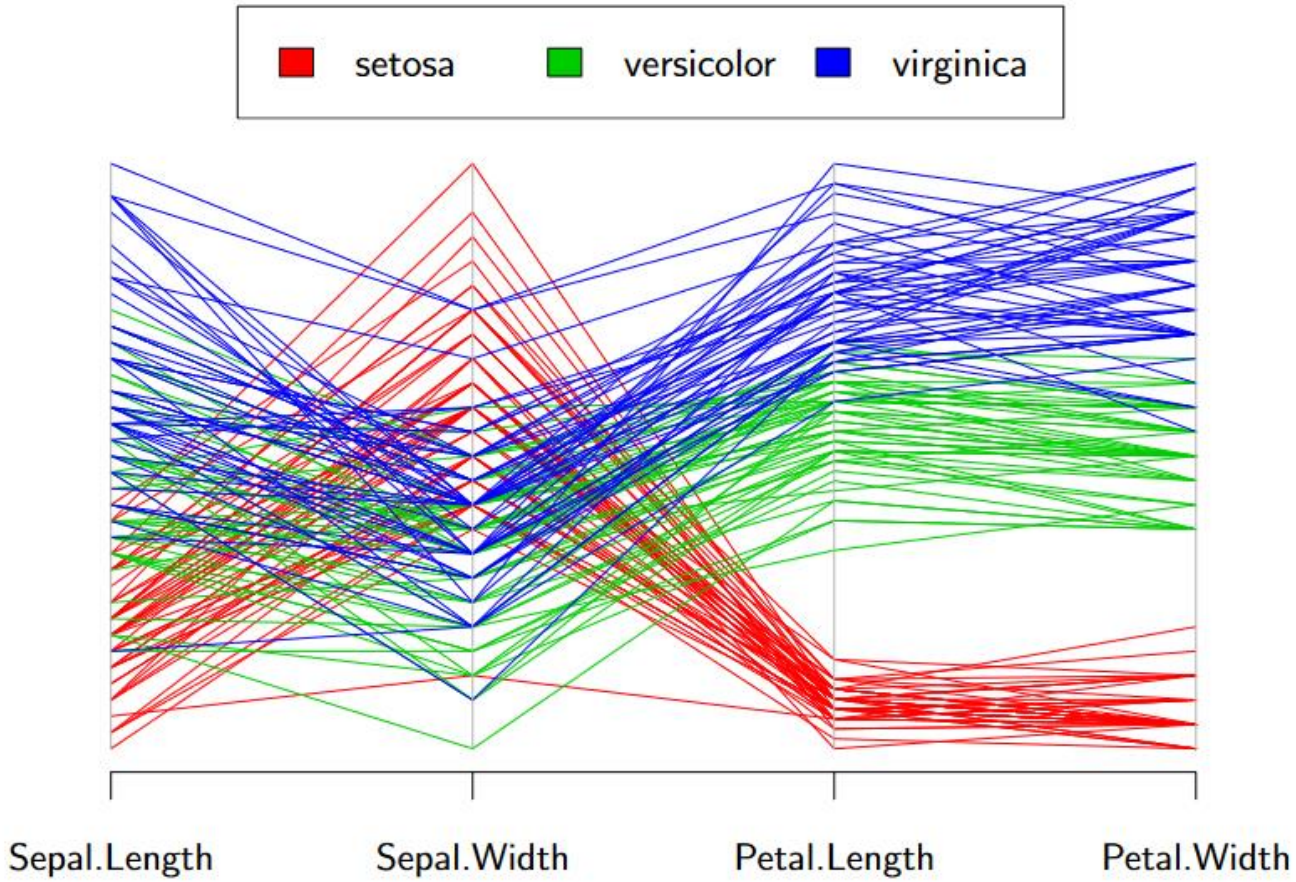
# Шашырау матрицасы



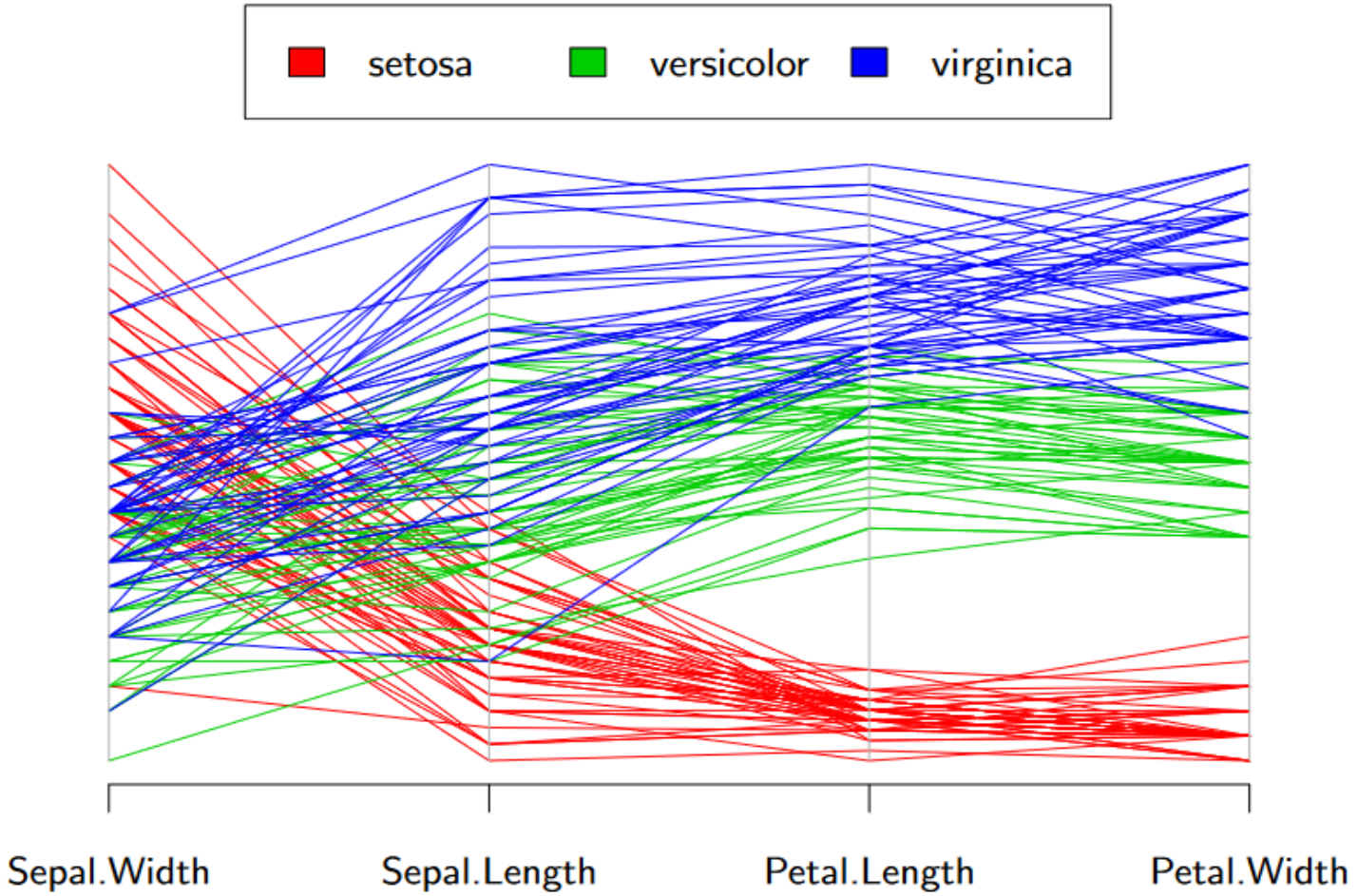
# Параллель координаталар сызбасы

- Көптеген айнымалыларды визуализациялаңыз, әдетте салыстырмалы түрде шағын үлгімен.
  - 1. Көлденең осьте айнымалылар және тік осьтерде мәндер (стандартталған) бар әр айнымалыдан әрбір бақылауды сызыңыз.
  - 2. Бір бақылауға сәйкес нүктелерді түзумен қосыңыз.
- R коды: parcoord(x) (x бірнеше бағандары бар)
- Айнымалы мәндердің реті маңызды болуы мүмкін.

# Параллель координаталар сызбасы (1 реттік)



# Параллель координаталар сызбасы (2 реттік)



# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - Сандық айнымалылар
  - **Категориялық айнымалылар**
    - Бір
  - Екі

# Мысал деректер жинағы: Статистика студенттерінің шашы мен көзінің түсі

Үш категориялық айнымалы бойынша 592 студенттен алынған

- Шаш түсі: қара, қоңыр, қызыл, аққұба
- Көздің түсі: қоңыр, көк, жаңғақ, жасыл
- Жынысы: Ер, Әйел

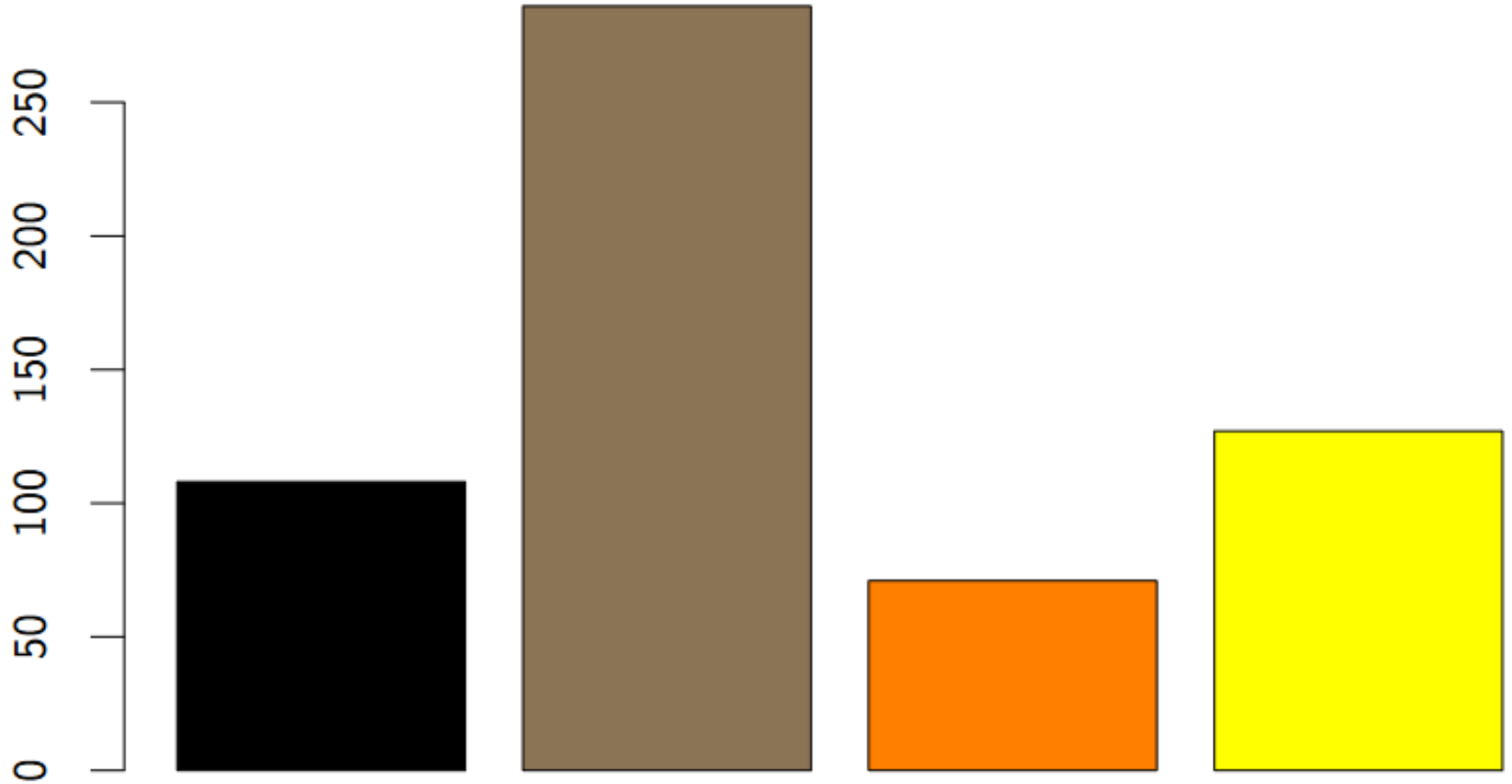
# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - Сандық айнымалылар
  - **Категориялық айнымалылар**
    - **Бір**
  - Екі

# Барплоттар

- Гистограммалардан категориялық көлденең осіне ие болу арқылы ерекшеленеді.
- Әдетте, жолақтардың арасындағы кішкене алшақтық.
- Жиіліктер немесе пропорциялар болуы мүмкін.
- R коды: barplot(x) (x а векторы)

# Шаш түсінің барплоты



# Жоспары

- Шолу
  - Визуализация дегеніміз не?
  - Неліктен визуализация?
  - Қалай визуализациялау керек?
- **Техникалар**
  - Сандық айнымалылар
  - Категориялық айнымалылар
  - **Екі**

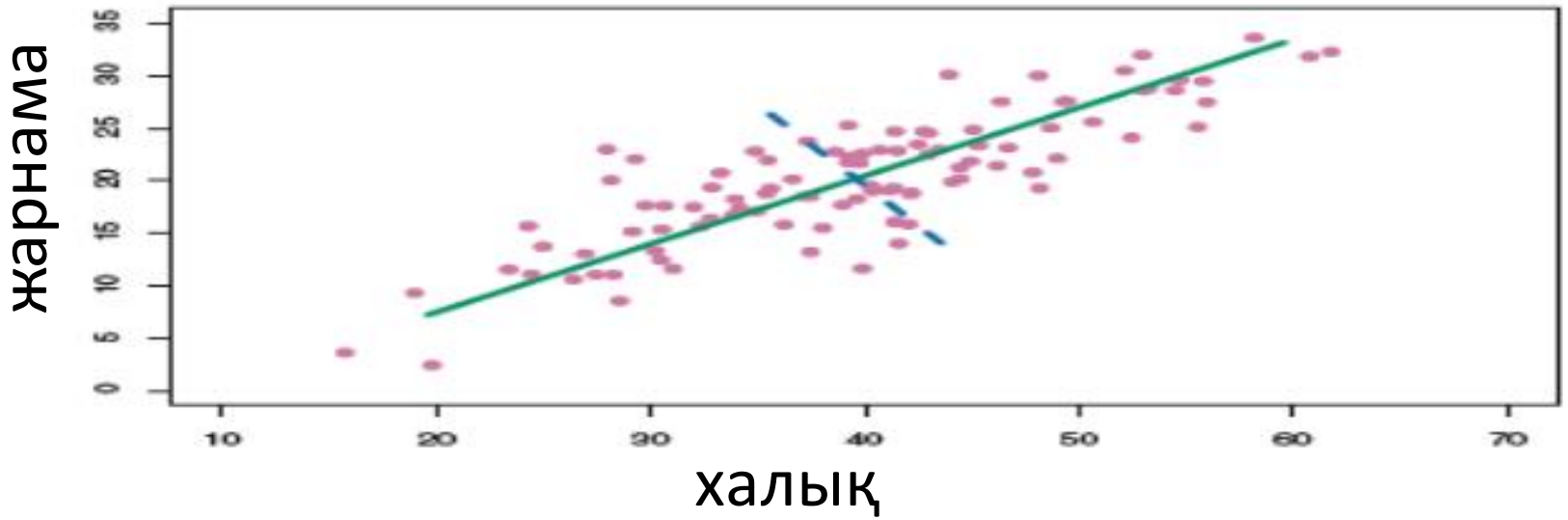
# Негізгі құрамдас талдау (PCA, өлшемді азайту техникасы)

- Барлау деректерін талдаудың бөлігі ретінде, X1, ..., Xp мүмкіндіктерінің жиыны бойынша өлшеулермен n бақылауды көргіміз келеді делік.
- Біз деректердің екі өлшемді шашырау графиктерін зерттей аламыз, олардың әрқайсысында екі мүмкіндік бойынша n бақылау өлшемдері бар.
- Дегенмен, p(p − 1)/2 мұндай шашырау сызбалары бар. Егер p үлкен болса, онда олардың барлығын қарау мүмкін болмайтыны сөзсіз.
- p үлкен болғанда n бақылауды визуализациялау үшін жақсырақ әдіс қажет.

# Негізгі құрамдастарды талдау

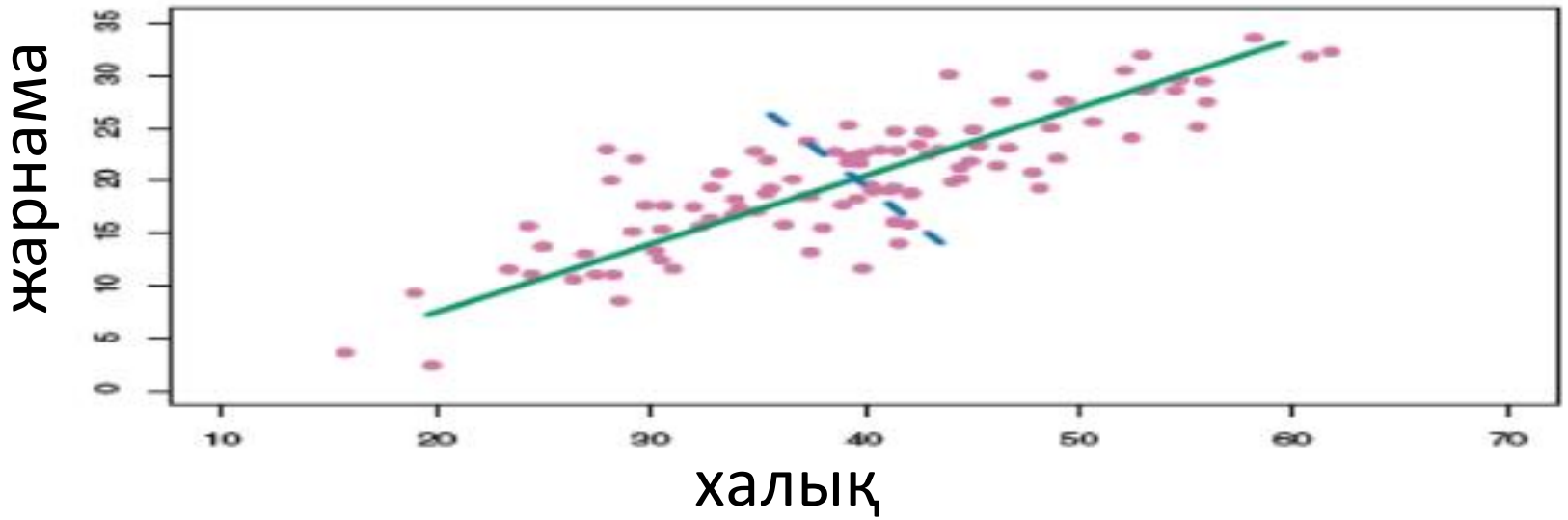
- Негізгі құрамдастарды талдау (PCA) - айнымалылардың үлкен жиынынан аз өлшемді мүмкіндіктер жинағын шығарудың танымал тәсілі.
- PCA әсіресе бір шкалада өлшенетін және жоғары корреляциялық өлшемдер жиыны болған кезде құнды.
- Ол бастапқы айнымалылардың өлшенген сызықтық комбинациялары болып табылатын және толық бастапқы жиынның ақпаратының көпшілігін сақтайтын бірнеше айнымалы мәндерді (көбінесе үшке дейін) қамтамасыз етеді.
- PCA сандық айнымалылармен пайдалануға арналған. Категориялық айнымалылар үшін корреспонденцияны талдау сияқты басқа әдістер қолайлырақ (мұнда қарастырылмаған).

# Негізгі құрамдастарды талдау




- Суретте он мыңдаған адамдағы халық саны және 100 қала үшін компанияның жарнамалық шығындары мыңдаған доллармен көрсетілген.
- Олардың жақсы қарым-қатынасы күшті сияқты.

# Негізгі құрамдастарды талдау




- Ақпараттың көп бөлігін сақтай отырып, айнымалылар санын азайту үшін осы фактіні пайдалана аламыз ба?
- Екі айнымалы бар ақпаратта артық болғандықтан, тым көп ақпаратты жоғалтпай екі айнымалыны бір айнымалыға дейін азайтуға болады.

# Негізгі құрамдастарды талдау

- PCA идеясы жаңа айнымалы екі бастапқы айнымалыны алмастыра алатындай ақпараттың көп бөлігін қамтитын екі айнымалының сызықтық комбинациясын табу болып табылады.
- Суретте ондаған мың адамдағы халық саны және 100 қала үшін компанияның жарнамалық шығындары мыңдаған доллармен көрсетілген.
- Жасыл тұтас сызық - деректердегі ең көп өзгергіштік бар немесе деректер ең көп өзгеретін бағыт. Бұл бірінші негізгі компонент.
- Көк үзік сызық екінші негізгі құрамдастарды көрсетеді.