Деректердің таралуын зерттеу

Біз қарастырған барлық бағалар орталық ережені немесе деректердің өзгергіштігін сипаттау үшін деректерді бір санмен қорытындылайды. Сонымен қатар, жалпы таратылған деректердің сипатын зерттеу пайдалы.

Негізгі терминдер

Қорап диаграммасы (boxplot) - бұл деректерді таратуды визуализациялаудың жылдам әдісі ретінде Тьюки қолданған График.

Синонимі: "мұртты қорап" типіндегі диаграмма.

Жиілік кестесі (frequency table) жиынтық жиілік интервалдарының сериясына бөлінген сандық мәндердің саны (себеттер, бинттер).

Гистограмма (histogram) жиілік кестесінің графигі, онда жиілік интервалдары X осіне, ал саны (немесе үлесі) y осіне қойылады.

Тығыздық графигі (densityplot) гистограмманың тегістелген нұсқасы, көбінесе ядролық тығыздықты бағалауға негізделген. Процентильдер және қораптық диаграммалар.

"Процентильдерге негізделген бағалау" осы тарауда біз процентильдерді деректердің таралуын өлшеу үшін қалай қолдануға болатындығын қарастырдық. Жалпы үлестірімді жалпылау үшін пайыз немесе маңызды. Квартильдер (25-ші, 50-ші және 75-ші перцентильдер) және декильдер (10-шы, 20-шы, ..., 90-шы процентили). Процентильдер таралудың құйрықтарын жалпылау үшін өте маңызды. Бұқаралық мәдениет байлықтың жоғарғы 99 пайызындағы адамдарға қатысты "бір пайыз" терминін қолданды. Кестеде. 1.4 мемлекеттік өлтіру деңгейінің кейбір процентильдерін көрсетеді. R-де оларды функцияның көмегімен алуға болады.

quantile:

```
quantile(state[["Murder.Rate"]], p=c(.05, .25, .5, .75, .95))   5%   25%   50%
75%   95% 1.600 2.425 4.000 5.550 6.510
```

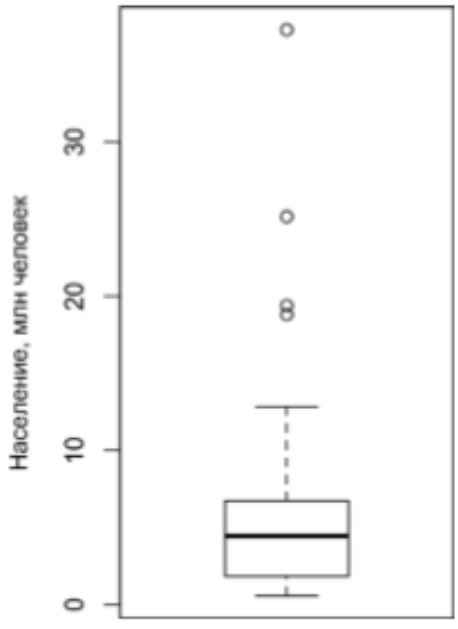
5%	25%	50%	75%	95%
1,60	2,42	4,00	5,55	6,51

1.4-кесте. Мемлекет бойынша кісі өлтіру пайызы

Медиан 100 мың адамға 4 кісі өлтіруге тең, бұл өте үлкен өзгергіштікке қарамастан: 5-ші процентиль тек 1,6, ал 95 — ші процентиль-6,51. Тьюки [Tukey-1977] қолданған қораптық диаграммалар процентильдерге негізделген және деректердің таралуын визуализациялаудың жылдам әдісін ұсынады. Суретте. 1.2 r-де алынған штат бойынша халықтың қораптық диаграммасы ұсынылған:

boxplot(state[["Population"]]/1000000, ylab="Население, млн человек")

Қораптың жоғарғы және төменгі жағы сәйкесінше 75 - ші және 25-ші процентильдер болып табылады. Медиан қорапта көлденең сызықпен көрсетілген. Мұрт деп аталатын нүктелі сызықтар жоғарғы және төменгі жағынан шығады және деректердің негізгі бөлігінің ауқымы туралы айтады. Қорап диаграммасының көптеген нұсқалары бар; мысалы, Boxplot R-функциясы [R-base-2015] бойынша құжаттаманы қараңыз. Әдепкі бойынша, бұл R функциясы мұртты қораптың сыртындағы ең алыс нүктеге дейін созады, егер ол 1,5-ке көбейтілген квартиль аралық ауқымнан (МКР немесе IQR) шықпаса (басқа бағдарламалық жүйелерде басқа ережелер қолданылуы мүмкін). Мұрттан тыс барлық деректер бір нүкте ретінде көрсетіледі.



1.2.-сурет. Мемлекеттің популяциялық қорапшасы

Жиілік кестесі және гистограммалар

Айнымалы жиілік кестесі айнымалы диапазонды тең сегменттерге бөледі және әр сегментке қанша мән түсетіні туралы хабарлайды. Кестеде. 1.5 штат бойынша халық санының жиілік кестесі көрсетілген, R-де есептелген

```
breaks <- seq(from=min(state[["Population"]]),
                    to=max(state[["Population"]]), length=11)
pop_freq <- cut(state[["Population"]], breaks=breaks,
                    right=TRUE, include.lowest = TRUE)
table(pop_freq)
```

Халқы ең аз штат - Вайоминг, халқы 563,626 адам (2010 жылғы халық санағы бойынша), ал халқы ең көп штат - Калифорния, халқы 37,253,956 адам. Бұл бізге серпіліс береді 37 253 956 563 626 36 690 330 −= , біз оны тең жиілік интервалдарына бөлуіміз керек - айталық, 10 in-terval. 10 тең өлшемді интервалдарда әр жиілік интервалының ені 3,669,033 болады, осылайша бірінші интервал 563,626-дан 4,232,658-ге дейін созылады. Керісінше, жоғарғы интервал, 33,584,923-37,253,956, тек бір штат бар: Калифорния. Калифорния штатынан төмен екі интервал біз Техас штатына жеткенше бос. Бос аралықтарды үйрету маңызды; бұл аралықтарда мәндердің болмауы пайдалы ақпарат болып табылады. Әр түрлі интервалдармен тәжірибе жасау пайдалы болуы мүмкін. Егер олар тым үлкен болса, онда маңызды тарату белгілері күңгірт болуы мүмкін. Егер олар тым кішкентай болса, онда нәтиже тым егжей-тегжейлі болады және үлкен суретті байқау мүмкіндігі жоғалады.

№ интервала	Диапазон интервала	Количество	Штаты
1	563 626– 4 232 658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV, NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4 232 659– 7 901 691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7 901 692– 11 570 724	6	VA,NJ,NC,GA,MI,OH
4	11 570 725– 15 239 757	2	PA,IL
5	15 239 758– 18 908 790	1	FL
6	18 908 791– 22 577 823	1	NY
7	22 577 824– 26 246 856	1	TX
8	26 246 857– 29 915 889	0	
9	29 915 890– 33 584 922	0	
10	33 584 923– 37 253 956	1	CA

Жиілік кестелері де, процентильдер де жиілік аралықтарын құру арқылы мәліметтерді жинақтайды. Жалпы алғанда, квартильдер мен декильдер әр аралықта бірдей болады (сандық интервалдарға тең), бірақ Интервалдардың мөлшері әр түрлі болады. Жиілік кестесінде, олардан айырмашылығы, интервалдарда әртүрлі сандар болады (тең өлшемді интервалдар).

Гистограмма-бұл жиілік кестесін визуализациялау әдісі, онда жиілік біліктері X осіне, ал мәліметтер саны y осіне қойылады. кестеге сәйкес келетін гистограмма жасау. 1.5 R тілінде аргументі бар hist функциясы қолданылады

breaks:

hist(state[["Population"]], breaks=breaks)

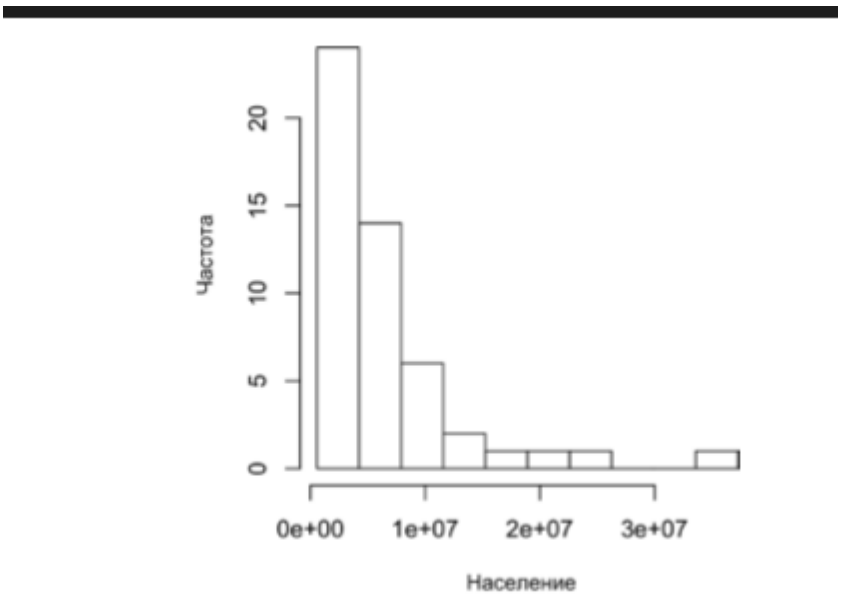
Нәтиже-суретте көрсетілген гистограмма. 1.3. Жалпы, гистограммалар келесідей көрсетіледі:

-бос аралықтар кестеге енгізілген;

-аралықтардың ені бірдей;

-пайдаланушы интервалдар санын (немесе сол сияқты интервал өлшемін) белгілейді;

гистограмма бағандары үздіксіз-бос интервал болмаса, олардың арасында бос орындар болмайды



Сурет. 1.3. Штат тұрғындарының гистограммасы

Статистикалық теориядағы статистикалық моменттер орталық позиция және өзгергіштік (тербеліс, өзгергіштік) бірінші және екінші ретті бөлу моменттері деп аталады. Үшінші және төртінші ретті сәттер-асимметрия және асып кету. Асимметрия дегеніміз — деректердің үлкен немесе кіші мәндерге ауысуы, ал асып кету дегеніміз-деректердің шекті мәндерге бейімділігі. Асимметрия мен асып кетуді өлшеу үшін, әдетте, метрикалық көрсеткіштер қолданылмайды; оның орнына олар сурет сияқты визаны көрсету кезінде анықталады. 1.2 және 1.3.

Тығыздықты бағалау

Тығыздықты бағалау гистограммамен тығыз сызық түрінде деректер мәндерінің таралуын көрсететін тығыздық графигі байланысты. Тығыздық графигін, әдетте, тығыздықты ядролық бағалау арқылы деректерден тікелей есептелетініне қарамастан, тегістелген гистограмма ретінде қарастыруға болады (Қысқаша нұсқаулық бойынша [Duong - 2001] қараңыз). Суретте. 1.4 гистограммаға салынған тығыздықты бағалау ұсынылған. R-де тығыздықты бағалау функцияның көмегімен есептеледі

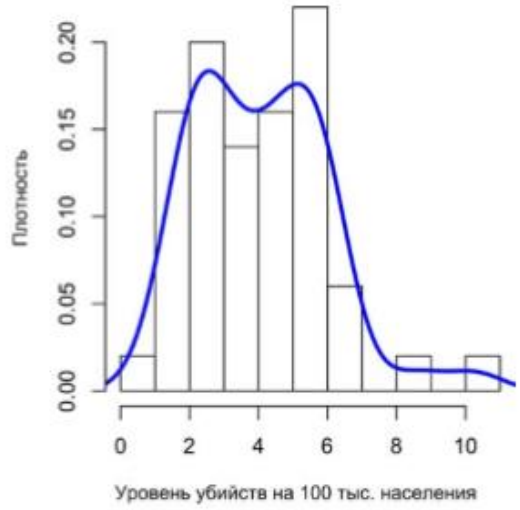
density:

hist(state[["Murder.Rate"]], freq=FALSE) lines(density(state[["Murder.Rate"]]), lwd=3, col="blue")

Суретте көрсетілген гистограммадан негізгі айырмашылық. 1.3, у осінің шкаласынан тұрады: тығыздық графигі гистограмманың санына емес, үлес

ретінде көрсетілуіне сәйкес келеді (в R она задается при помощи аргумента freq=FALSE).

Тығыздықты бағалау тығыздықты бағалау-статистикалық әдебиетте ұзақ тарихы бар кең тақырып. Шын мәнінде, тығыздықты бағалау функцияларын ұсынатын 20-дан астам R бағдарламалық пакеті жарияланды. [[Deng - Wickham-2011] ASH және KernSmooth бағдарламалық пакеттеріне ерекше назар аудара отырып, R пакеттеріне жан-жақты талдау жасайды. Деректер ғылымының көптеген міндеттері үшін тығыздықты бағалаудың әртүрлі түрлері туралы алаңдамаудың қажеті жоқ; негізгі функцияларды қолдану жеткілікті.



1.4.- Сурет. Мемлекеттегі кісі өлтіру деңгейінің тығыздығы

Деректерді таратуды зерттеудің негізгі идеялары •

Жиілік гистограммасы y осіндегі жиіліктерді және X осіндегі айнымалы мәндерді көрсетеді; ол деректердің таралуы туралы визуалды түсінік береді. •

Жиілік кестесі-бұл гистограммадан табуға болатын жиілікті есептеудің кестелік нұсқасы. •

Қорап диаграммасы-қораптың жоғарғы және төменгі жағы сәйкесінше 75-ші және 25-ші процентильде орналасқан диаграмма - сонымен қатар деректердің таралуы туралы тез түсінік береді; ол көбінесе үлестірімдерді салыстыру үшін жұптастырылған графиктерде қолданылады. •

Тығыздық графигі-бұл гистограмманың тегістелген нұсқасы; мәліметтер негізінде графиканы бағалау үшін арнайы функция қажет (әрине, көптеген бағалау мүмкін).

Екілік және категориялық деректерді зерттеу

Егер категориялық деректер туралы айтатын болсақ, онда олар туралы қарапайым үлестер немесе пайыздар ұсынылады.

Негізгі терминдер

Сән (mode) - бұл мәліметтер жиынтығындағы ең көп таралған санат немесе мән.

Математикалық күту (expected value) Санаттар сандық мәндермен байланысқан кезде, бұл сан санаттың пайда болу ықтималдығына негізделген орташа мәнді береді. Синонимі: күтілетін мән.

Баған диаграммалары (bar charts) тіктөртбұрыш түрінде көрсетілген әр санаттың жиілігі немесе үлесі.

Дөңгелек диаграммалар (pie charts) шеңбер секторы ретінде көрсетілетін әр санаттың жиілігі немесе үлесі.

Екілік немесе бірнеше категориялы категориялық айнымалы туралы жиынтық ақпарат алу өте қарапайым міндет: біз жай ғана (1) немесе маңызды санаттардың үлесін анықтаймыз. Мысалы, кестеде. 1.6 Даллас-Форт-Уэрт әуежайындағы проблемаларға байланысты кешіктірілген рейстердің пайызы көрсетілген, 2010 жылдан бастап кешіктірулер тасымалдаушы (тасымалдаушы), әуе қозғалысын басқарудағы жүйелік кідірістер (ATC), ауа - райы жағдайлары (ауа-райы), қауіпсіздік (қауіпсіздік) немесе қауіпсіздік (қауіпсіздік) факторларына байланысты жіктеледі ҚҚБ - келетін әуе кемесінің ғимараты (Inbound).

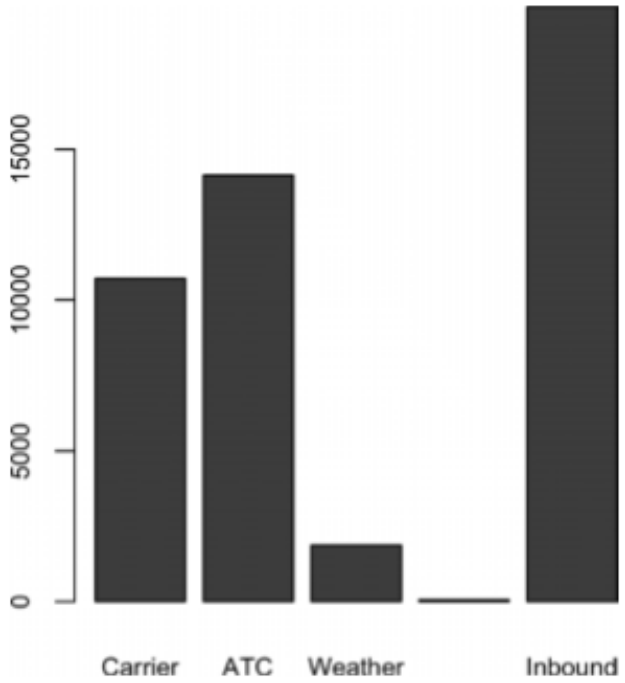
Carrier	ATC	Weather	Security	Inbound
23,02	30,40	4,03	0,12	42,43

1.6-кесте. Даллас-Форт-Уэрт әуежайындағы кешігу пайызы

Бағандық диаграммалар-бұл танымал баспасөзде жиі кездесетін жалғыз категориялық айнымалыны таңдауға арналған жалпы қабылданған визуалды құрал. Санаттар x осінде, ал жиіліктер немесе фракциялар y осінде көрсетілген. 1.5 Даллас-Форт-Уэрт әуе портындағы проблемаларға байланысты рейстердің жылдық кідірістерін көрсетеді; кесте R функциясы арқылы жасалады

barplot:

barplot(as.matrix(dfw)/6, cex.axis=.5)



1.5. – сурет. Даллас-Форт-Уэрт әуежайындағы мәселелерге байланысты рейстің кешігуінің бағаналы диаграммасы

Бағандық диаграмма гистограммаға ұқсайтынын ескеріңіз; бағандық диаграммада X осі факторлық айнымалының әртүрлі санаттарын білдіреді, ал гистограммада X осі сандық шкала бойынша бір айнымалының мәндерін білдіреді. Гистограммада тіктөртбұрыштар әдетте бір-біріне жақын орналасады, ал үзілістер мәліметтерде мәндер жоқ екенін көрсетеді. Бағандық диаграммада тіктөртбұрыштар бір-бірінен бөлек көрсетіледі. Дөңгелек диаграммалар бағандық диаграммаларға балама болып табылады, дегенмен статистика мамандары мен деректерді визуализациялау мамандары

әдетте дөңгелек диаграммаларды визуалды емес ақпарат ретінде қарастырады

Сандық деректер

Бөліктегі категориялық деректер ретінде. ""Жиілік кестесі және гистограмма" бұрын осы тарауда біз мәліметтерді жиілік интервалдарына бөлуге негізделген жиілік кестелерін қарастырдық, нәтижесінде сандық мәліметтер нақты түрде реттік факорге айналады. Бұл мағынада гистограммалар мен бағандық диаграммалар ұқсас, бір қоспағанда — бағандық диаграммадағы X осіндегі Санаттар тапсырыс берілмейді. Сандық деректерді категорияға айналдыру деректерді талдауда маңызды және кеңінен қолданылатын кезең болып табылады, өйткені бұл процедура деректердің күрделілігін (және мөлшерін) азайтады. Бұл белгілер арасындағы байланыстарды, әсіресе талдаудың бастапқы кезеңдерінде анықтауға көмектеседі.

Корреляция

Көптеген модельдеу жобаларында (деректер ғылымында немесе статистикалық зерттеуде) деректерді барлау болжаушылар мен болжаушылар мен мақсатты айнымалы арасындағы корреляцияны зерттеумен байланысты. Егер X - тің жоғары мәндері Y-нің Жоғары мәндерімен, ал X-нің төмен мәндері y-нің төмен мәндерімен бірге жүрсе, X және Y-нің өзгеруі (әрқайсысы өлшеу деректерімен) оң арақатынаста болады деп айтылады.егер X-тің жоғары мәндері Y-нің төмен мәндерімен бірге жүрсе және керісінше, айнымалылар теріс корреляцияланады.

Негізгі терминдер

Корреляция коэффициенті (correlation coefficient) - бұл сандық ауысулардың бір − бірімен байланысты дәрежесін өлшейтін метрикалық көрсеткіш (1-ден 1 + - ке дейінгі диапазонда).

Корреляциялық матрица (correlation matrix) жолдар мен бағандар айнымалылар болатын кесте және ұяшық мәндері осы айнымалылар арасындағы корреляция болып табылады.

Шашырау диаграммасы (scatterplot) — X осі бір айнымалының мәні, ал y осі екіншісінің мәні болатын График

Осы екі айнымалыны қарастырайық, олардың әрқайсысы төмен мәннен жоғары мәнге параллель жүреді:

v1: {1, 2, 3}

v2: {4, 5, 6}

Көбейтінділердің векторлық қосындысы 4+10+18=32. Енді олардың біреуін ауыстырып, көбейтінділердің векторлық қосындысын есептеп көрейік, енді 32-ден аспайды. Сондықтан туындылардың бұл сомасы метрикалық көрсеткіш ретінде пайдаланылуы мүмкін; яғни 32 - ге тең байқалатын соманы көптеген еркін араласулармен салыстыруға болады (іс жүзінде бұл идея қайта іріктеу негізінде бағалауға қатысты: қараңыз. 3 - тараудың" тоқтату тесті"). Осы метрикалық индикатормен өндірілген мәндер, алайда, екінші үлгілердің таралуына сүйенуден басқа, аса маңызды емес. Неғұрлым пайдалы стандартталған нұсқа-әрқашан бірдей өлшеу шкаласында болатын екі айнымалы арасындағы корреляцияны бағалау үшін корреляция коэффициенті. Пирсонның корреляция коэффициентін есептеу үшін 1 айнымалы үшін орташа мәннен ауытқуды 2 айнымалы үшін бірдей көбейтеміз, содан кейін нәтижені стандартты клондау көбейтіндісіне бөлеміз:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

Біз 1 n − және n-ге бөлетінімізді ескеріңіз ("еркіндік дәрежесі және n немесе n-1?"бұрын осы тарауда қосымша ақпарат алу үшін). Корреляция коэффициенті әрқашан 1 + (идеалды оң корреляция) және 1 − (идеалды теріс корреляция) арасында болады; 0 корреляцияның жоқтығын көрсетеді. Айнымалылардың сызықтық емес байланысы болуы мүмкін, бұл жағдайда корреляция коэффициенті пайдасыз метрикалық көрсеткіш болуы мүмкін. Салық ставкалары мен салықтар есебінен алынған түсімдер арасындағы байланыс қолданылады: салық ставкалары 0-ден жоғарылаған кезде алынған түсімдер де артады. Алайда, салық ставкалары жоғары деңгейге жетіп, 100% жақындаған сайын, салық төлеуден жалтару артып, салық түсімдері азаяды. Кестеде. Корреляциялық матрица деп аталатын 1.7 телекоммуникациялық компаниялардың күнделікті кірістерінің 2012 жылғы маусым мен 2015 жылғы арасындағы арақатынасын көрсетеді. Кестеден Verizon (VZ) және ATT (T) жоғары корреляцияға ие екенін көруге болады. Level Three (LVLT) инфрақұрылымдық компаниясы ең төменгі корреляцияға ие. Бірліктердің диагоналіне назар аударыңыз (акцияның өзі 1 - ге тең) және ақпараттың артық болуы диаграммадан жоғары және төмен.

Кестеде. Корреляция матрицасы деп аталатын 1.7-суретте 2012 жылдың шілдесінен 2015 жылдың маусымына дейінгі телекоммуникациялық қорлардағы күнделікті кірістер арасындағы корреляция көрсетілген. Кесте

Verizon (VZ) және ATT (T) арасындағы ең жоғары корреляцияға ие екенін көрсетеді. Үшінші деңгейлі инфрақұрылымдық компания (LVLT) ең төмен корреляцияға ие. 1-дің диагоналіне (қордың өзімен байланысы 1-ге тең) және диагональдың үстінде және астындағы ақпараттың артықтығына назар аударыңыз.

1.7-кесте. Күнделікті телекоммуникациялық акцияларының қайтарымы арасындағы корреляция

	T	CTL	FTR	VZ	LVLT
T	1,000	0,475	0,328	0,678	0,279
CTL	0,475	1,000	0,420	0,417	0,287
FTR	0,328	0,420	1,000	0,287	0,260
VZ	0,678	0,417	0,287	1,000	0,242
LVLT	0,279	0,287	0,260	0,242	1,000

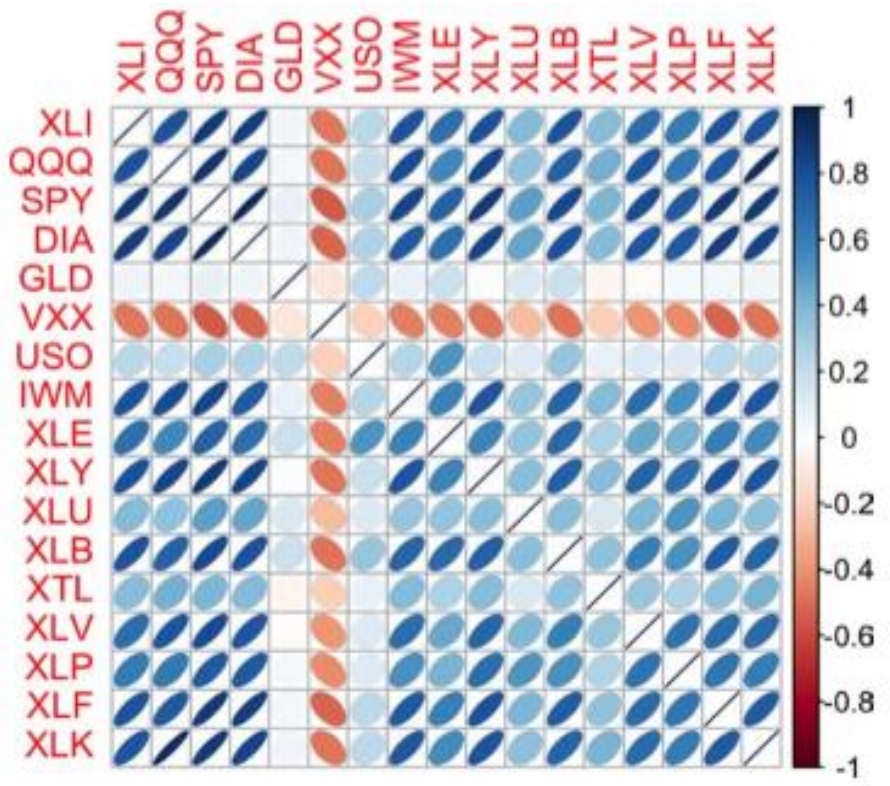
Кестеге ұқсас корреляция кестесі. 1.7, көптеген айнымалылар арасындағы байланысты көрсету үшін кеңінен қолданылады. Суретте. 1.6 ірі биржалық инвестициялық қорлардың (exchange traded funds, ETF) күнделікті кірістері арасындағы корреляцияны көрсетеді. Ол r-де пакеттің көмегімен оңай жасалады

corrplot:

```
etfs <- sp500_px[row.names(sp500_px)>"2012-07-01", sp500_sym[sp500_sym$sector=="etf", 'symbol']]

library(corrplot)

corrplot(cor(etfs), method = "ellipse")
```



1.6. –сурет. ETF кірістері арасындағы корреляция

S&P 500 (SPY) және Dow Jones (Dow Jones, DIA) индекстеріне арналған биржалық инвестициялық қорлар жоғары корреляцияға ие. Сол сияқты, негізінен технологиялық компаниялардан тұратын QQQ және XLK қорлары оң арақатынаста. Алтын бағасын (GLD), мұнай бағасын (USO) немесе нарықтың құбылмалылығын (VXX) бақылайтын қорғаныс биржалық инвестициялық қорлар басқа ETF - пен теріс байланысты болады. Эллипстің бағыты екі айнымалы оң (эллипс оңға бұрылады) немесе теріс (эллипс солға бұрылады) корреляцияны көрсетеді. Эллипстің толтырылуы мен ені байланыстың беріктігін көрсетеді: жұқа және қараңғы эллипстер күшті байланыстарға сәйкес келеді. Орташа және стандартты ауытқу сияқты, корреляция коэффициенті деректердің шығарылуына сезімтал. Бағдарламалық пакеттерде классикалық корреляция коэффициентіне балама нұсқалар ұсынылады. Мысалы, cor R-функциясында қысқартылған орташа мәнді есептеу үшін қолданылатынға ұқсас trim аргументі бар

Корреляцияның басқа бағалары статистикада корреляция коэффициенттерінің басқа түрлері бұрыннан ұсынылған, мысалы, спирманның ранг корреляциясы коэффициенті ρ (RO) немесе Кендаллдың ранг корреляциясының коэффициенті τ (тау).Бұл корреляция коэффициенттері мәліметтер жарасына, яғни жиынтықтағы бақылау

нөмірлеріне негізделген. Олар мәндермен емес, дәрежелермен жұмыс істейтіндіктен, бұл бағалар шығарындыларға төзімді және белгілі бір сызықтық емес түрлерін жеңе алады. Алайда, барлау талдауына арналған деректер талдаушылары әдетте Пирсонның корреляция коэффициентін және оның қатал баламаларын ұстануы мүмкін. Дәрежелік бағалау негізінен кішігірім мәліметтер жиынтығы мен статистикалық гипотезаларды белгілі бір тексерулер кезінде тартымды болады

Екі немесе одан да көп айнымалыларды зерттеу

Бізге бұрыннан таныс бағалау құралдары, мысалы, орташа бағалау және дисперсия, бір уақытта бір айнымалыдан есептеледі (бір өлшемді талдау). Корреляциялық талдау (бөлімді қараңыз. Бұрын осы тарауда" Корреляция") — екі айнымалыны салыстыратын маңызды әдіс (екі өлшемді талдау). Бұл бөлімде біз қосымша бағаларға, графиктерге және екіден көп айнымалыға (көп өлшемді талдау) жүгінеміз.

Негізгі терминдер

Конъюгация кестелері (contingency tables) екі немесе одан да көп категориялық айнымалылар санының қысқаша мазмұны.

Алтыбұрышты торлы графиктер (hexagonal binning) - екі сандық айнымалылардың графигі, онда жазбалар шес-тиугольдарда топтастырылған.

Контурлық графиктер (contour plot) - топографиялық карта түріндегі екі сандық айнымалылардың тығыздығын көрсететін График.

Скрипка графикасы (violin plots) қорап диаграммасына ұқсас, бірақ raft ұпайларын көрсететін График - жаңалықтар.

Екі өлшемді талдау бір өлшемді талдау сияқты жиынтық статистиканы есептеумен және оларды визуализациямен байланысты. Екі өлшемді немесе көп өлшемді талдаудың тиісті түрі деректердің сипатына байланысты: сандық немесе категориялық.

Шашырау диаграммасының алтыбұрышты торы мен контурлары (сандық және сандық деректерді көрсету)

Деректер мәндерінің салыстырмалы түрде аз саны болған кезде әдемі болады. Суреттегі акциялардың кірістілік кестесі. 1.7 шамамен 750 нүктені қамтиды. Жүздеген мың және миллиондаған жазбалары бар мәліметтер жиынтығы үшін шашырау диаграммасы тым тығыз болады, сондықтан

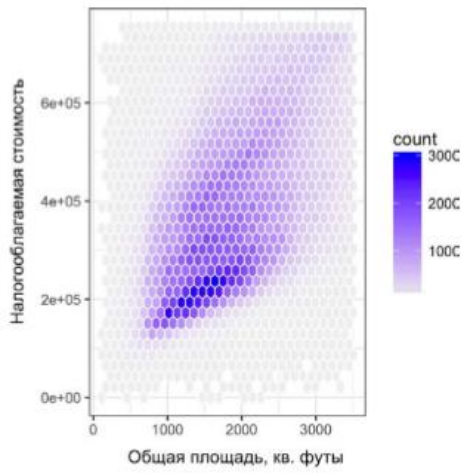
байланысты бейнелеудің басқа әдісі қажет. Мысал ретінде біз Вашингтон штатының Кинг округіндегі тұрғын үйдің салық салынатын құнын қамтитын kc_atx мәліметтер жиынтығын қарастырамыз. Деректердің негізгі бөлігіне назар аудару үшін біз өте қымбат тұрғын үйді, сондай-ақ функцияның көмегімен өте кішкентай немесе өте үлкен тұрғын үйді алып тастаймыз.

subset:

```
kc_tax0 <- subset(kc_tax, TaxAssessedValue < 750000 &
                  SqFtTotLiving>100 & SqFtTotLiving<3500)
nrow(kc_tax0)
```

[1] 432733

Суретте. 1.8 шаршы футтың жалпы ауданы мен Кинг округіндегі тұрғын үйдің салық салынатын құны арасындағы байланысты көрсету үшін алтыбұрышты тор кестесі көрсетілген. Монолитті қара бұлт ретінде пайда болатын нүктелерді көрсетудің орнына, біз жазбаларды алтыбұрышты сегменттерге және ото-бразильдерге топтастырдық алтыбұрыштар белгілі бір сегменттегі жазбалардың санын көрсететін түсті. Бұл диаграммада шаршы фут пен тұрғын үйдің салық салынатын құны арасындағы оң байланыс айқын көрінеді. Графиктің қызықты мәні - бұл бас бұлттың үстіндегі екінші бұлттың көлеңкесі, шаршы футпен бірдей ауданы бар, негізгі бұлтта орналасқан, бірақ салық салынатын құны жоғары үйлерге нұсқайды.



Сурет. 1.8 Хадли Уикхэм (Hadley Wickham) [ggplot2] әзірлеген қуатты ggplot2 R-пакетімен жасалды. Ggplot2 пакеті - бұл деректерді кеңейтілген визуалды талдауға арналған көптеген жаңа бағдарламалық кітапханалардың

бірі (бөлімді қараңыз. "Көптеген айнымалыларды визуализациялау" осы тарауда келтірілген).

```
ggplot(kc_tax0, (aes(x=SqFtTotLiving, y=TaxAssessedValue))) +

stat_binhex(colour="white") +

theme_bw() +

scale_fill_gradient(low="white", high="black") +

labs(x="Общая площадь, кв. футы", y="Налогооблагаемая стоимость")
```

Суретте. 1.9 екі сандық айнымалылар арасындағы байланысты визуализациялау үшін шашырау диаграммасына салынған контурлар қолданылады. Контурлар екі айнымалыға сәйкес келетін топографиялық картаға ұқсас; контурдың әр жолағы "шыңға"жақындаған сайын өсіп, нүктелердің тиісті тығыздығын білдіреді. Бұл график суреттегі графикпен бірдей. 1.8: негізгі шыңнан екінші "солтүстікке" шыңы бар. Бұл график сонымен қатар ggplot2 пакеті мен кіріктірілген функцияның көмегімен жасалды

geom_density2d.

```
ggplot(kc_tax0, aes(SqFtTotLiving, TaxAssessedValue)) +

theme_bw() +

geom_point( alpha=0.1) +

geom_density2d(colour="white") +

labs(x="Общая площадь, кв. футы", y="Налогооблагаемая стоимость")
```

Диаграммалардың басқа түрлері екі сандық айнымалылар арасындағы байланысты, соның ішінде жылу карталарын көрсету үшін қолданылады. Жылу карталары, торлы графиктер және контурлық графиктер - барлығы екі өлшемді тығыздық туралы визуалды түсінік береді. Бұл тұрғыда олар гистограммалар мен тығыздық графиктерінің табиғи аналогтары болып табылады.

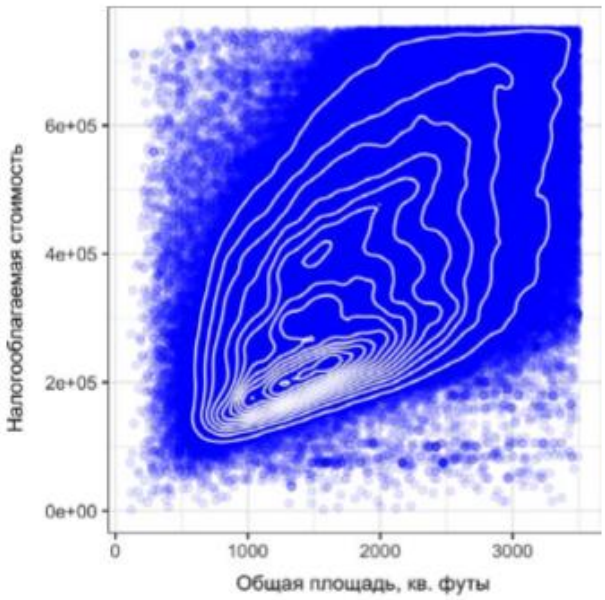


Рис. 1.9. Контурный график для налогооблагаемой стоимости против общей площади в квадратных футах