\#
Agriculture
Industry
Climate
Literacy (%)
Infant mortality (per 1000 births)
GDP ($ per capita)
Net migration

و تمامی آن هایی که بعد از اعمال تغییرات برای شناسایی در ستونشان NAظاهر می شود

| egion | Population | Area..sq..mi.. | Pop..Density..per.sq..mi.. | Coastline..coast.area.ratio. | Net.migration | Infant.mortality..per.1000.births. | GDP....per.capita. | Literacy.... | Pl |
|---|---|---|---|---|---|---|---|---|---|
| SIA (EX. NEAR EAST) | 31056997 | 647500 | 48,0 | 0,00 | 23,06 | 163,07 | 700 | 36,0 | |
| ASTERN EUROPE | 3581655 | 28748 | 124,6 | 1,26 | -4,93 | 21,52 | 4500 | 86,5 | |
| ORTHERN AFRICA | 32930091 | 2381740 | 13,8 | 0,04 | -0,39 | 31 | 6000 | 70,0 | |
| CEANIA | 57794 | 199 | 290,4 | 58,29 | -20,71 | 9,27 | 8000 | 97,0 | |
| /ESTERN EUROPE | 71201 | 468 | 152,1 | 0,00 | 6,6 | 4,05 | 19000 | 100,0 | |
| UB-SAHARAN AFRICA | 12127071 | 1246700 | 9,7 | 0,13 | 0 | 191,19 | 1900 | 42,0 | |
| ATIN AMER. & CARIB | 13477 | 102 | 132,1 | 59,80 | 10,76 | 21,03 | 8600 | 95,0 | |
| ATIN AMER. & CARIB | 69108 | 443 | 156,0 | 34,54 | -6,15 | 19,46 | 11000 | 89,0 | |
| ATIN AMER. & CARIB | 39921833 | 2766890 | 14,4 | 0,18 | 0,61 | 15,18 | 11200 | 97,1 | |
| .W. OF IND. STATES | 2976372 | 29800 | 99,9 | 0,00 | -6,47 | 23,28 | 3500 | 98,6 | |
| ATIN AMER. & CARIB | 71891 | 193 | 372,5 | 35,49 | 0 | 5,89 | 28000 | 97,0 | |
| CEANIA | 20264082 | 7686850 | 2,6 | 0,34 | 3,98 | 4,69 | 29000 | 100,0 | |
| /ESTERN EUROPE | 8192880 | 83870 | 97,7 | 0,00 | 2 | 4,66 | 30000 | 98,0 | |
| .W. OF IND. STATES | 7961619 | 86600 | 91,9 | 0,00 | -4,9 | 81,74 | 3400 | 97,0 | |
| ATIN AMER. & CARIB | 303770 | 13940 | 21,8 | 25,41 | -2,2 | 25,21 | 16700 | 95,6 | |
| EAR EAST | 698585 | 665 | 1050,5 | 24,21 | 1,05 | 17,27 | 16900 | 89,1 | |
| SIA (EX. NEAR EAST) | 147365352 | 144000 | 1023,4 | 0,40 | -0,71 | 62,6 | 1900 | 43,1 | |
| ATIN AMER. & CARIB | 270012 | 431 | 640.5 | 22.51 | 0.31 | 12.5 | 15700 | 97.4 | |

| | Country | Region | Population | Area (sq. mi.) | Pop. Density (per sq. mi.) | Coastline (coast/area ratio) | Net migration | Infant mortality (per 1000 births) | GDP ($ per capita) | Literacy (%) | Phones (per 1000) | Arable (%) | Crops (%) | Other (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | Belize | LATIN AMER. & CARIB | 287730 | 22966 | 12,5 | 1,66 | 0 | 2569 | 4900 | 941 | 115,7 | 2,85 | 1,71 | 9544 |
| 22 | Benin | SUB-SAHARAN AFRICA | 7862944 | 112620 | 69,8 | 0,11 | 0 | 85 | 1100 | 409 | 9,7 | 18,08 | 2,4 | 7952 |
| 23 | Bermuda | NORTHERN AMERICA | 65773 | 53 | 1241,0 | 194,34 | 2,49 | 853 | 36000 | 980 | 851,4 | 20 | 0 | 80 |
| 24 | Bhutan | ASIA (EX. NEAR EAST) | 2279723 | 47000 | 48,5 | 0,00 | 0 | 10044 | 1300 | 422 | 14,3 | 3,09 | 0,43 | 9648 |
| 25 | Bolivia | LATIN AMER. & CARIB | 8989046 | 1098580 | 8,2 | 0,00 | -1,32 | 5311 | 2400 | 872 | 71,9 | 2,67 | 0,19 | 9714 |
| 26 | Bosnia & Herzegovina | EASTERN EUROPE | 4498976 | 51129 | 88,0 | 0,04 | 0,31 | 2105 | 6100 | NA | 215,4 | 13,6 | 2,96 | 8344 |
| 27 | Botswana | SUB-SAHARAN AFRICA | 1639833 | 600370 | 2,7 | 0,00 | 0 | 5458 | 9000 | 798 | 80,5 | 0,65 | 0,01 | 9934 |
| 28 | Brazil | LATIN AMER. & CARIB | 188078227 | 8511965 | 22,1 | 0,09 | -0,03 | 2961 | 7600 | 864 | 225,3 | 6,96 | 0,9 | 9215 |
| 29 | British Virgin Is. | LATIN AMER. & CARIB | 23098 | 153 | 151,0 | 52,29 | 10,01 | 1805 | 16000 | 978 | 506,5 | 20 | 6,67 | 7333 |
| 30 | Brunei | ASIA (EX. NEAR EAST) | 379444 | 5770 | 65,8 | 2,79 | 3,59 | 1261 | 18600 | 939 | 237,2 | 0,57 | 0,76 | 9867 |
| 31 | Bulgaria | EASTERN EUROPE | 7385367 | 110910 | 66,6 | 0,32 | -4,58 | 2055 | 7600 | 986 | 336,3 | 40,02 | 1,92 | 5806 |
| 32 | Burkina Faso | SUB-SAHARAN AFRICA | 13902972 | 274200 | 50,7 | 0,00 | 0 | 9757 | 1100 | 266 | 7,0 | 14,43 | 0,19 | 8538 |
| 33 | Burma | ASIA (EX. NEAR EAST) | 47382633 | 678500 | 69,8 | 0,28 | -1,8 | 6724 | 1800 | 853 | 10,1 | 15,19 | 0,97 | 8384 |
| 34 | Burundi | SUB-SAHARAN AFRICA | 8090068 | 27830 | 290,7 | 0,00 | -0,06 | 6929 | 600 | 516 | 3,4 | 35,05 | 14,02 | 5093 |
| 35 | Cambodia | ASIA (EX. NEAR EAST) | 13881427 | 181040 | 76,7 | 0,24 | 0 | 7148 | 1900 | 694 | 2,6 | 20,96 | 0,61 | 7843 |
| 36 | Cameroon | SUB-SAHARAN AFRICA | 17340702 | 475440 | 36,5 | 0,08 | 0 | 6826 | 1800 | 790 | 5,7 | 12,81 | 2,58 | 8461 |

برای متغییر هایی که عددی هستند دستور is.na به درستی جواب میدهد و با دستور sum می توان تعداد داده های از دست رفته را بدست آورد ولی در فرمت های غیر عددی اسپیس یا " به عنوان کاراکتر برای R قابل قبول است درصورتی که می دانیم داده نداریم.

لذا برای حل این مشکل باید از دستوری دیگر یافت تا علاوه بر شناسایی کاراکتر ها و مقادیر نامطلوب آن ها را بتوان با نوع داده قابل تشخیص NA تغییر داد و سپس تعداد خانه هایی که متغیر NA دارند را شمارد.

```
na_strings <- c("NA", "N A", "N / A", "N/A", "N/ A", "Not Available", "NOt available","","  ")

dataofworlds <- readr::read_csv("countries of the world.csv", na = na_strings)
```

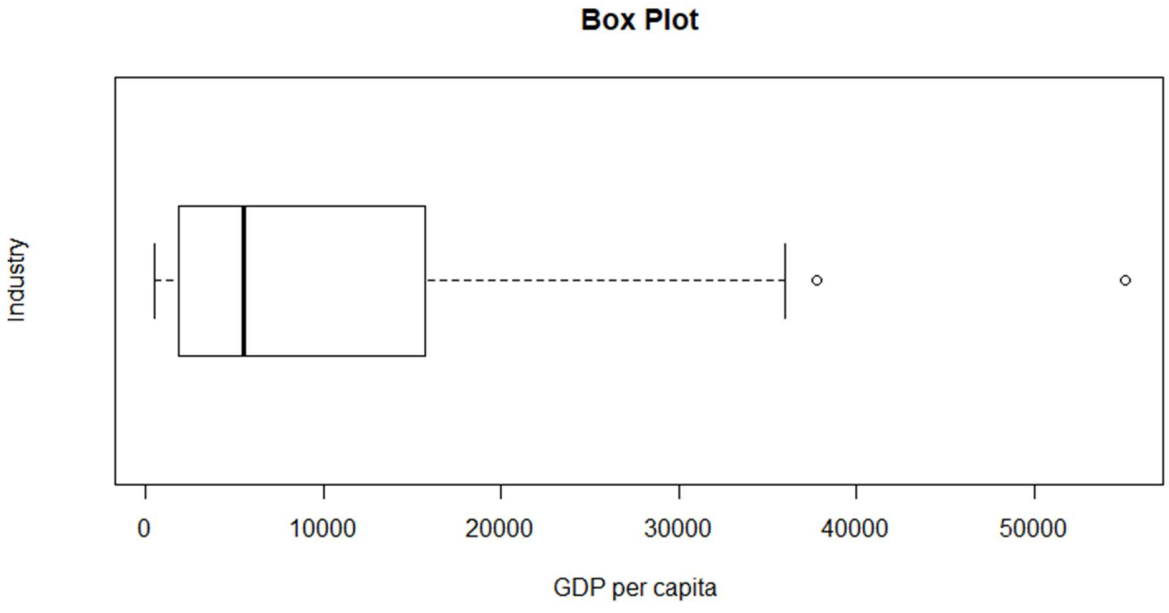
کد شناسایی داده های از دست داده شده .

| | |
|---|---|
| missed.Area | 0L |
| missed.climate | 22L |
| missed.Coastline | 0L |
| missed.Country | 0L |
| missed.migration | 230L |
| missed.mortality | 230L |
| missed.Pop.Density | 0L |
| missed.Population | 0L |
| missed.Service | 15L |
| na_strings | chr [1:9] "NA" "N A" "N / A" "N/A" "N/ A" "Not Availa… |
| onfidenc.interval | num [1:2] 8549 11178 |
| ratio.Area | 0 |
| ratio.climate | 0.0969162995594714 |
| ratio.Coastline | 0 |
| ratio.Country | 0 |
| ratio.Pop.Density | 0 |
| ratio.Population | 0 |
| ratio.Service | 0.066079295154185 |

برای بقیه مقادیر نیز به همین شیوه محاسبه می شوند (تعداد متغیر ها بسیار زیاد بود و فرایند محاسبه خسته کننده)

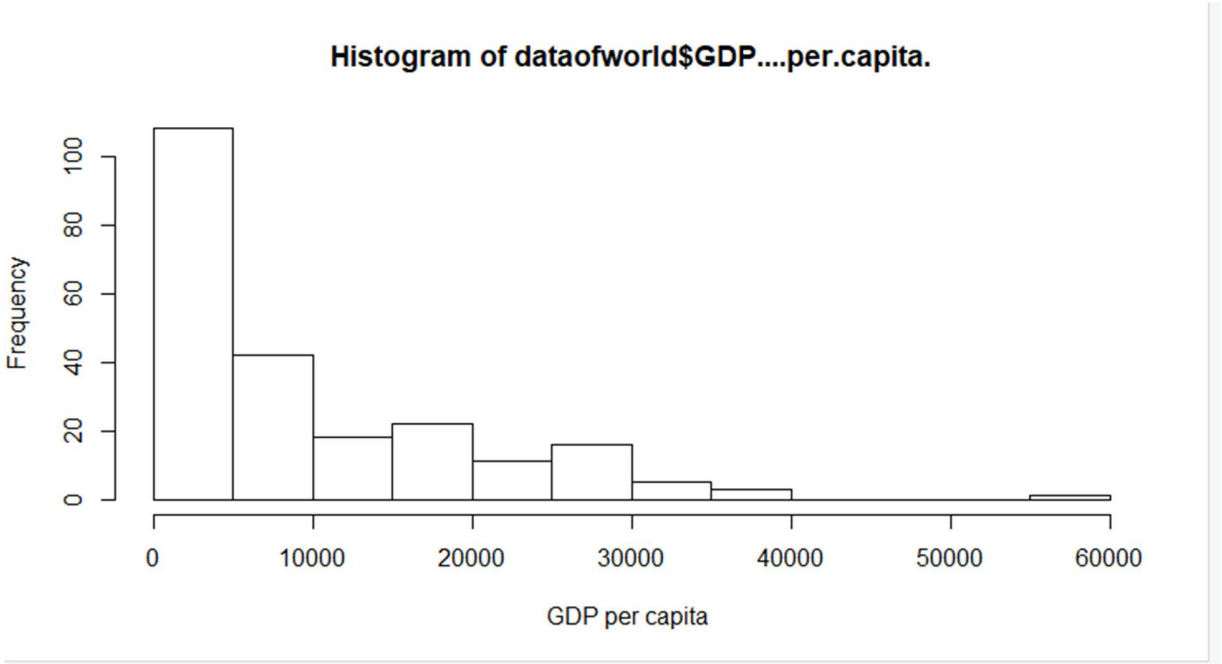
۲#

باکس پلات

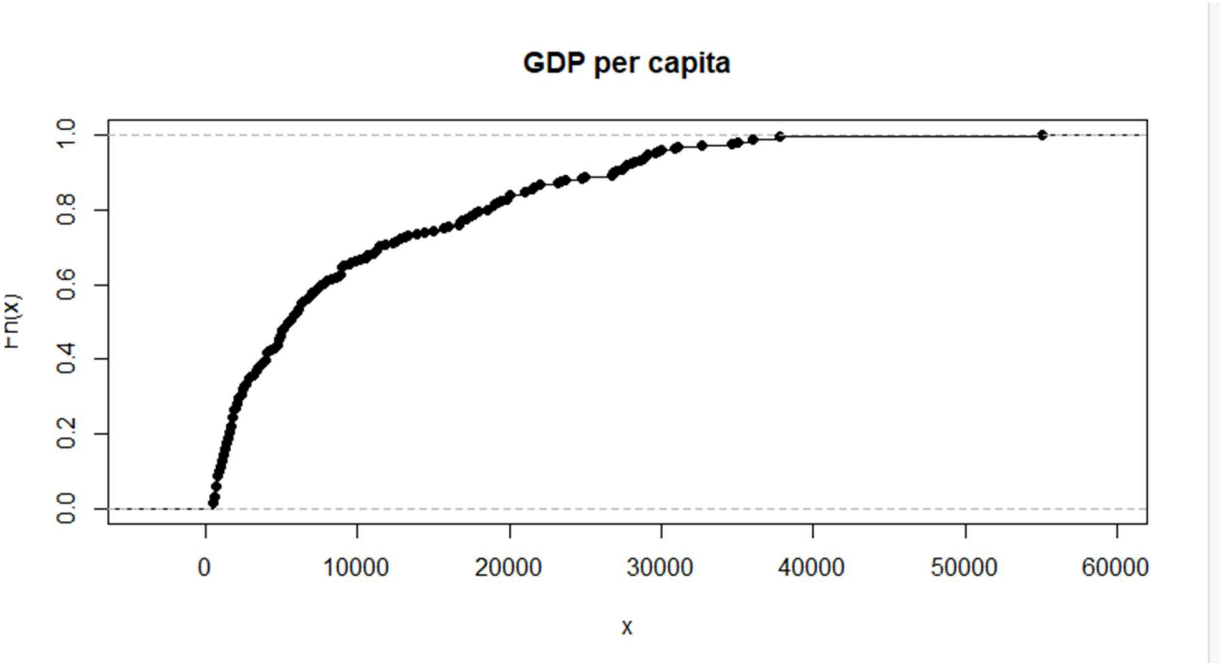


مقادیر چهارک ها

```
quantile(dataofworld$GDP....per.capita.,na.rm=TRUE)
  0%   25%   50%   75%  100%
 500  1900  5550 15700 55100
```

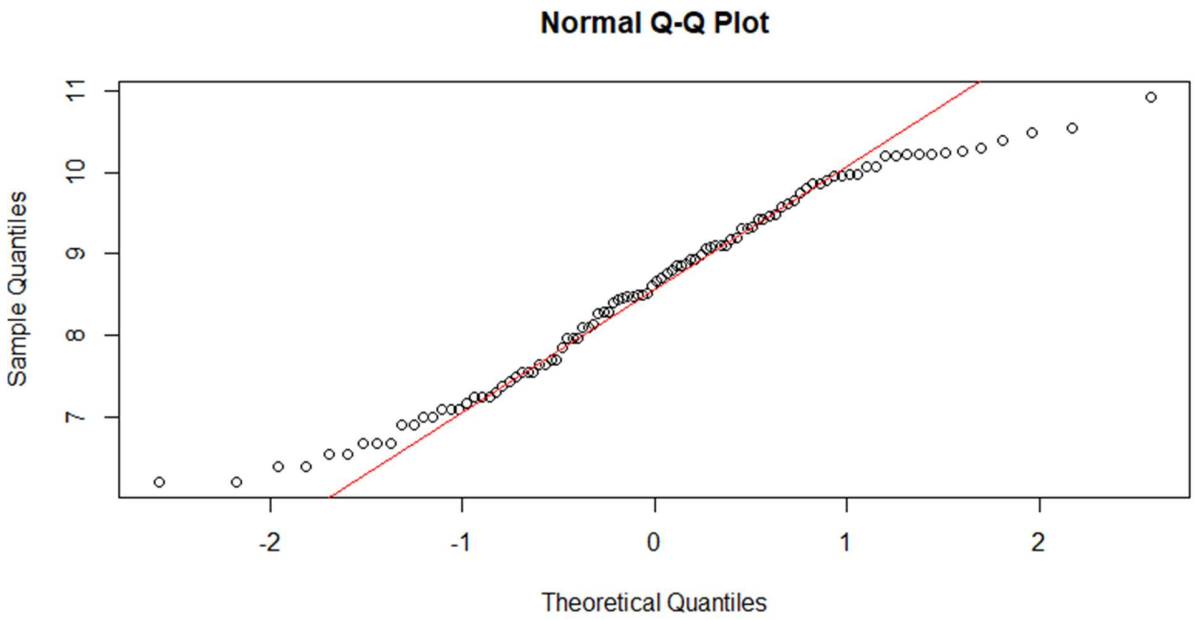



نمودار توزیع تجمعی



۳#

پس از محسابه مقادیر خواسته شده در سوال (نمونه گیری و پیدا کردن میانگین ، واریانس و انحراف معیار) توزیع را با نمودار نرمال مقایسه می کنیم برای این کار از نمودار Q-Q کمک میگیریم ومی دانیم هر چه داده ها به خط رسم شده نزدیک تر باشند داده ها توزیع نزدیک تری به نرمال دارند.اما پس از رسم رابطهی خطی خوبی مشاهده نمی شود و برای همین به جای متغییر انتخاب شده لگاریتم تابع را قرار می دهیم و تابع تا حد خوبی نرمال می شود.(برای داده " GDP per capita")




با توجه به تئوری زد اسکور را محاسبه و بر اساس آن بازه اطمینان را می یابیم.

```
> confidenc.interval=c(meanGDP-ci,meanGDP+ci)
> confidenc.interval
[1]  8549.332 11177.941
> 
```

با استفاده از دستور تی تست تست دو طرفه و یک طرفه را با فرض های صفر اولیه پیدا می کنیم

```
        One Sample t-test

data:  sampleData
t = 3.2071, df = 99, p-value = 0.001806
alternative hypothesis: true mean is not equal to 6500
95 percent confidence interval:
  7792.641 11987.359
sample estimates:
mean of x
     9890

> |
```

برای حالت دوطرفه مشاهده می کنیم فرض صفر به خوبی رد شده است و مقدار p-value کمتر از ۲/۵ درصد است و احتمال این که میانگین ۶۵۰۰ باشد بسیار کم است.

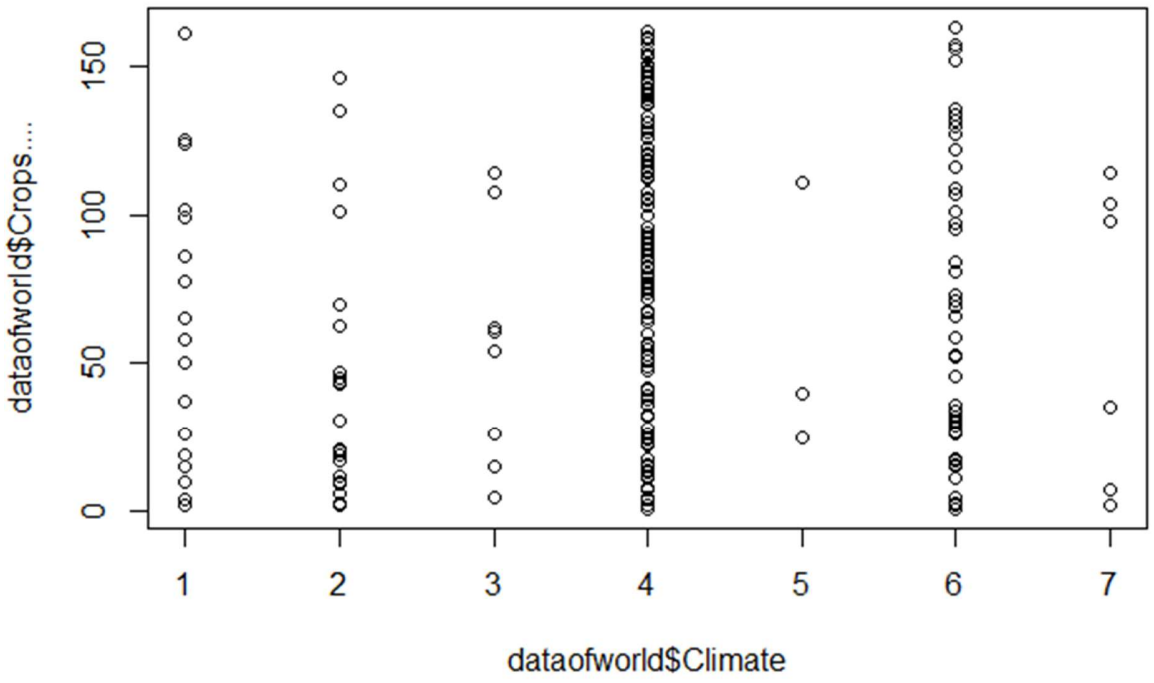
```
        One Sample t-test

data:  sampleData
t = 0.028382, df = 99, p-value = 0.4887
alternative hypothesis: true mean is greater than 9860
95 percent confidence interval:
 8134.931      Inf
sample estimates:
mean of x
     9890
```

برای حالت یک طرفه مشاهده می کنیم مقدار p-value کوچک تر از ۵درصد است و فرض صفر رد شده و میانگین بزرگ تر از مقداری است که در فرض صفر در نظر گرفته شده بود .

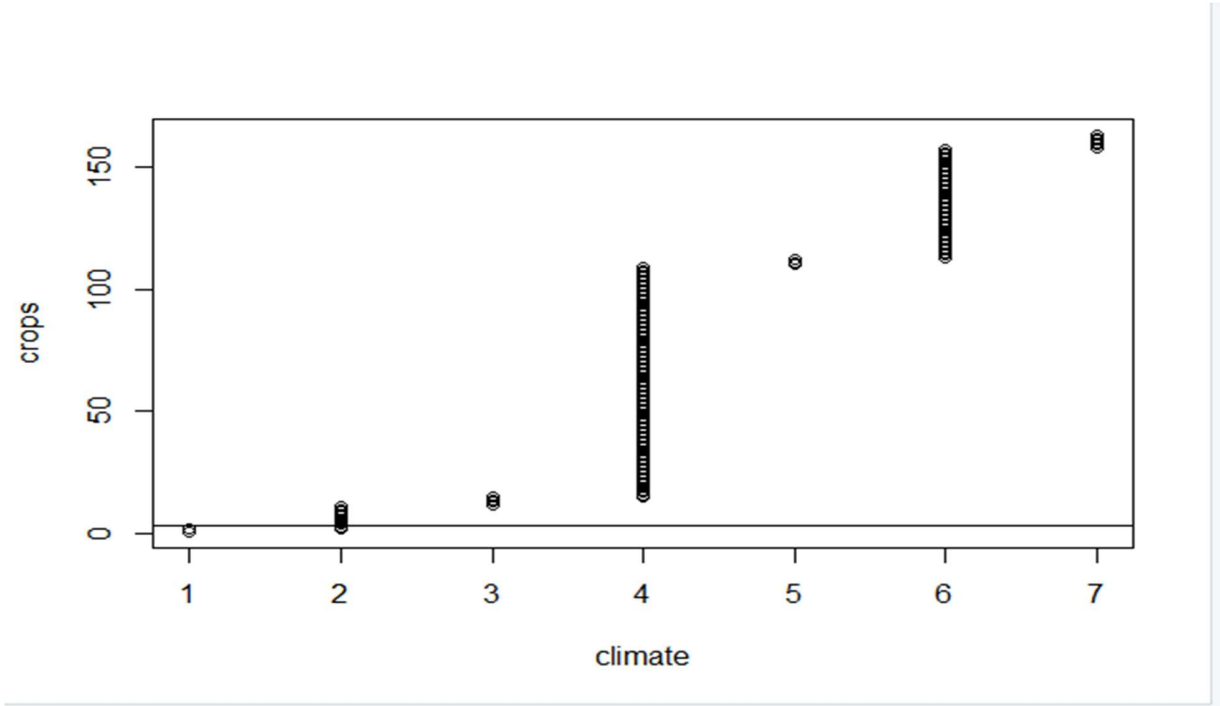
(مقدار p-value را با آلفا در یک طرفه مقایسه می کنیم اگر مقدار p-value کوچک تر باشد یعنی به درستی توانستیم فرض اولیه ای که داشتیم را رد کنیم ولی اگر بیشتر بود یعنی موفق نبودیم که فرض صفر را رد کنیم . )

#4

با امتحان چند متغییر به متغییر هایی به کورریلشن خوبی نرسیدم و این داده ها تقریبا ارتباطی با هم نداشته اند (چون بسیار به صفر نزدیک هست در هر دو روش محاسبه )(تشخیص ارتباط داده ها در این نمونه بسیار سخت بود و مانند قد وزن نمونه دیگر به راحتی قابل تشخیص نیست)




scatter plot




Q-Q plot

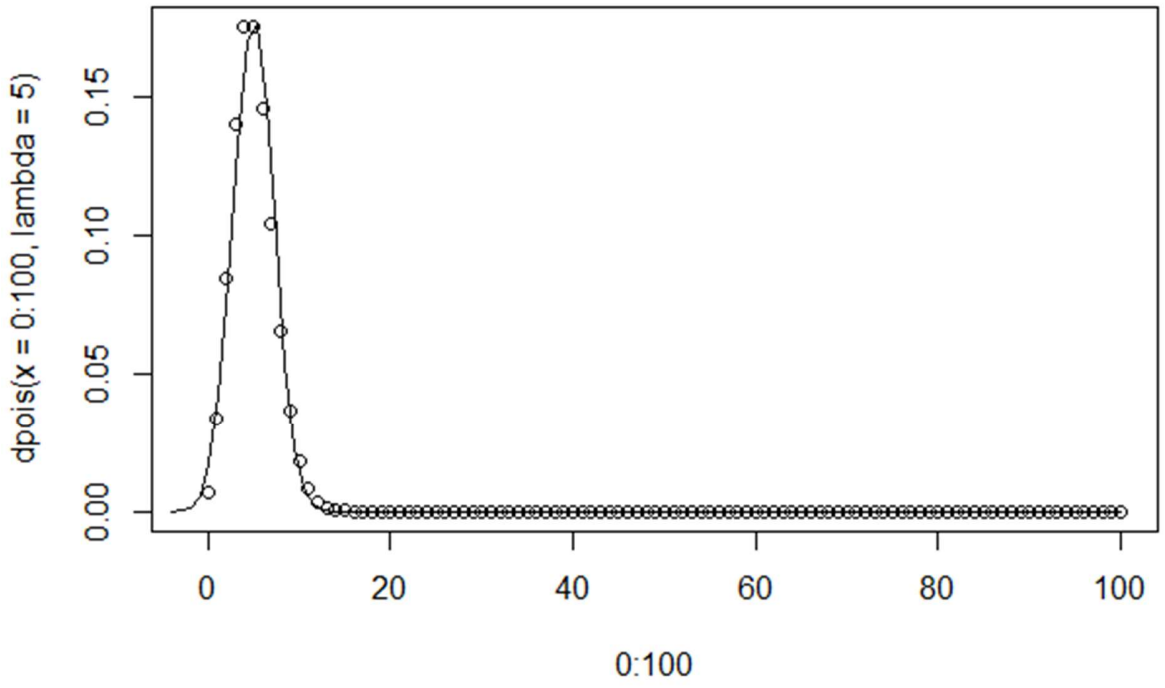
این دو متغیر تقریبا از هم مستقل هستند و تنها دلیل فشردگی اطلاعات محدود بودن حالت های تعریف شده برای آب وهوا است و خط ترسیم شده به خوبی عدم ارتباط این دو متغییر را نشان می دهد ( دیتا عمود بر خط است )

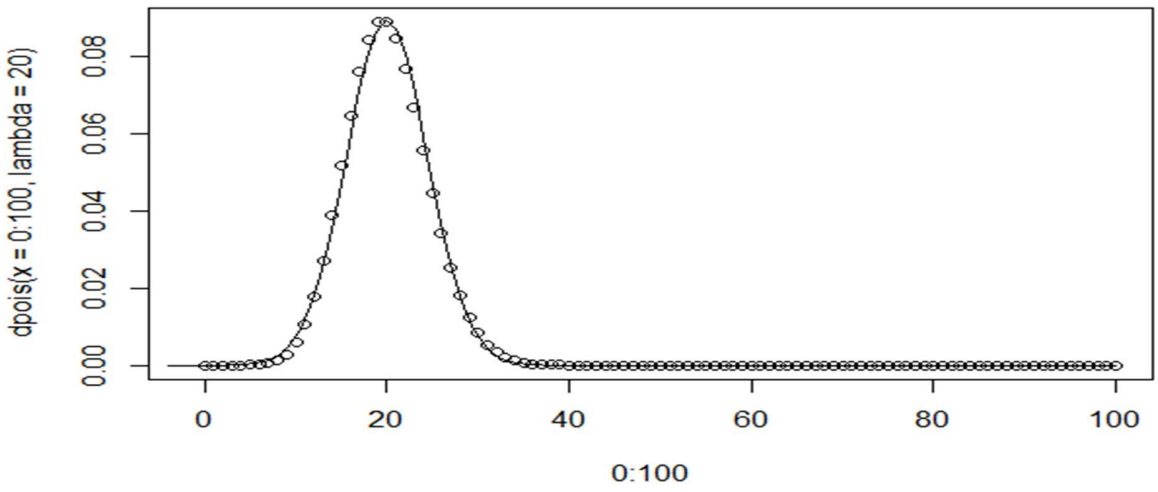
روش pearson برای سنجش میزان خطی بودن رابطه بین دو متغیر عددی استفاده می شود و spearman برای سنجش میزان یکنوایی که در هر صورت نشان میدهد داده ها روابط معنا دای با یک دیگر ندارند .

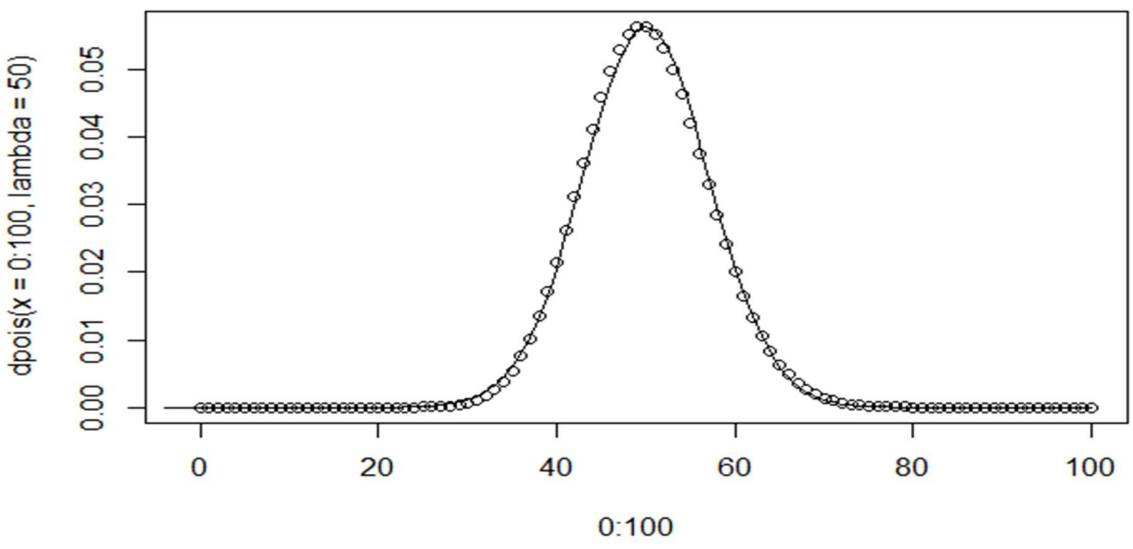
```
> cor(dataofworld$Climate,dataofworld$Crops....,use="complete.obs",method="spearman")
[1] 0.1436332
> cor(dataofworld$Climate,dataofworld$Crops....,use="complete.obs",method="pearson")
[1] 0.1429595
```

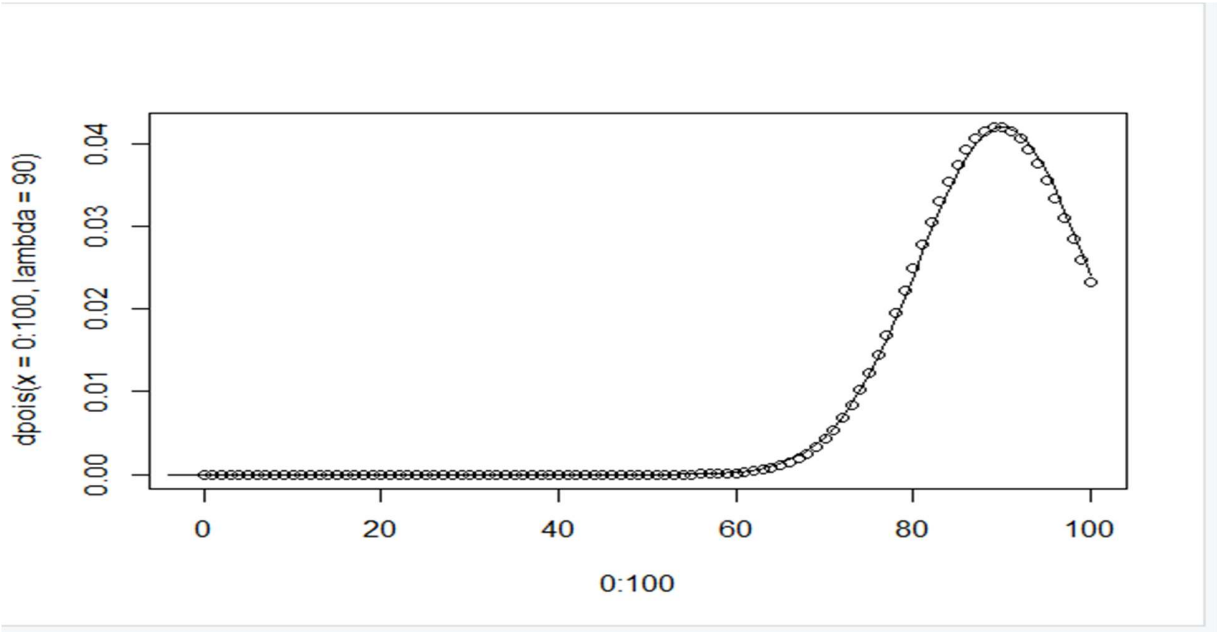
#5

نمودار توزیع بواسون برای مقادیر مختلف لاندا رسم شده است و هر با توجه به برابری لاندا با میانگین (در کد) قله با تغییر میانگین جابه جا می شپد و تنها نکته جالب این است که با افزایش لاندا نمودارقله کوتاه تر و عریض تری خواهد داشت .

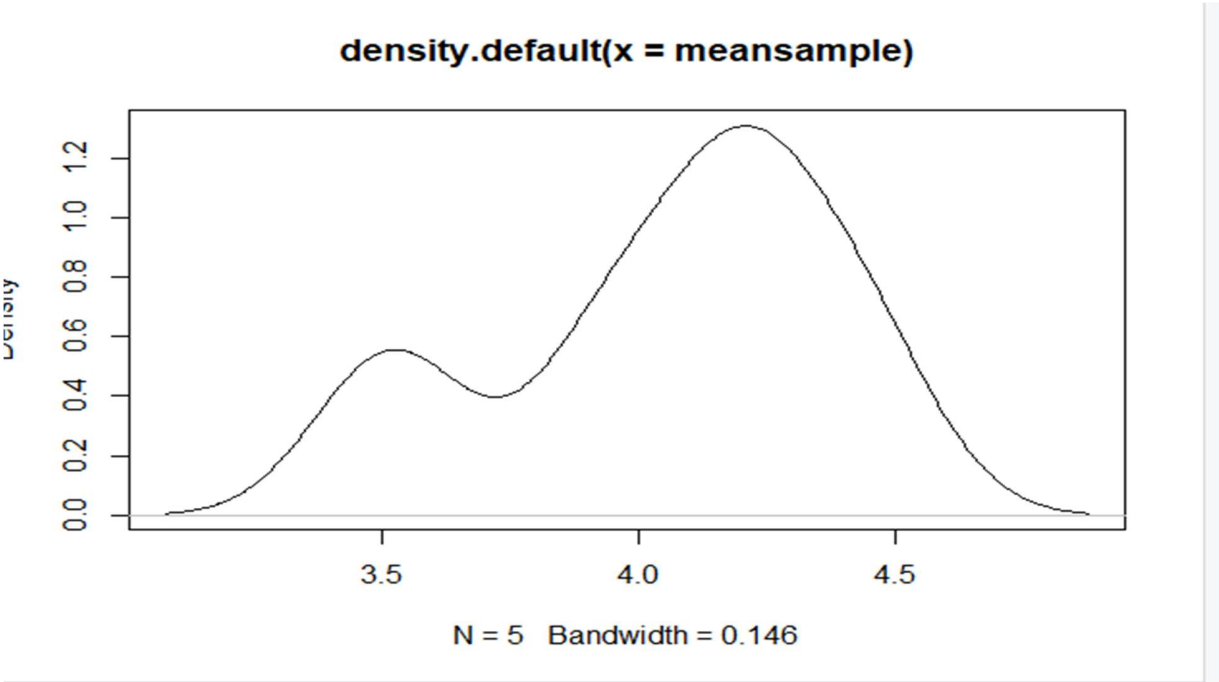


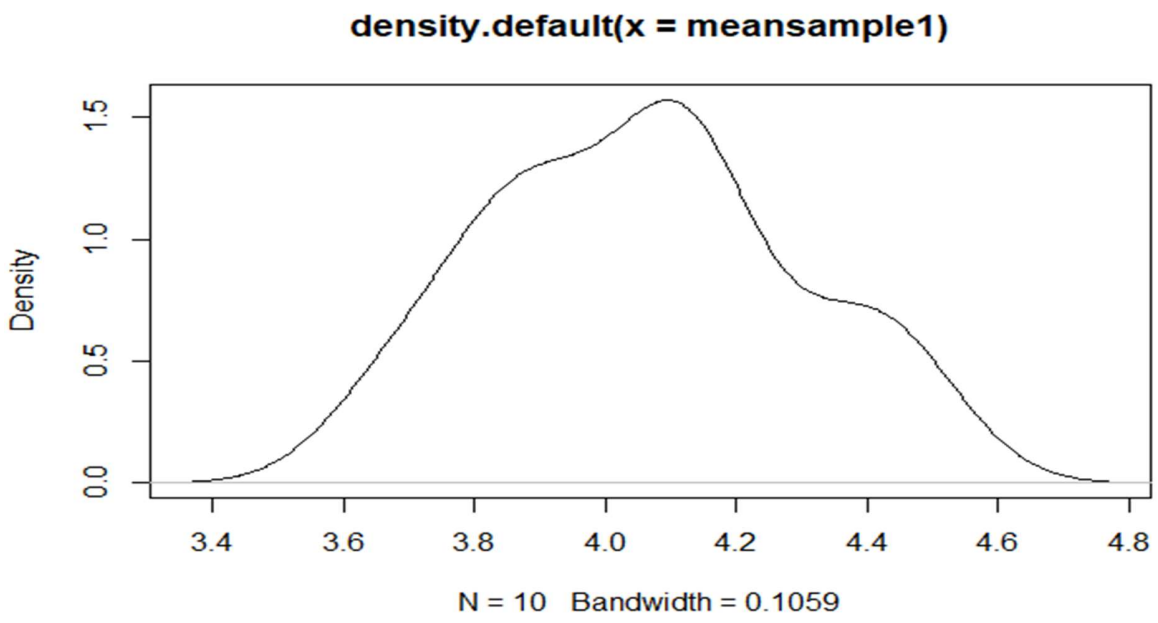

dpois(x = 0:100, lambda = 20)
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
0:100

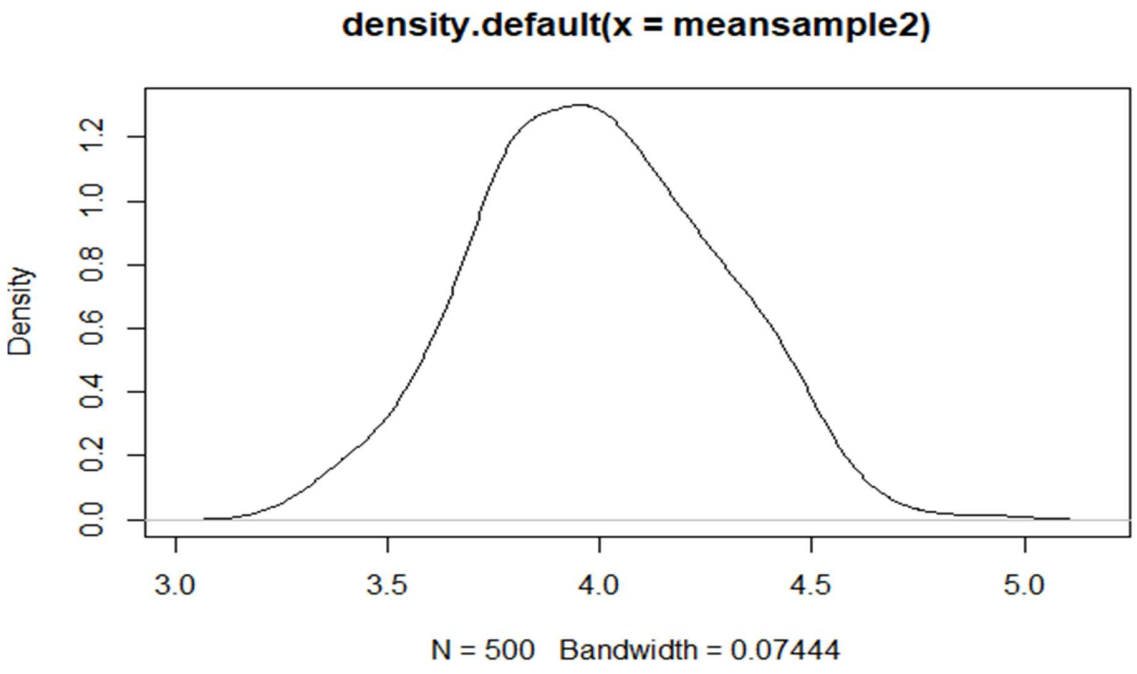

dpois(x = 0:100, lambda = 50)
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
0:100

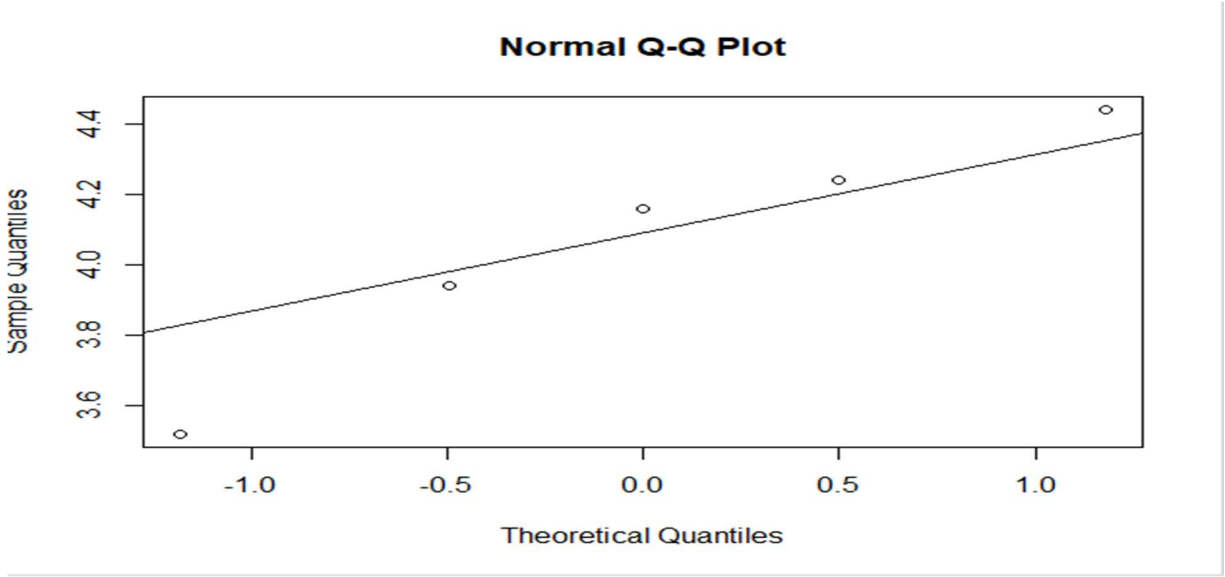


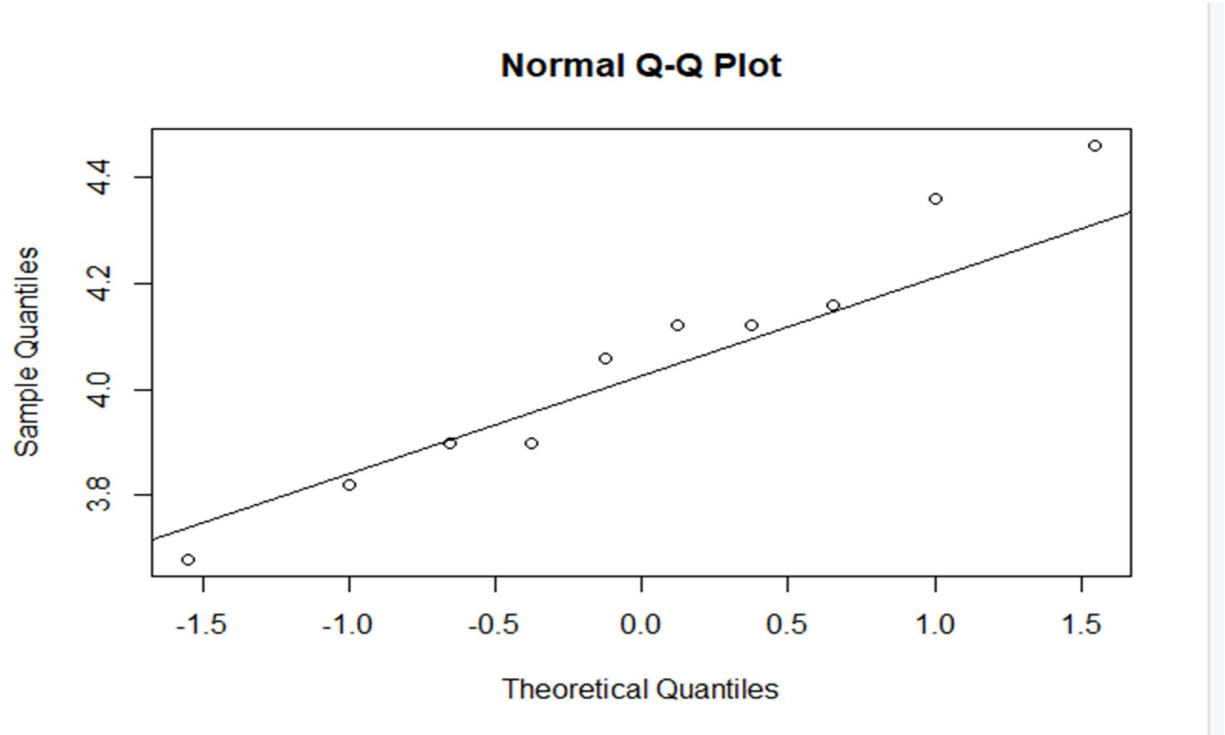

توزیع نمونه های توزیع های بالا

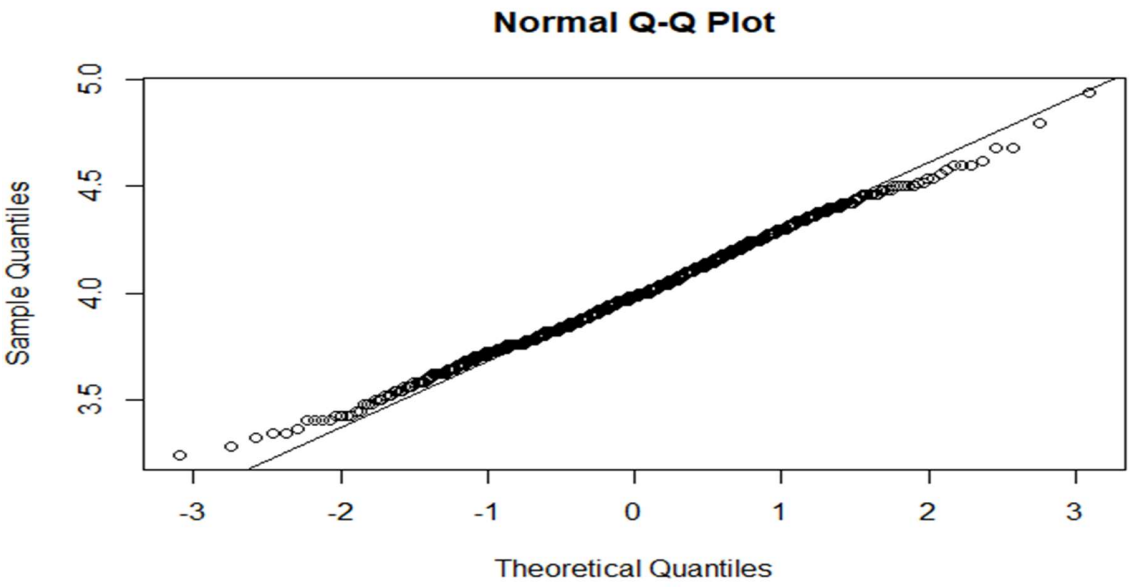


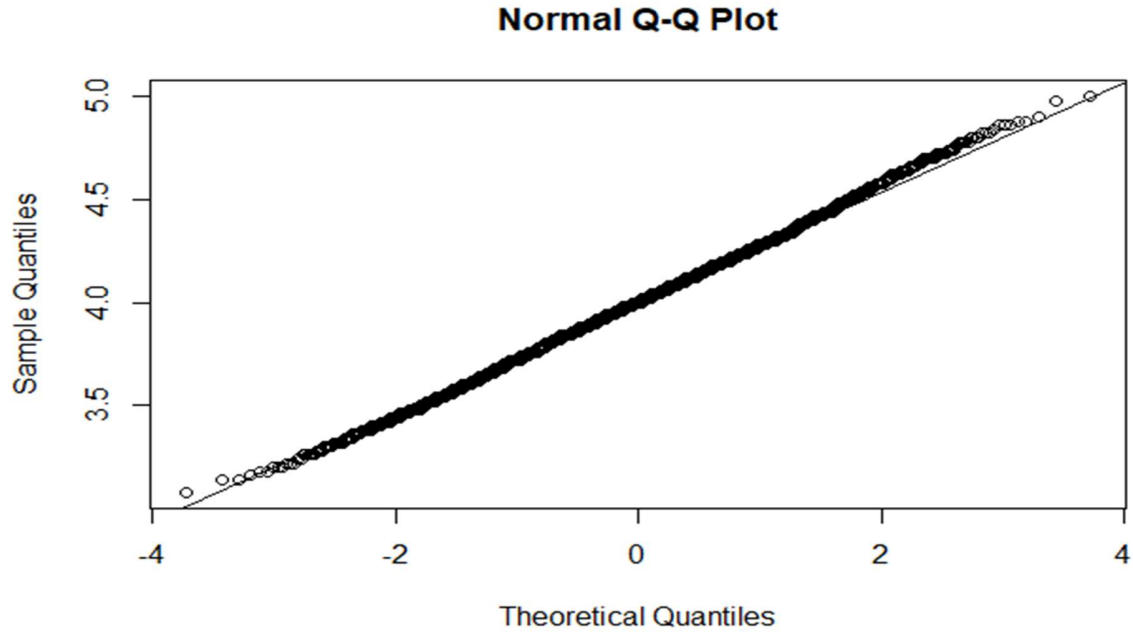

density.default(x = meansample1)
Density
0.0
0.5
1.0
1.5
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
N = 10 Bandwidth = 0.1059


density.default(x = meansample2)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
3.0
3.5
4.0
4.5
5.0
N = 500 Bandwidth = 0.07444


Normal Q-Q Plot
Sample Quantiles
3.6
3.8
4.0
4.2
4.4
-1.0
-0.5
0.0
0.5
1.0
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
3.8
4.0
4.2
4.4
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles




با توجه به قضیه حد مرکزی هر چه تعداد نمونه ها افزایش یابد (درصورت آی آی دی بودن نمونه ها) توزیه بیشتر به توزیع نرمال شبیه میشود بنابراین در نمودار Q-Q به خط ترسیم شده نزدیک و نزدیک تر می شود .

(با ثابت گرفتن لاندا و میل دادن تعداد نمونه ها به بینهایت به توزیع نرمال رسیدیم)

باز هم نمونه های کور ریلیشن کمی دارند ولی با همین داده ها خط رگرسیون برای آب و هوا و محصولات رسم شده اسن

فرمول خط رگرسیون

(بدیهتا منطقی تر بود نمودار برعکس رسم شود چون آب و هوا فاکتور است و مقدار های گسسته ای دارد ویافتن مقدار فاکتور براساس تعداد محصولات کمی غیرعقلانی است !!)

Climate= 0.003944 *crops+ 3.6762108