گزارش پروژه اول – مصور سازي داده هاي مربوط به خريد و فروش مسكن

استان خوزستان

محمد جواد ولي پور

در اين پروژه قصد داريم داده هاي مربوط به خريد و فروش مسكن در استان خوزستان را مصورسازي كنيم.

اطلاعات مربوط به كل استان ها از۴۳۱۳۳داده و۱۵متغير تشكيل شده است. كه۷۶۸عدد از اين داده ها اطلاعات مربوط به استان خوزستان است.

ابتدا به طور مختصر به معرفي ستون هاي داده ميپردازيم.

.۱كد قرارداد : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ كد قرارداد          : chr [1:43133] "19136381" "19140513" "19192568" "19202902" ...
```

اين ستون اطلاعات خاصي در اختيار ما قرار نميدهد.

.۲نوع قرارداد : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ نوع قرارداد         : chr [1:43133] "مبايعه نامه" "مبايعه نامه" "مبايعه نامه" "مبايعه نامه" ...
```

مقادير تمام داده هاي مربوط به اين ستون برابر با" مبايعه نامه "است و اطلاعات خاصي در اختيار ما قرار نميدهد.

.۳استان : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ استان               : chr [1:43133] "زنجان" "زنجان" "زنجان" "زنجان" ...
```

به داده هايي نياز داريم كه مقادير آن ها در اين ستون برابر با" خوزستان "باشد.

.٤شهرستان : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ شهرستان            : chr [1:43133] "ابهر" "ابهر" "ابهر" "ابهر" ...
```

.٥نوع ملك : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ نوع ملک             : chr [1:43133] "دستگاه آپارتمان" "دستگاه آپارتمان" "دستگاه آپارتمان" "دستگاه آپارتمان" ...
```

مقادير تمام داده هاي مربوط به اين ستون برابر با" دستگاه آپارتمان "است و اطلاعات خاصي در اختيار ما قرار نميدهد.

.٦منطقه شهرداري : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ منطقه شهرداري      : num [1:43133] NA NA NA NA NA NA NA NA NA 4 ...
```

اطلاعات اين ستون فقط براي برخي شهرستان ها در برخي استان ها مقدار گرفته است، به همين دليل نميتوانيم از اين ستون نيز استفاده كنيم

۷.نوع كاربري : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ نوع كاربري          : chr [1:43133] "مسكوني" "مسكوني" "مسكوني" "مسكوني" ...
```

مقادير تمام داده هاي مربوط به اين ستون برابر با" مسكوني "است و اطلاعات خاصي در اختيار ما قرار نميدهد.

۸.مساحت : مساحت ساختمان به متر مربع . نمونه اي از داده هاي اين ستون به شكل زير است

```
$ مساحت               : num [1:43133] 83 90 49 80.9 80 ...
```

۹.درصد : درصد فروش رفته از ساختمان در معامله . نمونه اي از داده هاي اين ستون به شكل زير است

```
$ درصد                : chr [1:43133] "100" "100" "100" "100" ...
```

۱۰.قيمت : قيمت كل ساختمان به هزار ريال. نمونه اي از داده هاي اين ستون به شكل زير است

```
$ قيمت                : num [1:43133] 3000000 1080000 10000000 3240000 750000 ...
```

۱۱.قيمت يك متر مربع : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ قيمت يک مترمربع    : chr [1:43133] "36144.58" "12000.00" "204081.63" "40039.55" ...
```

۱۲.عمر بنا : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ عمر بنا             : num [1:43133] 15 9 10 9 10 0 1 6 19 0 ...
```

۱۳.نوع اسكلت : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ نوع اسکلت           : chr [1:43133] "فلزي" "بتوني" "فلزي" "فلزي" ...
```

۱٤.تاريخ ثبت قرارداد : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ تاريخ ثبت قرارداد  : chr [1:43133] "1399/04/14" "1399/04/15" "1399/04/23" "1399/04/24" ...
```

اطلاعات خاصي در اختيار ما قرار نميدهد چون اين اطلاعات مربوط به يك بازه يك ماهه ميباشد و در اين بازه كوتاه نميتوان تاثير زمان بر تغييرات مختلف در داده ها را به درستي بررسي كرد و نتيجه گيري هاي مطمئن از آن ها استخراج كرد.

۱۵.شش رقم نخست كد پستي : نمونه اي از داده هاي اين ستون به شكل زير است

```
$ شش رقم نخست کد پستي: chr [1:43133] "456179" "456615" "456173" "456194" ...
```

همان طور که گفته شد تعداد زيادي از اين ستون ها اطلاعات خاصي به ما نميدهند و بايد قبل از کار با داده ها اين اطلاعات کنار بگذاريم.

ستون هايي که در ادامه کار به آن ها نياز داريم عبارتند از:

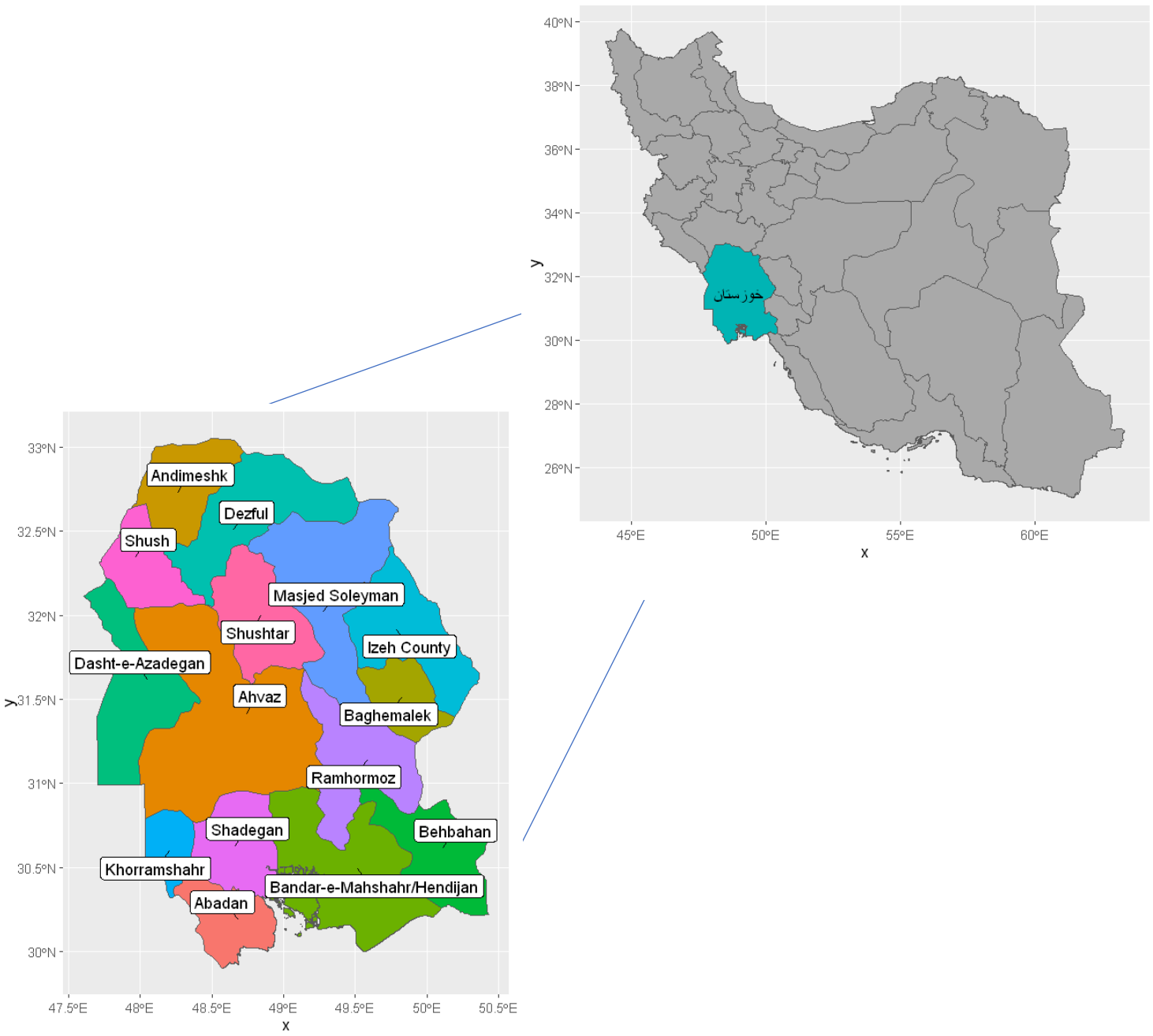
استان – شهر – مساحت – قيمت – قيمت هر متر مربع – عمر بنا – نوع اسکلت – شش رقم آخر کد پستي

که براي سادگي کار با داده ها نام اين ستون ها را به ترتيب به شکل زير در مي آوريم.

postal_code–skeleton_type–building_age–price_per_square–price–area–city–State

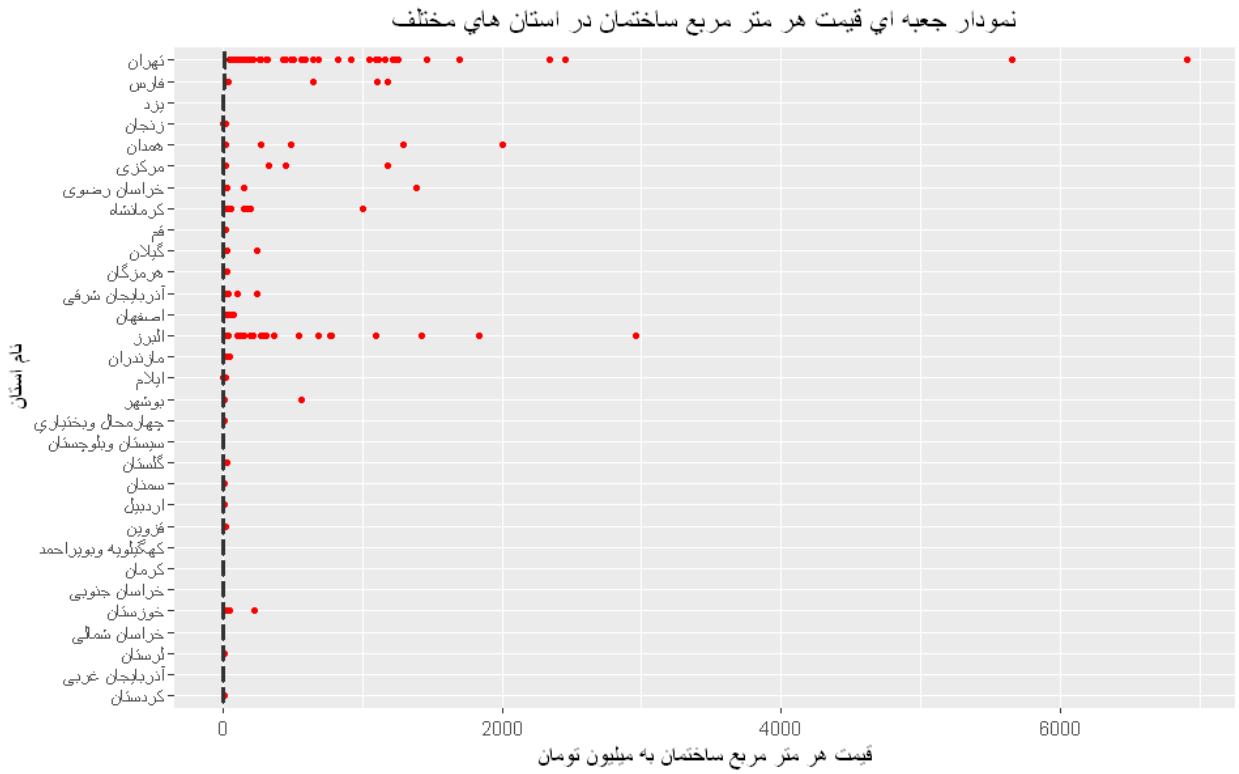
همچنين براي خوانا تر شدن اعداد موجود براي قيمت و قيمت هر متر مربع ، اين اطلاعات را به شکل واحد ميليون تبديل ميکنيم.

براي شروع کار ابتدا در نقشه ايران نگاهي به موقعيت مکاني استان خوزستان و شهرستان هاي اين استان مي اندازيم.



براي بررسي وضعيت قيمت در استان خوزستان نسبت به ديگر استان ها ابتدا نمودار جعبه اي قيمت ساختمان در استان هاي مختلف را رسم ميكنيم.

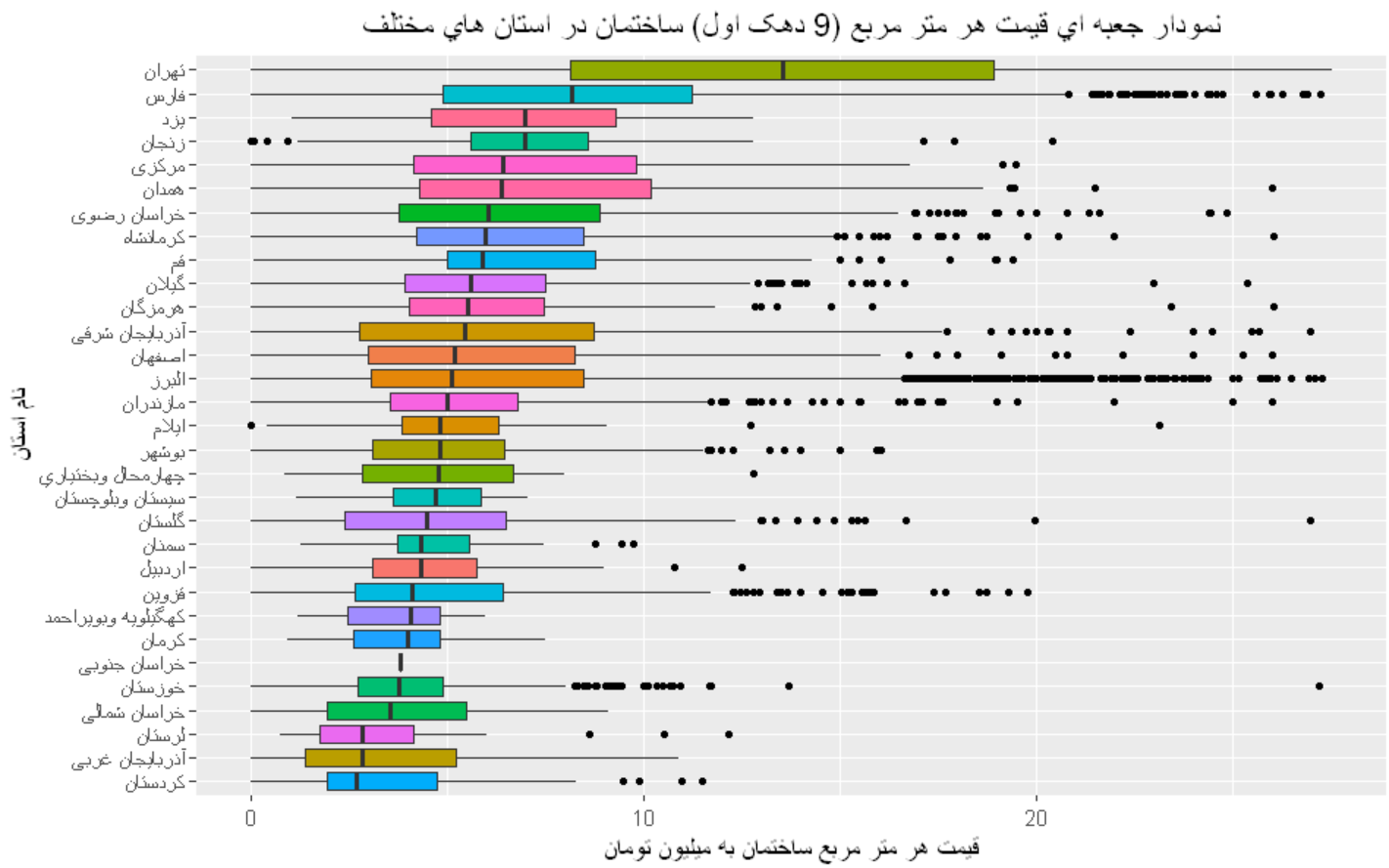
```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(data, aes(x=price_per_square, y=reorder(state, price_per_square, FUN=median) , fill=state)) +
    geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
    labs(title="نمودار جعبه ای قیمت هر متر مربع ساختمان در استان های مختلف", x="قیمت هر متر مربع ساختمان به میلیون تومان", y = "نام استان") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```



همان طور كه مشاهده ميكنيد به دليل وجود داده هاي پرت نميتوانيم درك درستي از داده هاي به دست آوريم، به همين دليل ميتوانيم موقتا برخي داده هاي پرت را كنار بگذاريم تا بتوانيم بخش اصلي داده ها را دقيق تر بررسي كنيم، به همين منظور اين نمودار را دوباره رسم ميكنيم با اين تفاوت كه به جاي رسم اطلاعات مربوط به تمام داده ها، اطلاعات داده هايي را رسم ميكنيم كه قيمت ساختمان در آن ها در۹دهك اول قيمتي وجود داشته باشد، به اين ترتيب داده هاي بسيار بزرگي كه وجود دارد را كنار ميگذاريم.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(data[data$price_per_square < quantile(data$price_per_square,0.9),],
    aes(x=price_per_square, y=reorder(state, price_per_square, FUN=median) , fill=state)) +
    geom_boxplot(outlier.colour="black", outlier.shape=16,outlier.size=1) +
    labs(title="نمودار جعبه ای قیمت هر متر مربع (9 دهک اول) ساختمان در استان های مختلف", x="قیمت هر متر مربع ساختمان به میلیون تومان", y = "نام استان") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```

Quantileدستوري است كه براي اين كار استفاده ميكنيم تا مرز بين دهك نهم و دهم را پيدا كنيم و داده هاي را از آن جا به دو بخش تقسيم كنيم. از دستور reorderنيز استفاده ميكنيم تا اين نمودار هاي جعبه اي را به ترتيب ميانه در نمودار بچينيم تا ديد بهتري به ما دهد.



همان طور که مشاهده ميکنيد با اين کار توانستيم درک بهتري نسبت به داده ها پيدا کنيم و همان طور که از نمودار واضح است با توجه به ميانه قيمت ها در استان هاي مختلف ، استان خوزستان در رده پنجم استان هاي ارزان کشور است.

براي ادامه کار قصد داريم داده هاي مربوط به استان خوزستان را دقيق تر بررسي کنيم بنابراين داده هايي که مقدار آن ها در ستونstateبرابر با" خوزستان "است را جدا ميکنيم.

```
khoozestan = data[data$state == 'خوزستان',]
```

در ابتدا قصد داريم همبستگي بين ستون ها را بررسي کنيم ، براي اين کار نياز داريم تا تمام داده هاي کيفي را به کمکdummy variableها به شکل داده هاي کمّي دربياوريم تا بتوانيم مقدارcorrelationرا براي آن ها به دست بياوريم. داده هاي کيفي اينdataframeدر ستون هايcityوbuilding_typeميباشد.

براي اين کار از دستور زير استفاده ميکنيم.

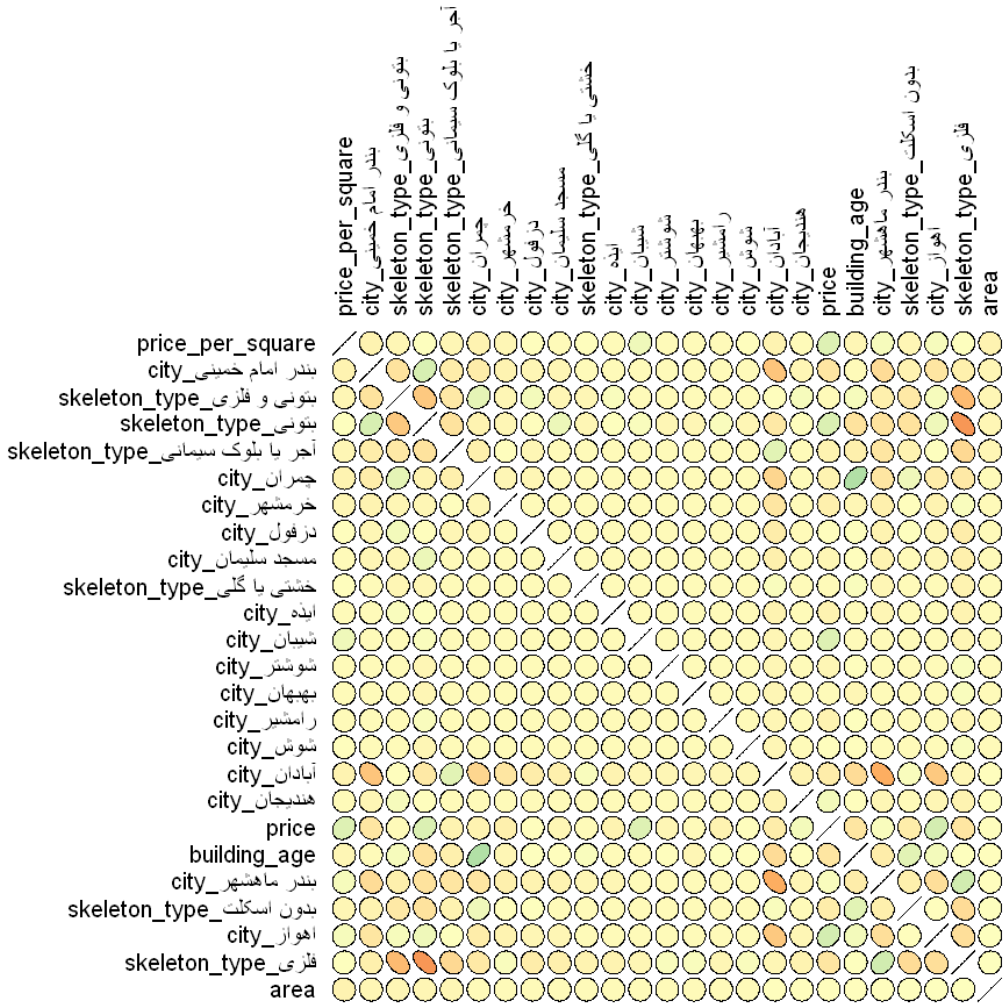
```
df<-khoozestan %>%
    select('city','area','building_age','skeleton_type','price','price_per_square')
```

```
df = dummy_cols(df, select_columns = c('city', 'skeleton_type'),remove_selected_columns = TRUE)
```

سپس با دستور `data <- cor(df)` ماتريسcorrelationرا براي اين داده ها به دست مي آوريم.

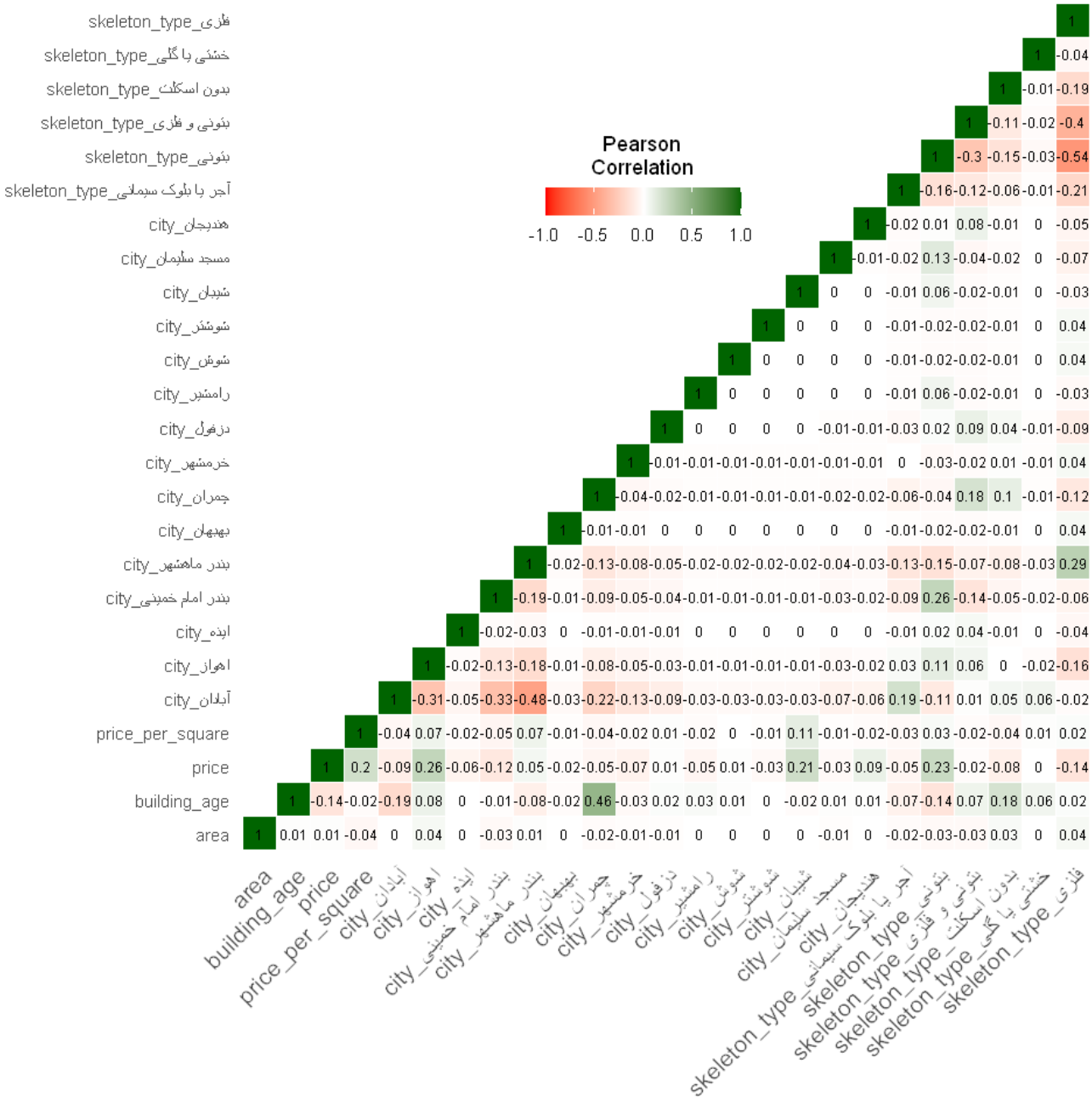
براي درك بهتر اين ماتريس ميتوانيم از نمودار زير استفاده كنيم.

```
options(repr.plot.width = 8, repr.plot.height = 8)
data <- cor(df)

# Build a Pannel of 100 colors with Rcolor Brewer
my_colors <- brewer.pal(5, "Spectral")
my_colors <- colorRampPalette(my_colors)(100)

# Order the correlation matrix
ord <- order(data[1, ])
data_ord <- data[ord, ord]
plotcorr(data_ord , col=my_colors[data_ord*50+50] , mar=c(0,0,0,0))
```



در اين نمودار هر چه رنگ بيضي مربوطه به سبز نزديك تر باشد دو متغير رابطه مستقيم قوي تري دارند و هر چه به رنگ قرمز نزديك تر باشد دو متغير رابطه معكوس قوي تري دارند و بخش هايي كه رنگ آن به زرد نزديك است يعني دو متغير هيچ رابطه اي با يكديگر ندارند.

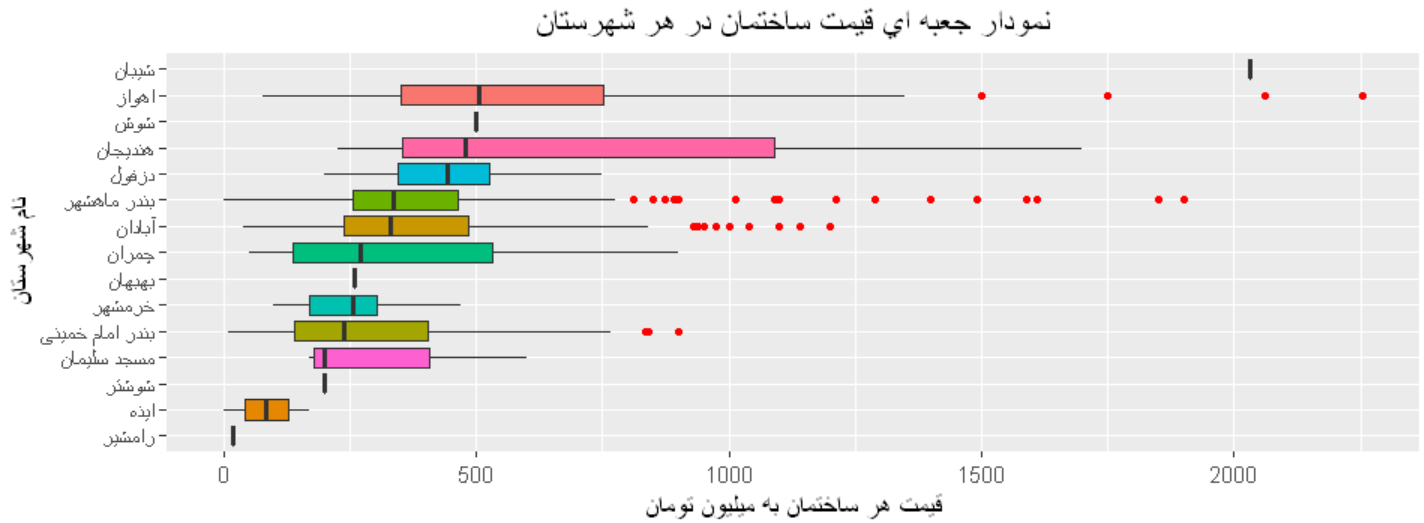
براي مشاهده دقيق تر اعداد شاخص همبستگي ميتوانيم از نمودار زير نيز استفاده كنيم.



حالا نمودار جعبه اي قيمت ساختمان در شهرستان هاي مختلف را با دستور زير رسم ميكنيم.

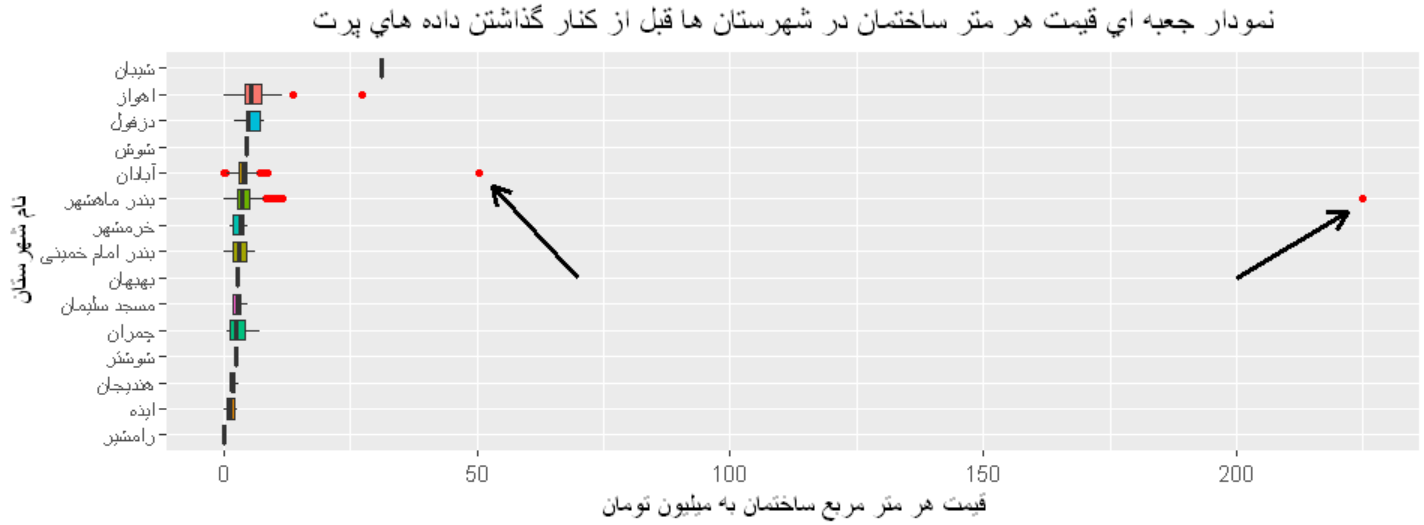
```
options(repr.plot.width = 8, repr.plot.height = 3)
ggplot(khoozestan, aes(x=price, y=reorder(city, price, FUN=median) , fill=city)) +
    geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
    labs(title="نمودار جعبه ای قیمت ساختمان در هر شهرستان", x="قیمت هر ساختمان به میلیون تومان", y = "نام شهرستان") +
    theme(legend.position="none", plot.title = element_text(hjust = 0.5))
```

و شهرستان ها را به ترتیب میانه قیمت در نمودار نمایش میدهیم.



سپس همین کار را با قیمت هر متر مربع نیز انجام میدهیم.

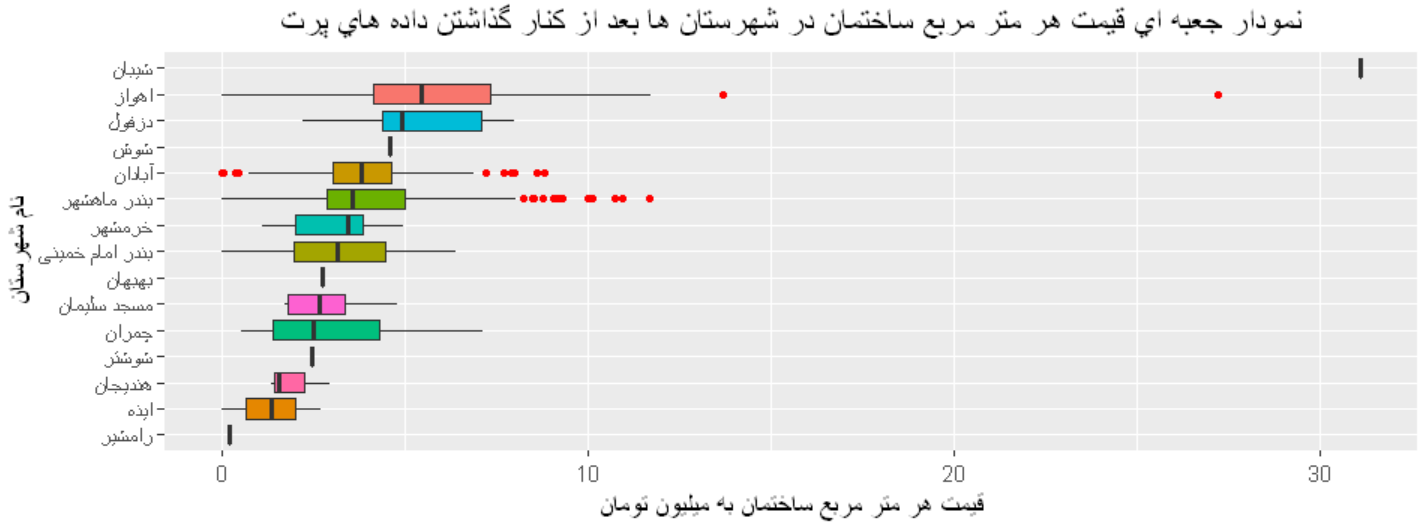
```
options(repr.plot.width = 8, repr.plot.height = 3)
ggplot(khoozestan, aes(x=price_per_square, y=reorder(city, price_per_square, FUN=median) , fill=city)) +
geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
labs(title="نمودار جعبه ای قیمت هر متر ساختمان در شهرستان ها قبل از کنار گذاشتن داده های پرت", x="قیمت هر متر مربع ساختمان به میلیون تومان", y = "نام شهرستان") +
geom_segment(aes(x = 200, y = 7, xend = 222, yend = 9.5),size = 1,arrow = arrow(length = unit(.3, "cm"))) +
geom_segment(aes(x = 70, y = 7, xend = 53, yend = 10.5),size = 1,arrow = arrow(length = unit(.3, "cm"))) +
theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```



همان طور که مشاهده میکنید وجود این دو داده ي پرت باعث میشود نتوانیم به خوبي اطلاعات بقیه داده ها را مشاهده و بررسي کنیم به همین دلیل موقتا این دو داده را از اطلاعاتمان کنار میگذاریم اما مشخصات مربوط به آن ها را حذف نمیکنیم تا بتوانیم در صورت نیاز به دقت این داده ها را بررسي کنیم، لازم به ذکر است که قصد ما از حذف کردن این داده ها غلط نشان دادن آن ها و یا بي اهمیت نشان دادنشان نیست و فقط براي درك بهتر بقیه داده ها این کار را انجام میدهیم.

کد قرار داد در این دو داده۱۹۱۸۹۱۹۸و۱۹۲۳۶۳۰۳بود.

مجددا این نمودار را رسم میکنیم.



به این شکل میتوانیم دقیق تر این داده ها را بررسي كنيم.

مشكلي كه در این جا ماوجه آن میشویم که در نمودار قبلي قطعا نمیتوانستیم متوجه آن شویم این است که تعداد داده ها در بعضي از شهرستان ها بسیار کم است و این موضوع سطح اطمینان ما به این داده ها را کم میکند. شهرستان هايي مانند شیبان ، شوش ، شوشتر ، بهبهان و رامشیر.

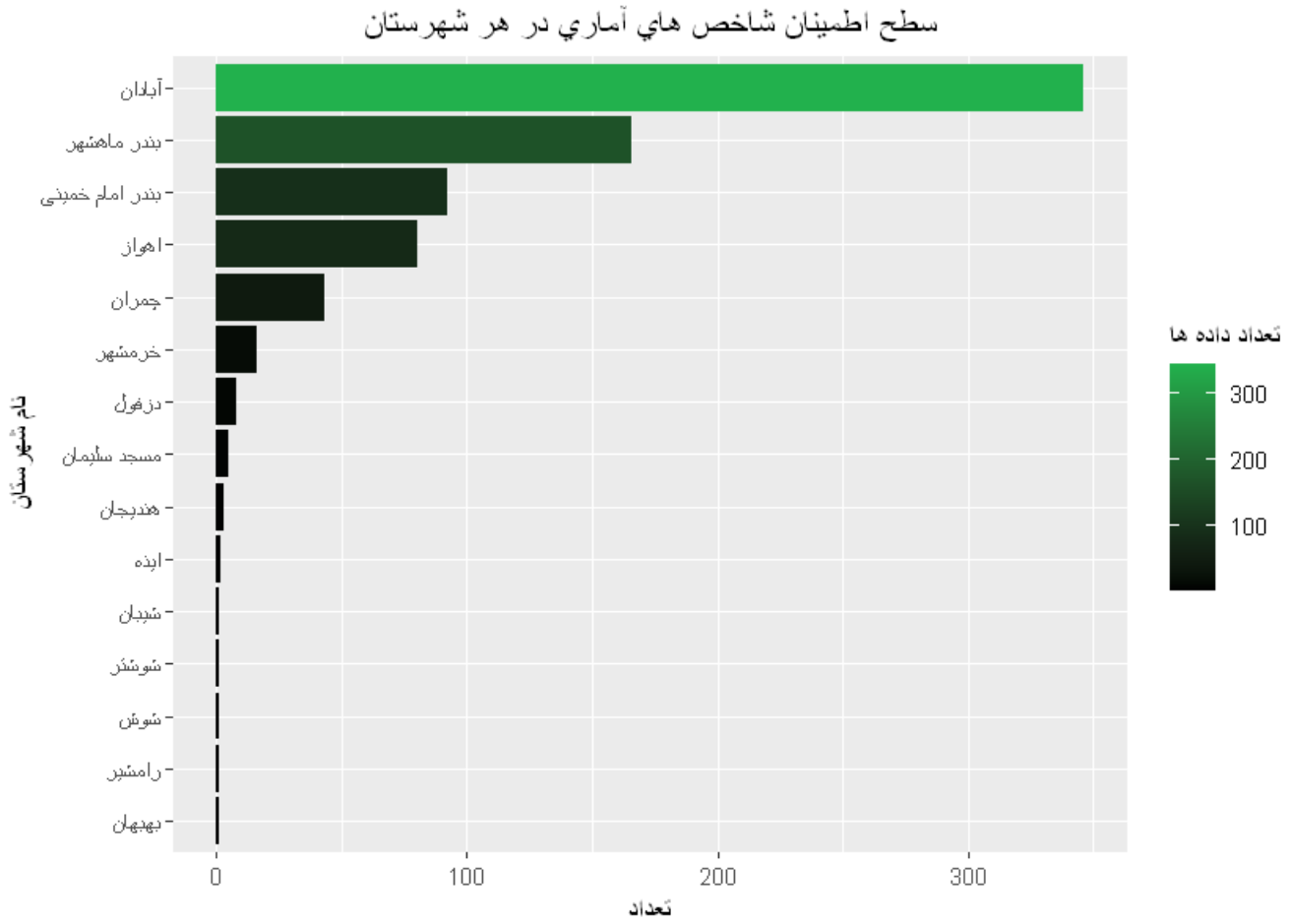
براي بررسي دقیق تر این موضوع ابتدا تعداد داده هاي مربوط به هر شهرستان را شمارش میکنیم.

```
confidence_level <- khoozestan %>%
    select(city) %>%
    group_by(city) %>%
    count() %>%
    arrange(n)
```

city	n
بهبهان	1
رامشیر	1
شوش	1
شوشتر	1
شیبان	1
ایذه	2
هندیجان	3
مسجد سلیمان	5
دزفول	8
خرمشهر	16
چمران	43
اهواز	80
بندر امام خمیني	92
بندر ماهشهر	166
آبادان	346

براي درك بهتر اين موضوع اين اعداد را در يك نمودار ميله اي نمايش ميدهيم. هر چه تعداد داده هاي مربوط به يك شهرستان بيشتر باشد، سطح اطمينان ما به شاخص هاي آماري مربوط به آن شهرستان نيز افزايش پيدا ميكند.

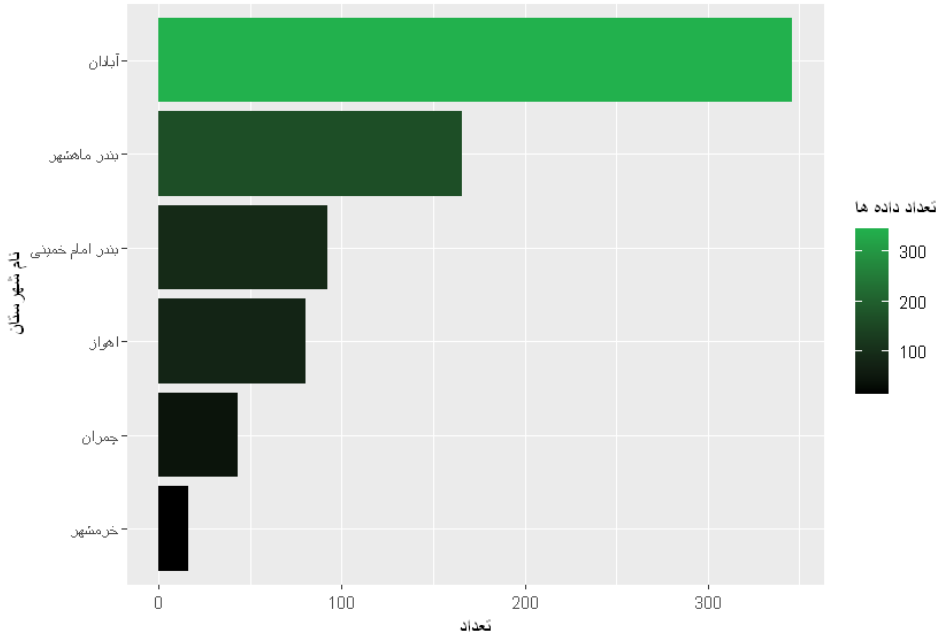
```
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(confidence_level , aes(x=n , y=reorder(city,n), fill = n)) +
    geom_bar(stat = "identity") +
    labs(title="سطح اطمینان شاخص های آماری در هر شهرستان", x="تعداد", y = "نام شهرستان") +
    scale_fill_gradient(name = "تعداد داده ها",low="black", high="#22b14d") +
    theme(plot.title = element_text(hjust = 0.5))
```



همان طور كه مشاهده ميكنيد علاوه بر بهبان ، رامشير ، شوش ، شوشتر و شيبان ، تعداد داده هاي شهرستان هاي ايذه ، هنديجان ، مسجد سليمان و دزفول نيز زير ۱۰ميباشد و به همين دليل سطح اطمينان شاخص هاي آماري در اين داده ها بسيار پايين است و به همين دليل داده هاي مربوط به اين شهرستان ها را كنار ميگذاريم تا بتوانيم بقيه داده ها را دقيق تر بررسي كنيم.

مجددا بر اين نكته تاكيد ميكنم كه تمام اين داده ها براي ما مهم هستند و بايد اطلاعات هر يك مورد بررسي قرار گيرد.

سطح اطمينان شاخص هاي آماري در هر شهرستان پس از كنار گذاشتن شهرستان ها با تعداد داده هاي زير 10



همچنين براي اين كار ميتوانيم از نمودارTreemapاستفاده كنيم.

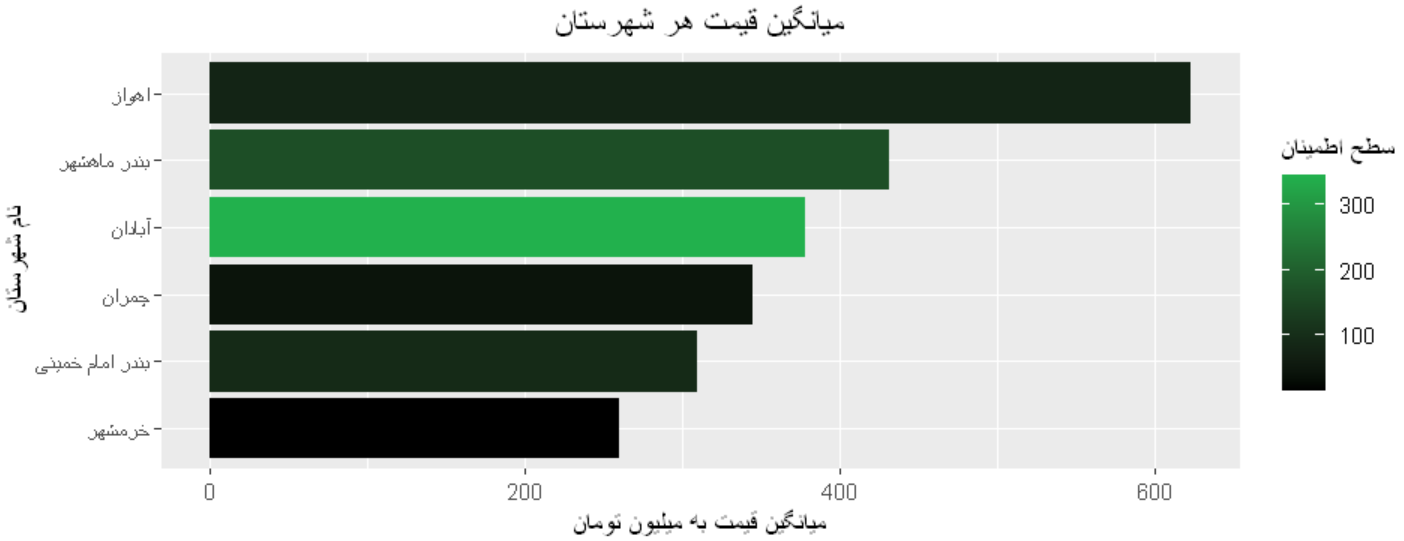
```
options(repr.plot.width = 5, repr.plot.height = 5)
ggplot(confidence_level, aes(area = n, fill = n,label = city)) +
    geom_treemap() +
    geom_treemap_text(colour = "white",size = 15, place = "centre",grow = FALSE) +
    labs(title="تعداد داده ها در هر شهرستان") +
    scale_fill_gradient(name = "تعداد داده ها",low="black", high="#22b14d") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```

براي مقايسه ميانگين قيمت بين اين چند شهرستان از نمودار ميله اي استفاده ميكنيم.

city	mean_price
اهواز	623.4250
آبادان	378.0111
بندر امام خميني	309.6426
بندر ماهشهر	431.9117
چمران	344.5581
خرمشهر	259.9812

```
df <- khoozestan %>%
    select(city,price) %>%
    group_by(city) %>%
    summarise(mean_price = mean(price))
df
```

```
options(repr.plot.width = 8, repr.plot.height = 3.1)
confidence = confidence_level[match(reorder(df$city , df$mean_price), confidence_level$city),]$n
ggplot(df , aes(x=mean_price , y=reorder(city,mean_price),fill = confidence)) +
    geom_bar(stat = "identity") +
    labs(title="ميانگين قيمت هر شهرستان", x="ميانگين قيمت به ميليون تومان", y = "نام شهرستان") +
    scale_fill_gradient(name='سطح اطمينان',low="black", high="#22b14d") +
    theme(plot.title = element_text(hjust = 0.5))
```

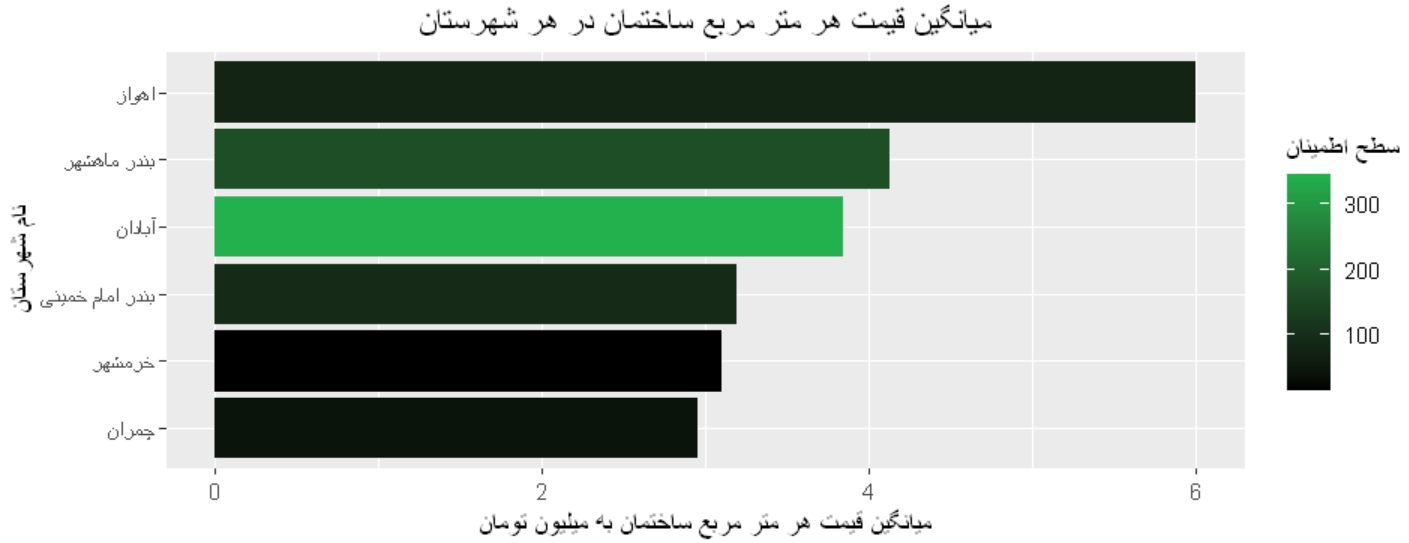


همين نمودار را براي اطلاعات مربوط به قيمت هر متر مربع نيز رسم ميكنيم.

city	mean_price_per_square
اهواز	6.002169
آبادان	3.839624
بندر امام خميني	3.195575
بندر ماهشهر	4.127006
چمران	2.949167
خرمشهر	3.095103

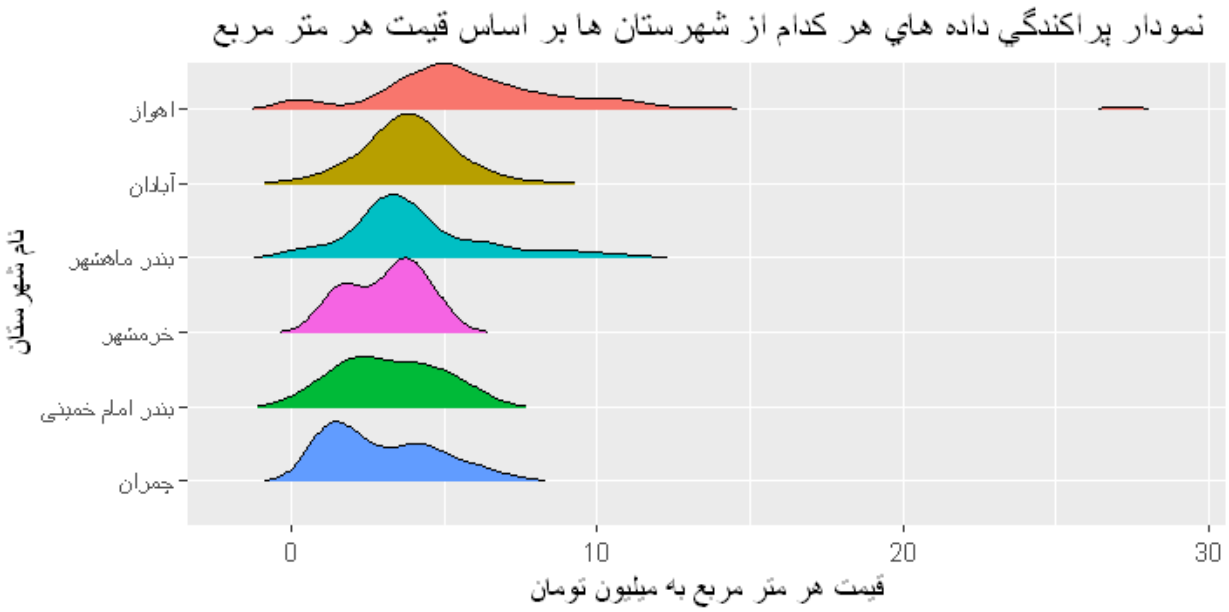
```
df <- khoozestan %>%
    select(city,price_per_square) %>%
    group_by(city) %>%
    summarise(mean_price_per_square = mean(price_per_square))
df
```

```
options(repr.plot.width = 8, repr.plot.height = 3.1)
confidence = confidence_level[match(reorder(df$city , df$mean_price_per_square), confidence_level$city),]$n
ggplot(df , aes(x=mean_price_per_square , y=reorder(city,mean_price_per_square),fill = confidence)) +
    geom_bar(stat = "identity") +
    labs(title="ميانگين قيمت هر متر مربع ساختمان در هر شهرستان", x="ميانگين قيمت هر متر مربع ساختمان به ميليون تومان", y = "نام شهرستان") +
    scale_fill_gradient(name='سطح اطمينان',low="black", high="#22b14d") +
    theme(plot.title = element_text(hjust = 0.5))
```



پراکندگي داده ها در هر يك شهرستان ها بر اساس قيمت هر متر مربع ساختمان به شكل زير است.

```
options(repr.plot.width = 6, repr.plot.height = 3)
ggplot(khoozestan, aes(x = price_per_square, y = reorder(city, price_per_square, FUN=median), fill = city)) +
  geom_density_ridges_gradient(scale = 1, rel_min_height = 0.01) +
    labs(title="نمودار پراکندگی داده های هر کدام از شهرستان ها بر اساس قیمت هر متر مربع", x="قیمت هر متر مربع به میلیون تومان", y = "نام شهرستان") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```

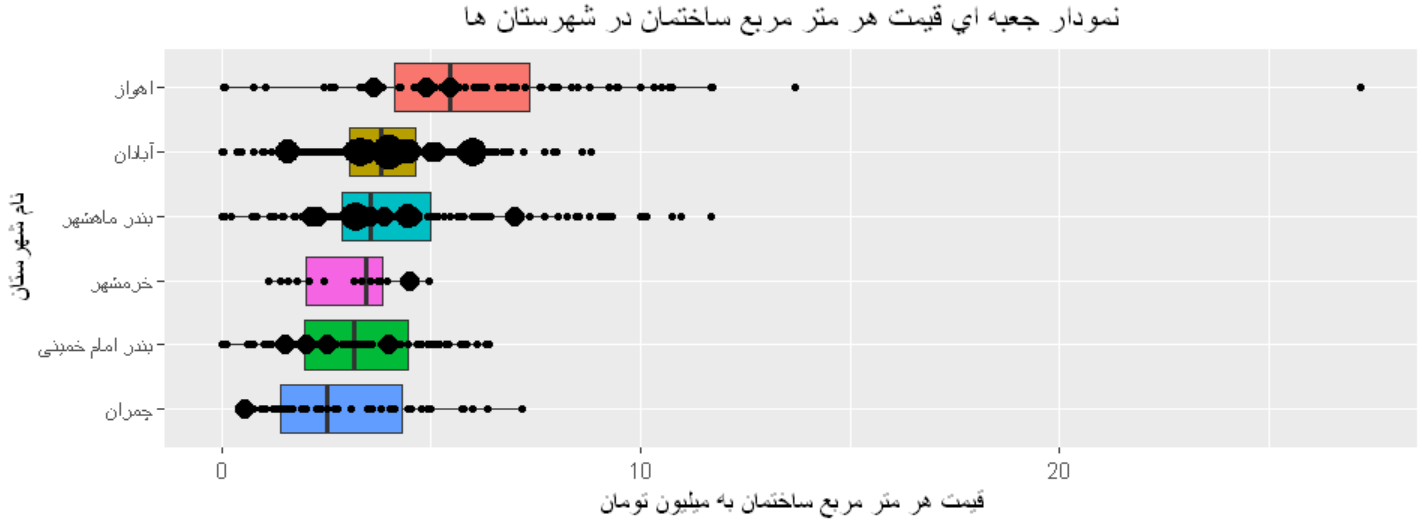


ميتوان تعداد داده هاي موجود در هر قيمت را به كمك دستورgeom_countروي نمودار جعبه اي نشان داد كه در نمودار زير ميتوانيد آن را مشاهده كنيد.

```
options(repr.plot.width = 8, repr.plot.height = 3)
ggplot(khoozestan, aes(x=price, y=reorder(city, price, FUN=median) , fill=city)) +
    geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
    geom_count(col="black", show.legend=F)+
    labs(title="نمودار جعبه ای قیمت ساختمان در هر شهرستان", x="قیمت هر ساختمان به میلیون تومان", y = "نام شهرستان") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```

نمونه ي ديگري از همين نمودار براي قيمت هر متر مربع را ميتوانيم به شكل زير رسم كنيم.

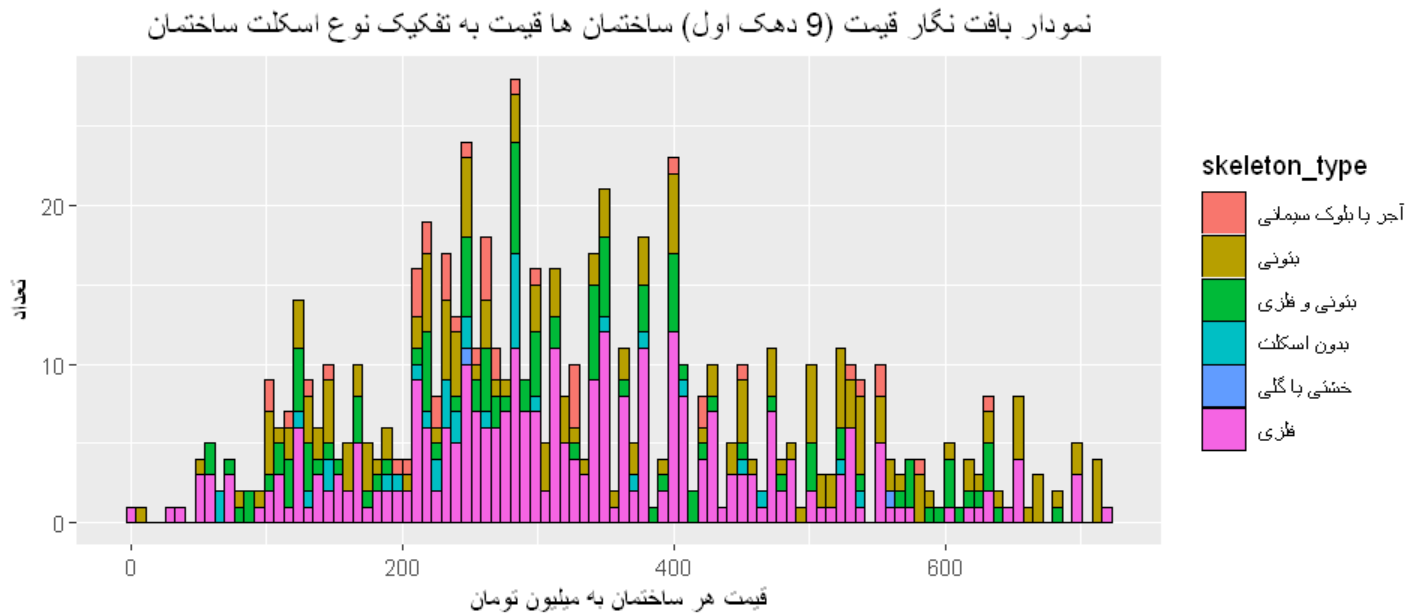
```
options(repr.plot.width = 8, repr.plot.height = 3)
ggplot(khoozestan, aes(x=price_per_square, y=reorder(city, price_per_square, FUN=median) , fill=city)) +
    geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
    geom_count(col="black", show.legend=F)+
    labs(title="نمودار جعبه ای قیمت هر متر مربع ساختمان در شهرستان ها", x="قیمت هر متر مربع ساختمان به میلیون تومان", y = "نام شهرستان") +
    theme(legend.position="none",plot.title = element_text(hjust = 0.5))
```



حالا قصد داريم تاثير اسكلت ساختمان بر قيمت بنا را بررسي كنيم.

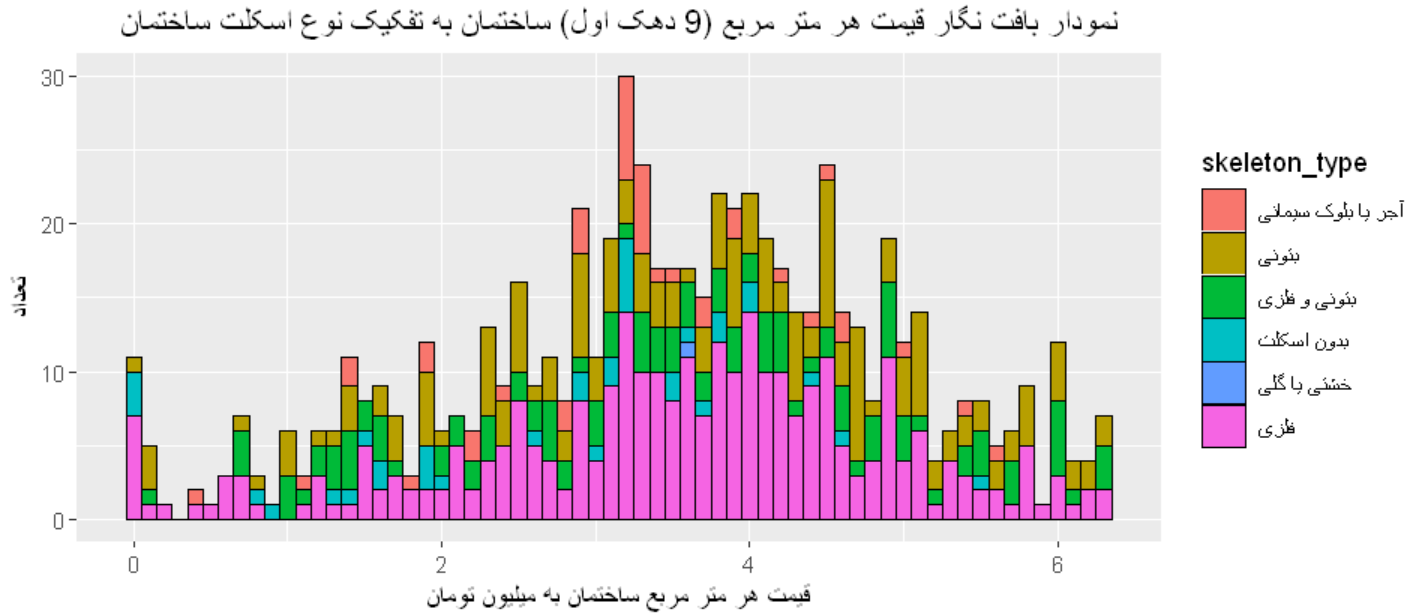
نمودار اولي كه براي اين منظور رسم ميكنيم كه نمودار پراكندگي قيمت ساختمان ها بر اساس قيمت آن ها ميباشد.

```
options(repr.plot.width = 8, repr.plot.height = 3.5)
ggplot(khoozestan[khoozestan$price < quantile(khoozestan$price,0.9),], aes(price)) +
#     geom_histogram(aes(fill=skeleton_type), binwidth = 100,col='black', size=.1) +
    geom_histogram(aes(fill=skeleton_type), bins=100, col="black", size=.1) +
    labs(title="نمودار بافت نگار قیمت (9 دهک اول) ساختمان ها قیمت به تفکیک نوع اسکلت ساختمان", x="قیمت هر ساختمان به میلیون تومان", y = "تعداد")+
    theme(plot.title = element_text(hjust = 0.5))
```



و سپس همين نمودار را براي قيمت هر متر مربع رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 3.5)
ggplot(khoozestan[khoozestan$price_per_square < quantile(khoozestan$price_per_square,0.9),], aes(price_per_square)) +
    geom_histogram(aes(fill=skeleton_type), binwidth = 0.1,col='black', size=.1) +
#     geom_histogram(aes(fill=skeleton_type), bins=100, col="black", size=.1) +
    labs(title="نمودار بافت نگار قیمت هر متر مربع (9 دهک اول) ساختمان به تفکیک نوع اسکلت ساختمان", x="قیمت هر متر مربع ساختمان به میلیون تومان", y = "تعداد")+
    theme(plot.title = element_text(hjust = 0.5))
```



در اين دو نمودار براي درك بهتر از داده ها تنها۹دهك اول داده در آن متغير (قيمت يا قيمت هر متر مربع) را رسم كرده ايم.

در نمودار اول كل۹دهك اول را به۱۰۰بازه از قيمت تقسيم كرده ايم و تعداد داده هاي مربوط به هر نوع از اسكلت ها را در آن نشان داده ايم.

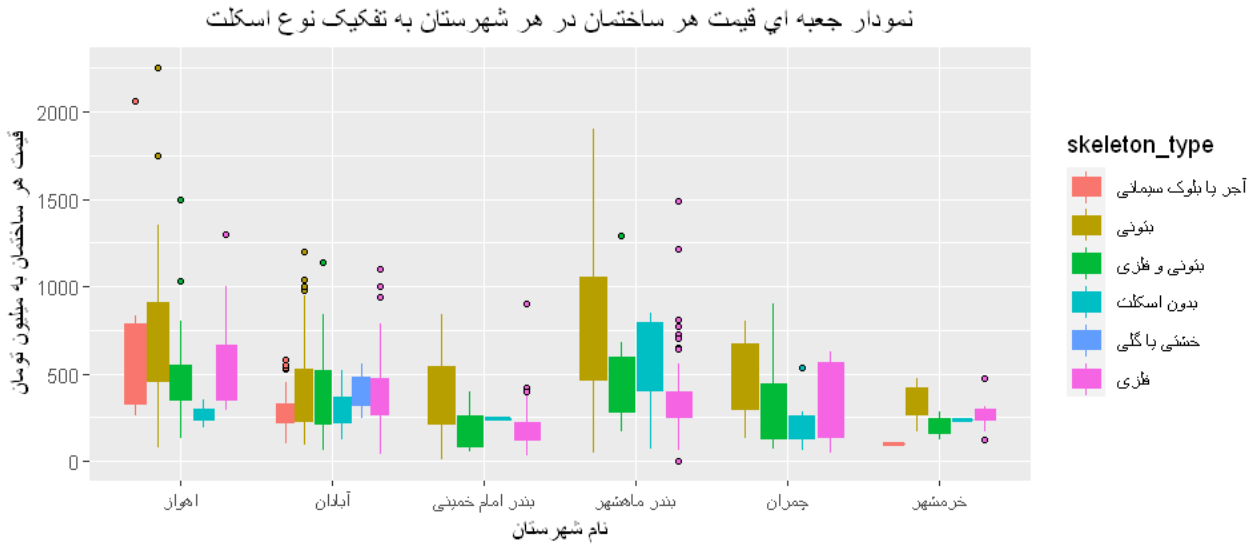
این کار را با قرار دادن مقدارbinsبرابر با۱۰۰انجام داده ایم.

در نمودار دوم کل۹دهك اول را به ازاي هر۱۰۰هزار تومن يك بازه جديد ايجاد كرده ايم و تعداد داده هاي مربوط به هر نوع از اسكلت ها را در آن نشان داده ايم.

این کار را با قرار دادن مقدارbinwidthبرابر با۱.۰انجام داده ایم چون واحد ما در این داده ها میلیون تومان است پس۱.۰آن برابر با۱۰۰هزار تومان است.

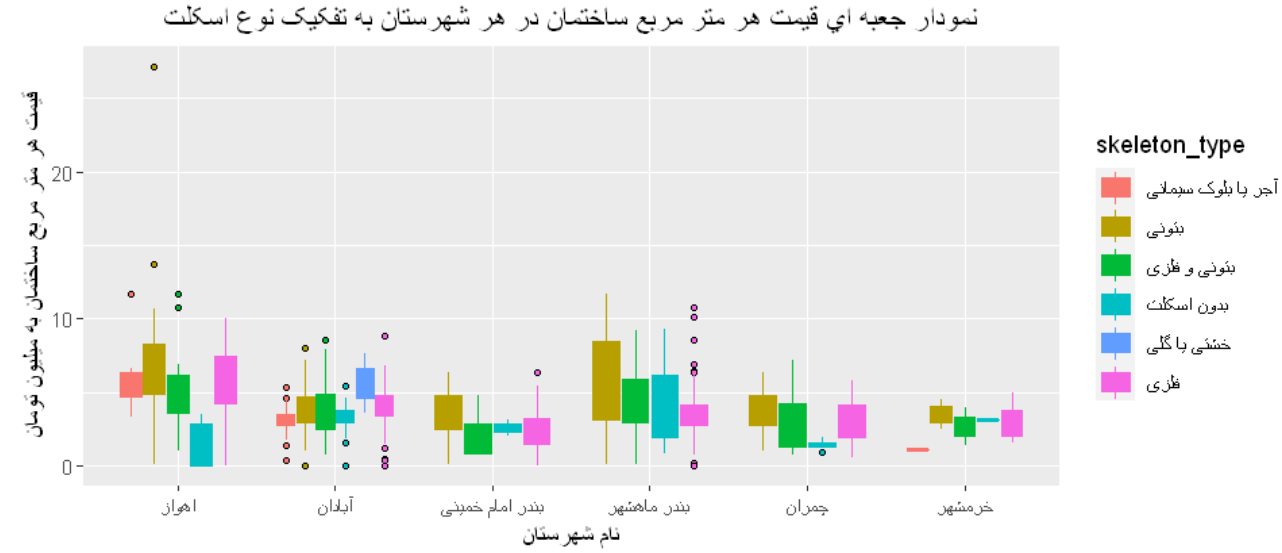
نمودار بعدي كه رسم ميكنيم نمودار جعبه اي قيمت شهرستان ها به تفكيك نوع اسكلت در آن ها ميباشد.

```
options(repr.plot.width = 8, repr.plot.height = 3.5)
ggplot(khoozestan, aes(x=city, y=price , fill = skeleton_type)) +
    geom_boxplot(aes(colour = skeleton_type) ,outlier.colour="black", outlier.shape=21,outlier.size=1) +
    labs(title="نمودار جعبه ای قیمت هر ساختمان در هر شهرستان به تفکیک نوع اسکلت", y="قیمت هر ساختمان به میلیون تومان", x = "نام شهرستان") +
    theme(plot.title = element_text(hjust = 0.5))
```



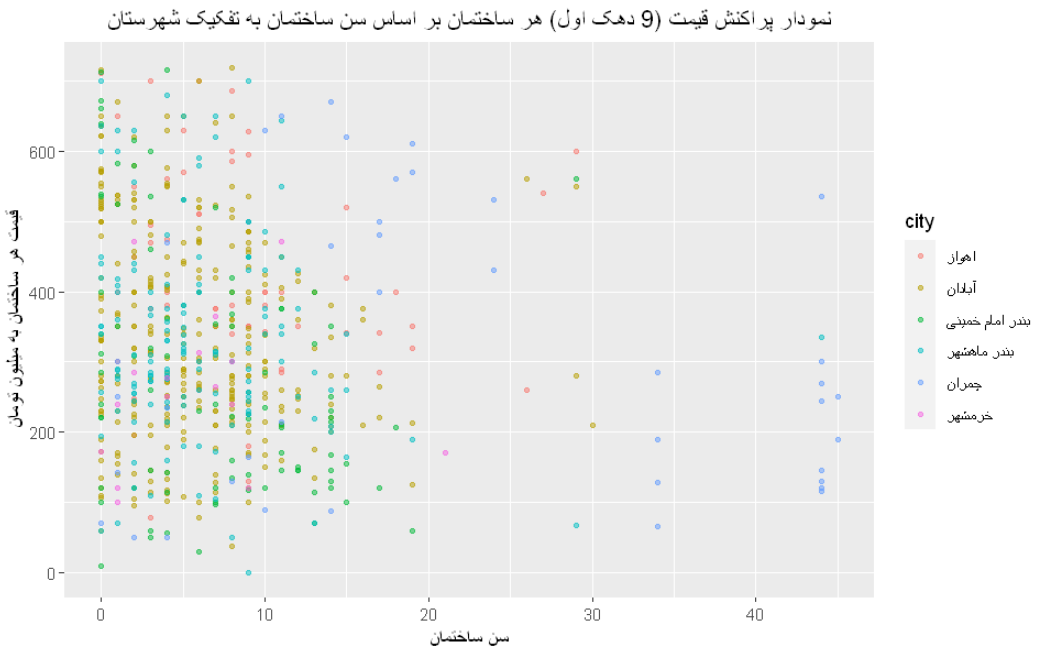
و سپس اين نمودار را براي قيمت هر متر مربع ساختمان نيز رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 3.5)
ggplot(khoozestan, aes(x=city, y=price , fill = skeleton_type)) +
    geom_boxplot(aes(colour = skeleton_type) ,outlier.colour="black", outlier.shape=21,outlier.size=1) +
    labs(title="نمودار جعبه ای قیمت هر ساختمان در هر شهرستان به تفکیک نوع اسکلت", y="قیمت هر ساختمان به میلیون تومان", x = "نام شهرستان") +
    theme(plot.title = element_text(hjust = 0.5))
```



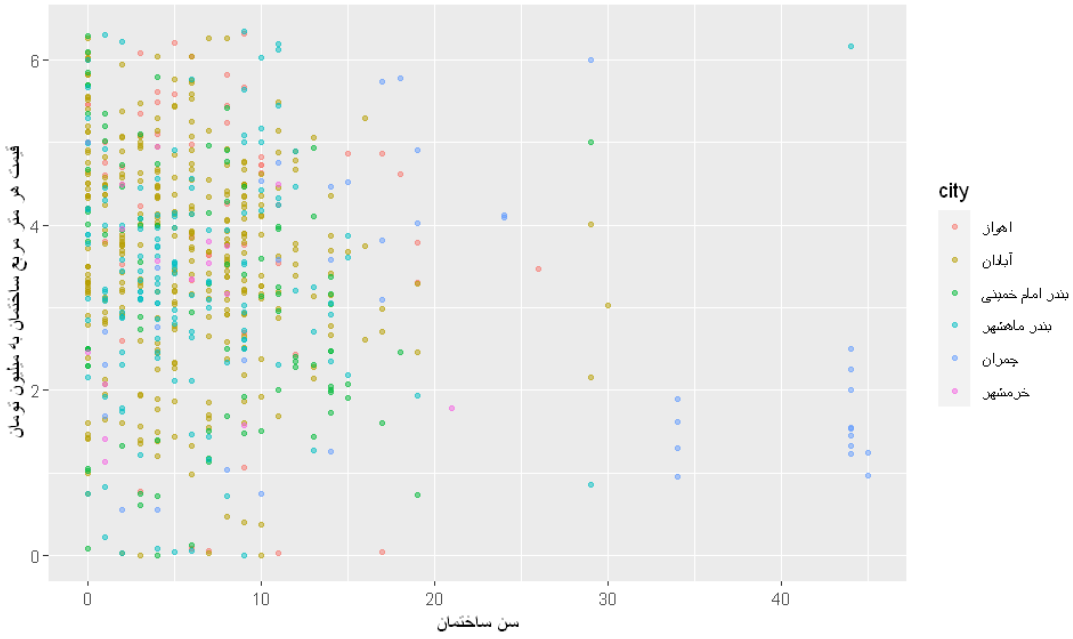
براي بررسي تاثير سن ساختمان روي قيمت ساختمان نمودار پراكنشي داده هاي اين دو ستون را به تفكيك شهرستان رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan[khoozestan$price < quantile(khoozestan$price,0.9),], aes(x=building_age, y=price , color = city)) +
    geom_point(size=1,alpha=0.5) +
    labs(title="نمودار پراکنش قیمت (9 دهک اول) هر ساختمان بر اساس سن ساختمان به تفکیک شهرستان", x="سن ساختمان", y = "قیمت هر ساختمان به میلیون تومان") +
    theme(plot.title = element_text(hjust = 0.5))
```



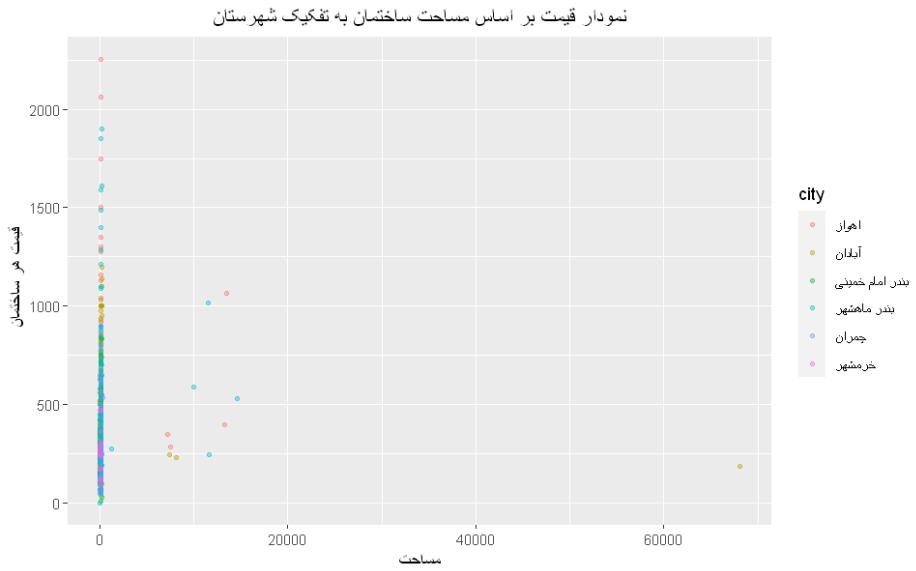
و سپس اين نمودار را براي قيمت هر متر مربع نيز رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan[khoozestan$price_per_square < quantile(khoozestan$price_per_square,0.9),],
    aes(x=building_age, y=price_per_square , color = city)) +
    geom_point(size=1,alpha=0.5) +
    labs(title="نمودار پراکنش قیمت هر متر مربع (9 دهک اول) ساختمان بر اساس سن ساختمان به تفکیک شهرستان",
    x="سن ساختمان", y ="قیمت هر متر مربع ساختمان به میلیون تومان") +
    theme(plot.title = element_text(hjust = 0.5))
```



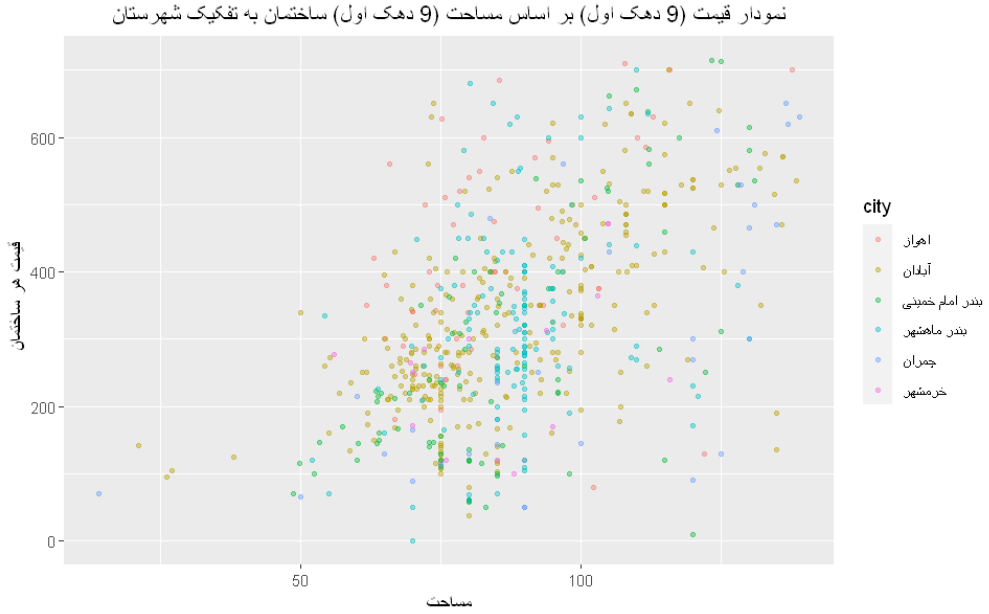
سپس براي بررسي تاثير مساحت روي قيمت ساختمان نمودار زير را رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan, aes(x=area, y=price , color = city)) +
    geom_point(size=1,alpha=0.4,) +
    labs(title="نمودار قيمت بر اساس مساحت ساختمان به تفكيك شهرستان", y=" قيمت هر ساختمان", x = "مساحت")+
    theme(plot.title = element_text(hjust = 0.5))
```



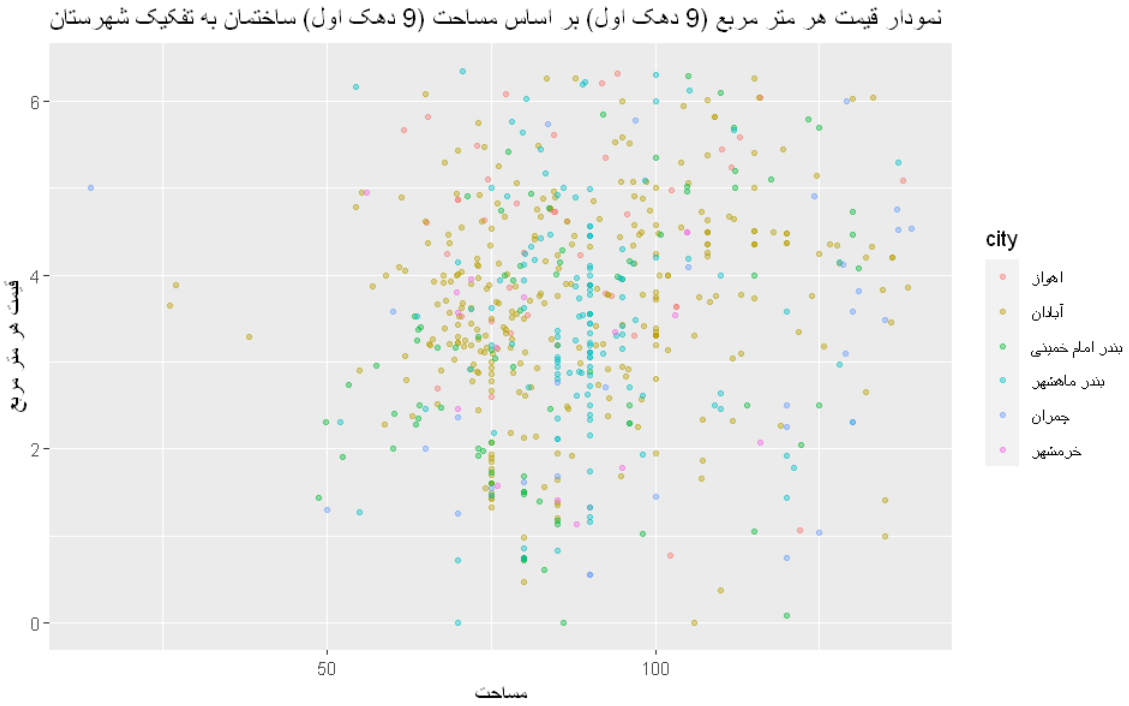
همان طور كه مشاهده ميكنيد در اين جا نيز به دليل وجود داده هاي پرت و ساختمان هايي با مساحت هاي بسيار بالا بررسي بخش عمده اي از داده ها عملا غير ممكن شده است بنابراين در اين جا نيز تنها۹دهك اول داده ها را بررسي ميكنم. در تمام بخش هاي اين گزارش كه تنها۹دهك از داده ها نمايش داده شده است به همين دليل است كه در غير اين صورت مصور سازي اين داده ها نميتوانست هيچ اطلاعاتي در اختيار ما قرار دهد.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan[(khoozestan$area < quantile(khoozestan$area,0.90))&(khoozestan$price < quantile(khoozestan$price,0.9)),],
    aes(y=price, x=area , color = city)) +
    geom_point(size=1,alpha=0.4,) +
    labs(title="نمودار قيمت (9 دهك اول) بر اساس مساحت (9 دهك اول) ساختمان به تفكيك شهرستان", y=" قيمت هر ساختمان", x = "مساحت")+
    theme(plot.title = element_text(hjust = 0.5))
```



سپس همين نمودار را براي قيمت هر متر مربع نيز رسم ميكنيم.

```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan[(khoozestan$area < quantile(khoozestan$area,0.90))&
    (khoozestan$price_per_square < quantile(khoozestan$price_per_square,0.9)),],
    aes(y=price_per_square, x=area , color = city)) +
    geom_point(size=1,alpha=0.4,) +
    labs(title="نمودار قيمت هر متر مربع (9 دهک اول) بر اساس مساحت (9 دهک اول) ساختمان به تفکيک شهرستان", y="قيمت هر متر مربع", x = "مساحت")
```



در اين گزارش سعي كرديم تا داده هاي مربوط به خريد و فروش مسكن در استان خوزستان را بررسي كنيم ، به عوامل تاثير گذار روي قيمت مسكن پرداختيم و سعي كرديم تاثير هر يك را بر شهرستان هاي مختلف استان خوزستان آشكار كنيم تا براي شخصي كه قصد دارد براي اولين بار با اين داده ها كار كند ديد كافي را ايجاد كنيم تا بتواند دقيق تر اين داده ها را مورد بررسي قرار دهد ، قطعا نكات مهم ديگري نيز در اين داده ها وجود دارد كه بنده نتوانستم به خوبي آن ها را نمايش دهم.

اين داده ها مشكلاتي نيز داشتند كه ميتوان به كم بودن تعداد داده ها در شهرستان هاي مختلف و كم بودن پراكندگي داده ها در برخي از ستون ها اشاره كرد كه باعث شد اطلاعات خيلي زيادي براي نمايش و بررسي موجود نباشد.

پينوشت : متاسفانه اطلاع نداشتم كه بايد براي اين تمرين گزارش تهيه كنم به همين دليل گزارش را با تاخير خدمتتان ارسال كردم.

با احترام كنعاني