

گزارش پروژه اول – مصور سازی داده های مربوط به خرید و فروش مسکن

استان خوزستان

مجتبی کنعانی سرچشمه

در این پروژه قصد داریم داده های مربوط به خرید و فروش مسکن در استان خوزستان را مصورسازی کنیم.

اطلاعات مربوط به کل استان ها از ۴۳۱۳۳ داده و ۱۵ متغیر تشکیل شده است. که ۷۶۸ عدد از این داده ها اطلاعات مربوط به استان خوزستان است.

ابتدا به طور مختصر به معرفی ستون های داده میپردازیم.

۱. کد قرارداد : نمونه ای از داده های این ستون به شکل زیر است

کد قرارداد \$: chr [1:43133] "19136381" "19140513" "19192568" "19202902" ...

این ستون اطلاعات خاصی در اختیار ما قرار نمیدهد.

۲. نوع قرارداد : نمونه ای از داده های این ستون به شکل زیر است

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

مقادیر تمام داده های مربوط به این ستون برابر با "مبايعه نامه" است و اطلاعات خاصی در اختیار ما قرار نمیدهد.

۳. استان : نمونه ای از داده های این ستون به شکل زیر است

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

به داده هایی نیاز داریم که مقادیر آن ها در این ستون برابر با "خوزستان" باشد.

۴. شهرستان : نمونه ای از داده های این ستون به شکل زیر است

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

۵. نوع ملک : نمونه ای از داده های این ستون به شکل زیر است

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

مقادیر تمام داده های مربوط به این ستون برابر با "دستگاه آپارتمان" است و اطلاعات خاصی در اختیار ما قرار نمیدهد.

۶. منطقه شهرداری : نمونه ای از داده های این ستون به شکل زیر است

منطقه شهرداری \$: num [1:43133] NA NA NA NA NA NA NA NA NA 4 ...

اطلاعات این ستون فقط برای برخی شهرستان ها در برخی استان ها مقدار گرفته است، به همین دلیل نمیتوانیم از این ستون نیز استفاده کنیم

۷. نوع کاربری : نمونه ای از داده های این ستون به شکل زیر است

نوع کاربری \$: chr [1:43133] "مسکونی" "مسکونی" "مسکونی" "مسکونی" ...

مقادیر تمام داده های مربوط به این ستون برابر با "مسکونی" است و اطلاعات خاصی در اختیار ما قرار نمیدهد.

۸. مساحت : مساحت ساختمان به متر مربع . نمونه ای از داده های این ستون به شکل زیر است

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

۹. درصد : درصد فروش رفته از ساختمان در معامله . نمونه ای از داده های این ستون به شکل زیر است

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

۱۰. قیمت : قیمت کل ساختمان به هزار ریال. نمونه ای از داده های این ستون به شکل زیر است

قیمت \$: num [1:43133] 3000000 1080000 10000000 3240000 750000 ...

۱۱. قیمت یک متر مربع : نمونه ای از داده های این ستون به شکل زیر است

قیمت یک مترمربع \$: chr [1:43133] "36144.58" "12000.00" "204081.63" "40039.55" ...

۱۲. عمر بنا : نمونه ای از داده های این ستون به شکل زیر است

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

۱۳. نوع اسکلت : نمونه ای از داده های این ستون به شکل زیر است

نوع اسکلت \$: chr [1:43133] "فلزی" "فلزی" "بتونی" "فلزی" ...

۱۴. تاریخ ثبت قرارداد : نمونه ای از داده های این ستون به شکل زیر است

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

اطلاعات خاصی در اختیار ما قرار نمیدهد چون این اطلاعات مربوط به یک بازه یک ماهه میباشد و در این بازه کوتاه نمیتوان تاثیر زمان بر تغییرات مختلف در داده ها را به درستی بررسی کرد و نتیجه گیری های مطمئن از آن ها استخراج کرد.

۱۵. شش رقم نخست کد پستی : نمونه ای از داده های این ستون به شکل زیر است

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

همان طور که گفته شد تعداد زیادی از این ستون ها اطلاعات خاصی به ما نمیدهند و باید قبل از کار با داده ها این اطلاعات کنار بگذاریم.

ستون هایی که در ادامه کار به آن ها نیاز داریم عبارتند از:

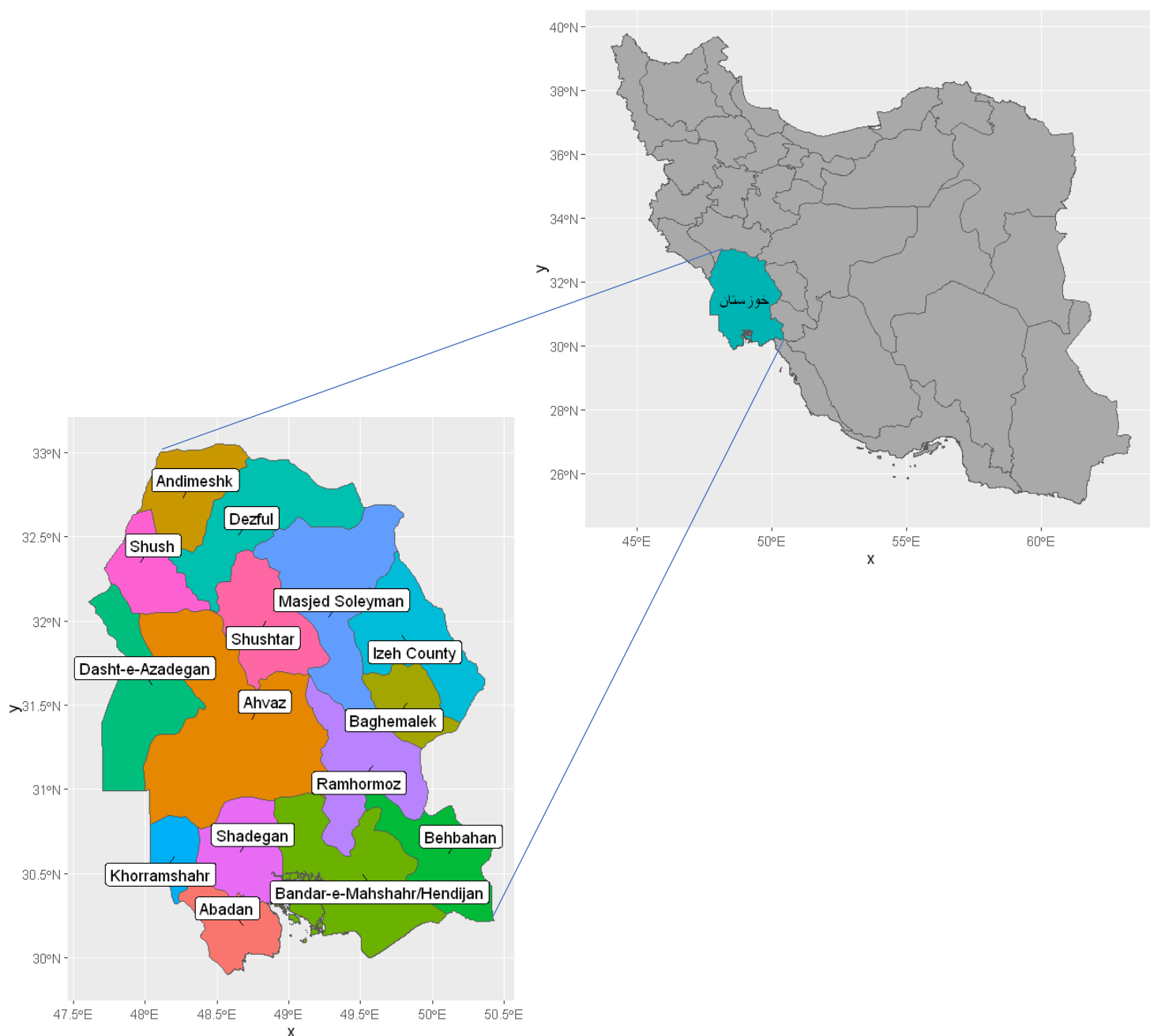
استان – شهر – مساحت – قیمت – قیمت هر متر مربع – عمر بنا – نوع اسکلت – شش رقم آخر کد پستی

که برای سادگی کار با داده ها نام این ستون ها را به ترتیب به شکل زیر در می آوریم.

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

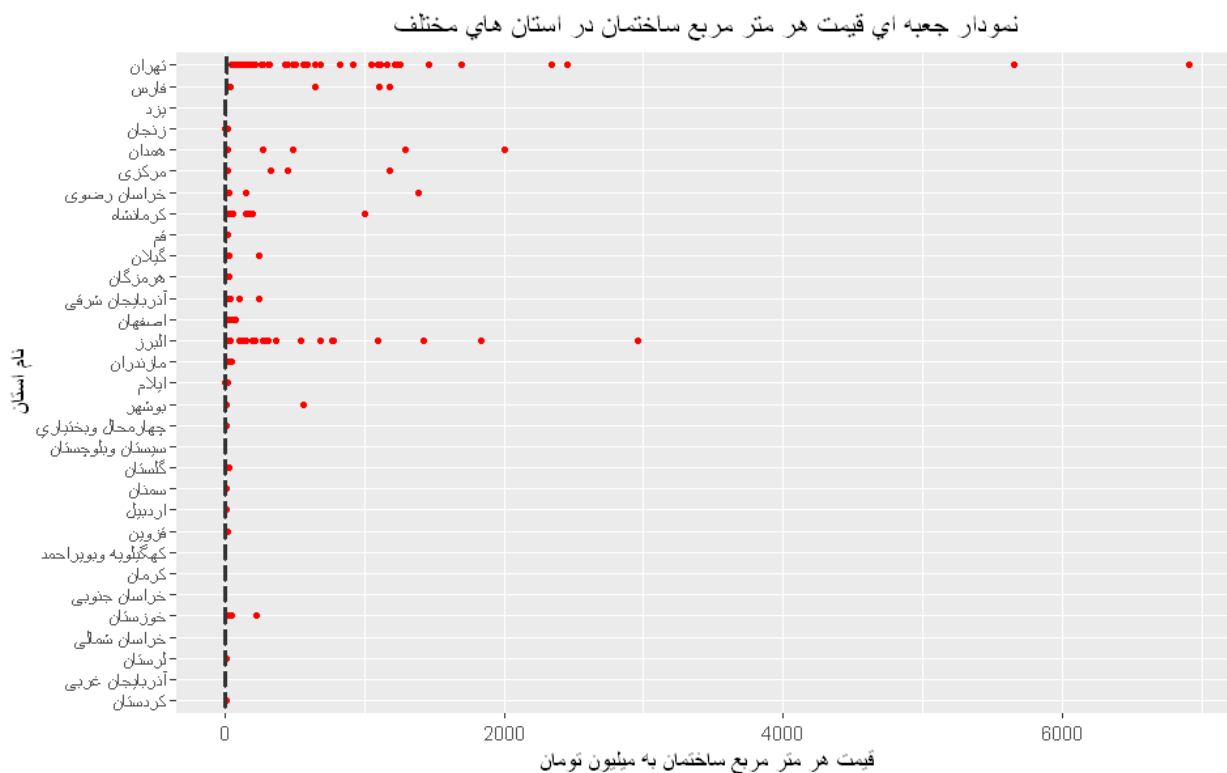
همچنین برای خوانا تر شدن اعداد موجود برای قیمت و قیمت هر متر مربع ، این اطلاعات را به شکل واحد میلیون تبدیل میکنیم.

برای شروع کار ابتدا در نقشه ایران نگاهی به موقعیت مکانی استان خوزستان و شهرستان های این استان می اندازیم.



برای بررسی وضعیت قیمت در استان خوزستان نسبت به دیگر استان ها ابتدا نمودار جعبه ای قیمت ساختمان در استان های مختلف را رسم میکنیم.

```
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))
```

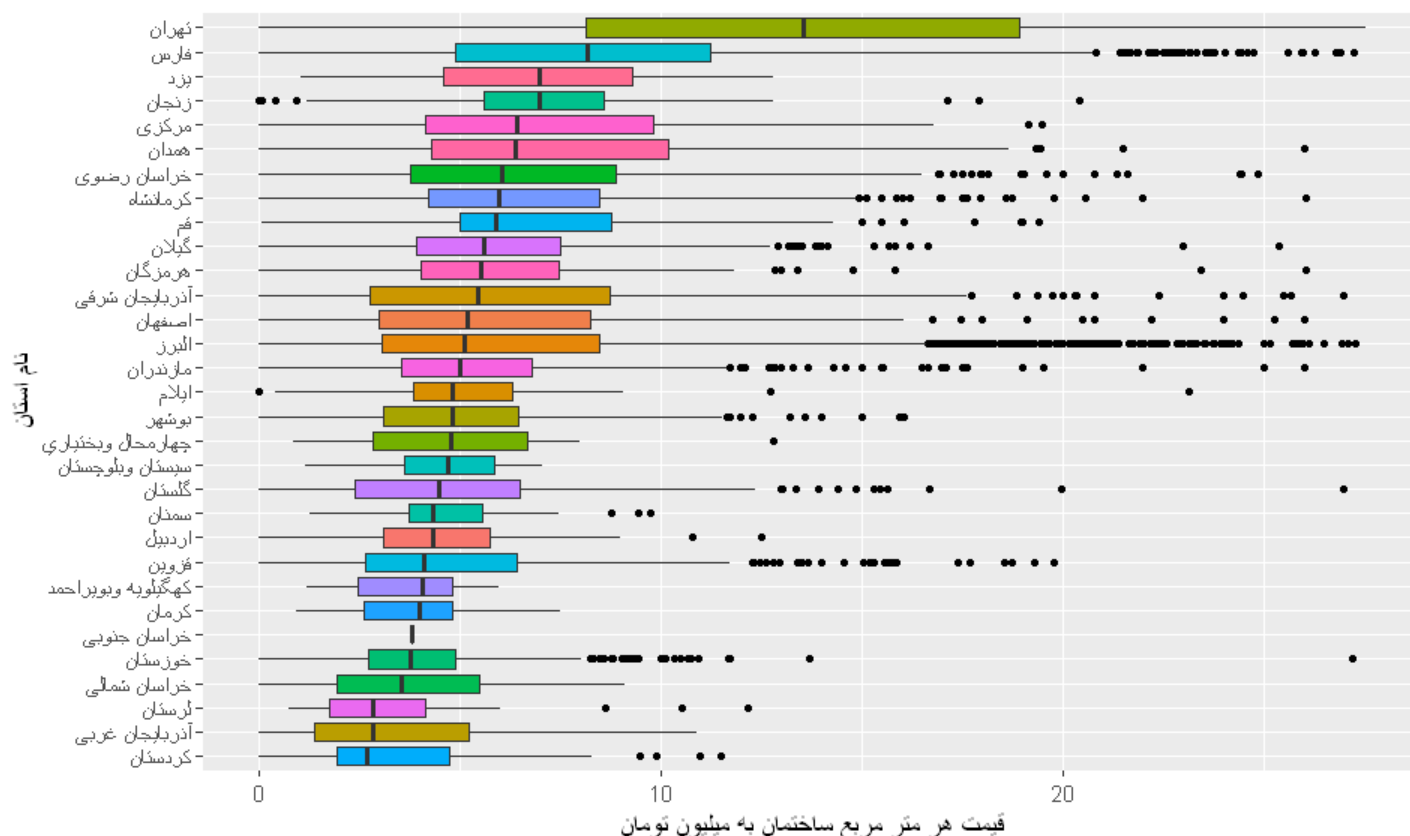


همان طور که مشاهده میکنید به دلیل وجود داده های پرت نمیتوانیم درک درستی از داده های به دست آوریم، به همین دلیل میتوانیم موقتا برخی داده های پرت را کنار بگذاریم تا بتوانیم بخش اصلی داده ها را دقیق تر بررسی کنیم، به همین منظور این نمودار را دوباره رسم میکنیم با این تفاوت که به جای رسم اطلاعات مربوط به تمام داده ها، اطلاعات داده هایی را رسم میکنیم که قیمت ساختمان در آن ها در ۹ دهک اول قیمتی وجود داشته باشد، به این ترتیب داده های بسیار بزرگی که وجود دارد را کنار میگذاریم.

```
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))
```

Quantile دستوری است که برای این کار استفاده میکنیم تا مرز بین دهک نهم و دهم را پیدا کنیم و داده های را از آن جا به دو بخش تقسیم کنیم. از دستور reorder نیز استفاده میکنیم تا این نمودار های جعبه ای را به ترتیب میانه در نمودار بچینیم تا دید بهتری به ما دهد.

نمودار جعبه‌ای قیمت هر متر مربع (9 دهک اول) ساختمان‌های مختلف



همان طور که مشاهده میکنید با این کار توانستیم درک بهتری نسبت به داده‌ها پیدا کنیم و همان طور که از نمودار واضح است با توجه به میانه قیمت‌ها در استان‌های مختلف، استان خوزستان در رده پنجم استان‌های ارزان کشور است.

برای ادامه کار قصد داریم داده‌های مربوط به استان خوزستان را دقیق‌تر بررسی کنیم بنابراین داده‌هایی که مقدار آن‌ها در ستون state برابر با "خوزستان" است را جدا میکنیم.

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

در ابتدا قصد داریم همبستگی بین ستون‌ها را بررسی کنیم، برای این کار نیاز داریم تا تمام داده‌های کیفی را به کمک dummy variable‌ها به شکل داده‌های کمی دربیاوریم تا بتوانیم مقدار correlation را برای آن‌ها به دست بیاوریم. داده‌های کیفی این dataframe در ستون‌های city و building_type میباشد. برای این کار از دستور زیر استفاده میکنیم.

```
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)
```

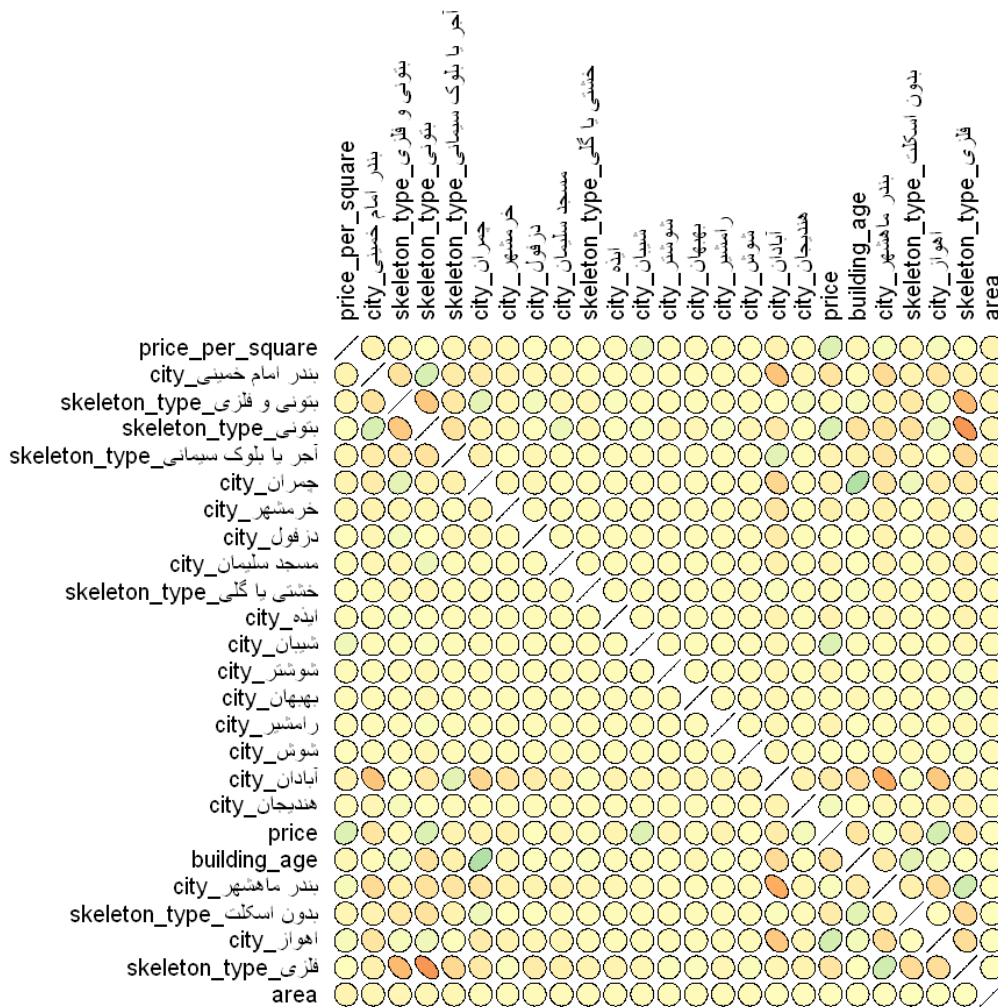
سپس با دستور `data <- cor(df)` ماتریس correlation را برای این داده‌ها به دست می‌آوریم.

برای درک بهتر این ماتریس میتوانیم از نمودار زیر استفاده کنیم.

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

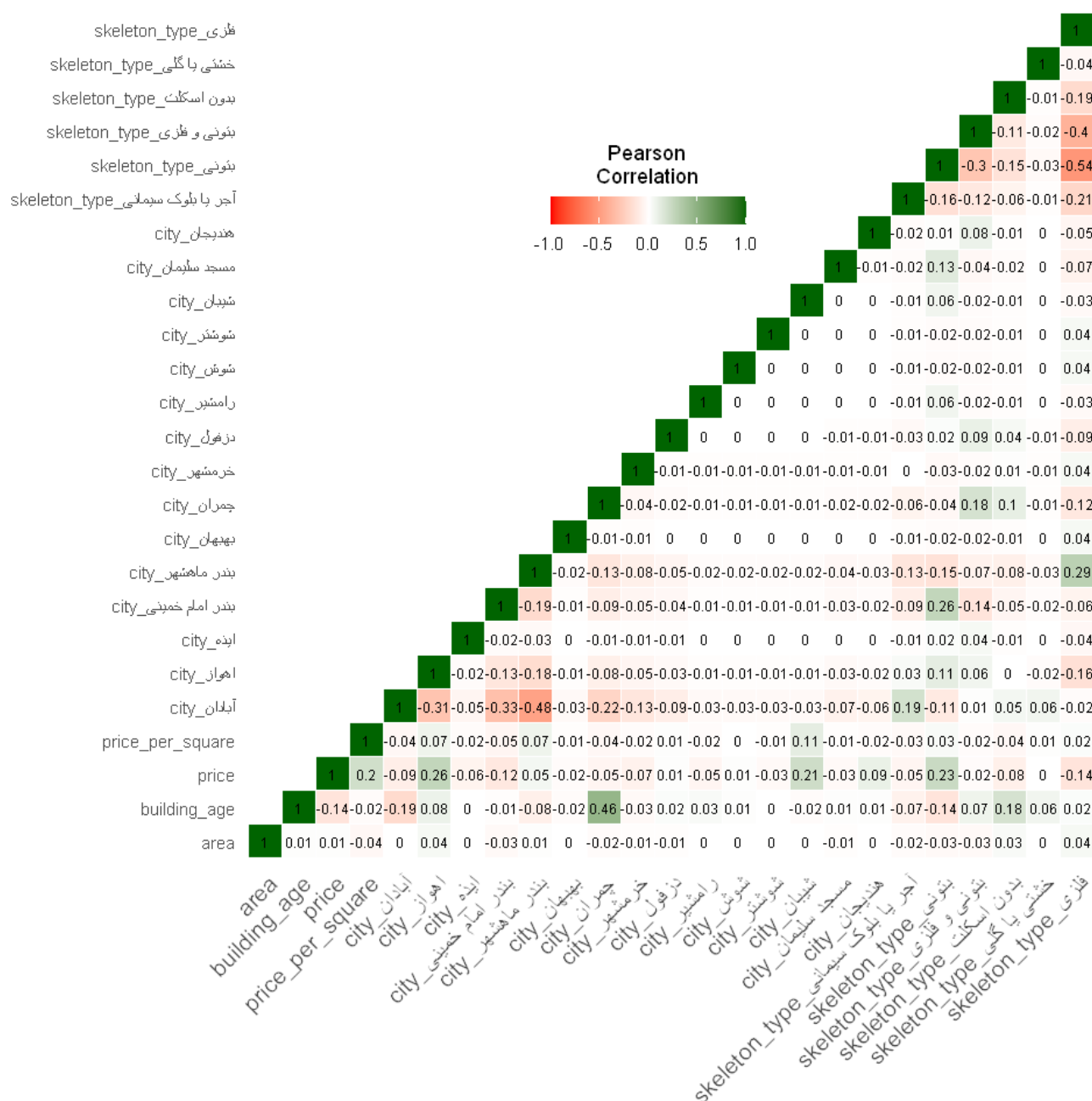
# Build a Pannel of 100 colors with Rcolor Brewer
my_colors <- brewer.pal(5, "Spectral")
my_colors <- colorRampPalette(my_colors)(100)

# 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 این نمودار هر چه رنگ بیضی مربوطه به سبز نزدیک تر باشد دو متغیر رابطه مستقیم قوی تری دارند و هر چه به رنگ قرمز نزدیک تر باشد دو متغیر رابطه معکوس قوی تری دارند و بخش هایی که رنگ آن به زرد نزدیک است یعنی دو متغیر هیچ رابطه ای با یکدیگر ندارند.

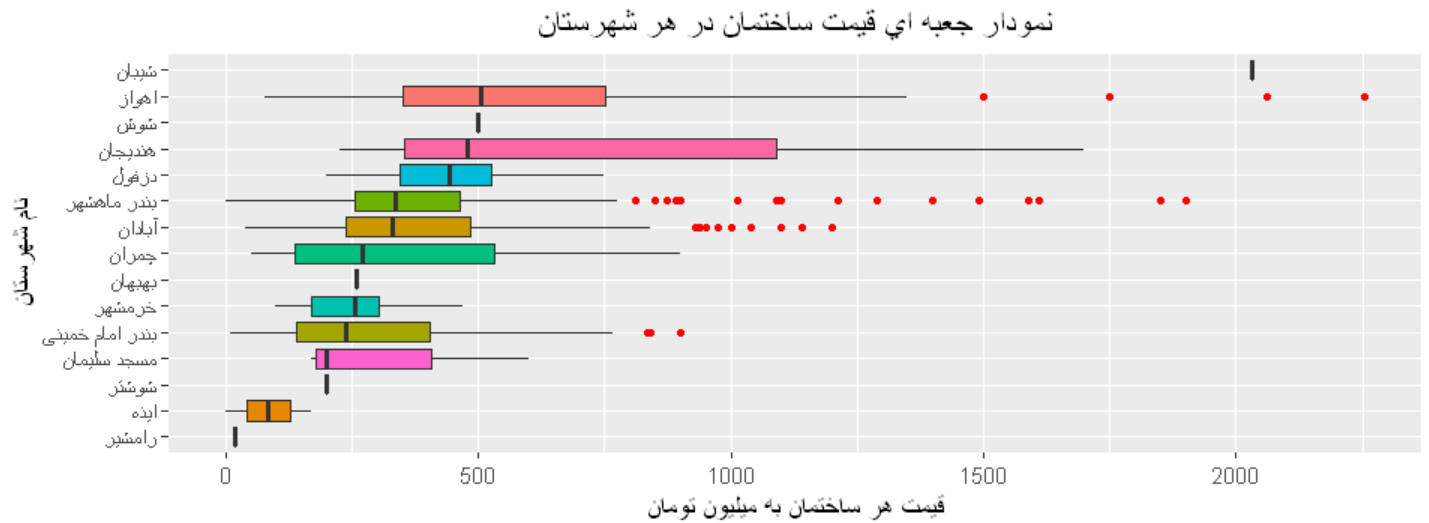
برای مشاهده دقیق تر اعداد شاخص همبستگی میتوانیم از نمودار زیر نیز استفاده کنیم.



حالا نمودار جعبه ای قیمت ساختمان در شهرستان های مختلف را با دستور زیر رسم میکنیم.

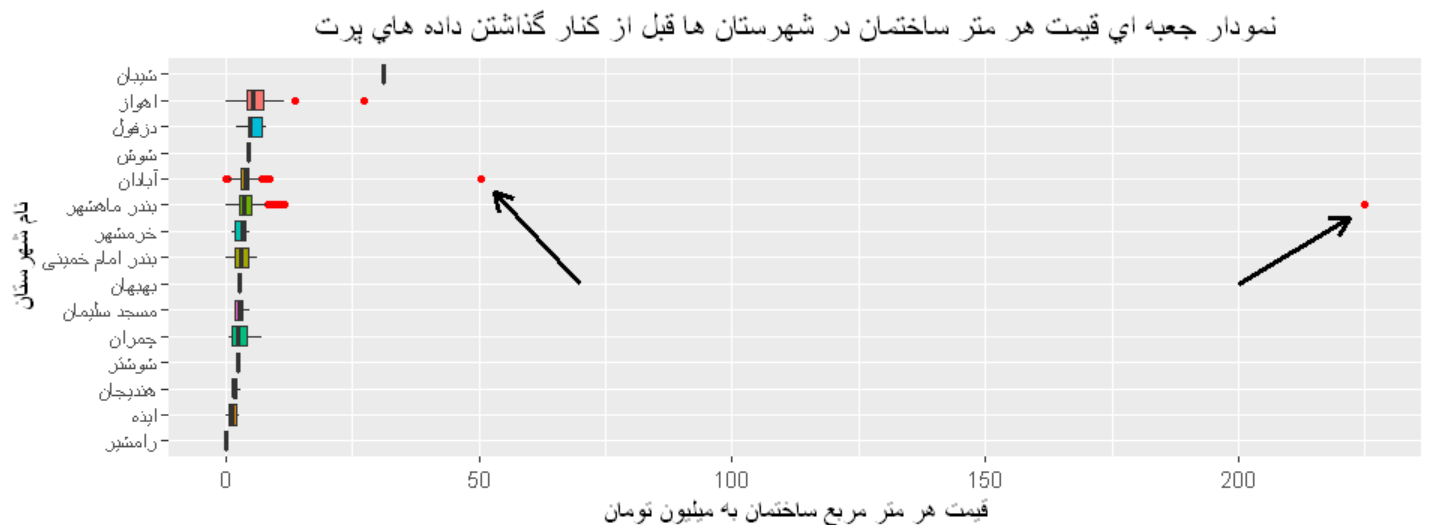
```
options(repr.plot.width = 8, repr.plot.height = 3)
ggplot(khoozestan, aes(x=price, y=reorder(city, price, FUN=median) , fill=city)) +
  geom_boxplot(outlier.colour="red", outlier.shape=16,outlier.size=1) +
  labs(title="نمودار جعبه ای قیمت ساختمان در هر شهرستان", x="قیمت هر ساختمان به میلیون تومان", 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 ی پرت باعث میشود نتوانیم به خوبی اطلاعات بقیه داده ها را مشاهده و بررسی کنیم به همین دلیل موقتا این دو داده را از اطلاعات کنار میگذاریم اما مشخصات مربوط به آن ها را حذف نمیکنیم تا بتوانیم در صورت نیاز به دقت این داده ها را بررسی کنیم، لازم به ذکر است که قصد ما از حذف کردن این داده ها غلط نشان دادن آن ها و یا بی اهمیت نشان دادنشان نیست و فقط برای درک بهتر بقیه داده ها این کار را انجام میدهیم.

کد قرار داد در این دو داده ۱۹۱۸۹۱۹۸ و ۱۹۲۳۶۳۰۳ بود.

مجددا این نمودار را رسم میکنیم.



به این شکل میتوانیم دقیق تر این داده ها را بررسی کنیم.

مشکلی که در این جا مواجه آن میشویم که در نمودار قبلی قطعاً نمیتوانستیم متوجه آن شویم این است که تعداد داده ها در بعضی از شهرستان ها بسیار کم است و این موضوع سطح اطمینان ما به این داده ها را کم میکند. شهرستان هایی مانند شیبان ، شوش ، شوشتر ، بهبهان و رامشیر.

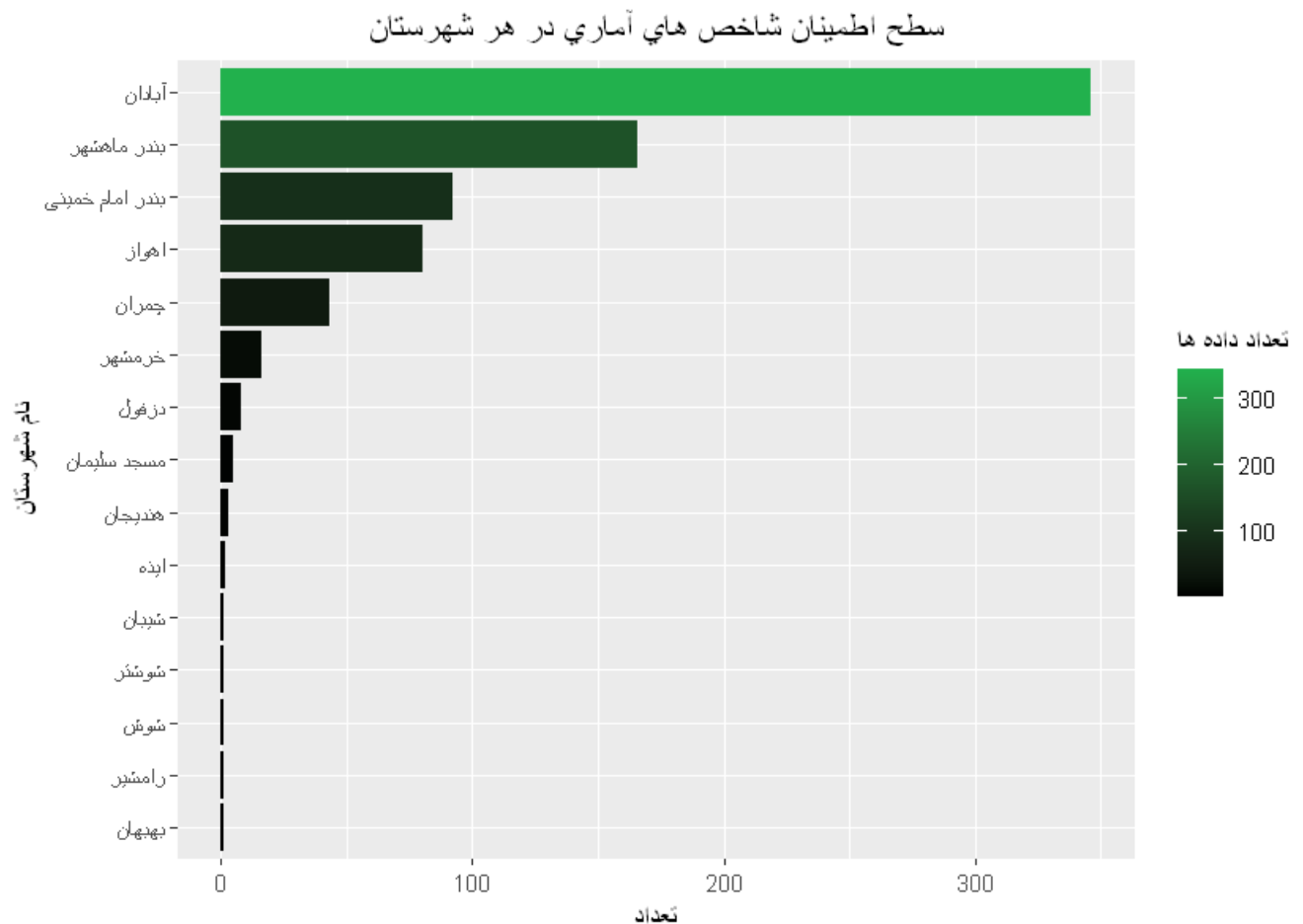
برای بررسی دقیق تر این موضوع ابتدا تعداد داده های مربوط به هر شهرستان را شمارش میکنیم.

```
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 خمینی	92
بندر ماهشهر	166
آبادان	346

برای درک بهتر این موضوع این اعداد را در یک نمودار میله ای نمایش میدهم. هر چه تعداد داده های مربوط به یک شهرستان بیشتر باشد، سطح اطمینان ما به شاخص های آماری مربوط به آن شهرستان نیز افزایش پیدا میکند.

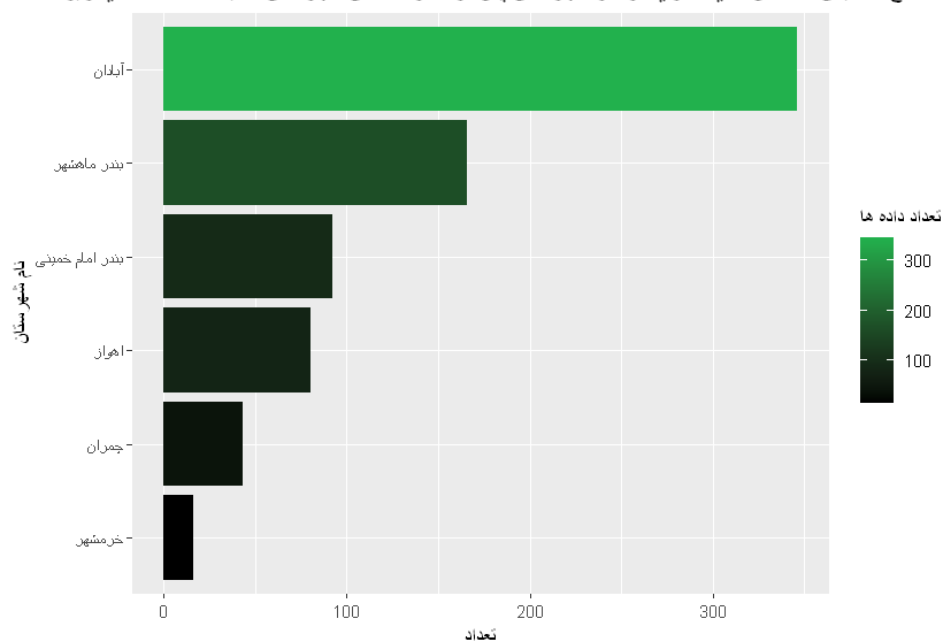
```
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(confidence_level , aes(x=n , y=reorder(city,n), fill = n)) +
  geom_bar(stat = "identity") +
  labs(title="سطح اطمینان شاخص های آماری در هر شهرستان", x="تعداد", y = "نام شهرستان") +
  scale_fill_gradient(name = "تعداد داده ها",low="black",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



همان طور که مشاهده میکنید علاوه بر بهبان ، رامشیر ، شوش ، شوشتر و شیبلیان ، تعداد داده های شهرستان های ایذه ، هندیجان ، مسجد سلیمان و دزفول نیز زیر ۱۰ میباشد و به همین دلیل سطح اطمینان شاخص های آماری در این داده ها بسیار پایین است و به همین دلیل داده های مربوط به این شهرستان ها را کنار میگذاریم تا بتوانیم بقیه داده ها را دقیق تر بررسی کنیم.

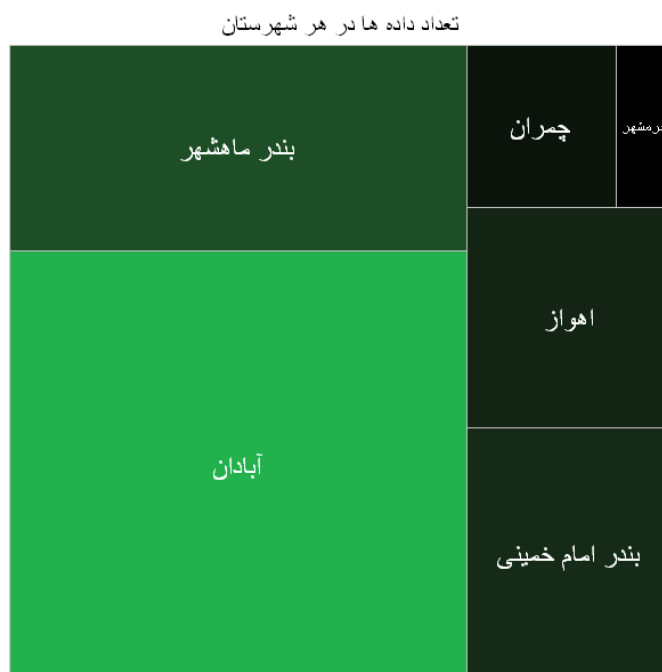
مجددا بر این نکته تاکید میکنم که تمام این داده ها برای ما مهم هستند و باید اطلاعات هر یک مورد بررسی قرار گیرد.

سطح اطمینان شاخص های آماری در هر شهرستان پس از کنار گذاشتن شهرستان ها با تعداد داده های زیر 10



همچنین برای این کار می‌توانیم از نمودار Treemap استفاده کنیم.

```
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))
```



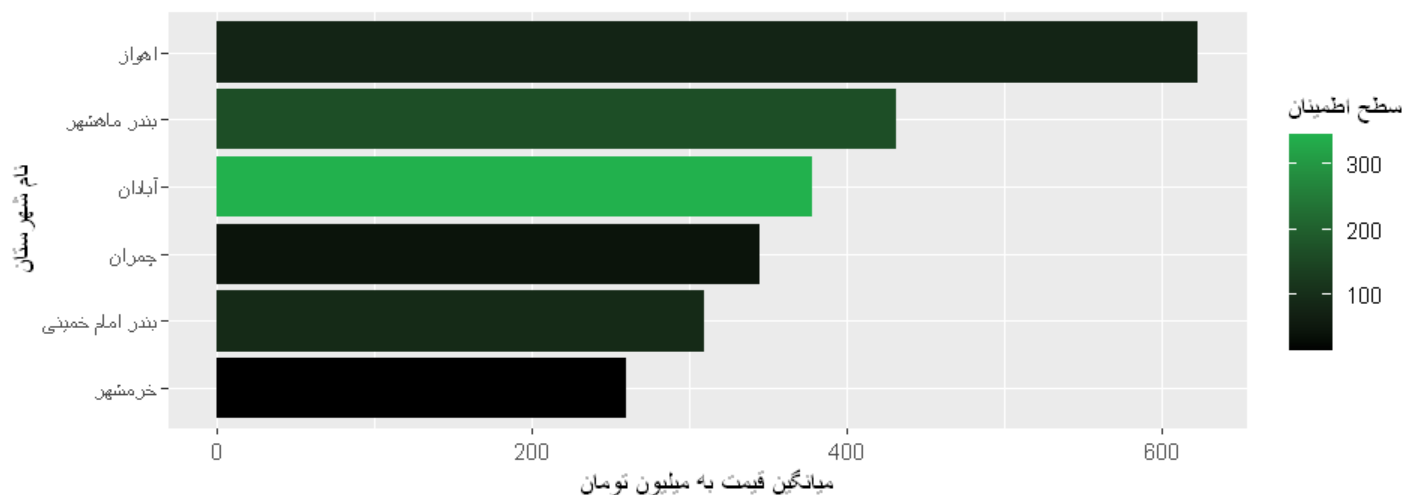
city	mean_price
اهواز	623.4250
آبادان	378.0111
بندر امام خمینی	309.6426
بندر ماهشهر	431.9117
چمران	344.5581
خرمشهر	259.9812

برای مقایسه میانگین قیمت بین این چند شهرستان از نمودار میله ای استفاده میکنیم.

```
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="میانگین قیمت هر شهرستان", x="میانگین قیمت به میلیون تومان", y = "نام شهرستان") +
  scale_fill_gradient(name="سطح اطمینان",low="black", high="#22b14d") +
  theme(plot.title = element_text(hjust = 0.5))
```

میانگین قیمت هر شهرستان

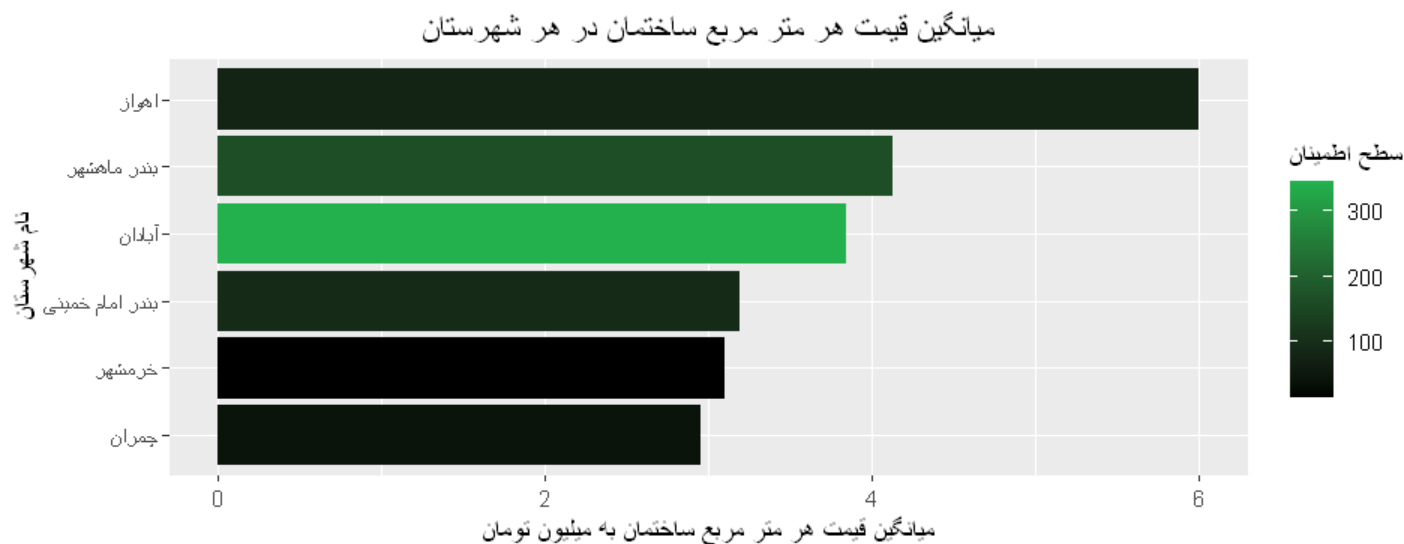


city	mean_price_per_square
اهواز	6.002169
آبادان	3.839624
بندر امام خمینی	3.195575
بندر ماهشهر	4.127006
چمران	2.949167
خرمشهر	3.095103

همین نمودار را برای اطلاعات مربوط به قیمت هر متر مربع نیز رسم میکنیم.

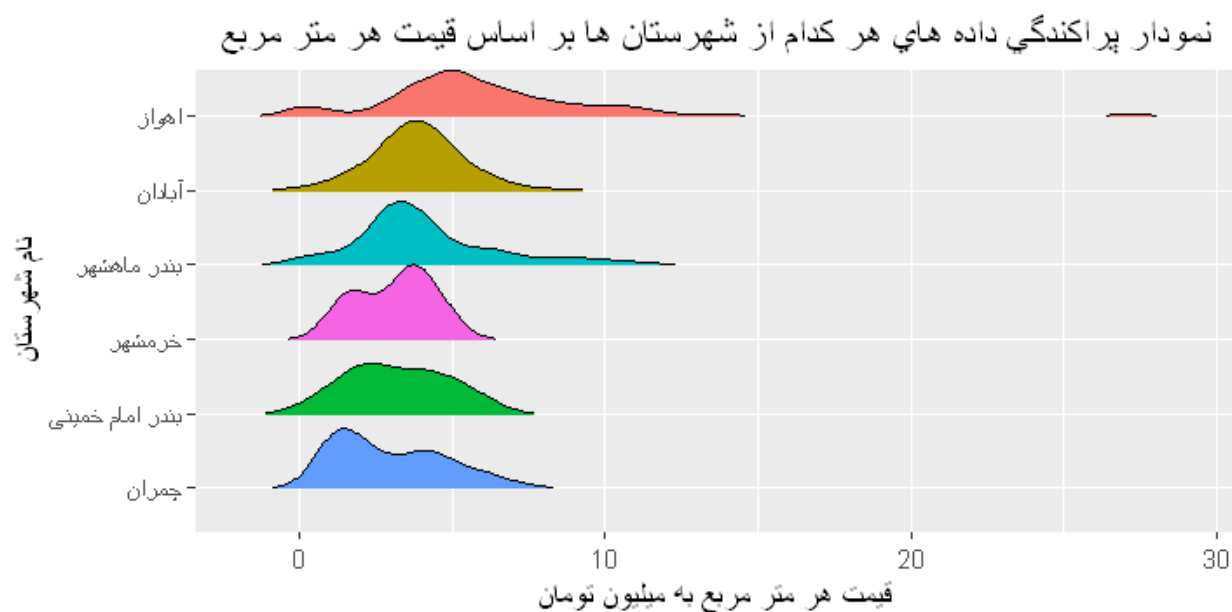
```
df <- khoozestan %>%
  select(city,price_per_square) %>%
  group_by(city) %>%
  summarise(mean_price_per_square = mean(price_per_square))
df
```

```
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="میانگین قیمت هر متر مربع ساختمان در هر شهرستان", x="میانگین قیمت به میلیون تومان", y = "نام شهرستان") +
  scale_fill_gradient(name="سطح اطمینان",low="black", high="#22b14d") +
  theme(plot.title = element_text(hjust = 0.5))
```



پراکندگی داده ها در هر یک شهرستان ها بر اساس قیمت هر متر مربع ساختمان به شکل زیر 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))
```



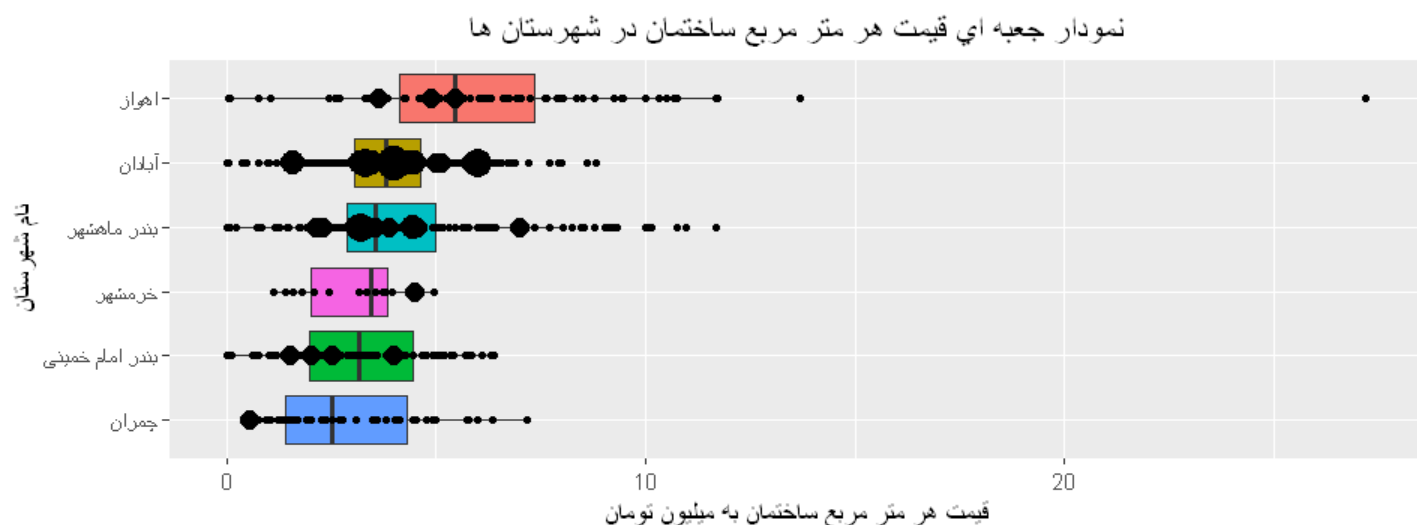
میتوان تعداد داده های موجود در هر قیمت را به کمک دسَنور geom_count روی نمودار جعبه ای نشان داد که در نمودار زیر میتوانید آن را مشاهده کنید.

```
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 ی دیگری از همین نمودار برای قیمت هر متر مربع را میتوانیم به شکل زیر رسم کنیم.

```
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))
```

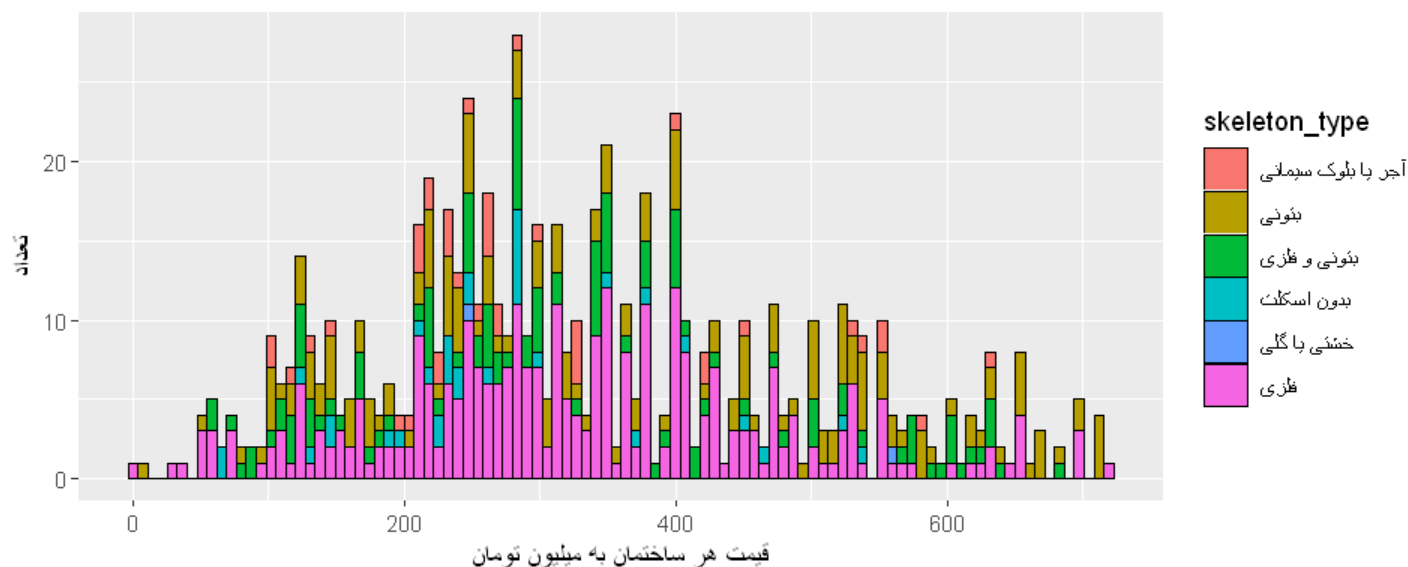


حالا قصد داریم تاثیر اسکلت ساختمان بر قیمت بنا را بررسی کنیم.

نمودار اولی که برای این منظور رسم میکنیم که نمودار پراکندگی قیمت ساختمان ها بر اساس قیمت آن ها میباشد.

```
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))
```

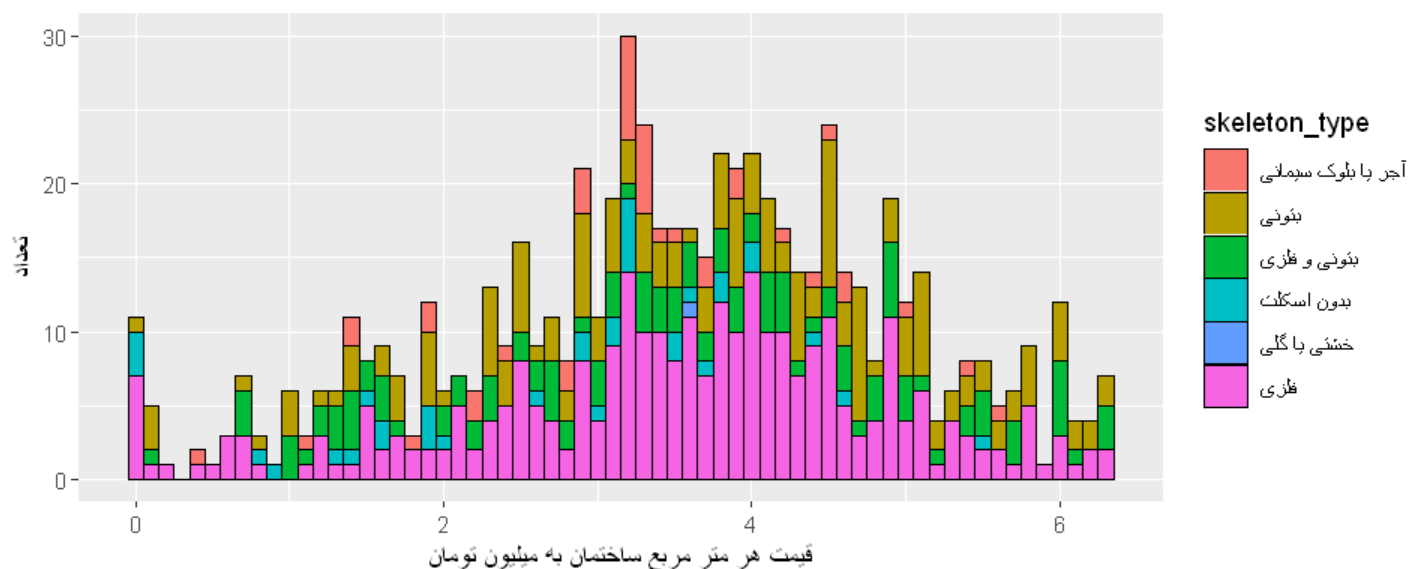
نمودار بافت نگار قیمت (9 دهک اول) ساختمان ها قیمت به تفکیک نوع اسکلت ساختمان



و سپس همین نمودار را برای قیمت هر متر مربع رسم میکنیم.

```
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))
```

نمودار بافت نگار قیمت هر متر مربع (9 دهک اول) ساختمان ها قیمت به تفکیک نوع اسکلت ساختمان



در این دو نمودار برای درک بهتر از داده ها تنها ۹ دهک اول داده در آن متغیر (قیمت یا قیمت هر متر مربع) را رسم کرده ایم.

در نمودار اول کل ۹ دهک اول را به ۱۰۰ بازه از قیمت تقسیم کرده ایم و تعداد داده های مربوط به هر نوع از اسکلت ها را در آن نشان داده ایم.

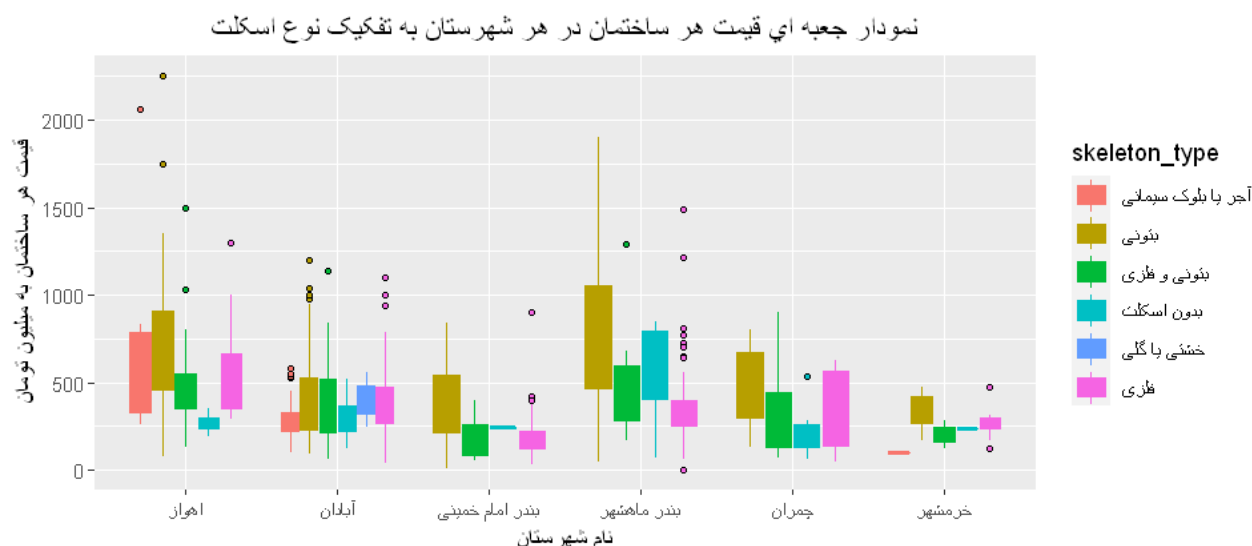
این کار را با قرار دادن مقدار bins برابر با ۱۰۰ انجام داده ایم.

در نمودار دوم کل ۹ دهک اول را به ازای هر ۱۰۰ هزار تومان یک بازه جدید ایجاد کرده ایم و تعداد داده های مربوط به هر نوع از اسکلت ها را در آن نشان داده ایم.

این کار را با قرار دادن مقدار binwidth برابر با ۰.۱ انجام داده ایم چون واحد ما در این داده ها میلیون تومان است پس ۰.۱ آن برابر با ۱۰۰ هزار تومان است.

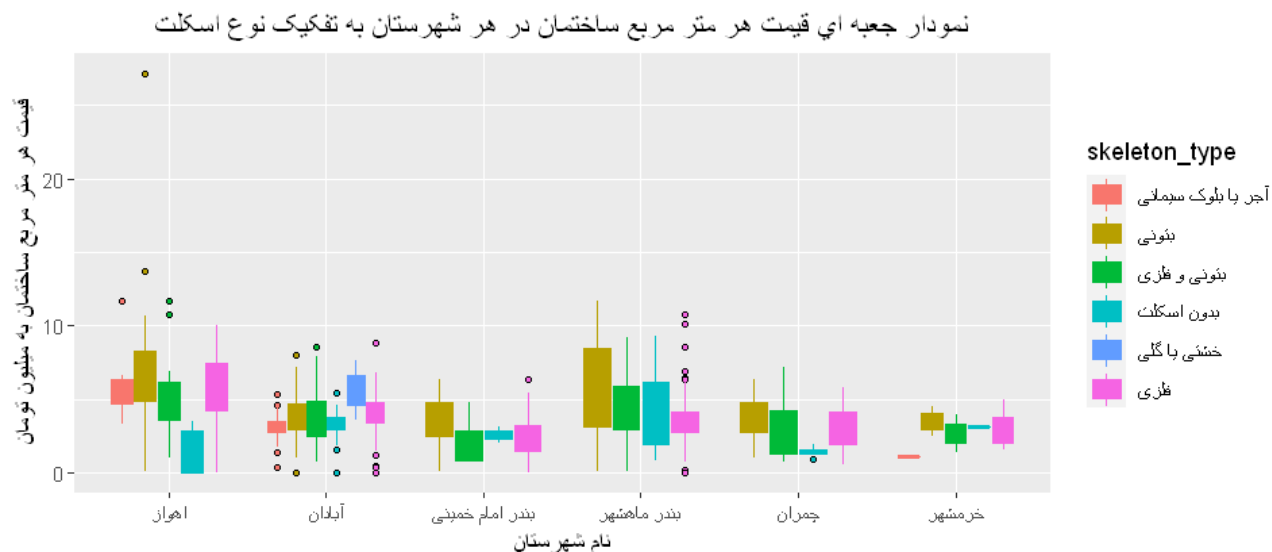
نمودار بعدی که رسم میکنیم نمودار جعبه ای قیمت شهرستان ها به تفکیک نوع اسکلت در آن ها میباشد.

```
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 این نمودار را برای قیمت هر متر مربع ساختمان نیز رسم میکنیم.

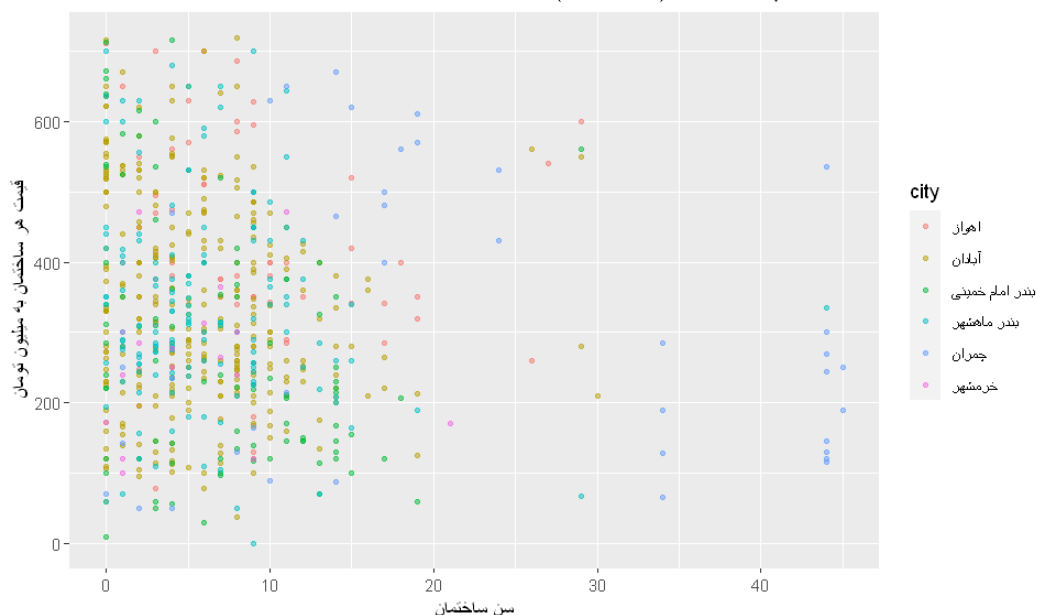
```
options(repr.plot.width = 8, repr.plot.height = 3.5)
ggplot(khoozestan, aes(x=city, y=price , fill = skeleton_type)) +
  geom_boxplot(aes(colour = skeleton_type), outlier.colour="black", outlier.shape=21, outlier.size=1) +
  labs(title="نمودار جعبه ای قیمت هر ساختمان در هر شهرستان به تفکیک نوع اسکلت",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



برای بررسی تاثیر سن ساختمان روی قیمت ساختمان نمودار پراکنشی داده های این دو ستون را به تفکیک شهرستان رسم میکنیم.

```
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))
```

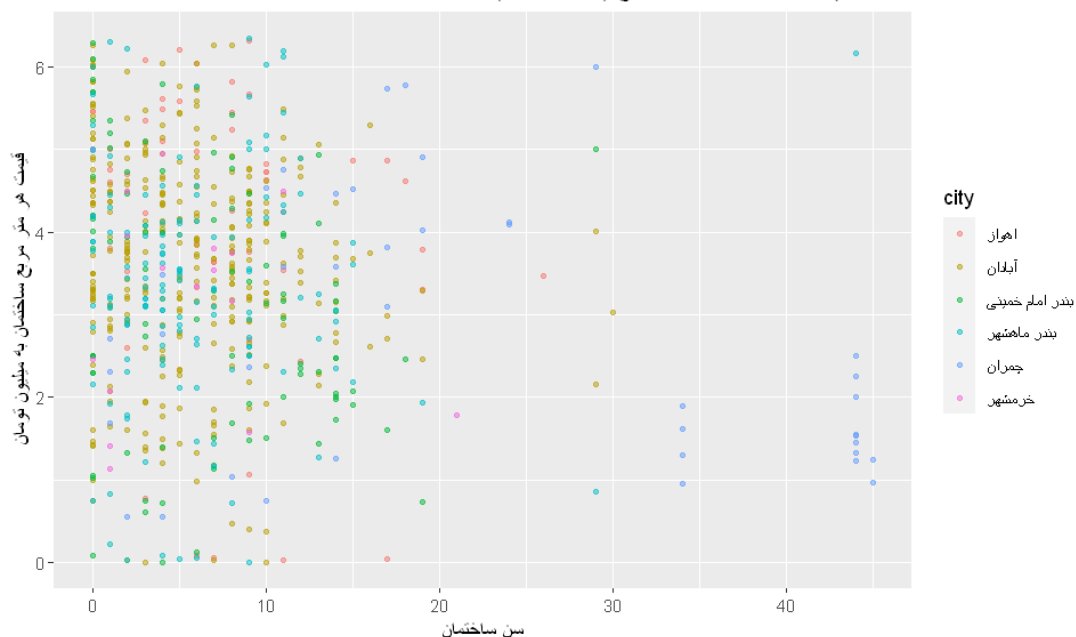
نمودار پراکنش قیمت (9 دهک اول) هر ساختمان بر اساس سن ساختمان به تفکیک شهرستان



و سپس این نمودار را برای قیمت هر متر مربع نیز رسم میکنیم.

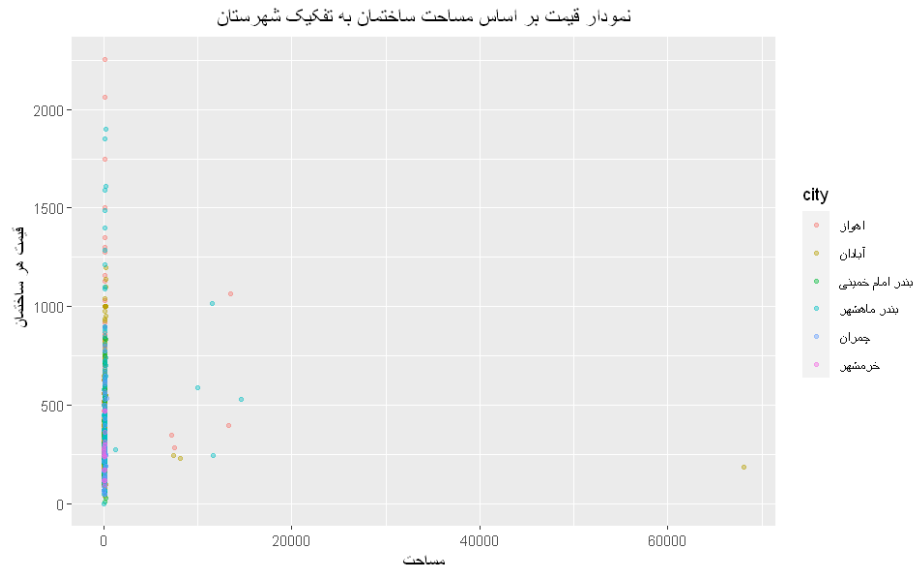
```
options(repr.plot.width = 8, repr.plot.height = 5)
ggplot(khoozestan[khoozestan$price_per_square < quantile(khoozestan$price_per_square,0.9)], aes(x=building_age, y=price_per_square , color = city)) +
  geom_point(size=1,alpha=0.5) +
  labs(title="نمودار پراکنش قیمت هر متر مربع (9 دهک اول) ساختمان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

نمودار پراکنش قیمت هر متر مربع (9 دهک اول) ساختمان بر اساس سن ساختمان به تفکیک شهرستان



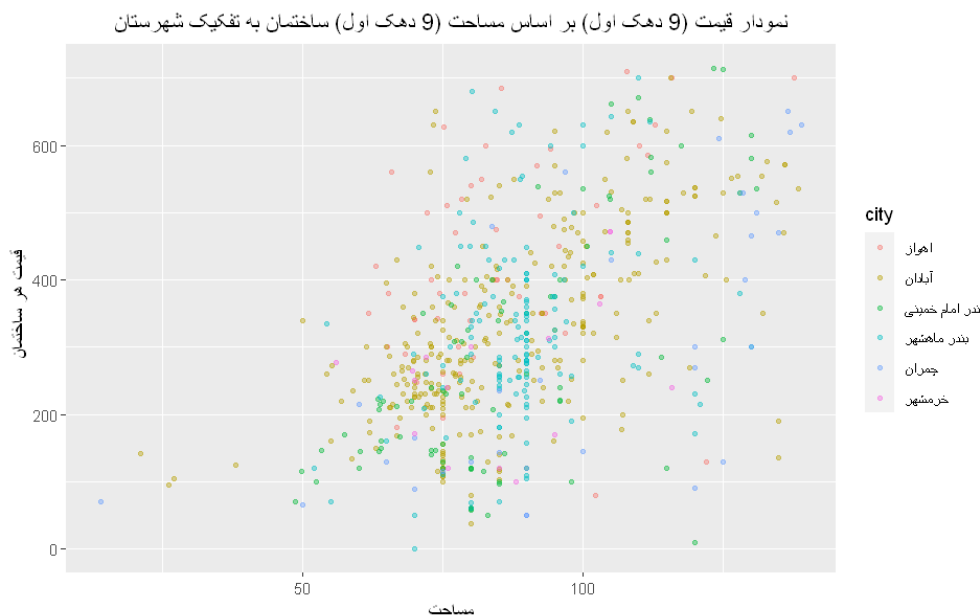
سپس برای بررسی تاثیر مساحت روی قیمت ساختمان نمودار زیر را رسم میکنیم.

```
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 قیمت بر اساس مساحت ساختمان به تفکیک شهرستان", y="قیمت هر ساختمان", x = "مساحت")+
  theme(plot.title = element_text(hjust = 0.5))
```



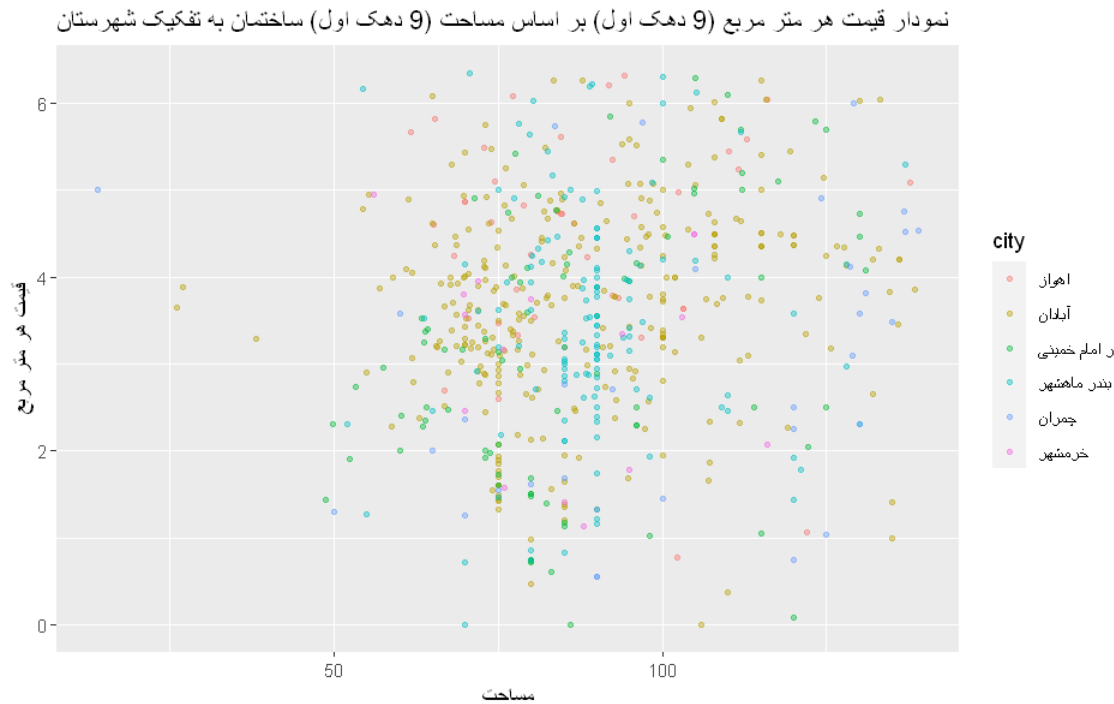
همان طور که مشاهده میکنید در این جا نیز به دلیل وجود داده های پرت و ساختمان هایی با مساحت های بسیار بالا بررسی بخش عمده ای از داده ها عملاً غیر ممکن شده است بنابراین در این جا نیز تنها ۹ دهک اول داده ها را بررسی میکنم. در تمام بخش های این گزارش که تنها ۹ دهک از داده ها نمایش داده شده است به همین دلیل است که در غیر این صورت مصور سازی این داده ها نمیتوانست هیچ اطلاعاتی در اختیار 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 قیمت (9 دهک اول) بر اساس مساحت (9 دهک اول) ساختمان به تفکیک شهرستان", y="قیمت هر ساختمان", x = "مساحت")+
  theme(plot.title = element_text(hjust = 0.5))
```



سپس همین نمودار را برای قیمت هر متر مربع نیز رسم میکنیم.

```
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 قیمت هر متر مربع (9 دهک اول) بر اساس مساحت (9 دهک اول) ساختن به تفکیک شهرستان", y="قیمت هر متر مربع", x = "مساحت")
```



در این گزارش سعی کردیم تا داده های مربوط به خرید و فروش مسکن در استان خوزستان را بررسی کنیم ، به عوامل تاثیر گذار روی قیمت مسکن پرداختیم و سعی کردیم تاثیر هر یک را بر شهرستان های مختلف استان خوزستان آشکار کنیم تا برای شخصی که قصد دارد برای اولین بار با این داده ها کار کند دید کافی را ایجاد کنیم تا بتواند دقیق تر این داده ها را مورد بررسی قرار دهد ، قطعاً نکات مهم دیگری نیز در این داده ها وجود دارد که بنده نتوانستیم به خوبی آن ها را نمایش دهیم. این داده ها مشکلاتی نیز داشتند که میتوان به کم بودن تعداد داده ها در شهرستان های مختلف و کم بودن پراکندگی داده ها در برخی از ستون ها اشاره کرد که باعث شد اطلاعات خیلی زیادی برای نمایش و بررسی موجود نباشد.

پینوشت : متأسفانه اطلاع نداشتیم که باید برای این تمرین گزارش تهیه کنم به همین دلیل گزارش را با تاخیر خدمتتان ارسال کردم. با احترام کنعانی