CÁC PHƯƠNG PHÁP ĐỊNH LƯỢNG 1

BÀI TẬP 1

Lời giải gợi ý

Câu 1. Anh/Chị hãy phân loại các biến trong bộ dữ liệu theo loại, bao gồm biến định tính (nominal), biến thứ tự (ordinal), biến khoảng (interval), và biến tỷ lệ (ratio).

Tên cột	**Mô tả**	**Biến**
Province	Tên tỉnh/thành	Định danh
Entry_cost	Chỉ số thành phần - Chi phí gia nhập thị trường thấp	Tỷ lệ
Access_to_land	Chỉ số thành phần - Tiếp cận đất đai dễ dàng	Tỷ lệ
Transparency	Chỉ số thành phần - Môi trường kinh doanh minh bạch và thông tin kinh doanh công khai	Tỷ lệ
Time_costs	Chỉ số thành phần - Thời gian thanh tra, kiểm tra và thực hiện các quy định thủ tục hành chính nhanh chóng	Tỷ lệ
Informal_charges	Chỉ số thành phần - Chi phí không chính thức thấp	Tỷ lệ
Business_environment	Chỉ số thành phần - Môi trường cạnh tranh bình đẳng	Tỷ lệ
Proactivity	Chỉ số thành phần - Chính quyền địa phương năng động sáng tạo trong giải quyết vấn đề cho doanh nghiệp	Tỷ lệ
Business_support	Chỉ số thành phần - Các chính sách hỗ trợ doanh nghiệp hiệu quả	Tỷ lệ
Labour_policy	Chỉ số thành phần - Chính sách đào tạo lao động tốt	Tỷ lệ
Law_and_order	Chỉ số thành phần - Thủ tục giải quyết tranh chấp công bằng, hiệu quả và an ninh trật tự được duy trì	Tỷ lệ
PCI_score	Điểm PCI - Tổng có trọng số các chỉ số thành phần	Tỷ lệ
Ranking	Xếp hạng PCI (theo điểm PCI)	Thứ tự
Tier	Xếp loại PCI (dựa trên độ lệch chuẩn của điểm PCI)	Định danh

Câu 2. Anh/Chị hãy tính các trị thống kê mô tả cơ bản của từng chỉ số thành phần của 63 tỉnh/thành và bình luận ngắn gọn về các trị thống kê này.

Sử dụng công cụ Data Analysis trong Excel, ta có bảng tổng hợp các trị thống kê mô tả cơ bản của từng chỉ số thành phần của 63 tỉnh/thành như sau:

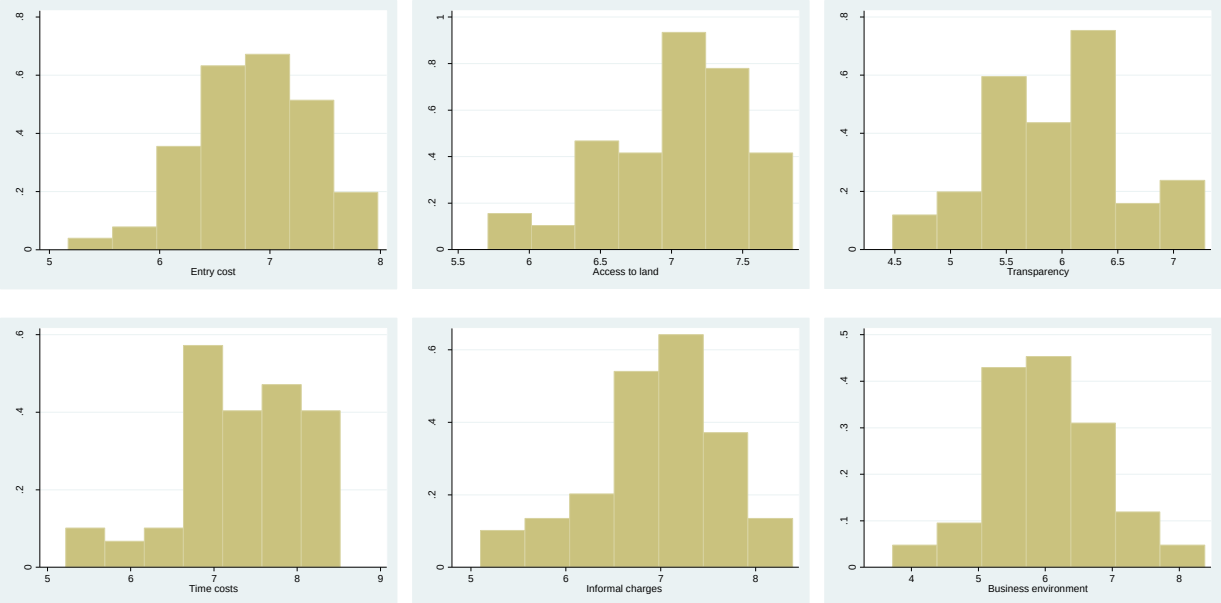
	Entry cost	Access to land	Trans-parency	Time costs	Informal charges	Business environment	Pro-activity	Business support	Labour policy	Law and order
Mean	6,86	7,05	5,99	7,44	7,07	6,06	6,88	6,78	5,93	7,15
Median	6,93	7,11	6,05	7,54	7,12	6,01	6,85	6,95	5,89	7,21

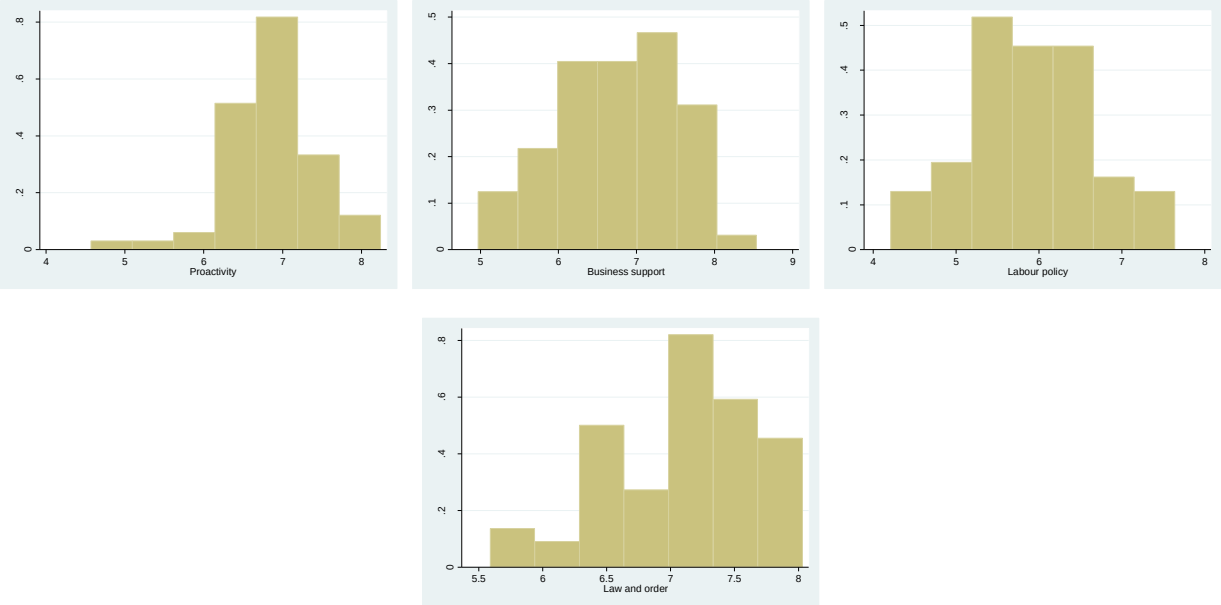
	Entry cost	Access to land	Trans-parency	Time costs	Informal charges	Business environment	Pro-activity	Business support	Labour policy	Law and order
Mode	7,02	7,56	5,43	7,64	6,70	6,96	7,27	7,57	6,21	7,87
Std	0,54	0,44	0,62	0,69	0,63	0,84	0,58	0,78	0,72	0,55
Variance	0,30	0,20	0,38	0,48	0,39	0,71	0,34	0,61	0,51	0,30
Kurtosis	0,52	-0,33	-0,15	0,50	0,74	0,78	3,74	-0,20	-0,08	-0,34
Skewness	-0,50	-0,48	-0,05	-0,52	-0,52	0,03	-0,98	-0,42	0,02	-0,53
Range	2,81	1,91	2,75	3,30	3,29	4,66	3,67	3,57	3,43	2,22
Minimum	5,17	5,94	4,53	5,22	5,10	3,72	4,57	4,97	4,21	5,81
Maximum	7,98	7,85	7,28	8,52	8,39	8,38	8,24	8,54	7,64	8,03

Sau đây là một số gợi ý bình luận, học viên có thể đưa ra các bình luận khác miễn sao hợp lý và thuyết phục:

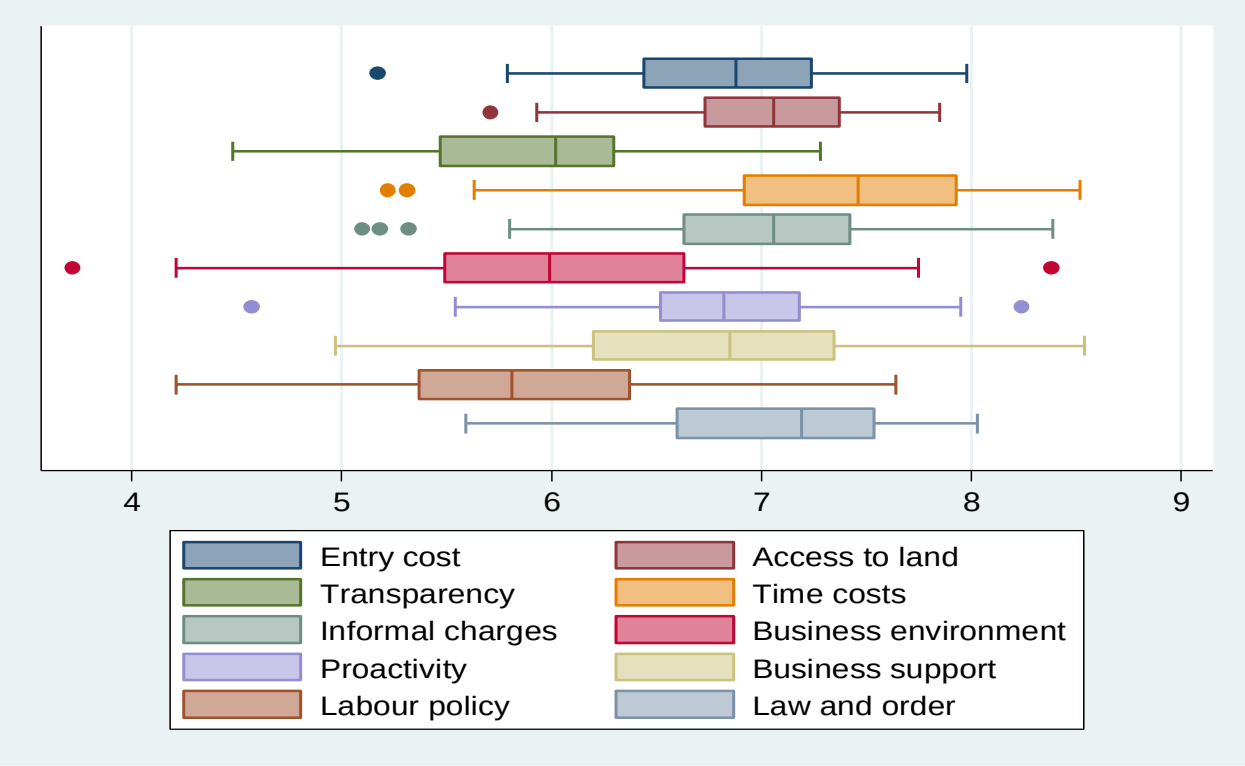
- Xét theo điểm trung bình: '*Thời gian thanh tra, kiểm tra và thực hiện các quy định thủ tục hành nhanh chóng*' là chỉ số có điểm trung bình cao hơn các chỉ số còn lại, kế đến lần lượt là các chỉ số: '*Thủ tục giải quyết tranh chấp công bằng, hiệu quả và an ninh trật tự được duy trì*'; '*Chi phí không chính thức thấp*'; '*Tiếp cận đất đai dễ dàng*'. '*Chính sách đào tạo lao động tốt'* là chỉ số có điểm trung bình thất, điều này cho thấy chất lượng đào tạo lao động chưa đáp ứng được các nhu cầu của doanh nghiệp.
- Xét theo độ lệch chuẩn và các giá trị min/max: Các chỉ số '*Môi trường cạnh tranh bình đẳng*'; '*Các chính sách hỗ trợ doanh nghiệp hiệu quả*'; '*Chính sách đào tạo lao động tốt*' có độ lệch chuẩn lớn hơn các chỉ số còn lại. Điều này cho thấy sự chênh lệch rất lớn về điểm số trong các chỉ số này giữa các tỉnh, thành. Trong đó, chỉ số '*Môi trường cạnh tranh bình đẳng*' có khoảng biến thiên rất lớn giữa các tỉnh, thành với điểm số thấp nhất là 3,72 và cao nhất là 8,24.

Câu 3. Anh/Chị vẽ đồ thị tần suất theo khoảng giá trị (histogram) và biểu đồ hộp (boxplot) của từng chỉ số thành phần của 63 tỉnh/thành. Anh/Chị có kết luận gì về phân phối xác suất của các chỉ số thành phần? Các kết luận này có phù hợp với các kết luận thu được từ các trị thống kê được tính ở Câu 2 không?





Biểu đồ histogram của từng chỉ số thành phần PCI của 63 tỉnh, thành



Boxplot của từng chỉ số thành phần PCI của 63 tỉnh, thành

Dựa theo đồ thị histogram, phần lớn phân phối xác suất của các chỉ số thành phần có xu hướng lệch trái, ngoại trừ 2 chỉ số: '*Môi trường cạnh tranh bình đẳng*' và '*Chính sách đào tạo lao động tốt*'. Hai chỉ số này tuy lệch phải nhưng gần như là phân phối chuẩn bởi các giá trị mean và median không chênh lệch nhau nhiều.

Dựa theo các boxplot, các chỉ số '*Chi phí gia nhập thị trường thấp*' và '*Tiếp cận đất đai dễ dàng'* dao động ở phạm vi hẹp và thấp hơn các chỉ số khác. Trong khi đó, các chỉ số '*Môi trường cạnh tranh bình đẳng*'; '*Các chính sách hỗ trợ doanh nghiệp hiệu quả*'; '*Chính sách đào tạo lao động*

tốt' có độ phân tán dữ liệu cao hơn. Ngoài ra, học viên có thể bình luận thêm về các điểm ngoại lệ (outliers) dựa trên các boxplot đối với từng chỉ số thành phần.

Các kết luận này là phù hợp với các trị thống kê đã được tính toán từ kết quả của Câu 2.

Câu 4. Lập bảng giá trị phân vị theo các mức 5%, 10%, 25%, 50%, 75%, 90%, 95% của từng chỉ tiêu thành phần của 63 tỉnh/thành. Dựa trên bảng giá trị phân vị, anh/chị hãy đưa ra một số nhận định về thành tích của các tỉnh/thành theo từng chỉ tiêu thành phần.

Sử dụng hàm PERCENTILE trong Excel để tính toán các giá trị phân vị theo yêu cầu của đề bài, ta lập thành bảng sau:

Giá trị Phân vị	Entry cost	Access to land	Trans-parency	Time costs	Informal charges	Business environment	Pro-activity	Business support	Labour policy	Law and order
5%	6,09	6,12	4,93	6,07	5,80	4,94	6,09	6,12	4,93	6,07
10%	6,18	6,34	5,22	6,45	6,04	5,11	6,18	6,34	5,22	6,45
25%	6,46	6,75	5,49	6,93	6,63	5,51	6,46	6,75	5,49	6,93
50%	6,88	7,06	6,02	7,46	7,06	5,99	6,88	7,06	6,02	7,46
75%	7,23	7,37	6,29	7,91	7,41	6,58	7,23	7,37	6,29	7,91
90%	7,45	7,56	6,81	8,34	7,84	7,05	7,45	7,56	6,81	8,34
95%	7,63	7,60	6,98	8,46	7,95	7,19	7,63	7,60	6,98	8,46

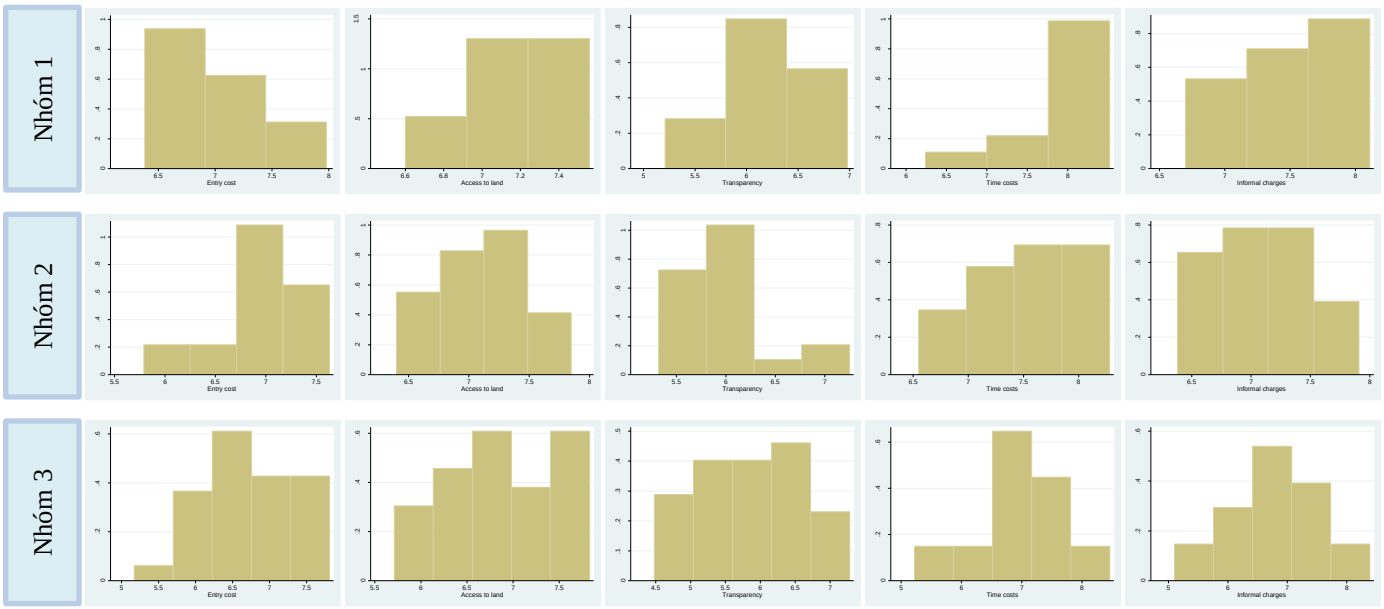
Học viên tự đưa ra các nhận định về thành tích của các tỉnh, thành theo từng chỉ tiêu thành phần.

Câu 5. Giả sử các tỉnh/thành được chia thành 3 nhóm. Nhóm 1 bao gồm các tỉnh/thành được xếp hạng 'Tốt' và 'Rất tốt'. Nhóm 2 bao gồm các tỉnh/thành được xếp hạng 'Khá'. Nhóm 3 bao gồm các tỉnh/thành được xếp hạng 'Trung bình', 'Tương đối thấp', và 'Thấp'.

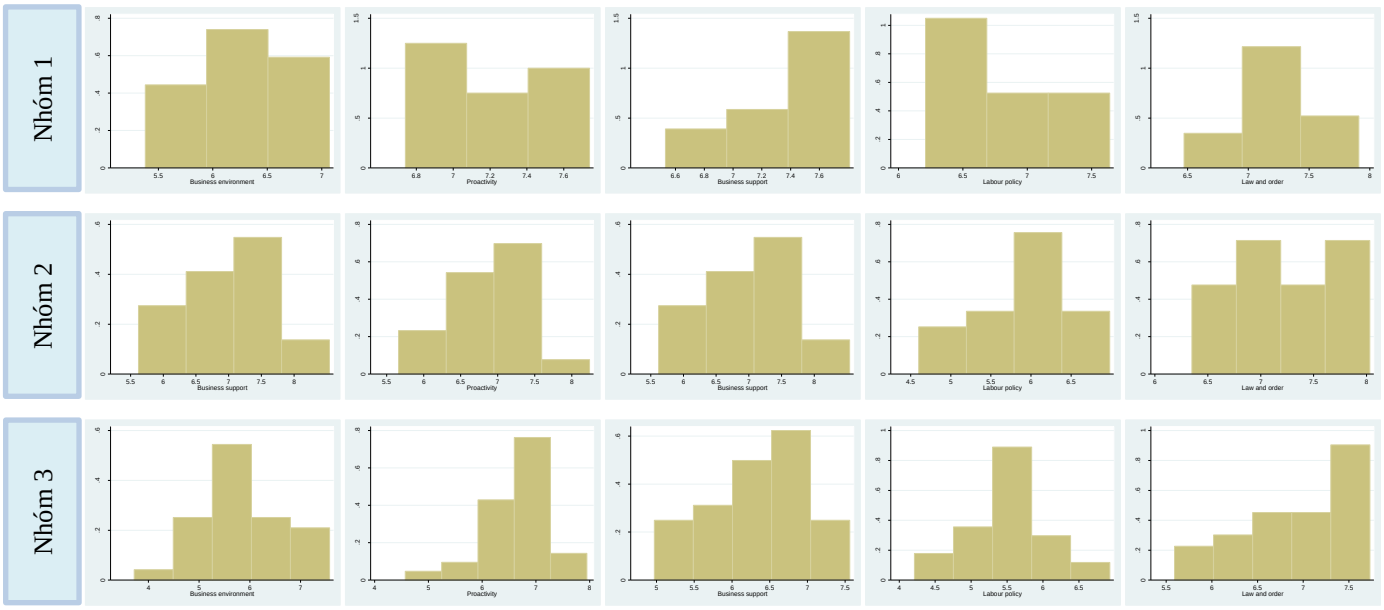
Đối với mỗi nhóm, anh/chị hãy vẽ đồ thị tần suất theo khoảng giá trị (histogram) và biểu đồ hộp (boxplot) của từng chỉ số thành phần và của tổng điểm có trọng số (cột 'PCI_score').

Dựa trên các đồ thị cho mỗi nhóm như trên, anh/chị so sánh và đưa ra nhận định ngắn gọn về phân phối điểm của 'PCI_score' và của từng chỉ số thành phần giữa 3 nhóm.

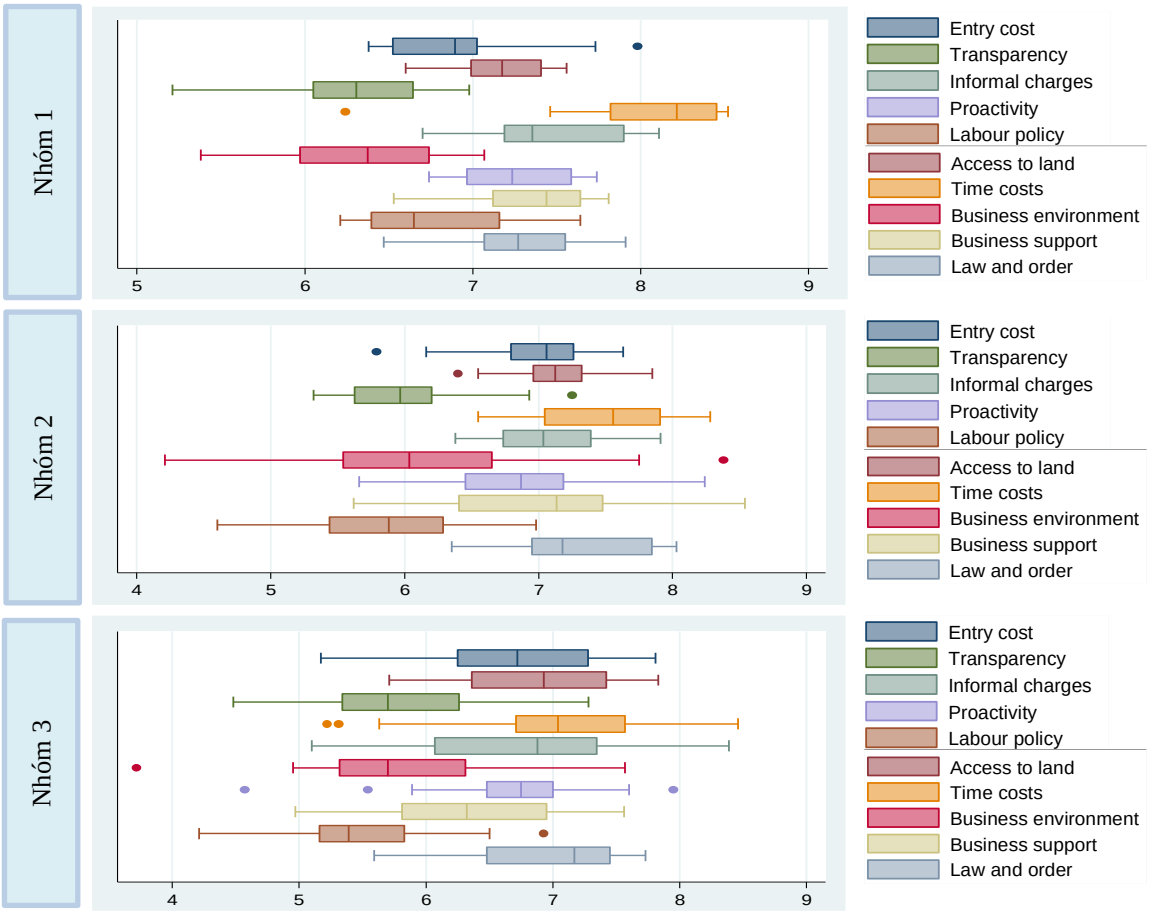
So sánh đồ thị histogram giữa 3 nhóm của 5 chỉ số thành phần: Entry cost, Access to land Transparency, Time costs và Informal charges



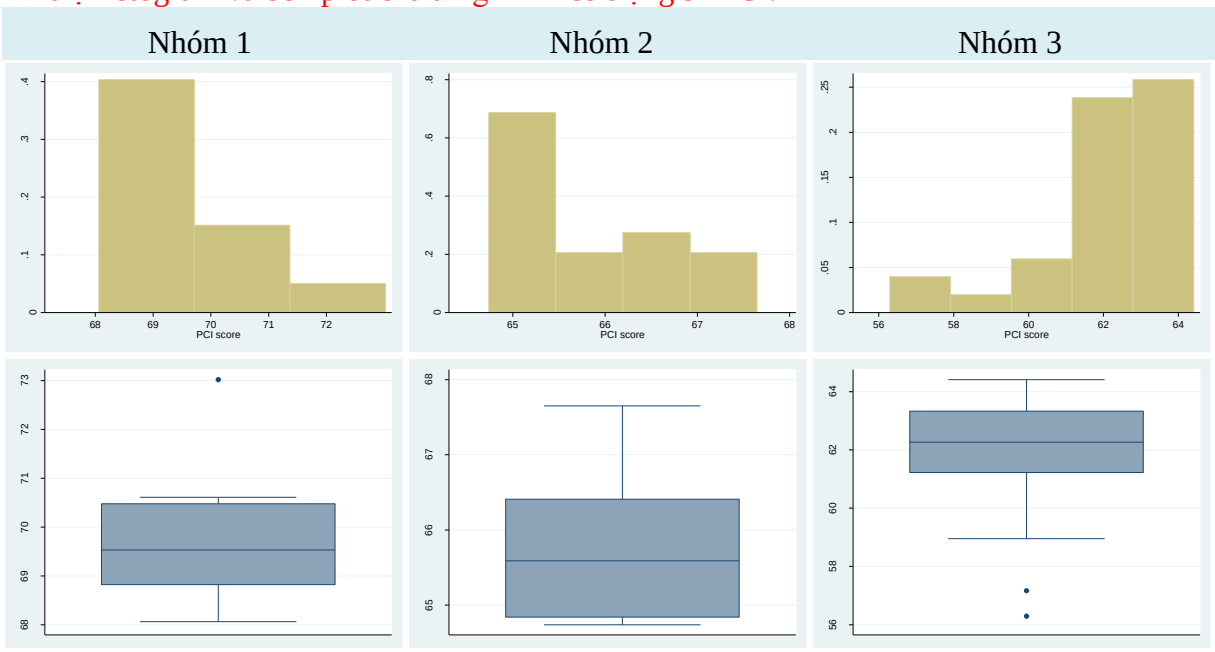
So sánh đồ thị histogram giữa 3 nhóm của 5 chỉ số thành phần: Business environment, Proactivity, Business support, Labour policy và Law and order



So sánh đồ thị histogram của từng chỉ số thành phần giữa 3 nhóm:



Đồ thị histogram và box plot của tổng điểm có trọng số PCI:



Sau đây là một số gợi ý bình luận, học viên có thể đưa ra các bình luận khác miễn sao hợp lý và thuyết phục:

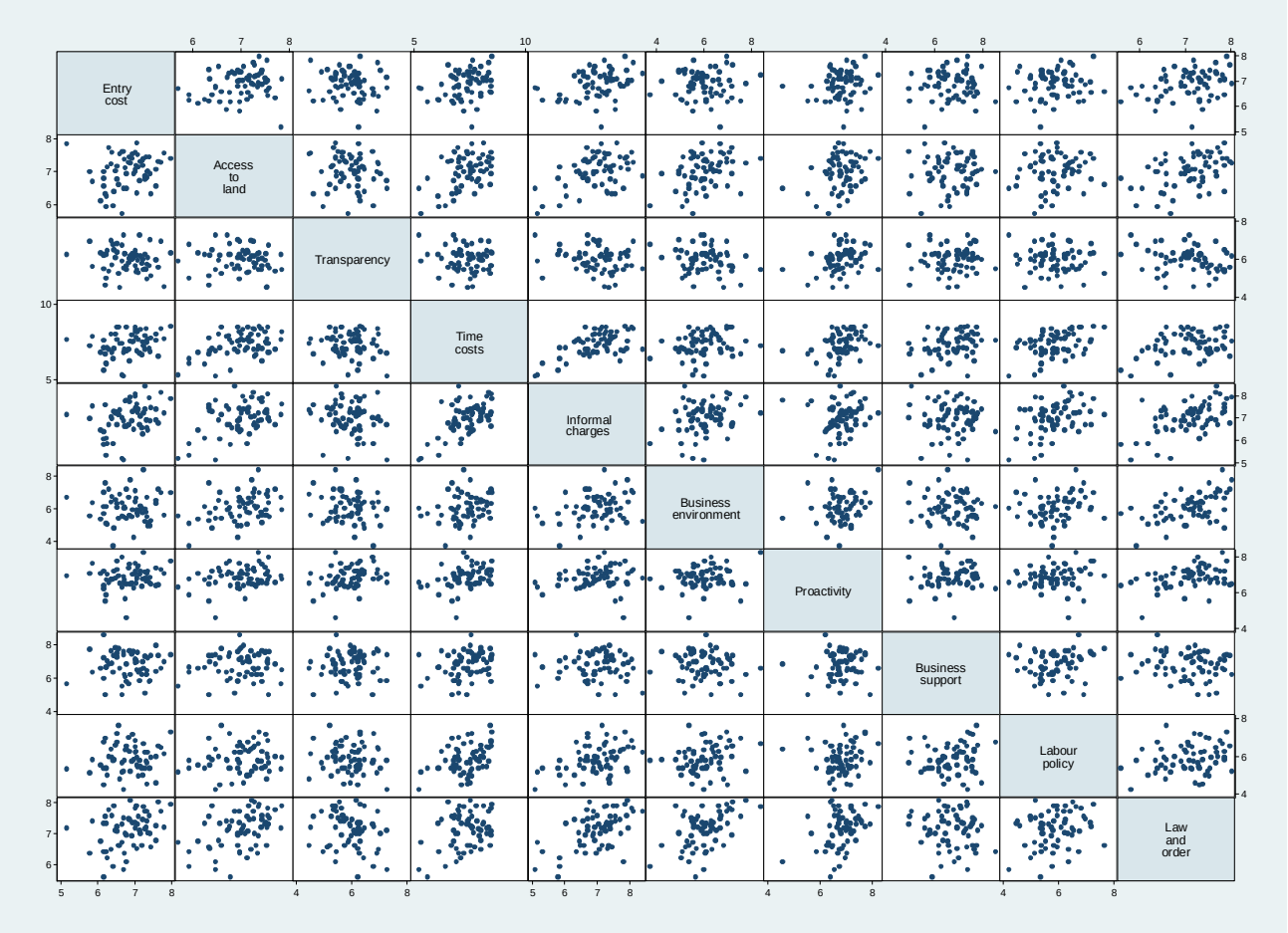
- Phân phối điểm của từng chỉ số thành phần là khác nhau giữa các nhóm. Một số chỉ số thành phần có phân phối khá tương đồng giữa nhóm 2 và nhóm 3, trong khi đó nhóm 1 thì hoàn toàn khác biệt, ví dụ như các chỉ số: '*Chi phí gia nhập thị trường thấp*'; '*Môi trường kinh doanh minh bạch và thông tin kinh doanh công khai*'.
- Phân phối điểm PCI_score của nhóm 1 và nhóm 2 có xu hướng lệch phải, trong khi đó nhóm 3 có xu hướng lệch trái.

Câu 6. Anh/Chị tính hệ số tương quan giữa các cặp chỉ số thành phần của 63 tỉnh/thành, kiểm chứng bằng các đồ thị scatter plot tương ứng, và nhận định ngắn gọn về các kết quả.

Sử dụng công cụ Data Analysis để tính ma trận hệ số tương quan của các cặp chỉ số thành phần của 63 tỉnh/thành, ta có bảng sau:

	Entry cost	Access to land	Trans-parency	Time costs	Informal charges	Business environment	Pro-activity	Business support	Labour policy
Entry cost	1,0000								
Access to land	0,1992	1,0000							
Transparency	-0,1243	-0,1110	1,0000						
Time costs	0,2901	0,4357	-0,1031	1,0000					
Informal charges	0,3389	0,4430	-0,1848	0,5839	1,0000				
Business environment	0,0165	0,3570	-0,1858	0,2355	0,2884	1,0000			
Proactivity	0,1419	0,2693	0,2955	0,3361	0,1311	0,1669	1,0000		
Business support	0,0257	0,0738	0,0908	0,2499	-0,0121	-0,1180	0,0815	1,0000	
Labour policy	0,1017	0,1269	-0,1093	0,4044	0,4083	0,2394	0,1007	0,1298	1,0000
Law and order	0,3358	0,4416	-0,2284	0,3416	0,5470	0,4865	0,3665	-0,1021	0,2304

Ma trận đồ thị scatter plot các chỉ số thành phần PCI của 63 tỉnh, thành:



Một số nhận định về kết quả từ bảng ma trận hệ số tương quan:

- Chỉ số '*Tiếp cận đất đai dễ dàng*' có mối tương quan đồng biến tương đối cao với các chỉ số '*Thời gian thanh tra, kiểm tra và thực hiện các quy định thủ tục hành chính nhanh chóng*'; '*Chi phí không chính thức thấp*' và '*Thủ tục giải quyết tranh chấp công bằng, hiệu quả và an ninh trật tự được duy trì'*. Trong đó, hai chỉ số '*Thời gian thanh tra, kiểm tra và thực hiện các quy định thủ tục hành chính nhanh chóng*' và '*Chi phí không chính thức thấp*' có mối quan tương quan đồng biến cao nhất.
- Chỉ số '*Các chính sách hỗ trợ doanh nghiệp hiệu quả*' gần như không có tương quan với các chỉ số '*Chi phí gia nhập thị trường thấp*'; '*Chi phí không chính thức thấp*' và '*Môi trường cạnh tranh bình đẳng*'.

Câu 7. Anh/Chị lập bảng tần suất 2 chiều phù hợp để trả lời các câu hỏi sau:

- Xác suất để chọn 1 tỉnh thuộc vùng Đồng Bằng Sông Cửu Long mà được xếp hạng PCI loại 'Khá' hoặc cao hơn.
- Xác suất để chọn 1 tỉnh thuộc vùng Bắc Trung Bộ và Duyên Hải Miền Trung, vùng Tây Nguyên và vùng Đông Nam Bộ mà được xếp hạng PCI loại 'Trung bình'.
- Xác suất để chọn 1 tỉnh thuộc vùng Đồng Bằng Sông Hồng và vùng Trung Du và Miền Núi Phía Bắc mà được xếp hạng PCI loại 'Tốt'.

Danh sách các tỉnh/thành nằm trong các vùng kinh tế xã hội trong Câu 7 được định nghĩa bởi Tổng cục Thống kê tại đường link https://tinyurl.com/2x9rpehu.

Sử dụng công cụ Pivot Table trong Excel, ta có bảng tần suất hai chiều như sau:

Vùng kinh tế xã hội	Rất tốt	Tốt	Khá	Trung bình	Tương đối thấp	Thấp	**Tổng cộng**
Trung du và miền núi phía Bắc			5	5	2	2	**14**
Đồng bằng Sông Hồng	1	4	2	3	1		**11**
Bắc Trung Bộ và Duyên hải Miền Trung		3	4	6	1		**14**
Tây Nguyên			2	2	1		**5**
Đông Nam Bộ		2	2	2			**6**
Đồng bằng Sông Cửu Long		2	5	4	2		**13**
Tổng cộng	**1**	**11**	**20**	**22**	**7**	**2**	**63**

Gọi A là biến cố "tỉnh thuộc vùng Đồng Bằng Sông Cửu Long"

B là biến cố "tỉnh được xếp hạng PCI loại Khá hoặc cao hơn"

C là biến cố "tỉnh thuộc vùng Đồng Bằng Sông Cửu Long mà được xếp hạng PCI loại Khá hoặc cao hơn"

$$P(C) = \frac{n(A \cap B)}{n(B)} = \frac{7}{20 + 11 + 1} = 0{,}21875 = 21{,}875\%$$

Gọi D là biến cố "tỉnh thuộc vùng Bắc Trung Bộ và Duyên Hải Miền Trung"

E là biến cố "tỉnh thuộc vùng Tây Nguyên"

F là biến cố "tỉnh thuộc vùng Đông Nam Bộ"

G là biến cố "tỉnh được xếp hạng PCI loại Trung bình"

H là biến cố "tỉnh thuộc vùng Bắc Trung Bộ và Duyên Hải Miền Trung, vùng Tây Nguyên và vùng Đông Nam Bộ mà được xếp hạng PCI loại Trung bình"

$$P(H) = \frac{n\,(D \cap G) + n\,(E \cap G) + n\,(F \cap G)}{n(G)} = \frac{6 + 2 + 2}{22} = 0{,}4545 = 45{,}45\%$$

Gọi I là biến cố "tỉnh thuộc vùng Đồng Bằng Sông Hồng"

J là biến cố "tỉnh thuộc vùng Trung Du và Miền Núi Phía Bắc"

K là biến cố "tỉnh được xếp hạng PCI loại Tốt"

L là biến cố "thuộc vùng Đồng Bằng Sông Hồng và vùng Trung Du và Miền Núi Phía Bắc mà được xếp hạng PCI loại Tốt"

$$P(L) = \frac{n\,(I \cap K) + n\,(J \cap K)}{n\,(K)} = \frac{4 + 0}{11} = 0{,}3636 = 36{,}36\%$$

---HẾT---