빅데이터와 통계학 HW2

2016314726 정영준

1. 데이터 수집과정

공공데이터 포털에서 '인구' 키워드로 검색을 하여 서울특별시 인구 현황 데이터셋을 찾았습니다. 서울 열린데이터 광장으로의 링크를 통해 데이터를 txt 파일로 다운로드 받을 수 있었습니다.

http://data.seoul.go.kr/dataList/10930/S/2/datasetView.do;jsessionid=0413C57E0233CB97D478E8019FE1BD64.new_portal-svr-11

2. 데이터에 대한 설명

사용한 데이터는 2014년부터 2020년 2/4분기까지 4분기별 서울특별시의 모든 구의 1인 ~ 10인 이상의 세대원수별 세대의 수가 기록된 데이터입니다. 구별로 분기마다 집계된 세대 수가 기록되어 있으며 합계를 계산한 열이 따로 있었습니다. 꺾은선 그래프로 나타낼 때는 2014년 1/4분기부터 2020년 2/4분기까지 데이터를 사용하였으며, 나머지 그래프 및 기술통계량은 가장 최신의 2020년 2/4분기 데이터를 사용하였습니다. 1인 세대, 2인 이상 세대로 나누어 확인하고자 2인 ~ 10인 이상 세대의 수를 합하여 새로운 열인 2인 이상 세대 열을 만들고 이를 활용하였습니다.

기간	자치구	1인세대	2인세대	3인세대	4인세대	5인세대	6인세대	7인세대	8인세대	9인세대	10인세대	세대순위	전체세대	다인세대
2020.2/4	종로구	37,503	14,341	10,697	8,837	2,364	524	164	38	13	16	24	74,497	36,994
2020.2/4	중구	31,643	13,498	9,121	6,839	1,660	429	108	26	12	18	25	63,354	31,711
2020.2/4	용산구	53,405	23,093	17,159	13,372	3,444	807	207	55	25	19	23	111,586	58,181
2020.2/4	성동구	56,420	29,972	24,552	19,422	4,264	965	248	62	17	15	21	135,937	79,517
2020.2/4	광진구	76,534	34,026	26,435	22,936	4,902	1,086	272	74	24	18	17	166,307	89,773
2020.2/4	동대문구	78,266	33,984	26,184	21,335	5,158	1,143	284	64	23	10	16	166,451	88,185
2020.2/4	중랑구	77,396	42,377	31,588	24,642	5,790	1,128	288	67	21	13	10	183,310	105,914
2020.2/4	성북구	77,396	41,884	35,441	31,054	7,217	1,535	382	90	27	38	7	195,064	117,668
2020.2/4	강북구	61,479	34,556	24,953	18,768	4,518	975	205	67	17	21	18	145,559	84,080
2020.2/4	도봉구	46,140	34,302	28,278	23,352	5,376	1,124	277	80	29	8	20	138,966	92,826
2020.2/4	노원구	70,945	49,730	44,406	41,982	8,733	1,595	345	109	32	20	5	217,897	146,952
2020.2/4	은평구	79,784	49,503	39,388	31,859	7,786	1,772	406	101	33	32	6	210,664	130,880
2020.2/4	서대문구	61,198	30,545	24,575	20,711	4,830	1,109	248	69	26	12	19	143,323	82,125
2020.2/4	마포구	81,156	36,658	28,568	24,268	5,371	1,110	270	93	28	22	14	177,544	96,388
2020.2/4	양천구	52,468	39,088	38,643	39,063	8,246	1,638	359	110	36	26	12	179,677	127,209
2020.2/4	강서구	107,620	59,609	46,970	39,512	9,097	1,828	456	110	32	18	3	265,252	157,632
2020.2/4	구로구	68,192	41,158	32,729	27,843	6,570	1,417	326	92	27	13	13	178,367	110,175
2020.2/4	금천구	52,320	24,765	17,378	13,770	3,348	742	161	57	15	14	22	112,570	60,250
2020.2/4	영등포구	84,507	36,497	28,444	23,612	5,208	1,115	283	74	20	27	11	179,787	95,280
2020.2/4	동작구	80,504	38,324	30,910	26,154	5,940	1,274	322	68	26	15	9	183,537	103,033
2020.2/4	관악구	157,514	50,306	32,489	25,540	6,058	1,328	345	82	28	25	2	273,715	116,201
2020.2/4	서초구	59,312	34,979	34,214	34,591	8,153	1,955	532	174	48	40	15	173,998	114,686
2020.2/4	강남구	93,397	44,204	41,235	42,381	9,350	2,089	561	159	67	41	4	233,484	140,087
2020.2/4	송파구	97,868	59,598	55,402	52,580	11,465	2,451	545	158	40	28	1	280,135	182,267
2020.2/4	강동구	67,911	43,223	37,911	34,317	7,670	1,522	398	100	21	22	8	193,095	125,184

3. 기술통계량과 그래프 첨부 및 해석

자치구	1인세대 비율	1인세대 비율 순위
종로구	50	2
중구	50	3
용산구	48	4
성동구	42	14
광진구	46	8
동대문구	47	5
중랑구	42	13
성북구	40	17
강북구	42	12
도봉구	33	23
노원구	33	24
은평구	38	19
서대문구	43	11
마포구	46	9
양천구	29	25
강서구	41	15
구로구	38	18
금천구	46	7
영등포구	47	6
동작구	44	10
관악구	58	1
서초구	34	22
강남구	40	16
송파구	35	21
강동구	35	20

먼저 1인 세대 수 /전체 세대 수 * 100을 함수로 설정한 열을 통해 1인세대 비율을 각 구별로 계산하였습니다. 관악구의 1인세대 비율이 58퍼센트로 가장 높았으며 양천구의 1인세대 비율은 29퍼센트로 가장 낮게 나타났습니다. 구에 따라 2배의 1인 세대 비율 차이가 난다는 것을 알게 되었습니다. 같은 서울특별시에 위치한 자치구인데 2배의 차이가 난다는 것이 의아하여 이유를 찾아보았습니다. 인터넷 검색을 통해 뉴스와 데이터 분석 보고서를 살펴본 결과 이는 관악구에 대학가 및 일자리가 밀집되어 있기 때문이고 학군 특수 지역으로 불리는 양천구, 서초구 등에는 인접 지역 대비 집값이 비싸 원룸 같은 1인가구가의 진입이 쉽지 않기 때문이라고 합니다.

기간	항목	1인세대	2인세대	3인세대	4인세대	5인세대	6인세대	7인세대	8인세대	9인세대	10인세대
2020.2/4	합계	1,810,878	940,220	767,670	668,740	152,518	32,661	7,992	2,179	687	531
2020.2/4	평균	72,435	37,609	30,707	26,750	6,101	1,306	320	87	27	21
2020.2/4	중앙값	70,945	36,658	30,910	24,642	5,790	1,143	288	80	26	19
2020.2/4	최빈값	77396	#N/A	#N/A	#N/A	#N/A	#N/A	248	74	17	18
2020.2/4	1th quantile	56420	33984	24953	20711	4830	1086	248	67	21	15
2020.2/4	2th quantile	70945	36658	30910	24642	5790	1143	288	80	26	19
2020.2/4	3th quantile	80504	43223	37911	34317	7786	1595	382	101	32	26
2020.2/4	최대값	157514	59609	55402	52580	11465	2451	561	174	67	41
2020.2/4	최소값	31,643	13,498	9,121	6,839	1,660	429	108	26	12	8
2020.2/4	범위	125,871	46,111	46,281	45,741	9,805	2,022	453	148	55	33

다음으로는 세대구성원수에 따른 기술통계량을 분석하겠습니다. 합계는 전체 서울특별시의 2020년 2/4 분기 각 구성원 수에 따른 세대의 총합을 나타냅니다. 1인 세대가 다른 세대보다 눈에 띄게 수가 많았습니다. 이는 2인세대는 2명의 사람이 있어야 1세대가 되지만 1인세대는 1명의 사람이 세대를 이루기 때문이라고 해석이 가능합니다. 실제 2인세대에 2를 곱한 경우 1인세대의 합계와 비슷한 값이 나오는 것을 볼 수 있습니다. 다음으로 평균, 중앙값, 최빈값을 구하였습니다. 평균과 중앙값에 모든 세대 분류에서 큰 차이가 있지 않다는 것을 관찰하였습니다. 최빈값의 경우 연속형 변수이기 때문에 모두 최대 1번 나타나 NULL값을 반환하는 셀이 있었습니다. 다음으로 1분위 2분위 3분위 수를 구하였고 최대값 및 최소값을 구하였습니다. 2분위 수는 상위 50%번째 수로 중앙값과 같

은 값을 가집니다. 각 세대 유형별 범위의 차이가 큽니다. 이는 각 유형별 세대의 평균이 1인세대는 72435 세대인 반면 10인 이상 세대는 21로 차이가 크기 때문입니다. 같은 스케일로 비교하기 위해 표본변동계수를 구하였습니다.

1인세대		2인세대		3인세대		4인세대		5인세대	
평균	72435.12	평균	37608.8	평균	30706.8	평균	26749.6	평균	6100.72
표준 오차	5076.149	표준 오차	2313.14	표준 오차	2158.828	표준 오차	2205.642	표준 오차	467.1538
중앙값	70945	중앙값	36658	중앙값	30910	중앙값	24642	중앙값	5790
최빈값	77396	최빈값	#N/A	최빈값	#N/A	최빈값	#N/A	최빈값	#N/A
표준 편차	25380.75	표준 편차	11565.7	표준 편차	10794.14	표준 편차	11028.21	표준 편차	2335.769
분산	6.44E+08	분산	1.34E+08	분산	1.17E+08	분산	1.22E+08	분산	5455816
첨도	4.310042	첨도	0.320245	첨도	0.33001	첨도	0.009781	첨도	-0.10968
왜도	1.462479	왜도	-0.15449	왜도	0.036269	왜도	0.337975	왜도	0.237477
범위	125871	범위	46111	범위	46281	범위	45741	범위	9805
최소값	31643	최소값	13498	최소값	9121	최소값	6839	최소값	1660
최대값	157514	최대값	59609	최대값	55402	최대값	52580	최대값	11465
합	1810878	합	940220	합	767670	합	668740	합	152518
관측수	25	관측수	25	관측수	25	관측수	25	관측수	25
표본변동계수	35.03928	표본변동계수	30.75264	표본변동계수	35.15227	표본변동계수	41.22756	표본변동계수	38.28677
범위	125871	범위	46111	범위	46281	범위	45741	범위	9805

6인세대		7인세대		8인세대		9인세대		10인세대 이상	
평균	1306.44	평균	319.68	평균	87.16	평균	27.48	평균	21.24
표준 오차	96.33878	표준 오차	23.57765	표준 오차	7.172559	표준 오차	2.34316	표준 오차	1.814093
중앙값	1143	중앙값	288	중앙값	80	중앙값	26	중앙값	19
최빈값	#N/A	최빈값	248	최빈값	74	최빈값	17	최빈값	18
표준 편차	481.6939	표준 편차	117.8883	표준 편차	35.86279	표준 편차	11.7158	표준 편차	9.070465
분산	232029	분산	13897.64	분산	1286.14	분산	137.26	분산	82.27333
첨도	0.143787	첨도	-0.0638	첨도	0.874971	첨도	4.554542	첨도	0.084951
왜도	0.402654	왜도	0.489064	왜도	0.932212	왜도	1.723417	왜도	0.859359
범위	2022	범위	453	범위	148	범위	55	범위	33
최소값	429	최소값	108	최소값	26	최소값	12	최소값	8
최대값	2451	최대값	561	최대값	174	최대값	67	최대값	41
합	32661	합	7992	합	2179	합	687	합	531
관측수	25	관측수	25	관측수	25	관측수	25	관측수	25
표본변동계수	36.87072	표본변동계수	36.87696	표본변동계수	41.14593	표본변동계수	42.63392	표본변동계수	42.70464
범위	2022	범위	453	범위	148	범위	55	범위	33

표준편차인 분산의 제곱근을 표본평균으로 나누고 여기에 100을 곱하여 함수를 만들었습니다. 1인 세대에서 분산이 가장 크지만 표본변동계수를 보면 10인세대 이상에서 가장 큰 표본변동계수를 관찰할 수 있었습니다. 분포의 그래프의 형태가 뾰족함을 나타내는 첨도는 1인세대에서 가장 높았으며 5인세대에서 가장 낮게 나타났습니다. 분포의 비대칭도를 나타내는 왜도는 9인세대에서 가장 높고 2인세대에서 가장 낮게 나타났습니다.

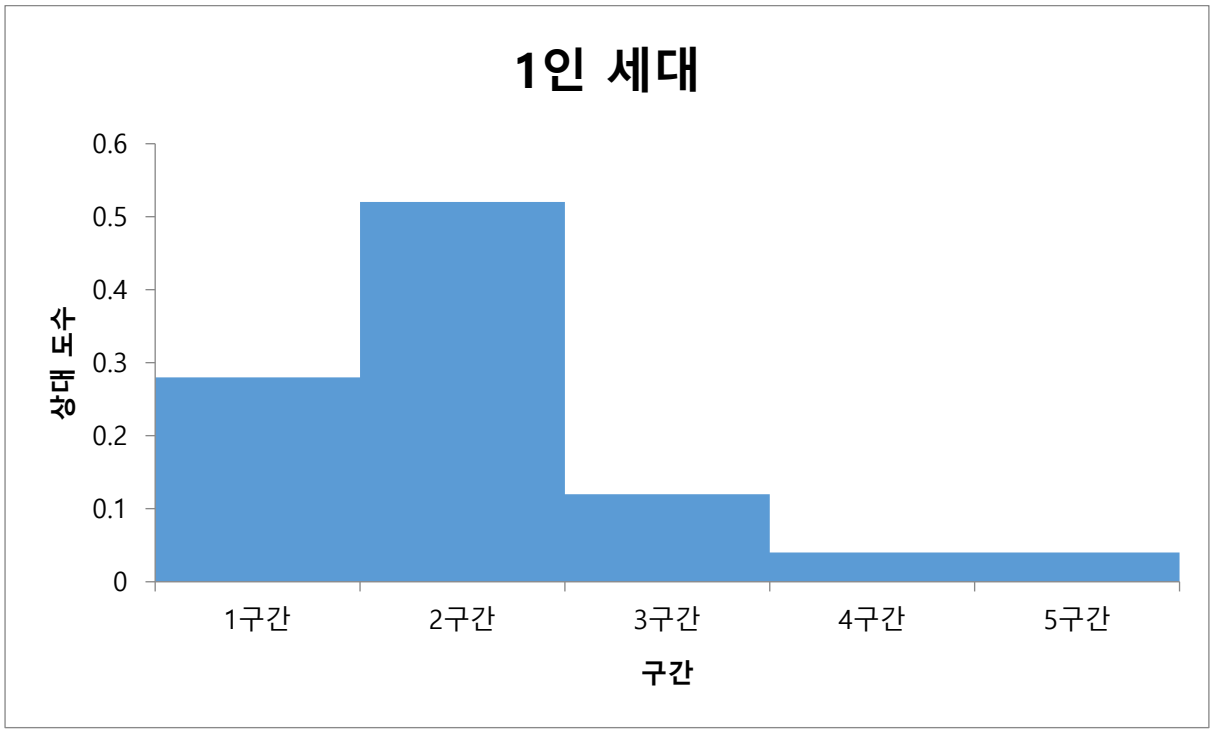
다음으로는 도수분포표, 히스토그램, 파이 차트, 상자 그림, 꺾은선 그래프를 각 세대 유형 또는 각 구별로 분석하였습니다.

	1인 세대	2인 이상 세대
범위	125,871	150,556
구간의길이	25174.2	30111.2
계급구간1	56817.2	61822.2
계급구간2	81991.4	91933.4
계급구간3	107166	122045
계급구간4	132340	152156
계급구간5	157514	182267

1인세대와 2인 이상 세대로 나누고 2를 밑으로 한 log를 전체 관측 수인 25에 적용하고 반올림을 통해 구간의 수인 5를 얻었습니다. 범위를 5개로 나누어 각 구간의 최대값을 통해 도수분포표와 히스토그램을 만들 수 있습니다.

계급	빈도수	상대도수
1구간	7	0.28
2구간	13	0.52
3구간	3	0.12
4구간	1	0.04
5구간	1	0.04

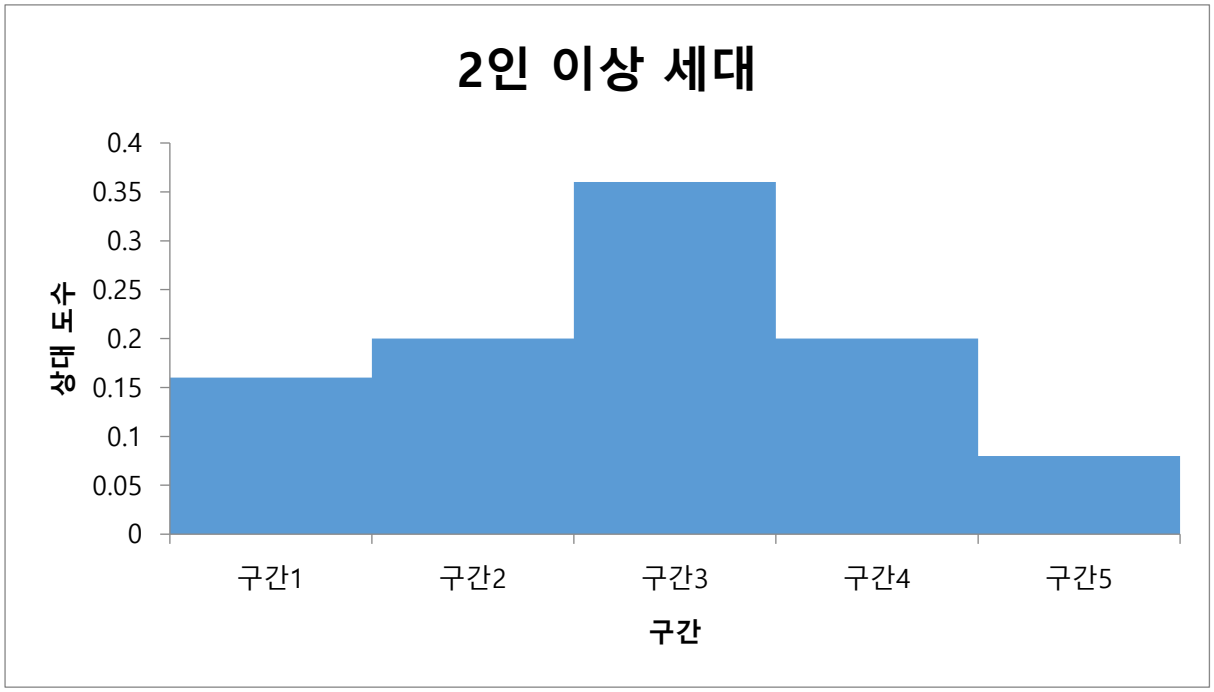
1인 가구의 도수분포표입니다. 빈도수의 총계는 25이고 상대도수의 총계는 1입니다. 1인 세대의 왜도는 1.4로 양수입니다. 이는 왼쪽으로 치우치고 오른쪽으로 꼬리가 긴 분포를 가집니다. 도수분포표와 히스토그램에서 이를 확인할 수 있었습니다.



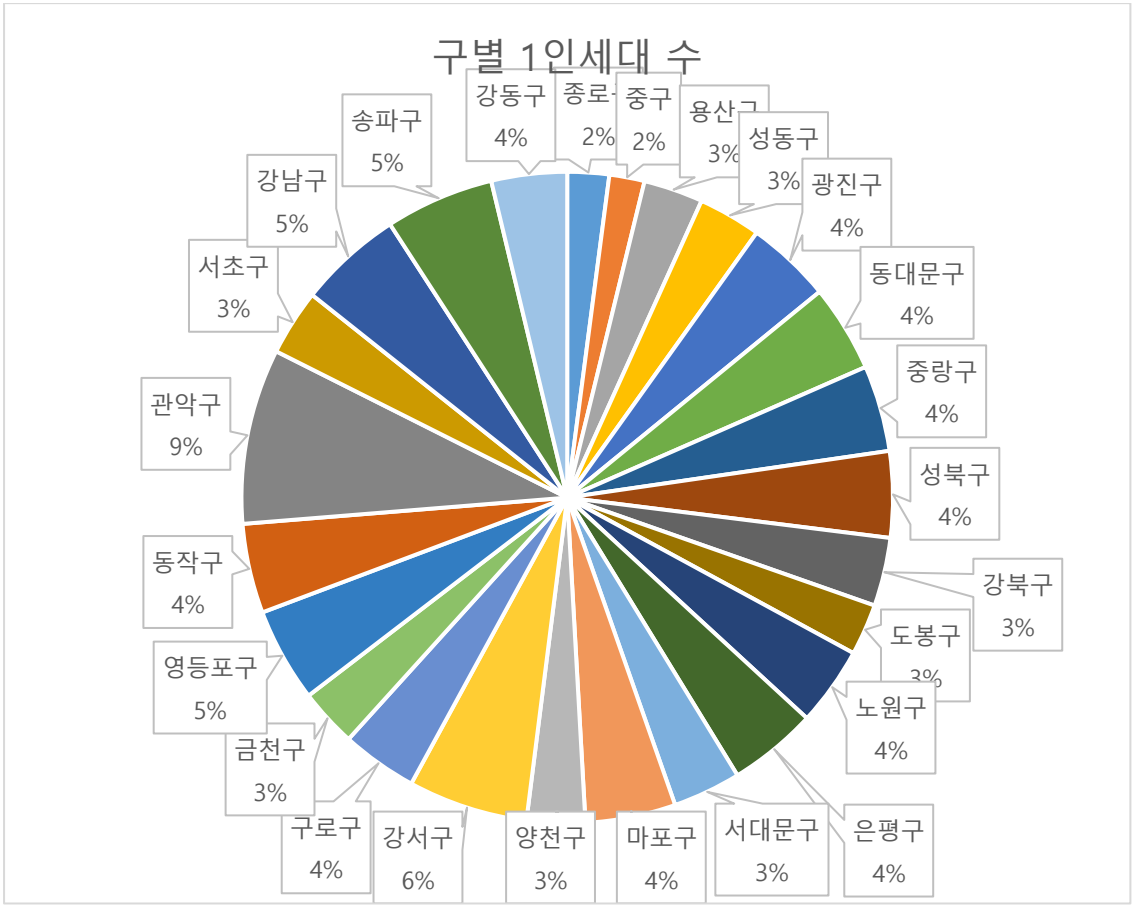
히스토그램을 통해 1인가구 수가 상위인 구는 평균과 많은 차이가 나는 1인가구 수가 나타난 다는 것을 알 수 있었습니다.

계급	빈도수	상대도수
구간1	4	0.16
구간2	5	0.2
구간3	9	0.36
구간4	5	0.2
구간5	2	0.08

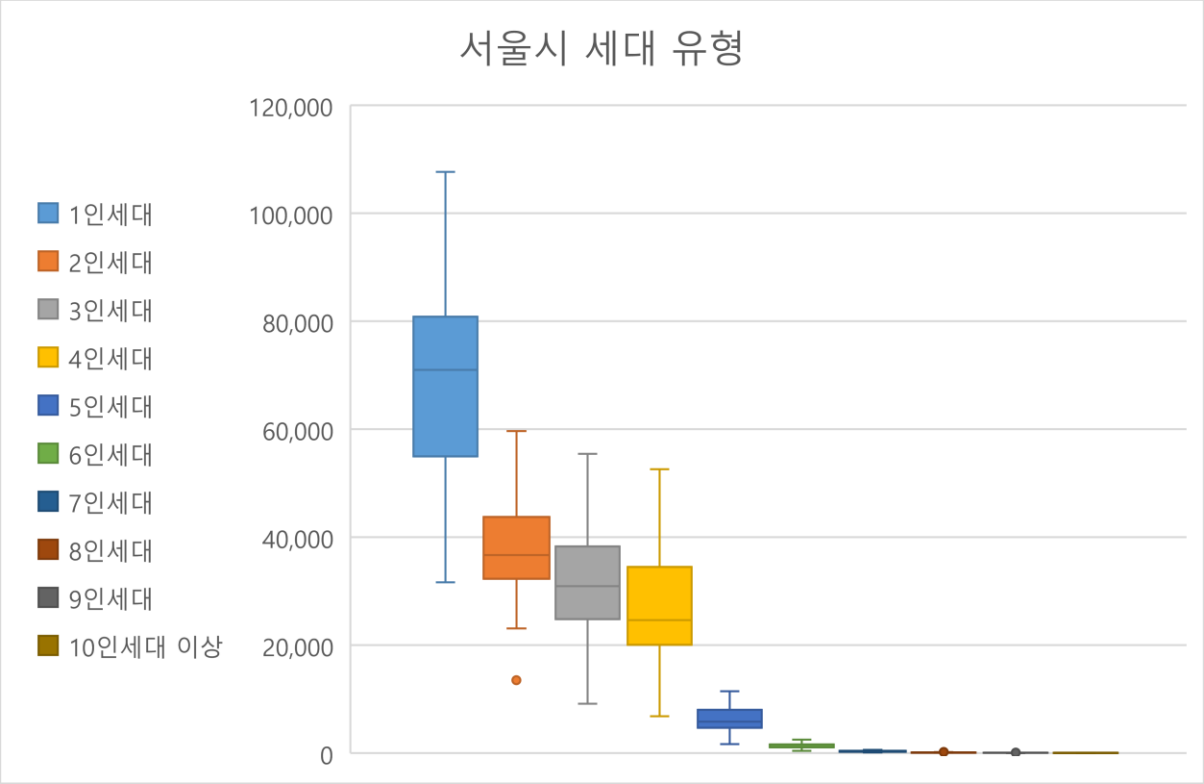
다음으로 2인 이상 세대의 수의 도수분포표입니다. 역시 전체 빈도수는 25이고 상대도수의 합은 1입니다. 2인가구를 제외한 모든 유형의 세대의 왜도는 양수로 역시 2인 이상 세대의 왜도는 양수로 나타났으며 분포도 마찬가지로 왼쪽으로 살짝 치우친 분포를 보였습니다.



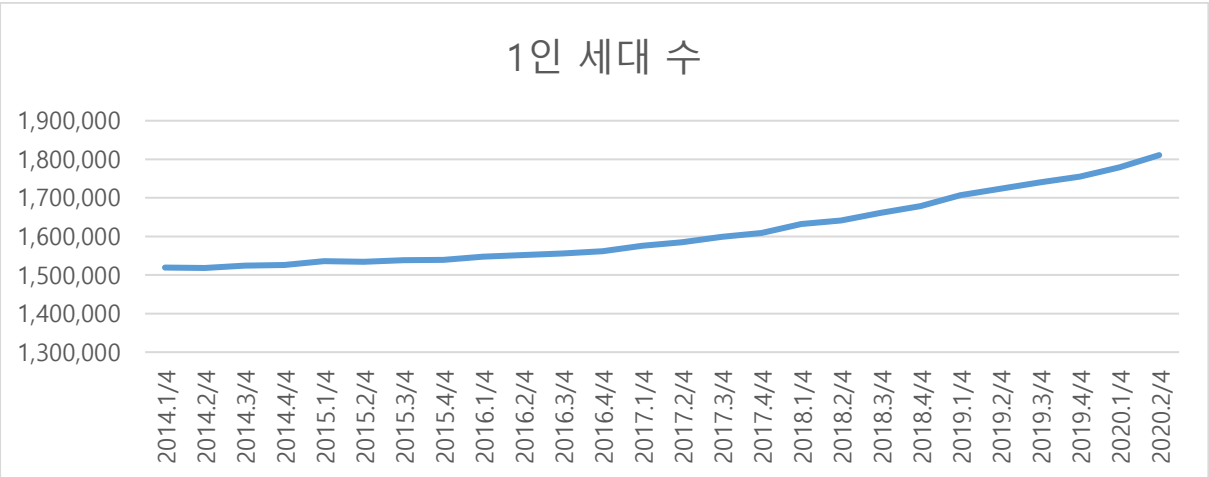
다음으로 파이 차트를 통하여 어느 구에 1인 세대가 많은지 보기 쉽게 시각화를 하였습니다.

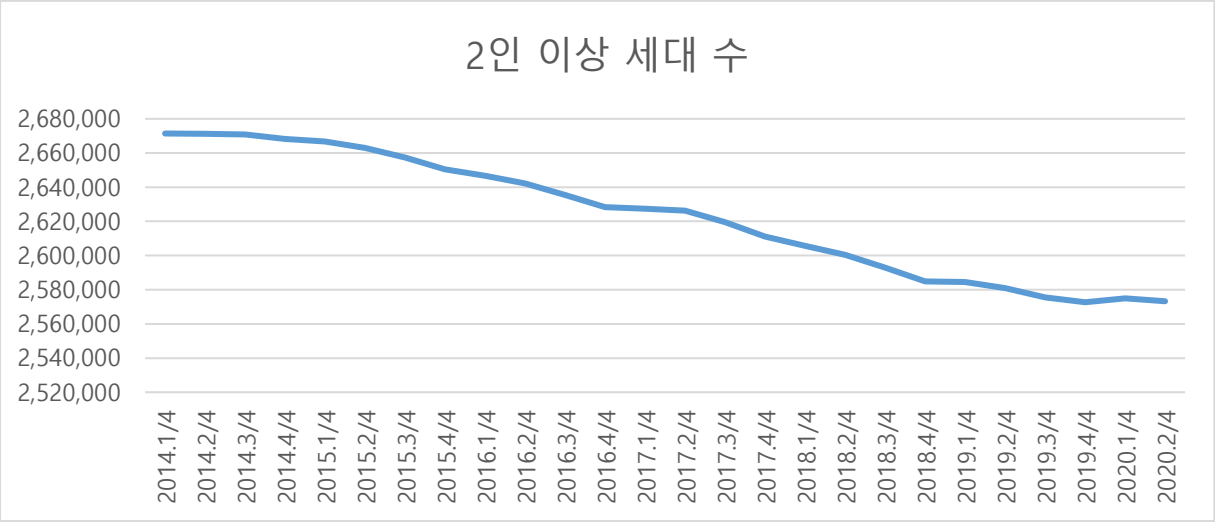


관악구에 총 1인세대의 9%가 위치하였으며 중구와 종로구에서는 총 1인세대의 2%가 위치하고 있었습니다. 1인 세대 비율과 위 총 1인 세대 수 파이 차트에서 순위가 다른 이유는 1인 세대 비율이 낮더라도 총 구의 인구수가 높으면 결국 1인 세대 수가 많아지기 때문입니다.



상자 그림으로 세대 구성원 수에 따라 분류한 세대 종류별 분포를 나타냈습니다. 역시 1인세대에서 가장 높은 평균값을 볼 수 있었습니다. 세대 구성원 수가 늘어날수록 세대의 수가 낮은 것을 알 수 있었습니다. 2인 세대에서는 하나의 outlier가 관찰 되었습니다. 이는 종로구로 전체 세대수가 63,354 세대로 다른 구의 1/3 또는 1/2 수준으로 2인 세대 또한 수가 낮게 나타나 이러한 결과가 나온 것을 확인하였습니다.





다음으로 2014년 1/4분기부터 2020년 2/4분기까지 서울시 전체 1인 세대 수, 2인 이상 세대 수를 사용하여 시간 순으로 꺾은선 그래프를 그렸습니다. 먼저 1인 세대 수 꺾은선 그래프를 보면 시간이 지날수록 점점 증가하는 것을 볼 수 있습니다. 2인 이상 세대 수를 보면 시간이 지날수록 점차 감소하는 것을 볼 수 있습니다. 전체 세대 수가 시간이 지날수록 증가하였는데 2인 이상 세대 수는 반대로 감소한 것을 보면 서울 전체 세대 수 증가는 1인가구가 늘어났기 때문임을 알 수 있습니다.

인구주택총조사에 따르면 서울에서 혼자사는 사람은 24%에 달한다고 합니다. 이는 세계 여러 대도시에서 나타나는 현상으로 비혼과 만혼의 증가, 여성의 경제활동 증가, 교육환경에 기인한 기러기 가족 증가, 이혼률 증가, 고령화로 인한 노인가구 증가를 그 원인으로 꼽는 전문가들이 많다고 합니다. 추가로 지금까지 분석한 그래프와 기술통계량을 보면 실제 1인가구가 많아지고 있다는 것을 알 수 있으며 구별로 환경이 달라 1인 가구 비율에 차이가 생긴다는 것을 분석을 통해 얻을 수 있었습니다.