[EDA] BOIDOT¸ AUGUSTIN / Lee, Hyeon Dong / Jeon Yeeun

Project: International Student Accommodation

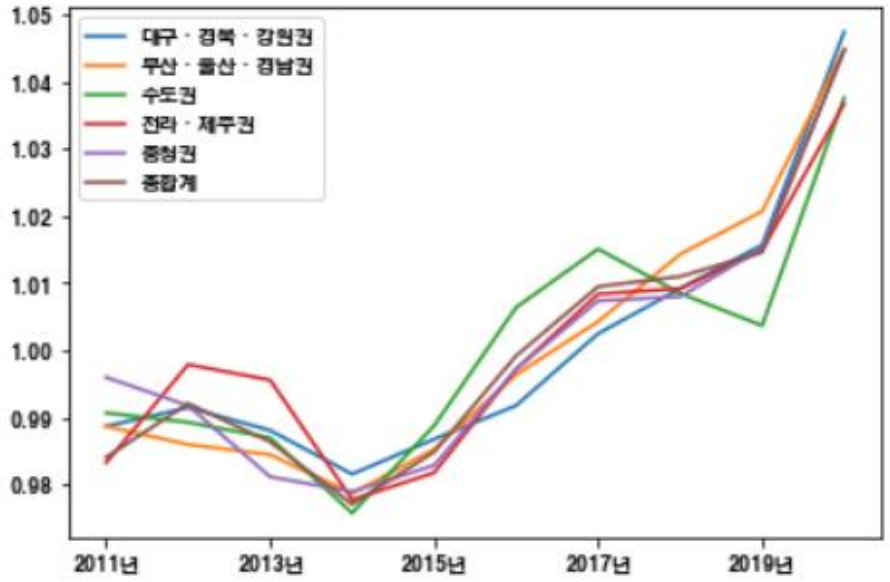
*This is just the report of final result, so if you want to look at the specific data or python code, you can look into each folder.

EDA1. Financial soundness of universities and trend of international students by year

EDA1 – Table 1

	대구·경북·강원권	부산·울산·경남권	수도권	전라·제주권	충청권	총합계
2020년	1.0474	1.0449	1.0376	1.0367	1.0446	1.0447
2019년	1.0157	1.0207	1.0037	1.0149	1.0152	1.0146
2018년	1.0092	1.0143	1.0085	1.0092	1.008	1.011
2017년	1.0025	1.0043	1.0151	1.0084	1.0074	1.0095
2016년	0.9918	0.9964	1.0064	0.9972	0.9972	0.9992
2015년	0.9868	0.9853	0.9888	0.9818	0.9829	0.9849
2014년	0.9816	0.9787	0.9758	0.9778	0.979	0.9771
2013년	0.9881	0.9845	0.987	0.9956	0.9812	0.9864
2012년	0.9915	0.986	0.9893	0.9979	0.9918	0.9921
2011년	0.9887	0.9887	0.9907	0.9833	0.996	0.9841

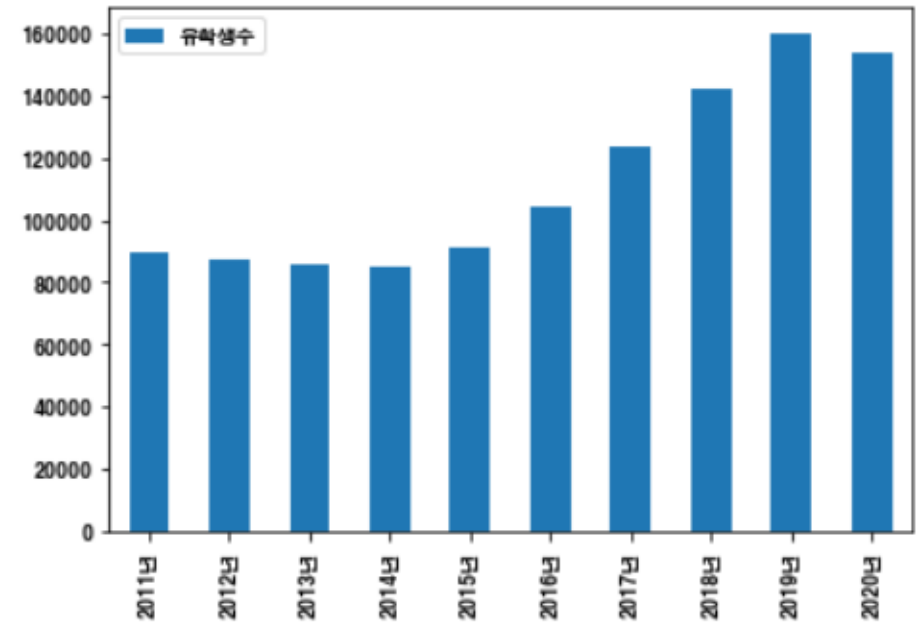
EDA1 – graph1



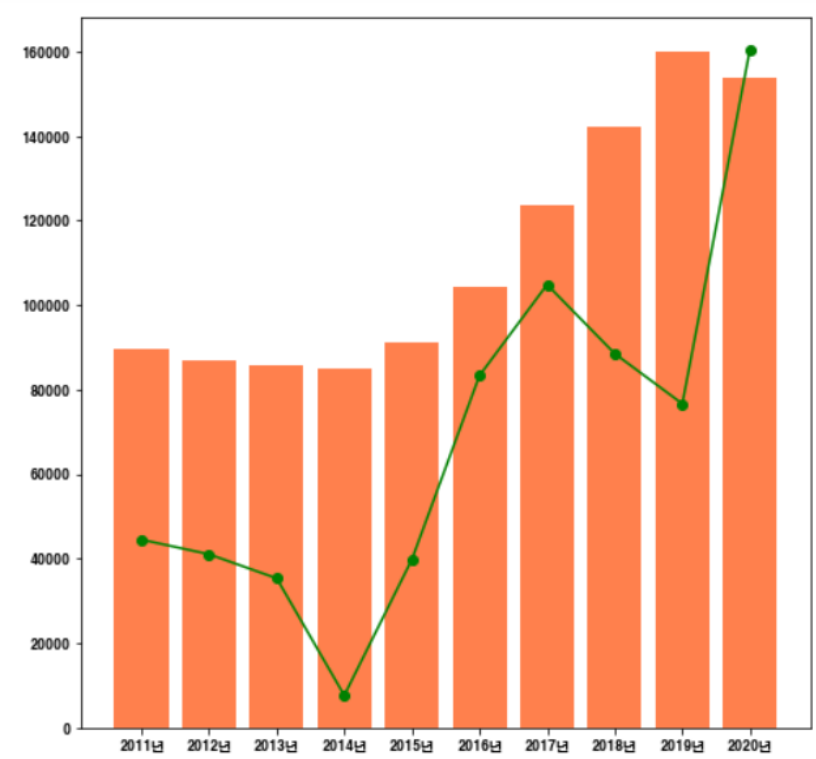
EDA1 – table2

	2011년	2012년	2013년	2014년	2015년	2016년	2017년	2018년	2019년	2020년
유학생수	89537	86878	85923	84891	91332	104262	123858	142205	160165	153695

EDA1 – graph2



EDA1 – graph3(graph1+graph2)



EDA1 interpretation

EDA1 used two datasets.The first is a database of indicators of the financial soundness of universities located in Korea. (table1) In the institution investigating this data, indicators of whether each university's finances are stable were shown through the 'University Risk Index by Region'. They calculated this as 175/{(new student enrollment rate index*1.5) + enrollment number index+tuition income index} and converted it into a standard score, and judged that the university's financial status was dangerous if this index exceeded 1. We tried to rearrange this index by region and show it as a curve graph, and as a result, we were able to obtain an upward right graph that increases financial risk as the year increases. (graph1)

The second is the data set showing the increasing trend of foreign students in Korea. (table2) Data from 2011 to 2020 were used, and this year's data has not yet been compiled because the agency investigating it conducts the survey every December. This data set was displayed in a bar graph to express the trend of increasing international students over the year at a glance, through which we were able to express that the number of international students increased over the year. (graph2) One thing that stands out is the fact that the steadily

increasing trend of international students slowed down by 2020. Rather, it did not increase but decreased, which can be said to be the most reasonable possibility to see as an outlier due to COVID-19.

This visualization plays a role in supporting the article that appears in financial part 6 in our biography process. The more the financial soundness of each university decreases, the more foreign college students are recruited, which can be seen as a positive correlation between the graphs. (graph3)

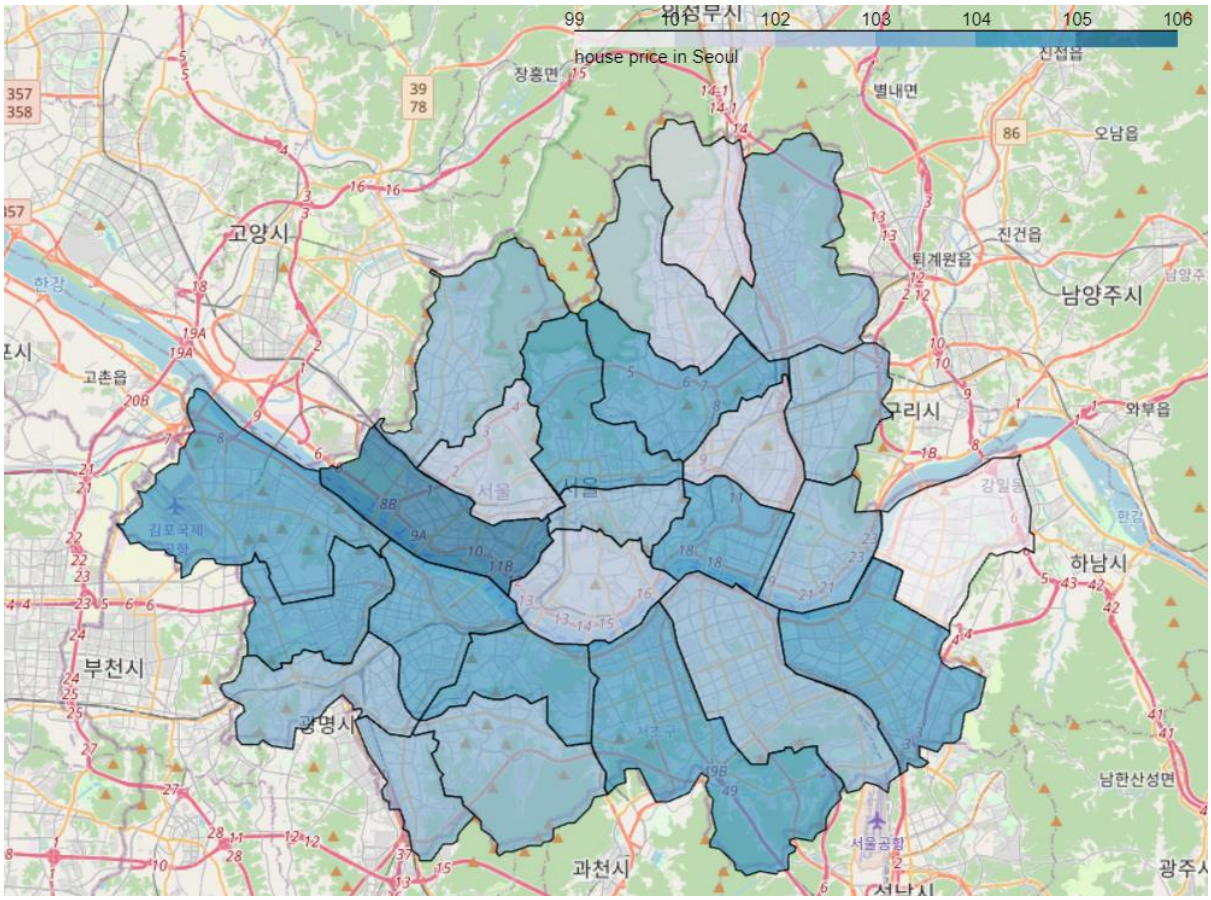
One question raised in this process, however, is that despite attracting as many foreign students as they have, the university's finances are still bad, and rather worse. If interpreted differently, it can lead to the opposite of the results we expect. However, this is a difficult part to judge hastily, and there is also a point where it is difficult to judge because the financial soundness of universities is not a factor related only to foreign students. You can also find an article saying that COVID-19 exists as a representative external factor and that the Korean University Association, which judged that it was actually threatened by this factor, asked the government for help. (https://www.yna.co.kr/view/AKR20210701075100530) Therefore, it is difficult to see this unconditionally as having a negative correlation, and as mentioned above, I think it is the most likely interpretation to think of it as a material that supports the particle of financial part 6.

EDA2. Housing price in Seoul

EDA2 – Table

	자치구	종합	아파트
1	서울시	103.6	105.0
2	종로구	104.8	106.1
3	중구	103.4	105.0
4	용산구	102.5	103.6
5	성동구	105.1	106.1
6	광진구	103.7	105.1
7	동대문구	102.5	103.5
8	중랑구	102.9	104.1
9	성북구	105.1	106.4
10	강북구	102.8	106.4
11	도봉구	101.4	101.6
12	노원구	103.4	103.9
13	은평구	103.7	103.3
14	서대문구	102.0	102.6
15	마포구	106.3	108.4
16	양천구	104.8	107.4
17	강서구	104.2	107.5
18	구로구	103.0	104.6
19	금천구	103.8	106.0
20	영등포구	104.6	105.4
21	동작구	104.7	107.2
22	관악구	103.3	106.6
23	서초구	104.2	105.3
24	강남구	103.9	105.4
25	송파구	104.3	106.1
26	강동구	99.4	99.4

EDA2 – Map data



EDA2 interpretation

We were able to find a lease price dataset of the autonomous district of the Republic of Korea(table), that is, a dataset for the overall distribution of housing prices. Using these datasets, we will be able to help determine the location of the dormitory, and in terms of cost, we will be able to create a means of persuading the mayor of Seoul and university presidents.

First, the comprehensive lease price data of each autonomous district were rearranged and displayed, and this was intended to be displayed on the map using the folium module. Like EDA3, which represents the number of foreign students in Seoul, colors are differentiated and painted according to the house price, and the darker the color, the higher the house price. As commonly known, the Gangnam area in Seoul shows a distinctly dark color, the area with a large number of international students also shows a distinctly dark color(We can know this in the previous EDA process.). One thing to note is that Cheongdam-dong, Daechi-dong, and Samseong-dong, which are known to be the most expensive housing prices in the country, show surprisingly light colors (outliers). Since these areas are areas where very high-income people live, it can be hypothesized that the number of samples in the lease model itself is small, and even a small number of samples are samples belonging to the lower classes.

By mapping this lease dataset, we were able to simply represent housing prices in Seoul (map data). Two regrets are that it is difficult to know the exact data because it shows housing prices for a large area based on the autonomous district of Seoul, and it is somewhat unreasonable to look at a wide area and determine the location. To compensate for this, we looked for a new dataset for administrative {-Dong}, but failed to find them. Next, since the data does not directly represent the data of the land price in building construction, it is difficult to accurately convey facts in terms of cost. This can also be said to be a problem caused by the inability to find a data set that directly investigated the land price.

EDA3. Distribution of international students in Seoul

EDA3-table1

	학교명	주소	행정구	행정동
0	서울시립대학교	서울 동대문구 서울시립대로 163 (전농동 90번지)	동대문구	휘경2동
1	서울여자간호대학교	서울 서대문구 홍제3동 서울여자간호대학	서대문구	홍제2동
2	서울여자대학교	서울특별시 노원구 화랑로 621 서울여자대학교	노원구	공릉2동
3	서일대학교	서울 중랑구 서일대학길 22(면목동 49-3) 서일대학교	중랑구	면목제3.8동
4	성공회대학교	서울 구로구 항동 성공회대학교	구로구	오류2동

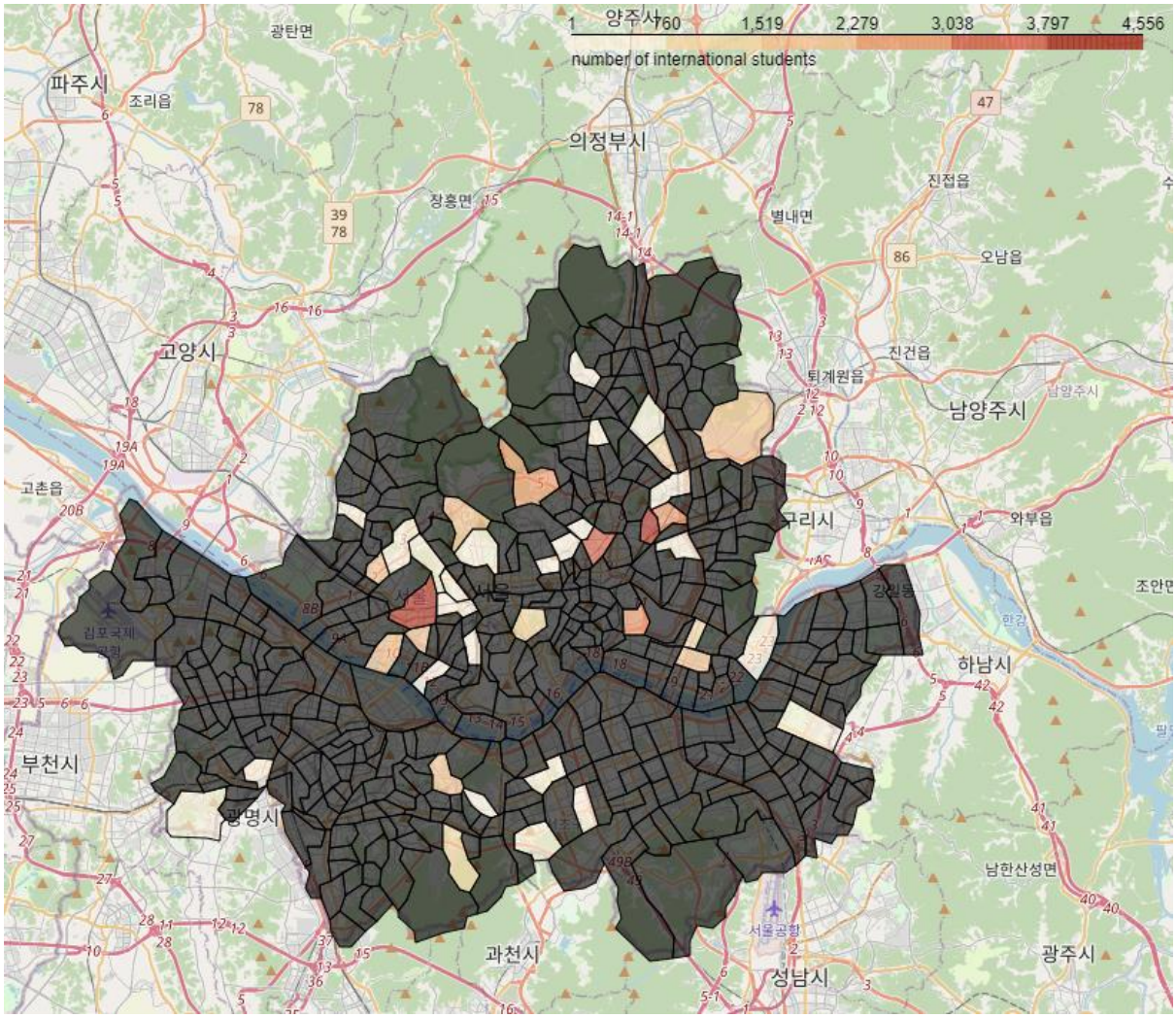
EDA3-table2

	Unnamed: 2	Unnamed: 8
51	서울대학교	1229
52	서울대학교 대학원	795
53	서울대학교 국제대학원	190
54	서울대학교 보건대학원	8
55	서울대학교 행정대학원	59

EDA3-table3

	학교명	주소	행정구	행정동	Unnamed: 2	Unnamed: 8	sum
0	서울시립대학교	서울 동대문구 서울시립대로 163 (전농동 90번지)	동대문구	휘경2동	서울시립대학교	456	460
1	서울여자대학교	서울특별시 노원구 화랑로 621 서울여자대학교	노원구	공릉2동	서울여자대학교	199	1882
2	서일대학교	서울 중랑구 서일대학길 22(면목동 49-3) 서일대학교	중랑구	면목제3.8동	서일대학교	13	13
3	성공회대학교	서울 구로구 항동 성공회대학교	구로구	오류2동	성공회대학교	133	133
4	성균관대학교	서울 종로구 명륜동3가 성균관대학교	종로구	종로1.2.3.4가동	성균관대학교	3376	3376

EDA3-map data



EDA3 interpretation

There were two datasets. The first one was about the information about the university's location such as administrative dong, real address of the university etc. (table1) The second one was about the number of international students in each university in 2021. (table2) So, we merged two data sets and made a new data set including the university's address and number of international students.(table3) We made assumption that the international students in each university will be living in somewhere in the same 'Dong' as the university. So using the folium library, we decided to make a map about the number of international students in each 'Dong'.

This map helps us to roughly see where the international students are living in Seoul. Also if we zoom the map, we can see where the university is locating specifically. It was interesting to find out that the universities are concentrated in the center of Seoul. Also, the 'Dong's that

many students were living in was ‘신촌동(Sinchon-dong)’, ‘회기동(Hoegi-dong)’, ‘종로1.2.3.4가동(Jongro1.2.3.4-dong), and ‘안암동(Anam-dong)’, which are pretty close to the center of Seoul. So when thinking of the distribution of international students, for their convenience, we think it is better to build accommodation near the center of Seoul rather than the areas far from the center of Seoul. If we build accommodation in ‘강서구 (Gangseo-gu)’, ‘양서구(Yangseo-gu)’, 송파구(Songpa-gu)’, ‘강남구(Gangnam-gu)’, 강동구(Gangdong-gu)’, especially near ‘경기도(Gyeonggi-do)’, students will have difficulty commuting to school.

There are two weakness in this EDA: 1) There were two datasets at first, one showing the university’s address, and the other one showing how many international students are in each university. I merged two datasets by ‘inner’, so there are universities that are not showed in the map. There was university that was in the first dataset, but not in second dataset, and also university that was in the second dataset, but not in first dataset. So when making the map, I could only use the dataset of universities that were in both dataset. 2) I only made a map including undergraduate students. It would have been better to include graduate students, so that the map gets more accurate.

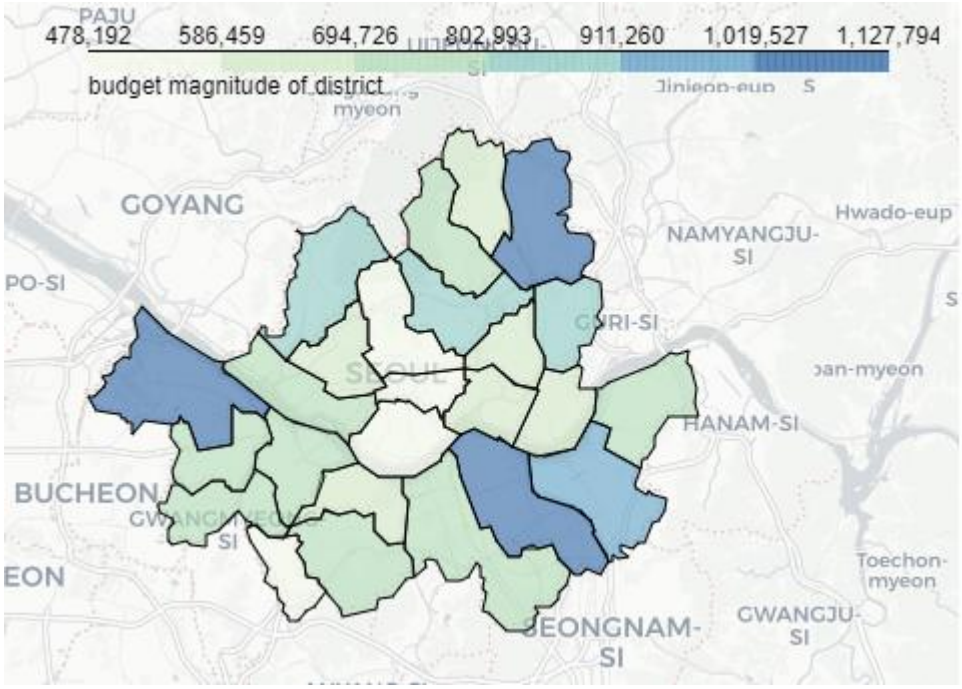
Due to 1),2) weakness, we know that the map is not accurate, but since it was EDA process, we just saw the brief distribution of international students, and we will make more accurate map in the later process if needed.

EDA4. Budget magnitude of each district

EDA4-table1

	자치구	예산
0	종로구	478192
1	중구	532170
2	용산구	502568
3	성동구	601066
4	광진구	610063

EDA4-map data



EDA4 interpretation

We had the dataset of showing the budget magnitude of each district(table1) According to the article about Nuri-Hall, Korea's first dormitory for international students(bibliography 11-accommodation), 8 universities in Daejeon invested 4.3 billion won and Daejeon City 4.3 billion won. So, we thought that we might get investment from Seoul City, and by studying about budget from the book that Seoul city made (bibliography1-financial aspects), we thought that we might get investment from the district in which we build the accommodation.

So, we used the folium library to make the map of budget magnitude of each district. The unit on the map is a million won, and as the color goes to blue, it means they have more budgets.

Also if the color goes to light-green, it means they have less budgets. Since, the budget of each district vary a lot, we thought it might be critical in getting the investment.

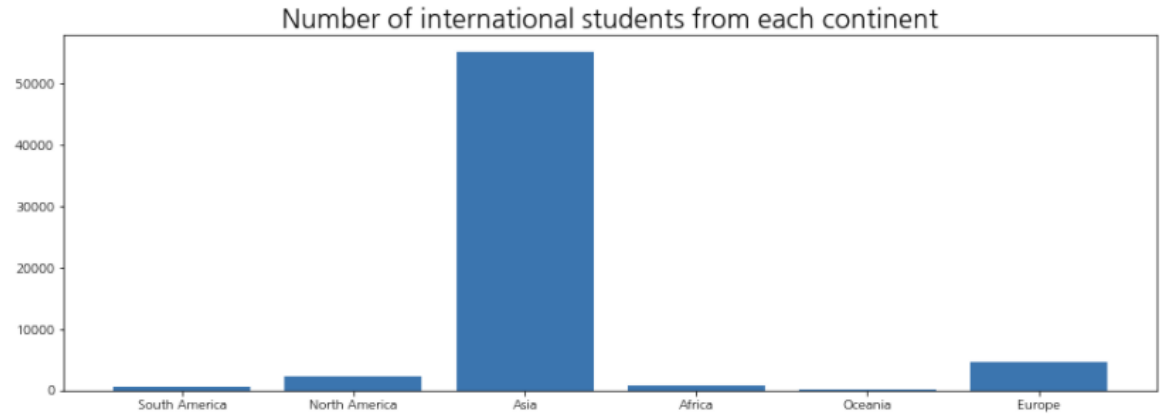
If we are building accommodation in certain district, we thought that we could get the financial investment from that district. So, in order to get as much financial investment as possible, we thought that we should build accommodation in the district that has more budget. So, we should consider building the accommodation in top3 district (노원구(Nowom-gu), 강서구(Gangseo-gu), 강남구(Gangnam-gu)). However we cannot simply think that the district with more budget can support us with more money, so in later process, we should consider which district has more money 'to invest in our accommodation'.

EDA5. Nationality of international students

EDA5-table1

연도	학제	학교명	학교상태	본분교	시도	시군구	설립	대륙	국가명	동포여부	총계
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	북아메리카	미국		1
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	네팔		7
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	라오스		2
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	몽골		7
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	베트남		143
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	우즈베키스탄		3
2021	대학교	강릉원주대학교	기존	본교	강원	강원 강릉시	국립	아시아	인도네시아		5

EDA5-graph1



EDA5-table2

	Unnamed: 9	sum1
717	중국	35097
702	베트남	7974
716	일본	2214
699	몽골	1829
692	미국	1785
769	프랑스	1229
696	대만	1098
710	우즈베키스탄	1010
749	독일	755
750	러시아	694

EDA5 interpretation

We saw a study that said the loneliness experienced by international students is due to the absence of the preferred cultural and/or linguistic environment. (bibliography2-international students) So, we thought that it might be good to build the international student accommodation where international students can meet the similar cultures. So, we just wanted to see the nationality of the international students in Korea.

We had the dataset showing the international student's nationality in each university. (table1). Rather than considering the specific university, we just wanted to know the nationality of whole international students. At first, we wanted to know the number of international students from each continent. We made a bar graph about it (graph1), so we could see that most of the students were from Asia.

We wondered why so many students are from Asia, so we analyzed the specific nationality of the students. We could see that so many students were from Asia because many students were from China(35097 students), Vietnam(7974 students), Japan(2214 students), and Mongolia(1829 students). We thought that many students are from those countries due to short distance between countries, race similarity, and cultural similarity.

EDA5 might not be directly connected to our topic, but it helped us to know more specifically

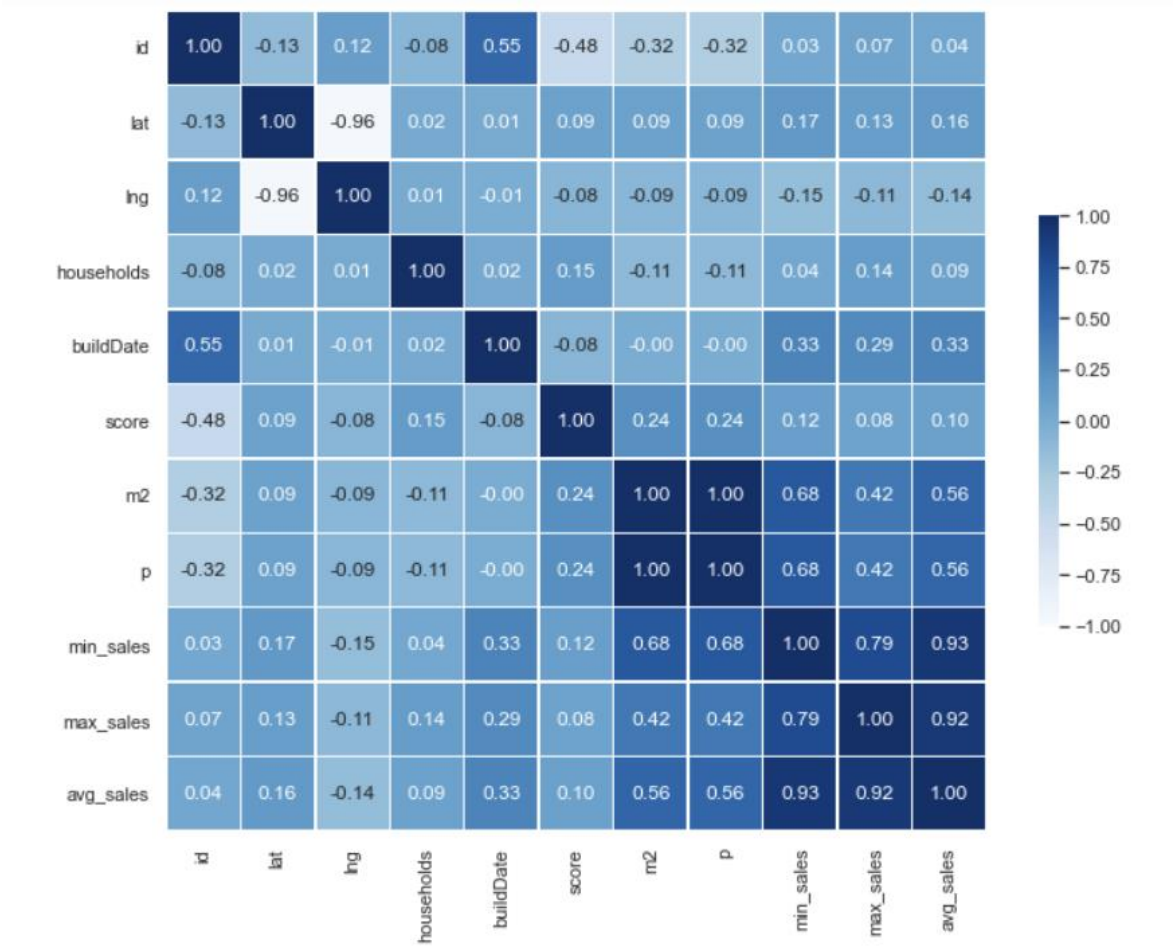
about the international students in Korea.

EDA6. Correlation between features of apartment in Seoul

EDA6-table1

	id	lat	lng	households	buildDate	score	m2	p	min_sales	max_sales	avg_sales
0	2766	37.681604	127.056592	492	200006	4.3	139	42	60100.0	62000.0	61000.0
1	5860	37.679290	127.057021	468	200105	4.1	105	32	48600.0	52200.0	51000.0
2	15564	37.676882	127.058075	57	200502	4.8	86	26	36000.0	46000.0	40500.0
3	3700	37.675277	127.060001	216	199509	4.8	102	31	34000.0	34800.0	34500.0
4	6204	37.676381	127.058361	165	200306	4.8	91	28	27900.0	50300.0	40000.0

EDA6-correlogram



EDA6 interpretation

We needed land price data for building a new accommodation, but we only had house price data. So, we thought that if we analyze the dataset of house price data, we might have some ideas of what are the features that give effects to house price. If we know how house prices are decided, then we thought that we might roughly estimate the land price.

We had the dataset of apartment prices in Seoul. This dataset had several features such as location, number of households in residence, build date, total evaluation(score), the area of the apartment(m2), the number of floors(p), and the descriptive statistics of sales price.

We just wanted to know if there's any correlation between the features. We followed the pearson's correlation coefficient.

Degree of correlation:

- **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
- **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
- **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
- **No correlation:** When the value is zero.

We thought we should not consider about id, latitude, longitude because they were not quantitative information.

There was moderate degree between build date and sales. Although build date was not quantitative feature, the bigger numbers meant later-built, so it seemed that later-built apartments were more expensive. It was reasonable because generally people like new things. There was high degree between the area of the apartment and sale. It seemed reasonable because more areas, more expensive. Also, since the correlation between area and other features, and the correlation between number of floors and other features are same, so we thought that there is direct proportion between the area and the number of floors. So we thought that the area in each floor is the same in this dataset. And lastly, the correlations between the sales were very high and we thought this was reasonable.

Also, it was interesting to find out that the correlation between some features (build date, area, floors) and the maximum sales were a little lower than correlation between features and minimum or average sales. But we could not find the reason to it.

So if we have to roughly estimate the land price, we should consider the build date and area. For example, let's say A apartment and B apartment price is same, and the build date is same. But if A has 12 floors, and B is 10 floors, then the land price of B is expensive than A.

However, there is still problem with data. In order to build a new accommodation, we need to find the dataset of the land with no buildings.

[Summary]

The first 4 EDAs were mostly focused on the geographical and financial aspects of the international accommodation campus. For the localization, the goal was to find districts of Seoul that are near most of the universities, that have high budgets and whose average rent are low. We know the rent price is not necessarily proportional to land price, but we assumed that if the rent is more expensive, the price of land will be too because it is well situated. Nevertheless, some data about the price of land and construction in each district would be useful in this case. If we fail to find the land price data during the project, then we should roughly estimate the land price based on EDA6. Furthermore, the number of students in each university as well as their location will prove useful in our opinion as we will be able to determine the optimal location regarding distance to universities to facilitate the greatest number of student's daily lives while also considering the two other criteria. The problem with this EDA was mostly the loss of data while joining the two tables that could be fixed by finding two tables that match or just adding individually data found as things progress. Regarding the financial aspect of the subject, we can see in EDA 1 the increase in the number of international students coming to study abroad in Korea. Unfortunately, we also see the financial instability of universities in the country increase, but it would be interesting to collect more data about the income brought to Korean universities by foreign students to determine whether internationalization of the studies has a positive impact or not on the universities and more broadly on the country's economy. Finally, the EDA5 brings insight on the nationality of foreign students, which can help the project as it will allow to create a more familiar environment for those students to reduce the feeling of loneliness and other issues that we talked about in our bibliography and the EDA. To conclude, the EDAs conducted helped a lot for the location of the campus, which is one of the major questions of the project as well as gave us leads to improve the response given to the financial question and on the ways to make this campus more attractive to international students.