**Title**

*Exploratory Data Analysis of the Seoul Bike*

**Student ID, Name**

2020310243, 김보현

**Introduction**

1) Seoul Bike is an unmanned public system to lend bikes. This system has been managed by Seoul since 2015. It was begun with 967 bicycles, but as more citizens use them, bicycles are increasing to about 37,000 bikes. However, as this program expands, several problems have happened. Firstly, although bikes are increasing gradually, the place where to rent them is quite a few. It means those who want to use the bike may wander to find it and even have trouble returning it. Secondly, bikes on the streets can hinder citizens’ movement when people do not return the bike to the rent spot accurately. Furthermore, more and more bikes are broken down since they are just left on the street without management.

2) Seoul Bike became one of the most important means of transportation for Seoul citizens. However, as more people use this system, the problems are getting serious. Without solving these problems, we can’t enjoy this system for a long time, so we should think of solutions about how to manage bikes and how people enjoy bikes freely without inconvenience.

3) By grasping trends in using bicycles annually with rental rates, I will look at system size growth rate. Also, I am planning to focus on 2020 data to check whether COVID-19 affects bike rental degrees. Especially comparing the amount of using bikes between regions in Seoul, I want to suggest where the city should put additional bikes and return spots. I hope my suggestion helps to maintain this system and relieve the inconvenience of citizens.

**Data**

1) The data is about Seoul Bike rental information for each spot (from January 2017 to January 2021). To compare the bike rental rate with COVID-19, I used the data about the number of confirmed patients. For the visualization of data, I used Seoul map JSON file, too.

2) I downloaded the dataset about Seoul Bike from Internet (<http://data.seoul.go.kr/dataList/OA-15249/F/1/datasetView.do>). In the case of COVID-19 dataset, it is from the website (<http://ncov.mohw.go.kr/>). I downloaded Seoul map file from Github(<https://github.com/southkorea/seoul-maps/blob/master/kostat/2013/json/seoul_municipalities_geo_simple.json>).

3) data description

* There are 10 dataset files about Seoul Bike since some files are organized on monthly basis. To process the data effectively, I combined and organized each file for months into three CSV files. The first dataset which contains bike return information from January 2017 to June 2018 has 15572 rows and 4 columns that are date, spot number, spot name, and the number of rentals. Another dataset that is from July 2018 to November 2018 has 8038 rows and equally 4 columns. The last dataset contains information from December 2018 to January 2021 has 45279 rows and 4 columns that are Seoul district group (like ‘Gangnam-gu’), spot, date, and the number of rentals. The reason why I didn’t combine three files into one CSV file is that each dataset has a different column name or data type. So, I used a total of 3 CSV files and manipulated each dataset on python.
* COVID-19 dataset has 500 observations and 4 columns. They are about date, total patients, domestic patients, and overseas patients.

**Methods**

1. Preliminary process

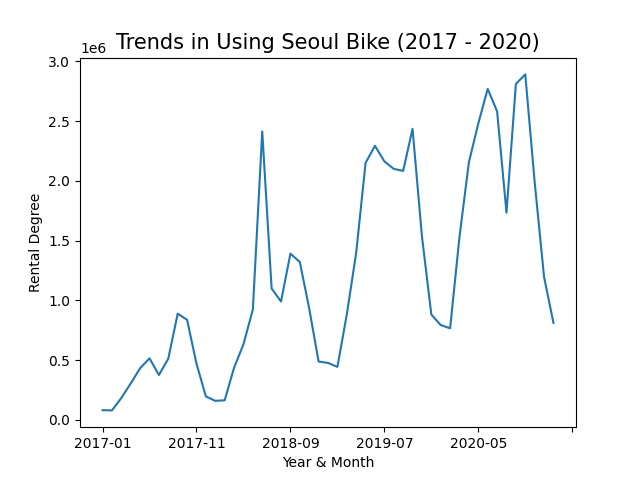
* Data loading: Using read\_csv function in pandas package, three CSV files are loaded. I changed its column name to English words and encoded each to ‘cp949’.
* Data cleaning: I removed NaN values by using dropna(). Although three datasets all have ‘date’ column, each data type was different. So, I changed each ‘date’ column to ‘integer’ type.
* Data frame: I made a new data frame that has total information, from January 2017 to January 2021. I used pd.concat() for combining the number of rentals and date and grouped this data frame by ‘date’ with the sum of the ‘rental’ column(how many people rented the bikes).

2. Visualizations for EDA

* I used matplotlib and seaborn package to plot several graphs. Focusing on each dataset, I created a new data frame by filtering original data.
* For making the map, I installed and used folium package. Through this package, I could draw the map with the data containing Seoul district information.

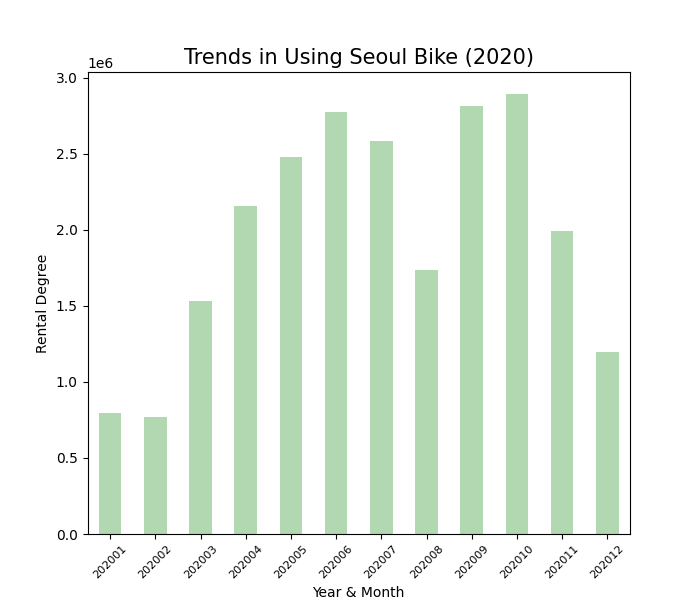
**Experimental Results**

* Trends in using Seoul Bike from 2017 to 2020



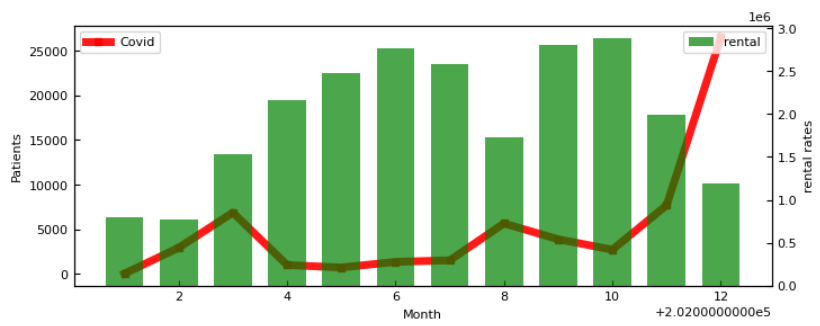
Firstly, I checked general trends of people who borrow the Seoul Bike. In this plot, X-axis is the date and Y-axis refers to a rental degree. I could see that the number of users has increased extremely. In 2017, there were not many users of the Seoul Bike but in 2018, people started to use it frequently. Until now, we can see that more people than in 2017 are using this bike.

* Trends in using Seoul Bike focusing on 2020



And then, I wondered whether COVID-19 affects to bike rental degree. In detail, I want to check if the number of confirmed patients is increasing, people do not use bikes to stay at home safely. So, I focused on only 2020 data. At this plot, I could see that more people rented bikes in June, September, and October.

* Hypothesis1: There is a specific correlation between COVID-19 and Seoul bike rental degree.
* Correlation between COVID-19 and bike rental

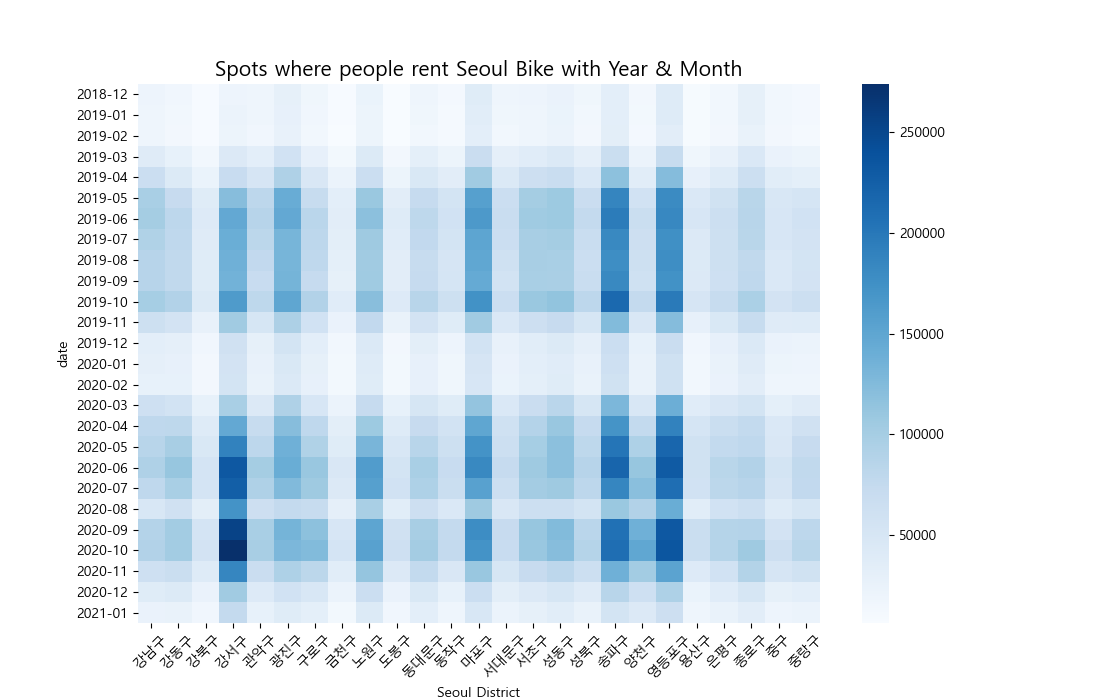
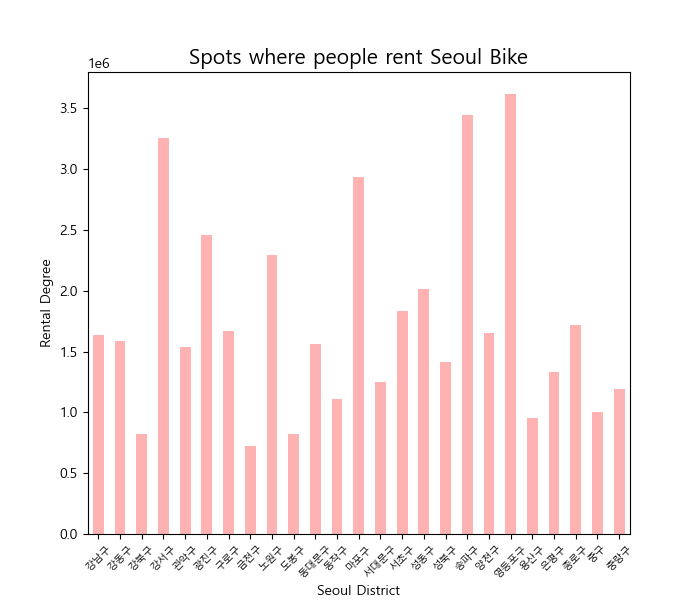


In this plot, X-axis refers to the month of 2020, and left Y-axis is the number of total patients, and the right Y-axis is the rental degree. When the number of patients increased especially in August and November, fewer people used bikes than over the past few months. In contrast, when the number of patients decreased like June or October, lots of people used bikes. It seems like there is a specific relationship. But, variation of bike rental degree can be because of season or other variable factors. Therefore, I could not conclude that variation of rental degrees is only because of COVID-19, but through this plot, I concluded that COVID-19 is one of the factors related to the Seoul bike rental degree.

* Hypothesis2: Bike rental degree will be the highest at spots near Han River Park.

Next, I wondered about the spots where most people rented bikes. By searching where people used the bikes most, I wanted to suggest that we should assign more bikes and spots to it. However, it is difficult to compare each rental spot on a graph because there are so many spots in Seoul. So, I divided it into districts. Before making related plots, I made another hypothesis that people will enjoy bikes especially in spots near the Han River. It’s because people mainly ride bicycles at the park, and Han River Park is one of the biggest parks in Korea with popularity.

* Spots where people rent Seoul Bike



By making bar plot and heatmap, I found specific Seoul Districts where people enjoyed renting bikes. From the left bar plot, I could know that ‘Gangseo-gu’, ‘Songpa-gu’, and ‘Yeongdeungpo-gu’ are the top 3 districts with the highest rental degree. According to the right heatmap, I could also know that especially in October, people enjoyed riding bikes.

지도이(가) 표시된 사진

자동 생성된 설명

With this map, I could look at where people rented bikes mainly at once. Deep blue parts are ‘Gangseo-gu’, ‘Songpa-gu’, and ‘Yeongdeungpo-gu’. Besides these three districts, I could find that ‘Mapo-gu’, ‘Gwangjin-gu’, and ‘Nowon-gu’ with deep green places show a quite high bike rental degree. Through these plots, I could prove that the hypothesis is valid. It’s because the top 3 districts all have Han River Park. There is Yeouido Han River Park in Yeongdeungpo-gu, Gamsil Han River Park in Songpa-gu, and Gangseo Han River Park in Gangseo-gu.

텍스트이(가) 표시된 사진

자동 생성된 설명

Finally, I checked specific spots with the highest rental rate in each strict. As a result, in Gangseo-gu 2701(spot number) spot, in Songpa-gu, 1210 spot, and in Yeongdeungpo-gu, 207 spot where is the front of Han River are showing the highest rental degree. So, I want to suggest that the city Seoul should assign more bikes and rental spots especially in 3 districts, Gangseo, Songpa, and Yeongdeungpo. Furthermore, the city should place additional people who manage and fix bikes, especially in these spots.

**Conclusions**

1. It was the first time to choose a dataset and topic, manipulate the dataset, and have EDA alone myself. This overall process let me understand things I studied through lectures well. Especially, by cleaning each dataset by removing missing values, converting data type, and deleting useless columns I felt the importance of data cleaning. I could learn lots of things through this project.
2. There are 3 findings of my findings. Firstly, I could verify the overall users of Seoul Bike for 4 years. Through the plot, I checked the scale of this rental system and its fast development speed. Second, I discovered the correlation between COVID-19 and bike rental degree. Of course, as I mentioned COVID-19 is just one of the factors that affect riding bikes. But, at the plot, I could see the correlation. Finally, I knew where people used bikes the most. Especially in Gangseo-gu, Songpa-gu, and Yeongdeungpo-gu, most people enjoyed bikes. This result verifies my second hypothesis that most people will use bikes near Han River Park. Through these findings, I suggest where we should put additional bikes and rental spots.

**References**

* <https://www.sisul.or.kr/open_content/traffic/bikeseoul.jsp>
* https://www.outsourcing.co.kr/news/articleView.html?idxno=85884
* <http://data.seoul.go.kr/dataList/OA-15249/F/1/datasetView.do>
* <https://github.com/southkorea/seoul-maps/blob/master/kostat/2013/json/seoul_municipalities_geo_simple.json>
* <http://ncov.mohw.go.kr/>