Innovative plan for improving bike services in Denmark with the focus on DTU students

Pooya Ghasemian Sorboni

Coursera capstone project, January 2021

Preface

The current work has been submitted as Capstone project to Coursera as a part of fulfilment of the IBM Data Science course organized by Coursera.

The thesis is aimed to optimize the current bike service system in Denmark with a great focus on attending the attention of DTU student community as the main benefit receivers of this plan. Such idea has been offered as per my personal experience of studying at the technical university of Denmark (DTU) and facing the real-life challenges of my student period.

Acknowledgments

I would like to appreciate Coursera team for providing the great opportunity to expand my knowledge in the field of data science. As a person with almost zero background in programming, this course gave me a great knowledge about the fundamentals required to obtain for being a good python user.

My big gratitude to Emad, Mette, and Amin jan for recommending me this course.

At the end, I hope to continue learning and implementing this knowledge in my professional field of study, which is Energy industry.

1 Introduction

1.1 Background

Denmark is a Scandinavian country located in the north of Europe and is known as one of the big leading nations toward green mobility and the reduction of fossil fuel consumption. Using bikes is one of the most common movements toward a green community in Denmark. In fact, nearly 60% of Dnish people use bike for their daily needs [1] Here I have provided some of the main reasons I assume that have made bikes very useful in Denmark:

- The geographic condition of Denmark as a country that does not have high variation of altitude has made it a good spot for bike lovers.
- In other words, having no hills and ups and downs in the streets make it very easy for different age range of people to use their bike for their daily commuting route to their school, work and etc.
- The high taxes of the fuel and rather expensive prices for public transportations, have turned bikes very valuable means of transportation for people

1.2 Problem

Despite the fact that using bikes is very common in Denmark, the bike services are not very reasonable. This means there are either not enough spots that you can have your bikes fixed, or the prices are very expensive.

Based on DTU website, there are currently 12894 active students studying in the technical university of Denmark (DTU) [2], which is one of the most famous engineering universities in the Europe with the QS world University ranking of 103 [3] When it comes to students, financial aspects become one of the most crucial points. However, there are two issues I have faced as a student in DTU:

- There is high price of bike services (it can be around 300 dkk for a normal service) [4]
- There are not sufficient service points around DTU
- There is a time limit

For me, the third one was very important. Imagine you are stuck somewhere in a cold day at the middle of the night, but you face a flat tire. Some areas are not covered by the public transport after some hours of the day (for example after 8 pm), imagine you face an issue in your bike. what can you do?? :)

1.3 Who can benefit

This idea can benefit both the business owner and the business user. Having access to the great business area of DTU students provide a big number of permanent users (there is always new students coming to DTU). Moreover, students can benefit the better and cheaper bike services.

2 Data usage (magic of data)

2.1 How to use data science to fix this issue

The best usage of data is to find the best spot for placing a bike repair shop or as stated earlier, setting a vending machine to sell the necessary emergency things required for having a self-service repair for the community of people using bikes around DTU. The topics that can be covered are:

- Finding the population of people living in the vicinity of DTU
- showing the best spots that can be spotted easier for people (around stations, DTU labs, bike shelters, etc), so you can do the repair in a relax area (no wind, rain, snow, etc)
- Showing the profit of giving discounts, outside of working hour services, etc, to attract costumers

2.2 Useful data

- Importing the map of Denmark, showing the location of DTU on the map
- Showing the neighborhoods (postal codes) that are covered around the university, so those can be benefited also
- Importing the population of each section as per postal code to determine the quantity of people for the potential business community
- Clustering the data to check the areas that can be the best choices as the business receivers
- Calculating the revenue of the business considering different scenarios of service price

2.3 Data sources

As stated in the previous section, the main data required for this project include:

- The coordination of Denmark, with focus on the technical University of Denmark (DTU)
- The postal code data for focusing on the areas that can have the best location of the business
- The population of each area to find out the quantity of the business users

The stated data are taken from different resources, mainly the information center of Denmark that provides the annual data regarding the latest changes in the population, post codes and so on. For the sake of a better understanding each reference has been stated in its particular section and all the list of references could be found in the Reference section of this report

2.4 Data cleaning

As the data has been taken from different resources a big attempt has been made on data cleaning. The first data set which is imported for the sake of data analysis is the postal code of the different areas in Denmark. As per the great job of the data center of Denmark, the data set contains the coordination data (longitude and latitude) for each postal code aswell. This made the data analysis job much more optimal as there was no requirement for an extra data importing and correspondingly possible data cleaning.

Table 1: Imported data for the postcode and coordination

```
df_DK.head()
```

	Postal Code	Place Name	Latitude	Longitude
0	800	Høje Taastrup	55.6500	12.2833
1	877	Valby	55.7333	12.1729
2	900	København C	55.6777	12.5709
3	910	København C	55.6777	12.5709
4	929	København C	55.6777	12.5709

```
df_DK.shape
```

```
(1182, 4)
```

After importing the data, a review of the data in pandas was made, to investigate how the data is reported. As the initial format of the data was well reported, for the sake of simplicity, the only adjustments made in this section was to update the column names that shows the area, to “Borough”.

Table 2: Adjusted postal code and coordination data of Denmark

	Postal Code	Borough	Latitude	Longitude
0	800	Høje Taastrup	55.6500	12.2833
1	877	Valby	55.7333	12.1729
2	900	København C	55.6777	12.5709
3	910	København C	55.6777	12.5709
4	929	København C	55.6777	12.5709
...	...	...	...	...
1177	9940	Læsø	57.5581	10.4927
1178	9970	Strandby	57.5000	10.5000
1179	9981	Jerup	57.5333	10.4333
1180	9982	Ålbæk	57.5956	10.4094
1181	9990	Skagen	57.7209	10.5839

1182 rows × 4 columns

In the next step, the second data file which includes the population per area/region in Denmark was imported. In order to keep everything with the same format, the column which indicate the region names was renamed to “Borough”. Then, both data sets were merged to have all the data in one table.

Table 3: the population data set for our investigated areas

	Borough	population
0	Copenhagen	637936
1	Frederiksberg	104351
2	Dragør	14497
3	Tårnby	42757
4	Albertslund	27500
...	...	...
94	Morsø	20135
95	Rebild	30438
96	Thisted	43217
97	Vesthimmerlands	36458
98	Aalborg	219310

99 rows × 2 columns

Table 4: Merging the data to have a single dataset

	Postal Code	Borough	Latitude	Longitude	population
0	2000	Frederiksberg	55.6794	12.5346	104351
1	2600	Glostrup	55.6667	12.4000	23342
2	2605	Brøndby	55.6500	12.4167	35215
3	2610	Rødovre	55.6806	12.4537	41025
4	2620	Albertslund	55.6569	12.3638	27500
...	...	...	...	...	...
69	9100	Aalborg	57.0480	9.9187	219310
70	9700	Brønderslev	57.2702	9.9410	36183
71	9800	Hjørring	57.4642	9.9823	64213
72	9900	Frederikshavn	57.4407	10.5366	59187
73	9940	Læsø	57.5581	10.4927	1767

74 rows × 5 columns

After that, a new data set that includes the area details (the size of the region) was imported. Please note that the main aim for doing such attempt was to provide a better idea for the readers of the report as some regions might be bigger that the rest and by having the size data, it makes it much easier to interpret the data and the results of the research. Following the same approach, the column showing the area name was renamed to "Borough". Moreover, the new data was merged with the original table to have a single data table. Please note, this data was imported from Wikipedia, therefore it was initially converted to the desired pandas format which is used in our analysis.

Table 5: Attaining the area information

```
<title>List of municipalities of Denmark - Wikipedia</title>
```

LAU-1 code [1]	Municipality	Administrative Center	Area (km²)	Population (2012-01-01)	Region
101	Copenhagen	Copenhagen	74.7	549,050	Capital
751	Aarhus	Aarhus	467.8	314,545	Central
851	Aalborg	Aalborg	1137.2	201,142	North
461	Odense	Odense	305.7	191,610	South
561	Esbjerg	Esbjerg	794.7	115,112	South
630	Vejle	Vejle	1058.8	108,021	South
147	Frederiksberg	Frederiksberg	8.1	100,215	Capital
730	Randers	Randers	747.6	95,756	Central
791	Viborg	Viborg	1408.7	93,819	Central
621	Kolding	Kolding	607.1	89,412	South
740	Silkeborg	Silkeborg	850.5	89,328	Central

Table 6: Attaining the area information

```
<title>List of municipalities of Denmark - Wikipedia</title>
```

LAU-1 code [1]	Municipality	Administrative Center	Area (km²)	Population (2012-01-01)	Region
101	Copenhagen	Copenhagen	74.7	549,050	Capital
751	Aarhus	Aarhus	467.8	314,545	Central
851	Aalborg	Aalborg	1137.2	201,142	North
461	Odense	Odense	305.7	191,610	South
561	Esbjerg	Esbjerg	794.7	115,112	South
630	Vejle	Vejle	1058.8	108,021	South
147	Frederiksberg	Frederiksberg	8.1	100,215	Capital
730	Randers	Randers	747.6	95,756	Central
791	Viborg	Viborg	1408.7	93,819	Central
621	Kolding	Kolding	607.1	89,412	South
740	Silkeborg	Silkeborg	850.5	89,328	Central

Table 7: Converting the data to the required format

	LAU-1code 1	Borough	Administrative Center	Area(km²)	Population(2012-01-01)	Region
0	101	Copenhagen	Copenhagen	74.7	549050	Capital
1	751	Aarhus	Aarhus	467.8	314545	Central
2	851	Aalborg	Aalborg	1137.2	201142	North
3	461	Odense	Odense	305.7	191610	South
4	561	Esbjerg	Esbjerg	794.7	115112	South
...	...	...	...	...	...	...
93	482	Langeland	Rudkøbing	288.8	13094	South
94	492	Ærø	Marstal	90.1	6636	South
95	741	Samsø	Tranebjerg	113.5	3889	Central
96	563	Fanø	Nordby	54.6	3251	South
97	825	Læsø	Byrum	118.9	1897	North

98 rows × 6 columns

Table 8: Merging the data to the initial data set

	postal_code	Borough	Latitude	Longitude	population	LAU-1code 1	Administrative Center	Area(km²)	Population(2012-01-01)	Region
0	2000	Frederiksberg	55.6794	12.5346	104351	147	Frederiksberg	8.1	100215	Capital
1	2600	Glostrup	55.6667	12.4000	23342	161	Glostrup	13.3	21650	Capital
2	2605	Brøndby	55.6500	12.4167	35215	153	Brøndbyvester	21.0	34084	Capital
3	2610	Rødovre	55.6806	12.4537	41025	175	Rødovre	12.1	36883	Capital
4	2620	Albertslund	55.6569	12.3638	27500	165	Albertslund	23.2	27864	Capital
...	...	...	...	...	...	...	...	...	...	...
69	9100	Aalborg	57.0480	9.9187	219310	851	Aalborg	1137.2	201142	North
70	9700	Brønderslev	57.2702	9.9410	36183	810	Brønderslev	633.0	35754	North
71	9800	Hjørring	57.4642	9.9823	64213	860	Hjørring	926.2	66178	North
72	9900	Frederikshavn	57.4407	10.5366	59187	813	Frederikshavn	650.3	61158	North
73	9940	Læsø	57.5581	10.4927	1767	825	Byrum	118.9	1897	North

74 rows × 10 columns

As some of the tables had some mutual data, in this step an attempt was made to delete the similar columns. For instance, most of the tables postal code as a part of their data. Therefore, the extra columns that show the similar data (postal code) were deleted. Similarly, some extra data, such as the specific code of each municipality (region) that has no usage in this study were deleted from our data set.

Table 9: Merging the data to the initial data set

	postal_code	Borough	Latitude	Longitude	population	Area(km²)	Region
0	2000	Frederiksberg	55.6794	12.5346	104351	8.1	Capital
1	2600	Glostrup	55.6667	12.4000	23342	13.3	Capital
2	2605	Brøndby	55.6500	12.4167	35215	21.0	Capital
3	2610	Rødovre	55.6806	12.4537	41025	12.1	Capital
4	2620	Albertslund	55.6569	12.3638	27500	23.2	Capital
...	...	...	...	...	...	...	...
69	9100	Aalborg	57.0480	9.9187	219310	1137.2	North
70	9700	Brønderslev	57.2702	9.9410	36183	633.0	North
71	9800	Hjørring	57.4642	9.9823	64213	926.2	North
72	9900	Frederikshavn	57.4407	10.5366	59187	650.3	North
73	9940	Læsø	57.5581	10.4927	1767	118.9	North

74 rows × 7 columns

3 Data analysis and visualization

After all the necessary data is investigated and data cleaning stage has been implemented. The data processing and visualization have been initiated.

In the first stage, the coordination data of the Hovedstaden region, which is the area that the Technical University of Denmark is located in, has been shown using Folium maps. In order to clearly depict the location of the university, its coordination was imported to the system and marked in the map.

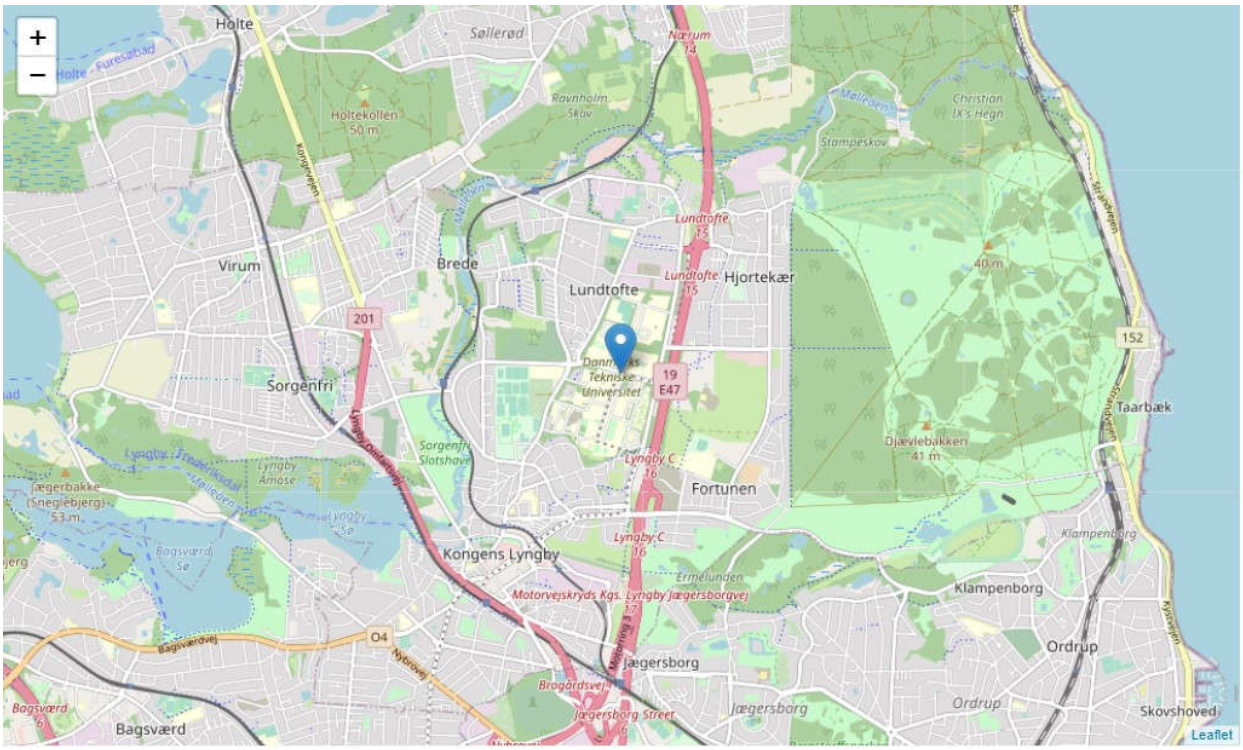

Figure 1: The location of the technical university of Denmark (DTU)

Now that a general overview of the location of the technical university of Denmark is shown, in the next stage the coordination of the other regions of Denmark in Hovedstaden state is depicted in the map. By doing this, a better overview of the possible business area or the future user business beneficial community group can be illustrated.

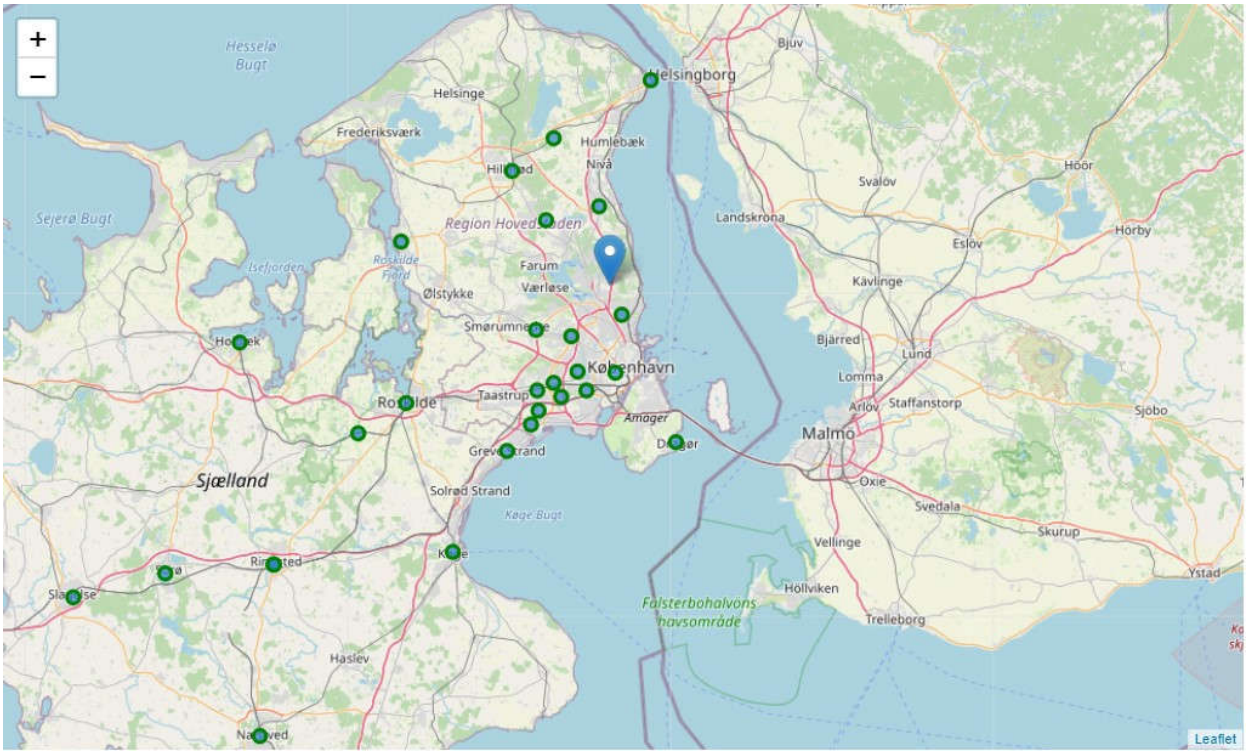

Figure 2: The location of the regions around the technical university of Denmark (DTU)

As stated in the introduction section, the focus of this project is to investigate the possible adjustments and the corresponding business that can be provided for DTU student community. As DTU is located in Hovedstaden state, one can simplify the data set to focus mainly on the data provided for Hovedstaden state. For such purpose, a filter has been created that simply filter the data merely for the illustrated state. By doing this, a better accuracy of the final result will be attained as there would be no noise data (data for the other states than Hovedstaden).

Table 10: Filtering the data to depict the data for Hovedstaden state (stated as Capital)

	postal_code	Borough	Latitude	Longitude	population	Area(km²)	Region
0	2000	Frederiksberg	55.6794	12.5346	104351	8.1	Capital
1	2600	Glostrup	55.6667	12.4000	23342	13.3	Capital
2	2605	Brøndby	55.6500	12.4167	35215	21.0	Capital
3	2610	Rødovre	55.6806	12.4537	41025	12.1	Capital
4	2620	Albertslund	55.6569	12.3638	27500	23.2	Capital
5	2625	Vallensbæk	55.6333	12.3667	16483	9.5	Capital
6	2635	Ishøj	55.6154	12.3518	23059	26.4	Capital
7	2650	Hvidovre	55.6572	12.4736	53448	23.0	Capital
9	2730	Herlev	55.7237	12.4400	28914	12.1	Capital
10	2750	Ballerup	55.7316	12.3633	48978	33.8	Capital
11	2791	Dragør	55.5945	12.6664	14497	18.3	Capital
12	2820	Gentofte	55.7500	12.5500	74623	25.6	Capital
13	2970	Hørsholm	55.8833	12.5000	24919	31.3	Capital
14	3000	Helsingør	56.0361	12.6136	63047	118.9	Capital
15	3400	Hillerød	55.9267	12.3109	51261	214.9	Capital
16	3450	Allerød	55.8667	12.3833	25871	67.5	Capital
17	3480	Fredensborg	55.9667	12.4000	40980	112.1	Capital
18	3600	Frederikssund	55.8396	12.0690	45489	247.0	Capital

4 Finding the best regions

4.1 Clustering and data segmentation

In order to find the best regions for the business community, data segmentation and clustering was implemented for the illustrated data. In this section, the K-means technique which is a partition-based clustering method was utilized for the sake of the aimed data analysis. The clustering was made using postal code and coordination as the initial input data. And the clustering parameter (k) was tried in the range of 1 to 10. The results shoed that the K value of 6 has the best segmentation result, therefore, the analysis has been made based on the cited K value.

The result of the clustering which segmented the region into 6 different clusters have been illustrated in the following.

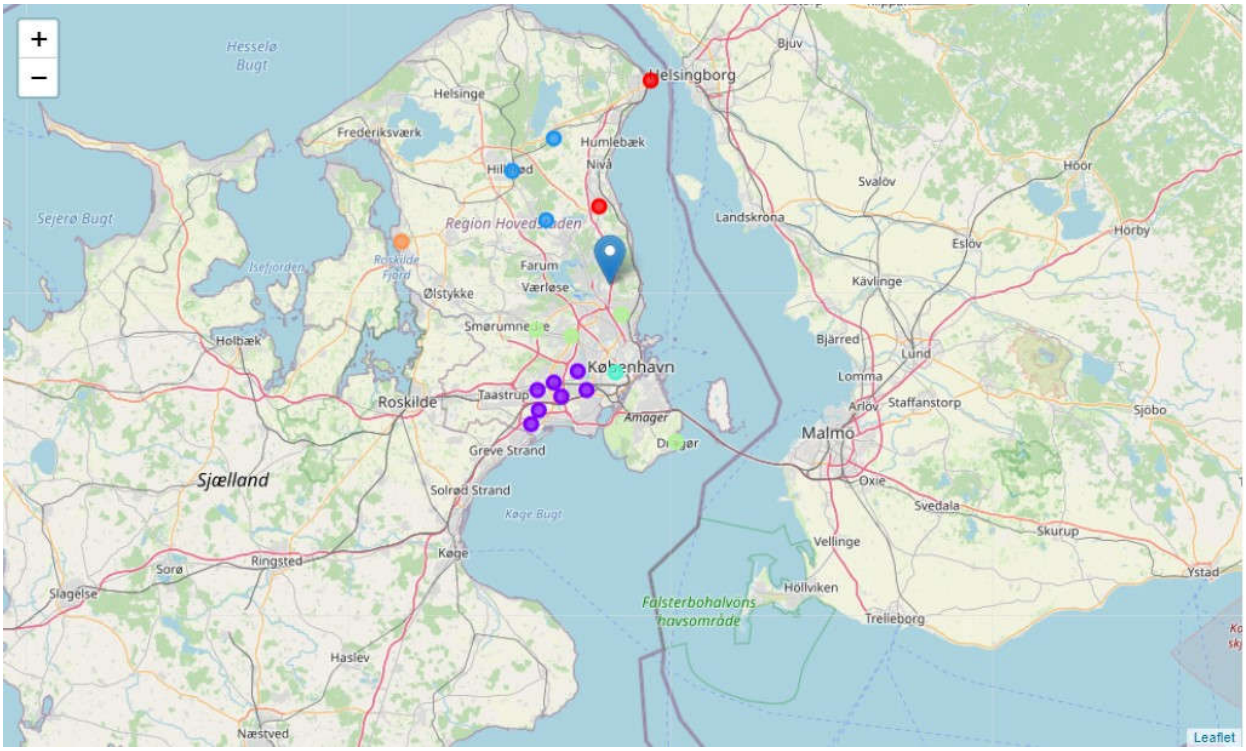

Figure 3: The clustering made using K_means technique

After doing the clustering and finding the best fit for the data, it can be noticed that the best regions for the new business that benefits the DTU community are regions 0 , 2 and 4, which are depicted with light blue, dark blue and red color respectively. As can be noticed, some areas seem to be quite far from DTU; however, this is mainly due to the fact that the postal codes used in this region do not include the sub areas (different neighborhoods within each region). Therefore, the processing has been made using the current data set. For a detailed analysis, the premium post data account in Denmark would be required.

4.2 Rebuilding the data set

Now that the desired locations for initiating the business that have the best possible business community, have been recognized, our data base can be filtered based on the selected cluster numbers.

Table 11: Filtering the data set as per selected clusters

	Cluster Labels	postal_code	Borough	Latitude	Longitude	population	Area(km²)	Region
9	4	2730	Herlev	55.7237	12.4400	28914	12.1	Capital
10	4	2750	Ballerup	55.7316	12.3633	48978	33.8	Capital
11	4	2791	Dragør	55.5945	12.6664	14497	18.3	Capital
12	4	2820	Gentofte	55.7500	12.5500	74623	25.6	Capital
13	0	2970	Hørsholm	55.8833	12.5000	24919	31.3	Capital
14	0	3000	Helsingør	56.0361	12.6136	63047	118.9	Capital
15	2	3400	Hillerød	55.9267	12.3109	51261	214.9	Capital
16	2	3450	Allerød	55.8667	12.3833	25871	67.5	Capital
17	2	3480	Fredensborg	55.9667	12.4000	40980	112.1	Capital

4.3 Business benefits

In order to make the comparison of how much profit would a startup or setting a vending machine can earn by having a better price of bike service / tools:

- The average value of 300 dkk for a normal service) has been used [4]
- The offered price by the startup has been chosen to be 10 dkk lower than the average price (which is not a big difference)
- As a rule of thumb, it is assumed that 60% of the population in these regions can use bike therefore require at least yearly bike services (assumption is based on the total percentage of people using bike in Denmark (60%) [1]

Table 12: Adding the proposed price and the comparison with current fees

	Cluster Labels	postal_code	Borough	Latitude	Longitude	population	Area(km²)	Region	Normal_price_dkk	Discounted_price_dkk
9	4	2730	Herlev	55.7237	12.4400	28914	12.1	Capital	5204520.0	5031036.0
10	4	2750	Ballerup	55.7316	12.3633	48978	33.8	Capital	8816040.0	8522172.0
11	4	2791	Dragør	55.5945	12.6664	14497	18.3	Capital	2609460.0	2522478.0
12	4	2820	Gentofte	55.7500	12.5500	74623	25.6	Capital	13432140.0	12984402.0
13	0	2970	Hørsholm	55.8833	12.5000	24919	31.3	Capital	4485420.0	4335906.0
14	0	3000	Helsingør	56.0361	12.6136	63047	118.9	Capital	11348460.0	10970178.0
15	2	3400	Hillerød	55.9267	12.3109	51261	214.9	Capital	9226980.0	8919414.0
16	2	3450	Allerød	55.8667	12.3833	25871	67.5	Capital	4656780.0	4501554.0
17	2	3480	Fredensborg	55.9667	12.4000	40980	112.1	Capital	7376400.0	7130520.0

In the next step, the top 5 locations that had the highest revenue were chosen.

Table 13: top five business locations as per revenue

	Cluster Labels	postal_code	Borough	Latitude	Longitude	population	Area(km²)	Region	Normal_price_dkk	Discounted_price_dkk
12	4	2820	Gentofte	55.7500	12.5500	74623	25.6	Capital	13432140.0	12984402.0
14	0	3000	Helsingør	56.0361	12.6136	63047	118.9	Capital	11348460.0	10970178.0
15	2	3400	Hillerød	55.9267	12.3109	51261	214.9	Capital	9226980.0	8919414.0
10	4	2750	Ballerup	55.7316	12.3633	48978	33.8	Capital	8816040.0	8522172.0
17	2	3480	Fredensborg	55.9667	12.4000	40980	112.1	Capital	7376400.0	7130520.0

As can be seen by just having a 10 dkk discount (which is less than 3% discount), a great profit is resulted. Please note :

- As per current assumption, each person purchases services only one time per year, which is a very tough decision
- As a person who studies at DTU, 4 times I just had flat tire during a year, which shows the significance of the possible profit

In order to have a better insight of the top 5 locations depicted in table 13, the areas are illustrated in the following map.

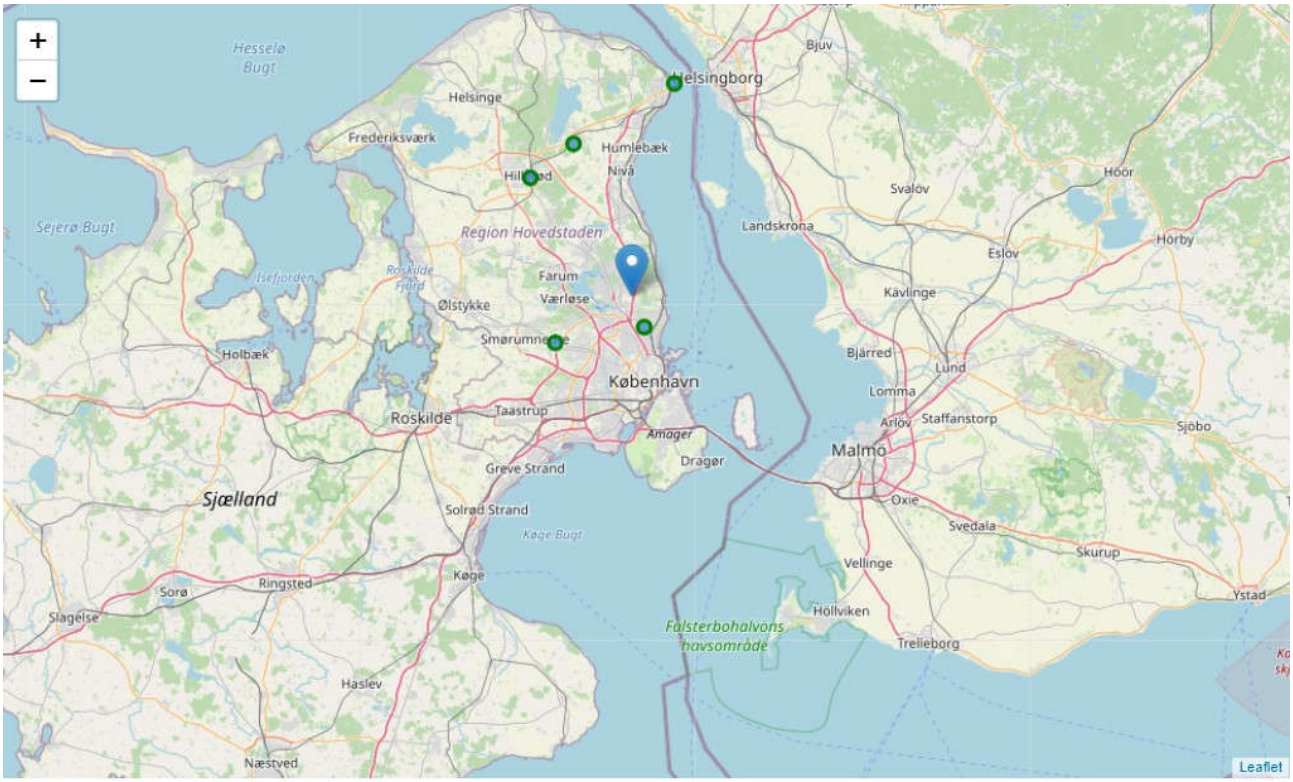

Figure 4: The top 5 business locations

4.3.1 A wider range of business offers

Let's try a wider range of offers to have a better idea of the great profit that this business can offer. For such purpose, 4 other discount ranges have been utilized:

- 5% discount for the normal price
- 10% discount for the normal price
- 15% discount for the normal price
- 20% discount for the normal price

By implementing different business offers, a better comparison of the possible business outcome can be implemented.

Please note that as per our initial assumption, the Normal price and discounted prices are assumed for the case that 60% of people use these services. which means for the current calculation, the quantity of people using both services are the same. This assumption is not realistic, therefore as a more realistic assumption, if people who were previously using the normal places find the discount, they would probably attract to the new system and won't use the previous one. so, the following assumptions now must take place:

- let's assume 10% of people using normal price now will start using the new service
- so, the community of users of normal shops will reduce by 10% and on the other hand, the community of people using the new service increases by 10%

Table 14: adjusted business user community for each offer

Normal_dkk_90%_user	discount_dkk_3%_10%userextra	discount_dkk_5%_10%userextra	discount_dkk_10%_10%userextra	discount_dkk_15%_10%userextra
12088926.0	14282842.2	14036586.3	13297818.6	12559050.9
10213614.0	12067195.8	11859140.7	11234975.4	10610810.1
8304282.0	9811355.4	9642194.1	9134710.2	8627226.3
7934436.0	9374389.2	9212761.8	8727879.6	8242997.4
6638760.0	7843572.0	7708338.0	7302636.0	6896934.0

4.4 Visualization

As a rule of thumb, the best route for understanding the business model, is to see how the data looks like. For such purpose, the visualization of the illustrated business offers have been made, which can be seen in the following.

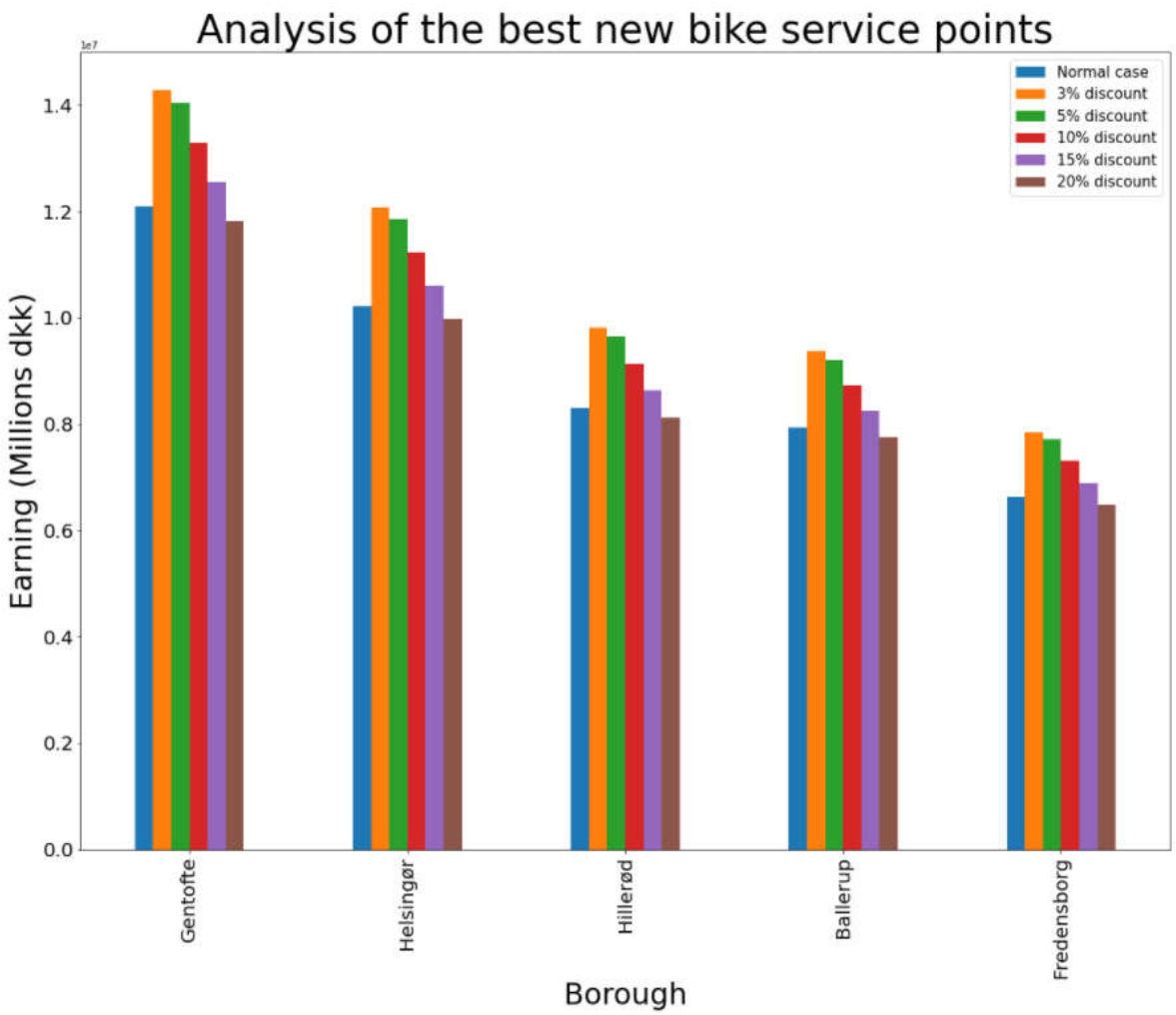

Figure 5: The comparison of the busines outcome for the different business cases in the top five locations of the possible business community

As can be seen, for the better illustration of the data, bar chart has been used. This provides us the great opportunity to know the outcome of different business offers and its comparison with the other cases. The result is as per the initial assumptions that was made and for a better review, have been provided in the following.

- 10% of people using normal price now will start using the new service
- so, the community of users of normal shops will reduce by 10% and on the other hand, the community of people using the new service increases by 10%

5 Conclusion

As can be seen from the previous section, having a very minor discount a great business income is resulted. It shows that having merely 3% discount can make a very big change in a business area with rather good quantity of users. As noted earlier, the calculations are made based on the very tough decision that only 10% of the people previously using the normal bike repair shops are now using the new services. Moreover, the simplicity of using a vending machine can definitely have a higher impact on the community of users and therefore a much higher number of users should be expected.

Having only the current result into account, one can state that even by offering 20% discount compared with the normal bike service shops, a very competitive business outcome will be resulted.

Moreover, it must be stated that all the other conditions are assumed similar for all cases; however, it is obvious that for the time periods that the normal bike repairing stores are closed, the only business provider would be the new innovative route that stated in this research. Therefore the impact and benefit of the new business should be considered much more than the illustrated output in this report.

6 Future studies

For the future studies, I highly recommend using a more detailed dataset that includes the subregion postal codes as well. As stated earlier in this report, the current data set used in this report does not have those details.

The impact of other parameters such as the staff fee and time periods could be also added to the business model. Because using a vending machine has the great benefit of much lower staff price (its un manned) and also there is not time limit for it (it can operate 24h)

References

1. Cycling in Copenhagen
2. The main webpage of the technical university of Denmark
3. QS world ranking of the universities
4. fribike shop website