Sociodemographic and Socioeconomic Analysis of Istanbul Residents from Survey Data

Harun Harman Kazım Halil Kesmük*

[1]Department of Computer Science, Hacettepe University

Abstract

This work proposes a comprehensive analysis of Istanbul' s socioeconomic and sociodemographic status from the surveys collected by Istanbul Metropolitan Municipality (IBB) and also aims the profiling of the Istanbul' s residents according to this survey. The dataset includes 50.390 person' s survey responses about different aspects of life circumstances such as the income, how long they have been living in Istanbul, their sport activities etc.

Introduction

Environmental and urban planning is a very important application to prevent possible problems in the future as well as affecting the present. These plans are usually made by municipalities, governorships or other institutions dealing with this field. The neighborhood/region may have specific needs, problems or requests, and these facts may not always be expressed by the public. Institutions' ability to detect these facts, divide the regions into profiles and take action according to these profiles accelerates the work process for environmental and urban planning.

The survey used in this project was collected by IBB between November 29, 2021, and March 7, 2022. The survey included face-to-face interviews with 50,000 households across Istanbul, ensuring a minimum of 12 interviews per neighborhood, proportionally distributed based on the number of households in each district. The selection of households was executed using a stratified random sampling method at the neighborhood level to enhance the representativeness of the findings.

The survey had four main sections focusing on actual household data: In the first phase, data on household demographics, age, sex, education, employment status and other chronic diseases and special conditions (156,689 unregistered) were collected The second category focused on socioeconomic status, including housing type, rent, income and car ownership data (50,390 unregistered). Phase III used the KISH method to assess migration history, language skills, neighborhood/city satisfaction, and urban challenges (50,390 records). The fourth phase examined lifestyle behaviors and attitudes, participation in activities, time management, digital literacy and safety attitudes (50,390 participants).

*Corresponding author: harunharman@hacettepe.edu.tr
halilkesmuk@hacettepe.edu.tr
Published: December 25, 2024

In the first part of the work, the data was visualized to get information about the distribution of features in the dataset and their relationships with each other. These visualizations were interpreted and information was obtained for the clustering task. In addition, the information obtained from visualization can be a guide for the work to be done by the authority.

For the second part of the work, after the detailed analysis, it is decided to proceed with making profiling for these neighbourhoods. To achieve this goal, clustering algorithms are preferred. The dataset is processed and different clustering algorithms are applied for two version of the data: Grouped by neighbourhood and indivually. The results show that grouping the data is more efficient and successful. To calculate the performance of the algorithms, "Silhouette Score" is preffered. As a result, 0.75 Silhouette score is achieved.

Methodologies

Data Analysis

In this part, Pandas, Matplotlib and Seaborn libraries were used for data analysis and visualization processes. The process consists of the following steps:

1. **Data Reading and Processing:** Data sets were read with the *read_data* function and made available for analysis.
2. **Analysis of Distributions:** Numerical variables were analyzed using histogram graphs with *plot_histogram* function. Frequencies of categorical variables were visualized with *plot_bar_chart* and *plot_countplot* functions.
3. **Investigation of Correlations:** The relationships between two categorical variables were visualized with *plot_stacked_bar_chart* and *plot_categorical_heatmap* functions.

4. **Saving Charts:** With the *save_path* option in functions, charts are exported and used for reporting.

Profiling

Profiling is the process of examining, analyzing, and summarizing datasets to understand their structure, quality, and content. The aim of this section is to find patterns that distinguish neighborhoods in terms of socioeconomic and sociodemographic characteristics and to profile neighborhoods and people according to these patterns.

Data Preprocessing

The initial phase involved importing data from several Excel files using the pandas library. Each dataset represented distinct yet interconnected facets of the problem domain. At this stage, the focus was on ensuring data integrity and consistency across all sources. Once the data was ingested, preprocessing played a crucial role in preparing the datasets for analysis. Categorical variables were transformed using encoding techniques such as One Hot Encoding, Label Encoding and Frequency Encoding to ensure they were machine-learning compatible.

After the encoding, missing values are handled. There were many columns with over 25.000 rows of missing values, so they are removed. Remaining 26 columns with missing values are handled with the help of Machine Learning. A function is written for imputing the missing values using Decision Tree Classifier and Decision Tree Regressor.

Features underwent scaling with MinMaxScaler, standardizing their range and preventing dominance of larger magnitude values during model training. After that PCA is applied to the data. For each clustering algorithm, different number of components are selected for PCA with min value of 1 and max value of 4.

Clustering

This section includes two approaches for clustering: clustering by person and clustering by neighbourhood. For both approaches, KMeans, DBSCAN, GaussianMixture, Spectral Clustering and Agglomerative Clustering for Neighbourhood Clustering are applied.

1. **KMeans:** It is an iterative centroid-based clustering algorithm that uses Euclidean distance to divide data into k clusters.
2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** It is a density-based clustering algorithm. It identifies dense regions as clusters and excludes low-density points (noise).
3. **Gaussian Mixture:** It is a statistical approach that models data points with Gaussian distributions. Clustering is performed with the Expectation-Maximization (EM) algorithm.
4. **Spectral Clustering:** It represents data points as nodes of a graph and creates clusters using the spectral properties of this graph (e.g. Laplacian matrix).
5. **Agglomerative Clustering:** It is a hierarchical clustering algorithm. It first initializes data points as individual clusters and then merges the most similar clusters to form a hierarchical structure.

To calculate the performance of the algorithms, "Silhouette Score" is preferred. Silhouette Score is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:
a(i) = Average distance of sample i to all other points in the same cluster.
b(i) = Minimum average distance of sample i to points in another cluster.

Results

Data Analysis

This section contains visualizations and interpretations of the distribution of features in the dataset or how they relate to each other.

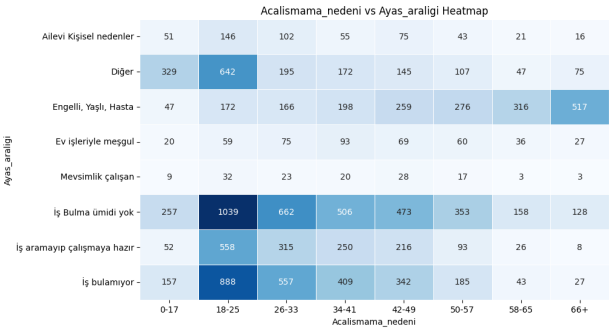

Figure 1: *Age-related reasons for not working*

Looking at this graph, it is seen that the majority of young and middle-aged people cannot work because they have no hope of finding a job or cannot find a job. At older ages, the reason for not being able to find a job is usually physical disabilities. Considering these, it can be suggested that the authorities should

open more job opportunities for young people and find less tiring jobs for older people.

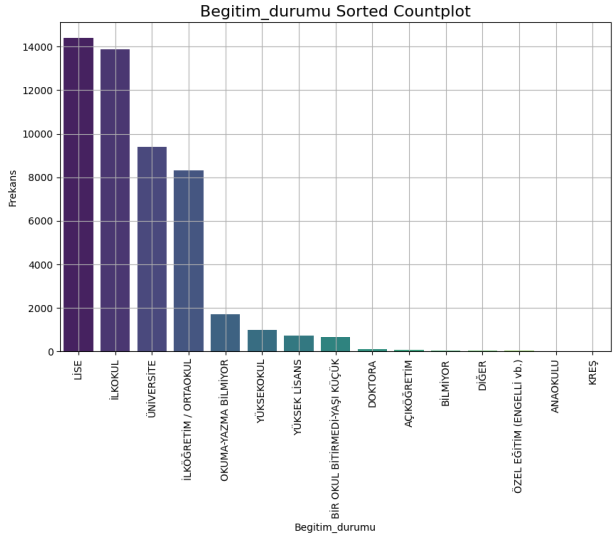

Figure 2: *Education status*

Looking at this graph, it is clear that the vast majority are high school and primary school graduates. The number of university graduates is below these. In today's conditions, the number of illiterate people is not small. Considering these situations, it may be suggested that authorities should take actions to encourage people to have a higher level of education.

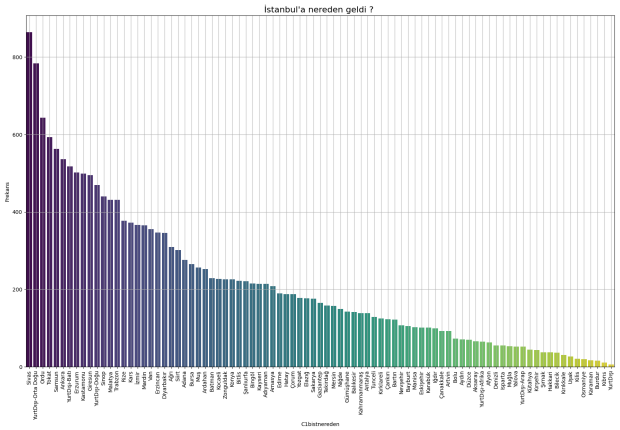

Figure 3: *Where people came to Istanbul from*

Looking at this graph, it can be seen that Istanbul has received migration from various parts of Anatolia. However, it is noteworthy that the provinces in the Black Sea region ranks high. In addition, the number of immigrants coming from abroad is not small. If we look at the ranking of migration from abroad, those from the Middle East are at the top. However, those from the West and East of the country are also among the regions with the highest number of migrants. Given these facts, it may be advisable for authorities to be ready for various measures to preserve the socio-demographic structure of the people of the region, if necessary.

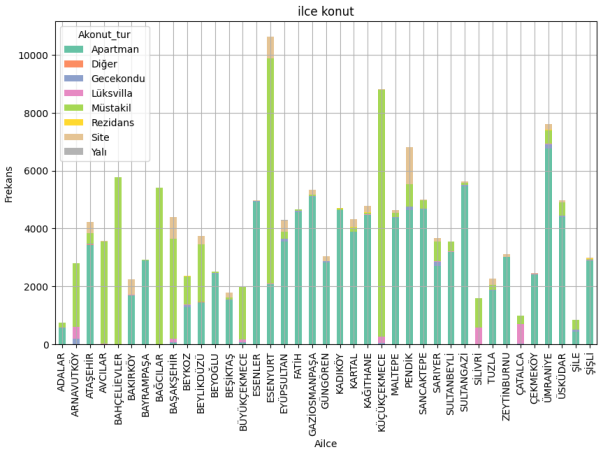

Figure 4: *Housing types by district*

Looking at this graph, Apartment and Detached housing types are quite high in the majority of districts. However, in some districts such as Arnavutköy, Silivri and Çatalca, luxury villa housing type also draws attention. On the other hand, the number of people living in housing complexes is not small. Authorities may be advised to consider these differences between housing types and plan according to regional needs.

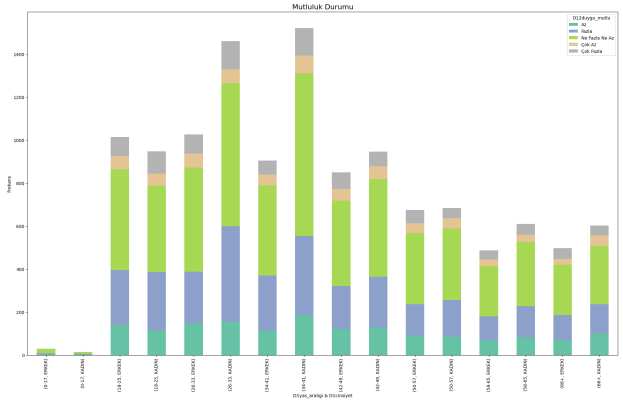

Figure 5: *Happiness status by age range and gender*

When we look at this graph, it is not possible to make a very sharp inference about the state of happiness in terms of age range and gender. In all age ranges and genders, the answer "neither less nor more happy" seems to be the majority. However, especially in the young-middle ages, the number of "less happy" responses is close to the number of "neither more nor less happy" responses. Even though the number of "less happy" responses ranks second in the older age group, it is clearly lower than the young-middle age group. Again, the number of "very happy" responses does not exceed the number

of "less happy" responses in any age group. All this shows that the vast majority of people are moderately or less happy.

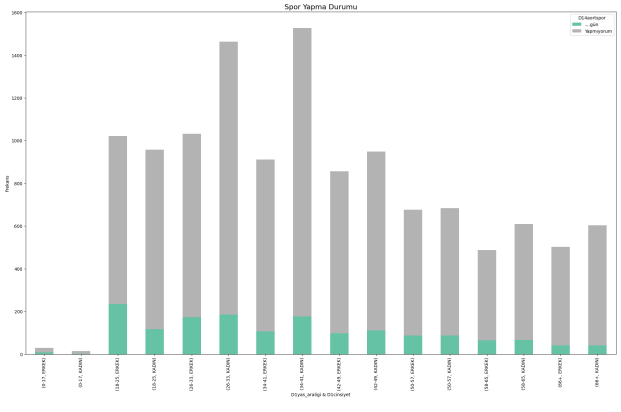

Figure 6: *Doing sports by age range and gender*

Looking at this graph, it is clear that most of the people who participated in the study do not do sports. In proportional terms, although it seems that middle-aged people do sports at a lower rate, it is not possible to distinguish young-middle-aged people from middle-aged people very sharply. However, it seems that men between the ages of 18-25 have the highest rate of doing sports. Although the rate for women in the same age range is not very low, the highest rates are generally among young-middle-aged men. Physical and temporal availability for sports seems to be effective in these rates. In order to increase these rates, authorities may be advised to create environments where people can do sports in a more comfortable and motivated way.

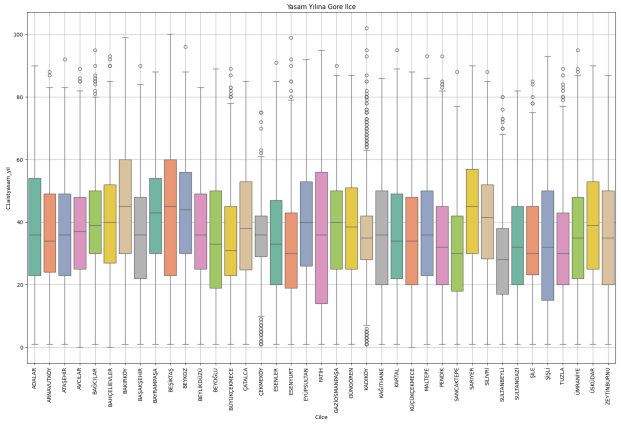

Figure 7: *Districts by life span*

This graph shows the differences in life span between districts. While life span is generally higher in districts such as Adalar, Beşiktaş, Bakırköy and Fatih, it is lower in districts such as Çekmeköy, Sancaktepe and Sultanbeyli. While living standards and socioeconomic conditions are one option to explain this, the extent to which districts are well-established or how recently they are inhabited can also be a factor. Authorities may be advised to make improvements in certain districts if they have significantly lower living standards.

Profiling

Table 1: *Silhouette Score of clustering algorithms without grouping by neighbourhood.*

Algorithm	Kmeans	Gaussian Mixture	Spectral Clustering	DBSCAN
Silhouette Score	0.51	0.47	0.1	0.33

Table 2: *Silhouette Score of clustering algorithms with grouping by neighbourhood.*

Algorithm	Kmeans	Gaussian Mixture	Spectral Clustering	DBSCAN	Agglomerative Clustering
Silhouette Score	0.75	0.92	0.33	0.61	0.74

As can be seen from the Table 1 and Table 2, results are better with grouping the data according to the neighbourhood. Especially Gaussian Mixture algorithm works pretty well. So, it is decided to proceed with this algorithm' s results for profiling. According to the model, there are 8 profiles. They will be discussed in the discussing section.

Discussion

Profiling Discussion

Profile 0

This neighborhood cluster is generally composed of individuals with low-medium socioeconomic status. In neighborhoods where income levels are limited, a family-oriented lifestyle is prominent, half of the population is female and the majority are married.

Although secondary and higher education graduates are predominant in terms of education level, the rate of primary school graduates is also noteworthy. This suggests that professional opportunities may be limited in some neighborhoods.

While the majority of housing types are in apartment buildings, it is seen that the buildings are old and the rental costs are low. This shows that the physical structure of the neighborhoods is generally in line with economic limits.

Infrastructure services (electricity, water, sewage) are generally accessible; however, green areas and environmental facilities are insufficient in some neighborhoods. In terms of social assistance needs, although a certain segment needs education and food assistance, the majority state that they do not need these assistances.

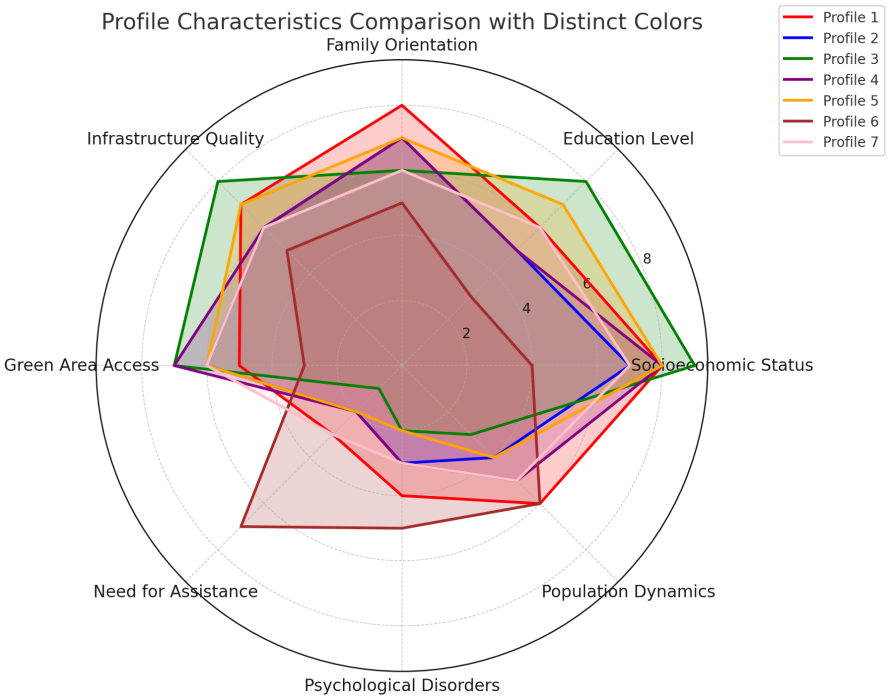

Figure 8: *A radar graph for demonstrating the profile characteristics comparison of profiles.*

The demographic structure indicates a dynamic population that includes young and middle-aged groups. The low rates of disability and chronic diseases reflect a positive picture in terms of health. The general profile of the neighborhoods represents a structure with basic infrastructure and open to economic and social development.

Şirinevler, Cevizli, Bağlarbaşı, Bahçelievler are some of the examples of this profile.

Profile 1

This neighborhood cluster has a high socioeconomic status, indicating that the economic situation is generally good and homogeneous. The gender ratio is balanced (0.49), and the fact that married individuals are predominant in terms of marital status (0.68) indicates that family structures are at the forefront. The low student ratio (0.06) suggests that the young population is small in these areas.

The property rate is around 40%, but the short duration of residence (0.13 years) may indicate that the neighborhoods have a dynamic population or are used as temporary settlements. The fact that the houses are generally 2+1 and 3+1 indicates that small and medium-sized families are common.

While basic infrastructure services (electricity, sewage, lighting) are welcomed, improvements may be needed in social areas such as green areas (0.53) and public transportation (0.57). Chronic disease (0.13) and disability (0.01) rates are low, but psychological disorders are prominent in some areas.

The level of education is generally at a good level, with the majority being primary school graduates (0.54), while the rates of high school (0.25) and higher education (0.13) are lower. In general, the neighborhoods have good economic opportunities, are family-oriented, but need development in social and environmental areas.

Güvercintepe, Hadımköy, Ferhatpaşa, Yavuz Selim are some of the examples of this profile.

Profile 2

This neighborhood has a fairly homogeneous socioeconomic structure, with all households having similar incomes and living standards. The gender distribution is balanced, the majority of married individuals are and the proportion of the young population is low. Apartments are common in housing, while detached houses come second, and the presence of luxury villas, albeit few, is striking.

The neighborhood infrastructure is generally satisfactory; services such as electricity, water and lighting are of high standards, but there is room for development in green areas and internet services. While the need for social assistance is low, requests for support for rent and household expenses are prominent. Primary and secondary education are predominant in terms of education, and the adult population is concentrated between the ages of 26-41.

The residents of the neighborhood generally live in a safe environment; the rate of dangerous incidents is low and recycling awareness is at a moderate level. This shows that a peaceful and orderly life is

maintained in the neighborhood.

Çamlık, Barbaros Hayrettin Paşa, Ambarlı, Ayazağa are some of the examples of this profile.

Profile 3

These neighborhoods are generally located in areas with high socioeconomic status and have high levels of economic prosperity. They mostly have modern infrastructure and quality living standards, the building age is young and environmental satisfaction is high.

Education levels are generally high, the higher education rate is remarkable. In these neighborhoods where the employment rate is high, individuals are generally economically self-sufficient.

Infrastructure services (electricity, water, internet) are complete and satisfaction with environmental factors such as green areas and clean air is prominent. The need for social assistance is very low and economic diversity is low.

Arnavutköy, Bebek, Levent, Etiler are some of the examples of this profile.

Profile 4

This neighborhood cluster has a high socioeconomic status (average 7.01) and a homogeneous economic structure. The population distribution is relatively balanced, the majority of individuals are married (68%) and the student rate is low (5%), indicating that family life is dominant. Housing structures are generally new (12.46%) and 2+1 to 3+1 types, and property ownership is high. Basic infrastructure and green space access are quite good; however, there is potential for improvement in public transportation and internet infrastructure. The rate of individuals in need of social assistance is low (13% need for food assistance), and the general welfare level is high. The level of education is mostly at primary and secondary education levels, while the rate of higher education is limited (10

Başak, Mimar Sinan, Atatürk , Altınşehir are some of the examples of this profile.

Profile 5

In this neighborhood cluster, the socioeconomic status is high and the economic power is generally balanced. The gender ratio in the population structure is balanced, the majority of individuals are married and the student rate is quite low. Although the housing is predominantly apartment type, detached houses also have an important place; the home ownership rate is high and the tenancy rate is low. Although the infrastructure services are generally at a good level, some deficiencies are observed in modern services such as internet and transportation. Although the need for assistance is generally low, rent, food and education support are requested in some neighborhoods. Satisfaction with environmental cleanliness and green areas is high, and the rates of chronic diseases and psychological disorders are low. While young and older age groups are less represented, mostly working-age individuals live there. In general, the neighborhoods are economically strong and have sufficient infrastructure, but they exhibit a structure open to development in terms of modernization and sustainability.

İnönü, Kamer Hatun, Nişantepe, İskenderpaşa are some of the examples of this profile.

Profile 6

This neighborhood cluster has a profile of low socioeconomic status, low income, and limited access to economic resources. Although gender and marital status are balanced, the low student ratio indicates a small young population. Residence periods are short, rents are low, apartment living is dominant, and luxury housing types are almost non-existent. The rate of higher education is high compared to other categories, but the rate of illiteracy is also noteworthy.

Although infrastructure services are generally accessible, social infrastructure such as green areas and sidewalks are inadequate in some places. While requests for social assistance are concentrated especially on food, rent, and bills, it is noteworthy that a large portion of the population states that they do not need assistance. Although chronic diseases and psychological disorders are at low levels in health indicators, such problems may occur in certain areas. In general, a neighborhood structure is observed where difficulties are experienced in accessing economic and social services.

Barbaros, Kılıçali Paşa, Gayrettepe are some of the examples of this profile.

Profile 7

This neighborhood cluster has an average socioeconomic status and the rate of households in need of assistance is quite low. The gender distribution is balanced, the rate of married individuals is slightly higher, and the student population is quite limited. Most of the housing is in apartment style, new construction is widespread, and population mobility is high. Environmental satisfaction is generally good, with positive evaluations especially regarding cleanliness, lighting, and electrical infrastructure. The rate of individuals with higher education and the rate of employment are remarkably high, indicating an economically active population. Health problems and social assistance needs are seen at low levels, and environmental awareness is supported by recycling opportunities. In general, it presents a dynamic, newly developing, and middle-upper income group-hosting regional profile.

Müeyyetzade, Zeytinlik, Nizam, Bozkurt are some of the examples of this profile.

Future Work

This data analysis and profiling proposes a comprehensive analysis of Istanbul' s socioeconomic and sociodemographic status, but with limited data. Although the survey system could be systematic, the answer might not fully represents the all residents of that area.

For the future work, this analysis might be used by the IBB experts in relevant fields to improve services. Also with better and bigger data, the already established system might work with better results for the future.

References

[1] Data-Driven Management Model (DBMM) Research Part A Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-a-bolumu-verileri.

[2] Data-Driven Management Model (DBMM) Research Part B Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-b-bolumu-verileri.

[3] Data-Driven Management Model (DBMM) Research Part C Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-c-bolumu-verileri.

[4] Data-Driven Management Model (DBMM) Research Part D1 Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-d1-bolumu-verileri.

[5] Data-Driven Management Model (DBMM) Research Part D2 Data,IBB https://data.istanbul/dataset/veriye-dayali-yonetim-modeli-arastirmasi-d2-bolumu-verileri.

[6] Data-Driven Management Model (DBMM) Research Part D3 Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-d3-bolumu-verileri.

[7] Data-Driven Management Model (DBMM) Research Part D4 Data,IBB https://data.ibb.gov.tr/dataset/veriye-dayali-yonetim-modeli-arastirmasi-d4-bolumu-verileri.

[8] Data-Driven Management Model Research,IBB https://data.ibb.gov.tr/dataset/87fb0a12-5566-45bb-956f-60ac802fa546/resource/884a1526-fe0f-4850-8c61-07b7d8447102/download/ddmm-project-description-file_en.pdf.