Fater

By Team Pariza

Table of Contents

1. Introduction

1.1 Preface

Understanding the potential revenue is essential for investors looking to maximize their returns in the retail sector, as it offers invaluable insights into the profitability and growth prospects of their investments. Accurate revenue estimates provide investors with a roadmap for assessing the viability and potential yield of their ventures, enabling them to make informed decisions that align with their financial objectives. By gaining clarity on revenue projections, investors can evaluate the risk-reward profile of different investment opportunities, identifying those with the highest likelihood of delivering substantial returns. Moreover, precise revenue forecasts empower investors to allocate their capital strategically, directing funds towards sectors or businesses with the greatest revenue-generating potential while minimizing exposure to underperforming assets. Additionally, by understanding the revenue potential of a retail venture, investors can gauge its attractiveness relative to alternative investment options, enabling them to optimize their portfolio composition and enhance overall returns. In essence, a thorough understanding of potential revenue equips investors with the knowledge and foresight needed to make sound investment decisions, capitalize on lucrative opportunities, and achieve their financial goals.

1.2 Real Problem Presentation

In today's competitive retail landscape, understanding the dynamics of consumer behavior and market demand is crucial for businesses aiming to maximize revenue and profitability. This is particularly true in the case of products catering to specific demographic needs, such as diapers, where market conditions are influenced by a myriad of socio-demographic, territorial, and commercial factors. In this essay, we undertake the task of estimating potential revenue for diaper sales in the Naples province of Italy, taking into account the diverse socio-demographic profile, territorial features, and points of interest relevant to the business.

2. Data Understanding

2.1 The Dataset

Below are the datasets used in the analysis, each serving a specific purpose:

- Socio Demo NA, which contains demographic information essential for understanding the population distribution across different administrative divisions.
- Shapes NA, which comprises spatial data representing the geographical shapes or boundaries of administrative areas.
- Gravitations NA, which is focused on the gravitational pull of different locations within the study area.
- Stores NA, which provides information about various stores within the study area.

Below, we provide an overview of each dataset, detailing its content:

Variable	**Description**	**Type**	**Levels / Unit**
Socio Demo NA			
microcode	ISTAT Microcell	Numerical	630010000001
district	Municipality	Categorical	ACERRA/NAPOLI/…
province	Microcell Province	Categorical	Napoli/…
region	Microcell Region	Categorical	CAMPANIA
population	Total Population	Numerical	54/15312/9/…
population_m	Male Population	Numerical	54/15312/9/…
population_f	Female Population	Numerical	54/15312/9/…
population_age_00_04_yr	Population Aged 0-4 Years	Numerical	54/15312/9/…
population_age_05_14_yr	Population Aged 5-14 Years	Numerical	54/15312/9/…
population_age_15_34_yr	Population Aged 15-34 Years	Numerical	54/15312/9/…
population_age_35_44_yr	Population Aged 35-44 Years	Numerical	54/15312/9/…
population_age_45_54_yr	Population Aged 45-54 Years	Numerical	54/15312/9/…
population_age_55_64_yr	Population Aged 55-64 Years	Numerical	54/15312/9/…
population_age_65_up_yr	Population Over 65 Years	Numerical	54/15312/9/…
Gravitation NA			
microcode	ISTAT Microcell	Numerical	630010000001
daytype	Weekday or Weekend	Categorical	1/2
fasciaoraria	Time Slot	Categorical	2/3/4/5/6
datatype	Population & Gender	Categorical	F1/F2/…/F6/Gf/Gm
media_annuale	Average Gravitated Amount	Numerical	53/14315/7/…
Shapes NA			
microcode	ISTAT Microcell	Numerical	630010000001

geometry	Microcell Polygon Geometry	Textual	POLYGON(...)
Stores			
Cod3HD	Store ID	Numerical	23/543/…
TipologiaPdv	Store Type	Categorical	LIS/SUP/…
MQVEND	Store Size	Numerical	250/300/…
Parking	Store Parking	Categorical	TRUE/FALSE
Indirizzo	Store Address	Textual	VIA CUCCARO 1/...
Lat	Store Latitude	Numerical	40.8787224
Long	Store Longitude	Numerical	14.1502586
Comune	Store Municipality	Categorical	POZZUOLI/…
Provincia	Store Province	Categorical	NA/…
Potenziale	Italian Diaper Market Share	Numerical	0.016/…

2.2 Pre Processing

2.2.1 Introduction

Data preprocessing stands as the pivotal initial stage in our data analysis process. Here, we embark on a journey through the key stages of data processing, where we delve into the significance and methodologies of data cleaning, feature creation, normalization, and merging. These critical steps serve as the cornerstone of data analysis, laying the groundwork for robust and insightful outcomes.

2.2.2 Data Cleaning

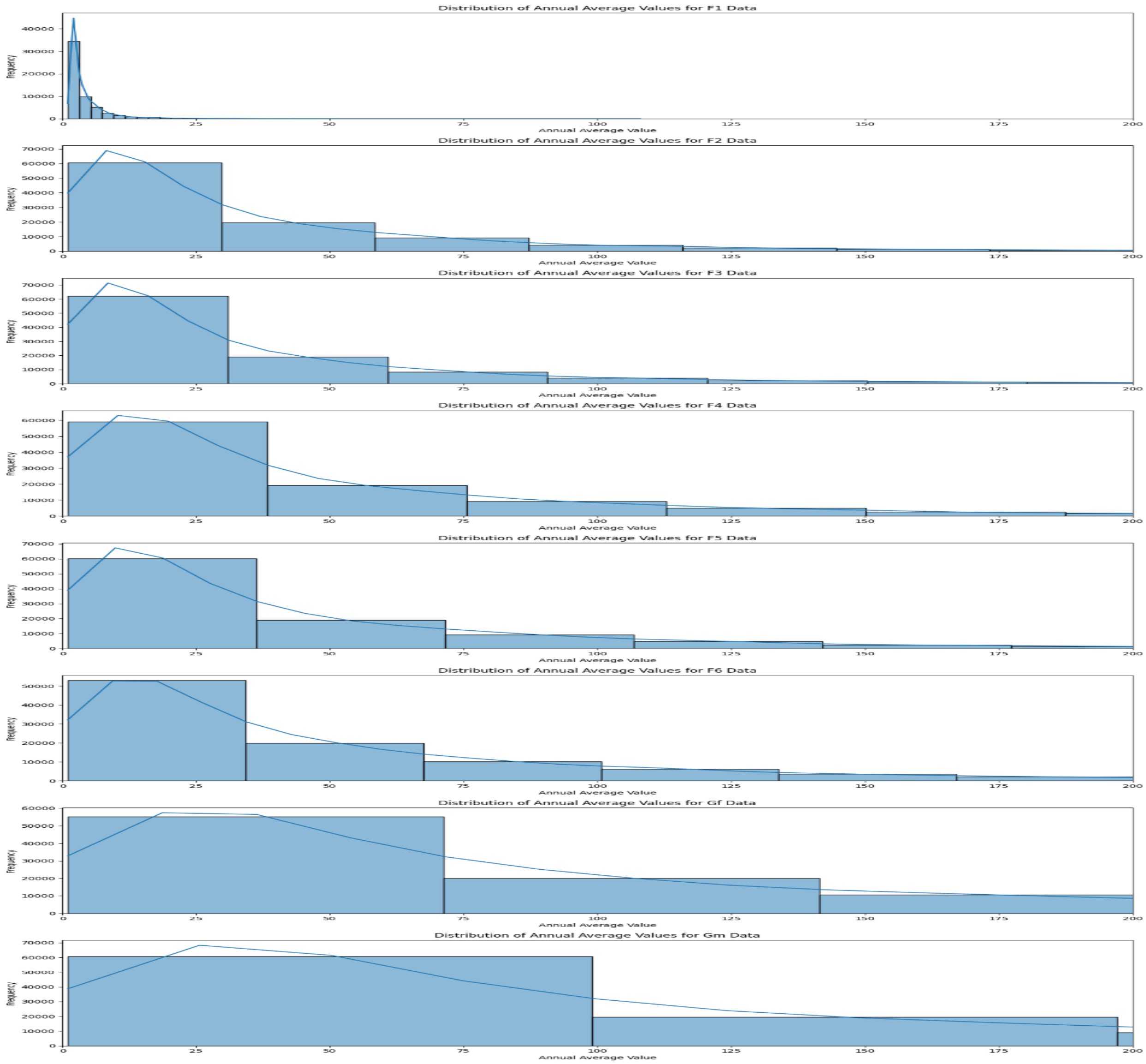
2.2.2.1 Handling Missing Values

In this section, our focus is on identifying any potential missing values within the datasets under examination. Upon thorough analysis, it becomes apparent that the dataset "Gravitation NA" is the sole dataset containing missing values.

```
       Socio Demo NA
microcode                 0
district                  0
```

```
          province                   0
          region                     0
          population                 0
          population_m               0
          population_f               0
          population_age_00_04_yr    0
          population_age_05_14_yr    0
          population_age_15_34_yr    0
          population_age_35_44_yr    0
          population_age_45_54_yr    0
          population_age_55_64_yr    0
          population_age_65_up_yr    0
                  dtype: int64

#################################################
                    Stores NA
              Cod3HD            0
              Insegna           0
              TipologiaPdV      0
              MQVEND            0
              Parking           0
              Indirizzo         0
              Lat               0
              Long              0
              Comune            0
              Provincia         0
              Potenziale        0
                  dtype: int64
#################################################
                 Gravitation NA
            Unnamed: 0           0
            microcode            0
            daytype              0
            fasciaoraria         0
            datatype             0
            media_annuale    48707
                 dtype: int64
#################################################
                    Shapes NA
            Unnamed: 0           0
            microcode            0
            geometry             0
```

Upon examining the distribution of the "media_annuale" variable, a discernible right skewness is observed. This skewness informs our decision to address missing values by imputing the median value, a strategy chosen to mitigate the influence of outliers and maintain the integrity of the dataset.



Distribution of Annual Average Values for F1 Data

Distribution of Annual Average Values for F2 Data

Distribution of Annual Average Values for F3 Data

Distribution of Annual Average Values for F4 Data

Distribution of Annual Average Values for F5 Data

Distribution of Annual Average Values for F6 Data

Distribution of Annual Average Values for Gf Data

Distribution of Annual Average Values for Gm Data

Fig.

In addition to handling missing values, we employ behavior detection to gain insights into group dynamics. By leveraging the categorical features "daytype", "fasciaoraria", and "datatype", we group the data, enabling a nuanced analysis of behavioral patterns within homogeneous subsets. Employing the median approach within each group allows us to discern trends and tendencies specific to distinct segments of our population. This method not only enhances our understanding of group behavior but also facilitates targeted interventions or optimizations tailored to each subgroup's unique characteristics.

Furthermore, given the logarithmic nature of the data, the possibility of missing records exists. However, thorough examination has revealed that each microcode consistently contains 80 records, indicating a proper data integrity within the dataset. This uniformity underscores the reliability of the dataset and enhances confidence in its accuracy for subsequent analyses and decision-making processes.

2.2.2.2 Outliers

In our decision-making process, we have chosen to retain all outliers within the dataset. This decision is based on our recognition that outliers are possible and may have a significant impact on our results. By keeping outliers, we ensure our analysis remains robust and reflective of the true nature of the data, allowing for a comprehensive exploration of potential influential factors. This approach enables us to maintain the integrity of our analysis and uncover valuable insights that outliers may provide.

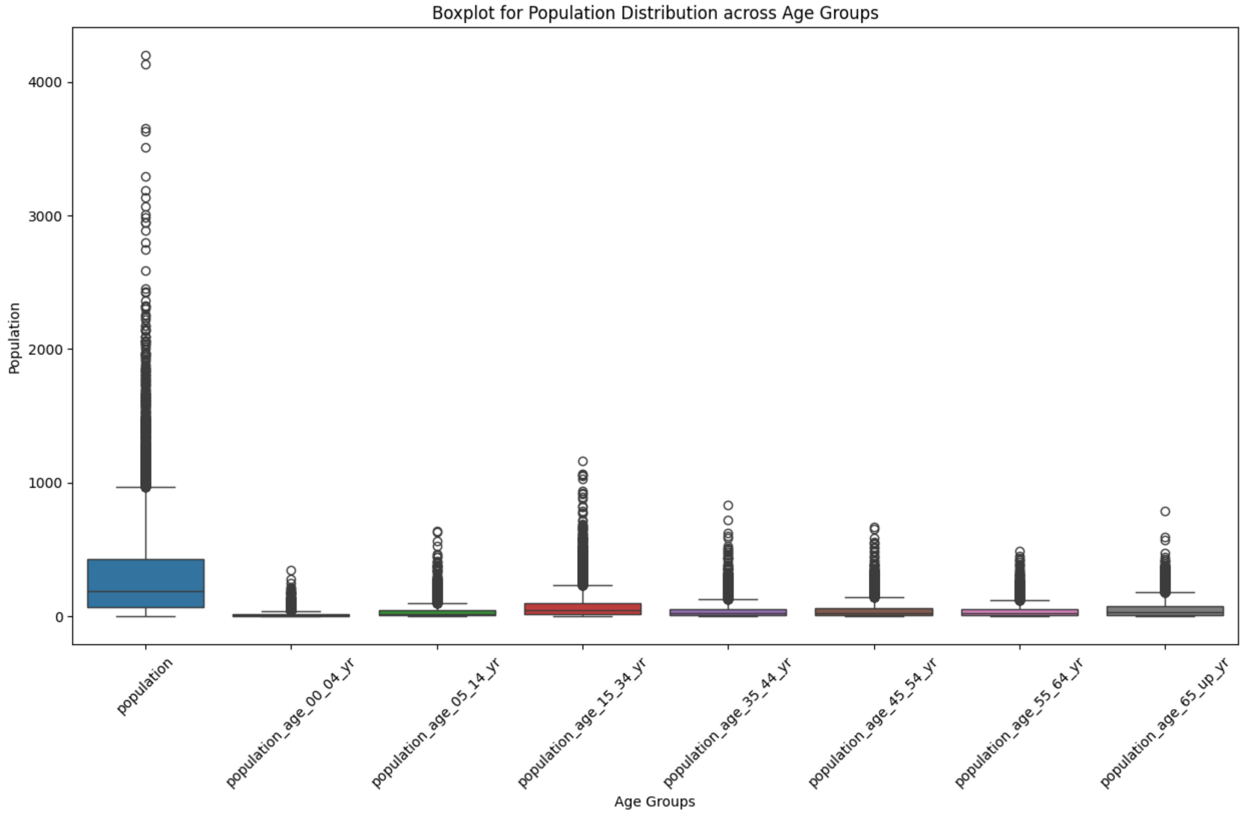


Fig.

This outlier diagram represents population data within our dataset. It is not unusual to observe a few densely populated zones within a given area, which should not be disregarded as anomalies. Instead, such instances are integral to our analysis and should be considered as part of the overall population distribution. These outliers provide valuable insights into localized population dynamics and are essential for a comprehensive understanding of the dataset.

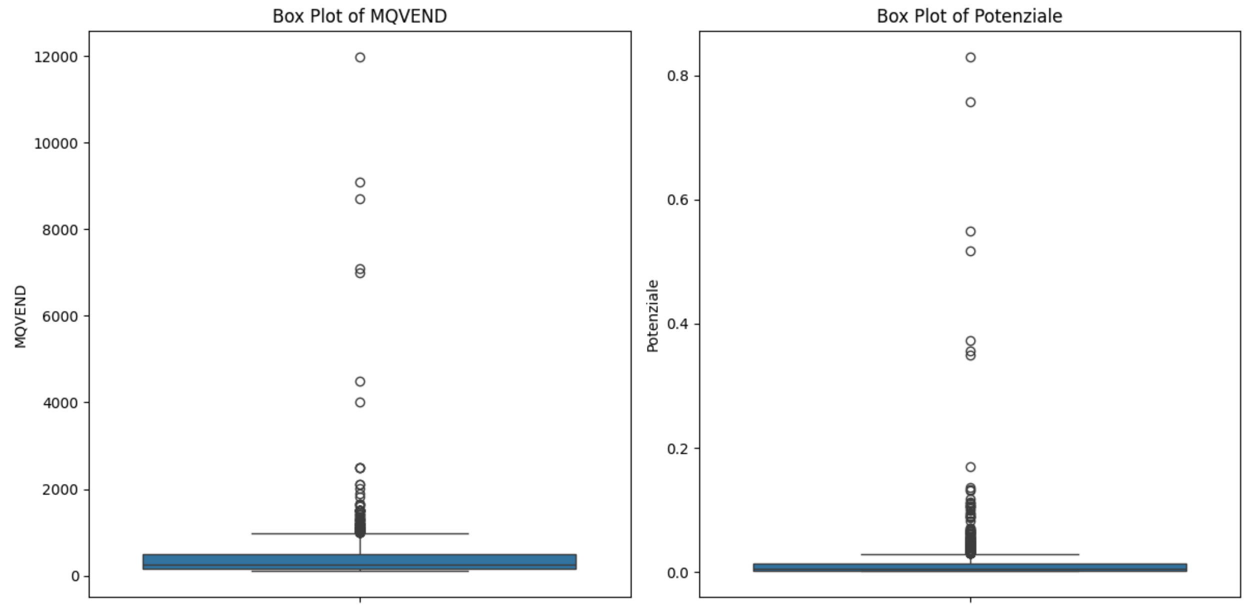
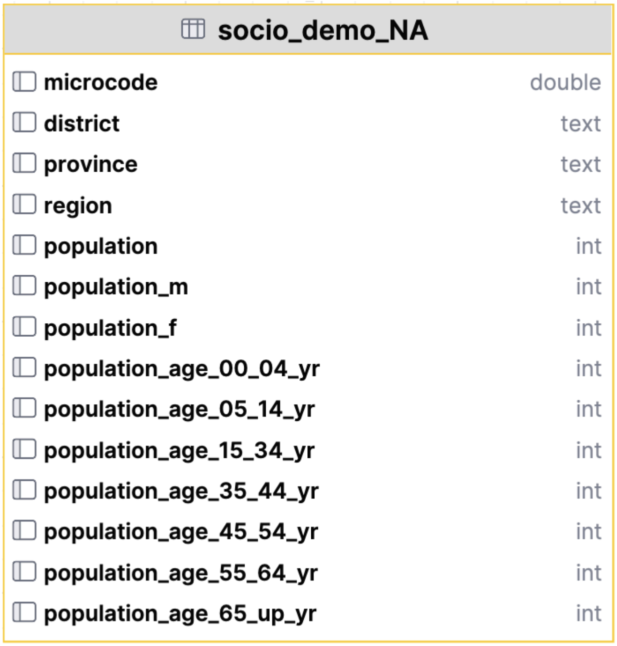


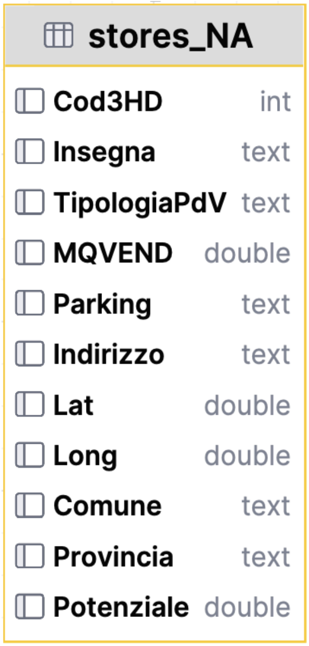
Fig.

2.2.2.3 Dealing Unnecessary Columns

In this section, we thoroughly review each column in our datasets to identify and remove any unnecessary ones, ensuring we retain only practical data for analysis. We only drop columns that are clearly useless, without needing any more study.



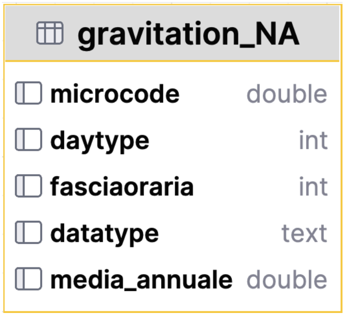
socio_demo_NA	
microcode	double
district	text
province	text
region	text
population	int
population_m	int
population_f	int
population_age_00_04_yr	int
population_age_05_14_yr	int
population_age_15_34_yr	int
population_age_35_44_yr	int
population_age_45_54_yr	int
population_age_55_64_yr	int
population_age_65_up_yr	int



stores_NA	
Cod3HD	int
Insegna	text
TipologiaPdV	text
MQVEND	double
Parking	text
Indirizzo	text
Lat	double
Long	double
Comune	text
Provincia	text
Potenziale	double

As we focus solely on data pertaining to the Napoli province within the Campania region, it is evident that the columns "province" and "region" serve no purpose and thus will be removed from the Socio Demo

NA dataset. For the same reason, we drop the column “Provincia” in the Stores NA dataset along with the “Indrizzo”. Also, the column “Unnamed: 0” is detected, which has been removed. The remaining columns will be retained for continued analysis.



gravitation_NA	
microcode	double
daytype	int
fasciaoraria	int
datatype	text
media_annuale	double

Moreover, having leveraged the columns "daytype" and "fasciaoraria" to discern patterns aiding in the completion of missing values within media_annuale, their relevance has diminished. Consequently, we will now proceed to discard these columns from the dataset, as they no longer serve a purpose.

2.2.3 Feature Creation

In this section, we'll be creating new columns from existing ones to aid our analysis. By extracting meaningful information and patterns from the data, we aim to enhance our understanding and uncover valuable insights. This process, known as feature creation or engineering, is essential in refining our analysis and enabling more effective decision-making.

2.2.3.1 Stores NA

Upon thorough investigation of the dataset, an immediate inconsistency emerges: the absence of a connection between the store dataset and others. Unlike the rest, the store dataset lacks the crucial "microcode" column. This omission obstructs seamless integration and analysis across datasets. However, leveraging our access to the polygons associated with each microcode and the coordinates of every store, we can infer whether a store falls within a specific zone. By determining the appropriate zone and subsequently matching it with the corresponding microcode for each store, we can rectify this discrepancy and establish a coherent framework for comprehensive analysis by the microcode column to the store dataset and revealing which zone each store belongs to.

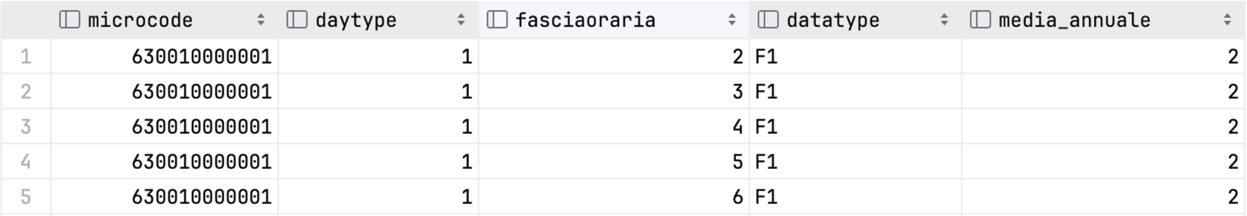
Following the execution of this operation, the columns labeled "Lat" and "Long" will be eliminated from the Stores database due to their lack of utility. Genuinely, the time of gravitation has no impact on the revenue of a store.

2.2.3.2 Gravitation NA

Upon first glance, we adopt a different approach. After addressing missing data, we pivot the dataset to analyze it from a new angle. This reorganization groups the data by microcode, allowing for a better understanding of each microcode's statistics and reducing the dataset's size.

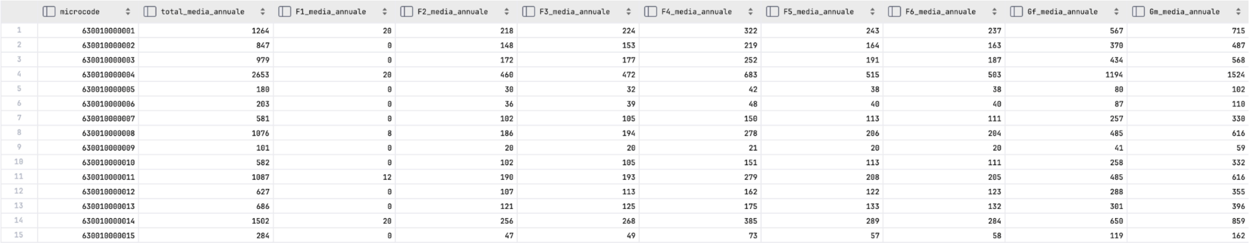
In this updated format, each "datatype" is displayed in its own column, showing the total "media_annuale" for its demographic group, regardless of time.

Here, a sample of data before and after this transformation is visualized:



	microcode	daytype	fasciaoraria	datatype	media_annuale
1	630010000001	1	2	F1	2
2	630010000001	1	3	F1	2
3	630010000001	1	4	F1	2
4	630010000001	1	5	F1	2
5	630010000001	1	6	F1	2

Fig. - Gravitation NA Before Transformation



	microcode	total_media_annuale	F1_media_annuale	F2_media_annuale	F3_media_annuale	F4_media_annuale	F5_media_annuale	F6_media_annuale	Gf_media_annuale	Gm_media_annuale
1	630010000001	1264	20	218	224	322	243	237	567	715
2	630010000002	847	0	148	153	219	164	163	370	487
3	630010000003	979	0	172	177	252	191	187	434	568
4	630010000004	2653	20	460	472	683	515	503	1194	1524
5	630010000005	180	0	30	32	42	38	38	80	102
6	630010000006	203	0	36	39	48	40	40	87	110
7	630010000007	581	0	102	105	150	113	111	257	330
8	630010000008	1076	8	186	194	278	206	204	485	616
9	630010000009	101	0	20	20	21	20	20	41	59
10	630010000010	582	0	102	105	151	113	111	258	332
11	630010000011	1087	12	190	193	279	208	205	485	616
12	630010000012	627	0	107	113	162	122	123	288	355
13	630010000013	686	0	121	125	175	133	132	301	396
14	630010000014	1502	20	256	268	385	289	284	650	859
15	630010000015	284	0	47	49	73	57	58	119	162

Fig. - Gravitation NA After Transformation

2.2.3.3 Socio Demo NA

In the Socio Demo dataset, the main focus is on how population is distributed across different zones. However, when it comes to products like diapers, which are mostly used by infants and elders, the raw population data doesn't tell the whole story. To address this, we need to rate each zone's population based on its relevance to diapers. We'll do this by introducing a new variable called "Population Score" for each zone. This score, represented as a new column in our dataset, will help us better understand the importance of each zone's population when it comes to products like diapers.

2.2.3.3.1 Introducing the Population Score

The concept of Population Score introduces a refined approach to assessing the value of our target market within the Naples province. By assigning specific weights (labeled as W) to distinct age ranges, we tailor our evaluation to the nuanced demographics of each population segment.

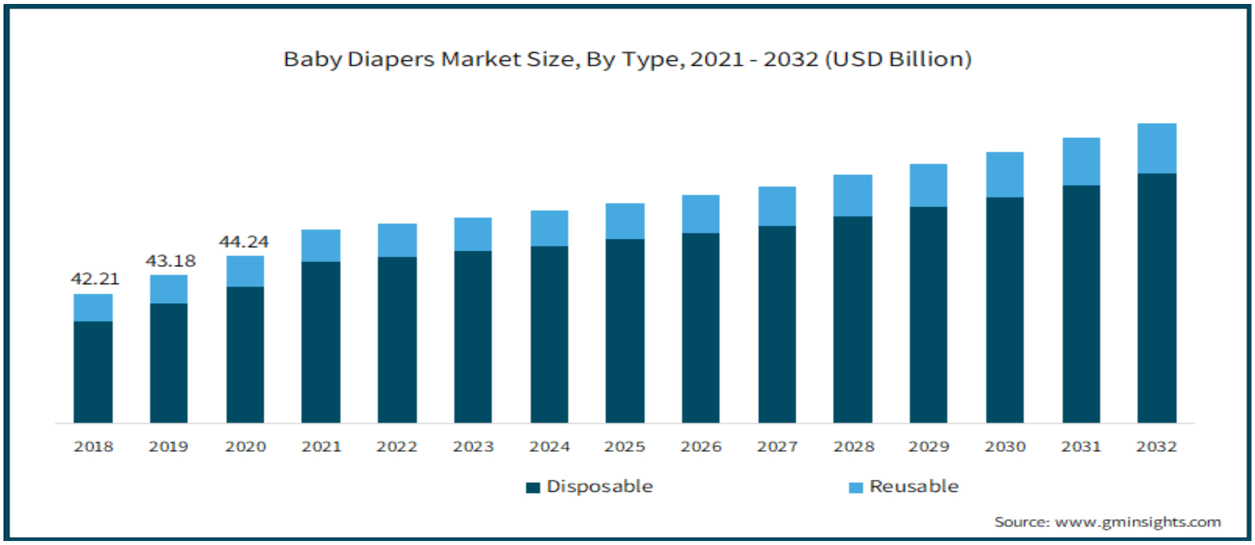
This strategic weighting is calculated through a precise formula:

$$Population\, Score\ = W_{0-4}P_{0-4} + W_{5-14}P_{5-14} + \ ...\ + W_{65-up}P_{65-up}$$

In this formula, W signifies the weight assigned to each age range, reflecting its significance, whereas P denotes the population count within the respective age group, providing crucial demographic context. The utilization of Population Score surpasses the simplicity of traditional population metrics, offering a more accurate and insightful method for evaluating the significance of different age groups within our targeted zones.

2.2.3.3.2 Determining Age-Range Weighting: Factors and Outcomes

While achieving pinpoint accuracy in determining the precise ratio between end-users remains challenging due to the absence of comprehensive data, we strive to attain a reliable approximation. According to Global Market Insight (GMI), the Baby Diapers Market was valued at approximately USD 46.47 billion in 2022 and is projected to soar to USD 70 billion by 2032, marking a notable 4.3% compound annual growth rate (CAGR)[1].



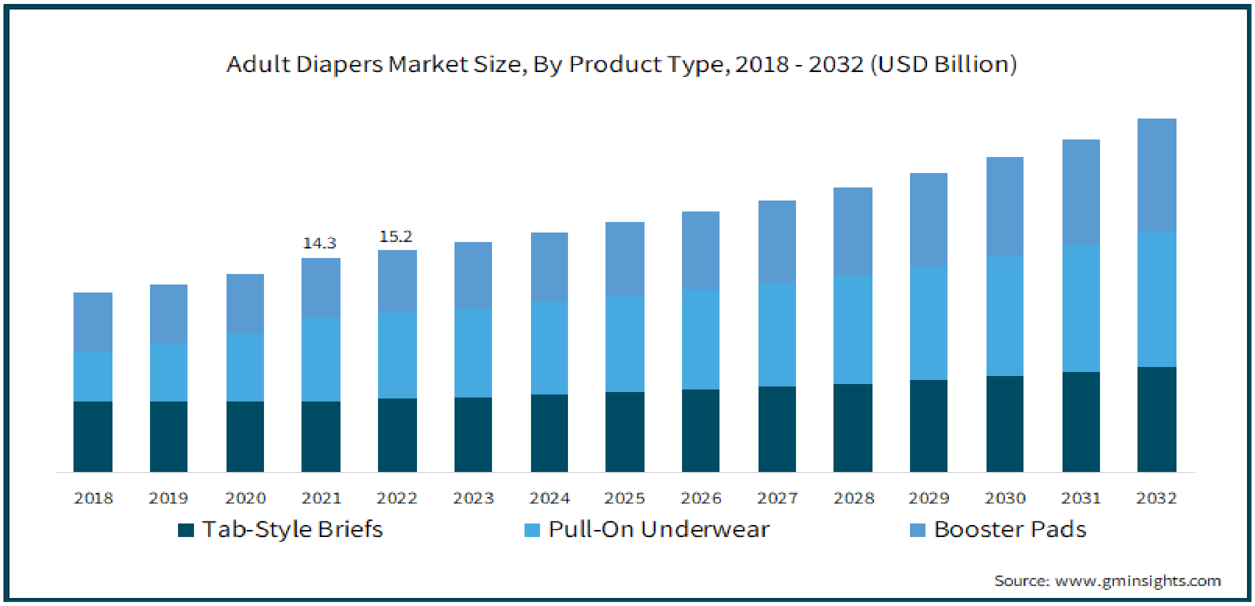
Source: www.gminsights.com

Fig. Baby Diapers Market Size by GMI

Meanwhile, the Adult Diapers Market, with a robust CAGR exceeding 6.8%, boasted a market size of approximately USD 15.2 billion, poised to escalate to USD 29.2 billion in revenue by 2032.[2]

[1] https://www.gminsights.com/industry-analysis/baby-diapers-market

[2] https://www.gminsights.com/industry-analysis/adult-diapers-market



Source: www.gminsights.com

Based on this information, we can derive ratios for various age ranges, providing a direct indication of their market significance.

Population Age Range	Population Weight (W)
0 - 4	0.046
5 - 14	0
15 - 34	0
35 - 44	0
45 - 54	0
55 - 64	0.1
65 - up	0.015

In conclusion, once we attain W and possess P, we can proceed to calculate the Population Score and incorporate it for each zone.

This heat map illustrates the population distribution across zones in Napoli province.

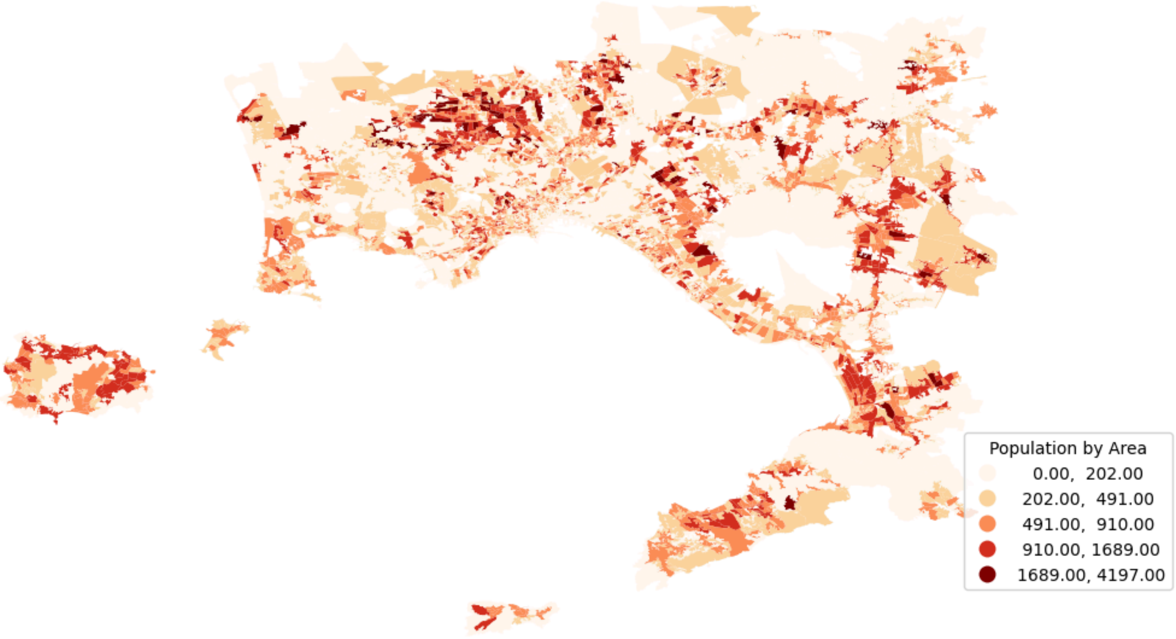


Fig. Napoli Province Zones Population

Below is a heatmap visualizing the Population Score.

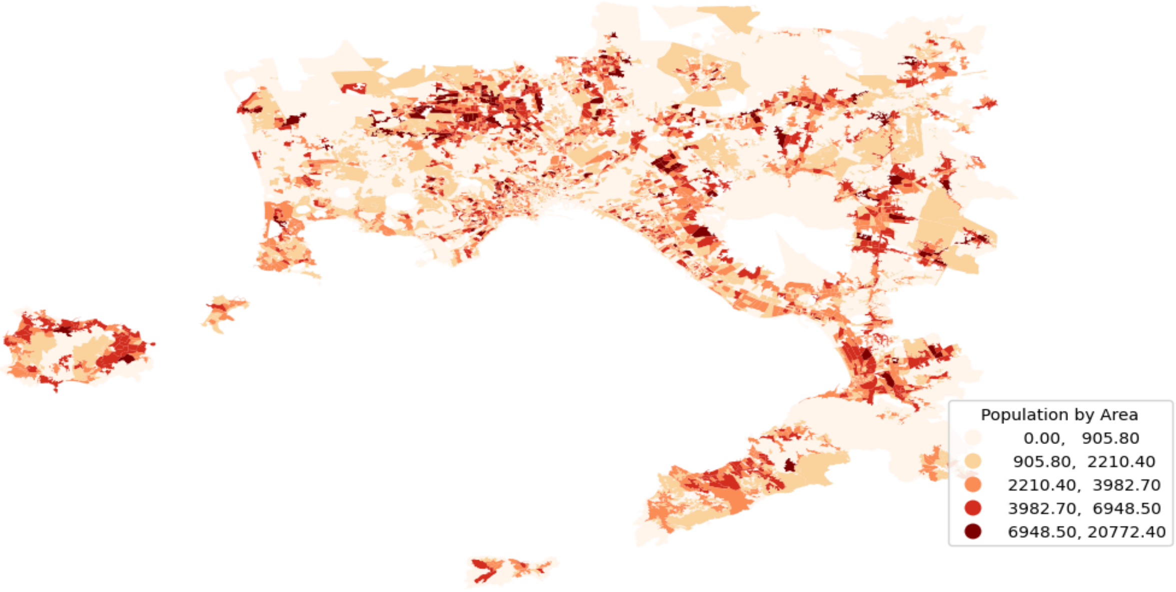


Fig. Napoli Province Zones Population Score

2.2.3.3.3 Classification Data

We've improved our dataset by adding a new column called "population_class" using the KMeans algorithm. This step categorizes our data based on values in the "population_score" column. To decide how many categories (or clusters) we should create, we used the Elbow method. This method helps us find the right balance—where adding more clusters doesn't give us much more detail. It suggested that three clusters (or groups) were just right. So, we went ahead and organized our data into these three distinct groups. This new classification makes our dataset richer and sets the stage for deeper analysis, allowing us to uncover more detailed insights into our population data.

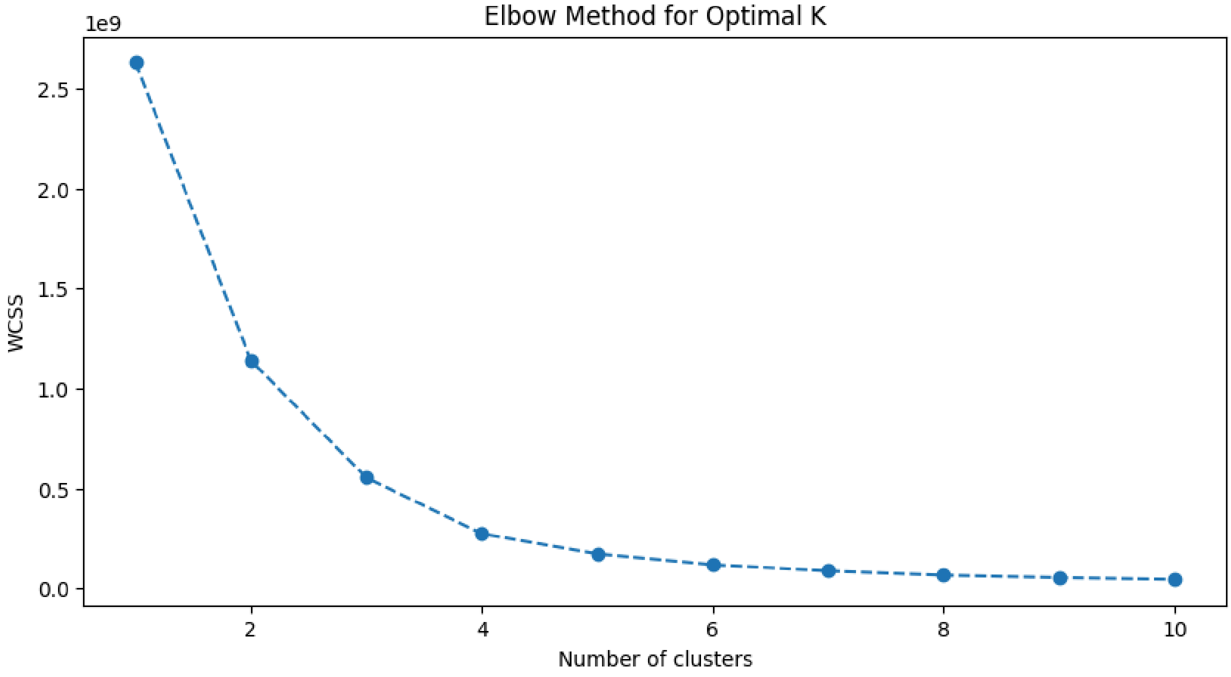


Fig. Elbow Method Result

2.2.3.4 Shapes NA

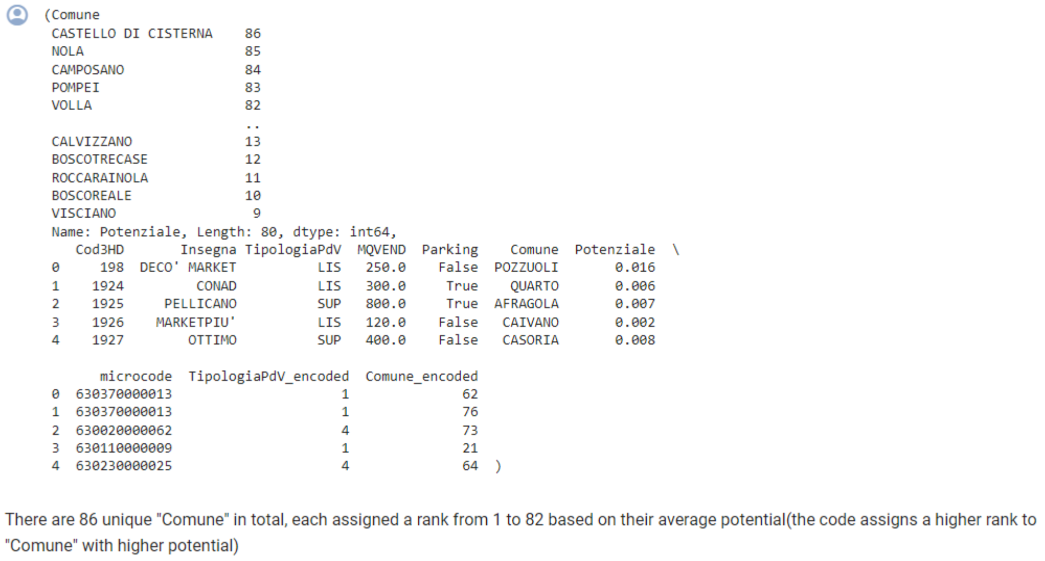
In this dataset, we explore zone polygons to identify neighboring areas for each zone. By leveraging geographic data, we add a "neighbors" column to the dataset, listing the microcodes of zones sharing common borders. Studying neighboring behaviors offers valuable insights for our analysis.

2.2.4 Data Encoding

Data encoding is the process of transforming categorical data into a numerical format so that machine learning algorithms can process it, as most of these algorithms require numerical input.

2.2.4.1 Comune Encoding

To encode the categorical values of the "Comune" feature, we employ a ranking-based method that emphasizes the potential associated with each "Comune." This process begins by calculating the mean potential for each "Comune," highlighting the varying levels of potential across different comunes. Following this, we assign ranks to each "Comune" based on their calculated potential, where a higher potential leads to a higher rank. This method ensures that "Comunes" with greater potential are distinctly recognized and given precedence in the encoding process, facilitating a more nuanced analysis that accounts for the inherent potential within each region.



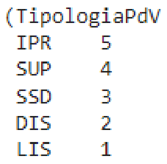
```
(Comune
 CASTELLO DI CISTERNA    86
 NOLA                    85
 CAMPOSANO               84
 POMPEI                  83
 VOLLA                   82
                         ..
 CALVIZZANO              13
 BOSCOTRECASE            12
 ROCCARAINOLA            11
 BOSCOREALE              10
 VISCIANO                 9
 Name: Potenziale, Length: 80, dtype: int64,
    Cod3HD       Insegna TipologiaPdV  MQVEND  Parking    Comune  Potenziale  \
 0     198  DECO' MARKET          LIS   250.0    False  POZZUOLI       0.016
 1    1924         CONAD          LIS   300.0     True    QUARTO       0.006
 2    1925     PELLICANO          SUP   800.0     True  AFRAGOLA       0.007
 3    1926    MARKETPIU'          LIS   120.0    False   CAIVANO       0.002
 4    1927        OTTIMO          SUP   400.0    False   CASORIA       0.008

       microcode  TipologiaPdV_encoded  Comune_encoded
 0  630370000013                     1              62
 1  630370000013                     1              76
 2  630020000062                     4              73
 3  630110000009                     1              21
 4  630230000025                     4              64  )
```

There are 86 unique "Comune" in total, each assigned a rank from 1 to 82 based on their average potential(the code assigns a higher rank to "Comune" with higher potential)

Fig.

2.2.4.2 TipologiaPdV Encoding

Following the same approach as for the "Comune" feature, the "TipologiaPdV" feature is also encoded based on the average potential associated with each store type ("TipologiaPdV"). The ranks are assigned inversely relative to the average potential



```
(TipologiaPdV
 IPR    5
 SUP    4
 SSD    3
 DIS    2
 LIS    1
```

2.2.4.3 Insegna Encoding

The encoding for the "Insegna" feature follows the same methodology as for "Comune" and "TipologiaPdV". This method calculates the average potential for each brand ("Insegna") and then ranks these brands based on their average potential. The brand with the highest average potential receives the lowest rank number (indicating a higher priority or importance), and the ranks increase as the average potential decreases. This ranking system effectively captures the contribution of each brand to the overall potential, allowing for a nuanced analysis of how different brands influence the potential outcomes.

2.2.4.4 Parking Encoding

For the "Parking" attribute, originally binary with true/false values, we transformed it into a numerical format, assigning 0 for false and 1 for true. This encoding step ensures uniform data types across all features, which is essential for improving the accuracy and interpretability of our model predictions.

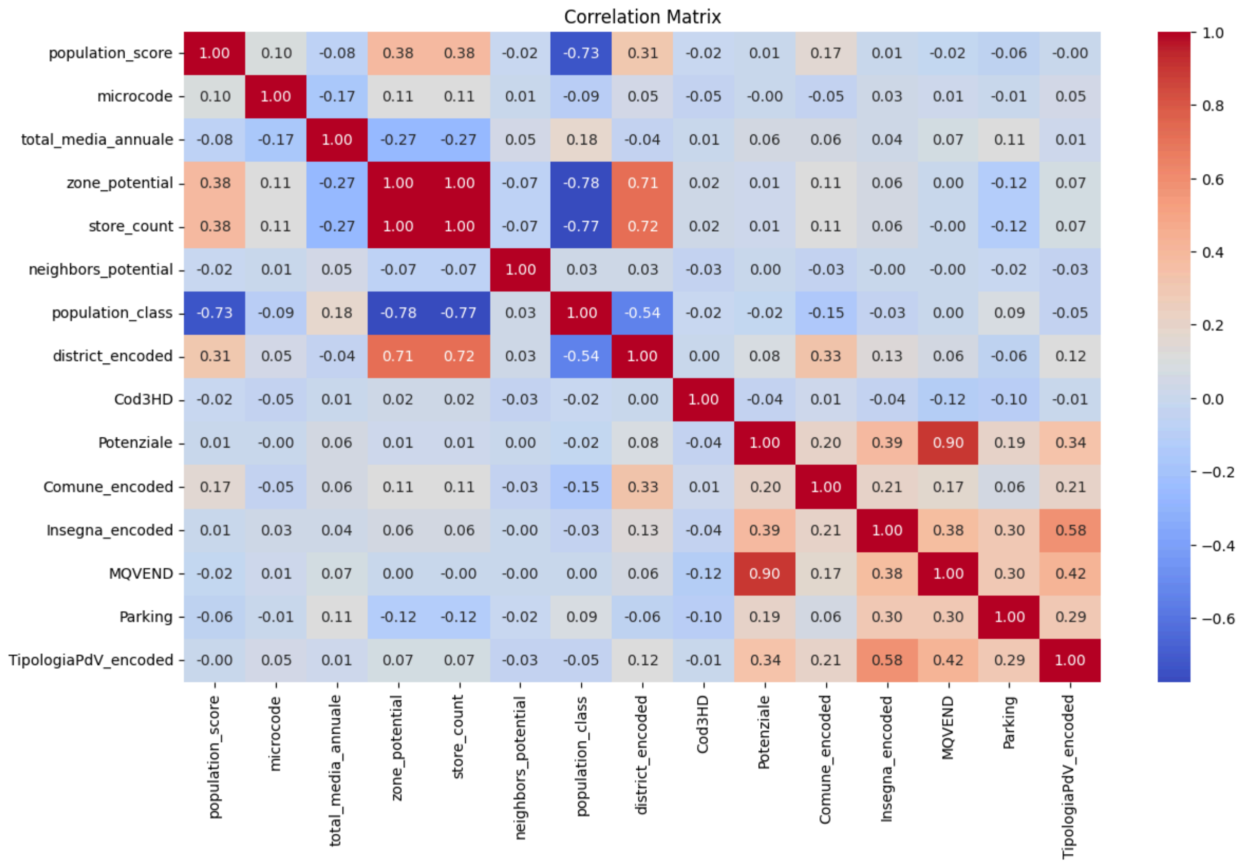
2.2.5 Data Normalization

In this project, we chose not to normalize the data before feeding it into the Random Forest model. This decision is based on the nature of Random Forest, which is a collection of decision trees. Decision trees, and consequently Random Forests, are less sensitive to the scale of the data. This is because they make decisions based on thresholds and do not rely on the distance between features, unlike models such as K-Nearest Neighbors or Gradient Descent-based algorithms where normalization can significantly impact performance. Therefore, normalizing the data is not a prerequisite for Random Forest, allowing us to maintain the original scale of our features without compromising the model's effectiveness.

3. Data Merging

We approach our data analysis from two perspectives: zone-based and store-based. The majority of our data focuses on zone attributes such as population, gravitation, and geometries, while the remaining data pertains to store characteristics like parking availability and size.

To consolidate our zone-related data, we merge various datasets using the microcode as a key identifier. This integration results in a comprehensive dataset encompassing all zone attributes.
Subsequently, we merge this consolidated zone dataset with the store dataset using the microcode added to the stores dataset as a common identifier. This combined dataset provides a holistic view of both zone and store attributes, facilitating thorough investigation and analysis in subsequent stages.



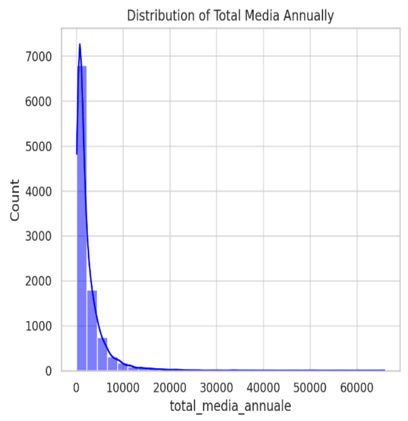
	population_score	microcode	total_media_annuale	zone_potential	store_count	neighbors_potential	population_class	district_encoded	Cod3HD	Potenziale	Comune_encoded	Insegna_encoded	MQVEND	Parking	TipologiaPdV_encoded
population_score	1.00	0.10	-0.08	0.38	0.38	-0.02	-0.73	0.31	-0.02	0.01	0.17	0.01	-0.02	-0.06	-0.00
microcode	0.10	1.00	-0.17	0.11	0.11	0.01	-0.09	0.05	-0.05	-0.00	-0.05	0.03	0.01	-0.01	0.05
total_media_annuale	-0.08	-0.17	1.00	-0.27	-0.27	0.05	0.18	-0.04	0.01	0.06	0.06	0.04	0.07	0.11	0.01
zone_potential	0.38	0.11	-0.27	1.00	1.00	-0.07	-0.78	0.71	0.02	0.01	0.11	0.06	0.00	-0.12	0.07
store_count	0.38	0.11	-0.27	1.00	1.00	-0.07	-0.77	0.72	0.02	0.01	0.11	0.06	-0.00	-0.12	0.07
neighbors_potential	-0.02	0.01	0.05	-0.07	-0.07	1.00	0.03	0.03	-0.03	0.00	-0.03	-0.00	-0.00	-0.02	-0.03
population_class	-0.73	-0.09	0.18	-0.78	-0.77	0.03	1.00	-0.54	-0.02	-0.02	-0.15	-0.03	0.00	0.09	-0.05
district_encoded	0.31	0.05	-0.04	0.71	0.72	0.03	-0.54	1.00	0.00	0.08	0.33	0.13	0.06	-0.06	0.12
Cod3HD	-0.02	-0.05	0.01	0.02	0.02	-0.03	-0.02	0.00	1.00	-0.04	0.01	-0.04	-0.12	-0.10	-0.01
Potenziale	0.01	-0.00	0.06	0.01	0.01	0.00	-0.02	0.08	-0.04	1.00	0.20	0.39	0.90	0.19	0.34
Comune_encoded	0.17	-0.05	0.06	0.11	0.11	-0.03	-0.15	0.33	0.01	0.20	1.00	0.21	0.17	0.06	0.21
Insegna_encoded	0.01	0.03	0.04	0.06	0.06	-0.00	-0.03	0.13	-0.04	0.39	0.21	1.00	0.38	0.30	0.58
MQVEND	-0.02	0.01	0.07	0.00	-0.00	-0.00	0.00	0.06	-0.12	0.90	0.17	0.38	1.00	0.30	0.42
Parking	-0.06	-0.01	0.11	-0.12	-0.12	-0.02	0.09	-0.06	-0.10	0.19	0.06	0.30	0.30	1.00	0.29
TipologiaPdV_encoded	-0.00	0.05	0.01	0.07	0.07	-0.03	-0.05	0.12	-0.01	0.34	0.21	0.58	0.42	0.29	1.00

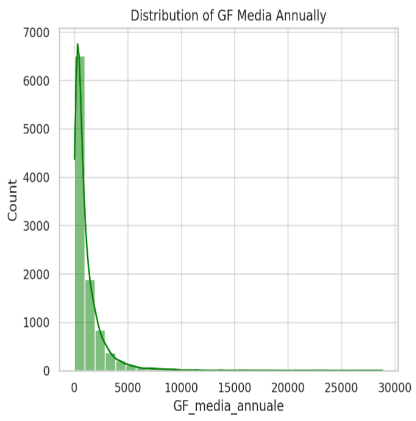
Fig. Correlation Matrix of the Merged Dataset

3. Exploratory Data Analysis

3.1 Gravitation NA

3.1.1 Distribution & Feature Analysis





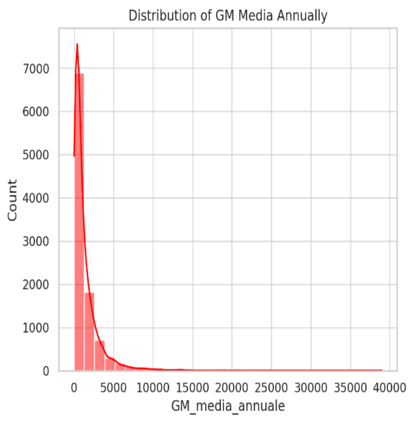


Fig. Distribution of the Total, Female, and Male media annuale in the Province of Napoli

The histograms for total_media_annuale, GF_media_annuale, and GM_media_annuale reveal key aspects of their distribution across the dataset:

- Total Media Annually shows a right-skewed distribution, indicating that most entries have lower annual media values, with a few exceptions having significantly higher values. This suggests variability in the dataset, with certain areas or cases exhibiting much higher media attention or relevance.
- GF Media Annually and GM Media Annually distributions are similarly right-skewed, highlighting that, like the total media values, these specific categories also have a few high-value outliers compared to the majority.

These distributions and statistical summaries highlight the diverse and skewed nature of the media values within the dataset, pointing towards a variety of gravitational phenomena or events captured in these measurements. The presence of outliers suggests specific instances or areas with unusually high gravitational activity or interest, warranting further investigation.

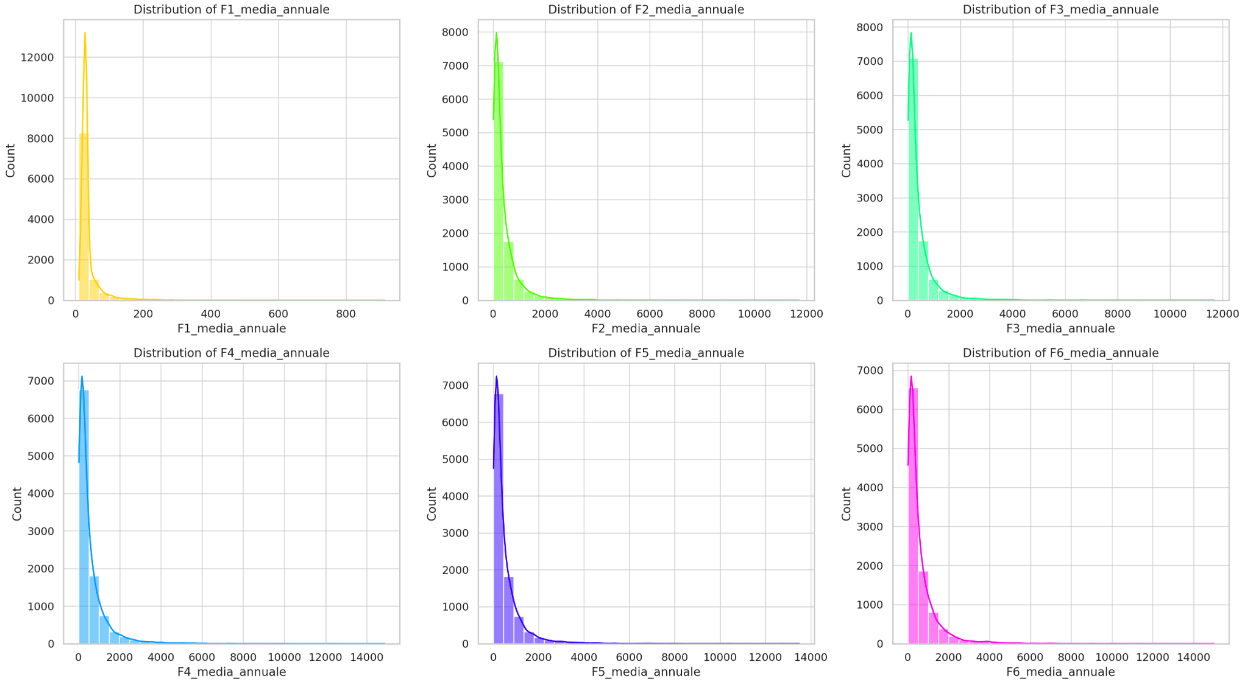


Fig. The Distribution of Different Age Groups Gravitation in the Province of Napoli

The histograms for features F1_media_annuale through F6_media_annuale reveal diverse distributions across these gravitational measurements, each with unique characteristics:

- F1_media_annuale shows a concentration of lower values with a long tail extending towards higher values. This pattern suggests a prevalence of lower gravitational measurements with a few exceptional high values.
- F2_media_annuale to F6_media_annuale similarly exhibit right-skewed distributions, indicating a general trend of lower measurement values while still possessing outliers on the higher end. The

extent of skewness varies across these features, with F4 and F5 showing particularly broad tails, suggesting significant variability in these gravitational measurements.

- Standard deviations are notably high for all features, especially for F2 through F6, underscoring the wide range of values within each feature. The large standard deviations point to significant variability in the gravitational measurements, reflecting diverse or dynamic gravitational environments.
- Range of values demonstrates the extensive variability within each feature, with maximum values (e.g., 14,861 for F4) significantly larger than the minimums (as low as 10), highlighting the presence of extreme gravitational measurements or anomalies within the dataset.

3.1.2 Correlation Analysis

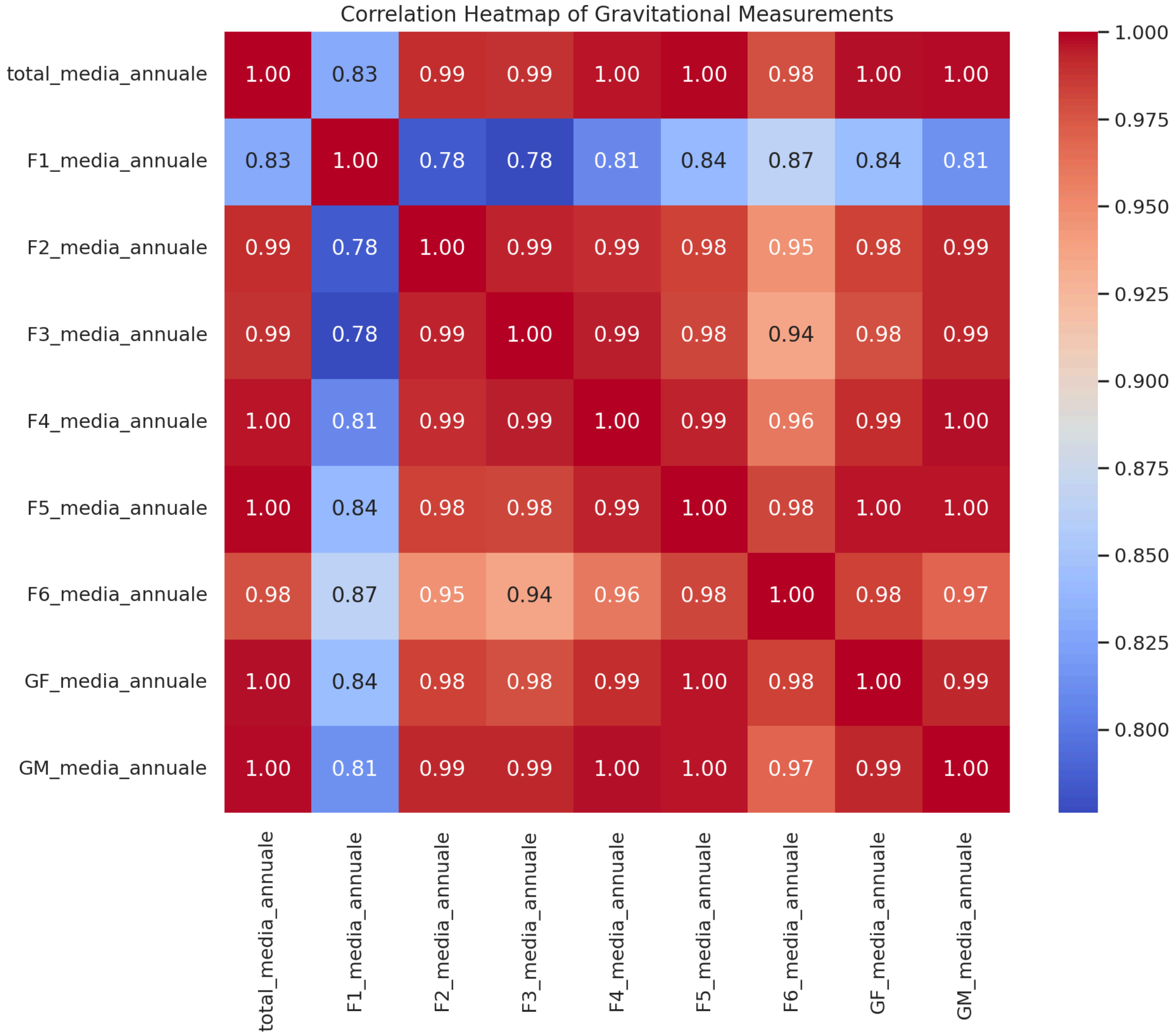


Fig. Gravitation Correlation Heatmap

The correlation heatmap reveals significant relationships between the gravitational measurements in the dataset. Key observations include:

- High Correlation Between Total Media and Other Measurements: The total_media_annuale has a very high correlation with GF_media_annuale and GM_media_annuale (0.997 and 0.998, respectively), indicating that as GF and GM values increase, the total media value also tends to increase. This suggests a strong linear relationship between these variables, implying that they likely contribute significantly to the total media value.
- Strong Inter-correlations Among Specific Measurements: Features F1_media_annuale through F6_media_annuale show strong correlations with each other and with GF_media_annuale and GM_media_annuale. Notably, F4_media_annuale, F5_media_annuale, and F6_media_annuale have particularly high correlations with the total media annual value, indicating that these features play a critical role in determining the total media value.
- Relationships Within Specific Features: The F2_media_annuale, F3_media_annuale, and F4_media_annuale variables exhibit very high correlations with each other (all above 0.99), suggesting that these measurements might be influenced by similar factors or conditions, or perhaps they measure related aspects of gravitational phenomena.

The high degree of correlation between the total media annual value and the specific gravitational measurements suggests that these variables are closely linked, potentially reflecting underlying physical phenomena or measurement methodologies that capture similar aspects of gravitational forces.

3.2 Socio Demo NA

3.2.1 Distribution & Feature Analysis

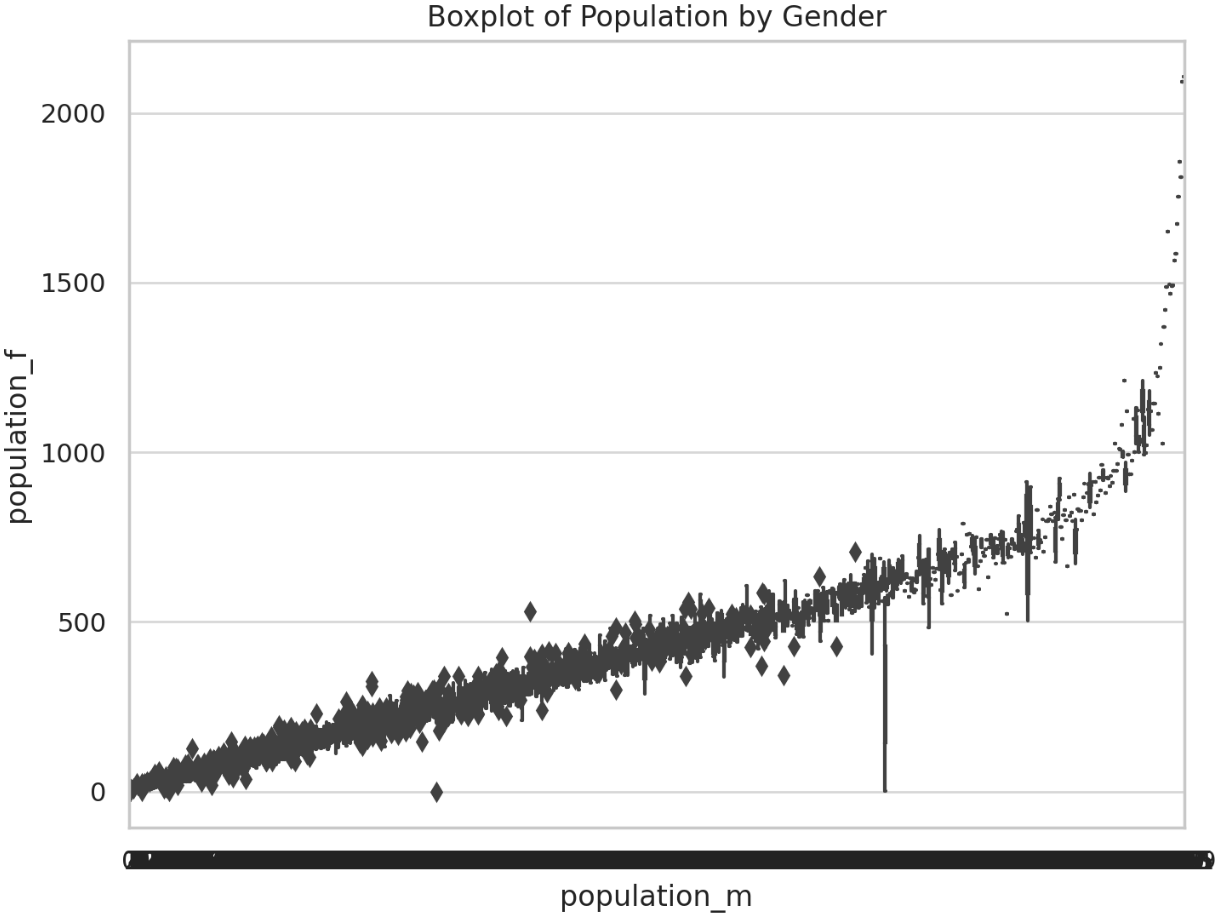


Fig. Boxplot Population by Gender

The boxplot comparing the male (X-axis) and female (Y-axis) populations across districts shows a balanced distribution between genders, with a slight variation in the interquartile range. This balance indicates that there is no significant gender disparity in the overall population numbers. However, the presence of outliers suggests that a few districts have unusually high or low ratios of male to female populations, which could be a point of interest for demographic studies focusing on gender distribution.

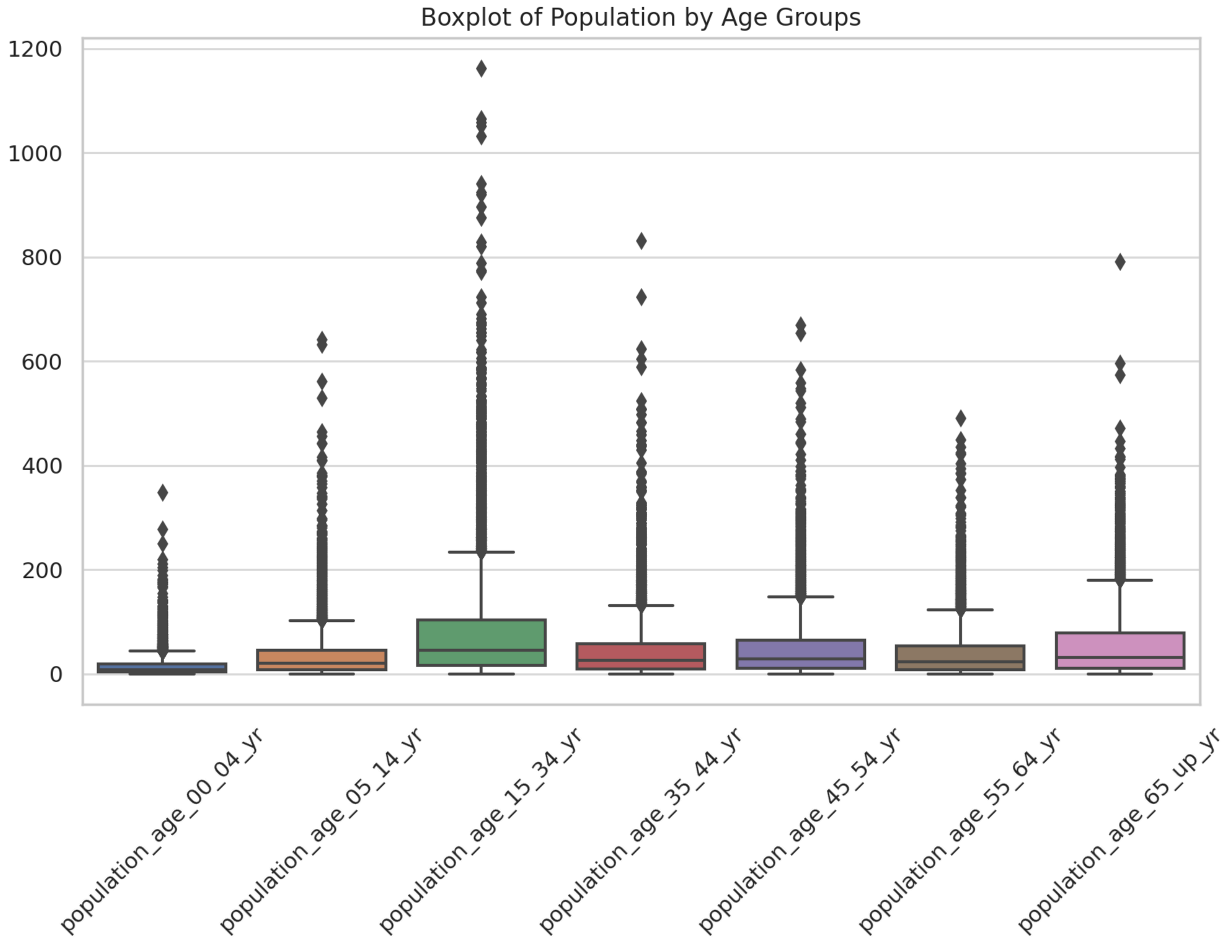


Fig. Boxplot Population by Age Groups

This chart illustrates the distribution of populations across different age groups, providing insight into the demographic structure. The age groups 15-34 and 35-44 years have wider interquartile ranges, indicating a significant proportion of the population within these age brackets. The presence of outliers in almost all age groups suggests that certain districts have higher concentrations of specific age demographics, potentially reflecting local socio-economic conditions, birth rates, or migration patterns.

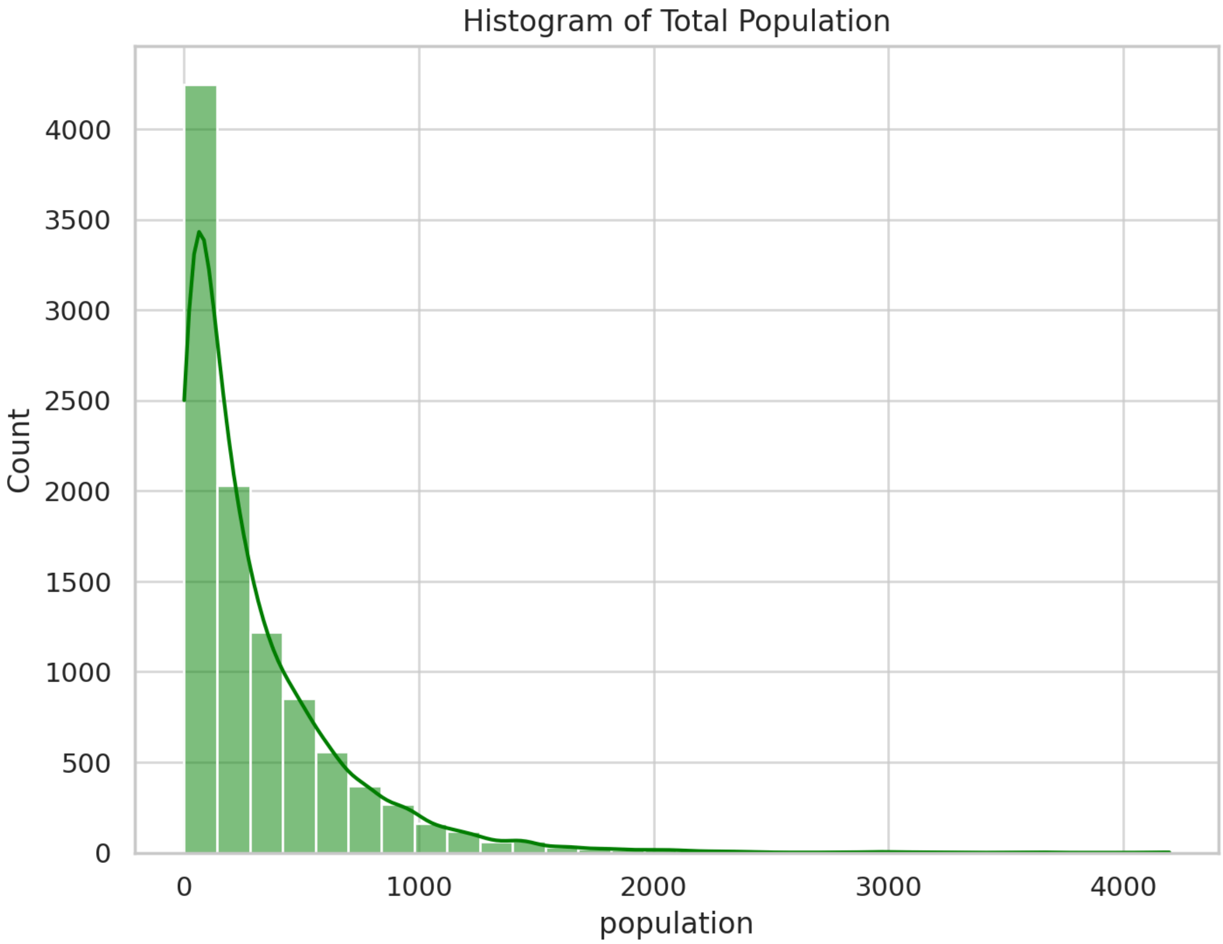


Fig. Histogram of Total Population

This chart shows a right-skewed distribution of the total population across districts, with a concentration of districts having smaller populations and a few districts with significantly larger populations. The skewness points to a pattern of population concentration in certain areas, possibly urban centers, while rural or less developed areas have lower populations. This distribution has implications for resource allocation, urban planning, and regional development policies.

3.2.2 Correlation Analysis

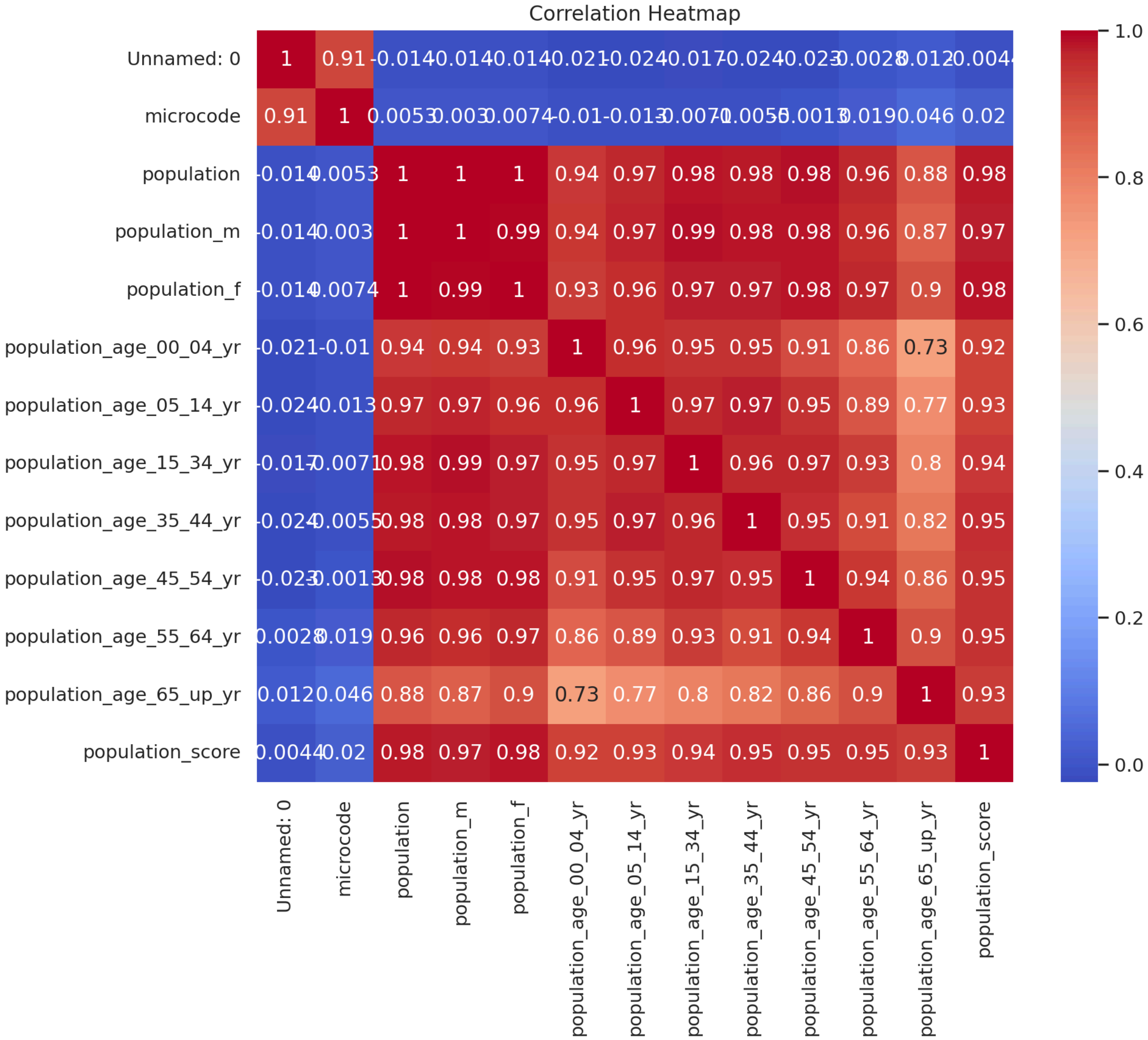


Fig. Correlation Heatmap of Socio Demo

The correlation heatmap provides a comprehensive view of the relationships between different demographic variables. Strong correlations between total population and specific age groups indicate that population size influences the distribution across age groups. The population score's strong correlation with population and age demographics suggests that these factors are closely linked to the socio-demographic status of districts.

3.3 Stores NA

3.3.1 Distribution & Feature Analysis

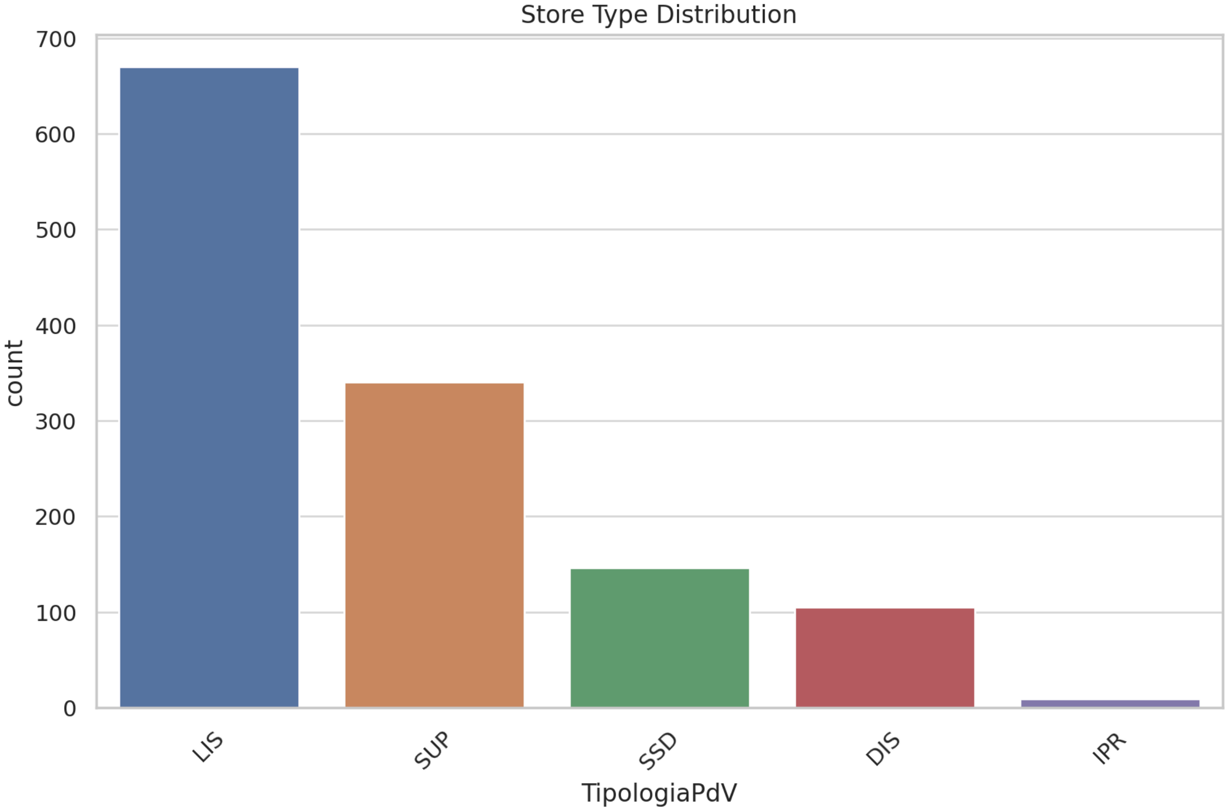


Fig. Store Type Distribution

The analysis of store types within the dataset reveals a variety of retail formats. The count plot shows the prevalence of different types of stores, with some types being more common than others. This distribution provides insight into the retail landscape, indicating a diverse array of store formats catering to various consumer needs.

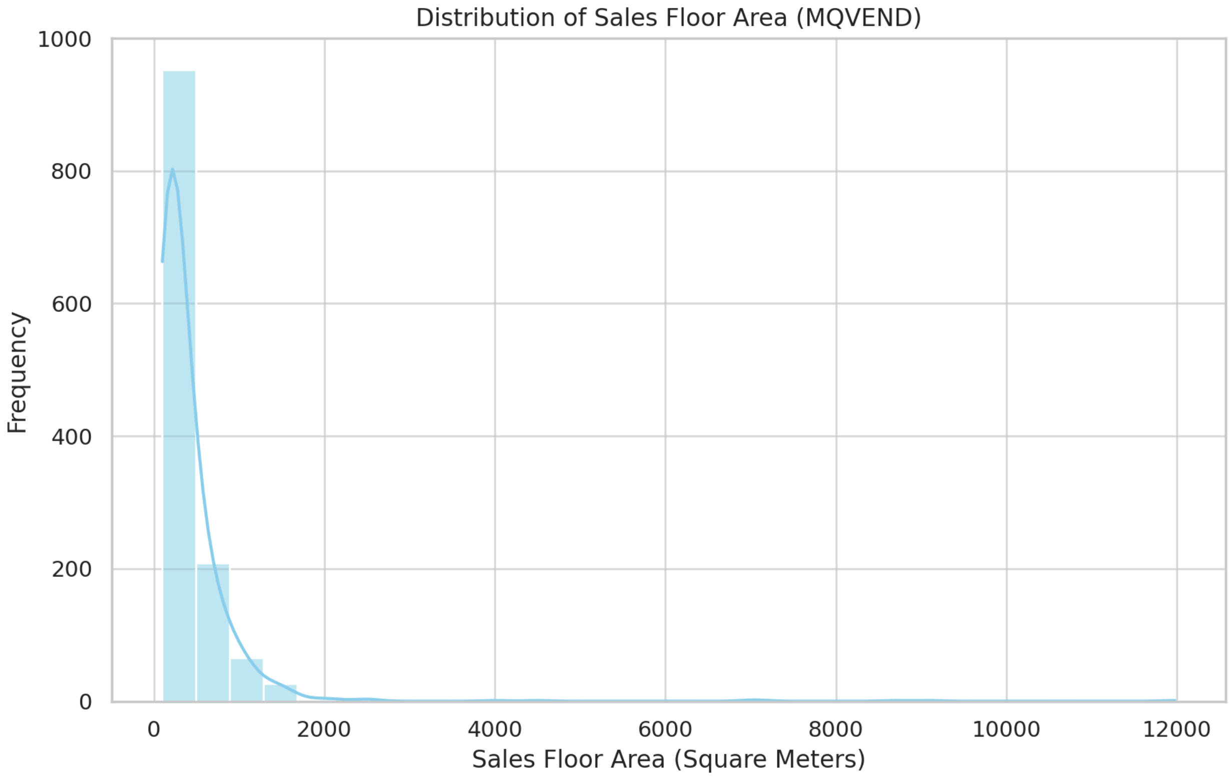


Fig. Distribution of Sales Floor Area

The sales floor area distribution highlights the range of store sizes within the dataset. The histogram shows a right-skewed distribution, indicating that most stores have smaller sales areas, with a few larger establishments. This pattern suggests that small to medium-sized stores dominate the retail environment, with larger stores being less common.

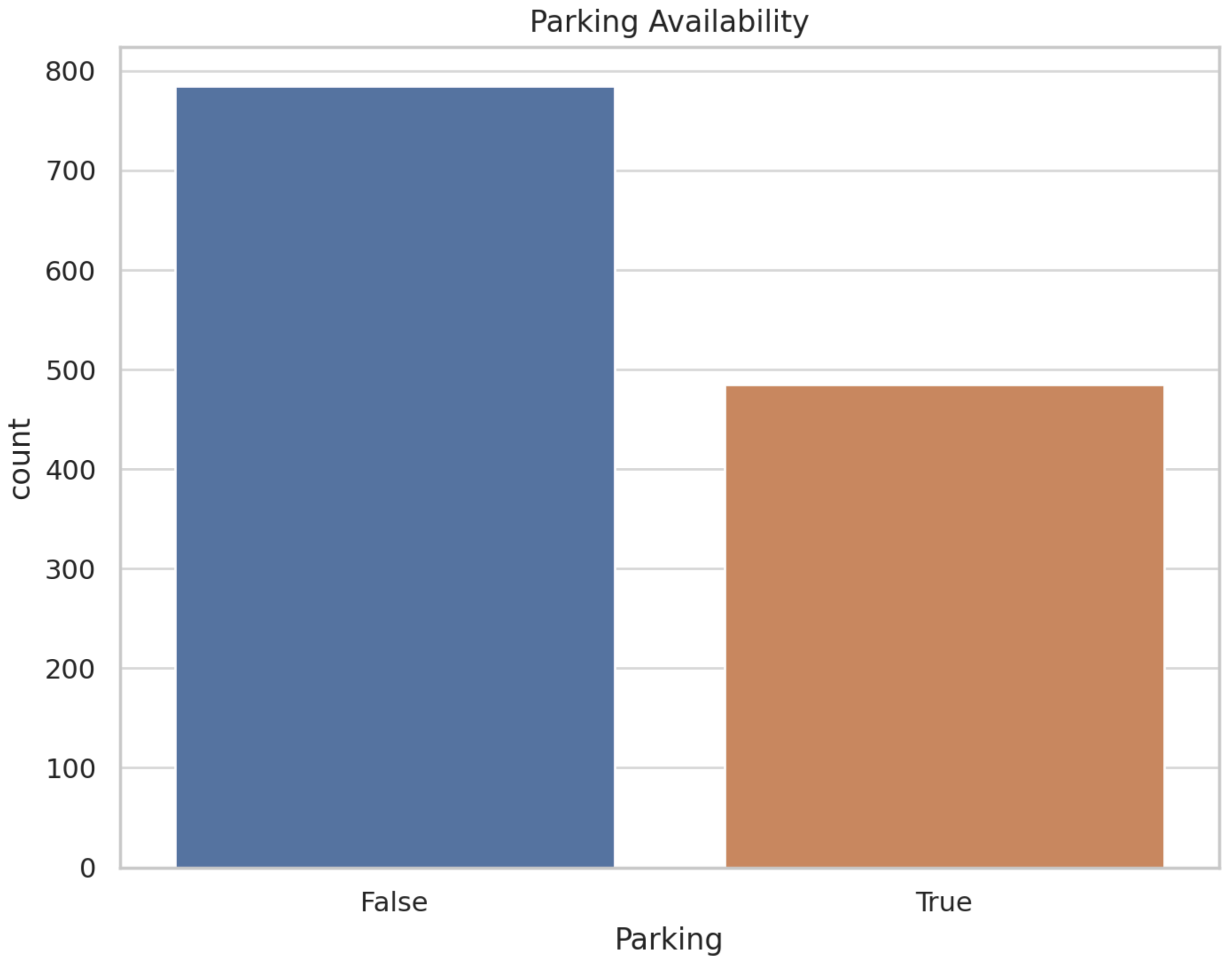


Fig. Parking Availability Distribution

The count plot for parking availability demonstrates the proportion of stores that offer parking facilities. This feature is crucial for customer convenience and can influence store accessibility. The plot indicates a mix of stores with and without parking, suggesting variability in the amenity offerings among the stores.

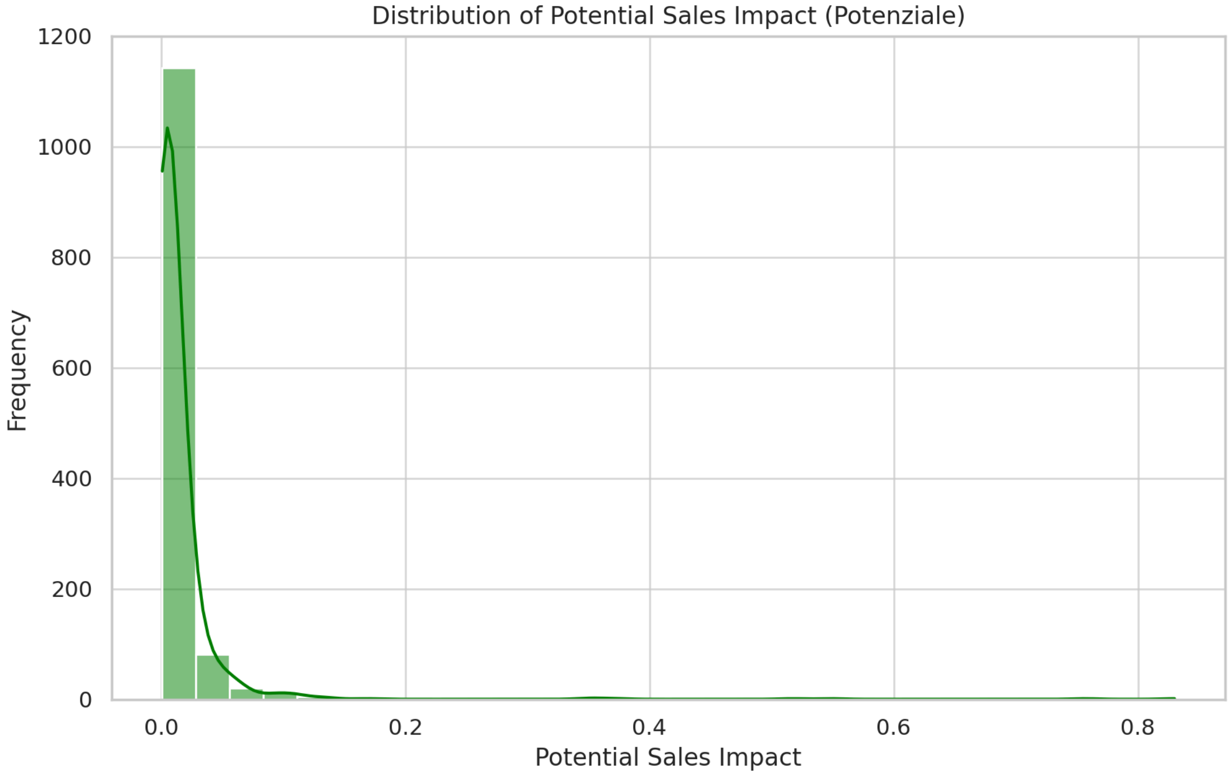


Fig. Distribution of Potential Sales Impact (Potenziale)

The potential sales impact score distribution provides insights into the market influence of stores. The histogram shows a right-skewed distribution, implying that most stores have lower potential sales impact scores, with fewer stores achieving higher scores. This distribution may reflect the competitive landscape and the varying capacities of stores to attract and retain customers.

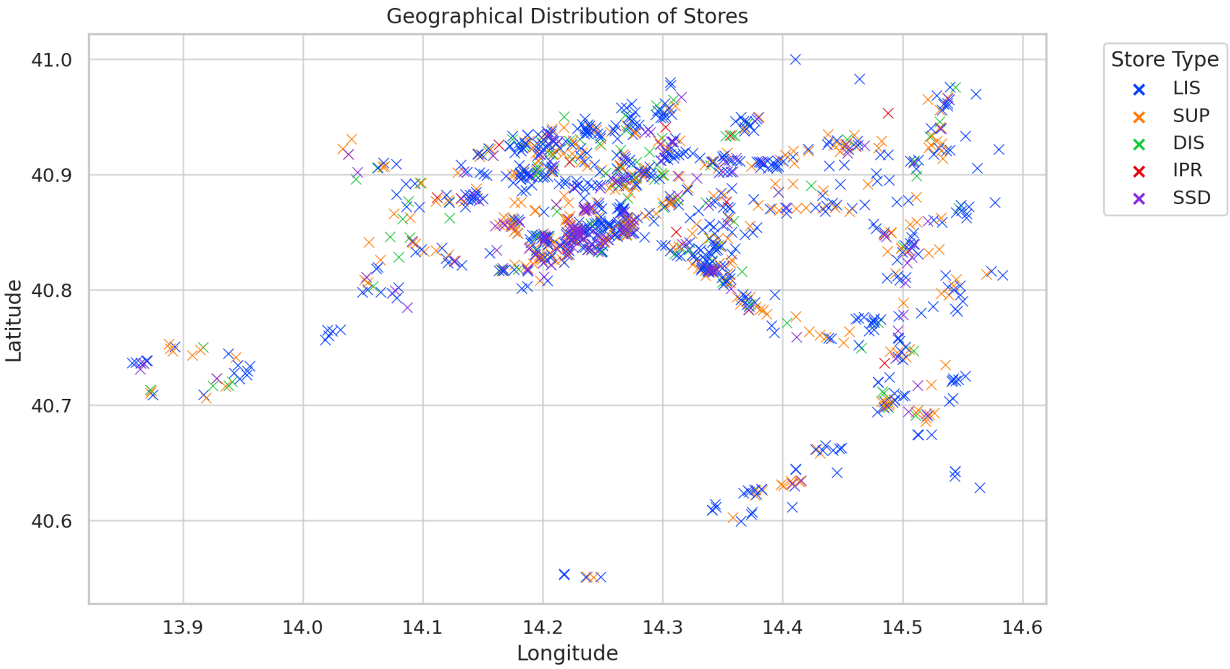


Fig. Geographical Distribution of Stores

The scatter plot of the geographical distribution of stores, categorized by store type, reveals patterns of store locations across the area. The visualization highlights areas with higher concentrations of stores and shows how different types of stores are distributed geographically. This analysis is essential for understanding market saturation, identifying underserved areas, and exploring the strategic placement of stores.

4. Data Modeling

4.1 Train & Test

For a model, as we care about our outliers, and the importance of feature selection, we utilize Random Forest Regression.Random Forest Regression stands out as a formidable machine learning technique, particularly in scenarios where the intricate balance between embracing outliers and the pivotal role of feature selection cannot be overlooked. This ensemble method, characterized by its construction of multiple decision trees during training and outputting the mean prediction of the individual trees, offers a nuanced approach to regression tasks that demand a high degree of accuracy and robustness against overfitting.

In the development of a robust Random Forest Regression model, special attention must be given to how the data is partitioned into training and testing sets. We adhere to a conventional split of 70% of the data for training and 30% for testing.

4.2 Feature Importance

Below are the key features predominantly utilized in our trained model. This report, generated by the random forest algorithm, highlights the primary features significantly contributing to the training process.

```
   Insegna_encoded: 0.5724459054002119
       MQVEND: 0.26499983280785616
TipologiaPdV_encoded: 0.08866885671969398
   Comune_encoded: 0.04076788856033855
total_media_annuale: 0.007655482295197377
   zone_potential: 0.007178245491033629
  population_score: 0.005300018503693835
      Parking: 0.0032769452289461536
     microcode: 0.002936755092131363
    store_count: 0.002640019535546785
 district_encoded: 0.0025725408020342505
 population_class: 0.0015575095633160688
```

5. Evaluating

The training and testing were done with two different levels of rigor, and the findings below are based on both phases.

	0.25	0.5
Main Squared Error	0.0006	0.0006
Root Mean Squared Error	0.025	0.025
R2 Score	0.832	0.832
Accuracy	0.77	0.91

6. Team Acknowledgement

- Parisa Rajaei Nezhad - D03000014
- Mohammad Mehradnia - D03000070
- Mahtab Taheri - D03000055
- Abdolhamid Saadi - D03000026