Städtenamenmuster Deutschland - A Data-Driven Exploration of Naming Conventions

Palwinder Singh
Philipps University
Marburg, Germany
Email: Singhpa@students.uni-marburg.de

***Abstract*—This report examines recurring patterns in place names through substring analysis. The results highlight structural consistencies and regional variations across naming conventions.**

I. Introduction

This report presents a comprehensive substring analysis of place names from the dataset, with a particular focus on identifying frequent linguistic patterns and their structural roles. German place names often contain recurring elements such as "stadt", "dorf", or "hausen" that convey geographic, historical, or cultural significance. By examining substrings of various lengths and their positional preferences within names, we uncover patterns that reflect naming conventions, regional traditions, and linguistic regularities.

The study is organized by substring length, ranging from single letters to longer multi-syllabic segments. Each section provides insights into frequency, position (start or end), and regional usage across federal states. This analysis aims to reveal not only the structure of names but also the deeper semantics behind the formation of place names in German-speaking regions.

Note: All analysis has been performed based on the Gemeniade names in the dataset.

II. Dataset and Tools

A. Dataset

The analysis presented in this report is based on data provided by the **Statistisches Bundesamt** (Federal Statistical Office of Germany). Specifically, the dataset originates from the official *Gemeindeverzeichnis* (municipal directory), with the following characteristics:

- **Dataset Title:** Gemeindeverzeichnis
- **Publisher:** Statistisches Bundesamt
- **Reference Date for Administrative Boundaries:** 09.05.2011
- **Publication Date:** May 2014
- **Population Data Reference:** 09.05.2011
- **Area Data Reference:** 31.12.2010

B. Technology Used

To extract, transform, and analyze substring patterns within place names, the following tools and technologies were utilized:

- **Python:** Used for text processing, substring extraction, and frequency analysis.
- **Power Query:** Employed for preliminary filtering, organization, and manual validation of data.
- **Power BI:** Used to visualize substring distributions across federal states (*Länder*) and districts (*Kreise*) with interactive charts and Maps.

These tools together enabled a multi-level exploration of naming patterns, from character-level analysis to geographic distribution. All analysis has been performed based on the Gemeniade names in the dataset.

III. Common Substring Findings

The substring analysis of Gemeniade names reveals distinct patterns, with particular emphasis on the frequency and distribution of individual characters. These results form the basis for deeper structural analysis of name construction and composition.

A. Analysis of Substrings of Length One

Analyzing single-letter substrings shows which characters appear most frequently in Gemeniade names. Among these, the letter **E** appears the most. It is followed by **R**, making them the top two most used letters in Gemeniade names overall.

Table I shows the frequency of the top single-letter substrings.

TABLE I
Frequency of Single-Letter Substrings

Substring	Count of Occurrences
E	9235
R	6988
N	6956
S	6474
A	6427

1) Positional Substring Preferences: Further analysis shows preferences in the placement of letters at the start and end of names. Most names start with the letter **B**, while **N** is the most common ending letter.

TABLE II
FREQUENCY OF SINGLE-LETTER SUBSTRINGS AT START

Starting Letter	Count of Occurrences
B	1209
S	1149
H	929

TABLE III
FREQUENCY OF SINGLE-LETTER SUBSTRINGS AT END

Ending Letter	Count of Occurrences
N	1100
M	736
H	677

B. Analysis of Umlaut Characters

In the context of German words, specific umlaut characters were analyzed for their frequency. The substring **Ü** was the most frequently occurring, followed closely by **Ö**. The character **ä** occurred less frequently, yet it still appeared a notable number of times.

Table IV presents the frequencies of these special characters.

TABLE IV
FREQUENCY OF GERMAN UMLAUT CHARACTERS

Character	Count of Occurrences
Ü	770
Ö	756
ä	155

1) State-wise Substring Distribution: A geographic breakdown of substring usage reveals that the state of **Rheinland-Pfalz** has the highest count, followed by **Bayern** and **Schleswig-Holstein**. This distribution indicates a regional pattern in name composition and frequency.

TABLE V
SUBSTRING COUNTS BY STATE

Land (State)	Count of Substrings
Rheinland-Pfalz	270
Bayern	265
Schleswig-Holstein	223

C. Analysis of Substrings of Length Two

The analysis of two-letter substrings in the dataset highlights which combinations of letters are most commonly found in the names. The substring **En** appears most frequently, followed by **St** and **Er**. This suggests that these letter pairs are particularly prominent in the names analyzed.

Table VI shows the top three most frequent two-letter substrings.

Fig. 2. State-wise Occurrence of German Characters *ä*, *ö*, and *ü*

TABLE VI
FREQUENCY OF TWO-LETTER SUBSTRINGS

Substring	Count of Occurrences
En	4311
St	3157
Er	2991

1) Positional Preferences for Substrings: Further analysis explores which two-letter substrings are preferred at the start or end of names. At the **start**, the substrings **Sc**, **Ba**, and **Gr** are most common. At the **end**, **dT** is the most frequent, followed by **En** and **Ch**.

TABLE VII
FREQUENCY OF TWO-LETTER SUBSTRINGS AT START

Starting Substring	Count of Occurrences
Sc	422
Ba	293
Gr	291

TABLE VIII
FREQUENCY OF TWO-LETTER SUBSTRINGS AT END

Ending Substring	Count of Occurrences
dT	1925
En	1397
Ch	663

D. Analysis of Substrings of Length Three

The analysis of three-letter substrings reveals significant patterns, particularly around the word **Stadt**, which is a com-

mon component in German place names. The most frequent substrings—**Sta**, **tad**, and **adt**—are all derived from this word.

Table IX shows the most frequent three-letter substrings.

TABLE IX
FREQUENCY OF THREE-LETTER SUBSTRINGS

Substring	Count of Occurrences
Sta	1921
tad	1855
adt	1849

1) State-wise Positional Substring Distribution: Substring positions provide further insight. At the **start** of words, the substrings **Sch**, **Neu**, and **Obe** are the most frequent. These likely originate from words such as *Schule*, *Neuburg*, or *Oberdorf*.

TABLE X
THREE-LETTER SUBSTRINGS AT WORD START

Starting Substring	Count of Occurrences
Sch	422
Neu	191
Obe	186

At the **end** of words, the substring **adt** remains dominant, again pointing to the widespread occurrence of the word **Stadt**. The substrings **Orf** and **Ach** also appear frequently, likely reflecting town name endings like *Dorf* or *Bach*.

TABLE XI
THREE-LETTER SUBSTRINGS AT WORD END

Ending Substring	Count of Occurrences
adt	1732
Orf	629
Ach	607

E. Analysis of Substrings of Length Four

The substring analysis of four-letter segments continues to show the dominance of fragments derived from the word **Stadt**, such as **Stad** and **tadt**, which still lead the chart. However, this level of granularity also reveals the presence of other meaningful words commonly found in German place names, such as **Bach**, **Dorf**, and **Berg**.

Table XII lists the most frequent four-letter substrings.

TABLE XII
FREQUENCY OF FOUR-LETTER SUBSTRINGS

Substring	Count of Occurrences
Stad	1855
tadt	1847
Bach	717
Dorf	715
nGen	609
Berg	584

1) Start and End Patterns in Four-Letter Substrings: At the **start** of names, substrings such as **Ober**, **Groß**, and **Schw** are frequently observed. These are typical prefixes in German place names, pointing to historical or geographical descriptors (e.g., *Oberhausen*, *Großröhrsdorf*, *Schwarzenberg*).

TABLE XIII
FOUR-LETTER SUBSTRINGS AT WORD START

Starting Substring	Count of Occurrences
Ober	186
Groß	143
Schw	106

At the **end** of names, the substring **tadt** continues its dominance, again reflecting the influence of "Stadt". However, meaningful geographic terms like **Dorf** and **Bach** also emerge with high frequency.

TABLE XIV
FOUR-LETTER SUBSTRINGS AT WORD END

Ending Substring	Count of Occurrences
tadt	1730
Dorf	570
Bach	526

Notably, **570 Gemeniade names end with the substring Dorf**. In a subsequent section, we will explore the geospatial distribution of such names to uncover potential regional naming trends

F. Analysis of Substrings of Length Five

The five-letter substring analysis confirms the continued dominance of **Stadt**, which appears **1847 times** in the dataset. Additionally, two new meaningful substrings—**ausen** and **Hause**—emerge. These substrings are often found within compound words like *Hausen*, a common component of German place names.

Table XV presents the most frequent five-letter substrings.

TABLE XV
FREQUENCY OF FIVE-LETTER SUBSTRINGS

Substring	Count of Occurrences
Stadt	1847
Ingen	499
ausen	375
Hause	371

1) Five-Letter Substrings at Start and End Positions: At the **start** of names, the substrings **Gross**, **Niede**, **Stein**, and **Schön** are frequently found. These are common prefixes in German locations, reflecting size, geographic placement, or descriptive elements.

The analysis of name endings reveals that **Stadt** continues to be a dominant pattern. Additionally, substrings such as **Ingen**, **ausen**, **sdorf**, and **ndorf** further support the trend of common suffixes in German town names.

TABLE XVI
FIVE-LETTER SUBSTRINGS AT WORD START

Starting Substring	Count of Occurrences
Gross	143
Niede	102
Stein	69
Schön	67

TABLE XVII
FIVE-LETTER SUBSTRINGS AT WORD END

Ending Substring	Count of Occurrences
Stadt	1730
Ingen	323
ausen	284
sdorf	250
ndorf	218

2) Tree Map of Substring Composition: The tree map in Fig. 3 provides a visual summary of the relative proportions of common five-letter substrings, illustrating the prominence of patterns like *Stadt*, *Hausen*, and *Dorf*-related suffixes.

Fig. 3. Tree Map Showing Frequency of Common Five-Letter Substrings

G. Analysis of Substrings of Length Six

At six letters, full place-name components begin to emerge. Substrings like **Hausen**, **Endorf**, and **enbach** are clearly identifiable and commonly used in German locality names. These units reflect meaningful linguistic and geographic patterns.

Table XVIII presents the most frequent six-letter substrings.

TABLE XVIII
FREQUENCY OF SIX-LETTER SUBSTRINGS

Substring	Count of Occurrences
Hausen	370
Endorf	221
enbach	208

1) Six-Letter Substrings at Word Start: At the beginning of names, the most frequent substrings include **Nieder**, **Langen**, and **Königs**, all of which are typical German place-name prefixes reflecting geography or heritage.

2) Six-Letter Substrings at Word End: A detailed look at word endings reveals that **Hausen** appears **281 times** at the end, despite its total frequency being 370. This indicates that while it is primarily a suffix, it can also appear within compound names or mid-word. Similarly, **Endorf** and **enbach** are often terminal elements.

TABLE XIX
SIX-LETTER SUBSTRINGS AT WORD START

Starting Substring	Count of Occurrences
Nieder	100
Langen	39
Königs	29

TABLE XX
SIX-LETTER SUBSTRINGS AT WORD END

Ending Substring	Count of Occurrences
Hausen	281
Endorf	184
enbach	162

H. Analysis of Substrings of Length Seven

At the seven-letter level, entire place-name units are clearly visible. The substring **Kirchen** is the most frequent, with 121 occurrences. This is a strong indication of the religious or historical relevance embedded in many German place names.

The analysis also shows substrings like **Schwarz**, **shausen**, and **ersdorf**, all of which reflect common components of German geography-based naming conventions. Table XXI presents the frequency of the top seven-letter substrings.

TABLE XXI
FREQUENCY OF SEVEN-LETTER SUBSTRINGS

Substring	Count of Occurrences
Kirchen	121

In terms of positioning within names, **Schwarz** is observed both at the start (22 occurrences) and more frequently at the end (40 occurrences), suggesting its versatility in compound names. Notably, **shausen**, **ersdorf**, and **Kirchen** commonly appear as suffixes, reinforcing their role as terminal naming elements in German place names.

TABLE XXII
SEVEN-LETTER SUBSTRINGS AT WORD END

Ending Substring	Count of Occurrences
Schwarz	40
shausen	99
ersdorf	79
Kirchen	76

I. Analysis of Longer Substrings

While shorter substrings highlight structural units like prefixes, suffixes, and roots, longer substrings in the dataset reflect complete compound names that are semantically rich and often represent official or administrative place types. Names like **Kreisstadt**, **Hansestadt**, and **Landeshauptstadt** demonstrate the compound nature of German toponymy, combining functional or historical roles with geographic identifiers. These extended substrings, though less frequent, offer high semantic density and are often tied to formal naming conventions in German municipalities.

TABLE XXIII
LONG SUBSTRINGS WITH THEIR LENGTHS

Substring	Count of Occurrences	Length
Neustadt	26	8
Ershausen	43	9
Kreisstadt	17	11
Hansestadt	16	11
Neukirchen	15	10
Schwarzwald	16	12
Neunkirchen	10	11
Reichenbach	12	12
Landeshauptstadt	8	17
Universitätsstadt	6	18

IV. HISTORICAL-LINGUISTIC ORIGIN OF COMMON SUBSTRINGS

The analysis revealed that many Gemeniade names are composed of meaningful linguistic units commonly found in German place names. Substrings such as Stadt, Dorf, Bach, Berg, Hausen, Endorf, Ingen, and Kirchen appear frequently and reflect real-world naming conventions. Prefixes like Gross, Nieder, Langen, Königs, Neu, Ober, Schön, and Schwarz often denote geographical or hierarchical attributes. Similarly, suffixes such as dorf, hausen, bach, stadt, ingen, kirchen, sdorf, and Universitätsstadt suggest settlement types or landscape features. These patterns indicate that the name constructions are not random but instead follow established linguistic structures typical in German toponymy.

A. Universitätsstadt

The substring Universitätsstadt appears six times within the dataset and is associated with cities known for their academic significance. All of these cities are located in the southern part of Germany, specifically within the federal states of Baden-Württemberg and Hessen. This geographic concentration suggests a regional naming convention related to the presence of universities.

TABLE XXIV
OCCURRENCES OF UNIVERSITÄTSSTADT BY STATE AND DISTRICT

State (Land)	District (Kreis Name)
Baden-Württemberg	Konstanz
Baden-Württemberg	Mannheim
Baden-Württemberg	Tübingen
Baden-Württemberg	Ulm
Hessen	Gießen
Hessen	Marburg-Biedenkopf

B. Dorf

The substring **Dorf**, meaning 'village', is a common element in German place names, reflecting the country's rural and agricultural history. It appears most frequently in **Bayern**, followed by **Schleswig-Holstein** and **Thüringen**. Notably, **Rendsburg-Eckernförde** in Schleswig-Holstein and **Eifelkreis Bitburg-Prüm** in Rheinland-Pfalz show particularly high concentrations, indicating strong regional naming patterns tied to traditional settlement types.

Fig. 4. Geographic distribution of Universitätsstadt substrings in southern Germany.

TABLE XXV
OCCURRENCES OF DORF BY STATE

State (Land)	Count of Substrings
Bayern	130
Schleswig-Holstein	124
Thüringen	95

C. Stadt

The substring **Stadt**, meaning "city" or "town," is widely used in German place names, often denoting urban status or municipal rights. The dataset reveals that **Stadt** appears most frequently in the southern and western regions of Germany. The highest number of occurrences is found in **Baden-Württemberg**, followed by **Nordrhein-Westfalen**, **Hessen**, **Sachsen**, and **Niedersachsen**. This widespread usage reflects both historical city privileges and modern administrative classifications.

TABLE XXVI
OCCURRENCES OF STADT BY STATE

State (Land)	Count of Substrings
Baden-Württemberg	323
Nordrhein-Westfalen	269
Hessen	192
Sachsen	176
Niedersachsen	165

D. Hausen

The substring **Hausen** is a frequent element in German place names, typically derived from 'Haus' (house) and historically associated with homesteads or settlements. It appears prominently in central and southern Germany, indicating long-standing patterns of residential or agricultural development.

Fig. 5. Geographic distribution of Dorf substrings across German federal states.

Fig. 6. Geographic distribution of Stadt substrings across German federal states.

Significantly, the district of **Westerwaldkreis** in Rheinland-Pfalz records the highest number of substring appearances among all districts in the dataset, underscoring its importance in the toponymic landscape. This suggests a strong historical presence of dispersed housing or small residential communities in this region.

E. Kirchen

The substring **Kirchen**, meaning "churches," is commonly found in German place names and often signifies historical or regional religious significance, particularly the presence of churches or ecclesiastical centers.

TABLE XXVII
OCCURRENCES OF HAUSEN BY STATE

State (Land)	Count of Substrings
Rheinland-Pfalz	101
Bayern	75
Thüringen	66
Baden-Württemberg	55
Niedersachsen	31

Fig. 7. Geographic distribution of Hausen substrings across German federal states.

A notable concentration is seen in the Bavarian district of **Mühldorf**, which accounts for 8 occurrences more than any other single district. This suggests a strong historical influence of religious institutions in shaping local place names within this region. The broader distribution of the substring across both southern and northern Germany reflects its consistent use in ecclesiastically influenced naming traditions.

TABLE XXVIII
OCCURRENCES OF KIRCHEN BY STATE

State (Land)	Count of Substrings
Bayern	55
Rheinland-Pfalz	16
Nordrhein-Westfalen	11

V. FURTHER WORK

This analysis provides a strong foundation for understanding the structural and linguistic patterns in German place names. However, many more insights could be uncovered through deeper investigation, particularly at the regional level.

One promising direction is a **hierarchical analysis at the Kreis (district) level**, which would allow for more localized pattern detection. While this report has mostly focused on

Fig. 8. Geographic distribution of Kirchen substrings across German federal states.

substring frequencies across all data or at the state level, district-level analysis could reveal:

- **Localized Naming Trends:** Certain substrings like *Dorf*, *Hausen*, or *Kirchen* may occur more frequently in specific districts, indicating regional linguistic or historical influences.
- **Comparison Between Districts:** Students can examine how naming patterns differ between neighboring districts within the same state, helping to identify regional naming boundaries or cultural influences.
- **Prefix and Suffix Patterns by Region:** Investigating which name elements tend to appear at the start or end of place names within each district could offer insights into formal naming conventions or geographic descriptors.
- **Geographic Concentration of Substrings:** By mapping where certain substrings are most concentrated, one can hypothesize connections to settlement history, migration, or administrative practices.

All of these analyses can be supported and made more efficient through the use of tools like **Power BI**, which allows filtering, slicing, and visualizing the data at various levels. However, the conceptual direction of this further work remains student-driven, focusing on discovering linguistic and cultural patterns through structured exploration of geographic name data.