A Data-Driven Guide to Berlin’s Neighborhoods for University Students

Why Berlin?

- Berlin is an international destination for both domestic and international students
  - Diverse culture
  - Lively social scenes
  - World-renowned universities & research institutions
  - A great selection of post-graduation opportunities
- Relocating to and settling in a new city is nerve-wracking, especially for young adults.
  - There is business potential behind developing a data product that facilitates the process of deciding where to live.

Source: Der Tagesspiegel

Business Problem:

Which neighborhoods in Berlin are most appealing to university students? How can we characterize the experiences of living in one of these neighborhoods?

Potential Stakeholders:

- **Prospective University Students:** Wants to make an informed decision when searching for apartments. Wants to find a place that suits their lifestyle choices and feels like home.
- **The City of Berlin:** Wants to brand Berlin as a student city to bring in more revenue and an educated workforce. Wants to know which neighborhoods to promote.
- **Commercial & residential real estate companies:** Wants to invest in and market areas that will appeal to students

Project Goals:

1) Develop evaluation criteria to select the most suitable neighborhoods for university students
2) Segment these neighborhoods into clusters and describe their appeal to a unique client profile

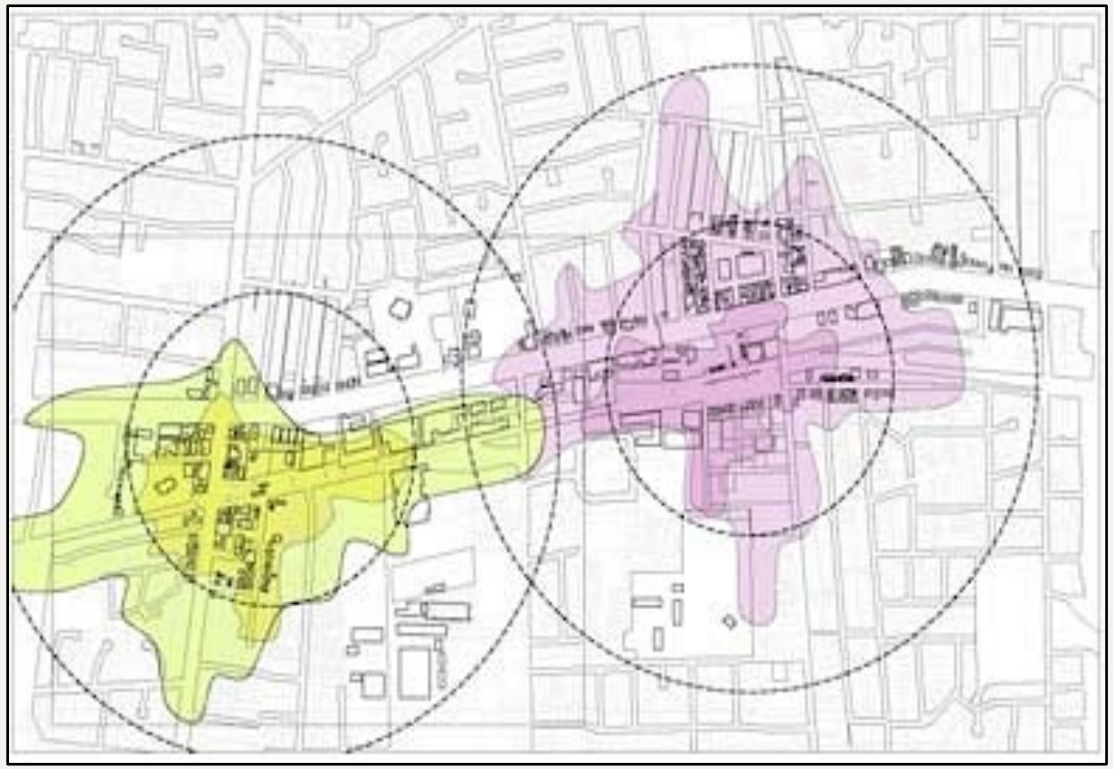
Data Sources

- Berlin’s 96 localities (Ortsteile) CSV file and GeoJSON for mapping
- CSV file with Berlin universities and their coordinates
- CSV file with S-Bahn (light rail) and U-Bahn (metro) stations and their coordinates used for clustering.
- CSV file with 2019 demographics by locality
- 2019 rent prices by locality webscraped from HomeDay
- Venue data obtained using Foursquare API

Note: Spatial data were geocoded using geopy’s Nominatim, processed using pyproj, haversine, and geopandas, and mapped with Folium.

Map of Berlin divided into 12 boroughs and 96 localities

How is a neighborhood defined in this project?

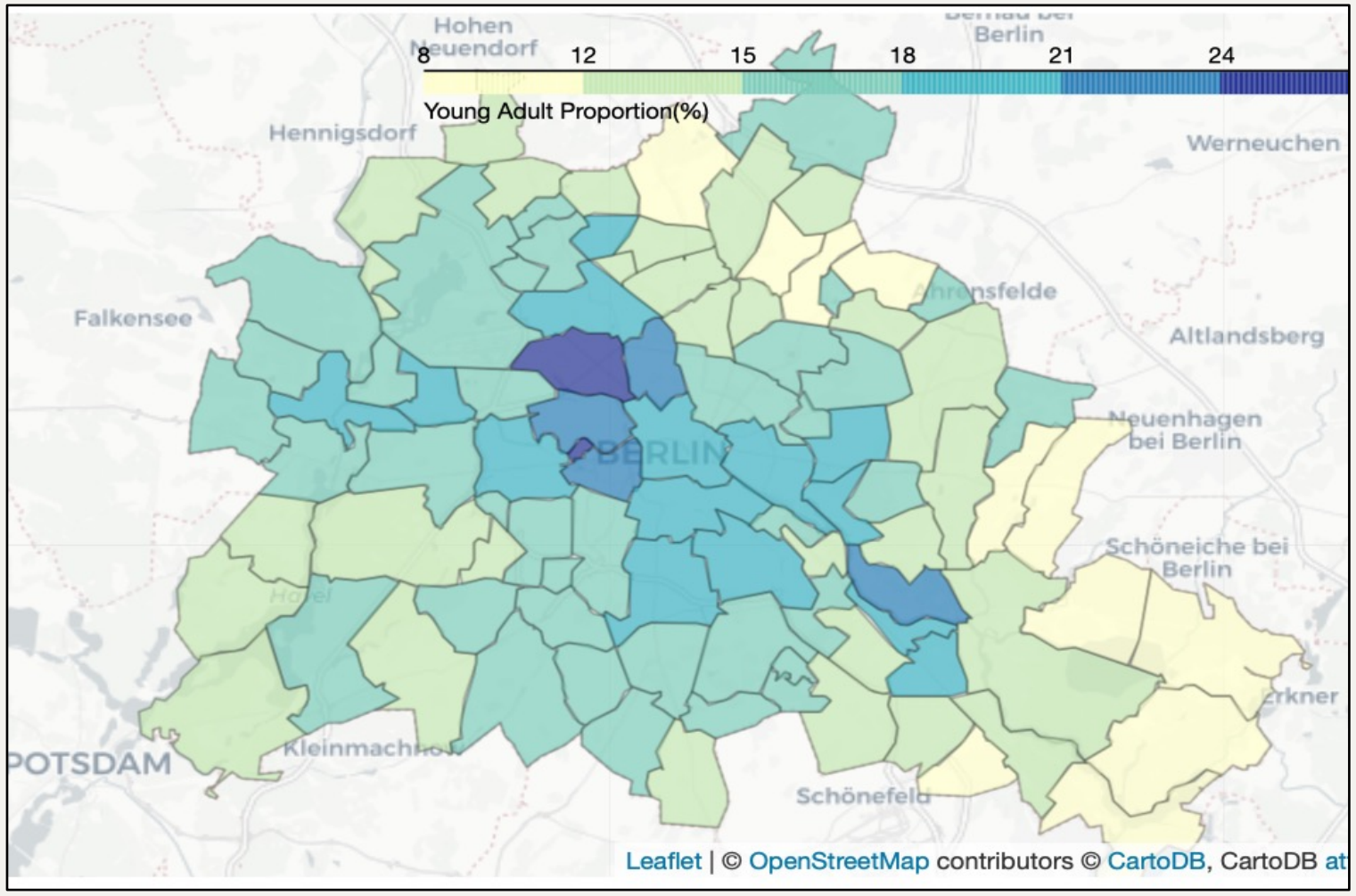


Source: Marcus White

- A neighborhood is not defined by localities in Berlin.
- We examine neighborhoods at a more micro-level, or the 700-meters walking catchment around each train station.
- Venue data will be collected using this 700-meters radius.
- This better resembles how people interact with their environment on a daily basis and allows for more variety in our findings.

1st Criterion: Proportion of Young Adults

- Raw CSV file was cleaned using pandas after referencing the metadata.
- Measure: Number of 15- to 30-year-olds divided by the total population of that locality in 2019
- Selection criterion: top 50% of localities with the highest young adult proportion
- Localities in central Berlin tend to have a higher proportion of young adults compared to localities on the periphery.

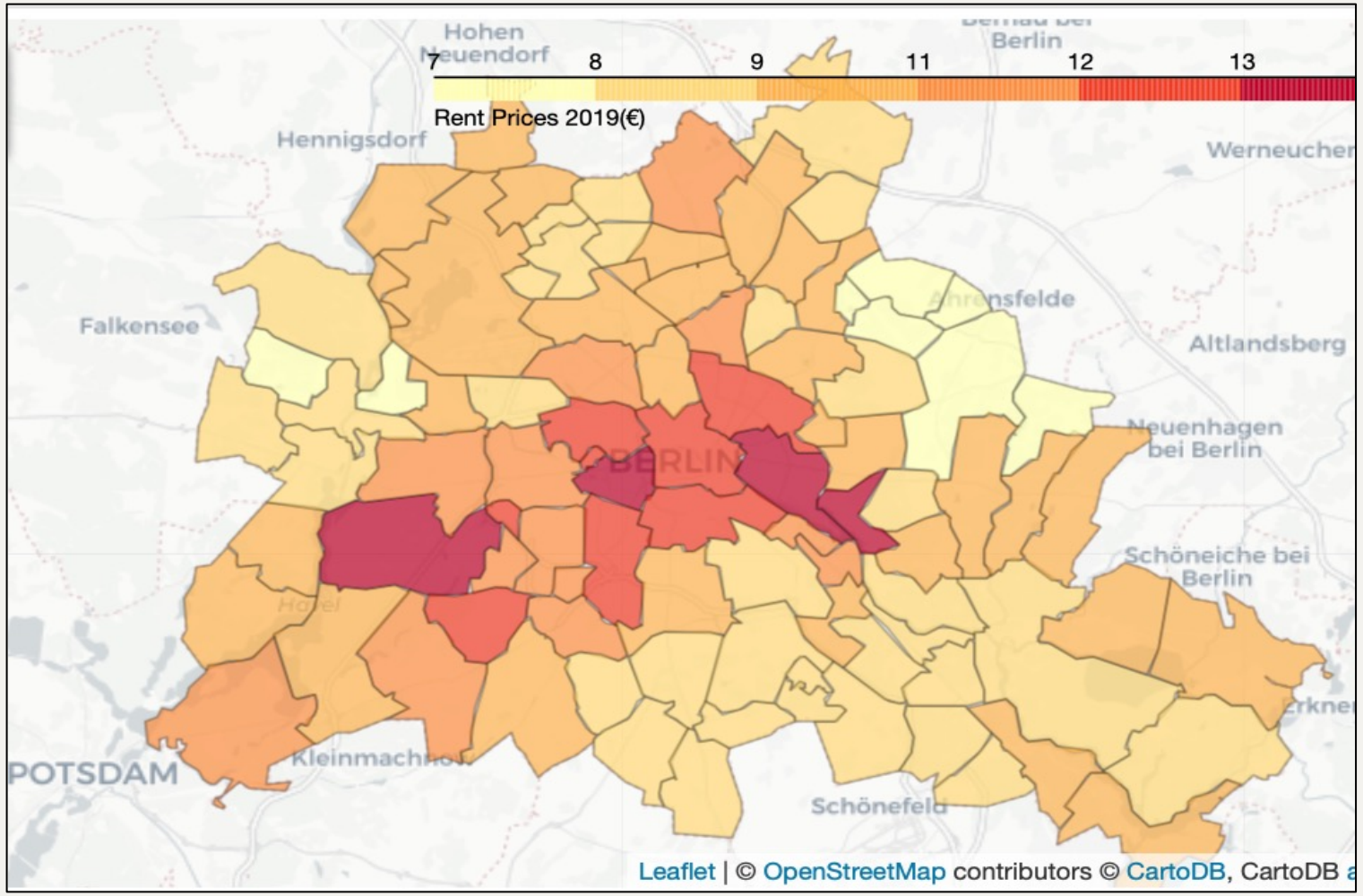


Proportion of Young Adults in 2019 by Locality

2nd Criterion: Housing Expenses

- Rent prices were webscraped from HomeDay using BeautifulSoup.
- Measure: 2019 rent prices (€/m^2) by locality
- Selection criterion: all localities minus the top 20% with extremely high rent prices

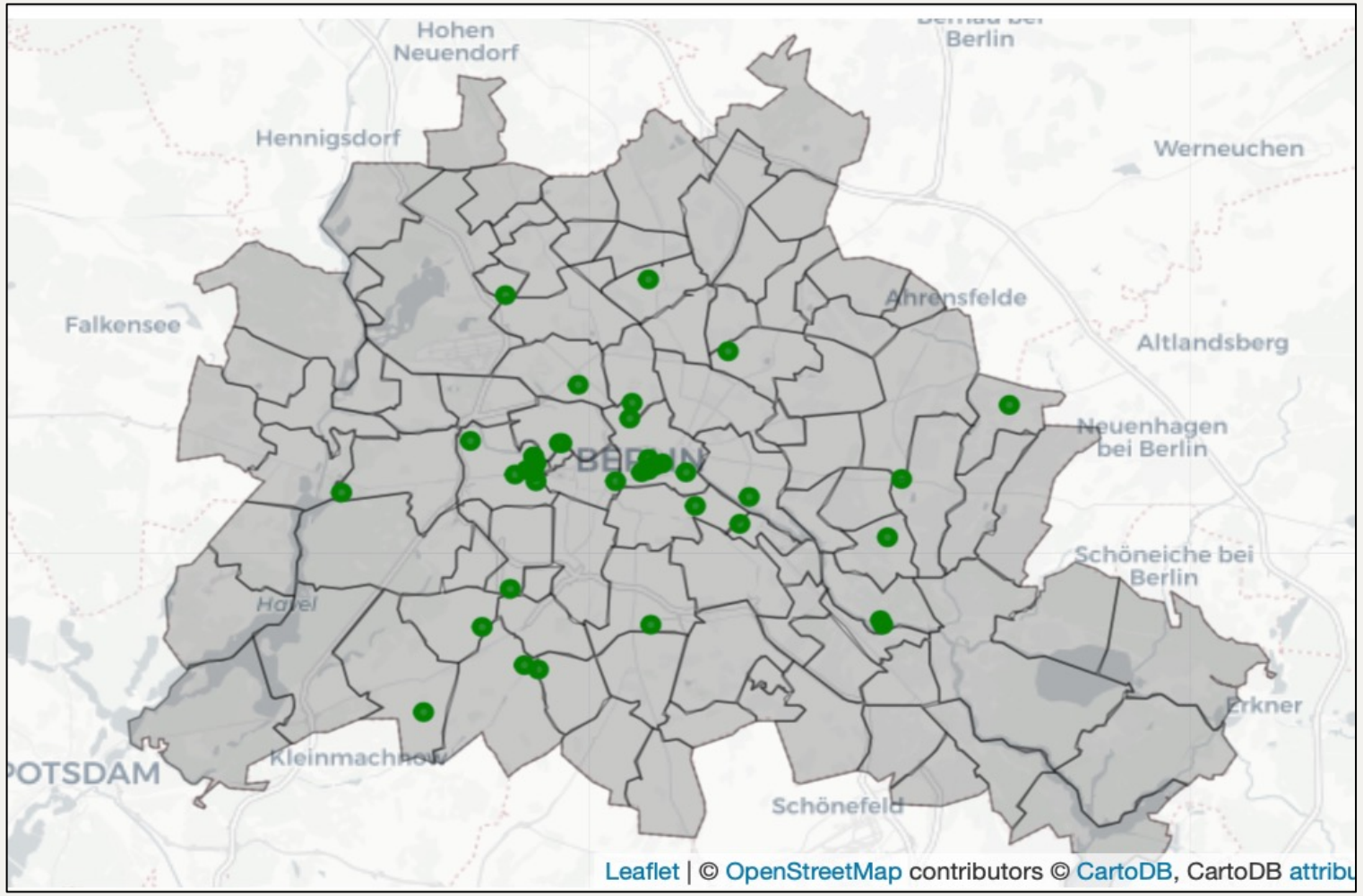
- Most expensive localities are unsurprisingly the most popular places to live such as Tiergarten, Prenzlauer Berg, and Friedrichshain. Cheaper localities are on the outer edges, especially on the Eastern side.



2019 Rent Prices by Locality

3rd Criterion: Commute Distance to Universities

- From the CSV file, localities' center points were geocoded using geopy.
- Coordinates of 39 Berlin universities were manually geocoded using Google Maps.
- Haversine distance from each locality to each university was calculated. 8.36 km was designated as the reasonable commute distance.
- Measure: Count of universities one would be able to reach from each locality within the 8.36-km radius
- Selection criterion: top 50% of localities with the highest count



Universities (Universitäten and Hochschulen) in Berlin

	Locality Name	Count	Rent Prices 2019(€)	Young Adult Proportion(%)
0	Gesundbrunnen	27	10.0	24.07
1	Charlottenburg	27	11.1	18.02
2	Tempelhof	26	10.2	18.27
3	Wedding	25	11.0	24.66
4	Charlottenburg-Nord	24	8.8	16.91
5	Reinickendorf	24	9.5	19.73
6	Neukölln	23	8.9	20.23
7	Steglitz	23	10.7	16.59
8	Fennpfuhl	17	9.5	17.88
9	Britz	16	9.3	17.51
10	Westend	16	11.0	17.13
11	Borsigwalde	14	9.2	17.91
12	Haselhorst	13	8.3	16.31
13	Baumschulenweg	13	9.6	15.58
14	Lichtenberg	13	10.0	20.39
15	Siemensstadt	13	10.0	19.12
16	Mariendorf	12	9.4	16.18
17	Friedrichsfelde	11	8.4	17.00
18	Wittenau	10	9.4	16.10
19	Niederschöneweide	8	9.3	20.33

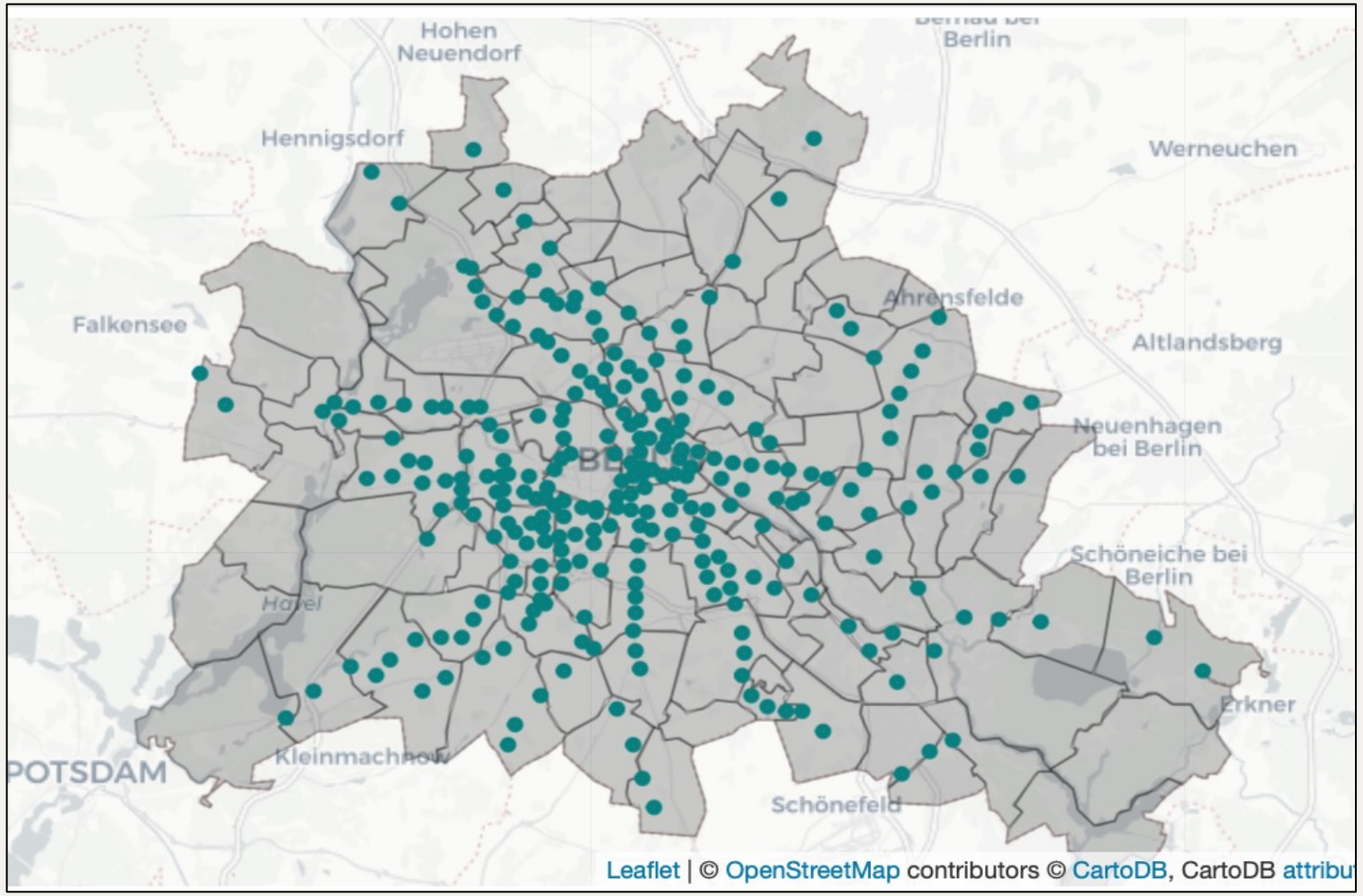
20 Most Suitable Localities for University Students

- Comparing the 48 localities from the 1st criterion, 77 from the 2nd, and 48 from the 3rd resulted in 20 matches.
- Highest university count: Gesundbrunnen
- Lowest rent prices in 2019: Haselhorst
- Highest young adult proportion in 2019: Wedding

- Note: Only 19 localities remained when factoring in the venue data. Borsigwalde was dropped due to the limited numbers of venues returned for train stations in this locality.

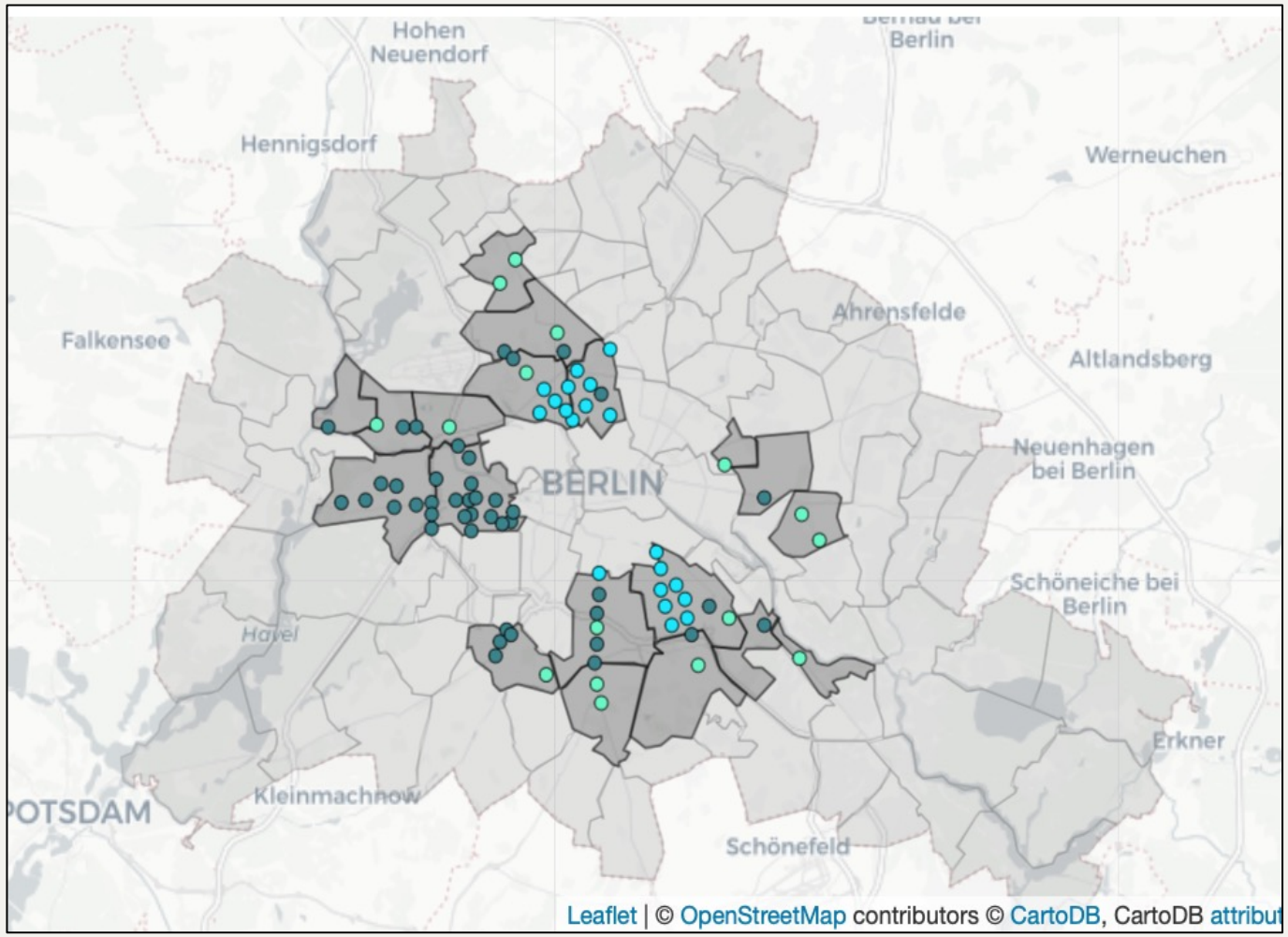
Getting Public Transport Stations

- Raw CSV file was cleaned using pandas.
- Coordinates transformed from Gauß-Krüger Zone 4 to WGS84 Decimal Degrees.
- Stations beyond the boundaries of Berlin, duplicated stations, bus & tram stops were dropped → 277 stations remaining
- Only stations that were in the 20 most suitable localities were subsetted for clustering → 97 stations remaining
- Stations = origin points for Foursquare queries



All S- and U-Bahn Stations in Berlin

Clustering Venue Data



Final Clusters (Light Blue: Cluster 0, Dark Blue: Cluster 1, Light Green: Cluster 2)

- All venues within a 700-meters radius from each station were obtained using the Foursquare API.
- Out of 3,963 venues, 195 that were categorized as transportation-related or outdoor objects were dropped.
- Exploratory data analysis was performed using Foursquare's metadata on venue categories.
  - 52.75% of venues were food-related, 19.55% were shops & services, 8.60% were nightlife spots.
- Stations with less than 10 venues were dropped → 79 stations remaining
- After one-hot encoding and aggregating, K-means was used to produce the three clusters.

Cluster 0

Source: Lonely Planet

- 20 stations in total

  + 8 in Neukölln, 6 in Wedding, 5 in Gesundbrunnen, 1 in Tempelhof

- Average of 63 venues within 700 meters of each station
- 53.6% of venues were food-related, 17.6% were nightlife spots, and 12.8% were shops & services
- 1st most common venue was cafes, 2nd most common was bars, 3rd most common was pubs
- Perfect for students who thrive in a lively environment and love going out

Cluster 1

Source: The Travelerr

- 43 stations in total, highest of all the clusters
  - + 17 in Charlottenburg, 7 in Westend, 4 in Tempelhof, 4 in Steglitz, 2 in Reinickendorf, 2 in Siemensstadt, and 1 each in Lichtenberg, Haselhorst, Neukölln, Gesundbrunnen, Baumschulenweg, Wedding, and Britz.
- Average of 49 venues within 700 meters of each station
- 55.1% of venues were food-related, 20.6% were shops & services, and 7.1% were outdoors & recreation.
- 1st most common venue was supermarkets, 2nd most common was Italian restaurants, 3rd most common was cafes
- Perfect for students who seek comfort and convenience and doesn’t put as much priority on the nightlife scene

Cluster 2

Source: Nomad and In Love

- 16 stations in total
  + 2 in Wittenau, 2 in Friedrichsfelde, 2 in Mariendorf, and 1 each in Fennpfuhl, Tempelhof, Charlottenburg-Nord, Reinickendorf, Niederschöneweide, Steglitz, Siemensstadt, Neukölln, Wedding, and Britz
- Average of 17 venues within 700 meters of each station
- 39.6% of venues were shops & services, 35.2% were food-related, and 13.9% were outdoors & recreation.
- 1st most common venue was supermarkets, 2nd most common was parks, 3rd most common was drugstores
- Perfect for students who prefer quieter neighborhoods and value their privacy

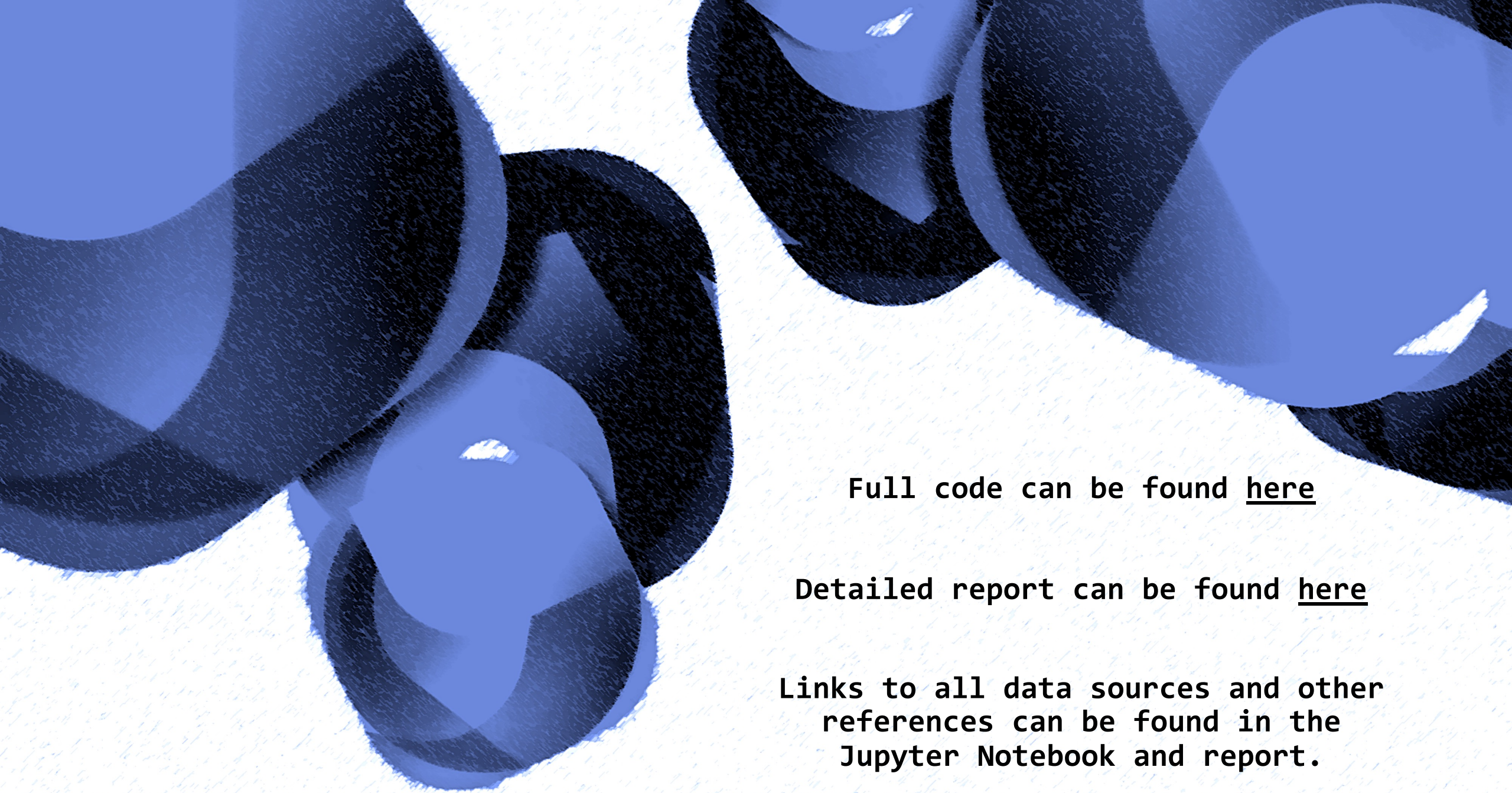
Future Directions & Conclusion

Potential improvements:

+ Incorporate additional criteria when selecting the most suitable localities such as crime rates, job opportunities, price range of surrounding venues, and socioeconomic characteristics of potential neighbors
+ Beyond just rent prices, approach the issue of housing by looking at shared apartments and student accommodations
+ Include bus and tram stops to expand the coverage to suburban localities
+ Optimize the number of clusters used for K-means via the elbow method or silhouette score

- Another possible data product is a classification model that takes user inputs. Users could select a specific venue category or university and get recommended a cluster label

- **Project goals were met. Stakeholders could apply a similar methodology to find the best neighborhoods for themselves or their clients.**
- **Workflow is easily generalizable to other location problems such as siting a youth hostel**



Full code can be found here

Detailed report can be found here

Links to all data sources and other references can be found in the Jupyter Notebook and report.