

Where to live in Berlin?

Coursera Capstone Project

Business Problem

- We are working for a real estate agency as a Data scientist.
- The agency finds properties for private individuals and public companies according to their needs.
- This time, a young professional is searching for a flat. He just finished his college degree and wants an area, which has a lot of cafés, bars and clubs.
- There is a budget constraint: the prices should be moderate at maximum.

Data Description I

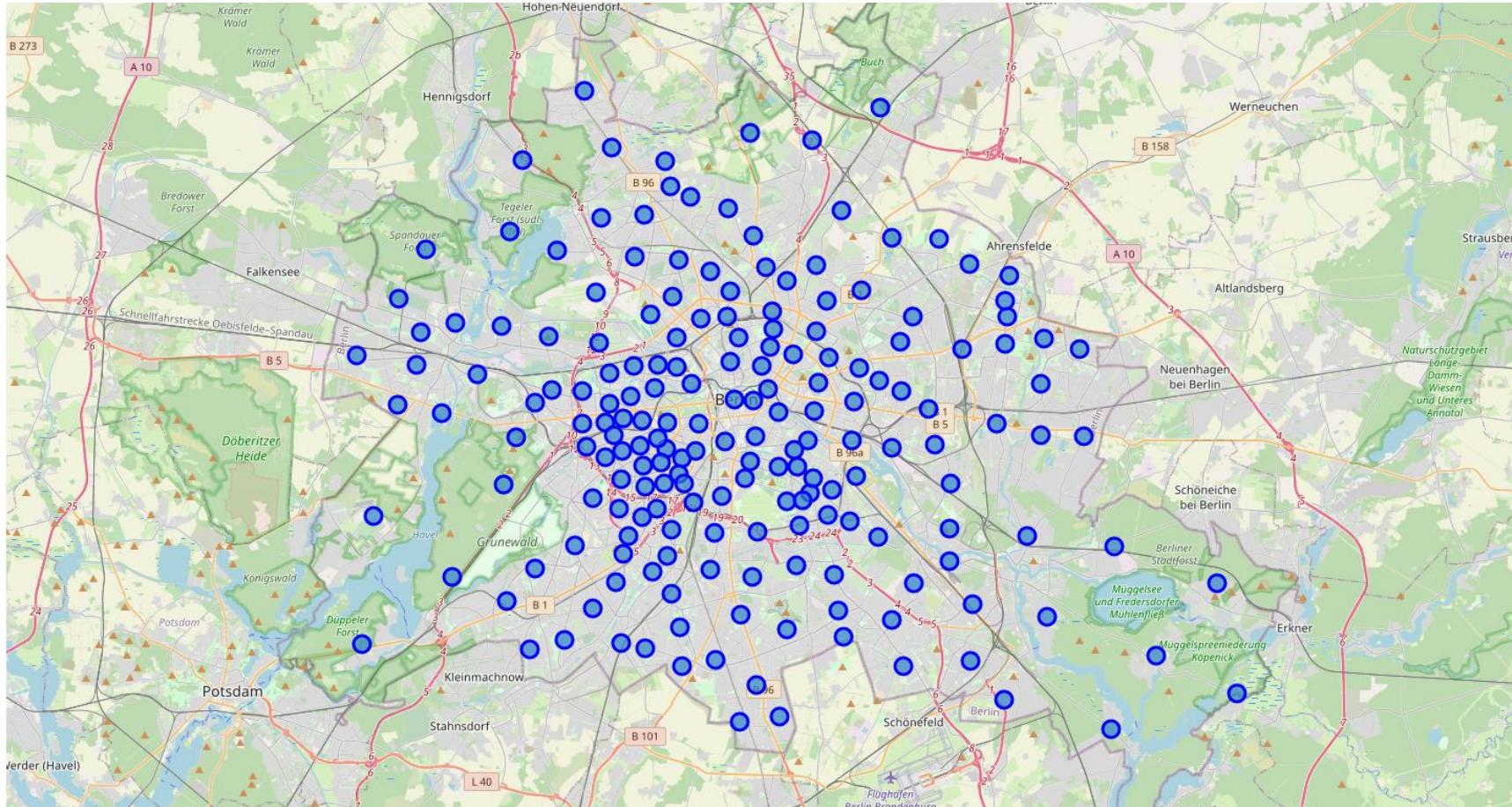
- The zip codes of Berlin are supplied by Simon Franzen (available via GitHub).
- The venue data is retrieved from the Foursquare API (number is limited by 100 venues per request).
- A map is drawn with the Folium package.
- The rental data is taken from miet-check.de for each zip code by hand.

Data Description II

The median price is 14.8€ per square meter. The maximum is around 27€ and the lowest around 7€. Most prices are between 11.5€ and 19€.

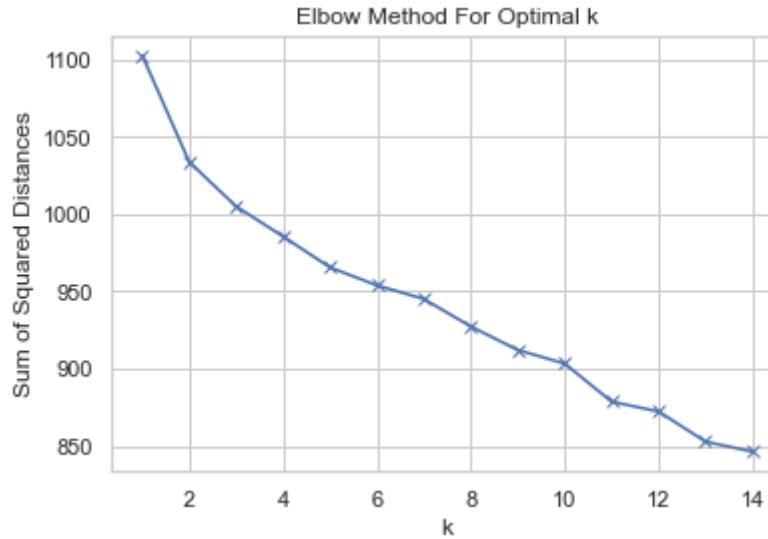


Map of Berlin – All zip codes

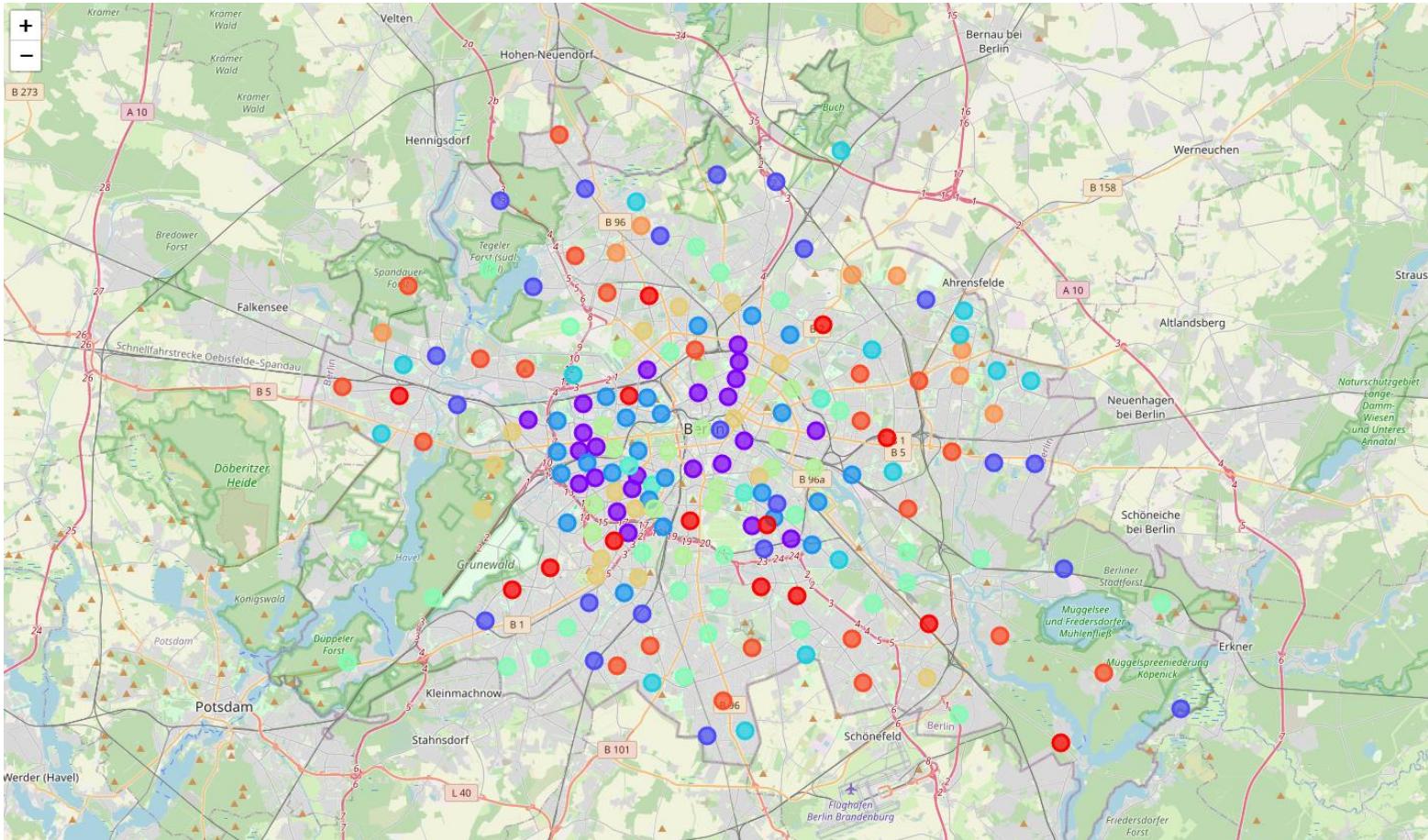


Cluster Analysis I

- As the clustering algorithm, k-means is used. The venue data, as well as the rental data, are our input variables.
- **Our goal is to have fine segments.**
- Deducted from the plot, five or eleven segments are appropriate to choose. For the following, $k=11$ is applied.



Map of Berlin – Clustered Areas



Cluster Analysis II: Chosen segments

Cluster 1

```
1
Café          7
Hotel         5
Italian Restaurant 3
German Restaurant 2
Bakery        2
Coffee Shop   2
Supermarket   1
Bar           1
Name: 1st Most Common Venue, dtype: int64
Café          5
Italian Restaurant 3
Vietnamese Restaurant 2
Pizza Place   2
Bar           2
Supermarket   2
Nightclub     1
Art Gallery   1
Gas Station   1
Coffee Shop   1
Bistro        1
German Restaurant 1
Cocktail Bar  1
Name: 2nd Most Common Venue, dtype: int64
```

Cluster 3

```
3
Café          6
Hotel         4
Supermarket   3
Bar           3
Italian Restaurant 2
Bus Stop      1
Park          1
Vietnamese Restaurant 1
Coffee Shop   1
Bistro        1
Plaza         1
Name: 1st Most Common Venue, dtype: int64
Café          5
Italian Restaurant 4
Coffee Shop   2
Bakery        1
Vietnamese Restaurant 1
Hotel         1
Bus Stop      1
Music Venue   1
Plaza         1
Zoo Exhibit   1
Pizza Place   1
Trattoria/Osteria 1
Bar           1
Gym / Fitness Center 1
Cocktail Bar  1
Ice Cream Shop 1
Name: 2nd Most Common Venue, dtype: int64
```

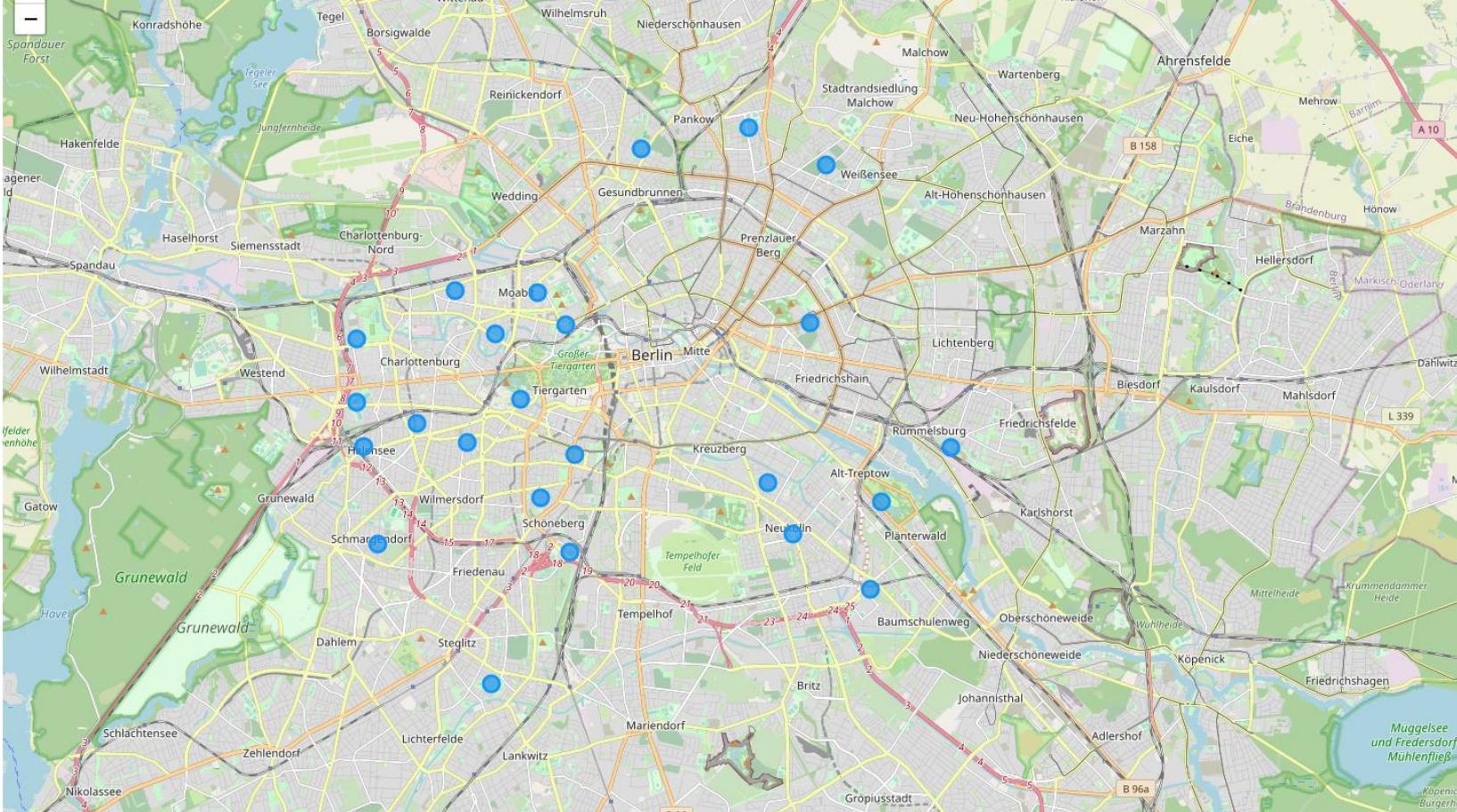
Cluster 7

```
7
Hotel         5
Café          4
Supermarket   2
Italian Restaurant 2
Bar           2
Burger Joint  1
Coffee Shop   1
Name: 1st Most Common Venue, dtype: int64
Nightclub     2
Bar           2
Supermarket   1
Vietnamese Restaurant 1
Park          1
Chinese Restaurant 1
Hotel Bar     1
Coffee Shop   1
Soccer Field  1
Zoo Exhibit   1
German Restaurant 1
Bistro        1
Italian Restaurant 1
Café          1
Bakery        1
Name: 2nd Most Common Venue, dtype: int64
```

Cluster Analysis II: Chosen segments

- After choosing, we are applying the lowest possible budget constraint on these clusters which is 20€ per square meter.
- With the <20€ budget constraint, every segment but 3 is dropped.
- According to the box plot, 20€ is between 50% and 75% of the price range which can be considered as a moderate pricing range.

Map of Berlin – Chosen segment (3)



Conclusion

As a data scientist, we are suggesting the following zip code areas to search:

- Prenzlauer Berg (10249)
- Rummelsburg (10317)
- Moabit (10553, 10555, 10557 and 10559)
- Charlottenburg (10629, 10711, 10719, 10787, 10787, 14057 and 14059)
- Schöneberg (10783, 10823 and 10829)
- Neukölln (12043, 12047 and 12057)
- Steglitz (12167)
- Alt-Treptow (12435)
- Weißensee (13086)
- Pankow (13189)
- Wedding (13359)
- Schmargendorf (14199)