

# IBM Coursera Data Science Capstone

Optimal Location for a Hookah Bar in Austin

Gabriel Aloba



# Austin Texas is ideal for situating a niche business with high profit margins

---

The City has a young population with high disposable income

It would be a great place to establish a new hookah bar

- There is an influx of talent
  - Because of presence of large companies
  - Lower tax burden and living cost than coastal tech hubs
- A hookah bar is a great place for a young multicultural crowd to mingle

With the right strategy in choosing a location, a new bar is bound to succeed

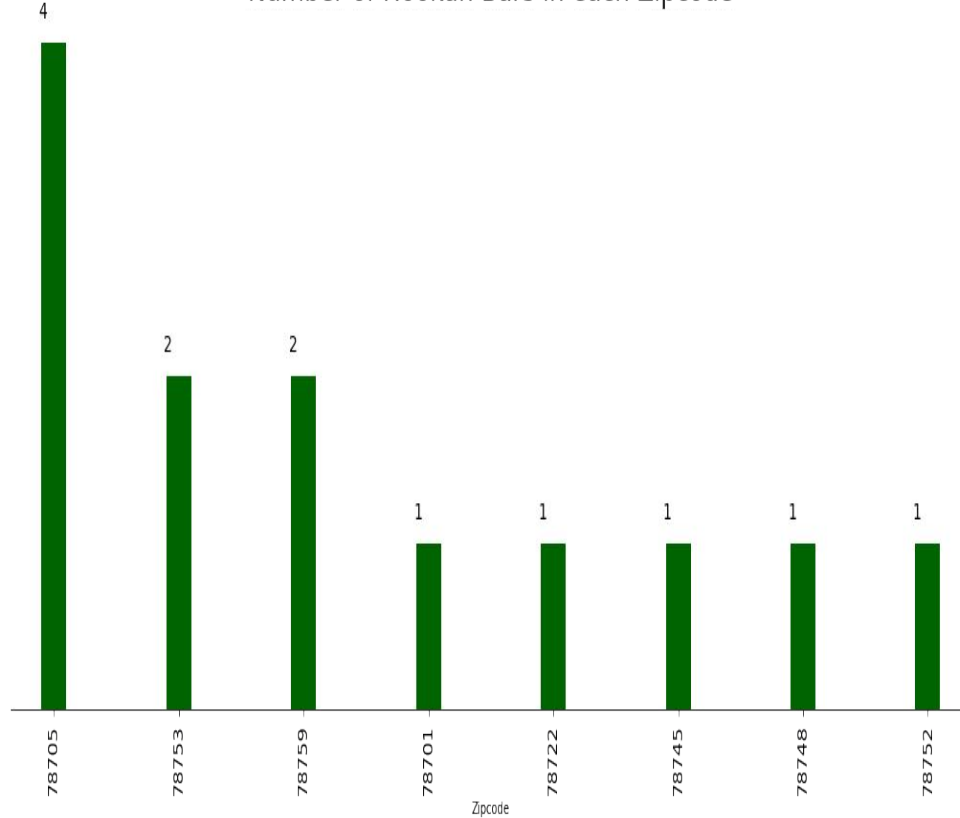
# Data requirements and acquisition

---

- List of zip codes along with their geographical coordinates. Acquired from [opendatasoft.com](https://opendatasoft.com)
- List of existing hookah bars and their corresponding zip codes. Acquired using API from [foursquare.com](https://foursquare.com)
- List of venues of all categories in all zip codes (up to specified limit). Acquired using API from [foursquare.com](https://foursquare.com)
- Distances of regions from city center and other target areas. Acquired using [Geopy](#) module

# Identify the region with most hookah bars: **78705**

Number of Hookah Bars in each Zipcode



- This region will be designated as the ideal zip code
- It's hidden characteristics will be analyzed and extracted using machine learning
- The recommended location will be strategically placed so as to avoid being in competition with it

# K-means clustering machine learning algorithm

---

- Get list of venues across all categories in all zip codes up to specified limit(100)
- Sort venues by most popular in their region
- Append list of 15 most popular venues to each zip code value
- Apply k-means cluster to split them into specified number of clusters (8)
- Identify cluster containing ideal zip code (78705) and extract zip code values

# Optimizing distances of zip codes acquired from cluster analysis

---

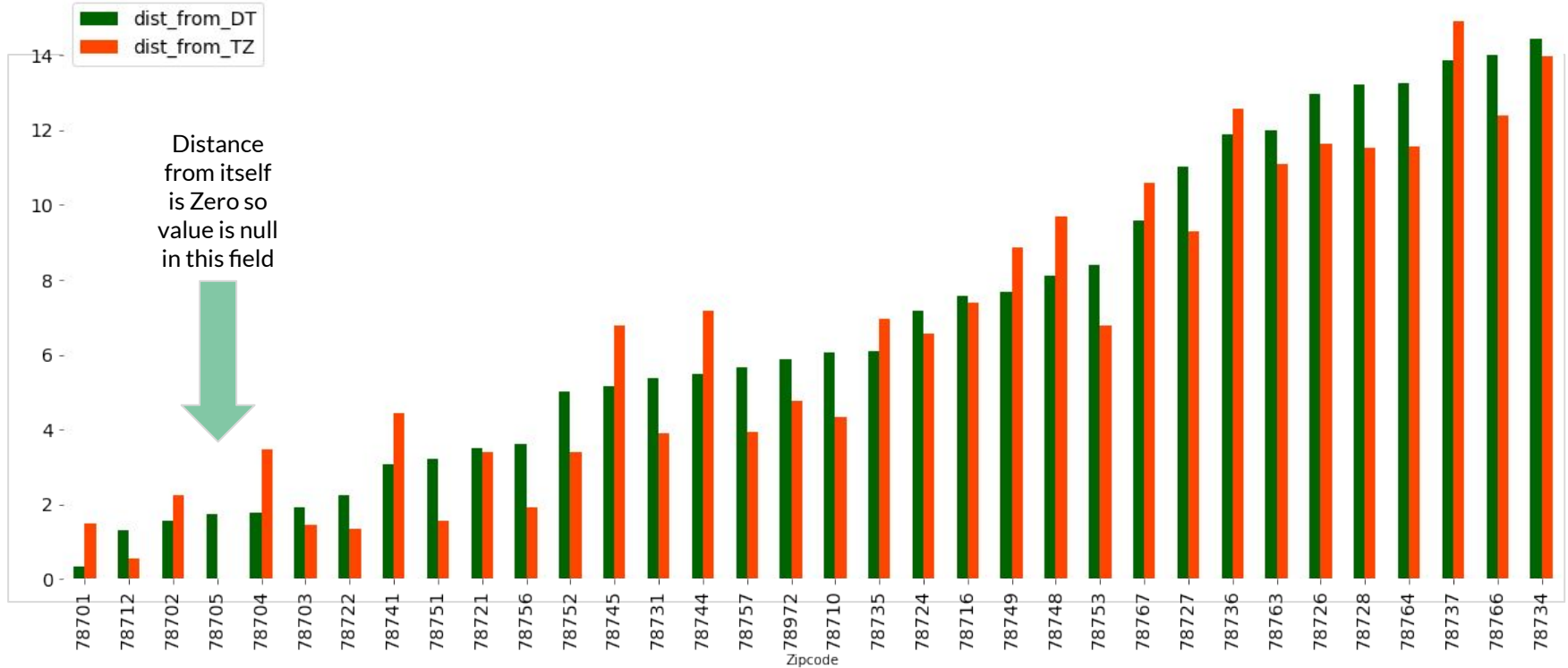
The regions identified in the previous process will be selected for being as close to the city-center as possible, yet as far away from the saturated zip code (78705) as possible

- Merge the zip codes back into the original table with geographical coordinates
  - The new table will now have much fewer rows because only rows with the selected zip codes are kept
- Extract the coordinates of the ideal/saturated zip code (78705)
- Get the coordinates of the city center (Downtown, Austin, TX) using Geopy

Calculate the distance of the selected zip codes from the points in the last 2 steps and append them to the table

# Distances from both city-center and saturated region

Distance from both Downtown and Exemplary Zipcode



# Results & Conclusion

---

2 zip codes are chosen based on arbitrary distance criteria that can later be widened if on-ground analysis requires more candidates.

For now, tentative high-confidence locations are: **78702** and **78704**