

Best Cities for Remote Workers

1. Introduction

1.1 Background

Remote work has increased significantly these couple of years and is taking a new turn with the coronavirus crisis. Companies have been hit brutally and had to send their employees working from home for those that have a business that can be handled remotely. From now on companies have to rethink their work models to face crisis. Remote work can benefit both employees and employers. One on side, employees will be able to avoid spending hours in transport and will be able to pick the location they want to work from. This will improve all in all the lifework balance and many employees will opt for this solution in the future. On the other hand, companies will have the possibility to hire the best people not only in a radius of 40 kms but from everywhere in the world. It will also allow them to reduce their real estate cost substantially and will be then more cost effective. With that in mind, many future employees will pick a destination, pack their bags and will work remotely. This is already a reality for the digital nomads that just need a good internet connection to work from their computer.

1.2 Problem

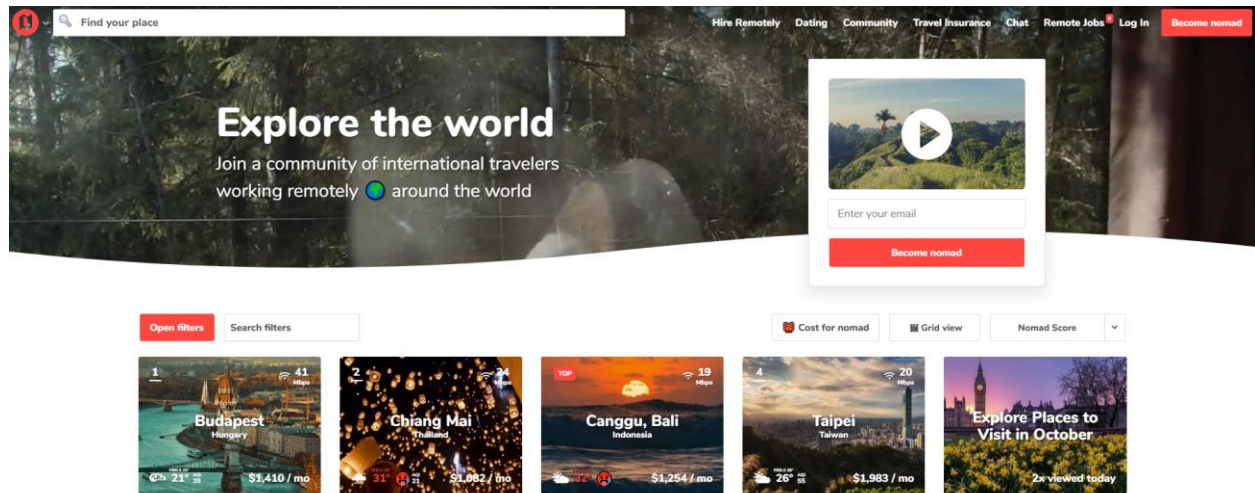
But how to choose the destination? What are the best cities to work remote? I am not saying to choose one place and stay there forever, but a guide to know those places is always useful, like TripAdvisor or foursquare when you are looking for a nice restaurant. But before getting on board, there are some criteria to consider. For example, making sure to be able to work and to have a great overall experience onsite. In other terms, a good internet connection, affordable cost of living, safety and fun. There are many others that can be included, and we will mention them during the analysis.

2. Data Acquisition and Cleaning

2.1 Data Sources

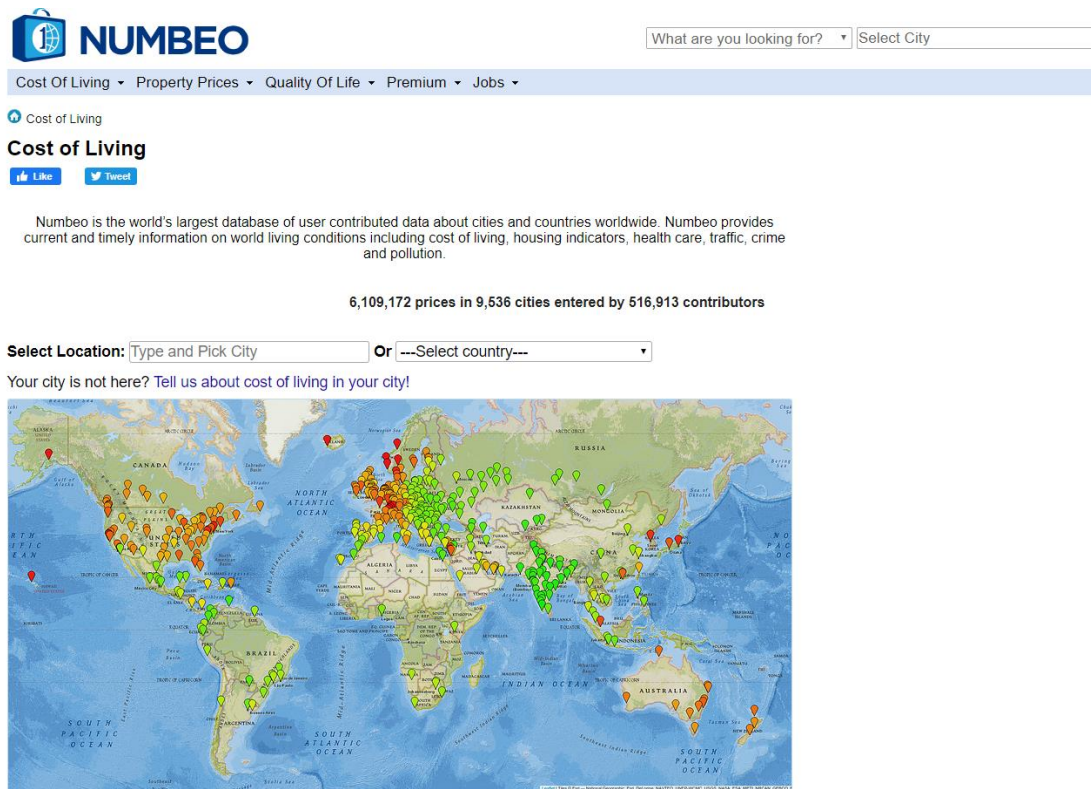
There is many resources on internet that can be used to solve our problems. Of course, we are going to use the foursquare API to gather quantitative data of the cities. Number of coffee places, coworking places by city. You will be using also meetup that lists the number of meetups by city and tells a lot about the dynamic of a city. To work thorough on the cost of living, safety, quality of life and so on, we are going to be using Numbeo that is the largest database of user contributed data about cities and countries worldwide and Nomadlist that classifies the best cities to work remote. To work with those data, we are going to extract thanks to BeautifulSoup, put them in DataFrame and create a scoring model based on the criterion chosen.

Nomadlist:



The Nomadlist homepage features a dark, forest-themed background. At the top, a navigation bar includes links for 'Hire Remotely', 'Dating', 'Community', 'Travel Insurance', 'Chat', 'Remote Jobs', 'Log In', and a red 'Become nomad' button. A search bar with the placeholder 'Find your place' is on the left. The main headline reads 'Explore the world' with the subtext 'Join a community of international travelers working remotely around the world'. A video player with a play button is on the right, with an email input field and a 'Become nomad' button below it. Below the hero section, there are filter buttons: 'Open filters', 'Search filters', 'Cost for nomad', 'Grid view', and 'Nomad Score'. A row of five location cards is displayed: 1. Budapest, Hungary (\$1,410/mo), 2. Chiang Mai, Thailand (\$1,082/mo), 3. Canggu, Bali, Indonesia (\$1,254/mo), 4. Taipei, Taiwan (\$1,983/mo), and a featured card 'Explore Places to Visit in October' (2x viewed today).

Numbeo:



The Numbeo homepage has a clean, white layout. At the top is the Numbeo logo and a search bar with the placeholder 'What are you looking for?' and a 'Select City' dropdown. Below this is a navigation bar with links: 'Cost of Living', 'Property Prices', 'Quality of Life', 'Premium', and 'Jobs'. The 'Cost of Living' section is active, showing social media share buttons for 'Like' and 'Tweet'. A paragraph describes Numbeo as the world's largest database of user-contributed data about cities and countries worldwide, providing information on cost of living, housing, health care, traffic, crime, and pollution. Below this, it states '6,109,172 prices in 9,536 cities entered by 516,913 contributors'. A 'Select Location' section includes a text input 'Type and Pick City' and a dropdown '---Select country---'. A link says 'Your city is not here? Tell us about cost of living in your city!'. At the bottom is a world map with numerous colored pins indicating data points across various countries.

2.2 Data Cleaning

Data downloaded and scraped are combined into one unique table. What we want to do first is to build a scoring model based on different factors that we selected like cost of living, quality of life, internet speed and so on.

To build our scoring model, we are going to use Numbeo. This website is the largest contributed database about cities and is a tremendous tool that we are going to use extensively to gather all the data that are interesting to define a good city for remote workers. On the technical side, we are going to use the package BeautifulSoup to scrape the data from the website. The data that we are going to scrape are in two different pages. Therefore, we are going to create two dataframes and merge them ultimately.

DataFrame 'Cost of Living':

| | City | Cost of Living Index | Rent Index | Cost of Living Plus Rent Index | Groceries Index | Restaurant Price Index | Local Purchasing Power Index |
|---|-----------------------|----------------------|------------|--------------------------------|-----------------|------------------------|------------------------------|
| 0 | Basel, Switzerland | 128.33 | 46.43 | 89.50 | 131.93 | 112.94 | 108.76 |
| 1 | Zurich, Switzerland | 125.57 | 62.96 | 95.88 | 126.29 | 109.21 | 124.61 |
| 2 | Lausanne, Switzerland | 122.55 | 52.62 | 89.40 | 129.41 | 104.10 | 104.97 |
| 3 | Geneva, Switzerland | 118.79 | 66.92 | 94.20 | 117.24 | 107.27 | 111.93 |
| 4 | Bern, Switzerland | 112.97 | 41.40 | 79.04 | 105.30 | 103.46 | 129.30 |

Some data are present in the first dataframe 'cost of living' are not present in the second one 'quality of life'. It seemed to us that it is not important for cities that don't have good metrics for our scoring but significant if it is about destinations that are appreciated by remote workers. Bali for example, which is an important destination for remote workers is missing in the second dataframe. That is why we decided to add a row for it with the same data as Jakarta, the capital of Indonesia (Bali is an island that belongs to Indonesia). We know that this might not be very accurate because of the difference between the two places but it seems the closest one. A mean or a median of the whole dataset would have been less accurate.

DataFrame 'Quality of Life':

| | City | Safety Index | Healthcare Index | Pollution Index | Climate Index |
|---|-----------------------------|--------------|------------------|-----------------|---------------|
| 0 | Canberra, Australia | 79.33 | 82.17 | 14.07 | 82.72 |
| 1 | Adelaide, Australia | 71.63 | 81.22 | 18.31 | 94.96 |
| 2 | Raleigh, NC, United States | 66.17 | 75.62 | 21.87 | 83.88 |
| 3 | Wellington, New Zealand | 70.78 | 74.90 | 13.66 | 97.68 |
| 4 | Columbus, OH, United States | 57.87 | 74.28 | 25.19 | 71.29 |

A good internet connection is essential for a remote worker. Therefore, we used the data from a Research designed and compiled by Cable.co.uk, and gathered by M-Lab, an open source project with contributors from civil society organizations, educational institutions, and private sector companies.

| | Country | Ranking 2019 | Speed | % change (2018-2019) |
|---|-----------|--------------|-----------|----------------------|
| 0 | Taiwan | 1.0 | 85.019167 | 2.02618 |
| 1 | Singapore | 2.0 | 70.856914 | 0.173352 |
| 2 | Jersey | 3.0 | 67.460708 | 1.18319 |
| 3 | Sweden | 4.0 | 55.176572 | 0.199407 |
| 4 | Denmark | 5.0 | 49.192330 | 0.118367 |

Social life is important also for the remote workers. In fact, it is nice to meet some likeminded people onsite to share experiences and have fun. The website NomadList gathered an impressive amount of data from digital nomads. The one we are going to be using is the percentage of nomads per population. Scraping the pages of NomadList is quite difficult, therefore we gathered the data manually in an Excel Sheet. It was long and fastidious, but much needed for our model.

| | City | Cntry | Pct Nomads |
|---|----------------|-----------|------------|
| 0 | Bali | Indonesia | 27.0 |
| 1 | Ko Pha Ngan | Thailand | 4.0 |
| 2 | Bocas del Toro | Panama | 2.0 |
| 3 | Chiang Mai | Thailand | 0.3 |
| 4 | Whistler | Canada | 0.2 |

Lastly, we merge all the DataFrame to only get one that contains all the data useful for our scoring model. This was done by splitting the city and country of the DataFrames 'Cost of Living' and 'Quality of Life' and joining the different DataFrame on the key 'City':

| | City | Country | Cost of Living Plus Rent Index | Groceries Index | Restaurant Price Index | Safety Index | Healthcare Index | Pollution Index | Climate Index | Internet Index | Pct Nomads |
|-----|---------------|---------------|--------------------------------|-----------------|------------------------|--------------|------------------|-----------------|---------------|----------------|------------|
| 0 | Zurich | Switzerland | 95.88 | 126.29 | 109.21 | 83.11 | 73.69 | 17.77 | 81.48 | 38.85 | 0.1 |
| 1 | Geneva | Switzerland | 94.20 | 117.24 | 107.27 | 71.41 | 68.89 | 29.26 | 82.61 | 38.85 | 0.1 |
| 2 | New York | United States | 100.00 | 100.00 | 100.00 | 54.32 | 63.33 | 56.34 | 79.66 | 32.89 | 0.1 |
| 3 | San Francisco | United States | 102.32 | 86.54 | 83.05 | 45.07 | 66.02 | 46.33 | 97.26 | 32.89 | 0.1 |
| 4 | Anchorage | United States | 64.49 | 87.60 | 67.81 | 37.21 | 60.62 | 16.62 | 41.61 | 32.89 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 235 | Bogota | Colombia | 19.29 | 21.85 | 19.04 | 38.61 | 64.36 | 69.61 | 97.12 | 3.48 | 0.0 |
| 236 | Medellin | Colombia | 17.53 | 23.26 | 17.52 | 48.41 | 76.59 | 63.02 | 99.76 | 3.48 | 0.1 |
| 237 | Lahore | Pakistan | 14.11 | 19.97 | 15.08 | 60.71 | 64.61 | 79.19 | 67.56 | 1.44 | 0.0 |
| 238 | Islamabad | Pakistan | 14.73 | 20.37 | 15.55 | 70.77 | 64.53 | 43.64 | 76.91 | 1.44 | 0.0 |
| 239 | Karachi | Pakistan | 13.68 | 19.37 | 16.71 | 45.42 | 58.51 | 89.92 | 71.32 | 1.44 | 0.0 |

3. Exploratory Data Analysis

3.1 Data Normalizing

The columns of our dataframe have completely different scales. Therefore, we rescaled the dataframe to change the values of its columns to a common scale, without distorting differences in the ranges of values. We will be using the normalization method that rescales the values into a range of [0.1]. For that purpose, we are going to use the data normalization method of scikit learn.

| | City | Country | Cost of Living Plus Rent Index | Groceries Index | Restaurant Price Index | Safety Index | Healthcare Index | Pollution Index | Climate Index | Internet Index | Pct Nomads |
|---|---------------|---------------|--------------------------------|-----------------|------------------------|--------------|------------------|-----------------|---------------|----------------|------------|
| 0 | Zurich | Switzerland | 0.928856 | 1.000000 | 1.000000 | 0.923212 | 0.731142 | 0.060070 | 0.784588 | 0.448057 | 0.003704 |
| 1 | Geneva | Switzerland | 0.910296 | 0.915357 | 0.980614 | 0.764762 | 0.630004 | 0.198387 | 0.797782 | 0.448057 | 0.003704 |
| 2 | New York | United States | 0.974370 | 0.754115 | 0.907964 | 0.533315 | 0.512853 | 0.524377 | 0.763339 | 0.376808 | 0.003704 |
| 3 | San Francisco | United States | 1.000000 | 0.628227 | 0.738583 | 0.408044 | 0.569532 | 0.403876 | 0.968827 | 0.376808 | 0.003704 |
| 4 | Anchorage | United States | 0.582081 | 0.638141 | 0.586290 | 0.301598 | 0.455752 | 0.046226 | 0.319089 | 0.376808 | 0.000000 |

3.2 Scoring Model

Some factors may be more important than others. We consider that the climate, the restaurant prices and the percentage of digital nomads in the city are the most important ones. That's why we are going to give them a weight twice more important than the others. Lastly, we sum the factors to get our final score.

| | City | Country | Total Score |
|----|-------------|----------------|-------------|
| 0 | Taipei | Taiwan | 7.944064 |
| 1 | Bursa | Turkey | 7.831079 |
| 2 | Izmir | Turkey | 7.596611 |
| 3 | Bali | Indonesia | 7.523524 |
| 4 | Valencia | Spain | 7.414207 |
| 5 | Medellin | Colombia | 7.400906 |
| 6 | Curitiba | Brazil | 7.357749 |
| 7 | Lisbon | Portugal | 7.336295 |
| 8 | Porto | Portugal | 7.313023 |
| 9 | Timisoara | Romania | 7.280140 |
| 10 | Prague | Czech Republic | 7.240398 |
| 11 | Islamabad | Pakistan | 7.216972 |
| 12 | Coimbatore | India | 7.184243 |
| 13 | Brno | Czech Republic | 7.159454 |
| 14 | Wellington | New Zealand | 7.130798 |
| 15 | Ankara | Turkey | 7.082159 |
| 16 | Cluj-Napoca | Romania | 7.066849 |
| 17 | Chiang Mai | Thailand | 7.035165 |
| 18 | Adelaide | Australia | 6.996851 |
| 19 | Zagreb | Croatia | 6.972736 |

There we are! We got finally our results. In the top 20, we can find some well-known cities for remote workers like Bali, Medellin, Lisbon, Prague or Chiang Mai. There is some surprise also, with cities like Bursa and Izmir in Turkey, Timisoara and Cluj in Romania, Islamabad in Pakistan or Coimbatore in India. In can be explained by the facts that cost of living is quite cheap and safety and healthcare are quite good and have improved substantially in those regions.

4. Machine Learning Model (Clustering)

4.1 Run the model

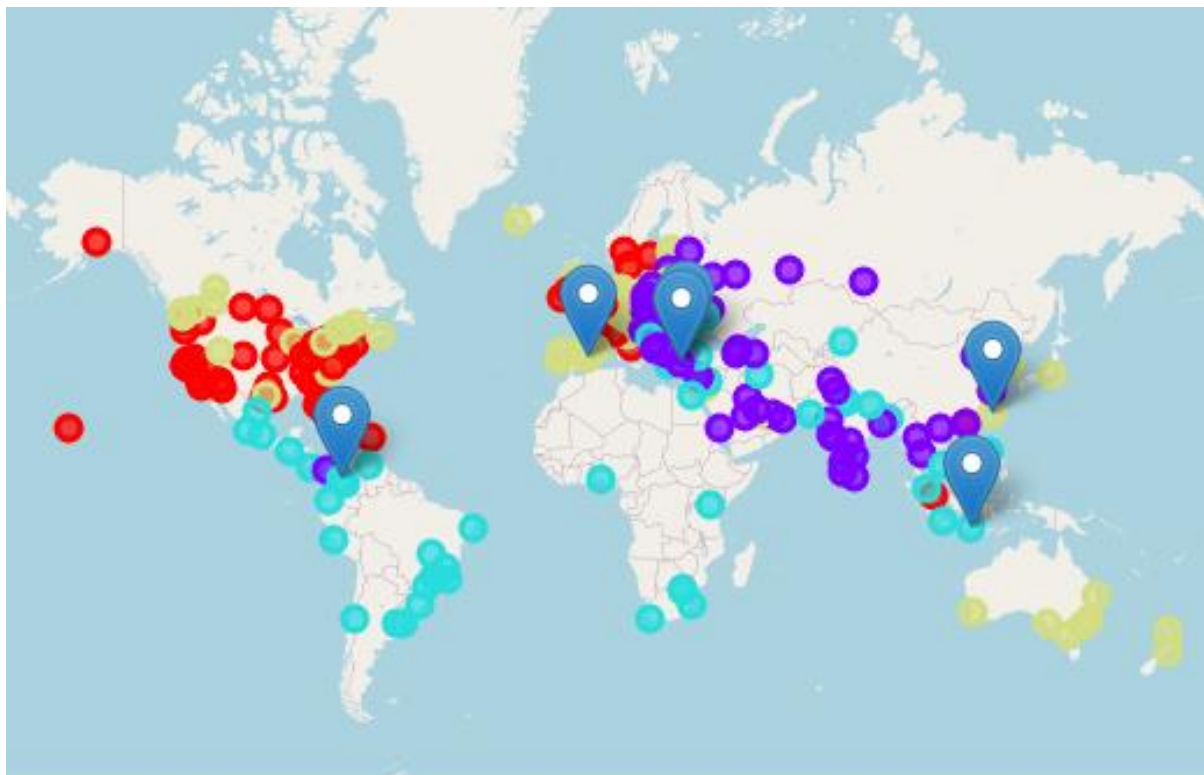
Now that we have all the data available, we can also think of find the similarities between the cities. To do so, we are going to cluster them thanks to k-means model. It is especially useful when it comes to quickly discover insights from unlabeled data. For that analysis, we considered only factors from the

Numbeo pages (without the % of digital nomads in the population). Once the model run, we added a label column with the cluster number of each city

| | City | Country | Cost of Living Plus Rent Index | Groceries Index | Restaurant Price Index | Safety Index | Healthcare Index | Pollution Index | Climate Index | Internet Index | Pct Nomads | Total Score | Labels |
|---|---------------|---------------|--------------------------------|-----------------|------------------------|--------------|------------------|-----------------|---------------|----------------|------------|-------------|--------|
| 0 | Zurich | Switzerland | 0.070858 | 0.000000 | 0.000000 | 0.923348 | 0.731142 | 0.939930 | 1.569177 | 0.448057 | 0.007407 | 4.689920 | 0 |
| 1 | Geneva | Switzerland | 0.089709 | 0.086678 | 0.039853 | 0.764762 | 0.630004 | 0.801613 | 1.595563 | 0.448057 | 0.007407 | 4.463647 | 0 |
| 2 | New York | United States | 0.021402 | 0.251765 | 0.138974 | 0.533315 | 0.512853 | 0.475623 | 1.526678 | 0.376808 | 0.007407 | 3.844826 | 0 |
| 3 | San Francisco | United States | 0.000000 | 0.373096 | 0.523605 | 0.408044 | 0.569532 | 0.596124 | 1.937653 | 0.376808 | 0.007407 | 4.792270 | 0 |
| 4 | Anchorage | United States | 0.417942 | 0.363341 | 0.828122 | 0.301598 | 0.455752 | 0.953774 | 0.638179 | 0.376808 | 0.000000 | 4.335515 | 0 |

4.2 Map of the result

To better understand the results, let's visualize it with a world map.



This map provides interesting characteristics. The data points from the same cluster are located next to each other. The green points are in majority in Asia. The purple points in Latin America and Africa. The red and blue points are in the so called 'developed countries'.

What is interesting too, is that the cities in the top 6 of our scoring model don't belong to only one clusters but they are spread out into different clusters:

- Taipei and Valencia in the blue cluster
- Bali in the green cluster
- Burza, Izmir and Medellin in the purple cluster

The only cluster not represented here is the red cluster where the cost of living is relatively higher than the quality of life.

5. Conclusions

In this study, I analyzed the several data that makes a city good for remote workers. I first identified the essential factors and pondered them subjectively according to what I think is most important for a remote worker. The scoring model gave some expected results like Medellin, Bali or Valencia and some unexpected ones like Burza, Islamabad. This proved that data can show you things that you were not aware. I realized that the safety and health in some regions are considerably improved or that big north American or European cities have a cost of living relatively higher than the quality of life.

The clustering model defined distinct categories of cities and that the world is way more widespread than the typical developed/undeveloped schema. Besides the cities in the top 6 of our scoring model don't belong to only one clusters but they are spread out into different clusters which shows that there is not a best cities to work remote but several and it depends on individual preferences.