Personal Quality Index Project

Business Intelligence & Big Data Management
Module 2

01 September 2023

Alessio David, Filippo Sibra, Gaspare Biondo

Agenda

→ **Project Goal**

→ **E2E Project Process**
  - Analysis
  - Data Architecture
  - Data Visualization

→ **New Possible Developments**
  - Data Architecture To-Be
  - New Features

Project Goal

Identify a **solution to analyze quality life** of several cities **based on different parameters** from economic prospectives to environments, life and health

Features

- Visualize and Analyze **Indexes by Cities**

- Visualize the main index components like **cost of living**

- Calculate a **Personalized quality index** based on selected parameters

Scope & Data Sources

- **Italian Cities**

- **Numbeo.com**

- **City Latitude and Longitude**

Target

- **People** interested to **change life city** (e.g. for a new job)

Analysis | Context

Starting from **Numbeo** web site, we **analyzed formula** behind the shown Indexes, how to **retrieve data** from the web site and how to use additional data to **calculate a personalized index**:

Data & Indexes

◆ Within Numbeo there is a collection of data regarding the **cost of living by city split by categories**:

Restaurants & Going out Food at Home Drinks at Home | Markets	Public Transport & Taxi Car and Gasoline Utilities | Sport	Childcare and School Clothing and Shoes Rent

◆**Indexes**:

Quality Of Life Purchasing Power Pollution Levels	Housing Affordability Cost Of Living Safety	Healthcare Quality Climate Condition Commute Times

Index recalculated to personalize it, based on **selected parameters**: **Salary and Number of household members**

Analysis | Logics Applied

Quality Index

◆ The **quality index** is calculated considering several indexes, it is necessary to consider different variables. More high is the value better is the quality of life:

$$100 + \frac{PurchasingPowerIncludesRentIndex}{2.5} - HousePriceToIncomeRation - \frac{CostOfLivingIndex}{10}$$
$$+ \frac{SafetyIndex}{2.0} + \frac{HealthIndex}{2.5} - \frac{TrafficTimeIndex}{2.0} - PollutionIndex * \frac{2.0}{3.0} + \frac{ClimateIndex}{3.0}.$$

◆ In order to perform an **Index personalized based on** specifics parameters we consider the quality index as a dependent function by **Salary** and **Number of household members**. We chosen these 2 variable for the following reasons:

1. Tech/Analysis prospective: **Data Availability | Clear impact on cost of living** calculation **and purchasing power**
2. Business prospective: **Interest for this kind of analysis considering the customer goal**
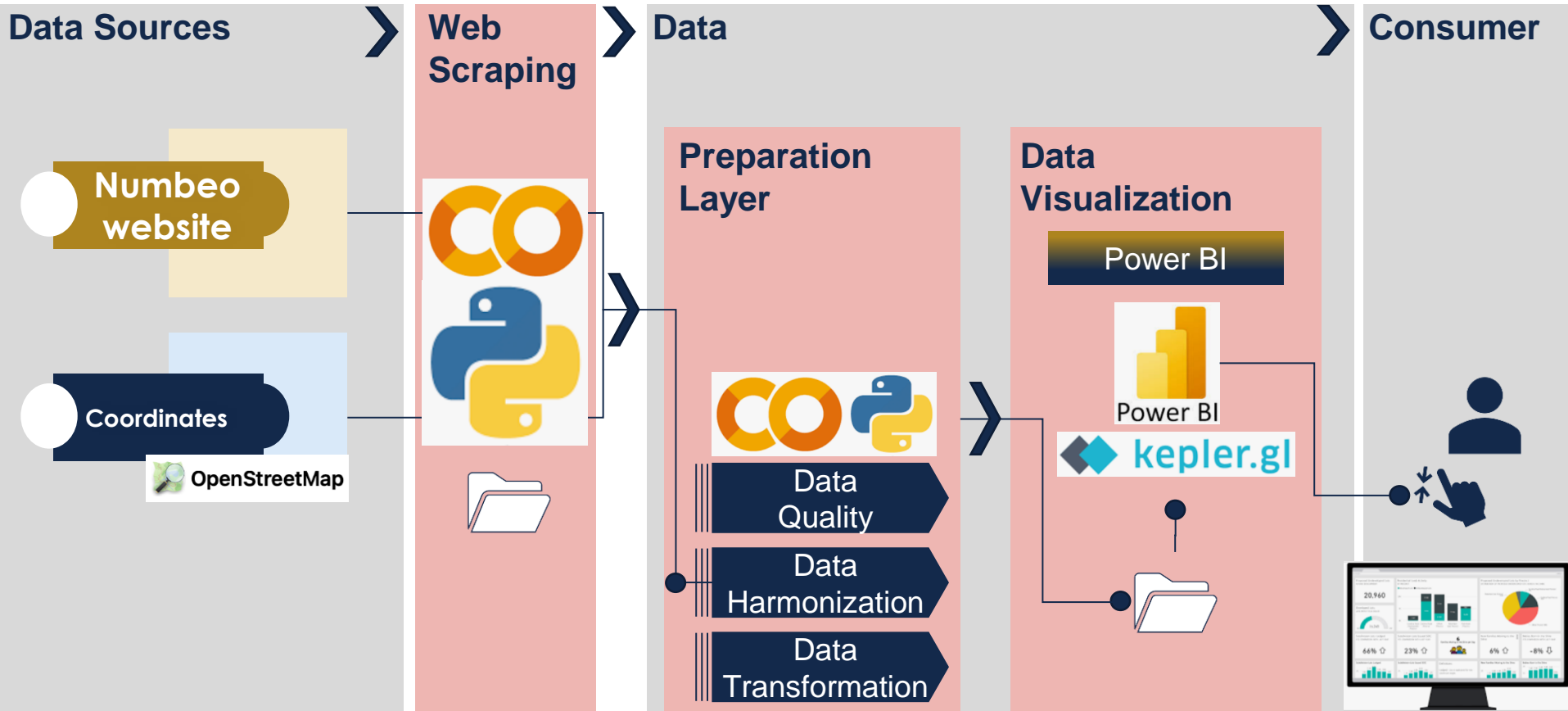
Purchasing Power $\dfrac{AverageDisposableSalary(City)}{BasketConsumerPlusRent(City)}$ $\quad Basket\ Consumer\ \sum \begin{array}{l} CostOfLiving(City) \\ * CostFactor \\ * CategoryFactor \end{array}$

Analysis | Data Architecture As-Is

Crawler | Navigate the pages related to Italian cities

List of available cities

https://www.numbeo.com/quality-of-life/in/Milan
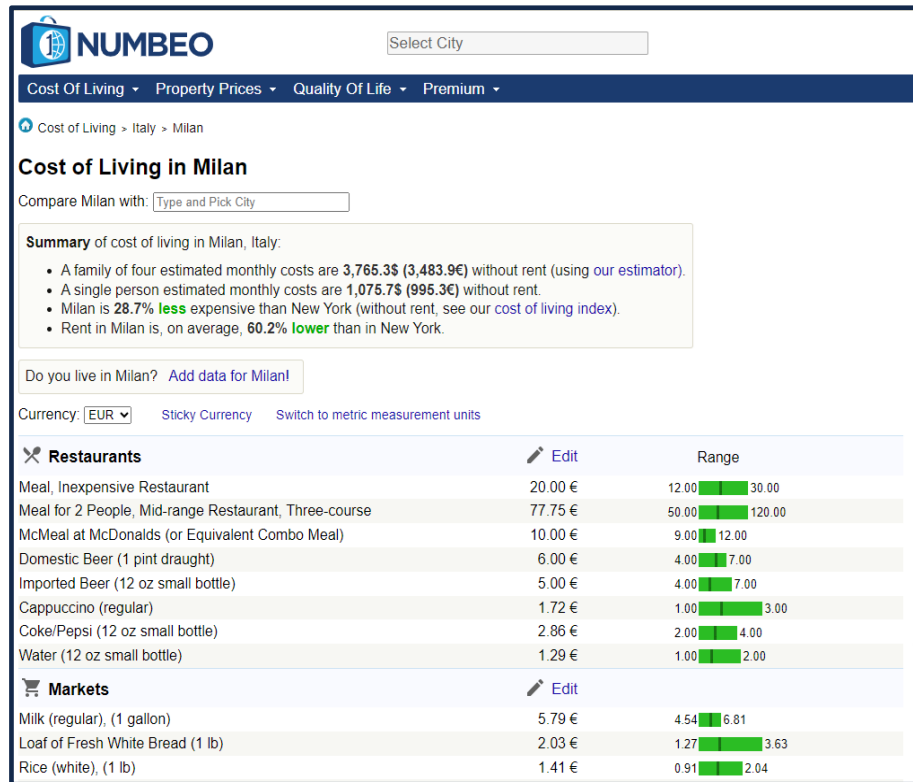(URL is different from API called on click)

Hybrid strategy: Use headless GET APIs instead of crawling the website
API called on click: *https://www.numbeo.com/quality-of-life/city_result.jsp?country=Italy&city=Milan*

Scraping | Get raw data to recalculate quality index

Strategy: tokenization of raw text

Issues: Missing data leads to missing tokens


```
wd.find_elements(By.CLASS_NAME, "data_wide_table")[0].text.split('\n')

'Coke/Pepsi (12 oz small bottle) 2.86 € 2.00-4.00',
'Water (12 oz small bottle) 1.29 € 1.00-2.00',
'Markets Edit',
'Milk (regular), (1 gallon) 5.79 € 4.54-6.81',
'Loaf of Fresh White Bread (1 lb) 2.03 € 1.27-3.63',
'Rice (white), (1 lb) 1.41 € 0.91-2.04',
'Eggs (regular) (12) 3.79 € 2.40-5.00',
'Local Cheese (1 lb) 7.23 € 4.31-9.07',
'Chicken Fillets (1 lb) 5.04 € 2.72-6.80',
'Beef Round (1 lb) (or Equivalent Back Leg Red Meat) 9.07 € 4.54-11.34',
'Apples (1 lb) 1.03 € 0.45-1.36',
```

Data structure

param	value	unit	lower	upper	city	latitude	longitude
Meal, Inexpensive Restaurant	11.00	€	10	30	Zambrone	38.699180	15.990094
Meal for 2 People, Mid-range Restaurant, Three...	65.00	€	60	70	Zambrone	38.699180	15.990094
McMeal at McDonalds (or Equivalent Combo Meal)	11.75	€	8.5	15	Zambrone	38.699180	15.990094
Domestic Beer (1 pint draught)	3.50	€	3.5	10	Zambrone	38.699180	15.990094
Imported Beer (12 oz small bottle)	4.50	€	3	6	Zambrone	38.699180	15.990094
...	...	...	...	...	...	...	...
1 Summer Dress in a Chain Store (Zara, H&M, ...)	15.00	€	15	15	Abano Terme	45.360314	11.789783
1 Pair of Nike Running Shoes (Mid-Range)	90.00	€	90	90	Abano Terme	45.360314	11.789783
1 Pair of Men Leather Business Shoes	100.00	€	100	100	Abano Terme	45.360314	11.789783
Average Monthly Net Salary (After Tax)	2028.67	€	NaN	NaN	Abano Terme	45.360314	11.789783
Mortgage Interest Rate in Percentages (%), Yea...	1.35	NaN	1	4.9	Abano Terme	45.360314	11.789783

API accepting city name with different spelling:

requests.get(f'https://nominatim.openstreetmap.org/search?q={Milan}&limit=2&format=json')

Personalized quality index final dataset

Highlights:

Starting from the main data:
- Cost Of Living
- Category Factor
- # Household members

We performed the **calculation** of the necessary variables and as final result of the **Quality Index value**, for all cases respect to **Cities**, **salary** and **number of household members**

Unstacked structure

city	salary	members	quality index
Verona	1000	1	173.1
Verona	1300	1	189.3
Verona	1500	1	200.1
Verona	1800	1	216.2
Verona	2000	1	227
Verona	2500	1	253.9
Verona	3000	1	280.8
Verona	3500	1	307.7
Verona	1000	2	146.7
Verona	1300	2	155
Verona	1500	2	160.4
Verona	1800	3	152.5
Verona	2000	3	156.2

Stepwise salary and members
to avoid calculation by the dashboard

Analysis | Data Model

Below the main tables used for the data visualization layer:

Retrieved Data by Web
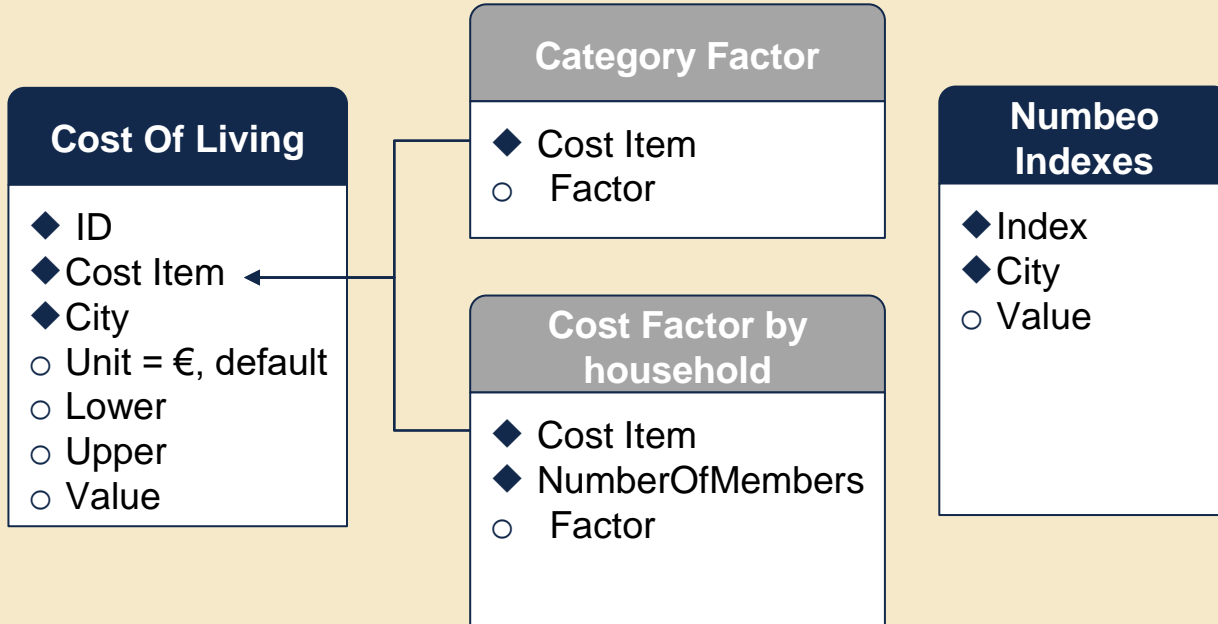
Category Factor
- ◆ Cost Item
- ○ Factor

Cost Of Living
- ◆ ID
- ◆ Cost Item
- ◆ City
- ○ Unit = €, default
- ○ Lower
- ○ Upper
- ○ Value

Cost Factor by household
- ◆ Cost Item
- ◆ NumberOfMembers
- ○ Factor

Numbeo Indexes
- ◆ Index
- ◆ City
- ○ Value

Calculated Data

Personal Indexes
- ◆ City
- ◆ Number Of Members
- ◆ Salary
- ○ LocalPurchasingIndex
- ○ BasketConsumerRent
- ○ QualityIndexPersonal

Data Visualization | Power BI *(1/3)*

A

Cities Comparison

1 Filtering two cities it is possible show the main indexes calculating to evaluate the difference using both numbers and graph visualization

2 The page is showing also the top 10 city list by quality index

Data Visualization | Power BI *(2/3)*

B **Cost Of Living Analysis**

1 Filtering a specific city it is possible show the costs split by category

2 It is possible, using the decomposition tree graph, to go through the cost item with a greater granularity

COST OF LIVING

city

Anzio (Rome)

🔍 rom

☑ **Anzio (Rome)**
☐ **Rome**

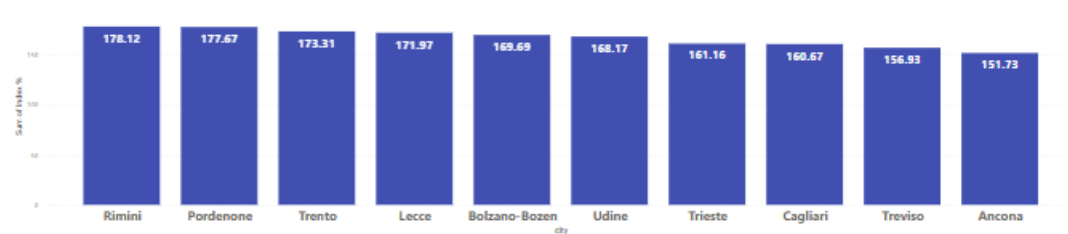|	Cost of Living
	3,100.00
	366.67
Clothing and Shoes	297.50
Utilities	221.89
Restaurants	75.88
Public Transport and Taxi	62.56
Leisure and Sport Memberships	56.67
Food at Home	27.06
Car and Gasoline	8.30
Going out	7.00
Drinks at Home	6.51
Cigarettes	5.90
Total	**4,235.94**

Cluster Parameter
Rent

Cost of living
4,236

Rent
3,100

Childcare and School Fees
367

Clothing and Shoes
298

Utilities
222

Restaurants
76

Public Transport and Taxi
63

Leisure and Sport Membe...
57

Food at Home
27

Car and Gasoline
8

Going out
7

Drinks at Home
7

Cigarettes
6

Apartment (3 bedroom...
1,000

Apartment (1 bedroom) i...
800

Apartment (3 bedrooms) ...
800

Apartment (1 bedroom) ...
500

Filters

Master - progetto sett v2, Cost of Living detail
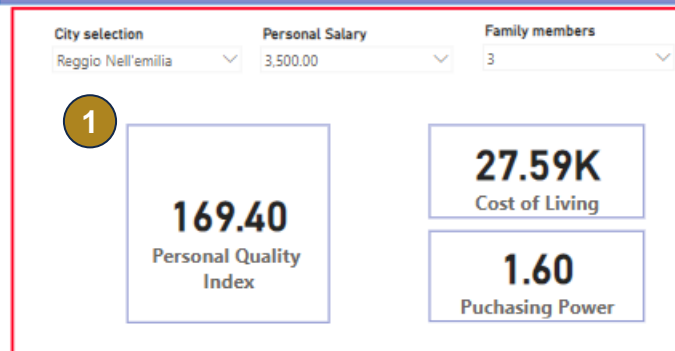Data updated on 8/30/23, 12:15 PM

Data Visualization | Power BI *(3/3) Core Feature*

C Cost Of Living Analysis

1 Filtering a specific city it is possible show the costs split by category

2 It is possible, using the decomposition tree graph, to go through the cost item with a greater granularity

Personal Quality Index

1

City selection: Reggio Nell'emilia
Personal Salary: 3,500.00
Family members: 3

169.40
Personal Quality Index

27.59K
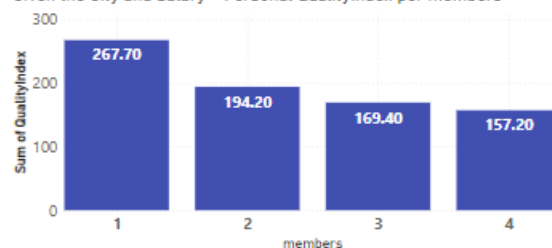Cost of Living

1.60
Puchasing Power

The Personal Quality Index is determined by utilizing a custom combination of:
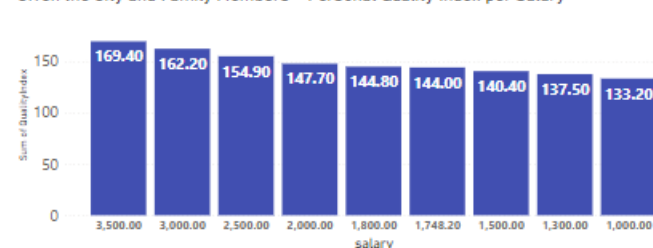*-Personal Salary*
-# of family members

potentially serving as a surrogate for case-specific simulation.

This calculation is derived from the same raw data used in computing the "standard" ratio.

«
Filters

Given the City and Salary - Personal QualityIndex per members

members	Sum of QualityIndex
1	267.70
2	194.20
3	169.40
4	157.20

Given the City and Family Members - Personal Quality Index per Salary

salary	Sum of QualityIndex
3,500.00	169.40
3,000.00	162.20
2,500.00	154.90
2,000.00	147.70
1,800.00	144.80
1,748.20	144.00
1,500.00	140.40
1,300.00	137.50
1,000.00	133.20

2

Master - progetto sett v2, Personal Quality Index
Data updated on 8/30/23, 12:15 PM

New Possible Developments | Data Architecture To Be

Data Sources

Numbeo website

Coordinates

OpenStreetMap

External Data
- Airbnb - Immobiliare.it
- Unemployment Rate | ISTAT
- Other Financial Data,…..

Data

AWS Batch **AWS S3** **AWS Athena**

Data Extraction & Ingestion

Scheduled jobs in order to retrieve data from the different data sources (Web Scraping, other DB, API,..) and ingest data on **S3**

Preparation Layer

Data Quality

Data Harmonization

Data Transformation

Consumption Layer

AWS S3

Amazon Athena

Data Visualization

kepler.gl

Power BI

Power BI

Consumer

Admin

Customer

Apps

Web APP

Flask

Thanks!!