42184 Data Science for Mobility

42577 Introduction to Business Analytics course

Challenge statement

Welcome to this year's challenge! :-)

The topic this year is *Sustainability in World Cities*. At a time when the world is facing unprecedented challenges of different kind, including climate change, pandemics, social inequality, degrading biodiversity, we need to be conscious of the impact and potential of cities as drivers of (positive) change. But what makes a city more sustainable? What best practices exist that could push poor performing cities to improve?

In this project, we invite you to appropriate these questions and use your best Data Sciences skills to explore them. We do not expect you to discover revolutionary knowledge and save the world with a single Data Sciences project, instead we want you to address the mandatory questions (below) but also seek yourself for new questions, new data, new insights.

You have access to a dataset from the "Urban Typologies" project[1], where you can find 65 indicators[2] that relate to demographics, mobility, economy, city form. This dataset was obtained by combining multiple sources and had the general objective of classifying the different cities of the world according to a *typology.* It is in itself an interesting Data Sciences exploration. We recommend that you go through the associated literature[1] to know more.

Project structure

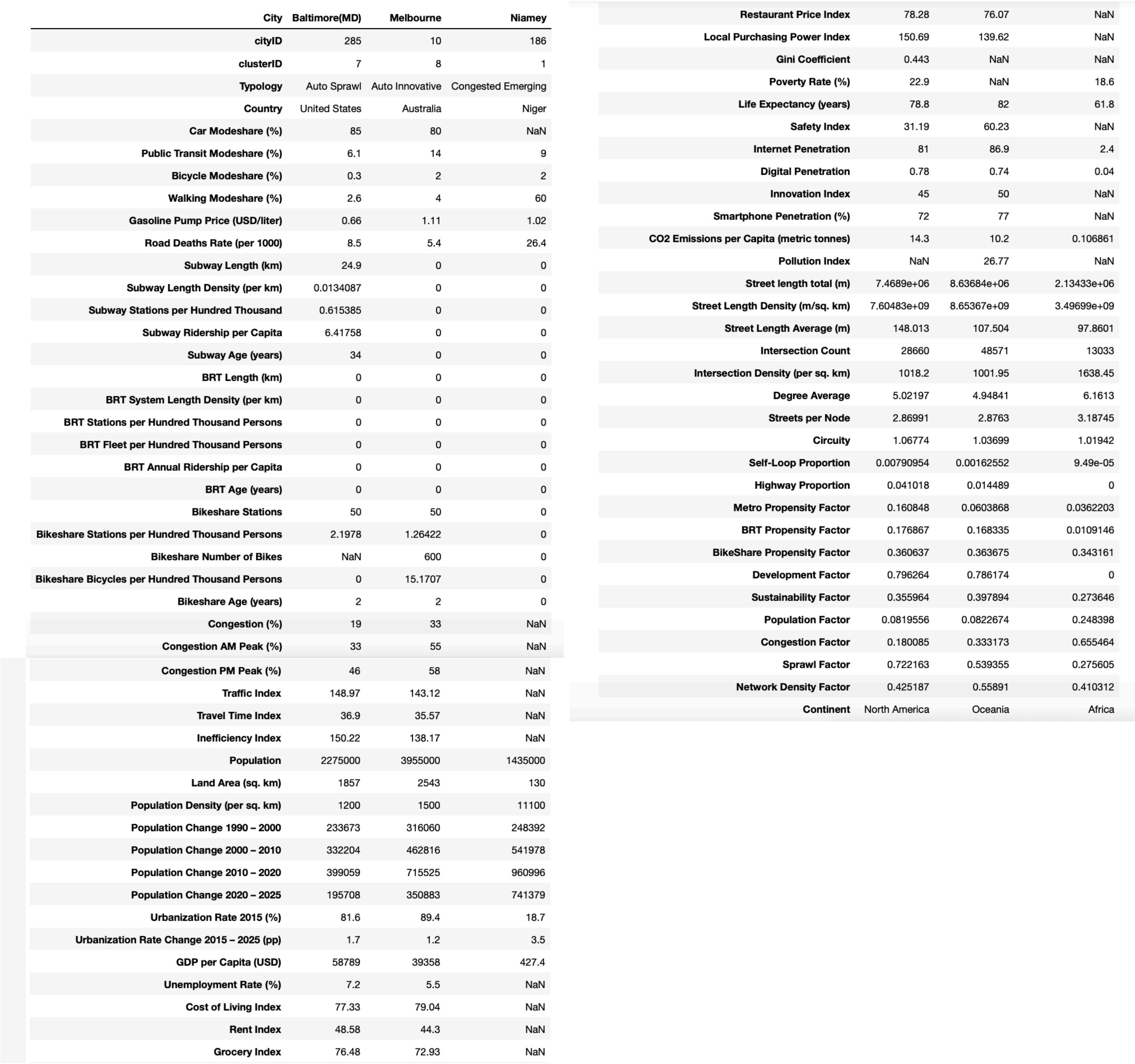
The project has three components:

- *Prediction challenge* - where all groups need to address the same problem (30%)
- *Exploratory component* - where each group is invited to choose their own research question and explore the data accordingly (40%)
- *Report* - Each group should deliver one or more jupyter-notebooks, that should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides the individual comments and reflections (30%)

[1] http://web.mit.edu/afs/athena.mit.edu/org/i/its-lab/www/dashboard/new%20dashboard/index.html

[2] https://www.dropbox.com/sh/1w131yft8tysndx/AADJcrNsu9m4wchBo69gOcBxa?dl=0&preview=0-Summary-of-Indicators-and-Sources.xlsx

Figure 1 shows the variables you will have in this dataset. The data is actually provided as an excel sheet. Notice that some variables require a lot of treatment in order to be usable (e.g. NaNs, categorical, different scales, IDs).



City	Baltimore(MD)	Melbourne	Niamey
cityID	285	10	186
clusterID	7	8	1
Typology	Auto Sprawl	Auto Innovative	Congested Emerging
Country	United States	Australia	Niger
Car Modeshare (%)	85	80	NaN
Public Transit Modeshare (%)	6.1	14	9
Bicycle Modeshare (%)	0.3	2	2
Walking Modeshare (%)	2.6	4	60
Gasoline Pump Price (USD/liter)	0.66	1.11	1.02
Road Deaths Rate (per 1000)	8.5	5.4	26.4
Subway Length (km)	24.9	0	0
Subway Length Density (per km)	0.0134087	0	0
Subway Stations per Hundred Thousand	0.615385	0	0
Subway Ridership per Capita	6.41758	0	0
Subway Age (years)	34	0	0
BRT Length (km)	0	0	0
BRT System Length Density (per km)	0	0	0
BRT Stations per Hundred Thousand Persons	0	0	0
BRT Fleet per Hundred Thousand Persons	0	0	0
BRT Annual Ridership per Capita	0	0	0
BRT Age (years)	0	0	0
Bikeshare Stations	50	50	0
Bikeshare Stations per Hundred Thousand Persons	2.1978	1.26422	0
Bikeshare Number of Bikes	NaN	600	0
Bikeshare Bicycles per Hundred Thousand Persons	0	15.1707	0
Bikeshare Age (years)	2	2	0
Congestion (%)	19	33	NaN
Congestion AM Peak (%)	33	55	NaN
Congestion PM Peak (%)	46	58	NaN
Traffic Index	148.97	143.12	NaN
Travel Time Index	36.9	35.57	NaN
Inefficiency Index	150.22	138.17	NaN
Population	2275000	3955000	1435000
Land Area (sq. km)	1857	2543	130
Population Density (per sq. km)	1200	1500	11100
Population Change 1990 – 2000	233673	316060	248392
Population Change 2000 – 2010	332204	462816	541978
Population Change 2010 – 2020	399059	715525	960996
Population Change 2020 – 2025	195708	350883	741379
Urbanization Rate 2015 (%)	81.6	89.4	18.7
Urbanization Rate Change 2015 – 2025 (pp)	1.7	1.2	3.5
GDP per Capita (USD)	58789	39358	427.4
Unemployment Rate (%)	7.2	5.5	NaN
Cost of Living Index	77.33	79.04	NaN
Rent Index	48.58	44.3	NaN
Grocery Index	76.48	72.93	NaN
Restaurant Price Index	78.28	76.07	NaN
Local Purchasing Power Index	150.69	139.62	NaN
Gini Coefficient	0.443	NaN	NaN
Poverty Rate (%)	22.9	NaN	18.6
Life Expectancy (years)	78.8	82	61.8
Safety Index	31.19	60.23	NaN
Internet Penetration	81	86.9	2.4
Digital Penetration	0.78	0.74	0.04
Innovation Index	45	50	NaN
Smartphone Penetration (%)	72	77	NaN
CO2 Emissions per Capita (metric tonnes)	14.3	10.2	0.106861
Pollution Index	NaN	26.77	NaN
Street length total (m)	7.4689e+06	8.63684e+06	2.13433e+06
Street Length Density (m/sq. km)	7.60483e+09	8.65367e+09	3.49699e+09
Street Length Average (m)	148.013	107.504	97.8601
Intersection Count	28660	48571	13033
Intersection Density (per sq. km)	1018.2	1001.95	1638.45
Degree Average	5.02197	4.94841	6.1613
Streets per Node	2.86991	2.8763	3.18745
Circuity	1.06774	1.03699	1.01942
Self-Loop Proportion	0.00790954	0.00162552	9.49e-05
Highway Proportion	0.041018	0.014489	0
Metro Propensity Factor	0.160848	0.0603868	0.0362203
BRT Propensity Factor	0.176867	0.168335	0.0109146
BikeShare Propensity Factor	0.360637	0.363675	0.343161
Development Factor	0.796264	0.786174	0
Sustainability Factor	0.355964	0.397894	0.273646
Population Factor	0.0819556	0.0822674	0.248398
Congestion Factor	0.180085	0.333173	0.655464
Sprawl Factor	0.722163	0.539355	0.275605
Network Density Factor	0.425187	0.55891	0.410312
Continent	North America	Oceania	Africa

Figure 1. Dataframe view

The <u>*prediction challenge*</u> consists of two parts, in both cases you are expected to **predict the 'CO2 Emissions per Capita (metric tonnes)'** for each city, conditioned on any other variable you choose, **except for "Pollution index** ". The difference between the two parts relies on how you split into train and test sets:

- Part 1 – The training set will correspond to the first 75% rows in the dataset and test set will be the last 25%, **without shuffling**. As a benchmark, we expect you to be able to predict the test set with an R^2 at least 0.60. You can use any sklearn regression model you want, including those not taught in the class.

- Part 2 – The test set shall correspond to all cities that belong to North America and South America, while the train set will be the remaining ones. The idea is for you to experiment (and discuss in the group) with the concept of generalizability/transferability. What would a model need to properly generalize to a very new datapoint? As a benchmark, we expect you to be able to predict the test set with an R^2 at least 0.30. You can use any sklearn regression model you want, including those not taught in the class.

In both part 1 and part 2, if you want to use a development set, you need to extract it from the train set, i.e. you should not change the test set as above proposed.

In the *exploratory component*, each group needs to address at least one new research question. Here, we expect you to formulate your own question, follow the data sciences cycle. The project will be positively valued with one or more of the following extensions:

- Extension of the dataset (preferably using Python APIs) with other relevant data on cities in the world;
- Generation and analysis of insightful visualizations;
- Usage of the breadth of techniques from the class beyond regression and data preparation (e.g. dimensionality reduction, clustering, classification, time series)

Some example research questions:

- Relationship with COVID (e.g. there's a lot of new recent datasets on covid-19, can you find relevant patterns with respect to the cities dataset?)
- Consider climate aspects (e.g. cities in hot or humid areas have different energy requirements than those in cold or dry ones)
- Consider other aspects, such as crime, industrialization, geographic location (near the sea? mountains?)
- Relationships with any other existing indicators (e.g. happiness indexes, freedom of press, country regime, population diversity)

Note: The ordering of tasks we mention is **not** mandatory. In other words, if you prefer to start with the exploratory component, and then go to the prediction challenge, this is very acceptable.

Evaluation

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions

- Honesty - While it's fine to use others' code (as starting point), these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to always reference the source of that code in you used.

Rules

- Each group should consist of 3-4 students. Exceptions are allowed for other forms, but only with strong justification.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the names of the group members (for example, for Pablo, Anders and Mila, it should be Pablo_Anders_Mila.zip).
- In the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defense
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

Report

The report can be in the form of a jupyter notebook. Below is the recommended structure (you are welcome to make your own structure):

- Introduction and problemstatement
- Preliminary data wrangling, cleaning, descriptive statistics
- Part 1: Prediction challenge
- Part 2: Exploratory component
- Reflections/Conclusion

Important dates

- October 5 – Announcement of this challenge statement
- October 19 – Communication of group members (to camara@dtu.dk and rodr@dtu.dk)
- November 15 – Descriptive statistics notebook – this notebook should present preliminary analysis on the dataset, and other datasets obtained by the group, including data preparation, data cleaning, initial analysis of patterns and insights from the data. Submit through CampusNet
- December 6 – Final submission – all materials, including report notebook. Submit through CampusNet