**Assessment outline**

**Pt.I**

Computational essay

* Find a dataset of your interest
* Prepare it for analysis
* Explore the dataset visually, identifying the key patterns.

**Pt.II**

* Perform a clustering exercise & analyse results
* Fit a regression model and:
  + Interpret the coefficients
  + Evaluate its predictive performance both with and without cross-validation
  + Reflect on the differences between the two approaches and the reasons for the potential divergences
* Consider potential predictive improvements of the model through alternative techniques

**Pt.III**

* HTML of Jupyter notebook
* Sumit through Turnitin
* Fully document your code
* Max. 2,000 words (code, comments)
* Deadline: May 15th

**Ideas**

Use the Dohono order (2017) in d1s1 toorganise the report

% Cycling

Age

Education

Slope

Infrastructure

Quietness

Ethnicity

Income

Car ownership

Density

|  |  |  |
| --- | --- | --- |
| **Social and geographical determinants of bicycle commuting in Leeds** | | |
| Research question | How do social and geographical factors influence the use of the bicycle to commute in Leeds? | |
| Unit of analysis | LSOA | |
| Data | Demographics (2011 census) | % of cycling trips to work, age, ethnicity, study level, income, car household, distance to work, health level, transport public… |
| Infrastructure (Open Street Data) | cycle\_path/area, 20\_mph\_area\_m/area, green\_areas\_m/area… |
| Other geographical variables (PCT) | Average slope |
| Scope | Leeds | |
| Cluster analysis | Groups of LSOA/% of trips cycling to work. The cluster should have a justification. For instance if I’m interested in the infrastructure, I can do a cluster in the variables related with friendly infrastructure. | |
| Regression analysis | Variable to predict: % of trips cycling to work  Predictors: Rest of variables | |
| Challenges | Calculate infrastructure indicators per area,  Join the datasets | |