**Thesis Paper Template - Luis**

**Main Theme: Using transactional data to measure health disparities between regions of the UK**

**Main differences with thesis report:**

* No discussion of data extraction from VisaNet
* No Clustering
* No transaction statistics at LAD level (e.g. average or stdev transaction value). These are thought to not be valuable from an analytical point of view as the variance of transactions statistics will be low when comparing LADs.
* No super-resolution transactional data (where features are broken down at market segment **and** enterprise level)
* Avoid using standard deviation as statistic, just average, sum and count to keep things simple.
* Only two “resolutions” for Visa Data [to be discussed]:
  + Low resolution:
    - number of cards in a year per LAD
    - transaction count in a year per LAD
    - sum spent over a year per LAD
    - average sum spent per card in a year per LAD,
    - average transaction count per card in year per LAD
  + High resolution: same stats as above but broken down
    - number of cards
      * ~~per funding source (debit/credit)~~
      * ~~per bank~~
      * activity filter count
    - transaction count
      * ~~attempted~~,
      * declined,
      * ~~contactless~~,
      * online/ecom
      * ~~mobile payment (pays)~~
    - sum spent
      * split into market segments
    - average sum spent per card in a year per LAD,
      * split into market segments
    - average transaction count per card in year per LAD
      * split into market segments
  + **Discussion point:** some features could be spared but which and why?
* **Model could change too, thesis’s model was LASSO but could be different.**

**Template**

1. **Introduction**
   1. Current ways to measure behaviour in Public Health and the problems with it
   2. Transactional data as a continuous proxy for behaviour
   3. Ecological Study
2. **Methods**
   1. **Data Sources – Data Context (England & Wales 2019)**
      1. Visa
         1. Filtering for inactivity
      2. ONS
   2. **Feature extraction**
      1. Aggregate stats by card
      2. Extract LAD stats from card aggregate data (e.g. avg spent per card, number of transactions…)
   3. **Outcomes**
   4. **Model:**
      1. LASSO with CV and subsampling OR sgsPLS (we’ll see)
3. **Results**
   1. **Barplots with several outcomes (# and nature of outcomes to be disccused)**
      1. ONS data as negative control
      2. Visa Low
      3. Visa High
      4. ONS + Visa Low
      5. ONS + Visa High (Visa High contains Visa Low)
   2. Heterogeneity across LADs