**Setup:**

1. Please download the dataset and unzip files into 'bicycle' folder then removing zip file.
2. Install packages: pandas, tqdm
3. Run: python bicycle\_insurnace.py

**1. Check missing files: Will fill the missing value later**

missing files: 2016-02: {'btp-street.csv'}

missing files: 2016-03: {'btp-street.csv'}

missing files: 2018-11: {'kent-street.csv', 'humberside-street.csv'}

**1.1 Build Northern Ireland areas and their coordinate.**

During data cleaning, I discover that Northern Ireland doesn't have LSOA code,

therefore, I used website's information to rebuilt its areas and coordinate.

* Area information: <https://www.police.uk/northern-ireland/>
* approximate coordinate: google map

**2. Load all data into single DataFrame:**

stage0-all\_df.csv

**3. Decide the price should base on which attribute**

The following files show the crime counts based on different area size.

1. **LSOA code:** stage1.0-group\_by\_column\_count\_LSOA\_code\_LSOA\_name.csv, stage1.0-group\_by\_column\_count\_LSOA\_code\_LSOA\_name.png
2. **Area which remove last code in LSOA name:**  stage1.0-group\_by\_column\_count\_area.csv, stage1.0-group\_by\_column\_count\_area.png
3. **File(Police force):** stage1.0-group\_by\_column\_count\_file.csv, stage1.0-group\_by\_column\_count\_file.png

**case a.** there are more than half of LSOA code crime counts which are less than 5 in this two years period.

**case c.** the area is too big, which cannot tell the different between different city or county

**case b.** is the case between case a and c. It has better data distribution.

**4. The area's coordinate is calculated using the mean coordinate of all bicycle theft crime within this area.**

The coordinate is an approximation in order to reduce the search time in this small project.

stage1.1-area\_coor.csv

**5. Check missing value: LSOA code**

1. **LOSA code is Nan:** stage1.2-nan\_losa.csv
2. **LOSA code is Nan, but has coordinate:** stage1.2-nan\_losa\_with\_coordinate.csv

**In case b,** the area can be recovery using mean coordinate, which is built in step 4.

**6. Recovery missing area using mean coordinate**

**7. Check missing value: Crime ID**

1. **Crime ID is Nan:** stage1.3-nan\_crimeid.csv
2. **Crime ID and LSOA code is Nan:** stage1.3-nan\_crimeid\_losa.csv
3. **Crime ID, LSOA code, area is Nan:** stage1.3-nan\_crimeid\_losa\_area.csv
4. **CrimeID, LSOA code, coordinate is Nan:** stage1.3-nan\_crimeid\_losa\_coor.csv

**Case a**. I discover British transport police(BTP) don't use Crime ID.

**Case b.** some LSOA code is empty,

**Case c,** d, because all case in BTP has coordinate, so all case can find its area

**8. Duplicate case:**

1. **Duplicate Crime ID:** stage1.4-dup\_crimeid.csv
2. **Check duplicate based on multiple columns value:** stage1.4-dup\_multi\_columns.csv
3. **Check again after remove duplicate cases:** stage1.4-remove\_dup\_multi\_columns.csv

**case a:** I discovery that in some cases have same crime ID, but may be the different cases based on their Month, location and coordinate.

**case b:** I use more columns ('Crime ID' ,'Reported by' ,'Falls within' ,'Longitude' ,'Latitude' ,'Location' ,'LSOA code' ,'LSOA name' ,'Crime type')

to find duplicate case, these cases can be removed from dataset.

**case c:** after remove duplicate cases, there are no duplicates in dataset

**9. Check missing value: LSOA code and coordinate**

1. **LSOA code and coordinate is Nan:** stage1.5-nan\_losa\_coordinate.csv

In this case, the only information is file(police force). We will distribute these case equally to coresponding file later.

**10. Summary monthly crime count based on each area**

**file:** stage2.0-monthly\_summary.csv

**11. Fill missing value.**

As there are missing files, so I need to interpolate monthly crime count based on previous/later monthly crime count.

I ignore missing BTP as it record the crimes cover whole UK. It make it meaningless using historical data to recover its crime count.

The interpolation method is spline with order 1 for the simplicity.

**file:** stage2.1-monthly\_summary\_fillna.csv

**12. Calculate final monthly bicycle insurace price based on each area using given formula in the email.**

I keep the minimum price to 1£ if bicycle theft keep decreasing or become 0.

**file:** stage2.2-monthly\_price.csv

**Discussion:**

1. Price in some area are very high. This is caused when the areas are very big. I can be solved by spliting into smaller areas.
2. Some decisions aree based on limited resources(time, computing power, scale of project, avaliable information): **(a)** The price results can be improved if I can get(defined) appropriate area boundary, which will be a time consuming task. **(b)** Recovering missing area using approximate area coordinate (mean coordainte), which can reduce search time.
3. Alternative solution: For a given location(coordinate), we can calculate the price based on crime counts within a range(radius).