Introduction

For this CA you will work on the **NIPostcode** and **NICrimeData** datasets. The NIPostcode dataset contains both character and numeric variables for all postcode data in Northern Ireland. And the NICrimeData contains information on crimes for Northern Ireland from January 2015 to December 2017. All datasets are available on Blackboard.

NI postcode dataset

The **NIPostcodes** dataset is constructed of the following data:

- Organisation Name
- Sub-building Name
- Building Name
- Number
- Primary Thorfare
- Alt Thorfare
- Secondary Thorfare
- Locality
- Townland
- Town
- County
- Postcode
- x-coordinates
- y-coordinates
- Primary Key (identifier)

Section 1

Complete the following tasks using the NIPostcodes dataset described above.

a) Show the structure and first 10 rows of the data frame containing all of the NIPostcode data.
b) Show the total number and mean missing values of the NIPostcode data.
c) Remove or replace missing entries with a suitable identifier. Decide whether it is best to remove missing data or to recode it.
d) Add a suitable title for each attribute of the data.
e) Modify the County attribute to be a categorising factor.
f) Move the primary key identifier to the start of the dataset.
g) Create a new dataset called **Limavady_data**. Store within it only information that has **locality**, **townland** and **town** containing the name **Limavady**. Store this information in an external csv file called **Limavady**.
h) Save the modified NIPostcode dataset in a csv file called **CleanNIPostcodeData**.

NI crime dataset

Now you will work with the NI crime dataset. Thirty six datasets contain all of the crime data for Northern Ireland from January 2015 to December 2017 in csv format. All csv files are located in the CA folder in Blackboard in the compressed folder **NICrimeData**. Each dataset is constructed of the following data:

- Crime ID
- Month
- Reported by
- Falls within
- Longitude
- Latitude
- Location
- LSOA code
- LSOA name
- Crime type
- Last outcome
- Context

Section 2

Complete these tasks using the NICrimeData datasets described above.

(a) Using R, amalgamate all of the crime data from each csv file into one dataset. Save this dataset into a csv file called **AllNICrimeData**. Count and show the number of rows in the **AllNICrimeData** dataset.

(b) Modify the structure of the newly created **AllNICrimeData** csv file and remove the following attributes: **CrimeID**, **Reported by**, **Falls within**, **LSOA code**, **LSOA name**, **last outcome** and **context**. Show the structure of the modified file.

(c) Factorise the **Crime type** attribute. Show the modified structure.

(d) Modify the **AllNICrimeData** dataset so that the Location attribute contains only a street name. For example, the attribute value "On or near Westrock Square" should be modified to only contain "Westrock Square". Modify the resultant empty location attributes with a suitable identifier.

(e) Create a function called **find_a_postcode** that takes as an input each **location** attribute from **AllNICrimeData** and finds a suitable postcode value from the postcode dataset. Use the **CleanNIPostcodeData** dataset you created in section 1 as the reference data to find postcodes. If there are several postcodes discovered with the same location, choose the most popular postcode for that location. Store the output from the **find_a_postcode** function in a suitably named variable. Show the structure and number of values in this variable.

(f) Append the data output from your **find_a_postcode** function to the **AllNICrimeData** dataset. Show the modified structure.

(g) Some location data in the **AllNICrimeData** dataset has missing location information eg it contains the identifier you added in task (d). Instead of deleting these from the dataset, create a function called **tidy_location** that takes as an input any data that does not have complete location information. Using longitude and latitude information from each missing record, find a close match to the missing location information. You can use a suitable R function such as the **goecode()** function available within the **ggmap** library to examine and extract relevant location data. Fully document the processes you followed to use this library. **Extract the first 500 records** with missing location information from the **AllNICrimeData** and determine correct location information for this data. Save the 500 records as **CrimeDataWithLocation** and show the first 10 rows of data for this new dataset.

(h) Append the **CrimeDataWithLocation** dataset with new attributes Town, County and Postcode. Use the **NIPostcode** dataset and match the location attribute to perform the join between both datasets. Modify Town and County attributes to become unordered factors. Show the modified **CrimeDataWithLocation** structure.

(i) Save your modified **CrimeDataWithLocation** dataset in a csv file called **FinalNICrimeData**.

Important Information

Use the following table structure to help you put the required steps together for each of the tasks.

Step Description	What does this step require you to do?
Snapshot of data before processing	Show several rows of data before processing it in this step.
Structure of dataset before change	Show the dataset structure.
R Code used to perform change	Show all of the relevant R code including comments.
Snapshot of data after application of change	Show several rows of data after processing it in this step.
Structure of dataset after change	Show the dataset structure.
Result	What has happened to the relevant datasets in this step?

You must store any required datasets such as the postcode data within your R project. Ensure that your code does not refer to local file locations to access your datasets.

Note: Save all of your R scripts on a public repo on GitHub. **Provide a link to this repo in your submitted document.**

Plagiarism will not be accepted and will result in an automatic mark of zero. If you use references, the Harvard referencing must be adopted. Please use the following link which might help you create the references required: http://www.neilstoolbox.com/bibliography-creator/.

Late submissions will not be accepted without a valid medical certificate.

Any deviation from the above project specification must be approved by myself before submission.

Due Date: Sunday 31st March. You should submit your work through Blackboard. A cover sheet should be contained in your submitted document. You must submit your work as a pdf document. The first page of your submission must be the cover page.