LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: Dr **James Connolly**

Assessment Title: **CA 2 - NI postcode and crime data**

Work to be submitted to: Blackboard

Date for submission of work: **April 14th**

Place and time for submitting work: **Home - sending via email as the link to upload is not open**

To be completed by the Student

Student's Name: **Michael McBride (L00143398)**

Class: **LY_KDATA_M: Data Science (2018/19)**

Subject/Module: Data Science

Word Count (where applicable): _____

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: **Michael McBride** Date: _____

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

The working code can be found on the following GITHUB ADDRESS

https://github.com/L00143398/ContinousAssignment_2.git

Step Description	Sec 1 - Step A – Description: Show the structure and first 10 rows of the data frame containing all of the NIPostcode data
Snapshot of dataset before processing	Not Applicable – reading the file is the first action so there is nothing to display before processing

R Code used to perform change	
	`# Reading in the NI Post Code file into the NIPostCodeSource dataframe`
	`# I am assuming that the file "NIPostcodes.csv" is in the current working directory`
	`# Please note I forced all blank spaces to "NA" to make it easier to manipulate as I move forward`
	`# This also take care of the Step C action as rather than dropping those rows I replace with NA`
	
	`NIPostCodeSource <- read.csv(file = "NIPostcodes.csv", header=FALSE, na.strings=c("","NA"))`
	
	`# The following 3 commands provide the row count, the structure and display the first 10 rows of the dataframe`
	`nrow(NIPostCodeSource)`
	`str(NIPostCodeSource)`
	`head(NIPostCodeSource, n =10L)`

Snapshot of data after application of change	`> nrow(NIPostCodeSource)`
	`[1] 943034`
	`> head(NIPostCodeSource, n =10L)`
	`  V1  V2  V3  V4  V5  V6  V7  V8  V9  V10  V11  V12  V13  V14 V15`
	`1 <NA> <NA> <NA> 17 HIGH ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557BG 281855 438598 1`
	`2 <NA> <NA> <NA> 15 CONVENTION AVENUE <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557BW 281892 438228 2`
	`3 <NA> <NA> <NA> 13 STATION ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HH 282306 438587 3`
	`4 <NA> <NA> <NA> 99 OLD COACH ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HW 282419 438387 4`
	`5 <NA> <NA> <NA> 20 BREDA COURT <NA> <NA> <NA> BREDA BELFAST DOWN BT86JB 335367 369985 5`
	`6 <NA> <NA> <NA> 11 UPPER HEATHMOUNT <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557AR 281719 438366 6`
	`7 <NA> <NA> <NA> 86 LEVER ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557EE 282080 438424 7`
	`8 <NA> <NA> <NA> 112 OLD COACH ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HW 282524 438243 8`
	`9 <NA> <NA> <NA> 134 WHITEPARK ROAD <NA> <NA> BALLINTOY BALLINTOY DEMESNE BALLYCASTLE ANTRIM BT546ND 303527 444150 9`
	`10 <NA> FLAT 3 <NA> 16 STATION ROAD <NA> <NA> <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557DA 282128 438612 10`
	`>`

Structure of dataset after change	`> str(NIPostCodeSource)`
	`'data.frame': 943034 obs. of 15 variables:`
	`$ V1 : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ V2 : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ V3 : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ V4 : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ V5 : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ V6 : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ V7 : Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ V8 : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ V9 : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ V10: Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ V11: Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ V12: Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ...`
	`$ V13: int 281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ V14: int 438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`$ V15: int 1 2 3 4 5 6 7 8 9 10 ...`
	`>`

Result	Simply read in the NIPostcodes.csv file and replaced any blanks with NA's into the NIPostCodeSource data frame. As can be seen from the structure, there is no headings and the majority of the data was defaulted to Factor although with that many levels that it isn't directly of value. This dataframe will be used to manipulate a lot of data subsequently.

Step Description	Sec 1 - Step B – Description: Add a suitable title for each attribute of the data

Snapshot of dataset before processing

```
> nrow(NIPostCodeSource)
[1] 943034
> head(NIPostCodeSource, n =10L)
     V1     V2   V3  V4                V5   V6   V7        V8                 V9         V10         V11     V12    V13    V14 V15
1  <NA>   <NA> <NA>  17          HIGH ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557BG 281855 438598   1
2  <NA>   <NA> <NA>  15 CONVENTION AVENUE <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557BW 281892 438228   2
3  <NA>   <NA> <NA>  13       STATION ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HH 282306 438587   3
4  <NA>   <NA> <NA>  99     OLD COACH ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HW 282419 438387   4
5  <NA>   <NA> <NA>  20       BREDA COURT <NA> <NA>      <NA>               BREDA     BELFAST        DOWN  BT86JB 335367 369985   5
6  <NA>   <NA> <NA>  11  UPPER HEATHMOUNT <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557AR 281719 438366   6
7  <NA>   <NA> <NA>  86        LEVER ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557EE 282080 438424   7
8  <NA>   <NA> <NA> 112     OLD COACH ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557HW 282524 438243   8
9  <NA>   <NA> <NA> 134     WHITEPARK ROAD <NA> <NA> BALLINTOY  BALLINTOY DEMESNE BALLYCASTLE      ANTRIM BT546ND 303527 444150   9
10 <NA> FLAT 3 <NA>  16       STATION ROAD <NA> <NA>      <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY BT557DA 282128 438612  10
>
```

R Code used to perform change

```r
#setting the column names
colnames(NIPostCodeSource) <- c("Org_name", "Sub_Building_name", "Building_No", "Number", "Primary_Thorfare",
                                "Alt_Thorfare", "Secondary_Thorfare", "Locality", "Townland", "Town",
                                "County", "Postcode", "x_coord", "y_coord",  "PK")
```

Snapshot of data after application of change

```
> head(NIPostCodeSource, 10)
   Org_name Sub_Building_name Building_No Number  Primary_Thorfare Alt_Thorfare Secondary_Thorfare  Locality           Townland        Town      County Postcode x_coord y_coord PK
1      <NA>             <NA>       <NA>     17          HIGH ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BG  281855  438598  1
2      <NA>             <NA>       <NA>     15 CONVENTION AVENUE        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BW  281892  438228  2
3      <NA>             <NA>       <NA>     13       STATION ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557HH  282306  438587  3
4      <NA>             <NA>       <NA>     99     OLD COACH ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557HW  282419  438387  4
5      <NA>             <NA>       <NA>     20       BREDA COURT        <NA>              <NA>     <NA>               BREDA     BELFAST        DOWN   BT86JB  335367  369985  5
6      <NA>             <NA>       <NA>     11  UPPER HEATHMOUNT        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557AR  281719  438366  6
7      <NA>             <NA>       <NA>     86        LEVER ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557EE  282080  438424  7
8      <NA>             <NA>       <NA>    112     OLD COACH ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557HW  282524  438243  8
9      <NA>             <NA>       <NA>    134     WHITEPARK ROAD        <NA>              <NA> BALLINTOY  BALLINTOY DEMESNE BALLYCASTLE      ANTRIM  BT546ND  303527  444150  9
10     <NA>           FLAT 3       <NA>     16       STATION ROAD        <NA>              <NA>     <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557DA  282128  438612 10
>
```

Structure of dataset after change

```
> str(NIPostCodeSource)
'data.frame':   943034 obs. of  15 variables:
 $ Org_name          : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...
 $ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...
 $ Building_No       : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...
 $ Number            : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...
 $ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...
 $ Alt_Thorfare      : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...
 $ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...
 $ Locality          : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA NA 54 NA ...
 $ Townland          : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...
 $ Town              : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...
 $ County            : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...
 $ Postcode          : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..
 $ x_coord           : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...
 $ y_coord           : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...
 $ PK                : int  1 2 3 4 5 6 7 8 9 10 ...
>
```

Result

As can be seen from the above structure snapshot, the file now has headings in the dataframe. This will allow for more ease of manipulating the data in later steps.

Step Description	Sec 1 - Step C – Description: Remove or replace missing entries with a suitable identifier. Decide whether it is best to remove missing data or to recode it.
Snapshot of dataset before processing	I applied this step when reading in the file by replacing blanks with NA. It did not make sense to remove the missing data as that would have impacted almost the whole file – and there is no logical value that can replace the missing values outside of NA. A Snapshot of the dataset is show above.
R Code used to perform change	`NIPostCodeSource <- read.csv(file = "NIPostcodes.csv", header=FALSE, na.strings=c("","NA"))`

Snapshot of data after application of change	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`$ Org_name          : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ Building_No       : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Number            : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ Alt_Thorfare      : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Locality          : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ Townland          : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ Town              : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ County            : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ Postcode          : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..`
	`$ x_coord           : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ y_coord           : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`$ PK                : int  1 2 3 4 5 6 7 8 9 10 ...`
	`>`

Structure of dataset after change	
Result	Although this step was applied at the time of reading the file the result is that all blank fields are populated with NA.

Step Description	Sec 1 - Step D – Description: Show the total number and mean missing values of the NIPostcode data

Snapshot of dataset before processing	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`$ Org_name          : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ Building_No       : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Number            : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ Alt_Thorfare      : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Locality          : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ Townland          : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ Town              : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ County            : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ Postcode          : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..`
	`$ x_coord           : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ y_coord           : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`$ PK                : int  1 2 3 4 5 6 7 8 9 10 ...`
	`>`

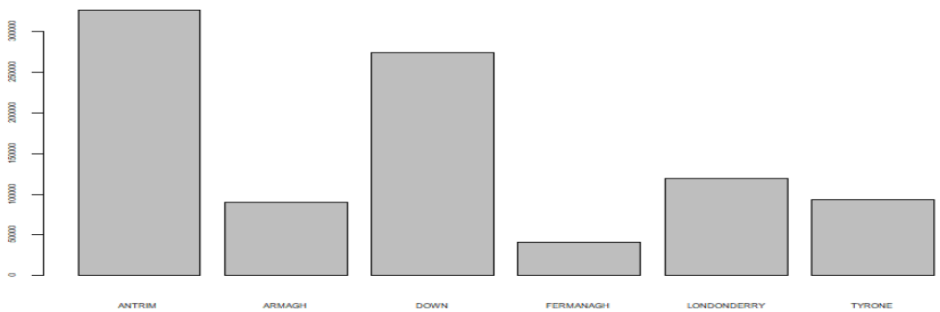
R Code used to perform change	`# The total number and mean of missing values is shown with the summary`
	
	`colSums(is.na(NIPostCodeSource))`
	`colMeans(is.na(NIPostCodeSource))`

Snapshot of data after application of change	`> colSums(is.na(NIPostCodeSource))`
	`PK  Org_name  Sub_Building_name  Building_No  Number  Primary_Thorfare`
	`0  890537  884099  895540  28753  470`
	`Alt_Thorfare  Secondary_Thorfare  Locality  Townland  Town  County`
	`921788  938400  856789  0  19872  0`
	`Postcode  x_coord  y_coord`
	`8900  0  0`
	`> colMeans(is.na(NIPostCodeSource))`
	`PK  Org_name  Sub_Building_name  Building_No  Number  Primary_Thorfare`
	`0.0000000000  0.9443318056  0.9375049044  0.9496370226  0.0304898869  0.0004983914`
	`Alt_Thorfare  Secondary_Thorfare  Locality  Townland  Town  County`
	`0.9774705896  0.9950860732  0.9085451850  0.0000000000  0.0210724110  0.0000000000`
	`Postcode  x_coord  y_coord`
	`0.0094376237  0.0000000000  0.0000000000`
	`>`

Structure of dataset after change	No Change
Result	This command show the number of columns with NA populated and the mean for all columns. This simply reports on the NIPostCodeSource dataframe and does not change it in any way.

Step Description	Sec 1 - Step E – Description: Modify the County attribute to be a categorising factor. There is no need to perform this action as the County attribute is already a Factor but I am displaying the structure showing this and the code that could be used in order to complete this command

Snapshot of dataset before processing	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`$ Org_name          : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ Building_No       : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA NA ...`
	`$ Number            : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ Alt_Thorfare      : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Locality          : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ Townland          : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ Town              : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ County            : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ Postcode          : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..`
	`$ x_coord           : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ y_coord           : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`$ PK                : int  1 2 3 4 5 6 7 8 9 10 ...`
	`>`

R Code used to perform change	`# Setting the County values as a categorizing factor using the as.factor command` `NIPostCodeSource$County <- as.factor((NIPostCodeSource$County))`
Snapshot of data after application of change	
Structure of dataset after change	No change as County was already categorized as a Factor.
Result	Although there was no change in this particular case by changing Factor to county it can be manipulated/used in different ways in R. I Included a Barplot diagram reflecting this.

Step Description	Sec 1 - Step F – Description: Move the primary key identifier to the start of the dataset.

Snapshot of dataset before processing	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`$ Org_name          : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ Building_No       : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA NA ...`
	`$ Number            : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ Alt_Thorfare      : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Locality          : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ Townland          : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ Town              : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ County            : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ Postcode          : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..`
	`$ x_coord           : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ y_coord           : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`$ PK                : int  1 2 3 4 5 6 7 8 9 10 ...`
	`>`

R Code used to perform change	`# moving the primary key to the start of the dataset using the following`
	`# First I do a head to show the order, then I reorder by moving the 15th`
	`# PK, to the first column followed by the next 14`
	
	`NIPostCodeSource<-NIPostCodeSource[,c(15, 1:14)]`

Snapshot of data after application of change	`> head(NIPostCodeSource, n = 2L)`
	`  Org_name Sub_Building_name Building_No Number  Primary_Thorfare Alt_Thorfare Secondary_Thorfare Locality  Townland  Town  County Postcode x_coord y_coord P`
	`1  <NA>  <NA>  <NA>  17  HIGH ROAD  <NA>  <NA>  <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BG  281855  438598`
	`2  <NA>  <NA>  <NA>  15 CONVENTION AVENUE  <NA>  <NA>  <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BW  281892  438228`
	`> NIPostCodeSource<-NIPostCodeSource[,c(15, 1:14)]`
	`> head(NIPostCodeSource, n = 2L)`
	`  PK Org_name Sub_Building_name Building_No Number  Primary_Thorfare Alt_Thorfare Secondary_Thorfare Locality  Townland  Town  County Postcode x_coord y_coo`
	`1 1  <NA>  <NA>  <NA>  17  HIGH ROAD  <NA>  <NA>  <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BG  281855  4385`
	`2 2  <NA>  <NA>  <NA>  15 CONVENTION AVENUE  <NA>  <NA>  <NA> MULLAGHACALL NORTH PORTSTEWART LONDONDERRY  BT557BW  281892  4382`

Structure of dataset after change	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`  $ PK  : int  1 2 3 4 5 6 7 8 9 10 ...`
	`  $ Org_name  : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`  $ Building_No  : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Number  : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`  $ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`  $ Alt_Thorfare  : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Locality  : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`  $ Townland  : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`  $ Town  : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`  $ County  : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`  $ Postcode  : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ...`
	`  $ x_coord  : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`  $ y_coord  : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`

Result	As can be seen from the before and after structure snapshot, the PK field has been moved from the last column in the dataframe to the first value. Other than that there is no change to the dataframe.

Step Description	Sec 1 - Step G – Description: Create a new dataset called Limavady_data.

Snapshot of dataset before processing	The Limavady dataframe does not exist and has yet to be created – so there is no snapshot to display.
	`> nrow(Limavady_Data)`
	`Error in nrow(Limavady_Data) : object 'Limavady_Data' not found`

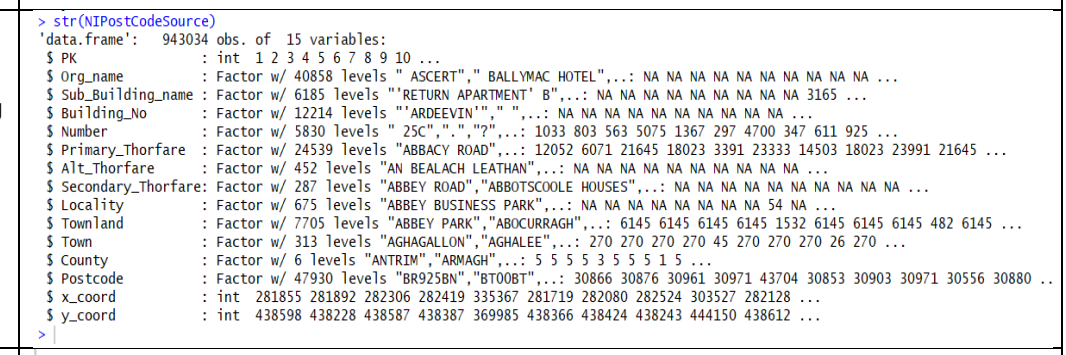
R Code used to perform change	`# To filter out all records that have the text LIMAVADY in either Town, Townland or Locality I had to use the dplyr::filter package`
	`# It wasn't clear on reading the requirements whether the condition was an AND or an OR - so I selected records if LIMAVADY was`
	`# populated in any of the 3 fields.`
	`# In order to use the dplyr function I installed the package and then called the library and ran the filter to populate Limavady_Data`
	
	`install.packages("dplyr")`
	`library(dplyr)`
	`Limavady_Data <- dplyr::filter(NIPostCodeSource, grepl('LIMAVADY', Town) | grepl('LIMAVADY', Townland) | grepl('LIMAVADY', Locality))`
	`nrow(Limavady_Data)`
	
	`# After creating Limavady data I run these commands to verify the population and then write the data out to Limavady.csv file`
	`head(Limavady_Data, n = 2L)`
	`str(Limavady_Data)`
	`write.csv(Limavady_Data, "Limavady.csv")`

Snapshot of data after application of change	`> nrow(Limavady_Data)`
	`[1] 8468`
	`> head(Limavady_Data, n = 2L)`
	`  PK Org_name Sub_Building_name Building_No Number Primary_Thorfare Alt_Thorfare Secondary_Thorfare  Locality  Townland  Town  County Postcode x_coord y_coord`
	`1 73  <NA>  <NA>  <NA>  30  LEIGHERY ROAD  <NA>  <NA>  <NA> BALLYLEIGHERY UPPER LIMAVADY LONDONDERRY  BT490JG  270404  431604`
	`2 75  <NA>  <NA>  <NA>  38  CURRAGH ROAD  <NA>  <NA> BALLYSCULLION  DUNCRUN LIMAVADY LONDONDERRY  BT490JE  268543  432534`

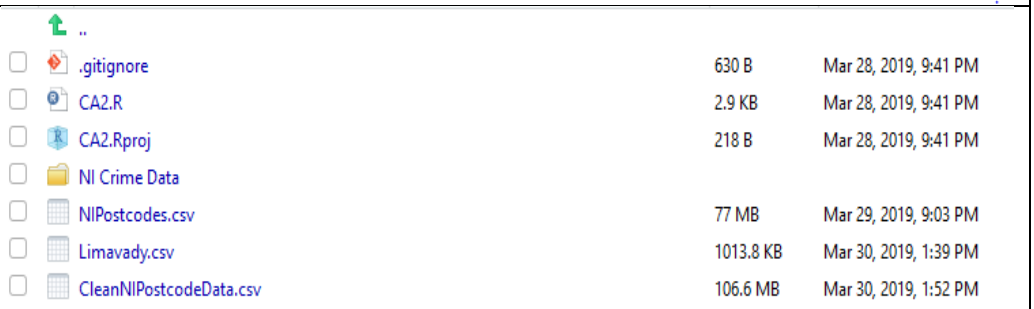
Structure of dataset after change	`> str(Limavady_Data)`
	`'data.frame':  8468 obs. of  15 variables:`
	`  $ PK  : int  73 75 76 89 92 99 226 236 310 455 ...`
	`  $ Org_name  : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 2336 ...`
	`  $ Building_No  : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA 10541 NA NA NA NA NA ...`
	`  $ Number  : Factor w/ 5830 levels " 25C",".","?",..: 2286 2773 2945 2721 NA 1491 1905 82 3271 3406 ...`
	`  $ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 14410 7034 20504 20504 20504 8015 20504 7824 7824 15043 ...`
	`  $ Alt_Thorfare  : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`  $ Locality  : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA 102 NA NA NA NA NA 308 308 NA ...`
	`  $ Townland  : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 830 3758 2199 1192 6386 3290 1362 1667 360 6345 ...`
	`  $ Town  : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 205 205 205 205 205 205 205 205 205 205 ...`
	`  $ County  : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 5 5 5 5 5 5 ...`
	`  $ Postcode  : Factor w/ 47930 levels "BR925BN","BT00BT",..: 28066 28064 28094 28078 28076 28092 28060 28054 28054 28033 ...`
	`  $ x_coord  : int  270404 268543 268640 267172 266864 267148 266391 267029 268259 266899 ...`
	`  $ y_coord  : int  431604 432534 433777 432608 431456 432852 429886 425935 424813 423197 ...`
	`>`

Result	Created a new dataframe called Limavady_data which had 8,468 records as I used the OR condition to select records if Limavady was in any of the 3 fields Town, Townland or Locality. I then wrote the results out to the Limavady.csv file on the current directory.

Step Description	Sec 1 - Step H – Description: Save the modified NIPostcode dataset in a csv file called CleanNIPostcodeData.

Snapshot of dataset before processing	
	`> str(NIPostCodeSource)`
	`'data.frame':  943034 obs. of  15 variables:`
	`$ PK  : int  1 2 3 4 5 6 7 8 9 10 ...`
	`$ Org_name  : Factor w/ 40858 levels " ASCERT"," BALLYMAC HOTEL",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Sub_Building_name : Factor w/ 6185 levels "'RETURN APARTMENT' B",..: NA NA NA NA NA NA NA NA NA 3165 ...`
	`$ Building_No  : Factor w/ 12214 levels "'ARDEEVIN'"," ",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Number  : Factor w/ 5830 levels " 25C",".","?",..: 1033 803 563 5075 1367 297 4700 347 611 925 ...`
	`$ Primary_Thorfare  : Factor w/ 24539 levels "ABBACY ROAD",..: 12052 6071 21645 18023 3391 23333 14503 18023 23991 21645 ...`
	`$ Alt_Thorfare  : Factor w/ 452 levels "AN BEALACH LEATHAN",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Secondary_Thorfare: Factor w/ 287 levels "ABBEY ROAD","ABBOTSCOOLE HOUSES",..: NA NA NA NA NA NA NA NA NA NA ...`
	`$ Locality  : Factor w/ 675 levels "ABBEY BUSINESS PARK",..: NA NA NA NA NA NA NA NA 54 NA ...`
	`$ Townland  : Factor w/ 7705 levels "ABBEY PARK","ABOCURRAGH",..: 6145 6145 6145 6145 1532 6145 6145 6145 482 6145 ...`
	`$ Town  : Factor w/ 313 levels "AGHAGALLON","AGHALEE",..: 270 270 270 270 45 270 270 270 26 270 ...`
	`$ County  : Factor w/ 6 levels "ANTRIM","ARMAGH",..: 5 5 5 5 3 5 5 5 1 5 ...`
	`$ Postcode  : Factor w/ 47930 levels "BR925BN","BT00BT",..: 30866 30876 30961 30971 43704 30853 30903 30971 30556 30880 ..`
	`$ x_coord  : int  281855 281892 282306 282419 335367 281719 282080 282524 303527 282128 ...`
	`$ y_coord  : int  438598 438228 438587 438387 369985 438366 438424 438243 444150 438612 ...`
	`>`

R Code used to perform change	`# Write out the NIPostCodeSource dataframe to CleanNIPostcodeData.csv on the current directory`
	
	`write.csv(NIPostCodeSource, "CleanNIPostcodeData.csv")`

Snapshot of data after application of change	 .. .gitignore 630 B Mar 28, 2019, 9:41 PM CA2.R 2.9 KB Mar 28, 2019, 9:41 PM CA2.Rproj 218 B Mar 28, 2019, 9:41 PM NI Crime Data NIPostcodes.csv 77 MB Mar 29, 2019, 9:03 PM Limavady.csv 1013.8 KB Mar 30, 2019, 1:39 PM CleanNIPostcodeData.csv 106.6 MB Mar 30, 2019, 1:52 PM
Structure of dataset after change	No change on the dataset
Result	This step was simply to write out the Clean NI Postcode data to a csv file. Above showing the files that were created during these steps including Limavady.csv and CleanNIPostcodeData.csv

Step Description	Sec 2 - Step A – Description: amalgamate all the crime data from each csv file into one dataset.
Snapshot of dataset before processing	I manually copied all of the CSV files into a single directory called “NI Crime Data” under the current directory and started processing the files from that point.

R Code used to perform change	
	`# In this step I navigate to the sub-directory NI Crime Data where all the csv files should be stored.`
	`getwd()`
	`setwd("NI Crime Data")`
	`getwd()`
	
	`# In the next steps read all of the files listed in the sub-directory NI Crime Data into a file called "crime_file_names"`
	`# This file is then fed into the rbind command which reads all of the seperte files into the AllNICrimeData variable.`
	`# The key to these 2 steps is to ensure that all files that you wish to read are in the same directory -`
	`# which I did by manually putting them into one directory`
	
	`crime_file_names <- list.files(full.names=TRUE)`
	`AllNICrimeData <- do.call(rbind,lapply(crime_file_names,read.csv))`
	`nrow(AllNICrimeData)`

Snapshot of data after application of change	`> head(AllNICrimeData, 5)`
	`  Crime.ID  Month  Reported.by  Falls.within Longitude Latitude  Location LSOA.code`
	`1  2015-01 Police Service of Northern Ireland Police Service of Northern Ireland -6.003289 54.55165  On or near Salisbury Place  NA`
	`2  2015-01 Police Service of Northern Ireland Police Service of Northern Ireland -5.707979 54.59231  On or near  NA`
	`3  2015-01 Police Service of Northern Ireland Police Service of Northern Ireland -5.815976 54.73161  On or near Milebush Park  NA`
	`4  2015-01 Police Service of Northern Ireland Police Service of Northern Ireland -6.393411 54.19788 On or near College Square North  NA`
	`5  2015-01 Police Service of Northern Ireland Police Service of Northern Ireland -6.251798 54.85970  On or near Staffa Drive  NA`
	`  LSOA.name  Crime.type Last.outcome.category Context`
	`1  NA Anti-social behaviour  NA  NA`
	`2  NA Anti-social behaviour  NA  NA`
	`3  NA Anti-social behaviour  NA  NA`
	`4  NA Anti-social behaviour  NA  NA`
	`5  NA Anti-social behaviour  NA  NA`
	
	`> nrow(AllNICrimeData)`
	`[1] 477696`
	`>`

Structure of dataset after change	`> str(AllNICrimeData)`
	`'data.frame':  477696 obs. of  12 variables:`
	`  $ Crime.ID  : Factor w/ 11667 levels "","0009d3218c478283888080303fed14c46c61e6b3b8f55963a2671dac3afb3907",..: 1 1`
	`  $ Month  : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`  $ Reported.by  : Factor w/ 1 level "Police Service of Northern Ireland": 1 1 1 1 1 1 1 1 1 1 ...`
	`  $ Falls.within  : Factor w/ 1 level "Police Service of Northern Ireland": 1 1 1 1 1 1 1 1 1 1 ...`
	`  $ Longitude  : num  -6 -5.71 -5.82 -6.39 -6.25 ...`
	`  $ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...`
	`  $ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...`
	`  $ LSOA.code  : logi  NA NA NA NA NA NA ...`
	`  $ LSOA.name  : logi  NA NA NA NA NA NA ...`
	`  $ Crime.type  : Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`  $ Last.outcome.category: logi  NA NA NA NA NA NA ...`
	`  $ Context  : logi  NA NA NA NA NA NA ...`

Result	The result of these steps is to create a dataframe AllNICrimeData that contains the combined data from all 36 csv crime statistics files. The total row count is 477,696. The 2 steps I followed was to use the list.files command to create a list of all of the csv files to be read in and stored this in crime_file_names. I then fed this into the rbind.lapply command which reads all of the data into a single datafrom.

Step Description	Sec 2 - Step B – Description: Modify the structure of the newly created AllNICrimeDatacsv file

Snapshot of dataset before processing	`> str(AllNICrimeData)`
	`'data.frame':  477696 obs. of  12 variables:`
	`$ Crime.ID  : Factor w/ 11667 levels "","0009d3218c478283888080303fed14c46c61e6b3b8f55963a2671dac3afb3907",..: 1 1`
	`$ Month  : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Reported.by  : Factor w/ 1 level "Police Service of Northern Ireland": 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Falls.within  : Factor w/ 1 level "Police Service of Northern Ireland": 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Longitude  : num  -6 -5.71 -5.82 -6.39 -6.25 ...`
	`$ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...`
	`$ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...`
	`$ LSOA.code  : logi  NA NA NA NA NA NA ...`
	`$ LSOA.name  : logi  NA NA NA NA NA NA ...`
	`$ Crime.type  : Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Last.outcome.category: logi  NA NA NA NA NA NA ...`
	`$ Context  : logi  NA NA NA NA NA NA ...`

R Code used to perform change	`# Removing columns from AllNICrimeData that we do not want - by using the -c command and subset as below.` `AllNICrimeData = subset(AllNICrimeData, select = -c(Crime.ID, Reported.by, Falls.within, LSOA.code, LSOA.name, Last.outcome.category, Context) )`

Snapshot of data after application of change	`> head(AllNICrimeData, n=10)`
	`  Month Longitude Latitude  Location  Crime.type`
	`1  2015-01 -6.003289 54.55165  On or near Salisbury Place Anti-social behaviour`
	`2  2015-01 -5.707979 54.59231  On or near  Anti-social behaviour`
	`3  2015-01 -5.815976 54.73161  On or near Milebush Park Anti-social behaviour`
	`4  2015-01 -6.393411 54.19788 On or near College Square North Anti-social behaviour`
	`5  2015-01 -6.251798 54.85970  On or near Staffa Drive Anti-social behaviour`
	`6  2015-01 -7.206893 54.62265  On or near Killyclogher Road Anti-social behaviour`
	`7  2015-01 -5.915793 54.59242  On or near Ravenhill Reach Anti-social behaviour`
	`8  2015-01 -5.535389 54.48792  On or near  Anti-social behaviour`
	`9  2015-01 -7.322812 54.99940  On or near Great James Street Anti-social behaviour`
	`10 2015-01 -5.954670 54.61568  On or near Jamaica Road Anti-social behaviour`

Structure of dataset after change	`> str(AllNICrimeData)`
	`'data.frame':  477696 obs. of  5 variables:`
	`$ Month  : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Longitude : num  -6 -5.71 -5.82 -6.39 -6.25 ...`
	`$ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...`
	`$ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...`
	`$ Crime.type: Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`>`

Result	In this step I have dropped a number of columns that are not required for further processing form the AllNICrimeData dataframe as can be seen from the Structure of the before and after shown above. The question requests to show the structure of the modified file but I understood this to mean the structure of the modified dataframe.

Step Description	Sec 2 - Step C – Description: Factorise the Crime type attribute. Show the modified structure.

Snapshot of dataset before processing	`> str(AllNICrimeData)`
	`'data.frame':  477696 obs. of  5 variables:`
	`$ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Longitude : num  -6 -5.71 -5.82 -6.39 -6.25 ...`
	`$ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...`
	`$ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...`
	`$ Crime.type: Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`>`

R Code used to perform change	`# Setting Crime Type as a factor` `AllNICrimeData$Crime.type <- as.factor((AllNICrimeData$Crime.type))`
Snapshot of data after application of change	As the Crime.type value was already Factorized there is no need for this step – as can be seen from the snapshot above

Structure of dataset after change	`> str(AllNICrimeData)`
	`'data.frame':  477696 obs. of  5 variables:`
	`$ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`$ Longitude : num  -6 -5.71 -5.82 -6.39 -6.25 ...`
	`$ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...`
	`$ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...`
	`$ Crime.type: Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...`
	`>`

Result	No real change as a result of this step – the crime.type attribute was already Factorized and although I re-ran the command it remains as is

Step Description	Sec 2 - Step D – Description: Modify the AllNICrimeDatadataset so that the Location attribute contains only a street name
Snapshot of dataset before processing	

```
> head(AllNICrimeData, n=10)
     Month Longitude Latitude                        Location            Crime.type
1  2015-01 -6.003289 54.55165      On or near Salisbury Place Anti-social behaviour
2  2015-01 -5.707979 54.59231                    On or near  Anti-social behaviour
3  2015-01 -5.815976 54.73161        On or near Milebush Park Anti-social behaviour
4  2015-01 -6.393411 54.19788 On or near College Square North Anti-social behaviour
5  2015-01 -6.251798 54.85970         On or near Staffa Drive Anti-social behaviour
6  2015-01 -7.206893 54.62265    On or near Killyclogher Road Anti-social behaviour
7  2015-01 -5.915793 54.59242      On or near Ravenhill Reach Anti-social behaviour
8  2015-01 -5.535389 54.48792                    On or near  Anti-social behaviour
9  2015-01 -7.322812 54.99940   On or near Great James Street Anti-social behaviour
10 2015-01 -5.954670 54.61568         On or near Jamaica Road Anti-social behaviour
>
```

R Code used to perform change

```
# Removing the text 'On or near ' from the Location value in AllNICrimeData
AllNICrimeData$Location <- gsub('On or near ', '', AllNICrimeData$Location)

# For the Location field, populating 'NA' for all fields that are blank.
# This will allow us to the filter/identify records where the location is blank.

AllNICrimeData$Location <- replace(AllNICrimeData$Location, AllNICrimeData$Location == '', NA)
head(AllNICrimeData, n=10L)
```

Snapshot of data after application of change

```
> AllNICrimeData$Location <- replace(AllNICrimeData$Location, AllNICrimeData$Location == '', NA)
> head(AllNICrimeData, n=10L)
     Month Longitude Latitude             Location            Crime.type
1  2015-01 -6.003289 54.55165      Salisbury Place Anti-social behaviour
2  2015-01 -5.707979 54.59231                 <NA> Anti-social behaviour
3  2015-01 -5.815976 54.73161        Milebush Park Anti-social behaviour
4  2015-01 -6.393411 54.19788 College Square North Anti-social behaviour
5  2015-01 -6.251798 54.85970         Staffa Drive Anti-social behaviour
6  2015-01 -7.206893 54.62265    Killyclogher Road Anti-social behaviour
7  2015-01 -5.915793 54.59242      Ravenhill Reach Anti-social behaviour
8  2015-01 -5.535389 54.48792                 <NA> Anti-social behaviour
9  2015-01 -7.322812 54.99940   Great James Street Anti-social behaviour
10 2015-01 -5.954670 54.61568         Jamaica Road Anti-social behaviour
```

Structure of dataset after change

```
> str(AllNICrimeData)
'data.frame':   477696 obs. of  5 variables:
 $ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Longitude : num  -6 -5.71 -5.82 -6.39 -6.25 ...
 $ Latitude  : num  54.6 54.6 54.7 54.2 54.9 ...
 $ Location  : Factor w/ 14984 levels "No Location",..: 3507 2 2790 1022 3724 2263 3363 2 1909 2179 ...
 $ Crime.type: Factor w/ 14 levels "Anti-social behaviour",..: 1 1 1 1 1 1 1 1 1 1 ...
>
```

Result

This step removed the text “On or near “ from the Location field. This step will allow us to use this filed to compare against the Postcode file to find the Postcode.
Another result is that some of the Locations being blank – but this was handled by replacing blanks with <NA>’s – as can be seen from the head command above.

Step Description	Sec 2 - Step E – Description: Choose 1000 random samples of crime data from the AllNICrimeData. Then create a function called find a postcode that takes as an input each location attribute from random crime sample and finds a suitable postcode value from the postcode dataset.

Snapshot of dataset before processing	The dataframe random_crime_sample did not exist – after creating it in the first step below we had the following dataset of 1000 records:
	`> random_crime_sample <- AllNICrimeData[ sample( which(AllNICrimeData$Location !='NA'), 1000 ), ]`
	`> random_crime_sample = as.data.frame(sapply(random_crime_sample, toupper))`
	`> nrow(random_crime_sample)`
	`[1] 1000`
	`> head(random_crime_sample, 10)`
	`  Month Longitude Latitude Location Crime.type`
	`1 2017-01 -5.912124 54.675554 HILLVIEW AVENUE ANTI-SOCIAL BEHAVIOUR`
	`2 2015-03 -5.985594 54.756053 HENRYVILLE MANOR CRIMINAL DAMAGE AND ARSON`
	`3 2015-01 -5.92591 54.598233 VICTORIA SQUARE VIOLENCE AND SEXUAL OFFENCES`
	`4 2015-06 -6.349373 54.170532 BARCROFT PARK CRIMINAL DAMAGE AND ARSON`
	`5 2016-01 -6.336651 54.462918 EDWARD STREET VIOLENCE AND SEXUAL OFFENCES`
	`6 2015-08 -5.874591 54.60365 EDGCUMBE GARDENS SHOPLIFTING`
	`7 2016-01 -6.230057 54.480218 CASTLEVUE GARDENS OTHER THEFT`
	`8 2016-03 -7.633041 54.344127 BELMORE STREET ANTI-SOCIAL BEHAVIOUR`
	`9 2017-12 -7.329566 54.993601 STANLEYS WALK CRIMINAL DAMAGE AND ARSON`
	`10 2015-12 -5.919951 54.647108 SHORE ROAD VIOLENCE AND SEXUAL OFFENCES`

R Code used to perform change	`# select 1000 random records from AllNICrimeData where the Location is not NA - using the following command.`
	`# I also used sapply to turn all the text to upper class - to allow for a cleaner comparison between the NI Postcode file`
	`random_crime_sample <- AllNICrimeData[ sample( which(AllNICrimeData$Location !='NA'), 1000 ), ]`
	`random_crime_sample = as.data.frame(sapply(random_crime_sample, toupper))`
	`nrow(random_crime_sample)`
	`head(random_crime_sample, 10)`
	Function code
	`# Function to find a post code based on the location in the crime_data file.`
	`# Firstly I read in the CleanNIPostcodeData.csv file and then I remove all rows to leave only the Primary_Thorfare and Postcode`
	`# I then populated most_frequent_postCode with the most frequent postcode found for the same Primary_Thorfare -`
	`# because you can have multiple different throughfare values`
	`# These 3 commands leave me with a list of street locations and their corresponding postcodes taking the most populate postcode`
	`# which will allow me to compare and populate the appropriate postcode by matching with the crime file`
	`find_a_postcode <- function(crime_data){`
	`  new_CleanNIPostCode <- read.csv(file = "CleanNIPostcodeData.csv", header=TRUE, na.strings=c("","NA"))`
	`  new_CleanNIPostCode = subset(new_CleanNIPostCode, select = -c(PK, Org_name, Sub_Building_name, Building_No, Number, Alt_Thorfare, Secondary_Thorfare,`
	`  most_frequent_PostCode <- new_CleanNIPostCode %>% group_by(Primary_Thorfare) %>% summarize(Postcode =names(which.max(table(Postcode))))`
	`  # In this next step I do a left join on the crime file against the most frequnet postcode by joining the Location and Primary_Thorfare.`
	`  # This appends the Postcode to the crime file based on the street address.`
	`  # In this case I put the results into match_result and the removed any records where the Postcode are NA which indicates that a match was not found`
	`  match_result <- dplyr::left_join(crime_data, most_frequent_PostCode, by=c("Location" = "Primary_Thorfare"))`
	`  match_result <- match_result[!is.na(match_result$Postcode), ]`
	`  return(match_result)`
	`}`
	Calling the function
	`# Calling the find_a_postcode function and passing the crime sample d`
	`# Populating the results into the crime_data_with_postcode dataframe.`
	`crime_data_with_postcode <- find_a_postcode(random_crime_sample)`

Snapshot of data after application of change	`> nrow(crime_data_with_postcode)`
	`[1] 982`
	`> head(crime_data_with_postcode, 10)`
	`  Month Longitude Latitude Location Crime.type Postcode`
	`1 2017-01 -5.912124 54.675554 HILLVIEW AVENUE ANTI-SOCIAL BEHAVIOUR BT274PP`
	`2 2015-03 -5.985594 54.756053 HENRYVILLE MANOR CRIMINAL DAMAGE AND ARSON BT399FP`
	`3 2015-01 -5.92591 54.598233 VICTORIA SQUARE VIOLENCE AND SEXUAL OFFENCES BT14QG`
	`4 2015-06 -6.349373 54.170532 BARCROFT PARK CRIMINAL DAMAGE AND ARSON BT358EW`
	`5 2016-01 -6.336651 54.462918 EDWARD STREET VIOLENCE AND SEXUAL OFFENCES BT666DD`
	`6 2015-08 -5.874591 54.60365 EDGCUMBE GARDENS SHOPLIFTING BT42EG`
	`7 2016-01 -6.230057 54.480218 CASTLEVUE GARDENS OTHER THEFT BT670JU`
	`8 2016-03 -7.633041 54.344127 BELMORE STREET ANTI-SOCIAL BEHAVIOUR BT746AA`
	`9 2017-12 -7.329566 54.993601 STANLEYS WALK CRIMINAL DAMAGE AND ARSON BT489HH`
	`10 2015-12 -5.919951 54.647108 SHORE ROAD VIOLENCE AND SEXUAL OFFENCES BT413NW`
	`>`

Structure of dataset after change	<pre>> str(crime_data_with_postcode)
	'data.frame': 982 obs. of 6 variables:
	$ Month : Factor w/ 36 levels "2015-01","2015-02",..: 25 3 1 6 13 8 13 15 36 12 ...
	$ Longitude : Factor w/ 937 levels "-5.450768","-5.486587",..: 204 432 246 641 622 142 546 919 877 225 .
	$ Latitude : Factor w/ 938 levels "54.062792","54.063553",..: 693 759 477 20 206 533 218 104 843 627 ..
	$ Location : chr "HILLVIEW AVENUE" "HENRYVILLE MANOR" "VICTORIA SQUARE" "BARCROFT PARK" ...
	$ Crime.type: Factor w/ 14 levels "ANTI-SOCIAL BEHAVIOUR",..: 1 4 14 4 14 11 7 1 4 14 ...
	$ Postcode : chr "BT274PP" "BT399FP" "BT14QG" "BT358EW" ...
	></pre>

Result	In this step we created a sampling of 1,000 records from the AllNICrimeData into the random_crime_sample dataframe. This selection ensured to exclude records with Location of NA. Then we passed this random_crime_sample into the function find_a_postcode. In this function we first read in the CleanNIPostCode.csv file, which will use to find the postcodes. But first this dataframe had to be stripped of all attributes that were not required leaving only Primary_Thorfare and Postcode combinations. The next step was to create a unique Primary_Thorfare and associate the most frequent Postcode associated with this location – so that we now have a clean file which contains a unique combination of Primary_Thorfare (or Location) and Postcode. We then did a left join into the match_result dataframe to select a postcode for every Location from the Crime data file. The final step is to drop any records that have no match for Location and for which there is no Postcode and then we return the results to the crime_data_with_postcode dataframe. Please note that because the select of the 1,000 records is random it is possible that the number of records returned into crime_data_with_postcode can also be 1,000 but most likely it will be less due to those locations for which a Postcode match was not found. In this example only 982 locations were matched with a Postcode

Step Description	Sec 2 - Step F – Description:

Snapshot of dataset before processing

```
> str(random_crime_sample)
'data.frame':   1000 obs. of  5 variables:
 $ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 25 3 1 6 13 8 13 15 36 12 ...
 $ Longitude : Factor w/ 937 levels "-5.450768","-5.486587",..: 204 432 246 641 622 142 546 919 877 225 ...
 $ Latitude  : Factor w/ 938 levels "54.062792","54.063553",..: 693 759 477 20 206 533 218 104 843 627 ...
 $ Location  : Factor w/ 829 levels "ABBEY GARDENS",..: 399 387 781 71 288 285 161 92 718 697 ...
 $ Crime.type: Factor w/ 14 levels "ANTI-SOCIAL BEHAVIOUR",..: 1 4 14 4 14 11 7 1 4 14 ...
> nrow(random_crime_sample)
[1] 1000
> head(random_crime_sample, 10)
     Month Longitude  Latitude          Location                   Crime.type
1  2017-01 -5.912124 54.675554   HILLVIEW AVENUE        ANTI-SOCIAL BEHAVIOUR
2  2015-03 -5.985594 54.756053  HENRYVILLE MANOR    CRIMINAL DAMAGE AND ARSON
3  2015-01  -5.92591 54.598233   VICTORIA SQUARE VIOLENCE AND SEXUAL OFFENCES
4  2015-06 -6.349373 54.170532     BARCROFT PARK    CRIMINAL DAMAGE AND ARSON
5  2016-01 -6.336651 54.462918     EDWARD STREET VIOLENCE AND SEXUAL OFFENCES
6  2015-08 -5.874591  54.60365  EDGCUMBE GARDENS                  SHOPLIFTING
7  2016-01 -6.230057 54.480218 CASTLEVUE GARDENS                  OTHER THEFT
8  2016-03 -7.633041 54.344127    BELMORE STREET        ANTI-SOCIAL BEHAVIOUR
9  2017-12 -7.329566 54.993601     STANLEYS WALK    CRIMINAL DAMAGE AND ARSON
10 2015-12 -5.919951 54.647108        SHORE ROAD VIOLENCE AND SEXUAL OFFENCES
>
```

R Code used to perform change

```
library(plyr)
random_crime_sample <- rbind.fill(random_crime_sample, crime_data_with_postcode)
write.csv(random_crime_sample, "random_crime_sample.csv")
```

Snapshot of data after application of change

```
> nrow(random_crime_sample)
[1] 1982
> head(random_crime_sample, 10)
     Month Longitude  Latitude          Location                   Crime.type Postcode
1  2017-01 -5.912124 54.675554   HILLVIEW AVENUE        ANTI-SOCIAL BEHAVIOUR     <NA>
2  2015-03 -5.985594 54.756053  HENRYVILLE MANOR    CRIMINAL DAMAGE AND ARSON     <NA>
3  2015-01  -5.92591 54.598233   VICTORIA SQUARE VIOLENCE AND SEXUAL OFFENCES     <NA>
4  2015-06 -6.349373 54.170532     BARCROFT PARK    CRIMINAL DAMAGE AND ARSON     <NA>
5  2016-01 -6.336651 54.462918     EDWARD STREET VIOLENCE AND SEXUAL OFFENCES     <NA>
6  2015-08 -5.874591  54.60365  EDGCUMBE GARDENS                  SHOPLIFTING     <NA>
7  2016-01 -6.230057 54.480218 CASTLEVUE GARDENS                  OTHER THEFT     <NA>
8  2016-03 -7.633041 54.344127    BELMORE STREET        ANTI-SOCIAL BEHAVIOUR     <NA>
9  2017-12 -7.329566 54.993601     STANLEYS WALK    CRIMINAL DAMAGE AND ARSON     <NA>
10 2015-12 -5.919951 54.647108        SHORE ROAD VIOLENCE AND SEXUAL OFFENCES     <NA>
> tail(random_crime_sample, 10)
       Month Longitude  Latitude             Location                 Crime.type Postcode
1973 2017-08 -5.935396 54.589374          DONEGALL ROAD                OTHER THEFT  BT126HN
1974 2016-05 -6.346952  54.17288            DORANS HILL CRIMINAL DAMAGE AND ARSON  BT358EJ
1975 2015-11 -6.262407 54.854913 CHICHESTER PARK EAST                   BURGLARY  BT424BH
1976 2017-05 -5.892509 54.213154            SHIMNA VALE     POSSESSION OF WEAPONS  BT330EF
1977 2016-05   -6.6422 54.552689       BALLYGITTLE ROAD     ANTI-SOCIAL BEHAVIOUR  BT715JS
1978 2016-03 -5.890862 54.214806              PARK LANE     ANTI-SOCIAL BEHAVIOUR  BT247PR
1979 2016-01 -5.879786 54.589472          DUNRAVEN PARK     ANTI-SOCIAL BEHAVIOUR   BT56BS
1980 2016-05 -5.926828 54.626138        SEAMOUNT PARADE     ANTI-SOCIAL BEHAVIOUR  BT153NS
1981 2017-07 -6.356031 54.179878    SPRINGFARM HEIGHTS                     DRUGS  BT358XA
1982 2017-11 -5.821693 54.718528               KILLALOE                OTHER THEFT  BT388FL
>
```

Structure of dataset after change

```
> str(random_crime_sample)
'data.frame':   1982 obs. of  6 variables:
 $ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 25 3 1 6 13 8 13 15 36 12 ...
 $ Longitude : Factor w/ 937 levels "-5.450768","-5.486587",..: 204 432 246 641 622 142 546 919 877 225 ...
 $ Latitude  : Factor w/ 938 levels "54.062792","54.063553",..: 693 759 477 20 206 533 218 104 843 627 ...
 $ Location  : chr  "HILLVIEW AVENUE" "HENRYVILLE MANOR" "VICTORIA SQUARE" "BARCROFT PARK" ...
 $ Crime.type: Factor w/ 14 levels "ANTI-SOCIAL BEHAVIOUR",..: 1 4 14 4 14 11 7 1 4 14 ...
 $ Postcode  : chr  NA NA NA NA ...
```

Result

I combined the two data frames, the random_crime_sample which I fed into the find_a_postcode function and the crime_data_with_postcode which was returned from this function with the Postcode added for 982 records. I stored the output from joining these 2 dataframes into the random_crime_sample file

As you can see from the details above, the size of the file has grown from 1,000 records to 1,982 records. Finally I wrote the data out to the csv file.

Step Description	Sec 2 - Step G – Description: Create another data frame called chart_data
Snapshot of dataset before processing	Chart data did not exist so I need to create it from the random crime sample. The directions stated to create a separate new dataframe called *updated_random_sample* but given that the attributes in the random_crime_sample already were the list as stated I did not create this separate dataframe as there is no obvious need – however it would have been a simple copy of the random_crime_sample dataframe if required

R Code used to perform change	`# Command to extract all those records from random_crime_sample that have BT1 in the Postcode`
	`# and sort by Postcode and then Crime.type`
	
	`chart_data <- dplyr::filter(random_crime_sample, grepl('BT1', Postcode) )`
	`chart_data <- chart_data %>% arrange(Postcode, Crime.type)`
	
	`summary(chart_data$Crime.type)`

Snapshot of data after application of change	`> summary(chart_data$Crime.type)`
	`ANTI-SOCIAL BEHAVIOUR BICYCLE THEFT BURGLARY CRIMINAL DAMAGE AND ARSON DRUGS`
	`57 1 7 32 2`
	`OTHER CRIME OTHER THEFT POSSESSION OF WEAPONS PUBLIC ORDER ROBBERY`
	`3 14 0 1 1`
	`SHOPLIFTING THEFT FROM THE PERSON VEHICLE CRIME VIOLENCE AND SEXUAL OFFENCES`
	`14 0 7 40`
	`> head(chart_data, 10)`
	`Month Longitude Latitude Location Crime.type Postcode`
	`1 2015-12 -5.986473 54.553504 ERINVALE AVENUE VIOLENCE AND SEXUAL OFFENCES BT100FP`
	`2 2017-11 -5.99083 54.563299 ASHTON AVENUE ANTI-SOCIAL BEHAVIOUR BT100JR`
	`3 2016-11 -5.987572 54.565104 ORCHARDVILLE CRESCENT OTHER THEFT BT100JT`
	`4 2016-02 -5.985197 54.56121 UPPER LISBURN ROAD VIOLENCE AND SEXUAL OFFENCES BT100LA`
	`5 2015-03 -6.350018 54.227011 GLEN ROAD BURGLARY BT118BU`
	`6 2017-05 -6.003331 54.57958 HAMILL PARK VIOLENCE AND SEXUAL OFFENCES BT118DQ`
	`7 2015-08 -5.993002 54.587369 MONAGH ROAD ANTI-SOCIAL BEHAVIOUR BT118EF`
	`8 2016-01 -5.986597 54.582052 DENEWOOD PARK ANTI-SOCIAL BEHAVIOUR BT118FS`
	`9 2017-06 -5.985395 54.57789 BENRAW ROAD VEHICLE CRIME BT118GQ`
	`10 2015-02 -5.993975 54.580485 COOLNASILLA PARK EAST CRIMINAL DAMAGE AND ARSON BT118LA`

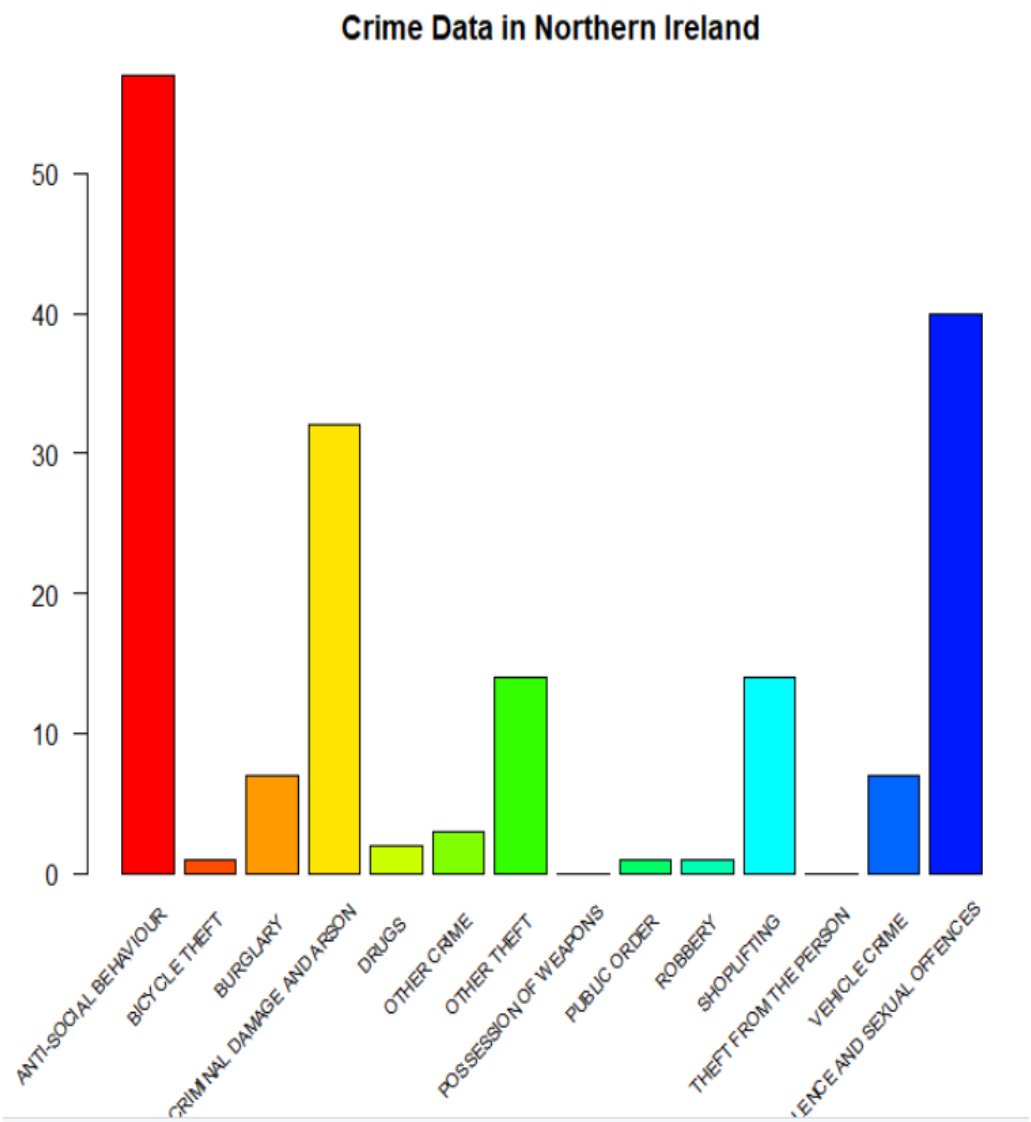
Structure of dataset after change	`> str(chart_data)`
	`'data.frame': 179 obs. of 6 variables:`
	`$ Month : Factor w/ 36 levels "2015-01","2015-02",..: 12 35 23 14 3 29 8 13 30 2 ...`
	`$ Longitude : Factor w/ 937 levels "-5.450768","-5.486587",..: 433 437 435 430 642 452 440 4`
	`$ Latitude : Factor w/ 938 levels "54.062792","54.063553",..: 301 313 314 310 57 353 387 36`
	`$ Location : chr "ERINVALE AVENUE" "ASHTON AVENUE" "ORCHARDVILLE CRESCENT" "UPPER LISBURN`
	`$ Crime.type: Factor w/ 14 levels "ANTI-SOCIAL BEHAVIOUR",..: 14 1 7 14 3 14 1 1 13 4 ...`
	`$ Postcode : chr "BT100FP" "BT100JR" "BT100JT" "BT100LA" ...`
	`> nrow(chart_data)`
	`[1] 179`
	`>`

Result	From the random_crime_data dataframe I extracted all records that had 'BT1' in the postcode and wrote them to chart_data – and from there I sorted by Postcode and Crime.type. I also showed a summary of the Crime.type as shown above.

Step Description	Sec 2 - Step H – Description: Create a bar plot of the crime type from the chart_datadata frame

Snapshot of dataset before processing	`> str(chart_data)`
	`'data.frame':  179 obs. of  6 variables:`
	`$ Month     : Factor w/ 36 levels "2015-01","2015-02",..: 12 35 23 14 3 29 8 13 30 2 ...`
	`$ Longitude : Factor w/ 937 levels "-5.450768","-5.486587",..: 433 437 435 430 642 452 440 434 431 441 ...`
	`$ Latitude  : Factor w/ 938 levels "54.062792","54.063553",..: 301 313 314 310 57 353 387 361 348 357 ...`
	`$ Location  : chr  "ERINVALE AVENUE" "ASHTON AVENUE" "ORCHARDVILLE CRESCENT" "UPPER LISBURN ROAD" ...`
	`$ Crime.type: Factor w/ 14 levels "ANTI-SOCIAL BEHAVIOUR",..: 14 1 7 14 3 14 1 1 13 4 ...`
	`$ Postcode  : chr  "BT100FP" "BT100JR" "BT100JT" "BT100LA" ...`
	
	`> summary(chart_data$Crime.type)`
	`ANTI-SOCIAL BEHAVIOUR  BICYCLE THEFT  BURGLARY  CRIMINAL DAMAGE AND ARSON  DRUGS`
	`57  1  7  32  2`
	`OTHER CRIME  OTHER THEFT  POSSESSION OF WEAPONS  PUBLIC ORDER  ROBBERY`
	`3  14  0  1  1`
	`SHOPLIFTING  THEFT FROM THE PERSON  VEHICLE CRIME  VIOLENCE AND SEXUAL OFFENCES`
	`14  0  7  40`
	`> head(chart_data, 10)`

R Code used to perform change	`# Plotting the Crime Type on a barplot`
	`# first I extracted the names of the various crimes into labelist`
	`# then I put the arguments of what to graph_to_plot but did not show the names.`
	`# Then the final arugment shows the labels - which are reduced to 70% so they can fit properly and slanted`
	`# at a 45 degree angle.`
	
	`labellist <- names(chart_table)`
	
	`grapht_to_plot <- barplot(table(chart_data$Crime.type),`
	`las=2,`
	`col= rainbow(20),`
	`names.arg = "",`
	`main="Crime Data in Northern Ireland")`
	
	`text(grapht_to_plot[,1], -3.7, srt = 45, adj= .9, xpd = TRUE, labels = labellist , cex=.7)`

Snapshot of data after application of change	
Structure of dataset after change	No change
Result	In this last step I show the various crime data in a barplot