Assignment 2

Ashish Mhatre

30/01/2022

Problem Statement

Perform data analysis on Farmer market data set and NY collision data set and perform grouping, summarizing, cleaning, pivoting, date time conversion etc as per the given task using libraries like tidyverse, dplyr, lubridate and magrittr.

Import Libraries

importing all the required libraries

Task 1

Finding the number of farmers market in the state of California city wise.

```
## # A tibble: 465 x 2
##    city             Number_of_Markets
##    <chr>                        <int>
##  1 "Los Angeles"                   27
##  2 "San Francisco"                 20
##  3 "Sacramento"                    18
##  4 "San Diego"                     15
##  5 "San Jose"                      14
##  6 "Oakland"                       10
##  7 "San Francisco "                 8
##  8 "Stockton"                       8
##  9 "Bakersfield"                    6
## 10 "Paso Robles"                    6
## # ... with 455 more rows
```

Calculated the number of farmers market in each city of California state and displayed the information in descending order as per market count.

Task 2

Finding the number of Farmers Market year wise in state of Massachusetts

```
## # A tibble: 9 x 2
##    Year Number_of_Farmers_Market
##   <dbl>                    <int>
## 1  2012                        3
## 2  2013                       15
## 3  2014                       48
## 4  2015                       15
## 5  2016                       72
## 6  2017                       27
## 7  2018                       20
## 8  2019                       12
## 9  2020                        1
```

The following dataset view shows the year wise count of markets in state of Massachusetts.

Task 3

Finding the top 15 states in terms of count of Farmers market.

```
## # A tibble: 15 x 2
##    State          Number_of_Markets
##    <chr>                      <int>
##  1 California                   760
##  2 New York                     673
##  3 Michigan                     343
##  4 Illinois                     338
##  5 Ohio                         336
##  6 Massachusetts                326
##  7 Wisconsin                    312
##  8 Pennsylvania                 311
##  9 Florida                      264
## 10 Virginia                     262
## 11 Missouri                     256
## 12 North Carolina               255
## 13 Texas                        236
## 14 Iowa                         227
## 15 Indiana                      201
```

Following dataset view shows top 15 states with highest farmer market count in descending order.

Task 4

Display in Long format Payment Method, Product Type, and count of Markets accepting the particular combination

```
## # A tibble: 155 x 3
##    Payment_system Products     Farmer_Market
##    <chr>          <chr>                <int>
##  1 WICcash        ""                     940
##  2 WICcash        "Bakedgoods"           222
##  3 WICcash        "Beans"                 27
##  4 WICcash        "Cheese"               146
##  5 WICcash        "Coffee"                73
##  6 WICcash        "Crafts"               107
##  7 WICcash        "Eggs"                 177
##  8 WICcash        "Flowers"              151
##  9 WICcash        "Fruits"               230
## 10 WICcash        "Grains"                30
## # ... with 145 more rows
```

The following view shows the count of Markets accepting a particular payment for respective product type in a Long Format.

Task 5

Split the string in Column Season1Date

```
##    Startdate    Enddate    FMID
## 1 06/14/2017 08/30/2017 1018261
## 2 06/24/2017 09/30/2017 1018318
## 3                  <NA> 1009364
## 4 04/02/2014 11/30/2014 1010691
## 5       July   November 1002454
## 6 05/05/2015 10/27/2015 1011100
##                                          MarketName
## 1  Caledonia Farmers Market Association - Danville
## 2                Stearns Homestead Farmers’ Market
## 3                106 S. Main Street Farmers Market
## 4              10th Steet Community Farmers Market
## 5                             112st Madison Avenue
## 6                          12 South Farmers Market
##                                                  Website
## 1 https://sites.google.com/site/caledoniafarmersmarket/
## 2                       http://www.StearnsHomestead.com
## 3                http://thetownofsixmile.wordpress.com/
## 4
## 5
## 6                   http://www.12southfarmersmarket.com
##                                                Facebook         Twitter Youtube
## 1 https://www.facebook.com/Danville.VT.Farmers.Market/
## 2                        StearnsHomesteadFarmersMarket
## 3
## 4
## 5
## 6                              12_South_Farmers_Market @12southfrmsmkt
##                                                             OtherMedia
## 1
## 2
## 3
## 4 http://agrimissouri.com/mo-grown/grodetail.php?type=mo-grown&ID=275
## 5
## 6                                                      @12southfrmsmkt
##                   street      city    County          State   zip
## 1                         Danville Caledonia        Vermont  5828
## 2        6975 Ridge Road    Parma   Cuyahoga           Ohio      
## 3     106 S. Main Street  Six Mile           South Carolina 29682
## 4 10th Street and Poplar    Lamar     Barton       Missouri 64759
## 5   112th Madison Avenue  New York  New York       New York 10029
## 6 3000 Granny White Pike Nashville  Davidson      Tennessee 37204
##                Season1Date                                   Season1Time
## 1 06/14/2017 to 08/30/2017                        Wed: 9:00 AM-1:00 PM;
## 2 06/24/2017 to 09/30/2017                        Sat: 9:00 AM-1:00 PM;
## 3                                                                      
## 4 04/02/2014 to 11/30/2014   Wed: 3:00 PM-6:00 PM;Sat: 8:00 AM-1:00 PM;
## 5         July to November Tue:8:00 am - 5:00 pm;Sat:8:00 am - 8:00 pm;
## 6 05/05/2015 to 10/27/2015                        Tue: 3:30 PM-6:30 PM;
##                Season2Date           Season2Time Season3Date Season3Time
## 1 09/06/2017 to 10/18/2017 Wed: 2:00 PM-6:00 PM;                        
## 2                                                                       
## 3                                                                       
## 4                                                                       
## 5                                                                       
## 6                                                                       
##   Season4Date Season4Time         x        y                     Location
## 1                         -72.14033 44.41104                             
## 2                         -81.73394 41.37480                             
## 3                         -82.81870 34.80420                             
## 4                         -94.27462 37.49563                             
## 5                         -73.94930 40.79390 Private business parking lot
## 6                         -86.79071 36.11837                             
##   Credit WIC WICcash SFMNP SNAP Organic Bakedgoods Cheese Crafts Flowers Eggs
## 1      Y   Y       N     Y    N       Y          Y      Y      Y       Y    Y
## 2      Y   N       N     Y    N       -          Y      N      Y       Y    Y
## 3      Y   N       N     N    N       -                                      
## 4      Y   N       N     N    N       -          Y      N      Y       N    Y
## 5      N   N       Y     Y    N       -          Y      N      Y       Y    N
## 6      Y   N       N     N    Y       Y          Y      Y      N       Y    Y
##   Seafood Herbs Vegetables Honey Jams Maple Meat Nursery Nuts Plants Poultry
## 1       N     Y          Y     Y    Y     Y    Y       N    N      N       Y
## 2       N     Y          Y     Y    Y     Y    N       N    N      N       Y
## 3                                                                          
## 4       N     Y          Y     Y    Y     N    Y       N    N      Y       Y
## 5       N     Y          Y     Y    Y     N    N       N    Y      N       N
## 6       N     Y          Y     Y    Y     Y    Y       N    N      N       Y
##   Prepared Soap Trees Wine Coffee Beans Fruits Grains Juices Mushrooms PetFood
## 1        Y    Y     Y    N      Y     Y      Y      N      N         Y       Y
## 2        N    Y     N    N      N     N      Y      N      N         N       N
## 3                                                                            
## 4        Y    Y     N    N      N     N      Y      N      N         N       N
## 5        Y    Y     N    N      N     N      N      N      N         N       N
## 6        Y    Y     N    N      Y     N      Y      N      Y         Y       Y
##   Tofu WildHarvested      updateTime
## 1    N             N 6/20/2017 22:43
## 2    N             N 6/21/2017 17:15
## 3                               2013
## 4    N             N 10/28/2014 9:49
## 5    N             N  3/1/2012 10:38
## 6    N             N  5/1/2015 10:40
```

Here we added the two new columns which are StartDate & EndDate

Task 6

Importing dataset

Subtask 1

Calculate all the statistical parameters for column NUMBER.OF.PEDESTRIANS.INJURED and group by BOROUGH

```
## # A tibble: 6 x 10
##   BOROUGH  total average minimum maximum   med mode  First_Quan  Mean Third_Quan
##   <chr>    <int>   <dbl>   <int>   <int> <dbl> <chr>      <dbl> <dbl>      <dbl>
## 1 ""       16284  0.0312       0       8     0 0              0     0          0
## 2 "BRONX"  11487  0.0682       0       9     0 0              0     0          0
## 3 "BROOKL~ 24063  0.0650       0       9     0 0              0     0          0
## 4 "MANHAT~ 16795  0.0601       0      27     0 0              0     0          0
## 5 "QUEENS" 16601  0.0523       0      15     0 0              0     0          0
## 6 "STATEN~  1929  0.0383       0       6     0 0              0     0          0
```

Following data set view shows the statistical parameter calculated for all the boroughs

Subtask 2

Display the number of accident by Vechicle Type code and Borough

```
## # A tibble: 1,660 x 3
##    VEHICLE.TYPE.CODE.1 BOROUGH   Number_of_Accident
##    <chr>               <chr>                  <int>
##  1 "\u007fomm"         MANHATTAN                  1
##  2 "?omme"             BROOKLYN                   1
##  3 "0"                 MANHATTAN                  1
##  4 "1"                 MANHATTAN                  1
##  5 "1"                 QUEENS                     1
##  6 "11111"             BROOKLYN                   1
##  7 "12 Pa"             QUEENS                     1
##  8 "12 PA"             BROOKLYN                   1
##  9 "15 Pa"             BROOKLYN                   1
## 10 "18 WH"             QUEENS                     1
## # ... with 1,650 more rows
```

Here we can see all the number of accidents segregated by Borough and Vehicle Type code.

Subtask 3

Show all the Contributing Factor for accident by Borough.

```
## # A tibble: 302 x 2
##    BOROUGH       CONTRIBUTING.FACTOR.VEHICLE.1
##    <chr>         <chr>
##  1 BRONX         Windshield Inadequate
##  2 BROOKLYN      Windshield Inadequate
##  3 MANHATTAN     Windshield Inadequate
##  4 QUEENS        Windshield Inadequate
##  5 BRONX         View Obstructed/Limited
##  6 BROOKLYN      View Obstructed/Limited
##  7 MANHATTAN     View Obstructed/Limited
##  8 QUEENS        View Obstructed/Limited
##  9 STATEN ISLAND View Obstructed/Limited
## 10 BRONX         Vehicle Vandalism
## # ... with 292 more rows
```

The following views show us all the contributing factor for accident broken up by respective borough.

Subtask 4

Display the number of accidents by each hour of the day

```
## # A tibble: 24 x 2
##    ‘hour(CRASH.TIME)‘ Number_of_Accidents
##                 <dbl>               <int>
##  1                  0               50392
##  2                  1               27063
##  3                  2               20517
##  4                  3               17821
##  5                  4               20532
##  6                  5               22741
##  7                  6               36447
##  8                  7               50445
##  9                  8               95606
## 10                  9               93453
## # ... with 14 more rows
```

The following dataset view shows the number of accidents taken place in 24 hrs grouped by each hour.

Subtask 5

Show the number of accidents taken place in each month of each year.

```
## # A tibble: 109 x 3
##    ‘year(CRASH.DATE)‘ ‘month(CRASH.DATE)‘ Number_of_Accidents
##                 <dbl>               <dbl>               <int>
##  1               2012                   1                3142
##  2               2012                   2                2948
##  3               2012                   3                3056
##  4               2012                   4                3125
##  5               2012                   5                3396
##  6               2012                   6                3437
##  7               2012                   7                3630
## 8              2012                   8                 3096
## 9              2012                   9                 3380
## 10             2012                  10                 3481
## # ... with 99 more rows
```

Here we can see the number of accidents happened during each month of a respective year.

Subtask 6

Present in long format showing Borough in first column, Type of outcome in second column and Injured/Killed in Third column.

```
## # A tibble: 2,963,112 x 3
##    BOROUGH  `Type of Outcome`             `injured/killed`
##    <chr>    <chr>                                    <int>
##  1 BROOKLYN NUMBER.OF.PERSONS.INJURED                    0
##  2 BROOKLYN NUMBER.OF.PERSONS.KILLED                     0
##  3 BROOKLYN NUMBER.OF.PEDESTRIANS.INJURED                0
##  4 BROOKLYN NUMBER.OF.PEDESTRIANS.KILLED                 0
##  5 BROOKLYN NUMBER.OF.CYCLIST.INJURED                    0
##  6 BROOKLYN NUMBER.OF.CYCLIST.KILLED                     0
##  7 BROOKLYN NUMBER.OF.MOTORIST.INJURED                   0
##  8 BROOKLYN NUMBER.OF.MOTORIST.KILLED                    0
##  9 BROOKLYN NUMBER.OF.PERSONS.INJURED                    0
## 10 BROOKLYN NUMBER.OF.PERSONS.KILLED                     0
## # ... with 2,963,102 more rows
```

The following long format shows all the data in long format where column 2 has Type of outcome and column 3 shows Injured/Killed

Conclusion

1. From task 1 we can say that Los Angeles with 27 tops the city with most no. of farmers market in state of California followed by San Francisco with 20 farmer markets.
2. In Massachusetts year 2016 had the most number of markets at 72.
3. California with 760, New York with 673 & Michigan with 343 are the top three states in terms of number of farmers market. here we used the slice function to display only the top 15 cities
4. Used regular express, pivot_longer, NA_if etc to show the view of count of Markets accepting a particular payment for respective product type in a Long Format.
5. We separated the Strings in column Season1Date and combined the new columns with original data using the separate function and bind_cols function.
6. As following –>

A. We calculated all the statistical parameters for column NUMBER.OF.PEDESTRIANS.INJURED using the group_by clause and summaries function, we also used inbuilt R function to calculate (total,mean,median,max,min etc)

B. Displayed the number of accidents segregated by Borough and Vehicle Type code.

C. Showed all the contributing factor for accident broken up by respective borough.

D. Calculated data set view to show the number of accidents taken place in 24 hrs grouped by each hour. by this we can infer that maximum number of accidents happen at 16:00 Hr of the day.

E. Calculated the number of accidents happened during each month of a respective year.

F. Using Pivot_longer function we pivoted the required columns of type of outcome in a single column and displayed the injured/killed in the third column by each borough in first column.

Reference :- https://tidyr.tidyverse.org/reference/pivot_longer.html https://dplyr.tidyverse.org/reference/na_if.html https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html https://stackoverflow.com/questions/2547402/how-to-find-the-statistical-mode