Programming Fundamentals for Data Engineers

Final Project Report

UN Cell Phone and Internet Data Wrangling and Analysis

Table of Contents

Initial Data Review:

Before choosing a dataset we reviewed various potential datasets provided; we ultimately landed on the UN data for cell phones and internet usage because they provided a large amount of relatively clean numeric data that we could use for data/stats work.

Preliminary Design/Planning:

Before coding we laid out the main blocks of our code as well as the user inputs we would request, as well as some high-level stats we would be providing.

Main code blocks:

- Main (also contains data/stats coding)
- User input
- Find nulls

User inputs:

- UN Sub-region
- Dataset (cell phones or internet usage)
- Years for specific data column operations

Data/Stats:

- Print summary of final merged dataframe
- Mean
- Sum

Program execution/flow of control overview:

main() → user_input(args) → main() → find_null(args) → main() → data/stats work (done in main)

Design Process:

Three (3) separate original dataframes were merged into a master dataframe with a multi-index for work on the project; Excel was not used to modify the files prior to loading. General overview of steps employed for each individual dataframe prior to merging:

1. High-level review of data in Jupyter Lab first to get a better understanding of NaNs, data types (e.g. df.info()).
2. Loaded data with pd.read_excel()
3. Dropped duplicate rows
4. Dropped unnecessary columns (e.g. sparse data)
5. Casted dataframe elements as needed (e.g. int to float)
6. Checked and changed column labels as necessary (e.g. “Country” was changed to “country” to facilitate dataframe merging later.

The final/master dataframe was merged in 2 steps:

First the two large dataset frames (cell phones and internet user usage data) were merged. Second the UN Codes dataframe was merged into the first merged dataframe. In both merges, Pandas pd.merge() was used. Following each merge duplicate rows and columns were checked for and removed if they existed.

The final/master dataframe (tech_df) was provided a multi-index in the following manner:

- The row multi-index consisting of UN sub-regions and countries was created by using df.set_index() to effectively move the “un sub-region” and “country” columns into a 2-level row index.
- The column multi-index was prepared using a list of valid years and a list of valid data type labels. These 2 lists where then used to construct a 2-level multi-index using Pandas pd.MultiIndex.from_product() function. (Note: an intermediate step of transposing the dataframe was required, since Pandas does not have a “set_column_index()” function, and the “set_index()” function is specific to rows).

user_input(args) function has primarily 4 main parts:

1. The 1st prompt is preceded with an overview and notes/instructions. The initial prompt than allows the user to “quit” or enter any other key to continue. (Also of note, user inputs are case insensitive).
2. If the user does not quit, a sub-region **mask** is created to be used in a nested while loop try/catch block for validation. In this first try and catch block, the user is prompted again to enter a valid sub-region name or index number – then this input is checked against the validation mask created (the “if” statement checks if a valid name was entered; and the “elif” checks whether a valid index number was entered). If an invalid entry was given, user has the option to get “help”.
3. Once a valid sub-region has been entered, the user is prompted for a valid data type (‘1’ for cell phone or ‘2’ internet usage data). This part of the user input is also implemented with a try-catch block, with if/elif /else statements in the try-body: acceptable entries are “1”, “2”, or “quit” and any other entries raise an exception to be handled giving the user another chance to enter.
4. The last 2 user input blocks are very similar as they both prompt/obtain a specific year input (e.g. 2000, 2001. These blocks also employ nested while loops with try-catch blocks with conditional if/elif/else statements. Valid entries include dates from 1995-2017, or “quit” – other entries cause an exception and user has another opportunity to make a valid entry.

Note: Nested while loops were used to ensure that the if/elif/else statements were evaluated/used first (i.e. to prevent the else-statement being associated with the try-catch.)

Data Statistics Design:

Masking:

We used a **masking** operation to find if any missing values were in the sub region - the user then chooses and a statement is printed out based off if there was any missing data. This notice to the user was just to let them know if there were missing data.

Dataframe Manipluation:

For the statistics portion of the project, the first thing we did was show the user the dataframe filtered down to their sub-region and data-type selections. Our first statistic was calculating the **mean** of their data type selected from all data from 1995 to 2017 and presenting it to the user.

Next we **created two columns** taking the differences from 1995 to 2017 for both number of cell phones and internet users and appended them to the final/master merged dataframe. We then presented these columns to the users based off the sub region they choose.

For the global data statistics, we used the **groupby** function to take the **mean and sum** of the number of cell phones and presented this data to the user. We then calculated the total number of cell phones across all un sub-regions and presented to the user.

Next the **.describe** function was used to show some basic statistics and percentiles for the merged dataframe. From the two (2) user year inputs (e.g. 2000, 2002), we took the average of these two columns individually and presented them back to the user based on their chosen sub region and data-type.

The **Pivot Table** was next printed (note: it was created above when the dataframes were being merged). The pivot table information isn't great due to the data set not being a good application for use with a pivot table, but we put something together to create one with the aggregate function of sum.

Charting:

A **Bar Plot** was created using the sub region the user chose and presented them the internet users % for 2017 in the chart. We added labels and a grid with a legend. This plot is saved to the folder of the .py file and then showed to the user.

A **Pie Chart** was created using the sub region the user chose and presented them the number of cell phones for 2017 in the chart. We added labels and a grid with a legend. This plot is saved to the folder of the .py file and then showed to the user.

To finish off the project we saved the merged dataframe to excel using the **.to_excel** function and choose its given name.

Quick Find Guide for Code (Where to find things):

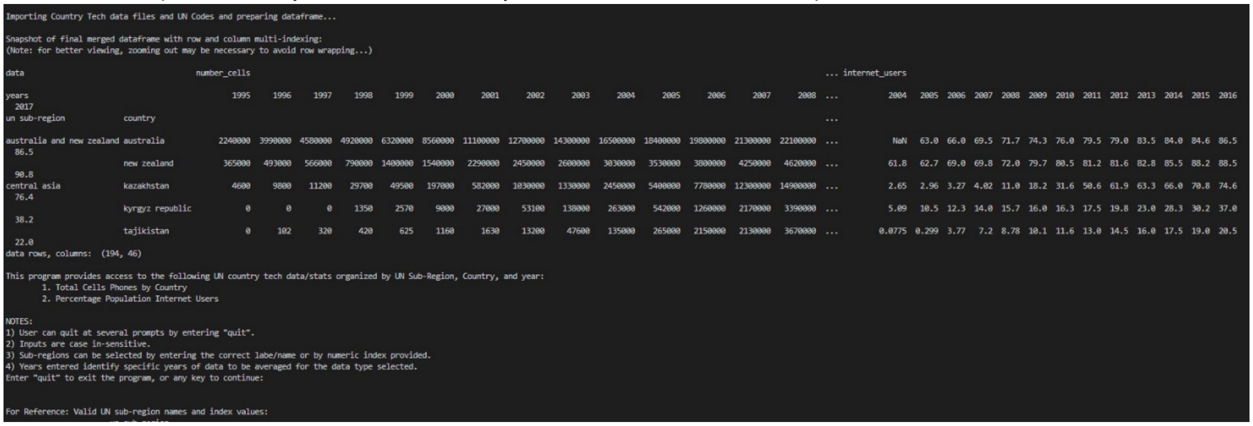
Citation:

Note these datasets where combined to create the given datasets

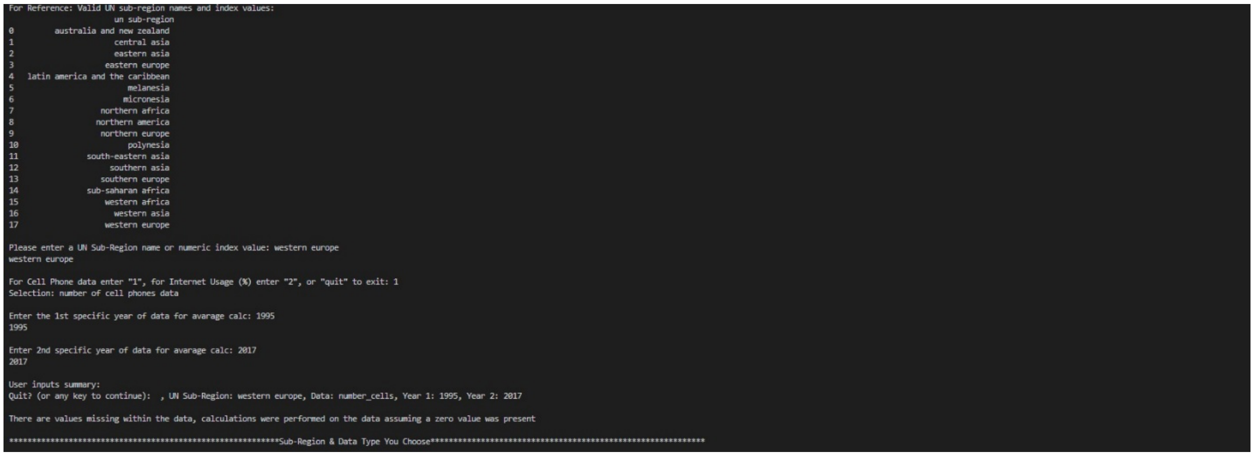
Cell Phones (per 100 people), United Nations, June 2021, [Online] Available: https://www.gapminder.org/data/
Cell Phones (total), United Nations, June 2021, [Online] Available: https://www.gapminder.org/data/
Individuals using the Internet (% of population), United Nations, June 2021, [Online] Available: https://www.gapminder.org/data/

Screenshots:

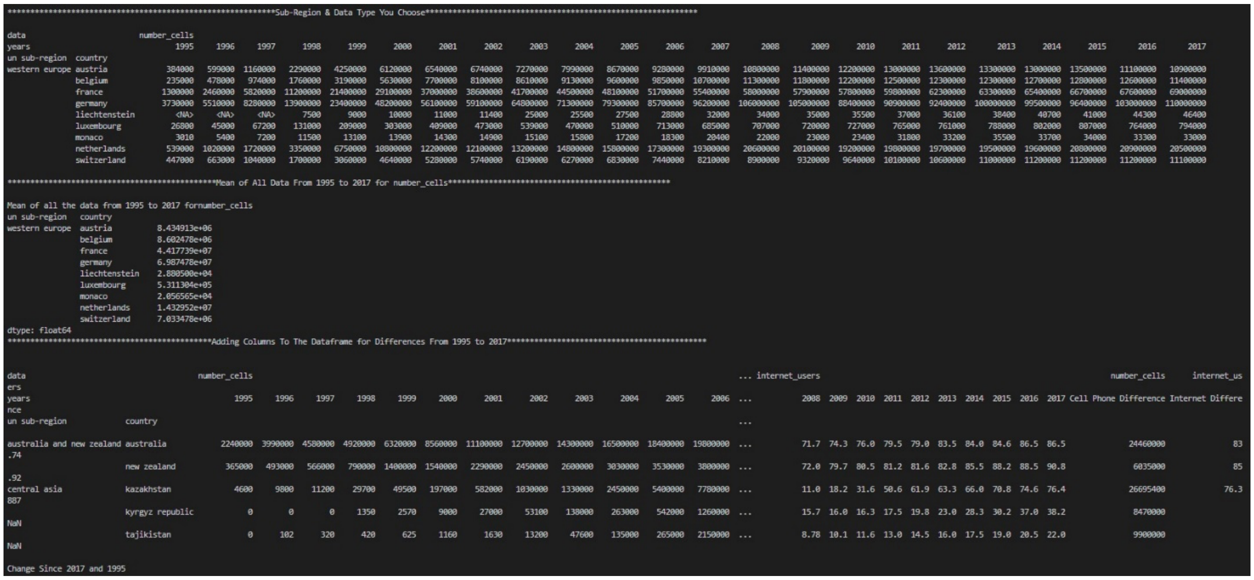
Execution: (Note: they are in a folder you downloaded as well)



```
Importing Country Tech data files and UN Codes and preparing dataframe...

Snapshot of final merged dataframe with row and column multi-indexing:
(Note: for better viewing, zooming out may be necessary to avoid row wrapping...)

data                                    number_cells                                                                                                                   ... internet_users
years                                       1995    1996    1997    1998    1999    2000     2001     2002     2003     2004     2005     2006     2007     2008 ...   2004  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014  2015  2016
  2017
un sub-region            country                                                                                                                                   ...
australia and new zealand australia      2240000 3990000 4580000 4920000 6320000 8560000 11100000 12700000 14300000 16500000 18400000 19800000 21300000 22100000 ...    NaN  63.0  66.0  69.5  71.7  74.3  76.0  79.5  79.0  83.5  84.0  84.6  86.5
  86.5
                          new zealand     365000  493000  566000  790000 1400000 1540000  2290000  2450000  2600000  3030000  3530000  3800000  4250000  4620000 ...   61.8  62.7  69.0  69.8  72.0  79.7  80.5  81.2  81.6  82.8  85.5  88.2  88.5
  90.8
central asia              kazakhstan        4600    9800   11200   29700   49500  197000   582000  1030000  1330000  2450000  5400000  7780000 12300000 14900000 ...   2.65  2.96  3.27  4.02  11.0  18.2  31.6  50.6  61.9  63.3  66.0  70.8  74.6
  76.4
                          kyrgyz republic      0       0       0    1350    2570    9000    27000    53100   138000   263000   542000  1260000  2170000  3390000 ...   5.09  10.5  12.3  14.0  15.7  16.0  16.3  17.5  19.8  23.0  28.3  30.2  37.0
  38.2
                          tajikistan           0     102     320     420     625    1160     1630    13200    47600   135000   265000  2150000  2130000  3670000 ... 0.0775 0.299  3.77   7.2  8.78  10.1  11.6  13.0  14.5  16.0  17.5  19.0  20.5
  22.0
data rows, columns:  (194, 46)

This program provides access to the following UN country tech data/stats organized by UN Sub-Region, Country, and year:
        1. Total Cells Phones by Country
        2. Percentage Population Internet Users

NOTES:
1) User can quit at several prompts by entering "quit".
2) Inputs are case in-sensitive.
3) Sub-regions can be selected by entering the correct labe/name or by numeric index provided.
4) Years entered identify specific years of data to be averaged for the data type selected.
Enter "quit" to exit the program, or any key to continue:


For Reference: Valid UN sub-region names and index values:
```



```
For Reference: Valid UN sub-region names and index values:
                  un sub-region
0         australia and new zealand
1                      central asia
2                      eastern asia
3                    eastern europe
4   latin america and the caribbean
5                         melanesia
6                        micronesia
7                   northern africa
8                  northern america
9                   northern europe
10                        polynesia
11               south-eastern asia
12                    southern asia
13                  southern europe
14               sub-saharan africa
15                   western africa
16                     western asia
17                   western europe

Please enter a UN Sub-Region name or numeric index value: western europe
western europe

For Cell Phone data enter "1", for Internet Usage (%) enter "2", or "quit" to exit: 1
Selection: number of cell phones data

Enter the 1st specific year of data for avarage calc: 1995
1995

Enter 2nd specific year of data for avarage calc: 2017
2017

User inputs summary:
Quit? (or any key to continue):  , UN Sub-Region: western europe, Data: number_cells, Year 1: 1995, Year 2: 2017

There are values missing within the data, calculations were performed on the data assuming a zero value was present

************************************************************Sub-Region & Data Type You Choose************************************************************
```



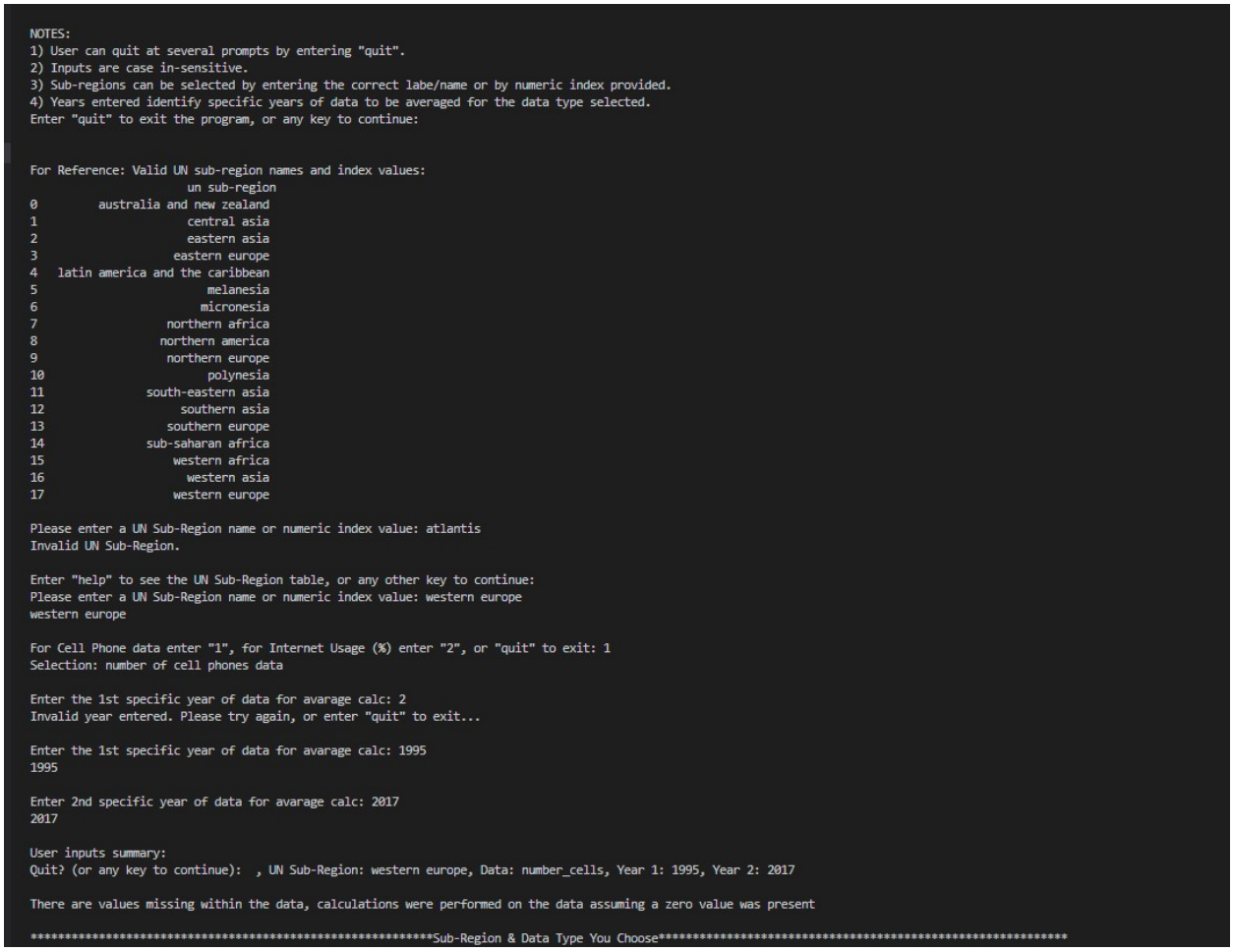
```
************************************************************Sub-Region & Data Type You Choose************************************************************

data                           number_cells
years                                  1995    1996    1997     1998     1999     2000     2001     2002     2003     2004     2005     2006     2007      2008      2009     2010     2011     2012      2013     2014     2015      2016      2017
un sub-region  country
western europe austria               384000  599000 1160000  2290000  4250000  6120000  6540000  6740000  7270000  7990000  8670000  9280000  9910000  10800000  11400000 12200000 13000000 13600000  13300000 13000000 13500000  11100000  10900000
               belgium               235000  478000  974000  1760000  3190000  5630000  7700000  8100000  8610000  9130000  9600000  9850000 10700000  11300000  11800000 12200000 12500000 12300000  12300000 12700000 12800000  12600000  11400000
               france               1300000 2460000 5820000 11200000 21400000 29100000 37000000 38600000 41700000 44500000 48100000 51700000 55400000  58000000  57900000 57800000 59800000 62300000  63300000 65400000 66700000  67600000  69000000
               germany              3730000 5510000 8280000 13900000 23400000 48200000 56100000 59100000 64800000 71300000 79300000 85700000 96200000 106000000 105000000 88400000 90900000 92400000 100000000 99500000 96400000 103000000 110000000
               liechtenstein           <NA>    <NA>    <NA>     7500     9000    10000    11000    11400    25000    25500    27500    28800    32000     34000     35000    35500    37000    36100     38400    40700    41000     44300     46400
               luxembourg             26800   45000   67200   131000   209000   303000   409000   473000   539000   470000   510000   713000   685000    707000    720000   727000   765000   761000    788000   802000   807000    764000    794000
               monaco                  3010    5400    7200    11500    13100    13900    14300    14900    15100    15800    17200    18300    20400     22000     23000    23400    31800    33200     35500    33700    34000     33300     33000
               netherlands           539000 1020000 1720000  3350000  6750000 10800000 12200000 12100000 13200000 14800000 15800000 17300000 19300000  20600000  20100000 19200000 19800000 19700000  19500000 19600000 20800000  20900000  20500000
               switzerland           447000  663000 1040000  1700000  3060000  4640000  5280000  5740000  6190000  6270000  6830000  7440000  8210000   8900000   9320000  9640000 10100000 10600000  11000000 11200000 11200000  11200000  11100000

**********************************************Mean of All Data From 1995 to 2017 for number_cells**************************************************

Mean of all the data from 1995 to 2017 fornumber_cells
un sub-region   country
western europe  austria          8.434913e+06
                belgium          8.602478e+06
                france           4.417739e+07
                germany          6.987478e+07
                liechtenstein    2.880500e+04
                luxembourg       5.311304e+05
                monaco           2.056565e+04
                netherlands      1.432952e+07
                switzerland      7.033478e+06
dtype: float64
**********************************************Adding Columns To The Dataframe for Differences From 1995 to 2017**********************************************


data                                    number_cells                                                                                            ... internet_users                                                            number_cells      internet_us
ers
years                                       1995    1996    1997    1998    1999    2000     2001     2002     2003     2004     2005     2006 ...           2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 Cell Phone Difference Internet Differe
nce
un sub-region            country                                                                                                                ...

australia and new zealand australia      2240000 3990000 4580000 4920000 6320000 8560000 11100000 12700000 14300000 16500000 18400000 19800000 ...           71.7 74.3 76.0 79.5 79.0 83.5 84.0 84.6 86.5 86.5              24460000               83
.74
                          new zealand     365000  493000  566000  790000 1400000 1540000  2290000  2450000  2600000  3030000  3530000  3800000 ...           72.0 79.7 80.5 81.2 81.6 82.8 85.5 88.2 88.5 90.8               6035000               85
.92
central asia              kazakhstan        4600    9800   11200   29700   49500  197000   582000  1030000  1330000  2450000  5400000  7780000 ...           11.0 18.2 31.6 50.6 61.9 63.3 66.0 70.8 74.6 76.4              26695400             76.3
887
                          kyrgyz republic      0       0       0    1350    2570    9000    27000    53100   138000   263000   542000  1260000 ...           15.7 16.0 16.3 17.5 19.8 23.0 28.3 30.2 37.0 38.2               8470000
NaN
                          tajikistan           0     102     320     420     625    1160     1630    13200    47600   135000   265000  2150000 ...           8.78 10.1 11.6 13.0 14.5 16.0 17.5 19.0 20.5 22.0               9900000
NaN

Change Since 2017 and 1995
```

```
Change Since 2017 and 1995
un sub-region   country
western europe  austria          10516000
                belgium          11165000
                france           67700000
                germany         106270000
                liechtenstein        <NA>
                luxembourg         767200
                monaco              29990
                netherlands      19961000
                switzerland      10653000
Name: (number_cells, Cell Phone Difference), dtype: object

**********************************************Mean of Number of Cell Phones across all UN Sub-Regions in 2017**********************************************
un sub-region
australia and new zealand          1.655000e+07
central asia                       1.575000e+07
eastern asia                       3.428800e+08
eastern europe                     4.089500e+07
latin america and the caribbean    2.089316e+07
melanesia                          1.435750e+06
micronesia                         2.380000e+04
northern africa                    4.027000e+07
northern america                   1.439377e+08
northern europe                    1.256410e+07
polynesia                          8.000000e+04
south-eastern asia                 8.327227e+07
southern asia                      1.825368e+08
southern europe                    1.338024e+07
sub-saharan africa                 1.566668e+07
western africa                     6.120000e+05
western asia                       1.507222e+07
western europe                     2.597482e+07
Name: (number_cells, 2017), dtype: float64

**********************************************Total number of cell phones across all UN Sub Regions in 2017**********************************************
```

```
**********************************************Total number of cell phones across all UN Sub Regions in 2017**********************************************
un sub-region
australia and new zealand            33100000
central asia                         78750000
eastern asia                       1714400000
eastern europe                      408950000
latin america and the caribbean     668581100
melanesia                             5743000
micronesia                              95200
northern africa                     241620000
northern america                    431813000
northern europe                     125641000
polynesia                              240000
south-eastern asia                  915995000
southern asia                      1642831000
southern europe                     187323300
sub-saharan africa                  736334000
western africa                         612000
western asia                        271300000
western europe                      233773400
Name: (number_cells, 2017), dtype: int64

****************************************Total number of Cell phones in the world under UN-Sub Regions in 2017********************************************
7697102000

*********************************************************Describe'd Data from Number_Cell Phones**********************************************************
years          1995          1996          1997          1998          1999          2000          2001          2002          2003  ...          2010          2011          2012          2013          2014          2015          2016          2017  Cell Phone Difference
count  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  ...  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01           1.800000e+01
mean   1.102660e+06  1.681392e+06  2.324167e+06  3.165700e+06  4.477693e+06  6.362844e+06  8.218179e+06  9.836579e+06  1.174858e+07  ...  3.637401e+07  4.029215e+07  4.280315e+07  4.541216e+07  4.804369e+07  4.918868e+07  5.179235e+07  5.398858e+07           5.372774e+07
std    2.876833e+06  3.944024e+06  5.153486e+06  6.717112e+06  8.801550e+06  1.178165e+07  1.501800e+07  1.796993e+07  2.126517e+07  ...  5.372767e+07  6.082340e+07  6.614046e+07  7.179184e+07  7.632656e+07  7.764204e+07  8.258637e+07  8.805482e+07           8.683329e+07
min    0.000000e+00  0.000000e+00  2.000000e+01  3.042500e+02  4.107500e+02  4.867500e+02  5.960000e+02  9.017500e+02  2.728500e+03  ...  1.470000e+04  1.587500e+04  1.830000e+04  2.310000e+04  2.153333e+04  2.242000e+04  2.843333e+04  2.380000e+04           2.360900e+04
25%    3.114083e+03  5.856383e+03  1.213133e+04  2.141550e+04  4.960757e+04  1.021361e+05  2.040218e+05  3.289628e+05  4.680576e+05  ...  8.639559e+06  1.024594e+07  1.143364e+07  1.252127e+07  1.280546e+07  1.283887e+07  1.276089e+07  1.276813e+07           1.219548e+07
50%    6.785714e+04  1.434094e+05  2.970314e+05  5.434697e+05  1.010368e+06  1.706570e+06  2.261685e+06  2.790520e+06  4.314390e+06  ...  1.381115e+07  1.432473e+07  1.445760e+07  1.476361e+07  1.515250e+07  1.566056e+07  1.587591e+07  1.615000e+07           1.584093e+07
75%    7.405753e+05  1.234563e+06  1.762216e+06  2.746639e+06  4.245200e+06  6.087775e+06  6.980450e+06  7.642350e+06  9.130700e+06  ...  3.860517e+07  3.760000e+07  3.934317e+07  4.096575e+07  4.111658e+07  4.109058e+07  4.097583e+07  4.073875e+07           4.069642e+07
max    1.213013e+07  1.583352e+07  1.983366e+07  2.485047e+07  3.097067e+07  3.924477e+07  4.979900e+07  6.392320e+07  7.812380e+07  ...  2.071484e+08  2.350880e+08  2.619360e+08  2.876000e+08  3.022260e+08  3.032420e+08  3.190560e+08  3.428800e+08           3.394860e+08

*******************************************Average of your two choosen sub-region, data type and years************************************************
Average of number_cells in 1995 and number_cells in 2017

un sub-region   country
western europe  austria          5642000.0
                belgium          5817500.0
                france          35150000.0
                germany         56865000.0
                liechtenstein         <NA>
                luxembourg        410400.0
                monaco             18005.0
                netherlands     10519500.0
                switzerland      5773500.0
dtype: object
```

```
*********************************************************************Pivot Table*********************************************************************
country                               afghanistan    albania   algeria  andorra      angola antigua and barbuda argentina    armenia  australia  ... united states   uruguay uzbekistan   vanuatu venezuela    vietnam       yemen      zambia    zimbabwe
2017                                     23900000    3630000  45800000  80300      13300000              184000  61900000    3490000   26700000  ...     400000000 5100000     24300000  228000     24500000  120000000    15400000    13400000    14100000
     un sub-region                                                                                                                              ...
1995 australia and new zealand                0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0  2240000.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0
     central asia                             0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0     3730.0       0.0       0.0        0.0         0.0         0.0         0.0
     eastern asia                             0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0
     eastern europe                           0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0
     latin america and the caribbean          0.0        0.0       0.0      0.0         0.0                 0.0  405000.0        0.0        0.0  ...           0.0   39900.0        0.0       0.0  404000.0        0.0         0.0         0.0         0.0
...                                           ...        ...       ...      ...         ...                 ...       ...        ...        ...  ...           ...       ...        ...       ...       ...        ...         ...         ...         ...
2016 southern europe                          0.0  3370000.0       0.0  76100.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0
     sub-saharan africa                       0.0        0.0       0.0      0.0  13000000.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0  12000000.0  12900000.0
     western africa                           0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0
     western asia                             0.0        0.0       0.0      0.0         0.0                 0.0       0.0  3430000.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0  16400000.0         0.0         0.0
     western europe                           0.0        0.0       0.0      0.0         0.0                 0.0       0.0        0.0        0.0  ...           0.0       0.0        0.0       0.0       0.0        0.0         0.0         0.0         0.0

*********************************************************************Plot 1 Data*********************************************************************
country
austria          87.9
belgium          87.7
france           80.5
germany          84.4
liechtenstein    99.5
luxembourg       97.4
monaco           97.1
netherlands      93.2
switzerland      89.7
Name: (internet_users, 2017), dtype: object

*********************************************************************Plot 2 Data*********************************************************************
country
austria           10900000
belgium           11400000
france            69000000
germany          110000000
liechtenstein        46400
luxembourg          794000
monaco               33000
netherlands       20500000
switzerland       11100000
Name: (number_cells, 2017), dtype: object
```

Handling of Incorrect Inputs Screenshot



```
NOTES:
1) User can quit at several prompts by entering "quit".
2) Inputs are case in-sensitive.
3) Sub-regions can be selected by entering the correct labe/name or by numeric index provided.
4) Years entered identify specific years of data to be averaged for the data type selected.
Enter "quit" to exit the program, or any key to continue:


For Reference: Valid UN sub-region names and index values:
                    un sub-region
0         australia and new zealand
1                    central asia
2                    eastern asia
3                  eastern europe
4   latin america and the caribbean
5                       melanesia
6                      micronesia
7                 northern africa
8                northern america
9                 northern europe
10                     polynesia
11             south-eastern asia
12                 southern asia
13               southern europe
14             sub-saharan africa
15                 western africa
16                   western asia
17                 western europe

Please enter a UN Sub-Region name or numeric index value: atlantis
Invalid UN Sub-Region.

Enter "help" to see the UN Sub-Region table, or any other key to continue:
Please enter a UN Sub-Region name or numeric index value: western europe
western europe

For Cell Phone data enter "1", for Internet Usage (%) enter "2", or "quit" to exit: 1
Selection: number of cell phones data

Enter the 1st specific year of data for avarage calc: 2
Invalid year entered. Please try again, or enter "quit" to exit...

Enter the 1st specific year of data for avarage calc: 1995
1995

Enter 2nd specific year of data for avarage calc: 2017
2017

User inputs summary:
Quit? (or any key to continue):  , UN Sub-Region: western europe, Data: number_cells, Year 1: 1995, Year 2: 2017

There are values missing within the data, calculations were performed on the data assuming a zero value was present

**********Sub-Region & Data Type You Choose**********
```