13th February 2020
Author: Verity Tether

Accessing and Analysing Census Data in Excel

Welcome to Accessing and Analysing Census Data in Excel! In this session, you will do the following:

- Learn how to access and download census data via InFuse
- Learn the basics of census geographies
- Learn key excel skills, fundamental to any excel beginner

This workshop will be broken down into two sections. The first will run through, step by step, how to obtain census data. The second will introduce a variety of Excel functions you can use to analyse this data. If you already have experience of either one of these, feel free to skip the half of the session you do not need.

Through this handout, you will analyse data relating to the long-term health of the population in Leeds, and whether they have dependent children in the family. This sort of data could be helpful for councils identifying needs for additional childcare in certain areas, or looking at where to build accessible facilities.

Part 1: Census Data

Introduction to the UK Census

A full census of the UK population is taken every 10 years, providing information on national demographics. Almost everyone in the UK takes part in the census, and this data is then used to plan local services. After it has been anonymised so that individuals cannot be identified, this data is then released at various different spatial scales to enable researchers to download it and analyse it.

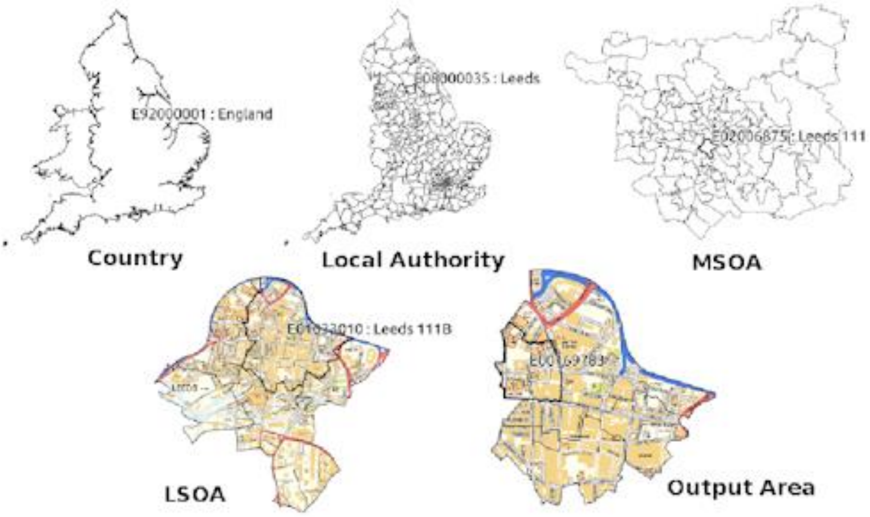
The most recent UK census was in 2011. Whilst this means that the data is a little out of date, it still provides a good idea of the demographics in local areas.

Census Geographies

Census data is released at a variety of spatial scales, allowing us to examine the data at different levels of granularity. Each area has a code (for example E00056946), to which the census data for that area will be assigned. This data can then be analysed using statistical tools (such as those in Excel, SPSS or

R), or mapped using GIS packages. The scale at which you are conducting your analysis may impact which data you are able to select because of data anonymisation practices.

The following image shows the 2011 Census Geography hierarchy. It goes from **Country** to **Local Authority** to **Middle Layer Super Output Area** (MSOA) to **Lower Layer Super Output Area** (LSOA) to **Output Area**, the smallest. However, keep in mind that these boundaries can change between censuses, so are not necessarily directly comparable between years.



To help understand this a bit more, the average number of households in each type of neighbourhood geography are as follows:

- MSOA – 7,200
- LSOA – 1,500
- OA - 125

Keep in mind that other geographies are available. For additional information on this, visit this page.

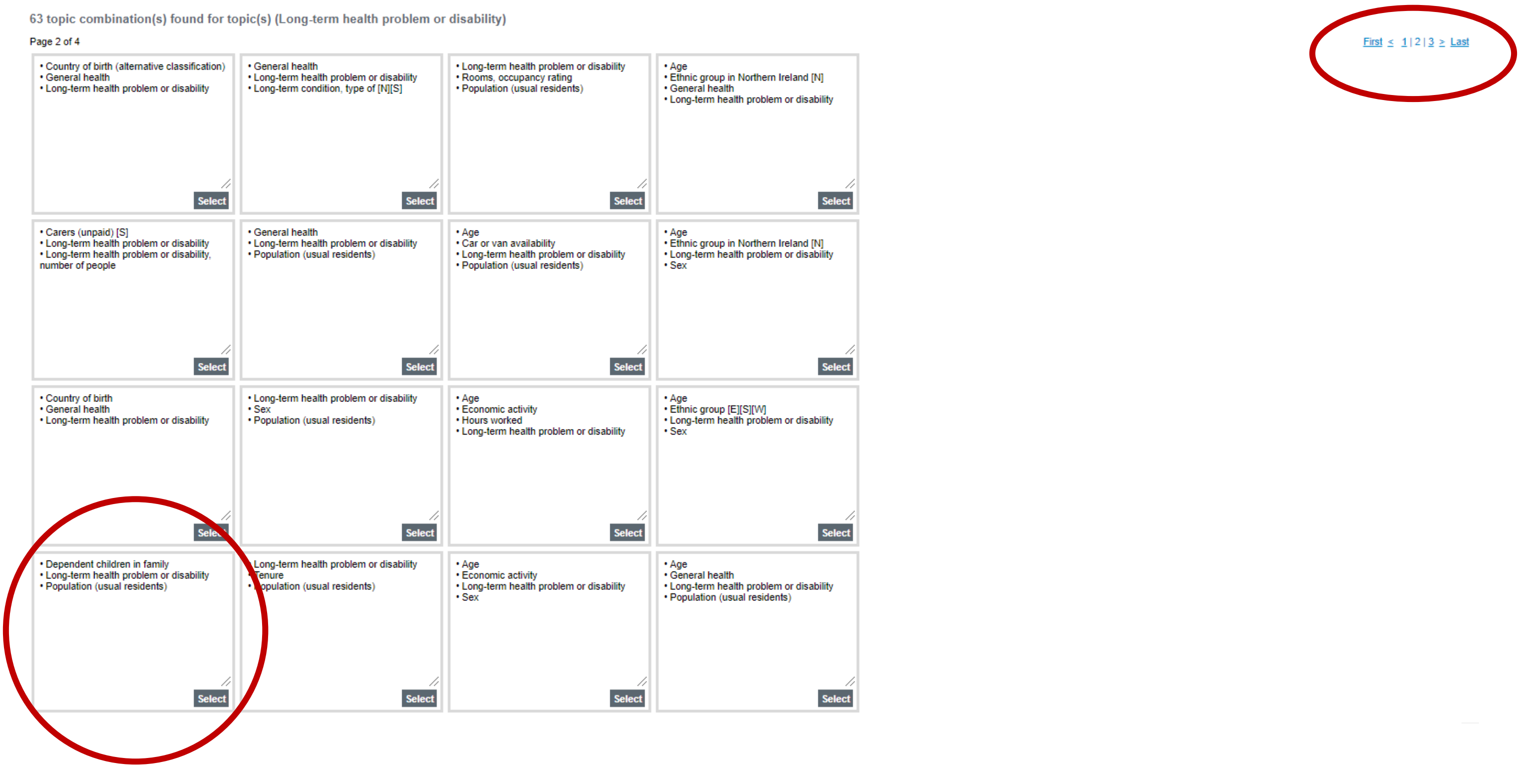
How to Access and Download the Data

1. Go to http://infuse.ukdataservice.ac.uk/

Here you can access and download data from the 2001 and 2011 censuses.

2. Click *2011 Census Data.*
3. Click *Topics.*

4. From here, you can see the data is separated into different themes. Feel free to have a browse to see what is available. For this workshop, however, we need the data on the long-term health of the population, and the dependent children in families. So, scroll down the list on the left hand side of the screen, until you reach *Long-term health problem or disability.* Select this, and it brings up a number of combinations of variables which may be relevant to this. Navigate to the second page (see screenshot below), and in the bottom corner of the second page there is an option which lists:
   a. Dependent children in family
   b. Long-term health problem or disability
   c. Population (usual residents)



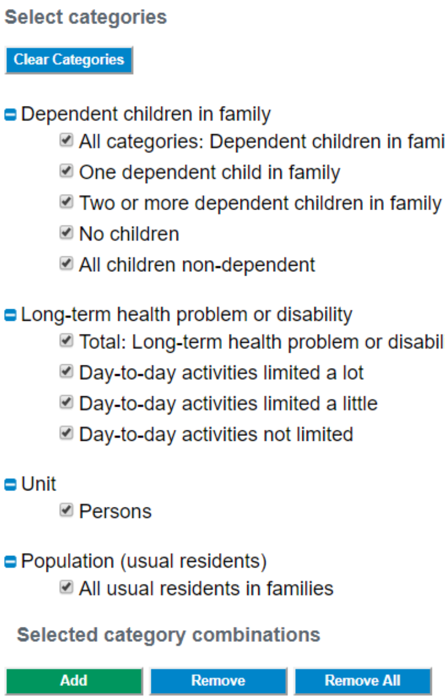
63 topic combination(s) found for topic(s) (Long-term health problem or disability)

Page 2 of 4

First < 1 | 2 | 3 > Last

• Country of birth (alternative classification) • General health • Long-term health problem or disability Select	• General health • Long-term health problem or disability • Long-term condition, type of [N][S] Select	• Long-term health problem or disability • Rooms, occupancy rating • Population (usual residents) Select	• Age • Ethnic group in Northern Ireland [N] • General health • Long-term health problem or disability Select
• Carers (unpaid) [S] • Long-term health problem or disability • Long-term health problem or disability, number of people Select	• General health • Long-term health problem or disability • Population (usual residents) Select	• Age • Car or van availability • Long-term health problem or disability • Population (usual residents) Select	• Age • Ethnic group in Northern Ireland [N] • Long-term health problem or disability • Sex Select
• Country of birth • General health • Long-term health problem or disability Select	• Long-term health problem or disability • Sex • Population (usual residents) Select	• Age • Economic activity • Hours worked • Long-term health problem or disability Select	• Age • Ethnic group [E][S][W] • Long-term health problem or disability • Sex Select
• Dependent children in family • Long-term health problem or disability • Population (usual residents) Select	• Long-term health problem or disability • Tenure • Population (usual residents) Select	• Age • Economic activity • Long-term health problem or disability • Sex Select	• Age • General health • Long-term health problem or disability • Population (usual residents) Select

5. Click *Select* in the box which has these variables, circled above.
6. The next page gives you a description of each of the variables you are about to be given access to. Have a read about them so that you fully understand the dataset, then click *Next*.
7. On the following page, click the box next to all the variables listed, so that it resembles the screenshot below. This ensures that we download each one. It is possible to only download one variable if you're examining a specific topic (so, for example, only people whose day-to-day activities were limited a lot by their illness), but we want them all for this workshop.



Select categories

Dependent children in family
- ☑ All categories: Dependent children in family
- ☑ One dependent child in family
- ☑ Two or more dependent children in family
- ☑ No children
- ☑ All children non-dependent

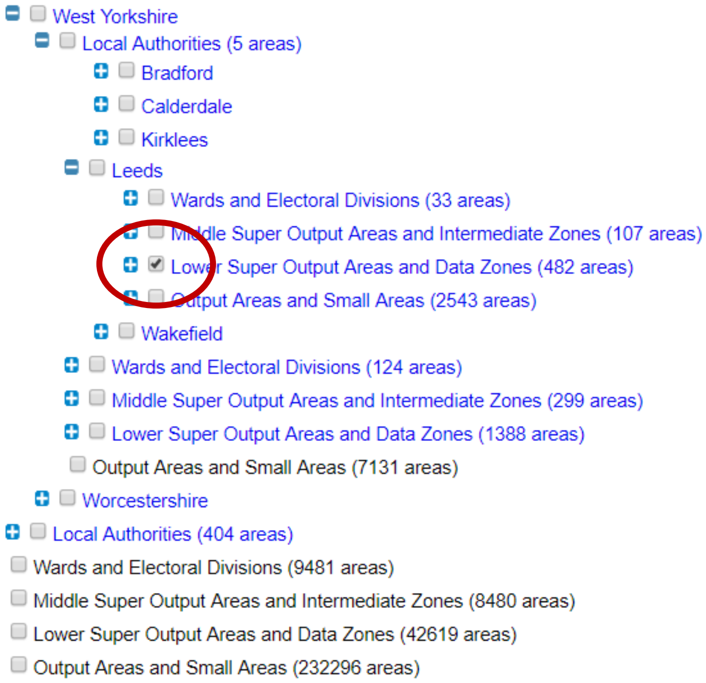
Long-term health problem or disability
- ☑ Total: Long-term health problem or disability
- ☑ Day-to-day activities limited a lot
- ☑ Day-to-day activities limited a little
- ☑ Day-to-day activities not limited

Unit
- ☑ Persons

Population (usual residents)
- ☑ All usual residents in families

Selected category combinations

8. When you have ticked them all, press *Add*. They will then be listed in the box underneath the variables, and it says *"You have selected 20 category combinations"*.
9. Press *Next*.
10. Now, we must select the georaphic level that we want to analyse this data at. In our case, that is LSOA, so follow these steps to select all LSOAs for Leeds:
    a. Press the + next to *Counties* to expand this list
    b. Scroll down, and press the + next to *West Yorkshire* to expand that list too
    c. Press the + next to *Local Authorities*
    d. Press the + next to *Leeds*
    e. Tick the box next to *Lower Super Output Areas and Data Zones (482 areas)*, to select all the LSOAs in Leeds, as seen in the screenshot below



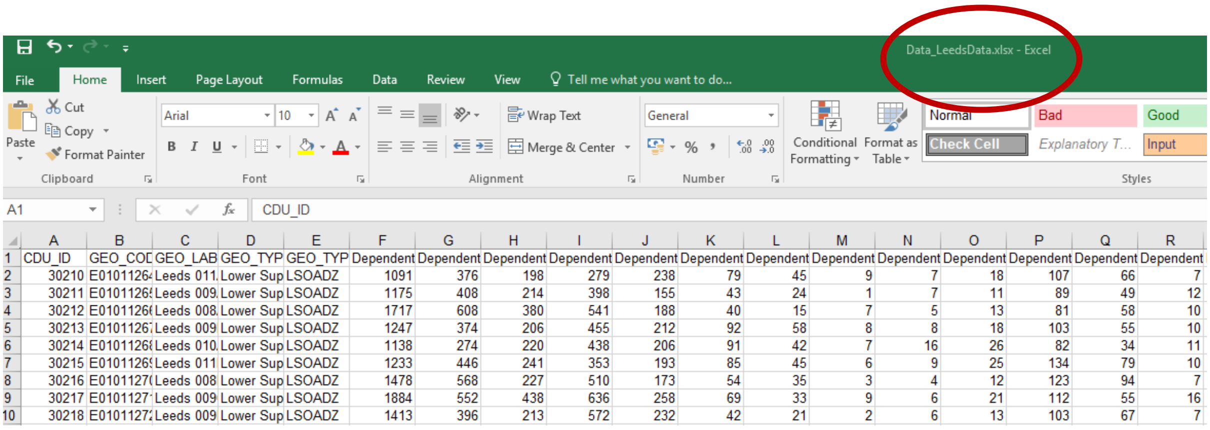
- ⊟ □ West Yorkshire
  - ⊟ □ Local Authorities (5 areas)
    - ⊞ □ Bradford
    - ⊞ □ Calderdale
    - ⊞ □ Kirklees
    - ⊟ □ Leeds
      - ⊞ □ Wards and Electoral Divisions (33 areas)
      - ⊞ □ Middle Super Output Areas and Intermediate Zones (107 areas)
      - ⊞ ☑ Lower Super Output Areas and Data Zones (482 areas)
      - ⊞ □ Output Areas and Small Areas (2543 areas)
    - ⊞ □ Wakefield
  - ⊞ □ Wards and Electoral Divisions (124 areas)
  - ⊞ □ Middle Super Output Areas and Intermediate Zones (299 areas)
  - ⊞ □ Lower Super Output Areas and Data Zones (1388 areas)
  - □ Output Areas and Small Areas (7131 areas)
- ⊞ □ Worcestershire
- ⊞ □ Local Authorities (404 areas)
- □ Wards and Electoral Divisions (9481 areas)
- □ Middle Super Output Areas and Intermediate Zones (8480 areas)
- □ Lower Super Output Areas and Data Zones (42619 areas)
- □ Output Areas and Small Areas (232296 areas)

11. Press *Add*, so it is listed in the box at the bottom of the screen.
12. Press *Next*.
13. You are now at the Downloads page, where it lists the data you will be downloading. At the bottom of the screen there is a *File Reference* option – give this file a logical name (mine is LeedsData), and press *Get the Data.*
14. When your download is ready, a red button appears saying *Download Data*. Press this, and save your downloaded file to a logical place in your directory.
15. Unzip the folder (right click on it, press *Extract All*, and then *Extract*).
16. You can see three files:
    a. *Data_LeedsData* contains the data we will anayse.
    b. *Meta_LeedsData* contains the metadata – information about the data we have downloaded (for example in this dataset, definitions of what constitutes a dependent child).
    c. *Citations* contains information of how to appropriately reference this data.
17. You have now successfully downloaded data from InFuse and are ready to move on to part 2!

13th February 2020
Author: Verity Tether

Part 2: Analysing Data in Excel

1. Firstly, we need to put your file into a format which is appropriate for editing as it is currently in a .csv format. Although this is a very useful file format and can be used with a variety of software packages, it does not always save changes made and can be a little glitchy when you just want to use excel on its own – as it's a text file it doesn't save things like formatting and formulae. For this reason, we are going to use a .xlsx file format – the excel spreadsheet.
    a. Open up your Data_LeedsData file.
    b. Go to File → Save As.
    c. Navigate to your current folder.
    d. From the drop down box labelled *Save as type* (currently listed as *CSV (Comma Delimited)(*.csv)*), select *Excel Workbook (*.xlsx)*.
    e. Press Save.
    f. Your spreadsheet should now have the title *Data_LeedsData.xlsx* at the top, as in the screenshot below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CDU_ID	GEO_COD	GEO_LAB	GEO_TYP	GEO_TYP	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
2	30210	E01011264	Leeds 011.	Lower Sup	LSOADZ	1091	376	198	279	238	79	45	9	7	18	107	66	7
3	30211	E01011265	Leeds 009.	Lower Sup	LSOADZ	1175	408	214	398	155	43	24	1	7	11	89	49	12
4	30212	E01011266	Leeds 008.	Lower Sup	LSOADZ	1717	608	380	541	188	40	15	7	5	13	81	58	10
5	30213	E01011267	Leeds 009	Lower Sup	LSOADZ	1247	374	206	455	212	92	58	8	8	18	103	55	10
6	30214	E01011268	Leeds 010.	Lower Sup	LSOADZ	1138	274	220	438	206	91	42	7	16	26	82	34	11
7	30215	E01011269	Leeds 011	Lower Sup	LSOADZ	1233	446	241	353	193	85	45	6	9	25	134	79	10
8	30216	E01011270	Leeds 008	Lower Sup	LSOADZ	1478	568	227	510	173	54	35	3	4	12	123	94	7
9	30217	E01011271	Leeds 009	Lower Sup	LSOADZ	1884	552	438	636	258	69	33	9	6	21	112	55	16
10	30218	E01011272	Leeds 009	Lower Sup	LSOADZ	1413	396	213	572	232	42	21	2	6	13	103	67	7

2. Now, open up your LeedsData file so we can have a look at it. There is a lot of data here, so let's pick it apart:
    a. Row 1: this contains unique codes relating to the data you have downloaded.
    b. Row 2: this contains a description of the data.
    c. Columns A-E: these contain details of each LSOA in your dataset. The data stored in columns A-C are unique to each LSOA, and can be used to identify specific areas.
    d. Columns F-J: these contain the counts of people in each LSOA who have a long-term health problem or disability (henceforth referred to as "disability").
        i. Column F: the total number of people with disabilities in the LSOA.

   ii. Column G: the number of people with disabilities in the LSOA who have no children.

   iii. Column H: the number of people with disabilities in the LSOA who have one dependent child in the family.

   iv. Column I: the number of people with disabilities in the LSOA who have two or more dependent children in the family.

   v. Column J: the number of people with disabilities in the LSOA who have children, but they are non-dependent.

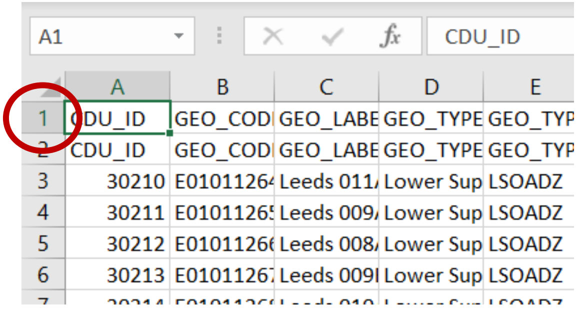
e. Columns K-O: these contain the counts of people in each LSOA who have disabilities whose day-to-day activities are “limited a lot” by their disability. These are broken down in the same manner as in d(i) – d(v) above.

f. Columns P-T: these contain the counts of people in each LSOA who have disabilities whose day-to-day activities are “limited a little” by their disability. These are broken down in the same manner as in d(i) – d(v) above.

g. Columns U-Y: these contain the counts of people in each LSOA who have disabilities whose day-to-day activities not limited by their disability. These are broken down in the same manner as in d(i) – d(v) above.

3. As we don’t need all the codes for the variables which are stored in row 1, we will delete this. However, before we do that, let’s copy and paste the titles for columns A-E from row 1 to row 2. Highlight them all by clicking on one and dragging it across to the other cells. Then click copy. Click on cell A2, right click and click paste.

4. We can then delete row 1. Right click on the 1 (as circled below) and click delete.



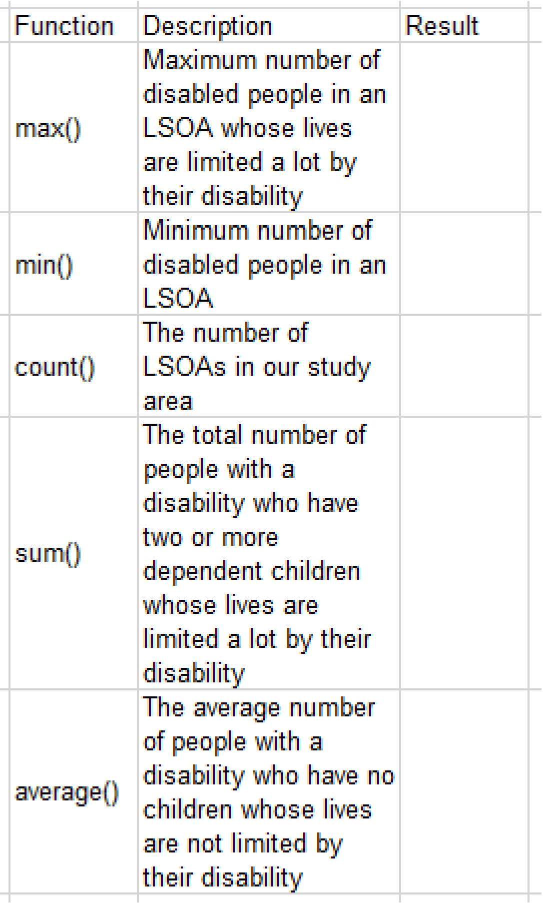
	A	B	C	D	E
1	CDU_ID	GEO_COD	GEO_LABE	GEO_TYPE	GEO_TYP
2	CDU_ID	GEO_COD	GEO_LABE	GEO_TYPE	GEO_TYP
3	30210	E0101126	Leeds 011	Lower Sup	LSOADZ
4	30211	E0101126	Leeds 009	Lower Sup	LSOADZ
5	30212	E0101126	Leeds 008	Lower Sup	LSOADZ
6	30213	E0101126	Leeds 009	Lower Sup	LSOADZ

5. If we scroll down in our spreadsheet to have a look at our data, we lose the top rows which contain the description of the data held in them, so it can be hard to remember which variables we are looking at. To resolve this, we use **Freeze Panes.**

6. On the top of the screen, navigate to *View* and click *Freeze Panes*. This gives us the option of freezing just the top row, so we always know which variables we are looking at.

Basic Functions

1. Excel has a great number of functions (predefined formulae) built in to facilitate analysis. Some of these are fairly basic, and some are very complex. To begin this analysis, we are going to use some of Excel's built-in functions to have a look at the data, as having a basic understanding of our dataset is a fundamental start to any project.
2. Firstly, let's make a small table where we will put our results from this exploration:
    a. Scroll towards the bottom of the spreadsheet, where there is no data.
    b. Select any cell, and create the following table (you don't have to type out the description, but it may help if you choose to come back to this workshop):



Function	Description	Result
max()	Maximum number of disabled people in an LSOA whose lives are limited a lot by their disability	
min()	Minimum number of disabled people in an LSOA	
count()	The number of LSOAs in our study area	
sum()	The total number of people with a disability who have two or more dependent children whose lives are limited a lot by their disability	
average()	The average number of people with a disability who have no children whose lives are not limited by their disability	

    c. The () after each function name represents the fact that we are going to insert data into the function. The description of the data reflects a specific column in the table, and it is the data in that column which we will insert into the function to get our results.
    d. In order to run these functions, click in the corresponding *Result* box and type the following. When you're done with each, press enter and a value will appear. I will put the results at the bottom of this file if you'd like to check them.

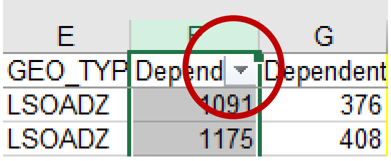
i. =MAX(K2:K483)

1. The equals sign starts the formula, and shows that you aren't just typing text into the box
2. MAX() is the formula we want to use
3. K2:K483 is the range of cells that we would like the maximum value for. I.e. the data in all cells from cell K2 (where the data starts), to cell K483 (where the data ends). If you are confused as to why we are using row K, ask one of us.

ii. Now that you have the structure to complete these functions, have a go at filling in the table with the rest of the functions. As before, the answers will be at the end of this document if you'd like to check.

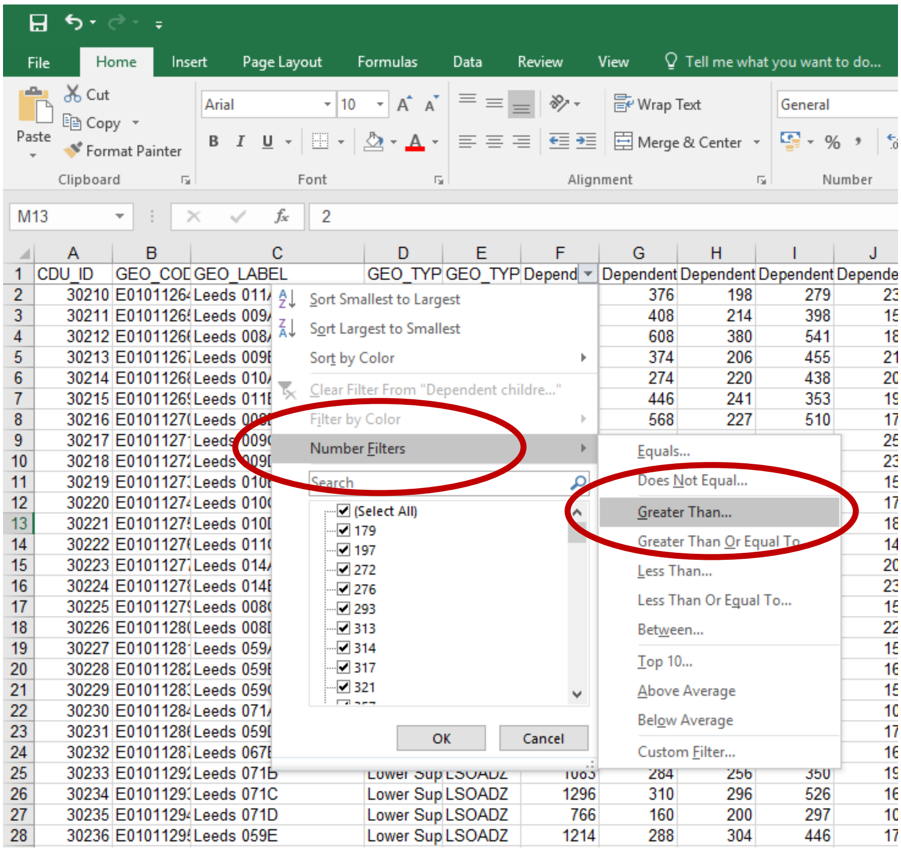
Filtering

1. Say we want to look at only those LSOAs which have more than 1,000 people with disabilities. To do this, we can add a *Filter*.
2. Highlight the whole of column F – the total number of people in the LSOA with a disability. You can select the whole column by clicking the F at the top.
3. On the top bar, go to *Data* → *Filter.* This will insert a small drop-down arrow onto cell F1, as below:



E	F	G
GEO_TYP	Depend	Dependent
LSOADZ	1091	376
LSOADZ	1175	408

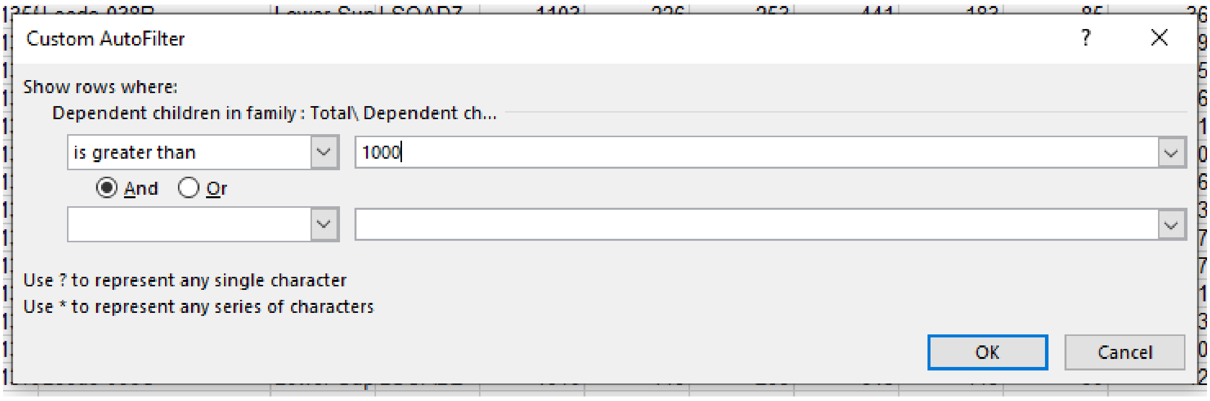
4. Click on this arrow, hover over *Number Filters*, then click on *Greater Than*:



M13 2

	A	B	C	D	E	F	G	H	I	J
1	CDU_ID	GEO_COD	GEO_LABEL	GEO_TYP	GEO_TYP	Depend	Dependent	Dependent	Dependent	Depende
2	30210	E0101126	Leeds 011				376	198	279	23
3	30211	E0101126	Leeds 009				408	214	398	15
4	30212	E0101126	Leeds 008				608	380	541	18
5	30213	E0101126	Leeds 009				374	206	455	21
6	30214	E0101126	Leeds 010				274	220	438	20
7	30215	E0101126	Leeds 011				446	241	353	19
8	30216	E0101127	Leeds 00				568	227	510	17
9	30217	E0101127	Leeds 009							25
10	30218	E0101127	Leeds 009							23
11	30219	E0101127	Leeds 010							15
12	30220	E0101127	Leeds 010							17
13	30221	E0101127	Leeds 010							18
14	30222	E0101127	Leeds 011							14
15	30223	E0101127	Leeds 014							20
16	30224	E0101127	Leeds 014							23
17	30225	E0101127	Leeds 008							15
18	30226	E0101128	Leeds 008							22
19	30227	E0101128	Leeds 059							15
20	30228	E0101128	Leeds 059							16
21	30229	E0101128	Leeds 059							15
22	30230	E0101128	Leeds 071							10
23	30231	E0101128	Leeds 059							17
24	30232	E0101128	Leeds 067							16
25	30233	E0101129	Leeds 071B	Lower Sup	LSOADZ	1083	284	256	350	19
26	30234	E0101129	Leeds 071C	Lower Sup	LSOADZ	1296	310	296	526	16
27	30235	E0101129	Leeds 071D	Lower Sup	LSOADZ	766	160	200	297	10
28	30236	E0101129	Leeds 059E	Lower Sup	LSOADZ	1214	288	304	446	17

5. In the box which appears, type 1000, and press *OK*:



Custom AutoFilter

Show rows where:

Dependent children in family : Total\ Dependent ch...

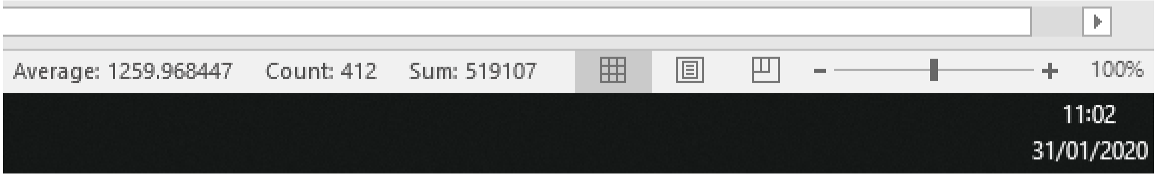
is greater than | 1000

● And ○ Or

Use ? to represent any single character

Use * to represent any series of characters

6. To check that this has worked, we will count how many rows are in the columns now. This will show us that it has removed LSOA rows with more than 1000 people with disabilities.
    a. Note: it is important to note the wording here. By doing this, we have selected any rows which have **more than 1000** people. It will therefore discount any row which has exactly 1000. If we wanted to include 1000, we would have used the *Greater than Or Equals To* option from the drop down box in step 4.
7. To count the number of rows in our sheet, we could use the count() function we were introduced to earlier. But, there is another (possibly easier) way:
    a. Click in any cell on the top row of data (I selected F2) – **not** the row with the column headers.
    b. At the same time, press *CTRL*, *shift* and the *down arrow* on your keyboard. This will highlight all the data in the column.
    c. In the bottom right-hand corner, Excel gives a summary of the cells you have selected – the average value, the count of cells, and the total sum within the cells:



Average: 1259.968447 Count: 412 Sum: 519107

    d. Given that, from our use of the functions before, we identified that there are 482 cells in total, we can tell that 70 LSOAs had fewer than 1,000 people with disabilities.

8. In order to use this data (with the 70 LSOAs removed) as our primary spreadsheet from now on, copy the data into a new sheet. The best way to do this is to use a similar keyboard shortcut as in 7b:
    a. Click in cell A1 (the top left hand corner).
    b. At the same time, press *CTRL*, *shift* and the *down arrow* on your keyboard. This will highlight all the data in the column.
    c. Then, press *CTRL*, *shift* and the *right arrow* all together. This will highlight all the columns to the right of the current cell.
    d. Press *File* → *New* → *Blank Workbook.*
    e. Go to the sheet with the highlighted data. Press *CTRL* and *c* to copy the data.
    f. Go to the blank sheet, click in cell A1 and press *CTRL* and *v* to paste the data.
    g. Press *File* → *Save As*, navigate to your folder and save this new file as *LeedsDataNew.*
    h. This *LeedsDataNew* file is the one we will be using from now on.

VLOOKUP ()

1. VLOOKUP is one of Excel's most useful functions. We can use it to retrieve information from a table, based on a unique identifier.
2. The syntax of VLOOKUP is as follows:

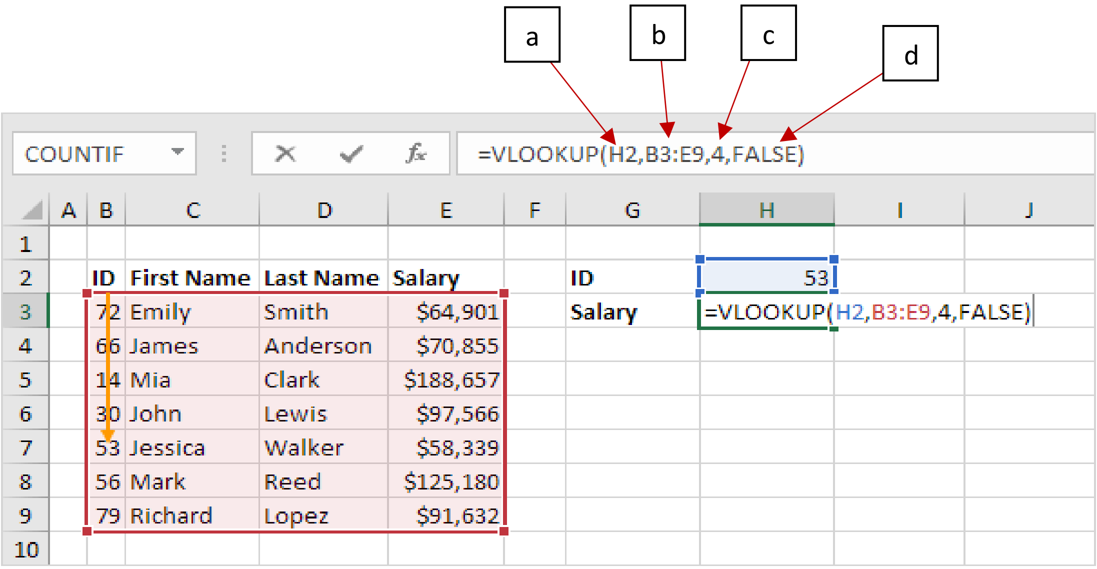
   =VLOOKUP (value, table, col_index, [range_lookup])

   Where:

    a. Value – the value that we are looking for (in this case, the LSOA code)
    b. Table – the table we are looking at
    c. Col_index – the column in the table from which we want to retrieve a value
    d. Range_lookup – if we put "TRUE", it retrieves an approximate match. If we put "FALSE", it retrieves an exact match.

   Although this can look quite confusing, it is not complex in person! Have a go at the next few steps, and if you are confused one of us can help you.

3. Before we have a go at doing a VLOOKUP, here is an example from this website. This example uses VLOOKUP to identify the salary of employee number 53.



=VLOOKUP(H2,B3:E9,4,FALSE)

	A	B	C	D	E	F	G	H	I	J
1										
2		**ID**	**First Name**	**Last Name**	**Salary**		**ID**	53		
3		72	Emily	Smith	$64,901		**Salary**	=VLOOKUP(H2,B3:E9,4,FALSE)		
4		66	James	Anderson	$70,855					
5		14	Mia	Clark	$188,657					
6		30	John	Lewis	$97,566					
7		53	Jessica	Walker	$58,339					
8		56	Mark	Reed	$125,180					
9		79	Richard	Lopez	$91,632					
10										

Let’s look at the structure of the formula:

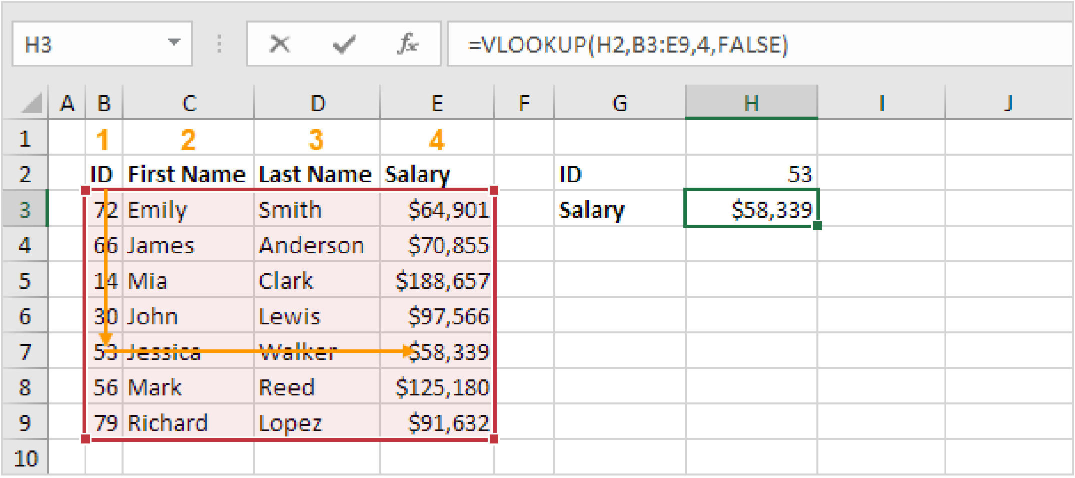
a. Value – the value that we are looking for. Here, cell H2 is called, as the ID we are looking for, 53, has been typed into this cell.

b. Table – the table we are looking at. Here, cells B3 – E9 are selected, as this covers the whole table we are looking at. Note – the ID value must be located in the left-hand column.

c. Col_index – the column in the table from which we want to retrieve a value. Here, the salary data is in column 4 of the table, so we put a 4.

d. Range_lookup – if we put “TRUE”, it retrieves an approximate match. If we put “FALSE”, it retrieves an exact match. As we want the exact salary of employee number 53, we put FALSE.

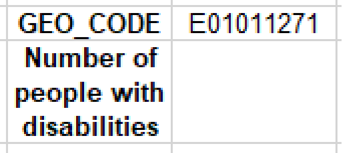
So, the formula reads across the row in which ID number 53 is stored, goes to column 4, and selects that data, as demonstrated below:



=VLOOKUP(H2,B3:E9,4,FALSE)

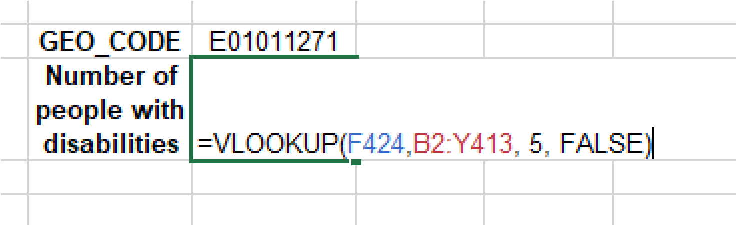
	A	B	C	D	E	F	G	H	I	J
1		1	2	3	4					
2		ID	First Name	Last Name	Salary		ID	53		
3		72	Emily	Smith	$64,901		Salary	$58,339		
4		66	James	Anderson	$70,855					
5		14	Mia	Clark	$188,657					
6		30	John	Lewis	$97,566					
7		53	Jessica	Walker	$58,339					
8		56	Mark	Reed	$125,180					
9		79	Richard	Lopez	$91,632					
10										

4. Now, we will try ourselves!
   a. Open your *LeedsDataNew* sheet.
   b. Say we want to identify how many people have a disability in LSOA Leeds 009C, which is in Guiseley, because we are considering updating the accessibility of buildings there.
   c. Locate this LSOA in your table – for me it is row 9.
   d. Find the geocode for it, which is in row B (so, for me, cell B9).
   e. Create the following table to the right of the table in your excel document:



GEO_CODE	E01011271
Number of people with disabilities	

   f. Enter the following formula into the blank cell, and press enter. Note, the value (here, F424), will correspond to the cell in which you have typed E01011271:



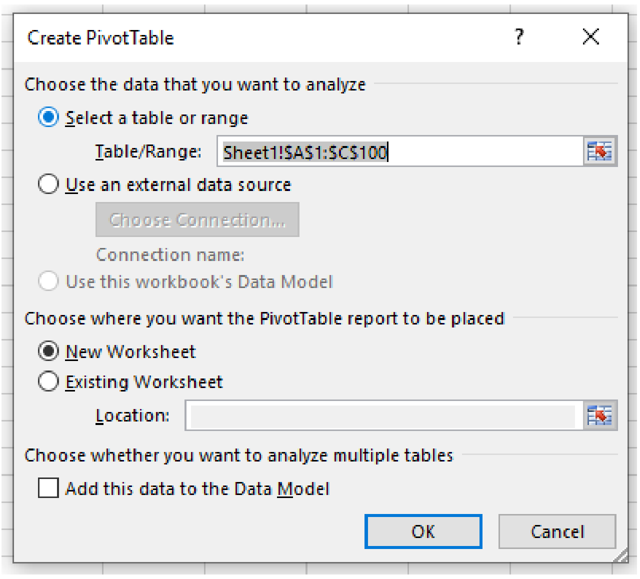
GEO_CODE	E01011271			
Number of people with disabilities	=VLOOKUP(F424,B2:Y413, 5, FALSE)			

   g. This gives us the answer 1,884 – there are 1,844 people with disabilities in that particular LSOA.
   h. If you have any questions about this formula, feel free to ask one of us!

5. If you'd like to try one for yourself, calculate the following (the answer is at the end of the document):
    a. How many people live in LSOA 011C, who have two or more children and whose day-to-day activities are limited a lot?

Pivot Tables

1. Pivot tables are another incredibly helpful excel function. They allow us to summarize large amounts of data into one more manageable table.
2. We are going to create a pivot table, using a dataset I have created which has been based on the census data we have been using. **Note: this data is not accurate and has been created purely for this learning exercise.**
3. Navigate to *PivotTableData.xlsx* on Github, and download it (click on the file, and then *Download* on the following page). Open the spreadsheet, and have a look at the data on the PivotTable sheet. It provides information on people living in an LSOA:
    a. Disability? – this column identifies whether the respondent has a disability or not.
    b. Limited? – this column identifies how much their day-to-day behaviour is limited by their disability. If they do not have a disability, this is listed as NA.
    c. Number of dependent children – this column counts the number of dependent children in the household.
4. We are going to create a pivot table that will enable us to see how many children live in households with parents who have a disability, as well as the average number of children to adults with a disability.
5. On Excel, click in a cell in the table. Then, go to *Insert* → *Pivot Table.* The box which opens automatically selects all the data in the table, and says that it will place the Pivot Table in a new worksheet. Press *OK*.


Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: Sheet1!$A$1:$C$100

○ Use an external data source

Connection name:

○ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

● New Worksheet

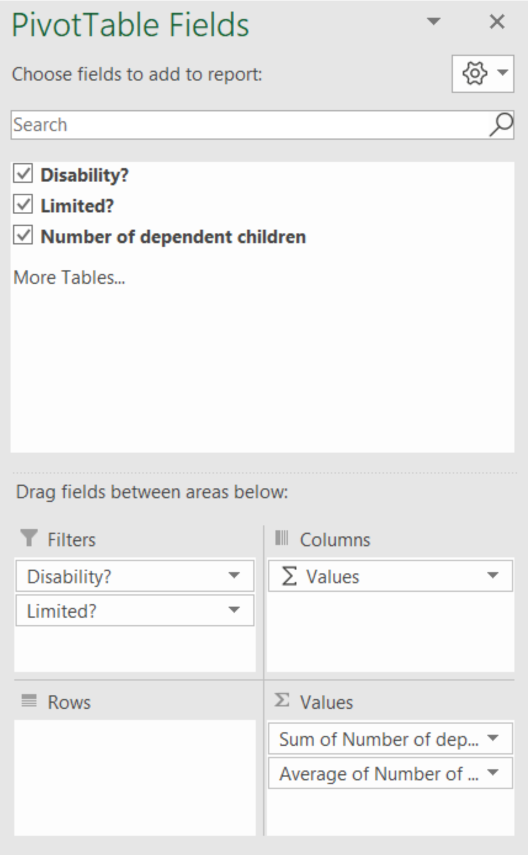
○ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

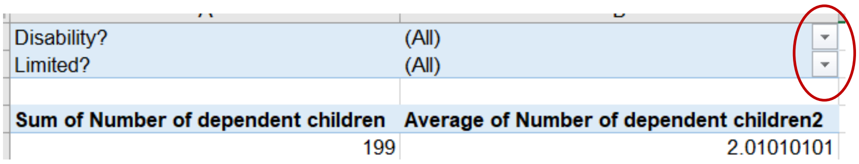
□ Add this data to the Data Model

6. To create the pivot table itself, we drag options on the *pivot table fields pane* which appears on the new sheet:
   a. For our table, we want to be able to filter both the *Disability?* and the *Limited?* fields. So, we drag these into the *Filters* box.
   b. We then drag the *Number of dependent children* field into the *Values* box. This automatically sums these values to create a total.
   c. We then drag this field into the *Values* box again. By default, this will calculate the sum again, so to get it to calculate the average:
      i. Left click on it to highlight it.
      ii. Select *Value Field Settings*.
      iii. In the *Summarize value field by…* option, select Average.
      iv. Press ok.
   d. When finished, the fields pane and the pivot table will look like this:


PivotTable Fields
Choose fields to add to report:
Search
More Tables...
Drag fields between areas below:

☑ **Disability?**
☑ **Limited?**
☑ **Number of dependent children**

Filters: Disability?, Limited?
Columns: Σ Values
Rows:
Values: Sum of Number of dep..., Average of Number of ...



Disability?	(All)
Limited?	(All)
Sum of Number of dependent children	**Average of Number of dependent children2**
199	2.01010101

7. We can use the highlighted buttons on the pivot table above to look at the total number of children, as well as the average number of children, by how much the parent are affected by their disability.

13th February 2020
Author: Verity Tether

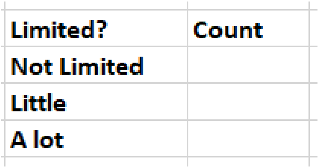
Data Visualisation in Excel

Being able to display your data clearly and concisely is a valuable tool for all data analysts. We will go through a few options that excel gives you for this.

Categorical Data: Preparing the Data

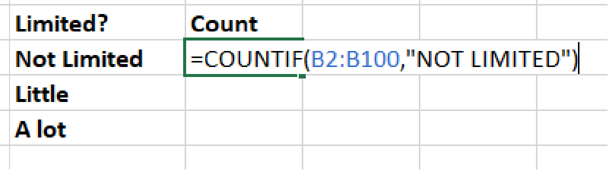
Categorical data is data which can be categorised into groups. In our case, a good example of categorical data is in the PivotTableData file – how limited people are by their disability. We are going to create pie charts and bar charts using this data, but first, we need to put it into a format where we can do so. To do this, we are going to use another function.

1. Open the PivotTableData file, and create the following table:



Limited?	Count
Not Limited	
Little	
A lot	

2. We are going to use the COUNTIF() function. This function does as its name suggests: it counts the frequency of cells which meet certain criteria. In the case below, it counts the number of times “NOT LIMITED” is listed in the data.



Limited?	Count
Not Limited	=COUNTIF(B2:B100,"NOT LIMITED")
Little	
A lot	

3. Using the screenshot above as a guide, can you fill in the function to get the frequency of the occurrence of “little” and “a lot”? The answers are at the end of the sheet.

Pie Charts and Bar Charts

1. To create a pie chart, highlight the small table we just made. Click *Insert*, and navigate to the pie chart symbol. This will create a 2D pie chart with some default settings. We can make the following changes:
    a. Double click on the default title “Count” to rename it.

   b. Under Design, click *Add Chart Element* (top left-hand corner), and select *Data Labels* to add values to the pie segments.
   c. Click *Quick Layout* (top left-hand corner) to look at a number of pre-designed pie chart layouts. *Chart Styles* also allows us to explore different designs.
2. A bar chart can be added in much the same way. Feel free to explore this option and ask if you have any questions.

Histogram

A histogram is used to demonstrate the distribution of continuous numerical data. It splits numbers into ranges, allowing us to see how many values fall into these ranges.

1. On the *LeedsDataNew* spreadsheet, highlight column F – that showing the total number of people with disabilities across Leeds.
2. Go to *Insert* → *Insert Statistics Chart* and click the symbol underneath *Histogram*
3. As before, the *Add Chat Element* button can be used to add things such as axes and grid lines.

Boxplots

A boxplot is used to graphically display distribution of a data set. It also shows us whether there are any outliers in the data (a value which differs significantly from the rest in the dataset). This can be very valuable when examining the data before you start analysing it.

1. On the *LeedsDataNew* spreadsheet, highlight row L – people who have no children whose day-to-day activities are limited a lot
2. Go to *Insert* → *Insert Statistics Chart* and click the symbol underneath *Box and Whisker*
3. The resulting plot can be edited in the same way, adding things like data labels which can be helpful when interpreting a boxplot.
4. If you are not sure how to interpret a boxplot, this website is very clear.

5. We can also create a boxplot with multiple variables, so that we can compare their distributions.
6. We are going to create a boxplot comparing the following variables:
   a. People who have no children whose day-to-day activities are limited a lot.
   b. People who have no children whose day-to-day activities are limited a little.
   c. People who have no children whose day-to-day activities are not limited.
7. Highlight the columns L, Q and V simultaneously by clicking the L, Q and V letters whilst keeping the control button on the keyboard pressed.

8. Repeat steps 2 and 3 above, adding a legend so we can see which boxplot refers to which variable.

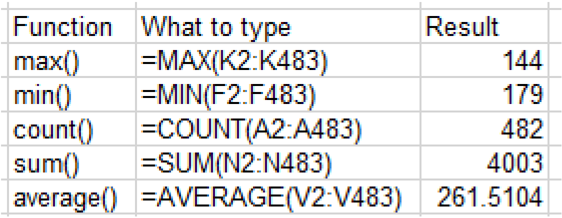
Wrap-Up

- This workshop has introduced you to census data, showing you how to access and download data from InFuse.
- You have gained understanding of UK census geographies.
- You have been introduced to fundamental excel tools, including basic functions, filtering, VLOOKUP and pivot tables.
- You have also explored data visualisation in excel. Note, there are a really wide number of graphs which can be created in excel, including scatterplots and line graphs. These have not been examined here as the data we have is not appropriate for them, but now that you understand how to visualise data in excel, you could do this style of graph on your own data with no trouble.

13th February 2020
Author: Verity Tether

Answers

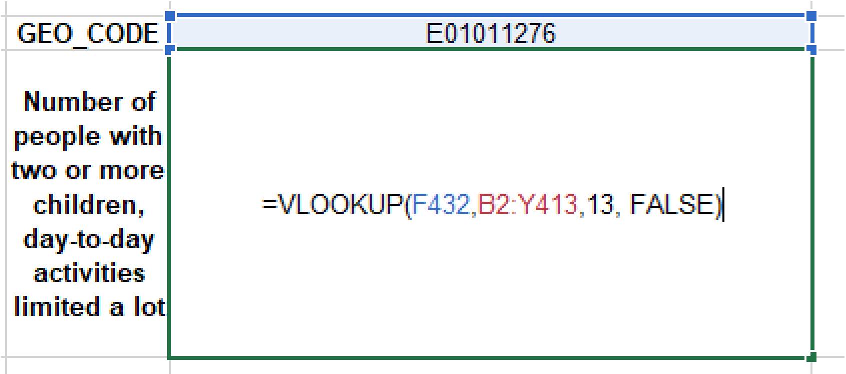
Functions



Function	What to type	Result
max()	=MAX(K2:K483)	144
min()	=MIN(F2:F483)	179
count()	=COUNT(A2:A483)	482
sum()	=SUM(N2:N483)	4003
average()	=AVERAGE(V2:V483)	261.5104

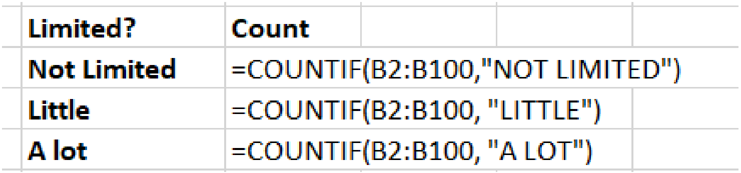
VLOOKUP

Answer = 5



GEO_CODE	E01011276
Number of people with two or more children, day-to-day activities limited a lot	=VLOOKUP(F432,B2:Y413,13, FALSE)

Categorical Data: COUNTIF()



Limited?	**Count**
Not Limited	=COUNTIF(B2:B100,"NOT LIMITED")
Little	=COUNTIF(B2:B100, "LITTLE")
A lot	=COUNTIF(B2:B100, "A LOT")