Week 3: Identify Risk Factors for Infection

UPDATE

Thank you again for the previous analysis. We will next be publishing a public health advisory that warns of specific infection risk factors of which individuals should be aware. Please advise as to which population characteristics are associated with higher infection rates. </span>

Your goal for this notebook will be to identify key potential demographic and economic risk factors for infection by comparing the infected and uninfected populations.

Imports

In [2]:

```
import cudf
import cuml
```

Load Data

Begin by loading the data you've received about week 3 of the outbreak into a cuDF data frame. The data is located at `./data/week3.csv` . For this notebook you will need all columns of the data.

In [6]:

```
gdf = cudf.read_csv('./data/week3.csv')
```

Calculate Infection Rates by Employment Code

Convert the `infected` column to type `float32` . For people who are not infected, the float32 `infected` value should be `0.0` , and for infected people it should be `1.0` .

In [7]:

```
gdf.astype({'infected': 'float32'})
```

Out[7]:

	age	sex	employment	infected
0	0	m	U	0.0
1	0	m	U	0.0
2	0	m	U	0.0
3	0	m	U	0.0
4	0	m	U	0.0
...	...	...	...	...
58479889	90	f	V	0.0
58479890	90	f	V	0.0
58479891	90	f	V	0.0
58479892	90	f	V	0.0
58479893	90	f	V	0.0

58479894 rows × 4 columns

Now, produce a list of employment types and their associated **rates** of infection, sorted from highest to lowest rate of infection.

NOTE: The infection **rate** for each employment type should be the percentage of total individuals within an employment type who are infected. Therefore, if employment type "X" has 1000 people, and 10 of them are infected, the infection **rate** would be .01. If employment type "Z" has 10,000 people, and 50 of them are infected, the infection rate would be .005, and would be **lower** than for type "X", even though more people within that employment type were infected.

In [9]:

```
infected = gdf.groupby(['employment']).agg({'infected':'sum'})
count = gdf.groupby(['employment']).agg({'infected':'count'})
(infected/count).sort_values(by=['infected'], ascending=False)
```

Out[9]:

	infected
employment	
Q	0.012756
I	0.010354
V	0.007590
P	0.006190
Z	0.005655
R, S, T	0.005390
O	0.005284
L	0.004970
G	0.004948
N	0.004784
M	0.004777
K	0.004772
X	0.004539
J	0.003939
C	0.003882
A	0.003853
B, D, E	0.003774
H	0.003388
F	0.003182
U	0.000217

Finally, read in the employment codes guide from `./data/code_guide.csv` to interpret which employment types are seeing the highest rates of infection.

In [10]:

```
ecg = cudf.read_csv('./data/code_guide.csv')
ecg
```

Out[10]:

	Code	Field
0	A	Agriculture, forestry & fishing
1	B, D, E	Mining, energy and water supply
2	C	Manufacturing
3	F	Construction
4	G	Wholesale, retail & repair of motor vehicles
5	H	Transport & storage
6	I	Accommodation & food services
7	J	Information & communication
8	K	Financial & insurance activities
9	L	Real estate activities
10	M	Professional, scientific & technical activities
11	N	Administrative & support services
12	O	Public admin & defence; social security
13	P	Education
14	Q	Human health & social work activities
15	R, S, T	Other services
16	U	Student
17	V	Retired
18	X	Outside the UK or not specified
19	Y	Pre-school child
20	Z	Not formally employed

Calculate Infection Rates by Employment Code and Sex

We want to see if there is an effect of `sex` on infection rate, either in addition to `employment` or confounding it. Group by both `employment` and `sex` simultaneously to get the infection rate for the intersection of those categories.

In [11]:

```
infected = gdf.groupby(['employment','sex']).agg({'infected':'sum'})
count = gdf.groupby(['employment','sex']).agg({'infected':'count'})
(infected/count).sort_values(by=['infected'], ascending=False)
```

Out[11]:

		infected
employment	sex	
I	f	0.015064
Q	f	0.014947
V	f	0.010852
B, D, E	f	0.007973
R, S, T	f	0.007748
O	f	0.007719
K	f	0.007672
M	f	0.007645
J	f	0.007645
C	f	0.007630
Z	f	0.007629
P	f	0.007584
F	f	0.007577
G	f	0.007556
A	f	0.007491
X	f	0.007391
N	f	0.007389
H	f	0.007385
L	f	0.007221
Q	m	0.005120
I	m	0.005117
V	m	0.003685
G	m	0.002596
P	m	0.002577
C	m	0.002569
J	m	0.002546
O	m	0.002543
Z	m	0.002543
R, S, T	m	0.002542
N	m	0.002538
F	m	0.002535
M	m	0.002520

		infected
employment	sex	
A	m	0.002514
K	m	0.002490
H	m	0.002482
B, D, E	m	0.002462
X	m	0.002435
L	m	0.002197
U	f	0.000329
	m	0.000110

Take the Assessment

After completing the work above, visit the *Launch Section* web page that you used to launch this Jupyter Lab. Scroll down below where you launched Jupyter Lab, and answer the question *Week 3 Assessment*. You can view your overall progress in the assessment by visiting the same *Launch Section* page and clicking on the link to the *Progress* page. On the *Progress* page, if you have successfully answered all the assessment questions, you can click on *Generate Certificate* to receive your certificate in the course.

launch_task_page

Optional: Restart the Kernel

If you plan to continue work in other notebooks, please shutdown the kernel.

In [ ]:
```
import IPython
app = IPython.Application.instance()
app.kernel.do_shutdown(True)
```