Arise Citizen Scientist 2020: Ask Questions

by M.A.Tucker 27APR20

**Motivation**

A favorite satirist and punster, Alexandra Petri (Twitter: @petridishes), told a story of how Newton, fleeing to the countryside during a plague, decided to use the time to work out the math leading to the gravitational mechanics we still use today. She encouraged us to do something different during this time, something equivalent to Newton's effort. I was inspired. I had just such a thought in early February as our plague picked up steam. Above the TODO list items such as "paint bathroom ceiling" or "fix leaky faucet", I had placed "discover anti-gravity". It turns out that anti-gravity is a REALLY HARD PROBLEM. So as an alternative, perhaps slightly more achievable goal, I decided to try to better understand our COVID travails. By asking and answering questions, I hoped to better judge the veracity of often conflicting pronouncements.

**Method**

Citizen Scientists have an important role to play in understanding our current COVID travails. There are hundreds of thousands of Citizen Scientists contributing worldwide. The following leverages their extensive knowledge.

We are blessed in the USA to have ready access to a vast amount of data and powerful computer resources (much of it free). Data capturing various aspects of the plague are maintained and updated daily by the CDC, Kaggle.com, Johns Hopkins and many other organizations. However, huge streams of numbers purportedly describing something become rather mind numbing. My goal as a budding Citizen Scientist is to understand the data and capture the essence in some visualizations (e.g. graphs, tables). In short, convert "data" into "facts". How hard could that be?

By looking at smaller sized areas, you may discern aspects not apparent if all of the USA or world is considered. Creating a Kaggle notebook at www.kaggle.com/imdevskp/corona-virus-report allows you to easily use the latest USA-County-Wise data. The answers below are based on a snapshot of this data as of April 27th, 2020. It changes by the hour with updates posted daily.

Here is a link to a Johns Hopkins dashboard focusing on the same USA county wise data (<https://coronavirus.jhu.edu/us-map>). If you reference the Johns Hopkins dashboard tomorrow, you'll likely see different results. It will have been updated while the data used for this post will be dated.

**Questions**

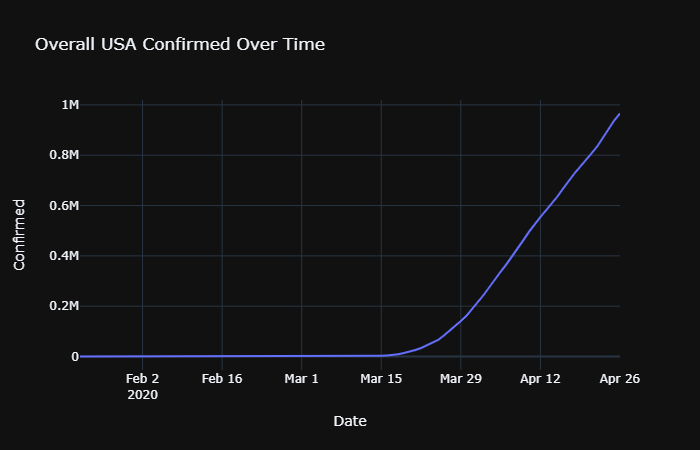
The following questions and some "answers" are based on our best (but undoubtedly flawed) available data (based on data gathered through April 27th, 2020).

Is it a hoax?

No, it's a modern day plague.

Is it almost over?

No, it's spreading rapidly with increasing numbers of confirmed cases & deaths. If the number of daily new cases begins to decline, we'll see the "curve flatten". In some areas (e.g. Pittsburgh, PA), there were no deaths reported yesterday after a steady drumbeat. Also, the "doubling rate" has extended from 2 days to 2 weeks in some areas. This is very good news. However, it's still doubling.



Do various US county sized areas "look alike"?

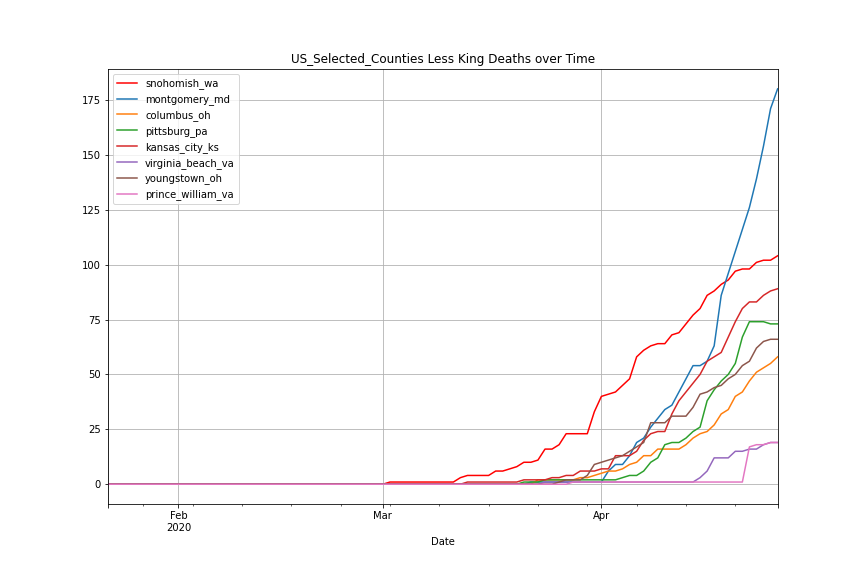
All areas are experiencing increasing numbers of confirmed cases. The meaningful differences relate to infection rates (confirmed cases given the population size of the area) and mortality rates (deaths given the number of infections). Choosing a mix of mid-sized cities and suburban areas vs. a baseline (first reported cases in USA) of King County, WA, leads to more questions than answers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **region** | **population** | **confirmed cases** | **infection rate** | **deaths** | **mortality rate** |
| King County, WA | 2,253,000 | 5863 | 0.26% | 408 | 6.96% |
| Pittsburgh, PA | 1,216,000 | 1211 | 0.10% | 73 | 6.03% |
| Kansas City, KS | 890,089 | 1043 | 0.12% | 67 | 6.42% |
| Youngstown, OH | 228,683 | 737 | 0.32% | 66 | 8.96% |
|  |  |  |  |  |  |
| Montgomery County, MD | 1,051,000 | 3645 | 0.35% | 180 | 4.94% |
| Snohomish, WA | 822,000 | 2311 | 0.28% | 104 | 4.50% |
|  |  |  |  |  |  |
| Columbus, OH | 1,683,751 | 2199 | 0.13% | 58 | 2.64% |
| Virginia Beach+, VA | 910,385 | 732 | 0.08% | 19 | 2.60% |
| Prince William, VA | 470,335 | 1265 | 0.27% | 19 | 1.50% |

King County, Pittsburgh, Kansas City and Youngstown, OH all have a very high mortality rate (while varying in population from 2.2 million to a couple hundred thousand). Why? Due to limited testing, the actual number of infections is likely much higher. Which leads to the mortality rates being inflated. Many cited mortality rates are much lower (~2%) due to modeling of infection rates in the population rather than based on the known data.

Montgomery County, MD and Snohomish, WA have similar mid-range mortality rates. Montgomery County, MD recently surged past Snohomish, WA. Since February, Snohomish, WA had been near the top of the leaderboard. Why is Montgomery County surging? It should be noted that this is not a contest you want to win. Montgomery County, MD (my home) does not appear to be leveling off like many areas. Why? We've closed restaurants, bars, venues, "non-essential" businesses since early on. We're wearing masks, social-distancing, etc. Are these measures inadequate?

Columbus, OH, greater Virginia Beach, VA and Prince William, VA are all remarkably low in both infection rates and mortality rates. Are they "behind the curve" and due to "catch up" or will their lucky state persist?



What's next? Can we predict well?

We don't know what's next. No, we can't predict well. General, wide-spread testing is required for any level of precision in our predictions (rather than the current rationed testing when severe symptoms present).

**Your Data Must Be Wrong!**

While I have cross-checked much of the data, I could very well have made mistakes leading to errors. Also, this analysis is obviously very basic and barely scratches the surface. I'd welcome the opportunity to share the Kaggle notebook that generated this & many more charts. (Kaggle.com is a free cloud based org that provides the computer resources that would allow you to ask your own questions). Moreover, you might bring unique insights to the problem. If you have a bit of time & inclination, it's a fascinating & timely time to study. Kids from 6-60+ can dig in and discover answers. Arise Citizen Scientists - Ask Questions!

Arise Citizen Scientist 2020: Ask Questions

by M.A.Tucker 20APR20

TODO: Intro?

Citizen Scientists have an important role to play in understanding our current COVID travails. There are hundreds of thousands of Citizen Scientists contributing worldwide. The following leverages their extensive knowledge.

**Motivation**

A favorite satirist and punster, Alexandra Petri (Twitter: @petridishes), told a story of how Newton, fleeing to the countryside during a plague, decided to use the time to work out the math leading to the gravitational mechanics we still use today. She encouraged us to do something different during this time, something equivalent to Newton's effort. I was inspired. I had just such a thought in early February as our plague picked up steam. Above the list items such as "paint bathroom ceiling" or "fix leaky faucet", I had placed "discover anti-gravity". It turns out that anti-gravity is a REALLY HARD PROBLEM. So as an alternative, perhaps slightly more achievable goal, I decided to try to better understand our COVID travails in general and specifically, better judge the veracity of often conflicting pronouncements.

**Method**

We are blessed in the USA to have ready access to a vast amount of data and powerful computer resources (much of it free). Data capturing various aspects of the plague are maintained and updated daily by the CDC, Kaggle.com, Johns Hopkins and many other organizations. However, huge streams of numbers purportedly describing something become rather mind numbing. My goal as a budding Citizen Scientist is to understand the data and capture the essence in some visualizations (e.g. graphs, tables). In short, convert "data" into "facts". How hard could that be?

Here is a link to a Johns Hopkins dashboard focusing on the same USA county wise data (<https://coronavirus.jhu.edu/us-map>). By looking at smaller sized areas, you may discern aspects not apparent if all of the US or world is considered.

**Questions**

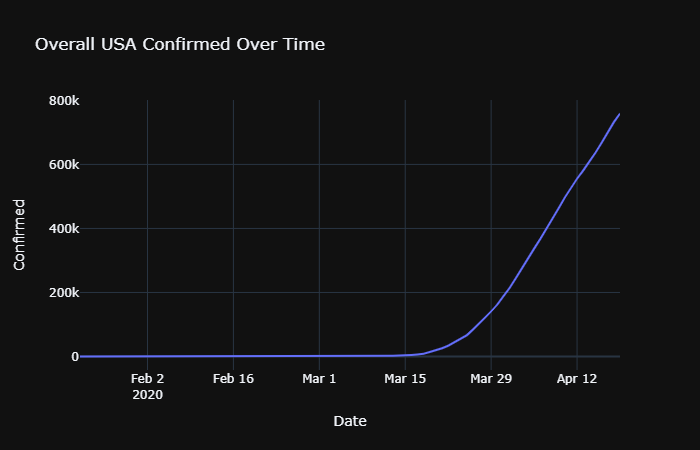
Here are some questions and some "answers" based on our best (but undoubtedly flawed) available data. The following is based on data gathered through April 20th, 2020.

Is it a hoax?

No, it's a modern day plague.

Is it almost over?

No, it's spreading rapidly with increasing numbers of confirmed cases & deaths.



Do various US county sized areas "look alike"?

All areas are experiencing increasing numbers of confirmed cases. The meaningful differences relate to infection rates (confirmed cases given the population size of the area) and mortality rates (deaths given the number of infections). Choosing a mix of mid-sized cities and suburban areas vs. a baseline, King County, WA, leads to more questions than answers.

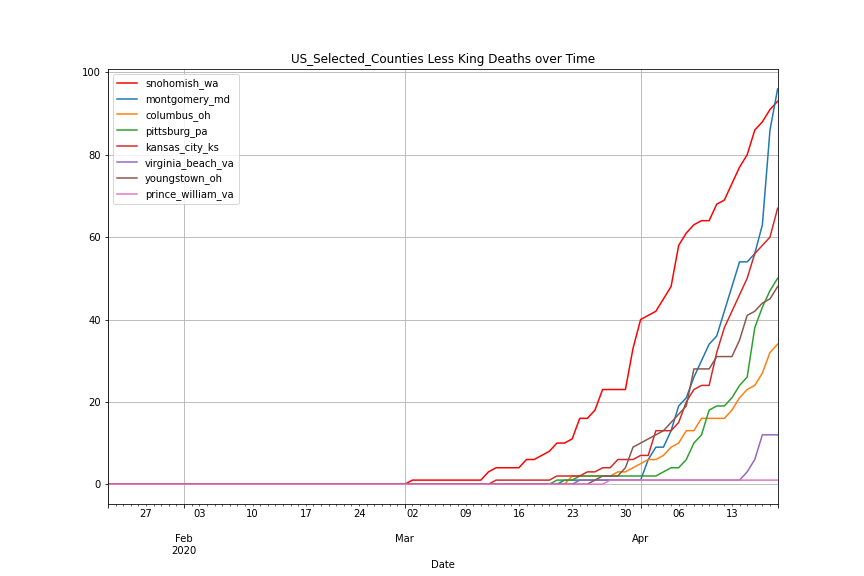
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **region** | **population** | **confirmed cases** | **infection rate** | **deaths** | **mortality rate** |
| King County, WA | 2,253,000 | 5174 | 0.23% | 346 | 6.69% |
| Kansas City, KS | 890,089 | 832 | 0.09% | 67 | 8.05% |
| Youngstown, OH | 228,683 | 589 | 0.26% | 48 | 8.15% |
|  |  |  |  |  |  |
| Montgomery County, MD | 1,051,000 | 2507 | 0.24% | 96 | 3.83% |
| Snohomish, WA | 822,000 | 2143 | 0.26% | 93 | 4.34% |
| Pittsburg, PA | 1,216,000 | 1035 | 0.09% | 50 | 4.83% |
|  |  |  |  |  |  |
| Columbus, OH | 1,683,751 | 1733 | 0.10% | 34 | 1.96% |
| Virginia Beach+, VA | 910,385 | 546 | 0.06% | 12 | 2.20% |
| Prince William, VA | 470,335 | 742 | 0.16% | 1 | 0.13% |

TODO: Expand?

King County, Kansas City and Youngstown, OH all have a very high mortality rate (while varying in population from 2.2 million to a couple hundred thousand). Why?

Montgomery County, MD, Snohomish, WA and Pittsburg, PA have similar mid-range mortality rates. Montgomery County, MD recently surged past Snohomish, WA. Since February, Snohomish, WA had been near the top of the leaderboard. Why is Montgomery County surging? It should be noted that this is not a contest you want to win.

Columbus, OH, greater Virginia Beach, VA and Prince William, VA (Haymarket and Manasas) are all remarkably low in both infection rates and mortality rates. Are they "behind the curve" and due to "catch up" or will their lucky state persist?



What's next? Can we predict well?

We don't know what's next. No, we can't predict well. General, wide-spread testing is required for any level of precision in our predictions (rather than the current rationed testing when severe symptoms present).

TODO: Wrap! engage & enlist?

**Your Data Must Be Wrong!**

This is obviously very basic and barely scratches the surface. However, you too could tackle this & maybe bring unique insights. If you have a bit of time & inclination, it's a fascinating & timely time to study. Kids from 6-60+ can play and learn. I'd be happy to share the Kaggle notebook that generated this & many more charts. Kaggle.com is a free cloud based org that would allow you to ask your own questions. Arise Citizen Scientists!



A useful first step for a citizen scientist tackling a problem is to understand and visualize the data she will work with to answer questions. US county-wise data (link!) has a line with two pieces of information for each county in the USA ("Confirmed Cases" and "Deaths") for every day since King County, WA recognized the plague in late January (January 23rd, 2020). Here is a link to a Johns Hopkins dashboard focusing on USA county wise data (<https://coronavirus.jhu.edu/us-map>). By looking at smaller sized areas, you may discern aspects not apparent if all of the US or world is considered. The following is based on examinations of the same USA county-wise data.

As of mid-April, this means there are ~250,000 lines of data - for too much to eye-ball effectively. So after finding the latest data, the initial task is to greatly reduce the shear volume by extracting only the data for areas of interest. These are the areas of interest I picked but you can easily extract areas more interesting to you:

* King County, WA
* Snohomish, WA
* Montgomery County, MD
* Columbus, OH
* Pittsburg, PA
* Kansas City, KS
* Virginia Beach/Norfolk/Chesapeake, VA
* Youngstown, OH
* Prince William, VA

These areas are selected for several reasons. King County is the first area impacted by the plague. It serves as a reliable "baseline". The other areas are mid-sized or suburban cities where I know, love and/or talk to people. I have at least anecdotal evidence of the timing various mitigation methods (or lack thereof).



Arise Citizen Scientist 2020: Know Your Data

These sections are intended to stand by themselves so please do skip to those that interest you:

* **Motivation** - what inspired this?
* **Method** - know your data!
* **Accuracy** - trust but verify
* **Questions** -
* **Findings**
* **Follow-on Questions**
* **Resources**
* **Broader Questions**
* **Easier Problems**
* **Acknowledgements**

**Motivation**

Reading a favorite satirist and punster, Alexandra Petri (Twitter: @petridishes), sometimes brings chuckles, often groans but a recent column near the end of March was nothing short of profoundly inspiring. She told a story of how Newton, fleeing to the countryside during a plague, decided to use the time to work out the math leading to gravitational mechanics. She encouraged us to do something different during this time, something equivalent to Newton's effort. Then, of course, she careened off into the weeds with groan-worthy puns, similes and analogies.

However, I was inspired. I had just such a thought in early February as our plague picked up steam. So above the list items such as "paint bathroom ceiling" or "fix leaky faucet", I placed "discover anti-gravity". It turns out that anti-gravity is a REALLY HARD PROBLEM. So as an alternative, perhaps slightly more achievable goal, I decided to try to become a Citizen Scientist. Here's the story so far.

The plague news is disturbingly short on science. The pronouncements, updates and prognostications are understandably filled with emotionally charged reporting. Stories about people facing fear, heartbreak, loneliness, destitution, powerlessness. All reminders of the old adage: "There, but for the grace of God, go I…". Many experts offered dramatically differing opinions ("it's a hoax!", "millions may die!", "it's the flu", "it's a death sentence..."). But I want to see if we could use science to better understand patterns, judge credibility and perhaps even support the adoption of new behaviors.

~~So I adopted a maxim from the old arms control negotiations - "Trust but verify".~~

We are blessed in the US to have ready access to a lot of data. Data capturing various aspects of the plague are maintained and updated daily by the CDC, Kaggle.com, Johns Hopkins, other orgs. However, huge streams of numbers purportedly describing something become rather mind numbing. My goal is to understand the data and capture the essence in some visualization (e.g. a graph). In short, convert "data" into "facts". How hard could that be?

I first decided looking at data from all the United States would likely "wash out" emerging trends and particularly, regions of interest. For example, confirmed cases in the US are ~600,000 while data from Seattle and New York with thousands of cases would mask trends in different regions of interest such as my area with "mere" 1000 cases.

**Method**

A useful first step for a citizen scientist tackling a problem is to understand and visualize the data she will work with to answer questions. US county-wise data (link!) has a line with two pieces of information for each county in the US - "Confirmed Cases" and "Deaths" for every day since King County, WA recognized the plague in late January (January 23rd, 2020).

As of mid-April, this means there are ~250,000 lines of data - for too much to eye-ball effectively. So after finding the latest data, the initial task is greatly reduce the shear volume by extracting only the data for areas of interest. These are the areas of interest but you can easily extract areas more interesting to you:

* King County, WA
* Snohomish, WA
* Montgomery County, MD
* Columbus, OH
* Pittsburg, PA
* Kansas City, KS
* Virginia Beach/Norfolk/Chesapeake, VA
* Youngstown, OH
* Prince William, VA

These areas are selected for several reasons. King County is the first area impacted by the plague. It serves as a reliable "baseline". The other areas are mid-sized or suburban cities where I know, love and/or talk to people. I have at least anecdotal evidence of the timing various mitigation methods (or lack thereof). Questions are asked for each of these areas.

**Accuracy**

Wherever possible, the findings are crossed checked with credible sources. Johns Hopkins (<https://coronavirus.jhu.edu/us-map>) just launched a site showing US county wise data that serves as a nice cross-check source. However, there are distinct possibilities that the findings are WRONG. The data may be incomplete or inaccurate. For example, with limited testing available, deaths may be mis-attributed to pneumonia or other respiratory ailments. Also (unavoidably), it's quite easy for a typo in the code to produce believable but utterly inaccurate results.

Moreover, generalizations will be made that may gloss over nuances. It is useful to compare & contrast various aspects but the comparisons may be mis-leading. The old arms negotiator slogan comes to mind: "Trust but verify!".

**Questions**

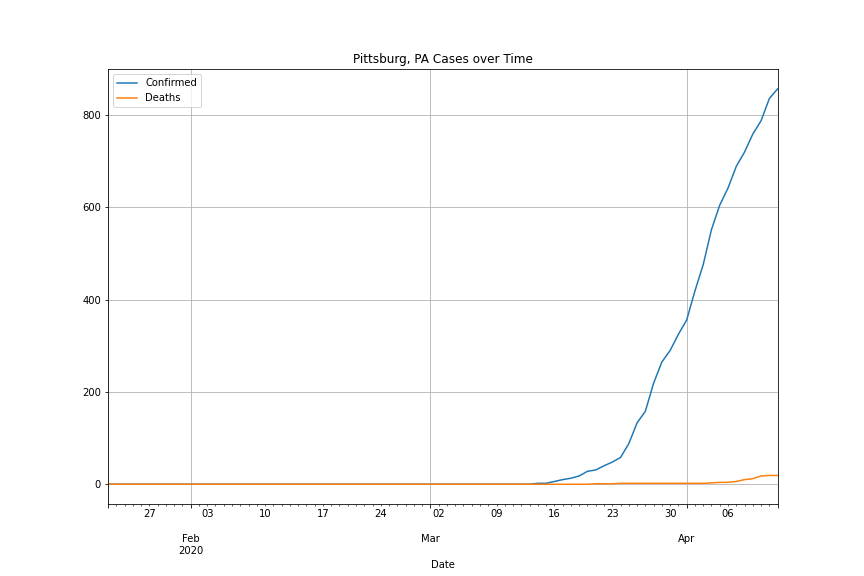
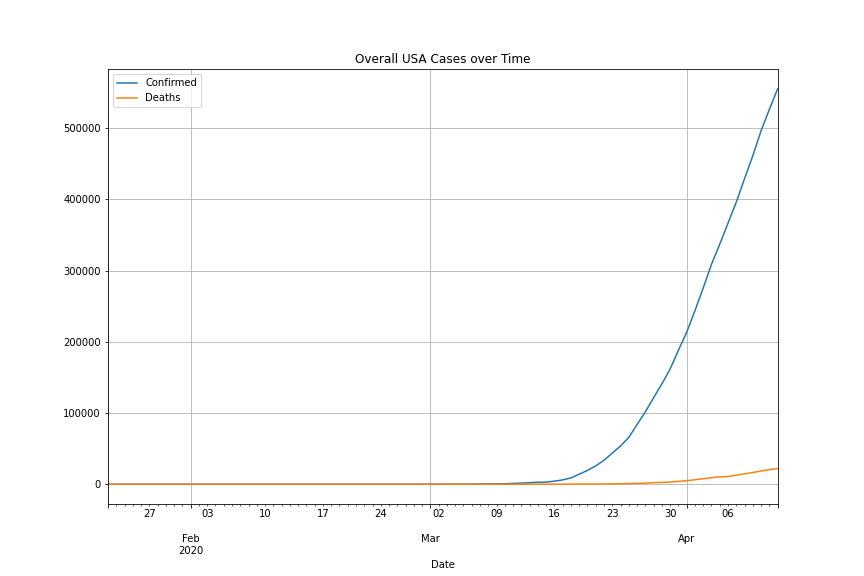
For each area:

1. Do confirmed cases and deaths "look" similar for each area?
2. Does population affect the magnitude of confirmed cases and deaths?

**Findings**

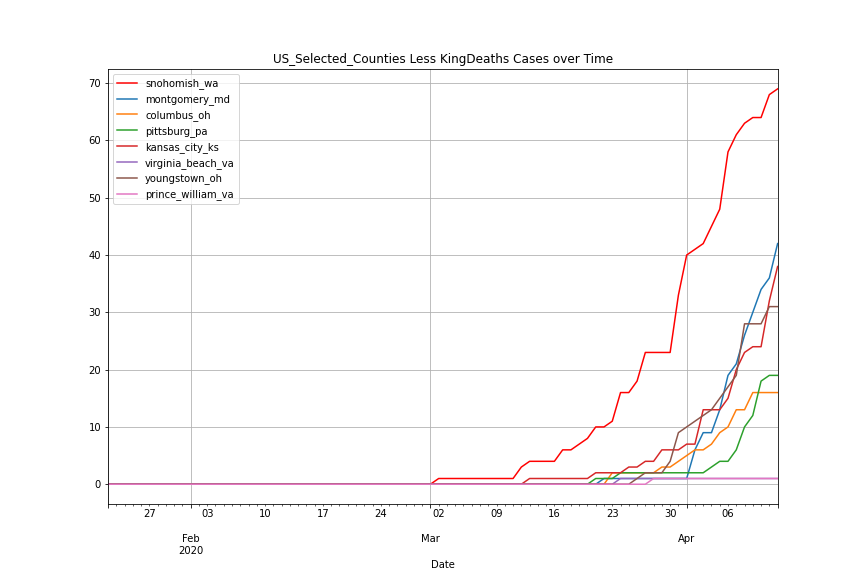
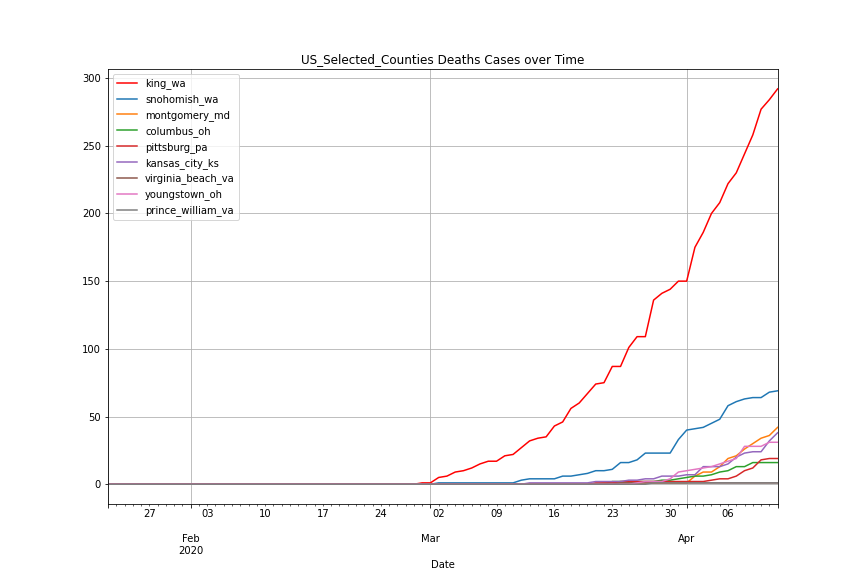
1. Do confirmed cases and deaths "look" similar for each area?

On the surface, all areas look remarkably similar. Confirmed cases rise dramatically (the classic "hockey stick") followed but a rise in deaths at a smallish percentage of the confirmed cases.



Overall US Pittsburg, PA

Comparing "Confirmed Cases" for all the areas together illustrates the similarities. It also highlights the early start in King County,WA and the comparatively large number of cases. Removing King County from the mix, shows the extent of the cases in Snohomish County, WA (King County's next door neighbor).



Confirmed Cases (All Areas) Confirmed Cases (less King)

2. Does population affect the magnitude of confirmed cases and deaths?

While the areas look similar at a high level, the population of the area can be factored into the data. If you have X cases, the dividing X by the total population supplies a rate rather than the raw count. After factoring in the population, the areas of interest look different.

Kansas City, KS and Youngstown, OH "look similar" to King County: high mortality rates (~5-7%) vs the mortality rates in other areas (~0-3%).

Snohomish, WA and Montgomery County, MD have similar infection and mortality rates. Watching this evolve over time, Montgomery County is "catching up" with Snohomish rapidly.

Pittsburg, PA and Columbus, OH are very similar with low infection and mortality rates.

Finally, Prince William, VA (including Haymarket and Manasas, VA) and the greater Virginia Beach area are also very similar with very low infection and mortality rates (1 death each).

As of April 12th, the details of the infection and mortality rates are summarized below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **region** | **population** | **confirmed cases** | **infection rate** | **deaths** | **mortality rate** |
| **King County, WA** | **2,253,000** | **4426** | **0.20%** | **292** | **6.60%** |
| **Youngstown, OH** | **228,683** | **417** | **0.18%** | **31** | **7.43%** |
| **Kansas City, KS** | **890,089** | **675** | **0.08%** | **38** | **5.63%** |
|  |  |  |  |  |  |
| **Snohomish, WA** | **822,000** | **1867** | **0.23%** | **69** | **3.70%** |
| **Montgomery**  **County, MD** | **1,051,000** | **1631** | **0.16%** | **42** | **2.58%** |
|  |  |  |  |  |  |
| **Pittsburg, PA** | **1,216,000** | **857** | **0.07%** | **19** | **2.22%** |
| **Columbus, OH** | **1,683,751** | **1118** | **0.07%** | **16** | **1.43%** |
|  |  |  |  |  |  |
| **Virginia Beach+,VA** | **910,385** | **454** | **0.05%** | **1** | **0.22%** |
| **Prince William, VA** | **470,335** | **389** | **0.08%** | **1** | **0.26%** |

**Follow-on Questions**

* Forecasting: Are the lucky "laggards" just behind the curve and will they "catch up" as the plague works through the population?
* Are "Social Distancing", business closures, etc. "flattening the curve"? If so, how many days until the impact is seen in an area?
* What is "flattening the curve" and what does it look like?

**Resources**

**Broader Questions**

* What should we do the next time we see a virus outbreak?
* Is a "coin-operated" healthcare system the best model for preventing plagues?

**Easier Problems**

**Acknowledgements**

Acknowledgements:

Alexandra Petri Twitter: @petridishes

CDC

Kaggle.com

Johns Hopkins CSSE