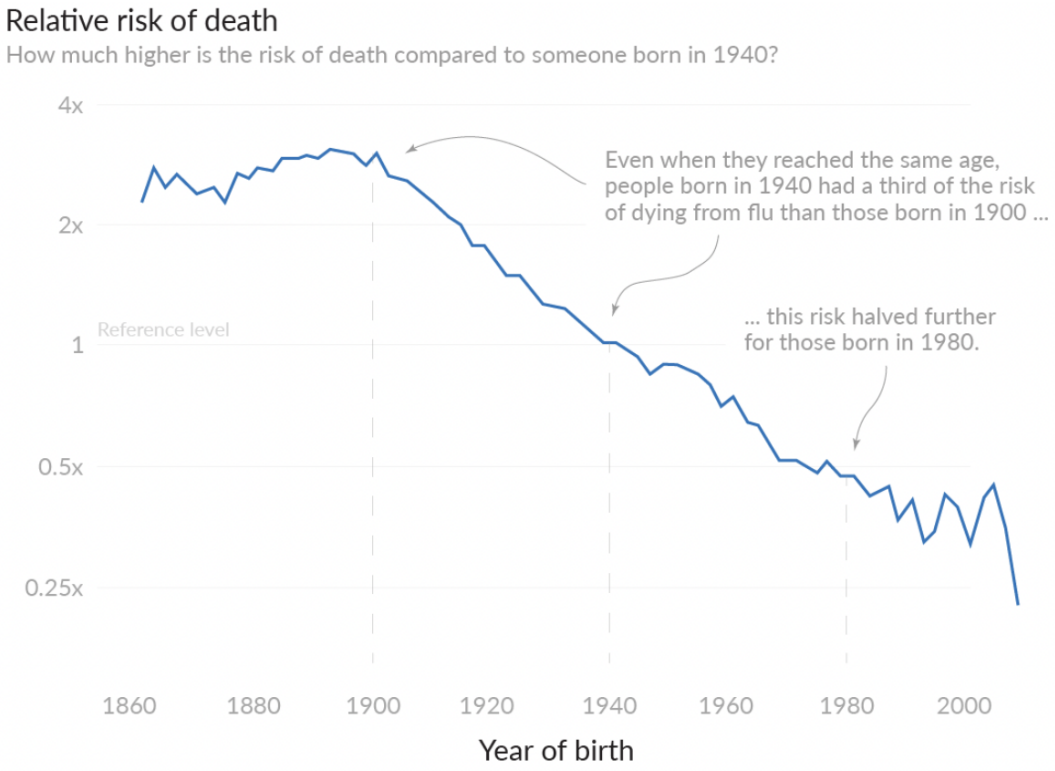
CAPP 30239: Project 2-Analyzing Existing Data Visualizations

Wesley Janson

October 24, 2022

1. **Visualization 1:** Relative risk of death from influenza



(a) Title: Relative risk of death. This visualization was created by Saloni Dattani and Fiona Spooner, researchers studying influenza deaths, and can be found at Our World in Data. They use data from Acosta et al's (2019) intrinsic estimator model using historical U.S. mortality data from their publication in *Demography*.

(b) The purpose of this data visualization is to display the relative decrease in mortality risk from influenza in the United States in the past ≈ 150 years, and how rapidly it has declined (however not monotonically).

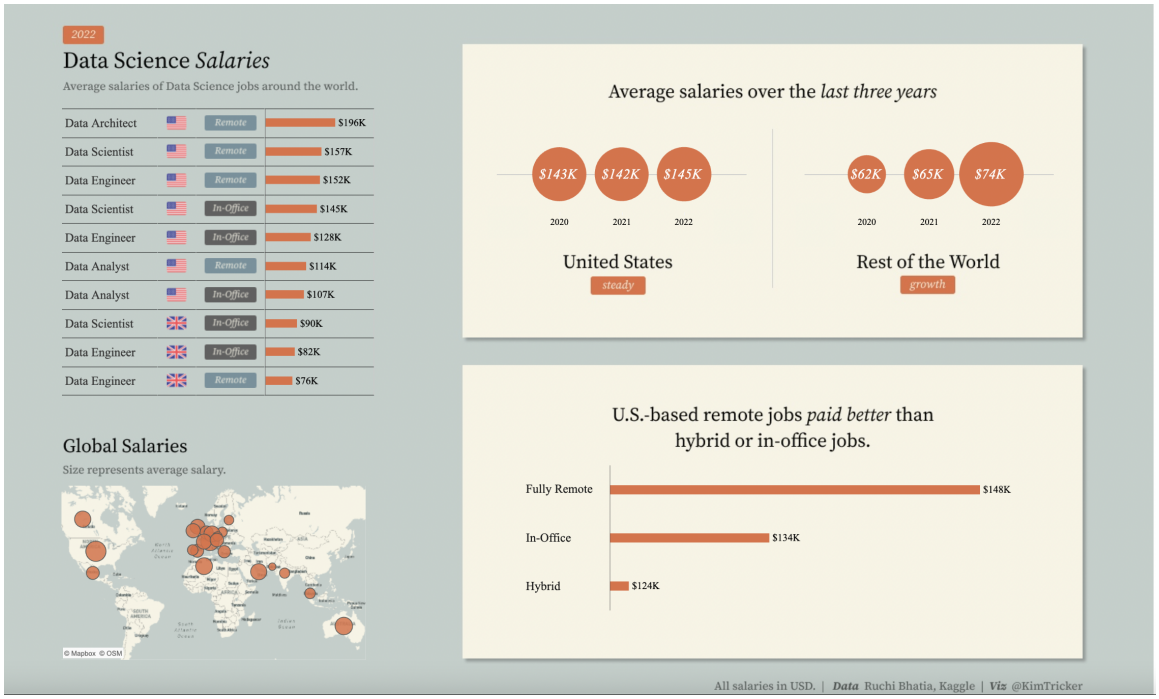
(c) This visualization is composed by creating using the underlying data from the model from Acosta et al (2019). This data is meant to show the mortality rates from influenza for individuals in the United States based on their birth year. Graphically, it is a simple line graph, where the main benefit is being able to see the trend (and levels) over time.

(d) Yes, I think this visualization does a great job of having a clear message, and that message is that over the past $\approx$ 150 years, the relative risk of dying of influenza (conditional on age) has drastically fallen, with those being born in 2000 having about 1/6 the risk of dying from influenza as those born in 1900. I imagine the intended audience is pretty broad, it could be the greater public, health care experts, or policymakers as all groups stand to gain from this evidence.

(e) Yes, this visualization is quite effective, as it displays a clear message, and is easy to follow. One does not need a technical background to get the message of this chart.

(f) There is one very stark change that could be made. The y-axis is manipulated such that the distance between each tick mark is not equal. If they were to expand it to be equidistant between each tick mark, it would visually look like an even more stark decrease in relative risk of death from influenza!

2. **Visualization 2:** 2022 Data Science Salaries



(a) Title: Data Science Salaries. This visualization was created by Kim Tricker, and can be found at Tableau Viz of the Day. Tricker uses data from Ruchi Batia at Kaggle.

(b) The purpose of this data visualization is to be informational, and provide the viewer with comparative information about data science salaries conditional on location, time dependency, title, and work model (in-person, hybrid, work-from-home).

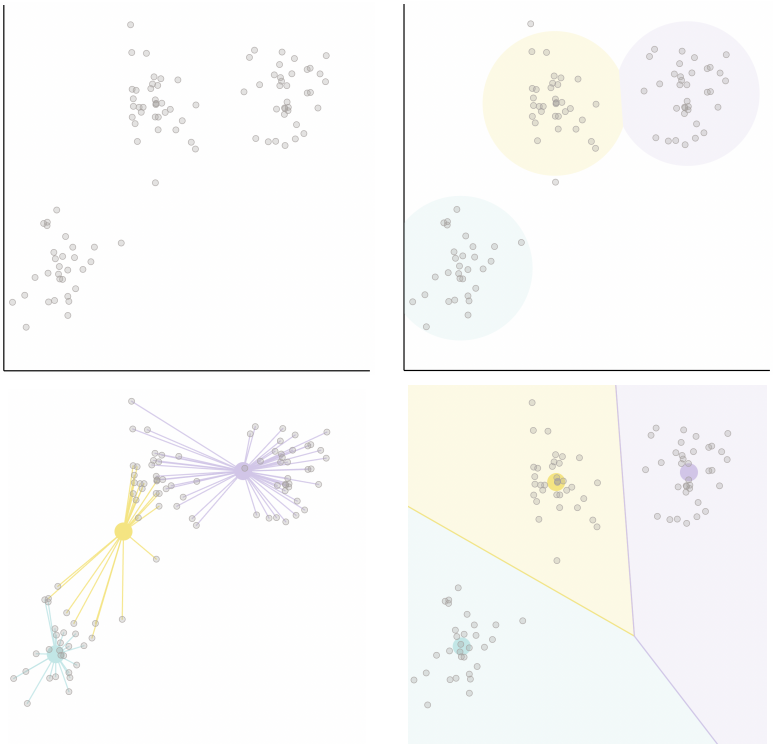
(c) This data visualization is created by panelling different graphs/visualizations. It uses 4 panels specifically, going clockwise they are: a sparkchart of some sort looking at the highest salaried positions conditional on work model and country; an area chart analyzing the trend in average salaries over the past three years; a horizontal bar chart of U.S.-based data science jobs conditional on work model; map analysis of average data science job salaries across the globe.

(d) It is hard to say if there is a clear message, as this visualization is meant to be informational. The intended audience could be those looking for data science jobs (like my self haha), or anybody interested in the conditional breakdown analysis of data science positions.

(e) Yes, I think this data visualization is very effective! It contains quite a bit of information, blending comparative data by categories and over time. They are able to convey a lot of information in an engaging and very digestible way for anybody.

(f) I think the one thing I noticed was the horizontal bar charts in the bottom right quadrant are a bit deceiving, with the x-axis not starting at 0. It makes it look like the "Fully Remote" positions are many, many times more lucrative than the "In-Office" and "Hybrid" work options while they are only 10% and 20% more high paying, respectively. This would strengthen the message in the sense that it would be a more accurate depiction of the information, as this is an informational data visualization.

3. **Visualization 3:** K-Means Clustering, An Explorable Explainer



(a) Title: K-Means Clustering, An Explorable Explainer. This visualization was created by

Yi Zhe Ang, and can be found via FlowingData. Ang uses generated data that is meant to display three distinct species of cats.

(b) The purpose of this set of visualizations is to illustrate what the k-means clustering algorithm does behind the scenes. While the algorithm can be relatively intuitive to explain, visualization definitely helps break down the steps that are taken. Furthermore, in reality, the data is seldom as separated into distinct groups as this example, but having the easily partitioned data helps even a novice viewer understand the concept.

(c) This page is quite engaging in the sense that it is one where the graph's next iteration is displayed depending on the user scrolling down the page. This is a nice design that allows the graph to be directly next to the text that is describing it. Different iterations of the same initial scatterplot (3 of which are seen above) are used to convey the steps in the k-means clustering algorithm.

(d) In a way, yes, this visualization has a clear message. The message is to convey the steps that the k-means clustering algorithm takes in order to get results. The intended audience would be any student/data scientist etc. that is trying to learn about the algorithm.

(e) Yes, I think this visualization is incredibly effective. It is a bit more interactive than the last two examples I've went over, but is simple while also being a great learning tool. This is a great example of where a more elaborate, interactive graph can help supplement learning.

(f) It is hard for me to think of a way to recommend improving this data visualization, as I think it is a great, informative, and all-encompassing visual. Perhaps the creator could have added some text to the graphs, which could have further hammered home the concepts they were trying to display. Also, since this data is generated, maybe adding a real-world example would help illustrate how the algorithm sometimes is less straightforward, and the visualizations could show that.