Data.gov Salary Data Instructional User Guide

May 3, 2024

Team Generic Open Dataset Project Charter #2

Herschel Combs, Kaden Harvey, Rachel Kwan, Khushboo Rathore and Victoria Stavish

Table of Contents

Introduction

Purpose of project

The purpose of this instructional user guide is to aid in the understanding and use of publicly available salary data from Data.gov. This user guide provides those who wish to access information from Data.gov with documentation on how to navigate and properly use the public data on the website. By making an accessible user guide to this data, we hope those interested in researching salary data, such as academics, economists, journalists, employees and/or employers as well as regular consumers, can understand what salaries look like in their industry or their geographic areas and empower them to use salary data to do their own analyses and draw their own conclusions.

Background

Data.gov is The United States Government's open data site and was originally made public in 2009 with the goal of publicizing data about the nation and improving transparency about the nation and its government, according to the site’s ‘About’ page. Data.gov is managed and hosted by the U.S. General Services Administration and is developed publicly on GitHub.

How the data is collected

Datasets on Data.gov are collected through the OPEN Government Data Act, which requires government data to be public and available in machine-readable formats while maintaining security and confidentiality, according to the website. Data.gov typically works with pinpoint people at various public agencies and offers contact information to government agencies in order to work collaboratively with consumers of the data to get more usable data on the website.

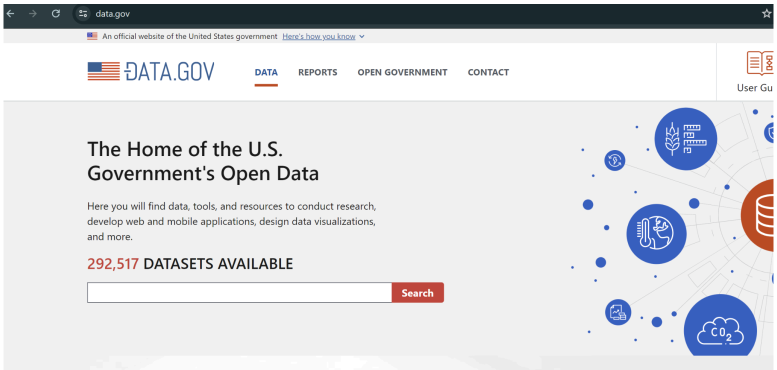
The salary data we collected

To obtain the salary data on Data.gov, we compiled every dataset that the site's data catalog returned when "Salaries" was typed into the search bar. This resulted in 180 data results. Some of these results contain no public datasets. Some of these results contain multiple public datasets in various electronic formats.

Using Data.gov filters

How we found our salary dataset

In order to find the raw salary data we used and analyzed, first type Data.gov into your search bar. You should find yourself on the home page, which is shown below. At the top of the home page is a navigation bar. The leftmost option on the navigation bar should say "Data". Click "Data".

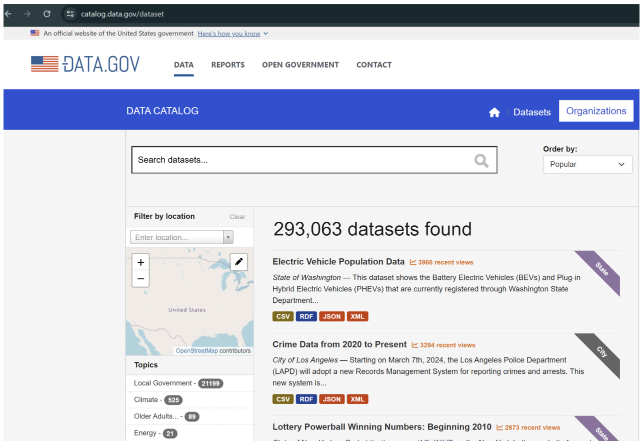


The Home of the U.S. Government's Open Data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

292,517 DATASETS AVAILABLE

Selecting "Data" will take you to the "Data Catalog" page, shown below, which includes nearly 300,000 datasets.



293,063 datasets found

Electric Vehicle Population Data 3966 recent views

State of Washington — This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department...

CSV RDF JSON XML

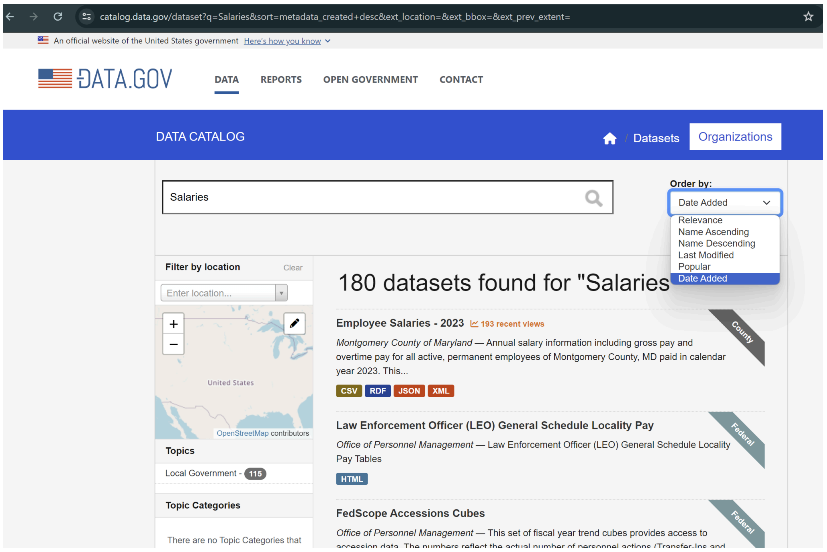
Crime Data from 2020 to Present 3294 recent views

City of Los Angeles — Starting on March 7th, 2024, the Los Angeles Police Department (LAPD) will adopt a new Records Management System for reporting crimes and arrests. This new system is...

CSV RDF JSON XML

Lottery Powerball Winning Numbers: Beginning 2010 2873 recent views

Once you have navigated to the data catalog, type “Salaries” into the search bar and click the magnifying glass to search. As of April 25, 2024, 180 datasets are returned for this search. As part of this project, we also sorted the 180 datasets to display by “Date Added”. You can adjust the order in which the datasets are displayed by toggling the drop-down menu on the right-hand side of the screen titled “Order by” and selecting your preferred ordering. This is demonstrated in the image below.



catalog.data.gov/dataset?q=Salaries&sort=metadata_created+desc&ext_location=&ext_bbox=&ext_prev_extent=

An official website of the United States government

Salaries

Order by: Date Added

180 datasets found for "Salaries

Employee Salaries - 2023 193 recent views

Montgomery County of Maryland — Annual salary information including gross pay and overtime pay for all active, permanent employees of Montgomery County, MD paid in calendar year 2023. This...

Law Enforcement Officer (LEO) General Schedule Locality Pay

Office of Personnel Management — Law Enforcement Officer (LEO) General Schedule Locality Pay Tables

FedScope Accessions Cubes

Office of Personnel Management — This set of fiscal year trend cubes provides access to accession data. The numbers reflect the actual number of personnel actions (Transfer-Ins and

Local Government - 115

There are no Topic Categories that

Using other available filters on Data.gov

While you can search almost any term in a similar way that we searched “Salaries” for this project, there are many other filters available to users on Data.gov to narrow down topics, types of data, types of files, data locations, publishers, bureaus and organizations that the data is from or describes. For example, a user can filter the datasets to only see data that is available in CSV format, that is only representative of the state of Illinois and/or has the tag “payroll”. Using these filters, you can parse through all of the datasets to find exactly what you want. This is not an exhaustive description of how to use the filters on Data.gov, but will give users a good starting point for understanding how the website works and its capabilities and limitations.

Working with the salary data

Dataset file types

Nearly all of the salary or salary-related data on Data.gov is downloadable in a structured, electronic format. There are just 10 datasets within all of our data that are not publicly available in a structured, electronic format.

The most common electronic format throughout the datasets is CSV, and most of our data — over 90 datasets — are available in multiple electronic format options. Below is a list of all the available file types we encountered in this data and a description of the capabilities and limitations of said file type, organized by the frequency with which we saw this file type occur.

CSV (160)

- Comma Separated Value files are text files in which information is separated by commas and are most common in spreadsheets and databases.
- This simple file type is able to be opened and manipulated in many programs, including but not limited to Excel, Google Sheets, Python, C, R, SQL, and even text editors.

JSON (88)

- A JavaScript Object Notation file is a text format for storing and transporting data.
- While created to work specifically in JavaScript, JSON files are readable and usable in many programs, including but not limited to Excel, Google Sheets, Python, C, R, SQL, and even text editors.

XML (85)

- An Extensible Markup Language file is used to store data in hierarchical elements, much like an HTML file.

- An XML file doesn’t do anything besides carry the data from one location to another. It stores information by wrapping the information in tags.

RDF (84)

- A Resource Description Framework file is primarily a data file used to store metadata.
- This file can be opened in most text editors and R.

HTML (33)

- Hypertext Markup Language files are primarily for creating and structuring documents originally intended for web browsers.
- This type of file typically only contains text data.
- In this dataset, an HTML file will usually take you to a landing page for the data rather than the actual dataset itself.

TEXT (19)

- A text file is a file containing lines of text.
- In this dataset, a text file will almost always lead you to a landing page for the data rather than the data itself.

API Endpoint (14)

- An Application Programming Interface Endpoint is a URL that acts as a specific location and point of contact within the API that accepts requests about the data from a client and sends back responses and the point of contact between.
- API endpoints are most accessible when coding in Python, but are also useful in R, SQL, C, Javascript, and most other programming languages.

NA (10)

- Some datasets are not publicly available and therefore have no electronic file formats available for download. These datasets are indicated by file format “na”.

GeoJSON (4)

- A GeoJSON file is a file format for encoding geographical data structures such as polygons, lines, and points through JavaScript Object Notation.
- This file type is able to be opened and manipulated in many programs, including but not limited to Python, C, R, SQL, and more.

XLSX/XLS (3)

- These files can be opened in Microsoft Excel and often contain data, most commonly numbers, text, and formulas.
- This simple file type is able to be opened and manipulated in many programs, including but not limited to Excel, Python, C, R, and SQL.

Shapefile (1)

- A shapefile is a format for storing geometric location and attribute information about features such as polygons, lines, and points.
- Working with shapefiles is often best in ArcGIS but is also functional in other programming languages such as Python, R, SQL, and more.

KML (1)

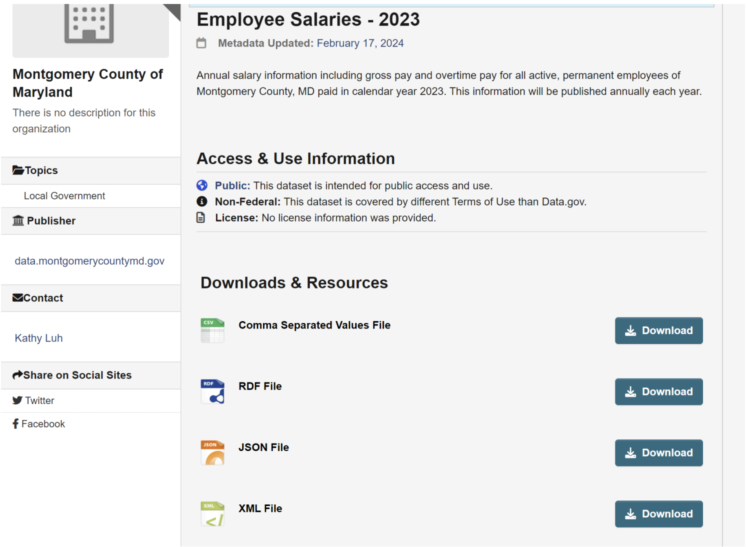
- A keyhole markup language file displays geographic data as well as additional annotations such as overlays, links, polygons, lines, points, and images.
- KML files are easiest to use with Google Earth, but it is also functional using ArcGIS. You can also convert KML files to shapefiles, GeoJSON files, and CSVs.

Downloading and exporting data

To download any of the datasets within this salary data, you will have to download the datasets one by one. There is no available option to download all of the datasets at once.

Click on the dataset you are interested in. You will see one or more downloadable file format options for most datasets, but some datasets won't have any downloadable file format options.

If the dataset you selected has downloadable file format options, decide which file format option best suits your needs and click the "Download" button on the right-hand side of the screen, as shown in the image below.



Montgomery County of Maryland

There is no description for this organization

Topics

Local Government

Publisher

data.montgomerycountymd.gov

Contact

Kathy Luh

Share on Social Sites

Twitter

Facebook

Employee Salaries - 2023

Metadata Updated: February 17, 2024

Annual salary information including gross pay and overtime pay for all active, permanent employees of Montgomery County, MD paid in calendar year 2023. This information will be published annually each year.

Access & Use Information

Public: This dataset is intended for public access and use.

Non-Federal: This dataset is covered by different Terms of Use than Data.gov.

License: No license information was provided.

Downloads & Resources

Comma Separated Values File — Download

RDF File — Download

JSON File — Download

XML File — Download

File conversion

If you are interested in a dataset but it isn't available in the electronic file format you need, there are a few file conversion options available to adjust the file types.

Excel

- Excel has a built-in conversion tool that supports the conversion of HTML, XML, and text files into XLSX or XLS files.

- To achieve this, open an Excel file, select the “Data” tab, select ‘Get and Transform Data’ and select the file type you are using.

CloudConvert

- CloudConvert is a file conversion website that can convert over 200 various file formats without fees. This specific file conversion website is capable of converting all of the file types within this dataset.

Salary dataset metadata

What is the metadata

The metadata log we created gathers significant information about all 180 datasets that we discovered. We also created a metadata data dictionary which defines and explains each column of the metadata log.

Possibilities of the metadata

The metadata of the salary and salary-related datasets available on Data.gov offers a lot of insight into the types of salary data available to the public and the scope of the salary data available on Data.gov. For example, the metadata shows us that the salary data on Data.gov largely consists of federal datasets, meaning much of the salary data you’ll find on the site are from federal agencies rather than state, county or private agencies.

Additionally, the metadata shows us that Data.gov is not a reliable source of salary data that is representative of the nation as a whole. The metadata shows that just 14 states have salary data on Data.gov and most of those states have fewer than 8 datasets to represent the state and the cities and countries within that state.

The metadata also shows us that much of the data on Data.gov does not have any publicly available documentation. This means that even if you get your hands on the data, unless you can contact someone who is an expert on this specific dataset, there is little confirmation that column names or data classifications mean exactly what you interpret them to mean.

The metadata helps us understand what kind of data we have and can also give us insights into what we don't yet know or understand about the data, which is equally important as knowing what we have.

Limitations of the metadata

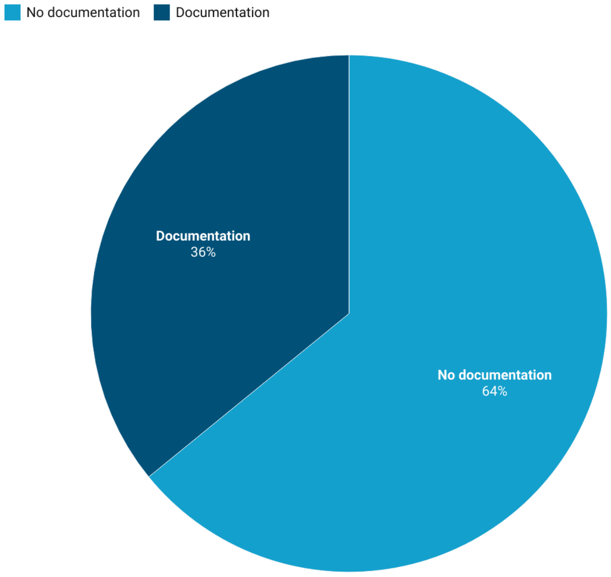
The metadata of the salaries dataset from Data.gov also leaves much to be desired. There is a lot of metadata that would be useful to have but doesn't exist or exists only for some datasets but is not consistent across all of the datasets. Some examples of this are the first published date and the last updated date columns of our metadata log. More than 40 of our salary datasets from Data.gov do not have a first published date and 32 of our salary datasets do not have a last updated date. Nearly 20 datasets also do not have any metadata on the dates the data represents, meaning that in some cases we know that salary data is recorded by calendar year, but we don't know which years are represented in the data.

While these limitations of the metadata severely impact our abilities to complete analysis of or about the data, these holes in the metadata can also help us understand what standards people, organizations, and agencies are/aren't prioritizing in publishing and documenting salary data.

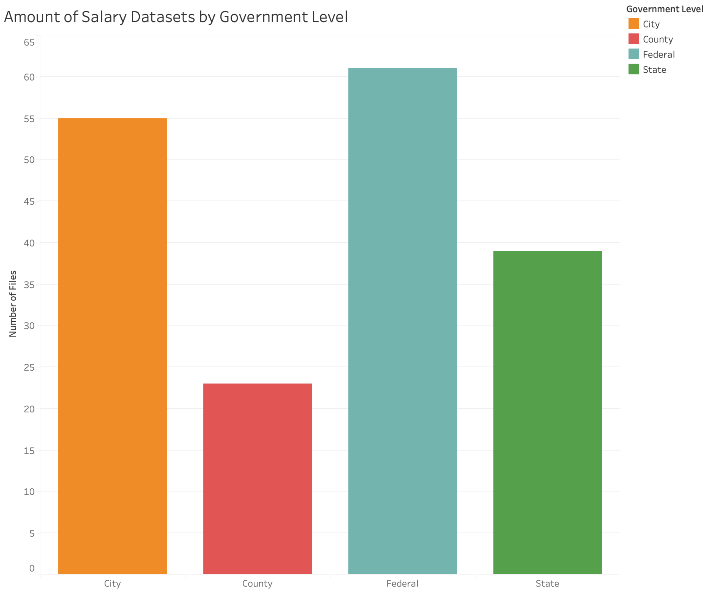
Metadata Visualizations

Most public salary datasets do not have data dictionaries or documentation of the data

Just 61 datasets have data dictionaries or documentation

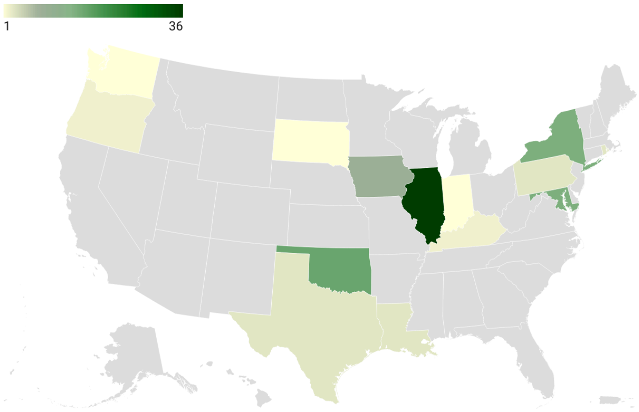


Created with Datawrapper



Just 14 states have salary data on Data.gov

On average, the states with salary data on Data.Gov have less than eight datasets available

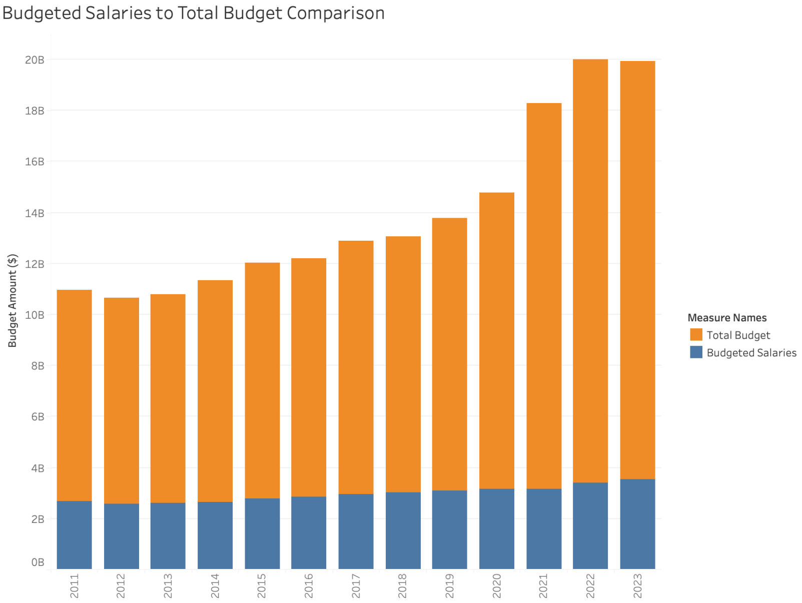


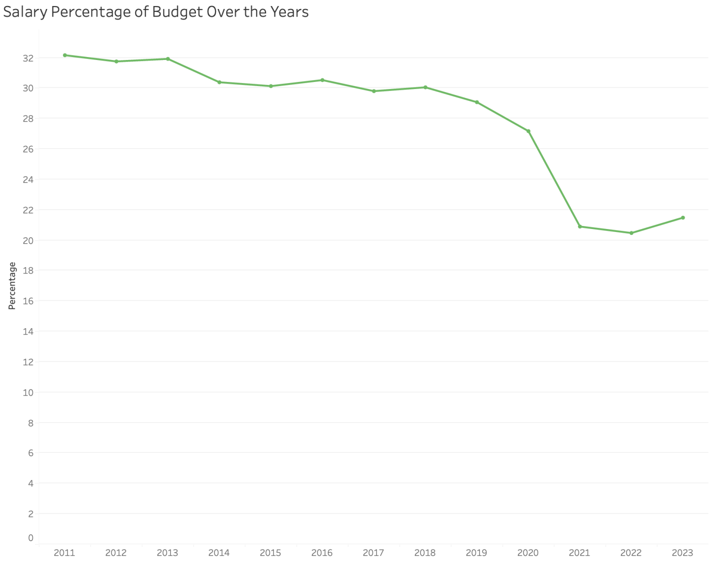
Created with Datawrapper

Salary data project potential

City of Chicago

One geographic area that has many years of consistent, cohesive, and comparable salary data is the city of Chicago. Of the 36 datasets that contain salary data from Illinois, 28 are from the city of Chicago or the Chicago data portal.

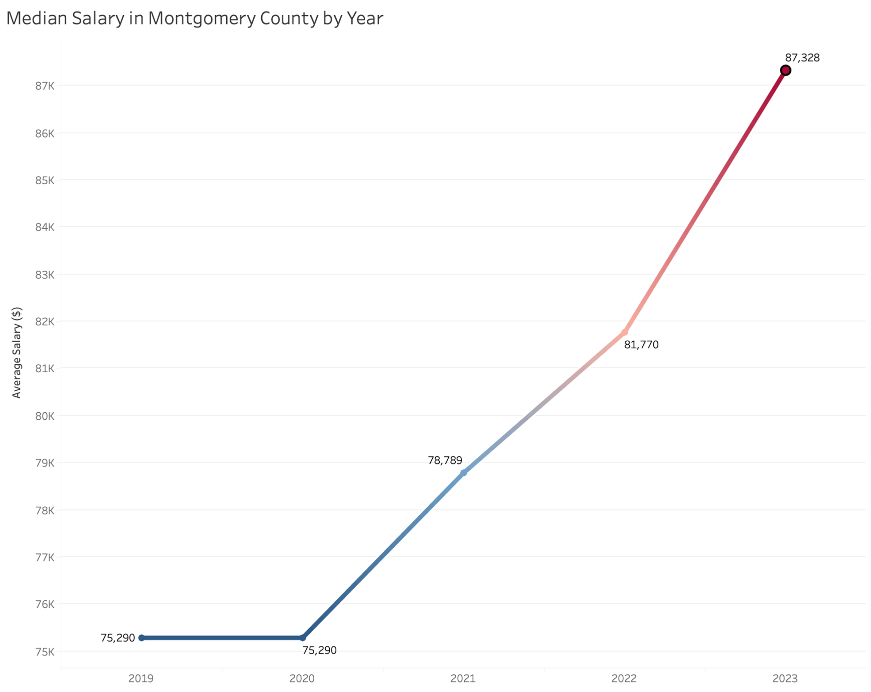




There are budget ordinance and recommendations datasets that outline the government positions and salaries in the City of Chicago each fiscal year from 2011 - 2023. Using this data, we can see how average salaries as well as salaries for specific positions have changed throughout the past decade. This analysis can offer a glimpse into how Chicago government employee salaries compare to the surrounding area and how they have grown and/or shifted over the past decade.

Montgomery County, Maryland

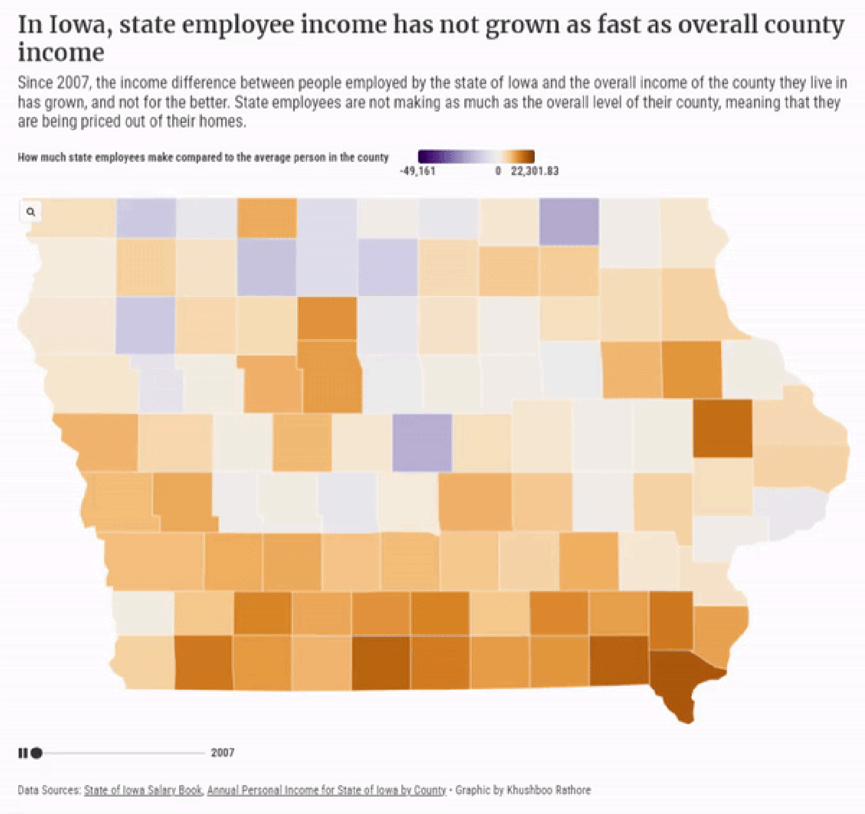
Multiple Montgomery County, Maryland datasets ranging from 2019 - 2023 provide individual data for employees in the county, including their base salary and the department they work in. There are about 10,000 employees within each dataset, so by using the median of the salaries for each year, a comparison can be made about the changes over the years.



We can see the drastic change from 2020 - 2023 as the color of the line changes. With this information, we can compare the salary medians with the economic factors that may have caused them for each year.

State of Iowa

The salary data contains multiple documents from the state of Iowa that document salary and income information. One of the possible comparisons is looking at state employee data and comparing it to overall income in each county. This can be accomplished using data analysis software and filtering the variables down, then getting a median value of the state employee salaries by year and joining it with the county income per capita.



In Iowa, state employee income has not grown as fast as overall county income

Since 2007, the income difference between people employed by the state of Iowa and the overall income of the county they live in has grown, and not for the better. State employees are not making as much as the overall level of their county, meaning that they are being priced out of their homes.

How much state employees make compared to the average person in the county
-49,161
0
22,301.83
2007

Data Sources: State of Iowa Salary Book, Annual Personal Income for State of Iowa by County · Graphic by Khushboo Rathore

This is at the most basic level and the data can be further cleaned in order to do more with it and get more accurate data. This data also contains a time variable in the year that the data was collected. This allows for analysis of patterns in the salaries over time, along with across locations. One example graphic, made in Flourish and converted into a gif, is shown above and can also be found here.

Constraints and Limitations of this project

Not exhaustive

While data.gov is a central location for a lot of salary data, it does not include all publicly available salary data across the United States. The way that data.gov collects datasets is often through people from public government agencies reaching out to data.gov contacts themselves, not the other way around. Because of this, the datasets on data.gov have a severe volunteer bias with which entities are/aren't included in the data.

For example, The city of Minneapolis, Minnesota has published their government employee salary data each fiscal year since 2010. This data is useful and relevant to our topic, but is not accessible through Data.Gov despite being publicly available data. This is just one example of how the dataset is simply a snapshot into some salary data across the country, but certainly is not a comprehensive collection of available salary data across the United States.

The salary dataset also skews heavily towards salary data and does not include much salary data for private companies, average salaries across different industries or employment groups outside of government entities. This severely limits the conclusions or findings one can draw from this data, as it represents a limited view of salaries in a geographic area.

Inconsistent data

While we have standardized much of the metadata related to the datasets within the salary data filter of Data.Gov, much of this data is missing years, data dictionaries and more. Of the nearly 200 datasets that make up all of the data we have documented, there are just a handful of legitimate findings one could and should use for research, comparison and/or data analysis. The salary data on Data.Gov should be used as a starting point for understanding public salaries, but not as a sole data source on the topic.

FAQs

How can I use this data?

The salary data from Data.gov can be used for a variety of purposes. Academics, economists, journalists, and anyone curious about salaries can utilize this data to understand salary trends across different industries or geographical areas. For example, one can investigate changes in government salaries over time or compare salaries across various states or counties. This data can be crucial for conducting research, informing policy decisions, or gaining insight into salaries in a particular sector.

Can I combine datasets I find on Data.Gov to do analyses?

There are some datasets from Data.gov that can be joined and analyzed jointly to produce sound conclusions. However, much of the salary data on Data.gov is not standardized and does not work well together, so it is essential that users vet the data they are interested in analyzing to ensure their analyses will be sound if they combine datasets. For example, the city of Chicago has published a decade of consistent salary data that can be used to understand Chicago city government employee salaries over the years. This data, however, is recorded by fiscal year, so you would not want to compare the salary data from Chicago with any salary data that is recorded quarterly or by academic year, for example.

Some datasets in this collection don’t include salary data. Why?

Not all datasets labeled as "Salary" on Data.gov contain salary data due to how data is categorized and uploaded on the platform. Sometimes, datasets might indirectly relate to salaries or be associated with salary data but only provide metadata or reference information instead of direct salary figures. Additionally, some datasets are restricted and private, which could explain their lack of salary data.

How often is this data updated?

The update frequency of the data on Data.gov can vary depending on the source agency and the nature of the data itself. While some datasets, particularly those involving government salaries, may be updated annually or in accordance with fiscal reporting periods, others might be updated less frequently. It's important to check the metadata associated with each dataset for specific information on its last update and the expected frequency of future updates.

Version History

Version 1.0	March 15, 2024
Version 2.0	April 26, 2024
Version 3.0	May 7, 2024