Scary Data Stories
(& Tips for Warding off Data Vampires)

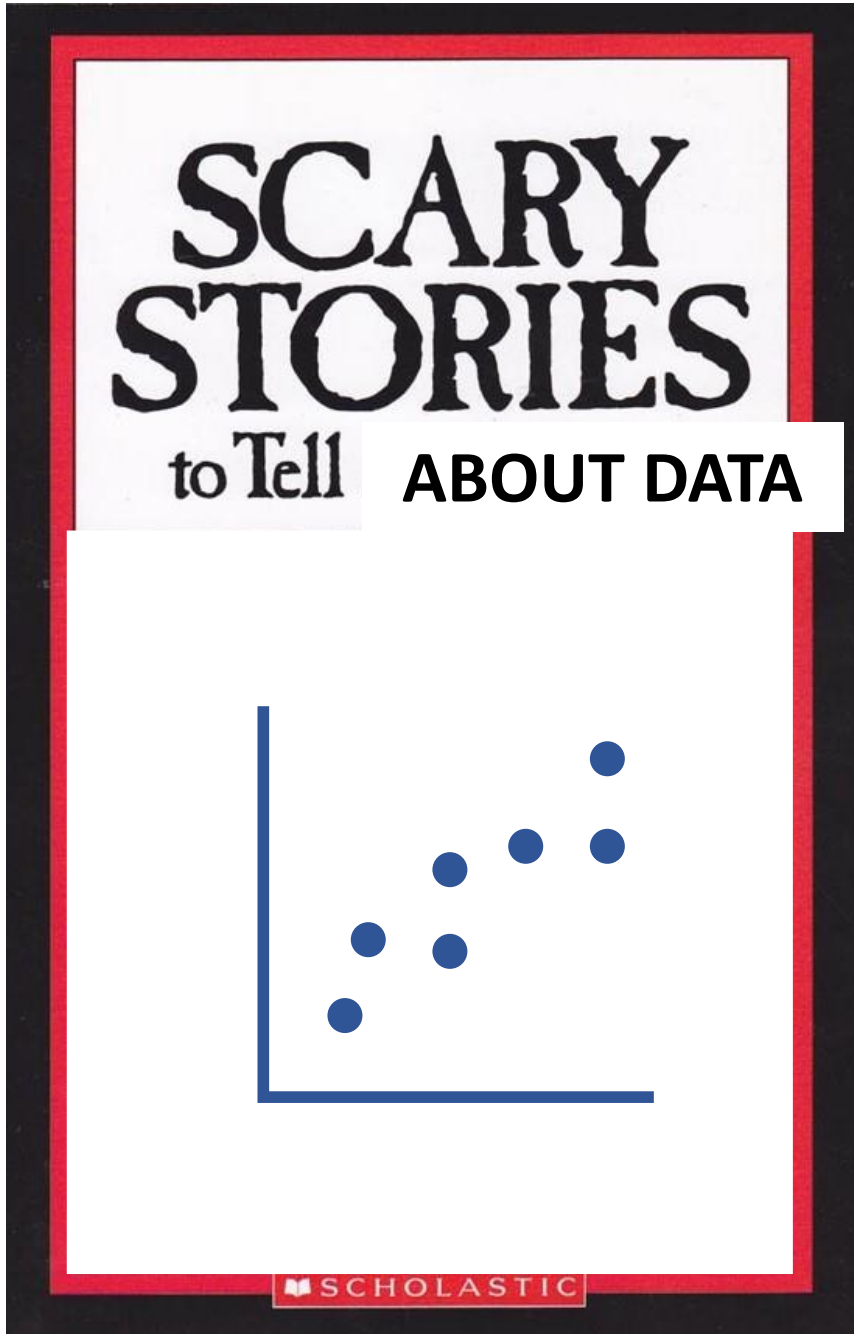
Ali Ruth | DAE-DLMAB

2024-10-31

Presentation note

- This presentation was originally prepared for a data science brown bag at the U.S. Department of State – Bureau of Global Health Security and Diplomacy, Presidential Emergency Program for AIDS relief (PEPFAR).
  - Commonly used data file format for analyses was **Excel**
  - Most common type of data was **aggregate clinical data from PEPFAR's global HIV treatment clinics**

- All data examples in this presentation are from public-use data sources.



1) Mysterious Missing Rows
2) Diabolical Date Formatting
3) Ghoulish Geolocations

Mysterious Missing Rows

Mysterious Missing Rows

- Public Health England (PHE) usually received Covid testing data from labs in a **.csv** file format

Mysterious Missing Rows

- Public Health England (PHE) usually received Covid testing data from labs in a **.csv** file format
- One week in October 2020, PHE imported testing data into pre-2007 Excel **.xls** files

Mysterious Missing Rows

- Public Health England (PHE) usually received Covid testing data from labs in a **.csv** file format
- One week in October 2020, PHE imported testing data into pre-2007 Excel **.xls** files
- Multiple rows per test result
- Excel has a row processing limit of:
  - 1,048,576 rows (.XLSX)
  - 65,000 rows (.XLS pre-2007)

Mysterious Missing Rows

- Public Health England (PHE) usually received Covid testing data from labs in a **.csv** file format
- One week in October 2020, PHE imported testing data into pre-2007 Excel **.xls** files
- Multiple rows per test result
- Excel has a row processing limit of:
  - 1,048,576 rows (.XLSX)
  - 65,000 rows (.XLS pre-2007)

→15,841 rows of case data disappeared from the document used for contact tracing!

Mysterious Missing Rows

Under-reported figures

From 25 Sept to 2 Oct

50,786

Cases initially reported by PHE

15,841

Unreported cases, missed due to IT error

---

8 days of incomplete data

1,980 cases per day, on average, were missed in that time

48 hours Ideal time limit for tracing contacts after positive test

Source: PHE and gov.uk

→15,841 rows of case data disappeared from the document used for contact tracing!

Tips to Address Mysterious Missing Rows

- Always know the specs of the **dataset**
  - Dimensions (rows x columns)
  - File size (GB; MB; KB)
- Always know the specs of the **software**
  - Dimension maximum
  - File size maximum
- Build **checks** into import/export scripted workflow to check for missing rows or columns

Tips to Address Mysterious Missing Rows

- Read and transfer very large datasets using scripted workflows for data wrangling and analysis (RStudio / posit, Python, Stata, SAS, etc.)
- If API is available: Also helpful to read in **only** rows / variables essential for an analysis

Tips to Address Mysterious Missing Rows

- For more spooky spreadsheet tales…

[European Spreadsheet Risks Interest Group – Spreadsheet Horror Stories Database]

Diabolical Date Formatting

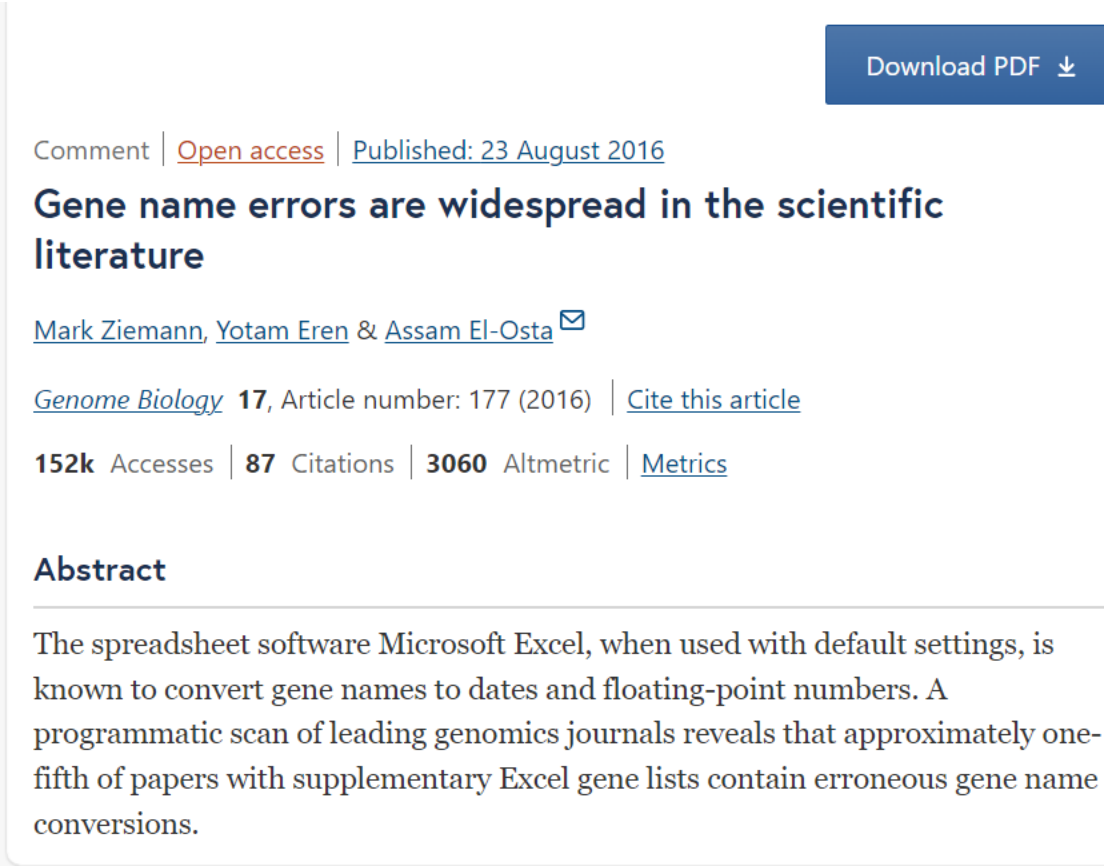
Diabolical Date Formatting

- Many gene names (SEPT2, MARCH1) have historically been text strings that autoconvert to dates in Excel

Gene name	Gene abbrev	Excel conversion
Septin 2	SEPT2	2-Sept
Membrane-Associated Ring Finger (C3HC4)1, E3 Ubiquitin Protein Ligase	MARCH1	1-Mar

Diabolical Date Formatting

- A 2016 study reviewed appendices for a sample of published articles in leading genome journals



Download PDF

Comment | Open access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

152k Accesses | **87** Citations | **3060** Altmetric | Metrics

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Diabolical Date Formatting

- A 2016 study reviewed appendices for a sample of published articles in leading genome journals
- **Approximately one-fifth of studies reviewed contained erroneous gene name conversions**

Download PDF

Comment | Open access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

Abstract

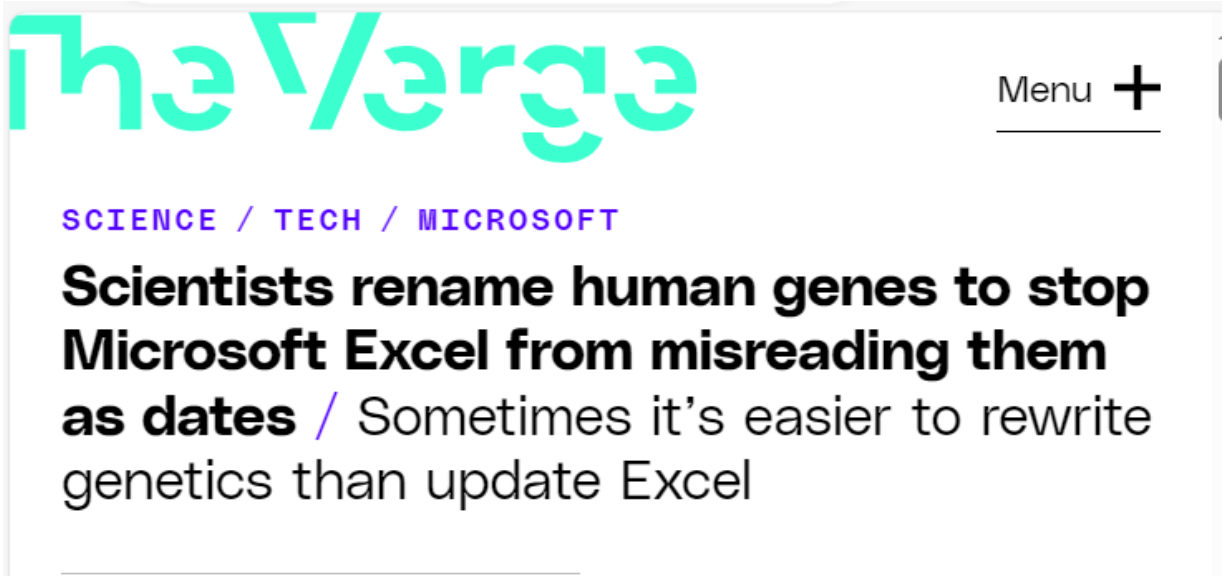
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Tips to Address Diabolical Date Formatting

- Solutions
  - Check and confirm variable types when importing data
- Best practice: utilize ISO 8601 for date formatting
  - Largest -> smallest time unit
  - “2023-10-31”
  - Facilitates easy sorting and reduces ambiguity

Tips to Address Diabolical Date Formatting

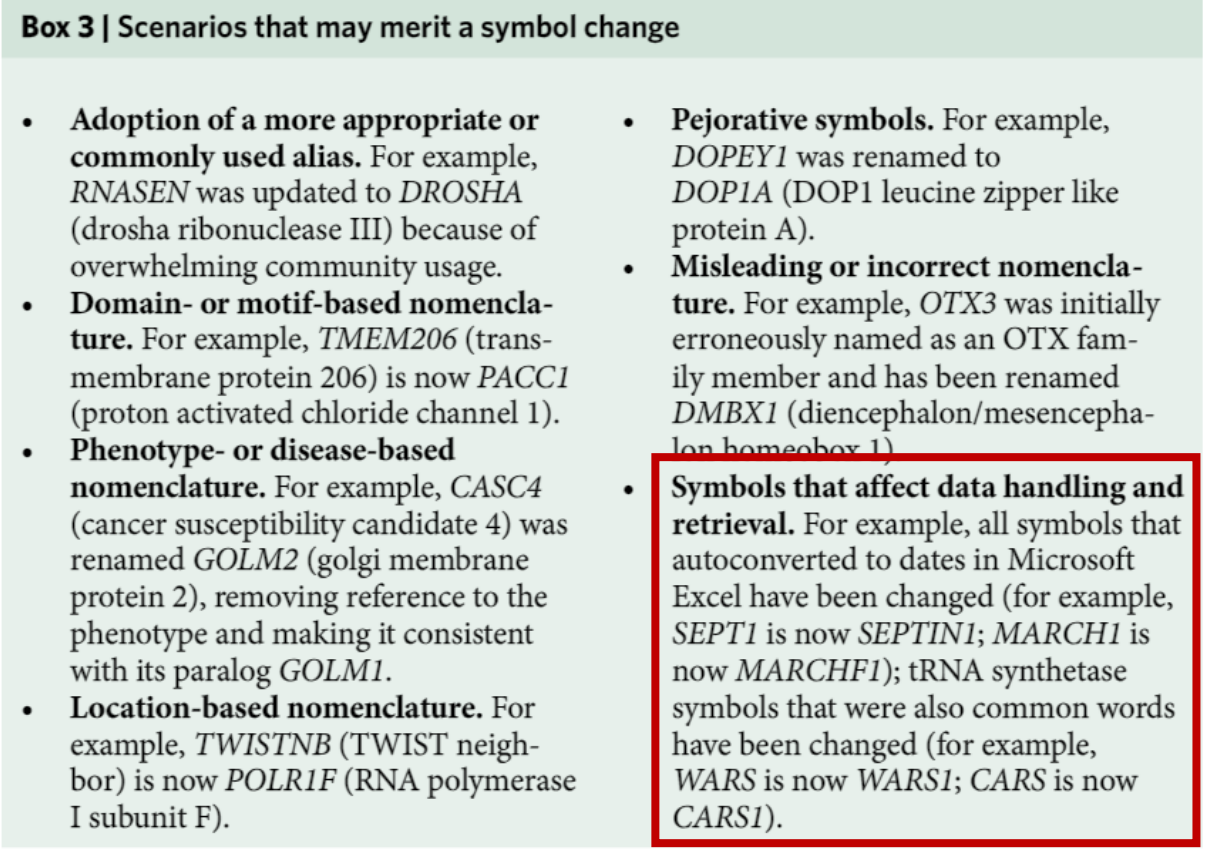
- Or: change all the gene names?!...
  - MARCH1 → MARCHF1
  - SEPT1 → SEPTIN1



Tips to Address Diabolical Date Formatting

- Or: change all the gene names?!..
  - MARCH1 → MARCHF1
  - SEPT1 → SEPTIN1

Human Gene Nomenclature Committee 2020 Guidance →



Box 3 | Scenarios that may merit a symbol change

- **Adoption of a more appropriate or commonly used alias.** For example, *RNASEN* was updated to *DROSHA* (drosha ribonuclease III) because of overwhelming community usage.
- **Domain- or motif-based nomenclature.** For example, *TMEM206* (transmembrane protein 206) is now *PACC1* (proton activated chloride channel 1).
- **Phenotype- or disease-based nomenclature.** For example, *CASC4* (cancer susceptibility candidate 4) was renamed *GOLM2* (golgi membrane protein 2), removing reference to the phenotype and making it consistent with its paralog *GOLM1*.
- **Location-based nomenclature.** For example, *TWISTNB* (TWIST neighbor) is now *POLR1F* (RNA polymerase I subunit F).
- **Pejorative symbols.** For example, *DOPEY1* was renamed to *DOP1A* (DOP1 leucine zipper like protein A).
- **Misleading or incorrect nomenclature.** For example, *OTX3* was initially erroneously named as an OTX family member and has been renamed *DMBX1* (diencephalon/mesencephalon homeobox 1).
- **Symbols that affect data handling and retrieval.** For example, all symbols that autoconverted to dates in Microsoft Excel have been changed (for example, *SEPT1* is now *SEPTIN1*; *MARCH1* is now *MARCHF1*); tRNA synthetase symbols that were also common words have been changed (for example, *WARS* is now *WARS1*; *CARS* is now *CARS1*).

Ghoulish Geolocation Coordinates

Ghoulish Geolocation Coordinates

- CMS public use Covid nursing home dataset had a string variable for geographic coordinates that combined latitude and longitude
- At first, there was no other info on “GEOLOCATION” field
- I split the string variable into what I presumed was latitude and longitude to use for mapping

FACILITY ID	GEOLOCATION
00176	(-115.121992, 36.107369)
23257	(-86.311653, 31.001042)
78791	(-77.05922, 38.734332)
45228	(-118.036817, 34.061096)

Ghoulish Geolocation Coordinates

FACILITY NAME	GEOLOCATION	latitude	longitude
WESTRIDGE HEALTH CARE CENTER	(84.24571, 39.26015)	84.24571	39.26015

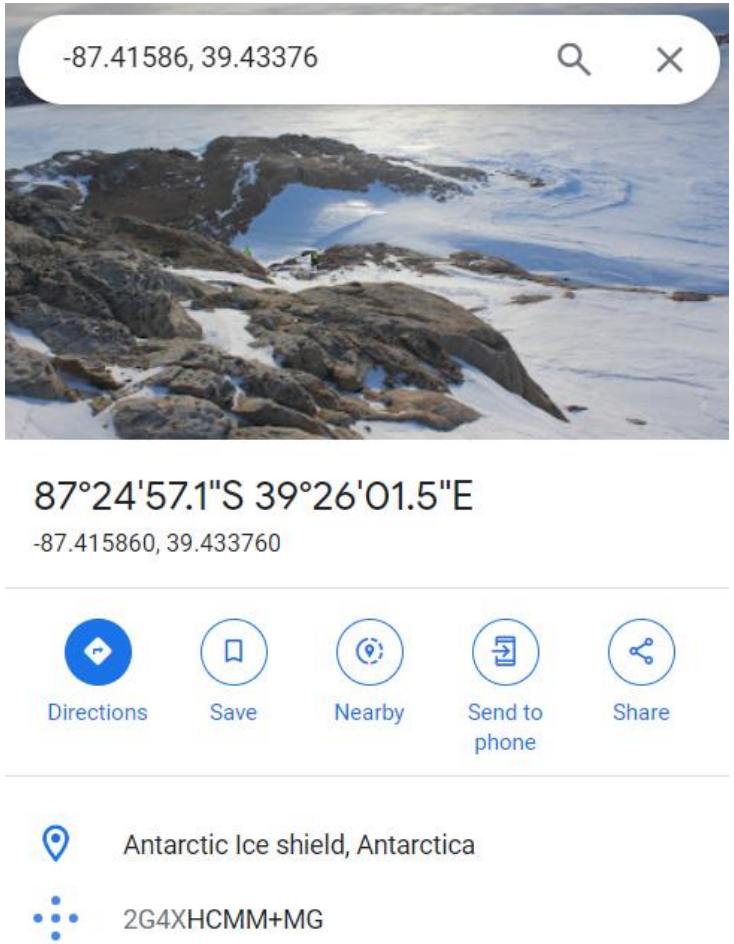
Ghoulish Geolocation Coordinates

FACILITY NAME	GEOLOCATION	latitude	longitude
WESTRIDGE HEALTH CARE CENTER	(84.24571, 39.26015)	84.24571	39.26015

- Source: CMS Covid Nursing Home public use dataset
- Wrote an R script to map facilities (~15,000 nursing homes in the U.S.) but RStudio kept crashing
- Subsetted by states (NV, IN) to try and troubleshoot for smaller lists of facilities

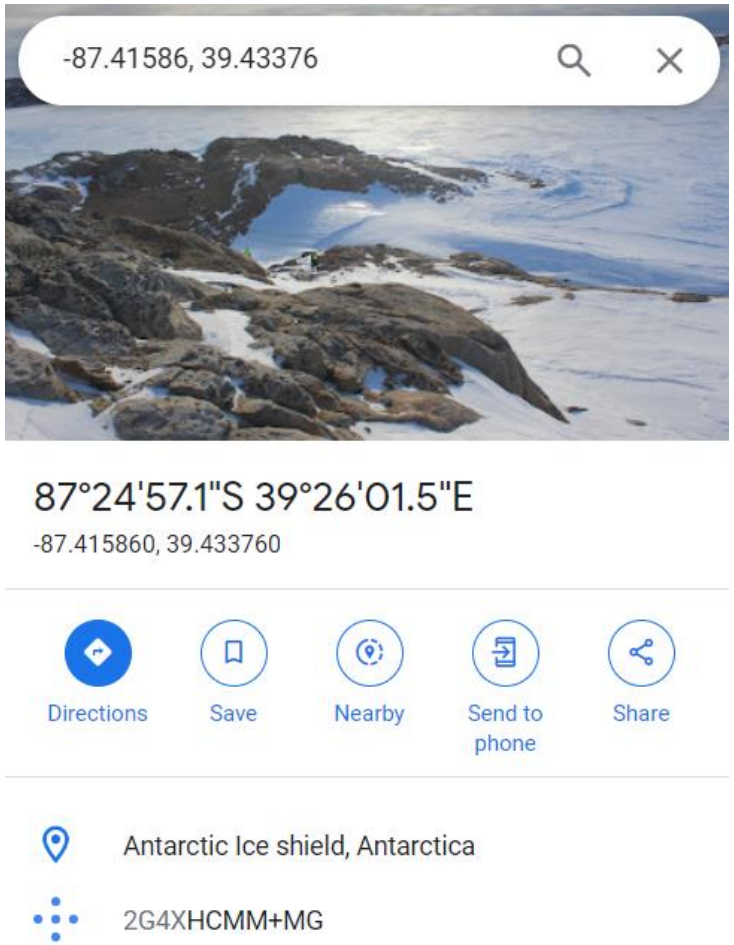
Ghoulish Geolocation Coordinates

- A number of the nursing home locations in Indiana had mysteriously flown off to Antarctica



Ghoulish Geolocation Coordinates

- A number of the nursing home locations in Indiana had mysteriously flown off to Antarctica
- Latitude and longitude were reversed in the original geolocation string!



FACILITY NAME	GEOLOCATION	latitude	longitude
WESTRIDGE HEALTH CARE CENTER	(-87.419773, 39.433756)	-87.419773	39.433756

Tips for Addressing Ghoulish Geolocation Coordinates

- **Data dictionaries** are the key to preventing this type of problem
- One option that can work well:
  1. High-level, succinct, machine-readable **csv** variable dictionary paired with
  2. Longer **pdf** technical documentation

- Also great to include contact info on pages with public use federal data so users can provide feedback on documentation!

Tips for Addressing Ghoulish Geolocation Coordinates

var_name	var_type	description	example	range	source	notes
county_name	String	County name	Alameda	NA	Census	
county_fips	String	County FIPS code	11003	NA	Census	
week_ending	Date	Ending date of measured week	2022-10-11	2021-01-01 - 2022-12-31	JHU Covid database	
covid_cases_num	Numeric	Number of covid cases in county, by week	137	0-3429	JHU Covid database	
covid_cases_rate	Numeric	Covid cases per 100k population	50.2	0-441	JHU Covid database; Census	Population denom derived from Census 5-year ACS survey

Tips for Addressing Ghoulish Geolocation Coordinates

- HHS CMS metadata standards for machine-readable price transparency data: [hospital-price-transparency/documentation/CSV at master · CMSgov/hospital-price-transparency]

hospital-price-transparency / documentation / CSV / ↑ Top

General Data Elements

These required general data elements about the MRF must be stated once at the top of the file (i.e. the first row).

Column Header (Tall format)	Column Header (Wide format)	Name	Type	Description	Blanks Accepted
hospital_name	**hospital_name**	Hospital Name	String	The legal business name of the licensee.	No
last_updated_on	**last_updated_on**	MRF Date	Date	Date on which the MRF was last updated. Date must be in an ISO 8601 format (i.e. YYYY-MM-DD). See additional last updated on notes	No
version	**version**	CMS Template Version	String	The version of the CMS Template used.	No
hospital_location	**hospital_location**	Hospital Location(s)	String	The unique name of the hospital location absent any acronyms.	No
				The geographic	

Tips for Addressing Ghoulish Geolocation Coordinates

- NCHS metadata standards:
  - https://intranet.cdc.gov/nchs/data-science/metadata.htm
- External resources that go into a nice level of detail about style and formatting advice for data dictionaries and metadata:
  - Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set
  - Improving Metadata Infrastructure for Complex Surveys: Insights from the Fragile Families Challenge - PMC (nih.gov)

Thank you!