Kate Barron
618 Individual Project--Meteorite Landings

1. Motivation

In the past, I have analyzed human-centered data sets--for example, data sets exploring how people use technology services, or exploring health in specific populations. For this project, I wanted to stretch myself by analyzing non-human-centered data. To this end, I have chosen a meteorite landing data set. Initially, I developed the following data questions:

1) What are typical qualities of valid and relict meteorites?
2) Where do found meteorites tend to land? What about fell meteorites?
3) How does time affect the class of meteorite?
4) Have meteorite falls increased over time? Decreased?

2. Data Source

My data set was Meteorite Landing Data (pulled as a CSV from NASA's Data Portal [https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh]). I originally found this dataset on (https://www.kaggle.com/nasa/meteorite-landings). Its variables were described by the kaggle page as follows:

Field	**Description**	**Data Type**
name	Name of meteorite	nominal
id	Unique, numerical identifier	nominal
nametype	*Valid:* typical meteorite *Relict:* meteorite which has been highly degraded by weather on Earth	nominal
recclass	Class of meteorite (based on physical, chemical characteristics, etc.)	nominal
mass	Mass of meteorite in grams	real-number
fall	*Fell:* meteorite's fall was observed *Found:* meteorite's fall was not observed	nominal
year	Year the meteorite fell or was found (depending on value of fall)	ordinal
Reclat	Latitude of meteorite's landing	real-number
Reclong	Longitude of meteorite's landing	real-number
GeoLocation	Parentheses-enclosed tuple containing reclat and reclong	real-number

The kaggle page included "notes on missing or incorrect data points," which I referred to while cleaning my data. The kaggle page also directed me to a meteorite classification guide (https://en.wikipedia.org/wiki/Meteorite_classification), so as to better understand the field "recclass." The original dataset included 45717 records. Years covered were: 860 CE to 2016 CE (some entries were attributed to years 2017-2029, which were errors).

3. Methods

Before approaching my four data questions, I prepared the data set using Python. I accomplished this task in four steps.

1) First, I deleted the GeoLocation field using regular expressions, since the data was redundant (i.e. I already had latitude and longitude, and did not need a field combining the two). Next, I converted each record into a list. I then added these lists to a data structure ("structure_1"), excluding any record with at least one blank field. This reduced my data set from 45717 records to 38117 records; I felt the amount of data would still be sufficient for analysis.
2) I discovered that some "recclass" entries had inadvertently been split into two during the first step. I identified these fields by selecting all items in structure_1 for which the length was greater than 9. I then merged the split entries (which I knew occurred on positions 3 and 4). I then appended all records (those which I had edited, and those which were left alone) a new data structure, "structure_2."
3) I parsed the records in structure_2, designating that records in which both latitude and longitude were 0 should be passed (these were error entries), and all other records should be appended to a new data structure, "structure_3." This step reduced my dataset from 38117 records to 31931 records, which I felt would still be sufficient for analysis. After structure_3 was formed, I used regular expressions to clean the "year" field, so that it would include only year (and exclude month, day, and time--these appeared to be arbitrarily chosen as January 1st, and 12:00am).
4) Finally, the kaggle page had identified any year before 860 or after 2016 as error entries. I created a final data structure, "structure_4," which excluded records with said years. My final data set contained 31723 rows, which I still considered sufficient for analysis. I then wrote these records to a fresh CSV, called "meteorite2.csv."

Regarding each data question:

Question #1:

I attempted to use faceting to discover qualities of valid and relict meteorites, but the resulting graphics were not telling. Further, I realized that my data set only contained 21 records of relict type (and 31702 records of valid type). It appeared that the question I was asking could not be

answered given the small number of relict records. I decided to make a new field, called "class_2," which would assign one of the following categories to a record (based on recclass):

- Achondrites (stony)
- Chondrites (stony)
- Magmatic (iron)
- Non-magmatic (iron)
- Pallasites (stony-iron)
- Mesosiderites (stony-iron)
- Ungrouped iron (iron)
- Ungrouped stone (stony)
- Unknown

...wherein each category is the 2nd tier of the stony, stony-iron and iron classification system. However, chondrites *still* dominated the data set (29381 out of 31723 records). For lack of time, I decided to proceed with these categories--but I think the best approach would have been to subdivide the chondrites themselves into their 3rd tier classifications, and to investigate trends therein.

As I played with faceting, I noticed that my extremely large and small mass values were masking mass trends. I decided to remove masses smaller than 1750000 and larger than 0. After running a few test plots, it appeared that the most telling variables for faceting would be mass, lat and long, based on "class_2" facets and year_group color-coding.

Question #2:

Originally, I tried to plot every record using a combination of ggplot2 and map library functions (color-coding each point by found or fell). However, this mapping was extremely cluttered and difficult to read. Further, the map didn't convey anything in terms of trends. As such, I added a new field to my R data frame called "year_group," which created 7 year categories (by century, pre-16th through 21st). I encoded year_group by color, and found/fell by shape. However, this new mapping was still difficult to read. Finally, I decided to create two separate maps: one which include "found" data only, and one which would include "fell" data only (with points still color-coded by year_group). During this process, I realized there were significantly more "found" records than "fell" records (972 vs. 30751, respectively).

Question #3:

I originally tried to accomplish this task by creating a scatterplot (x-axis=year, y=axis=count, color=class_2). However, the results were not very helpful, as two chondrite outliers were very apparent while all other points hovered together below 2500 count (and appeared more or less as a straight line of dots). First I tried to eliminate the outliers by modifying the y-axis to run from 0 to 2500. Even with this change, points before 1800 appeared as a straight line, and the

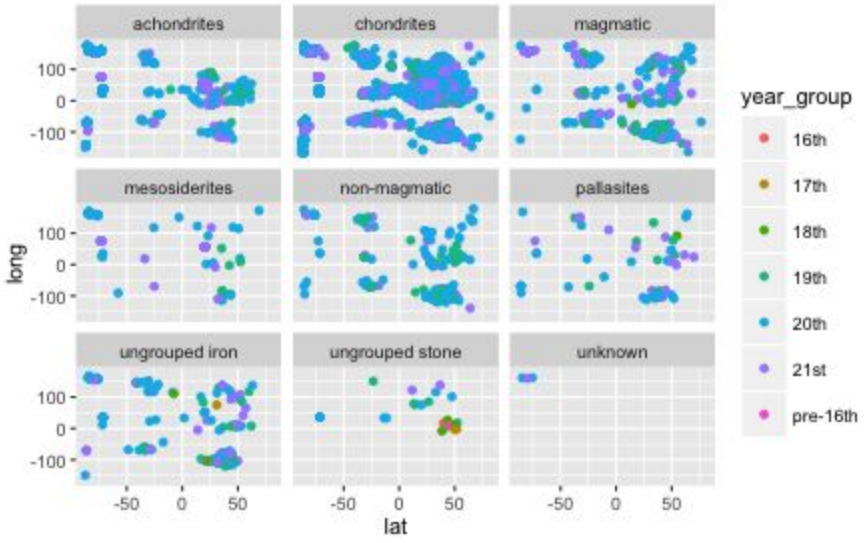
class_2 values “unknown,” “ungrouped stone” and “ungrouped iron” never appeared to increase over time. I created a new data frame (“reduced”), which subsetted my original data frame (df) to exclude class_2 values “unknown,” “ungrouped stone” and “ungrouped iron,” as well as years before 1800. Finally, I redrew my graphic as both a scatter plot and a line graph (wherein points/lines were color-coded by class_2 values).

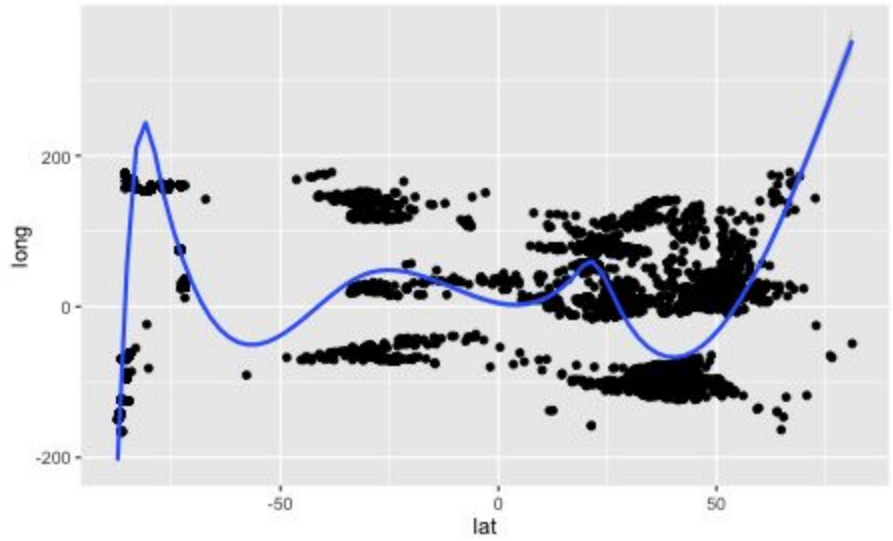
Question #4:

I was able to accomplish this task by plotting a histogram (x-axis=year_group [created for Questions #1 and #2], y-axis=count).

4. Analysis and Results (36 points): For <u>each</u> of the four questions, provide the following:

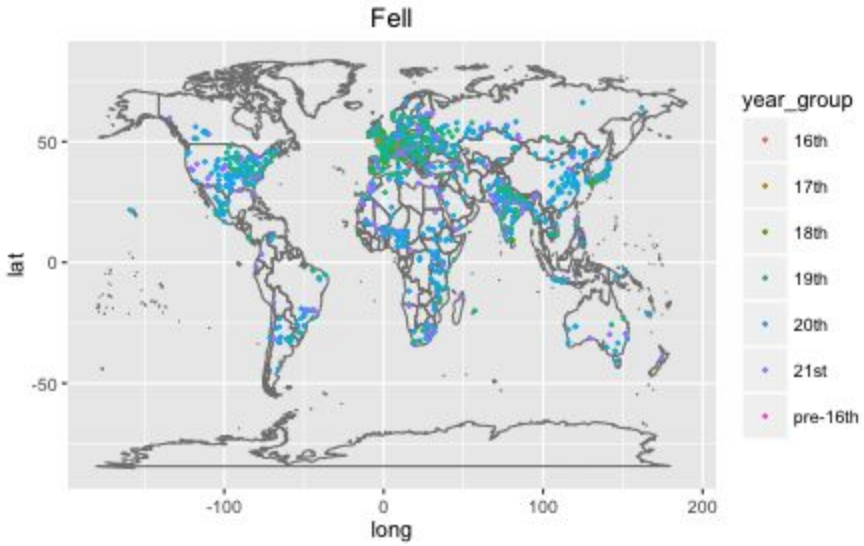
Question #1:

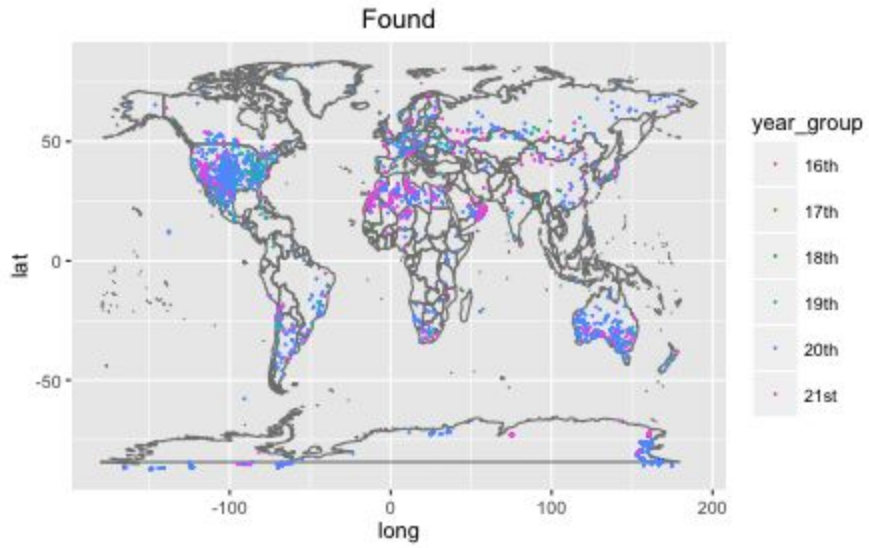




This question evolved into: what are typical qualities of each class_2 meteorite type? As it turned out, this was a difficult question to approach because there wasn't a lot of continuous data available. However, I was able to make some interesting longitude versus. latitude plots--an approach I had not before considered. These plots, as I found, would supplement my question #2 maps. Most landings occurred above 0 degrees latitude and 0 degrees longitude (with some clusters around -100 degrees longitude). These findings match those in the question #2 maps. My second plot (of aggregated class_2, latitude versus longitude) makes these findings more apparent.

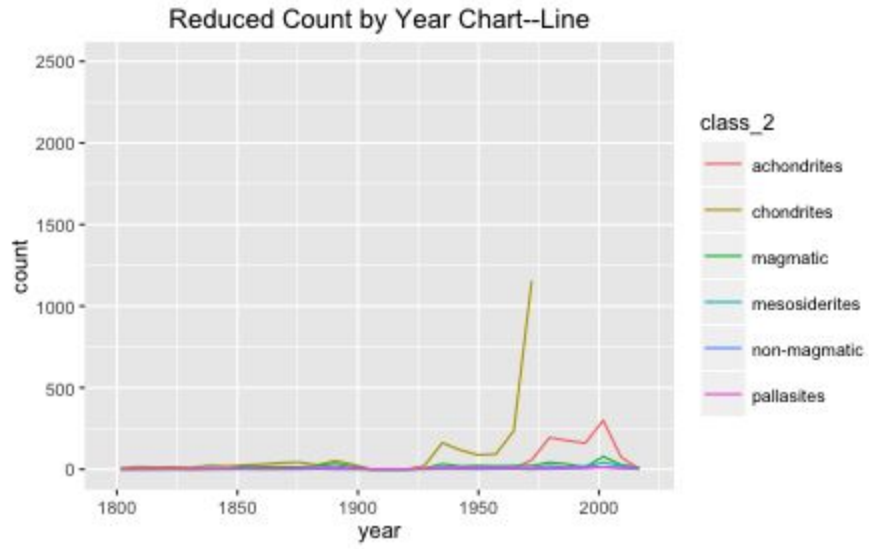
Question #2:

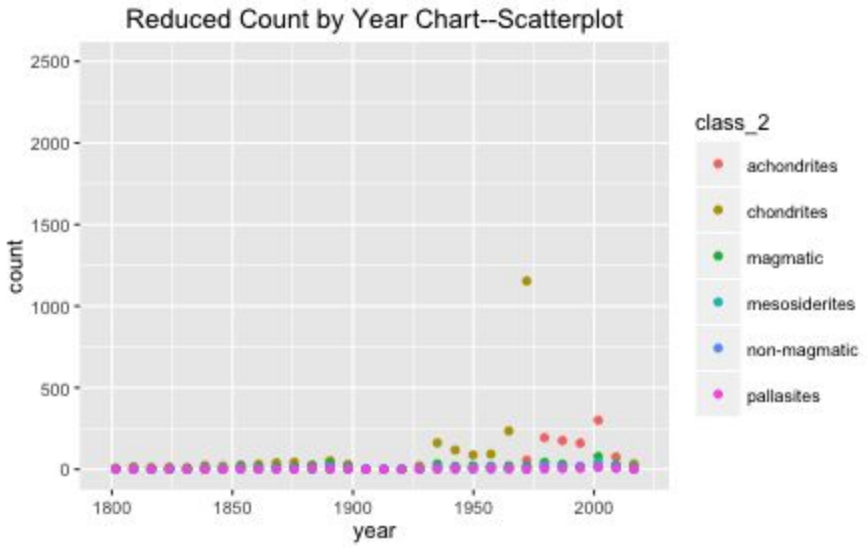




The greatest concentration of “fell” meteorites appears to be in Europe (ostensibly followed by South Asia and the eastern United States). Counterintuitively, a lot of “fell” meteorites were observed in the 18th and 19th centuries. On the contrary, of a large amount of found meteorites were discovered in the United States in the 20th and 21st centuries. This may reflect increasing spending on astronomical sciences in the last 75 years or so. In both maps, meteorites have not been identified (or reported) in similar regions: Canada, Greenland, Russia, sub-Saharan Africa, northern South America, Northern Australia and much of Antarctica. I am wondering if this is due to the kinds of terrain in these areas (rainforest, tundra, etc.). I struggle to understand why South Asia should have so many “fell” meteorites (and from the 18th/19th centuries to boot), but so few “found” meteorites.

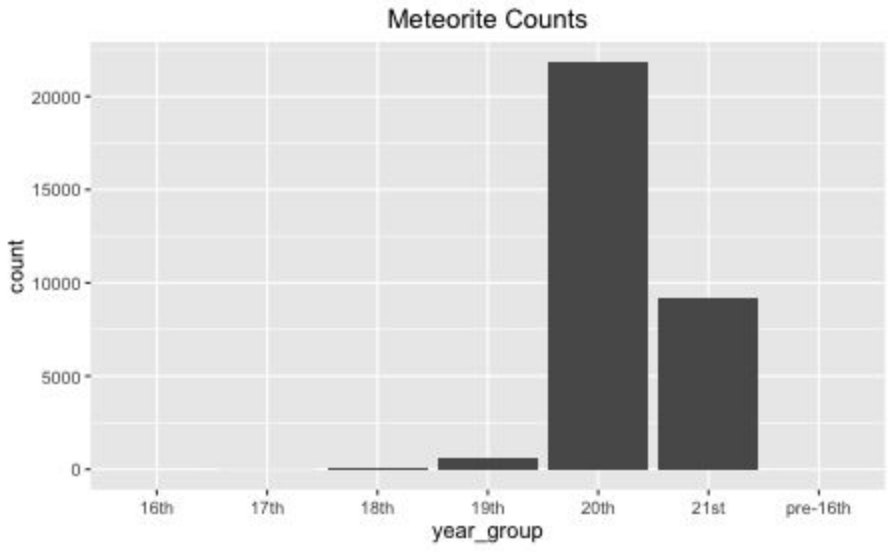
Question #3:

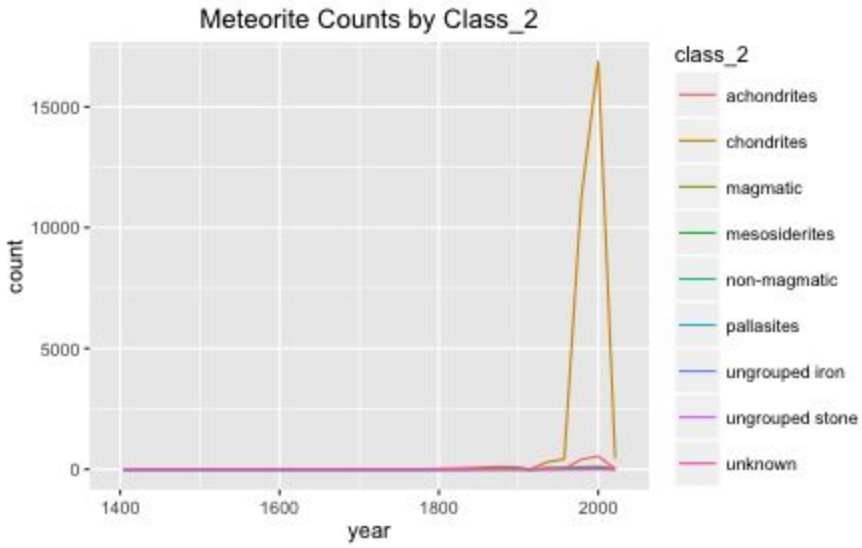




I'm not sure that this exploration is very telling, as findings for question #4 show that recorded meteorite landings have increased in the 20th/21st centuries. As such, it makes sense that some class_2 categories see an increase over the last 100 years or so. I'm not sure how this question could ever truly be answered, assuming that meteorite detection methods continue to change. For whatever reason, chondrite landings have increased the most over time (followed by achondrites--so stony meteorites represent the largest portion of landings).

Question #4:





Ostensibly, meteorite landings have increased dramatically from the 19th century to the 20th century, but I think that the histogram is misleading. Undoubtedly, meteorites are better recognized/detected in the 20th and 21st centuries due to more resources and better technology. Additionally, I imagine that methods for recording meteorite landings have not been consistent over the last 500 years. In summary, just because a meteorite landing was not recorded does not mean it didn't happen. Further, I'm not sure how this question could ever truly be answered, assuming that meteorite detection methods continue to change. The line graph shows that chondrite landings have increased the most in the 20th century--but if chondrites are simply the most common meteorite to land on earth, I'm not sure that the finding is profound.