PS1
Name: Yangyang Dai

Part 1: Write a data section for your assigned data set (5 points)

City of Chicago Food Inspections

The source can be found here https://webapps1.cityofchicago.org/healthinspection/inspection.jsp. The data are owned by the Chicago Department of Public Health.

One key paper that uses this data is the 2016 Chicago Department of Public Health Food Establishment Inspection Audit by the office of Inspector General of the City of Chicago. (Office of Inspector General, City of Chicago. (2016). Chicago Department of Public Health Food Establishment Inspection Audit. Chicago. Retrieved from http://chicagoinspectorgeneral.org/wp-content/uploads/2016/11/Audit-of-CDPH -Food-Establishment-Inspections.pdf)
Another one is Fudging the Nudge: Information Disclosure and Restaurant Grading by Daniel E Ho, which studies health inspections data from ten jurisdictions to examine the effectiveness of the restaurant letter grading system in alleviating actual manifestations of foodborne illness over time.
(Ho, D. E. (2012). Fudging the Nudge: Information Disclosure and Restaurant Grading. The Yale Law Journal, 122(3), 574688.)

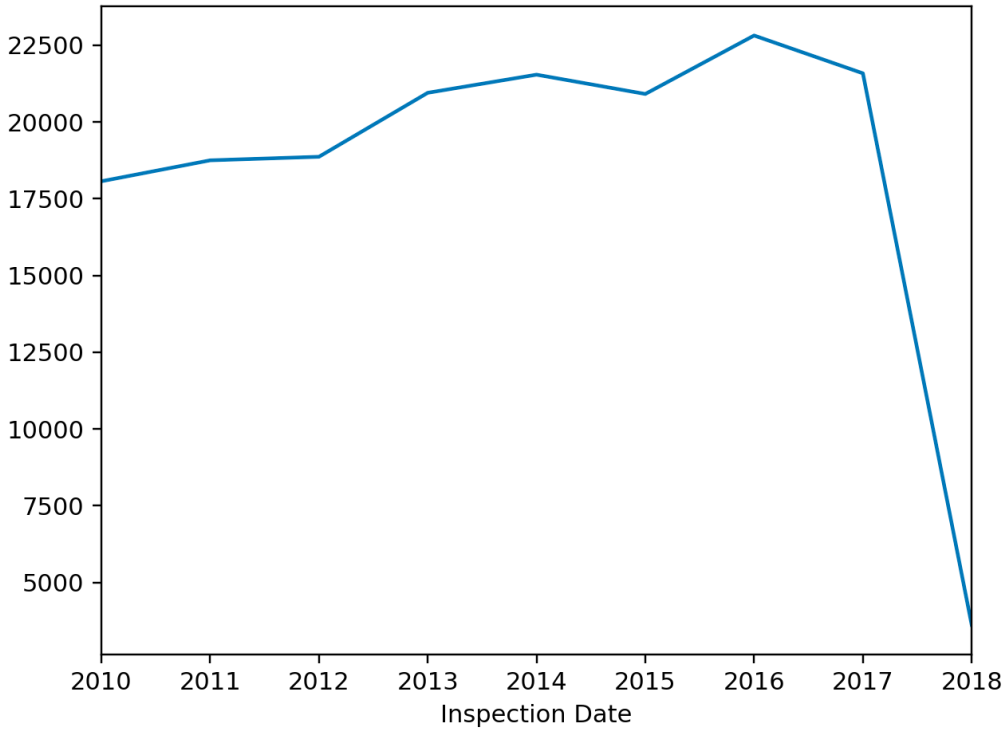
This information is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010 to the present. Inspections are performed by staff from the Chicago Department of Public Health's Food Protection Program using a standardized procedure. The results of the inspection are inputted into a database, then reviewed and approved by a State of Illinois Licensed Environmental Health Practitioner (LEHP).

	Inspection Type	Zip	Latitude	Longitude	Risk	Facility Type	DBA Name	AKA Name
count of categories	1.08E+02	100	166520	166520	4	453	25549	24401
min	1.00E+00	60007	41.64467	-87.914428	117586	1	1	1
max	8.90E+04	60827	42.021064	-87.525094	22	110828	2323	2800

The eight variables I include here are Inspection Type, Zip, Latitude, Longitude, Risk, Facility Type, DBA Name, and AKA name. And the statistics I include are count of categories, min of value, max of value in each category.

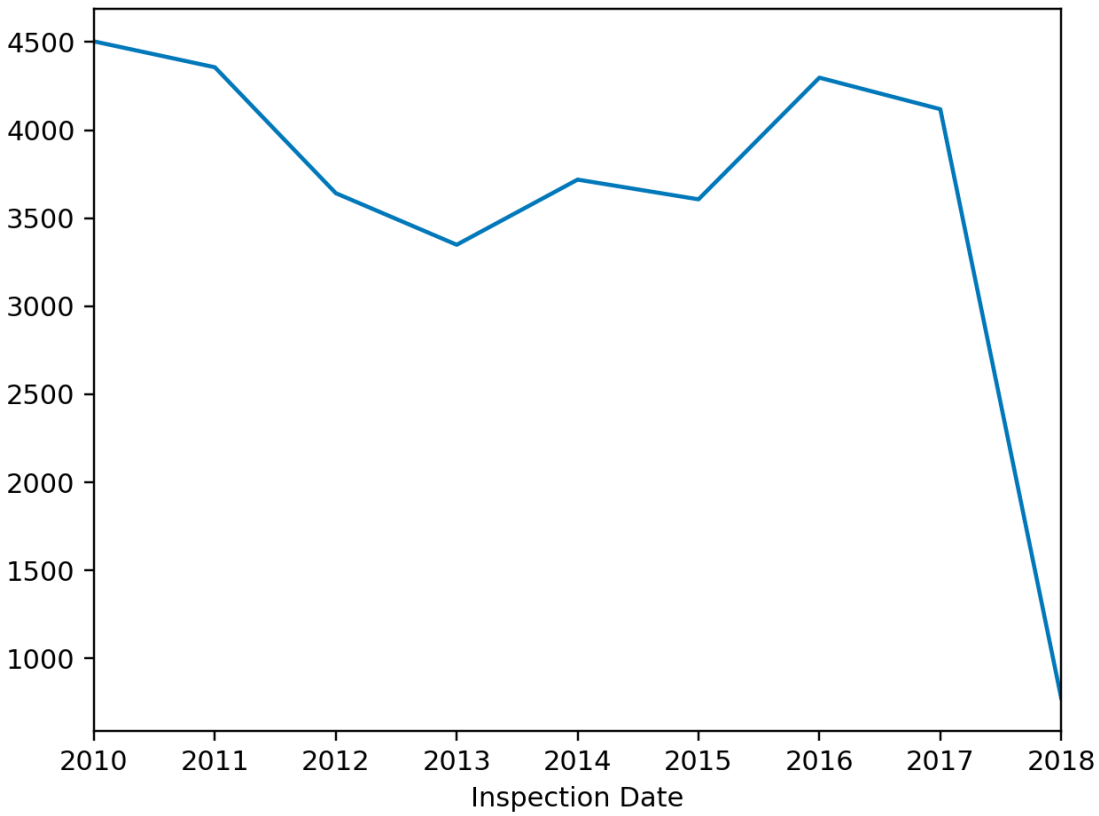
Simple visualization:

Fig 1: Inspection counts by year



I visualized the data of number of inspection counts versus year form 2010 to 2018. There is a sharp drop in 2018 simply because of lack of data since it is just the beginning of the year. We can see that the general rends increase over the years.

Conditional visualization:
Fig 2: Failed Inspection counts by year



This is a figure showing the failed inspection counts grouped by year. We can see that the number decreased after 2010 but rises a little in 2016.

Part 2: Critique a computational research paper (5 points)

Research Question:
How House members' expressed priorities respond to tumultuous changes in institutional and electoral contexts.

The paper uses a collection of nearly 170,000 House press releases: every press release, from each House office, from 2005 to 2010 as its data source.

Theory used to classify the data/ computational methods:
To construct the hierarchy of topics the model nests, or classifies, a set of granular topics into a set of coarse topics. This modeling strategy builds on Pachinko Allocation Models, that allow for a nesting of topics, while contributing a model that relies on a different distribution that allows for fast inferences. And Structural Topic Model(STM), clustering models, and task-specific tests are also used to conduct text analysis.

The type of this study is a combination of a descriptive study and an identification exercise. It firstly explores the dataset using unsupervised learning model and then classifies the data. And then, the study aims to identify and "classify granular topics into a set of coarse topics". At the end of the study, it reveals that this method and study is able to reveal how legislators' expressed priorities change.

Computational Methods:
The press release is stored using a document-to-vector model where each element of the 2,727 element-long count vector records the number of times a token occurs in a document from a specific legislator.
Secondly, a two-layer hierarchical model is applied to cluster the documents into granular topics, which are then clustered into coarse topics, which can be analyzed to determine change in communication styles.

Results:
Republicans abandon credit claiming. Instead, Republicans articulate criticisms of the Democratic party, the Obama administration, and Democratic policy proposals. In contrast, Democrats embrace credit claiming and defend Democratic policies—though less vocally than Republicans criticize those same proposals.
There is a strong year- to-year relationship in legislators' presentational styles. While legislators are responsive at the margin to changing conditions, the basic strategy remains the same.

Comments:

1) It might be hard to find, but it would be really helpful if more data or information can be added after year 2010, since there have been a lot of critical events happened after that time and might be highly relevant to the future paths of legislators.
2) If there are more evidences and channels showing the interaction between constituents and legislators, it would be great to add that aspect into the paper. Since the paper assumes a strong reliability of press release in capturing communication between legislators and constituents, more theories and past documents indicating this relation would give us a better understanding towards this matter.