Intro: 0.5p

* Business case
* Classification
* Enumerate methods

Data: 2p (2/3 per person)

What datasets

Data preparation

Analysis: 3p (1 per person

If time: discussion

Conclusion 0,5p

# Intro:

It is estimated that every year 48 million people fall sick, 128’000 are hospitalized and 3’000 die from food born illnesses (Centers for Disease Control and Prevention, 2018). These numbers clearly illustrate the necessity of clear and strict oversight over any organisation handling large quantities of food, thusly constantly operating under the danger of poisoning a significant part of the population. Among the four governmental agencies working on minimizing food posed risks, the ‘Food Safety and Inspection Service’ handles (among other things) the inspections of food retailers (Food Safety and Inspection Service, 2018). While agencies name might not be amongst the most recognizable ones, it is nevertheless employing over 9’000 people whilst presiding over an annual budget ranging from 1,03 – 1,05 billion US$. Its field operations in the State of New York alone incur a cost in the range of 13,3 – 15,3 million US$ each year (U.S. Census Bureau, 2018, p. 16)

The official budget report unfortunately does not include a breakdown of the costs directly caused by the inspections. Nonetheless, the classification by objects shows that the vast majority of funds, a range of 960 – 990 million US$ (2016 – 2019), are connected with inspections, suggestions them to be among the most central cost points of the organization (U.S. Census Bureau, 2018, p. 17).

Based on this information, we conclude that successful predictions of which food retailers pose the highest public health risk and the subsequently more efficient allocation of resources, could not only lead to cost reductions of millions of US$ every year but might even save lives by preventing the outbreak of dangerous foodborne illnesses.

The following paper will outline the steps that were taken in order of attempting to create an algorithm, with said predictive goal. In section 2, it discusses the steps tat were taken to gather the data for various independent variables, whereas section 3 describes the subsequent data analysis.

# Demographic data

Since the original dataset did not feature any demographic data, related to the inspected stores, this type of data had to be acquired otherwise and matched subsequently to the inspection dataset. The obvious choice for the matching was to match by location. Fortunately, the amount of demographic data on different sections of the state New York is ample, in the form of different datasets detailing the results of various census-endeavours. Unfortunately, a large percentage of these datasets turned out to be unusable. This dataset, for example, includes demographic data by zip code, which would have provided a straightforward way of matching with the inspection’s dataset, turned out to be based on sample size oftentimes just one or even zero questioned persons. We therefore chose to use two datasets based on US wide census studies, by the official census bureau of the USA. This information can be downloaded from the American FactFinder (US Census Bureau, 2017)

But are also available on Kaggle

<https://www.kaggle.com/muonneutrino/us-census-demographic-data> normally you are not supposed to cite Kaggle, how do we handle this?

. The first dataset featured the demographic makeup of every US-county; the matching between the two datasets was thus a very straightforward matter, since it could be done via the variable county. The disadvantage of this structure was however that its information did not necessarily exactly reflect the demographic makeup in the direct vicinity of the food retailers. The second dataset, which displayed the demographic makeup of the US, by Census Tract, the relatively small geographic locations specifically designed for the collection of census-data, offered notably more specific information about the retailers’ surroundings.

Since Census Tracts, however, do not follow the traditional pattern of location specification used in the inspection’s dataset, direct matching of the two sets was not possible. An approach of nevertheless matching the sets, was found in the creation of a third “translation” dataset AddTrac (short for Address to Census Tract) which matched the retailers’ locations with the Census Tract codes. To do so, we made use of the census bureau’s geocoding service (US Census Bureau), Geocoder)

This service added various geolocation-identifiers to specifically created csv files holding the addresses of our food retailer. Unfortunately, circa 20% of the addresses were not identified by the service (possible reasons include name changes and confidentiality concerns (U.S. Census Bureau, 2018, p. 7)) which led to some data loss. By combining the output of state FIPS , county FIPS and census block codes, as well as adding placeholder zeros where necessary, the Tract Ids could be recreated, which ultimately enabled the merging of the more exact Census Tract based data with the inspection’s dataset.

# Conclusion

Despite the utilization of an array of different variables and methods, the prediction of Inspection Grades proved to be exceedingly difficult.

#short recap by method of how well it went

It seems likely that the reason for this regrettable result lies in the low predictive quality of the here used independent variables. We have already showcased the, if anything, very weak correlation between the data that had been available to us and the dependent variable. It seems that the stores inspection grade is, if at all, only very weakly dependent on the customer experience as shown by google-ratings. A close look at deficiency descriptions illustrates why this might be the case: The deficiencies might be very minor or unnoticeable in the eye of the customer; utter filthiness certainly is no necessity for failure.

Whilst this has rendered the attempt of predicting inspection grades with publicly available ratings and demographic data largely in vain, it remains a possibility that the situation would be far less dire for the ‘Food Safety and Inspection Service’. With access to variables such as the track record of the owner and management (who influence the companies policy towards food safety) and the employees (who act on said policy) far superior results may be achievable.