**CS 573 FINAL PROJECT PROPOSAL**

**INTERACTIVE DATA VISUALIZATION**

**NYC RESTAURANT FOOD SAFETY**

**TEAM**

Shile Zhao ([szhao2@wpi.edu](mailto:szhao2@wpi.edu), 1nfo)

Zhaojun Yang ([zyang7@wpi.edu](mailto:zyang7@wpi.edu), yzj1212)

Shaowei Gong ([sgong@wpi.edu](mailto:sgong@wpi.edu), codescracker)

Cut Famelia ([cfamelia@wpi.edu](mailto:cfamelia@wpi.edu), CutFamelia)

**BACKGROUND AND MOTIVATION**

New York City (NYC) is one of the most diverse and cosmopolitan cities in the world, where many people come from many different backgrounds and cultures. Along with this fact, NYC offers many options for basic needs for the people living in the city. One of the needs is restaurants that sell different types of food based on the diverse needs such as American, Chinese, Indian, Mexican, Mediterranean, Asian, and so forth. All these different choices of restaurant are spread out across the city’s six boroughs.

Along with the increasing number of those thematic restaurants, the quality of the restaurants have become a serious concern for the government and the people. One of the quality aspects is the safety of the food they provide. To address this issue, the government under the Department of Health and Mental Hygiene (DOHMH) has been running official inspections towards all the restaurants.

“In July 2010, the Health Department began requiring restaurants in all five boroughs to post letter grades summarizing their sanitary inspection scores to help achieve three goals: to inform the public about a restaurant’s inspection results in a simple, accessible way; to improve sanitary conditions and food safety practices in restaurants; and to reduce illnesses associated with dining out. This report summarizes progress toward these goals.” (<http://www1.nyc.gov/assets/doh/downloads/pdf/rii/restaurant-grading-18-month-report.pdf>)

In fact, surprisingly, the inspections have found so many food safety rule violations in many of the restaurants in different ways. According to the report, New Yorkers eat food at restaurants almost a billion times every year. While most of them do not get sick, foodborne bacteria, viruses and other contaminants cause millions of cases of illness each year, it is estimated that more than 6,000 New Yorkers are hospitalized and 20,000 visit emergency rooms each year because of foodborne illnesses. Each year, NYC receives approximately 2,700 complaints about restaurant-acquired foodborne illnesses and another 3,000 complaints about restaurant hygiene.

Before letter grading, restaurants were motivated to practice food safety by their own desire to maintain healthful conditions and by the threat of fines for violations found at the time of inspection. Grading introduced a third and potentially more significant incentive: recognition with an A grade for excellent food safety practices. From the data they have collected, it shows that not every restaurant has done improvement based on the inspection findings and recommendations. This fact is indicated by the lower grades given to the restaurants after the last inspection. To evaluate the impact of the inspections, it is important for DOHMH to see the trend of the grades given to each restaurant after every inspection was done, whether it is going up or going down, or fluctuating. Thus, DOHMH can easily do some analysis on the trends to get some insights.

From all the above findings, we believe that it is important for New Yorkers, especially the customers who like to eat outside and care about the food quality to see the inspection data summary and analyze the results of the inspections in an easy and interactive way to help them make decisions on choosing better places to eat to maintain healthy lifestyle. Also, the visualization will be useful for food safety supervisors and analysts to support their evidence-based works.

**OBJECTIVES**

This project aims to address the following questions:

1. How much the inspections affect the restaurant food safety practices.
2. How the rank of every restaurant type looks like.
3. Which type of restaurants have the highest rate of food safety.
4. Which type of restaurants have the lowest rate of food safety.
5. How every restaurant type’s grade performance looks like based on location and period of time.

In this project, we expect that we will learn how to work on the followings tasks:

1. Creating attractive-animated different types of graphs to explore the data
2. Creating effective user interfaces
3. Manipulating data to fit the visualization

**DATA**

We will use DOHMH New York City Restaurant Inspection Results dataset retrieved from NYC Open Data website

(<https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>). This website provides data on restaurant inspections, violations, grades, and adjudication information of 439,301 records.

The structure of data is described as follow:

* CAMIS: This is an unique identifier for the entity (restaurant)
* DBA:This field represents the name (doing business as) of the entity (restaurant)
* BORO:Borough in which the entity (restaurant) is located.
* BUILDING:This field represents the building number for the entity (restaurant)
* STREET:This field represents the street name at which the entity (restaurant) is located.
* ZIPCODE:Zip code as per the address of the entity (restaurant)
* PHONE :Phone Number
* CUISINE DESCRIPTION: This field describes the entity (restaurant) cuisine.
* INSPECTION DATE: This field represents the date of inspection
* ACTION: This field represents the actions that is associated with each restaurant inspection.
* VIOLATION CODE:This field represents the violation codes that is associated with each restaurant inspection.
* VIOLATION DESCRIPTION:This field is the description that corresponds to the violation codes
* CRITICAL FLAG:This indicates if Violation is critical or not.
* SCORE:Total Score for a particular inspection. If there was adjudication a judge may reduce the total points for the inspection and this field will have the update amount.
* GRADE:• N = Not Yet Graded • A = Grade A • B = Grade B • C = Grade C • Z = Grade Pending • P= Grade Pending issued on re-opening following an initial inspection that resulted in a closure
* GRADE DATE:The date when the current grade was issued to the entity (restaurant)
* RECORD DATE: The date when the extract was run to produce this data set
* INSPECTION TYPE:The type of inspection. A combination of the program and inspection type.

The followings are the details of the data attributes along with each number of the value types:

CAMIS = 25,969  
DBA = 20,506  
BORO = 6  
BUILDING = 7,228  
STREET = 3,309  
ZIPCODE = 230  
CUISINE DESCRIPTION = 84  
INSPECTION DATE = 1,335  
ACTION = 6  
VIOLATION CODE = 97  
VIOLATION DESCRIPTION = 95  
CRITICAL FLAG = 3  
GRADE DATE = 1,256  
INSPECTION TYPE = 34

**DATA PROCESSING**

The dataset takes 160 MB in memory, which is too much to fit into a web page. The pre-processing work includes to reduce the redundant columns. Considering the way of representing the data, we must pre-processing the dataset. The original dataset contains a lot of information, and what we can do is to extract the right data for each question and reorganize them. Also, we have to validate the field values for each column, to filter out invalid values and outliers. For example, in the first visualization, we want to use “tree chart” so the data must be constructed as a tree structure, or json objects. On the other hand, there are blank rows and rows with P and G values in the “grade” column, meaning that the scores are not graded yet. So, we will remove these rows as we only need the data of graded restaurants.

Since there are 20506 restaurants in total in NYC, which is too many to be put in one visualization, we decided to shorten the number by putting the restaurants into a number of categories based on the themes/types. We will visualize the restaurants by categories, i.e. 84 categories, which is actually still so many, but we think it is still possible and make sense to be visualized.

**VISUALIZATION DESIGN**

**Prototype 1:**

In this design, wewanted to visualize each restaurant on a map as the dataset relates to locations. However, we then realize putting all the restaurants on a map does not give much useful information about the inspection results of the restaurants. As a result, we choose not to use map here in the project. Instead, we use clustered bar charts to show the summary of different types of restaurants in different boroughs.

**Prototype 2:**

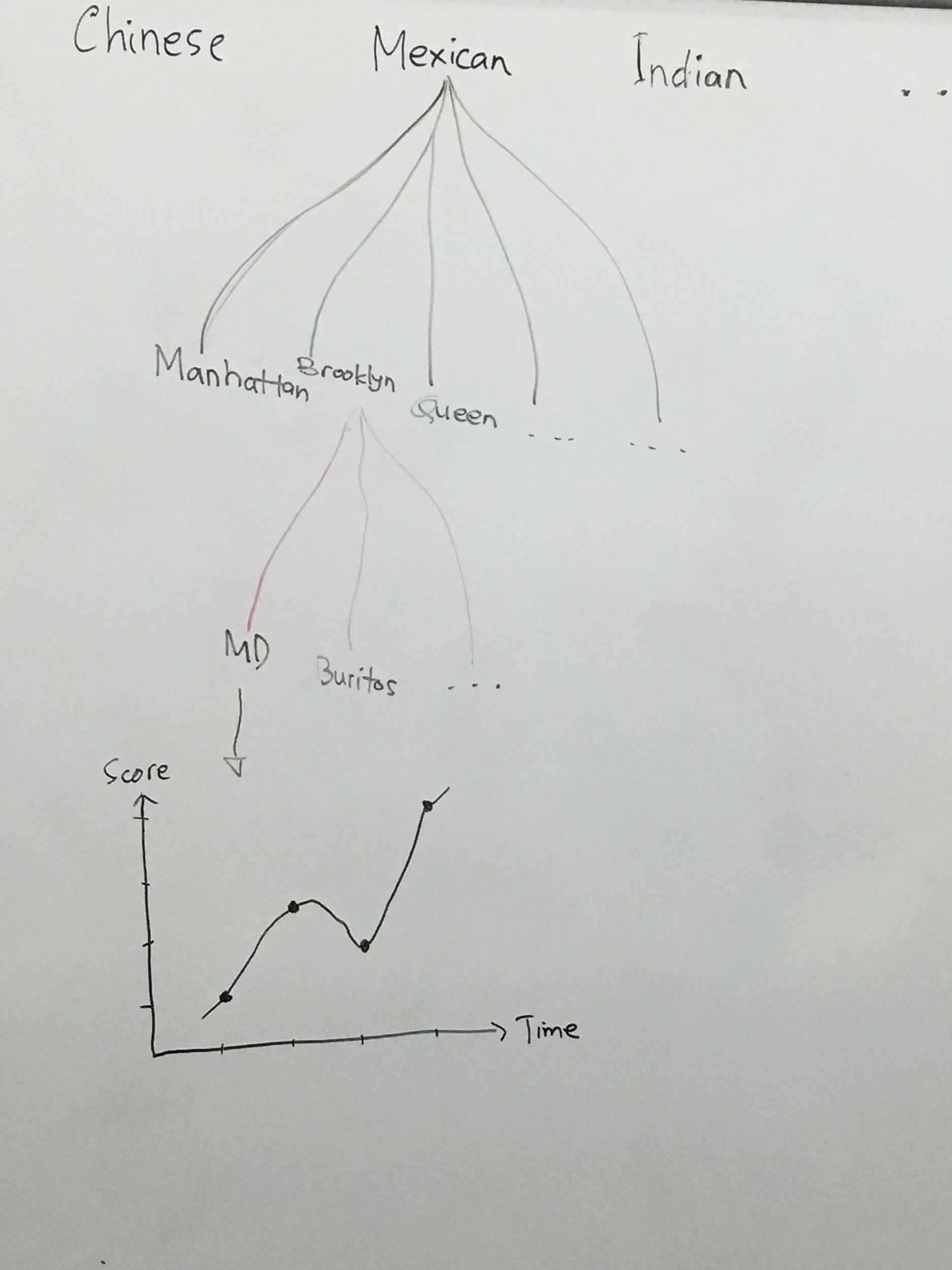
In this design, we came up with a tree graph to explore the attribute details because we need a well-structured visualization to show all the attributes without having to show everything at the same time but based on user needs. So the root is the types of restaurants, and the sub-trees are locations and restaurants names. For each restaurant, we will show a feature of the score trend across the time.

**Prototype 3:**

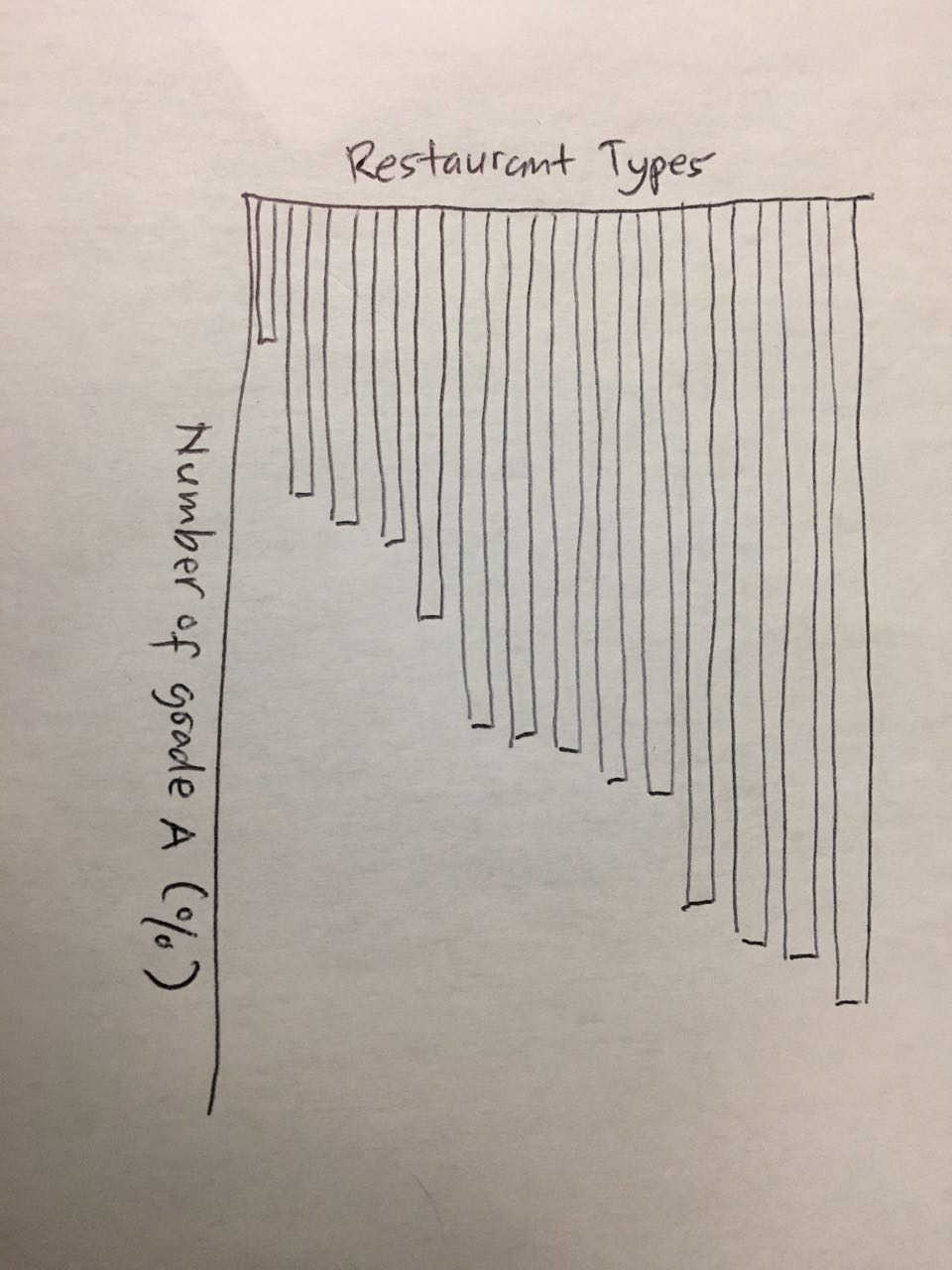
In this design, we used DNA chart to visualize the grade of each restaurant within a given street during each specific time period. We use the color gradient to represent the different grades, the length of bucket to stand for duration of a time period of each grade. Each gene bar to represent each restaurant belongs to a given street. The drawback of this visualization is, if we have too many restaurants, it will be hard to clearly see the grade of every single restaurant.

**The final prototype design with details:**

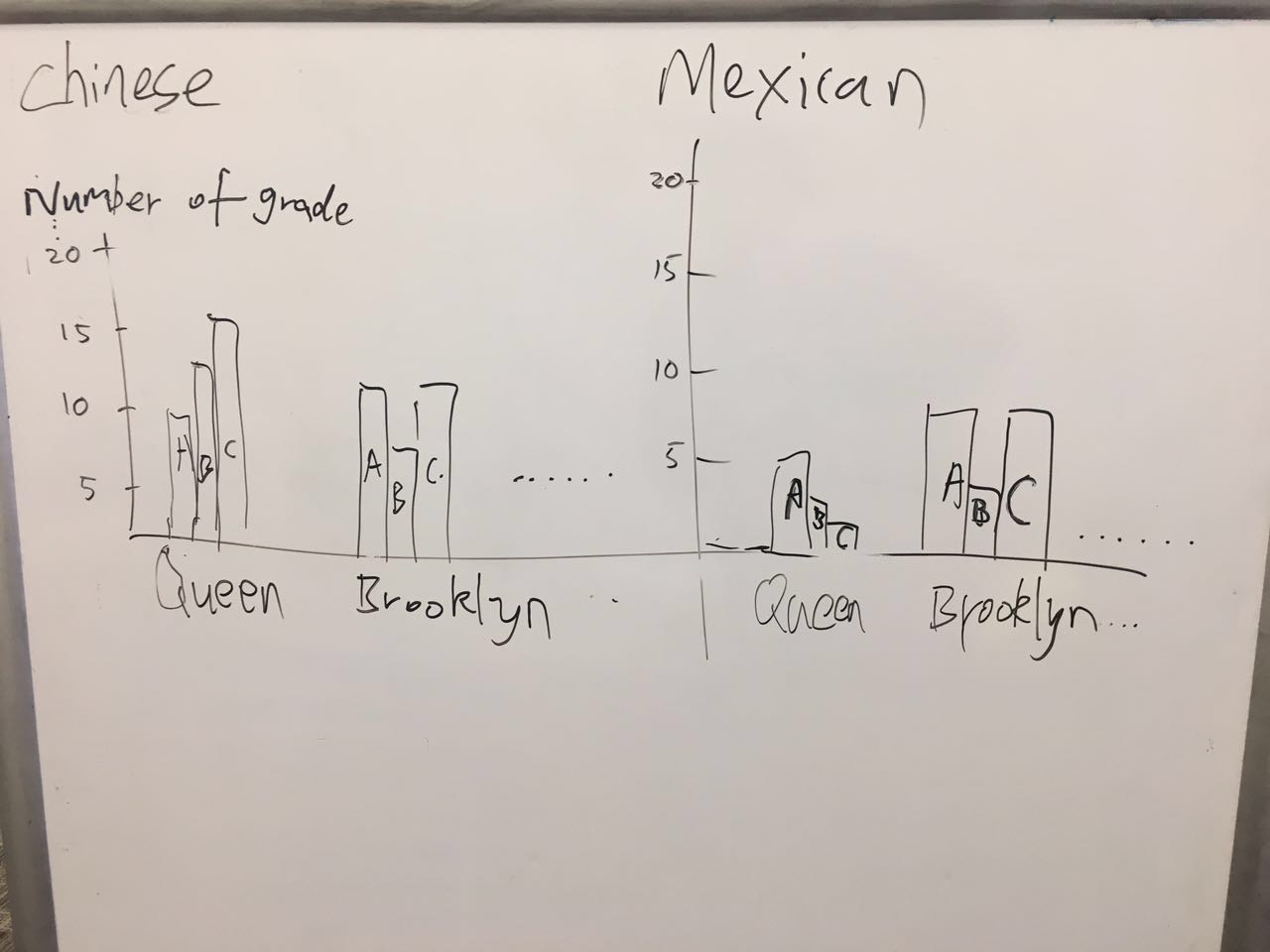
After reviewing all the three prototype designs carefully, there are some weaknesses and inefficiency that we found on the designs. Therefore, we came up with a revised design that combines some effective things from every design.



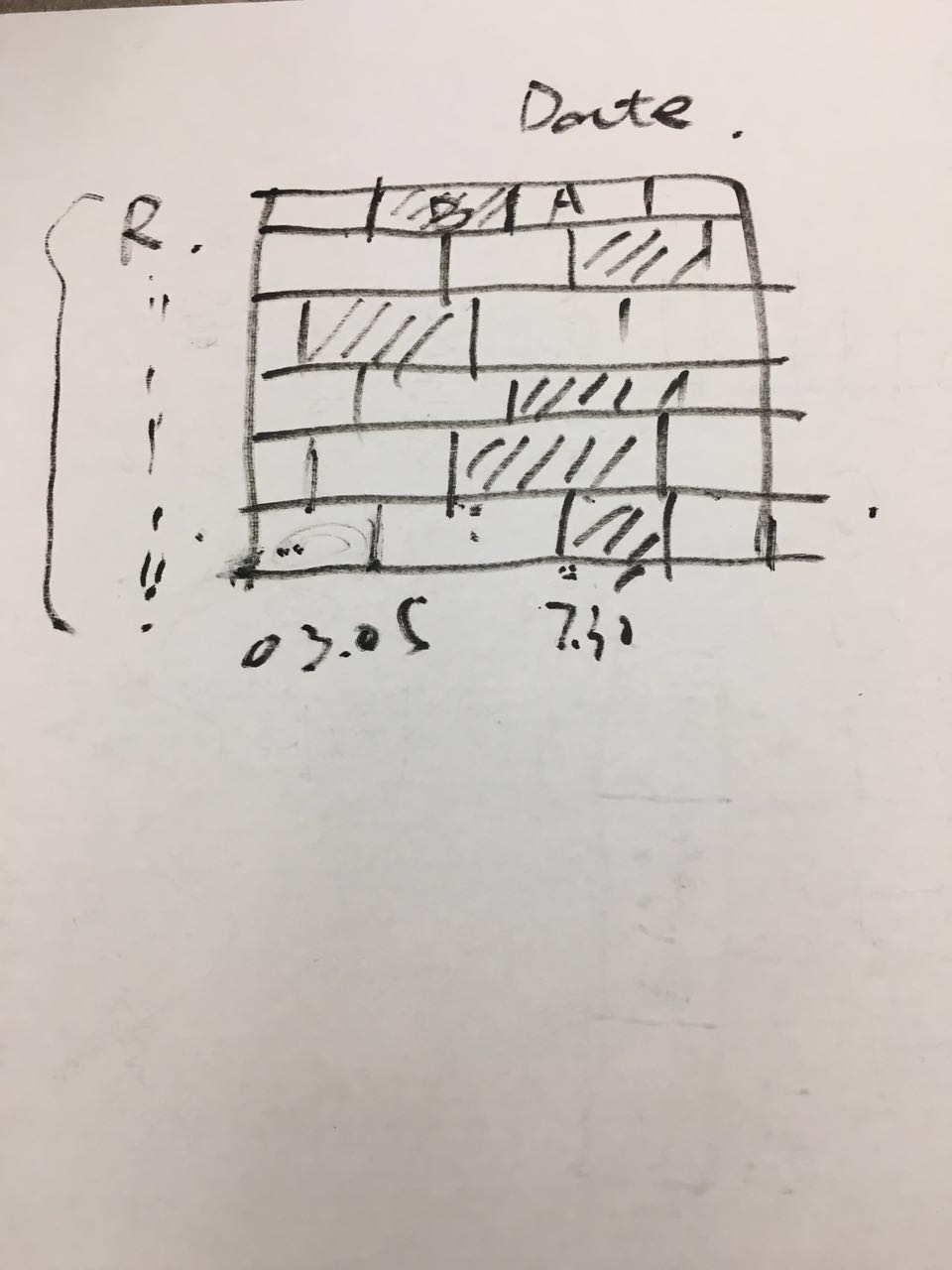
The root of the tree represents all the restaurants in NYC. In the first level of branch of the tree, each child represents a type of restaurants, like Chinese restaurant, Mexican restaurant, etc. The children of the first level branch stand for different boroughs that belong to the given types of restaurant. Then the children of the second level branch represent the different restaurant names that belong to given boroughs. In addition, we will display line charts to show the trend of scores of each inspection of a specific restaurant in each time the inspection was performed. Here, users, especially DOHMH as the inspection party, can see the progress of every restaurant has done so that they can analyze how much the inspections affect the food safety improvement made by the restaurants. This is important for DOHMH for further analysis and actions.



The second visualization is a horizontal bar chart showing the rank of every restaurant type based on the number of A grade that it has in percentage. We chose horizontal bar chart for this case because it is one of the easiest way for people to see the different gaps among the data. Thus, users can easily find which restaurant type has the highest rate of food safety, which restaurant type is the worst, or which are the top 10 best restaurant types in food safety.



In this visualization, we will use bar charts to show the grades summary of different types of restaurants by locations. Here, users can see in which location/borough a restaurant type has made the best performance or better progress in food safety practices. When users mouse over a selection bar, the tooltip will appear to show the type of the restaurants to be selected. Once a restaurant type is selected, the corresponding clustered bar chart will pop up to show the grade summary of this type of restaurant in different boroughs. The y axis is the number of each grade (A, B, and C) achieved by the restaurants. A is the highest grade indicating an excellent food safety practices and C is the lowest. The x axis will change according to which restaurant type is selected by a user.



The last visualization we will create is a DNA chart. The x position presents the date when the grade was made, while y position represents the different restaurants within a street. In each restaurant bar, we will use the different colors to indicate the grade of this restaurant in a period of time. This chart will help users to find out the restaurants with lower quality of food safety from the colors. This should be helpful for food safety supervisors and analysts in monitoring and evaluating food safety quality in NYC. Additionally, the chart can help users to identify in which period of time the business got lower grades.

**MUST-HAVE FEATURES**

1. A tree graph to navigate to the restaurants by selecting features like types of restaurant, or the location/boroughs the restaurants belong to.
2. Clustered bar charts to visualize the grade aggregations of restaurant based on the types of restaurant and the boroughs the restaurants belong to.
3. A chart to show the restaurant grade trend.

**OPTIONAL FEATURES**

1. Showing the grade rank of all restaurants.
2. Letting users select any two different restaurants and compare them based on the grades.

**PROJECT SCHEDULE**

|  |  |
| --- | --- |
| **Time** | **Activity** |
| 11/08/2016 - 11/13/2016 | Finish the data collection and preprocessing. |
| 11/14/2016 - 11/20/2016 | Complete the clustered bar chart of grade summary of different types of restaurants. Complete the DNA chart to show all the restaurants’ grade trend on a street. |
| 11/21/2016 - 11/29/2016 | Complete the visualization of tree chart showing every specific restaurant and the lice chart for the score trends. |
| 11/30/2016 - 12/08/2016 | Working on optional features. |
| 12/09/2016 - 12/12/2016 | Working on the documentation and video. |

**REFERENCE**  
<https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>

<http://www1.nyc.gov/assets/doh/downloads/pdf/rii/restaurant-grading-18-month-report.pdf>