

**Food Inspections in the City of Chicago Data Analysis using IBM Watson Analytics**

Bhagyashree Arvind Jadhav

California State University, Los Angeles

1. **Data sets**

URL of dataset:

<https://www.healthdata.gov/dataset/food-inspections>

Twitter Hashtags:

following tags are used to analyze twitter data:

1.FoodInspection

2.Foodsafety

3.Healthyfood

1. **Data Description**

The dataset contains information for the restaurants and other food establishments in Chicago. There are around 400 different types of facilities in the city. (Chicago, 2017) Every facility is categorized by its risk of affecting the public’s health. High risk is showed by 1, medium risk is showed by 2 and low risk has been showed by 3. The number of inspection in each area or facility are dependent on the risk. If the area is under high risk then frequent food inspections happen in that area. (Health, 2010) Violation type is mentioned for all the inspections which are failed. A facility can receive up to 45 distinct violations. Street address, city, state and zip code of each facility will help to display analyzed data on the geo map.

Along with this information Inspection Type, Inspection date and results of the inspection are also provided in the dataset. Every inspection receives result after inspection is done. The result can be fail, no entry, not ready, out of business, pass or pass with conditions. Facilities getting a ‘pass’ did not have critical violations. Businesses receiving a ‘pass with conditions’ were found to have critical or serious violations, but these violations got corrected during inspection. If facility receives a ‘fail’ means there were critical violations which could not be corrected during inspection. ‘Doing business as’ and ‘Also know as’ are the two fields which provide legal names of the facilities. (Health, 2010) License number mentioned in the dataset is assigned by the department of Business Affairs and Consumer Protection.

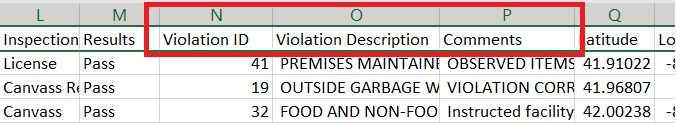
For each entry in the dataset facility type given which provides more information about the establishment description. There are different types of inspections performed based on the facility type. ‘Canvass’ is the most common type of inspection performed based on the risk of the facility. If the inspection is performed on the request of owner before opening the establishment then it is called consultation. The dataset provides all this necessary information from year 2010 to 2017 which can be used to find different trends from this data.

1. **Data Refinement**
2. Data refinement 4 examples
3. Data set has a column for violations in food inspections. The column contains violation number, violation description and comments provided by food inspector. Separated this column into 3 different columns in excel to work on violations in further data analysis.

Before Cleaning:

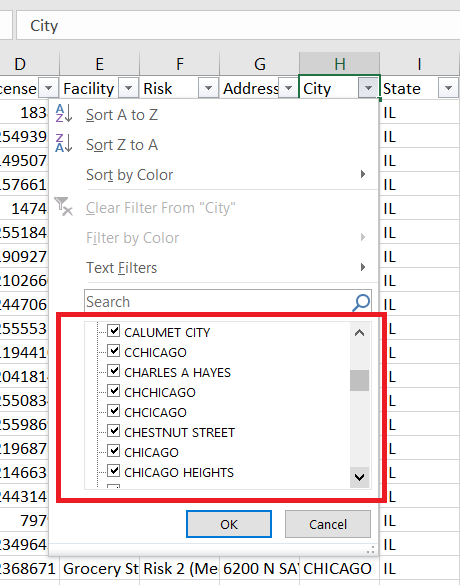
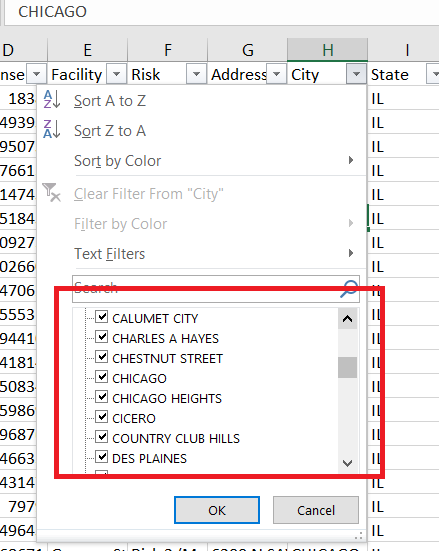


After Cleaning:



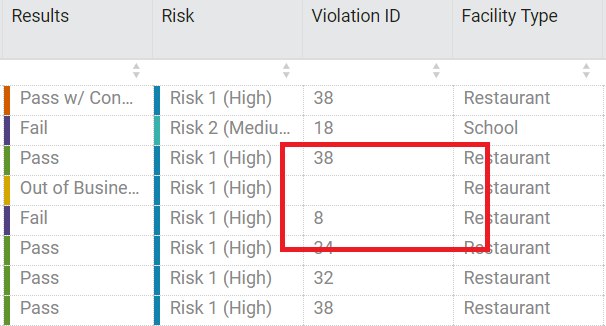
1. Misspelling: ‘Chicago’ was misspelled in city name column. Replaced all the values for Chicago city with correct spelling.

Before cleaning: After Cleaning

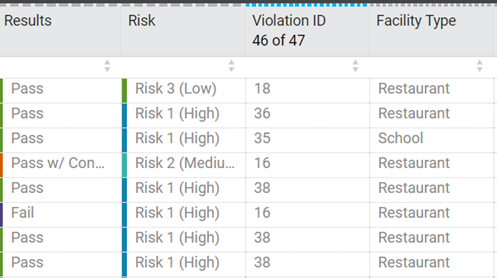
 

1. Missing Values: Removed missing values from Violation ID field

Before cleaning:



After cleaning:

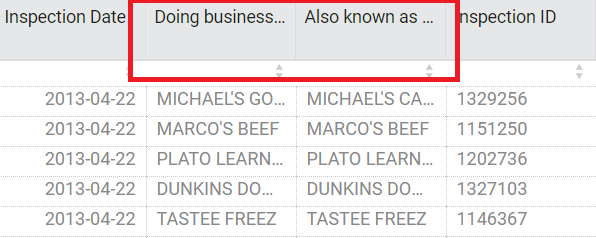


1. Data contained abbreviations for the ‘DBA Name’ and ‘AKA Name’ fields. Renamed both the columns to ‘Doing business as name’ and ‘Also known as name’ for better understanding of data and visualizations.

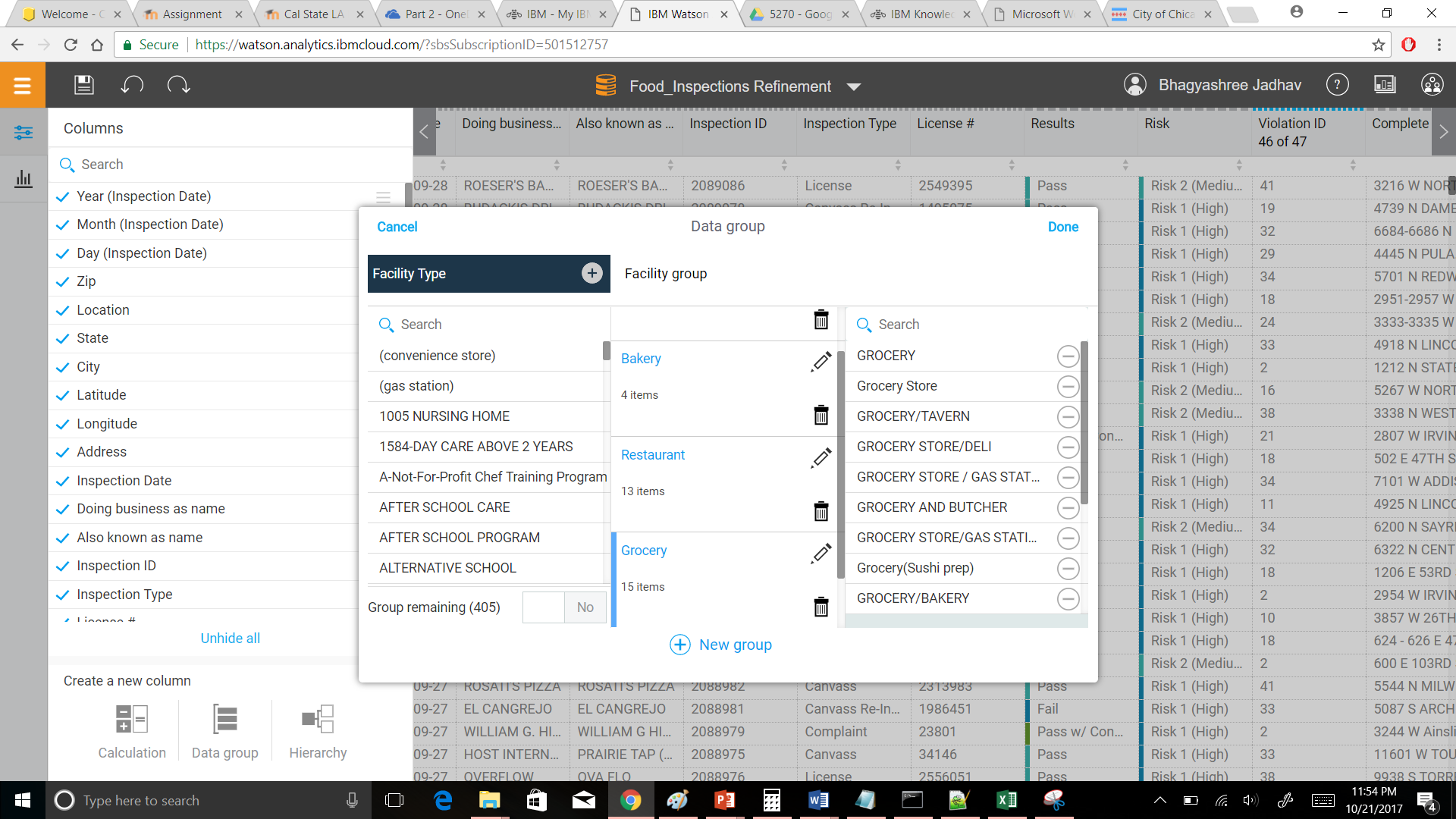
Before cleaning:

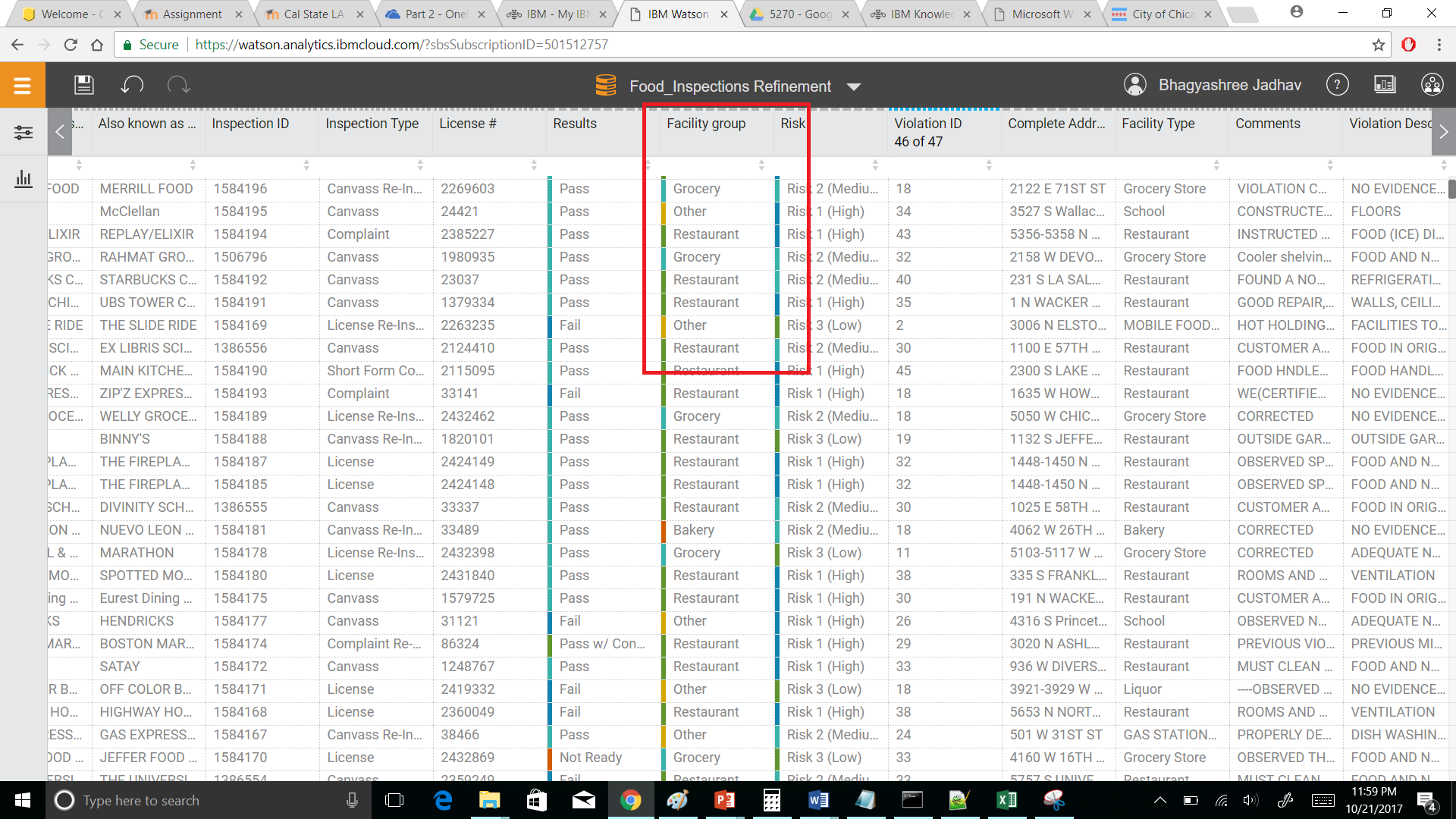


After cleaning:

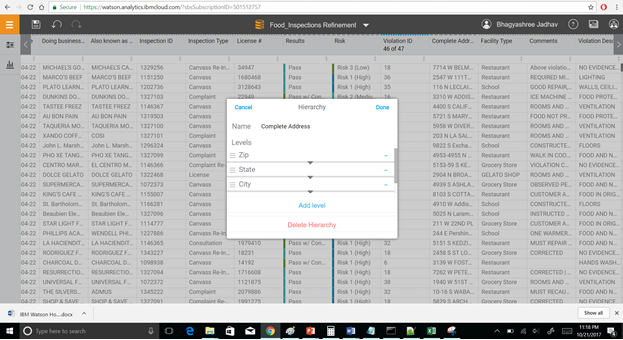


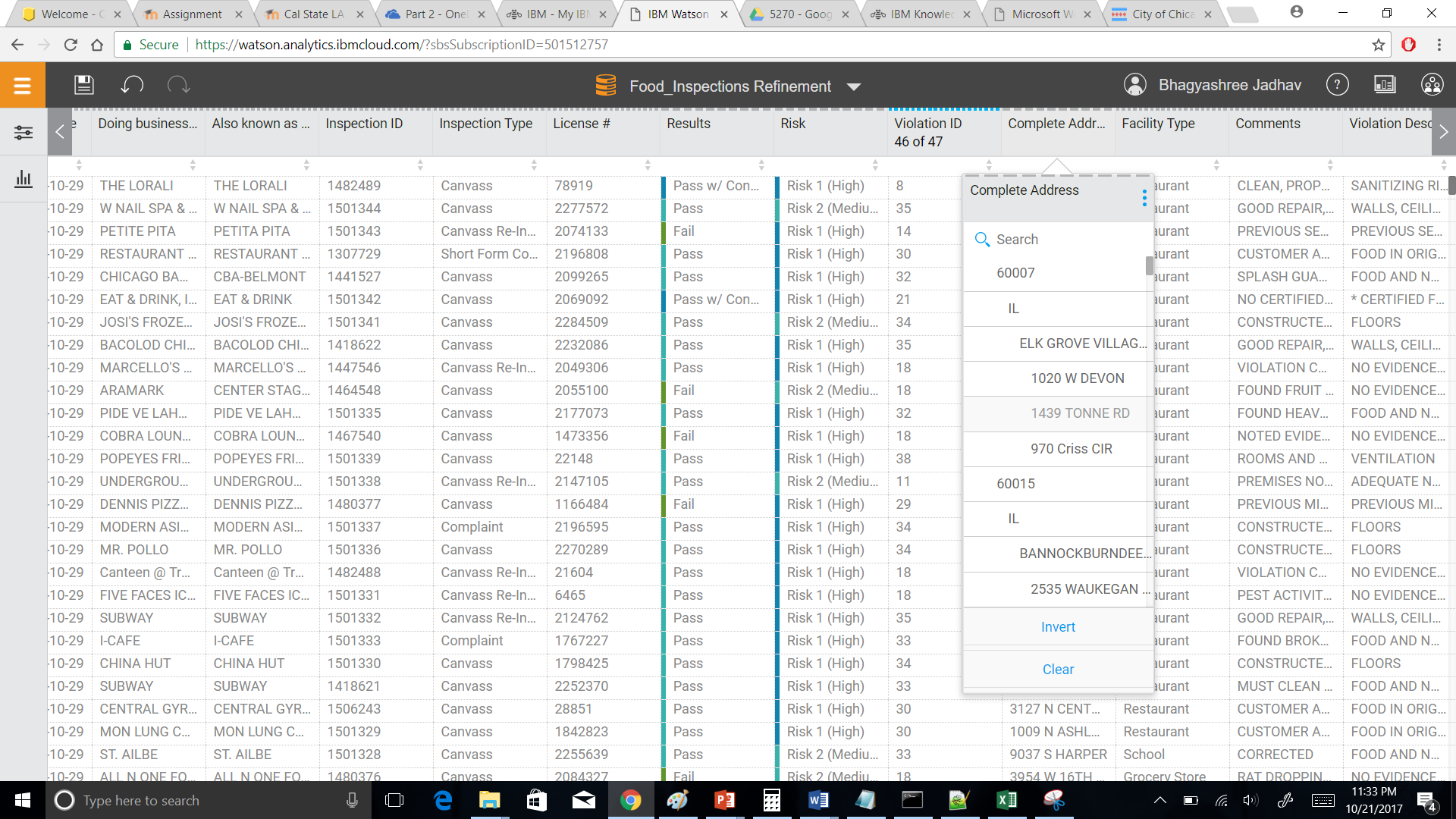
1. Create data grouping and data hierarchy (Note: Could not create calculated field with this dataset and had informed about the same)
2. Data grouping: There are different facility types in this data. Grouped this data under Children’s Service Facility, Grocery, Restaurants and other categories. This grouped data further used for the analysis of restaurants data.





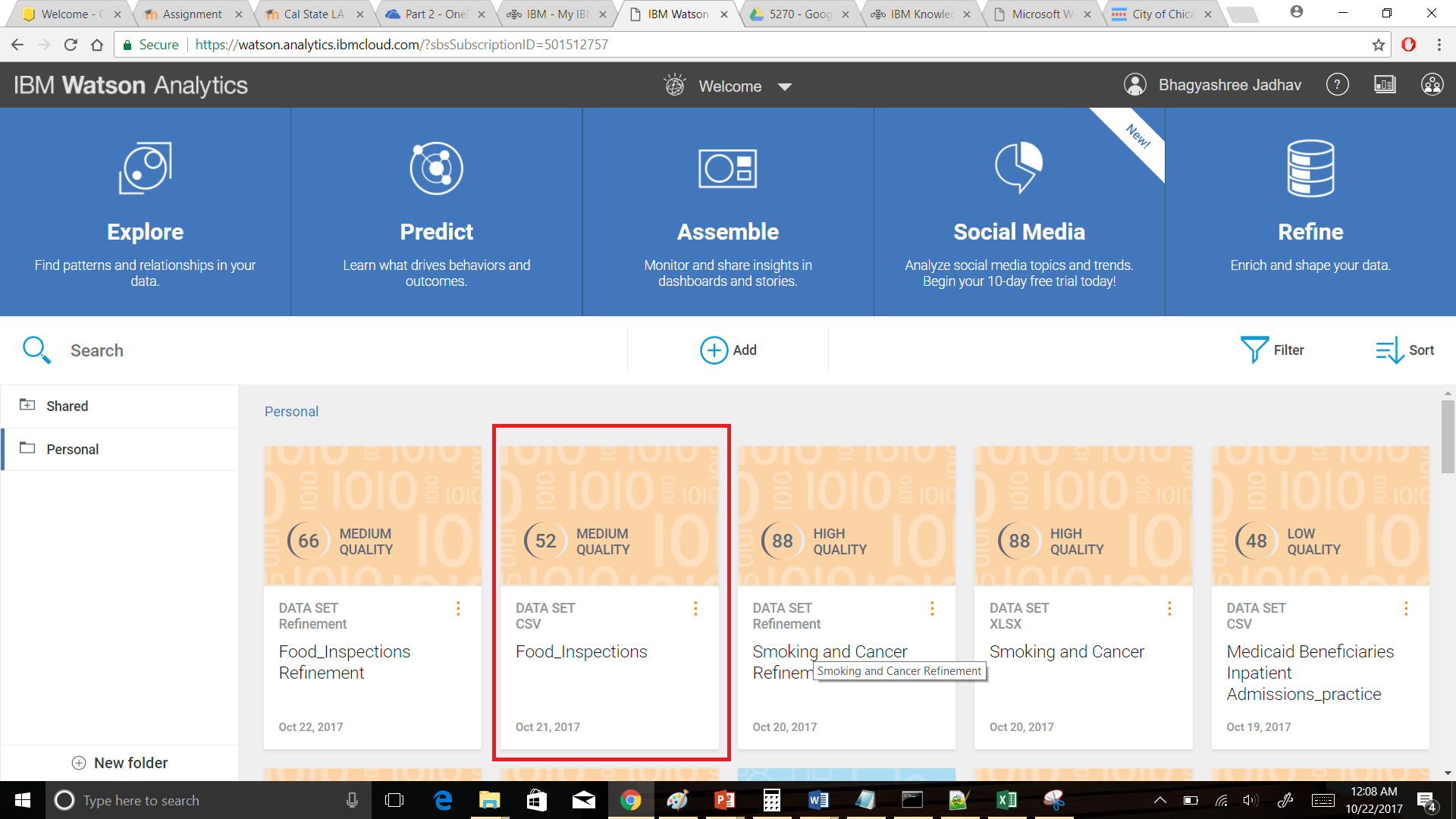
1. Data Hierarchy: Data set has zip, state, city and address field. Created hierarchy from this data to create a complete address field from it.



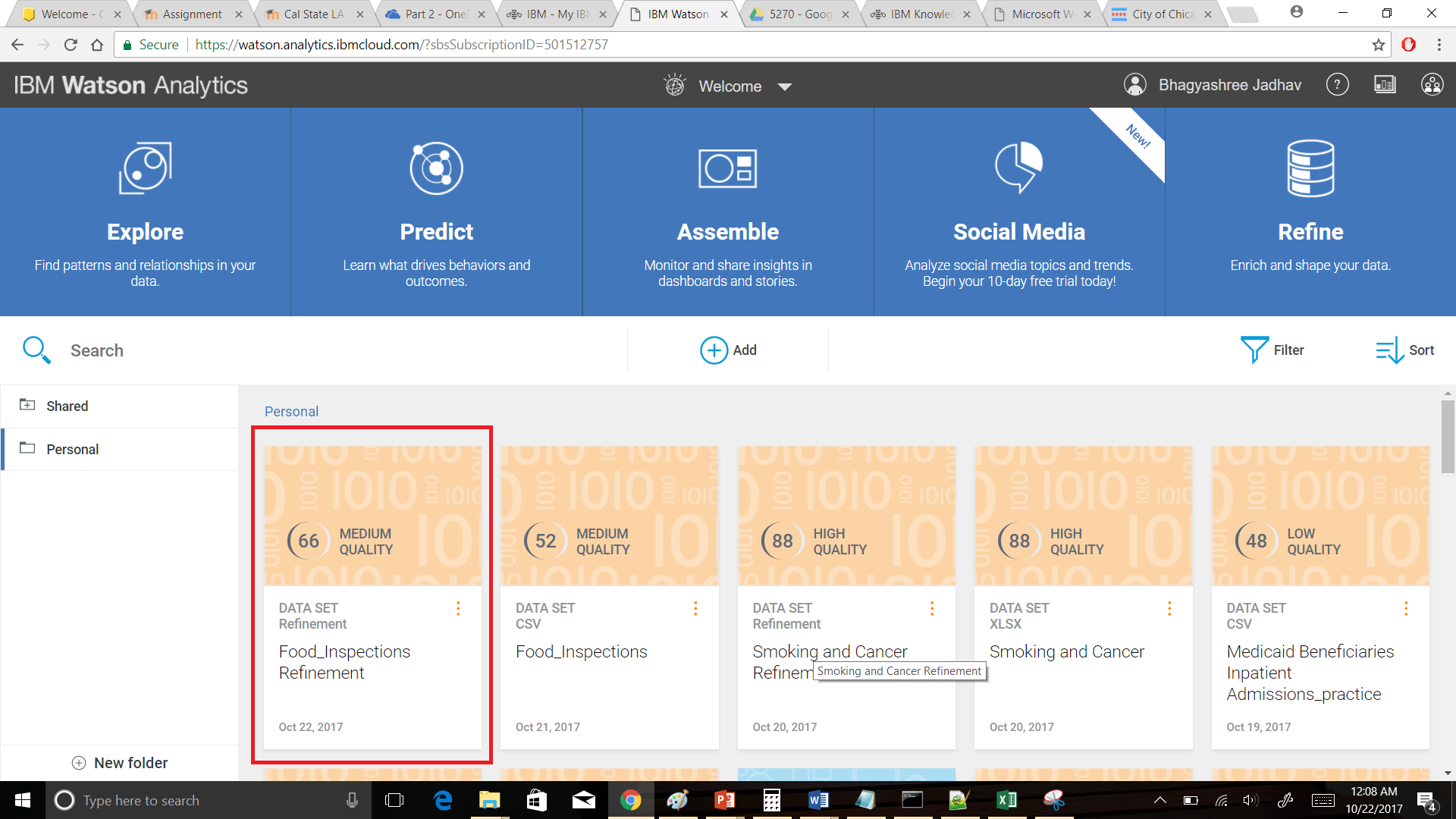


1. **Data Quality**

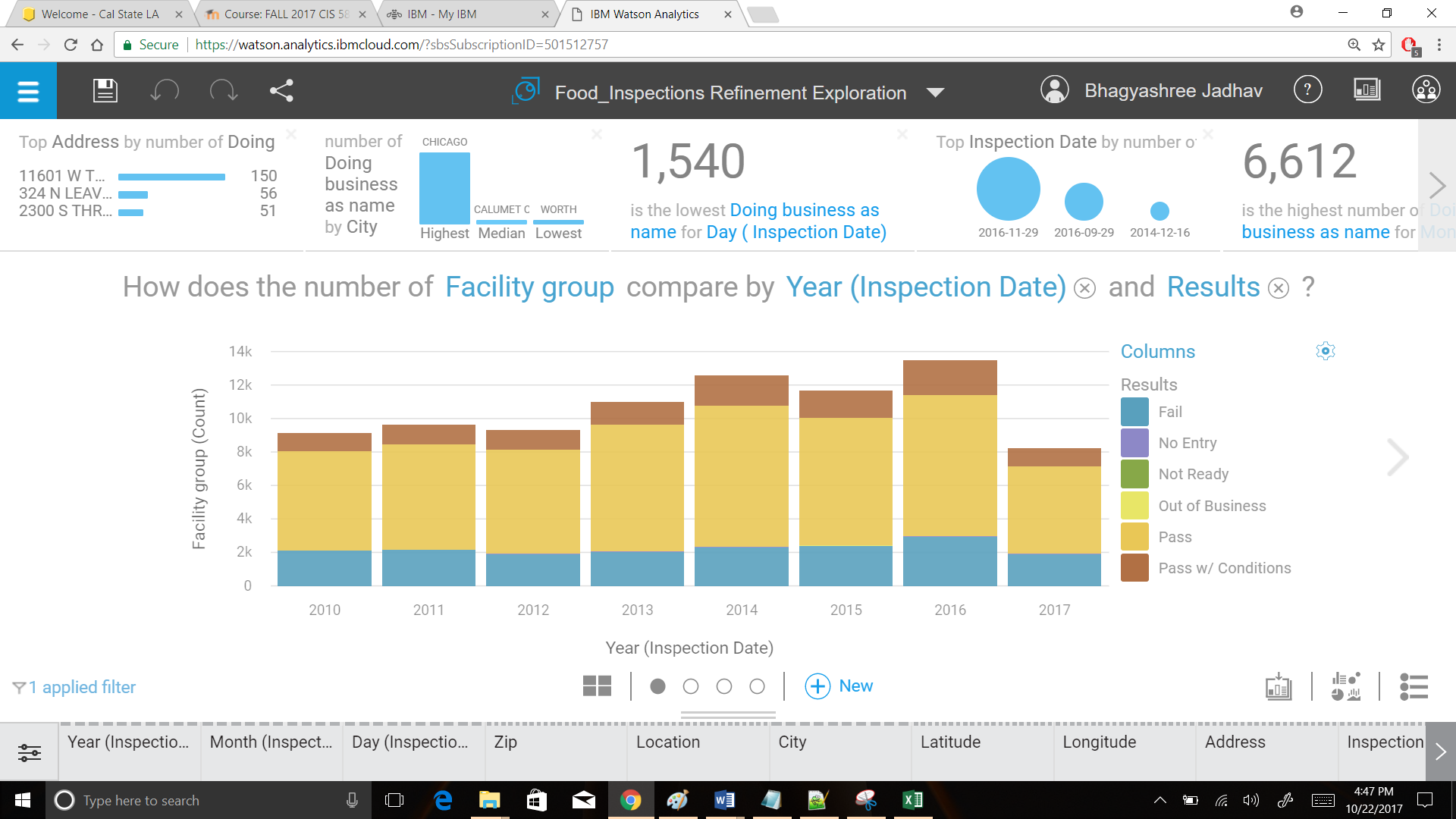
Original Dataset Quality

****

Refined dataset quality

****

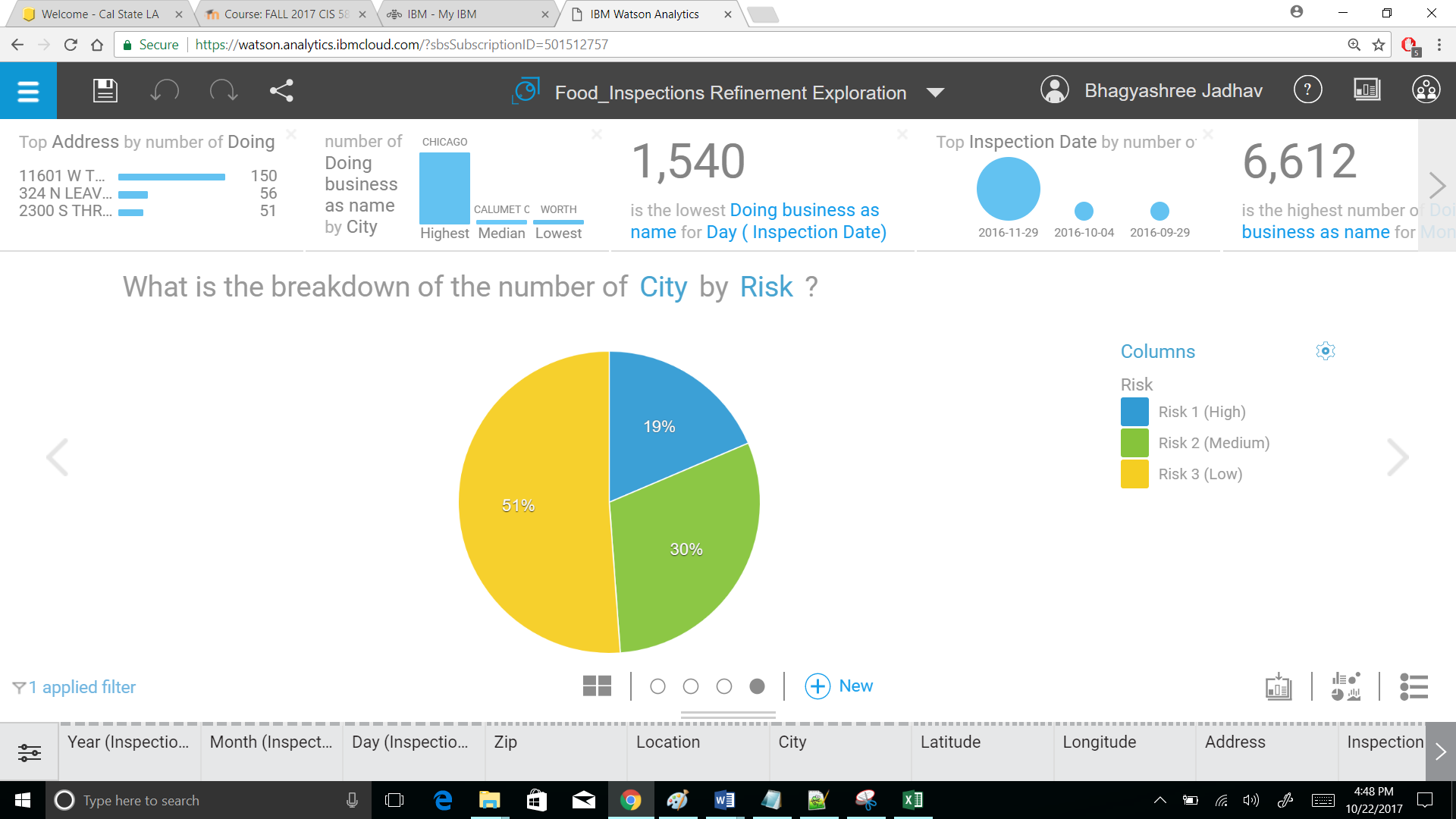
1. **Data Exploration**
2. **How does the number of Facility group compare by Year (Inspection Date) and Results?**



[Tools used: Stacked bar chart with filter on Restaurant Group]

The above stacked bar chart provides yearly count of inspection results for restaurants across the city of Chicago. It shows how many restaurants have failed and passed health or food inspections from 2010 to 2017. The data has been filtered for a specific Facility group which provides all the information for restaurants only. In 2016 maximum (2921) number of restaurants have failed food inspections. From 2012 to 2016 number of failures are increasing every year but the count has decreased in 2017. Number of failure results should decrease eventually as most of the people have food in restaurants on daily basis.

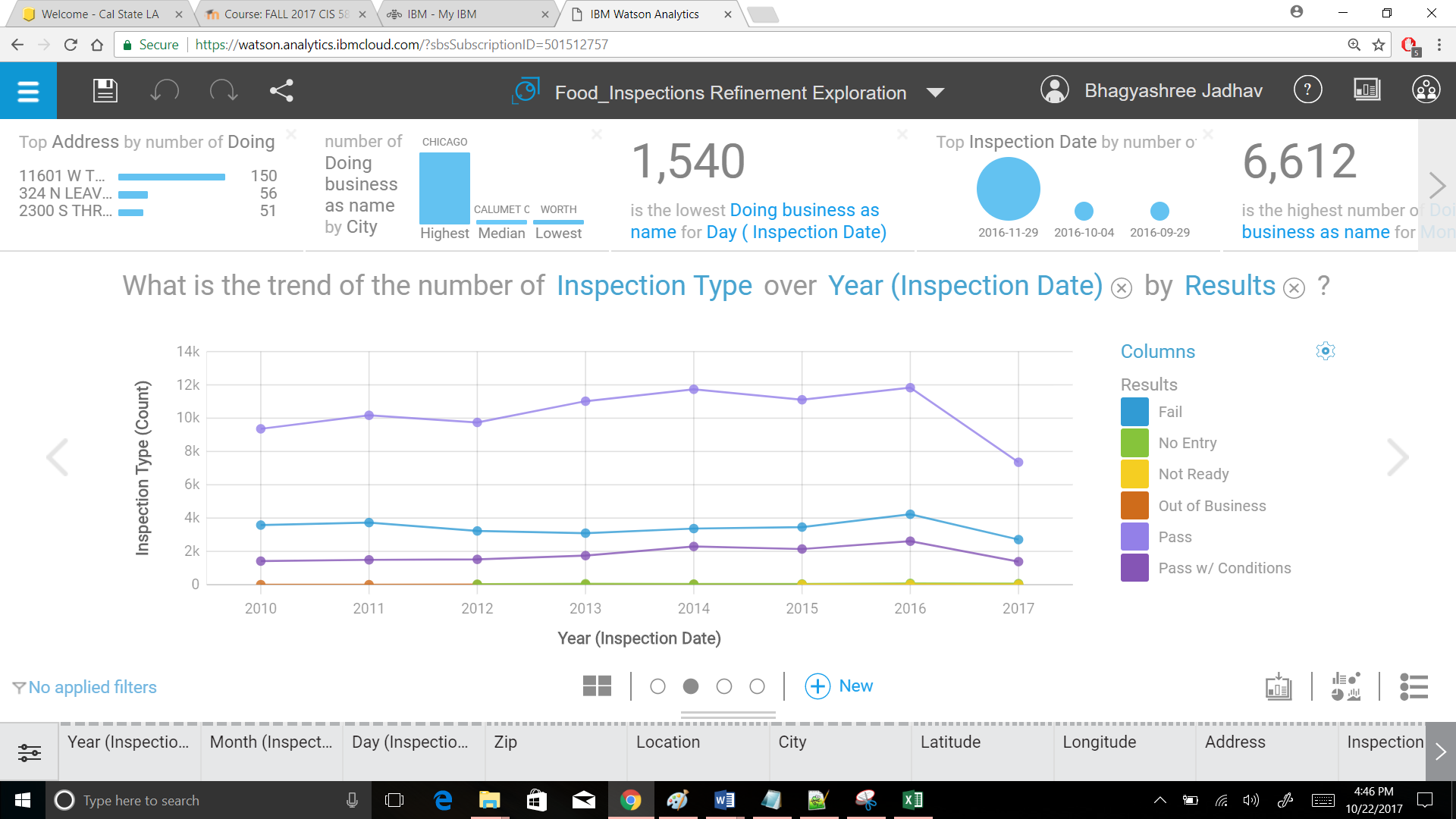
1. **What is the breakdown of the number of City by Risk?**



[Tools used: Pie Chart]

This pie chart tells us how many cities are under which level of food inspection risk. 51% of cities are under low risk, 30% cities are under medium risk and 19% cities are under high risk. Even though the number of cities under high risk are less the high risked restaurants or grocery stores can cause serious medical problems in society. Further analysis can provide specific cities which are under high risk and food inspectors can focus on those cities.

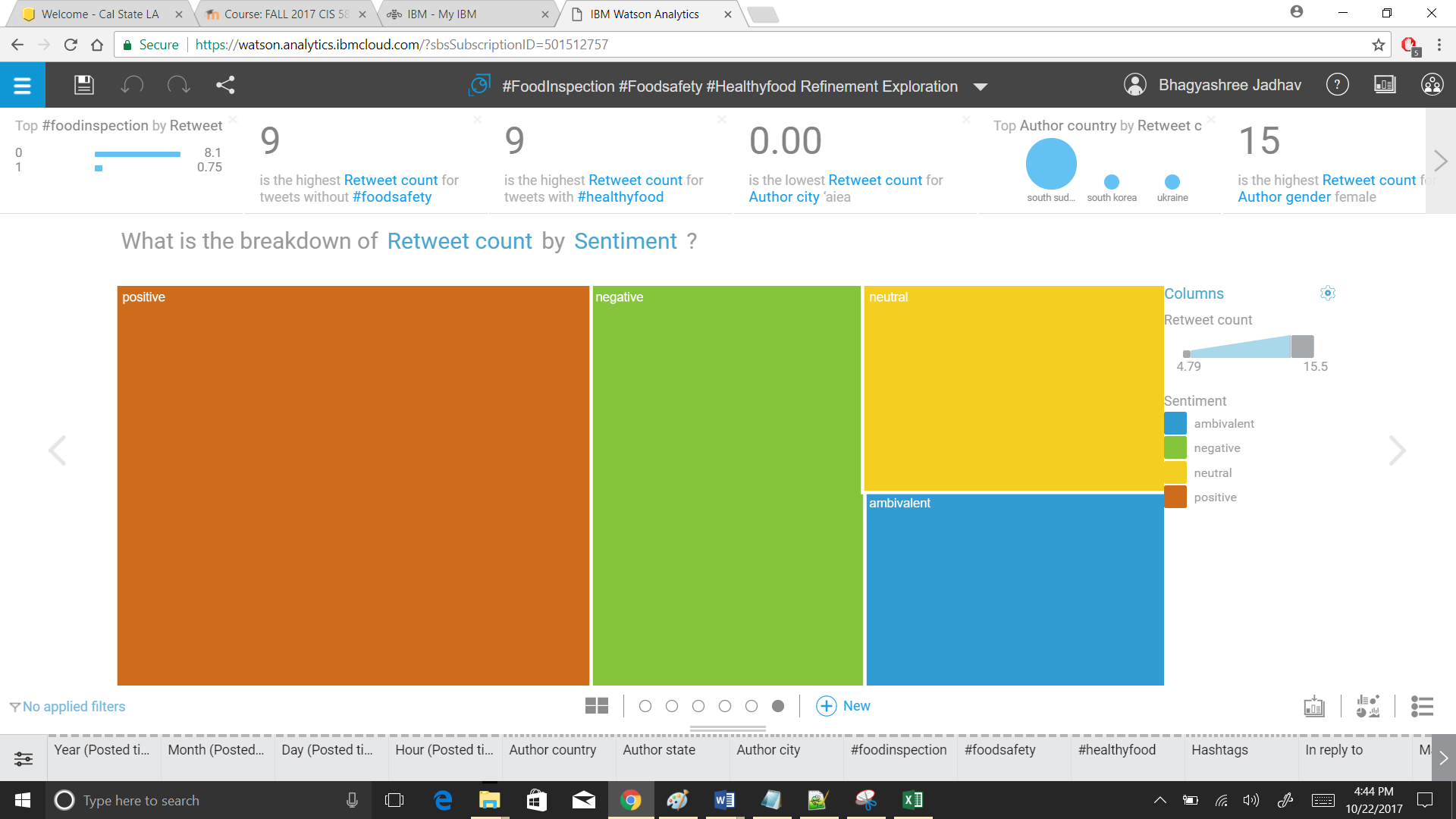
1. **What is the trend of the number of Inspection Type over Year (Inspection Date) by Results?**



[Tools used: Trend Line]

Data provides yearly information about type of inspections and their corresponding results. These trend lines compare all the results on the basis of count of inspection type. From 2013 to 2016 the number of failed inspections are increasing but it has started decreasing in 2017. From the graph it can be seen that in 2012 and 2015 number of restaurants passing food inspections have decreased but in the subsequent years the numbers have increased which ca be considered as the impact of previous inspection.

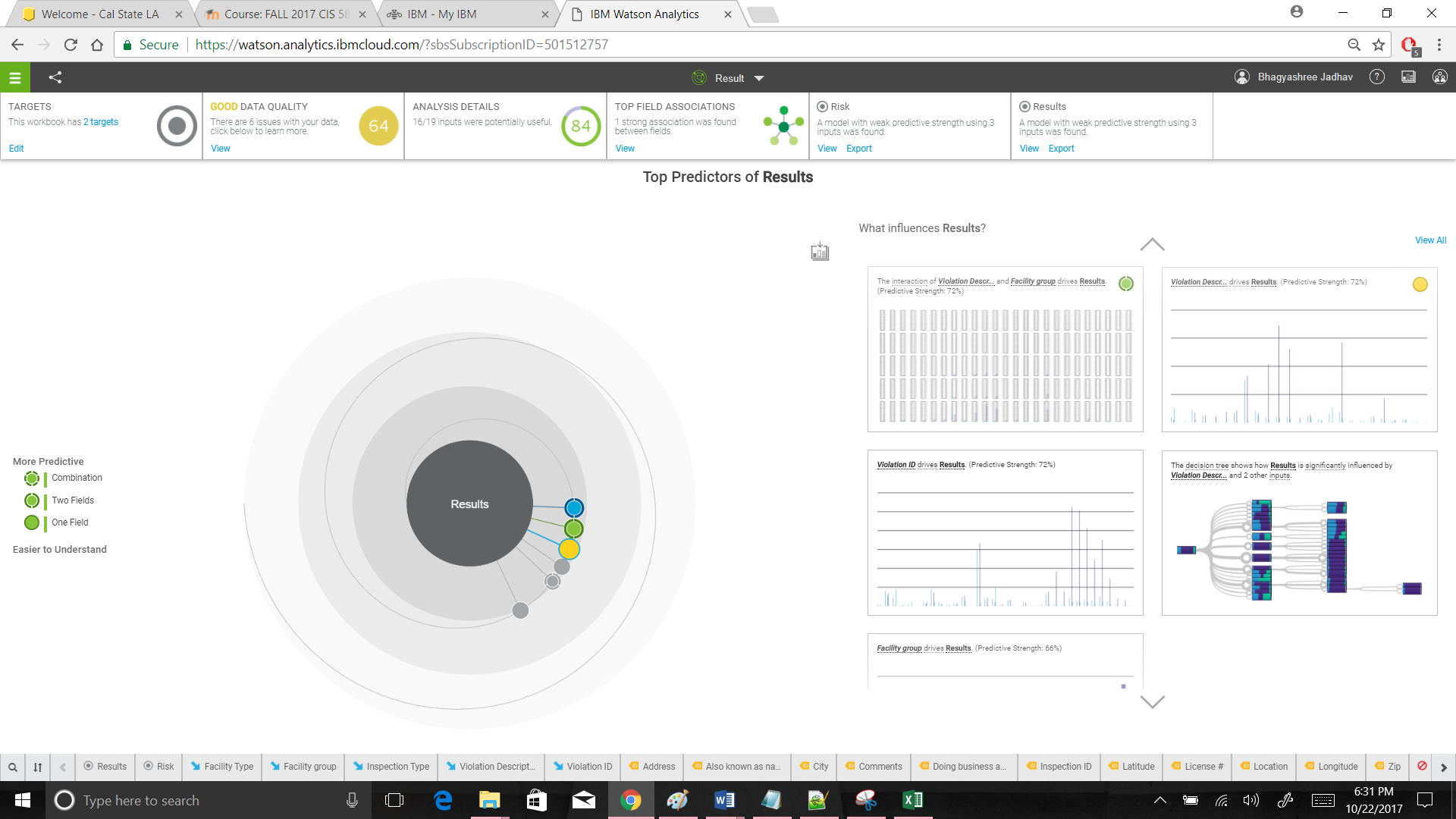
1. **What is the breakdown of Retweet count by sentiment?**

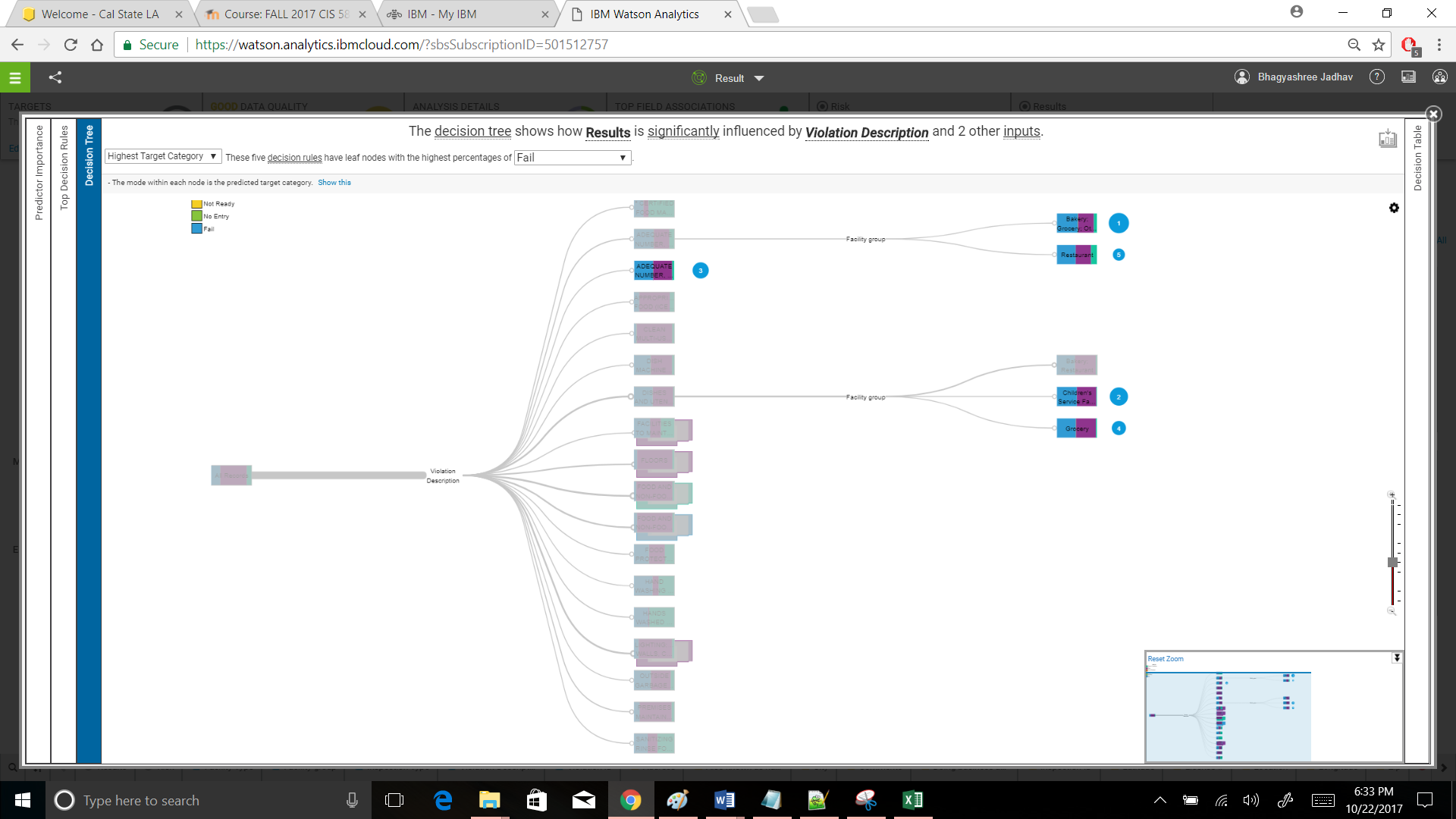


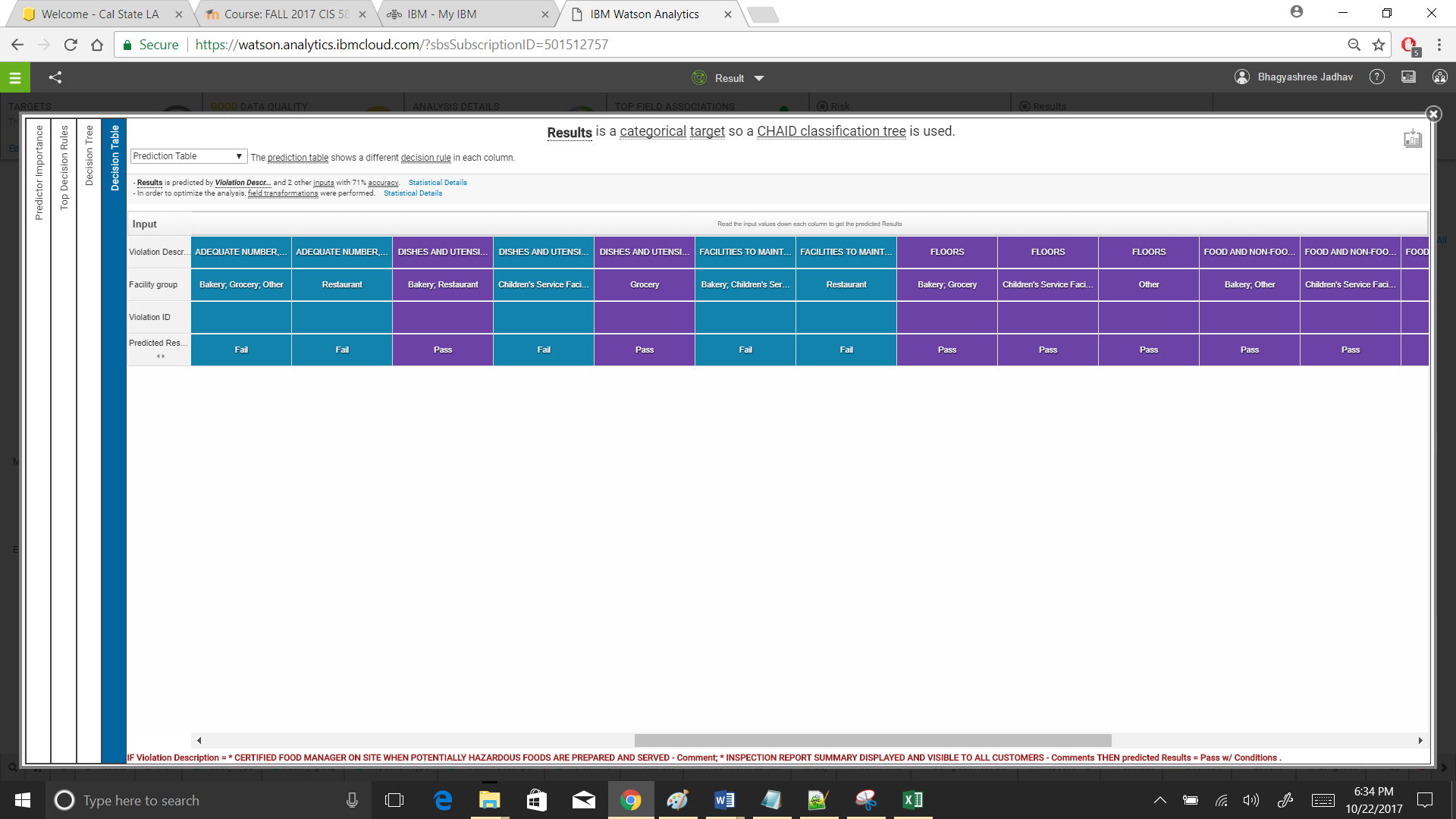
[Tools used: Treemap, Tweeter hashtags]

The above treemap shows the retweet count by sentiment. There are four types of sentiments positive, negative, neutral and ambivalent. The retweet count of positive sentiments is more as compared to other sentiments. From the treemap it can be concluded that people retweet the posts related to food inspections with positive sentiment. Tweeter can be used to reach out to society and business owners to make them understand the importance of food safety and hygienic food.

1. **Prediction**





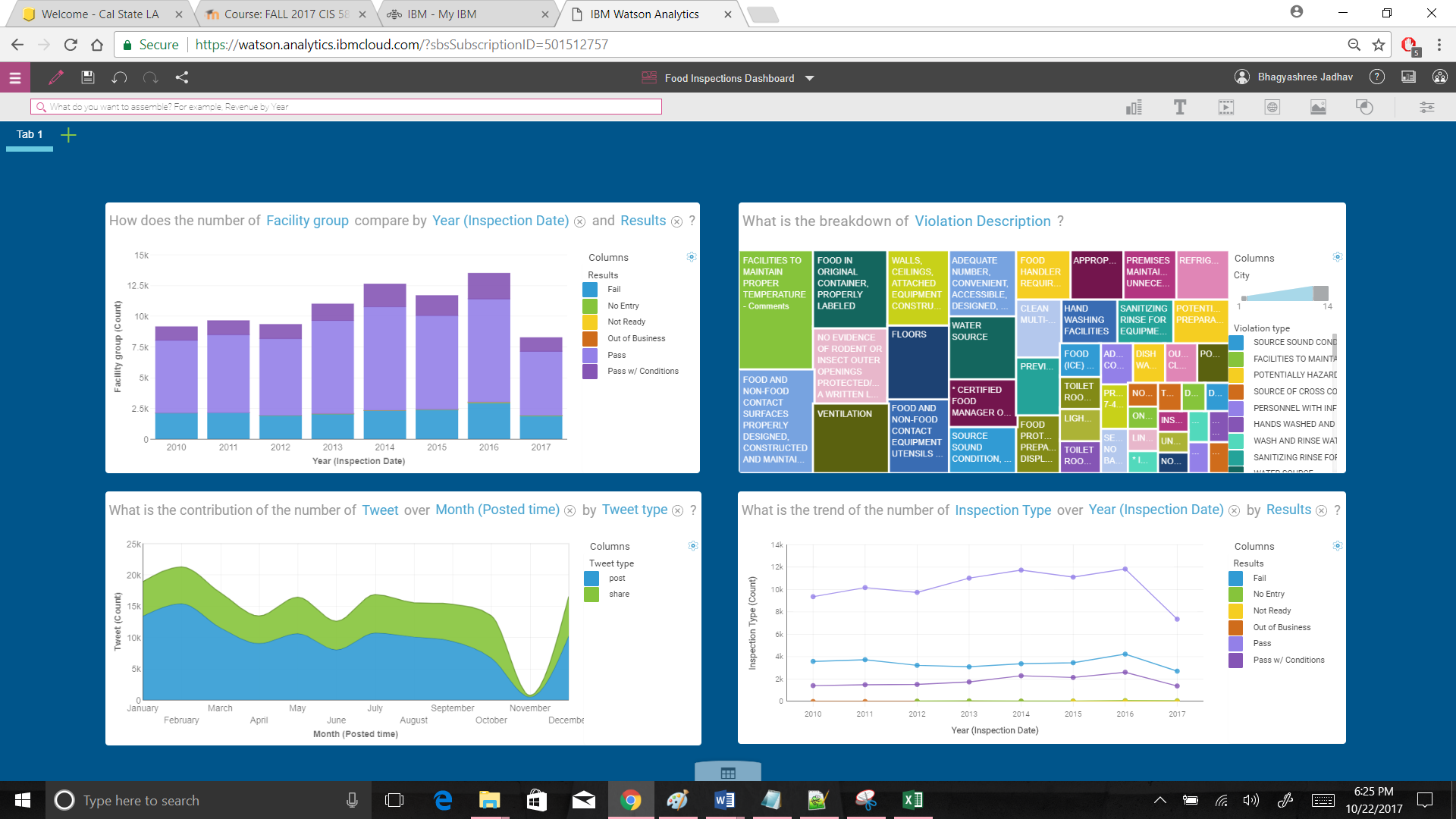


Top predictions of ‘Results’ are mainly driven by Violation Description, Facility Group and Violation ID. The prediction strength is 72%. There are six types of results for the food inspections and violations are associated with the establishments which failed the food inspection. All the results are affected by Violation Description. The decision tree shows that dishes and utensils flushed, scraped and soaked is a violation description which has highest number of failures.

Depending on the violation description the decision tree is further divided into three main facility groups which are Bakery Restaurant, Children’s service facility and Grocery. Children’s Service Facility leads to the second highest percentages of the failed results. As ‘Results’ is a categorical target, CHAID Classification tree is used for showing analyzed data. Decision table contains input values and predicted results for each column. In order to optimize the analysis field transformation is performed in decision table. The main insights of this prediction show that the interaction of Violation Description and Facility Group drives Results.

Violation ID also drives Results. Violation having description as food and non-food contact equipment utensils clean, Free of abrasive detergents has 90% of passed records. Violation number 18 has highest number of (46%) failed inspections whereas violation number 32 has highest number of passed inspections. Violation number 33 and 34 has also got significantly higher number of passed inspections. The overall prediction shows that Bakery, Restaurants, Children’s service facility and Grocery are major facility types which might fail the inspections.

1. **Dashboard**



**Dashboard Summary:**

The above food inspections dashboard provides the summary of the data analysis done. The stacked bar graph shows results of inspections for restaurants in the city of Chicago. The treemap provides information about number of cities by violation description. Highest number of cities have violation ‘Facilities to maintain proper temperature’. Trend analysis of tweeter data is done to display monthly total number of tweets on food inspection and food safety. The graph also divides data into total number of shares and posts. Another graph for trend lines compare all the results on the basis of count of inspection type from 2010 to 2017. It can be concluded from this data analysis that people have positive sentiment towards food safety but food inspection failure rate is almost increasing every year. Safe food saves life. Hence, it cannot be compromised.

# References

Chicago, City of. *Food Inspections*. 29 09 2017. dataset. 23 09 2017. <https://www.healthdata.gov/dataset/food-inspections>.

Health, Chicago Department of Public. *Food Inspection*. 2010. Web pdf. 25 09 2017. <https://data.cityofchicago.org/api/assets/BAD5301B-681A-4202-9D25-51B2CAE672FF>.