**A close up of text on a white background

Description generated with high confidence**

**CIS 5810: Healthcare Information System**

**Fall 2017**

**National Cardiovascular Disease Surveillance Data Analysis**

**using IBM Watson Analytics**

Submitted to: Dr. Shilpa Balan

**Submitted By:**

**Narendra Mali CIN:305848717**

1. **Dataset Link:**

<https://www.healthdata.gov/dataset/national-vital-statistics-system-nvss-national-cardiovascular-disease-surveillance-data-4>

**Twitter Hashtags:**

We are going to use following tags to analyze twitter data:

1. Cardiac arrest
2. Heart attack
3. Cardiovascular Diseases
4. **Data Description:**

Cardiovascular disease is also known as heart disease. Cardiovascular disease is heart conditions that include diseased vessels, structural problems, and blood clots. One in four deaths in the United States is because of heart disease and about 610,000 people die of heart disease every-year. (Services, 2017) National Vital Statistics System (NVSS) has collected cardiovascular disease related information and published it as a dataset. Dataset contains information for all the 50 states in United States and this information is categorized into 8 different indicators such as Rate of major cardiovascular disease mortality among US adults (18+), Rate of heart failure mortality among US adults (18+) etc.

Data also covers different topics such as Major Cardiovascular Disease, Heart Failure, Stroke, Heart Disease, Coronary Heart Disease and Heart Attack. This dataset includes information from the year 2000 till 2015. That is fifteen years of comprehensive data which also gives information about Confidence limit which helps us to get biological statistic. The data is organized by location (national and state) and indicator; NVSS mortality data include CVDs (e.g., heart failure). The data can be viewed by temporal trends and stratified by age group, sex, and race/ethnicity.

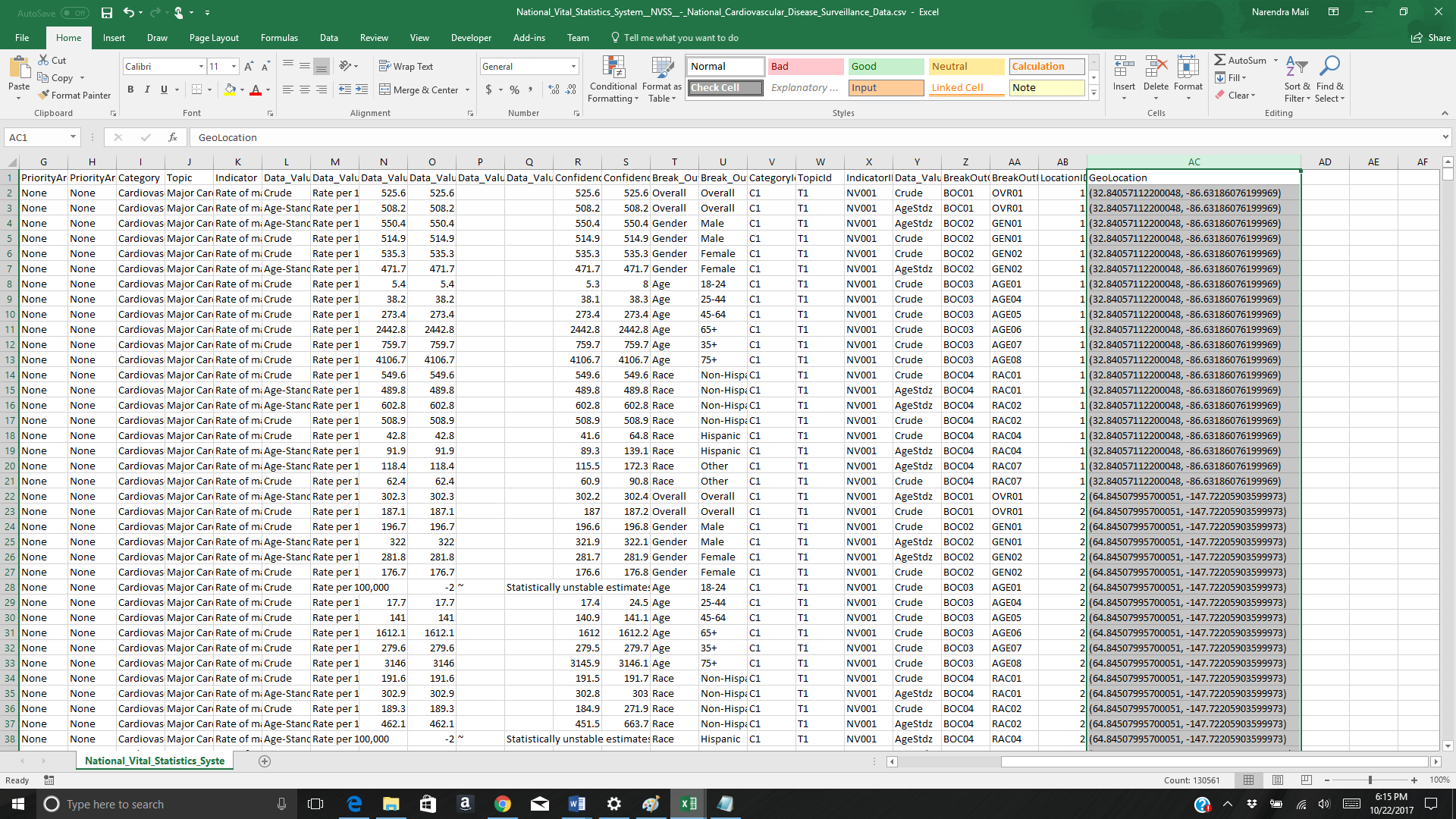
There are three twitter hashtags which will contribute to the people’s opinion in the related topics. Twitter hashtags are as follows, Heart attack, cardiovascular disease and cardiac arrest. These three hashtags are being analyzed from 1st April 2014 till 22nd October 2017. Hashtags in IBM Watson analytics also gives valuable information regarding sentiments and hence it is highly required to use such a platform for data analysis. Heart disease is the leading cause of death for both men and women. By analyzing this dataset, we will get insights of Cardiovascular Diseases across United States of America.

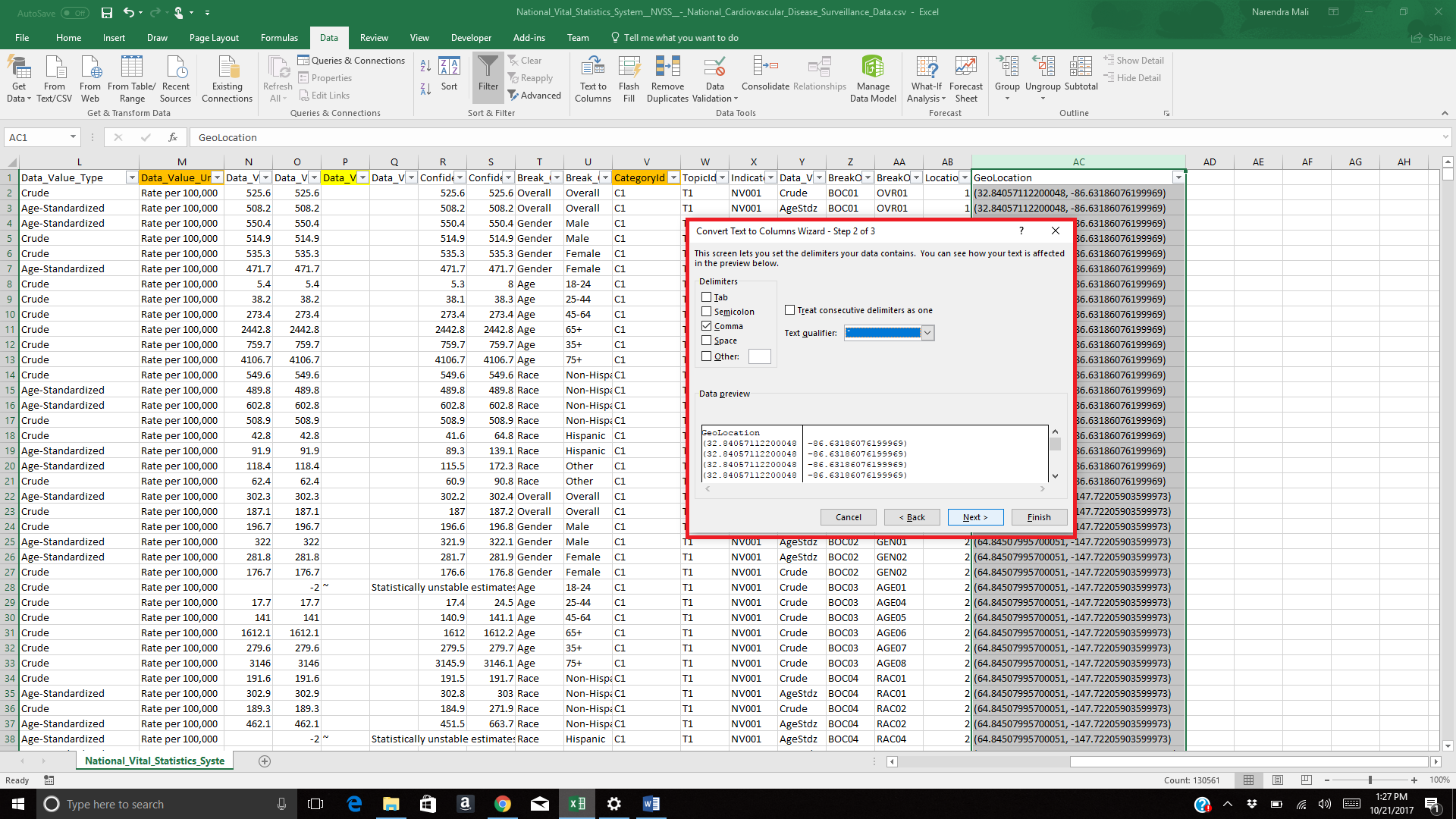
1. **Data Refinement**

**Excel:**

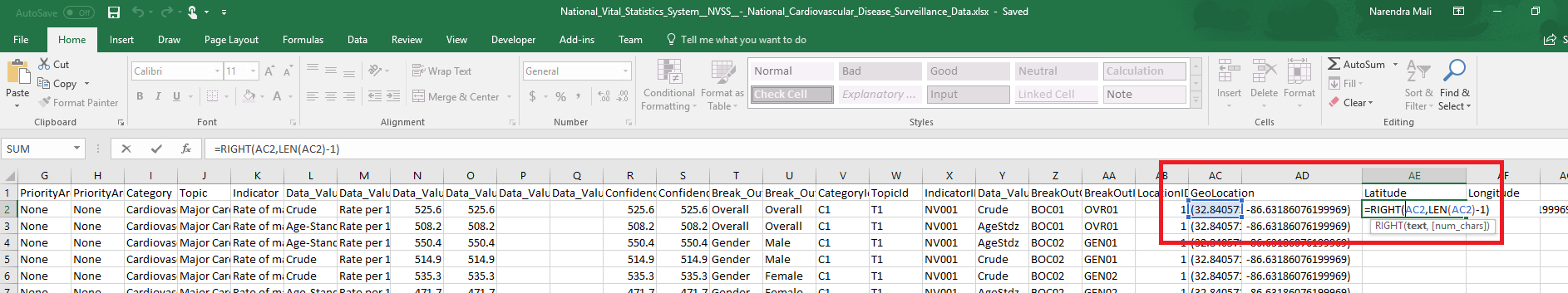
Geo Location was in the column ‘AC’ without separating latitude and longitude.

Before Refinement:

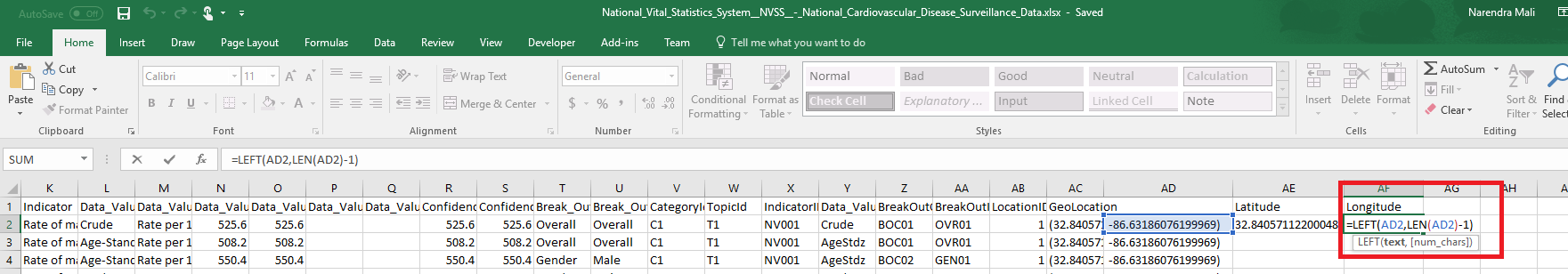


Separated two values i.e. latitude and longitude separated by comma. 

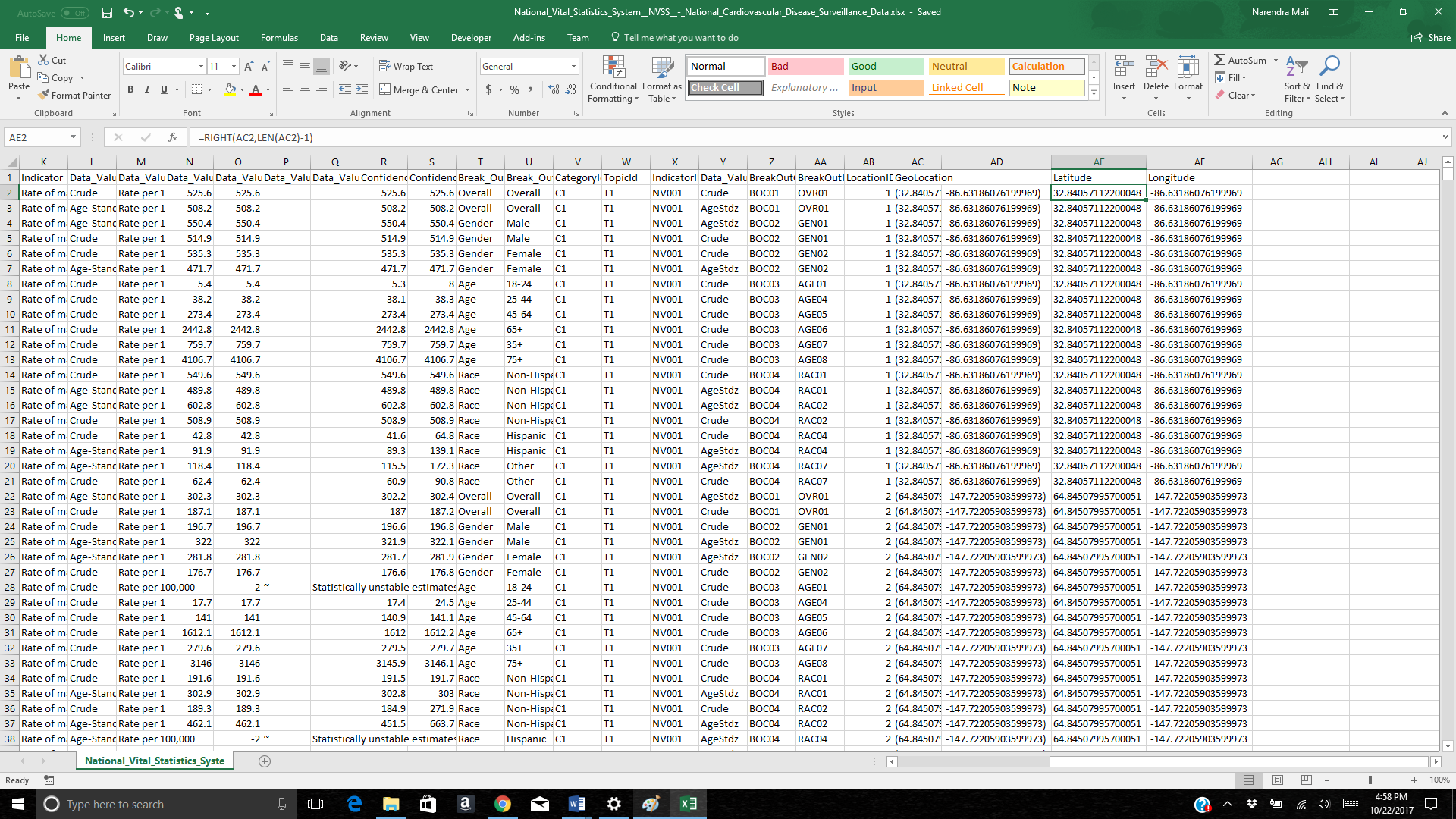
2. Adding Formula to Remove round bracket from left



Adding Formula to Remove round bracket from Right



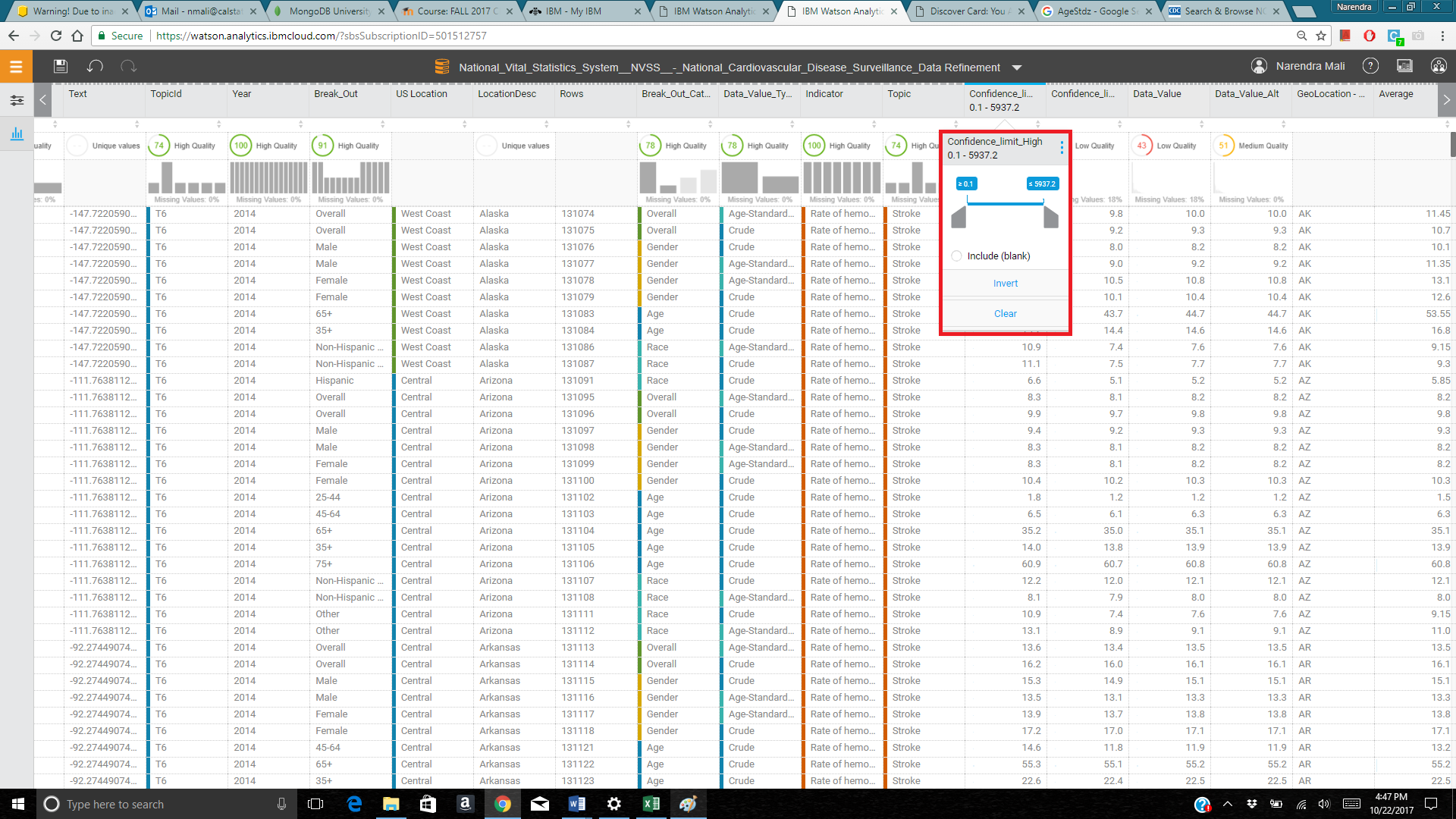
After Refinement:



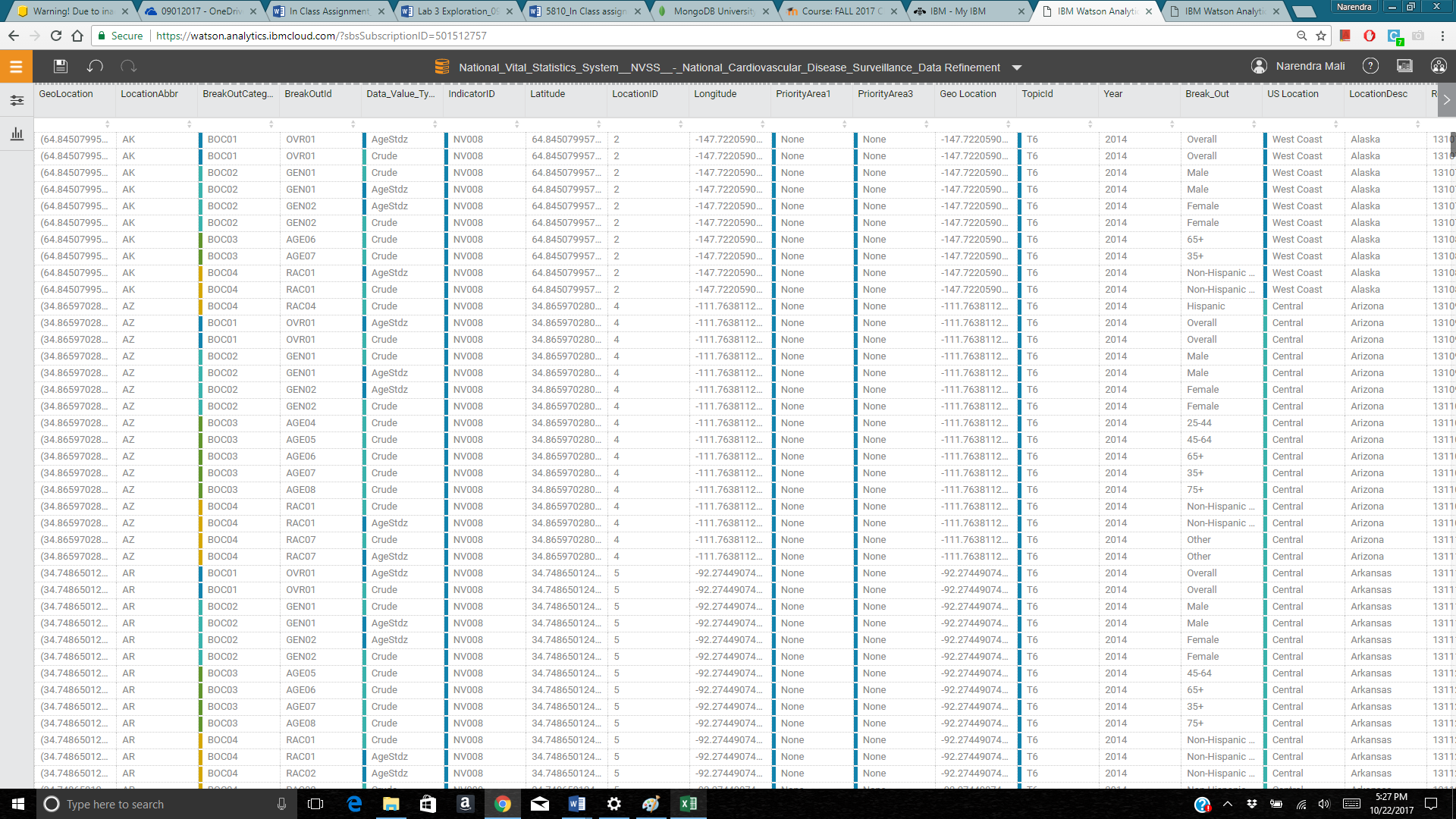
3. Saved the file in .CSV format again so that all excel formulas will be removed automatically.

**In Watson:**

Removing blanks: Removed Blanks in IBM watson

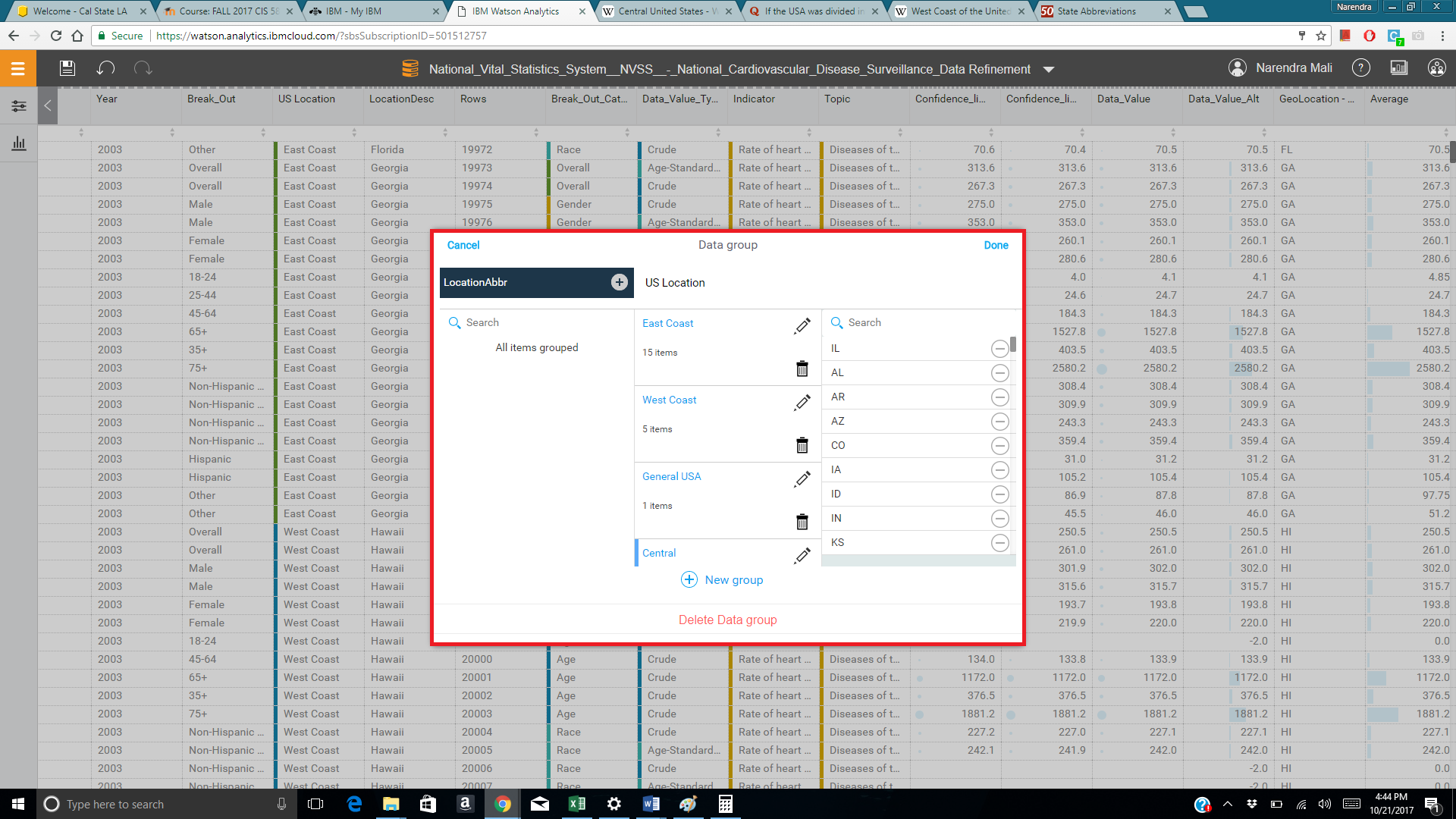


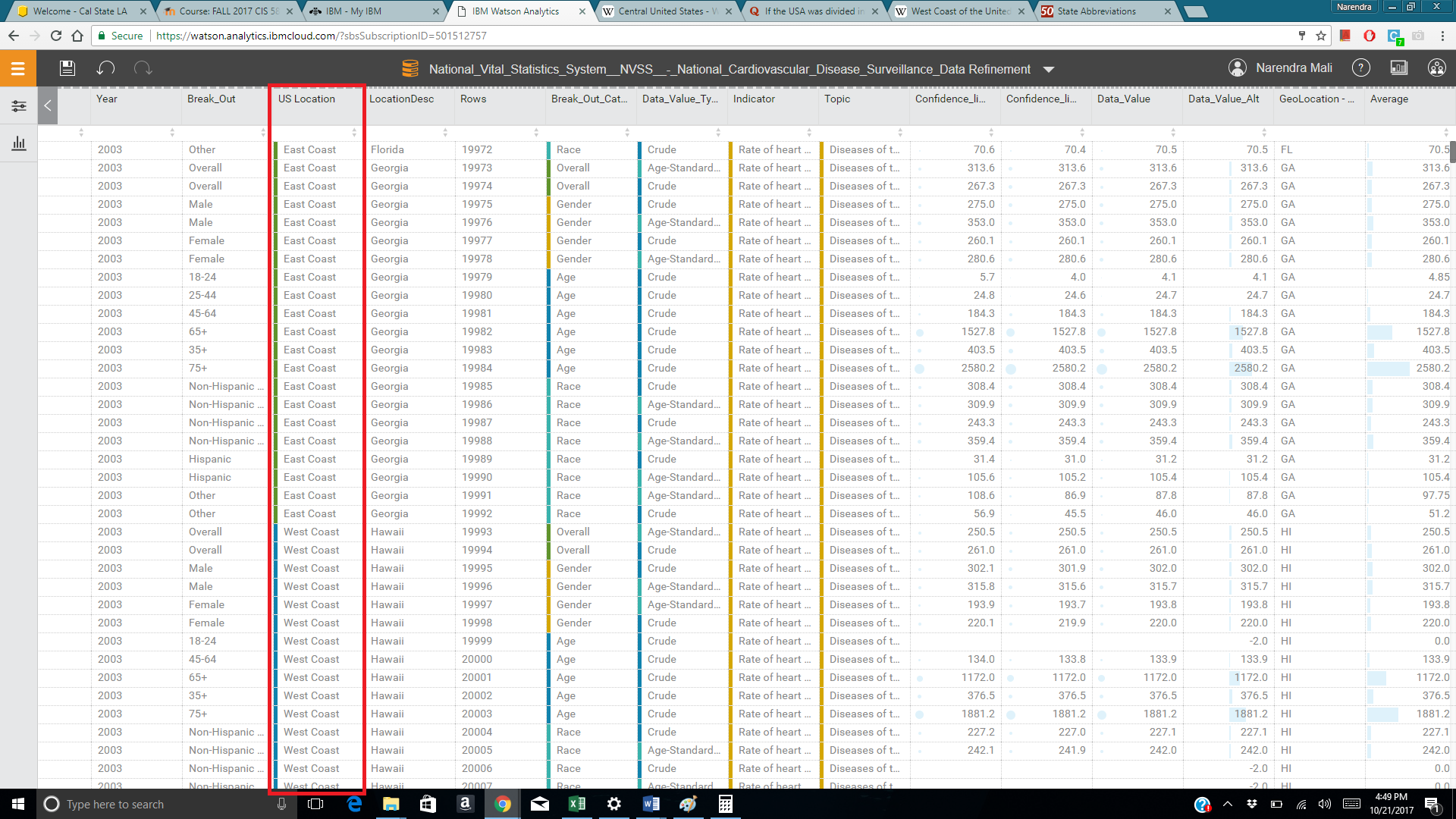
2. Row names, abbreviations and other fields were in proper format and without any spelling mistakes. Hence, no data cleaning field were found in IBM Watson Analytics.



**Grouping:**

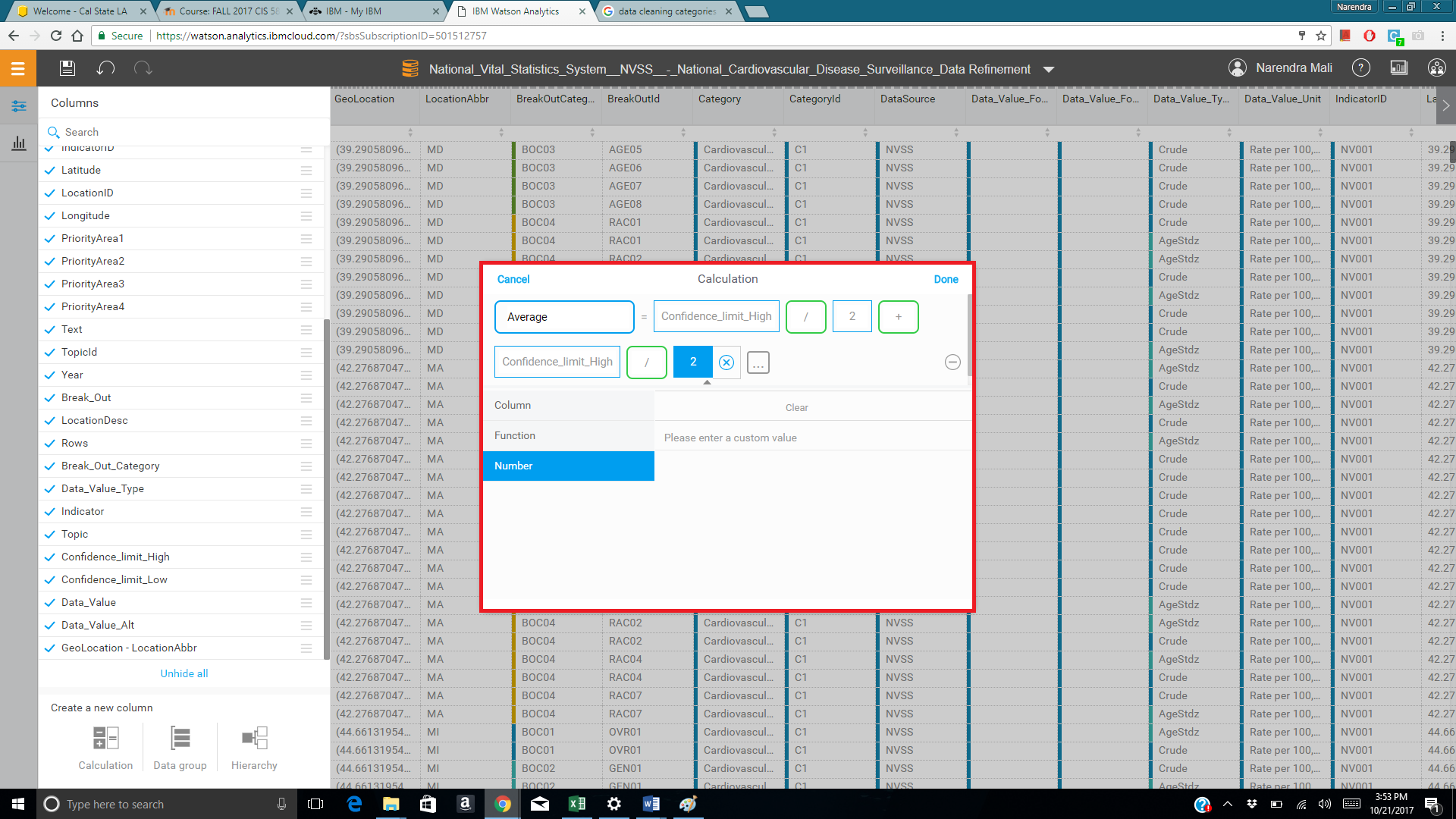
Grouped the column LocationAbbr in Four parts. Those four parts are as follows: East coast West Coast, Central USA and General USA. General USA contains not specified US location.

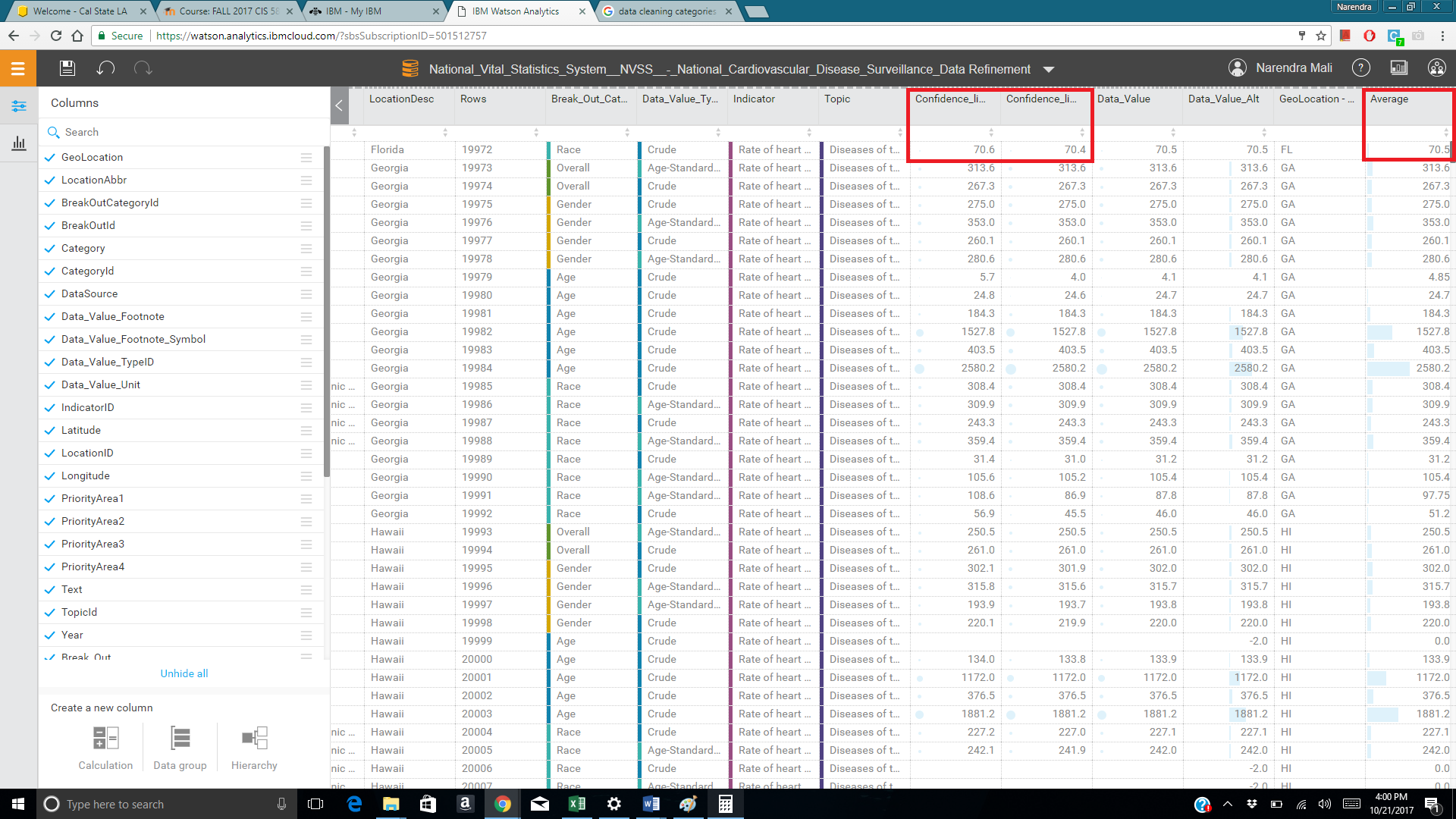




**Calculated field:**

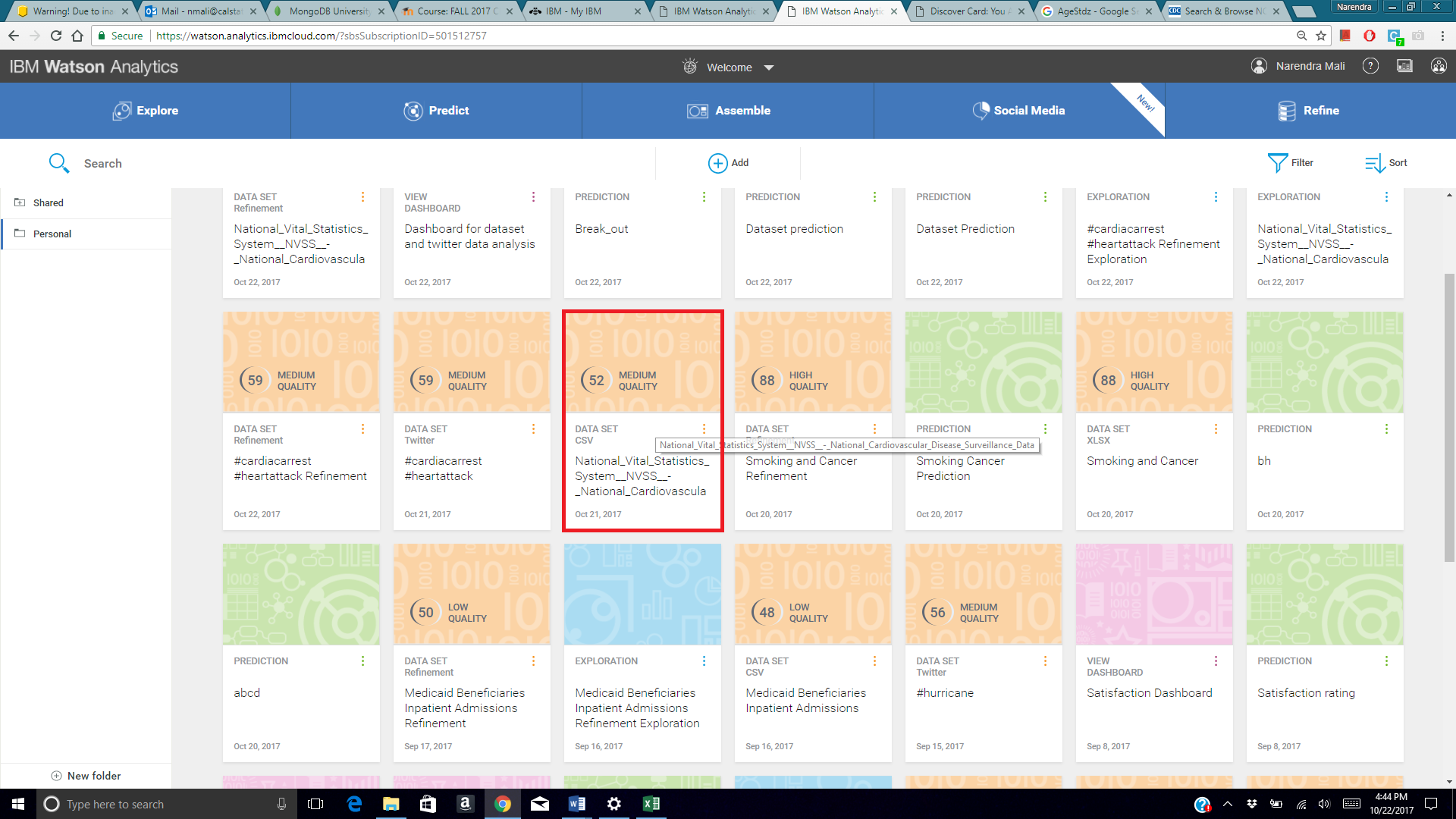
Average of Confidence limit high and Confidence limit low.

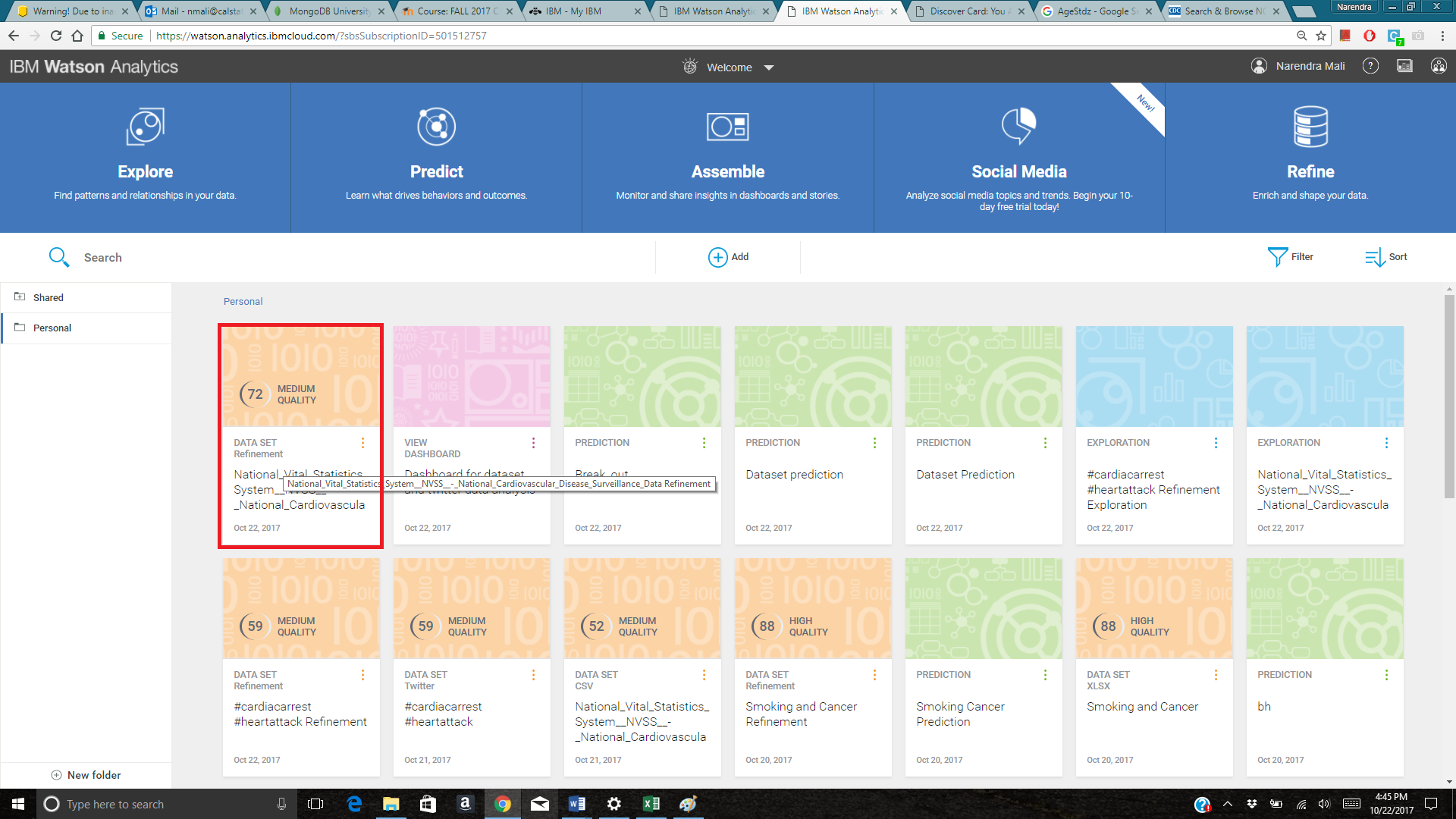




**D. Data Quality**

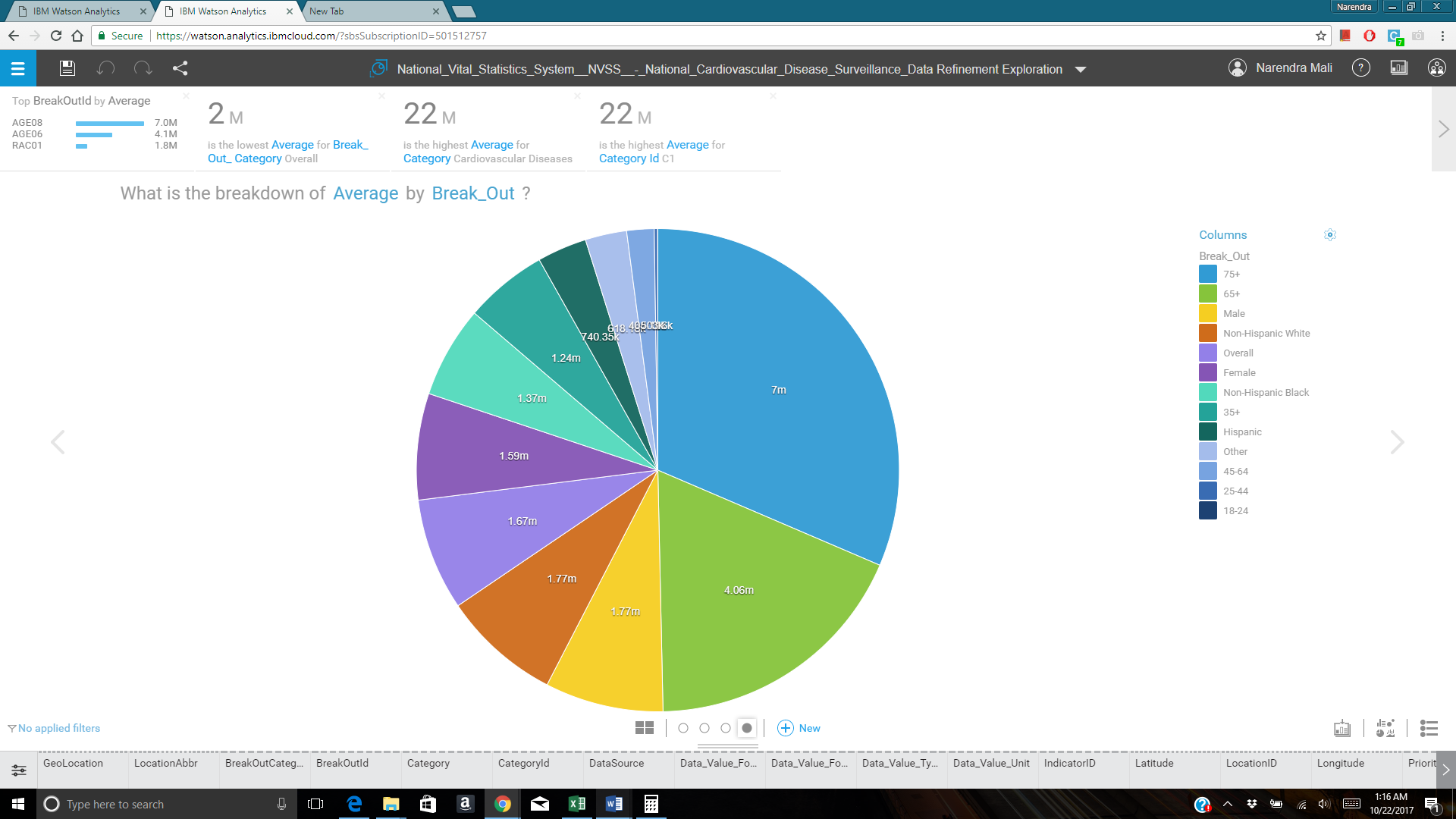
**Data quality strength of the entire data before refinement:52**



**Data Quality after refinement: 72**

1. **Data Exploration:**

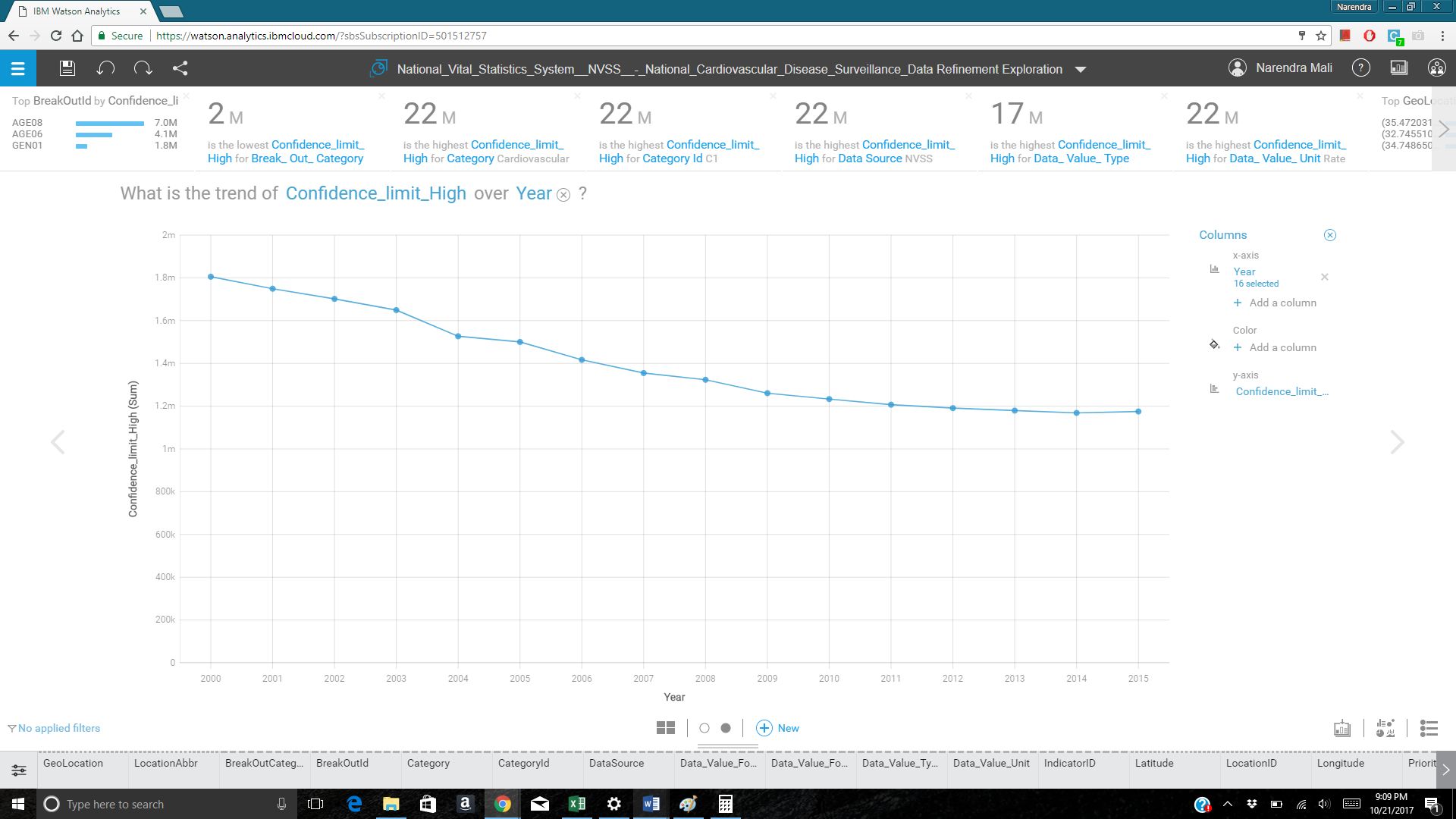
**Q1. What is the breakdown of Average by Break\_out?**



[Chart type: Pie]

Average is the field created to take average of high and low confidence limit. High and low confidence limit are nothing but rating which gives upper and lower value of how many people can have cardiovascular disease per 100,000 people. Break out has specific age and race categories. From the above pie chart, there are more than 7 million people who were in the age group of more than 75 years. Surprising finding from this pie chart is that there are more than 50,000 cases for people who are in the age group of around 25 years old. There were 1.77million and 1.59 million cases of heart related issues in men and women respectively. Heart diseases in men is more common than women. Also, over the year heart related issues are becoming common in women as well. Pie chart also tells us than there are 1.77 million non-Hispanic white people who suffered through cardiovascular disease.

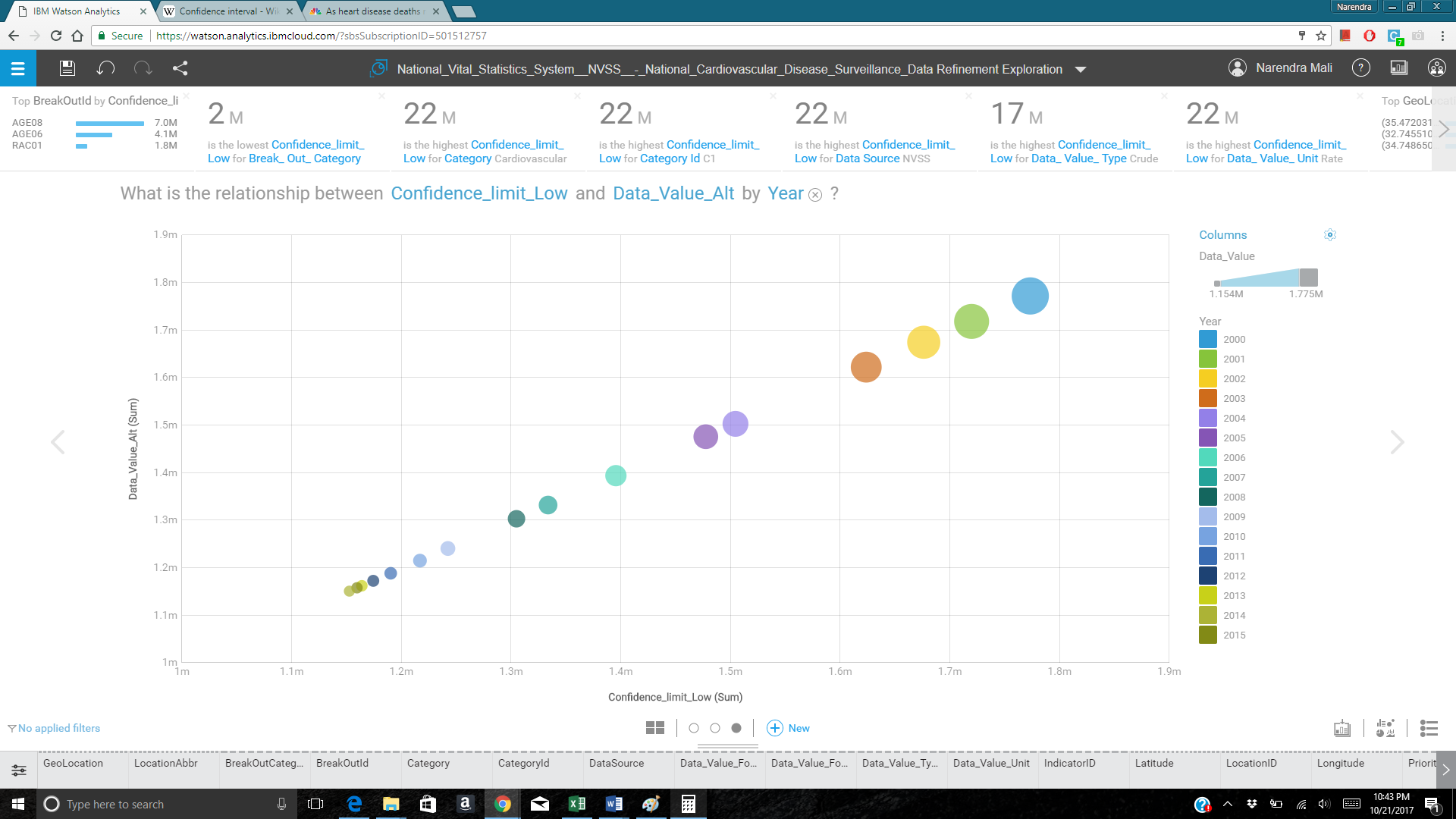
**Q2. What is the trend of Confidence\_limit\_high over Year?**



[Chart type: Trend line]

Confidence limit high category represents maximum number of people who can have cardiovascular disease per 100,000 people. The above trend line shows us that because of advancement in technology and public awareness the number is significantly decreasing. But after careful observation, the number is slightly increasing in the year 2015 and it will be on the same level as 2009-10.

**Q3.** **What is the relationship between Confidence\_limit\_low and Data\_value\_Alt by Year?**



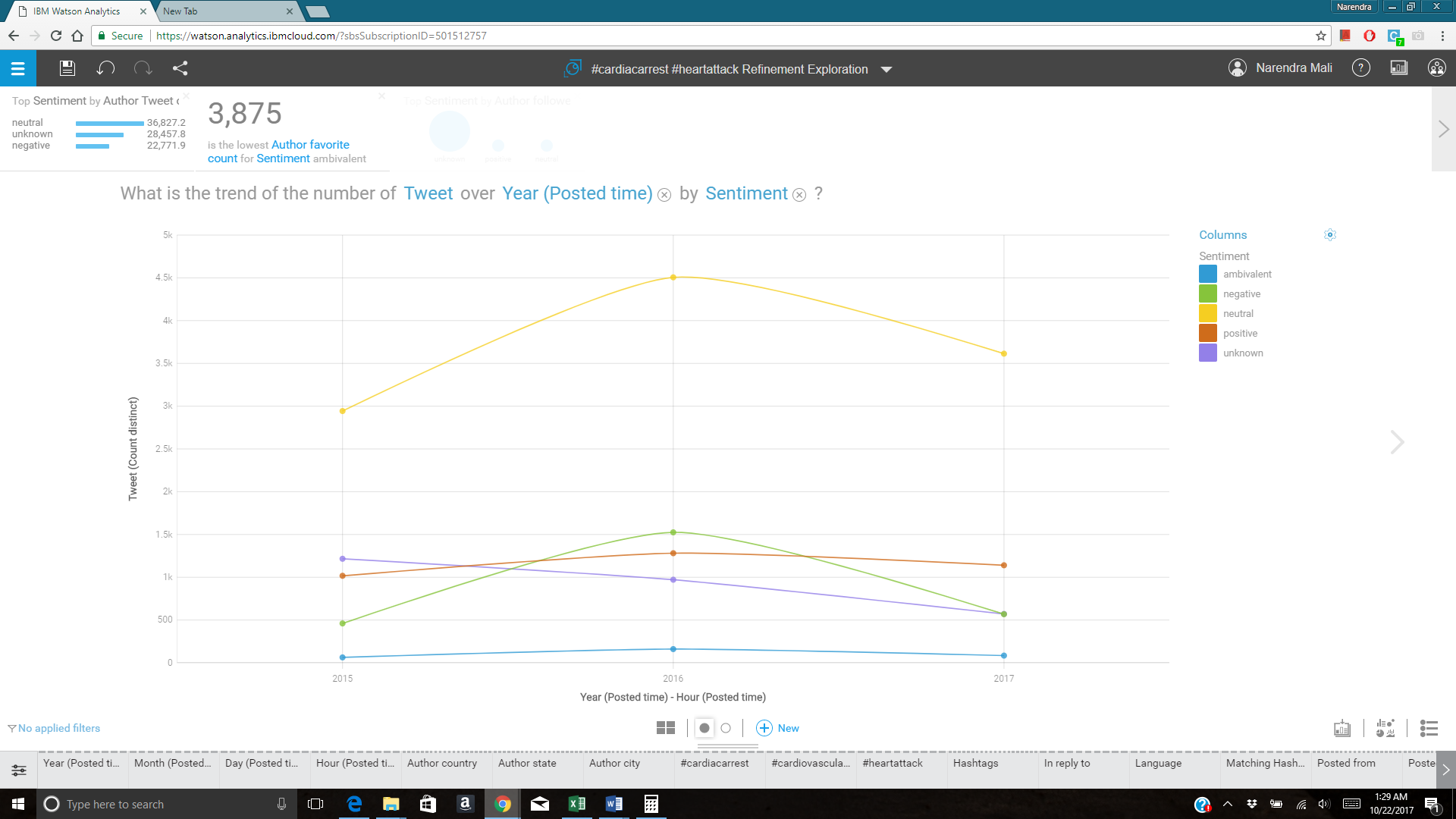
[Chart Type: Bubble]

The above data represents the actual data per 100,000 versus the confidence limit lower side over the years. As it can be seen from the above chart, the number were fairly close, and the confidence limit were accurate. There were more than 1.8 million people in danger with cardiovascular related diseases in year 2000, compared to 1.15 million people in year 2015.

**Twitter hashtags analysis:**



**Q1 twitter: What is the trend of the number of tweet over year (posted time) by Sentiment?**

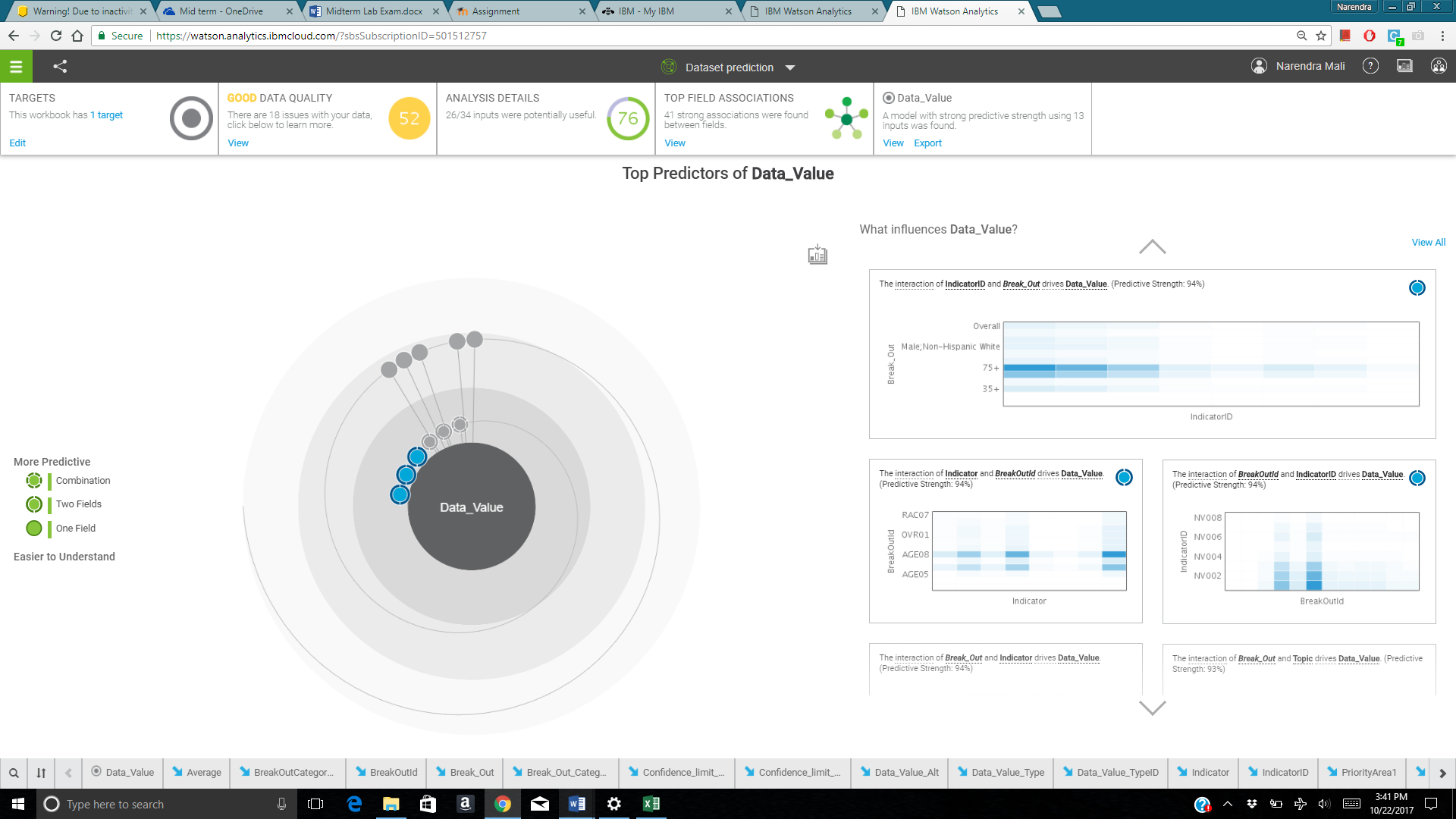


[Chart type: Trend Line]

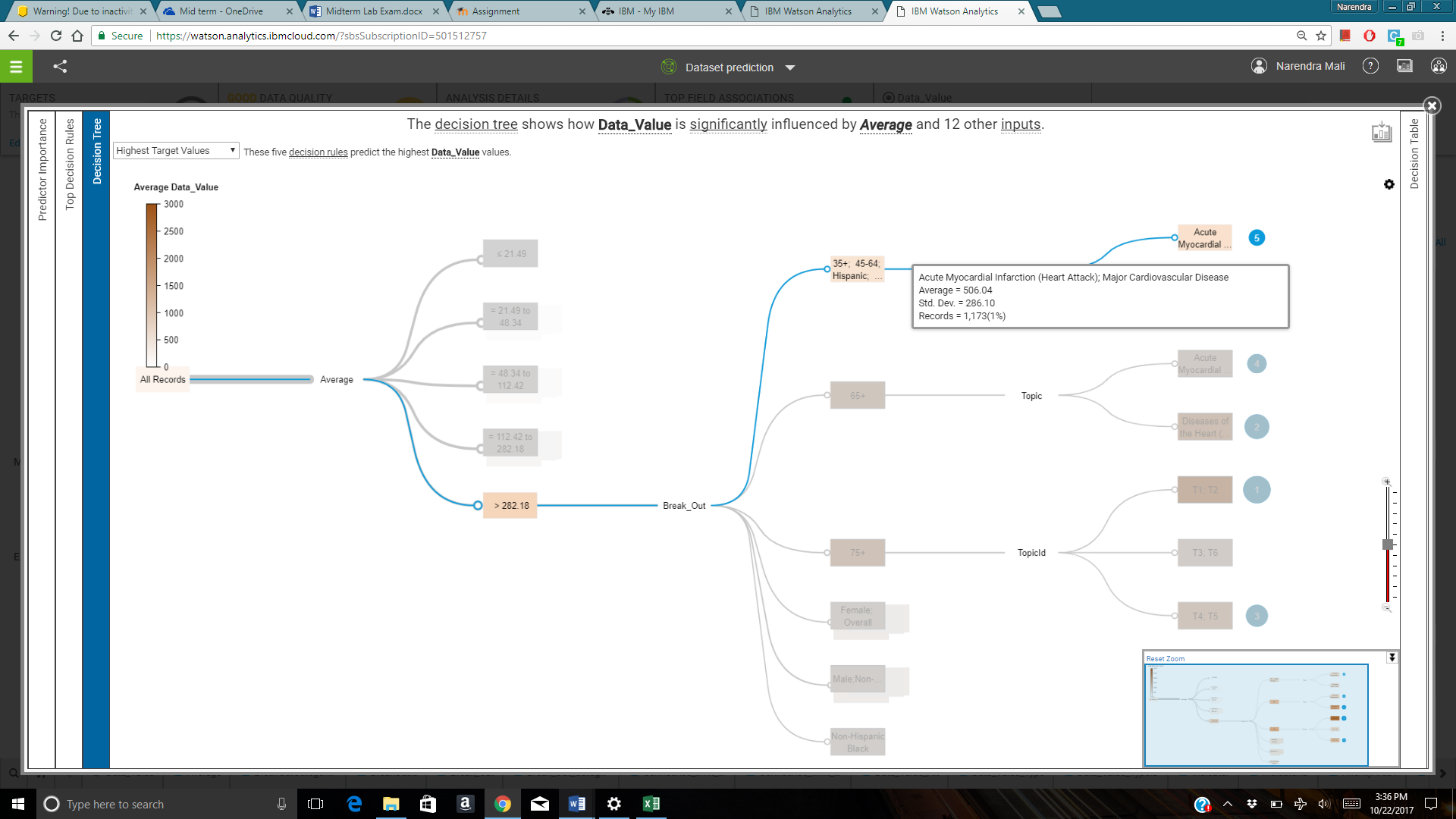
Twitter data is being taken for three important hashtags and in the above graph it has been analyzed against sentiment. There are five different types of sentiment predicted in the trend lines. It can be seen that there are more neutral tweets than positive or negative tweets. However, it was seen that in year 2016, negative tweets were in all time high. For the year 2017, data is available till October month only. Even though the number of negative tweets are significantly lower than year 2016. Also, there are 5000 tweets which belongs to unknown sentiment.

1. **Prediction**

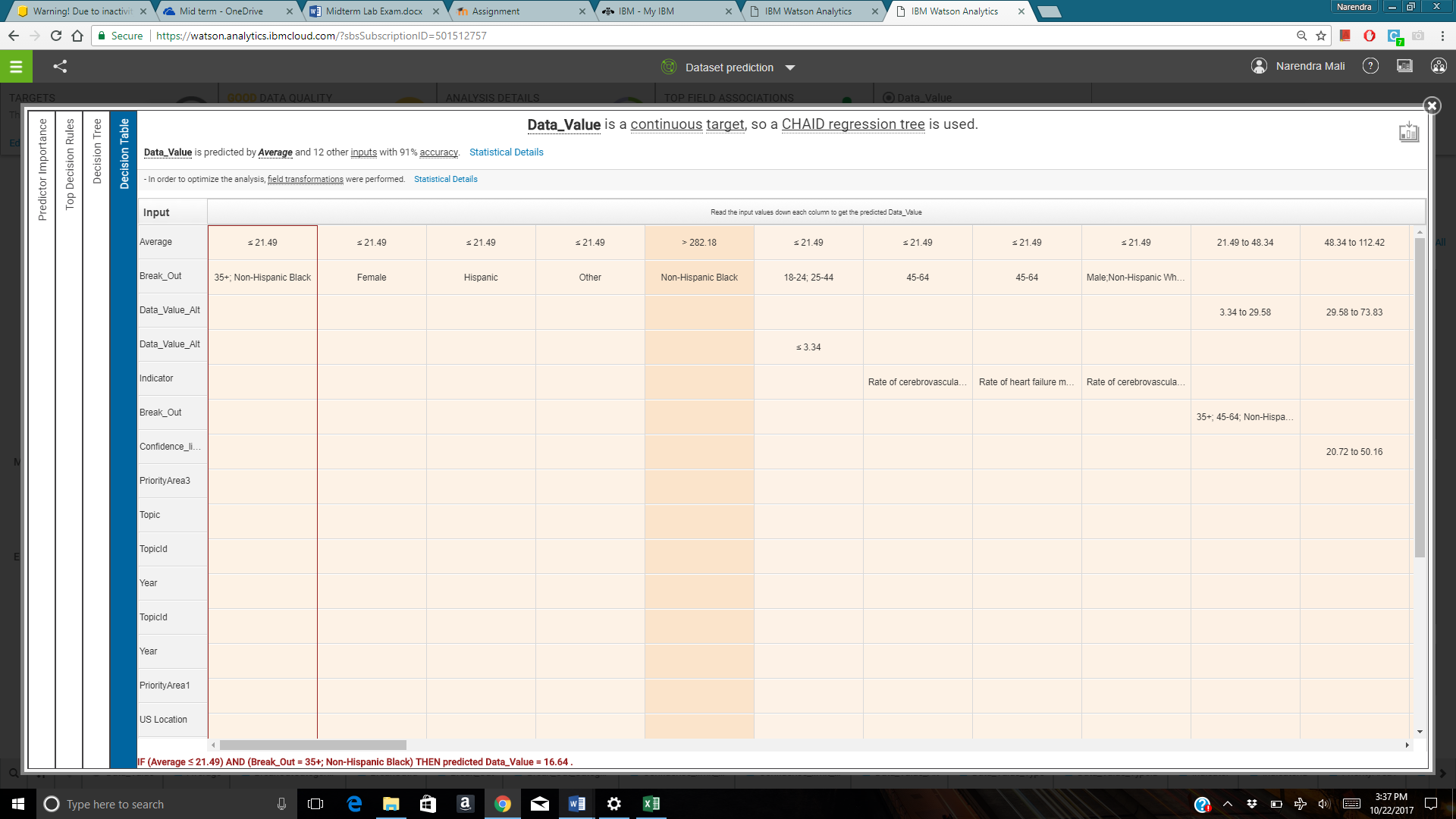
**Bull’s Eye**



**Decision tree**



**Decision Table**

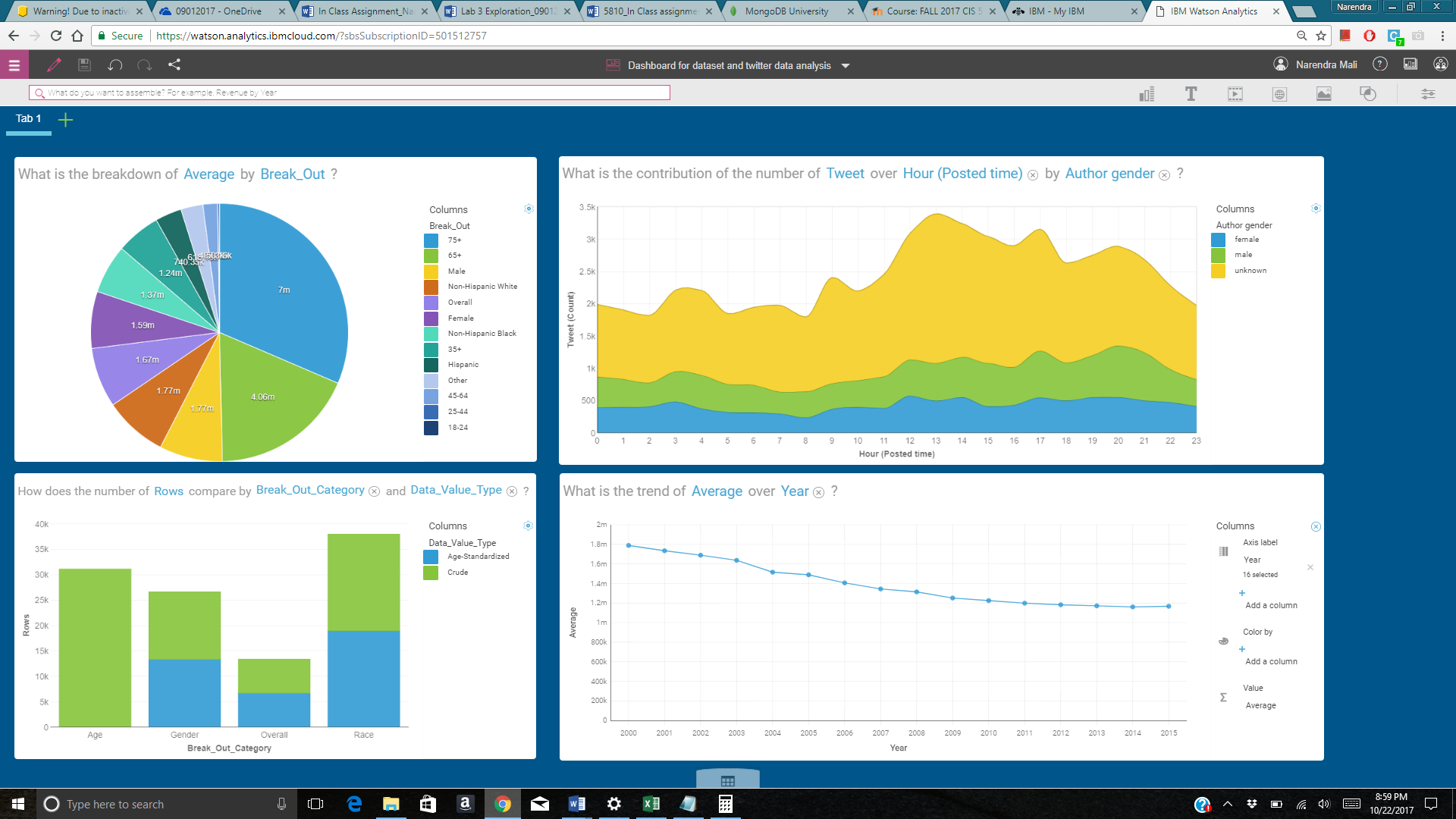


Data value is the actual value which represents number of people per 100,000 people who had the cardiovascular disease. The above prediction report predicts that there are five major categories of people. Fifth category has people in the age group of 35+ with Hispanic race. They have more number of chances to have cardiovascular related diseases. On the first place are the people in the age group of 75+ who has the highest possibility of cardiovascular disease. Possibility of them having a heart related disease is more compared to others. Also, there are more than 2000 people can be found per 100,000 normal human beings with possibility of heart disease.

Heart attack and heart failure are not the only two categories which should be considered at heart related problems. Others category such as Tachycardia which are increase in heart rate at irregular interval of time. Such diseases can also have adverse effect in longer term. Second, from the CHAID decision tree, there are more than 282 people from Non-Hispanic Black community who has higher possibility of heart diseases. Whereas Hispanic people in the age group of 35+ also has the same possibility. A different finding in this decision tree is that overall female can have heart diseases chances equal to Non-Hispanic black men and this probability is more than males who can have heart diseases.

The predictive strength is 93%. Overall, prediction shows that there will be more cardiovascular related diseases observed in the 75+ aged group people. Followed by 65+ aged group but it also states that depending on the topic and category of people, the numbers may differ.

1. **Dashboard**



**Summary:**

The above Cardiovascular disease related dashboard gives us the comprehensive information regarding heart diseases in different aged groups people and also people who belongs to different race in United States of America. Trend lines gives us the past data as well as the future prediction regarding number of people who has the possibility of having heart related disease.

Currently, heart related diseases are found in the aged group of 75+ but from this NVSS dataset, it is proven that heart disease is not just for old people. It has been found that every year there are more than 1000 entries listed in the age group of 25+ who are struggling with heart related disease. Possibility of men having heart related diseases is more but in the near future that number might change. As more and more women are facing heart related issues. Data is small but it covers the last fifteen years of data which is sufficient enough to raise a question regarding how one should take heart related issues seriously. Obesity is growing out of control. A big part of this is economic. Fake foods are more affordable. It's enticing people to eat more because they think they're saving money when they're really just buying heart disease.

# References

Services, U.S. Department of Health & Human. *Heart Disease*. 24 08 2017. 22 09 2017. <https://www.cdc.gov/heartdisease/facts.htm>.