**Week 4: Spark Assignment**

**Student Name:** Shivani Sharma

**Course No :** ALY6110

**Course Title:** Data Management and Big Data

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Mr. Daya Rudhramoorthi

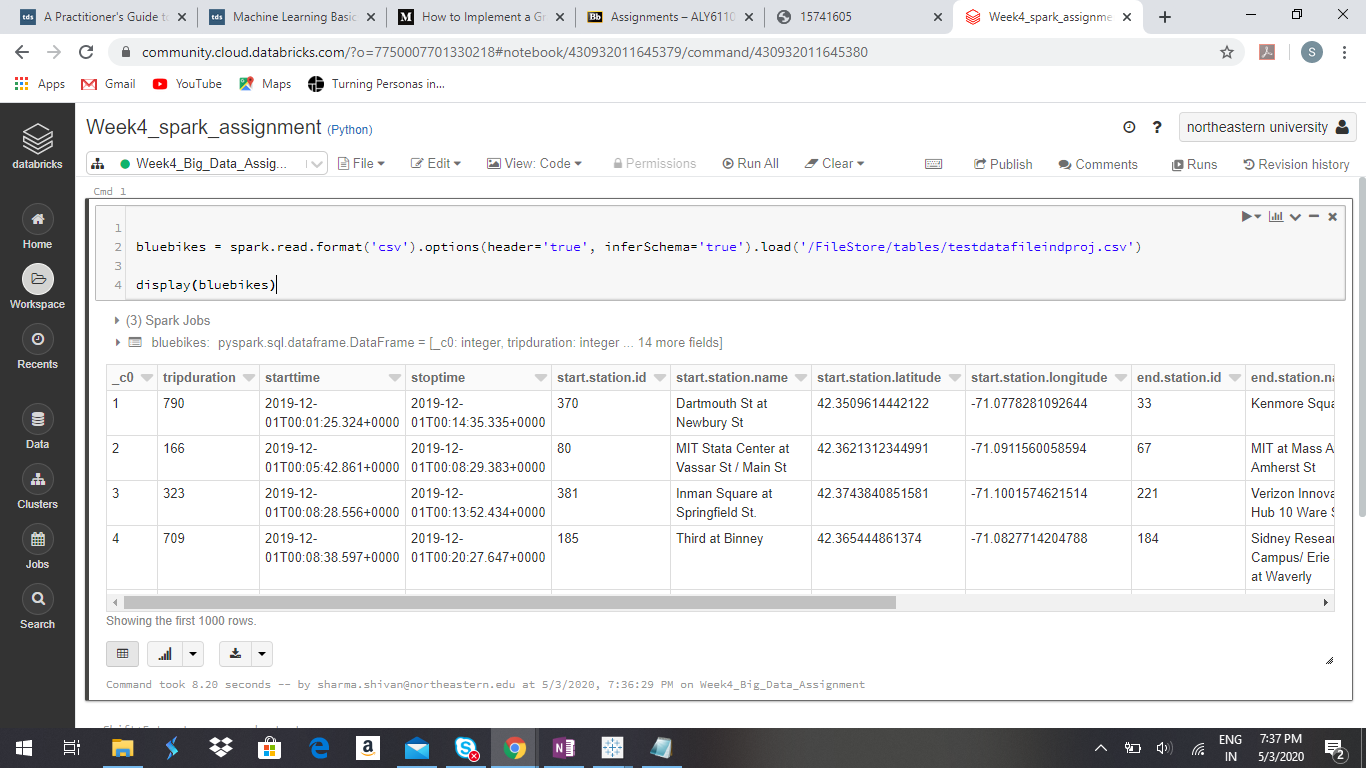
**Lab Work Completion Date:** 05-03-2020

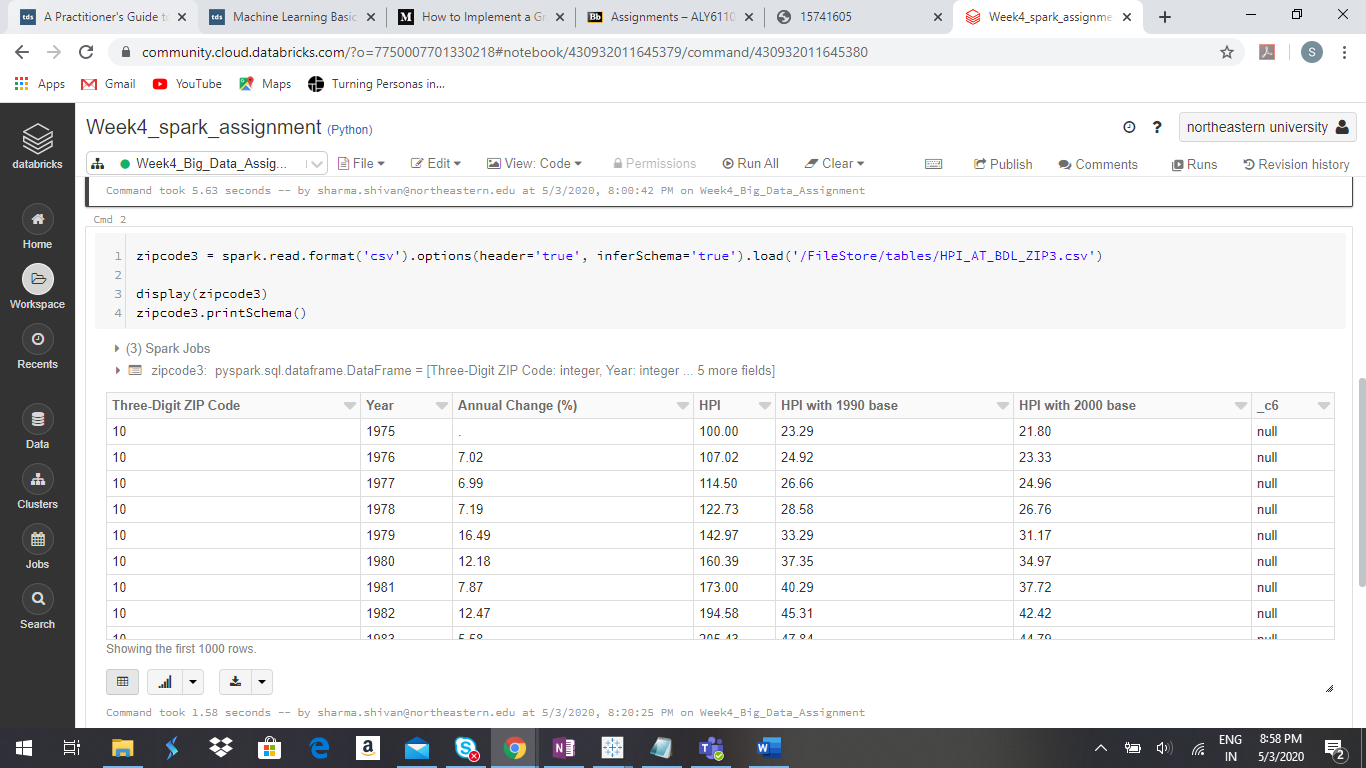


**Summary**

In my spark assignment I have chosen one dataset from my final project i.e bluebikes trip data and studied and analyzed the pattern for top 10 age groups who are having longer trip duration and shorter trip durations. Another data set that I have chosen is the HPI data for zip 3, where I studied the trend for top 10 annual % changes in the values and top 10 HPL producing zip codes of 3 digits.

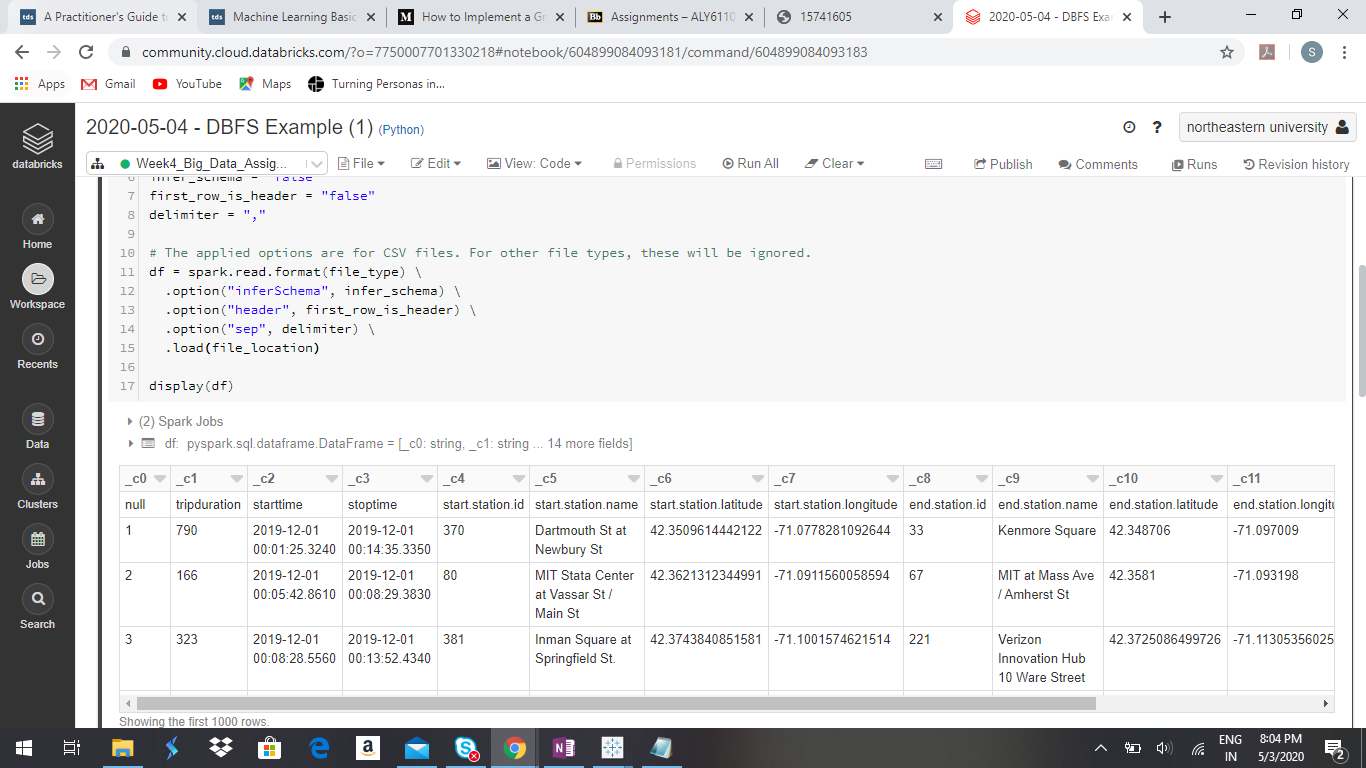
For this I have used spark 2.4.5 version. I created a cluster and hosted spark on it with python and then loaded the datasets into Spark . Created tables on top of that and then I performed analysis on the data. Here are the screenshots for loading the created datasets into spark.

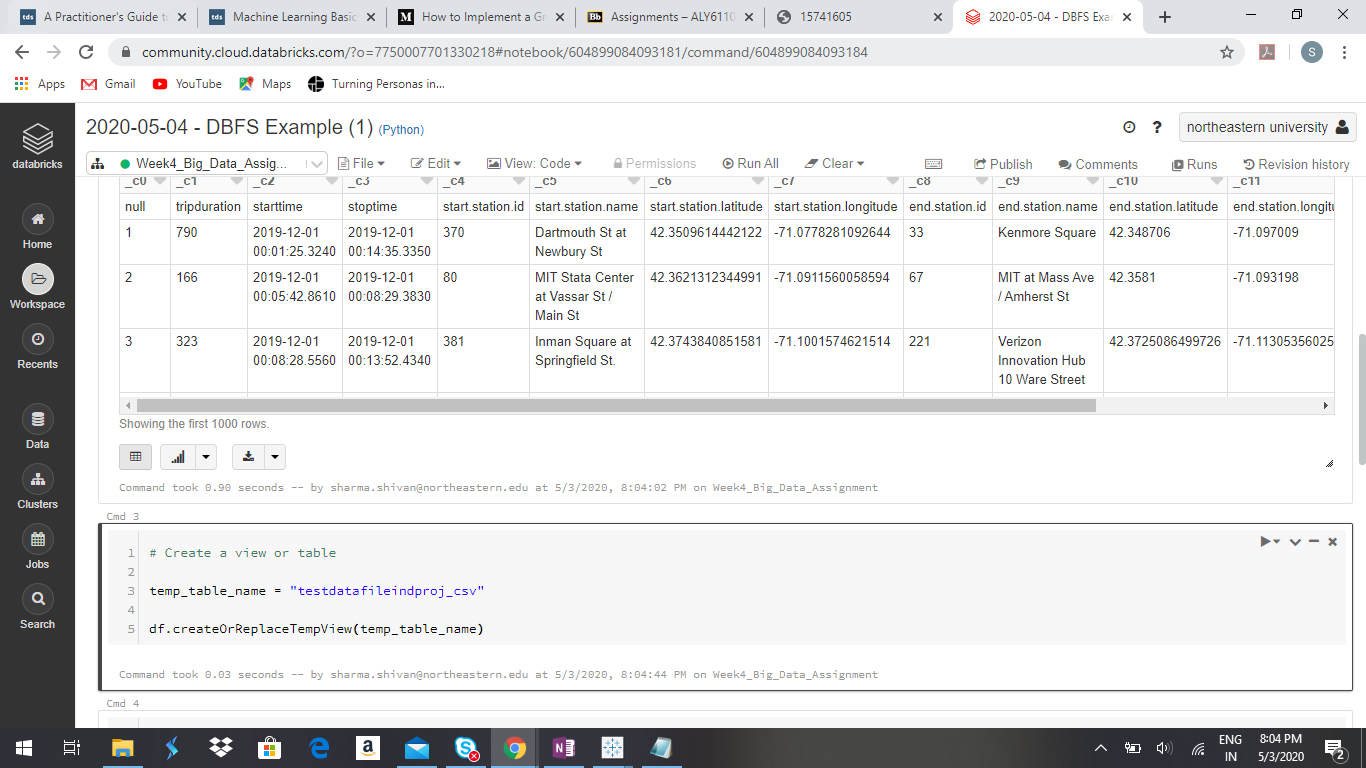


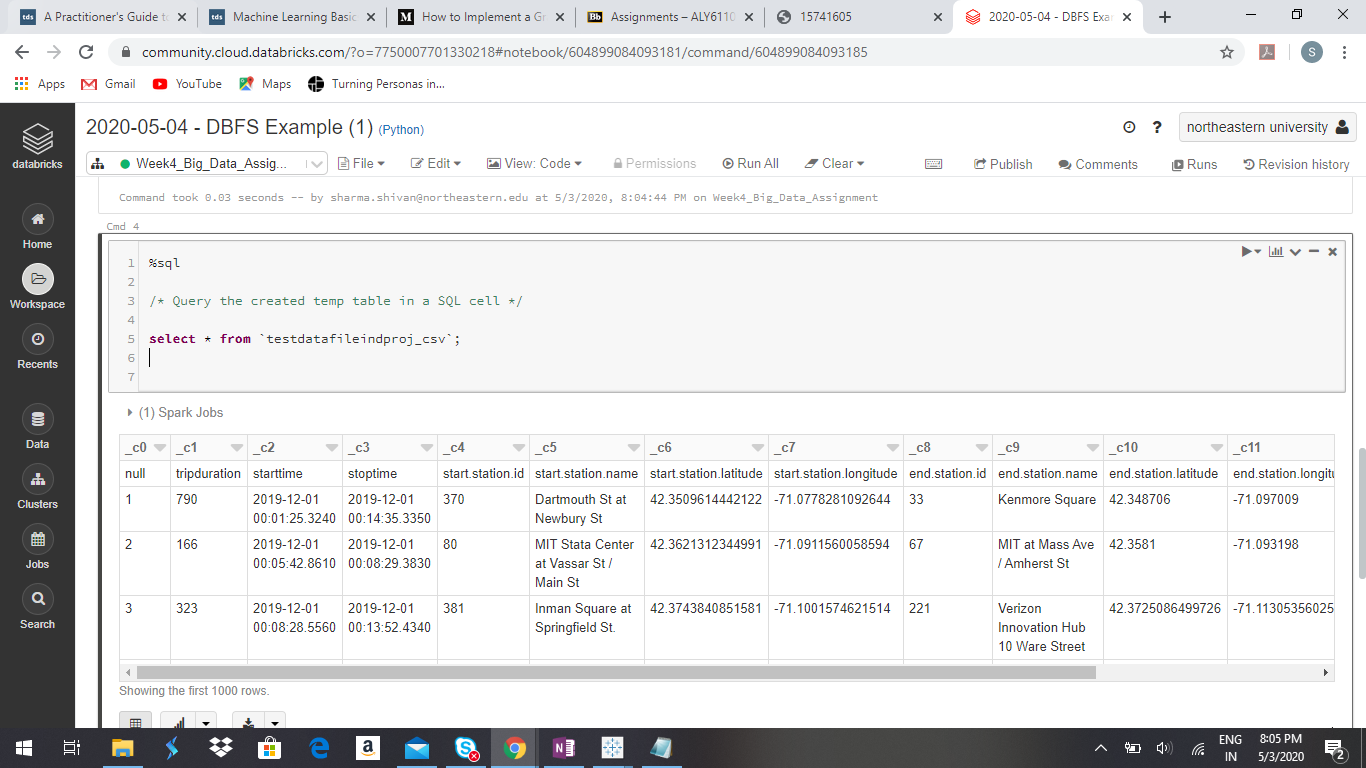


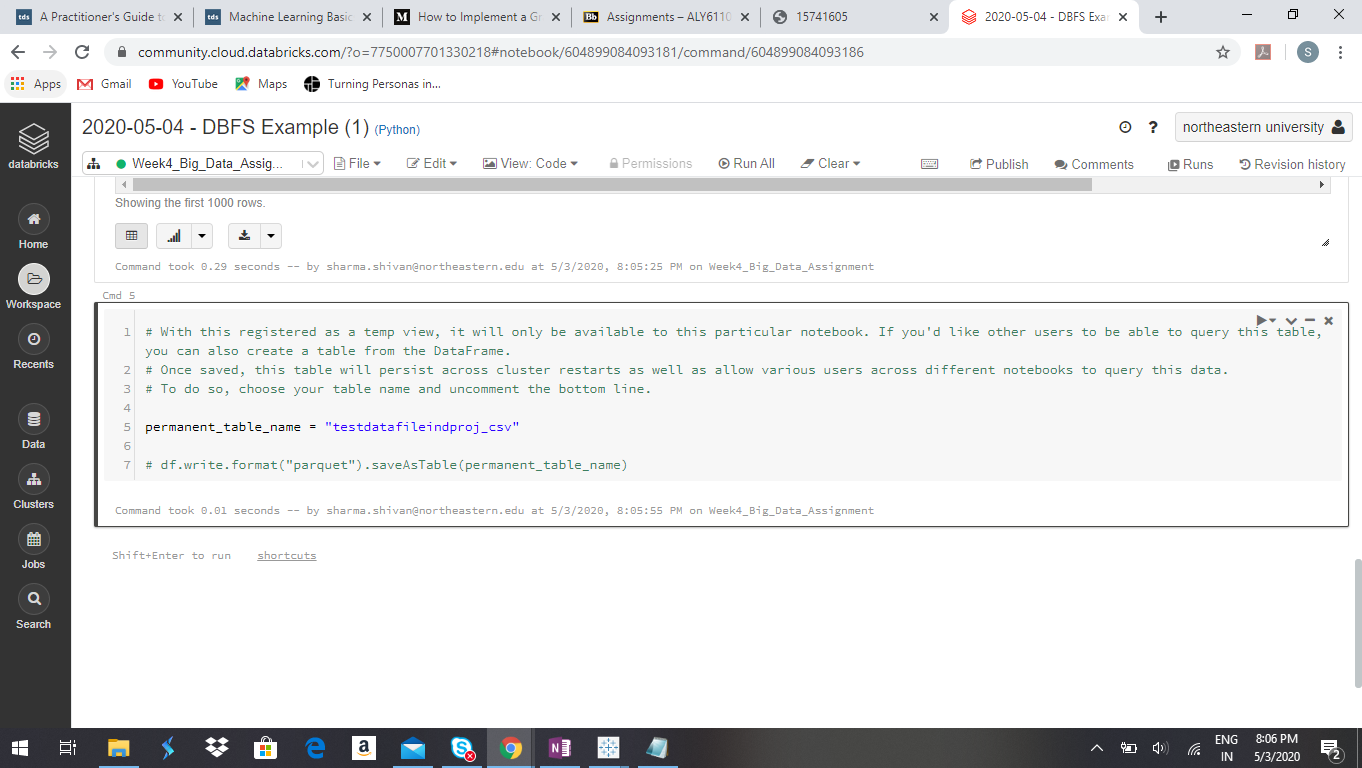
**Analysis**

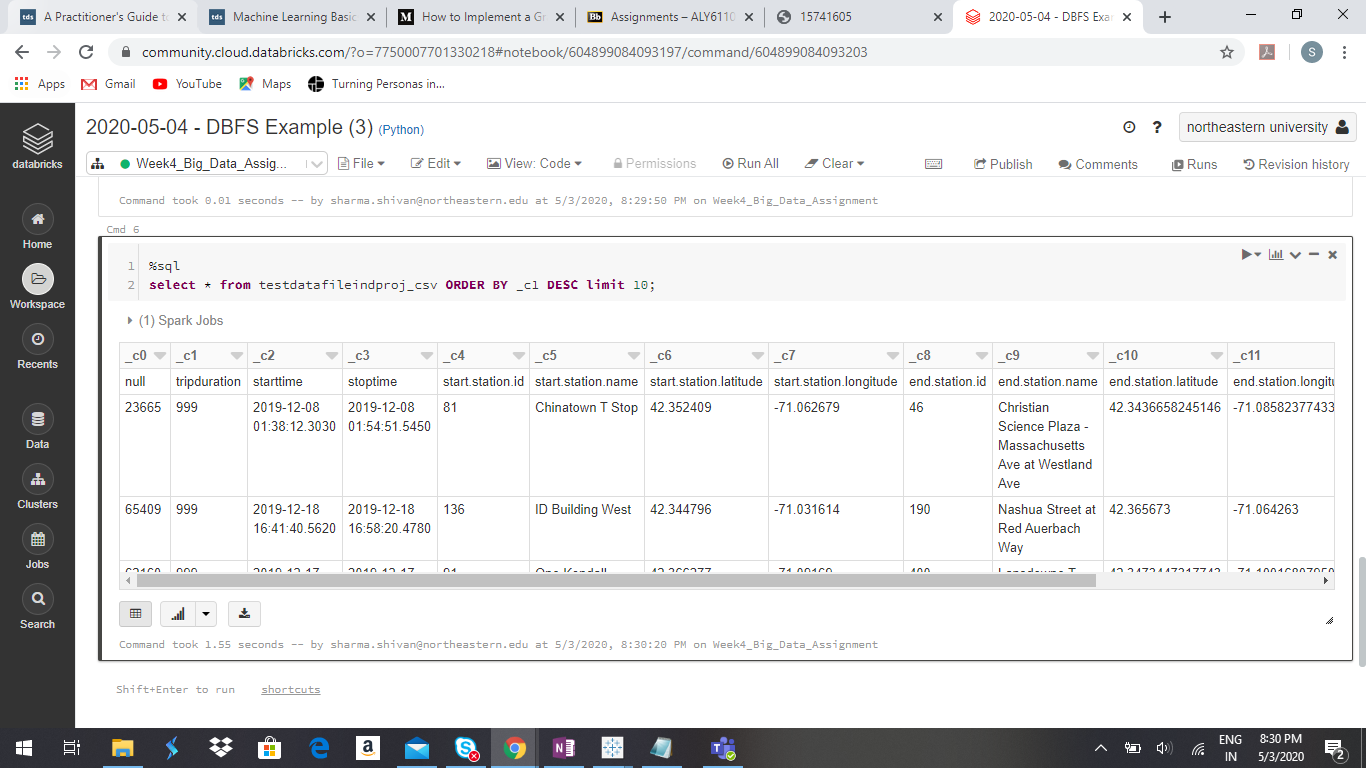
Firstly, I will start with “BlueBikes” trip data set analysis. So, I created tables and schema for the same and performed analysis on finding out longer and shorter trips based on age groups. Here are the screenshots for the same:

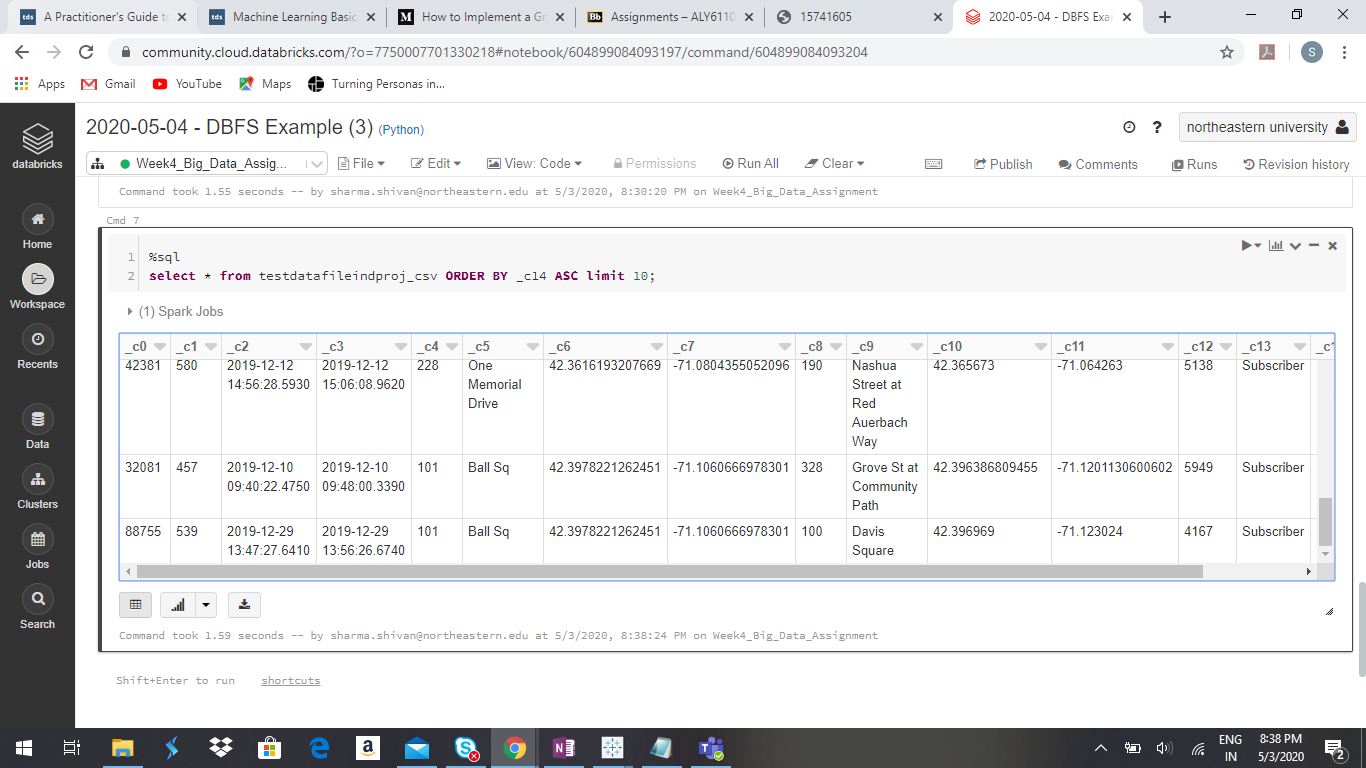






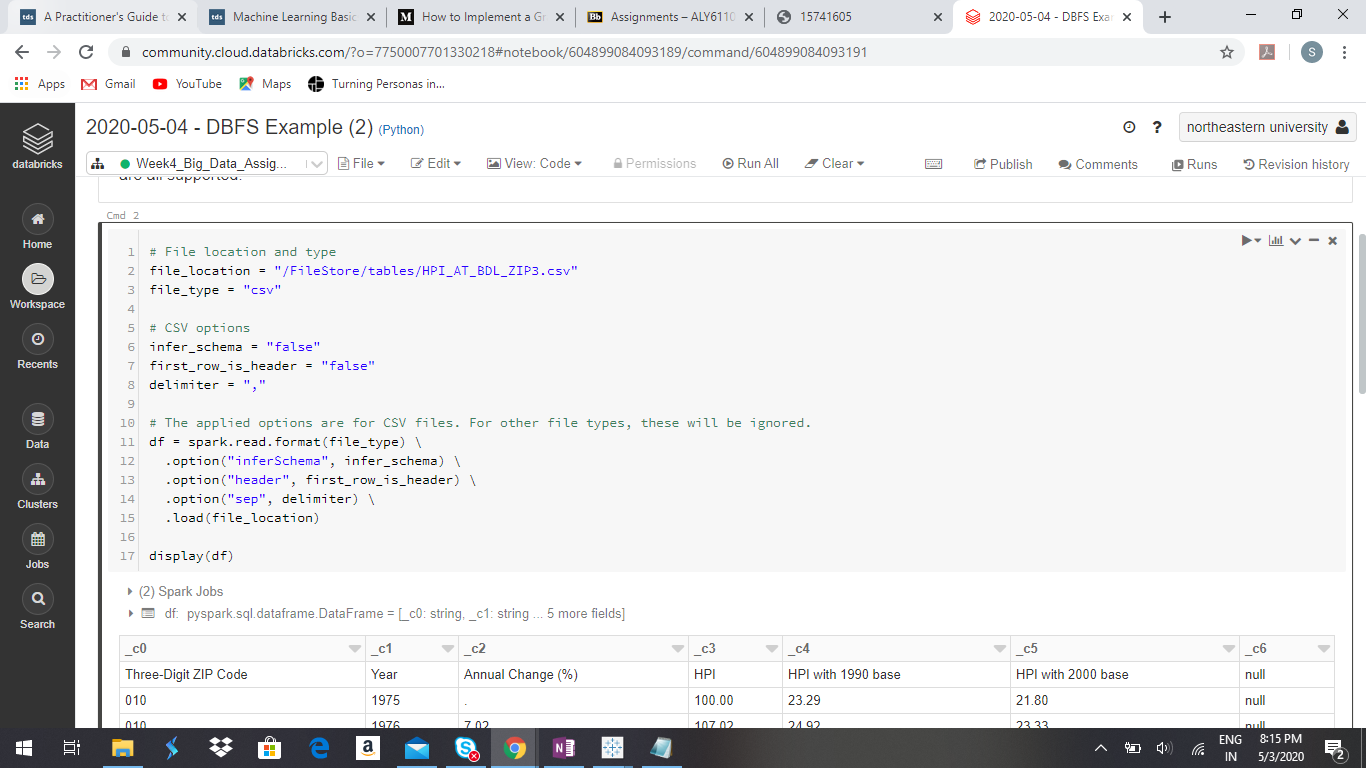


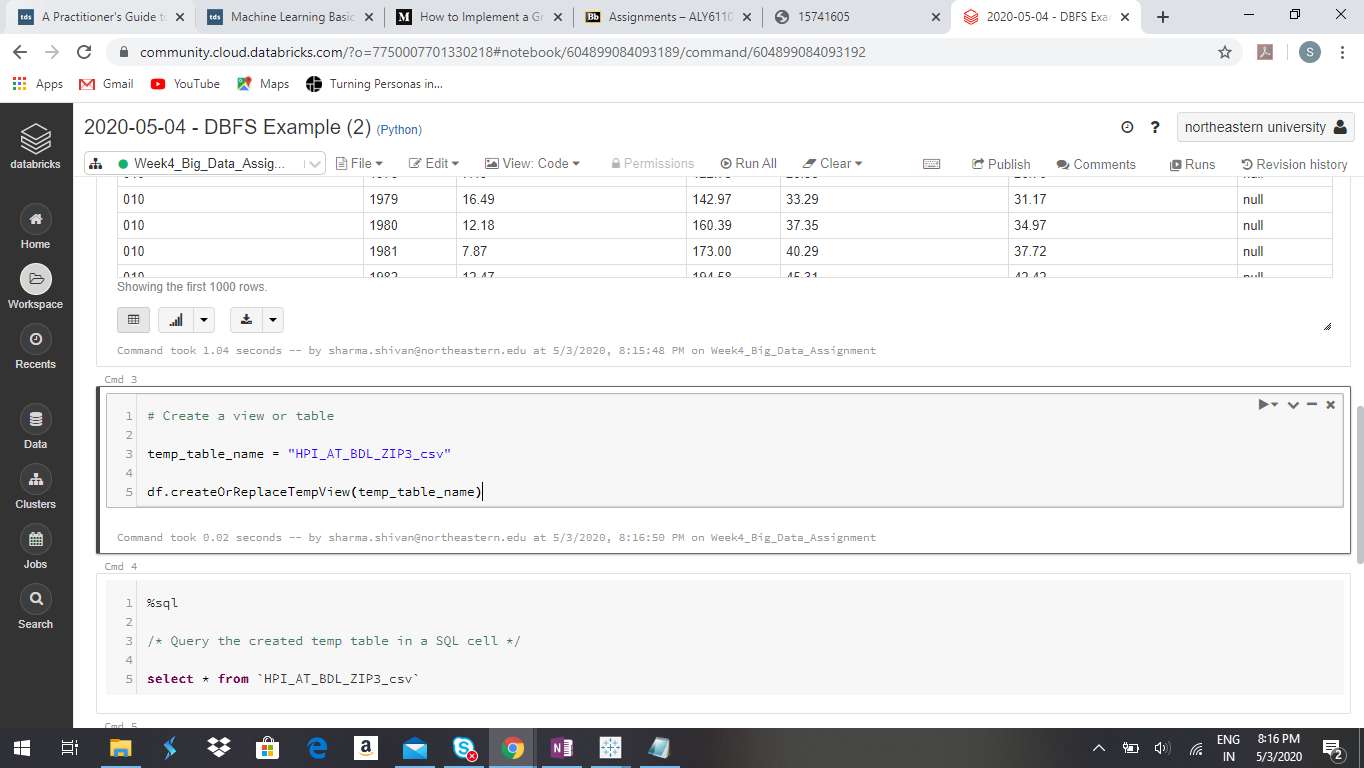


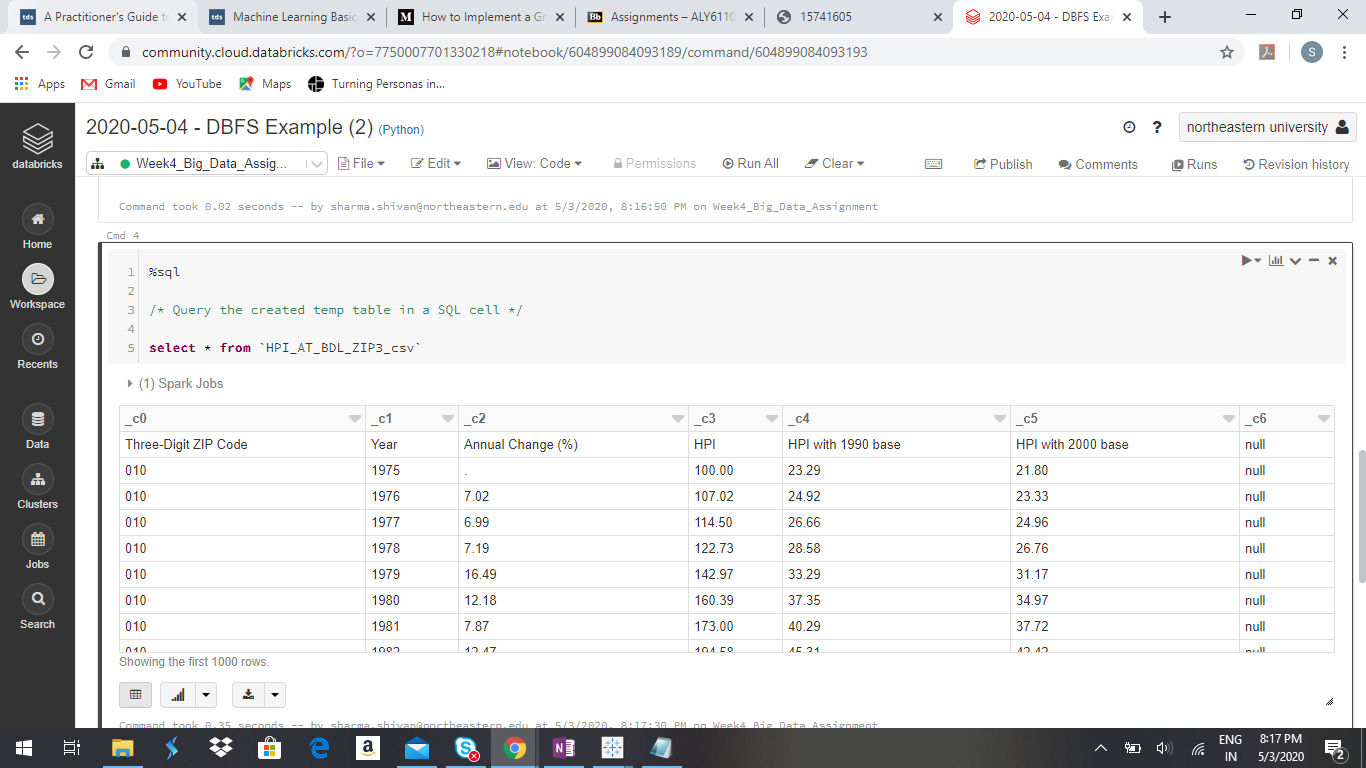


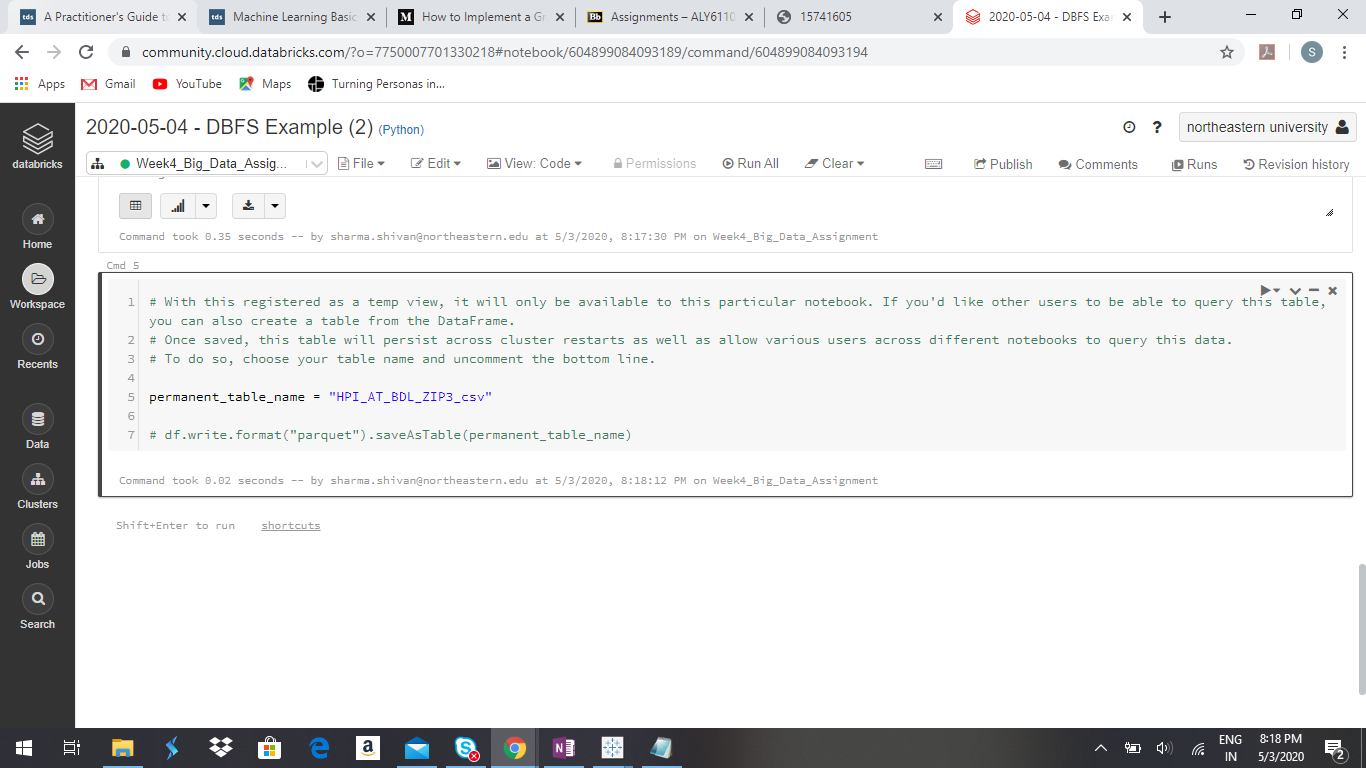
From the above results we can observe that the people who are born in 1969 are taking longer trips and maximum trips. Secondly people born in 1993 are taking maximum shorter trips. From this we can conclude that people above 50 are preferring longer trips and people of age group 25-30 are taking shorter trips. So these results can be used for business people by offering schemes for them on long duration trips and offering schemes for the young generation on shorter trips.

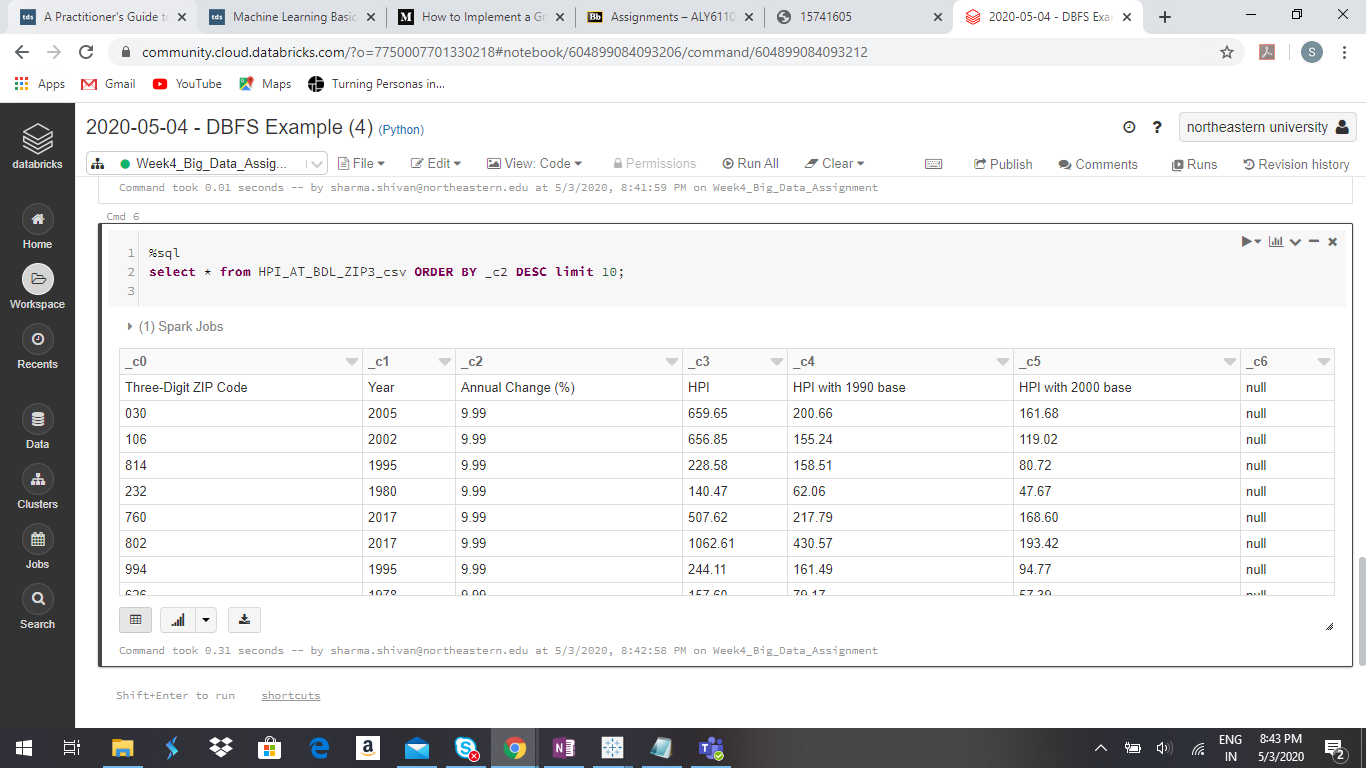
Now coming to data set of HPL with 3 figures of zip code. Following is the schema of the table I have created:

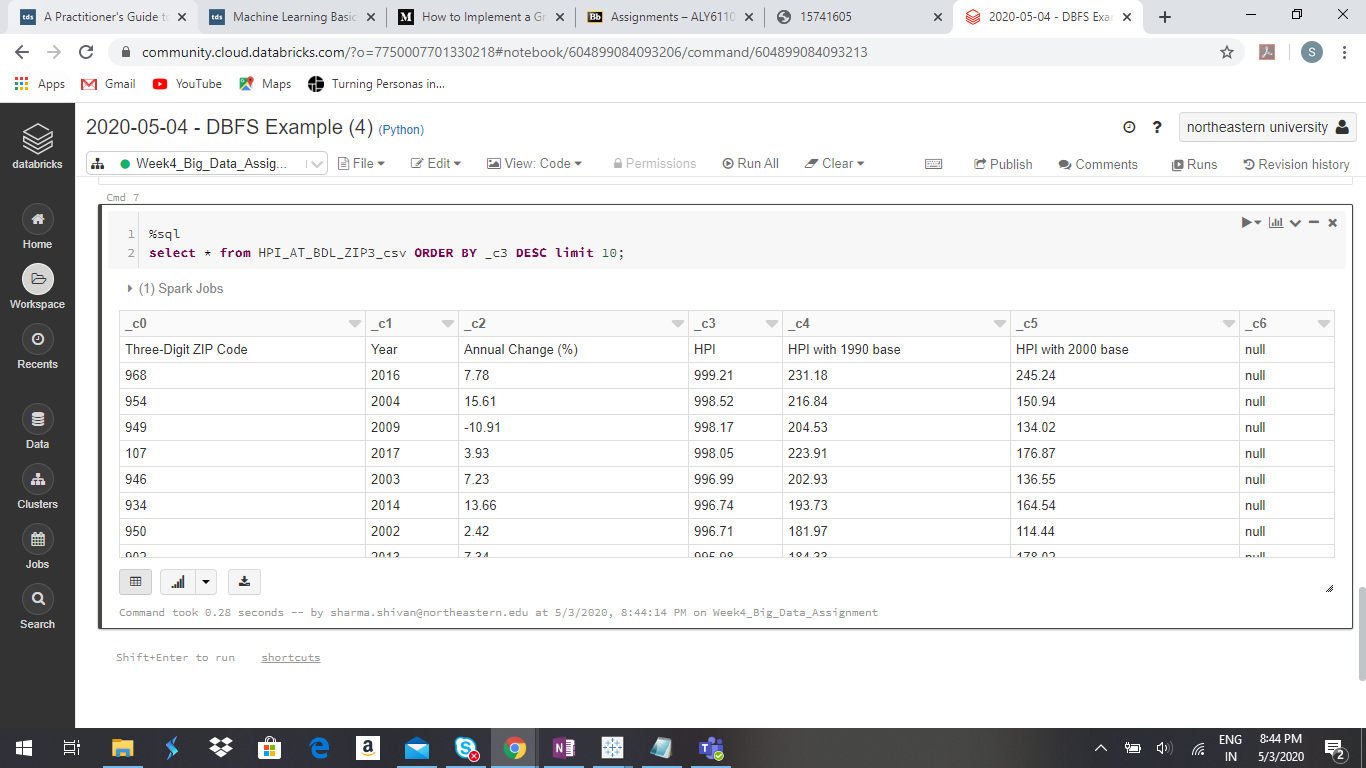












From this data set I have analyzed the top 10 Annual % changes zip codes and top 10 HPI areas and from this data we can conclude that the zip codes which are showing maximum annual % changes are different than the zip codes which are having high HPI values. This data can be used for making future business strategies to focus on maintaining the growth. Similarly we can find out the least 10 areas to prepare strategies for increasing the HPI and annual % changes.

# References

1. “CSV Files.” *CSV Files - Databricks Documentation*, docs.databricks.com/data/data-sources/read-csv.html.
2. “SQL TOP, LIMIT or ROWNUM Clause.” *SQL SELECT TOP, LIMIT, ROWNUM*, [www.w3schools.com/sql/sql\_top.asp](http://www.w3schools.com/sql/sql_top.asp).
3. Malhotra, Rajesh, et al. “Best Way to Get the Max Value in a Spark Dataframe Column.” *Intellipaat Community*, 18 Sept. 2019, intellipaat.com/community/4448/best-way-to-get-the-max-value-in-a-spark-dataframe-column.

“Introduction.” *Introduction | From Compstat to Gov 2.0 Big Data in New York City Management*, ccnmtl.columbia.edu/projects/caseconsortium/casestudies/127/casestudy/www/layout/case\_id\_127\_id\_884.html.