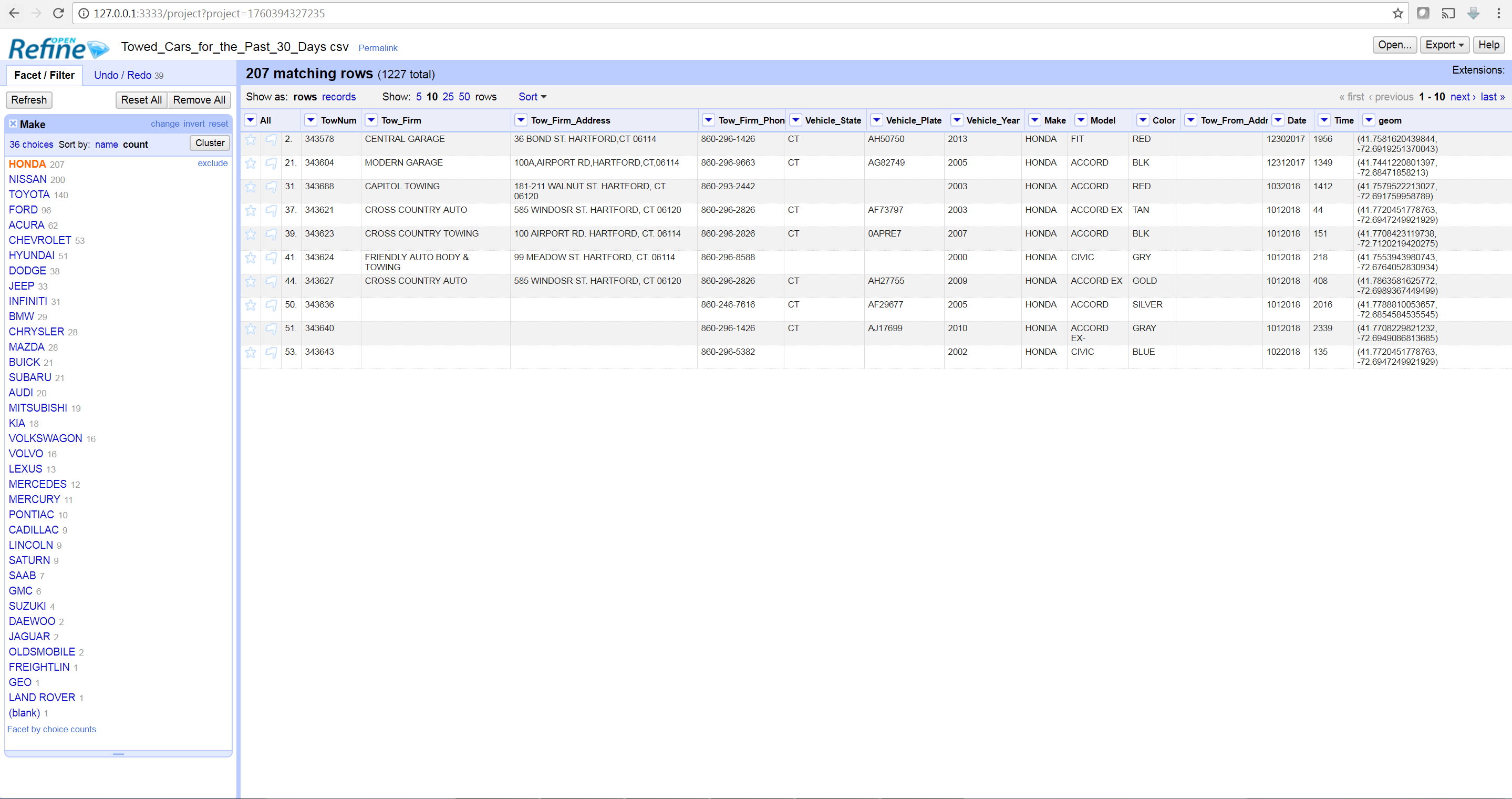
**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 2**

**Part 1**

**Using OpenRefine do the tutorial on Towed Cars here:** <http://goo.gl/tR8eOt>**. Provide screen capture of your finished file from within the OpenRefine environment.**



Use Clustering tool and replace tool I clean the “Make” column to 36 different car manufacturers.

**Read the OpenRefine documentation on their clustering algorithms here** <https://goo.gl/hYOxm>**. Nothing to turn in here, but be prepared to discuss in next class.**

**Part 2**

In Text Mining class, we use Stemming or Lemmatization to convert similar words to tokens. We can also remove stop words, emoji or other special characters. We can only do dictionary replacement on similar term (New York or NY or NYC).

OpenRefine is a nice tool to do data cleaning on text data. Because human typed text has lots of incorrect or variant terms OpenRefine can help us merging those data. The strength of OpenRefine is it has two Clustering Algorithm Key Collision and K-Nearest Neighbors to give good suggestions on terms that means the same thing.

We can perform the same clustering or KNN functions in OpenRefine using Python code. But this will be harder to do than using OpenRefine because we would have to write the code ourselves or find existing library to do the same thing.

The strength of the clustering function is fast and based on algorithm only (machine learning approach). Its computationally complexity is linear on number of input words. It can be a first option or combined with other semantically-aware dictionary approaches to get better result.

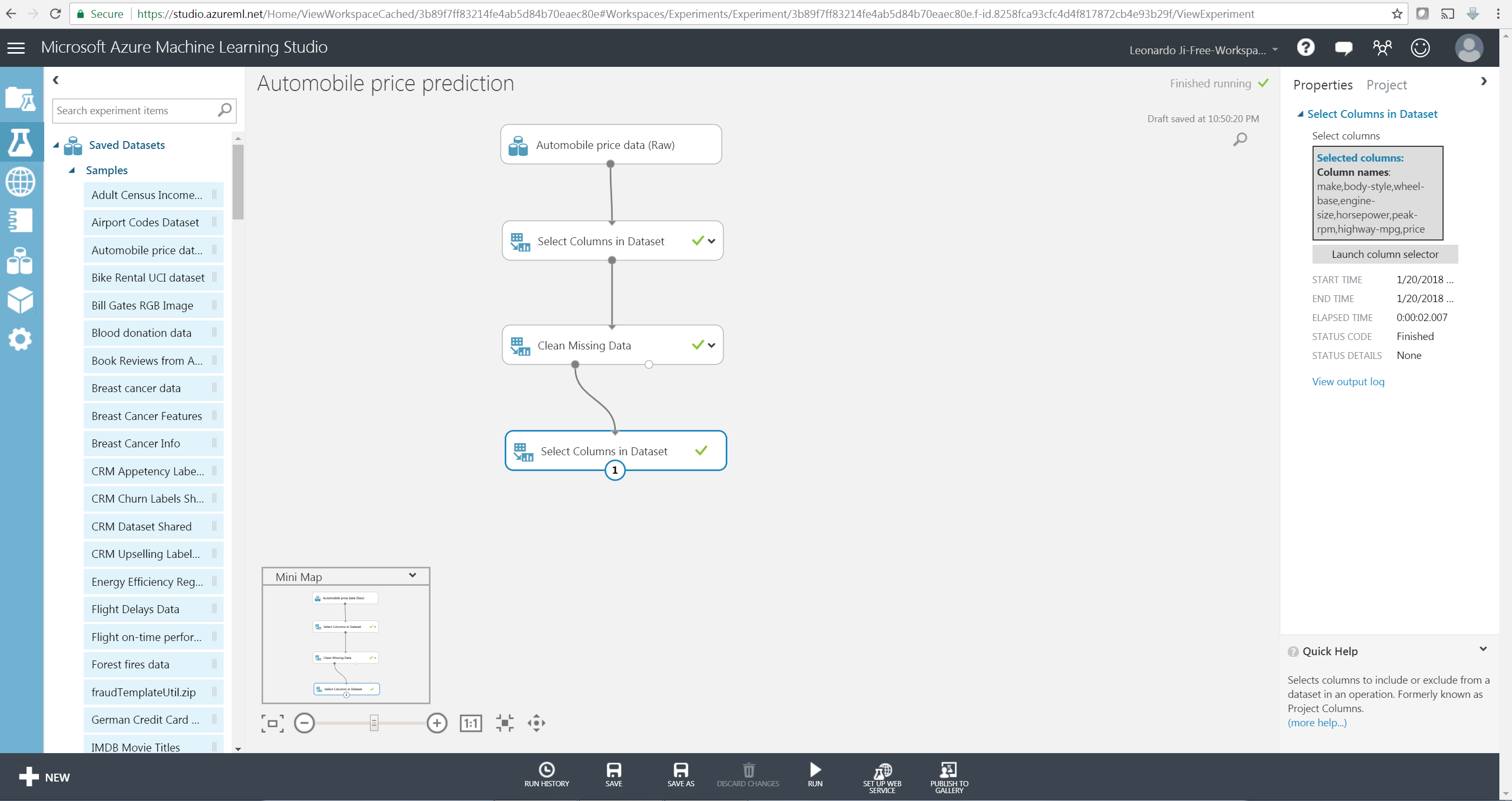
The weakness of the clustering function is it is not semantically-aware and it produces lots of false positive. It only finds words with similar alphabetical structure. It suggests merging “HONDA” and “HYUNDAI” even though these two cars are clearly different manufacture. Also, it would not work for other character set based languages.

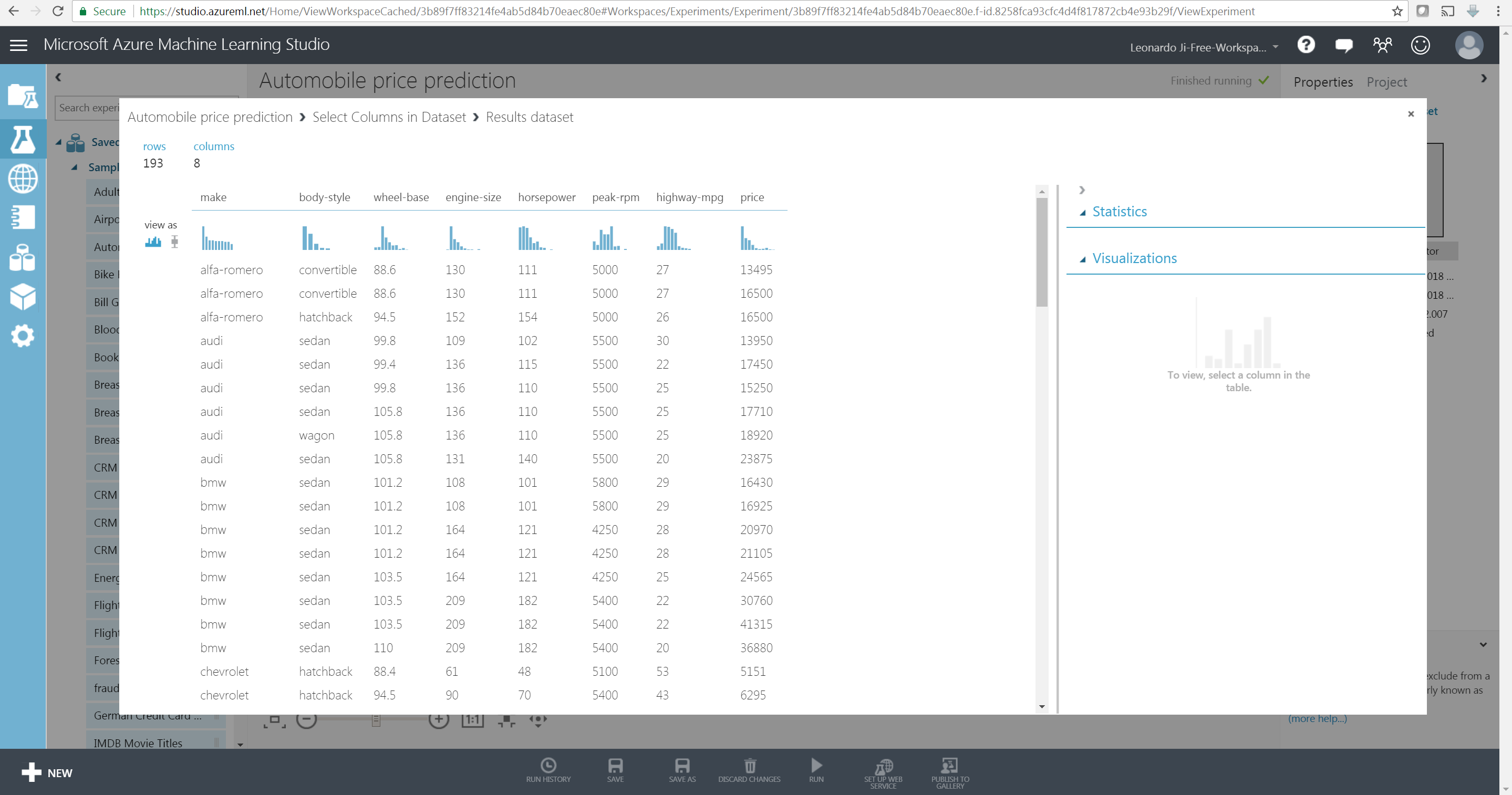
**Part 3 Azure ML Studio**

**Run through the first 3 steps in this tutorial:**

**https://docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment**

**Screen Shot the Final Result.**





After step 3 we have 193 rows and 8 columns of data.

**Part 4 Reading Assignment: Big Data Architecture Answer Questions 3, 8, 9,17 (Page 58)**

**3. What are 4V of Big Data?**

Volume, Variety, Velocity, Veracity or data quality.

Volume refers amount of data. Social media and wearable devices created large volume of data.

Variety refers structure or format of the data. The examples are XML, JSON, audio, video, text, or scanned document.

Velocity refers the speed data flow into the system. Velocity increases with number of data sources and number and speed data is produced from mobiles and sensors.

Veracity refers the quality of data. Human entered data typically have errors. Sensors might also produce incorrect data without calibration, and data could be lost during transmission.

**8. What forms of Data are called interaction data?**

Data generations from human interaction is called interaction data. For example, social media like Twitter generates lots of data from each tweet (user, location, time, topic, number of views…etc.).

**9. What forms of data are called observation data?**

Observation data are those generated by wearable devices or Internet of things (IOT) devices. These devices generate data frequently, and its velocity and volume dwarfs those interaction data or transaction data.

**17. Discuss an application of Big Data architecture.**

There are three components of Big Data architecture: Data Lake or Data Platform, Data Product (BI, Analytics dashboards, EDW), Data R&D (Data Analytics from Big Data). An integrated analytics application requires all three components of the Big Data architecture. For example, a ride sharing company wants to have a Big Data architecture platform supports not only a booking app, but also customer billing records, vehicle maintenance records, driver safety records, driver location and routes (CRM, ERP, SCM). Those applications require data from other places besides Data Lake. Integrated analytics applications can support all these by combining Data Lake and Data Product. This Big Data architecture would require support transporting and loading data in parallel (streaming). Client tools should have access to all three data platforms. Data connections between all three data storage platforms available.

**Part 5 Create a Databricks Trial Account**

