**MongoDB installation:**

[https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/#get-mongodb-community-edition (Links to an external site.)Links to an external site.](https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/#get-mongodb-community-edition)

Note: If you have version 3.2 installed, then upgrade to 3.4 first and then to 3.6. (I was working with MOngoDB 3.2 and wanted to upgrade to the latest version 3.6. But, an error was shown while running the mongod.exe saying all the data must be upgraded to version 3.4 first. So, installed version 3.4 first, cleaned the /data/db folder and then installed 3.6 again). The link to the installation guide is below:

**Loading and accessing our dataset into MongoDB:**

**Data source**: <https://www.kaggle.com/nsharan/h-1b-visa>

Command to load csv file into MongoDB:

"C:\Program Files\MongoDB\Server\3.6\bin\mongoimport" --db g5 --collection h1b --type csv --headerline --file "C:\Users\jaide\Downloads\4th Sem\INFO-I590 - SQL NoSQL\Project\h1b\_kaggle.csv"

database name (--db): g5  
collection (--collection): h1b  
file type (--type): csv  
file path (--file): path to the downloaded file.

To access the data, first switch to your desired database:

db.adminCommand( { listDatabases: 1 } ); // lists all databases

use g5; //switches to database g5

db.getCollectionNames(); // lists all collections (tables)

db.h1b.findOne(); //fetches first row or doc from collection h1b

**Python with MongoDB:**

We are using the python library**PyMongo:**

Command to install PyMongo: pip install pymongo

Guide: [https://docs.mongodb.com/getting-started/python/client/ (Links to an external site.)Links to an external site.](https://docs.mongodb.com/getting-started/python/client/)

Acessing data using **pyMongo** and **pandas** from **MongoDB**:

from **pymongo** import **MongoClient**

client = MongoClient('mongodb://localhost:27017/') # connect to MongoDB localhost

db = client.g5 # g5 is our database  
coll = db.h1b #h1b is our collection in g5

**Cleaning** the data:

import **pandas** as pd

h1b\_data = pd.DataFrame(list(coll.find({}))) # load the entire data from collection h1b

h1b\_data.isnull().sum() # check for null values. We founf none in our dataset

h1b\_data['CASE\_STATUS'].unique() # displays all unique values of Case\_Status

h1b\_data[h1b\_data['CASE\_STATUS']=="NA"] # displays all rows with Case\_Status = "NA"

We found many rows with "NA" values which won't help our analysis, hence need to be deleted (cleaned).

h1b\_data = coll.delete\_many({"CASE\_STATUS": "NA"}) # Command to delete rows with NA values from mongoDB collection.

h1b\_data[h1b\_data['CASE\_STATUS']=="NA"] # should show no rows with NA values.