Through this assignment I have got the access to a great learning experience as the assignment covered everything from the beginning to the end of the module. The way it has pushed me to the limits in designing data modeling by writing my own python scripts, creating database, performing queries, modelling nosql database, I felt it as a great preparation towards my next module big data and also towards professional career after studies. Since the dataset was a real time one, the more I gained knowledge theoretically, I performed better practically.

Coming to assignment, I was given a comma separated file (CSV) of the air quality readings of 18 stations of this beautiful city, Bristol with dates ranging from 2004 to 2021. The first task was about cropping and cleaning, where the data before Jan 1st, 2010 was cropped off for our further analysis. Then I created a dictionary holding station id and names, and using that I cleaned the cropped data by comparison. Also, if there was any mismatched data, then it was removed too using the for loop. In this task I faced problem with delimiter ‘;’ and used ‘,’ for my results.

In next task, I was asked to create an Entity-relationship(E-R) model for the dataset. For this I normalized the data and broke it into 3 tables named ‘readings’, ‘stations’ and ‘schema and created the E-R diagram. Then I did forward engineering and created database 'pollution\_db' by this e-r diagram. As I used ‘MySQLWorkbench’ tool for this task, I faced some problems connecting with database and also with forward engineering.

In 3rd task, same thing of second task was to be done but by writing python scripts. I used sqlalchemy and with the help of pandas created database with tables 'readings', 'stations' & 'schema' and populated with cleaned data extracted from 1st task. I used ‘mariadb’ to connect my scripts with the database. Then next part of this task is to create a csv file holding 100 entries to be inserted into readings table. Initially I faced issue in connecting to ‘mariadb’ &and ‘mysql.connector’ with finally settling with ‘mariadb’ which somehow worked. Since ‘MySQLWorkbench’ was not working then

I used ‘phpmyadmin’ to view the database created by running my python scripts.

Then in 4th task, we have to run 3 queries where 1st query is about finding the station name with the highest Nox value for year 2019. In 2nd query, we have to find the average mean of `PM2.5` & `VPM2.5` for every station for year 2019 & in 3rd query is same but time period is 2010 to 2019. Again ‘MySQLWorkbench’ was tried initially but no success and went back to ‘phpmyadmin’.

In next task, a NoSQL model for a station is to be created, implemented & queried, where I choose 'Rupert Street' and did requirements. The software I choose was MongoDB and created model using ‘key-value’ modelling. Whatever I did was included as images in the report.

In conclusion, I learned and experiences all the important data modelling structures and techniques that I would definitely implement in the near future, and learned more about NoSQL databases and their uses in various industries. As I advance to the optional module as big data in the next semester, I will learn more about the teachings and insights of this module and perform the same assignments to perform more analytical tasks.

127.0.0.1/pollution\_db2/ http://localhost/phpmyadmin/index.php?route=/database/sql&db=pollution\_db2

Showing rows 0 - 0 (1 total, Query took 58.2593 seconds.)

SELECT readings.Datetime , readings.NOx as HighestNOx ,stations.Location

FROM readings, stations

WHERE NOx=(SELECT MAX(NOx) FROM readings WHERE year(`DateTime`)="2019")

AND readings.SiteID = stations.SiteID;

2019-01-24 09:00:00+00:00 1403.5 Colston Avenue