Statistical Methods for Big Data (MATH70072)
Coursework Assignment

Introduction

This is the coursework assignment associated with the Big Data in Statistics module. A report response is to be submitted by **1800hrs on 8th April 2022**. An electronic copy of the report (in Word, PDF or iPython notebook format) should be submitted via the module's Blackboard page.

Data

The dataset is a modified version of the VAST 2016[1] Mini-Challenge 1 data. Some information taken from the VAST 2016 website is repeated here, to ensure that these instructions are self-contained.

At the end of 2015, a [fictitious] growing organisation, GAStech, moved into a new, state-of-the-art, three-story building near to their previous location. The new office is built to the highest energy efficiency standard, but as with any new building, there are still several heating, ventilation, and air conditioning (HVAC) issues to work out. The building is divided into several HVAC zones. Each zone is instrumented with sensors that report building temperatures, heating and cooling system status values, and concentration levels of various chemicals such as Carbon Dioxide (abbreviated CO2) and Hazium (abbreviated Haz), a recently discovered and possibly dangerous chemical. CEO Sten Sanjorge Jr. has read about Hazium and requested that these sensors be included. However, they are very new and very expensive, so GAStech can afford only a small number of sensors.

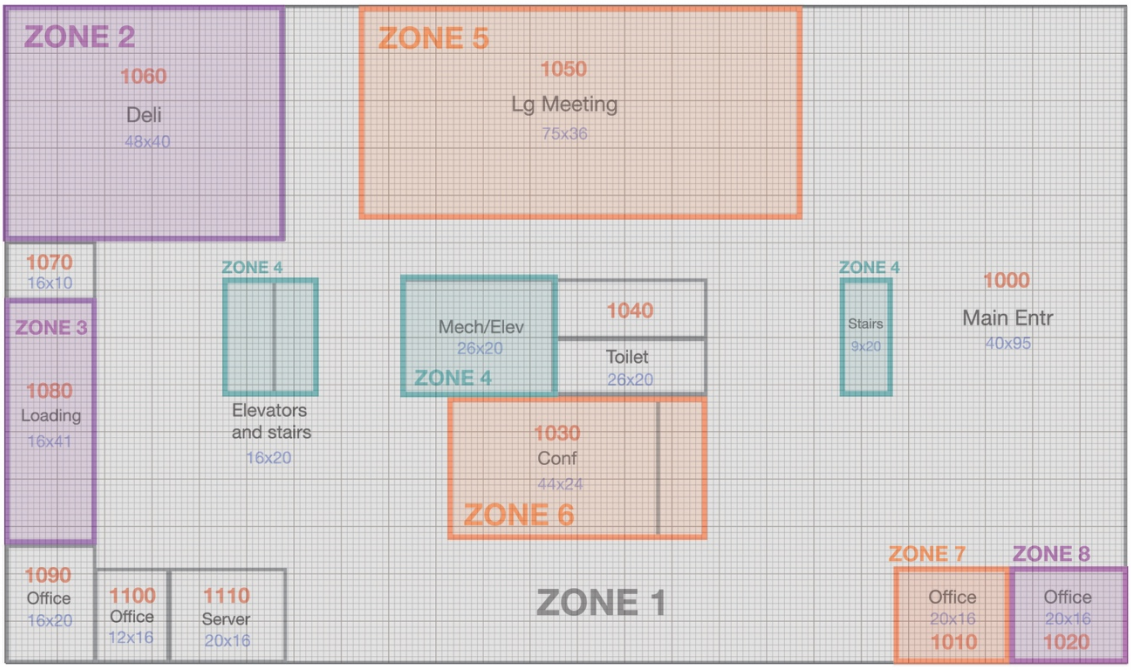
With their move into the new building, GAStech also introduced new security procedures, which staff members are not necessarily adopting consistently. Staff members are required to wear proximity (prox) cards while in the building. The building is instrumented with passive prox card readers that cover individual building zones. The prox card zones do not generally correspond with the HVAC zones. When a prox card passes into a new zone, it is detected and recorded. As part of the deal to entice GAStech to move into this new building, the builders included a free robotic mail delivery system. This robot, nicknamed Rosie, travels the halls periodically, moving between floors in a specially designed chute. Rosie is equipped with a mobile prox sensor, which identifies the prox cards in the areas she travels through.

The building is partitioned into different zones, across three floors, as depicted in the three figures below.

[1] http://vacommunity.org/VAST+Challenge+2016

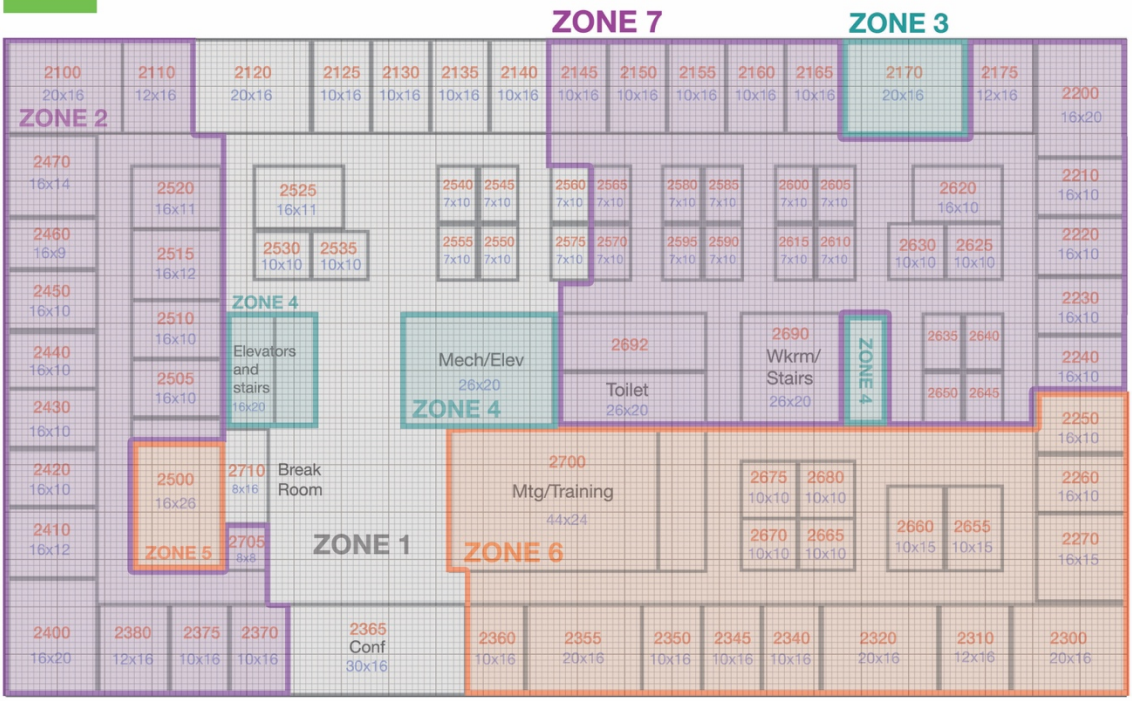
Floor 1

Proximity Zones

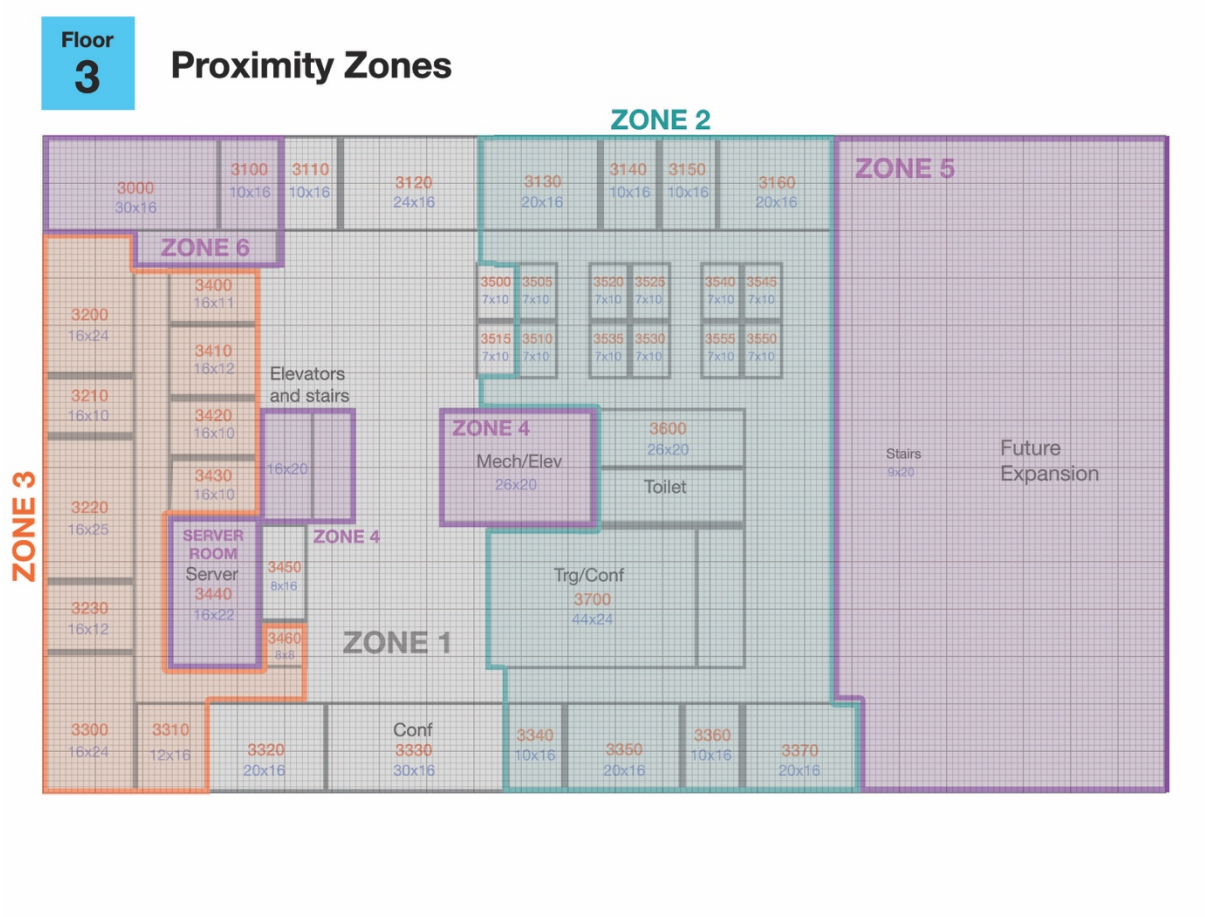


Floor 2

Proximity Zones



*Any room that is not labeled is an office



There are four datasets provided, covering May 31 to June 13, 2016. The data are as follows:

- Fixed proximity sensor data reading employees' prox cards (prox-fixed.csv);
- Mobile proximity sensor data (from Rosie) reading employees' prox cards (prox-mobile.csv);
- Environmental conditions of the building (bldg-measurements.csv) – see Annex A for further details;
- Hazium concentration within the building (f1z8-haz.csv), containing the Hazium concentration on floor 1, zone 8.

Acquiring the data

These instructions assume that you have successfully completed Exercise 1 of Week 1. If you have not done so then please complete this exercise before proceeding with the coursework.

Please log on to bazooka. A unique dataset is to be generated for each student, using the following commands:

```
$ cd ~/bd-sp-2017
$ cd coursework
$ chmod +x *.py
$ ./process_data.py /tmp/coursework/prox-fixed.csv \
/home/USERNAME/bd-sp-2017/data/prox-fixed.csv
$ ./process_data.py /tmp/coursework/prox-mobile.csv \
/home/USERNAME/bd-sp-2017/data/prox-mobile.csv
$ ./process_data.py /tmp/coursework/bldg-measurements.csv \
/home/USERNAME/bd-sp-2017/data/bldg-measurements.csv
$ ./process_data.py /tmp/coursework/f2z2-haz.csv \
/home/USERNAME/bd-sp-2017/data/f2z2-haz.csv
```

```
$ cd ../data
$ ls –la
```

Note that USERNAME will need to be replaced with your actual username.

You should see four new files in the data directory corresponding to the four data files (prox-fixed.csv, prox-mobile.csv, bldg-measurements.csv, f2z2-haz.csv). Please run the following commands and record the output of each command at the top of your coursework report submission.

```
$ md5sum prox-fixed.csv
$ md5sum prox-mobile.csv
$ md5sum bldg-measurements.csv
$ md5sum f2z2-haz.csv
```

Create a folder in HDFS called `coursework`. You should now upload these four data files to your coursework folder on HDFS.

Map Reduce

For questions 1-4 below, write a Map Reduce program to compute the required answer. Your response to each of these questions should consist of three components: (1) your answer to the question; (2) the Shell command used to execute the Map Reduce program; (3) Python code developed and used to compute the answer. The code will be checked for execution quality, so please ensure that the code is self-contained and executable. (Marks will be deducted for code that does not execute using the commands provided via component (2).)

1. Using both prox-fixed and prox-mobile datasets, produce a diagram that displays the number of staff members present in the building on each day (i.e. number of unique prox-ids on each day)? NB: The x-axis may be marked with day number (i.e. 0, 1, 2, …) from the beginning of the dataset. [8 marks]
2. Using the prox-fixed dataset, what is the (floor, zone) of the most visited location in the building? [5 marks]
3. Using both datasets, what is the prox-ID of the most active staff member (i.e. the staff member with the greatest number of prox card readings) on 2nd June 2016? [5 marks]
4. Using the bldg-measurements dataset, produce a time series plot of the average hourly "Total Electric Demand Power". (This should be a single plot, with the x-axis denoting hour of day, with a range of 0hrs-23hrs.) What does this plot indicate about power usage throughout the day? [5 marks]

Spark

With the exception of question 8, for the following questions, write a sequence of Spark commands (that are executed in the Spark REPL) to compute the required answer. For each question, the full sequence of Scala commands should be pasted into your submission, together with the computed answer, and any other information requested. The code will be checked for execution quality, so please ensure that the code is self-contained and executable. (Marks will be deducted for code that does not execute using the sequence of commands provided in your coursework submission.)

5. Parse the prox-fixed.csv data fie into an RDD[ProxReading], where ProxReading is defined as:
   `case class ProxReading(timeStamp: org.joda.time.DateTime, id: String, floorNum:`

`String, zone: String)`. In this class, `timestamp` corresponds to a joda DateTime object[2], `id` corresponds to prox-id, `floorNum` corresponds to the floor number, `zone` corresponds to the zone id. [2 marks]

6. Using the prox-fixed dataset, what is the (floor, zone) of the most visited location in the building across the complete dataset? [3 marks]
7. Using both datasets, what is the prox-ID of the most active staff member (i.e. the staff member with the greatest number of prox card readings) on 7th June 2016? [3 marks]
8. Provide a concise summary of your experiences writing Map Reduce programmes and Spark commands for questions 6 and 7. Comment on the differences between the two computational platforms. [2 marks]
9. Construct an appropriate RDD from the bldg-measurements.csv data, containing the "Date/Time" column (number 1) and "F_2_Z_1 VAV REHEAT Damper Position" column (number 193). What is the date and time of the first occurrence of the F_2_Z_1 VAV REHEAT Damper Position being fully open, i.e. the earliest date and time of variable "F_2_Z_1 VAV REHEAT Damper Position" being set to its maximum value of 1.0? [3 marks]
10. A rogue employee is believed to be increasing the Hazium concentration in the building by modifying the Reheat Damper position ("F_2_Z_1 VAV REHEAT Damper Position"). By using the Spark package MLlib[3] or other Spark command sequence, demonstrate a statistical association between the Hazium concentration (from f2z2-haz.csv) and the "F_2_Z_1 VAV REHEAT Damper Position" variable. Provide a concise summary of your statistical findings, using diagrams where appropriate.
    [10 marks]
11. By using the (fixed) proximity location data, determine the employee IDs for those that entered the Server Room (the HVAC control location) prior to the sudden increase in Hazium concentration at the end of the dataset (i.e. employees in the Server Room on 10th June 2016). [3 marks]

General

12. Write a question, that could appear in next year's coursework paper, which tests a student's understanding of the opportunities and problems with using Big Data technology. [10 marks]
13. Identify, download, and perform a statistical analysis of any suitable data that is available on the internet, and write a one-page summary of your findings. Your analysis should use Hadoop, Spark, or both tools. *Please note that the data need not be "big" – the question is intended to assess your approach to the analysis, and how you utilise Big Data technology in performing a statistical analysis.* [20 marks]
14. Please read the following research paper:

    https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf

    a. Write a short (less than one side of A4) synopsis of the paper, extracting the key statistical contributions of the paper. [15 marks]
    b. Discuss how the key points raised in the paper could be relevant (if at all) to the statistical analysis performed in question 13, linking to other research if and where appropriate. [25 marks]

[2] http://joda-time.sourceforge.net/apidocs/org/joda/time/DateTime.html
[3] https://spark.apache.org/docs/1.2.1/mllib-guide.html

Annex A – Building measurement information

Field	Units	Description
F_#_BATH_EXHAUST:Fan Power	[W]	Power used by the bathroom exhaust fan
F_#_VAV_SYS AIR LOOP INLET Mass Flow Rate	[kg/s]	Total flow rate of air returning to the HVAC system from all zones it serves
F_#_VAV_SYS AIR LOOP INLET Temperature	[C]	Mixed temperature of air returning to the HVAC system from all zones it serves
F_# VAV Availability Manager Night Cycle Control Status		On/off status of the HVAC system during periods when the system is normally scheduled off. The night cycle manager cycles the HVAC system to maintain night and weekend set point temperatures.
F_#_VAV_SYS COOLING COIL Power	[W]	Power used by the HVAC system cooling coil
F_#_VAV_SYS HEATING COIL Power	[W]	Power used by the HVAC system heating coil
F_#_VAV_SYS SUPPLY FAN OUTLET Mass Flow Rate	[kg/s]	Total flow rate of air delivered by the HVAC system fan to the zones it serves
F_#_VAV_SYS SUPPLY FAN OUTLET Temperature	[C]	Temperature of the air exiting the HVAC system fan
F_#_VAV_SYS SUPPLY FAN:Fan Power	[W]	Power used by the HVAC system fan
F_#_VAV_SYS Outdoor Air Flow Fraction		Percentage of total air delivered by the HVAC system that is from the outside
F_#_VAV_SYS Outdoor Air Mass Flow Rate	[kg/s]	Flow rate of outside air entering the HVAC system
COOL Schedule Value		The supply air temperature set point. Air exiting the HVAC system fan is maintained at this temperature during cooling operation
DELI-FAN Power	[W]	Power used by the deli exhaust fan
Drybulb Temperature	[C]	Drybulb temperature of the outside air
Wind Direction	[deg]	Direction of wind outside of the building

Wind Speed	[m/s]	Speed of wind outside of the building
HEAT Schedule Value		The supply air temperature set point. Air exiting the HVAC system fan is maintained at this temperature during heating operation
Pump Power	[W]	Power used by the hot water system pump
Water Heater Setpoint		Water heater set point temperature
Water Heater Gas Rate	[W]	Rate at which the water heater burns natural gas
Water Heater Tank Temperature	[C]	Temperature of the water inside the hot water heater
Loop Temp Schedule		Temperature set point of the hot water loop. This is the temperature at which hot water is delivered to hot water appliances and fixtures.
Supply Side Inlet Mass Flow Rate	[kg/s]	Flow rate of water entering the hot water heater
Supply Side Inlet Temperature	[C]	Temperature of the water entering the hot water heater
Supply Side Outlet Temperature	[C]	Temperature of the water exiting the hot water heater
F_#_Z_# REHEAT COIL Power	[W]	Power used by the zone air supply box reheat coil
F_#_Z_# RETURN OUTLET CO2 Concentration	[ppm]	Concentration of C02 measured at the zone's return air grille
F_#_Z_# SUPPLY INLET Mass Flow Rate	[kg/s]	Flow rate of the air entering the zone from its air supply box
F_#_Z_# SUPPLY INLET Temperature	[C]	Temperature of the air entering the zone from its air supply box
F_#_Z_# VAV REHEAT Damper Position		Position of the zone's air supply box damper. 1 corresponds to fully open, 0 corresponds to fully closed
F_#_Z_#: Equipment Power	[W]	Power used by the electric equipment in the zone
F_#_Z_#: Lights Power	[W]	Power used by the lights in the zone

F_#_Z_#: Mechanical Ventilation Mass Flow Rate	[kg/s]	Ventilation rate of the zone exhaust fan
F_#_Z_#: Thermostat Temp	[C]	Temperature of the air inside the zone
F_#_Z_#: Thermostat Cooling Setpoint	[C]	Cooling set point schedule for the zone
F_#_Z_#: Thermostat Heating Setpoint	[C]	Heating set point schedule for the zone
Total Electric Demand Power	[W]	Total power used by the building
HVAC Electric Demand Power	[W]	Total power used by the building's HVAC system including coils, fans and pumps.