Python project:

Azure VM Workload Data Analysis

Group members:

Student Name	Student ID
Dana Ghazal	0183507
Baraa AbuAsfar	0185264
Aya Al-Ali	0189980

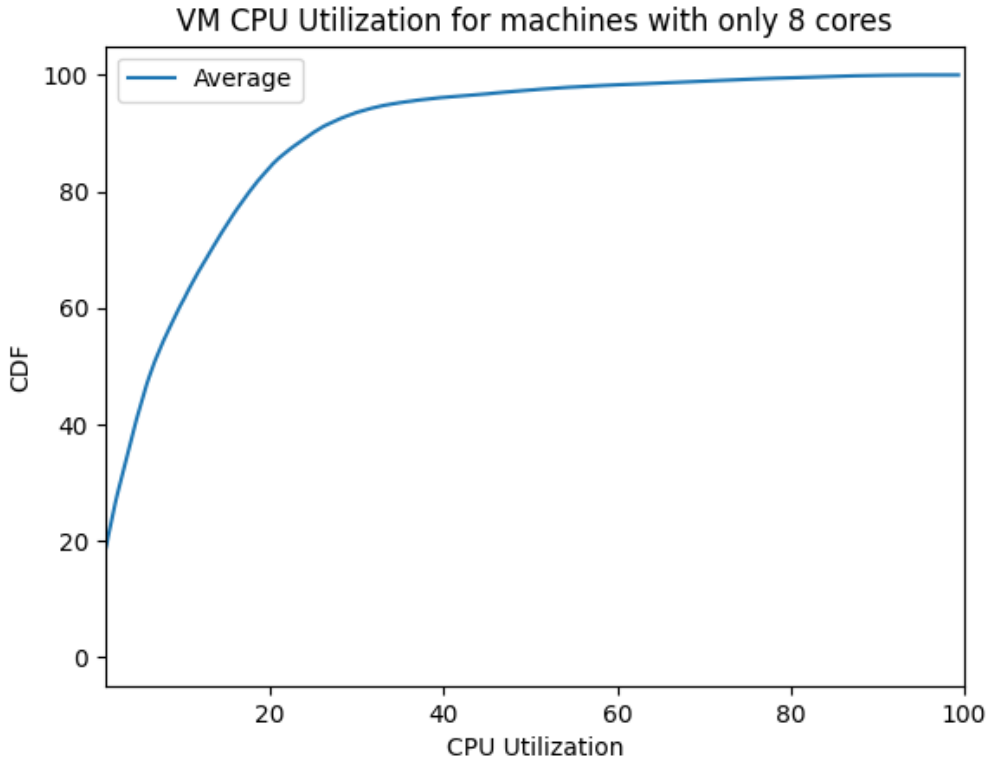
Brief Description

Our project is about Virtual Machines and it is a host computer emulates guest operating system and machine resources, and examples of Virtual Machines

Such as Windows virtual PC, IBM VM/370, VMWare.

In this project we use Azure that is a public cloud computing platform, and we analyzed the dataset that deployed on it.

1) The CDF for average CPU utilization for machines with only 8 cores

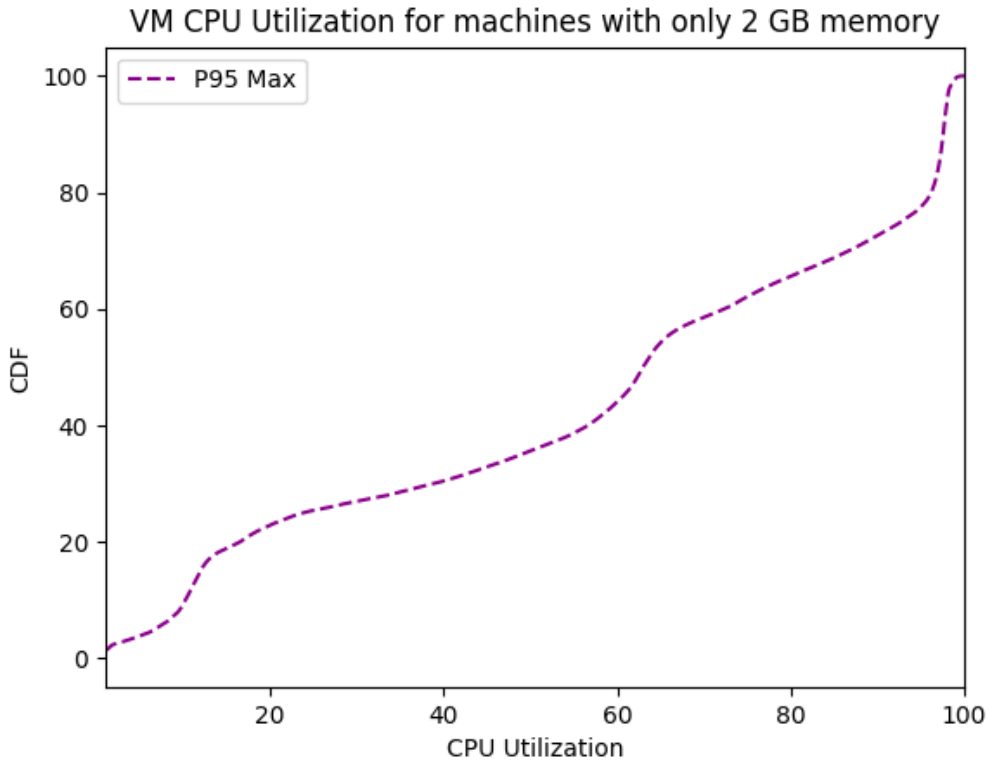


This is a graph that represents average CPU utilization for machines with only 8 cores. The X axis represents CPU utilization and the Y axis represents CDF, the relationship between X AND Y is average, the average as we see is increasing rapidly then slowly ,it even starts with stability when its value of CPU equal 90.

The highest value for CDF = 99 when CPU >= 98

The lowest value for CDF = 20 when CPU =0

2) The CDF for P95 CPU utilization for machines with only 2 GB memory

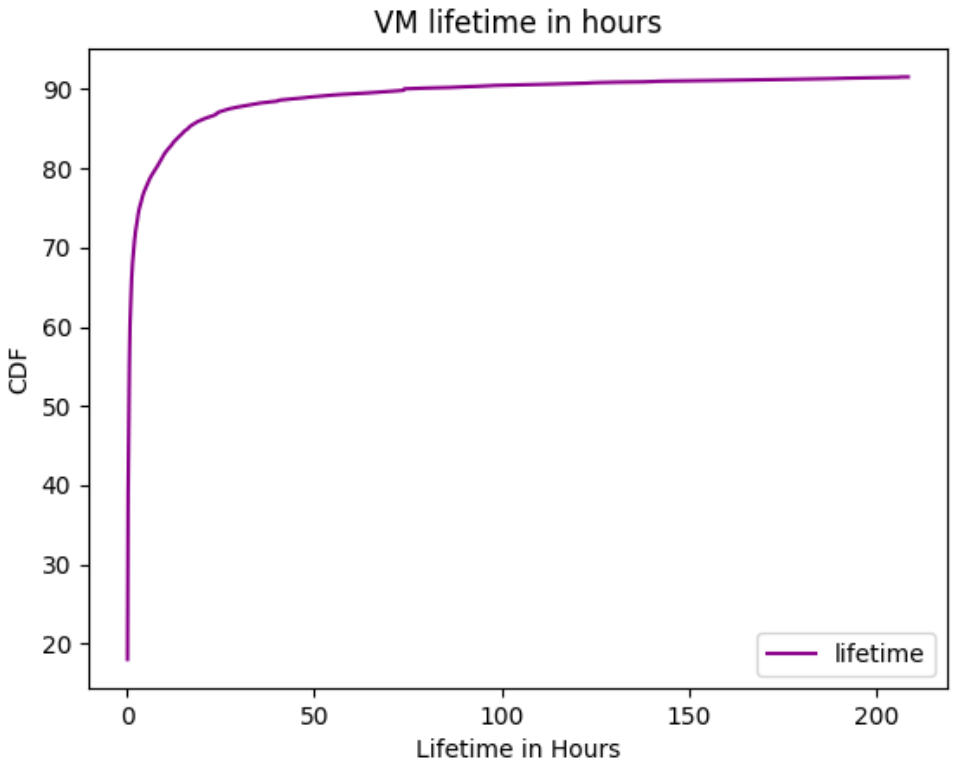


This is a graph that represents P95 CPU utilization for machines with only 2 GB memory. The X axis represents CPU utilization and the Y axis represents CDF, the relationship between X AND Y is P95 Max. The relationship is clearly non-linear, it's a growing zigzag line.

The highest value for CDF = 99 when CPU = 100

The lowest value for CDF = 0 when CPU =0

3) The CDF for lifetime in hours

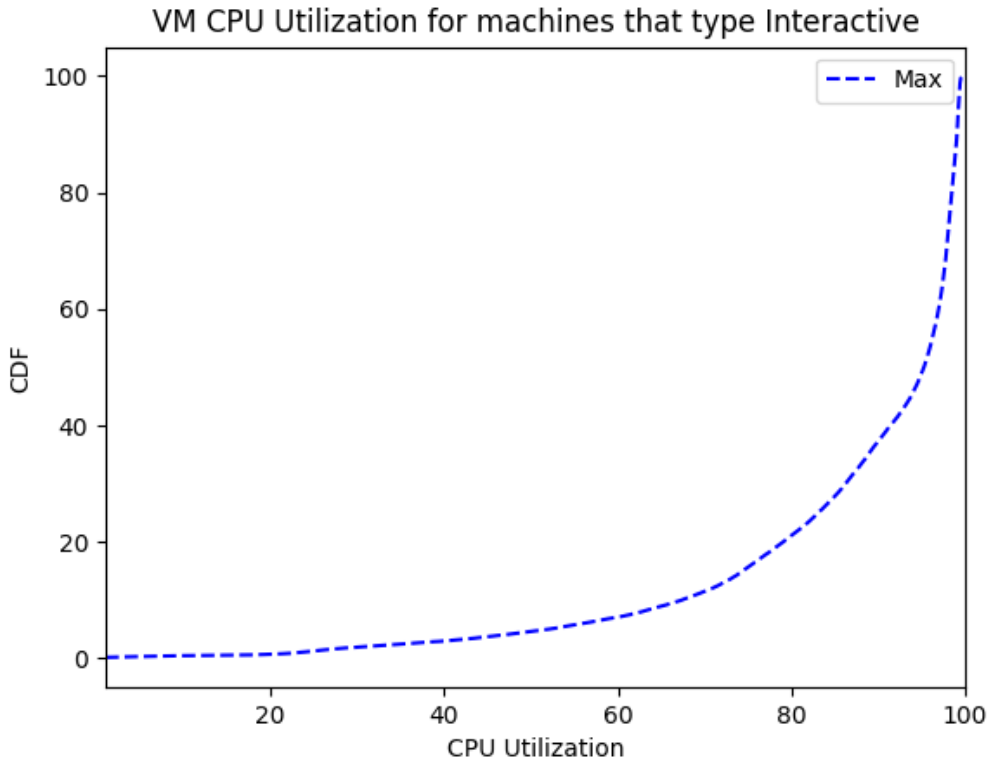


This is a graph that represents lifetime in hours. The X axis represents lifetime in hours and the Y axis represents CDF, the relationship between X AND Y is lifetime. The relationship starts increasing constantly with CDF (CPU=0) then the CPU value starts increasing and CDF begin to slow down when CPU reached 70.

The highest value for CDF = 90 when CPU = 200

The lowest value for CDF = 20 when CPU =0

4) The CDF for max CPU utilization for machines that type Interactive

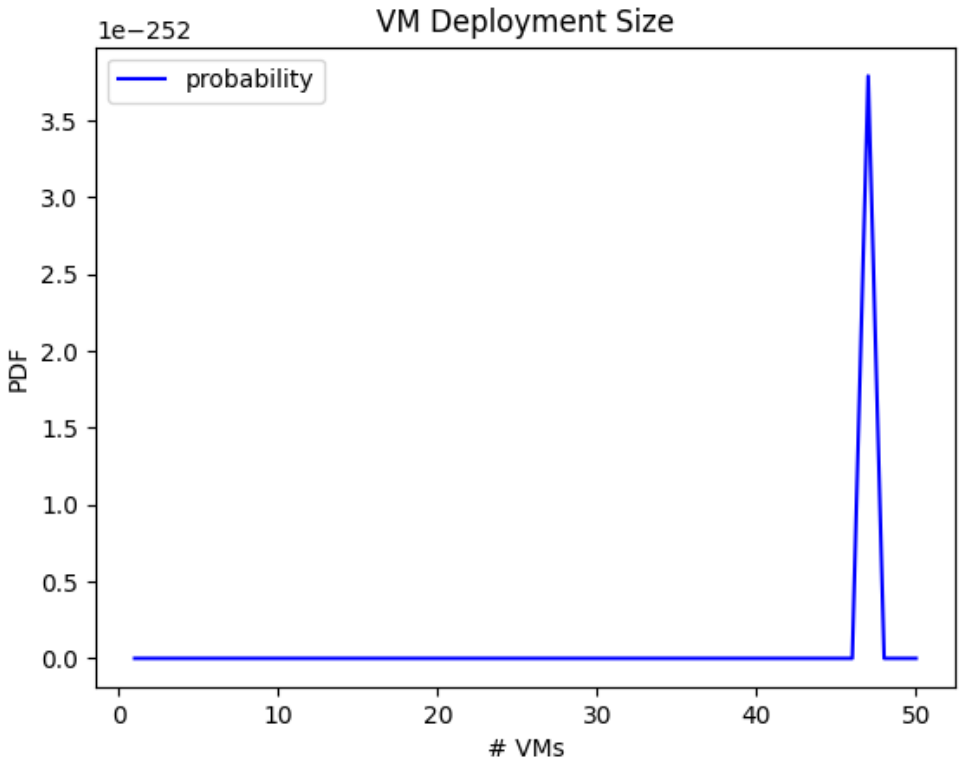


This is a graph that represents max CPU utilization for machines that type Interactive. The X axis represents CPU utilization and the Y axis represents CDF, the relationship between X AND Y is MAX. The relationship starts increasing but too slowly then starts increasing rapidly when CPU value = 50.

The highest value for CDF = 99 when CPU = 100

The lowest value for CDF = 0 when CPU <=19

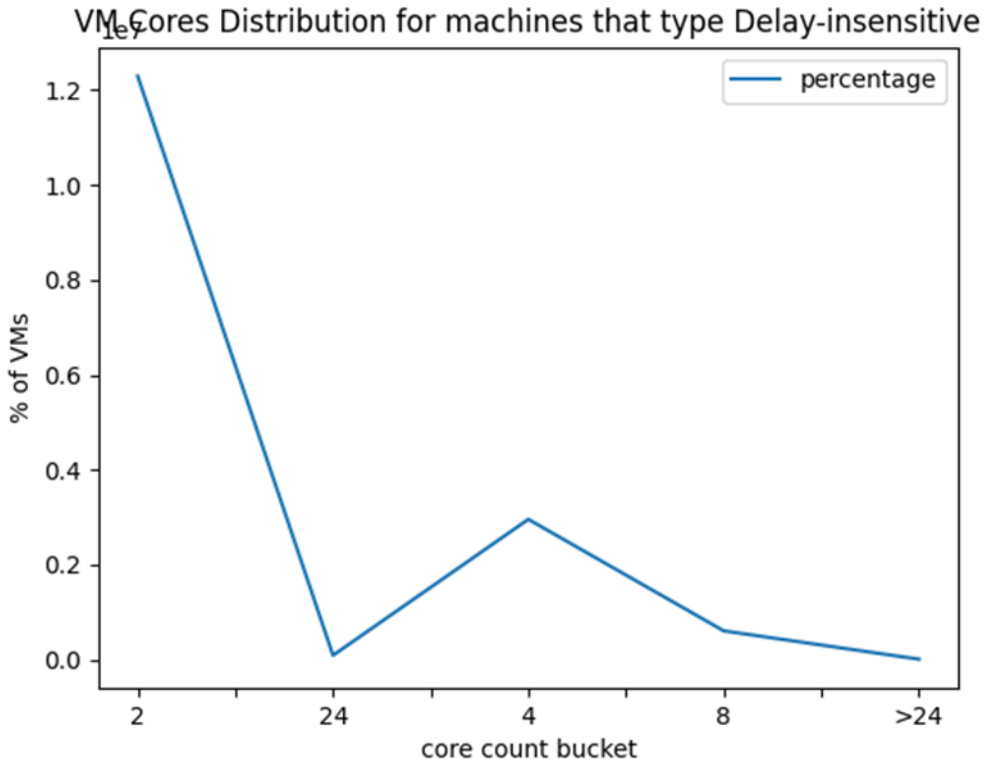
5) The PDF for deployment size



This is a graph that represents for deployment size, the Y axis represents PDF and the X axis represents the number of VMs, the relationship between Y and X is probability, the probability as we see the equilibrium is constant beyond the 40's, and then it rose as high as possible, and then it is balanced.

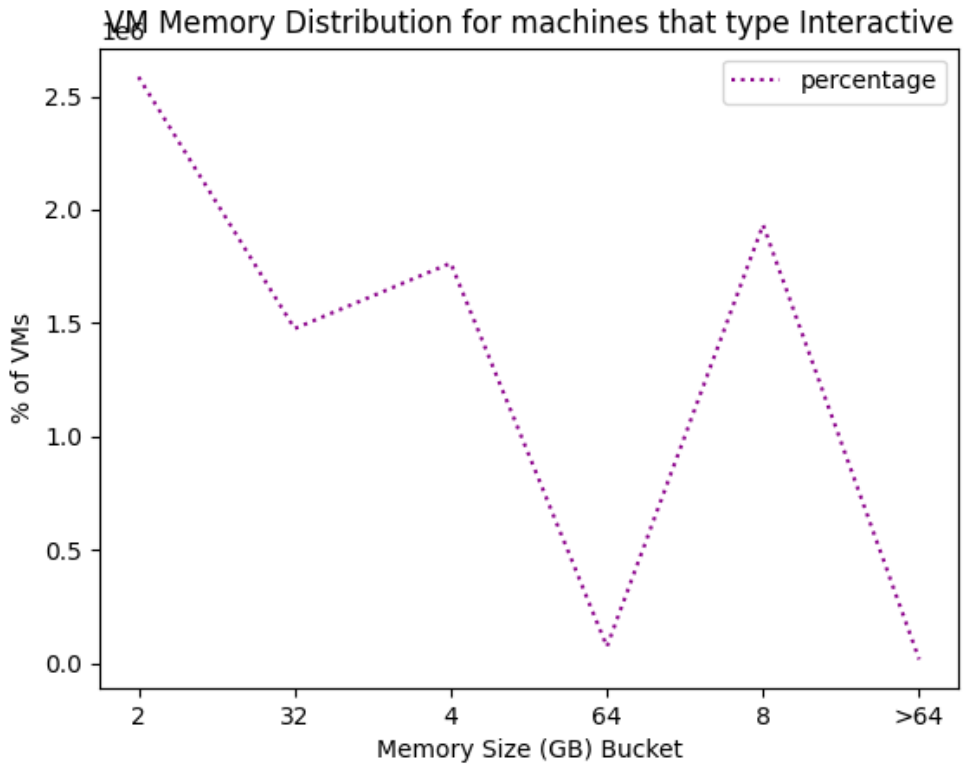
The highest value for PDF is more than 3.5 when the number of VMs is more than 40 and less than 50.

6) The percentage of Cores Distribution for machines that type Delay-insensitive



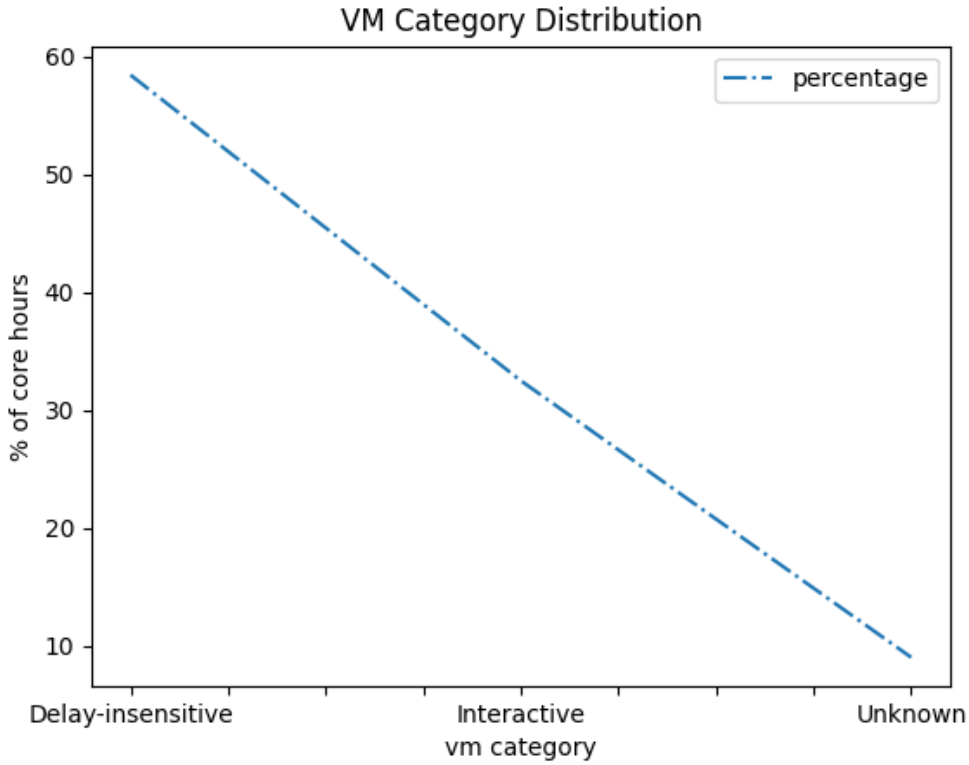
This is a graph that represents for the percentage of Cores Distribution for machines that type Delay-insensitive, the Y axis represents the percentage of VMs and the X axis represents core count bucket, the relationship between them is percentage of how many cores selected by VM creators when they created their virtual machine, the probability as we see, the highest value for percentage of VMs is 1.2% for the VMs that have 2 cores, and the lowest percentage of VMs for the core count 24 and greater than 24.

7) The percentage of Memory Distribution for machines that type Interactive



This graph that represents for the percentage of Memory Distribution for machines that type Interactive, the Y axis represents a percentage of VMs and the X axis represents memory size (GB) bucket, the relationship between them is percentage of how much memory was selected by VM creators when they created their virtual machine, the highest value for percentage of VMs is 2.5% for memory size 2 GB, and the lowest value for percentage of VMs for memory size 64 and greater than 64 GB.

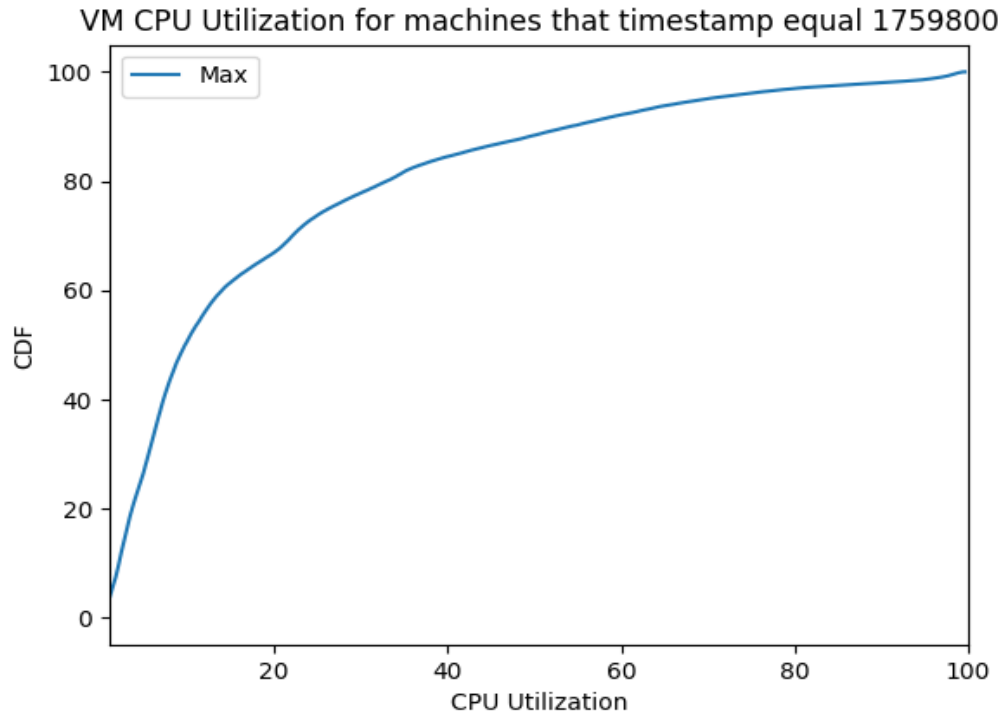
8) The percentage of Core hours for Category Distribution



This graph represents for the percentage of Core hours for Category Distribution, the Y axis is representing a percentage of core hours and the X axis represents a VMs category, the relationship between them is a percentage of core hours for each type. The highest value of percentage core hours is 60% for Delay-insensitive type, and the lowest value is 10% for the Interactive and unknown types.

The plots for the file that from 195 vm_cpu_readings files.

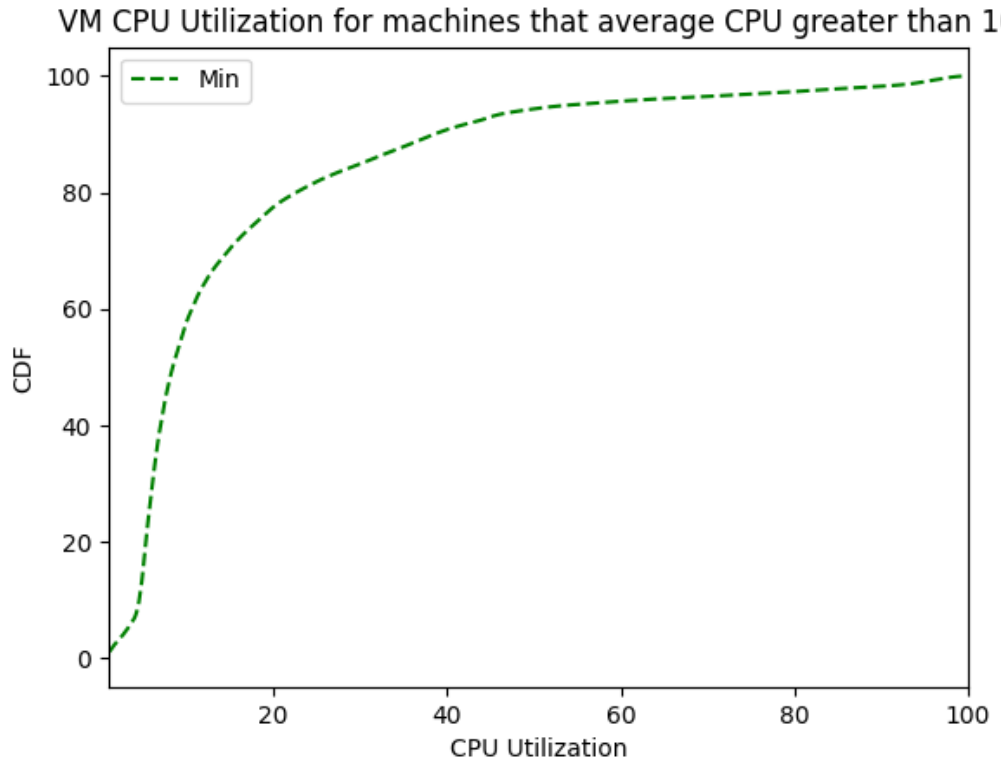
1) The CDF for max CPU utilization for machines that timestamp equal 1759800



This graph represents The CDF for max CPU utilization for machines that timestamp equal 1759800 and the Y axis represents (CDF) and the X axis represents CPU utilization, the relationship between them is max and as we see it starts increasing rapidly until it reached 100.

The lowest value is 0 at X axis and in Y axis is between 0 and 20 and the highest value is on 100 at X and Y axis.

2) The CDF for min CPU utilization for machines that average CPU greater than 10



This graph represents The CDF for min CPU utilization for machines that average CPU greater than 10 and the Y axis represents (CDF) and the X axis represents CPU utilization, the relationship between them is min and as we see it starts increasing rapidly clearly non-linear until it reached 100.

The lowest value is 0 at X axis and in Y axis is between 0 and 20 and the highest value is on 100 at X and Y axis.

Conclusion and references

We were able to apply almost everything we learned in this course in our project that about Virtual Machines we included many aspects such as Lifetime, Memory, CPU, Category, Deployment, and Cores.

In this project we used three files from Azure cloud and they are:

1) vmtable.csv.gz: the number of entries 2695548 rows, and the number of fields 13 columns.

2) deployments.csv.gz: the number of entries 33205 rows, and the number of fields 2 columns.

3) vm_cpu_readings-file-133-of-195.csv.gz: the number of entries 10000000 rows, and the number of fields 5 columns.