

Performance Evaluation Of Terapixel Rendering in Cloud(Super) Computing Project

Cahyadi/200511881

1/23/2021

Introduction

Cloud technology, a relatively new technology that is based on utilising the internet infrastructure to the fullest has seen a major increase in popularity in recent years. This success could be attributed to the huge growth of the Tech Industries that seen many new Unicorns; a term first mentioned by Aileen Lee on 2013, used to describe privately owned start-up companies that valued more than 1 billion USD. One of many reasons why the emergence of these “Unicorns” of Tech world brought popularity to Cloud technologies is that Cloud technologies provide a huge benefit to a start-up companies to further expanse and scale their business with greater flexibility and relatively minimum cost and commitment. Start-up companies now have an option to rent the necessary hardware infastructure instead of investing huge capital on it, since many of the established Tech companies such as Google and Amazon provide access to their extensive range of hardware to be rent. These infrastructures could then be run via virtualisation methods in which provide a great flexibility of scaling.

Cloud technologies have been utilised in many parts of the modern society, one of them being for rendering high-resolution images. The implementation of the cloud technologies in rendering task, would eliminate the hardware constraint for the client or the user. An example of this implementation could be found in rendering model in “Image-Based Network Rendering of Large Meshes for Cloud Computing” (Okamoto et al. 2010). This report would evaluate the performance of GPU nodes in a scalable cloud architecture that are used to produce a Terapixel 3D City visualization of Newcastle Upon Tyne.

To gain the information adauate to create an evaluation on the performance of the GPU nodes capabilities to render the terapixel visualizations, a data mining process following a CRISP-DM best data mining practice would be conducted. The tools that would be used to carry out the data mining process are Rstudio combined with the Project Template package to streamlined the data mining process.

CRISP-DM Methodology

Business Objective

The scalable cloud architecture that was created to support the terapixel images rendering in order for it to be accessible for users without putting a constraint on the users hardware. Therefore it might be crucial to be able to extract performance information out of the hardware, in this case, the GPU, to further re-evaluate the amount of GPU nodes and its type needed to run the current rendering process or to create a plan for future usage. The goal of this data mining project is to be able to extract the performance evaluation information out of the current cloud-architecture that is currently use to support the terapixel image rendering. The information extracted would then be able to be used to improve the cost efficiency of running the current cloud-architecture or create future planning to prepare for an expansion. To aid in

evaluating the performance of the cloud-architecture rendering process, these questions are provided as a guidance for the analyses. The questions are

- Which event dominate the task run time
- Are there any implication that GPU temperature affect the performance of rendering process
- Are there any correlation between the increased power draw and render time
 - Quantify the variation in computation requirements for particular tile
- Are there any particular GPU cards (based on their serial number) that is distinct from the other in terms of performance
- Are there any inefficiencies detected for the task scheduling process

To help achieving the goals mentioned above, this data mining project would be done using the best practice CRISP-DM methodology, and would be done entirely on R environment supported by the Project Template package and git version control to further streamline the data mining process.

Data understanding

There are three data frames available in which the data mining process could be implemented. These data frames are extracted from the application checkpoint and system metric output from the production of terapixel image. Each data frames are in the comma-separated value format in which it could be loaded directly into R by placing the data frames into the data folder of the Project Template, and loading the project. The three data frames mentioned consist of Application Checkpoint data, GPU data, and Task X Y data. Before any process could be made, an initial data exploration would be done on these data frames.

Application Checkpoints Data

Data Preview

```
## # A tibble: 6 x 6
##   timestamp      hostname      eventName  eventType jobId      taskId
##   <chr>          <chr>          <chr>      <chr>    <chr>      <chr>
## 1 2018-11-08T0~ 0d56a730076643~ Tiling      STOP    1024-lvl12-7e~ b47f0263-ba~
## 2 2018-11-08T0~ 0d56a730076643~ Saving Co~ START    1024-lvl12-7e~ 20fb9fcf-a9~
## 3 2018-11-08T0~ 0d56a730076643~ Saving Co~ STOP    1024-lvl12-7e~ 20fb9fcf-a9~
## 4 2018-11-08T0~ 0d56a730076643~ Render      START    1024-lvl12-7e~ 20fb9fcf-a9~
## 5 2018-11-08T0~ 0d56a730076643~ TotalRend~ STOP    1024-lvl12-7e~ 20fb9fcf-a9~
## 6 2018-11-08T0~ 0d56a730076643~ Render      STOP    1024-lvl12-7e~ 3dd4840c-47~
```

Data Summary

```
##   timestamp      hostname      eventName      eventType
##   Length:660400   Length:660400   Length:660400   Length:660400
##   Class :character Class :character Class :character Class :character
##   Mode  :character Mode  :character   Mode  :character Mode  :character
##   jobId          taskId
##   Length:660400   Length:660400
##   Class :character Class :character
##   Mode  :character Mode  :character
```

The Application Checkpoints Data consist of 660400 rows and 6 columns as could be seen from the first 6 rows of the data and the summary table above. For all the columns that are present on the data frame, all of them are character class, and there are no missing values detected. However, after checking for duplicates, there are 130821 duplicated individual within the data set. Therefore, moving onwards with the analyses, the data would be made sure to carry no duplicates.

Assessing each columns of the data, a following interpretations could be extracted.

- **Timestamp** = Contains the date and the current time of the event down to the milliseconds. The class of this column is that of a character, therefore for further analysis some processing might be needed to convert the timestamp into a numerical format.
- **Hostname** = Contains the Hostname of the virtual machine. The hostname is auto-assigned by the azure batch system. There are 1024 hostname recorded on the application checkpoints data.
- **Event Name** = Contains the name of the events that are happening on the task. There are 5 distinct events recorded on the application checkpoints data. These events are
 - Total Render = The entire task itself
 - Saving Config = Saving the configuration
 - Render = The rendering process of the image tile
 - Tiling = Post-processing of the rendered tile
 - Uploading = Uploading the output of the post-processed image tile to the Azure Blob Storage
- **Event Type** = Contains two distinct event Type which are STOP and START. This event type are meant to describe the checkpoint of the event name. Meaning, a data point with event name “Tiling” and event type “START” refer to the start of “Tiling” process of a certain task.
- **Job ID** = Contains the Azure batch job. There are three distinct value of the Job ID column, and it refers to the level of zoom (4,8,12) of the rendering proces.
- **Task ID** = Contains the ID of the Azure batch task. There are 65793 distinct task ID. Each task id would have all of the five event names respectively.

GPU Data

Data Preview

```
## # A tibble: 6 x 8
##   timestamp hostname gpuSerial gpuUUID powerDrawWatt gpuTempC gpuUtilPerc
##   <chr>      <chr>      <dbl> <chr>      <dbl>      <int>      <int>
## 1 2018-11-- 8b6a0ee~    3.23e11 GPU-1d~    132.        48        92
## 2 2018-11-- d824187~    3.24e11 GPU-04~    117.        40        92
## 3 2018-11-- db871cd~    3.23e11 GPU-f4~    122.        45        91
## 4 2018-11-- b9a1fa7~    3.25e11 GPU-ad~    50.2        38        90
## 5 2018-11-- db871cd~    3.23e11 GPU-2d~    142.        41        90
## 6 2018-11-- 265232c~    3.24e11 GPU-71~    120.        43        88
## # ... with 1 more variable: gpuMemUtilPerc <int>
```

Data Summary

```
##   timestamp      hostname      gpuSerial      gpuUUID
## Length:1048575 Length:1048575 Min. :3.201e+11 Length:1048575
## Class :character Class :character 1st Qu.:3.236e+11 Class :character
## Mode :character Mode :character Median :3.236e+11 Mode :character
```

```

##                               Mean    :3.240e+11
##                               3rd Qu.:3.250e+11
##                               Max.    :3.252e+11
## powerDrawWatt      gpuTempC      gpuUtilPerc      gpuMemUtilPerc
## Min.      : 22.55    Min.      :26.00    Min.      : 0.00    Min.      : 0.00
## 1st Qu.: 45.15    1st Qu.:38.00    1st Qu.: 0.00    1st Qu.: 0.00
## Median : 96.68    Median :40.00    Median : 89.00    Median :43.00
## Mean      : 89.32    Mean      :40.07    Mean      : 63.23    Mean      :33.47
## 3rd Qu.:121.34    3rd Qu.:42.00    3rd Qu.: 92.00    3rd Qu.:51.00
## Max.      :197.01    Max.      :55.00    Max.      :100.00    Max.      :83.00

```

Extracting the information from the data preview and the data summary above, the GPU data consist of 1048575 rows and 8 columns. The columns are consisted of Timestamp, Hostname, GPU serial number, GPU UUID, Power Draw on GPU in Watt, GPU Temperature in Celsius, Percent Utilisation of GPU Cores, and Percent Utilisation of the GPU Memory. For the timestamp, hostname, and GPU UUID, the class of column is listed as characters and for the rest of the columns, they are listed as continuous data. However, upon closer inspection it could be deducted that GPU Serial should not be regarded as a continuous data since it refers to serial number of the physical GPU card. It seems that this data set also suffers from duplicated values, and after some thorough look at the data set there are 662118 duplicated values found on the data. Similar to the application checkpoints data, the information regarding the duplicated values within the data should be taken into consideration for further analyses to make sure that these duplicated values are not carried over.

Assessing each columns of the GPU data, a following interpretations could be extracted.

- **Timestamp** = Similar to the Application Checkpoints Data, the timestamp column for the GPU data contains the current time in which the condition is being recorded. It shows the date as well as the time down to the milliseconds. There is an interesting feature that is found on the timestamp for the GPU data. The timestamp shows that for each hostname, the GPU condition of each hostname is recorded every **2 Seconds**.
- **Hostname** = Similar to the Application Checkpoints data, contains the hostname of the virtual machine that is auto-assigned by the Azure Batch system. The number of unique hostname recorded on this data is equal to the number of hostname that is recorded for the Application Checkpoints data, which is 1024 unique hostnames. Therefore, this information could prove to be useful for the analyses on the later part of this report.
- **GPU Serial** = Contains the serial number of the physical GPU card. Even though on the data summary the GPU serial column is treated as a continuous variable, on reality, this column should be treated as a categorical variable.
- **GPU UUID** = Contains the unique system id which is assigned by the Azure system to the GPU Unit. Note that GPU UUID is unique for each hostname since the number of unique GPU UUID is equal to that of the number of hostname at 1024 unique data points.
- **GPU Power Draw in Watt** = Contains the current recorded value of the Power Draw of the GPU in Watt. This variable is a continuous variable. It has a minimum recorded value of 22.55 Watt and a maximum recorded value of 197.01 Watt.
- **GPU Temperature in Celcius** = Contains the current recorded Temperature of the GPU in Celcius. This variable is a continuous variable in which the minimum recorded value for this variable is 26 degree Celsius and the maximum recorded value of 55 degree Celsius.
- **GPU Util Perc** = Contains the current recorded value of the Percent Utilisation of the GPU cores. This variable is a continuous variable ranging from 0% to 100%. It has a minimum recorded value of 0% and a maximum recorded value of 100%.
- **GPU Memory Util Perc** = Contains the current recorded value of the Percent Utilisation of the GPU memory. This variable is continuous variable ranging from 0% to 100%. It has a minimum recorded value of 0% and a maximum recorded value of 83%

Task X Y Data

Data Preview

```
## # A tibble: 6 x 5
##   taskId                      jobId                      x      y level
##   <chr>                      <chr>                      <int> <int> <int>
## 1 00004e77-304c-4fbd-88a1-1~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 116   178    12
## 2 0002afb5-d05e-4da9-bd53-7~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 142   190    12
## 3 0003c380-4db9-49fb-8e1c-6~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 142    86    12
## 4 000993b6-fc88-489d-a4ca-0~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 235    11    12
## 5 000b158b-0ba3-4dca-bf5b-1~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 171    53    12
## 6 000d1def-1478-40d3-a5e3-4~ 1024-lvl12-7e026be3-5fd0-48ee-b7~ 179   226    12
```

Data Summary

```
##   taskId                      jobId                      x      y
## Length:65793      Length:65793      Min.   : 0      Min.   : 0
## Class :character   Class :character   1st Qu.: 63     1st Qu.: 63
## Mode  :character   Mode  :character   Median :127     Median :127
##                                     Mean  :127     Mean  :127
##                                     3rd Qu.:191    3rd Qu.:191
##                                     Max.   :255     Max.   :255
##   level
## Min.   : 4.00
## 1st Qu.:12.00
## Median :12.00
## Mean   :11.98
## 3rd Qu.:12.00
## Max.   :12.00
```

Extracting information out of the Data Preview and the Data Summary above, the Task-X-Y Data consist of 6579 rows and 5 columns. The number of rows are equal to the number of task as shown by the taskId column. It has recurring columns from the Application Checkpoints data in taskId and jobId. This information should be taken into consideration when finding the correlation between the Application Checkpoints data and the Task-X-Y data. The Task-X-Y data columns that shows coordinate “X” and “Y” is deemed by the R system as a column of continuous variable, since the values inside the column are consisted of numerical value. However, note that since it only shows the coordinate of an object, the stats value extracted from these two columns might not be that beneficial for the analyses, although it still could be useful for determining whether a certain coordinates affect the performance of the rendering task. The “level” column is shown as a continuous variable on the data summary even though the “level” column represent zoom level of the rendered visualisation. There are no duplicates detected within this data frame.

Assessing each column of the Task-X-Y data, a following interpretations could be extracted.

- **Task ID** = Similar to the Application Checkpoints data, this column contains the ID of the Azure Batch task. Since it shows the same number of unique taskId as that of the Application Checkpoints data with 65793 unique task ID, it should be taken into consideration to link the task id from the Task-X-Y data with the task id in the Application Checkpoints data.
- **Job ID** = Similar to the Application Checkpoints data, this column contains Azure batch job Id
- **X** = Contains the X coordinate of the rendered image tile.
- **Y** = Contains the Y coordinate of the rendered image tile.
- **Level** = Contains the zoom levels of the render. The visualisation is created with 12 zoom levels starting with 1 until 12, however, for the data set in this project, only zoom level 4,8, and 12 that are

present since the intermediate level are derived in the tiling process. Note that the zoom level is also reflected in the JobId column, as jobId also mentions the level of zoom in its character.

Data Preparation

Before analyses would be made, the data frames mentioned and described on the Data Understanding section, would need to be processed first. The process includes but not limited to data wrangling, removing duplicates, merging data frames, and extracting new features from the data. In this part of the report, the processes that are applied to the data frames would be discussed and explanations would be given.

Before any data merging could be done, each individual data frame would be processed first. The data processing is done based on the information gained from the assessment of each data frame that is highlighted in the Data Understanding section of this report.

Application Checkpoints Data

Referring to the initial data assessment of the Application Checkpoints data frame, data points duplicates would be removed. Removing the duplicates on the Application Checkpoints would leave us with a data frame which has 657930 data points and removing 130821 duplicates. The reason to remove any duplicates is to avoid having any error during the analyses or producing an analysis that is skewed.

The “**Timestamp**” column of the Application Checkpoints could not be directly used as a method of analyses since the type of values that are present in the column are characters. Therefore we would process the “**Timestamp**” column and extract only the numerical time value out of the “**Timestamp**”. The dates that are recorded on the “**Timestamp**” are all showing the same date which is **2018-11-08** with varying times. Therefore, for the analyses, only the time would be extracted and would be converted into the Hour:Min:Sec format.

For the **Event Name** column, the number of data points for each event names seems to be the same, indicating that for all the task that is present in the data, each task includes all possible event names. The Distribution table of the **Event Name** column is shown as below

| | | | | | |
|----|--------|---------------|--------|-------------|-----------|
| ## | | | | | |
| ## | Render | Saving Config | Tiling | TotalRender | Uploading |
| ## | 131586 | 131586 | 131586 | 131586 | 131586 |

Each event name recorded on the data 131586 times, which if divided by the number of **Event Type** of each Event Name which is 2, would equal to the amount of distinct **Task ID**. With these information at hand, we would proceed with wrangling the data, to prepare it for analyses. The wrangling process are done in these manners

1. Ensure that the Application Checkpoints data does not contains any duplicated data points. If there are any, remove the duplicated data points.
2. Extract only the time characters out of the **Timestamp** column and change the format to that of an Hour:Minute:Second format.
3. Transform the **Event Type** row into two separate columns from the two distinct value of **Event Type** column which are **START** and **STOP**.
4. Fill these new columns **START** and **STOP** with the **Timestamp** for each **Event Name** which has **Event Type** that matched each of these two new columns.
5. Create a new feature called **duration** which would store the duration of each event, which is calculated by subtracting the **STOP** and **START** column. This newly made feature **duration** is what would be used in analyses as a metric that describe the length of an event.

The resulting data frame would be

```
## # A tibble: 6 x 7
##   hostname eventName jobId taskId STOP          START
##   <chr>      <chr>    <chr> <chr> <Period>    <Period>
## 1 0d56a73~ Tiling      1024~ b47f0~ 7H 41M 55.921S 7H 41M 55.2S
## 2 0d56a73~ Saving C~ 1024~ 20fb9~ 7H 42M 29.845S 7H 42M 29.842S
## 3 0d56a73~ Render       1024~ 20fb9~ 7H 43M 10.965S 7H 42M 29.845S
## 4 0d56a73~ TotalRen~ 1024~ 20fb9~ 7H 43M 13.957S 7H 42M 29.842S
## 5 0d56a73~ Render       1024~ 3dd48~ 7H 43M 56.239S 7H 43M 16.506S
## 6 0d56a73~ Uploading 1024~ 3dd48~ 7H 43M 57.245S 7H 43M 56.239S
## # ... with 1 more variable: duration <Duration>
```

To aid with the analyses that concerns with the usage of duration column, a horizontal data frame is created in which there are 5 new columns created by using the **Event Name** values as column name and the **duration** of the corresponding **Event Name** would be the values resided within these newly made columns. The resulting data frame would be

```
## # A tibble: 6 x 8
##   hostname jobId taskId Tiling          'Saving Config'      Render
##   <chr>    <chr> <chr> <Duration>    <Duration>    <Durat>
## 1 0d56a73~ 1024~ b47f0~ 0.7209999999999997s 0.00199999999999889s 23.501s
## 2 0d56a73~ 1024~ 20fb9~ 0.99s          0.00300000000000011s 41.12s
## 3 0d56a73~ 1024~ 3dd48~ 0.9769999999999997s 0.00199999999999889s 39.733s
## 4 0d56a73~ 1024~ c9e24~ 0.977s          0.00200000000000244s 33.156s
## 5 0d56a73~ 1024~ c5a7a~ 0.911s          0.00199999999999889s 40.571s
## 6 0d56a73~ 1024~ f600a~ 0.9279999999999997s 0.0019999999999978s 32.781s
## # ... with 2 more variables: TotalRender <Duration>, Uploading <Duration>
```

Do note that the newly made columns that are based on the **Event Name** column values are not set up in a chronological order in which for each task, the order of event should be, **Saving Config**, **Render**, **Tiling**, and **Uploading**. We would only use this newly made horizontal data to aid in extracting the average duration of each event name to be used for analyses.

GPU Data

The GPU Data as shown on the Data Understanding part of this report, has some duplicated data points, therefore before any analyses or wrangling is done on this data set, the GPU data set needs to be cleared of this duplicated data points. After removing the data points, the number of rows on the GPU Data would be 1048569 down from the original number of rows for the GPU data of 1048575 data points. Since there are not much information that could be extracted from the GPU data alone, therefore the wrangling that is done for this data set is merely to prepare it for merging this data frame with other data frame on this project. Other than removing any duplicates from the data, the **Timestamp** column of the GPU data would also be transformed into the Hour:Min:Sec format, just as has been done to the **Timestamp** column of the application checkpoints.

Task-X-Y Data

Similar to the GPU data frame, the Task-X-Y data by itself, does not have any noticeable information that could be extracted. And since the Task-X-Y data does not contain any duplicated data points, therefore there is no individual wrangling done to this data.

Merging Application Checkpoints Data with GPU Data

To extract necessary information out of the GPU data, the GPU data needs to be merged with the Application Checkpoints data to gather detailed information regarding the tasks or event tasks that are being run on the GPU nodes at a current point of time. The merged should be by **Hostname** and **Timestamp** columns, however the merge for the **Timestamp** column should not be done so directly, since the **Timestamp** of the GPU data only record the GPU condition every 2 seconds, and if we merged it directly, it would only bear 232 matching results. To tackle this problem, the GPU condition for a particular task is calculated by the average GPU condition during the length of the task itself. An example for this calculation if **Total Render** event started at **8H 1M 29.305S** and ended at **8H 0M 44.142S** we would apply a conditional that would extract all the GPU conditions recorded within this time frame, and for this particular event, we would use the average GPU condition between the two timestamps above to be recorded as the GPU condition of the particular event. With this conditional we are successful on creating the desired data. The data frame that are created are based on the horizontal data, and would bear 40 columns. Since not all the columns would be visible if shown here, below is only a preview of the first view columns of the data that are part of the **Total Render** event name. #### Data Preview

```
##                               hostname
## 1 0d56a730076643d585f77e00d2d8521a00000N
## 2 0d56a730076643d585f77e00d2d8521a00000N
## 3 0d56a730076643d585f77e00d2d8521a00000N
## 4 0d56a730076643d585f77e00d2d8521a00000N
## 5 0d56a730076643d585f77e00d2d8521a00000N
## 6 0d56a730076643d585f77e00d2d8521a00000N
##                               jobId
## 1 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 2 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 3 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 4 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 5 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 6 1024-1vl12-7e026be3-5fd0-48ee-b7d1-abd61f747705
##                               taskId      taskStart      taskStop
## 1 02d712d0-cd88-4d7b-83f9-ce0db22d8ede  7H 56M 50.424S  7H 57M 25.383S
## 2 0531106c-2412-4858-8806-f105759024fe  8H 11M 8.898S  8H 11M 52.639S
## 3 0cbcfac1-1913-44d3-8fb6-483da86d2bff  7H 57M 27.517S  7H 58M 5.097S
## 4 1bf9597f-77bd-4b37-aadd-a5af9baa2308  7H 53M 10.807S  7H 53M 51.943S
## 5 1f4f7e10-835b-435d-af6a-5d13370eb955  8H 25M 42.215S  8H 26M 27.751S
## 6 20fb9fcf-a927-4a4b-a64c-70258b66b42d  7H 42M 29.842S  7H 43M 13.957S
## duration_Total_Render                               gpuUUID      gpuSerial
## 1                               34.959s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## 2                               43.741s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## 3                               37.58s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## 4                               41.136s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## 5                               45.536s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## 6                               44.115s GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11
## totalAvgPowerDrawWatt totalAvgGPUTempC totalAvgGPUUtilPerc
## 1                               86.23100                               37.50000                               71.00000
## 2                               122.89857                               41.07143                               78.35714
## 3                               93.96556                               37.77778                               68.22222
## 4                               91.44923                               38.38462                               56.84615
## 5                               97.88267                               39.66667                               79.26667
## 6                               92.97000                               36.76923                               57.46154
## totalAvgGPUMemUtilPerc
## 1                               32.30000
```



```
## 2          43.07143
## 3          31.88889
## 4          28.46154
## 5          38.86667
## 6          31.69231
```

However, after carefully assessing the data, it is found that there is some problem detected on the wrangled data. Since the GPU data **Timestamp** only record the timestamp every 2 seconds, an event name that has duration less than 2 seconds, would return a **NaN** or not a number value. To mitigate this problem, for the GPU conditions that returns **NaN** value would be subjected to another conditional in which, the GPU conditions for this particular event name is average of GPU conditions of the two closest timestamp on the GPU Data to the START and STOP timestamp of this particular event. For example, a **Saving Config** event happen between **7H 56M 50.424S** and **7H 56M 50.426S** and there are no GPU condition that are recorded between that time for the particular hostname. Therefore, we would take the GPU condition of the particular hostname with the timestamp that is earlier than but is the closest to **7H 56M 50.424S** and the one which is latter but closest to **7H 56M 50.426S** and we would take the average between these two closest timestamps and record it as the GPU condition of the corresponding **Saving Config** event. There are both data that has been wrangled horizontally with each event name with its each corresponding features as columns and data that is wrangled vertically meaning that the event names are still contained within the **Event Name** column. But for the preview below, we choose to preview the data that is wrangled vertically so that it is more compact in width.

```
##                               hostname  eventName
## 1 0d56a730076643d585f77e00d2d8521a00000N TotalRender
## 2 0d56a730076643d585f77e00d2d8521a00000N TotalRender
## 3 0d56a730076643d585f77e00d2d8521a00000N TotalRender
## 4 0d56a730076643d585f77e00d2d8521a00000N TotalRender
## 5 0d56a730076643d585f77e00d2d8521a00000N TotalRender
## 6 0d56a730076643d585f77e00d2d8521a00000N TotalRender
##                               jobId
## 1 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 2 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 3 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 4 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 5 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
## 6 1024-1vl112-7e026be3-5fd0-48ee-b7d1-abd61f747705
##                               taskId  eventStart  eventStop duration
## 1 b47f0263-ba1c-48a7-8d29-4bf021b72043 7H 41M 31.687S 7H 41M 56.111S 24.424s
## 2 20fb9fcf-a927-4a4b-a64c-70258b66b42d 7H 42M 29.842S 7H 43M 13.957S 44.115s
## 3 3dd4840c-47f2-4dcc-a775-df2ef6498d71 7H 43M 16.504S 7H 43M 57.245S 40.741s
## 4 c9e249d8-52ed-40c6-8713-b5cbf02ea87e 7H 44M 47.555S 7H 45M 21.898S 34.343s
## 5 c5a7a2df-ddeb-4f54-9cc5-446e3a9ba1ba 7H 45M 24.026S 7H 46M 5.524S 41.498s
## 6 f600a589-3332-4408-a2bd-2b6833981407 7H 46M 7.958S 7H 46M 41.744S 33.786s
##                               gpuUUID  gpuSerial AvgPowerDrawWatt
## 1 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 72.38000
## 2 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 92.97000
## 3 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 124.86250
## 4 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 77.15364
## 5 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 89.50286
## 6 GPU-1265fef9-aea4-4a5e-8a63-cc5af7b19f4f 3.25217e+11 101.25600
## AvgGPUTempC AvgGPUUtilPerc AvgGPUMemUtilPerc
## 1 32.33333 69.16667 26.33333
## 2 36.76923 57.46154 31.69231
## 3 39.08333 81.75000 41.83333
```

| | | | |
|------|----------|----------|----------|
| ## 4 | 37.90909 | 71.72727 | 32.81818 |
| ## 5 | 38.07143 | 71.00000 | 33.00000 |
| ## 6 | 38.60000 | 86.20000 | 39.60000 |

Merging the Merged Data Frame with Task-X-Y Data

Merging the merged GPU and Application Checkpoints data frame with the Task-X-Y data frame is done by specifying the merge by **Task Id** and **Job Id** columns. For the Horizontal data, the merge would bear 65793 rows while for the vertical data the merge would bear 328965 rows.

Subsetting the data based on Event Name

To streamline the analyses, the main data would also be subset in terms of its event name and **GPU Serial** column values are extracted as a feature to be recorded into a new column **GPU Serial Feature**. The subset would produce these data frames

- totalRenderData = A data frame that contains only event name ***Total Render***.
- eventNamesData = A data frame that contains all event names apart from ***Total Render***. In this data frame, the **Event Name** column is also extracted as feature, creating a new feature **Event Name Feature** in which, the event name is turned into a factor, and turned into numerical class for analyses.
- savingConfigData = A data frame that contains only event name ***Saving Config***
- renderData = A data frame that contains only event name ***Render***
- tilingData = A data frame that contains only event name ***Tiling***
- uploadingData = A data frame that contains only event name ***Uploading***

 These data subsets are created in order for separate analyses for different event name could be executed.

Data Analyses and Modelling

To reach the goal and answer the questions mentioned in the earlier part of the report, analyses would be made on the data that we have gathered and processed. The analyses would be done in a structure based on the order of the questions mentioned on the Business Objective part of the report. However before any analyses could be made, and initial analysis would be done on the entire data to provide a broad overview of the data and extract some interesting features.

Numerical Summaries of the Data

 Extracting the numerical summaries out of the numerical value of the data would bear us

| | | | |
|----|----------------------------------|------------------|---------------|
| ## | duration | AvgPowerDrawWatt | AvgGPUTempC |
| ## | Min. :0.001999999999999534s | Min. : 23.45 | Min. :26.00 |
| ## | 1st Qu.:0.9020000000000001s | 1st Qu.: 42.50 | 1st Qu.:37.55 |
| ## | Median :1.066s | Median : 58.40 | Median :39.50 |
| ## | Mean :17.2364643685498s | Mean : 66.28 | Mean :39.70 |
| ## | 3rd Qu.:40.736s | 3rd Qu.: 91.80 | 3rd Qu.:41.45 |
| ## | Max. :93.697s (~1.56 minutes) | Max. :158.60 | Max. :52.43 |
| ## | AvgGPUUtilPerc AvgGPUMemUtilPerc | | |
| ## | Min. : 0.00 Min. : 0.00 | | |

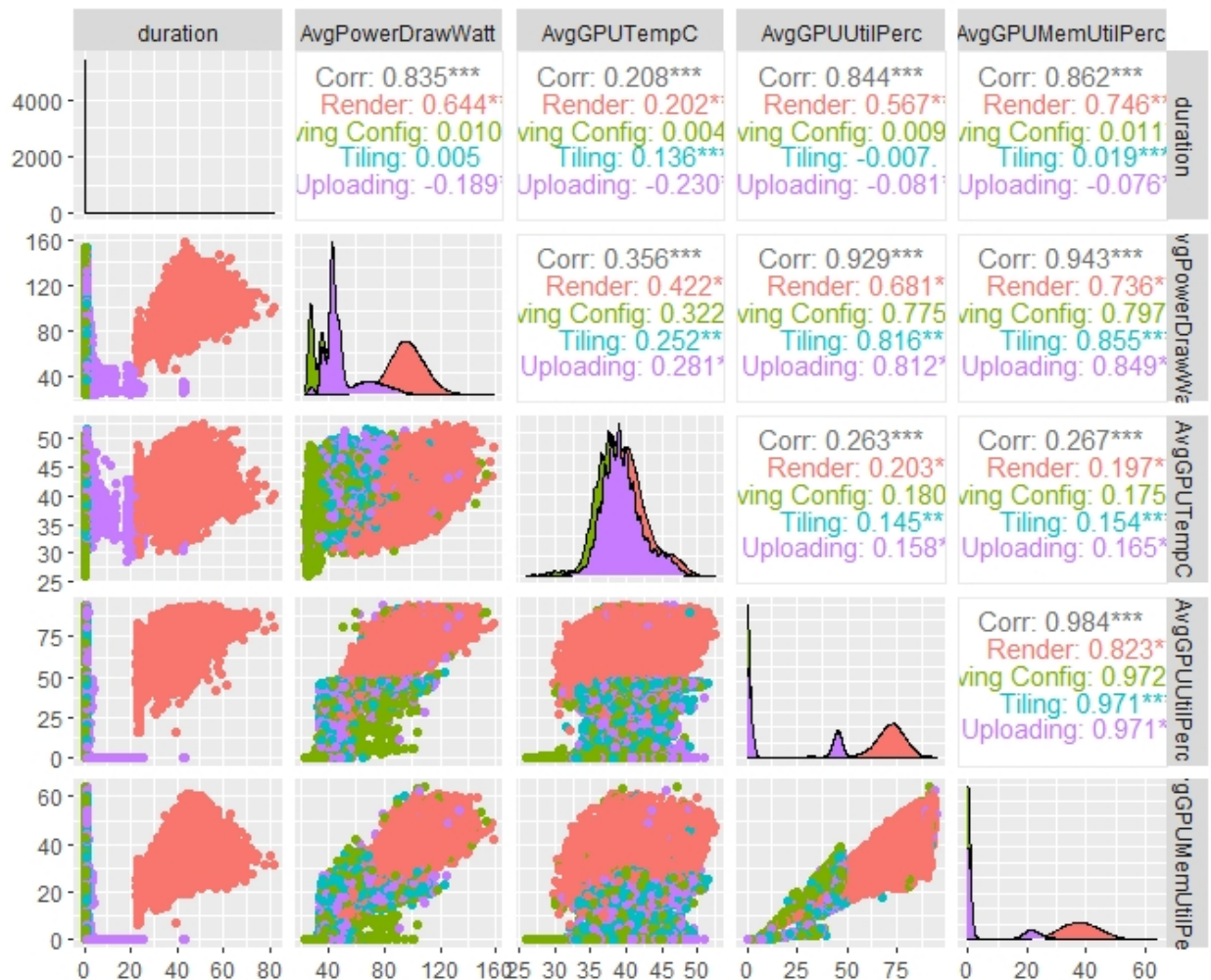
```
## 1st Qu.: 0.00 1st Qu.: 0.00
## Median :44.00 Median :18.90
## Mean :34.18 Mean :17.58
## 3rd Qu.:68.78 3rd Qu.:34.92
## Max. :94.50 Max. :64.00
```

With variation for each column

```
##          duration AvgPowerDrawWatt      AvgGPUTempC      AvgGPUUtilPerc
##      423.216000      757.905404          9.393483      1113.057298
## AvgGPUMemUtilPerc
##      320.918051
```

Based on the numerical summaries extracted from the data as shown above, it seems that out of the numerical columns on the data, the column or feature that has the most variation belong to the **Percent Utilisation of GPU Cores**. The **GPU Temperature** seems to have a really low variation. This information might be used to answer the questions stated on the Business Objective of the report in regards to the GPU Temperature.

Graphical Summaries of the Data



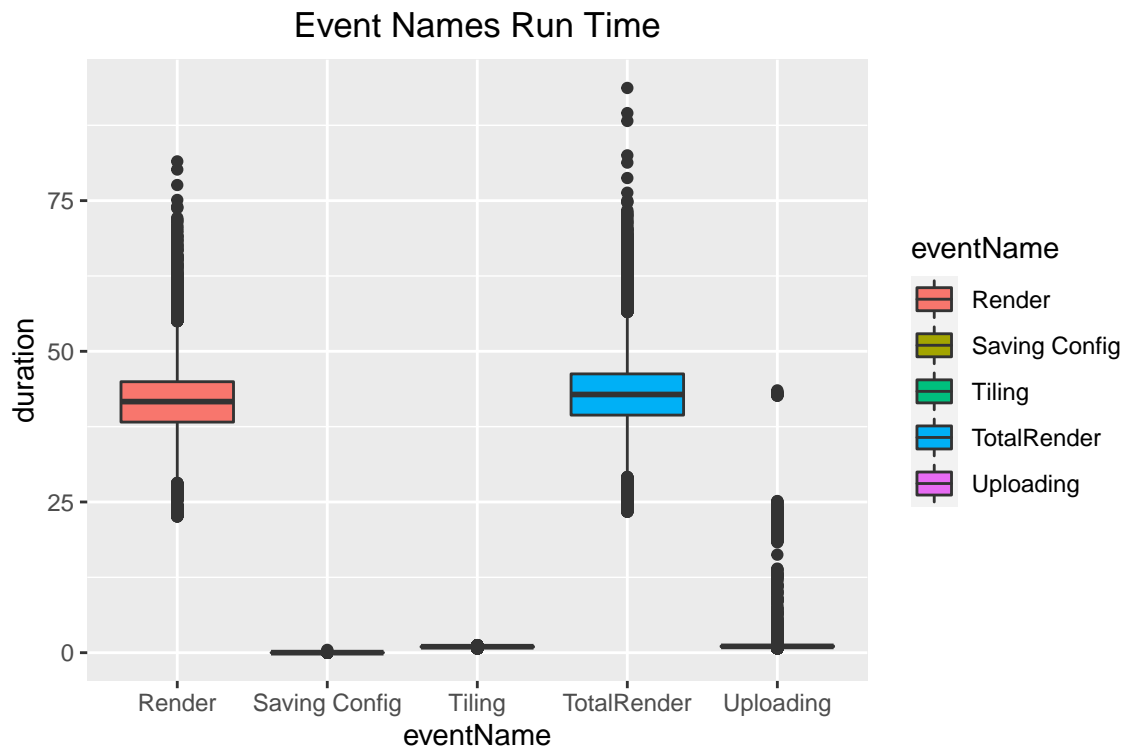
From the graph above, most of the feature that relates to GPU performance such as **duration**, **GPU Power Draw**, **Percent Utilisation of GPU Cores** and **Percent Utilisation of GPU Memory** show that data points that belong to **Render** event name tends to reside on the upper side of these graphs. Therefore it might be indicated that **Render** demand more performance on the GPU node than other event name. It seems there are interesting traits based on the correlation value on some of these combination of features plots. These interesting correlation traits are found between these variables

- **Duration** and **GPU Power Draw** with a correlation score between the two showing 0.835
- **Duration** and **Percent Utilisation of GPU Cores** with a correlation score between the two showing 0.844
- **Duration** and **Percent Utilisation of GPU Memory** with a correlation score between the two showing 0.862
- **Percent Utilisation of GPU cores** and **GPU Power Draw** with a correlation score between the two showing 0.929
- **Percent Utilisation of GPU memory** and **GPU Power Draw** with a correlation score between the two showing 0.943
- **Percent Utilisation of GPU cores** and **Percent Utilisation of GPU memory** with a correlation score between the two showing 0.984

Apart from the traits shown above, **GPU Temperature** feature doesn't show any strong correlation with any other features shown on the graph above since all the correlation between the **GPU Temperature** and other feature show correlation score below 0.4. These information gathered from plotting the graphs above, would be assessed further to answer the questions stated on the Business Objective part of this report.

Determining Event which Dominate the Task Run Time

To answer the question regarding which Event dominates the task run time, a box plot would be plotted to show the spread and mean of each event name duration.



Based on the graph above, a conclusion could be drawn that the **Render** event name dominates the task run time since the box plot of **Render** event is nearly identical with the **Total Render** event name that represent the whole length of the task run time. To quantify the result of the box plot graph above, a numerical summaries are extracted from the data that contains the duration of each event name.

Data Preview

```
## # A tibble: 6 x 8
##   hostname jobId taskId Tiling          'Saving Config'      Render
##   <chr>      <chr> <chr> <Duration>          <Duration>          <Durat>
## 1 0d56a73~ 1024~ b47f0~ 0.7209999999999997s 0.001999999999999889s 23.501s
## 2 0d56a73~ 1024~ 20fb9~ 0.99s                0.003000000000000011s 41.12s
## 3 0d56a73~ 1024~ 3dd48~ 0.9769999999999997s 0.001999999999999889s 39.733s
## 4 0d56a73~ 1024~ c9e24~ 0.977s                0.002000000000000244s 33.156s
## 5 0d56a73~ 1024~ c5a7a~ 0.911s                0.001999999999999889s 40.571s
## 6 0d56a73~ 1024~ f600a~ 0.9279999999999997s 0.00199999999999978s 32.781s
## # ... with 2 more variables: TotalRender <Duration>, Uploading <Duration>
```

Extracting the mean out of the data would give us a result of

```
##           Tiling Saving Config           Render   TotalRender   Uploading
##    0.973207210    0.002476266  41.208219552  42.604778289   1.393640524
```

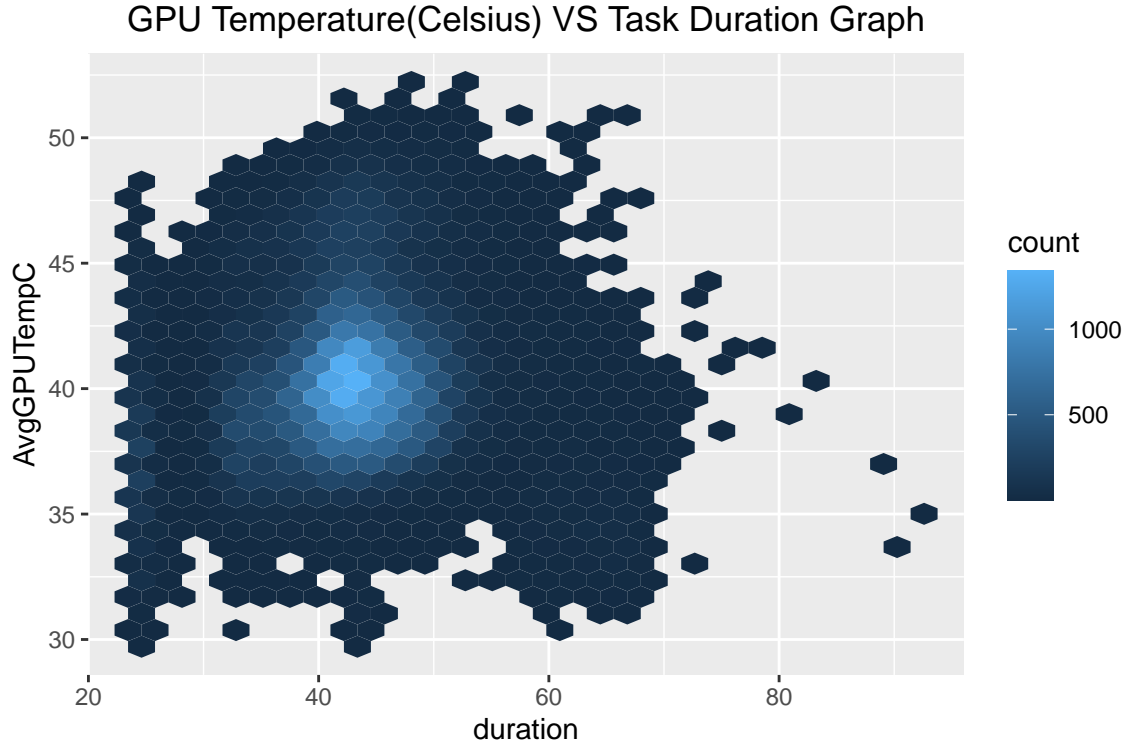
With a variance of

```
##           Tiling Saving Config           Render   TotalRender   Uploading
## 6.951450e-03  7.194054e-06  3.652248e+01  4.222425e+01  6.625830e+00
```

These result further cemented the narrative in which **Render** event dominate the task run time by more than 90%.

Implications on Temperature of GPU Affecting the Rendering Performance

GPU capabilities are measured by its performance to succesfully render an object. Since there are no information given on the complexity of the object being rendered, it is assumed that each tile of the rendered object are equal in memory size, therefore the metric in which the GPU performance is measured is in terms of the length of the render process. Therefore, to check on the implications on Temperature of GPU affecting the GPU performance, we would like to revisit the plot that shows the correlation between the Duration of rendering task and the temperature of the GPU during this process. To asses this, we would use only the **Total Render** event name that represent the whole length of the task itself.



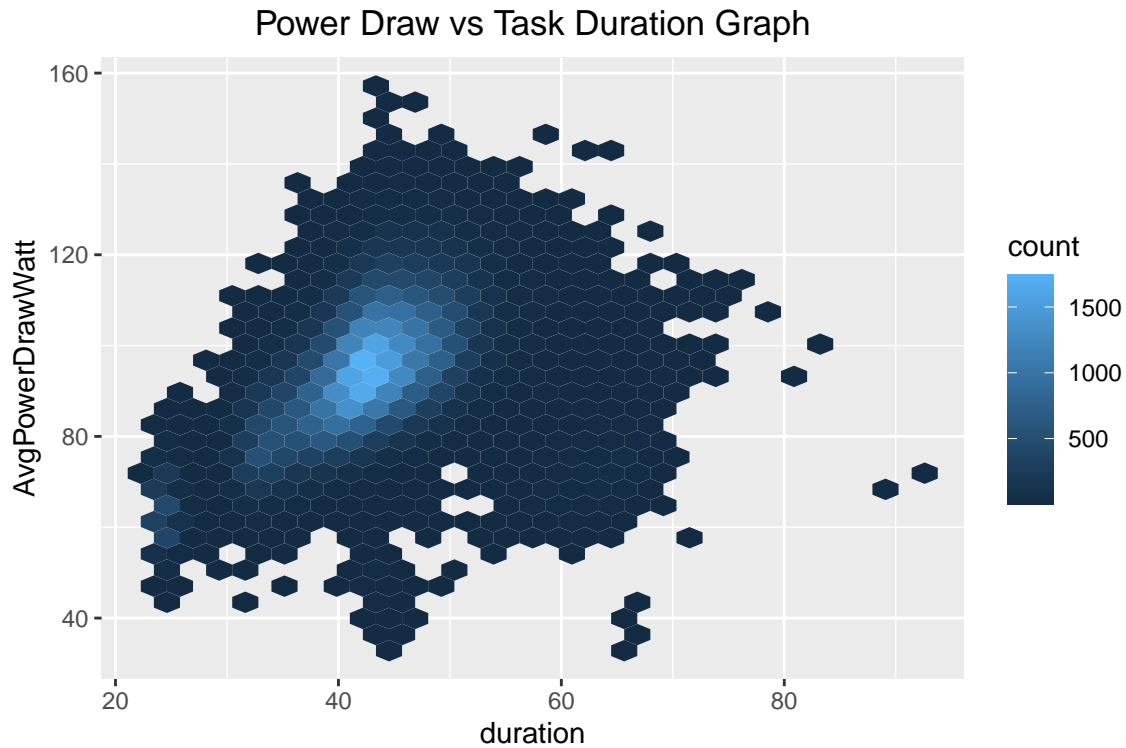
Based on the GPU Temperature vs Task Duration graph, it seems that there are no strong correlation between the two variables. The graph didn't show any noticeable trend and the spread of the data appeared randomly scattered and there seems to be a high volume of data points around the mean of both variable. To further quantify this deduction, the correlation matrix between the **GPU Temperature** and the **Duration** of the task show a score of

| | duration | AvgGPUPTempC |
|--------------|-----------|--------------|
| duration | 1.0000000 | 0.1037189 |
| AvgGPUPTempC | 0.1037189 | 1.0000000 |

The deduction from the GPU Temperature vs Task Duration Graph, and further supported by the correlation matrix between the two variables, shows that there are no implications that Temperature of GPU affect the Performance of the GPU rendering capabilities that are measure by the length of the render as performance metric.

Connection Between the Increased Power Draw and Render Time and Computational Variation for Particular Tile

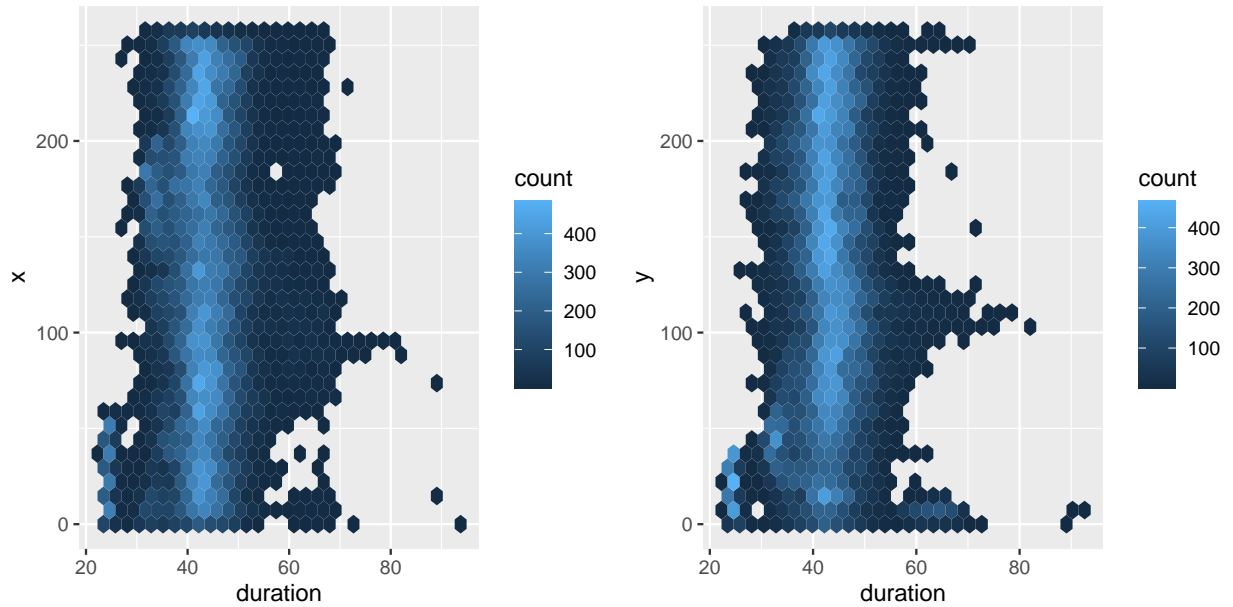
To answer the question regarding the correlation between an increased power draw and the duration of render, a graph between the two feature variable **GPU Power Draw** and **duration** would be plotted and produce result of



Looking at the graph above, there is a slight positive correlation trend is detected between the two variables. However, based on this graph alone, there is no clear indication that increased power draw resulted in longer duration or the other way around. Therefore to better understand the correlation between the two variables, a correlation matrix is produced.

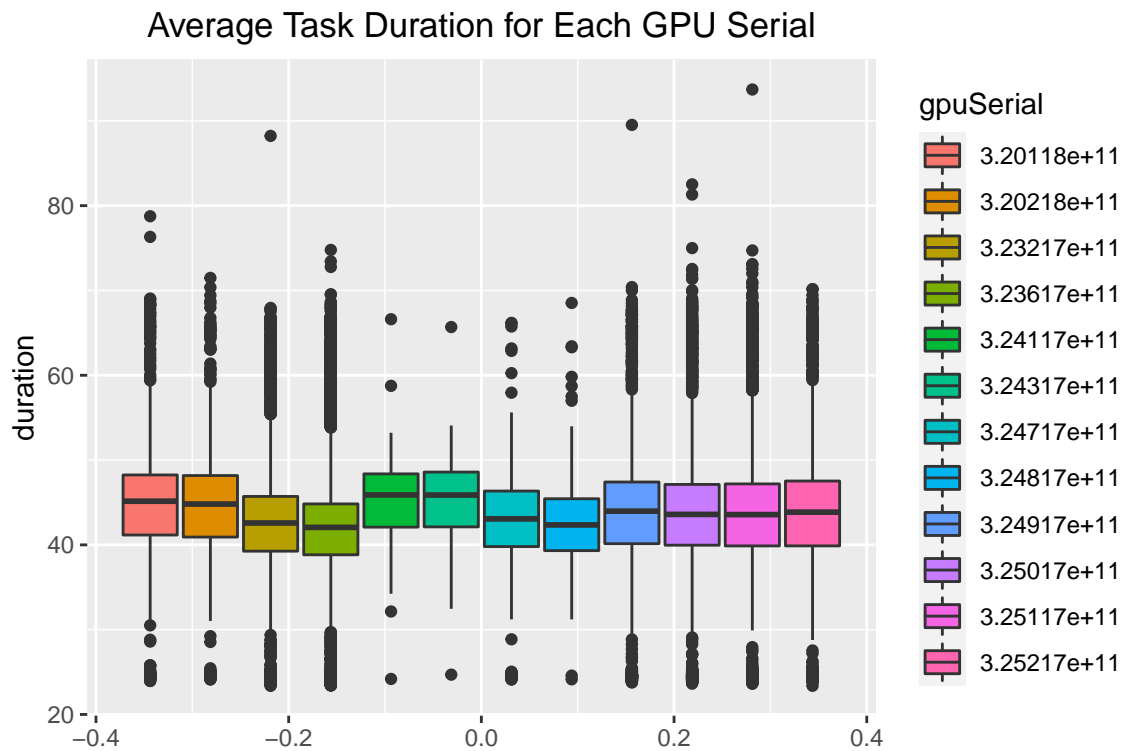
```
##           duration AvgPowerDrawWatt
## duration      1.0000000      0.5018929
## AvgPowerDrawWatt 0.5018929      1.0000000
```

The correlation matrix shows that there is some correlation between the **GPU Power Draw** and the **Duration** of the render task, however the correlation is not strong, therefore it could be deduced that the **Duration** of an event task does not influence the **Power Draw** or the other way around. Therefore, it prompt a question on what truly drives the increase in **Power Draw** or what caused the render task to run longer



Identifying GPU Card whose Performance Differ from Others

There are a total of 12 type of physical GPU cards used on the current Cloud Architecture, and we would like to gather the information regarding the performance of each type of GPU cards. To aid in identifying which GPU card perform better or worse than the rest, we would plot a graph in which it will show the average render run time for each type of GPU cards.



From the graph, we could observe that out of the 12 GPU cards type, there are 2 that performs slightly worse than the other which are GPU with serial number of **3.24117e+11** and **3.24317e+11** . To quantify the graph, we extract the average run time for each type of GPU cards.

| gpuSerial | duration |
|-------------|----------|
| 3.20118e+11 | 44.50196 |
| 3.20218e+11 | 44.26358 |
| 3.23217e+11 | 42.26390 |
| 3.23617e+11 | 41.68215 |
| 3.24117e+11 | 45.36851 |
| 3.24317e+11 | 45.29177 |
| 3.24717e+11 | 42.75674 |
| 3.24817e+11 | 42.67423 |
| 3.24917e+11 | 43.50489 |
| 3.25017e+11 | 43.24727 |
| 3.25117e+11 | 43.29552 |
| 3.25217e+11 | 43.42346 |

From both the graph and the numerical summaries above, we could deduct that out of the 12 distinct GPU Serials, the two GPU Serials that perform the worst are GPU Serial **3.24117e+11** and **3.24317e+11** with average render run time of **45.36851 Seconds** and **45.29177 Seconds** respectively. And the GPU Serial that perform the best is GPU Serial **3.23617e+11** with average render run time of **41.68215 Seconds**. The difference might not be much at glance but if we take into account the number of tiles that needs to be rendered which is 65025 tiles in total even after considering all 1024 hostnames working together at a same time, even 1 second difference in average rendering time would result in great inefficiency. Therefore it should be taken into consideration to opt for the fastest type of GPU cards.

Efficiency of the Task Scheduling

