**Part 3) Visualization (10%)**

Please download data in the link below:

<https://github.com/kaopanboonyuen/Python-Data-Science/blob/master/Dataset/vsl_data.zip>

You have to submit the following items to MyCourseVille:

* Source files from Power BI and Gephi
* File.docx – a capture screen of each question. It cannot be graded if we cannot map your answer to the question.
* All files must be renamed as “{student\_id}\_{firstname}\_Part3”, e.g., 6030133421\_Chaiyatad\_Part3.ipynb

**Part 3-1) Spatial Analysis and Visualization with Power BI**

1. ให้นิสิตใช้โปรแกรม PowerBI นำเข้าข้อมูลจากไฟล์ taxi\_od.csv ซึ่งเป็นข้อมูลการเดินทางด้วยรถแท็กซี่ภายในหนึ่งวัน ข้อมูลแต่ละแถวคือบันทึกเหตุการณ์การรับหรือส่งผู้โดยสาร ประกอบด้วยคอลัมน์ดังต่อไปนี้

|  |  |
| --- | --- |
| ชื่อคอลัมน์ | คำอธิบาย |
| taxi\_ID | หมายเลขของแท็กซี่ |
| trip\_ID | หมายเลข trip แท็กซี่หนึ่งคันมีได้หลาย trip |
| time | timestamp อยู่ในรูปแบบ yyyy-mm-dd hh:mm:ss เช่น 2016-09-23 06:03:46 |
| hour | timestamp เฉพาะส่วนชั่วโมง |
| status | O คือ origin หรือ pick-up แสดงว่าเป็นการรับผู้โดยสาร  D คือ destination หรือ drop-off แสดงเป็นการส่งผู้โดยสาร |
| latitude | พิกัดตำแหน่งละติจูดแบบทศนิยม |
| longitude | พิกัดตำแหน่งลองจิจูดแบบทศนิยม |
| rounded\_latitude | พิกัดตำแหน่งละติจูดแบบทศนิยม ที่ลดความละเอียดเหลือเพียงทศนิยมสองตำแหน่ง เพื่อทำให้จุดที่อยู่ในพื้นที่เดียวกันมีพิกัดเท่ากัน |
| rounded\_longitude | พิกัดตำแหน่งลองจิจูดแบบทศนิยม ที่ลดความละเอียดเหลือเพียงทศนิยมสองตำแหน่ง เพื่อทำให้จุดที่อยู่ในพื้นที่เดียวกันมีพิกัดเท่ากัน |
| distance | ระยะทางของทริป หน่วยเป็นเมตร มีค่าเฉพาะเมื่อ status เป็น D |
| duration | เวลาที่ใช้ในทริป หน่วยเป็นวินาที มีค่าเฉพาะเมื่อ status เป็น D |

ตัวอย่างข้อมูล

taxi\_ID,trip\_ID,time,hour,status,latitude,longitude,rounded\_latitude,rounded\_longitude,distance,duration

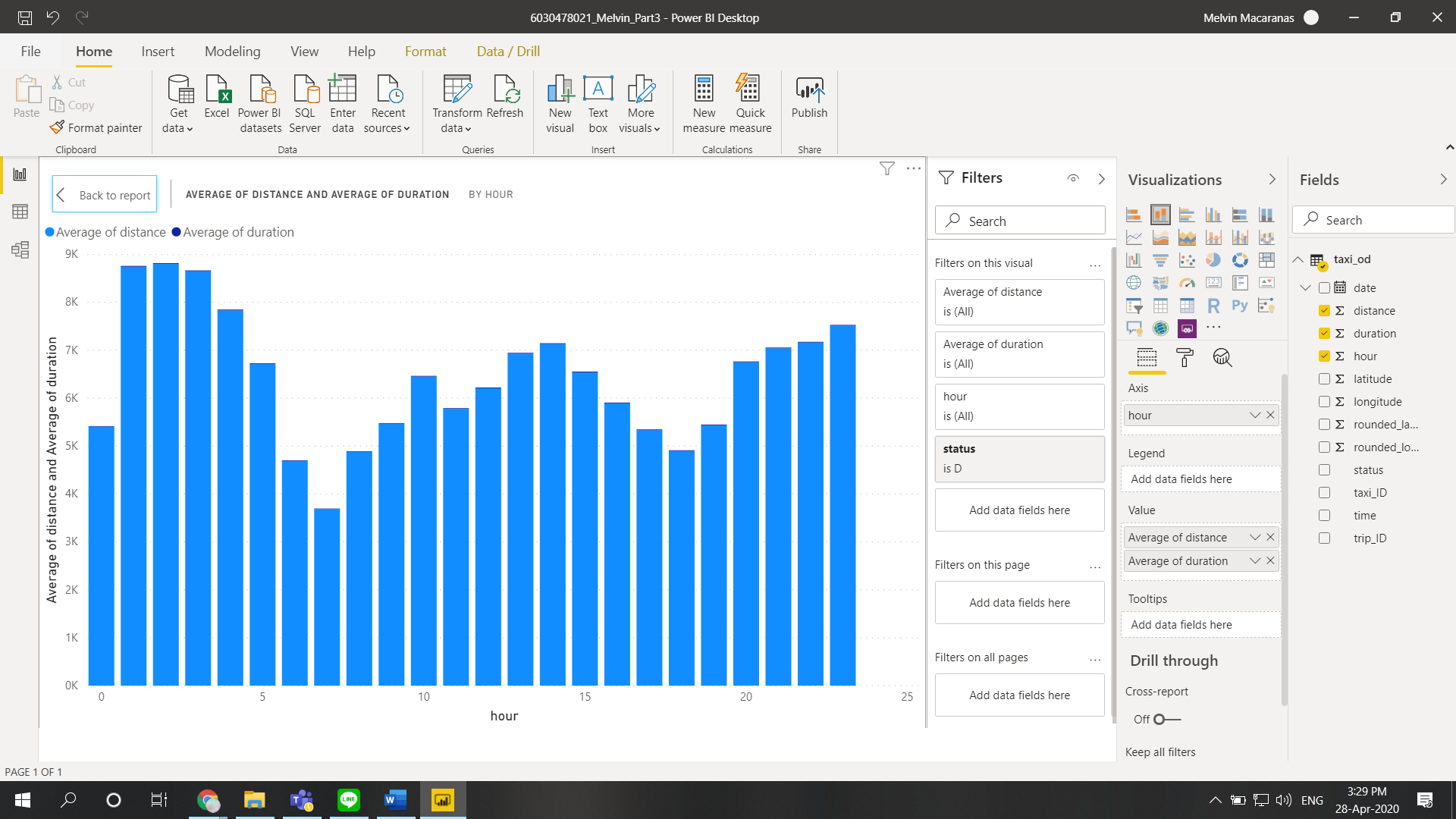
AK2306,AK2306\_T\_6203,2016-09-23 06:03:46,6,O,13.73136425,100.650322,13.73,100.65,,

AK2306,AK2306\_T\_6203,2016-09-23 06:14:46,6,D,13.73135662,100.6502838,13.73,100.65,5,11

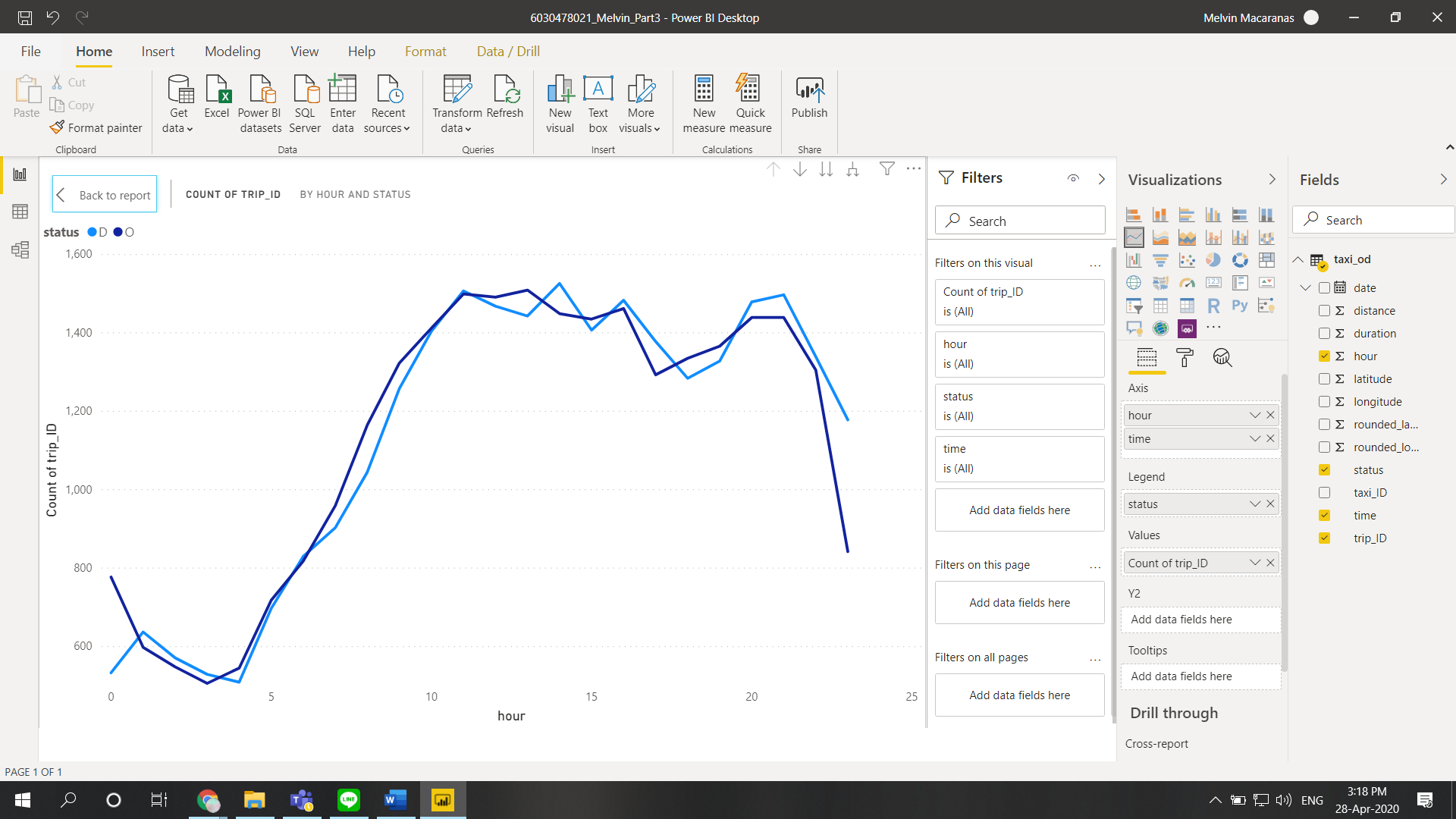
ให้นิสิตสร้าง visualization จากข้อมูล เพื่อตอบคำถามข้างล่างนี้ นิสิตสามารถเลือกใช้ visualization ที่คิดว่าช่วยตอบคำถามและสื่อสารได้อย่างมีประสิทธิภาพที่สุด กี่ visualization ก็ได้ โดยให้ capture หน้าจอมาแปะ แล้วเขียนคำอธิบายประกอบภาพ

* 1. **การเดินทางด้วยรถแท็กซี่ในแต่ละช่วงเวลาของวันมีปริมาณมากน้อยอย่างไร มีการเปลี่ยนแปลงอย่างไรบ้าง**

จะเห็นว่าตั้งแต่ช่วงเวลาประมาณ 10 โมงเช้าถึงสามทุ่มมีการใช้ taxi หนาแน่นมาก และมีการรับและส่งผู้โดยสารในปริมาณเท่าๆกัน เพราะคนน่าจะแค่นั่งไปที่ไม่ไกลมากเพราะรถติด ถ้าไปไกลก็นั่งรถไฟฟ้าเอา แต่จะมีตกลงมาเล็กน้อยในช่วง 5 โมงถึง 1 ทุ่มเพราะคนเลิกงาน โรงเรียนเลิกเรียน รถติดมาก คนจึงน่าจะเลือกเดินทางด้วยวิธีอื่น แต่ช่วงเวลาประมาณ 5 ทุ่มจะมีการส่งผู้โดยสารมากกว่ารับ น่าจะเป็นเพราะว่าช่วงเวลาประมาณ 3-4 ทุ่มผับยังไม่ทันปิด คนที่กลับช่วงนี้เป็นคนบ้านไกลเลยรีบกลับ เลยทำให้ช่วง 5 ทุ่มพวกเขาพึ่งถึงบ้าน แต่พอช่วงหลังเที่ยงคืนผับเริ่มปิดกันแล้ว การรับผู้โดยสารสูงมาก เพราะแท็กซี่จะไปรวมกันแถวข้าวสาร ทองหล่อเพื่อรอรับผู้โดยสาร และทยอยส่งเสร็จตอนเวลาประมาณตี 1 เป็นต้นไปซึ่งตัวกราฟจะกลับมาสมดุลพอๆกัน



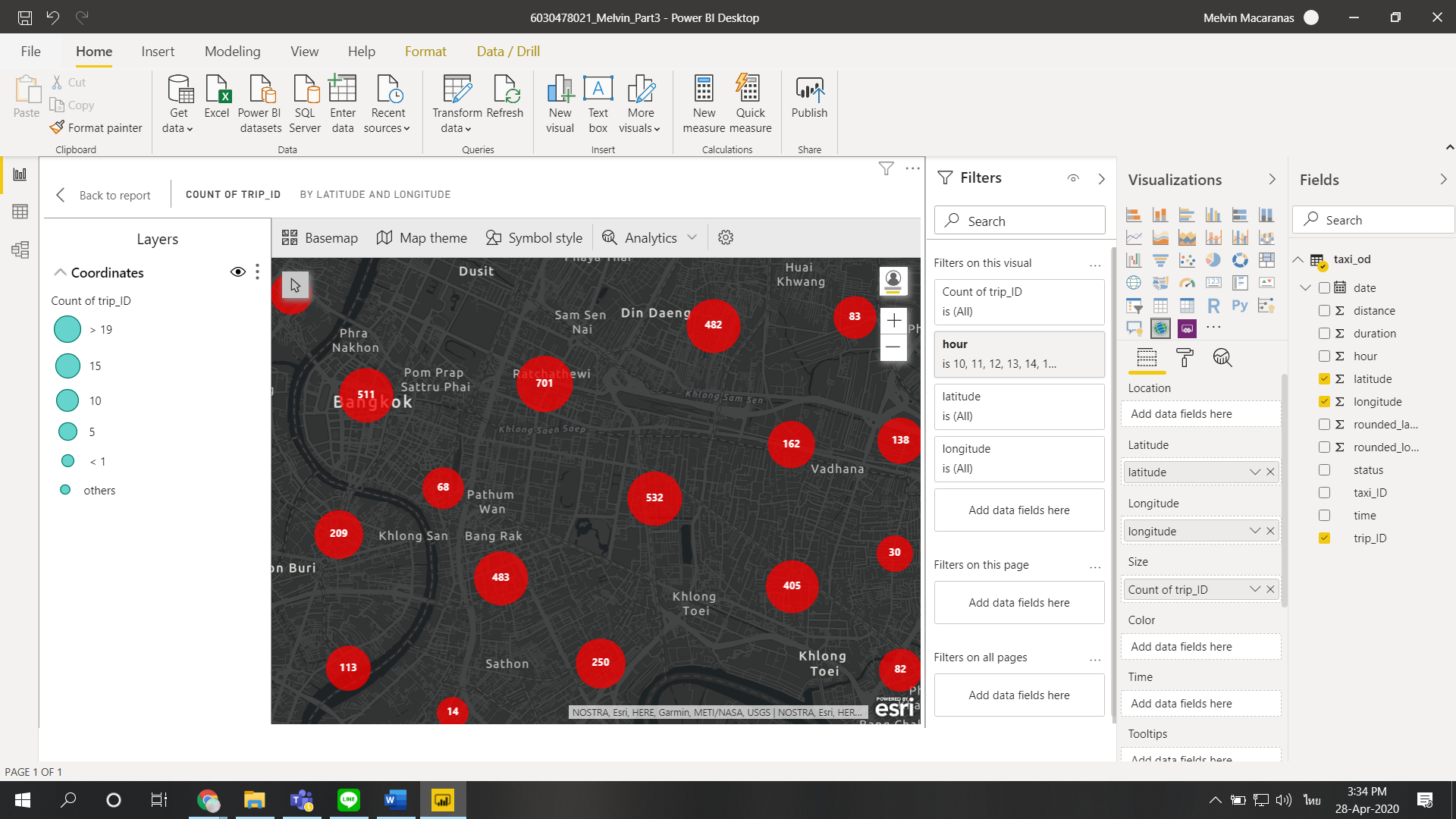
ภาพด้านบนแสดงระยะทางเฉลี่ยในแต่ละชั่วโมง



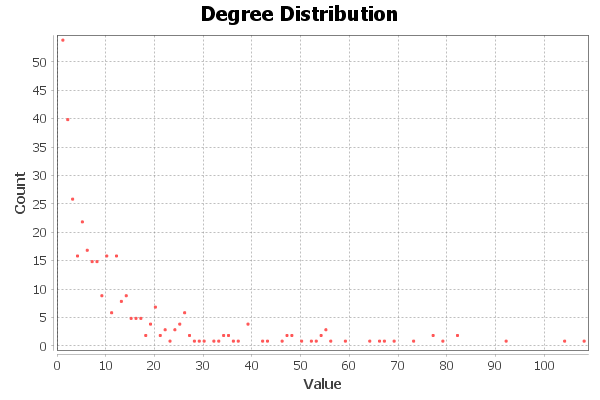
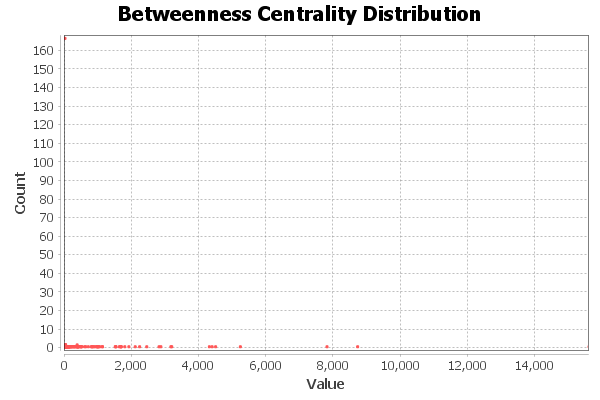
ภาพด้านบนแสดงจำนวนทริปในแต่ละชั่วโมง

* 1. **พื้นที่ย่านใดในกรุงเทพฯที่มีผู้ขึ้นลงแท็กซี่หนาแน่นที่สุด ในช่วงเวลาที่มีการใช้บริการมากที่สุด ตามข้อมูลที่ได้จากข้อที่แล้ว**

สำหรับในช่วงเวลาที่มีการใช้บริการมากที่สุด จะพบว่า บริเวณสยาม ราชเทวี จุฬา ชิดลม ราชประสงค์ จะมีการขึ้นลงหนาแน่นที่สุด



**Part 3-2) Network Analysis and Visualization with Gephi**

1. ให้นิสิตใช้โปรแกรม Gephi เปิดไฟล์ hero-social-network.gephi ซึ่งเป็นข้อมูลเครือข่ายความสัมพันธ์ของตัวละครในนิยายภาพแนวซุปเปอร์ฮีโร่ โดยแต่ละ node เป็นตัวละครแต่ละตัวและ edge แสดงความสัมพันธ์ระหว่างตัวละคร ให้นิสิตใช้เครื่องมือวิเคราะห์ในโปรแกรม Gephi และสร้าง network visualization เพื่อตอบคำถามข้างล่างนี้ โดยให้ capture หน้าจอที่แสดงผลลัพธ์ที่เกี่ยวข้องมาแปะกี่ภาพก็ได้ พร้อมทั้งเขียนคำอธิบายประกอบภาพ
   1. **เน็ตเวิร์คนี้มีคุณลักษณะใกล้เคียงกับเน็ตเวิร์คโมเดลประเภทใด (random, small-world, scale-free) เมื่อพิจารณาจาก degree distribution และ average path length**

คาดว่าน่าจะเป็น scale-free เพราะจะเห็นได้จาก degree distribution ว่า node ส่วนใหญ่จะมี degree น้อย จะมีแค่บาง node เท่านั้นที่มี degree มากโดดเด่นออกมา และเมื่อพิจารณา avg. path length ควบคู่ด้วยก็จะยิ่งสนับสนุนแนวคิดดังกล่าวอีกด้วย

* 1. **ใครคือตัวละครที่มีความสำคัญที่สุด 3 อันดับแรก เมื่อพิจารณาจาก centrality measures ที่แสดงถึงความเป็นศูนย์กลางของเน็ตเวิร์ค และเป็นตัวเชื่อมโยงกลุ่มต่าง ๆ ในเครือข่าย**

จากการวิเคราะห์โดยพิจารณาจาก Degree และ Betweeness Centrality จะพบว่า TOP3 คือ Spider-man, Captain America, Thor แต่เมื่อพิจารณาจาก Closeness Centrality จะมีความต่างคือจะได้ Beast แทนที่จะเป็น Thor

ดังนั้นจึงสรุปว่าตัวละครที่สำคัญที่สุด 3 อันดับแรกจึงเป็น Spider-man, Captain America และ Thor

