Natthanan Bhukan 61070507206

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

CPE 213 Data Models, 2/2020

LAB Lecture 4: Data Visualization I (Distributions)

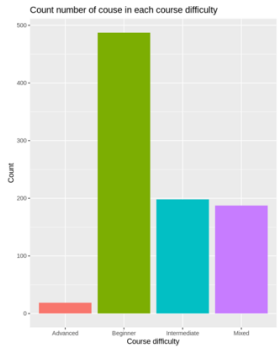
Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

---

1. Find the dataset you like
   - Dataset ที่เลือกคือ coursera courses เป็น Data ที่เกี่ยวกับ Course ที่มีอยู่ใน Coursera มาส่วนหนึ่ง โดยจะมีข้อมูล ชื่อ Course องค์กรที่สอน ระดับความยาก รูปแบบการเรียน คะแนนรีวิว และ จำนวนนักเรียนที่ลงเรียน
2. Perform 2 analyses of data distribution (1 categorical, 1 numerical)

   2.1 Categorical

   Categorical variable in number of course in each course difficulty แสดงจำนวน Course ที่มีอยู่ในข้อมูลโดยแยกเป็นแต่ละ Course difficulty



   Categorical variable in number of course in each course certificate type แสดงจำนวน Course ที่มีอยู่ในข้อมูลโดยแยกเป็นแต่ละ Course certificate type

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

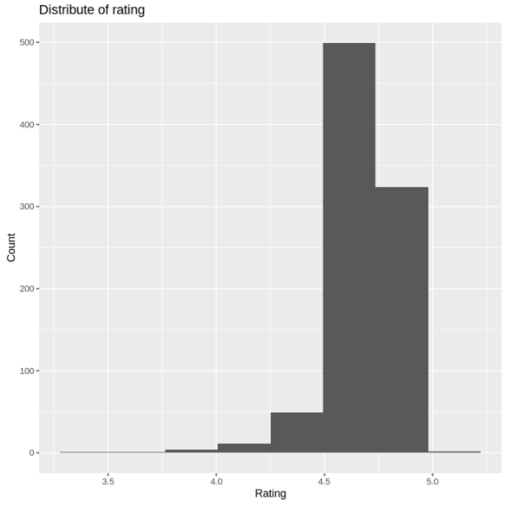
CPE 213 Data Models, 2/2020

LAB Lecture 4: Data Visualization I (Distributions)

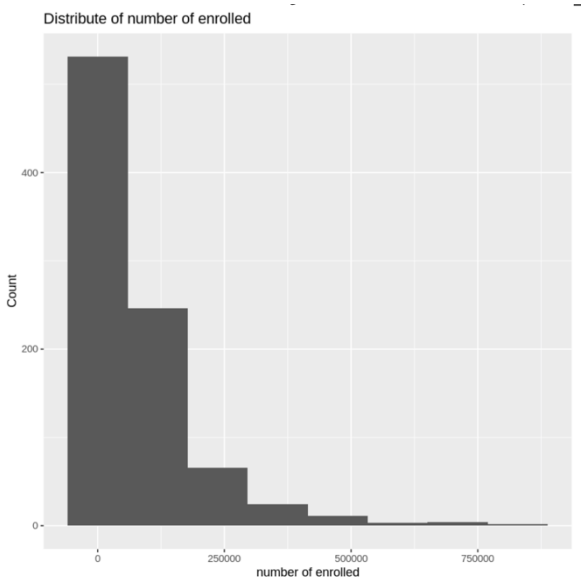
Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

2.2 Numerical

Distribute of rating แสดงการกระจายตัวข้อมูล ใน Column rating



Distribute of number of enrolled แสดงการกระจายตัวข้อมูล ใน Column Number of enrolled



King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

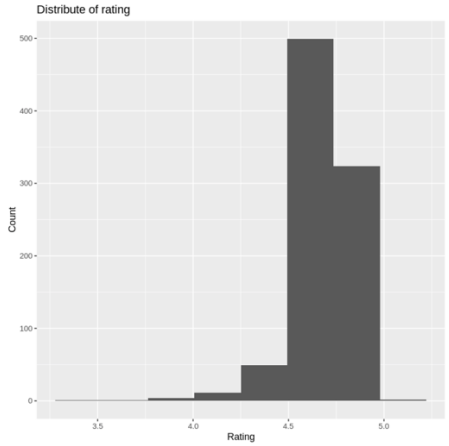
CPE 213 Data Models, 2/2020

LAB Lecture 4: Data Visualization I (Distributions)

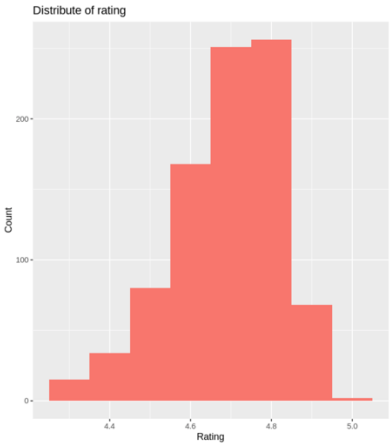
Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

---

3. Explain the results and your understanding of data and the process



จาก ภาพนี้ จะเห็นได้ว่า Distribute of rating มี Outliers เลยทำการหาค่า Q1 กับ IQR เพื่อนำเอาค่า Outliers ออก โดยเมื่อนำค่า Outliers ออก และ Plot Histogram ออก มาใหม่ จะได้ดังรูปข้างล่าง



โดยจะเห็นได้ว่า ข้อมูลมีกระจายตัวใกล้เคียงกับ Normal Distribution หรือ Bell Shape ซึ่งจะเห็นได้ว่า นักเรียนที่เรียน Coursera มีการให้คะแนน Course ที่ค่อนข้างกระจายตัวอยู่มาก และ ให้ค่อนข้างสูง เพราะ คะแนนต่ำสุดที่มีหลังจากตัด Outliner ออกคือ 4.2 เต็ม 5 ซึ่งเราก็อาจจะใช้

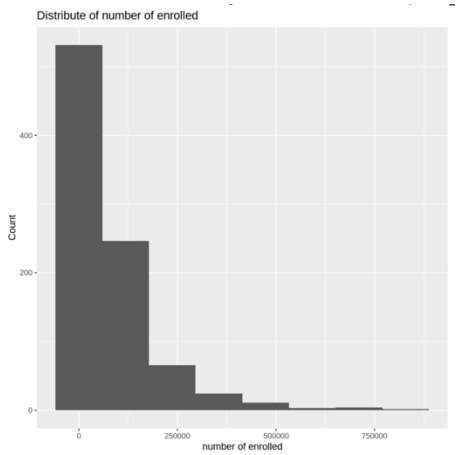
King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

CPE 213 Data Models, 2/2020

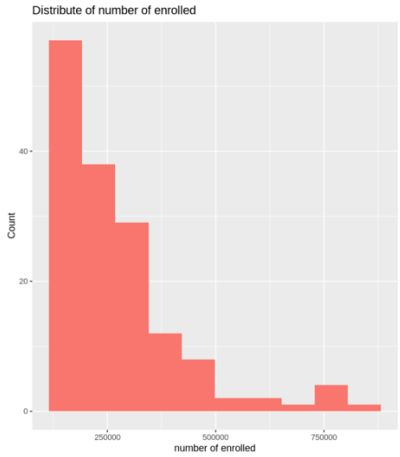
LAB Lecture 4: Data Visualization I (Distributions)

Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

ข้อมูลตรงนี้มาปรับปรุง Course ก็ได้โดย Course ที่คะแนนไม่ได้อยู่ช่วงของการกระจายตัวนี้ ต้องมีการปรับปรุงมากขึ้น



ในส่วนต่อมา จาก ภาพนี้ จะเห็นได้ว่า Distribute of number of enrolled มี Outliers เลยทำการหาค่า Q1 กับ IQR เพื่อนำ เอาค่า Outliers ออก โดยเมื่อนำค่า Outliers ออก และ Plot Histogram ออก มาใหม่ จะได้ดังรูปข้างล่าง



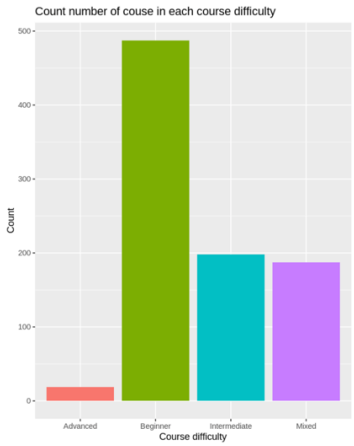
King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

CPE 213 Data Models, 2/2020

LAB Lecture 4: Data Visualization I (Distributions)

Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

โดยจะเห็นได้ว่า ข้อมูลมีกระจายตัวใกล้เคียงกับ Reverse J-Shaped ซึ่งจะเห็นได้ว่า นักเรียนที่เรียน Coursera ส่วนมากจะไม่ได้ลงทุก Course เพราะ จะเห็นได้ว่าจำนวนคนที่ลง Course ที่มีมากกว่า 500000 ค่อนข้างน้อย ซึ่งมันจะหมายความว่า จะมี แค่บาง Course ที่เป็น Course มีชื่อเสียง ของ Coursera ที่ทำให้ทุกคนอยากมาเรียนที่ Coursera โดยถ้าเราอาจจัด Promotion ให้กับ Course ที่มีคนลงน้อย ๆ เพื่อสร้างแรงจูงใจ เรียกคนมาเรียนมากขึ้นก็ได้



ในส่วนต่อมา เมื่อมาดู Bar Graph เราจะเห็นได้ว่าจำนวน Course ที่มีอยู่ใน Coursera ส่วนมาก ระดับความยาก จะเป็น Beginner ส่วน Advanced จะเป็นส่วนที่น้อยที่สุด ซึ่งถ้าเรามาดูจำนวนคนที่ลงเรียนเฉลี่ย Coursera ของแต่ละ ระดับความยาก จะได้ดังนี้

	course_difficulty	avg_enrolled
	<chr>	<dbl>
1	Advanced	205000.0
2	Beginner	268809.5
3	Intermediate	270000.0
4	Mixed	272222.2

ซึ่งก็คือ Advanced Course มีคนลงค่อนข้างใกล้เคียงกับ Course อื่น โดยถ้าเราสามารถหา Course ที่อยู่ในระดับ Advanced มาเปิดสอน ก็อาจจะเพิ่มยอดนักเรียนก็ได้ เพราะ Advanced

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

CPE 213 Data Models, 2/2020

LAB Lecture 4: Data Visualization I (Distributions)

Assign Date: 19 Feb 2021 Due Date: 4 Mar 2021

Course มี จำนวนน้อยที่สุด และห่างกับ Course ระดับอื่น อย่างมาก แต่มีจำนวนนักเรียนเฉลี่ยพอกับ ระดับอื่น ๆ

โดยเมื่อลองดู ข้อมูลของ Course certificate type เราจะเห็นได้ว่า Professional Certificate มีจำนวนน้อยที่สุด และ น้อยกว่า Course อื่น ๆ เป็นอย่างมาก เมื่อทำการดูจำนวนนักเรียนที่เรียนเฉลี่ย ในแต่ละ Course certificate type ก็จะได้ดังนี้

	course_Certificate_type	avg_enrolled
	<chr>	<dbl>
1	COURSE	249239.1
2	PROFESSIONAL CERTIFICATE	288571.4
3	SPECIALIZATION	297272.7

ซึ่งจะเห็นได้ว่า จำนวนนักเรียน ในหมวด Professional certificate มีจำนวนนักเรียนเฉลี่ย เป็นอันดับ 2 โดยจากข้อมูลตรงนี้ถ้าสามารถหา Course ที่เป็น Professional certificate มาเปิดสอน ก็อาจทำให้มีนักเรียนมาเรียน มากขึ้นก็ได้

เหมือนกับ ระดับความยากของ Course ที่เป็นระดับ Advance โดยถ้าทำการหา Course ในระดับ Advance และ อยู่ในหมวด Professional certificate มาเปิดสอน โดยมีการออก Promotions เล็กน้อย ก็น่าจะสามารถให้นักเรียนมาเรียนใน Coursera เพิ่มมากขึ้นได้