Hands-on Activity #2

The Secret Sauce of The Secret Sauce Podcast

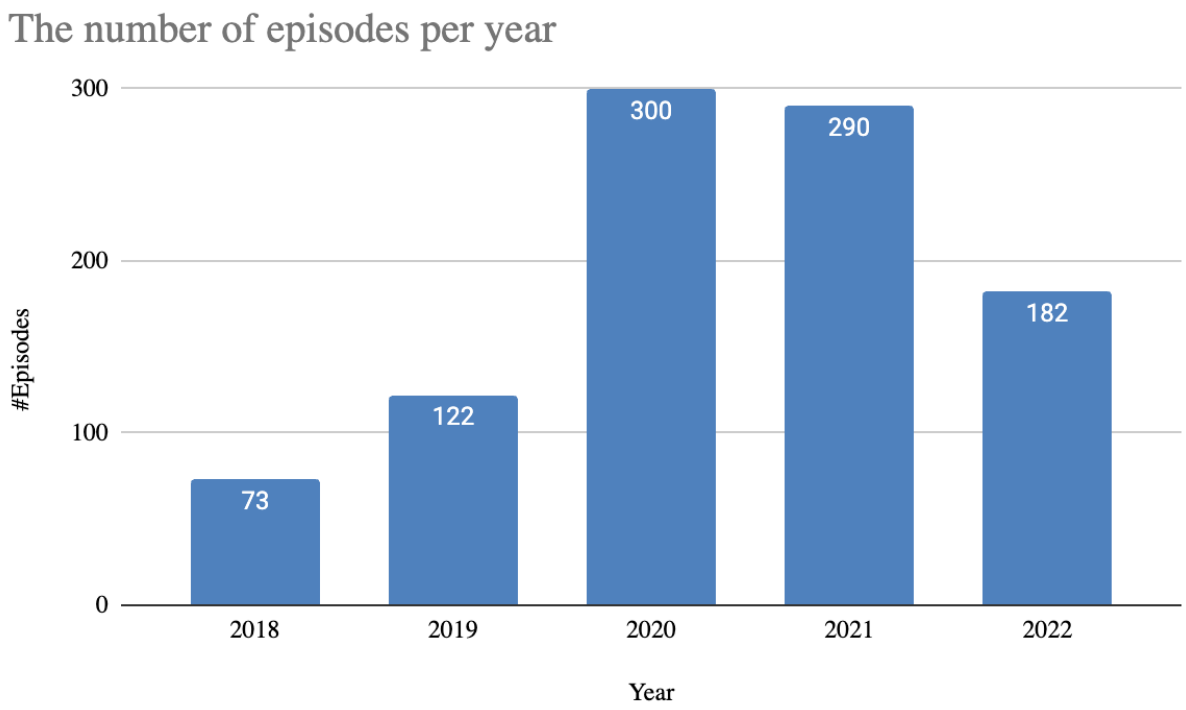
จากการบ้านใน Hands-on Activity #1 เราได้ผ่านขั้นตอนการเตรียมข้อมูลกันไปเรียบร้อยแล้ว ครั้งนี้เราจะเริ่มทำการวิเคราะห์ข้อมูลเพื่อตอบคำถามว่า

“อะไรคือ **เคล็ดลับความสำเร็จ** ของ The Secret Sauce”

โดยหวังว่าผลที่ได้จากการวิเคราะห์จะช่วยให้ทีมงาน The Secret Sauce เข้าใจพฤติกรรมของแฟนรายการ ทราบถึงปัจจัยที่ทำให้รายการประสบความสำเร็จในช่วงที่ผ่านมา และสามารถนำ Insights ที่ได้ไปใช้ในการวางแผนคอนเทนต์ หัวข้อพูดคุย การหาแขกรับเชิญ รวมไปถึงกลยุทธ์ในการแข่งขันกับช่องอื่น ๆ ได้ในอนาคต

Exploratory Data Analysis (EDA) เป็นกระบวนการที่จะช่วยให้เราสามารถทราบคำตอบเหล่านั้นได้ ผ่านการตรวจสอบข้อมูลเบื้องต้นที่จะช่วยให้เราเข้าใจเกี่ยวกับพื้นฐานของข้อมูลชุดนั้น เช่น ข้อมูลมีรูปแบบกระจายตัวรูปแบบใด, มีแนวโน้มหรือ trend ที่น่าสนใจหรือไม่ หรือมี outlier มากน้อยแค่ไหน เพื่อใช้ในการตั้งสมมติฐานก่อนจะนำไปสร้างแบบจำลองทางสถิติ (Statistical Modeling) ต่าง ๆ ต่อไป

Exploratory Data Analysis

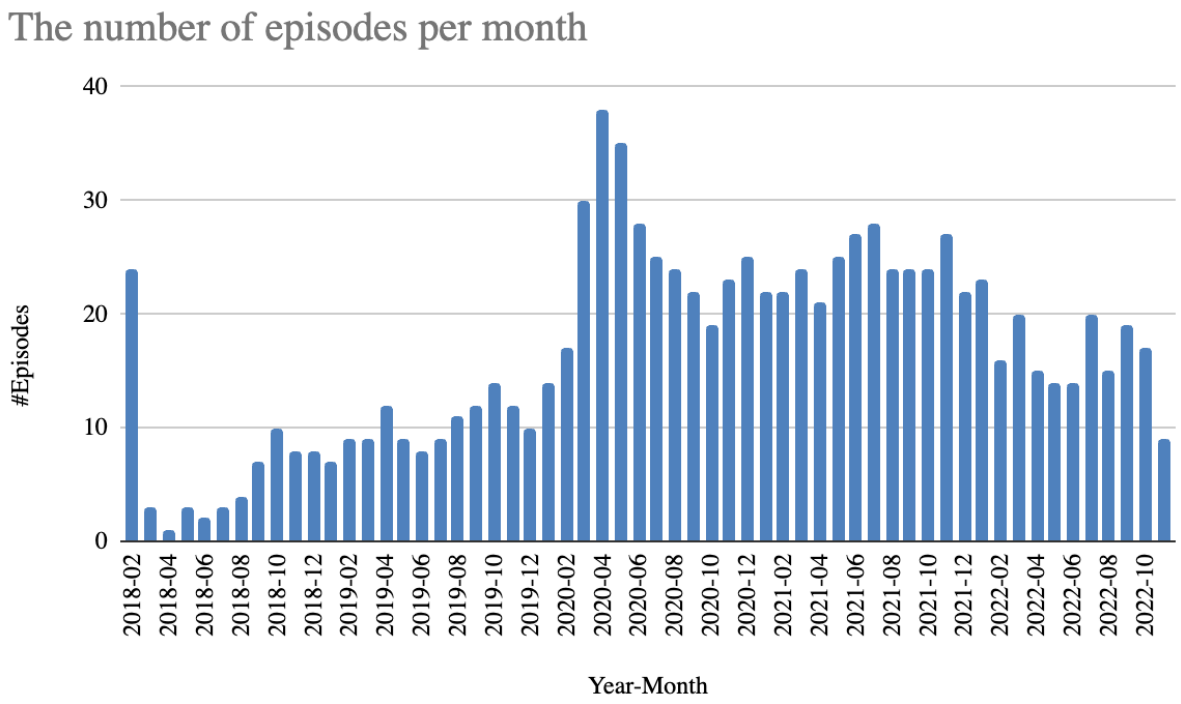


จำนวน Episode ของรายการ The Secret Sauce ที่ปล่อยบน YouTube ในแต่ละปี ตั้งแต่เริ่มต้นจนถึงปัจจุบัน

เป็นเวลามากกว่า 4 ปีแล้ว ที่ช่อง The Secret Sauce บน YouTube ผลิตเนื้อหาให้แฟนรายการได้ติดตาม หากพิจารณาจำนวน episode ที่ผลิตในแต่ละปี จะพบว่ารายการมีแนวโน้มผลิตคอนเทนต์เพิ่มขึ้นเรื่อย ๆ และเริ่มคงที่ในปี 2020 เป็นต้นมา

⚠️ **หมายเหตุ**: การเปรียบเทียบข้อมูลในกราฟข้างต้นต้องทำด้วยความระมัดระวัง เนื่องจากข้อมูลปี 2018 กับ 2022 เป็นข้อมูลที่ไม่ครบปี และในปี 2018 มีการเอา episode เก่า ๆ ตั้งแต่ปี 2017 มาอัปโหลดขึ้นพร้อม ๆ กันด้วย ดังนั้นเพื่อช่วยให้กราฟนี้ไม่ชี้นำไปในทางที่ผิด เราอาจจะต้องกรองข้อมูลและเลือกแสดงข้อมูลของปี 2019-2021 เท่านั้น

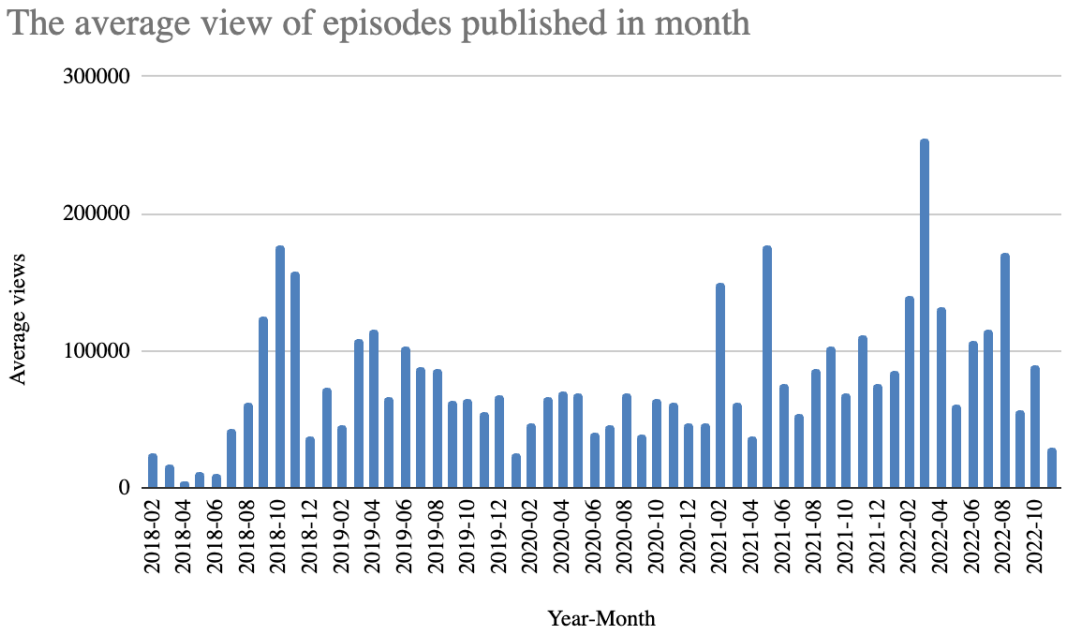
เพื่อให้เห็นแนวโน้มการผลิตรายการใหม่ละเอียดขึ้น เราอาจพิจารณาเป็นจำนวน episode ที่ผลิตรายเดือนแทน



กราฟแสดงจำนวน Episode รายเดือน ตั้งแต่เริ่มรายการจนถึงปัจจุบัน

จะพบว่าเดือนกุมภาพันธ์ ปี 2018 (เดือนแรก) มีจำนวน episode โดดจากเดือนอื่น ๆ ในปีเดียวกัน เพราะเป็นการอัปโหลดตอนเก่า ๆ จากนั้นจำนวน Episode ในแต่ละเดือนมีแนวโน้มที่จะผลิตมากขึ้น โดยเพิ่มขึ้นสูงสุดในปี 2020 และคงที่ในปีต่อ ๆ มา จนถึงปัจจุบัน

ด้วยกระแสตอบรับที่ดี ทำให้รายการมีจำนวนผู้ติดตามเพิ่มมากขึ้นตามลำดับ คำถามที่ตามมาคือ ยอดวิวมีแนวโน้มปรับตัวสูงขึ้นตามจำนวนแฟนรายการที่เพิ่มมากขึ้นหรือไม่ (ตามเวลา)



กราฟแสดงยอดวิวเฉลี่ยของวิดีโอที่ปล่อยในแต่ละเดือน ตั้งแต่เริ่มรายการจนถึงปัจจุบัน

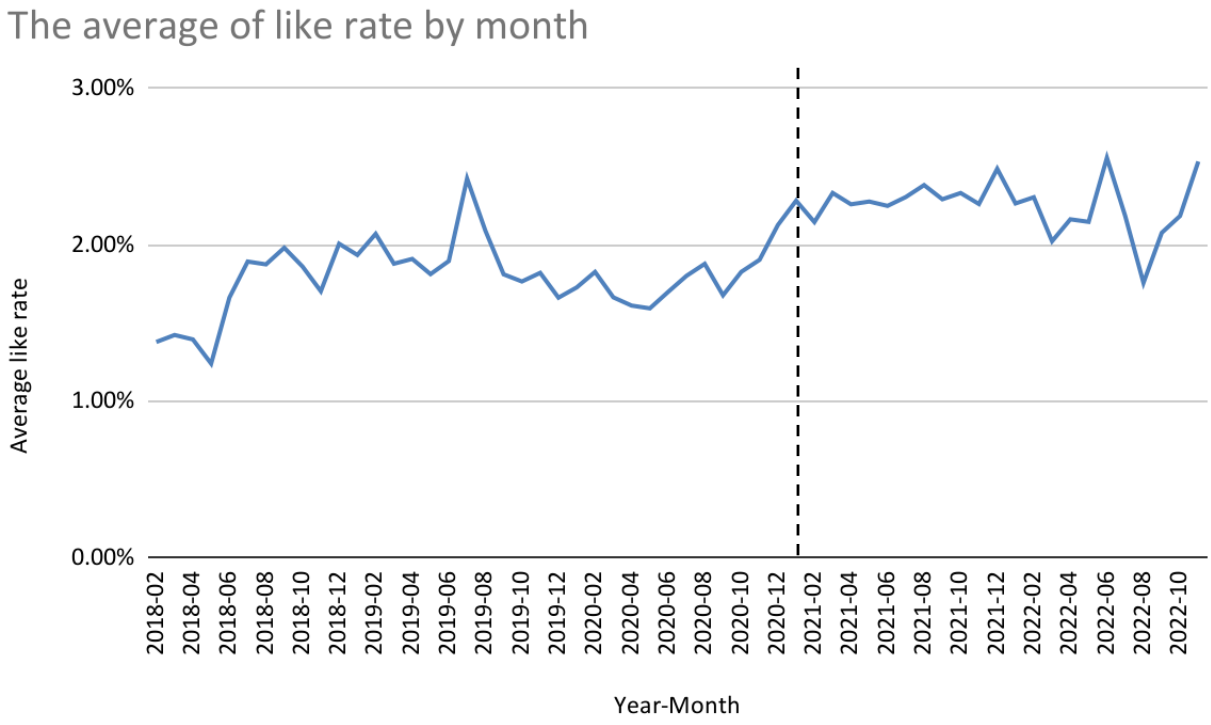
จากกราฟข้างต้น จะเห็นได้ว่ายอดวิวเฉลี่ยของวิดีโอที่ปล่อยในแต่ละเดือน ไม่มีแนวโน้มเพิ่มขึ้นหรือลดลงอย่างชัดเจน และบางเดือนนั้นมียอดวิวเฉลี่ยสูงกว่าเดือนอื่น ๆ หลายเท่าตัว เช่น เดือนมีนาคม ปี 2022 ซึ่งสาเหตุเกิดจากบาง episode ในเดือนนั้น ๆ ได้รับความนิยมสูงมากหรือเรียกว่า กลายเป็น “ไวรัล” ทำให้ค่าเฉลี่ยในเดือนนั้นสูงขึ้นผิดปกติ ในเชิงเทคนิคเราจะเรียกยอดวิวของตอนที่ไวรัลนี้ว่าเป็น Outlier ในข้อมูล

เมื่อเราลองพิจารณาค่าทางสถิติต่าง ๆ ของยอดวิวของ episode ทั้งหมดในเดือนมีนาคมปี 2022 พบว่ายอดวิวเฉลี่ยในเดือนนี้จะอยู่ที่ 254,503 views ซึ่งหากดูที่ค่ามัธยฐาน (Median) จะพบว่าอยู่ที่ 106,492 views หรือครึ่งหนึ่งของ episode ทั้งหมดในเดือนนี้ มียอดวิวน้อยกว่าค่านี้ ในขณะที่มี Episode ที่เป็น Outlier ที่มียอดวิวสูงถึง 1.3 ล้านวิว

2022-03	
AVERAGE	254,503
MEDIAN	106,492
PERCENTILE 25	55,691
PERCENTILE 75	277,441
MIN	11,822
MAX	1,343,306

ตารางแสดงค่าทางสถิติของ Episodes ต่าง ๆ ในเดือนมีนาคม ปี 2022

อีกหนึ่งปัจจัยที่แสดงถึงกระแสตอบรับของผู้ชม คือ จำนวนไลก์ (likes) ของแต่ละ episode แต่เนื่องจากยอดไลก์อาจจะแปรผันตามยอดวิว (ยิ่งยอดวิวสูงจำนวนการกดไลก์ก็จะยิ่งสูงขึ้นตาม) เราจะคำนวณค่าอัตราการกดไลก์ (likeRate) จากจำนวนไลก์ทั้งหมดหารด้วยยอดวิวของแต่ละ episode เพื่อใช้ในการเปรียบเทียบแทน



กราฟแสดงแนวโน้มของ likeRate เฉลี่ยของแต่ละ Episode ตามเวลาที่เปลี่ยนไป

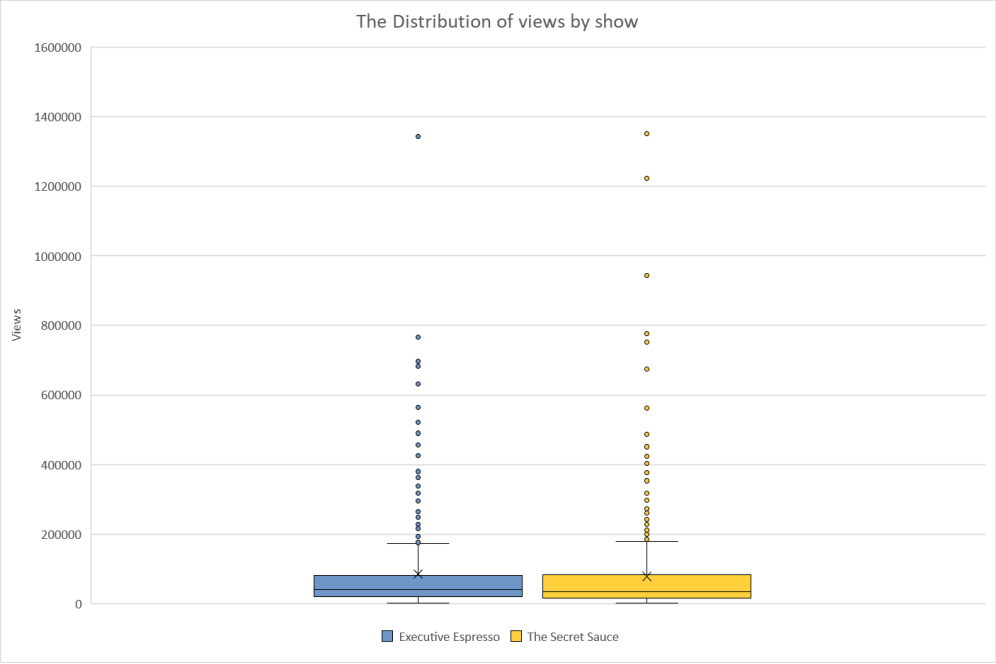
จากกราฟแสดงอัตราการกดไลก์เฉลี่ยรายเดือนจะพบว่าอัตราการกดไลก์มีแนวโน้มเพิ่มสูงขึ้นเล็กน้อยจากปีแรก ๆ (2018) และเริ่มคงที่อยู่ที่ประมาณ 2% ในช่วงปีหลัง ๆ (2021 เป็นต้นไป)

ช่วงปลายปี 2019 The Secret Sauce ได้มีการออกรายการย่อยใหม่ชื่อว่า Executive Espresso เป็นรายการที่เน้นการนำเสนอความเคลื่อนไหวในโลกเศรษฐกิจ-ธุรกิจ-เทรนด์ โดยจะย่อยมาเล่าให้ฟังแบบสั้น ๆ กระชับ แต่เข้มข้นเหมือนเอสเพรสโซหนึ่งถ้วย ดังนั้นในการวิเคราะห์ เราอาจจะอยากเปรียบเทียบความนิยมของทั้งสองรายการจากยอดวิวที่แต่ละ episode ได้รับ

	Executive Espresso	The Secret Sauce
AVERAGE	85,496	78,229
MEDIAN	40,375	35,190
PERCENTILE 25	19,980	16,759
PERCENTILE 75	81,795	83,227
MIN	2,244	3,164
MAX	1,343,306	1,351,575

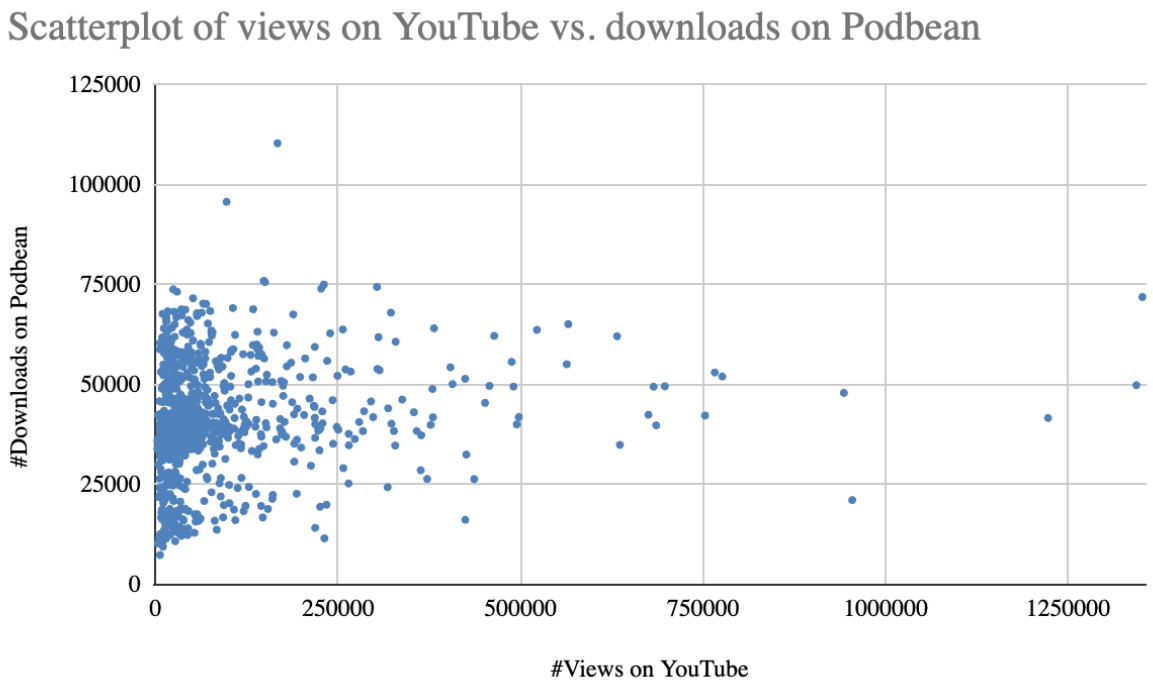
ตารางแสดงค่าสถิติของยอดวิวแยกตามรายการย่อย

จะเห็นว่า Executive Espresso ได้รับกระแสตอบรับพอ ๆ กันกับ The Secret Sauce สังเกตได้จากค่าสถิติต่าง ๆ ที่ใกล้เคียงกัน เพื่อให้เห็นภาพละเอียดมากขึ้น เราอาจจะใช้ Boxplot ในการเปรียบเทียบกากระจายตัวของยอดวิวของทั้งสองรายการได้ ตามกราฟด้านล่าง



กราฟแสดงการกระจายของยอดวิวแยกตามรายการ เรียกกราฟนี้ว่า ‘Boxplot’ หรือ ‘Box and Whiskers’

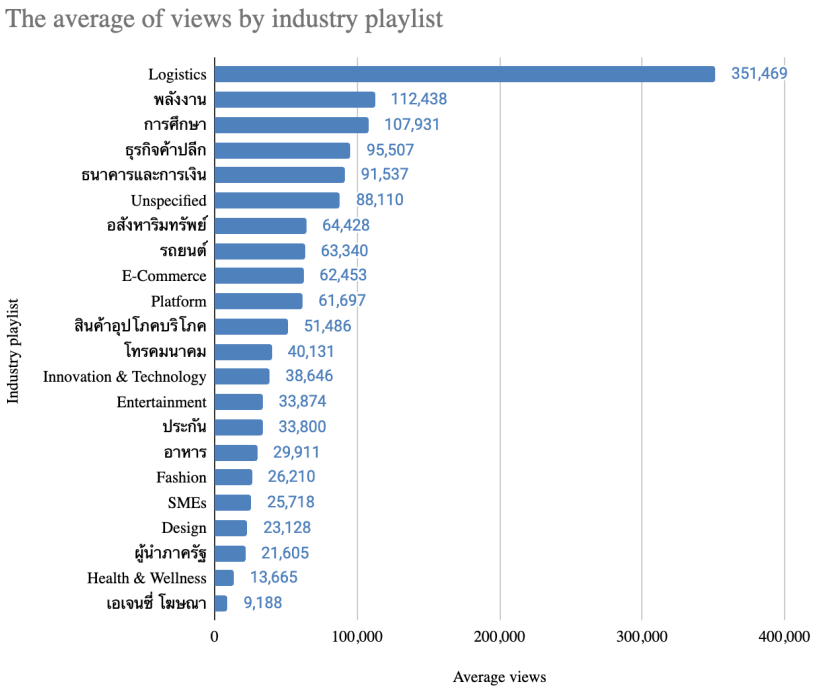
เนื่องจาก The Secret Sauce ได้รับกระแสตอบรับจากผู้ฟังที่ค่อนข้างดี จึงมีการเผยแพร่ในหลากหลายช่องทาง คำถามถัดมาคือ กระแสตอบรับที่ได้จากแต่ละ Platform นั้นแตกต่างกันหรือไม่ เพื่อตอบคำถามนี้ เราจึงได้ไปทำการเก็บข้อมูลเพิ่มเติม จากเว็บไซต์ Podbean ซึ่งเป็นอีกหนึ่งช่องทางที่รายการ The Secret Sauce ใช้ในการนำเสนอในรูปแบบ Podcast ให้ผู้ติดตามสามารถ download ฟังเนื้อหาในแต่ละตอนได้ โดยเราจะทำการเปรียบเทียบระหว่างยอด download ของแต่ละ episode ใน Podbean กับยอดวิวใน YouTube



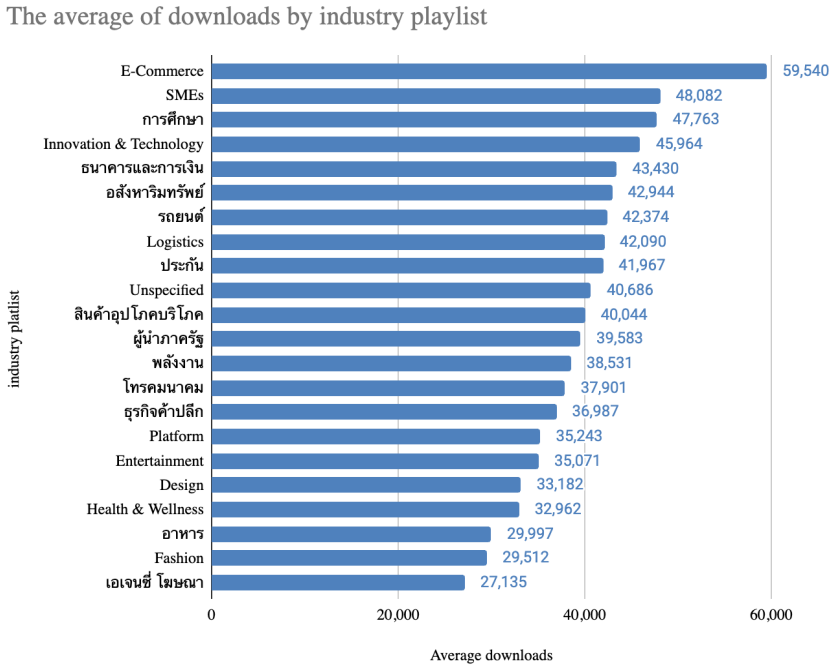
กราฟ scatter plot แสดงความสัมพันธ์ระหว่างยอดวิวบน YouTube และยอดดาวน์โหลดบน Podbean

จากกราฟจะพบว่า ทิศทางกระแสตอบรับของแต่ละ episode ไม่ได้มีความสัมพันธ์กันเท่าไรนัก บางตอนอาจได้รับความนิยมมากบน YouTube แต่ไม่ได้มียอดดาวน์โหลดสูงบน Podbean แสดงให้เห็นว่ากลุ่มผู้ฟังอาจมีความสนใจที่แตกต่างกัน หรือคอนเทนต์บางอย่างอาจได้รับความสนใจเฉพาะในกลุ่มผู้ฟังบางกลุ่มตามแต่ละ platform

นำไปสู่คำถามถัดไปว่า กลุ่มผู้ฟังนิยมฟังคอนเทนต์ประเภทไหนเป็นพิเศษ และในแต่ละ platform กลุ่มคนฟังมีความสนใจที่แตกต่างกันหรือไม่ โดยข้อมูลที่เราจะนำมาใช้ในการวิเคราะห์คือ ข้อมูลประเภทของธุรกิจที่ถูกจัดหมวดหมู่ตาม playlist



กราฟแสดงยอดวิว (YouTube) เฉลี่ยรายหมวดหมู่ธุรกิจ โดยพิจารณาเฉพาะหมวดหมู่ที่มี ตั้งแต่ 5 วิดีโอขึ้นไป



กราฟแสดงยอดดาวน์โหลด (Podbean) เฉลี่ยรายหมวดหมู่ธุรกิจ โดยพิจารณาเฉพาะหมวดหมู่ที่มี ตั้งแต่ 5 วิดีโอขึ้นไป

จาก 2 กราฟข้างต้นจะเห็นได้ว่าพฤติกรรมการเลือกฟังคอนเทนต์แต่ละ platform นั้นค่อนข้างแตกต่างกัน ตัวอย่างเช่นในหมวด SMEs ที่ได้รับความสนใจค่อนข้างต่ำบน YouTube แต่ได้รับความนิยมเป็นอันดับสองใน Podbean

Data Preparation (Solutions)

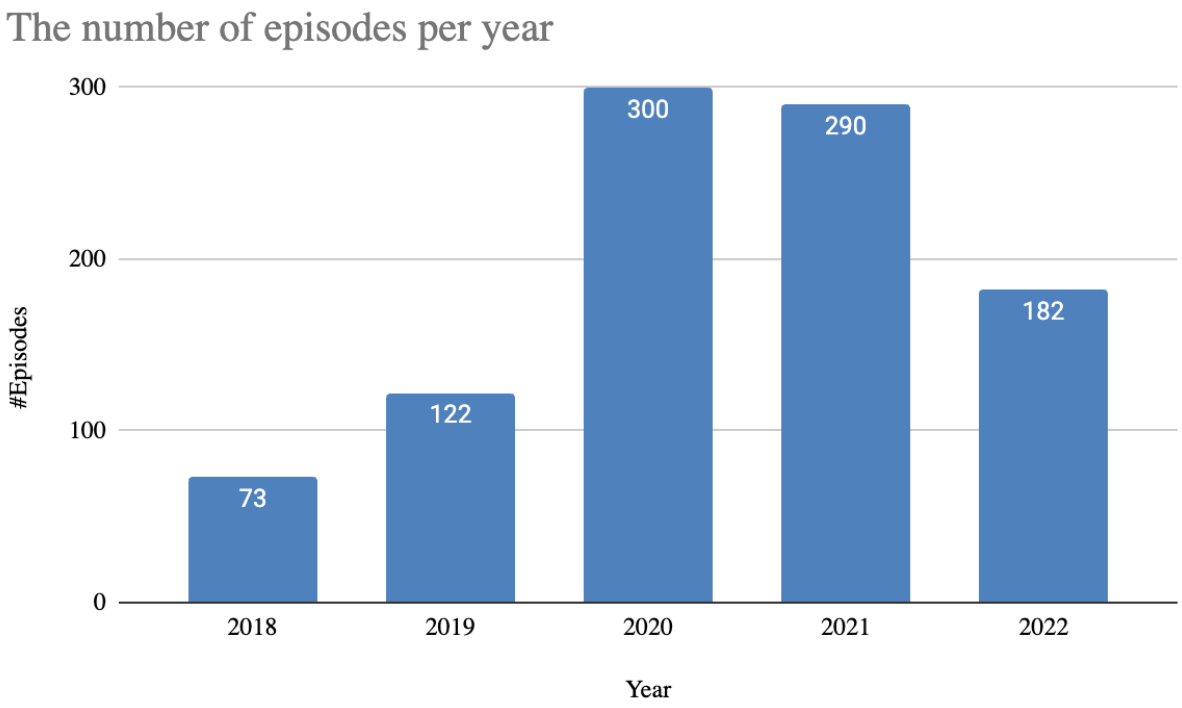
ชุดข้อมูลที่จะใช้ในการทำ Hands-on Activity #2: EDA วันนี้สามารถดาวน์โหลดได้ที่

- Google Sheets
  https://to.skooldio.com/dab5-data-prep-solution-google-sheets

- Excel
  https://to.skooldio.com/dab5-data-prep-solution-excel

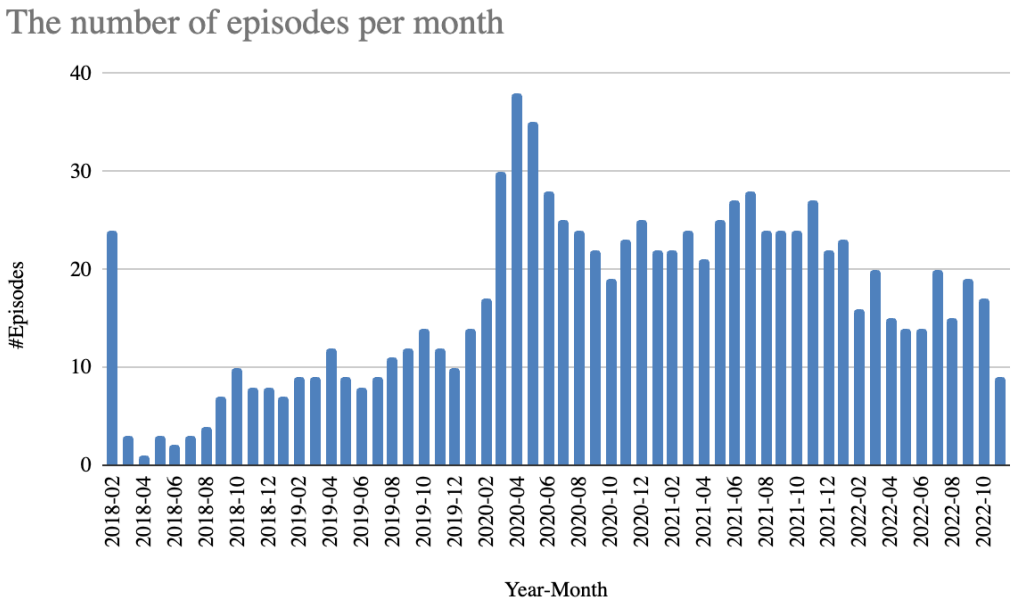
Instructions

1. สร้าง bar chart แสดงจำนวน episode ที่ผลิตรายปี เพื่อดูภาพรวมและแนวโน้มของการผลิตรายการ โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★)
    1. id
    2. year

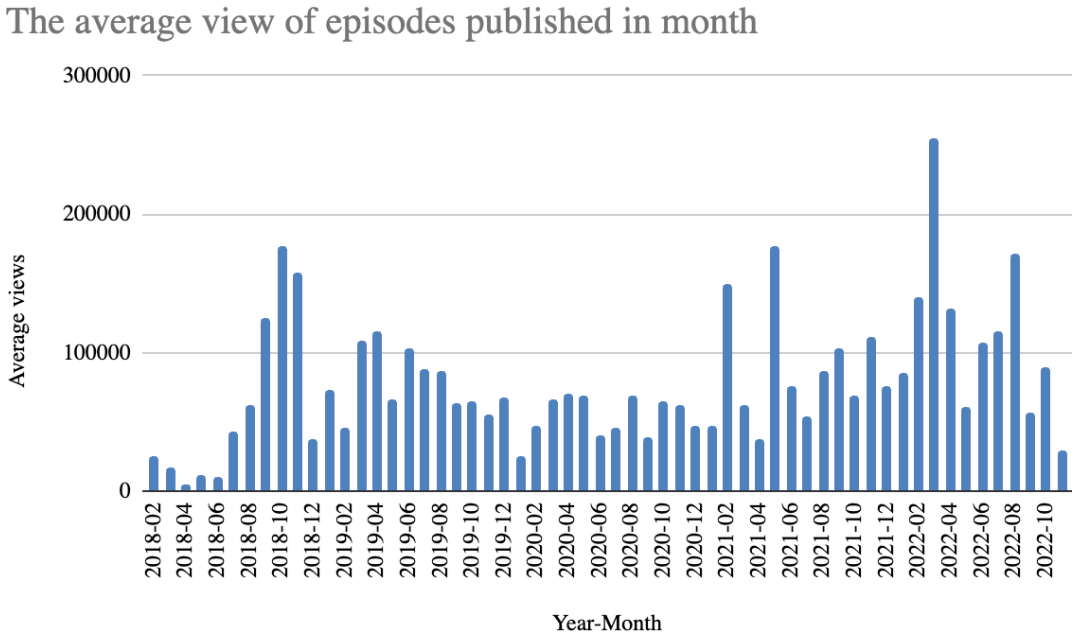


2. สร้าง bar chart แสดงจำนวน episode ที่ผลิตรายเดือน เพื่อดูแนวโน้มการผลิตรายการในระดับที่ละเอียดมากยิ่งขึ้น โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. id
    2. yearMonthStr เช่น ‘2021-02’ (ต้องสร้างขึ้นมาใหม่ hint: ลองดูฟังก์ชัน TEXT)

    💡 **หมายเหตุ**: แม้ว่าเราจะมีคอลัมน์เดือนที่สร้างเตรียมไว้แล้ว แต่หากเราใช้ค่าเดือนเพียงอย่างเดียวในการสรุปข้อมูล เราจะได้จำนวนรายการที่ออกอากาศในเดือนนั้น ๆ ของทุกปีรวมกัน เราจึงจำเป็นต้องสร้างคอลัมน์ yearMonthStr ขึ้นมา



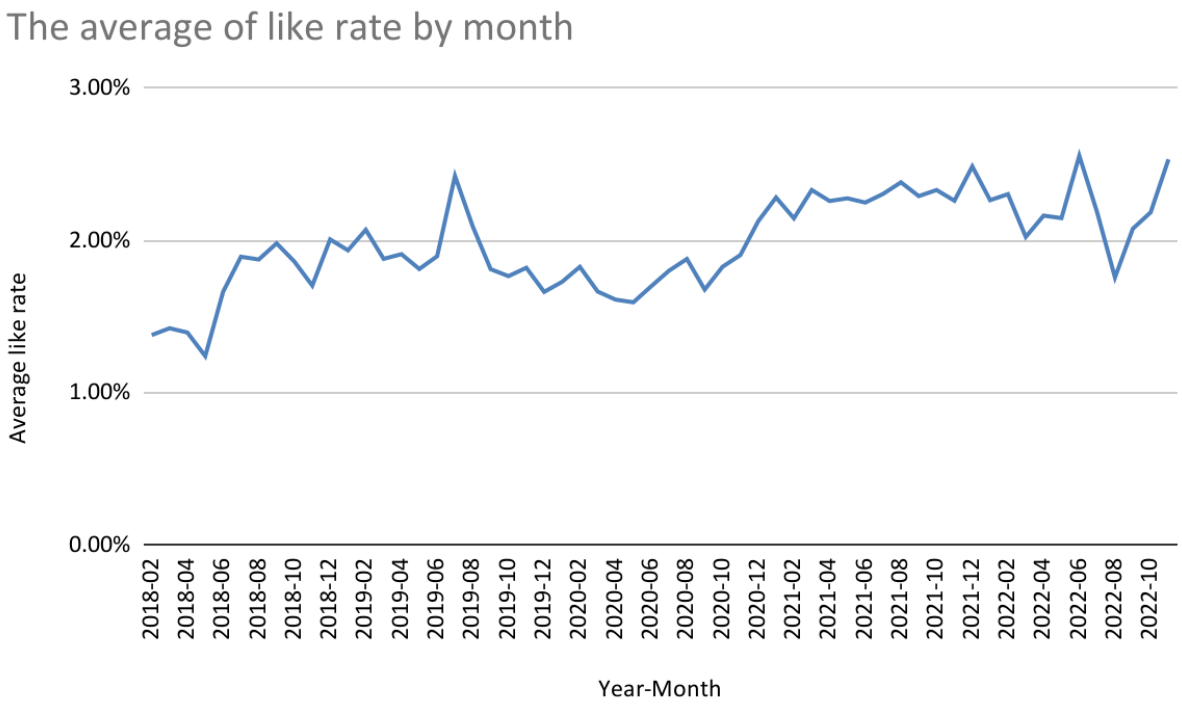
3. สร้าง bar chart แสดงยอดวิวเฉลี่ยของ episode ที่ปล่อยในแต่ละเดือน เพื่อพิจารณาว่ามีแนวโน้มปรับตัวสูงขึ้นหรือไม่ โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. view
    2. yearMonthStr เช่น ‘2021-02’ (ต้องสร้างขึ้นมาใหม่ Hint: ลองดูฟังก์ชัน TEXT)



4. จากข้อที่ 3 จะเห็นได้ว่าจำนวนการรับชมเฉลี่ยของเดือนมีนาคม 2022 มียอดสูงกว่าเดือนอื่น ๆ ซึ่งสาเหตุหนึ่งอาจเป็นเพราะว่าบาง episode ในเดือนนั้น ๆ ได้รับความนิยมสูงทำให้ค่าเฉลี่ยในเดือนนั้นสูงขึ้นผิดปกติ ลองสร้างตารางเพื่อวิเคราะห์การกระจายตัวของยอดวิวในเดือนมีนาคม 2022 จากข้อมูล id view และ yearMonthStr โดยคำนวณค่าสถิติ ดังต่อไปนี้ (★★)
    1. ค่าเฉลี่ย หรือ Mean
    2. ค่ามัธยฐาน หรือ Median
    3. ค่าเปอร์เซ็นไทล์ที่ 25
    4. ค่าเปอร์เซ็นไทล์ที่ 75
    5. ค่าต่ำสุด หรือ Min
    6. ค่าสูงสุด หรือ Max

2022-03	
AVERAGE	254,503
MEDIAN	106,492
PERCENTILE 25	55,691
PERCENTILE 75	277,441
MIN	11,822
MAX	1,343,306

5. สร้าง line chart แสดงแนวโน้ม likeRate เฉลี่ยของแต่ละ episode ตามเวลาที่เปลี่ยนไปเพื่อวิเคราะห์ว่ารายการมีทิศทางได้ดีขึ้นหรือไม่ โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. likeRate ต้องสร้างขึ้นมาใหม่โดยคำนวณจากยอดไลก์ (like) หารด้วยยอดวิว (view)
    2. yearMonthStr



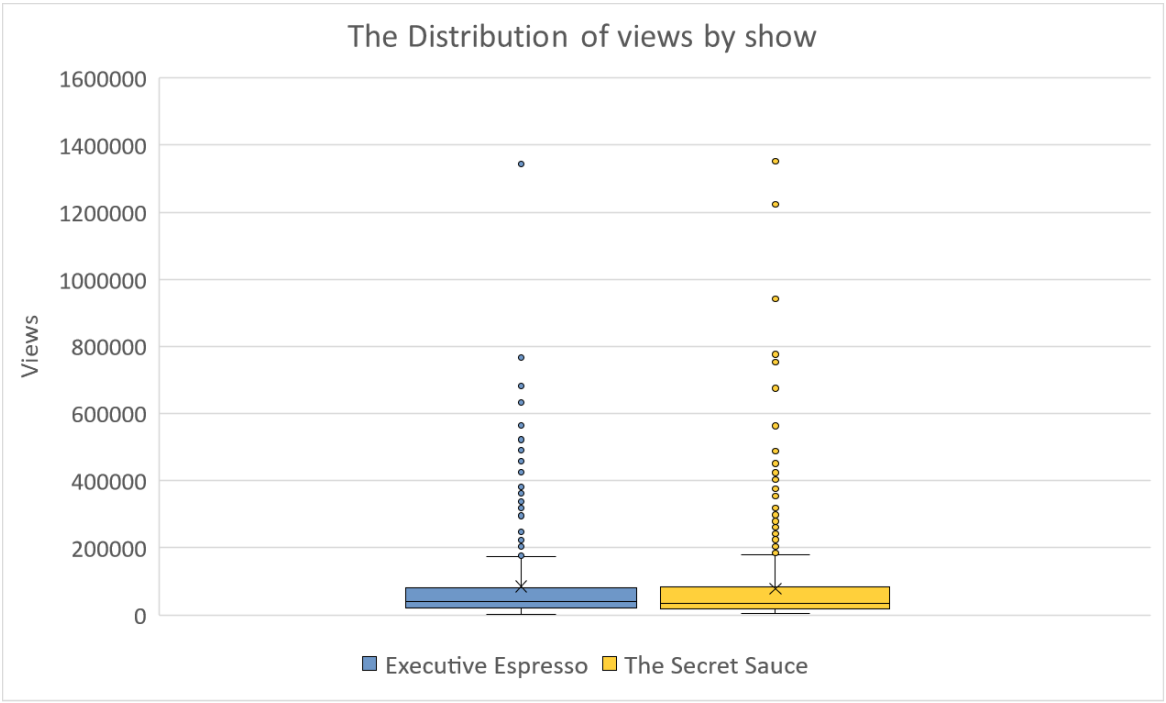
6. สร้างตารางเปรียบเทียบการกระจายตัวของยอดวิวระหว่างรายการ The Secret Sauce และ Executive Espresso โดยใช้ pivot table จากข้อมูล id view และ showTiltle ซึ่งมีรายละเอียดการคำนวณดังต่อไปนี้ (★★★)

    1. ค่าเฉลี่ย หรือ Mean
    2. ค่ามัธยฐาน หรือ Median
    3. ค่าเปอร์เซ็นไทล์ที่ 25
    4. ค่าเปอร์เซ็นไทล์ที่ 75
    5. ค่าต่ำสุด หรือ Min
    6. ค่าสูงสุด หรือ Max

	Executive Espresso	The Secret Sauce
AVERAGE	85,496	78,229
MEDIAN	40,375	35,190
PERCENTILE 25	19,980	16,759
PERCENTILE 75	81,795	83,227
MIN	2,244	3,164
MAX	1,343,306	1,351,575

💡 **หมายเหตุ:**

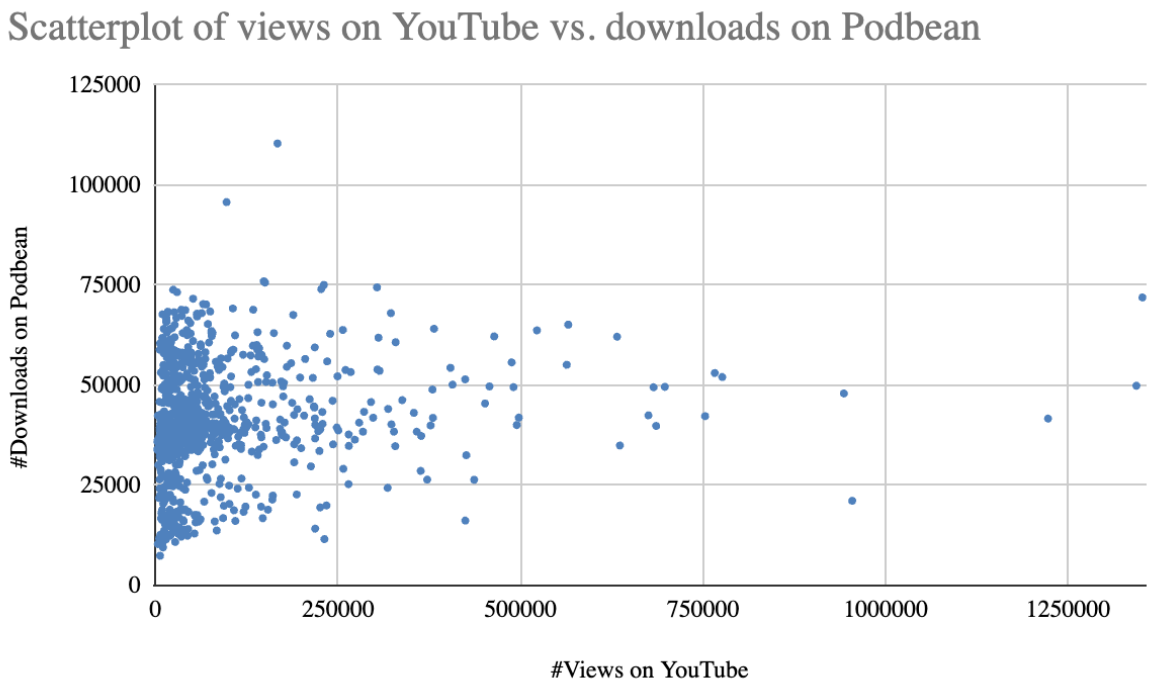
1. สำหรับ Excel อาจลองลากคลุมข้อมูลเพื่อสร้าง Boxplots ซึ่งเป็นกราฟรูปแบบหนึ่งที่ช่วยเปรียบเทียบการกระจายตัวของข้อมูลแต่ละกลุ่มได้ชัดเจน (ปัจจุบัน Google Sheets ยังไม่รองรับการสร้าง Boxplots)
2. สำหรับ Excel จะไม่สามารถสร้าง Boxplot จาก pivot table ได้จะต้อง copy ตารางออกมาวางอีกทีก่อน plot กราฟ



7. สร้าง scatter plot แสดงความสัมพันธ์ระหว่างยอดวิวบน YouTube และยอดดาวน์โหลดบน Podbean โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. id
    2. view
    3. download

    💡 **หมายเหตุ:**

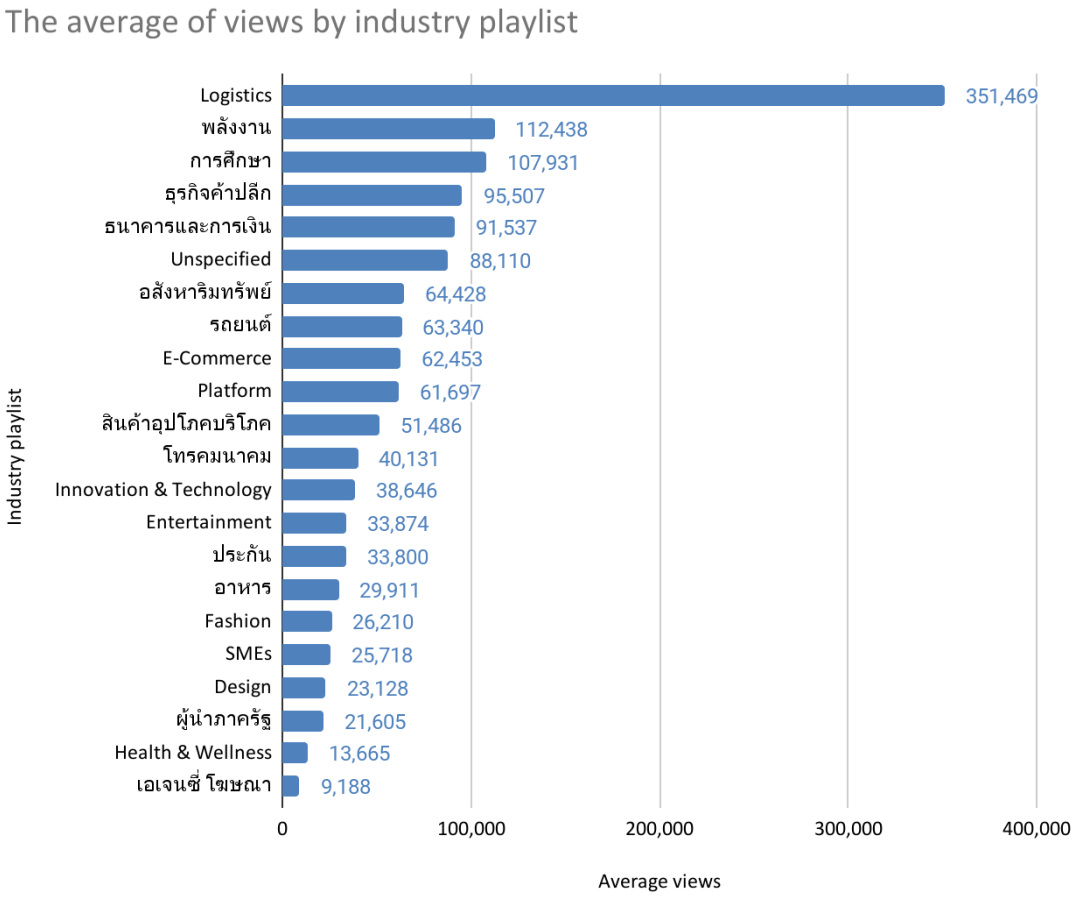
    1. สำหรับข้อมูลยอดดาวน์โหลดบน Podbean นั้นบาง episode ที่เป็น unspecified ลองใช้ Filter ใน pivot table เพื่อกรองข้อมูลเหล่านั้นออก
    2. สำหรับ Excel จะไม่สามารถสร้าง scatter plot จาก pivot table ได้จะต้อง copy ตารางออกมาวางอีกทีก่อน plot กราฟ



8. สร้าง bar chart แสดงยอดวิวเฉลี่ย แยกรายหมวดหมู่ธุรกิจ และพิจารณา **เฉพาะ playlist ที่มีจำนวนตอนรวมกันมากกว่า 5 ตอนขึ้นไป** เท่านั้น โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. industry
    2. view
    3. id

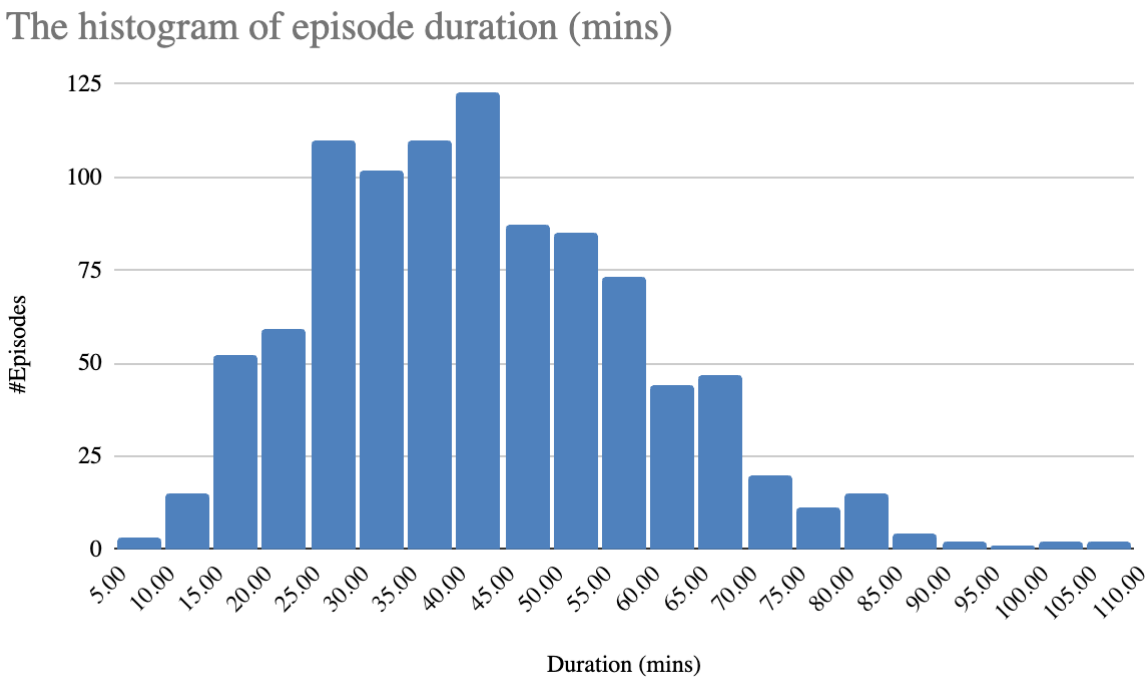
💡 **หมายเหตุ**: สำหรับ Google Sheets แนะนำให้ copy ตารางจาก pivot table ออกมาวางอีกทีก่อนทำการ filter ข้อมูล

💡 **หมายเหตุ**: สำหรับ Excel สามารถทำการ click ที่ filter จากนั้นกด More Sort Options และ Value Filters ได้เลย

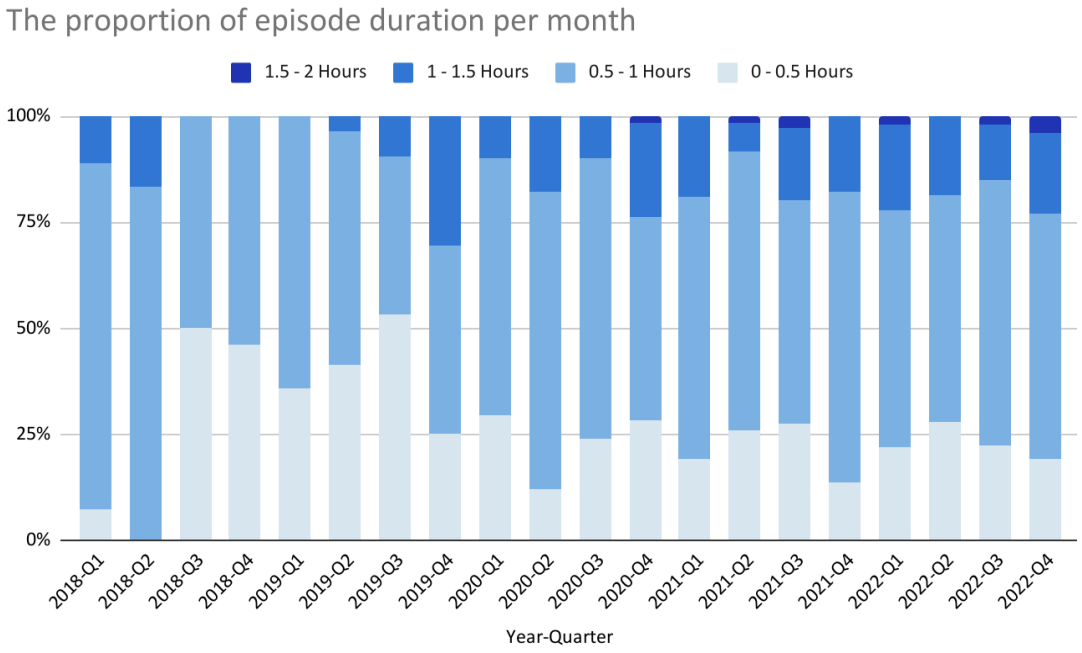


9. สร้าง histogram แสดงจำนวน video ตามความยาวของแต่ละคลิป (duration) ในหน่วยนาที เพื่อวิเคราะห์การกระจายตัวของความยาว video โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★)
    1. id
    2. duration_mins (ต้องสร้างขึ้นมาใหม่ Hint: ลองคำนวณจากคอลัมน์ secs mins hours)

💡 **หมายเหตุ**: สำหรับ Excel จะไม่สามารถสร้าง histogram จาก pivot table ได้จะต้อง copy ตารางออกมาวางอีกทีก่อน plot กราฟ

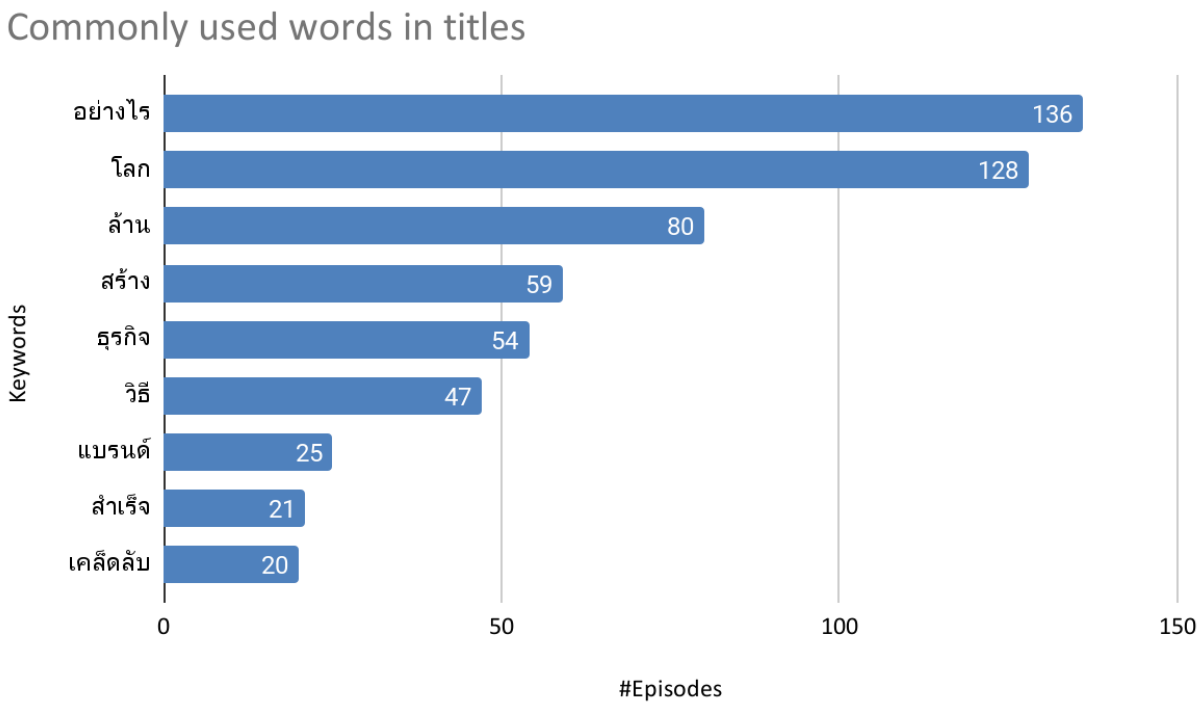


10. สร้าง stacked bar chart แสดงสัดส่วนจำนวน episode ตามช่วงความยาวของคลิปราย quarter (duration range) โดยแบ่งเป็น 4 ประเภทได้แก่ 0 - 0.5 hours, 0.5 - 1 hours, 1 - 1.5 hours และ 1.5 - 2 hours โดยใช้ pivot table จากข้อมูลดังต่อไปนี้ (★★★)
    1. id
    2. yearQtrStr (ต้องสร้างขึ้นมาใหม่ Hint: สามารถทำได้หลายวิธี ลองดูฟังก์ชัน ROUNDUP หรือ IF)
    3. duration_range (ต้องสร้างขึ้นมาใหม่ Hint: ลองดูฟังก์ชัน IFS)



11. สร้าง bar chart แสดงจำนวน keyword ที่ถูกใช้บ่อยในการตั้งชื่อ episode จากข้อมูล title โดยมีรายละเอียด keyword ดังต่อไปนี้
    1. วิธี
    2. สร้าง
    3. อย่างไร
    4. โลก
    5. ล้าน
    6. แบรนด์
    7. สำเร็จ
    8. เคล็ดลับ
    9. ธุรกิจ

💡 **หมายเหตุ**: ในการเลือก keyword ที่จะวิเคราะห์ เราควรเริ่มต้นจากการตัดคำและนำไปนับความถี่ที่เกิดขึ้นโดยใช้ tool อื่นก่อนจึงค่อยกรองคำที่น่าสนใจมาใช้ในการวิเคราะห์ แต่เพื่อความสะดวกในการทำ hands-on activity ครั้งนี้ทีมงานจึงได้เลือก keyword ไว้ให้แล้ว



Next Steps

ลองแลกเปลี่ยนความคิดเห็นกันในกลุ่มว่า จากการวิเคราะห์ เราสามารถเสนอแนะ action ที่เป็นประโยชน์ต่อธุรกิจได้บ้าง เพื่อให้รายการประสบความสำเร็จมากขึ้นกว่าเดิม

ในกิจกรรมกลุ่มครั้งถัดไป เราจะมาสร้าง Reports / Dashboard บน Business Intelligence tools เพื่อใช้ในการนำเสนอผลการวิเคราะห์ให้กับทีมกัน

Group Answer Sheets

Section A:

- Group 1A Answer Sheet
- Group 2A Answer Sheet
- Group 3A Answer Sheet
- Group 4A Answer Sheet
- Group 5A Answer Sheet
- Group 6A Answer Sheet
- Group 7A Answer Sheet
- Group 8A Answer Sheet

Section B:

- Group 1B Answer Sheet
- Group 2B Answer Sheet
- Group 3B Answer Sheet
- Group 4B Answer Sheet
- Group 5B Answer Sheet
- Group 6B Answer Sheet
- Group 7B Answer Sheet
- Group 8B Answer Sheet