Nattapong Tantrakul

6420422025

DADS5001

Journey

1. Data Set มีทั้งหมด 3 set คือรายการส่งออกปี 2019, 2020 และ 2021 ซึ่งจะต้องเอาทั้ง 3 ไฟล์มา Merge กัน เพื่อนำไปใช้งานต่อ แต่เมื่อ Merge กันครั้งแรกเกิด Error เพราะ Column name ของแต่ละไฟล์ไม่เหมือนกัน จริงต้องมีการปรับ Column name ใน Dataframe ให้เหมือนกันก่อน
2. หลังจากที่ Merge กันแล้ว จะต้องดูว่าในแต่ละเดือน มีการส่งออกไปที่ประทศไหนบ้าง และเป็นสินค้าประเภทไหน ปรากฏว่า ข้อมูลเดือน และปีการส่งออกเป็นประเภท Int แบบแยกคนละ column กัน จึงต้องแปลงข้อมูลใน Dataframe ใน column เดือน และปี ให้เป็น str และนำมา concat กัน และใช้ pd.to_datetime แปลงจาก str format เป็น datetime format และนำไปใส่ใน column ใหม่ใน Dataframe
3. เมื่อจัดการเรื่อง datetime format เรียบร้อย จะทำการรวมยอดการส่งออกแต่ละเดือนเพื่อดูว่า trend การส่งออกเป็นอย่างไร จึงต้องทำการสร้าง Dataframe ขึ้นมาใหม่ โดย column แรกจะเป็น เดือน/ปี ซึ่งจะเอา Datetime จากข้อ 2 มาทำให้เป็น set เพื่อตัดตัวซ้ำ และทำให้เป็น List เดือน และ Sort วันที่ หลังจากนั้นสร้าง List ว่างมาอีก 1 อันเพื่อเก็บผลรวมการส่งออกของแต่ละเดือน
4. ต่อมา จะหาว่ายอดส่งออกสินค้าประเภทไหนมากที่สุด โดยจะเทียบจากยอดส่งออก Jan-19 กับ Dec-21 โดยจะหายอดส่งออกที่เกิดขึ้นใน Jan-19 เพื่อเป็นจุดตั้งต้นก่อน และหายอดส่งออกรวมจาก Jan-19 – Dec-21 เพื่อเป็นตัวเทียบ ซึ่งจะต้องสร้าง Dataframe ใหม่ โดยการสร้างจะวิธีคล้ายๆกับข้อ 2
5. เมื่อทำ Dataframe ในข้อ 4 เรียบร้อย เมื่อนำมา sort ก็จะสามารถหาสินค่าที่ยอดส่งออกโตขึ้น แต่เมื่อหา %Growth จะมี Dataframe บางตัวที่มี %growth เป็น Infinity เพราะรายการสินค้ารายการนั้นยังไม่เคยส่งออกมาก่อนใน Jan-19 จึงทำให้ตัวหารเมื่อจะเป็น % เป็น 0 ซึ่งการที่เป็นแบบนี้ เมื่อ sort highest value จะทำให้รายการ Infinity ขึ้นมาอยู่หัวตาราง ไม่สามารถนำไปใช้งานต่อได้ จึงได้ทำการแก้ไข โดยราการดังกล่างส่วนใหญ่จะเป็นรายการที่มียอดส่งออกน้อย วิธีการแก้คือ เพิ่มเงื่อนไขการ filter เข้าไป โดย Data ที่จะนำมาคำนวน จะต้องการยอดขายใน Jan-19 มากกว่า 1 แสนบาท และยอดขายเมื่อจบ Dec-21 จะต้องทีมากกว่า 1 ล้านบาท หลังจากแก้ก็จะ Sort value ได้ และได้รายชื่อสินค้าที่มียอดส่งออกมากที่สุดออกมา และเนื่องจาก data มีปริมาณมาก จึงคัดมาเป็น Top 5 เท่านั้น ทั้งสินค้าที่ Top 5 ในส่วนยอดขาย และ %Growth
6. เมื่อมีข้อมูลอยู่ประมาณนึง จึงได้ทดลอง Plot graph ในครั้งแรก นำแค่ยอดส่งออก และ %Growth มา plot ตรงๆ และ Graph ที่ได้ จะ trend ค่อนข้างยาก จึงได้ทำการปรับ Dataframe ใหม่ โดยเพิ่ม column cumulative revenue และ cumulative %growth และนำข้อมูลใหม่มา Plot graph
7. จากข้อ 6 ได้ทดลอง Plot graph แบบ sub plot เพื่อเทียบ growth ทั้งสินค้าที่ growth ด้าน revenue และ สินค้าที่ growth ด้าน %growth แต่ประกฏว่า sub plot จะให้ดูข้อมูลยาก จึงเปลี่ยนมาเป็น Plot line chart ธรรมกา แต่นำ Data 3 ส่วนมา Plot รวมใน graph เดียว เพื่อให้เทียบกันได้ชัดๆ
8. หลังจากเทียบ %growth เรียบร้อย ต่อไปจะหาดูว่าสินค้าแต่ละประเภท ส่งออกไปประทศไหนบ้าง และประเทศใดที่ Thailand export ออกไปทางที่สุด ครั้งแรกจะเป็น Heatmap จึงทำการแปลง Dataframe มาใหม่ โดยการใช้ Pivot แต่เมื่อทดลอง plot ออกมา ปรากฏว่า ข้อมูลทั้งประเทศ และรายการสินค้ามีจำนวนที่เยอะมากๆ ทำให้ Heatmap visualize ออกมาได้ไม่เหมาะ (ตารางถี่มากๆ และ data label ซ้อนทับกัน) จึงได้เปลี่ยนมาดูข้อมูลเพราะรายการที่มี %growth สูงสุดแทน โดยการให้ Market cap ว่ารายการที่ว่า ส่งออกไปประเทศไหนเยอะที่สุด