Journey

วิธีการทำงาน, อุปสรรคและการแก้ปัญหาที่เจอในการทำโปรเจค เรียงตามเวลา

1. หา dataset ที่ต้องการในการทำ analysis

   เริ่มต้นจาก dataset ที่ใช้เป็น dataset หลักของโปรเจคในครั้งนี้คือ dataset 'opendata_project.csv' จาก Baania ซึ่งเป็นข้อมูลรายละเอียดของอสังหาริมทรัพย์ในประเทศไทย ซึ่งมีรายละเอียดของข้อมูลค่อนข้างมากและคิดว่าน่าจะเป็นข้อมูลที่มีประโยชน์สำหรับการตัดสินใจซื้ออสังหาริมทรัพย์ได้ (อีก dataset ที่ใช้จะมาหลังจากเจอสิ่งที่อยากรู้ระหว่างทำการ analyze จึงไปรวมอยู่ในหัวข้อ 2)

2. Cleansing data
   - ขั้นตอนแรกหลังจาก import ข้อมูล ผมเริ่มจากการตรวจสอบ data type ของแต่ละ column ว่าถูกต้องตามต้องการหรือไม่ โดยหลังจากนั้นได้ทำการ cast column ที่เป็นข้อมูล time stamp ให้เป็น type datetime ('date_created', 'date_finish', 'date_update') เพื่อที่จะนำไปสร้าง column ใหม่ไว้เก็บปีแยกกันอีกที ('year_create', 'year_finish') และเปลี่ยน type ของ column 'price_min' ให้เป็น numeric เพื่อจะได้นำไปคำนวนได้ในส่วนต่อไป
   - Drop column ที่ไม่ใช้ในการ analyze ออก ได้แก่ 'row_number', 'source', 'url_project'
   - Drop row ที่มีค่า Na ใน column ที่ต้องการนำมา analyze ได้แก่ 'neighborhood_name_th', 'subdistrict_id', 'date_finish'
   - Drop row ที่มี 'propertytype_name_th' ที่ไม่ได้ใช้ทิ้ง ได้แก่ 'โกดัง / โรงงาน', 'ที่ดิน', 'สำนักงาน', 'โรงแรม', 'อพาตเมนต์' เนื่องจากต้องการโฟกัสไปที่อสังหาริมทรัพย์ที่เป็นที่อยู่อาศัย
   - Fill Na value ใน column ที่สามารถเติมได้ 'price_min' เติม 0, 'developer_name_en' เติม '-' และ 'zipcode' เติม 0 ส่วน column ที่เหลือทั้งหมดจะเป็นพวก facility ซึ่งเป็น boolean ว่ามีหรือไม่มี จึงเติมด้วย 0 ทั้งหมดในช่องที่เป็น Na
   - รวม type 'บ้านแฝด' เข้าเป็น type เดียวกับ 'บ้าน'
3. Data analysis
   - เริ่มจาก group by 'propertytype_name_th' เพื่อดูว่าข้อมูลที่จะนำมาใช้มีอสังหาริมทรัพย์ทั้งหมดกี่แบบ
   - จากนั้นสร้าง bar plot เพื่อดูภาพรวมของราคาและปริมาณของอสังหาริมทรัพย์ที่สร้างขึ้นในปี 2018 – 2022
   - จากนั้นต้องการหาราคาเฉลี่ยของคอนโดแยกตามแต่ละเขตใน กทม. เพื่อใช้เลือกเขตที่ช่วงราคาเหมาะสมกับเงินทุนที่จะซื้อ จึงสร้าง bar plot ขึ้นมาเพื่อแสดงราคาเฉลี่ยของคอนโดในแต่ละเขตโดยผมเลือกที่จะแสดง 10 เขตที่มีราคาเฉลี่ยสูงที่สุด

- หลังจากนั้นผมคิดว่าถ้าจะเลือกพื้นที่ที่จะปล่อยเช่าได้ดี ควรจะดูความหนาแน่นของประชากรในแต่ละเขตด้วยในการประกอบการตัดสินใจ จึงไปหาข้อมูลประชากรใน กทม มาโดยแยกตามเขต แต่ dataset ที่หาได้เป็นไฟล์ pdf จึงนำข้อมูลแค่ column ที่ต้องการใช้ได้แก่ จำนวนประชากร และขนาดพื้นที่ของแต่ละเขตมาสร้างไฟล์ csv เองเพื่อนำมา import เข้า pandas
- สร้าง colum density โดยนำจำนวนคน / พื้นที่ของเขต
- นำข้อมูลจาก dataset opendata_project.csv มา join กับ density เพื่อดู 10 เขตที่เราเลือกไว้ว่าเขตไหนมีความหนาแน่นของประชากรที่น่าสนใจ นำมา plot bar plot
- หลังจากดูกราฟราคาและความหนาแน่นของประชากรแล้วผมตัดสินใจว่าผมจะเลือกซื้อคอนโดในเขตคลองสาน จึงนำข้อมูลคอนโดในเขตคลองสานมาช่วยตัดสินใจต่อ
- เริ่มจากลิสคอนโดทั้งหมดที่สร้างปี 2018 – 2022 ในเขตคลองสานและนำ dataset ที่มีส่วนในการตัดสินใจเลือกคอนโดมา 2 เรื่องคือ เรื่อง facility ในคอนโด และ เรื่องจำนวน unit ในคอนโด จึงนำข้อมูลข้างต้นมา plot bar plot แต่การเปรียบเทียบข้อมูล 2 factor โดยแยกกราฟกันค่อนข้างดูยาก จึงนำมา plot แบบ 2 แกน เพื่อให้ดูง่ายขึ้นและเปรียบเทียบได้ดีขึ้น