### Attrition\_Flag

1. ข้อมูล Attrition\_Flag มีความ imbalance โดยที่ Existing Customer มากกว่า Attried Customer จำนวนมาก
2. ข้อมูล Total\_Trans\_Amt, Total\_Trans\_Ct และ Total\_Ct\_Chng\_Q4\_Q1 ดูมีความสัมพันธุ์กับ output เนื่องจากสังเกตได้ถึงการแยกกลุ่มของข้อมูลเหล่านี้

### Customer\_Age

1. ข้อมูลมีลํกษณะเป็น normal distribution โดยลูกค้าส่วนมากมีอายุช่วง 40 - 50 ปี มีความแปรปรวนเท่ากับ 64 โดยที่ช่วงอายุที่มากกว่า 66 อาจพิจารณาเป็น Outlie
2. ข้อมูล Customer\_Age มีความสัมพันธุ์กับฟีเจอร์ month on book แต่ไม่เห็นถึงความสัมพันธุ์ต่อ output จึงอาจพิจารณาตัดข้อมูลใดข้อมูลหนึ่งออก

### Gender

1. จำนวนประชากรเพศชายกับหญิงมีจำนวนห่างกันเพียง 589 คนและไม่มีความสัมพันธุ์ใดๆกับ Output
2. เพศหญิงมี credit limit และ Avg\_open\_to\_buy ต่ำกว่าเพศชายอย่างเห็นได้ชัด
3. ในข้อมูลของ income กลุ่มที่มีรายได้ต่ำกว่า 40000 และรายได้ที่ไม่ทราบส่วนใหญ่เป็นเพศหญิง
4. กลุ่มรายได้ 60k-80k, 80k-120k, 120k+ ส่วนใหญ่เป็นเพศชาย
5. ข้อมูล Total\_Trans\_Amt ในเพศหญิงมีการแบ่งกลุ่มโดย Existing Customer มี Total\_Trans\_Amt อยู่ในช่วง 2500 – 500 แต่ Attrited Customer อยู่ในช่วง 2500 +- เล็กน้อยแต่ในเพศชายไม่เห็นการแบ่งกลุ่มโดยมี mean ที่ใกล้เคียงกัน
6. ข้อมูล Total\_Trans\_Ct ในเพศหญิงมีการแบ่งกลุ่มโดย Existing Customer มี Total\_Trans\_Ct อยู่ในช่วง 60-80 แต่ Attrited Customer อยู่ในช่วง 40-50 แต่ในเพศชายไม่เห็นการแบ่งกลุ่มอย่างชัดเจนแต่สังเกตุเห็นได้ว่ากลุ่ม Attrited Customer จะมีช่วง Totl\_Trans\_Ct ต่ำกว่ากลุ่ม Existing Customer

### Dependent\_count

1. ลูกค้าส่วนใหญ่มีบุตร 2-3 คนแต่จำนวนบุตรดูไม่มีความสัมพันธุ์ใดๆกับ Output

### Education\_Level

1. ลูกค้าส่วนมากเป็นลูกค้าที่เรียนจบแล้ว
2. จากการวิเคราะฟีเจอร์ Dependent count ซึ่งพบว่า total\_trans\_ct เมื่อ plot กับ Education Level แล้วพบว่าเกิดการแบ่งกลุ่มของการสูญเสียลูกค้าได้อย่างชัดเจนโดยที่ช่วงของ Total\_Trans\_Ct ในกลุ่ม Existing Customer ในทุกอาชีพมีแนวโน้มที่จะสูงกว่ากลุ่ม Attrited Customer โดยข้อมูลส่วนใหญ่ไม่มีการซ้อนทับกัน

### Marital\_Status

1. ข้อมูลไม่มีความสัมพันธุ์ใดๆกับ output ไม่ว่าจะเปรียบเทียบกับข้อมูลใดๆก็ตาม

### Income\_Category

1. ลูกค้าส่วนใหญ่มีรายได้ต่ำกว่า 40000
2. คนที่มีรายได้มากกว่า 120k มี credit limit และ avg\_open\_to\_buy สูงกว่ากลุ่มอื่นๆ

### Card\_Category

1. ผู้ใช้งานส่วนมากเป็นผู้ใช้งานระดับ Blue โดยผู้ใช้งานระดับอื่นมีข้อมูลน้อยมาก
2. ผู้ใช้งานระดับ Blue มี avg\_open\_to\_buy และ credit limit ต่ำกว่ากลุ่มอื่นๆแต่เนื่องจากข้อมูลที่ไม่ใช่ระดับ Blue น้อยมากจึงอาจพิจารณาที่จะตัดข้อมูลนี้ทิ้ง

### Months\_on\_book

1. ข้อมูลมีจำนวน Months\_on\_book ที่ค่อนข้างโดดคือมี 36ครั้ง จำนวนถึง 2463 คน
2. เมื่อ plot ค่า Months\_on\_book กับ Total\_Trans\_Amt พบว่าช่วงที่ Total\_Trans\_Amt มากกว่า 12000 เห็นกลุ่มของ Existing Customer ชัดเจน แต่ช่วงที่น้อยกว่า 12000 ยังคงเห็นการปนกันของ Existing Customer และ Attrited Customer แต่การที่เห็นการแบ่งกลุ่มชัดเจนเนื่องมาจากตัวข้อมูลของ Total\_Trans\_Amt ที่มีการแบ่งกลุ่ม output ในตัวของข้อมูลอยู่แล้ว

### Total\_Relationship\_Count

1. ข้อมูลไม่มีความสัมพันธุ์กับ ouput ไม่ว่าจะเป็นการเปรียบเทียบกับข้อมูลใด

### Months\_Inactive\_12\_mon

1. ข้อมูลมีความสัมพันธุ์กับ total revolving balance และ total\_trans\_ct โดยที่ข้อมูลที่เป็น Attrited Customer จะมีช่วงของข้อมูล total revolving balance และ total\_trans\_ct ที่ต่ำกว่า Existing Customer

### Contacts\_Count\_12\_mon

1. contacts\_count\_12\_mon จำนวน 6 ครั้ง พบว่าเป็นลูกค้าที่เสียไปทั้งหมด แต่เป็นเพียงข้อมูลเล็กน้อยของข้อมูลทั้งหมด

### Credit\_Limit

1. จำนวน credit limit มีลักษณะ right skewed จึงทำการ correct skewed data ด้วยวิธีการ Box-cox
2. Credit\_Limit มีความสัมพันธุ์กับ Avg\_Open\_To\_Buy แบบ linear จึงอาจะพิจารณาตัดข้อมูลใดข้อมูลหนึ่งออก
3. ข้อมูลมีความสัมพันธุ์กับ Avg\_Utilization\_Ratio เมื่อ Credit\_Limit เพิ่มมากขึ้น Avg\_Utilization\_Ratio ก็จะต่ำลงอย่างเห็นได้ชัดซึ่งอาจพิจารณาตัดข้อมูลใดข้อมูลหนึ่งออก

### Total\_Revolving\_Bal

1. เมื่อ plot Total\_Revolving\_Bal กับ Avg\_Utilization\_Ratio มองเห็นแนวโน้มที่มีความเป็นกลุ่มของข้อมูลที่ไม่ได้แบ่งแยก Output จึงอาจพิจารณาตัดข้อมูลใดออก

### Avg\_Open\_To\_Buy

1. ข้อมูลมีลักษณะ Right skewed
2. ข้อมูล Avg\_open\_To\_Buy มีความสัมพันธุ์กับ Avg\_Utilzation\_Ratio เมื่อค่า Avg\_open\_To\_Buy มากขึ้น Avg\_Utilzation\_Ratio ลดลงอย่างเห็นได้ชัดจึงอาจพิจารณาตัดข้อมูลใดข้อมูลนึงออก

### Total\_Amt\_Chng\_Q4\_Q1

1. ข้อมูลมีลักษณะใกล้เคียงกับ normal distribution
2. จากกราฟ histrogram จะสังเกตุได้ว่าเมื่อ Total\_Amt\_Chng\_Q4\_Q1 มากกว่า 1 ข้อมูลที่เหลือจะเป็น Existing Customer แต่เป็นเพียงข้อมูลส่วนน้อย

### Total\_Trans\_Amt

1. จากกราฟการกระจายตัวของข้อมูลพบว่าเมื่อ Total\_Trans\_Amt มีค่ามากกว่า 10000 ลูกค้าส่วนใหญ่เป็น Attrited Customer
2. ความสัมพันธุ์ระหว่าง Total\_Trans\_Amt และ Total\_Trans\_Ct มีการแบ่งกลุ่ม Output เนื่องมาจากการกระจายตัวของข้อมูลทั้งคู่ที่ส่งเสริมกันและ Total\_Trans\_Amt ที่มีค่ามากกว่า 10000 ส่วนใหญ่เป็น Attrited Customer
3. ความสัมพันธุ์ระหว่าง Total\_Trans\_Amt และ Total\_Ct\_Chng\_Q4\_Q1 มีการแบ่งกลุ่ม Output เนื่องจากตัว Total\_Trans\_Amt ที่มีค่ามากกว่า 10000 ส่วนใหญ่เป็น Attrited Customer
4. ความสัมพันธุ์ระหว่าง Total\_Trans\_Amt และ Avg\_Utilization\_ratio มีการแบ่งกลุ่ม Output เนื่องจากตัว Total\_Trans\_Amt ที่มีค่ามากกว่า 10000 ส่วนใหญ่เป็น Attrited Customer

### Total\_Trans\_Ct

### Total\_Ct\_Chng\_Q4\_Q1

### Avg\_Utilization\_Ratio

Summarize

1. ตัดข้อมูล Avg\_Utilization\_Ratio เนื่องจากข้อมูลมีความเป็น Right Skewed และมีความสัมพันธุ์กับข้อมูล Credit\_Limit, Total\_Revolving\_Bal และ Avg\_open\_to\_buy ในลักษณะ parabolic และไม่ได้แสดงถึงการแบ่งกลุ่มของ Output
2. ตัดข้อมูล Dependent\_Count, Marital\_Status, Income\_Category, Card\_Category, Months\_on\_book, Total\_Relationship\_Count และ Contact\_Count\_12\_mon เนื่องจากไม่มีความเกี่ยวข้องใดๆกับ Output ไม่ว่าจะเปรียบเทียบกับข้อมูลใดๆก็ตาม
3. ตัดข้อมูล Credit\_Limit และ Avg\_Open\_To\_Buy เนื่องจากมีความสัมพันธุ์กันแบบ parabolic และทั้งสองข้อมูลมีความเบ้ขวานอกจากนี้ยังไม่มีความสัมพันธุ์ใดๆกับข้อมูล Output
4. ตัดข้อมูล Total\_Revolving\_Bal, Total\_Amt\_Chng\_Q4\_Q1, Customer\_Age และ Gender ออกเนื่องจากไม่เห็นถึงความสัมพันธุ์ที่ชัดเจนกับข้อมูล Output เมื่อเทียบกับข้อมูลอื่นๆ
5. Graphical user interface

   Description automatically generated
6. A picture containing graphical user interface

   Description automatically generated
7. A picture containing graphical user interface

   Description automatically generated
8. A screenshot of a computer

   Description automatically generated with low confidence
9. A picture containing graphical user interface

   Description automatically generated

ลองตัดข้อมูลที่มองดูแล้วรู้สึกว่าไม่น่าจะส่งผลอะไรกับ Output ซึ่งได้แก่

1. Customer\_Age
2. Gender
3. Marital\_Status
4. Card\_Category
5. Months\_on\_book

ผลลัพธ์ที่ได้

A picture containing graphical user interface

Description automatically generated

ลองใช้ gridsearchCv ในการเลือก reduce dimension of dataset และ tune hyper parameter

Diagram

Description automatically generated with medium confidence