Journey

Step Clean Data

1) Install xlrd version 1.2.0 เนื่องจาก file ที่จะต้อง read เป็น file excel ที่มี font ภาษาไทย

2) ลำดับแรกข้อมูลที่ได้มา มี column ที่มีความซับซ้อน ยากในการจะไปใช้วิเคราะห์ ข้อมูลเป็น pivot มาอยู่แล้ว ต้อง split ข้อมูลที่เป็น pivot ออกมา ให้อยู่ในรูปของ tabula data ซึ่งการ split จะ split แบบตรงๆเลยก็ไม่ได้ ต้อง split ที่ละระดับการศึกษา ที่ละคณะ นำไป unpivot ให้อยู่ในรูปของ tabular แล้วนำไปเก็บไว้ แล้วขยับไปทำ ระดับการศึกษาถัดไป คณะถัดไป แบบนี้เรื่อยๆ จนครบ

3) อ่าน ไฟล์ excel ทั้ง 6 ไฟล์ 2558 2559 2560 2561 2562 2563 โดยใช้ เทคนิค regular expression โดยกำหนดให้อ่านชื่อไฟล์ที่เป็นตัวเลข 4 ตัว และลงท้ายด้วย xls

4) Insight ที่อยากได้ อีก 1 อันคือ อยากรู้กระจายตัวของงานตามภาคต่างๆ ในประเทศไทย แต่ข้อมูลที่ได้มาตอนแรกจะมีให้แค่จังหวัด จึงต้องไปหา file ข้อมูลอีกอันมา merge (เป็น file ที่บอกว่าในแต่ละภาคมีจังหวัดอะไรบ้าง)

5) ขั้นตอนที่จะ merge ก็พบว่า ข้อมูลที่อยู่ในไฟล์แรก ไม่มีคำว่า จังหวัด เช่น สมุทรปราการ แต่ข้อมูลไฟล์ภาค มีคำว่าจังหวัด “จังหวัดสมุทรปราการ” จึงต้องเอาคำว่าจังหวัด ในไฟล์แรกออกก่อน

6) เนื่องจาก file ข้อมูลภาคมีหลาย column อีก เพราะฉะนั้นก็เลือกเอา 2 column ไปใช้ แค่ coloumn ภาคและชื่อจังหวัด หลังจากนั้นก็ค่อย merge

7) ยังพบว่า ยังมีบาง field ใน column region หลังจากที่ merge แล้ว ยังเป็น NaN ที่เป็นเช่นนี้ก็เพราะว่าข้อมูล column จังหวัด ไม่ได้ระบุเป็นจังหวัด แต่ถูกระบุเป็น “ไม่มีที่ตั้งกิจการแน่นอน” จึง clean data ที่เป็น NaN ด้วยการ fillna ด้วย wording ('ไม่มีที่ตั้งแน่นอน')

เป็นอันได้ data ที่พร้อมจะนำไป visualize

Step Visualization

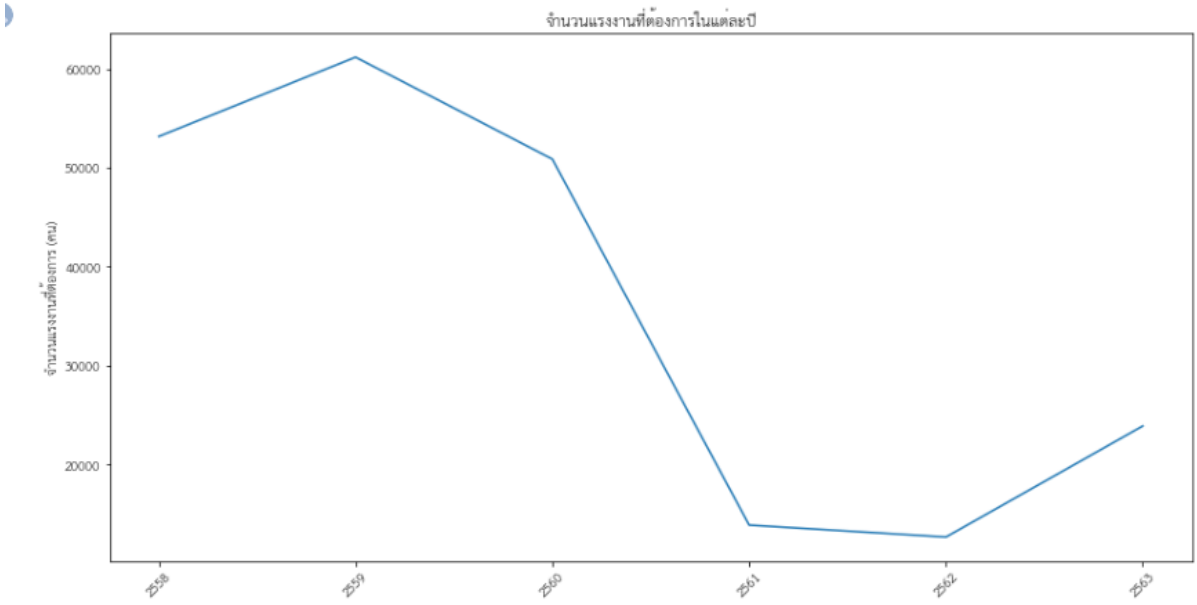
1) ต้องการหา insight คร่าวๆเกี่ยวกับผลิตภัณฑ์ (Industry ของผลิตภัณฑ์) ที่มีความต้องการแรงงานมากที่สุด แต่ชื่อผลิตภัณฑ์ส่วนใหญ่ก็จะเป็นวลีด้วย เช่น Electronics Parts *for* Vehicles, Electric Switches *for* Eco Cars ซึ่งทำการ set stopword เพื่อ remove คำที่เป็น stopword ออกไป แล้วเรียกใช้ `wordcloud` เพื่อดูเล่นๆ ว่ากลุ่ม `Industry` ไหนที่มีคำซ้ำเยอะ ก็ได้ความว่า `electricity power, software digital degital content enterprise software` มีคำซ้ำเยอะสุด แต่ก็ไม่ได้บอกว่าต้องการคนเยอะรึป่าว ฉะนั้นก็ยังเอาไปใช้อะไรไม่ได้

2) **ต้องการใช้** matplotlib ในการ plotgraph แต่การแสดงผลภาษาไทยในกราฟของ matplotlib บน Google Colab ไม่สามารถทำได้ ต้องไป load font จาก github ก่อน

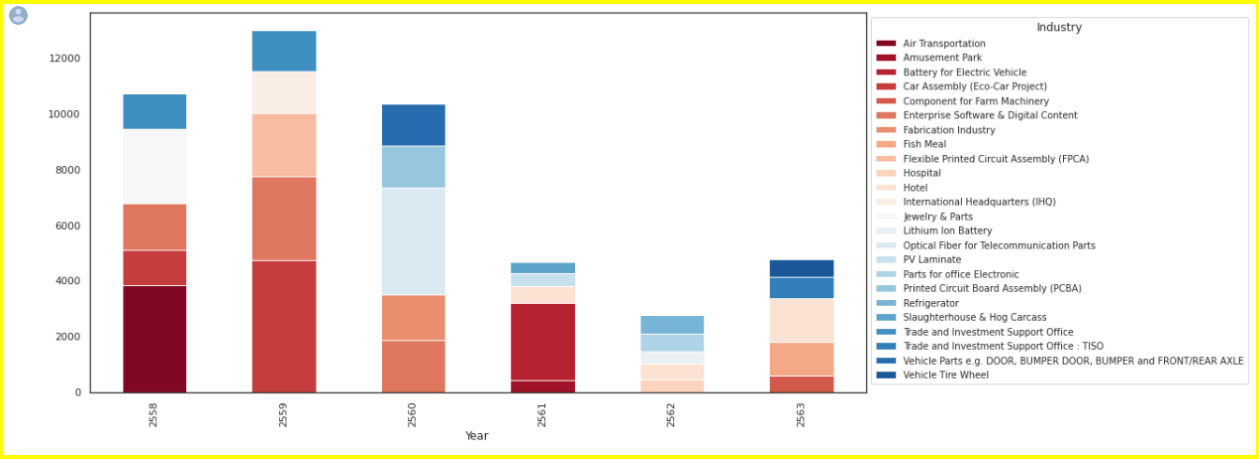
โจทย์ที่ตั้งไว้ตอนแรกเลยที่ตั้งใจจะหาก็คือ

Hyphothesis:
1. ความต้องแรงงานในปัจจุบันมีการชะลอตัวลงจริงหรือไม่
2. ต้องการหาผลิตภัณฑ์ (กลุ่ม Product) ที่มีความต้องการแรงงานมากไปน้อยในช่วงปี 2558-2563
3. กลุ่มจังหวัดที่อยู่ในภาคตะวันออกมีความต้องการแรงงานสูงสุด

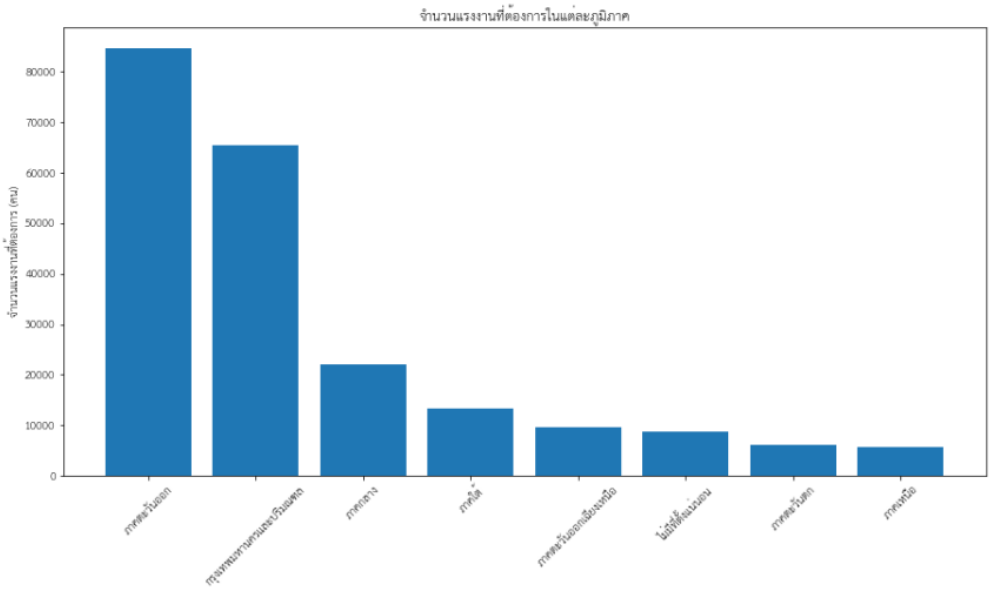
คำถามที่ 1 plot ความต้องการแรงงานทั้งหมดแบบยังไม่สนใจอะไร เปรียบเทียบในแต่ละปี
ตอน plot graph ปกติจะ sort โดยใช้ value แต่คราวนี้ต้องการให้เรียงปี จึงต้องเปลี่ยนเป็น sort โดยใช้ index แทน
results : แนวโน้มความต้องการแรงงานดูลดลง แต่ไม่แน่ใจว่ามาจากปัจจัยใดบ้าง อาจจะเป็นเรื่องของเก็บข้อมูลหรือผลกระทบช่วงโควิดก็เป็นได้



คำถามที่ 2 เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากที่สุด 5 อันดับในแต่ละปี เปรียบเทียบ trend เรียงไปตั้งแต่ปี 58-64 ว่ามีการเปลี่ยนแปลงหรือไม่
ต้อง group by product industry และ group by year ด้วย เพิ่มเติมคือ เอามาแค่ 5 อันดับพอ เพื่อไม่ให้เยอะเกินไป (ค่อนข้างซับซ้อน)
เขียน loop ทีละปี (year)
ใน 1 loop ใช้การ groupby ชื่อผลิตภัณฑ์ (product industry) , sort จากมากไปน้อย, และกำหนด nlargest() = 5 และเอาข้อมูลไปรวมกัน และก่อนจะเอาไปรวมกัน ใน loop ถัดไป แต่เราก็ไม่ไม่รู้อีกว่าข้อมูลเป็นของปีอะไร ทำให้ต้อง add column ปี (year) เข้าไปด้วยเพื่อเป็นการ identify ว่าที่กำลัง loop อยู่เป็นของปีไหน
results : ได้กลุ่ม product industry ที่ไม่ซ้ำกันเลย 25 กลุ่มในแต่ละปี (ดูแล้ว graph วุ่นวายมาก ไม่ได้มี trend อะไร) ซึ่งดูแล้วงงกว่าเดิม ไม่ work ที่จะเอาไปใช้อะไรได้



คำถามที่ 3 กลุ่มจังหวัดที่อยู่ในภาคตะวันออกมีความต้องการแรงงานสูงสุด หรือเปล่า
- เรียงลำดับ ภาคที่ต้องแรงมากงานที่สุด (เรียงลำดับมากไปน้อย)
result : ภาคตะวันออกเยอะสุด, ตามด้วยกรุงเทพ&ปริมณฑล,อันดับสามภาคกลาง ภาคที่เหลือไม่ต่างกันเยอะมาก ซึ่งก็เป็นไปตามที่คาดการณ์ เพราะภาคตะออกวันมีนิคมอุตสาหกรรมเยอะ น่าจะใช้แรงงานจำนวนเยอะ



ระหว่างทำ ก็คิดคำถามขึ้นมาใหม่ ???
เพราะข้อที่ 2 ที่คิดไว้ว่าจะมี Insight แต่เหมือนไม่มีเลย

คำถามที่ 4 **ลอง Plot graph โดยที่**

อยากทราบว่าแต่ละภาคมีอุตสาหกรรมอะไรที่โดดเด่น เผื่อเป็นแนวทางถ้าหากมีใครอยากทำงานใกล้บ้าน อาจเลือกหางานหรือเลือกเรียนในอุตสาหกรรมที่มีใกล้บ้าน

แรงงานใน กทม เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

แรงงานใน ปริมณฑล เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

แรงงานใน ภาคกลาง (ยกเว้น กทม-ปริมณฑล) เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

แรงงานใน ภาคตะวันออก เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

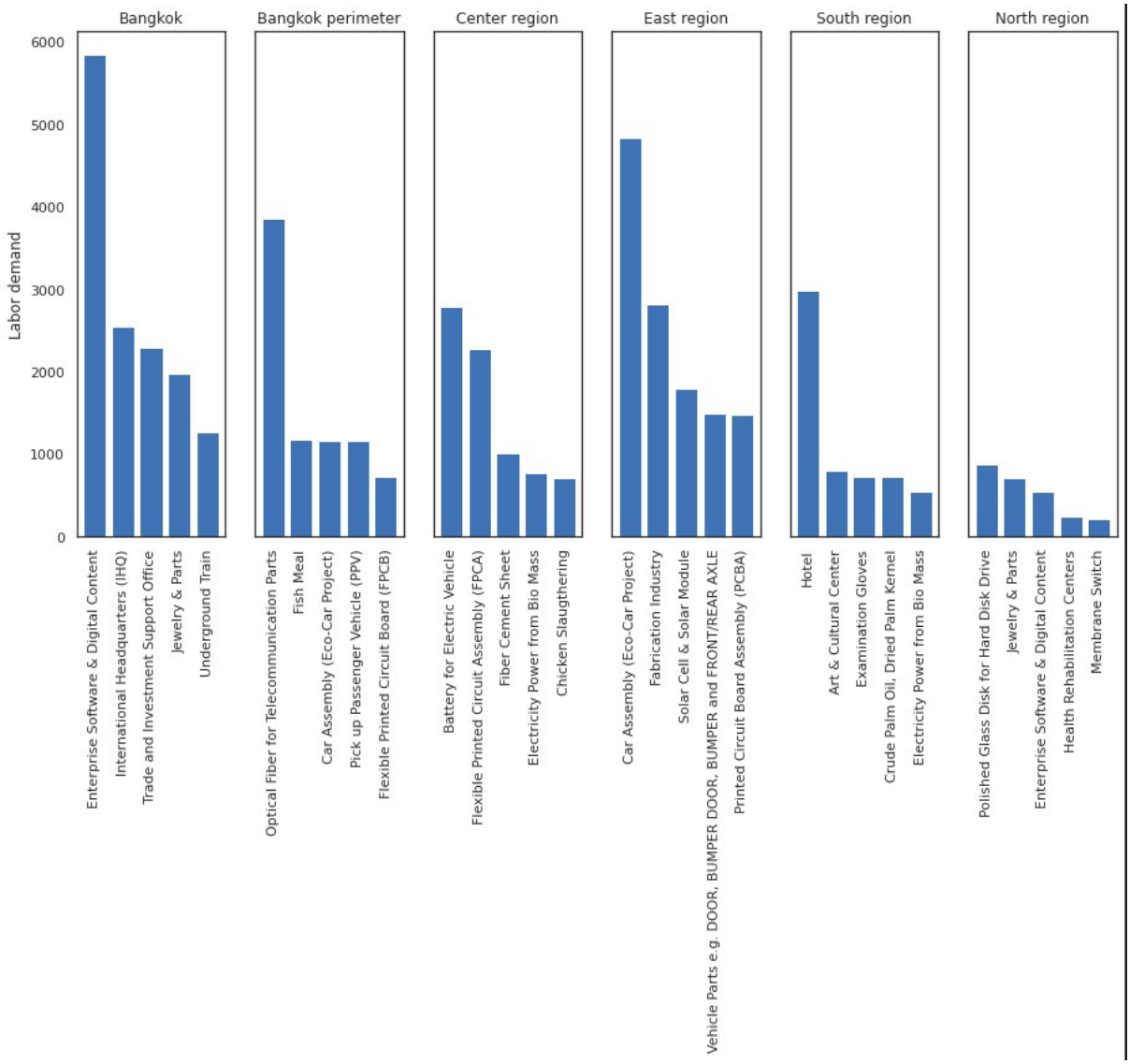
แรงงานในภาคใต้ เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

แรงงานในภาคเหนือ เรียงลำดับ กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากไปน้อย (เอา 5 อันดับแรก)

(ตรงนี้สิ่งที่เพิ่มมา คือต้องเอาภาคกลางหักลบ กทม. ออกก่อน เพื่อแยก กทม และปริมณฑล ออกจากกัน และค่อย groupby ผลิตภัณฑ์, nlarge = 5, sort จากมากไปน้อย)

result : จะเห็นว่าในแต่ละภาค จะมี Product (หรือกลุ่ม Industry) ที่มีเอกลักษณ์ต่างกันออกไป

กรุงเทพ : Enterprise Software & Digital Content, Trade&Invesment, International HQ, Jewelry, Underground Train
ปริมณฑล : Optical fiber for communication part , Fish meal, Car Assembly
ภาคกลาง : Battery for EV, Printed Circiut Board, Fiber Cement sheet
ภาคตะวันออก : Car Assembly, Fabrication , Solar Cell, Vehicle Part, Printed Circiut Board
ภาคใต้ : Hotel , Art & Culture , Examination Gloves , Crude Plam Oil, Electricity from Bio Mass
ภาคเหนือ : Polished Glass Disk for Hard Disk Drive, Jewelry, Enterprise Software & Digital Content, Health Rehabilitation, Membrane Switch

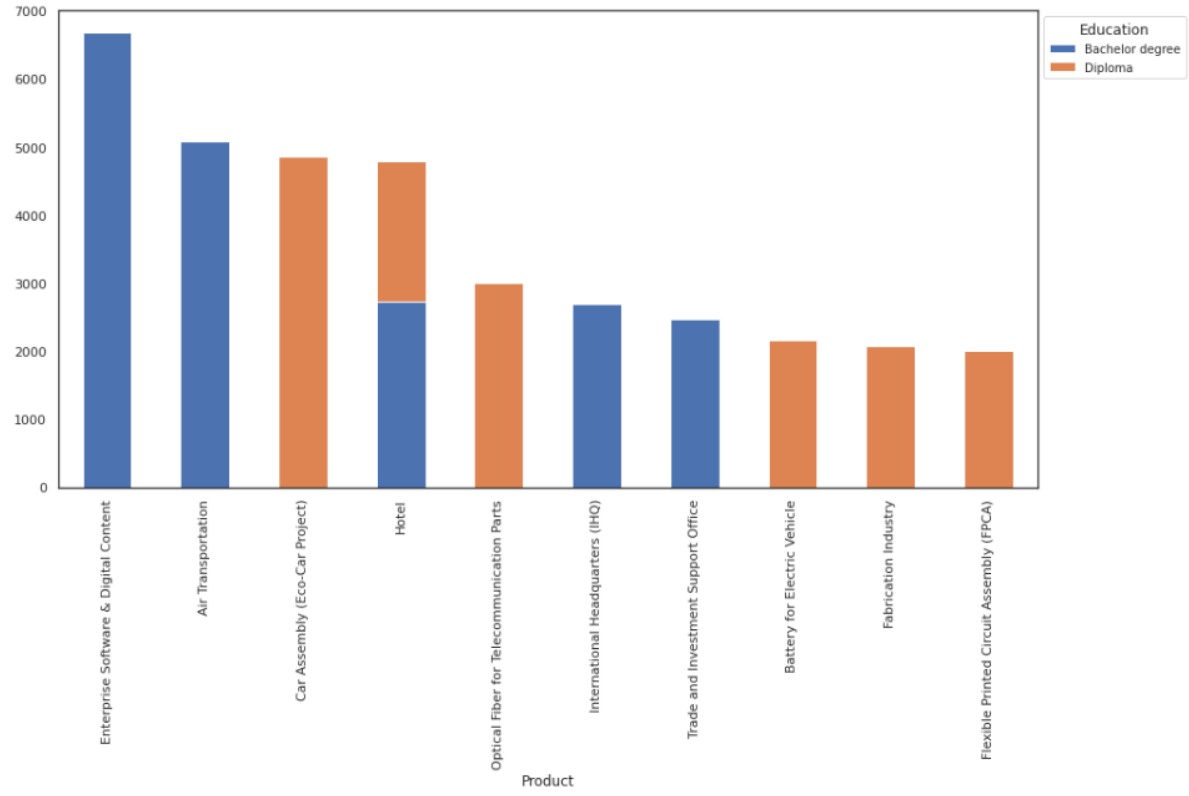


คำถามที่ 5

กลุ่มอุตสาหกรรมที่ต้องการแรงงานมากที่สุด เรียงลำดับจากมากไปน้อย group แยกระหว่าง กลุ่มปริญญาตรี , กลุ่มอาชีวะ เพื่อที่ว่าเป็น trand ว่าเราควรจะเรียนหรือฝึกงานฝึกประสบการณ์จากอุตสาหกรรมด้านไหน

result : กลุ่มปริญญาตรี : เป็น trend ของ Enterprise Software & Digital Content, Air Transpotation

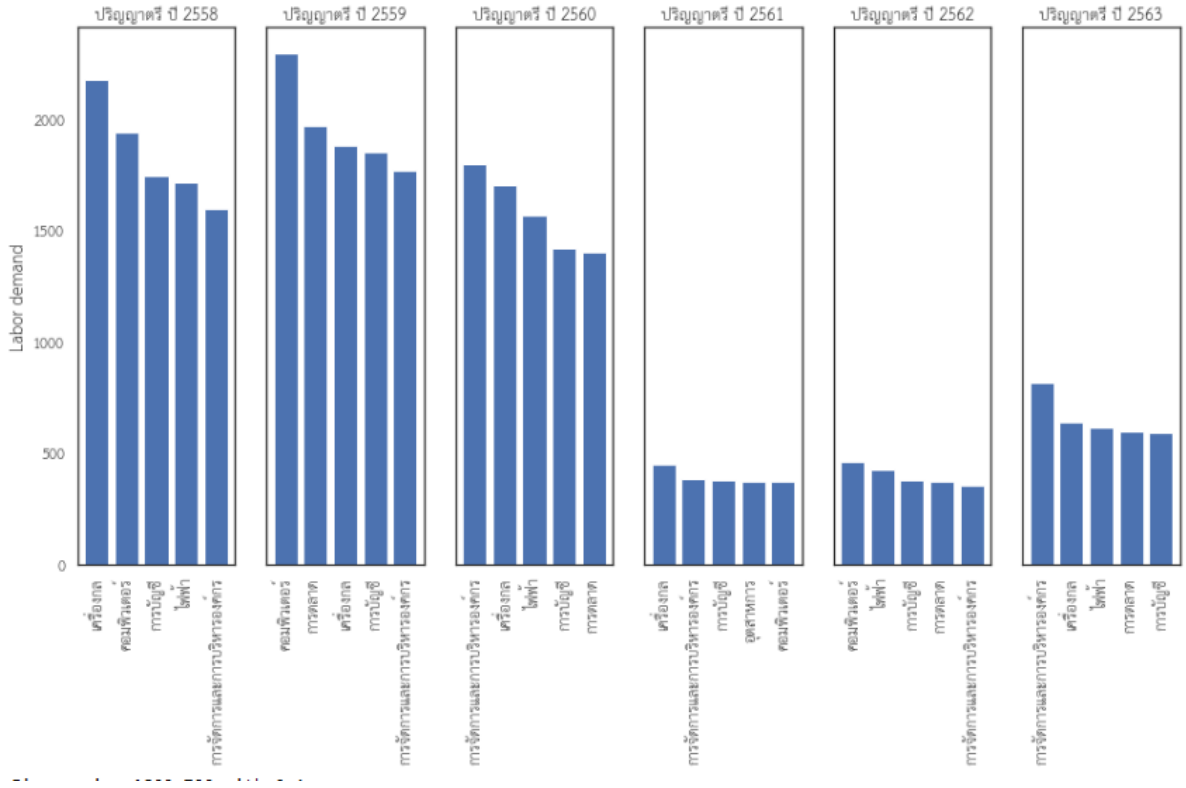
กลุ่มอาชีวะ : เป็น trend ของ Car Assembly, Optical fiber for communication part และ Battery for EV Car



คำถามที่ 6 สาขาของปริญญาตรีที่ต้องการแรงงานมากที่สุด 5 อันดับในแต่ละปี เปรียบเทียบ trend เรียงไปตั้งแต่ปี 58-63 ว่ามีการเปลี่ยนแปลงหรือไม่

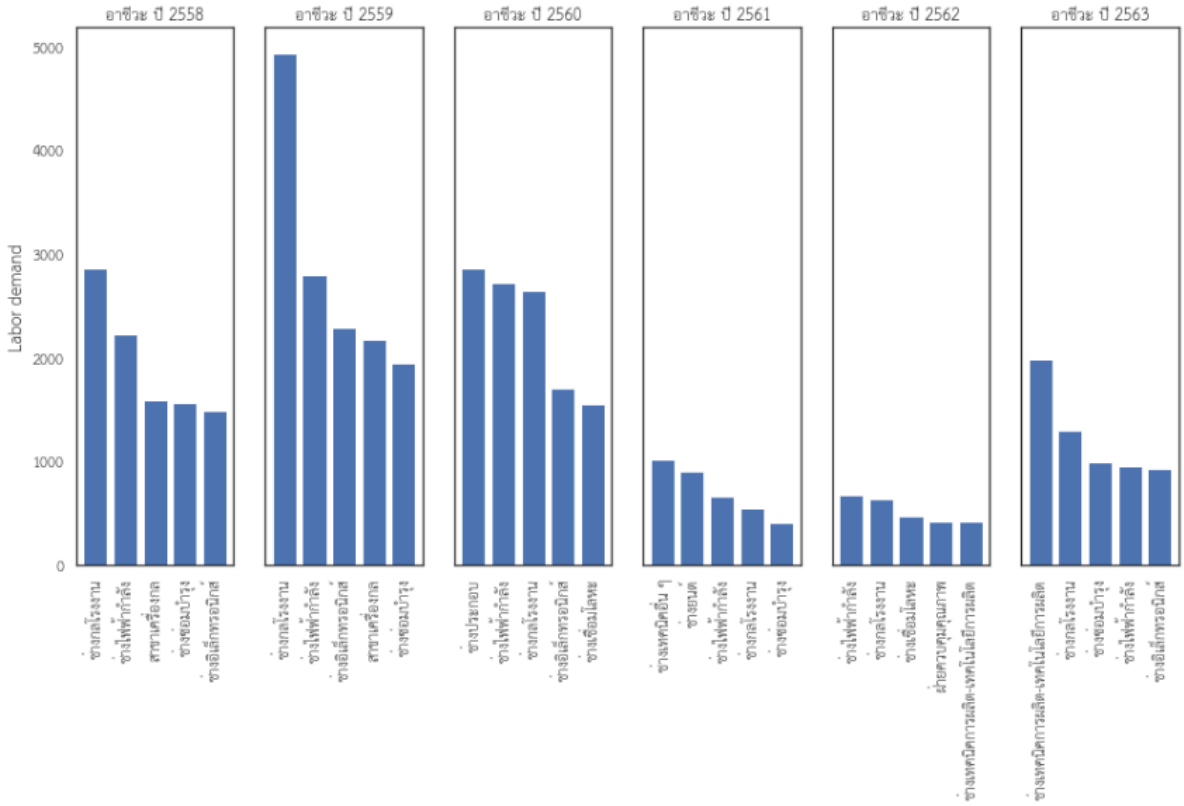
(ตรงนี้สิ่งที่เพิ่มมา ต้อง plot จำนวนปี ทั้งหมด 6 ปี จนครบ, ซึ่งก็จะใช้เวลา,จึง set ปี(year) เป็น index แล้ว loop เอา จนครบ)

result : ท็อปฟอร์มคือ : เครื่องกล คอมพิวเตอร์ ไฟฟ้า บัญชี การจัดการบริหารองค์กร ไม่ค่อยต่างมากในแต่ละปี



คำถามที่ 7 สาขาของอาชีวะที่ต้องการแรงงานมากที่สุด 5 อันดับในแต่ละปี เปรียบเทียบ trend เรียงไปตั้งแต่ปี 58-63 ว่ามีการเปลี่ยนแปลงหรือไม่

result : สาขาช่างกลโรงงาน สาขาช่างไฟฟ้ากำลัง เป็น 2 สาขา ที่ต้องการแรงงาน และ ติดอับดับตลอดในช่วงปี 58-64 , ที่เหลือจะมีความต้องการประปราย และมีอีก 2 ช่างที่มีความต้องการไม่แพ้กันก็คือช่างอิเล็กทรอนิกส์ และช่างซ่อมบำรุง

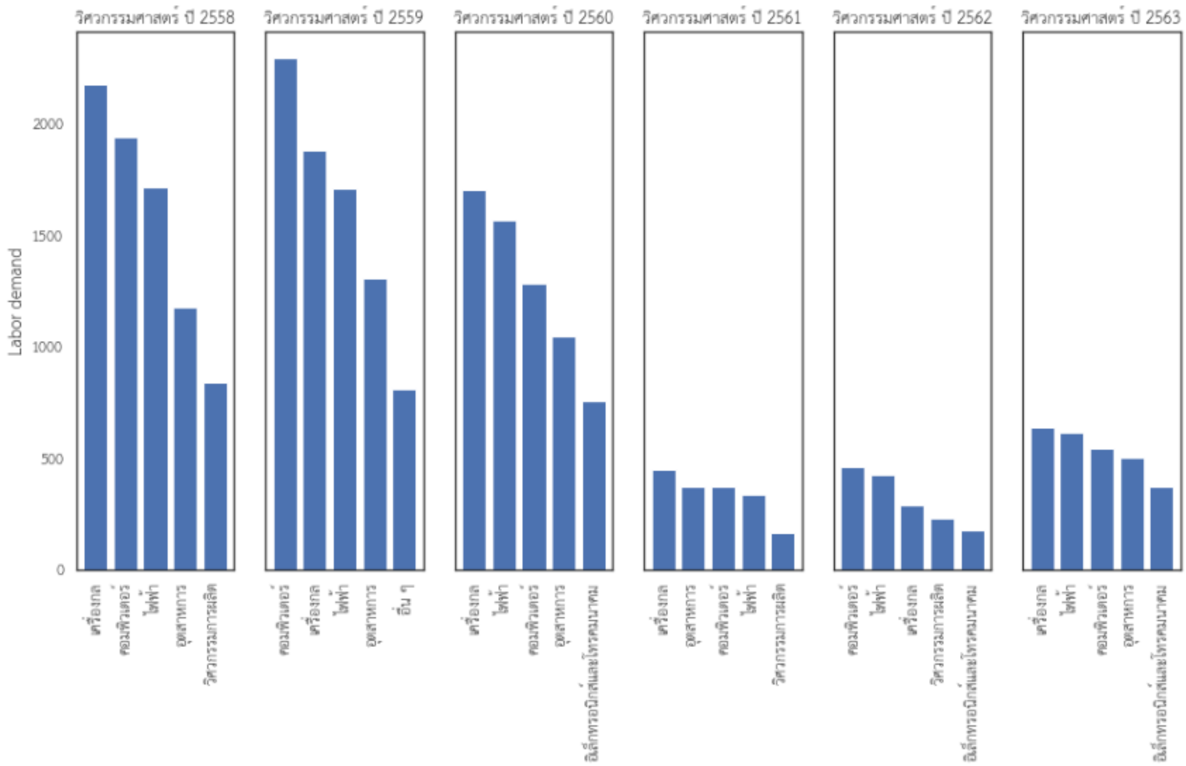


<Figure size 1800x720 with 0 Axes>

คำถามที่ 8

เฉพาะวิศวะ สาขาที่ต้องการแรงงานมากที่สุด 5 อันดับในแต่ละปี เปรียบเทียบ trend เรียงไปตั้งแต่ปี 58-64 ว่ามีการเปลี่ยนแปลงหรือไม่ เนื่องจากวิศวกรรมศาสตร์ เป็นคณะยอดฮิตที่คนอยากเรียน จึงอยากรู้ว่าแล้วในช่วงนี้ วิศวะสาขาไหนเป็นที่ต้องการของตลาดมากที่สุด เพื่อมีประโยชน์วางแผนในการเรียนต่อ

result : ในช่วงปี 58-63 ท็อปฟอร์ม 4 อันดับแรก : เครื่องกล คอมพิวเตอร์ ไฟฟ้า ยืนยันว่าไม้ตกงาน



<Figure size 1800x720 with 0 Axes>

.......จบเท่านี้ค่ะ.........