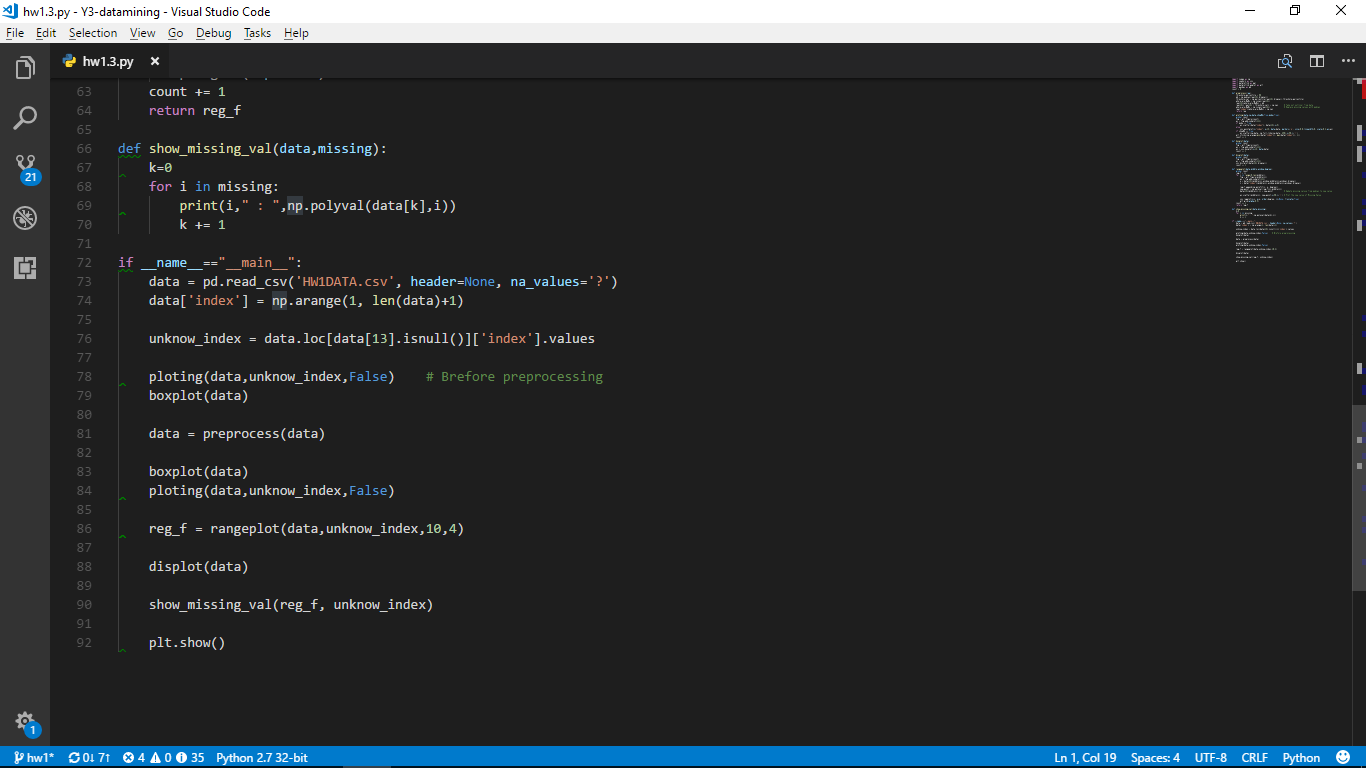
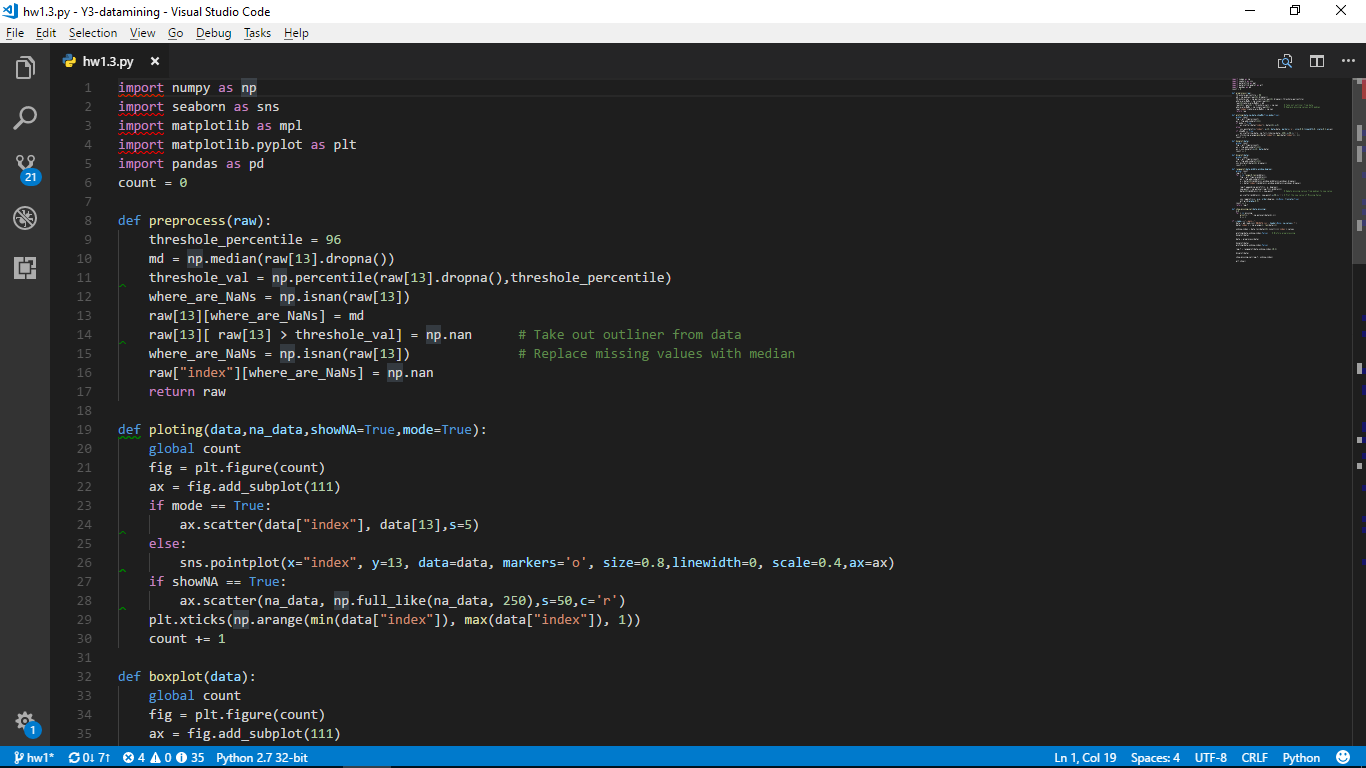
# 261448 Data Mining Homework 1

580610642\_CPE

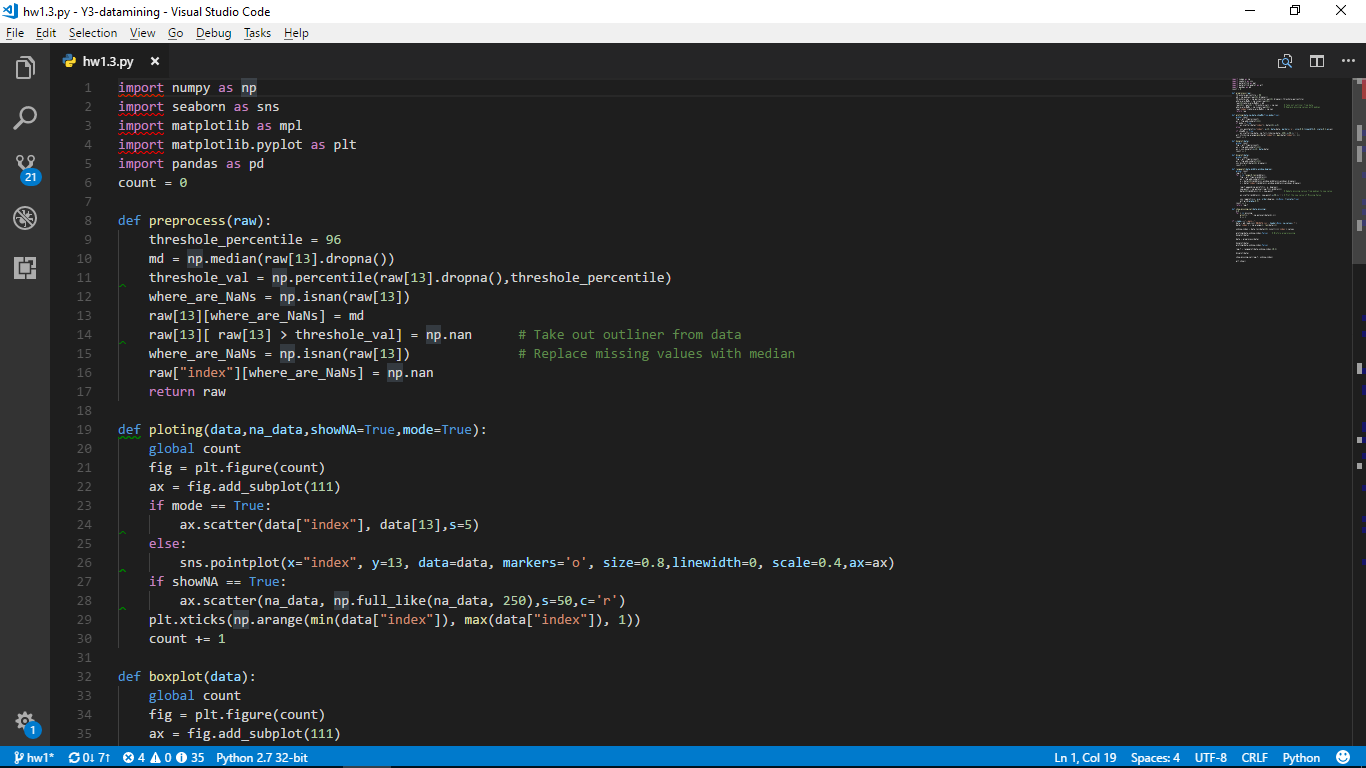
# ส่วน CODE ของโปรแกรม



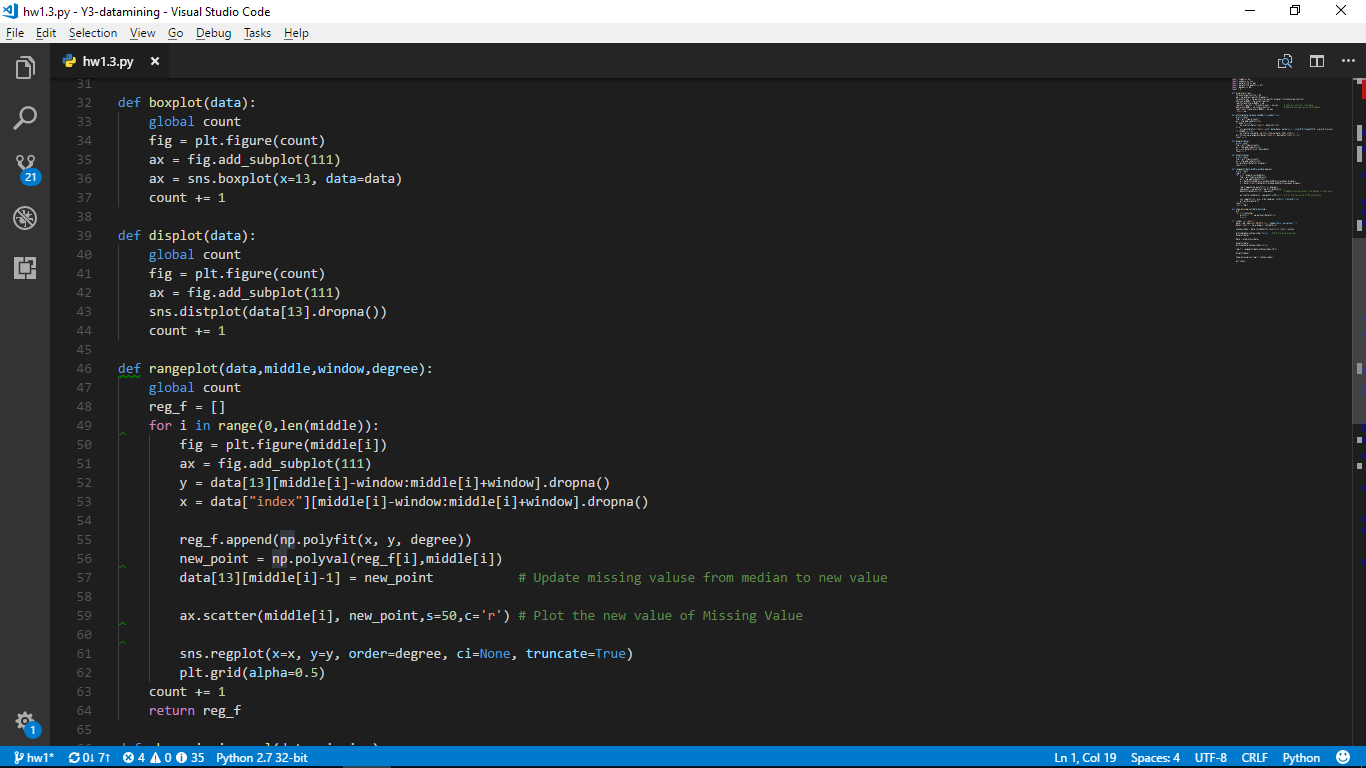
Code การทำงานหลักของโปรแกรม เริ่มจากการโหลด ดาต้าเซต และ มาร์กจุดที่เป็น missing\_values แล้วจึง แสดงการกระจายตัวของข้อมูล ผ่านฟังชั่น **plotting()** และ **boxplot()** เพื่อจะได้มองเห็นการกระจายตัวของข้อมูล เพื่อประการตัดสินใจว่าควรประมวลผลข้อมูลอย่างไร แล้วจึงนำไป **preprocessing** แล้วจึงแสดงการผระจายตัวอีกรอบ



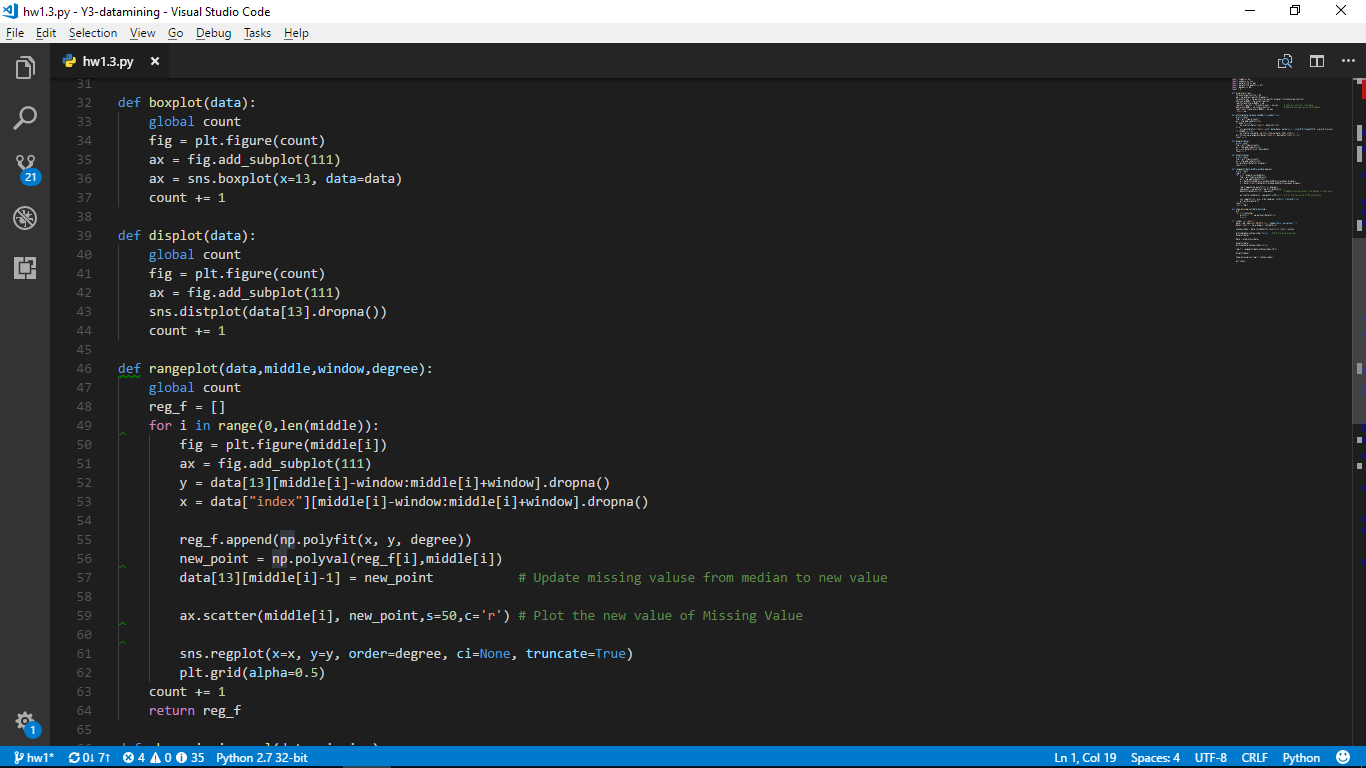
**Function preprocessing** ทำการประมวลผลข้อมูลเบื้องต้น



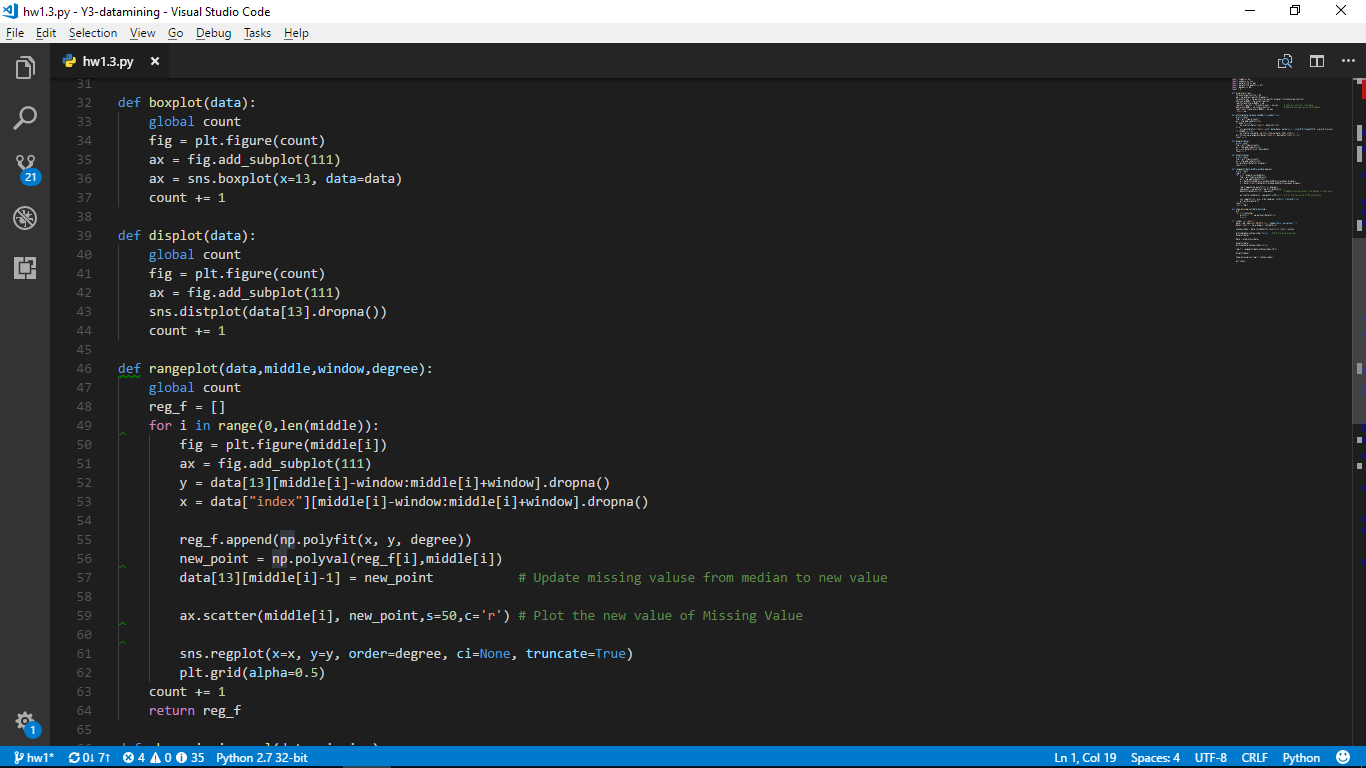
**Function plotting** แสดงการกระจายตัวของข้อมูล



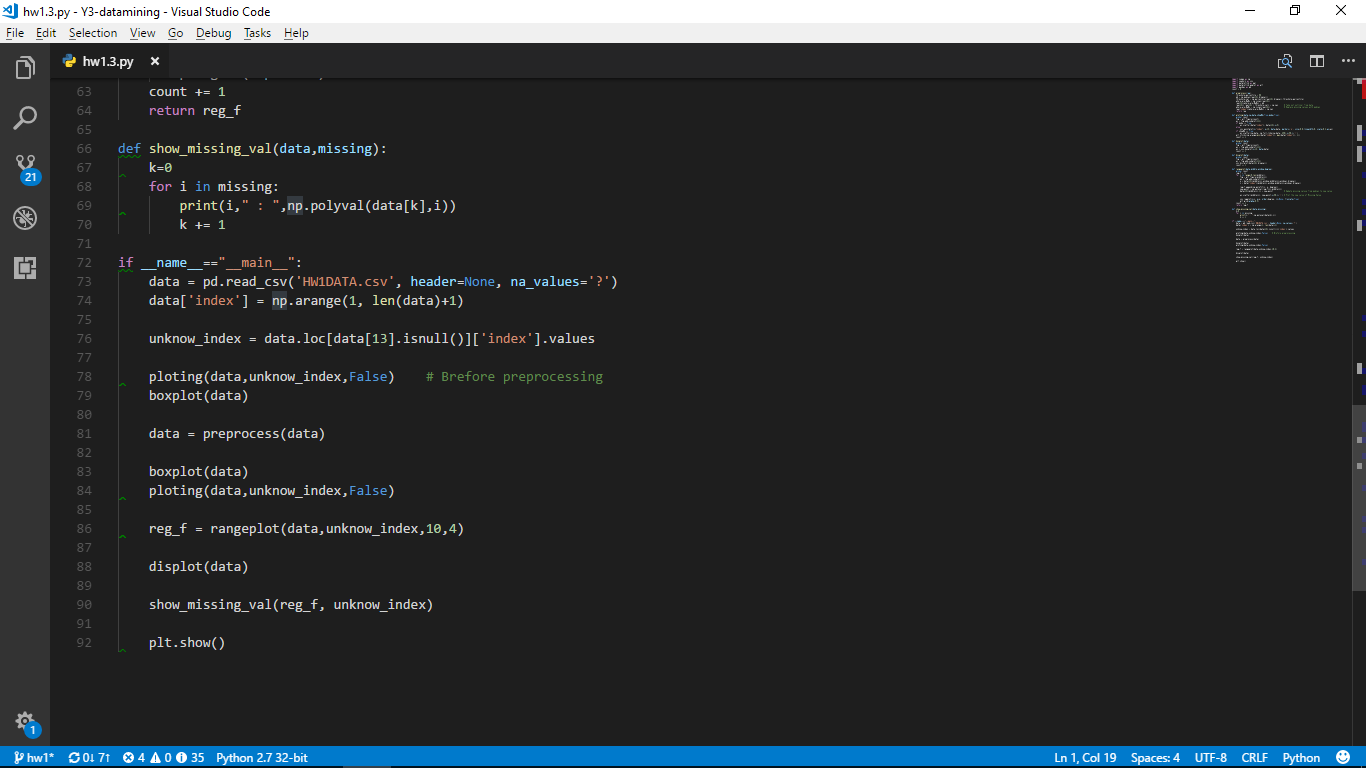
**Function boxplot** แสดงการกระจายตัวของข้อมูลด้วยboxplot



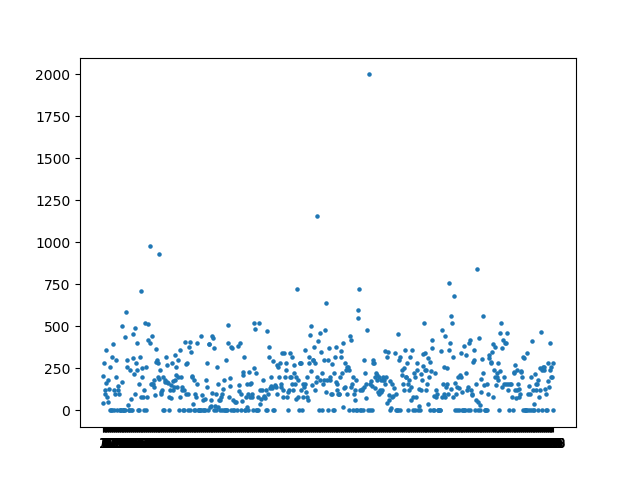
**Function displot** แสดงการกระจายตัวของข้อมูลด้วยhistrogram



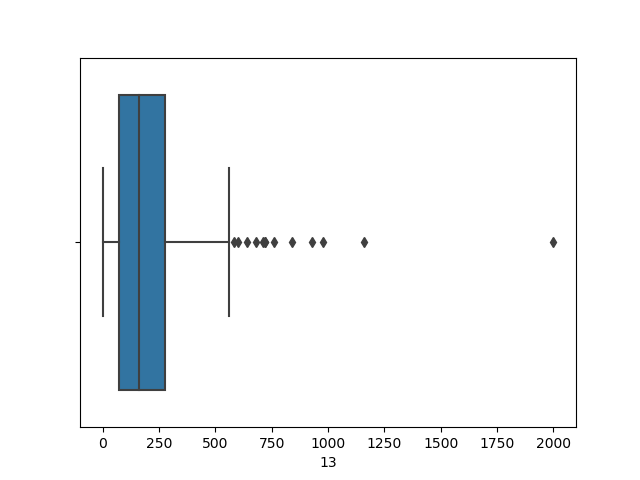
**Function displot** แบ่งช่วงข้อมูลเป็น ช่วงวินโดว์ขนาดเล็กๆ แล้วจึงทำการ **polyfit()** ข้อมูลในช่วง โดยมี missing value เป็นจุดศูนย์กลางของแต่ละวิโนดว

 **Function** แสดงค่าMissing value

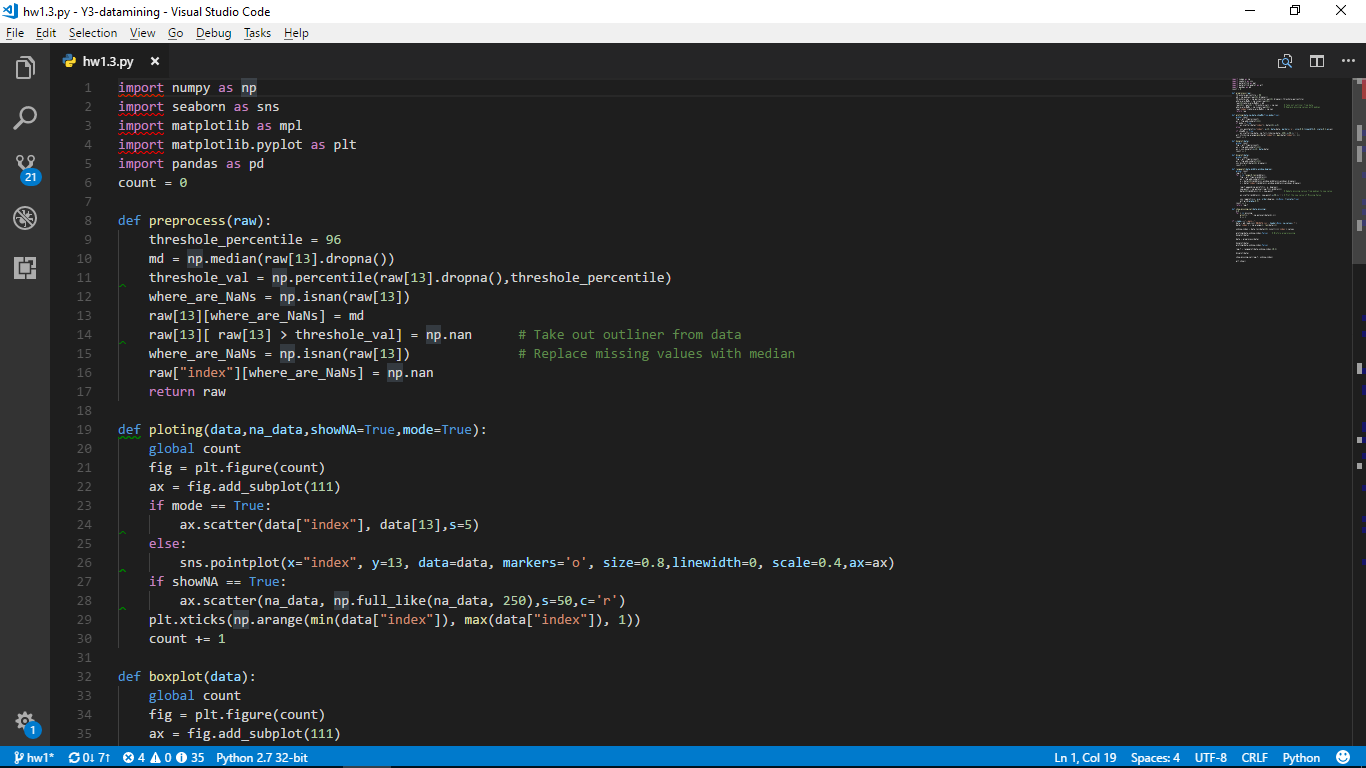
# ผลข้อมูลที่ได้



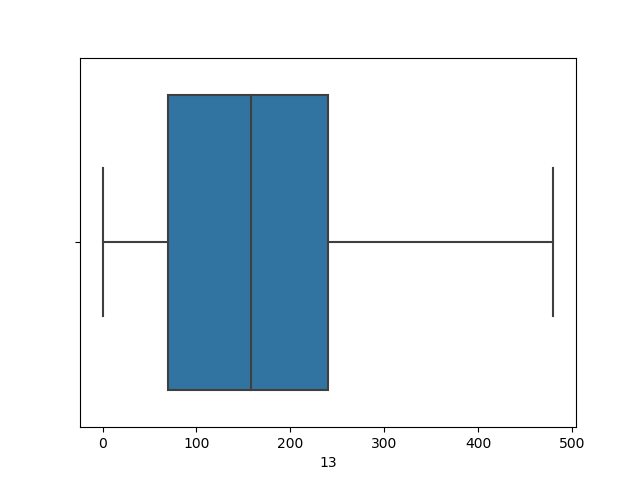
Scatter plot แสดงผลของข้อมูลดิบที่ได้มา จะเห็นได้ว่าข้อมูลการจายตัวอย่างไม่เป็นระเบียบ และมีบางข้อมูลที่กระโดดมากๆจากกลุ่ม

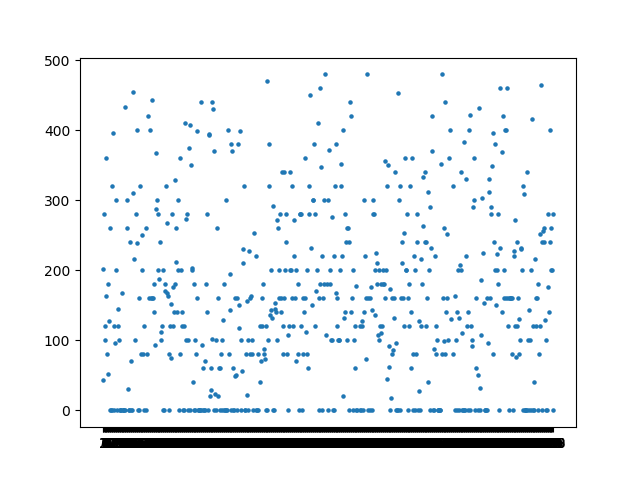


Boxplot แสดงให้เห็นว่าค่าส่วนมากจะกระจุกตัวอยู่ในช่วง 0 – 560 โดยประมาณ ดังนั้นจึงพิจารณาว่าข้อมูลหลังช่วงนี้ไปจะเป็น Outliner และจะไม่นำมาประมวลผล



จากcode การ preprocess จะเห็นว่าได้ทำการเลือก percentile ในการคัดข้อมูล ออกอยู่ที่ 96 ค่าที่เข้าเกณนี้จะจัดเป็น outliner และ สำหรับค่า missing เลือกที่ให้เป็นค่า median ไปก่อน ในช่วงประมวลผล





หลังจาก preprocessing เห็นได้ว่าข้อมูลไม่มี outliner แล้ว แต่ปัญหาต่อมาคือ หากจะใช้ข้อมูลทั้งหมดในการทำ regression จะไม่เป็นการดีแน่ จึงเลือกที่จะจะเพิจารณาข้อมูล รอบๆmiss value แทน เนื่องจากเป้าคือ การหาค่าmissing value ด้วย regression การใช้ข้อมูลรอบควรจะเหมาะสมกว่า

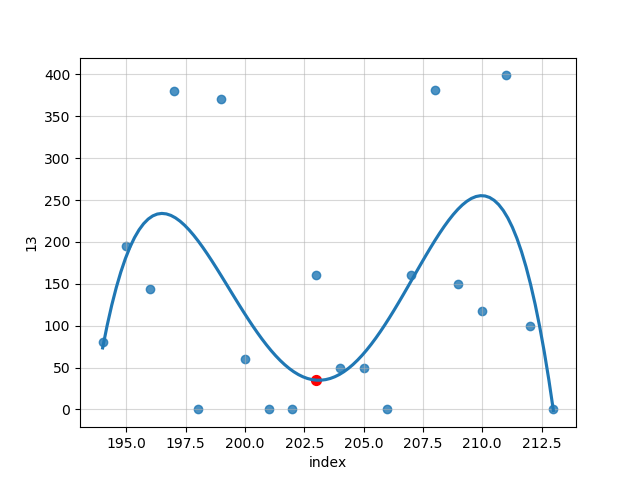
# ผลลัพธ์

ใช้ **window size** เป็นข้อมูล**10** ดัวทางด้านซ้ายและขวา ของ missing value และ degree ของฟังชั่น **polyfit** เป็น

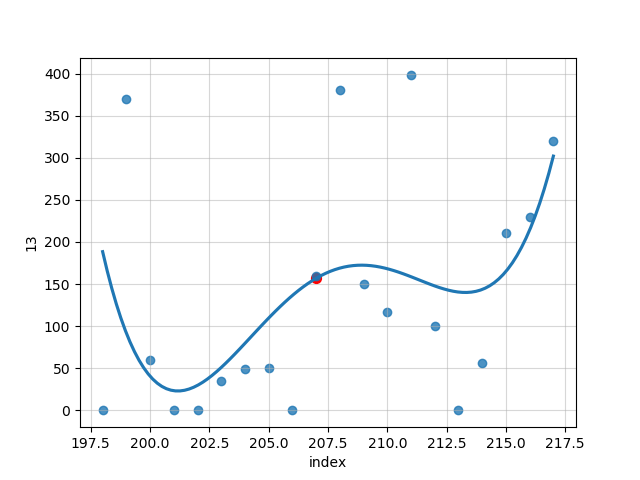
4F(72) = 298.05925257736817

# C:\Users\darunphop\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Figure_72.png

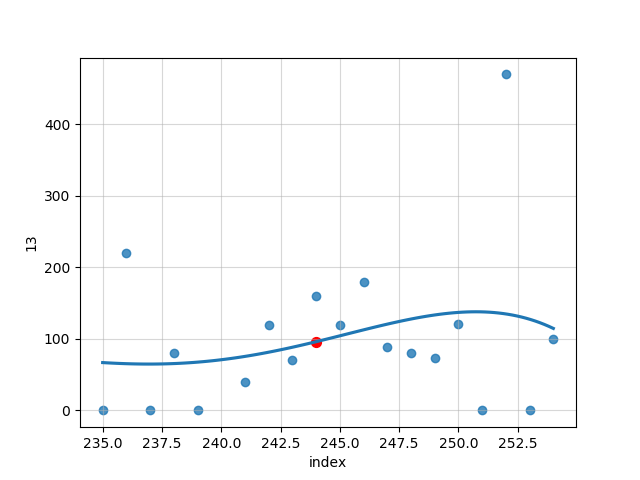
# F(203) = 35.0040285885334



# F(207) = 156.82328571379185



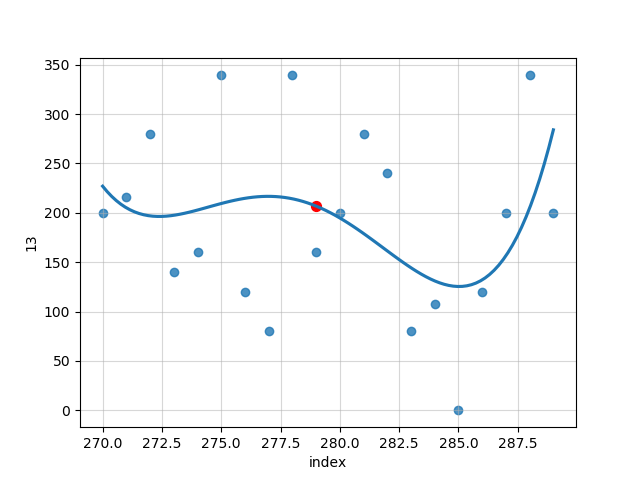
# F(244) = 96.07176165096462



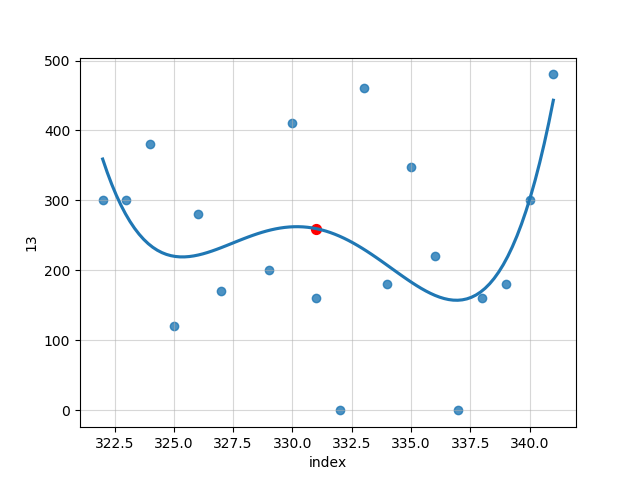
# F(271) = 215.8470665924251

# C:\Users\darunphop\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Figure_271.png

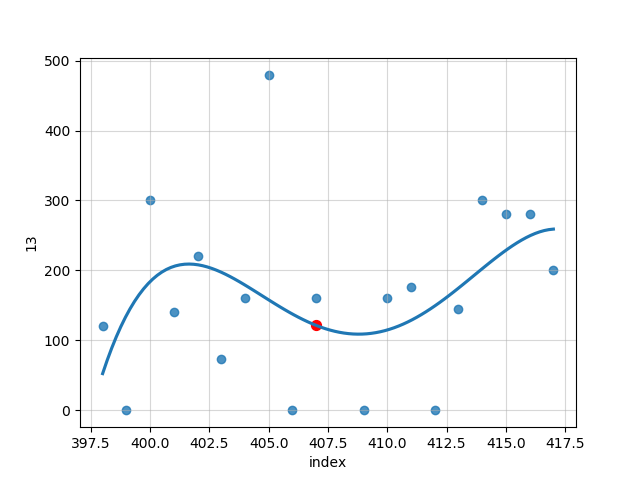
# F(279) = 206.7659551203251



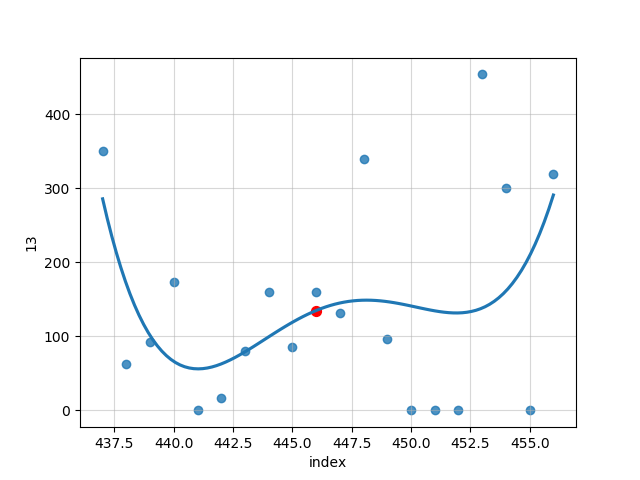
# F(331) = 259.68719732761383



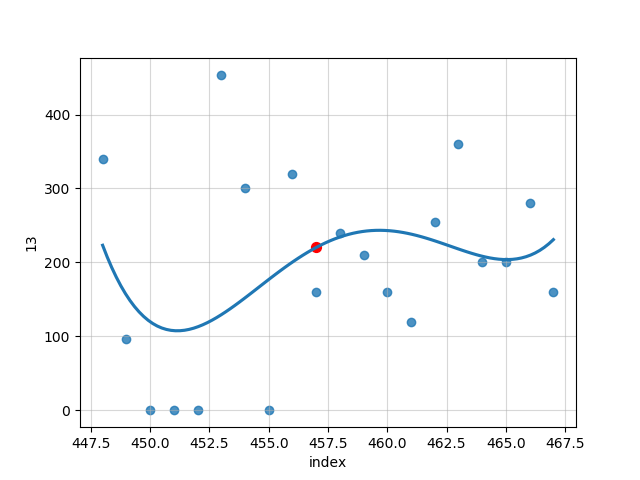
# F(407) = 121.25457167625427



# F(446) = 134.94110941886902

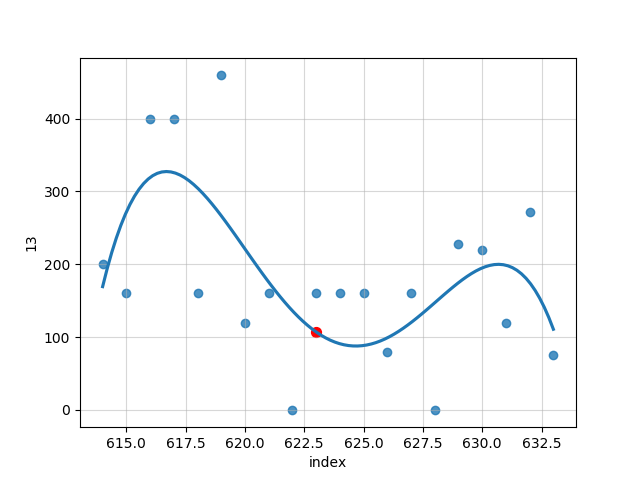


# F(457) = 220.1053376197815

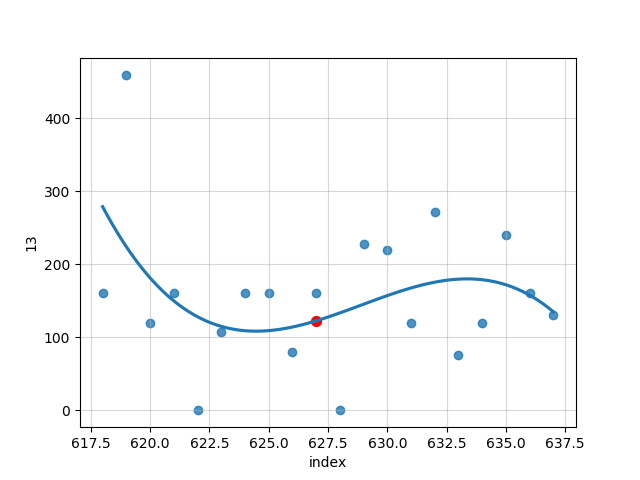


# F(593) = 280.9908380508423C:\Users\darunphop\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Figure_593.png

# F(623) = 107.02058982849121



# F(627) = 122.48888589441776



# สรุปผล

|  |  |  |  |
| --- | --- | --- | --- |
| **ตำแหน่ง** | **ค่า** | **Window size** | **degree** |
| 72 | 298.05925257736817 | 10 | 4 |
| 203 | 35.0040285885334 | 10 | 4 |
| 207 | 156.82328571379185 | 10 | 4 |
| 244 | 96.07176165096462 | 10 | 4 |
| 271 | 215.8470665924251 | 10 | 4 |
| 279 | 206.7659551203251 | 10 | 4 |
| 331 | 259.68719732761383 | 10 | 4 |
| 407 | 121.25457167625427 | 10 | 4 |
| 446 | 134.94110941886902 | 10 | 4 |
| 457 | 220.1053376197815 | 10 | 4 |
| 593 | 280.9908380508423 | 10 | 4 |
| 623 | 107.02058982849121 | 10 | 4 |
| 627 | 122.48888589441776 | 10 | 4 |