

**เรื่อง Analyzing Crime in San Francisco**

จัดทำโดย

นายธนดล สร้อยสน รหัสประจำตัว  61070297

อาจารย์ประจำวิชา

รศ. ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชา 06026117 FUNDAMENTALS OF DATA SCIENCE

หลักสูตรวิทยาศาสตร์บัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์เชิงธุรกิจ

ภาคเรียนที่  2 ปีการศึกษา  2562

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

**วัตถุประสงค์:**

-จำแนกประเภทอาชญากรรมโดยอัตโนมัติ แล้วส่งเรื่องไปให้กับเจ้าหน้าที่ที่เกี่ยวข้องกับอาชญากรรมประเภทนั้นๆ เพื่อลดขึ้นตอนในการตรวจสอบประเภทของอาชญากรรม

-เพิ่มความรวดเร็วในการจัดการคดีเพื่อให้หน่วยงานที่เกี่ยวข้องกับการรักษาความปลอดภัยสามารถวางแผนหรือรับมือกับเหตุการณ์ที่จะเกิดขึ้นได้

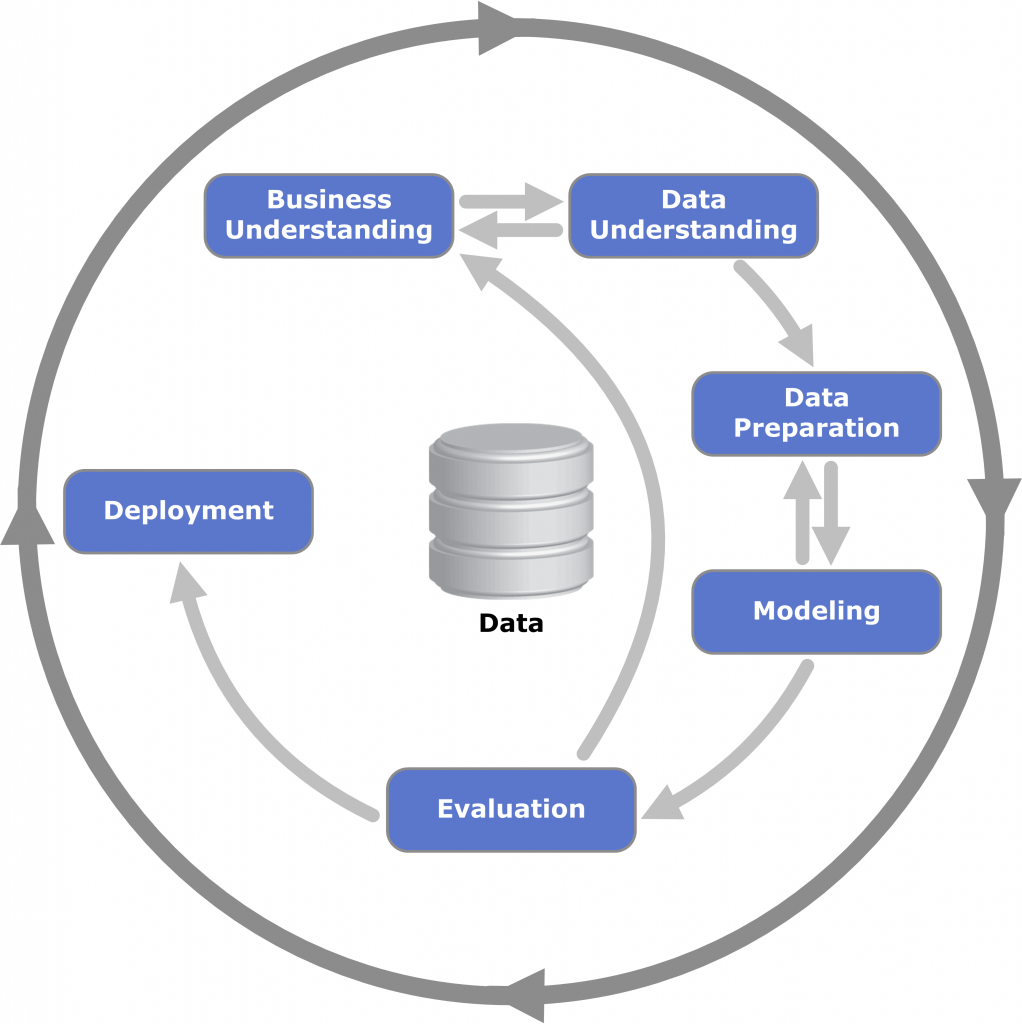
**ข้อมูลที่ใช้ ตัวแปรที่มี:**

เป็นข้อมูลเกี่ยวกับอัตราอาชญากรรมในภูมิภาคต่างๆ ของเมือง San Francisco ในปี 2016 พร้อมประเด็นสำคัญอื่นๆ ที่เกี่ยวข้องกับอาชญากรรม

ข้อมูลใน Data Set ประกอบด้วย

* IncidntNum : หมายเลขของเหตุการณ์
* Category : ประเภทของอาชญากรรม
* Descript : รายละเอียดของอาชญากรรม
* DayOfWeek : วันที่เกิดอาชญากรรม
* Date : วัน เดือน ปี ที่เกิดอาชญากรรม
* Time : เวลาที่เกิดอาชญากรรม
* PdDistrict : เขต
* Resolution : ชนิดของการลงโทษการทำอาชญากรรม
* Address : ที่อยู่ที่เกิดเหตุอาชญากรรม
* X : ละติจูดของสถานที่เกิดอาชญากรรม
* Y : ลองจิจูดของสถานที่เกิดอาชญากรรม
* Location : พิกัด
* PdId : หมายเลขอ้างอิง

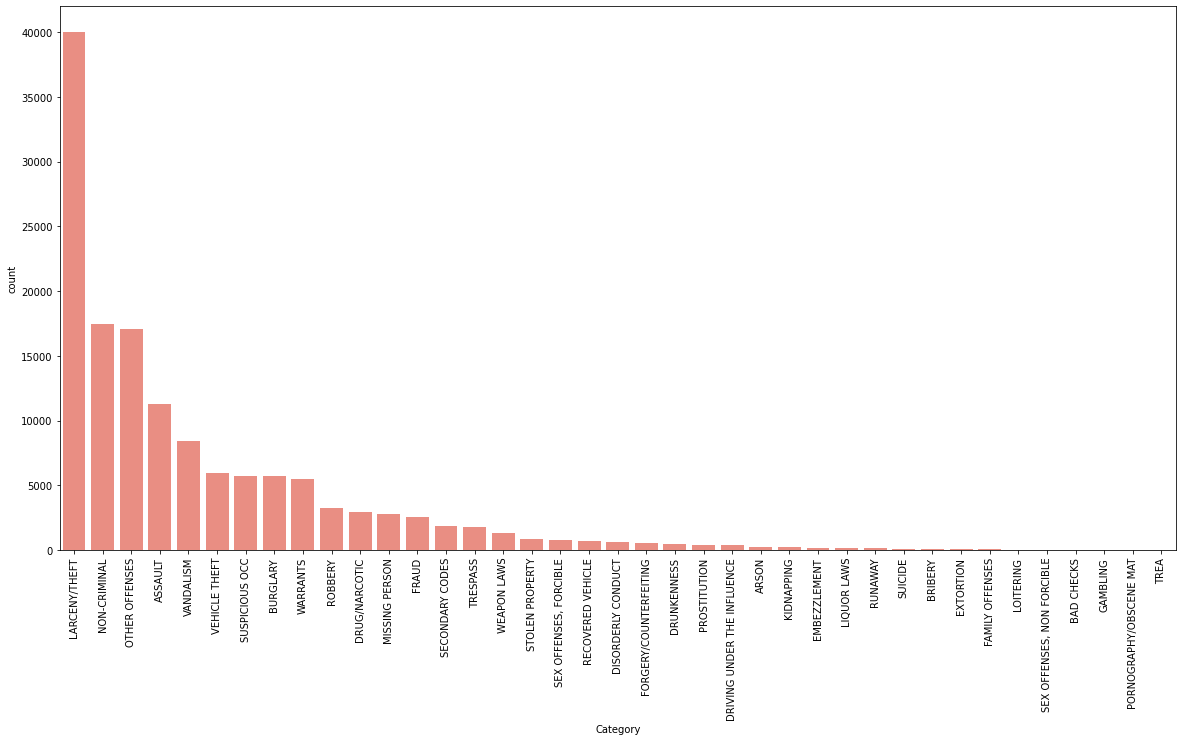
โดยในกระบวนการ CRISP-DM นี้จะประกอบด้วย 6 ขั้นตอนดังในรูปด้านล่าง แต่ละขั้นตอนในรูปจะเป็นขั้นตอนที่ต่อเนื่องกันนั่นคือขั้นตอนถัดไปจะรอผลลัพธ์จากขั้นตอนก่อนหน้าซึ่งแสดงด้วยลูกศรที่เชื่อมระหว่างกล่องสี่เหลี่ยมแต่ละกล่องตัวอย่าง เช่น เมื่อได้ผลลัพธ์จากขั้นตอนการเตรียมข้อมูล (Data Preparation) แล้วจะนำไปสร้างโมเดลจำแนกประเภทข้อมูลในขั้น Modeling และหลังจากนั้นอาจจะย้อนกลับมาเปลี่ยนแปลงข้อมูลให้ถูกต้องมากขึ้นเป็นต้น



**1.Business Understanding**

เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางดาต้าไมน์นิงพร้อมทั้งวางแผนในการดำเนินการคร่าวๆ

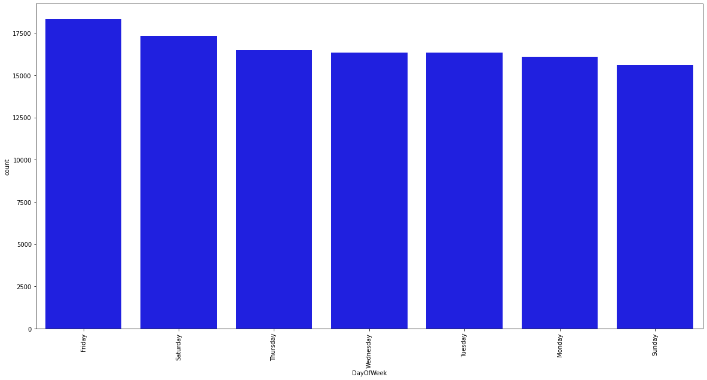
San Francisco คือ เมืองในรัฐแคลิฟอร์เนีย สหรัฐอเมริกา ซึ่งเป็นเมืองที่มีความหนาแน่นประชากรเป็นอันดับสองของประเทศ เนื่องจากในแต่ละวันแต่ละเดือนของปี 2016 อาจมีการเกิดอาชญากรรมหลายครั้งหลายประเภทซึ่งอาจจำแนกประเภทไม่ถูกและไม่เห็นภาพรวมในการเกิดอาชญากรรม เราจึงเสนอที่จะทำโมเดลจำแนกประเภทอัตโนมัติและทำนายอาชญากรรมที่เข้ามาว่าเป็นประเภทอะไรบ้างเพื่อที่จะได้ส่งไปให้เจ้าหน้าที่ที่เกี่ยวข้องเพื่อลดเวลาของเจ้าหน้าที่ ที่จะใช้ในการวิเคราะห์แนวโน้มการเกิดเหตุการณ์ที่เป็นอาชญากรรมเพื่อจัดการให้อาญชากรรมลดลงและให้หน่วยงานที่เกี่ยวข้องกับการรักษาความปลอดภัยสามารถวางแผนหรือรับมือกับเหตุการณ์ที่จะเกิดขึ้นได้

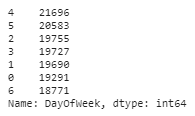
กราฟแสดงจำนวนการเกิดอาชญากรรมของแต่ละประเภท

กราฟแสดงการเกิดอาชญากรรมในแต่ละวัน ทำให้ทราบว่าวันศุกร์เกิดอาชญากรรมเยอะที่สุดจำนวน

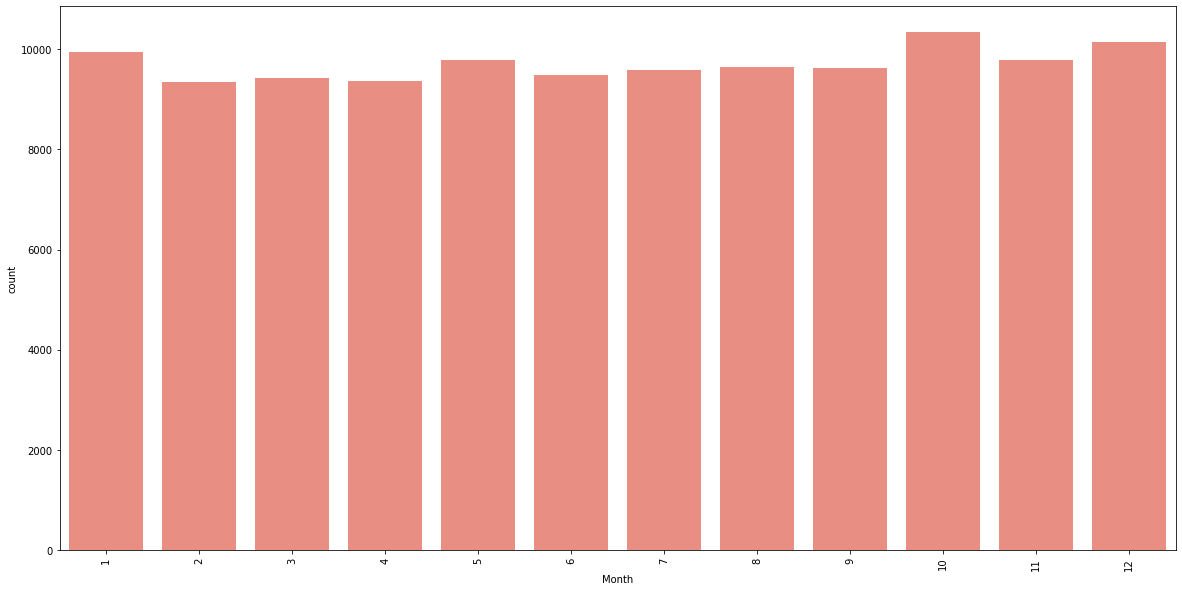
21696ครั้ง และ วันที่เกิดอาชญากรรมน้อยที่สุดคือวันอาทิตยจำนวน18771ครั้ง

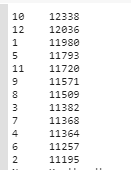
โดยเลข 0=จันทร์,1=อังคาร,2=พุธ,3=พฤหัสบดี,4=ศุกร์,5=เสาร์,6=อาทิตย์

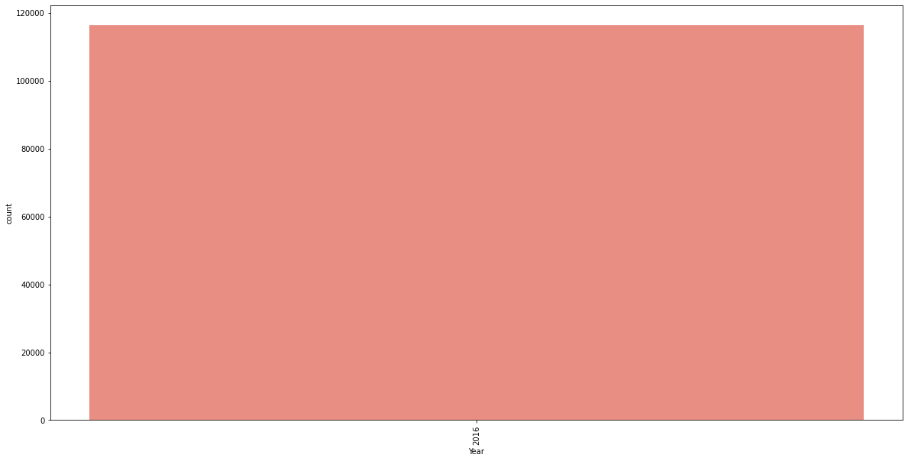




กราฟแสดงการเกิดอาชญากรรมในแต่ละเดือน ทำให้ทราบว่าเดือนที่10มีการเกิดอาชญากรรมเยอะที่สุดเท่ากับ12338ครั้ง และเดือนที่2เกิดอาชญากรรมน้อยที่สุดเท่ากับ11195ครั้ง

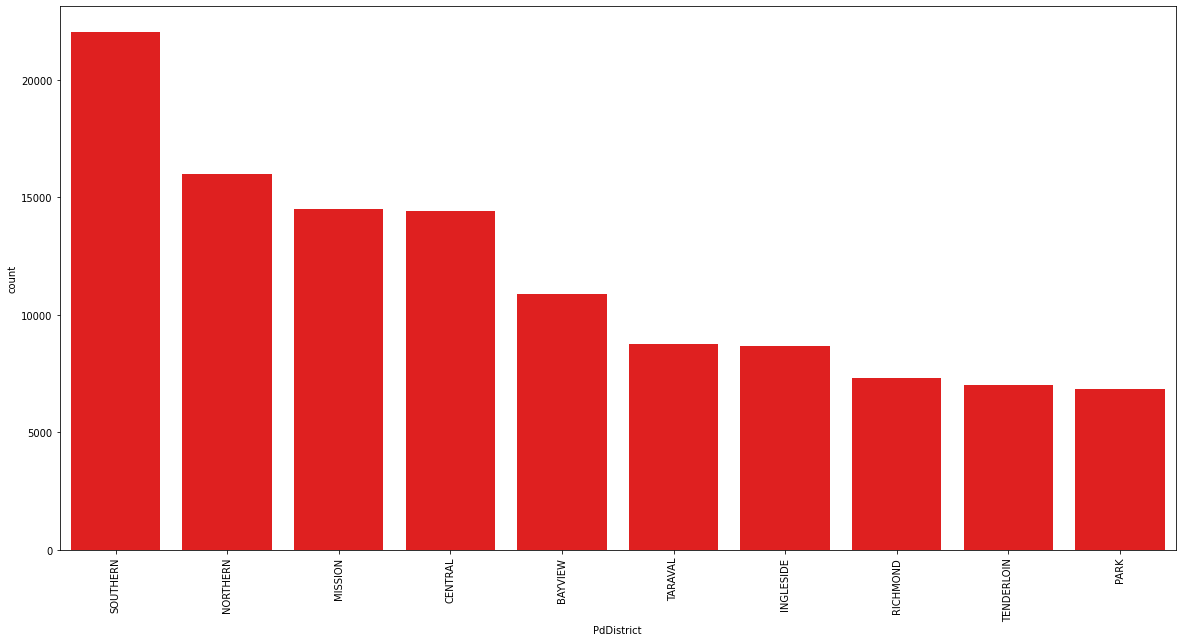


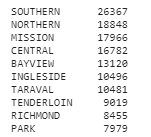


กราฟแสดงการเกิดอาชญากรรมในปี 2016 มีจำนวนทั้งหมด 139513 ครั้ง

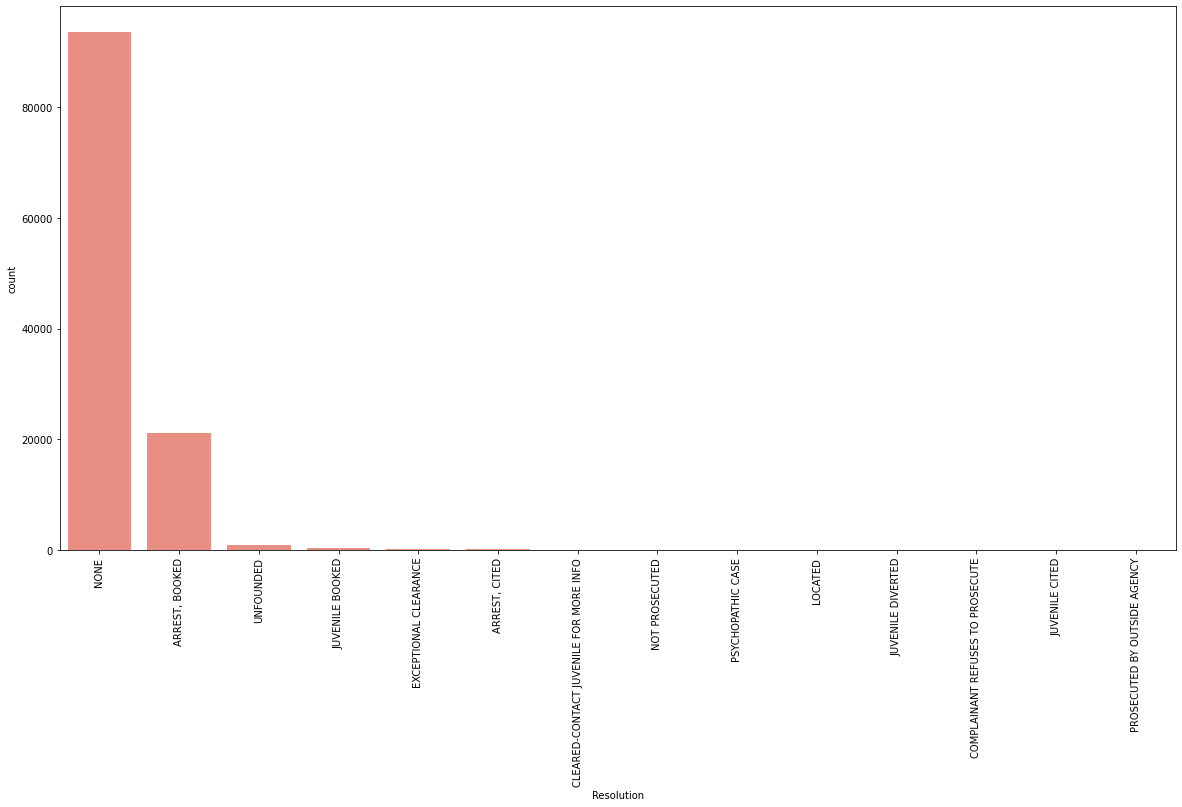


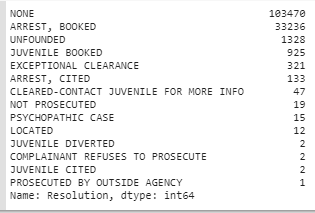
กราฟแสดงการเกิดอาชญากรรมในแต่ละเขต ทำให้ทราบว่าเขตSOUTHERNมีการเกิดอาชญากรรมเยอะที่สุดจำนวน26367ครั้ง และ เขตPARKมีการเกิดอาชญากรรมน้อยที่สุดจำนวน 7979ครั้ง

****



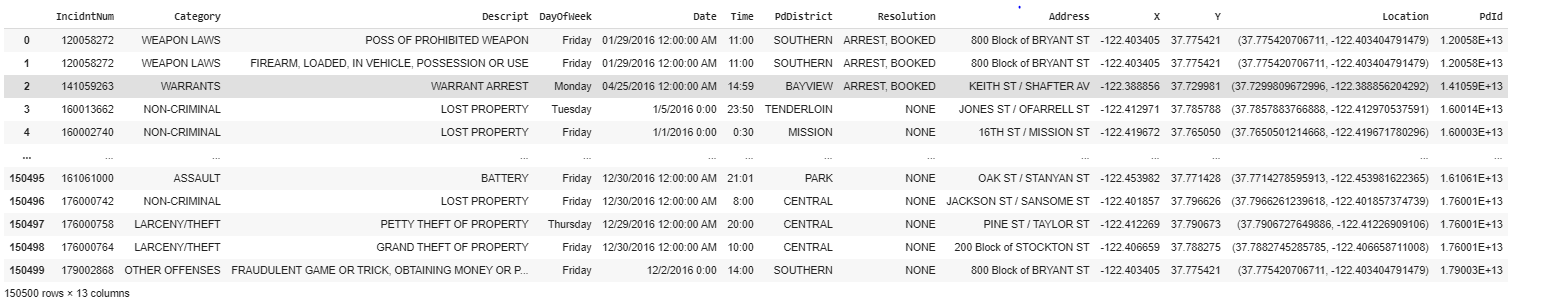
กราฟแสดงจำนวนชนิดของการลงโทษการทำอาชญากรรม

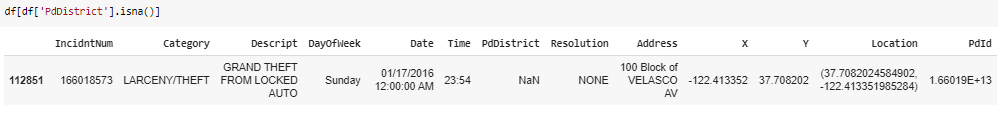
****

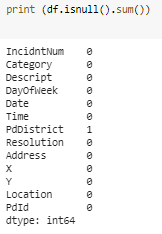
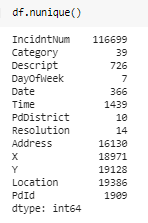


**2. Data Understanding**

ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์



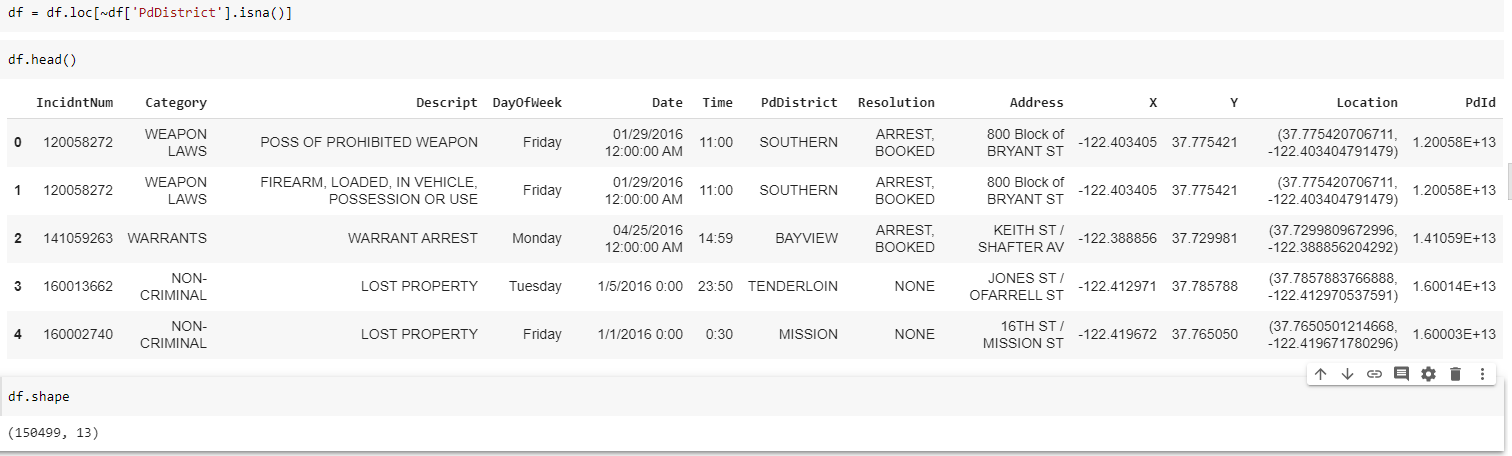
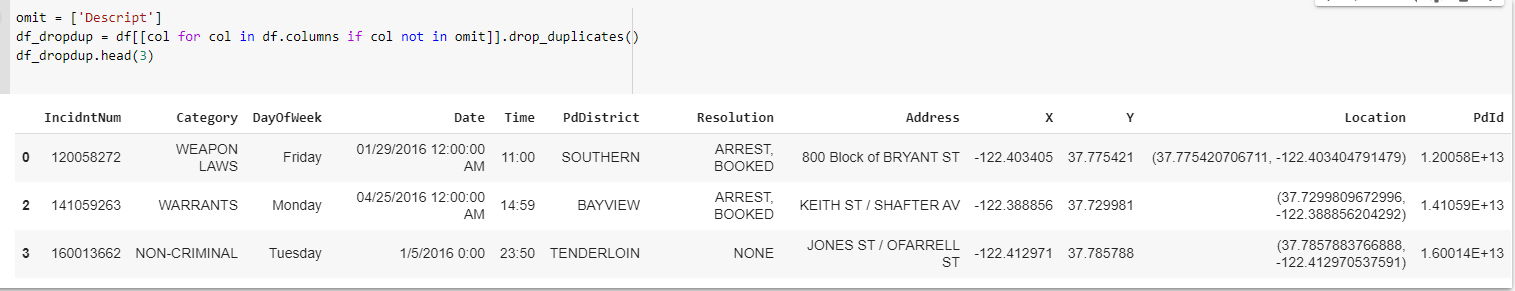
ทำการตรวจสอบข้อมูลของแต่ละตัวแปรที่มีค่าว่างเป็นค่าNan และตรวจสอบค่าที่ไม่ซ้ำกัน โดยพบว่ามี จำนวนเคส Incidntnum อยู่116699 แต่มีจำนวนแถวข้อมูลอยู่150500 ทำให้รู้ว่ามีข้อมูลที่ซ้ำกันอยู่



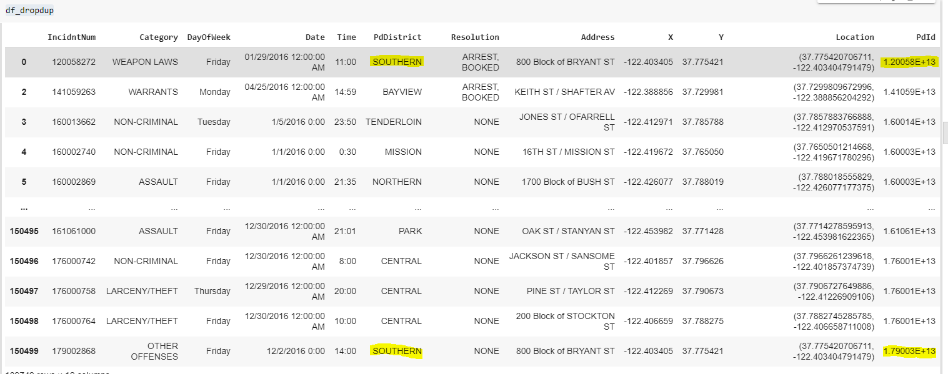
**3. Data Preparation**

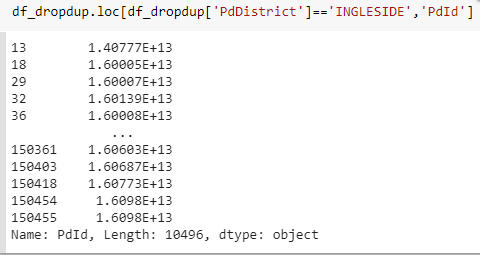
ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (raw data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (data cleaning)

จากขั้นตอน Data Understanding มีข้อมูลที่เป็นค่าNanจึงต้องนำออก ทำให้แถวข้อมูลลดลง1แถว

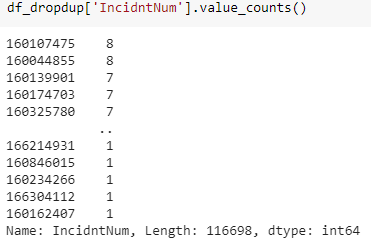
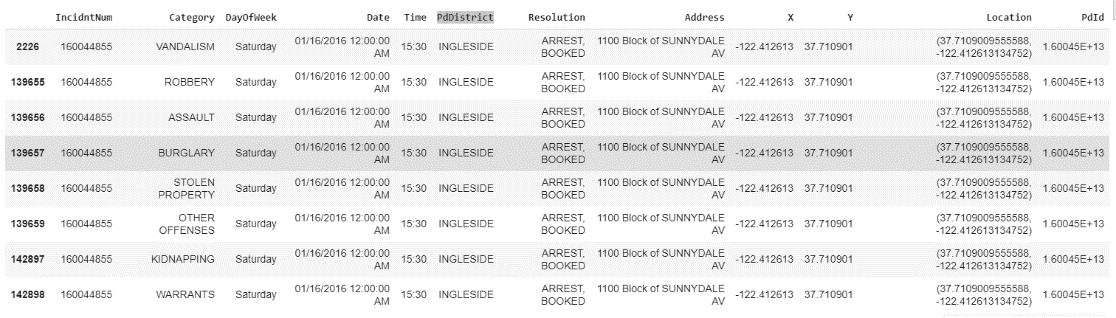


ทำให้เราทราบว่า 1ประเภทของคอลัมน์ Category มีได้หลาย Descript เราจึงไม่ใช้คอลัมน์ Descript โดยการเอาคอลัมน์ Descriptออก และใช้คำสั่ง drop\_duplicates() เพื่อลดความซ้ำซ้อนของข้อมูล

เช็คคอลัมน์ PdDistrict จากรูปทำให้ทราบว่า PdDistrict 1ค่า มี PdId หลายค่าเราจึงไม่ใช้ข้อมูลประเภทนี้

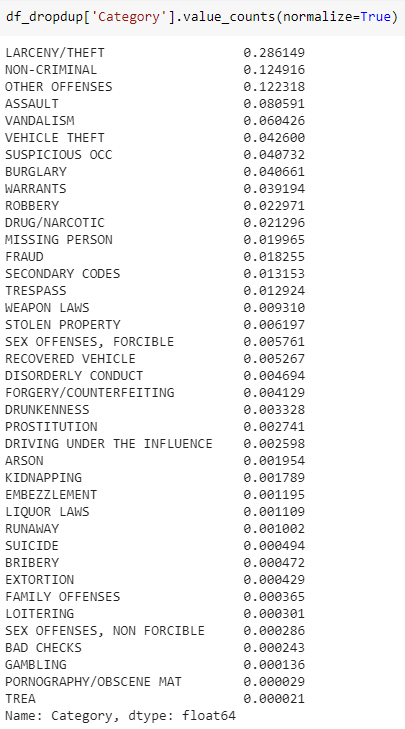
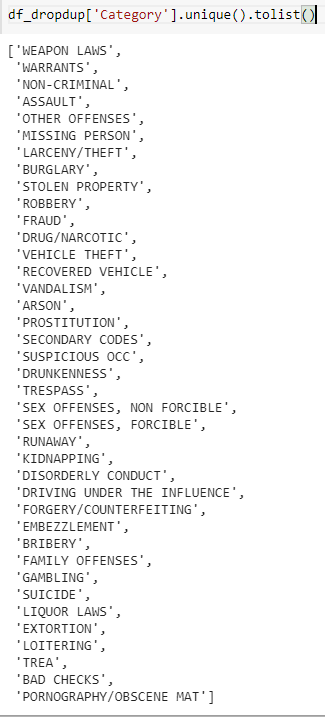


เช็คจำนวน IncidntNum โดยยกตัวอย่างค่า IncidntNum = 160044855 ทำให้เราทราบว่า 1 IncidntNum มีได้หลายประเภท แต่เราต้องการที่จะทำนาย Category โดยเลือกที่จะไม่สนใจคอลัมน์ IncidntNum



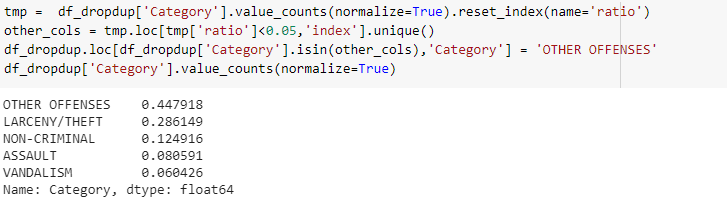
**Outlier**

ดูอัตราส่วนของแต่ละประเภทอาชญากรรม และ แปลงArrayให้เป็นlist

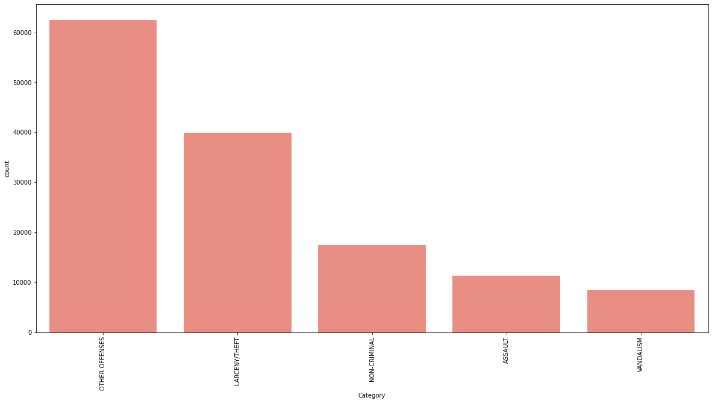
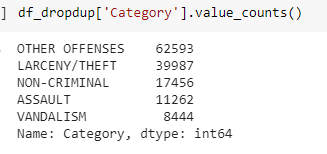


บางอัตราส่วนน้อยกว่า5%ทำให้โมเดลอาจจะไม่สามารถเรียนรู้หรือเรียนรู้ได้ยาก เราจึงนำอัตราส่วนของ Category ที่น้อยกว่า5% จากข้อมูลทั้งหมดไปรวมไว้ใน OTHER OFFENSESS

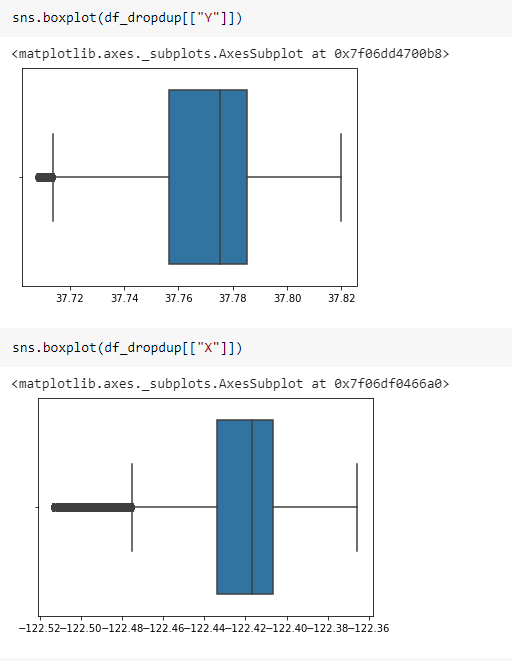
อัตราส่วนที่แสดงดังภาพคืออัตราส่วนที่มากกว่า5%



จะได้จำนวนข้อมูลที่อยู่ในประเภท Categoryต่างๆทั้งหมด5ประเภท ที่จะนำไปวิเคราะห์ในโมเดล

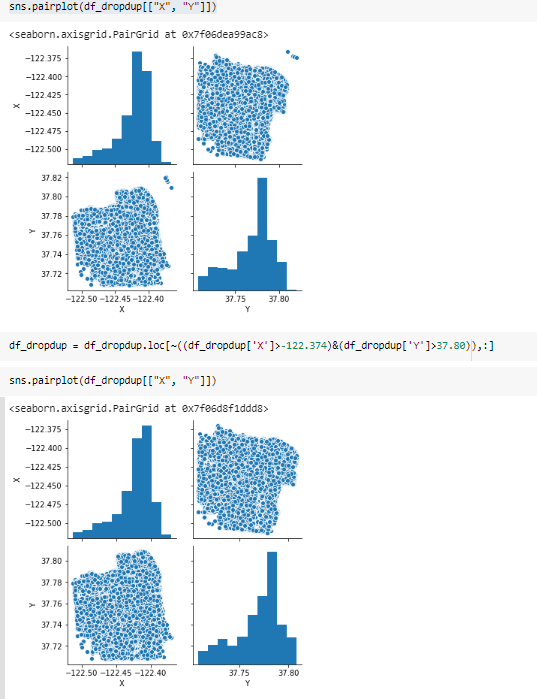


เราจึงนำข้อมูลมาพลอต Boxplot เพื่อดูการกระจายตัวของข้อมูล จากรูปไม่มีoutlierที่เห็นได้ชัด



ทำการพลอต pairplot อีกรอบ เพื่อเช็คความสัมพันธ์ระหว่างข้อมูล2คอลัมน์ ว่ามีกระจายตัวผิดปกติหรือไม่

จากรูป มีค่าการกระจายตัวบางค่าที่หลุดออกมาจากนอกกลุ่ม จึงทำการตัดข้อมูลส่วนนั้นออก



**Transform data**

คือการแปลงให้อยู่ในรูปที่โมเดลสามารถนำไปเรียนรู้ได้



จากรูป ข้อมูลอยู่ในtypeที่ไม่สามารถนำไปเรียนรู้ในโมเดลได้

เราจึงนำข้อมูลมาแปลงแปลงให้อยู่ในรูปที่โมเดลสามารถนำไปเรียนรู้ได้

DayOfYear คือ 1วันใน1ปี

Year คือ ปี

DayOfWeek คือ1วันใน1สัปดาห์

WeekOfYearคือสัปดาห์ใน1ปี

DayOfMonth คือ 1วันใน 1เดือน

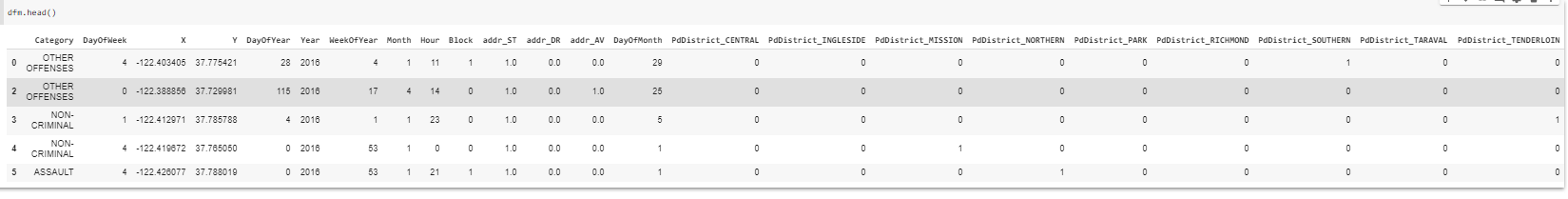
Month คือเดือน และต้องการแสดงข้อมูลเป็นชั่วโมง

St DR AV คือประเภทถนนที่ตั้ง

นำอา PdDistrictและประเภทถนนที่ตั้ง มาทำ one hot คือการนำ categorical

มาทำ Encoding ให้อยู่ในรูปแบบ binary

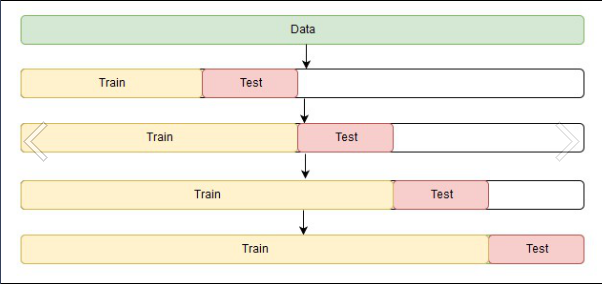
และเลือกคอลัมน์



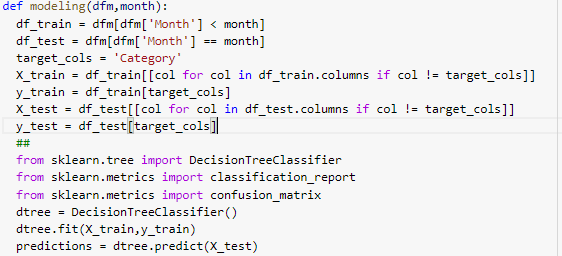
**4. Modeling**

ขั้นตอนนี้จะเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิง โดยวิธีการ การจำแนกประเภทข้อมูล (Classification) เทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนที่ **(3) Data Preparation** เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย ตัวอย่างเทคนิคในการวิเคราะห์ข้อมูลต่างๆ

ลำดับแรกแบ่งข้อมูลเป็นชุด train และ test ด้วยวิธีแบบ Time series Cross-ValidationโดยเราจะทำการเลือกTest เดือน9-12โดยใช้ข้อมูลชุดเทสไม่เหมือนกัน โดยสมมุตว่าถ้าเราจะทดสอบเดือนนั้น จะให้เดือนก่อนหน้าที่เหลือเป็นชุดเทรน เพื่อที่จำลองสถานการณ์จริงว่าข้อมูลที่ใช้ในเดือนนั้นสามารถเรียนรู้ได้จากข้อมูลในเดือนก่อนหน้า และใช้โมเดล Decision Tree

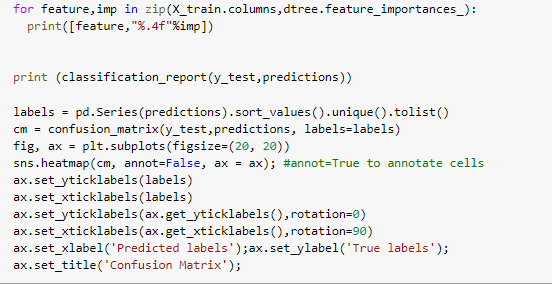


### Decision Tree คือ เป็นโมเดลทางคณิตศาสตร์ที่จะใช้ทำนายประเภทของวัตถุ โดยการพิจารณาจากลักษณะของวัตถุและใช้ต้นไม้ตัดสินใจในการสร้างโมเดลที่พยากรณ์ได้ ซึ่งจะเชื่อมโยงข้อมูลสังเกตการณ์เข้ากับข้อมูลปลายทาง การทำงานของต้นไม้การตัดสินใจจะทำการจัดกลุ่ม (classify) ชุดข้อมูลนำเข้าในแต่ละกรณี (Instance) แต่ละบัพ (node) ของต้นไม้การตัดสินใจคือตัวแปร ([attribute](https://www.mindphp.com/%E0%B8%84%E0%B8%B9%E0%B9%88%E0%B8%A1%E0%B8%B7%E0%B8%AD/73-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3/5050-what-is-attribute.html)) ต่างๆของชุดข้อมูล



**5. Evaluation**

ในขั้นตอนนี้เราจะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมน์นิงแล้วแต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไปก็จะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือ มีความน่าเชื่อถือมากน้อยเพียงใด



สี คือ แสดงจำนวนการทำนาย

True label คือ ผลจริง

Predicted labels คือ ผลที่เราทำนาย

Precision คือ modelสามารถตรวจจับ class ที่เราสนใจได้แค่ไหน

Recall คือ ดูความแม่นยำของ model ว่า model ทำนายถูกแค่ไหน

Accuracy คือ modelเราทำนายถูกทั้งหมดกี่ %

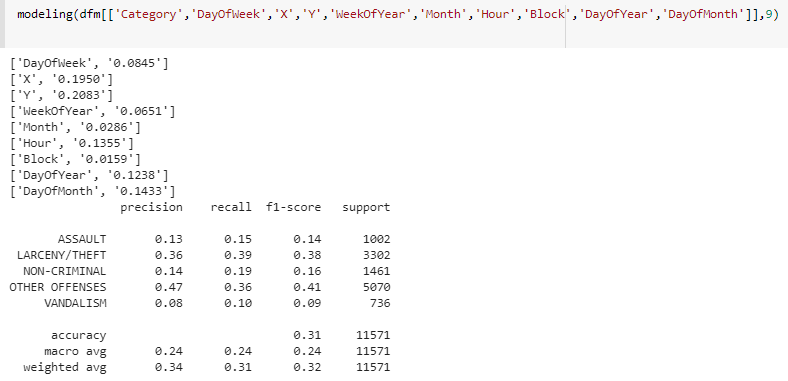
f1-score คือ วัดความสามารถของ model การหาค่าเฉลี่ยของทั้ง precision และ recall

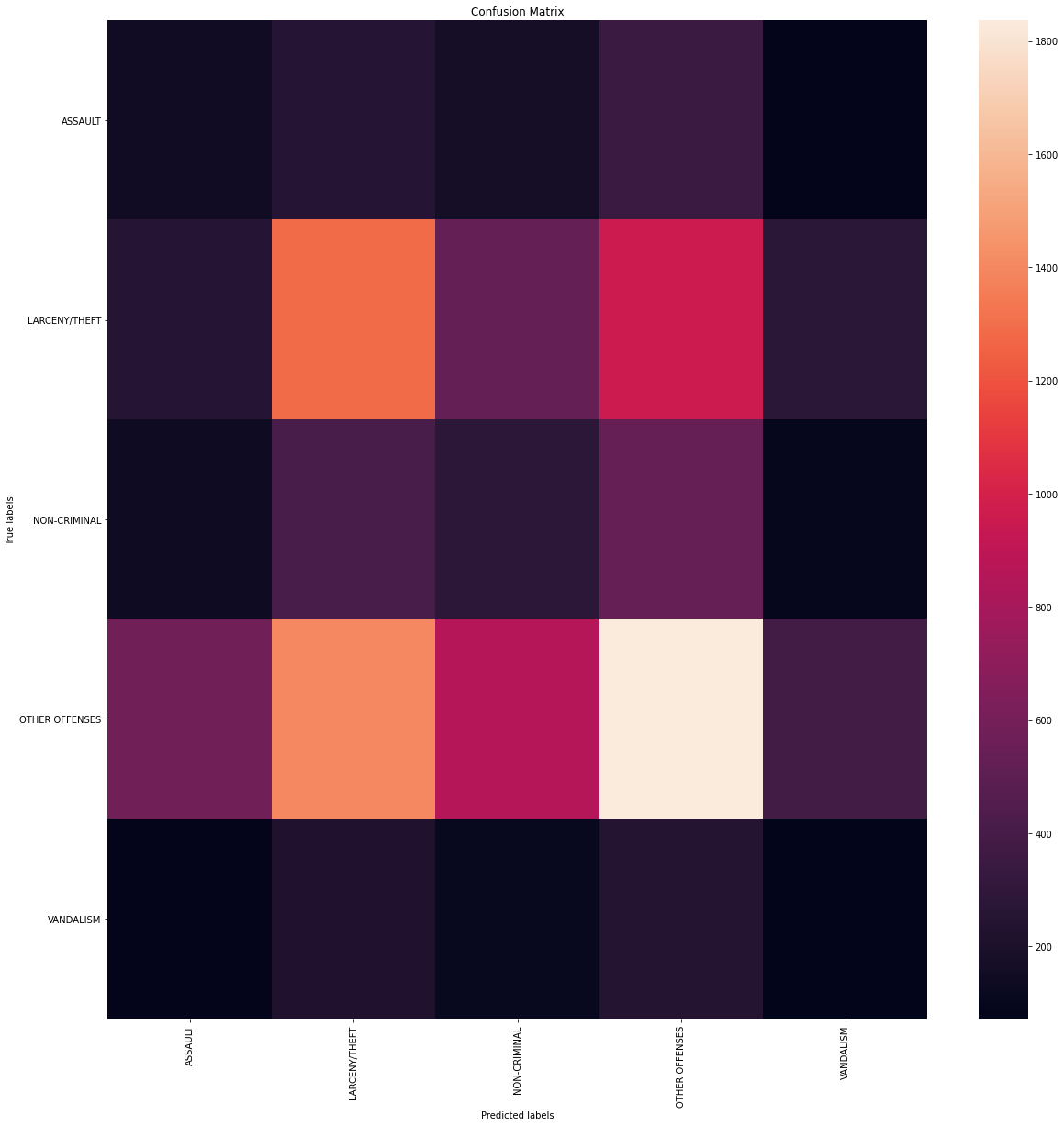
Support คือ จำนวนนวนข้อมูลที่อยู่ในคลาสนั้นในTest-set

macro avg คือ ผลเฉลี่ยของตัววัดนั้นจากทุกคลาส

weighted avg คือ ผลเฉลี่ยของตัววัดนั้นจากทุกคลาสโดยคำนึงถึงจำนวนข้อมูลของแต่ละคลาส

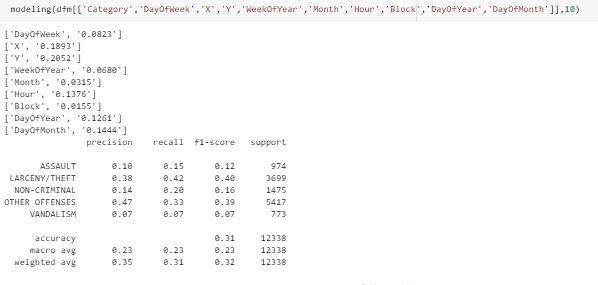
**ทดสอบเดือนที่ 9**

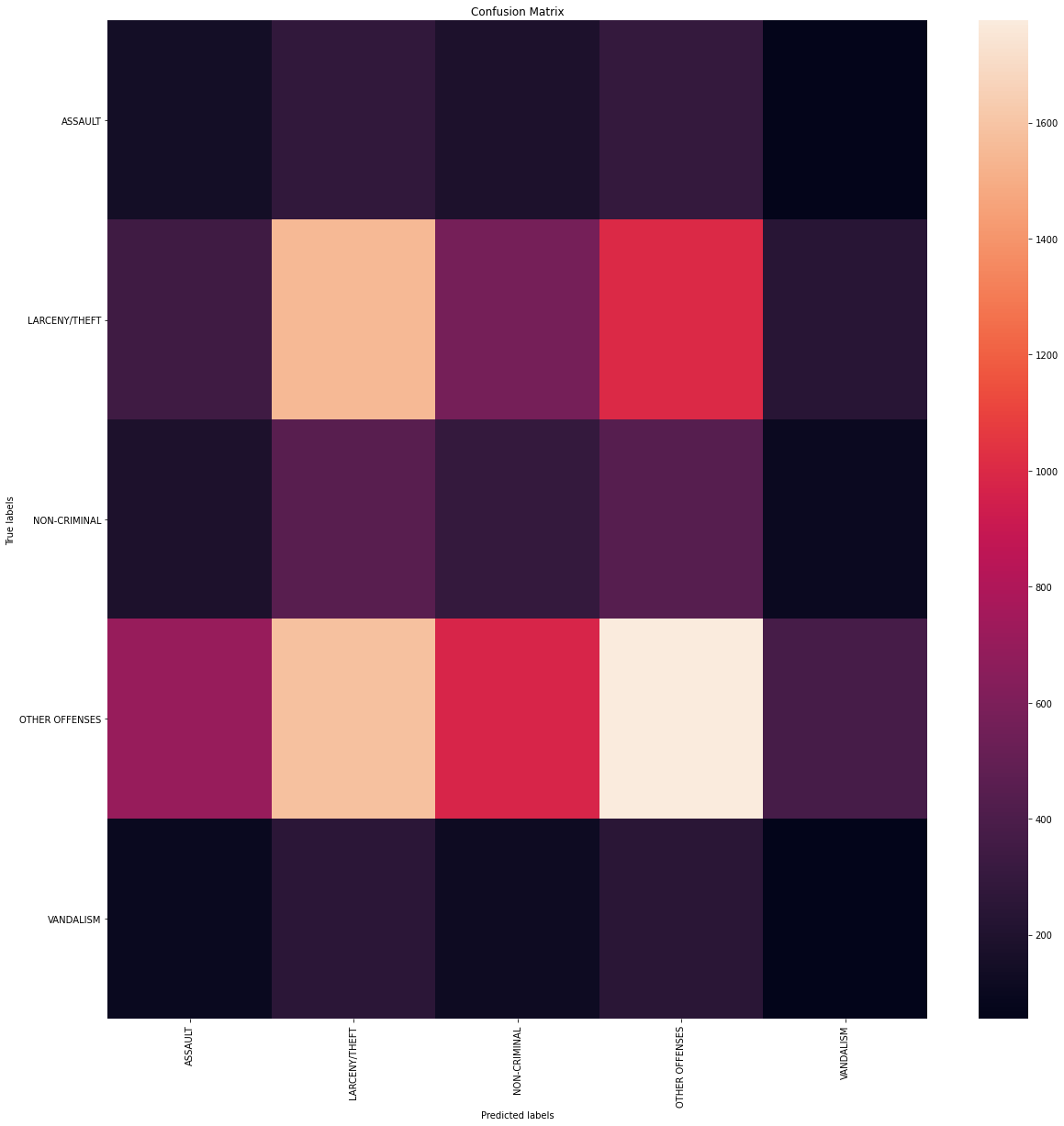




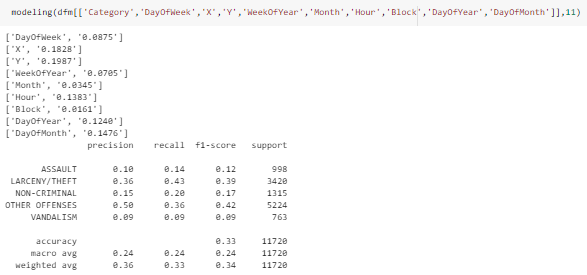
ยกตัวอย่าง เช่น พิจารณาแกน x Predicted labels ช่อง OTHER OFFENSES และ แกนy True labels ช่อง OTHER OFFENSES สังเกตได้ว่ามีสีขาวเด่นชัดขึ้นมา หมายความว่า ทำนายตรงกันมากที่สุดและถูกต้อง และที่ทำนายผิดส่วนใหญ่ทำนายเป็นLARCENY/THEFT ซึ่งแสดงเป็นสีโทนส้มอ่อน หมายความว่า LARCENY/THEFT ทำนายได้ดีรองลงมา ซึ่งถ้าไปดูแกน True labels ช่อง LARCENY/THEFT และไปดูแกน Predicted labels ช่อง OTHER OFFENSES จะเป็นสีชมพู หมายความว่า ความสามารถในการทำนายโมเดลผิดเป็น OTHER OFFENSES เพราะ สังเกตุช่องLARCENY/THEFT ทั้งแกน Predicted labels และ True labels มีการทำนายตรงกันมากที่สุดและถูกต้อง สรุปได้ว่า LARCENY/THEFT และ OTHER OFFENSES ทำนายสลับกันอยู่ แต่ก็ยังสามารถทำนายได้ดีกว่าคลาสอื่น ๆ

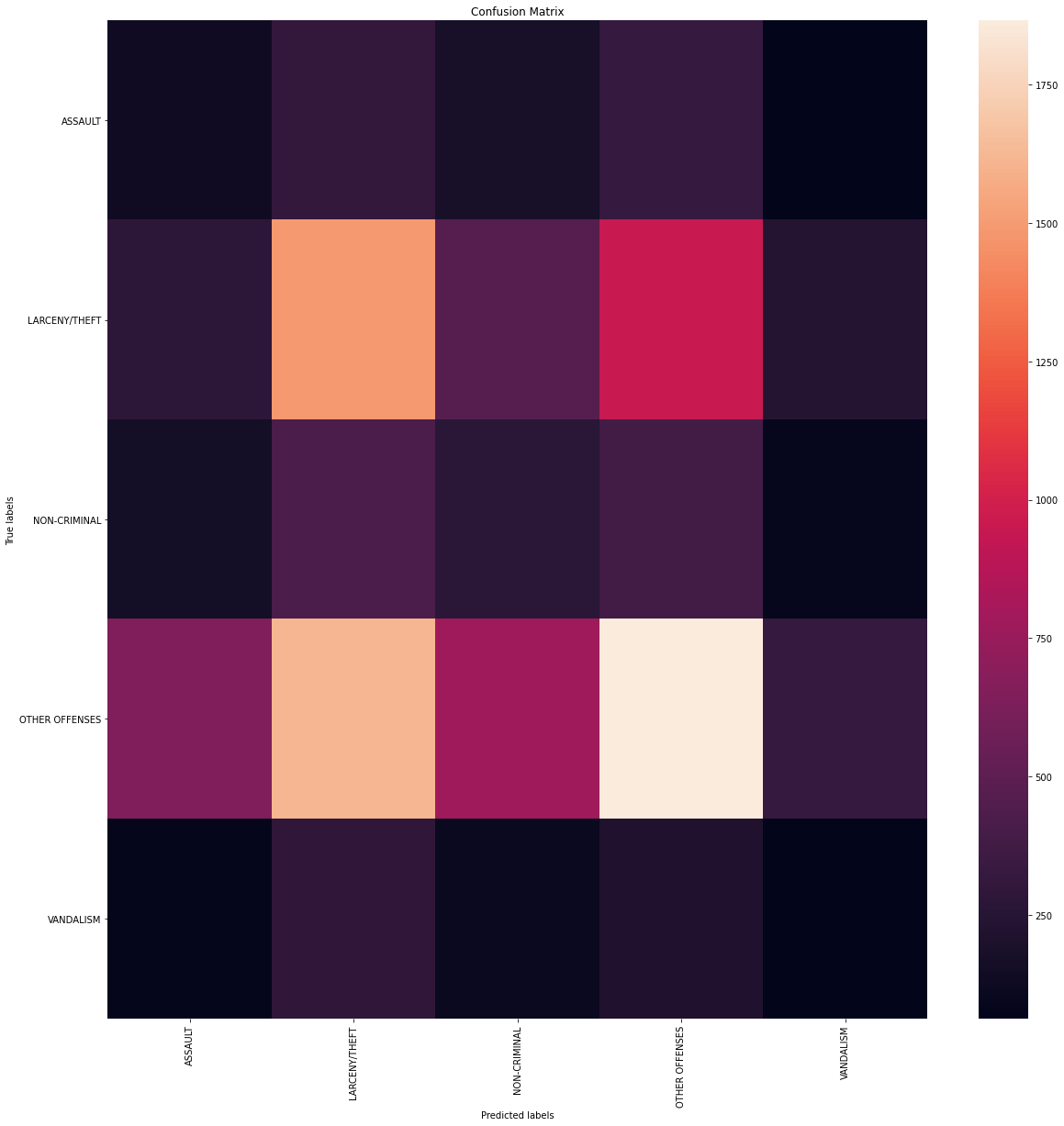
**ทดสอบเดือนที่ 10**

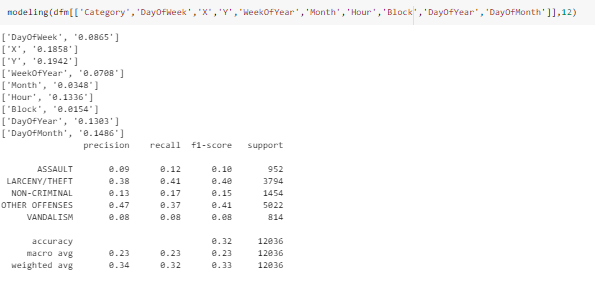


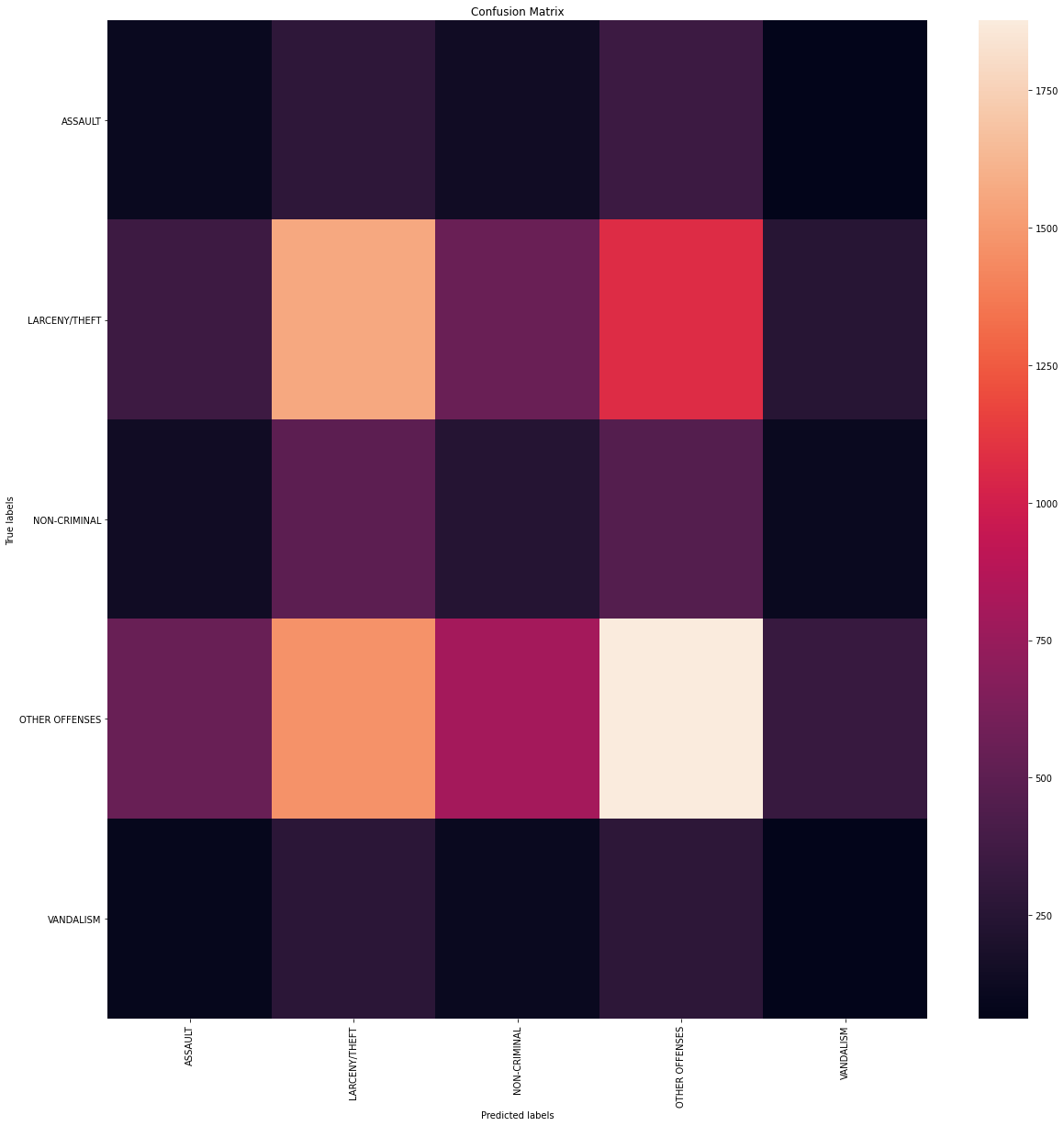


**ทดสอบเดือนที่ 11**





**ทดสอบเดือนที่ 12**



เมื่อเรานำทั้ง4เดือนมาเฉลี่ยกันจะได้ค่าดังนี้

precision recall f1-score support

ASSAULT 0.105 0.14 0.12 981.5

LARCENY/THEFT 0.37 0.4125 0.3925 3553.75

NON-CRIMINAL 0.14 0.19 0.16 1426.25

OTHER OFFENSES 0.4775 0.355 0.4075 5183.25

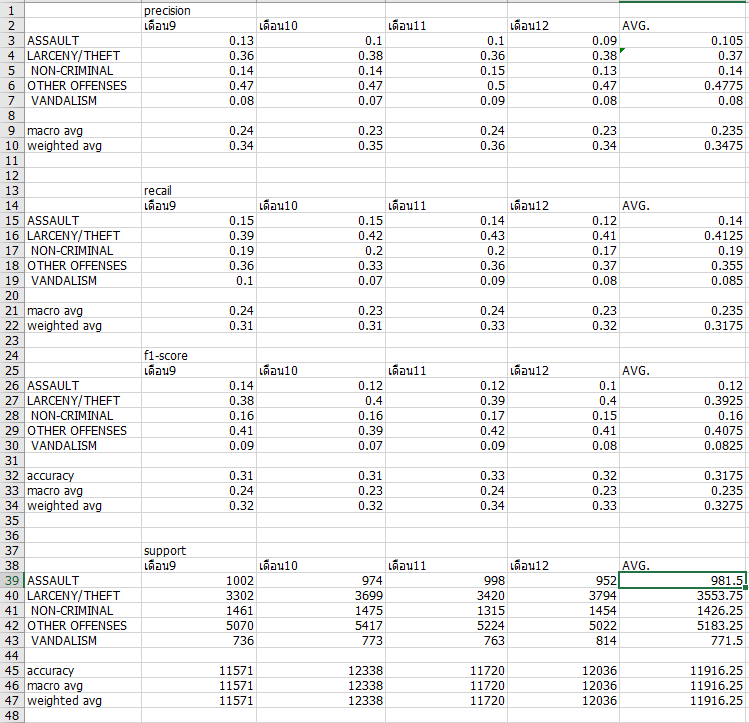
VANDALISM 0.08 0.085 0.0825 771.5

accuracy 0.3175 11916.25

macro avg 0.235 0.235 0.235 11916.25

weighted avg 0.3475 0.3175 0.3275 11916.25

หมายเหตุ:หาค่าเฉลี่ยทั้ง4เดือนจาก Excel



สรุป model ของเราทำนายถูกต้อง 31.75% โดยเฉลี่ยจากการวัดข้อมูล4เดือน โดยที่หากเราวัดโดยไม่คำนึงถึงจำนวนข้อมูลของแต่ละคลาสโดยเฉลี่ยจากทุกคลาสแล้วจะมีf1-score 23.5% ถ้าคำนึงจำนวนข้อมูลของแต่ละคลาสโดยเฉลี่ยจากทุกคลาสแล้วจะมีf1-score 32.75% โมเดลของเราสามารถทำนายคลาสที่มีค่าsupport เยอะได้ดี แต่อาจจะทำนายคลาสที่มีค่าsupport น้อยได้ไม่ดีและคลาสที่ทำนายได้ดีคือ อาชญากรรมประเภท LARCENY/THEFT และ OTHER OFFENSES

**6. Deployment**

จากโมเดล Decision Tree สามารถที่จะนำไปจำแนกประเภทอาชญากรรมโดยอัตโนมัติ โดยสามารถทำนายได้ 5 ประเภทคือ ASSAULT, LARCENY/THEFT, NON-CRIMINAL,OTHER OFFENSES, VANDALISM ซึ่งผลทำนาย สามารถส่งเรื่องไปให้กับเจ้าหน้าที่ที่เกี่ยวข้องกับอาชญากรรมประเภทนั้นๆได้โดยอัตโนมัติ เพื่อลดขึ้นตอนในการตรวจสอบประเภทของอาชญากรรม และเพิ่มความรวดเร็วในการจัดการคดีที่มีจำนวนมากได้

**Link ของ video ที่นำเสนอโครงงาน**

<https://www.youtube.com/watch?v=6J1drTF_qcQ>(ปรับความละเอียดหน้าจอสูงสุดเพื่อความคมชัด)