**Activity 5**

1. **Start & Visualize**

ก็จะลองดูข้อมูล shape info describe แล้วก็ลองเอาผลลัพธ์ (Survived) มาแสดงเป็น Pie chart และเอาข้อมูลที่เป็น numeric เช่น Pclass, Age, SibSp, Parch, Fare มาแสดงด้วย Histogram ส่วนข้อมูลที่เป็น Categorical เช่น Sex, Ticket, Cabin, Embarked ก็จะแสดงด้วย Bar chart ยกเว้น PassengerId, Name ที่ไม่ได้แสดงเพราะรู้อยู่แล้วว่า unique

1. **Narrowing down columns**

ก็จะ drop พวกที่ไม่เกี่ยวข้องและ unique เช่น PassengerId, Name, Ticket, Cabin นอกจากนั้นตัว Cabin ยังมี Missing value เยอะมาก เลยยิ่งน่า drop เข้าไปใหญ่ ซึ่งจะพบว่า Age และ Embarked ก็มี missing value เหมือนกันแต่ไม่เยอะมาก ซึ่งจะจัดการในขั้นตอนถัดไป

1. **Impute missing values**

ในที่นี้จะมี Age และ Embarked ที่มี missing value ซึ่ง Age เป็น numeric data จึงเลือกใช้ mean ในการเติม missing value ส่วน Embarked เป็น categorical data จึงเลือกใช้ mode ในการเติม

1. **Categorical to numerical variables**

ในส่วนนี้เราจะมี data ที่เป็น categorical อยู่ 2 อันคือ ‘Sex’ และ ‘Embarked’ ในส่วนของ ‘Sex’ นั้นมีค่าแค่ 2 แบบคือ male หรือ female ดังนั้นจึงเปลี่ยน male เป็น 0 และ female เป็น 1 ในส่วนของ ‘Embarked’ มีอยู่ 3 ค่า จึงใช้วิธี dummy code

1. **Outliers**

ข้อมูลส่วนใหญ่ค่อนข้าง normalize แล้ว จะมีแค่ Age และ Fare ที่มีการกระจายของข้อมูลอยู่ ซึ่งตัว Age นั้นค่อนข้างเป็น Normal form แต่จะมีข้อมูลที่เกินขอบเป็น Outlier ประมาณ 7% จึงทำการลบข้อมูลส่วนนั้นทิ้งไป ในส่วนของ Fare นั้นเนื่องจากเมื่อนำมา plot กราฟแล้วจะเห็นว่ากราฟนั้นเบ้ขวาอยู่ จึงทำการ take log เพื่อให้ตัวกราฟค่อนข้างเป็น normal form มากขึ้น ซึ่งก็ยังมี Outliers บางส่วนอยู่ จึงทำการลบ outliers ส่วนนั้นออกไป สุดท้ายจึงเหลือข้อมูลทั้งหมด 800 ชุด (จากเดิม 891)แล้วก็ export to .csv ก็เป้นอันเสร็จสมบูรณ์