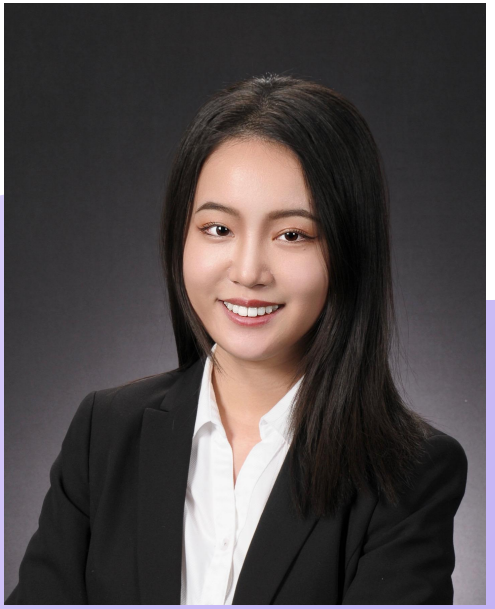
PE Location & Procedure Characteristics Exploratory Data Analysis

Katherine Tu

Presenter Introduction

- B.S Mathematics at University of California, Irvine
- M.S Analytics at University of Chicago



Agenda

1. Dataset Overview
2. Data Preprocessing Outliers
3. Data Preprocessing Missing Values
4. Methodology
5. Test Results

Dataset Overview

Sample Size

325 Patients in total

PE Location

Saddle
Bilateral
Unilateral

Time Characteristics

Procedure time
Catheter time
Fluoroscopy time

Data Preprocessing

1

Outliers

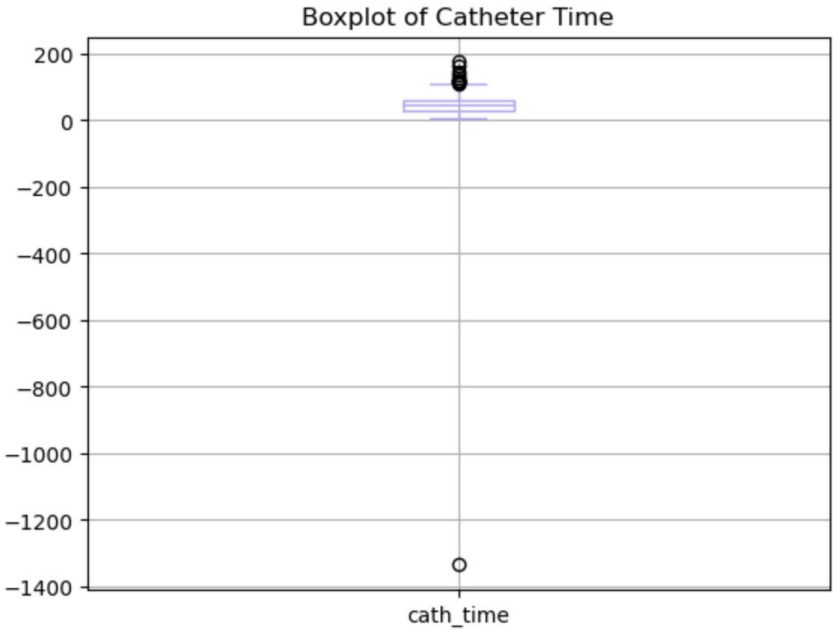
Found one extreme outlier in catheter time

2

Missing Value

Fill missing value with median or using linear regression

Outliers



- Cather time cannot be negative, maybe due to human data entry error
- Check whether there are errors in proc_time and fluoro_time for this patient
- Treat this outlier as null value and fill it using linear regression to avoid information loss

proc_time	cath_time	fluoro_time
105.0	-1335.0	31.3

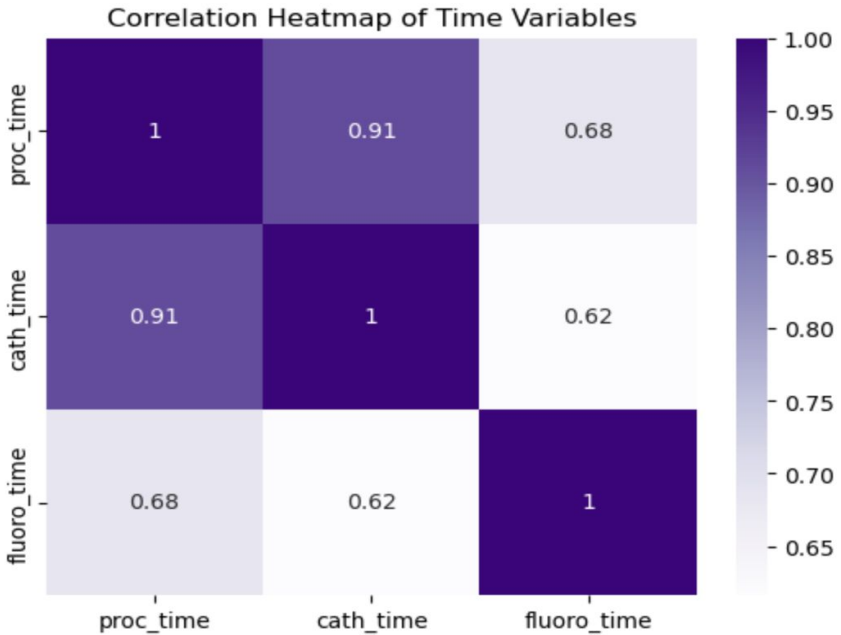
Missing Value

PATIENTID	0.000000
site	0.000000
age	0.000000
bmi	0.003115
CUPE	0.000000
location	0.000000
spesi	0.071651
rvlv_rat	0.065421
biomarkers	0.000000
dyspnea	0.000000
lytics_contra	0.000000
hx_pe	0.000000
hx_dvt	0.000000
cur_dvt	0.000000
proc_time	0.080997
cath_time	0.074766
fluoro_time	0.028037
T20	0.000000
Disks	0.000000
missing	0.000000

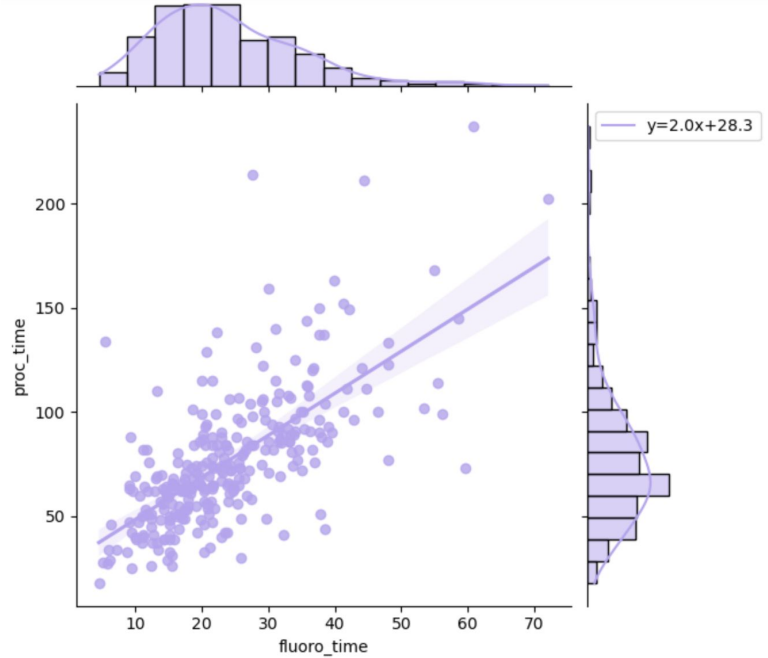
- Fill null values to avoid information loss
- Fill null values in bmi, spesi, and rvlv_rat columns with median
- Fill null values in procedure time, catheter time, and fluoroscopy time columns using linear regression

Missing Value

- Strong correlation between time variables



Missing Value — Procedure Time

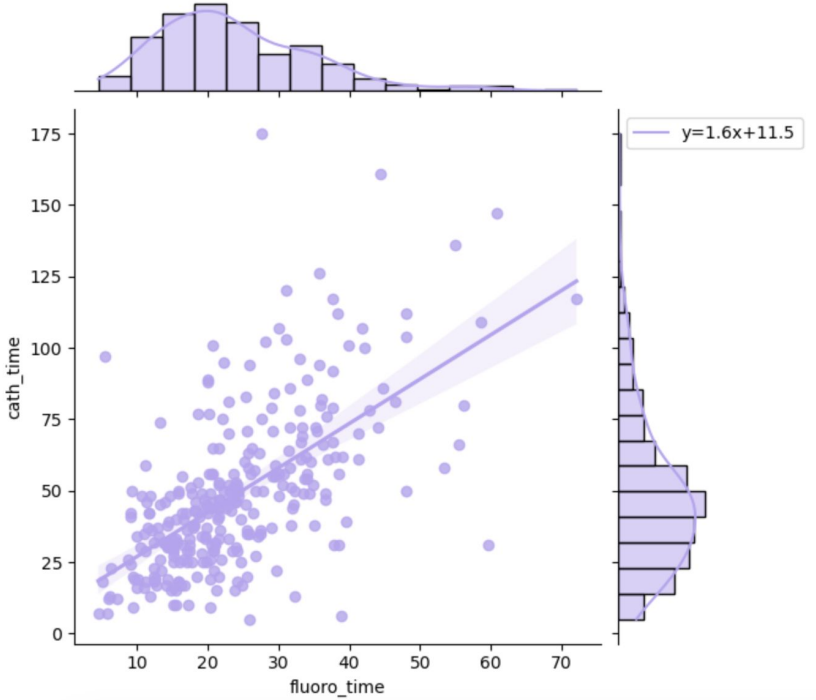


- Linear relationship between procedure time and fluoroscopy time
- Use original fluoroscopy time to generate procedure time and fill null values

	proc_time	cath_time	fluoro_time
3	NaN	47.0	19.5

	proc_time	cath_time	fluoro_time
3	67.3	47.0	19.5

Missing Value — Catheter Time

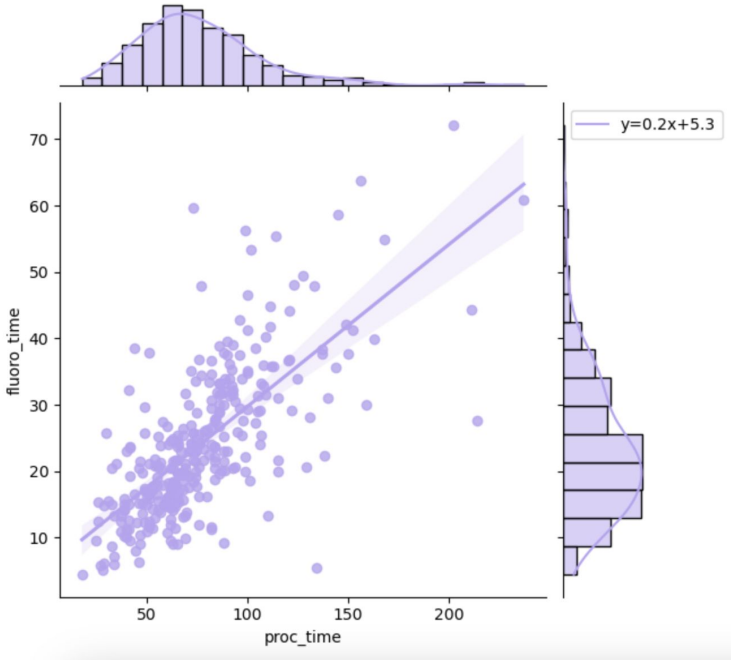


- Linear relationship between catheter time and fluoroscopy time
- Use original fluoroscopy time to generate catheter time and fill null values and one outlier

	proc_time	cath_time	fluoro_time
5	85.9	NaN	28.8

	proc_time	cath_time	fluoro_time
5	85.9	57.58	28.8

Missing Value — Fluoroscopy Time



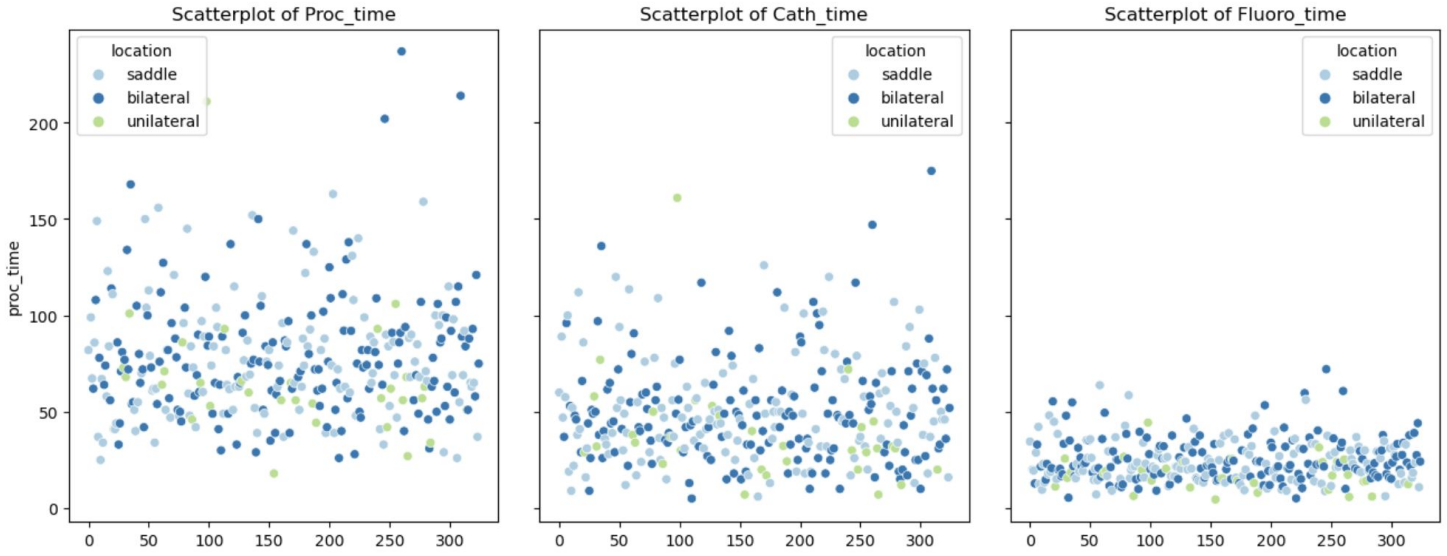
- Linear relationship between procedure time and fluoroscopy time
- Use original procedure time to generate fluoroscopy time and fill null values

	proc_time	cath_time	fluoro_time
38	51.0	36.0	NaN

	proc_time	cath_time	fluoro_time
38	51.0	36.0	15.5

Location & Time Variables

Question: Is Pulmonary Embolism location making a difference in procedure time, catheter time, and fluoroscopy time?



Methodology — Kruskal-Wallis Test

A nonparametric method for testing whether 3 or more groups are originated from the same distribution

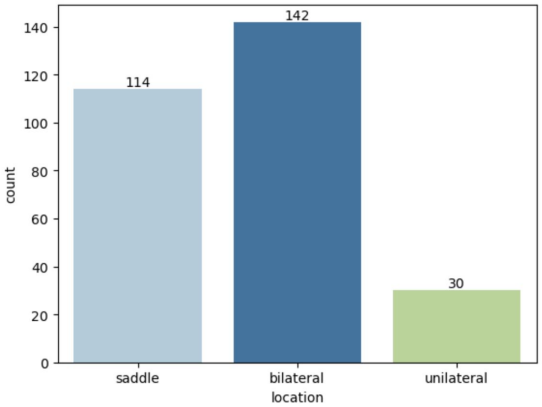
- Null Hypothesis: The 3 locations are not different in terms of time variables distribution/median
- Alternative Hypothesis: At least 1 location is different from the other 2 locations in terms of time variables distribution/median

- Post-hoc pairwise test — the Dunn's Test
  - Figure out the pairs of location having different distribution of time variables
  - Null hypothesis: there is no difference between groups
  - Alternate hypothesis: there is a difference between groups.

Methodology — Dataset Overview

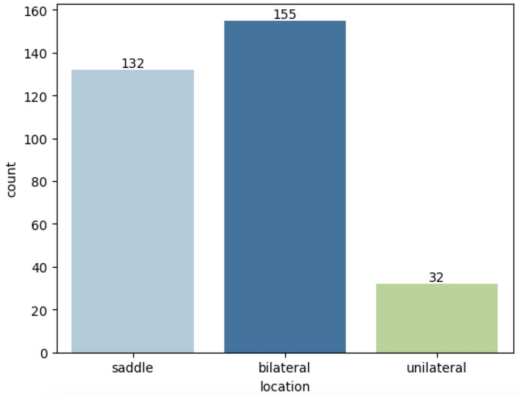
Data Without Filling NA

286 Patients Data



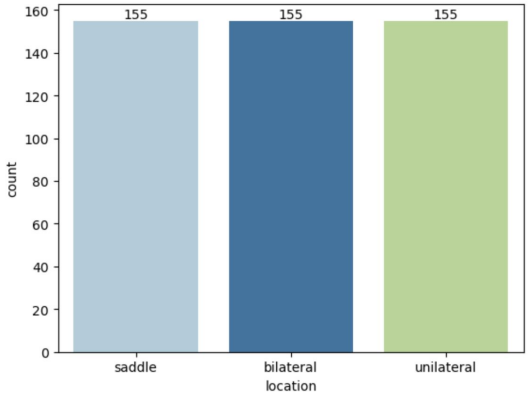
Data After Filling NA

319 Patients Data



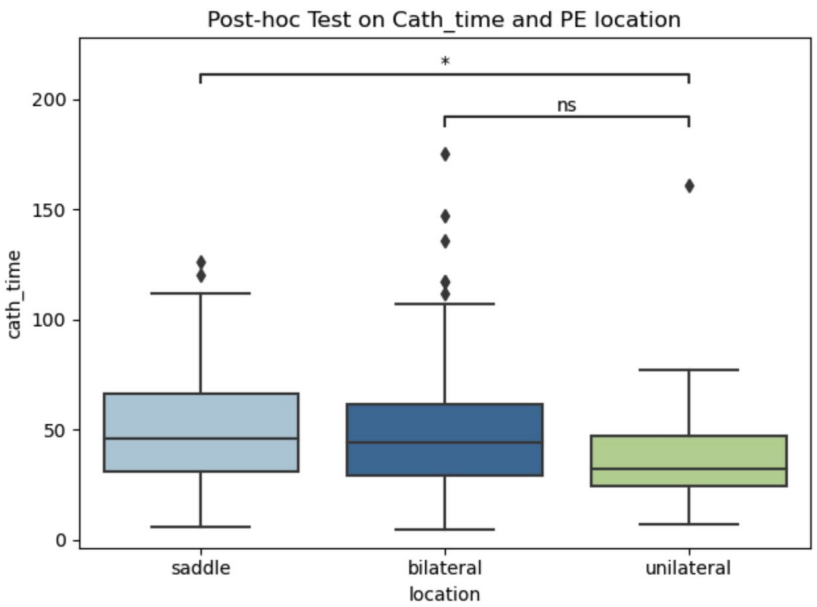
Data After Oversampling

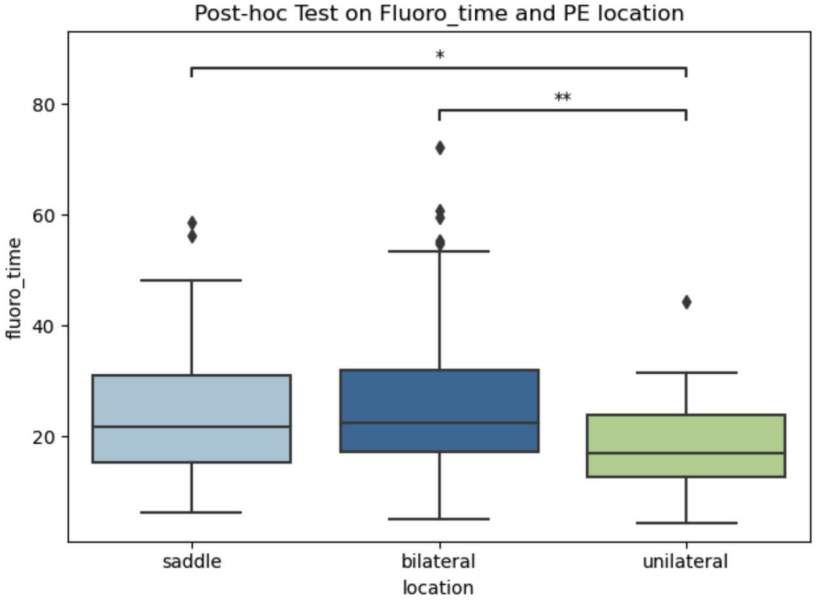
465 Patients Data



Test Result on Data Without Filling NA

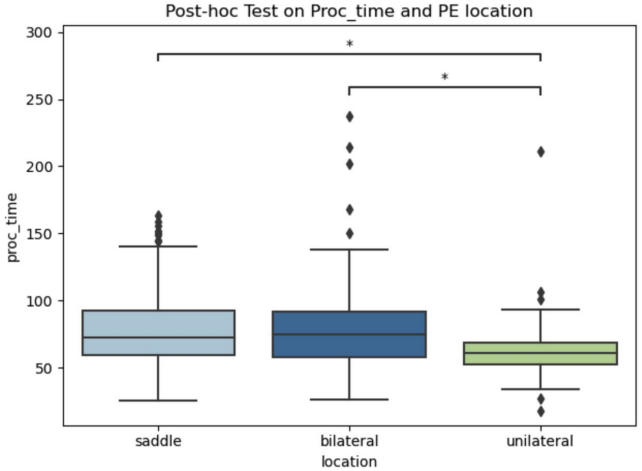
There is a statistically significant difference between the medians of cath_time and fluoro_time in these 3 PE locations

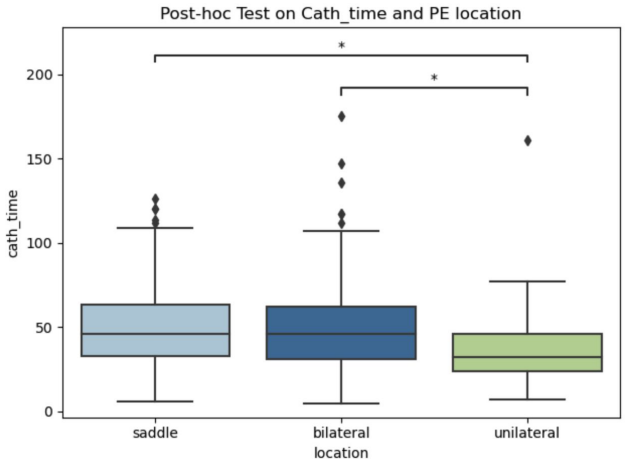


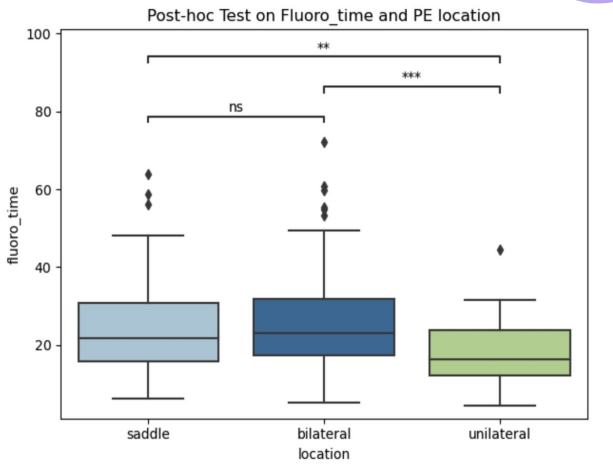


Test Result on Data After Filling NA

There is a statistically significant difference between the medians of proc_time, cath_time and fluoro_time in these 3 PE locations

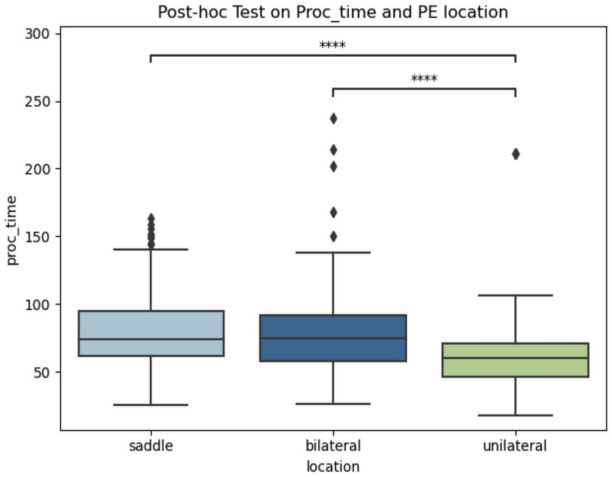


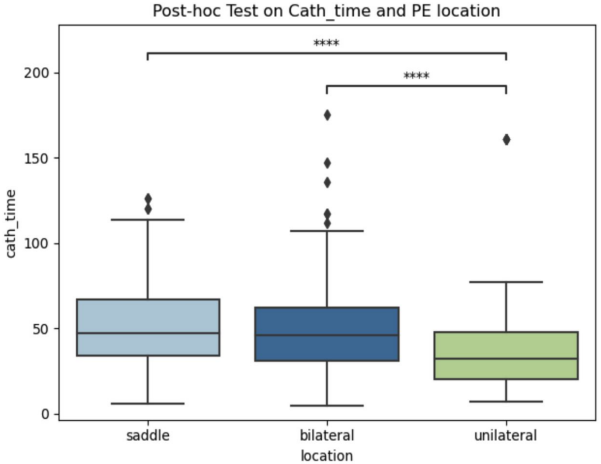


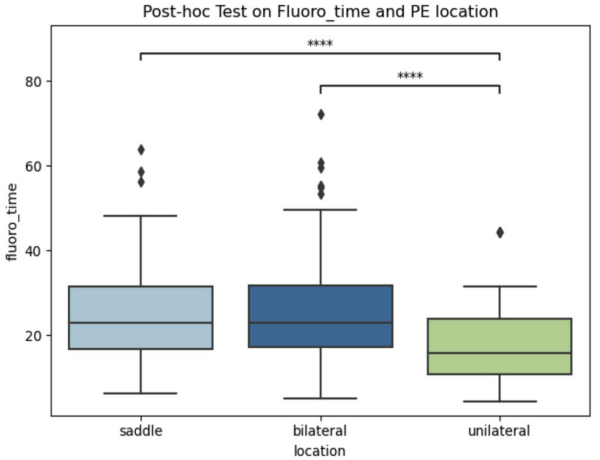


Test Result on Data After Oversampling

There is a statistically significant difference between the medians of proc_time, cath_time and fluoro_time in these 3 PE locations







Thank you for listening!