NYPD Shooting Incident Report

Jacky Luo

Introduction 💻

Purpose & Method

Purpose

Investigate relationships between victim and perpetrator of NYPD historic shooting incidents

Method

Exploratory Data Analysis in ***R***

Model validation with ***XGBoost***

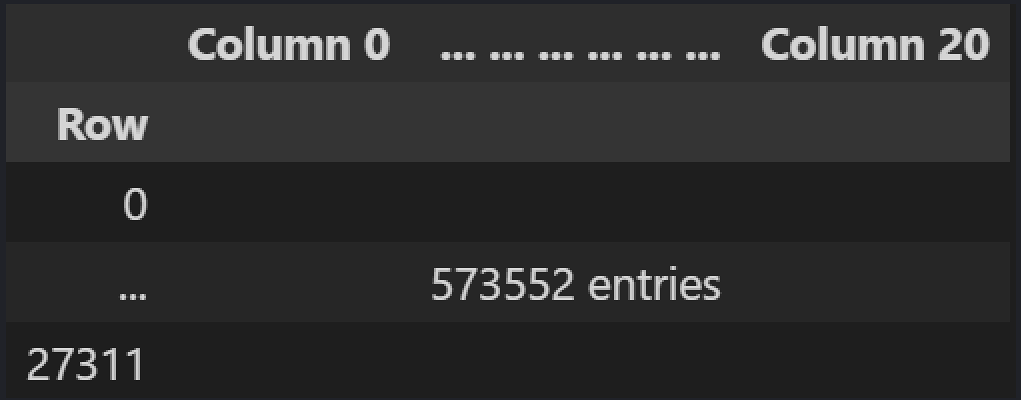
Data Investigation 🔍

Data Structure & Data Quality

Data Overview

21 Columns, **27312 Rows**, **573552 Entries**

12 String, 7 Numeric, 1 Date, 1 Boolean



	Column 0	... ...	Column 20
Row			
0			
...		573552 entries	
27311			

Date Time Columns

OCCUR_DATE (`chr`): Date in `MM/DD/YYYY` format

OCCUR_TIME (`time`): Time in `hh:mm` format

Incident Description Columns

INCIDENT_KEY (`dbl`): Unique incident identifier

BORO (`chr`): Geographic subdivision of NYC

LOC_OF_OCCUR_DESC (`chr`): Description of location

Incident Description Columns

PRECINCT (`dbl`): NYPD organizational subdivision

JURISDICTION_CODE (`dbl`): NYPD organizational subdivision

LOC_CLASSFCTN_DESC (`chr`): Description of location (street, vehicle, house, etc)

STATISTICAL_MURDER_FLAG (`lgl`): `TRUE` if victim died from incident

Perpetrator Description Columns

PERP_AGE_GROUP (`chr`): Binned age group of perpetrator

PERP_SEX (`chr`): Sex description of perpetrator (`M` , `F` , `U`)

PERP_RACE (`chr`): Race description of perpetrator

Victim Description Columns

VIC_AGE_GROUP (`chr`): Binned age group of victim

VIC_SEX (`chr`): Sex description of victim (`M` , `F` , `U`)

VIC_RACE (`chr`): Race description of victim

Latitude Longitude Columns

X_COORD_CD (`dbl`): FIPS3104 NY State X coord (ft)

Y_COORD_CD (`dbl`): FIPS3104 NY State Y coord (ft)

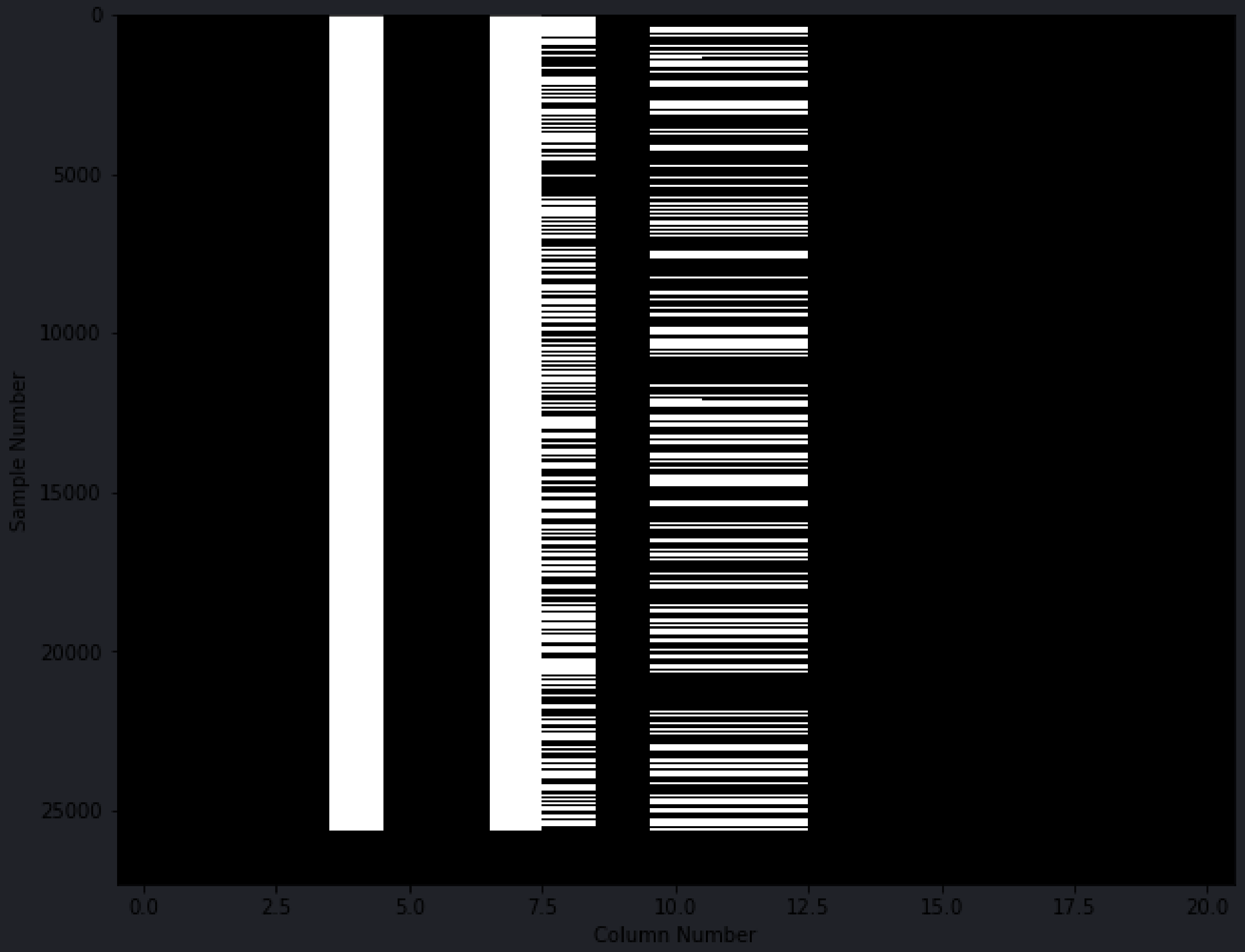
Latitude (`dbl`): EPSG 4326 decimal latitude coordinate

Longitude (`dbl`): EPSG 4326 decimal longitude coordinate

Lon_Lat (`chr`): `POINT (Long, Lat)` format longitude/latitude pair

Descriptive Statistics

Min, max, mean, median, IQR for each numeric column can be found in the written report



Missing Values

94165 total missing values, or **16.4%** of the dataset

Missing Value Columns

Column	Type	No Missing	% Missing
LOC_OF_OCCUR_DESC	chr	25596	93.7%
LOC_CLASSFCTN_DESC	chr	25596	93.7%
LOCATION_DESC	chr	14977	54.8%
PERP_AGE_GROUP	chr	9344	34.2%
PERP_SEX	chr	9310	34.1%
PERP_RACE	chr	9310	34.1%
Sum	-	94133	16.4%

Dropped Columns for Analysis

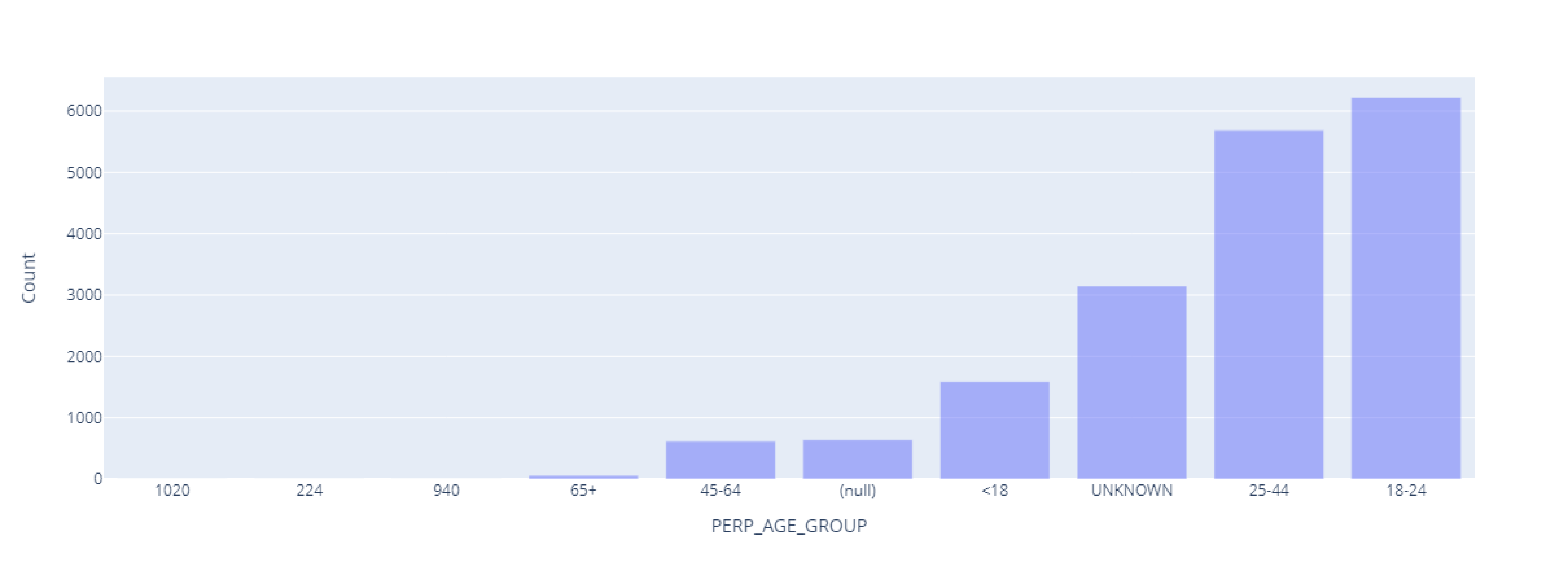
Latitude, Longitude, Lon_Lat duplicates, redundant

LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC,LOCATION_DESC too many missing values

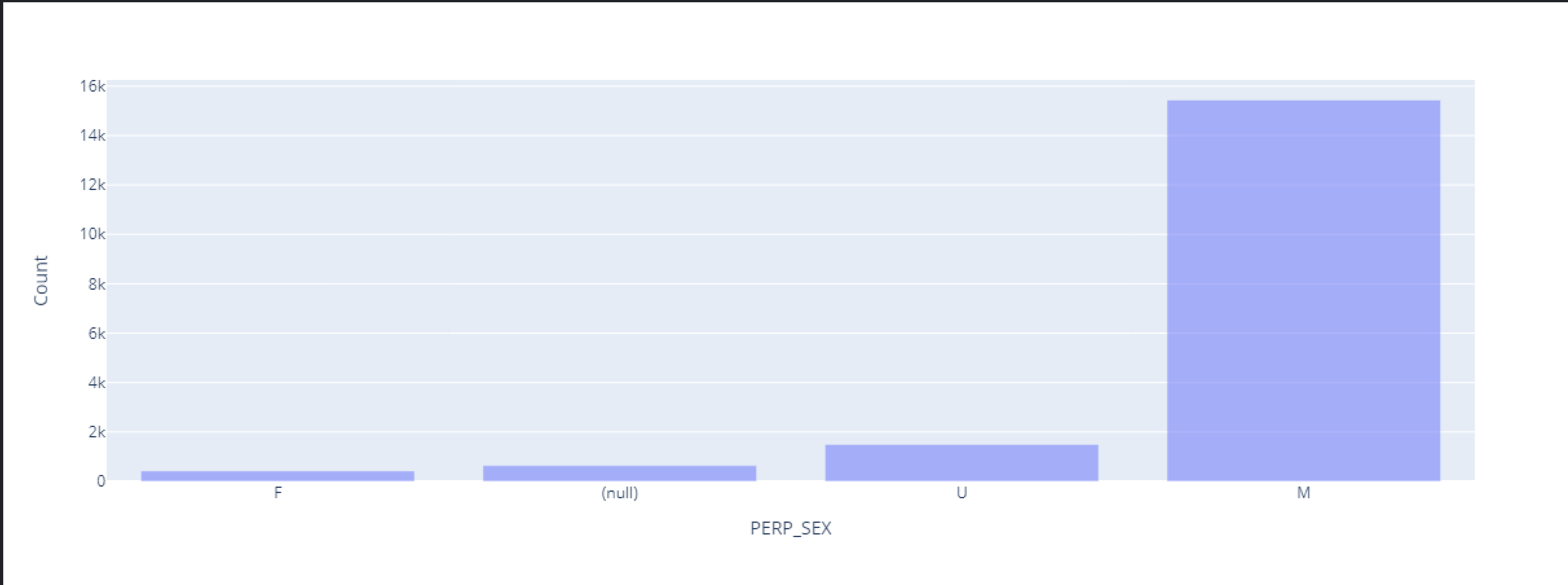
Analysis 📈

Demographics EDA

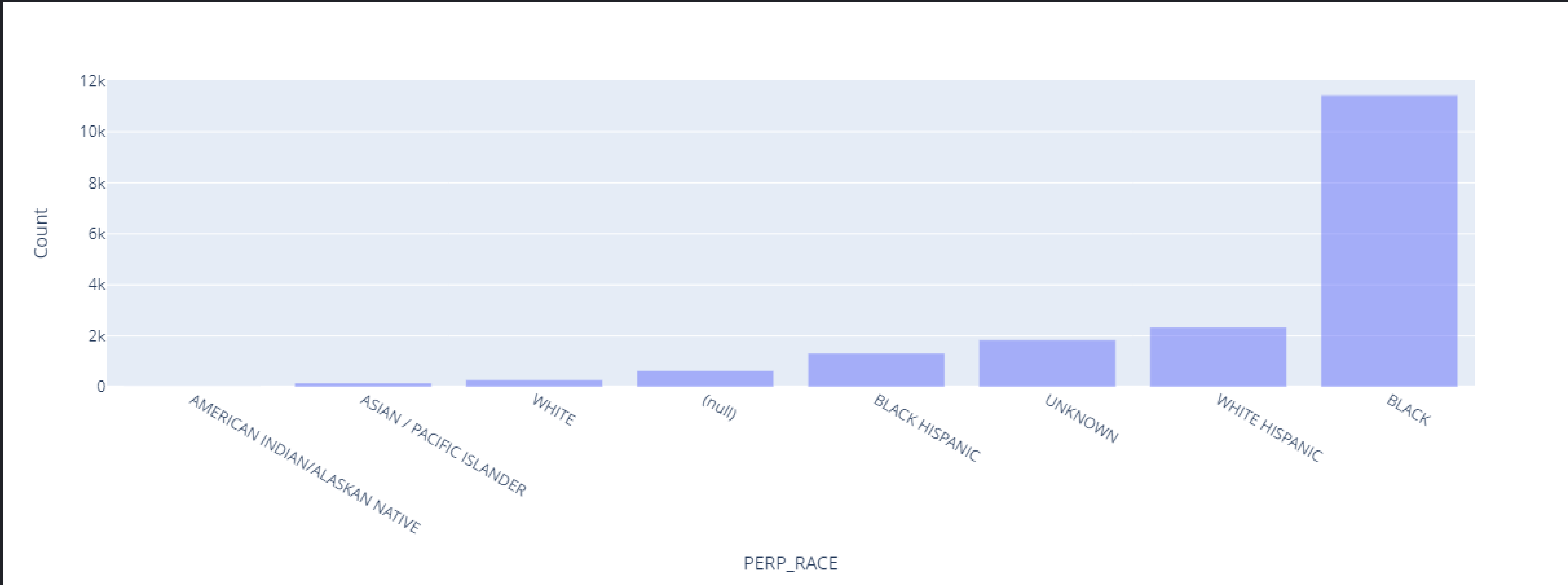
Perp Age Group



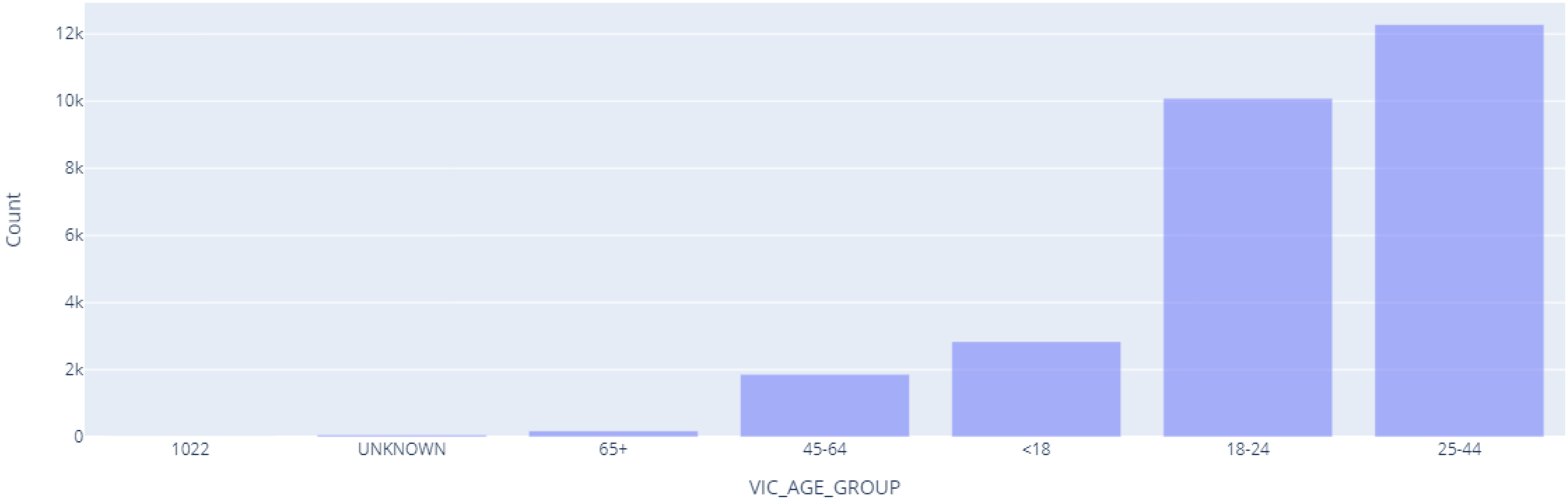
Perp Sex



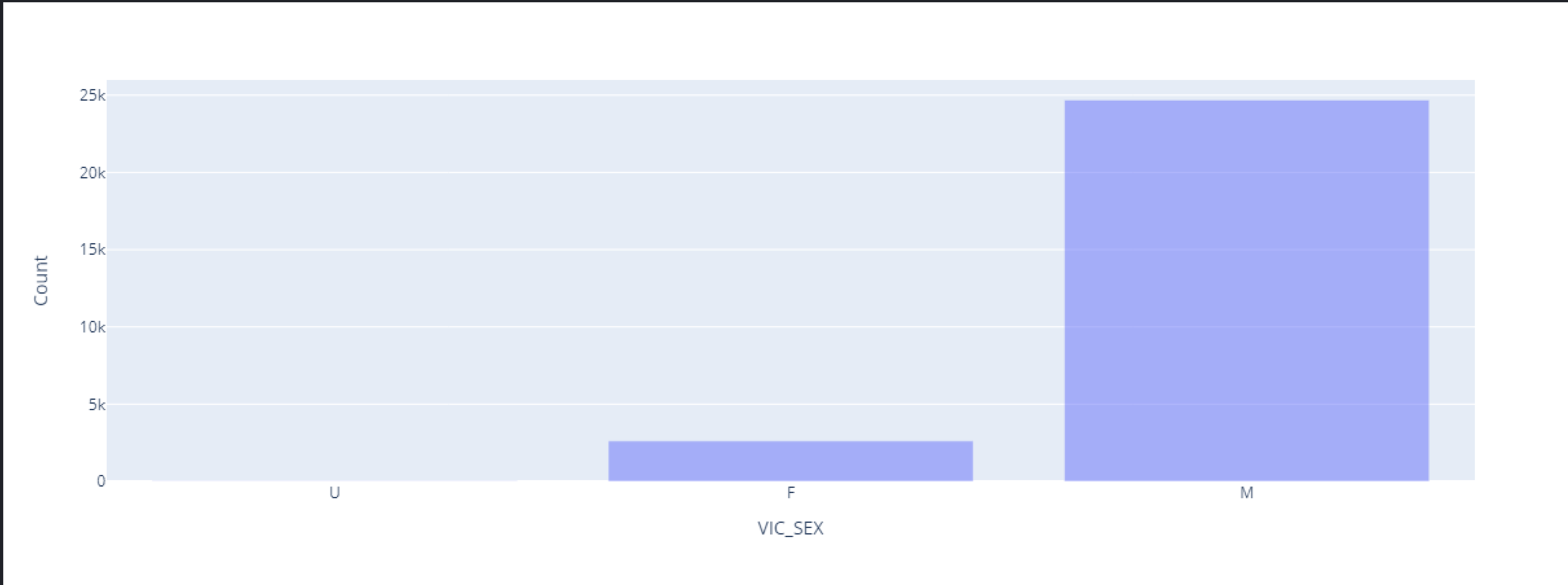
Perp Race


12k
10k
8k
6k
4k
2k
0
Count
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
WHITE
(null)
BLACK HISPANIC
UNKNOWN
WHITE HISPANIC
BLACK
PERP_RACE

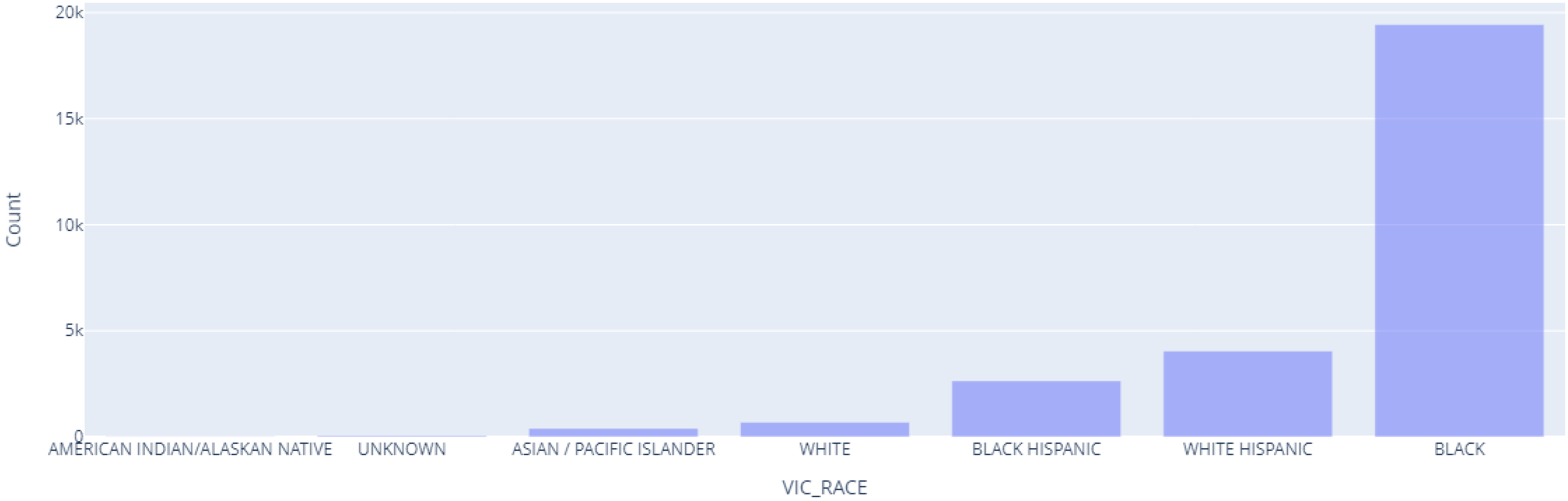
Vic Age


12k
10k
8k
6k
4k
2k
0
Count
1022
UNKNOWN
65+
45-64
<18
18-24
25-44
VIC_AGE_GROUP

Vic Sex



Vic Race


20k
15k
10k
5k
0
Count
AMERICAN INDIAN/ALASKAN NATIVE
UNKNOWN
ASIAN / PACIFIC ISLANDER
WHITE
BLACK HISPANIC
WHITE HISPANIC
BLACK
VIC_RACE

Modeling 🎓'

Preparing, Training, Validation

Model Details

XGBoost model

Tuned with 5-fold CV grid search

Use perp age/sex/race to predict victim age

Data Prep

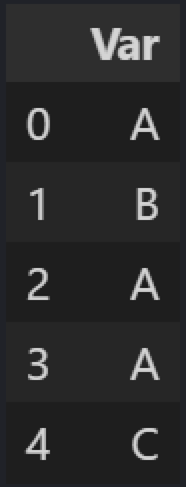
Remove all unknown, null, erroneous and missing values

Reduces dataset size to 14093 rows

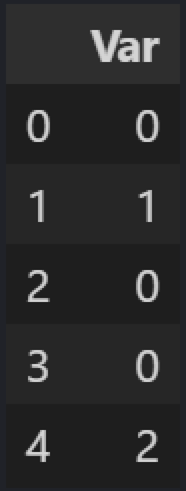
All features categorical, must encode

Feature Encoding Methods

Ordinal Encoding



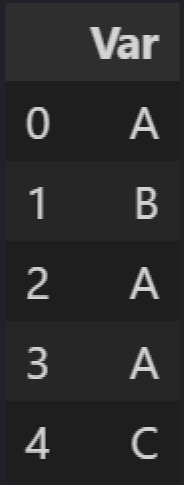
	Var
0	A
1	B
2	A
3	A
4	C



	Var
0	0
1	1
2	0
3	0
4	2

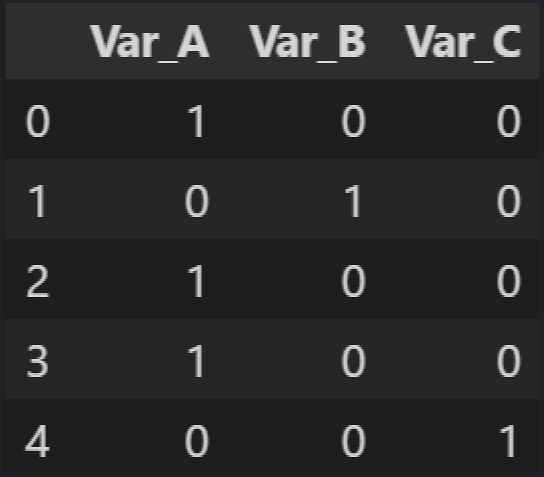
Feature Encoding Methods

Dummy/One Hot Encoding



	Var
0	A
1	B
2	A
3	A
4	C

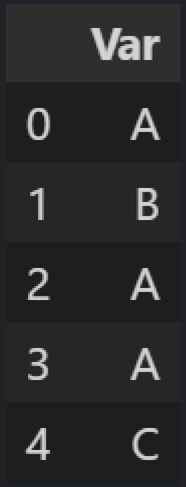
➡



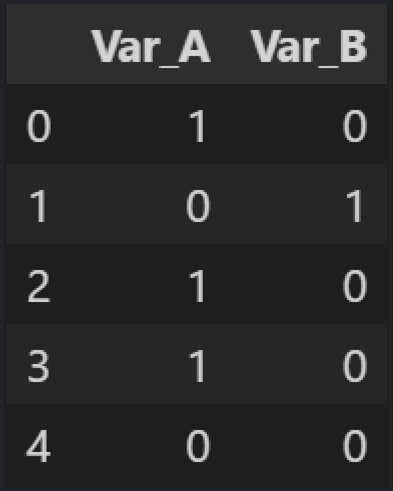
	Var_A	Var_B	Var_C
0	1	0	0
1	0	1	0
2	1	0	0
3	1	0	0
4	0	0	1

Feature Encoding Methods

Dummy/One Hot Encoding



	Var
0	A
1	B
2	A
3	A
4	C



	Var_A	Var_B
0	1	0
1	0	1
2	1	0
3	1	0
4	0	0

Feature Encoding Methods

Ordinal encode target variable (VIC_AGE_GROUP)

Dummy/One Hot encode features

Train/Test Split

70/30 training/test data split (9866, 4227)

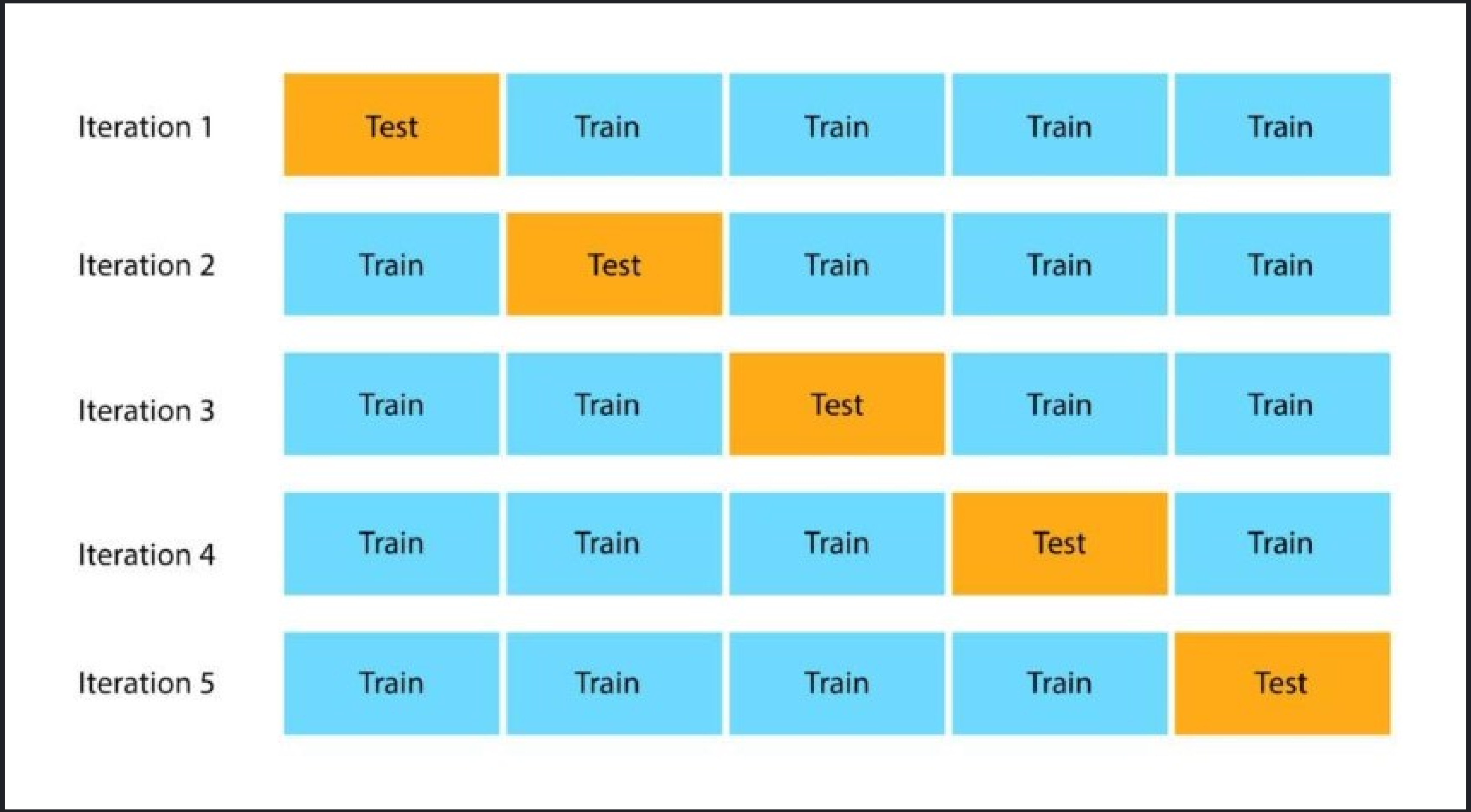
5 Fold CV for 2 hyperparameters

`max_depth` : [3,5,7]

`nrounds` : [25, 50, 75, 100, 125, 150, 175, 200, 225, 250]

Minimize Log Loss

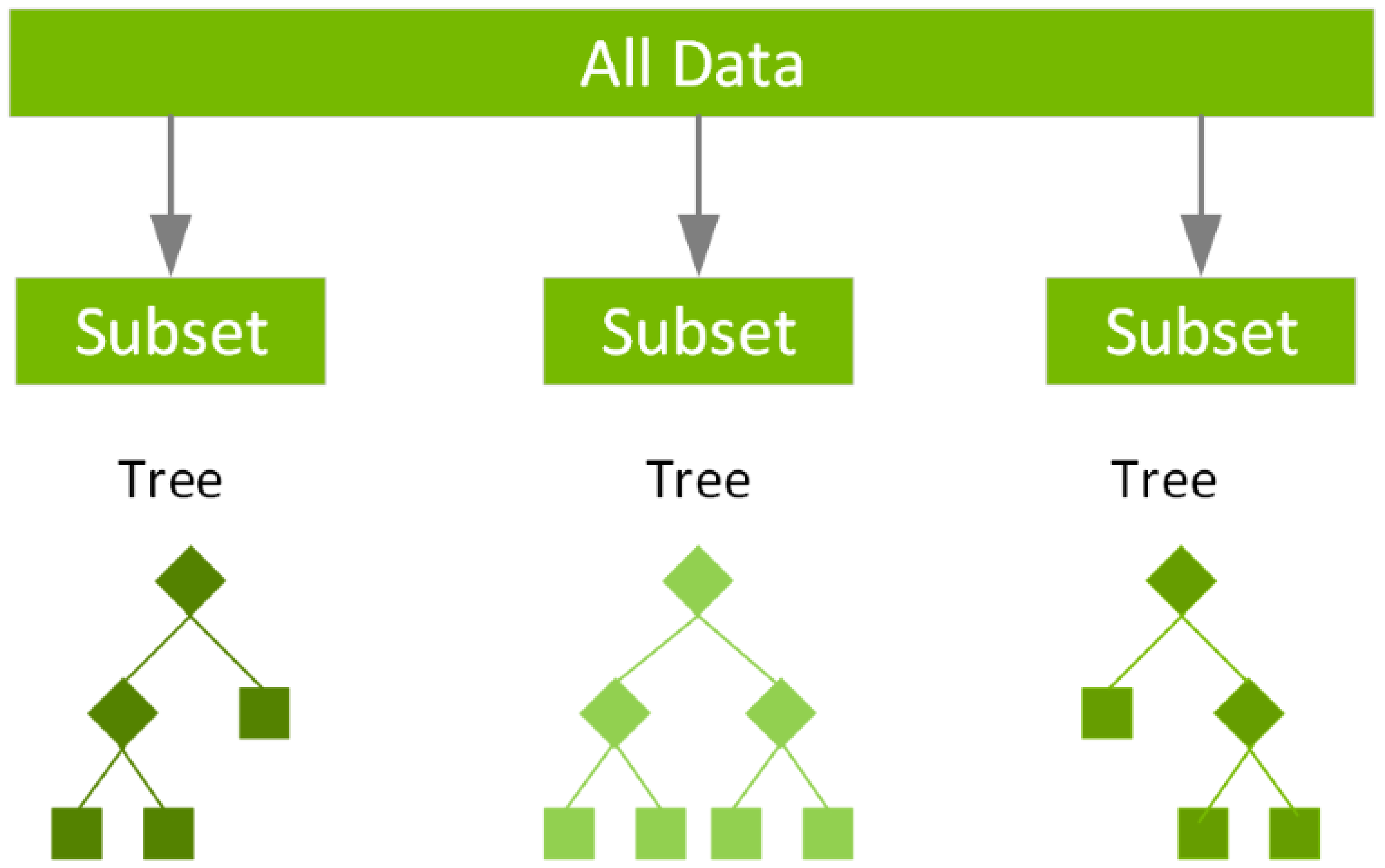
Cross Validation



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

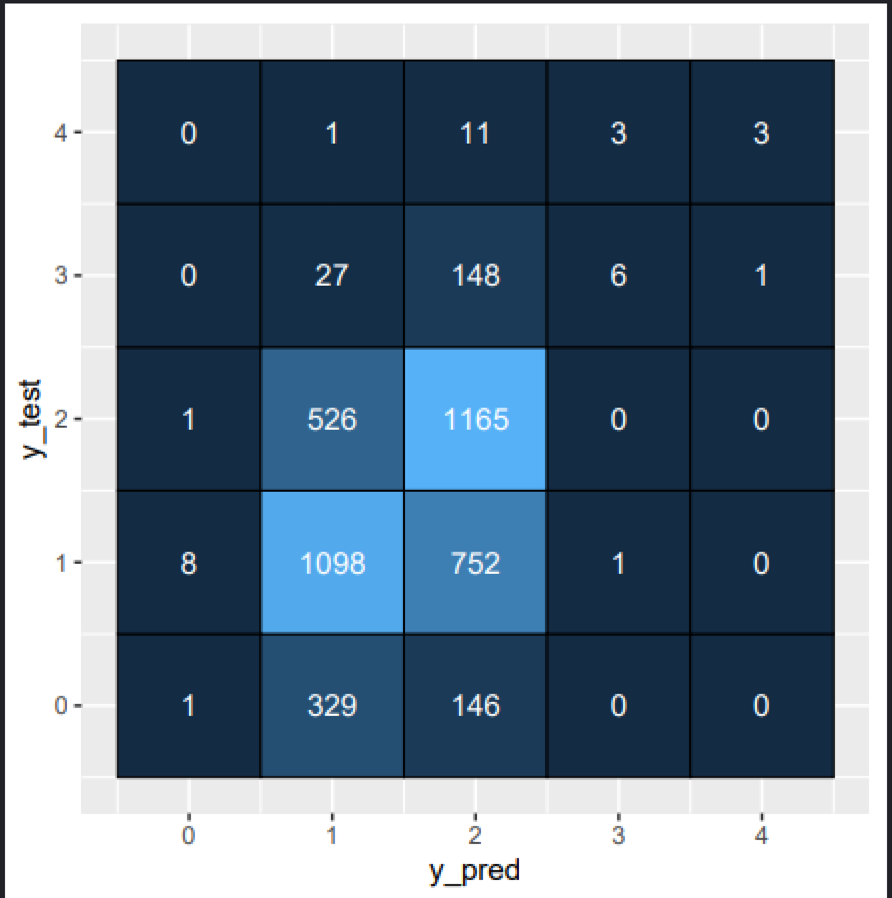
From: section.io

Model Architecture



From: Nvidia

Model Performance



y_test \ y_pred	0	1	2	3	4
4	0	1	11	3	3
3	0	27	148	6	1
2	1	526	1165	0	0
1	8	1098	752	1	0
0	1	329	146	0	0

Model Performance

C	True	Pred	Accuracy	Precision	Recall	F1
0	476	10	88.55%	0.10	0.0021	0.0041
1	1859	1981	61.11%	0.55	0.59	0.57
2	1692	2222	62.53%	0.52	0.69	0.60
3	182	10	95.74%	0.60	0.033	0.063
4	18	4	99.62%	0.75	0.17	0.27

Thanks For Listening!