FINAL PRESENTATION –

CRIME STATISTICS

HEATH THOMPSON
TAYLOR TOWNSEND
JIYIN ZHANG
SCOTT MARTIN
DAWSON HILL

University of Idaho

GOAL AND MODE OF COLLECTION

- **Data Driven Goal**
  - The datasets were generated to find statistical data about crime in the greater Los Angeles area from 2020 to present date. Our collection was driven by the need to better understand crime rates for 2021 in LA.
  - data.lacity.org

- **Collection Type**
  - The available data format for the dataset includes CSV, RDF, JSON, XML, and RSS. We chose to use CSV as it worked best for our group.
  - The metadata uses the PREMIS data model which consists of 4 interrelated entities: object, event, agent, and rights.

ANALYSIS STUDY FOR MACHINE LEARNING

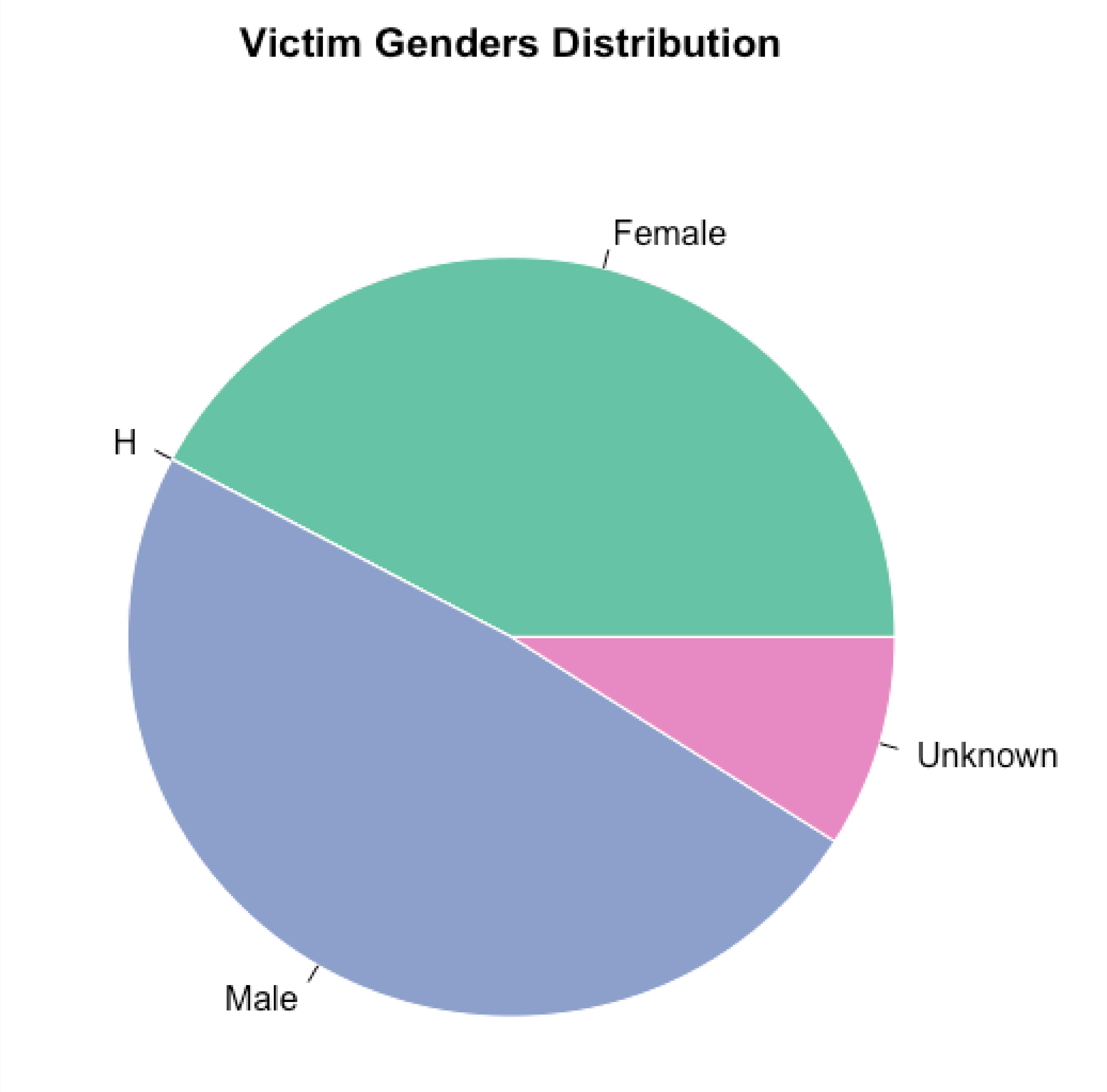
PRELIMINARY

- **Question 1 –**
  - Which victim descent (race) is most likely to be victims of a crime in Los Angeles?
- **Question 2 –**
  - Within the city of Los Angeles, at any particular time, is there an increase or decrease in crime rates, in comparison to the location where the crime took place?
- **Question 3 –**
  - Is there a correlation between the age of victims and their associated gender?

PRELIMINARY VISUALIZATIONS

MALE VERSUS FEMALE CRIME VICTIMS

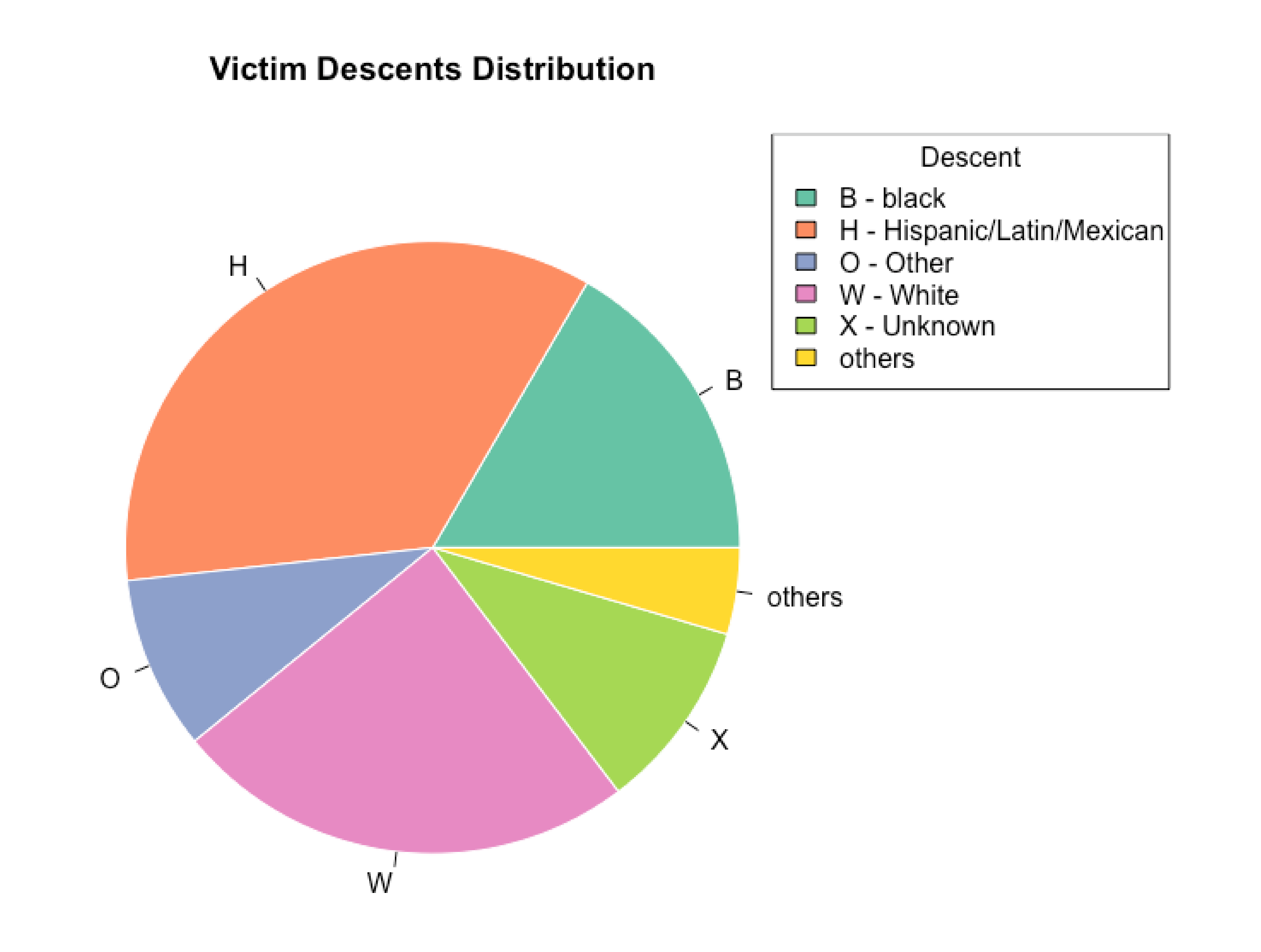
Gender	Amount
F - Female	75263
H	28
M - Male	86765
X - Unknown	16107



PRELIMINARY VISUALIZATIONS

RACE DISTRIBUTION

Descent	Amount
B - Black	29610
H – Hispanic/Latin/Mexican	62546
O - Other	16332
W - White	43553
X - Unknown	17970
others	8150



PROPOSED OUTCOME

ACCURATELY IDENTIFY PEOPLE WHO ARE MOST LIKELY TO BECOME A VICTIM OF CRIME IN LA

If there is a direction correlation considering location, time, age, race, and sex, then can we predict if certain individuals are most susceptible to being a victim of a crime in Los Angeles?

Analysis is based on the following categories –

- Race
- Age
- Sex
- Time of day
- Location

MACHINE LEARNING - KMODES

Step 1: Assign K observations as the K clusters/leaders

Step 2: Calculate the dissimilarities and assign observation to most similar cluster

Step 3: Update the clusters features

Step 4: Repeat steps 2 and 3 until the clusters no longer change

https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

Leaders			
P1	blonde	amber	fair
P7	red	green	fair
P8	black	hazel	fair

person	hair color	eye color	skin color
P1	blonde	amber	fair
P2	brunette	gray	brown
P3	red	green	brown
P4	black	hazel	brown
P5	brunette	amber	fair
P6	black	gray	brown
P7	red	green	fair
P8	black	hazel	fair

MACHINE LEARNING - KMODES

	Cluster 1 (P1)	Cluster 2 (P7)	Cluster 3 (P8)	Cluster
P1	0 ✓	2	2	Cluster 1
P2	3 ✓	3	3	Cluster 1
P3	3	1 ✓	3	Cluster 2
P4	3	3	1 ✓	Cluster 3
P5	1 ✓	2	2	Cluster 1
P6	3	3	2 ✓	Cluster 3
P7	2	0 ✓	2	Cluster 2
P8	2	2	0 ✓	Cluster 3

MACHINE LEARNING - KMODES

person	hair color	eye color	skin color
P1	blonde	amber	fair
P2	brunette	gray	brown
P3	red	green	brown
P4	black	hazel	brown
P5	brunette	amber	fair
P6	black	gray	brown
P7	red	green	fair
P8	black	hazel	fair

FULL ANALYSIS ML - PREPROCESSING

```
import pandas as pd
from kmodes.kmodes import KModes
import matplotlib.pyplot as plt
from seaborn import countplot
```

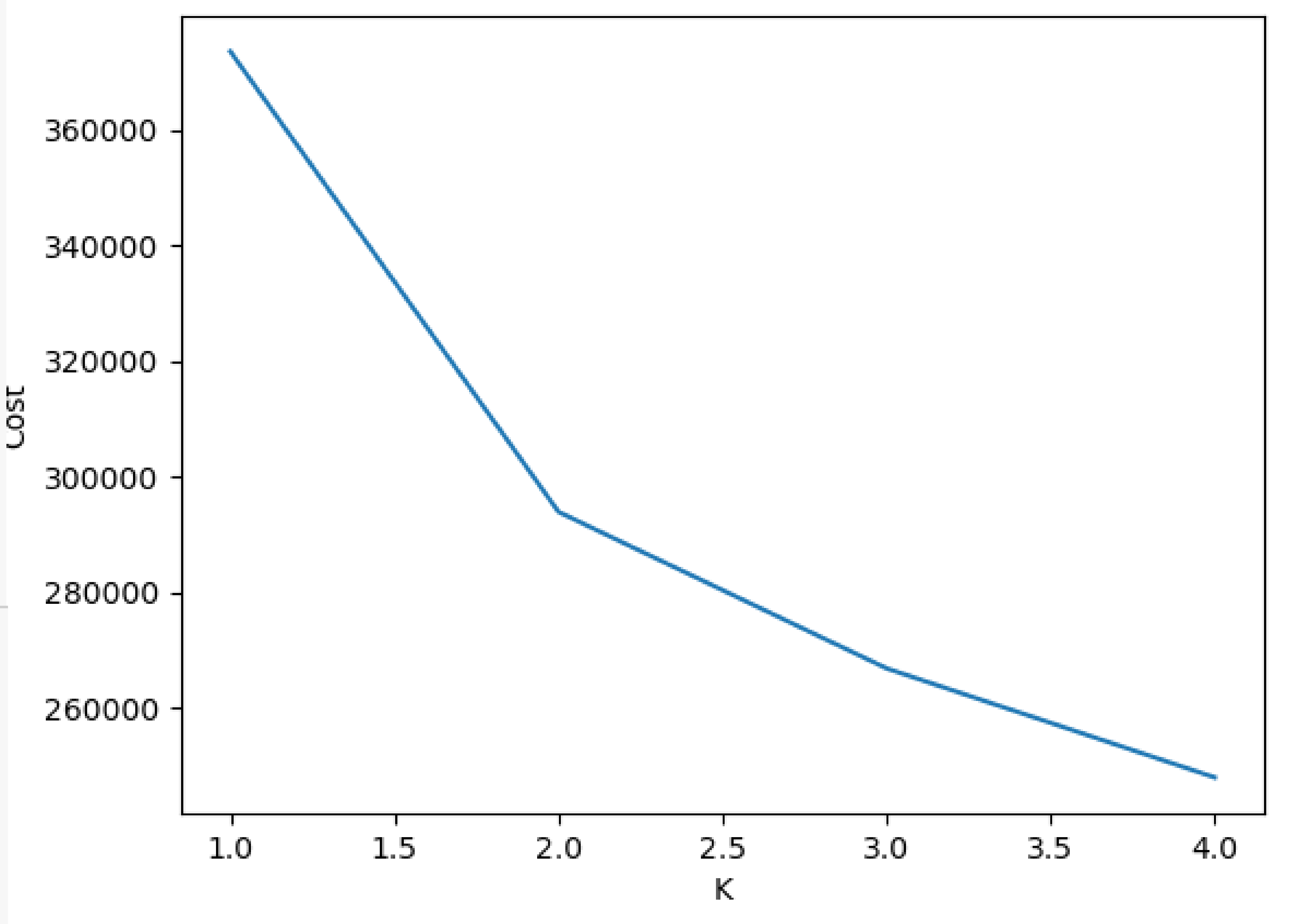
```
# read csv into pandas dataframe
data = pd.read_csv(file_path)
data = data.drop(columns=['DATE.OCC','DR_NO','AREA.NAME'])
data.head(10)
```

```
#Remove any rows with H in their Vict.Sex column
cleaned = data.drop(data[data['Vict.Sex']=='H'].index)
#Remove any rows without the specified characters below in their Vict.Descent column
toKeep = ['B','H','O','W','X']
cleaned = cleaned.drop(cleaned[cleaned['Vict.Descent'].isin(toKeep)==False].index)
encode.head(10)
#the table printed now only stores records with the most meaningful data
```

```
#here, we will implement Kmodes (similar to KMeans, but clusters categorical variables rather than numerical).
#to do this, for now, we will strip the loc (lat, lon) data. What's left is our categorical data
#but we will categorize age data in ranges (10-20, 21-30, 31-40, etc.) as well as time data
KmodesData = pd.DataFrame.copy(cleaned)
KmodesData.head()
KmodesData.drop(columns=['LAT', 'LON'], inplace=True)
#group ages into age bins
KmodesData['AgeBins'] = pd.cut(KmodesData['Vict.Age'], bins=[0,20,30,40,50,60,70,80,max(KmodesData['Vict.Age'])])
KmodesData.drop(columns=['Vict.Age'], inplace=True)
#group times into time bins
KmodesData['TimeOccBins'] = pd.cut(KmodesData['TIME.OCC'], bins=[0, 600, 1200, 1600, 2100, 2400])
KmodesData.drop(columns=['TIME.OCC'], inplace=True)
#convert it all to strings to ensure categories can be determined by KModes
KmodesData = KmodesData.astype('str', copy=True)
KmodesData.head(10)
```

FULL ANALYSIS ML – MODEL

```python
#Kmodes requires us to give it the number of clusters we wish to categorize
#we will use the Elbow method to determine this number of clusters K
cost = []
K = [1,2,3,4]
for i in K:
    kout = KModes(n_clusters=i, init='Cao', n_init=4)
    kout.fit_predict(KmodesData)
    cost.append(kout.cost_)
plt.plot(K, cost)
plt.xlabel('K')
plt.ylabel('Cost')
plt.show()
#we will select the farthest right significant bend...
#we can see the bend at K=2, so we will use 2 clusters
```




```python
#we will now implement our KModes algo
kout = KModes(n_clusters=2, init='Cao', n_init=4)
#and use the fitted model to assign clusters to each victim
clusters = kout.fit_predict(KmodesData)
#finally append the cluster values to our dataframe
KmodesData['Cluster'] = clusters
#KmodesData.head(10)
#make copy of of this data and add lat and long back to it.
#We will use this dataframe for further data analysis
csv = pd.DataFrame.copy(cleaned)
csv['Cluster'] = clusters
csv.to_csv(path)
```

MACHINE LEARNING – POST ANALYSIS

```
kout.cluster_centroids_
```

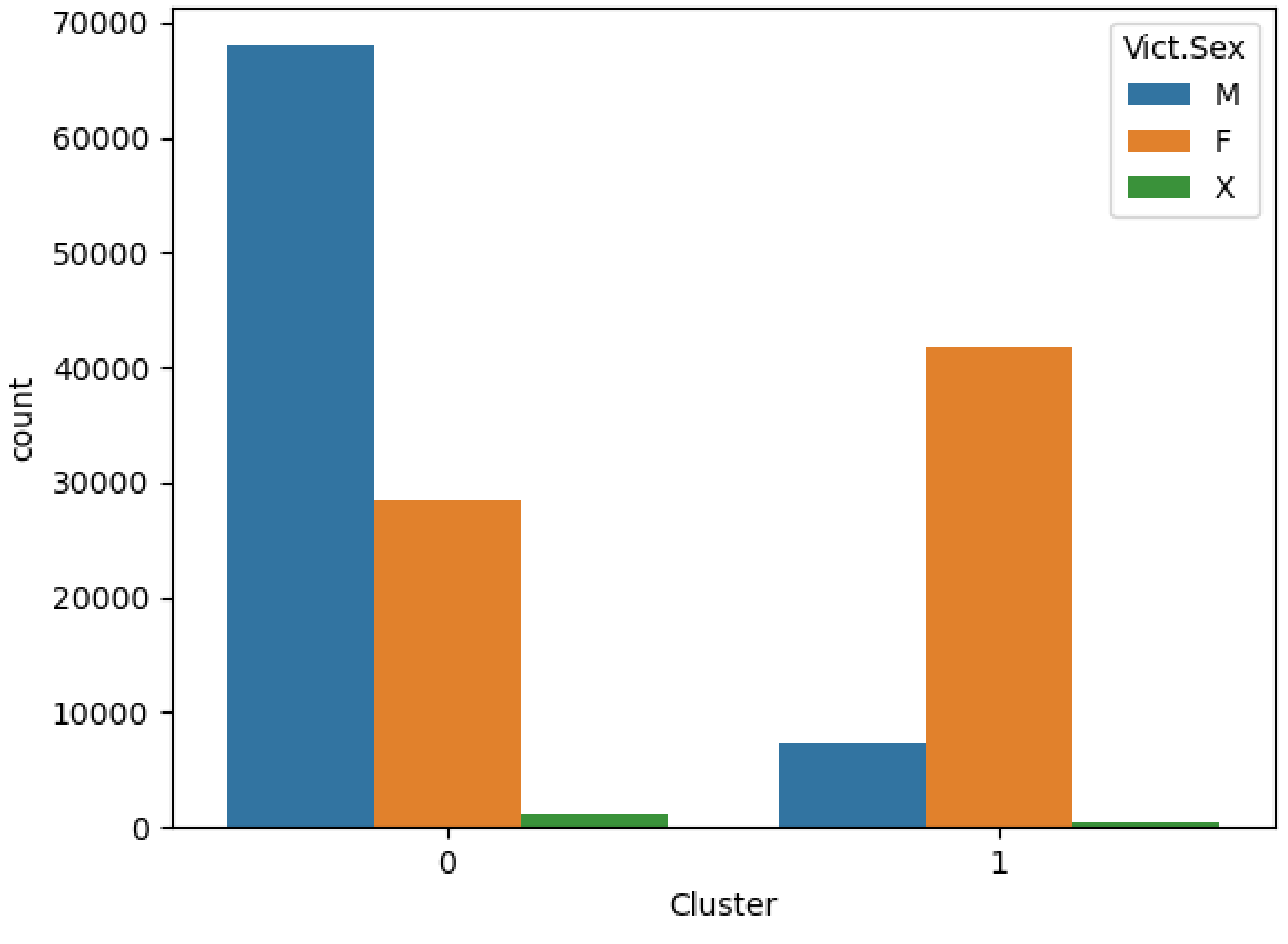
```
array([['M', 'H', '(30, 40]', '(1600, 2100]'],
       ['F', 'W', '(20, 30]', '(600, 1200]']], dtype='<U12')
```

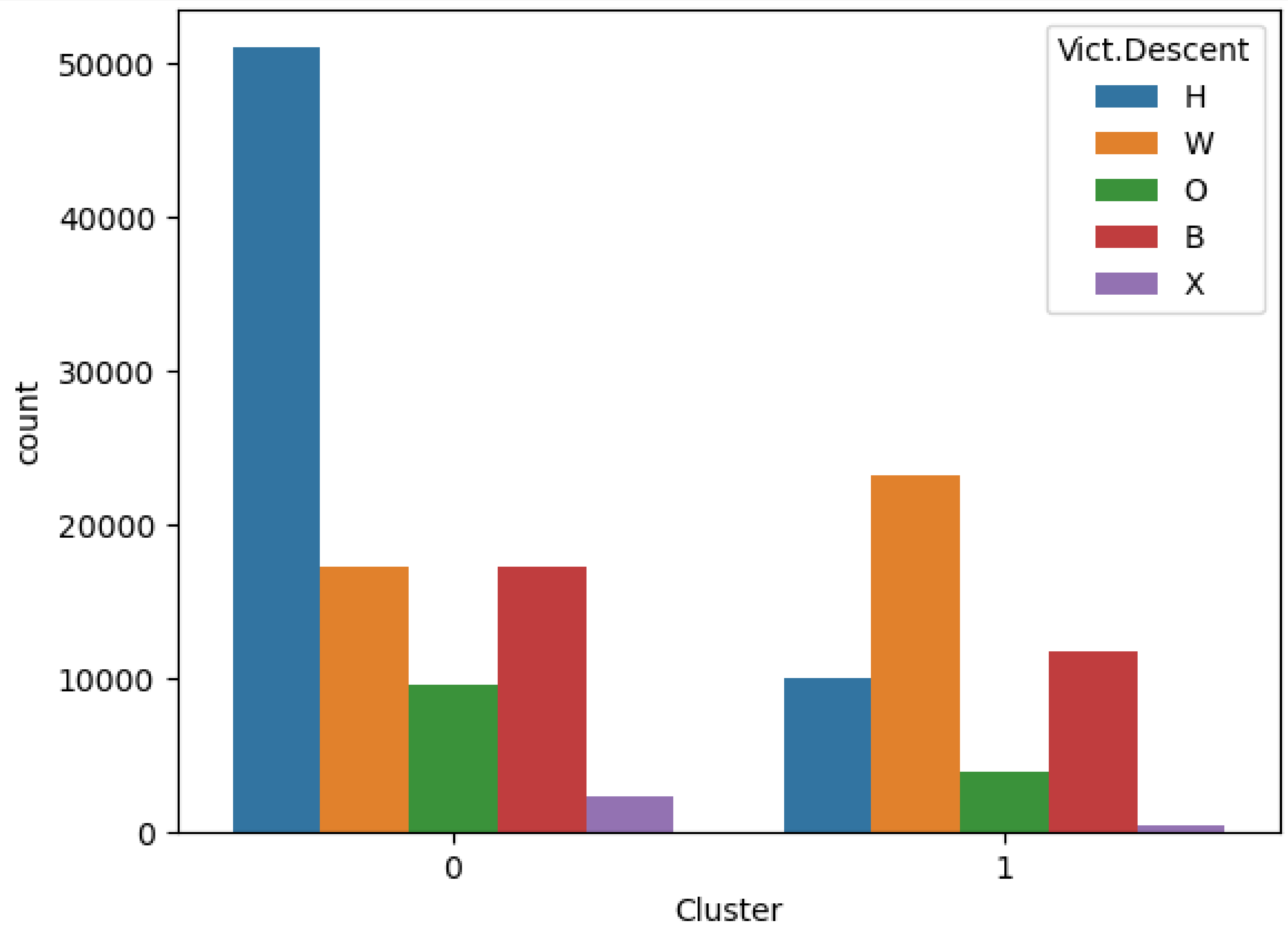
K-Modes Output:

- Initialized clusters:
- Each entry assigned a cluster
- Clusters update after each iteration

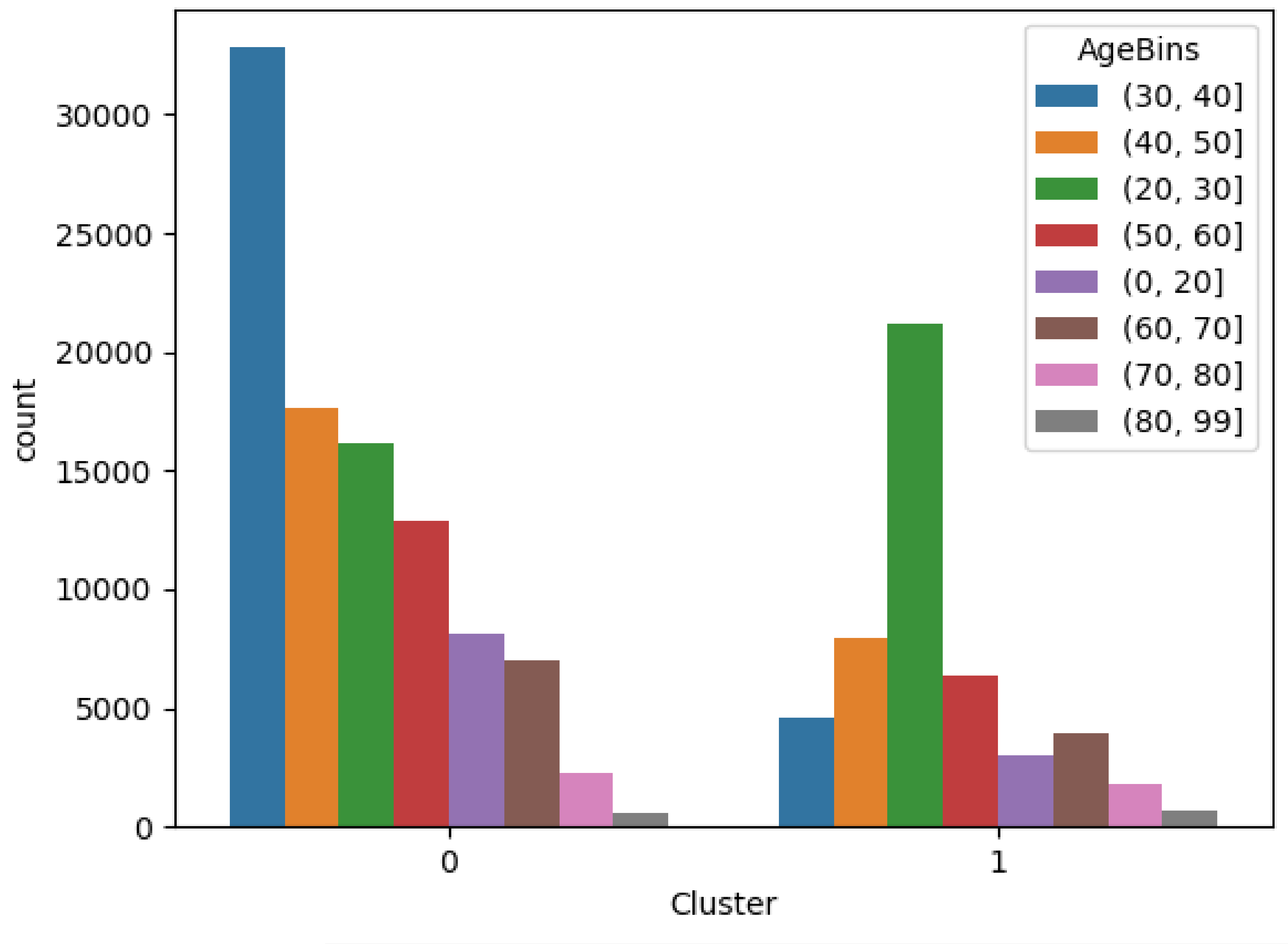
	Vict.Sex	Vict.Descent	AgeBins	TimeOccBins	Cluster
0	M	H	(30, 40]	(0, 600]	0
1	M	W	(40, 50]	(1200, 1600]	0
2	M	H	(20, 30]	(1600, 2100]	0
3	F	O	(50, 60]	(600, 1200]	1
4	F	B	(20, 30]	(1200, 1600]	1
5	F	B	(50, 60]	(2100, 2400]	1
6	M	B	(20, 30]	(1200, 1600]	0

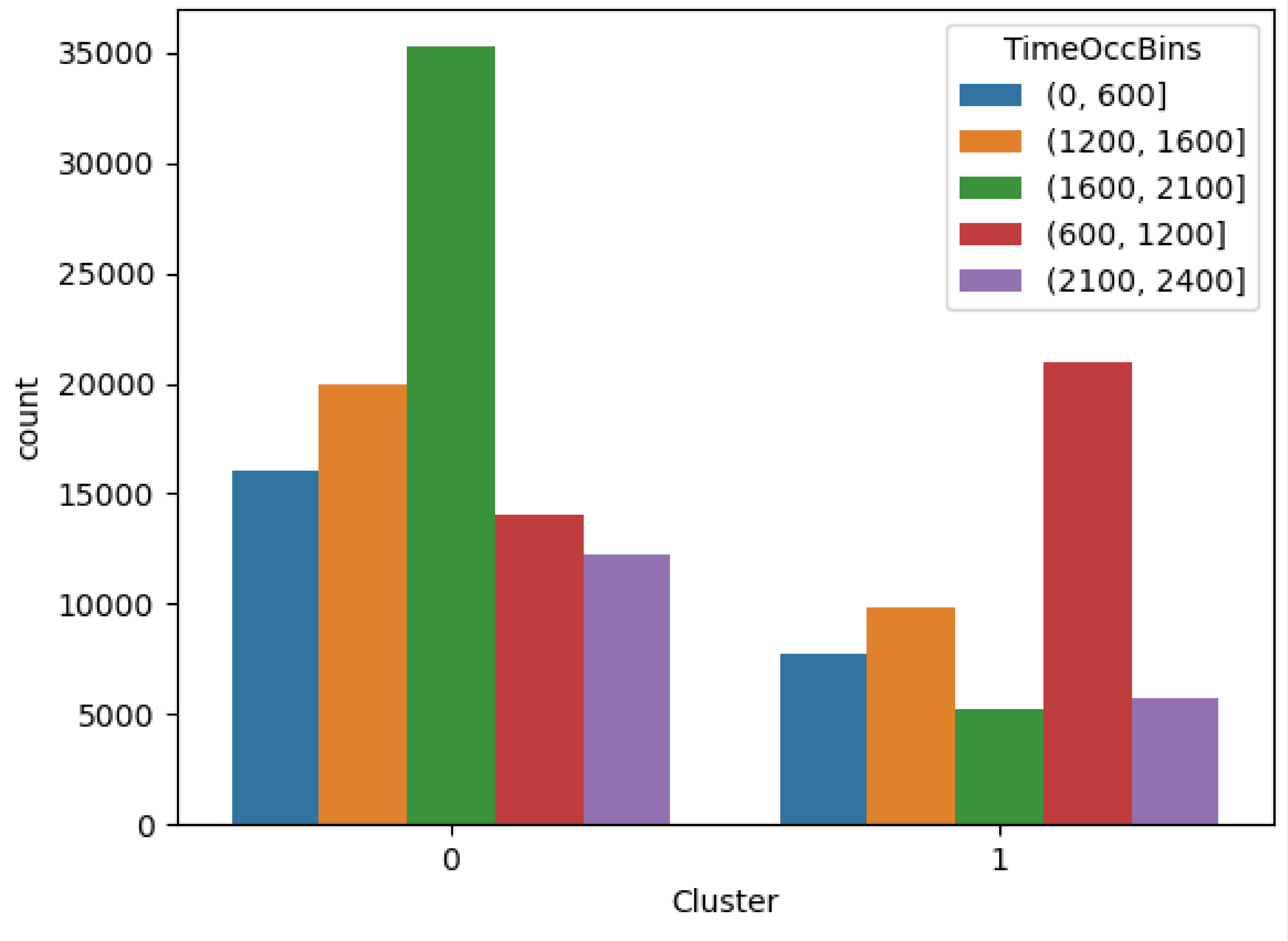
MACHINE LEARNING – POST ANALYSIS





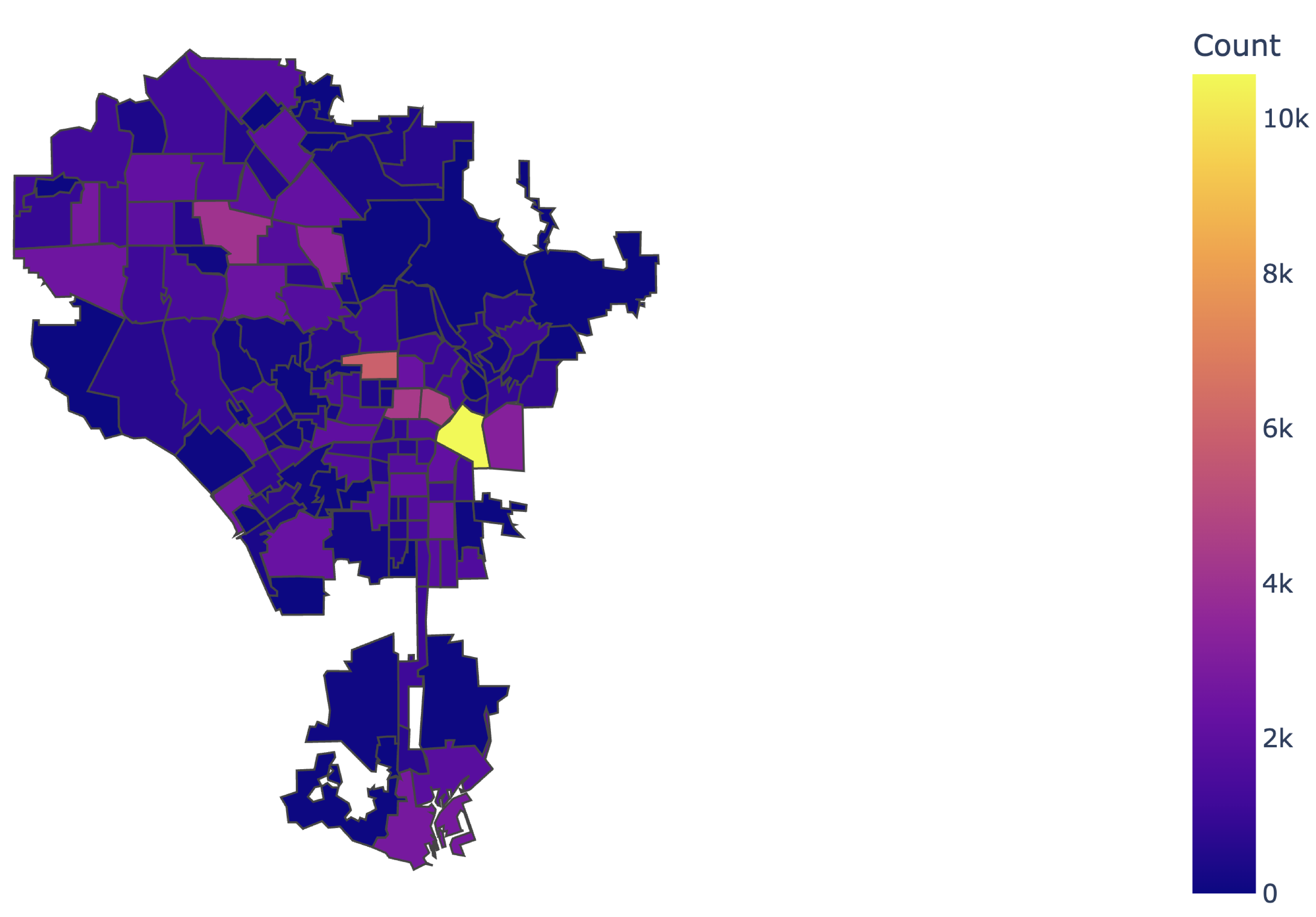
MACHINE LEARNING - POST ANALYSIS





VISUALIZATIONS CHOROPLETH MAP FOR INCIDENTS

AreaName	Count
Downtown	10577
Hollywood	5948
Westlake	4716
Koreatown	4424
Van Nuys	4126
North Hollywood	3389
Boyle Heights	3220
San Pedro	2782
Canoga Park	2760
Venice	2611



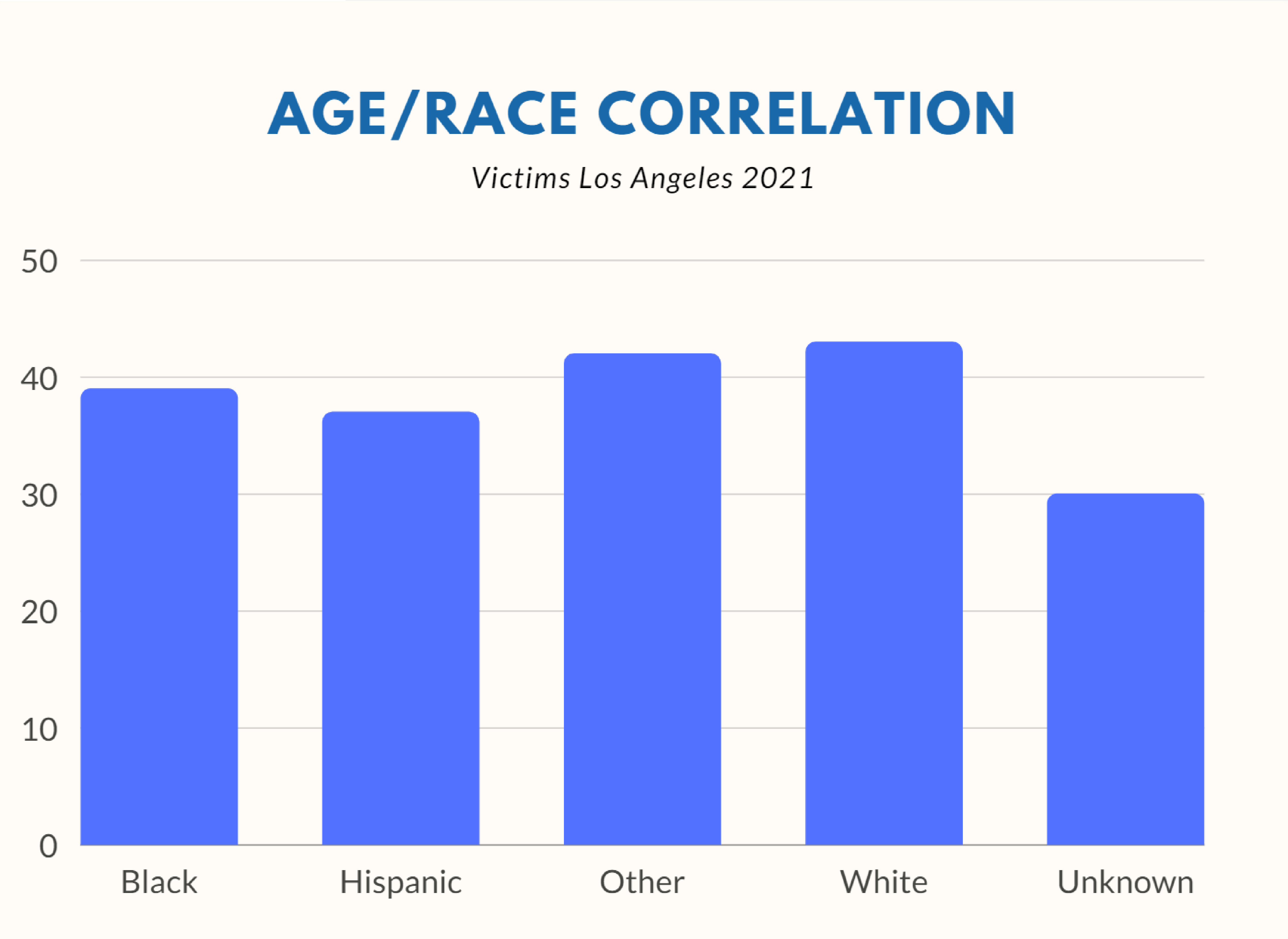
HYPOTHESES QUESTIONS – POST ANALYSIS

- Question 1 –
  - Which victim descent (race) is most likely to be victims of a crime in Los Angeles?
    - Answer – Hispanic
- Question 2 –
  - Within the city of Los Angeles, at any particular time, is there an increase or decrease in crime rates, in comparison to the location where the crime took place?
    - Answer – Downtown LA has greatest number of incidences between 4:00-9:00PM
- Question 3 –
  - Is there a correlation between the age of victims and their associated gender?
    - Answer – Yes, we found that male victims tend to be roughly 10 years older than their female counterparts

VALIDATION

CORRELATION OF AGE AND RACE

- Average age of each descent group

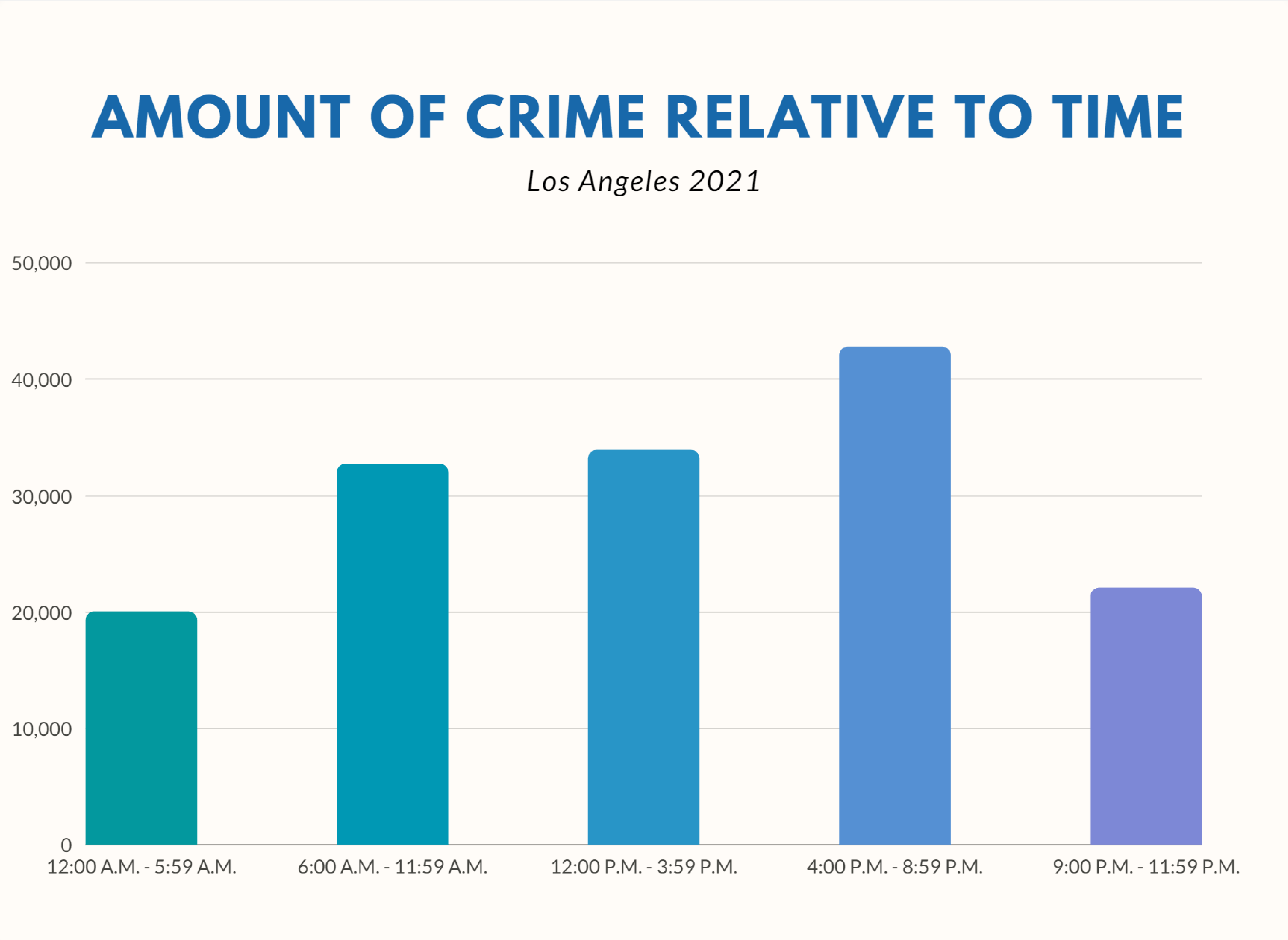


Race	Average Age
Black	39
Hispanic	37
Other	42
White	43
Unknown	30

VALIDATION

TIME AND LOCATION

Occurrences of crime during different times of the day



DATA MANAGEMENT PLAN

Logical Collections – 1 file with 28 columns

- Date.Rptd
- DATE.OCC
- Crm.Cd.1

Physical Data Handling

- Dataset – CSV file
- Metadata – no available file, only on webpage

Interoperability Support – available for download in multiple formats

- CSV, RDF, JSON, XML, and RSS

DATA MANAGEMENT PLAN

Security – None; available for public download

Data Ownership – Los Angeles Police Department

Metadata – 4 interrelated entities: object, event, agent, rights

DATA MANAGEMENT PLAN

Persistence – updated weekly online

Discovery

- Data.lacity.org
- Category (LAPD)
- Tags (lapd, crime, crime data, police, safe city, crimes)

Dissemination – publicly available online

THANK YOU! QUESTIONS?