Data Exploration and Visualization Assignment 1 Exploratory Visual Analysis

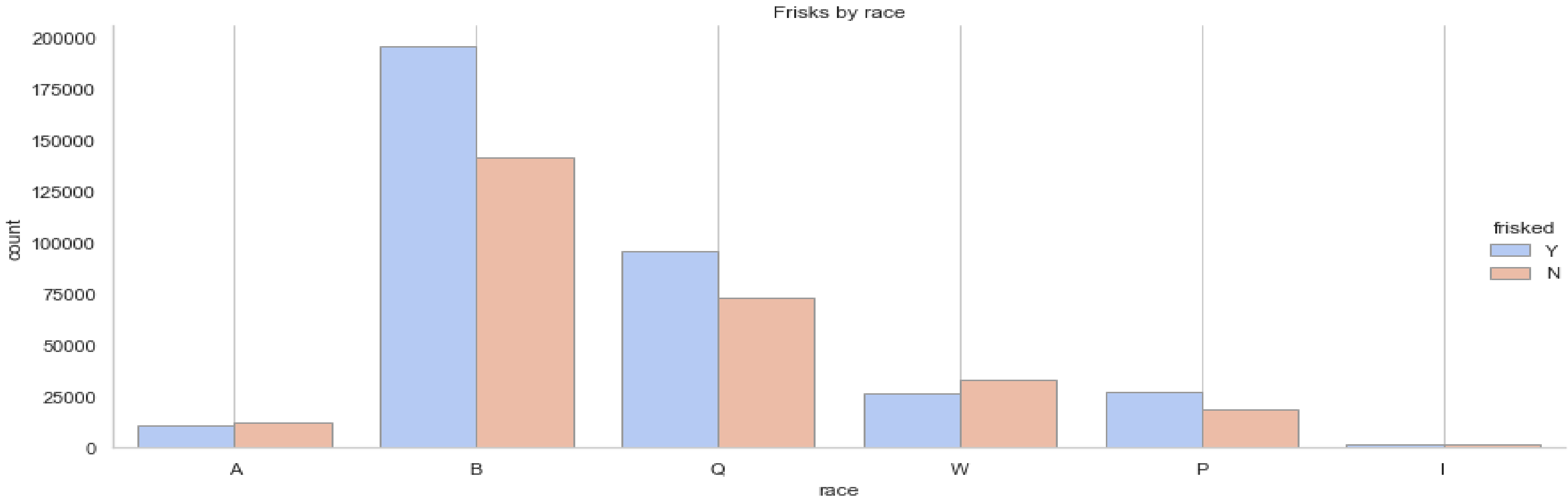
AYA MIGDADY

Task 2:

Investigate hypotheses (questions) and develop preliminary insights:

What is the effect of each of the following on a number of Frisk:……………….(Race)

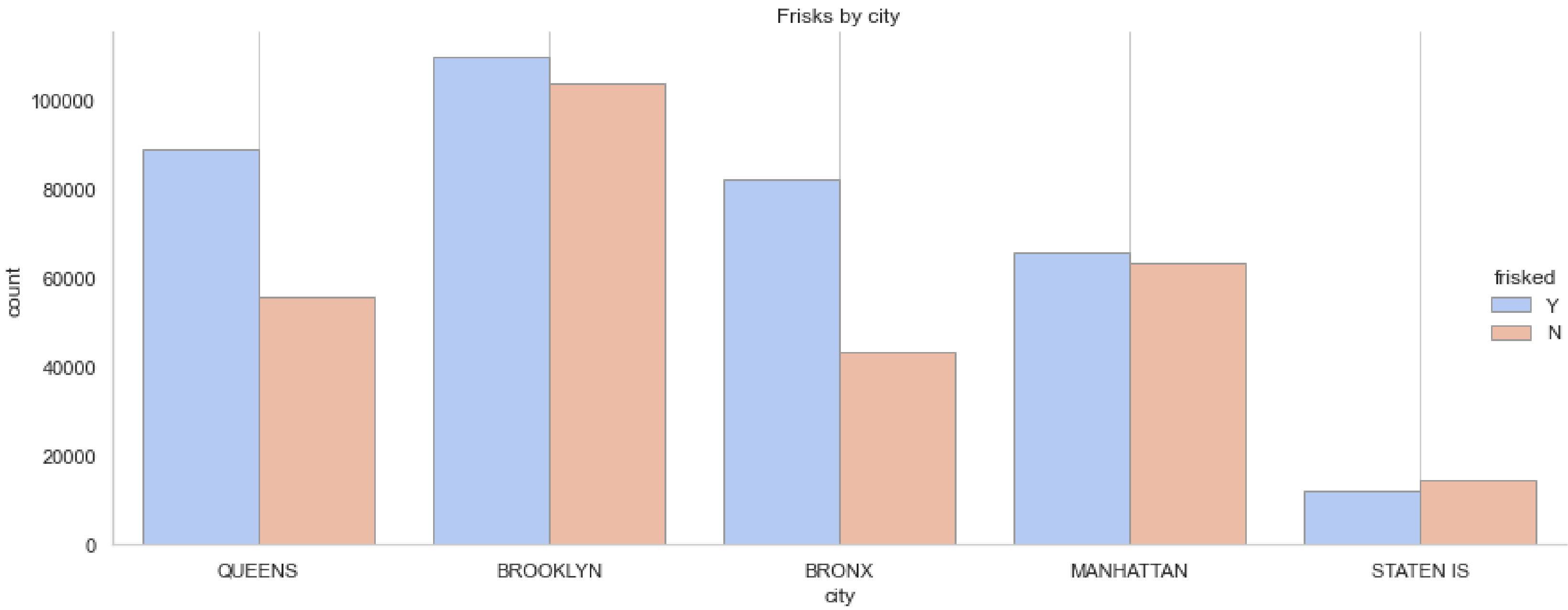
Individuals (stopped by police) who are black (B), white-hispanic (Q) and black-hispanic (P) are more likely to be frisked than not, with African-Americans being frisked the most. On the other hand, individuals who are white (W), Asian (A) and American Indian (I) are less likely to be frisked, with white individuals having the highest no-frisk to frisk ratio.



Task 2:

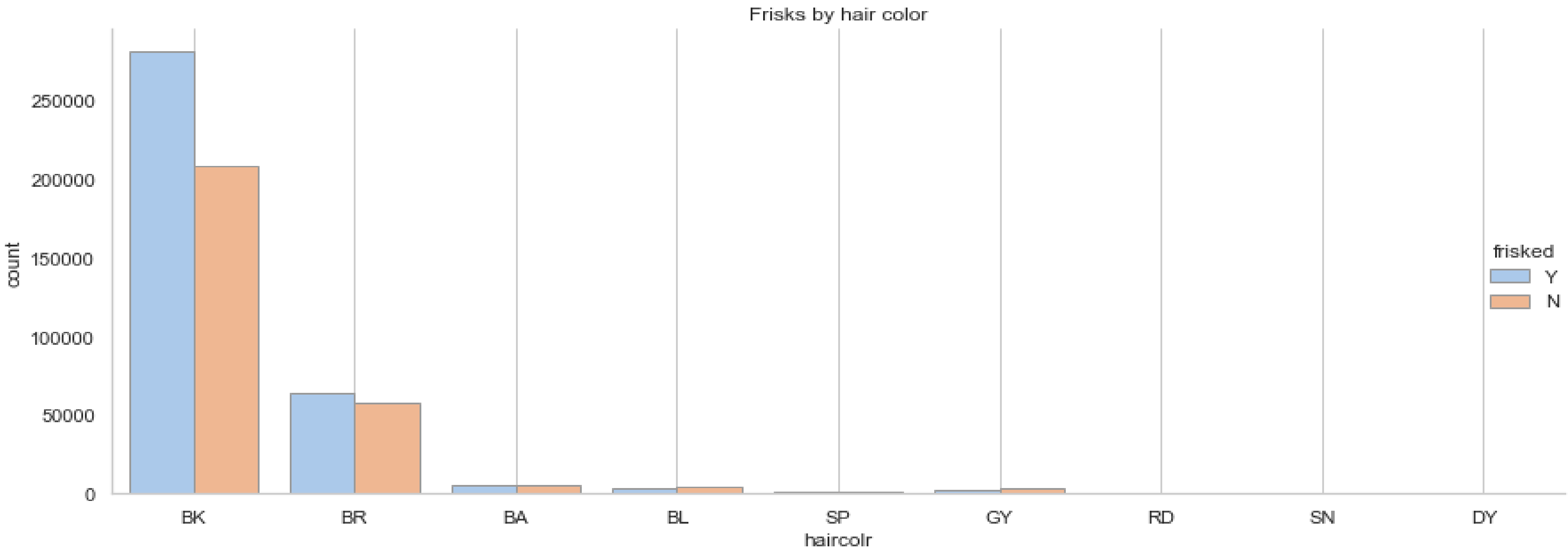
City

It looks like one is more likely to be frisked in most cities (if stopped) except in Staten Island. However, the discrepancy of no-frisk to frisk in Bronx and Queens is the greatest.



Task 2:

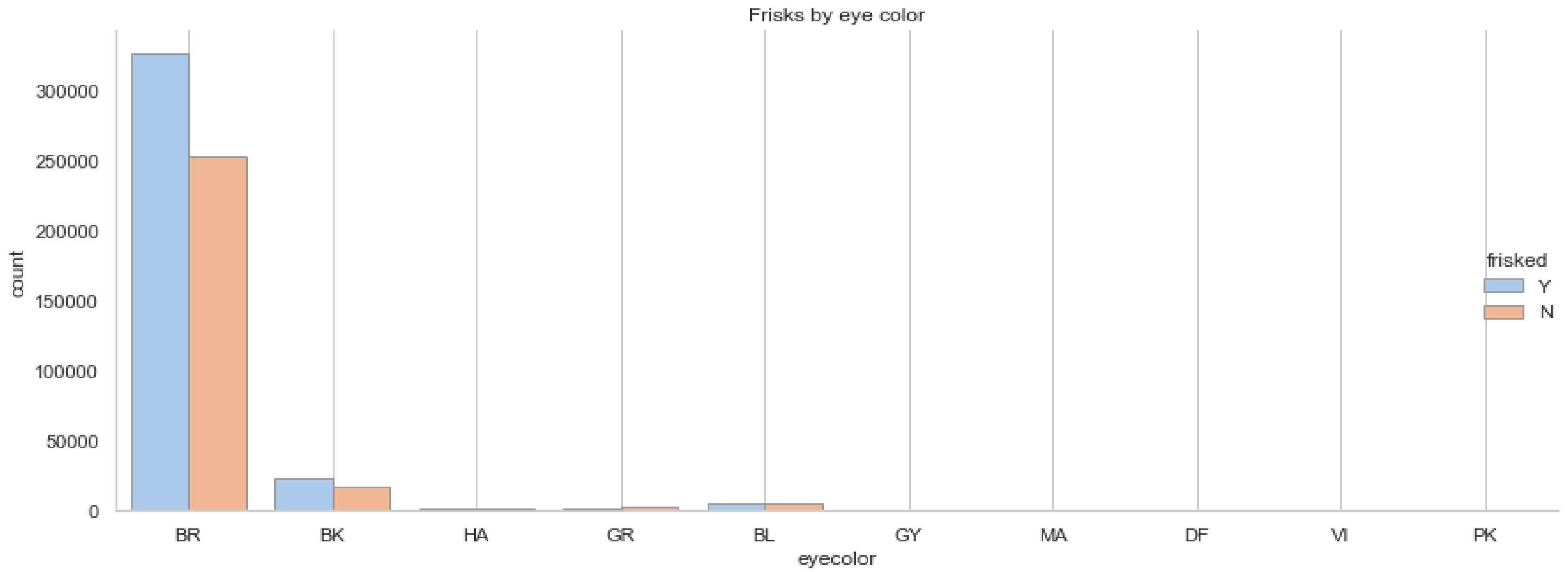
hair color



Task 2:

eye color

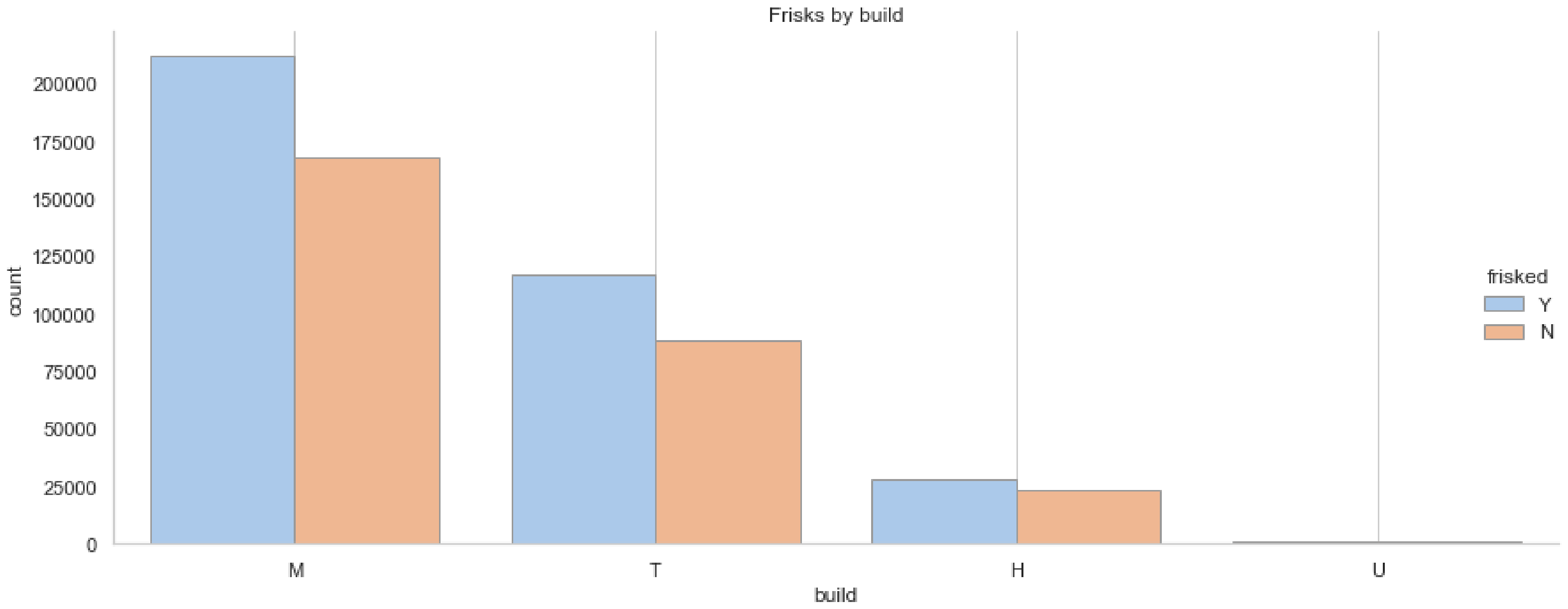
Individuals with black hair (BK) and brown eyes (BR) are more likely to be frisked, but given there isn't too much gap between Y/N for all colors of hair and eyes, these attribute don't seem to be significant factors in predicting if someone gets frisked.



Task 2:

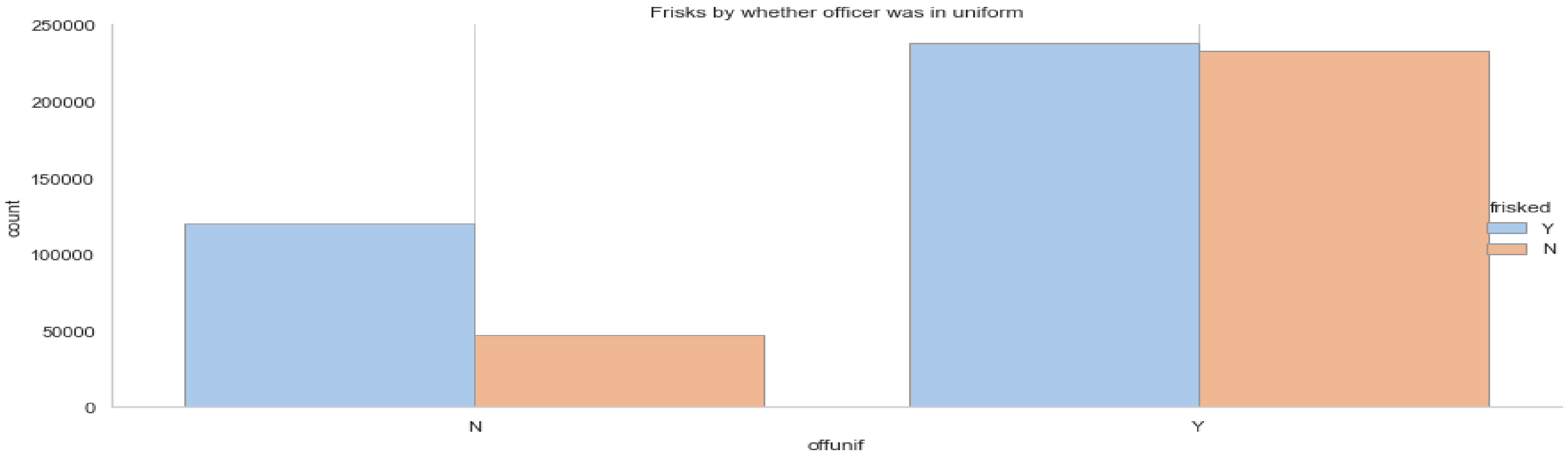
build of body

Individuals with medium (M) and thin (T) body builds are more likely to be frisked than heavy (H) or muscular (U) builds.



if officer is in uniform

This chart is interesting as it tells us that an individual is 2.5 times more likely to be frisked if the officer is not in a uniform. On the other hand, there isn't too much difference if the officer is in uniform. This is contradictory to what we would expect.

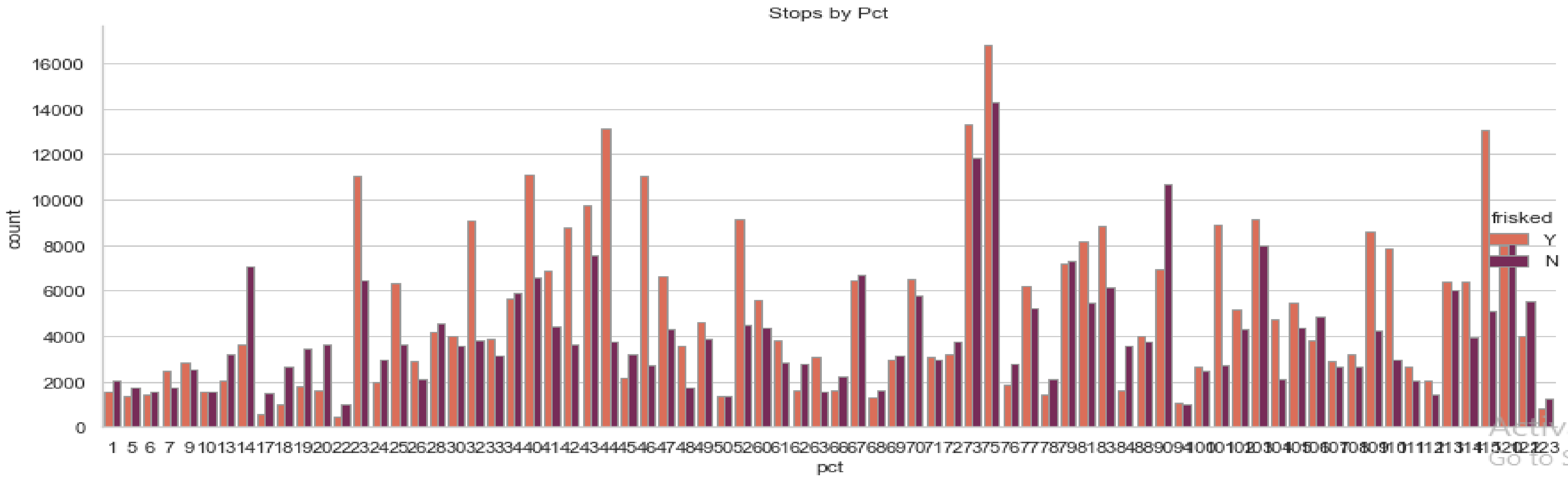


Task2:

What is number of stops per:

Precinct

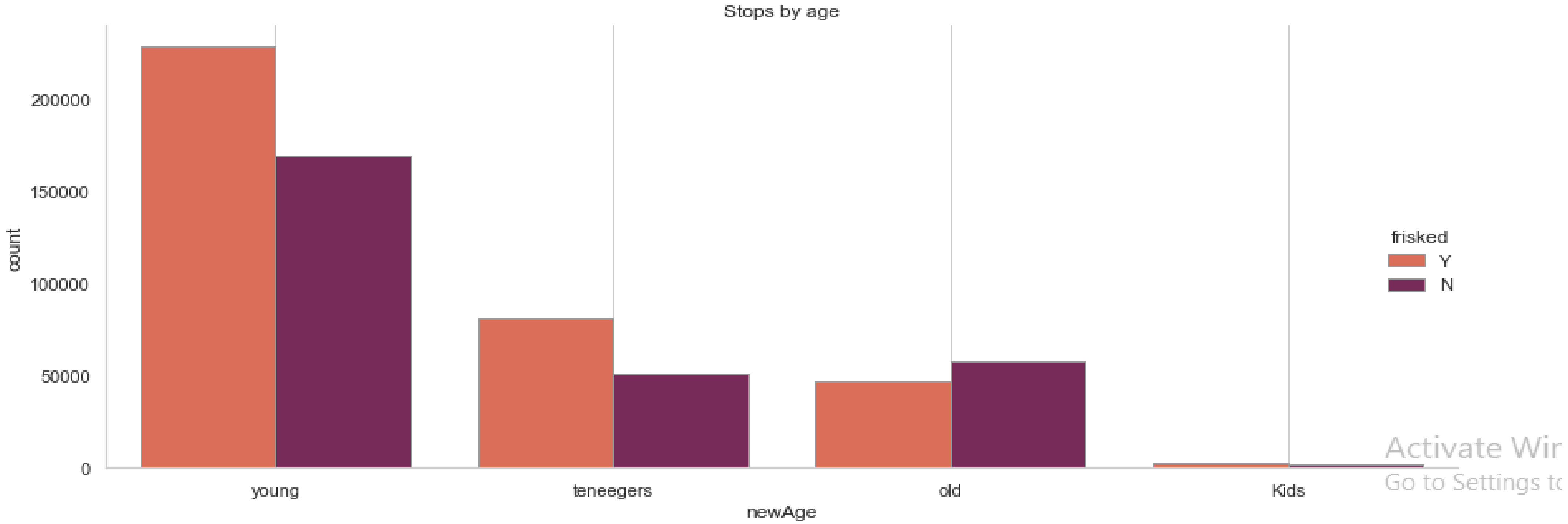
We conclude that the Precinct with number 75 have mostly stopped people than other Precinct.



Task2

Age

We conclude that the young people have been mostly stopped than other people followed by teenagers.



There is a correlation between age and timestop?

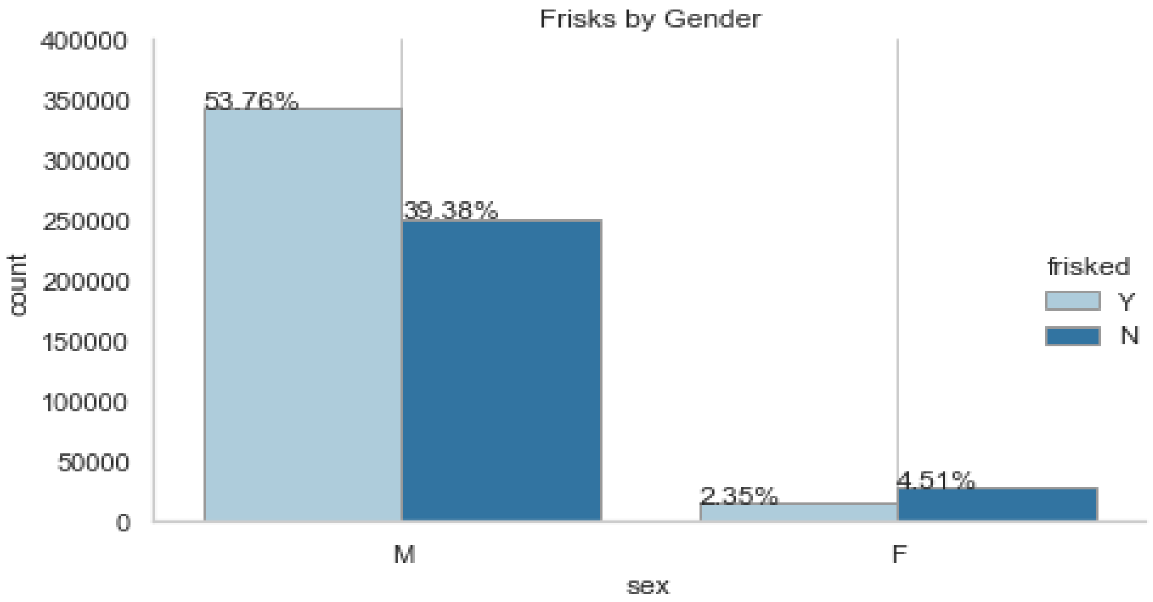
According to the value of R which close to Zero, we conclude theirs is not a correlation between Age and timeofstop variables.

Task 3:

Answer the following questions. Use visualization techniques to make the answers visible.

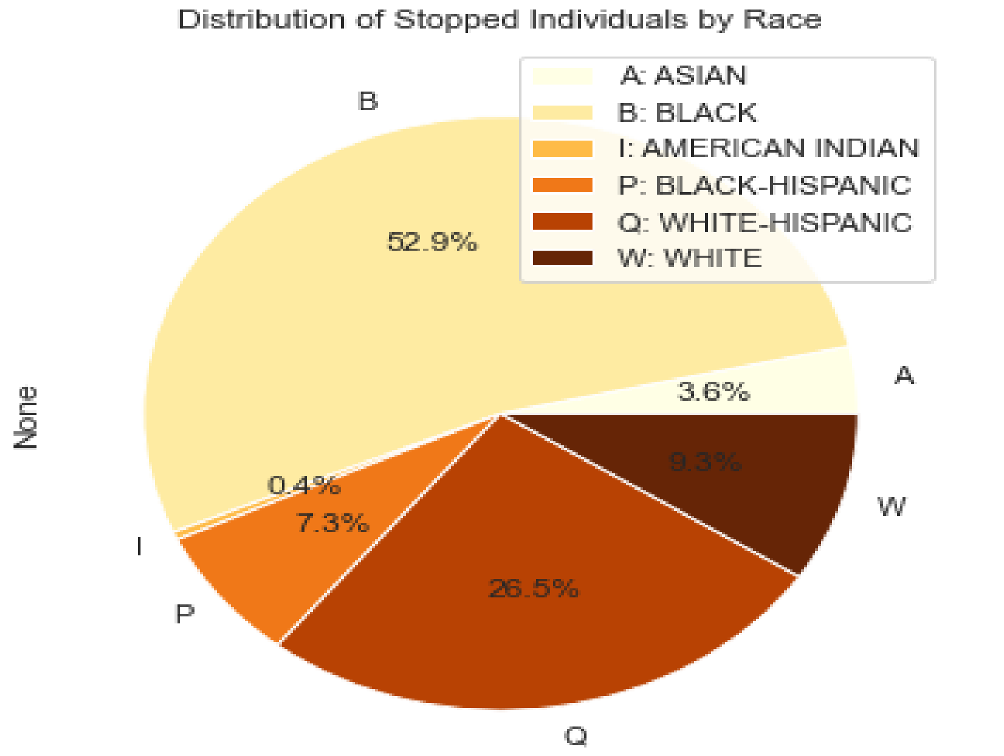
- Who is most affected by Stop and Frisk, and in what capacity are they affected?

From the bar chart below, it looks like males (M) are much more likely to be frisked than females (F), by more than 20 times. Women are also less likely to be frisked if stopped by an officer, whereas men are.



Task 3:

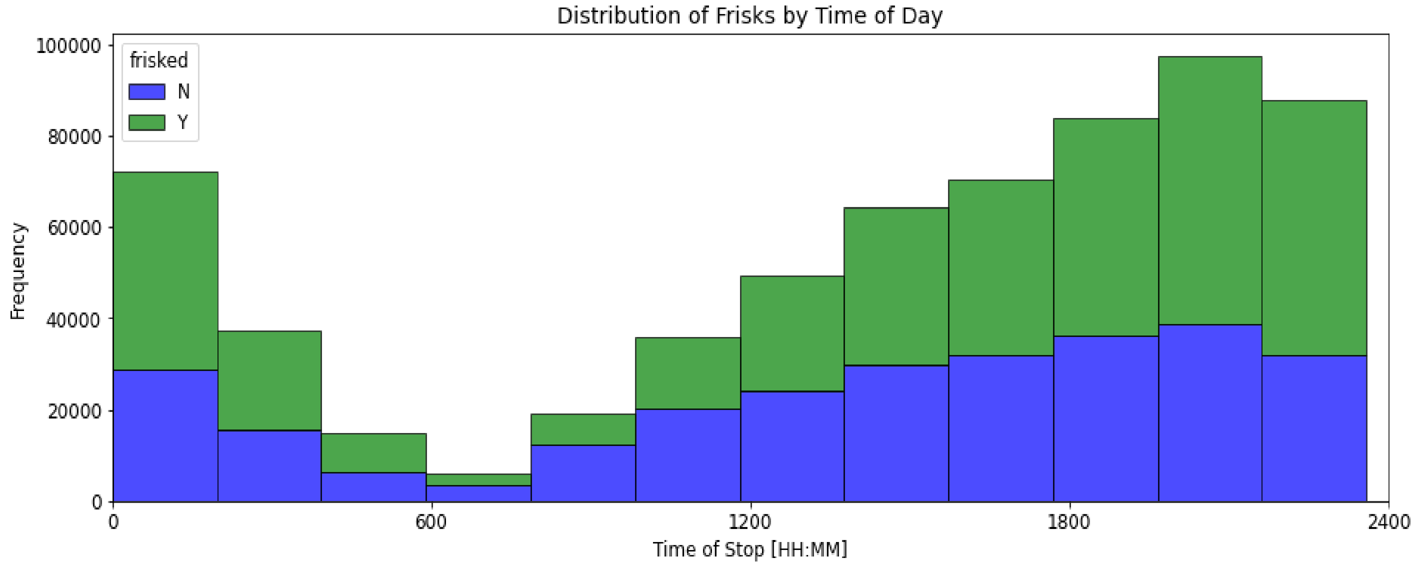
About 50% of individuals stopped by NYPD are black (B) or African American, followed by White-Hispanic (Q) as illustrated in figure below:



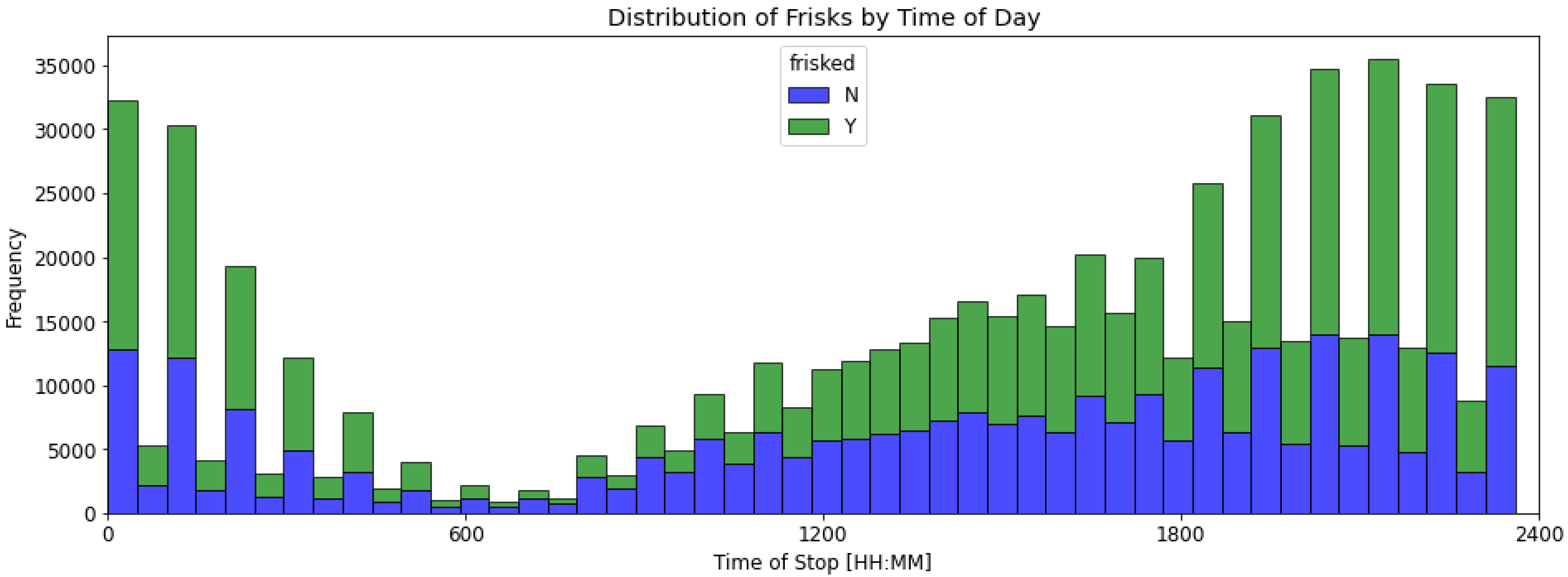
Task 3:

- When does Stop and Frisk happen the most?

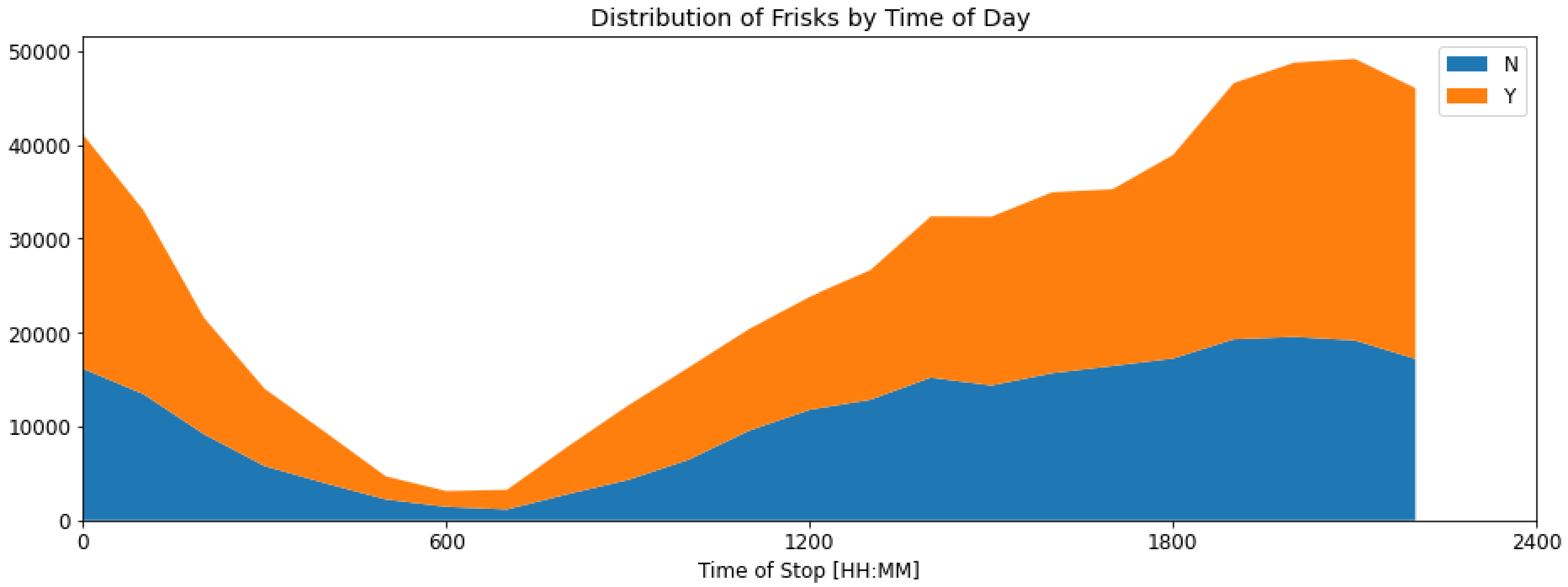
Based on the histogram bars, as shown below, the time of stop has a correlation with the likelihood of being frisked. Binning the time of stop into 12 bins would likely result in loss of information, as compared to 24 bins (but this is too many bins). Hence, I would leave them unbinned for now. It is also important to note that individuals are more likely frisked during the night (& midnight) than the day.



Task 3:


Distribution of Frisks by Time of Day
frisked
N
Y
Frequency
35000
30000
25000
20000
15000
10000
5000
0
0
600
1200
1800
2400
Time of Stop [HH:MM]

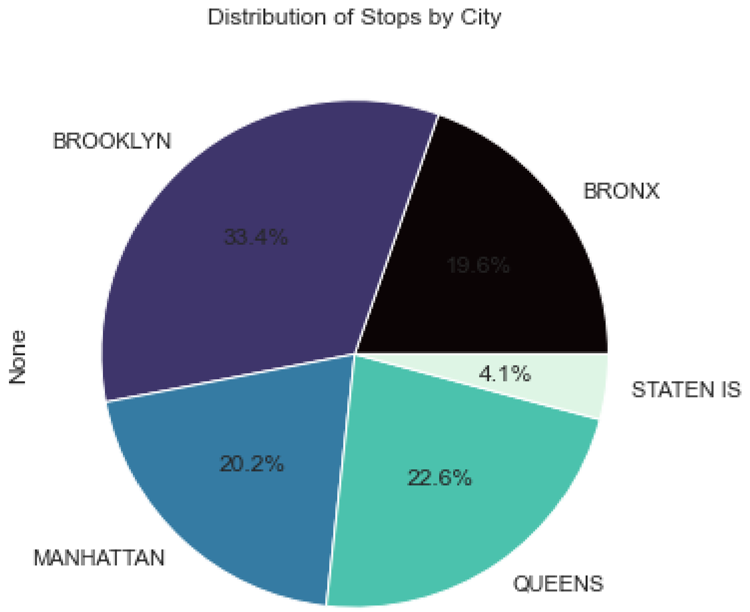
Task 3:


Distribution of Frisks by Time of Day
50000
40000
30000
20000
10000
0
0
600
1200
1800
2400
Time of Stop [HH:MM]
N
Y

Task 3:

- Where in New York City do the stops occur?

Most of the police stops occurred in the city of Brooklyn (30%) followed by Queens (25%), Manhattan (20%) and Bronx (20%).



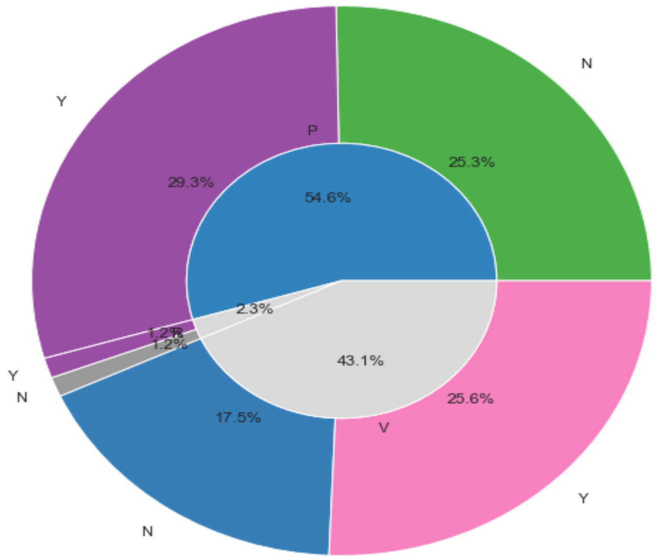
Task 3:

- Why do officers stop people?

There are many reasons for stop and frisked people below some of them with nested pie chart for explanation:

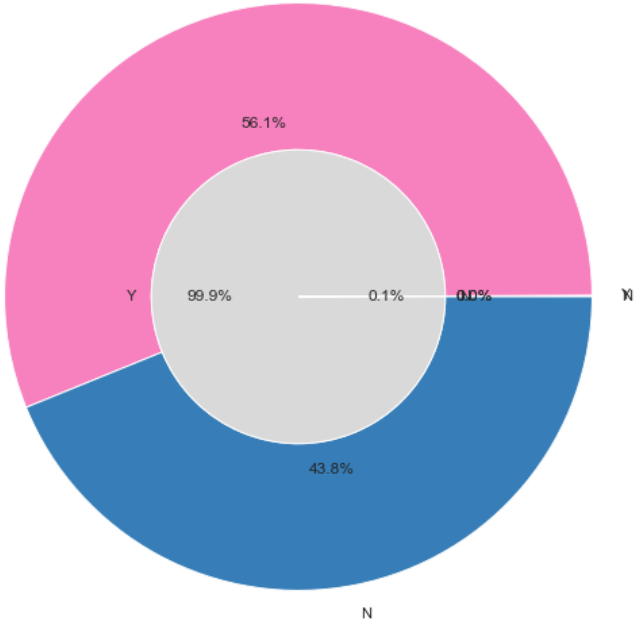
people who identify themselves (typeofid) photo id (P) is more likely to be frisked than if they had presented a verbally (V) or had refused (R).

typeofid	P	R	V
frisked			
N	0.252849	0.011544	0.174513
Y	0.292749	0.011876	0.256469
All	0.545599	0.023419	0.430982



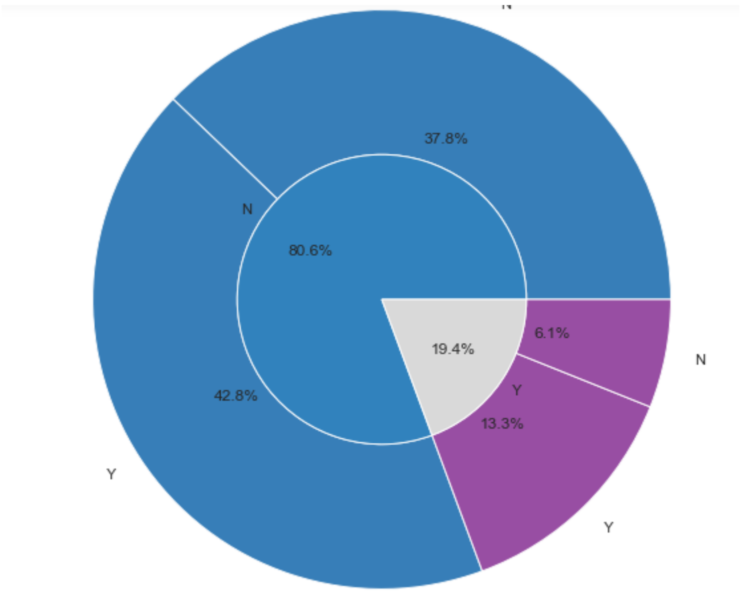
officers who explain their reason for stopping the individual (explnstp) are more likely to frisk them than if they had not explained.

explnstp	N	Y
frisked		
N	0.000409	0.438497
Y	0.000393	0.560701
All	0.000802	0.999198

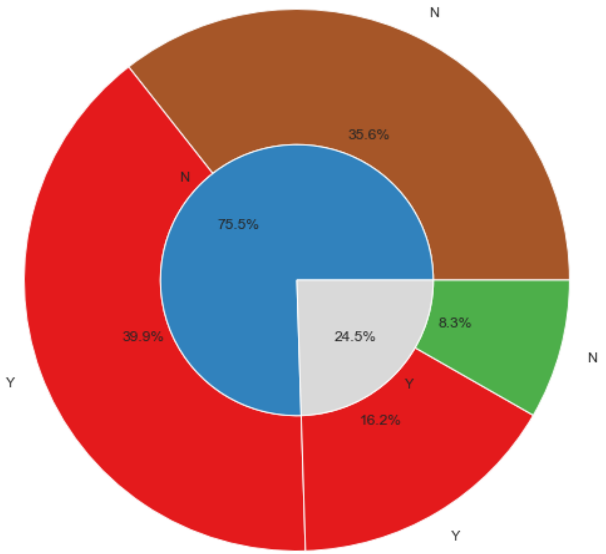


individuals who display evasive response when questioned (ac_evasv) and change direction at the sight of officer (ac_cgdir) are not more likely to be frisked.

ac_evasv	N	Y
frisked		
N	0.378399	0.060506
Y	0.428075	0.133020
All	0.806474	0.193526



ac_cgdir	N	Y
frisked		
N	0.355801	0.083105
Y	0.399230	0.161865
All	0.755031	0.244969

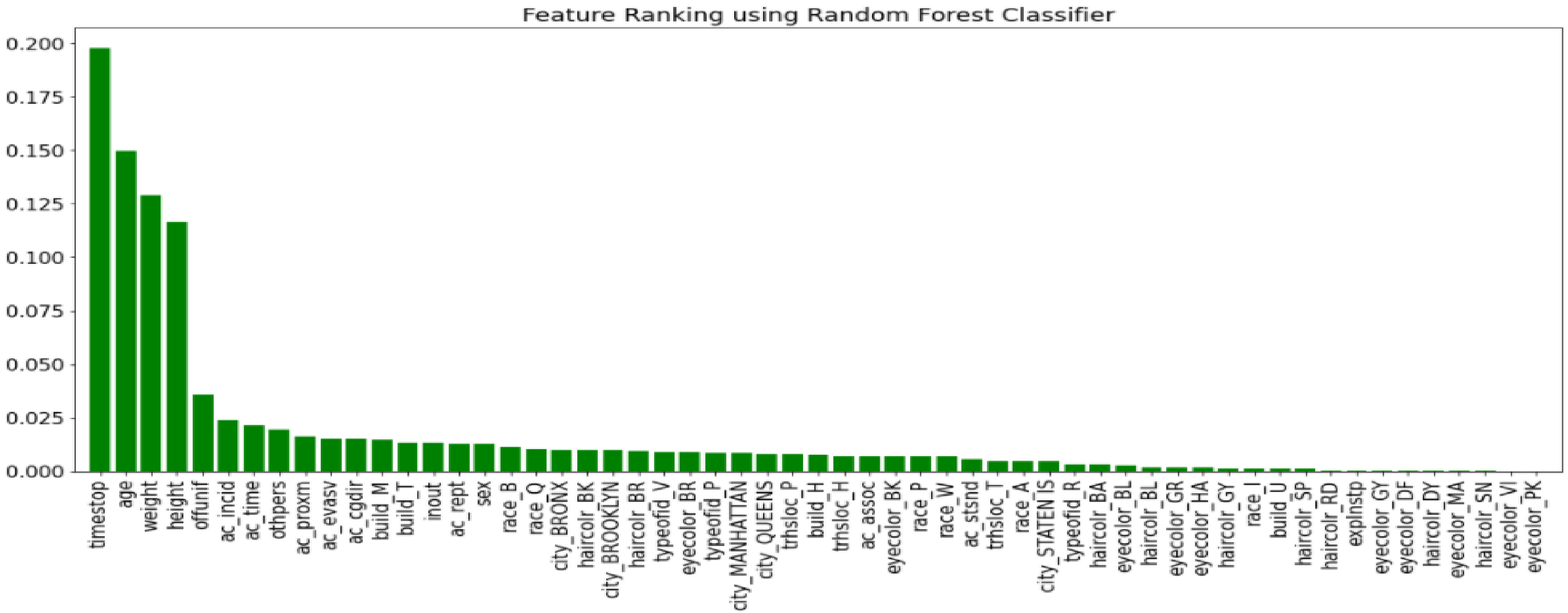


- What are the most important factors contributing to a person getting frisked?

We performed feature selection using random forest classifier to determine the most important factors contributing to a person getting frisked, the figure below shows the result after random forest classifier:

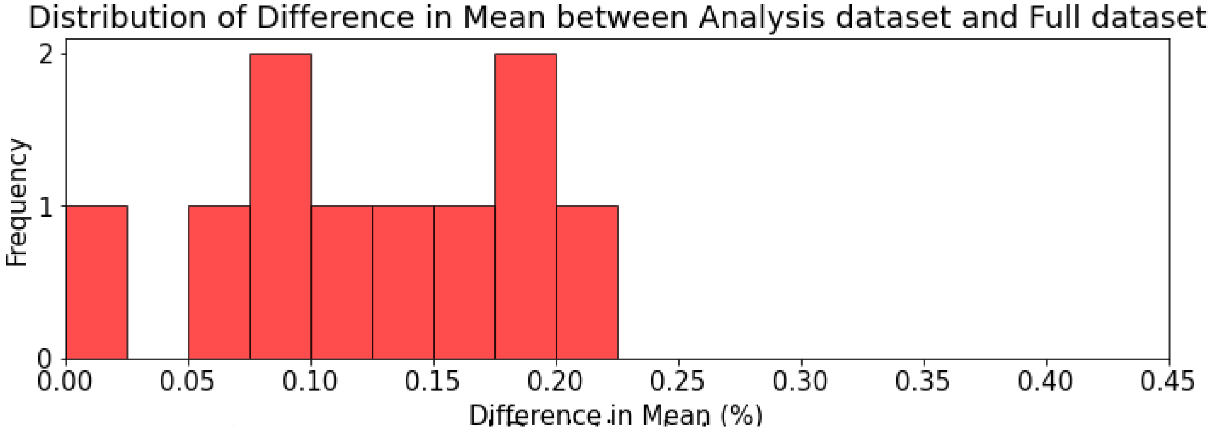
We can see that there is a jump in the feature importance score after the 4th-ranked feature

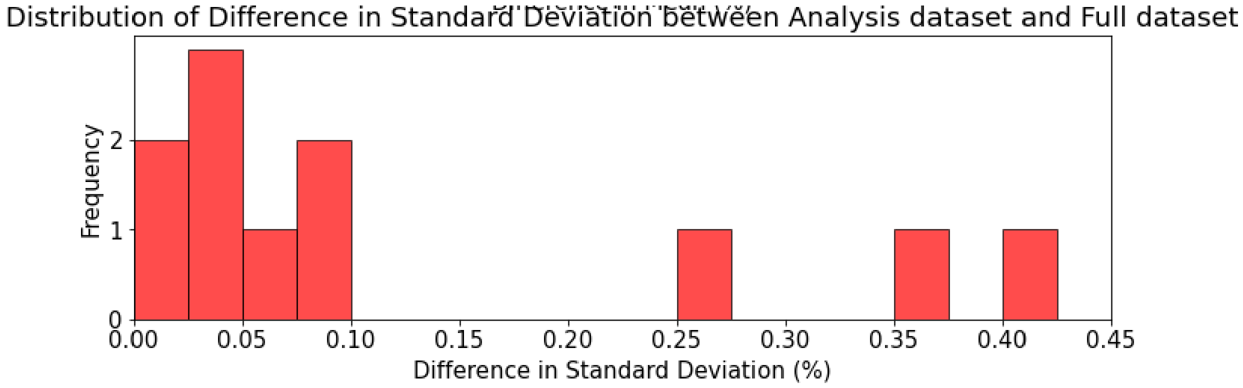
We decided to keep the top 10 features and drop the rest of the features, as these constitute more than 70% of the total feature importance. So, Number of features will have removed equal to 47



Task 3:

We performed data sampling due to large size of dataset to ensure our models can be run in a reasonable amount of time, given the dataset size of more than 600,000 records. The figures below illustrate the ratio differences in mean and slandered deviation between reduced and full dataset.

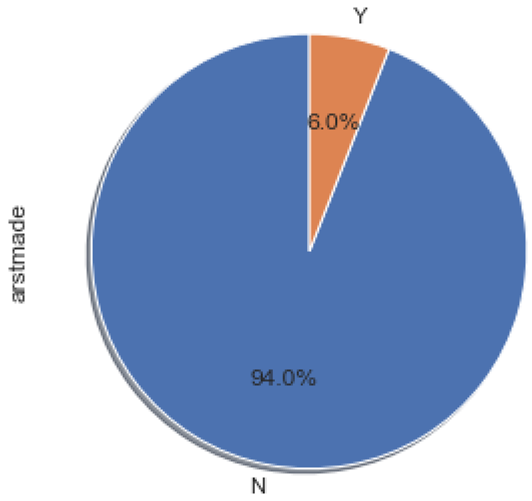




Task 3:

- Use a classification model to predict whether a stop is “effective”. Define “effective”. Report AUC and ROC curve.

We will define an effective stop as one where an arrest is made after an officer makes a stop. In the dataset, about 94% of civilians are unnecessarily stopped. Only about 6% of people are arrested after being stopped. The figure below illustrates the distribution of ‘arstmade’ column over the data.



Task 3:

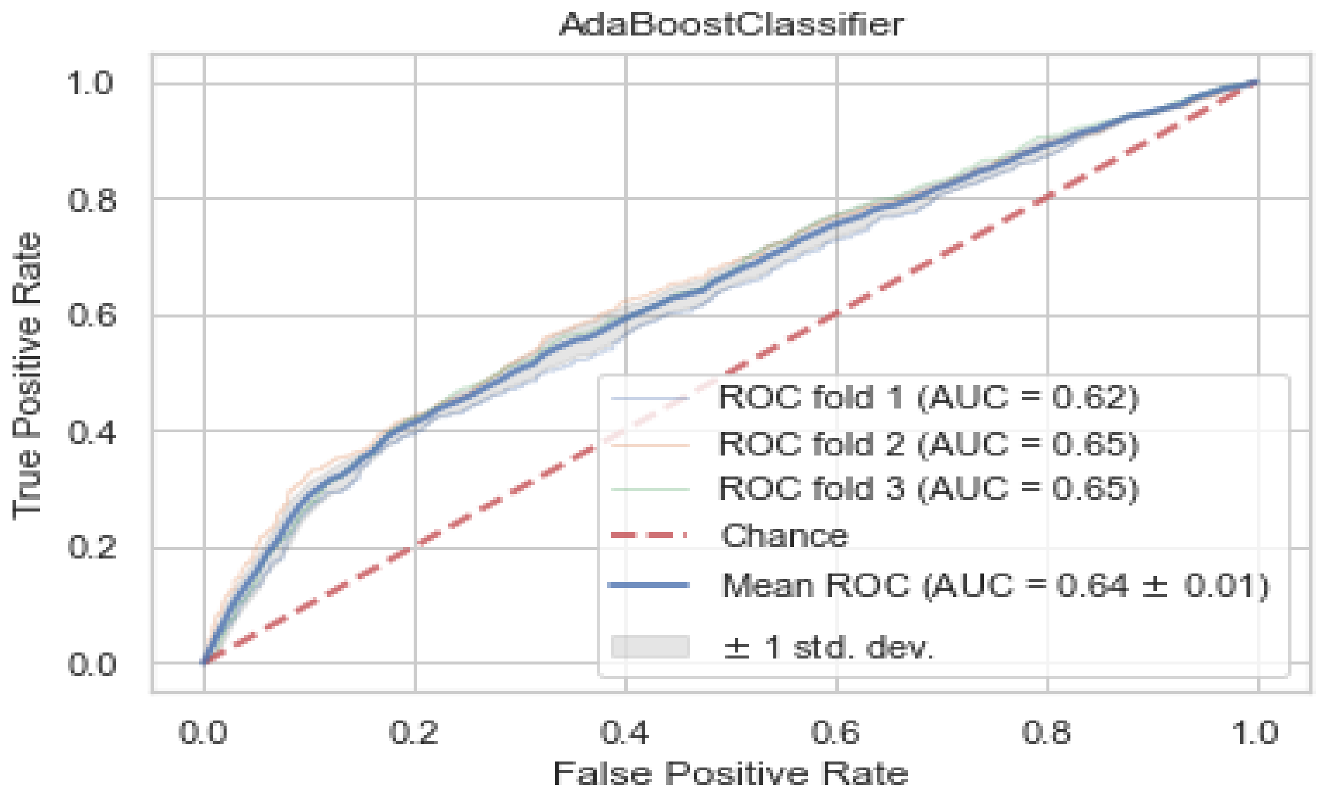
Given the imbalance of the data, with ~94% no-arrests, we realized that accuracy would not be the best metric for success. Instead, we chose to plot the ROC curve and compare the respective AUC values.

Additionally, we thought that the classifiers would perform better if we balanced the data. The reason for this assumption is that the prior distribution of the classes would lead to significantly higher posterior probabilities for no-arrest. We split the data into two groups: arrests and no-arrest. Then, we perform resampling technique in order to generate a balanced data set of arrests and no-arrests. After balancing the data, however, the AUC did not improve.

The classifiers used includes:

	accuracy	Precision	Recall_score	Specificity_list	True_pve Rate	False_pve Rate	F1_Score	AUC	Cohen's Kappa
ExtraTreesClassifier	0.584056	0.580245	61.006256	55.896564	61.006256	44.103436	0.594355	0.616113	0.523125
LogisticRegression	0.593090	0.563055	83.272993	35.436966	83.272993	64.563034	0.671497	0.619467	0.530109
KNeighborsClassifier	0.505733	0.505656	50.837643	50.358172	50.837643	49.641828	0.506720	0.507271	0.443628
DecisionTreeClassifier	0.536126	0.536325	52.675972	54.507323	52.675972	45.492677	0.531456	0.535914	0.474386
GaussianNB	0.592396	0.559939	86.043197	32.454433	86.043197	67.545567	0.678263	0.617728	0.529058
BaggingClassifier	0.565996	0.555751	65.992008	47.283042	65.992008	52.716958	0.603043	0.586482	0.503668
AdaBoostClassifier	0.603337	0.579717	75.455332	45.348043	75.455332	54.651957	0.655146	0.638975	0.541836
RandomForestClassifier	0.591873	0.585828	62.921446	55.545029	62.921446	44.454971	0.606327	0.627738	0.531120
QuadraticDiscriminantAnalysis	0.597085	0.564303	84.891024	34.537051	84.891024	65.462949	0.677823	0.631413	0.534152
VotingClassifier(DTC)	0.528484	0.528528	52.248204	53.406499	52.248204	46.593501	0.525435	0.528273	0.466608

Task 3:


AdaBoostClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 1 (AUC = 0.62)
ROC fold 2 (AUC = 0.65)
ROC fold 3 (AUC = 0.65)
Chance
Mean ROC (AUC = 0.64 ± 0.01)
± 1 std. dev.

Task 3:

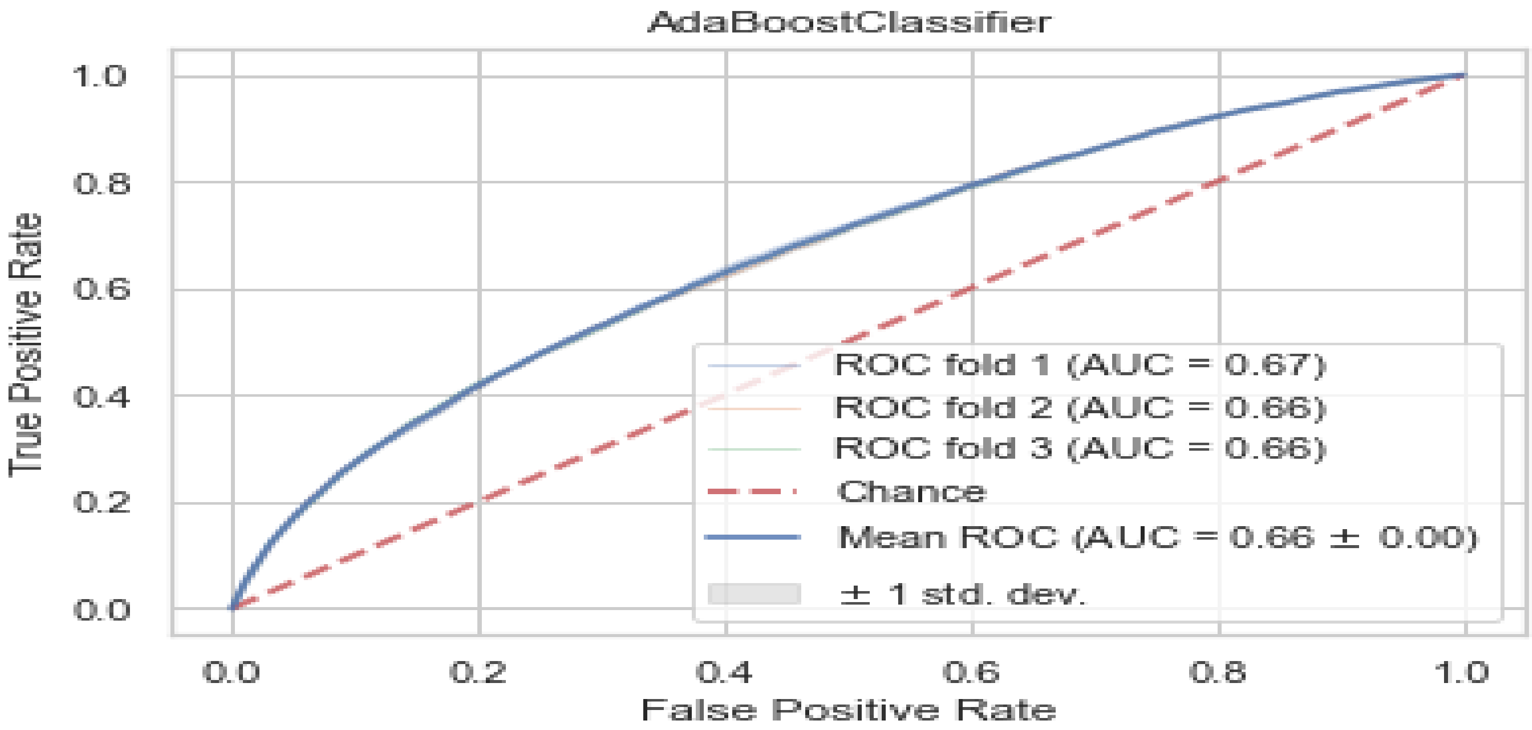
- Use a classification model to tell whether someone will be frisked or not? Report AUC and ROC curve

As we do in the previous task, we perform resampling technique in order to generate a balanced data set of arrests and no-arrests. After balancing the data, however, the AUC did not improve.

The classifiers used includes:

	accuracy	Precision	Recall_score	Specificity_list	True_pve Rate	False_pve Rate	F1_Score	AUC	Cohen's Kappa
ExtraTreesClassifier	0.587537	0.585115	60.777759	56.734547	60.777759	43.265453	0.596144	0.619638	0.531826
LogisticRegression	0.606960	0.591292	69.823799	51.541843	69.823799	48.458157	0.640239	0.649609	0.551793
KNeighborsClassifier	0.533027	0.534578	52.536620	54.081243	52.536620	45.918757	0.529882	0.542355	0.476480
DecisionTreeClassifier	0.544214	0.545300	54.262986	54.582130	54.262986	45.417870	0.543935	0.544263	0.487727
GaussianNB	0.608079	0.588961	72.069725	49.500723	72.069725	50.499277	0.648178	0.650341	0.552924
BaggingClassifier	0.574278	0.565304	65.079497	49.754923	65.079497	50.245077	0.604995	0.602529	0.518139
AdaBoostClassifier	0.613530	0.606620	65.040516	57.657653	65.040516	42.342347	0.627718	0.661406	0.558702
RandomForestClassifier	0.590988	0.588163	61.290187	56.912571	61.290187	43.087429	0.600182	0.629585	0.535377
QuadraticDiscriminantAnalysis	0.608174	0.589829	71.552743	50.039632	71.552743	49.960368	0.646579	0.650658	0.553030
VotingClassifier(DTC)	0.546166	0.547345	54.327056	54.905946	54.327056	45.094054	0.545288	0.546252	0.489699

Task 3:


AdaBoostClassifier
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 1 (AUC = 0.67)
ROC fold 2 (AUC = 0.66)
ROC fold 3 (AUC = 0.66)
Chance
Mean ROC (AUC = 0.66 ± 0.00)
± 1 std. dev.