Problem Set #3

Advanced Methods in Applied Statistics

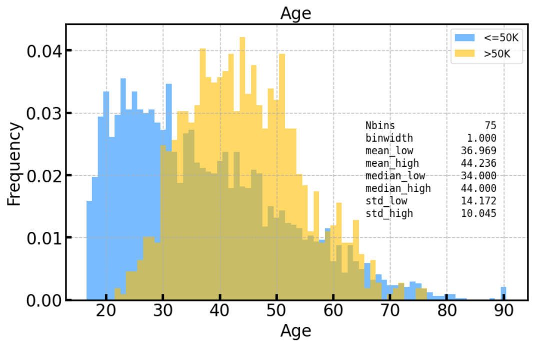
Cyan Yong Ho Jo

13th March 2024

EX1

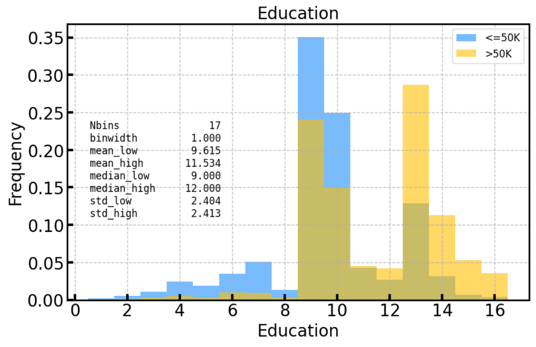
Quick look

The training dataset comprises four features: age, education, gender, and working hours per week. Initially, I opt to examine the data distribution concerning these features, which are depicted in Figures 1, 2, 3, and 4, respectively. Subsequently, I establish a null hypothesis for the classification, positing that the classified groups earn more than 50k. Consequently, the alternative hypothesis posits that the other group earns 50k or less. A quick investigation of the data impresses me that discerning the dataset into two distinct groups might not be simple or intuitive. So I shall consider utilising several classifiers instead of one. My selected classifiers include the following four: Adaptive Boosting, Logistic Regression, XGB, and Gradient Boosting.



```
Nbins            75
binwidth      1.000
mean_low     36.969
mean_high    44.236
median_low   34.000
median_high  44.000
std_low      14.172
std_high     10.045
```

Figure 1: The distribution of groups earning less than 50k and more than 50k, categorized by age. Each histogram bar is binned in intervals of 1



```
Nbins            17
binwidth      1.000
mean_low      9.615
mean_high    11.534
median_low    9.000
median_high  12.000
std_low       2.404
std_high      2.413
```

Figure 2: The distribution of groups earning less than 50k and more than 50k, categorized by education. Each histogram bar is binned in intervals of 1.

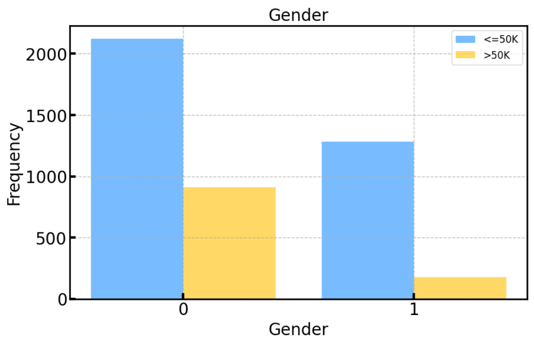
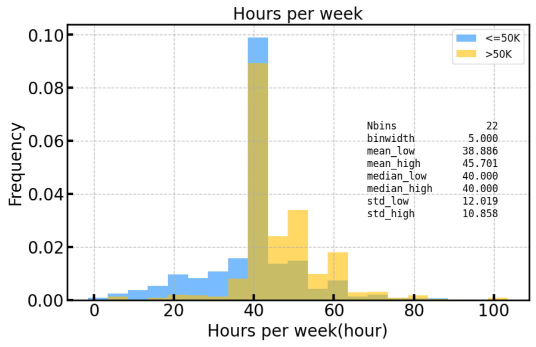


Figure 3: The distribution of groups earning less than 50k and more than 50k, categorized by gender.



```
Nbins            22
binwidth      5.000
mean_low     38.886
mean_high    45.701
median_low   40.000
median_high  40.000
std_low      12.019
std_high     10.858
```

Figure 4: The distribution of groups earning less than 50k and more than 50k, categorized by the working hours per week. Each histogram bar is binned in intervals of 5 hours.

Classification

To train a classifier, the dataset must be divided into two subsets, each serving a distinct purpose: a training set and a test set. This division is executed by the train_test_split function from the sklearn.model_selection module in Python, which utilises a random selection process to segregate the data. Both high-earners and low-earners must be represented by more than 500 data points each. Given the smaller number of high-earning individuals, particular attention is paid to the size of the high-earning group. The proportion of the test group is set to 45% of the total dataset. Consequently, within the testing group of 2025 individuals, there are 506 high-earners.

Overtraining check

Since there is no guarantee that the dataset is not overtrained, comparing the training and testing sets is crucial. The outcome of this comparison is illustrated in TableI. The classifier now scores each data item against a given threshold. Setting thresholds for classifiers is challenging due to the different distributions of decision scores produced by each classifier. A function is defined to find the optimal threshold, however, it is required to adjust the threshold manually based on visual inspection of histograms. The precision achieved by each classifier is then determined. Algorithms such as Adaptive Boosting, XGB, and Gradient Boosting meet the precision criteria. Logistic Regression, however, fails to meet the 85% condition. These precisions are summarised in TableII.

	> 50k		**≤ 50k**	
	Test Group	**Train Group**	**Test Group**	**Train Group**
Mean	43.70	44.70	36.67	37.21
Median	44	43	34	35
Std Dev	9.92	10.14	14.35	14.02
Mean	11.49	11.57	9.63	9.60
Median	12	12	9	9
Std Dev	2.39	2.43	2.43	2.38
Mean	45.77	45.64	38.98	38.81
Median	40	44	40	40
Std Dev	10.51	11.16	11.99	12.05

Table I: No sign of overtraining. Basic statistics of the test and train groups will be classified into two groups by four different classifiers. The training and testing groups have similar statistical characteristics. Comparing the statistical character helps prevent overtraining.

	Adaptive boost	**Logistic regression**	**XGB**	**Gradient boost**
Precision	85.71%	81.48%	88.89%	89.29%

Table II: Precision of the classifications by the four different classifiers.

Confusion matrix and decision scores

The confusion matrix results from classification tasks, comprising counts of true negatives, false positives, false negatives, and true positives. These quantities evaluate the classifier's performance, and certain combinations of these values are employed to assess metrics such as precision. The confusion matrices for all classifications are tabulated in Table?? and visualised in Figures5, 6, 7, and 8 respectively. For Adaptive boosting, working hours, age, and education show moderate differences in predicting the high-earners and low-earners, while gender is considered a bad variable to guess. For Logistic regression, however, gender marked the highest score for the differentiator. XGB says education is the most important feature while the second and the third most important features are gender and age having similar importance scores. Lastly, Gradient boost shows that the two most effective features are education and age in prediction. This is summarised in TableIV.

Adaptive Boosting

	Negative	Positive
False	470	6
True	1513	36

Logistic regression

	Negative	Positive
False	484	5
True	1514	22

XGB classifier

	Negative	Positive
False	482	3
True	1516	24

Gradient Boosting

	Negative	Positive
False	481	3
True	1516	25

Table III: Confusion matrices for the classifications using Adaptive Boosting, Logistic Regression, XGBoost, and Gradient Boosting.

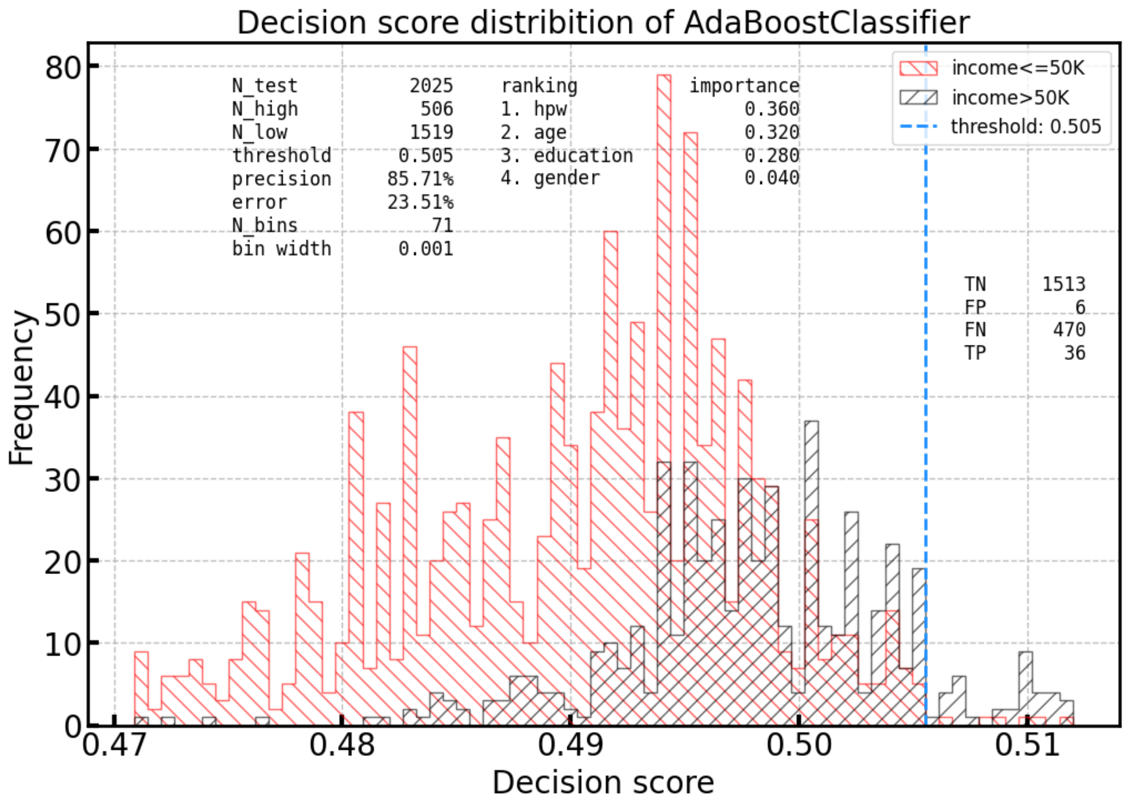


Figure 5: The Decision score distribution of the classification by Adaptive boost.

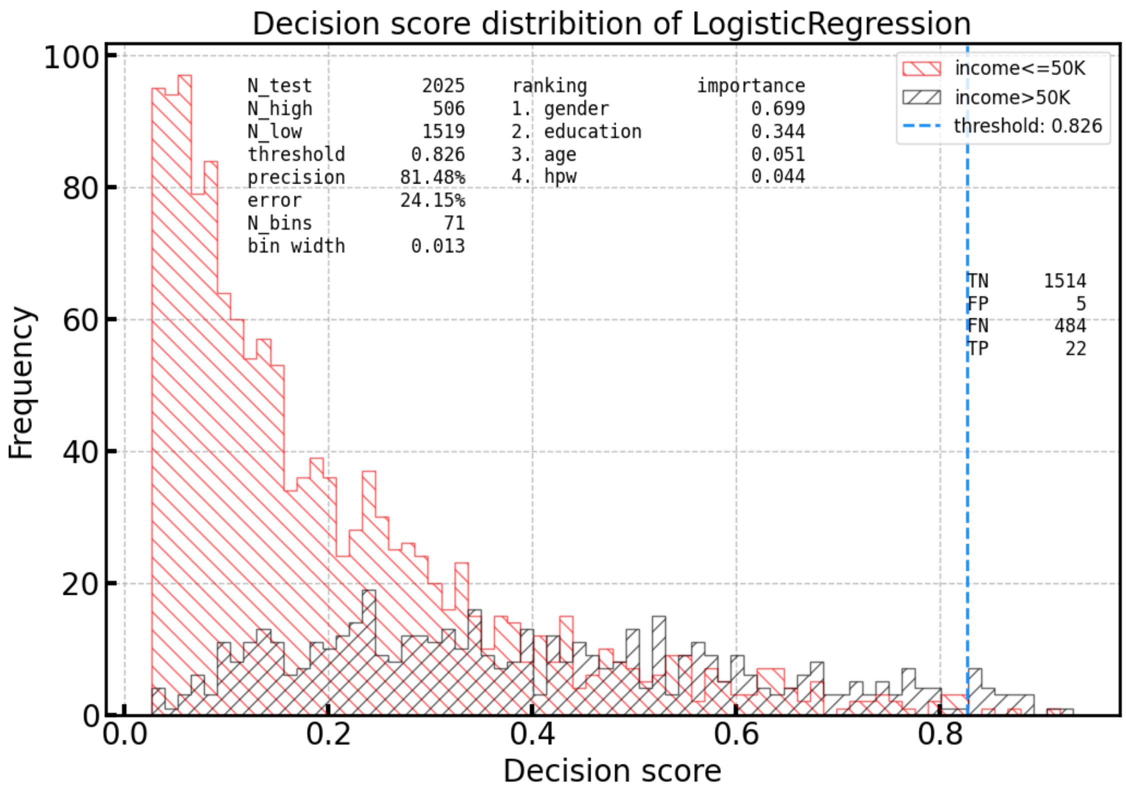


Figure 6: The Decision score distribution of the classification by Logistic regression.

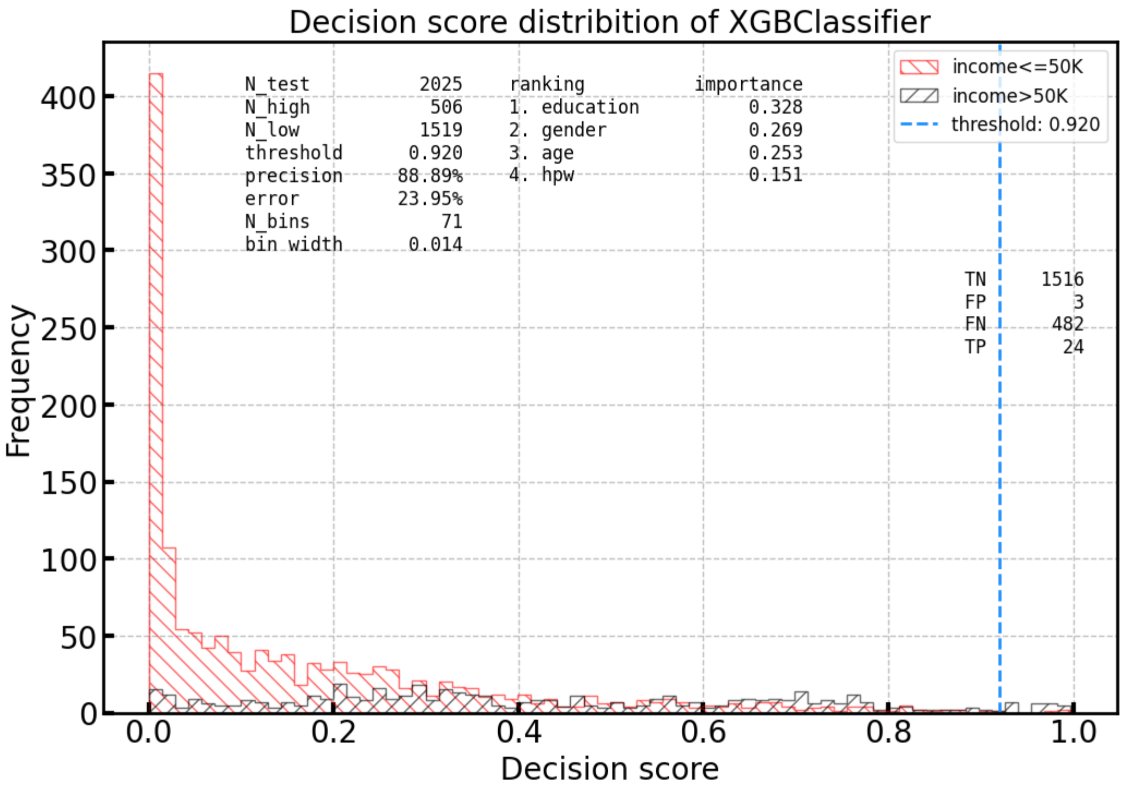


Figure 7: The Decision score distribution of the classification by XGB.

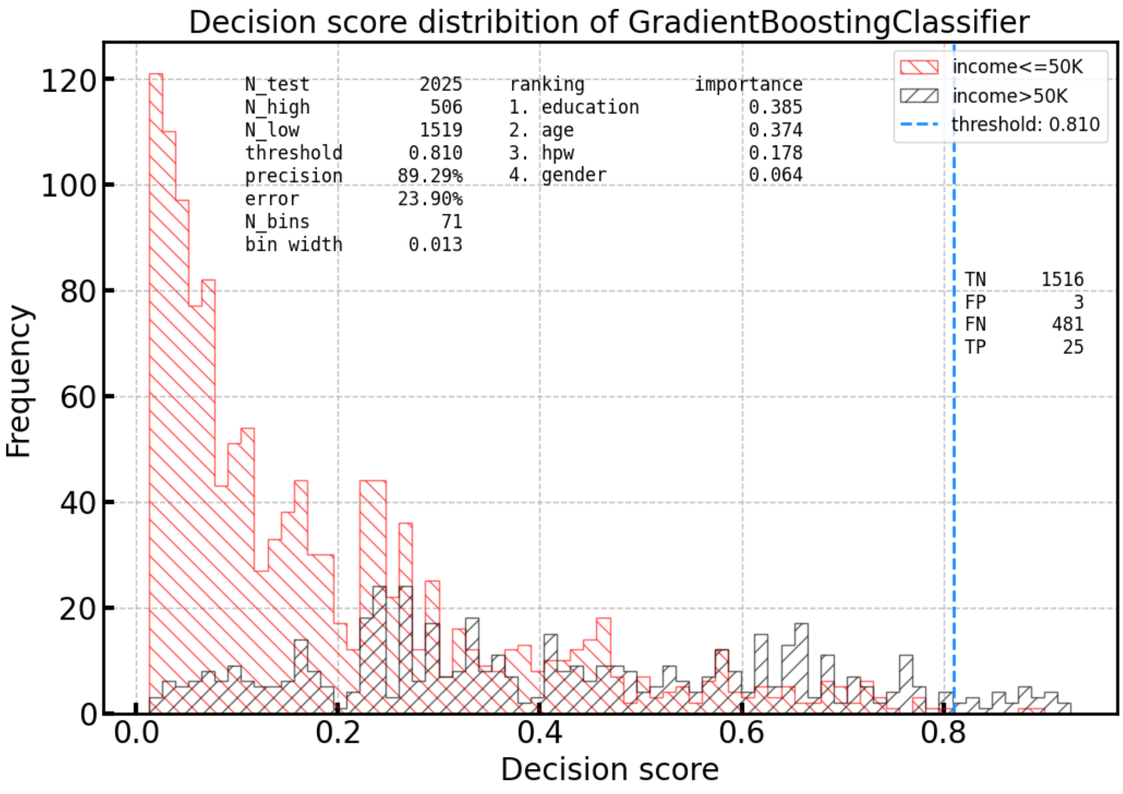


Figure 8: The Decision score distribution of the classification by Gradient boost.

	AdaBoost	**Logistic Regression**	**XGBoost**	**Gradient Boosting**
1	hpw: 0.360	gender: 0.699	education: 0.328	education: 0.385
2	age: 0.320	education: 0.344	gender: 0.269	age: 0.374
3	education: 0.280	age: 0.051	age: 0.253	hpw: 0.178
4	gender: 0.040	hpw: 0.044	hpw: 0.151	gender: 0.064

Table IV: Feature importance ranking and scores for different classifiers

Prediction on real data

The real dataset is processed using four classifiers having the same thresholds that they used in training and testing. The resulting prediction scores are presented in Figure9, 10, 11, and 12.

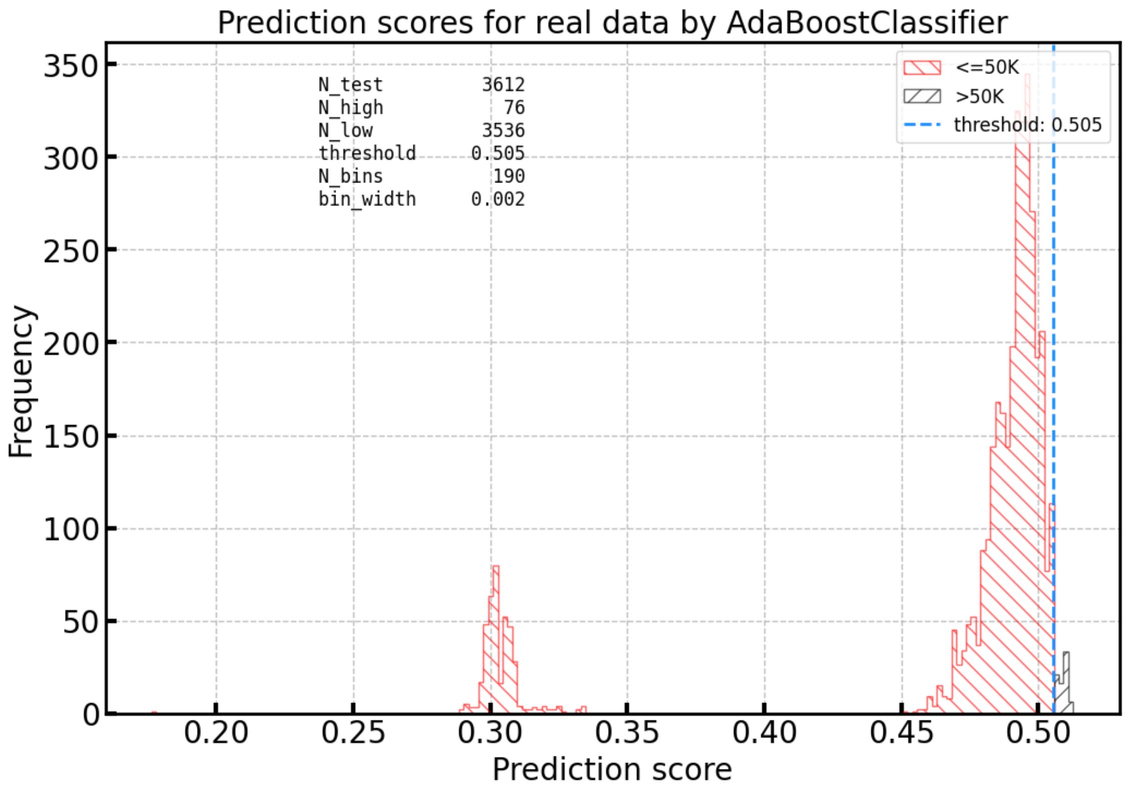


Figure 9: Distribution of prediction scores by Adaptive boost on real data, with the threshold indicating the decision boundary for income classification.

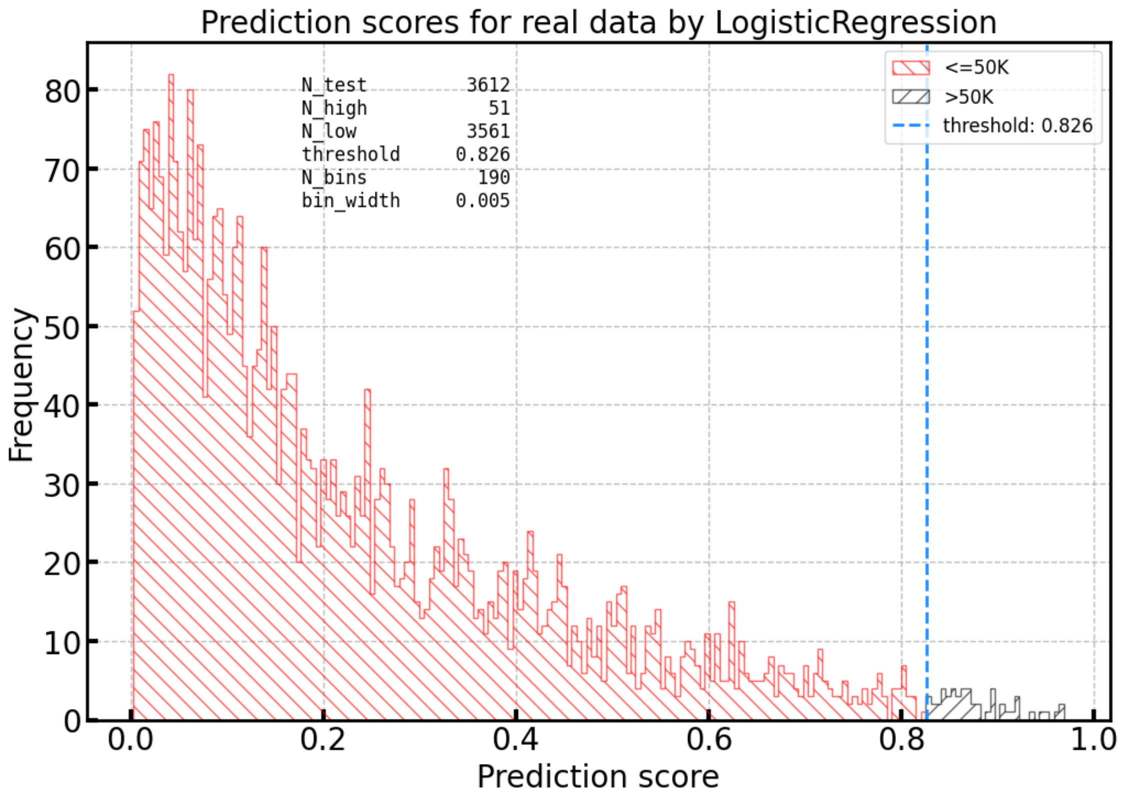


Figure 10: Logistic Regression's prediction score distribution on real data, with a marked threshold for classifying incomes.

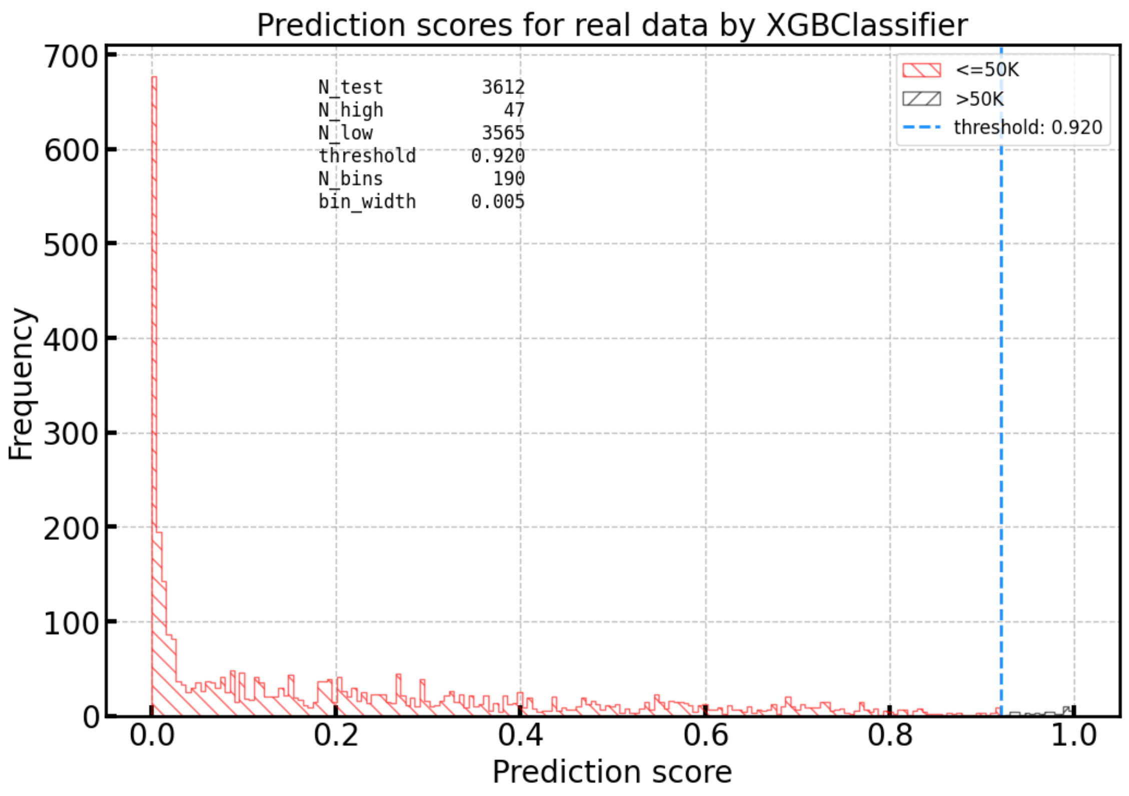


Figure 11: Prediction score distribution for real data using XGB, displaying the threshold for income predictions.

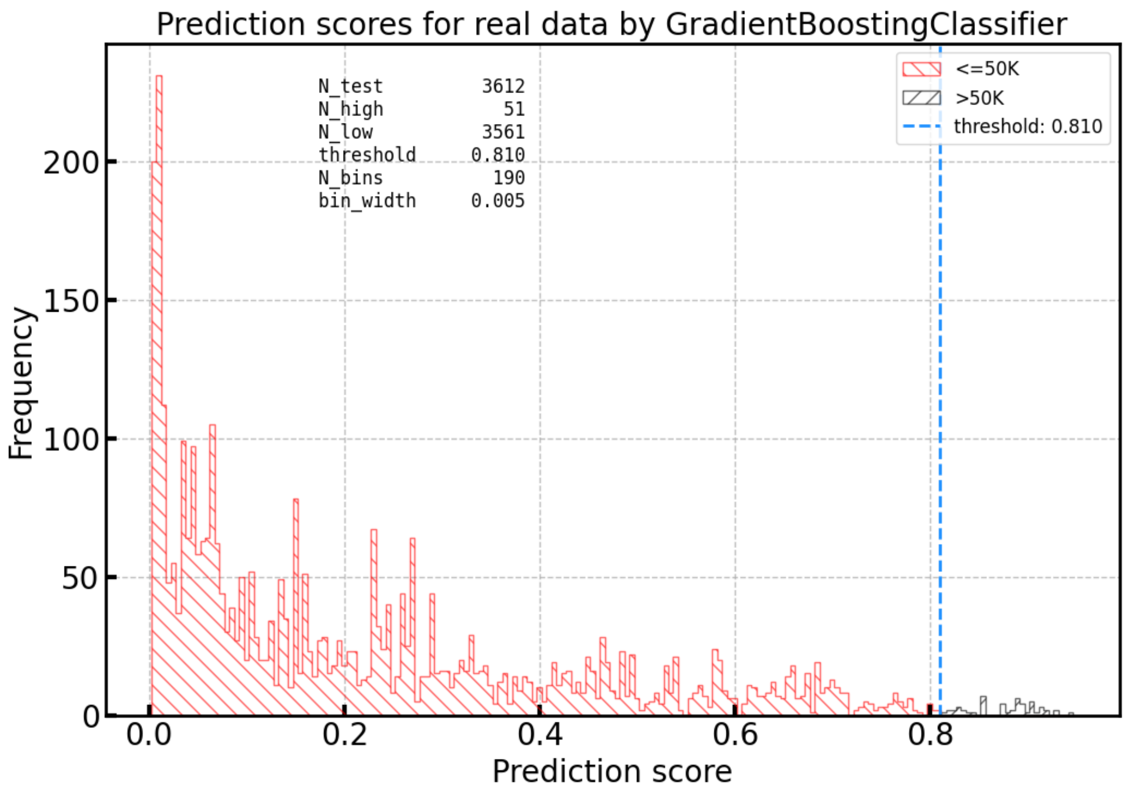


Figure 12: Gradient Boosting's prediction scores for real data, with the threshold delineating the classification of income levels.

EX2

Scatter of car accidents

The *lat*, *lon*, and *crash_date* columns of data contain some missing values, so the rows missing the longitude, latitude, and time should be removed before data processing. *crash_date* column has string format of date and time so the data in that column need to be parsed in numerical format.

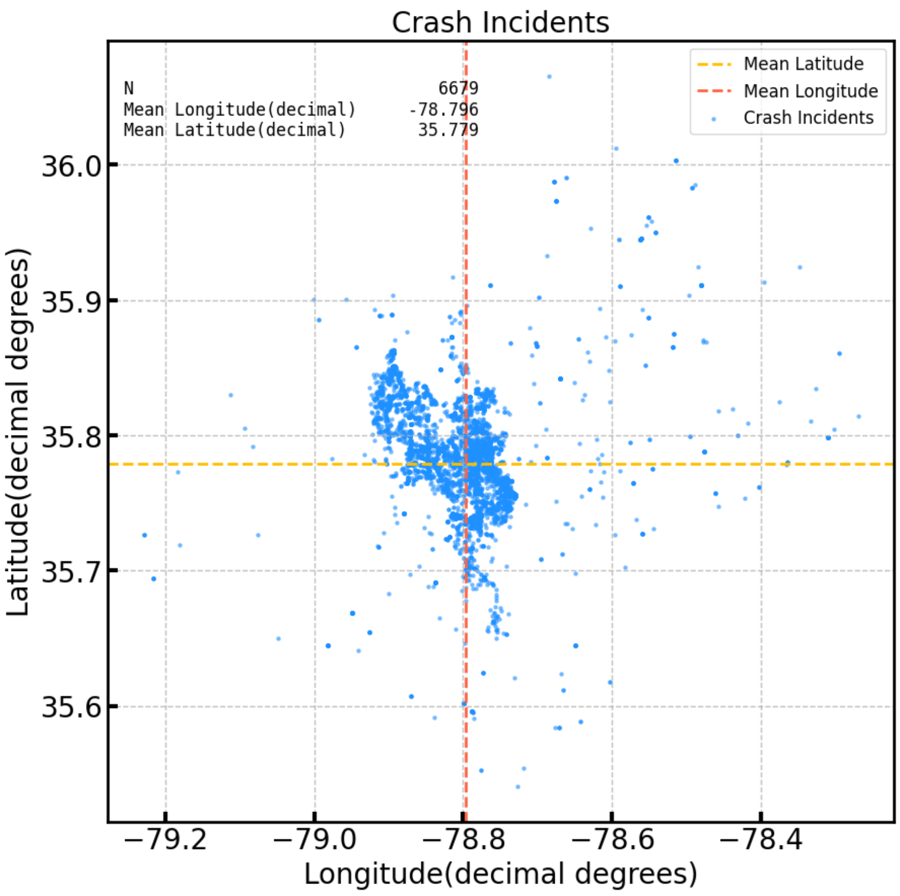


Figure 13: The scatter plot of the car accidents in the given data.

Histogram of time and Gaussian KDE

Since the daily time cycle repeats every 24 hours, the histogram's x-axis can be viewed as a loop. Thus, the data from either end of the axis, near 00:00 and 23:59, could be connected to the other side. Here the initially parsed data is plotted in Figure14. Manipulating the histogram to be adjusted such that its peak is centred horizontally. The adjusted histogram in Figure15 exhibits a dominant bell-shaped distribution with some hints of several minor peaks that may contribute to the total profile. By applying the Gaussian formula to each time data point in the adjusted histogram, we generate a Gaussian KDE that reflects the underlying distribution of the time-based events.

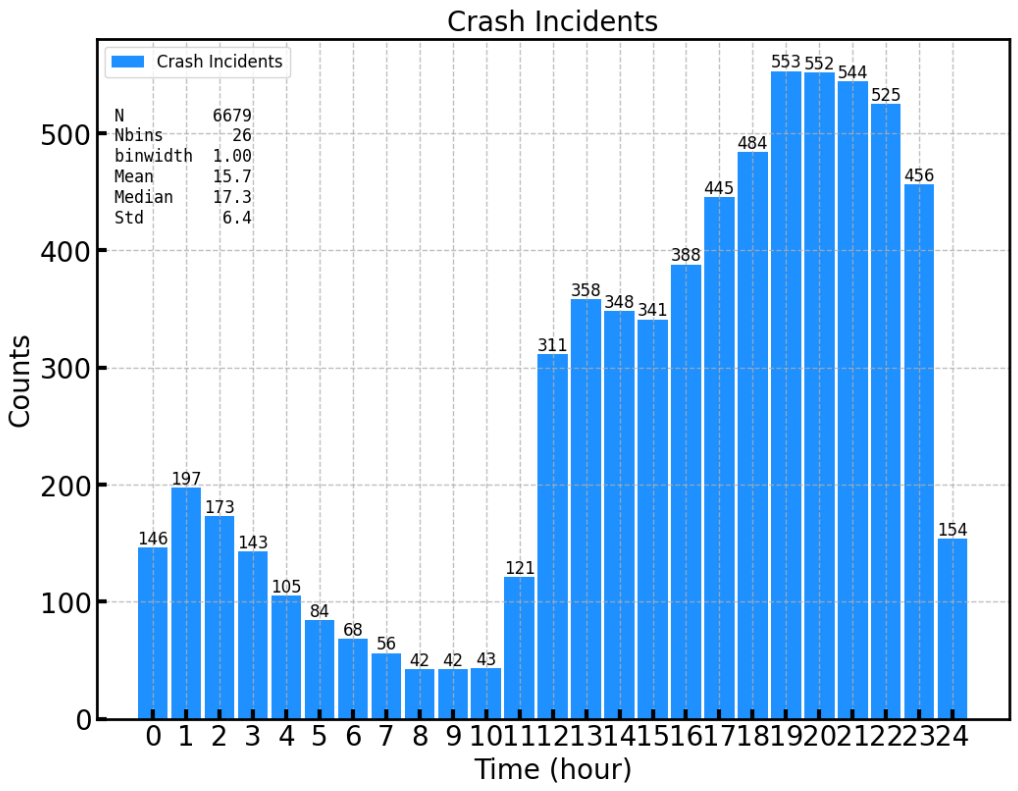


Figure 14: First figure

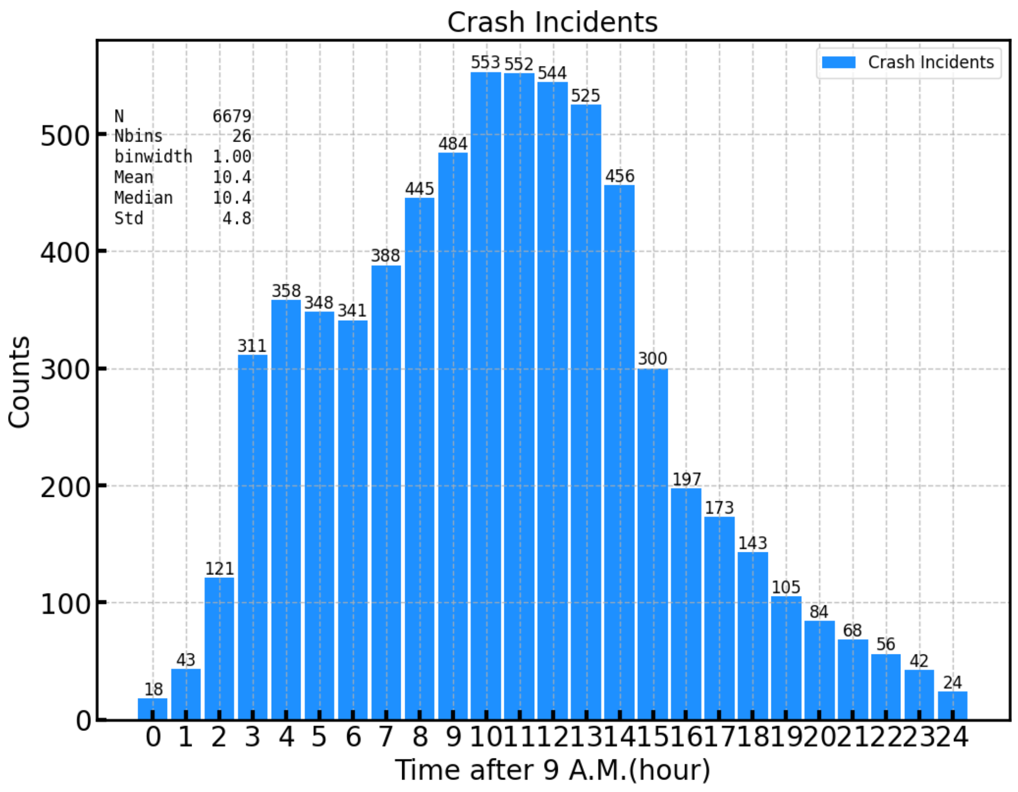


Figure 15: Second figure

Epanechnikov KDE for time

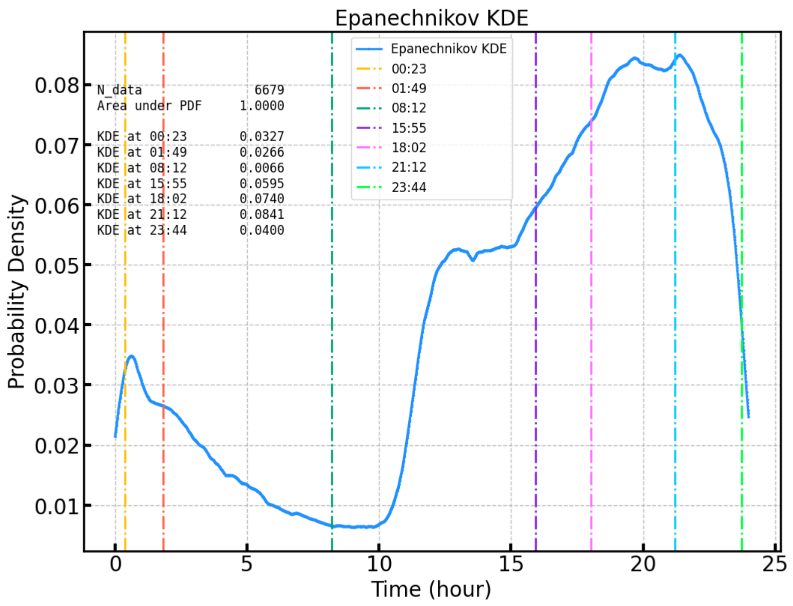
An Epanechnikov kernel density function is manually defined to calculate the values at each point. Summing up the values calculated by that function at each point and dividing by the total area will give a PDF or KDE. There are two data layouts that I used to give as the input to the function. One is the original distribution of the car accident and the other is the modification in the way I build Figure 15. As they have different distributions, KDE values are also slightly different. Figure16 shows the Epanechnikov KDE in its original distribution. Figure 17 shows the KDE of the modified data distribution. TableV as well as in the two figures.

Optimal patrol starting point

To optimise the start time for a two-hour patrol, I aggregated the KDE values within two-hour intervals. This is visualized in Figures 16 and 17, showing the sum of KDE values for each potential start time throughout the day. The analysis aimed to identify the period with the highest likelihood of activity, as indicated by the KDE values. The optimal patrol start time is 19:33 which is marked by a dashed vertical line in Figure 17, suggesting the most advantageous time to commence the patrol based on the data.

Time	KDE original	KDE modified
00:23	0.0327	0.0392
01:49	0.0266	0.0263
08:12	0.0660	0.0065
15:55	0.0595	0.0588
18:02	0.0740	0.0732
21:12	0.0841	0.0832
23:44	0.0400	0.0511

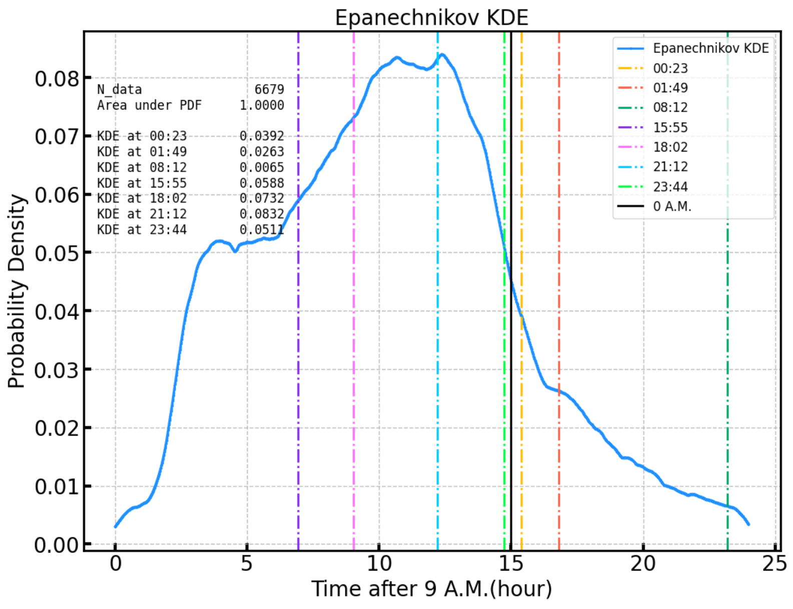
Table V: KDE Values at Specified Time Points



```
N_data                 6679
Area under PDF       1.0000

KDE at 00:23         0.0327
KDE at 01:49         0.0266
KDE at 08:12         0.0066
KDE at 15:55         0.0595
KDE at 18:02         0.0740
KDE at 21:12         0.0841
KDE at 23:44         0.0400
```

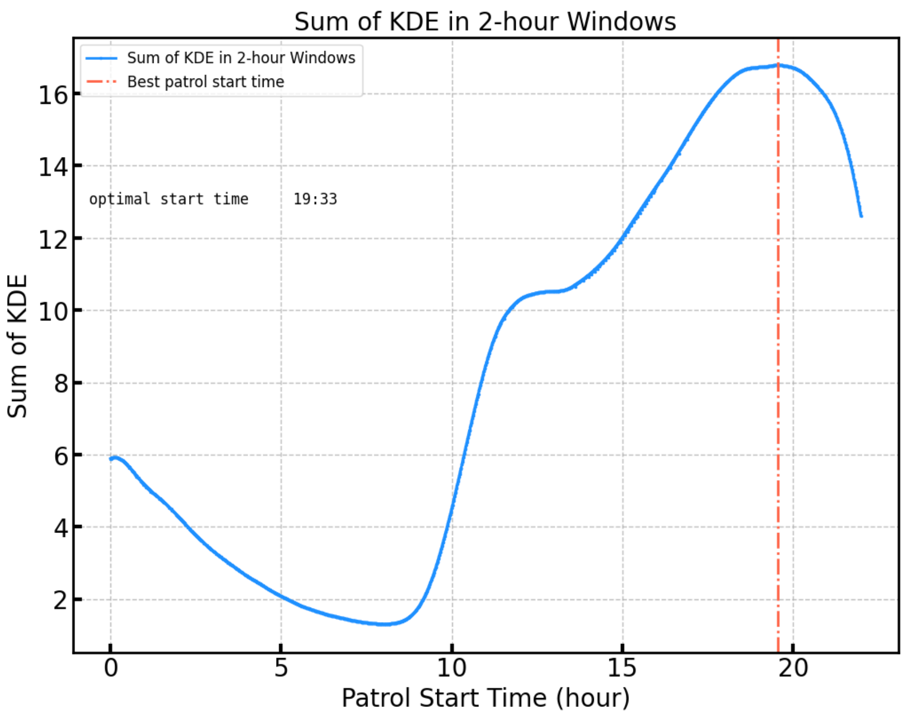
Figure 16: The original Epanechnikov KDE values over 24 hours. Each vertical line corresponds to a specific time point, and the height of the curve represents the KDE value, indicating the probability density of events.



```
N_data                   6679
Area under PDF         1.0000

KDE at 00:23           0.0392
KDE at 01:49           0.0263
KDE at 08:12           0.0065
KDE at 15:55           0.0588
KDE at 18:02           0.0732
KDE at 21:12           0.0832
KDE at 23:44           0.0511
```

Figure 17: The modified Epanechnikov KDE values over 24 hours. Each vertical line corresponds to a specific time point, and the height of the curve represents the KDE value, indicating the probability density of events.



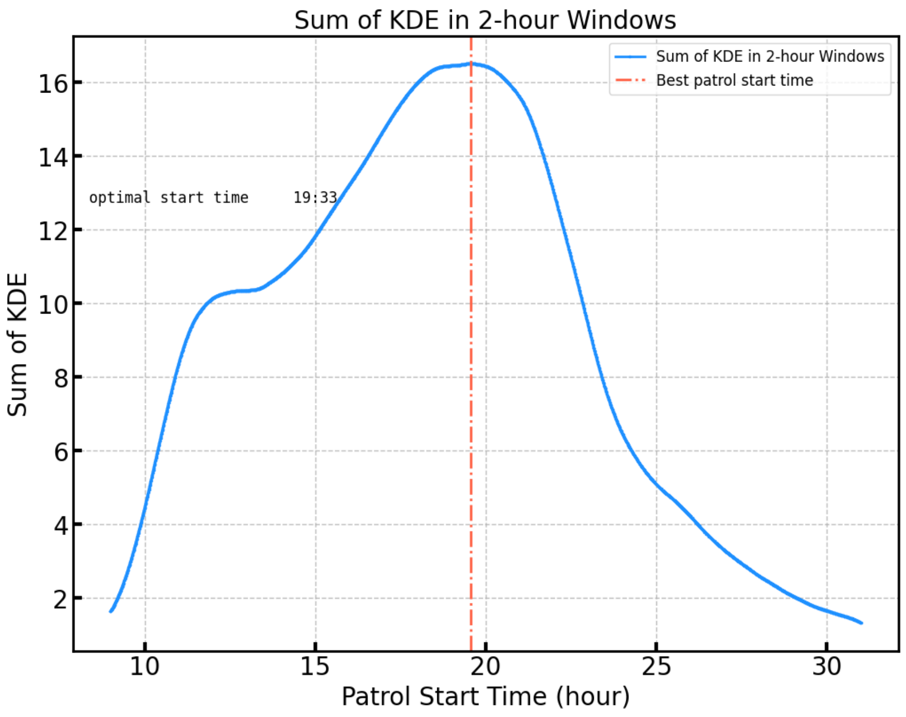


Figure 18: Cumulative sum of KDE values in two-hour windows plotted against potential patrol start times. The dashed vertical line indicates the optimal start time for a patrol, determined by the peak of the aggregated KDE values. The top panel is for the original KDE and the bottom is for the modified KDE.

Epanechnikov KDE map

The geographic distribution of Epanechnikov Kernel Density Estimate (KDE) values is depicted in Figure 19, which illustrates a broad view across various latitudes and longitudes. A region of interest is demarcated with a dashed line, where the concentration of KDE values is noticeably higher. In Figure 20, this region of interest is magnified to provide a closer examination of the spatial distribution of the KDE values. The proportion of the probability within the boxed region is significant at 13.39%, particularly when considering that this region covers only a small fraction of the map's total area, a mere 0.32%. This high probability density in a relatively minuscule area indicates a substantial clustering of the data points of interest, likely corresponding to a higher frequency or intensity of the studied phenomenon in that specific location.

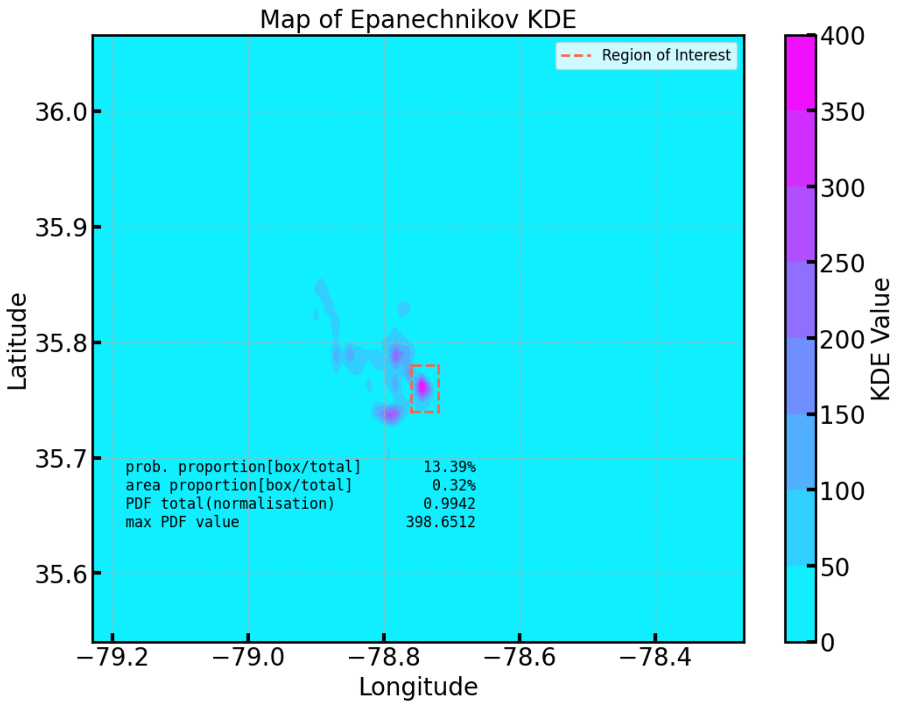


Figure 19: The geographical map representing the Epanechnikov KDE values. A dashed line encloses the region of interest, highlighting an area with a relatively high density of the accidents across the latitude and longitude coordinates.

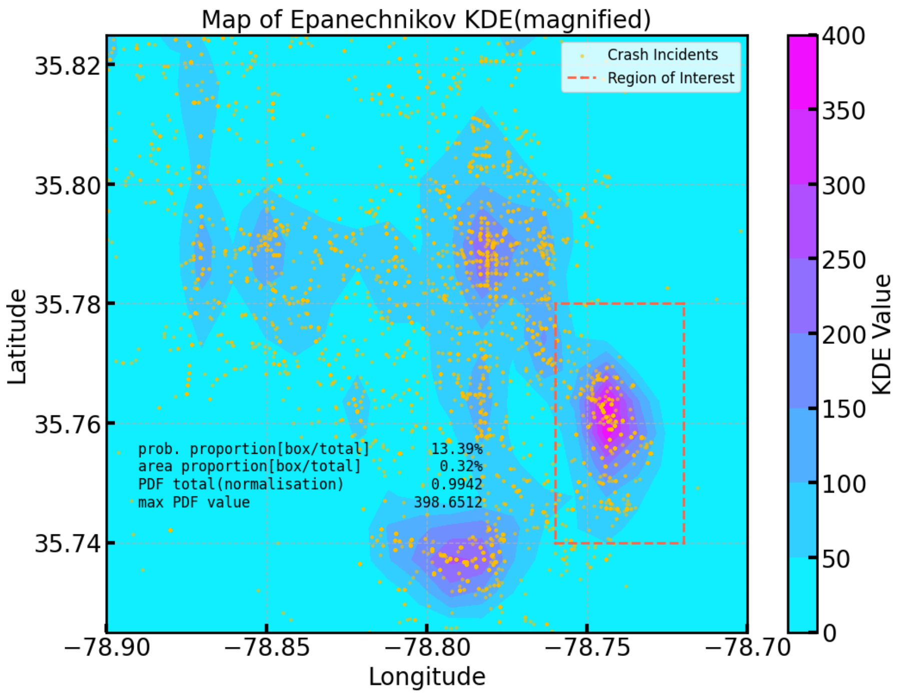
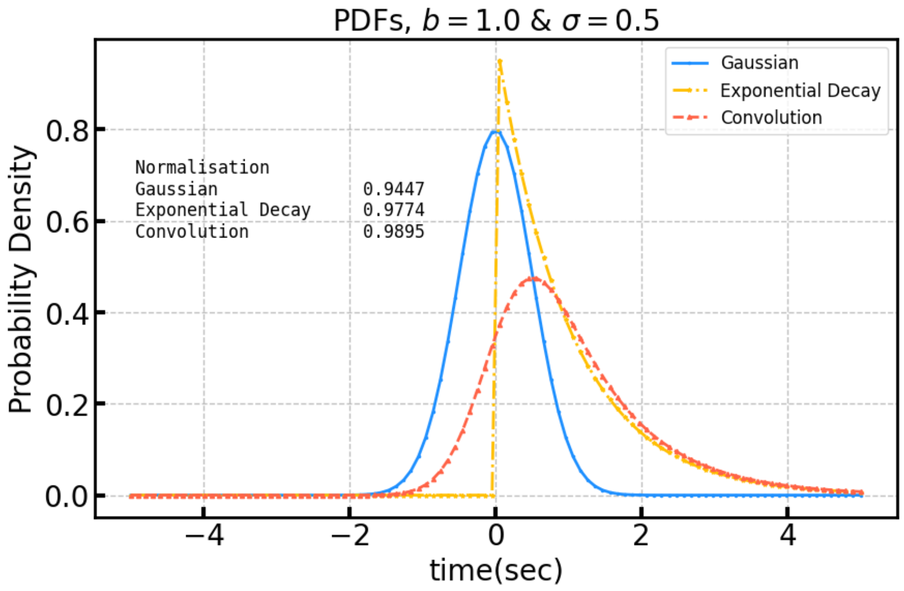


Figure 20: A magnified view of the region of interest within the geographical map of Epanechnikov KDE values. The coloured points represent the accidents, emphasising the concentration of events within this locale.

EX3

The convolution of the PDFs

The calculation of the convolution of two PDFs is computationally demanding since it entails definite integration. The Gaussian, and exponential decay and their convolution are plotted in Figure21. They all are PDFs so they must be normalised.



```
Normalisation
Gaussian                    0.9447
Exponential Decay           0.9774
Convolution                 0.9895
```

Figure 21: The figure displays three probability density functions (PDFs): the exponential decay and Gaussian distributions, considered seed PDFs with initial guess values of $b = 1$ and $\sigma = 0.5$. The red curve represents the convolution of these two seed PDFs, resulting in another PDF. The normalisation of all three PDFs has been verified.

Calculating likelihood and finding minimum

For hypothesis testing, I consider a null hypothesis where $b = 1$ and an alternative hypothesis where $b \neq 1$. To determine the most probable values of the remaining free parameters b(for the alternative) and σs(for both), I conduct a log-likelihood ratio test. Likelihood values are calculated for 100 different data sets, each containing 200 data points, employing the specified convolved PDF. The calculation of each likelihood value necessarily involves integration, making the processing time-consuming. In analysing the aggregate data for the likelihood computations, I search for the parameter(s) that minimizes the negative likelihood.

Minimisation: Minuit vs. Raster scan

The minimisation process, which utilised the *iminuit* package, was unsuccessful in locating the correct minimum of the negative ln-likelihood. Despite attempts with varying limits and step sizes, there appeared to be a missing configuration within the *migrad* function that I was not accounting for. I am quite keen to find what causes *iminuit* not to find the right minimum. Consequently, I changed my approach to employ raster scanning as the method of minimisation. The resulting mean, median, and standard deviation for the estimated parameters obtained from the raster scans for both the null and alternative hypotheses are summarized in TableVI.

The distribution of ln(likelihood ratio)

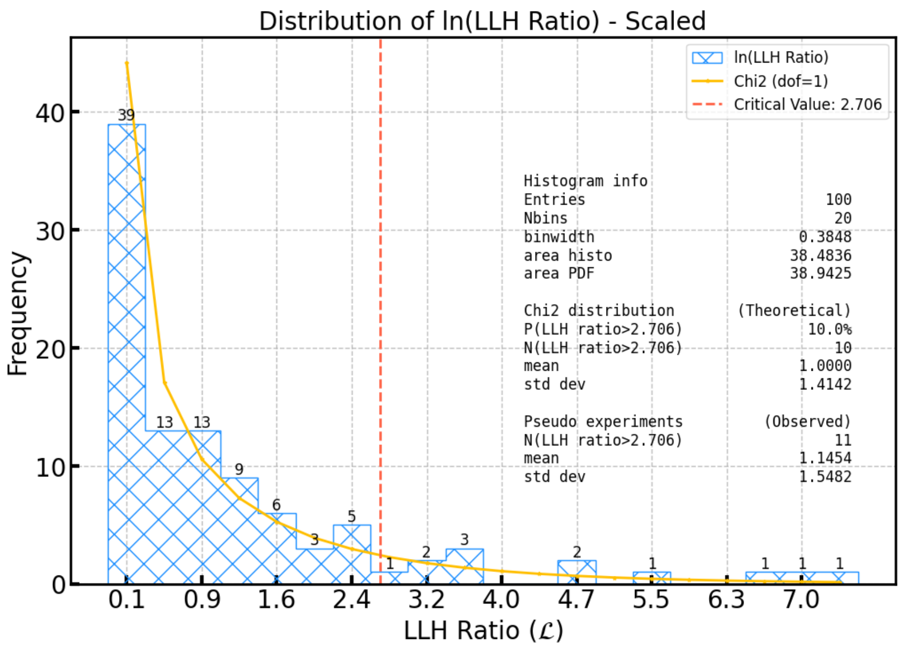
Now, I possess a set of 200 likelihood values for both the null hypothesis and the alternative hypothesis. From these sets, I compute 200 likelihood ratios, take the natural logarithm of each, and then multiply by -2. The resultant values are depicted in Figures22 and23. Given that the null hypothesis includes σ and the alternative hypothesis comprises both σ and b, and considering that the two σ values originate from the same parameters of the underlying Gaussian distribution and show high correlation as indicated by the statistics in TableVI, it is reasonable to insist that the degrees of freedom (DOF) for this analysis is one. The distribution of ln(likelihood ratio) seems to conform to a chi-squared distribution with DOF=1. By comparing the statistical properties of the ratio's distribution with that of the chi-squared distribution, they manifest similar traits. This comparison is illustrated in the blank spaces of Figures22 and 23.

Consistency above 2.706: Yes

Given the distribution of the ratio follows the χ^2 distribution with DOF = 1, I can calculate percentile position with respect to a specific value since the value has to do with the standard deviation of the normal distribution. For this instance, the proportion above 2.706 in the ratio or χ^2 distribution is the same as the proportion of the area of the Gaussian curve with standard deviation $std = DOF$ outside the interval $[-\sqrt{2.706} \times (DOF), \sqrt{2.706} \times (DOF)]$, where $\sqrt{2.706} \sim 1.644$. The theoretical expectation value from the χ^2 distribution (DOF=1) gives 10 while the ratio distribution result in giving 11. This discrepancy is rather minuscule considering the statistical characteristic of the two distributions that are shown in the blank space of Figure22 and 23.

Parameter	Mean	Median	Std Dev
σ_0 (Null)	0.6136	0.6263	0.0555
σ_A (Alternative)	0.6123	0.6263	0.0576
b_A (Alternative)	0.9953	0.9737	0.0901

Table VI: Statistical measures for the parameters of null and alternative hypotheses.

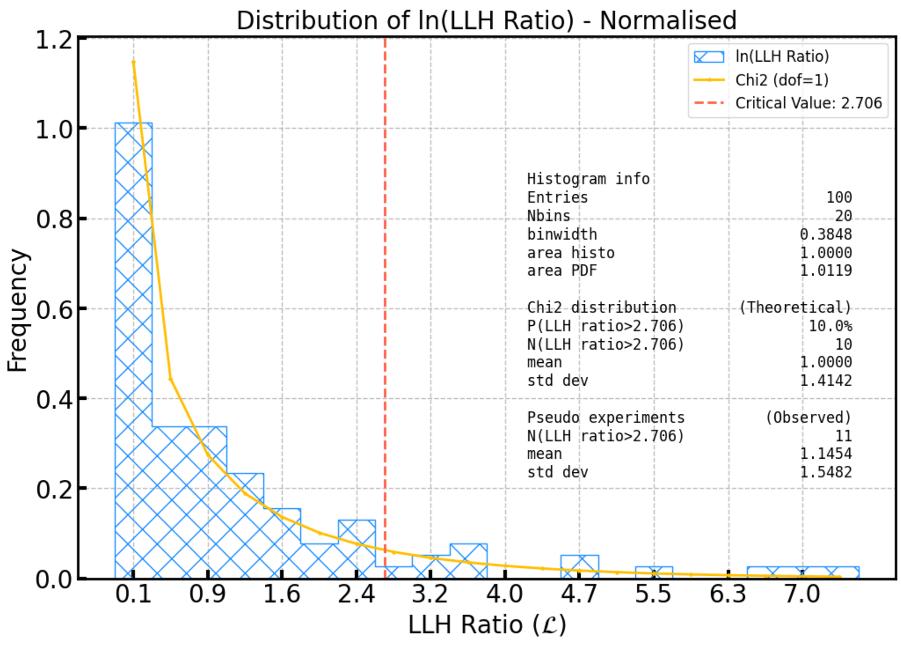


```
Histogram info
Entries                                100
Nbins                                   20
binwidth                            0.3848
area histo                         38.4836
area PDF                           38.9425

Chi2 distribution            (Theoretical)
P(LLH ratio>2.706)                   10.0%
N(LLH ratio>2.706)                      10
mean                                1.0000
std dev                             1.4142

Pseudo experiments              (Observed)
N(LLH ratio>2.706)                      11
mean                                1.1454
std dev                             1.5482
```

Figure 22: The histogram of the natural log of the likelihood ratio. The χ^2 distribution of degree of freedom = 1 is plotted on top.



```
Histogram info
Entries                               100
Nbins                                  20
binwidth                           0.3848
area histo                         1.0000
area PDF                           1.0119

Chi2 distribution           (Theoretical)
P(LLH ratio>2.706)                  10.0%
N(LLH ratio>2.706)                     10
mean                               1.0000
std dev                            1.4142

Pseudo experiments             (Observed)
N(LLH ratio>2.706)                     11
mean                               1.1454
std dev                            1.5482
```

Figure 23: Normalised distribution of ln(Likelihood Ratio), reflecting the frequency of the ratio values and their comparison to a χ^2 distribution with one degree of freedom (dof=1). The dashed red line indicates the critical threshold for the likelihood ratio, aiding in the visual determination of significance levels.