20171401, JAYITHA .C
20161066, Thapaswini Chowdary

DATA WAREHOUSING AND DATA MINING

OPTIONAL PROJECT PHASE
OUTLIER DETECTION

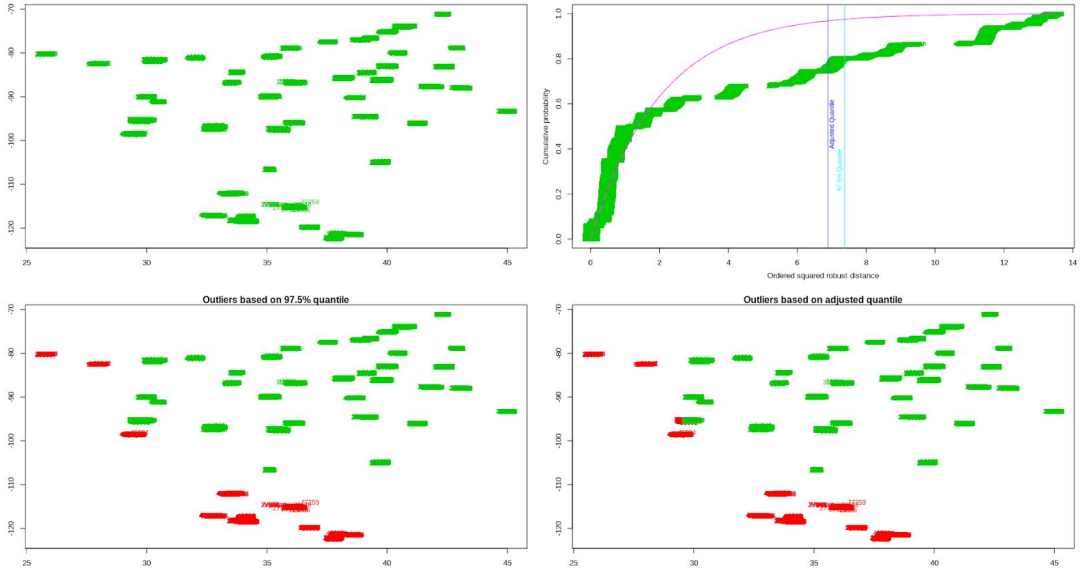
DATA SET DESCRIPTION

Year	Month	Day	Race	Age	Sex	City	State	Latitude	Longitude	Disposition
2007	1	1	Asian	10	Female	Albuquerque	AL	00.00	000.00	Closed by Arrest
2008	2	2	Black	20	Male	Atlanta	AZ	25.73	-122.51	Closed without Arrest
2009	3	3	Hispanic	30	S_Unknown	Baltimore	CA	25.74	-122.5	Open/ No Arrest
2010	4	.	White	40		.	CO	25.75	-122.49	
2011	5	.	Other	50		.	DC	.	.	
2012	6	.	R_Unknown	60		.	.	.	.	
2013	7	.		70		.	.	.	.	
2014	8	.		80		.	.	.	.	
2015	9	.		90		Stockton	.	45.03	.	
2016	10	29		100		Tampa	TX	45.04	-71.05	
2017	11	30		100		Tulsa	VA	45.05	-71.04	
	12	31		-500		Washington	WI	45.06	-71.02	

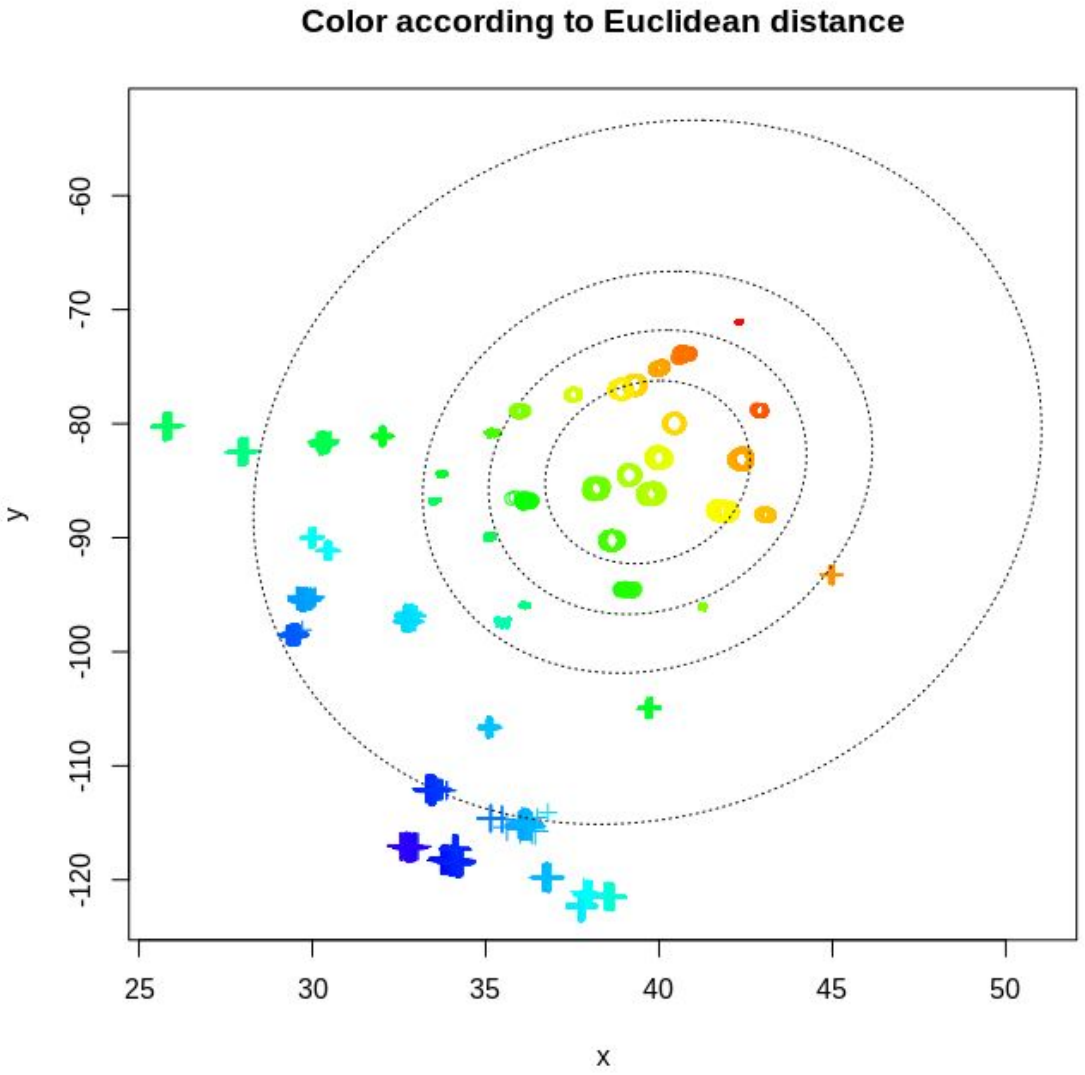
OBJECTIVE AND TOOLS USED:

We used R packages to plot histograms wrt the dates and we used euclidean distances and chi squared measures to check for outliers wrt the latitude and longitude of the detected victim. The objective was to see if the number of victims on a particular date were less, if it was say 1, this doesn't exactly constitute a homicide and is hence an outlier, similarly we were also trying to identify regions where the number was very low, again out focus is on identifying homicides that frequently occur in one region and we can eliminate all such regions from ouR dataset

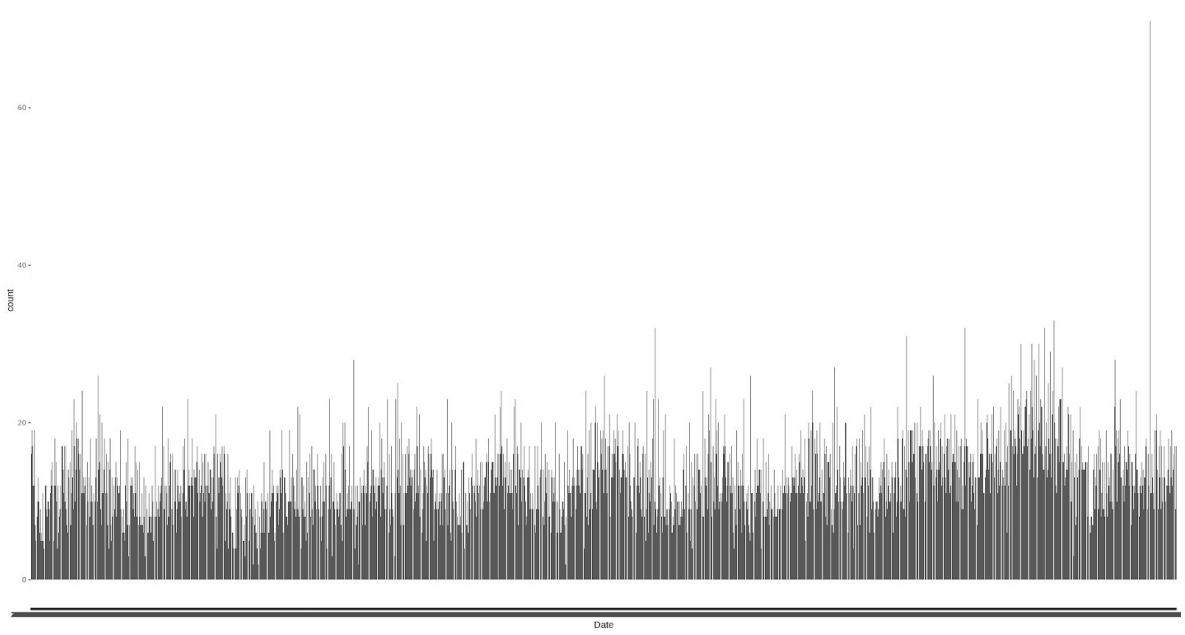
RESULTS:



In the second set of plots we see the results of the R package trying to identify the outliers based on the chi squared measure which it sets itself. But again this doesn't seem to perform very well.



In the graph above we have plotted the points based on their distances from the mean of all the latitudes and longitudes, the distance measures used are Euclidean and Mahalanobis distance represented by the +s and th os.



Shown above is a histogram of deaths by date, if the number of deaths are very low, it doesn't signify that that point is an outlier. It is only if the combination of all 3 is an anomaly but individually they do provide us with some information like most significant historical events.

CONCLUSION

Although we have used the 2 measure, location and date separately for a rough estimate of outliers, this isn't right because ideally we have to use a combination of these 2 parameters to get a very good estimate. While data preprocessing we have already removed all the unknown values because when included these showed up as the outliers. Upon manual analysis of the data we found no outliers as to deem something as an outlier we need to set a minimum threshold and non of the data points have gone below the minimum threshold we set for the count.