

Breaking Human Trafficking Network: An Analytics Approach

Raj Laxmi Prakash, Oklahoma State University, OK, US

ABSTRACT

Labor migration, illegal sex workers, child trafficking and many more impact our society in many ways. The need for developing a tool to discover and track human trafficking is extremely important and pulls the attention of many researchers. This exploratory paper describes the comprehensive analysis of the role of online classified advertisements in facilitating sex trafficking specifically and explores technological innovations to combat the increasing network of human traffickers.

With the growth of internet and social media, human trafficking networks are spreading from the ease of communication. On websites such as Backpage.com, different online advertisements are posted to lure men, women, teens and children. These ads are used for selling as well as recruiting potential victims by manipulations and false promises of a job etc. Traffickers have become more sophisticated in their methods resulting in being seemingly untraceable and hiding their identities. In 2018, Backpage.com was seized by the FBI for their participation in illegal prostitution and sex trafficking. However, this did not end the problem but shifted it to unknown places.

The recruiting ads are particularly valuable for investigators because they seem more likely to provide direct links to human traffickers than any of the other categories of ads. These ads often provide location information including the location where the ad appears, the poster's location, a destination city (e.g., Las Vegas), and the area code of the contact phone number. The purpose of this paper is to examine the relationships between different locations. This Paper uses network analytics to build models for identifying potential connected relationships including timing, locations, contact numbers and other features of the ads. The data was obtained by scrapping ads from sites like Backpage.com and analyzed using different SAS tools like SAS® Enterprise Guide™, SAS® Visual Analytics™ and Python for network and exploratory analysis of the data.

INTRODUCTION

The United Nations defines Human Trafficking as “the recruitment, transportation, transfer, harboring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labor or services, slavery or practices similar to slavery, servitude or the removal of organs”

A pictorial ball-park idea of possible network is shown below as an example of the complexity of the human trafficking network and the reason of human to enter and getting trapped into this net: [Pocock NS et al.]³

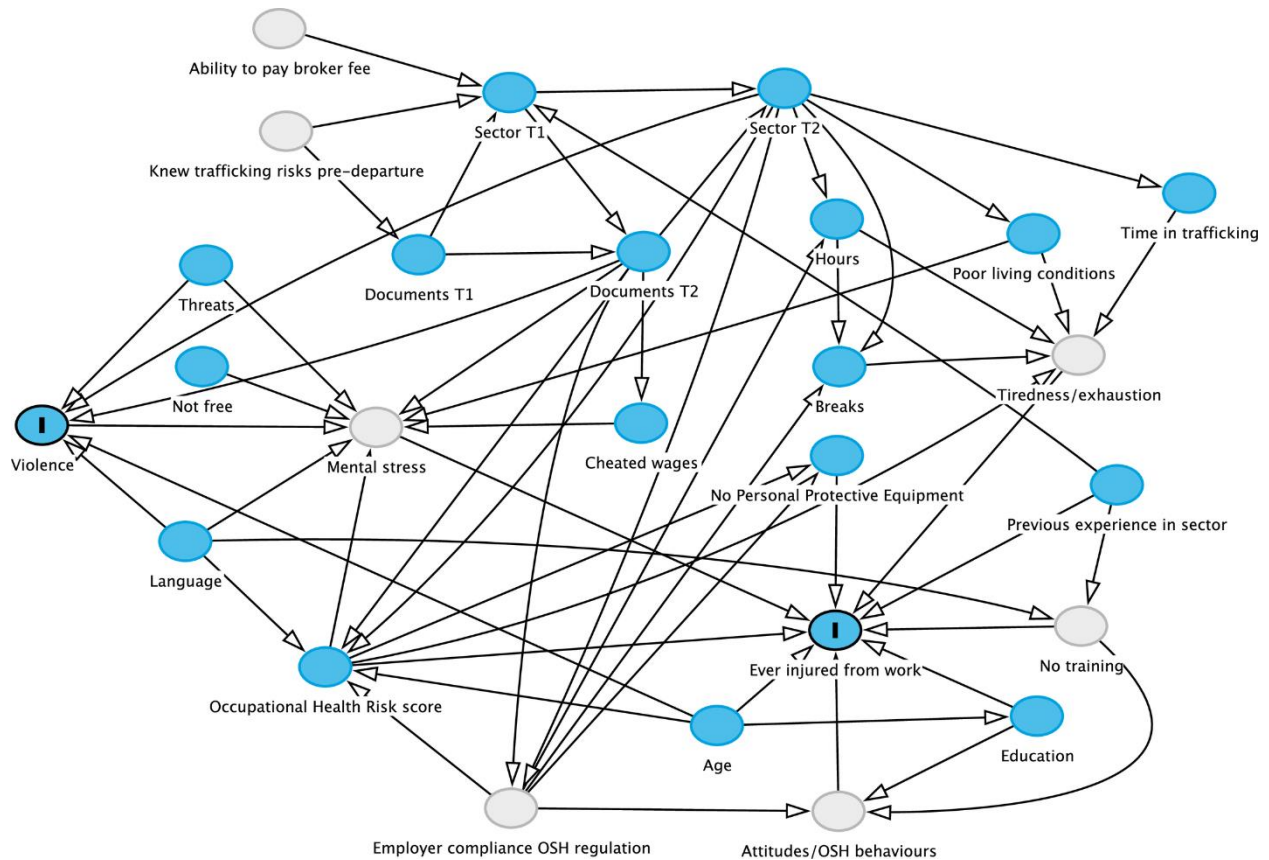


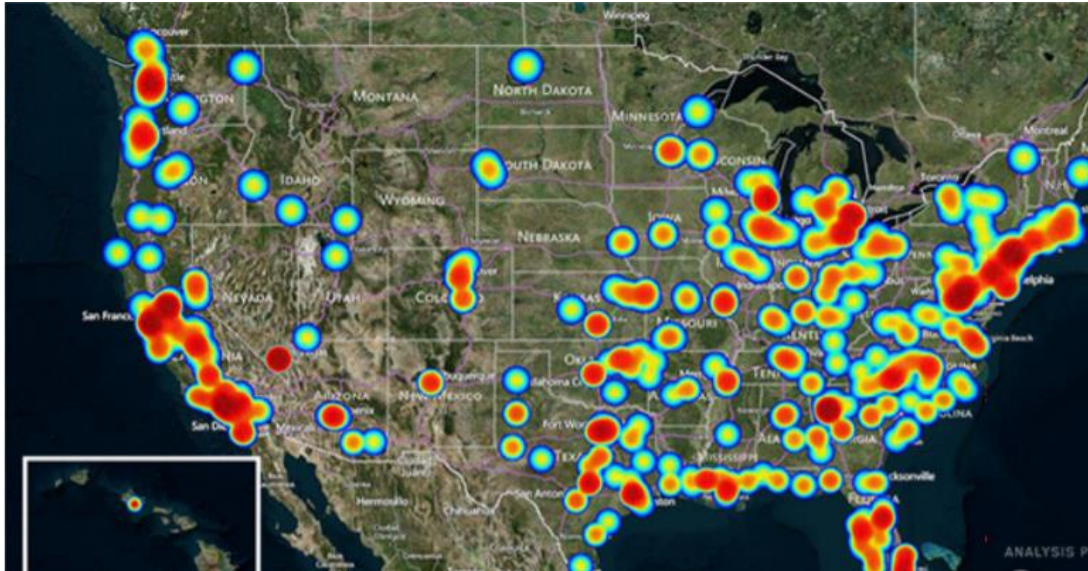
Fig 1 : Ball Park Pictorial representation of network

Human trafficking is the business of stealing freedom for profit. In some cases, traffickers trick, defraud or physically force victims into providing commercial sex. In others, victims are lied to, assaulted, threatened or manipulated into working under inhumane, illegal or otherwise unacceptable conditions. It is a multi-billion dollar criminal industry that denies freedom to 24.9 million people around the world [2].

Human trafficking is categorized as a most suitable multidimensional human rights violation and the center of this violation resides in the human exploitation. The United Nations defines trafficking in persons as “the recruitment, transportation, transfer, harboring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation” [4].

Human Trafficking is a form of modern day slavery and it is a grim reality of 21st century global landscape in developed as well as developing countries. In the United States, the US justice department estimates that 35,000 – 170,500 people are trafficked into the country every year. The 2016 global slavery index estimates that including US citizens and migrants 57,700 people are victims of human trafficking [2].

Following figure indicates the most prominent region of the United States in 2014, which had considerably good number of human trafficking cases registered [5]:



In October 2018, a Dallas based man was sentenced to life in federal prison for trafficking children for sex. This case gives a rare glimpse into a dark underworld whose existence is undercover because of the hidden nature of this crime. The 2017 Federal Human Trafficking Report found that there are 661 active sex trafficking cases and 68.8% involved children and teens and there are hundreds and thousands of victims beyond this number who have not been found due to its hidden network. They are lured by the online classified ads and captured by older men or women who have befriended them on social media.

Sex traffickers not only use ads on classified ad sites to attract customers, they use ads to attract new victims. These ads are made attractive to catch eyes of teens and children as for eg. these ads promise independence , fun , travel, flexible schedules, good working condition and high earnings. These ads typically list a variety of desired qualities including physical appearance, age, willingness to relocate and open mind. The recruiting ads are particularly valuable for investigators because they seem more likely to provide direct links to human traffickers than any of the other categories of ads. These ads often provide location information including the location where the ad appears, the poster's location, a destination city (e.g., Las Vegas), and the area code of the contact phone number. The purpose of this paper is to examine the relationships between different locations.

DATA DESCRIPTION

There are three types sets of files created from a collection of Backpage ads.

- 1) **Base file / Fact table-** SAS dataset "ModifiedFact". This dataset contains one record per ad, uniquely identified by "Fact_Key".
- 2) **Dimension files:** These datasets contain de-duplicated content. They contain one or more text fields, which are the data fields that were extracted from the ads, and the "[field]_Key" field, which uniquely identifies the record for linkage purposes. The current dimension files are "Location_Dim", "Location_City_Dim", "Posting_Body_Dim", and "Title_Dim".
- 3) **Multiple value files:** These datasets contain the variables that could have more than one occurrence per ad. They are the "Emails" and "Phones" datasets. They contain one or more text fields, which are the data fields that were extracted from the ads, and the "Fact_Key" field, which relates the records back to the associated Fact table records.

DATA PREPARATION

All these datasets contained huge information, so the required information were filtered with the help of 3 step data preparation and all these data preparation was done on SAS Enterprise Guide.

Step -1 – Data Joining

In the first step, data were merged with the common keys. For the dimension files, “ModifiedFact” contains four fields, "Title_Key", "Posting_Body_Key", "Location_Key", and "Location_City_Key". These join to the identical fields in the dimension tables. The relationship is many to one from “ModifiedFact” to the dimension table.

For the multiple value fields, the “Fact_Key” field joins to “ModifiedFact”. The relationship is one to zero, one, or many from “ModifiedFact” to the multiple value field.

Step - 2 - Variable Selection

In the second step, important variables were selected based on problem requirement and literature review. One calculated column of ‘Area Code’ was added by extracting information from the variable ‘Phone No’. As in the paper, only recruiting Ads were required, so all other categories of ads were filtered out. This resulted into selection of 13 variables

Step – 3 – Data Modification

In the last step of data preparation, missing values were treated, as data was huge, so the missing values were removed. Now the prepared dataset has 0.5 million observation and 13 variables of only recruitment Ad type categories.

Data Distribution & Characteristics

The most important variables for establishing the relationship among locations were – Area code associated with the phone no , the Ad_Location_City – cities where the ads appeared, Ad_Location_State – States where Ads appeared , Recruiting_Location- the recruitment location or destination city for which these ads were posted.

Name	Type	Length	Format
Fact_Key	Numeric	8	BEST12.
Post_ID	Character	37	\$CHAR37.
Ad_ID	Character	9	\$CHAR9.
Recruitment_Area	Numeric	8	BEST12.
Posted_DateTime	Date	8	DATETIME...
Location_Status	Numeric	8	BEST12.
Posting_Body_Key	Numeric	8	BEST12.
Location_Key	Numeric	8	BEST12.
Ad_location_City	Numeric	8	BEST12.
Phone	Character	10	\$CHAR10.
Area Code	Character	3	\$CHAR3.
Email	Character	52	\$CHAR52.
Ad_location_State	Character	12	\$CHAR12.

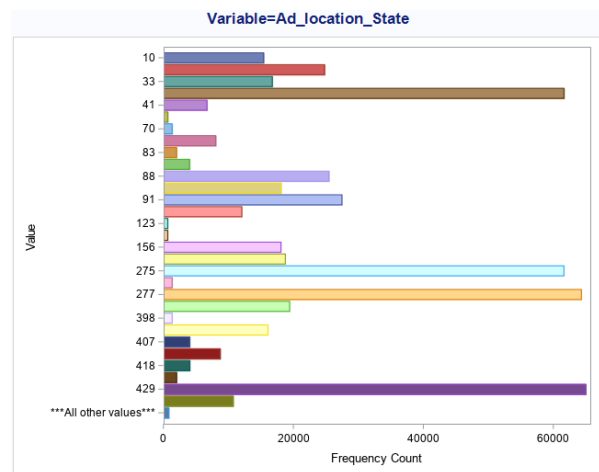
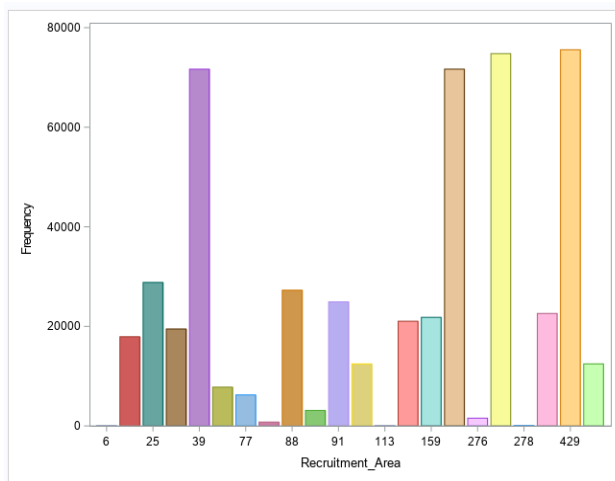
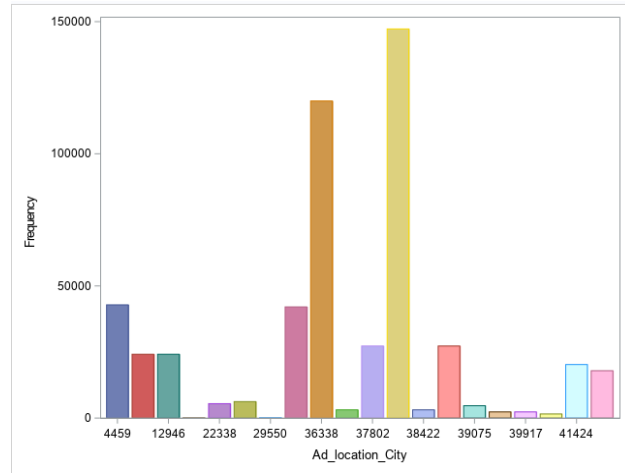
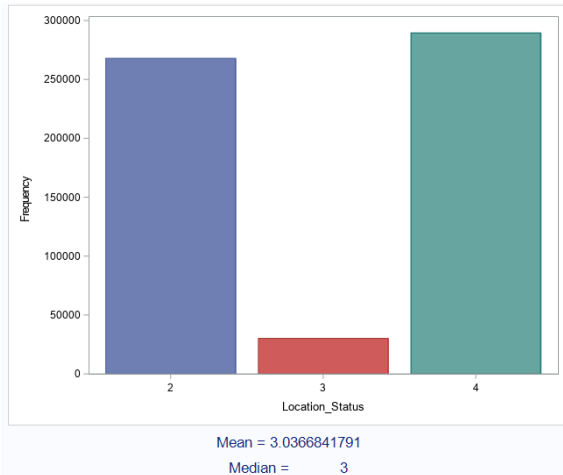
Fig: Description of 13 variables*

*Appendix shows complete description

Below is the distribution of important variables

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Email		monesetko@gmail.com	521930	99.9686
		Dream Builders@gmail.com, Dream	114	0.0218
		valiantservices13@gmail.com	50	0.0096

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Phone		7752393500	521930	99.9686
		2016809604	114	0.0218
		2024921429	50	0.0096



For the recruiting ad, it is found that there were only 3 emails and corresponding phone nos were circulated in recruiting ads. The major ads are posted in multi-cities or the cities are unknown.

From the above frequency analysis of Recruitment area and Ad location city, it can be interpreted that there are few states where the frequency of these ads are higher. As for eg 39 (Utah) 275 (Elko Nevada), 277(Reno Nevada) and 429 (Salt Lake City Utah) are found to have highest frequency of Ads.

ANALYSIS

In order to solve the objective of the project, network analysis was done using networkx in python. But before establishing connections, correlation analysis was observed among different variables which provided the following graph.

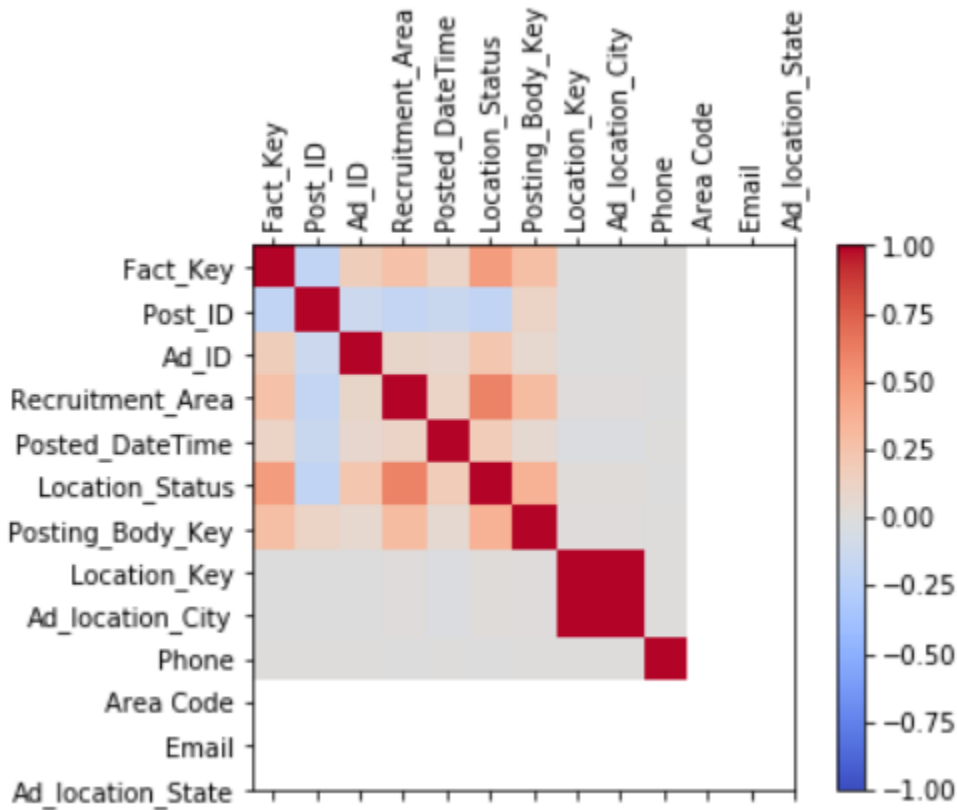


Fig: 2 Correlation matrix among variables

The above graph shows that there are some positive and negative correlation among different variables as for eg. Ad Location City key and phone are highly correlated. Location Status is correlated with Recruitment Area. There are some negative correlation was also observed between Post_ID and corresponding variables. However, for the purpose of this paper is to examine relationship among different locations , so further analysis was done on locations which includes Area code, Ad_location_City, Ad Location State, Recruitment Area.

In order to find connections among the cities and geographical area for the network of recruitment places and Ads posted, Network analysis was done using Networkx on python.

Network analysis was done between the Ad ID as source and Area code, which was taken out from phone no. mentioned in the Ad text, was taken as Target and network connection was observed. The following graph shows the possible connections among the Ad ID and Area Code.

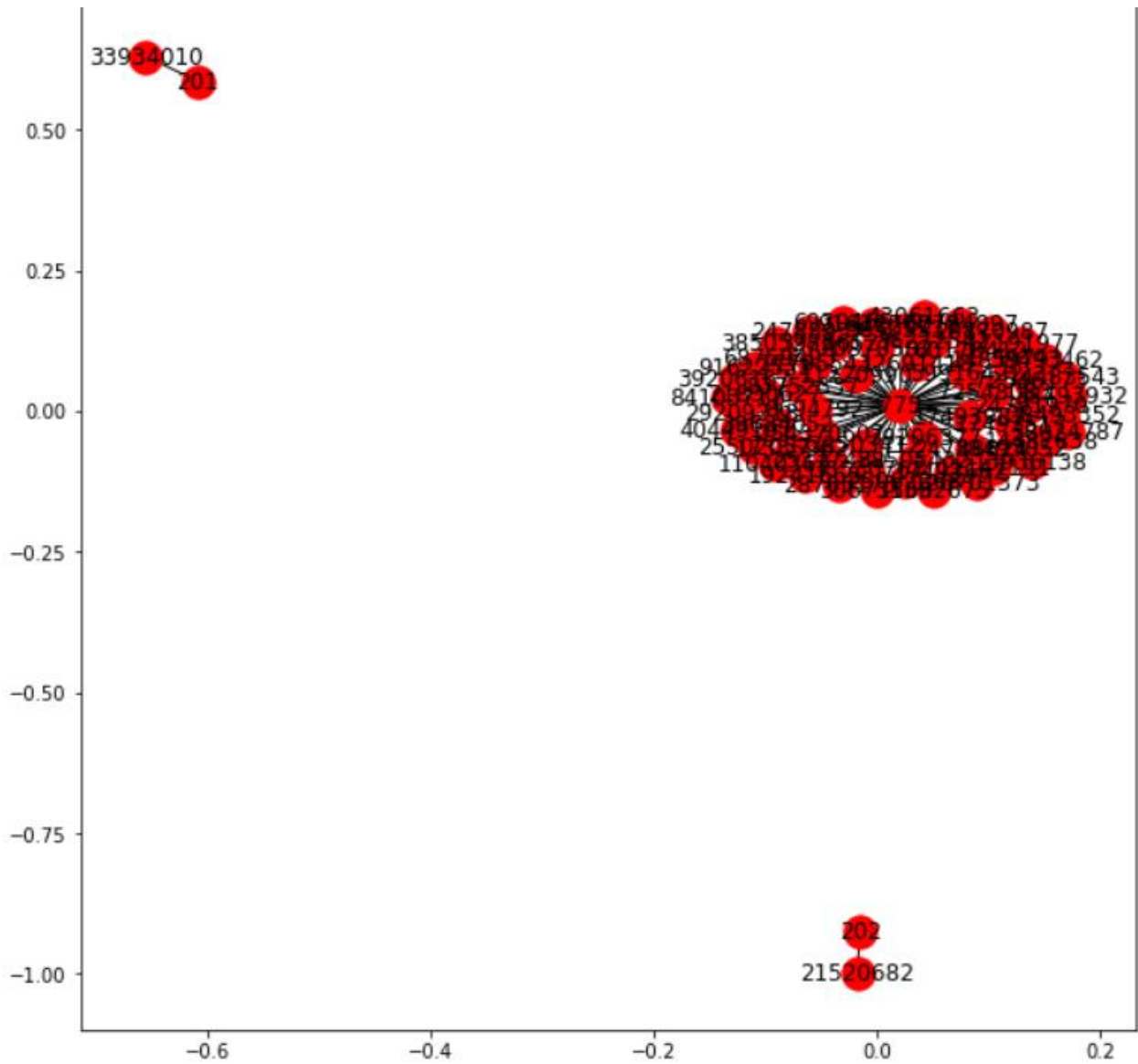


Fig 3: Network Analysis between Ad ID and Area Code

Most of the recruiting Ads were associated with area code 775 which is of Nevada. Only one Ad was associated each with 202 and 201 area code which is of Washington DC and New Jersey respectively.

The above graph shows that area code 775 (Nevada) has highest degree of centrality and is the most influential among all the three area code.

In order to understand with respect to the State variable, the network comes as shown in below fig.

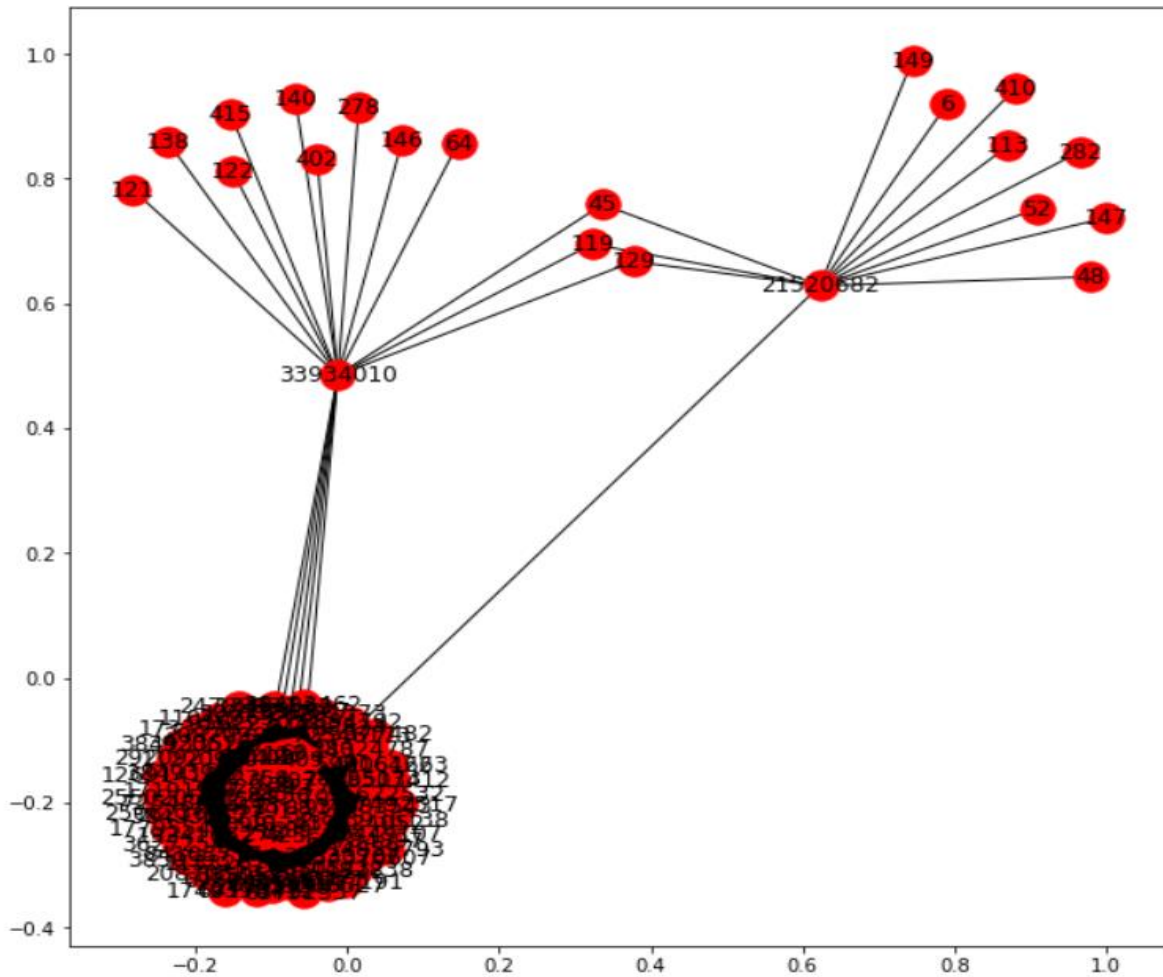


Fig 4: Network Analysis between AD ID and Ad Location State

Table 1:

State	Degree of Centrality
409	0.5405405405405406
403	0.5405405405405406
60	0.5405405405405406
418	0.5405405405405406
143	0.5405405405405406
123	0.5405405405405406
425	0.5315315315315315
453	0.5315315315315315
89	0.5315315315315315
88	0.5315315315315315
93	0.5315315315315315
77	0.5315315315315315
156	0.5315315315315315
159	0.5315315315315315
25	0.5315315315315315
275	0.5315315315315315
277	0.5315315315315315
10	0.5315315315315315

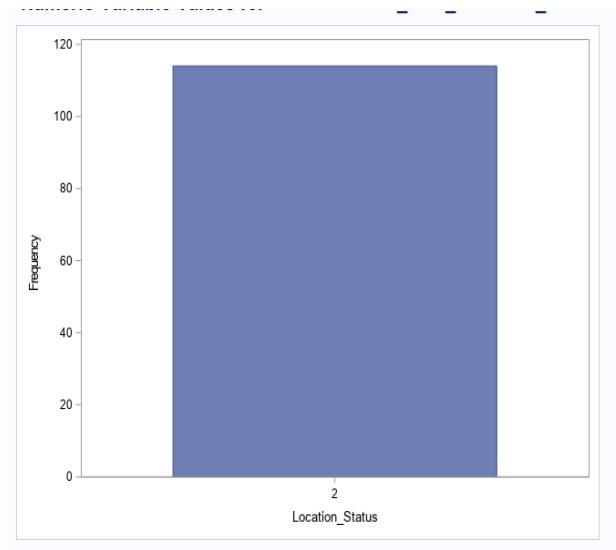
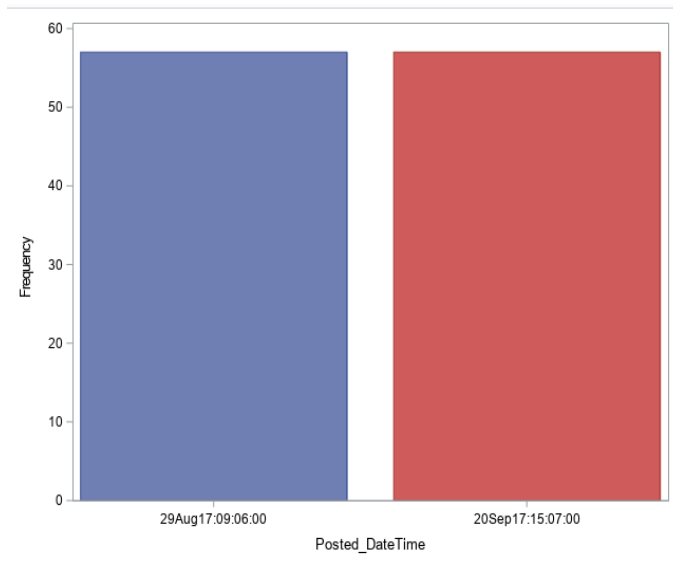
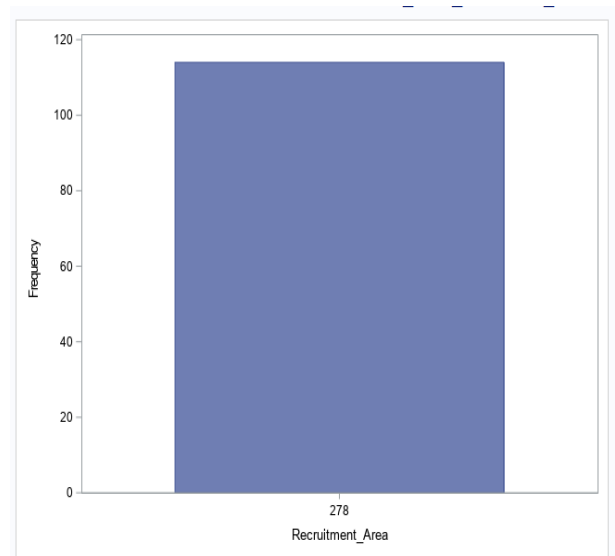
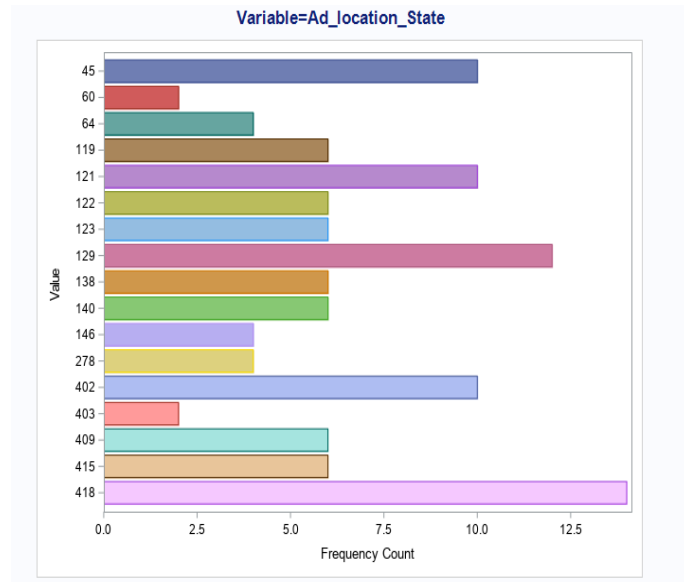
From Fig. 4 and Table 1 of result of degree of centrality, it can be said that the most influential locations are 409 (Houston Texas), 403(Dallas, Texas), 60 (Phoenix Arizona), 418(San Antonio Texas), 143(Athens Georgia), 123(Jacksonville Florida) and other states which are shown in the result of degree of centrality.**

The above graph shows that majority of Ads are centralized in majority of states but two Ads 33934010 and 21520502, which are linked to two specific area codes 201(New Jersey) and 202 respectively (Washington) (as described earlier) , have been posted specific locations in which 45 (Birmingham Alabama) , 119 (Daytona Florida) and 129 (Orlando Florida) are common.

Hence, these two Ad IDs can help in connecting the network among different locations and phone no. So, in order to understand the connection among different States, Location City, Recruitment Area, Ad IDs 33934010 and 21520502 was taken into consideration for a deeper analysis.

** Names for numerical coding for locations are in Appendix

Ad ID 33934010 shown the following frequencies for different variables.



Variable	Label	Value	Frequency Count	Percent of Total Frequency
Ad_ID		33934010	114	100.000
Area Code		201	114	100.000
Email		DreamDeals@gmail.com, Dream	114	100.000
Phone		2016869604	114	100.000

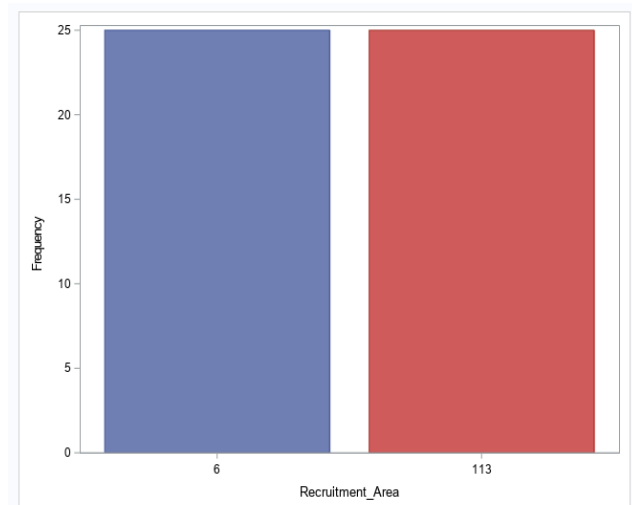
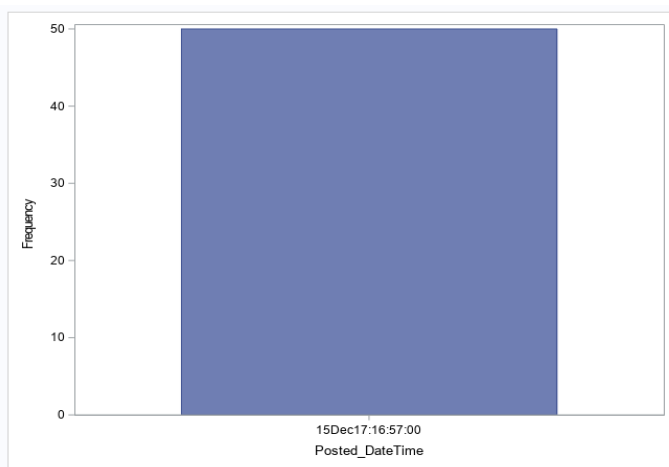
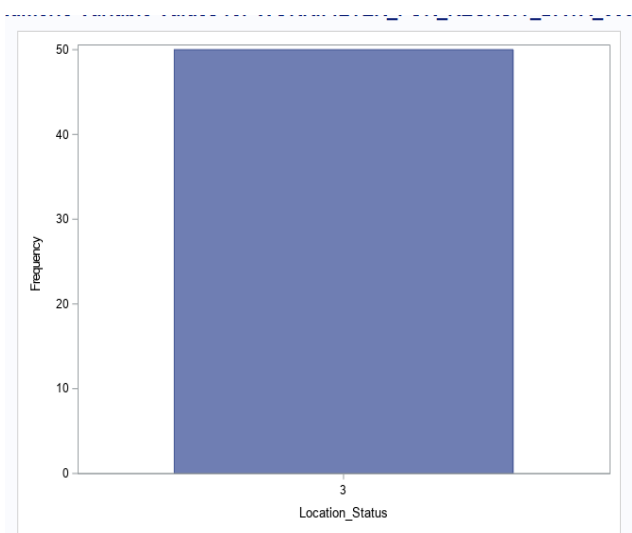
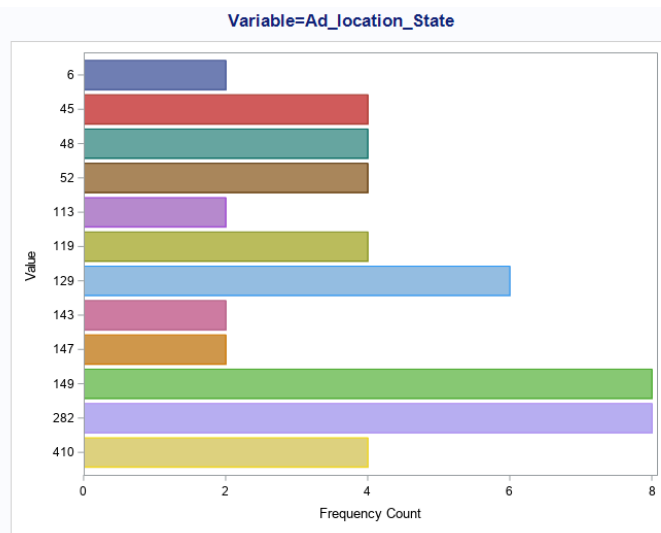
From the above graph and chart, it can be said that there count of 114 of this Ad ID and have single phone no and email ID and two posted date time.

From the graph, it can be observed that Ads was posted in 17 different states and the states are San Antonio TX (having highest frequency), Orlando FL, Birmingham AL, Fort Myers FL, Corpus Christi TX, Daytona FL, Phoenix AZ, Dallas TX (lowest frequency). So, it can be said that the Ads are posted in different States for the recruitment in a specific location.

The graph shows one geographical area for which Ads were posted for recruitment is Central Jersey and the Location status is 2 which implies that the ads appeared in multiple cities.

There is only one phone no and respective Email is used for this particular Ad. Also, the area code of the phone is of New Jersey as identified in the Ad text.

For Ad ID 21520682, following graph resulted for different variables.



Variable	Label	Value	Frequency Count	Percent of Total Frequency
Ad_ID		21520682	50	100

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Area Code		202	50	100

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Email		valiant services 18@gmail.com	50	100

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Phone		202 4921423	50	100

From the above graph, it can be stated that there are 12 states where this Ads was posted. The states were Savanna GA (highest frequency), Albuquerque NM, Orlando FL, Birmingham AL, Huntsville AL, Tuscaloosa AL, Daytona FL, Macon GA (lowest frequency).

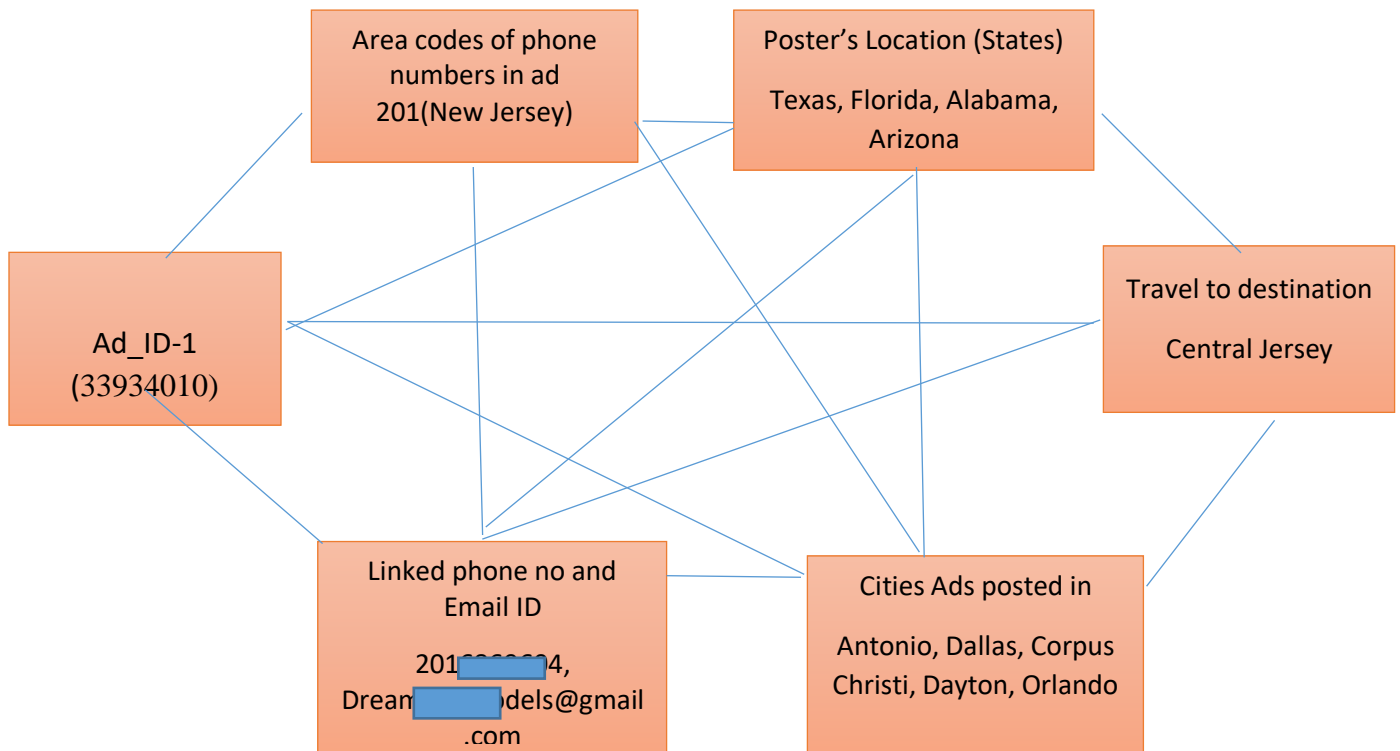
There were two recruitment area which are 6 (Connecticut) and 113 (Hartford Connecticut). There is one phone no which has area code of Washington, DC and one respective Email is observed in this particular recruiting Ad. The Ad was posted once on Oct 15th 2017 as per the data.

RESULT

From the above analysis, it can be interpreted that there are different geographical area where Ads are posted for recruitment in different location with the use of phone nos. of different area code.

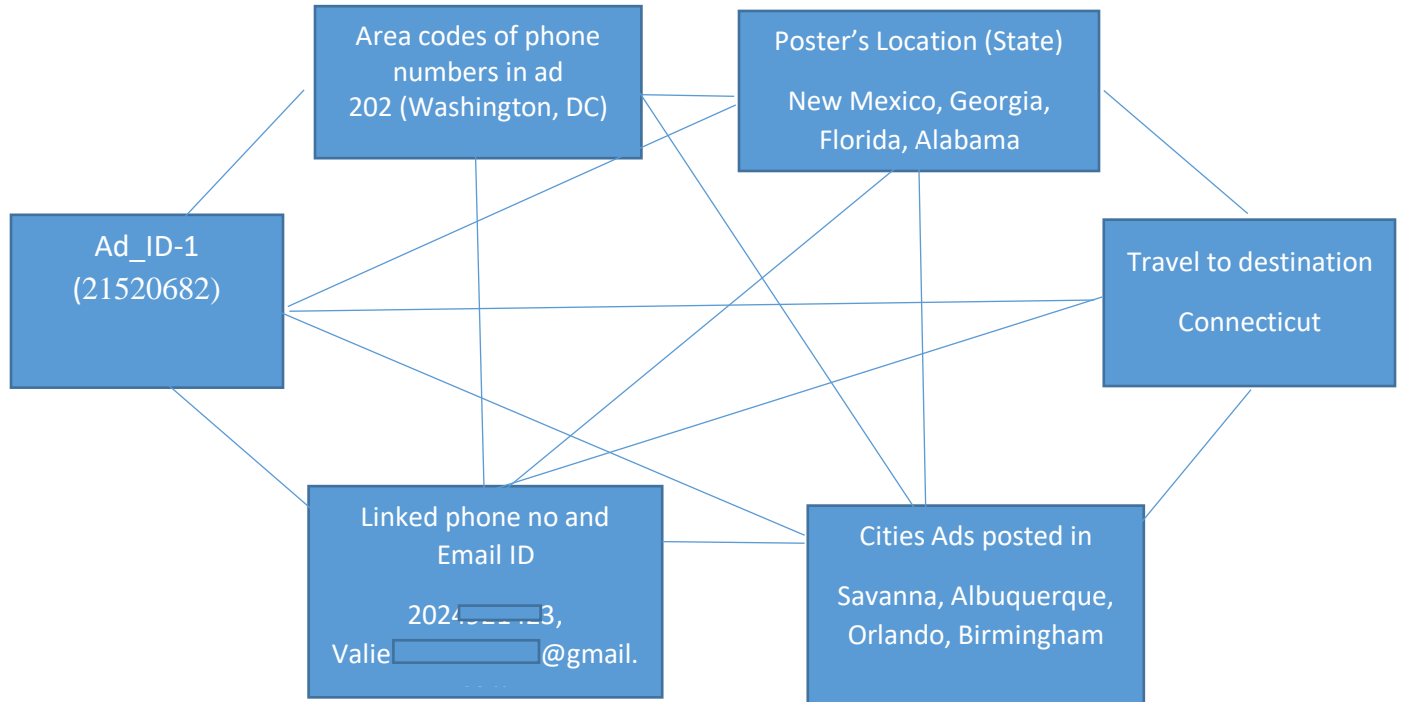
When the information for Ad ID 33934010 will be linked on graph then it resulted into following graph.

CITIES IN THE TRAFFICKING NETWORK



When connections of different locations were established for the Ad ID 21520682, then following network was established.

CITIES IN THE TRAFFICKING NETWORK



From the above network analysis and network, the result shows that there is a complex network of human traffickers and it is not easy to track them. However, development in technology in data science can help in tracking these traffickers. From the network graph, it can be easily interpreted that the Human Traffickers are using phone numbers of different area codes, while they are posting ads in multiple cities and surprisingly the recruitment is taking place far from these locations.

CONCLUSION

The recruitment Ads are found to be potential Ads to establish connections among different cities and geographical areas. There are lot of information available on the Ad text which can help us to reach the human traffickers and break their network across the globe. Data Sciences can help us to identify the types of Ads, the phone nos. and email ID linked to the Ads, the recruitment place, the locations where Ads are circulated and posted. It can be concluded from the analysis that the Phones no are of different place, and the recruitment place is in some other state and the same Ads are posted in different states and cities of US. Hence, in order to combat this problem, the network of Ads should be identified the work can be done to reach the traffickers.



- Locations for posting of Ads
- Location of Recruitment

Here in the picture the blue line is linking areas where the Ads were posted and Red line is connecting the place of recruitment and area codes of linked phone nos.

Future Scope

A lot of result came out from this analysis and this leads to a lot of future scope of this project as well. The characteristics of each of these locations can be analyzed futher and movement of human traffickers can be tracked. The criminal history and no of missing children/ adult can be analyzed and they can be tracked by applying such analysis. More information about location like the street address etc can help in better analysis and can make the process of tracking the human traffickers easy.

REFERENCES

1. Zimmerman C, Kiss L (2017) Human trafficking and exploitation: A global health concern. PLoS Med 14(11): e1002437. <https://doi.org/10.1371/journal.pmed.1002437>
2. <https://www.newsweek.com/sex-ads-website-backpagecom-co-founder-charged-after-fbi-raid-876333>
3. Pocock NS, Kiss L, Oram S, Zimmerman C (2016) Labour Trafficking among Men and Boys in the Greater Mekong Subregion: Exploitation, Violence, Occupational Health Risks and Injuries. PLoS ONE 11(12): e0168500. <https://doi.org/10.1371/journal.pone.0168500>
4. United Nations General Assembly. Optional Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention Against Transnational Organized Crime. 2000.[27 Oct 2017]. https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=XVIII-12-a&chapter=18&lang=en.
5. <https://polarisproject.org/sites/default/files/us-citizen-sex-trafficking.pdf>
6. <https://www.caclapeer.org/lapeercacblog/human-trafficking-the-internet>

CONTACT

Name: Raj Laxmi Prakash

Enterprise: Oklahoma State University

Address: Spears School of Business, Oklahoma State University, Stillwater, OK -74074

Email: rajlaxmi.prakash @okstate.edu

Appendix

Field Descriptions

Fact_Key:	SAS Numeric, contains an integer. It is the primary key for Ad_Fact, and the foreign key for the multiple-entry dimension files. It was added during the transformation process.
Post_ID:	Character. Composed of elements of other fields. Potentially, this could be a primary key, but at the current time duplicates exist so more work would be required.
Ad_ID:	Character. A series of digits or the string "geoBlock". The identifier of the ad in the Backpage environment.
Recruitment_Area:	Numeric Coded(G_ACSSf). The entries for Recruitment_Areas are the names of locations of recruitment or destination city found on Ads text including cities and states.
Posted_DateTime:	SAS DateTime. When the ad was posted. If the field consisted of a date only, the time is set to "00:00:00".
Location_Status:	Numeric Coded. Describes the nature of the Location or Location_City variable where 1 = " ", 2 = "Multi_Cities", 3 = "One_City", 4 = "Unknown"
Posting_Body_key:	Numeric. The body of the ad from the Backpage entry numerically coded.
Location_key:	Numeric. The contents of the "location" field from the Backpage entry, numerically coded
Ad_Location_City:	Numeric. City values that were found in the ad text numerically coded
Phone:	Character. Ten digit integer. Value of phone number or numbers, if found in the ad. Several phone numbers that contained emojis instead of numeric characters were dropped.
Area_Code:	Area code extracted from the phone
Email:	Character. Value of email address or addresses, if found in the ad. Note that the values don't always follow the rules for valid emails.
Ad_location_State:	Numeric. State values that were found in the ad text numerically coded

