MGT 6203 **Team 64** Group Project Proposal: **Uncovering Insights in Human Trafficking Worldwide**

TEAM INFORMATION (1 point)

Team #: **64**

Team Members:

1. **Roy Garcia; rgarcia90**

I'm Roy Garcia, currently living in Queens, NY. I'm currently a Sr. Customer Development Analyst for Johnson & Johnson, undergrad in Industrial Engineering, at work I have worked with some predictive models to detect if a customer will default payments, segmentation, fraud detection systems, and others.

2. **Lucas Chen; lchen687**

My name is Lucas and i live in Baltimore, MD. I am currently a Sr. Analyst for the data management department at Comcast. I help maintain and improve our database systems so that analysts have an easier time performing analyses.

3. **Garrison Winter; gwinter3**

My name is Garrison, and I live in Cumming, GA. I just recently finsihed my undergrad at Hendrix College and was a dual sport athlete for football and lacrosse. I held a record for most saves in a game for little over two years. I currently work full time in midtown Atlanta for Cushman and Wakefield as a senior data analyst.

4. **Amit Dommeti; adommeti6**

My name is Amit, and I live in Atlanta GA. I went to undergrad at Georgia tech and majored in Biomedical Engineering. I worked in a consulting company for a few years (did some cool data science projects, like building a clinical review model in SQL, and uncovering some insights during the peak of the Covid-19 pandemic) and now work at a Finance Technology company as a Product Analyst.

OBJECTIVE/PROBLEM (5 points)

Project Title: Uncovering Insights in Human Trafficking Worldwide

Background Information on chosen project topic:

There is an increasing amount of public attention that is being driven toward the problem of human trafficking. There is a shortage of analytics solutions that can assess decisions related to preventing human trafficking.

Our team, as a group of growing and developing data scientists, learned that we have two key traits in common with each other - we enjoy analytics, and we are passionate about social good. Human trafficking is a worldwide problem, and while there is a lot of data that exists around it, the solutioning thus far has been minimal.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

To predict how human trafficking patterns will change over the next few years, so that the right amount of resources can be allocated to help mitigate this issue.

State your Primary Research Question (RQ):

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Which areas of the world are the most susceptible to human trafficking?
2. How has human trafficking trended over the last 2 decades?
3. Has human trafficking leaned or shifted towards a specific group of people? (specific gender, age group, and others…)

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Human trafficking is a modern way of slavery which presents a global public concern, and is present in every nation. For decades, nations have tried to eradicate it, but as time goes by, they find new ways to get the traffic network back. We want to offer how this problem will evolve to prevent shortage of resources.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

- Main Data Source: [Global Human Trafficking]

Data Description (describe each of your data sources, include screenshots of a few rows of data):

The data is CTDC data that contains records from 2002 to 2019. The "dataset contains information on 48.8k victims of human trafficking, including the reason, means of control, origin and destination, as well as other variables." Because the data-set is very large and contains almost 50k rows across *63* columns (examples of columns are: "Year of Registration" , "Gender", "Age Group" , and "Country." 56 of the columns are integers (though many of these are binary indicators), while 7 contain strings (text fields). There are missing records throughout the columns, where a value of "-99" indicates a missing value. Below are several rows of data (there are many columns so only a few are shown here)

	yearOfRe	Datasour	gender	ageBroad	majorityStatus	majorit	majorityEntry	citizens	typeOfExploitConcate
2	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
3	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
4	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
5	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
6	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
7	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
8	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
9	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
10	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
11	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
12	2002	Case Manag	Female	18--20	Adult	-99	-99	CO	Sexual exploitation
13	2002	Case Manag	Female	18--20	Adult	-99	-99	MD	-99
14	2002	Case Manag	Female	18--20	Adult	-99	-99	MD	-99

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

The year will be the independent variable since it will ride along the y axis of our model graph while all of the factors are dependent since they will be feeding into the time series graph to show a trend over time. No new variables will need to be created.

APPROACH/METHODOLOGY (8 points)
Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

To start we will aggregate our data, which has data points from all over the world. Once its aggregated to the yearly level we will build a time series model such as ARIMA to help trend out future human trafficking trends across the world. Within each metric that is in our table we can spot out certain trends with certain factors that are highly correlated with increases/decreases in trafficking patterns. We may need to transform some of the categorical variables to dummy variables so that potential regression models can show us what the important factors in the dataset are. We can optimize the model hyper parameters by testing different factors to see how they contribute to a good time series model.

Due to the volume of the dataset, there are innumerable ways we can gather insights - it is important for us to choose a specific area of focus and dive deep into the data + findings. We plan to look at the data from a *global* lens (vs. state level or individual level). In doing so, we will be better equipped to solve a more targeted problem.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

Through the results, we hope to find clear trends that relate to human trafficking. This would be through identifying which of the predictors are statistically significant to the response and ultimately being able to predict whether or not a certain group, demographic, etc. could be a victim of human trafficking.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

By creating a tool like this we can detect where the agencies or organizations involved are lacking to detect criminal's activities on time before occurring. Analyzing each time series output should give us a great idea of where we can assign more resources and put more effort to make people feel more secure.

PROJECT TIMELINE/PLANNING (2 points)
Project Timeline/Mention key dates you hope to achieve certain milestones by:

Below is a high-level timeline of how the team will approach certain deadlines, milestones, and internal checkpoints of our team.

Date	Tasks to Complete + Submissions
06/05/23	**First Team meeting, Introductions, Overview**
06/12/23	**Finalize dataset to use**
06/18/23	**Introduction Meeting with David (TA)**
06/21/23	**Submit Project Proposal**
Week of 06/20/23	**Team Exploratory Analysis**
Week of 06/26/23	**Teammates select model options**
06/29/23	**Team create the proposal video from insights gathered so far**
End June/Early Jul	**Team sync with David on feedback**
07/01/23	**Submit Proposal Video**
Week of 07/03	**Team work on Progress Report**
07/05/23	**Submit Progress Report**
Week of 07/10/23	**Further analysis - begin write up of results, insights, findings, etc.**
07/15/23	(Around this time) **Final meeting with David to discuss results + final report**
07/19/23	**Submit Final Report**

Appendix (any preliminary figures or charts that you would like to include):