MGT 6203 Group Project Proposal Template

**Please edit the following template to record your responses and provide details on your project plan.**

**TEAM INFORMATION (1 point)**

**Team #:** 48

**Team Members:**

1. **Andrew Taylor, 902697379, ataylor44.** BS in Industrial Engineering from Georgia Tech, seven-year retail strategy consultant. Previously worked on many course-related projects (this is my last semester of classes), as well as a number of primarily supply chain-related projects in my work.
2. Team Member 2 Name; GT Id or EdX username
3. Team Member 3 Name; GT Id or EdX username
4. Team Member 4 Name; GT Id or EdX username
5. Team Member 5 Name; GT Id or EdX username

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:** COVID Impact on Country Happiness

**Background Information on chosen project topic:**

There is extensive data, by day, by country on the rates of COVID cases, COVID deaths, and vaccinations across two publicly available datasets. We intend to marry this up to a third dataset which contains happiness scores, by country, by year, as well as potential contributors to country happiness such as level of corruption, life expectancy, and GDP. This gives us a dataset with daily COVID data and annual happiness scores. From there, we can try to answer a number of questions about the impact of COVID on a country’s happiness.

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

We’d like to understand if and how rates of COVID cases, deaths, and vaccinations impacted the happiness of countries from 2020 to 2022.

**State your Primary Research Question (RQ):**

How and to what degree do the rates of COVID cases, deaths, and vaccinations impact the happiness score of countries from 2020 to 2022?

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

1. To what degree are COVID cases significant, if we also include other factors which are more stable?
2. Are COVID case rates, deaths, or vaccination rates more important when predicting country happiness?
3. What interaction terms may be significant? E.g. is the ratio of case rate to vaccination rate important?

**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.)**

This would be important to countries in order to better understand the mental health effects of COVID. It could provide motivation to governments to better publicize their COVID responses, as well as provide incentive for companies and governments to continue to administer and/or require COVID vaccinations. This would also be useful data to inform responses to future pandemics and disasters.

**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

1. Novel Coronavirus 2019 Dataset - <https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset?resource=download>
2. COVID-19 World Vaccination Progress - <https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>
3. World Happiness Report up to 2022 - <https://www.kaggle.com/datasets/mathurinache/world-happiness-report>

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

1. Novel Coronavirus 2019 Dataset - This is a country- and day-level dataset which includes COVID statistics that we plan to aggregate into features. This includes the cumulative cases, deaths, and recoveries.

***TODO – ADD SCREENSHOT***

1. COVID-19 World Vaccination Progress - This is a country- and day-level dataset which includes COVID statistics that we plan to aggregate into features. This includes both the daily vaccinations and daily vaccinations by manufacturer.

***TODO – ADD SCREENSHOT***

1. World Happiness Report up to 2022 - This is a country- and year-level dataset which includes our dependent variable, as well as a few other non-COVID-related features. The dependent variable is a measure of the ‘bliss’ level of a country, based on a measurement technique called the Cantril step. The other features include real GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption.

***TODO – ADD SCREENSHOT***

**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?)**

Our dependent variable will be measured as a change in YOY happiness score from 2019 to 2022. As detailed in (***TODO – add link)***, the happiness score is calculated using a Cantril ladder survey. This is a method whereby survey “respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale.” (***TODO – add citation***). We will then create a dependent variable which represents the % change in the happiness score observed from 2019 to 2022. This allows us to then create a model and see the relative significance and impact of COVID factors, when compared to other more general factors.

Our independent variables will be a combination of COVID data and other happiness factors, listed below. Because we’re trying to estimate the change in the happiness score, all values also will need to be manipulated and aggregated to represent an across-period metric as well.

***TODO – Add detail*** (range of values, data type, etc.)

***TODO – Review & confirm variables below***

* COVID cases, aggregated (% of country contracted? by EOY? Weighted daily avg? Trendline?)
* COVID vaccinations, aggregated (% of country contracted? by EOY? Weighted daily avg? Rate of response?)
* COVID recoveries, aggregated (As % of cases? By EOY? Weighted daily avg?)
* COVID deaths, aggregated (As % of cases? By EOY? Weighted daily avg?)
* Change in real GDP per capita (Compound? 2022 / 2019?)
* Change in social support (Compound? 2022 / 2019?)
* Change in healthy life expectancy at birth (Compound? 2022 / 2019?)
* Change in freedom to make life choices (Compound? 2022 / 2019?)
* Change in generosity (Compound? 2022 / 2019?)
* Change in perceptions of corruption (Compound? 2022 / 2019?)

**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?))**

We plan to break this problem into four main steps, which may be iterative depending on our findings.

***TODO – Add graphic with steps***

In the data cleansing and transformation step, our objective is to create the dataset which we’ll study in exploratory data analysis and ultimately feed to the machine learning model. The activities in this step include performing null and duplicate checks, confirming data types, aggregating the daily datasets to an annual level, joining the datasets and checking join veracity, and creating calculated fields for our variables as-needed.

Next, in the exploratory data analysis step, our objective is to summarize the data in a way that will help us understand anything we need to know before modeling. The activities in this step include univariate summaries (histograms, boxplots), dependent-independent variable summaries (pairplots, correlation heatmaps), and any other simple summaries to get to know the data better. During this step, we may find outliers or other phenomenon in the data which require additional transformation. We will perform any necessary transformations before proceeding to the next step.

In the model development and selection step, our objective is to test various model types and select the model which provides the best results. It’s worth noting that we’ll need to be careful about how we define ‘best results’. Because we want to understand the relationship between COVID features and the dependent variable, we plan to only test model types whose coefficients can be interpreted (linear regression, LASSO regression), and will likely leave out opaque models (Random Forest, Boosting). Activities here include splitting data into training/testing, scaling data as needed, performing PCA if needed for dimensionality reduction, training and scoring models from different families, selecting a ‘best’ model type, and fitting that model on the final data.

Finally, in the results evaluation step, our goal is to interpret the model to answer our research question. The main activity will be evaluating model performance metrics to validate results and analyzing coefficients to understand if and how COVID-19 rates and responses impact country happiness.

**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**

We expect that the results evaluation will provide us insights as to how and to what degree COVID-19 cases, vaccinations, and deaths impact the happiness of a country’s people. Ideally, we see significant coefficients for some or all of our COVID-19 variables, which would tell us how things such as vaccination rate, quickness of response to COVID, and deaths as a percentage of country population impact the bliss level of country. In addition, comparing these coefficients to those of other happiness factors may give us an idea of the level of impact COVID had compared to, for example, GDP or life expectancy.

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

Although COVID-19 specifically is in a much later phase, this could inform future government responses to pandemics from a mental health perspective, informing the degree to which the effectiveness of a response and spread of a disease effects the happiness of the whole country. In addition, it could inform privatized mental health companies (e.g. therapy apps) to prioritize countries with certain features in the event of a pandemic.

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

A picture containing timeline

Description automatically generated

In order to meet the project deadlines stated in the Project Instructions, we plan to work on two-week stints oriented around our Thursday meeting cadence. This would begin with completing the first round of data cleansing & transformation by 10/13, in order to begin EDA. Similarly, we would complete initial EDA by 10/20 in order to begin model development, select an initial model by 10/27 in order to begin results evaluation, and have all modeling complete by 11/10 to complete our final report. This approach frontloads the work, in case the inevitable roadblocks pop up. It also allows for a full 10 days at the end of the project to complete the final report and presentation.

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