

Misinformation and Covid Vaccine Uptake

Scoping Document

Team 32

Aizaan Anwar || Anne Wullenweber || Cori Campbell || Mariana Prazeres || Saran Davies || Yifan Zhang

Business problem, opportunity and impact

Overview of problem

Misinformation on the Covid vaccine has been circulating on the internet and especially on social media since the start of its rollout in December 2020. Although by now vaccines have been offered to entire populations in many countries, uptake has stagnated in different regions and among different demographics across various countries. This stagnation is thought to be partially due to the prevalence of vaccine misinformation ([Loomba et al. 2021](#)). Misinformation refers to *“misleading healthcare information, dangerous hoaxes with false conspiracy theories, and fraud that endangers public health.”* ([The Royal Society](#)). The World Health Organisation has defined windfalls of vaccine misinformation as infodemics: *“[an] overabundance of information - some accurate and some not - that occurs during an epidemic. It can lead to confusion and ultimately mistrust in governments and public health response.”* ([WHO](#)).

Why does this problem matter?

Understanding the factors that lead to different rates of vaccine uptake is crucial to improve vaccination programmes not only for the COVID vaccine, but also for future vaccine rollouts. Given recent advances in rapid vaccine development, the frequency of new vaccines entering the market may increase. Without careful consideration of how to achieve higher vaccine uptake, new means of rapid vaccine development will not be enough to combat future epidemics. Therefore, overcoming vaccine hesitancy and misinformation is of paramount importance for successful vaccination programmes.

Business impact:

The socio-economic impact of effective vaccination coverage cannot be overstated. For example, in the absence of vaccine hesitancy, 236-305 covid-related deaths per million population can be prevented ([Mesa et al.](#)). An economic analysis of 10 vaccines across 94 countries showed that \$586 billion in direct cost of illness can be saved with an investment of \$34 billion. The reduction in morbidity and mortality from vaccination not only leads to long-term savings from prevention of disease, but also significant returns in investment ([Rodrigues and Plotkin 2020](#)). It is estimated that every dollar invested in vaccines over a decade can lead to a return of 16 times the original investment ([Ozawa et al. 2016](#)).

Target Audience

1. **Health authorities and policy-makers** would likely be the primary audience for this report; as vaccine hesitancy has a direct bearing on effective public health policy and government expenditure.
2. **Social media companies** should also be a target audience, as the evidence shows that their platforms are being used to spread misinformation ([Wilson and Wiysonge 2020](#)). In addition to social duty, it is also in their best interest to ensure the continued health of their target user and workforce, so as to maintain their company revenues.

Data

Dataset	Fields	Details
index.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)	<ul style="list-style-type: none">• <code>location_key</code> (String) : identifier for location at national, provincial and/or local granularity• <code>country_name</code> (string) : name of country• <code>country_code</code> (string) : code of country• <code>sub_region_1</code> (String) : province, state or local equivalent• <code>sub_region_2</code> (String) : Municipality, county, or local equivalent• <code>sub_region_3</code> (String) : Locality which may not follow strict hierarchical order, such as "city" or "nursing homes in X location"• <code>place_id</code> (character) : ? Unique identifier for each location	Various names and codes, useful for joining with other datasets

<p>Global_vaccination_search_insights.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>date (Date)</code> : date of Google search • <code>country_region (String)</code> : country • <code>sub_region_1 (String)</code> : province, state or local equivalent • <code>sub_region_2 (String)</code> : Municipality, county, or local equivalent • <code>sub_region_3 (String)</code> : Locality which may not follow strict hierarchical order, such as "city" or "nursing homes in X location" • <code>place_id (character)</code> : ? Unique identifier for each location • <code>sni_covid19_vaccination (float)</code> : All searches related to COVID-19 vaccination, indicating overall search interest in the topic. For example, "when can i get the covid vaccine" or "cdc vaccine tracker". This parent category includes searches from the following 2 subcategories • <code>sni_vaccination_intent (float)</code> : Searches related to eligibility, availability, and accessibility of vaccines. For example, "covid vaccine near me" or "safeway covid vaccine" • <code>sni_safety_side_effects (float)</code> : Searches related to the safety and side effects of the vaccines. For example, "is the covid vaccine safe" or "pfizer vaccine side effects" 	<p>Google search trends</p>
<p>geography.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>location_key (String)</code> : identifier for location at national, provincial and/or local granularity • <code>openstreetmap_id (integer)</code> : ? • <code>latitude (float)</code> : latitude coordinate of location • <code>longitude (float)</code> : longitude coordinate of location • <code>elevation_m (integer)</code> : elevation in metres • <code>area_sq_km (integer)</code> : total area in square kilometers • <code>area_rural_sq_km (integer)</code> : rural area in square kilometers • <code>area_urban_sq_km (integer)</code> : urban area in square kilometers 	<p>Sourced from Wikidata</p>
<p>demographics.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>location_key (String)</code> : identifier for location at national, provincial and/or local granularity • <code>population (integer)</code> : total population • <code>population_male (integer)</code> : male population • <code>population_female (integer)</code> : female population • <code>population_rural (integer)</code> : rural population, available at country level • <code>population_urban (integer)</code> : urban 	<p>Various (current*) population statistics. Sourced from Wikidata, DataCommons, WorldBank, WorldPop, Eurostat</p> <p>*Contains the most recently reported information for each</p>

	<p>population, available at country level</p> <ul style="list-style-type: none"> • <code>population_largest_city (integer)</code>: population of largest city in region, available at country level • <code>population_clustered (integer)</code>: , available at country level • <code>population_density (float)</code>: • <code>human_development_index (float)</code>: HDI, available at country level • <code>population_age_00_09 (integer)</code>: population aged 0-9, available at variable levels of granularity depending on country • <code>population_age_10_19 (integer)</code>: population aged 10-19, available at variable levels of granularity depending on country • <code>population_age_20_29 (integer)</code>: population aged 20-29 available at variable levels of granularity depending on country • <code>population_age_30_39 (integer)</code>: population aged 30-39 available at variable levels of granularity depending on country • <code>population_age_40_49 (integer)</code>: population aged 40-49 available at variable levels of granularity depending on country • <code>population_age_50_59 (integer)</code>: population aged 50-59, available at variable levels of granularity depending on country • <code>population_age_60_69 (integer)</code>: population aged 60-69, available at variable levels of granularity depending on country • <code>population_age_70_79 (integer)</code>: population aged 70-79, available at variable levels of granularity depending on country • <code>population_age_80_and_older (integer)</code>: population aged >=80, available at variable levels of granularity depending on country 	datapoint to date
<p>economy.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>location_key (String)</code>: identifier for location at national, provincial and/or local granularity • <code>gdp_usd (integer)</code>: GDP, in USD • <code>gdp_per_capita_usd (integer)</code>: GDP per capita, in USD • <code>human_capital_index (float)</code>: human capital index 	<p>Various (current*) economic indicators</p> <p>*Contains the most recently reported information for each datapoint to date</p>
<p>epidemiology.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>date (Date)</code>: measurement date • <code>location_key (String)</code>: identifier for location at national, provincial and/or local granularity • <code>new_confirmed (integer)</code>: incident COVID cases confirmed on date • <code>new_deceased (integer)</code>: incident deaths recorded on date • <code>new_recovered (integer)</code>: incident 	COVID-19 cases, deaths, recoveries and tests

	<p>recovered covid cases on date</p> <ul style="list-style-type: none"> new_tested (integer): incident number of individuals tested on date cumulative_confirmed (integer): cumulative confirmed COVID cases to date cumulative_deceased (integer): cumulative deaths to date cumulative_recovered (integer): cumulative recovered COVID cases to date cumulative_tested (integer): cumulative number of COVID tests performed to date 	
<p>facility-boundary-us-all.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> facility_place_id (character): unique identifier for facility provider_place_id (character): unique identifier for provider facility_name (string): name of facility facility_latitude (float): latitude coordinate of facility facility_longitude (float): longitude coordinate of facility facility_country_region (string): country in which facility is located facility_sub_region_1 (String): province, state or local equivalent facility_sub_region_2 (String): Municipality, county, or local equivalent facility_sub_region_3 (String): Locality which may not follow strict hierarchical order, such as "city" or "nursing homes in X location" mode_of_transportation (string): ? primary mode of transport to facility travel_time_threshold_minutes (integer): ? facility_catchment_boundary (string): 	<p>Metrics quantifying access to COVID-19 vaccination sites. Sourced from Google</p>
<p>vaccinations.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> Date (string): ISO 8601 date (YYYY-MM-DD) of the datapoint Location_key (string): Unique string identifying the region New_persons_vaccinated (integer): count of new persons which have received one or more doses Cumulative_persons_vaccinated (integer): Cumulative sum of persons which have received one or more doses New_persons_fully_vaccinated (integer): Count of new persons which have received all doses required for maximum immunity Cumulative_persons_fully_vaccinated (integer): Cumulative sum of persons which have received all doses required for maximum 	<p>Trends in persons vaccinated and population vaccination rate regarding various Covid-19 vaccines.</p>

	<p>immunity</p> <ul style="list-style-type: none"> • <code>New_vaccine_doses_administered (integer)</code> : Count of new vaccine doses administered to persons • <code>Cumulative_vaccine_doses_administered (integer)</code> : Cumulative sum of vaccine doses administered to persons • <code>\${statistic}_\${vaccine} (integer)</code> : Statistic value corresponding to a specific vaccine such as <code>new_persons_vaccinated_moderna</code> 	
<p>health.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>key (string)</code> : Unique string identifying the region • <code>life_expectancy (double)</code> : Average years that an individual is expected to live • <code>smoking_prevalence (double)</code> : Percentage of smokers in population • <code>diabetes_prevalence (double)</code> : Percentage of persons with diabetes in population • <code>infant_mortality_rate (double)</code> : Infant mortality rate (per 1,000 live births) • <code>adult_male_mortality_rate (double)</code> : Mortality rate, adult, male (per 1,000 adults) • <code>adult_female_mortality_rate (double)</code> : Mortality rate, adult, female (per 1,000 male adults) • <code>pollution_mortality_rate (double)</code> : Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population) • <code>comorbidity_mortality_rate (double)</code> : Mortality from cardiovascular disease, cancer, diabetes or cardiorespiratory disease between exact ages 30 and 70 • <code>hospital_beds (double)</code> : Hospital beds (per 1,000 people) • <code>nurses (double)</code> : Nurses and midwives (per 1,000 people) • <code>physicians (double)</code> : Physicians (per 1,000 people) • <code>health_expenditure (double)</code> : Health expenditure per capita • <code>out_of_pocket_health_expenditure (double)</code> : Out-of-pocket health expenditure per capita 	Health indicators for the region
<p>lawatlas-emergency-declarations.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • <code>date (string)</code> : ISO 8601 date (YYYY-MM-DD) of the datapoint • <code>key</code> : Unique string identifying region • <code>lawatlas_... - (integer)</code> : range of boolean values relating to government legal mitigation (NB - individual data names not entered due to covering over 100 schema) 	Government emergency declarations and mitigation policies

<p>oxford-governme nt-response.csv (https://github.com/GoogleCloudPlatform/covid-19-open-data)</p>	<ul style="list-style-type: none"> • date (string) : ISO 8601 date (YYYY-MM-DD) of the datapoint • key : Unique string identifying region • Public measures: integers 0-3 'School_closing', 'workplace_closing', 'cancel_public_events', 'restrictions_on_gatherings', 'public_transport_closing', 'stay_at_home_requirements', 'restrictions_on_internal_movement', 'international_travel_controls' • Financial measures: USD & integers 1-3 'income_support', 'debt_relief', 'fiscal_measures', 'international_support', 'emergency_investment_in_healthcare', 'investment_in_vaccines' • Policy: integers 0-2, 0-3, & 0-5 'Public_information_campaigns', 'testing_policy', 'contact_tracing', 'facial_coverings', 'vaccination_policy' • stringency_index (integer) : Overall stringency index 	<p>Summary of government's response to the events, including a <i>stringency index</i>, collected from University of Oxford</p>
<p>COVID-19 misinformation_ final dataset.csv (https://osf.io/mf7qc/?show=revision) Downloaded: https://drive.google.com/file/d/1TEQbLkTB68u1J4RbKqg60VAIEfhAZ5pU/view?usp=sharing</p>	<ul style="list-style-type: none"> • Country (string): name of country • Gender (binary): gender, either 'male' or 'female' • Age (integer): age at time of survey • Education (string): highest obtained degree • Political affiliation (string): 'Centre right/slightly conservative', 'Middle of the road', 'Centre left/slightly liberal', 'Very left wing/liberal', 'Left wing/liberal', 'Right wing/conservative', 'Very right wing/conservative' • Compliance (string): ranging from 0 to 11 • Trust_in_politicians_approach_effectiveness (ordinal integer): rank ranging from 1 to 7 • Trust_in_WHO_approach_effectiveness (ordinal integer): rank ranging from 1 to 7 • Trust_in_scientists (ordinal integer): rank ranging from 1 to 5 • Trust_in_journalists (ordinal integer): rank ranging from 1 to 5 • Trust_in_govt (ordinal integer): rank ranging from 1 to 5 • Risk_perception (float): rank ranging from 1 to 6.17 • Social_media_exposure (binary): • WHO_media_exposure (binary): • Social_media_info_trust : rank ranging 	<p>Dataset contains information on susceptibility to Covid-related misinformation in studies conducted in April and May 2020. It contains data collected in Ireland, Spain, Mexico, the USA and the UK. It contains ranks for different misinformation types as well as ranks for trust in different organizations, gender, age, education and political affiliation and an indication on the willingness to get the vaccine. Dataset contains 5000 rows, 4473 of which containing at least one NaN value</p>

	<p>from 1 to 7</p> <ul style="list-style-type: none"> • WHO_info_trust : rank ranging from 1 to 7 • Vaccine_self (binary): indicator of whether participant would take the vaccine themselves • misinformation_5g (integer): rank ranging from 1 to 7 • misinformation_breath (integer): rank ranging from 1 to 7 • misinformation_bioengineering (integer): rank ranging from 1 to 7 • misinformation_hot-air (integer): rank ranging from 1 to 7 • misinformation_vaccination (integer): rank ranging from 1 to 7 • Misinformation (float): ranges 1 to 7 	
<p>orb_us.csv https://github.com/sloomba/covid19-misinfo/tree/main/dat</p>	<ul style="list-style-type: none"> • Trust: all boolean values (0 or 1), trust in different sources of information regarding Covid 'Trust:Television', 'Trust:Radio', 'Trust:Newspapers', 'Trust:White House Briefings', 'Trust:State Govt. Briefings', 'Trust:National Health Authorities', 'Trust:International Health Authorities', 'Trust:Healthcare Workers', 'Trust:Scientists', 'Trust:Govt. Websites', 'Trust:Social Media', 'Trust:Celebrities', 'Trust:Search Engines', 'Trust:Family and friends', 'Trust:Work Guidelines', 'Trust:Other', 'Trust:None of these', • Reason: all boolean values, reason not to get the vaccine if person said they were unsure about getting it or would not get it 'Reason:Unsure if safe', 'Reason:Unsure if effective', 'Reason:Not at risk', 'Reason:Wait until others', 'Reason:Won't be ill', 'Reason:Other effective treatments', 'Reason:Already acquired immunity', 'Reason:Approval may be rushed', 'Reason:Other', 'Reason:Do not know', • Misinformation Images: questions while participant sees each misinformation image (information provided for all 5 images) 'Image n:Vaccine Intent':rank ranging from -2 to 2, whether information shown in image makes participant less or more likely to get vaccine 'Image n:Agreement': rank ranging from -2 to 2, agreement with information shown in image, 'Image n:Trust': rank ranging from -2 to 2, how much do you think the information in image is trustworthy, 'Image n:Fact-check': rank ranging from -2 to 2, how likely are you to fact-check information with other sources, 'Image n:Share': rank ranging 	<p>Dataset from a study in the US on whether people would get the vaccine before and after having been shown images containing misinformation on the vaccines. Contains information on trust in different organizations, social media usage and demographics. Dataset contains 4001 rows, 1704 of which containing at least one NaN value</p>

	<p>from -2 to 2, how likely to share image, 'Image n:Vaccine Intent': rank ranging from -2 to 2, influence of information provided in image on vaccine intent</p> <ul style="list-style-type: none"> • Social media related: 'Social media usage': rank ranging from 1 to 7, how much social media the person uses per day, 'Seen such online content' {1, 2, 3} whether a person has seen content as in images online (yes, no, do not know), • Vaccine Intent: rank ranging from 1 to 4, corresponding to: [Yes; unsure, leaning yes; unsure, leaning no; no], pre and post images are shown Would you have the vaccine for yourself if it became available? 'Vaccine Intent for self (Pre)', 'Vaccine Intent for self (Post)', Would you have the vaccine to protect friends, family etc.'Vaccine Intent for others (Pre)', 'Vaccine Intent for others (Post)', • Personal Information: 'Age':int, 'Gender': int, 'Education':int between 1 and 6, 'Employment': int from 1 to 5, 'Religion': int from 1 to 5, 'Political': int from 1 to 3, 'Ethnicity': int from 1 to 5, 'Income': int from 1 to 6, 'Treatment': boolean <p>https://github.com/sloomba/covid19-misinfo/blob/main/doc/orb_questionnaire.pdf</p>	
<p>orb_uk.csv https://github.com/sloomba/covid19-misinfo/tree/main/dat</p>	<p>Very similar to orb_us.csv, only mentioning the differences here:</p> <ul style="list-style-type: none"> • 'Trust:White House Briefings': this column only exists in the US dataset • 'Political': int from 1 to 5 in the UK dataset • 'Ethnicity': int from 1 to 4 in this dataset <p>More information on values in questionnaire: https://github.com/sloomba/covid19-misinfo/blob/main/doc/orb_questionnaire.pdf</p>	<p>Dataset from a study in the UK on whether people would get the vaccine before and after having been shown images containing misinformation on the vaccines. Contains information on trust in different organizations, social media usage and demographics. Dataset contains 4000 rows, 2167 of which containing at least one NaN value</p>
<p>https://data.gesis.org/sharing/#!Detail/10.7802/2272 Downloaded: https://drive.google.com/file/d/1Fvb</p>	<p>Dataset was recorded in two survey waves, variables ending with 1 correspond to wave 1, variables ending with 2 to wave 2</p> <ul style="list-style-type: none"> • 'Vacc_voluntary1': participant would voluntarily get vaccine, int between 0 and 4 • 'Vacc_enforced1', participant would get vaccine if enforced, int between 0 and 4 	<p>Dataset on whether people would get the vaccine voluntarily or if they were forced to in Germany. Contains information on trust in</p>

9JUH2j1U3XAH-e-NPJ8YcPJqZEafd/view?usp=sharing	<ul style="list-style-type: none"> • 'Trust_gov1': int ranging from 1 to 7, trust in government • 'Trust_fed1': int between 1 to 7, trust in federal government • 'Trust_science1': int between 1 to 7, trust in science • 'Trust_media1': int ranging from 1 to 7, trust in media • 'Cov19_truth_gov1': int from 1 to 5, belief that government provides truthful information about covid • 'Public_trust1': int ranging from 1 to 7, trust in public institutions • 'Female1': int, gender • 'TotalCov19_per100k1': float, current number of covid cases in area • 'Vacc_voluntary2', : participant would voluntarily get vaccine, int between 0 and 4 • 'Vacc_enforced2': participant would get vaccine if enforced, int between 0 and 4 • 'Trust_gov2', int ranging from 1 to 7, trust in government • 'Trust_fed2', int between 1 to 7, trust in federal government • 'Trust_science2', int between 1 to 7, trust in science • 'Trust_media2', int ranging from 1 to 7, trust in media • 'Cov19_truth_gov2': int from 1 to 5, belief that government provides truthful information about covid • 'Public_trust2', int ranging from 1 to 7, trust in public institutions • 'Vacc_effective2': belief in vaccine effectiveness, int from 1 to 4 • 'Vacc_freedom2', belief that vaccine compromises individual freedom if enforced, int between 0 and 4 • 'Altruism2' people's willingness to help others, int from 1 to 7 • 'Age2' int, age • 'FederalState_childhood2' which federal german state the person grew up in, int from 1 to 16 • 'East_childhood2' bool, whether person grew up in east or west germany • 'FederalState_today2': which german federal state the person lives in, int from 1 to 16 • 'Female2': int, gender • 'High_education2': bool, whether person has higher education • 'Household_income2': int from 1 to 7 • 'N_household2': int, people in household • 'Single_household2': bool, whether it is a single household • 'Survey_day2': day of survey 	<p>media, science & politics.</p> <p>Dataset contains 2653 rows, 1704 of them containing at least one NaN value</p>
---	--	---

	<ul style="list-style-type: none"> • 'Cov19_risk_group2': bool, person belongs to risk group • TotalCov19_per100k2': float, current number of cases in area • 'Cov19_critical_locally2': whether covid status in area is critical, int from 1 to 9 	
--	---	--

Methods

The goal of this project is quantifying to what extent the uptake of the Covid vaccine has been influenced by misinformation on the internet and on social media across different regions and social groups. Further factors that influence the vaccine uptake (e.g. geopolitical, political sentiment, educational and health status and age/demographics) will be taken into account. To establish a base of comparison, the uptake of the Covid vaccine rollout can be compared to the uptake of previous vaccine rollouts (e.g. Polio, Influenza).

Visualisation

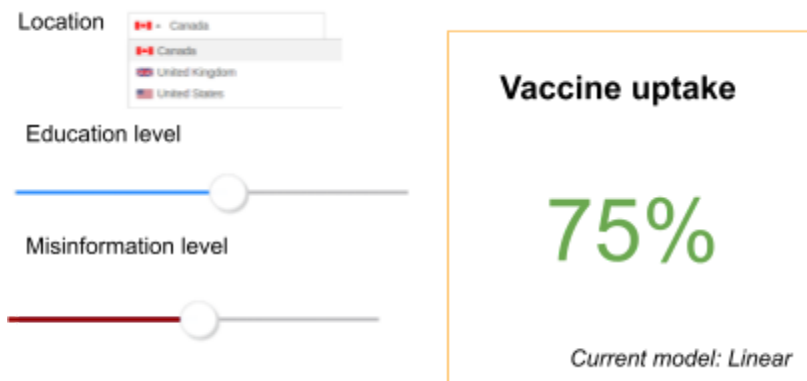
- Initial EDA: Histograms for each factor, for an initial assessment of distribution; Box plots and scatter plots for each factor that can influence vaccine uptake.
- Vaccine uptake per location with time slide and filters for different factors (political, educational, demographics, etc...)



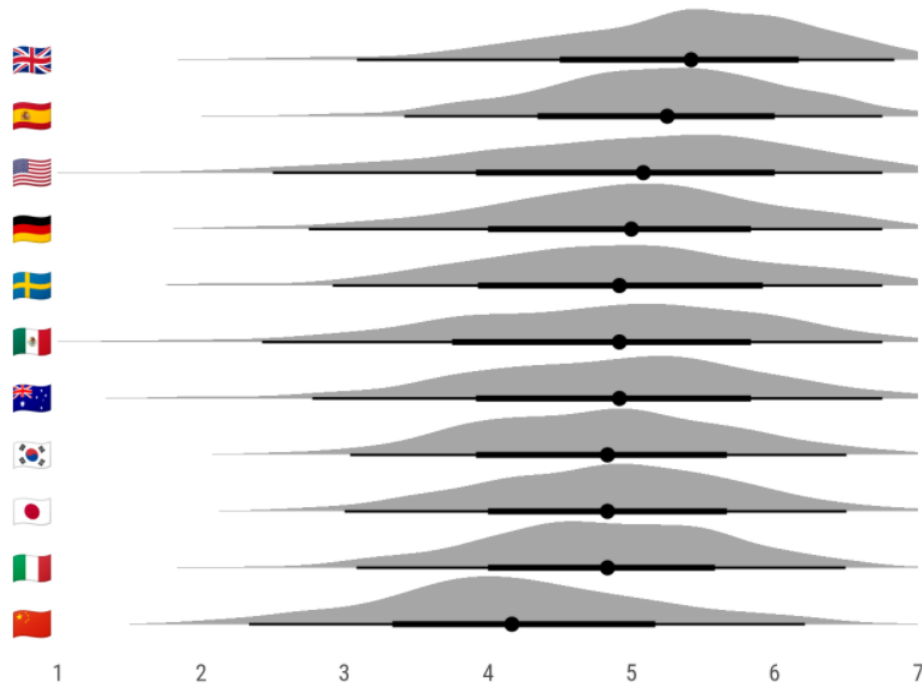
Extra: Clicking on a location will display the time series graph with filters applied and average behavior.

- Correlation heatmap for different factors and vaccine uptake, by prompting the user to select two variables of interest to compare.

- Sliders to “predict” vaccine uptake. This prediction might just be a filter on our existing data or if possible our own model.



- Distribution of the prevalence of vaccine misinformation across country



*note: just an example. Data to be updated.

Models

Next, we want to quantify how relevant misinformation is in predicting vaccine uptake. In particular, we want to answer three questions:

- Question 1: How does misinformation on vaccines affect the *probability* of people getting a vaccine, controlling for other individual-specific characteristics?
- Question 2: How does misinformation on vaccines affect the vaccine uptake rate, controlling for other region-specific characteristics?

- Question 3: How many excess deaths are caused by the misinformation in each country?

Preparation:

- We need to understand what factors are truly independent from our misinformation metrics and which are confounding variables. Our visualizations should help us towards this goal.
 - To examine if one of the variables is redundant, we can do several tests. For example, the correlation heatmap (this is visually vivid, but this can only detect if two variables are correlated), another method is to compute the variability of each variable that cannot be explained by a linear combination of the remaining variables, this is analogous to the R^2 coefficient of a regression of each variable on all the other statistics.
 - To look for instrument variables, related to misinformation but not related to vaccine uptake.

Traditional statistics modelling:

- Model: Logistic model.
 - Data (cross-section): Dataset on whether people would get the vaccine voluntarily or forced, with information on trust in media, science & politics
 - Logistic regression analysis is often used to model the probability of a certain class or event existing and in our case is 'get vaccine' and 'not get vaccine' (binary dependent variable).
 - The Logistic regression also allows us to estimate the marginal effect of each explanatory variable. Specifically, the regression allows us to say, for example, one unit higher exposure to misinformation (or distrust in government) reduces the probability of getting a vaccine by 0.1 percent.
 - Then, we can predict the probability of someone getting a vaccine, given his/her background information (such as education, etc).
- Model: Depending on the datasets, we can choose between t/z-testing for the effect of misinformation: or time-series models such as VAR and Granger Causality tests. We can also produce impulse response functions (IRF) which can be used to visualize the dynamics of vaccine take-up rate following a shock to misinformation.
 - Marginal analysis of polarity of misinformation between locations; age groups, etc...

ML modelling:

We can test different ML regression models on their prediction of vaccine uptake with all our available predictors normalized and scaled when needed:

- Polynomial regression: the simplest nonlinear model, as the relationships are quite unlikely to be linear; To avoid overfitting, Ridge and Lasso regression.
- K-Nearest Neighbors: the idea is to memorize the training set and then to predict the label of any new instance on the basis of the labels of its closest neighbors in the training sets.
- Ensemble tree models: ML models that have been empirically proven to capture most nonlinear relationships in tabular data, if our tree model is not good chances are we are missing important predictor data.
- Feature importance/explainability methods: After modelling, it is important to use typical ML methods to access feature insights such as: permutation importance, model coefficients, mutual information gain, SHAP values, partial plots.
 - Evaluation of prediction results: Confusion matrix/ROC Curve and AUC (overall predictability), Mean decrease accuracy (importance of input variables)

Milestones


In this section, we outline the details on the milestones we hope to achieve in this project. We have listed three different versions: we intend to complete version 1 with 99% probability, version 2 with 70% probability and version 3 with 20% probability.

Version 1: Perform EDA and create a static dashboard from available datasets described in Data to understand variability in vaccine uptake across regions

Version 2: Perform EDA and create an interactive dashboard with additional datasets obtained via crawling of social networks/media to understand temporal dependencies between misinformation spread and vaccine uptake

Version 3: Perform EDA, create an interactive dashboard with all the above data and add models for predicting vaccine uptake

Timeline

Date	Deliverables	Details
Week 1	Team formation	
Week 2	Idea formation & possible project descriptions	 Project Description D... Submitted Sep 25
Week 3	Project selection & Project scoping doc	This doc. Deadline Oct 2

Week 4	EDA, dataset cleaning and merging & first report draft	Deadline Oct 9 for Project report draft (until EDA)
Week 5	Missing EDA, Dashboard & modelling	
Week 6	Dashboard & modelling	Deadline Oct 23 for project report.
Week 7	Presentation and datafolio	Deadline Oct 26 final presentations and datafolios.

Concerns

- Some regions are missing key elements to include in our models and we will have to decide how to deal with missingness
- The subjectivity involved in quantifying misinformation may lead to conflicting conclusions. Also, the effect might be small and confounded, hence difficult to assess. We won't be able to discriminate between misinformation and disinformation