Examining Internet Use and Corruption Among Arab Spring Countries: A Case Study of Palestine, Lebanon & Jordan

[redacted]

Abstract:

This research proposed looks at the relationship between internet use and corruption in the Middle East. It focuses on countries in the Middle East that were involved in the Arab Spring, and protested against corruption, and whether after the Arab Spring, an increase use in internet could be associated with an increased view of corruption. While the Arab Spring is often written about Tunisia, Egypt, Syria and Libya, there is not a lot of reliable data for these countries. In this research I will be using Lebanon, Jordan and Palestine as proxies for countries that participated in the Arab Spring.The datasets used come from Waves 1,3,4,5 of the Arab Barometer. Following research also conducted on the Arab Barometer (Robbins 2017 & Falco et al. 2015), additional x variables that could be seen as confounding the relationship of internet use with corruption were included (elections, economic changes and government and demographic control variables). The research here included extensive model building to get a champion logistic regression (as corruption was coded into a binary variable), whilst also looking at the relationship of the certain interaction effects (age, government trust, wave, country, education and income level) with internet use on corruption. The results achieved showed that there was a statistically significant relationship between internet use and the log odd ratio of corruption. For a one unit increase in internet use is associated with an increase in the log odd ratio of corruption by 18.7%. Therefore, the results show that an increase can be associated with an increment in views of corruption, especially after the Arab Spring. There are also statistically significant effects on corruption between the interactions of Lebanon (compared to Palestine) with more internet use, great government trust (compared to no government trust) with more internet use and wave 3 (compared to wave 1) with more internet use.

Introduction:

Did the Arab Spring help reduce corruption in governments in the Middle East? Has an increase in the internet, and therefore, transparency helped show the corruption in certain Middle Eastern countries? Many have argued that democratic backsliding is an illusion (Levitsky & Way, 2015) as by the 2000s there has been a reduction in the level of authoritarian states and instability. Yet, in the early 2010s, there were multiple anti-governments protests across the Islamic world known as the Arab Spring. The Arab Spring led to regime changes in countries such as Tunisia, Egypt and Libya, however, democracy cannot necessarily be a measurement of success, rather the fact that the protests happened themselves could be seen as a sign of success in such a strong authoritarian area of the world. Within the past few years, the Middle East still seems to be an area with strong corruption and authoritarian regimes, igniting more protests (seen in Iran and in Lebanon in 2019). This leads to wonder if the Arab Spring helped at all, and whether the internet can help increase transparency of the corruption within certain countries?

Background

2.1. How the internet plays a role in politics:

Many have argued that the internet is a 'global village', however, there lies a 'digital divide,' an inequality between those who have access to the internet and those who do not (Burrell 2012). However, over time many more users are connecting from the margins of global economy and regimes. The Middle East itself has had a 38% internet penetration increase from 2009 to 2019 (Puri-Mirza 2019). Zuckerberg argued for internet in India as a route out of poverty for much of India's population (Zuckerberg 2015).

However, the problem that lies within the 'digital divide' framework is that the internet is seen as a black box in terms of access, yet, the internet rarely lacks neutrality (Burrell 2012). Furthermore, the internet is a place where people can post whatever they want, in large quantities. The regulation and structure of content on the internet is often a contentious matter. The organization of information on the internet often reinforces social differences of culture and interaction by just removing the physical boundaries (Chen 2014). Thus, as Burrell (2012) highlights, the internet has never been neutral in policy and the internet is always in flux.

2.2. How the internet plays into non-democracies:

The lack of neutrality that lies in non-democracies can be associated with how these non-democracies have greater control over the information online. The most striking example of control of flow of information today is seen in North Korea, a society that has been indoctrinated since Kim Jong-un's grandfather through the effects of propaganda (Demick 2010). North Korea have their radio dial is fused to state-channel, and since 2008 has set up signal jammers in its 3G network along the Chinese border (Winn 2011). Furthermore, in Russia, whilst not all news is censored like in North Korea, bad news is blamed on external factors whilst good news is systematically attributed to domestic politicians (Rozenas et al. 2019).

The internet in non-democracies is an important motivator for collective action.[1]Especially as many authoritarian countries do not have forms of civic engagement which encourages social trust and norms of reciprocity (Putnam 1995). Collective action shows that coordination is essential to revolution, and coordination requires some common knowledge of shared grievances (Tilly 1978). The problem with coordination is that people in non-democracies know about many grievances, and learning about another one has little impact on a potential revolution, what matters is whether enough people know about specific grievances (Shepsle 1985). The internet has allowed for online rumors and grievances to replace word of mouth transmission, making it easier for people to coordinate and share grievances, creating a larger chance for collective action to take place (Huang 2017).

The internet has been seen as an important factor for motivating collective action. Non-democratic governments try to censor the internet when it might motivate collective action. For example, in China, the CCP is more likely to censor posts that could represent collective action and spur social mobilization than those that are negative criticism of the government (King et al. 2013). Similarly, the Russian government censors information about political events, like popular protests, for fears of collective action (Mikhailov 2011). The reason these governments censor information about collective action, is that when a citizen learns that their private misery is shared by others, it allows them to coordinate on removing the cause of their current misery; the government (Egorov et al. 2009).

[1] "Collective action is any form of organized social or political act carried about by a group of people in order to address their needs." (Collective Action: Definition, Theory, Logic & Problems, 2016)

Literature

3.1. Authoritarian regimes in the Middle East:

The region of the middle East has a history of durable authoritarian regimes. The persistence in post-cold war era came from low internal pressure, due to oil allowing non-effort incomes for countries (Przeworki 1997). The authoritarian persistence can also be attributed to mechanisms that sustain elite cohesions, such as dynastic monarchies (as seen in Jordan and Saudi Arabia). The Middle East can also be seen as an area untouched by the third wave of democratization (Huntington 1991). Some of the cultural explanations for the lack of democratization, has mainly been attributed to Islam being hostile to the ideals of democracy (Bellin 2004). However, some of the non-cultural explanations emphasize that these regimes continue not only because of an international environment that promotes non-democratic values for oil, but because of the citizens themselves (Bellin 2004). Many authoritarian countries in the Middle East are seen as having a skeptical middle class, and a weak civil society (Bellin 2004).

3.2. Arab Spring and its effects:

On December 17th 2010, a street vendor, Bouazizi, lit himself on fire in response to corrupt officials harassing him in Tunisia. Bouazizi had unknowingly sparked the Jasmine revolution which led the Tunisian authoritarian President of 20 years to step down and the flee the country, and overturned the Bourguiba era (Stepan 2013). Furthermore, the events in Tunisia sparked further protests and revolts that defeated dictators in other countries such as Egypt and Yemen. However, not all countries underwent regime changes, as some oppressed their citizens greatly through their tightly controlled coercive-security apparatus (Stepan 2013).

So, one cannot look at democracy as a measurement of success, rather the fact that the protests happened themselves could be seen as a sign of success in such a strong authoritarian area of the world. These revolutions emphasize the overcoming of a weak civil society.[2] Cultural approaches have argued that civil society is rooted in liberalism, thus, making it harder for Arab countries to organize. Civil society is also crucial in the development of collective action, especially in forming connections against the state. This is because without civil society, it is hard to measure the belief that people want to participate in collective action in authoritarian regimes – also known as preference falsification (Kuran 1991). However, the Arab Spring shows that there was enough collective action to ignite protests not only in one country in the Middle East, but across the region, emphasizing that the grievances people faced and the belief that change was possible.

3.3. The internet in the Middle East after the Arab Spring:

The Arab Spring deeply impacted the use of social media and internet access in the Middle East. For example, since early 2011 it seems that official sources of government in Egypt have increased their media presence, as the Army Facebook page has responded to statements and rumors when necessary (Barnsby 2012). Authoritarian leaders, such as Erdogan in Turkey, seemed to have learned from the Arab spring as any post inciting to collective action is automatically deleted and when those posts start spreading, it is the whole social media platform that is shut down (Dogramaci et al. 2015).

The Arab Spring removed four long-standing presidents from different countries, but support for democratic change to regimes seems unlikely. In fact, the youth in Tunisia are seen as more anti-democratic in this new era, potentially highlighting the long-standing corruption within Tunisia's so-called democracy (Robbins 2017). Corruption in the Arab World is strong and needs to be tackled by major reforms to improve the effectiveness of governance (Robbins et al. 2016).

[2] A civil society is known as the sphere of activity that is independent from the state (Encarnacion 2000).

However, access to internet could be more likely to change the way the Arab World works in terms of public spheres. The internet holds a “complex” relationship with the information environment originally controlled by the state (Lynch 2011). In terms of types of regimes, it seems that the internet has no significant effect on its relationship to political Islam, emphasizing that the core of internet usage and grievances may not about challenging the current regime itself (Falco et al. 2015). Rather, access to the internet allows for a slow erosion of the ability of Arab states to monopolize information (Lynch 2011).

3.4. Related Works and Motivation:

Dr Robbins (2017)[3] looks at what are the chances of a shift towards democracy after the Arab Spring, especially in the cases of Egypt and Tunisia. This paper specifically looks at Waves 2 and 3 of the Arab Barometer. Dr Robbins has two models: Concern about democracy and support for political Islam. Concern about democracy is constructed from 3 questions that ask in both waves whether, “Democracy is good at managing the economy”, “Democracy is indecisive”, and “democracy is not good at maintaining stability.” This y variable about concern for democracy ranges from 1 (no concern) to 10 (high concerns). Dr Robbins’ second y variable, support for political Islam, is created from two questions, “it would be better if more religious people held public office” and “men of all religion should have influence in government.” Support for political Islam was coded with a range of 1 (low) to 7 (high).

Dr Robbins then chooses x variables that could influence both democracy and political Islam. The x variables are constructed from the responses for: “Rate your current government performance”, “What are your views about the previous elections”, “What is the degree of government crackdown against corruption”, “What is the current economic situation of your country” and household income. All these x variables were coded as into numerical categories as the responses for these questions ranged a scale between very bad to very good. Dr Robbins also includes standard demographic controls in his x variables; age (only focuses on those aged 18-29), education, sex and urban vs. rural.

Dr Robbins then runs two models, both OLS regression as both dependent variables are not binary, one on democracy and the other on political Islam. For each model, Dr Robbins runs an OLS regression separately for each country and each wave (i.e. in total, he runs 4 different regressions). Dr Robbins’ results for the first model measuring concern for democracy, saw support towards democracy being unchanged in Egypt, whereas in Tunisia there was higher concern for democracy. For the second model measuring political Islam, there is unchanged support in Tunisia, and declining support in Egypt. Dr Robbins’ strength is that he focuses his research on 18-29 as he believes that older generations seem to have fixed political and social preferences.

Dr Robbins’ paper has inspired this project, as the models run, highlight the need to view the results per country per wave, as he explains the variation on each model run between the countries. Thus, in this project, both wave and country are used as factor variables to try and account for differences between waves and countries. Another takeaway from this paper important to this project, are the x variables used, as the standard demographic controls are included in this project, and so are the x variables that are seen throughout the waves I am looking at (i.e. household income, current economic situation of the country, age, education, sex and participation in previous elections).

Falco et al. (2015)[4] investigate the relationship between political Islam, migration and Internet use by looking at the 2nd and 3rd wave of the Arab Barometer. They use data from the Arab Barometer and only focus on 9 countries.[5] Falco et al. have three different models: willingness to migrate, political Islam and Internet use. Willingness to migrate is created from the question, “Do you think about emigrating from your country?” with Yes coded as 1 and No coded as 0. Political Islam is based on three questions that are specifically about whether the governing laws should be in accordance with Islamic law (one is based on penal laws, another on inheritance laws and the last on statute laws), and is coded on a scale of zero (low

[3] Michael Robbins (2017). “Youth, Religion and Democracy after the Arab Uprisings: Evidence from the Arab Barometer.” Journal of the Muslim World.

[4] Chiara Falco & Valentina Rotondi. (2015). “Political Islam, Internet Use and Willingness to Migrate: Evidence from the Arab Barometer.” Peace Economics, Peace Science and Public Policy. DOI: https://doi.org/10.1515/peps-2015-0045

[5] Algeria, Egypt, Iraq, Jordan, Lebanon, Palestine, Sudan, Tunisia and Yemen

support) to 10 (high support). The last y variable used is internet use, coded from the specific question, "do you use the internet?" The answer was recoded as Yes (Daily/Weekly/Monthly/Yearly) = 1 and No (Never) = 0. Depending on the regression ran, the other two y variables were used as x variables.

Additional x variables used were generalized trust in the government (coded as 1 or 0), current government satisfaction (coded 1 to 10) and time spent abroad. Falco et al. also included demographic control variables such as gender (coded as M or F), age (between 18-65), marital status (coded to married or not), household income, employment (coded as employed or not), education (coded as none, primary, secondary and tertiary).

Falco et al. ran three main regressions, two of them being logistic probit models (due to the binary nature of internet use and willingness to migrate), and the last model being an OLS regression (because of the numerical scale of political Islam). In all three models, country and time (measured in waves) were kept as fixed effects to capture systematic differences across countries and to account for changes in surveys between waves. The main results of the regressions were the following; there was a positive significant relationship between Internet use and wiliness to migrate, a negative significant relationship between political Islam and willingness to migrate and no significant effect of Internet use on political Islam.

Falco et al. have inspired this project, as it highlights what Dr Robbins' paper suggested above, that fixed effects need to be run on country and on time to capture systemic differences of countries and of time. Another important inspiration in this paper, is that they code Internet use as a dummy variable, however, due to the increase in the internet use over time, it seems no longer feasible to reduce internet use to a binary variable and so my main y variable is coded to have 4 different levels (none, monthly, weekly and daily). Another takeaway from this paper important to this project, are the x variables used, as the standard demographic controls are included in this project, and so are the x variables that are seen throughout the waves I am looking at (i.e. gender, age, marital status, household income, employment status and education).

3.5. The internet and its potential influences on corruption in the Arab World:

All in all, access to the internet is a way of organizing and facilitating collective action that already exists. Nevertheless, we have seen that in the Arab Spring, it played a central role in the revolutionary movements. Authoritarian governments continue to strive to restrict internet access use when it is done to inspire political change but encourage it when it reinforces their own power (Rozenas et al. 2019). Furthermore, the implications of attitudes towards corruption has become more commonplace, rather than attacking the regime, potentially due to the internet that allows for more transparency and accountability (Robbins 2017, Robbins et al. 2016, Lynch 2011). Thus, this leads to a gap in research which I will explore of looking at the effects of the internet on corruption, whilst accounting for multi-variables including those specified as important in leading the internet transparency (Lynch 2011, Robbins 2017) but not limited to: age and education.

Explanatory Data Analysis

4.1. Initial Data:

I downloaded the data from the Arab Barometer website, and wanted to collect information about countries, that were involved in the Arab Spring. I downloaded wave 1, wave 3, 4 and 5. Wave 1 represents the first time period of 2006-2009, W3 represents 2012-2014, W4 represents 2016-2017 and W5 represents 2018-2019. I decided not to include wave 2 because the internet was used during the Arab Spring to co-ordinate people for protests (Bruns et al. 2013) and therefore did not want to confound the effort of coordinating for protests for the Arab Spring. By this I mean that people protesting against the corruption are more likely to use the internet to co-ordinate, and therefore those who would have used the internet in W2 are more likely to represent people who believe in corruption, as they would have coordinated online for protests. I then filtered for countries that were repeated in each wave that also participated in the Arab Spring which led to filtering for Algeria, Jordan, Morocco, Lebanon and Palestine. Each of these countries had protests during the Arab Spring, and most of these protests were in 2011.

However, after further exploration, both Algeria and Morocco did not participate in a key demographic question (education) and thus both were removed from the datasets.

These are the initial observations and variables for each wave for the three countries of Jordan, Lebanon & Palestine:

W1: 3608 observations with 13 variables

W3: 4195 observations with 15 variables

W4: 4200 observations with 15 variables

W5: 7293 observations with 13 variables

From the above datasets, the only data missing were 2 observations in wave 5 which were omitted so W5 now holds 7291 observations with 13 variables.

4.2. Potential Problems:

The problem with the data is that I want to look at individual responses, however, the Arab Barometer is not a panel study. This means that the Arab Barometer is actually a series cross-section: different samples at different times, even if they are intended to be nationally representative. Thus, by looking at repeated cross-sectional data, it is advantageous in that it allows the investigation of time-varying relationships. However, RCS (repeated cross-section) could affect inferences of causality as additional variables may affect the relationship of internet and corruption within a certain wave, but not in another (Almond et al. 2012). For example, if we look at Lebanon in wave 5 (2018), the country was commencing its spiral towards an economic crisis (seen in full-force today, triggering protests in Nov 2019). Thus, corruption in wave 5 might be more affected by the start of the economic crisis in Lebanon in 2018, than in 2006 (when Lebanon's economy was doing quite well). Therefore, to try and account for systematic differences across time and across country, that might affect causality, I decided to make the variables of wave and country categorical variables for my analysis.

4.3. Recoding Variables:

Firstly, I filtered for the relevant countries of, Jordan, Lebanon and Palestine. As mentioned above, education level was not collected at all in Algeria for wave 5, and was also not collected for Morocco in wave 1 so these two countries were removed from the dataset. I chose to include wave one in my analysis as I wish to look at pre and post Arab spring in terms of internet access and its impact on corruption. Then I selected the necessary variables from each wave. I then recoded the variables.

My independent variable was corruption. Corruption can be seen as subjective, as each respondent was asked if they thought there was corruption, meaning if they thought there was an abuse of government power for private gain. Corruption came from a question in the survey from each wave asking about the corruption with wave 1 being, "Here are some statements that describe how widespread corruption and bribe taking are in all sectors of your country. Which of the following statements reflects your own opinion the best?" with the answers being "Hardly anyone is involved", "Not a lot of officials are corrupt", "A lot of officials are corrupt" and "All officials are corrupt." For wave 3,4 and 5 the corruption question asked in the survey was "Do you think there is corruption within the state's institutions and agencies?" with the answers being "yes", "no", "refused" and "missing". To encompass both questions, corruption was recoded into a binary variable - 0 being none (or a little) and 1 representing corruption (or a lot of corruption).

My main x variable was the use of internet,[6] which was recoded into a categorical variable- 0 being no use, 1 being monthly use, 2 being weekly use, and 3 being daily use. Furthermore, within the use of internet, the

[6] Internet use is a variable accounting for whether individuals use the Internet and how much time they devote to this activity (Falco et al. 2015)

aggregated variable for wave 3 and wave 4 had access to a computer or device recoded as 2 - as increases the likelihood of internet use to weekly (Evans 2017).

Additional x variables were also recoded that might affect corruption. Trust in government could affect corruption, and so the subjective question of, “how much do you trust the government” was recoded into a variable called government trust, with a scale of 0 (no trust) to 3 (a great deal of trust) and included a category for NAs, which may exist because people might not want to publicize this information that might offend the government or be against popular opinion - The number in the Na category for government trust among all waves was 1062. The next x variable that might affect corruption, is economic condition. This variable was coded from “What is the current economic situation of your country?”, and the responses were coded into overall categories of bad and good, and there was also a category included for NAs, again accounting for potential respondents scared of criticizing the government, or criticizing popular public opinion - there was total of 405 number of observations of NAs for all the waves. The next x variable seen as a predictor that might influence corruption is whether a person voted in the previous election, and Yes was coded into voted, whilst all other responses were coded into no.

Control demographic variables were also recoded. Gender was recoded as a categorical variable, as M for male and F for female. Age was split into the following categories (due to W1 being collected in these categories); 18-24,25-34, 35-44, 45-54, 55-64,65-74 and 75+. Education was split into basic, secondary and higher in accordance with international recognized levels of education (ISCED). Education also was recoded to consist of a category for the Nas as there were 1903 observations missing, and this may be due to embarrassment of lack of education. Marital status was also recoded into a categorical variable of married, single or other (i.e. separated/divorced). Employment status was recoded as 0 for unemployed (including students and housewives) and 1 for currently employed. Income was split by the way the question was asked, i.e. whether expenses are covered or not. Income level also included a category for the NAs, with 664 observations, as this could be due to people not wanting to reveal their true income levels - embarrassment may come from it being too low to support a family. Religion was split into a categorical variable of Muslim, Christian and Other, as there was no consistent data collection of other religions throughout the waves (i.e. Druze in W1 was categorized as other, and Jews in W4 was also put in the other category). I then added a column to represent the wave each observation belongs to, and rbinded the data together. I also added the WGI (world governance index) corruption level for each country at the time of the survey as seen in Table 1, where 0 represents corruption and 100 represents clean governance.[7]

Table 1	W1	W3	W4	W5
Jordan	2006- 64.39	2013- 60.19	2016- 60.19	2018- 60.58
Lebanon	2006- 68.54	2013- 19.43	2016- 13.94	2018- 12.02
Palestine	2006- 44.88	2012- 50.24	2016- 51.92	2018- 48.56

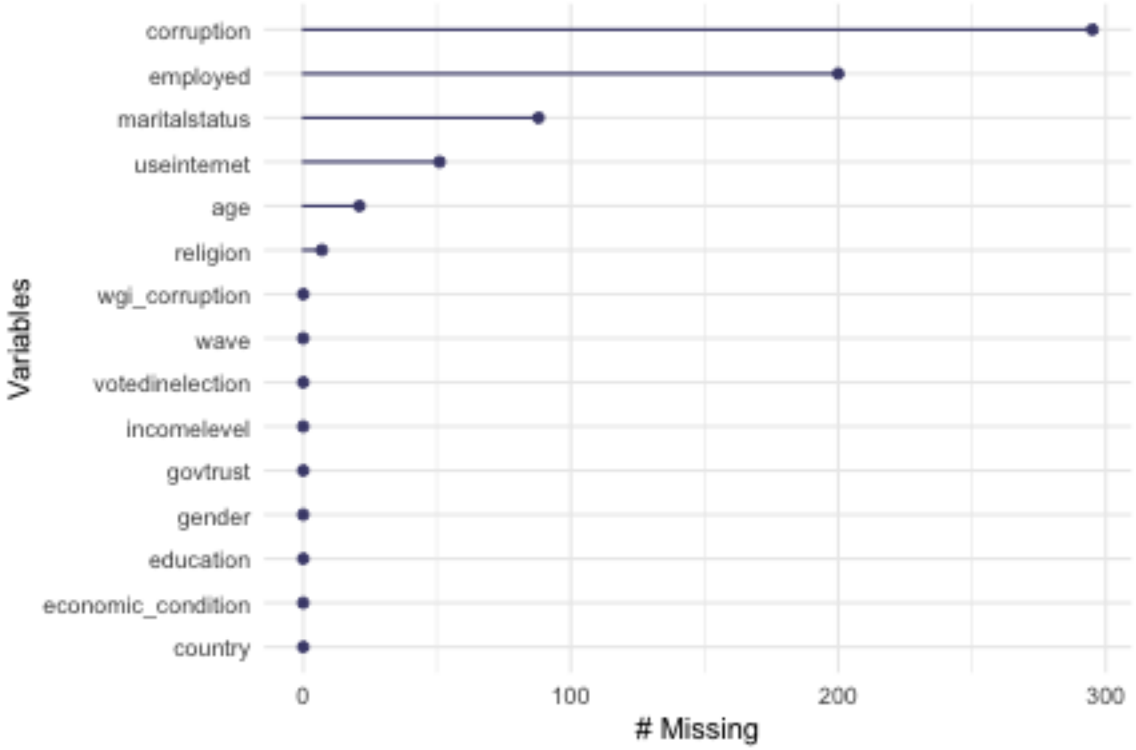
4.4. Missing Data:

After recoding the data includes the following observations of the 15 variables mentioned above (internet use, corruption, government trust, economic condition, voted in an election, age, education, gender, income level, religion, wave, country, employed, marital status, wgi corruption index) as seen in Table 2:

Table 2	# of Obs. (With NAs)	# of Obs.(W/O NAs)	# of Obs. dropped
W1	3608	3331	277
W3	4195	4114	81
W4	4200	4113	87
W5	7291	7086	205
All waves combined	19294	18644	650

[7] Data comes from official website of http://info.worldbank.org/governance/wgi/

In total, 650 observations were dropped from all the waves combined together, as they accounted for roughly 3.4% of the dataset. The graph below shows which variables the observations dropped belonged to, with the majority being from corruption (y variable). Whilst this is a small % of the overall dataset, the fact that the majority of the observations dropped come from corruption could affect what NA means on corruption (i.e. maybe those missing, were people too afraid to even be asked the question). However, since roughly 300 of the total observations dropped were from corruption, this affects less than 1.6% of the dataset and therefore in the big picture analysis does not hold as much as an effect as just mentioned.



4.5. Size of Y Variables:

The % of corruption in all five datasets as seen in Table 3:

Table 3	No Corruption (%)	Corruption (%)
W1	49	51
W3	17	83
W4	25	75
W5	13	87
All waves combined	23	77

These percentages show that in the all waves combined dataset, the corruption vs. no corruption is more or less balanced compared to majority of the waves. However, we also see an overall trend here in that as time passes, the % of corruption increases, emphasizing that wave is an important categorical variable. This is because at the time of wave 1, where the corruption is ~50%, corruption was not a widespread notion. Whereas, wave 3, was taken right after the Arab Spring, when corruption was seen as high as the governments used force to remove protestors. Wave 4 holds lower corruption than Wave 3 and Wave 5, and this could be because maybe the government implemented some of the anti-corruption measures that people protested for, and they have finally been taken into effect. By this I mean if the government implemented a law right after the Arab Spring that is against corruption, by the time of Wave 4, you can see the results of this law. Wave 5 also shows a high corruption %, highlighting that it might be less taboo to officially declare your country's government as corrupt than in all previous waves, and potentially the laws put in place were not upheld, as W5 holds the highest % of those who see their country as corrupt.

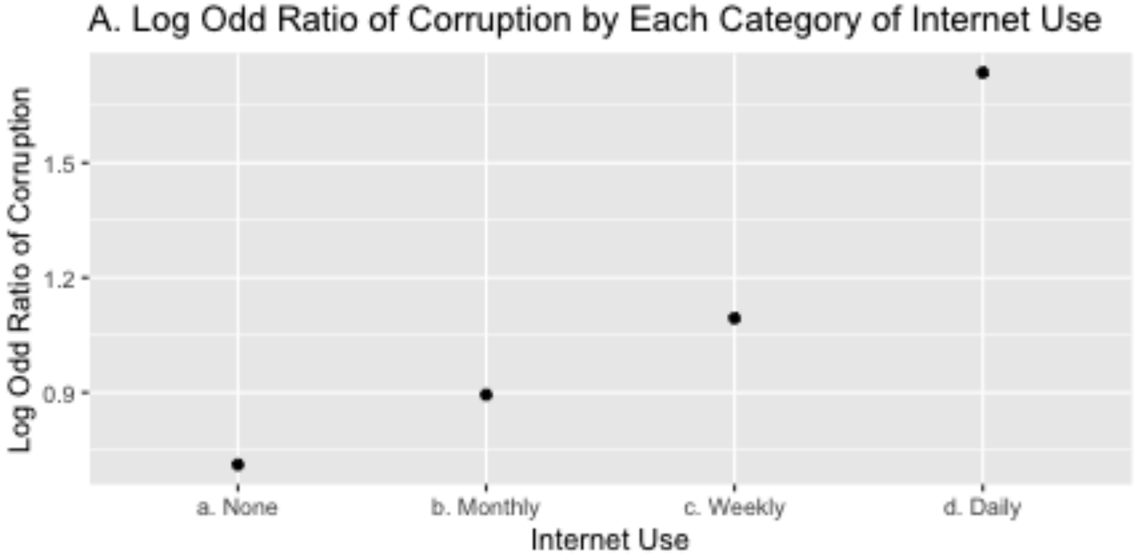
4.6. X Variables & Their Types:

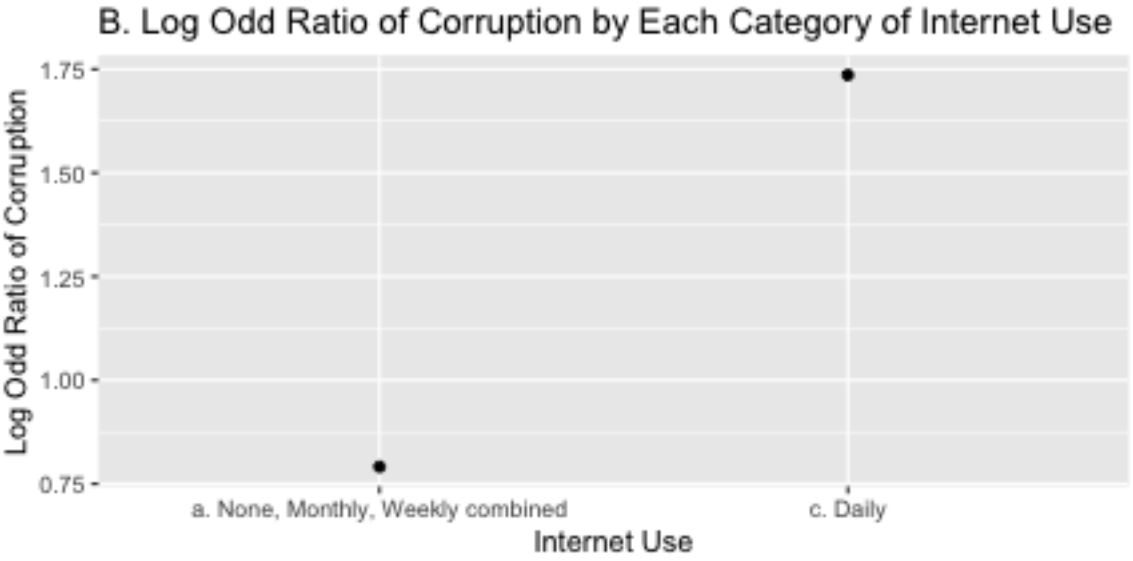
After finally transforming all the variables to have the same categories to be able to be binded in one data frame, the type of each x variable was determined and the balancing of each x variable was observed to see if any transformations could be applied. Upon review those categorical variables with 3 or less categories were coded as categorical (see in the table below). However, the predictors that contained more than 3 categories had to be deciphered whether it was best to keep them as categorical variables or numerical variables. Since I will be running a logistic regression (because my y variable is binary), to view the effect of each category within an x variable (that holds more than 3 categories), the log odd ratio of corruption was used. What this means is that for the x variables that had more than 3 categories, for each category within each x variable a subset was created. Then using this subset, the log was taken of:

$$= \frac{\%\ Corruption}{\%\ No\ Corruption} = \frac{P(Y=1)}{P(Y=0)}$$

This was then plotted for every category of a certain x variable, and if a linear transformation was available it was made, and that x variable was kept as a numeric variable. If it was not possible, then that x variable became a categorical variable.

For example, in the case of internet use, when internet use = 0 (i.e. no internet use), this category was subsetted. Then within this subset, the % of those who believe in corruption, was divided by those who do not. This number was then logged. This was repeated for each category within internet use, which gives the following graphs.





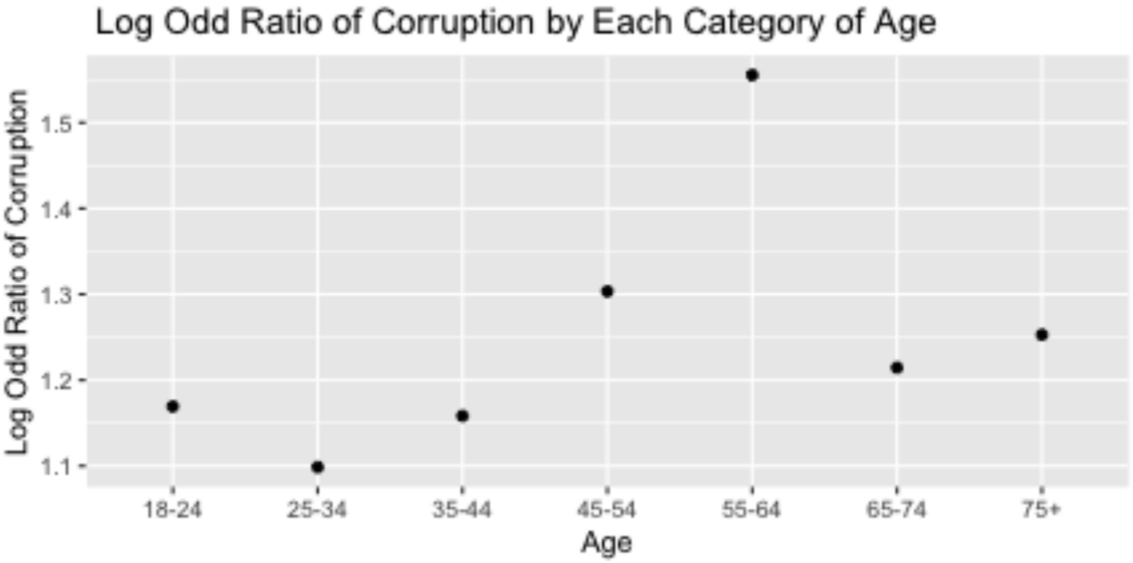
However, as we can see from the first graph of internet use (A), there is no linear relationship between internet use and the log odd ratio of corruption. Furthermore, when looking at the balancing of the observations internet use, it is unbalanced between categories no use, monthly use and weekly use compared to daily use.

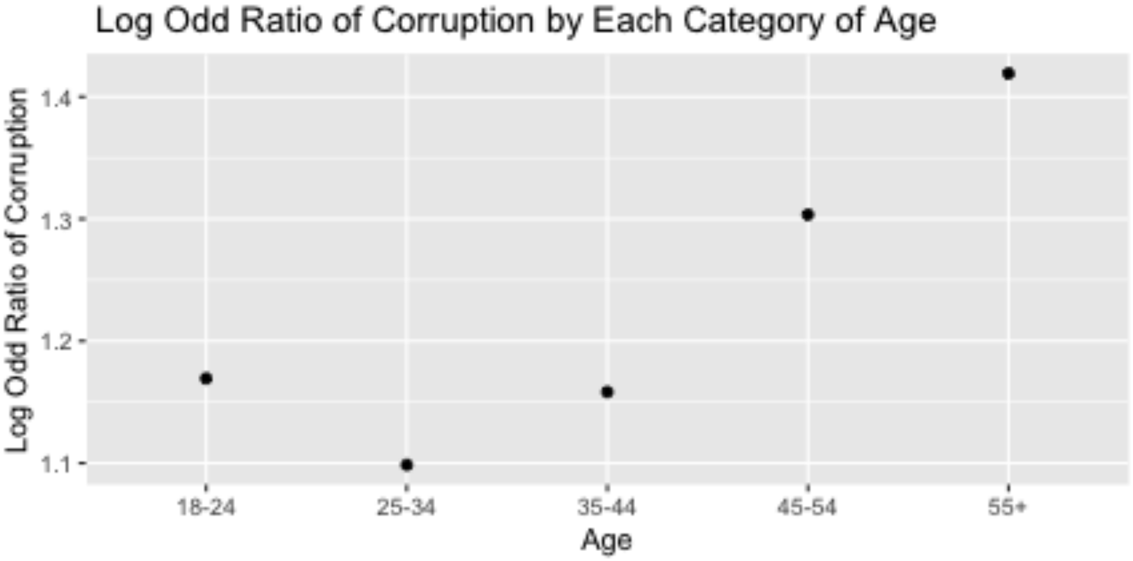
Therefore, I decided to combine the categories of no use, monthly use and weekly use which allows me to not only balance my data (49% no daily usage to 51% - which represents daily usage), but as we can see from the graph above (B), also provide a linear relationship with the log odd ratio of corruption. Thus, internet use was transformed into a numerical variable with 2 categories: 1 (non-daily internet use) & 2 (daily internet use). Table 4 and 5 show how combining the categories improved balancing of internet use:

Table 4 - Unbalanced	# of Observations	Table 5 - Balanced	# of Observations
No internet use	6869	Non Daily Use	9154
Monthly internet use	503	Daily Use	9490
Weekly internet use	1782		
Daily internet use	9490		

Whereas, for example, when it came to age, it was not possible to transform the categories of age to have a linear relationship with the log odd ratio of corruption (see graph C below) and so age was coded as a categorical variable. However, upon looking at the balancing of my data, I combined categories 5, 6 and 7 (which represent ages 55-64, 65-74 and 75+) for greater balance among different age categories. I considered combining just ages 6 and 7, however, this represented only 6% of the observations for age. Thus, upon combination of 5,6,7, they now represent 17% of age observations, and therefore there is now a greater balance between the number of observations for each age category (Table 7). I then reran the relationship of age with the log odd ratio of corruption (see graph D below), and while ages 35+ represent a more linear relationship, the whole graph is not linear and therefore age was coded as a categorical variable.

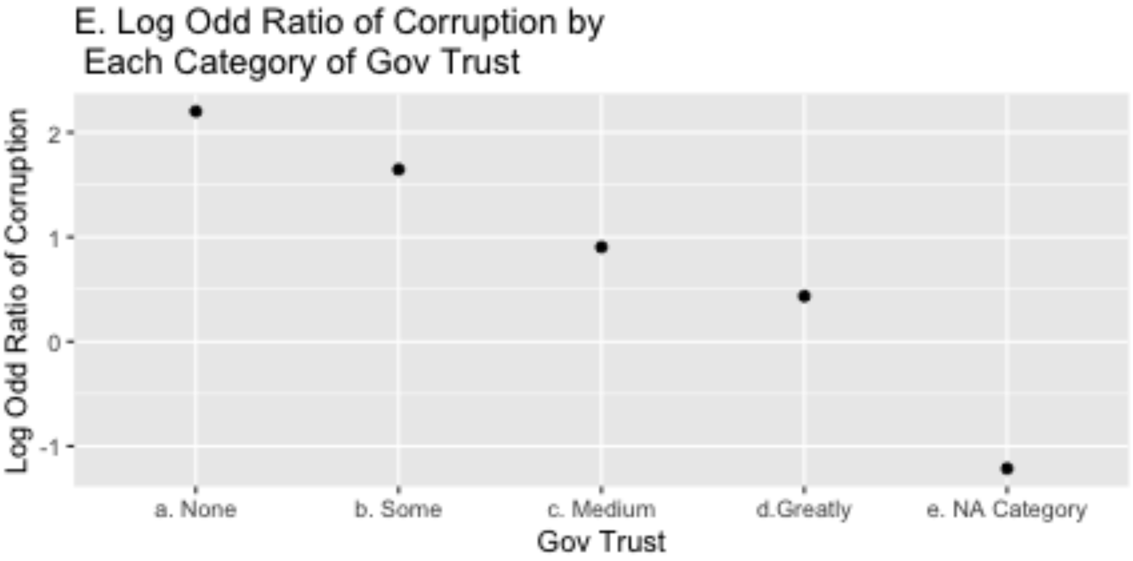
Table 6- Unbalanced	# of Observations	Table 7- Balanced	# of Observations
18-24	3376	18-24	3376
25-34	4604	25-34	4604
35-44	4239	35-44	4239
45-54	3268	45-54	3268
55-64	1939	55+	3157
65-74	948		
75+	270		



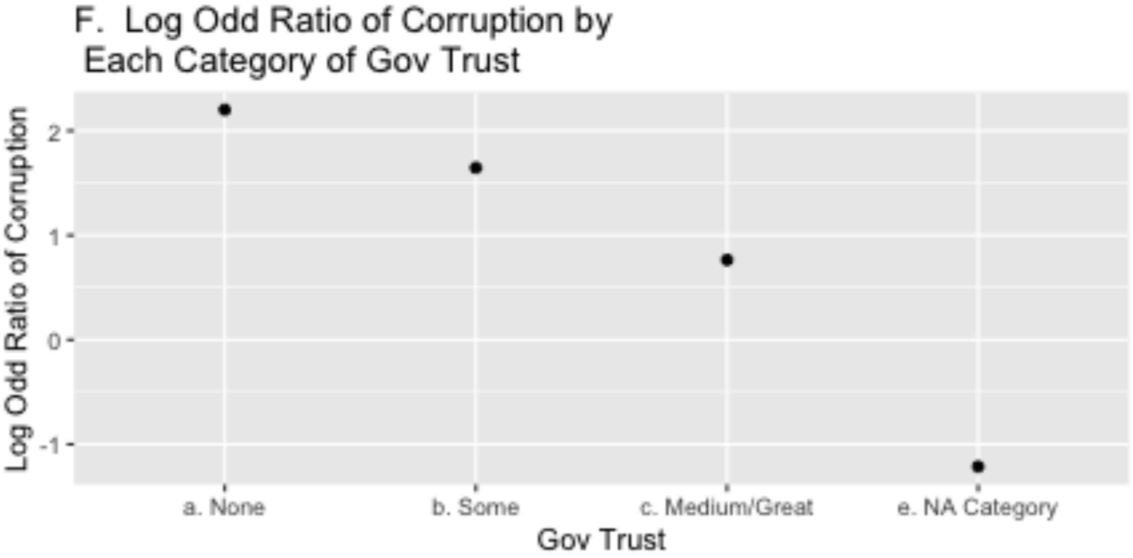


However, two of my variables that were transformed to numeric (gov trust and education) both held a category for NA. Rather than impute for the NAs to make the variables fully numeric, I decided to gain the best linear relationship with the categories within that variable (excluding the NA). This means that for gov trust there was no transformation needed when looking at the relationship of the variables that represent gov trust (None to Greatly without category NA), as it held a linear relationship with the log odd ratio of corruption (Graph E). Furthermore, I looked at the number of observations within each group for government trust, and for better balancing (Table 8 shows unbalanced and Table 9 shows balanced), grouped medium trust with great trust (to form medium-great trust), which also still gave me a linear relationship with the log odd ratio of corruption (Graph F).

Table 8 - Unbalanced	# of Observations	Table 9 - Balanced	# of Observations
No Gov Trust	6813	No Gov Trust	6813
Some Gov Trust	3941	Some Gov Trust	3941
Medium Gov Trust	4957	Medium/Great Gov Trust	6886
Great Gov Trust	1929	NA Category	1004
NA Category	1004		

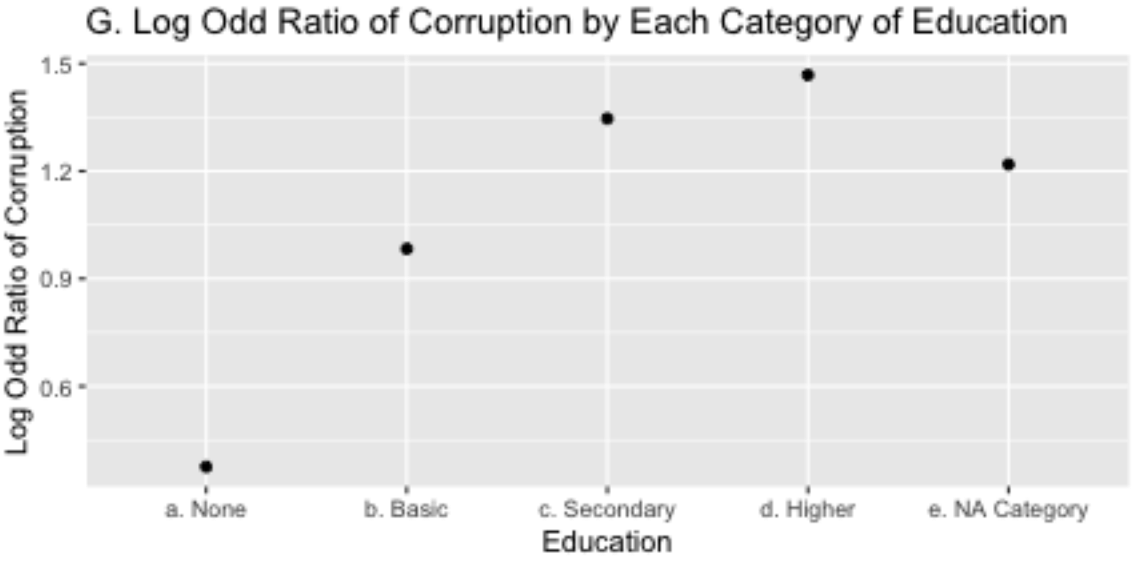


E. Log Odd Ratio of Corruption by Each Category of Gov Trust

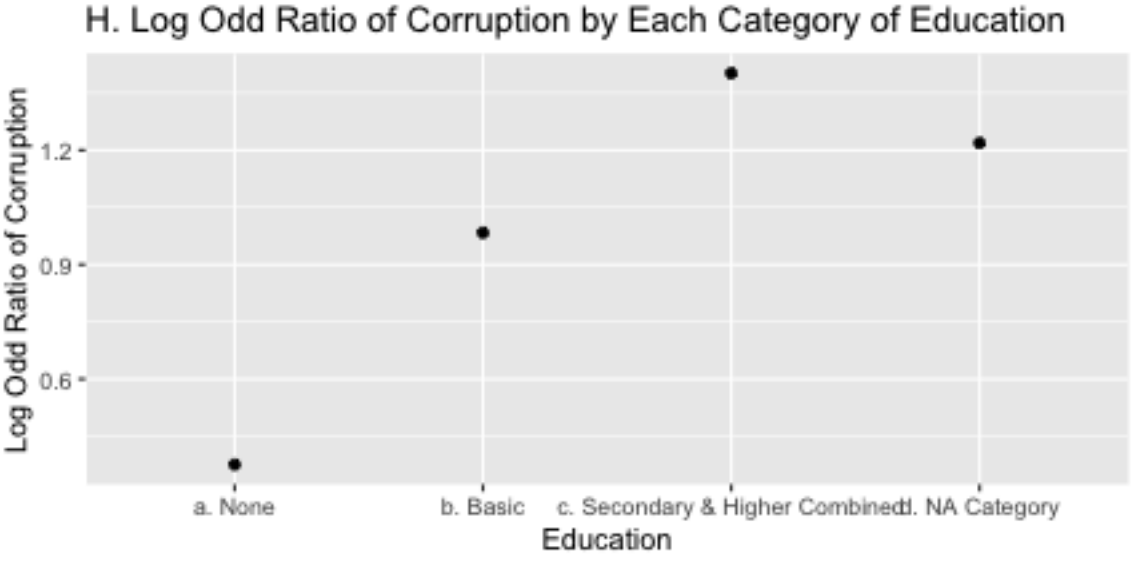


F. Log Odd Ratio of Corruption by Each Category of Gov Trust

Similarly, for education, a transformation of combining two and three into higher education was needed to make the relationship with the log odd ratio of corruption linear, and the NA category was coded as 4 (Graph G). Furthermore, the combination of these two categories, did not change much in the balancing of observations between each category of education (i.e. between no education, basic education and higher education).



G. Log Odd Ratio of Corruption by Each Category of Education

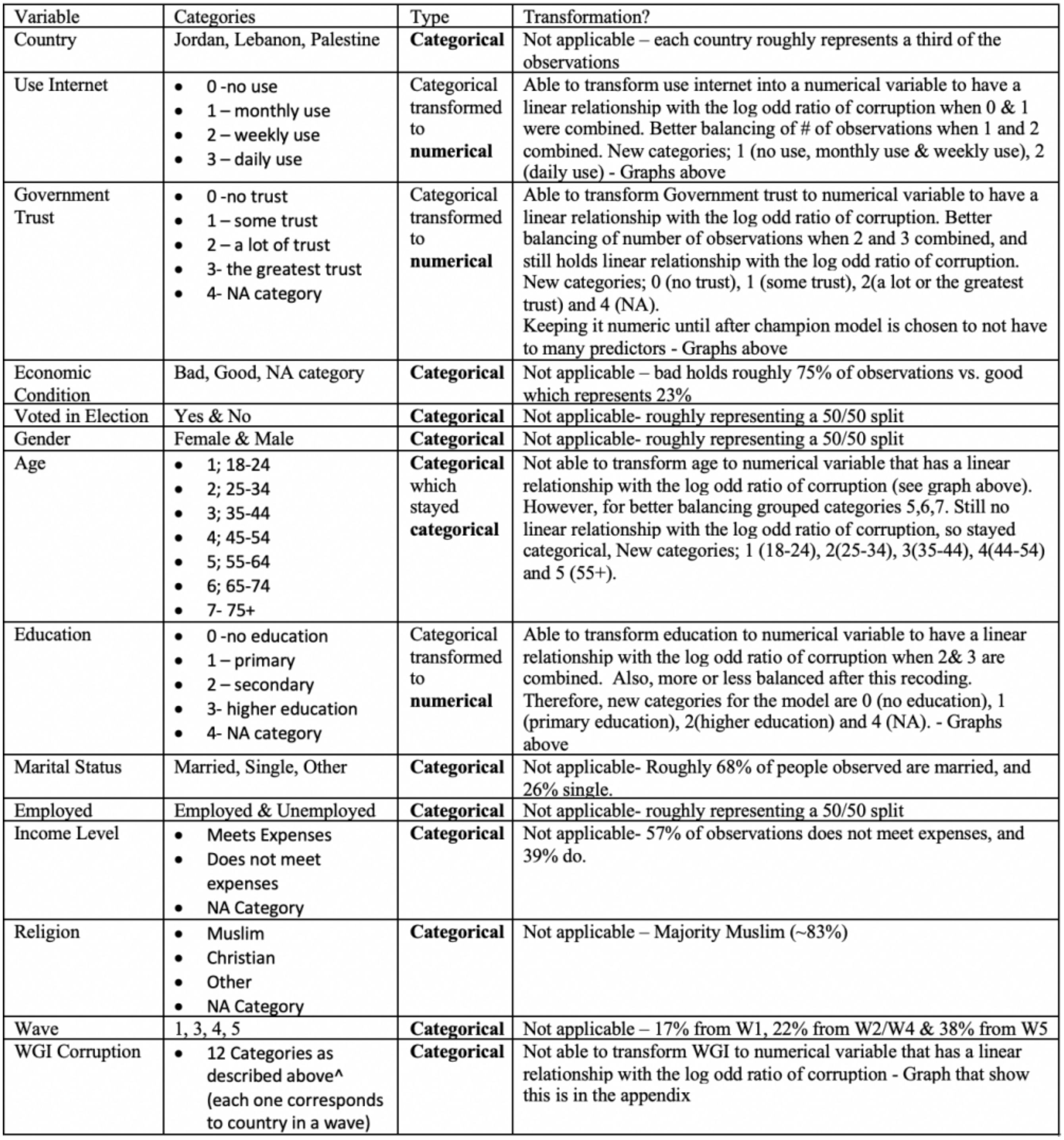


H. Log Odd Ratio of Corruption by Each Category of Education

The reason for keeping the NA in the education and government variables as a category, even if the variable becomes numeric, is that later on, after having chosen my champion model, I can change these variables to categorical variables and it improves the accuracy of my model. To choose my champion model, I need less than 30 predictors, and if I impute the NA categories for these two variables as well as include a dummy variable of whether an observation is in the NA category or not, then it will be too hard to choose the best logistic regression, as the function for choosing my champion model only accommodates for up to 30 predictors.[8]

Here is a table of all the x variables and their types following a potential transformation, that provided my champion model with the greatest accuracy possible (as will be explained in my methods):

[8] When I say more than 30 predictors, here I do not only mean the x variables themselves but the categories within their variables as I am running my champion model on a dummy variable dataset made from all the x variables (as will be explained in Methods Section 5.3)



Variable	Categories	Type	Transformation?
Country	Jordan, Lebanon, Palestine	**Categorical**	Not applicable – each country roughly represents a third of the observations
Use Internet	• 0 -no use • 1 – monthly use • 2 – weekly use • 3 – daily use	Categorical transformed to **numerical**	Able to transform use internet into a numerical variable to have a linear relationship with the log odd ratio of corruption when 0 & 1 were combined. Better balancing of # of observations when 1 and 2 combined. New categories; 1 (no use, monthly use & weekly use), 2 (daily use) - Graphs above
Government Trust	• 0 -no trust • 1 – some trust • 2 – a lot of trust • 3- the greatest trust • 4- NA category	Categorical transformed to **numerical**	Able to transform Government trust to numerical variable to have a linear relationship with the log odd ratio of corruption. Better balancing of number of observations when 2 and 3 combined, and still holds linear relationship with the log odd ratio of corruption. New categories; 0 (no trust), 1 (some trust), 2(a lot or the greatest trust) and 4 (NA). Keeping it numeric until after champion model is chosen to not have to many predictors - Graphs above
Economic Condition	Bad, Good, NA category	**Categorical**	Not applicable – bad holds roughly 75% of observations vs. good which represents 23%
Voted in Election	Yes & No	**Categorical**	Not applicable- roughly representing a 50/50 split
Gender	Female & Male	**Categorical**	Not applicable- roughly representing a 50/50 split
Age	• 1; 18-24 • 2; 25-34 • 3; 35-44 • 4; 45-54 • 5; 55-64 • 6; 65-74 • 7- 75+	**Categorical** which stayed **categorical**	Not able to transform age to numerical variable that has a linear relationship with the log odd ratio of corruption (see graph above). However, for better balancing grouped categories 5,6,7. Still no linear relationship with the log odd ratio of corruption, so stayed categorical, New categories; 1 (18-24), 2(25-34), 3(35-44), 4(44-54) and 5 (55+).
Education	• 0 -no education • 1 – primary • 2 – secondary • 3- higher education • 4- NA category	Categorical transformed to **numerical**	Able to transform education to numerical variable to have a linear relationship with the log odd ratio of corruption when 2& 3 are combined. Also, more or less balanced after this recoding. Therefore, new categories for the model are 0 (no education), 1 (primary education), 2(higher education) and 4 (NA). - Graphs above
Marital Status	Married, Single, Other	**Categorical**	Not applicable- Roughly 68% of people observed are married, and 26% single.
Employed	Employed & Unemployed	**Categorical**	Not applicable- roughly representing a 50/50 split
Income Level	• Meets Expenses • Does not meet expenses • NA Category	**Categorical**	Not applicable- 57% of observations does not meet expenses, and 39% do.
Religion	• Muslim • Christian • Other • NA Category	**Categorical**	Not applicable – Majority Muslim (~83%)
Wave	1, 3, 4, 5	**Categorical**	Not applicable – 17% from W1, 22% from W2/W4 & 38% from W5
WGI Corruption	• 12 Categories as described above^ (each one corresponds to country in a wave)	**Categorical**	Not able to transform WGI to numerical variable that has a linear relationship with the log odd ratio of corruption - Graph that show this is in the appendix

4.7. Initial Exploration of Data:

Upon initial exploration of the data, I looked at the correlations between all the variables. During this exploration, I discovered a correlation of 96% between two of my x variables which are wave and WGI Corruption Index. Since these two x variables are highly correlated, WGI was removed as wave seemed more important in terms of contextualizing the data. The reason for the removal of WGI as an x variable was that otherwise there would be no unique estimate for the coefficients (betas), as the coefficients of WGI and wave will cancel each other out. For that reason, since wave is more important, in terms of analysis as it gives the year and which original dataset the observation belongs to, it is WGI that is removed. This now means I have 14 variables in my dataset: internet use, corruption, government trust, economic condition, voted in an election, age, education, gender, income level, religion, wave, country, employed and marital status.

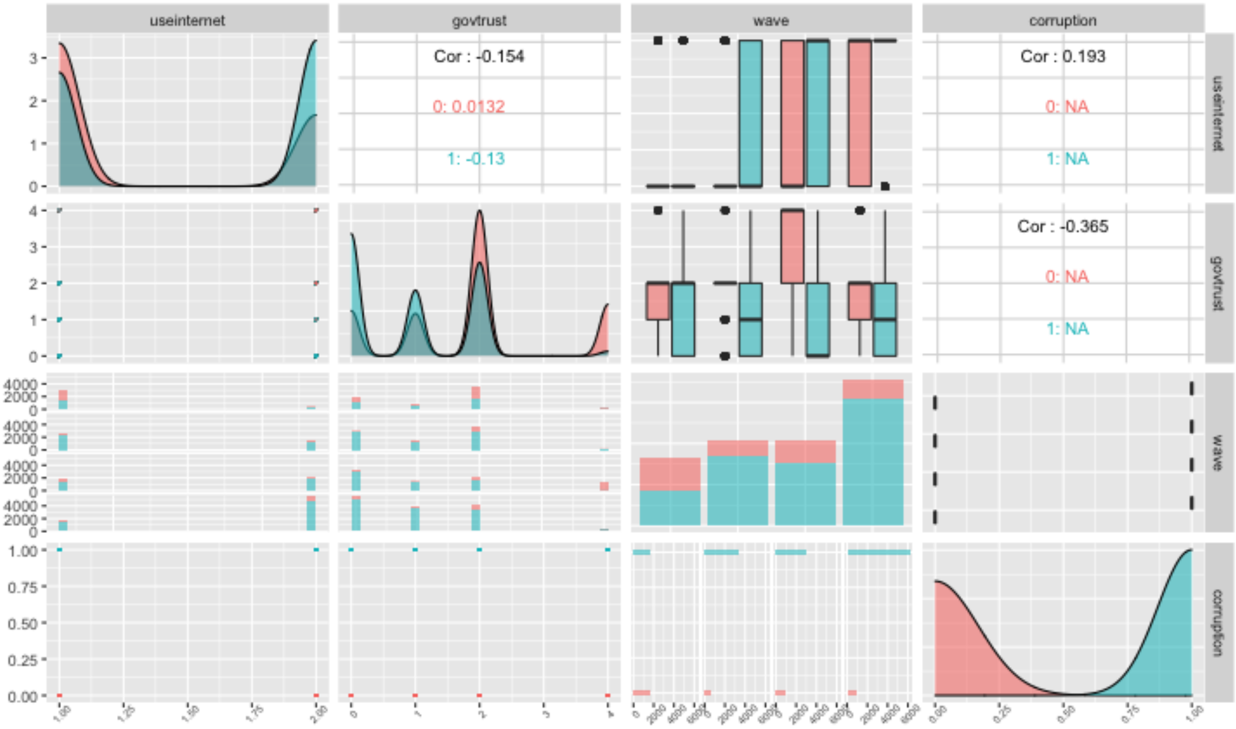
Upon running correlations between my y variable and x variables, the highest correlations among them were the following:

The value for the correlation between corruption and government trust was: -0.365

The value for the correlation between corruption and wave was: 0.243

The value for the correlation between corruption and internet use was: 0.193

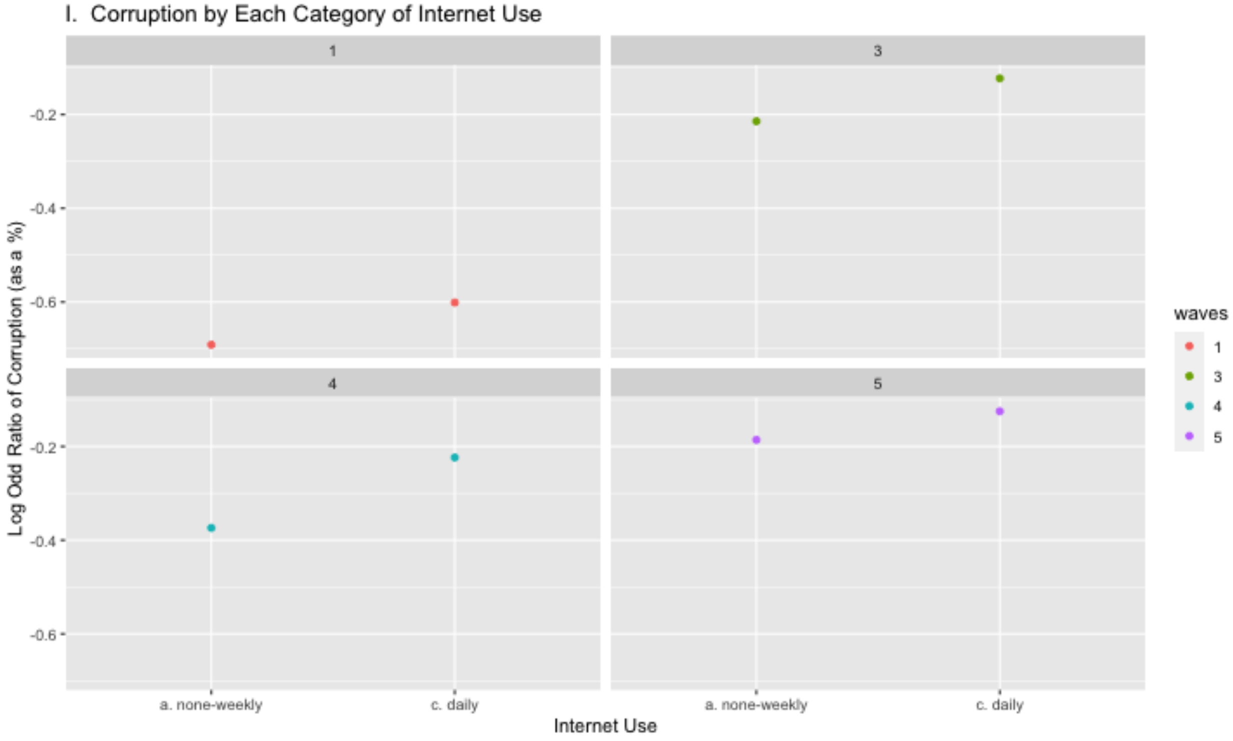
The importance of these correlations is that it suggests that internet use affects corruption could have a high predicting power in the logistic regressions because of its high correlation. The importance of the high negative correlation between government trust and corruption suggests that there is a strong negative trend between these two variables, and that government trust affects people's views on corruption which makes sense, as the more likely someone is to trust their government, the less likely they are to view corruption. Furthermore, the correlation between corruption and wave, also suggests that wave affects the value of corruption, as wave measures the year, and thus the time, which affects the views on corruption (as seen earlier, the % of those who view corruption is the most seen in the last wave vs. the lowest number of those who view corruption is seen in the first wave).



Then from the correlations, the x variables stated above that were seen to have the highest correlations with corruption, (internet use, government trust, wave) were then seen visually, through the function ggpairs, to try and gain further understanding of their initial relationship with corruption. From these graphs, we see that most of the data tends to come from W5, and the corruption vs. no corruption seems to be roughly 50/50 in wave 1, whilst through wave 3-5 have more or less the same balancing of corruption. This observation fits in accordance with Table 3 (size of Y variable). However, an interesting note from this bar chart (3,3 in ggpairs location), is that they seem to carry similar observations but less people believe there is corruption in w4 than in w3, re-emphasizing my earlier point the governments may have listened to the protests from the Arab spring and implemented some regulation of anti-corruption right after the Arab Spring. However, these anti-corruption measures have to be implemented and enforced which takes time and even years, and therefore might have taken years to show less corruption through these regulations, rather than immediately after these regulations go in place (right after the Arab Spring) and so Wave 4 could be capturing the full implementation of these regulations.

When it comes to government trust, (in 2,2 of the grid) it seems that, across the waves, as government trust increases, then belief in corruption decreases. This observation makes sense because if someone believed in corruption in their government, they would be less likely to trust their government. When it comes to visually understanding the relationship between internet use and corruption, we see that no use of internet and lots of use of internet have more beliefs in corruption. This could emphasize that a lack of internet might reduce a lack of awareness from government actions or even that too much internet means there could be a lack of awareness of navigation of the internet (i.e. government-controlled websites such as the CCP in China).

I then also decided to look at the relationship between internet use and the log odd ratio of corruption, by different waves. In the graph below, you can see that the there is little difference in movement between categories and the log odd ratio of corruption for W1. This makes sense as from the original balancing of the data, there was a more or less even split (49-no/ 51 -yes) of those who believe there is corruption. However, waves 3,4,5 seem to hold a similar overall movement in that as the internet usage increases, the log odd ratio of corruption also increases. Furthermore, from graph I, you can see that Wave 1 for both categories of internet use, they hold the lowest relationship with the log odd ratio of corruption, whereas Wave 3 & Wave 5 holds the highest relationship with the log odd ratio of corruption. This is definitely suggestive of a temporal aspect affecting corruption, as suggested above (such as right after the Arab Spring in W3, and lack of upholding anti-corruption regulations from W4 to W5).

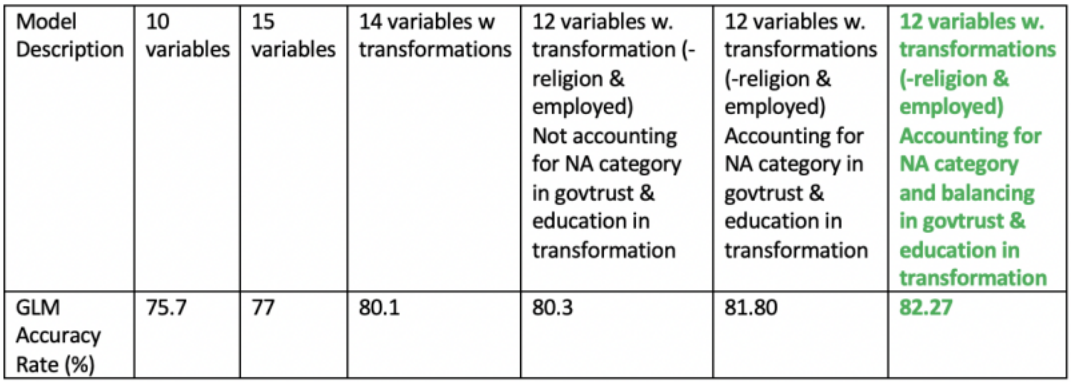


Methods

5.1. Model Building:

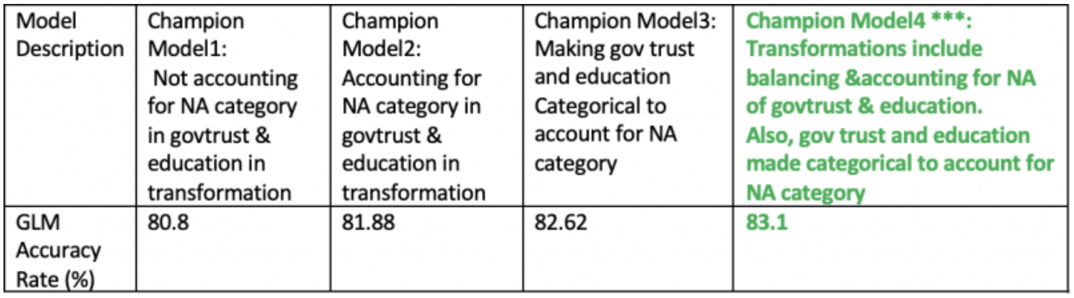
Originally, the following methods were done on 10 x variables (excluding economic condition, government trust, marital status voted in election and employed, but included WGI corruption). However, the accuracy rate for my model was pretty low for the full logistic regression with all of these x variables (75.7%). Therefore, I added the four variables (economic condition, government trust, marital status voted in election and employed) to help improve my accuracy rate for my full logistic model to 77%. To further increase accuracy, I removed WGI corruption (as it had a 96% correlation with wave), and did some transformations to my x variables (i.e. tried to make some of their relationships with the log odd ratio of corruption linear). This increased my accuracy of my full logistic model to 80.1%.

I also decided to drop religion and employed before doing my logistic regression, as these seemed the least important predictors in my random forest, and because I can only have 30 predictors in my dummy variables dataset to be able to choose my champion logistic model (more information below). This increased my accuracy rate to 80.3%. The next step in building my model, was that I needed to account for the NA category in government trust and education (I had recoded them with another level previously), and this increased in my accuracy rate to 81.80%. The final step in my x variable initial transformations was that I decided to balance some of my variables (as mentioned in x variables table) and this also increased the accuracy of my logistic model to 82.27%. To increase the accuracy of my champion model, all of the steps above were taken, with one final step of making education and government trust categorical (as both of these variables had an NA category). This increased my champion model accuracy to 83.1%. The first table below depicts the journey of reaching the best model (in terms of accuracy rate %) for the Full Logistic regression:



Model Description	10 variables	15 variables	14 variables w transformations	12 variables w. transformation (-religion & employed) Not accounting for NA category in govtrust & education in transformation	12 variables w. transformations (-religion & employed) Accounting for NA category in govtrust & education in transformation	**12 variables w. transformations (-religion & employed) Accounting for NA category and balancing in govtrust & education in transformation**
GLM Accuracy Rate (%)	75.7	77	80.1	80.3	81.80	**82.27**

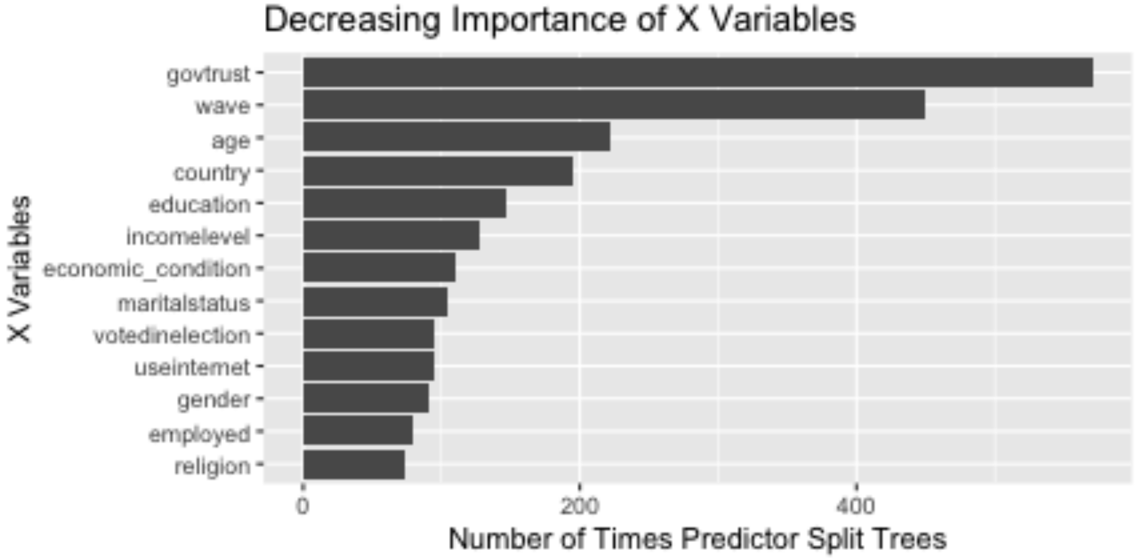
The next table below depicts the journey of reaching the best model (in terms of accuracy rate%) for the Champion Logisitc regression:



Model Description	Champion Model1: Not accounting for NA category in govtrust & education in transformation	Champion Model2: Accounting for NA category in govtrust & education in transformation	Champion Model3: Making gov trust and education Categorical to account for NA category	**Champion Model4 ***: Transformations include balancing &accounting for NA of govtrust & education. Also, gov trust and education made categorical to account for NA category**
GLM Accuracy Rate (%)	80.8	81.88	82.62	**83.1**

5.2. Random Forest:

I ran some initial data analysis through a random forest model to see the predictive power of the x variables. The way random forest works is that a sample of the variables is used at random out of the full set of predictors, and each split is then done through the best of the variables pulled from the data set at each node of the tree. The bottom two predictors (employed and religion) were dropped for the next steps of analysis, which were logistic regressions. The reason I dropped employed and religion is because they never appeared in my champion model and they always appeared at the bottom of the random forest, every time my data analysis was improved. Furthermore, I can only have 30 predictors in my dummy variables dataset to be able to choose my champion model.[9]



I decided to include both random forest and logistic regression. This is because a logistic regression can be interpreted more clearly in terms of the statistical significance of the x variables. Both the random forest and all logistic regressions were run on a non-test set for each method. All non-test sets were split to contain 75% of the observations, as I would like a lot of data to train the model and this is often the conventional split. Each test set then contained 25% of the observations for that method, and were used to measure the accuracy of each model.

5.3 Logistic Regressions:

Because I am estimating the effect of treatment (internet use) on a binary outcome (corruption), I ran logistic regressions. A logistic regression shows the sign, magnitude and significance of using the internet on the log odd ratio of predicted corruption. I decided to run multiple logistic regressions to measure the relationship between the independent and dependent variable. I ran one simple logistic regression for all the waves combined, keeping country and wave as a fixed effect - to account for systematic differences between countries and time. This most basic model, measures the effect of the explanatory variable of interest (use of internet) on the outcome of interest (corruption). The formula for the basic model is the following:

Logg Odd Ratio of Predicted Corruption $= \beta_0 + \beta_1(InternetUse) + \epsilon$

I also ran a full logistic regression to account for all the predictors (x variables). The full regression includes all the variables that might influence the outcome and that might have caused a confounding relationship with believing in corruption (excluding employed and religion). An Anova test was run on both logistic regressions and includes a 'ChiSq' test (as I am accounting for glms). An ANOVA test help to see which is the best model out of these two and which model holds the best predictive power. The ANOVA test below showed that my full model was better (seen by the pvalue less than 0.05). Therefore, this means that one of my additional x variables does not hold a value of 0 for its coefficient, and therefore at least one of additional

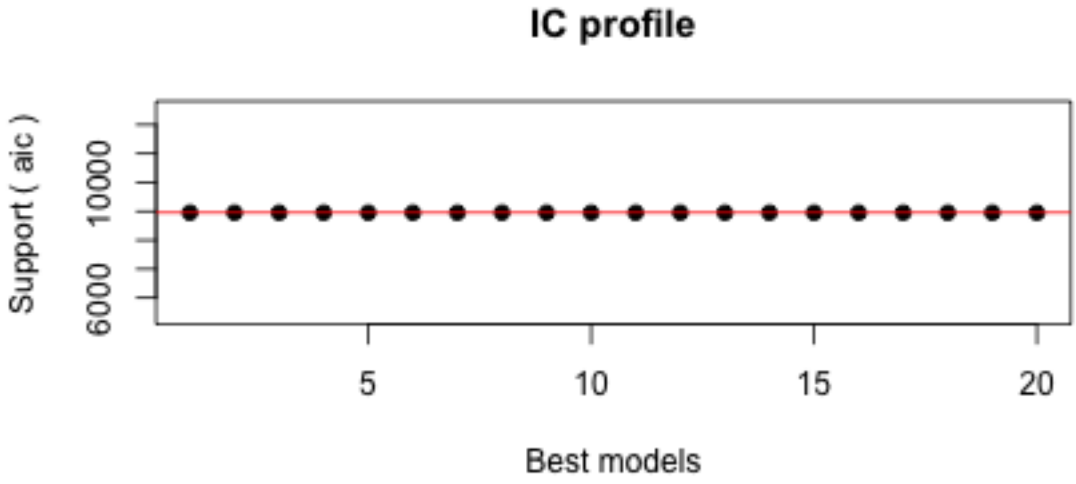
[9] When I say more than 30 predictors, here I do not only mean the x variables themselves but the categories within their variables as I am running my champion model on a dummy variable dataset made from all the x variables (as will be explained in section 5.3 of methods)

x variables is needed in my model. However, whilst the ANOVA test is useful in deciphering that my full model was better than my basic model, I wanted more specificity.

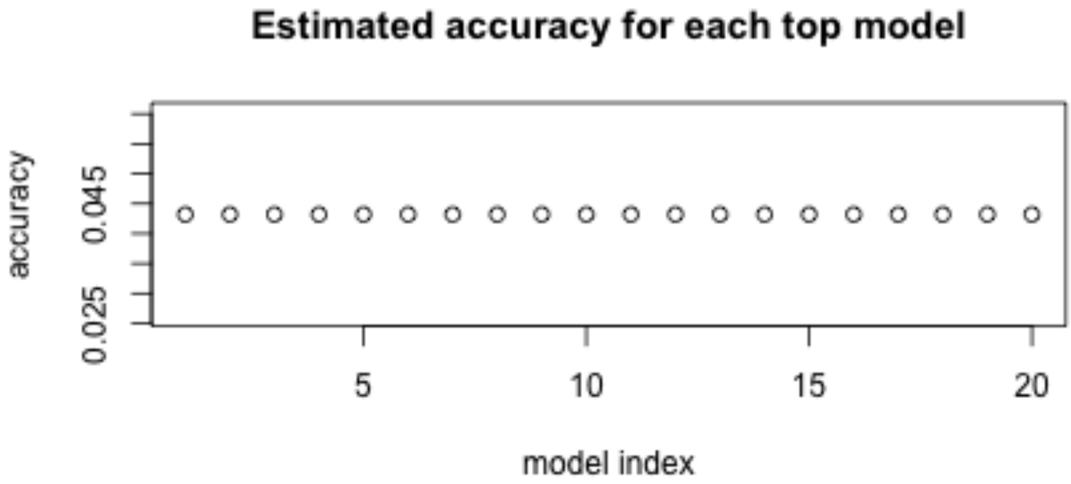
Table 7: ANOVA Table of Basic and Full Regressions of Use of Internet on Corruption

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	13982	14522.45			
2	13963	11861.66	19	2660.79	0.0000

For better model selection, I decided to do cross-validation to select my champion model. I wanted to see if my champion model could be a model with less predictors – by dropping some of the predictors from the full logistic regression- to improve the stability of the estimates given by the model. I started by creating a dummy variables dataset by converting all factor variables to dummy variables. This means that every category within each of my x variables became a predictor in my dummy variable dataset. I then divided the test set to include both a non-test and test set. Then I ran the glmmulti function on my non-test set. The glmmulti function was used to find the best 20 models using a genetic algorithm (seen in the Graph IC Profile below).



From the best 20 models returned, I then calculated the accuracy for each model.



My champion model was the model within these 20 models that had the highest accuracy. However, because this champion model is within the dummy variables' dataset, I recorded the variables, and from my original logistic dataset – that excluded employed and religion- I selected the corresponding variables (in my non-dummy variable dataset) to form my champion model. The champion model held the same accuracy rate as the best model from the dummy variables dataset, and improved the stability of the estimates as there were less predictors than the full logistic regression model. The champion model held the following predictors:

Logg Odd Ratio of Predicted Corruption = $\beta_0 + \beta_1(InternetUse) + \beta_2(GovTrust) + \beta_3(EconomicCondition) + \beta_4(VotedinLastElection) + \beta_5(Age) + \beta_6(Education) + \beta_7(IncomeLevel) + \beta_8(wave) + \epsilon$

Furthermore, I decided to then make government trust and education categorical variables in my champion model, as both included NA as a category. This also increased the overall accuracy of my champion model. I then decided to run a ROC curve of my champion model. A ROC curve shows how much the model is capable at distinguishing between classes (i.e. how good the model is at predicting no corruption as no corruption (0s), and corruption as corruption (1s)).

I also ran some interactive logistic regressions, using the variables that seemed the most important in the random forest and statistically significant in the logistic regressions. I ran the interaction effects to see if there would be any effect of using the internet on belief of corruption, depending on government trust, wave, age, country, education and income level. I decided to visually analyze the interaction effects.

Results:

6.1 Logistic Regressions:

The table below shows the results of the logistic regression for both the full model, and my champion model. We see that in the full model, the coefficient for internet use is 0.2427 and statistically significant the pvalue is less than 0.05 (0.0000324). So, a one unit increase in internet use is associated with an increase in log odds of corruption by 24.8%. The coefficient for internet use in the champion model is 0.187 and is statistically significant as pvalue is 0.01968 (less than 0.05). So, in the champion model, a one unit increase in internet use is associated with an increase in log odds of corruption by 18.7%. We also see statistically significant results in the full logistic model and the champion model with government trust, economic conditions, country, education, voted in previous election, income level and waves. All of the mentioned x variables, therefore have a statistically significant effect on corruption in the full logistic model. Some will further be explored through interaction effects with internet use.

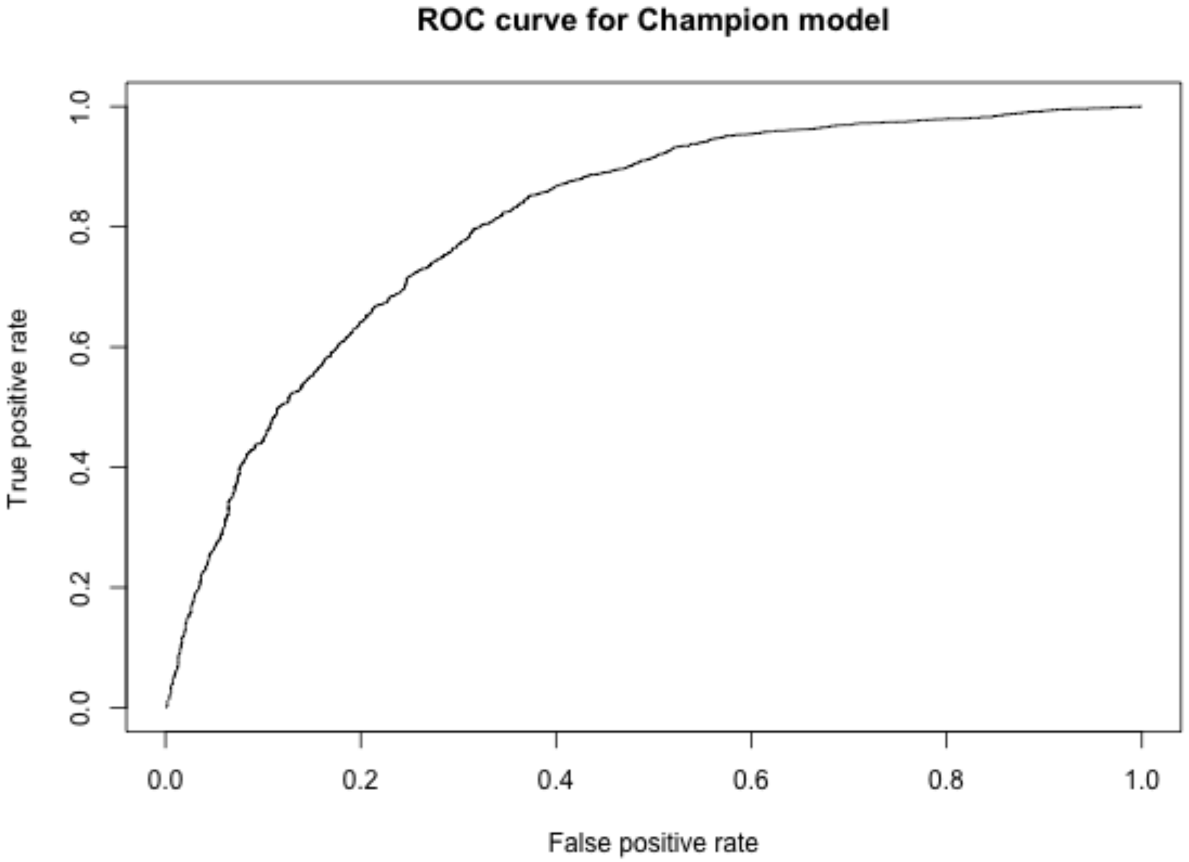
Table 8: Table of Logistic Regressions

	Dependent variable:	
	corruption	
	Full Model	Champion Model
	(1)	(2)
Lebanon	0.394^{***} (0.066)	0.499^{***} (0.068)
Palestine	-0.273^{***} (0.055)	-0.265^{***} (0.055)
Some Trust in Government		-0.579^{***} (0.073)
Great Trust in Government		-1.064^{***} (0.062)
NA Trust in Government		-3.306^{***} (0.108)
Internet Use	0.248^{***} (0.058)	0.187^{***} (0.059)
Government Trust (numeric)	-0.729^{***} (0.023)	
NA Economic Condition	-0.615^{***} (0.142)	-0.700^{***} (0.143)
Good Economic Condition	-0.364^{***} (0.054)	-0.418^{***} (0.055)
Voted in Previous Election	0.307^{***} (0.050)	0.213^{***} (0.051)
Male	-0.122^{**} (0.048)	
25-34	-0.112 (0.079)	-0.053 (0.072)
35-44	-0.077 (0.088)	-0.020 (0.076)
45-54	-0.007 (0.094)	0.067 (0.083)
55+	0.044 (0.099)	0.122 (0.088)
Education (numeric)	0.098^{***} (0.028)	
Marital Status (Other)	-0.038 (0.101)	
Single	0.007 (0.070)	
Basic Education		0.424^{***} (0.115)
Higher Education		0.642^{***} (0.115)
NA Education		0.493^{***} (0.134)
Income Level covers Expenses	-0.269^{**} (0.123)	-0.234^{*} (0.123)
Income Level does not cover Expenses	-0.397^{***} (0.124)	-0.346^{***} (0.123)
Wave 3	1.668^{***} (0.072)	1.727^{***} (0.073)
Wave 4	1.273^{***} (0.076)	1.516^{***} (0.081)
Wave 5	1.789^{***} (0.076)	1.849^{***} (0.077)
Constant	0.865^{***} (0.167)	0.185 (0.189)
Observations	13,984	13,984
Log Likelihood	−5,930.832	−5,874.593
Akaike Inf. Crit.	11,903.660	11,793.190
Note:		*p<0.1; **p<0.05; ***p<0.01

I also looked at the ROC curve of my champion model. The axis of my ROC curve can be defined as the following:

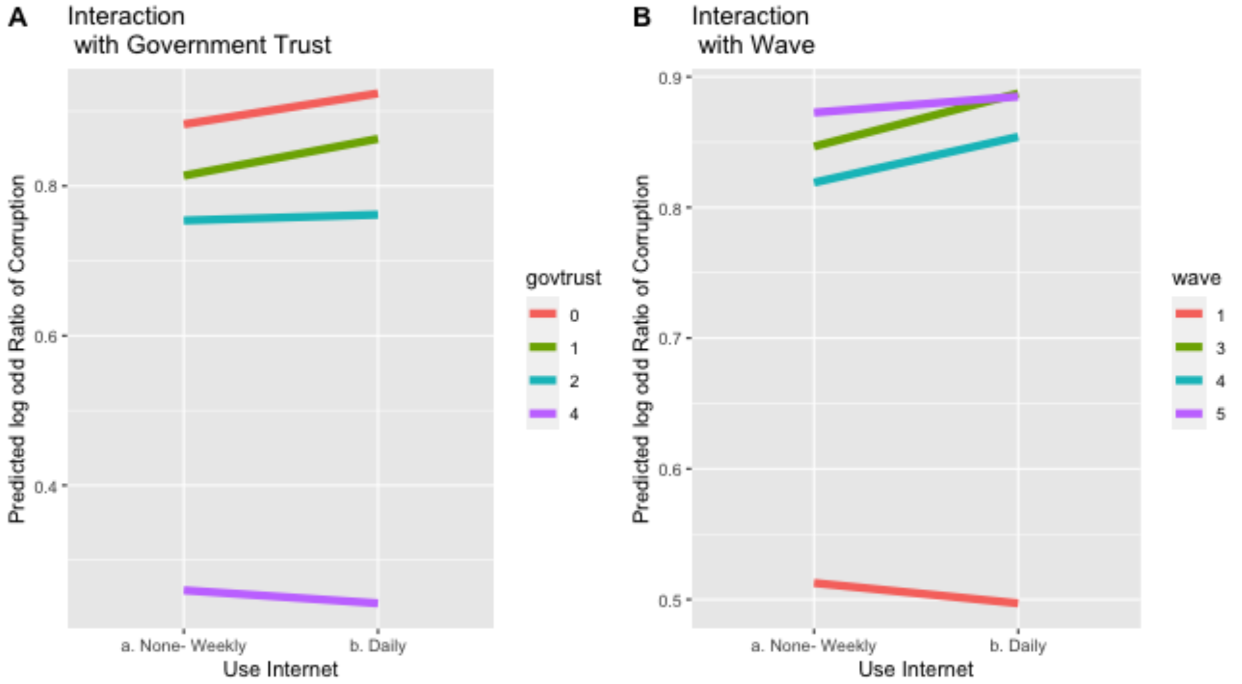
- X axis; Among all the people that are not corrupted; the percentage that my model says are corrupt
- Y axis: Among all the people that are corrupted; the percentage that my model says are corrupted

A ROC curve is seen as optimal when it is closer to the upper left corner, because it means that when the x value (Among all the people that are not corrupted; the percentage that my model says are corrupt) is small, then the y value (Among all the people that are corrupted; the percentage that my model says are corrupted) is high. From looking at my ROC curve, when I want to correctly identify roughly 60 % of people who say there is corruption then among the non-corrupted people I have roughly 20% chance that my model makes the mistake of identifying corruption from those who say there is no corruption.


ROC curve for Champion model
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

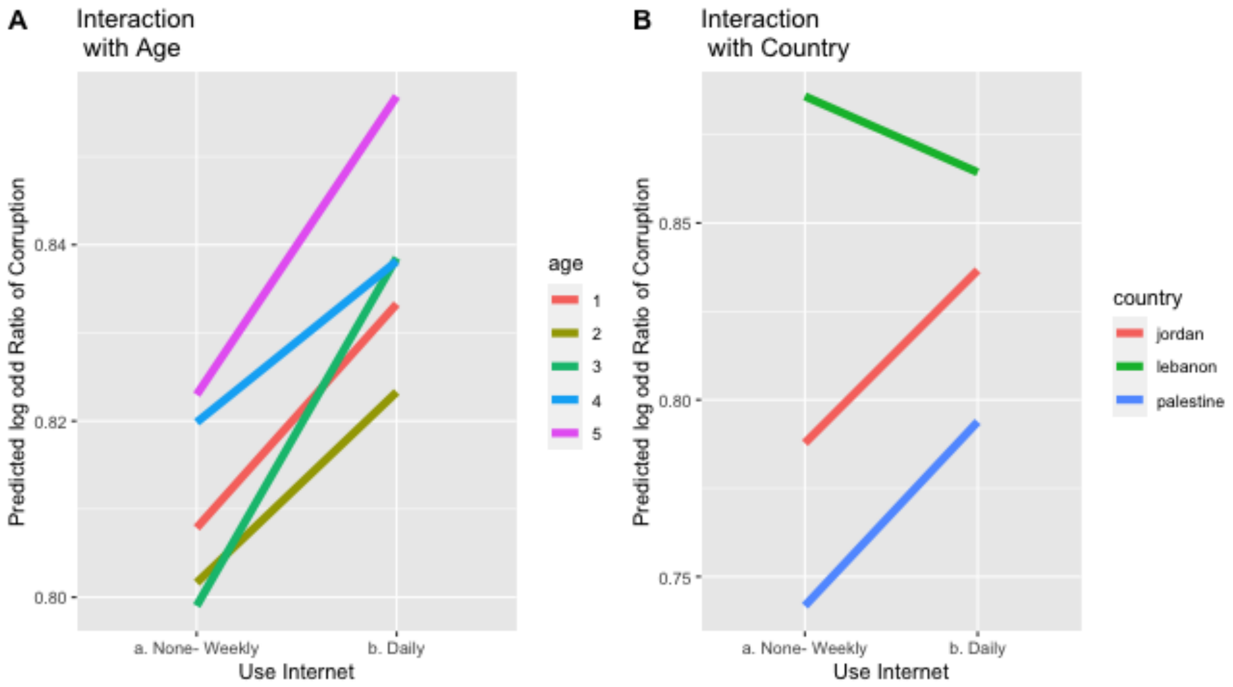
6.2 Interaction Effects:

The interaction effects of each of the variables of government trust, wave, age, country, education and income with internet use was explored. These relationships and trends can be seen in the following graphs.



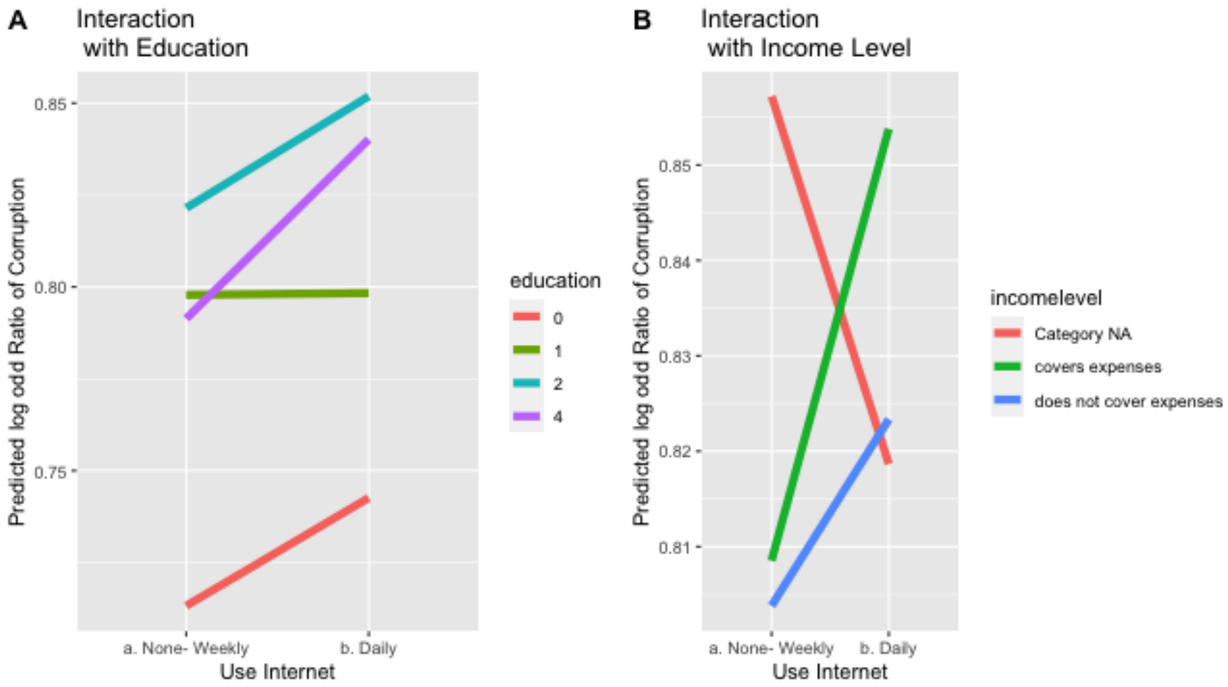
For the interaction of government trust, we see that no trust, some trust and high trust in government interacted with internet use is seen as perceiving high levels of predicted log odd ratio of corruption. Furthermore, the less trust in government one holds interacted with internet use, the higher their predicted perception of the log odd ratio of corruption. There is also a statistically significant interaction effect (pvalue less than 0.05 as it is 0.002706), of great government trust (blue line) and internet use on corruption. The coefficient is -0.36, and therefore, compared to no government trust, there is a significant interaction of a one unit increase internet use for great government trust, that reduces the log odd ratio of corruption by 36%. One interesting observation here, is that the NA category of government trust (purple line) actually hold less predicted log odd ratio of corruption when increasing internet use.

For the interaction of wave and internet use, the results were pretty similar to those expected. As wave 1 (the earliest time period) holds a negative relationship with increase in internet use and predicted log odd ratio of corruption. Whereas, waves 3,4,5 hold a positive relationship with increase in internet use and predicted log odd ratio of corruption. Furthermore, we visually see the potential trend highlighted earlier, that wave 4 holds less perceived predicted log odd ratio of corruption than waves 3 or 5. There is a statistically significant interaction effect of wave 3 with internet use on the predicted log odd ratio of corruption compared to wave 1 (pvalue is 0.0230) and a coefficient of 0.563. This means that compared to wave 1, the interaction of a one unit increase in internet use in wave 3, is associated with a 56.3% increase in in the log odd ratio of predicted corruption.



For the interaction of age and internet use, we only see strong linear positive trends, that as internet use increases, so do predicted perceptions of the log odd ratio of corruption. Those aged 55+, are seen to hold the highest predicted perceptions of the log odd ratio of corruption interacted with internet use. However, there are no statistical significance of any age category interacted with internet use. Therefore, there is little substantial evidence that different ages interacted with internet affects the perceived log odd ratio of corruption with any significance.

For the interaction of country, we see Palestine and Jordan as having positive relationships with increased internet use. Palestine also holds the lowest predicted log odd ratio of corruption.T here is a statistically significant interaction of Lebanon and increasing use internet on the log odd ratio of predicted corruption compared to Palestine, as it has a pvalue less than 0.05, and a coefficient of 0.617. Compared to Palestine, there is a 61.7% decrease in the log odd ratio of corruption, associated with a one unit increase in the use of internet in Lebanon.



For the interaction of education and internet use, the trends are more or less expected. Those with no education interacted with internet use, are seen as having a lot less predicted log odd ratio of corruption than those with any form of education. Those with no education also predict less log odd ratio of corruption with an increased use in internet. The NA category holds similar levels of predicted log odd ratio of corruption as those with some sort of education, suggesting that those within the NA category might hold some form of education. There was also no statistically significant interaction effect of education, and therefore there is little substantial evidence that education might be a big factor in explaining the perceived log odd ratio of corruption when interacted with internet use.

For the interaction of income level, for does not cover expenses, the overall predicted log odd ratio of corruption is lower than those who can cover expenses. There was also no statistically significant interaction effect of income level, and therefore there is little substantial evidence that income level might hold a big factor when interacted with internet use on the the perceived log odd ratio of corruption.

Discussion:

7.1 Limitations of Results and Interpretation:

One limitation, as previously stated in Section 4.2, was that the Arab Barometer is a series of cross-sectional data. To overcome this limitation, I fixed both country and wave by by making them categorical variables. Another limitation was that my models were only run on Lebanon, Jordan and Palestine. Whilst the Arab Spring is often written about Tunisia, Egypt, Syria and Libya (Blakemore 2019), there is not a lot of reliable data for these countries.Lebanon, Jordan and Palestine are good acting proxies for the Arab Spring, as they took part in the Arab Spring (Cannistraro 2011). However, these protests were not on the same scale as Libya or Syria, and are arguably some of the lesser oppressive regimes in the Middle East (Haines 2018). This research should be run again, in the future for the more oppressive regimes, that had a larger scale of demonstrations, when there is more reliable data available.

7.2 Conclusion:

On balance, my results show that an increase in use in internet tends to be associated with an increment of views on corruption, especially after the Arab Spring. The logistic regressions show there is evidence that those who use the internet, are more likely to believe in corruption within their country. This is seen by the positive coefficient in the logistics models, and the positive odds ratio, that suggests that the more you use the internet (treatment) the higher the odds of perceiving corruption.

From the interaction effects, we see that many different variables affect the relationship of increasing internet use on the log odd ratio of predicted corruption. The statistically significant interactions with internet use were respectively Great government trust (compared to no trust in government), Wave 3 (compared to wave 1), and Lebanon (compared to Palestine).

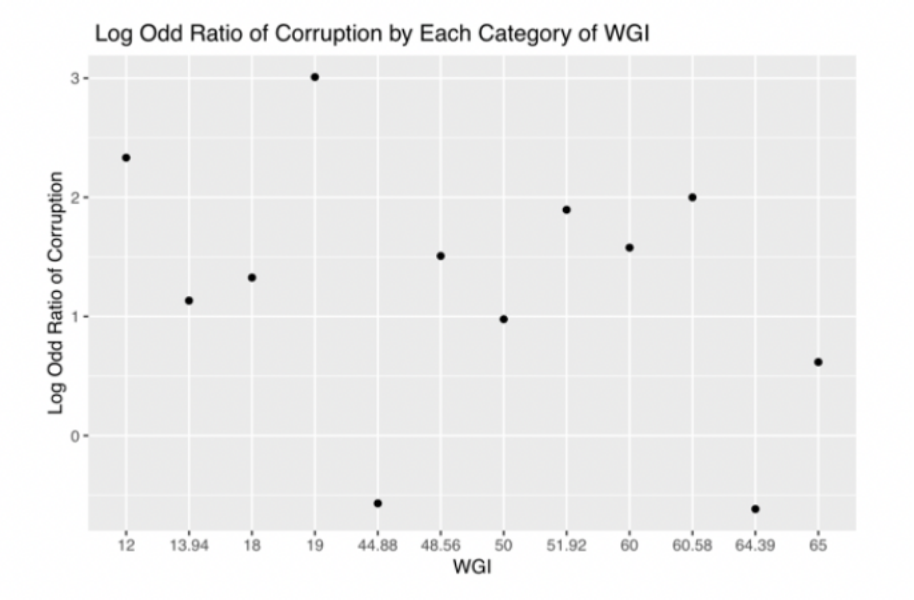
Great government trust is not surprising to significantly decrease the log odd ratio of corruption when using more internet, compared to no trust in government. This could be because, as more people hold a great trust in government, the use of internet actually helps reveal the depths of corruption as more people could be posting more negatively online. It would be interesting for further research to look into the number of negative posts online, and whether there has been an increase after the Arab Spring.

Wave 3 was expected to have such a significant difference. This is because Wave 3 was surveyed right after the Arab Spring, which was protesting corruption (Cannistraro 2011). So, the effects of corruption could be more commonplace as the large-scale demonstrations could have increased awareness of corruption by Wave 3.

Lebanon was also expected to hold the most corruption rate, because it is seen as the most corrupt country on the WGI index. However, what is interesting, is that unlike Palestine, an increase in internet usage in Lebanon actually decreases the log odd ratio of predicted corruption. Therefore, potential further research could look into how these regimes might control the internet, and whether they filter for posts (like the CCP does in China for collective action), or whether the internet is as ‘free’ as it is in the West, or if it might be controlled by the governments in the Middle East.

Appendix:

Transformation of WGI Corruption Index (or lack of)



Acknowledgement:

I thank Professor Daisy Yan Huang for her significant advice and guidance on this project.

This paper represents my own work in accordance with University Regulations 5/12/2020

References:

Almond, R. & Sinharay, S. 2012. “What Can Repeated Cross-Sectional Studies Tell Us About Student Growth?” ETS Research Report Series. 2012. i-20. 10.1002/j.2333-8504.2012.tb02299.x.

Barnsby, RE. 2012. “Social Media and the Arab Spring: How Facebook, Twitter, and Camera Phones Changed the Egyptian Army’s Response to Revolution.” Master’s thesis, Army Command and General Staff College, Fort Leavenworth, Kansas.

Bellin, E. 2004. “The Robustness of Authoritarianism in the Middle East: Exceptionalism in Comparative Perspective.” Comparative Politics 36, no. 2, pp. 139-57.

Blakemore, E. 2019. “What was the Arab Spring and what caused it to happen?” June 14th. Retrieved from https://www.nationalgeographic.com/culture/topics/reference/arab-spring-cause/

Bruns, A., Highfield, T., & Burgess, J. 2013. The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks. American Behavioral Scientist, 57(7), 871–898. https://doi.org/10.1177/0002764213479374

Burrell, J. 2012. ‘Is the digital divide a defunct framework?’ The Berkeley Blog, October 8. Retrieved from https://blogs.berkeley.edu/2012/10/08/is-the-digital-divide-a-defunct-framework/

Cannistraro, V.2011. Arab spring: A partial awakening Duke University Press. doi:10.1215/10474552-1471494

Chen, A. 2014. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed" October 23. Retrieved from https://www.wired.com/2014/10/content-moderation/#slide-id-1593151%20

Collective Action: Definition, Theory, Logic & Problems. (2016, October 15). Retrieved from https://study.com/academy/lesson/collective-action-definition-theory-logic-problems.html.

Countries and Territories. (n.d.). Retrieved from https://freedomhouse.org/countries/freedom-world/scores

Data Downloads. Arab Barometer. Retrieved from https://www.arabbarometer.org/survey-data/data-downloads/

Demick, B. 2010. "Nothing to Envy: Ordinary Lives in North Korea." New York: Spiegel & Grau.

Dogramaci, E., & Radcliffe, D. 2015. "How Turkey Uses Social Media." October 23. Retrieved from http://www.digitalnewsreport.org/essays/2015/how-turkey-uses-social-media/

Egorov, G. &, Guriev, S. & Sonin, K. 2009. "Why Resource-Poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." American Political Science Review 103 (4): 645–68.

Evans, M. 2017."5 Stats You Need To Know About Connected Consumers In 2017." August 22. Retrieved from https://www.forbes.com/sites/michelleevans1/2017/08/22/5-stats-you-need-to-know-about-connected-consumers-in-2017/#11af1cf41962

Falco, C. & Rotondi, V. 2015. "Political Islam, Internet Use and Willingness to Migrate: Evidence from the Arab Barometer." Peace Economics, Peace Science and Public Policy.

Haines, Gavin. 2016. "Mapped: The World's Most (and Least) Free Countries." The Telegraph. Telegraph Media Group, May 24. https://www.telegraph.co.uk/travel/news/the-worlds-most-authoritarian-destinations/.

Huang, H. 2017. "A War of (Mis)Information: The Political effects of Rumors and Rumor Rebuttals in an Authoritarian Country." 47(2) 283-31

Huntington, S. 1991. "Democracy's Third Wave," Journal of Democracy 2, no. 2, pp. 12-34.

International Standard Classification of Education (ISCED). (n.d.). Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED)

King, G. & Pan, J. & Roberts, M. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression," American Political Science Review 107, no. 2 , pp. 326-343.

Kuran, T. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989," World Politics 44, no. , pp. 7-48.

Levitsky, S & Way, L. 2015."The Myth of Democratic Recession," Journal of Democracy, vol. 26, no. 1 (January), pp. 45-58.

Lynch, M. 2011. "After Egypt: The limits and promise of online challenges to the authoritarian Arab state." Perspectives on Politics 9(02), 301-310

Marx, F.M. 1935. "Propaganda and Dictatorship," Annals of the American Academy of Political and Social Science, 179, pp211-218

Mikhailov, O. 2011. "Opposition Protests in Russia: Internet Is Boiling, TV Is Silent." BBC, December 7. Retrieved from: http://www.bbc.com/russian/russia/2011/12/111207_russia_protests_media_coverage.shtml.

Papaioannou, T. & Olivos, H. 2013. "Cultural identity and social media in the Arab Spring: Collective goals in the use of Facebook in the Libyan context." Journal of Arab & Muslim Media Research. 6. 10.1386/jammr.6.2-3.99_1.

Przeworski, A. & Limongi,F. 1997. "Modernization: Theories and Facts." World Politics 49, pp. 155-184.

Puri-Mirza, A. 2019. “Middle East: internet penetration 2019. “ May 28. Retrieved from https://www.statista.com/statistics/265171/comparison-of-global-and-middle-eastern-internet-penetration-rate/

Putnam, R. 1995. “Bowling Alone: America’s Declining Social Capital,” Journal of Democracy 6, no. 1, pp. 65-78.

Robbins, M. & Jamal, A. 2016. “The state of social justice in the Arab World.” Contemporary Readings in Law and Social Justice 8(1).

Robbins, M. 2017. “Youth, Religion and Democracy after the Arab Uprisings: Evidence from the Arab Barometer.” Journal of the Muslim World.

Rozenas A. & Stukal, D. 2019. “How Autocrats Manipulate Economic News: Evidence from Russia’s State controlled TV,” Journal of Politics

Shepsle, K. A. 1985. “Comment of Why the Regulators Chose to Deregulate.” In Regulatory Policy and the Social Sciences, ed. Roger Noll. Berkeley, CA: University of California Press, 231–39.

Stein, R.L. 2011. “The other wall.” Retrieved from:http://www.lrb.co.uk/blog/2011/04/19/rebecca.l.stein/the-other-wall/

Stepan, A., & Linz, J.J. (2013). Democratization Theory and the “Arab Spring”. Journal of Democracy 24(2), 15-30. doi:10.1353/jod.2013.0032.

Tilly, C. 1978. “From Mobilization to Revolution”. New York, NY: McGraw-Hill.

Volt, S. 2013. “Cedar Revolution 2005. Social Movement Theory and Political Opportunity Structure in Lebanon.” Munich, GRIN Verlag, https://www.grin.com/document/301243

Winn, P. 2011. “North Korea’s GPS Jammer Brigade.” September 16. Retrieved from https://www.pri.org/stories/2011-09-16/north-korea-s-gps-jammer-brigade

Zuckerberg, M. 2015. “Free Basics protects net neutrality.” The Times of India, December 28. Retrieved from http://blogs.timesofindia.indiatimes.com/toi-edit-page/free-basics-protects-net-neutrality/