STEPS IN THE ANALYSIS REQUIRED FOR THE ESSAY/PROJECT:

1. IDENTIFY YOUR QUESTION

- Is variable Y related to X1:X5?
- You need to decide what Y, X1:X5 are going to be. Y has to be focused on the topics identified in the assessment question and discussed already in class. All the variables you use have to be available in the datasets we provide. If not, you need to use one that is available.

2. IDENTIFY THE DATASET YOU WILL USE AND GET IT FROM BLACKBOARD

- Every year we provide details as to what datasets you can use and what particular topics you can explore with each dataset, make sure you check you are using an approved dataset/topic combination.

3. IDENTIFY THE SPECIFIC VARIABLES YOU WILL USE

- This will be closely linked to step 1, obviously.
- **You can use a dependent variable measured at any level. But** if you want to use categorical variables with more than 2 levels, you will have to recode for the purposes of your regression analysis. We only cover logistic regression for binary outcomes, so to be able to fit one of these models you will need to recode your choice of categorical variable into a binary outcome.
- **There are no restrictions regarding the level of measurement for your explanatory variables.**
- **The explanatory variables that you use will have to be justified**. This has two levels: the more conceptual one and the operational one.
- At the **conceptual level** you need to explain why the constructs that you use to explain Y should matter. Your literature review will have to do this job. For example, if you want to explain stop and search (as your Y) and want to use gender as one of your independent variables, you will need to be able to explain why gender should matter when understanding the likelihood that somebody is going to be stop and search. Looking at results from previous studies and thinking about theory may help you to provide an explanation for the two variables you select.
- At the **operational level**: you will soon discover that some of the things you are interested in studying may be measured by various variables in the available datasets. For example, in a given survey there may be various variables that focus on fear of crime. You will need to decide which one you prefer to use and have some very basic and explicity defensible rationale for your choice (i.e., the literature thinks one of those measures is better, you DO want to focus on fear of burglary, etc.). In other words you need to discuss why the specific variables you have selected are good measures for the constructs you are evaluating. Often this will be obvious or you won't have much choice (i.e., gender), and therefore you won't need the same level of discussion, but often it will not be so obvious and you will have some choice (i.e., fear of crime, drug use, etc.)
- If there are several measures available for your Y or X and you are unsure which you choose, you may want to do some basic descriptive analysis as

explained below before making a final choice. These early descriptive analysis may help you to understand problems or difficulties associated with using one versus some of the others.

- **Find out who was asked the questions.** This will be more of an issue if you are using some surveys than others. In most surveys some questions are only asked to subset of respondents. For example, in the British Offending Crime and Justice Survey many questions were stratified by aged (i.e., only asked to under 16s). Also some survey questions are “follow up” to previous questions, if you are going to use one of this follow up questions, you will possibly need to do some recoding. To find out who was asked the questions you are using you will need to look at the codebook as explained in class.

4. DELETE FROM THE DATASET ALL VARIABLES THAT YOU WON’T USE

- Once you know what you want you can delete everything else.This will make it easier for you to work with the file.
- **Always keep (DO NOT DELETE) the id variables**. Keeping this may help correct mistakes later on.
- Save the shorter subset of variables under a different filename in you pc/laptop or p: drive.

5. IDENTIFY THE LEVEL OF MEASUREMENT OF YOUR VARIABLES

- This should be pretty obvious, but it is fundamental. The level of measurement determines what tools are appropriate to carry out analysis, therefore, get this right and you’ll be in the right path. Get it wrong and everything else will be wrong. Sounds dramatic, but that’s the way it is. Don’t get it wrong. If in doubt ask.

6. STUDY YOUR VARIABLES

- If they are **categorical**:
  - Run frequency distributions and bar charts.
  - What is the mode? How are the cases distributed across the different categories?
  - Do you need to reorder the factor levels? If you are claiming the variable is ordinal you want them organised in a logical sense (more to less or less to more) that may help interpretation of your results. If you are dealing with an independent variable think about what you want the reference category to be in your regression model and reorder the factor accordingly. If you are dealing with a categorical level think about what is the level that you want to predict in your logistic model.
  - What’s the percentage of missing cases? Is it higher than 5%? What may be driving that? We will talk more about missing cases later on this semester
  - Are there any categories with few cases (say fewer than 30)? If so, does it make sense to collapse some categories? We will discuss collapsing and recoding later on this semester
- If they are **quantitative**:
  - Run summary statistics of central tendency and dispersion.
  - Run histograms and boxplots, for example,
  - What does the distribution look like? Is it normal? Multimodal? Skewed?
  - Are there any outliers?

    - What's the percentage of missing cases? Is it higher than 5%? What may be driving that? We will talk more about missing cases later on this semester
- Are you still sure you want to use these variables? Why not? If not go back to the previous step and select others you may rather work with.
- You may want to start thinking how you want to report the findings from these early descriptive analysis

7. STUDY VISUALLY THE RELATIONSHIP OF Y WITH EACH OF YOUR EXPLANATORY VARIABLES

- **If one variable is quantitative and the other is categorical** look at the section on comparative boxplots or other graphical displays we have used to compare distributions.
- **If both variables are categorical** refer to section on stacked bar charts.
- **If both variables are quantitative** refer to the scatterplot explanatios.
- Don't just produce the graphics, think about the story that the visual displays are telling you
- Think how you could make this graphics prettier. At the very least ensure you label the axis using meaningful descriptors. Experimenting is fun. If you are unsure about something get in touch.

8. SELECT THE RIGHT STATISTICAL TESTS OF SIGNIFICANCE AND OF ASSOCIATION FOR EXPLORING THE RELATIONSHIP BETWEEN Y AND YOUR EXPLANATORY VARIABLES

- **If one variable is dichotomous categorical and the other is quantitative**, this will have been covered when discussing the t test
- **If one variable is categorical (but has more than 2 categories) and the other is quantitative**, this will have been covered when discussing ANOVA and its alternatives
- **If both variables are categorical** this would have been discussed when discussing Chi Square and other measures of association for contingency tables
- **If both variables are quantitative** this would have been discussed when discussing linear regression
- Refer to the document in Blackboard entitled: *"Selecting the Right Tests"*
- Run the tests and interpret them
- Draft a preliminary write up of your results

9. EXPLORE THE MULTIVARIATE RELATIONSHIP BETWEEN Y AND YOUR EXPLANATORY VARIABLES

- Select the right type of regression analysis you need to use depending on the type of dependent variable than you have
- Run the tests and interpret them
- Draft a preliminary write up of your results
- Once you get here you are ready to start thinking about how to write your essay/report