Project Outline Running Draft

Presentation Guidelines

* Running notes to flesh out into a ppt

Title Slide

Include the name of the Project and Group Members

* Motivation & Summary Slide
* Define the core message or hypothesis of your project.
* Describe the questions you asked, and why you asked them
* Describe whether you were able to answer these questions to your satisfaction, and briefly summarize your findings

Initial story:

Overarching question: general curiosity on spending on cigarettes and their associated metrics such as mortality, and exploring these questions on a global scale

[why are we asking this question?]

Curious to see whether a country’s reported affluence has an impact on cigarette spending

And mortality rates in those countries

Though we were somewhat limited in our data exploration due to the nature of available and easily understood datasets, we felt that we’re able to ask good questions of our data

And exercise our current and growing data analytics skillset

[overall summary statement here once we get it]

* Questions & Data
* Elaborate on the questions you asked, describing what kinds of data you needed to answer them, and where you found it

Our questions [add more here as needed]

(main questions in numbers, sub questions/hypotheses in letters)

1. Finding relationship between a country’s affluence, and higher incidence of cigarette spending, and mortality
   1. Country’s affluence and higher incidence of cigarette spending (hypothesis: the greater the affluence, the higher incidence of spending)
   2. But this higher incidence of spending has no relation on mortality

To answer these questions, we needed data on global spending on cigarettes, percentages of reported smokers in given populations, how many packs smoked per day, the ages of those smokers, mortality rates from smoking, and overall country spending as a reference point, again with the knowledge that such datasets would be incomplete

Because data collection across countries is not as uniform due to uncontrollable issues such as reporting inconsistencies and missing data.

We found multiple aggregated csv datasets that addressed various aspects of these questions, but not in a uniform way and in a single file

We also got a dataset [from where, insert here] that already determined a country’s income level (high income generating, low income generating, etc.)

Looking beyond other sites

* Data Cleanup & Exploration
* Describe the exploration and cleanup process
* Discuss insights you had while exploring the data that you didn't anticipate
* Discuss any problems that arose after exploring the data, and how you resolved them
* Present and discuss interesting figures developed during exploration, ideally with the help of Jupyter Notebook

Exploration: we looked up how others visualized similar data in the past, and whether datasets were publicly available

Initial struggle trying to find openly usable and readable datasets, especially with the global scope of our data

So we found aggregated data from multiple sources

[list out sources]

And then combined them into a single csv to begin the data cleaning process

Data cleaning highlights:

with Yifu’s help, we’re merged our csv files using a function

we merged 6 csv files using outer joins, and we naturally get NaN’s for empty values as a result of datasets that don’t line up and with missing information

primarily Nick

* Data Analysis
* Discuss the steps you took to analyze the data and answer each question you asked in your proposal
* Present and discuss interesting figures developed during analysis, ideally with the help of Jupyter Notebook
* Discussion
* Discuss your findings. Did you find what you expected to find? If not, why not? What inferences or general conclusions can you draw from your analysis?
* Post Mortem
* Discuss any difficulties that arose, and how you dealt with them

The data we gathered, as we discussed before, does not line up due to incomplete datasets, with missing values for a variety of reasons

Learning git, git pushes

Talking out our plan of attack

To make sure that we’re not attempting to push and merge branches that would overwrite each other’s work

* Discuss any additional questions that came up, but which you didn't have time to answer: What would you research next, if you had two more weeks?

Another minor motivation that went into this project that we thought we could address was the rise of e-cigarette use

Wanted to see how spending, use, and mortality of e-cigarettes factors into our current findings

It’s probably a question best answered years down the line, due to a variety of factors and especially since e-cigarettes are still an emerging market that is going in various, possibly unpredictable, directions

Some other questions to ask:

Co-morbidity data: seeing how smoking and spending relates to mortality in relation to other conditions such as various related cancers (lung cancer, throat cancer, etc.).

[others]

Again, thinking big, but not being able to find publicly and freely available data for analysis on a global scale

* Questions
* Open-floor Q&A with the audience

Roles, so far:

Nick: data cleaning, initial data analysis, data exploration

Mauricio: data cleaning, data exploration

Emmanuel: data exploration, presentation building and presentation scripting

Bimal: data exploration