****

**Problem Statement**

Welcome to the Fall 2021 Central Regional Datathon! This document explains the topic of the Datathon, important details about the datasets you’ll be using, and guidance on how to submit your results.

**Background**

A recent study released by the “World Health Organization” and the “Institute for Health Metrics and Evaluation” shows that around 8 million people die prematurely every year from smoking. Almost 90% of those deaths are the result of direct tobacco usage, while the remaining 10% results from non-smokers being exposed to second-hand smoke. Tobacco usage is considered by many to be a global epidemic and one of the biggest public health threats the world has ever faced.

Smoking is so common in our society that people normally take it for granted, so it can be quite hard to analyze and measure the impacts it has. According to specialists, there is no safe amount of smoking. Any amount can lead to addiction and cause serious health consequences. Even if a person doesn’t smoke regularly, smoking a cigarette every now and then can put his or her health and future at risk as smoking leads to serious and life-threatening diseases and harms nearly every organ of the body. Such diseases include certain types of cancer (e.g., lung cancer), respiratory diseases such as chronic obstructive pulmonary disease (COPD) and heart disease.

However, the most important part to prevent all these smoking related illnesses is already done. We know that smoking kills; something that wasn’t exactly true in the middle of last century when smoking was not only a looked upon habit, but also encouraged by society, specially by the tobacco industry, through many unrestricted advertisements. In 1964, a report published by the surgeon general of the United States identified smoking as the major cause of the rise in lung cancer. This report is seen as a turning point in the history of smoking as it made clear to the public just how deadly it was.

Once people learned that smoking kills, they could act on it. Fortunately, in most parts of the world, people reacted to that new knowledge and the sales and consumption of tobacco started to decline. On the other hand, in the past decades, alternatives to smoking tobacco started to show up, especially e-cigarettes that claim to be “less unhealthy” than actual cigars. Their manufacturers hold to the fact that there isn’t enough research to conclude whether or not they are unsafe, so people shouldn’t be afraid to use them. A similar approach was taken by the cigarette’s manufacturers in the early last century.

**Your Task**

Your goal is to use tobacco related data in order to discover and analyze patterns associated with tobacco usage. More broadly, you should aim to highlight and discover hidden patterns in the data that could be used to draw any meaningful conclusion on the subject.

We have curated several datasets from a series of different sources. These datasets are related to different topics regarding tobacco and its consumption, since the production of tobacco, sales, usage (and unfortunate deaths from it), until global approaches to help people quit smoking.

We start with a dataset related to “**Tobacco Production”** sourced from [UNdata](http://data.un.org/) that shows the amount of metric tons and monetary value of tobacco produced by country each year. Following that, we present two datasets about “**Tobacco Usage**”. The first one, from [WHO](https://www.who.int/), shows usage at country level split by year and gender. The second one focuses on the United States and presents a few other subdivisions like race, age and education and the topics included are cigarette and e-cigarette use by demographics, cigarette and e-cigarette use frequency, and quit attempts. Be aware that some of these subdivisions (rows) can overlap. For example, “age” contains an “all ages category”. The second dataset was curated by the Center for Disease Control and Prevention (CDC) and it’s available on the [CDC](https://chronicdata.cdc.gov/) website.

Another dataset, sourced from the [ISS](http://www.pnlee.co.uk/ISS.htm), refers to the “**Sales per day**” of tobacco products. As the name implies, this dataset contains, for each country and year, the estimated number of sales of manufactured cigarettes, as well as of hand-rolled cigarettes, per adult (ages 15+) per day. We are also providing a dataset from [GHDX](http://ghdx.healthdata.org/gbd-results-tool) that displays death rates (per 100,000 individuals) due to smoking in many age groups. Similar to the above, we also curated from [IHME](http://www.healthdata.org/) a dataset of Mortality by Chronic Respiratory Disease in the US.

The last dataset relates to measures to help someone to “**Stop Smoking**”. It was curated from many individual datasets found on [GHO](http://apps.who.int/gho/data/node.home) and relates to the EMPOWER package, created by the World Health Organization (WHO) in 2008 to assist in the country-level implementation of six effective measures to reduce demand for tobacco products. This dataset will have information about the price of a pack of cigarettes over the years, the amount of taxes imposed on tobacco sales and the measures that governments are imposing to enforce bans on tobacco ads and how much they help individuals with programs to quit smoking.

All these datasets were curated to give you a starting point so not all of them need to be used. This is a very open-ended problem so feel free to use any other dataset you may find – just follow the guidelines on the “Additional Datasets” below and be aware that the quality of your analysis will also be judged by the reliability of the data being used.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important over breadth of the question posed**.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigour, which are far more important.**

Sample Question 1: According to the World Health Organization (WHO), Brazil and Turkey are the only two countries to fully implement all the MPOWER measures at the highest level of achievement. How do these two countries compare to the others in terms of results? Has MPOWER yielded any positive results? Furthermore, how do these 2 countries compare between themselves? If the difference is too big, can we conclude that MPOWER has not been successful?

Sample Question 2: How does tobacco consumption (or any other measurement) compare between the United States and the rest of the world? If, significantly different, can you find any reason why that would be the case?

Sample Question 3: Can you identify any group or segment (gender, age range, race…) whose tobacco consumption (or any other measurement) behaved differently than the rest? For example, considering overall sales of tobacco are decreasing in the world, is there any group where you can find an opposite trend? If yes, can you identify the reason?

Sample Question 4: A quick look at the data provided shows that Turkey has been increasing its production of tobacco since 2004. Considering that overall consumption of tobacco in the world (and Turkey) is trending down, what can explain that huge increase in production?

**Datasets**

The provided datasets are stored in the “Datathon Materials” folder on Google Drive. Your team should only use the datasets that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and partially pre-cleaned to be readily usable with popular data analysis libraries.

***us\_chronic\_resp\_disease***

Estimates for age-standardized mortality rates by county from chronic respiratory diseases

*﻿~3.4 million rows & 8 columns.* Size: ~350MB. Source: [IHME](http://www.healthdata.org/)

***tobacco\_production***Tobacco production by country and year

*1,207 rows & 5 columns.* Size: <1MB. Source: [UNdata](http://data.un.org/Data.aspx?q=tobacco&d=ICS&f=cmID%3a25090-0)

***tobacco\_use\_us***Tobacco usage in the US

*111,048 rows & 21 columns.* Size: ~30MB. Source: [CDC](https://chronicdata.cdc.gov/)

***tobacco\_use\_ww***Worldwide tobacco usage

*4,023 rows & 7 columns.* Size: <1MB. Source: [WHO](https://www.who.int/)

***sales\_per\_day***Average number of cigarettes sold per day on each year

*2,767 rows & 4 columns.* Size: <1MB. Source: [ISS](http://www.pnlee.co.uk/ISS.htm)

***death\_rates\_smoking\_age***Early deaths due to smoking per 100,000 individuals

*6,468 rows & 9 columns.* Size: <1MB. Source: [GHDX](http://ghdx.healthdata.org/gbd-results-tool)

***stop\_smoking***Indicators that contribute to an individual to stop smoking

*774 rows & 7 columns.* Size: <1MB. Source: [GHO](http://apps.who.int/gho/data/node.home)

**Additional Datasets**

Participants are welcome to scour the web for their own custom datasets to supplement their analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One’s R&D team via Slack if you believe your idea is worthy of an exception).

**Other Materials**

We will provide you the schema for each of the data tables in another packet.

**Submissions: Content**

Submissions should have two components:

1. Report – this should have two main sections:
   1. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
   2. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must “speak for itself”**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

**Submissions: Evaluation**

The competition will have multiple rounds of evaluation. Your Report will be judged as follows:

* **Non-Technical Executive Summary**
  + *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
* **Technical Exposition**
  + *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  + *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  + *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

**Submissions: Format**

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

**However, please also include the source file used to generate your report.** For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

**Submissions MUST be received by 5:00PM EST on Sunday, November 21st, 2021. Any submissions received after that time will NOT be evaluated by the judges**.

**Tips & Recommendations**

This will be a weeklong event, however, you should try to complete as much of your work as possible before the weekend. The extra time may lull you into a false sense of security. Additionally, with your extra time, you should really think about what problem you want to solve. The outcome of this Datathon for you will likely be decided by how well you planned your work.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard “terminal + text editor” environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

We’ve compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

| **Tips for Success** | **Try to Avoid** |
| --- | --- |
| **1.** Focus on hypothesis testing when brainstorming your research question | **1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy |
| **2.** Spend at least 3 hours on your report to ensure strong communication through visualizations and writing | **2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient |
| **3.** Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality | **3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile |

**Ask for Help**

Correlation One’s R&D team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.