**Big Data vs. Little Data: Using the macrocosm of personal iPhone data to speculate about social data collection**

Introduction: A passive versus an active future

A lot of us have heard about third party companies keeping a watch on our user data. A lot of data changes hands without our knowing. And yet, we steer clear, and evade the issue at hand. I say that as best we can, we embrace the data generation processes, and make the best possible use we can of it, without getting too bogged down.

As we all know, secondary data is no longer limited to acquiring data from alternative sources, having a survey about the same topic run by another company, or linking administrative data. While those are all forms of supplementing our primary data at hand, data about the data collection process are a lot more important than they ever were before. I am talking about paradata.

Not just metadata, wherein we label the types of variables in a dataset, and keep a tabulation of the like. Metadata is essentially akin to a data dictionary, a reference document that tells us the units of measurement, scales, and meanings of the variables. In essence, the metadata informs how we analyze the data. And yet, it is not part of the analytical process itself. In other words, we are not analyzing metadata.

No. The secondary data I am referring to here is the paradata. These are going to get ever more complex in the time to come, as neuroscientific processes are combined with UX design and research by commercial companies. At present however, just your ordinary vendors are able to tell you how long you lingered on a question, how fast your eyes glazed over an ad, and how much time you spent responding to a survey question. By gauging your time spent on a question, and equating that with you level of seriousness about the element of a survey, companies and researchers can draw inferences about your response to a stimulus.

Other variables in paradata might include the number of times you logged in to a questionnaire, or automatically logged out. By helping us think about such questions among others, survey researchers have helped us apply traditional psychological methods to modern day swathes of data, and make intelligible inferences. Unlike metadata, therefore, paradata can in fact be considered part of the analytical process, and be used as input variables when trying to determine a customer’s inclination towards a product, or a process.

Enough about trends in modern day survey research, and data collection, and on to today’s example.

In today’s example, we show you how to download and analyze iphone user health data. Many people are aware that they can do this, few actually do, and content themselves with the user friendly dashboards provided by apple on our iphones.

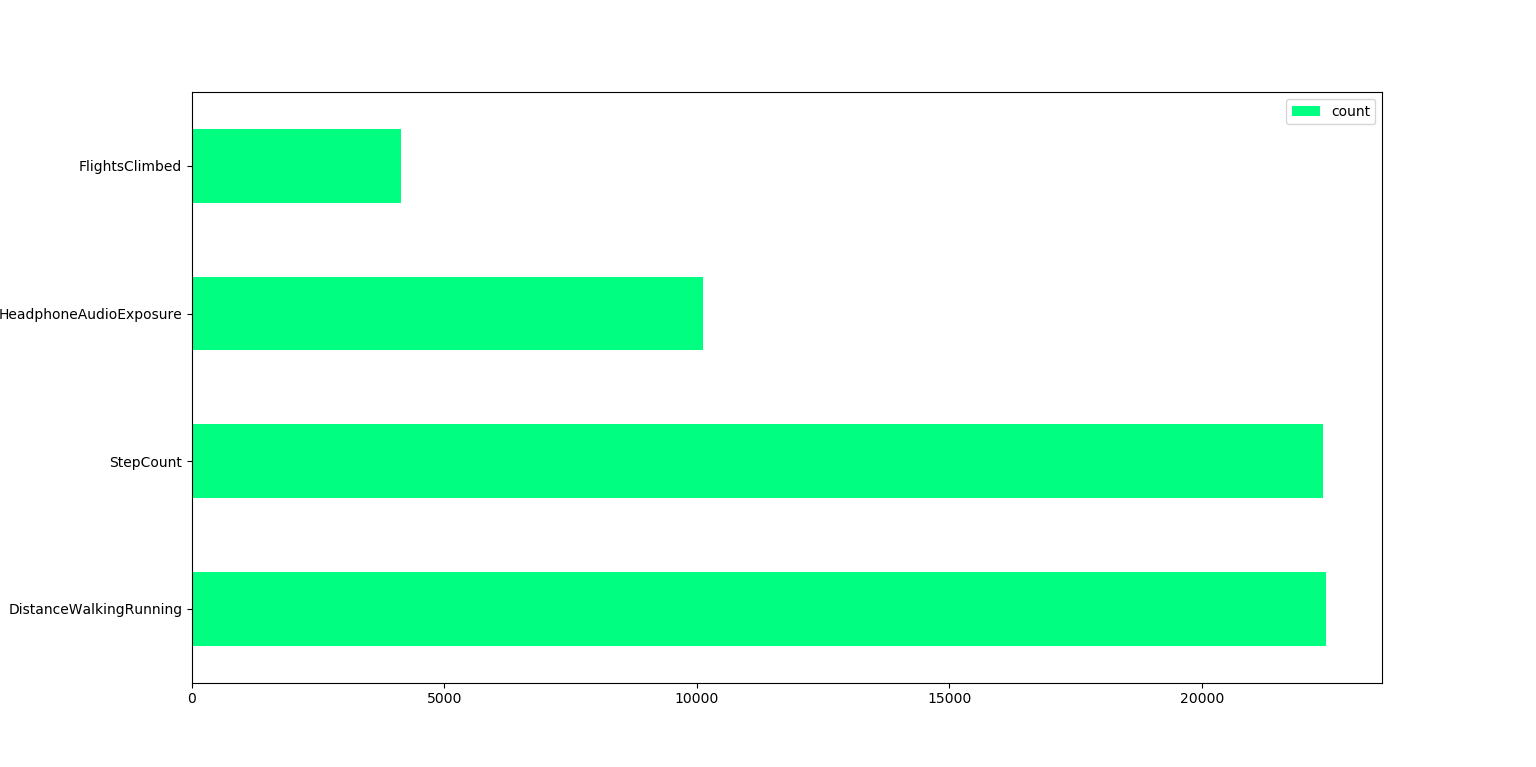
**Our Data:**

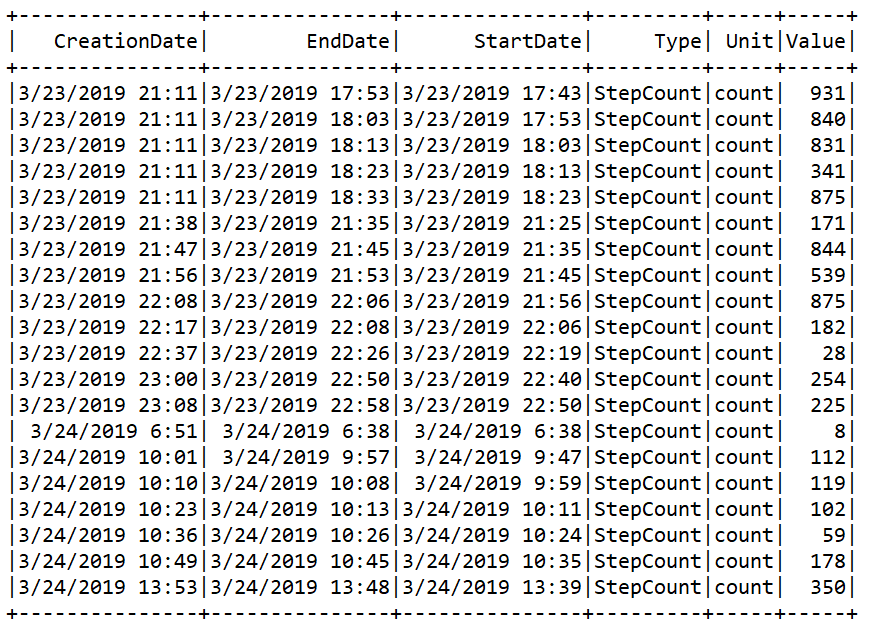
Well, to begin with, if you want to follow along, you can go to your iphone health data app. And navigate to the top right-hand corner of the landing page, or home page. Therein, scroll down, and click on export data. The data are/ is exported in a useless XML format, of no amicable use to your average data analyst / beginner data scientist. You can choose to drop these xml files into an empty excel worksheet and proceed from there, that is what I did!

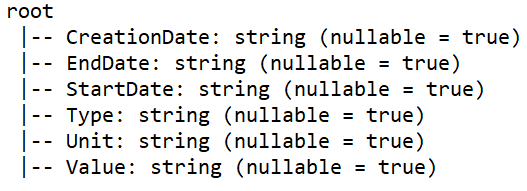
**Goal:**

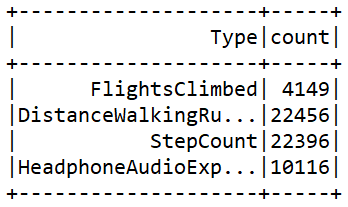
By studying our little data and putting it under spark instances and machine learning algorithms, we can make inferences about the macrocosm. And speculate about how we can combine our societies user data to make sociological inferences, and consequently recommendations. As we know, cell phone data is the mecca of secondary, subconsciously collected data. But few academics take it seriously. Certainly, while privacy is being protected using tech policy, no social policy recommendations are being made based on waivers and data volunteered by the public. Certain not on a large enough scale. We hope for this to be the case moving forward, wherein real time data collected are better informants of social policy and elaborate surveys requiring our full attention.

Figures:









A part of getting used to working with spark is getting used to how it talks to us – in forms of these unwieldly unaesthetic tables. And how it talks to python. Python, in itself, has trouble reformatting variable types and names. At least I find it to be so. Spark, with its added complexity, trades off user friendliness for compute power.

But I find it to be extremely useful in tabulating and plotting preliminary charts of larger swathes of data. Time will tell how the new kids on the block, Spark and Scala fare. But I think off to a good running start.