‘Clean your Room!’

The Philosophy of Dirty Data Prevention

Clean your room! … I mean data!

Why did our mother’s always ask us a million times to pick up our laundry and put it in the basket? Because Mom is a data scientist! She knows how busy a child’s life is and she is constantly trying to save time. She understands that while leaving clothing on the ground might feel like it’s saving time, in the long run, we eventually need to group all the dirty laundry together in the basket. If the clothing is dropped, it has to be picked up = 2 steps. If, from the start, the clothing is put in the basket = 1 step. Efficiency!!

My philosophy is **‘Due your best to ensure the data you are getting from the start is in the format you need to answer the question you are asking.’**

My journey into data science started like most journeys do, with a problem to solve. My problem was our biggest client could catch problems with our application’s before our company could. We had no monitoring system and barely analyzed the data we had.

There were 2 things that needed to be done.

1) Tools needed to be designed around the data we were capable of collecting
2) The data coming into the tools needed to be accurate.

After implementing several pipelines of data for visualization tools, clients became very happy. One day, I decided to revisit the dashboard’s design and that is when I learned, **‘Be careful and spend the time inspecting the data you plan on using before using it in any way.’**

I discovered engineers were ignoring task log protocols and we had several different ways of tagging if someone had canceled their order. From “Cancel” to “Canceled” to “Cancelled” to “Cancels”.

This meant that every employee who was using my tool was reporting potentially drastically inaccurate figures to clients.

I quickly investigated and found the mis wording task logs were application specific tags. Essentially, while one application might have called it ‘Cancel’ and the other ‘Cancelled’, there would be only one or the other tag used in the applications task log. The clients report would have always referenced their specific task accurately, however, the tool was also used to measure overall application performance company wide which <u>was</u> affected by the difference in word choices.

The project to fix every application was considered too massive of an undertaking to fix and while it was determined that because the clients wouldn’t know the difference, the issue would be left.

For the sake of efficiency and accuracy, the best defense against bad data is prevention… so put your dirty laundry in the basket the first time.