# Initial Posts

|  |
| --- |
| **Importance of Data Quality**  The supplemental article in week 2 really hits home for me and I believe it to be a very import consideration to take into account whether your a data scientist, BI developer, data architect, software developer, application analyst, operations leader, systems trainer, etc. Any weak leak in the data procurement chain will ultimately have downstream impacts on the data when it is eventually stored (or updated) in the database. For example:   * Bad end-user workflows = garbage data * Bad application build = garbage data * Poor data governance = garbage data * etc.   When you have garbage data in your database, like the article mentions, you get garbage. This tends to cause more work in the long run. Instead of using your time and effort working on building predictive models, you'll instead spend a great deal of time trying to get stakeholders to trust the data, investigating inconsistent data, and explaining every little data point to leaders in the organization.  Reference:  Garbage In, Garbage Out - The Impact of Data Quality on Data Insights | Databox Blog. (2021, November 29). Retrieved from https://databox.com/impact-of-data-quality-on-data-insights |
| **Recent Publication at Organization**  My organization was recently involved in a research publication that I thought was pretty cool and thought I would share. The aim was to reduce PPE (Personal Protective Equipment) by developing a set of questions that would aid clinicians in determining if a patient was truly symptomatic vs. asymptomatic for COVID-19. After the symptom questions were implemented in our test ordering process, analysis was then ran on data collected from our system to determine if the questions were statistically significant at determining if a patient was symptomatic vs. asymptomatic for COVID-19.    Harnessing the Electronic Health Record and Computerized Provider Order Entry Data for Resource Management During the COVID-19 Pandemic: Development of a Decision Tree - PubMed (nih.gov) |
| **GitHub Question**  Dr. Shankar Parajulee, does our GitHub repo need to be public or does it matter if it is private? |
| **F-String Formatting**  There are 3 main ways to insert expressions into string literals. Any way you do it is up to you but if you haven't found out about f-string formatting I would highly recommend using this method. Below are the 3 methods.  [pic here]    For me f-string formatting is way easier to read and is more intuitive when typing it out. Also, in previous years, str.format() was faster than f-string formatting but that is not the case anymore. "f-strings are faster than both %-formatting and str.format(). As you already saw, f-strings are expressions evaluated at runtime rather than constant values" (realpython.com). In a lot of other ways, f-strings are also a lot more flexible than the other ways of string formatting.  Reference:  Python 3's f-Strings: An Improved String Formatting Syntax (Guide). (2021). Retrieved 8 December 2021, from https://realpython.com/python-f-strings/ |
| **GitHub Process**  For those that are new to GitHub, in my previous course we used this to turn in assignments and I must say I was a bit confused at first. I created a diagram which really helped me a lot so figured I would share it with the rest of this class. Hope it helps! |

# Replies

|  |
| --- |
| *Regarding question about data sets allowed for project.*  Hi James, CDC Wonder is a pretty large data bank and allows you to create many different data sets. My assumption was that you could use pretty much any data set but it shouldn't be the same data set or hypothesis used in the book. This is a good question though because it is somewhat unclear. I'll defer to Dr. Shankar Parajulee. |
| *Regarding question about data sets allowed for project.*  Dr. Shankar Parajulee, for our project is it Ok to use data sets created from CDC Wonder? It looks like the book pulls data from the CDC as well so just wanted to make sure it was Ok. I was planning on pulling data from this website to try and find correlation, if any, regarding Autism Spectrum Disorder (ASD). The variables I would be using would be different than the variables used in the book and also the hypothesis being tested would be different as well. |
| *Regarding missing values for EDA*  Hi Madeleine! Ah yes, I've been there before for sure. This happens all the time where a stakeholder wants to use a particular field or data point for some initiative going on and come to find out, either the field is primarily NULL or is not being used as intended. If you can find some way to backfill the data or some way to generate a zip code from the data that you do have, that would be ideal but what my org typically has to do is a combination of workflow changes, application build changes, and end-user training to start generating the data moving forward. |
| *Regarding Project Topic on ASD*  Hi Yousof, Autonomous Vehicles as a topic sounds pretty interesting. What sort of theory will you be testing?  Yes, these traits and their severity vary from person to person. As the name suggests, Autism Spectrum Disorder (ASD) is a spectrum where doctors diagnose a patient based on their severity. "The DSM-5 [Diagnostic and Statistical Manual of Mental Disorders] incorporated a dimensional assessment approach, allowing clinicians to measure both the presence and the severity of ASD symptoms as: “very severe,” “severe,” “moderate”, or “mild” in the two symptom domains of ‘social communication’ and ‘fixated interests and repetitive behaviors". (Asatonline.org)  In previous years, the diagnosing criteria was rather complex and unlike a spectrum used today, there were many subcategories of the disorder that often overlapped with one another. The most recent version of the diagnosis criteria aimed to simplify the criteria. For example, Asperger's is no longer a sub-categorical diagnosis of the disorder. This does not mean that people who had Asperger's no longer are considered autistic but they are now included on the lower-severity range of the disorder. Over the years, the prevalence of ASD has been on the rise but some could argue that this rise in prevalence is due to the constant readjusting of the diagnosing criteria of ASD in the DSM as well as increased awareness about the disorder.    Hi Jahedur, I've seen first hand some of the challenges that people like you and your family go through and my heart goes out to all families affected by this disorder.      References:  Home. (2021). Retrieved 7 December 2021, from https://asatonline.org/research-treatment/resources/topical-articles/changes-to-the-dsm-autism-diagnostic-criteria/ |
| *Regarding Project Topic on ASD* Amelia, oh yes, I totally agree. My wife and I used to watch a kid that was very low functioning and self-injurious but could pieces together a puzzle faster than anyone I know. I'm talking about like a 500+ puzzle too. |