**The Simple Yet Practical Data Cleaning Codes**

**To solve the common scenarios of messy data**

In one of my articles — [My First Data Scientist Internship](https://towardsdatascience.com/my-first-data-scientist-internship-7f7aa2ee4040), I talked about how crucial data cleaning (data preprocessing, data munging…Whatever it is) is and how it could easily occupy 40%-70% of the whole data science workflow. The world is imperfect, so is data.

**Garbage in, Garbage out**

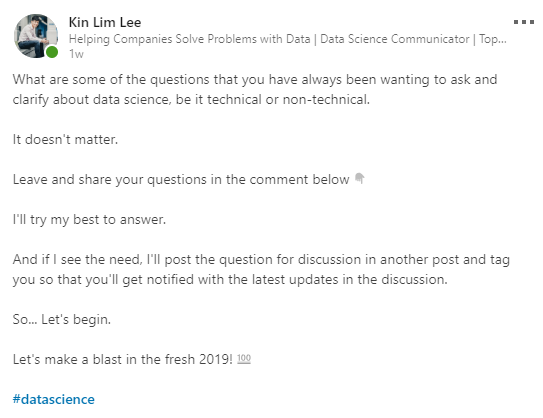
Real world data is dirty, and we as a data scientist — aka data cleaner sometimes — should be able to perform data cleaning prior to any data analysis or models building to ensure the the maximum quality of data.

Long story short, after being in data science field for quite some time, I do feel the pain of doing data cleaning before dealing with data analysis, visualization and models building.

Admit it or not, data cleaning is not an easy task and most of the time it is time-consuming and tedious, yet this process is too important to be neglected.

You’ll understand what I mean if you’ve been through the process. And this is exactly the reason why I wrote this article to help you perform data cleaning in a more smoother way.

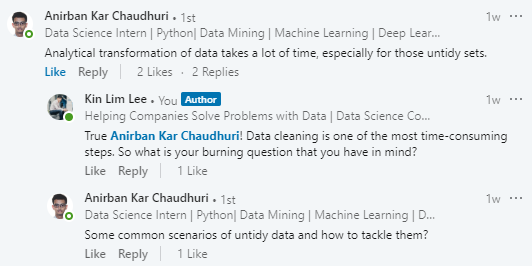
**Why is This Article Important to You?**



[My recent post on LinkedIn](https://www.linkedin.com/feed/update/urn:li:activity:6486373885376851968)

One week ago, I posted on LinkedIn to ask and answer some of the burning questions faced by aspiring data scientists and professionals about data science.

If you’ve been following my work, [I’m on a mission to democratize sharing-learning environment on LinkedIn](https://www.linkedin.com/pulse/2018-year-turning-point-my-life-kin-lim-lee/) with the focus particularly on data science by initiating discussions on LinkedIn among aspiring data scientists, data scientists and other data professionals in different expertise and background. Feel free to follow me on [LinkedIn](https://www.linkedin.com/in/admond1994/) if you’d like to participate in these discussions on interesting topics about data science. You’ll be amazed by how engaging and supportive the data science community is. *😊*



So I got a few interesting questions in the comments. However, there was one particular question posted by [Anirban](https://www.linkedin.com/in/anirban-kar-chaudhuri-7913737b/) that I eventually decided to write an article to answer that question since I’ve been getting the similar questions from time to time.

In fact, not long ago I realized that some data had similar patterns when it came to data cleaning. And this was when I started organizing and compiling some of the data cleaning codes that I thought they would be applicable to other common scenarios — **my little toolbox for data cleaning**.

Since the common scenarios here span across different types of datasets, this article focuses more on showing and explaining what the codes are used for so that you can plug and play easily.

At the end of this article, I hope you’ll find the codes useful and that would make your data cleaning process more faster and efficient.

Let’s get started!

**My Little Toolbox for Data Cleaning**

In the following code snippets, the codes are written in functions for self-explanatory purposes. You can always use the codes directly without putting them into functions with a small change of parameters.

**1. Drop multiple columns**

Sometimes, not all columns are useful in our analysis. Therefore, the df.drop comes in handy to drop the selected columns as specified by you.

**2. Change dtypes**

When a dataset gets larger, we need to convert the dtypes in order to save memory. If you’re interested in learning how to use Pandas to deal with large data, I strongly encourage you to check out this article — [Why and How to Use Pandas with Large Data](https://towardsdatascience.com/why-and-how-to-use-pandas-with-large-data-9594dda2ea4c).

**3. Convert categorical variable to numerical variable**

Some machine learning models require variables to be in numerical format. This is when we need to convert categorical variables to numerical variables before feeding them to the models. In terms of data visualization, I’d suggest to retain the categorical variables to have a more explicit interpretation and understanding.

**4. Check missing data**

If you want to check the number of missing data for each column, this is the fastest way to go with. This gives you a better understanding of which columns have higher number of missing data that determine your next action of data cleaning and analysis.

**5. Remove strings in columns**

There might be some time when you’d face the new line character or other weird symbols that appear in your columns of strings. This could easily be dealt with using df['col\_1'].replace where col\_1 is one of the columns in the dataframe df.

**6. Remove white space in columns**

Anything is possible when data is messy. It is not uncommon to see there are some white spaces at the beginning of the strings. Thus this approach is useful when you want to remove white spaces at the beginning of the strings in a column.

**7. Concatenate two columns with strings (with condition)**

This is helpful when you want to combine two columns with strings conditionally. For instance, you want to concatenate the 1st column with the 2nd column if the strings in the 1st column end with certain letters. The ending letters can also be removed after the concatenation, depending on your needs.

**8. Convert timestamp(from string to datetime format)**

When dealing with time series data, chances are we’ll encounter timestamp column in string format. This means we may have to convert the string format to datetime format — format to be specified based on our requirement — in order to give meaningful analysis and presentation using the data.

**Final Thoughts**

Thank you for reading.

The codes by nature are relatively simple to implement. I hope this little toolbox of data cleaning gave you more confidence to perform data cleaning and more broader perspective of how datasets typically look like based on my experience.

As always, if you have any questions or comments feel free to leave your feedback below or you can always reach me on [LinkedIn](https://www.linkedin.com/in/admond1994/). Till then, see you in the next post! 😄

**About the Author**

[**Admond Lee**](https://www.linkedin.com/in/admond1994/) is currently the Co-Founder/CTO of [**Staq**](https://www.trystaq.com) **—** the #1 business banking API platform for Southeast Asia.

Want to get free weekly data science and startup insights?

[**Join Admond’s email newsletter — Hustle Hub**](https://bit.ly/3pGF8jv), where every week he shares actionable data science career tips, mistakes & learnings from building his startup — Staq.