wrangle_report

September 6, 2022

0.1 Reporting: wrangle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

Briefly I will try to walk through every step and details I encountered during this tough but insightful project. First thing first I downloaded the three datasets and uploaded them on google sheets to have a detailed overview over the rows. I instantly started to see flaws that I automatically classified in my mind to quality and structural inaccuracies. Upon importing the needed packages as well as the datasets, I encountered a problem getting data through Tweepy's API, that I couldn't resolve for more than 10 days, since August, 24th, As mentioned in Udacity's platform, I followed all the steps, step up my developer account, leveraged the complementary files containing the code to scrape data Via the API, and of course I went through Stackoverflow, Github to get a deeper knowledge on how to get the data. Once ready and mastered the code, I went on the platform and started implementing it, to my surprise it didn't work! Once and twice and every time I tried! I went to my developer's account dashboard and changed the API's credentials, put them in the notebook, but again all I get is failed attempts. I was on the brink of a mental breakdown honestly; I contacted two of my classmates on LinkedIn to ask them if my code worked for them and they said yes! It works perfectly fine for us! I was in shock, I wasted so much time on that, maybe the problem is from my IP address or my PC's MAC address. I will gently ask you to verify this issue sir, I had to manually upload the tweet_json.txt, and then gathered data I needed from it. Then I started programmatically assessing them and eventually uncovering more inaccuracies that I judged important to work on to make the exploratory analysis later a lot easier. The first issue is imposed, tweets gathered beyond August 1st, 2017 need to be deducted. I had the choice between using query function or a mask and I opted for the mask. Second and third issues, I altered the timestamp column from str to datetime, next I converted other column's type from int to str since I can only use them later in the str form. For my next issue, while scrolling through expanded_urls column, my eyes landed on a gofundme link combined with the twitter link. As much as I tried to extract it using regex, I couldn't because some rows contained 3 links, I had to use Stackoverflow to help me, there was a thread where the same question as mine was asked, since it was public, I took the liberty to copy the code. 5th issue had me extracting the source of the tweet from a html tag, wasn't a big deal, I figured out that we had 4 sources using count function, I replaced each html with the exact source in a presentable way. Next issue I kept only original tweets with a link to a media, it was easy, used isnull() and a mask. 7th issue replaced none and inexistent dog names with unknown, 8th issue used the latest mask technique taught in our connect sessions including and "&", last issue was to uppercase the columns related to dog predictions. Those

were the quality issues, the tidiness issues revolved around dropping some columns, renaming others, and reduce dog stage into 1 single column to make the visualization easy.