Project Overview

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Image via

Boston Magazine

What Software Do I Need?

The entirety of this project can be completed inside the Udacity classroom on the **Project Workspace: Complete and Submit Project** page using the Jupyter Notebook provided there. (Note: This Project Workspace may not be available in all versions of this project, in which case you should follow the directions below.)

If you want to work outside of the Udacity classroom, the following software requirements apply:

- You need to be able to work in a Jupyter Notebook on your computer. Please revisit our Jupyter Notebook and Anaconda tutorials earlier in the Nanodegree program for installation instructions.
- The following packages (libraries) need to be installed. You can install these packages via conda or pip. Please revisit our Anaconda tutorial earlier in the Nanodegree program for package installation instructions.
    - pandas
    - NumPy
    - requests
    - tweepy
    - json
- You need to be able to create written documents that contain images and you need to be able to export these documents as PDF files. This task can be done in a Jupyter Notebook, but you might prefer to use a word processor like Google Docs, which is free, or Microsoft Word.
- A text editor, like Sublime, which is free, will be useful but is not required.

Project Motivation

Project Motivation

Context

Your goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "*Wow!*"-worthy analyses and visualizations.

The Data

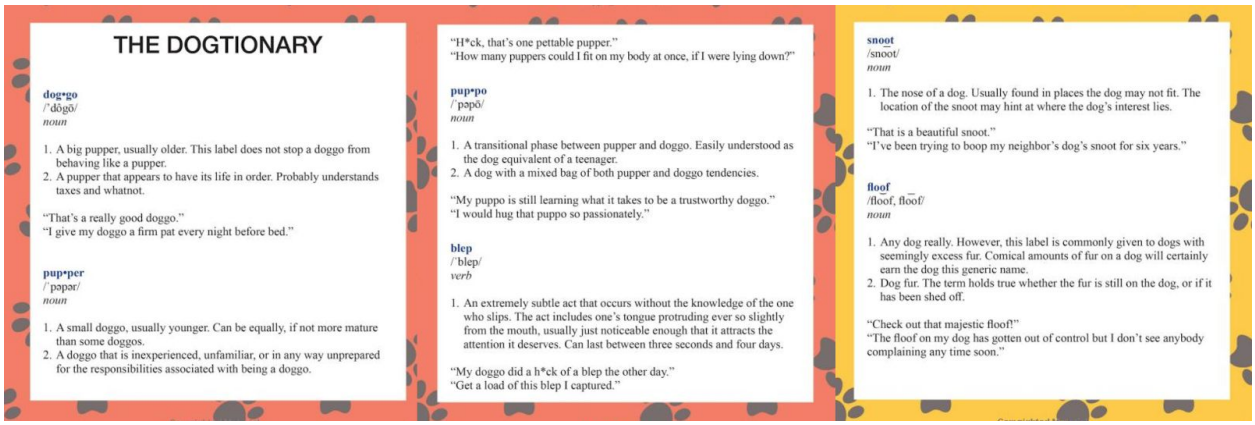
Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

text	rating_ numerator	rating_ denominator	name	doggo	floofer	pupper	puppo
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	13	10	Phineas	None	None	None	None
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10	13	10	Tilly	None	None	None	None
This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co	12	10	Archie	None	None	None	None
This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	13	10	Darla	None	None	None	None
This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such. 12/10 #BarkW	12	10	Franklin	None	None	None	None
Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breathtaking. 13/10 (IG: tucker_marlo) #Bar	13	10	None	None	None	None	None
Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more things by clicking below https://t.co/Zr4hWfAs1H https://t.co/tVJBRMnhxl	13	10	Jax	None	None	None	None
When you watch your owner call another dog a good boy but then they turn back to you and say you're a great boy. 13/10 https://t.co/	13	10	None	None	None	None	None
This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snuggly pettable boatpet. 13/10 #BarkWeek https://t.c	13	10	Zoey	None	None	None	None
This is Cassie. She is a college pup. Studying international doggo communication and stick theory. 14/10 so elegant much sophisticat	14	10	Cassie	doggo	None	None	None
This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13/10 would risk a petting #BarkWeek ht	13	10	Koda	None	None	None	None
This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifyingly good boy https://t.co/u1XPQMl29g	13	10	Bruno	None	None	None	None
Here's a puppo that seems to be on the fence about something haha no but seriously someone help her. 13/10 https://t.co/BxvuXk0UC	13	10	None	None	None	None	puppo
This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist https://t.co/f8dEDcrKSR	12	10	Ted	None	None	None	None
This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppared puppo #BarkWeek https://t.co/y70	13	10	Stuart	None	None	None	puppo

The extracted data from each tweet's text

I extracted this data programmatically, but I didn't do a very good job. The ratings probably aren't all correct. Same goes for the dog names and probably dog stages (see below for more information on these) too. You'll need to assess and clean these columns if you want to use them for analysis and visualization.



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

floof
/floof, flo͞of/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

The Dogtionary explains the various stages of dog: doggo, pupper, puppo, and floof(er) (via the #WeRateDogs book on Amazon)

Additional Data via the Twitter API

Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

Image Predictions File

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
892177421306343426	https://pbs.twimg.cor	1	Chihuahua	0.323581	TRUE	Pekinese	0.0906465	TRUE	papillon	0.0689569	TRUE
891815181378084864	https://pbs.twimg.cor	1	Chihuahua	0.716012	TRUE	malamute	0.078253	TRUE	kelpie	0.0313789	TRUE
891689557279858688	https://pbs.twimg.cor	1	paper_towel	0.170278	FALSE	Labrador_retriever	0.168086	TRUE	spatula	0.0408359	FALSE
891327558926688256	https://pbs.twimg.cor	2	basset	0.555712	TRUE	English_springer	0.22577	TRUE	German_short-haired_pointer	0.175219	TRUE
891087950875897856	https://pbs.twimg.cor	1	Chesapeake_Bay_retriever	0.425595	TRUE	Irish_terrier	0.116317	TRUE	Indian_elephant	0.0769022	FALSE
890971913173991426	https://pbs.twimg.cor	1	Appenzeller	0.341703	TRUE	Border_collie	0.199287	TRUE	ice_lolly	0.193548	FALSE
890729181411237888	https://pbs.twimg.cor	2	Pomeranian	0.566142	TRUE	Eskimo_dog	0.178406	TRUE	Pembroke	0.0765069	TRUE
890609185150312448	https://pbs.twimg.cor	1	Irish_terrier	0.487574	TRUE	Irish_setter	0.193054	TRUE	Chesapeake_Bay_retriever	0.118184	TRUE
890240255349198849	https://pbs.twimg.cor	1	Pembroke	0.511319	TRUE	Cardigan	0.451038	TRUE	Chihuahua	0.0292482	TRUE
890006608113172480	https://pbs.twimg.cor	1	Samoyed	0.957979	TRUE	Pomeranian	0.0138835	TRUE	chow	0.00816748	TRUE
889880896479866881	https://pbs.twimg.cor	1	French_bulldog	0.377417	TRUE	Labrador_retriever	0.151317	TRUE	muzzle	0.0829811	FALSE
889665388333682689	https://pbs.twimg.cor	1	Pembroke	0.966327	TRUE	Cardigan	0.0273557	TRUE	basenji	0.00463323	TRUE
889638837579907072	https://pbs.twimg.cor	1	French_bulldog	0.99165	TRUE	boxer	0.00212864	TRUE	Staffordshire_bullterrier	0.00149818	TRUE
889531135344209921	https://pbs.twimg.cor	1	golden_retriever	0.953442	TRUE	Labrador_retriever	0.0138341	TRUE	redbone	0.00795775	TRUE

Tweet image prediction data

So for the last row in that table:

- tweet_id is the last part of the tweet URL after "*status/*" → https://twitter.com/dog_rates/status/889531135344209921
- p1 is the algorithm's #1 prediction for the image in the tweet → **golden retriever**
- p1_conf is how confident the algorithm is in its #1 prediction → **95%**
- p1_dog is whether or not the #1 prediction is a breed of dog → **TRUE**
- p2 is the algorithm's second most likely prediction → **Labrador retriever**
- p2_conf is how confident the algorithm is in its #2 prediction → **1%**
- p2_dog is whether or not the #2 prediction is a breed of dog → **TRUE**
- etc.

And the #1 prediction for the image in that tweet was spot on:

WeRateDogs™ (author)
@dog_rates

Following

This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppared puppo #BarkWeek

1:02 PM - 24 Jul 2017

A golden retriever named Stuart

So that's all fun and good. But all of this additional data will need to be gathered, assessed, and cleaned. This is where you come in.

Key Points

Key points to keep in mind when data wrangling for this project:

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- You do *not* need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.

**Fun fact: creating this neural network is one of the projects in Udacity's Data Scientist Nanodegree, Machine Learning Engineer Nanodegree and Artificial Intelligence Nanodegree programs.*

Project Details

SEND FEEDBACK

Project Details

Your tasks in this project are as follows:

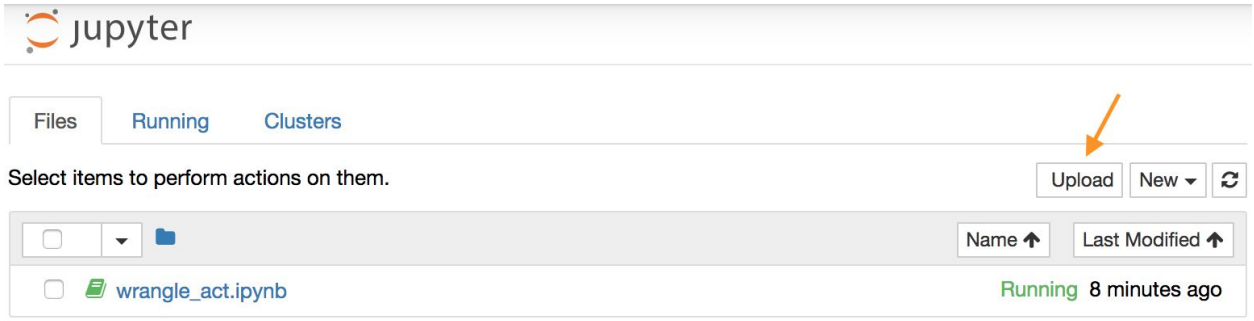
- Data wrangling, which consists of:
    - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Gathering Data for this Project

Gather each of the three pieces of data as described below in a Jupyter Notebook titled `wrangle_act.ipynb`:

1. The WeRateDogs Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. *Note: do not include your Twitter API keys, secrets, and tokens in your project submission.*

If you decide to complete your project in the Project Workspace, note that you can upload files to the Jupyter Notebook Workspace by clicking the "Upload" button in the top righthand corner of the dashboard.



Jupyter Notebook dashboard with arrow pointing to the "Upload" button

Assessing Data for this Project

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least **eight (8) quality issues** and **two (2) tidiness issues** in your `wrangle_act.ipynb` Jupyter Notebook. To meet specifications, the issues that satisfy the Project Motivation (see the *Key Points* header on the previous page) must be assessed.

Cleaning Data for this Project

Clean each of the issues you documented while assessing. Perform this cleaning in `wrangle_act.ipynb` as well. The result should be a high quality and tidy master pandas DataFrame (or DataFrames, if appropriate). Again, the issues that satisfy the Project Motivation must be cleaned.

Storing, Analyzing, and Visualizing Data for this Project

Store the clean DataFrame(s) in a CSV file with the main one named `twitter_archive_master.csv`. If additional files exist because multiple tables are required for tidiness, name these files appropriately. Additionally, you may store the cleaned data in a SQLite database (which is to be submitted as well if you do).

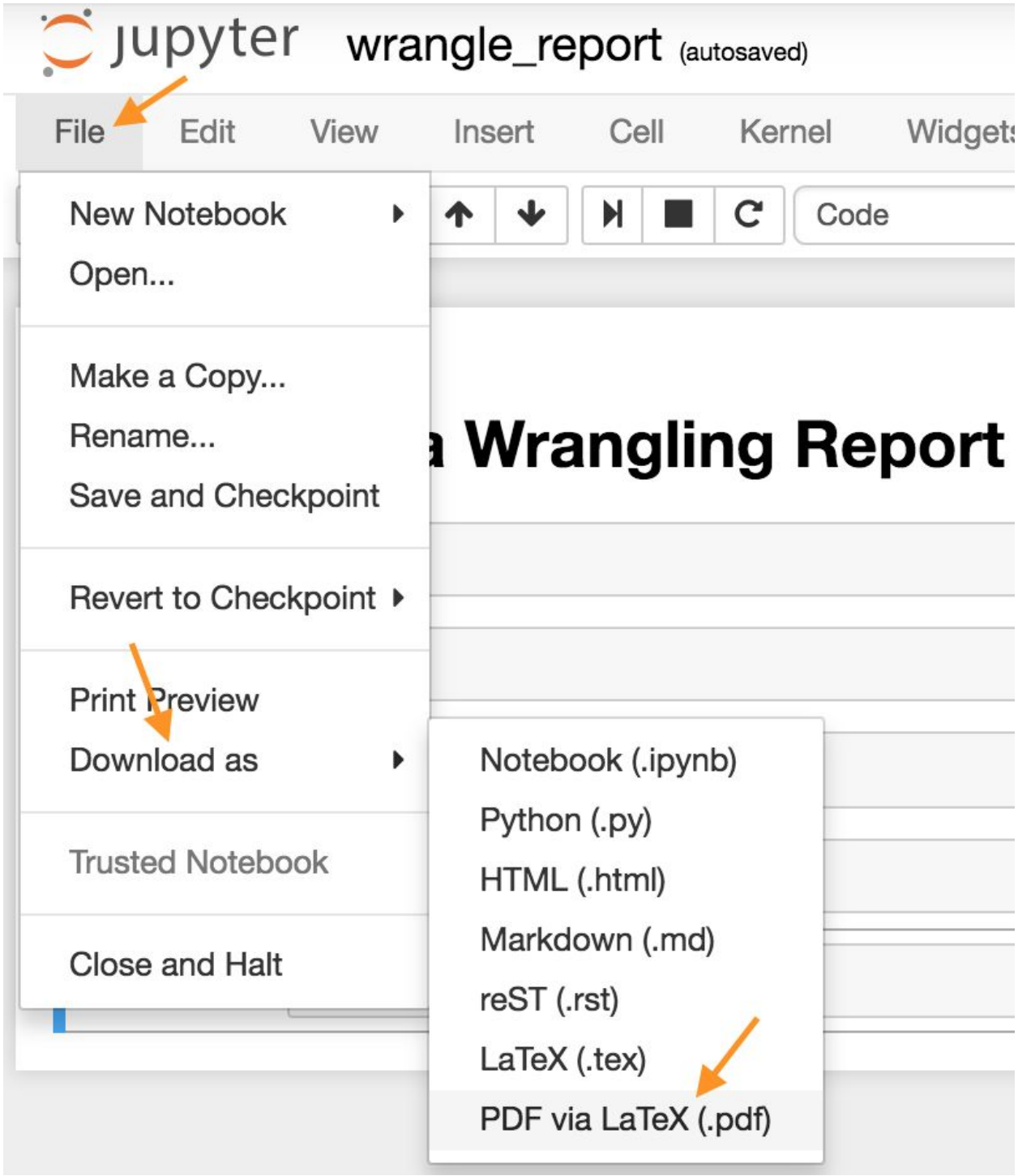
Analyze and visualize your wrangled data in your `wrangle_act.ipynb` Jupyter Notebook. At least **three (3) insights and one (1) visualization** must be produced.

Reporting for this Project

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.

Create a **250-word-minimum written report** called `act_report.pdf` or `act_report.html` that communicates the insights and displays the visualization(s) produced from your wrangled data. This is to be framed as an external document, like a blog post or magazine article, for example.

Both of these documents can be created in separate Jupyter Notebooks using the Markdown functionality of Jupyter Notebooks, then downloading those notebooks as PDF files or HTML files (see image below). You might prefer to use a word processor like Google Docs or Microsoft Word, however.



Download Jupyter Notebook in PDF form

NEXT

Toggle Sidebar

Twitter API

SEND FEEDBACK

How to Query Twitter Data

In this project, you'll be using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. This additional data will include retweet count and favorite count.

Some APIs are completely open, like MediaWiki (accessed via the wptools library) in Lesson 2. Others require authentication. The Twitter API is one that requires users to be authorized to use it. This means that before you can run your API querying code, you need to set up your own Twitter application. Here are the steps to do that on the Twitter site:

- First, if you do not already have one, you need to sign up for a Twitter account.
- Next, to set up a developer account, follow the directions on Twitter's Developer Portal, in the "How to Apply" section.
- You will be guided through the steps, and asked to describe in your own words what you are building. Here is some suggested language you can use: "As a Udacity student, I need to access the Twitter API in order to complete a Data Wrangling student project. In this project, I'll be using Tweepy to query Twitter's API for data included in the WeRateDogs Twitter archive. This data will include retweet count and favorite count. Before I can run my API querying code, I need to set up my own

Twitter application. Once I have this set up, I will develop some code to create an API object that I'll use to gather Twitter data. After querying each tweet ID, I will write its JSON data to a tweet_json.txt file with each tweet's JSON data on its own line. I will then read this file, line by line, to create a pandas DataFrame that I will assess and clean. I may post this completed project on my GitHub account, where it will get viewers. Otherwise there will be no other readers or users of my Twitter data or project analysis beyond the Udacity instructors and reviewers."

- Once you submit your application, you should soon receive an email from Twitter letting you know they have approved your new Twitter developer account. Follow the link in the email from Twitter to a page of directions to get started creating your app.
- If you are asked for an app name, it can be anything appropriate, and if you're asked for a Website URL, it can be anything in a standard URL format. You can do the same with other requested URLs, or perhaps leave them blank.
- If you're asked to explain how your app will be used, you could say something like "I'm creating this for a student Data Wrangling project with Udacity, where we need to query and analyze Twitter data from WeRateDogs."
- You should then be given a Success message, and a new developer page displayed to you where you can manage your app.
- You can then go to the Keys and Tokens tab on this page to find or generate the Consumer API keys, and the Access Token and Access Token Secret that you will need.

Note: If you have any trouble creating this Twitter account or accessing the data, please see the section at the bottom of this page "Accessing Project Data Without a Twitter Account."

Once you have your Twitter account and Twitter app set up, the following code, which is provided in the Getting started portion of the Tweepy documentation, will create an API object that you can use to gather Twitter data.

```
import tweepy

consumer_key = 'YOUR CONSUMER KEY'
consumer_secret = 'YOUR CONSUMER SECRET'
access_token = 'YOUR ACCESS TOKEN'
access_secret = 'YOUR ACCESS SECRET'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
```

Tweet data is stored in JSON format by Twitter. Getting tweet JSON data via tweet ID using Tweepy is described well in this StackOverflow answer. Note that setting the `tweet_mode` parameter to `'extended'` in the `get_status` call, i.e., `api.get_status(tweet_id, tweet_mode='extended')`, can be useful.

Also, note that the tweets corresponding to a few tweet IDs in the archive may have been deleted. Try-except blocks may come in handy here.

Do Not Include Your API Keys, Secrets, and Tokens in Your Submission

Do **not** include your API keys, secrets, and tokens in your project submission. This is standard practice for APIs and public code.

Twitter's Rate Limit

Twitter's API has a rate limit. Rate limiting is used to control the rate of traffic sent or received by a server. As per Twitter's rate limiting info page:

Rate limits are divided into 15 minute intervals

To query all of the tweet IDs in the WeRateDogs Twitter archive, 20-30 minutes of running time can be expected. Printing out each tweet ID after it was queried and using a code timer were both helpful for sanity reasons. Setting the `wait_on_rate_limit` and `wait_on_rate_limit_notify` parameters to `True` in the `tweepy.api` class is useful as well.

Writing and Reading Twitter JSON

After querying each tweet ID, you will write its JSON data to the required `tweet_json.txt` file with each tweet's JSON data on its own line. You will then read this file, line by line, to create a pandas DataFrame that you will soon assess and clean. This Reading and Writing JSON to a File in Python article from Stack Abuse, will be useful.

Accessing Project Data Without a Twitter Account

If you can't set up a Twitter developer account using the steps above, or you prefer not to create a Twitter account for some reason, you may instead follow the directions below to access the data necessary for the project. **Note**: We recommend that you follow the steps above to access the data using a Twitter developer account, because with the shortcut detailed below, you will miss practicing the valuable skill of gathering this data on your own. However, Twitter's updated process may not work for everyone, and we realize there are legitimate reasons that some students may prefer this approach, so we provide it to you here. **You choose the approach to access the data that works best for you. This shortcut approach will certainly work for you to pass the project equally well.**

Directions for accessing the Twitter data without actually creating a Twitter account:

At the bottom of this page you can find two files you can download:

- twitter_api.py: This is the Twitter API code to gather some of the required data for the project. Read the code and comments, understand how the code works, then copy and paste it into your notebook.
- tweet_json.txt: This is the resulting data from twitter_api.py. You can proceed with the following part of "Gathering Data for this Project" on the Project Details page: "Then read this tweet_json.txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count."

Supporting Materials

twitter_api.py

tweet_json.txt

Wrangle and Analyze Data

SUBMIT PROJECT

Wrangle and Analyze Data

Project Submission

ASK A MENTOR

DUE DATE

Mar 14

STATUS

Not submitted

Due at: Sat, Mar 14 9:37 pm

In this project, you'll gather, assess, and clean data then act on it through analysis, visualization and/or modeling.

Before you submit:

1. Ensure you meet specifications for all items in the Project Rubric. Your project "meets specifications" only if it meets specifications for all of the criteria.
2. Ensure you **have not** included your API keys, secrets, and tokens in your project files.

3. *If you completed your project in the **Project Workspace***, ensure the following files are present in your workspace, then click "Submit Project" in the bottom righthand corner of the **Project Workspace** page:
    - `wrangle_act.ipynb`: code for gathering, assessing, cleaning, analyzing, and visualizing data
    - `wrangle_report.pdf` or `wrangle_report.html`: documentation for data wrangling steps: gather, assess, and clean
    - `act_report.pdf` or `act_report.html`: documentation of analysis and insights into final data
    - `twitter_archive_enhanced.csv`: file as given
    - `image_predictions.tsv`: file downloaded programmatically
    - `tweet_json.txt`: file constructed via API
    - `twitter_archive_master.csv`: combined and cleaned data
    - any additional files (e.g. files for additional pieces of gathered data or a database file for your stored clean data)
4. *If you completed your project outside of the Udacity Classroom*, package the above listed files into a zip archive or push them from a GitHub repo, then click the "Submit Project" button on this page.

As stated in *point 4* above, you can submit your files as a zip archive or you can link to a GitHub repository containing your project files. If you go with GitHub, note that your submission will be a snapshot of the linked repository at time of submission. It is recommended that you keep each project in a separate repository to avoid any potential confusion: if a reviewer gets multiple folders representing multiple projects, there might be confusion regarding what project is to be evaluated.

Project Submission Checklist

Before submitting your project, please review and confirm the following items.

I am confident all rubric items have been met and my project will pass as submitted. (If not, I will discuss with my mentor prior to submitting.)

Project builds correctly without errors and runs.

All required functionality exists and my project behaves as expected per the project's specifications.

Once you have checked all these items, you are ready to submit!

It can take us up to a week to grade the project, but in most cases it is much faster. You will get an email once your submission has been reviewed. In the meantime, you should feel free to proceed with your learning journey by continuing on to the next module in the program.

Supporting Materials

Twitter-Archive-Enhanced (2)

Image-Predictions (3)

Tweet-Json

Tweet-Json

PROJECT SPECIFICATION

Wrangle and Analyze Data

Code Functionality and Readability

CRITERIA	MEETS SPECIFICATIONS
The student's code is functional.	All project code is contained in a Jupyter Notebook named wrangle_act.ipynb and runs without errors.
The student's code is readable, i.e., uses good coding practices.	The Jupyter Notebook has an intuitive, easy-to-follow logical structure. The code uses comments effectively and is interspersed with Jupyter Notebook Markdown cells. The steps of the data wrangling process (i.e. gather, assess, and clean) are clearly identified with comments or Markdown cells as well.

Gathering Data

CRITERIA	MEETS SPECIFICATIONS

The student is able to gather data from a variety of sources and file formats.	Data is successfully gathered: From at least the three (3) different sources on the Project Details page. In at least the three (3) different file formats on the Project Details page. Each piece of data is imported into a separate pandas DataFrame at first.

Assessing Data

CRITERIA	MEETS SPECIFICATIONS
The student is able to assess data visually and programmatically for quality and tidiness.	Two types of assessment are used: Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g. Excel, text editor). Programmatic assessment: pandas' functions and/or methods are used to assess the data.

The student is able to thoroughly assess a dataset.	At least eight (8) data quality issues and two (2) tidiness issues are detected, and include the issues to clean to satisfy the Project Motivation. Each issue is documented in one to a few sentences each.

Cleaning Data

CRITERIA	MEETS SPECIFICATIONS
The student uses the steps in the data cleaning process to guide their cleaning efforts.	The define, code, and test steps of the cleaning process are clearly documented.
The student is able to thoroughly clean a dataset programmatically.	Copies of the original pieces of data are made prior to cleaning. All issues identified in the assess phase are successfully cleaned (if possible) using Python and pandas, and include the cleaning tasks required to satisfy the Project Motivation.

	A tidy master dataset (or datasets, if appropriate) with all pieces of gathered data is created.

Storing and Acting on Wrangled Data

CRITERIA	MEETS SPECIFICATIONS
The student is able to store a gathered, assessed, and cleaned dataset.	Students will save their gathered, assessed, and cleaned master dataset(s to a CSV file or a SQLite database.
The student is able to act on their wrangled data to produce insights (e.g. analyses, visualizations, and/or models).	The master dataset is analyzed using pandas or SQL in the Jupyter Notebook and at least three (3) separate insights are produced. At least one (1) labeled visualization is produced in the Jupyter Notebook using Python's plotting libraries or in Tableau.

	Students must make it clear in their wrangling work that they assessed and cleaned (if necessary) the data upon which the analyses and visualizations are based.

Report

CRITERIA	MEETS SPECIFICATIONS
The student is able to reflect upon and describe their data wrangling efforts.	The student's wrangling efforts are briefly described. This document (wrangle_report.pdf or wrangle_report.html) is concise and approximately 300-600 words in length.
The student is able to describe some insights found in their wrangled dataset.	The three (3) or more insights the student found are communicated. At least one (1) visualization is included. This document (act_report.pdf or act_report.html) is at least 250 words in length.

Project Files

CRITERIA	MEETS SPECIFICATIONS
Are all required files included in the student's submission?	The following files (with identical filenames) are included: ● wrangle_act.ipynb ● wrangle_report.pdf or wrangle_report.html ● act_report.pdf or act_report.html All dataset files are included, including the stored master dataset(s), with filenames and extensions as specified on the Project Submission page.

Suggestions to Make Your Project Stand Out!

Assess and clean more than the required issues (eight and two, respectively, for quality and tidiness).

In act_report.pdf (or act_report.html), additional images beyond any visualizations are encouraged to make this report more engaging. Make this report as detailed as you'd like!

Gather additional data beyond the required pieces.

Create a model that can make predictions based on your wrangled data.