DATA ANALYST NANODEGREE

PROJECT 5 : Wrangle & Analyze Data

Fatimah Ahmed Al-Matar

Riyadh, Saudi Arabia

Contents

Abstract 3

Introduction 4

Gathering Data 5

Assessing Data 7

Cleaning Data 8

Storing Data 8

Analyzing & Visualizing Data 8

List of Figures

Figure 1: Twitter Archive CSV File 5

Figure 2: Tweet Image Prediction 5

Figure 3: Twitter Data using Twitter's API 6

Abstract

Data are generated from everywhere and in various formats. To understand what that data is all about and to have some clear insights, data should be gathered, cleaned and analyzed. Here comes the role of the Data Analyst; where he/she will gather the needed data from different and multiple resources, clean it if necessary and then apply the required analyses and draw some conclusions.

In this project, I am applying one of Data Analysis approaches known as “Data Wrangling” on data from Twitter about dogs, I will start this process by first gathering the required data, assessing and finding problems in the structure and quality of that data, solve and clean those problems, store that data and finally draw some conclusions and insights about this dog data.

Introduction

"WeRateDogs", a Twitter account created in November 2015. The purpose of this account is to let people share photos about their dogs and let other people rate them. Usually, the rating is out of 10, but most of the time people rate the dogs with more than 10/10. However, in this project, I will apply the Data Wrangling three steps: Gathering, Assessing and Cleaning on this dog data and try to figure out some interesting outcomes from these lovely dogs.

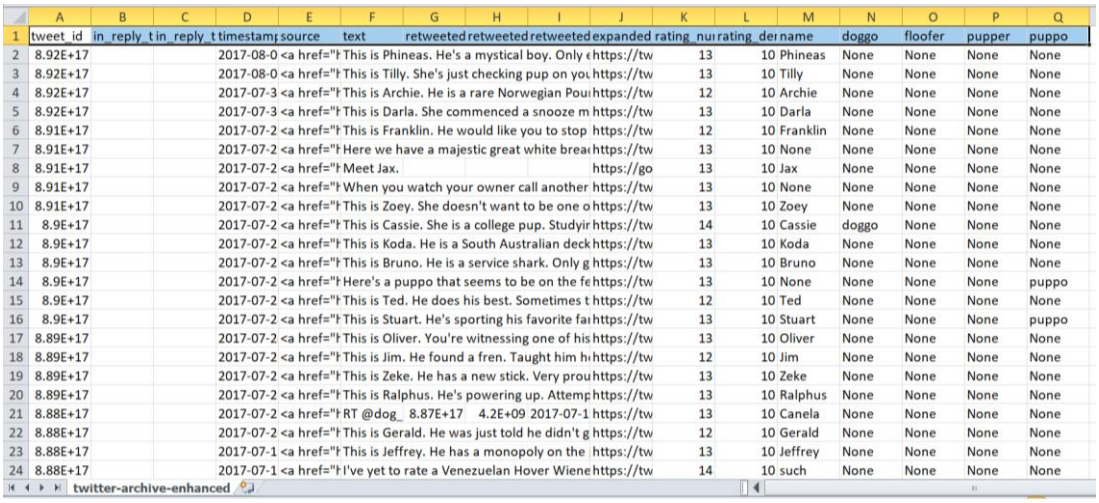
I have used Jupyter Notebook and Python codes to make all of this happen. Some necessary Python libraries have been used like; Pandas, Numpy, Matplotlib, etc. Most importantly, Tweepy library has been installed as well so we can use the Twitter API to fetch in some more data directly from Twitter.

Gathering Data

The very first step in Data Wrangling is to gather data. Many criteria can be used; we can bring the data in via Web Scraping, requesting APIs, having ready-made CSV or TSV files, and many more others.

In this project, I have gathered the dogs data from 3 resources:

1- The “WeRateDogs” Twitter archive: this file was provided by Udacity and I only downloaded it and used it in my notebook. The file contains data from the Twitter account “WeRateDogs” and has different attributes about them.



tweet_id	in_reply_t	in_reply_t	timestamp	source	text	retweeted	retweeted	retweeted	expanded	rating_nu	rating_de	name	doggo	floofer	pupper	puppo
8.92E+17			2017-08-0	<a href="h	This is Phineas. He's a mystical boy. Only c				https://tw	13	10	Phineas	None	None	None	None
8.92E+17			2017-08-0	<a href="h	This is Tilly. She's just checking pup on you				https://tw	13	10	Tilly	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Archie. He is a rare Norwegian Pou				https://tw	12	10	Archie	None	None	None	None
8.92E+17			2017-07-3	<a href="h	This is Darla. She commenced a snooze m				https://tw	13	10	Darla	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Franklin. He would like you to stop				https://tw	12	10	Franklin	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Here we have a majestic great white brea				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	Meet Jax.				https://go	13	10	Jax	None	None	None	None
8.91E+17			2017-07-2	<a href="h	When you watch your owner call another				https://tw	13	10	None	None	None	None	None
8.91E+17			2017-07-2	<a href="h	This is Zoey. She doesn't want to be one o				https://tw	13	10	Zoey	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Cassie. She is a college pup. Studyir				https://tw	14	10	Cassie	doggo	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Koda. He is a South Australian deck				https://tw	13	10	Koda	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Bruno. He is a service shark. Only g				https://tw	13	10	Bruno	None	None	None	None
8.9E+17			2017-07-2	<a href="h	Here's a puppo that seems to be on the fe				https://tw	13	10	None	None	None	None	puppo
8.9E+17			2017-07-2	<a href="h	This is Ted. He does his best. Sometimes t				https://tw	12	10	Ted	None	None	None	None
8.9E+17			2017-07-2	<a href="h	This is Stuart. He's sporting his favorite fa				https://tw	13	10	Stuart	None	None	None	puppo
8.89E+17			2017-07-2	<a href="h	This is Oliver. You're witnessing one of his				https://tw	13	10	Oliver	None	None	None	None
8.89E+17			2017-07-2	<a href="h	This is Jim. He found a fren. Taught him h				https://tw	12	10	Jim	None	None	None	None
8.89E+17			2017-07-2	<a href="h	This is Zeke. He has a new stick. Very prou				https://tw	13	10	Zeke	None	None	None	None
8.89E+17			2017-07-2	<a href="h	This is Ralphus. He's powering up. Attemp				https://tw	13	10	Ralphus	None	None	None	None
8.88E+17			2017-07-2	<a href="h	RT @dog_	8.87E+17	4.2E+09	2017-07-1	https://tw	13	10	Canela	None	None	None	None
8.88E+17			2017-07-2	<a href="h	This is Gerald. He was just told he didn't g				https://tw	12	10	Gerald	None	None	None	None
8.88E+17			2017-07-1	<a href="h	This is Jeffrey. He has a monopoly on the				https://tw	13	10	Jeffrey	None	None	None	None
8.88E+17			2017-07-1	<a href="h	I've yet to rate a Venezuelan Hover Wiene				https://tw	14	10	such	None	None	None	None

twitter-archive-enhanced

Figure 1: Twitter Archive CSV File

2- The tweet image predictions

This data was downloaded programmatically using Requests library, it contains images of the dogs and has related attributes related to them using Neural Networks.

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

Figure 2: Tweet Image Prediction

3- The final data used was fetched in directly from Twitter using Twitter's API that was requested with Python's Tweepy library. I have only taken the "Tweet ID","Retweet Counts" and "Favorite Counts" from each tweet and then saved them in a "Tweet_json.txt" file.

Out[12]:

	id	retweet_count	favorite_count
0	892420643555336193	7173	34478
1	892177421306343426	5391	29909
2	891815181378084864	3552	22507
3	891689557279858688	7396	37722
4	891327558926688256	7932	36020

In [13]:
```
tweet_info = tweet_info.rename(columns = {'id':'tweet_id'})
tweet_info.head(1)
```

Figure 3: Twitter Data using Twitter's API

Assessing Data

Data was assessed visually and programmatically.

- **Visual Assessment:**

First of all, data cannot be assessed 100% accurately from only a look, so further programmatic assessment is a must. However, with a quick inspection over the data, I have found that the “Twitter Archive” file has so many NaN values, also the source attribute has only html format which is much harder to do further analyses.

- **Programmatic Assessment:**

Many Python functions can be used to assess the data like: info(), describe(), value_counts(), etc.
The problems that I have come out with are as follow:

Tidiness Problems:

1) Dog Stage is seperated into 4 columns

2) Data in all 3 dataframe are related but they are seperated

Quality Problems:

1) Invalid "tweet_id" data type (int instead of string or object)

2) Invalid "timestamp" data type (object instead of datetime)

3) Invalid "name" for the some dogs (a, an, None, etc..)

4) The "Source" column is in an HTML format.

5) Delete the columns that will not be necessary for further analysis.

6) Some columns has attributes seperated by underscore instead of a space

7) Drop 66 jpg_url duplicated

8) Some p values come with upercase letter and other start with lowercase

Cleaning Data

In this stage, I have applied the three stages of cleaning; Define, Code and Test. I have also used different Python functions like: copy(), merge(), drop(), np.isnan(), etc.

Storing Data

After completing all the Data Wrangling steps, I have saved the cleaned data in a new CSV file and called it “twitter_archive_master”.

Analyzing & Visualizing Data

In this part of the project, I have tried to find some trends and insights over the dogs data. “this will be provided in detail in the “act_report.pdf” file.