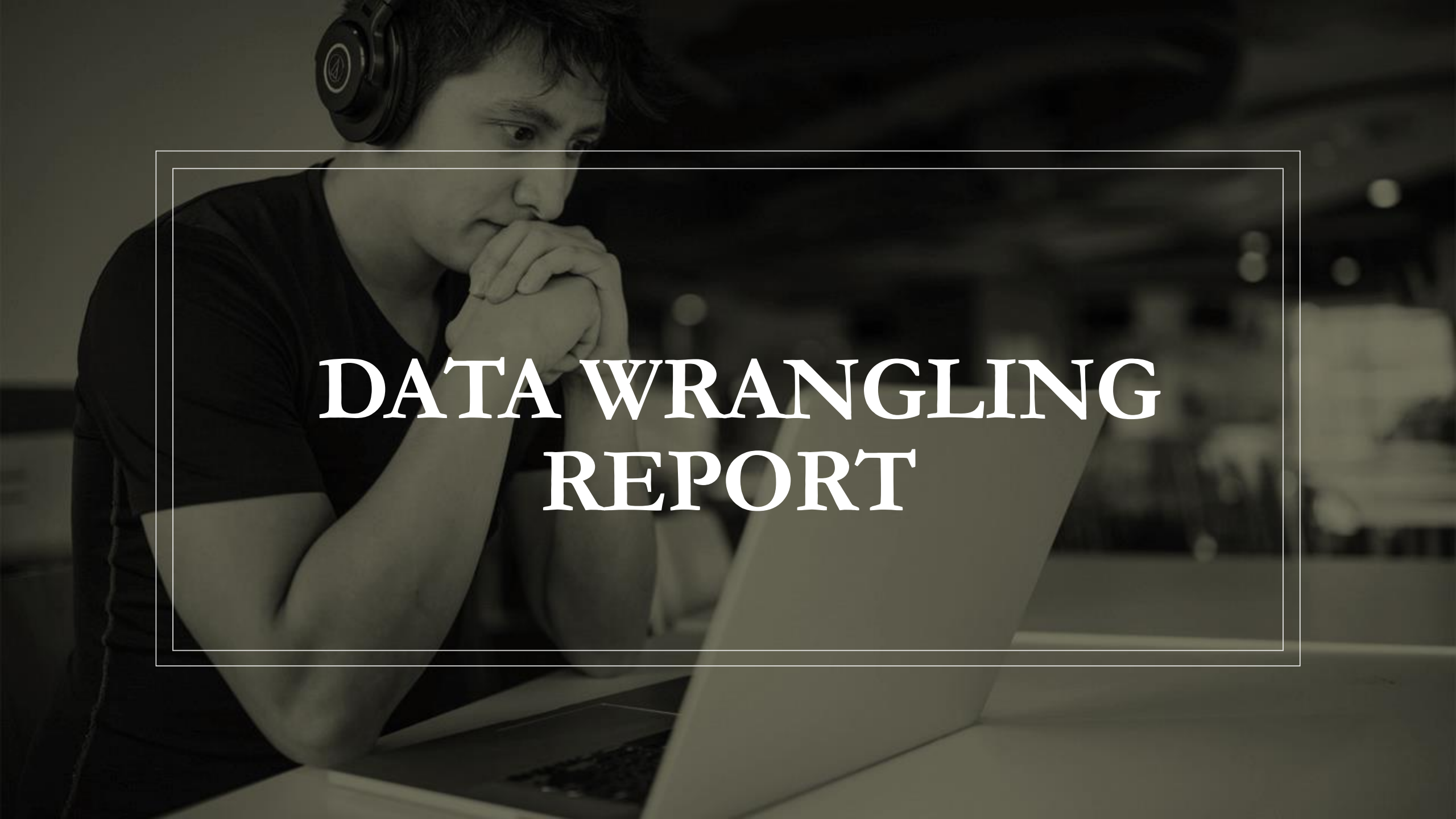


DATA WRANGLING REPORT

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data is not clean. And It is a first 3 process of main steps:

Gathering.	Assessing.	Cleaning.	Storing	Analysis & Visualization
It's about collecting data	After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness.	It is the process of fixing and resolving issues identified in the Cleaning process.	Final process before analysis and visualizations	Here we analys and show our concolutions by plots and grahs

Data Gathering

From WeRateDogs Twitter archive given by Udacity in csv format:

Using panda's method 'read_csv', I managed to read the data stored in the file 'twitter-archive-enhanced.csv'. I stored it in a DataFrame called 'twitter_archive'. The data has many issues that will be cleaned and resolved later.

2. Image prediction file downloaded programmatically using Requests library and the URL provided by Udacity in tsv format:

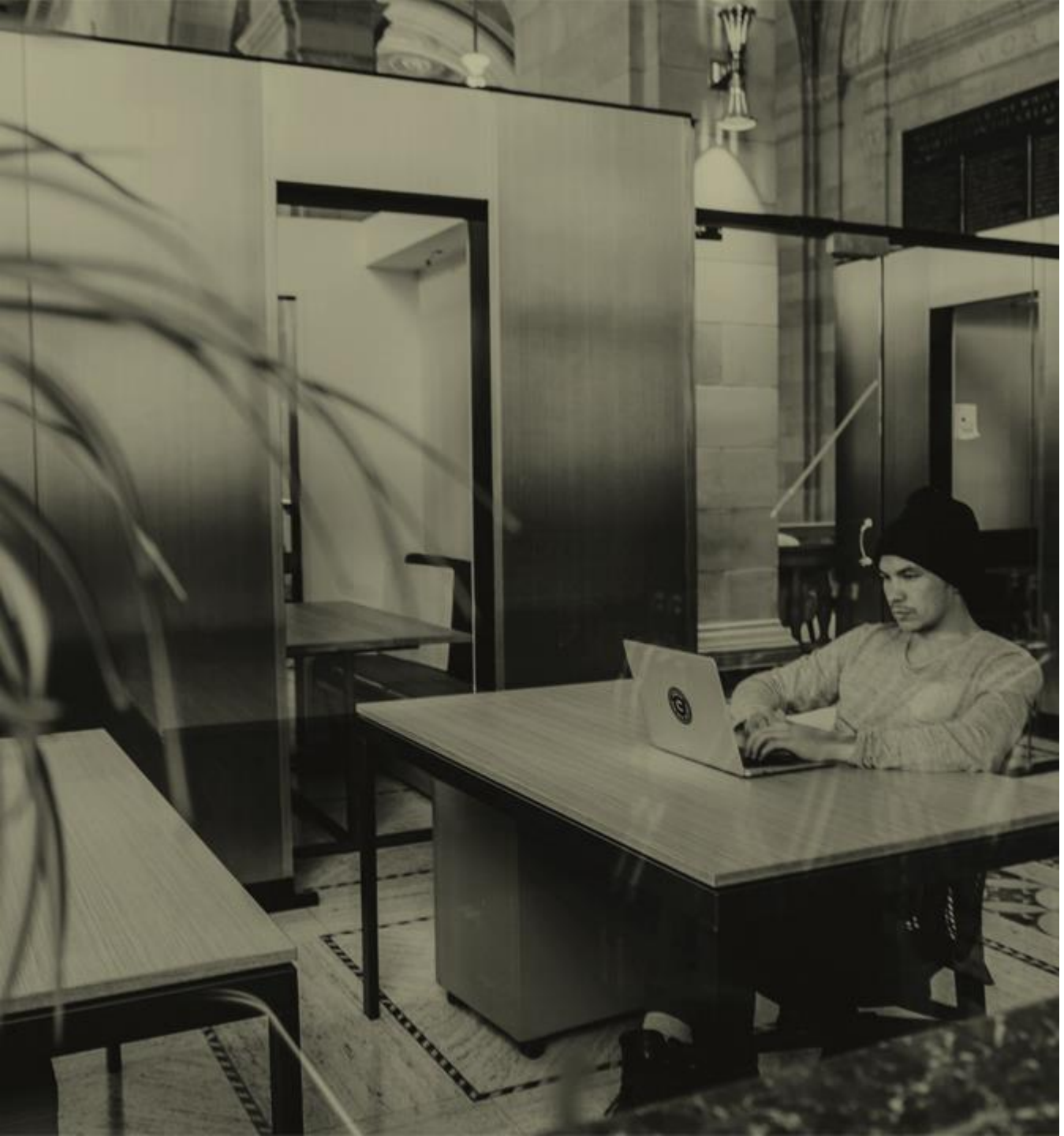
Using Requests library and 'get' method, data was downloaded in a file 'image_predictions.tsv'. Then, the content was stored in a DataFrame called 'image_predictions' using pandas' method 'read_csv'.

3. Data retrieved by querying Twitter's APIs and using Tweepy library.

Using the list of tweet_id's in dataframd 'twitter_archive', I made a loop through each tweet and query Twitter's APIs with the tweet ID to get each tweet's JSON data. Then, I retrieved the required data ('favorite_count', 'retweet_count', 'followers_count', 'favourites_count', 'created_at') and store it in a list called 'df_list'. There were some errors, and the tweet_id of each error was stored in list called 'error_list'. Finally, I created a DataFrame called 'tweet_data' using the list.

Assessing

- After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness. Data were assessed programmatically and visually.



Quality:

Issues with content. Low quality data is also known as dirty data. Identified quality issues are:

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float.
- retweeted_status_timestamp, timestamp should be datetime.
- The numerator and denominator columns have invalid values.
- In several columns null objects are non-null (None to NaN).
- There are invalid names (a, an and less than 3 characters).- We only want original ratings tweets, not retweets.
- We might change the type of columns: (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) to string since we aren't doing any actions on them.
- Sources are not readable.
- Missing values from images dataset (2075 rows instead of 2356)
- Some tweet_ids have the same jpg_url
- Some tweets have 2 different tweet_id, that are retweets.

Tidiness

- Issues with structure that prevent easy analysis. Untidy data is also known as messy data.
- Identified tidiness issues are:
- • Dog stage is in 4 columns (doggo, floofer, pupper, puppo), no need for that.
- • Merge 'tweet_info' and 'image_predictions' into 'twitter_archive'.

Cleaning

1. Create one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp

2. Delete retweets

3. Remove columns no longer needed

4. Change tweet_id from an integer to a string

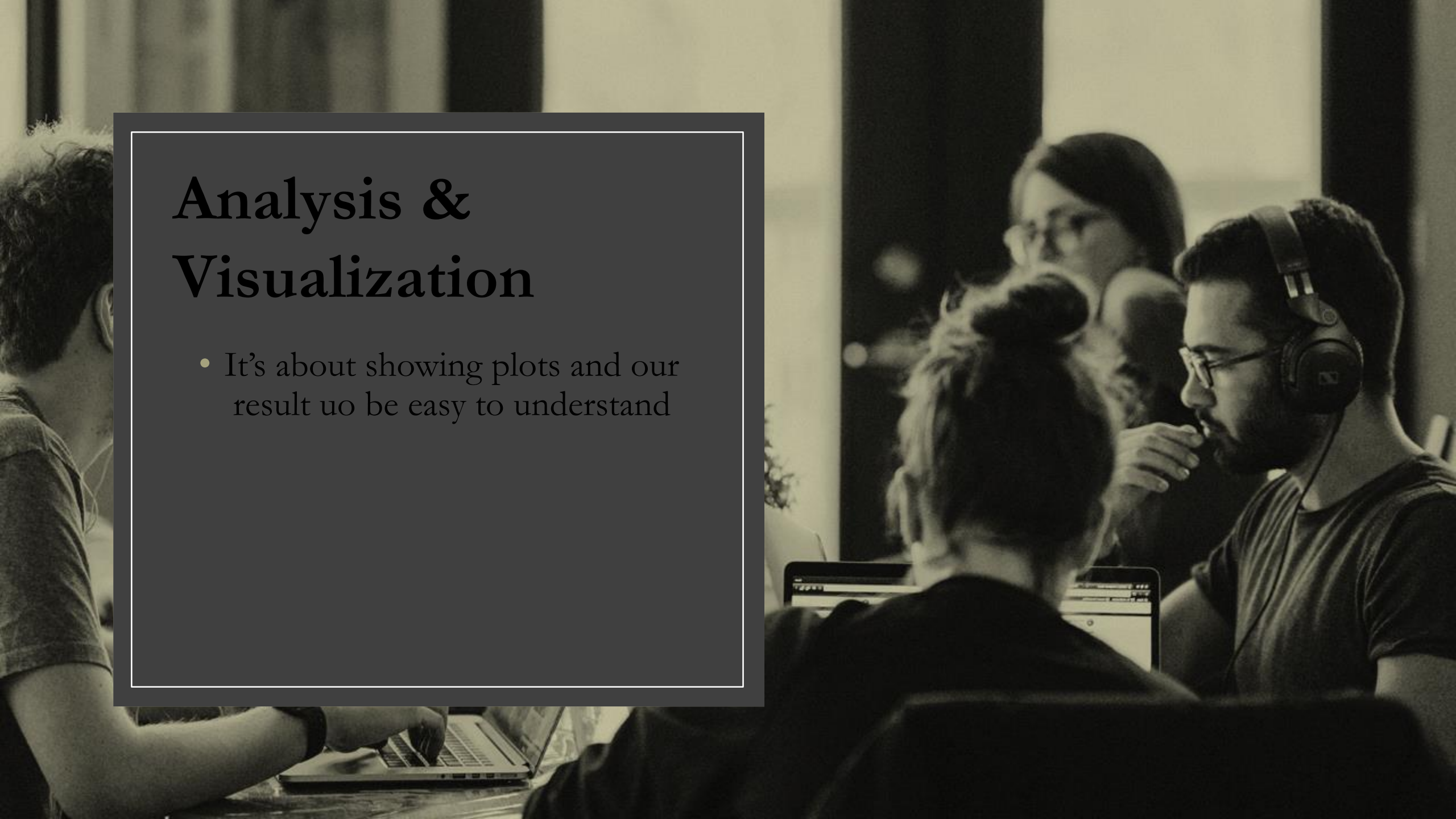
5. Change the timestamp to correct datetime format

6. Correct naming issues

7. Standardize dog ratings

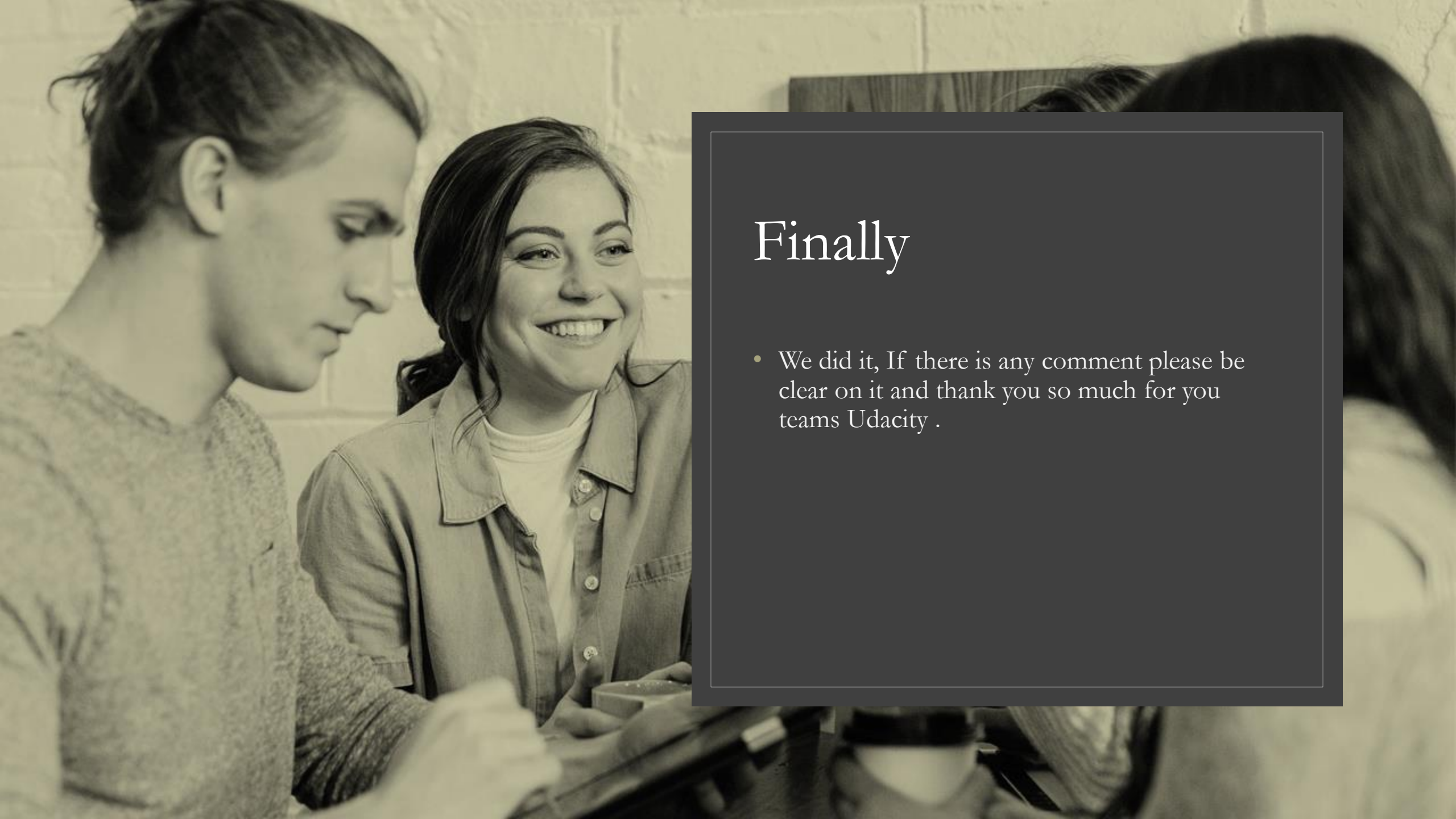
8. Creating a new dog_breed column using the image prediction data

9. Merge the clean versions of archive, images, and twitter_counts_df dataframes Correct the dog types



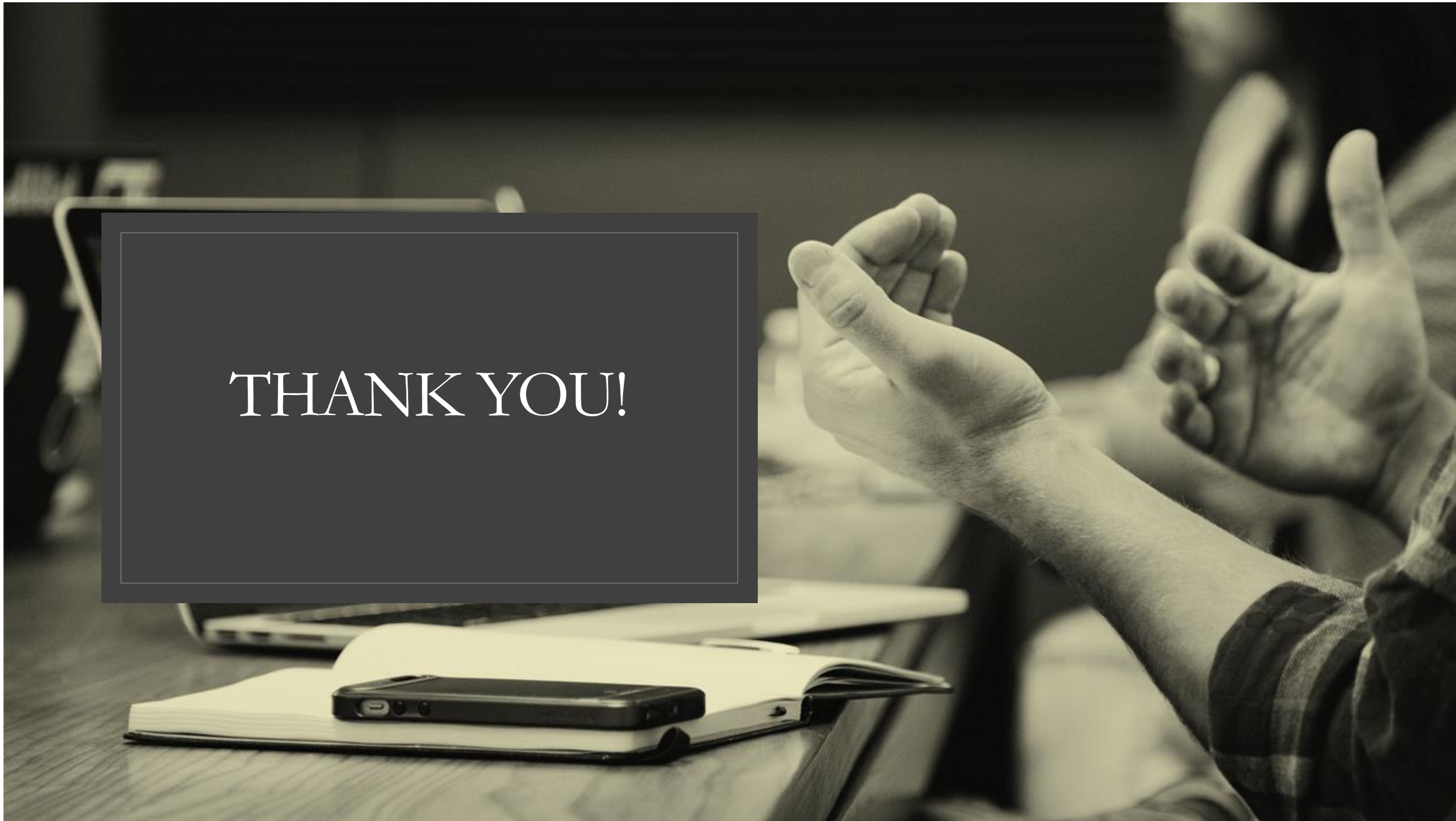
Analysis & Visualization

- It’s about showing plots and our result uo be easy to understand



Finally

- We did it, If there is any comment please be clear on it and thank you so much for you teams Udacity .


THANK YOU!