Capstone 2

Wrangling, Cleaning, Visualization, EDA, and Statistical Analysis
Of the Street Fashion Data Set

The problem

Street Fashion Data

Street fashion consists of entire outfits that have been put together by real people

Clothing pieces are not isolated

Entire outfits can offer extra data about fashion

Uses

Classification of styles and clothing elements in street fashion images could have wide ranging uses in the fashion industry

Image tagging could help for recommendation engines and provide style insights

Problem Statement

This project will use self-tagged and styled images from the website Chictopia to create classification models to tell the types and styles of clothing found in each image

Data Cleaning

Examine Variables

The street fashion data set contains almost 3000,000 images with metadata

Metadata includes username, location, upload time, photo name, tags, and styles

Clean NaNs

Entries with NaNs in the username, location, time, or photo name were removed

This left approximately 284,000 entries

Time data was converted into datetime objects

Clean Tags and Styles

Tags and styles were separated into discrete columns for each word

This left most entries with three styles

Brand names were removed from tags, leaving only colors and articles of clothing

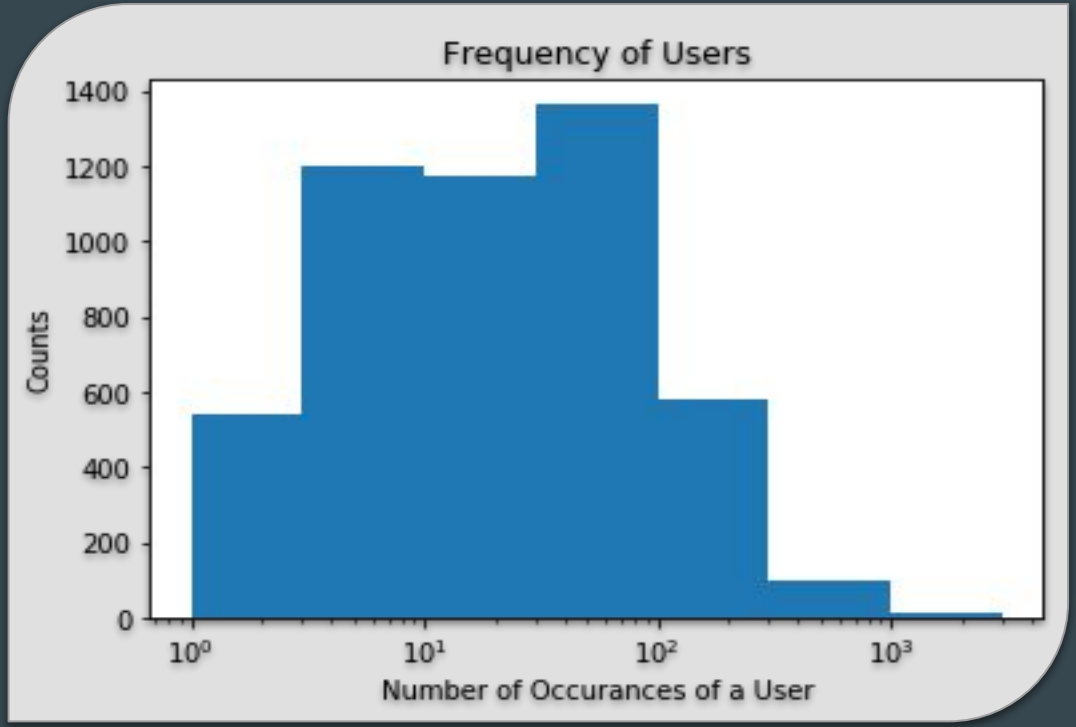
Variables

Users

Most users uploaded 20 or fewer images

Over 100 users uploaded over 1000 images, potentially fashion photographers

The most images uploaded by one user was 8672



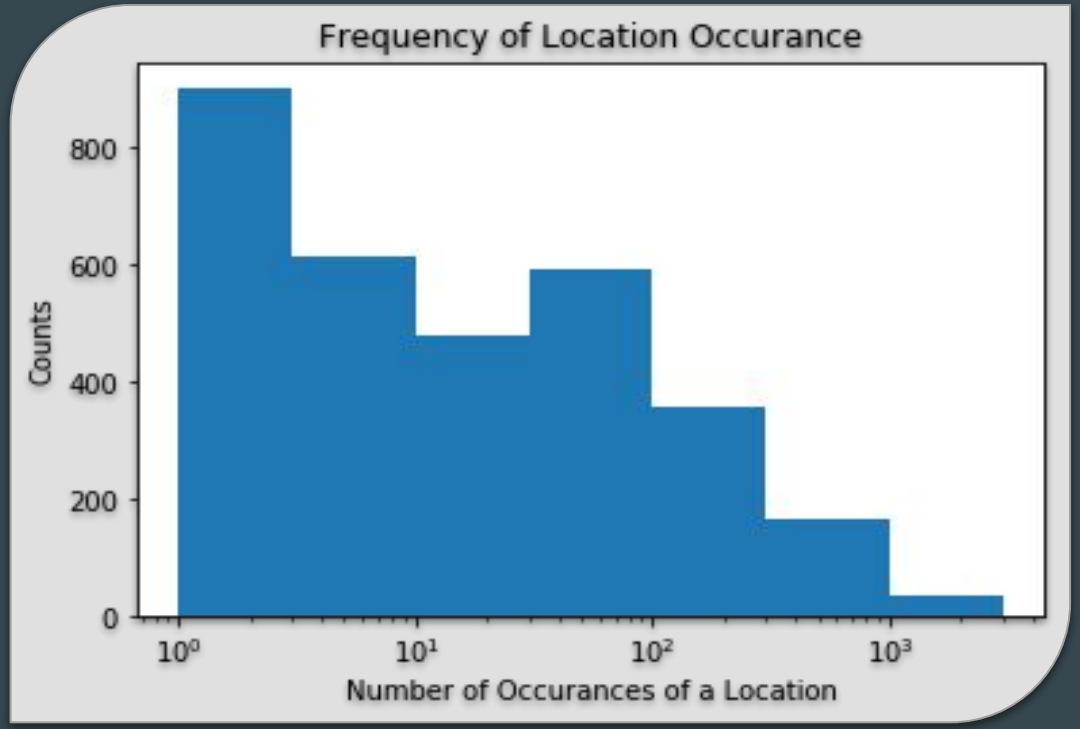
Variables

Location

3150 unique locations occur in the data

Most locations occur less than 12 times

The location with the most uploads is Los Angeles, CA



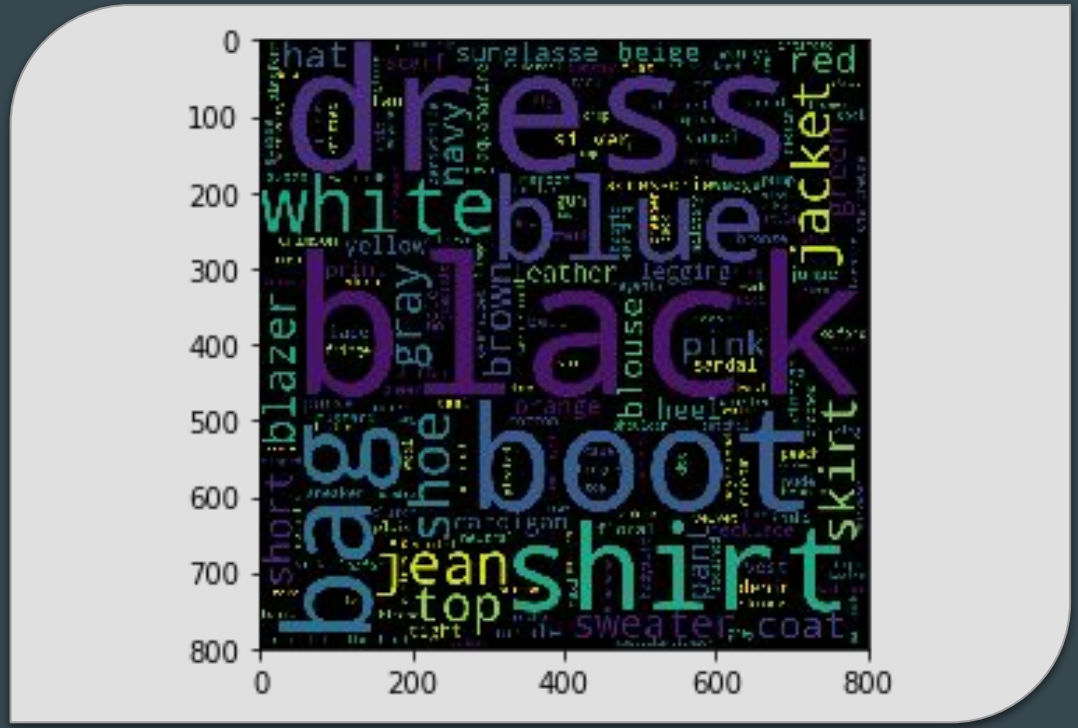
Variables

Tags

Most tags occur in the data only one time

In total, 258 tags occur more than 100 times in the data

These high frequency tags will be used for the classification



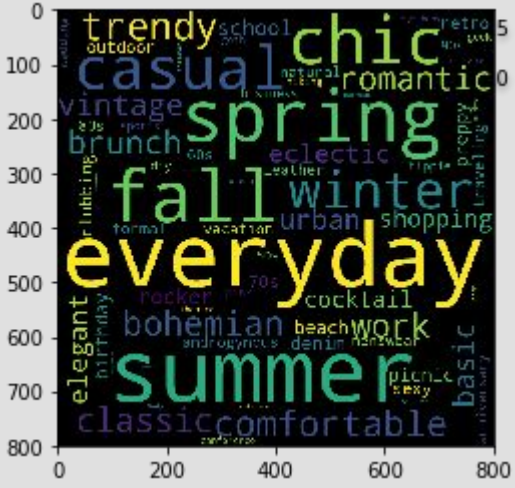
Variables

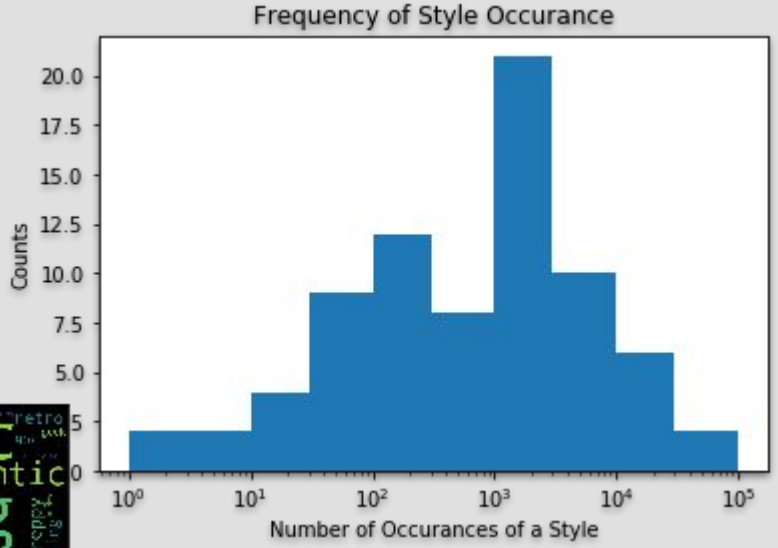
Styles

Styles are more common than tags, with a median of 1123.5 occurrences for each style

68 style occur more than 100 times in the data

These 68 styles will be used to classify the images

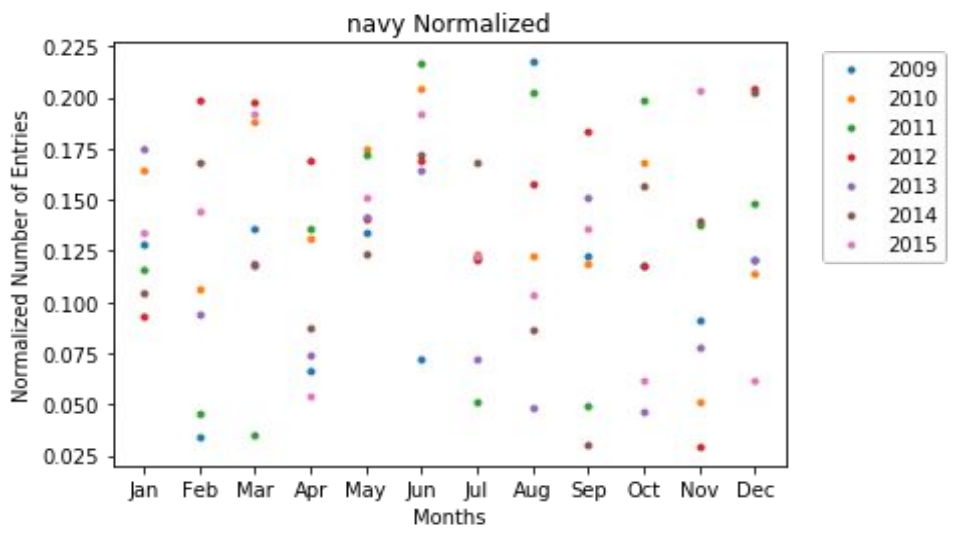


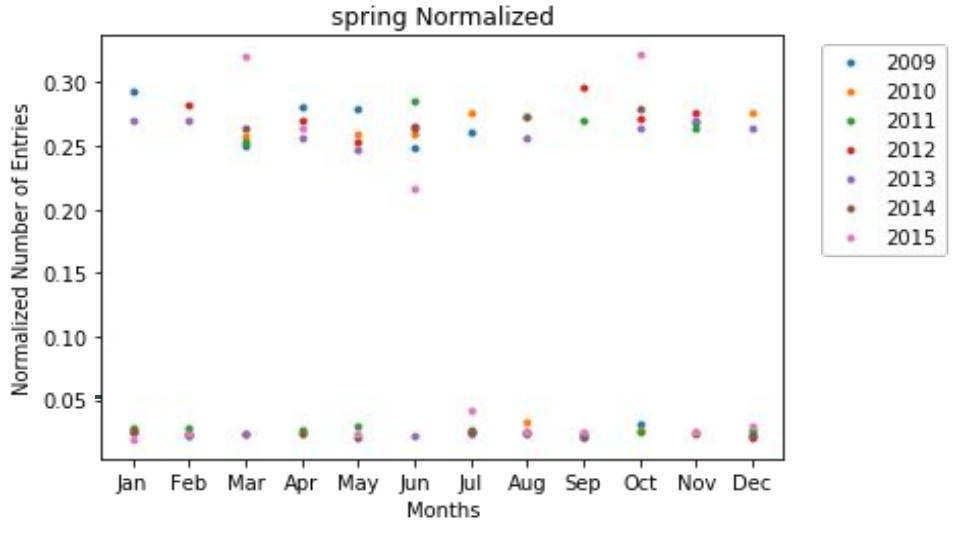


Time Data

- No correlation was found between months of the year and tags or styles
- All time data normalized before being plotted to remove the trends of the total number of uploads per month or year

- Time data may lose meaning because of the larger range of locations covered in the data
- These locations mitigate seasonal differences





Statistical Testing

Tags and styles

Compare the occurrences of styles “summer” and “winter” with tags “shorts” and “coat.”

The table shows percent of entry with summer or winter which contains shorts or coat

	Summer	Winter	Chi^2
Shorts	7.25%	1.10%	423
Coat	1.30%	13.16%	2126

Chi Squared

A chi squared test was used to test the significance of these two correlations

Each resulted in $p < 0.001$

Therefore, the null hypothesis could be rejected

Next Steps

Cleaned and Shaped

At this point, the data is cleaned and shaped to create an image classification model using the tags

Extend to Styles

If the image classification is successful, the model will be extended to use the images and tags to predict the styles of the outfit

Using other data

At this point, time data, user data, and location do not appear to be useful for the analysis

This can be revisited if the classifiers need improvement