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 300,000 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

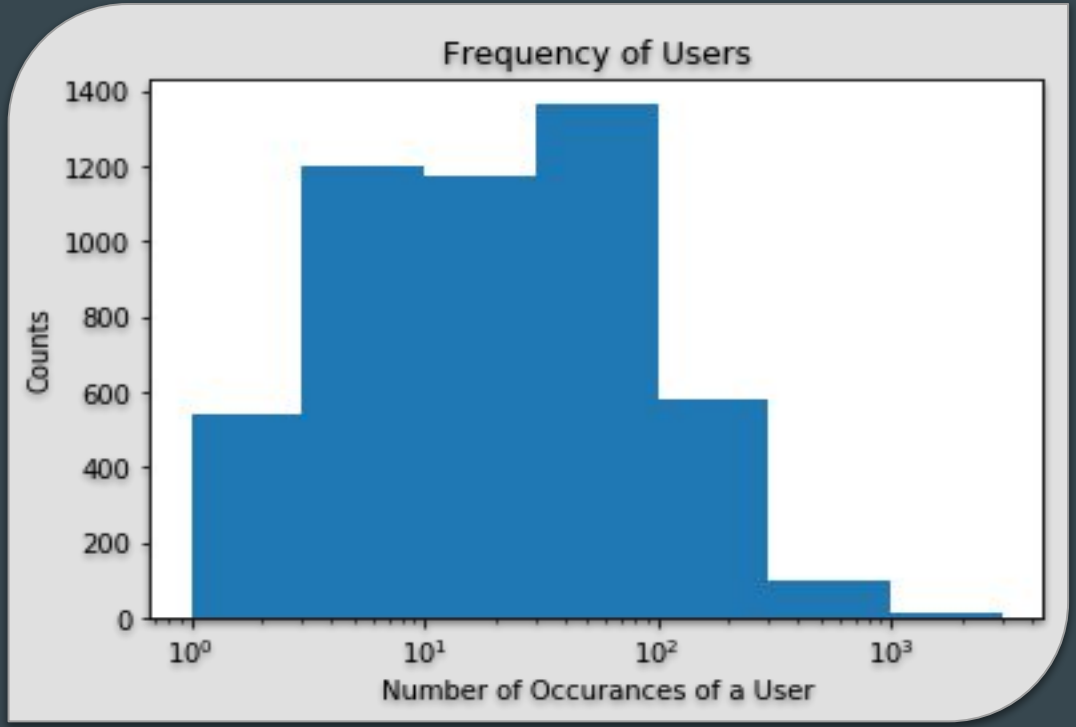
Variables

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



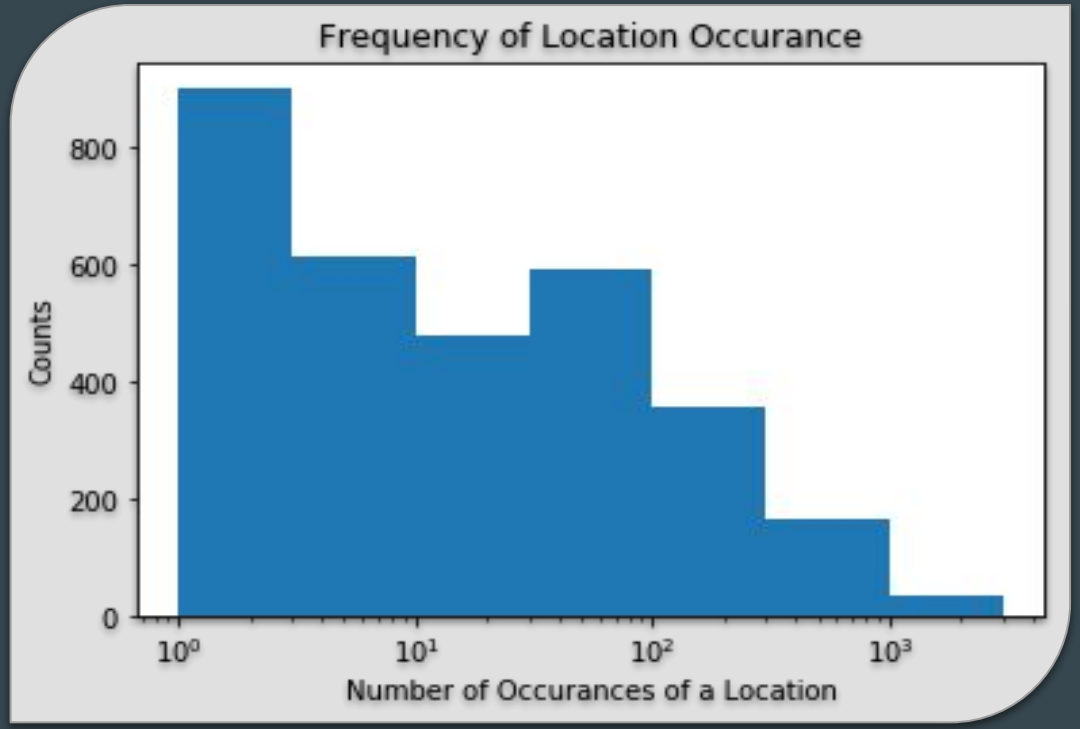
Variables

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



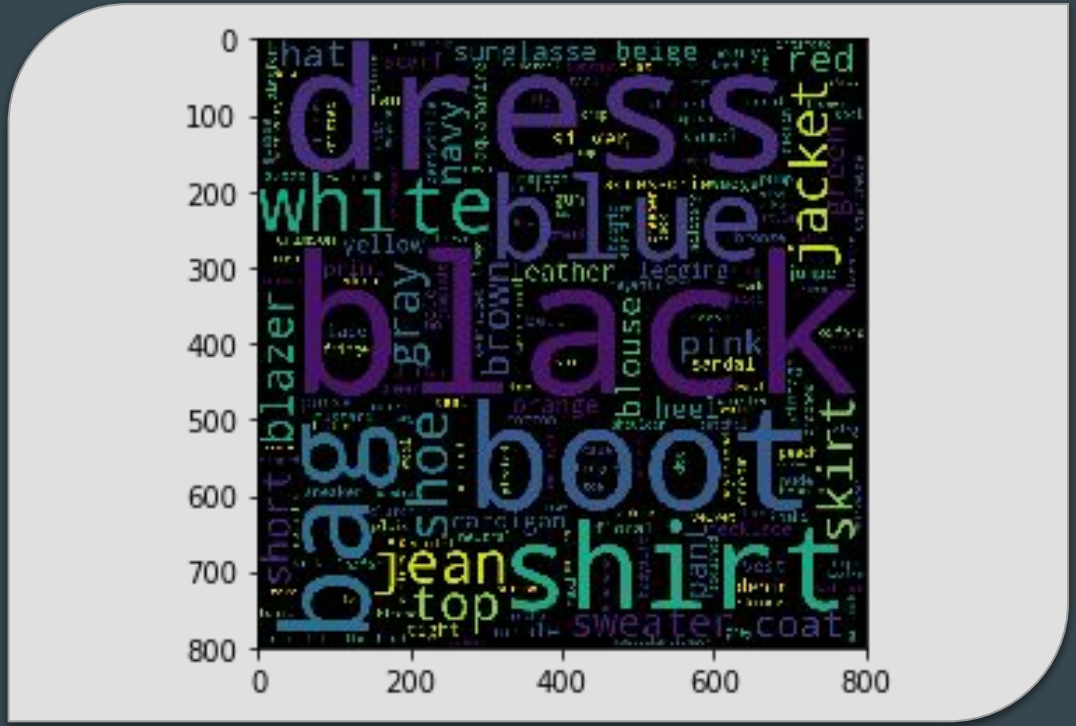
Variables

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



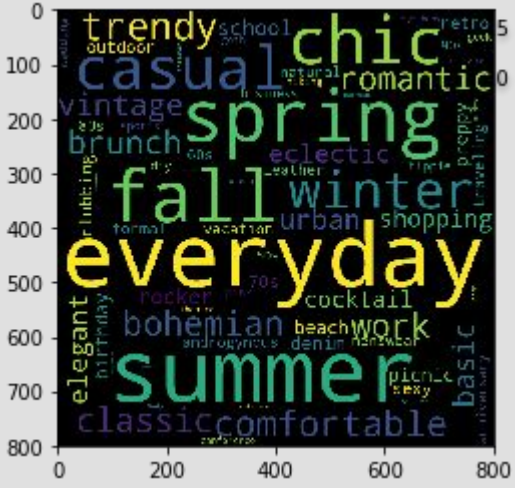
Variables

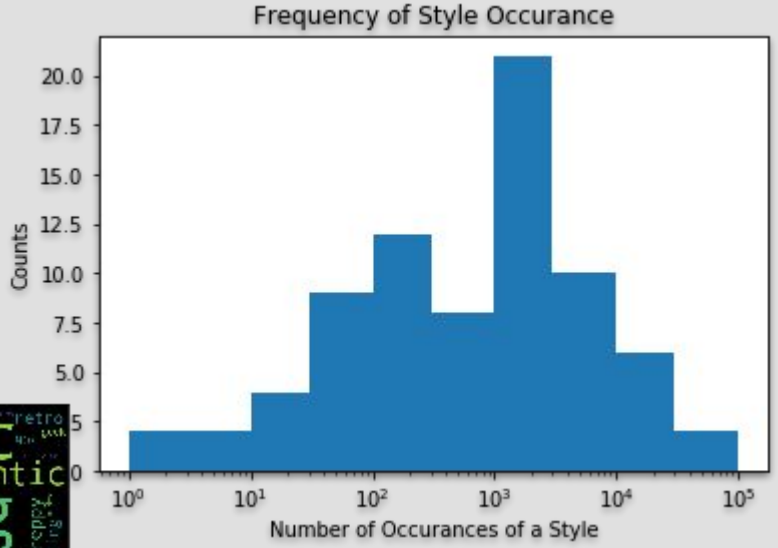
Styles

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

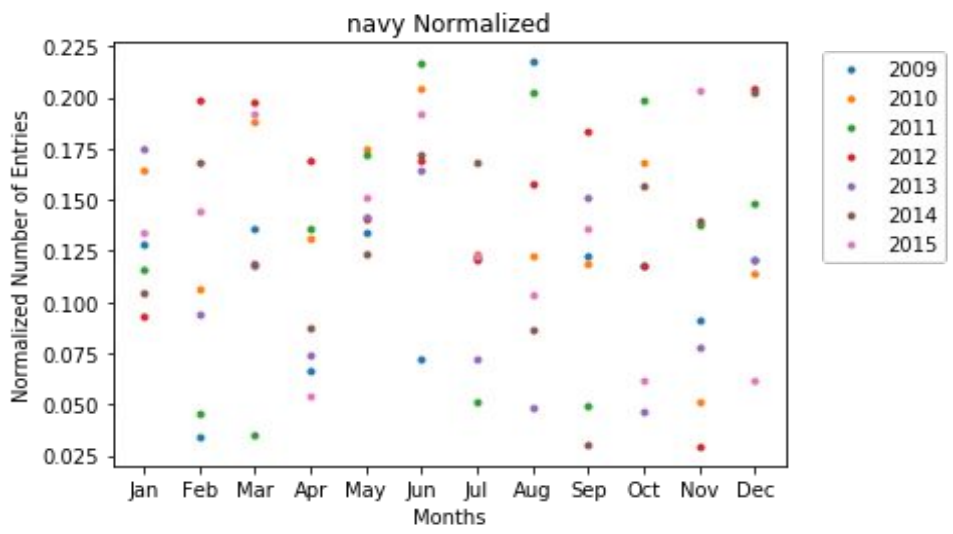


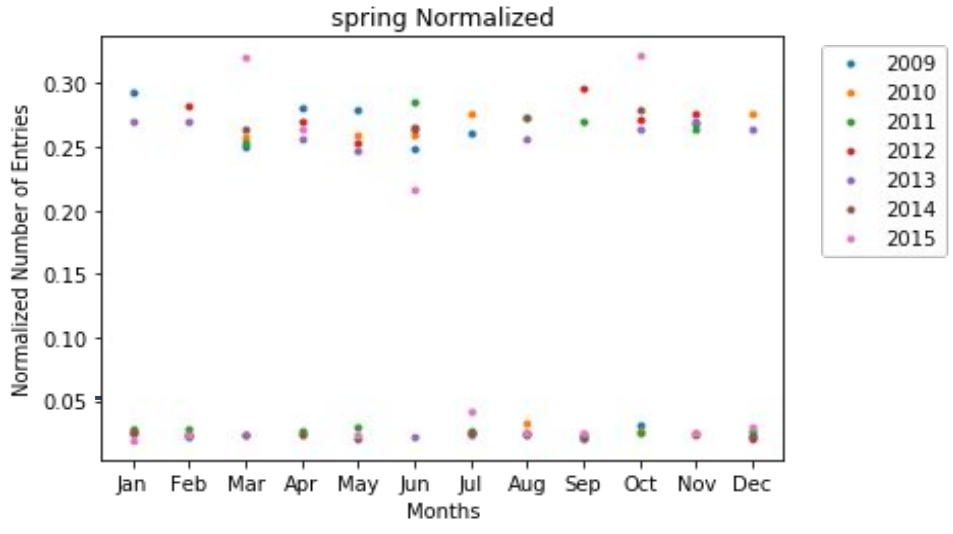


Time Data

- 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

Models - Image Classification

Image Classification

Image Classification was unsuccessful due to limit on computational power

Attempting 21 tag multilabel classification was unsuccessful on 10,000 data points

Model Structure

Layers:

- 2D Convolution
- Batch Normalization
- 2D Max Pooling
- Dropout
- Output

ReLU activation on all layers except output

Recall And Precision

Result was a high recall and low precision

Model was “succeeding” by classifying every image as having every tag

Models - Season Classification

New Model - Seasons

Tags used to predict the season of each outfit

Multiclass classification problem with four outputs: winter, spring, summer, or fall

Training and Testing

Training/testing split of 90%/10% used to maximize training sample

Close to 400,000 training samples after resampling

20,000 testing samples

Models

PCA, random forest classifier, and deep neural network classifier tested

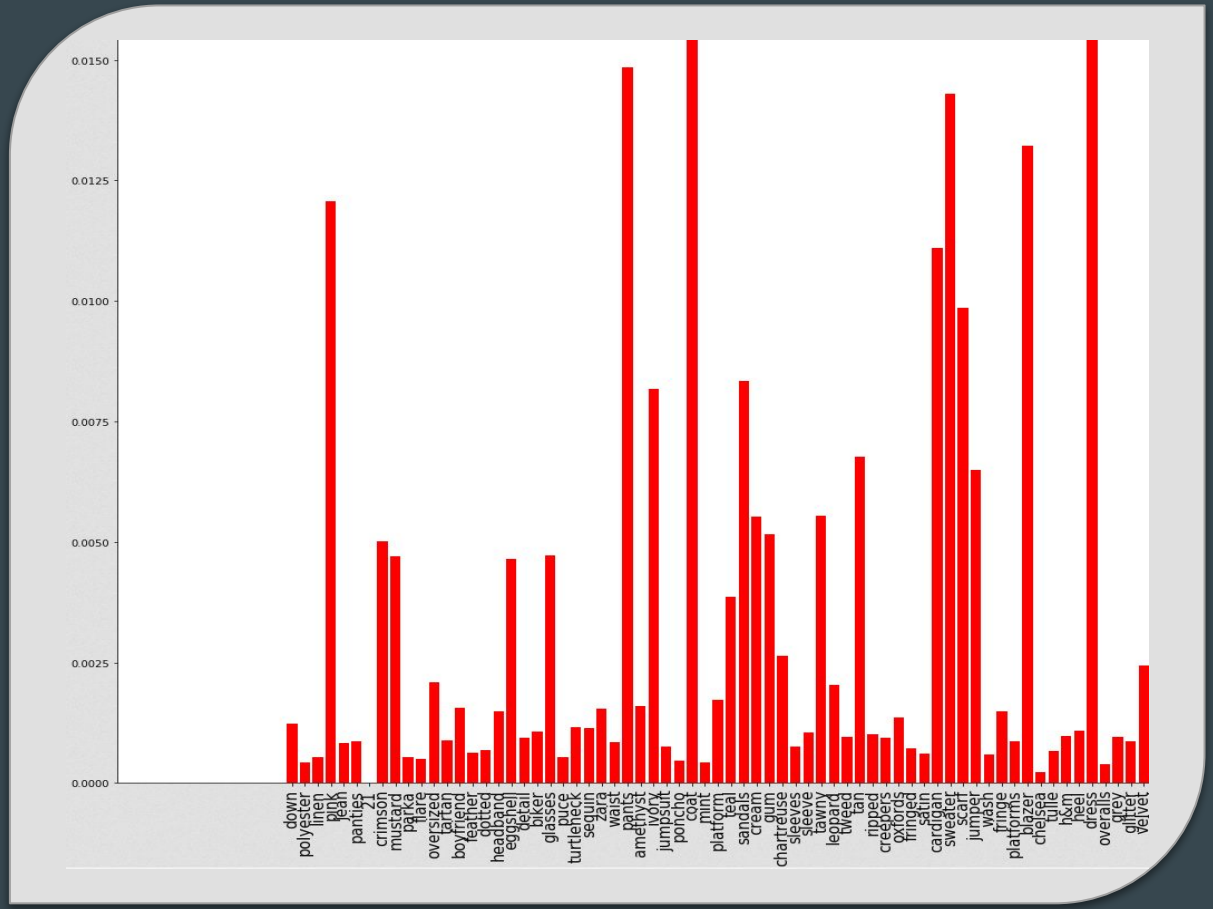
PCA did not find useful principal components

Feature Importance

Random Forest

Random forest was used to find the feature importance

Intuitive results - tags strongly associated with seasons had more importance that generic features

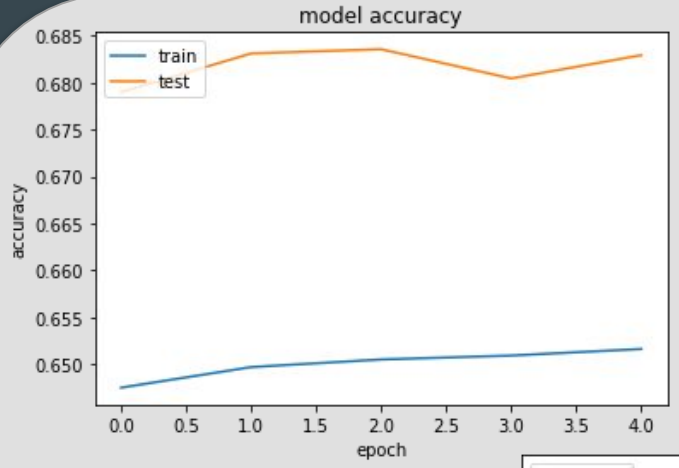


Deep Network

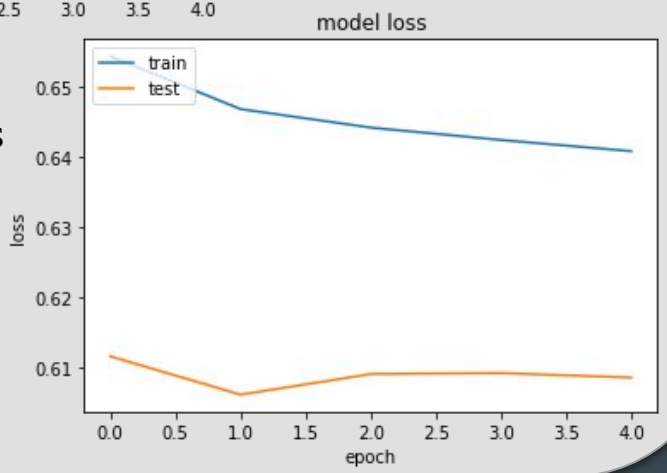
Random Forest

Most successful was a collection of densely connected feed forward layers using ReLU activation and Adam optimizer with adaptive learning rate

Softmax used for output layer



Training accuracy always lower than testing accuracy due to resampling



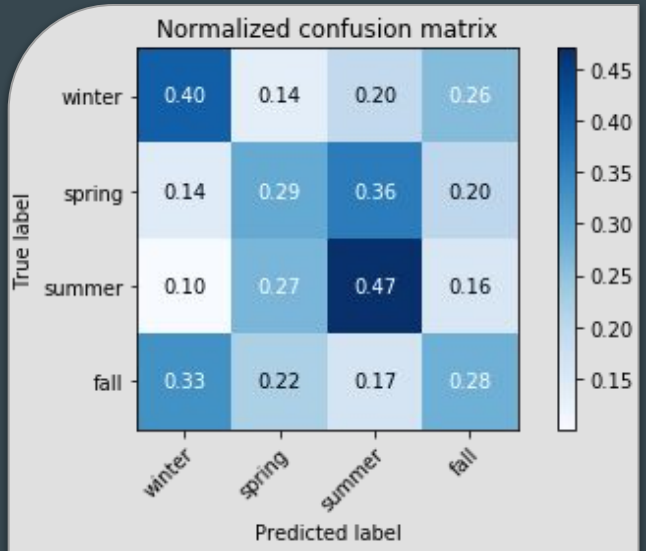
Results

Wide Vs Deep

Both wide and deep models were tested

Similar performance, best performance from deep model seven hidden dense layers, three layers of dropout at 0.2

Accuracy = 68.5%



Confusion between spring/summer and fall/winter

Metrics

The category “summer” had the highest precision and recall

Accuracy = 68.5%

Log-loss = 28

Generalizability and Business Applications

Recommendation Engines

Classifying the season of outfits useful for recommendation engines

Customers should be exposed to clothes matching the upcoming season to encourage purchases

This model could be integrated into recommendation pipeline

Feature Importance

Feature importances can tell what should be tagged on an image to best predict the season of the clothing

Future pipeline could include image classification for these most important features

Future Steps

Use Image Data

Dataset is robust, being able to use the image data is important

More computational power necessary to advance to this step

Feature Importance

Refine the model further to only use the tags with the highest feature importance

Easier to put this model into a pipeline

Resampling

Better resampling methods could be used to prevent overfitting on resampled data

Grouping winter/fall and spring/summer

Conclusions

...

Results show that this technique merits further investigation

Promising business applications for clothing recommendation systems