Automating data preprocessing with Reinforcement Learning for improved text detection and recognition in videos

Exploratory Data Analysis

CSCI S-597 Data Science Capstone (Fall 2021)

Kacper Krasowiak & Lekshmi Santhosh

Objectives of Exploratory Data Analysis

The goal of this analysis is to compute the image statistics for the training set (SynthText in the Wild) and assess if we have enough image variance that covers small text, low resolution and low light images. This will inform us about potentially including additional distortions in the image pre-processing stage.

The analysis is organized into 4 parts:

Text Prevalence

An assessment of the word and character distribution on images

Image Complexity

Evaluation of how complex the image backgrounds are

Font size

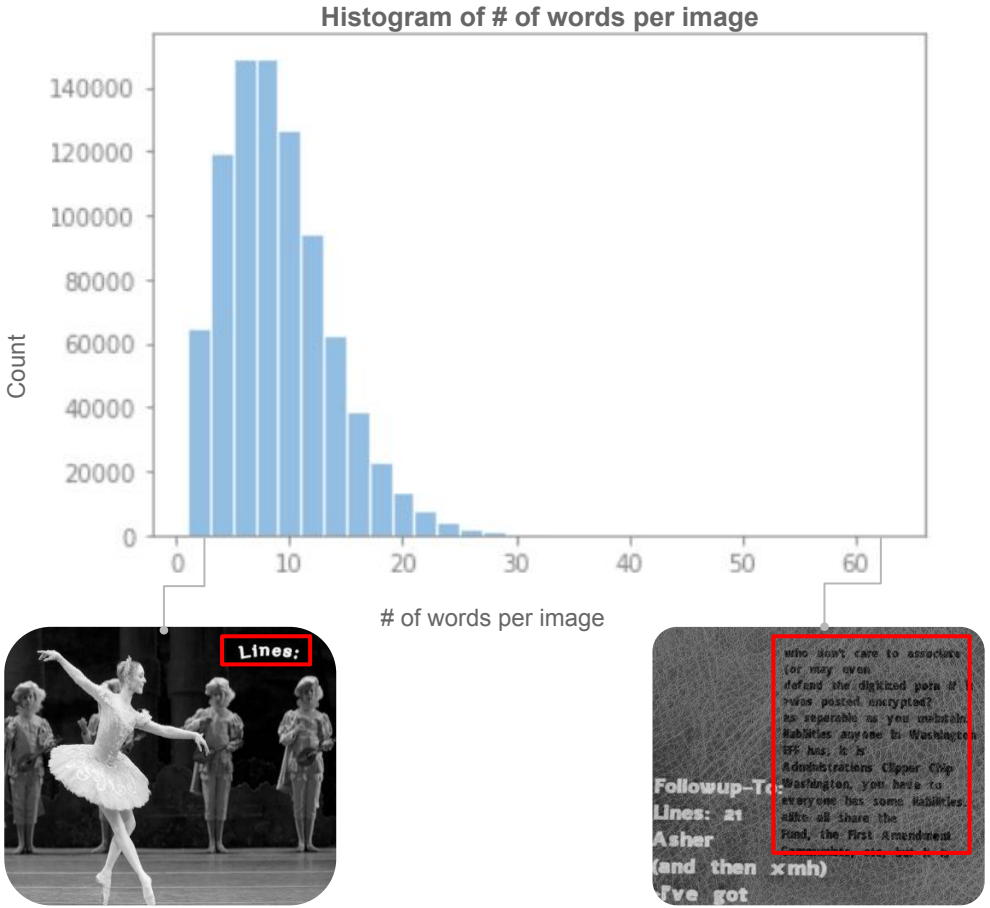
Summarization of how text size varies across all images

Brightness & Resolution

Analysis of image luminance and resolution

Section 1: Text Prevalence Assessment

How many words are there in an image?



Observations:

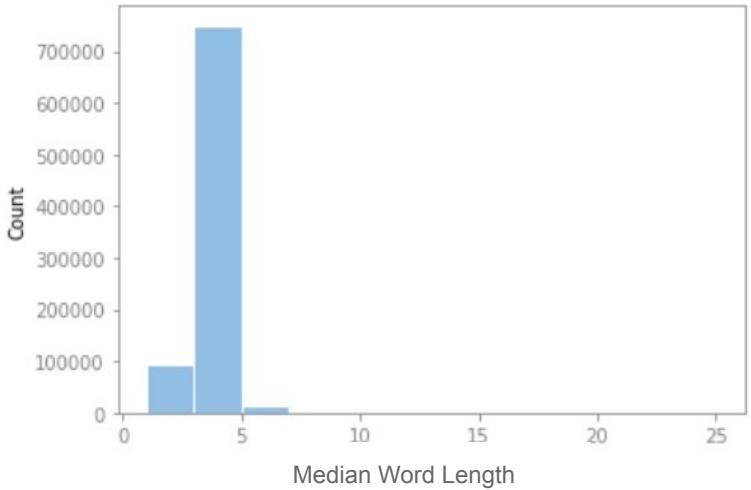
- All images have at least one word in them.
- The distribution looks log normal with median of 8 words, P75 of 11 words and P95 of 17 words.

Conclusions:

- 95% of the images have 17 words or less. Based on initial visual inspection, higher the number of words on an image, the smaller their font size. We will examine if this association holds true in Section 3.

What is the typical word length in images?

Histogram of the median word length per image



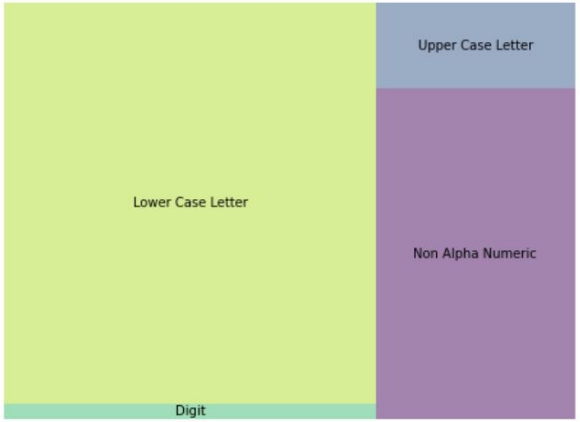
Word Cloud of the text annotations on images

Observations:

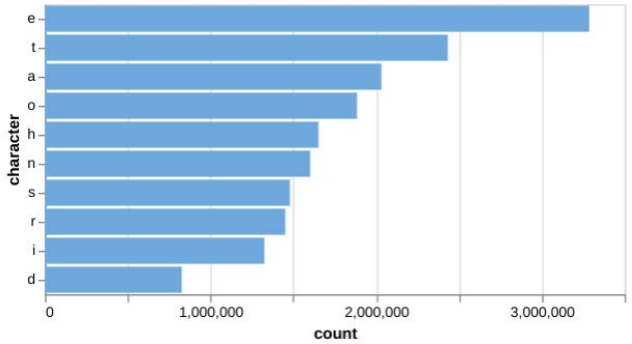
- 95% of images have median word length of 4 or less.
- We did not observe any statistically significant correlation between the median word length and the number of words per image. Given the finite image area, we were expecting images with more text to include shorter words.

How many characters are there in an image?

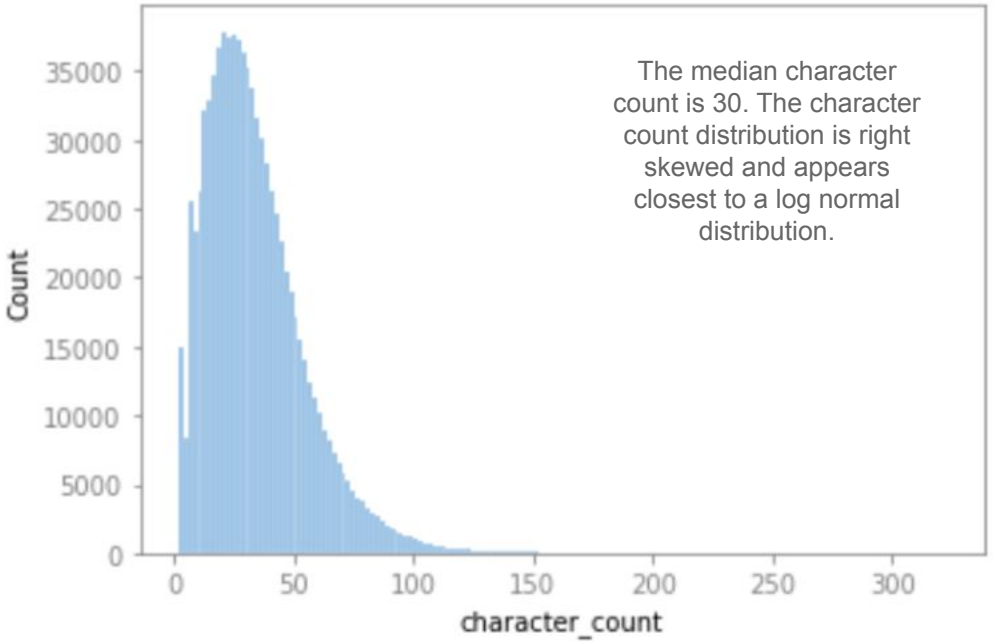
Character type prevalence



Top 10 most prevalent characters



Histogram of # of characters per image



The median character count is 30. The character count distribution is right skewed and appears closest to a log normal distribution.

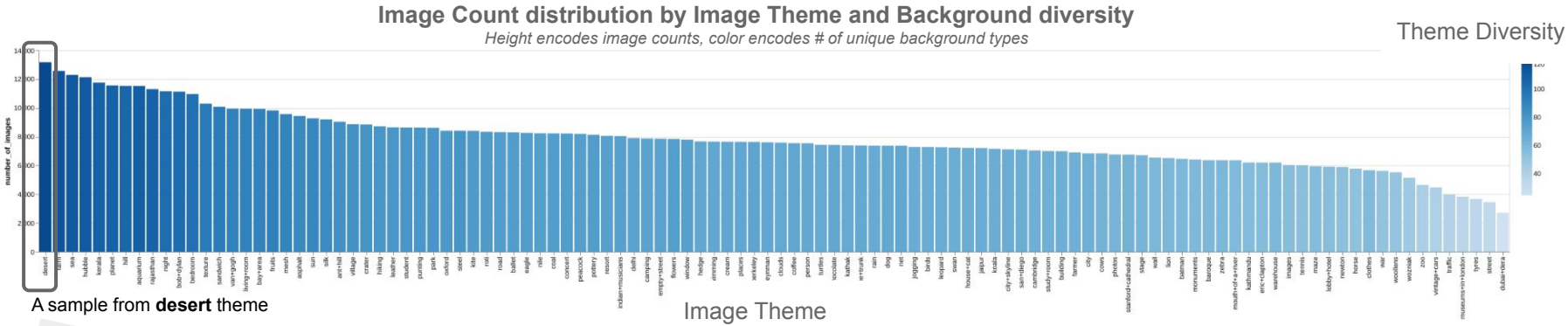
Observations:

Lower case letter is the most prevalent character type in our dataset. It is more difficult to distinguish between lowercase characters in OCR.

Section 2: Image Complexity Analysis

How much image variance do we have?

We have 858750 images in total, with **7944** unique backgrounds organized by **110** themes.



A sample from **desert** theme

There are multiple backgrounds within the same theme with varying background complexity

Multiple images share the same background.

The distribution of theme is not uniform with some themes under-represented than others. Themes with a large number of examples have more diverse backgrounds. This indicates that theme diversity is roughly consistent across themes.

How complex are image backgrounds?

Problem:
Each image background have varying complexities. Some images have plain background whereas others have busier backgrounds. During our literature review, we learnt that OCR performance deteriorates on images with complex backgrounds.

Approach:
We selected image entropy as a measure to characterize the background complexity of images. An image with lower entropy scores tend to have simpler monochromatic backgrounds.

Observations:
Image entropy distribution is left skewed with a median value of **9**. About **5%** percentage of images have entropy lower than 7.
We expect OCR word and character error rate to be positively correlated with the image entropy since it is more difficult to identify characters against complex backgrounds.

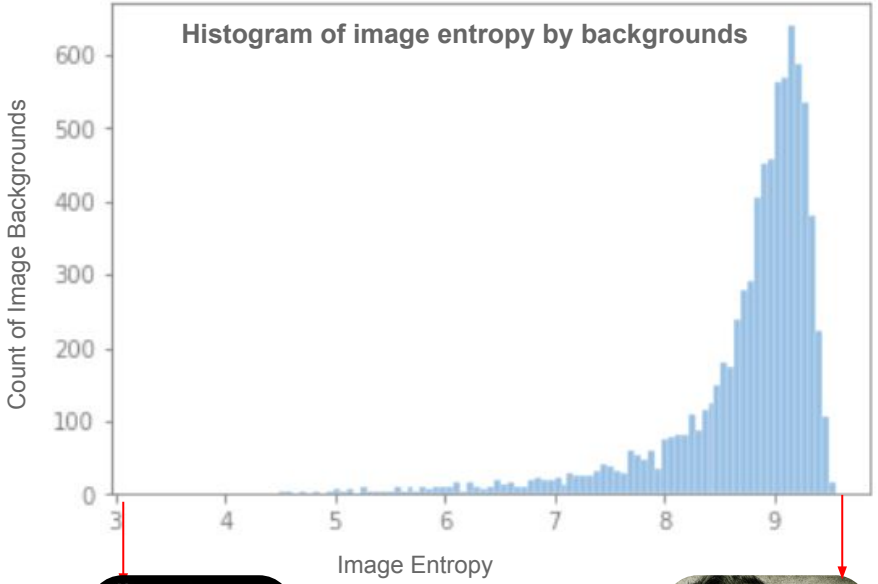


Image with the lowest entropy

Image with the highest entropy

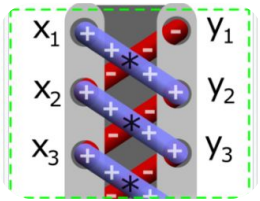
Section 3: Font Size Analysis

Methodology for font area estimation

In our dataset, we have bounding box coordinates as part of the image labels. We are computing the text areas geometrically and is using it as a proxy for font size and font coverage per image.

This metric tends to overestimate the text occupancy area but it still serves as a good proxy for evaluating how font size varies across images.

Bounding boxes are quadrilaterals. Each text label has a bounding box.

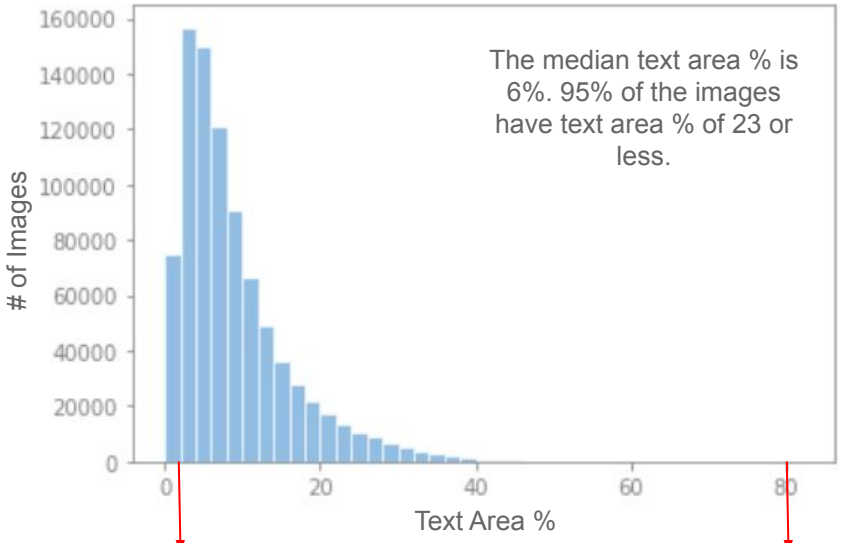


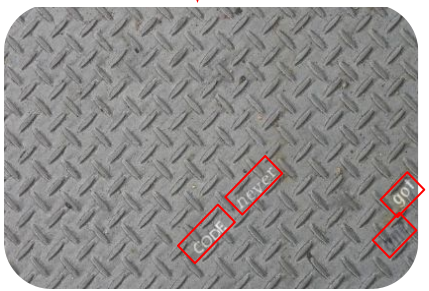
We use shoelace algorithm to compute the area of each bounding box.

$$\textit{Text to Total Area Ratio} = \frac{Sum\ of\ bounding\ box\ areas}{Image\ height * Image\ width}$$

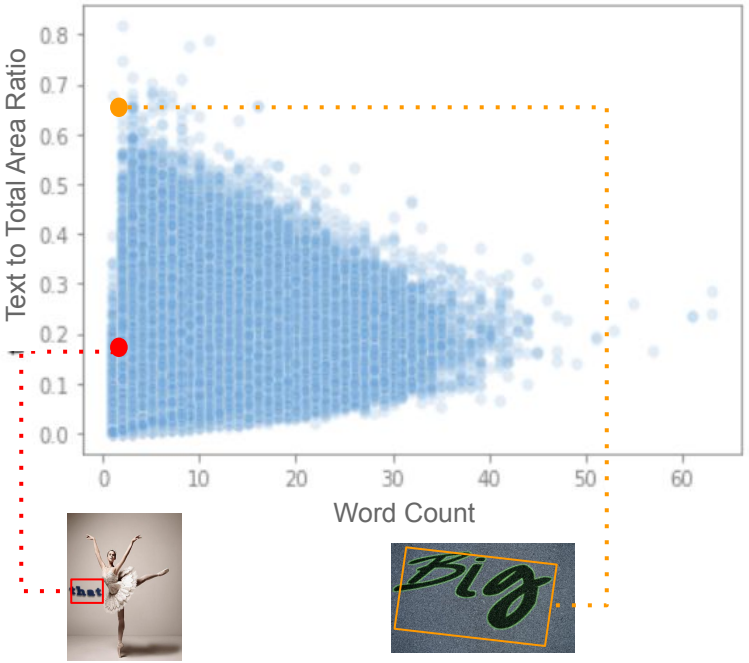
How much area does the font occupy in an image?

Distribution of Text Area %





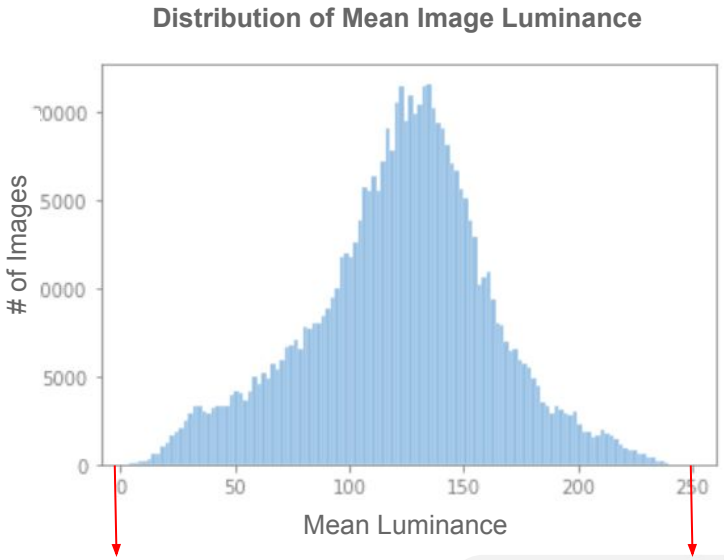
Scatterplot of Word Count & Text Area Ratio



As seen above, the variance of text to total area ratio is high for the same word count.

Section 4: Brightness and Resolution Analysis

How bright are the images?



Images in "**night**" theme have low luminance

Images in "**cream**" theme have high luminance

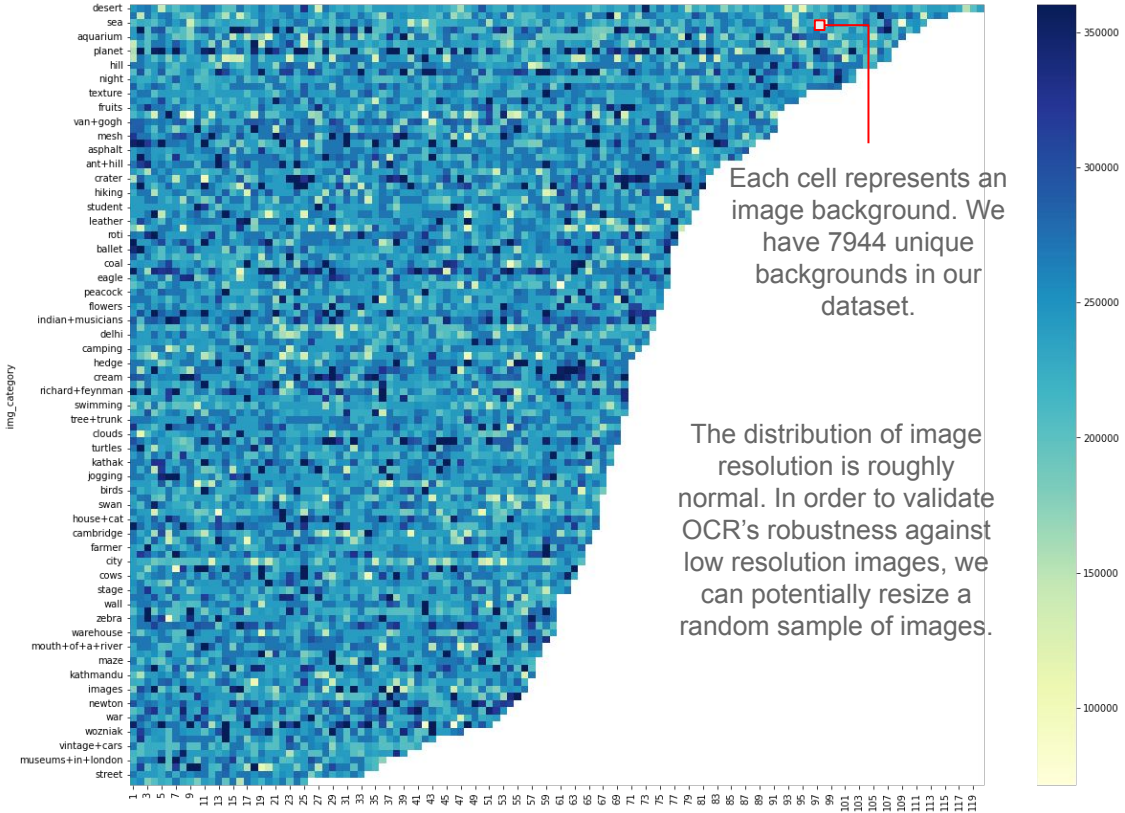
Problem:
One of the goals of our project is to recommend an image pre-processing framework which will enable OCR to be robust against low lighting conditions. Hence, We need enough low luminance images for model evaluation.

Approach:
We computed image luminance by first converting the image from RGB to LAB color space. We then extract the L layer and take the mean value of luminance per image.

Observations:
Image luminance distribution looks mostly normal with a median value of **125**. The normal distribution attributes to Central Limit Theorem since we are showing the sampling distribution of the mean luminance.
We have enough examples in our training set to evaluate the performance of OCR in low light conditions. We expect OCR word and character error rate to be positively correlated with the image luminance.

Do we have low resolution images in the dataset?

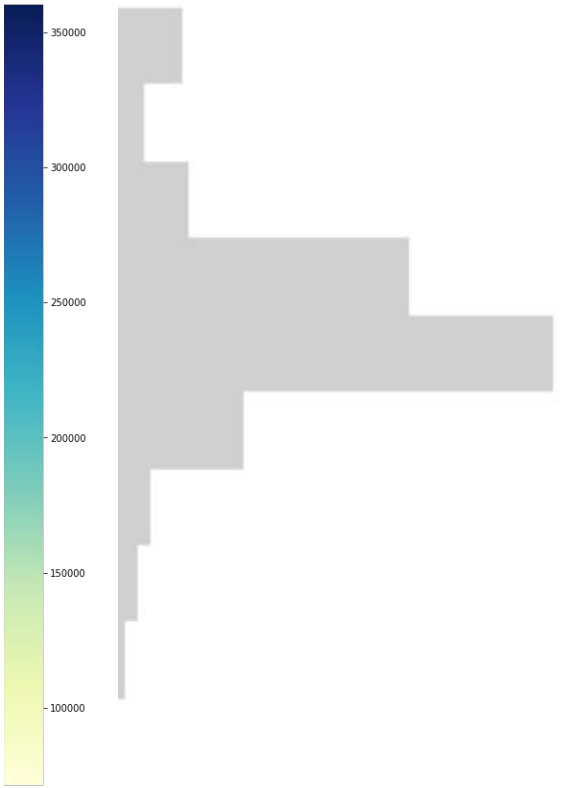
A Heatmap of pixel counts per image by theme & background variant



Each cell represents an image background. We have 7944 unique backgrounds in our dataset.

The distribution of image resolution is roughly normal. In order to validate OCR's robustness against low resolution images, we can potentially resize a random sample of images.

Distribution of Image Resolution (# of Pixels per image)



Conclusions

Findings and Next Steps

- Through the EDA, we learnt that our images have ~ 8000 unique backgrounds of varying complexity and image resolution. If sampling of images is necessary, a **stratified sample** (stratified by background) will best preserve the image variance in the population.

- We have sufficient variance by image resolution, low lighting conditions and text font size. We do not need to synthetically generate datasets for class imbalance.

- Through this exercise, we were able to characterize metrics to measure image complexity, luminance, text font size and resolution. This will help us define appropriate **image cohorts** for Reinforcement Learning framework evaluation.