Group - 6
Piyush Chaudhary - pc1905
Rewa Jayant Kale - rjk422
Samet Taspinar - st89

Image Exif Data Analysis Update

Background

It is interesting to analyze the metadata of images that can reveal various crucial information about images. Exif headers contain information about the name, camera brand (make) and model, date & time, resolutions, orientation, Huffman table, Quantization Table and so on. Other than these, several interesting information can be obtained from them, some of which are whether the image is modified, if yes, which program is used, or is in its original dimensions which the camera allows. Moreover, having many images, one can learn about people and their behaviors more. For that reason, we downloaded more than 6 million images (12 TB) in their original resolutions belonging to all Flickr users from Abu Dhabi who shared public pictures. Among these images, 77% of them are marked as blacklisted as they are modified by one of the image editing software and the rest of the images were marked as whitelisted whose detailed information we extracted and will be shown in our visualization. Note that, this data is collected until September 2014, therefore, the latest image is expected to be at that time.

Problem Statement/Objective

Understanding trends in digital photography and camera usage is critical for camera companies and digital photography experts. In order to understand these trends, the question of how consumers are using their products must be answered. This project will allow camera companies and experts to understand how their products are utilized and therefore what direction the industry is headed.

The solution of this problem can be useful to not only the experts in digital photography and camera producers but also image forensics researchers and people who are interested in learning more detailed information about images and cameras. Thanks to our visualization, these people can easily view and understand different aspects of trends in imaging and camera industry.

Analytical Questions

We calculated the entropies for each of the attributes. Looking at the entropy values, we checked which attributes have more than 90% similar values in the data like the Huffman value of images and also which have got all distinct values. We won't be considering those attributes in our analysis as they won't give meaningful results. Accordingly as was mentioned in the feedback we reduced the number of analytical questions we going to answer.

The initial list of analytical questions that we had come up with were:

1) What is the distribution of commonly used camera models in Abu Dhabi over time, is there any trend?
2) What are the most popular settings for cameras?
3) What are popular timings of the day for clicking pictures?
4) What is the popular dimension used for pictures?
5) What is the distribution of different Huffman and quantization tables?
6) Which of the EXIF tags have the highest entropy?

But after elimination of unimportant attributes, limitation of data collected and time, we reduced the number of questions. We came up with a list of 3 final questions which are listed below.

Some of the analytical questions that can be useful to understand trends are:

1. What is the distribution of commonly used camera models over time, is there any trend?
2. What are the most popular settings for cameras (resolution, brightness, white balance)?
3. What are the most popular brands?

Data Attributes

The initial data attribute list that we felt would help answer the analytical questions was as follows:

Attribute Name	**Type**	**Purpose**
Make	Categorical	Brand details
Model	Categorical	Model details
ISO Speed	Categorical	Camera's sensitivity to light
Brightness	Quantitative	Brightness of camera
Exposure Time	Ordinal	Total time of sensor's exposure to light
Flash	Categorical	Is flash ON or OFF
DateTime	Ordinal	Date time when picture clicked
XResolution	Quantitative	X value of resolution
YResolution	Quantitative	Y value of resolution
Huffman Table	Categorical	Huffman values table
Quantization Table	Categorical	Quantization values table
Image Quality	Quantitative	The quality of image [0-100]
Focal Length	Quantitative	Distance between lens & curved mirror
Digital Zoom Ratio	Categorical	Zoom factor for digital zoom
Geolocation	Quantitative	Geographical location of image

The final list of attributes that we used to answer the final list of three analytical questions was as follows:

Attribute Name	**Type**	**Purpose**
Make	Categorical	Brand details
Model	Categorical	Model details
Brightness	Quantitative	Brightness of camera
Exposure Time	Ordinal	Total time of sensor’s exposure to light
DateTime	Ordinal	Date time when picture clicked
Image Quality	Quantitative	The quality of image [0-100]
Focal Length	Quantitative	Distance between lens & curved mirror

What have others done to solve this or related problems?

http://diffractionlimited.com/choose-ccd-camera/

The above link provides the analysis as in how to choose a better CCD camera. Analysis of cost/size, focal length, sensitivity, resolution, software, etc is done and it helps in choosing a better camera.

https://tysonrobichaudphotography.wordpress.com/2012/07/16/how-do-aperture-and-focal-length-affect-the-dof-or-exposure-on-different-sized-sensors/

The above project explains how focal length, apertures affect the exposure of cameras.

http://www.cambridgeincolour.com/tutorials/digital-camera-sensor-size.htm

The above project explains the comparison with help of sensor size and focal length, lens size, pixel size.

Describe related works and explain how they are related to your work.

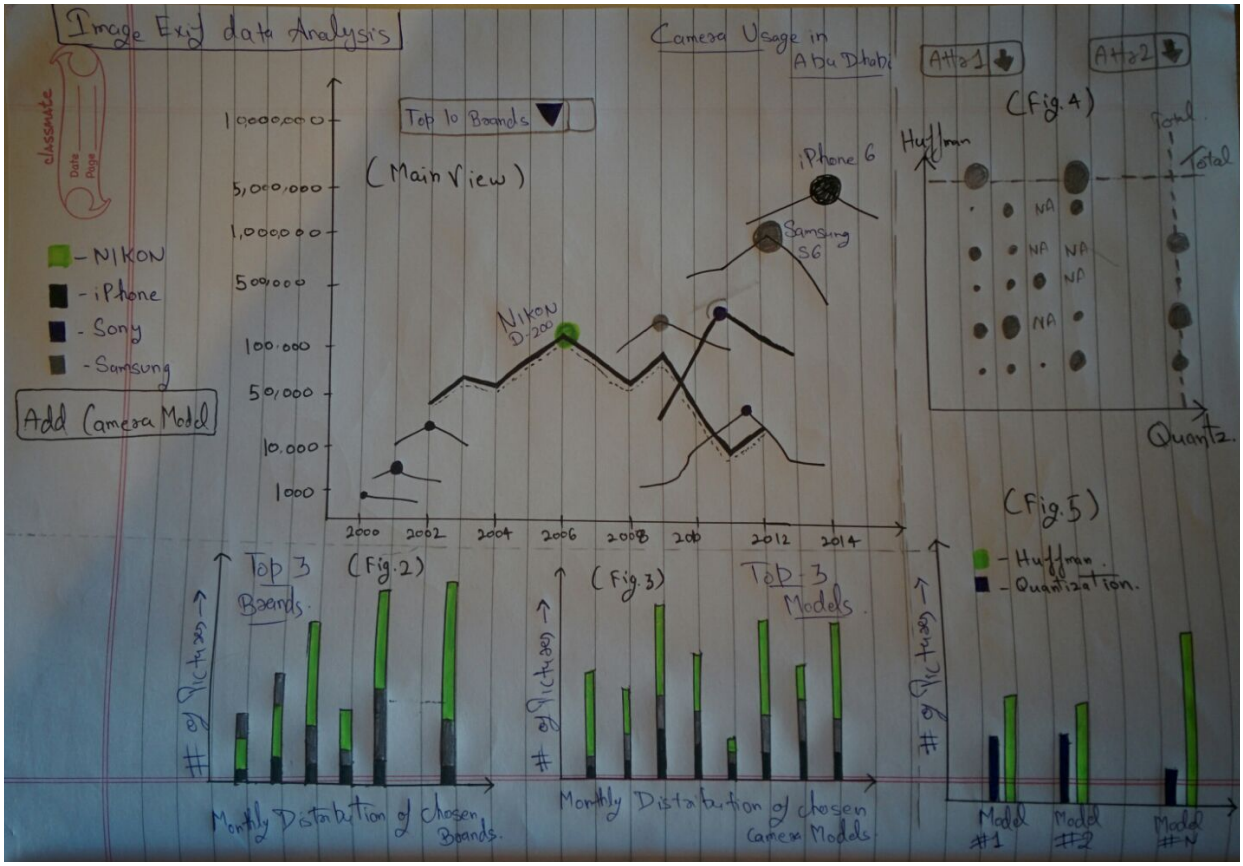
The related works and project links above have done some camera settings and parameters analysis. It helped us in understanding what camera settings can affect a camera image. Similarly, from the available data collected, we performed the analysis of camera settings. Along with that we also determined the top performing camera models and brands.

Initial mockup

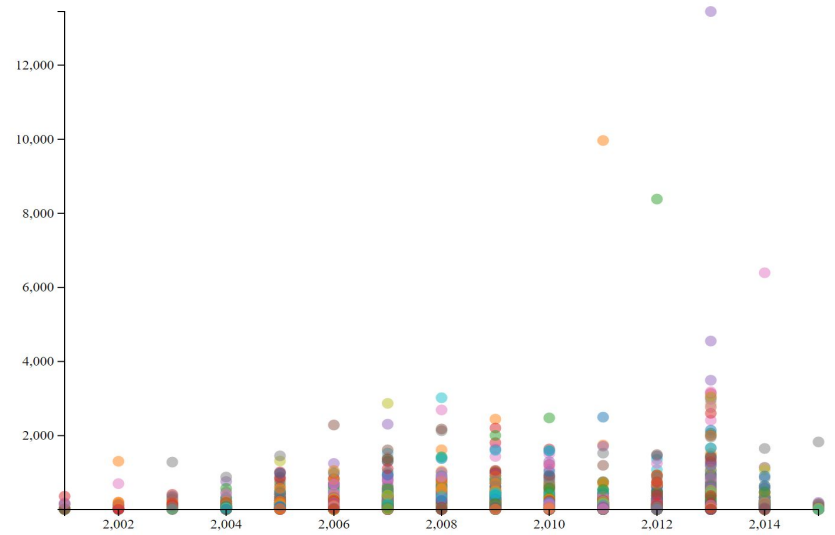
In this mockup, we had decided on 5 different figures and a sidebar that contains the brand names whose goal was to answer the main question and the analytical questions.

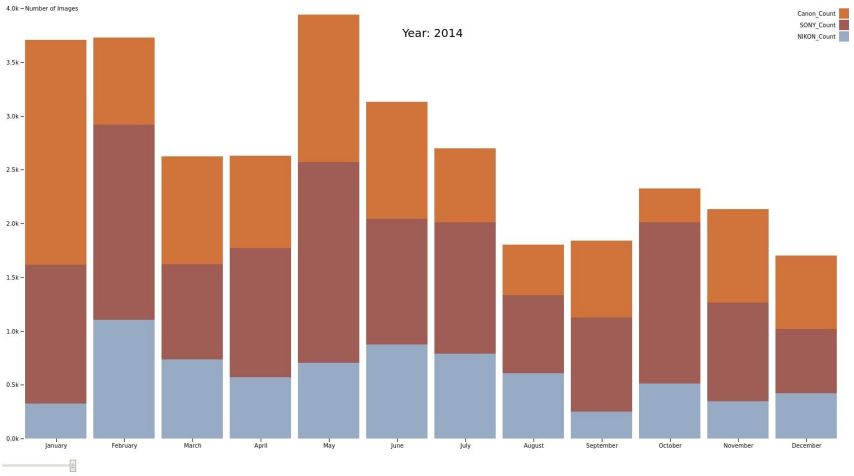
The Figure 1 was the main figure which would give the yearly distribution of models. It was proposed to have a drop down menu for selection of brands. Below the main figure, we had 2 bar charts for the monthly distribution of brand and model analysis respectively. On the right-hand side of the main figure, we had two figures, one for Huffman versus quantizations values analysis of images and below that a bar chart for huffman and quantization values count for models.

The figure was proposed to be made interactive by mouse hover, like when we hover on a particular model in figure 1 the model and brand distribution for that selection will be shown in the two figures, Figure 2 and 3 below the main figure. The figure 4, the figure on the right-hand side of the main figure would also highlight that model and the Figure 5, below the Figure 4 would show the distribution of Huffman and quantization value counts.



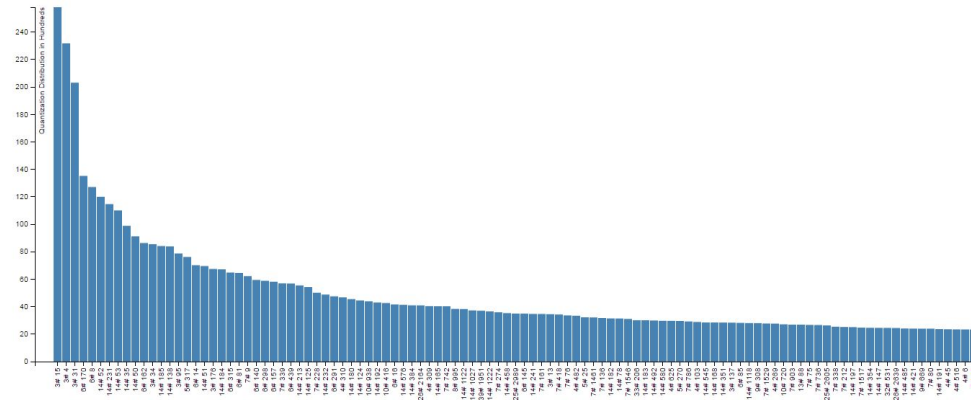
Mockup implementation figures:





Bar Chart : Quantization Distribution for Make#Model

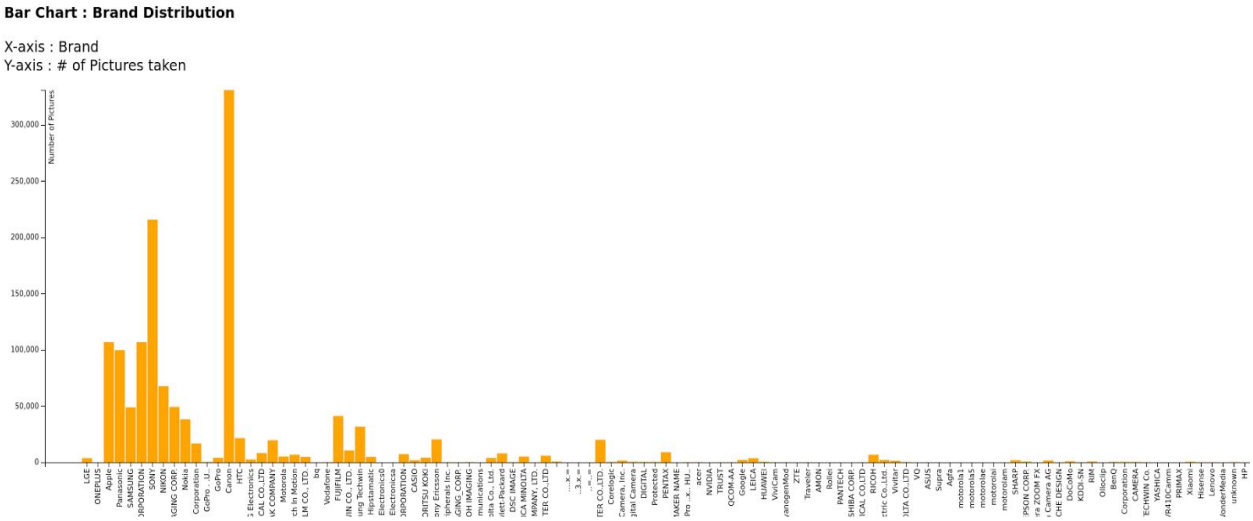
X-axis : Make#Model
Y-axis : Total Quantization Distribution in Hundreds



The problems in implementation and analysis of mockup:

1. In the initial mockup, figure 2 and 3 were slightly off-topic.. This two information was not contributing as figure 1 was already showing yearly distributions. Along with that, we realized we missed two very crucial question about the camera trends: "Which cameras are used most?" and “How is the overall usage of these brands over time”. Because of that, we decided to change these two graphs.

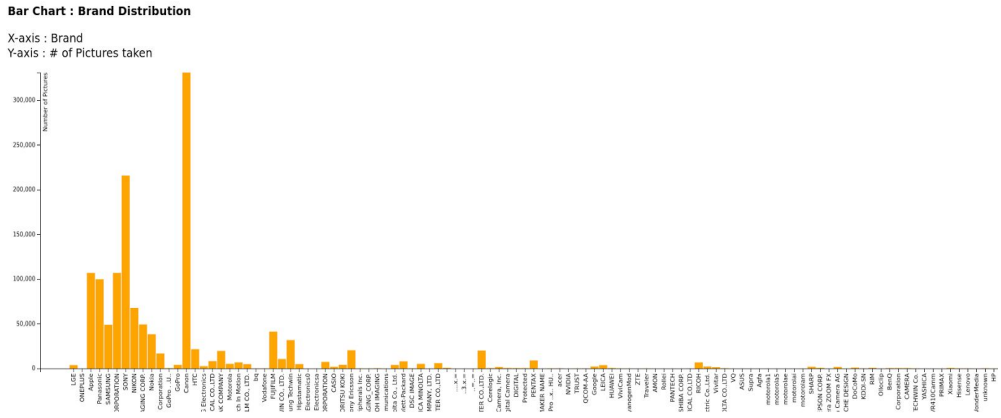
2. The stacked bar chart for monthly distribution of brands was redundant after the yearly distribution of models in figure 1. Also, we got feedback from the professor that we could convert it to brand distribution as below:



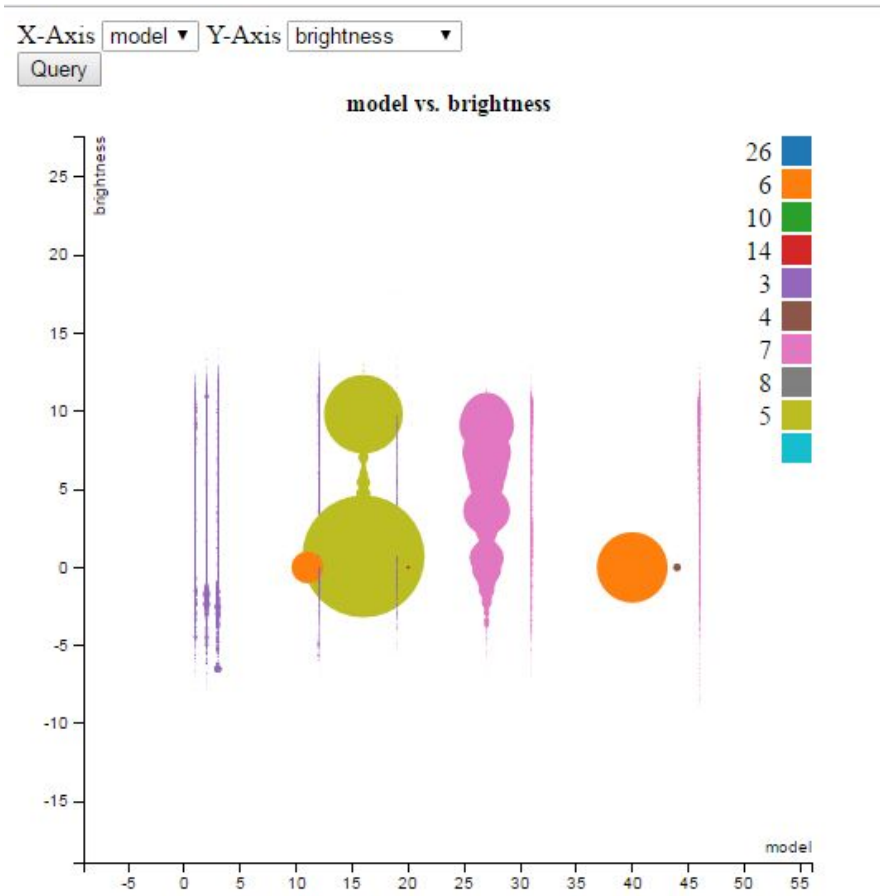
3. Figure 4 and 5 were too complicated for users to understand and there were a lot of unnecessary information provided. For example, Huffman and Quantization tables are too difficult for anyone to understand and find the relationship between different values. Therefore, we analyzed the whole data to choose only meaningful information which has enough entropy that users can relate. Our data analysis indicated that the best settings to show to users are "average exposure time", "average focal length", "average brightness" and "average image quality" for each model.

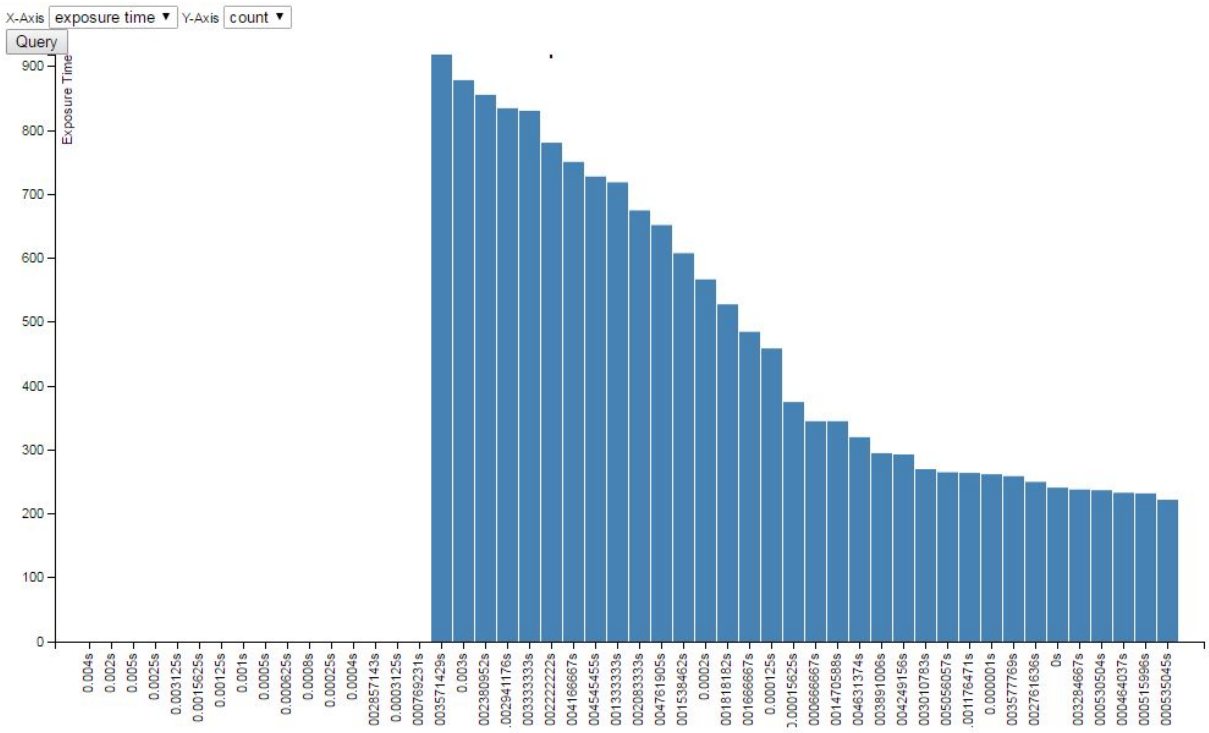
Second Mockup

The Figure 1, the main view remained the same but we changed the figure 2 as follows and were going to have figure 3 similar to for model distribution:



The figure 4 was as follows having a dropdown for selection of each of the camera settings and the comparison of model versus its each of the camera settings for top 48 models.





The figure 5 was as above, a simple bar chart for each of the selected camera settings from the dropdown and the comparison of its value and the number of images for it.

The problems with this implementation and analysis:

1. Our mentor, Josua, had suggested us a replacement for figure 3 which was initially model distribution bar chart to an area chart for checking the performance of top 5 models over the years. He told us to refer the following link for that stacked area chart in Figure 3: https://www.google.com/get/videoqualityreport/
2. Also, we took advice from Prof. Enrico for Figure 4 and Figure 5, he told us to plot a boxplot for distribution of all average values of the camera settings. Alternatively, he suggested we could have a line chart for their distributions. He told us that the current Figure 4 and Figure 5 would remain disconnected from the main view and are hard to understand for a user. Also, they don't give better insights.Our initial mockup was too complicated and instead of 2 scatterplots, we would simply add 4 small box plots, one for each setting that could serve users more meaningful information.

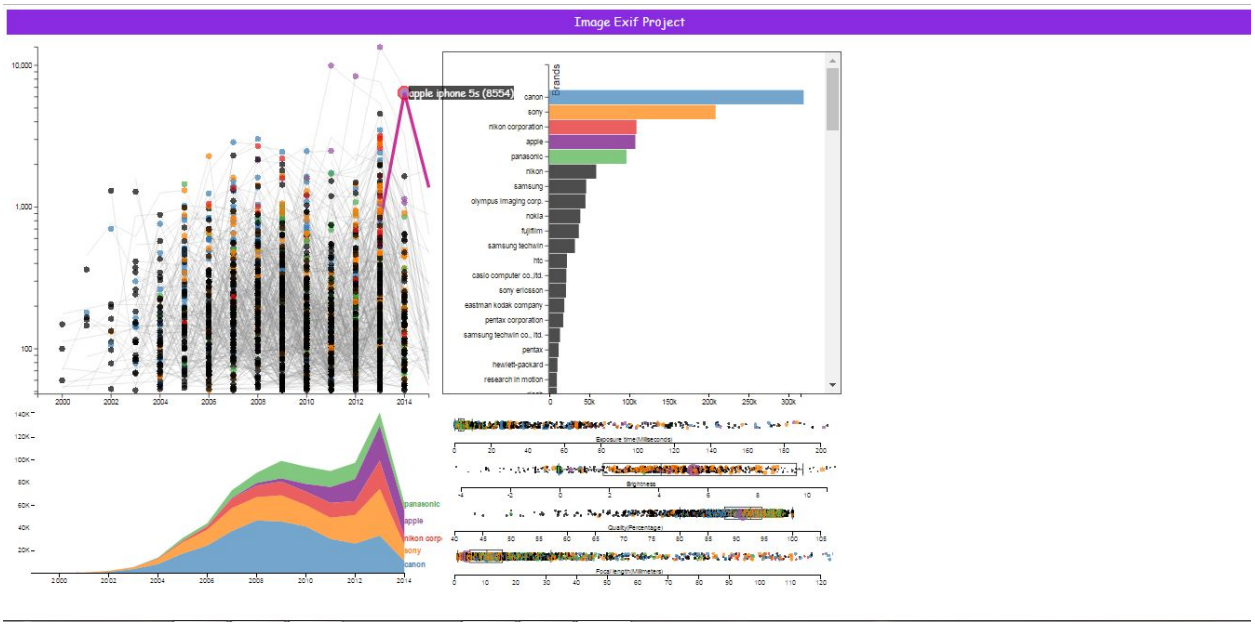
Third mockup

In this step, we implemented all the modifications mentioned above and did slight modifications but general characteristic was the same.

Our biggest focus on this step was color coding of the data. In the previous steps, the color synchronization of the graphs wasn't well thought and there were some problems. Also, we tried to color many all the dots which were causing confusions. In this step, we chose top 5 brands to be color-coded and the rest of the brands to be gray. This was a very crucial step that made the graphs highly related and easily understandable. We took help of Color Brewer and selected the following colors for the top 5 brands:

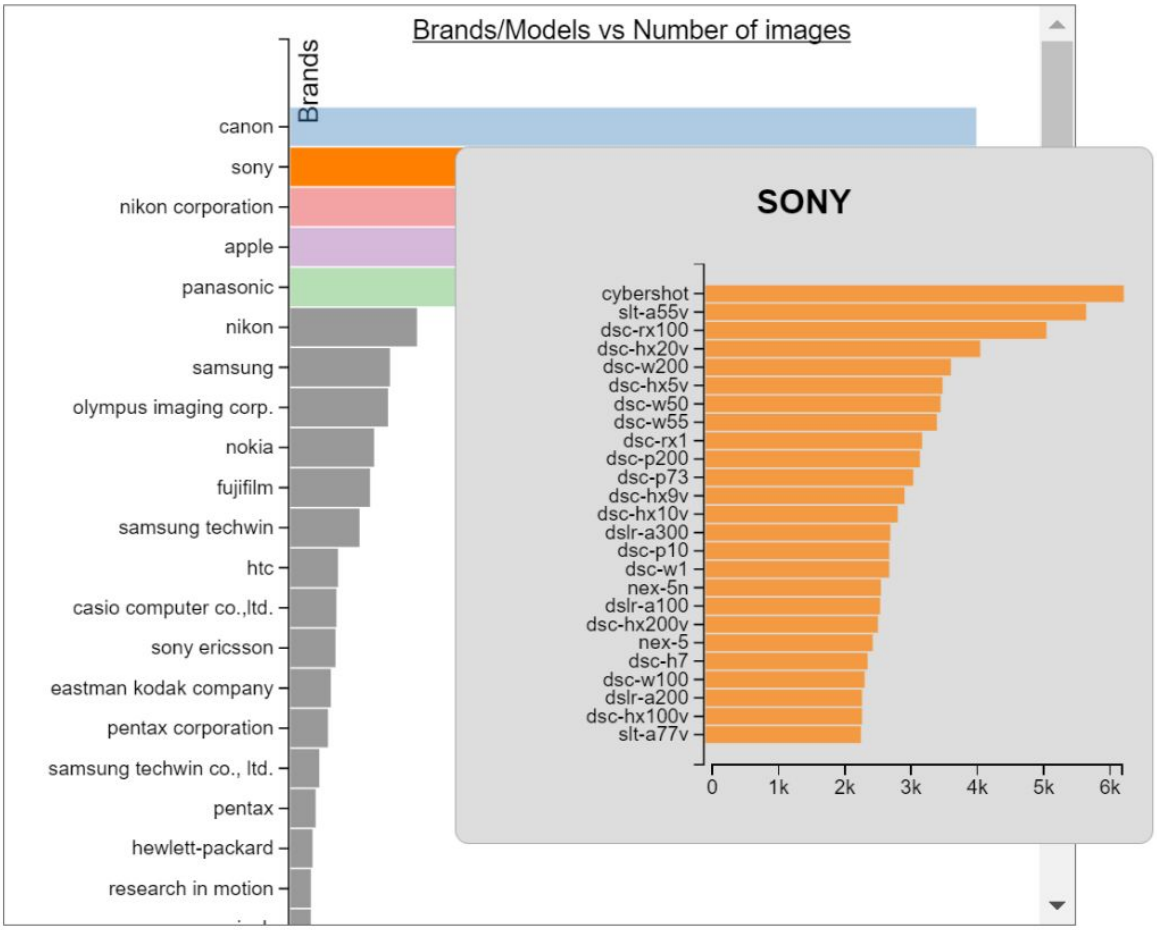
'Canon'	= "#377eb8"
'Nikon Corporation'	= "#e41a1c"
'Sony'	= "#ff7f00"
'Panasonic'	= "#4daf4a"
'Apple'	= "#984ea3"

Another improvement was the interaction between the figures which was relatively difficult to implement as we had many figures. When we hover on a circle on fig 1 or fig4, the corresponding brand or models in the other graphs were also highlighted.



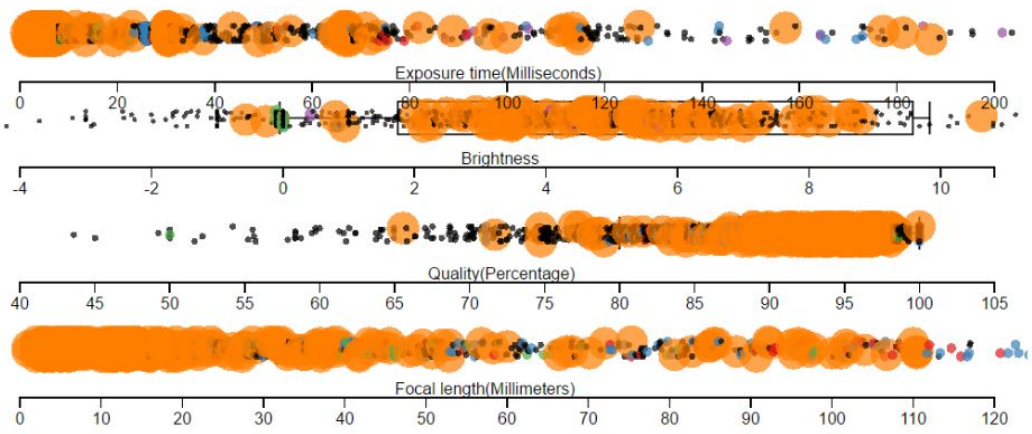
Forth Mockup

The second figure was quite weak and it was not showing the model-wise distribution of brands. For that reason, we added the capability to see up to top 25 models for each brand when the corresponding bar is hovered.

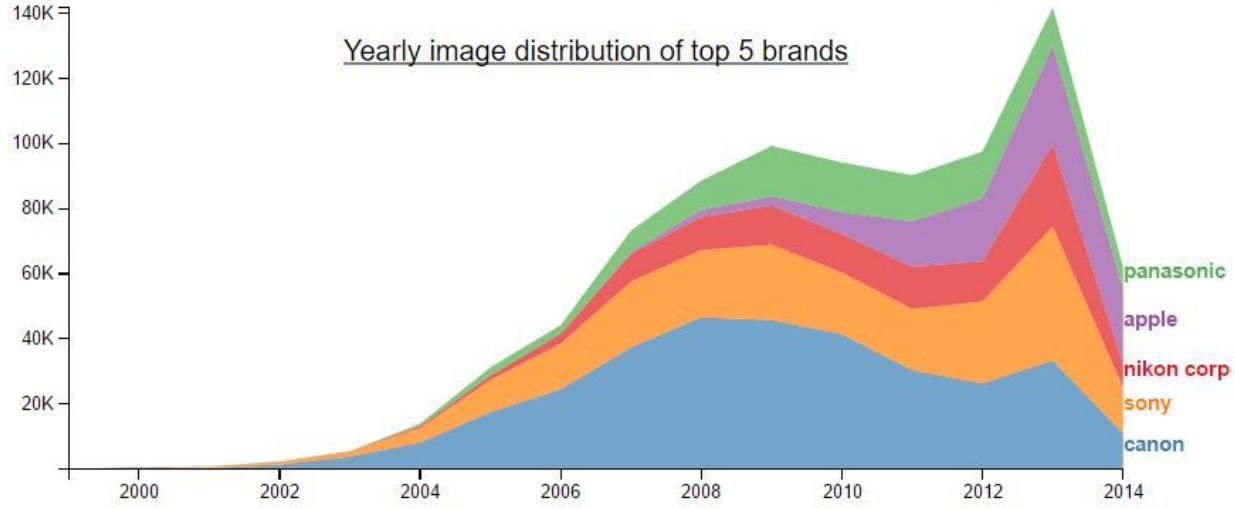


Also, we increased the size of circles when hovered with mouse pointer on figure 4.

Since there were approximately 2000 circles within a very small area in the box plots, it was extremely difficult for users to understand the general characteristic of different brands in box plots, therefore, we increased the size of the circles when they are hovered. This allowed us to see the results a lot clearer.

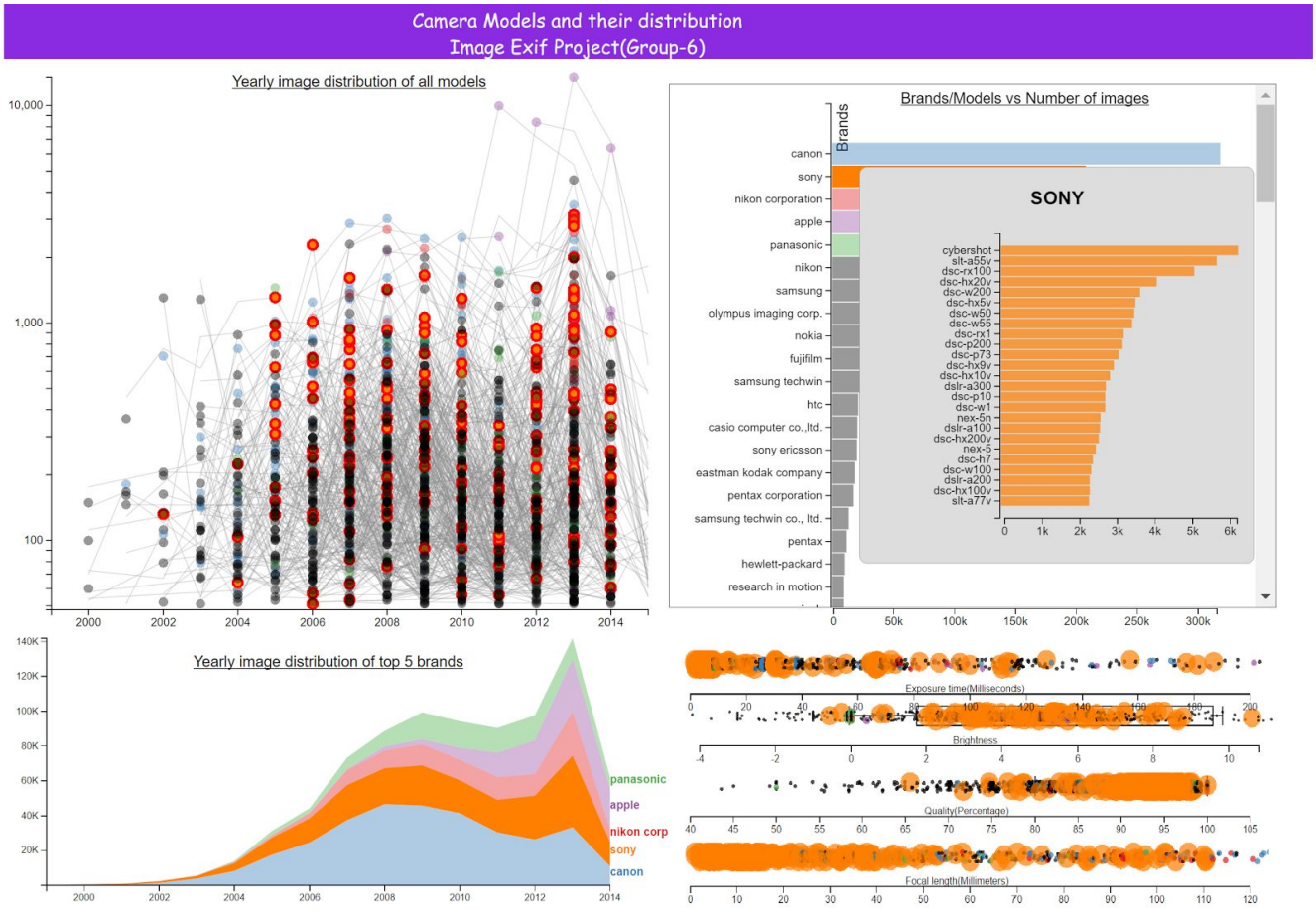


We added a stacked area chart as figure 3 which shows the yearly distribution of top 5 brands overall as was suggested by Josua, our mentor.



Final mockup

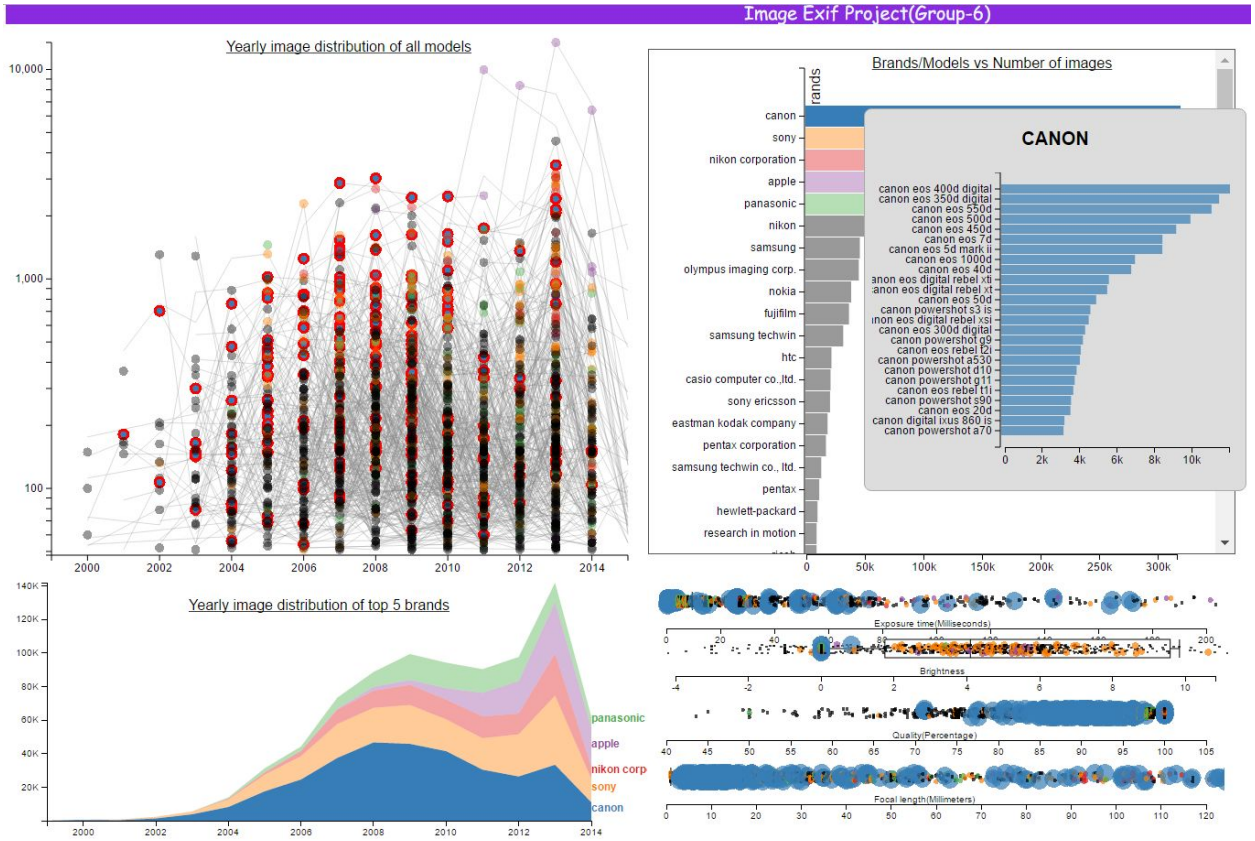
After fourth modification, we finalized whole figure and their interactions. The final interface became as follows:



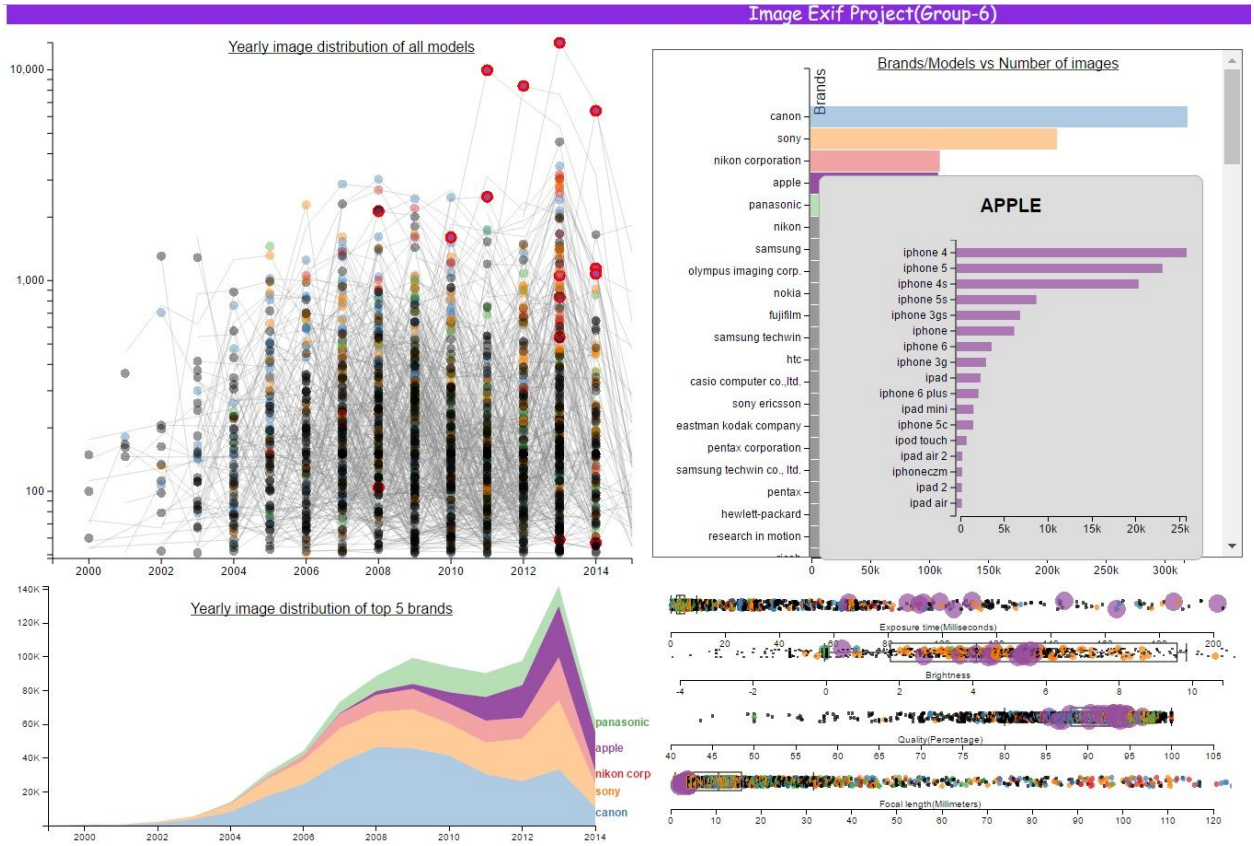
Data Analysis Results and Findings

Since our goal was to visualize the trends on camera usage and preferred settings, we decided to represent those with 4 figures each of which is visualizing one or multiple different aspects of the whole problem. Some of the discoveries we have:

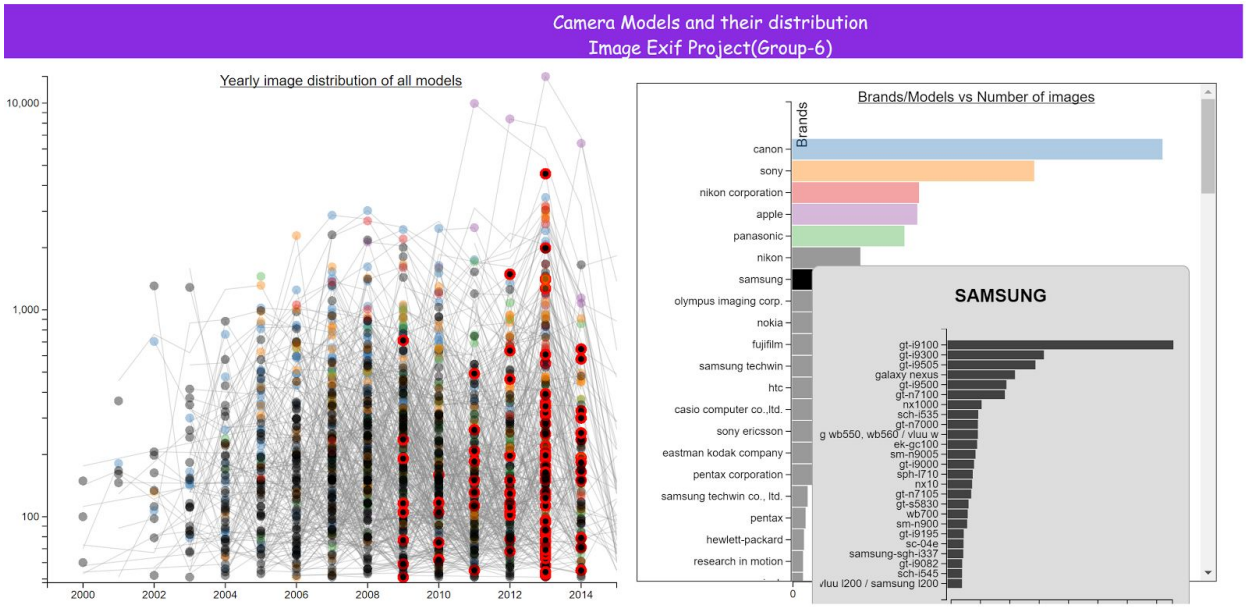
- Canon and Sony are dominating the market until now with over 200 models each. images taken by Canon and Sony cameras are approx. 330K and 220K, respectively.



- Apple is the fastest growing brand in the market: the first apple data is seen at 2007 and it is the 4th most used brand with only 17 models in Flickr in Abu Dhabi. Figure xxx show that the top 4 models which took most images are all Apple.



- Since there is not enough data for 2014 and 2015, there seems to be a decline in all cameras but it is not because the market is getting smaller, it is because there are less data for those two years.
- For exposure time we can see that majority of the values lie between 0 to 1 milliseconds. Similarly, for brightness, the values lie between 2 to 6. The quality of the majority of the images was between 85 to 100. Also, focal length was between 0 to 30 mm for most of the images.
- Observing the trend of top brands, Apple seemed to have higher exposure time values than the average values. Panasonic and Nikon Corporation had most of the values below the average brightness value of all the images. All the rest camera settings for top brands were near to the average values.
- Samsung is not as popular as Apple in Abu Dhabi, but the 5th most popular model is Samsung S3 which comes right after 4 Apple models.



Future Work and limitation

Our last step is to add selection option to the whole interface. When a user clicks on one of the bars in Figure 2 or Figure 3, that selection will remain until s/he unselects it. However, due to the time limit, we couldn't be able to finish that part. Also, since the code is written by 3 people, it is harder to manipulate all the interactions, therefore, it was not possible to add that capability on time.

Conclusion

Understanding the trends in cameras and images is a highly valuable issue that can help us keep ahead of the technology. Through this understanding, researchers and camera companies can know more about the users' expectations and needs, therefore, help technology enhance accordingly. For this project, we collected public pictures of Flickr users over the years until 2014, therefore, we can see the trends of image settings and trends of camera technologies.

This visualization would be beneficial for the camera companies in understanding how well their models are doing and how their competitors are doing. It will also help experts and researchers in studying the current demand and trends for camera brands, models and settings. The initial mockup has evolved many times and got its final version