Happy Little Data Project

Sandra “Liz” Barnes, Dan Parilla, and Derek Rouch

LLO 8200: Introduction to Data Science

Vanderbilt University

Happy Little Data Project

Robert Norman Ross, the famous American painter, is perhaps as well known for his soothing tones, happy little accidents, and signature perm as he was for the peaceful landscapes he painted for television audiences across the globe. His show, *The Joy of Painting*, enjoyed a 31-season run on the Public Broadcast System from 1983 to 1994[1]. Over those 11 years, there were 403 paintings of snow-capped mountains, thick forests, and furry friends. Each episode started with a blank canvas which, over the course of half an hour, was brought to life as the fabric met Ross's brush.

Bob Ross by the Numbers

Ross's oil paintings shared similar formats, with subtle stylistic changes occuring over the years. Our team was curious as to whether Ross's style could be quantified. Although no forgotten Ross work has been theorized, we wondered whether statistical analysis could be employed to predict the composition of a final (and to-date hypothetical) "lost work."

Our goal was to use data science to compare elements within the paintings, accounting for changes over time, to predict what elements would be included in Ross' lost masterpiece. The analysis provided in this study uses clustering techniques of elements to enable us to predict the last painting.

[1] Ross, R.N. (1983–1994). The Joy of Painting. Public Broadcast System.

About the Data

This study is an extension of a FiveThirtyEight article titled *A Statistical Analysis of the Work of Bob Ross*[2]. For academic transparency, FiveThirtyEight publishes all of the data files that support their articles in an online GitHub repository. The data used for this study was pulled from a comma-separated file hosted within that repository[3]. Much of the Bob's work becomes an exercise in conditional probability.

403 individual episodes, denoted by standard television season-and-episode syntax (e.g. S01E01), made up the individual observations. Aside from the text contained within the file's first two columns (EPISODE and TITLE, respectively), the majority of data was in binary format. That is, if a painting included the specific element (e.g. CABIN, FOG, MOUNTAIN, etc.), a "1" was placed in that column; if an element was not featured, a "0" was placed in that column. Figure 1 illustrates the first few rows and columns of data from the original comma-separated file.

Figure 1
Snapshot of the FiveThirtyEight .csv file

EPISODE	TITLE	APPLE_FRAM	AURORA_BO	BARN	BEACH	BOAT	BRIDGE	BUILDING	BUSHES	CABIN	CACTUS
S01E01	"A WALK IN	0	0	0	0	0	0	0	1	0	0
S01E02	"MT. MCKIN	0	0	0	0	0	0	0	0	1	0
S01E03	"EBONY SUN	0	0	0	0	0	0	0	0	1	0
S01E04	"WINTER MI	0	0	0	0	0	0	0	1	0	0
S01E05	"QUIET STRE	0	0	0	0	0	0	0	0	0	0
S01E06	"WINTER MC	0	0	0	0	0	0	0	0	1	0
S01E07	"AUTUMN M	0	0	0	0	0	0	0	0	0	0
S01E08	"PEACEFUL V	0	0	0	0	0	0	0	1	0	0
S01E09	"SEASCAPE"	0	0	0	1	0	0	0	0	0	0
S01E10	"MOUNTAIN	0	0	0	0	0	0	0	1	0	0
S01E11	"WINTER GL	0	0	0	0	0	0	0	0	0	0
S01E12	"SNOWFALL'	0	0	0	0	0	0	0	0	0	0
S01E13	"FINAL REFLE	0	0	0	0	0	0	0	1	0	0
S02E01	"MEADOW L	0	0	0	0	0	0	0	1	0	0
S02E02	"WINTER SU	0	0	0	0	0	0	0	0	0	0
S02E03	"EBONY SEA'	0	0	0	0	0	0	0	0	0	0
S02E04	"SHADES OF	0	0	0	0	0	0	0	1	1	0
S02E05	"AUTUMN S	0	0	0	0	0	0	0	1	0	0

[2] Hickey, W. (2014). A statistical analysis of the work of Bob Ross. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/
[3] https://github.com/fivethirtyeight/data/tree/master/bob-ross

While most of the data referred to the natural and human-made elements that were featured within the paintings, other data spoke to features of the paintings themselves. For example, the dataset had various columns to identify the "frame" for each painting, where appropriate. Two examples of Ross's frames are presented below in Figure 2.

Figure 2
Examples of Ross's frames

Oval frame

Apple frame

In total, 69 different variables—including episode and title—constituted the elements that FiveThirtyEight identified for their analysis. While all of the collected data was originally encoded qualitatively, our analysis sought to make quantitative predictions about Ross's hypothetical lost work.

Exploratory Data Analysis

Our preliminary analysis consisted of reviewing the statistical prevalence of various landscape elements within Bob Ross's collection of televised paintings, before extending the analysis to account for co-occurring elements within the paintings to predict Ross's final painting.

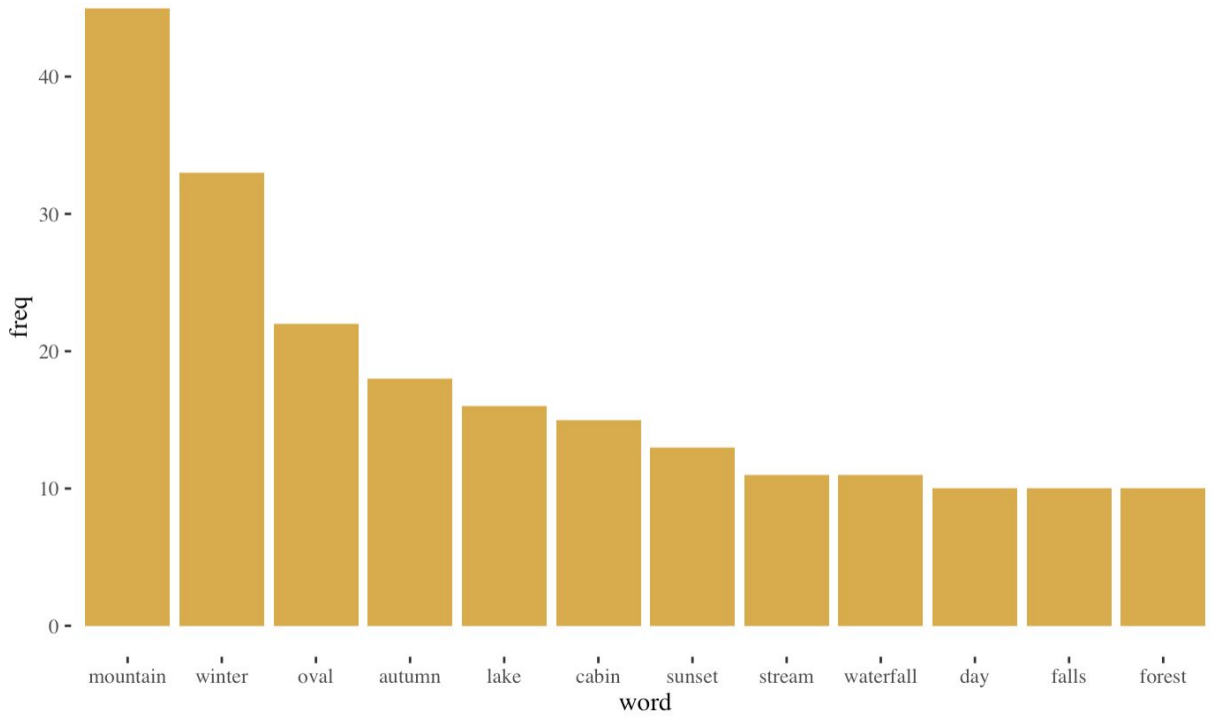
Searching for insight within the titles of Ross's paintings felt like a good place to start, so we parsed the titles into their individual words and created a word cloud from the resulting tokens. Ross's prolificness made it difficult to fit onto one visual, so we created a second word cloud that included only those words that were used in titles at least twice. Figure 3 contains the final result.

Figure 3
Word cloud from the titles of Ross's paintings (min. freq. = 2)

While some conclusions were immediately obvious, such as Ross's affinity for mountains, wintertime, and his propensity to frame his paintings within an oval, it was difficult to draw a conclusive line to separate elements within the next tier of usage. To simplify matters, we

visualized the titles another way. Figure 4 contains a bar plot of the frequency of usage for the top twelve terms within Ross's titles.

Figure 4
Top 12 words used within The Joy of Painting titles



Although the word cloud and bar plot provided a clue as to what Ross might *title* his final painting (something with "mountains" and "winter" is as good a bet as any), it failed to provide much insight about the *composition* of the work. To get us closer to an answer to our initial question, we felt five sub-questions needed to be answered:

1. What was Ross's most used frame?
2. How many different elements did he use on average?
3. What was the most popular element?

4. For the most popular element, what other elements tend to cluster with it?
5. Does this change over time?

In our quest to understand the progression and implementation of elements within Ross' paintings, we set out to tackle the above questions in order.

Sub-question 1: What was Ross's most-used frame? To find Ross's most-used frame, we selected those variables that contained the string "_FRAME," grouped them by element, and summarized their total usage. Figure 5 contains the results.

Figure 5
Frame popularity

element <chr>	n <int>
OVAL_FRAME	38
CIRCLE_FRAME	2
APPLE_FRAME	1
DOUBLE_OVAL_FRAME	1
FLORIDA_FRAME	1
HALF_CIRCLE_FRAME	1
HALF_OVAL_FRAME	1
RECTANGLE_3D_FRAME	1
RECTANGULAR_FRAME	1
SEASHELL_FRAME	1
SPLIT_FRAME	1
TOMB_FRAME	1
TRIPLE_FRAME	1
WINDOW_FRAME	1
WOOD_FRAMED	1

15 rows

In total, 53 of Ross's paintings utilized a frame. There were 15 different frames used in all, and 72% of all frames employed were full ovals. While oval frames reigned supreme, these 53 instances only accounted for 13% of Ross's entire *Joy of Painting* library. Somewhat surprisingly, the answer to "What was Ross's most used frame?" seemed to be *No frame at all*.

Sub-question 2: How many different elements did Ross use, on average? To determine how many different elements a typical Ross scene contained, we added a new variable to the dataset, TOTAL_ELEMENTS. This new column was calculated by adding all of the elements from the original dataset—with the exception of episode and title, which included summing columns 3 through 69. The binary nature of the data simplified the arithmetic, and R made easy work of the creation of our 70th variable.

We then took the mean of TOTAL_ELEMENTS and concluded that *the average Ross painting incorporated 7.99 different elements.*

Sub-question 3: What element was most popular? We now knew that Ross used eight elements on average. But even a man of Ross's omnibenevolence plays favorites, the happiest trees and puffiest clouds were not created equal, so what were Ross's elements of choice?

To find our answer, we applied a function to sum all the instances where each element was equal to one. Concerning ourselves only with those elements of highest use, we sorted the counts in descending order and listed the top 12, shown in Figure 6.

Figure 6
Element popularity

	.id <chr>	V1 <int>
61	TREE	361
62	TREES	337
19	DECIDUOUS	227
17	CONIFER	212
16	CLOUDS	179
40	MOUNTAIN	160
35	LAKE	143
30	GRASS	142
51	RIVER	126
10	BUSHES	120
55	SNOWY_MOUNTAIN	109
41	MOUNTAINS	99

12 rows

However, as the table illustrates, we encountered a problem. FiveThirtyEight, the giants on whose shoulders we were standing, encoded the elements with a degree of specificity we had not anticipated. It was obvious that trees were the *pièces de résistance* of Ross' work, but four of the most used elements were variants of trees. A closer examination of the data revealed what every *Joy of Painting* viewer has long known: Bob Ross does not create lonely trees. In fact, only 24 episodes throughout his 31-season career featured solitary trees. His other 337 episodes made

sure the trees had company. Put another way, 84% of Ross's paintings featured a forest of multiple trees.

While he switched back and forth between leafy friends and piney ones, deciduous trees held a slight edge. Table 1 illustrates the breakdown of Ross's tree usage.

Table 1
Tree usage

No trees	Deciduous only	Conifer only	Deciduous & Conifer
42 (10.4%)	145 (36.0%)	130 (32.3%)	82 (20.3%)

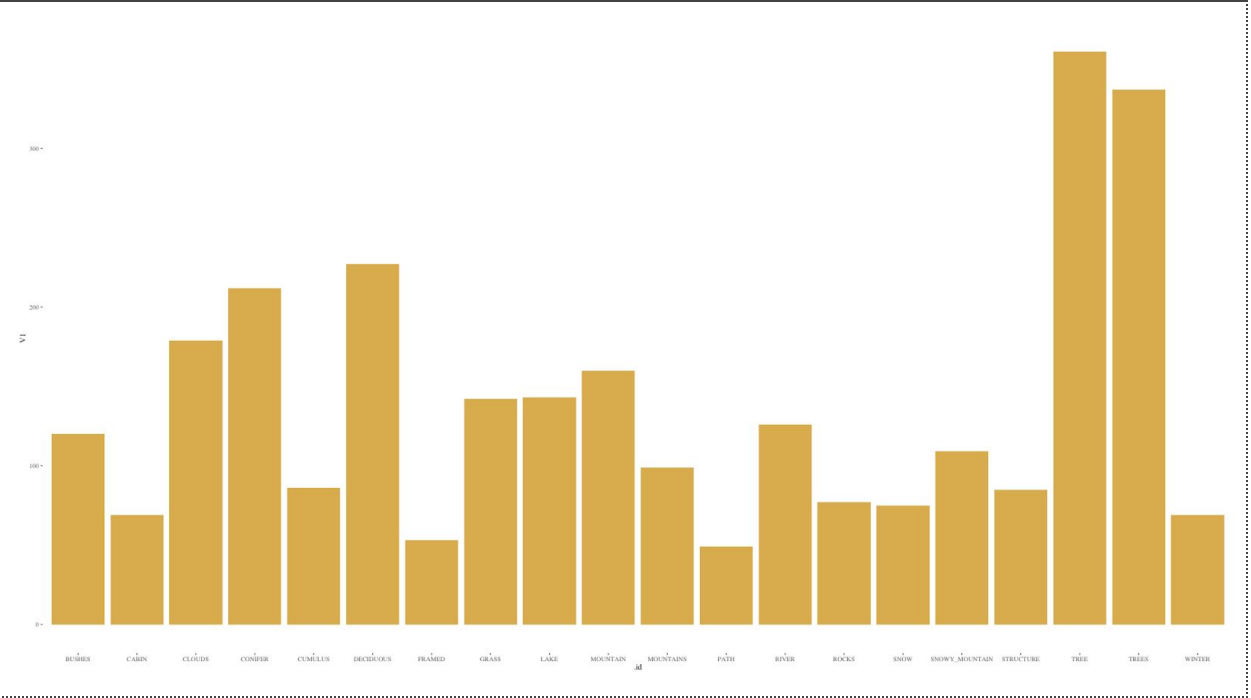
When Ross painted a tree, he painted them in bunches and typically stuck with one type. Though it felt a little unfair to exclude any group of trees from a hypothetical lost Ross masterpiece, the data led us to predicting that Ross's final work would feature *multiple deciduous trees*. Our confidence was certainly not as high in this prediction as some of the others, but this was no time to waffle.

Sub-question 4: For the most popular element, what other elements tend to cluster with it? We would have liked to run the k-means algorithm to cluster the elements that occurred frequently alongside deciduous trees, but the categorical nature of the data made little sense for a function that relies on Euclidean distance. Instead, we opted for an intuitive approach to clustering. For example, there were some variables that clearly were not worth exploring such as palm trees that were featured in only two percent of Ross's paintings, and never alongside deciduous trees. Additional potential pairings, such as the beaches, cliffs, and cacti felt similarly

fruitless. Fortunately, Ross was an artist of habit and many of the other elements in the top 12 felt very natural alongside their leaf-shedding friends.

To confirm our intuition, we reviewed the data from the previous questions and hypothesized what elements would be included for Ross' final painting. Wanting to identify eight elements in total (his mean per painting), we searched for seven more to accompany deciduous trees. We stored the data as a matrix and ran a cross product to look for large numbers and, naturally, found a match between the output of the cross product and that of Figure 7, a plot based on the element popularity data we discovered in sub-question 3.

Figure 7
Element popularity plot



In order of popularity, Ross's eight most-used elements were: (i) deciduous trees, (ii) clouds, (iii) mountain, (iv) lake, (v) grass, (vi) river, (vii) bushes, and (viii) a structure.

Ultimately, and to no surprise, Ross's most-used structure was the solitary cabin. All of these elements followed our intuitive clustering, and thus made sense as selections for our imagined missing piece.

Sub-question 5: Does this change over time and what would Ross' final painting be?

Now that we had an answer to our first four questions, we wanted to determine if Ross's style stayed consistent over time. To accomplish this, we divided his *Joy of Painting* work into three eras, which we called EarlyRoss, MiddleRoss, and LateRoss. EarlyRoss consisted of the first ten seasons of the show, MiddleRoss the next eleven, and LateRoss the final ten. We then ran through the same process described in the previous four sub-questions to see if there was any discernible change over the course of his illustrious career. Our findings are presented below in Table 2.

Table 2
Ross through the years

Sub-question	EarlyRoss (Seasons 1–10)	MiddleRoss (Seasons 11–21)	LateRoss (Seasons 22–31)
Frame of choice	No frame (Circle and Oval when used)	No frame (Oval when used)	No frame (Oval when used)
Average # of elements per painting	8.1	8.2	7.7
Element of choice	Deciduous trees	Deciduous trees	Conifer trees
Top 8 elements	1. Deciduous trees 2. Clouds 3. Mountain 4. Lake 5. Bushes 6. Grass 7. River	1. Deciduous trees 2. Clouds 3. Mountain 4. River 5. Grass 6. Lake 7. Bushes	1. Coniferous trees 2. Lake 3. Mountain 4. Grass 5. River 6. Clouds 7. Bushes

	8. Cabin	8. Rocks	8. Snow

When Ross used a frame, he definitely favored oval ones. In general, frames seemed to emerge in the middle and latter portions of his career, but he still preferred to use no frame at all, instead painting his landscapes edge-to-edge. His paintings became slightly simpler in the later episodes, but they still rounded out to about eight per painting. The scenes also become more wintry, with snow cracking the top eight in the latter years.

Perhaps in conjunction with the weather, Ross moved to painting more coniferous trees in his later career. Did he also group trees differently? Table 3 demonstrates that, while conifers held an edge in the latter years, he still typically stuck to one type.

Table 3
Tree usage, late career

No trees	Deciduous only	Conifer only	Deciduous & Conifer
13 (11.1%)	42 (35.9%)	50 (42.7%)	25 (21.4%)

Overall, throughout the 11 years of *The Joy in Painting*, Ross's paintings did slightly change in their composition. Assuming his undiscovered work was truly his *last* painting, chronologically speaking, we can predict that Ross's final painting would include a scene with eight elements: (i) conifers, (ii) a lake, (iii) at least one mountain, (iv) grass, (v) a river, (vi) clouds, (vii), bushes, and (viii) snow.

So, does such a painting exist? Unfortunately, not. Out of all of his works during the 31-season run of *The Joy of Painting*, there's no one painting that matches all of these criteria. There are, however, some that gets close. In total, ten paintings meet six of the eight criteria. In

1993, Ross painted one such landscape. During the third episode of the show's 29th season, Ross painted *Seasonal Progression*. Oddly enough, it was a framed painting, and his only on-air execution of what FiveThirtyEight called the "split frame." But Ross was a man of many surprises, and never one who would let statistical models govern spontaneity.

Figure 8
Ross posing next to "Seasonal Progression"

Conclusion

Over the 11 years, 403 paintings, and countless tubes of paint, canvas and trowel kits, Bob Ross brought mountain and forest landscapes, snowy peaks, happy little trees, and furry friends to life through *The Joy of Painting*. Reviewing data from a FiveThirtyEight article, *A Statistical Analysis of the Work of Bob Ross*, we completed an extension of their data and

reviewed Ross's paintings by the numbers to understand how many elements were used. Through statistical and data analysis and intuitive clustering, we are able to predict that Ross's last painting would look something like this:

Seasonal Progression (1993)