BLUETHUMB CANVAS SUCCESS PROJECT

2023-10-15

Introduction

Creating a brand name in the art industry is indeed a challenge on its own. Bluethumb was a company established in 2012 with a mission to empower Australian Artists. They did not agree with the habit of having to wait for an artist to reach the level of having their own gallery or exhibition, instead started Australia's online art gallery which today represents over 20,000 emerging and established artists (Bluethumb, 2023). I, myself, having the passion for art, stumbled upon their website which was very well made indeed. But as a new artist and user myself, there were some points I realized it lacked. They have an enormous customer base and an even larger artwork portfolio. Data science could revolutionize the way Bluethumb operates, if the website data could be utilized to provide insights into what are the current market trends and where do one's artwork stands before the artist can start selling. This will enable emerging artists to increase the artwork sales potential by aligning with market trends whilst improving their own profiles to the level of those who are well established. For Bluethumb, this will not only increase their sales due to more artist engagement but also it will be a significant value add to their corporate social responsibility (CSR).

"Bluethumb Canvas Success Project" aims to provide insights on the market trends in the art industry within Australia to guide new artists to areas with high demand. Further summary insights into making better data driven decisions when it comes to sizing, texture, topic…etc. The artwork data and artists profile information coupled with these insights will be incorporated into developing an art growth score model to indicate an artwork's potential to sell. The model initially will not take into account the image of the artwork itself but rather other variables that influence a buyer's decision. For example, art style, topic, size, price, frame, artist popularity, follower, count..etc. It could be later extended to include image analysis in phase 2.

The objective of a growth score model is to provide a comparison with the artists who are selling and to highlight weak areas to improve on or to make better artwork wise decisions. As per Martin et al (2020), the moment the human brain encounters a mismatch between the goal and capacity it initiates a learning process. This development will provide new artists with a guide they can refer to in order to improve and focus their efforts into achieving the end goal of selling and becoming a more established artist. Whilst for Blue thumber it will drive more artist engagement which converts to better sales and increased CSR for the brand name. The combinations of advanced analytics for market analysis, artist comparison, propensity to buy mapped into an artist growth journey in a user friendly platform will be a novel data science initiative that stands out in the art industry.

Goal and Objectives

The goal of this project is to enable artists to drive growth which eventually converts to increase in sales. The project is developed particularly for Bluethumb. The objectives are briefly mentioned below:

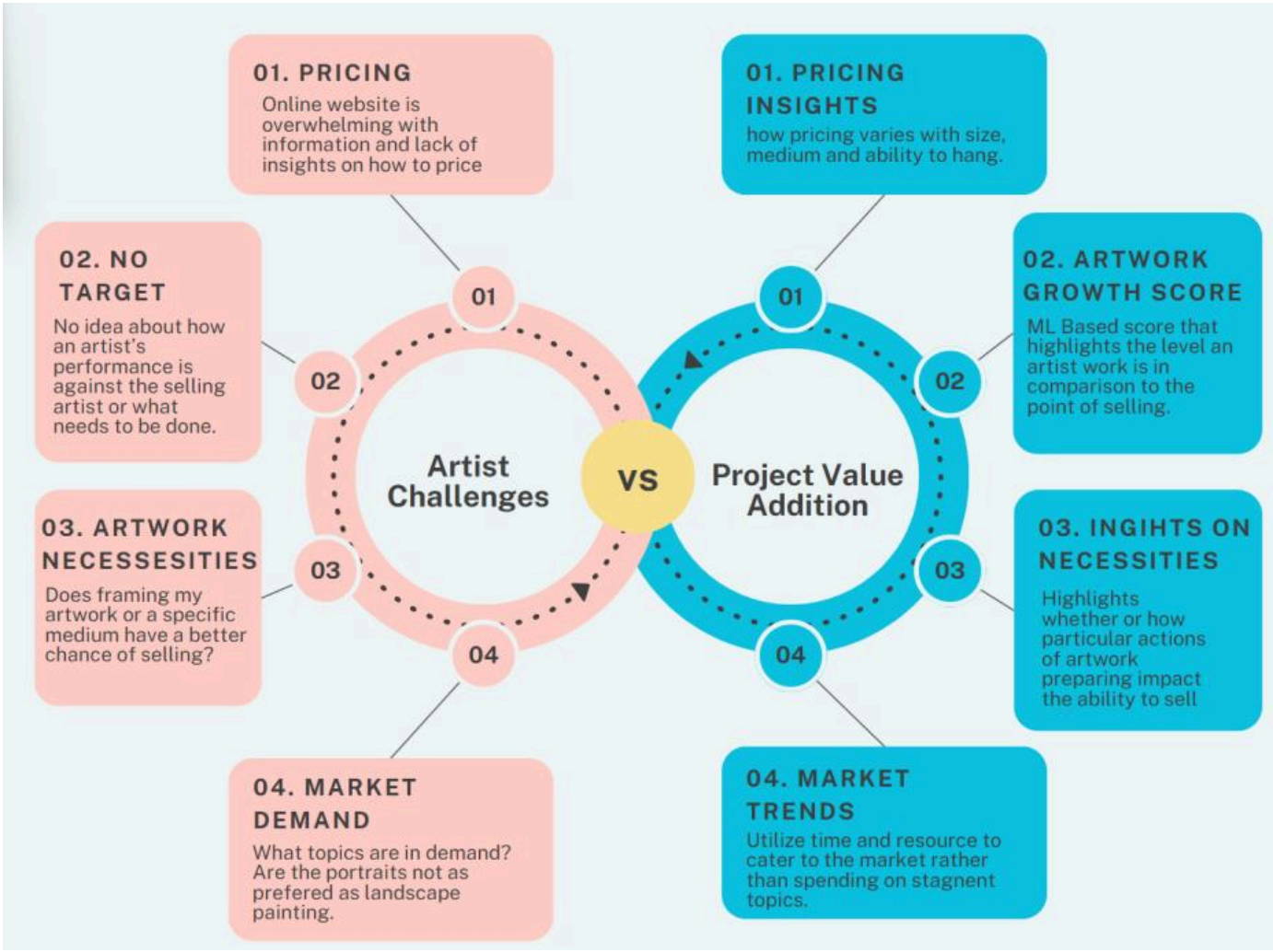
Efficient use of resources As it will enable emerging artists to focus their efforts on making creatives in areas of market interest.

Drive growth As the artwork grade would change along with the artist profile and other variables which will act as a growth indicator.

Drive Artist Interaction The more an artist can see a quantified indication of results from their efforts and not just sales, the more motivated they are to continue interacting on Bluethumb.

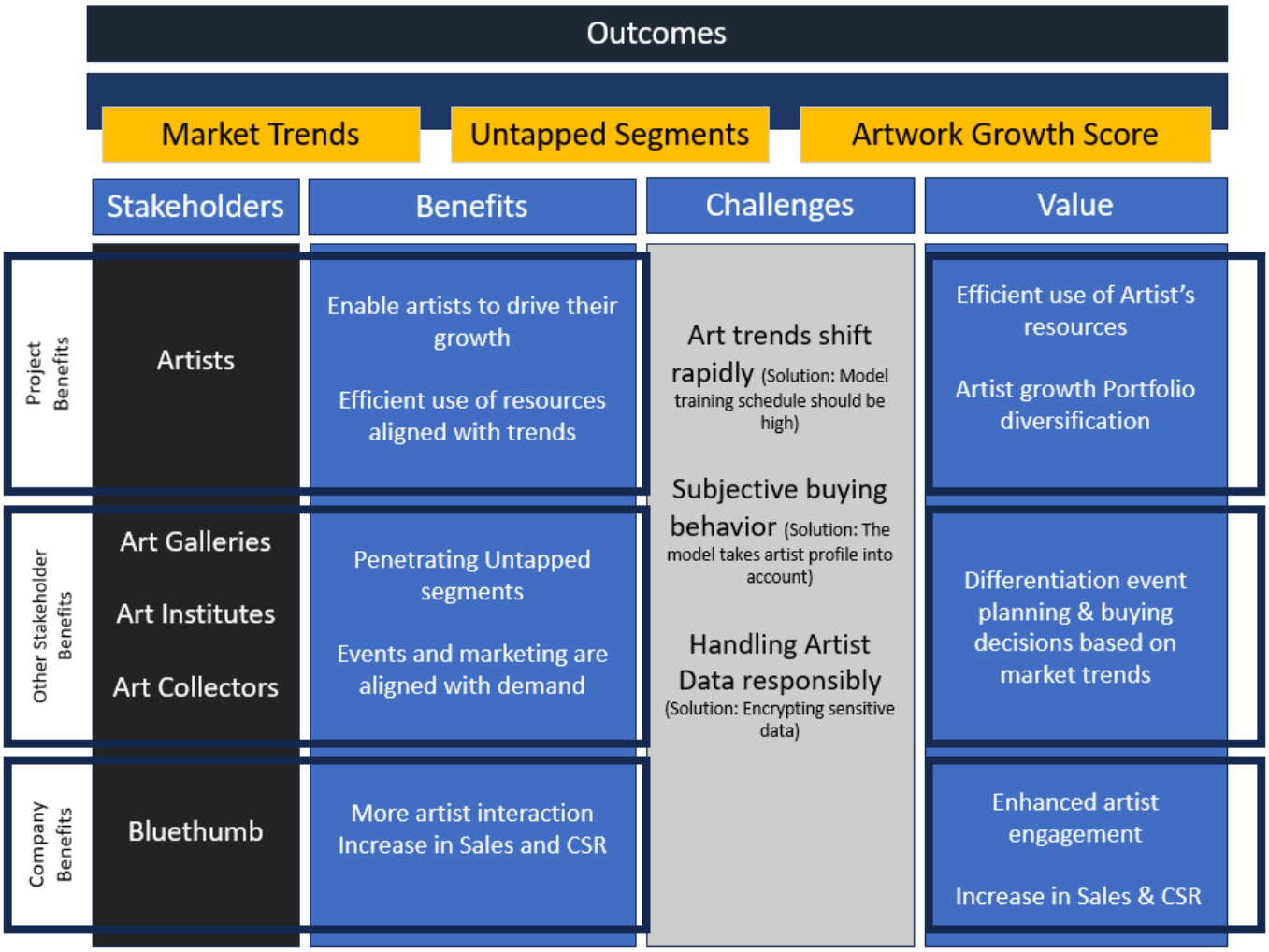
Drive sales Support emerging artists into making their first sale and much more as they grow.

As of today, there are several challenges artists utilizing this online platform face, the project outcome on addressing them are shown below:



Alt text

Business Model

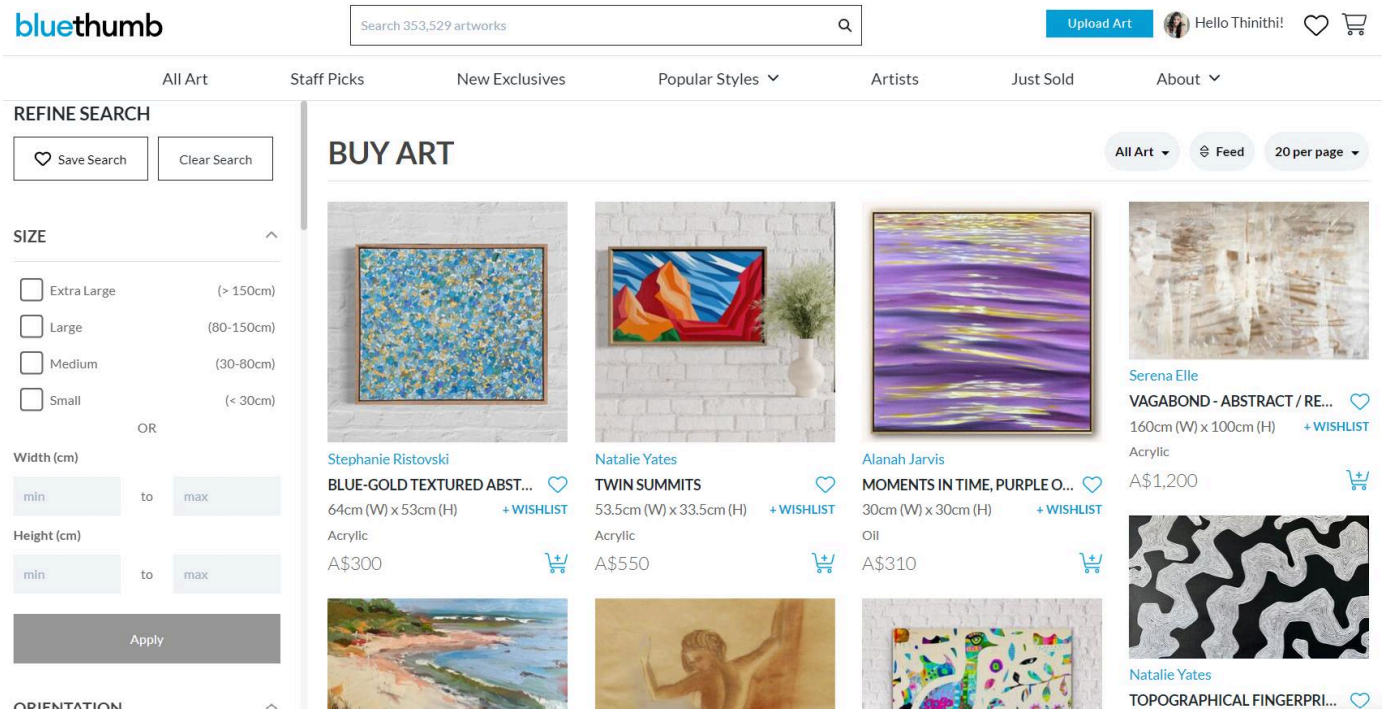


Alt text

Data Sources and Collection

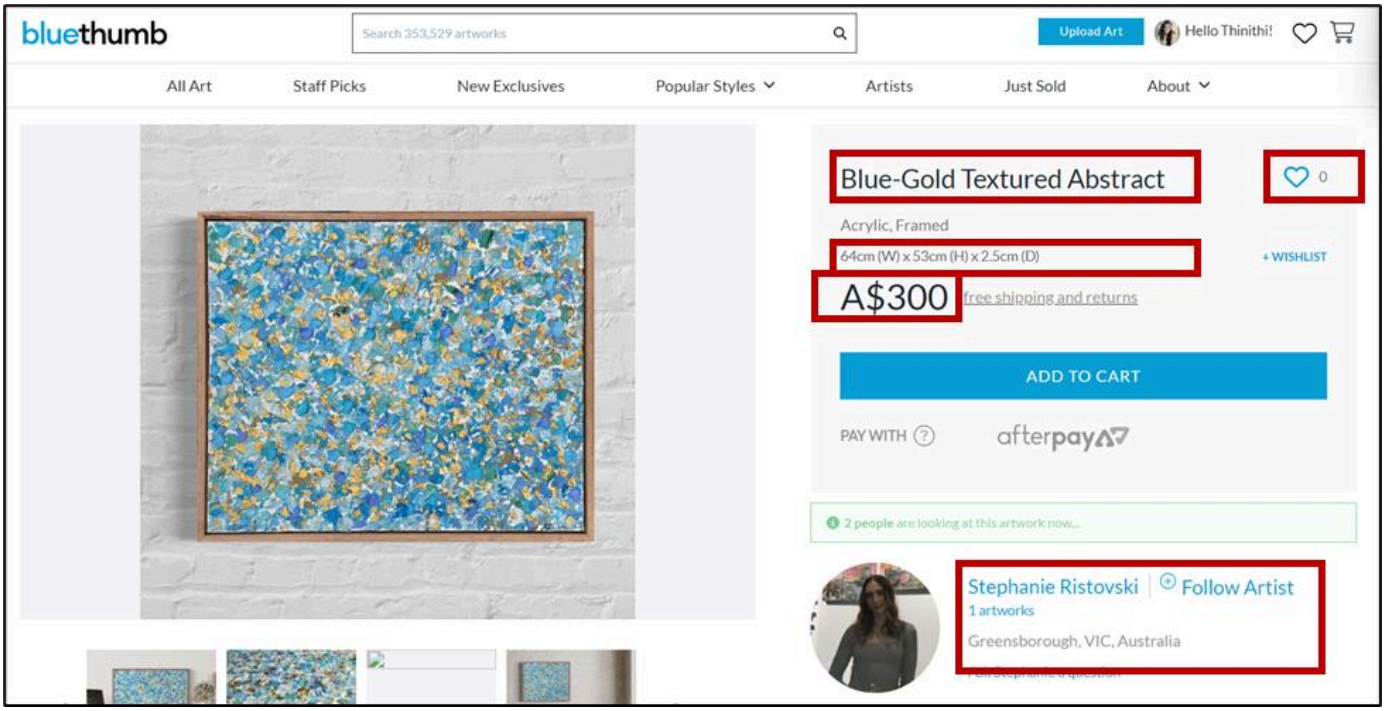
The required for this project is already available on the Bluethumb website. When developing this project within Bluethumb it is a matter of tapping into their own website data utilizing APIs which will be enabled by the data engineer, data architect and the data scientist through collaboration. In order to develop a prototype the data is obtained from the website using a web scraper known as Octoparse. The format of the feed page is shown below where all published art is posted in the order it was published. The points that have been extracted are indicated by the red marker

Arts for sale page



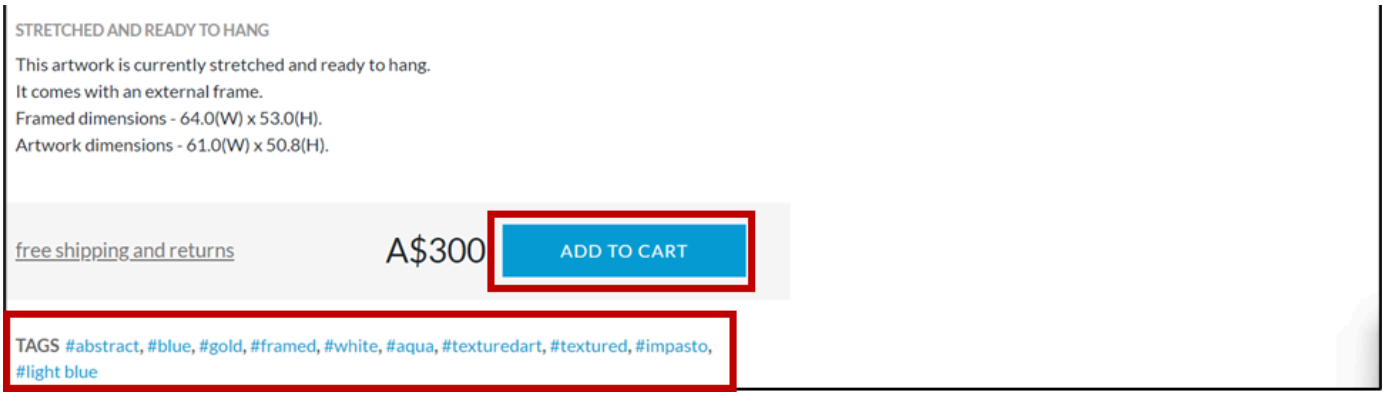
Alt text

Artwork page



Alt text

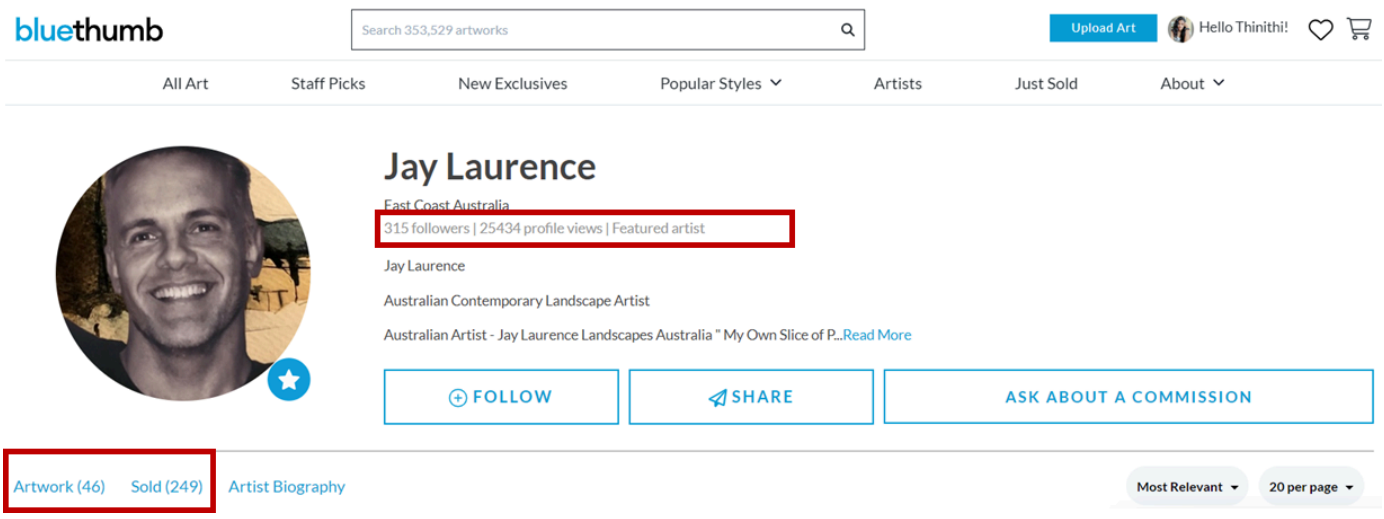
Artwork details



Alt text

Alt text

Artist profile page



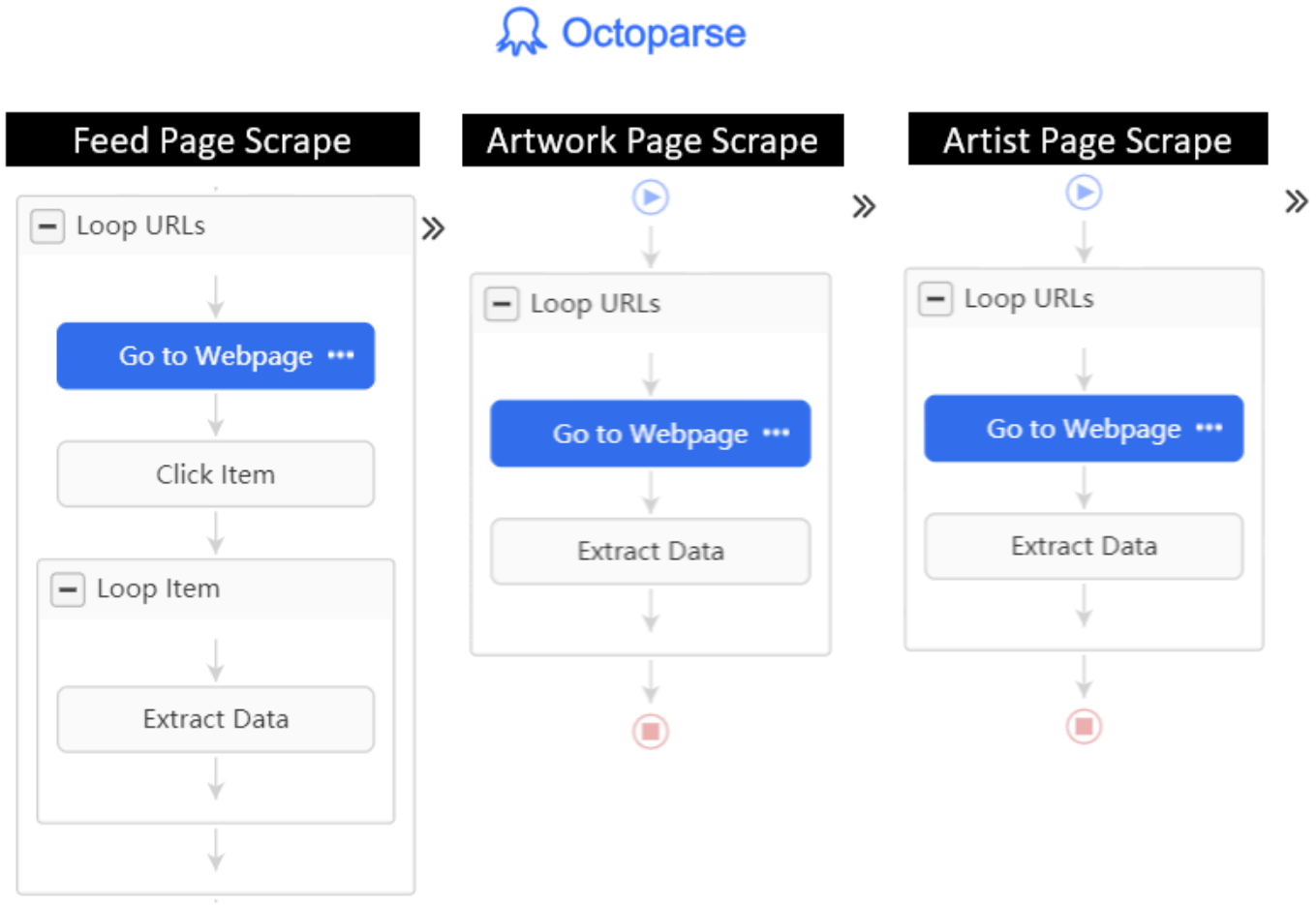
Alt text

Octoparse Webscrapper

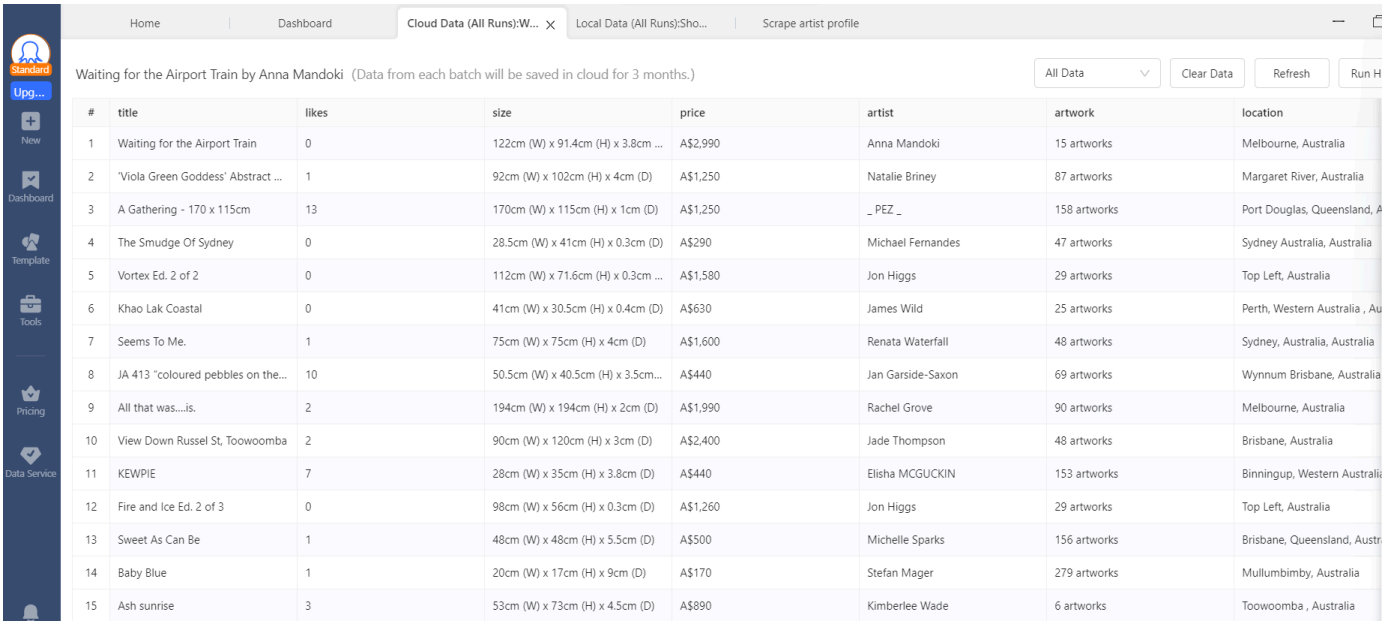
The data is extracted in three stages to speed up the process. Initially, the artwork url of all the artworks in the feed page is obtained after which, the workflow in the scraper is set to visit each artwork webpage and extract the relevant data. Finally, the url of each artwork is trimmed to obtain the artist page.

link from the extracted data and is fed into the scraper to visit each artist profile and obtain the artist data, after which it is stored in Artist_profile_data.csv. The three datasets are compiled into a single dataset named final_data.csv via excel and loaded into R Studio for data pre-processing and analysis. The octoparse

workflows for each stage are shown below.



Alt text



Home | Dashboard | Cloud Data (All Runs):W... | Local Data (All Runs):Sho... | Scrape artist profile

Waiting for the Airport Train by Anna Mandoki (Data from each batch will be saved in cloud for 3 months.)

All Data | Clear Data | Refresh

#	title	likes	size	price	artist	artwork	location
1	Waiting for the Airport Train	0	122cm (W) x 91.4cm (H) x 3.8cm ...	A$2,990	Anna Mandoki	15 artworks	Melbourne, Australia
2	'Viola Green Goddess' Abstract ...	1	92cm (W) x 102cm (H) x 4cm (D)	A$1,250	Natalie Briney	87 artworks	Margaret River, Australia
3	A Gathering - 170 x 115cm	13	170cm (W) x 115cm (H) x 1cm (D)	A$1,250	_ PEZ _	158 artworks	Port Douglas, Queensland, A
4	The Smudge Of Sydney	0	28.5cm (W) x 41cm (H) x 0.3cm (D)	A$290	Michael Fernandes	47 artworks	Sydney Australia, Australia
5	Vortex Ed. 2 of 2	0	112cm (W) x 71.6cm (H) x 0.3cm ...	A$1,580	Jon Higgs	29 artworks	Top Left, Australia
6	Khao Lak Coastal	0	41cm (W) x 30.5cm (H) x 0.4cm (D)	A$630	James Wild	25 artworks	Perth, Western Australia , Au
7	Seems To Me.	1	75cm (W) x 75cm (H) x 4cm (D)	A$1,600	Renata Waterfall	48 artworks	Sydney, Australia, Australia
8	JA 413 "coloured pebbles on the...	10	50.5cm (W) x 40.5cm (H) x 3.5cm...	A$440	Jan Garside-Saxon	69 artworks	Wynnum Brisbane, Australia
9	All that was....is.	2	194cm (W) x 194cm (H) x 2cm (D)	A$1,990	Rachel Grove	90 artworks	Melbourne, Australia
10	View Down Russel St, Toowoomba	2	90cm (W) x 120cm (H) x 3cm (D)	A$2,400	Jade Thompson	48 artworks	Brisbane, Australia
11	KEWPIE	7	28cm (W) x 35cm (H) x 3.8cm (D)	A$440	Elisha MCGUCKIN	153 artworks	Binningup, Western Australia
12	Fire and Ice Ed. 2 of 3	0	98cm (W) x 56cm (H) x 0.3cm (D)	A$1,260	Jon Higgs	29 artworks	Top Left, Australia
13	Sweet As Can Be	1	48cm (W) x 48cm (H) x 5.5cm (D)	A$500	Michelle Sparks	156 artworks	Brisbane, Queensland, Austr
14	Baby Blue	1	20cm (W) x 17cm (H) x 9cm (D)	A$170	Stefan Mager	279 artworks	Mullumbimby, Australia
15	Ash sunrise	3	53cm (W) x 73cm (H) x 4.5cm (D)	A$890	Kimberlee Wade	6 artworks	Toowoomba , Australia

Data is extarcted from the Bluethumb website and compiled into a single file named final_data.csv and loaded into R Studio for data pre-processing and analysis. File line: https://drive.google.com/file/d/1BMEo4PNLfArJddziqBzarIIscB4vpA7o/view?usp=sharing (https://drive.google.com/file/d/1BMEo4PNLfArJddziqBzarIIscB4vpA7o/view?usp=sharing)

Data collected summary

After the dataset was imported into R, the structure of the dataset was viewed via the glimpse() function . There are total of 17 columns in the dataset and 8,714 rows of unique artwork from which 253 are sold by now (Approx 3% conversion rate)

Column Name	Description
title	Artwork title
likes	Artwork likes count
size	Artwork dmiensions (Height x Width x Diameter)
price	Artwork price
artist	Artist name
artwork	Total count of artworks done by artist
location	Artist city location
medium	Artwork medium used(Ex: Acryllic, canvas, board..etc)
hand	Ready to hang, not framed..etc
sold_tag	"add_to_cart" or "sold"
sold	Binary 1-sold,0-add_to_card (Not sold)
description	Artwork category and description hashtags
page_url	Artwork page url
artist_url	Artist profile url
follow	Artist follower count
featured_artist	Artist Bluethumb recongition status
artwork_sold	Total artworks sold

Alt text

```
> glimpse(data)
Rows: 8,714
Columns: 17
$ title          <chr> "Waiting for the Airport Train", "'Viola Green Goddess' Abstract Green Pansy Flower", "A Gathering - 170 x 11…
$ likes          <dbl> 0, 1, 13, 0, 0, 0, 1, 10, 2, 2, 7, 0, 1, 3, 0, 0, 0, 4, 0, 1, 13, 7, 0, 4, 0, 2, 6, 1, 14, 8, 16, 0, 1, 0, 0,…
$ size           <chr> "122cm (W) x 91.4cm (H) x 3.8cm (D)", "92cm (W) x 102cm (H) x 4cm (D)", "170cm (W) x 115cm (H) x 1cm (D)", "2…
$ price          <chr> "A$2,990", "A$1,250", "A$1,250", "A$290", "A$1,580", "A$630", "A$1,600", "A$440", "A$1,990", "A$2,400", "A$44…
$ artist         <chr> "Anna Mandoki", "Natalie Briney", "_ PEZ _", "Michael Fernandes", "Jon Higgs", "James Wild", "Renata Waterfal…
$ artwork        <chr> "15 artworks", "87 artworks", "158 artworks", "47 artworks", "29 artworks", "25 artworks", "48 artworks", "69…
$ location       <chr> "Melbourne, Australia", "Margaret River, Australia", "Port Douglas, Queensland, Australia", "Sydney Australia…
$ medium         <chr> "Oil, acrylic, soil, bitumen and image transfer on canvas", "Acrylic paint, moulding and crackle paste", "Lar…
$ hang           <chr> "This artwork is currently stretched and ready to hang.", "This artwork is currently stretched and ready to h…
$ sold_tag       <chr> "free shipping and returns  A$2,990  Add to Cart", "free shipping and returns  A$1,250  Add to Cart", "free s…
$ sold           <chr> "add_to_cart", "add_to_cart", "add_to_cart", "add_to_cart", "add_to_cart", "add_to_cart", "add_to_cart", "add…
$ description    <chr> "POLITICAL ART, PEOPLE & PORTRAIT ART, BIRD ART\n#birds, #people, #group, #suitcase, #travel, #texture, #ligh…
$ Page_URL       <chr> "https://bluethumb.com.au/anna-mandoki/Artwork/waiting-for-the-airport-train", "https://bluethumb.com.au/nata…
$ artist_url     <chr> "https://bluethumb.com.au/anna-mandoki", "https://bluethumb.com.au/natalie-briney", "https://bluethumb.com.au…
$ follow         <chr> "9 followers | 1390 profile views", "10 followers | 1390 profile views", "11 followers | 1390 profile views",…
$ featured_artist <chr> "featured-ico.6af4c6f5.svg", "featured-ico.6af4c6f5.svg", "established-ico.205c386c.svg", "0", "photograph-ic…
$ artwork_sold   <chr> "Sold (1)", "Sold (37)", "Sold (98)", "Sold (11)", "Artist Biography", "Sold (9)", "Sold (15)", "Sold (13)", …
```

Alt text

Data cleaning steps

Cleansing Activity	Description
Column separation into sub columns	1) split "size" column into separate dimensions. 2) split "follow" into followers and profile_views 3) split "featured" column into status and image code. 4) split "location" into city location and other information. 5) split "description" into category and hashtags
Data cleaning using regex pattern identification	Use R function str_remove() with the pattern functionality where regex is used to remove unnecessary text within each column data. Ex: "A$" in price, sold(41) in artworks sold count..etc
Drop unnecessary columns	Such as columns named feature status image code, page_url, other information of city location and artist_url
Identifying and treating missing values	1) artwork_sold "NA" signifies no sales, replace with zero 2) followers "NA" signifies no followers, replace with zero 3) profile view "NA" signifies no views, replace with zero 4) status view "NA" signifies artist without special status, replace with zero
Convert "sold" column to binary	"sold" is indicated by 1 and "add_to_cart" is indicated by 0
Outlier Removal	To further analyze price and its relationship with other attribute, its extreme values are removed using "IQR Outlier Removal" approach.

Alt text

```
# Install libraries
#install.packages("readr")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("stringr")
#install.packages("visdat")
#install.packages("ggplot2")
#install.packages("wordcloud")
#install.packages("RColorBrewer")
#install.packages("wordcloud2")
#install.packages("tm")
#install.packages("tidymodels")
#install.packages("glmnet")

# Load libraries
library(readr)
library(dplyr)
library(tidyr)
library(stringr)
library(visdat)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(tm)
library(caret)
library(tidymodels)
library(glmnet)
```

Load dataset

```
## Rows: 8714 Columns: 17
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (16): title, size, price, artist, artwork, location, medium, hang, sold_...
## dbl  (1): likes
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Rows: 8,714
## Columns: 17
## $ title           <chr> "Waiting for the Airport Train", "'Viola Green Goddess…
## $ likes           <dbl> 0, 1, 13, 0, 0, 0, 1, 10, 2, 2, 7, 0, 1, 3, 0, 0, 0, 4…
## $ size            <chr> "122cm (W) x 91.4cm (H) x 3.8cm (D)", "92cm (W) x 102c…
## $ price           <chr> "A$2,990", "A$1,250", "A$1,250", "A$290", "A$1,580", "…
## $ artist          <chr> "Anna Mandoki", "Natalie Briney", "_ PEZ _", "Michael …
## $ artwork         <chr> "15 artworks", "87 artworks", "158 artworks", "47 artw…
## $ location        <chr> "Melbourne, Australia", "Margaret River, Australia", "…
## $ medium          <chr> "Oil, acrylic, soil, bitumen and image transfer on can…
## $ hang            <chr> "This artwork is currently stretched and ready to hang…
## $ sold_tag        <chr> "free shipping and returns  A$2,990  Add to Cart", "fr…
## $ sold            <chr> "add_to_cart", "add_to_cart", "add_to_cart", "add_to_c…
## $ description     <chr> "POLITICAL ART, PEOPLE & PORTRAIT ART, BIRD ART\n#bird…
## $ Page_URL        <chr> "https://bluethumb.com.au/anna-mandoki/Artwork/waiting…
## $ artist_url      <chr> "https://bluethumb.com.au/anna-mandoki", "https://blue…
## $ follow          <chr> "9 followers | 1390 profile views", "10 followers | 13…
## $ featured_artist <chr> "featured-ico.6af4c6f5.svg", "featured-ico.6af4c6f5.sv…
## $ artwork_sold    <chr> "Sold (1)", "Sold (37)", "Sold (98)", "Sold (11)", "Ar…
```

Load dataset

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 2 rows [8505,
## 8628].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 3229 rows [4, 8, 15, 19,
## 22, 23, 28, 32, 35, 41, 42, 44, 45, 47, 48, 50, 51, 53, 58, 60, ...].
```

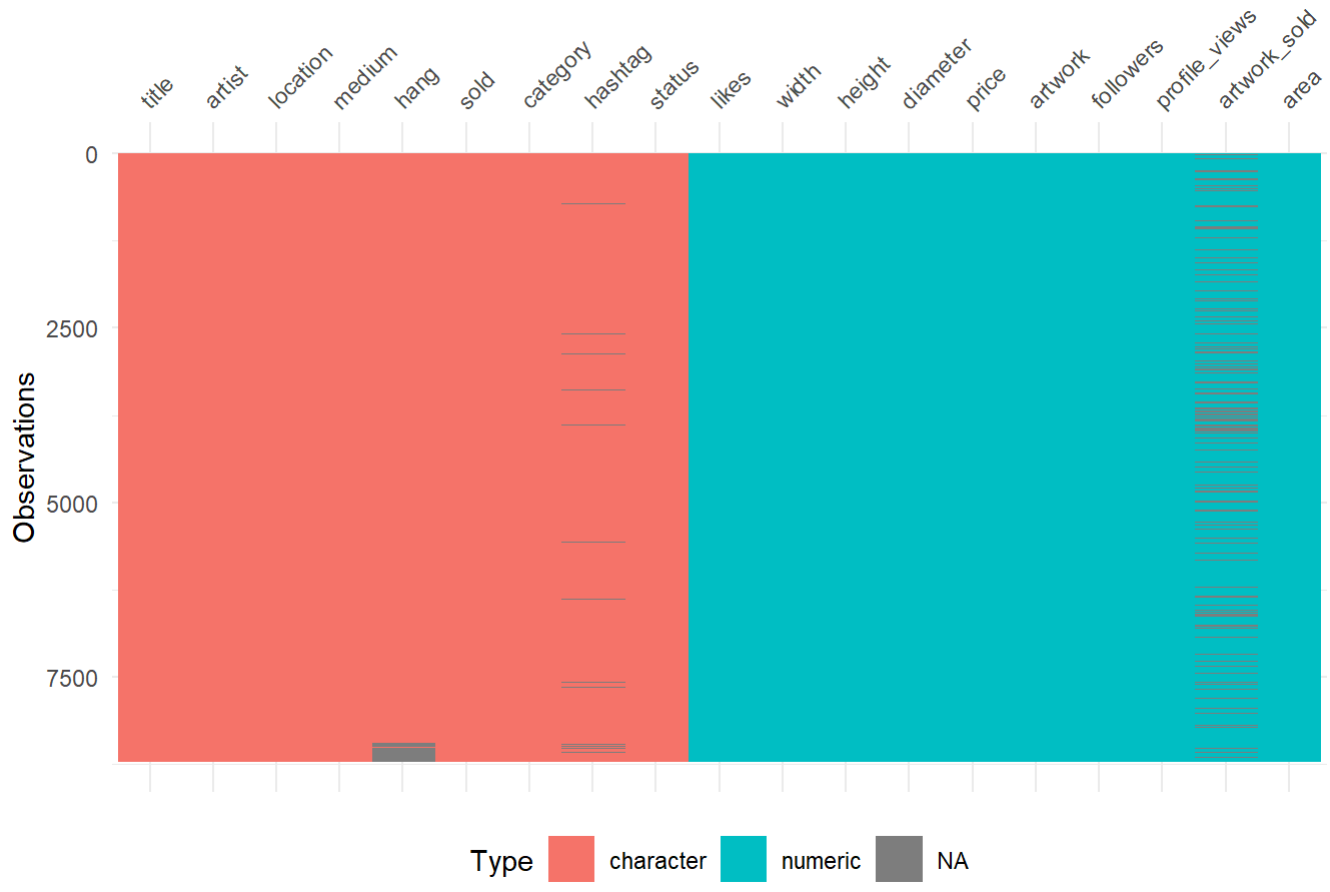
```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 303 rows [5, 12, 78, 89,
## 97, 132, 135, 179, 260, 264, 269, 335, 348, 399, 467, 492, 532, 635, 697, 717,
## ...].
```

Load dataset

```
## Warning: There were 2 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `followers = as.numeric(str_trim(str_remove(string = followers,
##   pattern = " follower")))`.
## Caused by warning:
## ! NAs introduced by coercion
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
## Rows: 8,714
## Columns: 19
## $ title         <chr> "Waiting for the Airport Train", "'Viola Green Goddess' …
## $ likes         <dbl> 0, 1, 13, 0, 0, 0, 1, 10, 2, 2, 7, 0, 1, 3, 0, 0, 0, 4, …
## $ width         <dbl> 122.0, 92.0, 170.0, 28.5, 112.0, 41.0, 75.0, 50.5, 194.0…
## $ height        <dbl> 91.4, 102.0, 115.0, 41.0, 71.6, 30.5, 75.0, 40.5, 194.0,…
## $ diameter      <dbl> 3.8, 4.0, 1.0, 0.3, 0.3, 0.4, 4.0, 3.5, 2.0, 3.0, 3.8, 0…
## $ price         <dbl> 2990, 1250, 1250, 290, 1580, 630, 1600, 440, 1990, 2400,…
## $ artist        <chr> "Anna Mandoki", "Natalie Briney", "_ PEZ _", "Michael Fe…
## $ artwork       <dbl> 15, 87, 158, 47, 29, 25, 48, 69, 90, 48, 153, 29, 156, 6…
## $ location      <chr> "Melbourne", "Margaret River", "Port Douglas", "Sydney A…
## $ medium        <chr> "Oil, acrylic, soil, bitumen and image transfer on canva…
## $ hang          <chr> "stretched and ready to hang.", "stretched and ready to …
## $ sold          <chr> "add_to_cart", "add_to_cart", "add_to_cart", "add_to_car…
## $ category      <chr> "POLITICAL ART, PEOPLE & PORTRAIT ART, BIRD ART", "NATUR…
## $ hashtag       <chr> "birds, people, group, suitcase, travel, texture, light …
## $ followers     <dbl> 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 2…
## $ profile_views <dbl> 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 13…
## $ status        <chr> "featured", "featured", "established", "0", "photograph"…
## $ artwork_sold  <dbl> 1, 37, 98, 11, NA, 9, 15, 13, 11, 46, 79, NA, 110, NA, N…
## $ area          <dbl> 11150.80, 9384.00, 19550.00, 1168.50, 8019.20, 1250.50, …
```

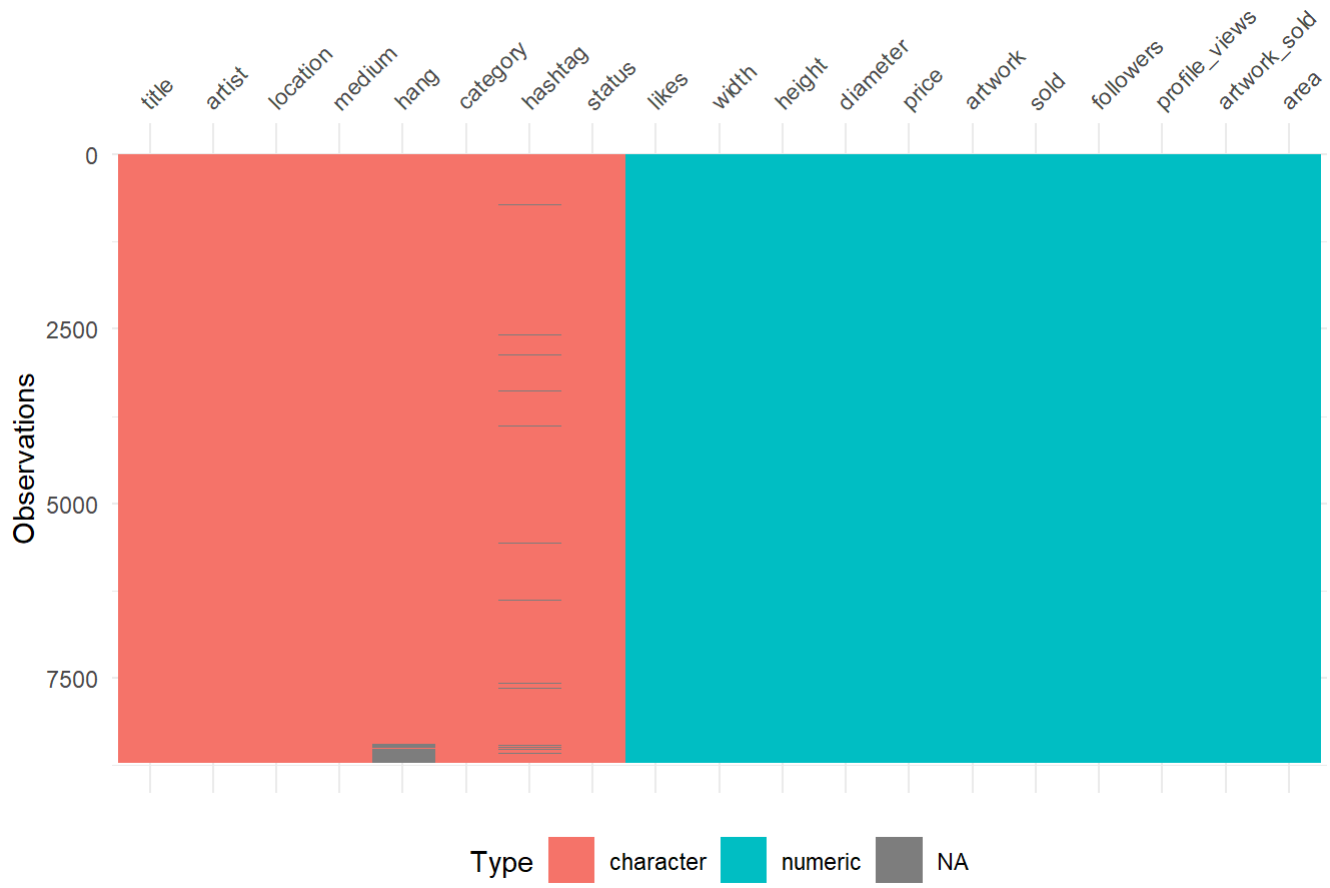
Load dataset



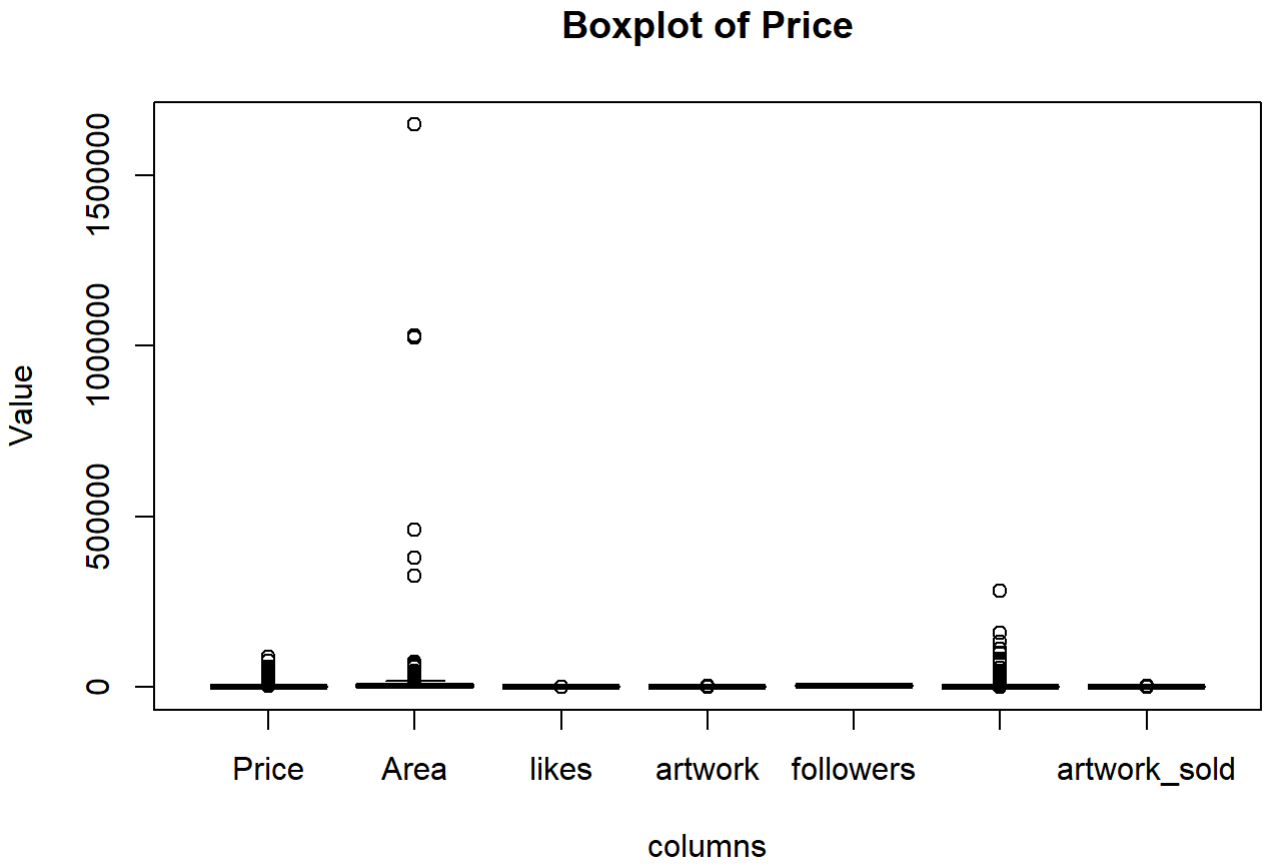
Load dataset

```
## Rows: 8,714
## Columns: 19
## $ title         <chr> "Waiting for the Airport Train", "'Viola Green Goddess' …
## $ likes         <dbl> 0, 1, 13, 0, 0, 0, 1, 10, 2, 2, 7, 0, 1, 3, 0, 0, 0, 4, …
## $ width         <dbl> 122.0, 92.0, 170.0, 28.5, 112.0, 41.0, 75.0, 50.5, 194.0…
## $ height        <dbl> 91.4, 102.0, 115.0, 41.0, 71.6, 30.5, 75.0, 40.5, 194.0,…
## $ diameter      <dbl> 3.8, 4.0, 1.0, 0.3, 0.3, 0.4, 4.0, 3.5, 2.0, 3.0, 3.8, 0…
## $ price         <dbl> 2990, 1250, 1250, 290, 1580, 630, 1600, 440, 1990, 2400,…
## $ artist        <chr> "Anna Mandoki", "Natalie Briney", "_ PEZ _", "Michael Fe…
## $ artwork       <dbl> 15, 87, 158, 47, 29, 25, 48, 69, 90, 48, 153, 29, 156, 6…
## $ location      <chr> "Melbourne", "Margaret River", "Port Douglas", "Sydney A…
## $ medium        <chr> "Oil, acrylic, soil, bitumen and image transfer on canva…
## $ hang          <chr> "stretched and ready to hang.", "stretched and ready to …
## $ sold          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ category      <chr> "POLITICAL ART, PEOPLE & PORTRAIT ART, BIRD ART", "NATUR…
## $ hashtag       <chr> "birds, people, group, suitcase, travel, texture, light …
## $ followers     <dbl> 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 2…
## $ profile_views <dbl> 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 13…
## $ status        <chr> "featured", "featured", "established", "0", "photograph"…
## $ artwork_sold  <dbl> 1, 37, 98, 11, 0, 9, 15, 13, 11, 46, 79, 0, 110, 0, 0, 0…
## $ area          <dbl> 11150.80, 9384.00, 19550.00, 1168.50, 8019.20, 1250.50, …
```

Load dataset

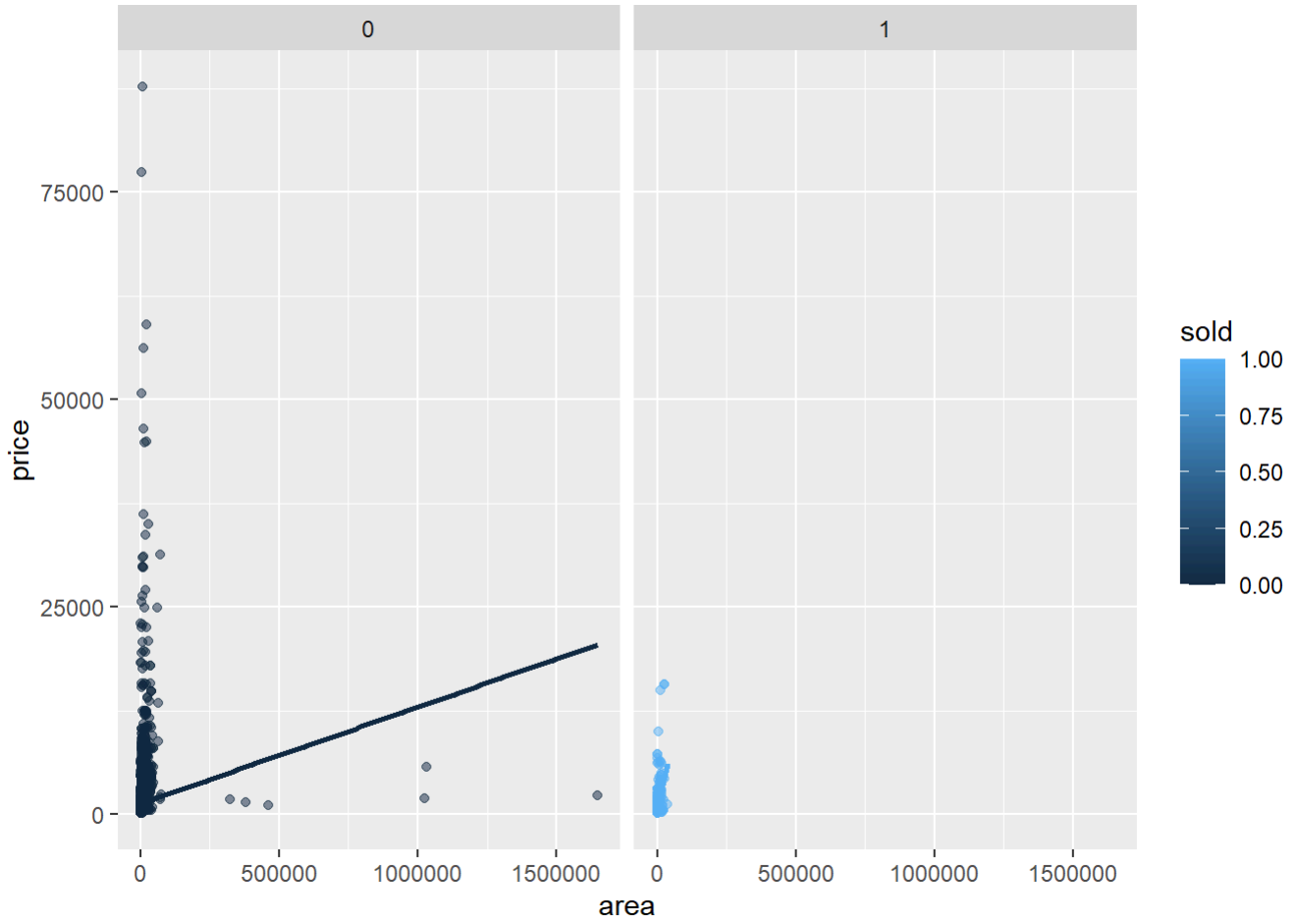


check for outliers



Before Outlier Treatment

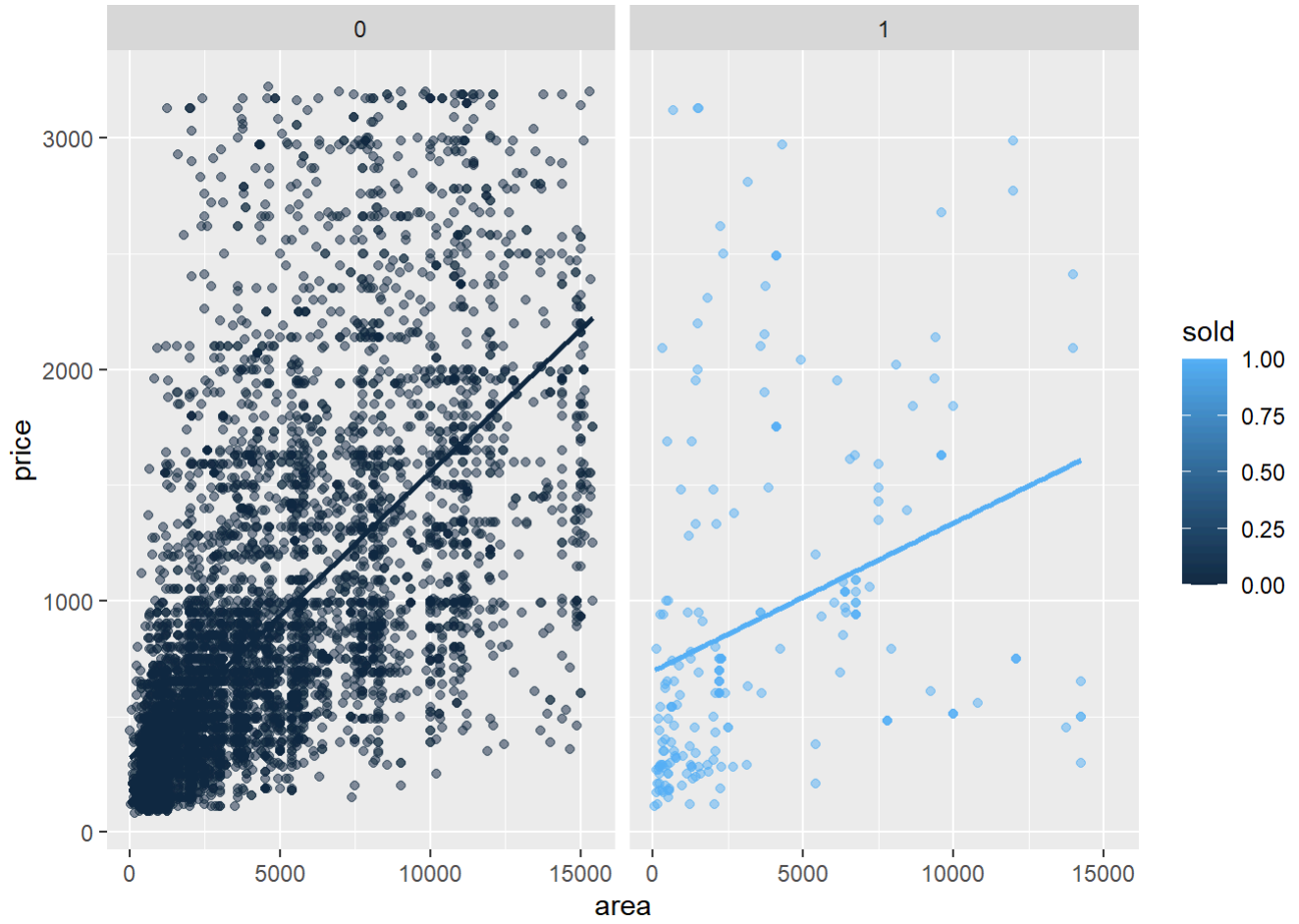
```
## `geom_smooth()` using formula = 'y ~ x'
```



Outlier Treatment

After Outlier Treatment

```
## `geom_smooth()` using formula = 'y ~ x'
```

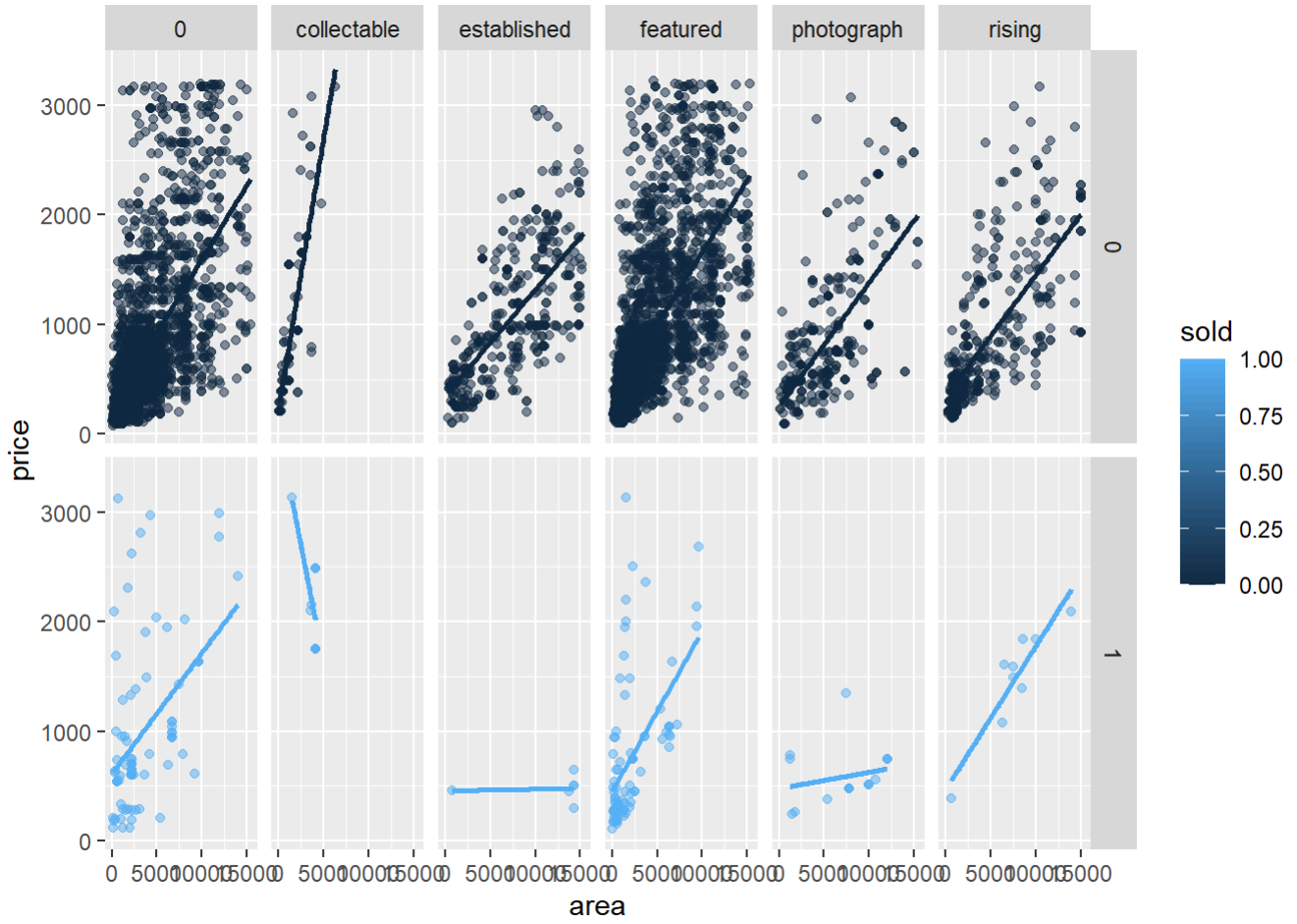


Confusion Matrix

```
##                    price       area     height       width artwork_sold
## price          1.000000000 0.64799132 0.61046742 0.602554969   0.03385989
## area           0.647991316 1.00000000 0.89376734 0.908309380   0.16601886
## height         0.610467420 0.89376734 1.00000000 0.704326149   0.11068518
## width          0.602554969 0.90830938 0.70432615 1.000000000   0.16311725
## artwork_sold   0.033859893 0.16601886 0.11068518 0.163117255   1.00000000
## followers      0.020284306 0.00519591 0.01827813 0.008518096  -0.02248001
## profile_views -0.001025278 0.05519091 0.03338166 0.048208127   0.16552747
## likes          0.077608513 0.05818019 0.05433367 0.057522276   0.11761012
##                  followers profile_views       likes
## price          0.020284306  -0.001025278  0.07760851
## area           0.005195910   0.055190910  0.05818019
## height         0.018278126   0.033381663  0.05433367
## width          0.008518096   0.048208127  0.05752228
## artwork_sold  -0.022480007   0.165527466  0.11761012
## followers      1.000000000  -0.090547033 -0.01017243
## profile_views -0.090547033   1.000000000  0.01479253
## likes         -0.010172434   0.014792530  1.00000000
```

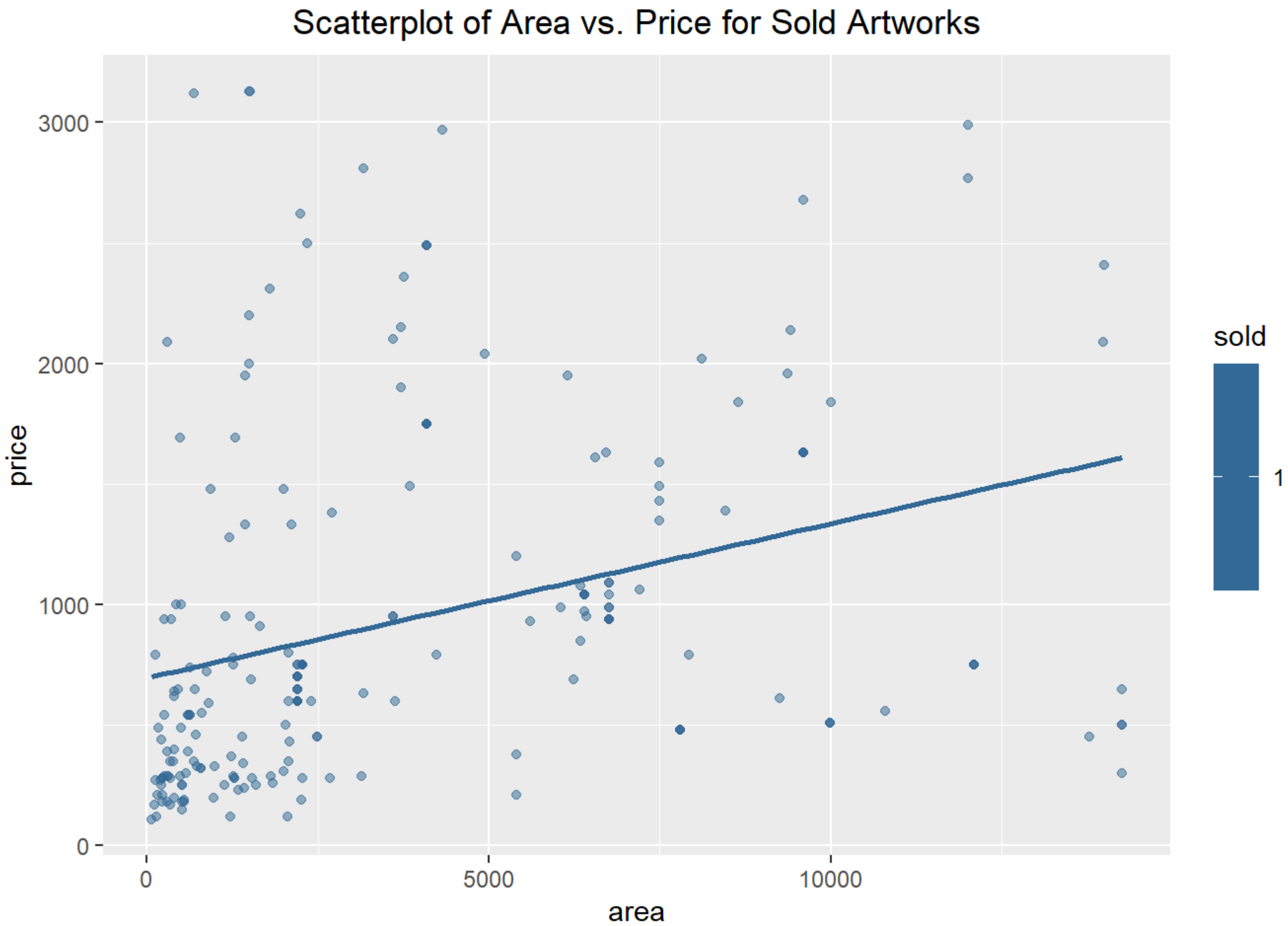
scatterplot of price vs area across different artist status

```
## `geom_smooth()` using formula = 'y ~ x'
```



Price vs Artwork Area

```
## `geom_smooth()` using formula = 'y ~ x'
```



Artwork decription word cloud - Not sold portfolio

Artwork decription word cloud - sold portfolio

Artwork topic wise - Conversion rate

```
##            word freq.x freq.y    percent
## 11         city     14    147 0.09523810
## 1        aerial     11    186 0.05913978
## 30     patterns     38    683 0.05563690
## 21    interiors     21    396 0.05303030
## 16      fashion      6    139 0.04316547
## 37        space      5    116 0.04310345
## 3  architecture     11    267 0.04119850
## 7          body      7    181 0.03867403
## 15      fantasy     16    425 0.03764706
## 44        water     15    446 0.03363229
## 31       people     33   1015 0.03251232
## 34     portrait     33   1018 0.03241650
## 36     seascape     30   1034 0.02901354
## 4         beach     30   1037 0.02892960
## 24          men      4    145 0.02758621
## 12      culture     13    483 0.02691511
## 32       places     10    406 0.02463054
## 38        still     17    697 0.02439024
## 23         life     17    700 0.02428571
## 22    landscape     56   2367 0.02365864
```

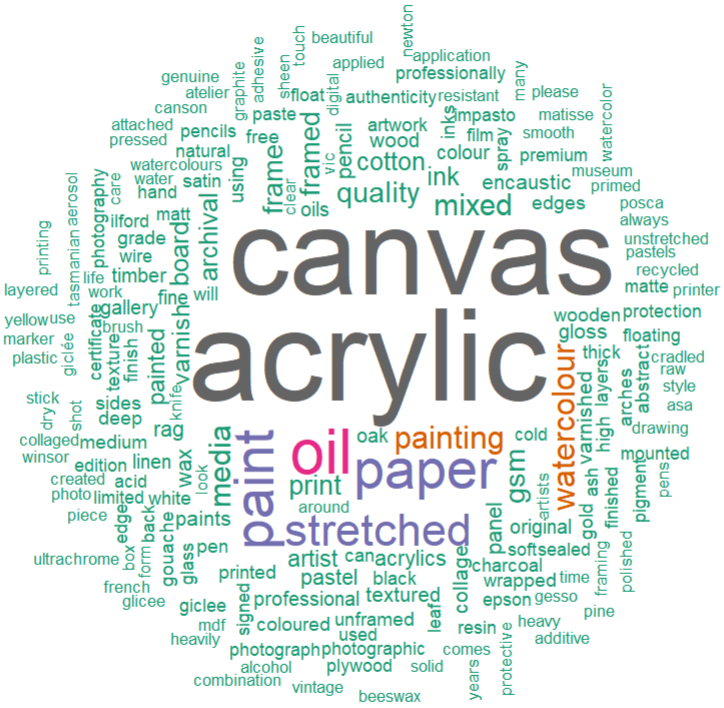
Medium word cloud- art portfolio

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("en")):
## transformation drops documents
```



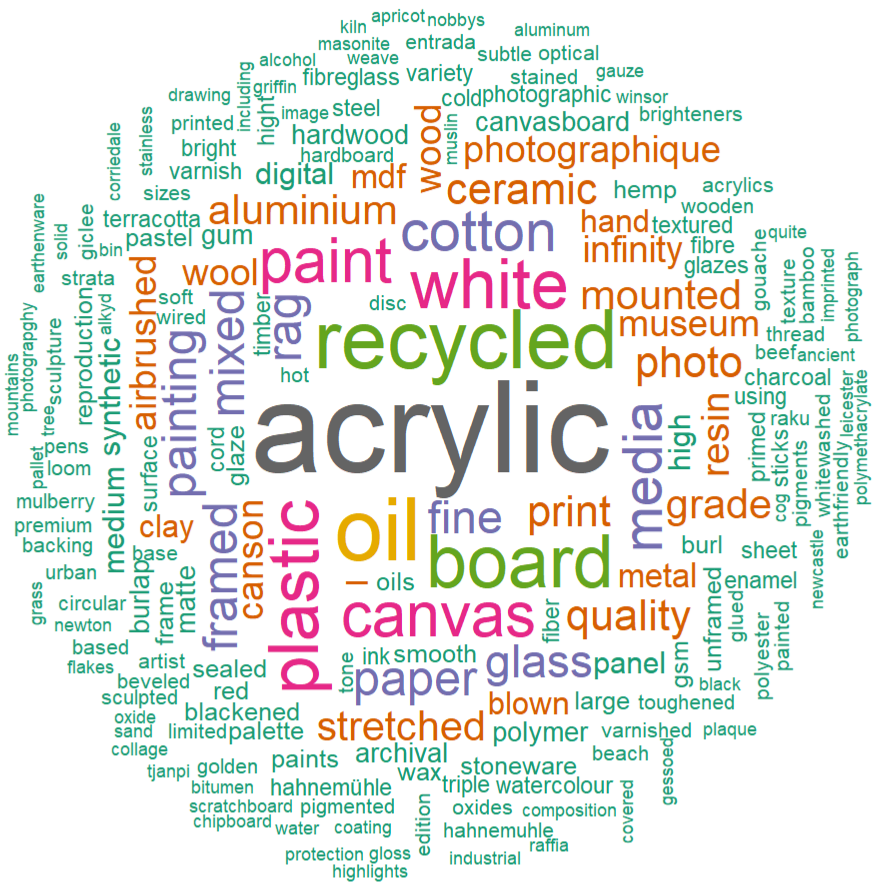
Medium word cloud- sold art portfolio

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("en")):
## transformation drops documents
```



Medium conversion rates

```
##          word freq.x freq.y    percent
## 31      board    333     25 0.07507508
## 215       rag    224     14 0.06250000
## 60     cotton    345     18 0.05217391
## 151     media    469     17 0.03624733
## 88     framed    392     14 0.03571429
## 207     print    364     12 0.03296703
## 183     panel    200      6 0.03000000
## 155     mixed    482     14 0.02904564
## 212   quality    438     12 0.02739726
## 179  painting    550     14 0.02545455
## 168       oil   1599     31 0.01938712
## 186    pastel    223      4 0.01793722
## 284       wax    223      4 0.01793722
## 177     paint   1366     20 0.01464129
## 13   archival    343      5 0.01457726
## 2     acrylic   3879     49 0.01263212
## 251 stretched   1009     11 0.01090188
## 87      frame    379      4 0.01055409
## 185     paper   1381     14 0.01013758
## 275   varnish    299      3 0.01003344
```

Scaling data between 0-1

```
## Rows: 7,637
## Columns: 8
## $ sold          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ likes         <dbl> 0.000000000, 0.002450980, 0.000000000, 0.000000000, 0.00…
## $ area          <dbl> 0.72393871, 0.60922105, 0.07579170, 0.52060509, 0.081115…
## $ price         <dbl> 0.92675159, 0.37261146, 0.06687898, 0.47770701, 0.175159…
## $ artwork       <dbl> 0.0065975495, 0.0405278040, 0.0216776626, 0.0131950990, …
## $ followers     <dbl> 0.001006711, 0.001118568, 0.001342282, 0.001454139, 0.00…
## $ profile_views <dbl> 0.01046112, 0.01046112, 0.01046112, 0.01046112, 0.010461…
## $ artwork_sold  <dbl> 0.0004852014, 0.0179524503, 0.0053372149, 0.0000000000, …
```

Logistic model training

```
## # A tibble: 8 × 3
##   term          estimate penalty
##   <chr>            <dbl>   <dbl>
## 1 (Intercept)      -2.08       0
## 2 likes            -3.11       0
## 3 area             -1.30       0
## 4 price             1.21       0
## 5 artwork          -3.08       0
## 6 followers        -5.45       0
## 7 profile_views    22.3        0
## 8 artwork_sold     -2.94       0
```

```
## # A tibble: 6 × 4
##   sold  .pred_class .pred_0 .pred_1
##   <fct> <fct>         <dbl>   <dbl>
## 1 0     0             0.871  0.129
## 2 0     0             0.907  0.0927
## 3 0     0             0.871  0.129
## 4 0     0             0.894  0.106
## 5 0     0             0.923  0.0766
## 6 0     0             0.910  0.0896
```

model performance

```
##           Truth
## Prediction    0    1
##          0 1492   25
##          1    2    9
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.982
```

```
## [1] 0.8928571
```

```
## [1] 0.5813953
```

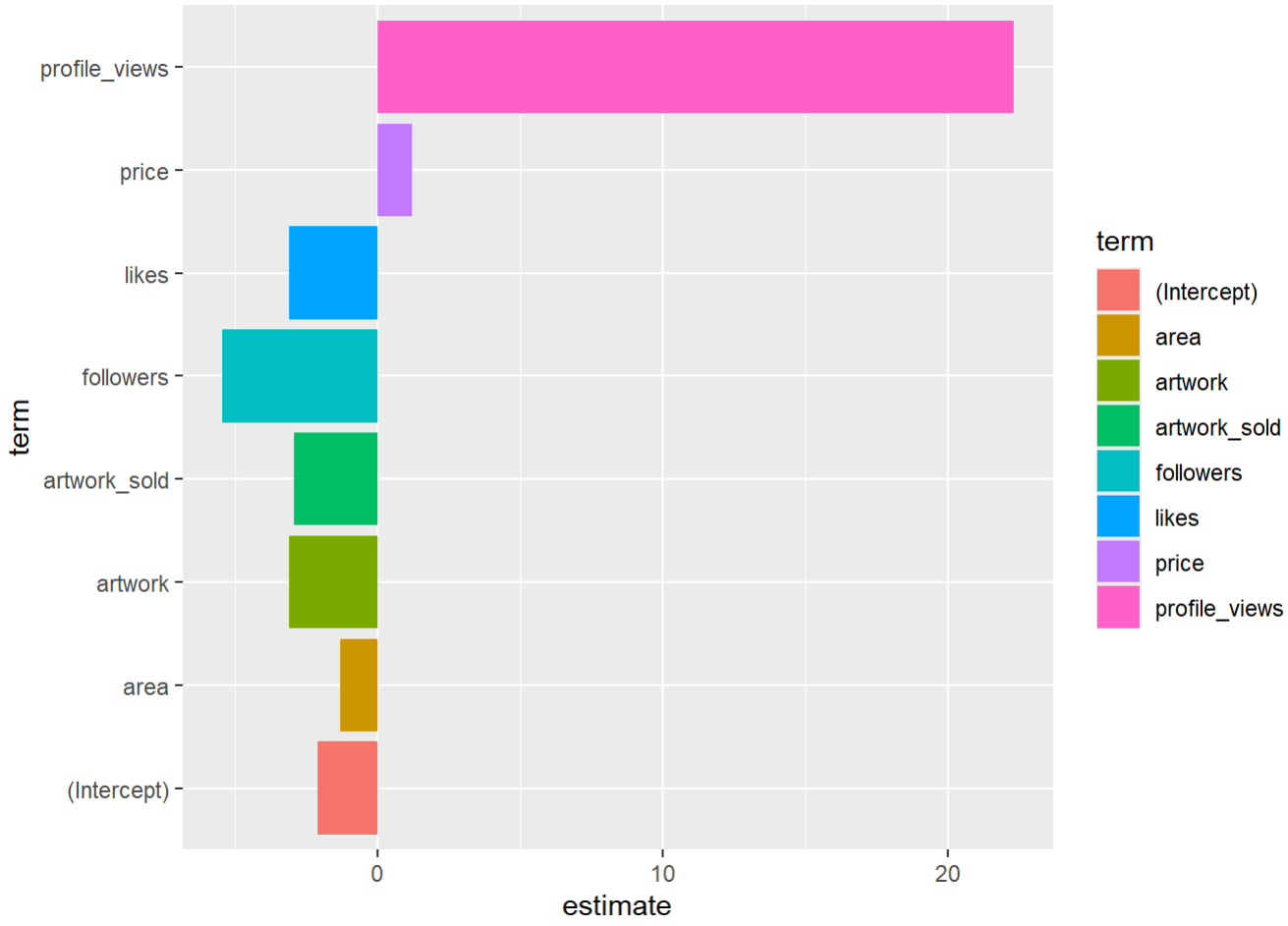
Derive probabilities for the full dataset using the model

Derive artwork growth level

```
##  [1] (-0.000999,0.1] (0.1,0.2]       (0.2,0.3]       (0.3,0.4]
##  [5] (0.4,0.5]       (0.5,0.6]       (0.6,0.7]       (0.7,0.8]
##  [9] (0.8,0.9]       (0.9,1]
## 10 Levels: (-0.000999,0.1] (0.1,0.2] (0.2,0.3] (0.3,0.4] ... (0.9,1]
```

```
## Rows: 7,637
## Columns: 20
## $ title          <chr> "Waiting for the Airport Train", "'Viola Green Goddess'…
## $ likes          <dbl> 0, 1, 0, 0, 0, 1, 10, 2, 7, 0, 1, 3, 0, 0, 0, 4, 0, 1, …
## $ width          <dbl> 122.0, 92.0, 28.5, 112.0, 41.0, 75.0, 50.5, 90.0, 28.0,…
## $ height         <dbl> 91.4, 102.0, 41.0, 71.6, 30.5, 75.0, 40.5, 120.0, 35.0,…
## $ diameter       <dbl> 3.8, 4.0, 0.3, 0.3, 0.4, 4.0, 3.5, 3.0, 3.8, 0.3, 5.5, …
## $ price          <dbl> 2990, 1250, 290, 1580, 630, 1600, 440, 2400, 440, 1260,…
## $ artist         <chr> "Anna Mandoki", "Natalie Briney", "Michael Fernandes", …
## $ artwork        <dbl> 15, 87, 47, 29, 25, 48, 69, 48, 153, 29, 156, 6, 3, 29,…
## $ location       <chr> "Melbourne", "Margaret River", "Sydney Australia", "Top…
## $ medium         <chr> "Oil, acrylic, soil, bitumen and image transfer on canv…
## $ hang           <chr> "stretched and ready to hang.", "stretched and ready to…
## $ sold           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ category       <chr> "POLITICAL ART, PEOPLE & PORTRAIT ART, BIRD ART", "NATU…
## $ hashtag        <chr> "birds, people, group, suitcase, travel, texture, light…
## $ followers      <dbl> 9, 10, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, …
## $ profile_views  <dbl> 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1390, 1…
## $ status         <chr> "featured", "featured", "0", "photograph", "featured", …
## $ artwork_sold   <dbl> 1, 37, 11, 0, 9, 15, 13, 46, 79, 0, 110, 0, 0, 0, 69, 1…
## $ area           <dbl> 11150.80, 9384.00, 1168.50, 8019.20, 1250.50, 5625.00, …
## $ artwork_growth <chr> "Level_02", "Level_01", "Level_02", "Level_02", "Level_…
```

Visualizing model coefficent importance



References:

https://stackoverflow.com/questions/70522236/combine-lapply-and-gsub-to-replace-a-list-of-values-for-another-list-of-values (https://stackoverflow.com/questions/70522236/combine-lapply-and-gsub-to-replace-a-list-of-values-for-another-list-of-values)

https://www.digitalocean.com/community/tutorials/normalize-data-in-r (https://www.digitalocean.com/community/tutorials/normalize-data-in-r)

https://www.datacamp.com/tutorial/logistic-regression-R (https://www.datacamp.com/tutorial/logistic-regression-R)

https://stackoverflow.com/questions/53357700/cleaning-a-column-in-a-dataset-r (https://stackoverflow.com/questions/53357700/cleaning-a-column-in-a-dataset-r)

https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a (https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a)