THE UNIVERSITY OF
MELBOURNE

MAST90106/MAST90107

Project Report

GROUP 7

Haonan Zhong *867492*

Samy Allouache *1210426*

Supanuth Amorntiyanggoon *1211674*

Xuan Hung Ho *1276655*

Haocong Chen *987916*

Abstract

Painting conservation is broadly practised by museums around the globe, and it is an effort to maintain and preserve the inherent value of a work of art. This project aims to build an interactive dashboard for a dataset consisting of the conservation records of 208 Southeast Asia canvas paintings from the early 20^{th} century. Additionally, we attempted to test the independence between categorical features and to use machine learning models to classify missing values of selected features.

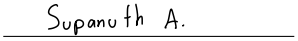
In this report, we will cover our approach to data cleaning and preprocessing, provide a general summary of the implemented interactive dashboard, discuss the approach and result of independence testing between painting attributes, and summarise the attempt on predictive modelling for some missing values of the dataset.

Signed Declaration

I certify that this report does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 8300 words in length (excluding text in images, tables, bibliographies, and appendices).

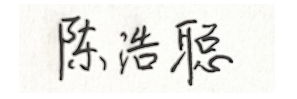
Haonan Zhong

Samy Allouache



Supanuth Amorn.

Xuan Hung Ho



Haocong Chen

Acknowledgements

We would first would like to express our special thanks of gratitude to our project supervisor, Vivek Katial. The completion of this project would not be possible without your guidance and support throughout the year.

Secondly, we would like to express our sincere gratitude to our client, Nicole Tse and the Grimwade Centre for Cultural Materials Conservation, for providing us with the dataset and giving us insightful information about the domain knowledge of the project. We are also thankful for your guidance and support throughout the year. It has been a pleasure working with you.

Additionally, we would like to pay our special thanks to the subject coordinators, Prof. Michael Kirley and Dr. Joyce Zhang for providing us with experience to the industry working environment.

Contents

1 Introduction

1.1 Project Background

Canvas painting can be thought of as a composite object, and it generally consists of five components. The first one is *auxiliary support*, it is the framework which a canvas is stretched on, which can be categorised into stretcher and strainers. Stretcher are rigid frames with flexible and expandable corners, which allows the tension of canvas to be adjustable, while strainer is rigid with fixed and non-expandable corners. Secondly, the *paint support*, which is the canvas fabric made with, for example, cotton, linen, and bast fibre. Following that is the *ground layer*, a priming layer applied on top of the flexible support, and its purpose is to seal and prepare the surface to accept paint. On top of that, it is the *paint layer*, which is the colour layer of a painting. Finally, the frame, the protective and decorative edging for a painting [1].

The practice of oil paint was spread from Western Europe to Southeast Asia in the nineteenth and early twentieth centuries, reflecting the colonial development and religious conversation in the region at the time. Conservation is an important procedure to preserve cultural heritage. However, the conservation of Southeast Asia paintings poses two main challenges, the effect of tropical climate and the used materials. These challenges present a significant source of problems for the storage of the paintings due to the high relative humidity (RH) and the elevated temperature in the tropical region, since components react differently to various temperatures and humidity. For example, the sizing layer with rabbit skin glue are considered as highly responsive materials to RH, which might swell and cause the paint layer on top to fall off. And due to war and expense limitations in the early 20^{th} century, the supply of imported artists' materials is limited. Local artists without access to imported materials started to source painting materials locally instead [2]. Therefore, it is important to study what materials are used in a painting and learn their expected behaviour in the tropical climate.

1.2 Project Overview and Description

Our client is *Dr. Nicole Tse* from *the Grimwade Centre for Cultural Materials Conservation*. Her study aims to develop regional relevant conservation solutions for Southeast Asia paintings, given most of the conservation protocols and practice were originally developed from research taken from the northern hemisphere. The dataset for this project consists of 208 Southeast Asia canvas painting condition reports sourced from four museums. Condition attributes related to the five components are investigated and collected by the client during her PhD studies. As discussed with the client, the expected outcome of this project can be categorised as:

- **Clean and transform the raw dataset**, which contains the conservation record of 208 early to mid 20^{th} century canvas paintings from Southeast Asia, **into an interactive dashboard for visualisation**.
- **Identify the relationship** between different painting conditions, test whether they are independent or associated.
- **Implement machine learning models** to predict missing values for selected attributes that are hard to identify via visual inspection.

To sum up, our project involves data cleaning, interactive data visualisation, independence testing, and predictive modelling from the data science perspective. The challenge we faced at the beginning of this

project was that none of our team members had a profound knowledge of canvas painting and painting conservation. But as the year progressed, the group gained a better understanding of the domain through literature readings and participating in the client's lab demonstration.

Another challenge is related to the given dataset. Our client put together the dataset almost twenty years ago using a legacy version of the software FileMaker Pro. Throughout the past two decades, the software has gone through multiple updates. Therefore, we noticed some inconsistencies and errors when exporting the dataset, making the cleaning process more laborious than we expected. This challenge will be discussed further in the data cleaning and preprocessing section.

2 Related Work

In this section, we will be focusing on reviewing past studies and works that are relevant to our project.

2.1 Dashboard

A dashboard is an information management tool with a graphical interface that allows users to visualise and understand the data more efficiently. Before we can build our dashboard, we must first identify the suitable tool among the many dashboard packages. Our supervisor has suggested that we use the R Shiny package to construct the dashboard; due to its simplicity, the dashboard user does not need to understand and master R to use it, which is perfect for the client. Shiny dashboards have been implemented for various areas, such as health care and the environment. In the early stage of the COVID-19 pandemic, the interactive interface developed by the London School of Hygiene and Tropical Medicine provides the latest information on the spread [3]. It enables the user to make comparisons with other recent disease outbreaks. Another example is the New Zealand Trade Intelligence Dashboard, which presents a full picture of New Zealand's trading profile through interactive charts [4]. These examples effectively helped and gave us inspiration on how to implement and improve the user interface of our proposed dashboard.

2.2 Independence Testing of Categorical Attributes

To identify the association between categorical variable, C. Fraser [5] suggested tabulating the frequencies of categorical variables as contingency tables, where χ^2 test allows us to quantify the association between two categorical variables, if the two are related, then the probability of one will depend on the probability of the other. However, due to the size of our dataset, we may need to look for alternative solutions, as their research also stated that contingency analysis with χ^2 test requires a large and balanced dataset to ensure a stable and reliable approximation of the test statistics. On the other hand, J.H. McDonald [6] proposed to use Fisher's exact test for independence testing, which is more accurate than χ^2 test when the sample size is smaller than 1000.

3 Data Cleaning and Preprocessing

Data preprocessing and feature engineering plays an important role in the data science pipeline, which helps us to get the best out of our data. This section will discuss our approach on data cleaning and preprocessing.

3.1 Presentation of the Dataset and Manual Cleaning

As we mentioned in the project overview, the given dataset consists of 208 canvas painting condition reports gathered in a single FileMaker Pro file, each record being a condition report. An example of the condition report is shown in Figure 1.

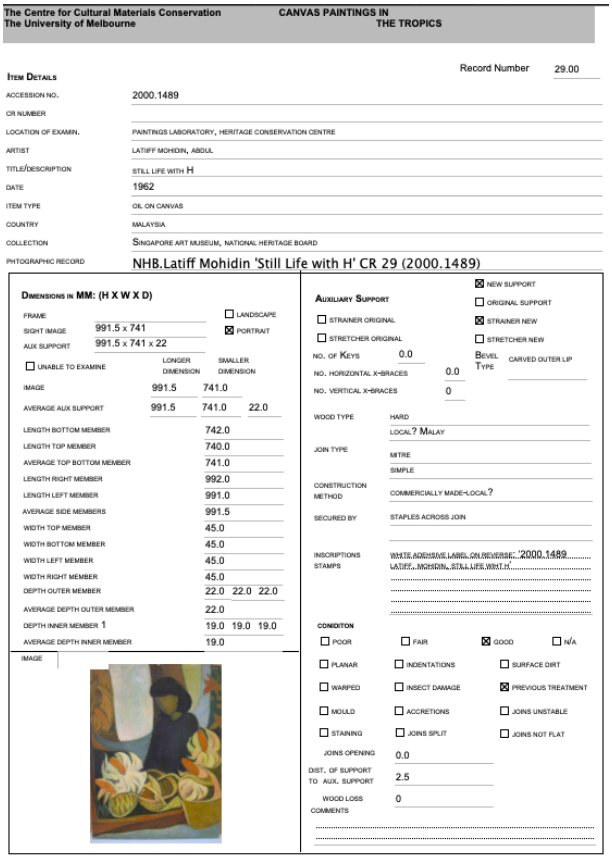


Figure 1: Example of a FileMaker Pro Condition Report

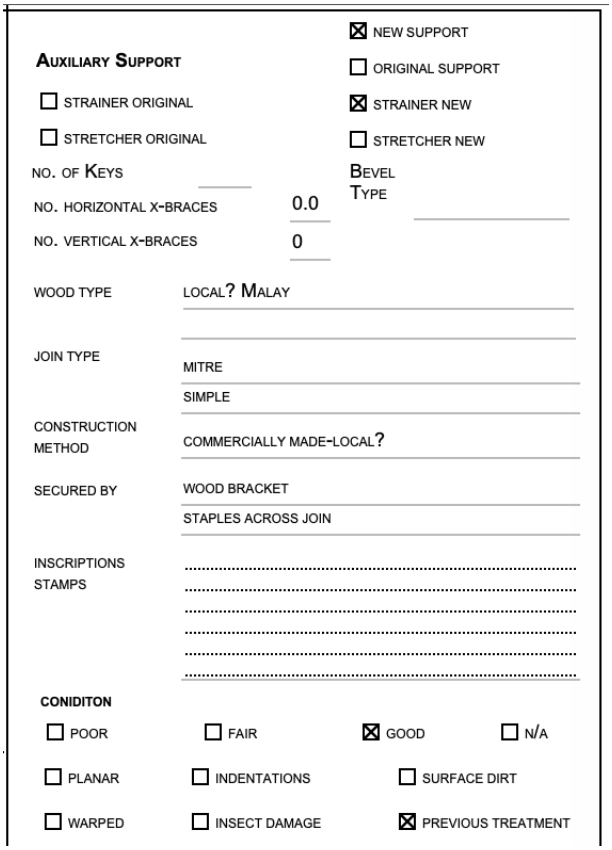


Figure 2: Zoom on the auxiliary support part

A condition report starts with general information such as the painting's accession number (identifier of the painting), the name of the painting, the name of the artist, the country and the museum collection it belongs to, the dimensions and a thumbnail of the painting if available. It then focuses on the condition of every main component of the piece of art: the auxiliary support, the painting support, the ground layer, the painting layer, the surface coating and some miscellaneous information.

For each of these main components, the report offers a large number of attributes that need to be filled in by the author of the report. There are two different types of attributes: they can either be text fields that the reporter can fill or boxes that can be ticked. An example for the text fields would be the "comment" attribute from the auxiliary support, and one for the boxes would be the condition (poor, fair, good, excellent or N/A) of the ground layer shown in Figure 2.

However, the FileMaker Pro file is not directly usable in the data preprocessing pipeline. It was exported into an excel file using the "export to excel" option from FileMaker Pro. Unfortunately, the excel file could not directly be used either. Indeed, it presented two types of problem:

- Some of the exported values were just not correct as shown in the red rectangle in Figure 3.

- Categorical and ordinal variable that have several possible values were exported as separated variables in the excel file (cf. blue rectangle in Figure 3)

	A	C	J	N	X	Y	Z
1	accession no	artist	stretcher original	strainer original	poor (aux support)	fair (aux support)	good (aux support)
2	UPVMA-III.00240	Gallardo, Alladin		strainerstrainer original			good
3	UPVMA-III.00085	Ancheta, isidro		strainerstrainer original	poor		
4	UPVMA-III.00292	Lagniton, Jose		strainerstrainer original			good
5	UPVMA-III.00126	Buenaventura y Espana, Oscar		strainerstrainer original		fair	
6	UPVMA-III.00161	Cristobal, bonifacio		strainerstrainer original			good
7	UPVMA-III.00289	Jervosa, Fortunato		strainerstrainer original			good
8	UPVMA-III.00180	Dumlao, Antonio	stretcher original	strainer			good
9	UPVMA-III.00177	dizon, vicente alvarez		strainerstrainer original			good
10	UPVMA-III.00205	Enriquez, Romeo		strainerstrainer original	poor		
11	UPVMA-III.00430	Trinidad, Jose V. Jr		strainerstrainer original			good
12	UPVMA-III.000452	Zablan, Felix A		strainerstrainer original			good
13	UPVMA-III.00293	Lagniton, Jose		strainerstrainer original			good
14	UPVMA-III.00269	Gonzales, Felix y Pinto		strainerstrainer original		fair	
15	UPVMA-III.00338	Navarro, Oscar		strainerstrainer original		fair	
16	UPVMA - III.00296	laxa, elias		strainerstrainer original			good

Figure 3: raw export from FileMakerPro

We could not find any explanation nor any pattern for the first problem even after discussion with the client. This resulted in the impossibility to implement an algorithm to solve it. Therefore, we had to clean it manually by correcting all the explainable export errors and the typos due to manual entries. To do that, each member was assigned a section of the raw excel file and we had to correct these typos or pieces of text that were not supposed to be here. We ended up with a manually cleaned excel file that we could process using a python script (cf. Figure 4)

	A	C	AU	AV	AW	AX	BN	BO	BP
1	accession no	artist	strainer new	stretcher new	strainer original	stretcher original	poor (aux support)	fair (aux support)	good (aux support)
2	UPVMA-III.00240	Gallardo, Alladin			strainer original				good
3	UPVMA-III.00085	Ancheta, isidro			strainer original		poor		
4	UPVMA-III.00292	Lagniton, Jose			strainer original				good
5	UPVMA-III.00126	Buenaventura y Espana, Oscar			strainer original			fair	
6	UPVMA-III.00161	Cristobal, bonifacio			strainer original				good
7	UPVMA-III.00289	Jervosa, Fortunato			strainer original				good
8	UPVMA-III.00180	dumlao, antonio gonzales				stretcher original			good
9	UPVMA-III.00177	dizon, vicente alvarez			strainer original				good
10	UPVMA-III.00205	Enriquez, Romeo			strainer original		poor		
11	UPVMA-III.00430	Trinidad, Jose V. Jr			strainer original				good
12	39/2521	krua-in-khong							
13	UPVMA-III.000452	Zablan, Felix A			strainer original				good
14	UPVMA-III.00293	Lagniton, Jose			strainer original				good
15	UPVMA-III.00269	Gonzales, Felix y Pinto			strainer original			fair	
16	UPVMA-III.00338	Navarro, Oscar			strainer original			fair	
17	UPVMA - III.00296	laxa, elias			strainer original				good
18	UPVMA - III.00003	ALANO, BEN			strainer original				good
19		unknown				stretcher original			

Figure 4: manually cleaned data

3.2 Metadata

Data Dictionary

No	Category	Field	Description	Example Data	Require to cl	Important(Y/I	Remarks
27	Painting support	canvas	Flexible support Type, cotton is most preferable		Y	Y	If the value is cotton it means good, What is the diff cotton and cotto
28	Painting support Condition	painting support comments	Comment			Y	Meeting on 29/7/2022, we plan to include all comment as search key
29	Painting support Condition	planar (painting support)				Y	
30	Painting support Condition	warped (painting support)				Y	
31	Painting support Condition	good tension (painting support)	positive tension		Y	Y	What is adequate tension? Only have positive tension field treat the s
32	Painting support Condition	loose (painting support)				Y	
33	Painting support Condition	taut (painting support)				Y	
34	Painting support Condition	corner distortions painting support				Y	Issue p0820 all field blank , but data mark as distortion -> treat as uni
35	Painting support Condition	indentations (painting support)				Y	Highly damage the condition
36	Painting support Condition	holes (painting support)				Y	
37	Painting support Condition	tears (painting support)				Y	
38	Painting support Condition	surface dirt (painting support).				Y	
39	Painting support Condition	staining (painting support)				Y	
40	Painting support Condition	painting support condition	comment			Y	Text based need to split the text
41	Painting support Condition	poor (painting support)	In general painting support condition			Y	
42	Painting support Condition	fair (painting support)	In general painting support condition			Y	
43	Painting support Condition	good (painting support)	In general painting support condition			Y	
44	Painting support Condition	excellent (painting support)	In general painting support condition		Y	Y	Issue if all respond is blank (p0820,UPVMA-III00260) -> Removed the
181	Painting support	solid support type			Y	Y	Logic solid support type is opposite of flexible support type If the value mark both flexible and solid, what should we treat this fie
184	Painting support	weave type			Y	Y	If value is blank? Remove ? Ex p0820 Do we still need this field
189	Painting support Condition	mould				Y	
191	Painting support Condition	overall distortions				Y	
192	Painting support Condition	insect damage				Y	Highly damage the condition
194	Painting support Condition	bottom distortions				Y	
195	Painting support Condition	rust stains on support				Y	
196	Painting support Condition	top distortions				Y	
197	Painting support Condition	deformation around tacks staples				Y	
199	Painting support Condition	tears around tacks staples				Y	
200	Painting support Condition	loss of tacks insecure support				Y	

Figure 5: Metadata

We have also developed metadata for the given dataset, shown in Figure 5, to summarise information on each data field. Each field would collect the index of a feature from the exported CSV file since we found out that the columns fields were not sorted correctly. We also described the logic of some essential fields in order to analyse them when we did some in-depth analysis. Hence, our team could locate and work with particular attributes easier.

3.3 Preprocessing Steps

The preprocessing script had two goals :

1. Select the columns we needed for the dashboard only
2. Reduce the number of features by fusing the columns that represented the same categorical/ordinal data

In order to do it, we first created four python dictionaries at the beginning of the script. Each dictionary had keys which were the names of the final features we had selected, the argument of the keys would depend on the dictionary. Two dictionaries are represented in figure 6, the complete list is :

- `BooleanColumnsDict` : A dictionary for which the argument of each key is the position of the column in the manually cleaned file.
- `CategoricalColumnsDict` : A dictionary for categorical data. The argument of each key is a list of column indexes representing all the columns in the manually cleaned file that refer to the same categorical feature.
- `MultipleValuesCatColDict`: A dictionary for columns that represented several categorical values in a single cell. The argument of each key is a tuple presenting the index of the original cell, the maximum number of values that are in the original cell and the list of all possible values.
- `OrdinalColumnsDict`: A dictionary for ordinal data. The argument of each key is a list of column indexes representing all the columns in the manually cleaned file that refer to the same ordinal feature. The indexes are sorted from the lowest ordinal value to the highest.

```python
# We regroup text fields and categorical data
# The text columns we wanted to keep will only have a list of 1 index.
CategoricalColumnsDict = {
    "accession_number": [0],
    "artist": [2],
    "canvas": [210],
    "title": [12],
    "date": [13],
    "country": [14],
    "collection": [15],
    "sight": [7],
    "support_type": [46, 47, 48, 49],
    "commentary_auxiliary_support": [64],
    "wood_type_hardness": [76],
    "wood_type_country_locality": [77, 78, 79],
    "media_type_1": [112],
    "media_type_2": [113],
    "media_type_3": [114],
    "ground_layer_application": [120, 121],
    "ground_layer_limit": [129, 130],
    "ground_layer_thickness": [123, 124],
    "frame_material": [163, 164],
    "slip_presence_frame": [170, 171],
    "glazed_frame": [172, 173],
    "frame_affixed_to_wall_by": [178, 179, 180, 181],
    "frame_hanging_system": [182, 183],
    "frame_strand_wire": [187, 188],
    "backing_board_type": [190, 191, 192],
}
```

```python
OrdinalColumnsDict = {
    "auxiliary_support_condition": [65, 66, 67, 68],
    "media_condition": [90, 91, 92, 93],
    "ground_condition": [115, 116, 117, 118],
    "painting_support_condition": [132, 133, 134, 135],
    "frame_condition": [205, 206, 207, 208],
}
```

Figure 6: Example of dictionaries

The first step in the preprocessing script was to fuse columns of the same categorical feature. This usually involved the fuction `fuseCategColumns` defined in Algorithm 2. There are five specific features that were treated differently :

- `canvas` : which was stripped of all the information that was not directly the canvas type (leaving only the values: `cotton`, `linen` or `bast`)
- `wood_type_country_locality` : which was separated into three new columns `wood_type`, `wood_country` and `locality`
- `collection` : because the names of each collection are changed from the raw data after a request from the client
- `sight` : because it is used to create a `area` feature
- `date` : which is used to create a `decade` feature

Then the code tackled `MultipleValuesCatColDict`. This consisted in creating a binary column for each possible value for this feature. This was done by checking in a `for` loop on the possible values if the long `string` in the original data contained the said value or not.

After that the script fused together the ordinal data. This was done with `fuseOrdinalColumns`, the algorithm of which is described in annex 3. It considered the columns in the original data that referred to the same ordinal feature and for a specific row it gave the level of the first non empty column it had encountered. If another column had a value, it then computed the mean with the former assigned level and took the closest upper integer. The lowest level was always 0 and the rows without any value were filled with NaNs.

Finally, the script handled binary data. In the original dataframe, these columns were either filled with the name of the column itself when the attribute was present or they were left empty. The script only checked if for a given column some cells were empty or not. It then transformed empty cells into 0 and the others into 1.

4 Exploratory Data Analysis

Exploratory data analysis is an important process to analyse and summarise the characteristic of a dataset. In this section, we will be exploring the provided dataset to present a general idea of the data that we are dealing with.

4.1 How is the Data Distributed?

In our analysis, we explored various categorical features of the dataset to get more depth about the data. First, we have explored the painting distribution across the four museums, as shown in Figure 7. We saw that the National Heritage Board of Singapore has the highest number of paintings at 63, followed by JB Vargas Museum and Balai Seni Negara of Malaysia. In comparison, the National Gallery of Thailand has the least number of paintings at 33.

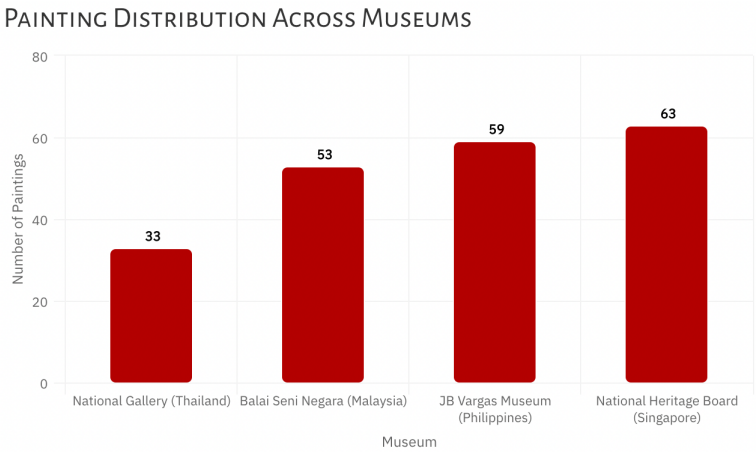


Figure 7: Distribution of Paintings Across Museums

Next, we explored the painting distribution across the timeline as shown in Figure 8. We can observe that the distribution of the paintings were heavily left skewed as most of the paintings were between the 1930s and 1960s. Another point worth mentioning is that all the Balai Seni Begara (Malaysia) paintings were from after 1920.

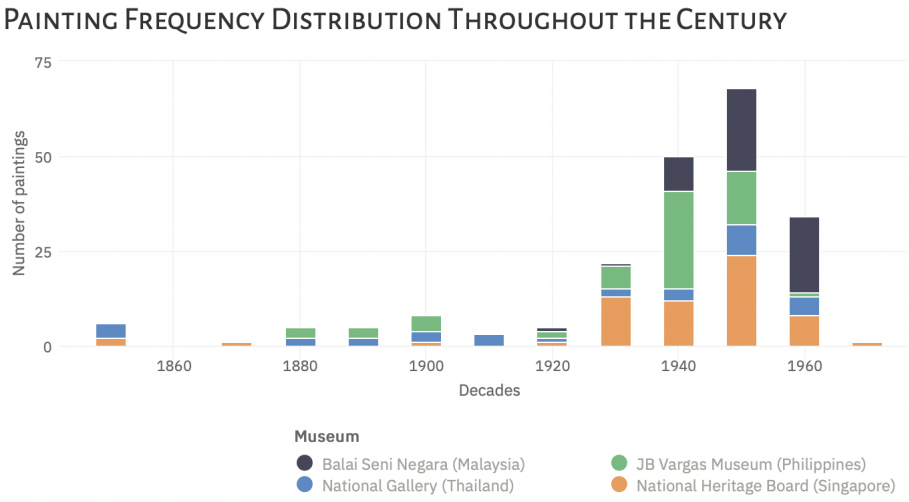


Figure 8: Distribution of Paintings Across the Timeline

In Figure 9, we explored the commonly used materials across all museums. Apart from the paintings with unspecified materials, we can see that the number of paintings with cotton canvas is significantly higher than other used fabrics in Balai Seni Negara and JB Vargas Museum. In contrast, the number of paintings with cotton or linen canvas is relatively the same in the National Gallery of Thailand and the National Heritage Board of Singapore.

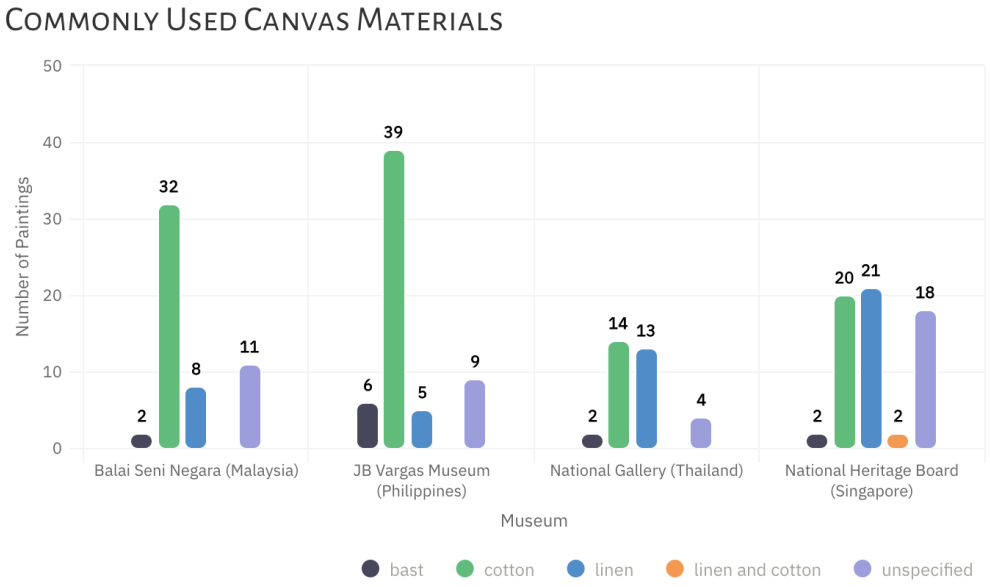


Figure 9: Commonly Used Canvas Materials

Moreover, Figure 10 shows that oil paint is the most popular media/paint materials across the four museums. As over 90% of all paintings were painted with oil.

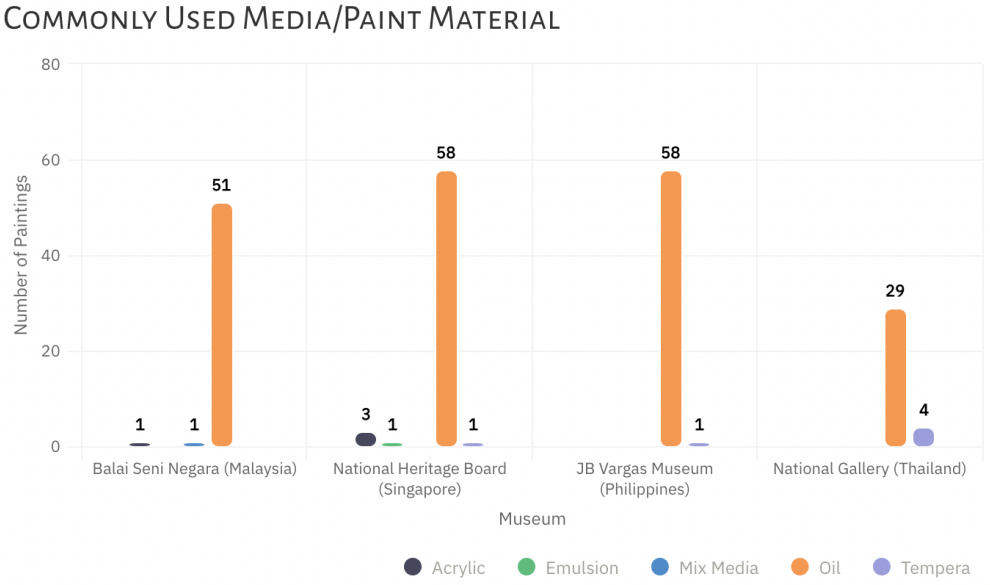


Figure 10: Commonly Used Media/Paint Materials

Furthermore, we have explored the distribution of ground type as shown in Figure 11. Apart from the paintings with unspecified or unsure ground type, we can see that over 60% of paintings' ground layers in the National Heritage Board were commercially applied. In contrast, 59% of paintings from Balai Seni Negara has artist applied ground. While the number of paintings with commercial applied or artist applied

ground is relatively the same for JB Vargas Museum and the National Gallery of Thailand.

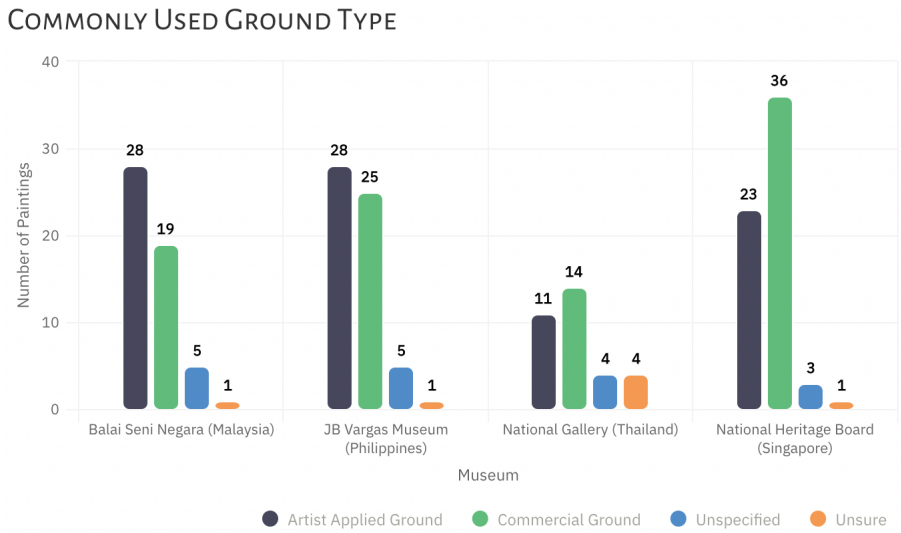


Figure 11: Commonly Used Ground Type

4.2 How is the Condition Rating of Paintings from each Museum?

As mentioned in the project domain section, paintings generally consist of five components. A condition rating score (Poor, Fair, Good, or Excellent) was assigned to each component of the painting by the client during the study. Here we explored the condition rating distribution for each museum to get more depth about the data.

4.2.1 Auxiliary Support Condition Rating

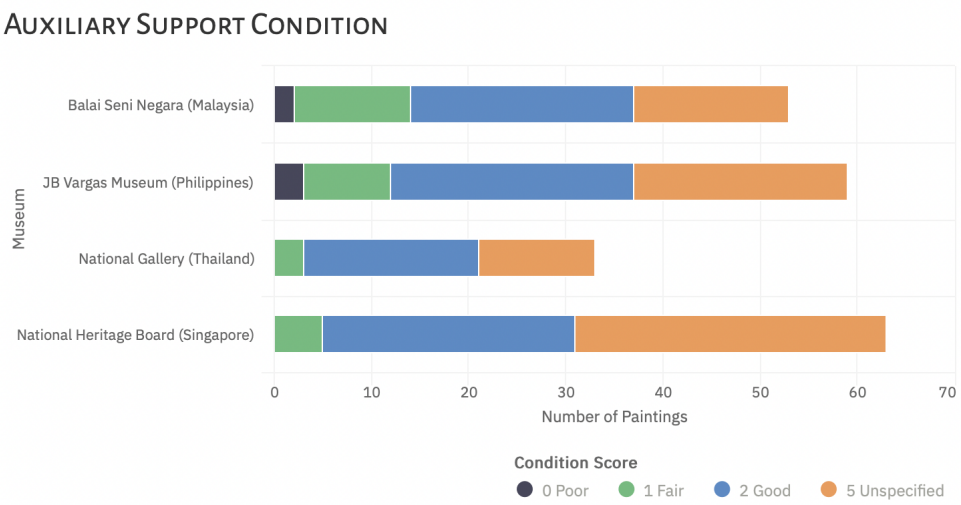


Figure 12: Auxiliary Support Condition Overview

As per the stacked bar chart shown in Figure 12, each museum has a fair amount of paintings with an unspecified condition score for auxiliary support. This is either because the auxiliary support could not be evaluated or because the piece of art did not present any sort of auxiliary support. Moreover, each museum has approximately the same amount of paintings in good condition for auxiliary support, and no paintings

with an excellent rating can be found. However, we notice a few paintings from JB Vargas Museum and Balai Seni Negara that were assigned to poor.

4.2.2 Paint Support Condition Rating

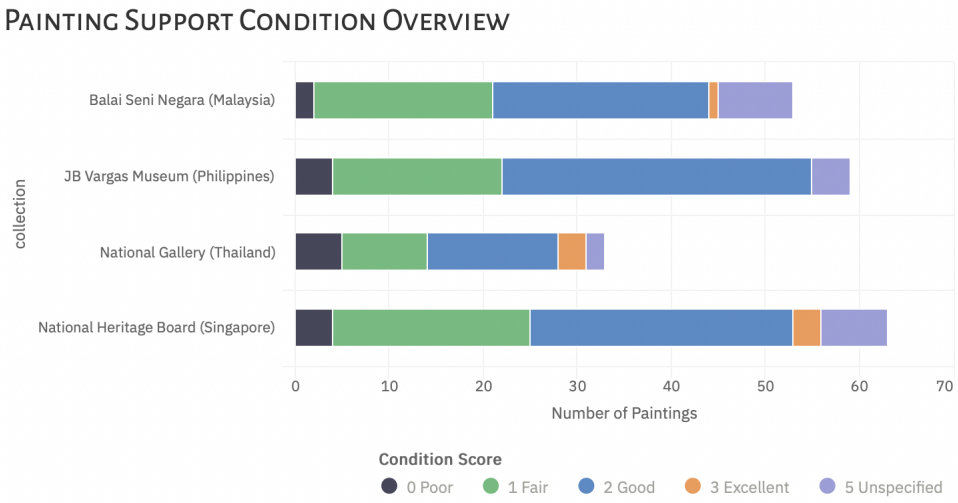


Figure 13: Paint Support Condition Overview

As for the paint support condition shown in Figure 13, painting with unspecified ratings were found across all the museums. And we saw that most paintings were assigned with a fair and good condition for paint support, while only a tiny portion of paintings was assigned to poor or excellent.

4.2.3 Ground Layer Condition Rating

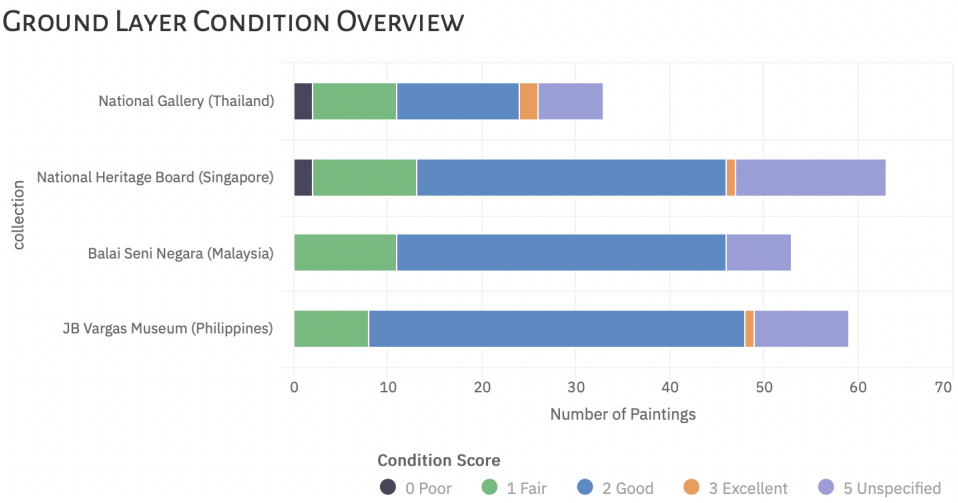


Figure 14: Ground Layer Condition Overview

The ground layer condition of the paintings across the museum were shown in Figure 14. Again, paintings with unspecified ground ratings were found across each museum. We saw that most of paintings were clustered between fair and good. Only a tiny portion of paintings were classified as poor or excellent.

4.2.4 Paint/Media Layer Condition Rating

Similarly, we have created a stacked bar chart showing the number of paintings in each museum based on their paint layer condition which is shown in Figure 15. Paintings with unspecified paint layer condition can be observed across all museums. The story is relatively the same with previous layers, where most paintings were classified with fair and good ratings. It is also worth noting that most paintings with excellent ratings were from the National Heritage Board of Singapore, while the number of paintings with poor ratings in the National Gallery of Thailand are relatively higher than the other museums.

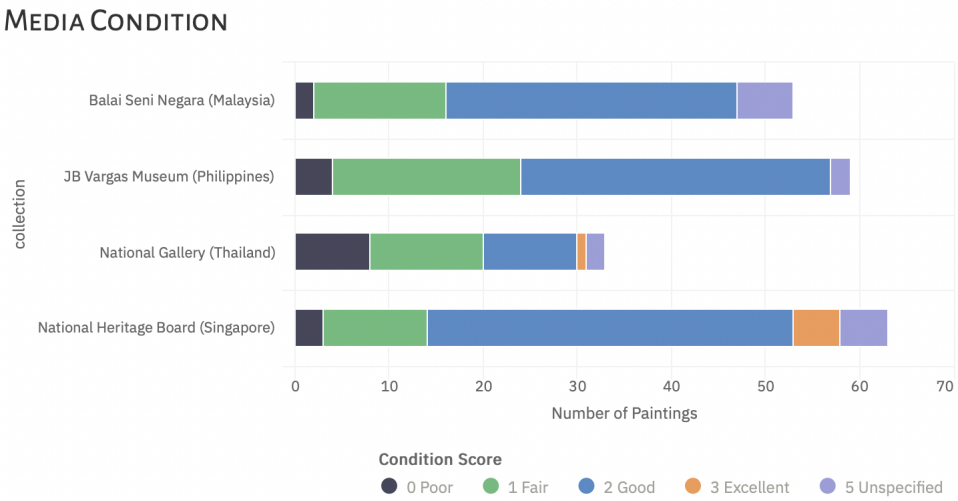


Figure 15: Paint/Media Layer Condition Overview

4.2.5 Frame Condition Rating

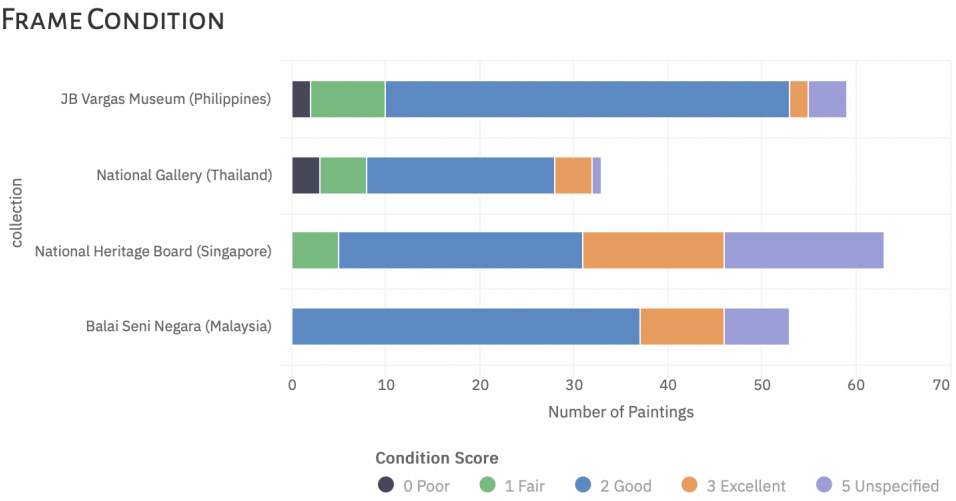


Figure 16: Frame Condition Overview

Frame condition across the four museums is shown in Figure 16. Paintings with unspecified frame condition can be identify across all museums, which is either missed or does not consist of a frame. We observed that the number of paintings with unspecified frame condition are comparatively higher in the National Heritage Board of Singapore. And it is noteworthy that the portion of paintings with excellent condition is considerably higher than other component mentioned above.

4.3 How is the Painting Condition Among the Museums?

As discussed with the client, the client provided the scope of requirements to the team to explore the condition attributes on paint support, ground layer, and paint layer rather than auxiliary support and frame, given that the client is more interested in those layers. Here, we have selected a few condition attributes of those layers to explore their general distribution among the museums.

4.3.1 Holes Condition

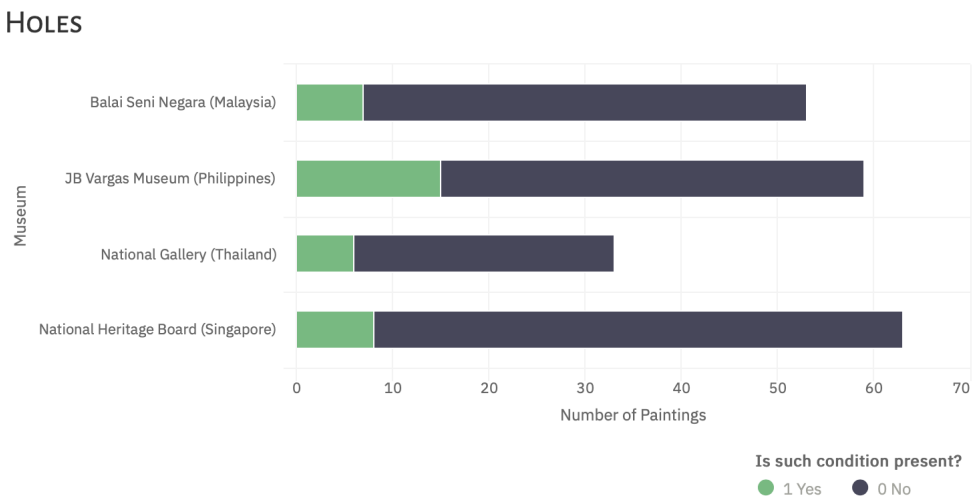


Figure 17: Holes condition

As the Holes condition stacked bar chart shown in Figure 17, each museum has a similar amount of holes condition. In contrast, JB Vargas museum has nearly double the holes condition of other museums (15 paintings).

4.3.2 Positive Tension Condition

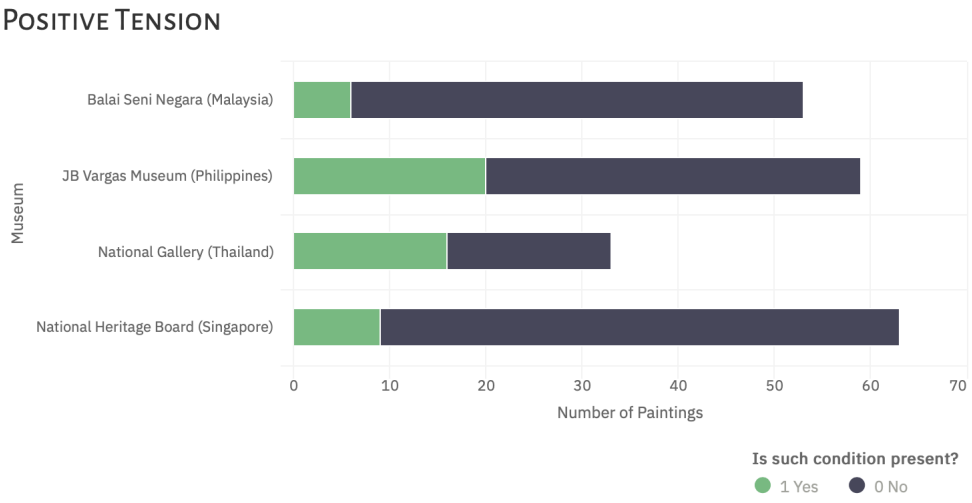


Figure 18: Positive Tension condition

As the Positive Tension condition stacked bar chart shown in Figure 18, Balai Seni Negara (Malaysia) and National Heritage Board (Singapore) have a similar amount of Positive tension condition. In contrast, JB Vargas museum and National Gallery (Thailand) have nearly double the holes condition of other museums (20 and 16 paintings, respectively).

4.3.3 Are Ground layer Thinly or Thickly Applied

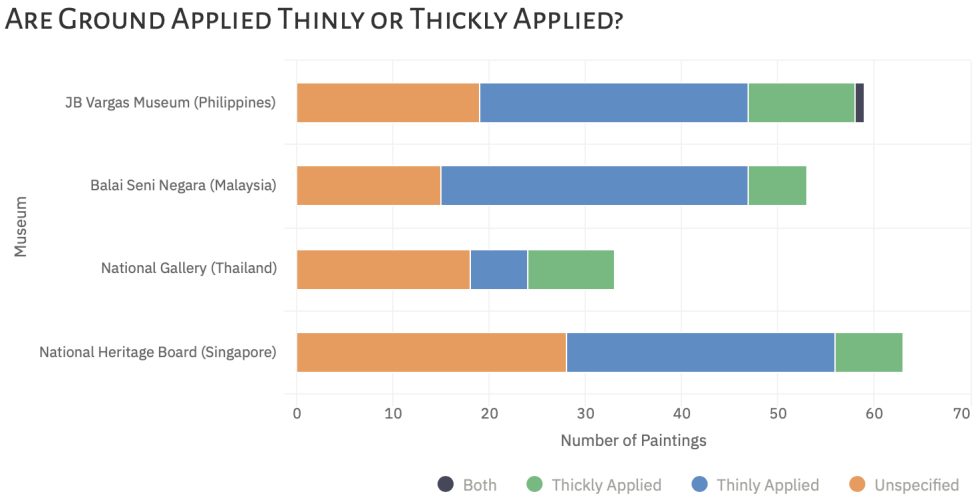


Figure 19: Are ground layer Thinly or Thickly Applied

As the ground layer Thinly or Thickly Applied condition stacked bar chart shown in Figure 19, most museums have Thinly applied rather than Thickly applied expecting Nation Gallery (Thailand). However, we found that unspecified conditions are significant huge among collections. Therefore, our team was inquisitive about filling unspecified conditions by predictive models.

4.3.4 Uniform application

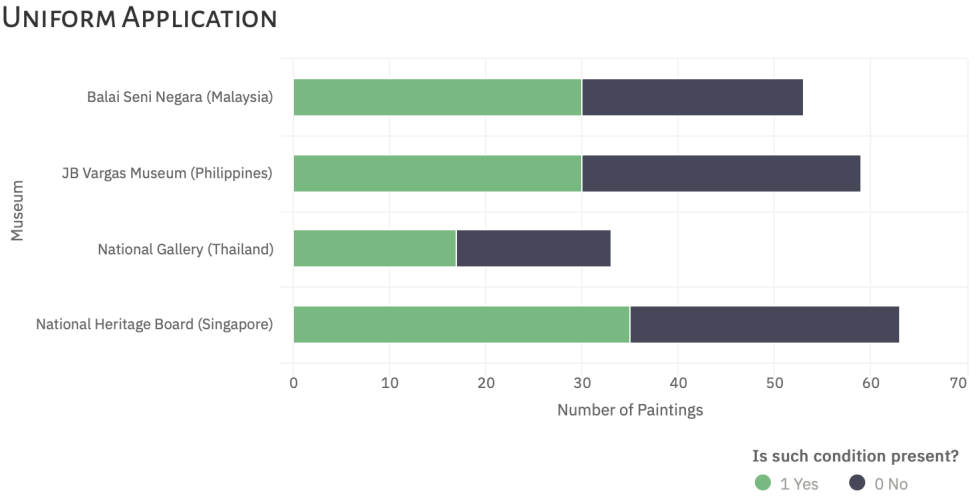


Figure 20: Uniform application

According to the stacked bar chart as shown in Figure 20, Every museum has a similar proportion of yes and no for uniform application. Our team suspected this condition is related to other conditions such as

ground type, Thinly or Thickly Applied ground.

4.3.5 Painting Plastic Behaviour

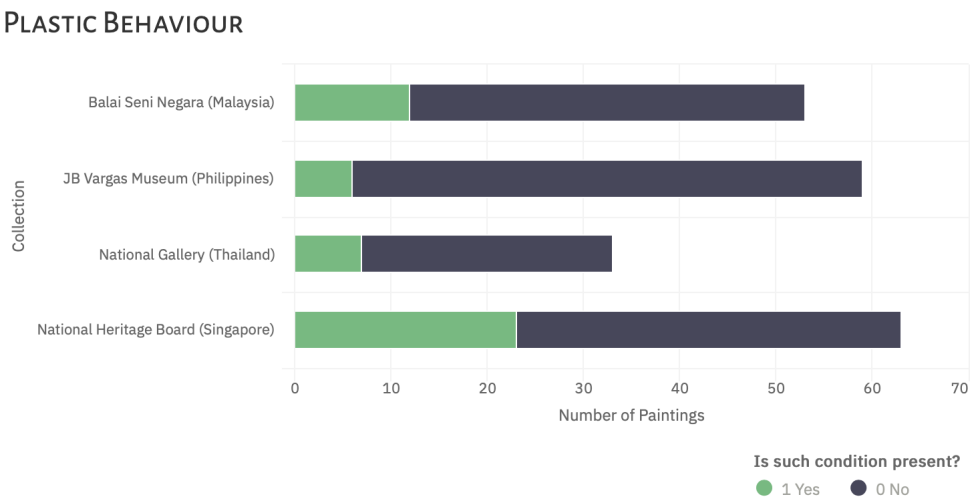


Figure 21: Plastic Behaviour

According to the stacked bar chart as shown Figure 21, National Heritage Board (Singapore) has the most Plastic Behaviour on paintings layers among collections, 23 paintings. In contrast, the JB Vargas Museum and National Gallery (Thailand) have minor Plastic Behaviour on layers 6 and 7, respectively. In addition, we examined the relationship with the painting layer condition. We discovered that there is an opposite relationship with painting layer condition when painting appears Plastic Behaviour is less likely to have good conditions.

4.3.6 Painting Elastic Behaviour

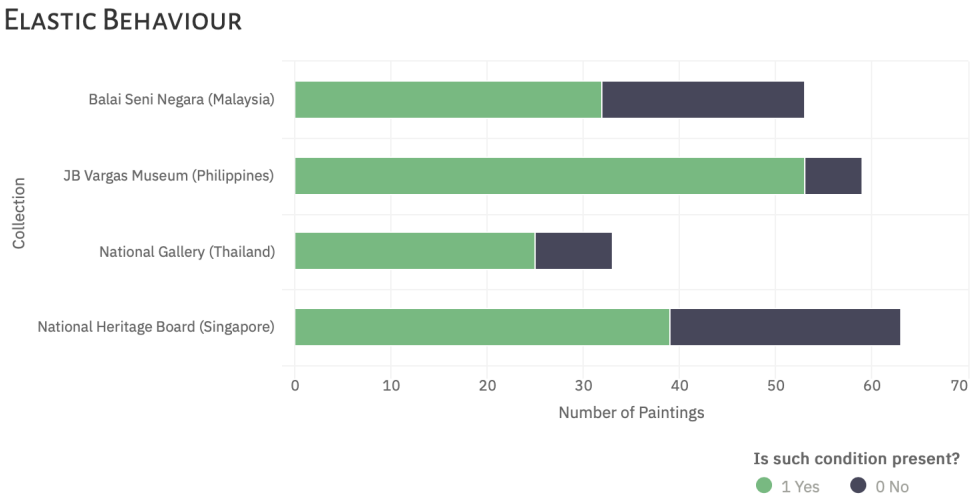


Figure 22: Elastic Behaviour

As the Elastic Behaviour condition stacked bar chart shown in Figure 22, most museums present Elastic Behaviour. Furthermore, JB Vargas Museum has the most Elastic Behaviour condition on 53 paintings. In addition, we examined the relationship with the painting layer condition. We discovered that there is a

positive relationship with painting layer condition when painting appears Elastic Behaviour is more likely to have good conditions.

5 Interactive Dashboard

As discussed in the data preprocessing section, although FileMaker Pro is an excellent tool for collecting and recording data, it does not provide data visualisation for exploration and comparison. Therefore, transforming the given dataset and creating an interactive dashboard for the client has been our primary focus this semester. In this section, we will focus on covering our approach to creating the dashboard.

5.1 Used Tools and Packages

R Shiny, *Leaflet*, *Highcharter*, and *Polished* are the four main tools we make use of when implementing the dashboard. *Shiny* is an R package that enables people to build interactive web interface straight from R, while the *Leaflet* package enables us to deploy interactive map visualisation in the Shiny app. Moreover, the *Highcharter* package allows us to add interactivity into visualisation figure to display more information when the user interacts with the plot. Finally, the *Polished* package allows us to establish authentication to the Shiny application to ensure confidentiality.

Along the way, we also considered tools like Tableau. However, to increase the technical challenge of this project, this idea was soon given up.

5.2 Dashboard Design Summary

5.2.1 Homepage

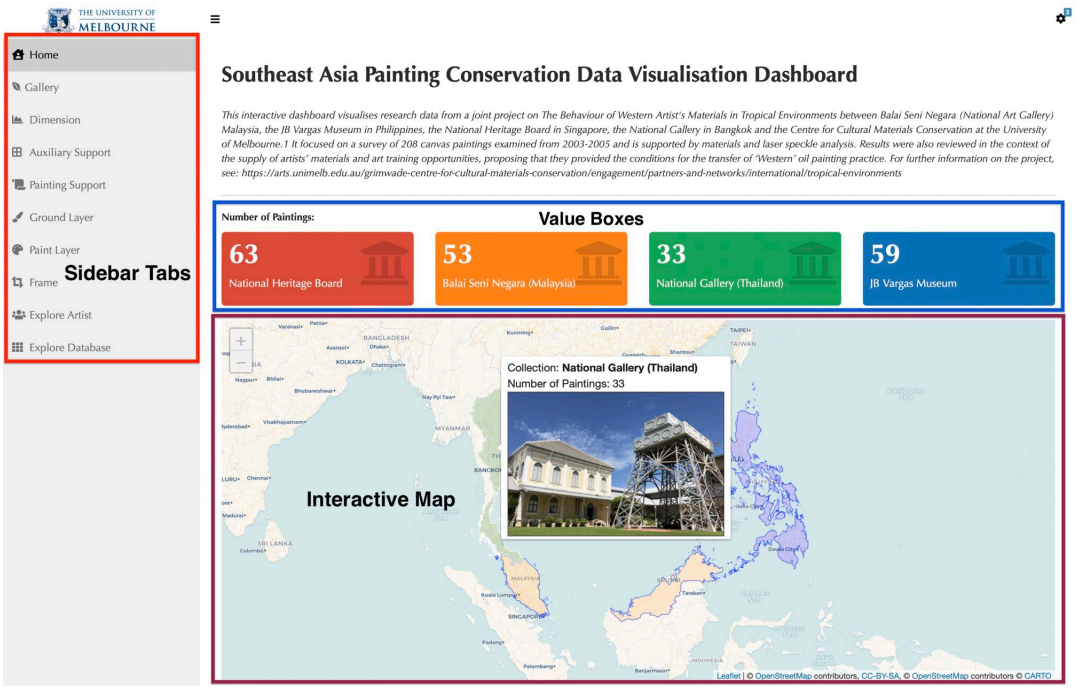


Figure 23: Dashboard Homepage

Our dashboard homepage are shown in Figure 23. The homepage displays general information about the study behind the dashboard. The value boxes on top show the number of paintings in each of the four

museums. We also used an interactive map to highlight the country in which each museum is located. Moreover, when the user hovers the mouse over the highlighted country's polygon, the tooltip will display the museum name, the number of paintings that participated in the study, and the photo of the museum. Furthermore, the sidebar tabs on the left allow users to navigate to the section they want to view.

5.2.2 The Gallery Tab

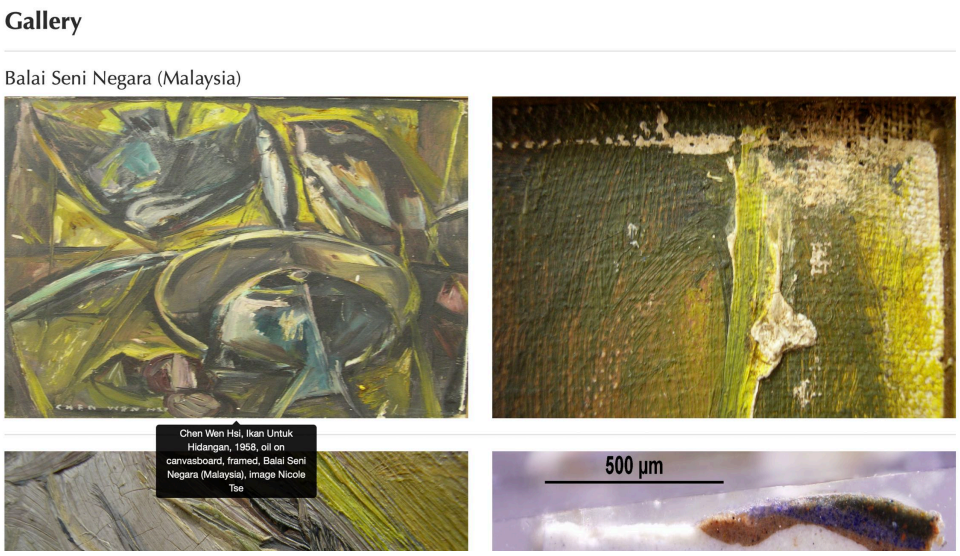


Figure 24: Gallery Tab

The gallery tab is dedicated to showcasing the paintings that participated in the study. Six paintings from each museum were selected for the showcase by the client. As shown in Figure 24, painting information is displayed via the tooltip when the user hovers above the image.

5.2.3 The Dimension Tab

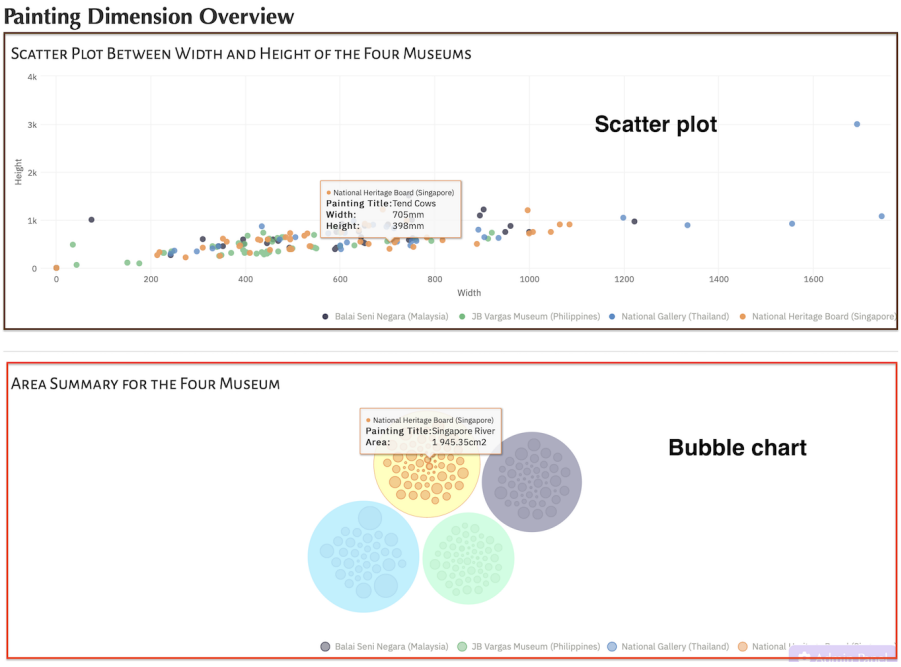


Figure 25: Dimension Tab

As shown in Figure 25, this tab is created to give the user a clear picture of the dimension of the paintings from each collection. In the scatter plot shown above, we plotted the height against the width for each painting. Paintings from the different museums were separated by hue. Additionally, the tooltip will display information such as the painting's name, width, and height when the user hovers above the dots. Furthermore, the legend below the chart allows users to filter out museums that they do not wish to be included in the comparison.

The packed bubble chart below serves the same purpose, but instead of comparing paintings using width and height, it directly shows the calculated area. Each painting was represented by a bubble, while the area of the painting determined the radius of the bubble. Furthermore, paintings were grouped by the museum, creating a visual hierarchy. Moreover, the tooltip displays general information about a painting: its name and calculated area, which provides a clear picture of the size of the paintings among the four museums.

5.2.4 Tabs for the Painting Components

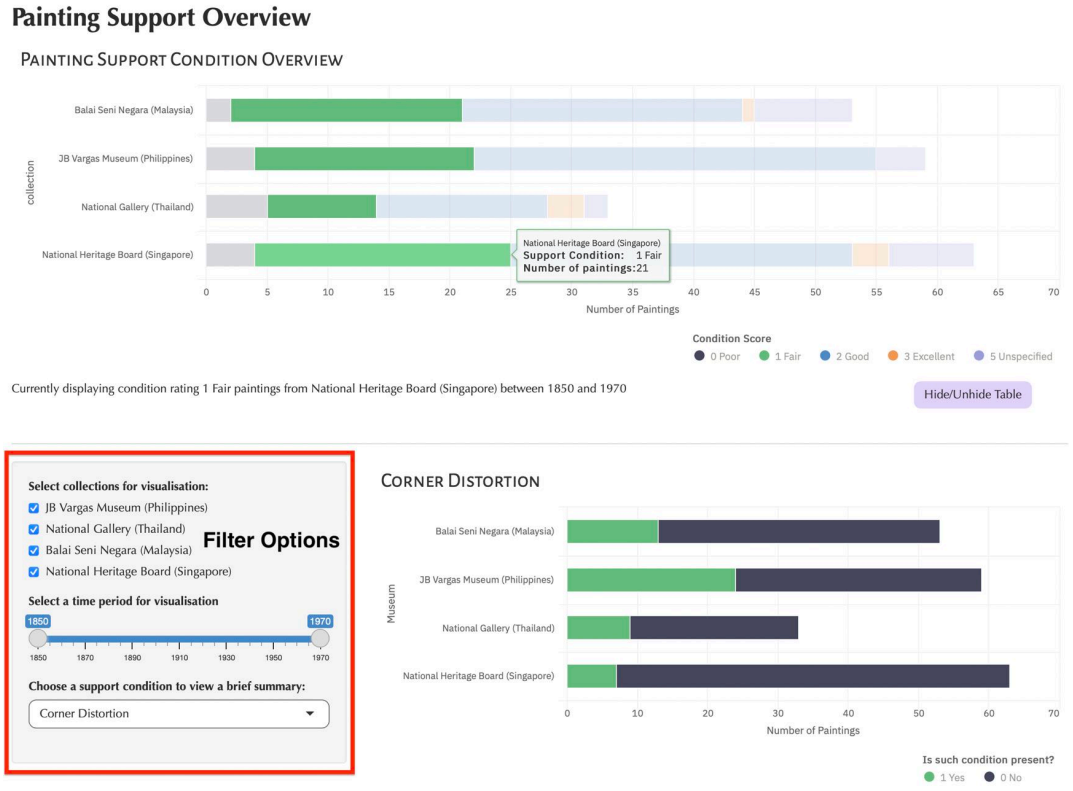


Figure 26: Painting Support Tab

The interface of the painting support tab is shown in Figure 26. The design is shared among the tabs for the five components. The stacked bar chart above depicts the number of paintings in each museum, and each bar is further broken down into sub-amounts based on the component condition ratings. Moreover, the tooltip displays a brief summary of the hovering region: the condition and number of paintings belonging to the condition.

Furthermore, the filter option shown in Figure 27 introduces interactivity, allowing the user to dynamically filter the museum to be included in the comparison, the period for visualisation, and the support condition

for visualisation. Therefore, instead of laying charts for every support condition in the interface, the filter allows the user to choose which condition to visualise on the fly, keeping the interface tidy.

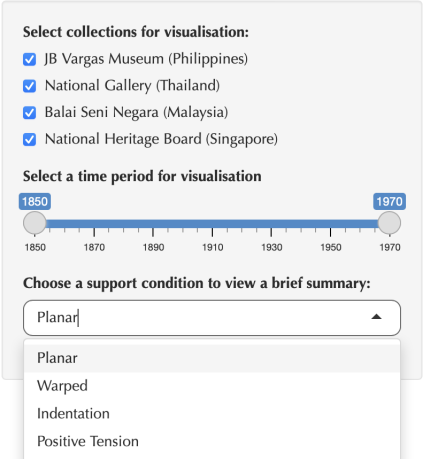


Figure 27: Filter Option

5.2.5 The Dataset Exploration Tab

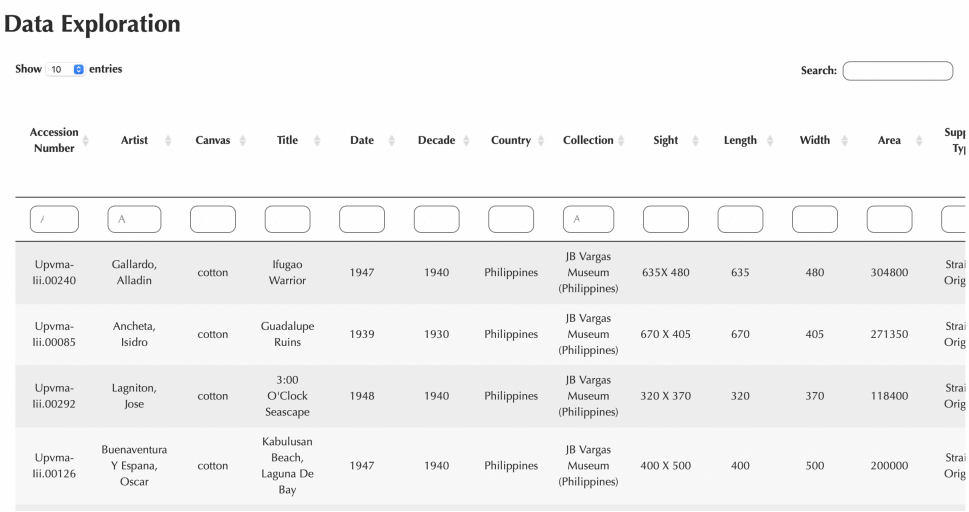


Figure 28: Explore Database Tab

As shown in Figure 28, the data exploration tab was dedicated to dataset exploration. It allows the user to browse a cleaned version of the dataset within the dashboard. In addition, a filter and search option have been added for each attribute, allowing the user to locate specific content by inputting a search string or selecting filter options.

5.2.6 Reactivity and Click Event

The main problem we encountered while building the dashboard was that basic interactive visualisation only provides limited information through the popup labels, such as the number of paintings under a specific condition. As a result, it loses a lot of information because it only tells the user how many paintings were under a particular condition, but it does not tell the user what paintings are actually under those conditions. Therefore, our proposed solution to this problem is to use the reactive functionality of Shiny and Highcharter's ability to capture the user's click events.

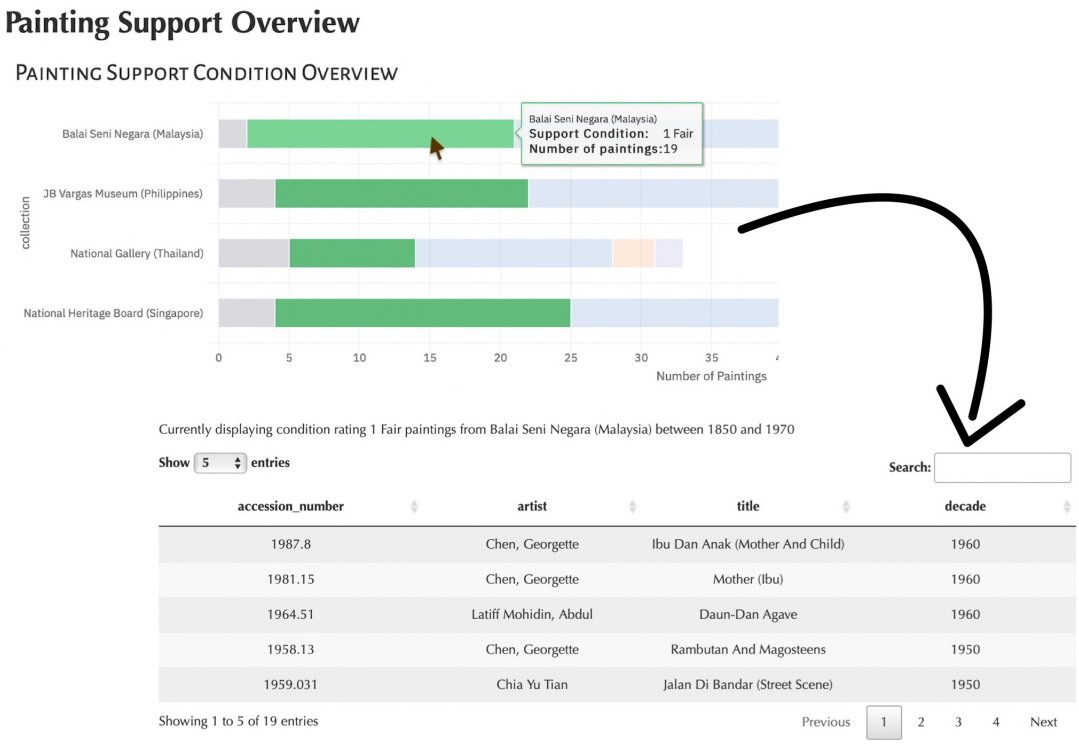


Figure 29: Click Event Example

An illustration of our approach to this problem is shown in Figure 29. Typically, when the mouse hovers above a specific section of the interactive plot, a tooltip will pop up and display some information about the selected part. However, with the introduction of reactivity and click events, information on the hovered section will be recorded when the user clicks on the hovered section. This effectively allows the dashboard application to filter the dataset and produce a summary table below the plot that contains all the paintings corresponding to the conditions of the selected section. Therefore, adding this functionality to our dashboard allows us to create visualisation without sacrificing the information of the original dataset. It also effectively groups paintings that belong to a given category, allowing users to query the data based on painting conditions.

5.2.7 Polished for Authentication

To ensure the confidentiality of the dashboard, we have utilised the polished package. The login page, as shown in Figure 30, was created by using the Polished package, which allows us to customise the colour styling and add logos on login page.

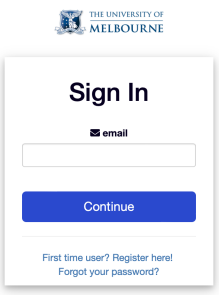


Figure 30: Login page

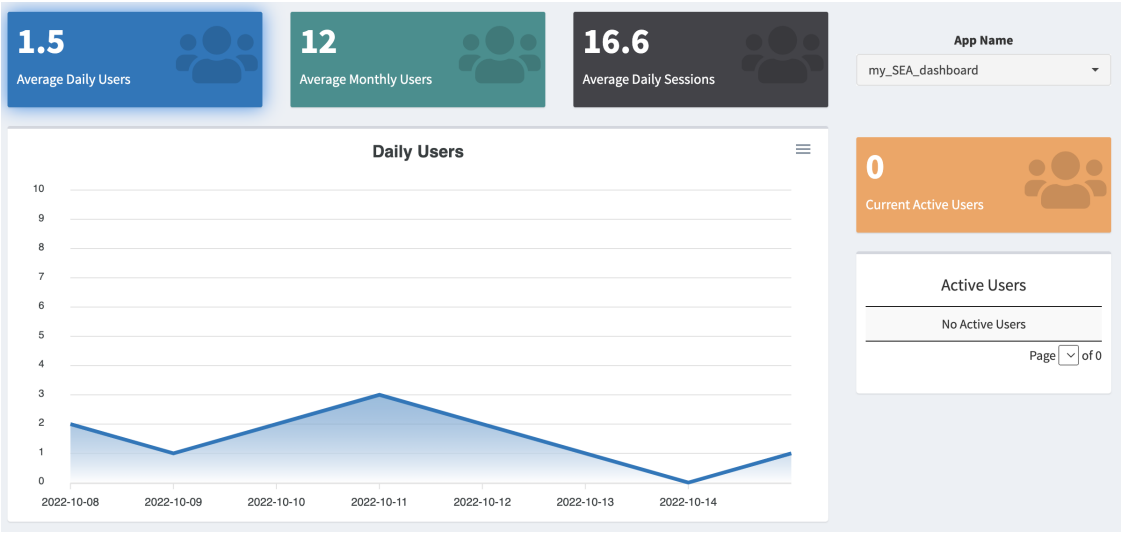


Figure 31: Polished Usage Monitor

Furthermore, the polished package also comes with an easy-to-use dashboard website to monitor users' usage as shown in Figure 31. This site has functionality allowing admin users to maintain user roles and authorisation of the application, as exhibited in Figure 32.

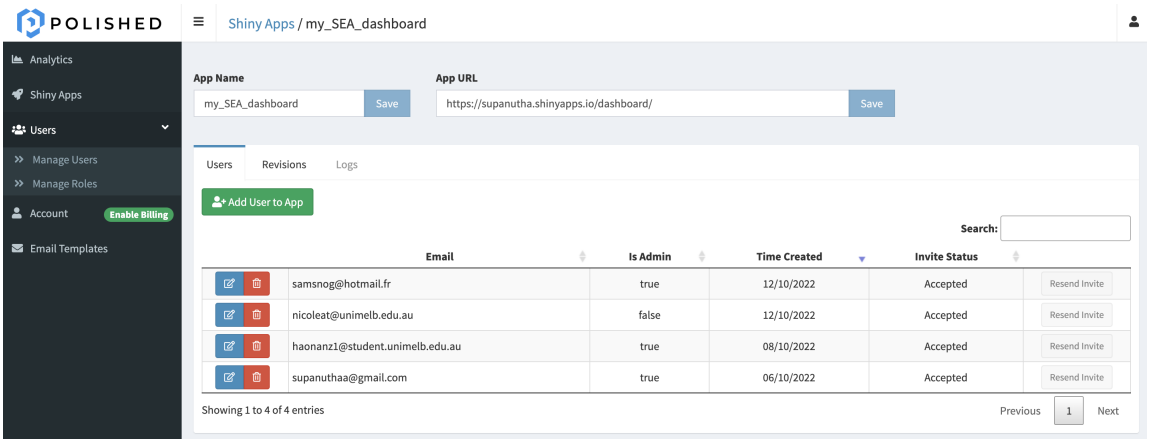


Figure 32: Polished User Management

5.3 Section Summary

In this section, we have mainly presented our approach and a summary of the constructed dashboard. Along the way, some challenges arose, but solutions were found. As an achievement, the client has shared the dashboard with museums involved in the study behind this project, and we have received beneficial feedback that is useful future improvements of the dashboard.

6 Independence Testing and Relationship Visualisation

As discussed in the data preprocessing section, given that all the painting condition attributes in the dataset are mostly categorical; therefore, it is impossible to run correlation analysis to identify the relationship between two selected categorical variables since, by definition, they cannot yield a mean, and we cannot compute the covariance between two categorical variables. For that reason, we've mainly applied three methods to test the relationship between the attributes.

6.1 Contingency Table Analysis with χ^2 Test of Independence and Log-linear Model

One method for dealing with categorical data is Log-linear model and χ^2 test of independence, which is a hypothesis testing method that is used to check whether the two categorical variables are associated or not. As with all other hypothesis testing, we've defined the null (H_0) and alternative (H_1) hypothesis as,

- H_0: There is no association between the two categorical variables
- H_1: There is association between the two categorical variables

We will first introduce the concept of test statistics and *p-value*. Test statistics measures the degree of agreement between a sample of data and the null hypothesis, while *p-value* tells us how likely the test statistics could have occurred under the null hypothesis. The general testing rule is if the test statistics are higher than the critical value, or the p-value is less than the significance level of 0.05. We would reject the null hypothesis and prefer the alternative.

	Indentation (No)	Indentation (Yes)	Total
Holes (No)	142	20	162
Holes (Yes)	30	16	46
Total	172	36	208

Table 1: Contingency Table of Indentation vs Holes (Observed)

To formulate the χ^2 test, we summarise the two selected categorical variables as a contingency table, an example was given in Table 1. Cells of the table represents the observation frequency fallen into the given combination. The test itself works by comparing the observation frequencies to the calculated expected frequencies under the null hypothesis using Pearson's χ^2 statistics, where the expected frequency of a cell and χ^2 test statistic is calculated as followed,

$$Expected_k = \frac{\text{Column Total} \times \text{Row Total}}{\text{Table Total}} \quad \text{and} \quad \hat{\chi}^2 = \sum_{k=1}^{n} \frac{(Observed_k - Expected_k)^2}{Expected_k}$$

If the corresponding *p-value* of $\hat{\chi}^2$ is less than the significance level of 0.05. In which the difference between the observed and expected distributions is statistically significant. We would then have sufficient evidence to reject the null hypothesis, claim the two variables are associated and vice versa. The test statistics $\hat{\chi}^2$ could also be approximated by using the residual deviance obtained from the Log-linear model, which effectively allows us to perform the test using more than two categorical variables.

6.2 Fisher's Exact Test

The problem with χ^2 test is that, for a reliable test, we need a large balanced dataset and the expected frequency of cells in the contingency table are required to be at least 5 [6]. But due to the size of the dataset, this assumption could not be satisfied.

An alternative method is to use the Fisher's exact test, where the test procedure are similar to the χ^2 test, comparing the *p-value* against the significance level. It is more effective than the χ^2 test when the expected frequencies and the dataset are small as suggested by related work, since Fisher's exact test does not rely on an approximation of the test statistics that becomes exact when the data size approaches infinity, instead

it directly assesses the null distribution by calculating the probability of getting the observed data using the hypergeometric distribution [6].

6.3 Mosaic Plot

Furthermore, we have used mosaic plot to visualise the relationship between two related categorical attributes. Mosaic plot is an effective way to visualise contingency tables. The general idea is to represent each cell of the table with a box, where the corresponding cell counts determine the size of the box. The height of each box is the proportion of individuals in the column which fall into that cell, while the width is the same for all boxes of the same column and is equal to the total count in that column. As an example, the mosaic plot for Table 1 is shown in Figure 33.

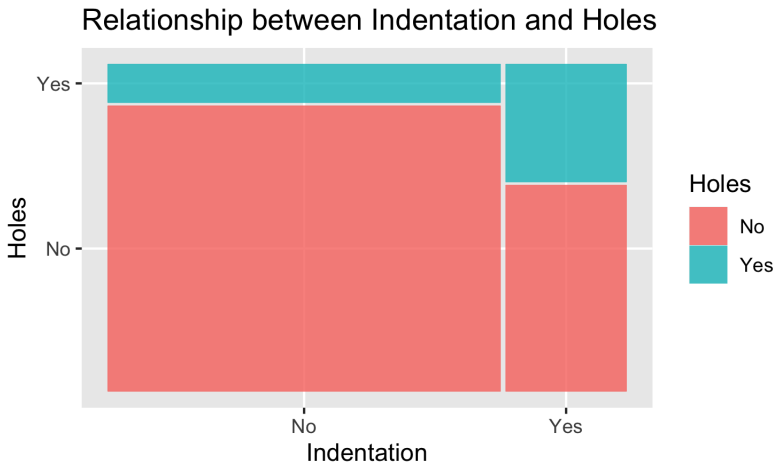


Figure 33: Example Mosaic Plot of Holes vs. Indentation

6.4 Limitation of the Test

During the exploratory data analysis, we noticed that some paintings were not assigned with a condition ratings. Therefore, paintings without an condition ratings on paint support, ground layer, and paint layer will be excluded from testing.

Furthermore, given paintings were sourced from four different museums, thus we would like to test the condition independence of each attributes given different museums. To do this, we've fitted the log-linear model for each combination of the categorical variables, and performed χ^2 test to determine whether the are likely to be related or not given different museum. The fitted log-linear model is given below, where `M`, `C1`, and `C2` is the museum, first condition attribute, second condition attribute, respectively. And λ_{ijk} is the cell observed frequency of the contingency table.

$$log(\lambda_{ijk}) = \mu + \mathtt{M_i} + \mathtt{C1_j} + \mathtt{C2_k} + (\mathtt{M} \cdot \mathtt{C1})_{ij} + (\mathtt{M} \cdot \mathtt{C2})_{ik}$$

However, due to the sparse nature and the small size of the dataset, zero cells have been a significant issue when constructing the table and fitting the model. Zero cells in the contingency table can be distinguished as sampling zeros and structural zeros. The former is due to sampling variability, and the latter are cells known to have zero values. Thus, in our case, the zeros can be treated as structural zeros, given that the cell values are simply missing because it has no paintings from the sample population that fit into the category.

One solution will be testing quasi-independence, where zero cells will be excluded from analysis when testing independence. However, when the table was overcrowded with zero cells, there was not enough data to provide a trustworthy model fit and test result, where the model will be saturated as the residual deviance and degrees of freedom were reduced to zeros, causing overfitting. Therefore, the results discussed only considers the painting collection as a whole, without the interaction of museums. Last but not least, it is important to note that the test of independence only tells whether two attributes are related to one another, and it does not necessarily imply any causal effect of a variable to another.

6.5 Testing Algorithm Description

We have implemented an algorithm for testing pairwise independence between the attributes of each layer, utilising the hypothesis testing function of R. The algorithms starts with the first attribute v_1 of the input dataframe that consists of k attributes to be tested, then it's relationship with the other $k-1$ attributes of the dataframe were then tested using Fisher's Exact test.

The obtained *p-value* of each test were then used to assess the two attribute's relationship. If the *p-value* is less than the significance level 0.05, we will reject the null hypothesis and conclude that the two are associated, vice versa. We repeat the steps above until each combination is tested. The pseudo-code of our algorithm is shown in Algorithm 1.

Algorithm 1: Independence testing for pairwise attributes

```
Input: Data frame consists of testing attributes
Output: Data frame consists of testing results
attributes ← colNames(Input);
for v_i in attributes do
    for v_j in attributes do
        if v_i = v_j then
            do nothing;
        else
            table ← ftable(v_i, v_j);
            p-value ← fisher.test(table);
            if p-value < 0.05 then
                v_i and v_j is associated;
            else
                v_i and v_j is independent;
            end
        end
    end
end
```

6.6 Result and Visualising Relationships

6.6.1 Relationship Between Paint Support, Ground Layer and Paint Layer Ratings

Firstly, we have tested the association between the paint support, ground layer and paint layer condition ratings. As mentioned in the introduction, these three components were stacked on top of each other in a painting. Therefore, we are interested to see if these conditions are closely related with each other.

Attribute 1	Attribute 2	*p-value*
Paint Support Condition	Ground Condition	2.907e-06
Paint Support Condition	Paint Layer Condition	2.661e-06
Ground Condition	Paint Layer Condition	5.658e-06

Table 2: Fisher's Exact Test Result of the Three Condition Ratings

As we can see from the result table 2 above, the low p-values suggest the three condition ratings were highly associated with each other. The mosaic plot shown in Figure 34 depicts the relationship between paint support condition ratings and paint layer condition ratings, we can see a clear trend between the two condition ratings, the proportion of paintings having good and excellent ground layer condition rating increases as the paint support condition rating increases from poor to excellent. While the proportion of paintings with poor and fair ground layer condition rating gradually declines as paint support condition rating increases.

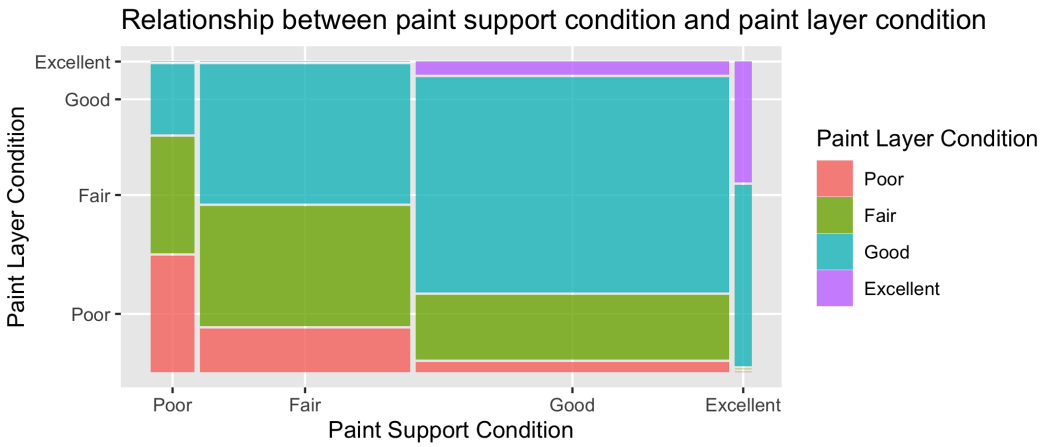


Figure 34: Relationship Between Paint Support Condition and Paint Layer Condition

It is no surprise that a similar trend could also be seen with the other two combination, as shown in Figure 35 and Figure 36.

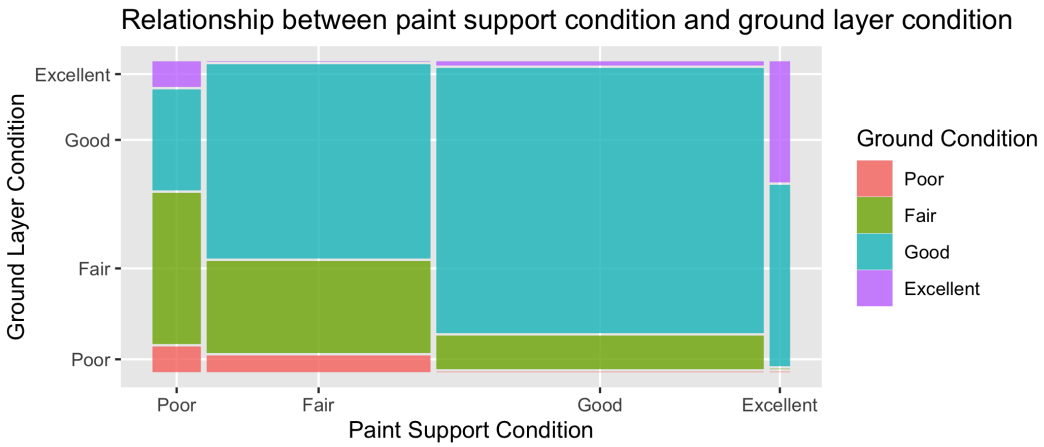


Figure 35: Relationship Between Paint Support Condition and Ground Condition

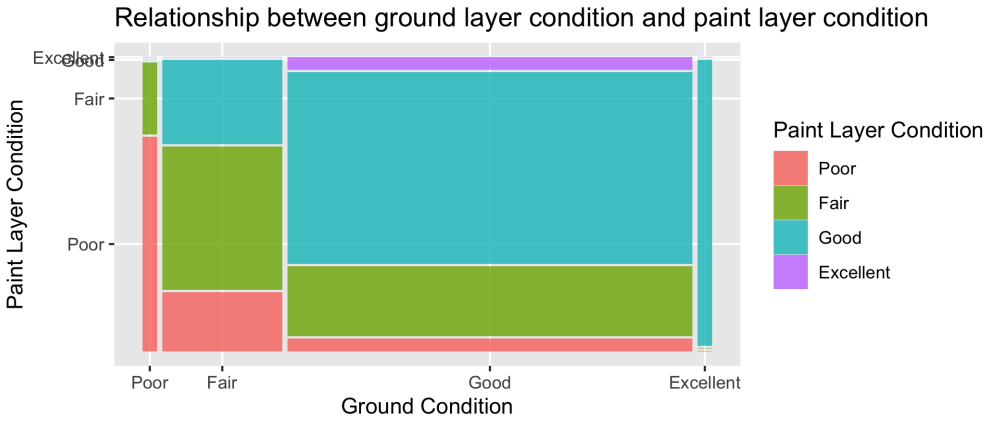


Figure 36: Relationship Between Ground Condition and Paint Layer Condition

6.6.2 Relationship Between the Paint Support Condition Attributes

This section will discuss the independence testing result between the paint support condition attributes. The testing results are presented using a tile plot as shown in Figure 37, which tries to mimic the correlation heatmap. For those whose p-value returned by the Fisher's exact test is greater than the significance level of 0.05 were labeled as blue tiles, indicating that the two condition attributes were independent. In contrast, the green tiles indicate the two attributes were related.

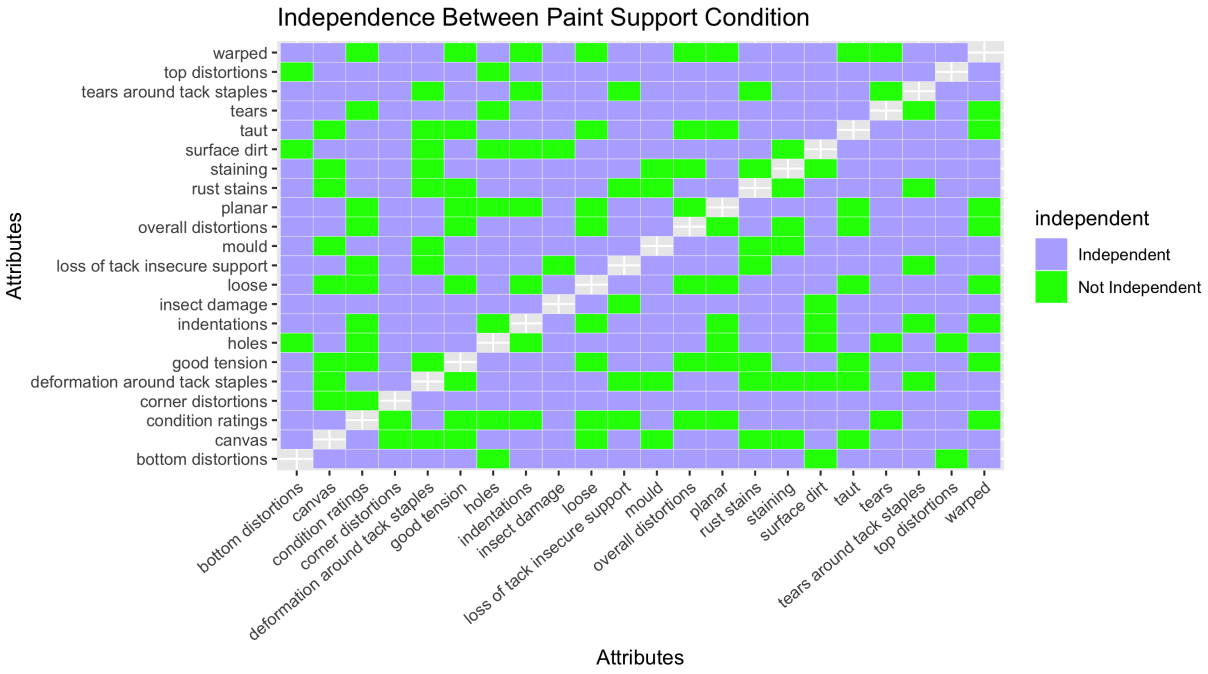


Figure 37: Independence Between Paint Support Conditions

As tile plot shown in Figure 37, we can see that the condition ratings of paint support are associated with corner distortion, good tension, holes, indentations, loose, loss of tack insecure support, overall distortions, planar, tears, and warped. Additionally, we tested relationship between the used canvas material and resulted condition of paint support to check if there is an effect to condition ratings due to the use of different materials. However, there's no statistical evidence that the two are associated. Next, we further analysed the relationship between condition ratings and the associated attributes by plotting them against

each other using the mosaic plot.

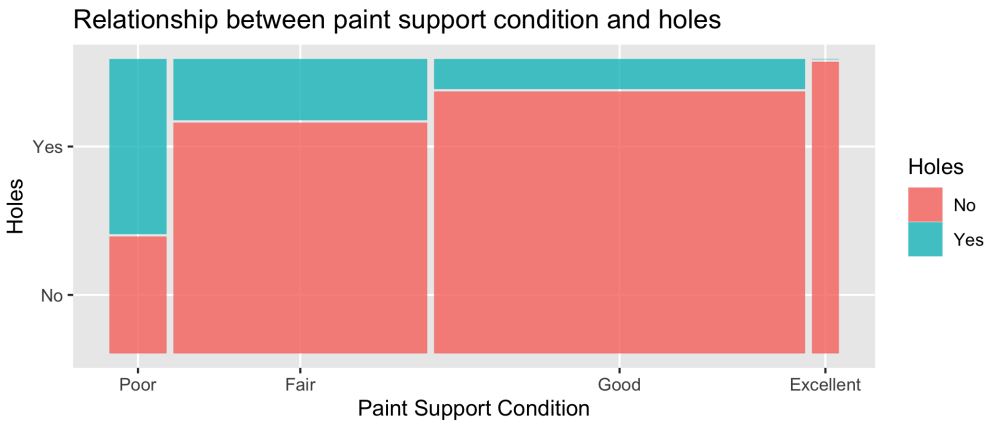


Figure 38: Relationship Between Paint Support Condition Ratings and Holes

As Figure 38 illustrates, there is a clear distinction between different support condition ratings, the proportion of paintings with no holes grows as the condition ratings increases from poor to excellent. And it is worth noting that most of the paintings were classified with fair and good condition, while poor and excellent only account for a small proportion. On top of that, similar relationship with support condition ratings could also be identified with tears, loose, corner distortion, overall distortion, indentation, and warped.

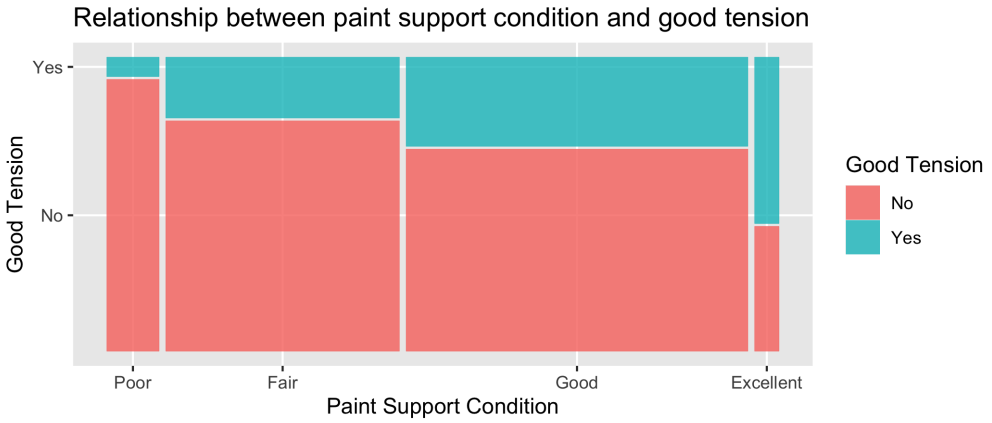


Figure 39: Relationship Between Paint Support Condition Ratings and Good Tension

As for the relationship between paint support condition and good tension shown in Figure 39, we can see a clear differentiation with the previous plot, it is no surprise that, as the support condition ratings increases from poor to excellent, the proportion of paintings with good tension also increases. Although the Fisher's exact test did not show that the support condition ratings were associated with taut, but similar trend could also be seen with the two.

Here we further investigate the relationship between insect damage and surface dirt in paint support shown in Figure 40. It is worth noting that when a painting has no insect damage, there is an even chance to have surface dirt present. On the other hand, when there is insect damage present in a painting, likely, surface dirt will also appear. However, we need to mention that paintings with insect damage only account for a small proportion of the sample population. To make the conclusion, we may need more samples to reduce the non-response bias.

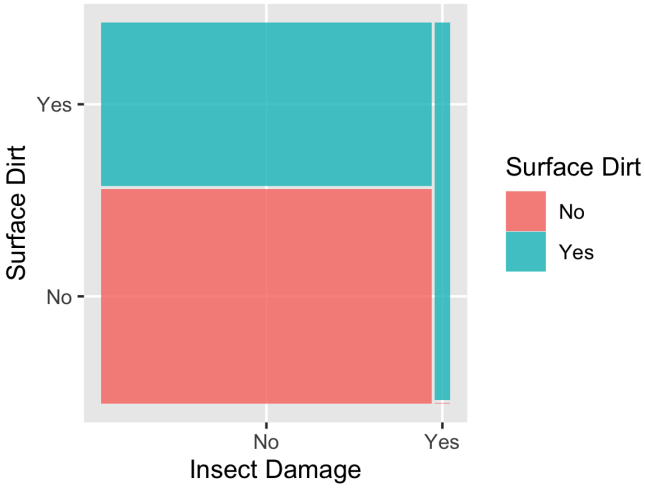


Figure 40: Relationship Between Insect Damage and Surface Dirt

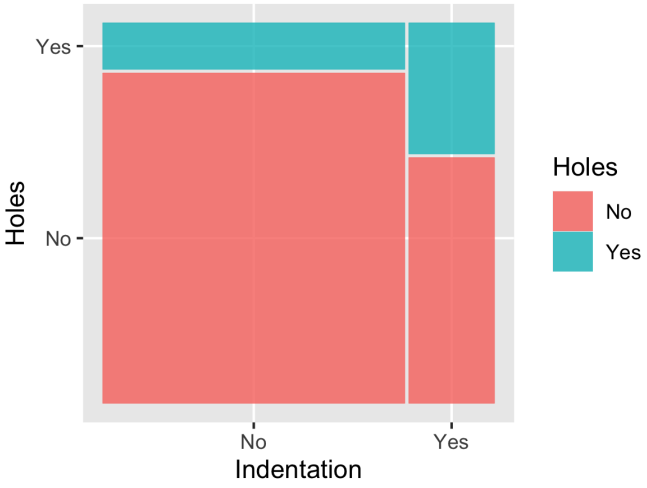


Figure 41: Relationship Between Indentation and Holes

As for the association between indentation and holes shown in Figure 1, a painting with indentation will also have a higher chance to have holes than those paintings without indentation. In fact, after estimating odds, under the study condition, paintings with indentation having holes are almost 3.78 times higher than for paintings with no indentation.

6.6.3 Relationship Between the Ground Layer Condition Attributes

This section will discuss the associations between the ground layer condition attributes. Again, the outcome of the Fisher's exact test will be presented using the tile plot.

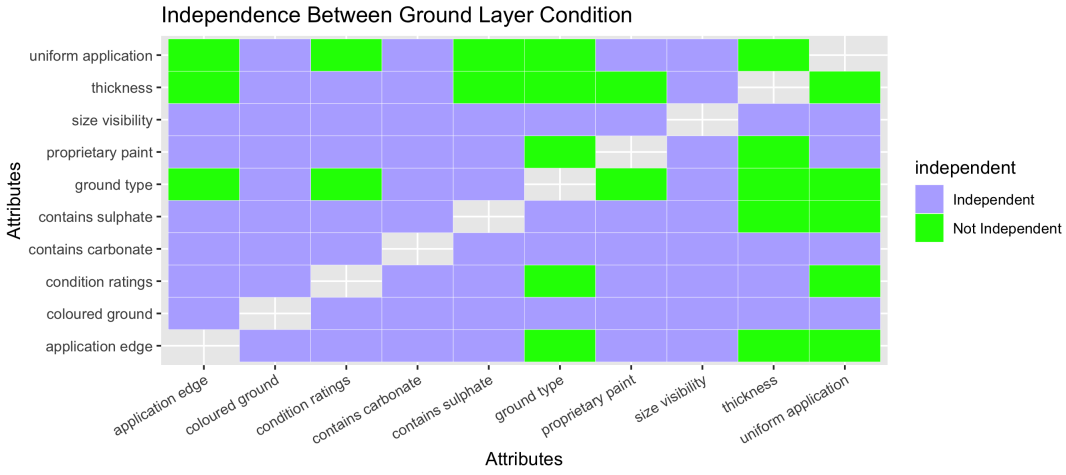


Figure 42: Independence Between Ground Layer Conditions

As Figure 42 depicts, there are only two attributes that are associated with the ground condition ratings, namely, ground type and uniform application. Figure 43 shows the relationship between different ground type and the ground condition ratings. As mentioned in the introduction, due to material shortage in early 20th century, local artists without access to commercial materials starts to source their own material instead. Here we can see that commercial ground tend to have higher proportion paintings being classified as good and excellent. After computing the odds ratio, we found that the odds of a painting having a good

or excellent condition in the commercial ground group is about two times higher than in the artist applied ground group.

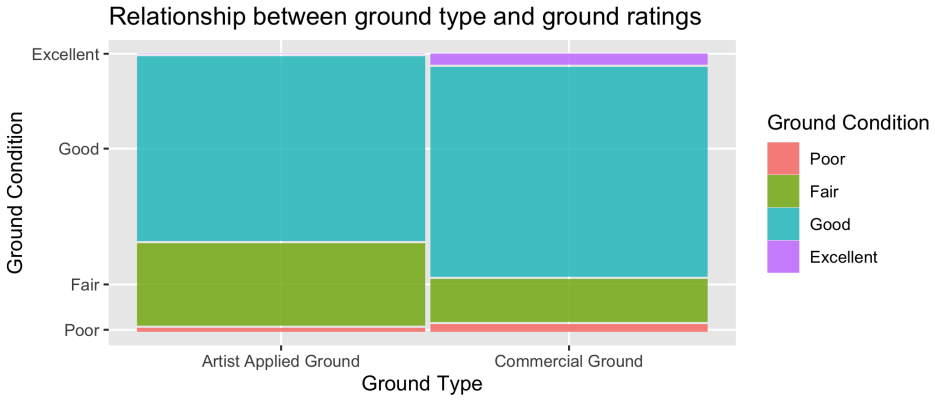


Figure 43: Relationship Between Ground Condition and Different Ground Type

Moreover, we can notice that the ground types are associated with the thickness of the applied ground. As we can see from Figure 44, the proportion of the ground being thickly applied for artist applied ground is almost three times that on commercial ground.

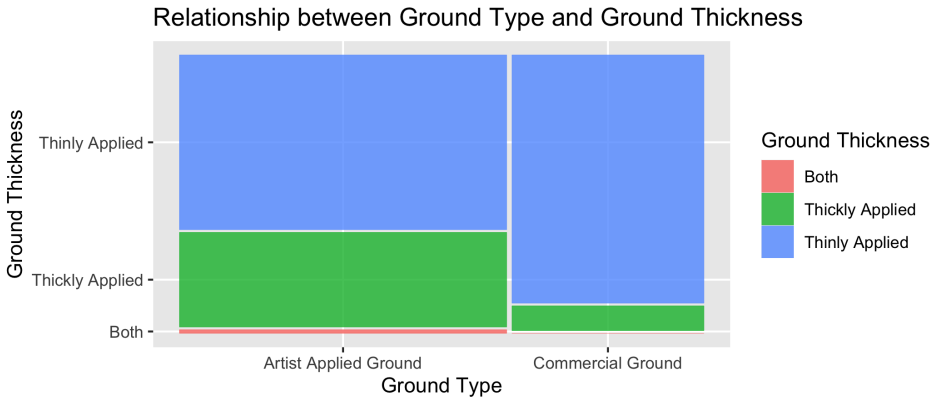


Figure 44: Relationship Between Ground Type and Ground Layer Thickness

Since the ground type and uniform application are related to each other. We can see that there is a clear contrast between paintings with artist applied ground and paintings with commercial ground in Figure 45, as paintings with commercial ground are about seven times more likely to have their ground to be uniformly applied on than the one with artist applied ground.

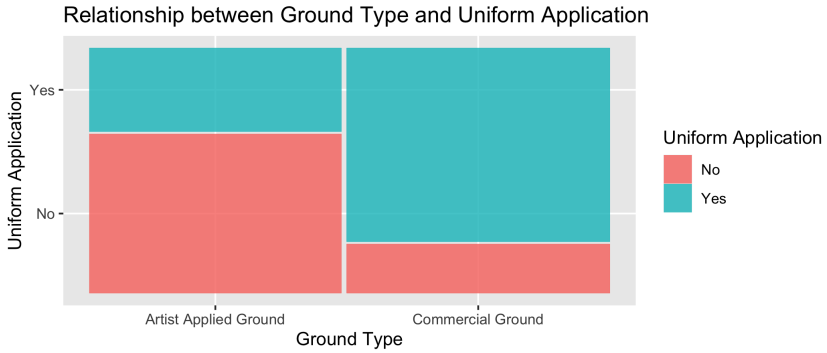


Figure 45: Relationship Between Uniform Applied Ground and Ground Type

6.6.4 Relationship Between the Paint Layer Attributes

In this section, we will further testing the association between the paint layer attributes. As per previous independence testing, the testing results were presented as tile plot, where green tile indicates association, while blue tiles indicates independence.

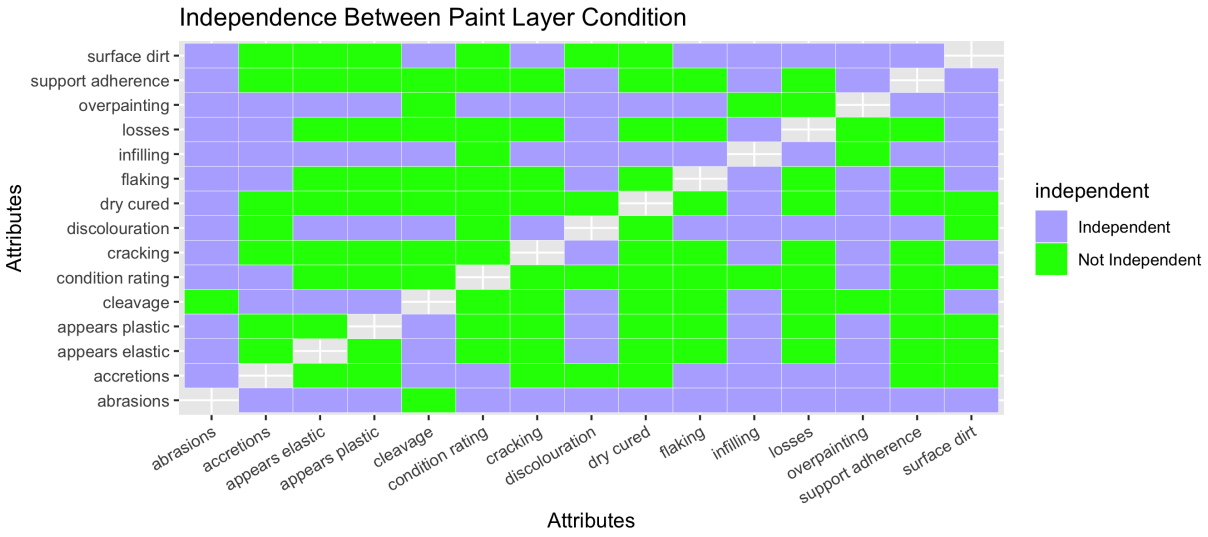


Figure 46: Independence Between Paint Layer Conditions

As we can see from Figure 46, the test results suggests that paint layer condition ratings are associated with elasticness and plasticness of the layer, cleavage, cracking, discolouration, dry cured, flaking, infilling, losses, support adherence, and surface dirt of the layer.

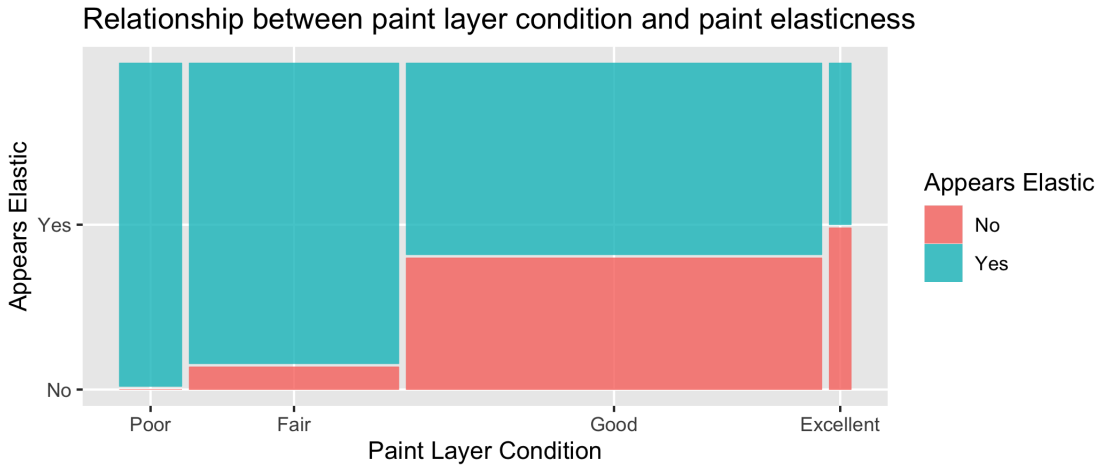


Figure 47: Relationship Between Paint Layer Condition and Paint Elasticness

Figure 47 visualises the relationship between paint layer condition and the elasticness of the painting. The height of the box shows the conditional proportion of paintings with elastic paint layer given the paint layer condition ratings, the plot clearly brings out paintings with poor layer condition will highly likely to appear elastic. Nevertheless, a clear trend could be seen as the proportion of paintings that appears to be elastic decreases as the paint layer condition ratings increases from poor to excellent, where the proportion of paintings does not appear elastic is considerably higher when condition ratings is excellent.

And it is worth pointing out that similar relationship with the layer condition can also be seen on cleavage,

paint cracking, paint discolouration, dry cured, flaking, losses, and surface dirt of the layer.

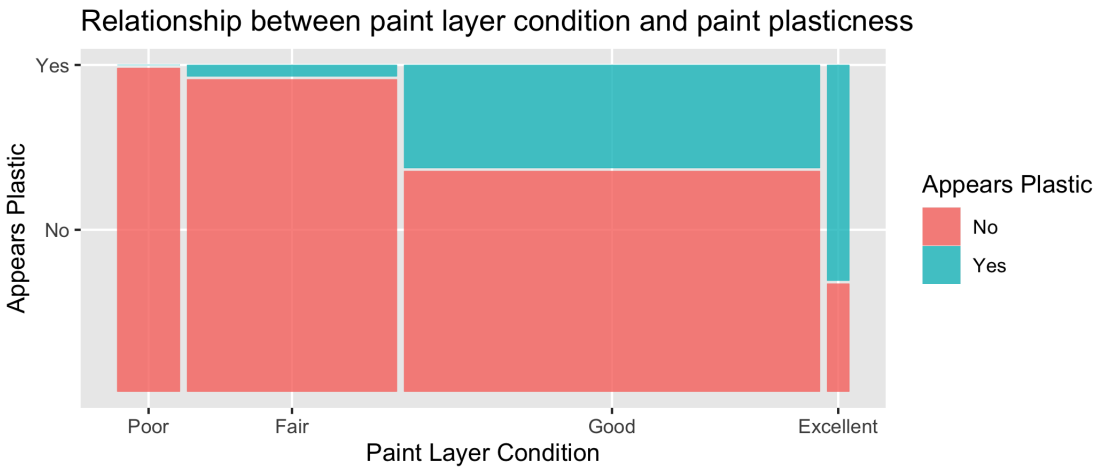


Figure 48: Relationship Between Paint Layer Condition and Paint Plasticness

On the other hand, it is a complete different story with plasticness of the paint. Figure 48 here visualises the relationship between paint layer condition and the plasticness of the painting, where we can see an decreasing trend of the proportion of paintings that does not appears to be plastic as the condition ratings increases from poor to excellent. And it is to be noted that similar relationship with the layer condition ratings could also be seen with support adherence of the paint layer.

Furthermore, we aim to explore the relationship between some of the condition attributes that are non-independent, especially the relationship between cracking and it's associated attributes, as the client suggests.

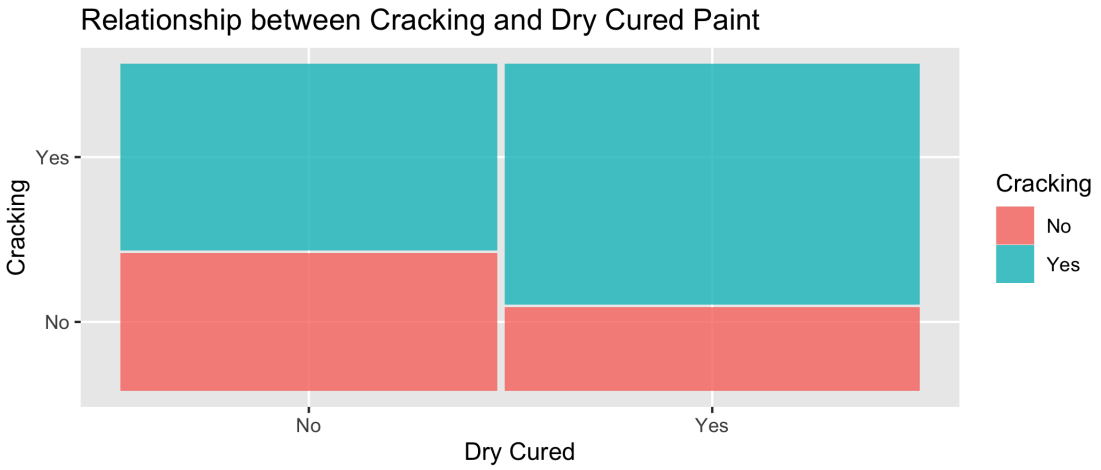


Figure 49: Relationship Between Dry Cured and Paint Discolouration

Figure 49 above visualises the relationship between dry cured and paint crack. The height the boxes shows the condition proportion of paintings with paint crack given whether not the painting has dry cured condition. As we can see, under the studied samples, if the painting has dry cured condition, there is a higher chance it will has paint crack than those without. In fact, after computing the odds ratio, the odds of having paint crack for a painting with dry cured condition are about 2.5 times higher than paintings without dry cured condition.

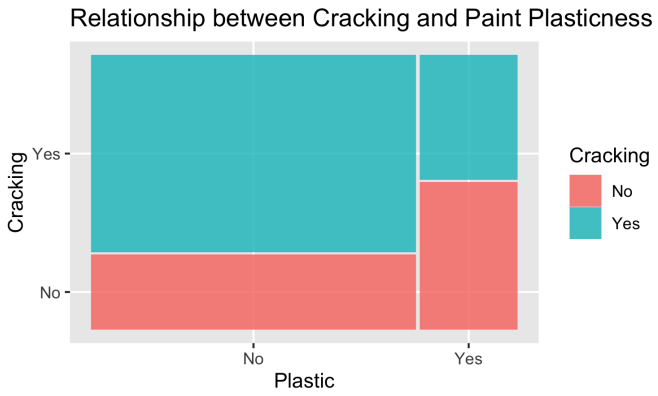


Figure 50: Relationship Between Paint Crack and Paint Plasticness

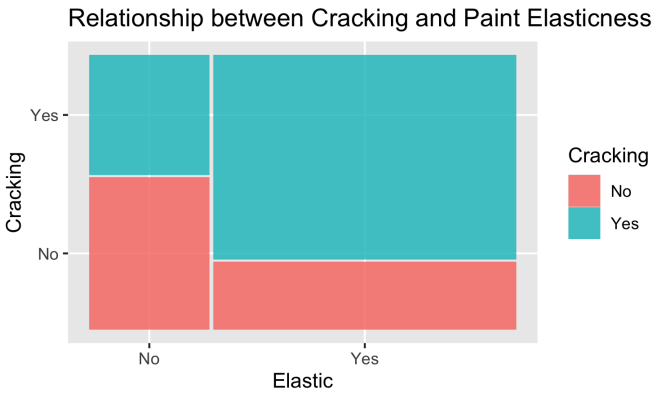


Figure 51: Relationship Between Paint Crack and Paint Elasticness

To assess the relationship between paint cracking and paint plasticness, Figure 50 depicts, paintings without signs of plasticness has higher proportion of cracking present. While the proportion of cracking is much lower among those paintings with plasticness present. In contrast, the relationship of paint crack and paint elasticness is complete opposite as shown in Figure 51.

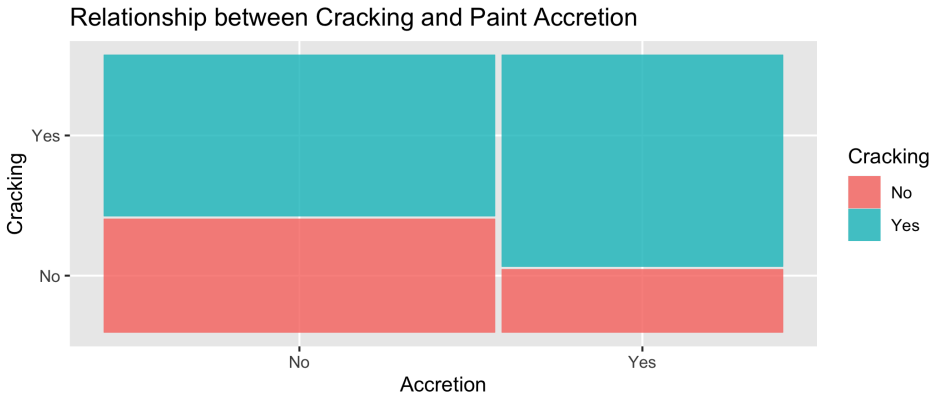


Figure 52: Relationship Between Paint Cracking and Paint Layer Accretion

Moreover, the relationship between paint cracking and paint layer accretion was shown in Figure 52. It can be seen from the figure that the proportion of paintings with paint cracks are moderately higher among those with accretion in paint layer compare to those does not have accretion.

6.7 Section Summary

In this section, we presented the independence testing result of pairwise condition attributes for the three layers, namely, the paint support layer, the ground layer, and the paint layer. Additionally, mosaic plots were used to visualise the relationship between the dependent attributes. Unfortunately, we cannot test the association for higher-order contingency tables due to the small size of the dataset, some cells in the constructed contingency tables were not captured by the sample population, causing the Log-linear model to become saturated, producing an unreliable fit to the data. Additionally, throughout the discussion, we have been quite conservative in trying not to deduce the relationship between the attributes, as dependence does not imply causality. Therefore, a more detailed analysis, such as cohort and case-control studies, would be required to further establish causality between the condition attributes.

7 Predictive Modelling of Missing Values

In this section, we will be focusing on covering our approach to use machine learning models to predict some of the missing values from the dataset. More specifically, we tried to predict missing values for three features: `ground_layer_application`, `ground_layer_limit` and `ground_layer_thickness`. The first one specifies if the ground layer was applied by the artist or if it was already applied in an industrial way when the artist bought the canvas. The second feature states if the ground layer is only applied on the visible part of the canvas or if it goes to the side edges as well. Finally, the third one states if the ground layer is thick or not.

7.1 Data Preprocessing for Prediction

7.1.1 One-Hot Encoding

As discussed in the data cleaning section, most of the categorical attributes record the presence of a particular paint condition. Hence, these attributes were treated as binary attributes, corresponding to the presence (1) or absence (0) of a condition. Moreover, the One-Hot encoding method was applied to categorical variables with more than two levels, where one new binary attribute is added for each level in the original variable.

7.1.2 Mutual Information for Feature Selection

Once all the features we are interested in were either numerical or binary (we did not transform the Accession numbers or the artists with One-hot encoding), we then compute the mutual information (MI) between the features and the target variable. Mutual information is used to identify features that are well related to the target variable, which is based on the idea of knowing a variable will let us predict the target with more confidence, a higher MI score means higher dependency between the target and features.

For each of the three features we tried to predict, we selected the 20 features that presented the highest mutual information score. We then fed several algorithms with these values.

7.1.3 Data Splitting

Supervised machine learning algorithms require the data to be splitted into a training and a testing set. We used `train_test_split` from the `sklearn` package to do it. However, by looking at the distribution of the data in Figure 53, we could see the distribution of `ground_layer_limit` and `ground_layer_thickness` were biased and splitting at random could lead to an even more over-representation of the dominating category in the training set. This is why we used the `stratify` option when splitting. It consists in making sure the training set has all the possible categories represented with the same ratio as in the original population.

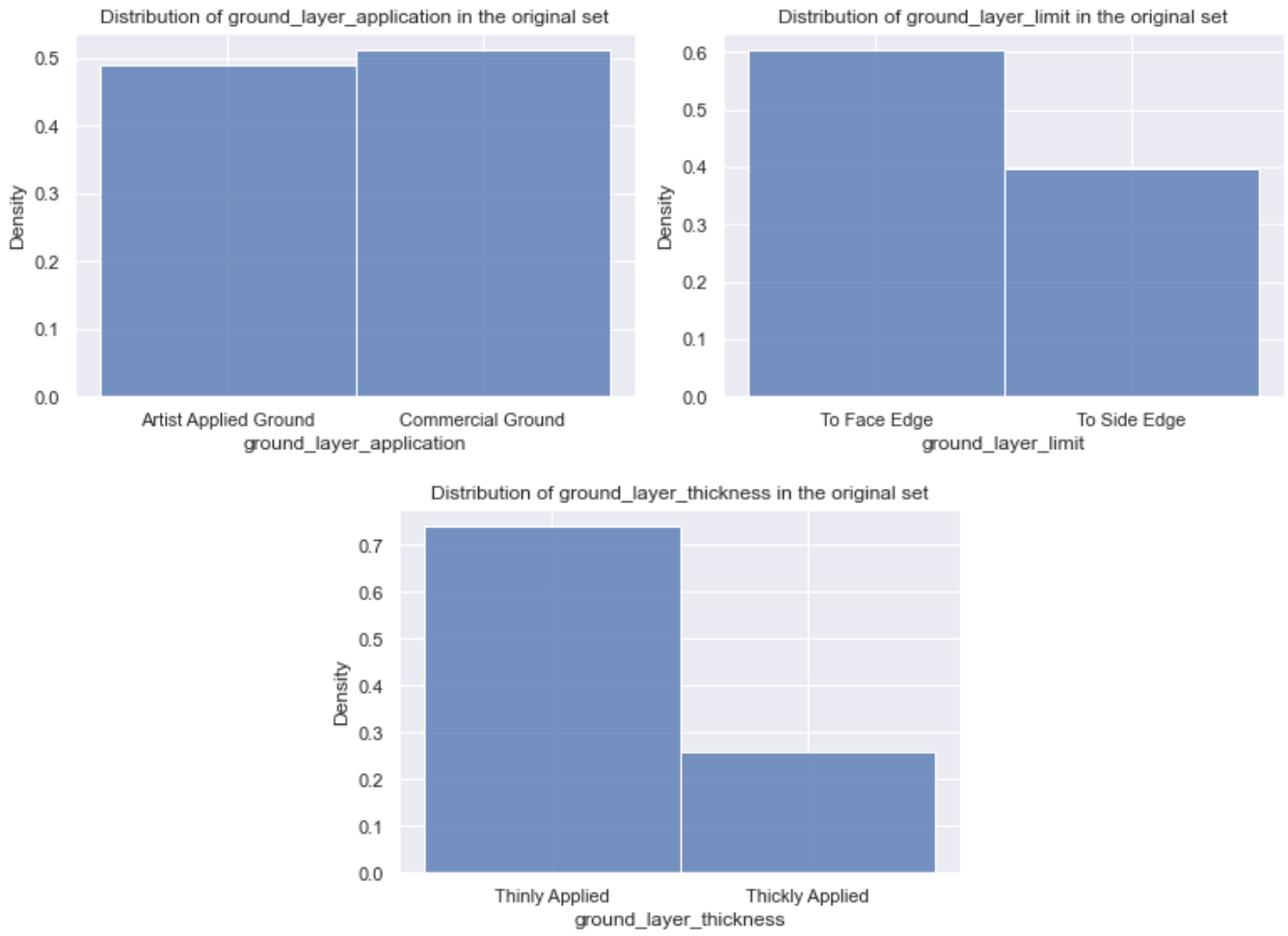


Figure 53: Distribution of each data set before the train/test split

7.2 Predictive Modelling

We have chosen to use a Bernoulli Naive Bayes classifier as the baseline model. We hope that we would be able to improve the prediction performance as we introduce more advanced machine learning models. Given the small size of the dataset, neural network models were not used due to their high complexity; as we do not have enough data to produce an adequate fit, models will incorporate the noise of the samples leading to overfitting (high variance).

7.2.1 Bernoulli Naive Bayes Classifier

Bernoulli Naive Bayes is a classification model based on the Bayes Theorem. It assumes all attributes are generated from a *Bernoulli* distribution, which is similar to our dataset, where 1 indicates the presence of a condition in the painting and 0 indicates the absence. Training is done by computing the prior probability of each class and the conditional probability of each attribute given the class. Classification is done by calculating the probability of a sample being each class, and the class with the highest posterior probability is the prediction outcome, and can be represented by the equation below. Although it is important to note that this classifier has a strong assumption of the attributes do not interact.

$$\hat{c} = \arg\max_{c_j \in C} \mathrm{P}(c_j) \prod_i \mathrm{P}(x_i \mid c_j)$$

7.2.2 Random Forest Classifier

Random Forest is an ensemble learning method for classification, it works by constructing multiple decision trees using different sets of bootstrap data (randomly sampled dataset, with replacement), which aims to reduce the correlation among the trees, so that trees can protect each other from their individual errors. Classification is done by aggregating the prediction of each individual tree, and the prediction outcome is the class with the highest vote.

7.2.3 Logistic Regression

A logistic regression is a statistical model which attempts to model the cumulative distribution function of a binary variable Y. It does it by assuming, if we are considering N features as predictors, there exist a vector of parameters $\boldsymbol{\beta} = (\beta_i)_{0 \leq i \leq N}$ such that for any set of features $\mathbf{X} = (1,\ x_1,\ \ldots,\ x_N)'$:

$$p(Y = 1) = \frac{1}{1 + \exp\left(-\boldsymbol{\beta}'\mathbf{X}\right)}$$

Unlike linear regression, there are no closed form solution for the parameters $\boldsymbol{\beta}$. Thus, parameters here are estimated by the method of coordinate descent.

7.2.4 Stacking

The last model that we implement is the stacking ensemble classifier, where the predictions of the previous classifier are given as input features for the meta-classifier. And the final prediction is done via majority voting.

7.3 Model Evaluation and Suggestion

7.3.1 Hyperparameter Validation

The logistic regression could take an extra argument when called: C, the inverse of the regularization strength. To find the best value for this parameter, we fitted several logisitc regressions for each of the features we wanted to predict with its complete non empty data set (the union of the training and the testing set). Each regression had a specific value of C that spaned from 0.01 to 100. For each of these models, we compared the training accuracy and we selected the value of C which yielded the best accuracy. Once this value selected, we fitted a new logistic regression with the training set only.

Regarding the random forest, we had the freedom to specify several arguments: the number of trees, the maximum depth of each tree, the criterion to measure the quality of a split and the maximum number of features to look at when deciding for the best split. Choosing the best value for these tuning parameter was done by doing a grid search. This simply consists on testing every possible combination within a pool of potential values we gave as extra input.

7.3.2 Evaluation

We can see in table 3 the summary results of our models. The accuracies are overall pretty satisfying (all are above 0.75) but the testing sets are pretty small as we can see in the table 4. This makes us think the holdout accuracies cannot be used as a base for a general result. Moreover, we can see for the logistic regression on `ground_layer_limit` the holdout accuracy was greater than the training accuracy. This is

not that surprising considering the small size of the training and the testing datasets. It is especially true for this feature because the number of unspecified category was even larger than the size of the training set with only 81 training data points and 35 testing ones for 91 to predict.

		Training accuracy	Holdout accuracy
ground_layer_application	Naives Bayes	0.852	0.821
	Random Froest	1.0	**0.839**
	Logistic Regression	0.852	0.821
	Stacking	0.914	**0.839**
ground_layer_limit	Naives Bayes	0.815	0.829
	Random Froest	0.975	0.829
	Logistic Regression	0.852	**0.886**
	Stacking	0.922	0.821
ground_layer_thickness	Naives Bayes	0.830	0.769
	Random Froest	0.989	**0.846**
	Logistic Regression	0.875	0.795
	Stacking	0.914	0.821

Table 3: Model Performance

	Training set	Testing set	To predict
ground_layer_application	128	56	24
ground_layer_limit	81	35	91
ground_layer_thickness	88	39	80

Table 4: Size of the Different Datasets

The previous paragraph was focusing on the holdout accuracy. However we know this metric can be oversimplifying, this is why we kept trace the confusion matrices for every model and left them available in the appendix. We did not know enough domain knowledge to gain much insight with the matrices, this is why we are not doing some deep analysis of these matrices. Moreover, this part was more of an exercise for us since it was not requested by the client at all. We were still able to present the results to the client but they were of no interest for her, since it is not the main focus of the project.

8 Conclusion and Future Work

In this report, we have covered our approach and discussed the outcome on data preprocessing, interactive dashboard, independence testing for categorical data, and predictive modelling for missing values. As an achievement, our developed dashboard has been shared to the participating museums of the original study. Now both the client and the museums have access to the dashboard, they can base new work on it, trying to add new features for instance, if a new team happens to work on this project.

Regarding the independence testing result of the painting condition attributes, we are unable to deduce the causal effect between the conditions due to the limitation of the test as mentioned earlier. Therefore, one might consider to implement case control or cohort studies to further establish the causal relationship between two attributes. As for the models trained for missing value predictions, since the models were trained using limited samples, the data might not be a general representation of the overall population.

We would not suggest using them for future predictions as the models could be highly biased. For future reference, if more samples can be included, models can be retrained to yield better generalisation.

We have proposed last semester to map the relationship between artists and their commonly used painting materials. However, as we explored the dataset, most of the artists only contributes one to two paintings to the study. Thus, if more paintings from the artists were included, we could perhaps develop the relationship and connection between the artists and the used materials.

Lastly, we have also proposed to identify the relationship between tropical climate and painting conditions. However, since the data was collected almost twenty years ago no detailed storage climate history could have been sourced. Therefore, for future references, if more detailed storage climate data can be found, the team working on this project can try to measure how different climate conditions affect paintings.

9 Appendix

9.1 Algorithms & Pseudo-code

Algorithm 2: fuseCategColumns

Input: D_I: Dataframe of manually cleaned data

L_{index} : A list of column indexes

colName: name of the feature

Output: D_O: a single column Dataframe

D_O := empty dataframe with one column named after colName;

$D_O \leftarrow D_I[:, L_{index}[0]$;

for *index in* $L_{index}[1:]$ **do**

$NewCat := D_I[:, index]$;

$nonEmptyNew$:= the boolean serie telling if the cells of $NewCat$ are different from "";

$nonEmptyExisting$:= the boolean serie telling if the cells of D_O are different from "";

if *colName = "ground_layer_application"* **then**

$D_O[(nonEmptAddCat)\&(nonEmptExistingCat)]$ = "unsure";

else

$D_O[(nonEmptAddCat)\&(nonEmptExistingCat)]$ = "both";

end

$D_O[(nonEmptAddCat)\&\neg(nonEmptExistingCat)] = NewCat[\neg(nonEmptExistingCat)]$;

end

Return: D_O

Algorithm 3: fuseOrdinalColumns

Input: D_I: Dataframe of manually cleaned data

L_{index} : list of the indexes of the columns we want to fuse, from the lower level to the highest

colName: name of the feature

Output: D_O: a single column Dataframe

D_O := empty dataframe with one column named after colName;

$level := 0$;

Fill D_0 with 0s;

for *index in* $L_{index}[1:]$ **do**

$level \leftarrow level + 1$;

$existsInOrigin$:= the boolean serie telling if the cells of $NewCat$ are different from 0 and "NaN";

$NaNInOrigin$:= the boolean serie telling if the cells of $NewCat$ are "NaN";

$nonZeroExisting$:= the boolean serie telling if the cells of D_O are different from 0;

$(D_O[(existsInOrigin)\&(nonZeroExisting)], D_O[(existsInOrigin)\&\neg(nonZeroExisting)]) \leftarrow \left(\lceil \frac{D_O[(existsInOrigin)\&(nonZeroExisting)]+level}{2} \rceil, level\right)$;

$D_0[NaNInOrigin] \leftarrow Nan$;

end

Replace 0 values of D_0 by NaN ;

Substract 1 to all the values of D_0;

Return: D_O

9.2 Confusion Matrices for the Predictive Modelling

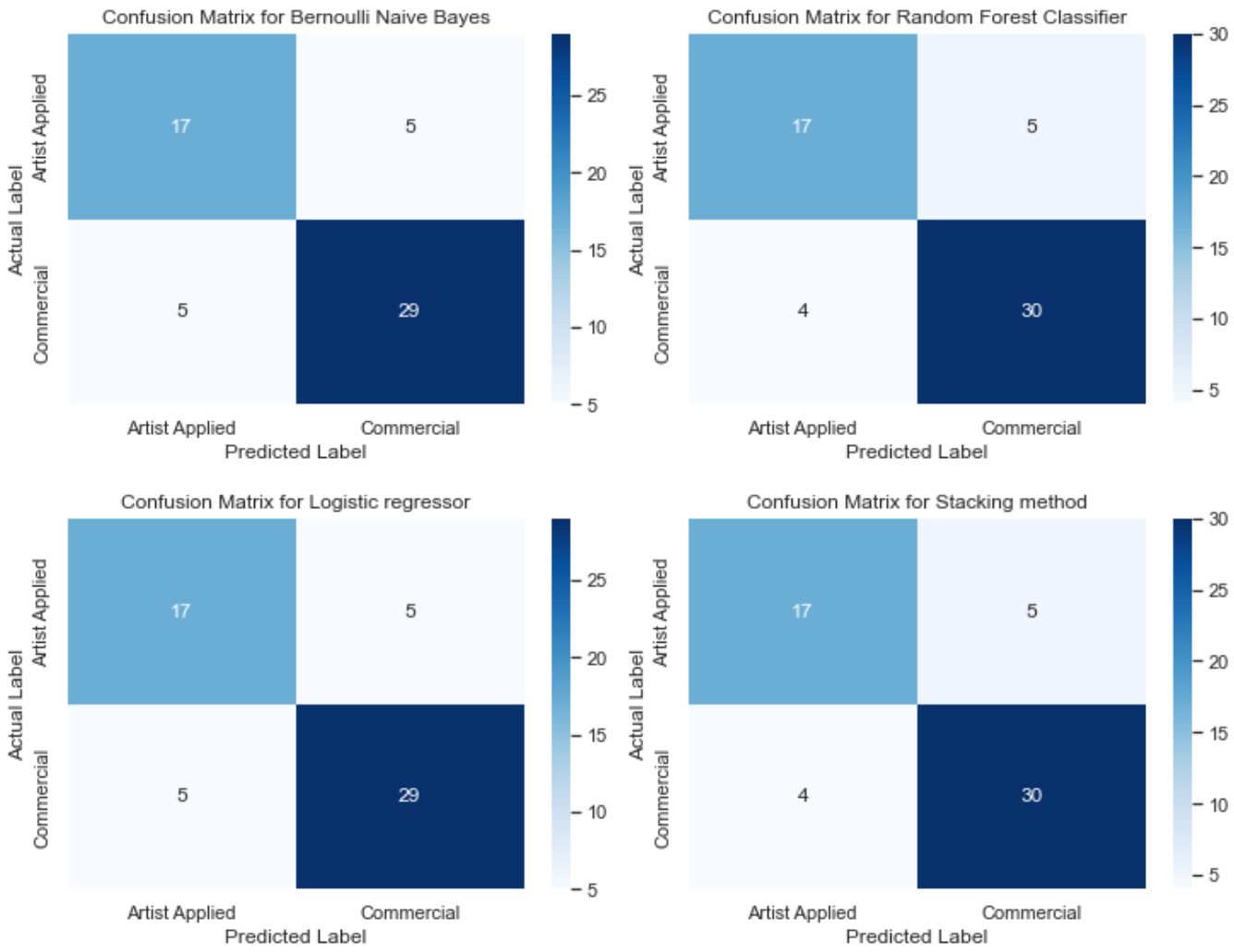


Figure 54: Confusion matrices for `ground_layer_application`

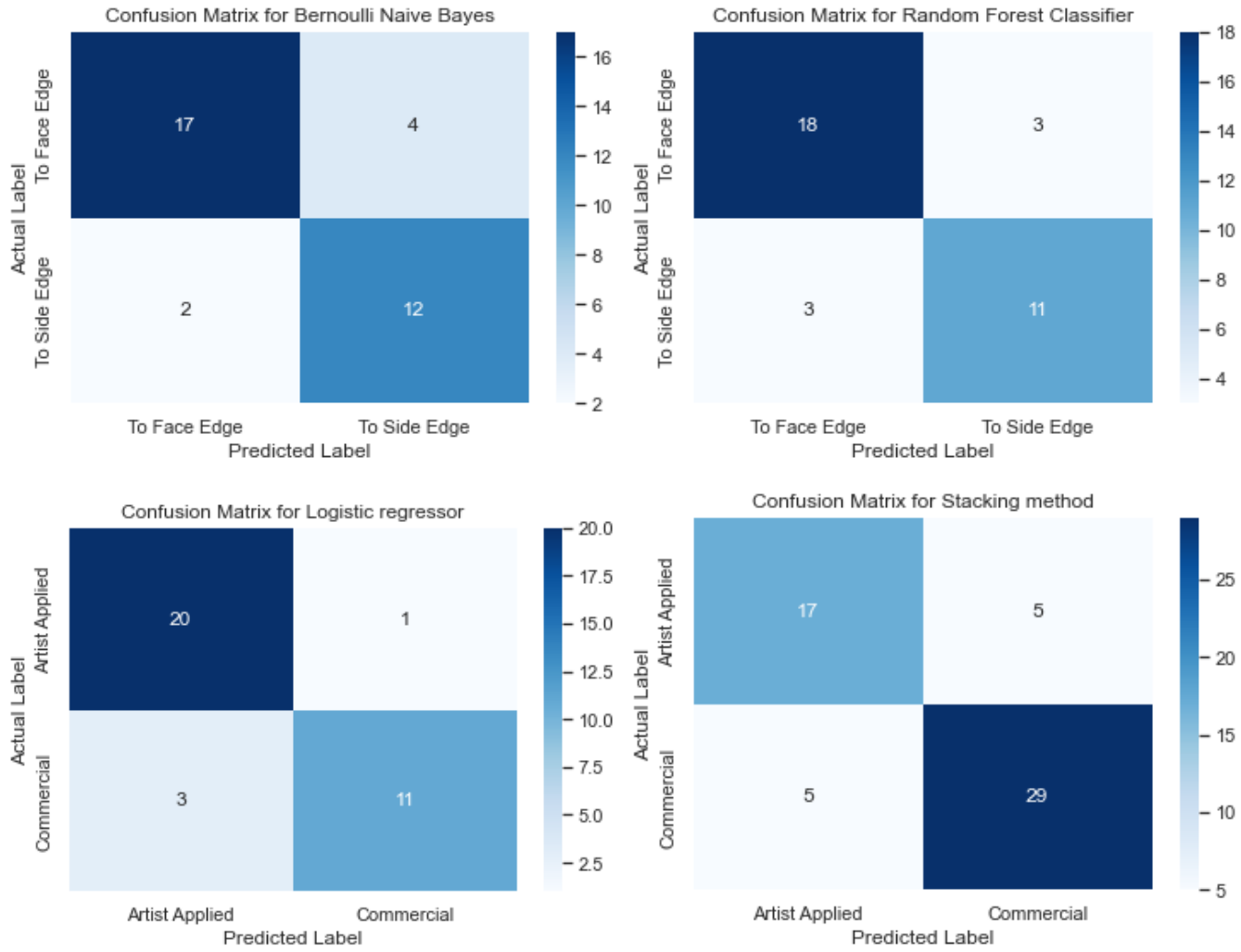


Figure 55: Confusion matrices for `ground_layer_limits`

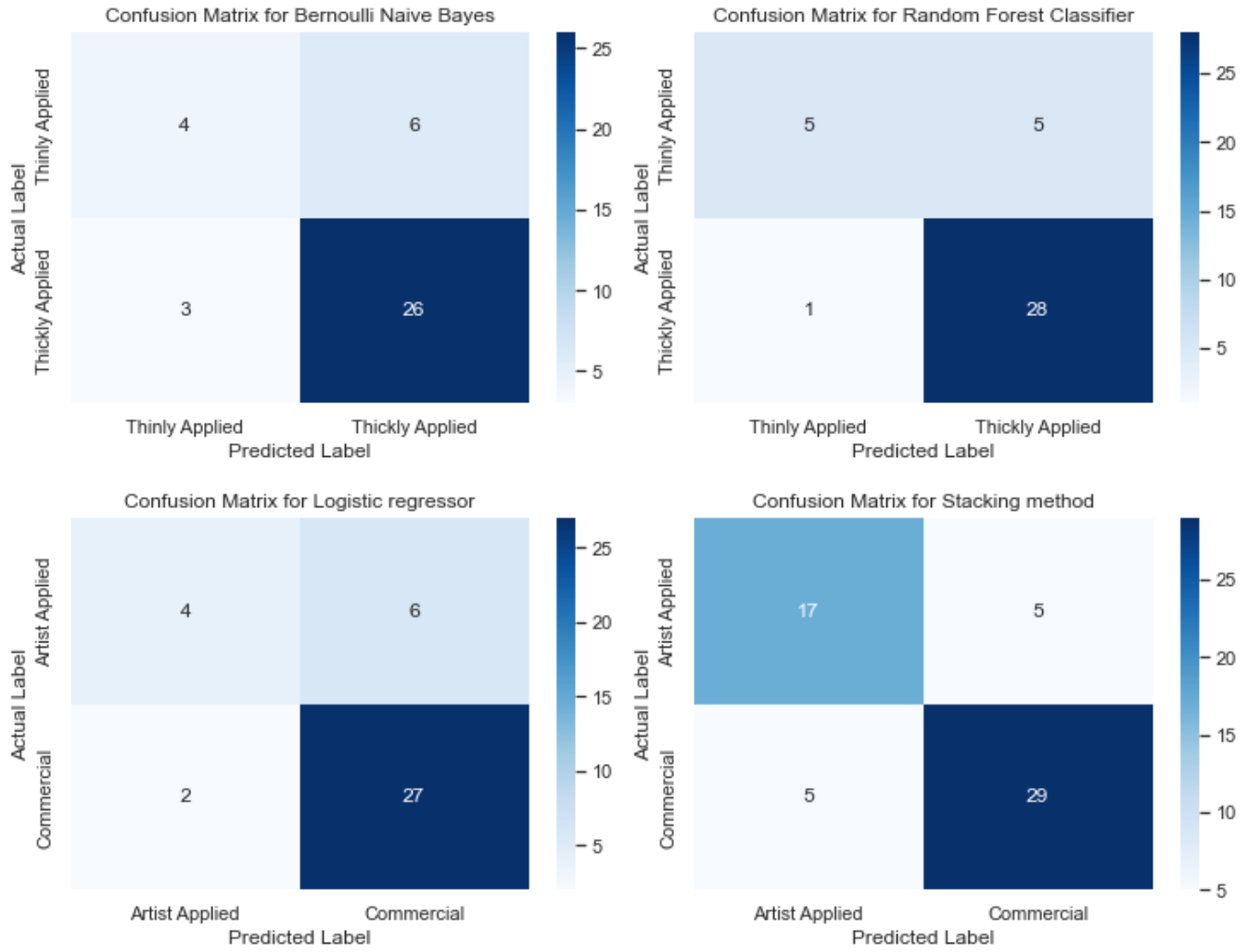


Figure 56: Confusion matrices for `ground_layer_thickness`

9.3 Project Management

9.3.1 GitHub Repository

Throughout the year, all our code and document were maintained with the following GitHub repository, https://github.com/DS-Project-Group-7/Data-Science-Project

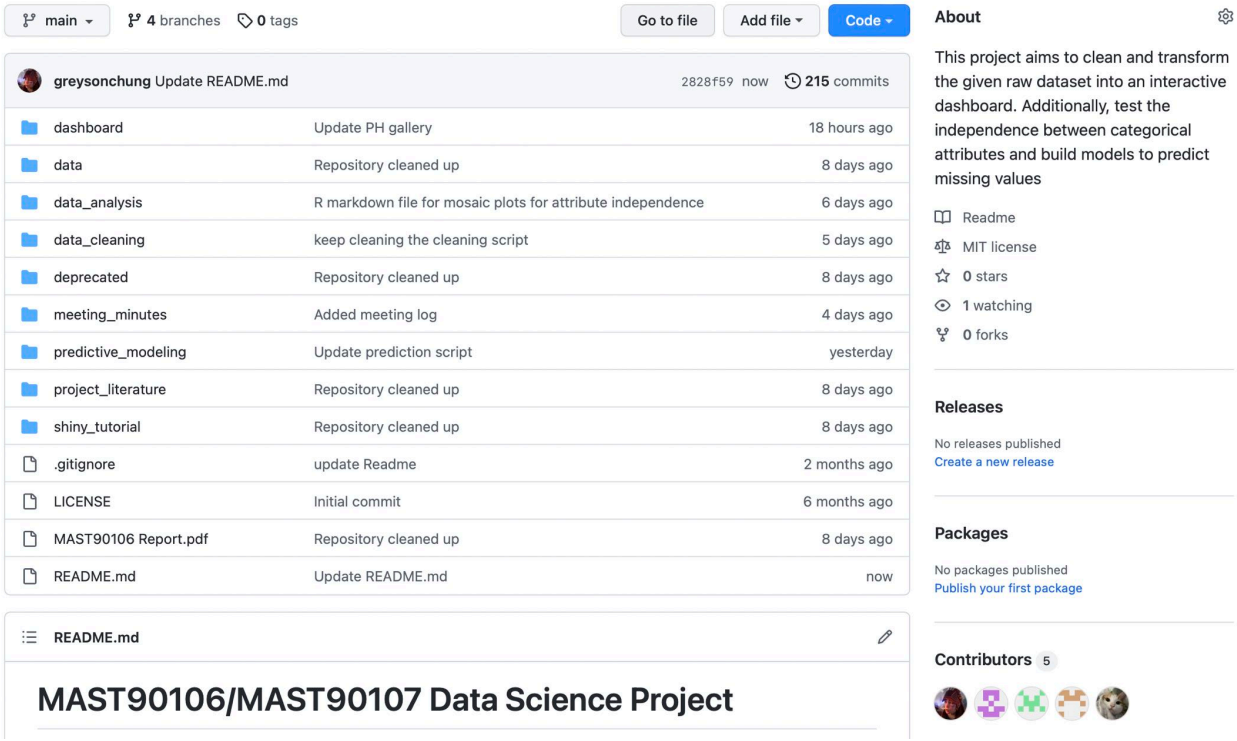


Figure 57: Project Repository

For confidentiality and privacy concerns, all group members have mutually agreed not to share project files with outsiders, and access to the GitHub repository is limited to group members, project supervisor Vivek Katial, and our client Nicole Tse. Furthermore, two factor authentication is set up for our GitHub accounts.

Branching is used throughout the project to maintain stability of the repository and isolates changes made to the code as shown in Figure 58.

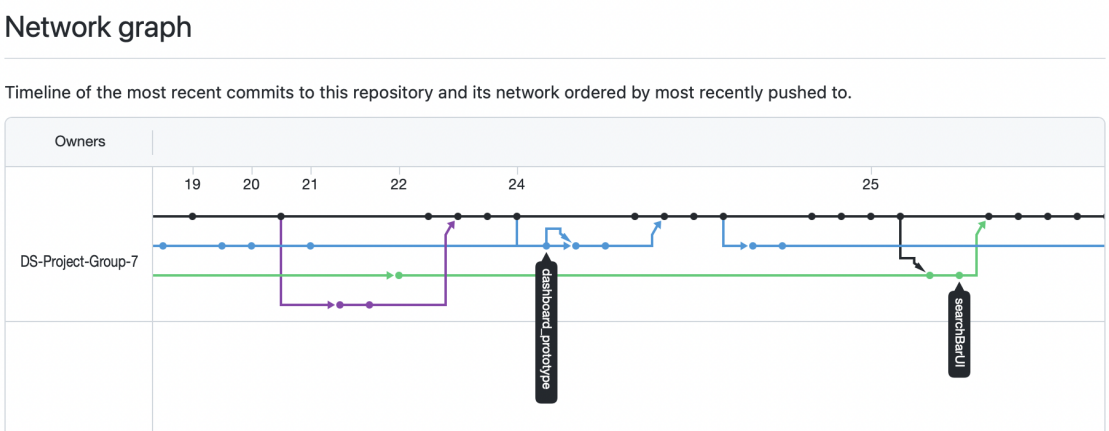


Figure 58: GitHub Branching

Group member contributions to the repository are tracked using GitHub's built-in contribution log functionality as shown in Figure 59.

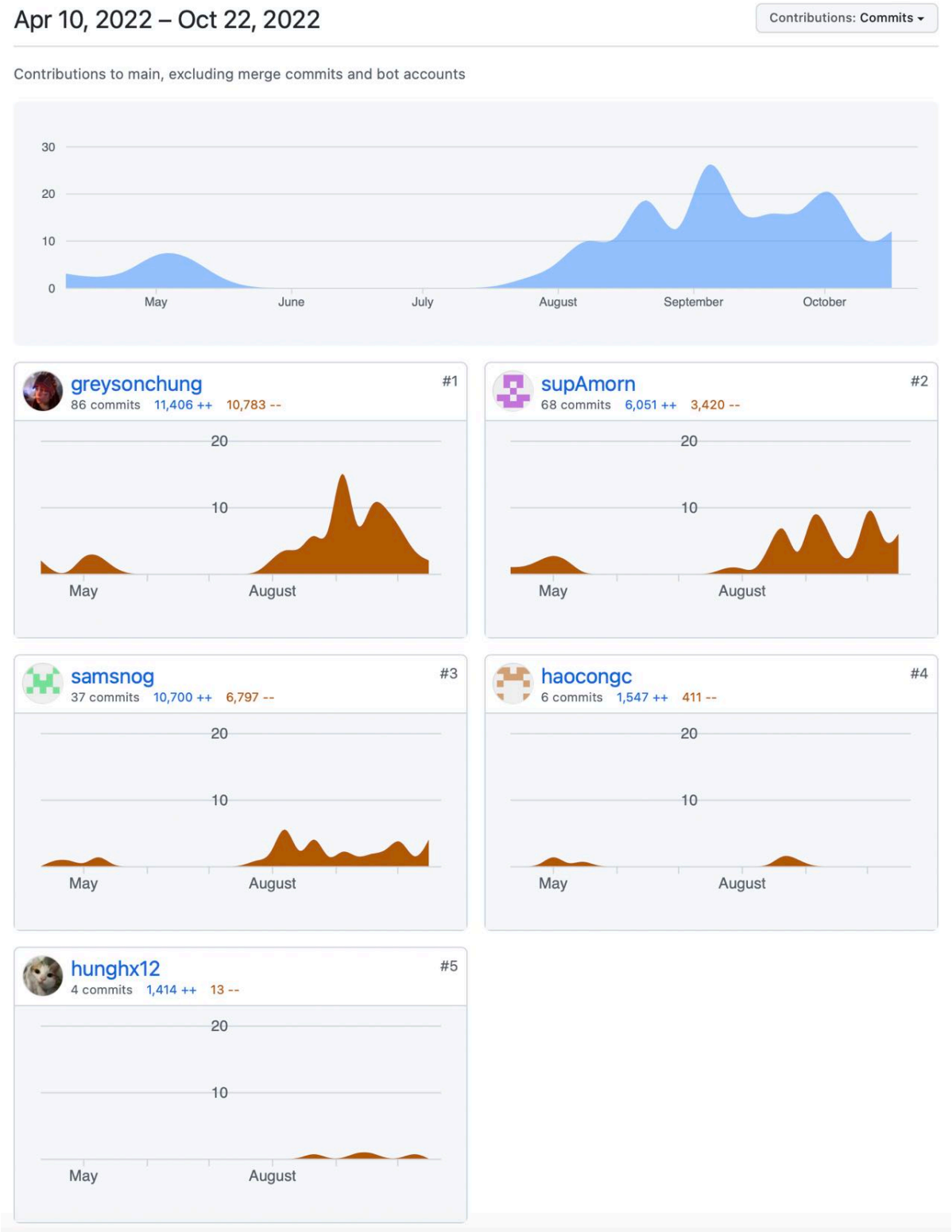


Figure 59: Repository Contribution

9.3.2 Project Task Tracking

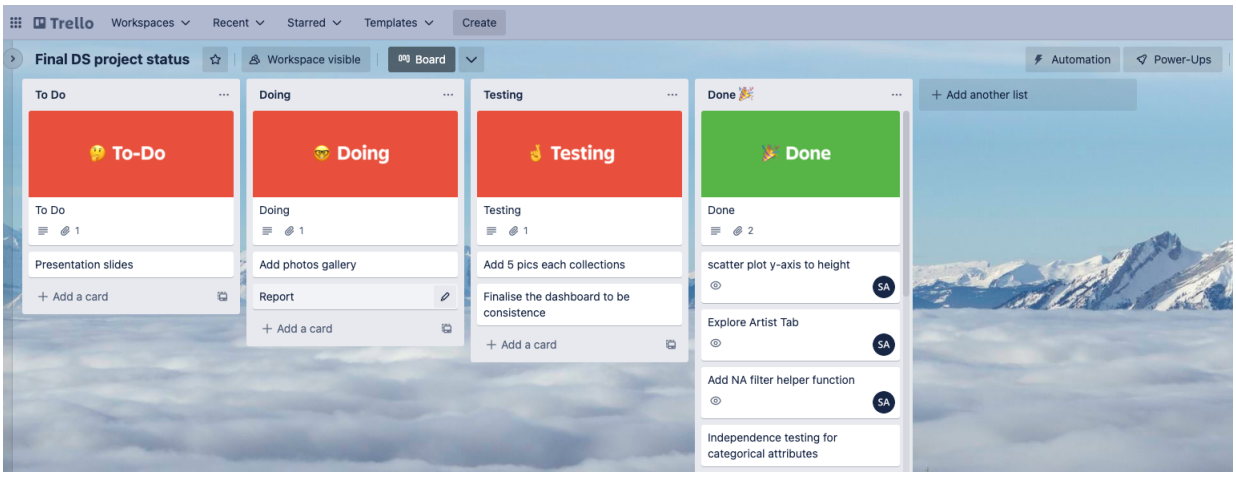


Figure 60: Trello Project Progress Tracking

As a result of a year-long project, our team encountered that there were so many tasks to work on and track the status along with studying other subjects; Hence, we created the Trello template of tracking status into 4 steps consisting of To-do, Doing, Testing and Done. Then, we discussed the tasks and allocated tasks by volunteers. The result of our work is shown in Figure 60.

One of the benefits of this approach is that everyone contributed their idea to prioritise the task and assign the task equally.

9.3.3 Client and Supervisor Meeting Logs

Week	Time	Hosts	Agenda
Week 4	24/3/2022, 10:00 - 10:30 AM	Vivek Katial (Super)	- Kick-off meeting with the group supervisor - Group member introduction
Week 5	29/3/2022, 10:00 - 11:00 AM	Nicole Tse (Client) Vivek Katial (Super)	- Kick-off meeting with client - Project plan discussion
Week 6	7/4/2022 , 10:00 - 10:30	Vivek Katial (Super)	- Project approach discussion with supervisor
Week 7	13/4/2022, 10:30 - 11:30 AM	Nicole Tse (Client)	- Question about literature - Question about data and cleaning process - R Shiny presentation
Non-teaching period	21/4/2022, 10:00 - 11:30 AM	Vivek Katial (Super)	- Data cleaning progress sharing - Question about R Shiny
Week 9	5/5/2022, 11:00 - 11:30 PM	Vivek Katial (Super)	- Dashboard demo discussion - Question about R Shiny - Data cleaning progress sharing
Week 10	11/5/2022, 10:30 - 11:30 AM	Nicole Tse (Client)	- Presenting dashboard demo to client - Question about dataset - Data cleaning progress sharing - Schedule for lab demonstration
Week 11	18/5/2022, 10:30 - 11:30 AM	Nicole Tse (Client)	- Lab demonstration at Grimwade Centre - Proposal for semester 2
Week 11	19/5/2022, 11:00 - 11:30 PM	Vivek Katial (Super)	- Semester 1 progress sharing - Proposal for semester 2

Table 5: Semester 1 Meeting Logs

Week	Time	Hosts	Agenda
Winter break	20/7/2022, 1:00 - 2:00 PM	Nicole Tse (Client)	- Semester 2 kick-off meeting - Discuss proposal for semester 2 - Feedback on data cleaning
Week 2	3/8/2022, 1:00 - 2:00 PM	Nicole Tse (Client)	- Data cleaning progress sharing - Question about dashboard interface - Ask about storage climate data
Week 2	4/8/2022 , 10:00 - 10:30 AM	Vivek Katial (Super)	- Question about Shiny interface - Dashboard implementation suggestion
Week 5	26/8/2022, 10:30 - 11:30 AM	Vivek Katial (Super)	- Dashboard progress sharing and feedback - Dashboard implementation suggestion
Week 6	31/8/2022, 12:30 - 1:30 PM	Nicole Tse (Client)	- Dashboard progress sharing and feedback - Additional question regarding the data - Raise concern about storage climate data
Week 7	9/9/2022, 10:30 - 11:30 AM	Vivek Katial (Super)	- Dashboard progress sharing and feedback - Dashboard implementation suggestion - Discuss possible task after complete dashboard
Week 8	14/9/2022, 12:30 - 1:30 PM	Nicole Tse (Client)	- Dashboard progress sharing and feedback - Discuss idea on independence testing - Discuss idea on missing value prediction
Week 10	5/10/2022, 12:30 - 1:30 PM	Nicole Tse (Client)	- Dashboard progress sharing and feedback - Independence testing progress and result sharing - Predictive modeling progress and result sharing - Project report progress sharing
Week 10	7/10/2022, 10:30 - 11:30 AM	Vivek Katial (Super)	- Dashboard progress sharing and feedback - Independence testing result sharing and discussion - Predictive modeling result sharing and discussion - Project report progress sharing
Week 11	12/10/2022, 12:30 - 1:30 PM	Nicole Tse (Client)	- Finalising dashboard - Independence testing result and limitation discussion - Predictive modeling result sharing - Discuss possible future work

Table 6: Semester 2 Meeting Logs

Our group has scheduled fortnight meeting with both the supervisor and the client. Most of the meeting were held on zoom due to COVID-19 and the availability of location of group members. However, our group had to skip two meeting during August because the client was not be available.

9.4 Dashboard User Manual

To access the dashboard as shown in Figure 61, the user requires username and password to access the dashboard (username: supanuthaa@gmail.com, password: SEADashboard07).

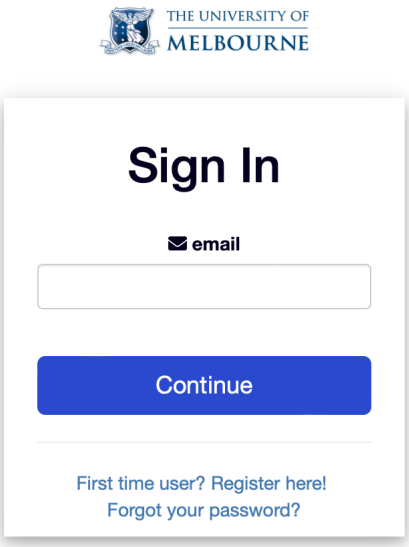


Figure 61: Login page

Once the user access into the dashboard, the system redirect on the homepage as displayed in Figure 62.

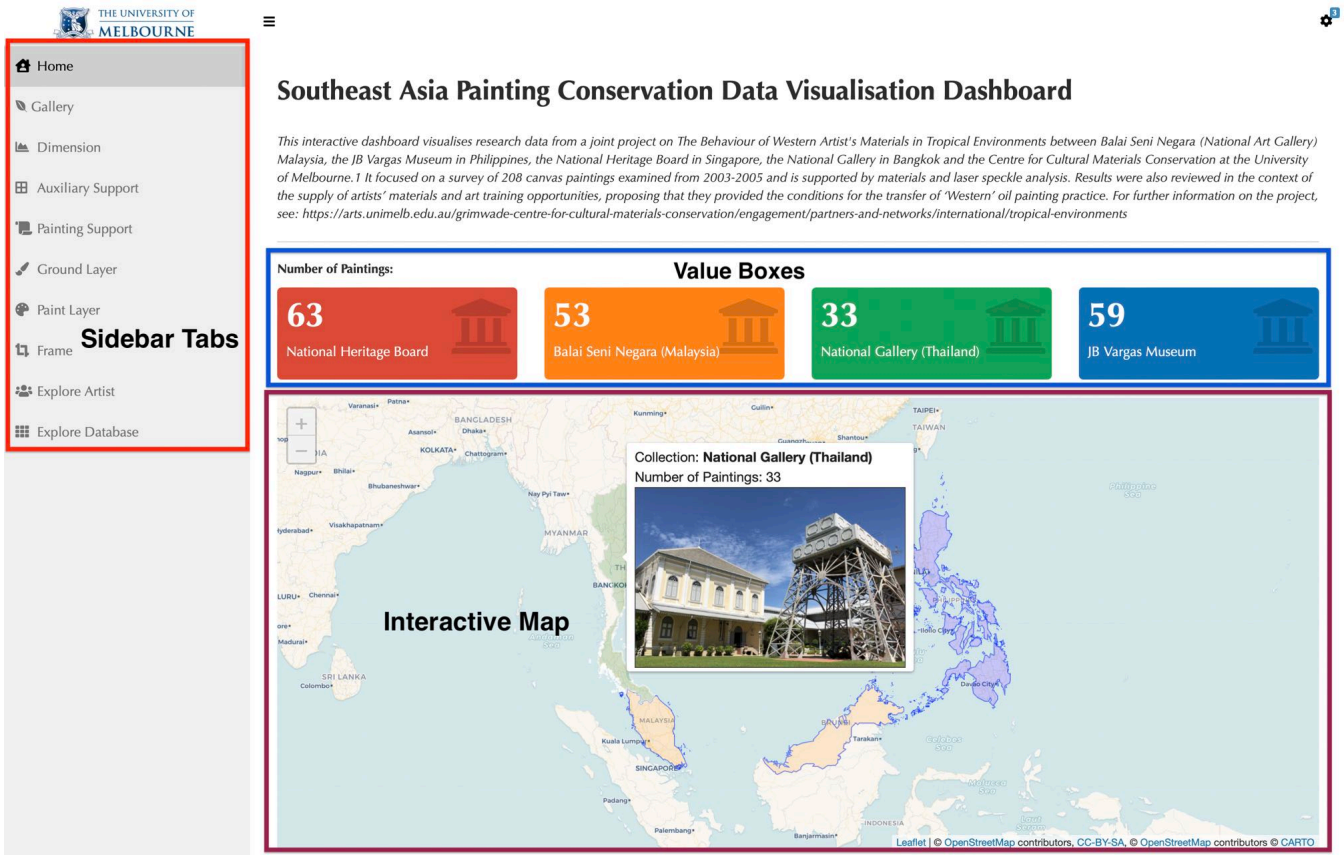


Figure 62: Dashboard Homepage

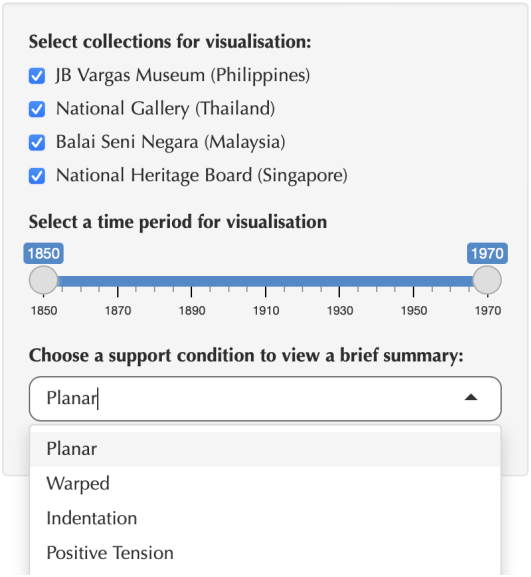


Figure 63: Filter Option

Moreover, when user selected each component, the filter option is able to real-time filter information as user desired as shown in Figure 63.

- The museum group checkbox will dynamically filter the museum on the overall condition plot and the condition presents on each museum plot.
- The slider bar of time period will dynamically filter the decade of paintings for the whole plotting.
- The select input condition will dynamically filter the the condition presents on each museum plot.
- The select input two conditions will dynamically filter the the heat map for comparing two attributes plot on painting support tab as displayed in Figure 64.

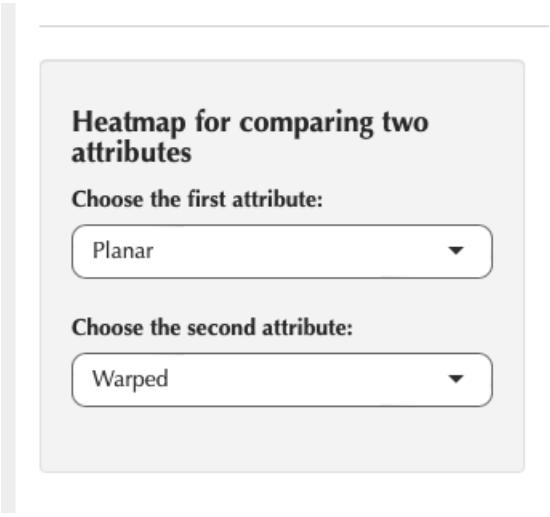


Figure 64: select input two conditions Filter Option

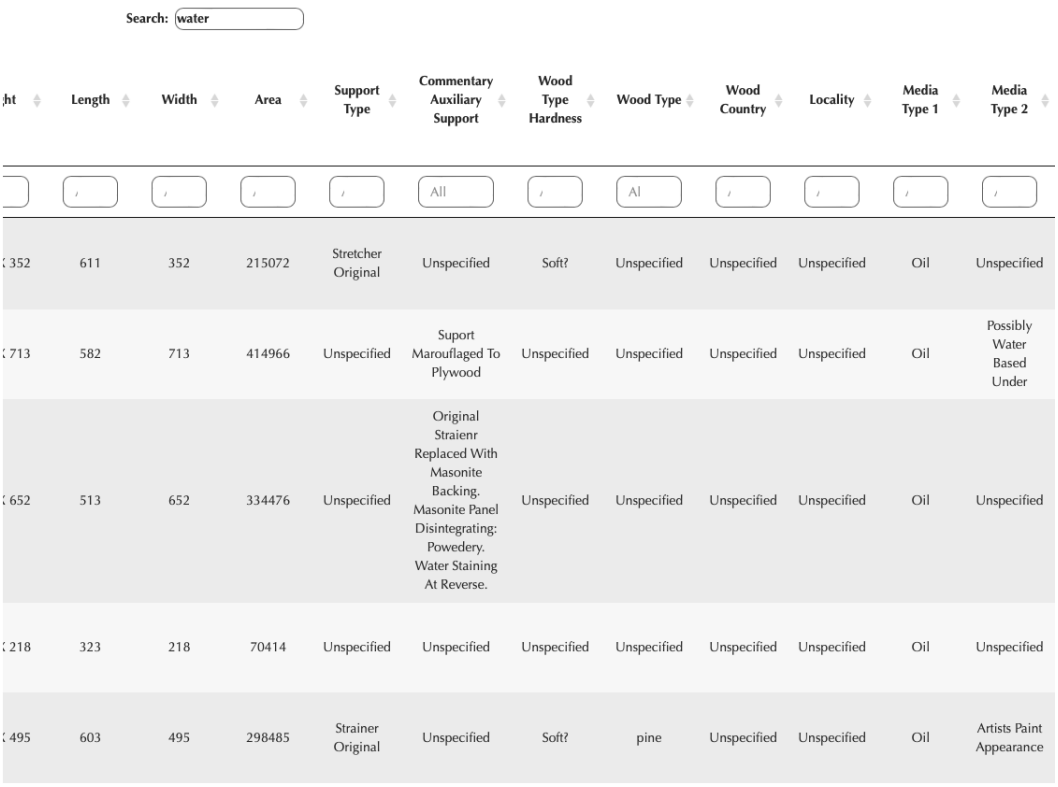


Figure 65: Search function for the whole dataset

The data exploration tab was designed for dataset exploration. The search bar at top right allows the user to browse the specific key word a cleaned version of the dataset as shown in Figure 65. In addition, a filter and search option have been added for each attribute, allowing the user to locate specific content by inputting a search string or selecting filter options as shown in Figure 66.

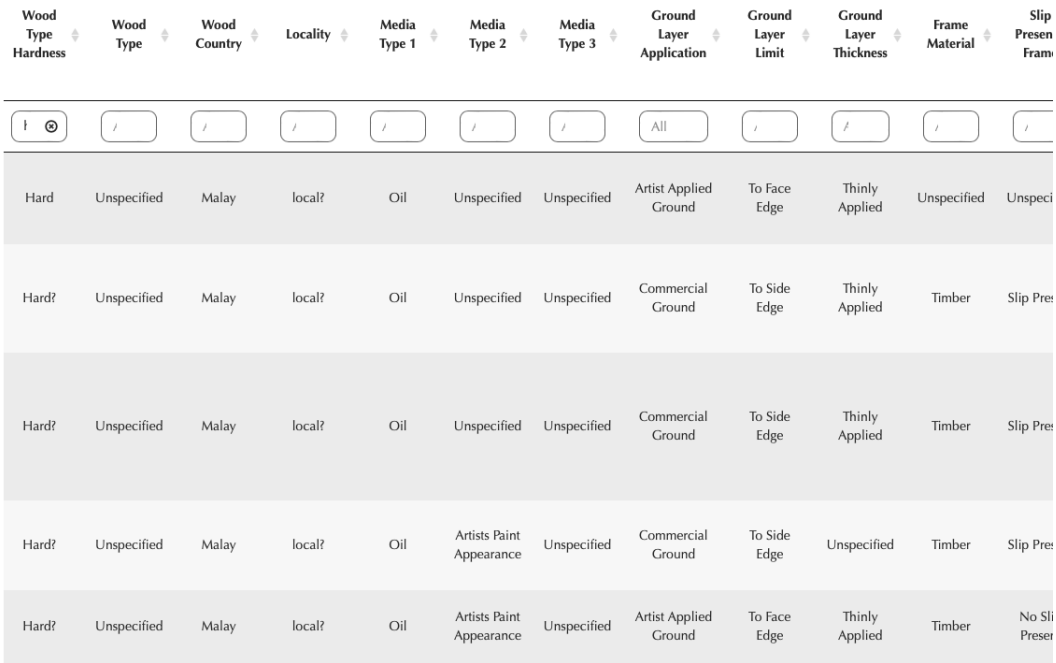


Figure 66: Search function for each column attribute

References

[1] N TSE and R SLOGGETT. "Southeast Asian Oil Paintings: Supports and Preparatory layers". In: *Preparation for painting: the artist's choice and its consequence*. GB (London): Archetype Publications, 2008, pp. 161–170.

[2] Claire Grech et al. "A Preliminary Investigation into the Behavior of Modern Artists' Oil Paints in a Hot and Humid Climate". In: Jan. 2019, pp. 419–435. ISBN: 978-3-030-19253-2. DOI: 10.1007/978-3-030-19254-9_33.

[3] Edward Parker. *Coronavirus outbreak: A new mapping tool that lets you scroll through timeline*. Sept. 2022. URL: https://theconversation.com/coronavirus-outbreak-a-new-mapping-tool-that-lets-you-scroll-through-timeline-131422.

[4] Wei Zhang. *New Zealand Trade Intelligence Dashboard*. URL: https://shiny.rstudio.com/gallery/nz-trade-dash.html.

[5] "Association between Two Categorical Variables: Contingency Analysis with Chi Square". In: *Business Statistics for Competitive Advantage with Excel 2007: Basics, Model Building, and Cases*. New York, NY: Springer New York, 2009, pp. 171–199. ISBN: 978-0-387-74403-2. DOI: 10.1007/978-0-387-74403-2_7. URL: https://doi.org/10.1007/978-0-387-74403-2_7.

[6] John H McDonald. *Fisher's exact test of independence*. URL: http://www.biostathandbook.com/fishers.html.