CS 448B Assignment 2 Exploratory Data Analysis

Yifan Shen

Select and Prepare the Data

Dataset

The dataset I choose to use is the Tate Gallery Collection in the Awesome Public Datasets. The repository contains the metadata for over 70, 000 artworks Tate owns. It also has a record of over 3,500 associated artists. The repository is no longer updated after October 2014. In the following data wrangling and exploratory data analysis process, both the artist data and the artwork data source files are used.

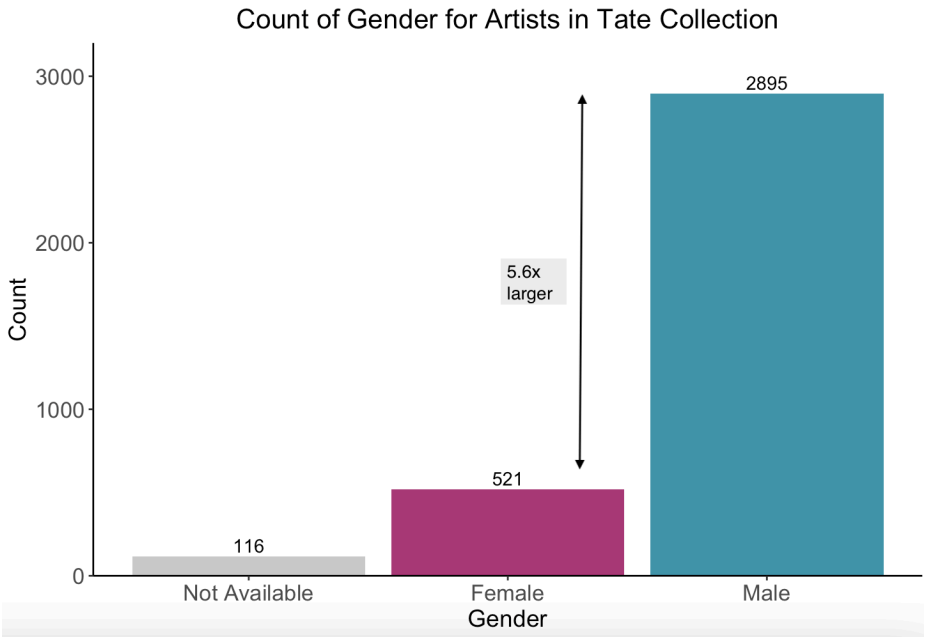
Guiding questions

- What is the general trend for artwork creation and acquisition for the artworks collected by Tate?
- Who are the top 10 most popular artists in the Tate collection? What are their demographics (gender)?
- What are the top 10 most common mediums for artworks?
- How has the aspect ratio (width/height) of artworks changed over time?

Data Wrangling

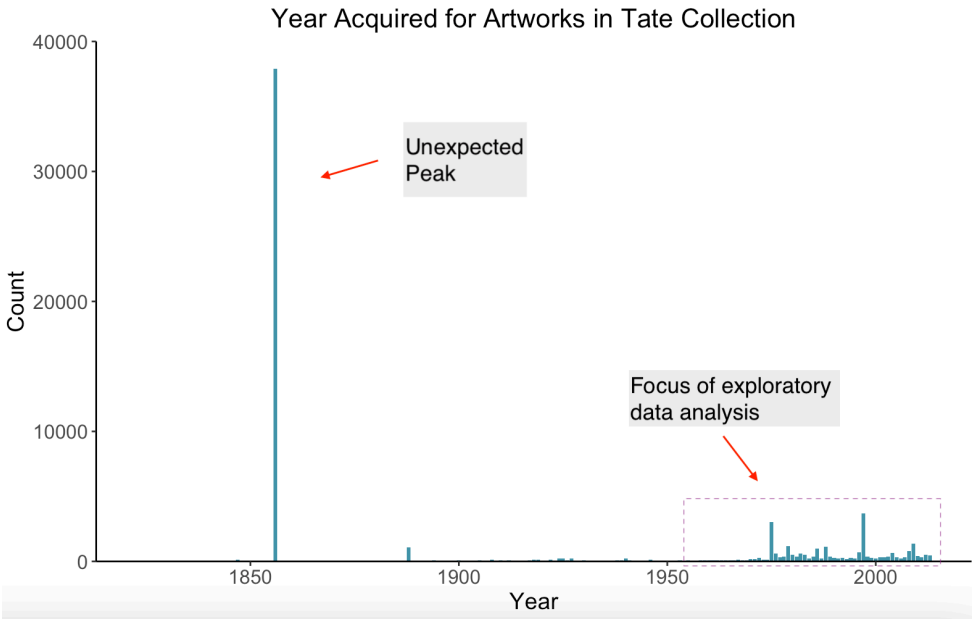
For the artist dataset: I first select only `id`, `name`, `gender`, `yearOfBirth` out of 9 features. In total, there are 3532 records. The features I am especially interested in are `gender` and `yearOfBirth`.

- For gender, as we can see from the figure below on the left, 116 out of 3532 (3.3%) records are missing. Moreover, the number of male artists is disproportionately higher (5.6 times) than that of female artists. This is an expected, however, disturbing fact. Due to this observation, I plan to add one more question about artists' gender in the exploratory data analysis part.
- For the year of birth, 60 out of 3532 (1.7%) records are missing. Even though artists who were born earlier are more likely to have a missing record of birth year, I will assume that the missing data will not have an effect on the following exploratory analysis. The general trend of year of birth also makes sense. The year of birth is sparsely scattered before 1700. After that, it has gradually increased since 1850. There are no observable outliers.



For the artwork dataset, I select `id`, `artist`, `artistId`, `medium`, `year`, `acquisitionYear`, `width`, `height`, and `units` out of 20 features with 69201 records. In the process of data wrangling, I check that all the width/height are of the same unit. As I am especially interested in the temporal trend of artwork creation and acquisition, I first check that there are no outliers in the artwork creation year.

After that, I explore the acquisition timeline for the collections. However, the peak around Year 1856 is especially unusual. This could be attributed to the establishment of an earlier institution before Tate was open to the public. To avoid confusion, I will narrow down the scope of 'What is the general trend for artwork acquisition' to the year after 1900.



Then, I transform the data frame and add a feature: `yearOut`, which is the number of years created before acquired by Tate for each piece of artwork. After transformation, a few records have negative values. Therefore, I choose to get rid of them.

For width and height, I check that they are all of the same unit, which is `mm`. However, I should point out that 3342 out of 69201 (4.8%) records are missing. Moreover, from the descriptive statistics, I notice several outliers that are abnormal. For instance, one artwork has a height of 37500 mm and a width of 10 mm, while the other has a height of 150 mm and a width of 11960 mm. Therefore, I remove the outliers accordingly.

In summary, in the data wrangling step, the following is completed:

- Select 4 out of 9 features from the artist dataset and 9 out of 20 features from the artwork dataset.
- Conduct feature engineering to calculate `yearOut`.
- Check that missing values in the data set could only have a negligible effect on the exploratory analysis.
- Narrow down the year acquisition temporal analysis to the Year after 1900.
- Identify outliers in `width` and `height`, which will affect the exploratory analysis process.

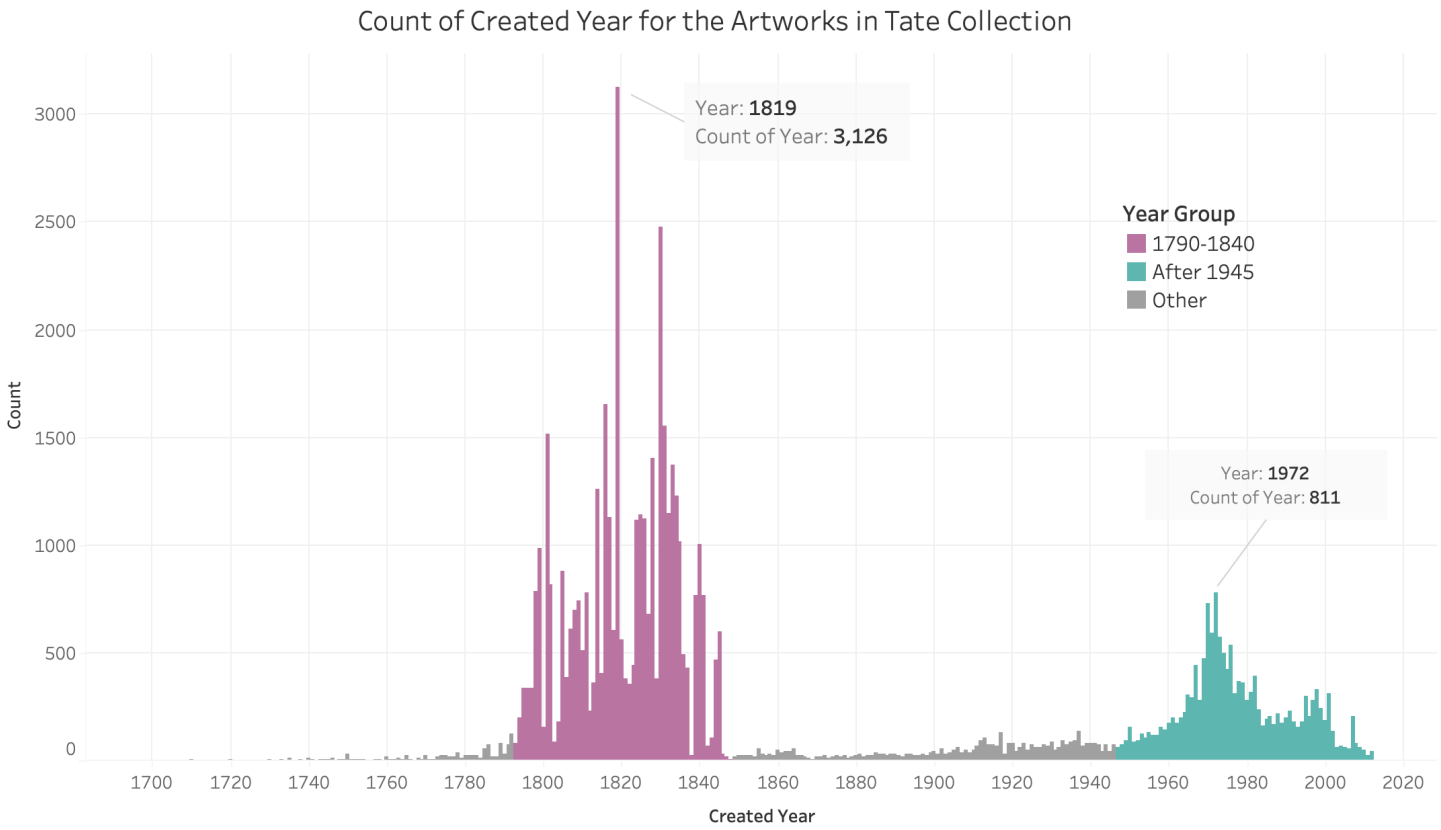
The final list of questions I would like to answer in the following exploratory analysis are:

- What is the general trend for artwork creation and acquisition for the artworks collected by Tate?
- How long will artworks get acquired by Tate after it has been created?
- Who are the top 10 most popular artists in the Tate collection? What are their demographics (gender)?
- Has female artists been recognized during recent years?
- How has the aspect ratio (width/height) of artworks changed over time?

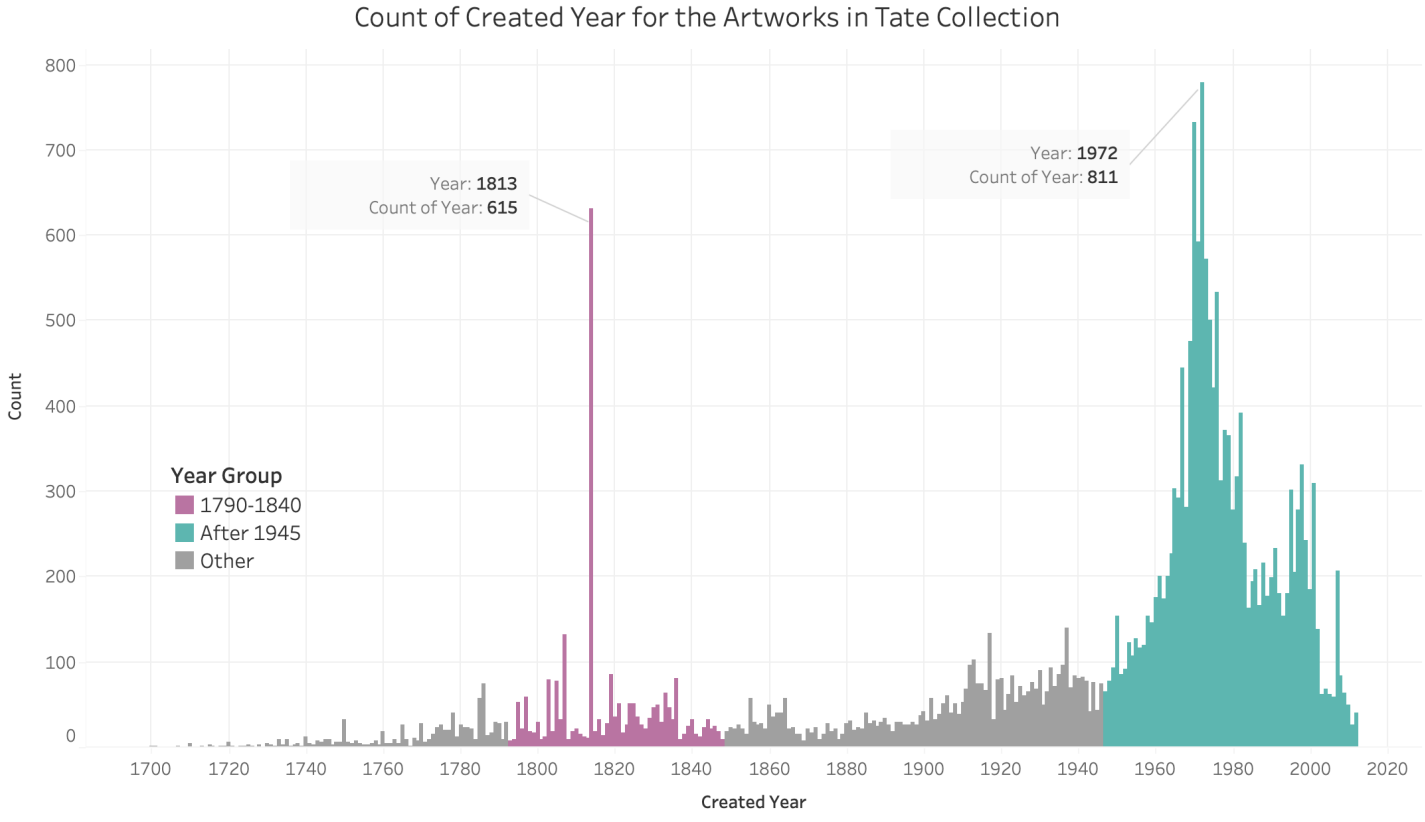
Exploratory Analysis

What is the general trend for artwork creation and acquisition for the artworks collected by Tate?

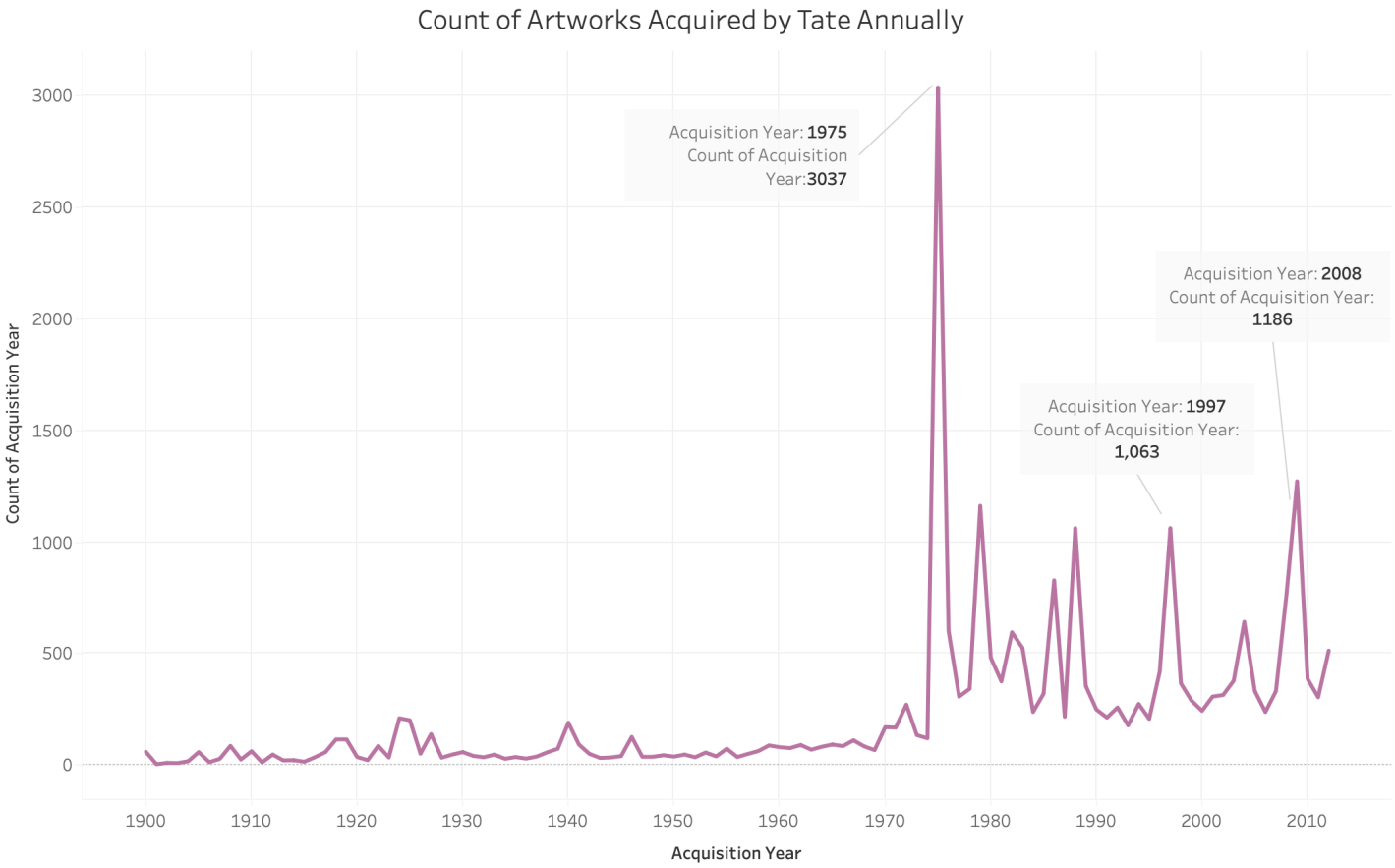
From the following figure, it can be seen that Tate has a preference to collect artworks created between 1790 and 1840 as well as artworks created after 1945. As annotated on the figure, Tate has collected 3147 pieces of artworks created in Year 1819. During the first period, there is a sharp increase of the number of artworks around 1790, then the number fluctuates from year to year. After Year 1840, we witness a sharp decrease in the number of artworks. For the second period, the number of artworks is lower in magnitude compared to the first period. With later analysis, we can discover at hind sight that a major proportion of the artworks collected during 1790-1840 belong to Joseph Turner (who created over 30,000 pieces of artworks in Tate collection).



The following figure illustrates the creation year of artworks in Tate collection, after filtering out the artworks produced by Joseph Turner.

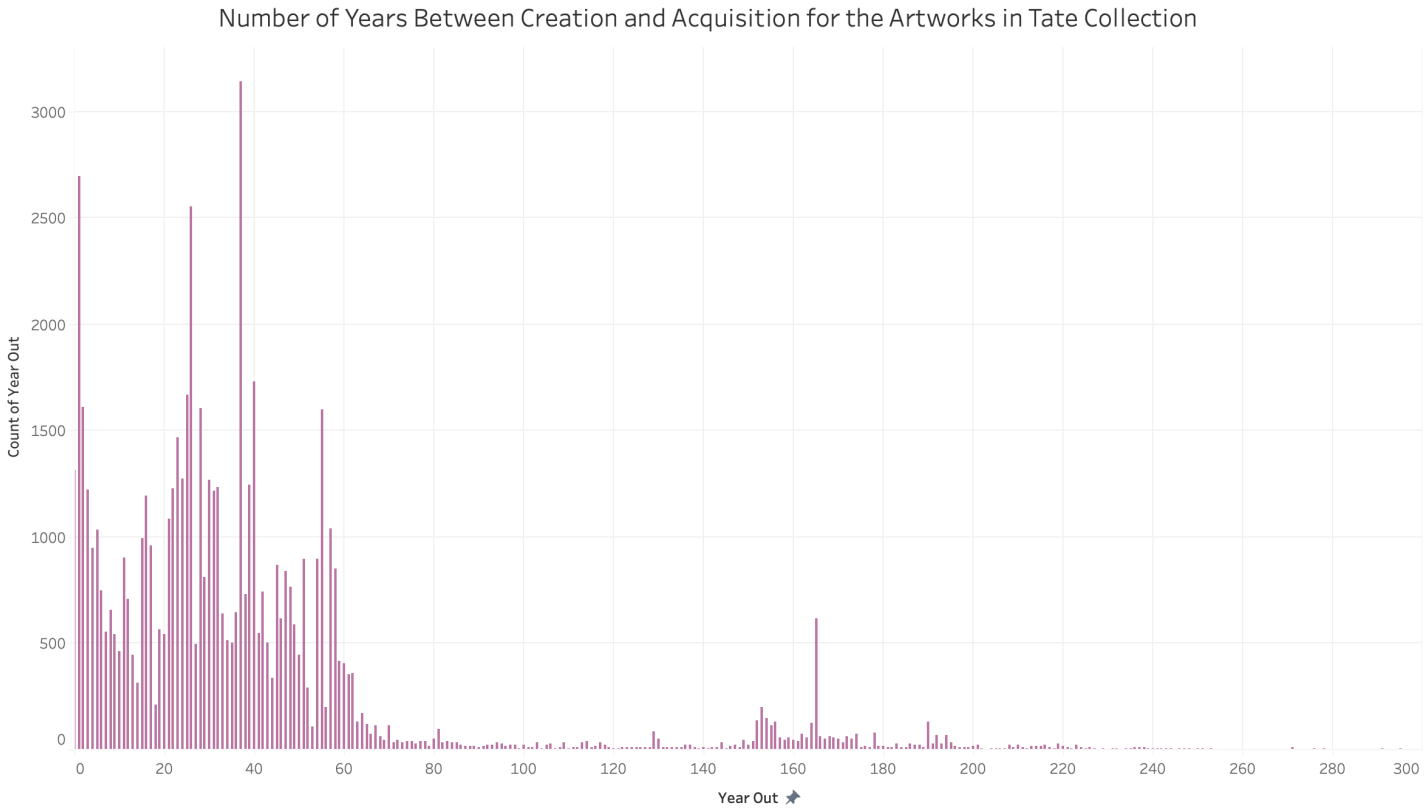


For artworks acquisition, if we disregard the unusual peak around Year 1856 (which can be due to acquisition of Turner's works), Tate gradually increase its volume of acquisition after Year 1970. The peak is at Year 1975, when Tate acquired over 3037 pieces of artworks. After that, the acquisition volume decreases and stabilized around 300-1200 pieces of artworks with variations from year to year.



How long will artworks get acquired by Tate after it has been created?

From the following illustration, we can see that most of the works are acquired within 60 years after its creation. However, for a small amount of works, it hasn't been acquired until over 160 years after creation.

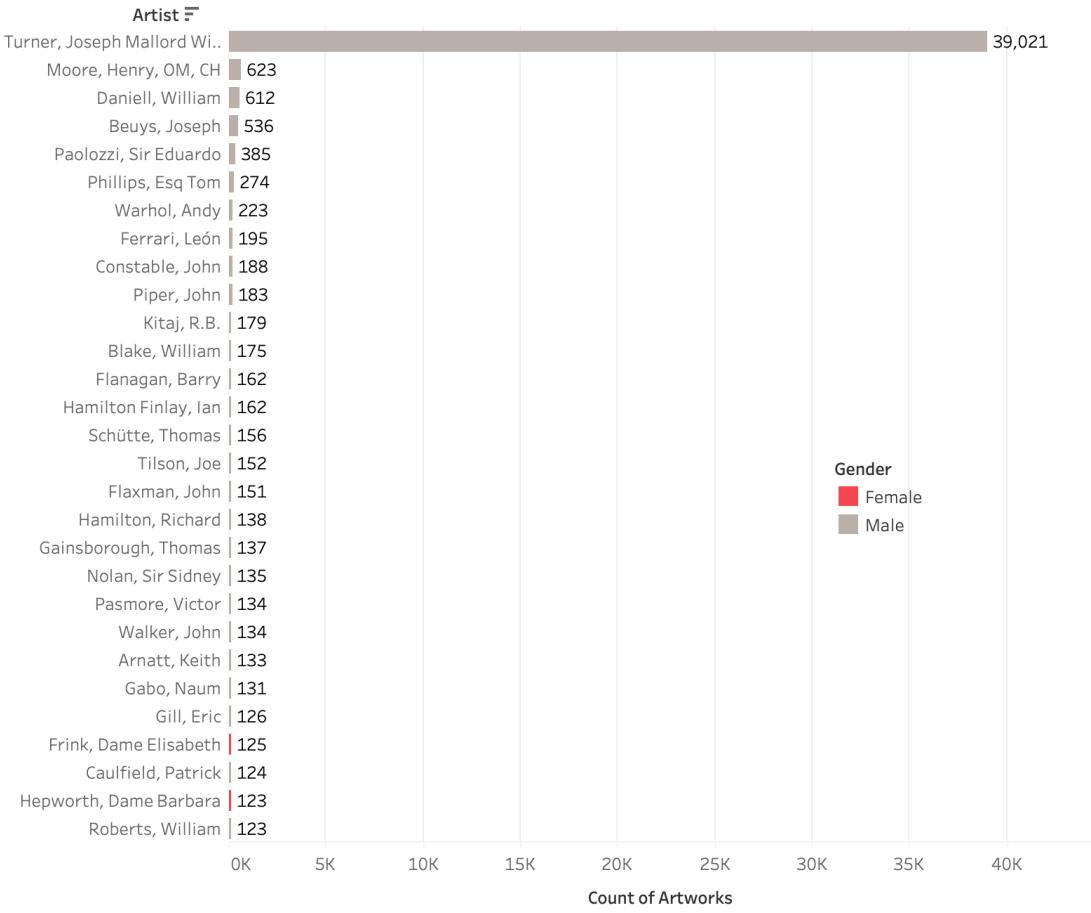


Who are the top 10 most popular artists in the Tate collection? What are their demographics (gender)?

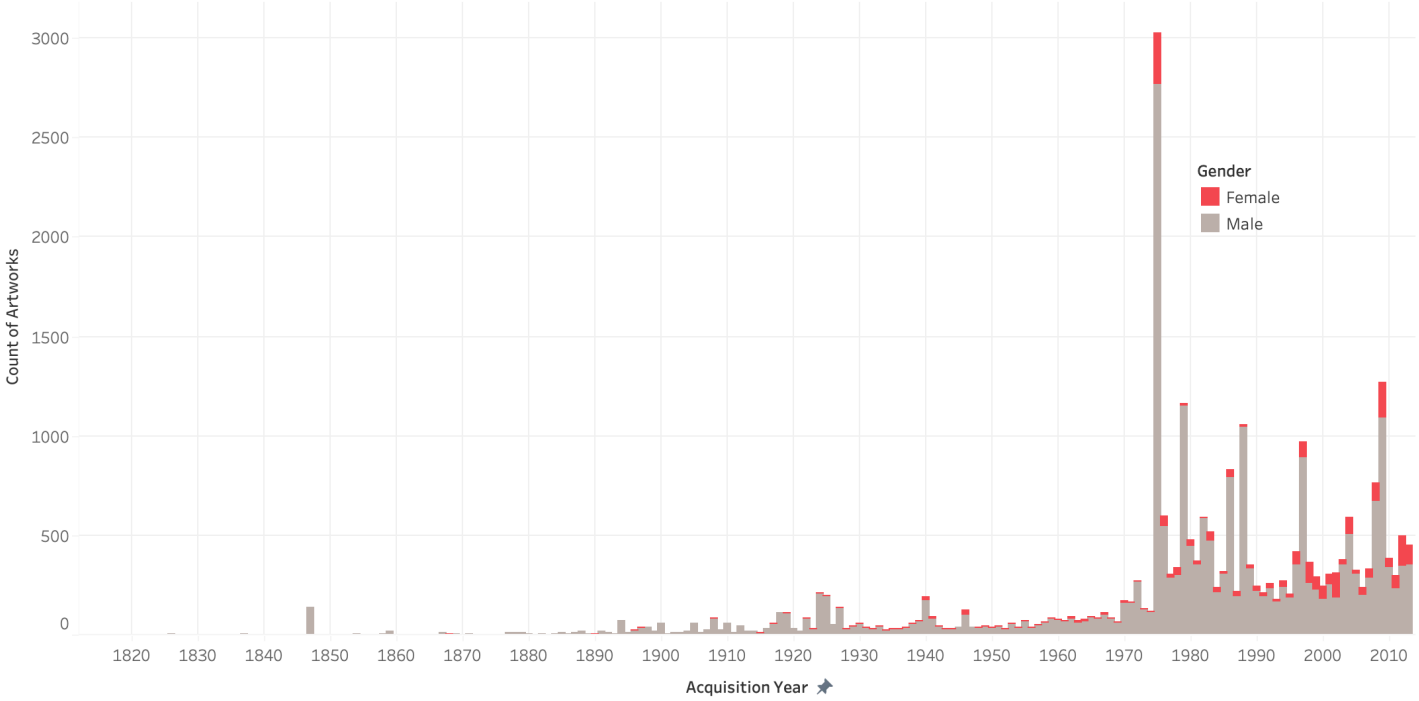
As the following illustration shows, Joseph Turner is the most popular artist (defined by the number of artworks collected by Tate) with over 39,000 pieces of arts collected by Tate. To verify this, I searched online and found out he 'He left behind **more than 550 oil paintings, 2,000 watercolours, and 30,000 works on paper**.' Therefore, this could make sense if most of the works by Turner are collected by Tate. However, this is still a surprising insight. Followed by Turner, it's Henry Moore, William Daniell, and Joseph Beuys. The demographics (gender) of the artists, however, reveals that in the top 30 most popular artist in Tate collection ranked by the pieces of artworks), only 2 female artist present in the 27th and 29th position.

To further investigate the artworks creation and acquisition by gender, I visualized the count of acquisition annually by gender. To make the figure becomes more readable, I temporarily removed the acquisition for Turner's works in 1856. As we can see from the visualization, only a small proportion of artworks acquired by Tate each year are created by female artists.

Gender for Artists in Tate Collections (Ranked by the Pieces of Artworks)

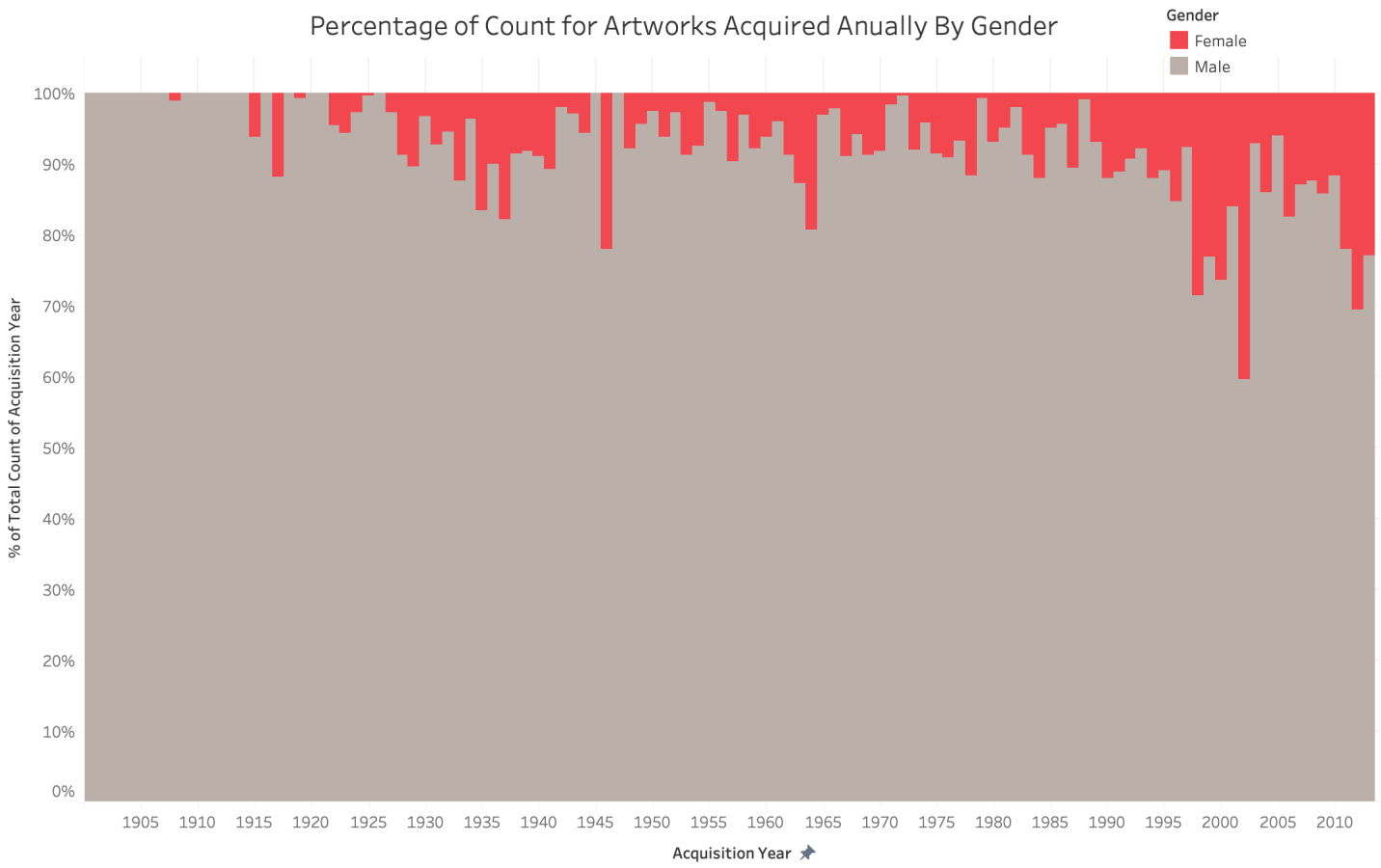


Count of Artworks Acquired Anually By Gender



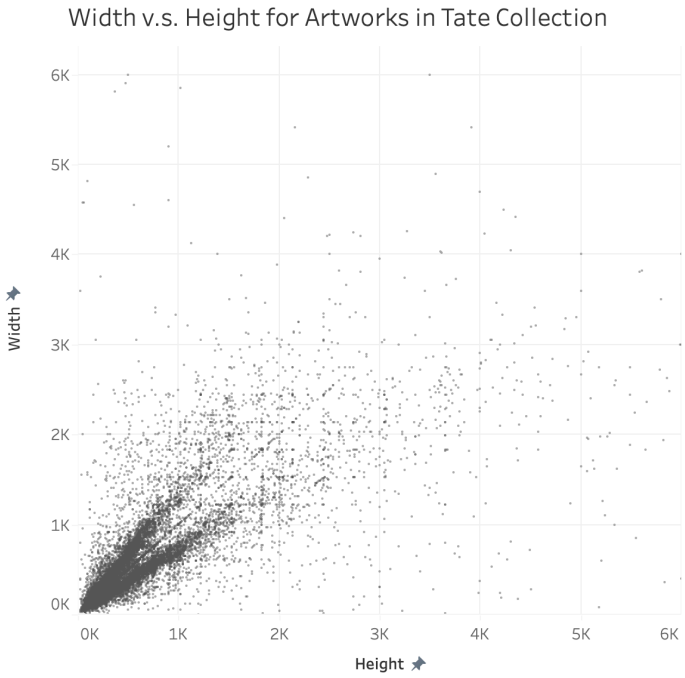
Has female artists been recognized during recent years?

To have a more intuitive understanding about whether female artists have been recognized in the recent years, I visualized the following figure with the percentage of count for artworks acquired by Tate each year aggregated by gender. From the illustration below, we can see that, starting from 1900 (I filtered out the data before 1900 as almost none of the artworks acquired before 1900 were created by female artists), the general trend is that more and more often, female artists' artworks are acquired by Tate. The percentage has been increased from 3-5% to around 20% in the last few years. Therefore, if judging from the percentage of count for artworks acquired by Tate each year aggregated by gender, female artists have been recognized more and more often by Tate collections.



How has the aspect ratio (width/height) of artworks changed over time?

Firstly, I explore the relation between width and height for artworks in Tate collection. As we can see from the figure below, the ratio of artworks are clustered along two lines, which correspond to the width/height ratio of 3:4 and the width/height ratio of 4:3.



Then, I continue to explore how the width-height ratio has changed over the year for the artworks collected in Tate according to their creation year. I choose to keep a record of median width-height ratio as median is more robust to outliers and skewed data. From the following figure, we can see that over the years, the ratio are fluctuating between 0.5 and 1.6, centering around 1.25.