Tutorial 6

Waseem

10/3/2021

The files artist_data.csv and artwork_data.csv contain information on approximately 70,000 artworks in the Tate museum in Britain. The dataset comes from this repository: https://github.com/tategallery/collection Use read.csv() to read the two tables into R as artist_info and artist_work. If you use read_csv(), the columns will be converted automatically, discarding some data in the process. Note that the two tables can be joined by the id and artistId columns. Convert to numeric In artist_work, the year column corresponds to the year in which the artwork was created. Convert this to a numeric column with the same name. What warnings did you observe? How did you deal with them?

#"NAs introduced by coercion" warning was observed. changed no year to spaces

```
temp<- which(str_detect(artist_work$year, "[^0-9] "))
artist_work$year[temp[1:18]]<-""
artist_work$year[temp[19]]<-"1998"

artist_work<-mutate(artist_work,year= as.numeric(year))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

Place of Birth The placeOfBirth column in artist_info is typically of the format "city, birth_area", or "birth_area". However, the birth_area uses the old name for the country. Use the table in artists_nation.xlsx to create a new column with the modern name for the nation in which the artist was born. Clue: You can use separate() here.

```
artist_info <- separate(artist_info, placeOfBirth, into=c(NA, "country"),
                        sep=",",extra="merge",fill="left",remove=FALSE)
manual_ids<-which(str_detect(artist_info$country,","))
artist_info$country[manual_ids]<-c("France","Columbia","Israel","Israel","Hrvatska","Colombia")
artist_nation<-read_excel("data/artists_nation.xlsx")
artist_info<-left_join(artist_info, artist_nation,by=c("country"="birth_area"))
artist_info
```

```
## # A tibble: 3,532 x 11
##       id name     gender dates  yearOfBirth yearOfDeath placeOfBirth    country
##    <int> <chr>    <chr>  <chr>        <int>       <int> <chr>           <chr>
##  1 10093 Abakano~ Female born ~         1930          NA Polska          "Polska"
##  2     0 Abbey, ~ Male   1852-~         1852        1911 Philadelphia, ~ " Unite~
##  3  2756 Abbott,~ Female 1898-~         1898        1991 Springfield, U~ " Unite~
##  4     1 Abbott,~ Male   1760-~         1760        1803 Leicestershire~ " Unite~
##  5   622 Abraham~ Male   born ~         1935          NA Wigan, United ~ " Unite~
##  6  2606 Absalon  Male   1964-~         1964        1993 Tel Aviv-Yafo,~ " Yisra~
##  7  9550 Abts, T~ Female born ~         1967          NA Kiel, Deutschl~ " Deuts~
##  8   623 Acconci~ Male   born ~         1940          NA New York, Unit~ " Unite~
##  9   624 Ackling~ Male   1947-~         1947        2014 Isleworth, Uni~ " Unite~
## 10   625 Ackroyd~ Male   born ~         1938          NA Leeds, United ~ " Unite~
## # ... with 3,522 more rows, and 3 more variables: placeOfDeath <chr>,
```

```
## #   url <chr>, nation <chr>
```

No information Retrieve the unique artist and artistId columns for the works of art for which the Tate Museum does not have information regarding the artist.

```
no_info<- anti_join(artist_work, artist_info, by= c("artistId"="id"))%>%
select("artist", "artistId") %>%
unique()
no_info
```

```
## # A tibble: 4 x 2
##   artist                           artistId
##   <chr>                               <int>
## 1 ?British School                     19232
## 2 Hullmandel, Charles                  5265
## 3 Kabakov, Ilya                        3462
## 4 The Leach Pottery (St. Ives, UK)    12951
```

Remove an Artist Outlier Identify the artist with the most number of artworks, and remove him/her from the artist_works tibble.

```
artist_work%>% group_by(artistId)
```

```
## # A tibble: 69,201 x 20
## # Groups:   artistId [3,342]
##       id accession_number artist  artistRole artistId title    dateText medium
##    <int> <chr>            <chr>   <chr>         <int> <chr>    <chr>    <chr>
##  1  1035 A00001           Blake,~ artist           38 A Figur~ date no~ Waterco~
##  2  1036 A00002           Blake,~ artist           38 Two Dra~ date no~ Graphit~
##  3  1037 A00003           Blake,~ artist           38 The Pre~ ?c.1785  Graphit~
##  4  1038 A00004           Blake,~ artist           38 Six Dra~ date no~ Graphit~
##  5  1039 A00005           Blake,~ artist           39 The Cir~ 1826-7,~ Line en~
##  6  1040 A00006           Blake,~ artist           39 Ciampol~ 1826-7,~ Line en~
##  7  1041 A00007           Blake,~ artist           39 The Baf~ 1826-7,~ Line en~
##  8  1042 A00008           Blake,~ artist           39 The Six~ 1826-7,~ Line en~
##  9  1043 A00009           Blake,~ artist           39 The Ser~ 1826-7,~ Line en~
## 10  1044 A00010           Blake,~ artist           39 The Pit~ 1826-7,~ Line en~
## # ... with 69,191 more rows, and 12 more variables: creditLine <chr>,
## #   year <dbl>, acquisitionYear <int>, dimensions <chr>, width <chr>,
## #   height <chr>, depth <dbl>, units <chr>, inscription <chr>,
## #   thumbnailCopyright <chr>, thumbnailUrl <chr>, url <chr>
```

```
artist_work <- filter(artist_work, artistId != 558)
```

Gender counts By Century Use the acquisitionYear column in the artist_work table to com- pute the century in which the artwork was acquired by the Museum. By joining with the artist_info table, recreate the following ta- ble. The artworks with missing or NA gender, and the artworks with missing acquisition year were removed prior to creating the table.

```
library(knitr)
artist_work %>% left_join(artist_info,by= c("artistId"="id")) %>%
filter(!is.na(gender),!is.na(acquisitionYear),gender!="")%>%
mutate(century= acquisitionYear%/% 100 + 1)%>%
group_by(century,gender) %>%
count() %>%
  pivot_wider(id_cols= "gender",names_from="century",values_from="n")%>%
  kable(col.names=c("","19th C.","20th C.","21th C."))
```

	19th C.	20th C.	21th C.
Female	6	1584	1133
Male	1939	18968	5465

What does the code below do? Install the packages you need and look up the help pages if you need to. The code takes a list of words, splits them into separate words, then converts it all to lowercase. It then removes words that are common occurences with stopwords,then prints out the words, with their sizes according to the count.

library(wordcloud) library(tm) wordcloud(ww, vv, scale=c(1.5,.3), min.freq = 30) workswatercolour Include a paragraph on markdown text to explain the code briefly.

Versatile artists Which artists have had more than 10 art works acquired by Tate, and used a different medium each time? Balka, Alys

```
versatile_artists <- artist_work %>% group_by(artistId)%>%
  filter(n()>=10, n_distinct(medium)==n()) %>%
  ungroup() %>% select(artist,medium,year)
versatile_artists
```

```
## # A tibble: 25 x 3
##    artist          medium                                                year
##    <chr>           <chr>                                                <dbl>
##  1 Balka, Miroslaw Concrete, plaster, brick, fabric, light bulb and print~  1986
##  2 Balka, Miroslaw Steel, felt and ash                                    1995
##  3 Balka, Miroslaw Steel, salt and linoleum                               1995
##  4 Balka, Miroslaw Steel, soap, linoleum and felt                         1995
##  5 Balka, Miroslaw Steel, linoleum, glass and ash                         1995
##  6 Balka, Miroslaw Steel                                                  1995
##  7 Balka, Miroslaw Wood, steel, polyester foam and salt                   1991
##  8 Balka, Miroslaw Wood, metal, plastic, water pump, motor and milk       1989
##  9 Balka, Miroslaw Steel and wax                                          1998
## 10 Balka, Miroslaw Soap and stainless steel                               2000
## # ... with 15 more rows
```

Popular artists Suppose we use the number of years between creation and acquisition by Tate as an indication of popularity. An artwork by a popular artist would presumably be acquired soon after creation. Follow these steps to identify artists who were popular. 1. Compute the difference: acquisition year minus year. 2. Keep only those artists whose first created artwork in the dataset is on 1970 or after, and who have 10 or more artworks acquired by Tate. 3. Compute the median difference between acquisition year and cre- ation year for each of these artists. 4. Keep only those for whom this difference is 1 year or less. To Explore

```
artist_work %>%
  mutate(yr_to_acq = acquisitionYear - year) %>%
  group_by(artistId) %>%
  filter(min(year)>=1970, n() >=10) %>%
  summarise (med1= median (yr_to_acq),.groups="drop")%>%
  filter(med1 <=1) -> popular_artists
popular_artists
```

```
## # A tibble: 47 x 2
##    artistId  med1
##       <int> <dbl>
##  1      687     1
##  2      717     1
##  3      784     1
```

```
##  4      805     1
##  5      891     1
##  6      917     1
##  7      953     1
##  8     1133     1
##  9     1212     1
## 10     1242     0
## # ... with 37 more rows
```

Having worked with this data for a little under a week, what question of interest would you want to explore about it? Enter your idea here, and we can execute it in a subsequent tutorial. The github link above has some links to ideas that others have tried.