Artist Project 2

Diana

February 17, 2019

This is a continuation of the Artist project. Previously, I learned to create a density plot to show the distribution of birth decades for genders (male/female). This time, I am going to create a density plot to show the distribution of birth decades for various ethnicities of the artists:

Step 1: Set working directory, load libraries, and read the .csv file

```
## [1] "C:/Users/diana/DataViz"
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

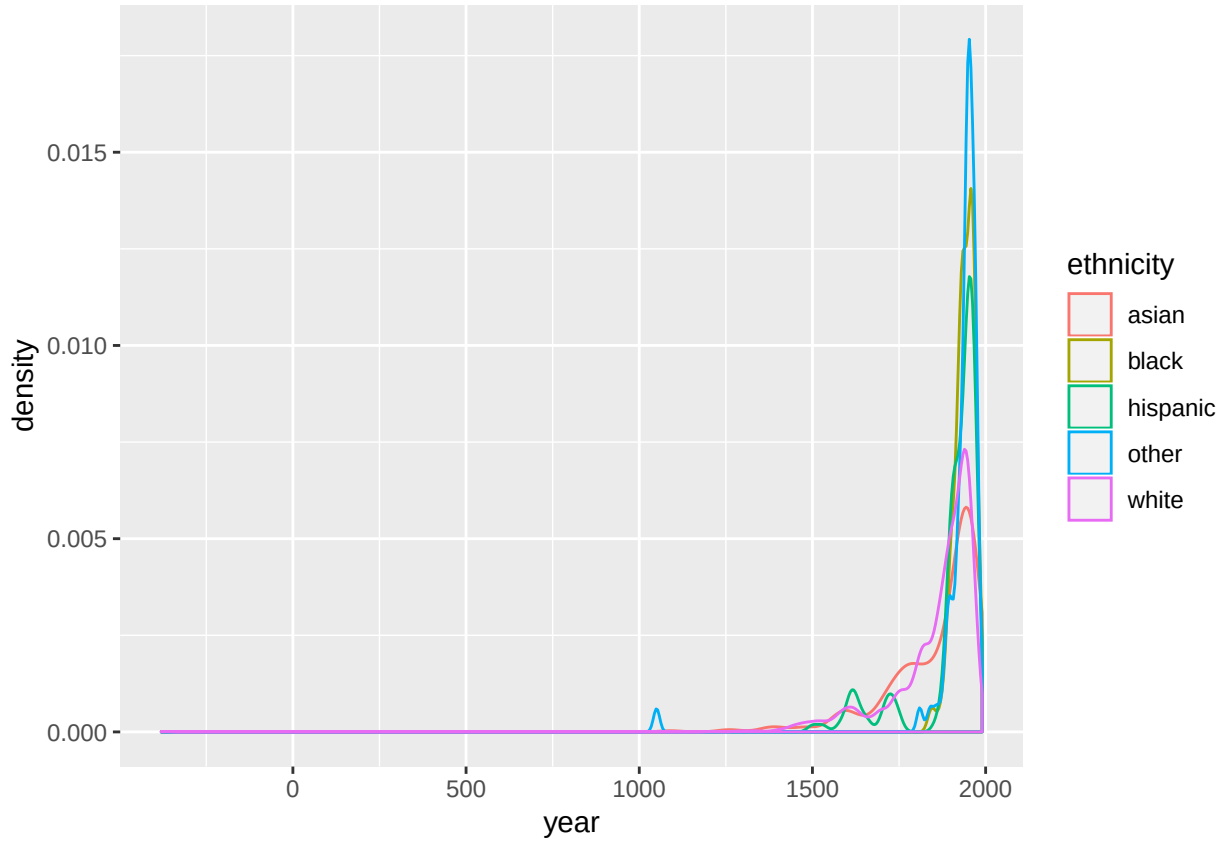
Step 2: Omit NA values and create a dataframe to store the ethnicity and year info

```
dist <- artist %>%  na.omit() %>% select(ethnicity, year)
head(dist)
```

```
##   ethnicity year
## 2     white 1880
## 3     white 1930
## 4     white 1790
## 6     white 1910
## 8     white 1810
## 9  hispanic 1900
```

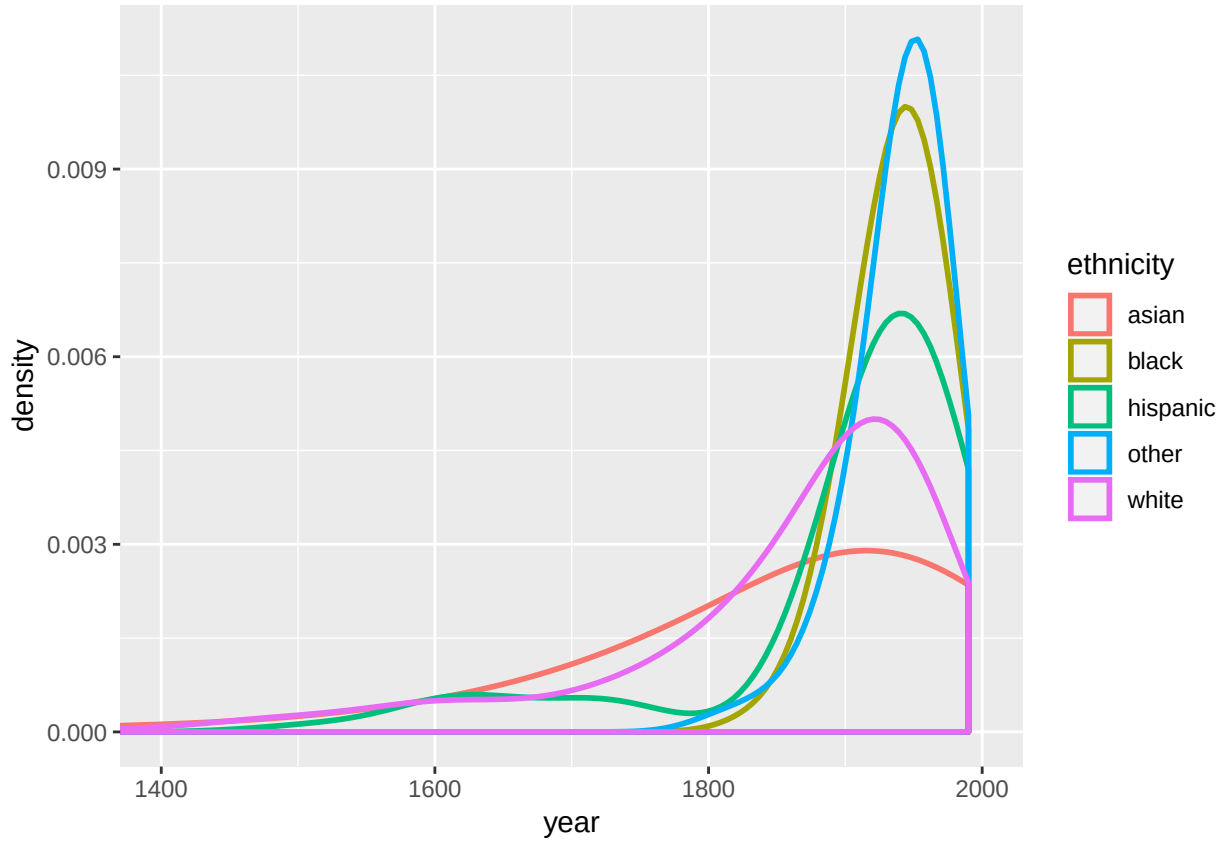
Step 3: Create a simple density plot to show the distribution of birth decade for various ethnicities of the artists

```
ggplot(data=dist, aes(x=year, color=ethnicity)) + geom_density()
```



Step 4 : Create a smoother density plot

```
ggplot(data=dist, aes(x=year, color=ethnicity)) + geom_density(size=1, adjust=3) +
  coord_cartesian(xlim=c(1400, 2000))
```



Step 5 : Conclusion

So, how do we interpret this plot? What does the height of each plot represent?

People often misinterpret the height of the density plot. A lot of them think that it represents probabilities/frequencies of the artists born in a particular birth decade. For example, we cannot say that the density of artists coming from other ethnicity (color in blue) in 1900 are 0.0045.

This is what a density plot for:

A Density Plot visualizes the distribution of data over a continuous interval or time period. An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram).

Therefore, the correct interpretation is that the plot shows a left-skewed distribution of birth decades for various artists coming from different ethnicities. Overall, we have more artists being born between 1800 - 2000.