MONASH University

ETC3250 Business Analytics: Plotting Data

Souhaib Ben Taieb, Di Cook, Rob Hyndman

August 31, 2015

What makes a data plot?

- data
- aesthetics: mapping of variables to graphical elements
- geom: type of plot structure to use
- layers

Example: Oscars

Extracted from The Oscars Database. `https://oscars.silk.co`. Best actors, actresses and directors, only.

Motivation "Tonight we honor Hollywood's best and whitest– sorry, brightest" (Neil Patrick Harris, Oscars 2015)

```
##              Var1 Freq
## 1           Asian    2
## 2           Black   14
## 3        Hispanic    8
## 4 Middle Eastern    1
## 5     Multiracial    2
## 6           White  306
```

Variables

- Categorical: Name, Sex, Birthplace, CityofBirth, State, Ethnicity, SexualOrientation, Religion, AwardCategory, Movie, Country
- Quantitative: Age, NumberofAwards
- Temporal: DOB, Year

Mapping hierarchy

Cleveland & McGill (1984)

Data element to graphical element in rank order of accuracy in returning data value, is as follows:

1. Position - common scale
2. Position - nonaligned scale
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color

Results corroborated by Heer & Bostock (2010).

Mapping hierarchy

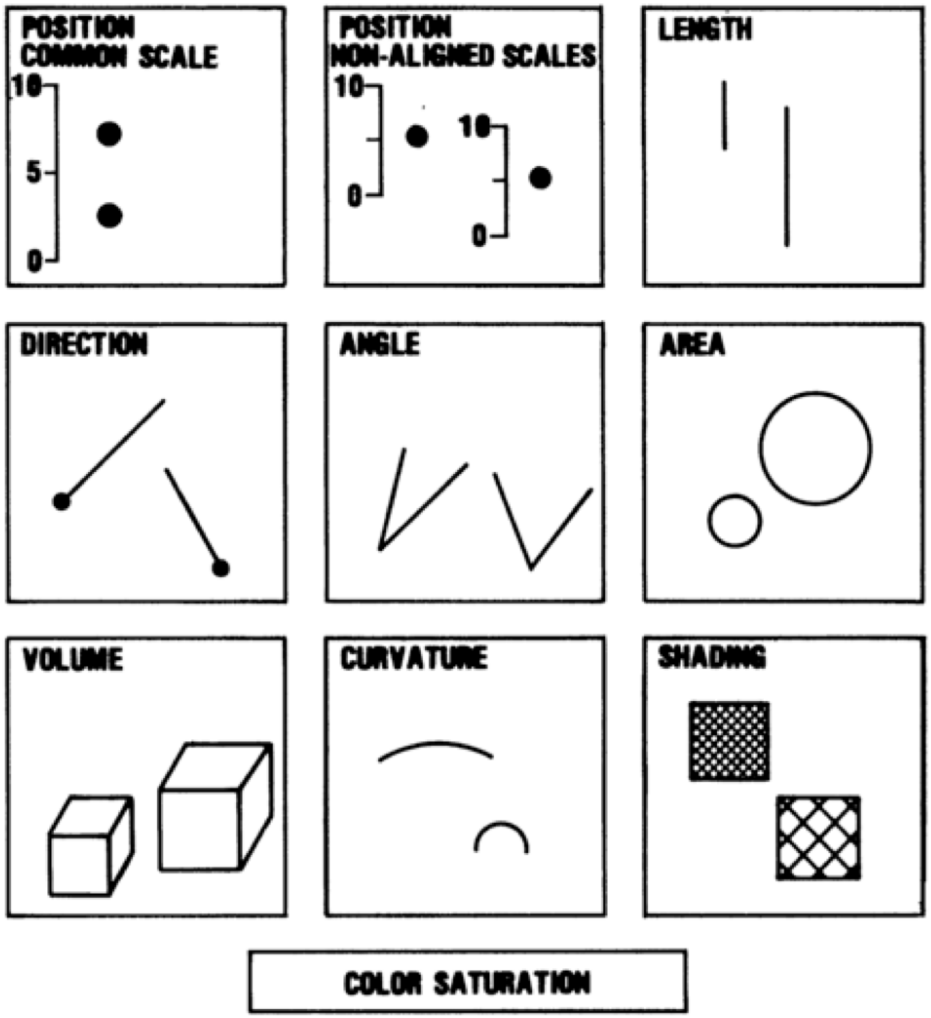


Figure 1. Elementary perceptual tasks.

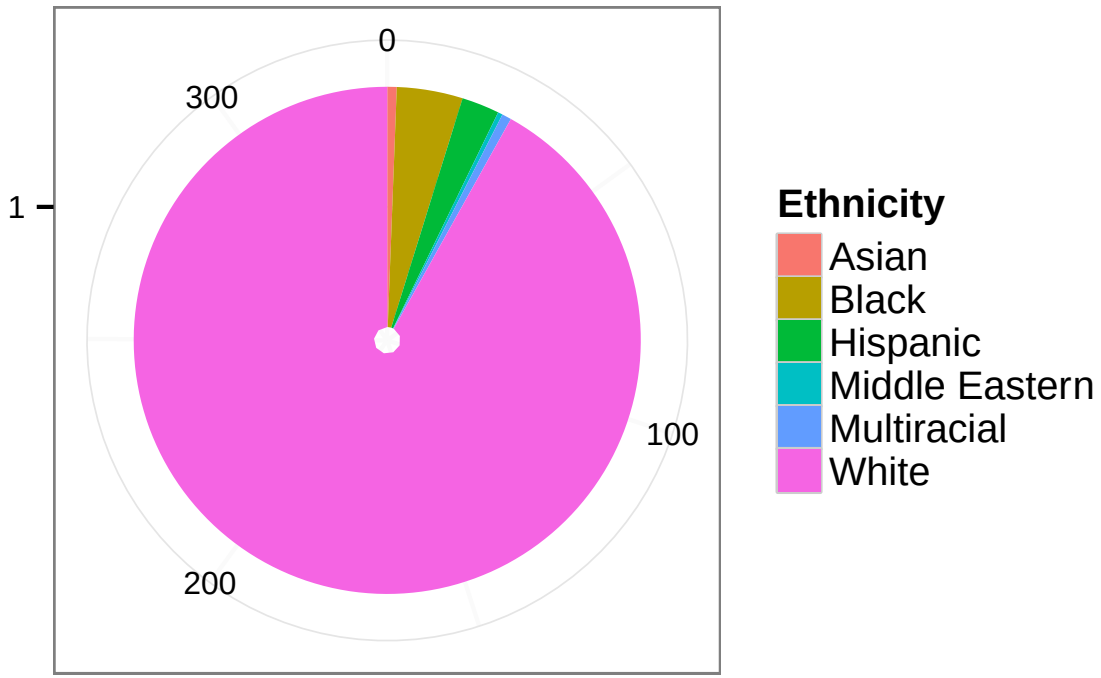
Basic mappings

Quantitative information should be mapped to position along a line, as first preference. Only two quantitative variables in the data: Age, NumberofAwards. Could treat year as quantitive to begin, too.

Unit of the data is person, award winner, aggregate this for categorical variables. Then we will have counts for categories, which is quantitative elements that should be mapped to position along a line.

Let's look at Ethnicity.

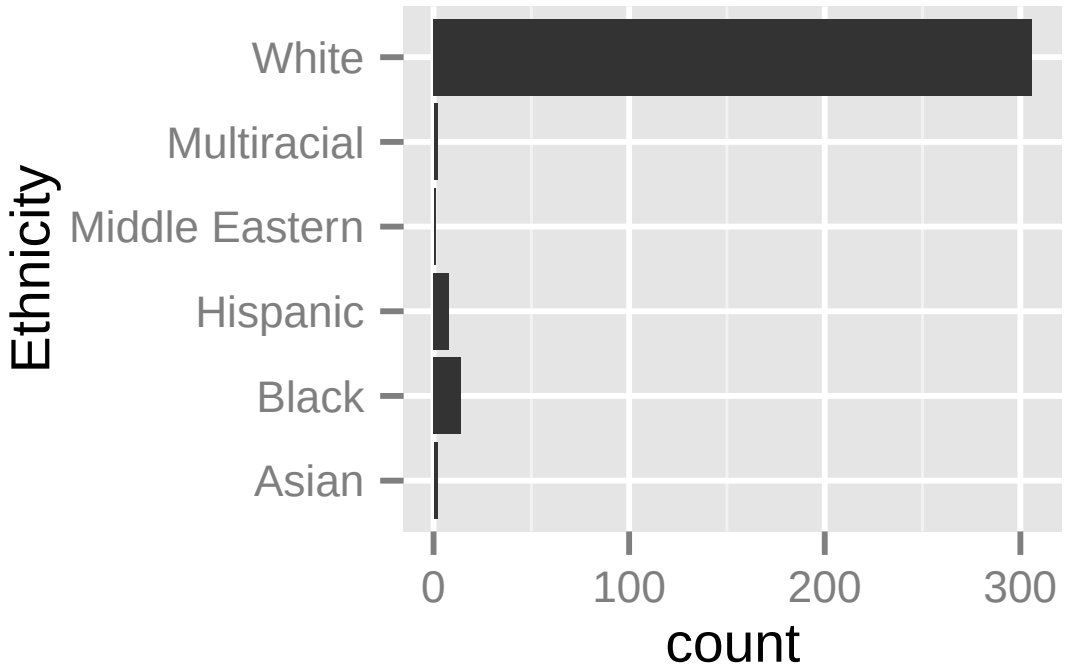
- In R, the best plotting package is `ggplot2`, good defaults (mostly) and conceptual structure underlying drawing data.
- How are data elements mapped to graphical elements?
- Are the number of Asian Oscar recipients higher than Hispanic?



Bar chart

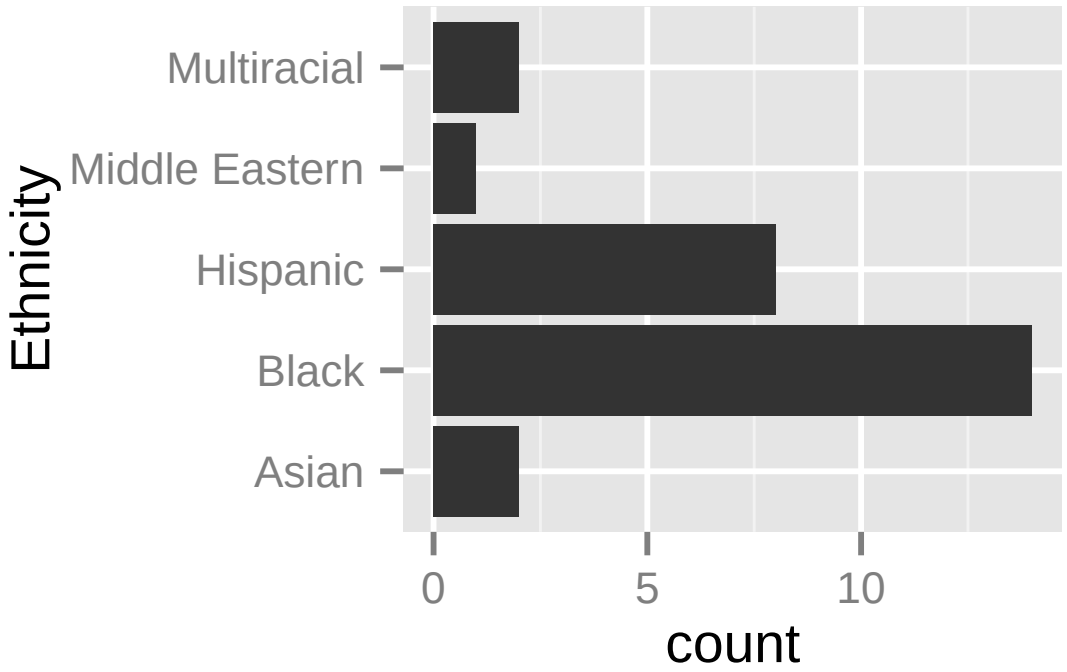
```
library(ggplot2)
qplot(Ethnicity, data=acting) + coord_flip()
```

- Are the number of Asian Oscar recipients higher than Hispanic?



Drill down

- The bar for whites is SO big, remove to focus on other categories.



Grammar

```
qplot(Ethnicity, data=acting)
```

- Data and variable are supplied. Implicit to these instructions, is the mapping and geom.

```
ggplot(data=acting) + geom_bar(mapping=aes(x=Ethnicity,
y=..count..))
```

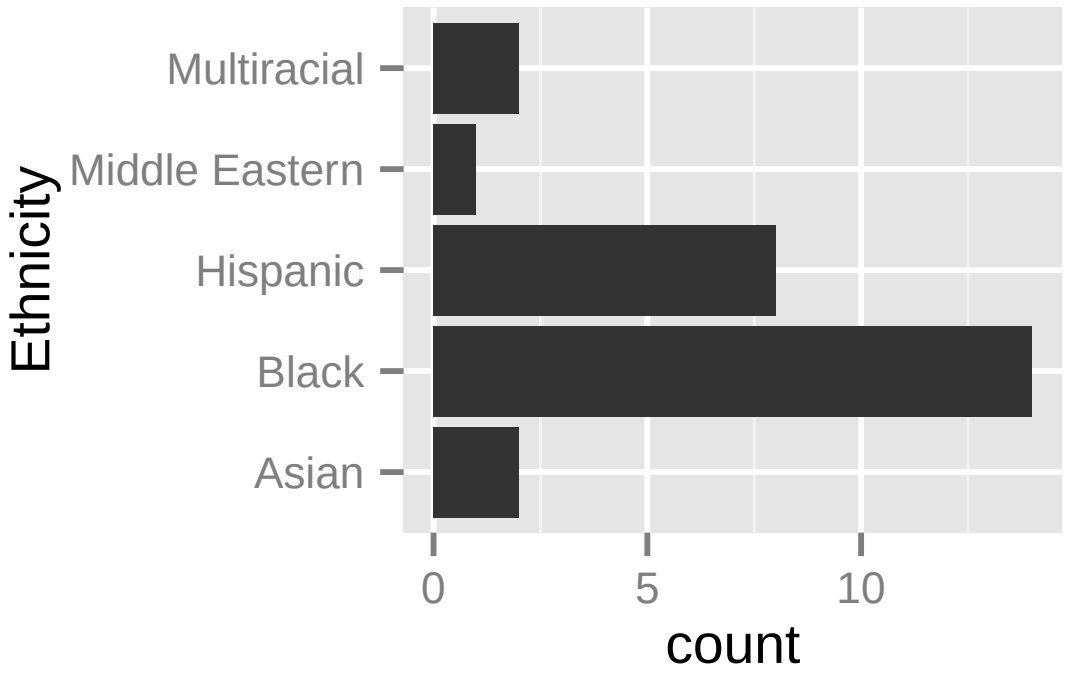
- `x` is mapped to categorical variable
- `y` is automatically calculated as the count of each level
- `count` is represented as a bar (read off using position along a line)

Order matters

- Sort categories by total count.

```
acting$Ethnicity <- factor(acting$Ethnicity,
      levels=c("White", "Black",
               "Hispanic", "Asian",
               "Multiracial",
               "Middle Eastern"))
qplot(Ethnicity, data=subset(acting,
                Ethnicity != "White"))
```

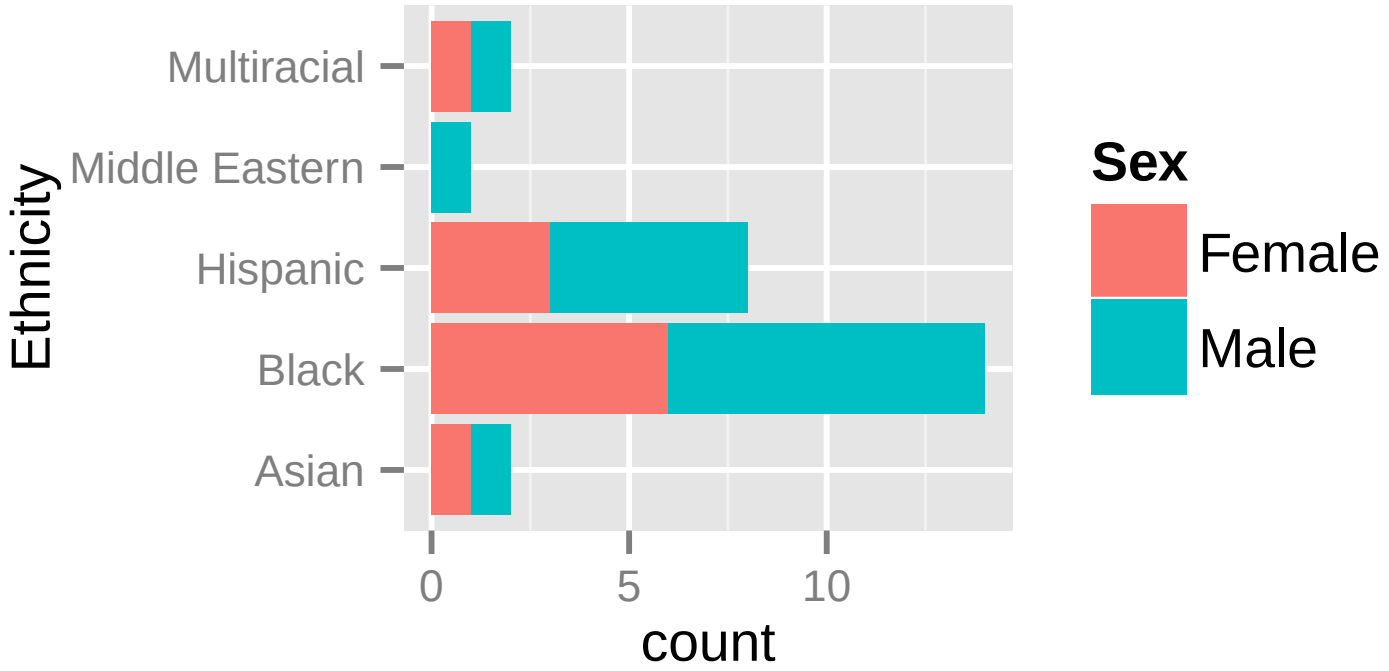
Ordered bars



Adding color

- Color bars by gender

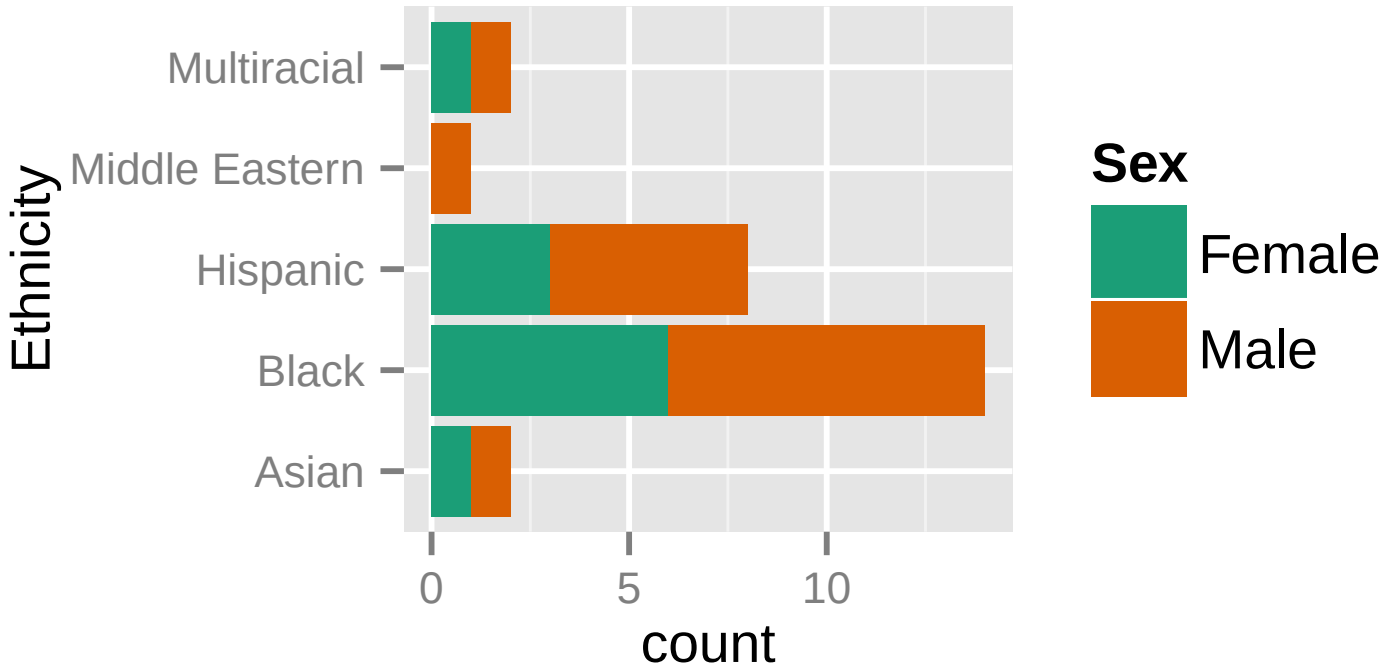
```
qplot(Ethnicity, data=subset(acting, Ethnicity != "White"),
  fill=Sex) + coord_flip()
```



Color choice

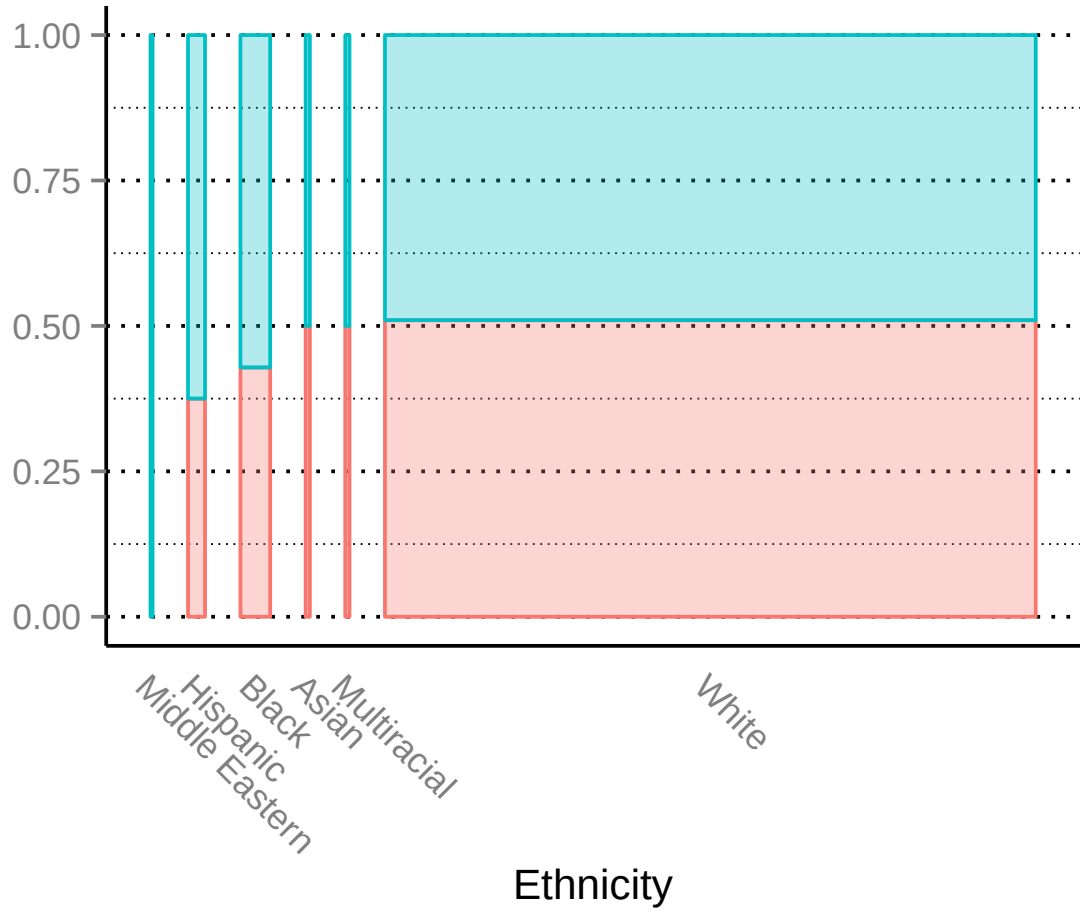
- There are several choices of color schemes: RColorBrewer developed for maps, and is good for area plots.

```
qplot(Ethnicity, data=subset(acting, Ethnicity != "White"),
  fill=Sex) + scale_fill_brewer(type="qual", palette=2) +
  coord_flip()
```



Mosaic

- To read proportions, along with counts of major variable, it is better to replace stacking with a mosaic plot (using `plotluck` package.
- Notice that bars are now sorted by proportion.

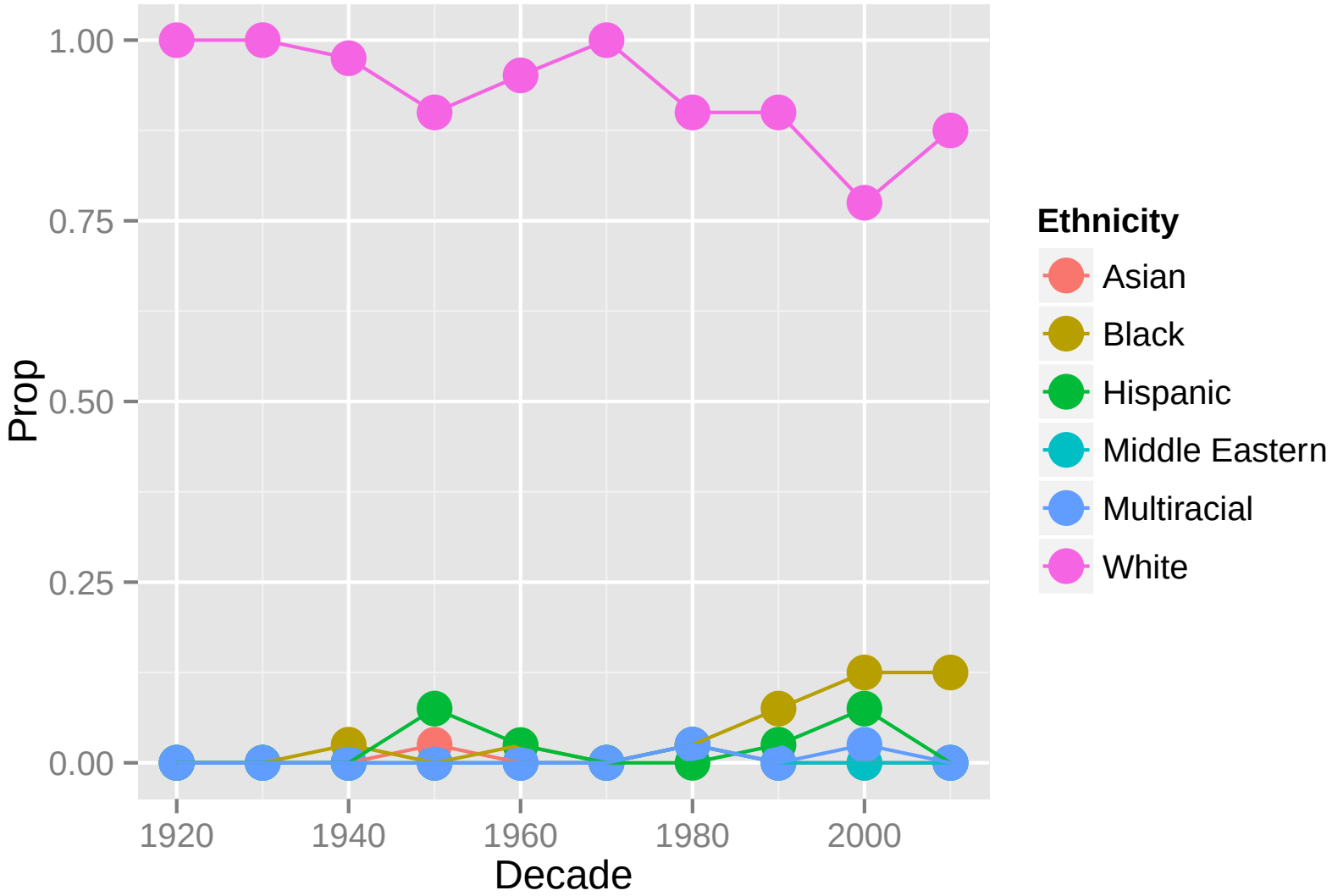


Temporal Trend

- These numbers are aggregated across almost 100 years of Oscar winners.
- Maybe the proportions have changed over time.
- We will aggregate by decade, and compute proportions for each ethnic class, and take a look at these.

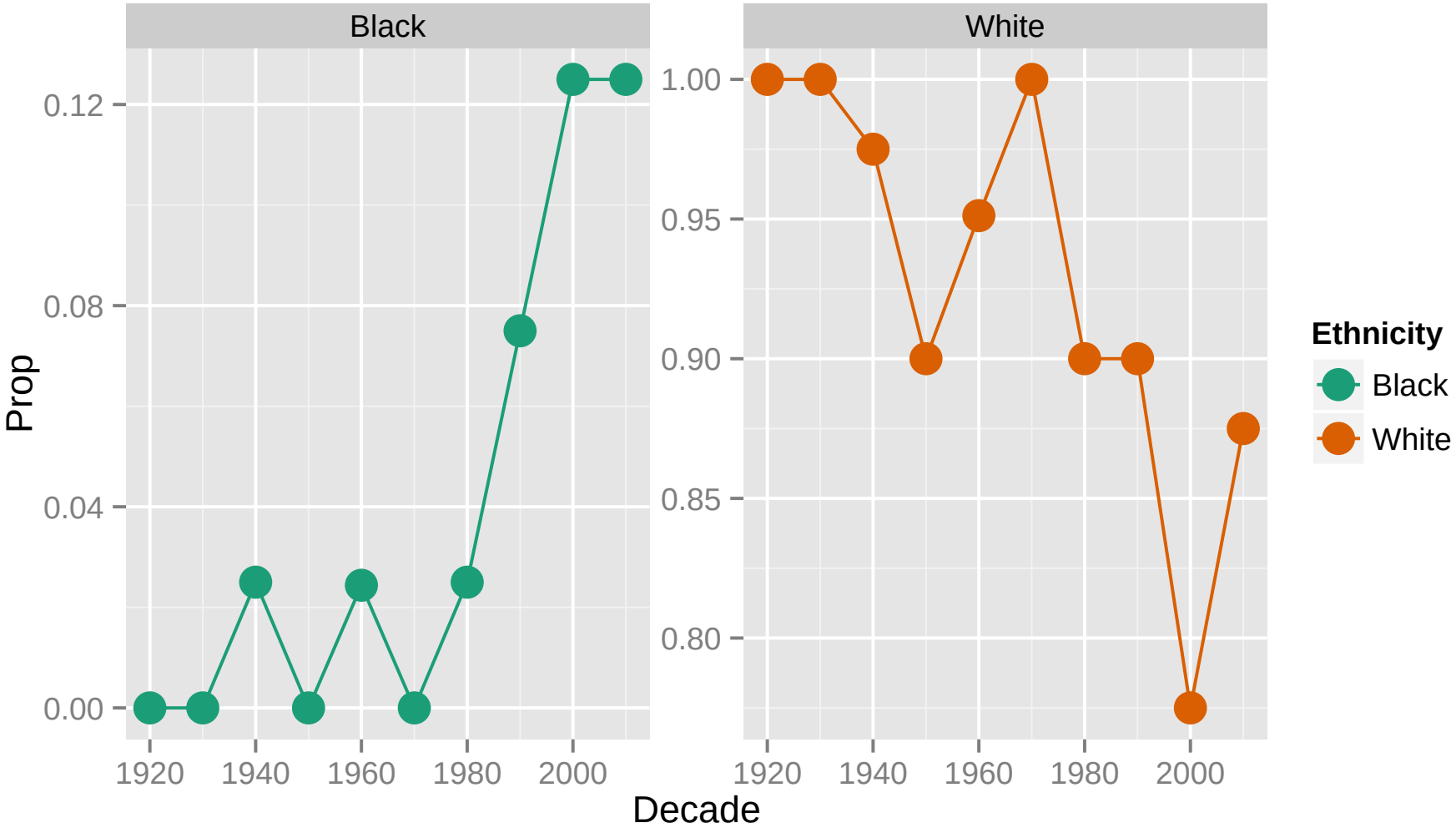
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1929    1953    1974    1974    1995    2015
```

Temporal Trend



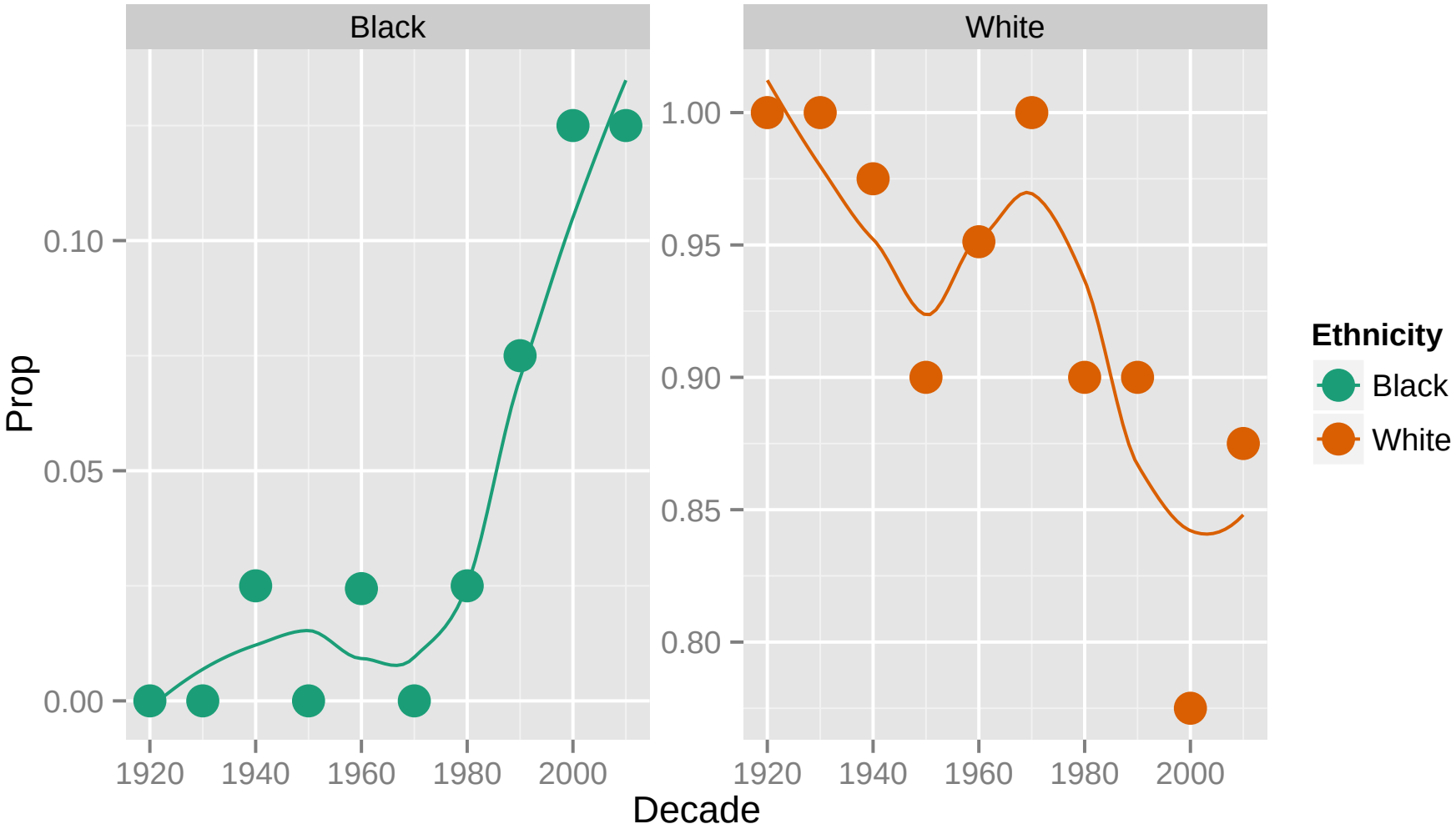
Facetting

- Whites dominate again, so lets use facetting to zoom in. Focus only on ethnicities with reasonable numbers.



Examine trend

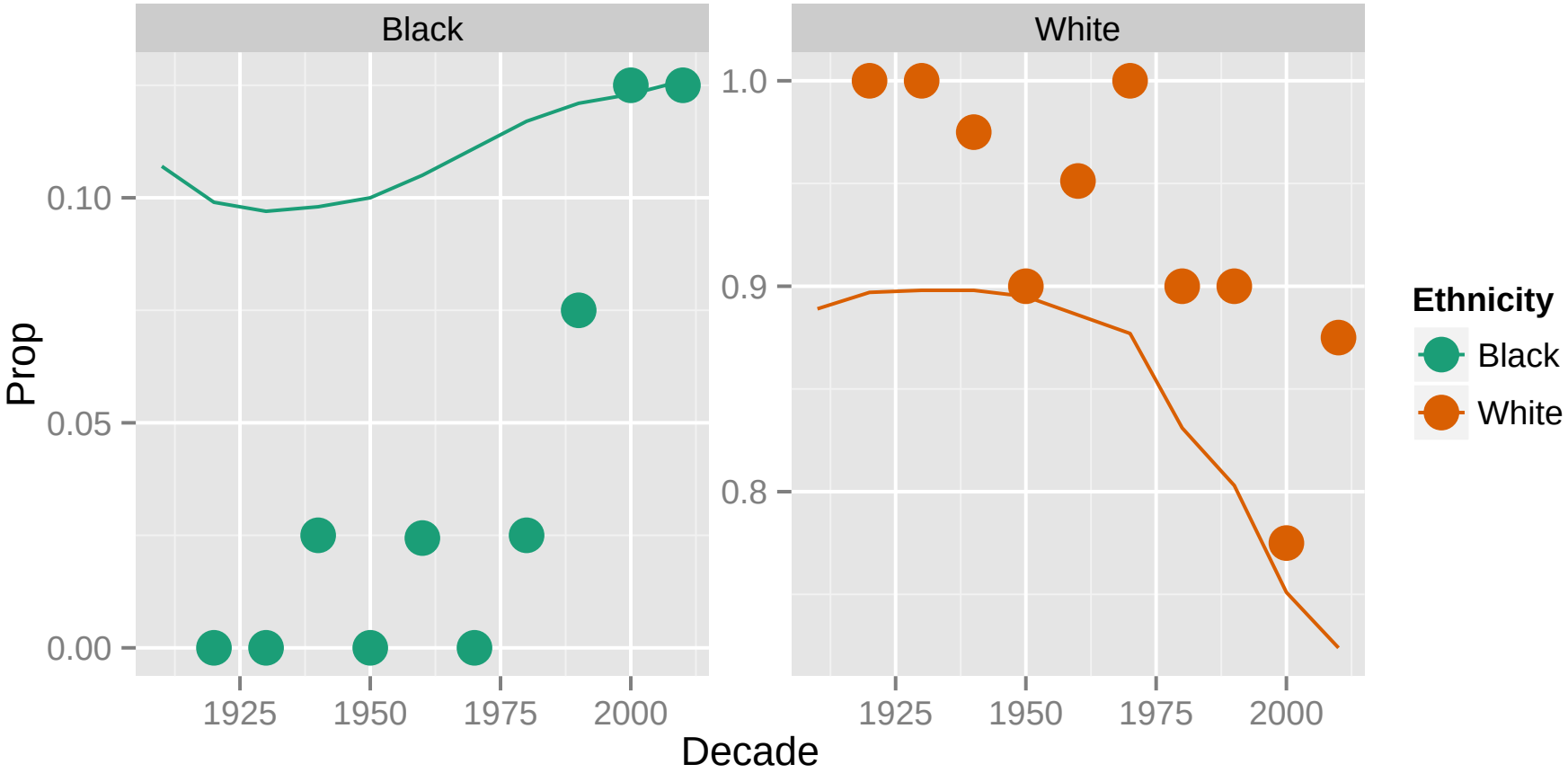
- Use a trend line, loess, instead of connecting the dots.



Layering

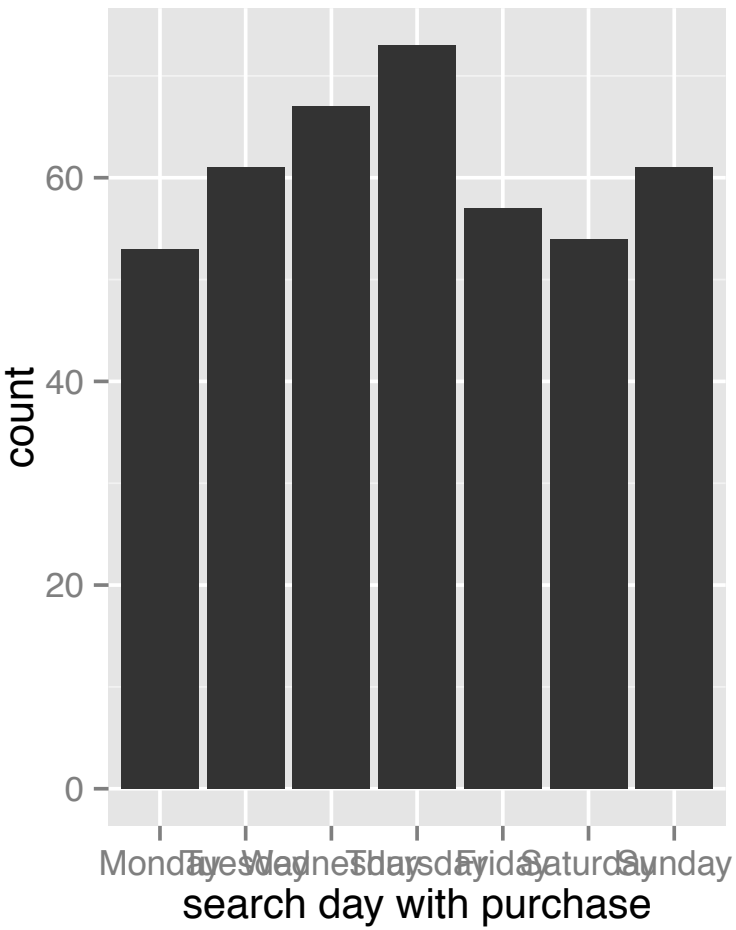
- Pull demographics data from `http://en.wikipedia.org/wiki/Historical_racial_and_ethnic_demographics_of_the_United_States`.
- Overlay these values as a line plot on the Oscars proportions.

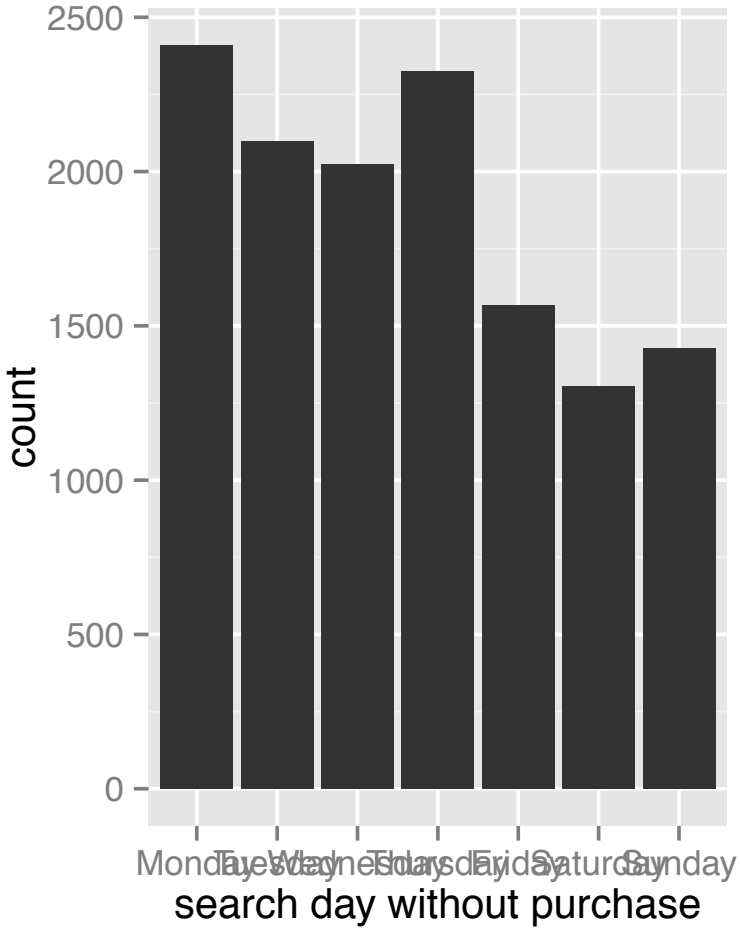
Layering


Black
White
Prop
0.10
0.05
0.00
1.0
0.9
0.8
1925
1950
1975
2000
Decade
Ethnicity
Black
White

Perceptual principles: proximity

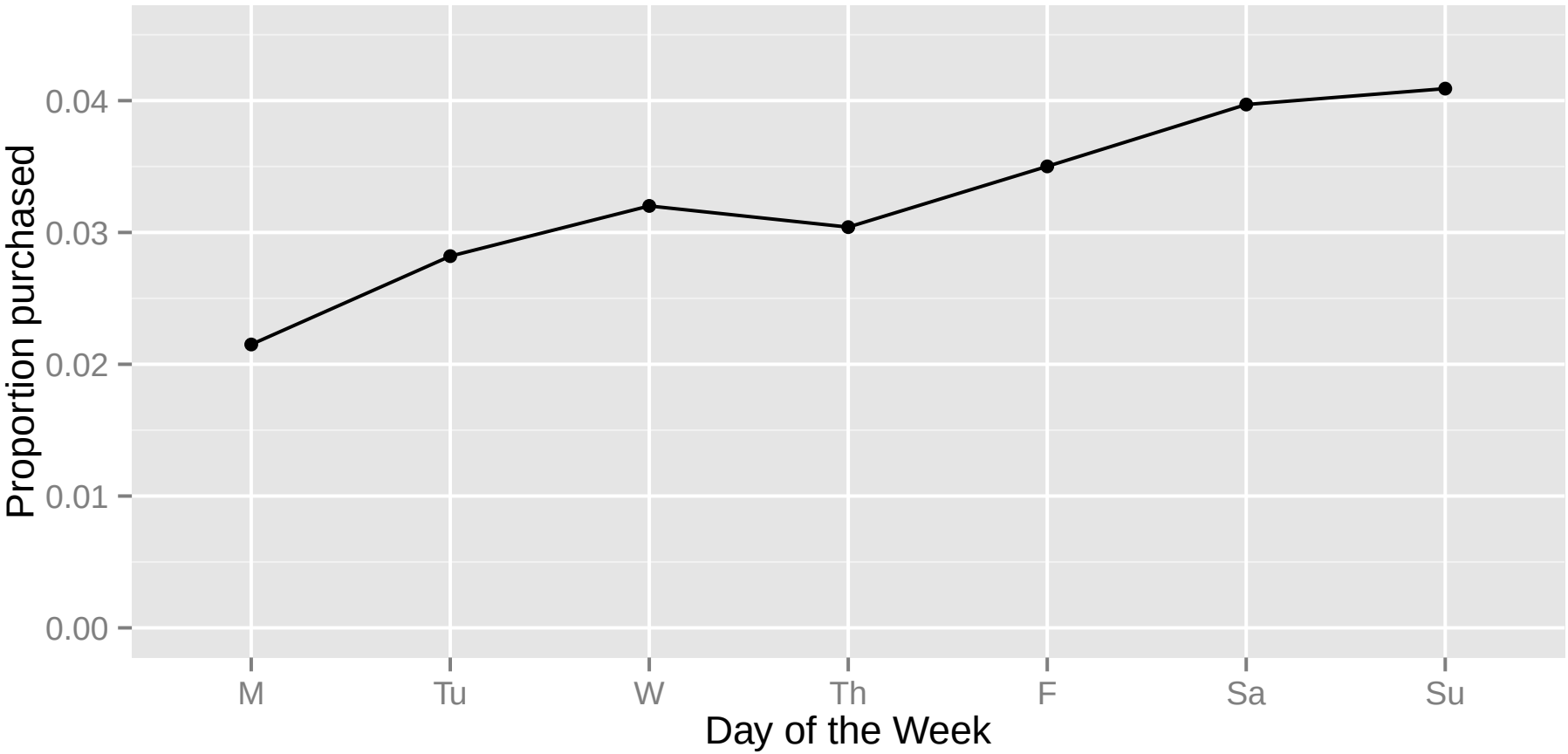
- Place elements for primary comparison close together.
- Is a week day/weekend day pattern?
- Better to plot proportions.





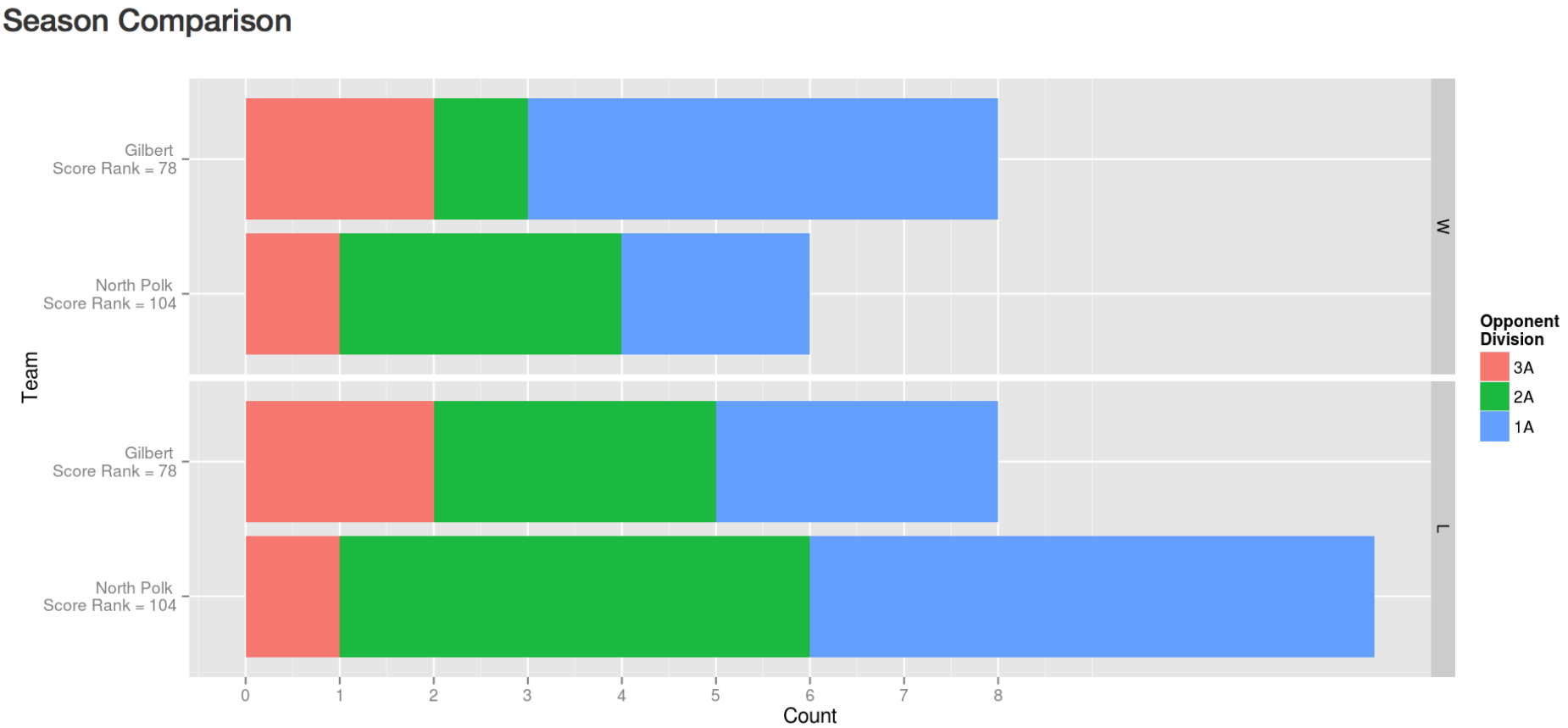
Perceptual principles: proximity

Rearrange into proportions, and plot these as points. The question "Is there a difference in purchase rate by day of search?" Proportions better answers this question. (Numbers for each day of the week are large enough for proportion to be reliably interpreted.)



Perceptual principles: proximity

Primary purpose is to examine wins and losses of two teams. Closest bars are the two teams, faceted by win/loss.



Perceptual principles: color

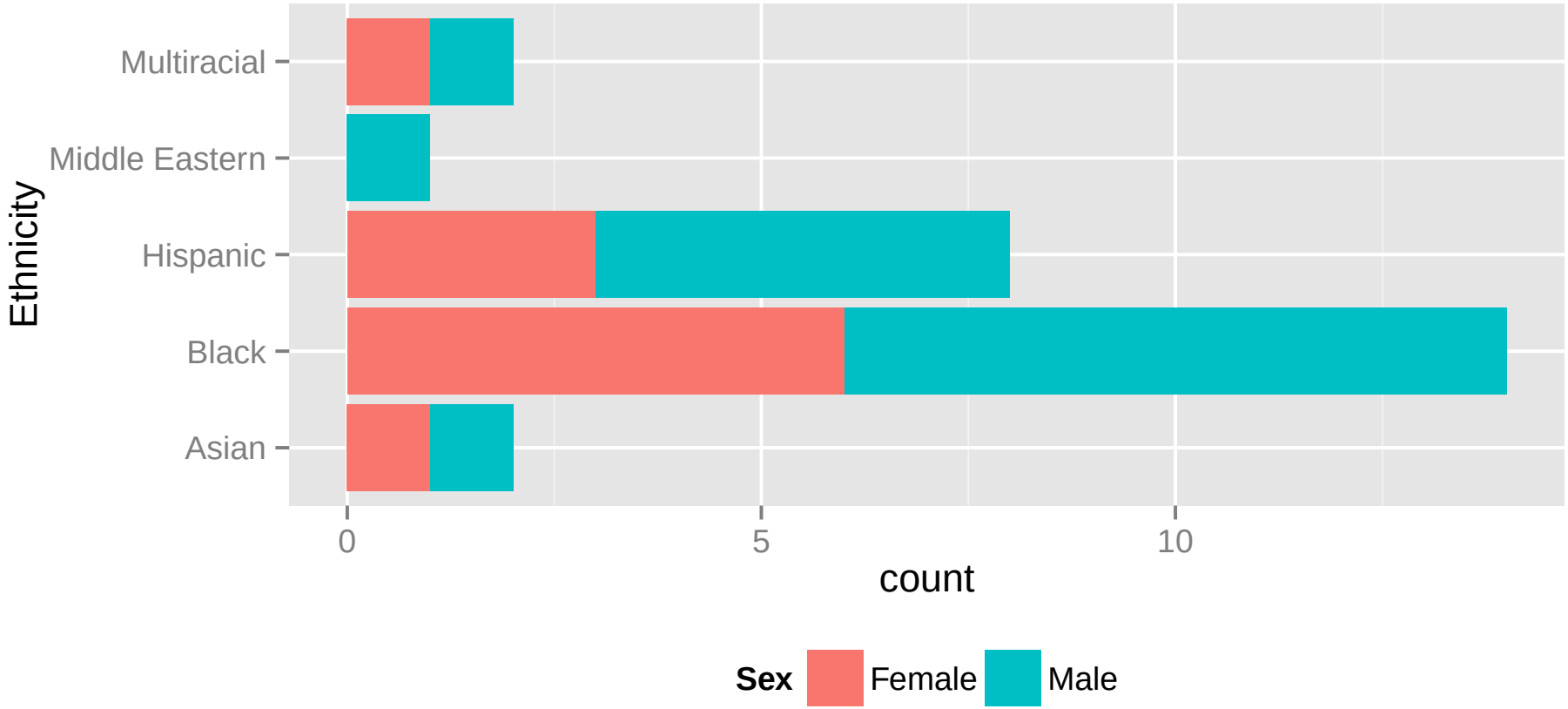
Applying color takes some care:

1. Color is a pre-attentive graphical element, which means people see it before they realize they see it. If most elements are blue and one is red we will pick the red element out SO QUICKLY.
2. Color is low on the scale of accurate reading. Use sparingly.
3. Map the appropriate information to the appropriate scale:

- Use QUALITATIVE scales to display categorical data, small number of levels,
- SEQUENTIAL scales are used to represent a gradient of quantitative information, e.g. 0-100
- DIVERGING scales are used to represent negative to positive quantitative information.

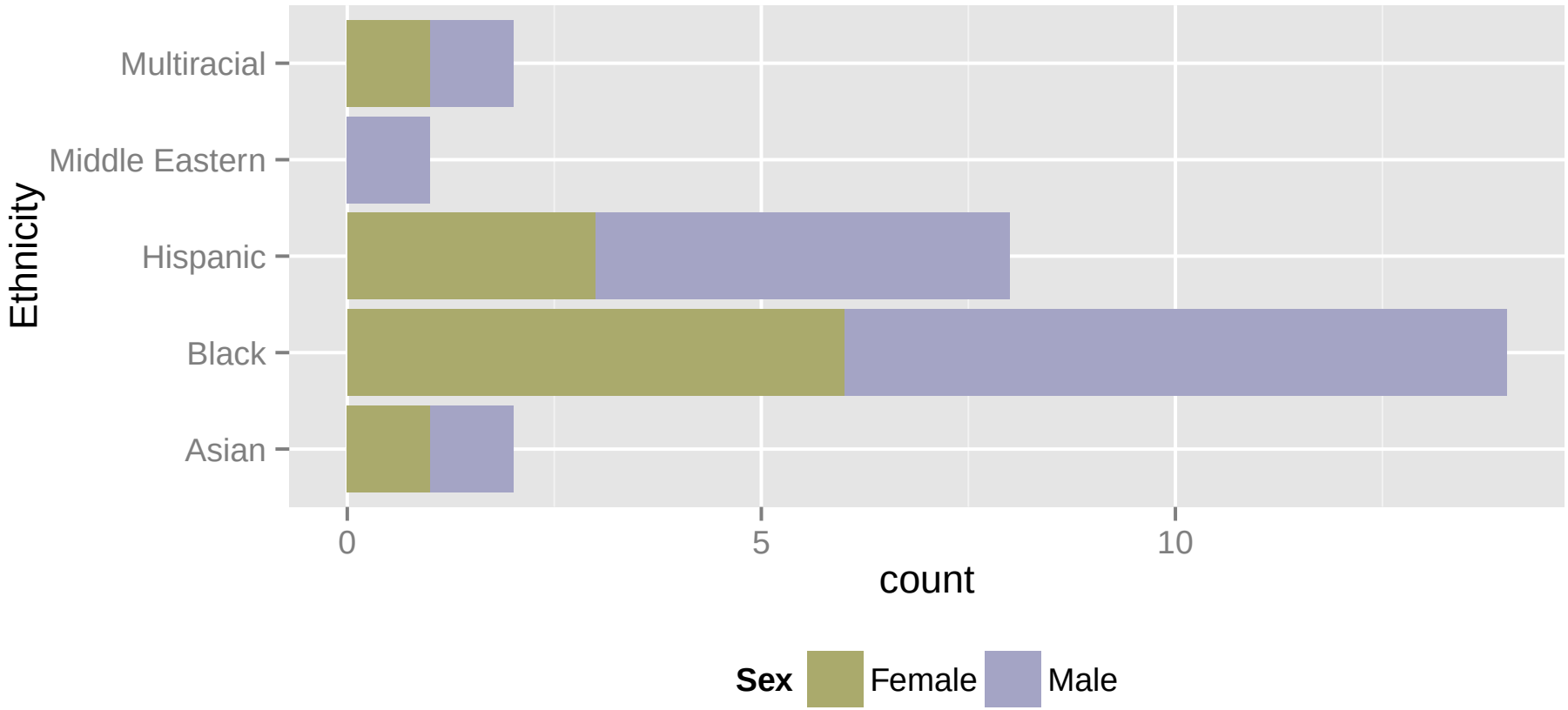
Color blindness

Color choice can affect whether all your audience can see what you see.



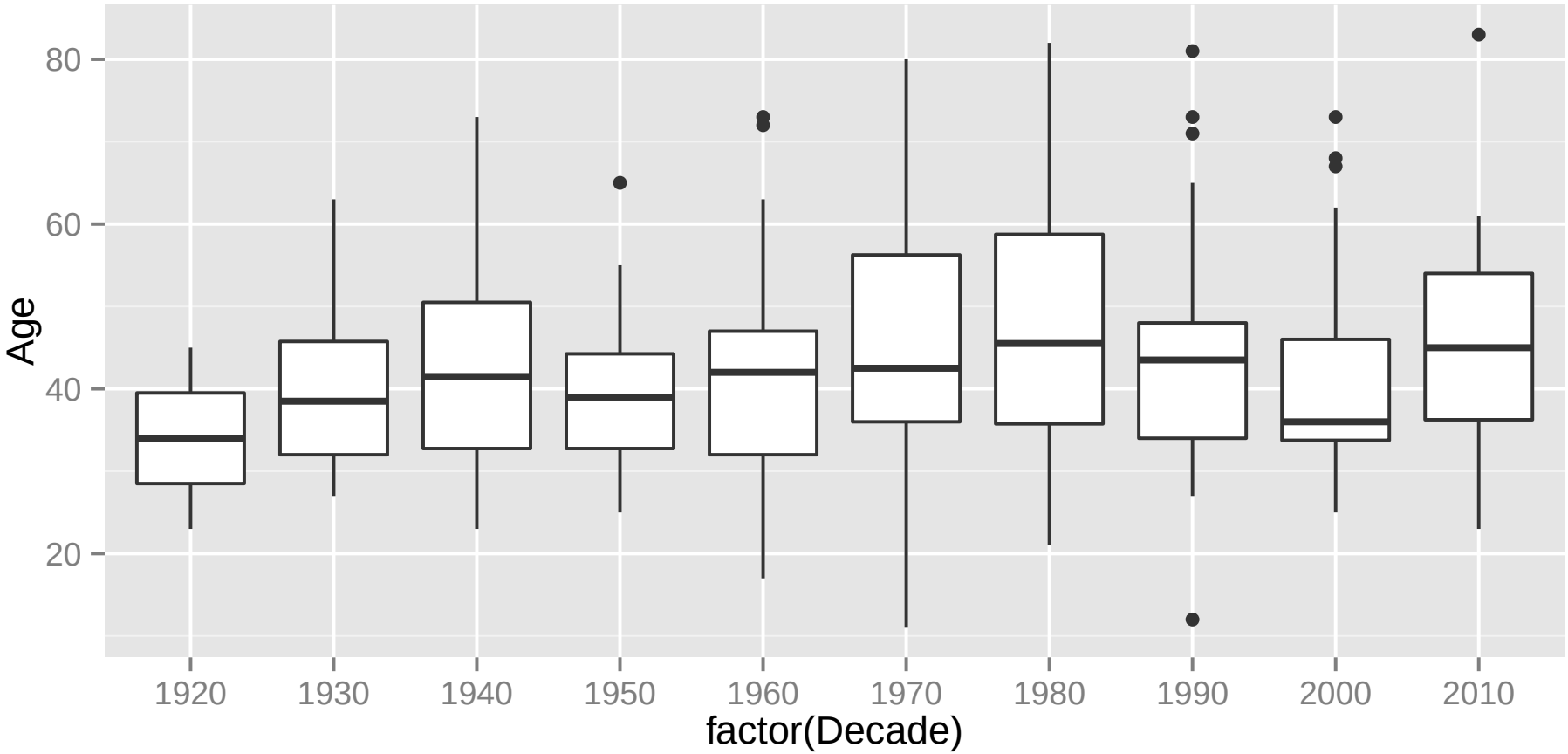
Color blind-proofing

Using the package `dichromat`, to see what this looks like to a red-green color blind eye.

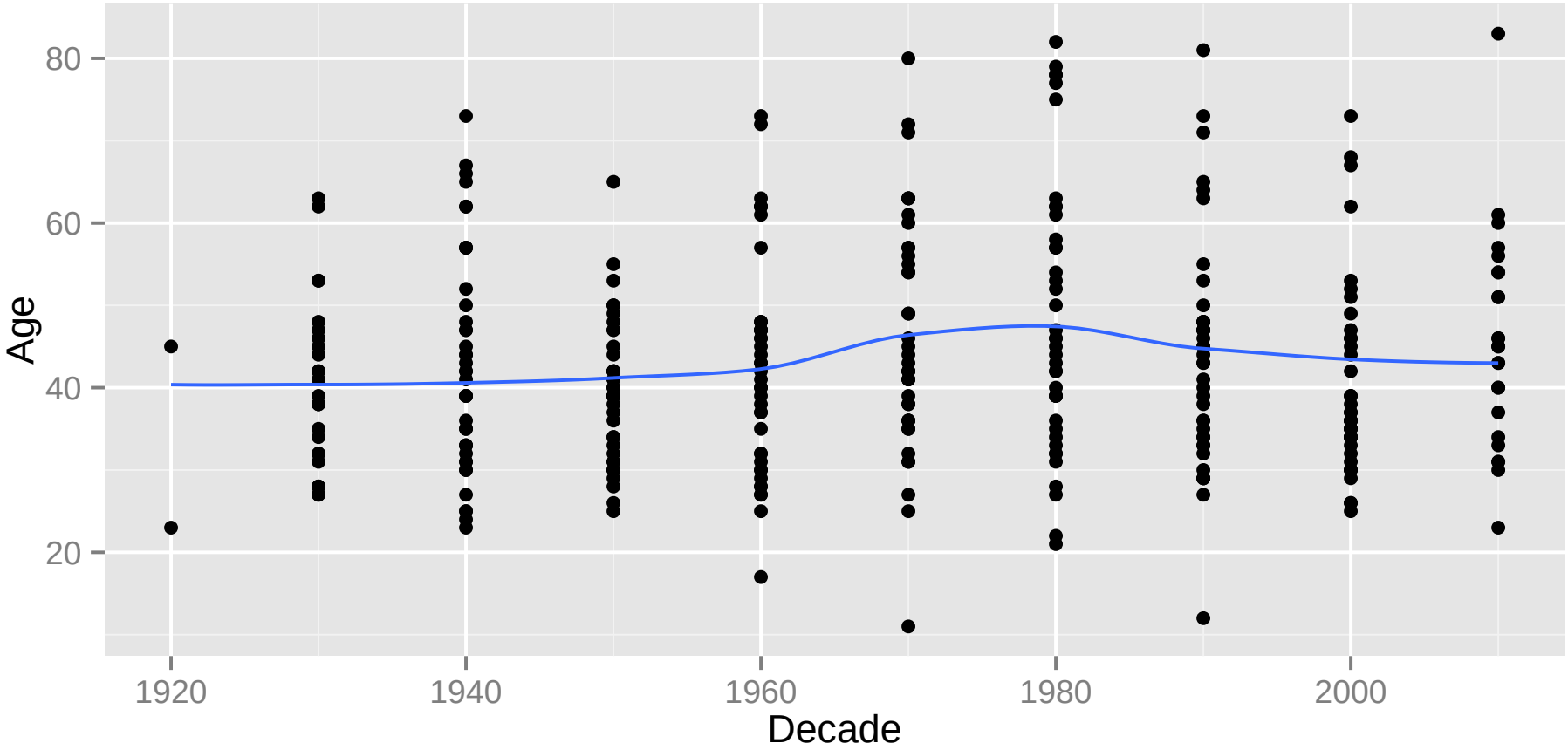


Exploring the Oscars

- How does the age of the winners change, on average by gender, over time?

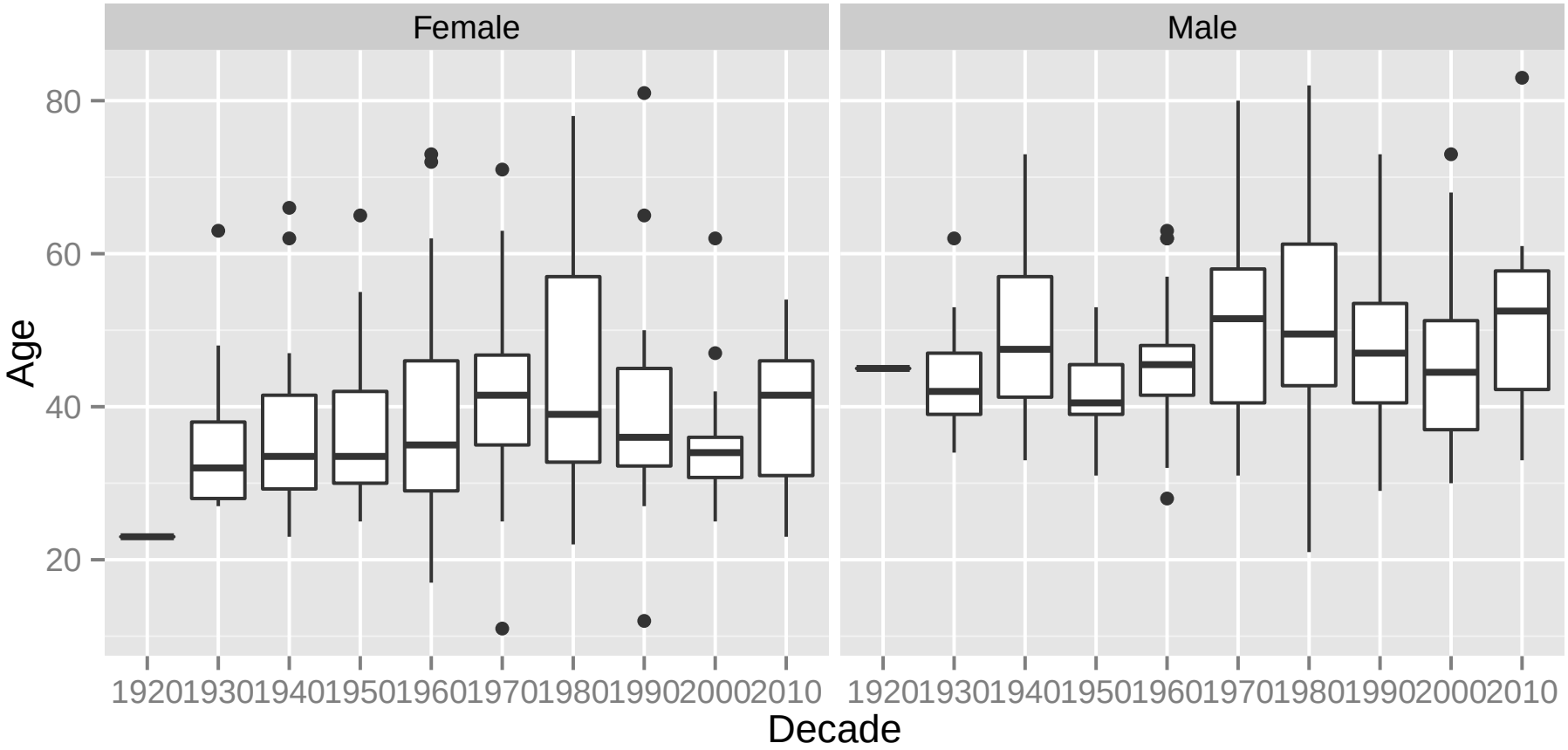


Exploring the Oscars



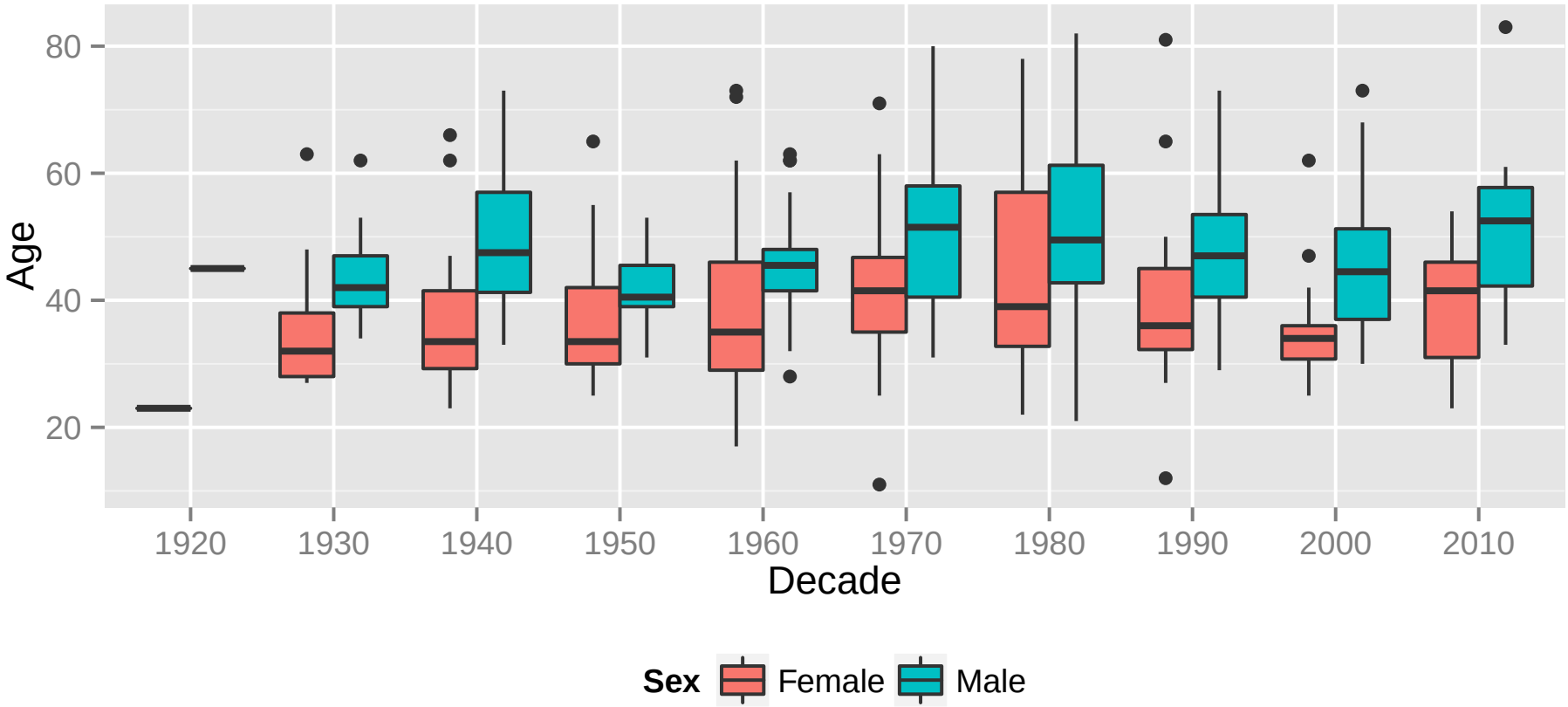
Exploring the Oscars

- Is there a difference between best actor vs best actress?



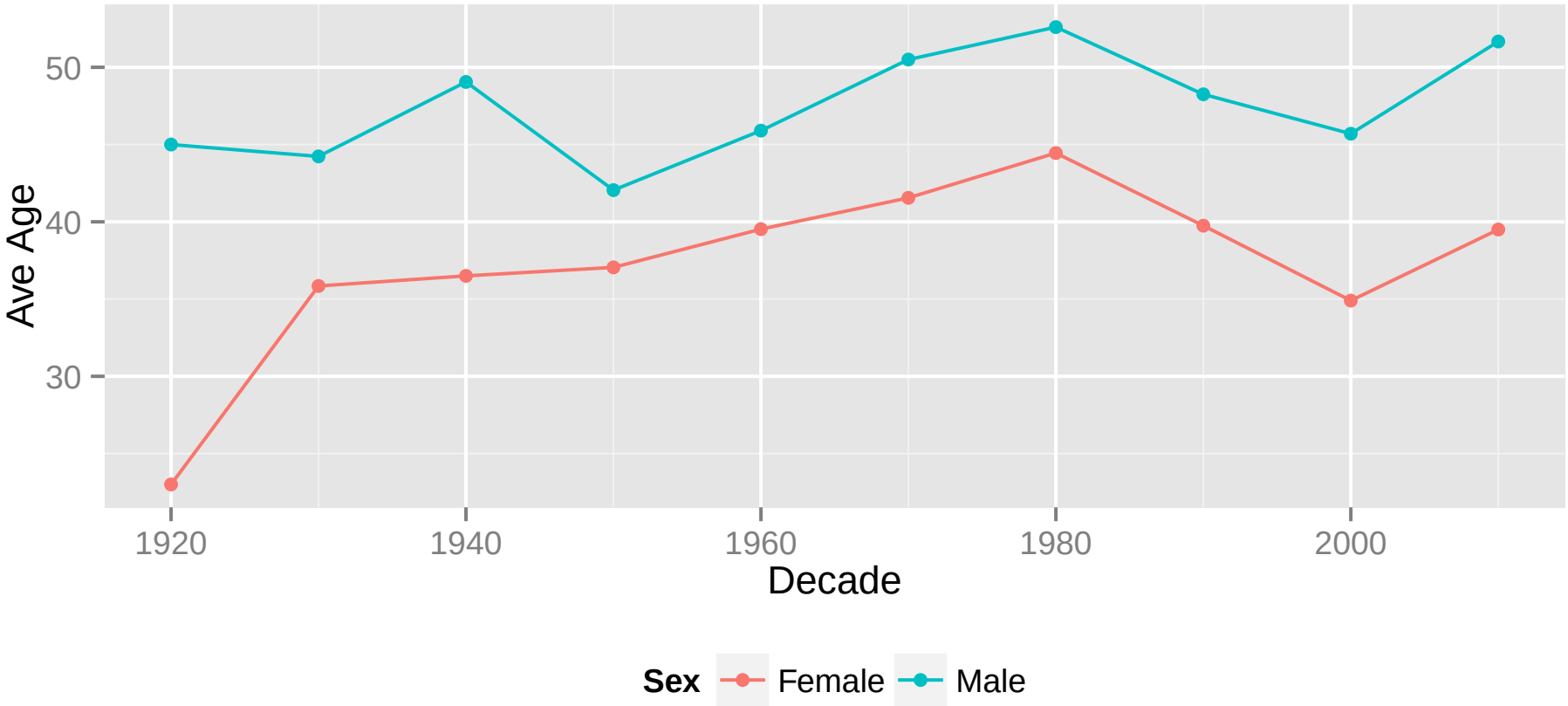
Exploring the Oscars

- Difference between best actor vs best actress ages, by decade?



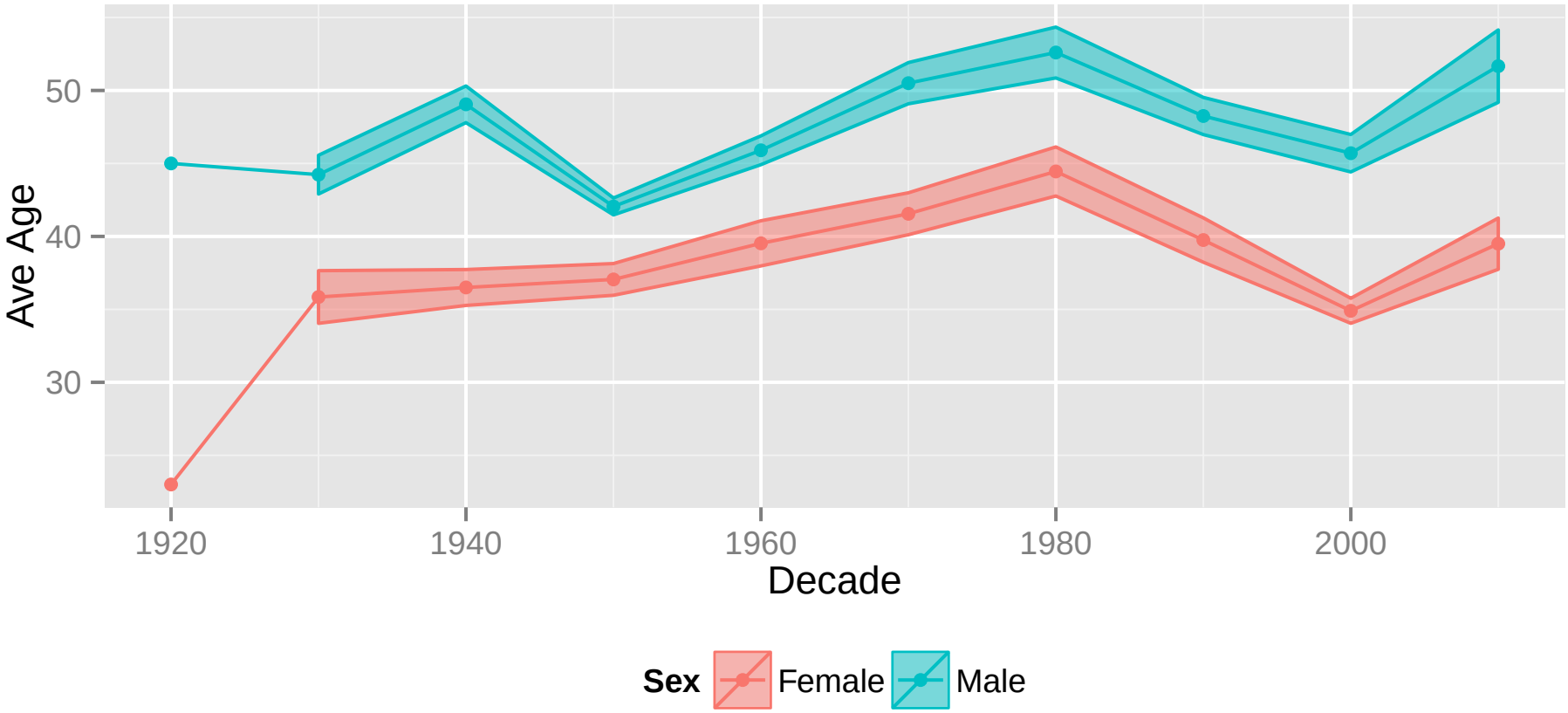
Showing the statistics

- Compute the mean for each decade and show this



Showing the statistics

- Add a 95% point-wise confidence interval



Resources

- 'R Graphics Cookbook' Winston Chang `http://www.cookbook-r.com/Graphs/`
- ggplot2 `http://ggplot2.org/` (On the graphics package specifics)
- `https://github.com/garrettgman` (Basic R usage)
- `http://yihui.name/knitr/` (Rmarkdown usage)
- `http://stackoverflow.com` (Q/A site with LOTS of useful info)
- 'Creating More Effective Graphs' Naomi Robbins `http://www.nbr-graphs.com` (Conceptual)
- `http://www.csc.ncsu.edu/faculty/healey/PP/` (Perceptual principles)
- `http://www.simonslab.com/videos.html` (Perception)