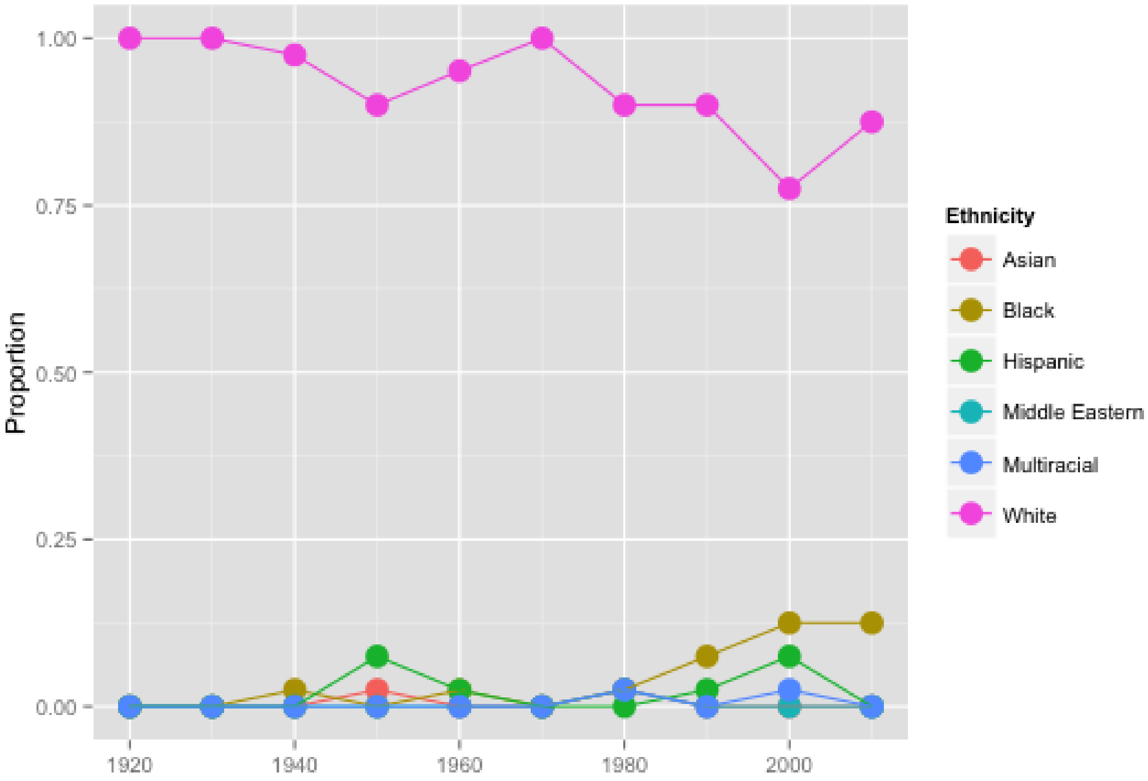
Plotting Data

Di Cook, Eric Hare

May 14, 2015



California Dreaming - ASA Travelling Workshop

Using the package ggplot2

Elements of a plot

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

Data - Oscar winners

Extracted from The Oscars Database. https://oscars.silk.co.

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

```
oscars <- read.csv("data/oscars.csv", stringsAsFactors=FALSE)
acting <- subset(oscars, AwardCategory=="Actor")
table(acting$Ethnicity)
```

```
## 
##           Asian          Black       Hispanic Middle Eastern    Multiracial 
##               2             14              8              1              2 
##           White 
##             306 
```

Variables

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

Mapping principles

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).

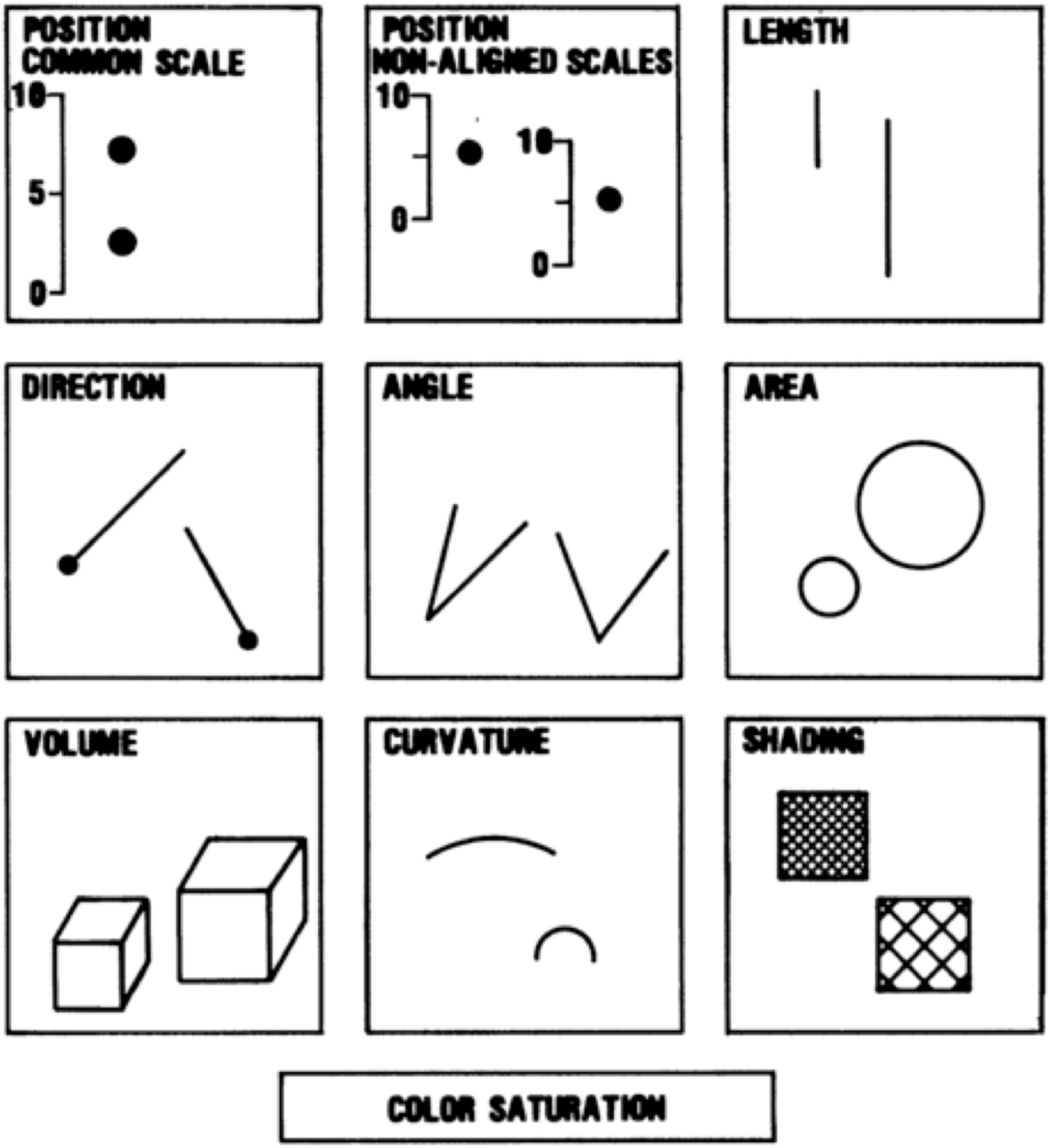


Figure 1. Elementary perceptual tasks.

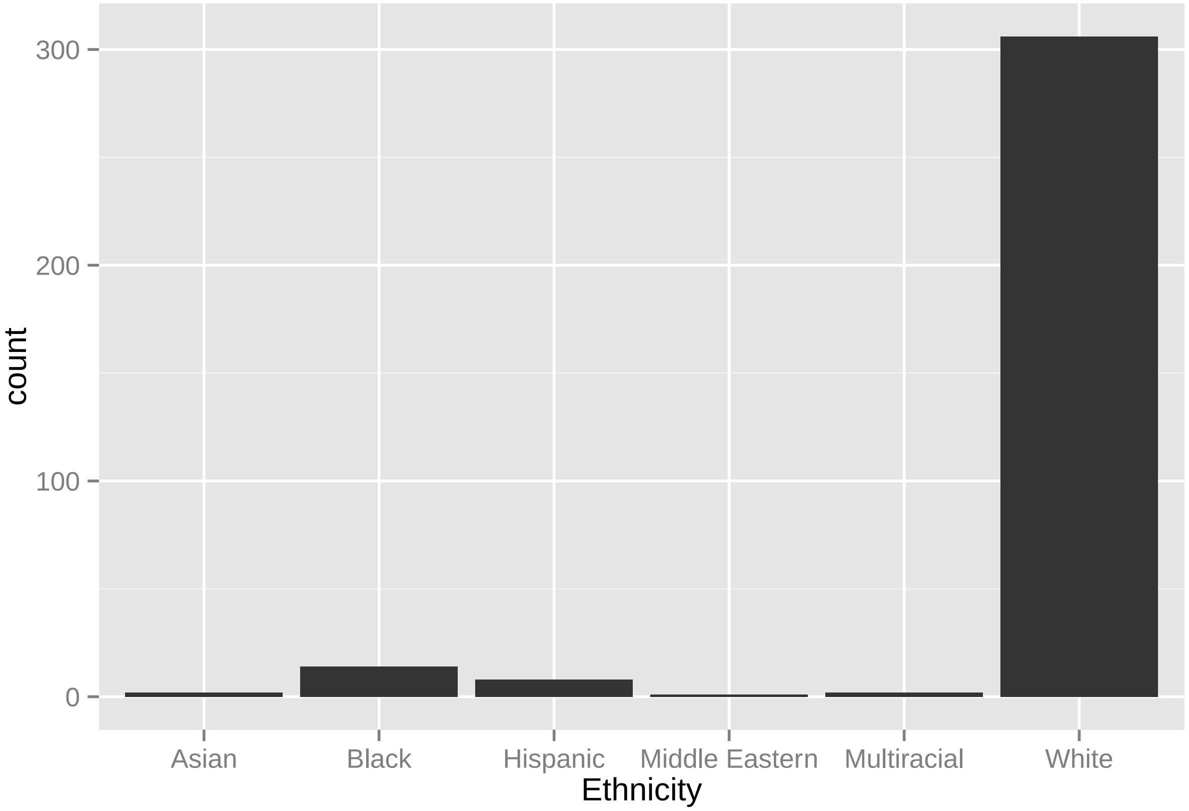
Basic mappings

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.

But unit of the data is person, award winner, need to aggregate based on categorical variables. Then were will have counts for categories, which is quantitative information.

Let's look at Ethnicity.

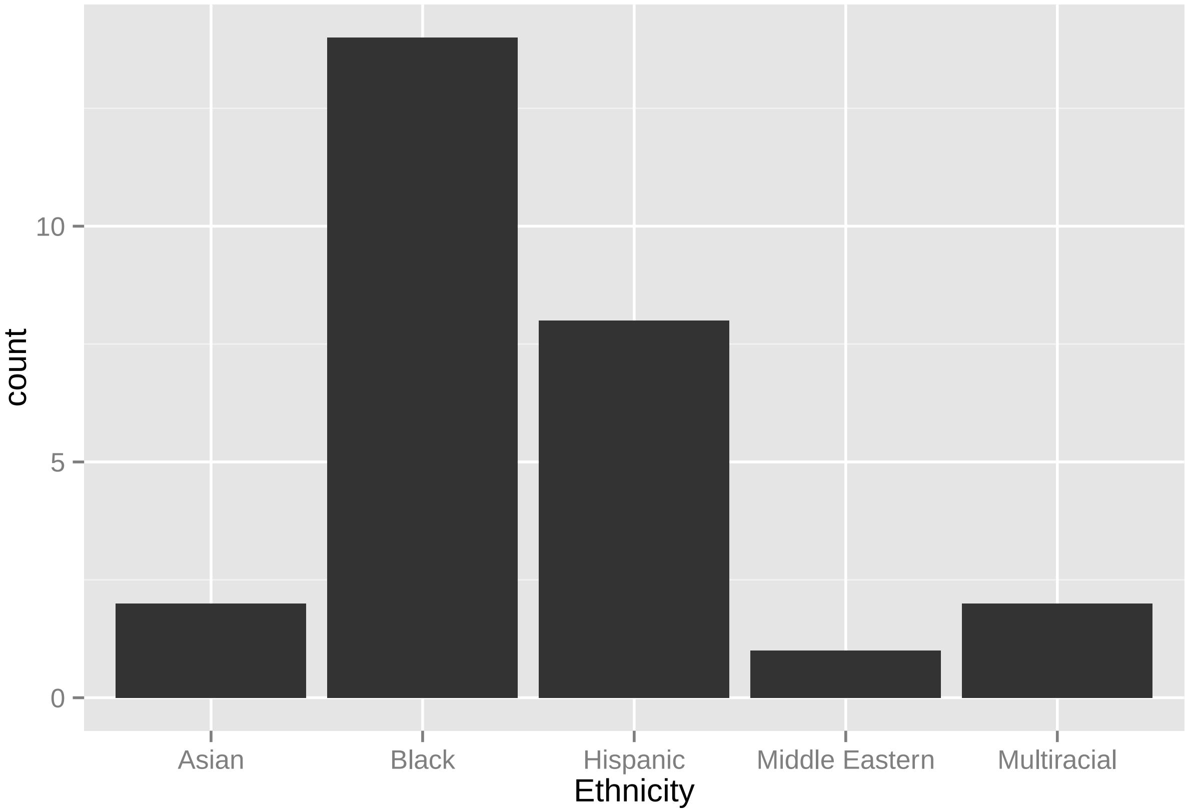
```r
library(ggplot2)
qplot(Ethnicity, data=acting)
```



Drill down

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

```r
qplot(Ethnicity, data=subset(acting, Ethnicity != "White"))
```



Grammar

`qplot(Ethnicity, data=acting)`

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

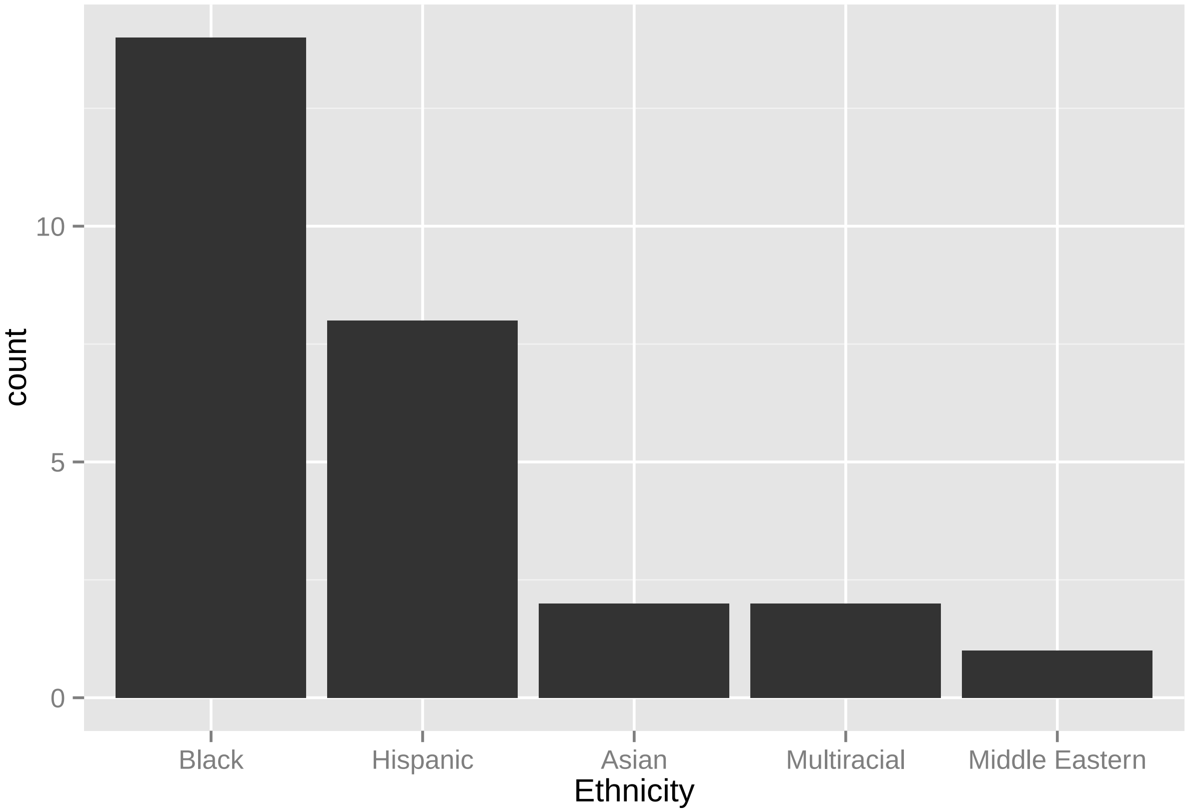
`ggplot(data=acting) + geom_bar(mapping=aes(x=Ethnicity, y=..count..))`

- `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.

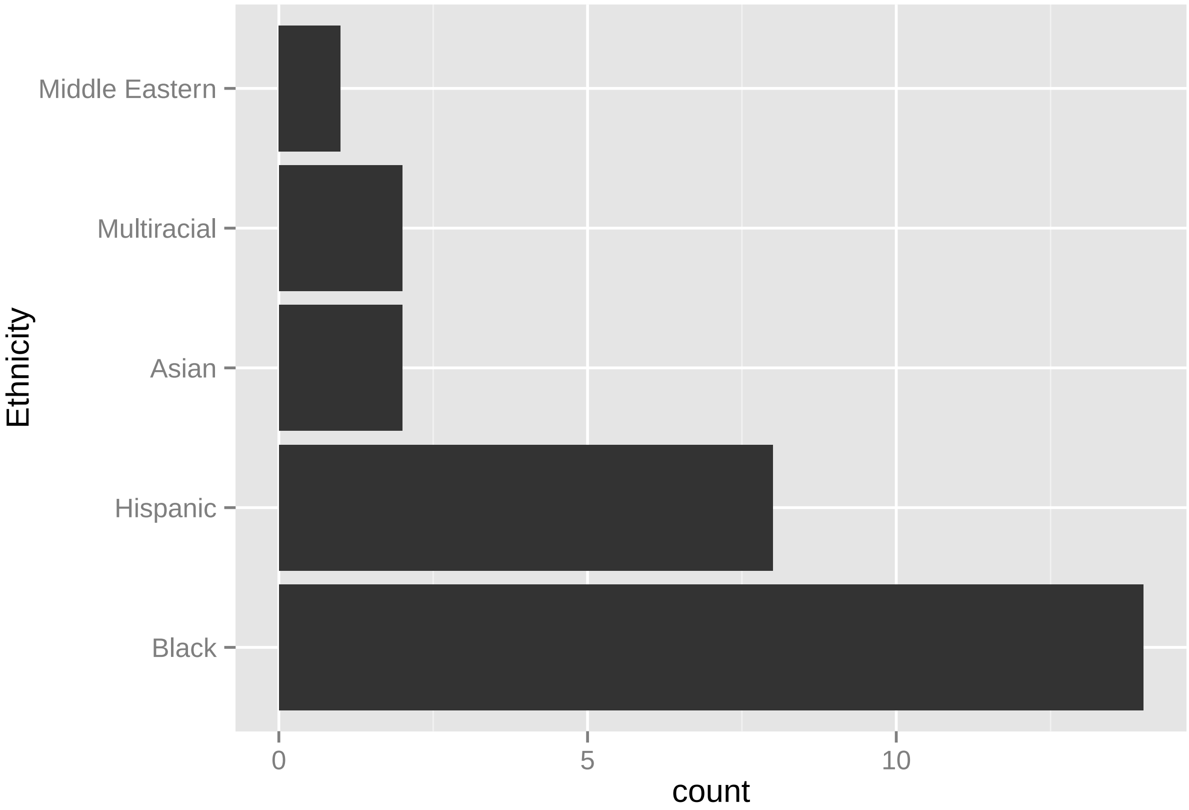
```
acting$Ethnicity <- factor(acting$Ethnicity,
      levels=c("White", "Black",
               "Hispanic", "Asian",
               "Multiracial",
               "Middle Eastern"))
qplot(Ethnicity, data=subset(acting,
                Ethnicity != "White"))
```



Flipping axes, adding color

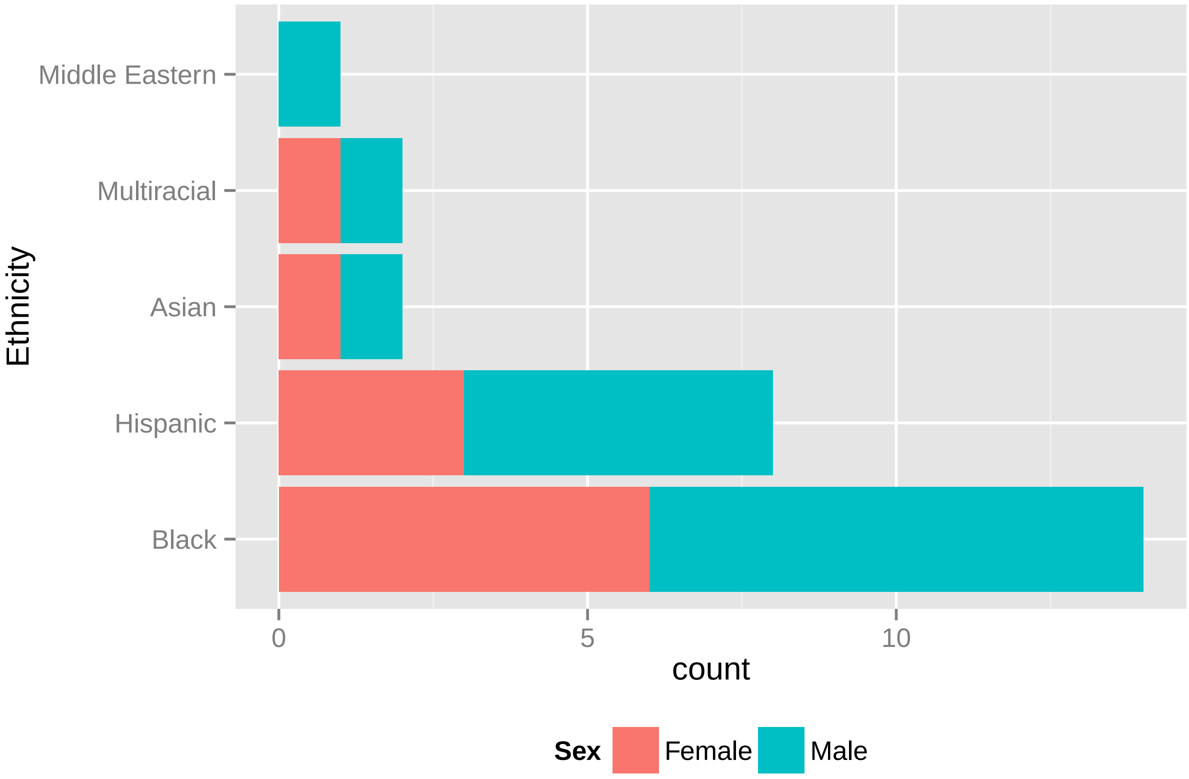
Flip axes to get better read on long labels

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



Color bars by gender

```
qplot(Ethnicity, data=subset(acting, Ethnicity != "White"), fill=Sex) +
    coord_flip() + theme(legend.position="bottom")
```

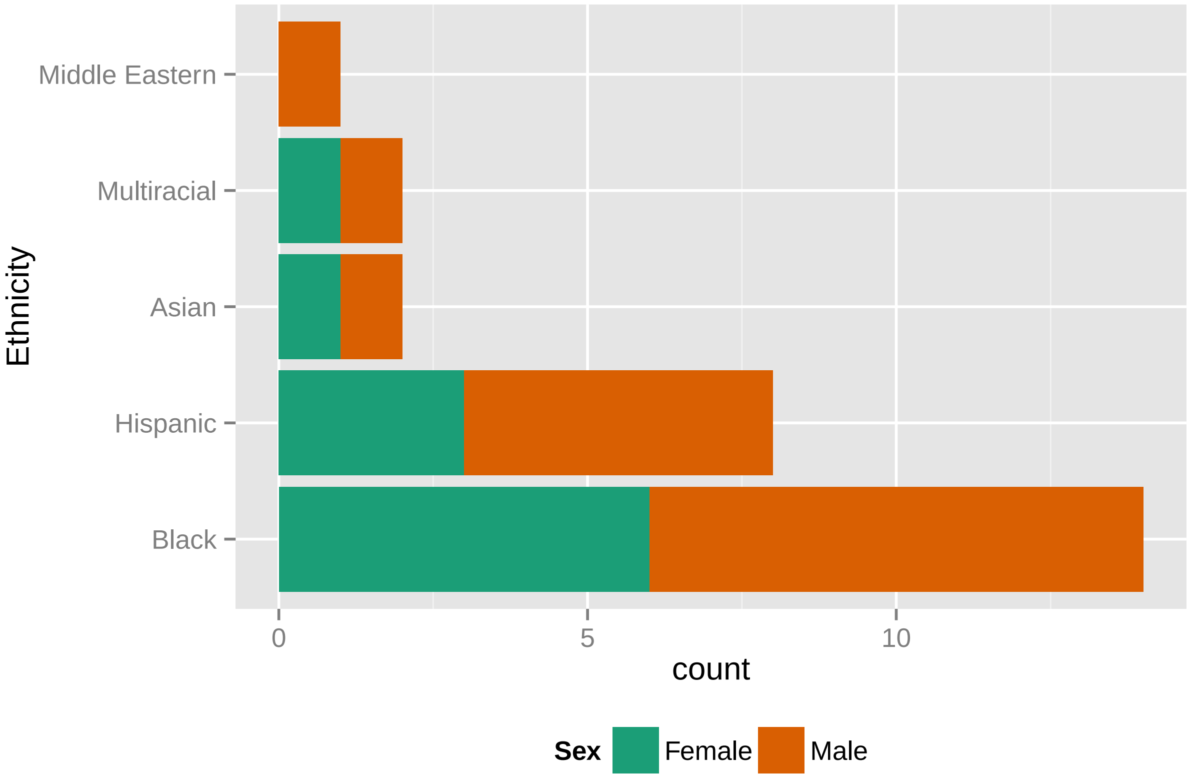


Color choice

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

```
library(RColorBrewer)
display.brewer.all()
```

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



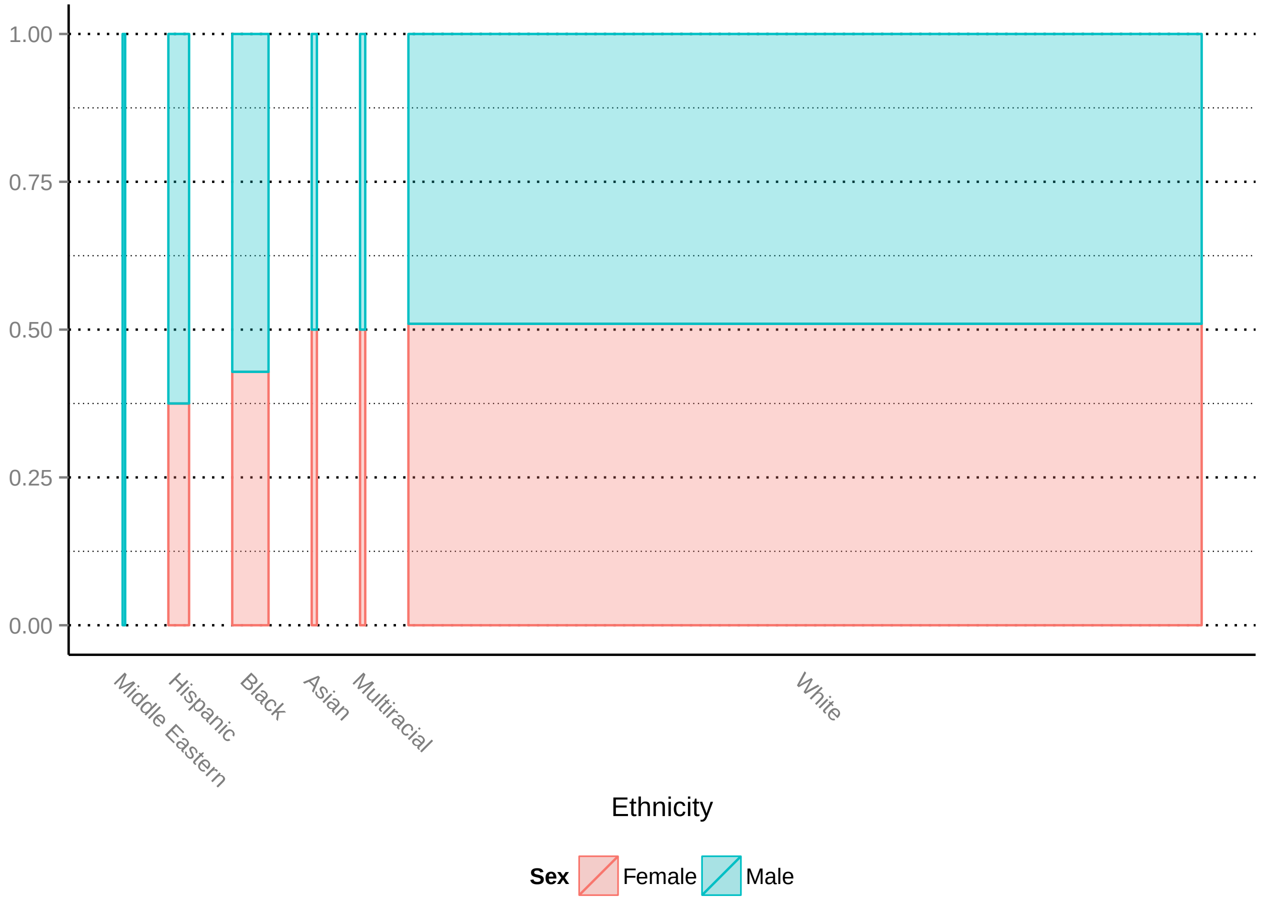
Mosaic

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

```
library(plotluck)
```

```
##
## Attaching package: 'plotluck'
##
## The following object is masked from 'package:ggplot2':
##
##     ggplotGrob
```

```
plotluck(acting, Ethnicity, Sex)
```



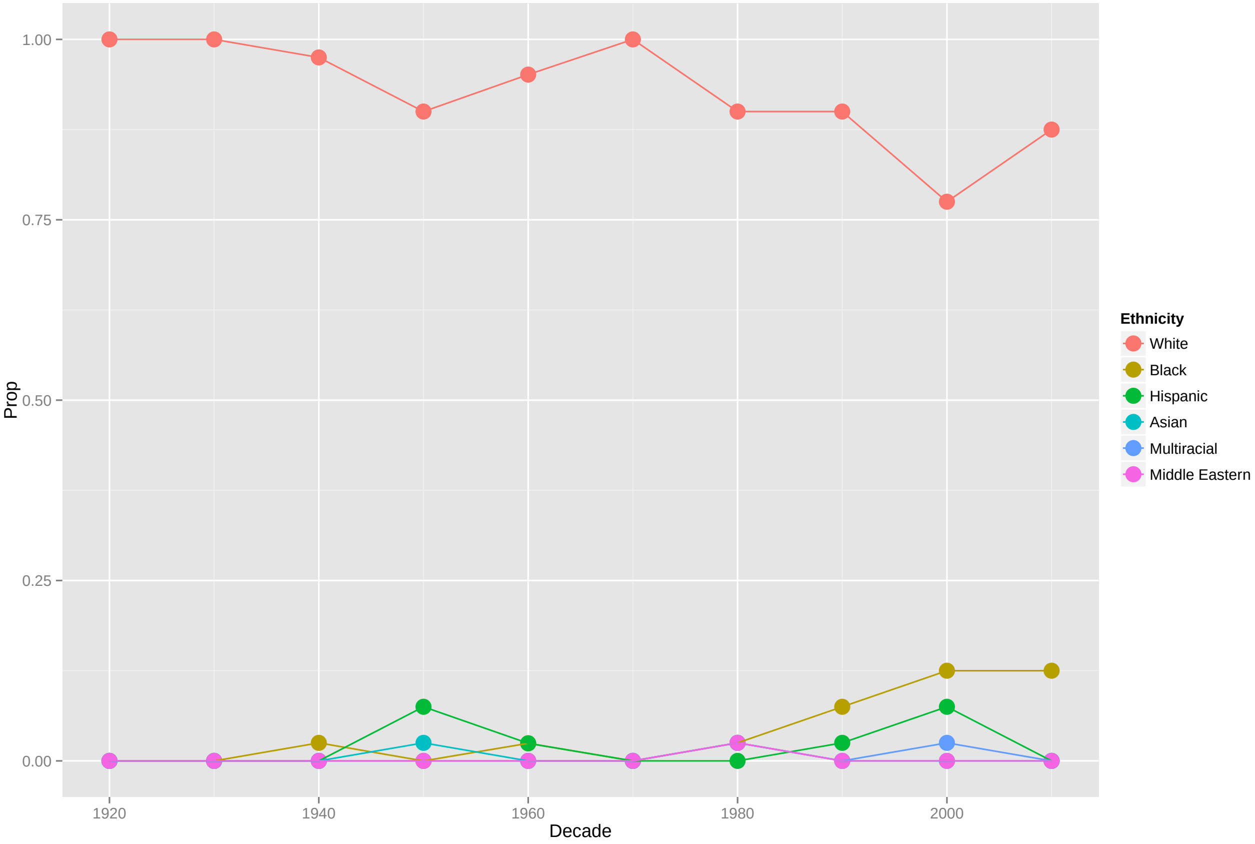
Temporal Trend

These numbers are aggregated across almost a hundred years. 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.

```
acting$Decade <- floor(acting$Year/10)*10
ptable <- data.frame(prop.table(table(acting$Decade, acting$Ethnicity), 1))
colnames(ptable) <- c("Decade", "Ethnicity", "Prop")
ptable$Decade <- as.numeric(as.character(ptable$Decade))
```

Temporal Trend

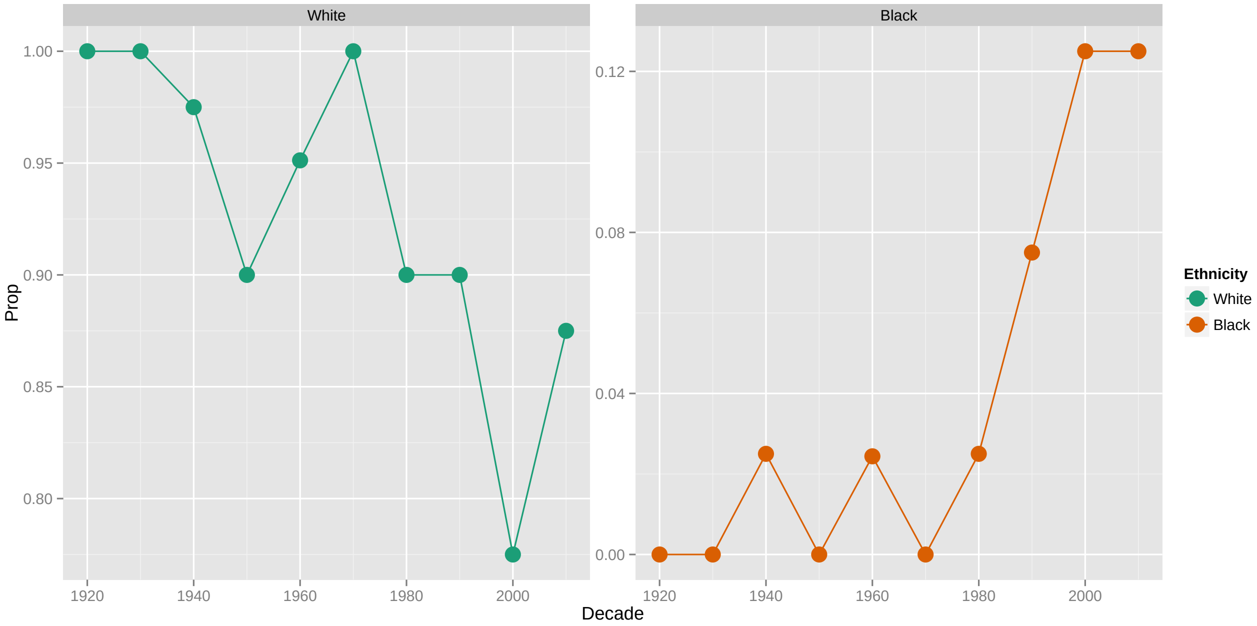
```
qplot(Decade, Prop, data=ptable, colour=Ethnicity, group=Ethnicity, size=I(5)) + geom_line()
```



Facetting

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

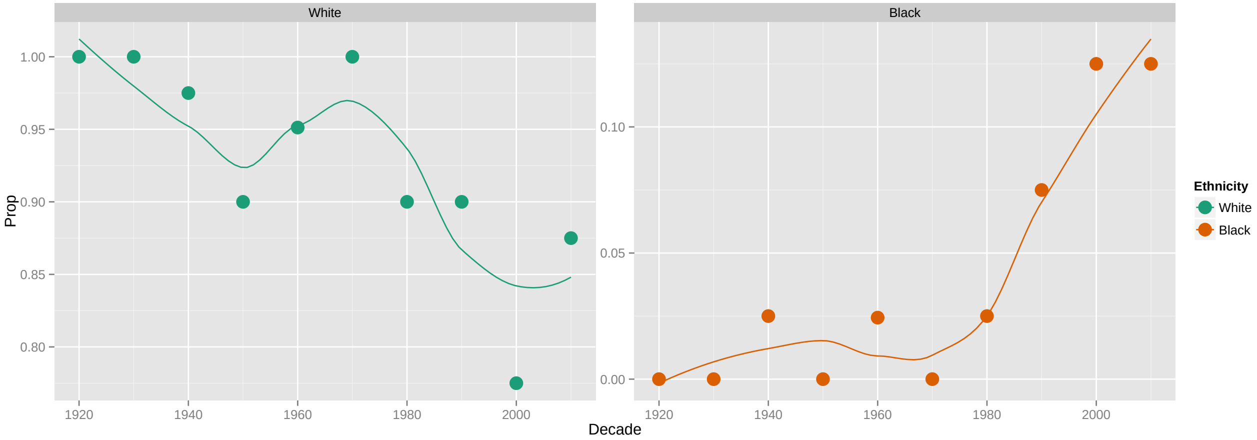
```
qplot(Decade, Prop, data=subset(ptable, Ethnicity=="White"|Ethnicity=="Black"),
      colour=Ethnicity, group=Ethnicity, size=I(5)) +
    geom_line() + scale_colour_brewer(type="qual", palette=2) +
    facet_wrap(~Ethnicity, scales="free_y")
```



Examine trend

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

```
qplot(Decade, Prop, data=subset(ptable, Ethnicity=="White"|Ethnicity=="Black"),
      colour=Ethnicity, size=I(5)) + geom_smooth(se=F) +
  scale_colour_brewer(type="qual", palette=2) +
  facet_wrap(~Ethnicity, scales="free_y")
```



Layering

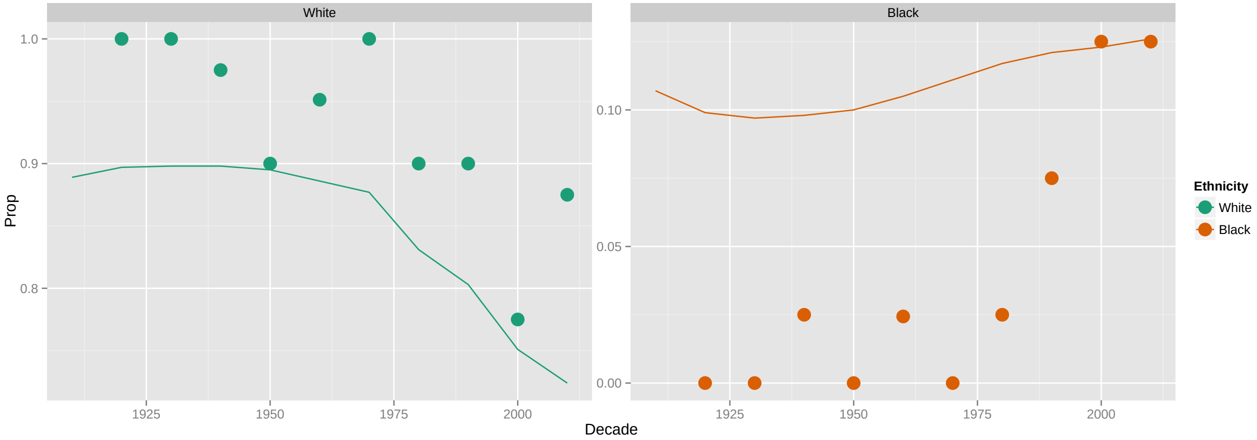
Pull demographics data from http://en.wikipedia.org/wiki/Historical_racial_and_ethnic_demographics_of_the_United_States.

Overlay these values as a line plot on the Oscars proportions.

```
library(reshape2)
pop.prop <- read.csv("data/wiki-race.csv")
pop.prop.m <- melt(pop.prop[,1:3], id="Year")
colnames(pop.prop.m) <- c("Decade", "Ethnicity", "Prop")
pop.prop.m$Prop <- pop.prop.m$Prop/100
```

Layering

```
qplot(Decade, Prop, data=subset(ptable, Ethnicity=="White"|Ethnicity=="Black"),
      colour=Ethnicity, group=Ethnicity, size=I(5)) +
  facet_wrap(~Ethnicity, scales="free_y") +
  geom_line(data=pop.prop.m, aes(x=Decade, y=Prop, colour=Ethnicity, group=Ethnicity)) +
  scale_colour_brewer(type="qual", palette=2)
```



Perceptual principles: proximity

Place elements for primary comparison close together.

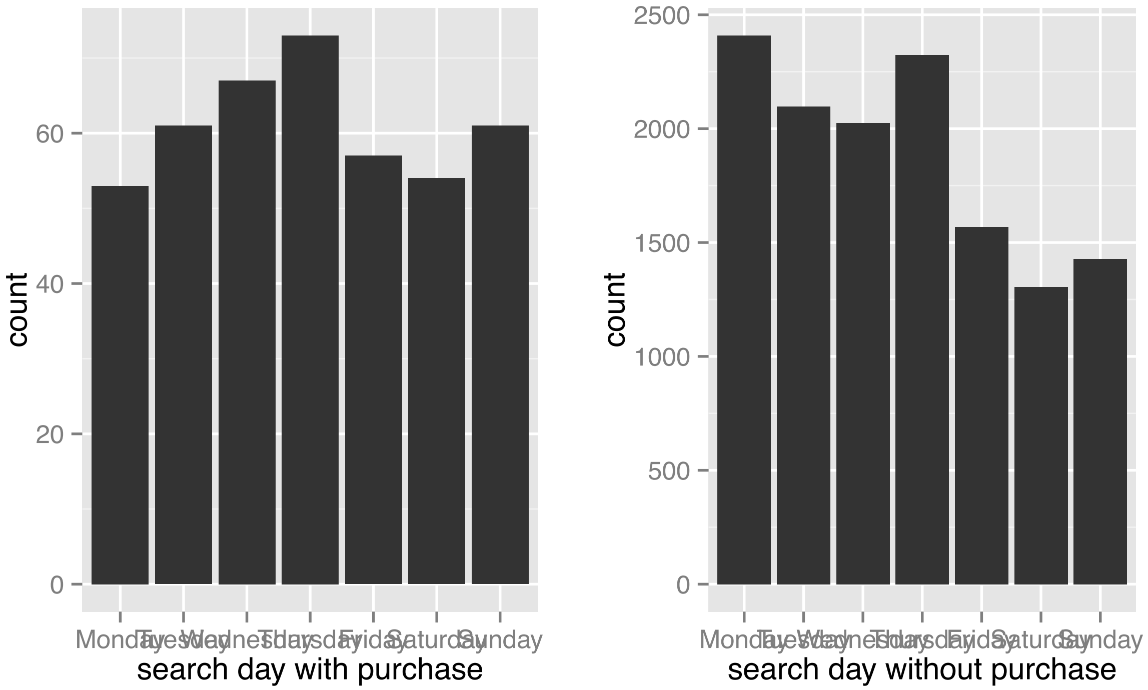
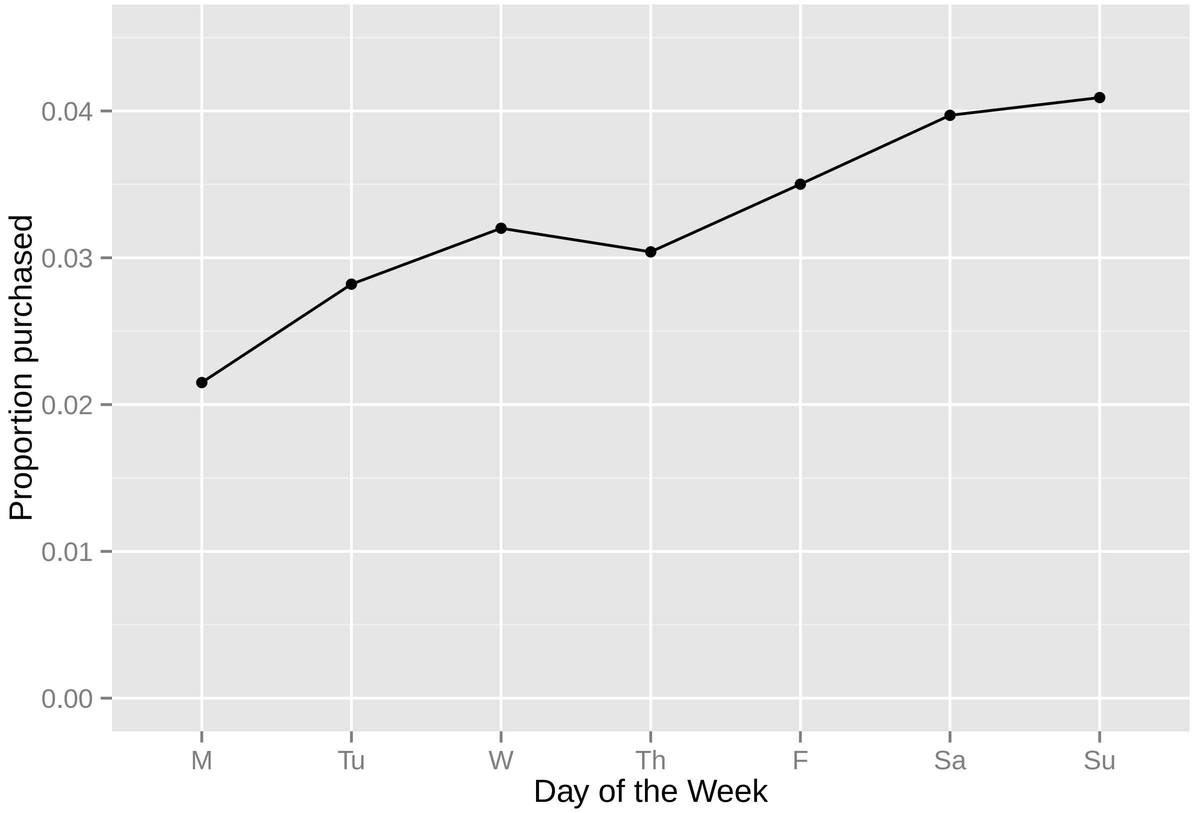


Figure 7: Histogram of day of week of first search by purchase or not. It is a little curious to see that for non-purchase group more people start to search on weekdays instead of weekends. Maybe some are for bussiness trips, or maybe people think more about going on vovations on weekdays.

***

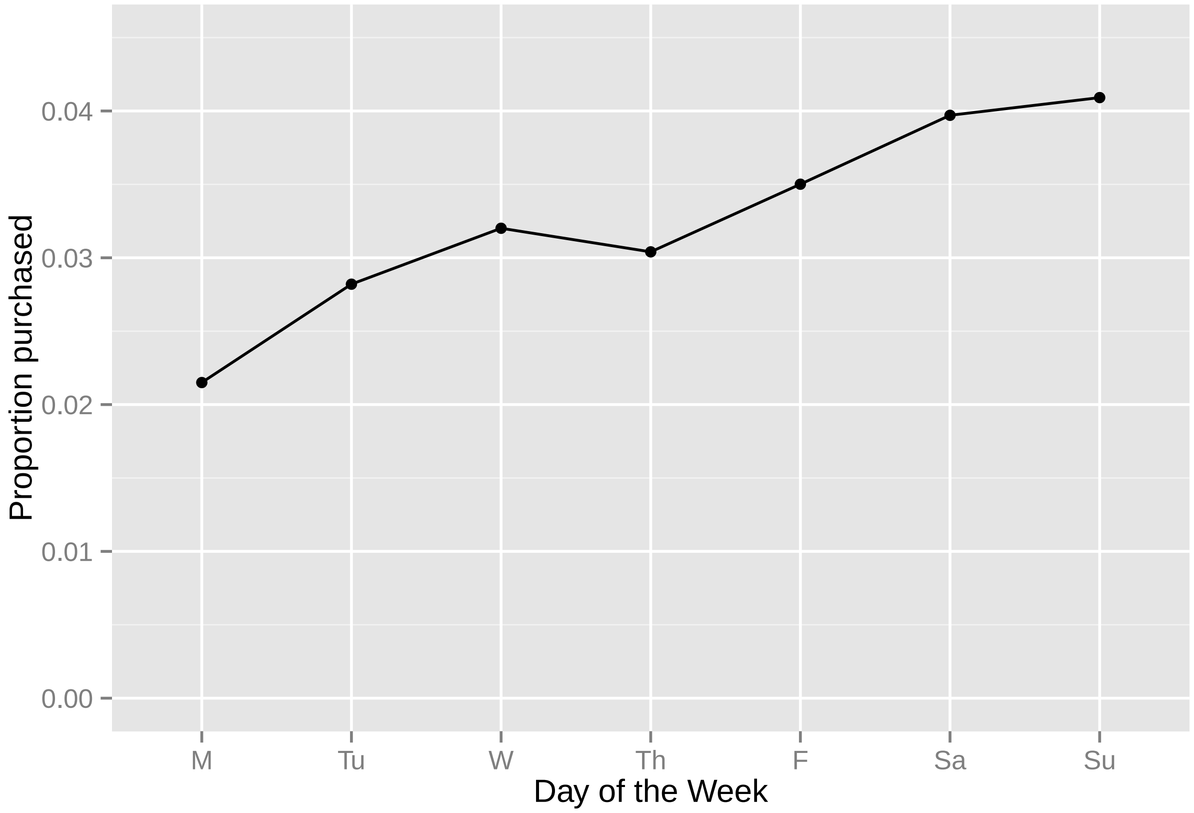
Having a hard time telling whether there is a week day/weekend day pattern? Plot proportions.

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



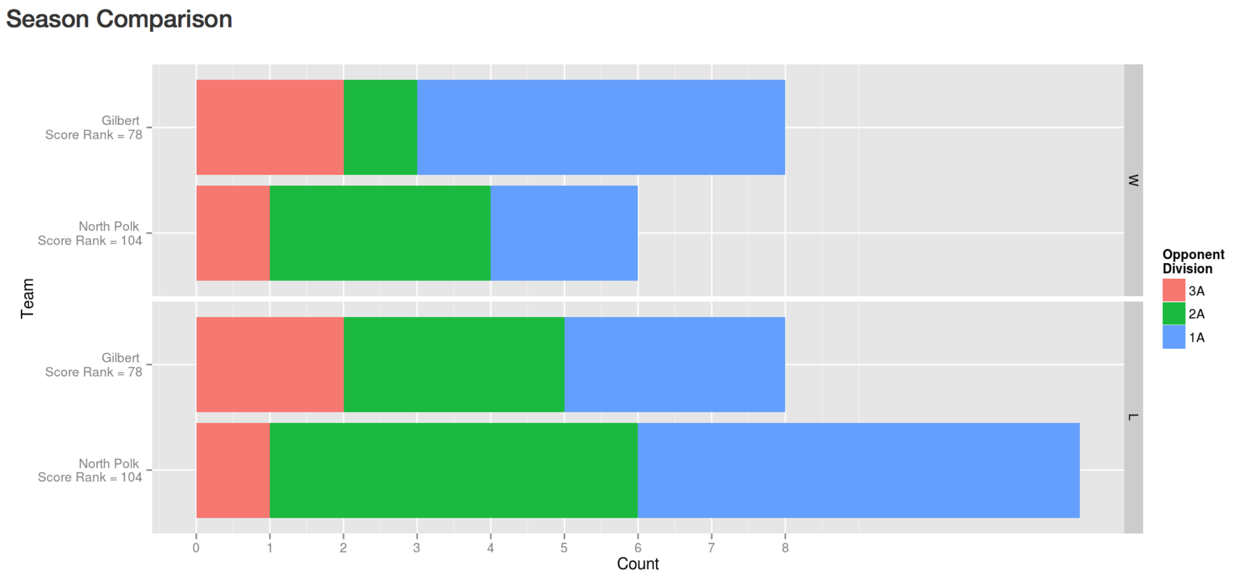
Perceptual principles: proximity

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

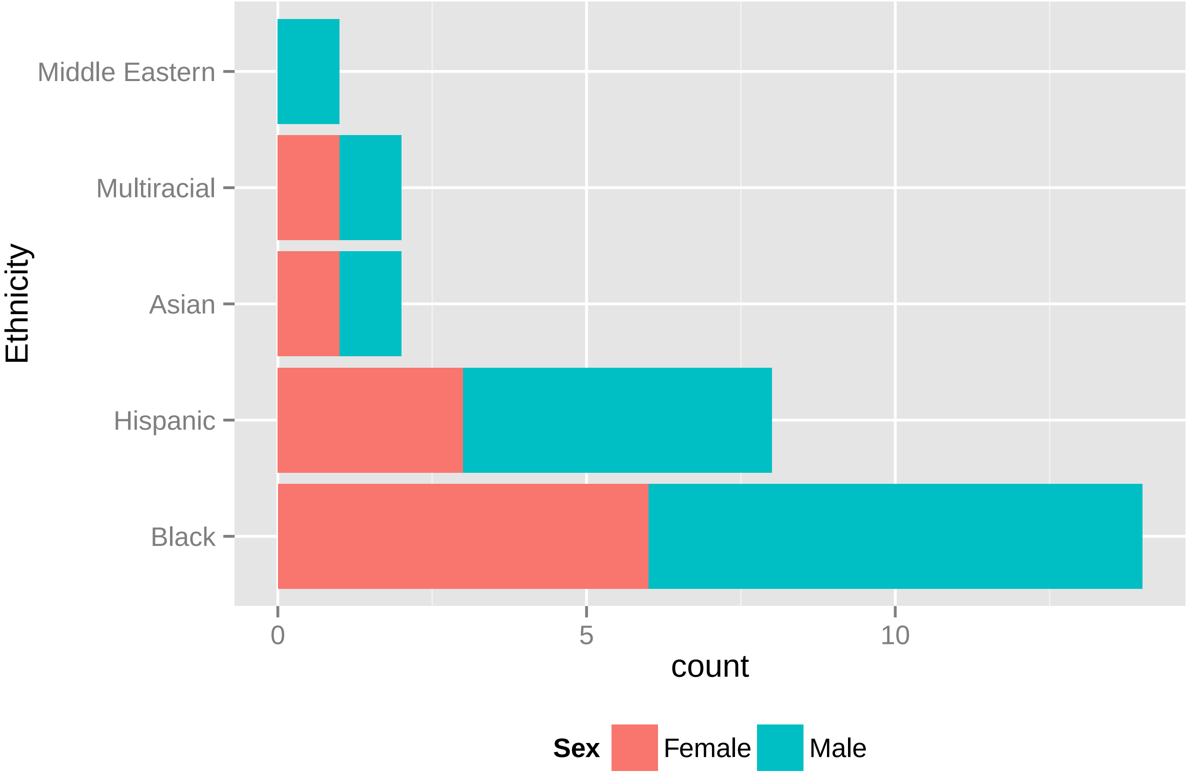
Applying color takes some care:

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:

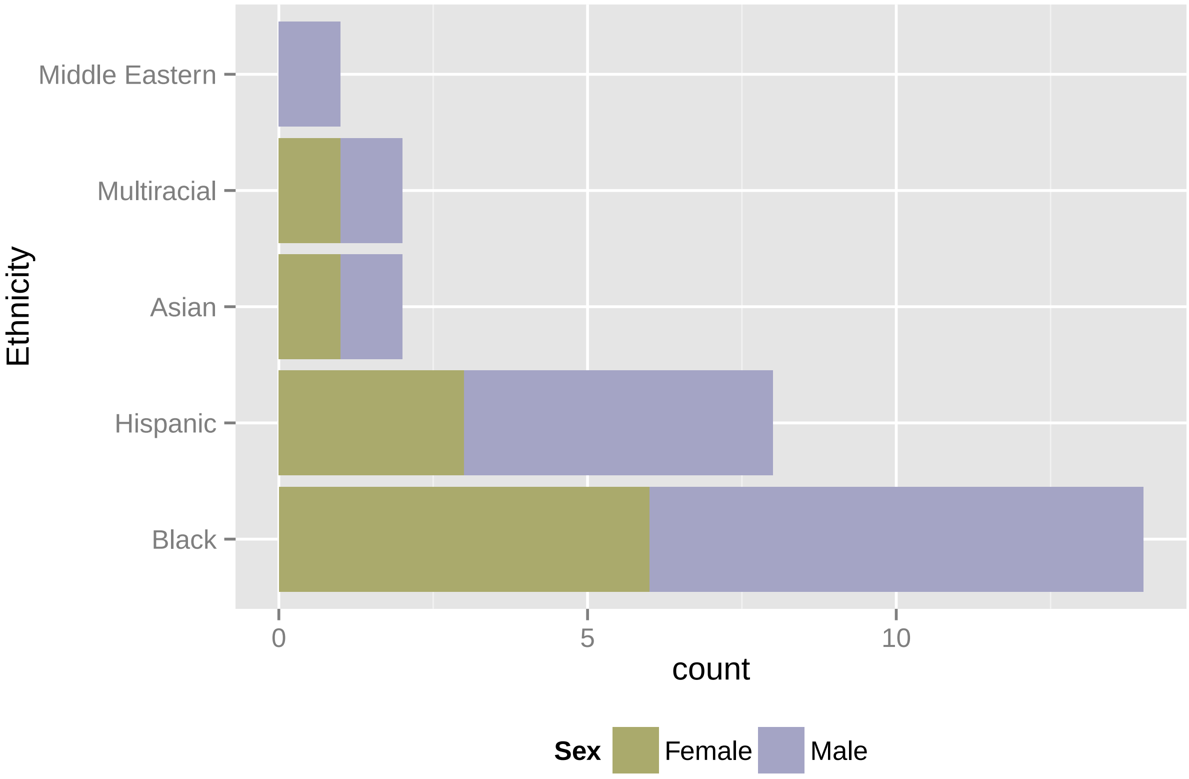
- Use QUALITATIVE scales to display categorical data, small number of levels,
- SEQUENTIAL scales are used to represent a gradient of quantitative information, e.g. 0-100
- DIVERGING scales are used to represent negative to positive quantitative information.

Color blindness

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



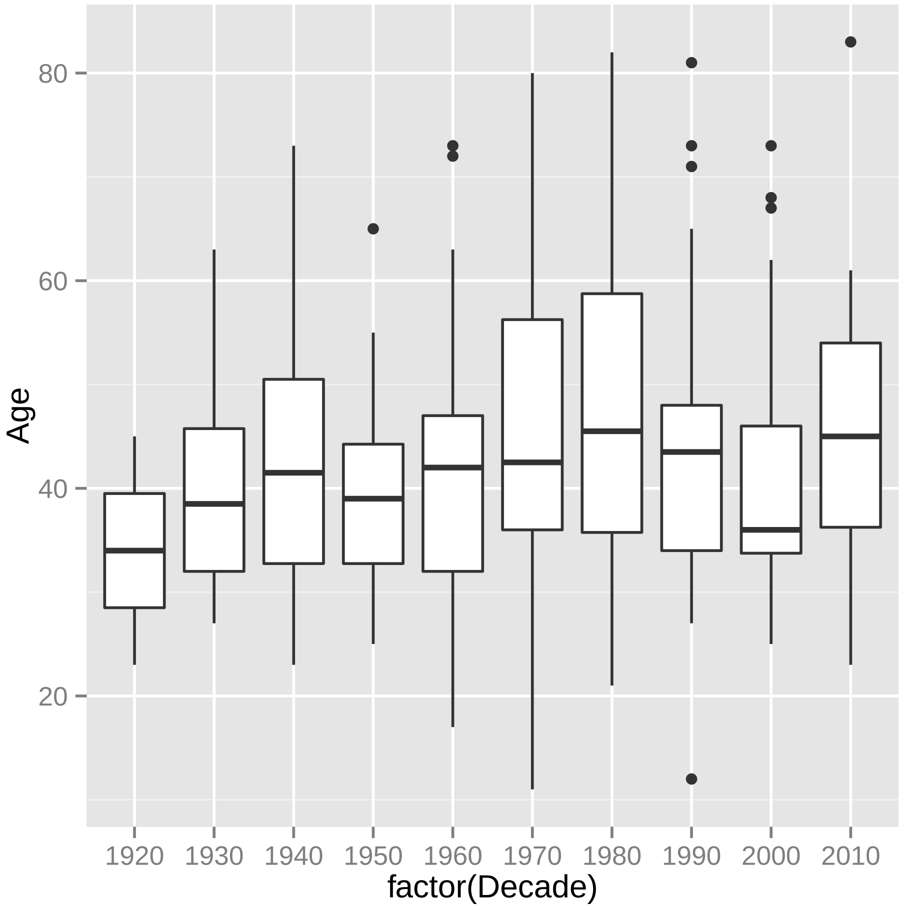
Using the package `dichromat`, to see what this looks like to a red-green color blind eye.



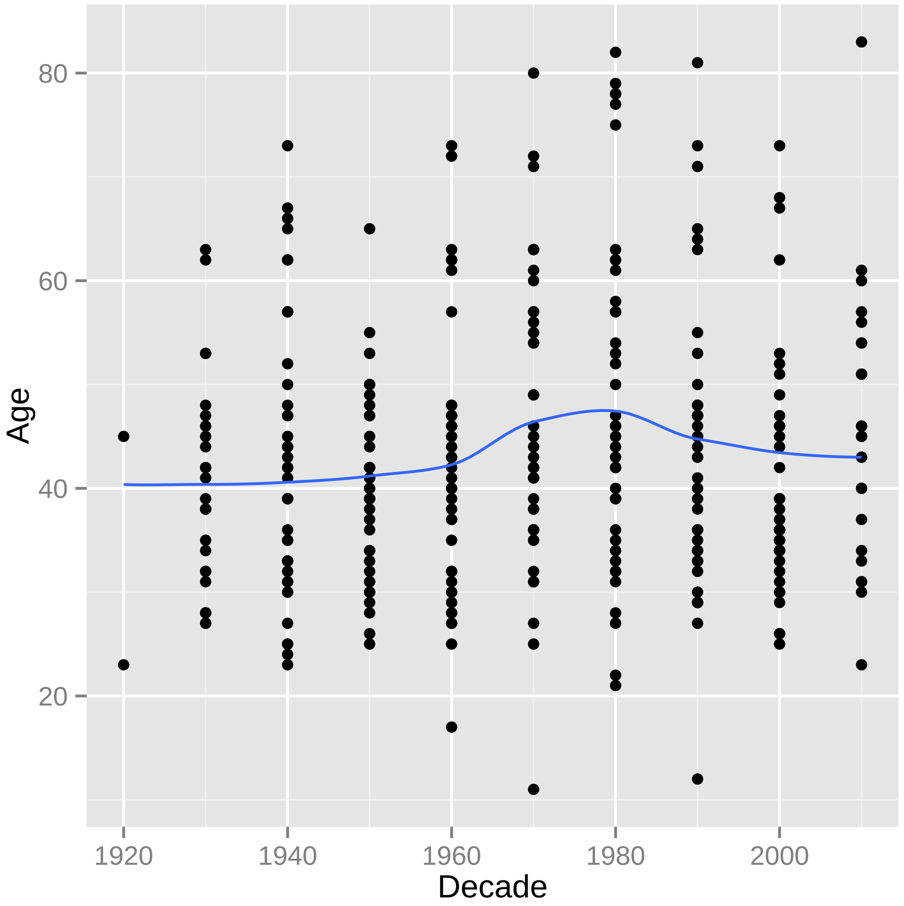
Exploring the Oscars

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

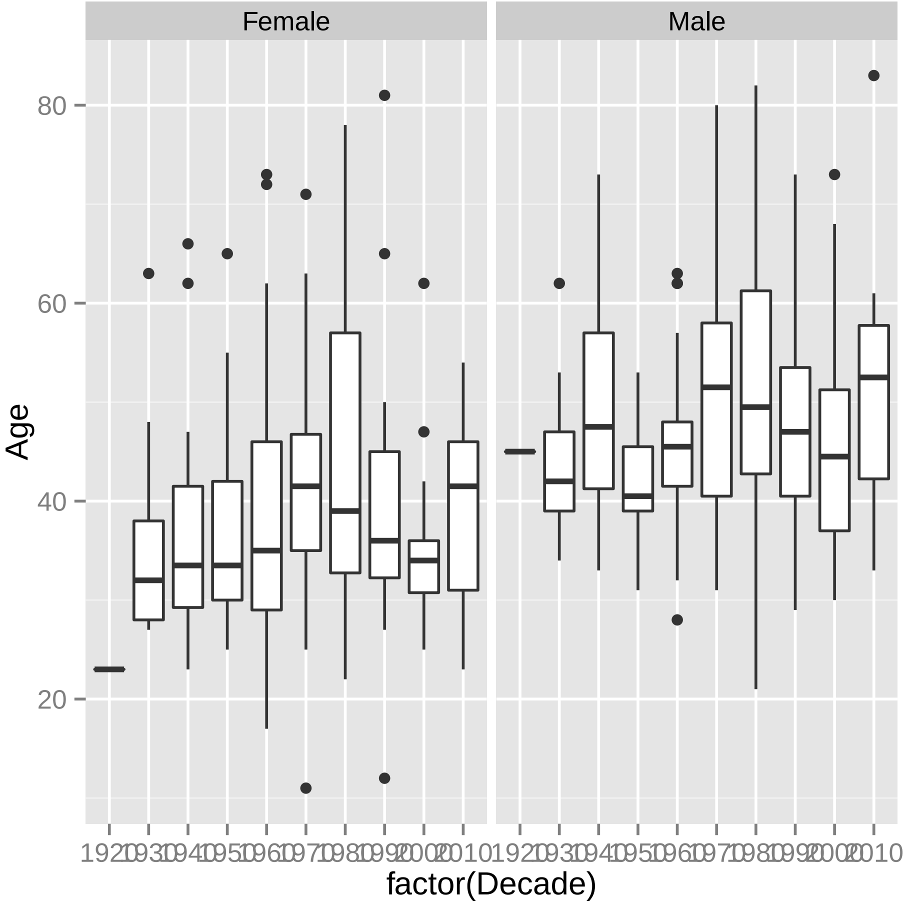
```
qplot(factor(Decade), Age, data=acting, geom="boxplot")
```



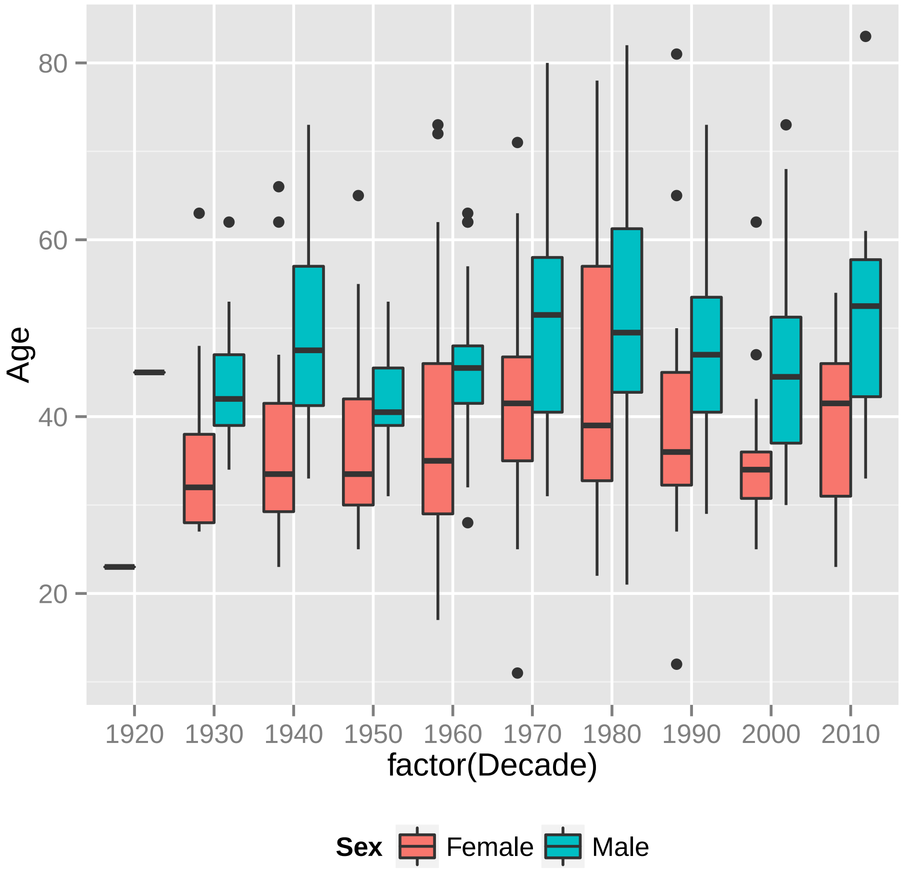
```
qplot(Decade, Age, data=acting, geom=c("point", "smooth"), se=F)
```



```
qplot(factor(Decade), Age, data=acting, geom="boxplot", facets=.~Sex)
```



```
qplot(factor(Decade), Age, data=acting, geom="boxplot", fill=Sex) + theme(legend.position="bottom")
```



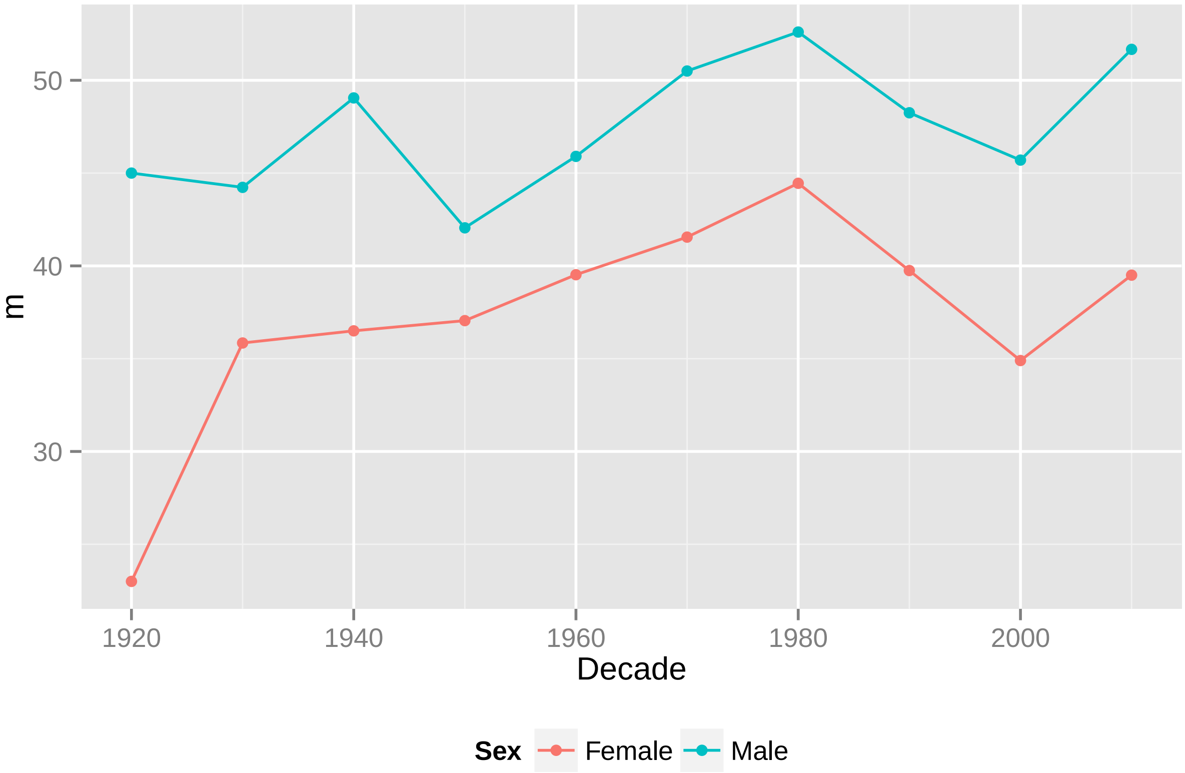
Showing the statistics

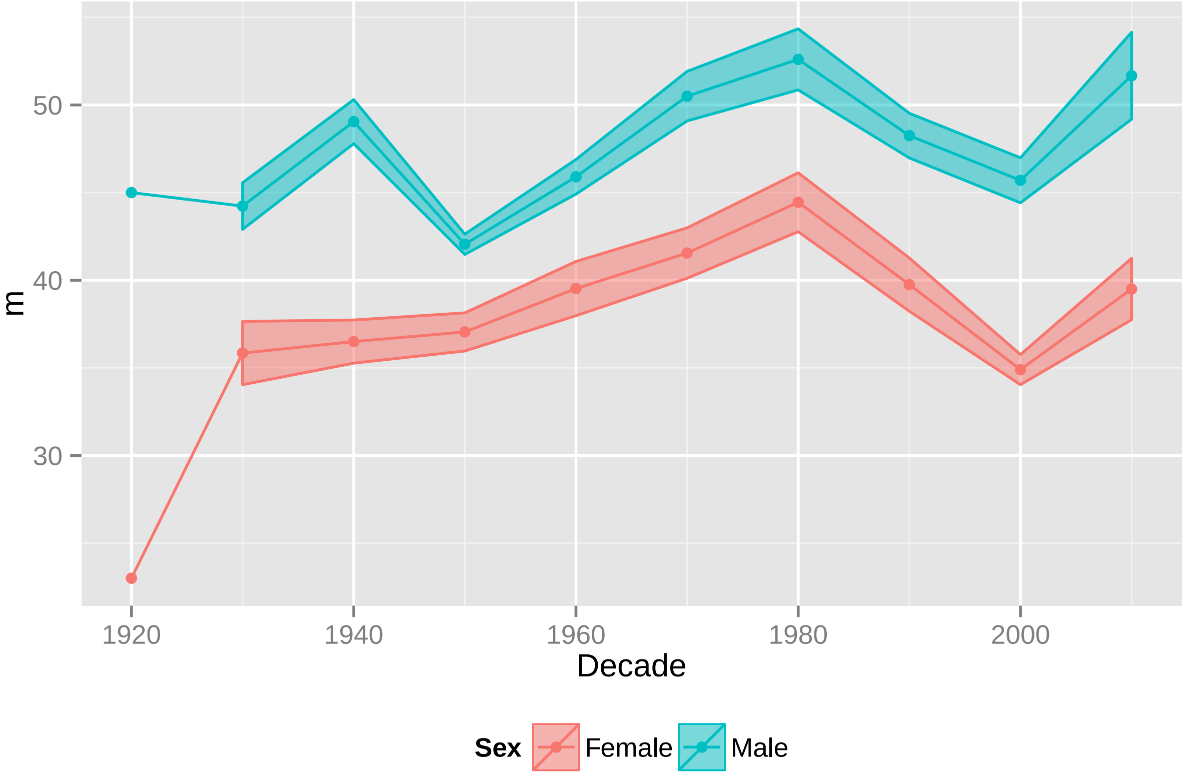
```
agesex <- summarise(group_by(acting, Decade, Sex), m=mean(Age, na.rm=T),
                    s=qt(0.975, length(Age)-1)*sd(Age, na.rm=T)/length(Age))
```

```
## Warning in qt(0.975, length(c(54L, 61L, 37L, 51L, 83L, 40L, 57L, 56L,
## 43L, : NaNs produced

## Warning in qt(0.975, length(c(54L, 61L, 37L, 51L, 83L, 40L, 57L, 56L,
## 43L, : NaNs produced
```

```
qplot(Decade, m, data=agesex, group=Sex, colour=Sex, geom=c("point", "line")) +
  theme(legend.position="bottom")
qplot(Decade, m, data=agesex, group=Sex, colour=Sex, geom=c("point", "line")) +
  geom_ribbon(mapping=aes(ymin=m-s, ymax=m+s, fill=Sex), alpha=0.5) +
  theme(legend.position="bottom")
```





YOUR TURN

Make plots to answer these questions:

- How do the numbers of winners vary by declared sexual orientation?
- Have the proportions in each of these categories changed over time?
- Are these different by gender?
- Is there are predominant religion?

Resources

- ‘R Graphics Cookbook’ Winston Chang http://www.cookbook-r.com/Graphs/
- ggplot2 http://ggplot2.org/
- https://github.com/garrettgman
- http://yihui.name/knitr/