Code Sample: Data Visualization with R

Sara Gong

This document, which contains a demonstration of my data visualization skills, was created by the R Markdown file `run.Rmd` (included in my submission). The provided dataset has been extracted from the Survey of Consumer Finances (1989-2016). In the following, I illustrate my work through three sample questions.

Preliminary Steps

I begin by initializing my environment and loading the data (located in my working directory).

```
# reset environment
rm(list=ls())

# turn off scientific notation
options(scipen=999)

# set option to display code in output
knitr::opts_chunk$set(echo = TRUE)

# load packages
library(dplyr)
library(magrittr)
library(knitr)
library(ggplot2)
library(tidyr)

# load data
data <- as_tibble(read.csv("RA_21_22.csv"))
```

In order to get a basic feel for this dataset, I summarize the data from the most recent year, 2016.

```
# summarize 2016 data
data %>%
  filter(., year == 2016) %>%
  select(., -weight, -year) %>%
  lapply(., function(x) kable(t(summary(x)), align='l', digits=0))
```

$age

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18	40	54	53	64	95

$sex

female	male
1457	4791

$education

college degree	no college	some college
2695	2015	1538

$race

black	Hispanic	other	white
834	612	321	4481

$asset_total

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	39432	300800	12001647	1366825	1723910000

$asset_housing

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	140000	622353	425000	67230000

$debt_total

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	28000	250250	161610	148670000

$debt_housing

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	114455	109000	12520000

$income

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	32404	68961	798282	163034	287690006

Next, I create the variable `wealth` = `asset_total` - `debt_total`. As before, I also summarize `wealth` in the year 2016.

```
# create "wealth" variable
data$wealth <- data$asset_total - data$debt_total

# summarize wealth in 2016
subset(data$wealth, data$year == 2016) %>%
  summary(.) %>%
  t(.) %>%
  kable(., align='l', digits=0)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2042900	16330	189936	11751397	1153720	1703800000

Now, I turn to the questions.

Question 1

Please summarize key trends in median wealth over the last 30 years by race and education using plots and in writing.

First, I construct group-level panels containing the data that I would like to plot. I also display the first 10 observations from each of these new dataframes.

```
# create data for median wealth by race
mw_by_race <- data %>%
  group_by(., year, race) %>%
  summarise(., median_wealth = median(wealth), n = n())

# create data for median wealth by education
mw_by_educ <- data %>%
  group_by(., year, education) %>%
  summarise(., median_wealth = median(wealth), n = n())

# print first observations from each dataset
kable(head(mw_by_race, 10))
```

year	race	median_wealth	n
1989	black	13965.97	308
1989	Hispanic	20524.46	162
1989	other	169588.02	115
1989	white	313604.43	2558
1992	black	24015.62	357

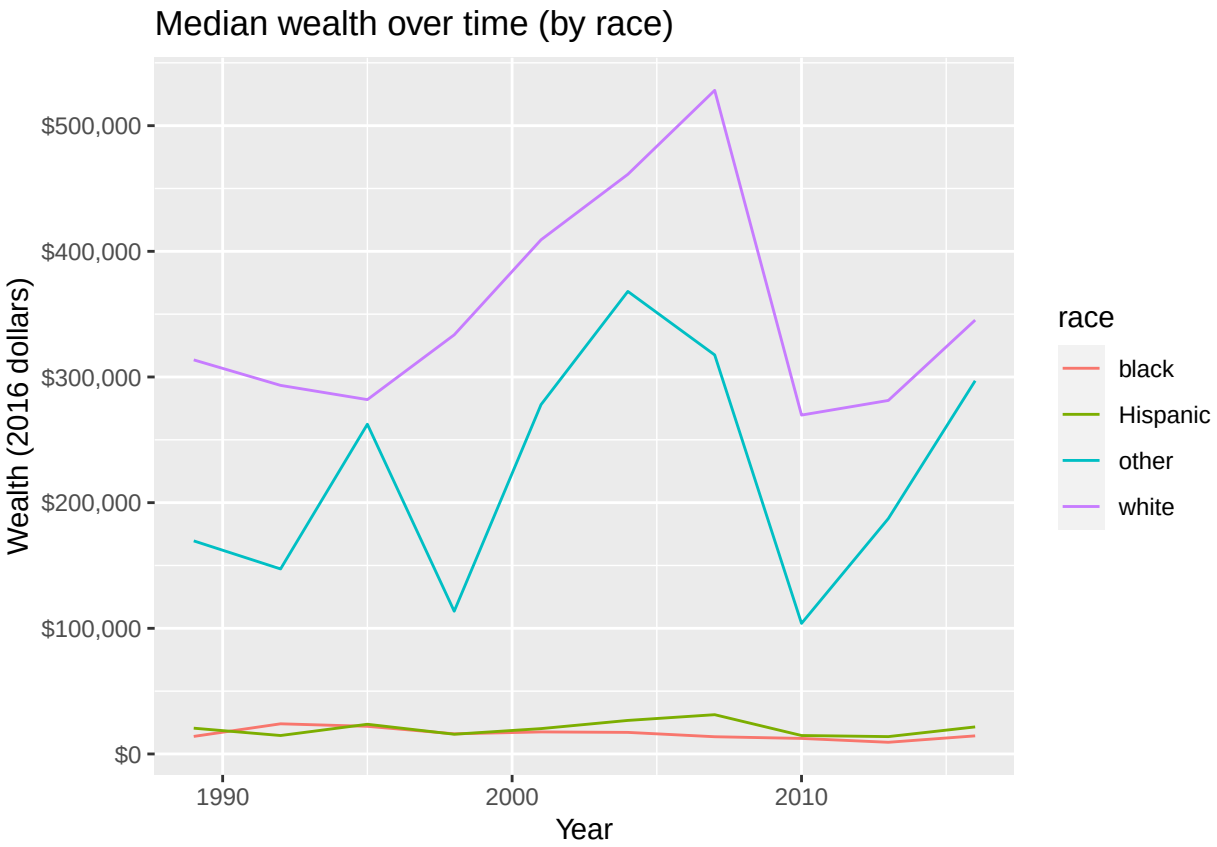
year	race	median_wealth	n
1992	Hispanic	14691.12	217
1992	other	147238.23	182
1992	white	293319.28	3150
1995	black	22002.36	380
1995	Hispanic	23646.67	178

```
kable(head(mw_by_educ, 10))
```

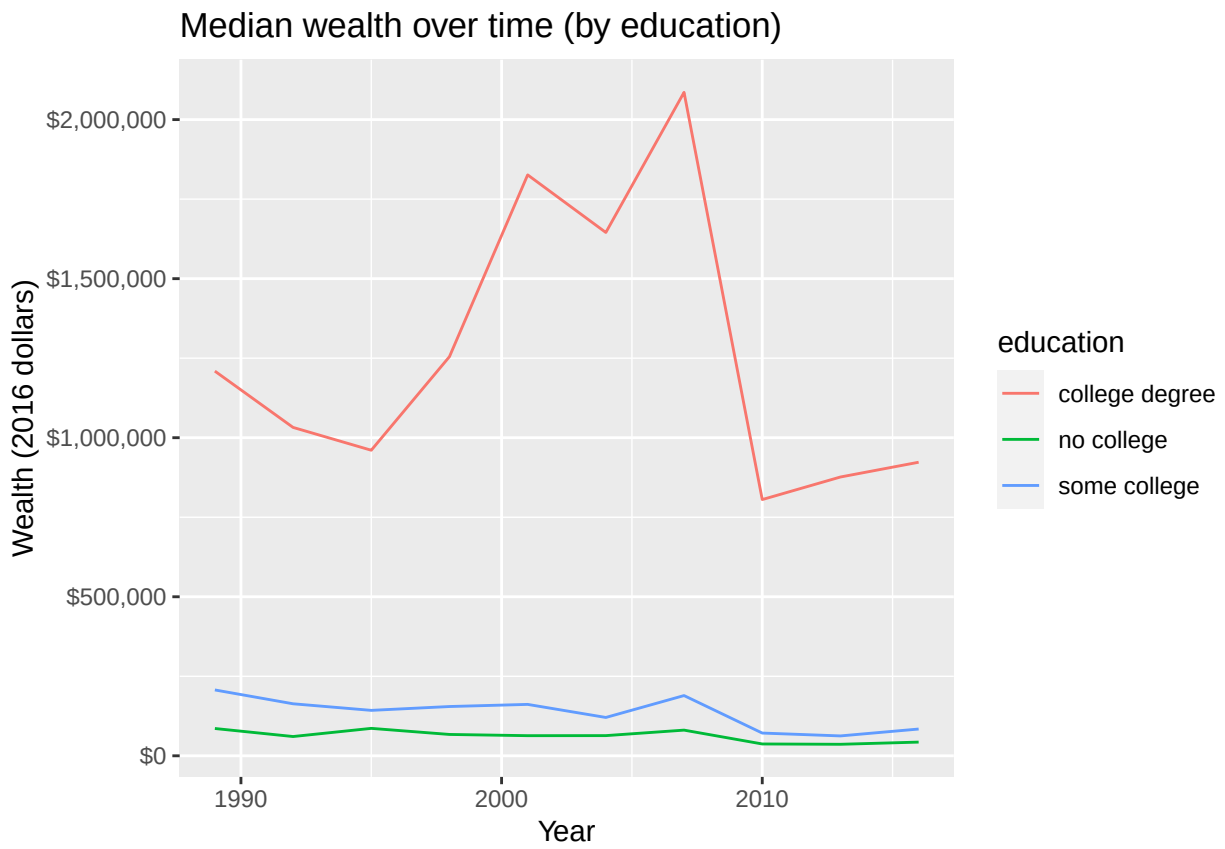
year	education	median_wealth	n
1989	college degree	1209310.56	1124
1989	no college	85829.56	1453
1989	some college	207110.47	566
1992	college degree	1032403.41	1589
1992	no college	60676.34	1532
1992	some college	163547.73	785
1995	college degree	960587.11	1627
1995	no college	86310.33	1690
1995	some college	142843.09	982
1998	college degree	1254802.90	1694

From the following graphs, it is clear that households that are Black or Hispanic have typically held less wealth than those that are not, and that respondents who do not have college degrees are less wealthy than those who do. These trends have persisted strongly over time, despite the large nominal loss in wealth for white and educated households during the 2007-2008 financial crisis.

```
# graph median wealth by race
mw_by_race %>%
  ggplot(.,aes(x = year, y = median_wealth, group = race, color = race)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Median wealth over time (by race)") +
  scale_y_continuous(name = "Wealth (2016 dollars)",
                     breaks = c(0, 100000,200000,300000,400000,500000,600000),
                     labels = c('$0', '$100,000','$200,000','$300,000',
                                '$400,000','$500,000','$600,000'))
```

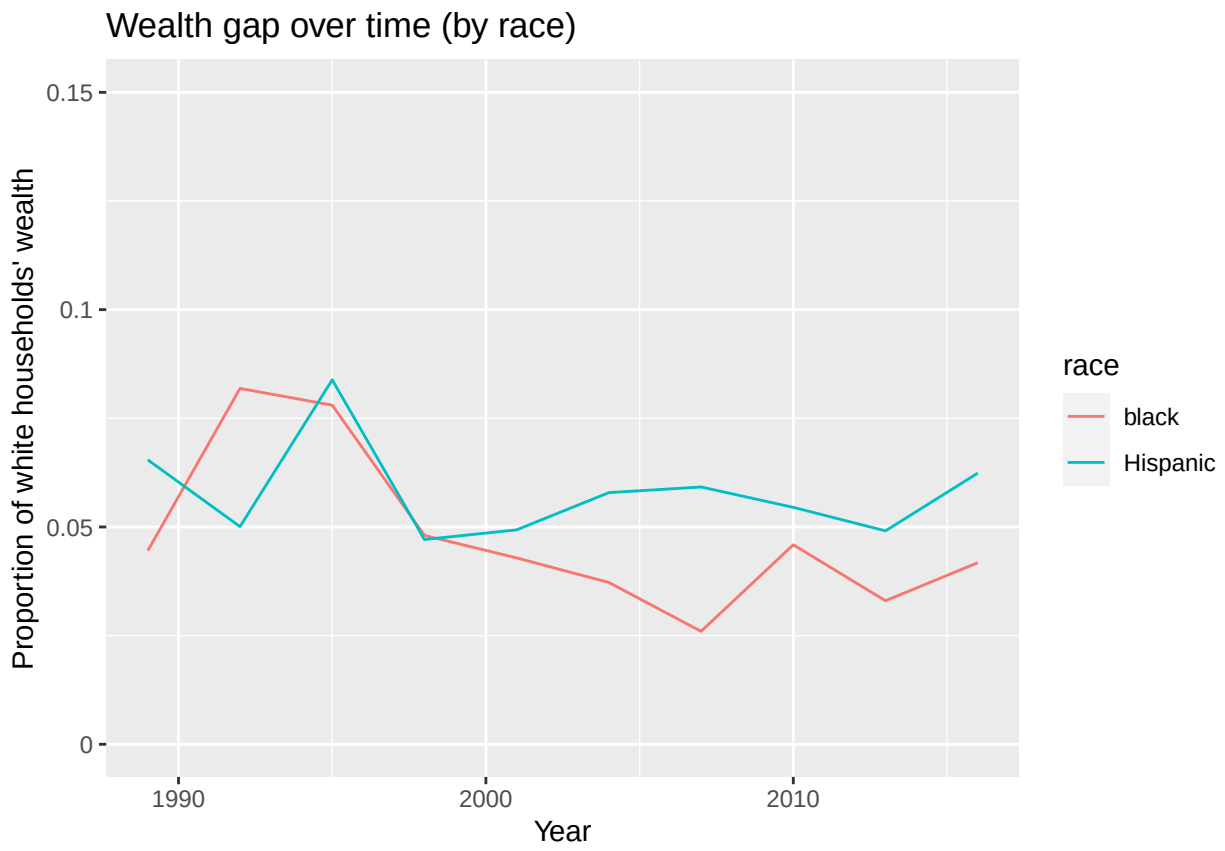



```
# graph median wealth by education
mw_by_educ %>%
  ggplot(.,aes(x = year, y = median_wealth, group = education, color = education)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Median wealth over time (by education)") +
  scale_y_continuous(name = "Wealth (2016 dollars)",
                     breaks = c(0, 500000,1000000,1500000,2000000),
                     labels = c('$0', '$500,000','$1,000,000','$1,500,000','$2,000,000'))
```

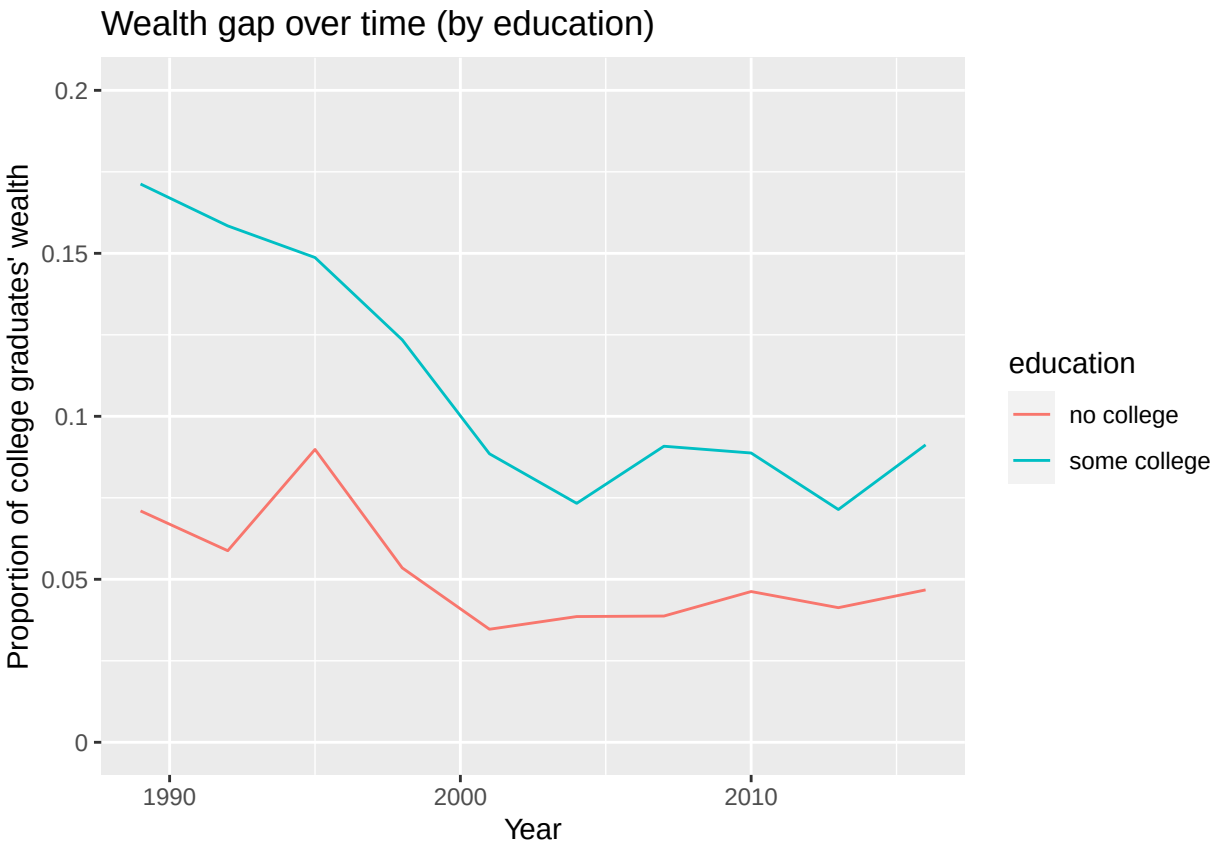


To further illustrate the wealth gap over time, I examine the ratio of the median wealth among groups. While both Black and Hispanic households have typically held less than a tenth of the wealth of white households, inequality has worsened since the mid-1990s, especially for Black families. Inequality has also worsened with respect to education level: from 2001 to at least 2016, the median college graduate has held twenty times the wealth of the median person without any college education.

```
# create variable to represent the wealth gap by race as a ratio
mw_by_race$median_ww <- mw_by_race$median_wealth %>%
  subset(., mw_by_race$race == "white") %>%
  rep(., each = 4)

mw_by_race$wealth_gap <- mw_by_race$median_wealth/mw_by_race$median_ww

# graph wealth gap by race
mw_by_race %>%
  filter(., race == "black" | race == "Hispanic") %>%
  ggplot(.,aes(x = year, y = wealth_gap, group = race, color = race)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Wealth gap over time (by race)") +
  scale_y_continuous(name = "Proportion of white households' wealth",
                     breaks = seq(from = 0, to = 0.15, by = 0.05),
                     labels = as.character(seq(from = 0, to = 0.15, by = 0.05)),
                     limits = c(0, 0.15))
```




```r
# create variable to represent the wealth gap by education as a ratio
mw_by_educ$median_cdw <- mw_by_educ$median_wealth %>%
  subset(., mw_by_educ$education == "college degree") %>%
  rep(., each = 3)

mw_by_educ$wealth_gap <- mw_by_educ$median_wealth/mw_by_educ$median_cdw

# graph wealth gap by education
mw_by_educ %>%
  filter(., education != "college degree") %>%
  ggplot(.,aes(x = year, y = wealth_gap, group = education, color = education)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Wealth gap over time (by education)") +
  scale_y_continuous(name = "Proportion of college graduates' wealth",
                     breaks = seq(from = 0, to = 0.2, by = 0.05),
                     labels = as.character(seq(from = 0, to = 0.2, by = 0.05)),
                     limits = c(0, 0.2))
```



Question 2

Repeat your analysis for just housing wealth for black and white households.

Similarly to the previous work, I first construct a new variable, `wealth_housing`, as `asset_housing - debt_housing`, and summarize its values in 2016.

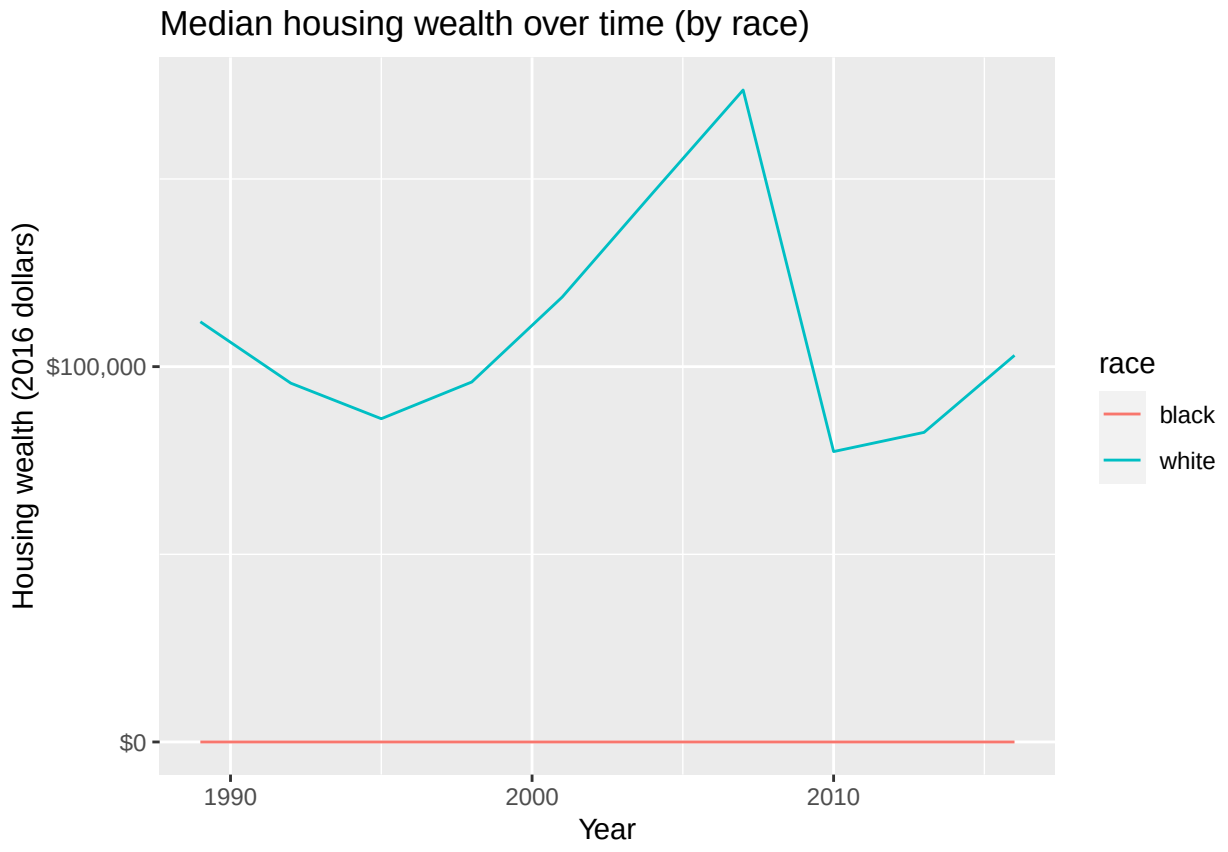
```
# create "wealth" variable
data$wealth_housing <- data$asset_housing - data$debt_housing

# summarize wealth in 2016
subset(data$wealth_housing, data$year == 2016) %>%
  summary(.) %>%
  t(.) %>%
  kable(., align='l', digits=0)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-706000	0	64000	507898	280000	67230000

We can easily see that the distribution of housing wealth between Black and white households is even more

unequal than that of total wealth: between 1989 and 2016, the median Black household had zero dollars in housing wealth.

```
# create data for median housing wealth by race
mhw_by_race <- data %>%
  filter(., race == "black" | race == "white") %>%
  group_by(., year, race) %>%
  summarise(., median_wealth_housing = median(wealth_housing), n = n())

# graph median wealth by race
mhw_by_race %>%
  ggplot(.,aes(x = year, y = median_wealth_housing, group = race, color = race)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Median housing wealth over time (by race)") +
  scale_y_continuous(name = "Housing wealth (2016 dollars)",
                     breaks = c(0, 100000,200000),
                     labels = c('$0', '$100,000','$200,000'))
```



Moreover, even after considering only households with nonzero housing wealth, the wealth gap remains prominent. Black households with nonzero housing wealth have typically held less than half of the housing wealth of white households. Note, however, that the previous recession caused a larger relative decline in white households' housing wealth, perhaps due the geographical segregation of majority-Black and majority-white neighborhoods.

```r
# create data for median housing wealth by race (among households with nonzero housing wealth)
nz_mhw_by_race <- data %>%
  filter(., race == "black" | race == "white") %>%
  filter(., wealth_housing > 0) %>%
  group_by(., year, race) %>%
  summarise(., median_wealth_housing = median(wealth_housing), n = n())

# graph median wealth by race (among households with nonzero housing wealth)
nz_mhw_by_race %>%
  ggplot(.,aes(x = year, y = median_wealth_housing, group = race, color = race)) +
  geom_line() +
  xlab("Year") +
  ggtitle("Median housing wealth over time, among households with nonzero housing wealth") +
  theme(plot.title = element_text(size = 10)) +
  scale_y_continuous(name = "Housing wealth (2016 dollars)",
                     breaks = c(0, 100000,200000, 300000),
                     labels = c('$0', '$100,000','$200,000', '$300,000'),
                     limits = c(0, 300000))
```

Question 3

Many households are not homeowners and so your analysis for the prior question includes many zeros for housing wealth. Let's dig deeper by focusing just on homeowners age 25 or older. Please summarize trends in for black and white households for both housing and non-

housing wealth. Which group had the largest loss in housing wealth, where 2007 is defined as the base period? Please answer this question both in dollar terms and in proportional terms.

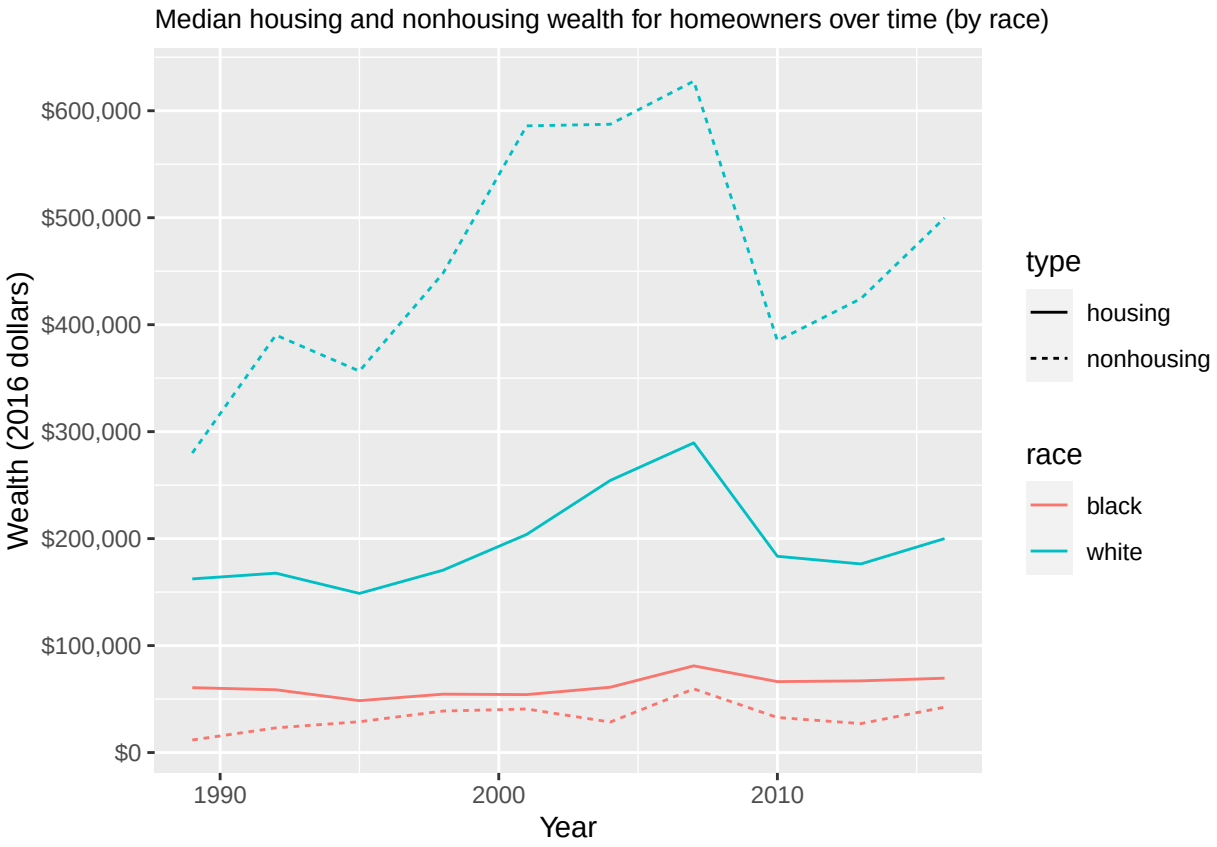
Once again, I begin by generating and summarizing the variable `wealth_nonhousing`.

```
# create "wealth" variable
data$wealth_nonhousing <- data$wealth - data$wealth_housing

# summarize wealth in 2016
subset(data$wealth_nonhousing, data$year == 2016) %>%
  summary(.) %>%
  t(.) %>%
  kable(., align='l', digits=0)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2614900	7200	91000	11243499	845450	1684810000

I choose the convenience of a plot in order to represent the four trends of interest; this allows us to see that the nonhousing wealth gap (among households with nonzero housing wealth) is historically even larger than the housing wealth gap. At the nonhousing wealth gap's peak in 2007, white households in this subpopulation owned over ten times more in nonhousing assets than Black households.

```
# create data for median total and housing wealth by race (for respondents with age >= 25)
mw2 <- data %>%
  filter(., race == "black" | race == "white") %>%
  filter(., age >= 25 & wealth_housing > 0) %>%
  group_by(., year, race) %>%
  summarise(., housing = median(wealth_housing),
            nonhousing = median(wealth_nonhousing)) %>%
  gather(., type, value, housing, nonhousing)

# graph median total and housing wealth by race
mw2 %>%
  ggplot(.,aes(x = year)) +
  geom_line(aes(y = value, group = interaction(race, type), color = race, linetype=type)) +
  xlab("Year") +
  ggtitle("Median housing and nonhousing wealth for homeowners over time (by race)") +
  theme(plot.title = element_text(size = 10)) +
  scale_y_continuous(name = "Wealth (2016 dollars)",
                     breaks = c(0, 100000,200000, 300000,
                                400000, 500000, 600000),
                     labels = c('$0', '$100,000','$200,000', '$300,000',
                                '$400,000', '$500,000', '$600,000'))
```



Finally, to examine the impact of the previous recession on each of these groups, I examine the decline in housing and nonhousing wealth from 2007 to 2010.

```
hw_white_2007 <- mw2 %>%
  filter(., year == 2007 & race == "white" & type == "housing")
hw_black_2007 <- mw2 %>%
  filter(., year == 2007 & race == "black" & type == "housing")
hw_white_2010 <- mw2 %>%
  filter(., year == 2010 & race == "white" & type == "housing")
hw_black_2010 <- mw2 %>%
  filter(., year == 2010 & race == "black" & type == "housing")

nhw_white_2007 <- mw2 %>%
  filter(., year == 2007 & race == "white" & type == "nonhousing")
nhw_black_2007 <- mw2 %>%
  filter(., year == 2007 & race == "black" & type == "nonhousing")
nhw_white_2010 <- mw2 %>%
  filter(., year == 2010 & race == "white" & type == "nonhousing")
nhw_black_2010 <- mw2 %>%
  filter(., year == 2010 & race == "black" & type == "nonhousing")
```

While white households experienced larger nominal losses in wealth, as well as a larger proportional loss in housing wealth, Black households suffered a larger proportional loss in nonhousing wealth. More precisely, between 2007 and 2010, white homeowners' median housing wealth dropped by 36.6 percent (from $289503.8

to $183484.6) and white homeowners' median nonhousing wealth dropped by 38.7 percent (from $627644.1 to $384986). Meanwhile, Black homeowners' housing and nonhousing wealth dropped by 18.2 percent (from $81061.1 to $66319.7) and by 44.9 percent (from $59522 to $32773), respectively.