Data Manipulation

Nora Quick

Learn

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dog_licenses <- readr::read_csv("https://github.com/merely-useful/novice-r/raw/master/data/nyc-dog-licen
```

```
## Rows: 118600 Columns: 15

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): animal_name, animal_gender, breed_name, borough, neighborhood_tabu...
## dbl  (7): row_number, zip_code, community_district, census_tract_2010, city_...
## date (3): animal_birth_month, license_issued_date, license_expired_date

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#View(dog_licenses)
#dog_licenses
```

```
#arrange(dog_licenses, animal_birth_month)
#arrange(dog_licenses, desc(animal_birth_month))
```

```
#arrange(dog_licenses, license_issued_date)
```

4.3.2 Exercise:

```
arrange(dog_licenses, animal_name)
```

```
## # A tibble: 118,542 x 15
##    row_number animal_name animal_gender animal_birth_month breed_name    borough
##         <dbl> <chr>       <chr>         <date>              <chr>         <chr>
##  1      24785 'RUSTY      M             2014-03-01          Cavalier Kin~ Queens
##  2      84389 'RUSTY      M             2014-03-01          Cavalier Kin~ Queens
##  3      85513 (LEELA)LILA F             2005-01-01          German Sheph~ Manhat~
##  4      64584 0HSO        M             2010-06-01          Unknown       Manhat~
##  5      94698 1           M             2014-06-01          Maltese       Brookl~
##  6      46118 166Y        M             2013-05-01          Pug           Manhat~
##  7      29226 2J          M             2014-11-01          American Pit~ Queens
##  8      32292 A           M             2015-06-01          Yorkshire Te~ Bronx
##  9      44557 A.          M             2014-08-01          Beagle        Queens
## 10     105430 A.J         M             2012-02-01          Maltese       Queens
## # ... with 118,532 more rows, and 9 more variables: zip_code <dbl>,
## #   community_district <dbl>, census_tract_2010 <dbl>,
## #   neighborhood_tabulation_area <chr>, city_council_district <dbl>,
## #   congressional_district <dbl>, state_senatorial_district <dbl>,
## #   license_issued_date <date>, license_expired_date <date>
```

It shows that it will provide puncuation first, then numbers, then letters. In other words, puncuation and numbers are given first priority.

```
#select(dog_licenses, animal_name)
#select(dog_licenses, animal_name, breed_name)
```

```
#dog_by_date <- arrange(dog_licenses, license_issued_date)
#select(dog_by_date, license_issued_date)
```

```
#select(dog_licenses, starts_with("Animal"))
```

```
#?dplyr::select
```

```
#arrange(dog_licenses, license_issued_date)
#dog_licenses %>% arrange(license_issued_date)
```

```
#dog_licenses %>%
#  arrange(license_issued_date) %>%
#  select(license_issued_date)
```

4.4.5 Exercise:

```
#select(dog_licenses, animal_name, breed_name)
dog_licenses %>% select(animal_name, breed_name)
```

```
## # A tibble: 118,542 x 2
##    animal_name breed_name
```

```
##    <chr>       <chr>
##  1 BONITA      Unknown
##  2 ROCKY       Labrador Retriever Crossbreed
##  3 BULLY       American Pit Bull Terrier/Pit Bull
##  4 COCO        Labrador Retriever
##  5 SKI         American Pit Bull Terrier/Pit Bull
##  6 CHASE       Shih Tzu
##  7 CHEWY       Shih Tzu
##  8 CHASE       Labrador Retriever
##  9 MILEY       Boxer
## 10 KENZI       Schnauzer, Miniature
## # ... with 118,532 more rows
```

```
#name_and_breed <- select(dog_licenses, animal_name, breed_name)
#arrange(name_and_breed, breed_name)
dog_licenses %>%
  select(animal_name, breed_name) %>%
  arrange(breed_name)
```

```
## # A tibble: 118,542 x 2
##    animal_name breed_name
##    <chr>       <chr>
##  1 LOKI        Affenpinscher
##  2 UNKNOWN     Affenpinscher
##  3 IVY         Affenpinscher
##  4 FRANKLIN    Affenpinscher
##  5 BONNIE      Affenpinscher
##  6 KUBIAK      Affenpinscher
##  7 KING        Affenpinscher
##  8 TINKERBELLE Affenpinscher
##  9 KENZIE      Affenpinscher
## 10 JULES       Affenpinscher
## # ... with 118,532 more rows
```

```
#dog_licenses %>% filter(animal_name == "BRUNO")
```

```
#dog_licenses %>% filter(license_issued_date == animal_birth_month)
```

```
#dog_licenses %>% filter("animal_gender" == "M")
#dog_licenses %>% filter(animal_gender == "M")
```

4.4.7 Exercise:

```
dog_licenses %>% filter(animal_name == "SPOCK")
```

```
## # A tibble: 10 x 15
##    row_number animal_name animal_gender animal_birth_month breed_name    borough
##         <dbl> <chr>       <chr>         <date>              <chr>         <chr>
##  1      94591 SPOCK       M             2010-01-01          American Pit~ Brookl~
##  2      94592 SPOCK       M             2010-01-01          American Pit~ Brookl~
```

```
##  3      84225 SPOCK       M             2012-02-01         Bichon Frise  Brookl~
##  4      97976 SPOCK       M             2013-05-01         Great Dane    Brookl~
##  5     116491 SPOCK       M             2016-08-01         Bull Dog, En~ Manhat~
##  6     110039 SPOCK       M             2012-06-01         Chihuahua     Manhat~
##  7      58033 SPOCK       M             2005-06-01         Jack Russell~ Manhat~
##  8      99451 SPOCK       M             2009-07-01         Unknown       Manhat~
##  9      22629 SPOCK       M             2011-11-01         Unknown       Bronx
## 10      76635 SPOCK       M             2011-11-01         Unknown       Bronx
## # ... with 9 more variables: zip_code <dbl>, community_district <dbl>,
## #   census_tract_2010 <dbl>, neighborhood_tabulation_area <chr>,
## #   city_council_district <dbl>, congressional_district <dbl>,
## #   state_senatorial_district <dbl>, license_issued_date <date>,
## #   license_expired_date <date>
```

```
dog_licenses %>% filter(animal_name == "PICARD")
```

```
## # A tibble: 1 x 15
##   row_number animal_name animal_gender animal_birth_month breed_name borough
##        <dbl> <chr>       <chr>         <date>             <chr>      <chr>
## 1      95046 PICARD      M             2014-06-01         Maltese    Staten Isl~
## # ... with 9 more variables: zip_code <dbl>, community_district <dbl>,
## #   census_tract_2010 <dbl>, neighborhood_tabulation_area <chr>,
## #   city_council_district <dbl>, congressional_district <dbl>,
## #   state_senatorial_district <dbl>, license_issued_date <date>,
## #   license_expired_date <date>
```

```
dog_licenses %>% filter(animal_name == "JANEWAY")
```

```
## # A tibble: 0 x 15
## # ... with 15 variables: row_number <dbl>, animal_name <chr>,
## #   animal_gender <chr>, animal_birth_month <date>, breed_name <chr>,
## #   borough <chr>, zip_code <dbl>, community_district <dbl>,
## #   census_tract_2010 <dbl>, neighborhood_tabulation_area <chr>,
## #   city_council_district <dbl>, congressional_district <dbl>,
## #   state_senatorial_district <dbl>, license_issued_date <date>,
## #   license_expired_date <date>
```

```
dog_licenses %>% filter(animal_name == "HARRY")
```

```
## # A tibble: 146 x 15
##    row_number animal_name animal_gender animal_birth_month breed_name    borough
##         <dbl> <chr>       <chr>         <date>             <chr>         <chr>
##  1       9310 HARRY       M             2009-11-01         Jack Russell~ Queens
##  2      56325 HARRY       M             2013-02-01         Tibetan Terr~ Queens
##  3       4333 HARRY       M             2014-10-01         Unknown       Queens
##  4      10604 HARRY       M             2013-12-01         Yorkshire Te~ Queens
##  5      56326 HARRY       M             2014-11-01         Labrador Ret~ Queens
##  6      67303 HARRY       M             2013-12-01         Yorkshire Te~ Queens
##  7      96650 HARRY       M             2015-07-01         Miniature Sc~ Queens
##  8     101658 HARRY       M             2007-11-01         Chihuahua     Queens
##  9     106173 HARRY       M             2012-06-01         French Bulld~ Queens
## 10      32115 HARRY       M             2011-05-01         Papillon      Queens
```

```
## # ... with 136 more rows, and 9 more variables: zip_code <dbl>,
## #   community_district <dbl>, census_tract_2010 <dbl>,
## #   neighborhood_tabulation_area <chr>, city_council_district <dbl>,
## #   congressional_district <dbl>, state_senatorial_district <dbl>,
## #   license_issued_date <date>, license_expired_date <date>
```

```
#dog_licenses %>% filter(animal_name == "BRUNO")
#dog_licenses %>% filter(borough == "Brooklyn")
#dog_licenses %>% filter((animal_name == "BRUNO") & (borough == "Brooklyn"))
```

```
#dog_licenses %>% filter((animal_name == "BRUNO") | (animal_name == "BRUCE"))
```

```
#dog_licenses %>% filter((animal_name == "BRUNO") | (animal_name == "BRUCE") | (animal_name == "BRADY"))
```

```
#dog_licenses %>% filter(animal_name %in% c("BRUNO", "BRUCE", "BRADY"))
```

4.4.10 Exercise:

```
start_of_2016 <- as.Date("2016-01-01")
endt_of_2016 <- as.Date("2016-12-31")

dog_licenses %>% filter(license_expired_date %in% c(start_of_2016, endt_of_2016))
```

```
## # A tibble: 127 x 15
##    row_number animal_name animal_gender animal_birth_month breed_name    borough
##         <dbl> <chr>       <chr>         <date>              <chr>         <chr>
##  1      45217 BLAZE       M             2012-10-01          American Pit~ Queens
##  2      45221 MARLEY      M             2013-01-01          Yorkshire Te~ Queens
##  3      45248 CHARLIE     M             2015-09-01          Rat Terrier   Queens
##  4      45288 PHOENIX     F             2004-01-01          Pug           Queens
##  5       1920 SASHA       F             2014-03-01          Border Collie Queens
##  6      42203 MOOSE       M             2008-12-01          Pug           Queens
##  7       1912 POLLIE      F             2012-12-01          American Pit~ Queens
##  8      23640 MAX         M             2009-07-01          Unknown       Queens
##  9      42799 NAPOLEON    M             2004-01-01          Pug           Queens
## 10      45235 JASPER      M             2015-04-01          Boxer         Queens
## # ... with 117 more rows, and 9 more variables: zip_code <dbl>,
## #   community_district <dbl>, census_tract_2010 <dbl>,
## #   neighborhood_tabulation_area <chr>, city_council_district <dbl>,
## #   congressional_district <dbl>, state_senatorial_district <dbl>,
## #   license_issued_date <date>, license_expired_date <date>
```

```
#dog_licenses %>%
#  mutate(called_chase = animal_name == "CHASE") %>%
#  select(animal_name, called_chase)
```

```
#dog_licenses %>%
#  mutate(called_chase = animal_name == "CHASE") %>%
#  filter(called_chase)
```

```
#dog_licenses %>%
#  mutate(called_chase = ifelse(animal_name == "CHASE", "called chase", "not called chase")) %>%
#  select(animal_name, called_chase)
```

```
#dog_licenses %>%
#  mutate(is_chase = animal_name == "CHASE",called_chase = ifelse(is_chase, "called chase", "not called
#  select(animal_name, is_chase, called_chase)
```

```
#dog_licenses %>%
#  mutate(license_duration = license_expired_date - license_issued_date) %>%
#  select(license_duration)

#dog_licenses %>%
#  mutate(license_duration = license_expired_date - license_issued_date,
#    avg_duration = mean(license_duration)) %>%
#  select(license_duration, avg_duration)
```

4.5.3 Exercise:

```
dog_licenses %>%
  mutate(name_length = stringr::str_length(animal_name)) %>%
  arrange(desc(name_length))
```

```
## # A tibble: 118,542 x 16
##    row_number animal_name    animal_gender animal_birth_mon~ breed_name  borough
##         <dbl> <chr>          <chr>         <date>            <chr>       <chr>
##  1      23714 CARLYAPPLEWHI~ F             2013-11-01        Havanese    Manhat~
##  2      48757 JEFFERSONBARN~ M             2011-07-01        Jack Russe~ Manhat~
##  3      93279 PIPLONGFELLOW~ M             2013-08-01        Jack Russe~ Manhat~
##  4      29051 SAMSONMAXWELL~ M             2014-03-01        Yorkshire ~ Manhat~
##  5      51152 BUDDYVONYANKE~ M             2007-11-01        Pointer, G~ Brookl~
##  6      78023 EMILIE.BUNNEL~ M             2008-04-01        Jack Russe~ Brookl~
##  7      44088 SHAWN-MICHAEL~ M             2012-09-01        Cocker Spa~ Staten~
##  8     118591 FLYNN-(BILLYG~ M             2008-11-01        Greyhound   Staten~
##  9      11999 DANGERFIELDS-~ M             2010-09-01        Bull Dog, ~ Manhat~
## 10      92696 EUNICETHOMPSO~ F             2010-03-01        Pug         Manhat~
## # ... with 118,532 more rows, and 10 more variables: zip_code <dbl>,
## #   community_district <dbl>, census_tract_2010 <dbl>,
## #   neighborhood_tabulation_area <chr>, city_council_district <dbl>,
## #   congressional_district <dbl>, state_senatorial_district <dbl>,
## #   license_issued_date <date>, license_expired_date <date>, name_length <int>
```

There are four dogs with 30 character names with the classified longest name of CARLYAPPLEWHITE-CRAWFORDCOLEMAN.

Apply

```
us_pop <- read_rds("http://data.cwick.co.nz/us-population.rds")
us_pop
```

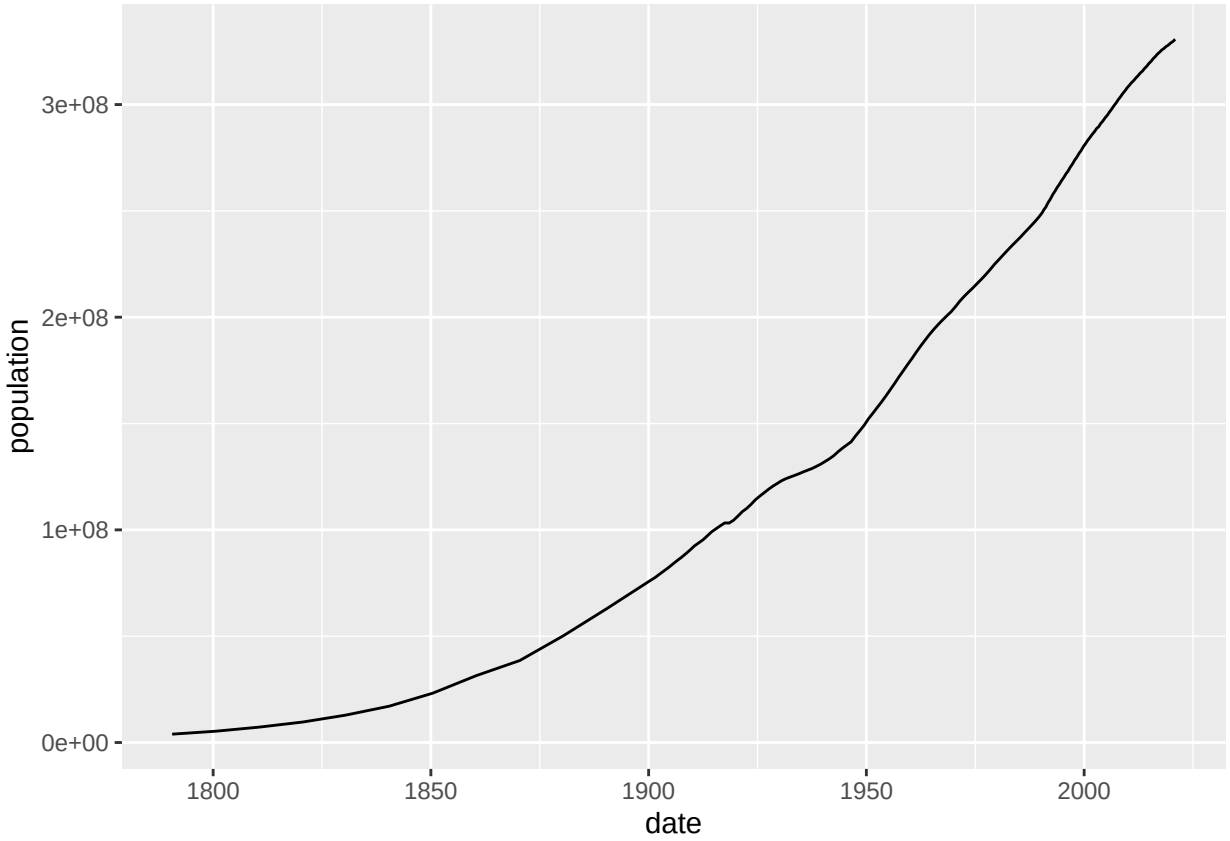
```
## # A tibble: 470 x 3
##    date       population  year
##    <date>          <dbl> <dbl>
##  1 2004-09-01  293309033  2004
##  2 1991-03-01  251681661  1991
##  3 2017-11-01  325663325  2017
##  4 2017-05-01  324636728  2017
##  5 2013-08-01  316202222  2013
##  6 2005-02-01  294380349  2005
##  7 1890-06-01   62979766  1890
##  8 2016-03-01  322196377  2016
##  9 1944-07-01  138397345  1944
## 10 1880-06-01   50189209  1880
## # ... with 460 more rows
```

1.

```
us_pop <- arrange(us_pop, date)
```

2.

```
ggplot(us_pop, aes(date, population)) +
  geom_line()
```



3.

```
us_pop <- us_pop %>%
  mutate(prev_population = lag(population))
```

This compares the "current" population to the previous population. In other words, it compares each rows population to the population of the row above.

4.

```
us_pop <- us_pop %>%
  mutate(change_rate = (population/prev_population))
us_pop
```

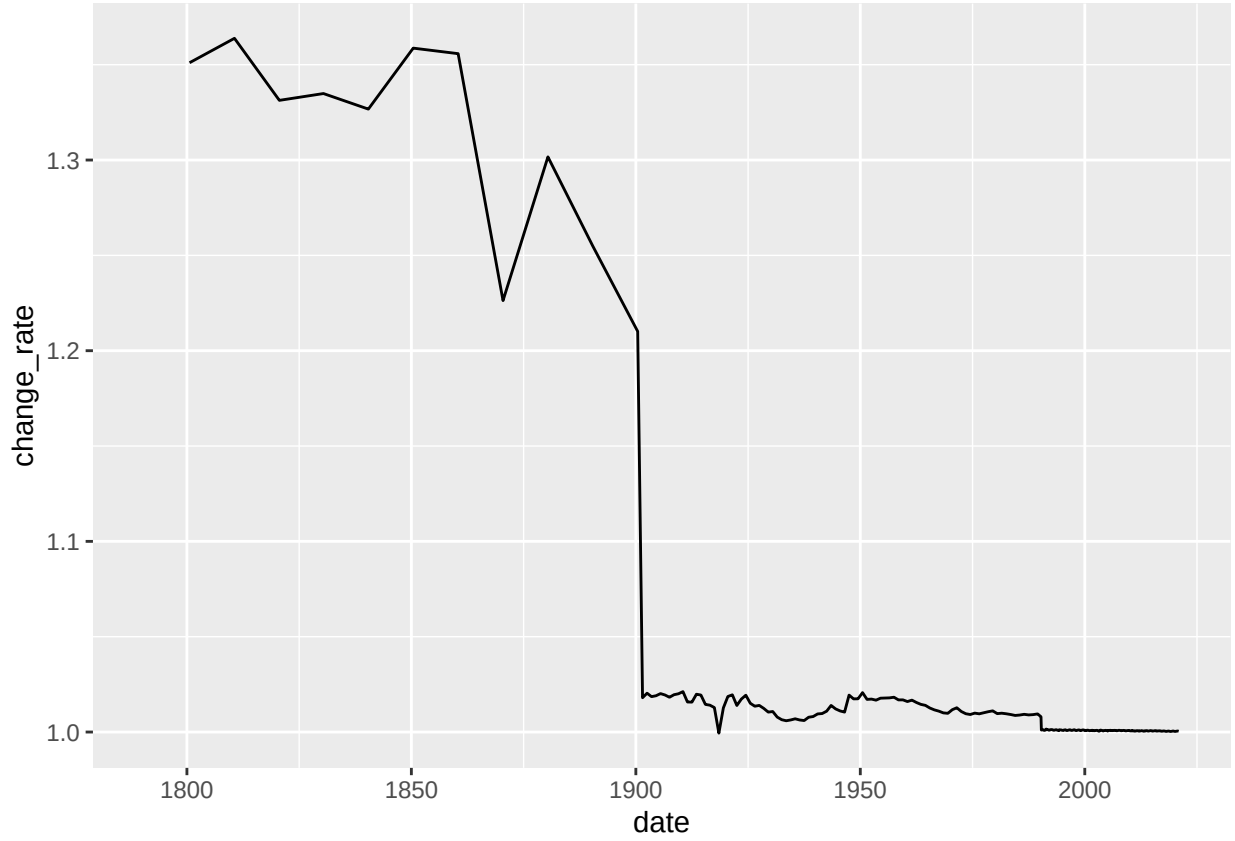
```
## # A tibble: 470 x 5
##    date       population  year prev_population change_rate
##    <date>          <dbl> <dbl>           <dbl>       <dbl>
##  1 1790-08-02    3929214  1790              NA       NA
##  2 1800-08-04    5308483  1800         3929214        1.35
##  3 1810-08-06    7239881  1810         5308483        1.36
##  4 1820-08-07    9638453  1820         7239881        1.33
```

```
##  5 1830-06-01   12866020  1830         9638453        1.33
##  6 1840-06-01   17069453  1840        12866020        1.33
##  7 1850-06-01   23191876  1850        17069453        1.36
##  8 1860-06-01   31443321  1860        23191876        1.36
##  9 1870-06-01   38558371  1870        31443321        1.23
## 10 1880-06-01   50189209  1880        38558371        1.30
## # ... with 460 more rows
```

5.

```
ggplot(us_pop, aes(date, change_rate)) +
  geom_line()
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

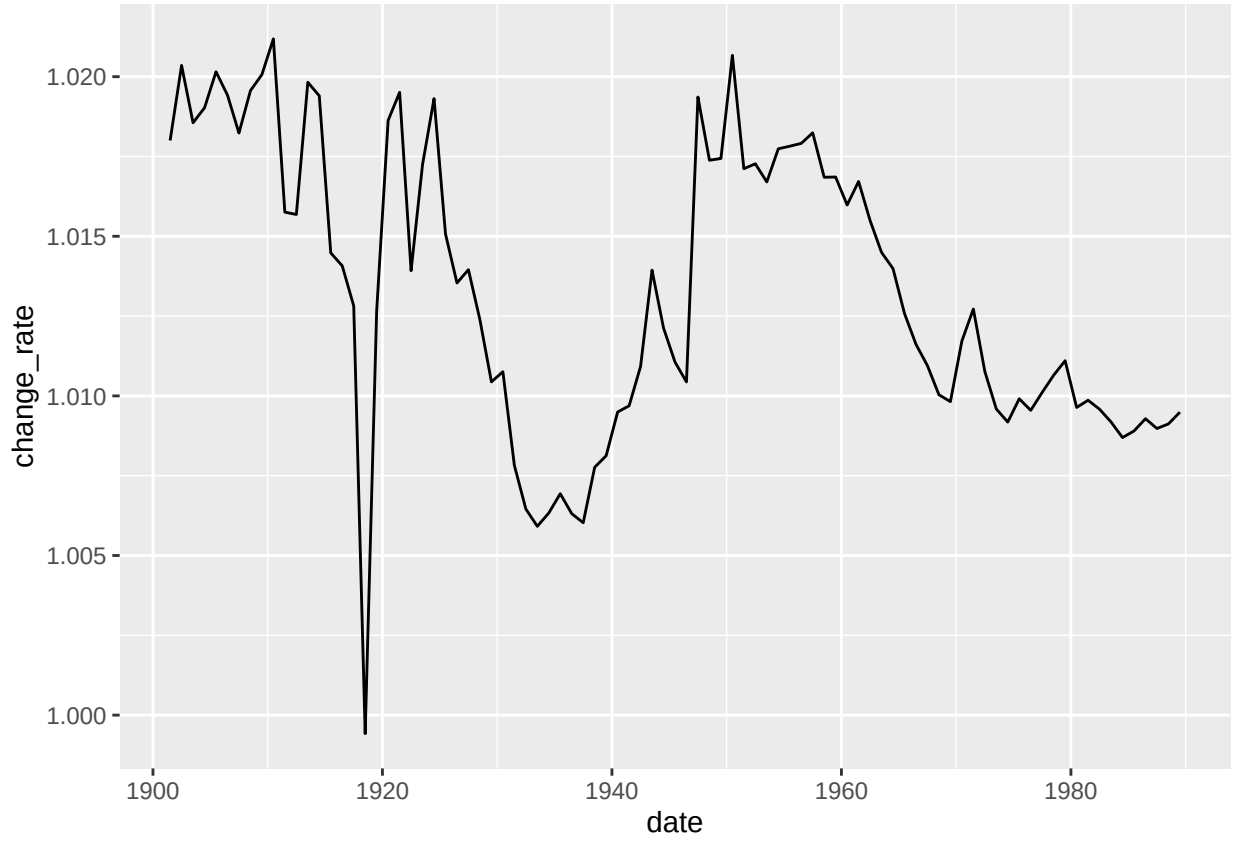



```
  #geom_point()
```

6.

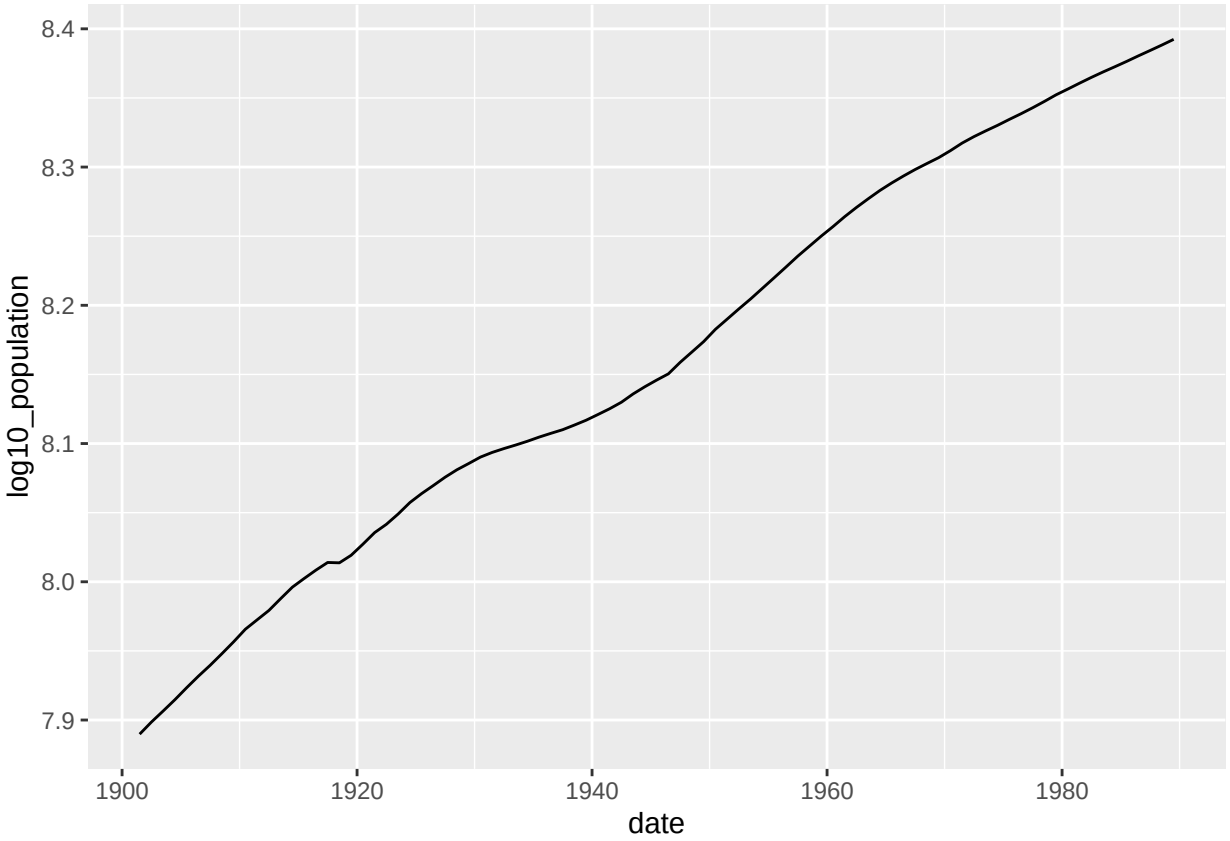
```
us_pop <- us_pop %>% filter(year %in% ("1901":"1989"))
```

```
ggplot(us_pop, aes(date, change_rate)) +
  geom_line()
```



7.

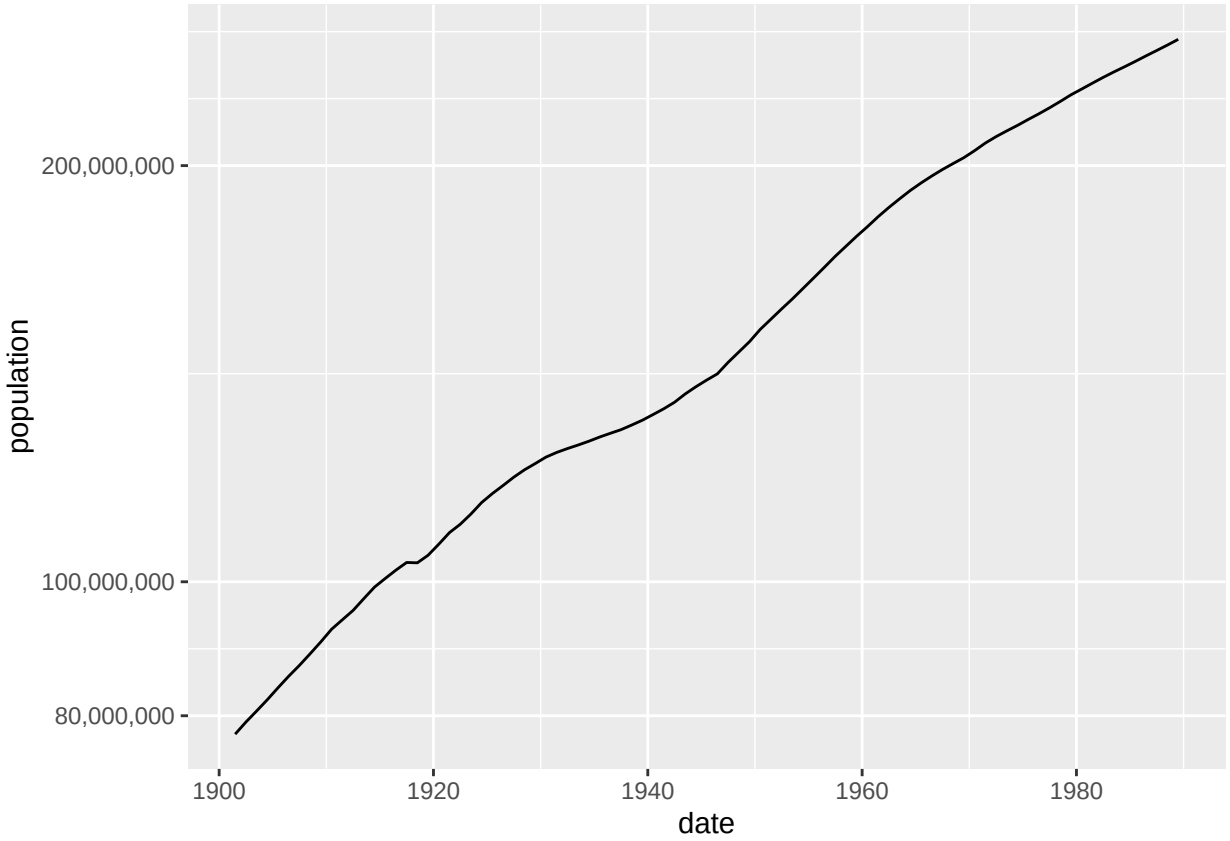
```
us_pop <- us_pop %>%
  mutate(log10_population = log10(population))

ggplot(us_pop, aes(date, log10_population)) +
  geom_line()
```



The population has steadily increased over the years.

8.

```
us_pop %>%
  ggplot(aes(date, population)) +
    geom_line() +
    scale_y_log10(labels = scales::label_comma())
```



They both have a similar trend in increase rate. I find the second one easier to understand because the I find the populatoin number easier to understand.