|  |
| --- |
| It’s clear that there are some economic shifts happening in the world, if not the US itself.  In light of this, I decided to do some simple investigation into the economic performance of US cities.  This is, by the way, one of the critical reasons to master data science. One you know a few critical skills, you will be able to very rapidly get some basic information about (almost) any topic.  In a case such as this (when you’re just personally interested), you can just scrape some data and plot it.  A large part of your job will be gathering data and quickly plotting it in ways that generate insight …  **Plotting GDP data for top US cities**  In the following code, we’ll scrape some data about US cities and plot a line chart using ggplot2.  There’s actually quite a bit more that we could do with this data, so feel free to create your own plots and leave the code in the comments below.  #=================  # INSTALL PACKAGES  #=================  library(tidyverse)  library(stringr)  library(forcats)  library(rvest)  library(ggthemes)  #============  # SCRAPE DATA  #============  df.metro\_gdp <- read\_html('https://en.wikipedia.org/wiki/List\_of\_U.S.\_metropolitan\_areas\_by\_GDP') %>%  html\_nodes('table') %>%  .[[1]] %>%  html\_table() %>%  as.tibble()  #=======================  # REMOVE 'Rank' VARIABLE  #=======================  df.metro\_gdp <- df.metro\_gdp %>%  select(-Rank)  #================  # RENAME VARIABLE  #================  df.metro\_gdp <- df.metro\_gdp %>% rename(metro\_area = `Metropolitan area`)  # inspect  df.metro\_gdp  # REMOVE 'MSA' FROM metro\_area  df.metro\_gdp <- df.metro\_gdp %>% mutate(metro\_area = str\_replace(metro\_area, ' MSA', ''))  # COERCE TO 'metro\_area' FACTOR  df.metro\_gdp <- df.metro\_gdp %>% mutate(metro\_area = metro\_area %>% as\_factor())  #========================================================  # CREATE NEW VARIABLE:  # - the original 'metro\_area' variable is rather long  # because it's a full 'metropolitan statistical area'  # - we can abbreviate these as the plain city name  # - we'll call the new variable 'metro\_brief'  #========================================================  # get unique values  df.metro\_gdp %>%  select(metro\_area) %>%  unique()  #---------------------------------------------------  # RECODE VALUES  # here we will create the new variable 'metro\_brief'  #---------------------------------------------------  df.metro\_gdp <- df.metro\_gdp %>%  mutate(metro\_area\_brief = recode(metro\_area,'New York–Northern New Jersey–Long Island, NY–NJ–PA' = 'New York'  ,'Los Angeles–Long Beach–Santa Ana, CA' = 'Los Angeles'  ,'Chicago–Joliet–Naperville, IL–IN–WI' = 'Chicago'  ,'Dallas–Fort Worth–Arlington, TX' = 'Dallas'  ,'Washington–Arlington–Alexandria, DC–VA–MD–WV' = 'Washington DC'  ,'Houston–Sugar Land–Baytown, TX' = 'Houston'  ,'San Francisco–Oakland–Fremont, CA' = 'San Francisco'  ,'Philadelphia–Camden–Wilmington, PA–NJ–DE–MD' = 'Philadelphia'  ,'Boston–Cambridge–Quincy, MA–NH' = 'Boston'  ,'Atlanta–Sandy Springs–Marietta, GA' = 'Atlanta'  ))  # INSPECT VALUES  df.metro\_gdp %>% glimpse()  df.metro\_gdp %>% select(metro\_area\_brief)  # CHECK TABLE OF CROSS-VALUES  df.metro\_gdp %>%  #select(metro\_area, metro\_brief) %>%  group\_by(metro\_area, metro\_area\_brief) %>%  summarise()  #======================  # RESHAPE: WIDE TO LONG  #======================  df.metro\_gdp <- df.metro\_gdp %>% gather(key = year, value = gdp\_nominal, -metro\_area, -metro\_area\_brief)  #========================  # COERCE 'year' TO FACTOR  #========================  df.metro\_gdp <- df.metro\_gdp %>% mutate(year = year %>% as.factor())  #===========================================  # WRANGLE AND COERCE 'gdp\_nominal' TO DOUBLE  #===========================================  df.metro\_gdp <- mutate(df.metro\_gdp, gdp\_nominal = str\_remove\_all(gdp\_nominal, ",") %>% as.double())  #================  # PLOT BASIC PLOT  #================  ggplot(df.metro\_gdp, aes(x = year, y = gdp\_nominal, group = metro\_area\_brief)) +  geom\_line(aes(color = metro\_area\_brief))  #==========  # FORMATTED  #==========  df.metro\_gdp %>%  mutate(highlight\_flag = if\_else(metro\_area\_brief == 'New York', T, F)) %>%  ggplot(aes(x = year, y = gdp\_nominal, group = metro\_area\_brief)) +  geom\_line(aes(color = highlight\_flag, alpha = highlight\_flag), size = 1.5) +  scale\_color\_manual(values = c('grey', 'red')) +  scale\_alpha\_manual(values = c(.7, 1)) +  labs(title = 'New York is the best performing US city by metro GDP'  ,subtitle = str\_c("Consistently, New York has a much higher GDP than other metro areas."  ,"\n77% higher than next highest metro in 2017.")  ,y = "Nominal GDP\n(metro area, millions of dollars)"  ,x = 'Year') +  theme(legend.position = 'none'  ,text = element\_text(color = '#3A3A3A'  ,family = 'sans')  ,plot.title = element\_text(margin = margin(b = 10)  ,face = 'bold'  ,size = 20)  ,axis.title = element\_text()  ,plot.subtitle = element\_text(size = 12)  ) +  scale\_y\_continuous(labels = scales::comma\_format())  And here is the finalized chart: |