Data Analysis of Flow of Commodities in U.S.

Mohit Bansal and Karan Jain

2017-08-03

```r
library(tidyverse)
library(readxl)
library(stringr)
library(reshape2)
library(rvest)
library(purrr)

# read data from local file
data <- read.table("csr.txt", header = TRUE)
```

Loading the data, tidy and transform it

```r
# load data into a tibble
cfs <- as.tibble(data) %>%
  # tidying the data
  # separating into different columns
  separate('SHIPMT_ID.ORIG_STATE.ORIG_MA.ORIG_CFS_AREA.DEST_STATE.DEST_MA.DEST_CFS_AREA.NAICS.QUARTER.S
           into = c("SHIPMT_ID", "ORIG_STATE", "ORIG_MA", "ORIG_CFS_AREA",
                    "DEST_STATE", "DEST_MA", "DEST_CFS_AREA",
                    "NAICS", "QUARTER", "SCTG", "MODE", "SHIPMT_VALUE",
                    "SHIPMT_WGHT", "SHIPMT_DIST_GC","SHIPMT_DIST_ROUTED",
                    "TEMP_CNTL_YN", "EXPORT_YN", "EXPORT_CNTRY",
                    "HAZMAT", "WGT_FACTOR"), sep = ",") %>%
  # selecting the relevant columns and transforming them
  select(ORIG_STATE, DEST_STATE, NAICS, SCTG, MODE, SHIPMT_VALUE, SHIPMT_WGHT, QUARTER) %>%
  filter(ORIG_STATE != "00") %>%
  mutate(ORIG_STATE = str_replace_all(ORIG_STATE, c("01" = "Alabama", "02" = "Alaska",
                                                    "04" = "Arizona", "05" = "Arkansas",
                                                    "06" = "California", "08" = "Colorado",
                                                    "09" = "Connecticut", "10" = "Delaware",
                                                    "11" = "District of Columbia", "12" = "Florida",
                                                    "13" = "Georgia", "15" = "Hawaii",
                                                    "16" = "Idaho", "17" = "Illinois",
                                                    "18" = "Indiana", "19" = "Iowa",
                                                    "20" = "Kansas", "21" = "Kentucky",
                                                    "22" = "Louisiana", "23" = "Maine",
                                                    "24" = "Maryland", "25" = "Massachusetts",
                                                    "26" = "Michigan", "27" = "Minnesota",
                                                    "28" = "Mississippi", "29" = "Missouri",
                                                    "30" = "Montana", "31" = "Nebraska",
                                                    "32" = "Nevada", "33" = "New Hampshire",
                                                    "34" = "New Jersey", "35" = "New Mexico",
                                                    "36" = "New York", "37" = "North Carolina",
                                                    "38" = "North Dakota", "39" = "Ohio",
                                                    "40" = "Oklahoma", "41" = "Oregon",
                                                    "42" = "Pennsylvania", "44" = "Rhode Island",
                                                    "45" = "South Carolina", "46" = "South Dakota",
                                                    "47" = "Tennessee", "48" = "Texas",
                                                    "49" = "Utah", "50" = "Vermont",
                                                    "51" = "Virginia", "53" = "Washington",
                                                    "54" = "West Virginia", "55" = "Wisconsin",
                                                    "56" = "Wyoming")),
         DEST_STATE = str_replace_all(DEST_STATE, c("01" = "Alabama", "02" = "Alaska",
                                                    "04" = "Arizona", "05" = "Arkansas",
                                                    "06" = "California", "08" = "Colorado",
                                                    "09" = "Connecticut", "10" = "Delaware",
                                                    "11" = "District of Columbia", "12" = "Florida",
                                                    "13" = "Georgia", "15" = "Hawaii",
                                                    "16" = "Idaho", "17" = "Illinois",
                                                    "18" = "Indiana", "19" = "Iowa",
                                                    "20" = "Kansas", "21" = "Kentucky",
                                                    "22" = "Louisiana", "23" = "Maine",
                                                    "24" = "Maryland", "25" = "Massachusetts",
                                                    "26" = "Michigan", "27" = "Minnesota",
                                                    "28" = "Mississippi", "29" = "Missouri",
                                                    "30" = "Montana", "31" = "Nebraska",
                                                    "32" = "Nevada", "33" = "New Hampshire",
                                                    "34" = "New Jersey", "35" = "New Mexico",
                                                    "36" = "New York", "37" = "North Carolina",
                                                    "38" = "North Dakota", "39" = "Ohio",
                                                    "40" = "Oklahoma", "41" = "Oregon",
                                                    "42" = "Pennsylvania", "44" = "Rhode Island",
                                                    "45" = "South Carolina", "46" = "South Dakota",
                                                    "47" = "Tennessee", "48" = "Texas",
                                                    "49" = "Utah", "50" = "Vermont",
                                                    "51" = "Virginia", "53" = "Washington",
                                                    "54" = "West Virginia", "55" = "Wisconsin",
                                                    "56" = "Wyoming")),
         SCTG = str_extract(SCTG, "^\\d+"),
         SCTG = as.numeric(SCTG),
         SHIPMT_VALUE = as.double(SHIPMT_VALUE),
         SHIPMT_WGHT  = as.double(SHIPMT_WGHT),
         NAICS = as.integer(NAICS),
         QUARTER = as.integer(QUARTER))

cfs %>%
  head(10)
```

```
## # A tibble: 10 × 8
##       ORIG_STATE    DEST_STATE NAICS  SCTG  MODE SHIPMT_VALUE SHIPMT_WGHT
##            <chr>         <chr> <int> <dbl> <chr>        <dbl>       <dbl>
## 1  Massachusetts Massachusetts   333    35    14         2178          11
## 2   Pennsylvania    California   311    35    14          344          11
## 3       Michigan     Tennessee   322    27    04         4197        5134
## 4         Kansas        Kansas   323    29    04          116           6
## 5        Florida       Florida  4235    33    05          388         527
## 6       Maryland       Montana   337    40    04         3716        1132
## 7           Iowa          Iowa   337    26    05        43738       13501
## 8     California    California  4239    40    14           77           4
## 9           Iowa          Iowa   327    31    05          338       12826
## 10       Georgia       Georgia  4237    34    05          145          22
## # ... with 1 more variables: QUARTER <int>
```

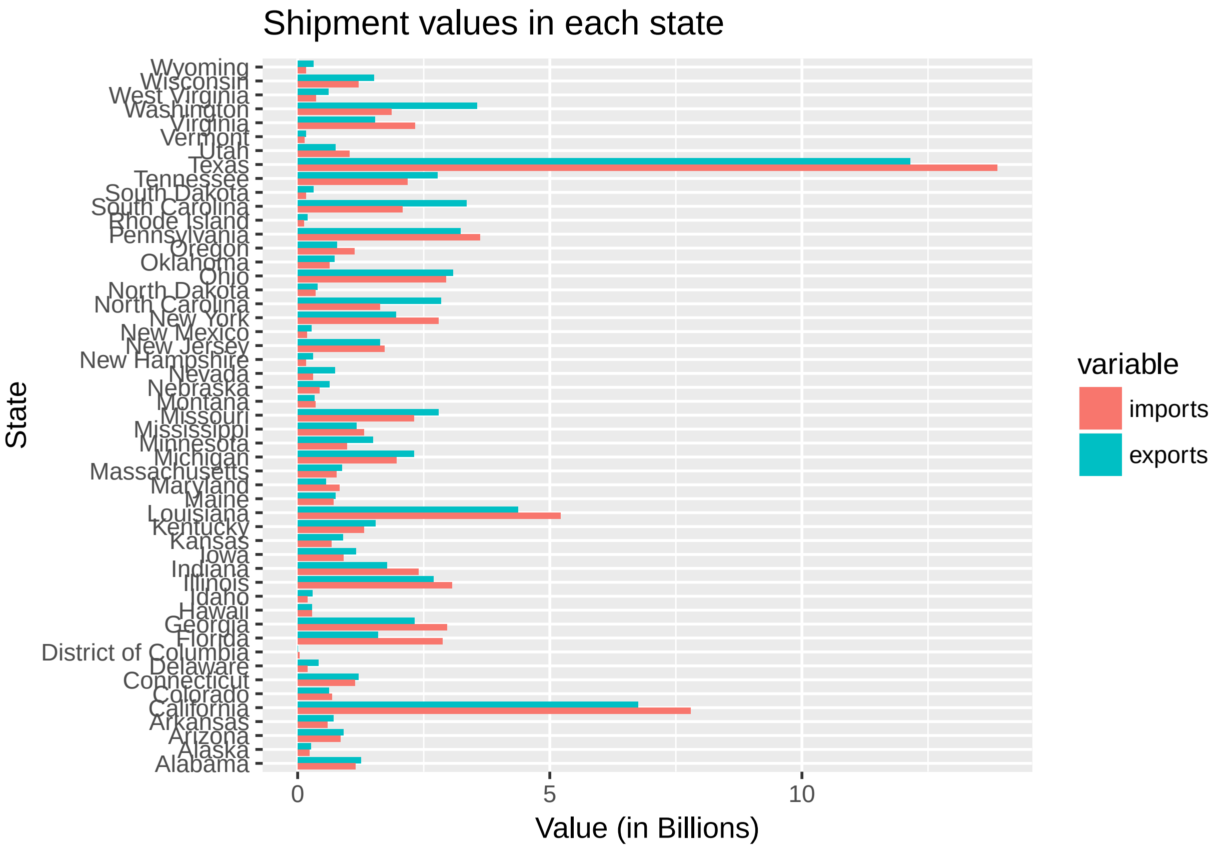
Creating the webscraping function

```
# create function to scrape gdp of various states
gdp_scrape <- function(year = 2007) {

  url1 <- "http://www.usgovernmentspending.com/compare_state_spending_%sbZ0a"
  url1 <- sprintf(url1, year)
  url <- read_html(url1)

  states <- url %>%
    html_nodes("td.lbltier") %>%
    html_text(trim = T)

  spending <- url %>%
    html_nodes(".lbltier+ .sptiera") %>%
    html_text(trim = T)


  debt <- url %>%
    html_nodes(".sptier") %>%
    html_text(trim = T)

  gsp <- url %>%
    html_nodes(".sptier+ .sptiera") %>%
    html_text(trim = T)

  rgr <- url %>%
    html_nodes(".sptiera+ .sptiera") %>%
    html_text(trim = T)

  pop <- url %>%
    html_nodes(".sptiera+ td:nth-child(7)") %>%
    html_text(trim = T)


  tibble(
    state = states,
    spending = spending,
    debt = debt,
    gsp = gsp,
    rgr = rgr,
    pop = pop
  )
}

# run scraping fuction and transform the data
gdp <- gdp_scrape(year = 2012) %>%
  filter(state != "All states combined") %>%
  # transforming data using strings and regular expressions
  mutate(spending = str_extract(spending, "\\w+.\\w"),
         spending = as.double(spending),
         debt = str_extract(debt, "\\w+.\\w"),
         debt = as.double(debt),
         gsp = str_extract(gsp,"(\\w.)?\\w+\\.\\w+"),
         gsp = str_replace(gsp, ",",""),
         gsp = as.double(gsp))

gdp %>%
  head(10)
```

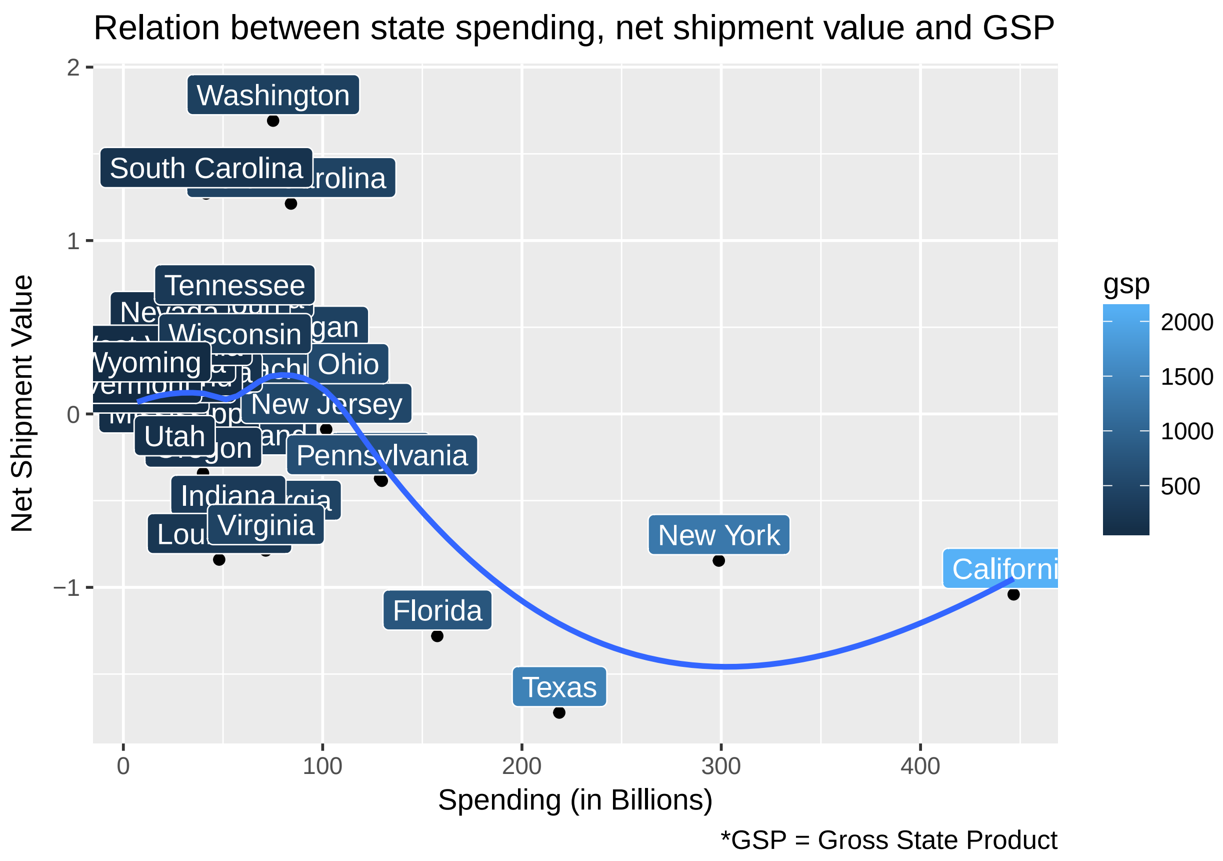
```
## # A tibble: 10 × 6
##                    state spending  debt    gsp   rgr   pop
##                    <chr>    <dbl> <dbl>  <dbl> <chr> <chr>
## 1                Alabama     41.7  29.3  185.9  1.0%   4.8
## 2                 Alaska     14.7   9.5   60.9  5.3%   0.7
## 3                Arizona     51.2  49.5  264.7  2.1%   6.5
## 4               Arkansas     24.3  14.0  109.2 -0.1%   3.0
## 5             California    446.7 420.1 2131.2  2.6%  38.0
## 6               Colorado     49.7  51.4  272.8  2.1%   5.2
## 7            Connecticut     41.2  42.8  239.5 -0.1%   3.6
## 8               Delaware     10.5   8.2   60.6 -1.6%   0.9
## 9   District of Columbia     13.8  11.6  109.7  0.2%   0.6
## 10               Florida    157.5 146.9  764.1  0.8%  19.3
```

Data Transformation and Visualization

```
# calculate sum of exported shipments from each state
exp <- cfs %>%
  group_by(ORIG_STATE) %>%
  summarise(exports = sum(SHIPMT_VALUE)) %>%
  mutate(exports = exports/1000000000,
         exports = round(exports, digits = 4)) %>%
  rename(state = ORIG_STATE)

# calculate sum of imported shipments for each state
imp <- cfs %>%
  group_by(DEST_STATE) %>%
  summarise(imports = sum(SHIPMT_VALUE)) %>%
  mutate(imports = imports/1000000000,
         imports = round(imports, digits = 4)) %>%
  rename(state = DEST_STATE) %>%
  # joining the imported shipments with the exported ones
  inner_join(exp, by = "state")

# combine the previous two tables
eximp <- melt(imp, id.vars = "state")

# plot shipment value of each state
ggplot(data = eximp, aes(x = state, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(title = "Shipment values in each state",
       y = "Value (in Billions)",
       x = "State")
```

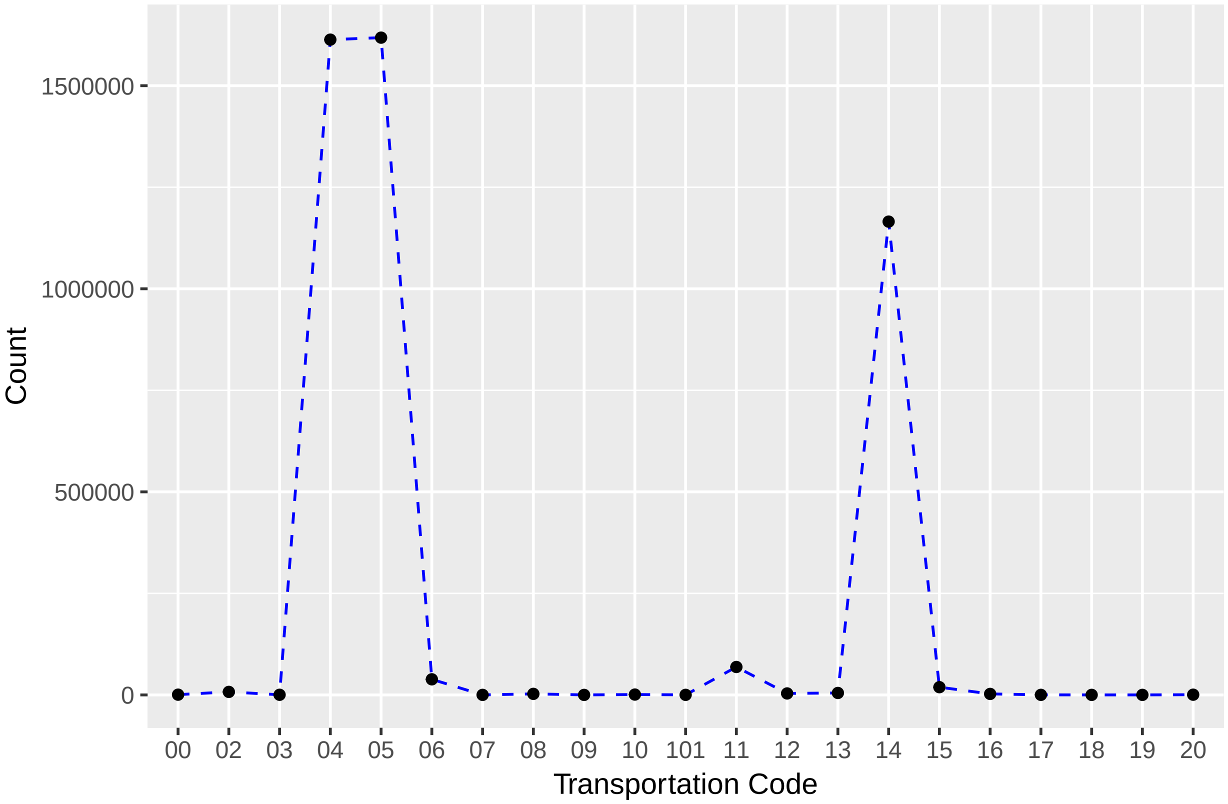


```
# join gdp data with export and import shipment data
gdp_net <- imp %>%
  inner_join(gdp, by = "state") %>%
  mutate(net = exports - imports) %>%
  # transforming the first column of dates as rownames
  remove_rownames %>%
  column_to_rownames(var="state")

# plot the graph to determine relation between gsp and shipment value
ggplot(gdp_net, aes(x = spending, y = net, fill = gsp)) +
  geom_jitter() +
  geom_label(label = rownames(gdp_net), color="white", nudge_x = 0.15, nudge_y = 0.15, check_overlap = '
  geom_smooth(se = FALSE) +
  labs(title = "Relation between state spending, net shipment value and GSP",
       x = "Spending (in Billions)",
       y = "Net Shipment Value",
       caption = "*GSP = Gross State Product")
```



Exploratory Data Analysis

```
# find out the most used mode of shipment across the country
maxmode <- group_by(cfs,MODE) %>%
  summarise(count= n())

# plot the graph for modes of shipment across the country
ggplot(maxmode, aes(x = MODE, y = count)) +
  geom_line(group = 1, color = "blue", linetype = "dashed") +
  geom_point() +
  labs(x = "Transportation Code",
       y = "Count",
       caption = "*04, 05 and 14 are various types of Trucks")
```



*04, 05 and 14 are various types of Trucks

```
# determine the most commodities shipped for the entire country
maxcom <- cfs %>%
  group_by(SCTG) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

maxcom
```

```
## # A tibble: 5 × 2
##    SCTG  count
##   <dbl>  <int>
## 1    35 319505
## 2    24 288078
## 3    43 283551
## 4    34 265539
## 5    40 264089
```

```
# 35 is Electronics
# 24 is Plastics and Rubbers
# 43 is Mixed Freight
# 34 is Machinery
# 40 is Miscellaneous Products

# determine the industry most shipped to and from across the whole country
maxind <- cfs %>%
  group_by(NAICS) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

maxind
```

```
## # A tibble: 5 × 2
##   NAICS  count
##   <int>  <int>
## 1   325 221721
## 2   332 209425
## 3  4238 199767
## 4   311 186452
## 5  4244 175812
```

```
# 325 is Chemical Manufacturing
# 332 is Fabricated Metal Industry
# 4238 is Machinery and Equipment Wholesalers
# 311 is Food Manufacturing
# 4244 is Grocery Wholesalers

# avg value of shipments across the country
avg_value <- cfs %>%
  summarise(avg = mean(SHIPMT_VALUE))

avg_value
```

```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 18279.68
```

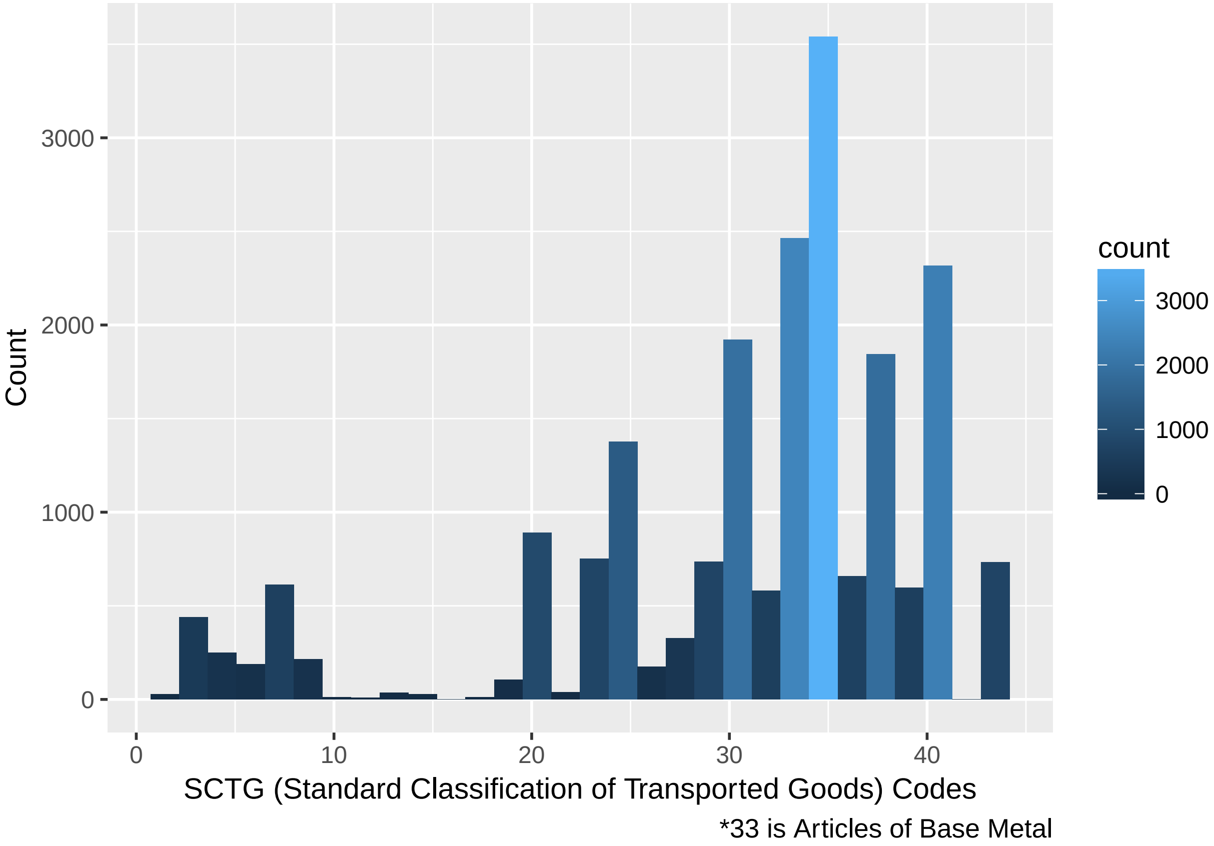
```
# avg weight of shipments across the country
avg_wght <- cfs %>%
  summarise(avg = mean(SHIPMT_WGHT))

avg_wght
```

```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 37587.62
```

```
# filtering out shipments between California and Texas
caltex <- cfs %>%
  filter((ORIG_STATE == "California" & DEST_STATE == "Texas") |
        (ORIG_STATE == "Texas" & DEST_STATE == "California"))

# determine the commodities most shipped between California and Texas
caltex_maxcom <- caltex %>%
  group_by(SCTG) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

caltex_maxcom
```

```
## # A tibble: 5 × 2
##    SCTG count
##   <dbl> <int>
## 1    35  3541
## 2    40  2234
## 3    30  1622
## 4    24  1361
## 5    34  1269
```

```
# 35 is Electronics
# 40 is Miscellaneous Products
# 30 is Textiles and Leather Products
# 24 is Plastics and Rubbers
# 34 is Machinery


# plot for the commodities shipped between Cal and TX
ggplot(caltex, aes(x = SCTG)) +
  geom_histogram(aes(fill = ..count..)) +
  labs(x = "SCTG (Standard Classification of Transported Goods) Codes",
       y = "Count",
       caption = "*33 is Articles of Base Metal")
```



```
# determine the industry most shipped to and from Cal and TX
caltex_maxind <- caltex %>%
  group_by(NAICS) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

caltex_maxind
```

```
## # A tibble: 5 × 2
##   NAICS count
##   <int> <int>
## 1   334  1556
## 2   325  1388
## 3   332  1332
## 4   339  1260
## 5  4236  1096
```

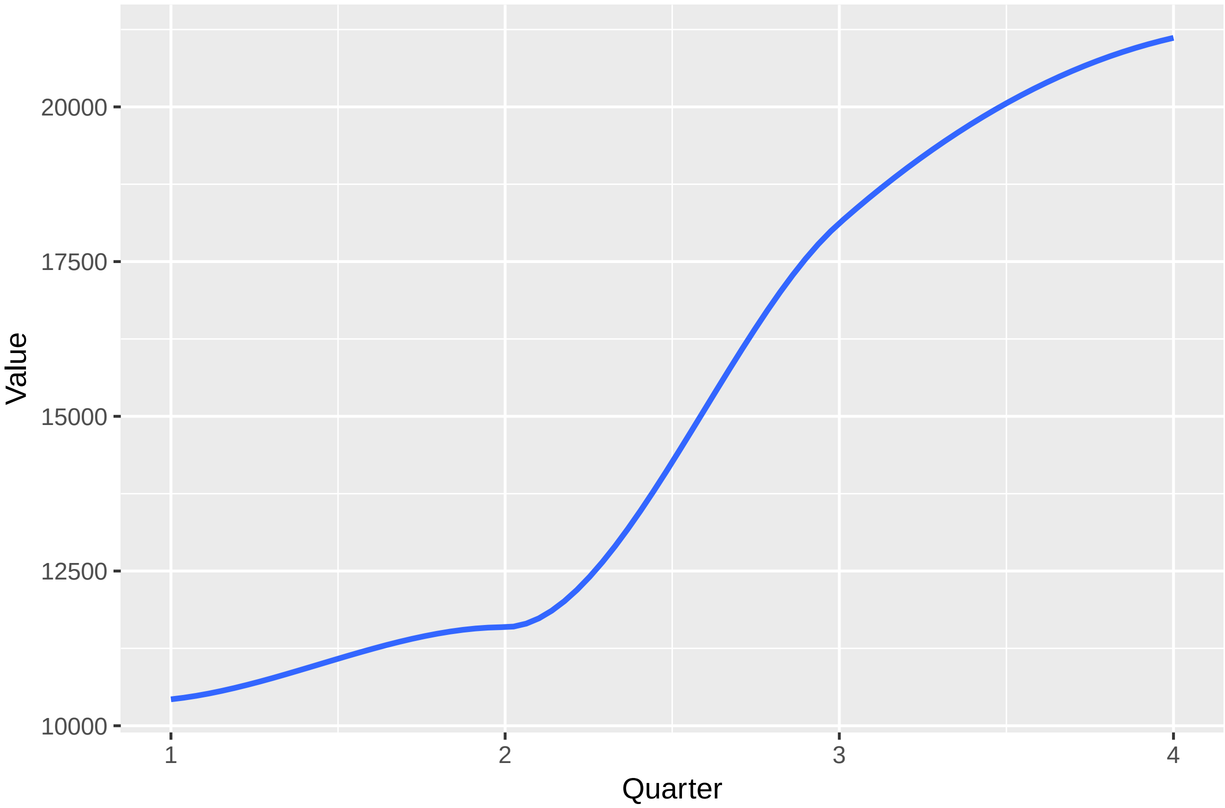
```
# 334 is Computers
# 325 is Chemical Manufacturing
# 332 is Fabricated Metal Industry
# 339 is Miscellaneous
# 4236 is Electrical

# avg value of shipments between CA and TX
caltex_avg_value <- caltex %>%
  summarise(avg = mean(SHIPMT_VALUE))

caltex_avg_value
```

```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 21610.47
```

```
# avg weight of shipments between CA and TX
caltex_avg_wght <- caltex %>%
  summarise(avg = mean(SHIPMT_WGHT))

caltex_avg_wght
```

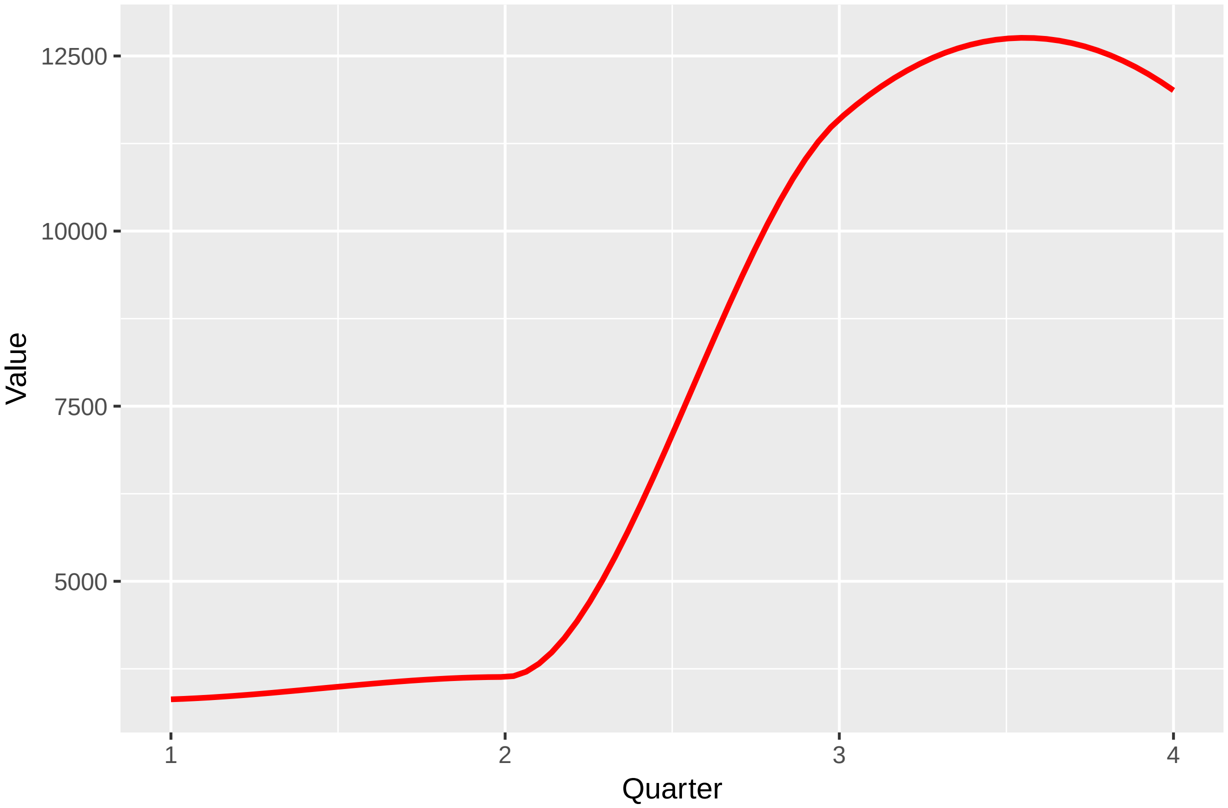
```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 19149.97
```

```
# sorting out the most served industry with shipments
# from California to Texas
caltex_4236 <- caltex %>%
  filter(ORIG_STATE == "California" & DEST_STATE == "Texas",
         NAICS == 4236)

# plot the variation of most served industry served across quarters of 2012
ggplot(caltex_4236, aes(x = QUARTER, y = SHIPMT_VALUE)) +
  geom_smooth(se = FALSE) +
  labs(x = "Quarter",
       y = "Value",
       caption = "*Graph for Electronic goods industry")
```



*Graph for Electronic goods industry

```
# sorting out the most served industry in other direction i.e,
# from Texas to California
texcal_4236 <- caltex %>%
  filter(ORIG_STATE == "Texas" & DEST_STATE == "California",
         NAICS == 4236)

# plot the variation of industry served across quarters of 2012
ggplot(texcal_4236, aes(x = QUARTER, y = SHIPMT_VALUE)) +
  geom_smooth(color = "red", se = FALSE) +
  labs(x = "Quarter",
       y = "Value",
       caption = "*Graph for Electronic goods industry")
```

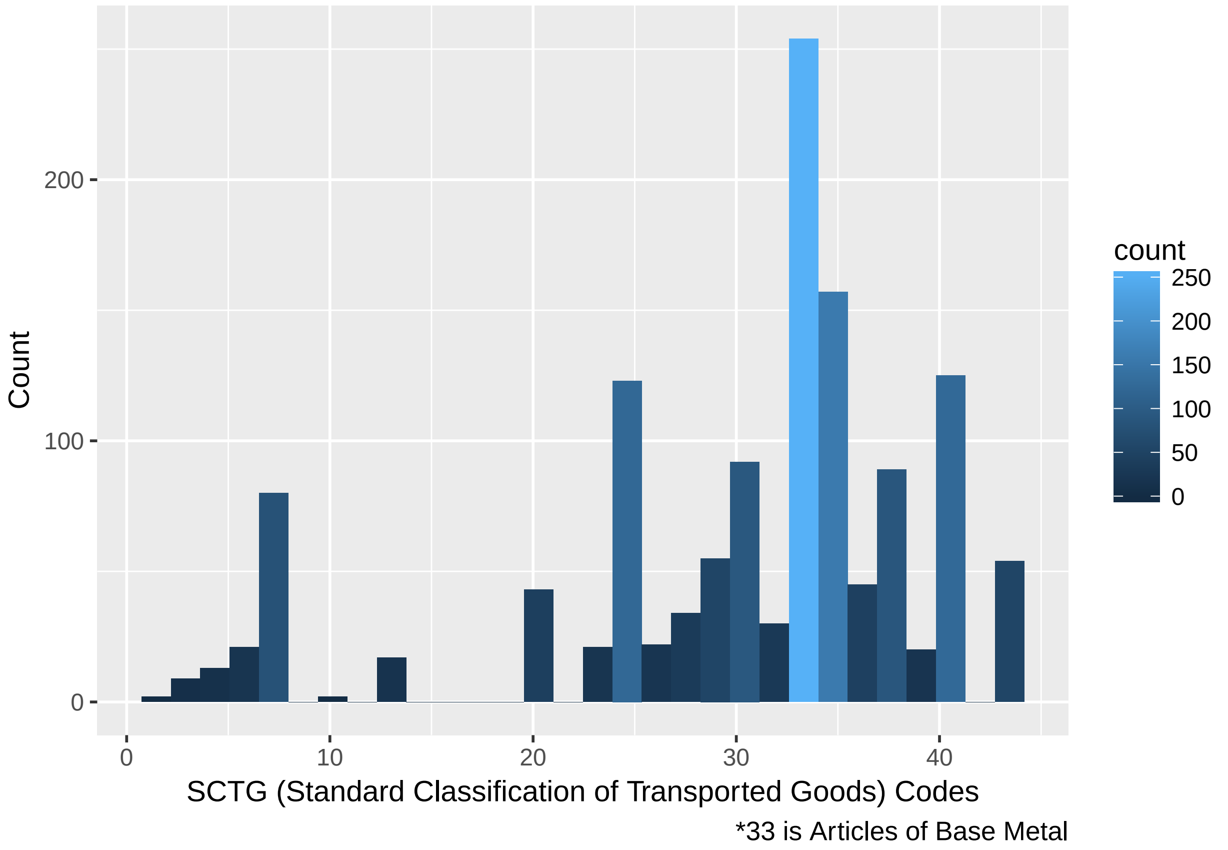


*Graph for Electronic goods industry

```r
# filtering out shipments between UT and OH
utoh <- cfs %>%
  filter((ORIG_STATE == "Utah" & DEST_STATE == "Ohio") |
           (ORIG_STATE == "Ohio" & DEST_STATE == "Utah"))

# determine the commodities most shipped between UT and OH
utoh_maxcom <- utoh %>%
  group_by(SCTG) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

utoh_maxcom
```

```
## # A tibble: 5 × 2
##    SCTG count
##   <dbl> <int>
## 1    35   157
## 2    34   156
## 3    40   123
## 4    24   122
## 5    33    98
```

```r
# 35 is Electronics
# 34 is Machinery
# 40 is Miscellaneous Products
# 24 is Plastics and Rubbers
# 33 is Base Metal Articles
# plot the commodities shipped between UT and OH
ggplot(utoh, aes(x = SCTG)) +
  geom_histogram(aes(fill = ..count..)) +
  labs(x = "SCTG (Standard Classification of Transported Goods) Codes",
       y = "Count",
       caption = "*33 is Articles of Base Metal")
```



```r
# determine the industry most shipped to and from UT and OH
utoh_maxind <- utoh %>%
  group_by(NAICS) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)

utoh_maxind
```

```
## # A tibble: 5 × 2
##   NAICS count
##   <int> <int>
## 1   333   140
## 2   332   116
## 3   311   113
## 4   334    86
## 5  4541    86
```

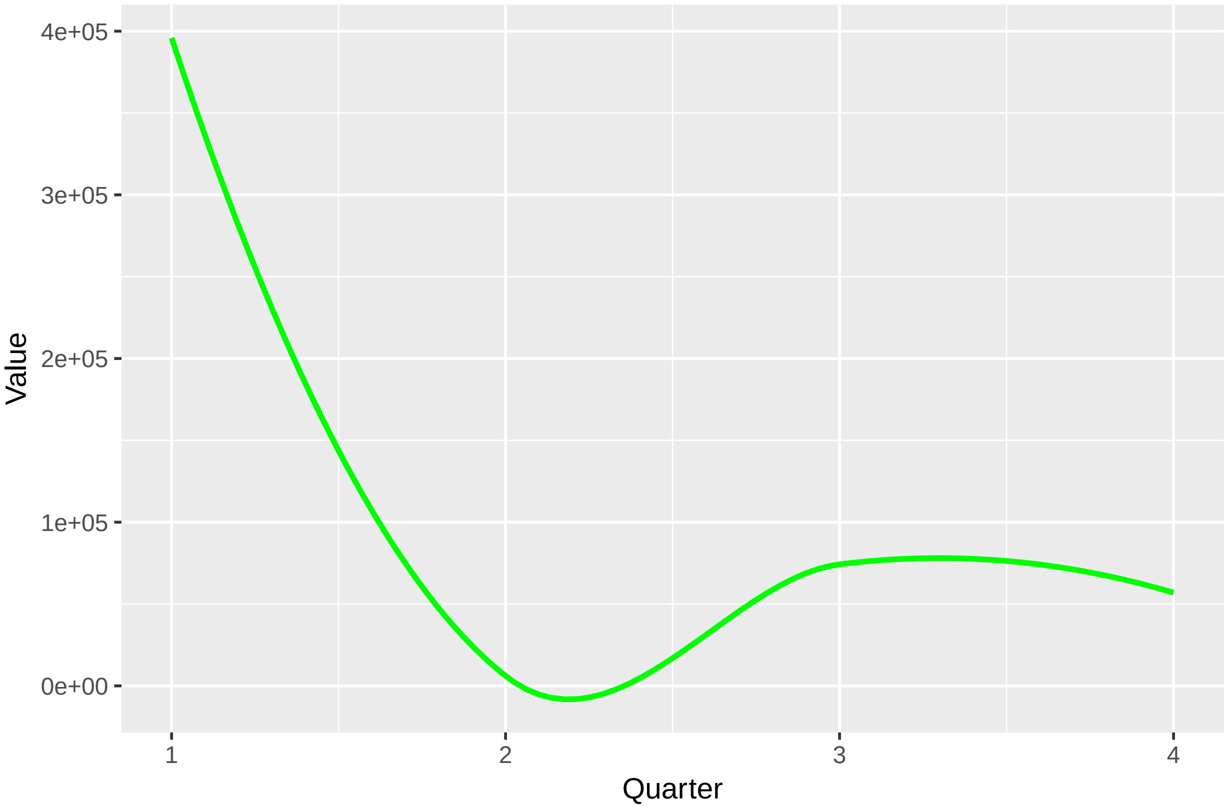
```
# 333 is Machinery
# 332 is Fabricated Metal Industry
# 311 is Food Manufacturing
# 334 is Computers
# 4541 is Electronic Shopping

# avg value of shipments between UT and OH
utoh_avg_value <- utoh %>%
  summarise(avg = mean(SHIPMT_VALUE))

utoh_avg_value
```

```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 58623.74
```

```
# avg weight of shipments between UT and OH
utoh_avg_wght <- utoh %>%
  summarise(avg = mean(SHIPMT_WGHT))

utoh_avg_wght
```

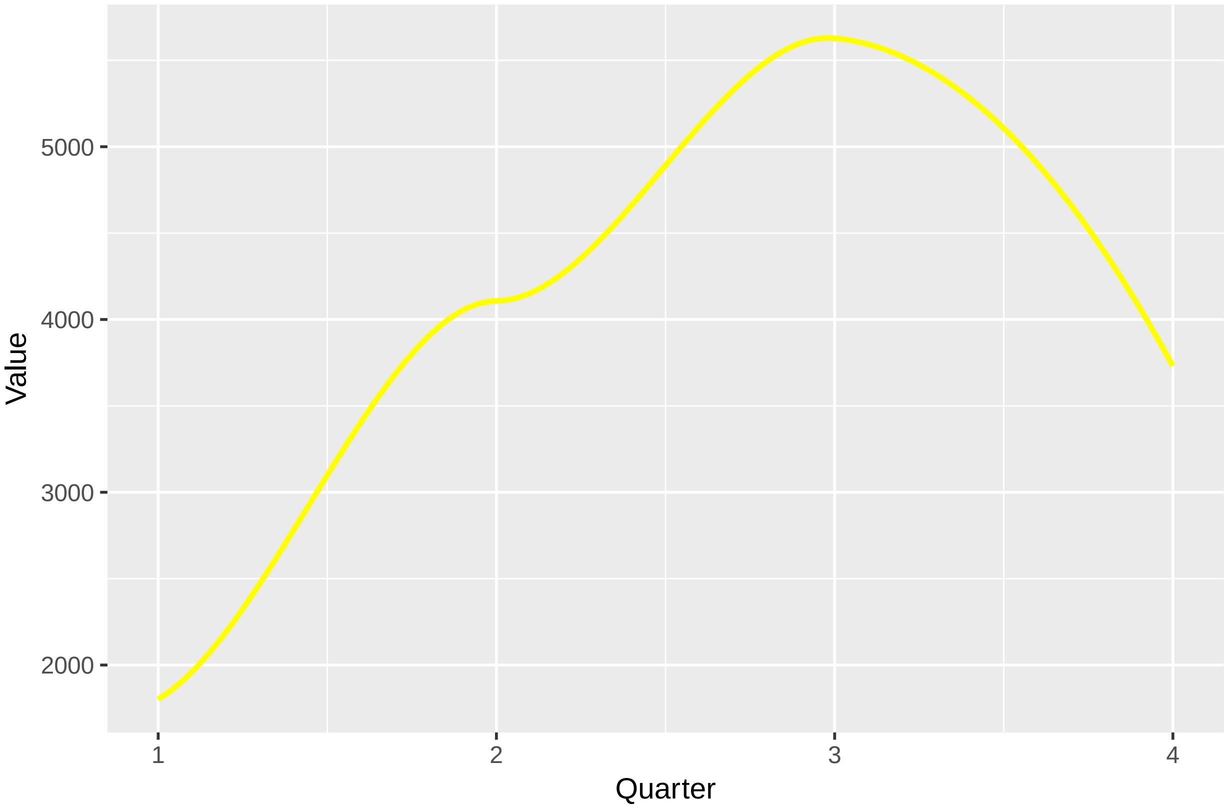
```
## # A tibble: 1 × 1
##        avg
##      <dbl>
## 1 7630.724
```

```
# sort out the most served industry from
# Utah to Ohio
utoh_333 <- utoh %>%
  filter(ORIG_STATE == "Utah" & DEST_STATE == "Ohio",
         NAICS == 333)

# plot the variation of industry served across quarters of 2012
ggplot(utoh_333, aes(x = QUARTER, y = SHIPMT_VALUE)) +
  geom_smooth(color = "green", se = FALSE) +
  labs(x = "Quarter",
       y = "Value",
       caption = "*Graph for Machinery manufacturing industry")
```



*Graph for Machinery manufacturing industry

```r
# sort out the most served industry in other direction i.e.
# from Ohio to Utha
ohut_333 <- utoh %>%
  filter(ORIG_STATE == "Ohio" & DEST_STATE == "Utah",
         NAICS == 333)

# plot the variation of industry served across quarters of 2012
ggplot(ohut_333, aes(x = QUARTER, y = SHIPMT_VALUE)) +
  geom_smooth(color = "yellow", se = FALSE) +
  labs(x = "Quarter",
       y = "Value",
       caption = "*Graph for Machinery manufacturing industry")
```



*Graph for Machinery manufacturing industry