Group 4 Project for CPSC 441

Gemma Clark, Raina Monaghan, Suzanna Storms

Fall 2020

Introduction

For the final project in CPSC 441 Fall 2020, our group has selected the Dairy production and consumption data from the tidytuesday R package. The data originates from the United States Department of Agriculture and appeared in an article from NPR (National Public Radio). The data consist of five csv files including data on (1) cows and their milk production, (2) fluid milk sales, (3) facts about milk products including cheese and butter, (4) consumption of various cheeses, and (5) milk production by state. Each member of the group analyzed a category: milk (Gemma Clark), cheese (Raina Monaghan), and cows (Suzanna Storms).

System Setup

The following R chunk was used to create variables which we would all use and to install packages needed for our analysis.

```
library(tidytuesdayR)
```

```
## Warning: package 'tidytuesdayR' was built under R version 3.6.2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

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

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.2
```

```
library(usmap)
```

```
## Warning: package 'usmap' was built under R version 3.6.2
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:ggpubr':
##
##     mutate
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.6.2
```

```
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp

## For best results, restart R session and update pander using devtools:: or remotes::install_github('ra

##
## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':
##
##     view
```

```
library(ggthemes)
library(nortest)


dairy = tt_load(2019,5)
```

```
## --- Compiling #TidyTuesday Information for 2019-01-29 ----

## --- There are 5 files available ---

## --- Starting Download ---

##
##  Downloading file 1 of 5: `clean_cheese.csv`
##  Downloading file 2 of 5: `fluid_milk_sales.csv`
##  Downloading file 3 of 5: `milk_products_facts.csv`
##  Downloading file 4 of 5: `milkcow_facts.csv`
##  Downloading file 5 of 5: `state_milk_production.csv`

## Only 10 Github queries remaining until 2020-10-15 10:55:17 PM CDT.

## --- Download complete ---
```

```
clean_cheese = dairy$clean_cheese
fluid_milk_sales = dairy$fluid_milk_sales
milk_products_facts = dairy$milk_products_facts
milkcow_facts = dairy$milkcow_facts
state_milk_production = dairy$state_milk_production
cow = dairy$milkcow_facts
```

Milk

In this section, Gemma Clark analyzes milk production by state and by region and milk sales of various milk types.

Milk Production by State and Region

The 'state_milk_production' file contained four columns: region, state, year, and milk produced. The years ranged from 1970 to 2017 and the data included all 50 states. The states were divided into the following ten regions: Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA, DE, MD); Lake States (MI, WI, MN); Corn Belt (OH, IN, IL, IA, MO); Northern Plains (ND, SD, NE, KS); Appalachian (VA, WV, NC, KY, TN); Southeast (SC, GA, FL, AL); Delta States (MS, AR, LA); Southern Plaines (OK, TX); Mountain (MT, ID, WY, CO, NM, AZ, UT, NV); and Pacific (WA, OR, CA, AK, HI).

Summary Statistics and Boxplots of Milk Production by State and by Year

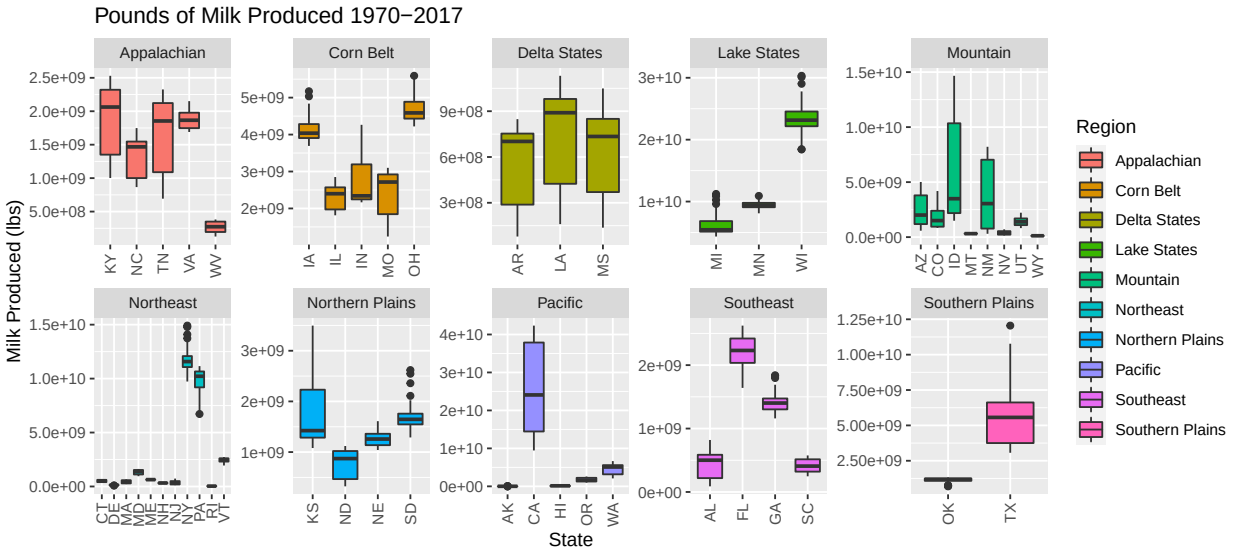
Here, I am calculating some summary statistics and boxplots for milk production by state and by year. These statistics include the mean, median, standard deviation, minimum, maximum, and sample size of each sub-group. This R chunk is used for calculating summary statistics by state.

```
# Preparing summary statistics by state
prep = state_milk_production[,-1]
state_df = pivot_wider(prep, names_from = state, values_from = milk_produced)

state_stats = matrix(data = NA, nrow = (ncol(state_df) - 1), ncol = 6)
cols = c('mean', 'median', 'standard deviation', 'minimum', 'maximum', 'sample size')
colnames(state_stats) = cols
state_stats_rows = c()
j = 1

for (i in 2:(ncol(state_df))) {
  state_stats_rows = append(state_stats_rows, colnames(state_df)[i])
  state_stats_data = as.numeric(as.matrix(state_df[i]))
  state_stats[j, 'mean'] = mean(state_stats_data)
  state_stats[j, 'median'] = median(state_stats_data)
  state_stats[j, 'standard deviation'] = sd(state_stats_data)
  state_stats[j, 'minimum'] = min(state_stats_data)
  state_stats[j, 'maximum'] = max(state_stats_data)
  state_stats[j, 'sample size'] = length(state_stats_data)

  j = j + 1
}

rownames(state_stats) = state_stats_rows

# Summary Statistics Milk Production by State
kable(signif(state_stats, digits = 3), caption = 'Summary Statistics of Milk Production by State')
```

Table 1: Summary Statistics of Milk Production by State

	mean	median	standard deviation	minimum	maximum	sample size
Maine	6.38e+08	6.34e+08	3.66e+07	5.74e+08	7.36e+08	48
New Hampshire	3.21e+08	3.24e+08	2.82e+07	2.72e+08	3.79e+08	48
Vermont	2.43e+09	2.50e+09	2.33e+08	1.94e+09	2.73e+09	48
Massachusetts	4.36e+08	4.52e+08	1.48e+08	2.11e+08	6.58e+08	48
Rhode Island	3.59e+07	3.30e+07	1.70e+07	1.30e+07	7.50e+07	48
Connecticut	5.12e+08	5.25e+08	1.04e+08	3.53e+08	6.61e+08	48
New York	1.17e+10	1.16e+10	1.18e+09	9.73e+09	1.49e+10	48
New Jersey	3.48e+08	3.41e+08	1.72e+08	1.19e+08	7.30e+08	48
Pennsylvania	9.70e+09	1.02e+10	1.34e+09	6.72e+09	1.12e+10	48
Delaware	1.27e+08	1.30e+08	1.84e+07	9.00e+07	1.59e+08	48
Maryland	1.33e+09	1.37e+09	2.25e+08	9.53e+08	1.62e+09	48
Michigan	6.21e+09	5.44e+09	1.78e+09	4.35e+09	1.12e+10	48
Wisconsin	2.34e+10	2.32e+10	2.80e+09	1.84e+10	3.03e+10	48
Minnesota	9.44e+09	9.42e+09	6.64e+08	8.10e+09	1.09e+10	48
Ohio	4.71e+09	4.59e+09	3.89e+08	4.22e+09	5.59e+09	48
Indiana	2.69e+09	2.34e+09	6.38e+08	2.17e+09	4.26e+09	48
Illinois	2.31e+09	2.40e+09	3.45e+08	1.81e+09	2.85e+09	48
Iowa	4.14e+09	4.04e+09	3.36e+08	3.69e+09	5.17e+09	48
Missouri	2.38e+09	2.71e+09	6.37e+08	1.24e+09	3.10e+09	48
North Dakota	7.63e+08	8.72e+08	2.81e+08	3.24e+08	1.12e+09	48
South Dakota	1.69e+09	1.65e+09	2.79e+08	1.29e+09	2.62e+09	48
Nebraska	1.27e+09	1.26e+09	1.54e+08	1.04e+09	1.61e+09	48
Kansas	1.78e+09	1.42e+09	6.68e+08	1.08e+09	3.50e+09	48
Virginia	1.87e+09	1.87e+09	1.29e+08	1.69e+09	2.15e+09	48
West Virginia	2.70e+08	2.74e+08	8.10e+07	1.27e+08	3.82e+08	48
North Carolina	1.32e+09	1.47e+09	2.89e+08	8.66e+08	1.75e+09	48
Kentucky	1.86e+09	2.06e+09	5.14e+08	1.00e+09	2.53e+09	48
Tennessee	1.63e+09	1.85e+09	5.63e+08	6.93e+08	2.33e+09	48
South Carolina	4.13e+08	4.08e+08	1.06e+08	2.47e+08	5.76e+08	48
Georgia	1.41e+09	1.40e+09	1.61e+08	1.16e+09	1.84e+09	48
Florida	2.22e+09	2.23e+09	2.45e+08	1.64e+09	2.62e+09	48
Alabama	4.33e+08	5.01e+08	2.22e+08	8.90e+07	8.20e+08	48
Mississippi	6.19e+08	7.34e+08	2.79e+08	1.37e+08	1.05e+09	48
Arkansas	5.43e+08	7.02e+08	2.70e+08	7.90e+07	8.48e+08	48
Louisiana	7.33e+08	8.90e+08	3.28e+08	1.60e+08	1.13e+09	48
Oklahoma	1.13e+09	1.18e+09	1.71e+08	6.92e+08	1.32e+09	48
Texas	5.82e+09	5.56e+09	2.41e+09	3.06e+09	1.21e+10	48
Montana	3.19e+08	3.14e+08	2.36e+07	2.75e+08	3.72e+08	48
Idaho	6.04e+09	3.49e+09	4.70e+09	1.49e+09	1.47e+10	48
Wyoming	1.13e+08	1.23e+08	2.47e+07	5.40e+07	1.40e+08	48
Colorado	1.80e+09	1.50e+09	9.66e+08	8.35e+08	4.19e+09	48
New Mexico	3.70e+09	3.05e+09	3.10e+09	3.04e+08	8.21e+09	48
Arizona	2.44e+09	2.00e+09	1.45e+09	5.85e+08	5.01e+09	48
Utah	1.43e+09	1.39e+09	4.27e+08	8.19e+08	2.22e+09	48
Nevada	3.95e+08	3.83e+08	1.86e+08	1.41e+08	7.09e+08	48
Washington	4.52e+09	5.09e+09	1.45e+09	2.09e+09	6.65e+09	48
Oregon	1.71e+09	1.63e+09	5.34e+08	9.70e+08	2.59e+09	48
California	2.54e+10	2.41e+10	1.17e+10	9.46e+09	4.23e+10	48
Alaska	1.41e+07	1.40e+07	6.86e+06	3.00e+06	3.50e+07	48
Hawaii	1.09e+08	1.36e+08	4.91e+07	1.90e+07	1.60e+08	48

```r
# Summary Boxplots Milk Production by State

# Add column with state abbreviations for labeling
milk_plot_data = state_milk_production
for (i in 1:nrow(milk_plot_data)) {
  milk_plot_data[i,'state_abb'] = state.abb[which(state.name == milk_plot_data$state[i])]
}

ggplot(milk_plot_data, aes(x = state_abb, y = milk_produced, fill = region)) +
  geom_boxplot() +
  labs(x = 'State', y = 'Milk Produced (lbs)', title = 'Pounds of Milk Produced 1970-2017', fill = 'Reg
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  facet_wrap(~region, scales='free', nrow = 2, ncol = 5)
```



This R chunk is used for calculating summary statistics by year.

```r
# Preparing summary statistics by year
year_df = pivot_wider(state_milk_production, names_from = year, values_from = milk_produced)

year_stats = matrix(data = NA, nrow = (ncol(year_df) - 2), ncol = 6)
cols = c('mean', 'median', 'standard deviation', 'minimum', 'maximum', 'sample size')
colnames(year_stats) = cols
year_stats_rows = c()
j = 1

for (i in 3:(ncol(year_df))) {
  year_stats_rows = append(year_stats_rows, colnames(year_df)[i])
  year_stats_data = as.numeric(as.matrix(year_df[i]))
  year_stats[j, 'mean'] = mean(year_stats_data)
  year_stats[j, 'median'] = median(year_stats_data)
  year_stats[j, 'standard deviation'] = sd(year_stats_data)
  year_stats[j, 'minimum'] = min(year_stats_data)
  year_stats[j, 'maximum'] = max(year_stats_data)
  year_stats[j, 'sample size'] = length(year_stats_data)
```

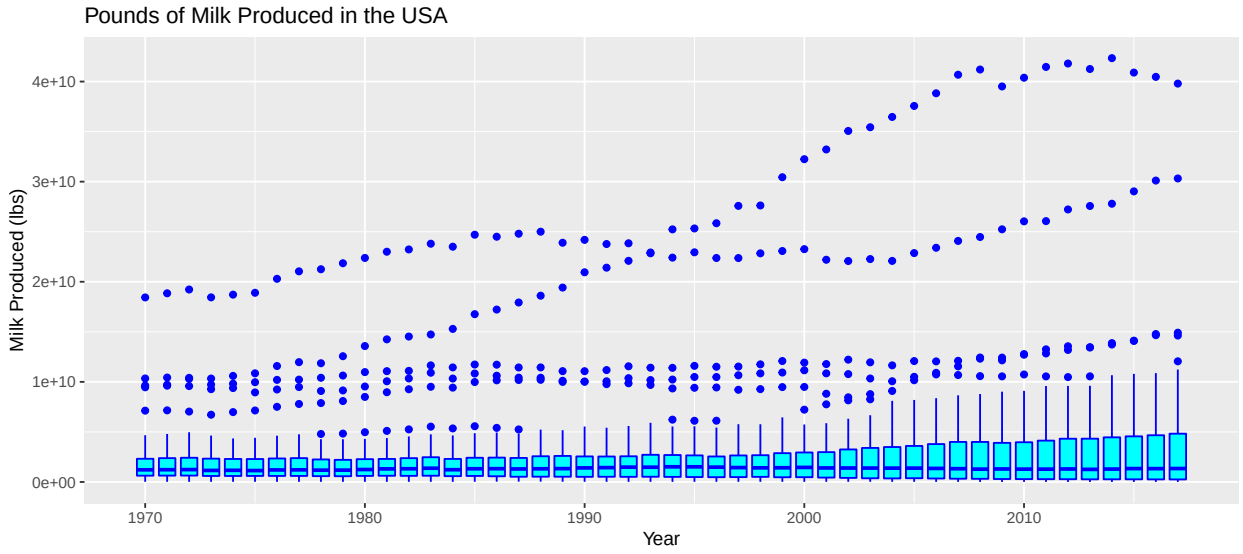
```
  j = j + 1
}

rownames(year_stats) = year_stats_rows

# Summary Statistics Milk Production by Year
kable(signif(year_stats, digits = 3), caption = 'Summary Statistics of Milk Production by Year')
```

Table 2: Summary Statistics of Milk Production by Year

	mean	median	standard deviation	minimum	maximum	sample size
1970	2.34e+09	1.22e+09	3.36e+09	1.9e+07	1.84e+10	50
1971	2.37e+09	1.22e+09	3.42e+09	1.7e+07	1.88e+10	50
1972	2.40e+09	1.24e+09	3.48e+09	1.8e+07	1.92e+10	50
1973	2.31e+09	1.15e+09	3.34e+09	1.8e+07	1.84e+10	50
1974	2.31e+09	1.16e+09	3.39e+09	1.8e+07	1.87e+10	50
1975	2.31e+09	1.14e+09	3.42e+09	1.7e+07	1.89e+10	50
1976	2.40e+09	1.20e+09	3.63e+09	1.6e+07	2.03e+10	50
1977	2.45e+09	1.21e+09	3.74e+09	1.6e+07	2.10e+10	50
1978	2.43e+09	1.18e+09	3.76e+09	1.5e+07	2.13e+10	50
1979	2.47e+09	1.18e+09	3.87e+09	1.3e+07	2.18e+10	50
1980	2.57e+09	1.24e+09	4.02e+09	1.3e+07	2.24e+10	50
1981	2.66e+09	1.31e+09	4.15e+09	1.3e+07	2.30e+10	50
1982	2.71e+09	1.32e+09	4.21e+09	1.4e+07	2.32e+10	50
1983	2.79e+09	1.37e+09	4.33e+09	1.4e+07	2.38e+10	50
1984	2.71e+09	1.22e+09	4.31e+09	1.4e+07	2.35e+10	50
1985	2.86e+09	1.32e+09	4.56e+09	2.2e+07	2.47e+10	50
1986	2.86e+09	1.33e+09	4.56e+09	3.2e+07	2.45e+10	50
1987	2.85e+09	1.30e+09	4.63e+09	3.5e+07	2.48e+10	50
1988	2.90e+09	1.32e+09	4.70e+09	3.1e+07	2.50e+10	50
1989	2.88e+09	1.33e+09	4.62e+09	2.3e+07	2.39e+10	50
1990	2.95e+09	1.41e+09	4.77e+09	1.7e+07	2.42e+10	50
1991	2.95e+09	1.43e+09	4.76e+09	1.3e+07	2.38e+10	50
1992	3.02e+09	1.48e+09	4.85e+09	1.2e+07	2.38e+10	50
1993	3.01e+09	1.48e+09	4.82e+09	1.2e+07	2.29e+10	50
1994	3.07e+09	1.51e+09	4.99e+09	1.3e+07	2.52e+10	50
1995	3.11e+09	1.51e+09	5.06e+09	1.2e+07	2.53e+10	50
1996	3.08e+09	1.48e+09	5.07e+09	1.4e+07	2.58e+10	50
1997	3.12e+09	1.44e+09	5.23e+09	1.5e+07	2.76e+10	50
1998	3.15e+09	1.41e+09	5.29e+09	1.4e+07	2.76e+10	50
1999	3.25e+09	1.42e+09	5.61e+09	1.4e+07	3.04e+10	50
2000	3.35e+09	1.45e+09	5.82e+09	1.3e+07	3.22e+10	50
2001	3.31e+09	1.40e+09	5.84e+09	1.4e+07	3.32e+10	50
2002	3.40e+09	1.39e+09	6.04e+09	1.8e+07	3.51e+10	50
2003	3.41e+09	1.39e+09	6.10e+09	1.7e+07	3.54e+10	50
2004	3.42e+09	1.38e+09	6.20e+09	1.5e+07	3.65e+10	50
2005	3.54e+09	1.38e+09	6.42e+09	1.4e+07	3.76e+10	50
2006	3.64e+09	1.35e+09	6.63e+09	1.0e+07	3.88e+10	50
2007	3.71e+09	1.32e+09	6.91e+09	9.0e+06	4.07e+10	50
2008	3.80e+09	1.28e+09	7.04e+09	7.0e+06	4.12e+10	50
2009	3.78e+09	1.30e+09	6.91e+09	6.0e+06	3.95e+10	50
2010	3.86e+09	1.28e+09	7.09e+09	7.0e+06	4.04e+10	50
2011	3.93e+09	1.28e+09	7.23e+09	7.0e+06	4.15e+10	50

	mean	median	standard deviation	minimum	maximum	sample size
2012	4.01e+09	1.30e+09	7.36e+09	6.0e+06	4.18e+10	50
2013	4.02e+09	1.26e+09	7.34e+09	3.0e+06	4.13e+10	50
2014	4.12e+09	1.29e+09	7.50e+09	4.0e+06	4.23e+10	50
2015	4.17e+09	1.34e+09	7.47e+09	4.0e+06	4.09e+10	50
2016	4.25e+09	1.33e+09	7.56e+09	4.0e+06	4.05e+10	50
2017	4.31e+09	1.34e+09	7.55e+09	3.0e+06	3.98e+10	50

```
# Summary Boxplots Milk Production by Year
ggplot(state_milk_production, aes(year, milk_produced, group = year)) +
  geom_boxplot(color = 'blue', fill = 'cyan') +
  labs(x = 'Year', y = 'Milk Produced (lbs)', title = 'Pounds of Milk Produced in the USA')
```

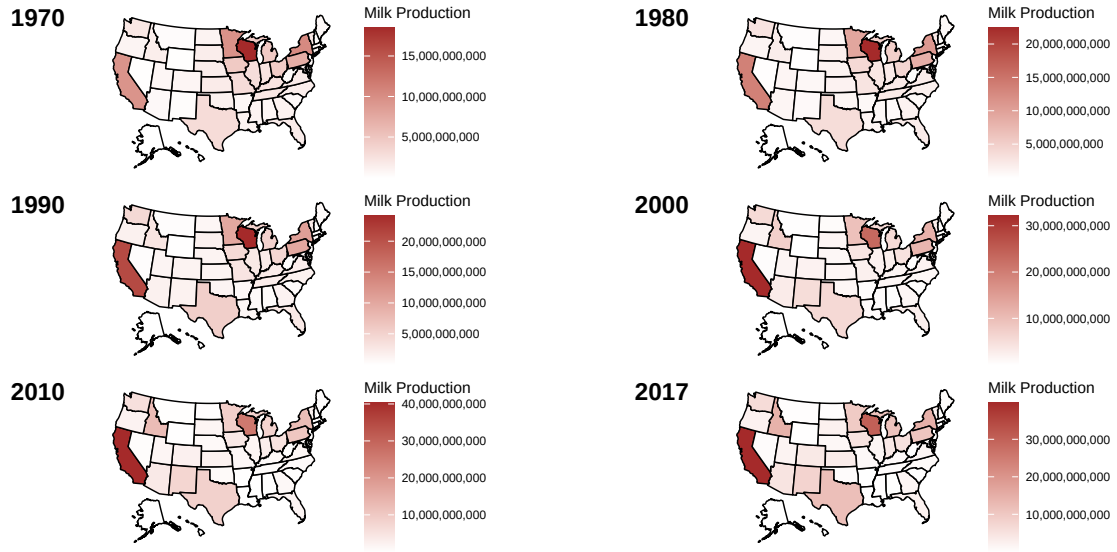


Visual Representation of State Milk Production Across the Decades

In this section, I am providing several choropleth maps of the United States showing milk production in each state from 1970 to 2017. As shown in the maps, California and Wisconsin tended to produce more pounds of milk than other states.

```
# Organize states data so that year has columns with milk production
states_by_year = pivot_wider(state_milk_production, names_from = year, values_from = milk_produced)


# Create choropleth figures for milk production at each decade
fig70 = plot_usmap(data = states_by_year, values = '1970', color = 'black' ) +
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')

fig80 = plot_usmap(data = states_by_year, values = '1980', color = 'black' ) +
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')

fig90 = plot_usmap(data = states_by_year, values = '1990', color = 'black' ) +
```

```
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')

fig00 = plot_usmap(data = states_by_year, values = '2000', color = 'black' ) +
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')

fig10 = plot_usmap(data = states_by_year, values = '2010', color = 'black' ) +
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')

fig17 = plot_usmap(data = states_by_year, values = '2017', color = 'black' ) +
  scale_fill_continuous(low = 'white', high = 'brown', name = 'Milk Production', label = scales::comma) +
  theme(legend.position = 'right')


# Plot the choropleths together
ggarrange(fig70, fig80, fig90, fig00, fig10, fig17,
          labels = c('1970', '1980', '1990', '2000', '2010', '2017'),
          ncol = 2, nrow = 3)
```

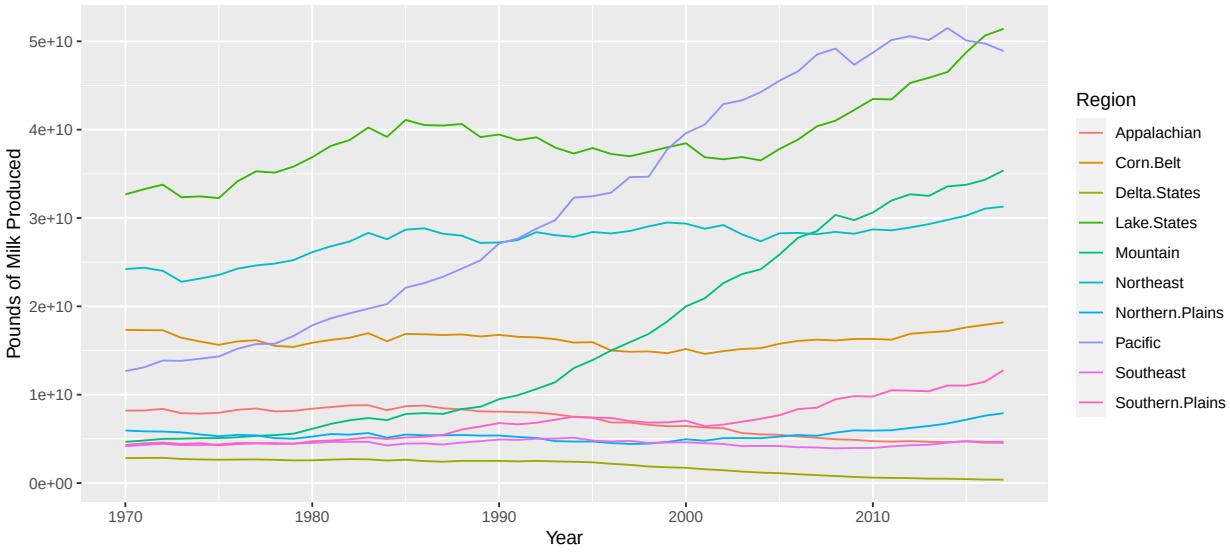


Visual Representation of Milk Production by Region over Time

In this section, I am providing a graph containing the total pounds of milk produced in each region over the 47-year period of available data. Note that some regions had more states than other regions which could have affected the graph in this section. As shown below, the Lake States (MI, WI, MN) were consistently high producers of milk, and the production of milk in the Pacific states (WA, OR, CA, AK, HI) grew to become the highest regional producers of milk over time.

```
interim = data.frame(year = state_milk_production$year,
                     state = state_milk_production$state,
                     milk_produced = state_milk_production$milk_produced)

states_by_states = pivot_wider(interim, names_from = state, values_from = milk_produced)
```

```r
summed_regions = states_by_states %>%
  mutate(`Northeast` = `Maine` + `New Hampshire` + `Vermont` + `Massachusetts` + `Rhode Island` + `Conne
  mutate(`Lake States` = `Michigan` + `Wisconsin` + `Minnesota`) %>%
  mutate(`Corn Belt` = `Ohio` + `Indiana` + `Illinois` + `Iowa` + `Missouri`) %>%
  mutate(`Northern Plains` = `North Dakota` + `South Dakota` + `Nebraska` + `Kansas`) %>%
  mutate(`Appalachian` = `Virginia` + `West Virginia` + `North Carolina` + `Kentucky` + `Tennessee`) %>%
  mutate(`Southeast` = `South Carolina` + `Georgia` + `Florida` + `Alabama`) %>%
  mutate(`Delta States` = `Mississippi` + `Arkansas` + `Louisiana`) %>%
  mutate(`Southern Plains` = `Oklahoma` + `Texas`) %>%
  mutate(`Mountain` = `Montana` + `Idaho` + `Wyoming` + `Colorado` + `New Mexico` + `Arizona` + `Utah` +
  mutate(`Pacific` = `Washington` + `Oregon` + `California` + `Alaska` + `Hawaii`)



inter_total = summed_regions %>%
  mutate(`Total Milk` = `Northeast` + `Lake States` + `Corn Belt` + `Northern Plains` + `Appalachian` +


milk_total = data.frame(year = inter_total$year,
                        total_milk = inter_total$`Total Milk`)



regional_milk = data.frame(year = summed_regions$year,
                           `Northeast` = summed_regions$Northeast,
                           `Lake States` = summed_regions$`Lake States`,
                           `Corn Belt` = summed_regions$`Corn Belt`,
                           `Northern Plains` = summed_regions$`Northern Plains`,
                           `Appalachian` = summed_regions$Appalachian,
                           `Southeast` = summed_regions$Southeast,
                           `Delta States` = summed_regions$`Delta States`,
                           `Southern Plains` = summed_regions$`Southern Plains`,
                           `Mountain` = summed_regions$Mountain,
                           `Pacific` = summed_regions$Pacific)


reg_milk = pivot_longer(regional_milk,
                        cols = c(Northeast, Lake.States, Corn.Belt, Northern.Plains, Appalachian, South
                        names_to = 'region',
                        values_to = 'milk_production')

ggplot(reg_milk, aes(x = year, y = milk_production, colour = region)) +
  geom_line() +
  labs(x = 'Year', y = 'Pounds of Milk Produced', colour = 'Region')
```



Statistical Analyses of Milk Production

Here, I am performing a two-way ANOVA where the independent variables are state and year, and the dependent variable is pounds of milk produced. As shown by the p-values below, the differences between some of the means are statistically significant. However, the residuals of the ANOVA model are not normally distributed (as shown by the Shapiro-Wilk test) which may have had some impact on the ANOVA outcome. Further analysis using non-parametric tests would be recommended for this dataset.

```
# Create ANOVA model and present the results

# ANOVA using state and year as independent variables
aov.milk = aov(milk_produced ~ state + year, data = state_milk_production)
summary(aov.milk)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## state         49 6.121e+22 1.249e+21   344.1 <2e-16 ***
## year           1 8.027e+20 8.027e+20   221.1 <2e-16 ***
## Residuals   2349 8.528e+21 3.630e+18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
milk.residuals = aov.milk$residuals
shapiro.test(milk.residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  milk.residuals
## W = 0.60048, p-value < 2.2e-16
```

Fluid Milk Sales

The 'fluid_milk_sales' file contained three columns: year, milk type, and pounds of milk product sold per year. The years ranged from 1975 to 2017 and the milk types included whole milk, reduced fat (2%) milk,

low fat (1%) milk, skim milk, flavored whole milk, flavored (not whole) milk, buttermilk, and eggnog. There was an additional column that included the total milk product.

Summary Statistics of Fluid Milk Sales

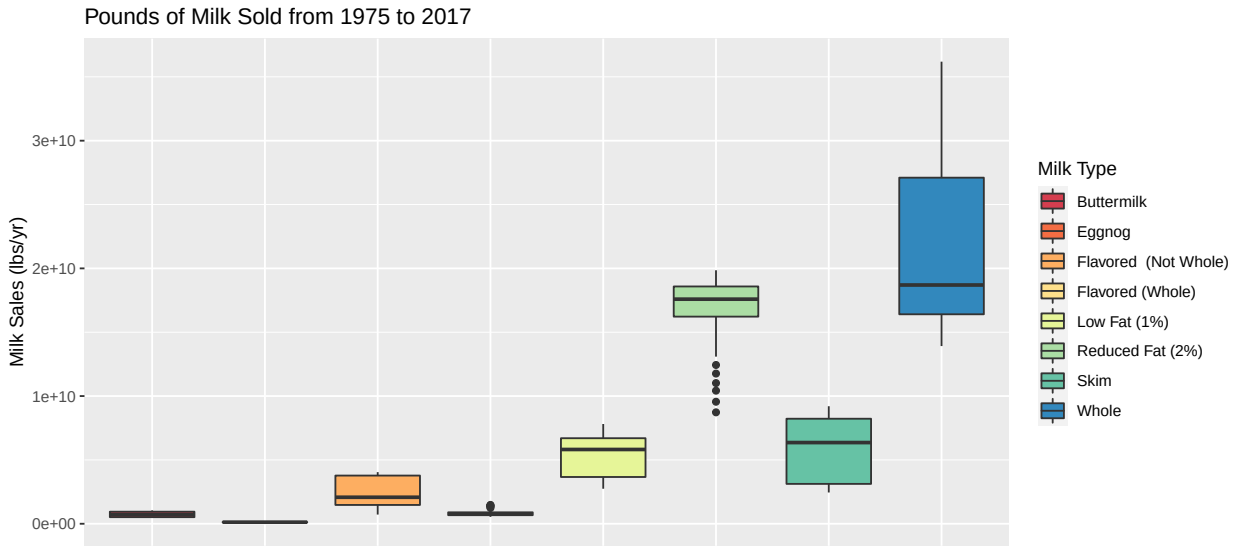
```
# Preparing summary statistics by fluid milk type
fluid_df = pivot_wider(fluid_milk_sales, names_from = milk_type, values_from = pounds)
fluid_df = fluid_df[-c(10)] # gets rid of total production column

fluid_stats = matrix(data = NA, nrow = (ncol(fluid_df) - 1), ncol = 6)
cols = c('mean', 'median', 'standard deviation', 'minimum', 'maximum', 'sample size')
colnames(fluid_stats) = cols
fluid_stats_rows = c()
j = 1


for (i in 2:(ncol(fluid_df))) {
  fluid_stats_rows = append(fluid_stats_rows, colnames(fluid_df)[i])
  fluid_stats_data = as.numeric(as.matrix(fluid_df[i]))
  fluid_stats[j, 'mean'] = mean(fluid_stats_data)
  fluid_stats[j, 'median'] = median(fluid_stats_data)
  fluid_stats[j, 'standard deviation'] = sd(fluid_stats_data)
  fluid_stats[j, 'minimum'] = min(fluid_stats_data)
  fluid_stats[j, 'maximum'] = max(fluid_stats_data)
  fluid_stats[j, 'sample size'] = length(fluid_stats_data)

  j = j + 1
}

rownames(fluid_stats) = fluid_stats_rows


# Summary Statistics of Milk Sales by Milk Type
kable(signif(fluid_stats, digits = 3), caption = 'Summary Statistics of Milk Sales by Milk Type')
```

Table 3: Summary Statistics of Milk Sales by Milk Type

	mean	median	standard deviation	minimum	maximum	sample size
Whole	2.16e+10	1.87e+10	6.71e+09	1.39e+10	3.62e+10	43
Reduced Fat (2%)	1.67e+10	1.76e+10	2.94e+09	8.73e+09	1.99e+10	43
Low Fat (1%)	5.43e+09	5.82e+09	1.63e+09	2.74e+09	7.81e+09	43
Skim	5.95e+09	6.36e+09	2.48e+09	2.45e+09	9.20e+09	43
Flavored (Whole)	8.22e+08	7.49e+08	2.45e+08	5.39e+08	1.48e+09	43
Flavored (Not Whole)	2.43e+09	2.08e+09	1.13e+09	7.19e+08	4.04e+09	43
Buttermilk	7.36e+08	7.11e+08	2.15e+08	4.53e+08	1.05e+09	43
Eggnog	1.21e+08	1.23e+08	1.96e+07	7.60e+07	1.53e+08	43

```
# Summary Boxplots Milk Sales by Milk Type
milk_type_graph_data = subset(fluid_milk_sales, milk_type != 'Total Production')
  # Get rid of Total Production data
```

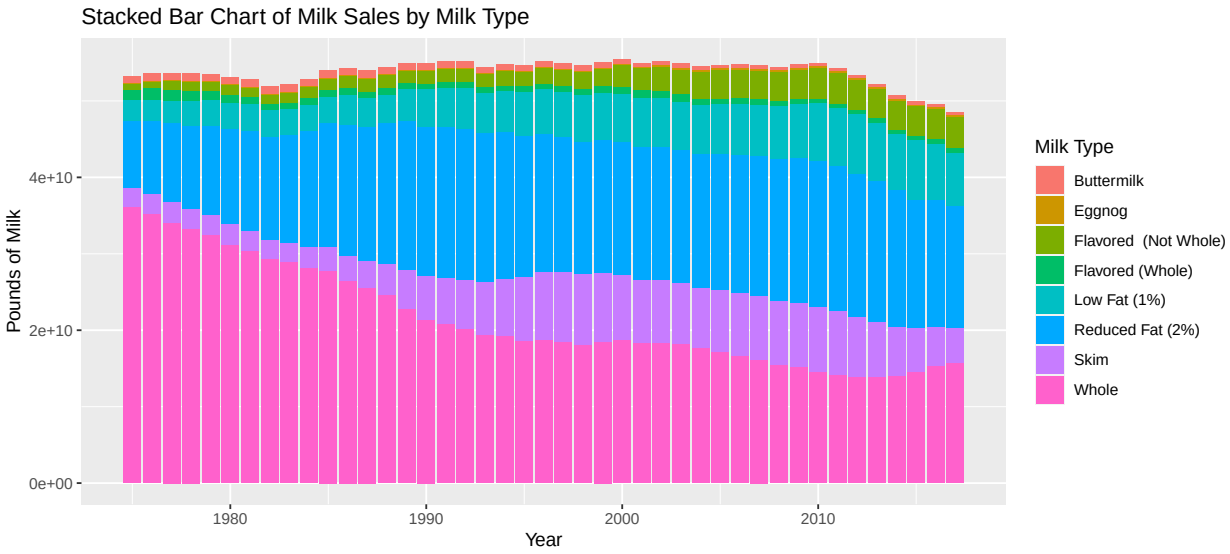
```
ggplot(milk_type_graph_data, aes(milk_type, pounds, group = milk_type, fill = milk_type)) +
  geom_boxplot() +
  labs(y = 'Milk Sales (lbs/yr)', title = 'Pounds of Milk Sold from 1975 to 2017', fill = 'Milk Type') +
  theme(axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank()) +
  scale_fill_brewer(palette = 'Spectral')
```



Visual Representation of Fluid Milk Sales

In this section, I am using a stacked bar chart to show the proportions of each type of milk that were produced from 1975 to 2017.

```
graph_data = filter(fluid_milk_sales, milk_type != 'Total Production')


ggplot(graph_data, aes(x = year, y = pounds, fill = milk_type)) +
  geom_bar(position = 'stack', stat = 'identity') +
  labs(title = 'Stacked Bar Chart of Milk Sales by Milk Type', x = 'Year', y = 'Pounds of Milk', fill =
```



Statiscal Analyses of Fluid Milk Sales

Here, I am performing a two-way ANOVA where the independent variables are milk type and year, and the dependent variable is pounds of milk sold per year. As shown by the p-values below, milk sales among milk type are statistically different, but milk sales across the years are not. However, the residuals of the ANOVA model are not normally distributed (as shown by the Shapiro-Wilk test) which may have had some impact on the ANOVA outcome. Further analysis using non-parametric tests would be recommended for this dataset.

```
milk_type_stats_data = subset(fluid_milk_sales, milk_type != 'Total Production')
  # Get rid of total proudction

# ANOVA using milk type and year as independent variable
aov.milk_type = aov(pounds ~ milk_type + year, data = milk_type_stats_data)
summary(aov.milk_type)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## milk_type     7 1.961e+22 2.801e+21 349.702 <2e-16 ***
## year          1 6.203e+17 6.203e+17   0.077  0.781
## Residuals   335 2.683e+21 8.010e+18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
milk_type.residuals = aov.milk_type$residuals
shapiro.test(milk_type.residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  milk_type.residuals
## W = 0.83961, p-value < 2.2e-16
```

Milk Data Conclusions

As shown in this document, milk trends varied by state over time and whole milk was the most commonly type of milk sold per year. In both of the ANOVA models created for milk production by state and milk sales by type, the residuals were not normally distributed. Additional non-paremetric tests could be performed in future such as Wilcoxon signed rank tests to compare two samples or sign tests to compare the data to an expected median value. The following chunk of code saves the means from the variables produced in the milk section of the report and saves them as a text file.

```
write.table(state_stats[1:50,1], file = 'mean_milk_production_by_state.txt')
write.table(year_stats[1:48, 1], file = 'mean_milk_proudction_by_year.txt')
write.table(fluid_stats[1:8, 1], file = 'mean_milk_sales_by_type.txt')
```

Cheese

In this section, Raina Monaghan analyzes cheese consumption trends in the United States.

The "clean_cheese" Dataset

The "clean_cheese" dataset contains 17 columns in total, the first dedicated to years spanning from 1970 to 2017, with 11 allocated to various cheese types. Cheese types include: Cheddar, American Other, Mozzarella, Italian other, Swiss, Brick, Muenster, Cream and Neufchatel, Blue, Other Dairy Cheese, Processed Cheese, Foods and Spreads. Four of the 17 columns are dedicated to cumulative values: Total American Cheese, Total Italian Chese, Total Natural Cheese, Total Processing Cheese Products. Data is gathered in the dataset as a measure of cheeses consumed, in pounds, per person.

```
clean_cheese <- as.data.frame(clean_cheese)
```

Summary statistics for cheese types are presented in this portion. They include the: mean, standard deviation, minimum and maximum, Q1 and Q3, median, MAD and IQR.

```
# Selecting only the cheese types in preparation for summary statistics
all_cheese <- select(clean_cheese, "Cheddar":"Foods and spreads")
summary_whole <- summarytools::descr(all_cheese)
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
summary_whole
```

```
## Descriptive Statistics
## all_cheese
## N: 48
##
##                     American Other    Blue    Brick   Cheddar   Cream and Neufchatel
## ----------------- ---------------- ------- -------- --------- ----------------------
##              Mean             2.62    0.20     0.05      8.89                   1.74
##           Std.Dev             0.65    0.06     0.03      1.52                   0.69
##               Min             1.20    0.15     0.01      5.79                   0.61
##                Q1             2.20    0.16     0.03      7.88                   1.08
##            Median             2.58    0.17     0.05      9.52                   2.05
##                Q3             3.00    0.18     0.08      9.95                   2.36
##               Max             3.99    0.32     0.12     11.07                   2.64
##               MAD             0.57    0.01     0.03      0.77                   0.63
##               IQR             0.75    0.02     0.04      1.62                   1.25
##                CV             0.25    0.31     0.58      0.17                   0.40
##          Skewness             0.30    1.22     0.44     -0.88                  -0.39
##       SE.Skewness             0.34    0.39     0.34      0.34                   0.34
##          Kurtosis            -0.33   -0.38    -0.99     -0.71                  -1.49
##           N.Valid            48.00   36.00    48.00     48.00                  48.00
##         Pct.Valid           100.00   75.00   100.00    100.00                 100.00
##
## Table: Table continues below
##
##
##
##                     Foods and spreads   Italian other   Mozzarella   Muenster   Other Dairy Cheese
## ----------------- ------------------ --------------- ------------ ---------- --------------------
##              Mean                3.16            2.15         6.86       0.34                 1.02
##           Std.Dev                0.44            0.73         3.43       0.09                 0.36
##               Min                1.91            0.87         1.19       0.17                 0.41
##                Q1                2.99            1.50         3.12       0.28                 0.73
##            Median                3.22            2.19         7.70       0.34                 0.97
##                Q3                3.44            2.77        10.00       0.40                 1.30
##               Max                3.98            3.49        11.73       0.53                 1.59
##               MAD                0.33            0.95         4.34       0.10                 0.47
##               IQR                0.45            1.23         6.78       0.12                 0.54
##                CV                0.14            0.34         0.50       0.26                 0.36
##          Skewness               -0.80           -0.04        -0.27       0.27                 0.02
##       SE.Skewness                0.34            0.34         0.34       0.34                 0.34
##          Kurtosis                0.56           -1.08        -1.45      -0.81                -1.35
##           N.Valid               48.00           48.00        48.00      48.00                48.00
##         Pct.Valid              100.00          100.00       100.00     100.00               100.00
##
## Table: Table continues below
##
##
##
##                     Processed Cheese    Swiss
## ----------------- ------------------ --------
##              Mean                4.33     1.16
```

```
##           Std.Dev               0.63     0.11
##               Min               3.31     0.88
##                Q1               3.80     1.06
##            Median               4.42     1.17
##                Q3               4.82     1.24
##               Max               5.44     1.35
##               MAD               0.79     0.12
##               IQR               0.99     0.17
##                CV               0.15     0.10
##          Skewness               0.04    -0.21
##       SE.Skewness               0.34     0.34
##          Kurtosis              -1.25    -0.76
##           N.Valid              48.00    48.00
##         Pct.Valid             100.00   100.00
```

Summary statistics are presented in this portion for "cumulative" type columns.

```
#Subsetting the variables of interest
sel_vars_total <-
  select(clean_cheese,
         "Total American Chese":"Total Processed Cheese Products")

sum_1 <- summarytools::descr(sel_vars_total)
sum_1
```

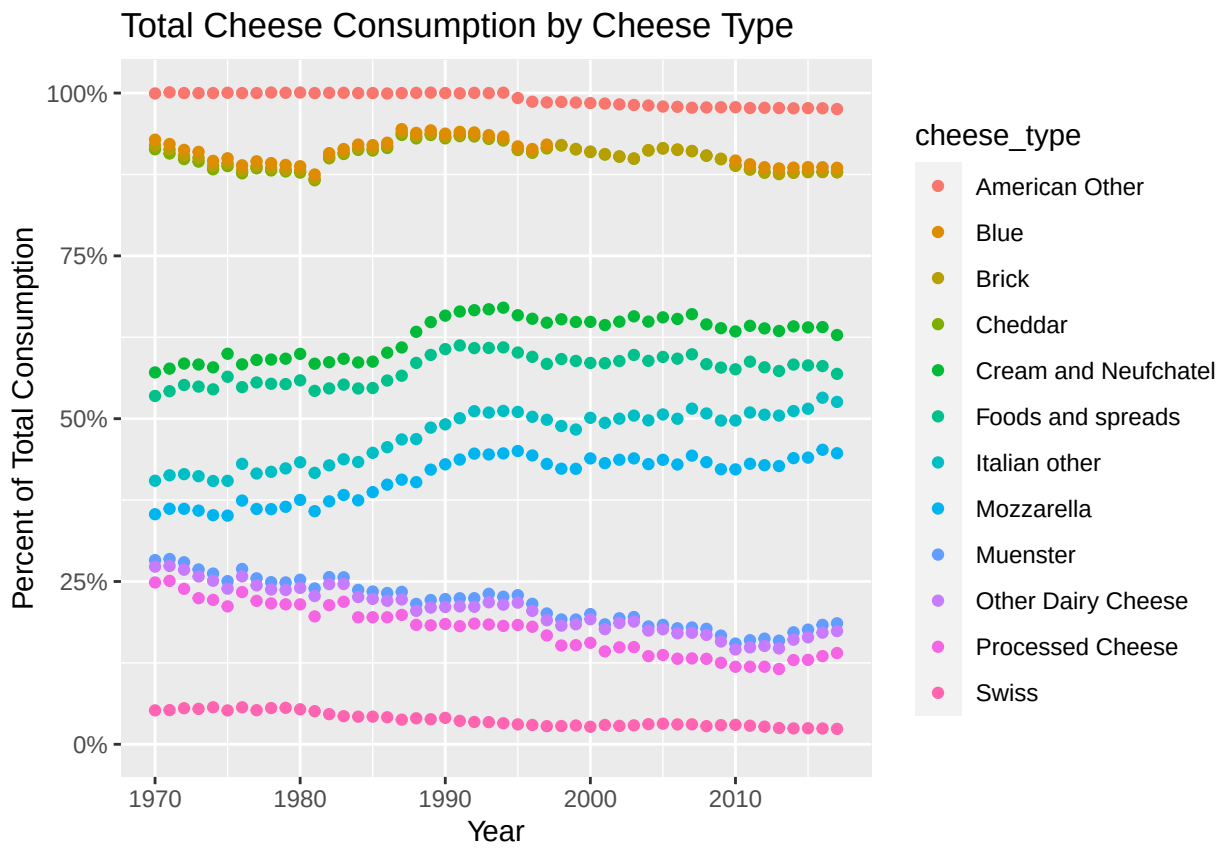
```
## Descriptive Statistics
## sel_vars_total
## N: 48
##
##                     Total American Chese   Total Italian Cheese   Total Natural Cheese
## ----------------- ---------------------- ---------------------- ----------------------
##              Mean                  11.51                   9.01                  25.35
##           Std.Dev                   1.95                   4.15                   7.37
##               Min                   7.00                   2.05                  11.37
##                Q1                  10.59                   4.60                  19.04
##            Median                  11.83                   9.95                  26.29
##                Q3                  12.84                  12.80                  31.98
##               Max                  15.06                  15.21                  37.23
##               MAD                   1.50                   5.29                   9.07
##               IQR                   2.03                   8.04                  12.31
##                CV                   0.17                   0.46                   0.29
##          Skewness                  -0.69                  -0.24                  -0.31
##       SE.Skewness                   0.34                   0.34                   0.34
##          Kurtosis                  -0.39                  -1.40                  -1.14
##           N.Valid                  48.00                  48.00                  48.00
##         Pct.Valid                 100.00                 100.00                 100.00
##
## Table: Table continues below
##
##
##
##                     Total Processed Cheese Products
```

```
## ----------------- ----------------------------------
##              Mean                              7.49
##           Std.Dev                              0.87
##               Min                              5.53
##                Q1                              6.86
##            Median                              7.59
##                Q3                              8.22
##               Max                              8.75
##               MAD                              0.99
##               IQR                              1.30
##                CV                              0.12
##          Skewness                             -0.28
##       SE.Skewness                              0.34
##          Kurtosis                             -0.98
##           N.Valid                             48.00
##         Pct.Valid                            100.00
```

This scatter plot will graph all cheese types as a measure of the total percent of consumption throughout the years. Mozzarella and Cheddar are among the highest consumer-selected cheeses in the country.

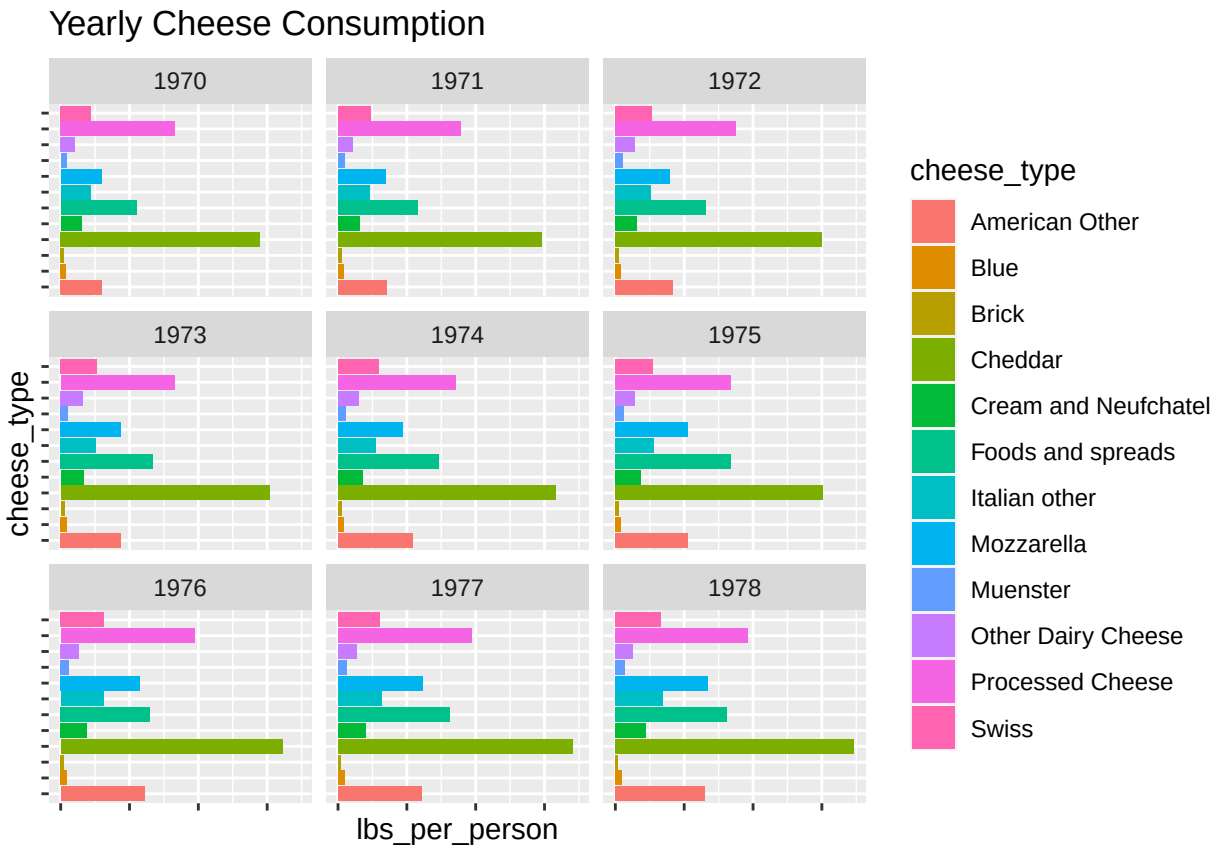
```
total_cheese_consumption = clean_cheese$`Total Natural Cheese` + clean_cheese$`Total Processed Cheese P

# The pipe command is intended to select the variables of interest and adjust the pivot of the table in
all_cheese_long <- clean_cheese%>%
  select(Year: "Foods and spreads")%>%
  gather(cheese_type, lbs_per_person, -Year) %>%
  mutate(total_consumption = lbs_per_person/total_cheese_consumption)

ggplot(all_cheese_long, aes(x = Year, y = total_consumption, col = cheese_type)) +
    geom_point(stat = "identity", position = "stack") +
    scale_y_continuous(labels = scales::percent) +
    labs(title = "Total Cheese Consumption by Cheese Type", y = "Percent of Total Consumption")
```

```
## Warning: Removed 12 rows containing missing values (position_stack).
```

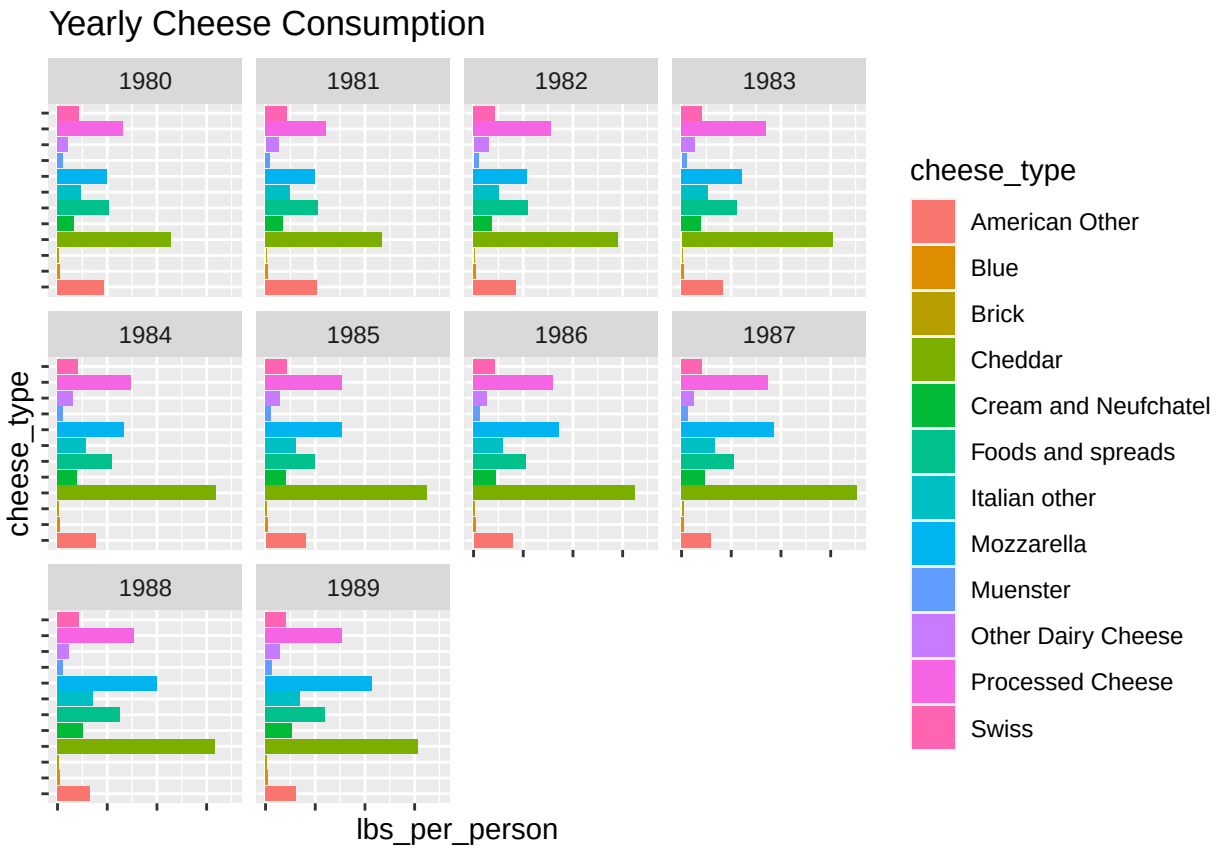


In order to better understand the trends within this time period, the bar plots below display the trends by decade. Expounding on observations made on the previous scatterplot, this set of bar plots show that Mozzarella became the more popular cheese in the 2000-2010 year bracket.

```
all_cheese_long %>%
filter(Year < "1979") %>%
  ggplot(aes(x = lbs_per_person, y = cheese_type, fill = cheese_type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Year) +
  labs(title = "Yearly Cheese Consumption") +
  theme(axis.text.y=element_blank(), axis.text.x=element_blank())
```

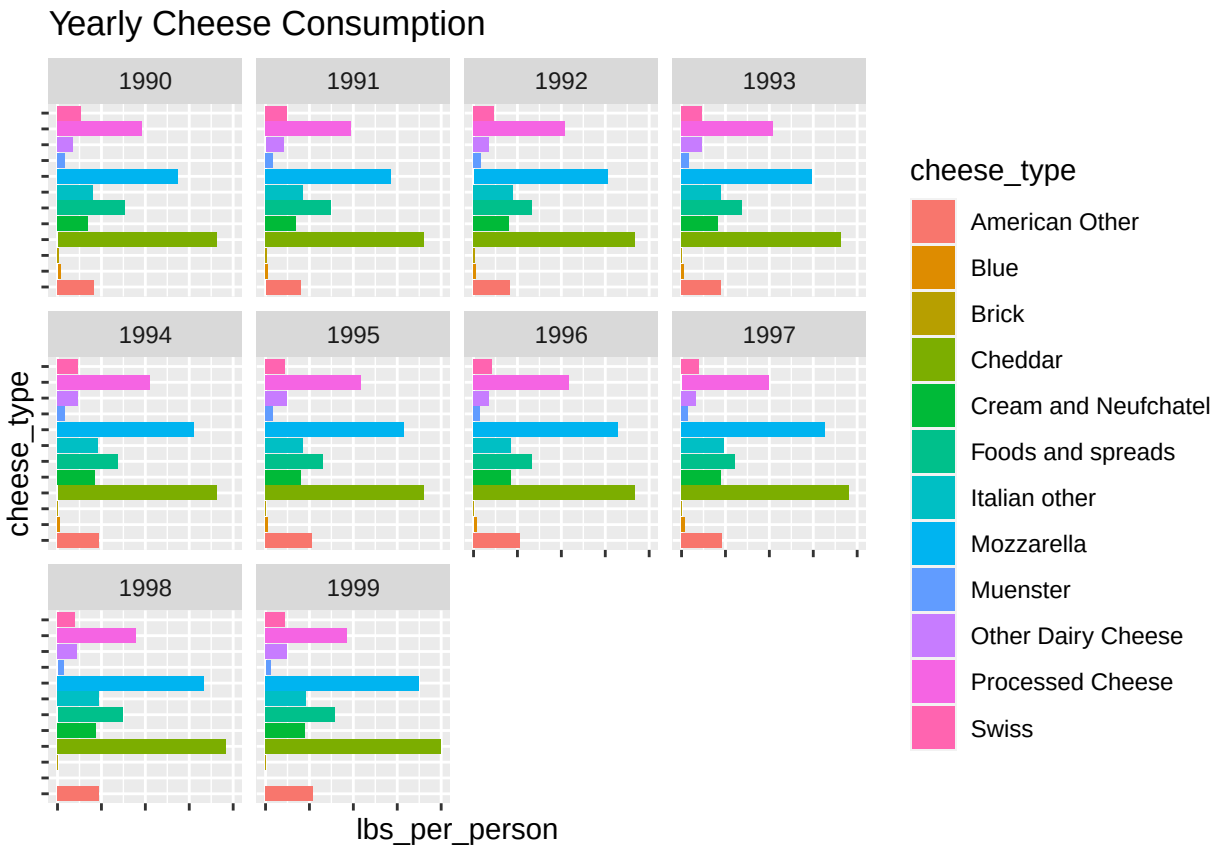



```
all_cheese_long %>%
filter(Year > "1979" & Year < "1990") %>%
  ggplot(aes(x = lbs_per_person, y = cheese_type, fill = cheese_type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Year) +
  labs(title = "Yearly Cheese Consumption") +
  theme(axis.text.y=element_blank(), axis.text.x=element_blank())
```



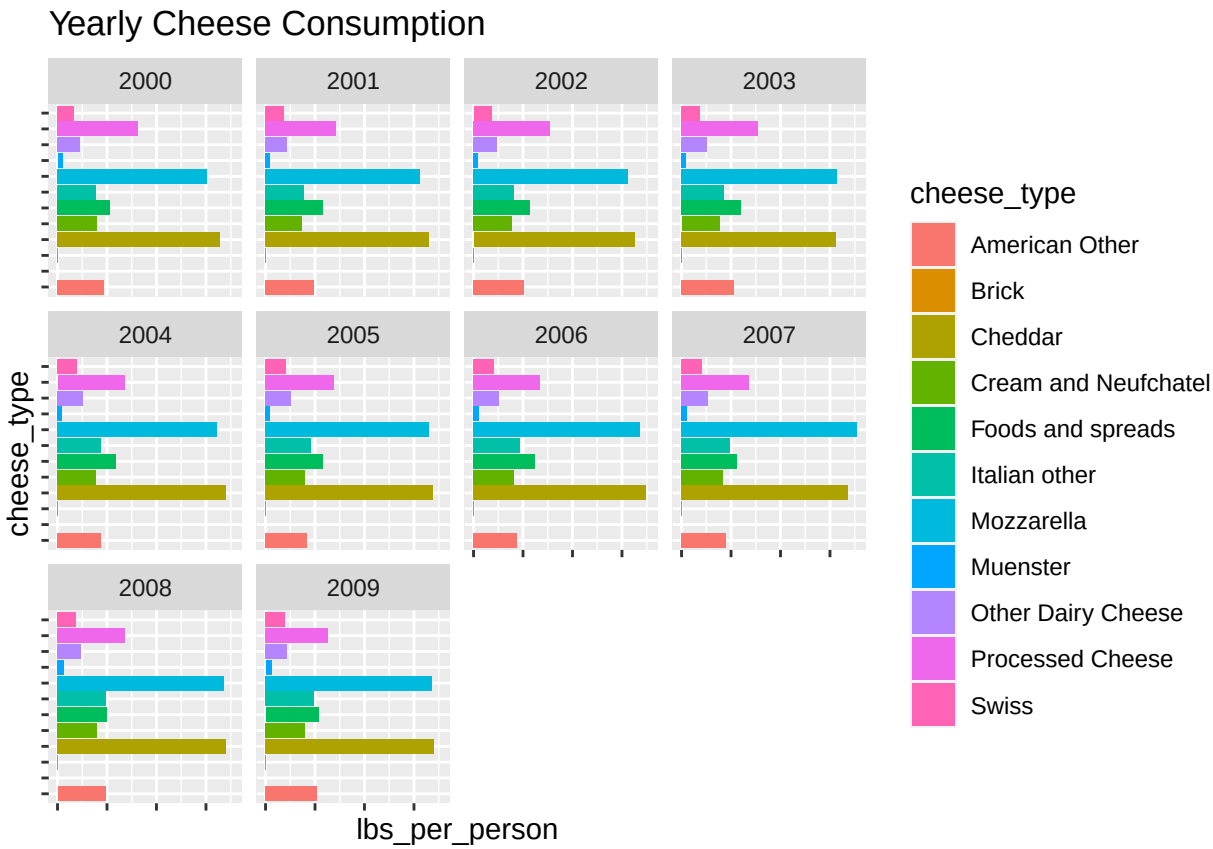

```
all_cheese_long %>%
filter(Year > "1989" & Year < "2000") %>%
  ggplot(aes(x = lbs_per_person, y = cheese_type, fill = cheese_type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Year) +
  labs(title = "Yearly Cheese Consumption") +
  theme(axis.text.y=element_blank(), axis.text.x=element_blank())
```

```
## Warning: Removed 2 rows containing missing values (position_stack).
```




```
all_cheese_long %>%
filter(Year > "1999" & Year < "2010") %>%
  ggplot(aes(x = lbs_per_person, y = cheese_type, fill = cheese_type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Year) +
  labs(title = "Yearly Cheese Consumption") +
  theme(axis.text.y=element_blank(), axis.text.x=element_blank())
```

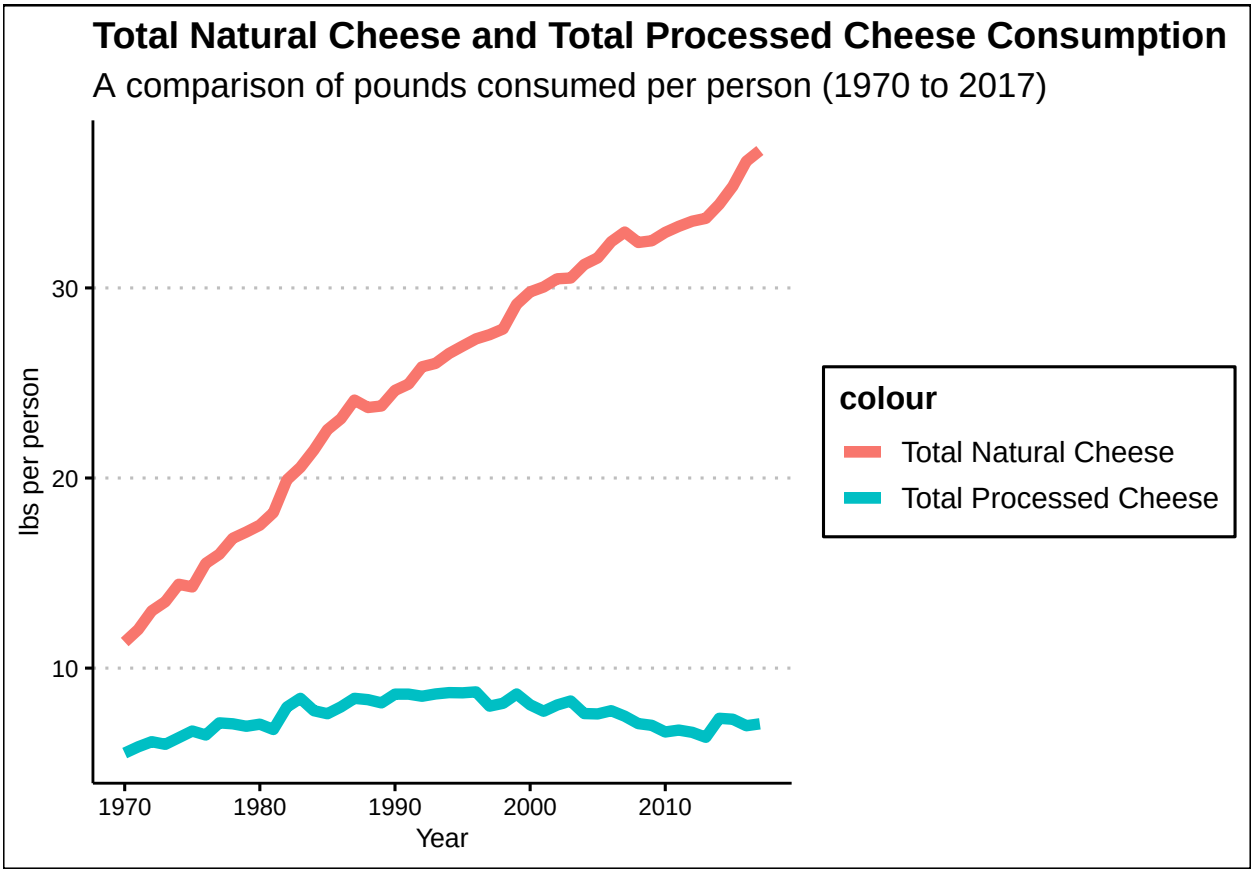
```
## Warning: Removed 10 rows containing missing values (position_stack).
```




```
all_cheese_long %>%
filter(Year > "2009") %>%
  ggplot(aes(x = lbs_per_person, y = cheese_type, fill = cheese_type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Year) +
  labs(title = "Yearly Cheese Consumption") +
  theme(axis.text.y=element_blank(), axis.text.x=element_blank())
```

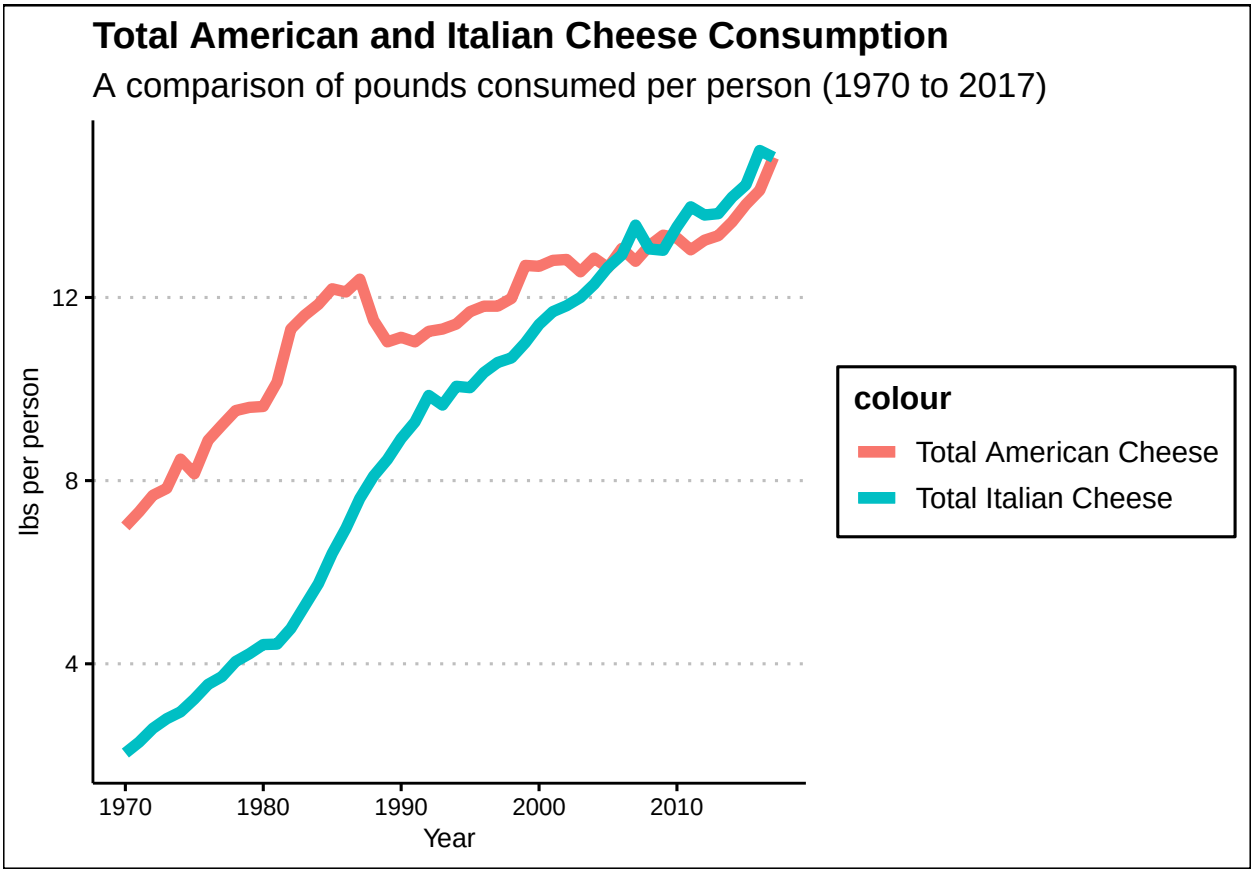
This portion seeks to examine the relationship between natural cheese trends and processed cheese trends. While they apparently lack a direct correlation in trends, total natural cheese consumption has continued to rise while total processed cheese consumption remains largely unchanged or in a slight decline.

```
ggplot(clean_cheese) +
  geom_line(aes(x = clean_cheese$Year,
                y = clean_cheese$`Total Natural Cheese`,
                 color = "Total Natural Cheese"),
                size = 2) +
  geom_line(aes(x = clean_cheese$Year, y =
                clean_cheese$`Total Processed Cheese Products`,
                color = "Total Processed Cheese"),
            size = 2) +
                labs(
                  title = "Total Natural Cheese and Total Processed Cheese Consumption",
                  subtitle = "A comparison of pounds consumed per person (1970 to 2017)",
                  y = "lbs per person",
                  x = "Year"
                ) + theme_clean()
```



In order to visualize other trends that may correspond with the decline of processed cheese, this portion is dedicated to a line graph which depicts the relationship between Italian cheese and American cheese. Due to the high consumption of cheddar and mozzarella, both retain a positive correlation.

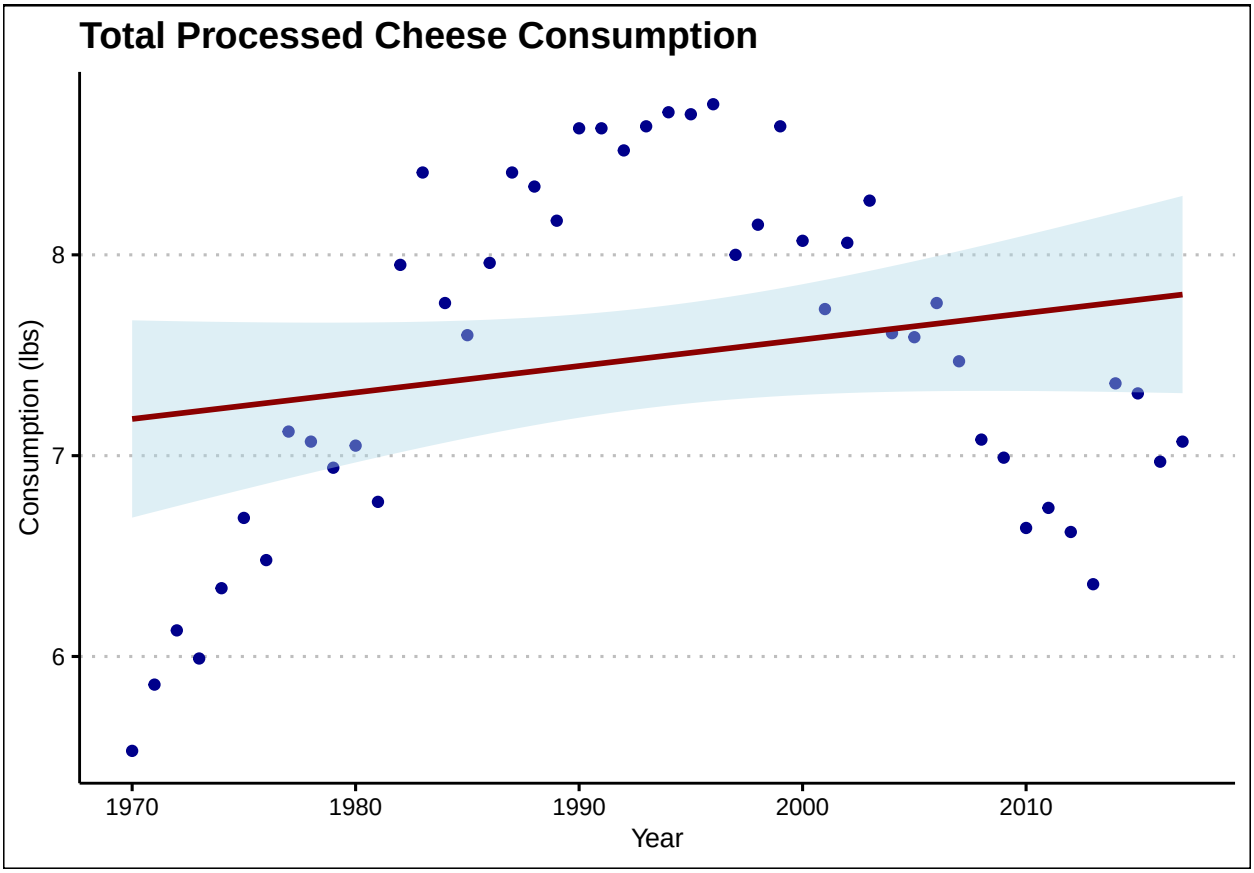
```
ggplot(clean_cheese) +
  geom_line(aes(x = clean_cheese$Year,
                y = clean_cheese$`Total American Chese`,
                 color = "Total American Cheese"),
                size = 2) +
  geom_line(aes(x = clean_cheese$Year, y =
                clean_cheese$`Total Italian Cheese`,
                color = "Total Italian Cheese"),
            size = 2) +
                labs(
                  title = "Total American and Italian Cheese Consumption",
                  subtitle = "A comparison of pounds consumed per person (1970 to 2017)",
                  y = "lbs per person",
                  x = "Year"
                ) + theme_clean()
```



A closer look at processed cheese consumption shows that it peaked around the year 2000 followed by a swift decline.

```
ggplot(clean_cheese,
       aes(x = clean_cheese$Year,
           y = clean_cheese$`Total Processed Cheese Products`,
           color = "Total Processed Cheese")) + geom_point(color = "darkblue") +
  geom_smooth(method = lm,
              se = TRUE,
              fullrange = TRUE, color = "darkred", fill = "lightblue") +
  labs(title = "Total Processed Cheese Consumption", x = "Year", y = "Consumption (lbs)") + theme_clean
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The following are simple statistical analyses, including a two-way ANOVA and simple linear regression. Due to a low significance value for both, other tests should be performed for a better statistical understanding of the data.

```
#Linear regression for total processed cheese products
tpcp.reg <-
  lm(clean_cheese$Year ~ clean_cheese$`Total Processed Cheese Products`)
summary(tpcp.reg)
```

```
##
## Call:
## lm(formula = clean_cheese$Year ~ clean_cheese$`Total Processed Cheese Products`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.917 -12.143  -4.044  11.271  24.945
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                 1967.875     17.503 112.428
## clean_cheese$`Total Processed Cheese Products`    3.420      2.321   1.474
##                                             Pr(>|t|)
## (Intercept)                                   <2e-16 ***
## clean_cheese$`Total Processed Cheese Products`    0.147
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 46 degrees of freedom
## Multiple R-squared:  0.04508,    Adjusted R-squared:  0.02432
## F-statistic: 2.172 on 1 and 46 DF,  p-value: 0.1474
```

```
res.aov <- aov(clean_cheese$Year ~ clean_cheese$`Total Processed Cheese Products`+ clean_cheese$`Total
summary(res.aov)
```

```
##                                              Df Sum Sq Mean Sq F value Pr(>F)
## clean_cheese$`Total Processed Cheese Products`  1    415     415   387.5 <2e-16
## clean_cheese$`Total Natural Cheese`             1   8749    8749  8162.8 <2e-16
## Residuals                                      45     48       1
##
## clean_cheese$`Total Processed Cheese Products` ***
## clean_cheese$`Total Natural Cheese`            ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Saving cheese type averages into a text file

```
cheese_type_only <- clean_cheese %>% select("Cheddar":"Foods and spreads")
means <- cheese_type_only %>% summarise_each(funs(mean))
```

```
## Warning: `summarise_each_()` is deprecated as of dplyr 0.7.0.
## Please use `across()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
write.csv(means, file = 'cheese_type_means.txt')
```

Cows

In this section, Suzanna Storms analyzes dairy cow data, focusing on milk production improvement, and the price of cows over time.

knitr::opts_chunk$set(echo = TRUE) library(tidytuesdayR) library(dplyr) tt_datasets(2019) library(ggplot2) library(data.table)

dairy = tt_load(2019,5) cheese = dairy$clean_cheese milk_fluid = dairy$fluid_milk_sales milk_facts = dairy$milk_products_facts cow = dairy$milkcow_facts

‘‘‘

Cow data

##Basic stats functions

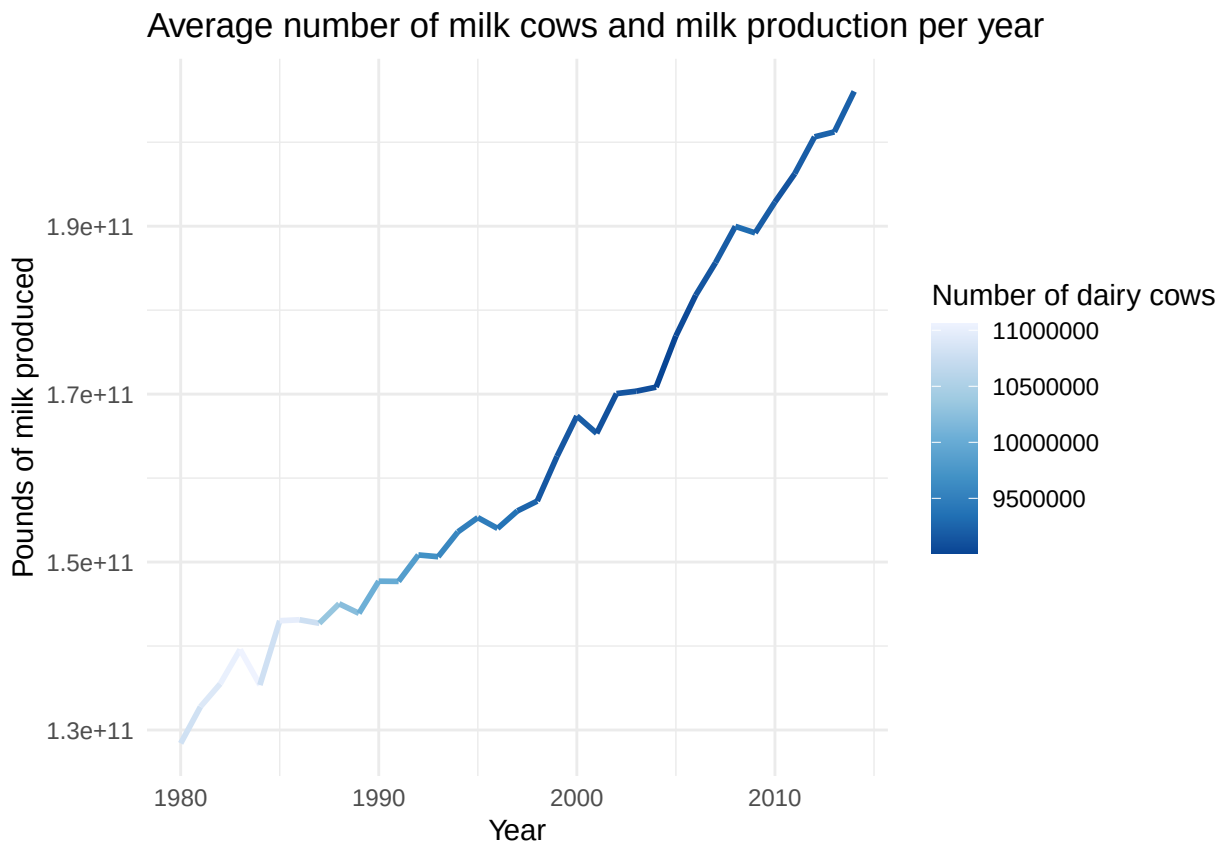
Here I will extract put my data into a summary table using the average milk price, average milk cow number, milk production, and slaugher (cull) cow price. Trying again:

```
newcow <- cow %>% select(avg_price_milk, avg_milk_cow_number, milk_production_lbs, slaughter_cow_price,
cow_sum_data <- summarytools::descr(newcow)
View(cow_sum_data)
df_csd <- as.data.frame(cow_sum_data)
```

##Graphs

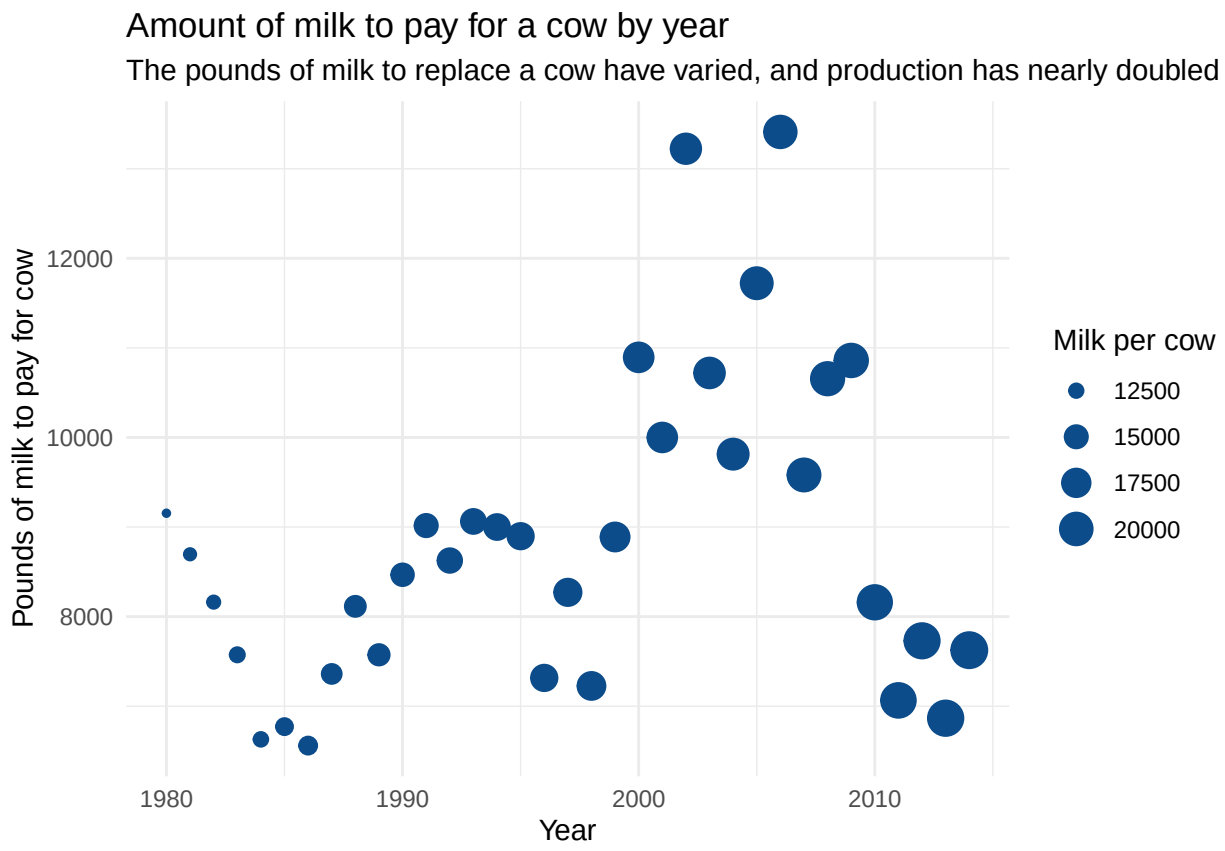
The first figure displays as that over time, fewer cows are needed to produce more milk.

```
ggplot(cow) +
 aes(x = year, y = milk_production_lbs, colour = avg_milk_cow_number) +
 geom_line(size = 1L) +
 scale_color_distiller(palette = "Blues") +
 labs(x = "Year", y = "Pounds of milk produced", title = "Average number of milk cows and milk producti
 theme_minimal()
```



The second figure displays pounds of milk to pay for a cow, and milk produced per cow over time. We can see that, except for in the years 2000-2010, the pounds of milk needed to pay for a cow has remained steay, and cows have become more efficient, paying for themselves in half the time. In 1980, it took nearly one year for a cow to pay for herself in milk volume. However, in the last 5 years, a cow could pay for herself in one third of the time due to the volume of milk modern cows produce.

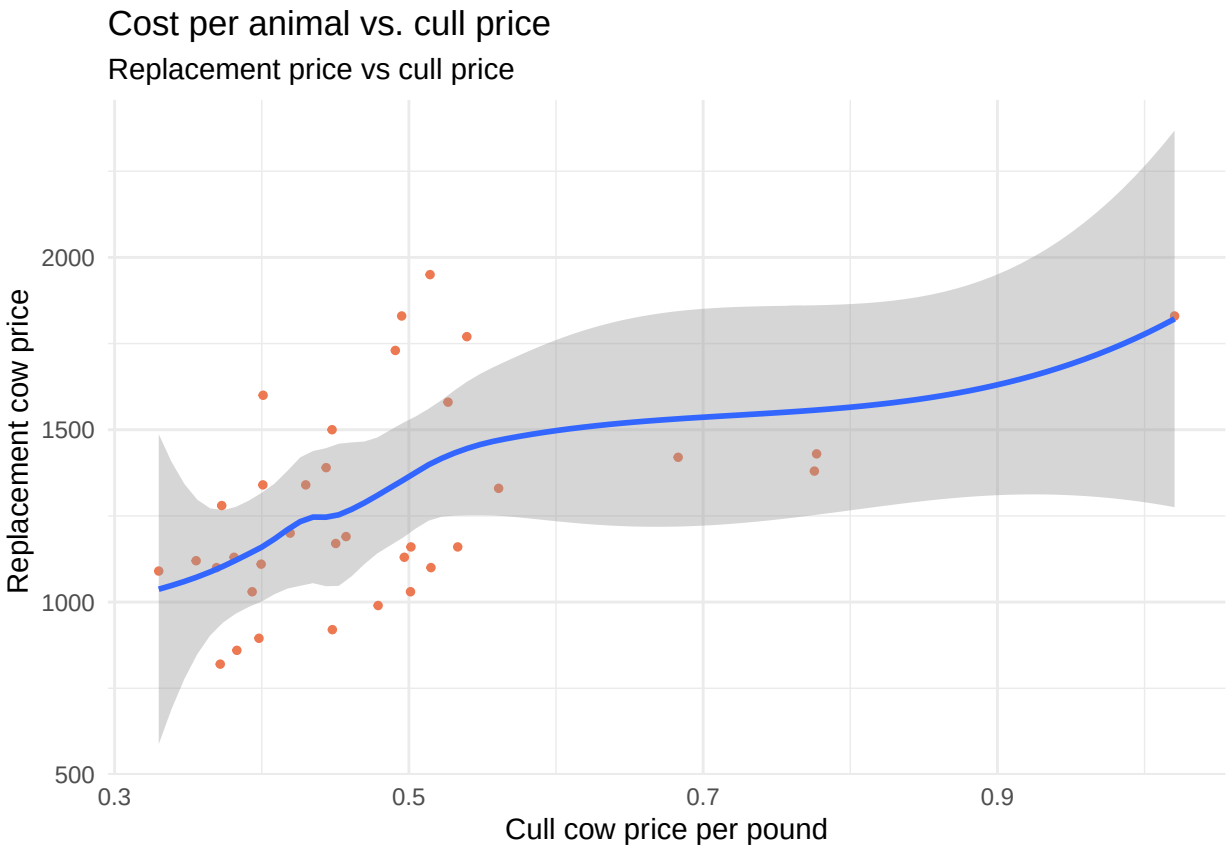
```
library(ggplot2)

ggplot(cow) +
 aes(x = year, y = milk_volume_to_buy_cow_in_lbs, size = milk_per_cow) +
 geom_point(colour = "#0c4c8a") +
 labs(x = "Year", y = "Pounds of milk to pay for cow", title = "Amount of milk to pay for a cow by year
 theme_minimal()
```



The third plot shows the slaughter cow price per pound by the total price for a replacment cow. Here we can see that often, seldom do high cull prices correlate with low replacement costs.

```
library(ggplot2)

ggplot(cow) +
 aes(x = slaughter_cow_price, y = milk_cow_cost_per_animal) +
 geom_point(size = 1L, colour = "#ed7953") +
 geom_smooth(span = 0.75) +
 labs(x = "Cull cow price per pound", y = "Replacement cow price ", title = "Cost per animal vs. cull p
 theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



##ANOVA Here I have provided two examples of ANOVA comparing the pounds of milk produced by cow number, the milk production and the price of milk, and then the slaughter price for a cow compared to buying a new cow.

```
aov.cow_mp <- aov(cow$milk_production_lbs ~ cow$avg_milk_cow_number)
summary(aov.cow_mp)
```

```
##                           Df    Sum Sq   Mean Sq F value   Pr(>F)
## cow$avg_milk_cow_number    1 9.717e+21 9.717e+21      47 7.96e-08 ***
## Residuals                 33 6.823e+21 2.067e+20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.cow_mprice <- aov(cow$milk_production_lbs ~ cow$avg_price_milk)
summary(aov.cow_mprice)
```

```
##                     Df   Sum Sq   Mean Sq F value   Pr(>F)
## cow$avg_price_milk   1 8.97e+21 8.970e+21    39.1 4.59e-07 ***
## Residuals           33 7.57e+21 2.294e+20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.cow_sl<- aov(cow$slaughter_cow_price ~ cow$milk_cow_cost_per_animal)
summary(aov.cow_sl)
```

```
##                               Df Sum Sq Mean Sq F value  Pr(>F)
## cow$milk_cow_cost_per_animal  1 0.1561 0.15606   10.16 0.00313 **
## Residuals                    33 0.5067 0.01535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##Verifying assumptions Here I am verifying some assumptions in order to appropriately use an ANOVA. I have several examples of QQ plots.

```
cow_mp.residuals <- aov.cow_mp$residuals
cow_mprice.residuals <- aov.cow_mprice$residuals
cow_sl.residuals<- aov.cow_sl$residuals
#boxplot not appropirate for my type of data.
#testing for normality
shapiro.test (cow_mp.residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cow_mp.residuals
## W = 0.91689, p-value = 0.01161
```

```
nortest::lillie.test(cow_mp.residuals)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cow_mp.residuals
## D = 0.16849, p-value = 0.01319
```

```
shapiro.test (cow_mprice.residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cow_mprice.residuals
## W = 0.96628, p-value = 0.3496
```

```
nortest::lillie.test(cow_mprice.residuals)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cow_mprice.residuals
## D = 0.12594, p-value = 0.1717
```

```
shapiro.test (cow_sl.residuals)
```

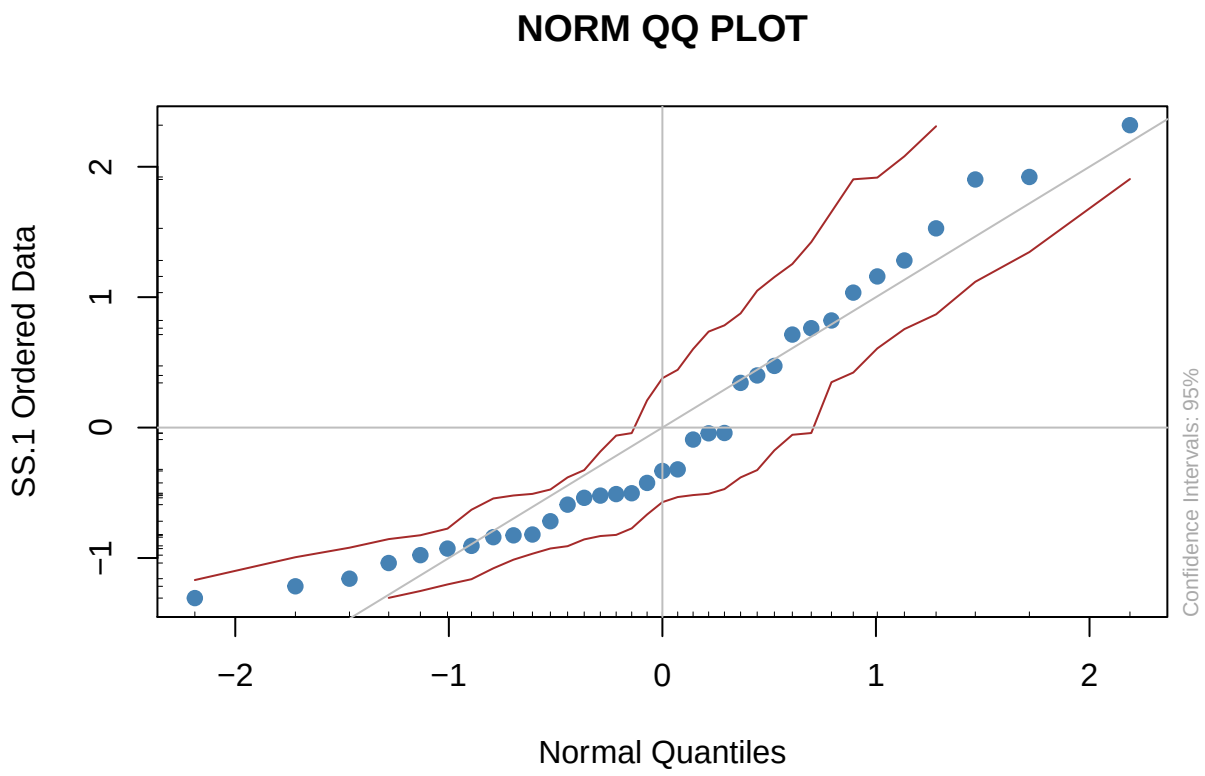
```
##
##  Shapiro-Wilk normality test
##
## data:  cow_sl.residuals
## W = 0.85898, p-value = 0.0003711
```

```
nortest::lillie.test(cow_sl.residuals)
```

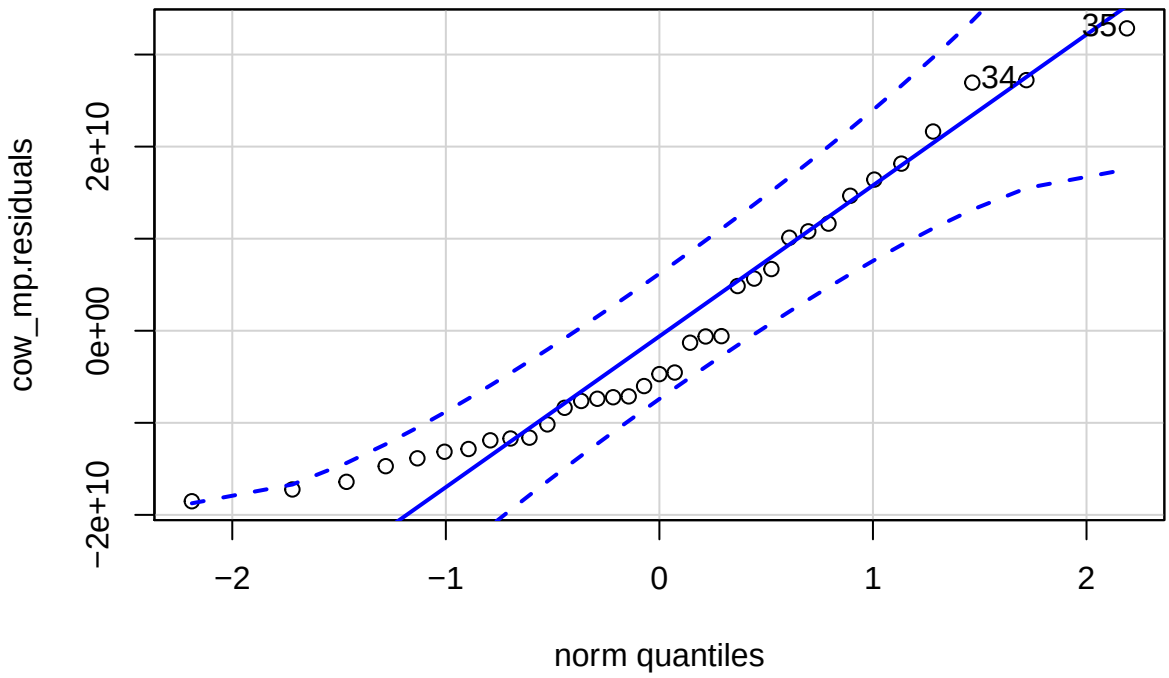
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cow_sl.residuals
## D = 0.15786, p-value = 0.0272
```

```
#testing QQ plot
fBasics::qqnormPlot(cow_mp.residuals)
```

```
## Warning in whichFormat(charvec[1]): character string is not in a standard
## unambiguous format
```

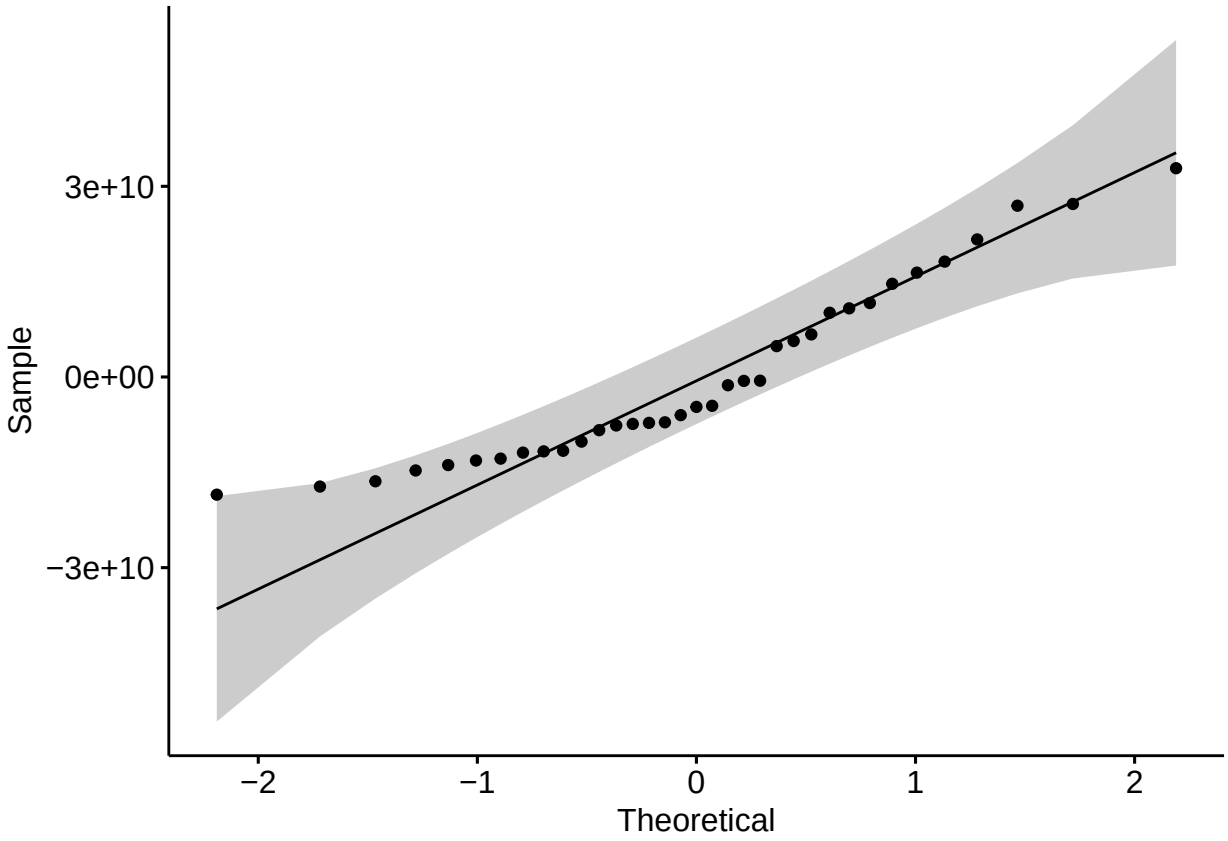



```
car::qqPlot(cow_mp.residuals)
```



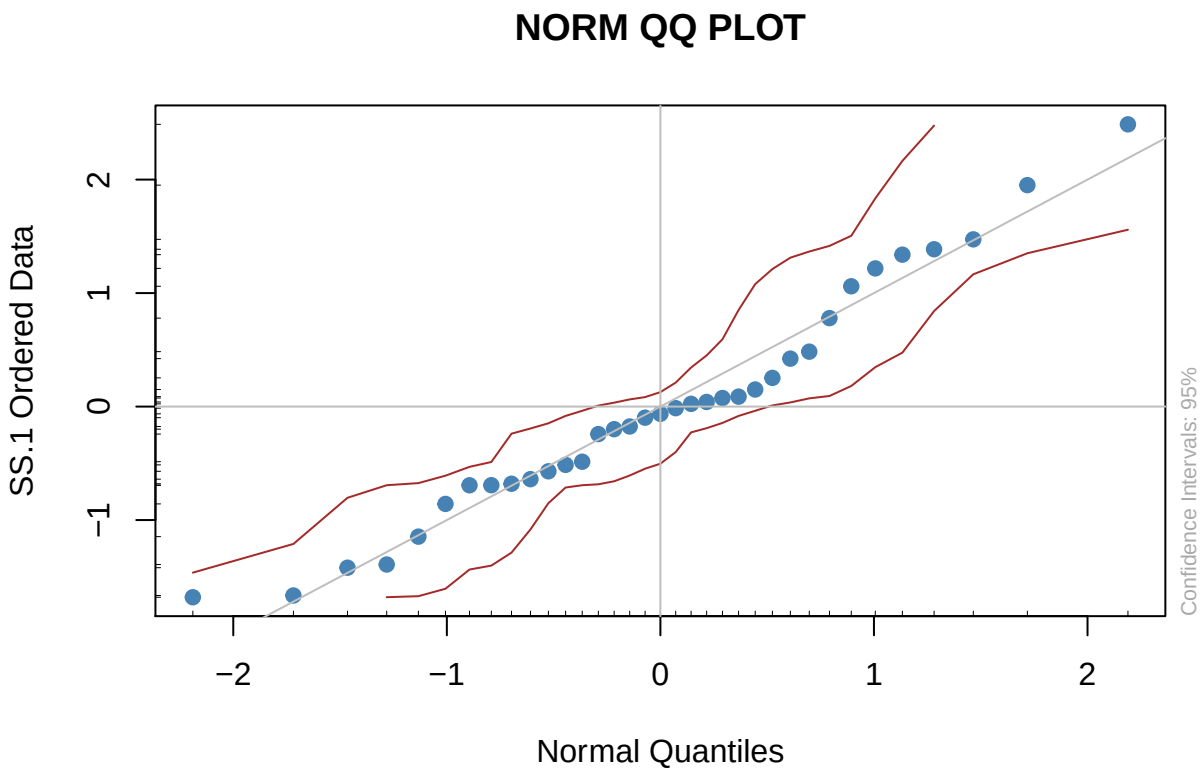

```
## [1] 35 34
```

```
ggpubr::ggqqplot(cow_mp.residuals)
```

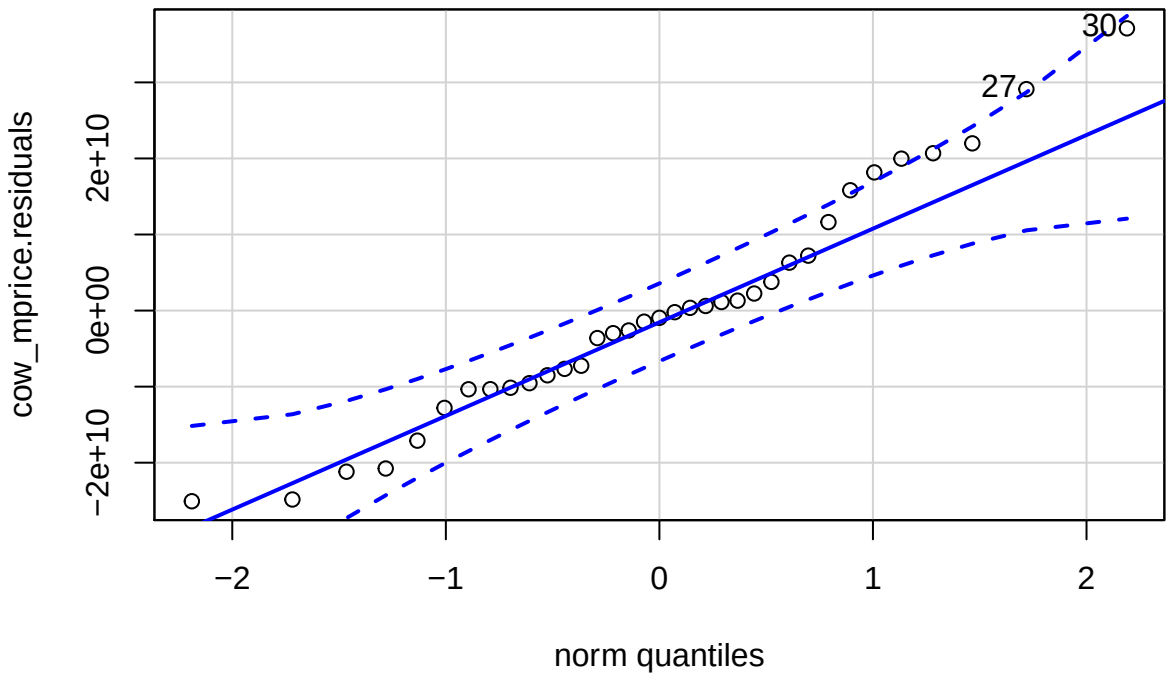



```
fBasics::qqnormPlot(cow_mprice.residuals)
```

```
## Warning in whichFormat(charvec[1]): character string is not in a standard
## unambiguous format
```

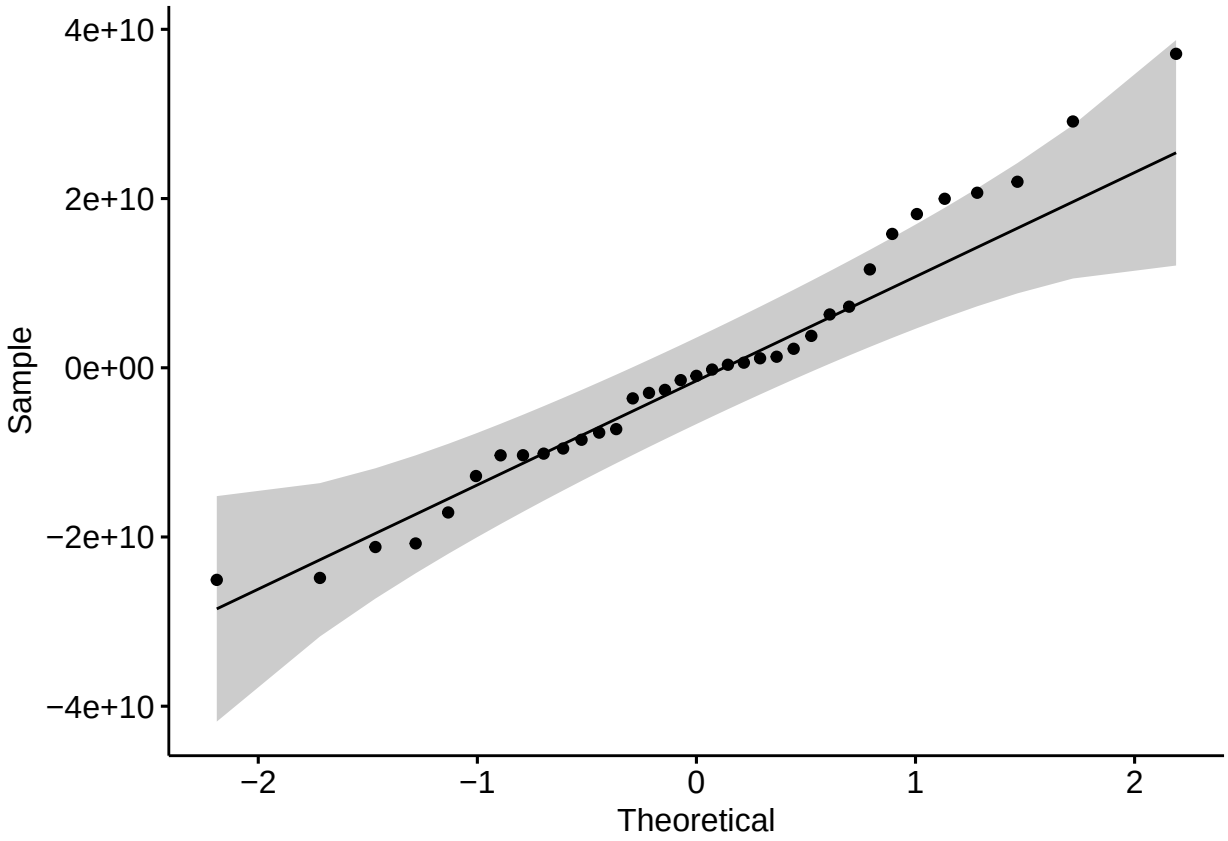



```
car::qqPlot(cow_mprice.residuals)
```



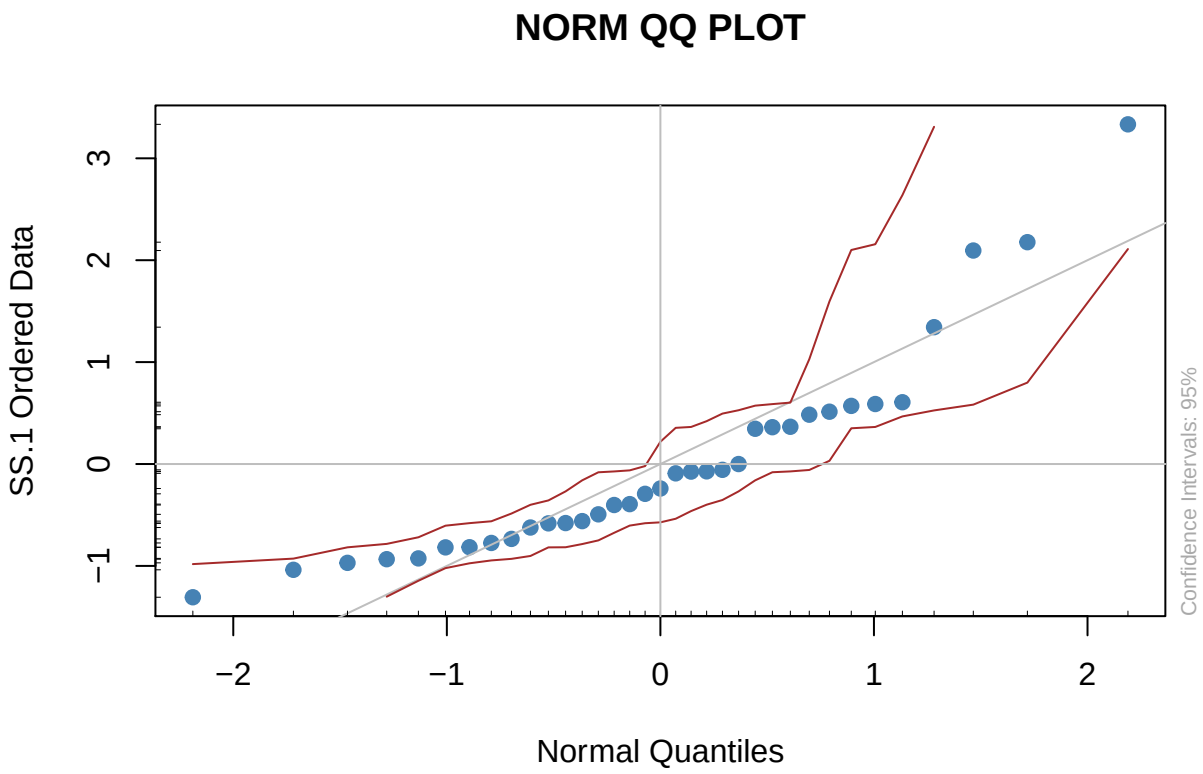

```
## [1] 30 27
```

```
ggpubr::ggqqplot(cow_mprice.residuals)
```

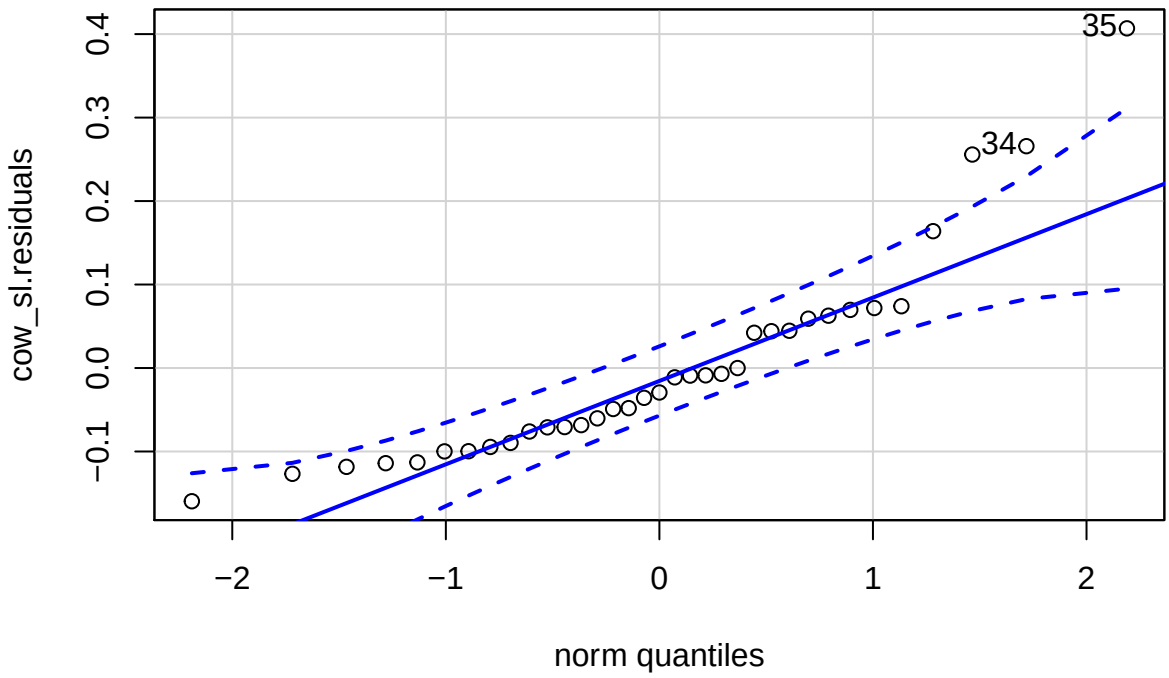



```
fBasics::qqnormPlot(cow_sl.residuals)
```

```
## Warning in whichFormat(charvec[1]): character string is not in a standard
## unambiguous format
```

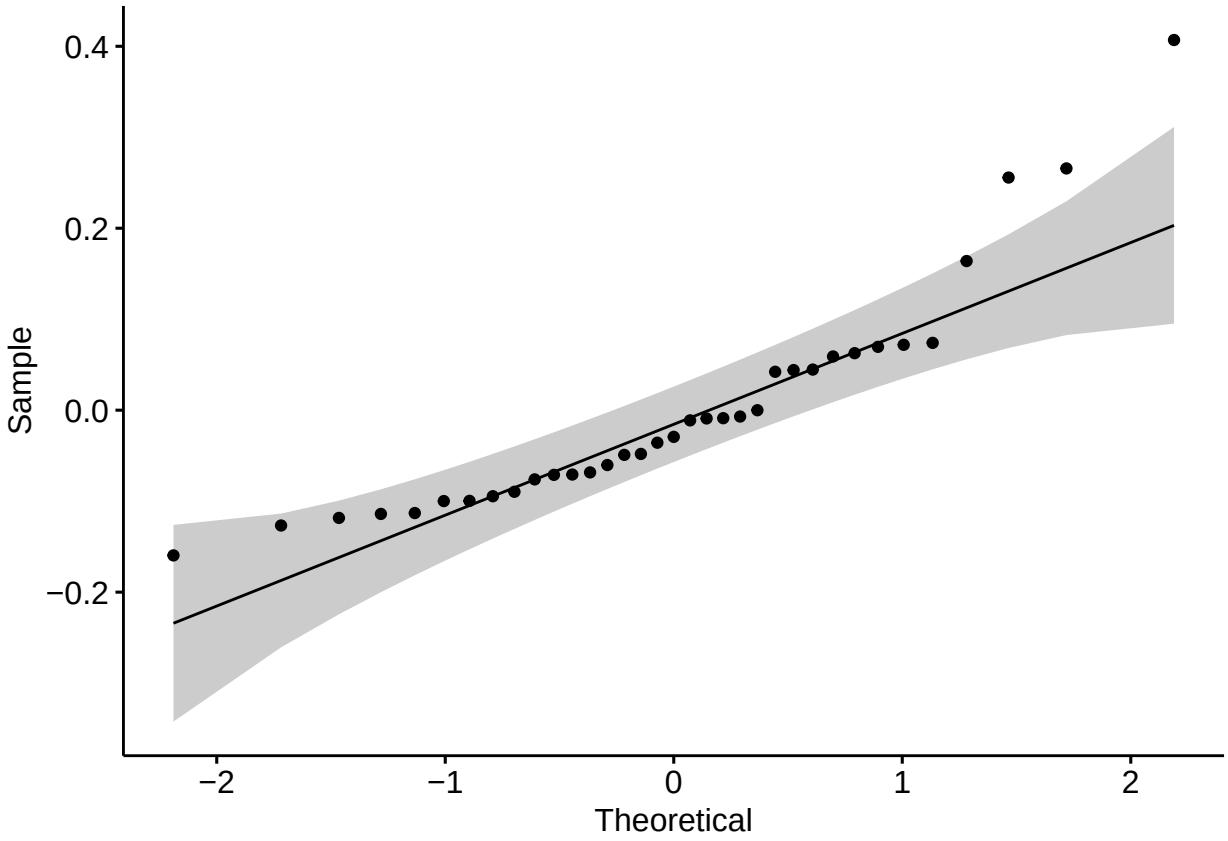



```
car::qqPlot(cow_sl.residuals)
```




```
## [1] 35 34
```

```
ggpubr::ggqqplot(cow_sl.residuals)
```



##Linear regression Here we are looking at a linear regression of the average milk cow number and milk production in pounds by year, followed by a linear regression of the slaughter cow price and the cost of a dairy cow animal by year.

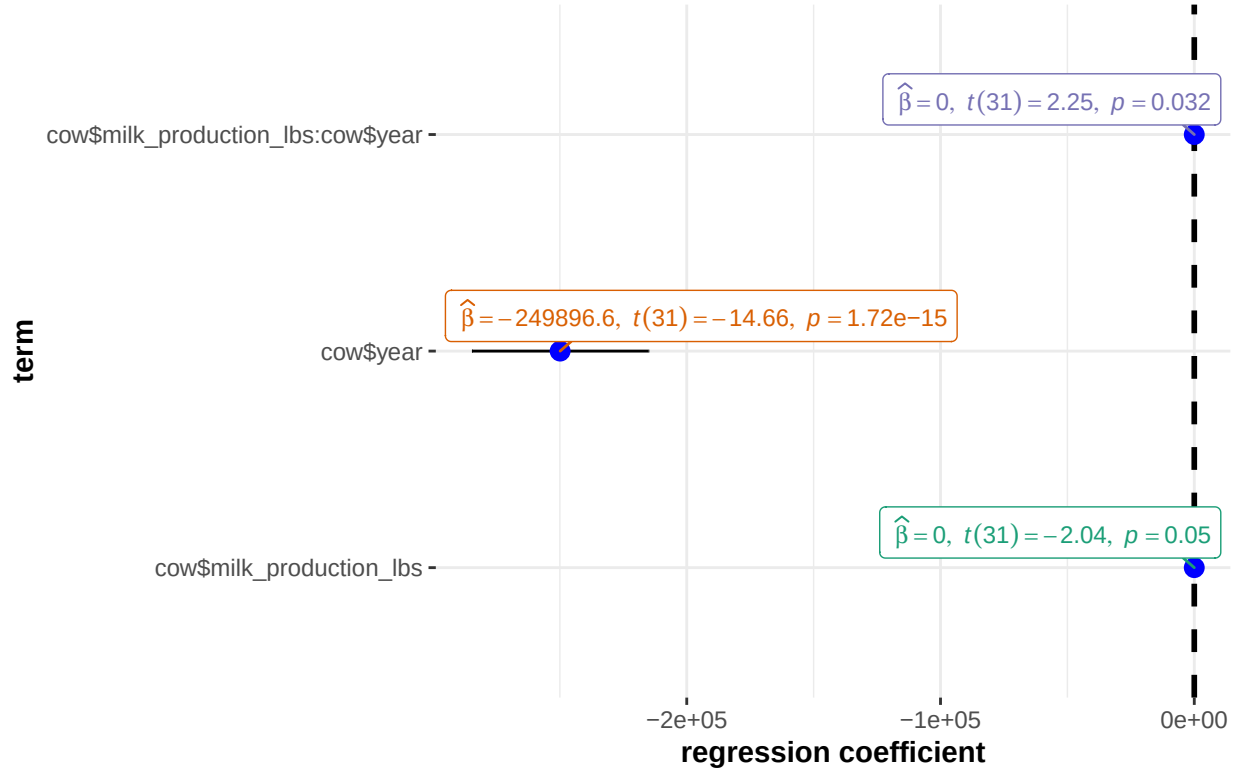
```
data(cow)
```

```
## Warning in data(cow): data set 'cow' not found
```

```
mod <- lm(cow$avg_milk_cow_number ~ cow$milk_production_lbs *cow$year )
ggstatsplot::ggcoefstats(mod)
```

```
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

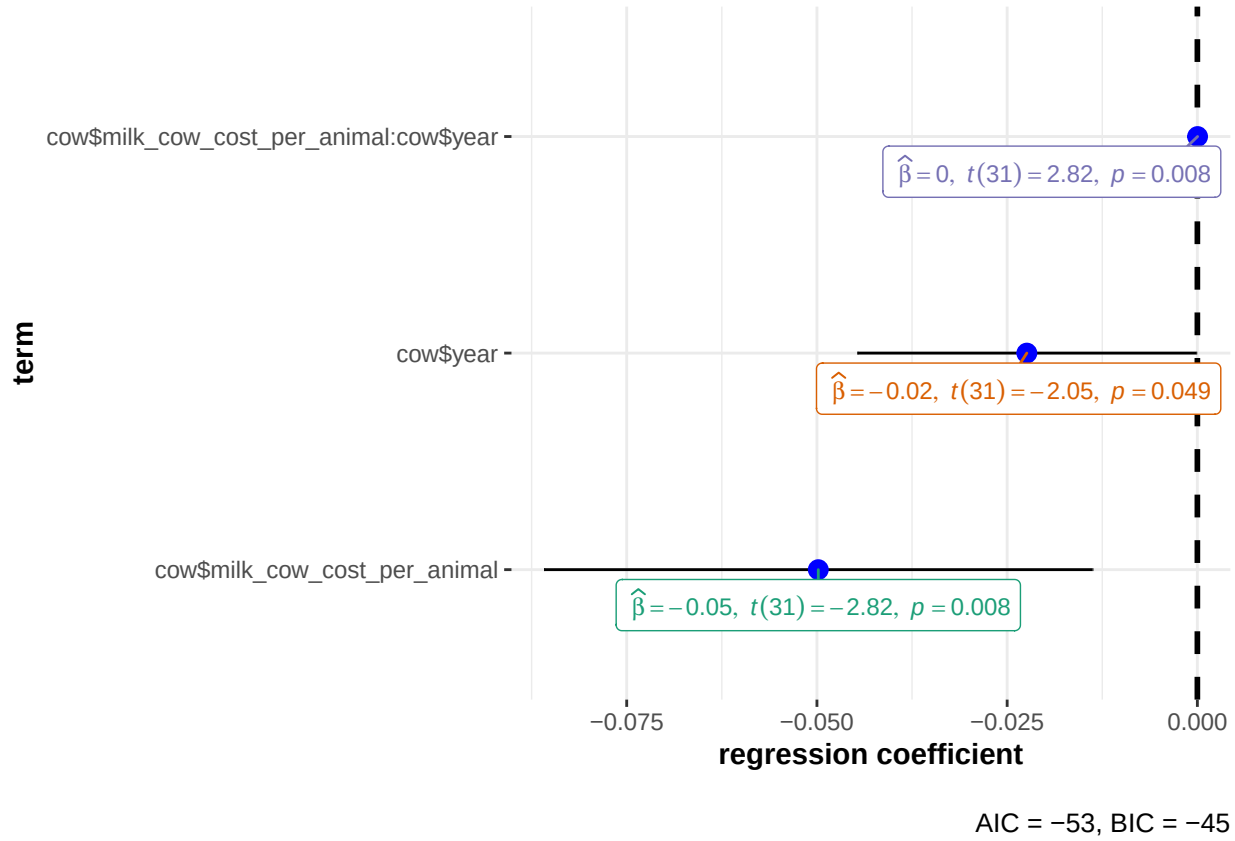


AIC = 919, BIC = 927

```
data(cow)
```

```
## Warning in data(cow): data set 'cow' not found
```

```
mod2<- lm(cow$slaughter_cow_price ~cow$milk_cow_cost_per_animal *cow$year)
ggstatsplot::ggcoefstats(mod2)
```



fwrite(cow_sum_data,"cow_sum_txt.txt", sep=";",col.names = FALSE, row.names = FALSE)

Conclusion

In this final project for CPSC 441, we used data from the USDA to calculate and create statistical summaries, graphically visualize data, perform statistical tests, and save the means of selected variables as a text file.