# Data_Analysis

*Terry Slenn*

*November 23, 2019*

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.
```

```
## v ggplot2 3.1.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
## Warning: package 'readr' was built under R version 3.5.3
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
## Warning: package 'forcats' was built under R version 3.5.3
```

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

```
library(gsheet)
```

```
## Warning: package 'gsheet' was built under R version 3.5.3
```

```
library(ggplot2)
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 3.5.3
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
Prices <- gsheet2tbl('https://docs.google.com/spreadsheets/d/1LC1dk9HlzCEqVPN2AGI6trGaXpwlP8wtyYheK6k70

Prices <- Prices %>%
  separate(LAT_LONG, into = c("LAT", "LONG"), sep = ", ") %>%
  mutate(LAT = as.numeric(LAT), LONG = as.numeric(LONG))

Stores <- Prices %>%
  distinct(ADDRESS, LAT, LONG, STORE, STATE)

Prices %>%
  distinct(`ZIP CODE`, STATE)
```

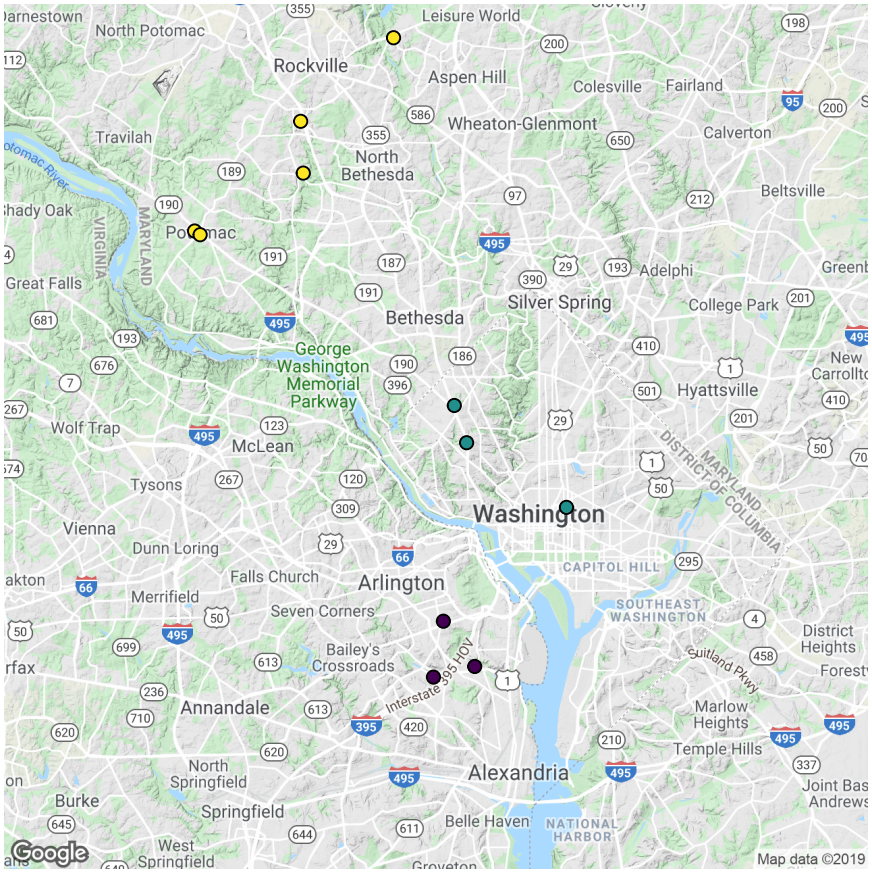
```
## # A tibble: 6 x 2
##   `ZIP CODE` STATE
##        <dbl> <chr>
## 1      20854 MD
## 2      20853 MD
## 3      20016 DC
## 4      20001 DC
## 5      22206 Arlington
## 6      22204 Arlington
```

```
DC_map <- get_map(location = c(lat = 38.937495, lon =-77.088846), zoom = 11)
```

```
## note : locations should be specified in the lon/lat format, not lat/lon.
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=38.937495,-77.088846&zoom=11&size=640
```

```
ggmap(DC_map) +
  geom_point(data = Stores, mapping = aes(x = LONG, y = LAT, fill = STATE),
             pch = 21, size = 2) +
  theme_void() +
  theme(legend.position = "none")+
  scale_fill_viridis_d()
```




## Data Cleaning

```
OJ <- Prices %>%
  filter(PRODUCT == "Orange Juice") %>%
  mutate(PRICE_STD = PRICE/parse_number(SIZE)*64) ## Standardize to 64 oz

Milk <- Prices %>%
  filter(PRODUCT == "Milk") %>%
  mutate(PRICE_STD = PRICE) ## SHOULD all be the same

Eggs <- Prices %>%
  filter(PRODUCT == "Eggs") %>%
  mutate(PRICE_STD = PRICE/parse_number(SIZE)*64)

Beef <- Prices %>%
  filter(PRODUCT == "Ground Beef") %>%
  mutate(PRICE_STD = PRICE)

Potatoes <- Prices %>%
  filter(str_detect(PRODUCT, "ota")) %>%
  mutate(PRICE_STD = PRICE/parse_number(SIZE),
         PRICE_STD = coalesce(PRICE_STD, PRICE))
```

```
## Warning: 8 parsing failures.
## row col expected actual
##   1  -- a number     lb
##   2  -- a number     lb
##   3  -- a number     lb
##   4  -- a number     lb
##   5  -- a number     lb
## ... ... ........ ......
## See problems(...) for more details.
```

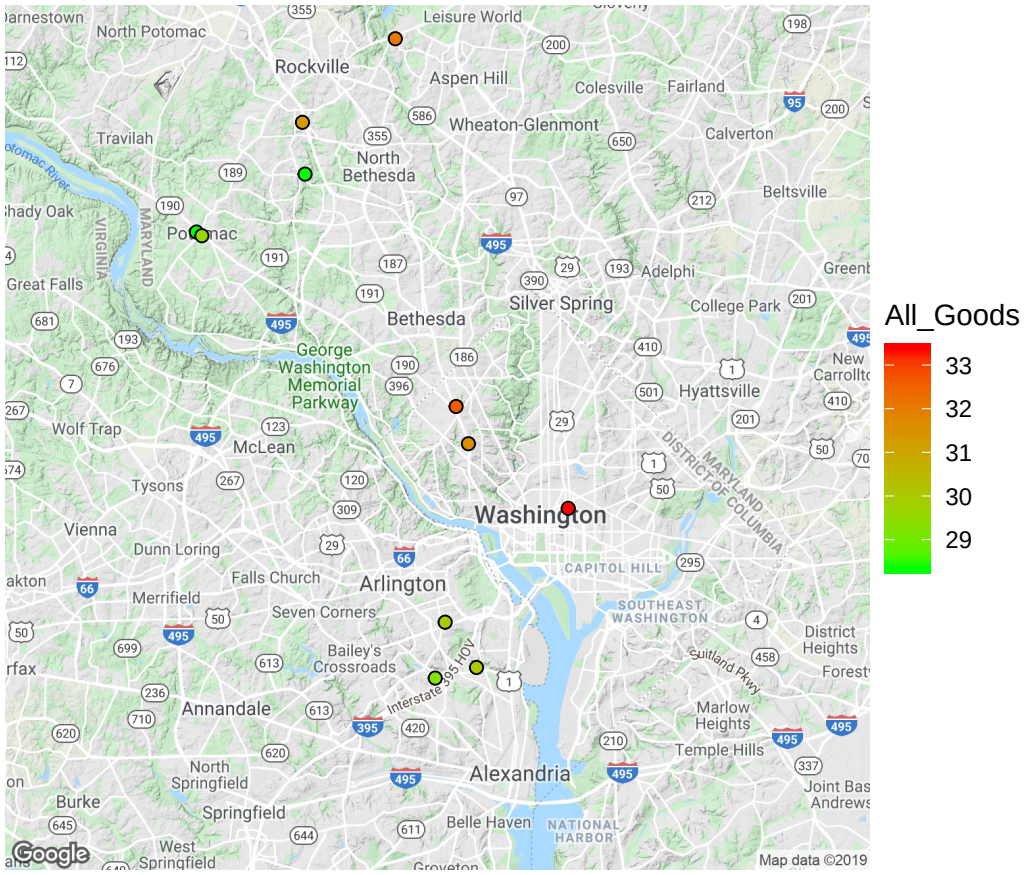
```
Cola <- Prices %>%
  filter(PRODUCT == "Cola") %>%
  mutate(PRICE_STD = PRICE)

Tortilla <- Prices %>%
  filter(str_detect(PRODUCT, "illa")) %>%
  mutate(PRICE_STD = PRICE/parse_number(SIZE)*12)

Prices <- bind_rows(OJ, Milk, Eggs, Beef, Potatoes, Cola, Tortilla) %>%
  group_by(ADDRESS, LAT, LONG, `ZIP CODE`, STATE, STORE) %>%
  summarize(All_Goods = sum(PRICE_STD),
            Luxury = sum(PRICE_STD[PRODUCT %in% c("Cola", "Tortilla Chips")]),
            Essential = All_Goods - Luxury)
```

## Exploratory Analysis

```
Store_Price <- Stores %>%
  left_join(Prices)
```

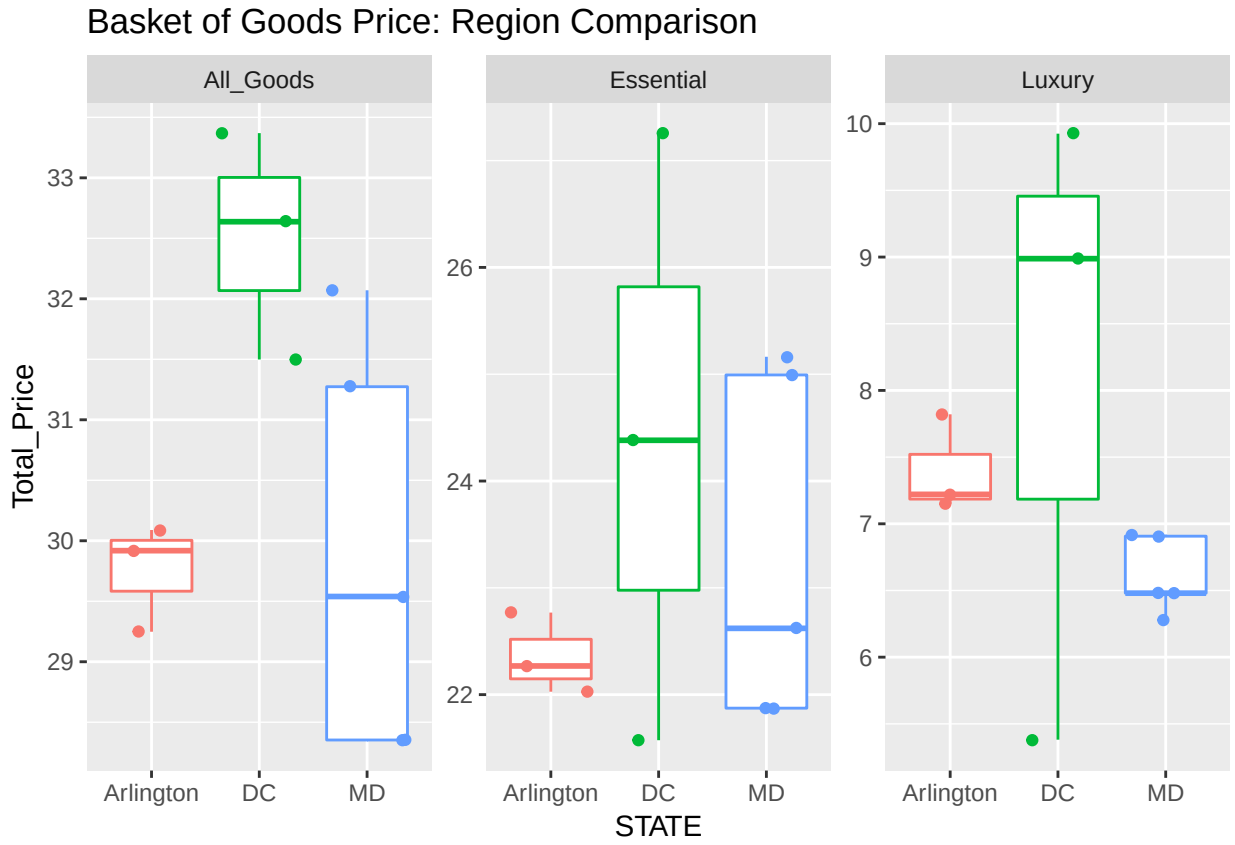
```
## Joining, by = c("ADDRESS", "LAT", "LONG", "STORE", "STATE")
```

```
ggmap(DC_map) +
  geom_point(data = Store_Price, mapping = aes(x = LONG, y = LAT, fill = All_Goods),
             pch = 21, size = 2) +
  theme_void() +
  scale_fill_gradient(low = "green", high = "red")
```




```
  #scale_fill_viridis_d()
```

```
Prices %>%
  gather(key = "Basket", value = "Total_Price", All_Goods, Essential, Luxury) %>%
  ggplot(aes(x = STATE, y = Total_Price, color = STATE)) +
  geom_boxplot() +
  geom_jitter() +
  facet_wrap(~ Basket, scales = "free_y") +
  ggtitle("Basket of Goods Price: Region Comparison") +
  theme(legend.position = "none")
```




```
Prices %>%
  gather(key = "Basket", value = "Total_Price", All_Goods, Essential, Luxury) %>%
  ggplot(aes(x = STORE, y = Total_Price, color = STORE)) +
  geom_boxplot() +
  geom_jitter() +
  facet_wrap(~ Basket, scales = "free_y", nrow = 2) +
  #coord_flip() +
  ggtitle("Basket of Goods Price: Store Comparison") +
  theme(legend.position = "none")
```


## Further work needed

Need to perform formal statistics tests between groups and figure out how cluster sampling impacts that.

Need to pull in demographic info to evaluate any trends that may exist with demographics and income.

**Problems:** Milk was supposed to be gallon price. Arlington data is throwing us way off. 8 lb bag of potatoes is also messing up data