A Study of Marijuana Prices Across the US in 2014

Max Shumway, Max Troilo, Dhanya Karra, Michael Varrone Jr.

2023-04-29

Setup

We will begin by loading in the necessary packages and the data set of interest for our project. We believe that the `weedprices` data will prove to be very useful for our research and cleaning needs.

```
pacman::p_load(tidyverse, readxl, knitr, kableExtra, ggmap, mapdata, ggthemes, viridis)
weed_prices <- read_csv("./weedprices.csv")
```

```
## Rows: 612 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): State, HighQ, MedQ, LowQ, Month
## dbl (3): HighQN, MedQN, LowQN
##
## 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.
```

Description and Basic Exploration of the Dataset

```
table <- tibble(weed_prices)
table
```

```
## # A tibble: 612 x 8
##    State                HighQ   HighQN MedQ    MedQN LowQ    LowQN Month
##    <chr>                <chr>    <dbl> <chr>   <dbl> <chr>   <dbl> <chr>
##  1 Alabama              $339.06   1042 $198.64   933 $149.49   123 Jan. 2014
##  2 Alaska               $288.75    252 $260.60   297 $388.58    26 Jan. 2014
##  3 Arizona              $303.31   1941 $209.35  1625 $189.45   222 Jan. 2014
##  4 Arkansas             $361.85    576 $185.62   544 $125.87   112 Jan. 2014
##  5 California           $248.78  12096 $193.56 12812 $192.92   778 Jan. 2014
##  6 Colorado             $236.31   2161 $195.29  1728 $213.50   128 Jan. 2014
##  7 Connecticut          $347.90   1294 $273.97  1316 $257.36    91 Jan. 2014
##  8 Delaware             $373.18    347 $226.25   273 $199.88    34 Jan. 2014
##  9 District of Columbia $352.26    433 $295.67   349 $213.72    39 Jan. 2014
## 10 Florida              $306.43   6506 $220.03  5237 $158.26   514 Jan. 2014
## # ... with 602 more rows
```

```
summary(weed_prices)
```

```
##     State               HighQ              HighQN          MedQ
##  Length:612         Length:612         Min.   :   93   Length:612
##  Class :character   Class :character   1st Qu.:  570   Class :character
##  Mode  :character   Mode  :character   Median : 1359   Mode  :character
##                                        Mean   : 2156
##                                        3rd Qu.: 2800
##                                        Max.   :16127
##      MedQN             LowQ              LowQN           Month
##  Min.   :  134.0   Length:612         Min.   :  11.0   Length:612
##  1st Qu.:  508.5   Class :character   1st Qu.:  48.0   Class :character
##  Median : 1249.5   Mode  :character   Median : 133.0   Mode  :character
##  Mean   : 2035.3                      Mean   : 193.2
##  3rd Qu.: 2482.5                      3rd Qu.: 252.2
##  Max.   :18472.0                      Max.   :1127.0
```

As summarized above, this data set contains information on states in the US for prices of marijuana by quality monthly from January of 2014 until December of 2014. There are noticeable missing values, however, for the prices of `LowQ` marijuana ranging from September until December.

Three character variables are indicative of the average of average prices of marijuana by quality in U.S. dollars, corresponding to high, medium, and low respectively. These are the `HighQ`, `MedQ`, and `LowQ` variables within our data set. An example being, in Alabama in Jan. 2014 the average price of high quality marijuana as indicated by `HighQ` was $339.06 while the price of `MedQ` was $198.64 and `LowQ` $149.09. These values range heavily within and between states.

Note: While it is not explicitly mentioned in the data, through personal investigation, inquiry, and inference we have concluded that it is quite possible that these price points represent a quantity of one ounce (28.5g) of marijuana.

The final three variables are all of type `double` and represent the number of reported prices from buyers respective to each quality: high, medium, low. These are the `HighQN`, `MedQN`, and `LowQN` variables. For example, the value **1042** of `HighQN` in Alabama in Jan. 2014 represents that there were 1042 reported prices from buyers of high quality marijuana.

Research Question & Data Dictionary

Given the data available in `weedprices.csv`, important questions to answer are: How does the average cost of marijuana vary across different regions? and How much does the average marijuana price change over time in particular states? To answer these, we intend to address **how do high quality marijuana prices vary on weighted average across the continental United States in 2014?**, as well as **what was the average change in prices of marijuana by quality across each region in the United States for an eight month period in 2014?**

In order to answer the first question, we will create a map that visualizes the entire continental United States with states colored individually with the weighted average price of high quality marijuana in 2014.To answer the second question, we will create a scatter plot that shows values representing the average prices of marijuana colored by quality and faceted by the respective region over and an eight month period in 2014.These analyses will answer our questions, demonstrating both how average prices differ between quality in US regions over time and between states in a single year. In order to do this we will use the following variables.

Variable Name	Description
State	State or territory name
HighQ	Character variable representing average high quality marijuana price
HighQN	Double variable representing number of reported prices from buyers of high quality
MedQ	Character variable representing average medium quality marijuana price
MedQN	Double variable representing number of reported prices from buyers of medium quality
LowQ	Character variable representing average low quality marijuana price
LowQN	Double variable representing number of reported prices from buyers of low quality
Month	Character value indicating month and year

```
d_dict <- read_excel("./data_dictionary.xlsx")
d_dict %>%
  kbl() %>%
  kable_styling()
```

Data Cleaning Methods

To better serve the data set we are intending to tidy it in the following manner:

1. Use `pivot_longer()` to reshape the data so as to see prices of marijuana respective of quality (`HighQ`,`MedQ`,`LowQ`) in `avg_chart`.
2. Splitting the `Month` variable into `month` and `year` using the `separate()` function then re-coding month abbreviations to numbers using `recode()`.
3. Cleaning names using `janitor:: clean_names()` for all data set variables.
4. Identify, create, and remove missing or `NA` values within `LowQ` variable for plotting purposes.
5. Cleaning strings in the marijuana quality variables (`HighQ`,`MedQ`,`LowQ`) using the `stringr` package and regular expressions to remove unnecessary punctuation.
6. Coercing (`HighQ`, `MedQ`, `LowQ`) variables from type character to type double using the `as.numeric()` function.

Data Cleaning

```
weed_price_cleaned <- weed_prices %>%
  separate(Month, into = c("month", "year"), sep = " ") %>% #separate month into month, year
  mutate(across(c(HighQ, MedQ, LowQ), ~ as.numeric(str_remove_all(.x, "\\$")))) #mutate across price va

weed_price_cleaned
```

```
## # A tibble: 612 x 9
##    State               HighQ HighQN  MedQ MedQN  LowQ LowQN month year
##    <chr>               <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>
##  1 Alabama              339.   1042  199.   933  149.   123 Jan.  2014
##  2 Alaska               289.    252  261.   297  389.    26 Jan.  2014
##  3 Arizona              303.   1941  209.  1625  189.   222 Jan.  2014
##  4 Arkansas             362.    576  186.   544  126.   112 Jan.  2014
```

```
##  5 California            249.  12096  194. 12812  193.   778 Jan.  2014
##  6 Colorado              236.   2161  195.  1728  214.   128 Jan.  2014
##  7 Connecticut           348.   1294  274.  1316  257.    91 Jan.  2014
##  8 Delaware              373.    347  226.   273  200.    34 Jan.  2014
##  9 District of Columbia  352.    433  296.   349  214.    39 Jan.  2014
## 10 Florida               306.   6506  220.  5237  158.   514 Jan.  2014
## # ... with 602 more rows
```

Our first order of cleaning of the data requires us to create a two variables out of the original `Month` variable found in `weed_prices`. This creates a `month` and `year` variable that are still respective of observations. Next, we first mutate across our price variables `HighQ,MedQ,LowQ` to use the `stringr` package and regex syntax to remove dollar signs in the variable names. Once completed we continue to mutate these variables by coercing them to type `numeric` for our data manipulation purposes.

```
weed_price_cleaned2 <- weed_price_cleaned %>%
  mutate(month = recode(month, "Jan." = "01", "Feb." = "02", "Mar." = "03", "Apr." = "04", "May." = "05
  mutate(State = tolower(State)) %>%
  janitor::clean_names()

weed_price_cleaned2
```

```
## # A tibble: 612 x 9
##    state               high_q high_qn med_q med_qn low_q low_qn month year
##    <chr>                <dbl>   <dbl> <dbl>  <dbl> <dbl>  <dbl> <chr> <chr>
##  1 alabama               339.    1042  199.    933  149.    123 01    2014
##  2 alaska                289.     252  261.    297  389.     26 01    2014
##  3 arizona               303.    1941  209.   1625  189.    222 01    2014
##  4 arkansas              362.     576  186.    544  126.    112 01    2014
##  5 california            249.   12096  194.  12812  193.    778 01    2014
##  6 colorado              236.    2161  195.   1728  214.    128 01    2014
##  7 connecticut           348.    1294  274.   1316  257.     91 01    2014
##  8 delaware              373.     347  226.    273  200.     34 01    2014
##  9 district of columbia  352.     433  296.    349  214.     39 01    2014
## 10 florida               306.    6506  220.   5237  158.    514 01    2014
## # ... with 602 more rows
```

This chunk is used to mutate the newly created `month` variable and re-code it to its corresponding number i.e., Jan. = 01, Feb., = 02 etc. for plotting purposes. We continue mutating, this time on the state variable to make all values lower case for joining purposes later on with our `map` data. Finally, we use the `janitor` package to clean all variable names.

Data Transformation

```
regionlist <- list(
  northeast = c("maine", "new hampshire", "vermont", "massachusetts", "new york", "connecticut", "rhode

  midwest = c("north dakota", "south dakota", "minnesota", "wisconsin", "michigan", "ohio", "indiana",

  southeast = c("virginia", "west virginia", "north carolina", "south carolina", "kentucky", "tennessee

  southwest = c("arizona", "new mexico", "oklahoma","texas"),
```

```
  west = c("washington", "oregon", "california", "idaho", "nevada", "utah", "wyoming", "montana", "colo
)

regionframe <- stack(regionlist)
colnames(regionframe)<- c("state", "region") #specify column names

head(regionframe)
```

```
##           state    region
## 1         maine northeast
## 2 new hampshire northeast
## 3       vermont northeast
## 4 massachusetts northeast
## 5      new york northeast
## 6   connecticut northeast
```

We create a list object called `regionlist` to include each region of interest which is equal to a character vector of respective state names. We then convert this list object to a data frame and label column names as state and region. This is for plotting purposes later on.

```
year_avg <- weed_price_cleaned2 %>%
  select(1:3, month) %>% #only state, high_q,high_qn,month
  group_by(state) %>% #show by state
  summarize(statemean = weighted.mean(high_q, w = high_qn)) #weighted by num of reported prices

year_avg
```

```
## # A tibble: 51 x 2
##    state                statemean
##    <chr>                    <dbl>
##  1 alabama                   340.
##  2 alaska                    289.
##  3 arizona                   301.
##  4 arkansas                  350.
##  5 california                246.
##  6 colorado                  238.
##  7 connecticut               343.
##  8 delaware                  368.
##  9 district of columbia      349.
## 10 florida                   303.
## # ... with 41 more rows
```

```
USA_states <- map_data("state") #state map data

weed_map <- USA_states %>%
  left_join(year_avg, by = c("region" = "state")) #join using adjusted `by`

head(weed_map)
```

```
##        long      lat group order  region subregion statemean
## 1 -87.46201 30.38968     1     1 alabama      <NA>  340.0571
```

```
## 2 -87.48493 30.37249     1     2 alabama      <NA>  340.0571
## 3 -87.52503 30.37249     1     3 alabama      <NA>  340.0571
## 4 -87.53076 30.33239     1     4 alabama      <NA>  340.0571
## 5 -87.57087 30.32665     1     5 alabama      <NA>  340.0571
## 6 -87.58806 30.32665     1     6 alabama      <NA>  340.0571
```

For the first output, we begin by creating an object called `year_avg` that is grouped by state and is summarized by a `statemean` variable we create by taking a weighted mean of high quality marijuana prices weighted by the number of reported prices for that quality of marijuana.

Note: When creating this weighted average, given that the value of the (`high_q`,`med_q`,`low_q`) variables are already averages of averages we will be referring to this weighted average of an average of averages simply as a weighted average with the purpose of ease of understanding in the following descriptions.

The second output utilizes `map_data()` for all the continental United States in a `USA_states` object. We compile this map data with a `left_join()` of the year_avg to have complete state map data associated with respective weighted averages of high quality marijuana prices.

```
weedregions <- weed_price_cleaned2 %>%
  left_join(regionframe, by = "state") %>%
  select(1:8, region) #select on state all weed data and region
weedregions
```

```
## # A tibble: 612 x 9
##    state                high_q high_qn med_q med_qn low_q low_qn month region
##    <chr>                 <dbl>   <dbl> <dbl>  <dbl> <dbl>  <dbl> <chr> <fct>
##  1 alabama                339.    1042  199.    933  149.    123 01    southeast
##  2 alaska                 289.     252  261.    297  389.     26 01    west
##  3 arizona                303.    1941  209.   1625  189.    222 01    southwest
##  4 arkansas               362.     576  186.    544  126.    112 01    southeast
##  5 california             249.   12096  194.  12812  193.    778 01    west
##  6 colorado               236.    2161  195.   1728  214.    128 01    west
##  7 connecticut            348.    1294  274.   1316  257.     91 01    northeast
##  8 delaware               373.     347  226.    273  200.     34 01    northeast
##  9 district of columbia   352.     433  296.    349  214.     39 01    northeast
## 10 florida                306.    6506  220.   5237  158.    514 01    southeast
## # ... with 602 more rows
```

```
month_average <- weedregions %>%
  filter(!is.na(low_q)) %>% #remove na in lowq
  group_by(region, month) %>%
  summarize(
    high_avg = weighted.mean(high_q, w = high_qn),
    med_avg = weighted.mean(med_q, w = med_qn),
    low_avg = weighted.mean(low_q, w = low_qn))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
month_average #filtered NA values out only has 8 months
```

```
## # A tibble: 40 x 5
## # Groups:   region [5]
```

```
##    region    month high_avg med_avg low_avg
##    <fct>     <chr>    <dbl>   <dbl>   <dbl>
##  1 northeast 01        359.    275.    213.
##  2 northeast 02        359.    275.    213.
##  3 northeast 03        358.    274.    214.
##  4 northeast 04        356.    274.    219.
##  5 northeast 05        355.    274.    221.
##  6 northeast 06        354.    273.    219.
##  7 northeast 07        354.    272.    220.
##  8 northeast 08        353.    272.    220.
##  9 midwest   01        347.    260.    180.
## 10 midwest   02        345.    260.    180.
## # ... with 30 more rows
```

The first output of this code chunk is data frame `weedgregions` that is a copy of our cleaned data set and joins it with `regionframe` by state to add a region respective of the state variable. We select all prices of marijuana across quality, region, and state.

The second output utilizes `weedregions` to create a monthly weighted average by region and month. We first filter out all `NA` values within the `low_q` price variable which removes the last 4 months of the calendar year. We summarize these state and month groupings similarly to above by calculating weighted averages by number of reported prices by quality of marijuana.

```
avg_chart <- month_average %>%
  pivot_longer(
    cols = c(high_avg, low_avg, med_avg), #quality variable respective of price
    names_to = "quality",
    values_to = "price"
  )

avg_chart
```

```
## # A tibble: 120 x 4
## # Groups:   region [5]
##    region    month quality  price
##    <fct>     <chr> <chr>    <dbl>
##  1 northeast 01    high_avg  359.
##  2 northeast 01    low_avg   213.
##  3 northeast 01    med_avg   275.
##  4 northeast 02    high_avg  359.
##  5 northeast 02    low_avg   213.
##  6 northeast 02    med_avg   275.
##  7 northeast 03    high_avg  358.
##  8 northeast 03    low_avg   214.
##  9 northeast 03    med_avg   274.
## 10 northeast 04    high_avg  356.
## # ... with 110 more rows
```

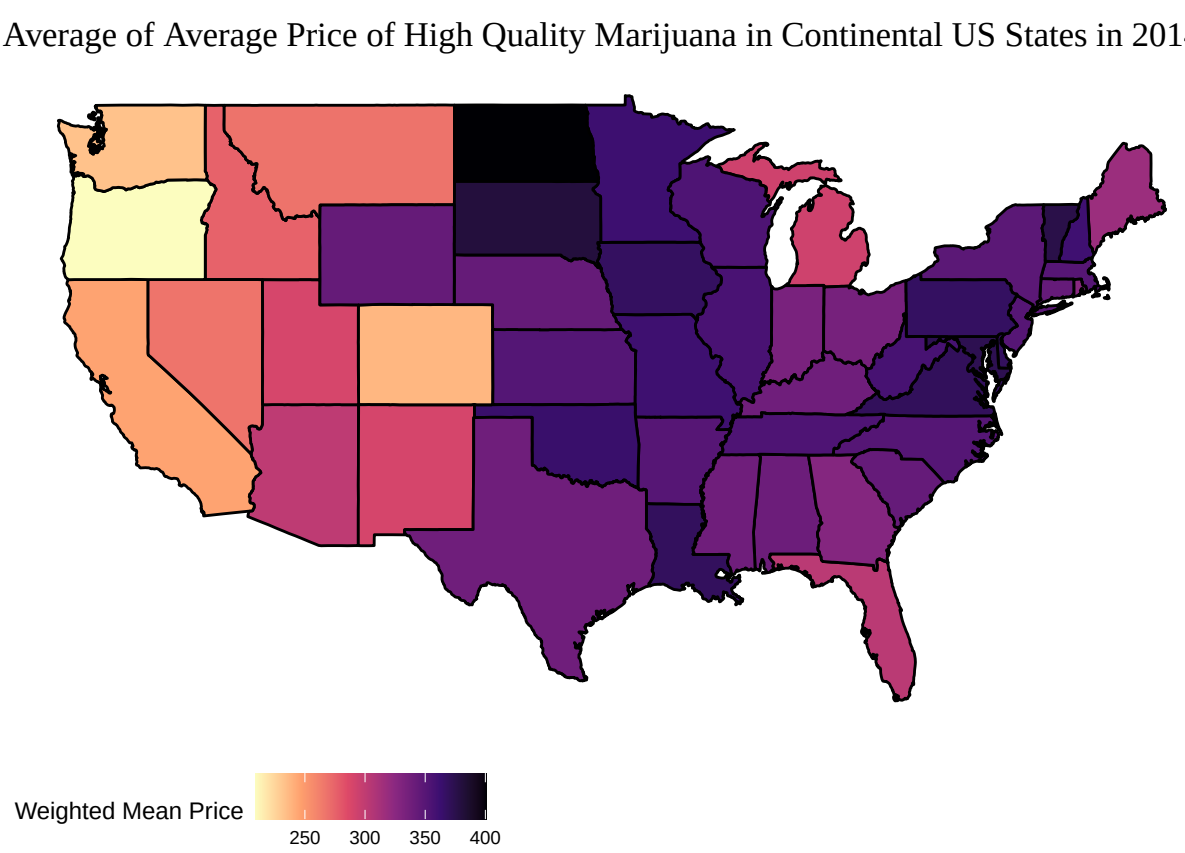
We created a final dataframe for our charting purposes called `avg_chart` that would be pivoted longer (`pivot_longer()`) to include a quality variable for each price in each month by region.

Data Visualization

High Quality Marijuana Prices in the United States

A plot that displays a map of the United States at the state level colored by the weighted average of average price of high quality marijuana for the year 2014.

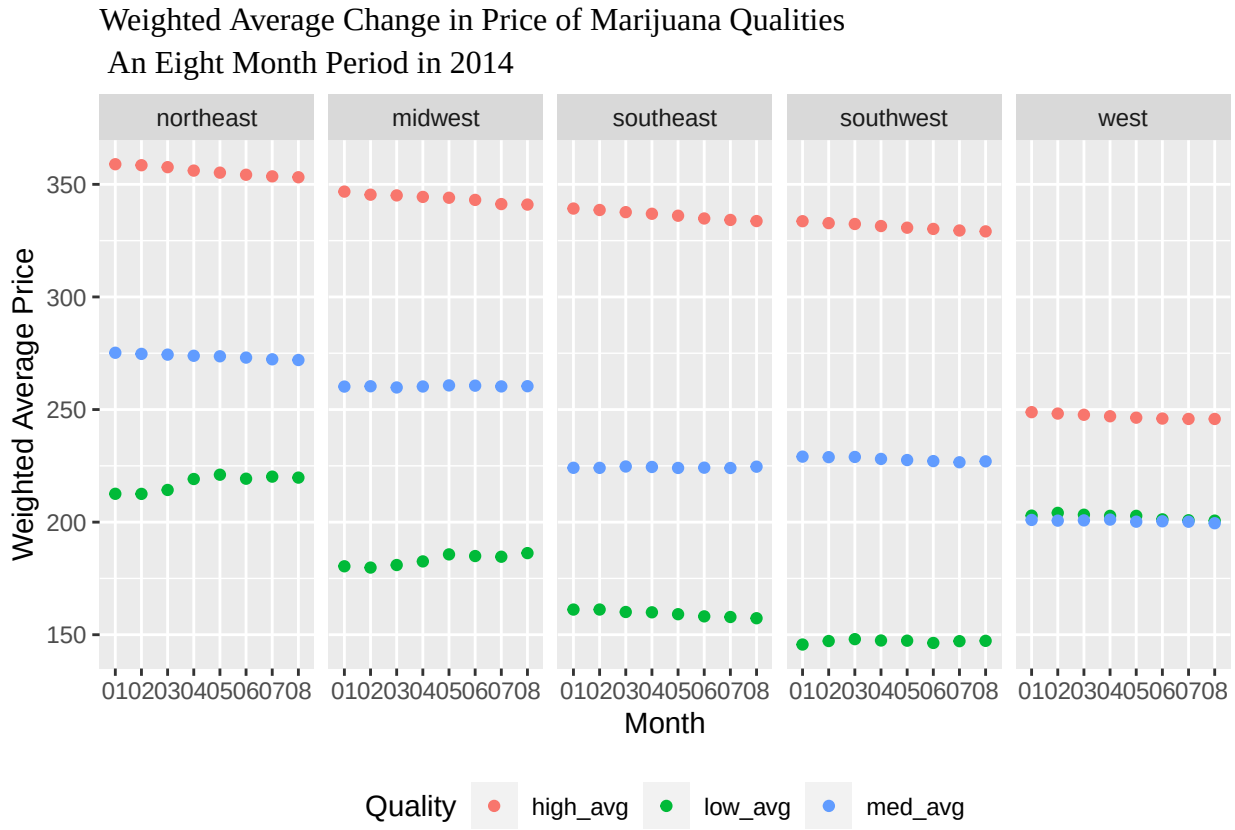
```
ggplot(weed_map, aes(long, lat, fill = statemean)) +
  geom_polygon(aes(group=group),
  color="black") +
  coord_fixed(1.3) +
  ggtitle("Average of Average Price of High Quality Marijuana in Continental US States in 2014") +
  labs(fill = "Weighted Mean Price") +
  theme_map() +
  theme(legend.position = "bottom") +
  theme(plot.title = element_text(size = 12.5, family = "serif")) +
  scale_fill_viridis_c(option = "magma", direction = -1)
```



From this first visualization, it is possible to see how the states largely differ in their respective price. The west has substantially lower prices and greater variation in prices than the mid-west, east, and south. The darkest coloring representative of highest prices can be found in North Dakota, Lousiana, and Vermont. The entire east coast tends to be around $300-$400 in price as opposed the west coast that boasts averages closer to $200-$250.

Marijuana Price Trends in the US by Region and Quality

```
avg_chart$quality <- as.factor(avg_chart$quality)
ggplot(avg_chart, mapping = aes(x = month, y = price, color = quality))+
  geom_point() +
  facet_wrap(~ region, nrow = 1) +
  labs(title = "Weighted Average Change in Price of Marijuana Qualities", subtitle = " An Eight Month Pe
  theme(plot.title = element_text(size = 11.5, family = "serif"),
        plot.subtitle = element_text(size = 11.5, family = "serif")) +
  labs(color = "Quality") +
  theme(legend.position = "bottom")
```



This plot gives the viewer a comprehensive understanding of the changes experienced within and between American regions respective of weed quality. There is a visible negative trend in prices for high quality weed over time and across regions, whereas both medium and low quality weighted average prices stay about constant. The northeast contains the highest high, medium, and low quality price for weed in comparison to other regions. Interestingly, the west contains the lowest average price over time for high quality weed but an almost identical trend in prices for both medium and low quality weed.

Prices of Low Quality Marijuana in The West

```
west_states<- c("washington", "oregon", "california", "idaho", "nevada", "utah", "wyoming", "montana",
```

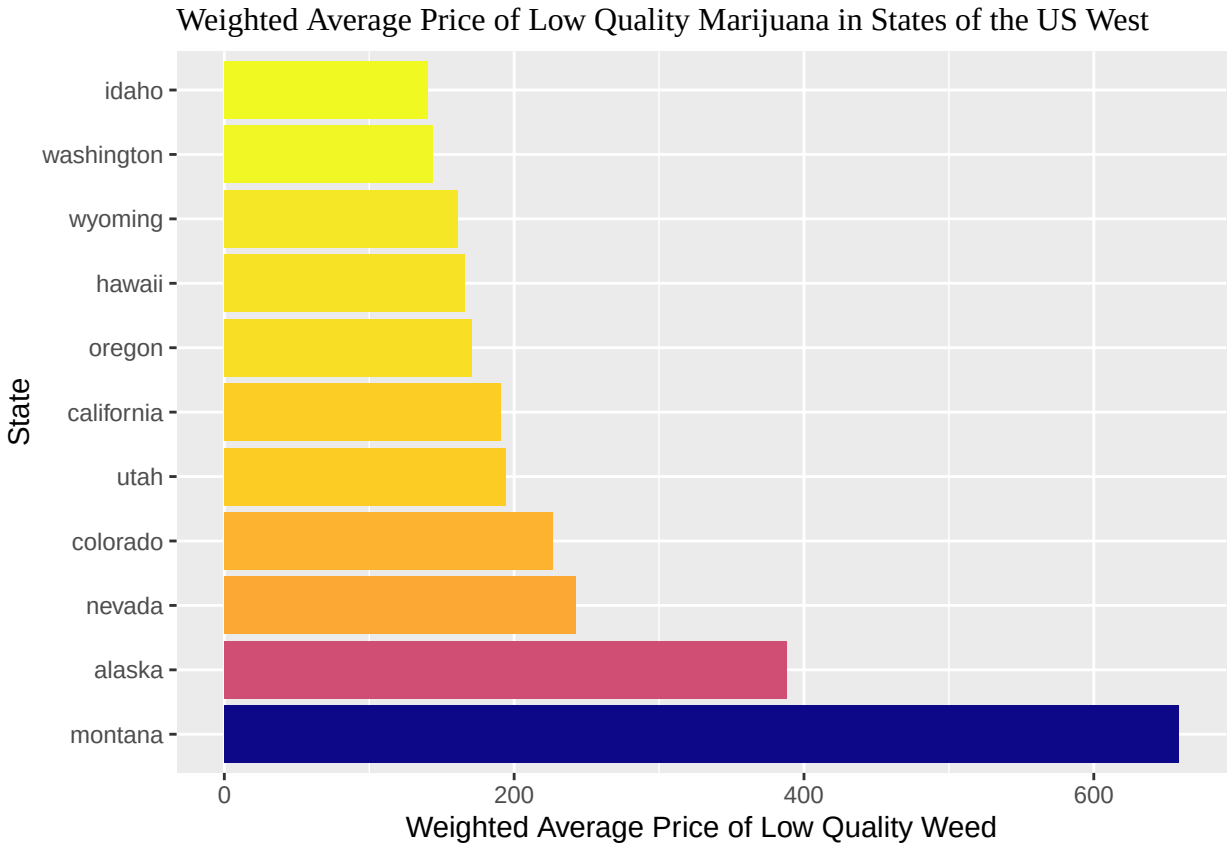
```
pricetable<- weedregions %>%
  select(state, 6:9) %>%
  filter(!is.na(low_q),
         state %in% west_states) %>%
  group_by(state) %>%
  summarize(lowmean = weighted.mean(low_q, w = low_qn)) %>%
  arrange(desc(lowmean))

pricetable
```

Supplementary Visual for Price Trends by Region

```
## # A tibble: 11 x 2
##    state      lowmean
##    <chr>        <dbl>
##  1 montana       659.
##  2 alaska        388.
##  3 nevada        242.
##  4 colorado      227.
##  5 utah          194.
##  6 california    191.
##  7 oregon        171.
##  8 hawaii        166.
##  9 wyoming       161.
## 10 washington    144.
## 11 idaho         140.
```

In the above map we saw that states in the West had many of the lowest means of weighted average prices for high quality marijuana, and yet in the following graph we demonstrated that despite having the lowest weighted average prices at that quality, the Western region has among the highest weighted average prices for low quality marijuana. This is surprising, so it is important to demonstrate how this occurred. To do so, we will take the `weedregions` frame we used earlier, filter it to isolate states in the West, and create a plot of the weighted mean prices of low quality marijuana for each state. This, hopefully, will provide some context to the above results.

```
ggplot(pricetable, mapping = aes(x = reorder(state, -lowmean), y = lowmean, fill = lowmean))+
  geom_col() +
  labs(title = "Weighted Average Price of Low Quality Marijuana in States of the US West",
       x = "State", y = "Weighted Average Price of Low Quality Weed") +
  labs(fill = "") +
  theme(legend.position = "none") +
  theme(plot.title = element_text(size = 11.5, family = "serif")) +
  coord_flip() +
  scale_fill_viridis(option = "plasma", direction = -1)
```



As the plot highlights, Montana and Alaska have far higher weighted average prices of low quality marijuana than the rest of the states in the West. Montana particularly, which has a fairly low `high_avg` value, has a low quality marijuana mean of average prices (`statemean`) of over $600. Alaska has a value of just under $400. These greatly impact the mean price for the Western region, which would otherwise be about $180. This helps to explain the effect we see in the faceted graph of the low quality mean of averages in the West, where several `low_q` and `med_q` points intersect, which does align with the general trend of weed price in western states.

Conclusion

High quality marijuana prices vary quite heavily on weighted across the continental United States in 2014. From our choropleth map we demonstrated that states in the Midwest such as North and South Dakota have high weighted average prices approaching 400 U.S. dollars per ounce. By contrast, many western states such as Oregon and Washington have weighted average prices that are closer to 200-250 U.S. dollars per ounce. This is a disparity of almost 200 U.S. dollars. The east coast weighted average prices are about 350 U.S. dollars with slight variation in states such Florida, Maine, and Michigan with prices closer to 300 U.S. dollars. Generally, western states tend to have lower weighted average prices of marijuana per ounce and trends in this weighted average appear to be regional across the entire continental United States.

Within each region of the United States there is little to no change in the weighted average prices of every quality of marijuana over an eight month period in 2014. Northeastern states tend to have the highest weighted average prices of marijuana across all qualities, while the west tends to have the lowest prices across high and medium qualities with low quality prices being notably high in states of this region. This is caused by the high outlier states of Alaska and Montana in this region. The only noticeable average change in prices over time is the relatively weak negative trend in the high quality marijuana across regions.