Term Project Step 2

Burden Joshua

2022-05-22

Legalization of Marijuana

Introduction

With more and more states adding medical and recreational cannabis to their ballots, questions are now being asked more openly about the impact of cannabis on public health, crime statistics, and popularity of legalization. The data out there is sparse, but the data that has been collected could show trends and insights on the future of legislation and public sentiment of legalization, along with showing the potential for lowered crime and raised tax revenue in states and at the federal level.

Research questions

I want to know, based on the data that is out there, if legalization of marijuana:

- Reduces the number of arrests of all nonviolent offenses significantly
- Would increase the tax income for states and government entities
- Would drug usage go down by making it legal
- What is the state tax collections for legal rec states vs med only vs no rec/med
- What is the public support for legalization

Approach

I will look at the data that is out there for sales, crime, and public sentiment and see if there is anything that stands out to answer any of the questions above.

How your approach addresses (fully or partially) the problem.

The data will tell me whether or not the any or all of my questions can be answered and if there is discrepancies in the data, more research might be conducted to see if there is correlation.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

- https://data.world/denver/marijuana-related-crime
- https://www.kaggle.com/datasets/tunguz/drug-use-by-age
- https://www.kaggle.com/datasets/mykeysw/marijuana-sales-forecasting-in-tx
- https://data.world/sya/marijuana-laws-by-state
- https://data.world/denver/marijuana-gross-sales
- https://data.world/health/support-for-legal-marijuana
- https://www.kaggle.com/datasets/terenceshin/historical-prices-for-biggest-weed-stocks

Bad dataset

- https://data.world/opensavannah/cannabis-justice

Required Packages

- Gplot2

Plots and Table Needs

- Histograms
- Scatter plots
- CDF
- Linear Regressions

Questions for future steps

Is there more data out there that would help with this research, and how deep should these models go in terms of covering the questions?

Step 2

How to import and clean my data

Google is very good actually for making suggestions on cleaning datasets. That is my first step, by bringing these datasets into google sheets, I am effectively having google help me get the datasets clean. One other thing that I needed to do was to also parse what is important and what isn't based on the datasets. There were quite a number of datasets that were not helpful or would be considered a large reach for the dataset to fit into the story, so I removed those datasets from the list of datasets I had. I also made the decision early on to remove header column names that either didn't make sense or were not required for the dataset story.

What does the final data set look like?

Summaries of the dataset, as to not get overwelmed by the vast amounts of data.

```
setwd('/Users/joshua/Documents/PERSONAL_GITHUB_REPOS/dsc520/TermProject')
cosalesrev <- read.csv("data/cleaned_COMJSalesMonthReports2019.csv")
cotaxrev <- read.csv("data/cleaned_COMJTaxMonthReports2019.csv")
costatepop <- read.csv("data/cleaned_COCountyPop20142018.csv")
drug_use <- read.csv("data/cleaned_drug-use-by-age.csv")
crime_denver <- read.csv("data/cleaned_crime_marijuana.csv")
legal_support <-read.csv("data/cleaned_legal_marijuana_support.csv")
```

Summaries of datasets

```
# summaries of the datasets
summary(cosalesrev)
```

```
##     Month            MedicalSales       MSYearToDate        RetailSales
##  Length:65          Min.   :24082927   Min.   : 25680596   Min.   : 14022213
##  Class :character   1st Qu.:29238678   1st Qu.:100238704   1st Qu.: 44590853
##  Mode  :character   Median :32686869   Median :193094556   Median : 76018423
##                     Mean   :32736868   Mean   :204611449   Mean   : 70423116
##                     3rd Qu.:36010893   3rd Qu.:302926736   3rd Qu.: 96642441
##                     Max.   :41056948   Max.   :445616062   Max.   :114317739
##   RSYearToDate         TotalSales         TSYearToDate
##  Min.   :1.402e+07   Min.   : 45987045   Min.   :4.656e+07
##  1st Qu.:1.599e+08   1st Qu.: 78199969   1st Qu.:2.831e+08
##  Median :3.088e+08   Median :109222331   Median :5.399e+08
##  Mean   :4.086e+08   Mean   :103159984   Mean   :6.132e+08
##  3rd Qu.:5.775e+08   3rd Qu.:127706166   3rd Qu.:8.809e+08
##  Max.   :1.214e+09   Max.   :143107279   Max.   :1.546e+09
##   TotalToDate
##  Min.   :4.656e+07
##  1st Qu.:1.047e+09
##  Median :2.644e+09
##  Mean   :2.895e+09
##  3rd Qu.:4.612e+09
##  Max.   :6.705e+09
```

```
summary(cotaxrev)
```

```
##     Month              SalesTax        LicenseFees       SalesTaxFees
##  Length:65          Min.   : 628947   Min.   : 592661   Min.   :1570401
##  Class :character   1st Qu.:1014752   1st Qu.: 972122   1st Qu.:2137462
##  Mode  :character   Median :1808419   Median :1063563   Median :3067394
##                     Mean   :1970529   Mean   :1097654   Mean   :3068184
##                     3rd Qu.:2763721   3rd Qu.:1221521   3rd Qu.:4016545
##                     Max.   :3692930   Max.   :1663120   Max.   :4892115
##  RetailSalesTax     RetailExciseTax    TotalTaxFees        YearToDate
##  Min.   : 1401568   Min.   : 195318   Min.   : 3519756   Min.   :  3519756
##  1st Qu.: 4394550   1st Qu.:2796865   1st Qu.:10856584   1st Qu.: 37511919
##  Median : 7746575   Median :4683825   Median :17694953   Median : 76306924
##  Mean   : 8877802   Mean   :4114956   Mean   :16060942   Mean   : 93255239
##  3rd Qu.:14608085   3rd Qu.:5598581   3rd Qu.:21622509   3rd Qu.:134971077
##  Max.   :18698640   Max.   :7867853   Max.   :26841073   Max.   :266529637
##   TotalToDate
##  Min.   :3.520e+06
##  1st Qu.:1.260e+08
##  Median :3.555e+08
##  Mean   :4.161e+08
##  3rd Qu.:6.818e+08
##  Max.   :1.044e+09
```

```
summary(costatepop)
```

```
##     County             X2014Pop         X2015Pop         X2016Pop
##  Length:64          Min.   :   703   Min.   :   689   Min.   :   690
##  Class :character   1st Qu.:  5692   1st Qu.:  5736   1st Qu.:  5636
##  Mode  :character   Median : 14288   Median : 14358   Median : 14592
##                     Mean   : 83526   Mean   : 85076   Mean   : 86472
##                     3rd Qu.: 41946   3rd Qu.: 41975   3rd Qu.: 42592
##                     Max.   :664715   Max.   :683081   Max.   :696347
##     X2017Pop         X2018Pop
##  Min.   :   714   Min.   :   762
##  1st Qu.:  5837   1st Qu.:  5876
##  Median : 14747   Median : 15014
##  Mean   : 87648   Mean   : 88993
##  3rd Qu.: 43202   3rd Qu.: 43666
##  Max.   :705651   Max.   :716492
```

```
summary(drug_use)
```

```
##      age             sample_size   marijuana_use_by_percentage
##  Length:17          Min.   :2223   Min.   : 1.10
##  Class :character   1st Qu.:2469   1st Qu.: 8.70
##  Mode  :character   Median :2798   Median :20.80
##                     Mean   :3251   Mean   :18.92
##                     3rd Qu.:3058   3rd Qu.:28.40
##                     Max.   :7391   Max.   :34.00
##  marijuana_frequency_over_12_months
##  Min.   : 4.00
##  1st Qu.:30.00
##  Median :52.00
##  Mean   :42.94
##  3rd Qu.:52.00
##  Max.   :72.00
```

```
summary(crime_denver)
```

```
##   REPORTDATE         OFFENSE_CODE  OFFENSE_TYPE_ID    OFFENSE_CATEGORY_ID
##  Length:1254        Min.   :1006   Length:1254        Length:1254
##  Class :character   1st Qu.:2203   Class :character   Class :character
##  Mode  :character   Median :2203   Mode  :character   Mode  :character
##                     Mean   :2249
##                     3rd Qu.:2206
##                     Max.   :7399
##  NEIGHBORHOOD_ID
##  Length:1254
##  Class :character
##  Mode  :character
##
##
##
```

```
summary(legal_support)
```

```
##       Year         Month           Asked_half_sample    Yes_Legal
##  Min.   :1969   Length:20          Length:20          Min.   :12.00
##  1st Qu.:1980   Class :character   Class :character   1st Qu.:25.00
##  Median :2002   Mode  :character   Mode  :character   Median :34.00
##  Mean   :1997                                         Mean   :35.95
##  3rd Qu.:2011                                         3rd Qu.:48.50
##  Max.   :2016                                         Max.   :60.00
##    No_Illegal      No_Opinion   Percent_Yes         Percent_No
##  Min.   :39.00   Min.   :1.00   Length:20          Length:20
##  1st Qu.:49.25   1st Qu.:2.00   Class :character   Class :character
##  Median :63.00   Median :4.00   Mode  :character   Mode  :character
##  Mean   :60.50   Mean   :3.45
##  3rd Qu.:70.75   3rd Qu.:4.25
##  Max.   :84.00   Max.   :6.00
##  Percent_No_Opinion
##  Length:20
##  Class :character
##  Mode  :character
##
##
##
```

glimpse of datasets

```
glimpse(cosalesrev)
```

```
## Rows: 65
## Columns: 8
## $ Month        <chr> "Jan 2014", "Feb 2014", "Mar 2014", "Apr 2014", "May 2014~
## $ MedicalSales <int> 32541720, 31738572, 34821878, 32686869, 31355208, 2995030~
## $ MSYearToDate <int> 32541720, 64280292, 99102170, 131789039, 163144247, 19309~
## $ RetailSales  <int> 14022213, 14248473, 19881631, 20765986, 21375001, 2397808~
## $ RSYearToDate <int> 14022213, 28270686, 48152317, 68918303, 90293304, 1142713~
## $ TotalSales   <int> 46563933, 45987045, 54703509, 53452855, 52730209, 5392839~
## $ TSYearToDate <int> 46563933, 92550978, 147254487, 200707342, 253437551, 3073~
## $ TotalToDate  <dbl> 46563933, 92550978, 147254487, 200707342, 253437551, 3073~
```

```
glimpse(cotaxrev)
```

```
## Rows: 65
## Columns: 9
## $ Month           <chr> "Feb 2014", "Mar 2014", "Apr 2014", "May 2014", "Jun 2~
## $ SalesTax        <int> 1330209, 1460429, 1569405, 1559710, 1569454, 1530968, ~
## $ LicenseFees     <int> 592661, 857615, 902995, 761687, 940028, 1547853, 13795~
## $ SalesTaxFees    <int> 1922870, 2318044, 2472400, 2321397, 2509482, 3078821, ~
## $ RetailSalesTax  <int> 1401568, 1434916, 1898685, 2217607, 2070577, 2473627, ~
## $ RetailExciseTax <int> 195318, 339615, 609907, 734351, 1135648, 969637, 13979~
## $ TotalTaxFees    <int> 3519756, 4092575, 4980992, 5273355, 5715707, 6522085, ~
## $ YearToDate      <int> 3519756, 7612330, 12593322, 17866677, 23582384, 301044~
## $ TotalToDate     <int> 3519756, 7612330, 12593322, 17866677, 23582384, 301044~
```

```
glimpse(costatepop)
```

```
## Rows: 64
## Columns: 6
## $ County   <chr> "Adams", "Alamosa", "Arapahoe", "Archuleta", "Baca", "Bent", ~
## $ X2014Pop <int> 479477, 15758, 617498, 12240, 3576, 5777, 312588, 61617, 1845~
## $ X2015Pop <int> 489774, 15854, 628951, 12401, 3544, 5885, 318071, 64713, 1857~
## $ X2016Pop <int> 497419, 16006, 637266, 12839, 3522, 5664, 321363, 66399, 1907~
## $ X2017Pop <int> 503375, 16056, 643257, 13316, 3539, 5866, 322854, 68169, 1962~
## $ X2018Pop <int> 511868, 16683, 651215, 13765, 3585, 5882, 326078, 69267, 2002~
```

```
glimpse(drug_use)
```

```
## Rows: 17
## Columns: 4
## $ age                                <chr> "12", "13", "14", "15", "16", "17",~
## $ sample_size                        <int> 2798, 2757, 2792, 2956, 3058, 3038,~
## $ marijuana_use_by_percentage        <dbl> 1.1, 3.4, 8.7, 14.5, 22.5, 28.0, 33~
## $ marijuana_frequency_over_12_months <int> 4, 15, 24, 25, 30, 36, 52, 60, 60, ~
```

```
glimpse(crime_denver)
```

```
## Rows: 1,254
## Columns: 5
## $ REPORTDATE          <chr> "2/27/2012", "8/6/2012", "9/18/2012", "8/19/2012",~
## $ OFFENSE_CODE        <int> 2203, 2203, 2203, 5707, 2203, 2203, 2203, 2203, 22~
## $ OFFENSE_TYPE_ID     <chr> "BURGLARY - BUSINESS BY FORCE", "BURGLARY - BUSINE~
## $ OFFENSE_CATEGORY_ID <chr> "Burglary", "Burglary", "Burglary", "All Other Cri~
## $ NEIGHBORHOOD_ID     <chr> "montclair", "five-points", "hampden-south", "mont~
```

```
glimpse(legal_support)
```

```
## Rows: 20
## Columns: 9
## $ Year               <int> 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 200~
## $ Month              <chr> "Oct", "Oct", "Oct", "Oct", "Nov", "Oct", "Oct", "O~
## $ Asked_half_sample  <chr> "no", "no", "no", "no", "no", "no", "yes", "yes", "~
## $ Yes_Legal          <int> 60, 58, 51, 58, 48, 50, 46, 44, 36, 34, 34, 31, 25,~
## $ No_Illegal         <int> 39, 40, 47, 39, 50, 46, 50, 54, 60, 64, 62, 64, 73,~
## $ No_Opinion         <int> 1, 2, 2, 3, 1, 3, 4, 2, 4, 2, 4, 5, 2, 4, 5, 5, 6, ~
## $ Percent_Yes        <chr> "60%", "58%", "51%", "58%", "48%", "51%", "46%", "4~
## $ Percent_No         <chr> "39%", "40%", "47%", "39%", "51%", "46%", "50%", "5~
## $ Percent_No_Opinion <chr> "1%", "2%", "2%", "3%", "1%", "3%", "4%", "2%", "4%~
```

Questions for future steps.

I am in a situation where there is almost too much data, but not enough connections as there

What information is not self-evident?

I was going to add this datasets to other states, but for many states there is not an aggregate set of information out there for just marijuana related offenses, and without being specific, it becomes difficult to see patterns and trends when the net is cast to wide.

What are different ways you could look at this data?

Data like this is not very connected. Making a story out of the datasets requires looking at each dataset as a piece of a puzzle and not trying to force the datasets to work with each other but rather, answer a question and check if the answer relates to the next question.

How do you plan to slice and dice the data?

For one set of data, I plan to see the sales trend for Colorado based on sales, and tax data. I will also look at if crime went up specifically in Denver, and then see if there is another data set out there to see if there is a relationship with higher sales, with uprising crime and youth uses.

What types of plots and tables will help you to illustrate the findings to your questions?

barcharts and lineplots seem to make the most sense in these instances. I might also look at doing some data plotting with a scatterplot.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I am unsure at this time I will add a machine learning algorithm to these datasets. I would like to say yes, but I am still parsing data, and narrowing my search queries to something smaller than my original scope.

Questions for future steps.

Is there more data out there that I just have not found yet that might give me a better and more up to date dataset for the questions I am posing?