Sunday Sales

X

DUI

ITP 449
Spring 2021

Exploratory Data Analysis Project

Frances Lawley

Introduction

Sunday laws (Blue laws) originally banned alcohol sales on Sundays to protect worker's families and to observe a day of rest.

Though not enforced equally across the United States, many participated in strict prohibited Sunday sales while other states permitted local breweries for carryout.

This dataset can be used to analyze the correlation between different types of Sunday Sales and an increase in DUI or vehicle-related fatalities due to alcohol.

State	DUI	Fatalities	Population	Sunday.Sales
50 unique values	386 – 141k	16 – 1323	586k – 39.1m	Local 38% Permitted 34% Other (14) 28%
Alabama	7863	247	4858979	Prohibited
Alaska	3163	24	738432	Permitted
Arizona	22367	272	6828065	Permitted
Arkansas	6919	149	2978204	Local
California	141458	914	39144818	Permitted
Colorado	25562	151	5456574	Permitted
Connecticut	8148	103	3590886	Local
Delaware	386	42	945934	Permitted
Florida	31783	797	20271272	Local
Georgia	19217	365	10214860	Local

- [Dataset] -

-[Initial Inquiries]-

Sunday Sales

Are DUIs higher in areas with permitted Sunday Sales compared to any other type of Sunday Sale (local, restricted, prohibited)?

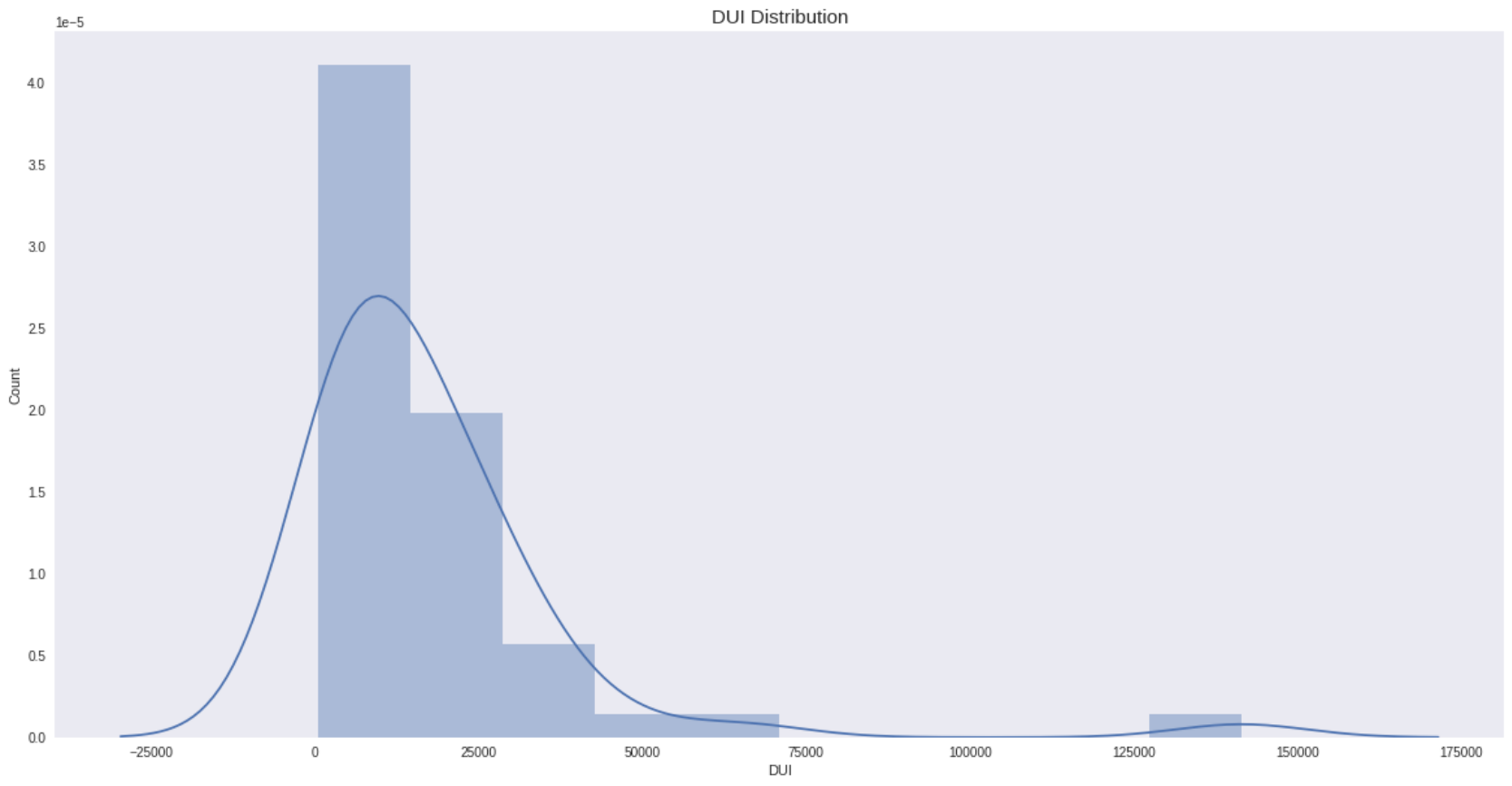
Population

How can different levels of population relate to the number of DUIs and vehicle-related deaths due to alcohol?

Region

Are DUIs higher in states of a certain region? Based on region, is it likely DUIs and alcohol-related deaths will be higher?

-[Dui Distribution]-



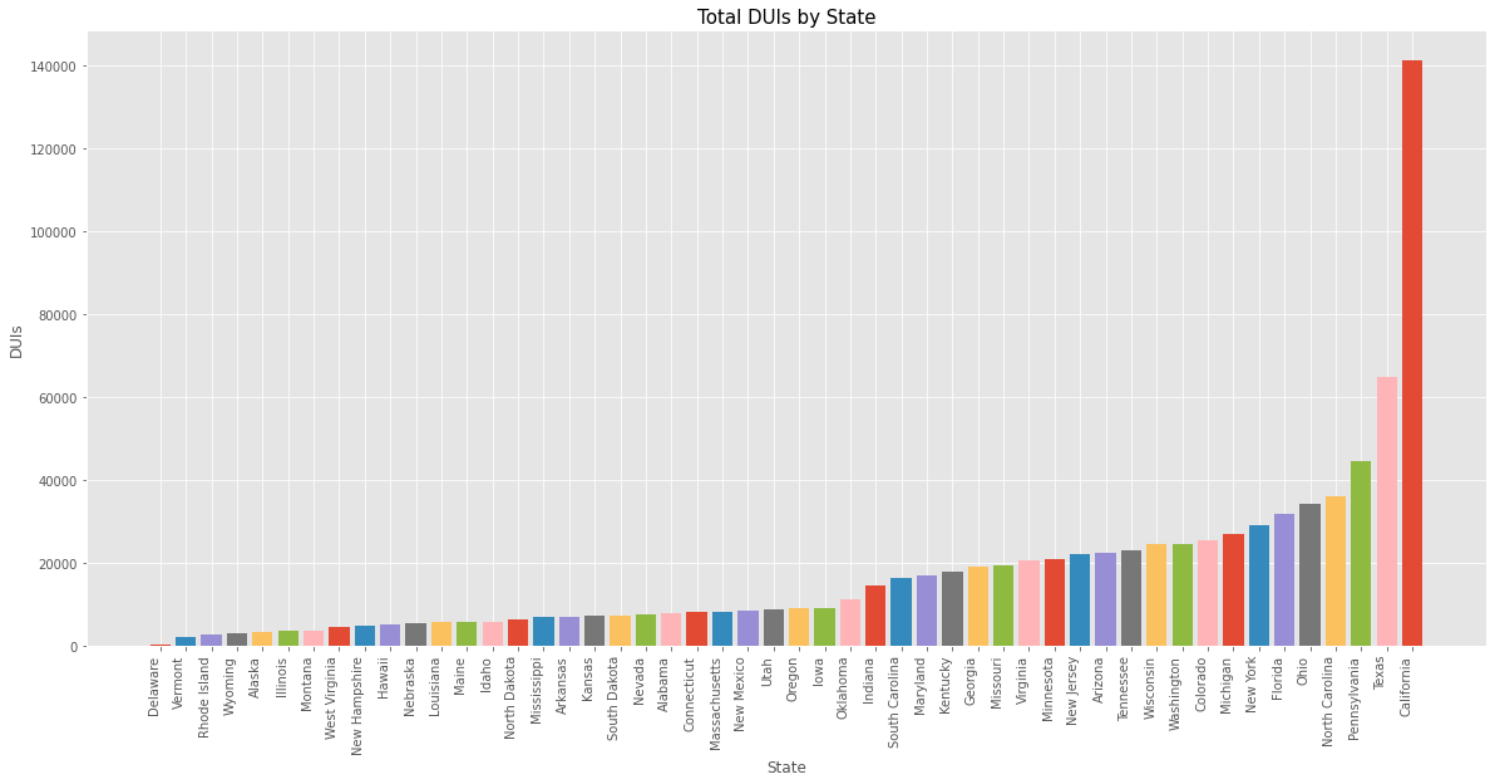
Average number of DUIs is 17,317

Most of the DUIs range from 0 to 27,500

50% of states have less than 8,916 DUIs

count	50.000000
mean	17317.340000
std	21914.054979
min	386.000000
25%	5778.000000
50%	8916.000000
75%	22325.500000
max	141458.000000

-[Total DUIs]-



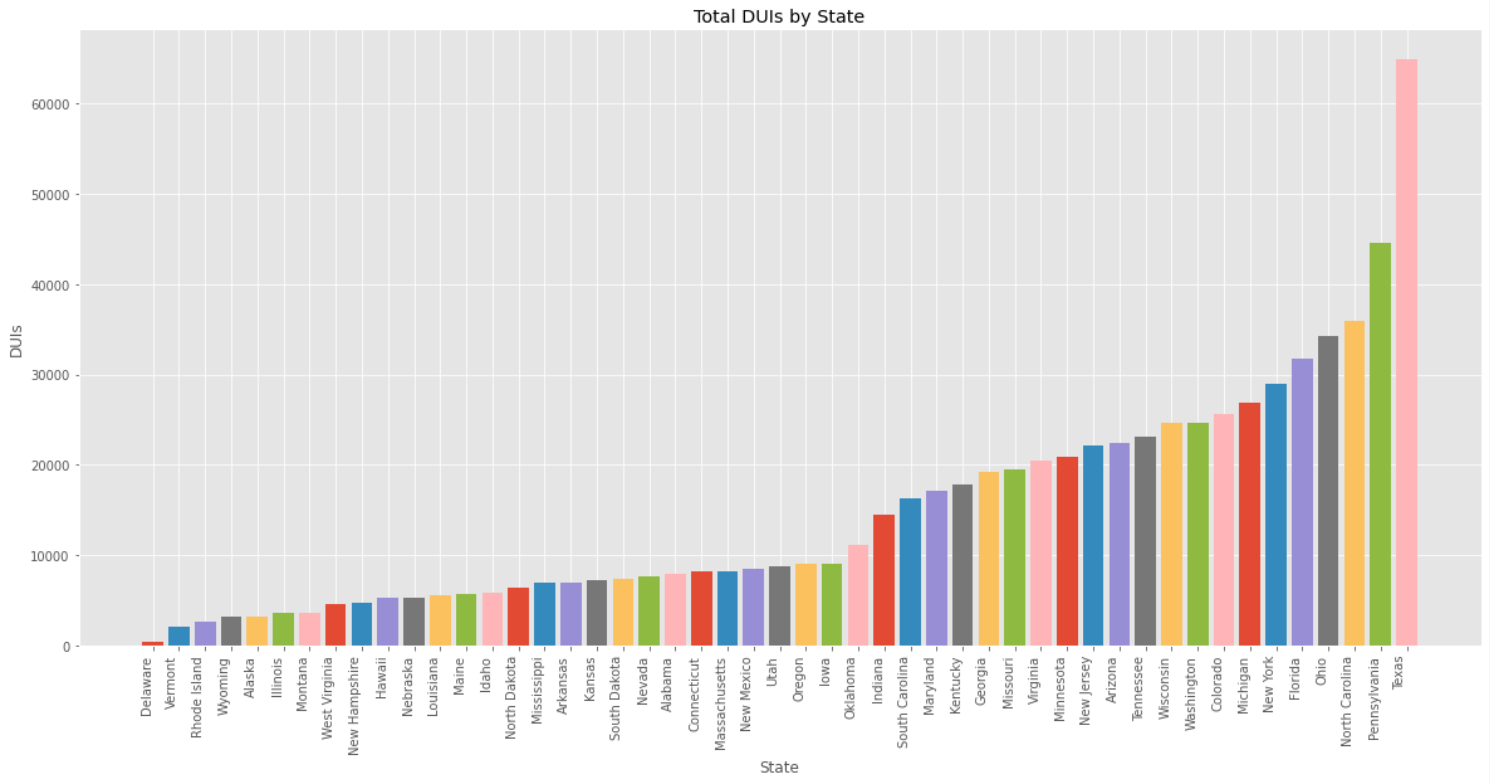
California, Texas, and Pennsylvania yield the highest number of DUIs

Delaware, Vermont, Rhode Island yield the lowest number of DUIs

California is the main outlier at max DUI of 141,458

count	50.000000
mean	17317.340000
std	21914.054979
min	386.000000
25%	5778.000000
50%	8916.000000
75%	22325.500000
max	141458.000000

-[Total DUIs]-



Removing California as the outlier significantly lowers standard deviation from 21914 to 12752

Average DUIs also lowers from 17317 to 14784

Texas receives the second highest DUIs at 64,971

count	49.000000
mean	14783.857143
std	12752.167116
min	386.000000
25%	5756.000000
50%	8813.000000
75%	22201.000000
max	64971.000000

Hypothesis 1

I hypothesize that there is a correlation between states with higher populations and the amount of DUIs reported by drunk drivers.

I hypothesize that there is a correlation between death rate per population and the number of DUIs.

Hypothesis 2

I hypothesize that states with Sunday Sales that are either permitted or local receive more DUIs and fatalities than states with Sunday Sales that are restricted or prohibited.

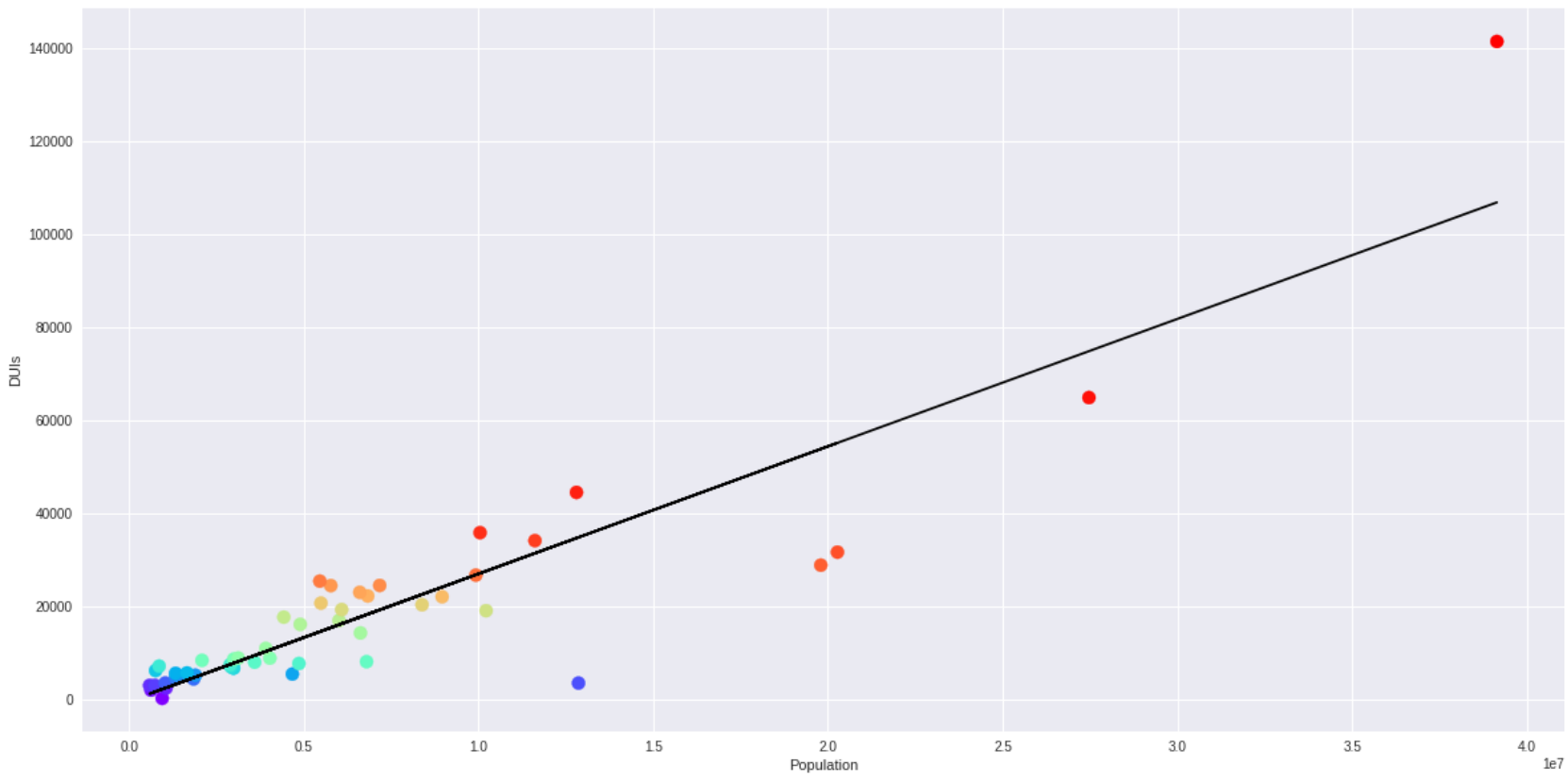
I hypothesize that States with prohibited Sunday Sales receive the least number of DUIs.

Hypothesis 3

I hypothesize that states in the South region have the highest number of DUIs.

I hypothesize that that states with populations below the median population (4,547,908) have fewer DUIs reported by drunk drivers.

-[Correlation Between Population & DUIs]-

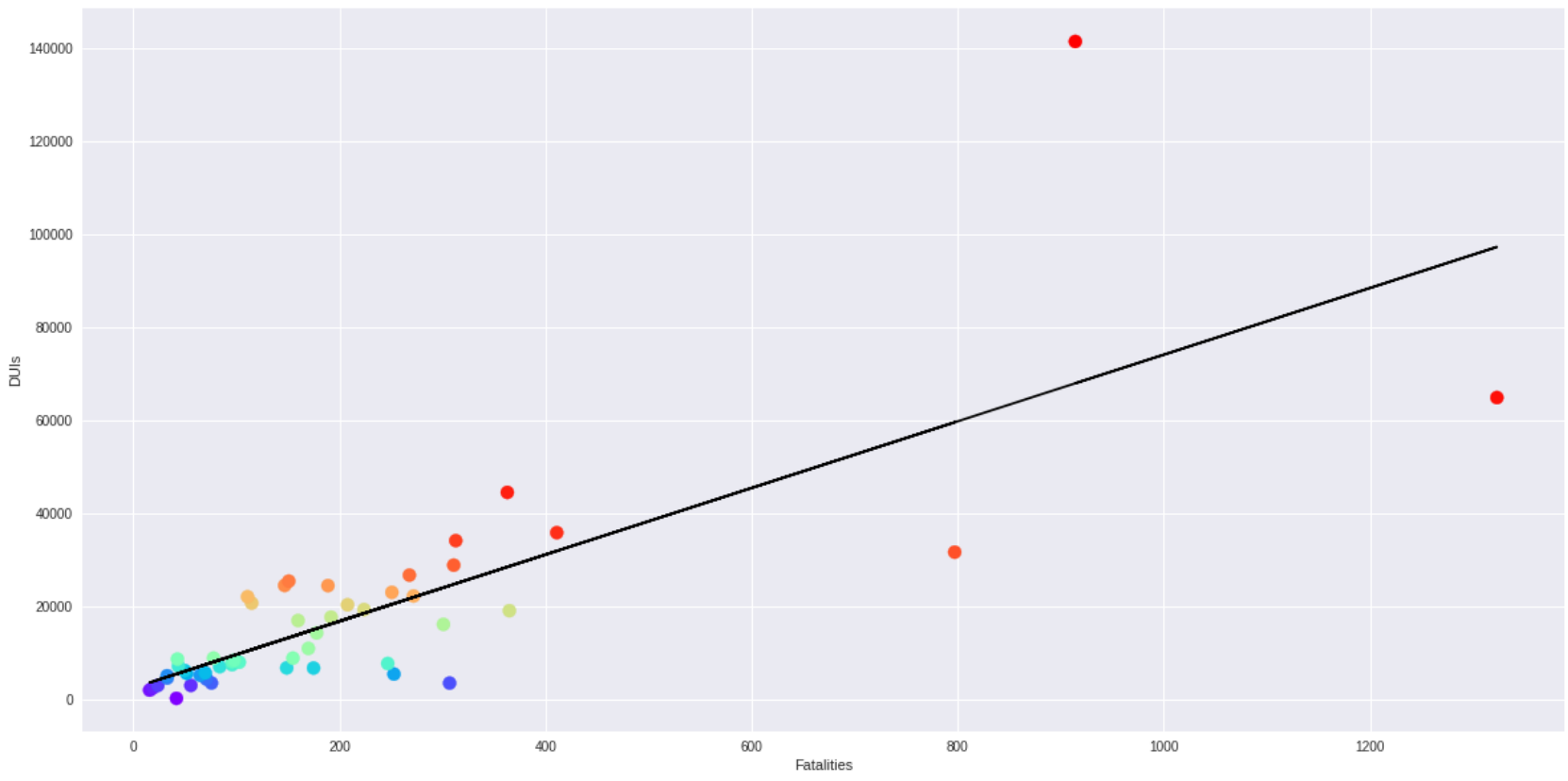


(0.9028849759179716, 3.178718762098287e-19)

There is a strong positive (close to 1) and significant correlation between Population and number of DUIs.

Hypothesis 1

-[Correlation Between Fatalities & DUIs]-

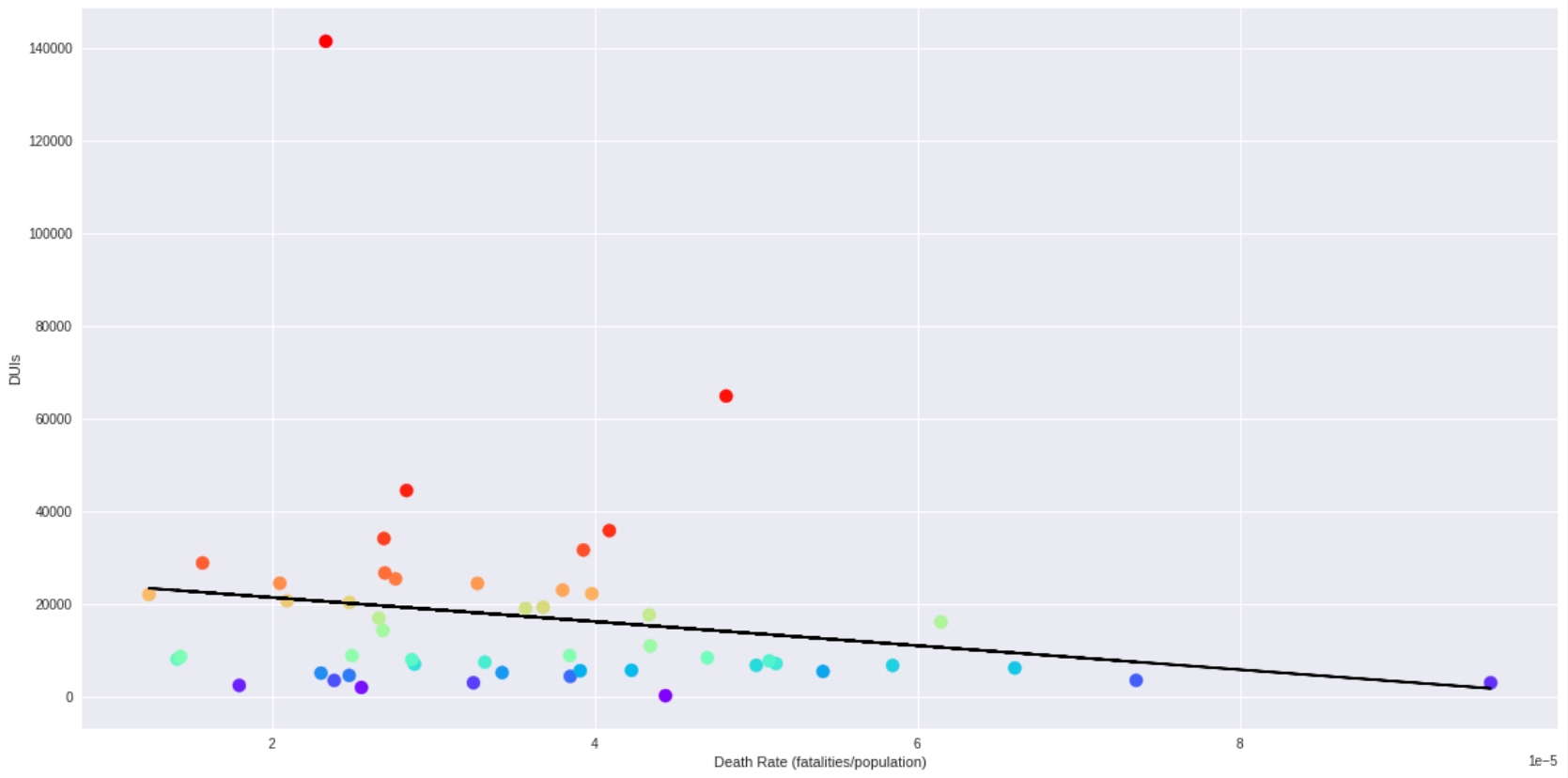


(0.7734952553713388, 4.5574345730745174e-11)

There is a strong positive and significant correlation between Fatalities and number of DUIs.

Hypothesis 1

-[Correlation Between Death Rate & DUIs]-



(0.7734952553713388, 4.5574345730745174e-11)

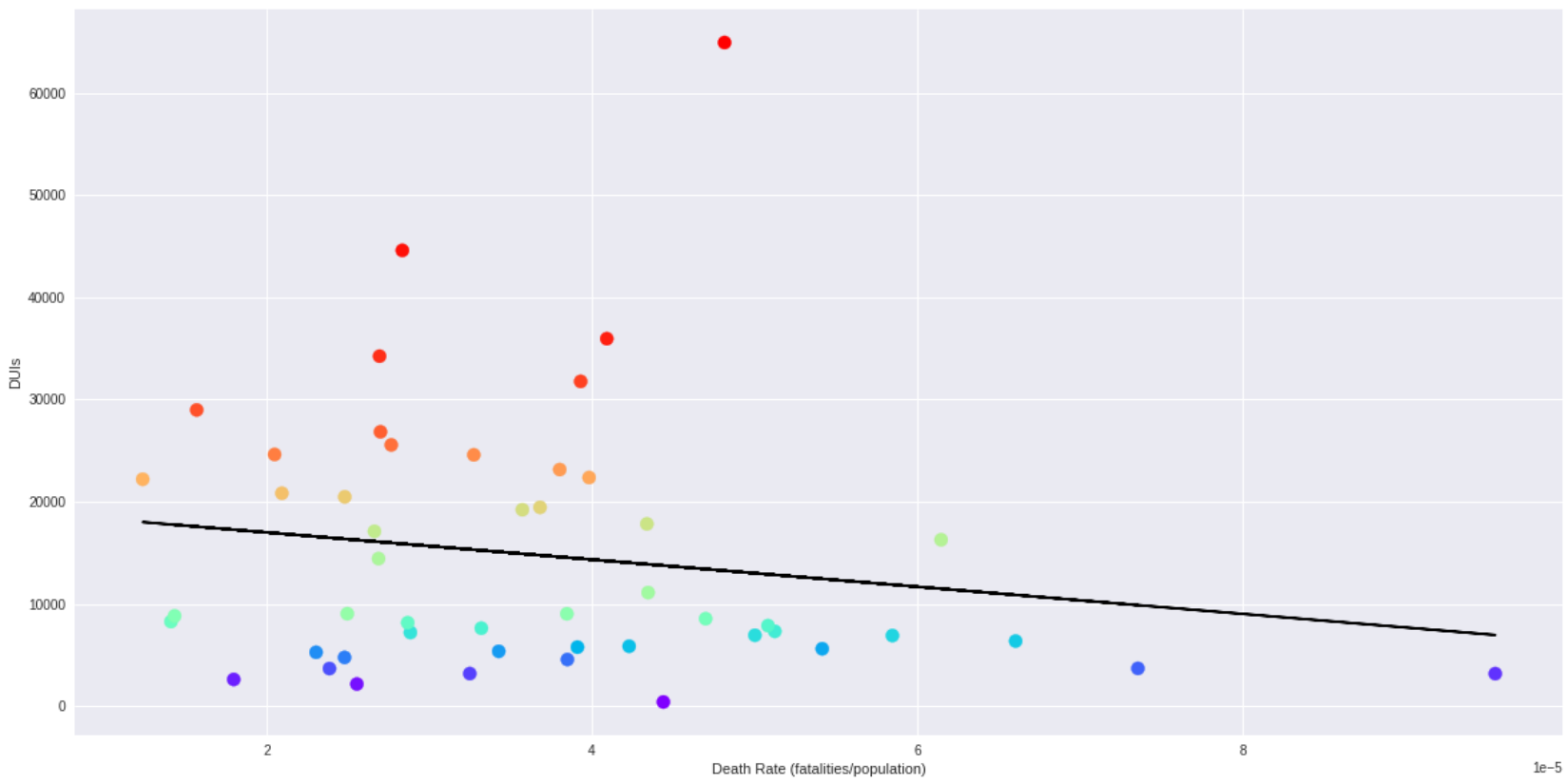
Based on the statistics, there is a strong positive and significant correlation between Death Rate (fatalities/population) and number of DUIs

With a p-value still less than 0.5, we are confident to reject the null hypothesis and accept the alternate hypothesis that there is a correlation between Death Rate and DUIs

However, the graph shows a downward trend, indicating a negative correlation – Why? (further analysis below)

Hypothesis 1

-[Correlation Between Death Rate & DUIs]-



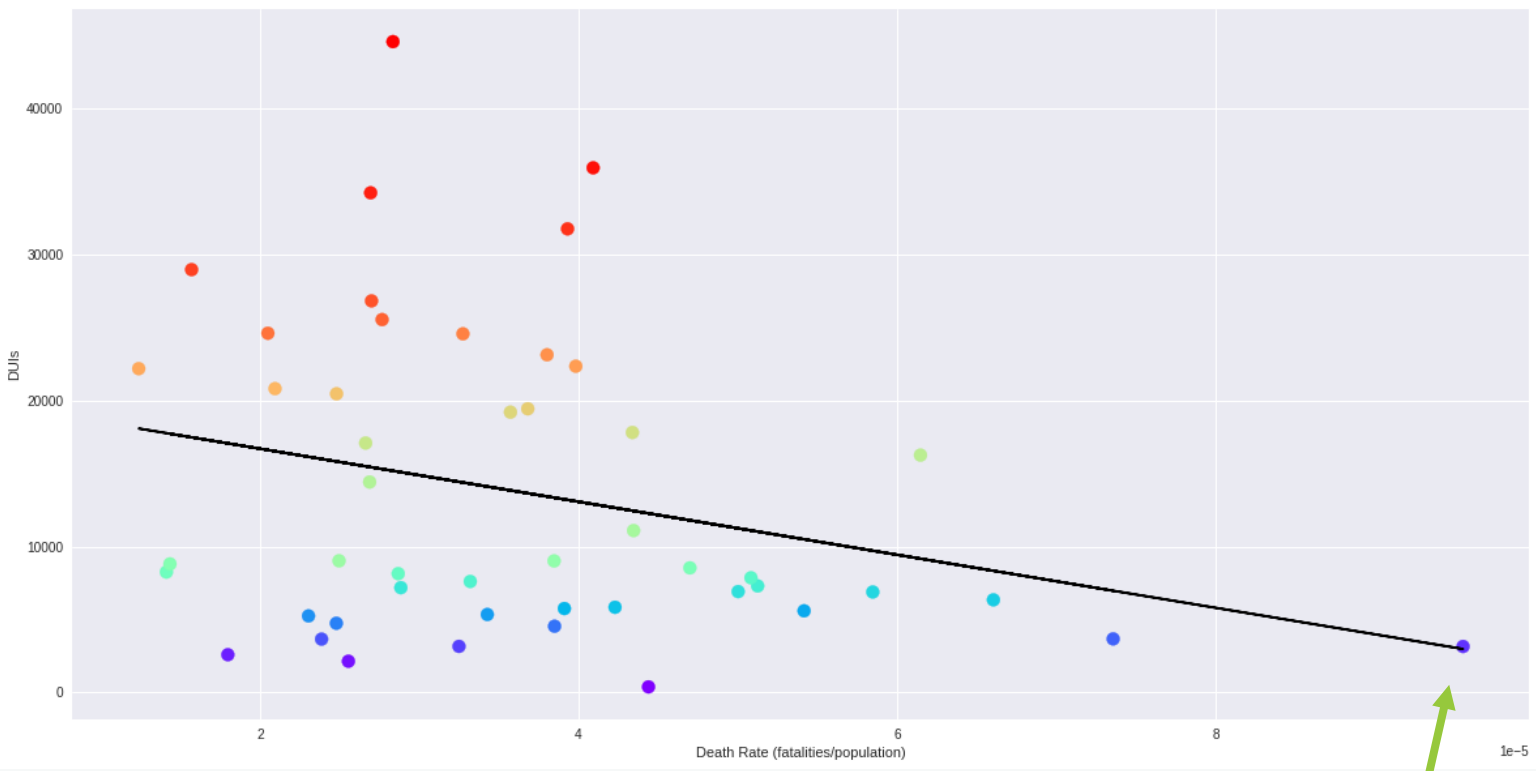
1st Graph: One outlier removed (California)
(0.8093749387080006, 1.936556866774875e-12)

Removing California as an outlier improves our correlation (0.77 raises to 0.81) and lowers our p-value, despite the graph still displaying a downward trend.

California has a large population (39,144,818) relative to fatalities (914) leading to a lower death rate. Its high number of DUIs (max 141,458) therefore offsets this correlation.

Hypothesis 1

-[Correlation Between Death Rate & DUIs]-



2nd Graph: Two outliers removed (California and Texas)

(0.7021804013988964, 2.6726199819089808e-08)

Removing the second outlier, Texas, along with California does not necessarily improve our correlation, but still displays a strong correlation coefficient.

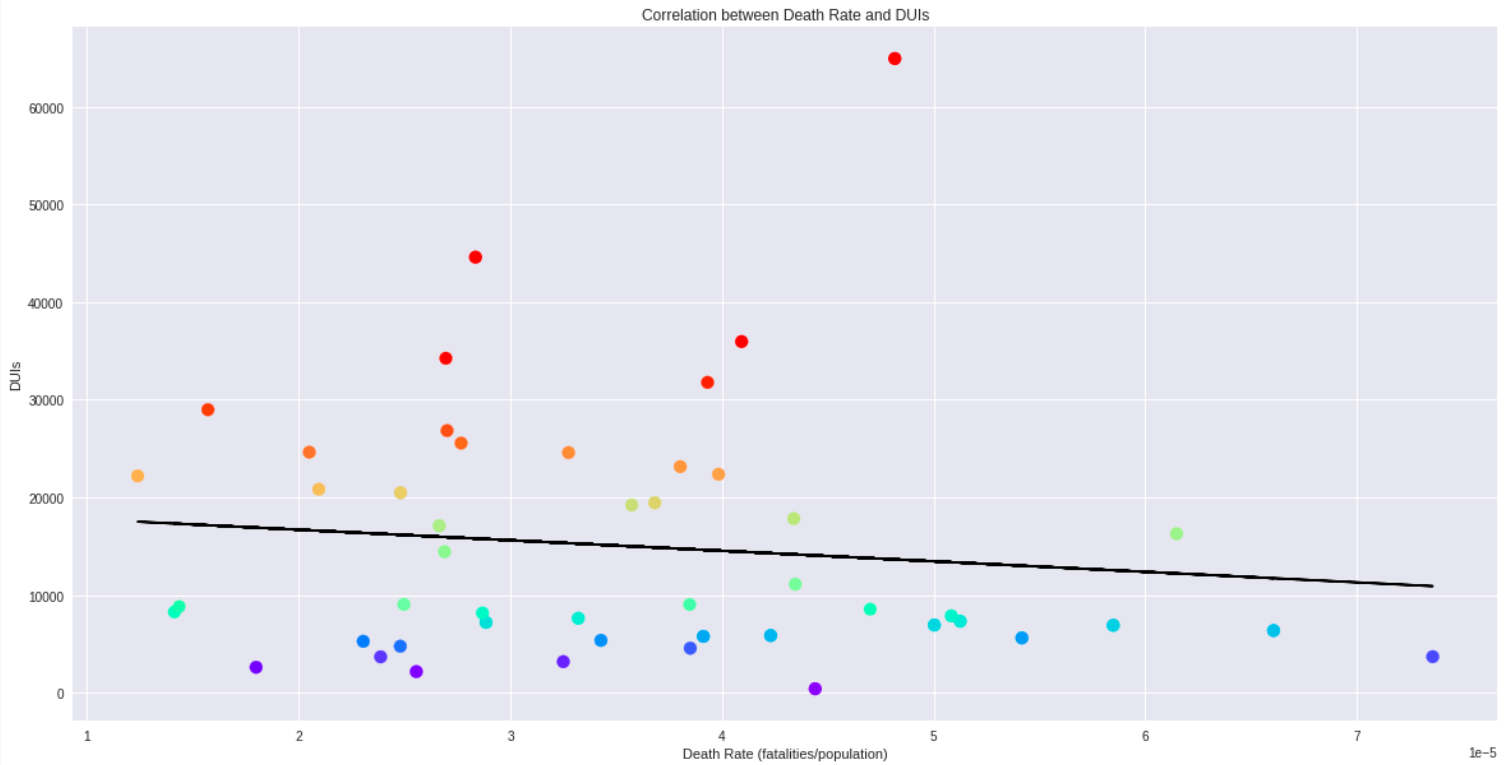
Why is the graph still displaying a downward trend?

What if we removed the bottom right outlier instead (the highest death rate)?

Wyoming displays the highest death rate at 0.000096

Hypothesis 1

-[Correlation Between Death Rate & DUIs]-



2nd Graph: Two outliers removed (California and Wyomg)

(0.8077579816422312, 3.97154847757383e-12)

Removing both outliers of California (highest DUIs) and Wyoming (highest Death Rate) raises the correlation coefficient back to 0.81 and lowers the p-value

With a p-value still less than 0.5, we are confident to reject the null hypothesis and accept the alternate hypothesis that there is a correlation between Death Rate and DUIs

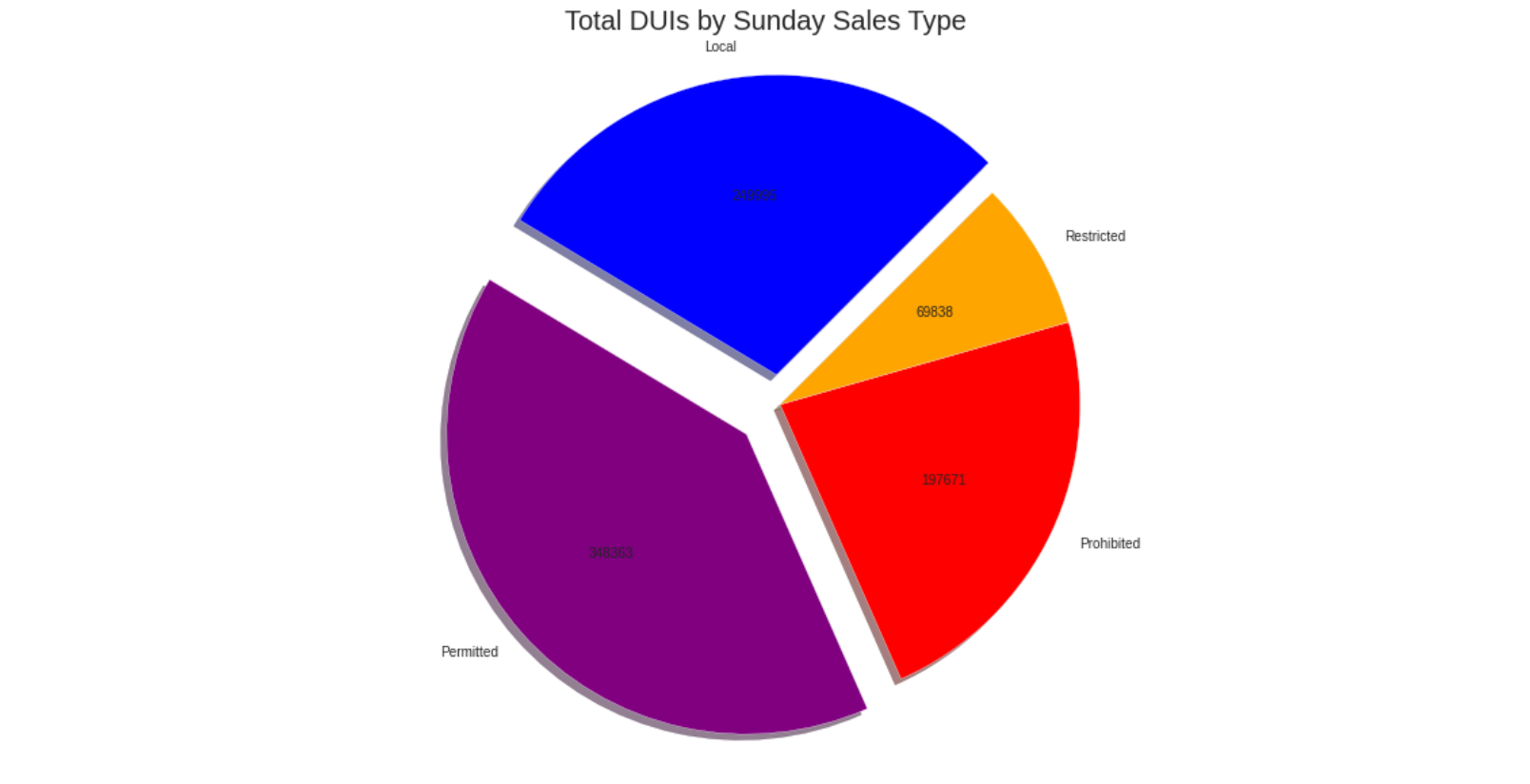
-[Further Analysis]-

States with higher populations yield lower death rates (fatalities/population) as the larger populations offset number of deaths.

States with higher populations generally yield more total DUIs and fatalities (California is an outlier with fatalities = 914)

Hypothesis 2

-[Sunday Sales & DUIs]-



From Hypothesis #2, we can reject that prohibited Sunday Sales states receive the lowest DUIs as Restricted Sunday Sales states yield the lowest total DUIs

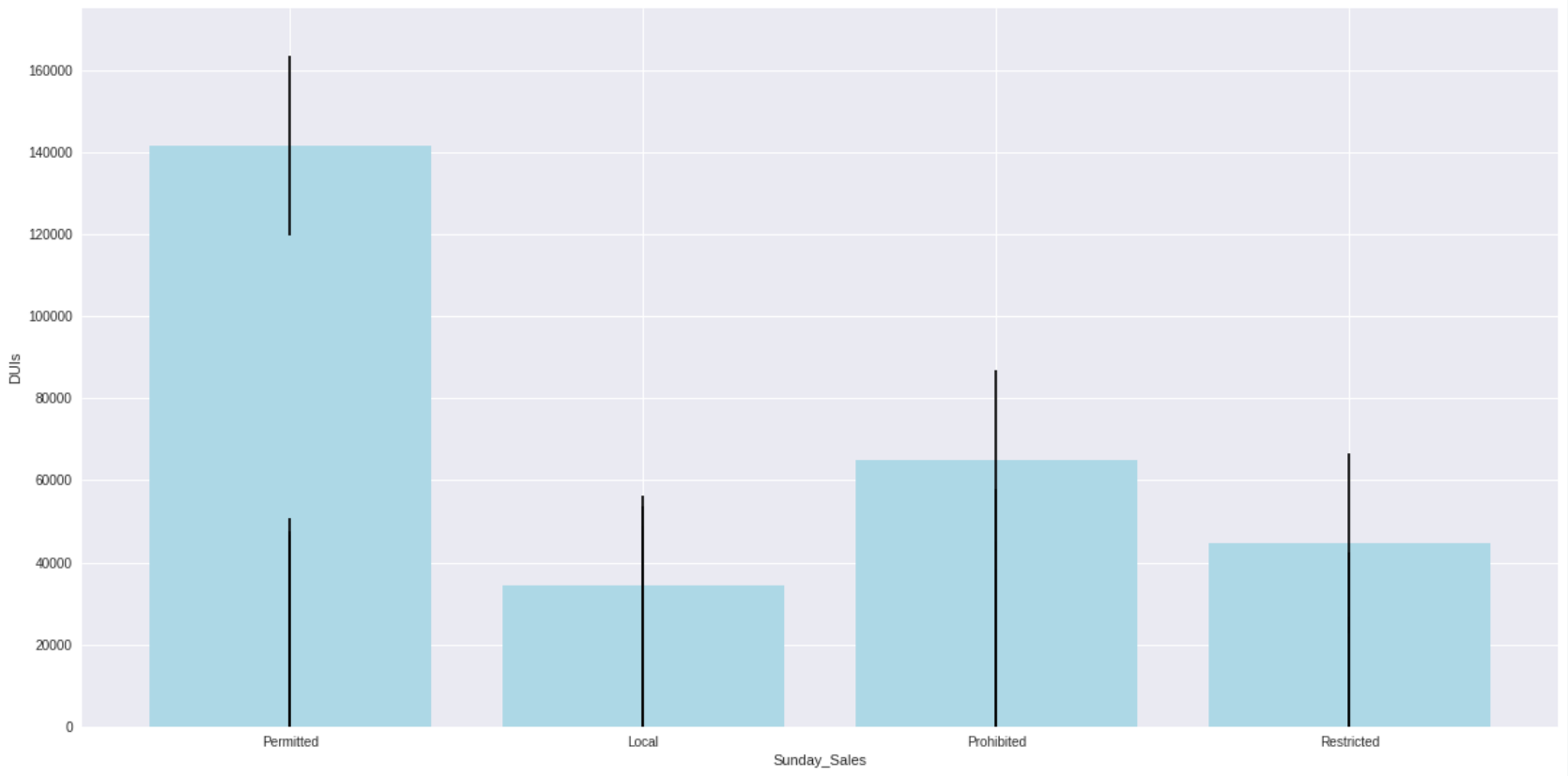
From Hypothesis #2, we can accept that Local and Permitted Sunday Sales types receive the most number of DUIs

Sunday_Sales	DUI	Fatalities	Population	Death_Rate
Local	249995	3621	115359979	0.000685
Permitted	348363	2798	109606822	0.000559
Prohibited	197671	3246	73263687	0.000495
Restricted	69838	604	22516104	0.000078

Hypothesis 2

-[Sunday Sales & DUIs]-

Bar Chart



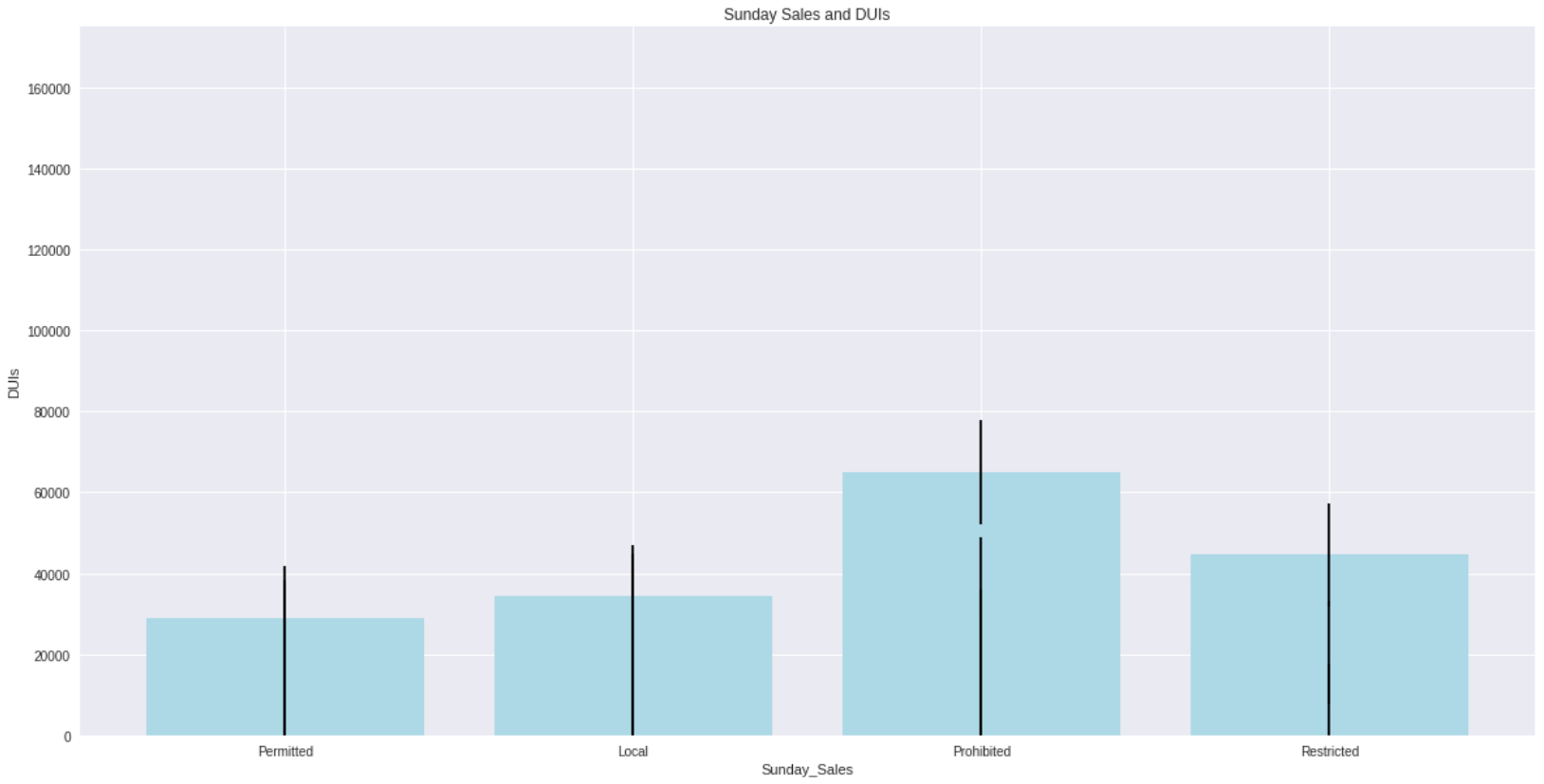
count	50.000000
mean	17317.340000
std	21914.054979
min	386.000000
25%	5778.000000
50%	8916.000000
75%	22325.500000
max	141458.000000

Even though total DUIs are highest in Permitted and Local states, Local Sunday Sales show the lowest spread while Permitted Sunday sales display the largest spread among values of DUIs

Hypothesis 2

-[Sunday Sales & DUIs]-

Bar Chart



count	49.000000
mean	14783.857143
std	12752.167116
min	386.000000
25%	5756.000000
50%	8813.000000
75%	22201.000000
max	64971.000000

When removing the outlier of California (permitted Sunday Sales), we see the the largest spread shift to the Prohibited states mainly due to the second highest outlier of Texas (max DUI 64,971 and prohibited Sunday Sales)

Hypothesis 2

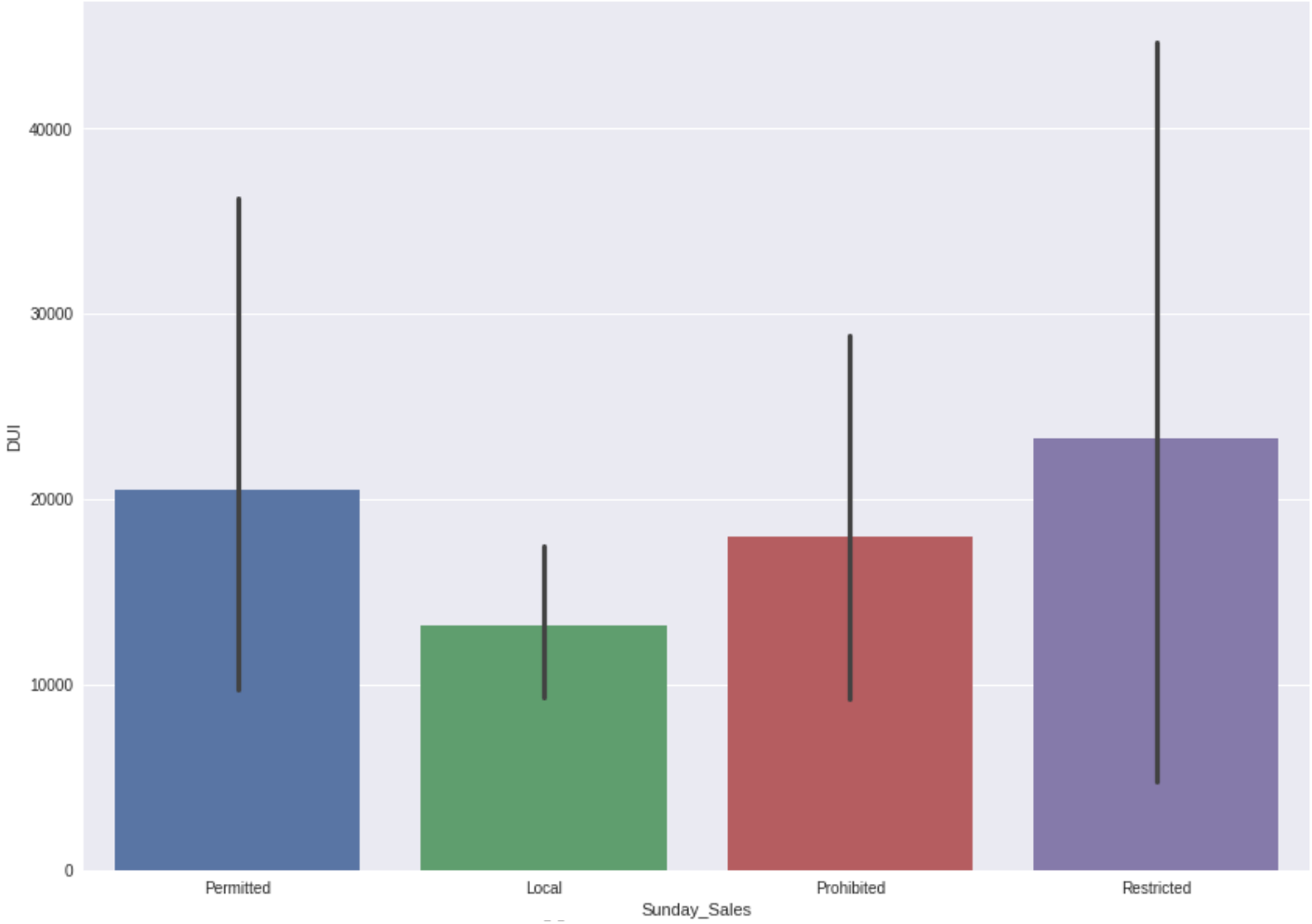
-[Further Analysis]-

Catplot

Even though Restricted Sunday Sales yields the lowest *total DUIs, it also* yields the highest *average* number of DUIs (23,279) relative to Prohibited Sunday Sales due to its low count of only 3 states

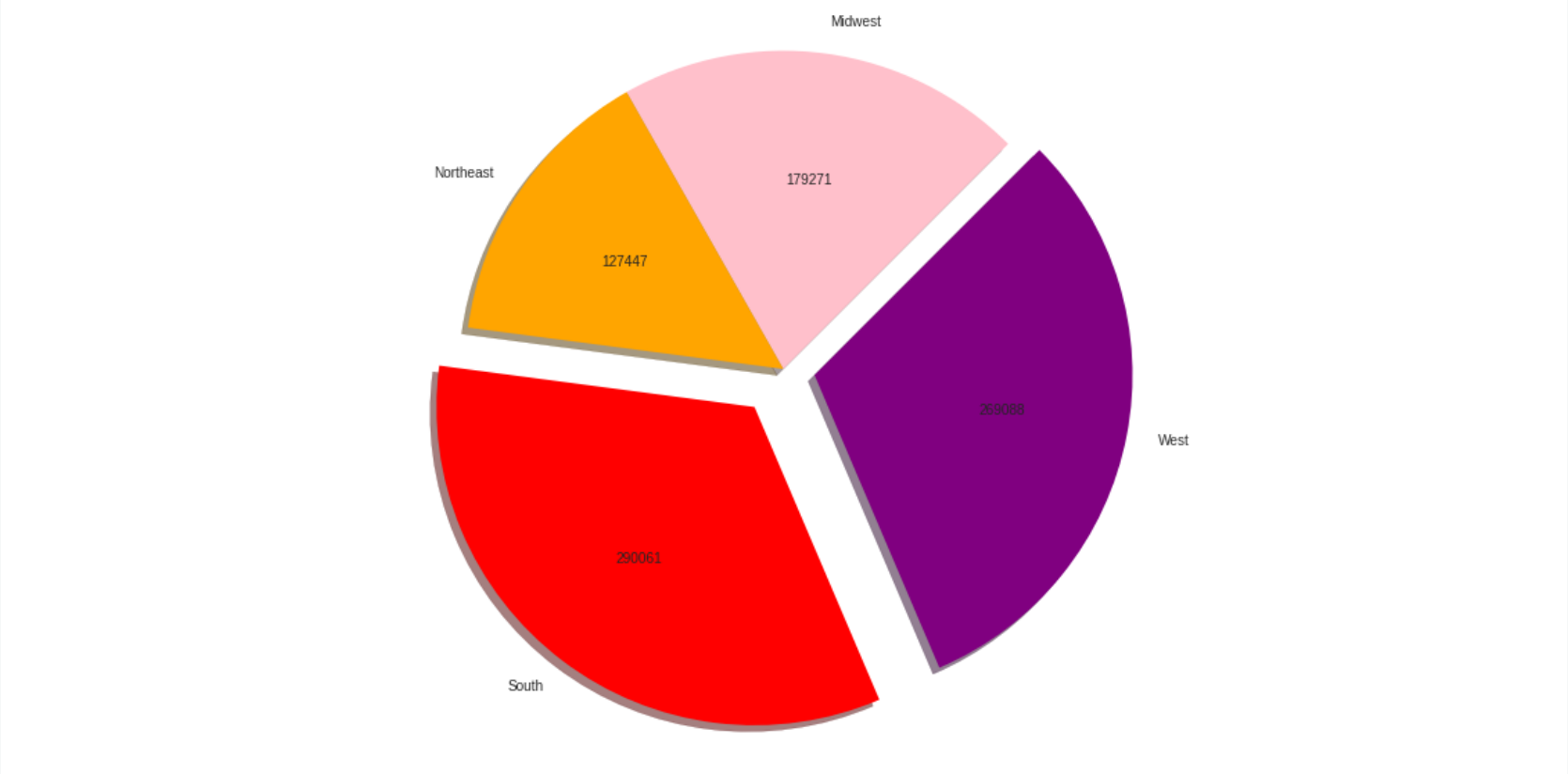
It is interesting to note that states with Local Sunday Sales yield the lowest *average* of DUIs (13,158), even compared to Prohibited Sunday Sales (17,970)

We can confirm that states with Permitted Sunday Sales have the highest standard deviation (32,734), while Local Sales have the minimum standard deviation (9,584)



count	17.000000	count	19.000000	count	11.000000	count	3.000000
mean	20491.941176	mean	13157.631579	mean	17970.090909	mean	23279.333333
std	32733.695486	std	9583.861210	std	18209.561068	std	20081.685545
min	386.000000	min	3659.000000	min	3674.000000	min	4746.000000
25%	3163.000000	25%	6635.000000	25%	7376.000000	25%	12611.500000
50%	9028.000000	50%	8148.000000	50%	11101.000000	50%	20477.000000
75%	24588.000000	75%	18521.000000	75%	19711.000000	75%	32546.000000
max	141458.000000	max	34254.000000	max	64971.000000	max	44615.000000

-[Region & DUIs]-



From Hypothesis #3, we can confirm that the South region yields the most DUIs at 290,061

The West region ranks 2nd highest at 269,088 DUIs

How can we further investigate the South and West regions?

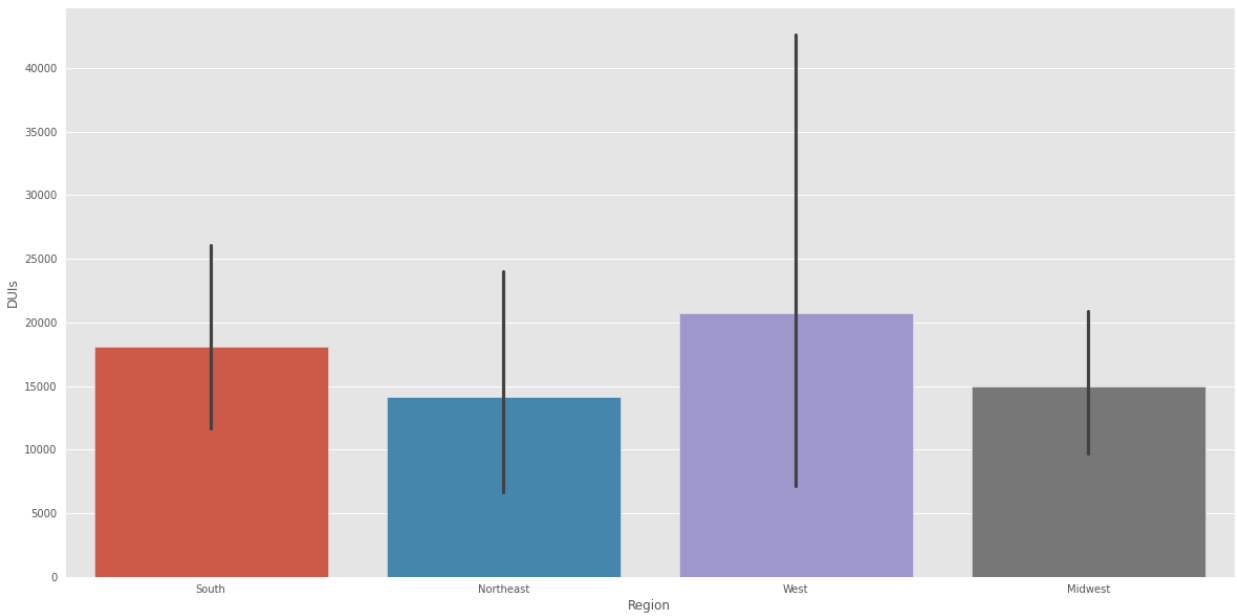
Region	DUI	Fatalities	Population
Midwest	179271	1915	67907403
Northeast	127447	1104	56283891
South	290061	5115	120510619
West	269088	2135	76044679

Hypothesis 3

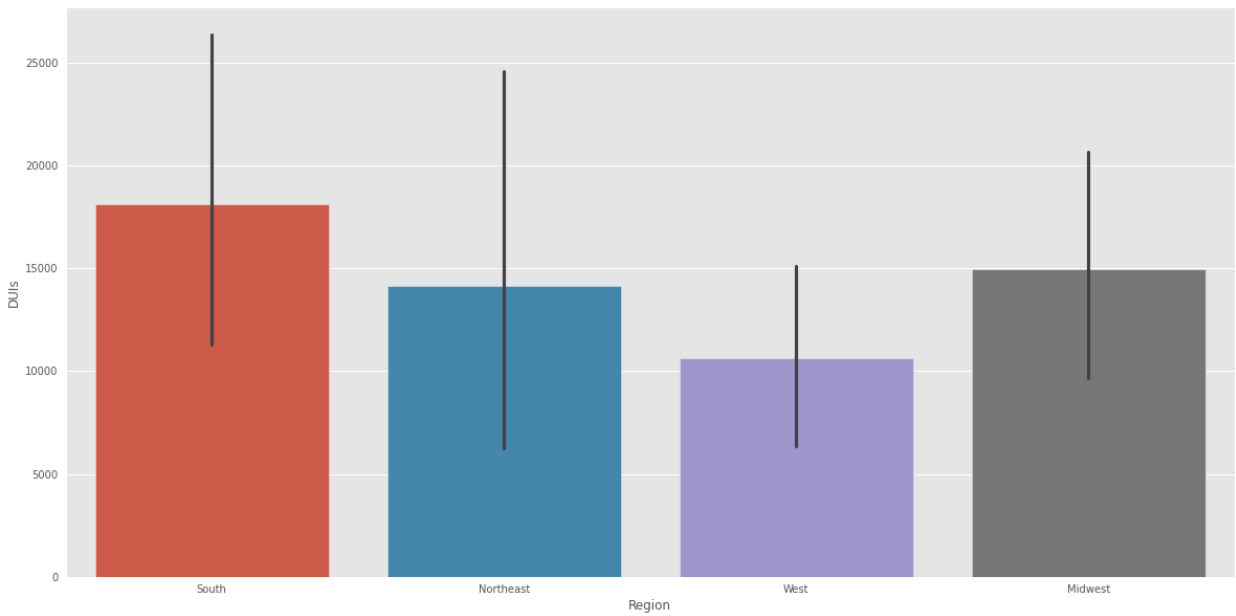
-[Region & DUIs]-

Box Plot

With California outlier



Without California outlier



With all data points, the West Region shows the highest average number of DUIs, followed by the South Region → Why?

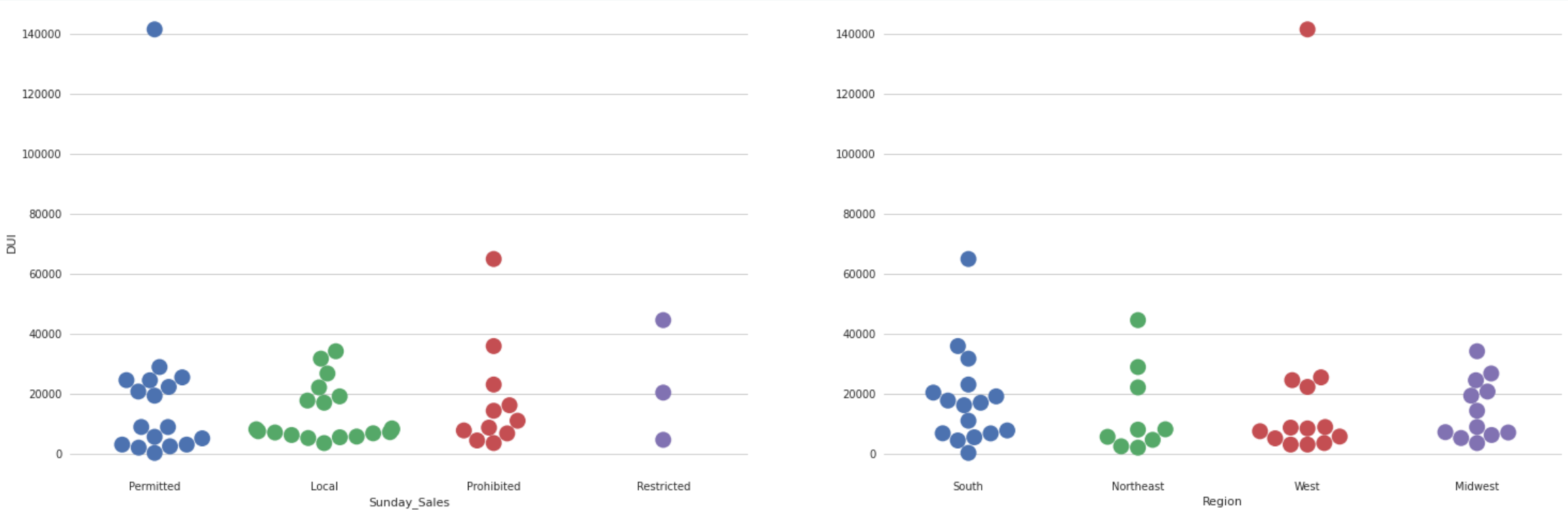
How is it that the West region yields *higher average* number of DUIs, while the South region yields the *highest total* DUIs?

Removing the outlier (California), the West average falls below all other regions. The South Region now displays the highest average of DUIs

Without confounding variable (California), we see a significant impact on the West region's average DUIs (jumps from highest to lowest), signaling that California was a large contributor

-[Region & DUIs]-

Swarm Plot



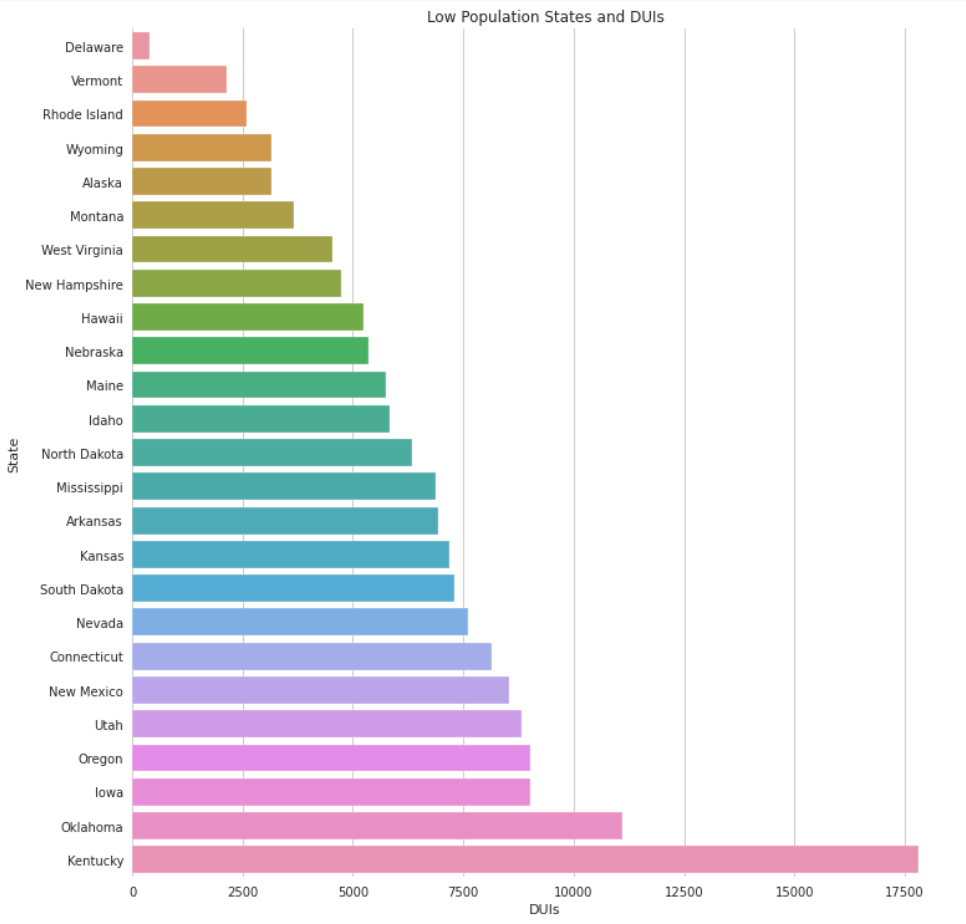
Combining Hypothesis #2 and #3, the most DUIs occur in states with Permitted Sunday Sales, and from states in the Southern region.

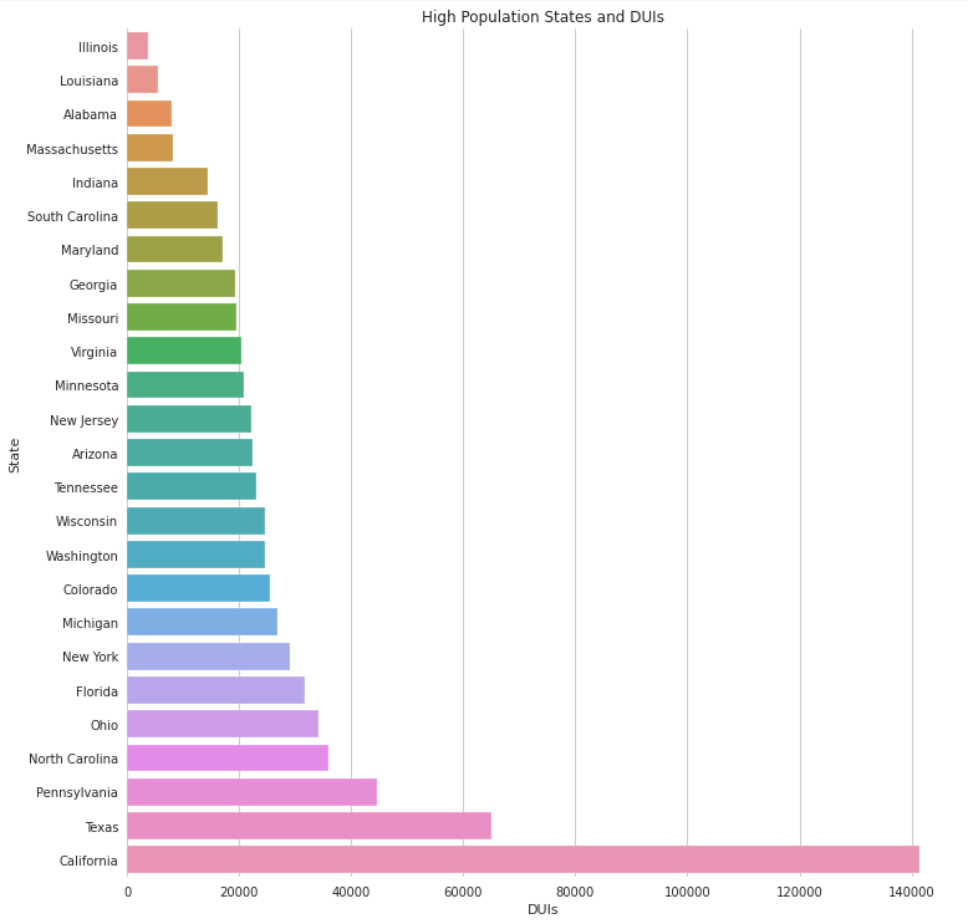
The outlier on both Swam plots is California (Permitted Sunday Sales and West Region)

Hypothesis 3

-[Low vs. High Population]-

CatPlot





When running a test statistic between high / low populations and total DUIs, we receive a p-value less than 0.5

Ttest_indResult(statistic=-4.008284646954876, pvalue=0.000212370204891677)

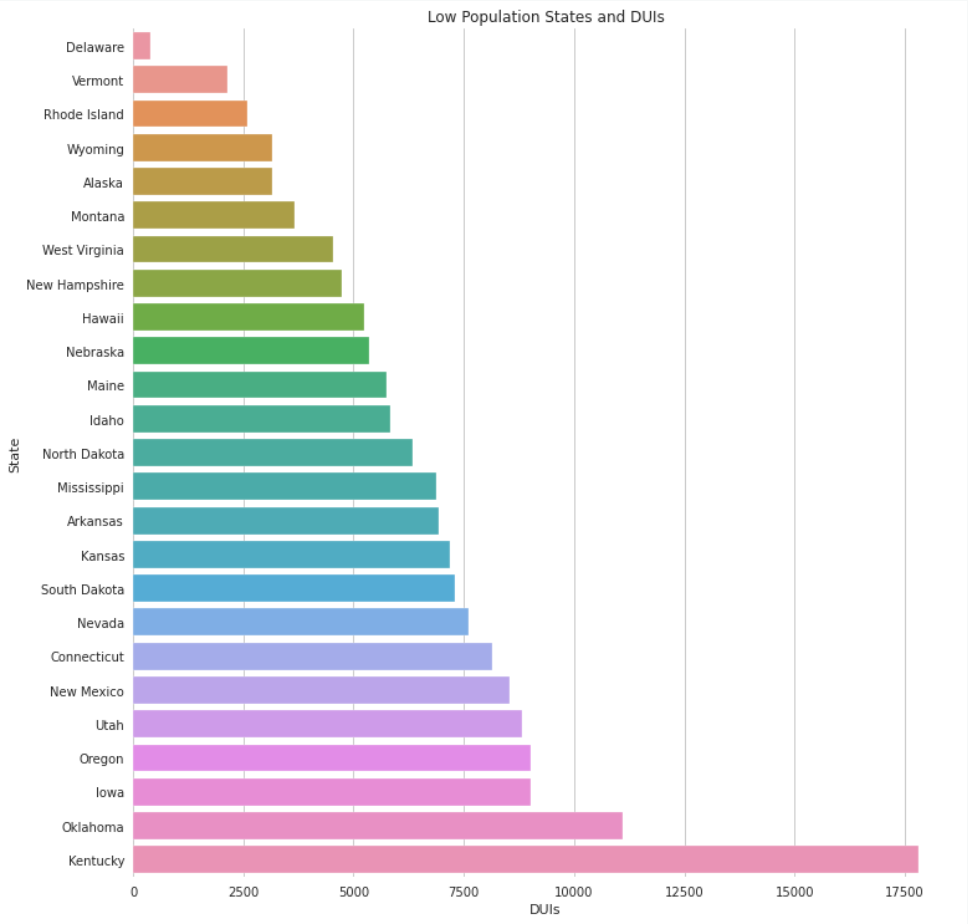
Thus, we reject the null Hypothesis #3 and accept the alternate that higher population states receive more total DUIs

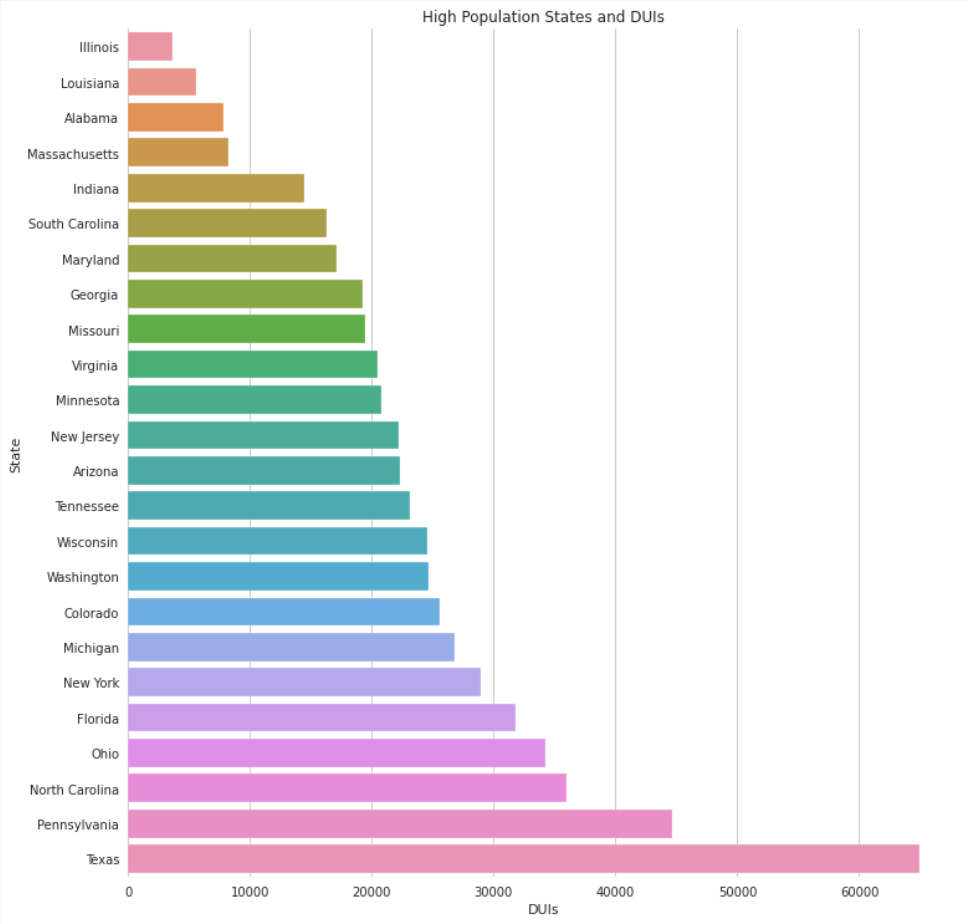
What if we were to remove the California outlier? Would this lead to less statistically significant results? (meaning high population states do not necessarily yield more DUIs than low population states)

Hypothesis3
2

-[Low vs. High Population]-

CatPlot





Ttest_indResult(statistic=-6.24850298585821, pvalue=1.1309687692689064e-07)

Removing the California outlier yields a different test statistic between high / low populations and total DUIs

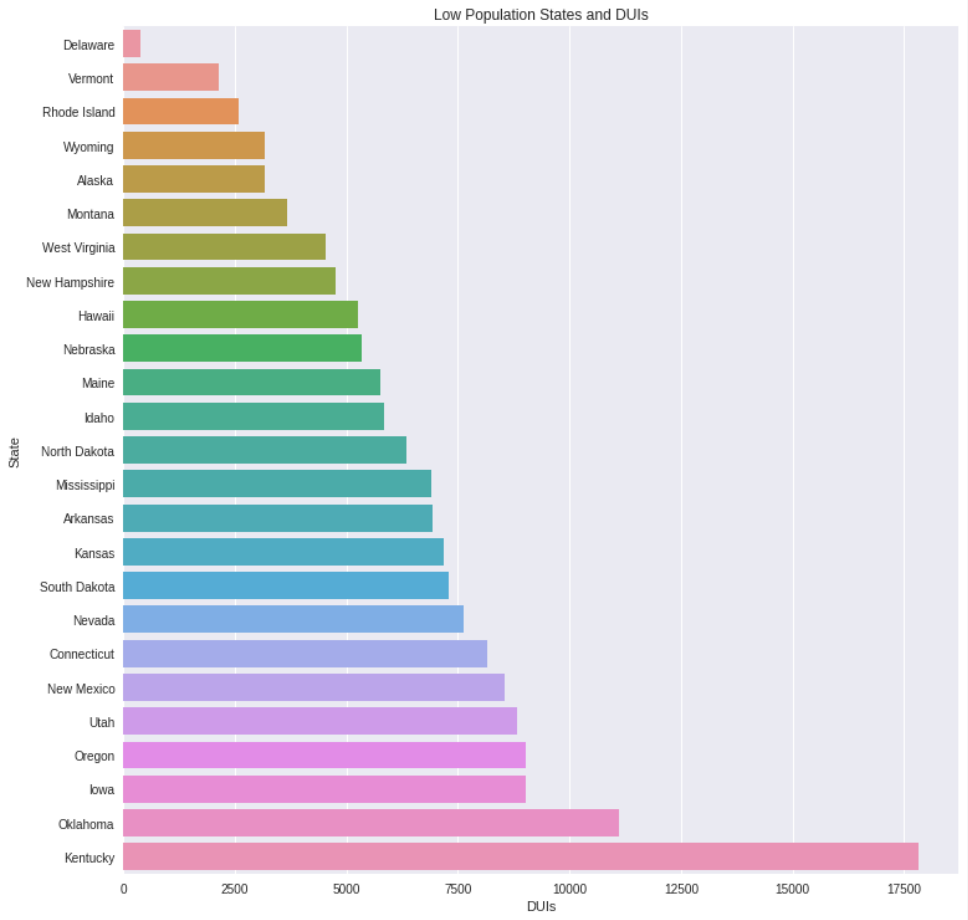
We receive an even smaller p-value that is still < 0.5

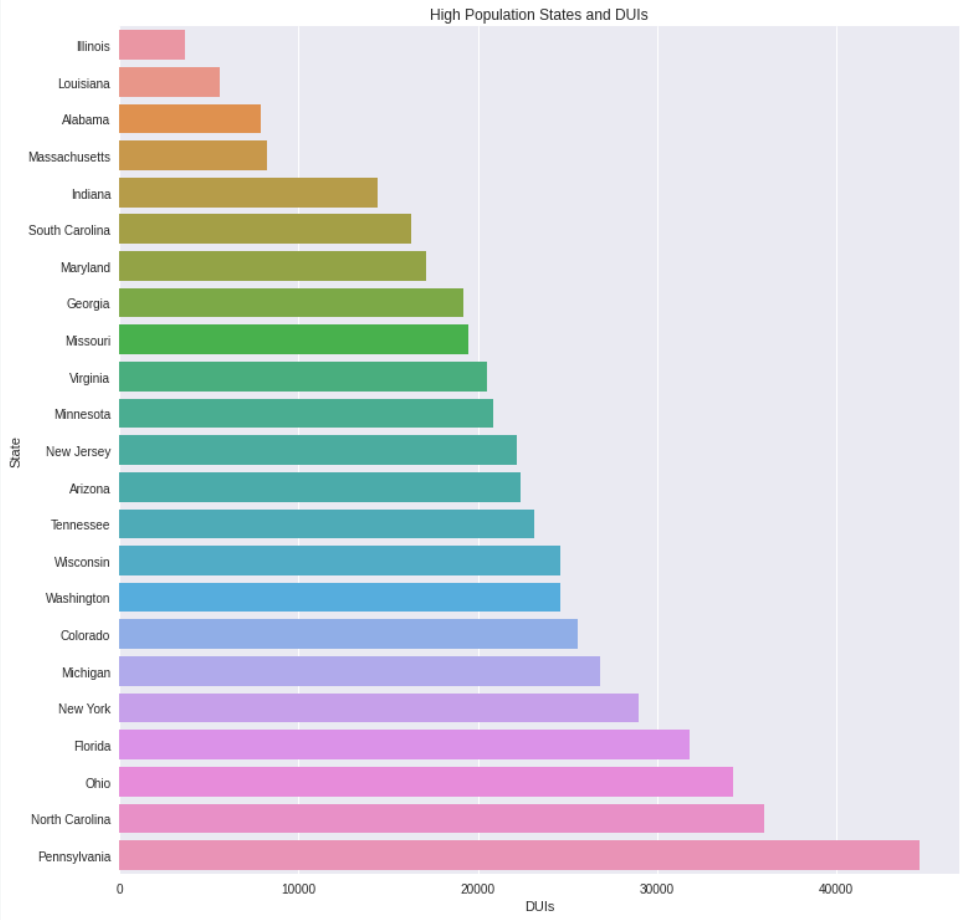
Thus, even without the high outlier of California, we still reject the null Hypothesis #3 and accept the alternate that higher population states receive more total DUIs

Hypothesis 3
2

-[Further Analysis]-

CatPlot





What if we remove the top *two* outliers of California and Texas from the high population class to see if there is a different result?

Ttest_indResult(statistic=-7.186939650753479,

pvalue=4.781042750005557e-09)

We still receive a p-value of less than 0.5

Thus, even without the two highest outliers, we are still confident to reject the null Hypothesis #3 and accept the alternate that higher population states receive more total DUIs

CONCLUSIONS

California displays an abnormally high number of DUIs, representing a confounding variable and skewing some outputs to favor the West region/high population states.

There is generally a strong positive correlation between total fatalities and total DUIs. There is also a positive correlation between death rate (fatalities/population) and total DUIs.

Permitted Sunday Sales yield the most DUIs, followed by Local and Prohibited. However, Local Sunday Sales return the lowest average number DUIs and smallest standard deviation, indicating a more even distribution among its 19 states.

The most DUIs occur in the South region and in states with high population (>4,547,908)

- DUIs
- Fatalities
- Population
- Region