Capstone Project1: Identifying counties that are the most fire-prone and predicting the cause of a fire wildfire.

We have seen what has caused more number of fires over the years.Also we did not see any direct increase or decrease of the number of fires over years. Now the question I am trying to answer is "Is the cause of wildfire related to season?" To answer this question I am performing a chi-square Test.

Following are the null and alternative hypothesis I have chosen.

H0(null hypothesis): The features Cause of Fire and Season are independent (which means they are not associated).

H1(alternate hypothesis): Cause of Fire and Season are not independent (which means they are associated).

Firstly, I created a Contingency table using the features in question. Also I choose significance level **α = 0.05**

SEASON	Autumn	Spring	Summer	Winter
STAT_CAUSE_DESCR				
Arson	56818	118176	48426	54785
Campfire	18560	21397	27054	7778
Children	9320	26595	15606	9225
Debris Burning	73292	189394	60094	103616
Equipment Use	31556	41534	53747	19785
Fireworks	834	1076	9037	387
Lightning	31007	24724	213776	1640
Miscellaneous	63120	109267	99036	49440
Missing/Undefined	22874	50117	43742	27831
Powerline	2801	5092	4644	1847
Railroad	4152	14749	6906	7467
Smoking	10984	18179	15002	8274
Structure	833	1373	984	572

Then I applied the Chi-Square test using the method *chi2_contingency.* This method returned **Chi-Square Statistic value,p-value,degree of freedom and expected value.**

- Mathematical formula to create **Expected Frequency(value)** Is :

```
E = (row total * column total)/grand total
```

- The formula to calculate **Chi-square value or X ²** is:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where X is not the english alphabet we know but the 22nd Greek alphabet Chi.

And X ² is the summation of the squared difference between Observed and Expected frequencies divided by the Expected frequency for all the cells

- The **degrees of freedom** can be calculated as:

```
df = (total_rows - 1) * (total_cols - 1)
```

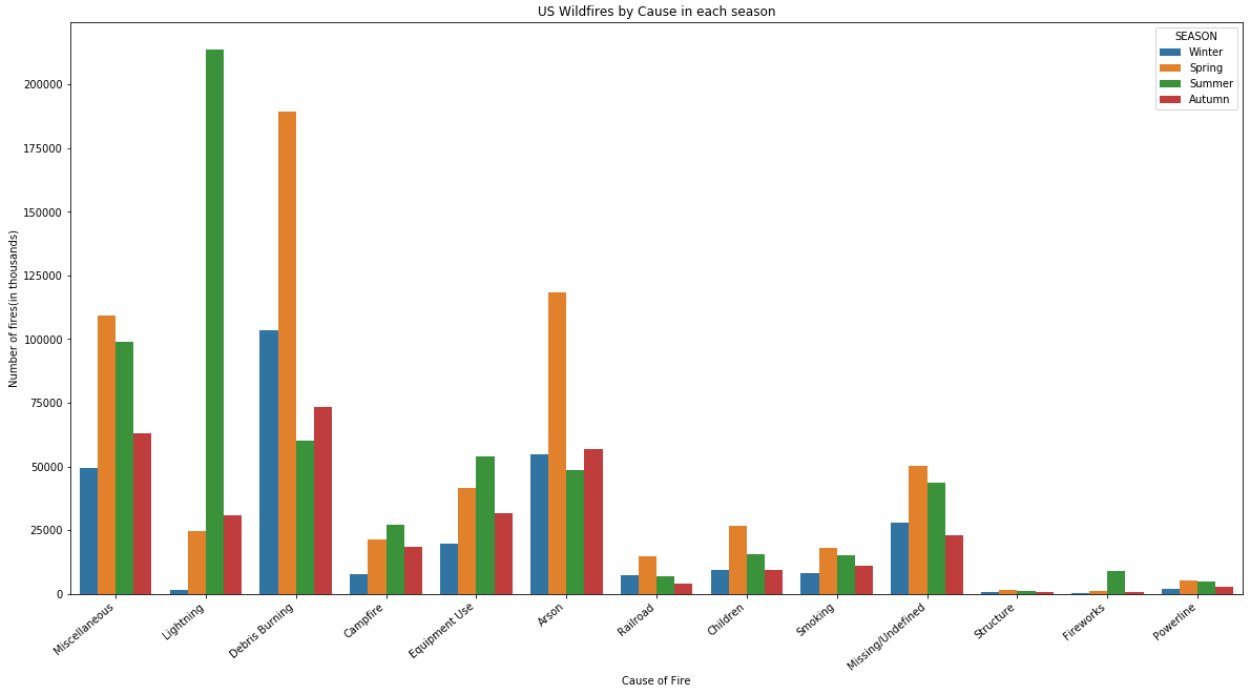
```
The Chi-Square Statistic Value is  410623.75290660973
The p-Value is  0.0
The degree of fredom is  36
[[ 49353.06234889  94071.3544635   90497.33513006  44283.24805755]
 [ 13267.43293619  25288.9147534   24328.12205763  11904.53025278]
 [ 10776.23020954  20540.45936716  19760.07303899   9669.23738432]
 [ 75641.87693722 144180.18819869 138702.40186236  67871.53300173]
 [ 26010.47683442  49578.29706205  47694.68654927  23338.53955426]
 [  2010.63104064   3832.44273643   3686.8381099    1804.08811303]
 [ 48100.98595178  91684.78477638  88201.43753172  43159.79174012]
 [ 56920.51416924 108495.59500088 104373.56065433  51073.33017555]
 [ 25645.39136754  48882.41148312  47025.23950232  23010.95764703]
 [  2551.69550808   4863.76004243   4678.97294625   2289.57150324]
 [  5902.74724249  11251.16460315  10823.70312941   5296.38502495]
 [  9302.58347806  17731.55678982  17057.8880929    8346.97163922]
 [   667.3719759    1272.070723     1223.74139487    598.81590623]]
```

Once we have all the values required to make a decision, we make a decision based on p-value. If our p value is less than the significance value we reject the Null Hypothesis and if our p value is greater than the significance value we do not reject it.

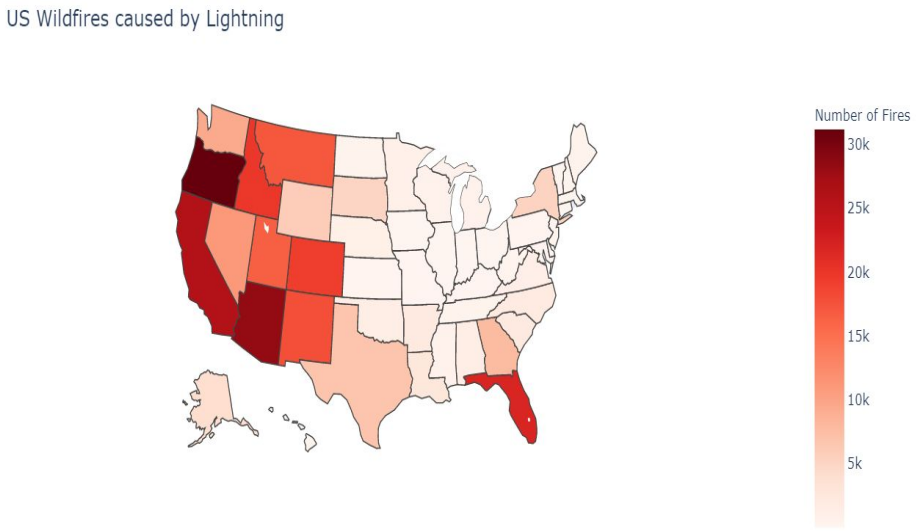
p-value=0.000000, significance=0.05

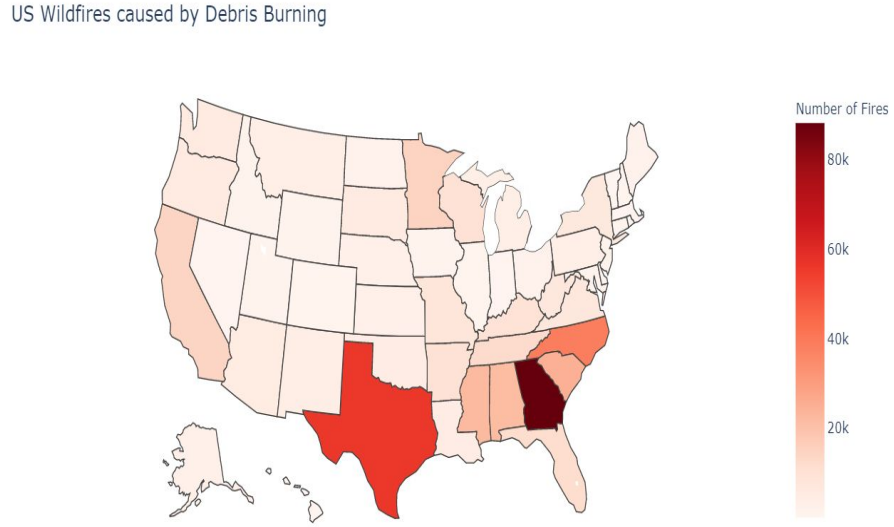
At 0.05 level of significance, we reject the null hypothesis and accept H1.
They are not independent.

As "Cause of Fire" and "Season" are not independent, let us see visualized data for the same



We already knew that wildfires caused by "Lightning" were more damaging.From the above observation we can also note that Lightning during Summer causes most wildfires.





We can see that there is a visible relationship between States and Cause of fire when we visualized data. However to establish a statistical relationship between these two variable I did apply chi-square test on STATE_NAME and Cause of Fire

H0: The features STAT_CAUSE_DESCR and STATE_NAME are independent (which means they are not associated).

H1: STAT_CAUSE_DESCR and STATE_NAME are not independent (which means they are associated).

After performing chi-square tests on these two features and choosing significance level as 0.05, we can conclude that STATE_NAME and Cause of Fire are associated with each other.

p-value=0.000000, significance=0.05

At 0.05 level of significance, we reject the null hypothesis and accept H1.

They are not independent.

I want to point out that in both tests performed as described above, p-value calculated is exactly 0.0. We have to treat this value with caution due to the large sample size.

Below is the link to the Github repository of jupyter notebook file with Statistical analysis code.

https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/blob/master/Capstone1_US-Wildfire-Prediction/Capstone1_Statistical_analysis.ipynb