# Analysis of Tobacco Supplement Uses

# INDEX

**5. Challenges and Barriers.……...…….**

# 1. Introduction

Many key tobacco control approaches are executed so as to (1) decrease cigarette utilization among current smokers and (2) dishearten tobacco utilization among non-smokers, particularly youth. The best method to accomplish these objectives is to build the cost of cigarettes. The higher the cost of buying a pack of cigarettes, the more uncertain it is that individuals will purchase and devour cigarettes.

In any case, there are numerous components that can upset the basic connection between the cost of cigarettes and utilization. For instance, buyers can counterbalance more expensive rates by buying cigarettes in mass, for example, in containers or in multipacks instead of as single packs. Also, smokers can change to bring down estimated cigarette brands, change to brands offering value limits and shop for cigarettes in areas where cigarettes are more affordable. At long last, smokers can react to higher cigarette costs by diminishing their day by day admission of cigarettes or stop their cigarette utilization all together. While not all smokers will essentially take part in cost limiting practices, the unfaltering ascent in cigarette costs combined with expanding rates of joblessness, stale or potentially declining wages, and higher family costs for things like fuel and nourishment have consolidated in the course of recent years to make cigarettes more expensive. An ongoing article from the International Tobacco Control (ITC) United States Survey announced an expansion in the utilization of rebate cigarettes by US smokers after the 2009 increment of $0.61 in the government extract charge (FET) on cigarettes.

For further information, please refer to: https://www.ncbi.nlm.nih.gov/books/NBK99239/

# 2. Problem

## 2.1 Problem Statement

Our aim is to analyze the relationship between tobacco sales and tobacco prices over time using Exploratory Data Analysis(EDA).

## 2.2 Problem Goal

# 3. Literature Review

# 4. Methodology

## 4.1. Dataset

Here for the analysis of Tobacco Supplement Uses, we are using two datasets.The datasets that we are using was downloaded from Division of Cancer Control and Population Sciences (DCCPS), National Cancer Institute, USA. (For more info, please visit: https://cancercontrol.cancer.gov/brp/tcrb/tus-cps/info.html)

It was organized in to lead efforts in cancer control research. The data and technical documentation for the 2014-2015 Tobacco Use Supplements are available for download from the Current Population Survey FTP Page(Source: https://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpssupps).

First dataset has mainly fifteen variables(columns). They are as follows:

| Year | LocationAbbr | TopicType | TopicDesc | MeasureDesc | Response |
|---|---|---|---|---|---|
| Year when the data is collected. Eg . 2003 | States from wherethe data is collected. Eg. AL(Alabama, USA) | Topic like Tobacco Use – Survey Data | Description of how tobacco is used likeCigarette Use, pipe use | Percent of Former Smokers Among Ever Smokers | Response from Smokers like they are former or current smoker. |

| Data_Value_Type | Data_Value | Data_Value_Std_Err | Low_Confidence_Limit |
|---|---|---|---|
| Percentage | Tobacco use percentage | How much data is scattered | Smallest value of feature from where data distribution begins. |

| High_Confidence_Limit | Sample_Size | Gender | Race | Age | Education |
|---|---|---|---|---|---|
| Largest value of feature from where data distribution ends. | Total number of samples taken | Gender of Tobacco user | Race of tobacco user | Age Group of tobacco user | Qualification of tobacco user |

For tobacco sale analysis, we are using different dataset. It has mainly eleven columns.

| LocationAbbr | LocationDesc | Year | Datasource | TopicDesc |
|---|---|---|---|---|
| Abbreviation of location like AL for | Name of the location. Eg. | Year of data collection like | Source from where the data is | Description of topic The Tax Burden on |

| | | | | |
|---|---|---|---|---|
| Alabama. | Alabama | 2006. | collected | Tobacco. |

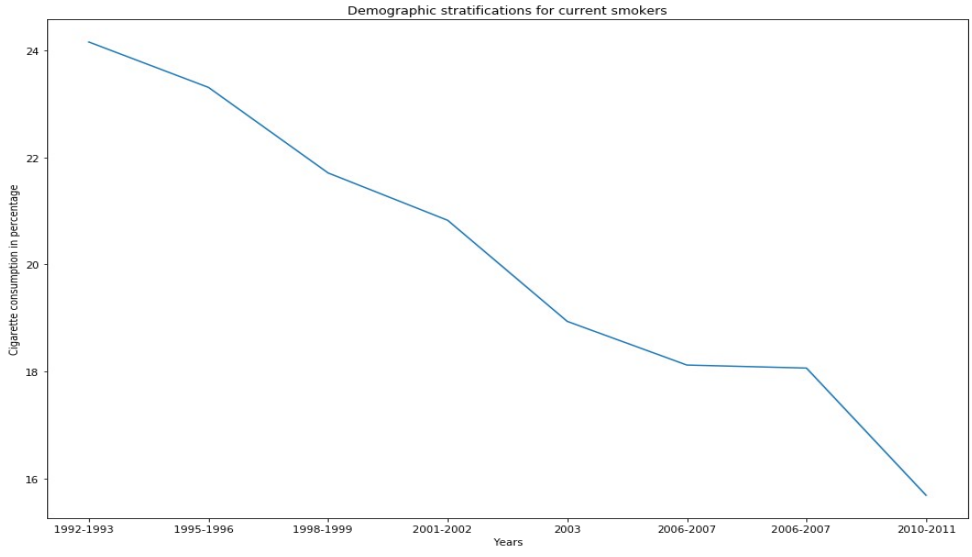
| SubMeasureDesc | Data_Value | Data_Value_Unit | Data_Value_Type |
|---|---|---|---|
| Consumption (Pack Sales Per Capita) | Average cost | $ | Dollar |

| GeoLocation | Source |
|---|---|
| Latitude, Longitude (32.84057112200048, -86.63186076199969) | Source from where dataset is downloaded. |

# 4.2. Implementation

## 4.2.1. Demographic stratifications for current smokers

The following plot shows the consumption of Cigarette over years. The plot clearly shows that there is a decline in cigarette consumption over years.

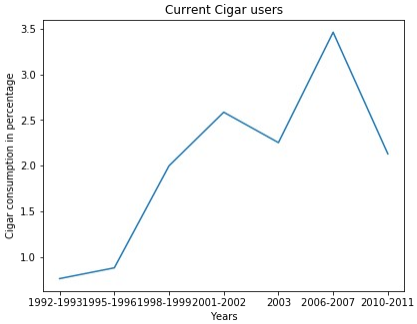


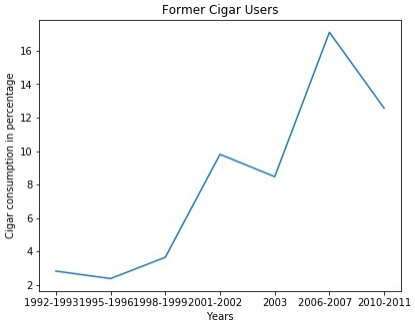

## 4.2.2. Consumption of different tobacco products

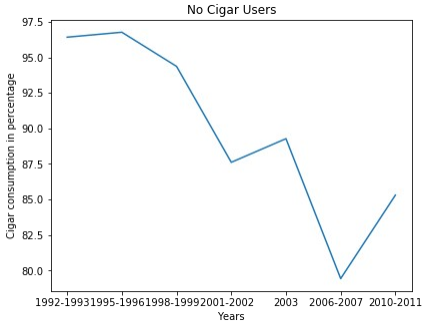
### 4.2.2.1. Cigar Use (Adults)

Analysing data of consumption of Cigar. Plots for Cigar Use for three smoking status (current, former and never).

According to the plots, t*he consumption is gradually increasing from 1992 to 2006. Then it starts decreasing from 2006-2007.*

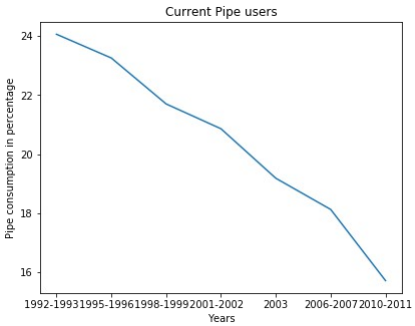


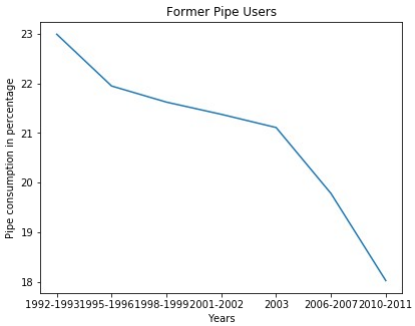



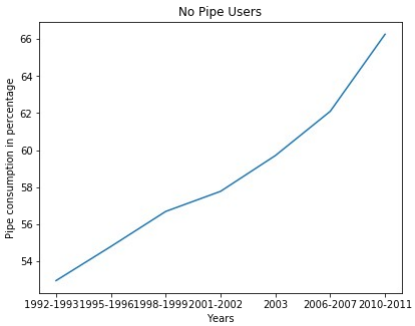




### 4.2.2.2. Pipe Use (Adults)

Analysing data of consumption of pipes Below are the plots for Pipe Use for three smoking status (current, former and never).
According to the plots, *the consumption is gradually decreasing from 1992 to 2011.*

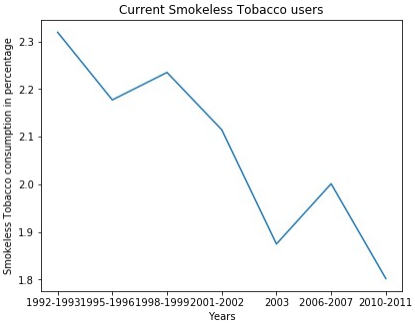


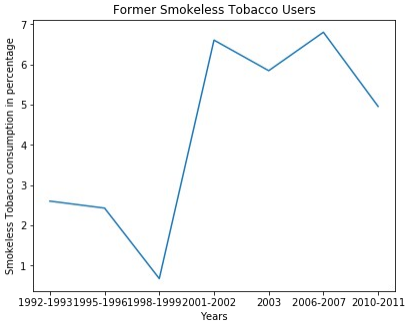



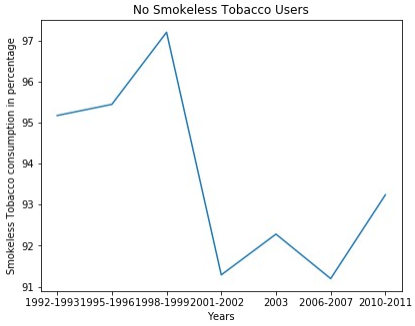




### 4.2.2.3. Smokeless Tobacco Use (Adults)

Analysing data of smokeless tobacco use. Below are the plots for Smokeless Tobacco Use for three User status (current, former and never).



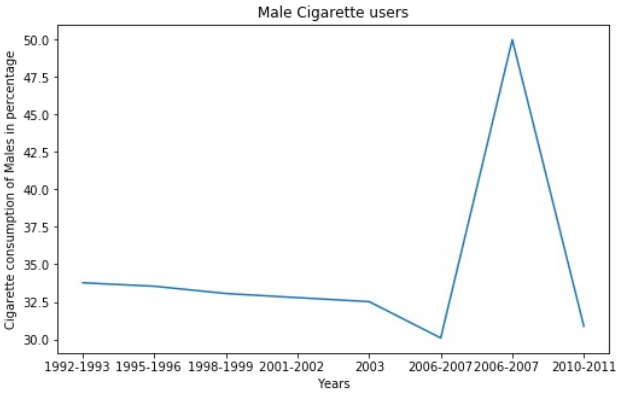



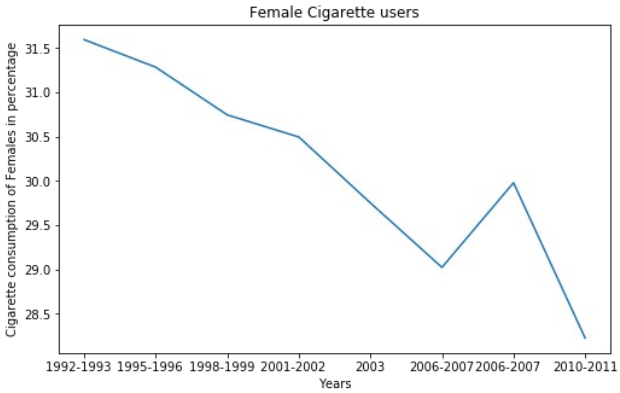

According to the plots, the consumption is gradually decreasing from 1992 to 2011 with a few exceptions at points 2003 and 2006-2007. Overall, the consumption has *decreased with lot many hiccups.*

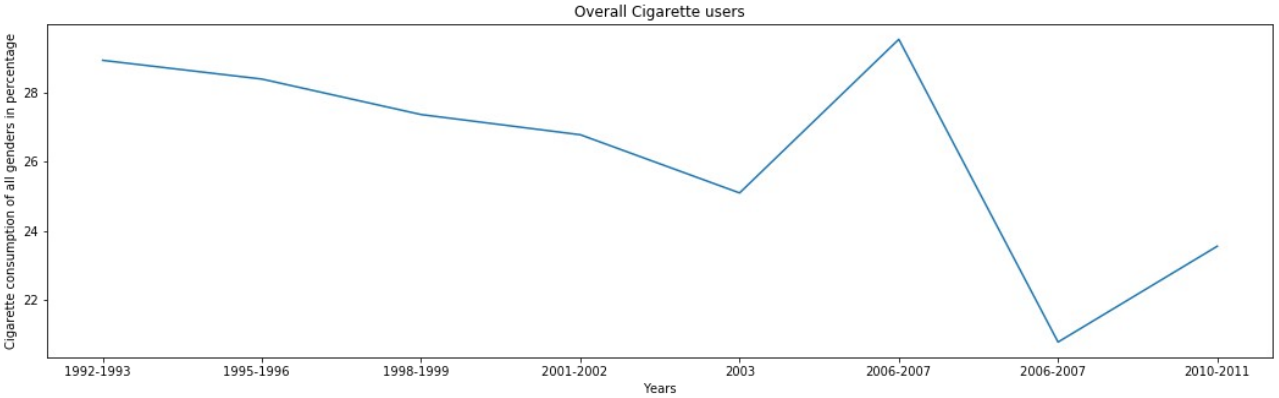
## 4.2.3. Cigarette use analysis (For Gender)

Analysing cigarette use data for different genders. According to the plots below, cigarette consumption for males was almost same with a slight decrease from 1992-2006. Then a *sudden increase on 2006-2007 then again a decrease.*

Similar case is with female plot *except that it has a gradual decrease from 1992-2006.*



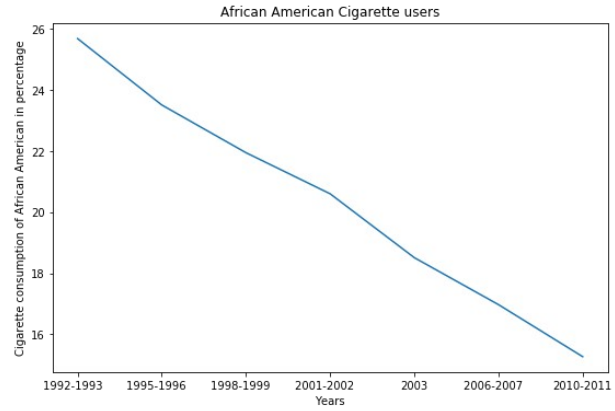



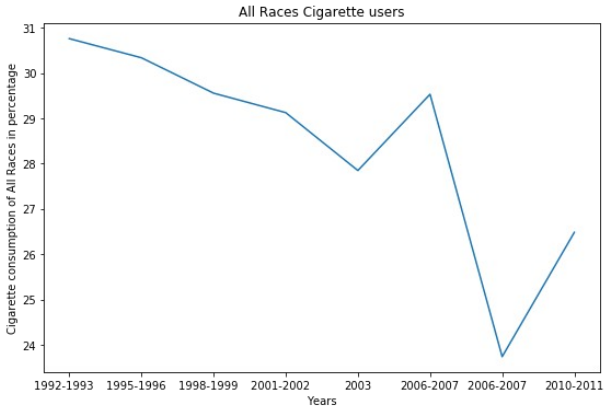



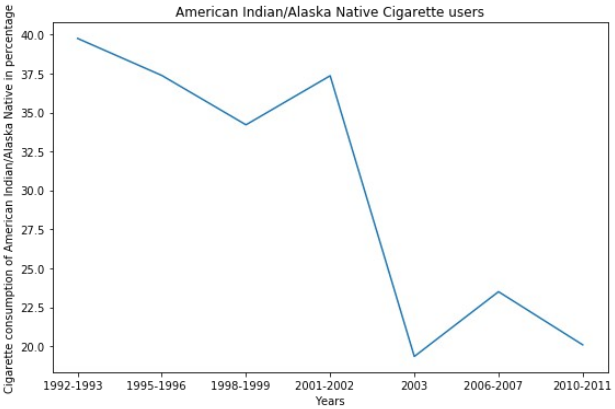

## 4.2.4. Cigarette Consumption Analysis (for Races)

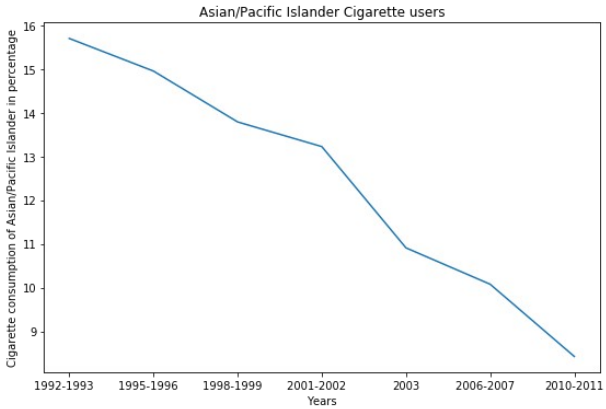
Analysing cigarette consumption data for different races. According to the plots below, the consumption of all the races are gradually decreasing with little bumps here and there.

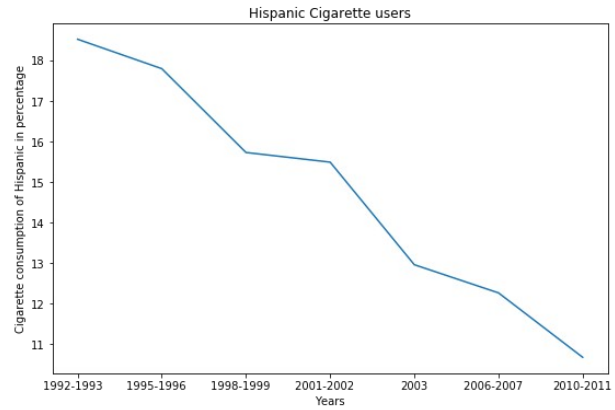
But the cigarette consumption of All Races d*ecreases from 1992-2006 and then starts to increase from there.*

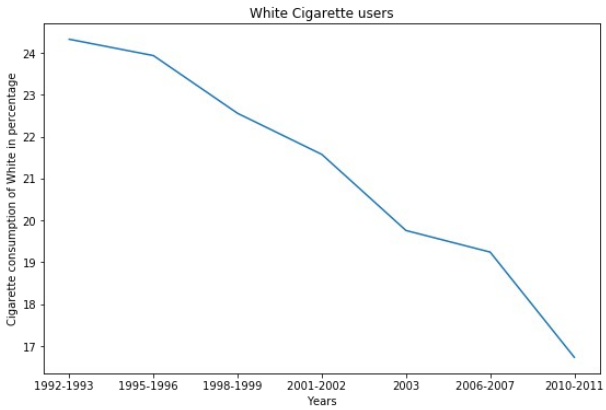


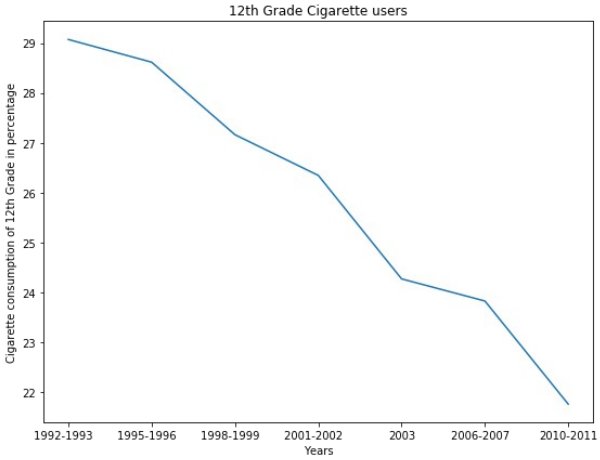



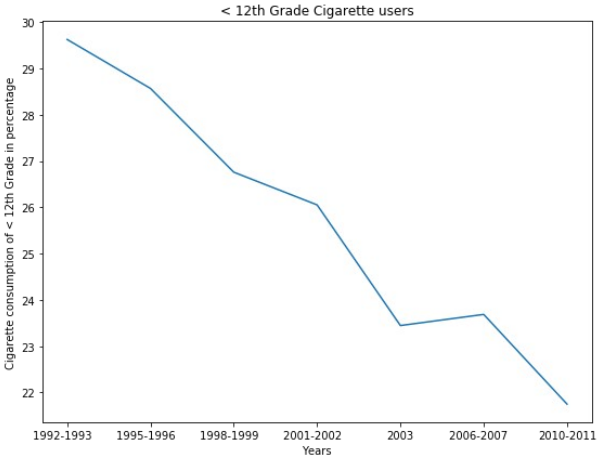



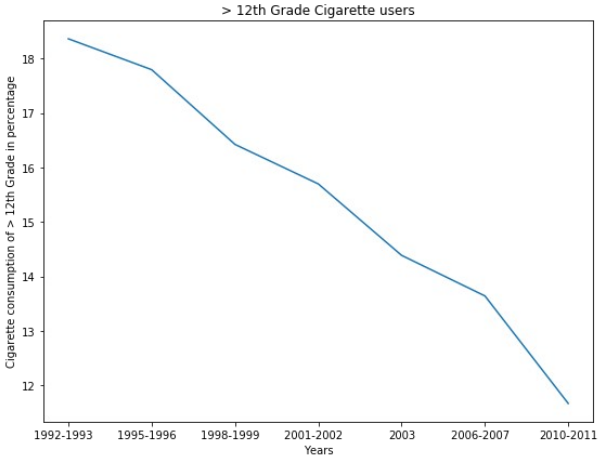

## 4.2.5. Cigarette Consumption Analysis (for Education)

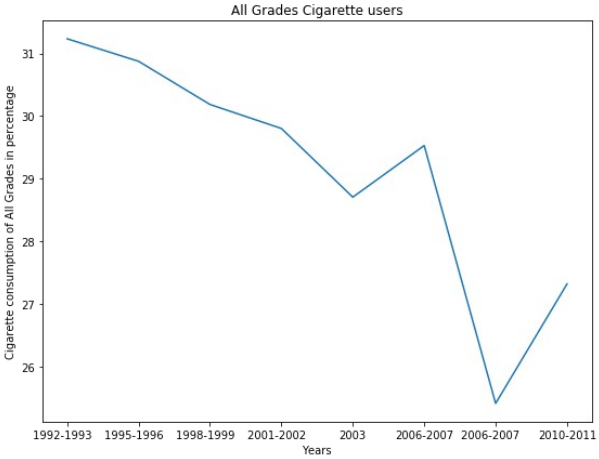
Analysing cigarette consumption data for different educations. According to the plots below, the consumption of all the educational backgrounds are gradually *decreasing with little bumps here and there.*

But the cigarette consumption of All Grades *decreases from 1992-2006 and then starts to increase from there.*



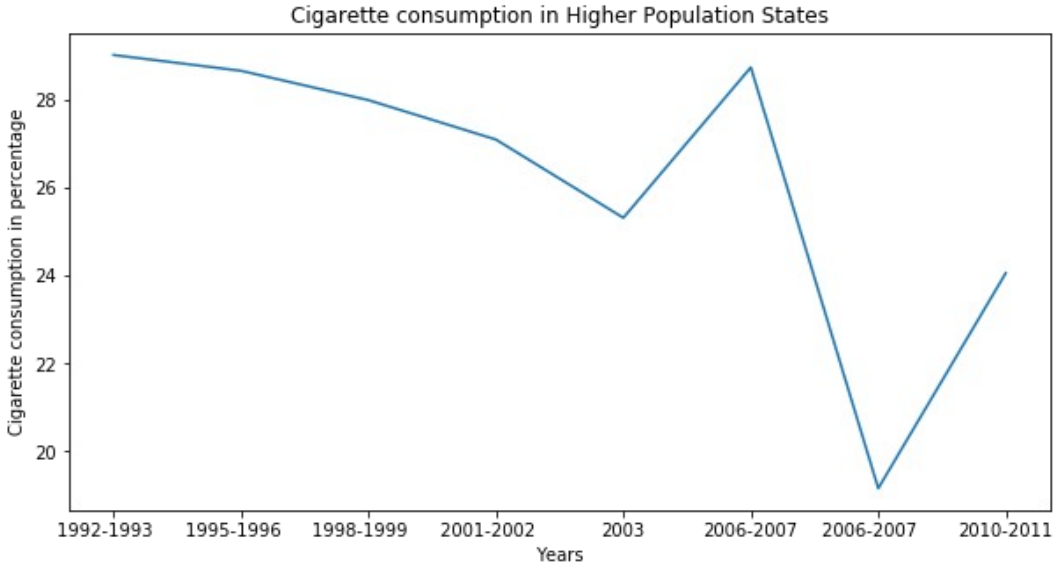




## 4.2.6. Location Wise Analysis

Analysing cigarette consumption in Higher Population States. According to the plots below, the consumption of Higher Population States are gradually *decreasing with little bumps here and there.*

*Lower Population States, East area, West Area, North Area, and South Area also have similar trend.*




## 4.2.7. Model Score

I am using random forest regressor to predict the model score. A **random forest** is a meta estimator that fits a number of classifying **decision** trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Here, I have split the dataframe into two random samples i.e train sample and test sample with 75% and 25% sample size respectively. The calculated score came out to be 0.9999540248243328, which is fairly good.

**Most important feature**: Low_Confidence_Limit
**Least important feature:** Data_Value_Type_Mean

## 4.2.8. Tobacco Sales Analysis

Now, I am going to look into a correlation between cigarette sales and cigarette prices over time, so I picked general information columns like 'LocationDesc' and 'Year' and then columns with cigarette sales and price.

| LocationDesc | Year |
|---|---|
| Alabama | 1970 |
| Alaska | 1970 |
| Arizona | 1970 |
| Arkansas | 1970 |
| California | 1970 |

| SubMeasureDesc | Data_Value | Data_Value_Unit | Data_Value_Type |
|---|---|---|---|
| Average Cost per | | | |

| | | | |
|---|---|---|---|
| pack | 0.427 | $ | Dollars |

However, the original dataframe does not contain columns that show price and sales of cigarettes only. They are included in 'SubMeasureDesc' and 'Data_Value' columns.

### 4.2.8.1. Data Processing

I created a new dataframe called 'data' with the useful columns for the later data analysis. Also, I prepared some dataframes based on their 'SubMeasureDesc' column values. The '**SubMeasureDesc**' column actually includes descriptions of value types in the '**Data_Value' column**, so we need to take note of when the column has certain string values. Since we need only cigarette sales and price from the column, we can just check if the 'SubMeasureDesc' column contains string values of '**Average Cost**' and '**Consumption**' and copy values from 'Data_Value' when it has the matching string values. Lastly, we need to make sure that there are no missing values. There are many ways to deal with missing values. Here I will just make sure to delete all missing values from the dataframe.

**Data Table**

```
   LocationDesc  Year          SubMeasureDesc  Data_Value
0      Alabama  1970  Average Cost per pack       0.427
1       Alaska  1970  Average Cost per pack       0.418
2      Arizona  1970  Average Cost per pack       0.385
3     Arkansas  1970  Average Cost per pack       0.388
4   California  1970  Average Cost per pack       0.397

                        GeoLocation
0    (32.84057112200048, -86.63186076199969)
1   (64.84507995700051, -147.72205903599973)
2  (34.865970280000454, -111.76381127699972)
3    (34.74865012400045, -92.27449074299966)
4   (37.63864012300047, -120.99999953799971)
```

Average Cost Table

```
   LocationDesc  Year          SubMeasureDesc  Data_Value  \
0      Alabama  1970  Average Cost per pack       0.427
1       Alaska  1970  Average Cost per pack       0.418
2      Arizona  1970  Average Cost per pack       0.385
3     Arkansas  1970  Average Cost per pack       0.388
4   California  1970  Average Cost per pack       0.397

                        GeoLocation
0    (32.84057112200048, -86.63186076199969)
1   (64.84507995700051, -147.72205903599973)
2  (34.865970280000454, -111.76381127699972)
3    (34.74865012400045, -92.27449074299966)
4   (37.63864012300047, -120.99999953799971)
```

Pack Sales Table

| | LocationDesc | Year | SubMeasureDesc | Data_Value | GeoLocation |
|---|---|---|---|---|---|
| 51 | Alabama | 1970 | Cigarette Consumption (Pack Sales Per Capita) | 89.8 | (32.84057112200048, -86.63186076199969) |
| 52 | Alaska | 1970 | Cigarette Consumption (Pack Sales Per Capita) | 121.3 | (64.84507995700051, -147.72205903599973) |
| 53 | Arizona | 1970 | Cigarette Consumption (Pack Sales Per Capita) | 115.2 | (34.865970280000454, -111.76381127699972) |
| 54 | Arkansas | 1970 | Cigarette Consumption (Pack Sales Per Capita) | 100.3 | (34.74865012400045, -92.27449074299966) |
| 55 | California | 1970 | Cigarette Consumption (Pack Sales Per Capita) | 123.0 | (37.63864012300047, -120.99999953799971) |

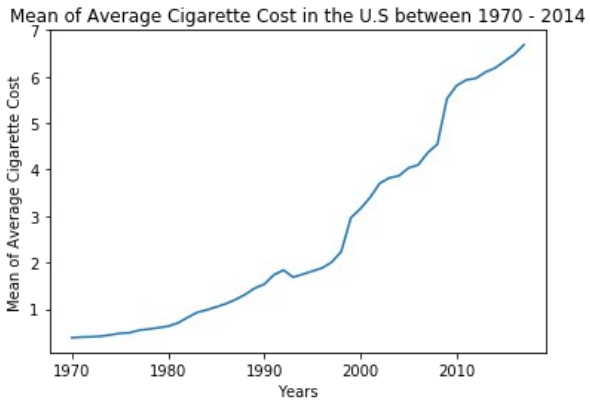
### 4.2.8.2. Exploratory Data Analysis and Data Visualization

In this step, we will use exploratory data analysis and data visualization. In exploratory data analysis, it is important to understand the central tendency of data, so we will calculate the mean values of each year.

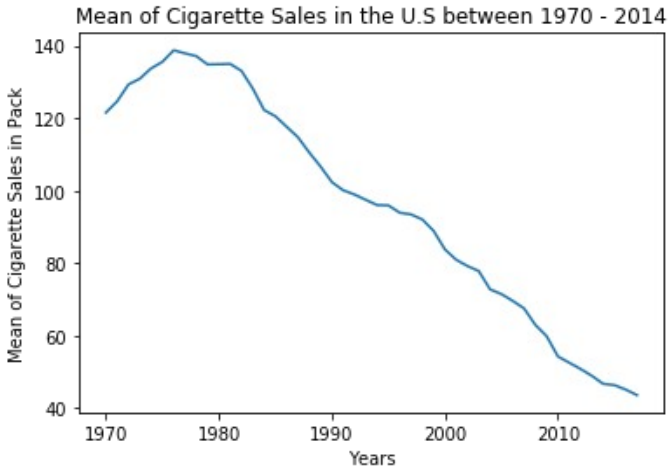
However, it is not enough to simply get mean values from the data because there are multiple same year values in the column. I used 'groupby' function to assign mean values to the rows with multiple identical year values as well.

Average Cigarette Price per Pack

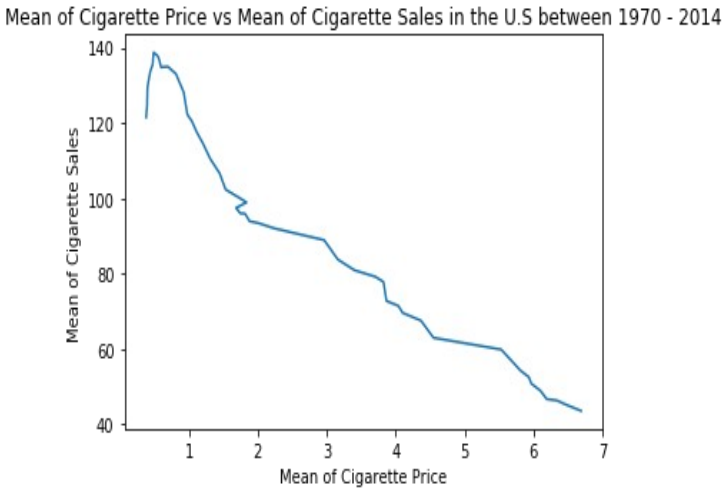
Now, let's start to plot graphs. The first graph of 'Average Cigarette Price per Pack in the U.S. between 1970 - 2014' shows that the average cigarette price in the U.S. has been gradually increased over time between 1970 and 2014.




The second graph of 'Cigarette Sales in the U.S. between 1970 - 2014' tends to decrease over time and it has an approximately inverse relationship with the average cigarette price per pack.

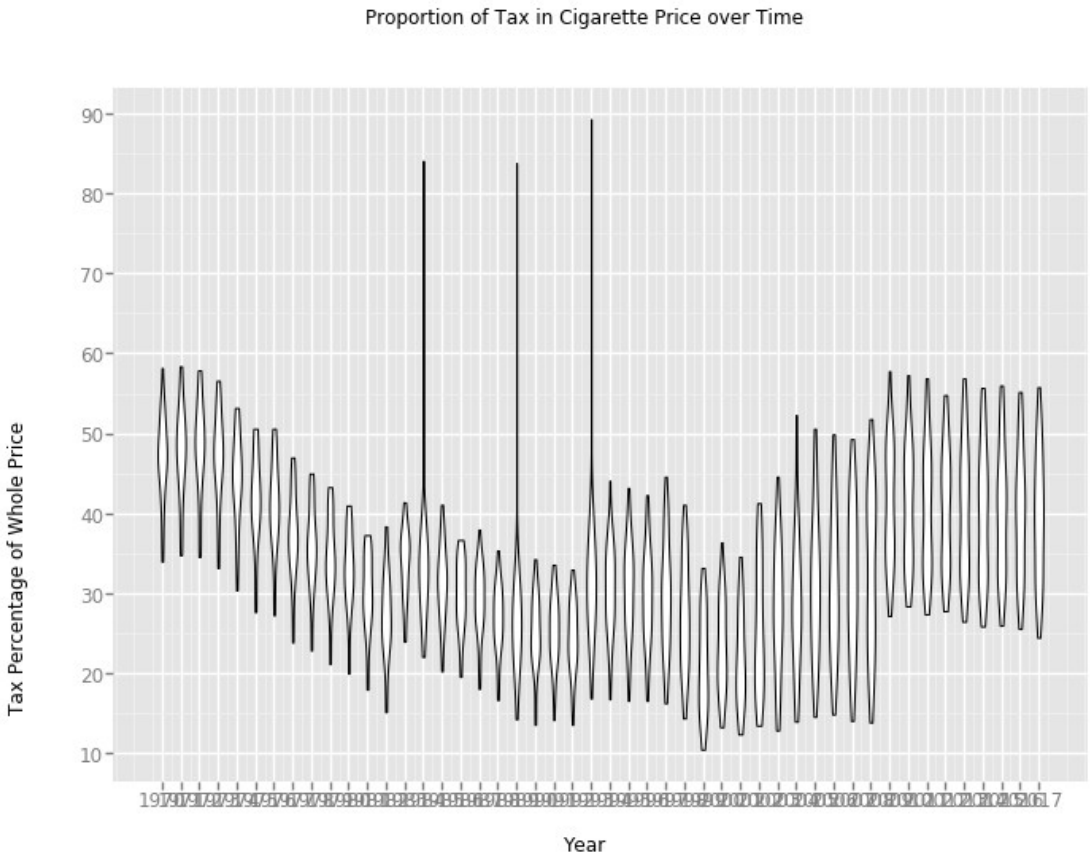



The third graph shows that the correlation between a mean of cigarette price and a mean of cigarette sales is negative. As the mean of cigarette price goes up, the mean of cigarette sales decreases, and vice and versa.




Tax Proportion (GGPlot)

I am going to use visualization tool 'ggplot' to show the proportion of tax in a cigarette price with violin plot. Below distribution is the distribution of tax percentage in the U.S. between 1970 and 2014.




According to the graphs from the previous process, I can tell that **there is a correlation between price and sales of cigarettes in the U.S**. At this point, we need to think about the validity of data. If the price and sales values from the graphs change over time naturally, do you think the result is distorted? Because of that reason, we need to transform the values into a unitless scale with standardization.

#### 4.2.8.3. Data Transformations

After the data processing part, we need to calculate standardized cost. From the previous graph, a price is the only value necessary for standardization, because a price changes over time by inflation. We cannot compare the price values in different years by the numbers because price values in 1970 and 2014 are different, so we cannot put them into one plot all together. That's why we need a unitless and proportional scale. Here is a formula to get the standardized cost.

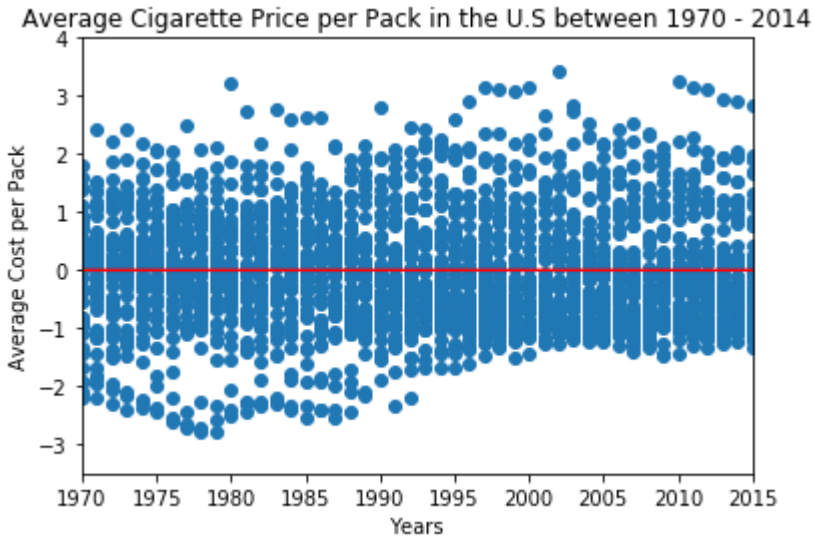
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Now let's look at the new plots with standardized cost. The first plot is standardized average prices by years. It seems well standardized since it has a **flat linear regression line across the average price**. Now we can see the unskewed correlation plot. The plot of standardized average cigarette prices and cigarette sales in the U.S. shows that there is still a similar correlation between the two.

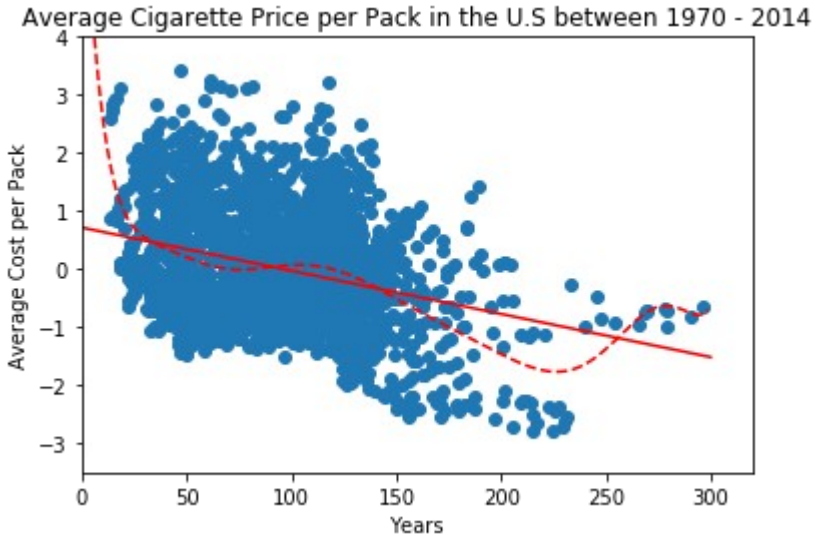
| | LocationDesc | Year | SubMeasureDesc | Data_Value | \ |
|---|---|---|---|---|---|
| 0 | Alabama | 1970 | Average Cost per pack | 0.427 | |
| 1 | Alaska | 1970 | Average Cost per pack | 0.418 | |
| 2 | Arizona | 1970 | Average Cost per pack | 0.385 | |
| 3 | Arkansas | 1970 | Average Cost per pack | 0.388 | |
| 4 | California | 1970 | Average Cost per pack | 0.397 | |

| | GeoLocation | Mean | STD | standard |
|---|---|---|---|---|
| 0 | (32.84057112200048, -86.63186076199969) | 0.380745 | 0.041286 | 1.120360 |
| 1 | (64.84507995700051, -147.72205903599973) | 0.380745 | 0.041286 | 0.902367 |
| 2 | (34.865970280000454, -111.76381127699972) | 0.380745 | 0.041286 | 0.103060 |
| 3 | (34.74865012400045, -92.27449074299966) | 0.380745 | 0.041286 | 0.175724 |
| 4 | (37.63864012300047, -120.99999953799971) | 0.380745 | 0.041286 | 0.393717 |

According to the linear regression line, the cigarette consumption tends to go down when the price values increase and vice versa.




Here I draw a linear regression line to show roughly how the results look like. I used first degree polynomial and 10th degree polynomial to compare how they are different by using polyfit and poly1d functions.

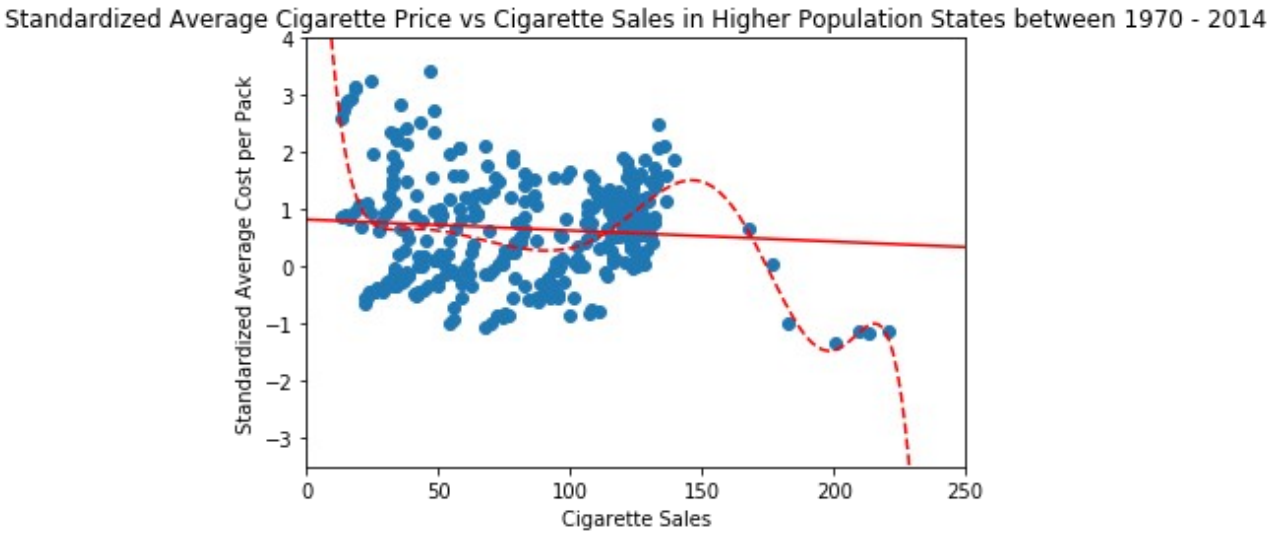


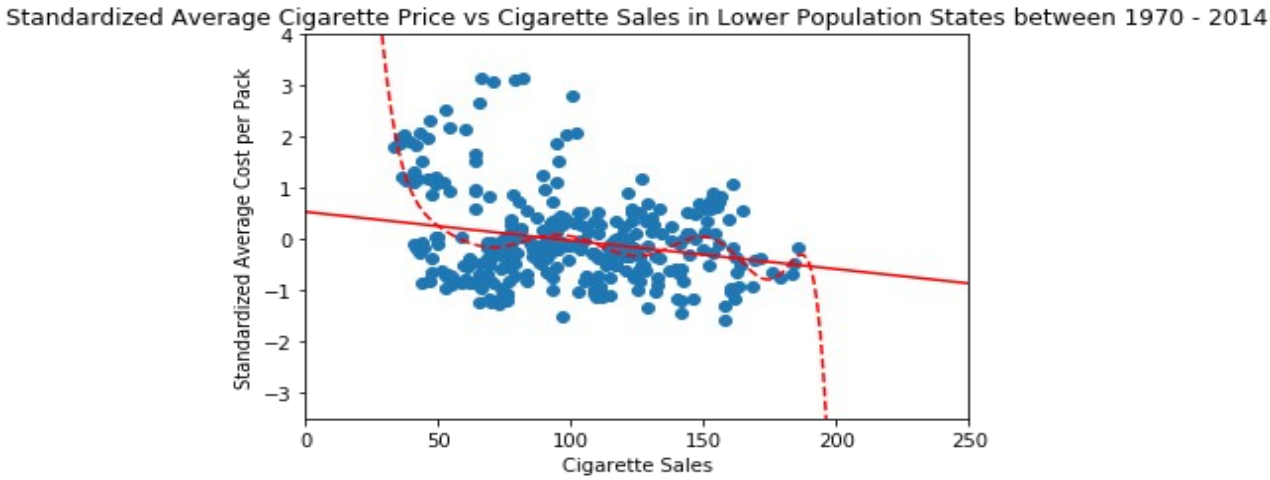

#### 4.2.8.4 Sales Trend by Location

At this point, I need to check if the data analysis is different based on location in order to prove that the result is not dependent on a location, because **a location could be what determines the amount of cigarette consumption the most**. I will make different groups of regions, compare the data and see if they have similar trends on price vs sales.

The first group deals with city and rural areas to compare the distribution in higher populated areas versus country sides. Here I made 'location_dataframe' function to do the data processing part faster and more effectively. The function will take the name of states and return dataframes after putting the input states.

We also need 'location_plot' function to plot the multiple pairs of dataframes. This will take two dataframes and a comment string and return a graph plot with a linear regression model.

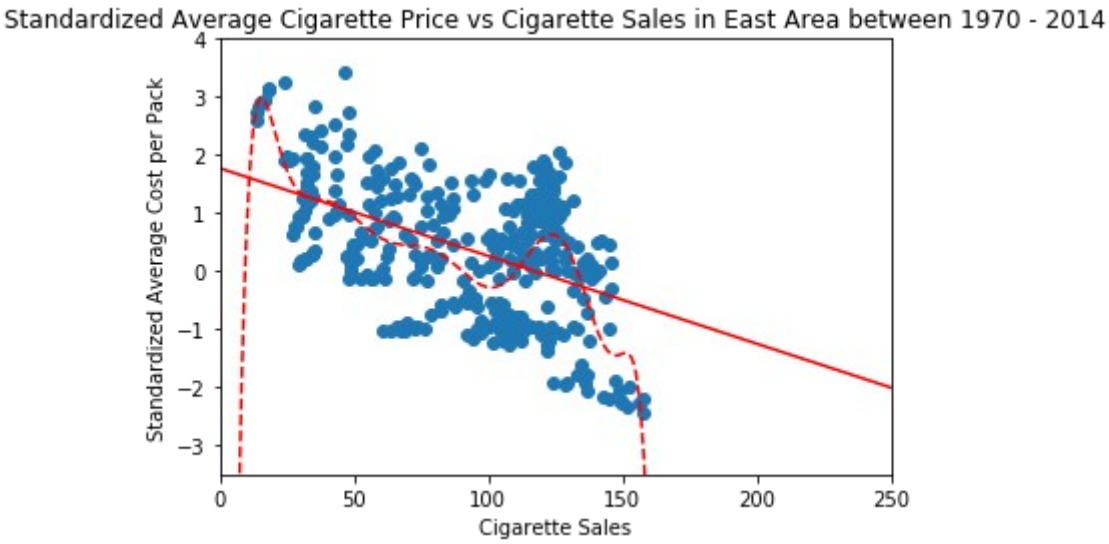


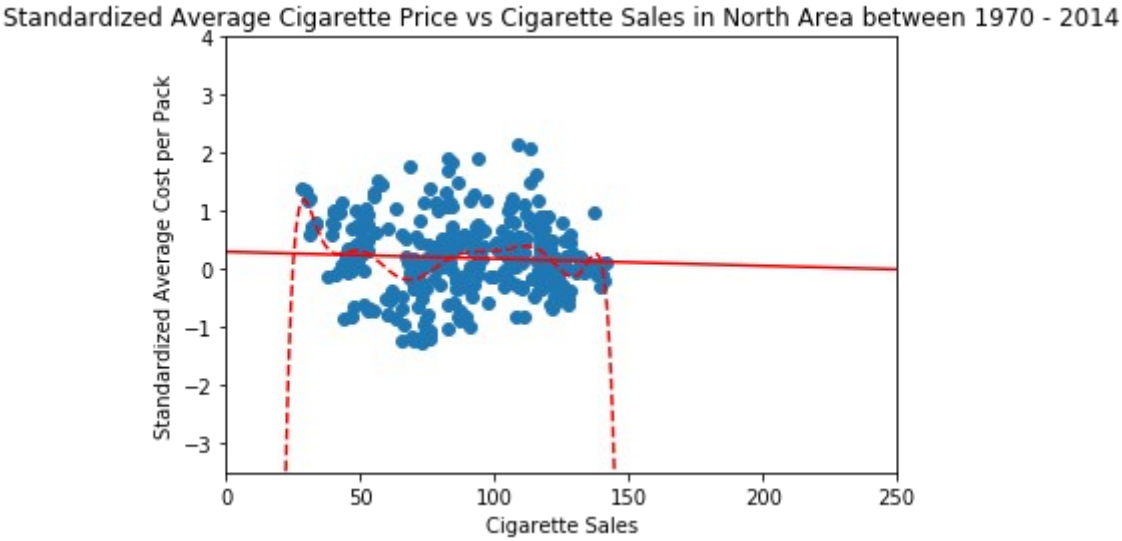



Now let's look at the results. *Both city and rural areas have a similar plot of the overall plot with a negative correlation.*

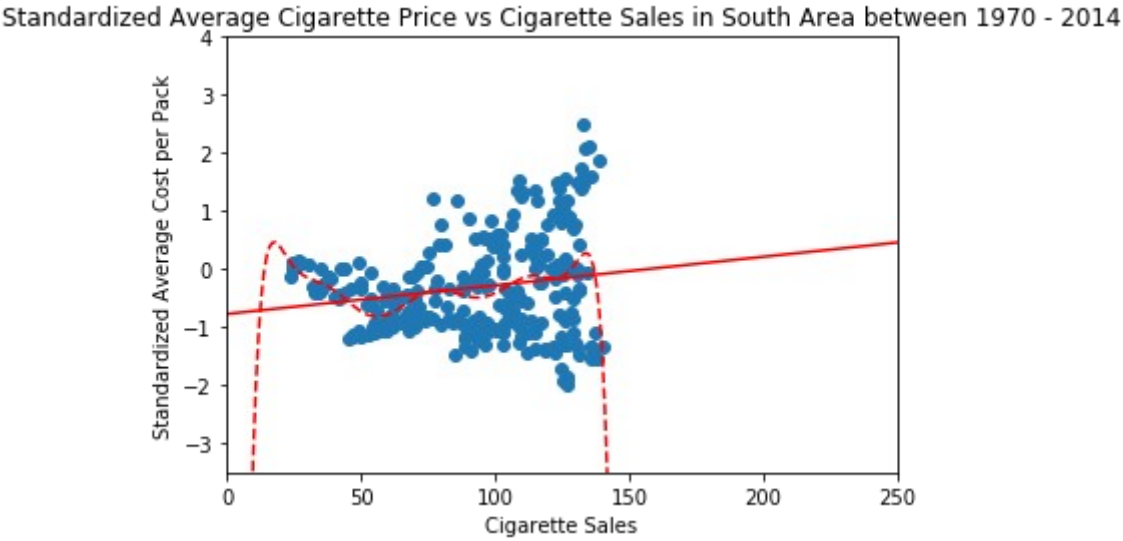
However, the **higher populated areas tend to have more extreme values on the top left and bottom right side than the lower populated areas.** I expect that this is because the higher populated areas have a bigger sample size so that this provides more distinct differences on the cigarette consumers' behaviour.

After the comparison of city and rural areas, I performed additional tests on four different regions.

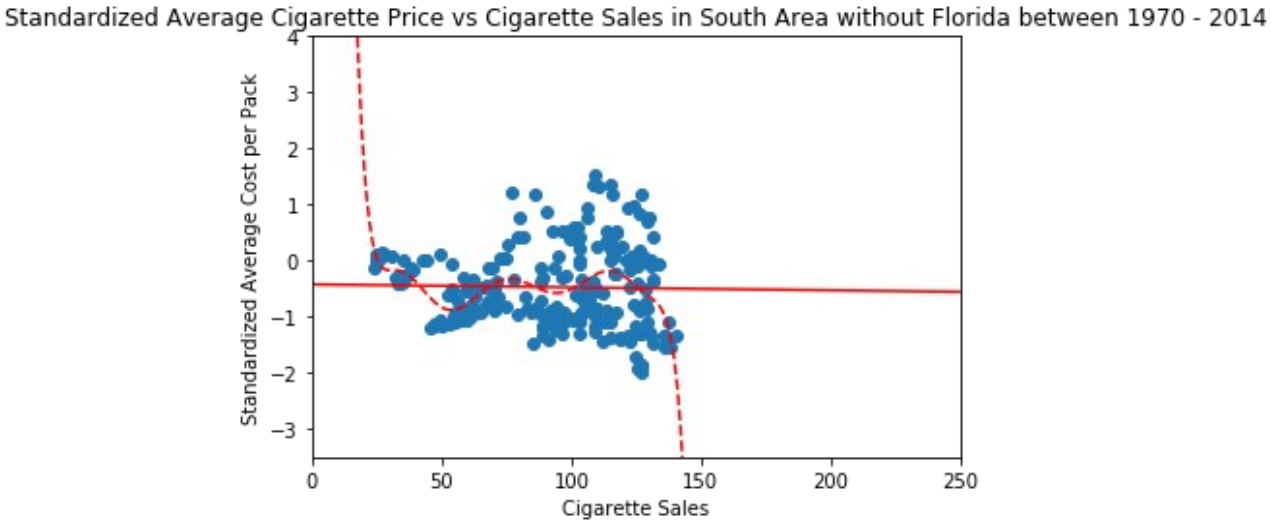



Overall, they have an inverse tendency in 1 dimensional polyfitting, except southern states.



I found out that this is because the cigarette consumers in Florida are not sensitive to the price. I expect that this is because there are lots of tourists visiting the state and many retirees live in the state. Typically, tourists and retirees are not sensitive to the price of cigarettes than other types of consumers.




The southern area shows a slight decreasing plot like other three samples without Florida. Also, the consumers in the east side are most sensitive to the price of cigarettes. It is surprising that most tested areas showed inverse correlations between cigarette prices and sales.

Also, it is interesting that we can see the different reactions from consumers on the changes of price based on the location.

### 4.2.8.5. Hypothesis Test and Machine Learning¶

In machine learning, hypothesis function is a candidate and most likely true formula y = h(x) in a machine learning algorithm and we put training data(historical records) (x,y) sets into the algorithm. The algorithm will return a new example x in y = h(x) from a prediction phase. *A null hypothesis of the hypothesis question is that there is no correlation between cigarette price and sales*. According to the result, **the test rejects the null hypothesis**( a correlation of price and sales = 0 ) since the p-value is less than significance alpha 0.05.

Here I have used built-in machine learning functions from sklearn library to execute the machine learning tests. Since the tests will not accept float values, we need to convert the float values into int values and reshape the format of the dataframe.

By using a cross_validation function, we can split test and training sets for each variable before we run the test. Use classifier function in sklearn and fit the train sets.

A score function returns the coefficient of determination $R2$ which tells the extent of the correlation. I tested with KNN, SVC, and Logistic Regression. Every test gives around 0.7 ~ 0.8 of accuracy of the test and SVC is slightly more effective than the others.

- KNN = 0.7254901960784313
- SVC = 0.7222222222222222
- Logistic Regression = 0.7222222222222222

This statistic test provides regression results which include the p-value. We have a p-value of 0.022 on the original price and sales data. This means that the null hypothesis ( a correlation between price and sales = 0 ) is rejected since the p-value is less than significance alpha 0.05.

### 4.2.9. Conclusion:

Finally, we are in the process of making a conclusion on the entire data analysis process. We performed the Exploratory Data Analysis(EDA) to analyze the relationship between tobacco sales and tobacco prices over time.

As already mentioned, we have used two datasets for undersanding the analysis of tobacco supplement. From the first dataset, we understood the consumption of different tobacco products, and its consumption on different section of people, based on their gender, race, education and location. One thing which is common in every case is that the number of persons consuming tobacco products are decreasing over time, whether they are male or female, american or african, educated or uneducated, belong to Higher Population States or lowers one. This may be a good sign as tobacco has predominatly negative effects on human health and concern about health effects of tobacco has a long history.

One thing which we can also conclude from this EDA analysis that the ladies smokers are diminished more radically than men smokers. A few explanations behind this gendered slack incorporate social procedures, for example, proscriptive standards against smoking take-up by ladies, credit of negative social generalizations to ladies who smoke, and a solid socially informed contradiction with ladies' conceptive jobs.

For the second dataset, we have tried multiple different input values with a given hypothesis function to find an optimized function. Now, we can determine whether our hypothesis is right or wrong with the best test results. According to the linear regression and machine learning test, we can conclude that the mean price of tobacco products and the mean sales of it have a negative correlation.

However, we cannot say that they are in a causation relationship since a correlation has nothing to do with a causation. Even though we cannot tell one causes the other, the result of the analysis is valuable because it could give a hint of ways to control cigarette sales. However one thing we will say to smokers that thety need to quit completely rather than cut down if they wish to avoid most of the risk associated with heart disease and stroke, two common and major disorders caused by smoking.

## 4.3 Technologies

### 4.3.1 Python

Python is a general-purpose programming language that is becoming more and more popular for doing data science. Companies worldwide are using Python to harvest insights from their data and get a competitive edge.

### 4.3.2 Jupyter Notebook

The Jupyter Notebook is an open-source web application that enables you to make and share records that contain live code, conditions, representations and story content. Utilizations include: information cleaning and change, numerical reproduction, factual displaying, information representation, machine learning, and substantially more.

### 4.3.3 Libraries

#### 4.3.3.1. Numeric and Scientific Computation (NumPy and SciPy):

NumPy or Numeric and Scientific Computation provides with fast precompiled functions for mathematical and numerical routines. In addition, NumPy optimizes Python programming with powerful data structures for efficient computation of multi-dimensional arrays and matrices.
Scientific Python also is known as SciPy is inextricably linked with NumPy. SciPy lends a competitive edge to NumPy, by enhancing useful functions for regression, minimization, Fourier-transformation, and many more which is used in this project.

Source: http://www.numpy.org/

#### 4.3.3.2. PANDAS:

Pandas is an open source tool that provides high-performance, easy-to-use data structures and data analysis tools for Python programming. It is used to add data structures and tools designed for practical data analysis in multiple streams such as finance, statistics, social sciences, and engineering. The best part of Pandas is its easy adaptability, which makes it one of the top Python Libraries for Data Science. It can work perfectly well with incomplete, unstructured, messy, and uncategorized data. It can, at the same time provide tools for shaping, merging, reshaping, and slicing of datasets.

Source:

http://pandas.pydata.org/

https://www.scipy.org/scipylib/index.html

#### 4.3.3.3. Matplotlib:

Matplotlib is a plotting library, capable of producing publication quality figures in a wide variety of hardcopy formats and interactive environments across platforms. Matplotlib, which is also one of the top Python Libraries for Data Science, is used for generating plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with fewer codes. For simple plotting, the pyplot module provides a MATLAB-like interface, particularly when combined with IPython.

Source:

https://matplotlib.org/

#### 4.3.3.4. SciKit-Learn:

Scikit-learn is a module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. It is one of the best-known machine-learning libraries for python. The Scikit-learn package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. The primary emphasis is upon ease of use, performance, documentation, and API consistency.

Source:

https://scikit-learn.org/stable/

### 4.3.5 GG PLOT

Ggplot is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics . It is built for making profressional looking, plots quickly with minimal code.

Source:
http://ggplot.yhathq.com/

### 4.3.6 Standard Scalar

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. Given the distribution of the data, each value in the dataset will have the sample mean value subtracted, and then divided by the standard deviation of the whole dataset.

StandardScaler will normalize the features (each column of X) so that each column/feature/variable will have mean = 0 and standard deviation = 1.

### 4.3.6 Extra Tree Classifier

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

### 4.3.7 Logistic Regression

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

### 4.3.8 Random Forest Regression

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models.

Different kinds of models have different advantages. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

# 5. Challenges and Barriers

## 5.1 Choosing the technology & frameworks

## 5.2. Integration

## 5.3. Barriers

Oversampling, undersampling or hybrid techniques on training set.
Unstructured Dataset