QMBE 1320 Final Project

Purpose

The purpose of this **individual/pair** final project is to put to work the tools and knowledge that you gain throughout this course. This provides you with multiple benefits.

1. It will provide you with more experience using data wrangling tools on real life data sets.
2. It helps you become a self-directed learner. As a data scientist, a large part of your job is to self-direct your learning and interests to find unique and creative ways to find insights in data.
3. It starts to build your data science portfolio. Establishing a data science portfolio is a great way to show potential employers your ability to work with data.

I plan to have you work on the project and use some of the in-class time to do peer evaluation of your projects.

Project Goal

The principal goal of this project is to import a real-life data set, clean and tidy the data, and perform basic exploratory data analysis; all while using Excel and Word to create a report that is clean and professional.

Project Data

You will need to select one data set from the four that I have supplied below. All four data sets contain key attributes that will demonstrate the data science capabilities that you have learned throughout this course. You may even need to learn new skills not taught to accomplish your mission. These include working with:

- multiple data types (numerical, characters, dates, etc)
- non-normalized characteristics (may contain punctuations, upper and lowercase letters, etc)
- data sets that need to be merged
- unclean data (missing values, values that do not align to the data dictionary)
- variables that need to be created (i.e. the data may contain income and expense variables but you want to analyze savings such that you need to create a savings variable out of the income and expense variables)
- data that needs to be filtered out
- and much more!

Available data sets include:

You can choose from one of the following four data sets. Each dataset has its own challenges and strengths.

- Petfinder.com Dog Data
- Hotel Bookings Data
- NFL Attendance Data
- Spotify Genre Data

Report
You will write a Word Document report that provides the sections in the grading rubric below. You will need to import, assess, clean & tidy the data, and then come up with your own research questions that you would like to answer from the data by performing exploratory data analysis (if you'd like to perform a predictive model to answer your hypothesis i.e. regression etc, that is fine)

Some thoughts to help you:

- Make a storyboard. Your project should be a logical, cohesive story–not simply a bunch of graphs created for the sake of making them. The story may change as you dive deeper into the data and find insights, but a storyboard gives you direction and purpose for developing insights. Clear writing means a clear mind, and a storyboard is vital to producing a good story.
- Speaking of insights, keep in mind that your project should follow the chain of data -> insights -> actions. As a future data analyst (or data scientist, or statistician, or whatever is trendy next year), you work to create insights that lead to actions, not to waste 40 hours on a awe-inspiring visualization that is ignored directly after a presentation and never used again.
- Simple descriptive statistics can (and usually) yield more of an immediate impact than a complicated model. Brooke Watson gave a compelling and enlightening presentation at the 2019 RStudio Conference on how the ACLU used various R packages to count and reunite families.
- Do subgroups matter in your data?
- Why are data missing?
- Are trends over time important?

Although each data set's data dictionary contains some additional questions worth pursuing, try to be creative in your analysis and investigate the data in a way that your classmates most likely will not. Creativity is an essential ingredient for a good data scientist!

Other Expectations
Upon submission you will upload the final Word file and the final Excel file. Your submitted files should be named with year, course number, lastname(s), first name(s) and then "finalproject."

For example my file name would be: 2024_QMBE1320_Odongo_finalproject.doc.

I expect your report to tell a story with the data. I do not want you to just report some statistics that you find but, rather, to provide a coherent narrative of your findings. Examples for what I am looking for are included below.

Any additional details regarding the final project will be provided in class.

Stage 1 Sample Format

1. **What data did you choose and why? (1 paragraph)**

2. **Write 2-3 Problem Statements**
    a. Problem statement 1: Example – I want to find out how customer satisfaction affects retention and revenue for our online store

    b. Problem Statement 2: Example – I want to increase sales by identifying the factors that influence customer behavior.

    c. Problem Statement 3: Example – I want to understand the relationship between air pollution and respiratory diseases.

What is your plan to analyze these problem statements? How will this analysis help the consumer or business? (1 paragraph)

3. **How did you clean your data? (1-2 paragraphs)**

4.

[Neat Descriptive Statistics (Table) Here]

Have you thought about creating new variables or combining data sets? What variables can you create that you might use for your analysis? How would you create them? (1 paragraph)

5. **Visualize Data: (Several graphs or tables depending on your "important variables)**

[Neat/Professional Table or Graph Here]

[Neat/Professional Table or Graph Here]

[Neat/Professional Table or Graph Here]

What insights or observations can you make from your data? Do you notice patterns or trends?

Rubric: STAGE 1

Section	Standard	Possible Points
Introduction	**1.1** Provide an explanation of what data you chose and why. **1.2** List 2- 3 problem statements you may want to address. **1.3** Why should I be interested in this? **1.4** Provide a short explanation of how you plan to address this problem statement (the data used and the methodology employed) **1.5** Explain how your analysis will help the consumer of your analysis. (**You may want to write this last after visualizing and studying your data)**	25
Professionality	**2.1** The written report submission is professional, neat, and clear (as if turning in to your manager or to a company who is paying for your data analysis services.)	15
Data Preparation	**3.1** Clean your data. **3.2** Cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process. **3.3** Provide summary information about the variables of concern in your cleaned data set. **3.4** What are ways you can slice and dice the data, create new variables, or join separate data frames to create new summary information.	20
Exploratory Data Analysis	**4.1** Visualize your data. Apply visualization tips and techniques discussed in class. **4.2** Provide several plots and tables summarizing the variables of concern (Minimum 5). **4.3** Graph(s) are carefully tuned for desired purpose. One graph illustrates one primary point and is appropriately formatted (plot and axis titles, legend if necessary, scales are appropriate, etc.). **4.4** Table(s) carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features. Size of table is appropriate. **4.5** What insights do the tables and graphs give you?	25
Formatting & Other Requirements	**5.1** Include the Excel File in your submission. **5.2** The Excel File is neat and easy to follow. 5.3 Use of space and additional sheets (tabs at the bottom) is effective. **5.4** Several techniques and tools from the course are used.	15

Total possible points: 100

Due no later than: Sunday, November 26, 11:59PM

Stage 2 – Sample Format

Introduction:

(1 Paragraph) What problem statement(s) are you addressing. Why should I be interested in this? Why is it important?

(1-2 Paragraphs) How do you plan to address this? What data are you using? (Cite this source) What methods or approach are you going to use? (e.g. creating new variables to analyze, linear regressions, forecasting and regression etc.)

(1 Paragraph) How will this analysis help the consumer or business?

Data: **(1 Paragraph)** Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

(1 Paragraph) Data importing and cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.

Neatly and briefly show final and cleaned data:

[Short, Neat, Professional Table of Data Here]

[Short, Neat, Professional Descriptive Statistics of Data Here]

Briefly explain the variables you have displayed and summarized.

Exploratory Data Analysis:

Present your findings in a logical way:

Example: (Paragraph 1) We found that …….

[Here is the graph or table to show it]

(Paragraph 2) We found that advertising expenditures affect the number of sales by (this) much…

[Here is the regressions results to show it – Notice the p-value is statistically significant]

Our model is a good fit and can be trusted because the R squared is….. Etc……

Summary:

(1-2 Paragraphs) Summarize the problem statement, how you addressed this problem statement, the interesting insights you found, and the implications for consumers or the business.

(1 Paragraph) Discuss the limitations of the analysis. How could someone improve on it.

Rubric: STAGE 2

Section	Standard	Possible Points
Introduction	**1.1** Provide an introduction that explains the problem statement you are addressing. Why should I be interested in this? **1.2** Provide a short explanation of how you plan to address this problem statement (the data used and the methodology employed) **1.3** Discuss your current proposed approach/analytic technique you think will address (fully or partially) this problem. **1.4** Explain how your analysis will help the consumer of your analysis.	15
Professionality	**2.1** Each submission is professional, neat, and clear (as if turning in to your manager or to a company who is paying for your data analysis services.)	10
Data Preparation	**3.1** Original source where the data was obtained is cited. **3.2** Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.). **3.3** Data importing and cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process. **3.4** Once your data is clean, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible. (Short Table of first few rows) **3.5** Provide summary information about the variables of concern in your cleaned data set. Provide me with a consolidated explanation, either with a table that provides summary info for each variable or a nicely written summary paragraph for each variable.	25
Exploratory Data Analysis	**4.1** Uncover new information in the data that is not self-evident (i.e. do not just plot the data as it is; rather, slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information). **4.2** Provide findings in the form of plots and tables. Show me you can display findings in different ways. **4.3** Graph(s) are carefully tuned for desired purpose. One graph illustrates one primary point and is appropriately formatted (plot and axis titles, legend if necessary, scales are appropriate, etc.). **4.4** Table(s) carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features. Size of table is appropriate. **4.5** Insights obtained from the analysis are thoroughly, yet succinctly, explained. Easy to see and understand the interesting findings that you uncovered.	25

Section	Standard	Possible Points
Summary	**6.1** Summarize the problem statement you addressed. **6.2** Summarize how you addressed this problem statement (the data used and the methodology employed). **6.3** Summarize the interesting insights that your analysis provided. **6.4** Summarize the implications to the consumer of your analysis. **6.5** Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.	15
Formatting & Other Requirements	**7.1** Include the Excel File in your submission. **7.2** The Excel File is neat and easy to follow. **7.3** Achievement, mastery, cleverness, creativity: Tools and techniques from the course are applied very competently and, perhaps, somewhat creatively. Perhaps student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, unusually sophisticated application of tools from course.	10

Total possible points: 100
Due no later than: The day of your Final Exam!

Section 1-- 10:20am-11:20am: Tuesday, December 12, 11:59PM
Section 2-- 11:30am-12:30am: Wednesday, December 13, 11:59PM

Example QMBE 1320 – Final Project

By

Group Member 1 and Group Member 2

Introduction:

Per capita income and expenditures provide crucial insight into the average standard of living in specified areas. Disposable per capita income measures the average income earned after taxes per person in a given area (city, state, country, etc.) in a specified year. It is calculated by dividing the area's total income after tax by its total population. Per capita expenditures, on the other hand, measures the average outlay for goods and services by person and provides insight into spending patterns across a given area. Together, the assessment of per capita income versus expenditures can provide better understanding of regional economies, differences in standard of living, and approximate savings rates.

This project involves exploring Bureau of Economic Analysis data regarding per capita disposable income (hereafter referred to as PCI) and per capita personal expenditures (hereafter referred to as PCE). The PCI data provides annual (non-inflation adjusted) per capita disposable income at the national and state-level from 1948-2015 and the PCE data provides annual (non-inflation adjusted) per capita personal consumption expenditures at the national and state-level from 1997-2014. Consequently, this research seeks to identify how the national and state-level savings rates defined as

$$Savings = PCI - PCE$$

has changed over time and by geographic location.

The analysis finds that the national-level and average state-level savings rates have remained around 7-8% since 1997. Furthermore, we find that American's are not making fundamental shifts in their earnings and expenditure rates. However, the analysis does uncover a noticeable shift in the disparity of savings rates across the states in recent years with much of the growth in savings rates being concentrated in the central U.S. states - from the Dakotas down to Oklahoma, Texas and Louisiana. Consequently, it appears that the often-neglected fly-over states offer Americans greater opportunities to save than the eastern and western states.

Data Preparation:

Prior to assessing how PCI, PCE, and savings rates have behaved over time and by geographic location we must acquire and clean the data.

The data for this project originated from the following sources:

- PCI data: http://bit.ly/2dpEPY1
- PCE data: http://bit.ly/2dhC89U

Cleaning Data: Once the basic data has been acquired we need to pre-process it to get the data into a tidy format. This includes removing punctuations, changing the income and expenditure data from character to a numeric data type, reducing the data sets to the same time period (1997-2014), making sure the common variables share the same names, and changing the data from a wide format to a long format. Once this has been done for both the PCI and PCE data we can merge the

clean data frames into one common data frame (titled *data_clean*) and create a new *Savings* variable (*Savings = Income − Expenditures*). I also remove the District of Columbia location since this is more comparable to metropolitan-level geographic areas than state-level geographic areas. We now have the data cleaned, in a tidy format, and ready to analyze as Table 1 illustrates.

Table 1: Clean and tidy data.

	Fips	Location	Year	Income	Expenditures	Savings
1	0	United States	1997	22536	20384	2152
2	1000	Alabama	1997	19050	17243	1807
3	2000	Alaska	1997	24803	23320	1483
4	4000	Arizona	1997	19956	19223	733
5	5000	Arkansas	1997	17954	16151	1803
6	6000	California	1997	23430	20848	2582
7	8000	Colorado	1997	23593	22605	988
8	9000	Connecticut	1997	29178	25122	4056
9	10000	Delaware	1997	23032	22335	697
10	12000	Florida	1997	22538	20445	2093

Showing 1 to 10 of 918 entries Previous 1 2 3 4 5 ... 92 Next

Exploratory Data Analysis:

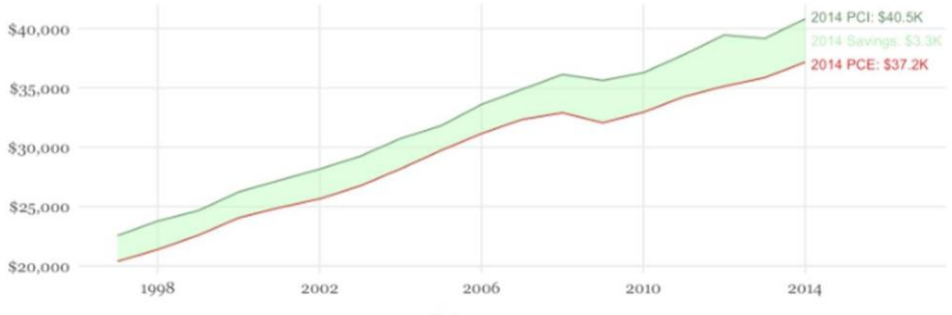
The primary purpose of this analysis is to assess how national and state-level PCI, PCE, and savings rates have changed over time and by geographic location. Thus, we will proceed by first assessing the national-level trends and then move on to assessing state-level trends.

National Level Patterns:

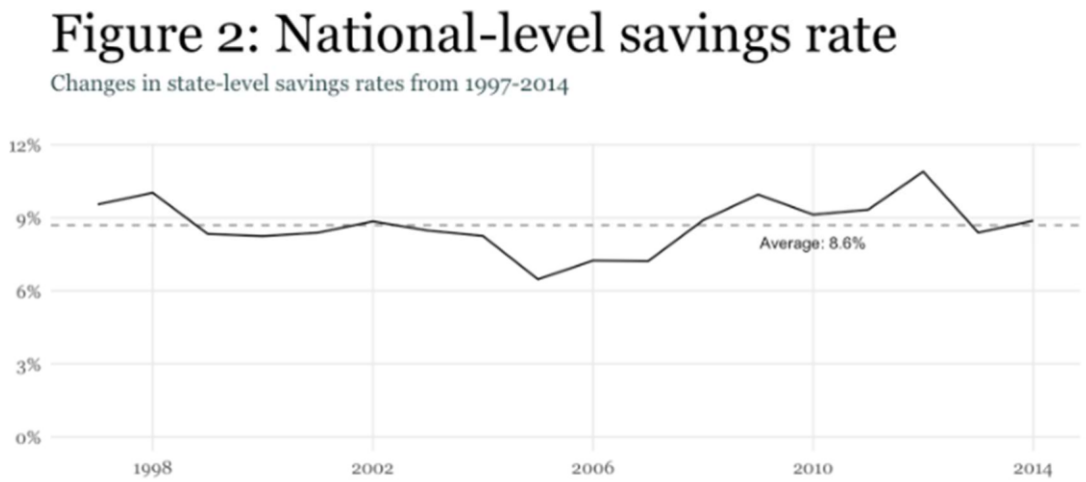
At the national-level PCI grew by 79.6% from $22,536 in 1997 to $40,471 in 2014. Expenditures (PCE), on the other hand, grew 82.5% from $20,384 in 1997 to $37,186. Although we are assessing non-inflation adjusted dollars, we can still observe that since 1997 the rate of growth in PCE has only slightly outpaced PCI. Figure 1 illustrates the growing trends (not surprising since inflation has not been removed) and also captures the decrease in both PCI and PCE from 2008 to 2009 due to the Great Recession.

Figure 1: Growth in PCI and PCE

Growth represented as current year dollars from 1997-2014 (not adjusted for inflation)



However, a closer look at just the savings rate ($SavingsRate = \frac{Savings}{Income}$) depicted in Figure 2 illustrates that no consistent trend has been established. In other words, the aggregate per capita savings rate has not consistently increased or decreased year-over-year. In 1998 the savings rate was 10% but reduced over the next few years to 6.5% in 2005 before peaking at 10.9% in 2012 and then dipping back down to about 8-9% in recent years. Bottom-line is that since 1997 the national-level per capita savings rate has ranged between 6.5% and 10.9% with an average of 8.6%.



However, understanding aggregate ratios and trends provides limited insight regarding lower-level activity [1]. Consequently, next we turn to investigating state-level trends.

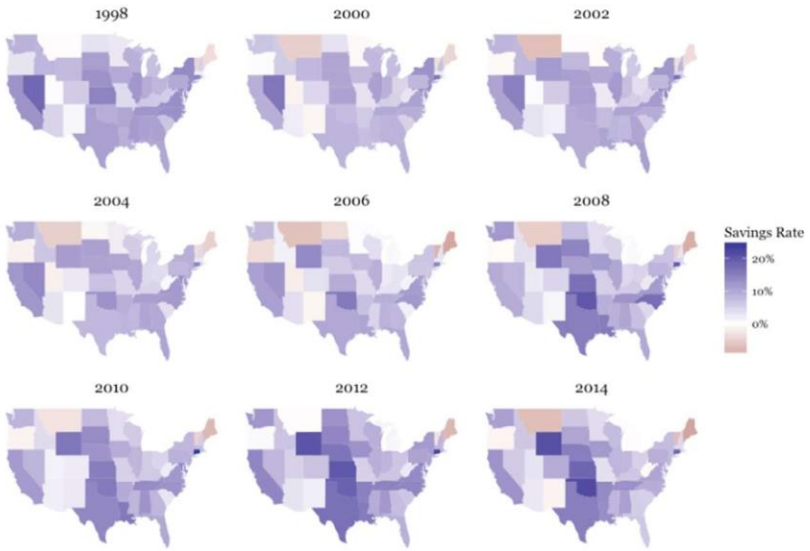
State Level Patterns:

To get a quick understanding of how U.S. states have progressed over the years we can map the savings rates over time. Figure 3 highlights a few attributes:

1. Note how the earlier years have less diverging colors suggesting that there was more "equality" in the savings rates across the states; however, the latter years appear to have more disparity in the savings rates
2. As the years have progressed it appears that a growth in savings rates has been concentrated in the central states; primarily from the Dakotas down to Texas
3. A few individual states stand out:
    - Main, Vermont & Montana for savings rates that are consistently less than 0%
    - Massachusetts for consistently being a top savings rate state

Figure 3: Savings rate changes over time

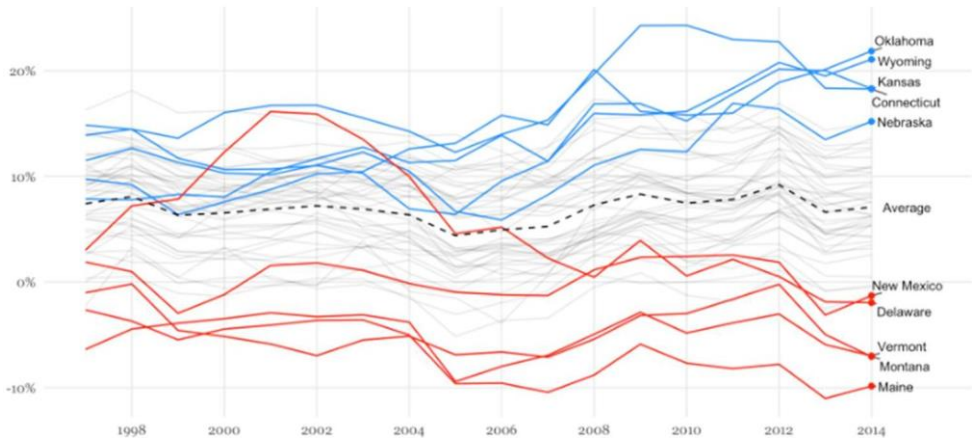
Temporal map assessment of state-level savings rates (1998-2014)



A closer look at the state-level trends provides more insight. We can see that the average savings rate over time has remained around 7%; however, confirming our assessment of the map it appears that the variance (or disparity in savings rates) has increased in recent years. Moreover, the trend lines illustrate that with a few exceptions, states that are leading the way as top or bottom savings rate states have, historically, always been near the top or bottom. However, this should not be too surprising as it takes decades for states to change their industrial and economic infrastructure.

Figure 4: Savings rate changes over time

Temporal assessment of state-level savings rates (1997-2014)



However, we can also look at those states that have had the largest change in their savings rate since 1997. As Table 2 displays, three of the four states with the largest change in their savings rate were Wyoming, Oklahoma and North Dakota; all having a savings rate increase close to, or more than, 10%. The remaining states with the largest changes have all experienced declining savings rates, led by Nevada.

Table 2: Top 10 states with the largest change in their savings rate since 1997

Location	1997	2014	Change
Wyoming	7.9%	21.1%	13.2%
Oklahoma	9.7%	21.9%	12.1%
Nevada	16.3%	5.9%	-10.4%
North Dakota	-2.2%	6.3%	8.5%
Maine	-2.6%	-9.8%	-7.2%
Michigan	6.7%	-0.3%	-7.1%
New York	14.4%	7.5%	-6.9%
West Virginia	6%	-0.5%	-6.5%
Montana	-1%	-7%	-6%
New Jersey	13.9%	8%	-5.9%

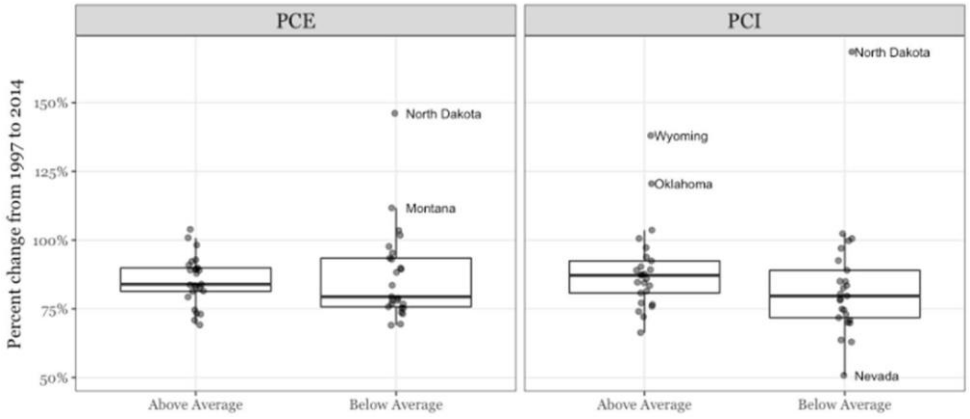
This may lead us to wonder if one component (PCI vs PCE) is driving the changes in savings rate. In other words, for those states that are growing above the average level, is their PCI level growing at a greater level than those states below the average? Or could it be that the those states with above average savings rates are experiencing a slower increase in their expenditures than those states below average. Figure 5 helps to illustrate this issue.

Figure 5 shows that, concerning PCE (left pane), the states that have had above average savings rates have not experienced, on average, any difference in PCE growth since 1997. However, the states with below average savings rates have experienced greater variance in their PCE growth rates. Concerning PCI (right pane), the states that have had above average savings rates have experienced, on average, slightly greater PCE growth since 1997; however, this difference is likely not to be statistically significant (though validation would be required to confirm). Again, those states with below average savings rates have experienced slightly greater variance in their growth rates than the above average savings rate states.

Thus, it appears that those states with below average savings rates have greater variability in their PCI and PCE growth rates whereas those states with above average savings rates have more consistency. However, no significant differences appear to exist in the average PCI & PCE growth rates among states with above versus below average savings rate. This is likely why we are seeing the average savings rate remain relatively steady but the variability in savings rates among the states growing.

Figure 5: Percent change in PCE & PCI

Comparing the change in PCE & PCI from 1997 to 2014 for those states with above versus below average savings rates



Summary:

Consequently, our analysis finds that the national-level and average state-level savings rates have remained around 7-8% since 1997. Furthermore, we find that PCI and PCE have grown at a relatively similar rates at the national, state-levels, and among those states that have experienced above versus below average savings rates. This suggests that the U.S. has not experienced a fundamental shift in PCI or PCE behavior.

The noticeable change that we have seen is a greater disparity in savings rates among the states in recent years. Although the average savings rate has remained around 7-8%, the variance in state-level savings rates has grown since 1997. Moreover, much of the above average growth in savings rates has been concentrated in the central U.S. states from the Dakotas down to Oklahoma, Texas and Louisiana; whereas much of the below average growth has been concentrated in more eastern and western states. Thus, if you are looking to save more of your hard-earned income you may have greater opportunities by seeking refuge in one of the fly-over states.