**Question 2.3:** What do the values in the "Description" column represent? Are the categories in the "Description" column mutually exclusive, or are they potentially subsets of each other? Give an example to illustrate your point.

```
In [11]: print(rgdp['Description'].unique())
```

```
['All industry total' ' Private industries'
 '  Agriculture, forestry, fishing and hunting'
 '  Mining, quarrying, and oil and gas extraction' '  Utilities'
 '  Construction' '  Manufacturing' '  Wholesale trade' '  Retail trade'
 '  Transportation and warehousing'
 '  Finance, insurance, real estate, rental, and leasing'
 '  Educational services, health care, and social assistance'
 '  Arts, entertainment, recreation, accommodation, and food services'
 '  Other services (except government and government enterprises)'
 'Government and government enterprises']
```

It is seen from the array above how the first listed value is **All industry total**. This means the column cannot be mutually exclusive as this column is the sum of all the other categories in the description. Further we have the category **Private industries** which also seems to be a supercategory only excluding the category **Government and gorvernment enterprises**. The remaining descriptions, however, seem to be mutually exclusive as they are subcategories of the private industry.

**Question 2.4:** What are the data types of columns `GeoFIPS`, `GeoName`, `Unit`, and `2021`? Are they integers, floats, strings, objects, or mixed types? Do you find any data types in these columns problematic? Why?

*Hint:* Look into `df.dtypes`.

```
In [12]: rgdp[['GeoFIPS','GeoName','Unit','2021']].dtypes
```

```
Out[12]: GeoFIPS    object
         GeoName    object
         Unit       object
         2021       object
         dtype: object
```

From the code above, it is apparent that the columns `GeoFIPS`, `GeoName`, `Unit`, and `2021` are all type objects. This could be problematic, as, for example, the GDP data in `2021` should be either integer or float. Additionally, we should store the others as strings for performance, which uses less memory.

**Question 3.1:** Look up the footnote of this dataset, what does each of these two types of missing values represent? What do you think is a good way to handle these two types of missing values respectively? This is an open-ended question.

The values marked **(D)** are *not shown to avoid disclosure of confidential information* but *estimates are included in higher-level totals*. From the excerpt, there is no clear trend in the missing values; it is only some industry classifications for some areas. My guess is that only a few companies work within that industry, data on the companies could be inferred from the data. Values marked with **(NA)**, on the other hand, are not available. Here, from the excerpt, this is often a location where no data is available.

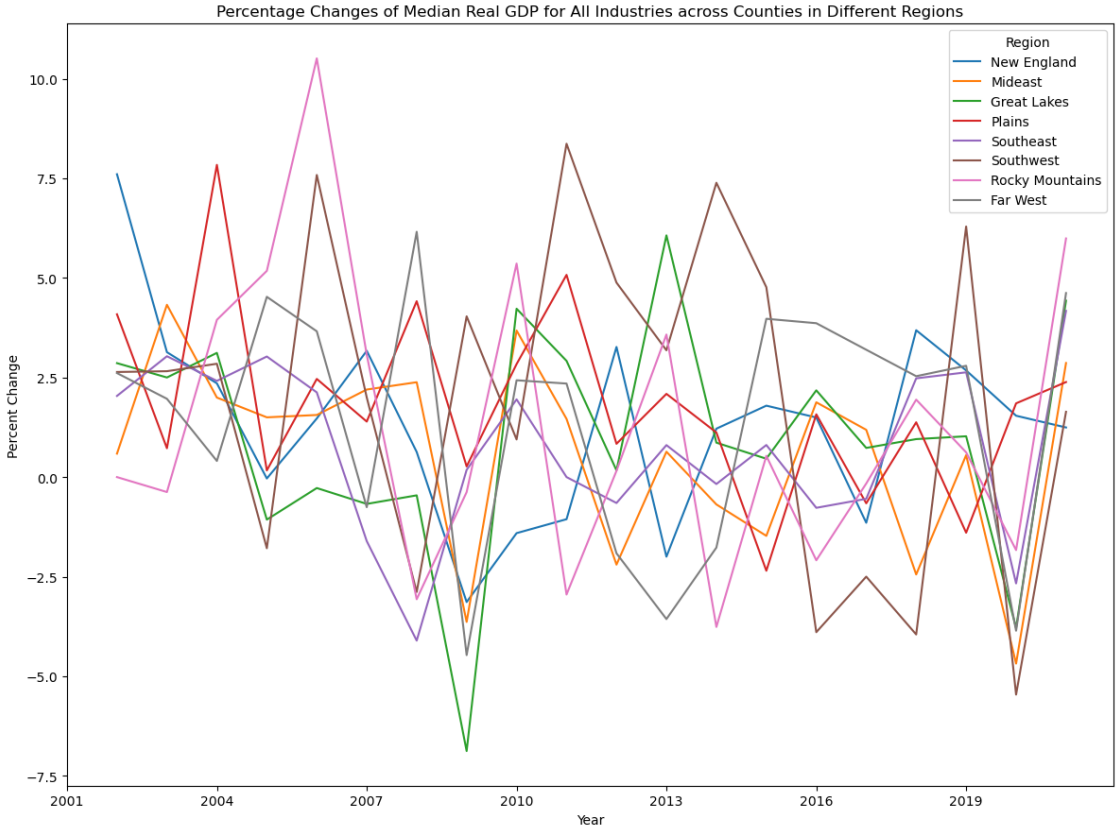
If we do not need data on such a granular level, we could drop rows marked **(D)**, while for the **(NA)**, we could either estimate them from other similar areas or drop them. If we want to analyze at a more granular level, the more appropriate thing could be to drop the entire area if it contains **(D)** or **(NA)** values. This way, the missing values will not mess with any analysis. The other solution could again be to estimate their value from similar areas. However, this would, at best, be a qualified guess.

**Question 7.2:** Now, we want to find the percent changes of median real GDP in each region with `pd.DataFrame.pct_change`. So, write the code that computes the percent changes of median real GDP in each region.

*Hint:* start by copying your code from question 7.1.

```
In [55]: plt.figure(figsize=(12, 9))
         for region in np.sort(rgdp_county_allindustry["Region"].unique()):
             rgdp_region_groupby = rgdp_county_allindustry[rgdp_county_allindustry['Region'] == region]
             rgdp_region_groupby = rgdp_region_groupby[['year','value']].groupby('year').median().reset_
             rgdp_region_pct_chg = rgdp_region_groupby.copy()
             rgdp_region_pct_chg['value'] = rgdp_region_pct_chg['value'].pct_change()
             rgdp_region_pct_chg = rgdp_region_pct_chg.dropna().reset_index(drop=True)
             plt.plot(rgdp_region_pct_chg["year"],
                      rgdp_region_pct_chg["value"] * 100, # as percentages
                      label=bea_regions[region]
                     )

         plt.xticks(np.arange(2001, 2022, 3))
         plt.xlabel("Year")
         plt.ylabel("Percent Change")
         plt.title("Percentage Changes of Median Real GDP for All Industries across Counties in Differe
         plt.tight_layout()
         plt.legend(title="Region", loc='upper right');
```


Percentage Changes of Median Real GDP for All Industries across Counties in Different Regions
Region
New England
Mideast
Great Lakes
Plains
Southeast
Southwest
Rocky Mountains
Far West
Percent Change
Year
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
2001
2004
2007
2010
2013
2016
2019

**Question 9.1:** Comment on the results above. Are the economic performance similar or different in each region? Do you find it surprising?

In **2002**, the map differentiated economic performance across the US. The distribution seems fairly even, with some experiencing growth and others experiencing a decline. There seems to be no clear tendency within each region.

In **2008** we are in the midst of the global financial crisis, and several countries are experiencing negative growth. However, interestingly, the figure shows that the *plains region* except Minnesota, fared better than the other regions and actually were growing during this period.

Lastly, in **2020** during the first year of COVID-19, there was a general tendency of negative growth across the board. Only a few counties in the *plains* and *rocky mountain* regions seem to have had positive growth.

**Question 9.2:** Let's look at the plot for 2008 that shows the regional economic performance in the midst of the Great Recession. The causes of the Great Recession include a combination of vulnerabilities that developed in the financial system, along with a series of triggering events that began with the bursting of the United States housing bubble in 2005–2012. As a sidenote, many empirical works suggest that housing crises usually accompany high levels of mortgage delinquencies (people default on their mortgage).

Look at the county-level change in mortgage delinquency figure that was used in Ben Bernanke's (the chairman of the Federal Reserve at that time) speech *Mortgage Delinquencies and Foreclosures* at the Columbia Business School's 32nd Annual Dinner in May 2008. What is the association between this mortgage delinquency graph and the regional real GDP graph we have above? How can this result potentially inform us about the causes of the Great Recession?

When comparing the plot of 2008 with the mortgage delinquency rates in the *plains region* they seem to be average at best and not better than the *southeast region* – except Florida – though the plains seem to have fared better when looking just isolated at the real GDP growth in 2008. Thus, this shows how there is a vast spillover effect between states as interest rates across the US went up. Not only did the states with high delinquency rates experience a tough 2008.