**HW 1**

**Part 1:**

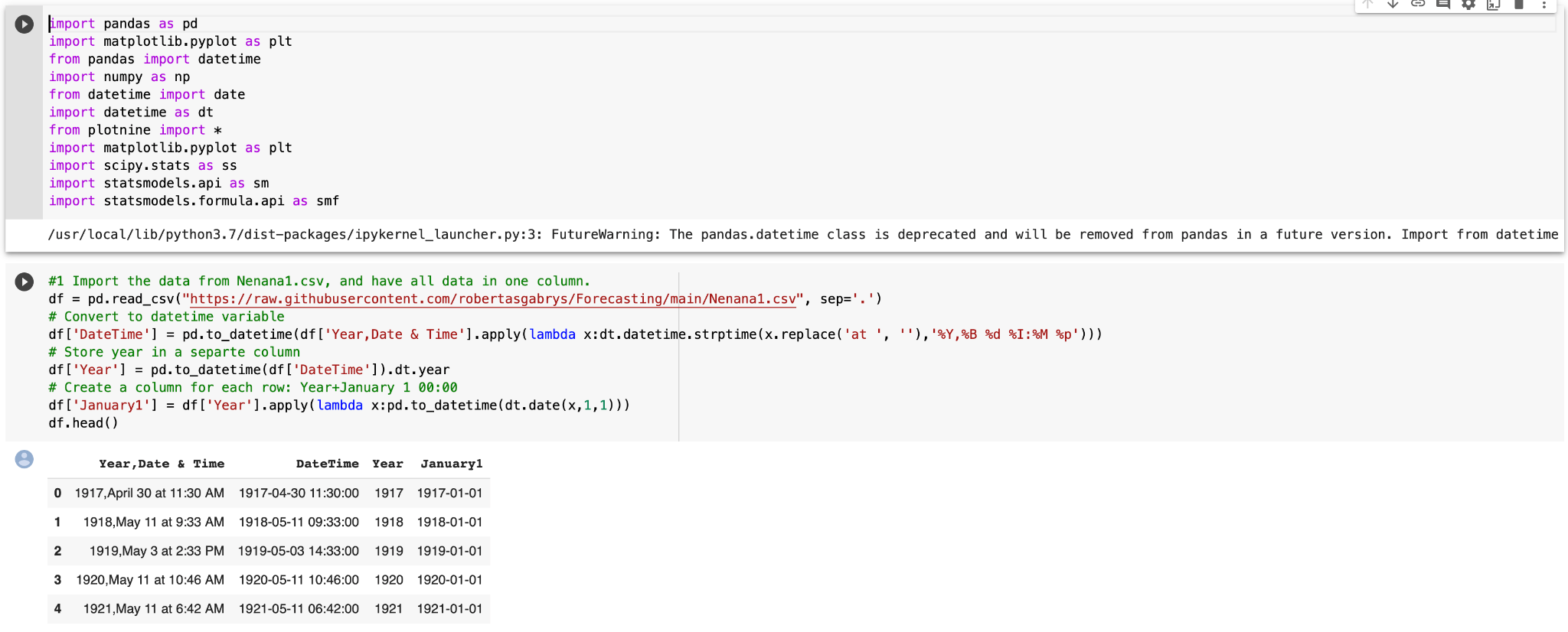
1. **In DSO 424 I started teaching about intervention and change point modeling in 2018. I was inspired by a talk at PyData LA in 2018. Please watch the video of that talk and provide 3 – 5 insights you gained (feel free to provide more!)** [**https://youtu.be/uuo8SwA1HO8**](https://youtu.be/uuo8SwA1HO8)

**Answer:**

* We often face two problems when performing traditional AB testing and using the counterfactual method: nonstationarity and autocorrelation. Both violate the assumption that the data is independent and identically distributed.
* When the model doesn’t match reality well, we can build and compare several different models developed individually and generate insights on which one should be trusted more.
* There are causal factors that analysts do not account for, which are called confounding variables; if we would like to mitigate that, we can add covariates to the model.
* The Time Series model allows us to use historical data up until the intervention to train the model to predict future trends.

**Part 2:**

**Case 1**

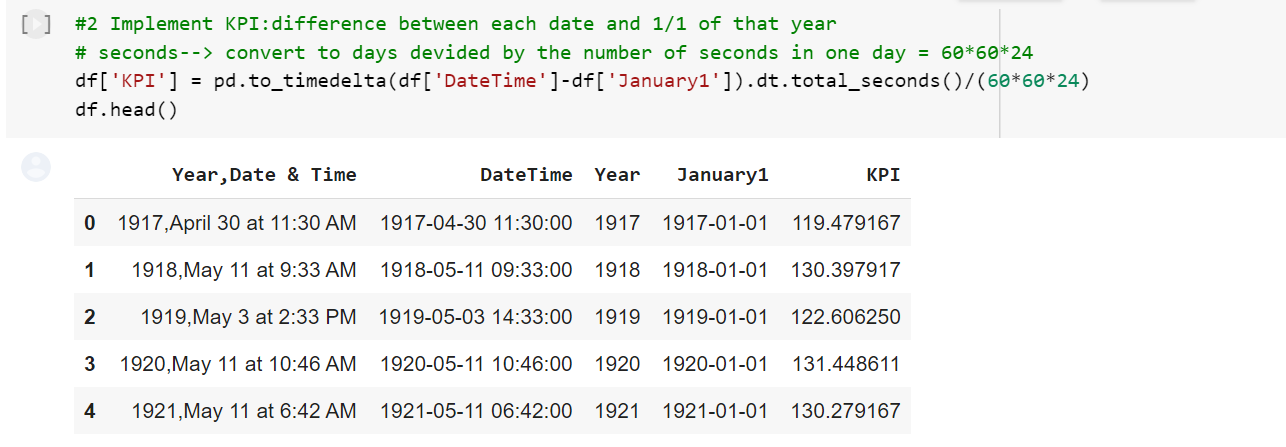
1. **Import the data from Nenana1.csv in python.**  
   You can import data from my GitHub: <https://raw.githubusercontent.com/robertasgabrys/Forecasting/main/Nenana1.csv>  
   **Answer:** 
2. To assess the global warming hypothesis let’s define a KPI as the amount of time in days between the moment ice breaks in Nenana and January 1 each year. **Implement your KPI in python.**

*Remarks:*

*The current data format will require us to perform data preparation. To give heads up: working with date/time variables is not the most pleasant task because of inconsistencies across different software, different formats in different countries, etc.*

*Working on a similar problem outside of class, you would need to come up with a meaningful KPI on your own. Understanding the questions and business objectives should help you to create a KPI. For global warming, you may use actual temperature data and not this data set. Nevertheless, this is an interesting data set and more importantly, it illustrates how one can be creative about solving a problem using data science tools.*

**Answer:**

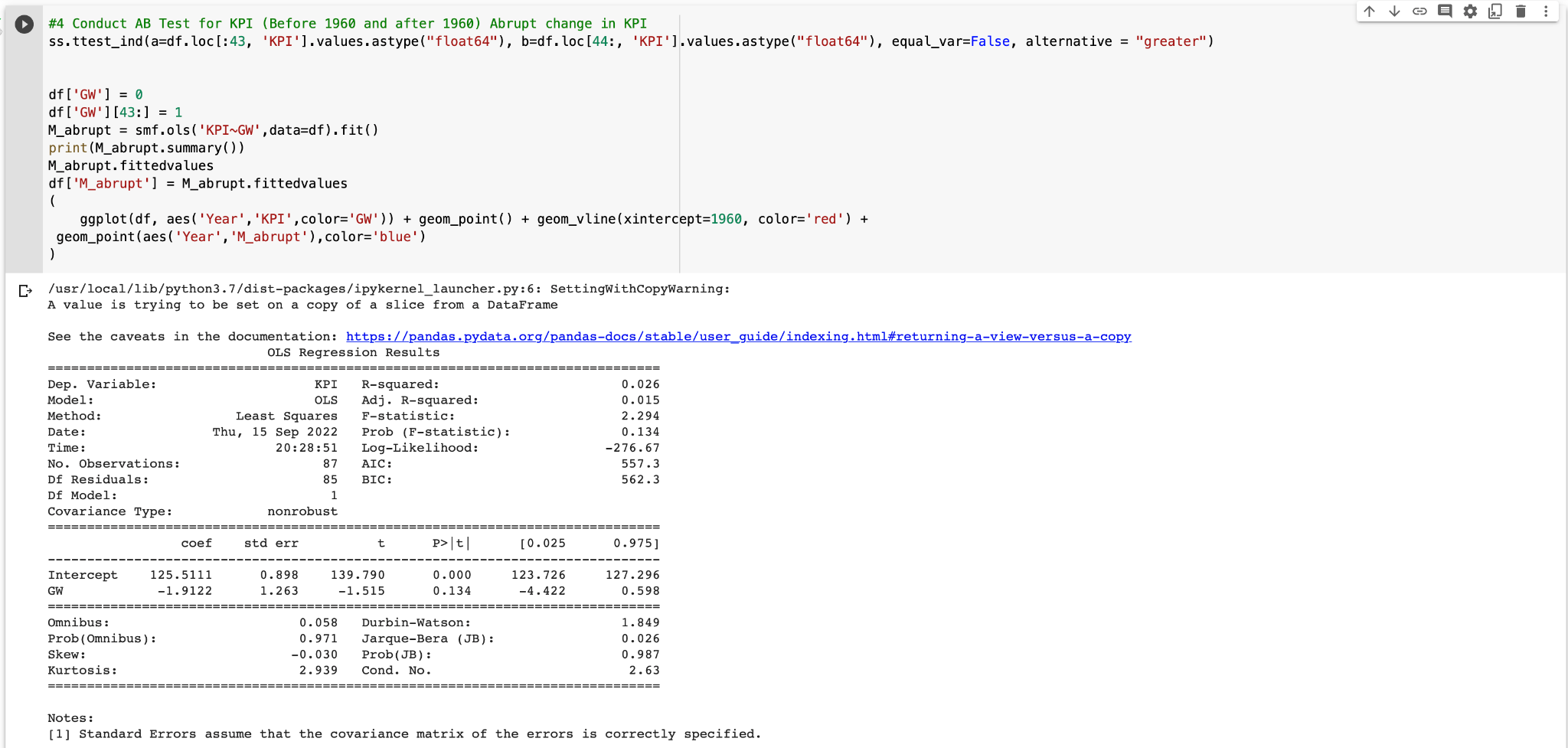


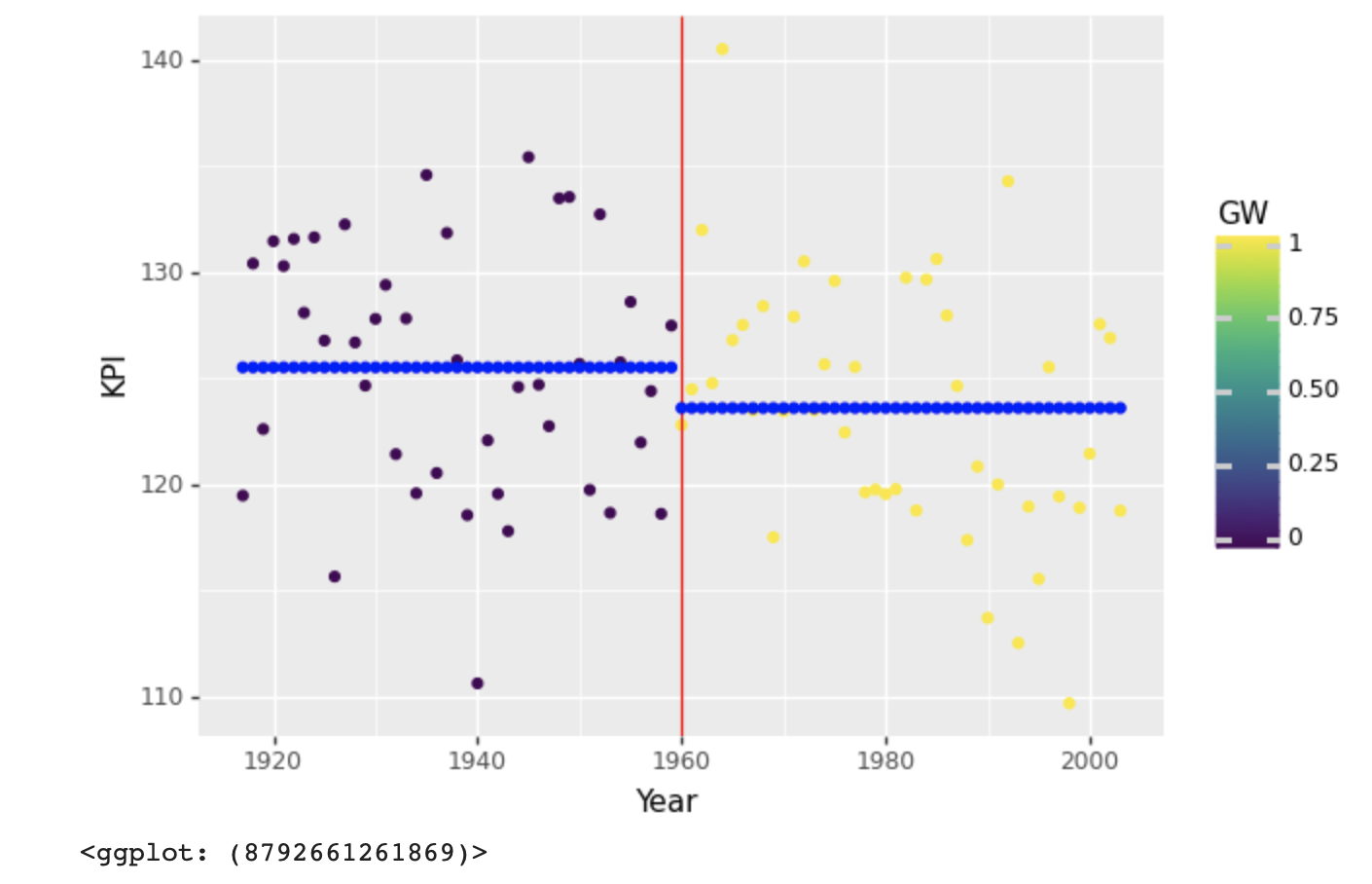
1. **Create a graph of your KPI and describe the pattern(s) you see in the data.**

**Answer:**

The mean KPI (days between Nenana ice break and January 1st) is lower for after 1960 than before. After 1960, there is also a downward trend in KPI.

1. It had been known since 1960 (judgmental call) that humans were increasing the amount of heat-trapping greenhouse gases in the atmosphere which led to the question: Does this human activity warm the climate noticeably? Judgmentally1960 has been selected to be a year of a long-term temperature change (you may ask your parents or grandparents why 1960). **Does the data provide statistically significant evidence of global warming?** Answer this question by assessing the significance of the abrupt change in KPI. **Provide the interpretation of the abrupt change.**





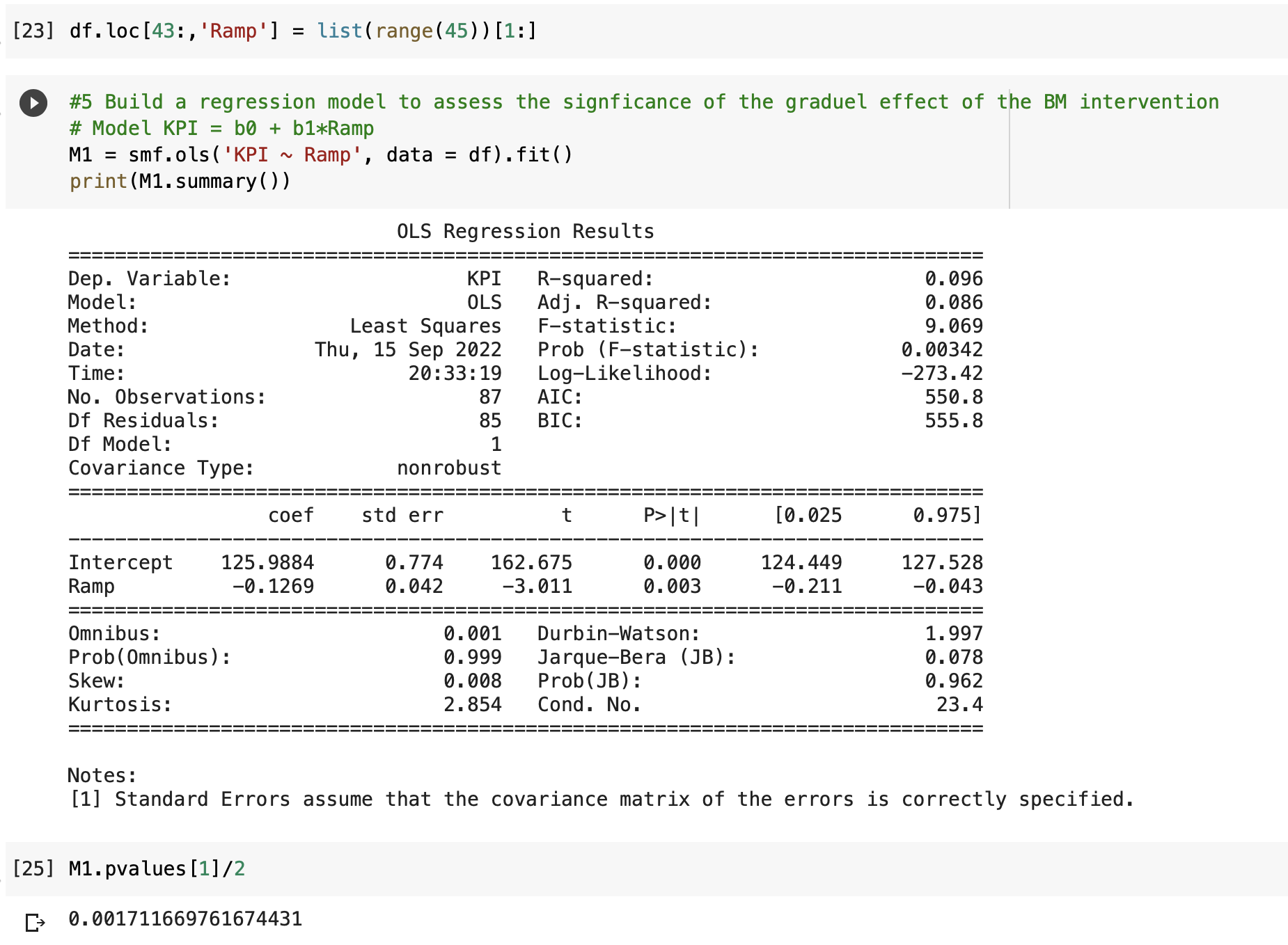
**Answer:**

Since the **p-value (0.067) > alpha (0.05)**, we conclude that the data **does not provide statistically significant evidence** of global warming.

H0: The number of days between Nenana ice break and Jan 1st before 1960 is not lower than the number of days after 1960. We performed an independent T-Test between the KPI of the samples before and after 1960. The results show that the one-sided p-value is 0.067 (0.134/2), which is greater than 0.05. We fail to reject the null hypothesis. We do not have statistically significant evidence to conclude that human activity leads to global warming noticeably.

The mean KPI after 1960 appears to be lower than that before 1960 on the graph, meaning that ice broke earlier after 1960 than before 1960, which signals global warming. However, this observation is not statistically significant and therefore could have happened by chance.

1. Does the data provide statistically significant evidence of global warming? Answer this question by assessing the significance of the gradual change in KPI. Provide **the interpretation of the gradual change**



**Answer:**

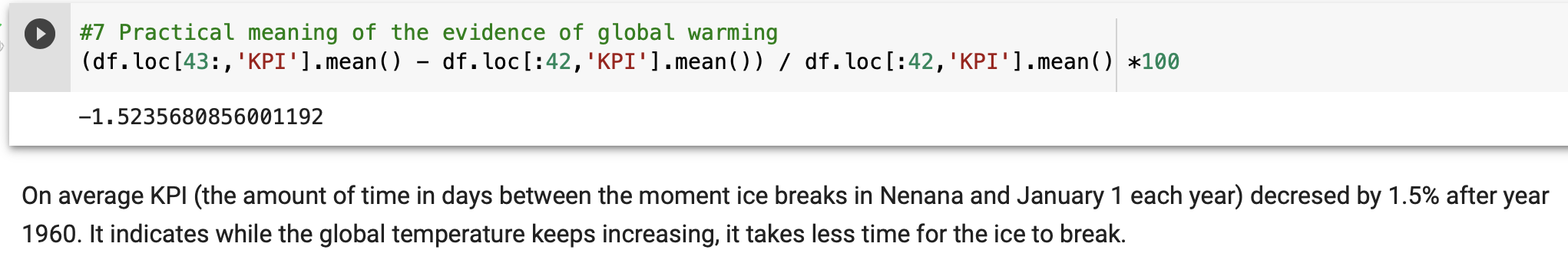
The one-sided p-value is 0.0017 which is less than 0.05. Therefore, the data provides statistically significant evidence that KPI was gradually decreasing, meaning that human activities warmed the climate.

1. **Which one – abrupt or gradual effect of the intervention in 1960 is more appropriate?** Briefly justify your answer.   
   **Answer:**

The gradual effect is more appropriate given that

1. the overall fitness of the trend: abrupt change of two horizontal lines does not fit the pattern and the trend of the data well
2. the p-value of gradual effect is 0.0017, which is less than 0.05; however, the p-value of abrupt effect is 0.067, which is larger than 0.05
3. In 1960, Charles Keeling accurately measured atmospheric concentrations and showed that they were definitely increasing, ushered in a new age of expanding climate research, and the atmospheric concentration kept accumulating since the oceans would not absorb all industrial CO2 emissions. Therefore, the effect of excessive CO2 emissions and accumulation should be considered **incremental**, and the gradual effect of the intervention of 1960 is more appropriate.
4. You may recall from your intro to business statistics course, BUAD 310, that statistical significance ≠ practical importance. Suggest and implement a way(s) to **quantify the evidence of global warming from a practical perspective.** If you use terms such as “result is statistically significant or in/nonsignificant”, many people won’t have a clue what they mean and will ask you to put it in Layman’s terms.

**Answer:**



1. Compare **your approach with the approach we used to assess the effectiveness of the batmobile program. Have you observed any differences?** Briefly comment.

**Answer:**

The metric we come up with in this case, which is the difference between the date of ice break and the first day of the year, is very straightforward and thus sufficient for our analysis and modeling. However, the first metric we thought of in the BM case, which was ‘ACC’, was against our expectation that BM reduced the accident rate, and thus might be flawed. Therefore, we had to take an extra step to come up with ‘KPI’ excluding the effect of the increased number of drivers after the BM policy, which could affect the accident rate. Our takeaway is that we should always go back and contemplate on the metric and approach we take after we get the results. We should constantly assess the effectiveness and reasonableness of the metric and see if we need any modifications.

1. Read the original article published by two Stanford professors:   
   <https://science.sciencemag.org/content/sci/294/5543/811.full.pdf>

And also a Critique of the article and the comments below the critique:   
 <https://www.john-daly.com/nenana.htm>

**Answer:**

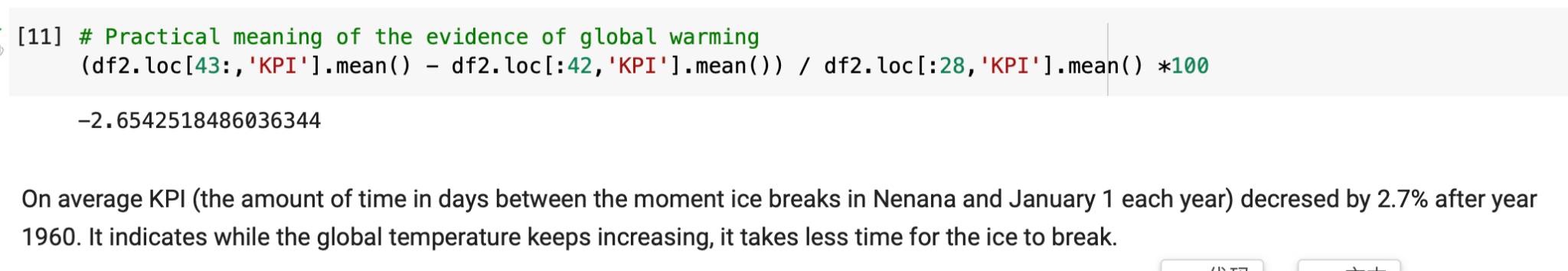
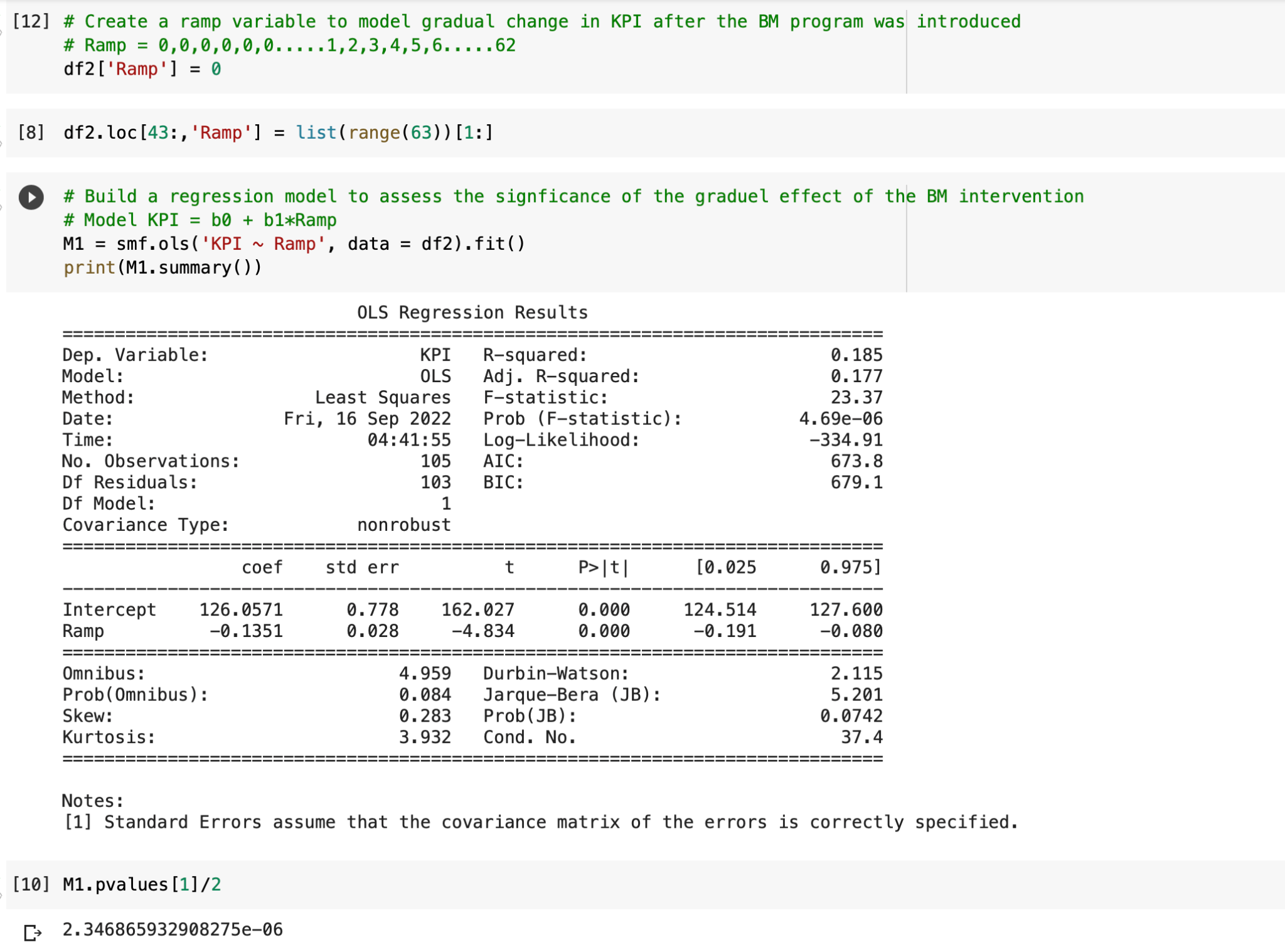
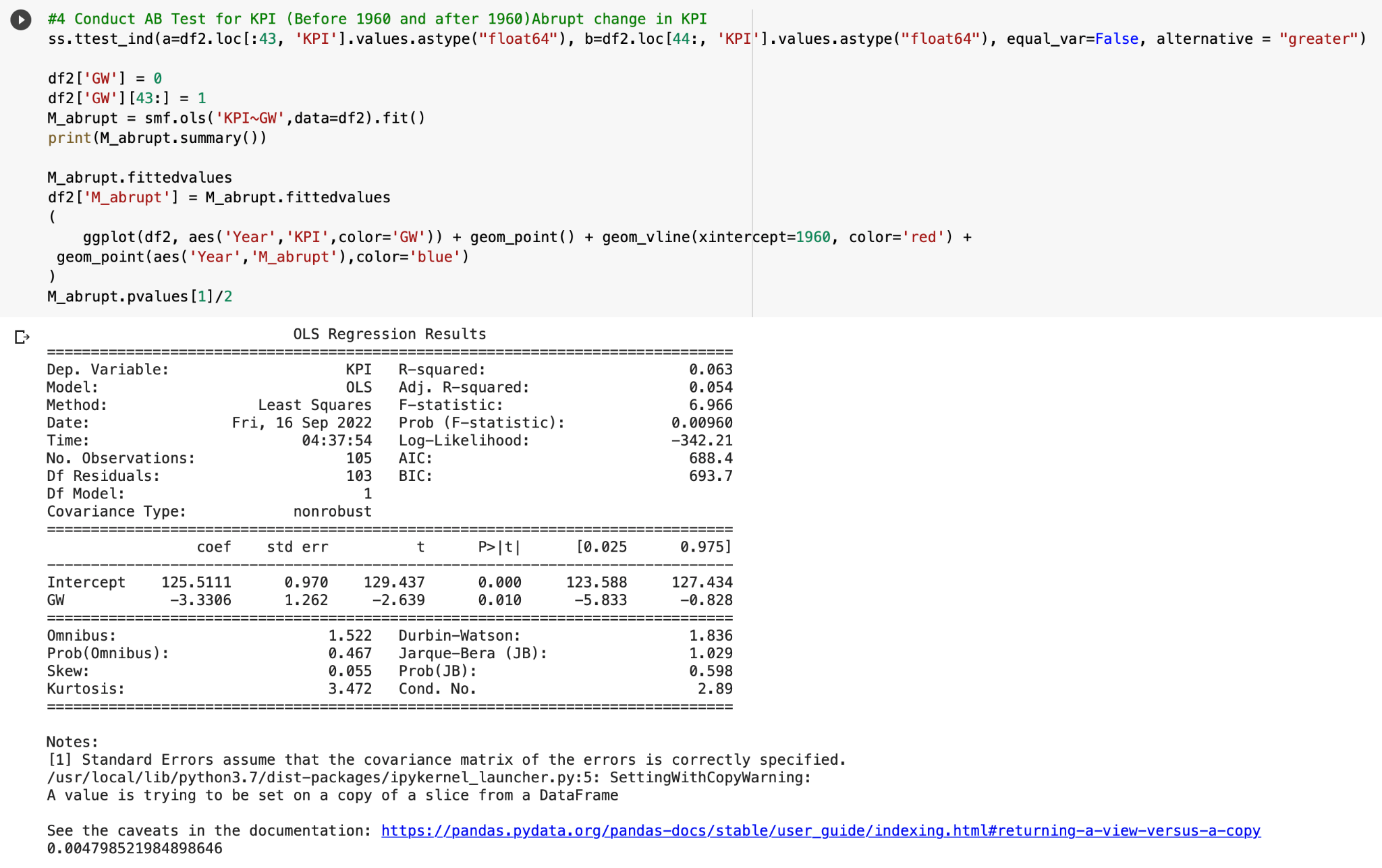
According to the authors of “Climate change in nontraditional data sets”, they compared the ice breakup record from 1949-2000 with available climatic data for Nenana and Fairbanks. And they made three main statements which were countered in the other article. First, they speculated that heat-island effects may affect the Fairbanks record because of urbanization. What’s more, they insisted that trends for precipitation and snowfall were non-significant for both cities. In addition, they quoted observation conclusions of Nenana residents as saying that hermal breakups have been more common than dynamic breakups in recent years. Then they made a regression and concluded that there was a significant negative trend reflecting an advancement of breakup relative to the vernal equinox by 5.5 days.

However, John L. Daly questioned a lot about the article. First, he proposed that the Nenana temperature record was too fragmented to mean anything. He also came up with the idea that the choice of the start point (1949) was inappropriate, which is known as “end date distortion”, for they chose a cold year as the start of the period to prove climate warming. In addition, he also proved the existence of heat-island effects in Fairbanks by comparing climate data from inside and outside urban areas. What’s more, he found a significant increase in total snowfall during the 20th century, causing ice to break earlier than would be the case with thermal melt alone. Last but not least, he thought it less rigorous to quote the saying of some residents only. In conclusion, he doubted the conclusions of the article above.

When we are performing statistical analysis, we should be cautious about “end date distortion”, where a long-term trend can be profoundly affected by the choice of start and end dates, particularly where those terminal dates are outside the mean. We should be aware of the omission of intervention points which, in this case, are the 1940 warm period and the 2001 cold period, since the omission could distort the data in a way that satisfies the researchers’ biases but fails to depict reality. For example, in this case, the original research reflected an upward trend in global warming. However, were the two intervention points taken, the trend of climatic warming would have been slighter.

1. Last year NASA National Snow and Ice Data Center updated nenan1.csv by including more recent data. The updated data set action is in **Nenana2.txt**. **Carry out abrupt and gradual change point analyses using Nenena2.csv.** Briefly summarize your findings. Does Nenana2.csv lead to different conclusions?

You can import data from my GitHub:  
<https://raw.githubusercontent.com/robertasgabrys/Forecasting/main/Nenana2.csv>  
**Answer:**

For Nenana.txt, there is only statistically significant evidence of global warming with a gradual change in KPI.Whereas for Nenana2.txt, the data provide statistically significant evidence of global warming with both abrupt and gradual changes in KPI. ****