A Statistical Analysis of Global Temperature Trends

Project - Team 10-7: The Outliers

By Morgan Pruchniewski, Scott Burstein, Katie Zhou

11/22/2020

Introduction

One of the biggest environmental issues facing our planet has been climate change. As fossil fuel emissions have increased, greenhouse gases have begun trapping heat in our atmosphere that has created severe impacts. These include increased global temperature and rising ocean levels, in addition to increased natural disasters. NASA.gov Source. USGS.gov Source

Climate change is a significant issue that has also become a hot-button political topic in recent years because of the failure of certain individuals, political parties, and corporations to take action: some people do in fact deny the existence of climate change/global warming despite a strong body of scientific support. Basis of climate change denial Source The main argument behind this is that the Earth is always changing, and that the recent increase in global land and ocean temperatures is not significant enough to support global warming.

In this report, we will be analyzing a dataset containing recorded land and ocean temperatures over the last 100+ years to examine whether there is evidence to suggest a statistically significant increase in land temperature over the last 30 years.

Research Questions

Is there significant evidence to support the existence of climate change?

To evaluate this, we will conduct comprehensive hypothesis tests on four sub-questions.

1. Is there evidence to suggest a statistically significant increase in mean earth surface temperature from early-20th-century levels to what the data show for more recent years?.

We predict that there will be sufficient evidence to support a statistically significant increase as detailed above, which brings us to the next question:

2. Is the earth changing/increasing temperature at a faster rate now than it was in the early 20th century?

For this question, we will record the observed average rates of change (slope) for every five-year period, and test whether the mean slope after 1980 is greater than the mean slope before 1980 by a statistically significant margin. Because there are only 23 five-year segments between 1900 and 2010, we will use a higher α level of 0.05.

Relating to the previous question, which looks at the rates of presumed increase, we will also evaluate net change (positive of negative).

3. Does the data provide evidence of a greater degree of net annual temperature fluctuation after 1980 than before 1980?

Finally, we will evaluate how the changes in earth surface temperature may differ geographically. With Europe and North America having had large roles in industrialization, globalization, and technological developments, all of which have been cited as major contributors to climate change, we wanted to consider if this might be manifested in the amount of climate change exhibited in the last century in those continents.

4. Has North America experienced a greater change in annual average temperatures from the first half of the 20th century to 2010 than other continents? Has Europe?

Data Sources:

We found this dataset on Kaggle climate-change-earth-surface-temperature-data.

The data was originally compiled by the Berkeley Earth Data Lab from 16 pre-existing archives, and it was updated to Kaggle in 2017. It is important to note that many of the earlier years include NA values, which is why we focused on analysis on the 1900s.

Types of Information Present in Data:

By Time and Location (for `GlobalLandTemperaturesByMajorCity`)

Each observation in `GlobalLandTemperaturesByMajorCity` is a city and its respective land temperature, coordinates, and Country, which will be used to investigate changes in climate over time. There are 239,177 rows in the major city dataset.

By Time (for `GlobalTemperatures`)

Each observation in `GlobalTemperatures` is a numeric date/time value (described below) and its respective land temperature, max temperature, min temperature, ocean temperature, and relevant uncertainties to said variables, which will be used to investigate changes in climate over time. There are 3,192 rows in the global dataset.

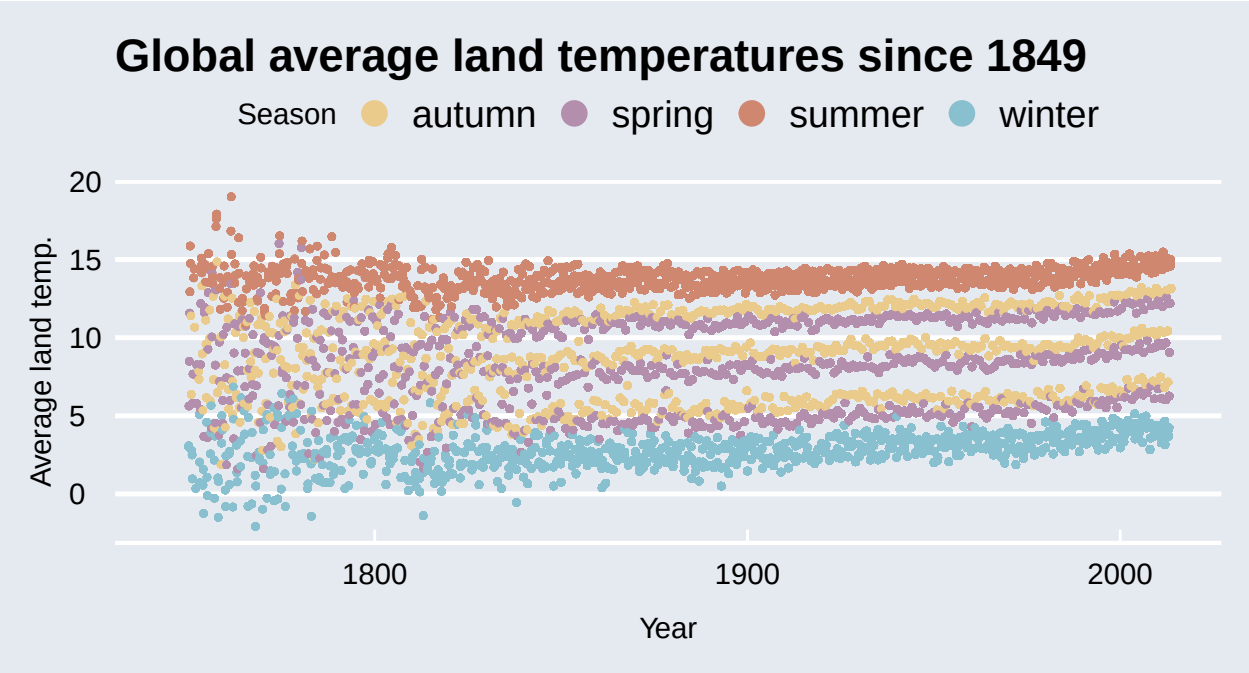
Relevant Variables:

The relevant variables in the `major_city` dataset include dt (date and time of the recorded temperature, a discrete numeric), AverageTemperature (the average temperature for each city, a continuous numeric), city (nominal categorical), country (nominal categorical), latitude (continuous numeric), and longitude (continuous numeric).

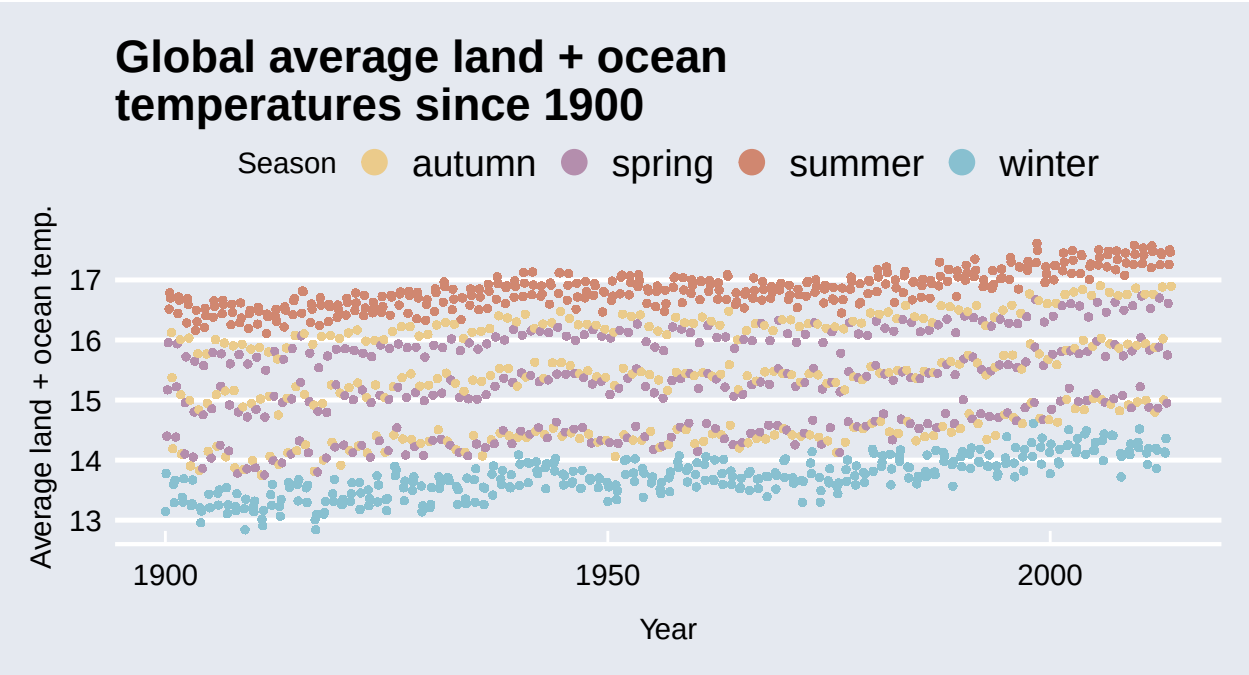
The relevant variables in the `global` dataset include dt (date and time of the recorded temperature, a discrete numeric), LandAverageTemperature (the average global land temperature at that time, continuous numeric), LandMaxTemperature (the highest recorded land temperature of that year, a continuous numeric), LandMinTemperature (the lowest recorded land temperature of that year, a continous numeric), and LandAndOceanAverageTemperature (the average of land and ocean temperature averages for that year, a continuous numeric).

Methodology

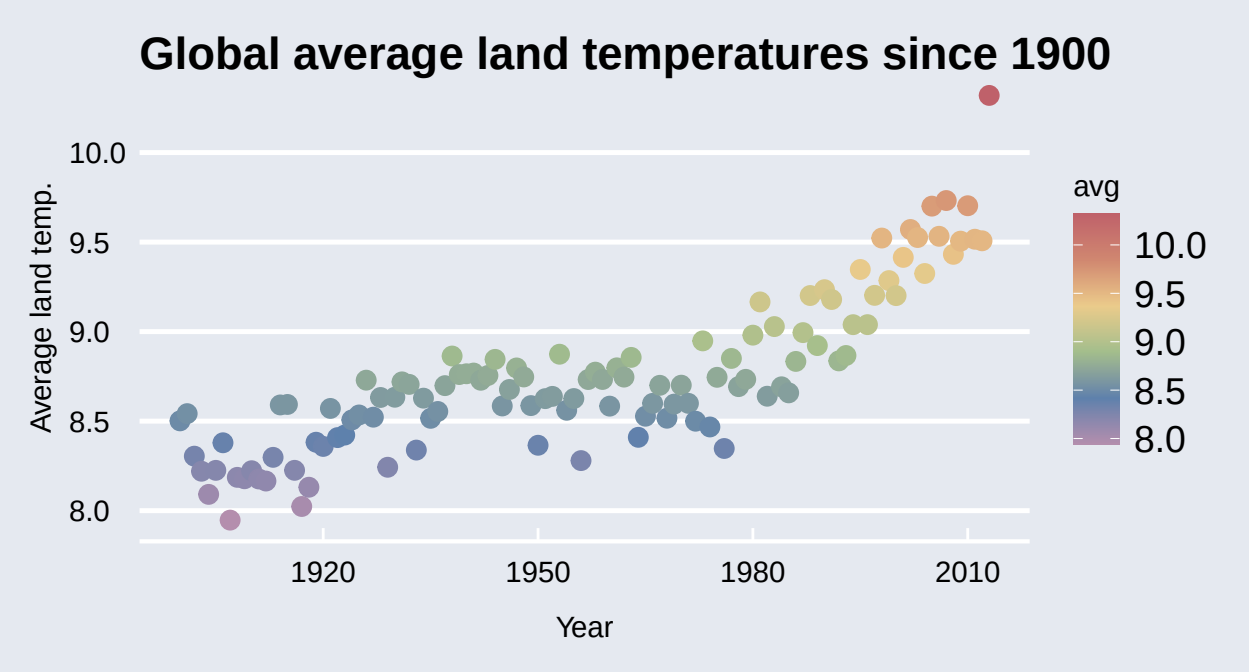
Initial Visualizations:



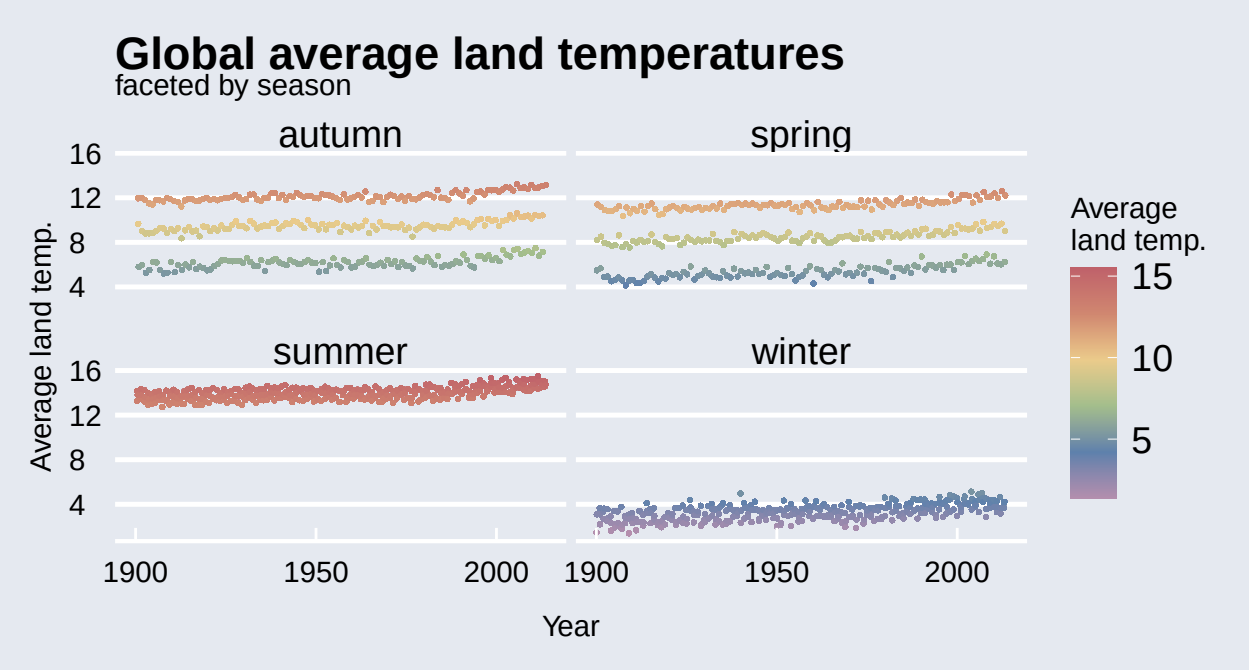
Based on this initial visualization, there seems to be a high amount of uncertainty in the average land temperature in the earlier data. Going forward, we will only use the data after 1900. Overall, across all seasons, there seems to be an increasing trend in average land temperature after roughly 1975.

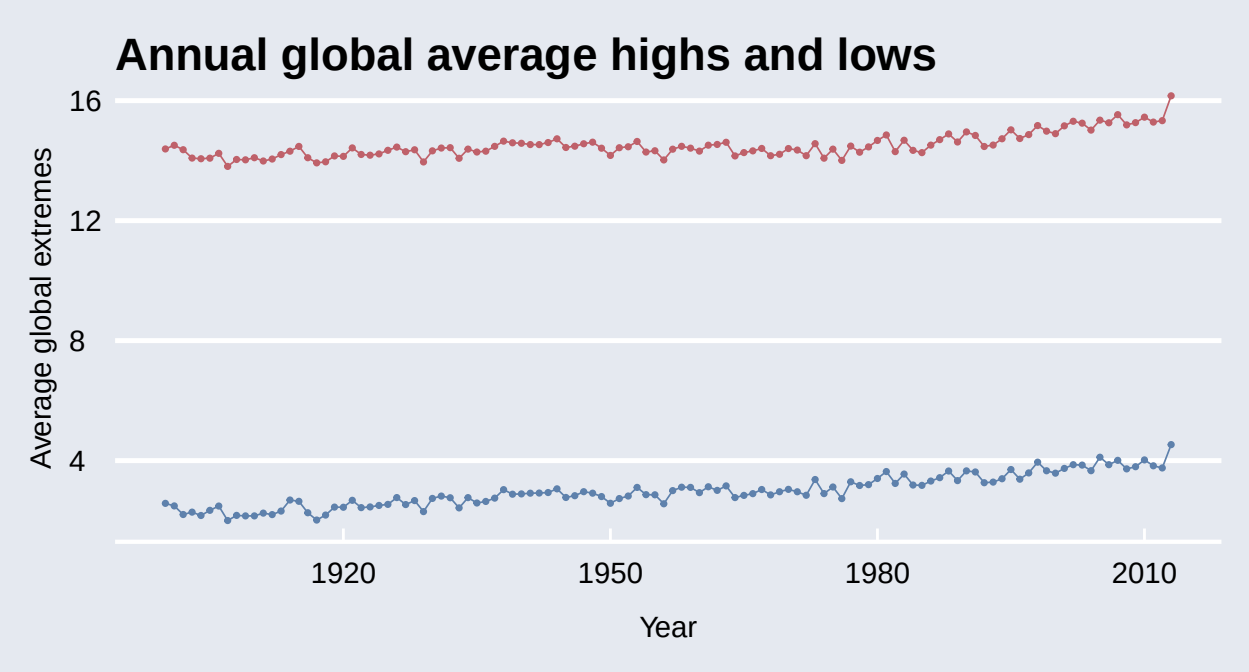


Based on the second visualization, the average land/ocean temperature seems to increase more steeply after 1975 than the average land temperature. The values for spring and autumn in the land/ocean data are also much more similar than they are for just land.

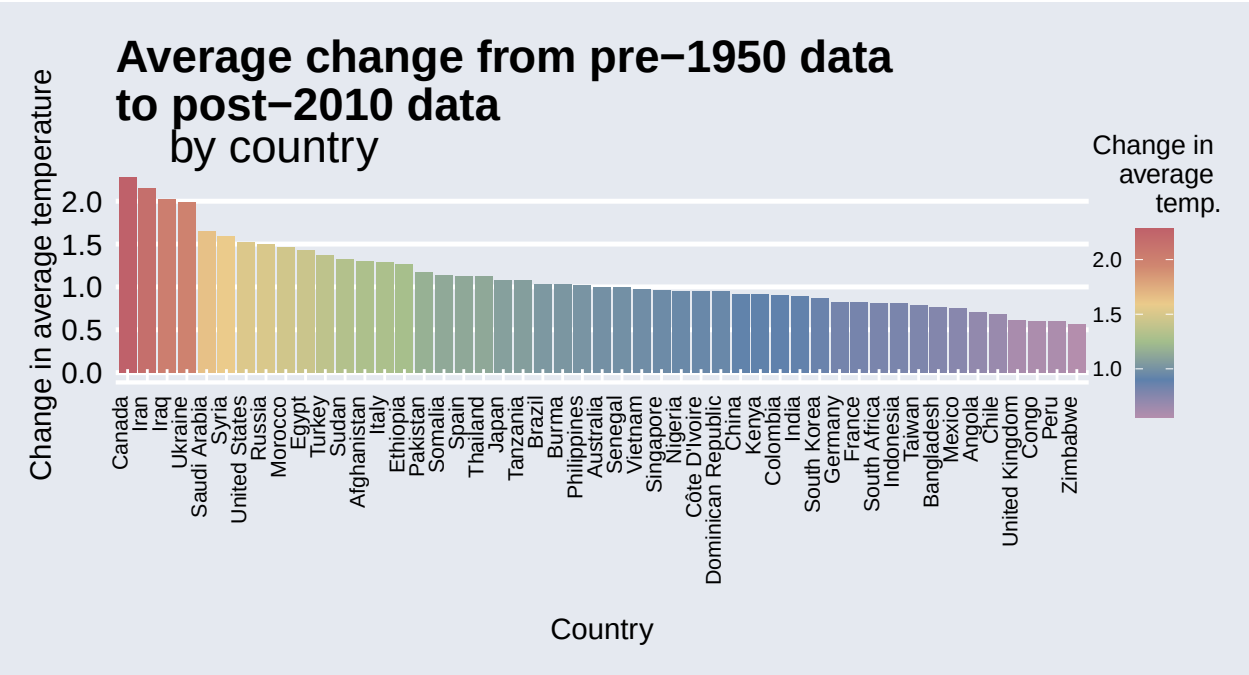


When we plot the average annual land temperature (an average of the data points from each month), we can visualize a more extreme change in average land temperature over time than when each month is plotted individually. This is marked by a more severe uptick since 1980.





The upward trend visualized in the first few graphs is consistent over each season, as well as for the annual average temperature highs and lows.



In this visualization we can see the countries that have had the greatest change in average temperature from before 1950 to 2010-2013.

Results

Question 1

Is there evidence to suggest a statistically significant increase in mean earth surface temperature from early-20th-century levels to what the data show for more recent years?

Null Hypothesis:

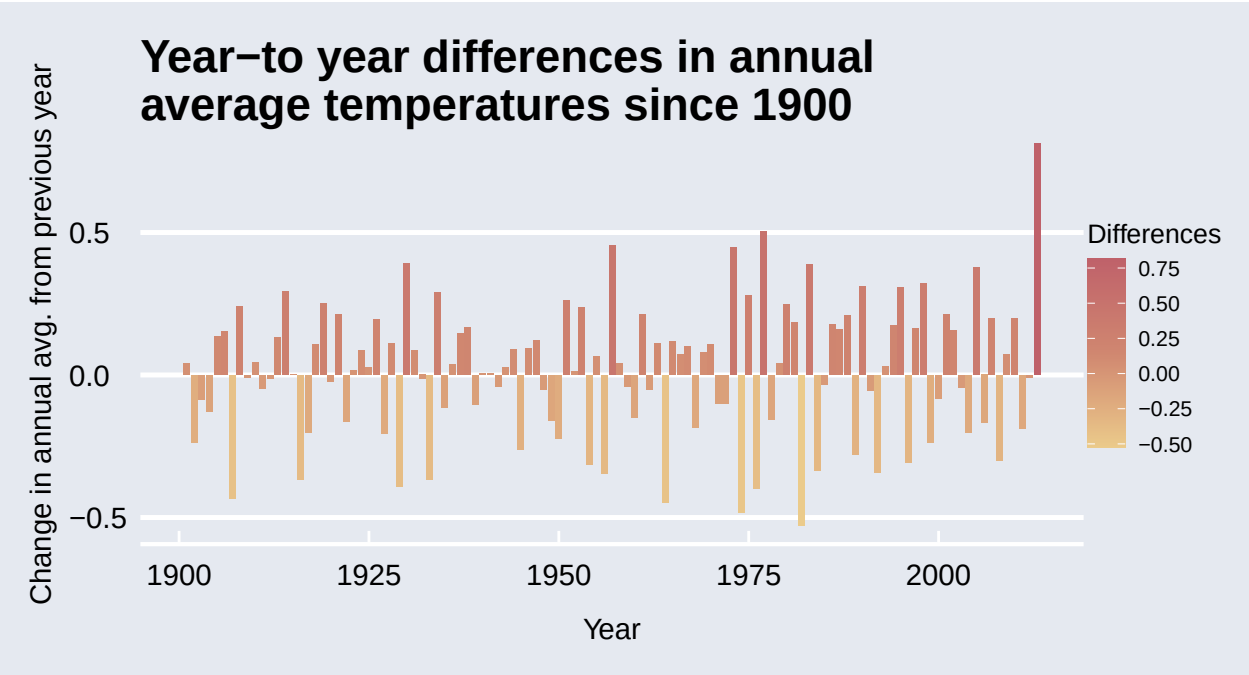
$H_o : \mu_{post1980} \leq \mu_{pre1980}$

The mean anomaly of annual temperature averages with respect to the average temperature before 1980 is not greater than 0 (the mean value from before 1980 is used as the reference value for calculating anomalies, so the mean anomaly of pre-1980 values is assumed to be 0).

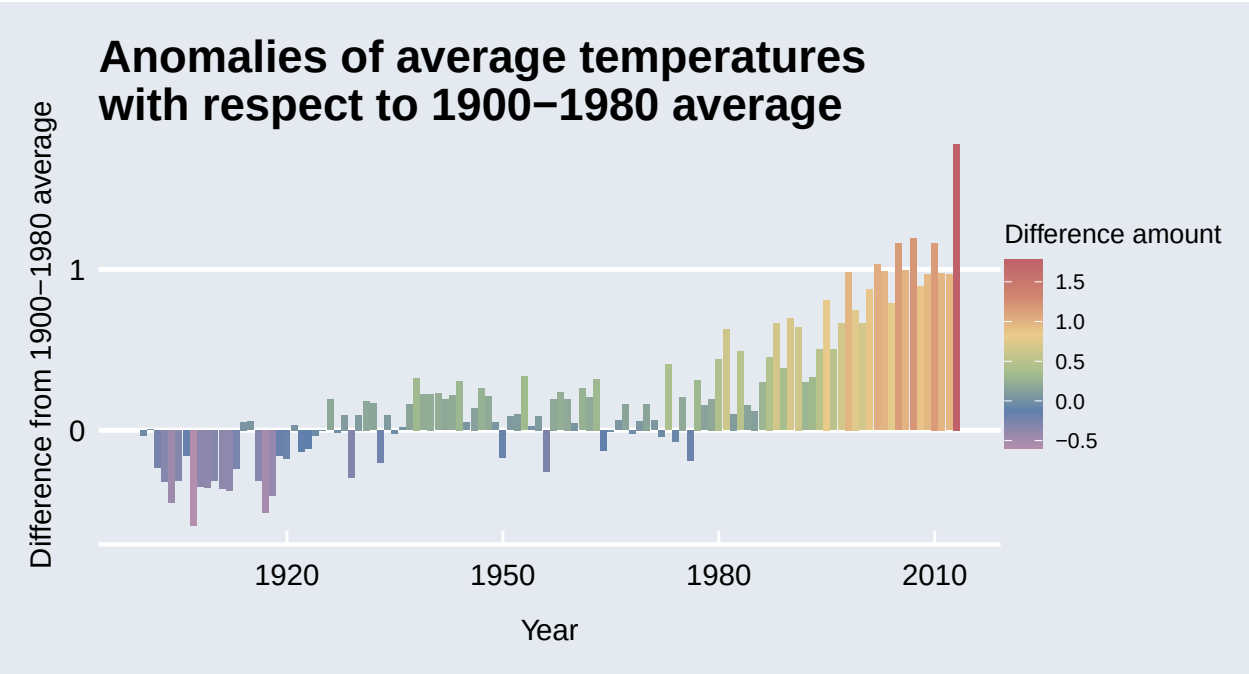
Alternative Hypothesis:

$H_a : \mu_{post1980} > \mu_{pre1980}$

The mean anomaly of annual temperature averages with respect to the average temperature before 1980 is greater than 0 (the mean anomaly of data before 1980 with respect to its average temperature).



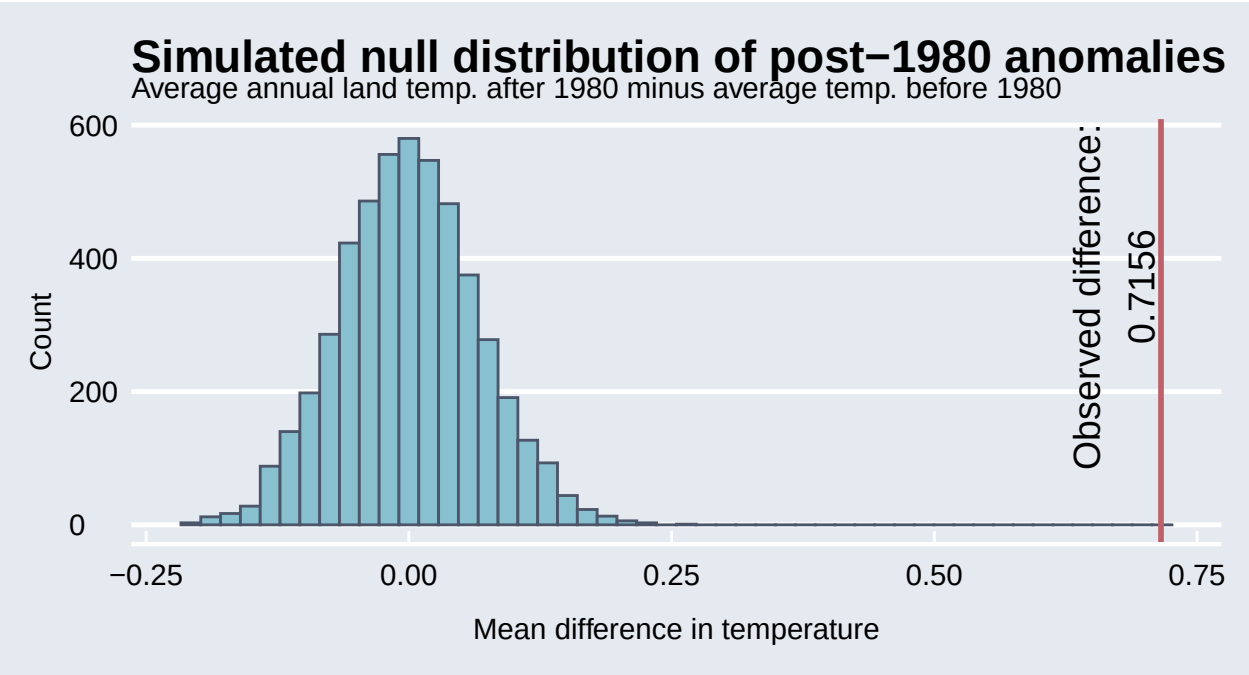
Based on this visualization, someone could potentially try to counter the argument for the existence of climate change, because there is a large fluctuation of positive and negative changes from year to year. Therefore, we then created a new data-frame to measure the overall difference between each year's average temperature and the average temperature across the 20th century, as shown below, and are using the comparison to the 20th century average for our hypothesis test.




```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 0.551 0.894
```

Our 99% confidence interval for this hypothesis test is (0.5514, 0.8941). We can be 99% confident that the true mean anomaly of annual temperature averages after 1980 with respect to the average temperature before 1980 is between these bounds.

```
## [1] 0
```




```
##
##  Welch Two Sample t-test
##
## data:  anomaly by post1980
## t = -10.673, df = 44.389, p-value = 3.866e-14
## alternative hypothesis: true difference in means is less than 0
## 99 percent confidence interval:
##        -Inf -0.5538219
## sample estimates:
## mean in group FALSE  mean in group TRUE
##        6.106227e-16        7.156482e-01
```

In both approaches used in this analysis, the p-value is well under $\alpha = 0.01$, so we reject this null hypothesis.

The p-value in this approach is the probability of obtaining results in which the post-1980 data exhibit an anomaly with respect to the pre-1980 average temperature as large as observed or greater, assuming under the null hypothesis that the post-1980 temperatures are not significantly different from the pre-1980 temperatures. From both tests, the p-value was extremely small: we observed a 0 p-value in the simulation-based approach.

In context of our question, this indicates that there is evidence to support a statistically significant increase in temperature over the 20th century. This is important because this is one of the main arguments used against climate change.

Question 2

Is the earth changing/increasing temperature at a faster rate now than it was in the early 20th century?

Null Hypothesis:

$H_o : \beta_{post1980} \leq \beta_{pre1980}$

The rate at which the average global land temperature increased after 1980 is not greater than the rate of temperature increase starting from 1900 based on their respective linear models.

Alternative Hypothesis:

$H_a : \beta_{post1980} > \beta_{pre1980}$

The rate at which the average global land temperature increased after 1980 is greater than the rate of temperature increase starting from 1900 based on their respective linear models.

```
## # A tibble: 2 x 2
##   term        estimate
##   <chr>          <dbl>
```

```
## 1 (Intercept) -12.2
## 2 year          0.0107
```

```
## # A tibble: 4 x 2
##   term               estimate
##   <chr>                 <dbl>
## 1 (Intercept)        -2.84
## 2 year                0.00586
## 3 post1980TRUE      -48.2
## 4 year:post1980TRUE   0.0243
```

Overall Regression:

$$\widehat{TempDiff} = -12.204 + 0.0107 \times year$$

Interaction Model for Before and After 1980:

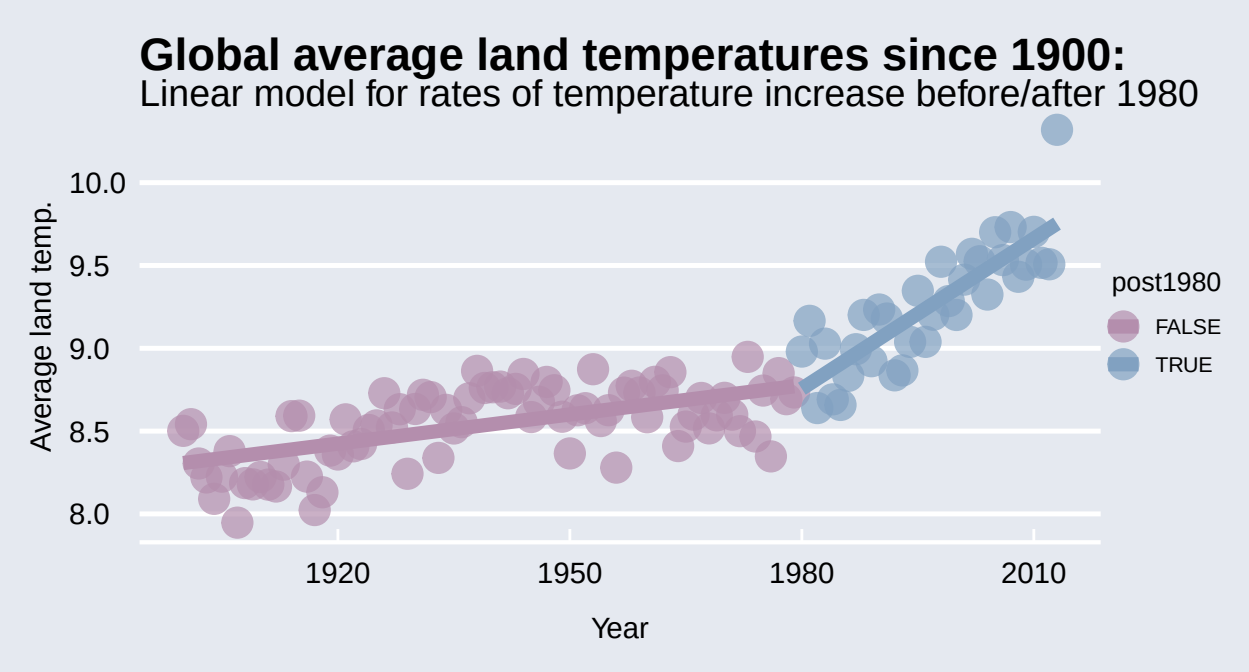
$$\widehat{TempDiff} = -2.836 + 0.00587year - 48.226post1980 + 0.0243year \times post1980$$

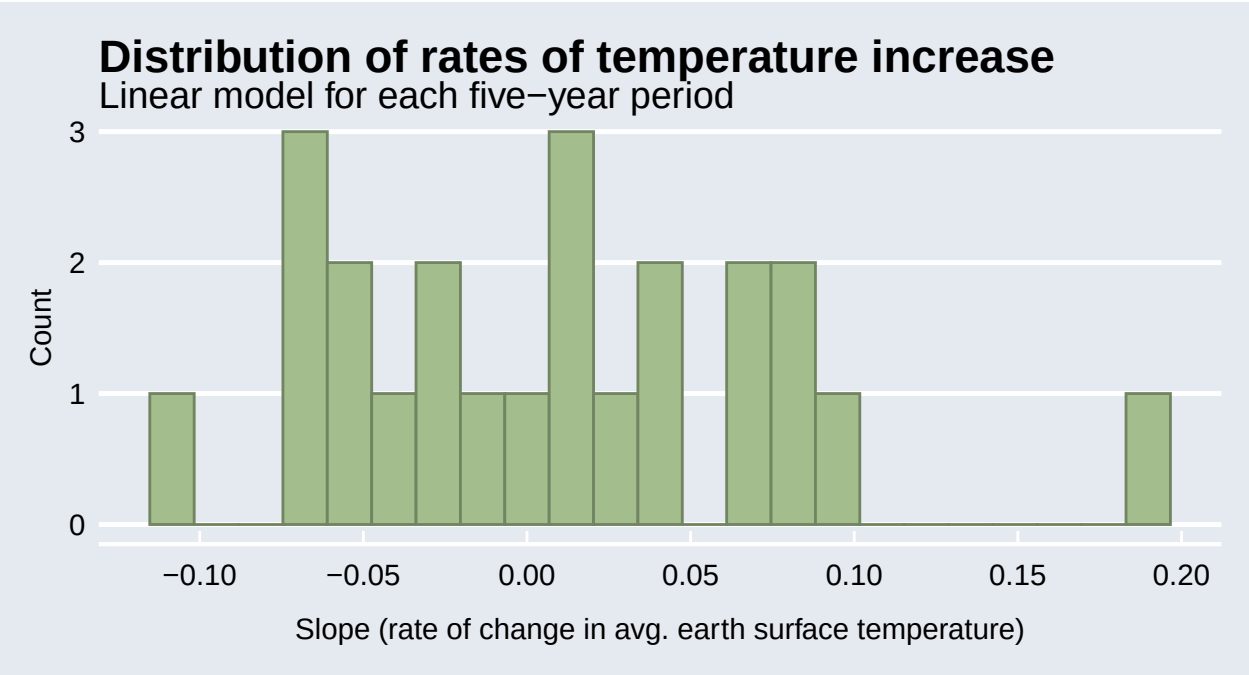
Pre-1980:

$$\widehat{TempDiff} = -2.836 + 0.00587year$$

Post-1980:

$$\widehat{TempDiff} = -51.0624 + 0.0302year$$





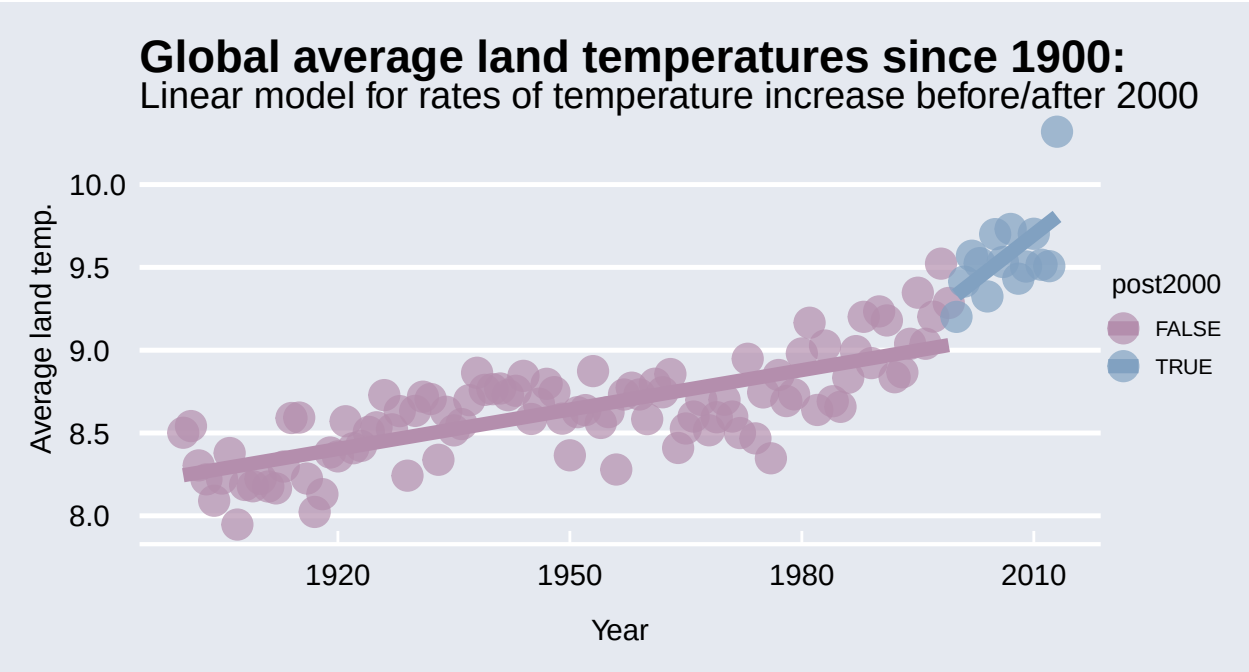

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0538
```

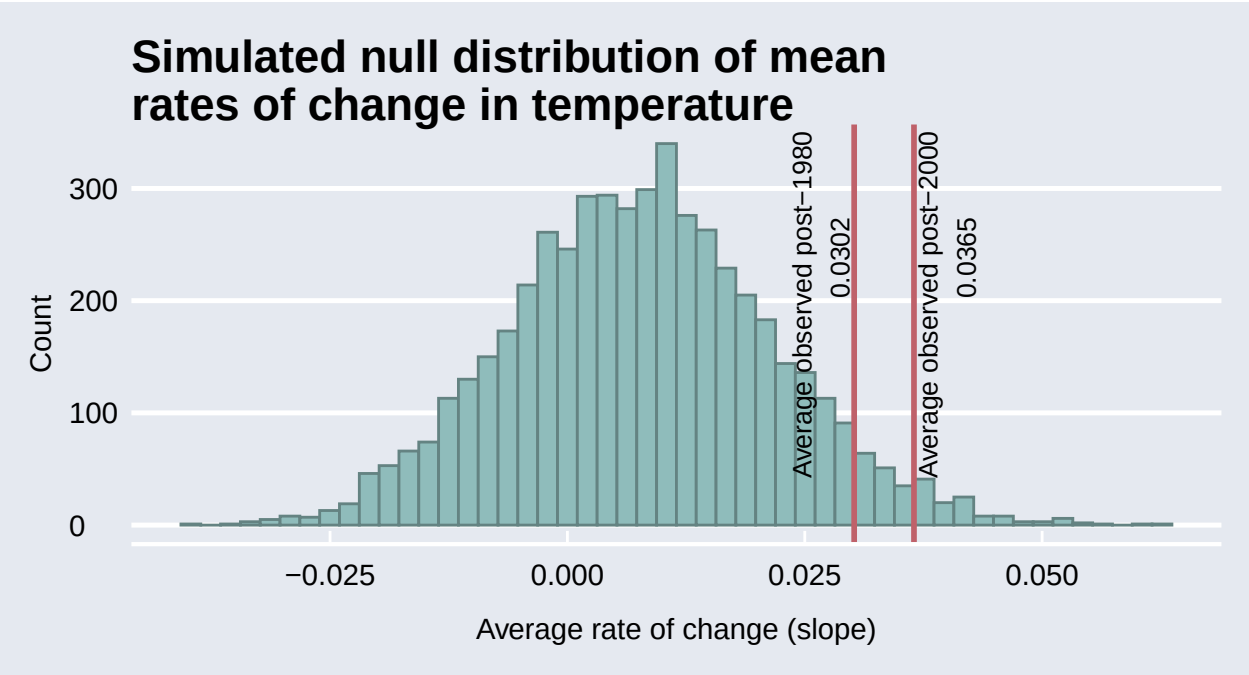
From the above simulation-based test, we found a p-value of 0.0592. Because this is not under $\alpha = 0.05$, we fail to reject this null hypothesis.

Here, the p-value is the probability of obtaining results in which the difference in average rate of change between post- and pre-1980 temperatures is at least as great as what was observed in the data.

In context of our question, this suggests we may have insufficient evidence of a statistically significant increase in the average degree of temperature change from before 1980 to after 1980.

After failing to reject the null hypothesis above, we then considered performing the same test for the data before and after the year 2000. The linear model as shown below suggests a steeper incline than 1980.





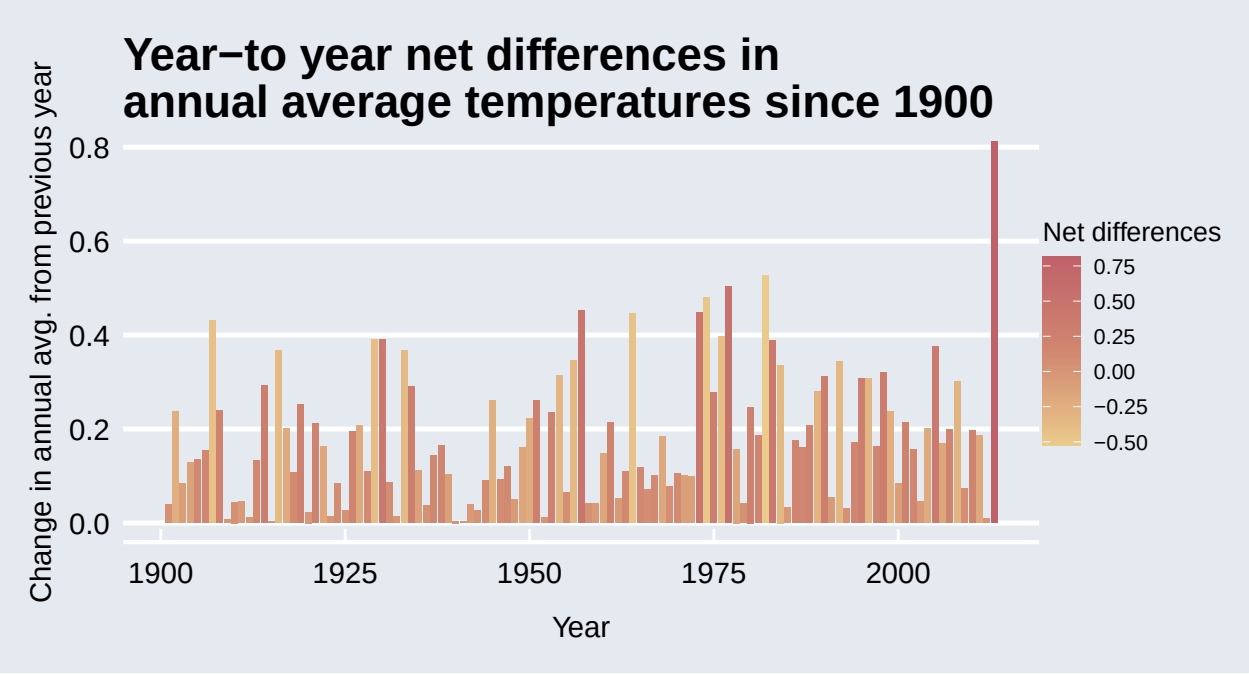

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0238
```

In the hypothesis test for years after 2000, the p-value is 0.0214, so we can reject this null hypothesis. This p-value indicates a 0.0214 probability of obtaining results in which the post-2000 data points have a mean rate of change equal to or greater than the one observed, assuming that the rates of change truly are equal.

In context of our original research question, this suggests that the earth surface temperature changes that we examined in part 1 might be occurring at a quickening pace over time, with a decreasing probability of obtaining results as or more extreme as observed (the p-values for the 1980s and 2000s tests). However, it should be noted that there is very limited data after 2000, so this p-value should not necessarily be interpreted as a practically significant value, and this test alone does not adequately answer our overall research question.

Question 3

Does the data provide evidence of a greater degree of net fluctuation in annual global mean temperatures (positive or negative variability) for years 1980-2013 than years 1900-1980?



Null Hypothesis:

$H_o : \mu_{pre1980net} \leq \mu_{post1980net}$

The mean net (absolute value) year-to-year fluctuation in global average surface temperatures after 1980 is not greater than the mean year-to-year fluctuation before 1980.

Alternative Hypothesis:

$H_a : \mu_{pre1980net} > \mu_{post1980net}$

The mean net (absolute value) year-to-year fluctuation in global average surface temperatures after 1980 is greater than the mean year-to-year fluctuation before 1980.

```
## # A tibble: 1 x 2
##      lower upper
##      <dbl> <dbl>
## 1 -0.00639 0.148
```

Our 99% confidence interval for this hypothesis test is (-0.00639, 0.14843). We can be 99% confident that the difference in means of average year-to-year net fluctuation after 1980 and before 1980 lies within these bounds.

```
## [1] 0.0196
```

```
##
##  Welch Two Sample t-test
##
## data:  net_diff by post1980
## t = -2.1105, df = 55.189, p-value = 0.01968
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.0134531
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           0.1649367           0.2298260
```

The p-values from both the simulation-based and CLT-based hypothesis tests performed here are both less than $\alpha = 0.05$, so we can reject the null hypothesis. In this case, the p-value represents the likelihood of obtaining results where the difference in mean net fluctuation of average temperatures from year to year after 1980 versus the mean net fluctuation before 1980 is as great as observed or greater. In context of our question, this suggests we have sufficient evidence to argue that not only have temperatures increased over the 20th century, but they have become more extreme overall from year to year for both cold and hot temperatures compared to earlier in the 20th century.

Question 4

Has North America experienced a greater change in annual average temperatures from the first half of the 20th century to 2010 than other continents?

Null Hypothesis:

$H_o : \mu_n \leq \mu_g$

The change in mean temperature from before 1950 to post-2010 in North America is not greater than the global change.

Alternative Hypothesis:

$H_a : \mu_n > \mu_g$

The change in mean temperature from before 1950 to post-2010 in North America is greater than the global change.

```
##
##  Welch Two Sample t-test
##
## data:  change by is_northamerica
## t = -0.96619, df = 2.0612, p-value = 0.2166
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.8458473
```

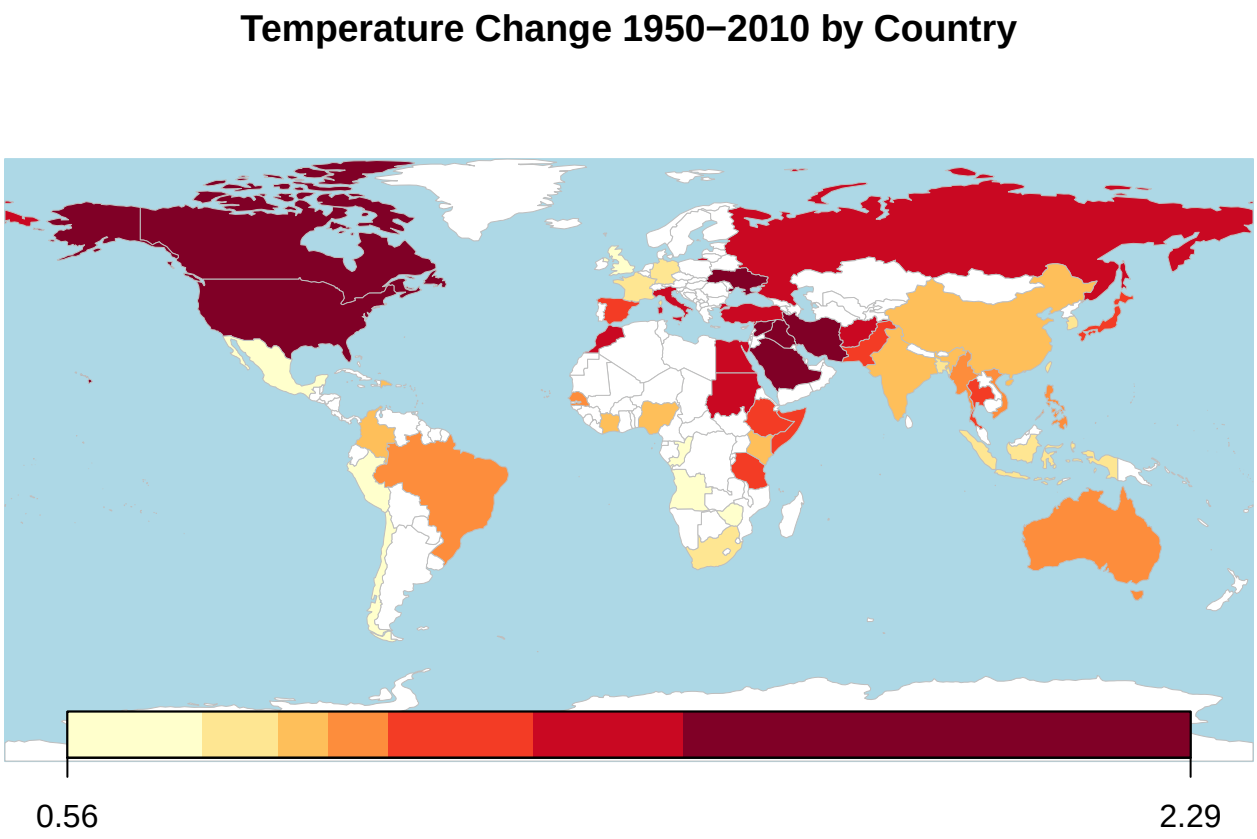
```
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1.087234            1.518391
```

From this first hypothesis test, we fail to reject the null hypothesis that North America has a different change in mean temperature from 1950-2010 than the rest of the world. The p-value was 0.2166, indicating a .2166 probability of obtaining results in which North America had an average difference of overall temperature change with respect to the average difference for other continents that was as great or greater than was observed as shown in this data set, assuming the null hypothesis.

```
##
##  Welch Two Sample t-test
##
## data:  change by is_europe
## t = -0.22096, df = 7.5932, p-value = 0.4155
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.3051081
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1.107800            1.148621
```

We also constructed a similar t-test for Europe, and when comparing the confidence intervals for the two, comparing Europe to the rest of the world produces a much smaller confidence interval (-0.4708, 0.3892), versus that for North America: (-2.2975, 1.4352). Additionally, the p-value for Europe's t-test was 0.4155, indicating a higher likelihood of Europe having the same true mean change in land temperature as the rest of the world.

```
## 49 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 194 codes from the map weren't represented in your data
```



Discussion

Throughout our analysis, we saw clear evidence that climate change exists and that global land and ocean temperatures are increasing to an unprecedented extent at an extraordinary rate. As seen in our linear regression models and associated scatter plots, there is meaningful evidence that the rate of temperature change has accelerated, especially since 1975.

Additionally, we were able to use hypothesis testing to support our claim that mean annual temperatures are increasing. In

our first hypothesis test, we concluded that there is a statistically significant increase in mean earth surface temperature from early-20th-century levels to the present. Furthermore, the mean temperature from years 1980 through 2013 is greater than the mean Earth temperature of the entire 20th century.

Our second hypothesis test found insufficient evidence of a statistically significant increase in the average degree of temperature change from before 1980 to after 1980, but did find a statistically significant increase in average temperature change since 2000. While the data sets have limited values for years after 2000, it still shows evidence that global warming trends are becoming more severe. This could be due to numerous confounding effects, including human causes such as increased greenhouse gas emissions in recent years.

Our third hypothesis test found that there is also sufficient evidence of an increase in global, annual year-to-year temperature change in more recent decades (post-1980) compared to the earlier 20th century. This result is supported by other findings from the scientific community, such as NOAA, which establishes that the rate of annual temperature change has increased in recent decades: climate.gov Source

The last hypothesis test found that both Europe and the United States of America do not significantly differ from the mean change in temperature globally. However, it also confirms that the U.S. is subject to more variable temperature fluctuations. This could provide evidence towards the effectiveness of initiatives in the European Union against climate change - such as the Paris Climate Accord - which the USA withdrew from in 2017.

While our findings supported the existence of climate change, it is important to also critique our data and methods. The data sets, `GlobalLandTemperaturesByMajorCity.csv` and `GlobalTemperatures.csv` attempt to comprehensively describe wide ranging global effects from 1849 to 2013. However there are inherent limitations in attempting this, as it fails to adequately or equally monitor all regions over this time period. For example, there were less European and South American cities included than Asian cities. This limits our ability to draw conclusions about trends in temperature change across Europe and South America for the same time period. Another limitation was the lack of data from earlier years. The data set included many NA values in the years 1849-1900, which is why we had to focus our analysis more specifically on the past 100 years. We also need to express that temperature collection and recording methods have improved over time. Thus, we have more confidence that recent observations are generally more accurate than observations preceding the digital revolution.

There were some limitations with our methods. In our difference of means hypothesis tests, while we found that there was not a difference between Europe and America's respective average change in temperature and the global change in temperature, our tests could not explain the reasons why the differences did not exist. Another limitation of the difference of means test is that while we did not find enough evidence to reject the null hypothesis that there was no difference, it does not mean that a difference does not exist. A final limitation in our methods was that our hypothesis tests could not actually make decisions about what they were testing, they could only be used as aids in decision making.

If we were able to restart the project, we would have found a data set that already contained all of the data we wanted to analyze. While we were able to join the two data sets together and perform our analysis, this was tedious compared to just using one data set. If we used one data set, we also could have performed more focused analysis on the data.