Climate Change EDA

```
#Load Required Libraries
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.2     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

```
library(ggplot2)
library(infer)
```

```
#Load Datasets
GlobalLandTemperaturesByCity <- read.csv("~/Desktop/RStudio/Projects/First Analytics EDA
GlobalLandTemperaturesByCountry <- read.csv("~/Desktop/RStudio/Projects/First Analytics E
GlobalLandTemperaturesByState <- read.csv("~/Desktop/RStudio/Projects/First Analytics EDA
GlobalTemperatures <- read.csv("~/Desktop/RStudio/Projects/First Analytics EDA project/Cl
```

This project uses four historical temperature datasets from the Berkeley Earth dataset on Kaggle, which track average land temperatures globally, by country, state, and city from the 18th century to the present.

Exploratory Data analysis

```
GlobalTemperatures |> head(5)
```

```
          dt LandAverageTemperature LandAverageTemperatureUncertainty
1 1750-01-01                  3.034                             3.574
2 1750-02-01                  3.083                             3.702
3 1750-03-01                  5.626                             3.076
4 1750-04-01                  8.490                             2.451
5 1750-05-01                 11.573                             2.072
  LandMaxTemperature LandMaxTemperatureUncertainty LandMinTemperature
1                 NA                            NA                 NA
2                 NA                            NA                 NA
3                 NA                            NA                 NA
4                 NA                            NA                 NA
```

```
5                 NA                            NA                 NA
  LandMinTemperatureUncertainty LandAndOceanAverageTemperature
1                            NA                             NA
2                            NA                             NA
3                            NA                             NA
4                            NA                             NA
5                            NA                             NA
  LandAndOceanAverageTemperatureUncertainty
1                                        NA
2                                        NA
3                                        NA
4                                        NA
5                                        NA
```

```r
range(GlobalTemperatures$dt)
```

```
[1] "1750-01-01" "2015-12-01"
```

```r
range(GlobalLandTemperaturesByCountry$dt)
```

```
[1] "1743-11-01" "2013-09-01"
```

```r
range(GlobalLandTemperaturesByState$dt)
```

```
[1] "1743-11-01" "2013-09-01"
```

```r
range(GlobalLandTemperaturesByCity$dt)
```

```
[1] "1743-11-01" "2013-09-01"
```

Data ranges from mid 1700's to early 2000's

```r
glimpse(GlobalTemperatures)
```

```
Rows: 3,192
Columns: 9
$ dt                                        <chr> "1750-01-01", "1750-02-01", …
$ LandAverageTemperature                    <dbl> 3.034, 3.083, 5.626, 8.490, …
$ LandAverageTemperatureUncertainty         <dbl> 3.574, 3.702, 3.076, 2.451, …
$ LandMaxTemperature                        <dbl> NA, NA, NA, NA, NA, NA, NA, …
$ LandMaxTemperatureUncertainty             <dbl> NA, NA, NA, NA, NA, NA, NA, …
$ LandMinTemperature                        <dbl> NA, NA, NA, NA, NA, NA, NA, …
$ LandMinTemperatureUncertainty             <dbl> NA, NA, NA, NA, NA, NA, NA, …
$ LandAndOceanAverageTemperature            <dbl> NA, NA, NA, NA, NA, NA, NA, …
$ LandAndOceanAverageTemperatureUncertainty <dbl> NA, NA, NA, NA, NA, NA, NA, …
```

```r
colSums(is.na(GlobalTemperatures))
```

```
                                     dt
                                      0
                   LandAverageTemperature
                                     12
        LandAverageTemperatureUncertainty
                                     12
                       LandMaxTemperature
                                   1200
            LandMaxTemperatureUncertainty
                                   1200
                       LandMinTemperature
                                   1200
            LandMinTemperatureUncertainty
                                   1200
           LandAndOceanAverageTemperature
                                   1200
LandAndOceanAverageTemperatureUncertainty
                                   1200
```

Missing data values for, seems like a majoirty of thse missing values are comming from the earlier record dates when accuracy and datacollection may not be as easy to collect.

```
nrow(GlobalTemperatures)
```

```
[1] 3192
```

```
GlobalTemperatures |> summarize(percent_null = sum(is.na(LandMaxTemperature))/n() * 100 )
```

```
  percent_null
1     37.59398
```

LandMaxTermpuatues Uncertainy has 1200 null values, same as 5 other features: LandMaxTemperatureUncertainty, LandMinTemperature ,LandMinTemperatureUncertainty, LandAndOceanAverageTemperature, LandAndOceanAverageTemperatureUncertainty. All of these come out toa 37.6% null proportion, meaning 37.6 percent of the data is missing.

This represetns a probnlem. 37.6% is a vvery signiiifacnt proportion, and due to modeling goals in the future and the importance of the variables, I am choosing to keep these columns rather then drop them.

```
GlobalTemperatures_na = GlobalTemperatures |> filter(is.na(LandAverageTemperature))
range(GlobalTemperatures_na$dt)
```

```
[1] "1750-11-01" "1752-09-01"
```

ƒAnother thing to notice is that a majoirty of these null values are in the earilier date ranges. This is most likey due to the fact that pre 1850's climate data often came from spares historical records (ship's journals, ect..). This tells me that removing these null values would be the best course of action as missingness in the early rows can introduce noise.

```
for (feature in colnames(GlobalTemperatures)) {
  # filter rows where current column has NA
  GlobalTemperatures_na = GlobalTemperatures |> filter(is.na(get(feature)))
  #checks if any rows returned, some features do not have null vals
  if (nrow(GlobalTemperatures_na) > 0) {
    # Find the date range where the current column is NA
    range_temp <- range(GlobalTemperatures_na$dt, na.rm = TRUE)
    cat(feature, "NA date range:", range_temp[1], "to", range_temp[2], "\n")
  }
}
```

```
LandAverageTemperature NA date range: 1750-11-01 to 1752-09-01
LandAverageTemperatureUncertainty NA date range: 1750-11-01 to 1752-09-01
LandMaxTemperature NA date range: 1750-01-01 to 1849-12-01
LandMaxTemperatureUncertainty NA date range: 1750-01-01 to 1849-12-01
LandMinTemperature NA date range: 1750-01-01 to 1849-12-01
LandMinTemperatureUncertainty NA date range: 1750-01-01 to 1849-12-01
LandAndOceanAverageTemperature NA date range: 1750-01-01 to 1849-12-01
LandAndOceanAverageTemperatureUncertainty NA date range: 1750-01-01 to 1849-12-01
```

Data Cleaning and Preparation

Dealing with Null values

Now that we know the null values has a high chance of introducing noise into our data, and that the null values are largelty pre 1850's, we should completely remove the null values to reflect more modern data trends.

```
GlobalTemperatures = GlobalTemperatures |> filter(dt >= as.Date("1850-01-01"))
```

Now we check if there are any more null values within the dataset

```
GlobalTemperatures |> summarise(across(everything(), ~sum(is.na(.))))
```

```
  dt LandAverageTemperature LandAverageTemperatureUncertainty
1  0                      0                                 0
  LandMaxTemperature LandMaxTemperatureUncertainty LandMinTemperature
1                  0                             0                  0
  LandMinTemperatureUncertainty LandAndOceanAverageTemperature
1                             0                              0
  LandAndOceanAverageTemperatureUncertainty
1                                         0
```

Looks good, now we have no nan values

```
glimpse(GlobalTemperatures)
```

```
Rows: 1,992
Columns: 9
$ dt                                        <chr> "1850-01-01", "1850-02-01", …
$ LandAverageTemperature                    <dbl> 0.749, 3.071, 4.954, 7.217, …
$ LandAverageTemperatureUncertainty         <dbl> 1.105, 1.275, 0.955, 0.665, …
$ LandMaxTemperature                        <dbl> 8.242, 9.970, 10.347, 12.934…
$ LandMaxTemperatureUncertainty             <dbl> 1.738, 3.007, 2.401, 1.004, …
$ LandMinTemperature                        <dbl> -3.206, -2.291, -1.905, 1.01…
$ LandMinTemperatureUncertainty             <dbl> 2.822, 1.623, 1.410, 1.329, …
$ LandAndOceanAverageTemperature            <dbl> 12.833, 13.588, 14.043, 14.6…
$ LandAndOceanAverageTemperatureUncertainty <dbl> 0.367, 0.414, 0.341, 0.267, …
```

However, notice how DT, representing the date time is a chr. We want to convert the DT column to Date format.

```
GlobalTemperatures <- GlobalTemperatures |> mutate(dt = as.Date(dt))
glimpse(GlobalTemperatures)
```

```
Rows: 1,992
Columns: 9
$ dt                                        <date> 1850-01-01, 1850-02-01, 185…
$ LandAverageTemperature                    <dbl> 0.749, 3.071, 4.954, 7.217, …
$ LandAverageTemperatureUncertainty         <dbl> 1.105, 1.275, 0.955, 0.665, …
$ LandMaxTemperature                        <dbl> 8.242, 9.970, 10.347, 12.934…
$ LandMaxTemperatureUncertainty             <dbl> 1.738, 3.007, 2.401, 1.004, …
$ LandMinTemperature                        <dbl> -3.206, -2.291, -1.905, 1.01…
$ LandMinTemperatureUncertainty             <dbl> 2.822, 1.623, 1.410, 1.329, …
$ LandAndOceanAverageTemperature            <dbl> 12.833, 13.588, 14.043, 14.6…
$ LandAndOceanAverageTemperatureUncertainty <dbl> 0.367, 0.414, 0.341, 0.267, …
```

Summary

Now that we have filtered and dealt with the nan values we are ready to look are some statistics.

```
GlobalTemperatures = GlobalTemperatures |> mutate(year = as.numeric(format(dt, "%Y"))) |>
```

```
GlobalTemperatures |> group_by(decade) |>
  summarise(decade_mean = mean(LandAverageTemperature))
```

```
# A tibble: 17 × 2
   decade decade_mean
    <dbl>       <dbl>
 1   1850        8.06
 2   1860        8.10
 3   1870        8.28
```

```
 4   1880          8.05
 5   1890          8.15
 6   1900          8.26
 7   1910          8.28
 8   1920          8.49
 9   1930          8.64
10   1940          8.73
11   1950          8.62
12   1960          8.63
13   1970          8.66
14   1980          8.91
15   1990          9.16
16   2000          9.49
17   2010          9.62
```

Here is the average temp per decade

```
GlobalTemperatures |> group_by(year) |>
  summarise(top_hottest = mean(LandAverageTemperature)) |>
  arrange(desc(top_hottest)) |> head(5)
```

```
# A tibble: 5 × 2
   year top_hottest
  <dbl>       <dbl>
1  2015        9.83
2  2007        9.73
3  2010        9.70
4  2005        9.70
5  2013        9.61
```

Top 5 average hottest YEARS

```
GlobalTemperatures |> group_by(year) |>
  summarise(top_coldest = mean(LandAverageTemperature)) |>
  arrange(top_coldest) |> head(5)
```

```
# A tibble: 5 × 2
   year top_coldest
  <dbl>       <dbl>
1  1862        7.56
2  1857        7.76
3  1884        7.77
4  1861        7.85
5  1875        7.86
```

Top 5 average Coldest YEARS

```r
decade_variance <- GlobalTemperatures |>
  group_by(decade) |>
  summarise(
    temp_variance = var(LandAverageTemperature, na.rm = TRUE)
  )
decade_variance
```

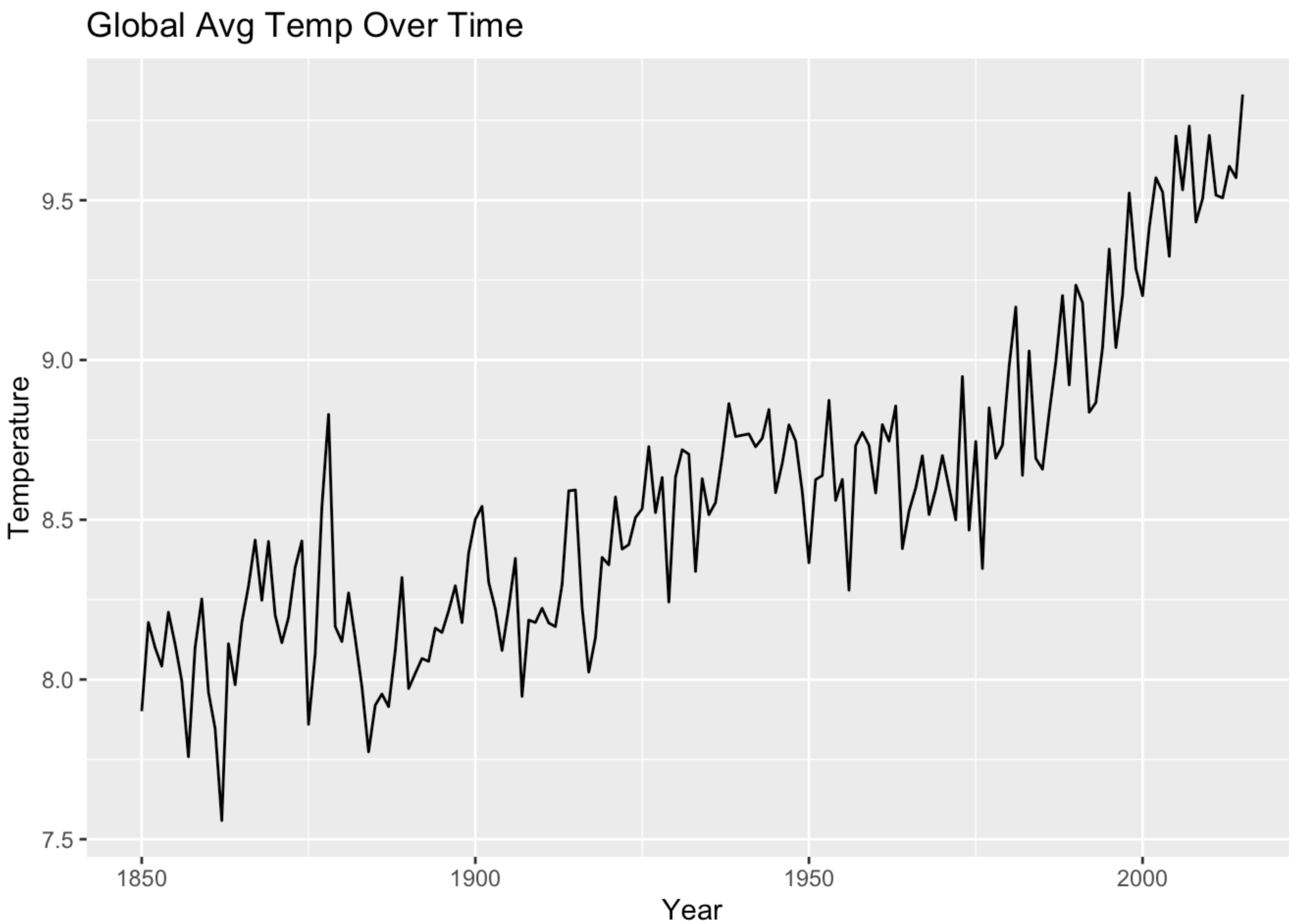
```
# A tibble: 17 × 2
   decade temp_variance
    <dbl>         <dbl>
 1   1850          19.1
 2   1860          19.5
 3   1870          19.6
 4   1880          19.0
 5   1890          19.3
 6   1900          18.5
 7   1910          18.4
 8   1920          17.7
 9   1930          17.8
10   1940          17.6
11   1950          17.7
12   1960          17.4
13   1970          17.5
14   1980          17.2
15   1990          17.0
16   2000          17.0
17   2010          17.7
```

Variance Per decade

Visualizations

```r
#Create dataframe with avg temp per year
global_temp_avg = GlobalTemperatures |> group_by(year) |>
  summarise(avg_temp = mean(LandAverageTemperature))
```

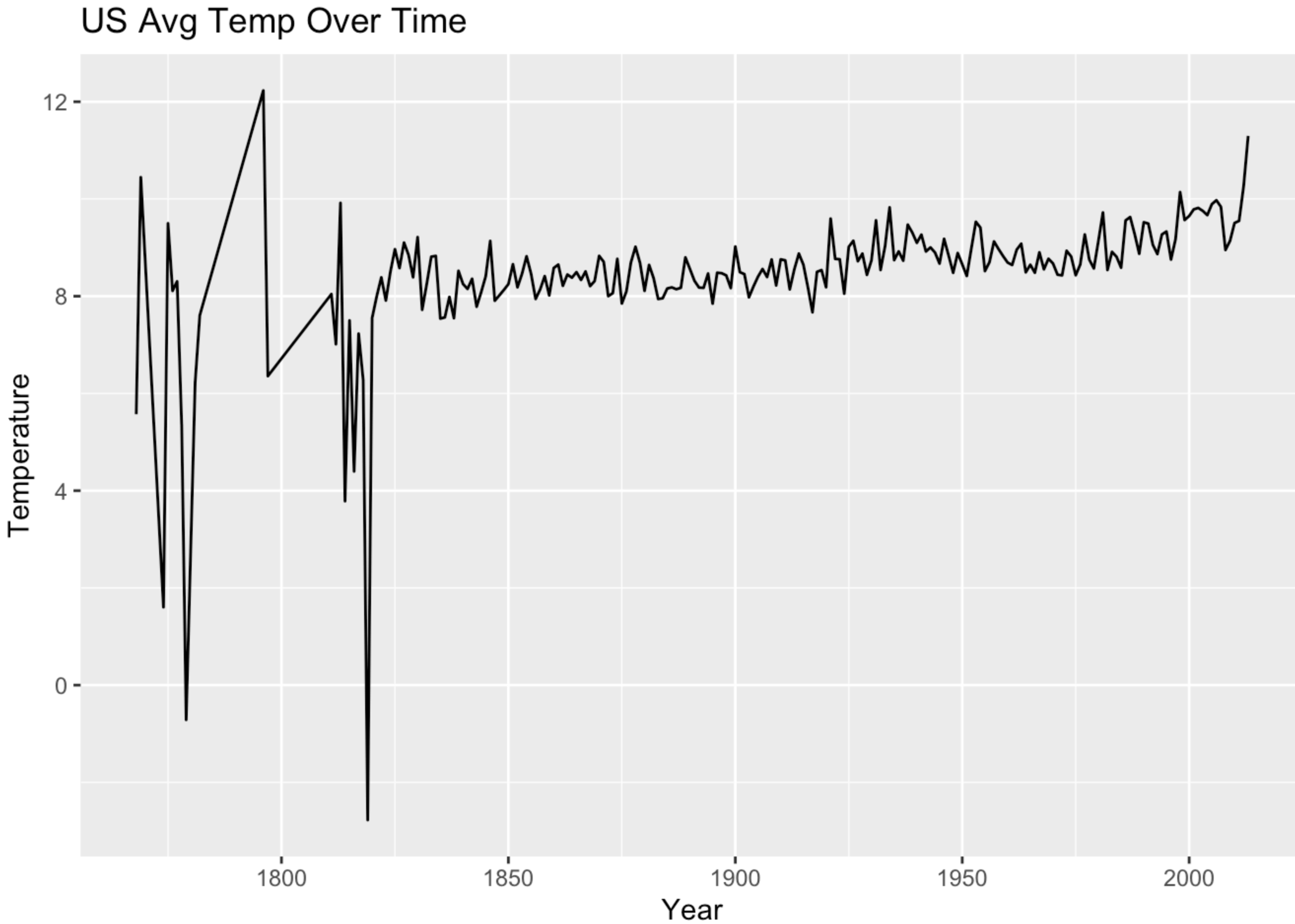
```r
global_temp_avg |> ggplot(aes(year, avg_temp)) +
  geom_line() +
  labs(title = "Global Avg Temp Over Time", x = "Year", y = "Temperature")
```



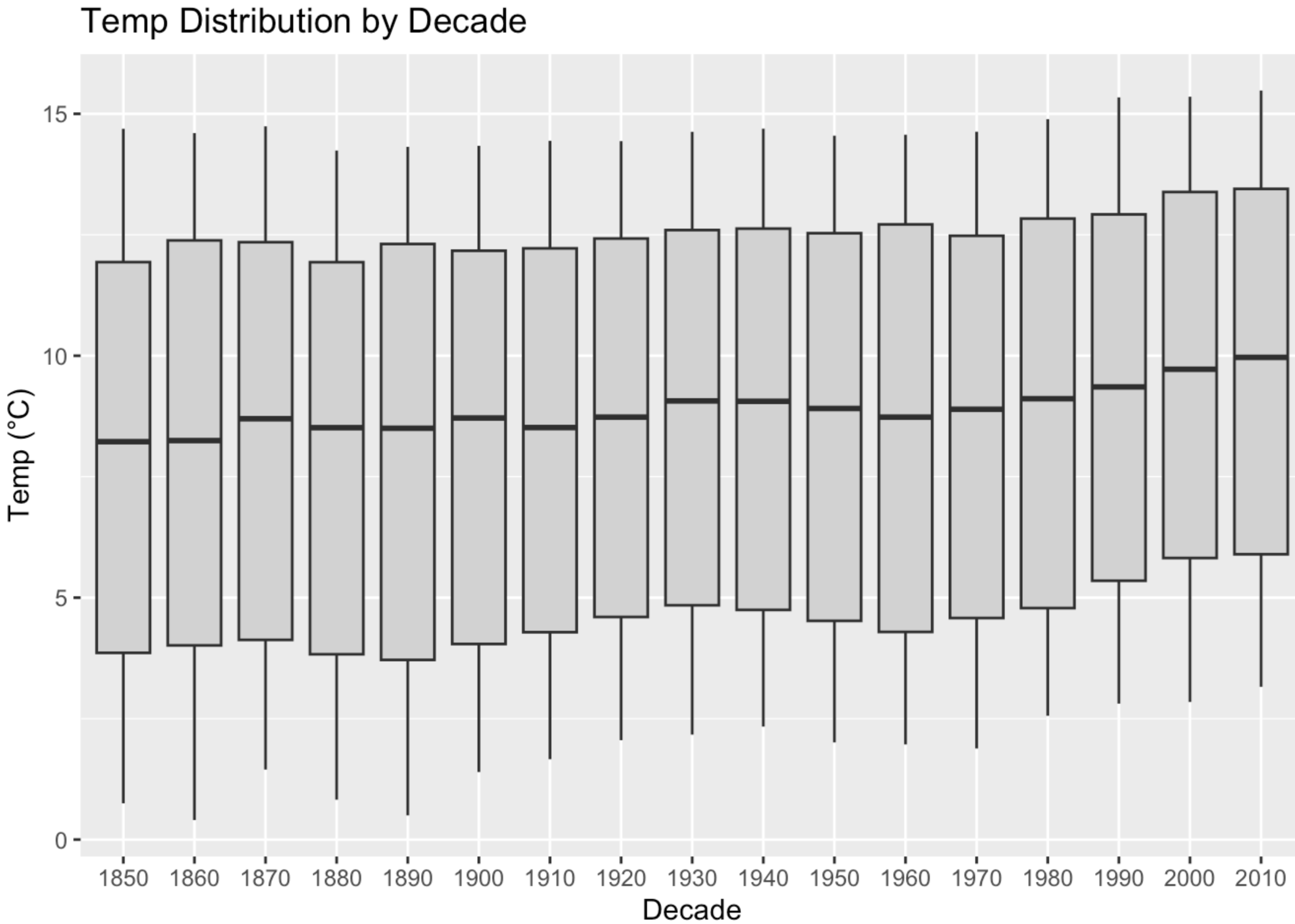

```r
GlobalLandTemperaturesByCountry_year_US = GlobalLandTemperaturesByCountry |>
  filter(!is.na(AverageTemperature), Country == "United States") |>
  mutate(dt = as.Date(dt)) |>
  mutate(year = as.numeric(format(dt, "%Y")))
```

```r
df = GlobalLandTemperaturesByCountry_year_US |> group_by(year) |>
  summarise(avg_temp = mean(AverageTemperature))
```

```r
df |> ggplot(aes(year, avg_temp)) + geom_line() +
  labs(title = "US Avg Temp Over Time", x = "Year", y = "Temperature")
```




```
GlobalTemperatures |> ggplot(aes(x = factor(decade), y = LandAverageTemperature)) +
  geom_boxplot(fill = "lightgray") +
  labs(title = "Temp Distribution by Decade", x = "Decade", y = "Temp (°C)")
```



Hypothesis Testing

The purpose of this hypothesis test is to determine whether the observed increase in average temperatures in the **United States** can be explained by **random chance**, or if it reflects a **statistically significant shift** in climate patterns. I specifically focus on comparing temperatures **before and after 1950**, as the post-1950 period marks the onset of notably sharper year-over-year temperature increases.

- **Null Hypothesis (H_0)**: There is **no difference** in the mean annual temperature before and after 1950; any observed difference is due to **random variation**.

- **Alternative Hypothesis (H_1)**: The **mean annual temperature after 1950 is higher** than before 1950, and this increase is **not due to random chance**.

A **significance level of α = 0.05** is used for this test. This means I will reject the null hypothesis if the p-value is **less than 0.05**, indicating strong evidence against the idea that the temperature increase is due to randomness.

```
#filter out Null vals and label pre/post 1950's
us_data = GlobalLandTemperaturesByCountry |>
  filter(!is.na(AverageTemperature)) |>
```

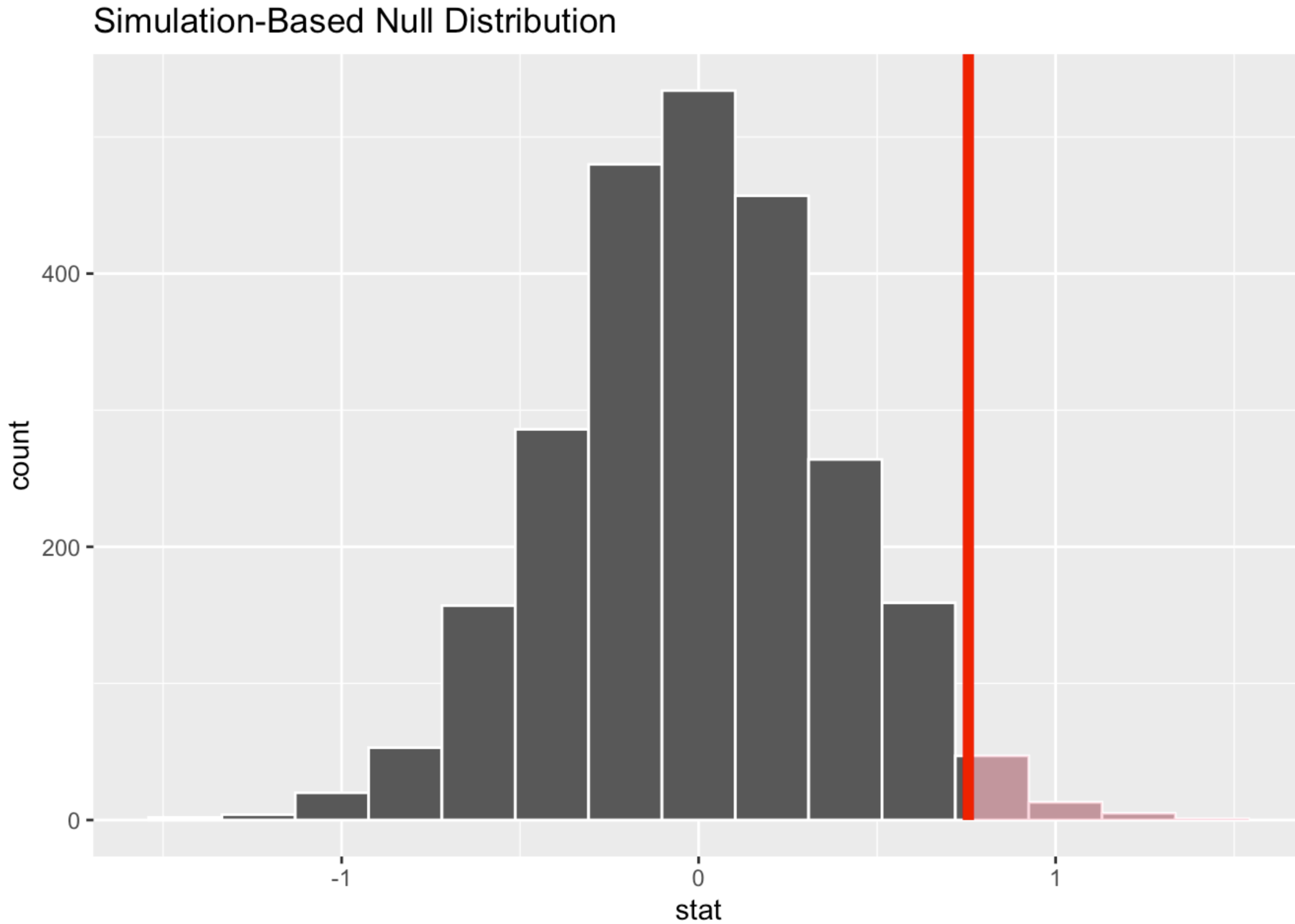
```
  mutate(dt = as.Date(dt),
    year = as.numeric(format(dt, "%Y")),
    period = ifelse(year < 1950, "Pre1950", "Post1950")
  ) |> filter(Country == "United States")
us_data |> head(5)
```

```
          dt AverageTemperature AverageTemperatureUncertainty       Country
1 1768-09-01             15.420                         2.880 United States
2 1768-10-01              8.162                         3.386 United States
3 1768-11-01              1.591                         3.783 United States
4 1768-12-01             -2.882                         4.979 United States
5 1769-01-01             -3.952                         4.856 United States
  year  period
1 1768 Pre1950
2 1768 Pre1950
3 1768 Pre1950
4 1768 Pre1950
5 1769 Pre1950
```

```
mean_pre = us_data |> filter(period == "Pre1950") |> summarize(mean_temp_pre = mean(Avera
mean_post = us_data |> filter(period == "Post1950") |> summarize(mean_temp_pre = mean(Ave

obs_test_stat = mean_post - mean_pre
```

```
perm_us = us_data |>
  specify(response = AverageTemperature, explanatory = period) |>
  hypothesize(null = "independence") |>
  generate(reps = 2482, type = "permute") |>
  calculate(stat = "diff in means", order = c("Post1950","Pre1950")) |>
  visualize() +
  shade_p_value(obs_test_stat, direction = "greater")
perm_us
```




```
pp = us_data |>
  specify(response = AverageTemperature, explanatory = period) |>
  hypothesize(null = "independence") |>
  generate(reps = 2000, type = "permute") |>
  calculate(stat = "diff in means", order = c("Post1950","Pre1950")) |>
  get_p_value(obs_stat = obs_test_stat, direction = "greater")

pp
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1  0.0305
```

With a P_val of 0.0275 we exceed the our significance level α = **0.05** and reject the null hypothesis. This indicates that there is a **2.75% chance** of observing a result as extreme or more extreme than yours **if the null hypothesis were true**.

The final verdict: The **mean annual temperature after 1950 is higher** than before 1950, and this increase is **not due to random chance**.