Buoy_HW

Jonathan Neimann

2024-09-27

```
library(data.table)
library(lubridate)
library(dplyr)
library(ggplot2)
```

#Question B

I think most of time it is appropriate to convert missing data to NA. This is because there are built-in R functions that will allow you to omit them when running statistic calculations on the data. If the cell is simply blank, finding things like the mean and the median of certain variables would not work. However, there are some situations where converting them to NA's might not be applicable. For example, maybe a missing cell is serving as a placeholder for data that will be collected at another date, or maybe certain inputs (ie 999) mean something to the data collector and represent a key that we are unaware of.

For this data set, there seems to be a lot of NA"s when we replace the 99's and 999's it's hard to determine any patterns other than some years had certain data collected while other years did not (if you add up the sum of all the NA's in the data frame it is 2,314,343). However, you can see that the more recent the dates get, the more data seems to be filled in, indicating that the buoy has been upgraded over time. If there is a section that was once collecting data and is no longer, maybe that part of the buoy broke or malfunctioned and was fixed at a later date. Below is the code i used to replace all 99's and 999's with NA's in the data frame.

```
bd_list[bd_list == 99 | bd_list == 999] <- NA
```

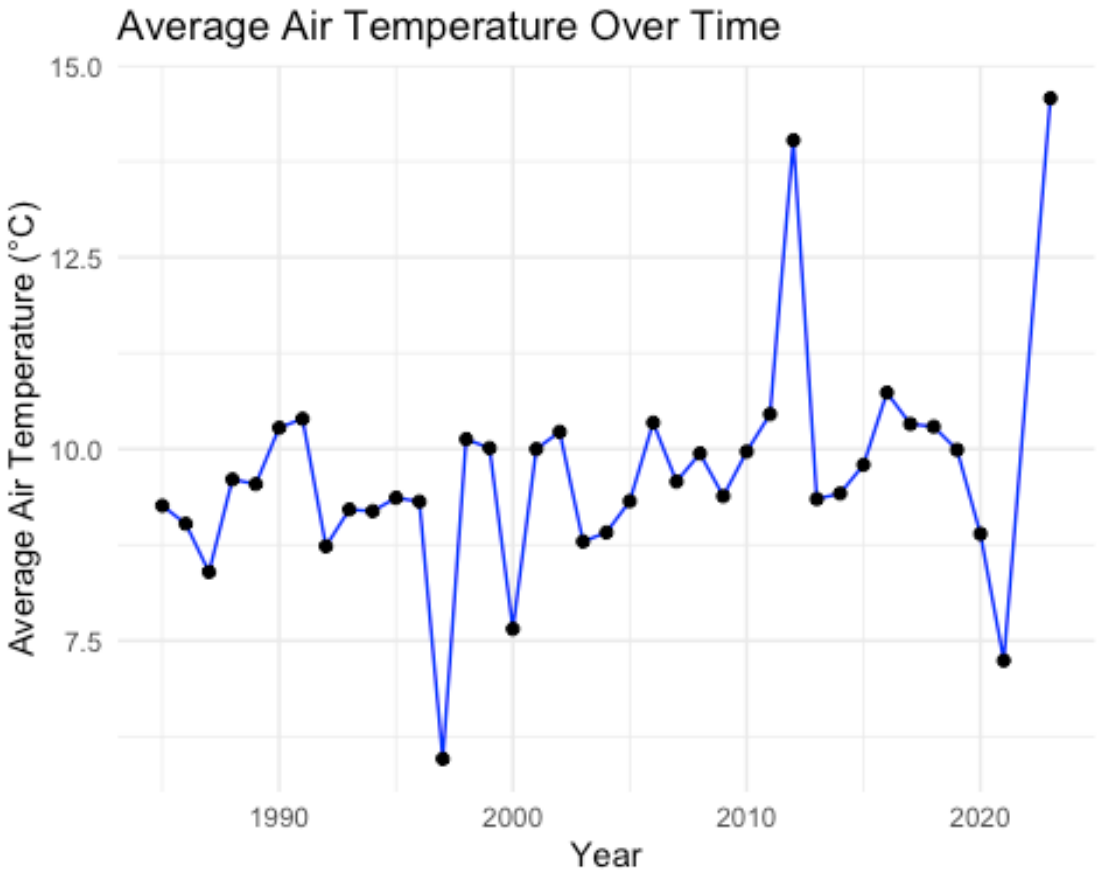
#Question C

Yes, this buoy data can show trends of climate change in both air temperature and water temperature data. First, we can look at air temperature. What I did here was create a new data frame called buoy_year_air which shrinks down the dates to just one row per each year and then averages out the average air temperature for that year. I then plotted it out using a simple scatterplot with years on the x-axis and average temperature on the y-axis and connected them with a line to show the trend. Although there is an upward overall trend, i was somewhat surprised to see it fairly level throughout all the years, both rising and falling fairly consistently. There is however, 2 huge spikes in average in both 2012 and 2023. This data may be a bit misleading however as the buoy is taking the average air temperature of it's location over water, where air temperature can remain rather temperate. Where we really start to see a bigger indicator of global warming is when we

look at the average water temperature. (note* the year 2022 came up as NA for its average air temperature so I excluded it from the plot).

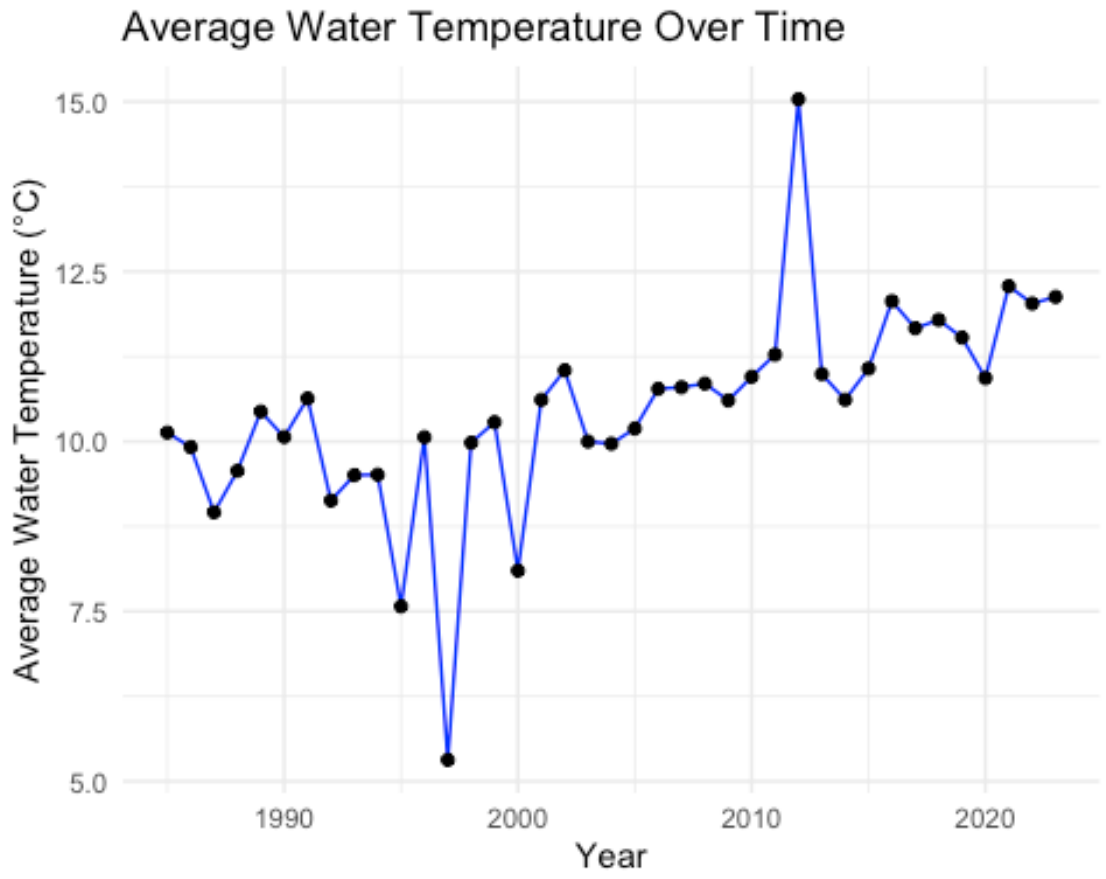
```
buoy_year_air <- bd_list %>%
  mutate(Year = year(Date)) %>%
  group_by(Year) %>%
  summarize(Avg_ATMP = mean(ATMP, na.rm = TRUE), .groups = 'drop') %>%
  filter(!is.na(Avg_ATMP))

ggplot(buoy_year_air, aes(x = Year, y = Avg_ATMP)) +
  geom_line(color = "blue") +
  geom_point() +
  labs(title = "Average Air Temperature Over Time",
       x = "Year",
       y = "Average Air Temperature (°C)") +
  theme_minimal()
```



```r
buoy_year_water <- bd_list %>%
  mutate(Year = year(Date)) %>%
  group_by(Year) %>%
  summarize(Avg_WTMP = mean(WTMP, na.rm = TRUE), .groups = 'drop') %>%
  filter(!is.na(Avg_WTMP))

ggplot(buoy_year_water, aes(x = Year, y = Avg_WTMP)) +
  geom_line(color = "blue") +
  geom_point() +
  labs(title = "Average Water Temperature Over Time",
       x = "Year",
       y = "Average Water Temperature (°C)") +
  theme_minimal()
```



Here in the water temperature graph (constructed the same way) we can see a clear upward trend in average temperature from 1985 to 2023, indicating the effects of climate change in the oceans. It is worth noting that in both graphs, the years 1997 and 2012 seem to have significant spikes in either direction. This may be worth examining more with further EDA.

In terms of statistical significance, we can run regression models on both data frames with average air and water temperature as our target variable and year as our predictor. When we create these models and display the results we can see that the slope coefficient for the year is positive, further showing an upward linear trend in both water and air temperature.

```
air_model = lm(Avg_ATMP~Year, data = buoy_year_air)
water_model = lm(Avg_WTMP~Year, data = buoy_year_water)

summary(air_model)

##
## Call:
## lm(formula = Avg_ATMP ~ Year, data = buoy_year_air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4397 -0.5040  0.0139  0.5168  4.0793
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -74.66171   41.04015  -1.819   0.0772 .
## Year          0.04209    0.02048   2.055   0.0472 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.39 on 36 degrees of freedom
## Multiple R-squared:  0.105,  Adjusted R-squared:  0.08013
## F-statistic: 4.223 on 1 and 36 DF,  p-value: 0.04719

summary(water_model)

##
## Call:
## lm(formula = Avg_WTMP ~ Year, data = buoy_year_water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5863 -0.2855  0.0819  0.3935  3.9124
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -152.88599   34.68132  -4.408 8.62e-05 ***
## Year          0.08152    0.01731   4.710 3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.216 on 37 degrees of freedom
## Multiple R-squared:  0.3749, Adjusted R-squared:  0.358
## F-statistic: 22.19 on 1 and 37 DF,  p-value: 3.44e-05
```

#Question D

What I am attempting to show here is if air temperature has an effect on precipitation on a given day. First, I loaded in the rain data, adjusted the date column to with lubridate, and took a look at the summary for the precipitation variable (HPCP)

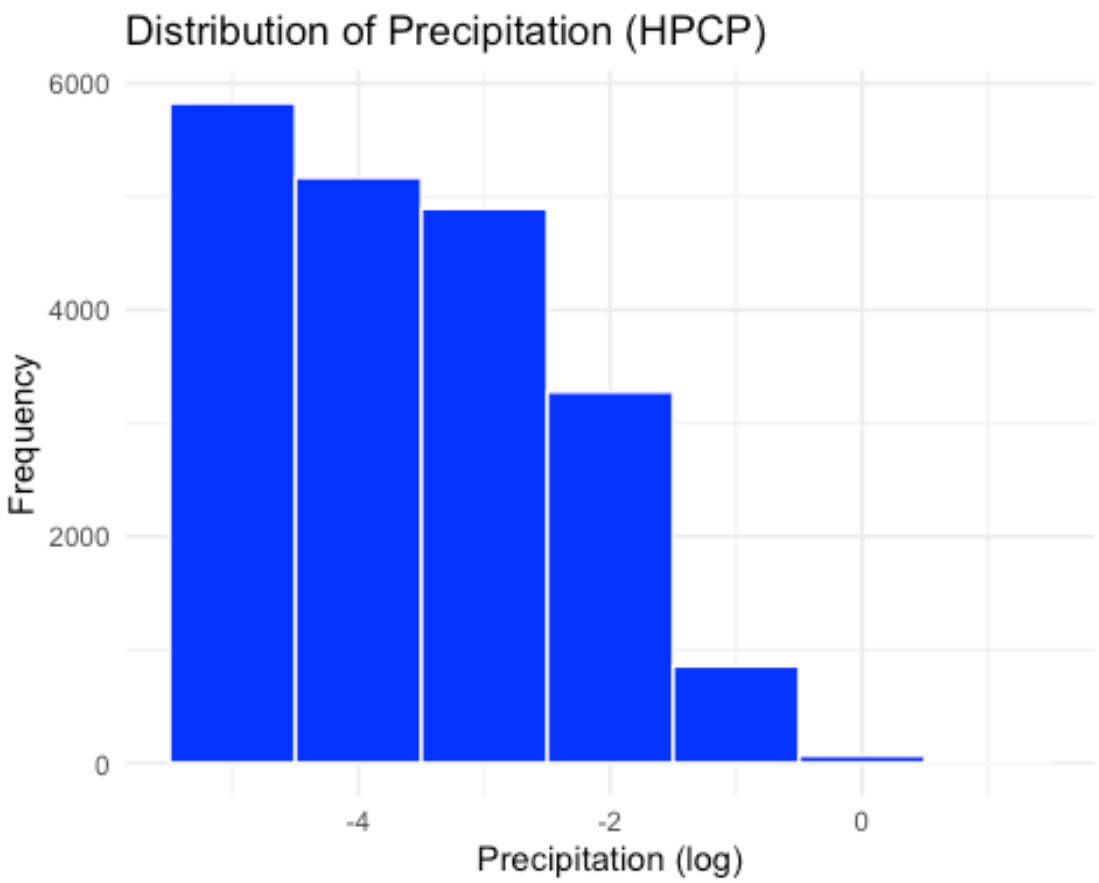
```
library(dplyr)
rain_data = read.csv('Rainfall.csv')
rain_data$DATE <- ymd_hm(rain_data$DATE)
summary(rain_data$HPCP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.01000 0.03875 0.04000 2.03000
```

We can see that the data for HPCP shows some pretty small numbers as the mean and median, with a maximum that is very far away relatively speaking. This indicates a right skewed distribution which I show here in a histogram. However, since the x-axis is so spread apart, I decided to take the log of HPCP to bring it closer together.

```
mean_precipitation <- mean(rain_data$HPCP)
ggplot(rain_data, aes(x = log(HPCP))) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of Precipitation (HPCP)",
       x = "Precipitation (log)",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Removed 11632 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

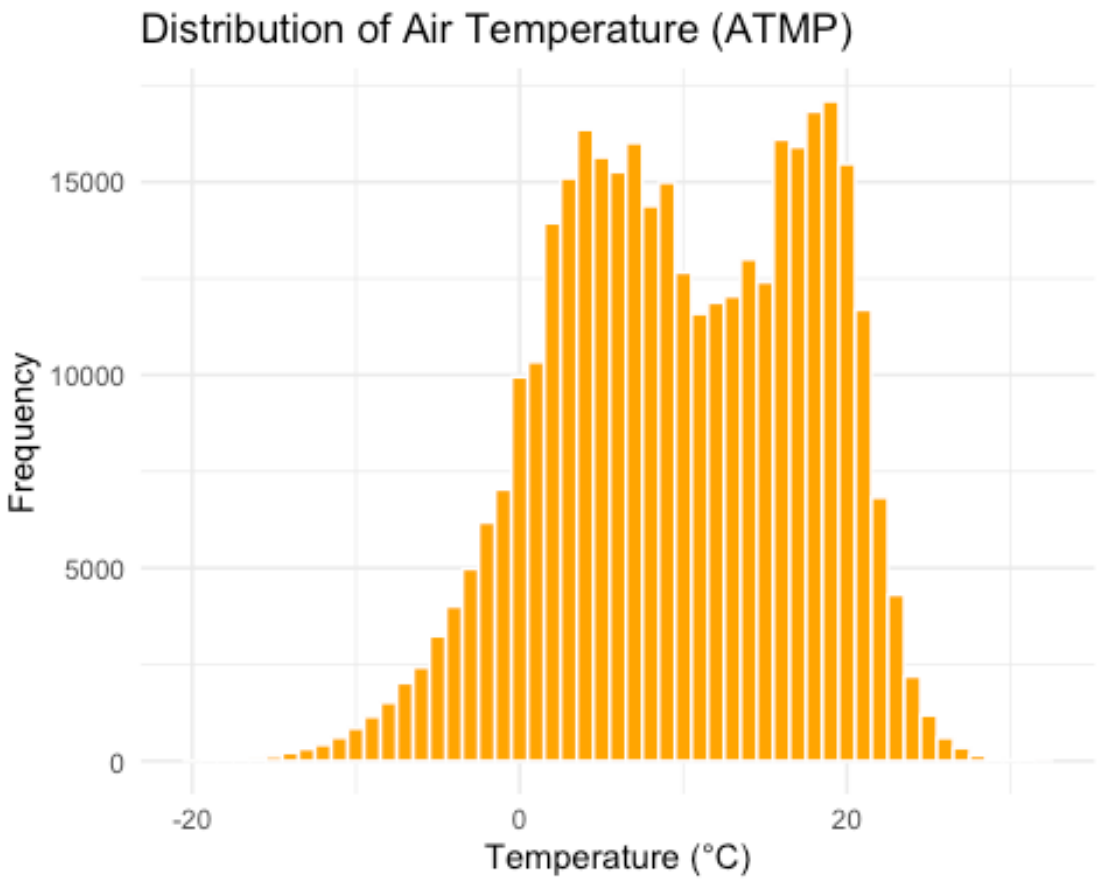


you can indeed see that this is a right skewed distribution, with the overwhelming majority of the data in the first three bins (lower amount of rainfall).

Next i wanted to see the distribution of air temperature, which I also used a hitagram for.

```
ggplot(bd_list, aes(x = ATMP)) +
  geom_histogram(binwidth = 1, fill = "orange", color = "white") +
  labs(title = "Distribution of Air Temperature (ATMP)",
       x = "Temperature (°C)",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Removed 102761 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Here you can see the data is a bit more normally distributed, but skewed slightly to the left. This makes sense based on the trends we saw with climate change earlier, as higher temperatures would be more prevelant in the data.

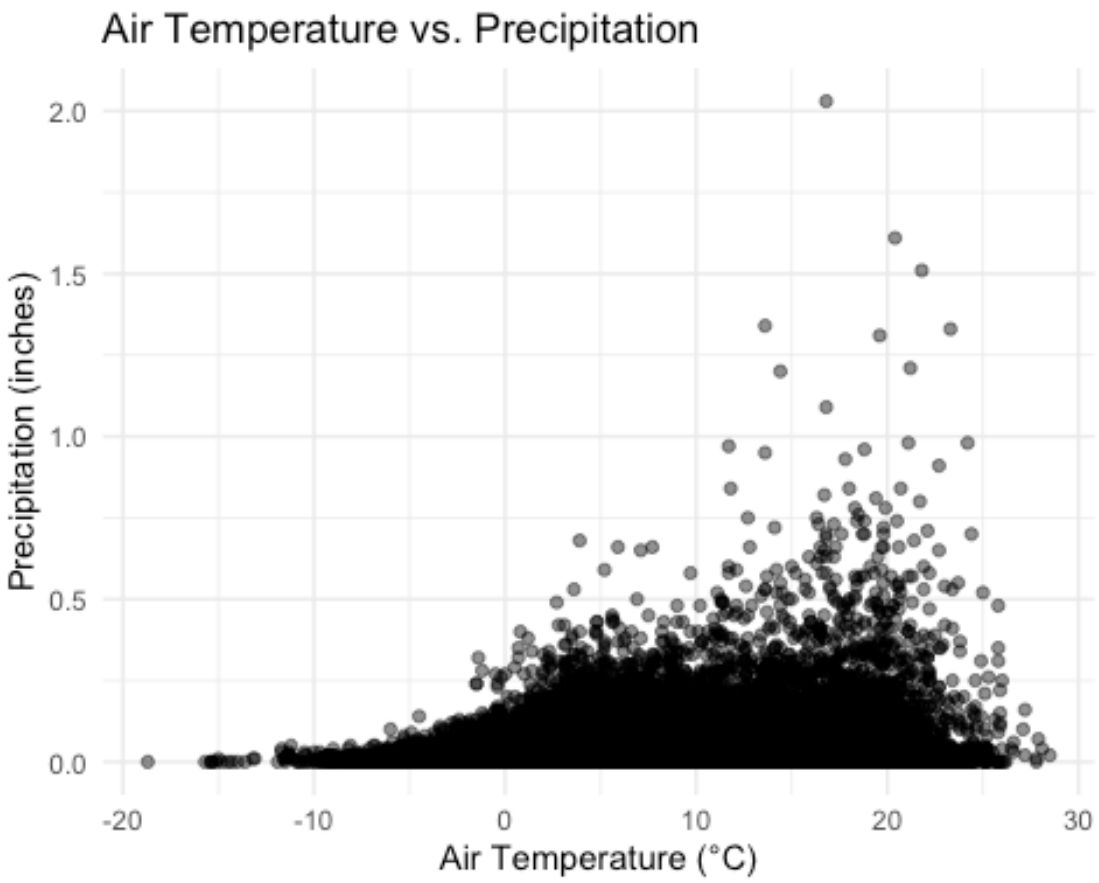
Next I wanted to see the relationship air temperature has on precipitation, so I combined the two data frames on rows where the dates matched. Because the rain_data does not count specifically by every hour, there are many rows in the buoy data that gor dropped. However the rain_data still provided a significant amount of rows to look at and the join was easily done because we made the date format the same between the two tables using lubridate

```
rain_data <- rain_data %>% rename(Date = DATE)
combined_data <- inner_join(rain_data, bd_list, by = "Date")
```

Lastly, I created a scatter plot to show the relationship between air temperature and precipitation from the new data frame called combined_data

```
ggplot(combined_data, aes(x = ATMP, y = HPCP)) +
  geom_point(alpha = 0.5) +
  labs(title = "Air Temperature vs. Precipitation",
       x = "Air Temperature (°C)",
       y = "Precipitation (inches)") +
  theme_minimal()
```

```
## Warning: Removed 148 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



As you can see in this plot, there is certainly an increase in the amount of rain as the temperature gets higher, with some pretty huge precipitation days around the 15-25 degree celsius range. This indicates that the highest chance of a really big rainfall day would be on a day that is in this temperature range. It is also worth noting that when the temperature is above 25 degrees celsius, there is a steep drop off in precipitation in general. This indicates that on really warm days, it is not very likely to rain at all.

I decided to run two simple inverse models. One looking at how air temperature affects precipitation and one on how precipitation affects air temperature using the combined_data data frame that I made.

```
temp_model1 = lm(HPCP ~ ATMP, data = combined_data)
temp_model2 = lm(ATMP ~ HPCP, data = combined_data)

summary(temp_model1)

##
## Call:
## lm(formula = HPCP ~ ATMP, data = combined_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07032 -0.03588 -0.02249  0.00641  1.97763
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.494e-02  7.075e-04   35.25   <2e-16 ***
## ATMP        1.632e-03  6.449e-05   25.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07568 on 29482 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.02127,    Adjusted R-squared:  0.02124
## F-statistic: 640.8 on 1 and 29482 DF,  p-value: < 2.2e-16
```

```
summary(temp_model2)
```

```
##
## Call:
## lm(formula = ATMP ~ HPCP, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.7740  -5.1869  -0.7043   5.3654  20.1654
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.07400    0.04419  182.71   <2e-16 ***
## HPCP        13.03068    0.51478   25.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.762 on 29482 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.02127,    Adjusted R-squared:  0.02124
## F-statistic: 640.8 on 1 and 29482 DF,  p-value: < 2.2e-16
```

You can see that the coefficients for the first model are very small, which makes sense as we are dealing with inches of rain. Yet there is still a positive slope meaning that everytime the precipitation increases, we can also expect the temperature to increase ever so slightly.

The second model tells a similar story but in a different way. Now the slope coefficient for the HPCP variable is massive. This also makes sense because it determines what happens if the precipitation increases by one full inch, which is a large amount in this data set. Nonetheless, if that were to happen this model is telling us we can expect a temperature increase of 13 degrees celsius.