A 50 Year Analysis of Temperature and Precipitation Data at San Francisco International Airport- A Deeper Dive

Author: Hilary Hickingbotham

Introduction

In the first project, I did an analysis of 50 years of temperature and precipitation data from San Francisco International Airport. For this second project, I want to expand on my findings, and answer some additional questions:

- Are temperature and precipitation normal?
- What is the correlation between temperature and precipitation?
- What are the different correlation tests that can be performed?
- Which is the best to use in this particular instance?

Gathered Data

The data was gathered using the NOAA database, looking specifically at the weather station located at San Francisco International Airport (Station USW00023234). Data pulled was for Max and Min Temperature and Measured Precipitation from 1Jan1970-31Dec2020. In total, there were 18,628 entries in the data set.

Data Cleaning and Organization

Before any analysis could be performed, a few data cleaning and organizational tasks were completed in preparation. These tasks included:

- Remove NAs in the data if any.
- Fixing the Date in R from a “character” to “date” variable for time analysis.
- Added a column that would assign TRUE/FALSE if there was any measured precipitation on that day.

```
#Identifying NA values and removing them
(cols_withNA <- apply(sfoweather50, 2, function(x) sum(is.na(x))))
sfoweather50_2 <- sfoweather50[complete.cases(sfoweather50),]
View(sfoweather50_2)

#Fixing the date variable
sfoweather50$date_fixed <- as.Date(sfoweather50$DATE, format = "%Y-%m-%d")

#adding the True/False indicator for precipitation
sfoweather50$RAIN[sfoweather50$PRCP > 0] <- TRUE
sfoweather50$RAIN[sfoweather50$PRCP == 0] <- FALSE


dailysfo50 <- sfoweather50_2 %>%
  mutate(
    year = year(date_fixed),
    yday = yday(date_fixed),
    month = month(date_fixed, label = TRUE))
View(dailysfo50)
```

Data Summary

A call to R will give a nice summary of the data as seen below.

```
Summary(sfoweather50)
```

```
STATION               NAME                DATE                PRCP        
 Length:18608       Length:18608       Length:18608       Min.   :0.00000  
 Class :character   Class :character   Class :character   1st Qu.:0.00000  
 Mode  :character   Mode  :character   Mode  :character   Median :0.00000  
                                                          Mean   :0.05331  
                                                          3rd Qu.:0.00000  
                                                          Max.   :5.59000  
      TMAX             TMIN          RAIN          date_fixed        
 Min.   : 37.00   Min.   :24.00   FALSE:15356   Min.   :1970-01-01  
 1st Qu.: 60.00   1st Qu.:47.00   TRUE : 3252   1st Qu.:1982-09-26  
 Median : 65.00   Median :51.00                 Median :1995-06-22  
 Mean   : 65.84   Mean   :50.23                 Mean   :1995-06-23  
 3rd Qu.: 71.00   3rd Qu.:54.00                 3rd Qu.:2008-03-17  
 Max.   :105.00   Max.   :72.00                 Max.   :2020-12-31  
```

Compared to other places in the United States, it only precipitated 17.45% of days from 1970-2020 in San Francisco. The median TMAX was 65° and the median TMIN was 51°.

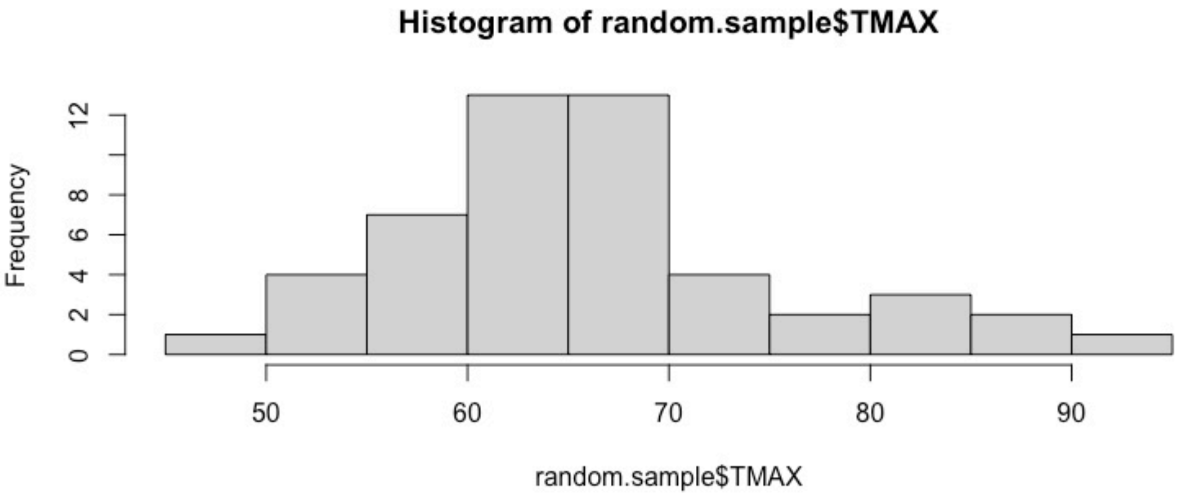
Test for Normality

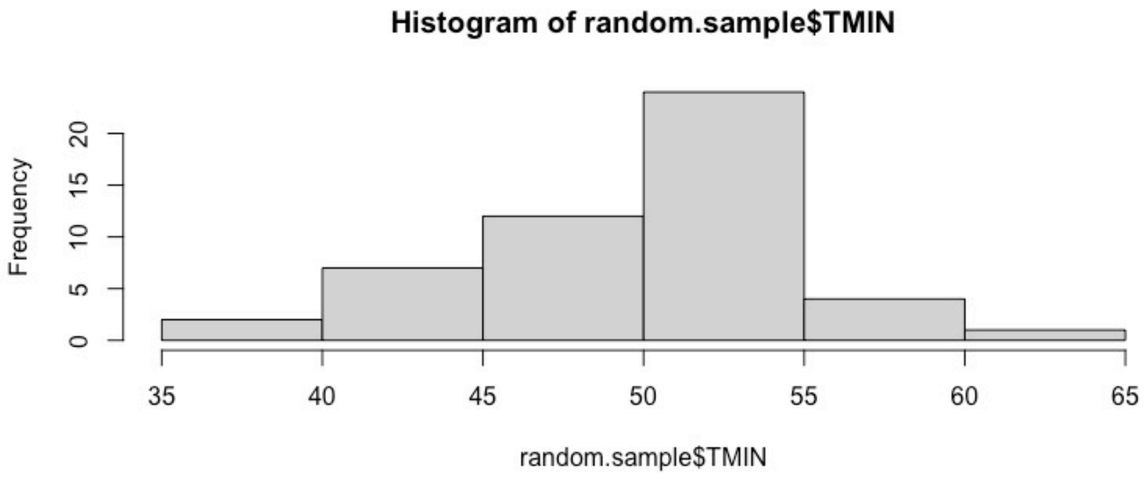
It is safe to assume that temperatures follow a normal distribution. From the first project, precipitation looked like an exponential, so model fitting will be done to determine the parameters. The Central Limit Theorem establishes that, in many situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. A random sample of 50 was generated and tested for normality. The Shapiro-Wilks test was performed as the normality test and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores.

```
random.sample <- dailysfo50[sample(nrow(dailysfo50), 50), ]
View(random.sample)
shapiro.test(random.sample$TMAX)
#p-value = 0.4195
shapiro.test(random.sample$TMIN)
#p-value = 0.08724
```

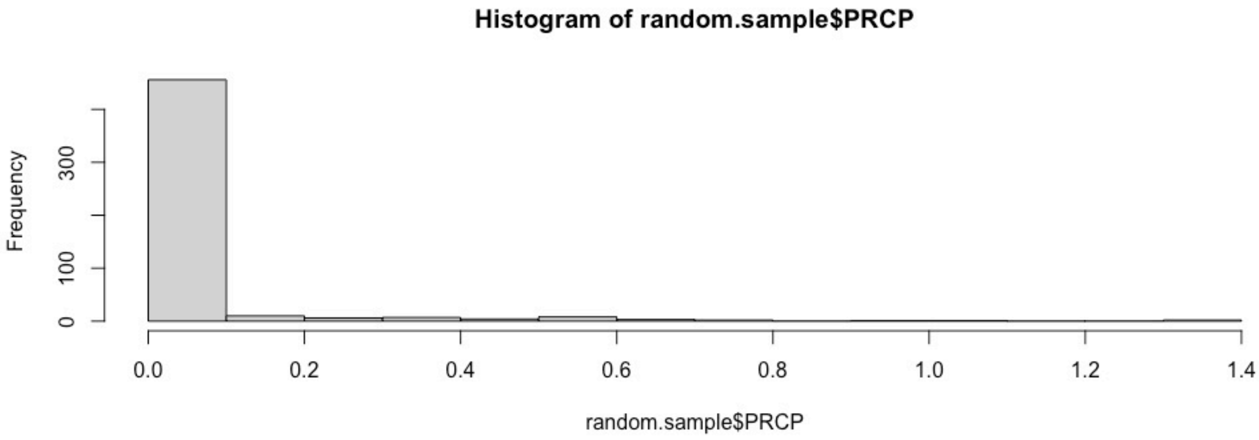
If the p-value > 0.05 implying that the distribution of the data is not significantly different from normal distribution. In other words, it can be assumed that TMIN and TMAX is normal.

```
#fitting normal data for TMIN and TMAX
tmaxfit <- fitdistr(random.sample$TMAX, densfun="normal")
      mean          sd
  66.520       9.918
tminfit <- fitdistr(random.sample$TMIN, densfun="normal")
      mean          sd
  50.740       5.257
```



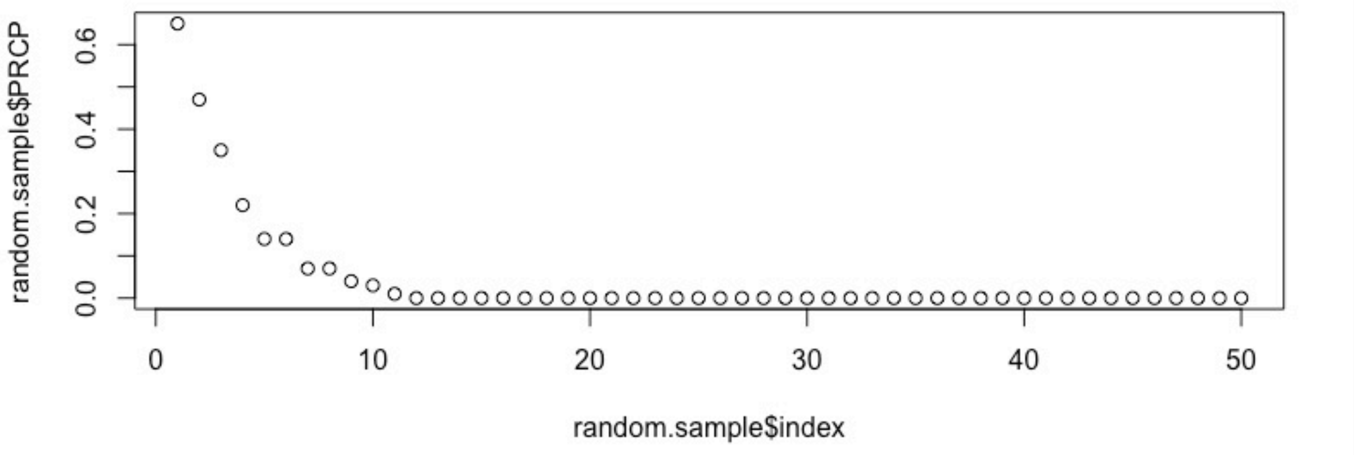


When looking at the random sample of 50 data points, the precipitation data does not look normal, but rather, exponential, which was expected.



The same random sample of 50 data points was then indexed and plotted with precipitation in descending order.

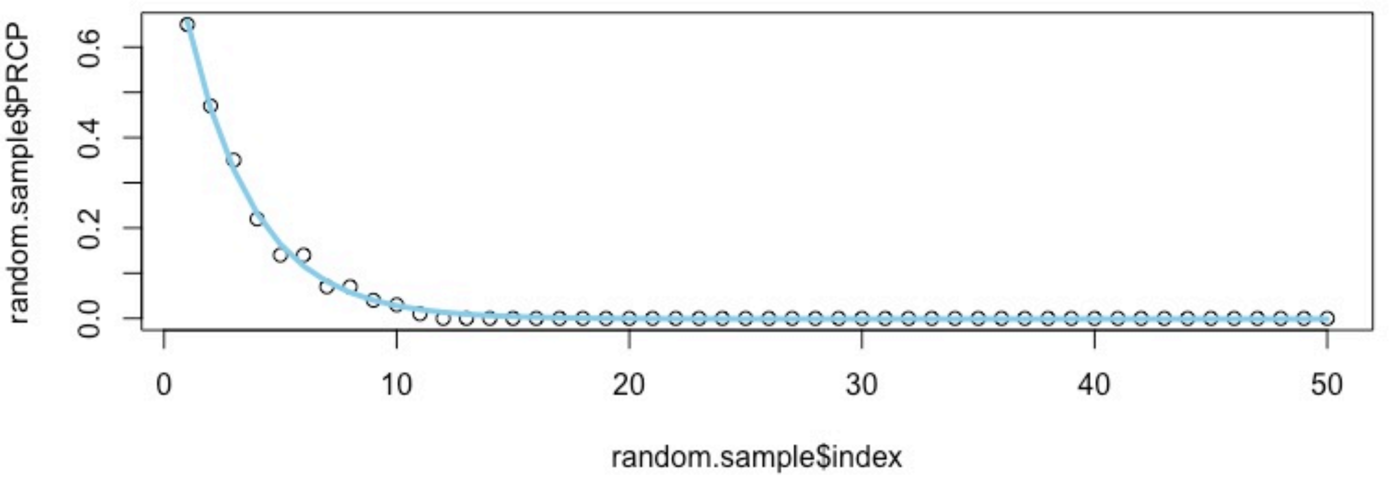
```
random.sample$index <- 1:nrow(random.sample)
random.sample <- random.sample[order(-random.sample$PRCP), ]
plot(random.sample$index, random.sample$PRCP)
```



It’s clear from the graph, that the random sample of precipitation is not normally distributed, but rather exponentially distributed. A model was created to predict the parameters. The model was:

$$E(y)=\alpha e^{\beta x}+ \theta$$

```
#estimating parameters
theta.0 <- min(random.sample$PRCP) * 0.5

# Estimate the rest parameters using a linear model
model.0 <- lm(log(random.sample$index - theta.0) ~ random.sample$PRCP,
data=random.sample)
alpha.0 <- exp(coef(model.0)[1])
beta.0 <- coef(model.0)[2]

# Starting parameters
start <- list(alpha = alpha.0, beta = beta.0, theta = theta.0)
start

model <- nlsLM(random.sample$PRCP ~ alpha * exp(beta * random.sample$index) + theta ,
data = random.sample, start = start)

# Plot fitted curve
plot(random.sample$index, random.sample$PRCP)
lines(random.sample$index, predict(model, list(x = random.sample$PRCP)), col = 'skyblue',
lwd = 3)
```



```
Formula: random.sample$PRCP ~ alpha * exp(beta * random.sample$index) +
    theta

Parameters:
                         Estimate Std. Error t value Pr(>|t|)
alpha.(Intercept)        0.930468   0.012926  71.983   <2e-16 ***
beta.random.sample$PRCP -0.345667   0.006049 -57.140   <2e-16 ***
theta                   -0.001266   0.001203  -1.052    0.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007449 on 47 degrees of freedom

Number of iterations to convergence: 21
Achieved convergence tolerance: 1.49e-08
```

From the model, the estimated parameters for the exponential are:

α= 0.93
β= -0.346
θ= -0.0012

Correlation

I wanted to deep dive into the correlation between TMIN, TMAX and Precipitation. There are three common ways to calculated correlation, and I wanted to analyze all three options to determine the best for this situation. I was familiar with the Pearson's Correlation Coefficient, but unfamiliar with the Spearman's Rank Correlation Rho, and Kendall's Rank Correlation Tau, and how those may be applicable in this case.

Pearson's, Spearman's and Kendall's correlation coefficients are the most commonly used measures of monotone association, with the latter two usually suggested for non-normally distributed data. These three correlation coefficients can be represented as the differently weighted averages of the same concordance indicators. The weighting used in the Pearson's correlation coefficient could be preferable for reflecting monotone association in some types of continuous and not necessarily bivariate normal data.

The first test is the default correlation test in R, the Pearson Correlation Coefficient. For the Pearson Correlation both variables should be normally distributed, but this is not a requirement.

The correlation is calculated as:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

m_x and m_y are the means of x and y variables.

If the p-value is < 5%, then the correlation between x and y is significant.

```
#Pearson Correlation Coefficient
cor.test(dailysfo50$PRCP, dailysfo50$TMIN)
```

```
                    Pearson's product-moment correlation

                data:  dailysfo50$PRCP and dailysfo50$TMIN
               t = -5.4696, df = 18606, p-value = 4.567e-08
        alternative hypothesis: true correlation is not equal to 0
                      95 percent confidence interval:
                         -0.05440363 -0.02571329
                            sample estimates:
                                   cor
                               -0.04006672
```

```
cor.test(dailysfo50$PRCP, dailysfo50$TMAX)
```

```
                    Pearson's product-moment correlation

                data:  dailysfo50$PRCP and dailysfo50$TMAX
                t = -29.589, df = 18606, p-value < 2.2e-16
        alternative hypothesis: true correlation is not equal to 0
                      95 percent confidence interval:
                          -0.2256718 -0.1982265
                            sample estimates:
```

```
        cor
 -0.2119909
```

Using all 18,000+ data points, it can be concluded that the correlation between TMIN and PRCP is not significant, but the correlation between TMAX and PRCP is.

The second correlation test in R is the Spearman's Correlation Coefficient. Spearman's correlation coefficient is often denoted rho ρ and measures the monotonic relationship of the variables rather than the linear association in the Pearson setting. Thus, Spearman's correlation coefficient is more reliable with non-linear data compared to Pearson's r. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

It is calculated as:

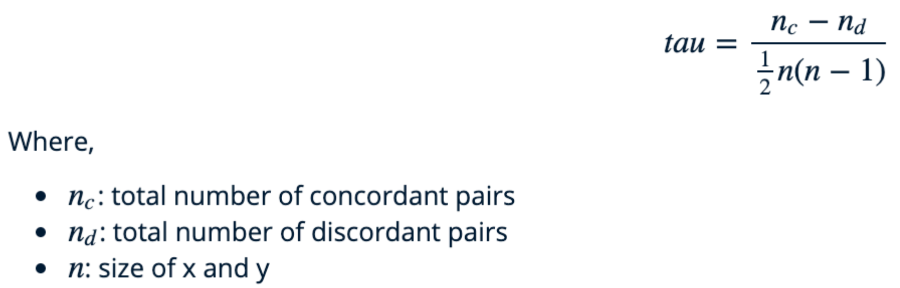
$$rho = \frac{\sum(x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum(x' - m_{x'})^2 \sum(y' - m_{y'})^2}}$$

Where $x' = rank(x_{)}$ and $y' = rank(y)$.

```
#Spearman's Correlation
cor.test(dailysfo50$PRCP, dailysfo50$TMIN, method = "spearman")
                             Spearman's rank correlation rho

                        data:  dailysfo50$PRCP and dailysfo50$TMIN
                             S = 1.2168e+12, p-value < 2.2e-16
                    alternative hypothesis: true rho is not equal to 0
                                    sample estimates:
                                           rho
                                        -0.1331047

cor.test(dailysfo50$PRCP, dailysfo50$TMAX, method = "spearman")
                             Spearman's rank correlation rho

                        data:  dailysfo50$PRCP and dailysfo50$TMAX
                             S = 1.4953e+12, p-value < 2.2e-16
                    alternative hypothesis: true rho is not equal to 0
                                    sample estimates:
                                           rho
                                        -0.3924535
```

The Spearman's correlation between TMIN and Precipitation is -0.133 and between TMAX and Precipitation is -0.392.

The last correlation test in R is the Kendall's Correlation Tau. The Kendall correlation method measures the correspondence between the ranking of x and y variables. The total number of possible pairings of x with y observations is n(n−1)/2n(n−1)/2, where n is the size of x and y. The data does not need to be normally distributed, which we know precipitation is not. It is calculated as:



$$tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Where,

- n_c: total number of concordant pairs
- n_d: total number of discordant pairs
- n: size of x and y

```
#Kendall's Tau
              cor.test(dailysfo50$PRCP, dailysfo50$TMIN, method = "kendall")
                              Kendall's rank correlation tau

                        data:  dailysfo50$PRCP and dailysfo50$TMIN
                              z = -18.06, p-value < 2.2e-16
                    alternative hypothesis: true tau is not equal to 0
                                    sample estimates:
                                           tau
                                       -0.1061326

cor.test(dailysfo50$PRCP, dailysfo50$TMAX, method = "kendall")
                              Kendall's rank correlation tau

                        data:  dailysfo50$PRCP and dailysfo50$TMAX
                              z = -53.573, p-value < 2.2e-16
                    alternative hypothesis: true tau is not equal to 0
                                    sample estimates:
                                           tau
                                       -0.3124775
```

The Kendall’s correlation between TMIN and Precipitation is -0.106 and between TMAX and Precipitation is -0.312.

All together:

```
           PRCP and TMIN   PRCP and TMAX
Pearson's    -0.04006672    -0.2119909
Spearman's   -0.13310470    -0.3924535
Kendall's    -0.10613260    -0.3124775
```

Conclusions

At the beginning of this report, there were a few questions I wanted to attempt to answer from the data analysis. Revisiting these questions:

- Are temperature and precipitation normal?
    - From the analysis, the temperature data is normally distributed. The precipitation data is exponentially distributed.
- What are the different correlation tests that can be performed?
    - There are three main correlation tests, including the Pearson's Correlation Coefficient, Spearman's Rank Correlation Rho, and Kendall's Rank Correlation Tau.
- What is the correlation between temperature and precipitation?
    - The correlation coefficients were calculated as:

```
           PRCP and TMIN   PRCP and TMAX
Pearson's    -0.04006672    -0.2119909
Spearman's   -0.13310470    -0.3924535
Kendall's    -0.10613260    -0.3124775
```

- Which is the best to use in this particular instance?
    - Even though Precipitation is not normally distributed, using the Pearson's Correlation Coefficient would be the best option. This is because Spearman and Kendall's correlation is dependent on the rank of the data, and should not be used when measurements are available.

Overall, this project deepened my understanding of model fitting within R, along with the different correlation tests that could be performed, when determining the correlation between variables.