Weather Data Analysis

Sai Nanda Vihari

2025-02-18

Performing Weather Data Analysis on NYC Flights Departures using R

```
# importing weather dataset using nycflights13 paackage
#import.packages("nycflights13")
library(nycflights13)
library(ggplot2)
head(weather)
```

```
## # A tibble: 6 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 EWR     2013     1     1     1  39.0  26.1  59.4      270      10.4         NA
## 2 EWR     2013     1     1     2  39.0  27.0  61.6      250       8.06        NA
## 3 EWR     2013     1     1     3  39.0  28.0  64.4      240      11.5         NA
## 4 EWR     2013     1     1     4  39.9  28.0  62.2      250      12.7         NA
## 5 EWR     2013     1     1     5  39.0  28.0  64.4      260      12.7         NA
## 6 EWR     2013     1     1     6  37.9  28.0  67.2      240      11.5         NA
## # i 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
## #   time_hour <dttm>
```

```
# Getting Top 5 rows from the dataset
```

```
#View column names
colnames(weather)
```

```
##  [1] "origin"     "year"       "month"      "day"        "hour"
##  [6] "temp"       "dewp"       "humid"      "wind_dir"   "wind_speed"
## [11] "wind_gust"  "precip"     "pressure"   "visib"      "time_hour"
```

```
#checking any null values are present or not
sum(is.na(weather))
```

```
## [1] 23974
```

```r
# removing these null values by omitting them
weather <- na.omit(weather)
```

Univarient analysis

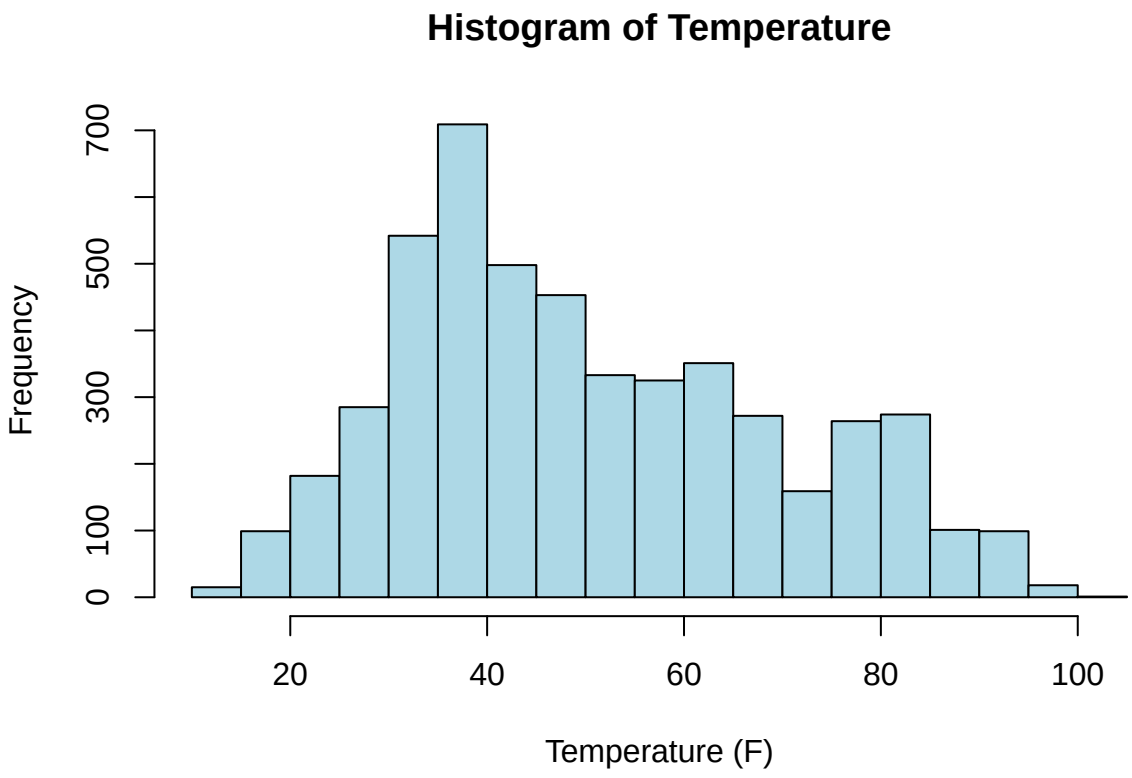
```r
# Summary statistics for Temperature

summary(weather$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.02   35.06   46.94   50.60   64.94  100.04
```

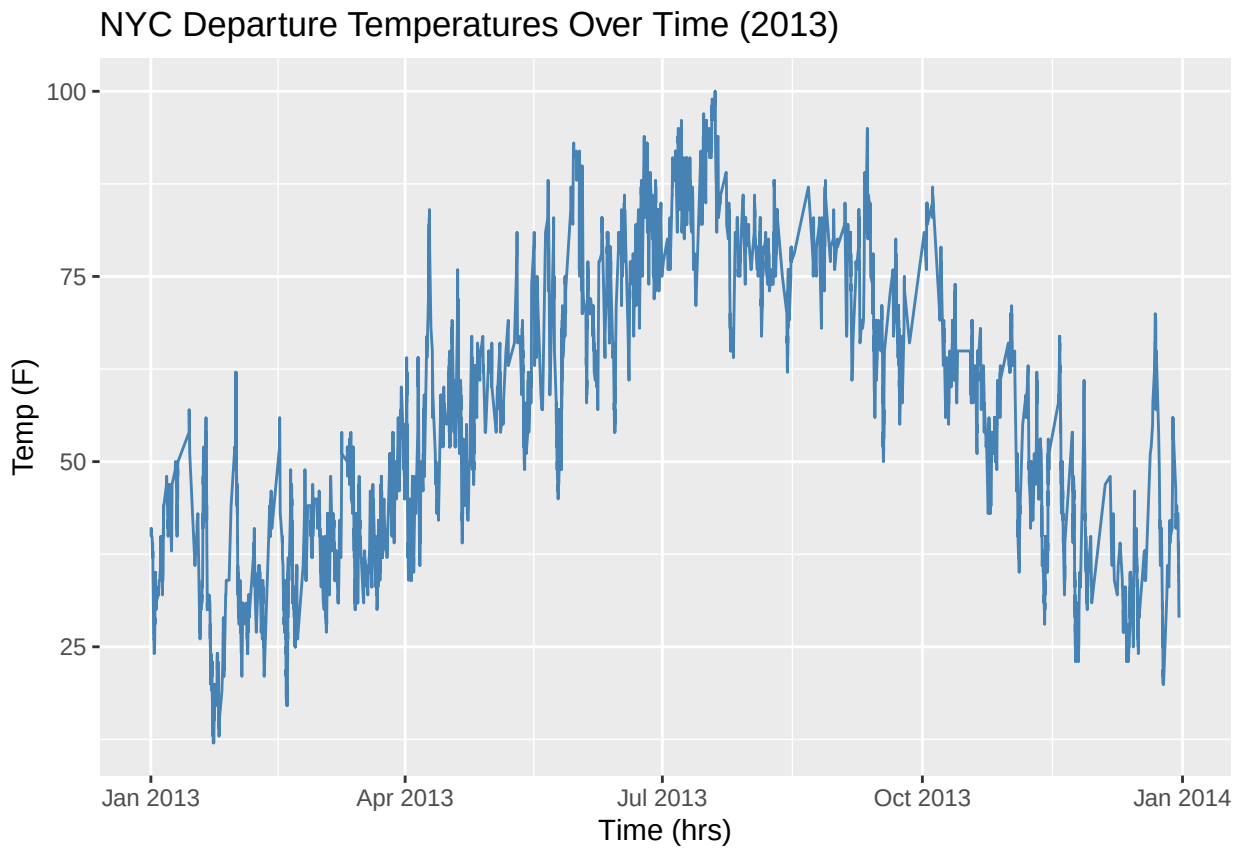
```r
# Summary statistics for Humidity

summary(weather$humid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.95   37.29   45.92   48.70   57.04  100.00
```

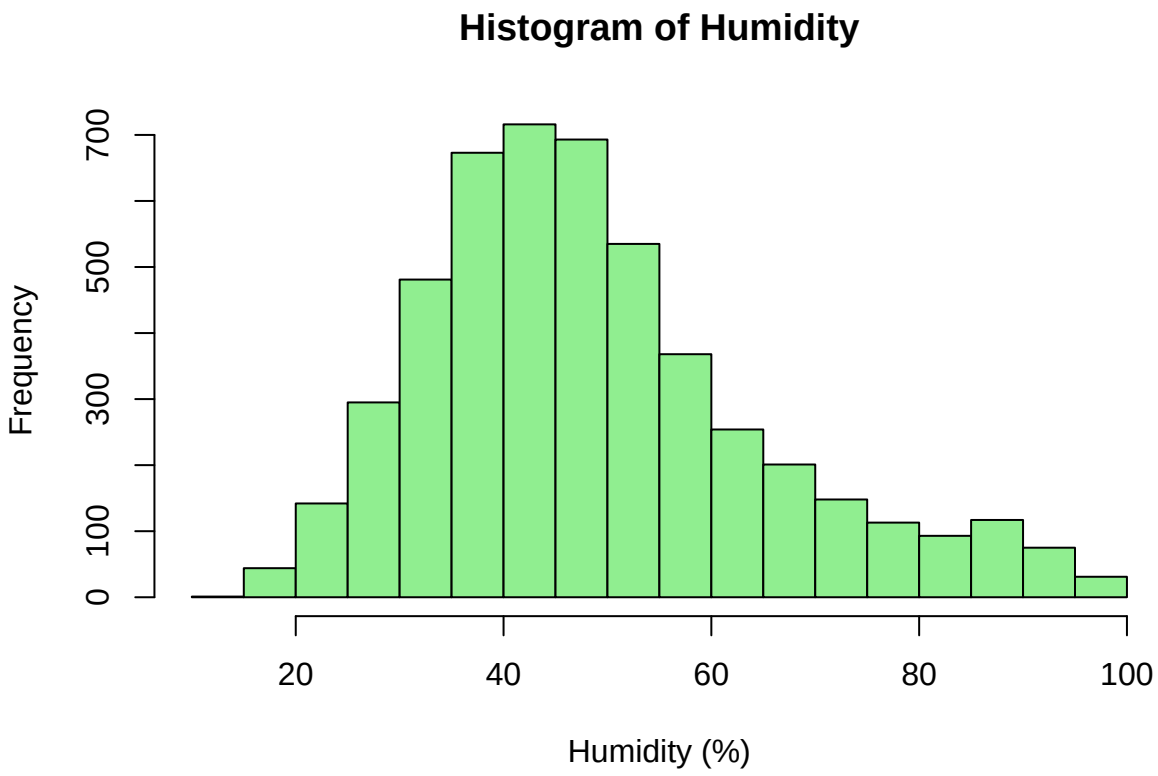
```r
hist(weather$temp,
     main = "Histogram of Temperature",
     xlab = "Temperature (F)",
     col = "lightblue",
     border = "black")
```



```
# Using ggplot2, plotting the line plot of 'temp' as a function of 'time_hour'
ggplot(weather, aes(x = time_hour, y = temp)) + geom_line(color = "steelblue") +
  labs(title = "NYC Departure Temperatures Over Time (2013)",
       x = "Time (hrs)",
       y = "Temp (F)")
```

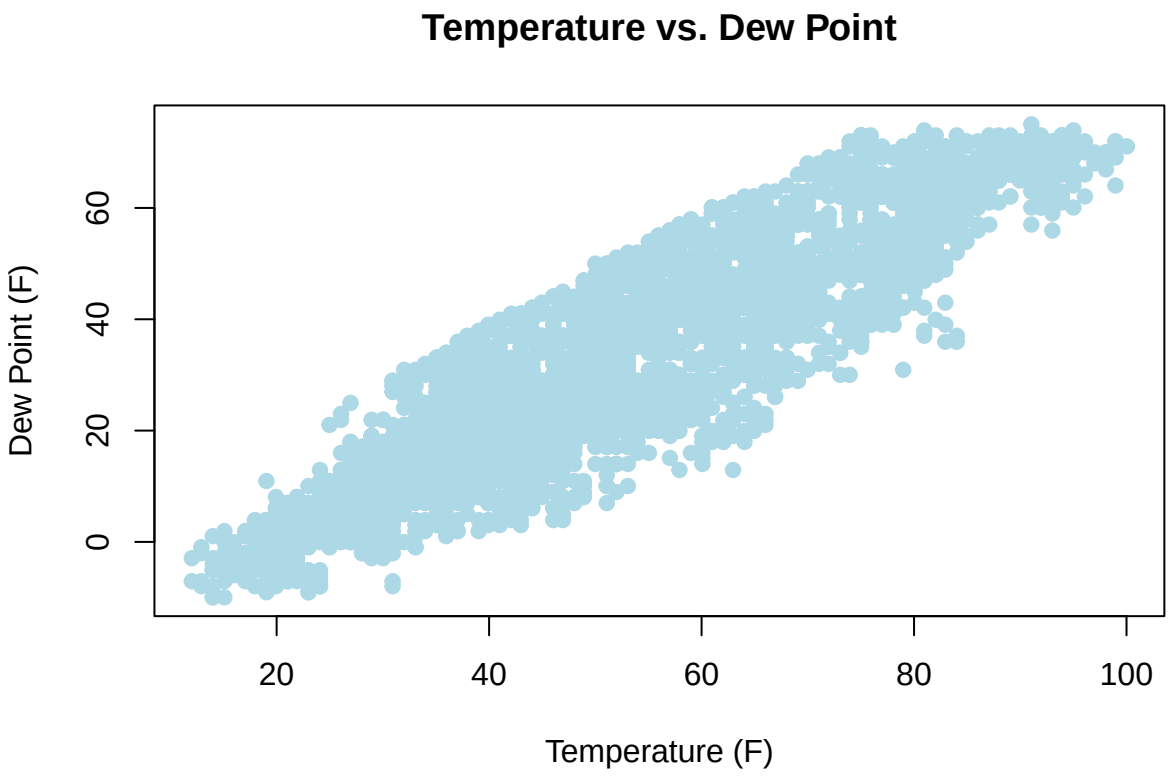



```
hist(weather$humid,
     main = "Histogram of Humidity",
     xlab = "Humidity (%)",
     col = "lightgreen",
     border = "black")
```

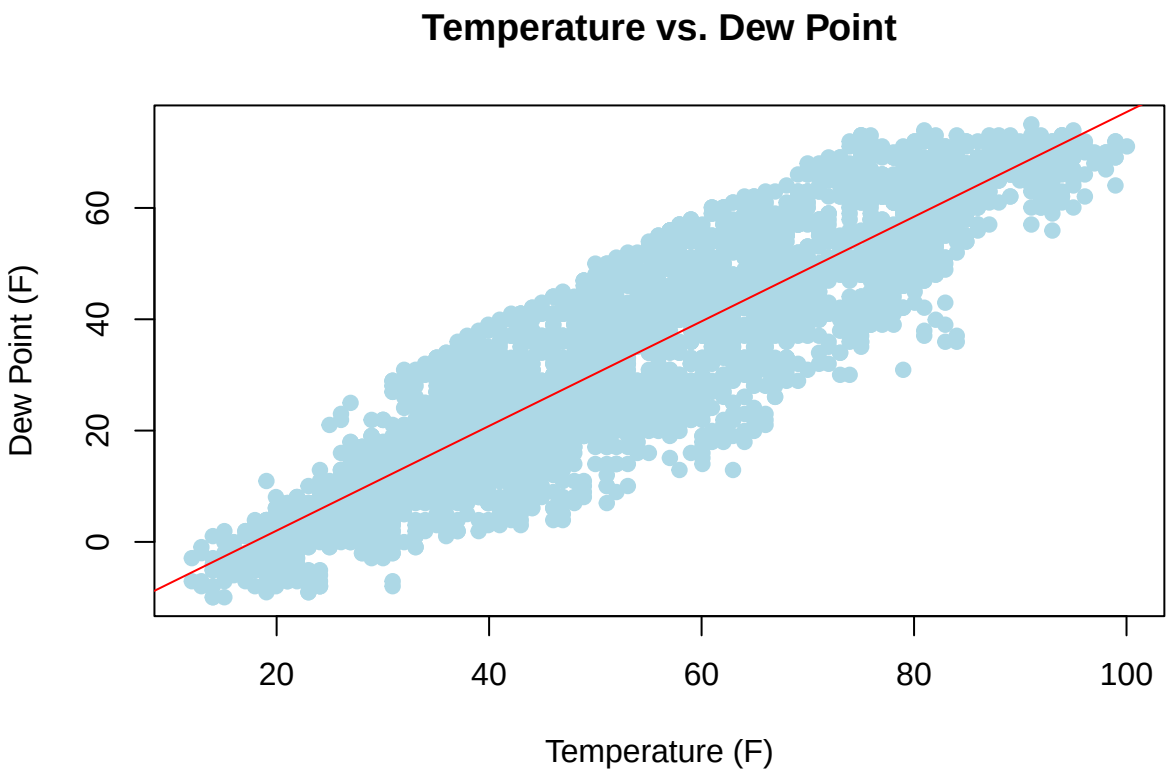


Bivarient Analysis

```
# Scatter plot of Temperature vs Dew Point
plot(weather$temp, weather$dewp,
     main = "Temperature vs. Dew Point",
     xlab = "Temperature (F)",
     ylab = "Dew Point (F)",
     col = "lightblue",
     pch = 19)
```




```r
# Fit and add a regression line to the scatter plot
plot(weather$temp, weather$dewp,
     main = "Temperature vs. Dew Point",
     xlab = "Temperature (F)",
     ylab = "Dew Point (F)",
     col = "lightblue",
     pch = 19)
abline(lm(dewp ~ temp, data = weather), col = "red")
```




```
# correlation between Temperature and Dew Point
correlation <- cor(weather$temp, weather$dewp, use = "complete.obs")
print(paste("Correlation between temperature and dew point:", round(correlation, 2)))
```

```
## [1] "Correlation between temperature and dew point: 0.91"
```

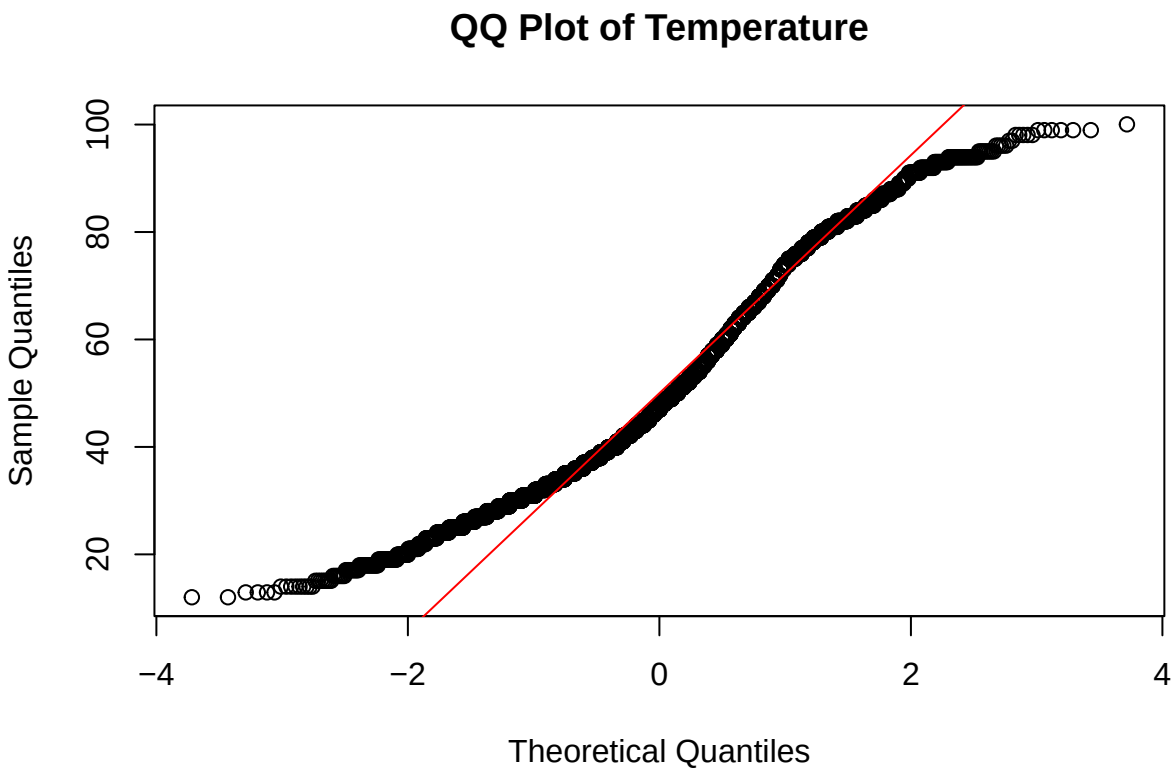
Hypothesis Testing

Null and Alternative Hypothesis:

Null Hypothesis H_0: The true mean temperature of the dataset is 50°F. Alternative Hypothesis H_A: The true mean temperature of the dataset is not equal to 50°F.

Normality Checking:

```
qqnorm(weather$temp, main = "QQ Plot of Temperature")
qqline(weather$temp, col = "red")
```



From the above Q-Q plot, we can say that most of the tail points are deviated from the line .So we are doing shapiro Wiki Test for further checking normality.

```
set.seed(17122000)
temp_sample <- sample(weather$temp, 500)
shapiro_test_result <- shapiro.test(temp_sample)
print(shapiro_test_result)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  temp_sample
## W = 0.95833, p-value = 1.098e-10
```

From the above test , P-value is greater than significant level $\alpha = 0.05$ .So fail to reject null hypothesis . So Normality is approximately satisfied.

Test Statistics:

```
t_test_result <- t.test(weather$temp, mu = 50)
print(t_test_result)
```

```
##
##  One Sample t-test
##
## data:  weather$temp
```

```
## t = 2.232, df = 4979, p-value = 0.02566
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  50.07305 51.12789
## sample estimates:
## mean of x
##  50.60047
```

Decision Making: From above Test Statistics , $P - Value = 0.025$ is less than significant level $\alpha = 0.05$ So Reject Null Hypothesis.

Conclusion: We Can Conclude that The true mean temperature of the weather dataset is not equal to 50°F.