# Seattle Rain Analysis

Josh Houlding

2023-08-13

# The objective

I found this neat dataset on Kaggle called Did it rain in Seattle? (1948-2017) (https://www.kaggle.com/datasets/rtatman/did-it-rain-in-seattle-19482017), comprised of weather data from sensor(s) at SeaTac International Airport near Seattle, Washington, USA from January 1st, 1948 to December 12th, 2017. I was very excited to delve into the data and see what insights came out of it.

# The data

The dataset contains a single CSV file with data on amounts of precipitation (inches) and the maximum and minimum recorded temperatures for each day from 1948-01-01 to 2017-12-12, which is over 25,000 days' worth of data. It also has a true/false column for whether it rained or not on a specific day.

# Limitations of the data

- The data is only concerned with rain, so there is no data on sleet, hail, snow, or other types of precipitation.
- The data does not include measurements such as humidity, wind speed, wind direction, or atmospheric pressure, all of which would be useful for a deeper analysis.

# Loading the data

We start by loading all the necessary R packages.

```
library(tidyverse)
library(dplyr)
library(sqldf)
library(ggplot2)
library(readr)
library(lubridate)
library(knitr)
```

Then, load the dataset and take a brief glance:

```
rainData <- read_csv("seattleWeather_1948-2017.csv")
```

```
## Rows: 25551 Columns: 5
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## dbl  (3): PRCP, TMAX, TMIN
## lgl  (1): RAIN
## date (1): DATE
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
kable(head(rainData), caption = "rainData")
```

rainData

| DATE | PRCP | TMAX | TMIN | RAIN |
|---|---|---|---|---|
| 1948-01-01 | 0.47 | 51 | 42 | TRUE |
| 1948-01-02 | 0.59 | 45 | 36 | TRUE |
| 1948-01-03 | 0.42 | 45 | 35 | TRUE |
| 1948-01-04 | 0.31 | 45 | 34 | TRUE |
| 1948-01-05 | 0.17 | 45 | 32 | TRUE |

| DATE | PRCP | TMAX | TMIN | RAIN |
|---|---|---|---|---|
| 1948-01-06 | 0.44 | 48 | 39 | TRUE |

```
# Display number of entries in the dataframe
nrow(rainData)
```

```
## [1] 25551
```

# Cleaning the data

## Removing entries with missing values

To ensure the analysis was as accurate as possible, I decided to remove entries from the data that had missing values.

```
# Analyze entries with missing values
entriesWithMissingValues <- rainData %>%
  filter(rowSums(is.na(.)) > 0)
kable(head(entriesWithMissingValues), caption = "Entries with missing values")
```

Entries with missing values

| DATE | PRCP | TMAX | TMIN | RAIN |
|---|---|---|---|---|
| 1998-06-02 | NA | 72 | 52 | NA |
| 1998-06-03 | NA | 66 | 51 | NA |
| 2005-09-05 | NA | 70 | 52 | NA |

```
# Show how many entries have missing values
nrow(entriesWithMissingValues)
```

```
## [1] 3
```

```
# Remove missing values from the dataframe
rainData <- na.omit(rainData)
nrow(rainData)
```

```
## [1] 25548
```

The entries for June 2nd, 1998, June 3rd, 1998, and September 5th, 2005 were all removed because they were missing precipitation values. This was likely due to an error with the sensor(s) that prevented them from getting a proper reading. As such, these entries were not very relevant to my analysis.

## Making titles of the dataframe lowercase

I planned to use the `sqldf` R library to run SQL queries directly in R, and SQL commands are usually in all capital letters. Thus, I believed it best to change all column names within `rainData` to lowercase so the SQL queries later on would be more readable.

```
colnames(rainData) <- tolower(colnames(rainData))
colnames(rainData)
```

```
## [1] "date" "prcp" "tmax" "tmin" "rain"
```

## Removing duplicate entries

Removing duplicates is a common practice in the data cleaning process, and improves the accuracy of the data.

```
# Number of entries before removing duplicates
nrow(rainData)
```

```
## [1] 25548
```

```
rainData <- distinct(rainData)
# Number of entries after removing duplicates
nrow(rainData)
```

```
## [1] 25548
```

Clearly there were no duplicate entries, since removing duplicates did not change the length of the dataframe. Thus, everything seems to be good here.

# Correcting inconsistent or erroneous values

I decided to find the max and min values for `prcp`, `tmax`, and `tmin` to see if there were any extreme values that did not make sense in the context of the data.

```
# Find the max and min prcp (in), tmax (°F), and tmin (°F)
max_values <- sqldf("SELECT MAX(prcp) AS max_prcp, MIN(prcp) AS min_prcp, MAX(tmax) AS max_tmax, MIN(tmax) AS min_tmax, MAX
(tmin) AS max_tmin, MIN(tmin) AS min_tmin FROM rainData")
kable(head(max_values),align="l")
```

| max_prcp | min_prcp | max_tmax | min_tmax | max_tmin | min_tmin |
|---|---|---|---|---|---|
| 5.02 | 0 | 103 | 4 | 71 | 0 |

The maximum precipitation in a day was just over 5 inches, which could indicate a flash flood, but is within the realm of possibility. The minimum was 0 inches, which makes sense, and a maximum high of 103°, minimum high of 4°, maximum low of 71° and minimum low of 0° are all sensible as well. One concern I had was over the lowest-ever minimum temperature being 0°. I am not sure if this was actually because the lowest-ever-recorded temperature at SeaTac was 0°F, or if this indicates a limitation of the dataset. Other than that, though, the numbers all seem reasonable, so there are no erroneous values that need to be corrected.

# Standardizing formats and data type conversion

I made sure the data types of the columns were in a sensible format.

```
str(rainData)
```

```
## tibble [25,548 × 5] (S3: tbl_df/tbl/data.frame)
##  $ date: Date[1:25548], format: "1948-01-01" "1948-01-02" ...
##  $ prcp: num [1:25548] 0.47 0.59 0.42 0.31 0.17 0.44 0.41 0.04 0.12 0.74 ...
##  $ tmax: num [1:25548] 51 45 45 45 45 48 50 48 50 43 ...
##  $ tmin: num [1:25548] 42 36 35 34 32 39 40 35 31 34 ...
##  $ rain: logi [1:25548] TRUE TRUE TRUE TRUE TRUE TRUE ...
##  - attr(*, "na.action")= 'omit' Named int [1:3] 18416 18417 21068
##   ..- attr(*, "names")= chr [1:3] "18416" "18417" "21068"
```

Everything in the `rainData` dataset is already in a sensible format, so no changes need to be made.

# The analysis

My analysis consisted of several distinct questions I wanted to find the answers to:

- What percentage of days did it rain in Seattle?
- What was the average precipitation, average maximum temperature and average minimum temperature during each month?
- How have average precipitation levels changed over time?
- What were the wettest and driest years on record?
- What were the hottest and coldest years on record?
- Is there a correlation between temperature and precipitation level?
- Which day of the week does it rain the most?
- Is there a correlation between average wind speed and precipitation level?
- Which wind direction is associated with the most precipitation?

## What percentage of days did it rain in Seattle?

```
# Find % of days that were rainy
num_days <- nrow(rainData)
num_rainy_days <- nrow(sqldf("SELECT * FROM rainData WHERE RAIN == TRUE"))
percent_rainy_days <- (num_rainy_days / num_days) * 100
print(percent_rainy_days)
```

```
## [1] 42.66479
```

About 42.66% of Seattle days were rainy. Contrary to popular belief, it is most certainly *not* always raining in Seattle!

# What was the average precipitation, average maximum temperature and average minimum temperature during each month?

```
# Add new columns for year and month
rainData <- mutate(rainData, year = substr(date, 1, 4))
rainData <- mutate(rainData, month = substr(date, 6, 7))
# Reorder columns so year and month are next to date
rainData <- rainData[, c(1,6,7,2,3,4,5)]
head(rainData, 3)
```

```
## # A tibble: 3 × 7
##   date       year  month  prcp  tmax  tmin rain
##   <date>     <chr> <chr> <dbl> <dbl> <dbl> <lgl>
## 1 1948-01-01 1948  01     0.47    51    42 TRUE
## 2 1948-01-02 1948  01     0.59    45    36 TRUE
## 3 1948-01-03 1948  01     0.42    45    35 TRUE
```

```
# Find the average values for each month
  averageByMonth <- rainData %>%
    group_by(month) %>%
    summarize(average_prcp = mean(prcp, na.rm = TRUE), average_tmax = mean(tmax, na.rm = TRUE), average_tmin = mean(tmin, n
a.rm = TRUE))
# Reorder the columns for readability
averageByMonth <- mutate(averageByMonth, month_name = c("January", "February", "March", "April", "May", "June", "July", "Aug
ust", "September", "October", "November", "December"))
averageByMonth <- averageByMonth[, c(1,5,2,3,4)]
kable(head(averageByMonth, 12), caption = "Average values for each month")
```

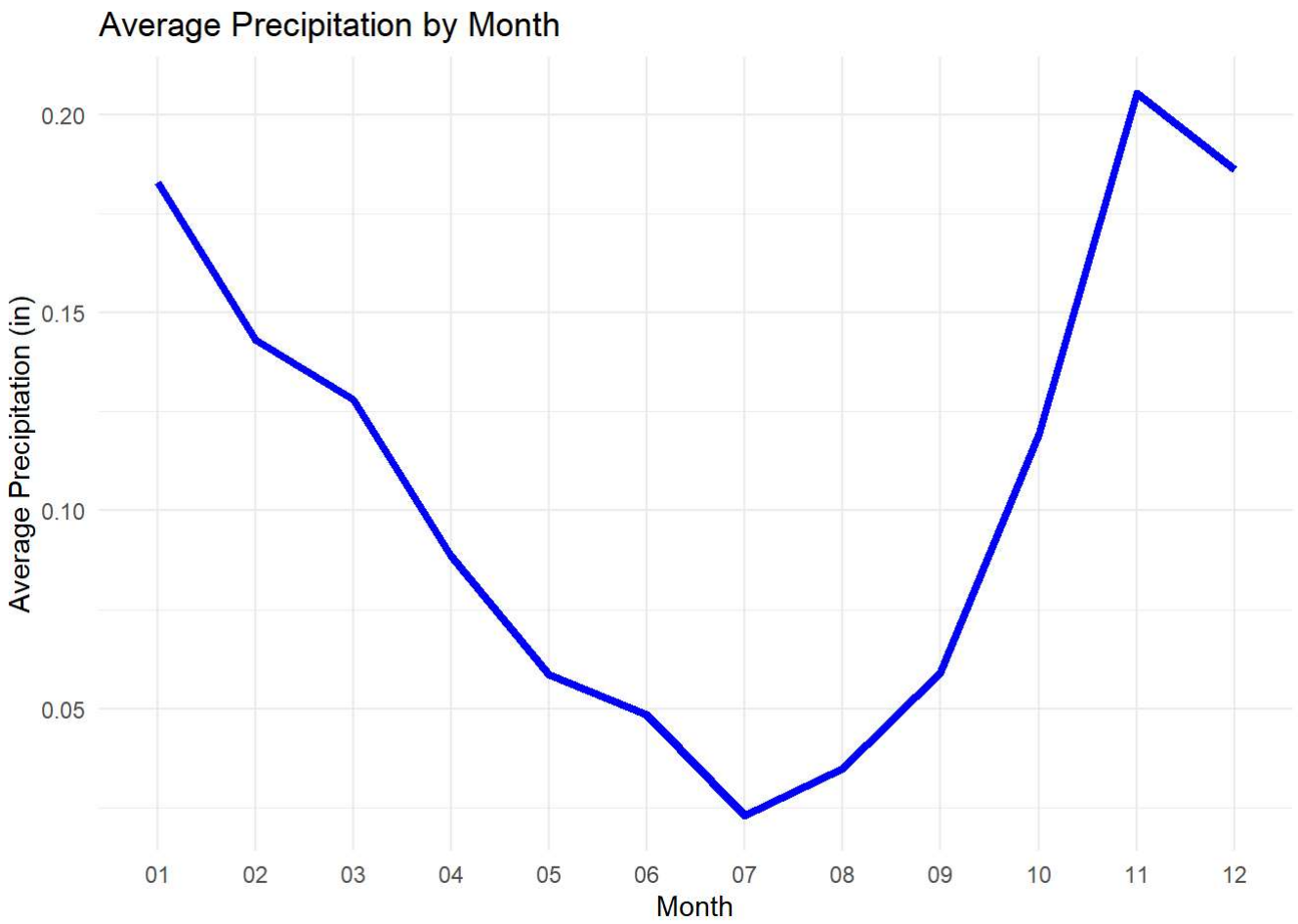
Average values for each month

| month | month_name | average_prcp | average_tmax | average_tmin |
|---|---|---|---|---|
| 01 | January | 0.1830553 | 45.13180 | 35.20876 |
| 02 | February | 0.1430233 | 48.99494 | 36.78766 |
| 03 | March | 0.1279631 | 52.32166 | 38.39355 |

| month | month_name | average_prcp | average_tmax | average_tmin |
|---|---|---|---|---|
| 04 | April | 0.0885048 | 57.52190 | 41.46095 |
| 05 | May | 0.0585023 | 64.30922 | 46.73318 |
| 06 | June | 0.0485224 | 69.65396 | 51.62869 |
| 07 | July | 0.0231060 | 75.49770 | 55.00553 |
| 08 | August | 0.0349677 | 75.23318 | 55.31751 |
| 09 | September | 0.0590567 | 69.63173 | 51.75417 |
| 10 | October | 0.1193502 | 59.45253 | 45.66728 |
| 11 | November | 0.2055238 | 50.58762 | 39.72762 |
| 12 | December | 0.1861403 | 45.44821 | 35.94519 |

Visualized in a line graph, we have the following:

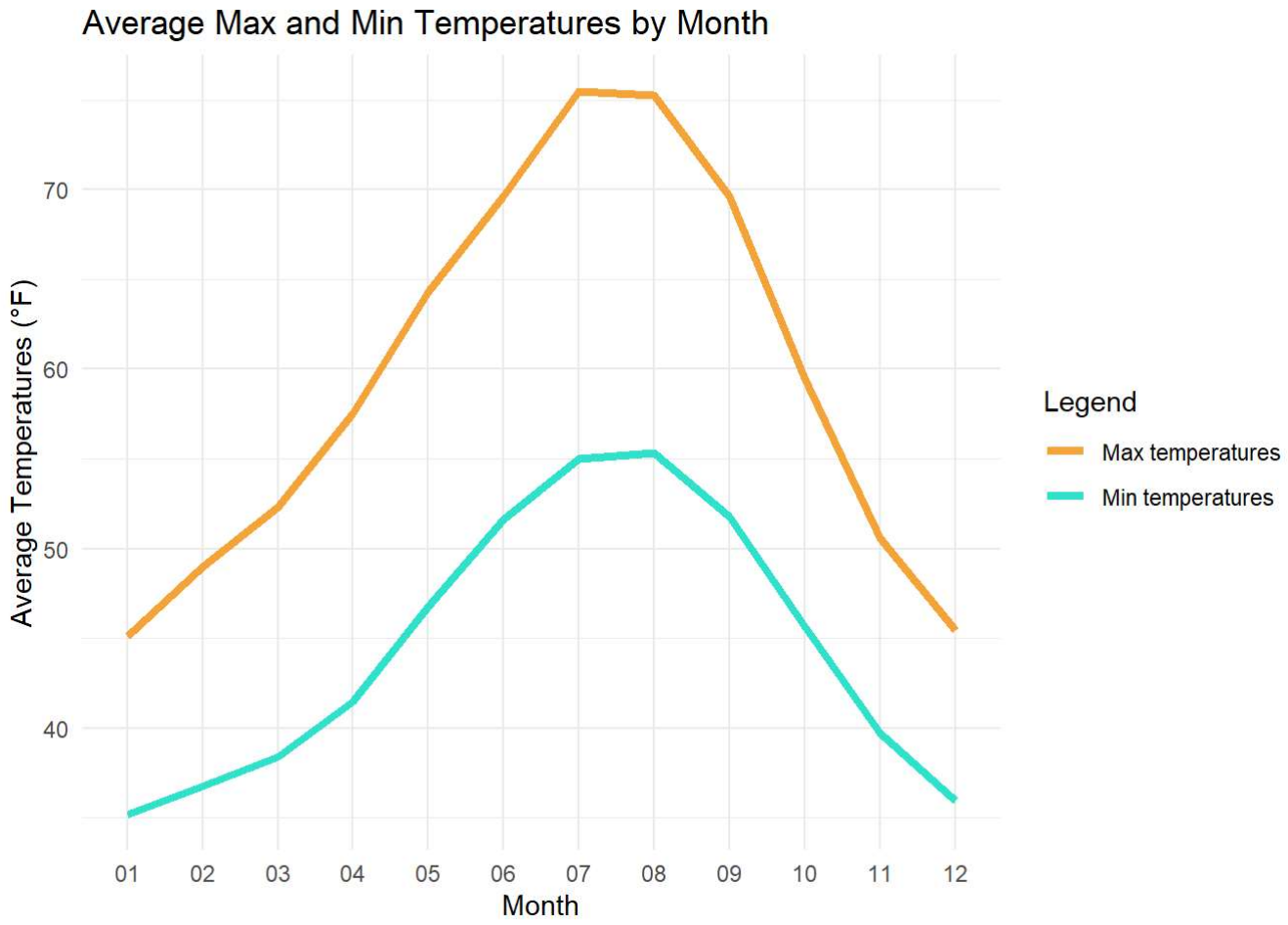
```
# Graph average precipitation level for each month
ggplot(data = averageByMonth) +
  geom_line(aes(x = month, y = average_prcp), linewidth = 1.5, color = "blue", linetype = "solid", group=1) +
  labs(title = "Average Precipitation by Month",
       x = "Month",
       y = "Average Precipitation (in)") +
  theme_minimal()
```




As expected, average monthly precipitation is high at the start and end of the year during winter, early spring and late fall, and very low during the late spring and summer months. Precipitation hits its maximum in November and minimum in July.

Looking at the average maximum and minimum temperatures, we see a similar pattern:

```
  # Graph average max and min temperatures for each month
  ggplot(data = averageByMonth) +
    geom_line(aes(x = month, y = average_tmax, color = "Max temperatures"), linewidth = 1.5, linetype = "solid", group = 1)
+
    geom_line(aes(x = month, y = average_tmin, color = "Min temperatures"), linewidth = 1.5, linetype = "solid", group = 2)
+
    labs(title = "Average Max and Min Temperatures by Month",
         x = "Month",
         y = "Average Temperatures (°F)") +
    scale_color_manual(name = "Legend", values = c("Max temperatures" = "#F4A438", "Min temperatures" = "#30E1CB")) +
    theme_minimal()
```



July had the highest average maximum temperature, and August had the highest average minimum temperature.

## How have average precipitation levels changed over time? What were the wettest and driest years on record?

First, I found the average precipitation levels for each year. Then, I looked for the wettest and driest years on record.

```
# Find average values for each year
  averageByYear <- rainData %>%
    group_by(year) %>%
    summarize(average_prcp = mean(prcp, na.rm = TRUE), average_tmax = mean(tmax, na.rm = TRUE), average_tmin = mean(tmin, n
a.rm = TRUE))
kable(head(averageByYear), caption="Averages for each year")
```

Averages for each year

| year | average_prcp | average_tmax | average_tmin |
|---|---|---|---|
| 1948 | 0.1251093 | 57.01366 | 41.19672 |
| 1949 | 0.0889315 | 59.14795 | 41.39178 |
| 1950 | 0.1510685 | 57.03562 | 41.00000 |
| 1951 | 0.1104110 | 58.54521 | 41.05205 |
| 1952 | 0.0649727 | 58.74317 | 41.46721 |
| 1953 | 0.1353973 | 58.44384 | 43.12603 |

```
# Find driest and wettest year on record
wettestYear <- sqldf("SELECT year, MAX(average_prcp) AS max_average_prcp FROM averageByYear")
  kable(head(wettestYear), align="l")
```

| year | max_average_prcp |
|---|---|
| 1950 | 0.1510685 |

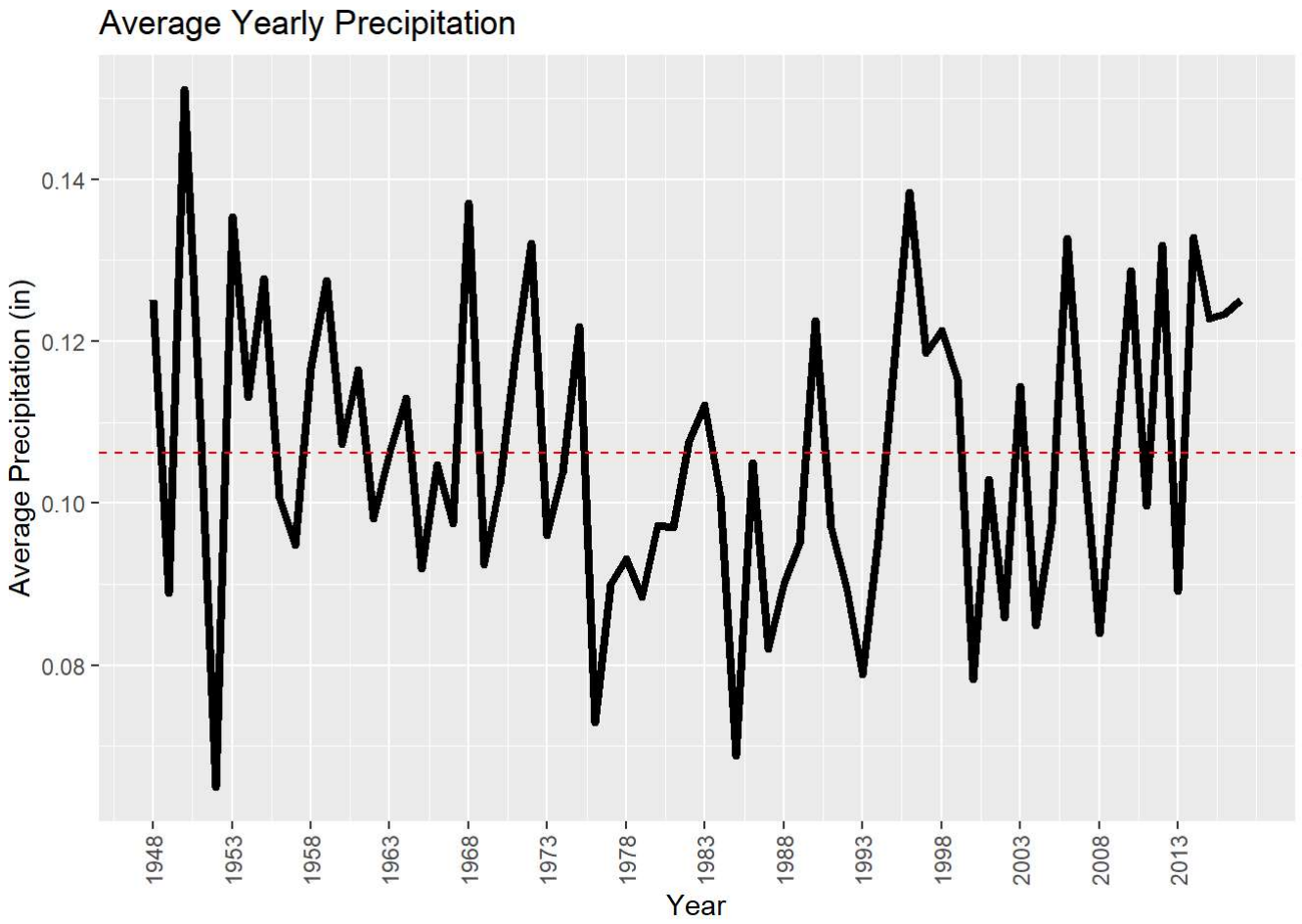
```
driestYear <- sqldf("SELECT year, MIN(average_prcp) AS min_average_prcp FROM averageByYear")
  kable(head(driestYear), align="l")
```

| year | min_average_prcp |
|---|---|
| 1952 | 0.0649727 |

We can see that the wettest year on record was 1950, with an average rainfall of ~0.065 inches/day, while 1952 was the driest with an average rainfall of ~0.151 inches/day.

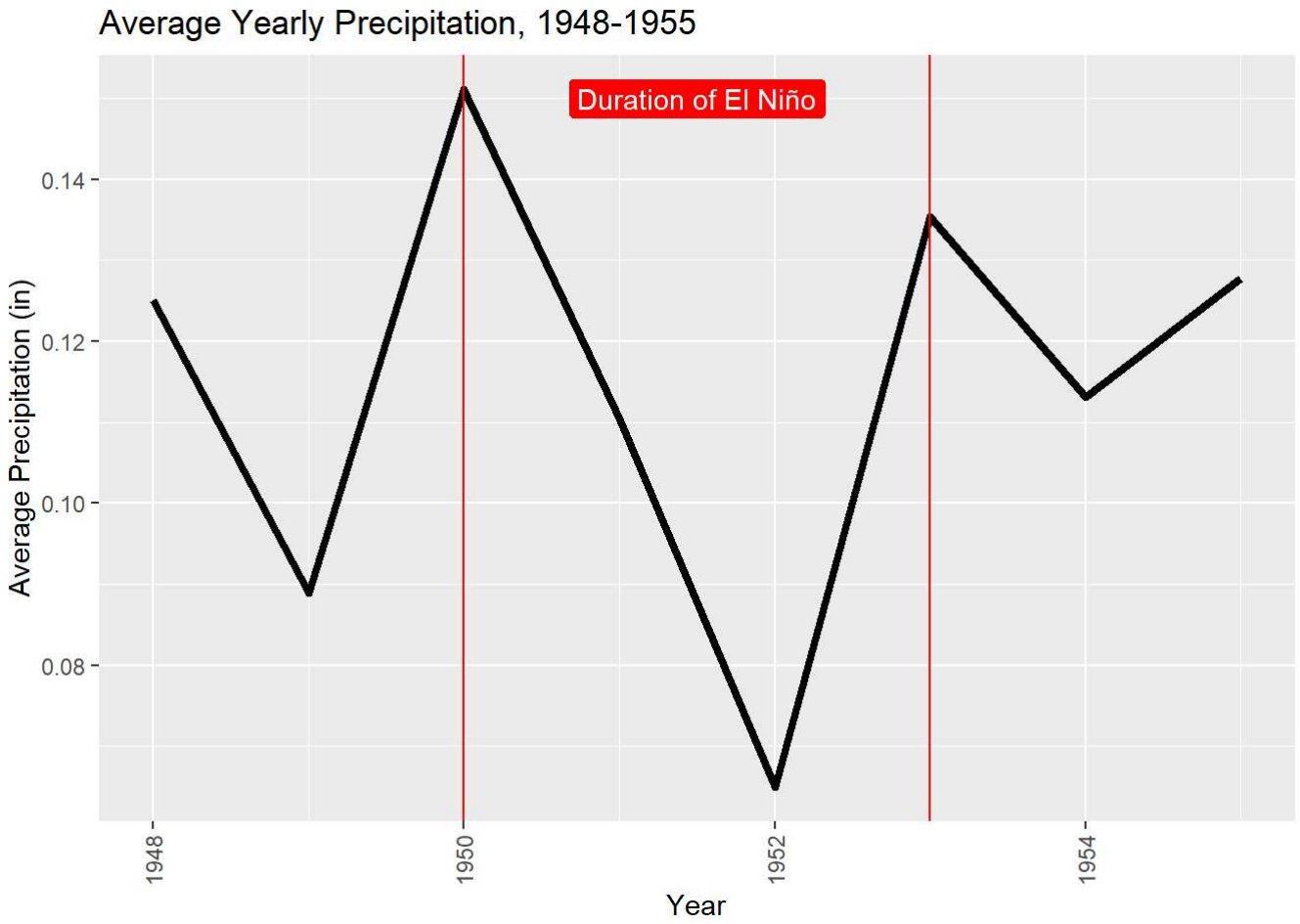
Graphing the average precipitation levels for each year, we get this:

```
# Graph average precipitation for each year
    ggplot(data = averageByYear) +
    geom_line(aes(x = 1948:2017, y = average_prcp), linewidth = 1.5, color = "black", linetype = "solid", group=1) +
    geom_hline(yintercept=sqldf("SELECT AVG(average_prcp) FROM averageByYear")[1,1], linetype="dashed",color="red") +
    labs(title = "Average Yearly Precipitation",
         x = "Year",
         y = "Average Precipitation (in)") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
    scale_x_continuous(breaks = seq(min(averageByYear$year), max(averageByYear$year), by = 5))
```



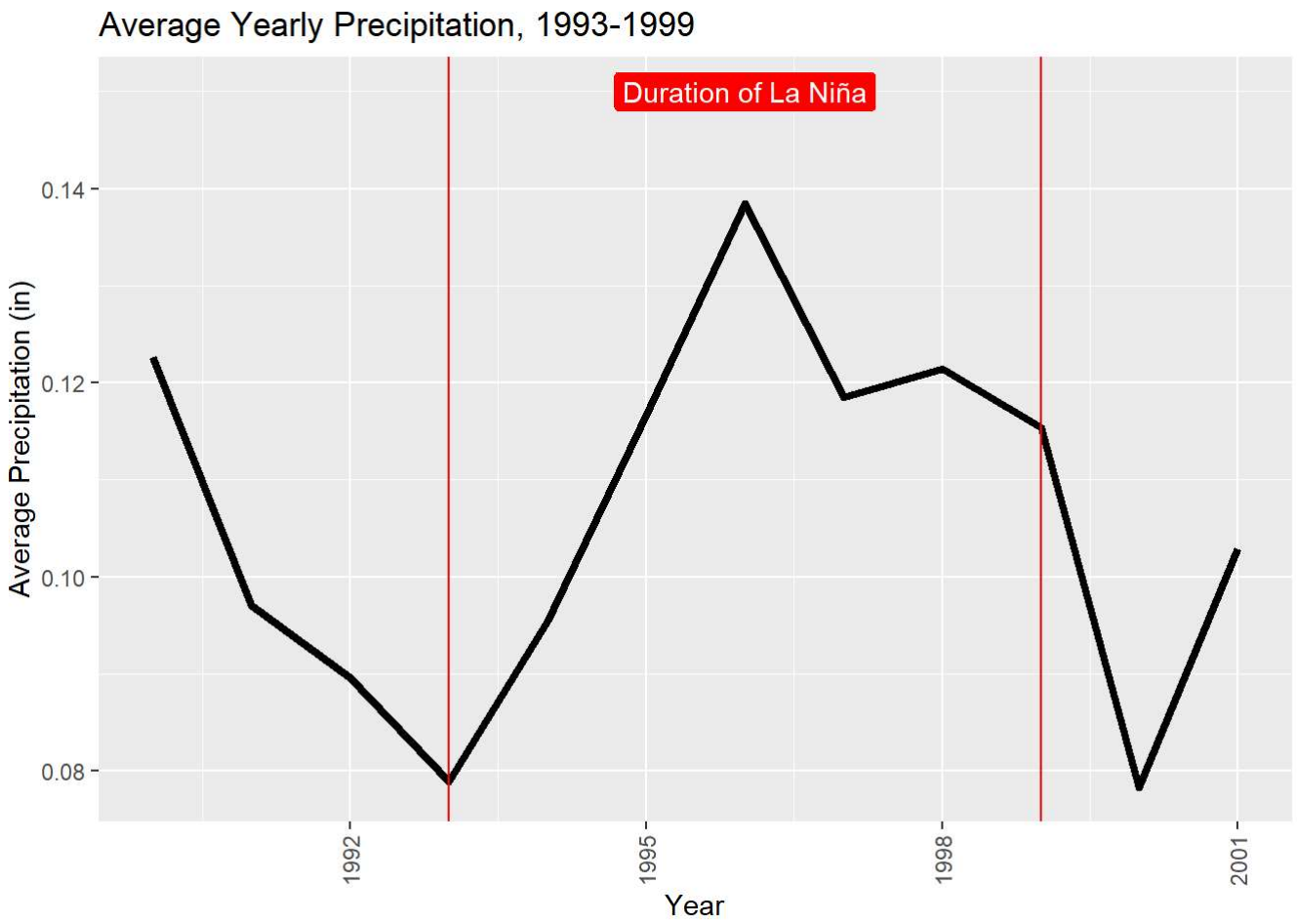

Clearly, the largest drop in average precipitation levels occurs from about 1950 to 1953, and a large, sustained period of high average precipitation levels occurs from about 1993 to 1999. Thus, I graphed the data from 1948 to 1955 and 1990 to 2001 to examine further.

```
# Graph average precipitation from 1948 to 1955
  averageByYear48to55 <- head(averageByYear, 8)
  ggplot(data = averageByYear48to55) +
    geom_line(aes(x = 1948:1955, y = average_prcp), linewidth = 1.5, color = "black", linetype = "solid") +
    labs(title = "Average Yearly Precipitation, 1948-1955",
         x = "Year",
         y = "Average Precipitation (in)") +
    geom_vline(xintercept = c(1950,1953), color = "red", linetype = "solid") +
    annotate(geom="label", 1951.5, 0.15, label="Duration of El Niño", fill="red", color="white") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



Looking at the graph, we can see a massive drop in average precipitation between 1950 and 1953. After some research, I found out that a cycle of the El Niño climate pattern started in 1950, leading to very dry winters.

```
# Graph average precipitation from 1990 to 2001
  averageByYear90to01 <- sqldf("SELECT * FROM averageByYear WHERE CAST(year AS REAL) >= 1990 AND CAST(year AS REAL) <= 200
1")
  ggplot(data = averageByYear90to01) +
    geom_line(aes(x = 1990:2001, y = average_prcp), linewidth = 1.5, color = "black", linetype = "solid") +
    labs(title = "Average Yearly Precipitation, 1993-1999",
        x = "Year",
        y = "Average Precipitation (in)") +
    geom_vline(xintercept = c(1993,1999), color = "red", linetype = "solid") +
    annotate(geom="label", 1996, 0.15, label="Duration of La Niña", fill="red", color="white") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```




Average precipitation increases significantly during the period of 1993-1999, suggesting a La Niña cycle.

## What were the hottest and coldest years on record?

I calculated the hottest year on record by taking the maximum sum of average maximum and minimum temperatures, and the coldest year on record by taking the minimum sum. I ended up with the following:

```
# Find hottest and coldest years on record
hottestYear <- sqldf("SELECT year, average_tmax, average_tmin, SUM(average_tmax + average_tmin) AS temp_sum FROM averageByYe
ar GROUP BY year ORDER BY temp_sum DESC LIMIT 1")
kable(hottestYear, align="l")
```

| year | average_tmax | average_tmin | temp_sum |
|---|---|---|---|
| 2015 | 63.36986 | 47.90137 | 111.2712 |

```
coldestYear <- sqldf("SELECT year, average_tmax, average_tmin, SUM(average_tmax + average_tmin) AS temp_sum FROM averageByYe
ar GROUP BY year ORDER BY temp_sum ASC LIMIT 1")
kable(coldestYear, align="l")
```

| year | average_tmax | average_tmin | temp_sum |
|---|---|---|---|
| 1955 | 55.32329 | 40.65205 | 95.97534 |

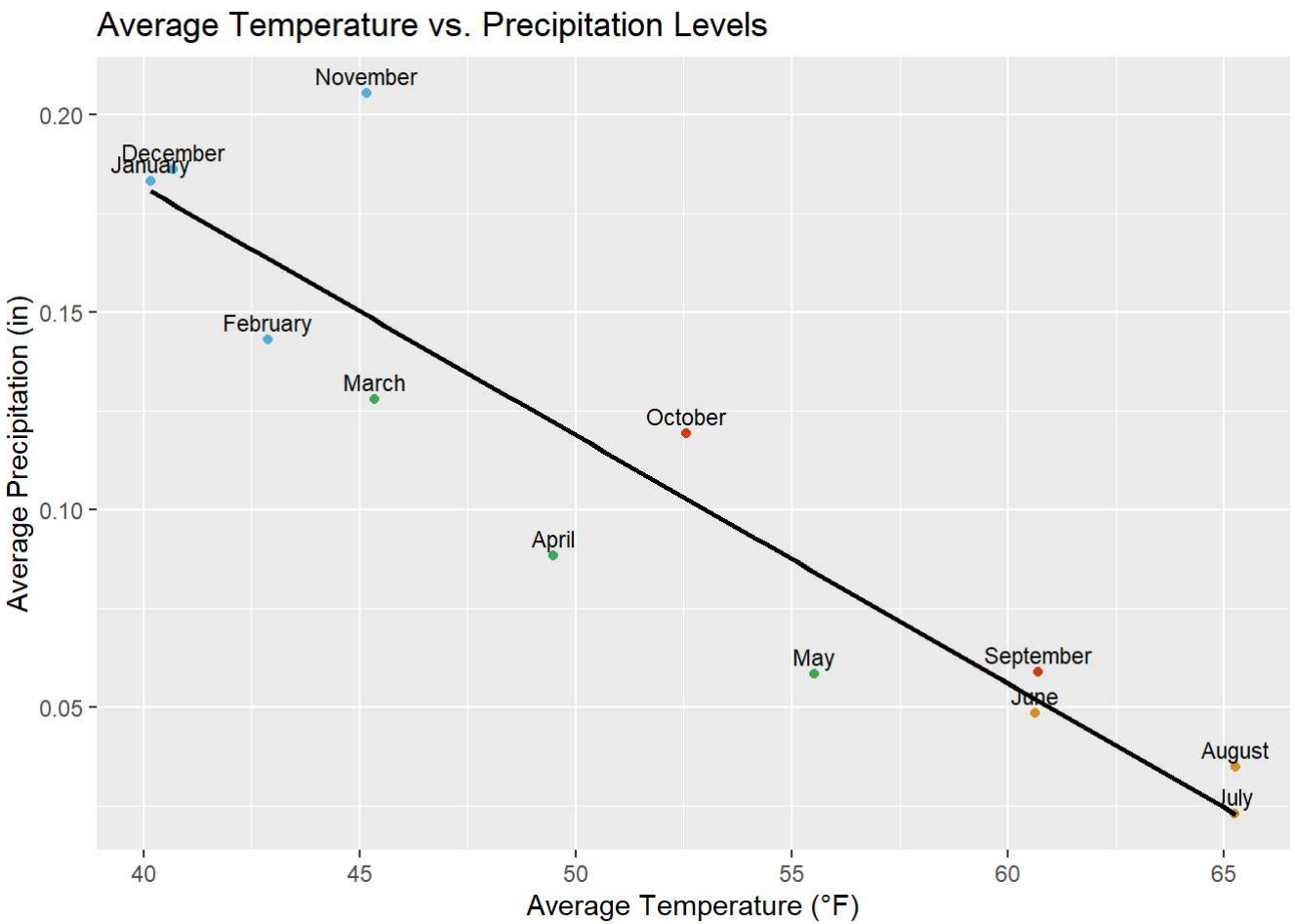
The hottest year on record was 2015, with a temperature sum of ~111.27 degrees, and the coldest year was 1955, with a temperature sum of ~95.98 degrees.

## Is there a correlation between temperature and precipitation level?

I wanted to find out if the average temperature of a month bore any relationship to the average precipitation levels seen that month. My prediction was that a strong correlation exists between the two.

```
# Add new column for average temperature
  averageByMonth <- mutate(averageByMonth, average_temp = (average_tmax + average_tmin)/2)

# Add month_name column and rearrange for readability
    months <- c("January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "Novembe
r", "December")
    averageByMonth <- mutate(averageByMonth, month_name = months)

# Add colors to each month's point for visual appeal
    month_colors <- c("#59acd9", "#59acd9", "#3ea852", "#3ea852", "#3ea852", "#d68f24", "#d68f24", "#d68f24", "#cc3f10", "#c
c3f10","#59acd9","#59acd9")

# Generate the plot
    ggplot(averageByMonth, aes(x = average_temp, y = average_prcp, label = month_name)) +
      geom_point(aes(color = month_colors)) +
      geom_smooth(method = "lm", se = FALSE, color="black") +
      scale_color_identity() +
      geom_text(aes(vjust = -0.5), size = 3) +
      labs(title = "Average Temperature vs. Precipitation Levels", x = "Average Temperature (°F)", y = "Average Precipitatio
n (in)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




I then calculated the regression coefficient, $r^2$, to see how strong the correlation is.

```
# Find the regression coefficient
month_regression <- lm(average_prcp ~ average_temp, data = averageByMonth)
r_squared <- summary(month_regression)$r.squared
print(r_squared)
```

```
## [1] 0.856791
```

As we can see, there is an ~85.7% correlation between average temperature and average rainfall. This is very strong, which confirmed my initial hypothesis.

# Which day of the week does it rain the most?

```
# Add a column to rainData for day of the week
  install.packages("lubridate")
```

```
## Warning: package 'lubridate' is in use and will not be installed
```

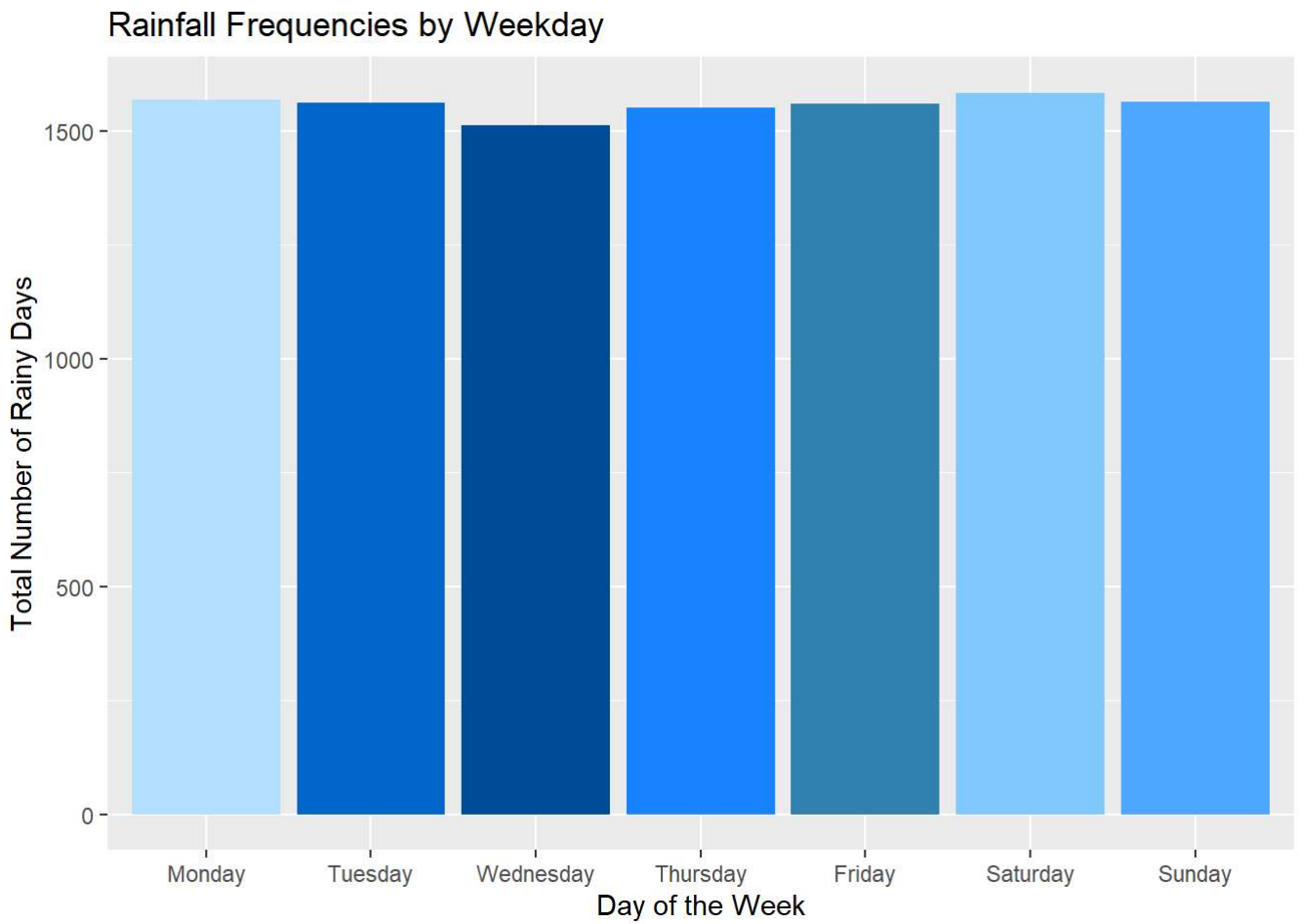
```
  library(lubridate)
  rainData <- mutate(rainData, day_of_week = weekdays(date))

  # Ensure each day occurs with roughly the same frequency (error checking)
  day_counts <- rainData %>%
    group_by(day_of_week = weekdays(date)) %>%
    summarise(count = n())

  # Filter rainy days
  rainy_days <- rainData %>%
    filter(rain == TRUE)

  # Count the number of rainy days for each day of the week
  rainy_day_counts <- rainy_days %>%
    group_by(day_of_week = weekdays(date)) %>%
    summarise(rainy_count = n())

  # Order days of the week chronologically and display the table
  day_num <- c(5,1,6,7,4,2,3)
  rainy_day_counts <- rainy_day_counts %>%
  mutate(day_num = day_num) %>%
  arrange(day_num)
  rainy_day_counts <- rainy_day_counts[, c(1,3,2)]
  kable(head(rainy_day_counts,12), caption="Rainy day counts")
```

Rainy day counts

| day_of_week | day_num | rainy_count |
|---|---|---|
| Monday | 1 | 1568 |

| day_of_week | day_num | rainy_count |
|---|---|---|
| Tuesday | 2 | 1561 |
| Wednesday | 3 | 1512 |
| Thursday | 4 | 1552 |
| Friday | 5 | 1559 |
| Saturday | 6 | 1584 |
| Sunday | 7 | 1564 |

```
  # Create a bar chart of rainy day counts using ggplot2
  ggplot(rainy_day_counts, aes(x = factor(day_of_week, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "S
aturday", "Sunday")), y = rainy_count, fill=day_of_week)) +
    geom_bar(stat = "identity") +
    scale_fill_manual(values = c("#3280ad", "#B3E0FF", "#80C9FF", "#4DA6FF", "#1A83FF", "#0066CC", "#004C99")) +
    labs(title = "Rainfall Frequencies by Weekday", x = "Day of the Week", y = "Total Number of Rainy Days") +
    guides(fill="none")
```




Saturday was the day of the week with the most rainy days, while Wednesday had the least rainy days, though it is extremely close. Saturday had 72 more than Wednesday, which is a difference of less than 5% of Saturday's total. Still, it was a fun question to answer.

## Is there a correlation between wind speed and precipitation, or wind direction and precipitation?

I wanted to deepen my analysis by including wind data as well. The Kaggle page) (https://www.kaggle.com/datasets/rtatman/did-it-rain-in-seattle-19482017) for the original dataset pointed to the NOAA website) (https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00024233/detail), where extra data was available. I downloaded a CSV file containing extra data, such as average wind speed `awnd` in mph, peak gust wind speed `wsfg` (the highest wind speed recorded for the day) in mph, and the direction of peak wind gust `wdfg` in degrees, with 0 and 360 degrees being north.

```
# Clean the dataframe
colnames(windData) <- tolower(colnames(windData))
windData <- sqldf("SELECT date, awnd, wdfg, wsfg FROM windData WHERE date NOT NULL AND awnd NOT NULL AND wdfg NOT NULL AND w
sfg NOT NULL")
windData <- distinct(windData)

# Merge the rainData and windData dataframes
rainData <- merge(windData, rainData, by="date", all=TRUE)
windData <- sqldf("SELECT date, awnd, wsfg, wdfg, prcp FROM rainData WHERE awnd NOT NULL AND wdfg NOT NULL AND wsfg NOT NUL
L")
kable(head(windData), caption="Merged dataframe for wind analysis")
```

Merged dataframe for wind analysis

| date | awnd | wsfg | wdfg | prcp |
|---|---|---|---|---|
| 1984-01-01 | 4.92 | 17.2 | 90 | 0.00 |
| 1984-01-02 | 7.61 | 17.2 | 180 | 0.62 |
| 1984-01-03 | 13.42 | 24.2 | 180 | 0.61 |
| 1984-01-04 | 10.07 | 27.5 | 180 | 0.19 |
| 1984-01-05 | 4.70 | 11.4 | 135 | 0.02 |
| 1984-01-06 | 5.14 | 11.4 | 360 | 0.03 |

```
nrow(windData)
```

```
## [1] 4654
```

Upon inspection of the new data, I found that wind readings were only available from 1984-01-01 to 1996-09-30. This isn't a big deal since there are still 4500+ days available, which is more than enough for a solid analysis.

# Is there a correlation between average wind speed and precipitation level?

In order to make a scatterplot, I took a random sample of 200 days when wind data was recorded.

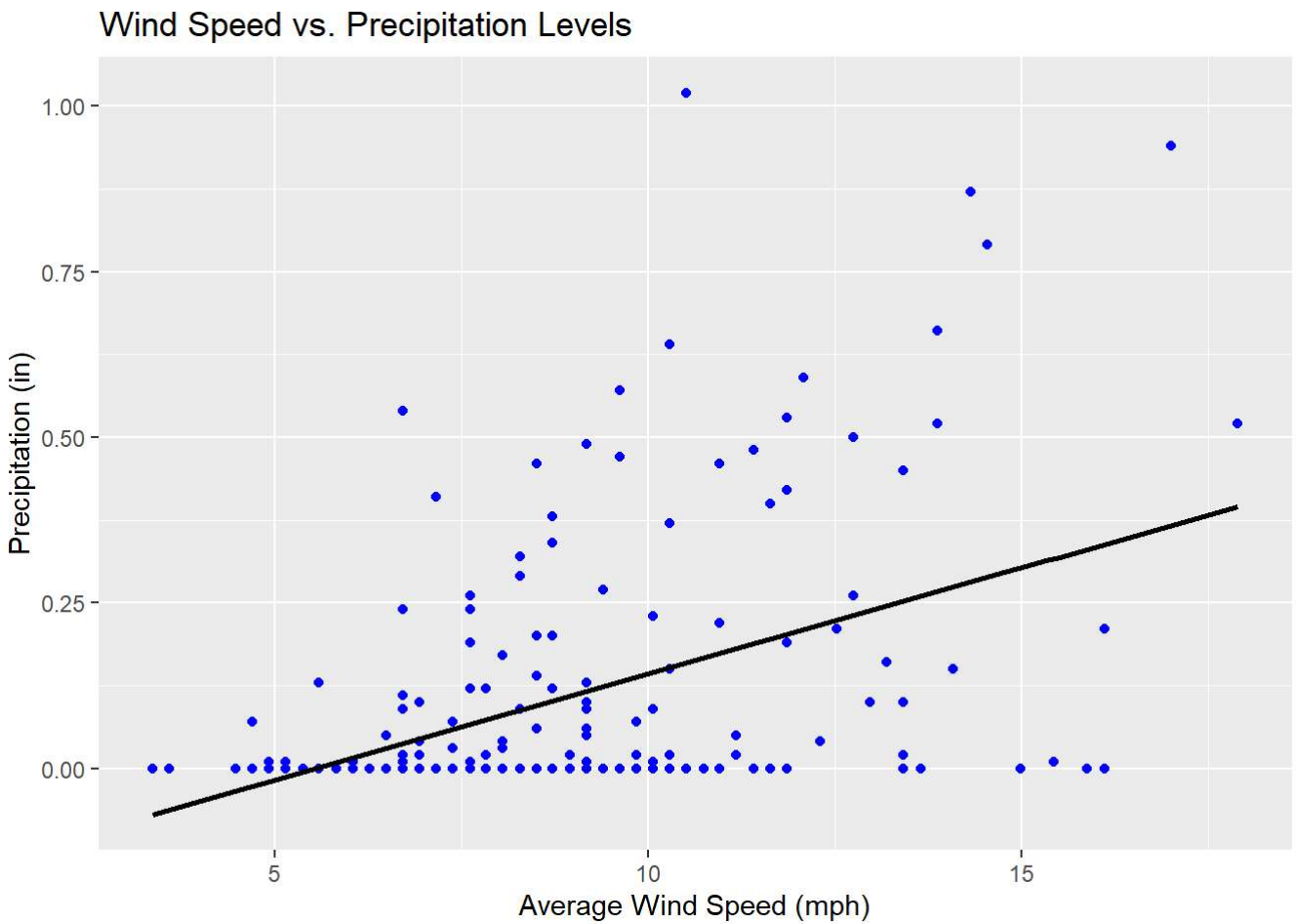
```
# Look for correlation between avg wind speed and precipitation level
windSample <- sample_n(windData, 200)
windSample <- sqldf("SELECT date, awnd, prcp FROM windSample")
kable(head(windSample), caption="Sample of wind data", align="l")
```

Sample of wind data

| date | awnd | prcp |
|---|---|---|
| 1986-08-18 | 8.50 | 0.00 |
| 1996-09-10 | 7.61 | 0.00 |
| 1992-10-08 | 9.17 | 0.06 |
| 1985-11-08 | 7.83 | 0.02 |
| 1989-04-09 | 16.11 | 0.00 |
| 1987-02-02 | 9.17 | 0.10 |

```
# Create the scatterplot
ggplot(windSample, aes(x = awnd, y = prcp)) +
  geom_point(color="blue") +
  geom_smooth(method = "lm", se = FALSE, color="black") +
  labs(title = "Wind Speed vs. Precipitation Levels", x = "Average Wind Speed (mph)", y = "Precipitation (in)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Find the regression coefficient
r_squared_windSpeed <- summary(lm(awnd ~ prcp, data = windSample))$r.squared
print(r_squared_windSpeed)
```

```
## [1] 0.1999758
```

The correlation coefficient $r^2$ is very small, so there is no meaningful correlation between wind speed and precipitation levels.

# Which wind direction is associated with the most precipitation?

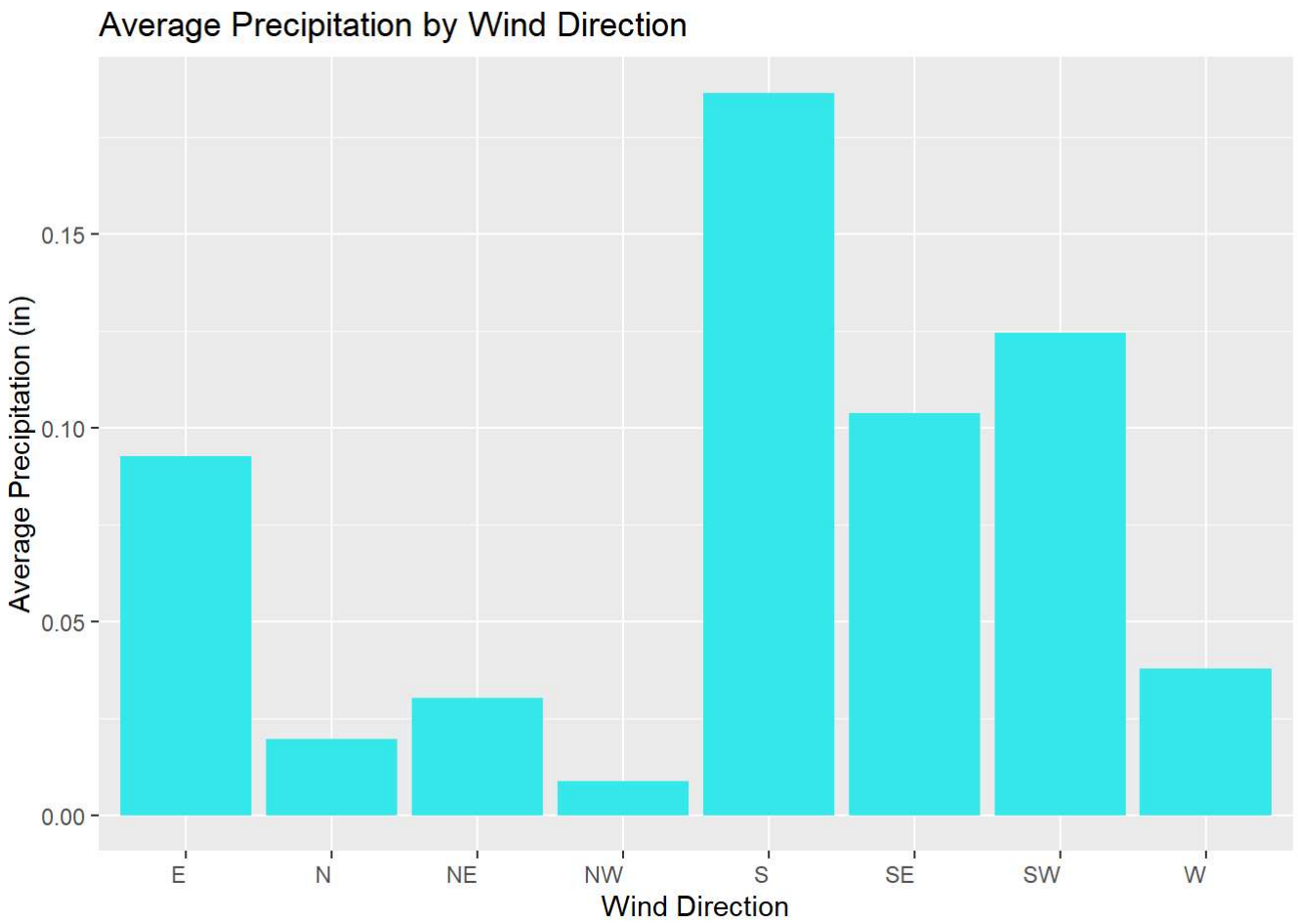
My hypothesis was that it rains more when the wind is blowing east, since it was my understanding that rain clouds often form over the ocean or Puget Sound and subsequently get blown over land where they can then dump their precipitation on Seattle.

First, I was suspicious that some of the values in wdfg might not be multiples of 45, which they would need to be in order to be a valid direction (0/360 is north, 45 northeast, 90 east, 135 southeast, 180 south, 225 southwest, 270 west, and 315 northwest). I filtered those out so only the values that made sense were retained.

```
# Find average precipitation for each wind direction
prcpPerWindDirection <- sqldf("SELECT wdfg AS wind_angle, AVG(prcp) AS avg_prcp FROM windData WHERE wdfg%45 == 0 GROUP BY wd
fg")
prcpPerWindDirection <- mutate(prcpPerWindDirection, dir=c("NE", "E", "SE", "S", "SW", "W", "NW", "N"))
prcpPerWindDirection <- prcpPerWindDirection[, c(1,3,2)]
kable(head(prcpPerWindDirection, 8))
```

| wind_angle | dir | avg_prcp |
|---:|---|---:|
| 45 | NE | 0.0301176 |
| 90 | E | 0.0926042 |
| 135 | SE | 0.1036458 |
| 180 | S | 0.1863621 |
| 225 | SW | 0.1244648 |
| 270 | W | 0.0377567 |
| 315 | NW | 0.0087273 |
| 360 | N | 0.0197022 |

```
# Show data in a bar graph
ggplot(prcpPerWindDirection, aes(x = dir, y = avg_prcp)) +
  geom_bar(stat = "identity", fill = "#34e8eb") +
  labs(title = "Average Precipitation by Wind Direction", x = "Wind Direction", y = "Average Precipitation (in)") +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))
```




Contrary to my prediction, winds blowing south were associated with the most rainfall.

# Wrap-up

I am surprised by how many insights I uncovered from this Seattle rain data. It was fun to dive in and see what I could find. I hope you enjoyed this analysis, and be sure to check out my Kaggle page for more projects!