EDA Long Form

Shay Lebovitz

11/23/2020

Introduction

This project is based on weather data for major U.S. and Canadian cities over the five year period from October 1, 2012 to November 30, 2017. The data was compiled from two sources. The first (link: https://www.kaggle.com/selfishgene/historical-hourly-weather-data?select=humidity.csv) gave hourly weather data over this time period for 27 United States cities, 3 Canadian cities, and 6 Israeli cities. The weather variables included were temperature, pressure, humidity, wind direction, wind speed, and a weather description, such as 'cloudy' or 'clear sky'. A separate data set for each of these weather attributes was provided, along with a final data set describing the cities. The second (link: https://kilthub.cmu.edu/articles/dataset/Compiled_daily_temperature_and_precipitation_data_for_the_U_S_cities/7890488) gave daily precipitation data for many United States and Canadian cities. However, it did not provide data for 5 cities that were in the original data set. These cities are Toronto, Montreal, San Francisco, Las Vegas, and Saint Louis.

Around 60% of the time working on this project was spent compiling and cleaning the data into one tidy data set. I ran into many problems trying to join the large data sets, because my computer would run out of memory. Thus, I went through a very long winded and tedious process to try to overcome this issue. However, I realized in the end that a simple fix in the way I was joining the data sets allowed me to join them with ease. Essentially, my computer could not handle joining the data sets via a single key (which is what I was originally attempting), but could join them via two keys. After this problem was fixed, I was able to easily clean the data and proceed with creating visualizations.

I have one additional note about a key assumption made in this project. Because I could only find daily precipitation data, but wanted to keep my data in hourly form for certain explorations, I resorted to simply dividing the daily precipitation value by 24 to achieve an average hourly precipitation value. While this is obviously not very accurate, I make sure that all analyses regarding precipitation data is done via daily precipitation, not hourly. Thus, the inaccuracy is not a problem.

Data

```
glimpse(all_data)
```

```
## Rows: 1,357,590
## Columns: 14
## $ city                <chr> "Vancouver", "Vancouver", "Vancouver", "Vancouver…
## $ country             <chr> "Canada", "Canada", "Canada", "Canada", "Canada",…
## $ latitude            <dbl> 49.24966, 49.24966, 49.24966, 49.24966, 49.24966,…
## $ longitude           <dbl> -123.1193, -123.1193, -123.1193, -123.1193, -123.…
## $ datetime            <dttm> 2012-10-01 12:00:00, 2012-10-01 13:00:00, 2012-1…
## $ temperature         <dbl> NA, 52.93400, 52.93227, 52.92860, 52.92492, 52.92…
## $ humidity            <dbl> NA, 76, 76, 76, 77, 78, 78, 79, 79, 80, 81, 81, 8…
## $ pressure            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ precipitation       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ weather_description <chr> NA, "mist", "broken clouds", "broken clouds", "br…
## $ wind_direction      <dbl> NA, 0, 6, 20, 34, 47, 61, 75, 89, 102, 116, 130, …
## $ wind_speed          <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ daily_precip        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ day                 <date> 2012-10-01, 2012-10-01, 2012-10-01, 2012-10-01, …
```

```
unique(all_data$city)
```

```
##  [1] "Vancouver"     "Portland"      "San Francisco" "Seattle"
##  [5] "Los Angeles"   "San Diego"     "Las Vegas"     "Phoenix"
##  [9] "Albuquerque"   "Denver"        "San Antonio"   "Dallas"
## [13] "Houston"       "Kansas City"   "Minneapolis"   "Saint Louis"
## [17] "Chicago"       "Nashville"     "Indianapolis"  "Atlanta"
## [21] "Detroit"       "Jacksonville"  "Charlotte"     "Miami"
## [25] "Pittsburgh"    "Toronto"       "Philadelphia"  "New York"
## [29] "Montreal"      "Boston"
```

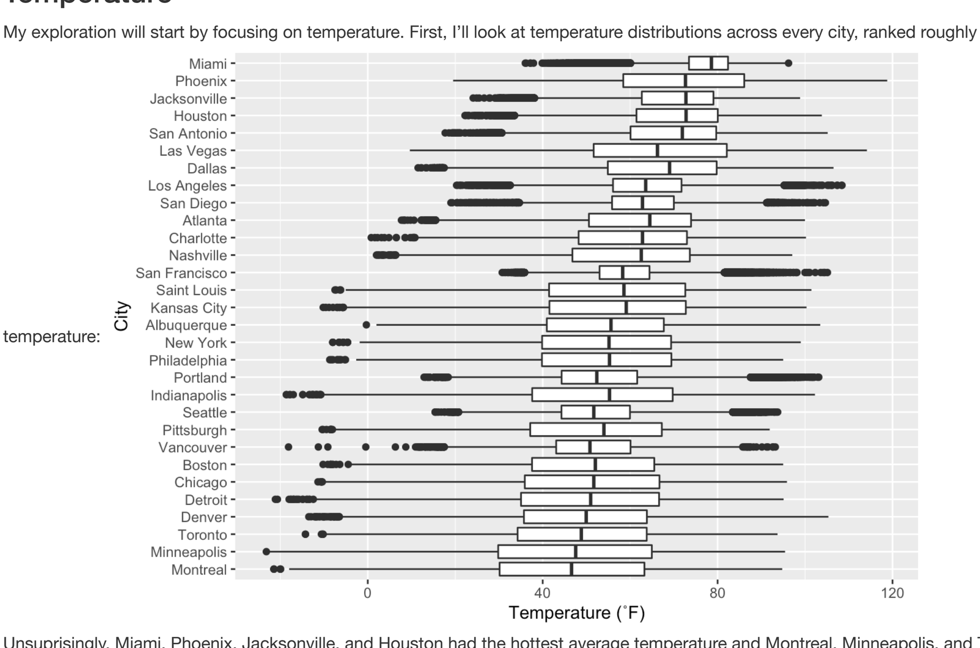
The tibble `all_data` contains all the weather data compiled from the two sources, cleaned and tidied. It contains 4 city-identifier variables: `city`, `country`, `latitude`, and `longitude`; 2 time-identifier variables: `datetime` and `day` (created for ease in certain visuals); and 8 weather attribute variabls: `temperature` (˚F), `humidity`, `pressure` (torr), `precipitation` (hourly, inches), `weather_description`, `wind_direction` (degrees), `wind_speed` (mph), and `daily_precip` (inches, and again, used for ease in certain visuals). The data set runs from noon on October 1, 2012 to midnight on November 30, 2017. Because the amount of cities in the data set was already large enough, I removed the 6 Israeli cities. The remaining cities are Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Diego, Las Vegas, Phoenix, Albuquerque, Denver, San Antonio, Dallas, Houston, Kansas City, Minneapolis, Saint Louis, Chicago, Nashville, Indianapolis, Atlanta, Detroit, Jacksonville, Charlotte, Miami, Pittsburgh, Toronto, Philadelphia, New York, Montreal, and Boston.

As mentioned earlier, Toronto, San Francisco, Saint Louis, Montreal, and Las Vegas do not have any precipitation data.

Data Explorations

Temperature

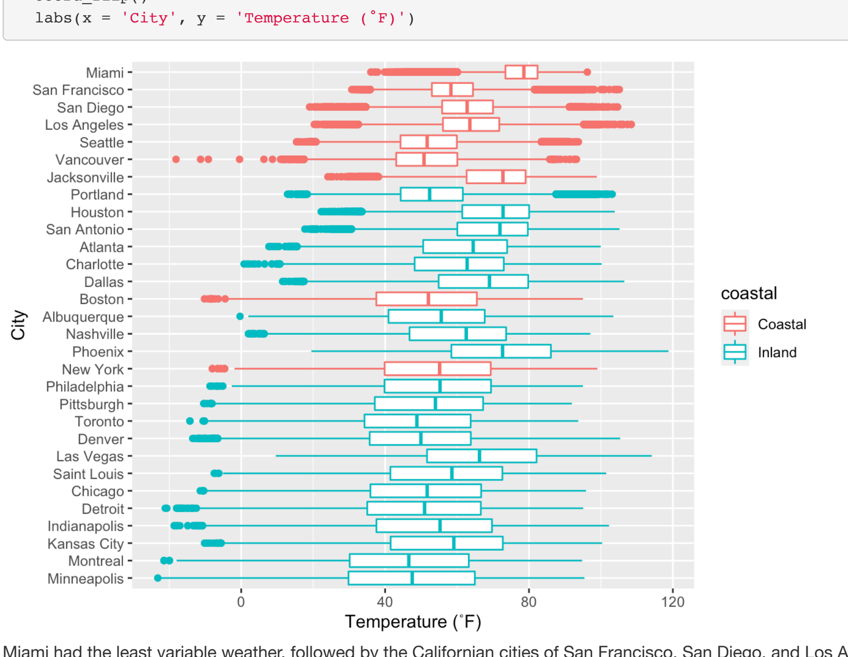
My exploration will start by focusing on temperature. First, I'll look at temperature distributions across every city, ranked roughly by average temperature:



Unsuprisingly, Miami, Phoenix, Jacksonville, and Houston had the hottest average temperature and Montreal, Minneapolis, and Toronto had the coldest. An interesting point is that Denver has a colder average temperature than Chicago, which may be a little surprising for some.

Using this same plot but ranking by standard deviation allows us to see the temperature variability of cities:

```
all_data %>%
  mutate(coastal = ifelse(city %in% coastal_city, 'Coastal', 'Inland')) %>%
  group_by(city) %>%
  mutate(sd_temp = sd(temperature, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(city_by_temp = fct_reorder(city, desc(sd_temp))) %>%
  ggplot(aes(x = city_by_temp, y = temperature, color = coastal)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'City', y = 'Temperature (˚F)')
```

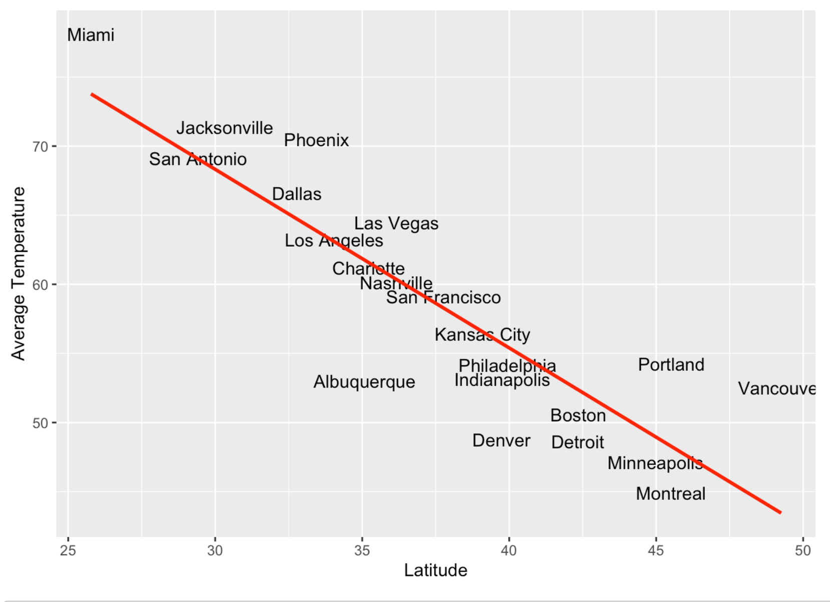


Miami had the least variable weather, followed by the Californian cities of San Francisco, San Diego, and Los Angeles. Minneapolis, Montreal, and Kansas city had the most variable weather. For the most part, we see that coastal cities tend to have much less variable weather than inland cities, with the top 7 all being coastal and the bottom 12 being inland.

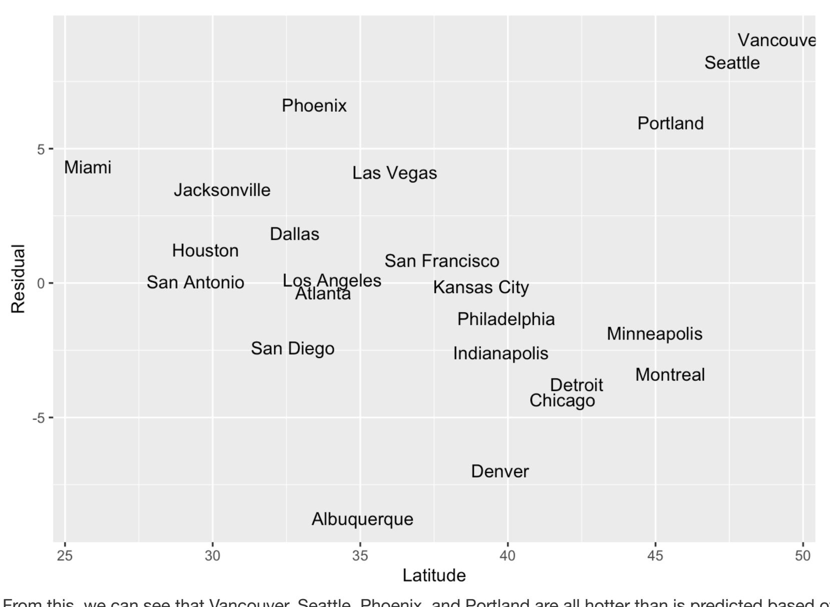
From these graphs, we can also determine the maximum and minimum temperature rankings. Minneapolis, Montreal, and Detroit had the lowest temperatures recorded over the time period, while Miami, San Francisco, and Jacksonville had the highest minimum temperatures. Phoenix, Las Vegas, and Los Angeles had the hottest temperatures recorded, and Pittsburgh, Vancouver, and Toronto had the coolest maximum temperatures. There are a few interesting points here: 1) Las Vegas had a lower minimum termperature than Seattle and Portland; 2) Chicago had a higher minimum termperature than Vancouver and Indianapolis (although this is likely false based on external sources); and 3) Miami had the highest average temperature, but is 20th out of 30 for overall maximum temperature.

We know that on average, the further towards the poles a place is, the colder it will be. However, due to the interesting geography of the United States, this is not always the case. By fitting a model of average temperature on latitude, we can see which cities break this rule.

```
all_data2 %>%
  ggplot(aes(x = latitude, y = avg_temp)) +
  geom_text(aes(label = city), check_overlap = TRUE) +
  geom_line(data = grid, color = 'red', size = 1) +
  labs(x = 'Latitude', y = 'Average Temperature')
```



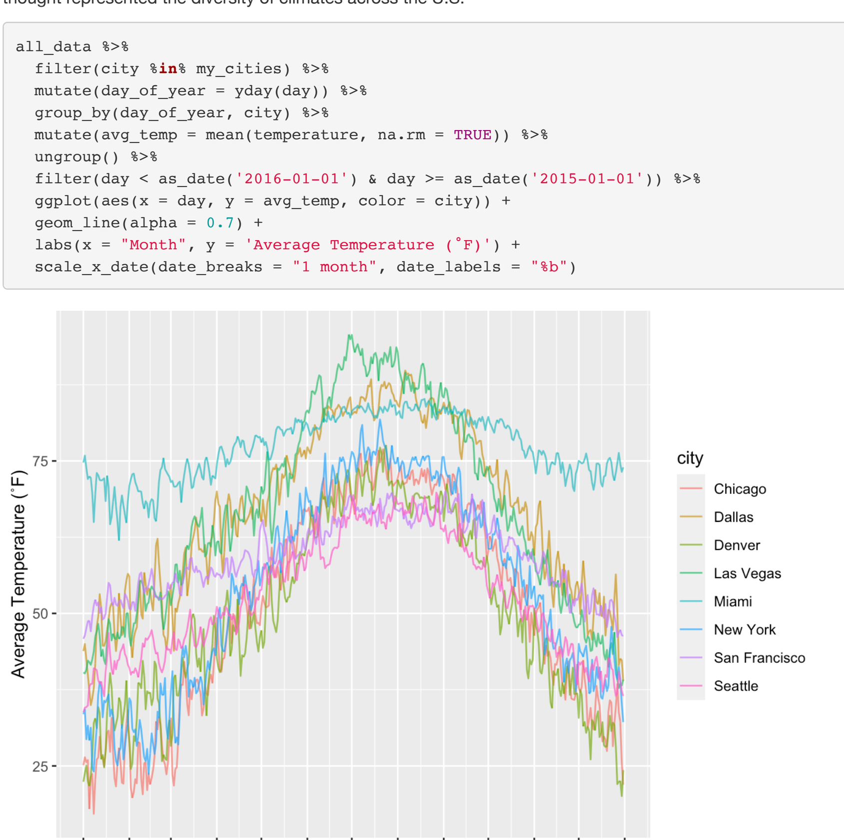

```
all_data2 %>%
  ggplot(aes(x = latitude, y = resid)) +
  geom_text(aes(label = city), check_overlap = TRUE) +
  labs(x = 'Latitude', y = 'Residual')
```



From this, we can see that Vancouver, Seattle, Phoenix, and Portland are all hotter than is predicted based of their latitude. This especially makes sense for Vancouver, Seattle, and Portland, which are all actually north of Minneapolis, yet have significantly warmer weather due to the Pacific Ocean. On the other end, Albuquerque and Denver have significanly lower average temperatures than would be expected based on their latitude, which is most likely due to elevation (both cities are above 5000 ft). Most of the other cities have relatively low residuals.

I'd also like to see how temperature ranges over different time frames. First, I'll start with how temperature varies over the year. Note that a smaller subsection of the cities were chosen for this analysis, because including all of them made the graphs too hard to read. I chose cities that I thought represented the diversity of climates across the U.S.

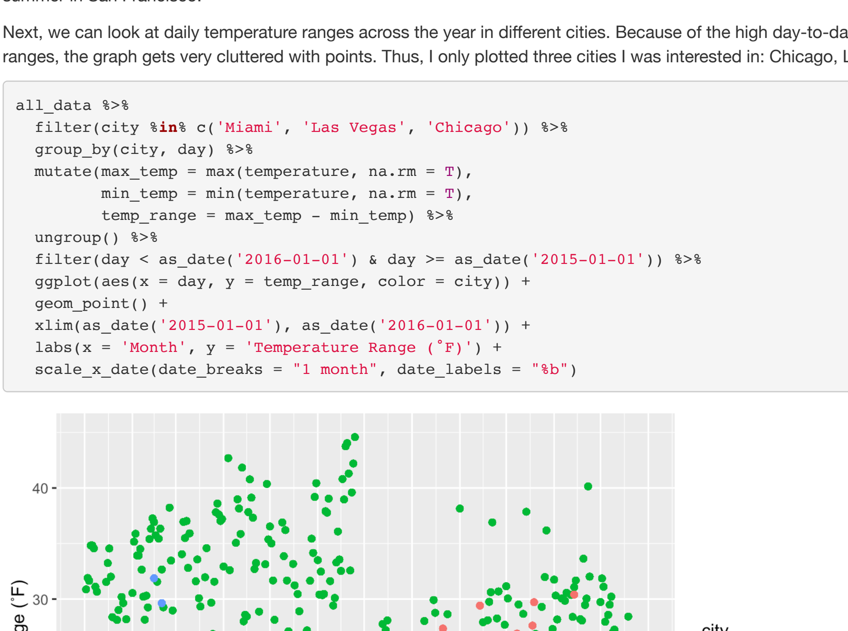
```
all_data %>%
  filter(city %in% my_cities) %>%
  mutate(day_of_year = yday(day)) %>%
  group_by(day_of_year, city) %>%
  mutate(avg_temp = mean(temperature, na.rm = TRUE)) %>%
  ungroup() %>%
  filter(day < as_date('2016-01-01') & day >= as_date('2015-01-01')) %>%
  ggplot(aes(x = day, y = avg_temp, color = city)) +
  geom_line(alpha = 0.7) +
  labs(x = "Month", y = 'Average Temperature (˚F)') +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



Some interesting points here: Miami has by far the least variation in temperature, with daily temperatures in January staying moslty above 60˚. Las Vegas has the hottest summers followed by Dallas. Chicago has the coldest winters. Although San Francisco has the second warmest winters, it is practically tied with Seattle as having the coldest summers. This is what prompted Mark Twain to state "The coldest winter I ever spent was a summer in San Francisco."

Next, we can look at daily temperature ranges across the year in different cities. Because of the high day-to-day variability of daily temperature ranges, the graph gets very cluttered with points. Thus, I only plotted three cities I was interested in: Chicago, Las Vegas, and Miami.

```
all_data %>%
  filter(city %in% c('Miami', 'Las Vegas', 'Chicago')) %>%
  group_by(city, day) %>%
  mutate(max_temp = max(temperature, na.rm = T),
         min_temp = min(temperature, na.rm = T),
         temp_range = max_temp - min_temp) %>%
  ungroup() %>%
  filter(day < as_date('2016-01-01') & day >= as_date('2015-01-01')) %>%
  ggplot(aes(x = day, y = temp_range, color = city)) +
  geom_point() +
  xlim(as_date('2015-01-01'), as_date('2016-01-01')) +
  labs(x = 'Month', y = 'Temperature Range (˚F)') +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



Las Vegas has consistently higher daily temperature ranges than Chicago and Miami. Miami seems to have greater daily temperature ranges in the first part of the year but Chicago does in the second part of the year.

Precipitation

Now, we'll look at precipitation. Becasue the majority of days see 0 precipitation, even the rainiest cities have average daily precipitation totals very close to 0. To better visualize how average rainfall varies across cities, I will only compute the average for days that had non-zero precipitation.

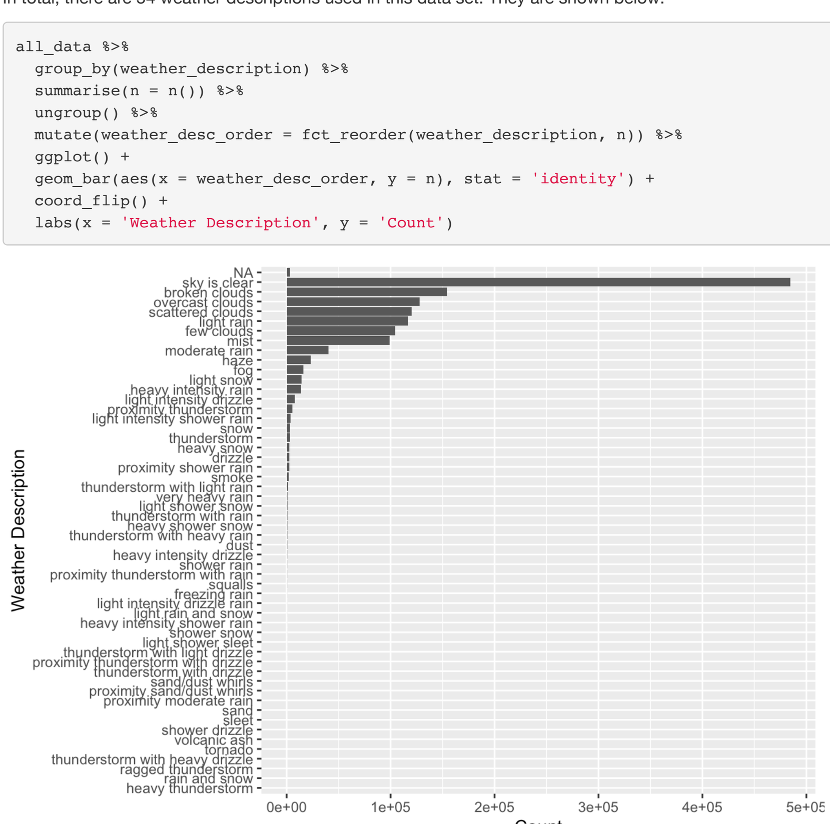
```
all_data %>%
  group_by(city) %>%
  mutate(mean_prec = mean(daily_precip, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(no_zero_prec = ifelse(daily_precip == 0, NA, daily_precip)) %>%
  mutate(city_by_prec = fct_reorder(city, mean_prec)) %>%
  ggplot(aes(x = city_by_prec, y = no_zero_prec)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'City', y = 'Daily Precipitation (in)')
```

Even when disregarding the zero precipitation days when taking the average of precipitation of each city, we still see that the average and standard deviation are very small compared to the huge outliers for each city. We see that Miami, Houston, and Atlanta are the rainiest cities (or more technically, rain the most when it does rain) and Phoneix, Albuquerque, and San Diego are the least rainy. The day that it rained upwards of 16 inches in Houston was on August 27, 2017, during Hurricane Harvey.

Weather Descriptions

In total, there are 54 weather descriptions used in this data set. They are shown below:

```
all_data %>%
  group_by(weather_description) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  mutate(weather_desc_order = fct_reorder(weather_description, n)) %>%
  ggplot() +
  geom_bar(aes(x = weather_desc_order, y = n), stat = 'identity') +
  coord_flip() +
  labs(x = 'Weather Description', y = 'Count')
```



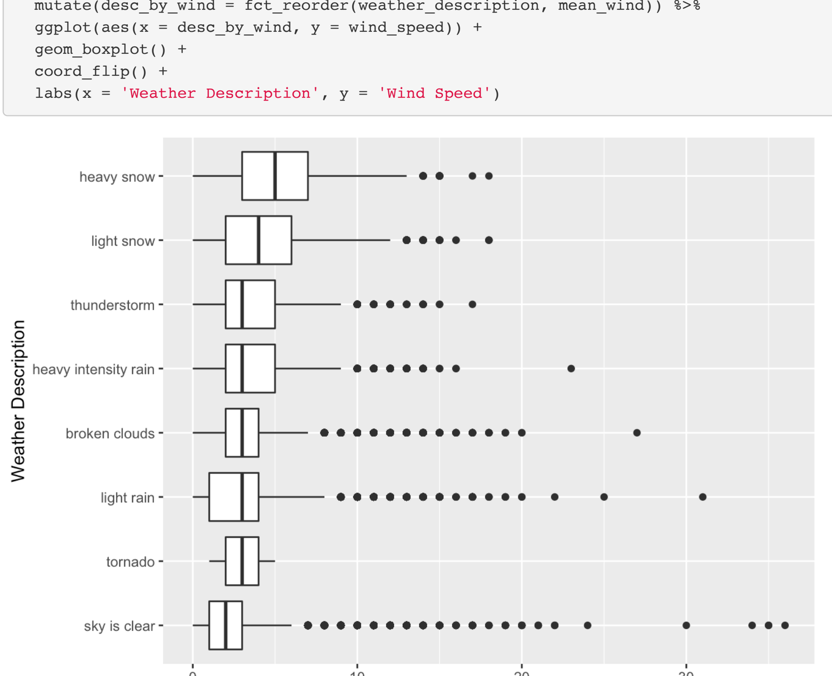
`sky is clear` is by far the most common weather description, followed by `broken clouds`, `overcast clouds`, `scattered clouds`, `light rain`, `few clouds`, and `mist`. After that, there is a steep decline in number of observations. It looks like the top 12 weather types (up until `heavy intensity rain`) will account for ~99% of the weather descriptions, so further analysis will likely be restricted to them.

```
rel_descs <- c('sky is clear', 'broken clouds', 'overcast clouds', 'scattered clouds',
               'light rain', 'few clouds', 'mist', 'moderate rain', 'haze', 'fog',
               'light snow', 'heavy intensity rain', 'snow', 'thunderstorm')
all_data %>%
  filter(weather_description %in% rel_descs) %>%
  group_by(weather_description) %>%
  mutate(mean_temp = mean(temperature, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(desc_by_temp = fct_reorder(weather_description, mean_temp)) %>%
  ggplot(aes(x = desc_by_temp, y = temperature)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'Weather Description', y = 'Temperature (˚F)')
```

Plotting average temperature against the weather description shows us some interesting data. We see that the average temperature for a thunderstorm is roughly 80˚F, likely because thunderstorms predominantly occur in the summer. Most of the rest hover around 50-70˚F, except for the two snow descriptions with average temperature of around 25˚F. Some interesting observations is that there were times when they labeled `snow` when it was over 70˚, and a few instances of some type of rain when it was well below 32˚F. However, I have found some other very strange occurences in this data set (such as a day in Kansas City where it went from 5˚F to 95˚F in one day), which tells me that these are likely mistakes.

We can also look to see how weather description is associated with wind speed:

```
rel_descs2 <- c('sky is clear', 'broken clouds', 'light rain', 'light snow',
                'thunderstorm', 'tornado', 'heavy snow', 'heavy intensity rain')
all_data %>%
  filter(weather_description %in% rel_descs2) %>%
  group_by(weather_description) %>%
  mutate(mean_wind = mean(wind_speed, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(desc_by_wind = fct_reorder(weather_description, mean_wind)) %>%
  ggplot(aes(x = desc_by_wind, y = wind_speed)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'Weather Description', y = 'Wind Speed')
```

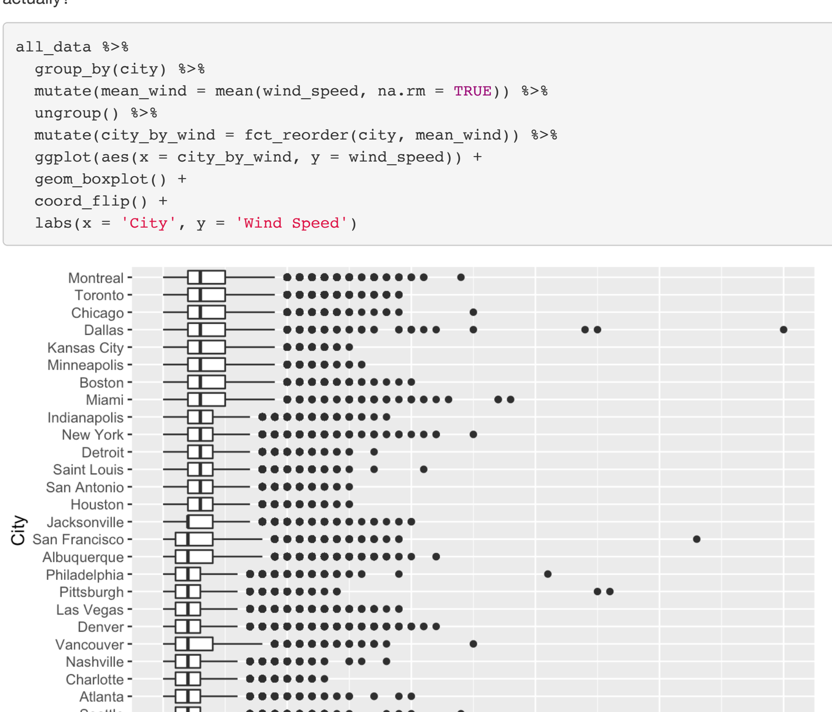


Interestingly, `tornado` had the second lowest average wind speed, but this is likely because of lack of data points for this weather description. `heavy snow` had the highest average wind speed, followed by `light snow` and `thunderstorm`, which makes sense. `sky is clear` had the lowest average wind speed, but also the highest individual instance. This is likely just because it has so many more data points than any other weather description.

Wind

We all know Chicago as being the "Windy City" (although I know this doesn't actually have to do with the wind), but how windy is Chicago actually?

```
all_data %>%
  group_by(city) %>%
  mutate(mean_wind = mean(wind_speed, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(city_by_wind = fct_reorder(city, mean_wind)) %>%
  ggplot(aes(x = city_by_wind, y = wind_speed)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'City', y = 'Wind Speed')
```



Here, we can see that Montreal, Toronto, and Chicago are the windiest cities. The least windy are Los Angeles, San Diego, and Phoneix. Phoenix, however, had the second windiest day with roughly 48 mph winds, second only to a day in Dallas with 50 mph winds. So, the title of the "Windy City" does seem appropriate for Chicago, as it is the windiest of all the cities in America in this data set.

Conclusion

This exploration covers only a fraction of what could be done with this expansive data set. Some ideas for future exploration include observing precipitation across the year, examining wind directions in different cities, or examining how changes in pressure are related to changes in weather patterns. Additionally, other data sets could be used to explore other relationships, such as how temperature seems to be related to crime rates. Overall, this project was a great way to explore weather patterns in the United States, something I have been interested in for a long time. It was also a great learning experience in dealing with large data sets, visualizing data, joining data sets, and dealing with dates and times.