

New Dehli Weather Portfolio Project

Olivia Lund

This project is an analysis of a dataset of detailed weather records from New Dehli, India, starting in November 1996 and ending in April 2014. This dataset was chosen because of its size and detail, featuring over one hundred thousand entries and over twenty different fields. These data can be used to perform a variety of analyses linking different attributes, for example finding what times of the year are more prone to smog, or if wind direction can be a good indicator for presence of precipitation.

Before being able to conduct any effective data analysis, as with any other skilled project, the right tools have to be selected. This project will be using the library TidyVerse because of its extensive data organization and the library GGPlot because of its concise and dynamic visualization options.

```
library("tidyverse")
```

```
## — Attaching packages — tidyverse 1.2.1 —
```

```
## ✓ ggplot2 3.2.1    ✓ purrr   0.3.2
## ✓ tibble  2.1.3    ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0    ✓ stringr 1.4.0
## ✓ readr   1.3.1    ✓ forcats 0.4.0
```

```
## — Conflicts — tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
```

```
library("ggplot2")
```

In order to read and modify the dataset, it first has to be read into this workspace, done here via local filepath. This data is sourced from user mahirkukreja who posted it to kaggle.com on the page <https://www.kaggle.com/mahirkukreja/delhi-weather-data/> (<https://www.kaggle.com/mahirkukreja/delhi-weather-data/>).

```
weather <- read.csv(file="~/Documentos/Escuela/19fa/DSCI_385/P01/dehliweather.csv",)
```

```
[1] "datetime_utc" "X_conds" "X_dewptm"
[4] "X_fog" "X_hail" "X_heatindexm" [7] "X_hum" "X_precipm" "X_pressurem" [10] "X_rain" "X_snow"
"X_tempm"
[13] "X_thunder" "X_tornado" "X_vism"
[16] "X_wdird" "X_wdire" "X_wgustm"
[19] "X_windchillm" "X_wspdm"
```

The initial column labels of the dataset were copied and pasted here from a terminal output. If the terminal output was shown instead, they could be updated as the dataset is changed. This shows that the original label format is strange and confusing at times, and uses strange or arbitrary notations and abbreviations. In order to help make the dataset tidy and readable, some of these names will be changed to be more clear:

```

colnames(weather)[colnames(weather)=="datetime_utc"]<- "datetime_utc"
colnames(weather)[colnames(weather)=="X_conds"]<- "air_condition"
colnames(weather)[colnames(weather)=="X_dewptm"]<- "dew_point"
colnames(weather)[colnames(weather)=="X_fog"]<- "any_fog"
colnames(weather)[colnames(weather)=="X_hail"]<- "any_hail"
colnames(weather)[colnames(weather)=="X_heatindexm"]<- "heat_index"
colnames(weather)[colnames(weather)=="X_hum"]<- "humidity"
colnames(weather)[colnames(weather)=="X_precipm"]<- "precipitation"
colnames(weather)[colnames(weather)=="X_pressurem"]<- "air_pressure"
colnames(weather)[colnames(weather)=="X_rain"]<- "any_rain"
colnames(weather)[colnames(weather)=="X_snow"]<- "any_snow"
colnames(weather)[colnames(weather)=="X_tempm"]<- "temperature"
colnames(weather)[colnames(weather)=="X_thunder"]<- "any_thunder"
colnames(weather)[colnames(weather)=="X_tornado"]<- "any_tornadoes"
colnames(weather)[colnames(weather)=="X_vism"]<- "visibility"
colnames(weather)[colnames(weather)=="X_wdird"]<- "wind_angle"
colnames(weather)[colnames(weather)=="X_wdire"]<- "wind_direction"
colnames(weather)[colnames(weather)=="X_wgustm"]<- "wind_gust_speed"
colnames(weather)[colnames(weather)=="X_windchillm"]<- "wind_chill"
colnames(weather)[colnames(weather)=="X_wspdm"]<- "wind_average_speed"
head(weather, n = 5)

```

datetime_utc <fctr>	air_condition <fctr>	dew_point <int>	any_f... <int>	any_hail <int>	heat_index <dbl>	humidity <fctr>	pre
1 19961101-11:00	Smoke	9	0	0	NA	27	
2 19961101-12:00	Smoke	10	0	0	NA	32	
3 19961101-13:00	Smoke	11	0	0	NA	44	
4 19961101-14:00	Smoke	10	0	0	NA	41	
5 19961101-16:00	Smoke	11	0	0	NA	47	
5 rows 1-9 of 21 columns							

These data labels make the data easier to analyze by clearly describing what they represent. It requires no experience with data science to understand what each of these values is generally trying to say. Specifics on each field will be explained below, which will give us an idea what fields warrant further changes:

datetime_utc This is simply the date and time of day using the Coordinated Universal Time system. It is continuous and should be changed to the date format built into R.

air_condition This field contains the primary air condition at the time of the reading. It is a categorical form variable that can take one of many different types of values. Here are the first ten encountered in the list: [2]

Clear

[3] Haze

[4] Unknown

[5] Scattered Clouds

[6] Shallow Fog

[7] Mostly Cloudy

[8] Fog

[9] Partly Cloudy

[10] Patches of Fog

dew_point This is the dew point at the current time, which is the temperature at which dew will form, given the current temperature and pressure conditions. It is a decimal with one point of precision.

any_fog This attribute represents whether or not there currently is fog, represented as a boolean value.

any_hail This is a boolean attribute representing if there currently is hail. There are no data points during which it was hailing, making this a prime candidate for removal.

heat_index This is a continuous number variable representing the current heat index, which takes into account humidity to more accurately represent perceived temperature.

humidity This is a continuous decimal number variable that represents the current humidity.

precipitaton All entries for this value contain missing entries, making it a prime candidate for removal.

air_pressure This value represents air pressure. The values are all around 1000, which isn't in a reasonable range for any of the most common units, but this variable could still be kept around to compare relative air pressure within the dataset.

any_rain A boolean value expressing the presence of rain.

any_snow A boolean value expressing the presence of snow. Only true for one single reading in May of 2014. Being virtually empty makes it a candidate for removal.

temperature A decimal number representing the current temperature in centigrade.

any_thunder A boolean value expressing the presence of thunder.

any_tornadoes A boolean value expressing the presence of tornadoes. Only true for two entries, one during August of 1999, and one during February of 2007. Curiously, snow is less common in New Delhi than tornadoes. Because this is mostly empty, it is a candidate for deletion.

visibility A decimal number representing the current visibility, in km.

wind_angle An integer representing the degrees to the right of north that the wind is blowing.

wind_direction A string representing the direction of the wind in 3-letter cardinal direction combinations.

wind_gust_speed A decimal representing the max speed in km/hr that the wind is blowing in periods of fluctuation.

wind_chill An integer representing the apparent temperature due to the speed of the wind in centigrade.

wind_average_speed A decimal representing the current average speed that the wind is going.

Now that we have looked at our data, we can modify it and clean it up.

```
weather <- subset(weather, select = -c(any_snow, any_hail, any_tornadoes, precipitati  
on))
```

This project holds a lot of potential for identifying relationships between many different aspects of the climate of New Delhi over the past twenty years, and displaying the power of Data Science in the process.