PM 566

Lab 3 - Exploratory Data Analysis

Learning Goals

- Read in and get familiar with the meteorology dataset
- Step through the EDA "checklist" presented in the class slides
- Practice making exploratory graphs

As you do this, think about what questions you would like to ask regarding this data. What would you ask a collaborator who was more familiar with it?

Lab Description

We will work with the meteorological data presented in lecture. Recall the dataset consists of weather station readings in the continental US.

The objective of the lab is to find the weather station with the highest elevation and look at patterns in the time series of its wind speed and temperature.

Steps

1. Read in the data

The data for this lab is available at https://github.com/USCbiostats/data-science-data/tree/master/02_met. There's some additional information about the dataset, but the main file you'll need to download is `met_all.gz`.

Once you've downloaded the file, you can read it into R with:

```
met <- read.csv("~/Downloads/met_all.gz")
```

2. Check the dimensions, headers, footers

- How many columns, rows are there?

```
dim(met)
head(met)
```

```
tail(met)
```

3. Take a look at the variables.

```
str(met)
```

4. Take a closer look at the key variables.

```
table(met$year)
table(met$day)
table(met$hour)
summary(met$temp)
summary(met$elev)
summary(met$wind.sp)
```

It looks like the elevation variable has observations with 9999.0, which is probably an indicator for missing. We should take a deeper look at the data dictionary to confirm. The wind speed variable is OK but there are a lot of missing data.

After checking the data we should make the appropriate modifications. Replace elevations with 9999 as `NA`.

```
met[met$elev==9999.0] <- NA
summary(met$elev)
```

- At what elevation is the highest weather station?

We also have the issue of the minimum temperature being -40C, so we should remove those observations.

```
met <- met[temp>-40]
met2 <- met[order(temp)]
head(met2)
```

We again notice that there is a -17.2C temperature reading that seems suspicious.

5. Check the data against an external data source.

We should check the suspicious temperature value (where is it located?) and validate that the range of elevations make sense (-13 m to 4113 m).

Google is your friend here.

Fix any problems that arise in your checks.

```
met <- met[temp>-15]
met2 <- met[order(temp)]
head(met2)
```

- Summarize that we removed temperatures colder than -15C. The new dataset has a minimum temp of -3C, which is reasonable.

6. Calculate summary statistics

Remember to keep the initial question in mind. We want to pick out the weather station with maximum elevation and examine its windspeed and temperature.

Some ideas: select the weather station with maximum elevation; look at the correlation between temperature and wind speed; look at the correlation between temperature and wind speed with hour and day of the month.

```
elev <- met[elev==max(elev), ]
summary(elev)
```

```
cor(elev$temp, elev$wind.sp, use="complete")
cor(elev$temp, elev$hour, use="complete")
cor(elev$wind.sp, elev$day, use="complete")
cor(elev$wind.sp, elev$hour, use="complete")
cor(elev$temp, elev$day, use="complete")
```

7. Exploratory graphs

We should look at the distributions of all of the key variables to make sure there are no remaining issues with the data.

```
hist(met$elev, breaks=100)
hist(met$temp)
hist(met$wind.sp)
```

One thing we should consider for later analyses is to log transform wind speed and elevation as they are very skewed.

Look at where the weather station with highest elevation is located.

```
leaflet(elev) %>%
  addProviderTiles('OpenStreetMap') %>%
  addCircles(lat=~lat,lng=~lon, opacity=1, fillOpacity=1, radius=100)
```

Look at the time series of temperature and wind speed at this location. For this we will need to create a date-time variable for the x-axis.

```
library(lubridate)
elev$date <- with(elev, ymd_h(paste(year, month, day, hour, sep= ' ')))
summary(elev$date)
elev <- elev[order(date)]
head(elev)
```

With the date-time variable we can plot the time series of temperature and wind speed.

```
plot(elev$date, elev$temp, type='l')
plot(elev$date, elev$wind.sp, type='l')
```

- Summarize any trends that you see in these time series plots.

8. Ask questions

By now, you might have some specific questions about how the data was gathered and what some of the different variables and values mean. Alternatively, maybe you have an idea for how some of the variable should be related and you want to explore that relationship. In a real-world analysis, these questions could potentially be answered by a collaborator, who may have been part of the team that collected the data. What questions do you have about the data?

If you haven't already, now would be a good time to look at the accompanying data dictionary for this dataset and see if it can answer any of your questions. If you have questions about the nature of the dataset and how it was gathered, this might be able to help.

For questions about variables in the dataset or relationships between them, try making some more exploratory plots. Do you see the patterns you would expect? There are many different types of summaries and visualization strategies that we have not discussed, but which could provide interesting perspectives on the data.

Some other useful plotting functions include: - `pairs` for making all pairwise scatter plots in a dataset with >2 dimensions. - `heatmap` and/or `corrplot` (from the `corrplot` package) for visualizing matrices in general or correlation matrices in particular. - `image` a low-level matrix visualization function - `barplot`, especially with `table`, for visualizing frequencies of categorical variables.

University of Southern California
Department of Population and Public Health Sciences