Weather Station Analysis

Jake Berberian

Purpose

- Using NOAA’s weather data, can we accurately predict the temperature at our location?
  - Multiple linear regression
- Data imputation
- Explore the datasets
- Make use of available data
  - Phones find station from vast network

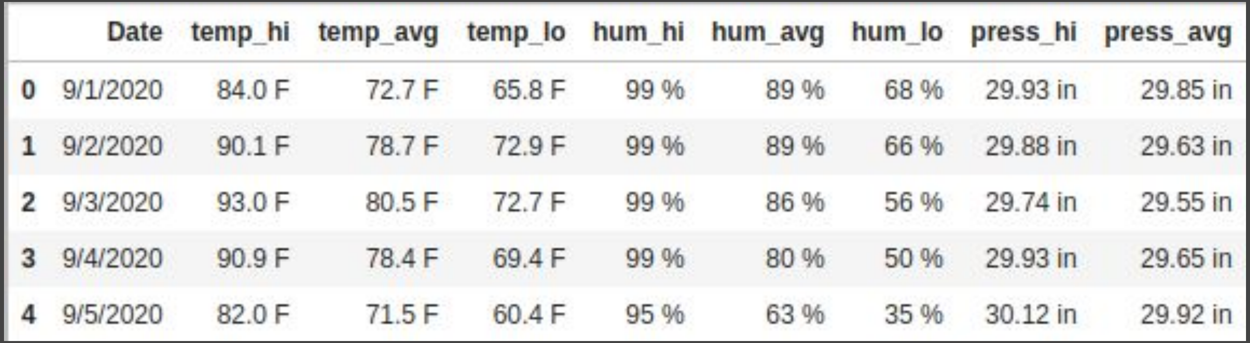
Setup

- We’ll use a personal weather station (PWS) at our location as our “response” variable
- Split data into train and test sets
  - Randomly sample
- Look at daily maximums and daily minimums
  - Not average due to data available
- Use September data due to NOAA constraints/completeness of data

The Weather Station

- Raspberry Pi 3 + Pi Sense HAT
  - Temperature, pressure measurements
- Weather Underground
  - Network of PWS

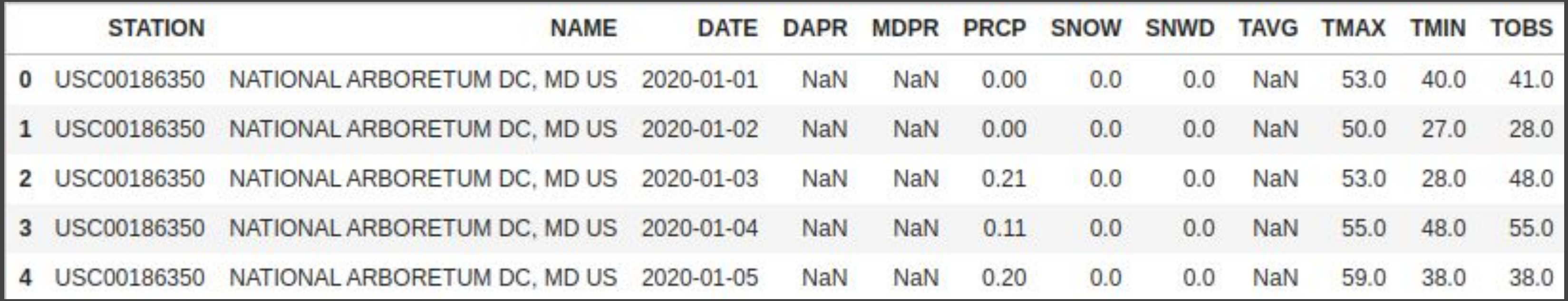
PWS Data



	Date	temp_hi	temp_avg	temp_lo	hum_hi	hum_avg	hum_lo	press_hi	press_avg
0	9/1/2020	84.0 F	72.7 F	65.8 F	99 %	89 %	68 %	29.93 in	29.85 in
1	9/2/2020	90.1 F	78.7 F	72.9 F	99 %	89 %	66 %	29.88 in	29.63 in
2	9/3/2020	93.0 F	80.5 F	72.7 F	99 %	86 %	56 %	29.74 in	29.55 in
3	9/4/2020	90.9 F	78.4 F	69.4 F	99 %	80 %	50 %	29.93 in	29.65 in
4	9/5/2020	82.0 F	71.5 F	60.4 F	95 %	63 %	35 %	30.12 in	29.92 in

- Collects hourly temperature, humidity and pressure readings
    - Use daily numbers because that’s what NOAA offers us
- Could add precipitation and dew points with more equipment

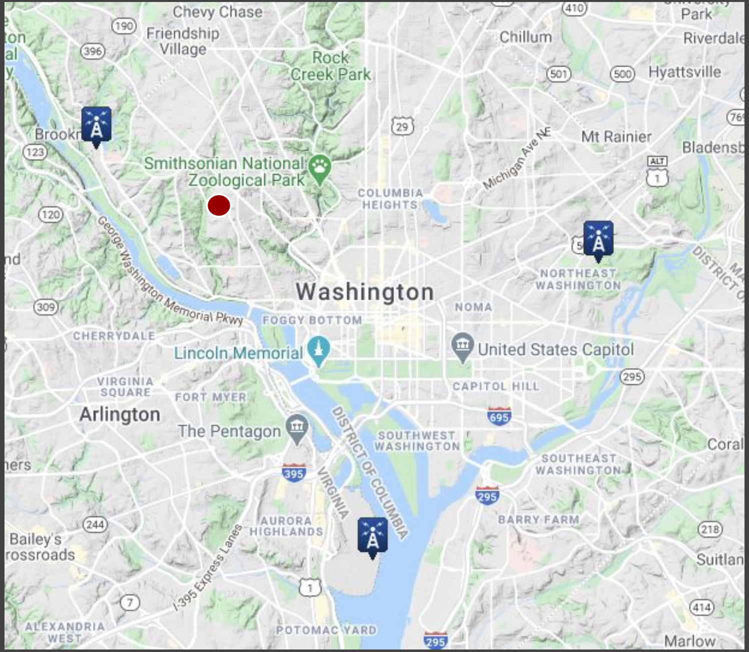
NOAA’s Data



	STATION	NAME	DATE	DAPR	MDPR	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	TOBS
0	USC00186350	NATIONAL ARBORETUM DC, MD US	2020-01-01	NaN	NaN	0.00	0.0	0.0	NaN	53.0	40.0	41.0
1	USC00186350	NATIONAL ARBORETUM DC, MD US	2020-01-02	NaN	NaN	0.00	0.0	0.0	NaN	50.0	27.0	28.0
2	USC00186350	NATIONAL ARBORETUM DC, MD US	2020-01-03	NaN	NaN	0.21	0.0	0.0	NaN	53.0	28.0	48.0
3	USC00186350	NATIONAL ARBORETUM DC, MD US	2020-01-04	NaN	NaN	0.11	0.0	0.0	NaN	55.0	48.0	55.0
4	USC00186350	NATIONAL ARBORETUM DC, MD US	2020-01-05	NaN	NaN	0.20	0.0	0.0	NaN	59.0	38.0	38.0

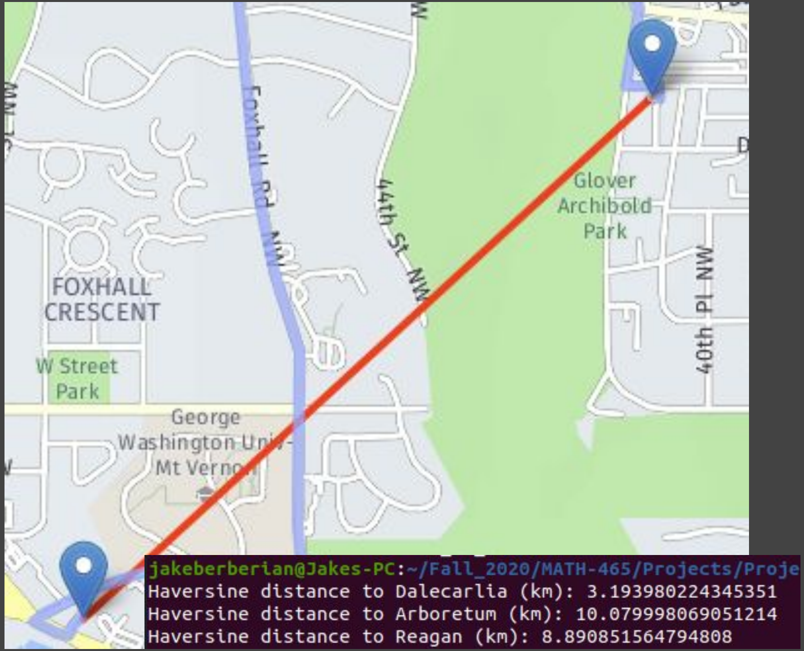
- Numerous weather stations collecting publically available data
- Find three data sources near 20007
  - National Arboretum
  - Dalecarlia Reservoir
  - Reagan National Airport

Stations



- Sources going clockwise starting at top left
  - Dalecarlia Reservoir
  - National Arboretum
  - Reagan National Airport
- Red dot indicates approximate location of PWS.
- Hypothesis about weights of regression

Return of Haversine

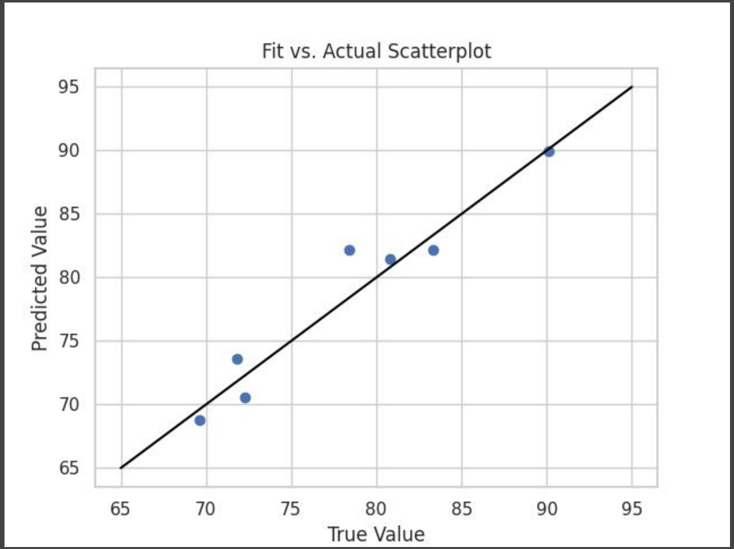


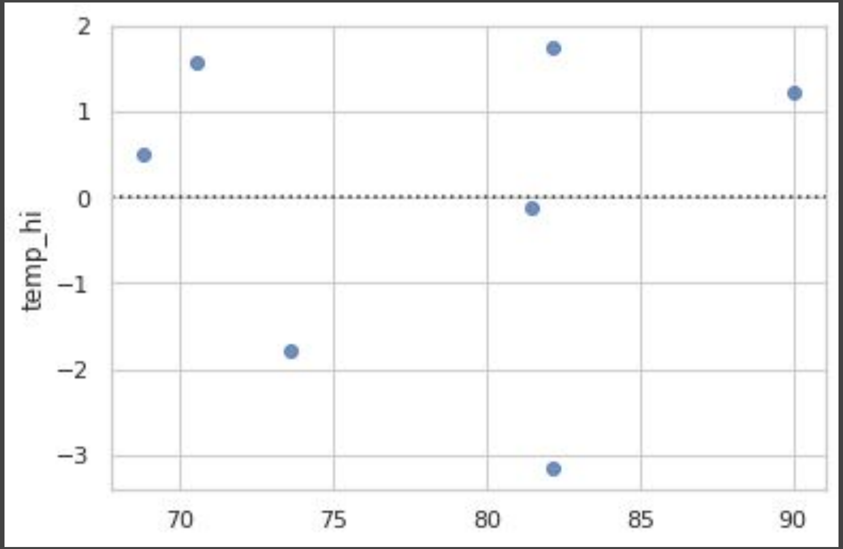
```
jakeberberian@Jakes-PC:~/Fall_2020/MATH-465/Projects/Proje
Haversine distance to Dalecarlia (km): 3.193980224345351
Haversine distance to Arboretum (km): 10.079998069051214
Haversine distance to Reagan (km): 8.890851564794808
```

- Hypotheses = which station will provide the most info (largest β_i in our regression)
- Why not Euclidean distance?
- Haversine distances
    - Dalecarlia Reservoir (3.19 km)
    - Reagan National Airport (8.89 km)
    - National Arboretum (10.08 km)

Multiple Linear Regression Output

$$\widehat{temp\_hi} = -5.29024 + 0.10998TMAX\_1 - 0.317634TMAX\_2 + 1.14581TMAX\_3$$





Further Studies

- More data
  - With n = 28, the regression output doesn’t tell us a whole lot.
  - Figuring out how to scrape sites would be great but that’s proved very difficult, if impossible
- Cross-validation*
- Other predictive techniques
- Building out a precipitation collector
- Exploratory data analysis*
- Finding other cities to do this in

* included in full report

Jupyter Notebooks and Python Scripts

- Converting .ipynb to .py files
  - Dependencies: must have jupyter installed on machine
  - Command line on Linux: jupyter nbconvert --to script [FILE NAME]

Sources

Massaron, Luca, and Alberto Boschetti. *Regression Analysis with Python: Learn the Art of Regression Analysis with Python*. Packt Publishing Ltd., 2016.

"Global Daily Summaries." *National Centers for Environmental Information*, NOAA, 1988. gov.noaa.ncdc:C01318.

StackOverflow, as always