WHAT MAKES A CITY POPULOUS ?

BY: EDMUND CHITWOOD

PHOTO BY NOFACE / PUBLIC DOMAIN

PHOTO BY NOFACE / PUBLIC DOMAIN

GOAL

STATISTICAL INFERENCE TO DETERMINE:

- FACTORS ASSOCIATED WITH HIGH POPULATION

Climate data for New York (Belvedere Castle, Central Park), 1981–2010 normals							
Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Record high °F (°C)	72 (22)	78 (26)	86 (30)	96 (36)	99 (37)	101 (38)	106 (41
Mean maximum °F (°C)	59.6 (15.3)	60.7 (15.9)	71.5 (21.9)	83.0 (28.3)	88.0 (31.1)	92.3 (33.5)	95.4 (35.2
Average high °F (°C)	38.3 (3.5)	41.6 (5.3)	49.7 (9.8)	61.2 (16.2)	70.8 (21.6)	79.3 (26.3)	84.1 (28.9
Average low °F (°C)	26.9 (−2.8)	28.9 (−1.7)	35.2 (1.8)	44.8 (7.1)	54.0 (12.2)	63.6 (17.6)	68.8 (20.4
Mean minimum °F (°C)	9.2 (−12.7)	12.8 (−10.7)	18.5 (−7.5)	32.3 (0.2)	43.5 (6.4)	52.9 (11.6)	60.3 (15.7
Record low °F (°C)	−6 (−21)	−15 (−26)	3 (−16)	12 (−11)	32 (0)	44 (7)	52 (11
Average precipitation inches (mm)	3.65 (92.7)	3.09 (78.5)	4.36 (110.7)	4.50 (114.3)	4.19 (106.4)	4.41 (112)	4.60 (116.
Average snowfall inches (cm)	7.0 (17.8)	9.2 (23.4)	3.9 (9.9)	0.6 (1.5)	0 (0)	0 (0)	0 (0)
Average precipitation days (≥ 0.01 in)	10.4	9.2	10.9	11.5	11.1	11.2	10.4
Average snowy days (≥ 0.1 in)	4.0	2.8	1.8	0.3	0	0	0
Average relative humidity (%)	61.5	60.2	58.5	55.3	62.7	65.2	64.2
Mean monthly sunshine hours	162.7	163.1	212.5	225.6	256.6	257.3	268.
Percent possible sunshine	54	55	57	57	57	57	59

Source: NOAA (relative humidity and sun 1961–1990)[213][

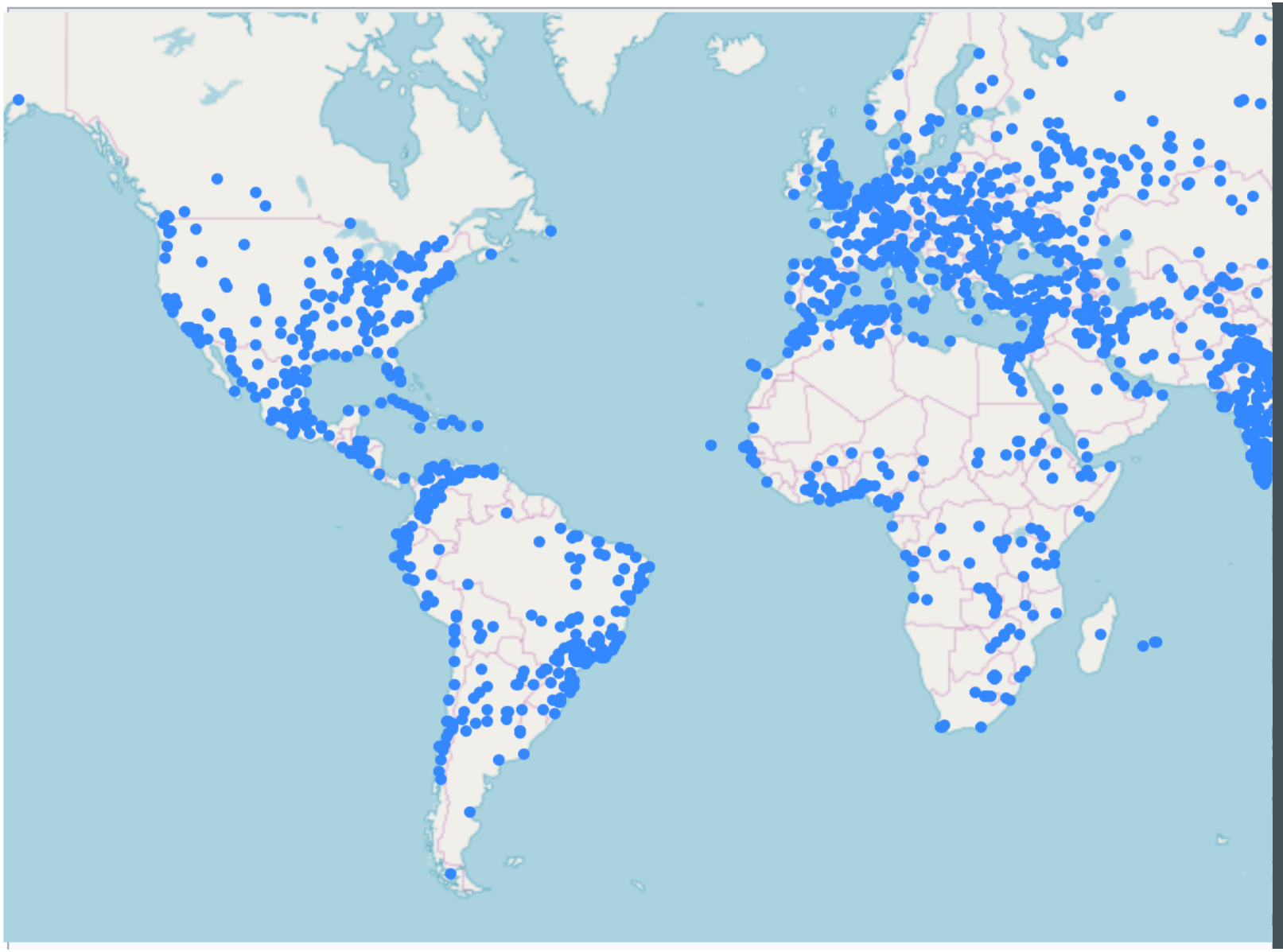
See Geography of New York City for additional climate information from

DATA SOURCE

WIKIPEDIA PAGES

- CITIES WITH POPULATION OVER 100,000
- DATA FROM INFOBOXES AND CLIMATE TABLES

PHOTO BY NOFACE / PUBLIC DOMAIN

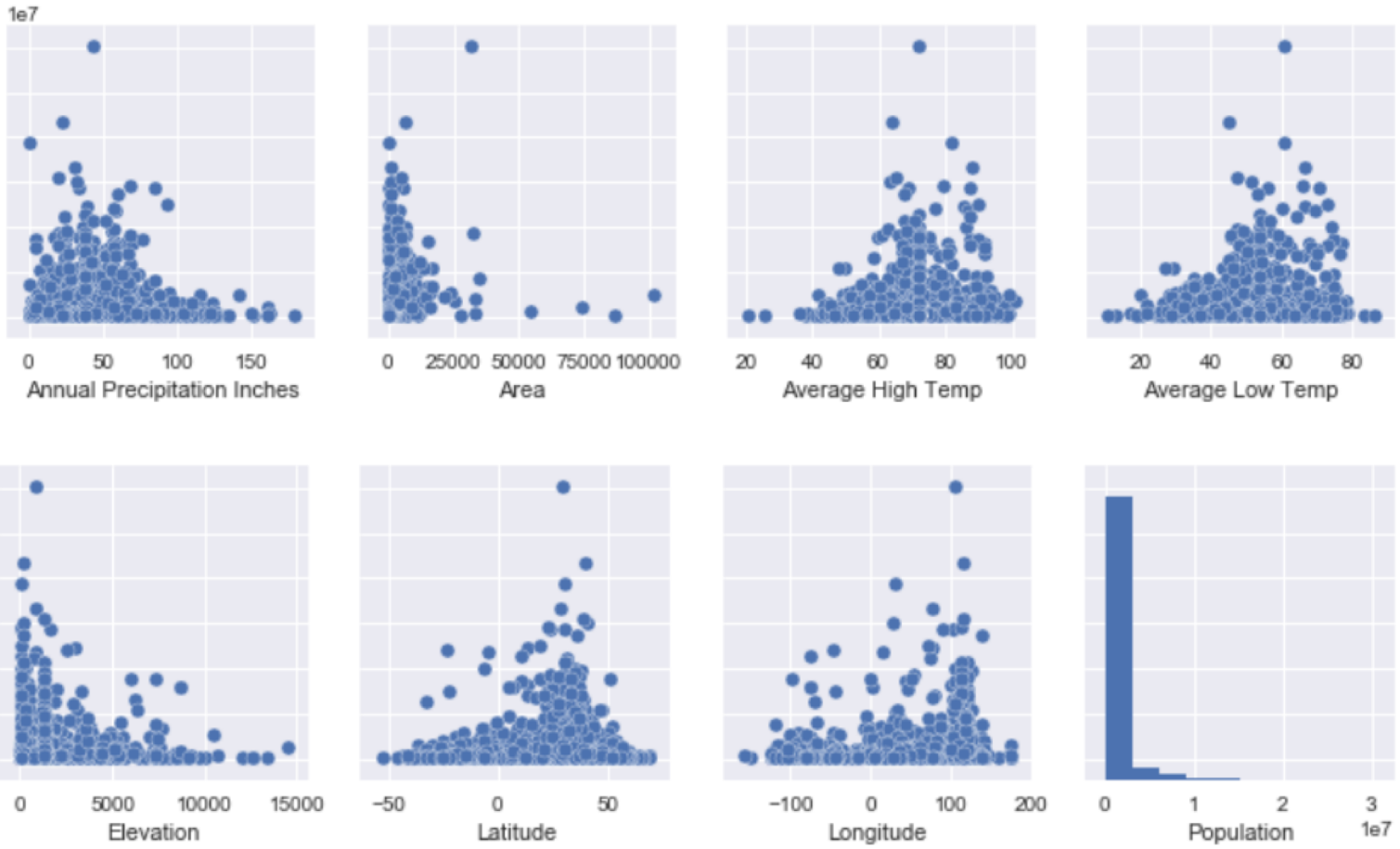


SCRAPED DATA

~3,000 PAGES

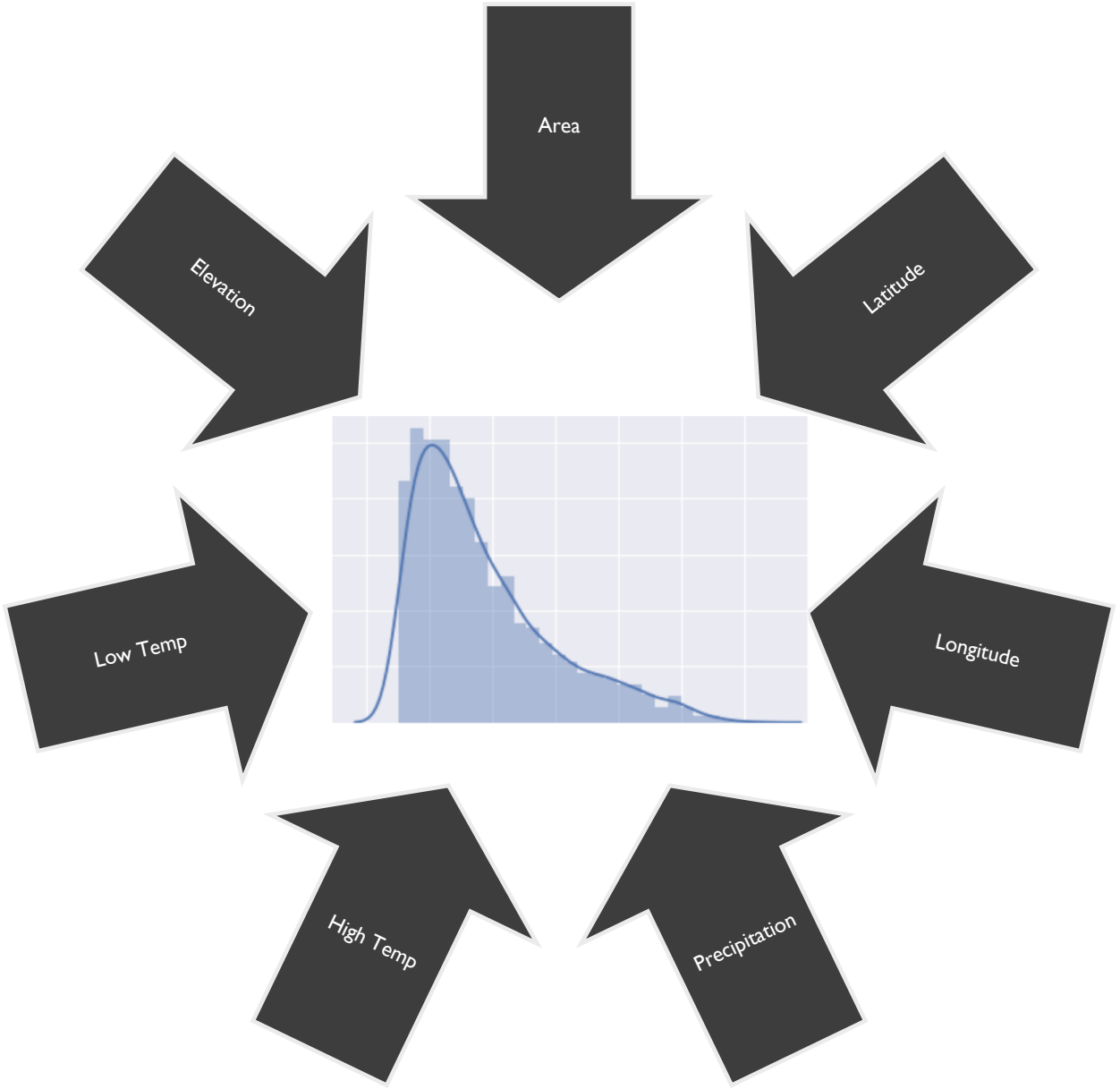
- 9 FEATURES
- TARGET: POPULATION

PHOTO BY NOFACE / PUBLIC DOMAIN



FEATURES VS POPULATION

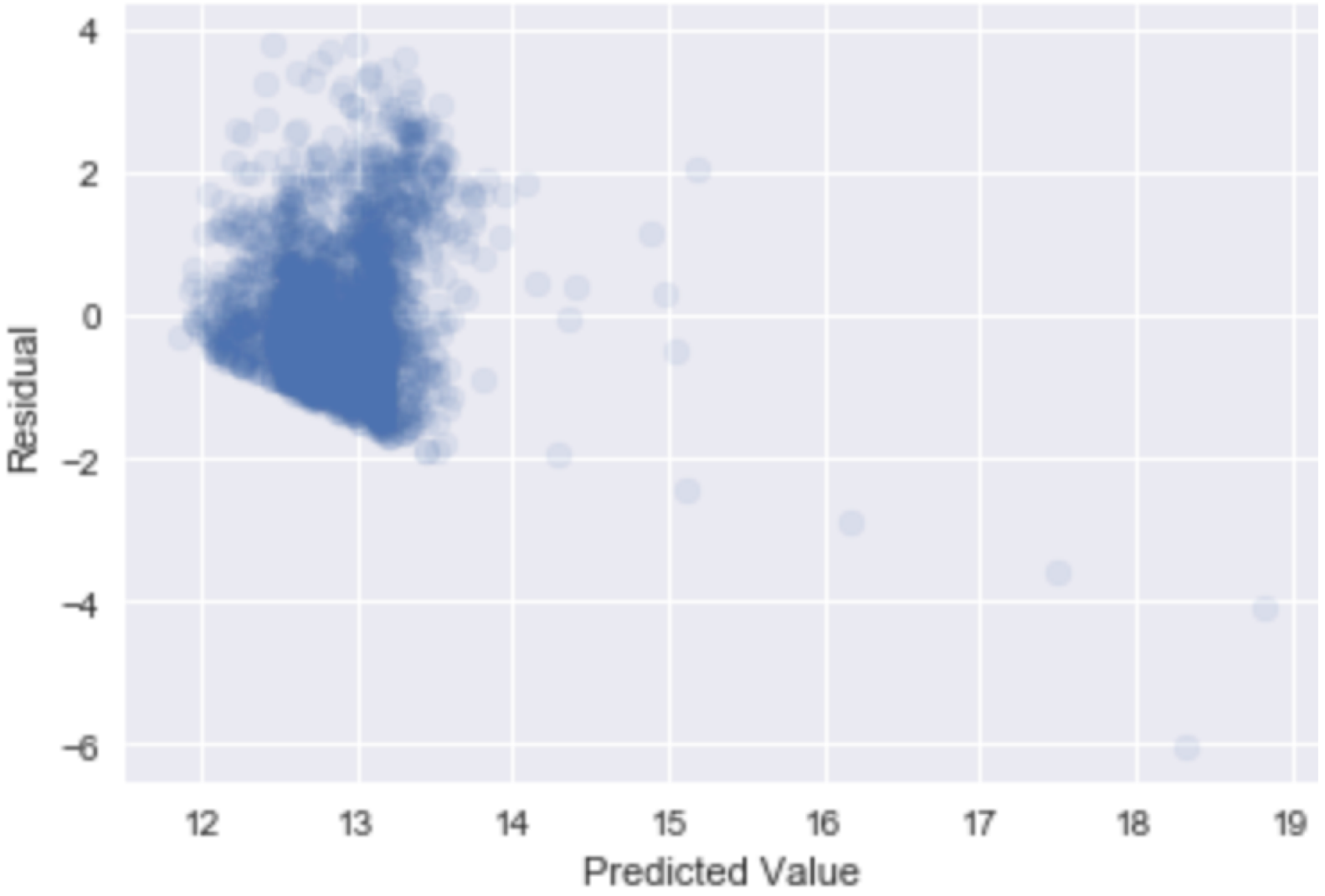
PHOTO BY NOFACE / PUBLIC DOMAIN



BASIS FOR MODELS

- 7 FEATURES
- TARGET: LOG POPULATION

PHOTO BY NOFACE / PUBLIC DOMAIN

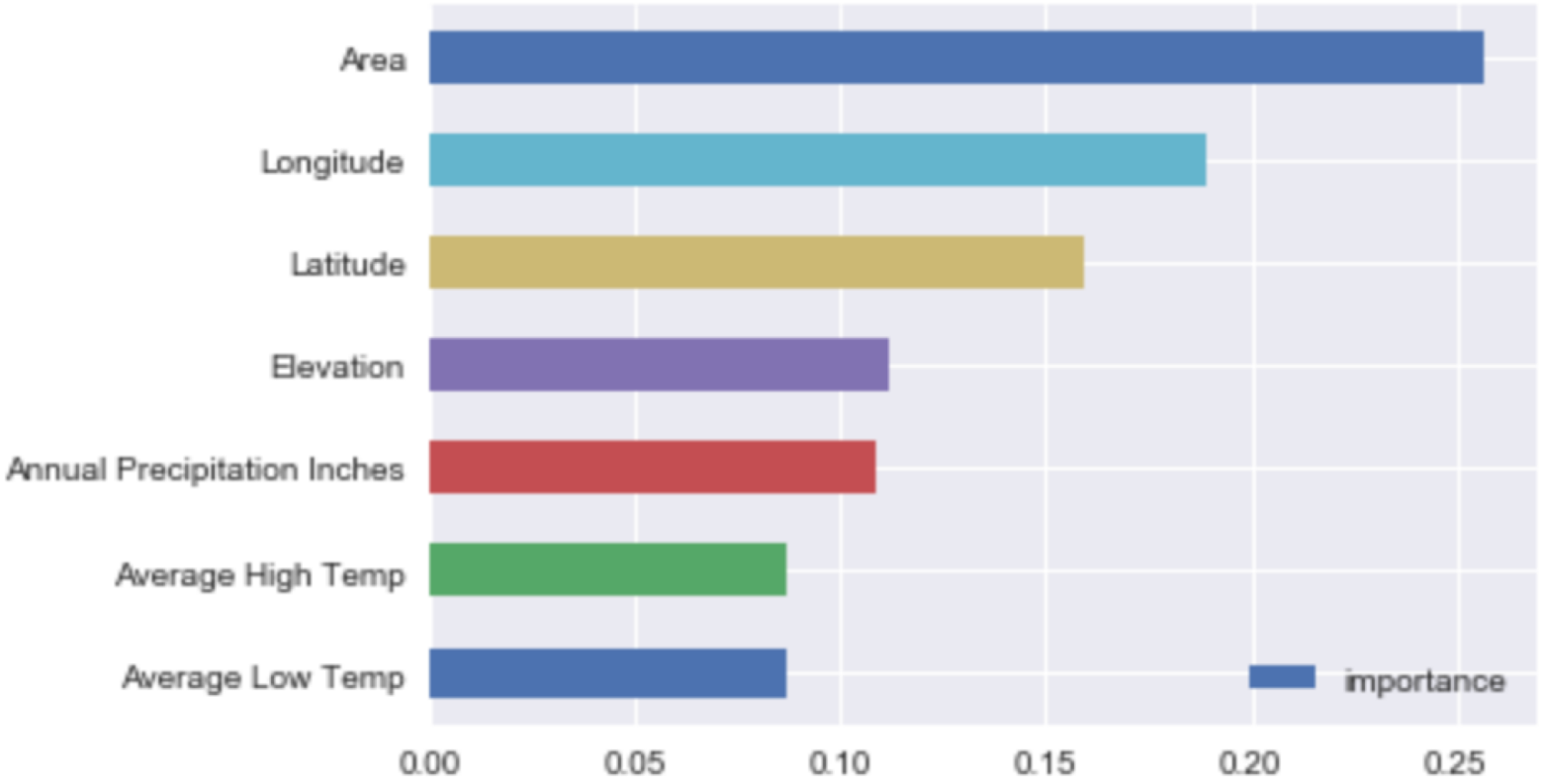


INITIAL PREDICTIONS

LINEAR REGRESSION

- R SQUARED 0.14

PHOTO BY NOFACE / PUBLIC DOMAIN

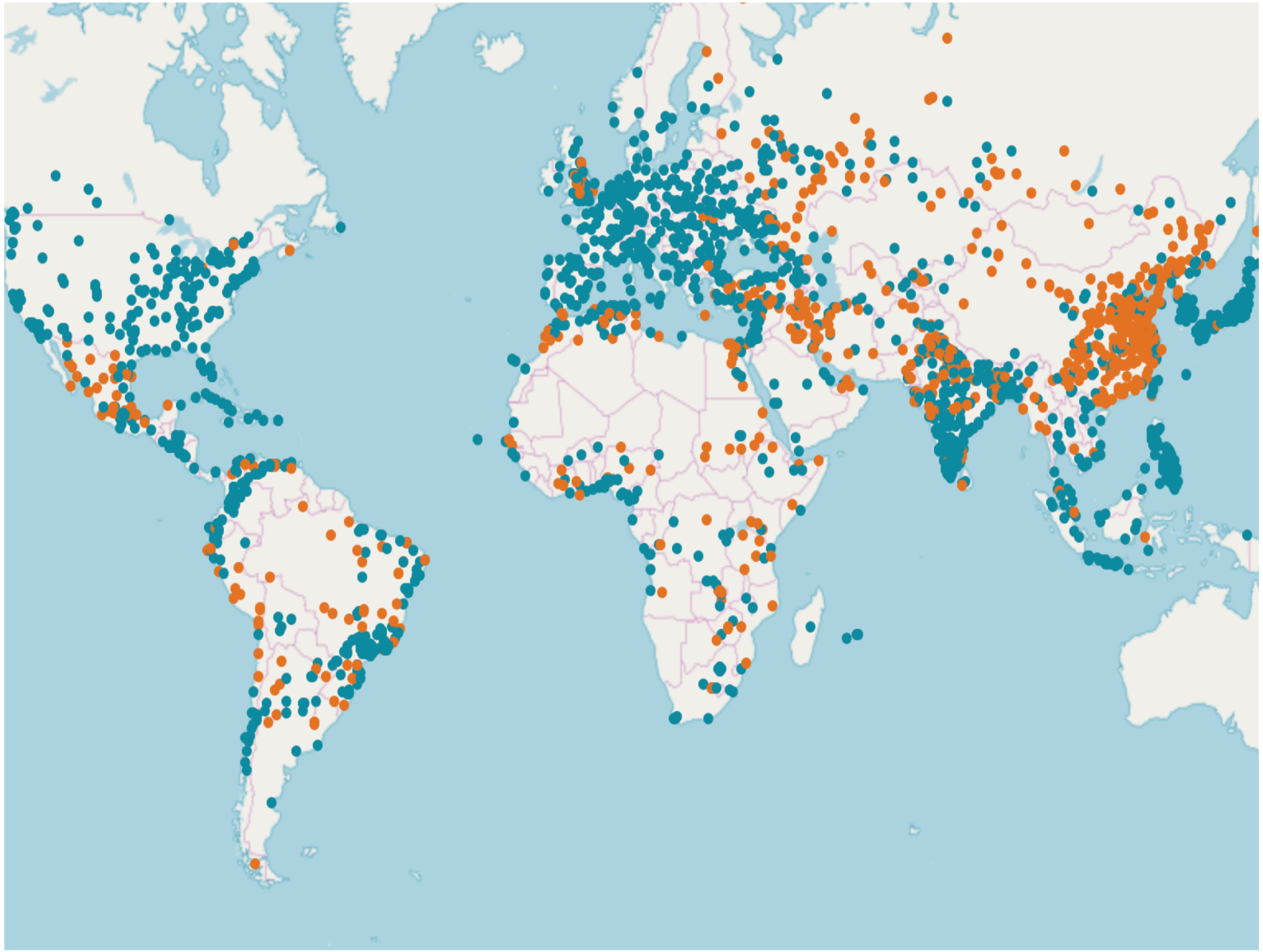


EXPLORING NONLINEAR RELATIONSHIPS

RANDOM FOREST

- R SQUARED 0.53

PHOTO BY NOFACE / PUBLIC DOMAIN

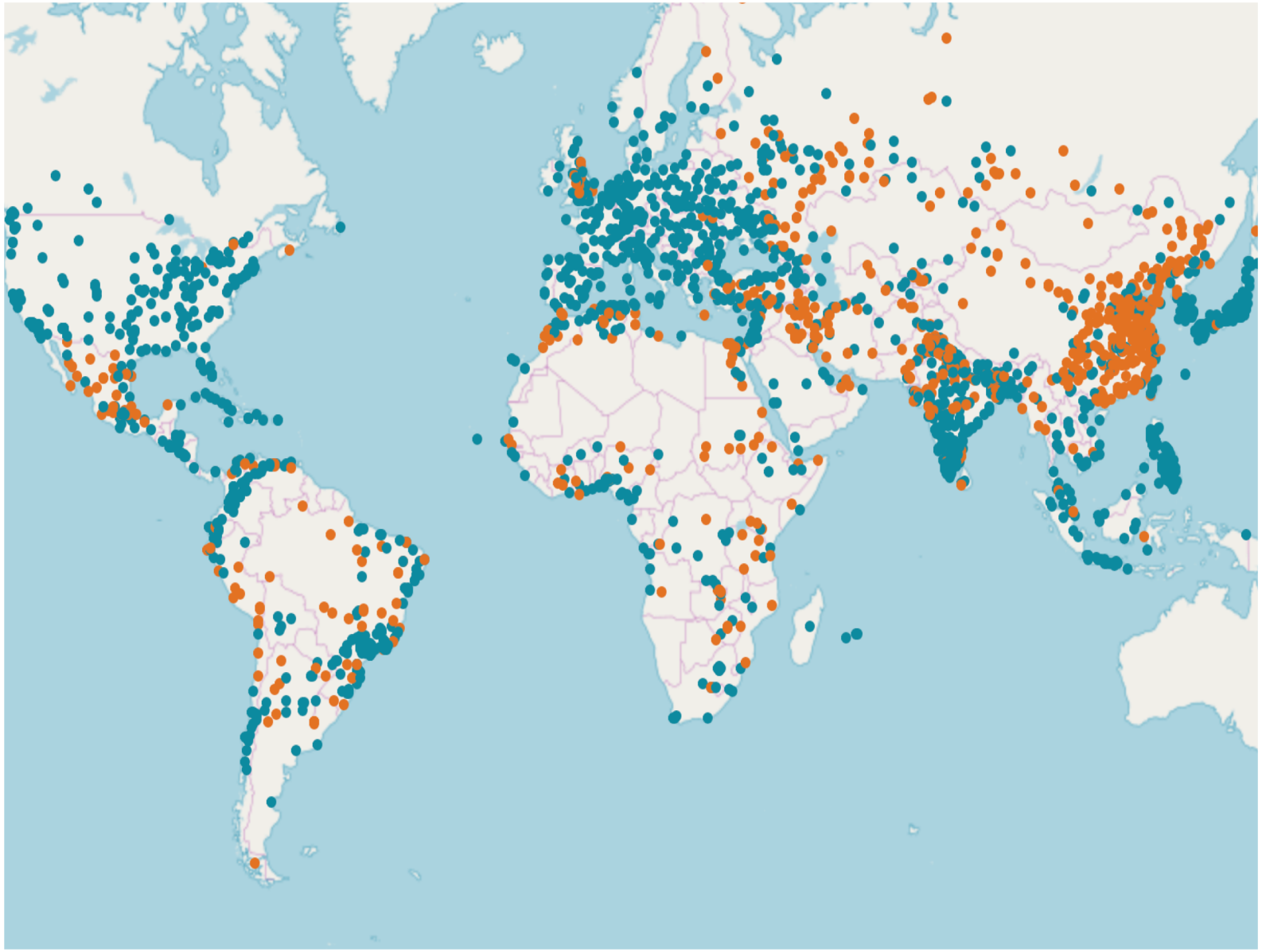


VISUALIZING THE DATA

CITIES BY AREA

- CONCENTRATIONS OF CITIES WITH LARGE AREAS (E.G. CHINA)

PHOTO BY NOFACE / PUBLIC DOMAIN



CONCLUSION AND NEXT STEPS

INFERRING POPULATION

- AREA AND LOCATION ARE MOST USEFUL AMONG PROJECT FEATURES FOR INFERRING POPULATION

EXPANDED ANALYSIS

- SOCIAL DATA FEATURES
- SMALLER LOCALITIES
- REGION SPECIFIC ANALYSIS

PHOTO BY NOFACE / PUBLIC DOMAIN