Paytm Labs Data Engineer Challenge

Read the instructions below, and complete all the challenges. Your submission must include the code used to solve this challenge.

*Disclaimer

We know that this dataset is small and can fit into any modern laptop with 2GB of ram. Before you decide to use Pandas (for good reasons), please use any modern distributed ETL (spark, flink, beam, dask, etc; spark preferred) because we expect your solution to scale to larger datasets (TB or more)!

The Challenge

The Data

The weather data is available in this repo under: `/data/2019/`. Sometimes a weather station will not take readings for every field, in those situations, the station reports all 9's for that field (exact value for missing fields are provided in the table below). Make sure to deal with the missing values correctly.

FIELD	TYPE	DESCRIPTION	MISSING
STN---	Int	Station number (WMO/DATSAV3 number) for the location.	
WBAN	Int	WBAN number where applicable--this is the historical "Weather Bureau Air Force Navy" number - with WBAN being the acronym	
YEARMODA	Int.	The year, month and day. yyyyMMdd	
TEMP	Real	Mean temperature for the day in degrees Fahrenheit to tenths	9999.9
DEWP	Real	Mean dew point for the day in degrees Fahrenheit to tenths	9999.9
SLP	Real	Mean sea level pressure for the day in millibars to tenths	9999.9
STP	Real	Mean station pressure for the day in millibars to tenths	9999.9
VISIB	Real	Mean visibility for the day in miles to tenths	999.9
WDSP	Real	Mean wind speed for the day in knots to tenths.	999.9
MXSPD	Real	Maximum sustained wind speed reported for the day in knots to tenths	999.9
GUST	Real	Maximum wind gust reported for the day in knots to tenths	999.9
MAX	Real	Maximum temperature reported during the day in Fahrenheit to tenths--time of max temp report	9999.9

		varies by country and region, so this will sometimes not be the max for the calendar day	
MIN	Real	Minimum temperature reported during the day in Fahrenheit to tenths--time of min temp report varies by country and region, so this will sometimes not be the min for the calendar day	9999.9
PRCP	Real	Total precipitation (rain and/or melted snow) reported during the day in inches and hundredths; will usually not end with the midnight observation--i.e., may include latter part of previous day. .00 indicates no measurable precipitation (includes a trace)	99.99
SNDP	Real	Snow depth in inches to tenths--last report for the day if reported more than once.	999.9
FRSHTT	String	Indicators (1 = yes, 0 = no/not reported) for the occurrence during the day of: Fog ('F' - 1st digit). Rain or Drizzle ('R' - 2nd digit). Snow or Ice Pellets ('S' - 3rd digit). Hail ('H' - 4th digit). Thunder ('T' - 5th digit). Tornado or Funnel Cloud ('T' - 6th digit). For example, 100011 means there was fog, thunder, and tornado or funnel cloud, while 010000 means there was only rain.	

Step 1 - Setting Up the Data

1. Load the global weather data into your big data technology of choice.
2. Join the stationlist.csv with the countrylist.csv to get the full country name for each station number.
3. Join the global weather data with the full country names by station number.

We can now begin to answer the weather questions!

Step 2 - Questions

Using the global weather data, answer the following:

1. Which country had the hottest average mean temperature over the year?
2. Which country had the most consecutive days of tornadoes/funnel cloud formations?
3. Which country had the second highest average mean wind speed over the year?

What are we looking for?

We want to see how you handle:

- Code quality and best practices
- New technologies and frameworks
- Messy (ie real) data
- Understanding of data transformation. This is not a pass or fail test, we want to hear about your challenges and your successes with this challenge.