assignment3

October 28, 2022

0.1 Assignment 3: Spark

Name: Bruno C. Gonzalez

0.2 Learning Outcomes

In this assignment, you will do the following:

- Import a dataset into the Databricks Spark environment
- Create tables for the data imported
- Perform basic data analysis using transformations and Spark SQL

Questions (12 marks)

1. Explain the main differences between RDDs, Dataframes and Datasets (4 marks) *Answer: The RDDs was the primary user-facing API in Spark since its inception and are defined as the distributed collection of the data elements without any schema. The Dataframes is an immutable distributed collection of data organized into named columns. The Dataset is an extension of the Dataframe with more added features like type-safety and strongly-typed.*
2. Answer the following questions:

2.1 How many sensor pads are reported to be from Poland (2 marks)

Answer: 12,744

2.2 How many different LCDs (distinct colors) are present in the dataset (2 marks)

Answer: 3

2.3 Find 5 countries that have the largest number of MAC devices used (2 marks)

Answer:

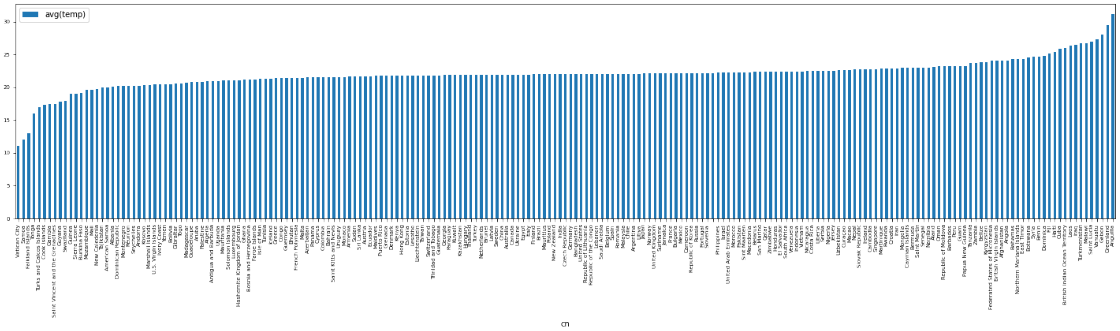
Country	Count
United States	70405
China	14455
Japan	12100
Republic of Korea	11879
Germany	7942

2.4 Propose and try an interesting statistical test or machine learning model you could use to gain insight from this dataset. Note, you don't have to use Machine Learning for this question. You can apply any analysis to the data even using SparkSQL, Python visualization libraries to analyze the data. Another example cloud be to apply correlation functions or other Spark functions to analyze the data. (2 marks)

Answer:

Summary table

summary	battery_level	c02_level	device_id	humidity	latitude	lcd	longitude	temp
count	198164	198164 198164	198164	198164	198164	198164	198164	
mean	4.4997	1199.7639	99082.5	null	61.9921	36.5211	-0.6459	22.0127
stddev	2.8733	231.06	57205.1637	null	21.6723	17.9077	88.7275	7.2098
min	0	800	1	25	-51.75	green	-175.0	10
25%	2	1000	49522	43	35.69	null	-87.69	16
50%	5	1199	99092	62	38.0	null	4.89	22
75%	7	1400	148629	81	47.0	null	100.52	28
max	9	1599	198164	99	72.0	yellow	178.42	34



Temperature by country

Correlation humidity and temperature:

−0.001 There's no correlation.

Link to databrick notebook:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/14053396234

```
[ ]:
```