**Problem statement:** Analysing changes in signal data occur due to multipath effects in both indoor and outdoor environments.

The main objectives are as follows:

1. To provide a continuous processing system (CPS) by using Spark SQL engine for structural streaming data.
2. To identify new changes in time series that occurs for streaming data.

**Reported outcomes:** Used sparklyr package in Rstudio to establish the spark connection and survey about the R packages for Data Science.

**Specific Aims:**

1. Downloaded Rstudio and installed required packages for importing the csv files and connect to apache spark.
2. Identifying the R packages for Data science.
3. Installed python packages to establish a spark session.

**Key Accomplishments:**

1. Implemented Spark Session in python notebook and able to integrate the csv file with apache spark.
2. Able to perform data filtering in python i.e., writing a SQL query to retrieve a particular attributes.
3. Implemented Spark Session in Rstudio by using sparklyr and also able to load the csv file in spark cluster.

**Red Flags:**

1. **Error occurred while connecting spark session in python as it requires winutils application and should be stored inside the spark Hadoop and path should be set properly.**
2. **Setting the valid version of java is required for spark session to work (i.e., JDK 1.8) in Rstudio and tried to use JDK 1.8 above versions but does not work in Rstudio.**
3. **Error occurred that dataset table not found in spark cluster, while filtering a particular attribute by using dplyr in Rstudio.**

**Future Work:**

1. To identify the Machine Learning packages present, especially in non-parametric approaches and try to implement in Rstudio.

**Timeline (tentative timeline for the upcoming week)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Future Specific aims** | **10/09** | **10/10** | **10/11** | **10/07** | **10/14** | **10/15** |
| Implementation of MLLib package on spark streaming (Gaussian mixture and KMeans) |  |  |  |  |  |  |
| Implementing Tidyverse package for pre-processing |  |  |  |  |  |  |
| Implementation of change detection based on non-parametric approaches |  |  |  |  |  |  |

**References:**

[1] S. Aminikhanghahi and D. J. Cook, “A survey of methods for time series change point detection,” *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, Aug. 2016.

[2] A. Bifet, G. Holmes, and B. Pfahringer, “MOA-TweetReader: Real-Time Analysis in Twitter Streaming Data,” *Discovery Science Lecture Notes in Computer Science*, pp. 46–60, 2011.

[3] “Analysis of real-time data with spark streaming,” *Journal of Advances in Technology and Engineering Research*, vol. 3, no. 4, 2017.

**Appendix A**

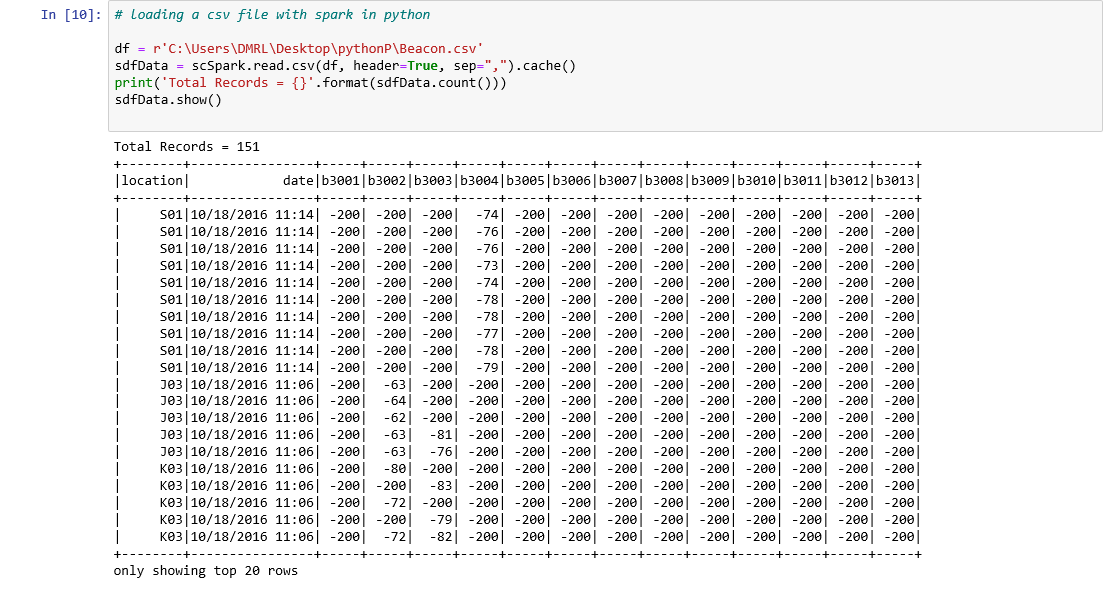
**pyspark: Python Interface for Apache Spark**

**Required Packages:** pyspark, sparkSession and SQLContext.

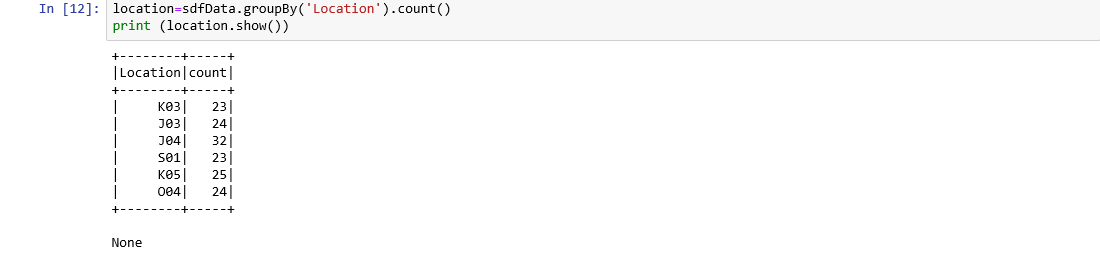
**Screenshots for Spark Session in python:**



**Loading csv file with apache spark in python:**



**Simple SQL query to retrieve number of locations present in BLE dataset:**

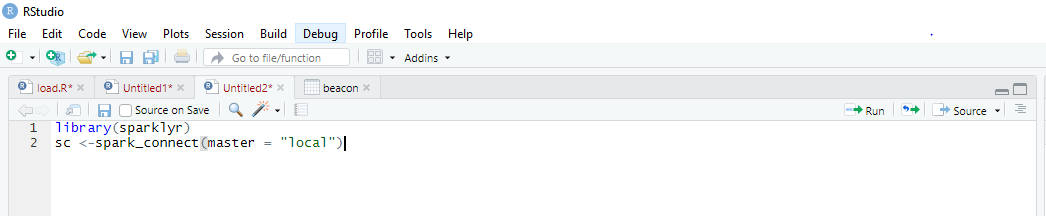


**Appendix B**

**sparklyr: R interface for Apache Spark**

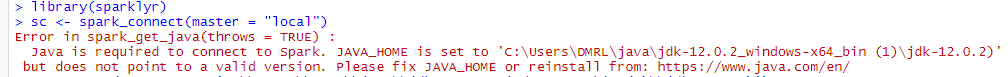
**Prerequisites:**

1. Install the sparkly package from CRAN in Rstudio.
2. Install a local version of spark for development purposes.



1. Need to connect to a local instance of Spark via the **spark\_connect** function.

**Issues Raised:**



**Issue Solved:**

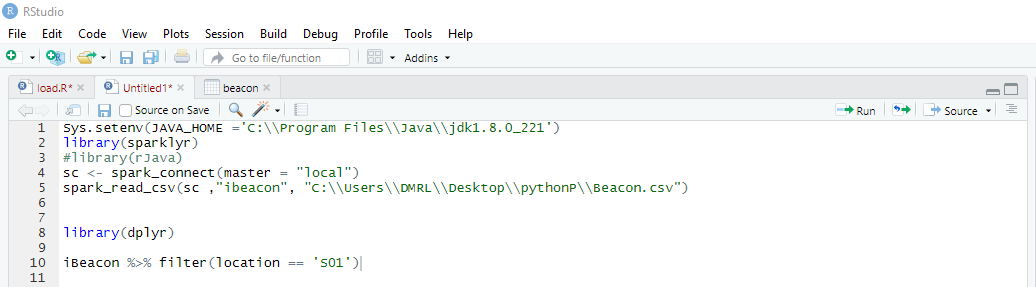
By providing the java directory in Rstudio by using Sys.setenv(JAVA\_HOME =’ java path’) and make sure Java(JDK 1.8 ) version is installed as it works for only particular version.

**Using dplyr – A grammar for Data Manipulation:**

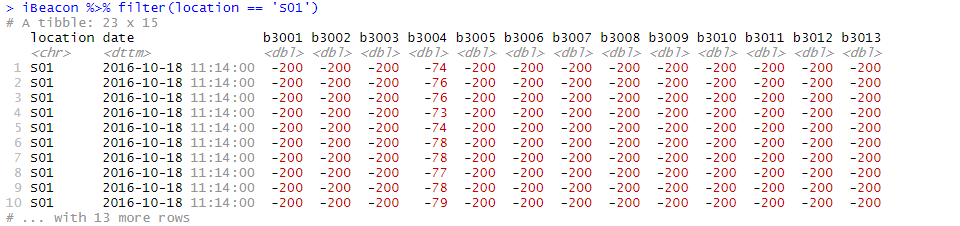
A fast and consistent toll for working with data frame like objects both in memory and out of memory.

**List of functions dplyr offers:**

* 1. **Select():Select columns from the dataset.**
  2. **Filter():filter out certain rows and columns.**



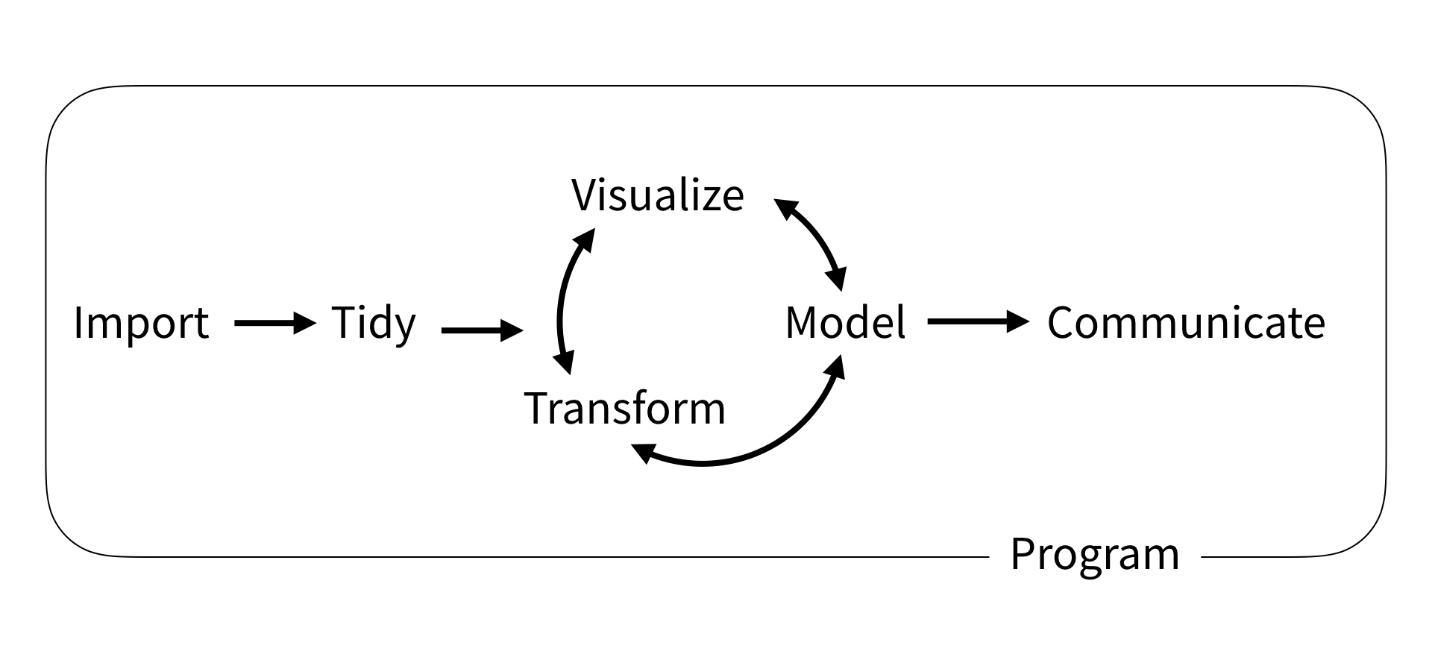
**Output:**



**Appendix C**

**R packages for Data Science**

**Tidyverse** is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar and data structures.

****

**Fig 1,** Structure of Tidyverse packages

**Required packages to be installed:**

