Process WiFi

Folder – SCRIPTS

Source File – Mod3\_Task03\_Script00\_Project

Subfolder -

1 – Preparation

Script01\_1\_00-Import

Script01\_2\_00-Inspection

Script01\_3\_00-Cleaning --->

* + - Change WAPS with value 100 to -110
    - Eliminate Duplex rows
    - Eliminate Duplex Columns
    - Set and take out zero vars
    - Trim white spaces
    - Check missing Values
    - Set the Correct Data type

2 – Pre-Process

Script02\_1\_00-Manipulation

* Recoding Building and Floor IDs
* Create ID Field
* Create df with only WAP fields – wap\_df
* Record tries, valid, invalid accesses by Row
* Change Data type to characters FLOOR, BUILDINGID, POSITION, SPCEID, USERID
* Change column name from RELATIVEPOSITION to POSITIONID
* Drop PHONEID
* Create new char columns
* Create SIGNAL\_QUALITY column as namespace with default value 1
* Create ZoneID and LocID levels
* Compute MAX and MIN valid access columns
* Create COORD\_POINT from LONG, LAT and FLOOR
* Assign levels to SIGNAL\_QUALITY (10 = excellent,

8 = Good,

6 = Fair,

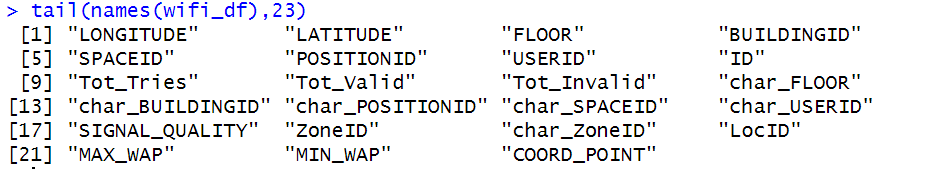
5 = Poor,

3 = No Signal,

1 = Not accessed

* Summarize SIGNAL\_QUALITY and MAX\_WAP + SIGNAL\_QUALITY
* New Datasets with specific information
* rm\_df – Takes out all MAX\_WAP with records of -110
* eda\_qnt\_df – Stores all quantative vars
* eda\_qlt\_df – Stores all qualitative vars
* **wifi\_df** – Base df for further modelling and processing. To be created after Manipulation
* Write a clean and modified Dataset

Fig 1 – wifi\_df fields in addition to WAPs



Script02\_Plots

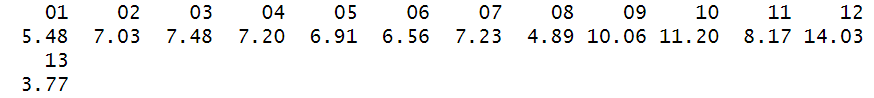
Testing-Code\_File

3 - EDA (Engineering Data Analysis)

* CLUSTERING ANALYSIS
* K-MEANS
* HIERARCHY

Script03\_1\_00-Understand\_The\_Data

* Quantify number of Levels for possible response vars (Bld, Flr, Zone, Loc, Quality, Space, Position, Coord)
* Number of instances (rows) that belong to each possible predictor class



Possible response vars: BUILDING, FLOOR, POSITIONID, SPACEID, SIGNAL QUALITY, USERID for classification (LocID and CoordID have too many levels)

Possible response vars for regression: Longitude, Latitude, Max\_WAP, Total\_Valid, Total\_Invalid

Distribution of Users by Space ID – Strong concentration of Users within space id 100 ~150

Distribution of users by Floors – Greater number concentrated on 4th Floor

Highest count of accessed WAPs by ZoneId – Highest recording in Zone 12, indicating Building 3 .. 4th Floor

Run Plots on Script03\_1\_01\_Plotting to view conclusions

Explore the relationship between the quantative vars

Long and Latitude, Tot\_Valid and FLOOR, SPACEID and USERID, FLOOR and USERID

Plotting Categorical Count Variables

Identify Zones by their ID

View USERID based on Position and sized by Tot\_Valid. Facet by Bld and FLr

Summarize the qualitative data - charSummary(eda\_qlt\_df)

Summarize the quantative data - numSummary(eda\_qlt\_df)

Assign the USERID within the ZoneID Area - bivariate(eda\_qlt\_df,"char\_USERID","ZoneID")

Create an NA df

valid\_WAP\_df- Gather NA df to long form and remove the NAs

nonValid\_WAP\_df - Gather NA df to long form and leave the NAs

invalid\_WAP\_df – Select with is.na all invalid Accesses

vw\_df- Count number of valid Accesses by WAP Point and transform into dataframe

Visualize top 10 valid WAPS

Visualize distribution of original wifi\_df valid accesses – left skewed

Boxplot all Valid accesses to view statistical distribution of Valid accesses

Perform the same operations with the non valid accesses

Merge Valid and non-valid data frames and obtain a table for confrontation for each WAP

Plot Percentage over 30 of valid accesses

signal\_WAP\_df – Long Form of gathered WAPs

valid\_signal\_WAP\_df – Filter only valid signals (363973 Total valid accesses)

Visualize and note that it is left skewed

Drop any columns that exceed the observed lower and upper whiskers adjustments based on the observation of the graphs. Limits set at -95 and -33

**base\_df** - Create a new normally distributed set to perform the modelling