Indoor Positioning Data Exploration

STAT4/510 Wasabees Group

2023-11-30

1. Describe your data
2. Describe the basic variable component of your data
3. Report the various findings you have established so far, with interpretation (include discussion on what you find and add how useful it is to your project objective)
4. Discuss any challenges you encountered, and ways by which you handled these.

1. Introduction - Data Description

Indoor position systems (IPS) development is an active area of research that can be used in numerous settings. An area of interest is the use of using signal strength from WIFI routers to estimate the location of a device.

The following report, describes and characterizes a large data set compiled in a 15 by 36 meter area that contains the data obtained from a handheld device connected to a WIFI network, in different locations and orientations, in order to create a model to predict indoor positioning.

The data is subdivided in two sub-sets, one denominated `offline` data, which corresponds to a testing device connected to the network at different locations and orientations, and the other an `online` data, where 60 locations and orientations of the devices were selected at random.

The `offline` data, intended to train a model, was collected designing a 1 meter resolution grid, resulting in 166 locations. In each of these locations, the device was oriented starting at 0 degrees inclination and at 45 degrees increments (for a total of 8), and the strength signal measured for each access point was measured 110 times. That is, per each location (x,y) we have 110 samples at each angle, for a total of 880 samples per location, and a total of 1.4608×10^5 observations.

The `online` data was designed to simulate real-world data (i.e., locations that are not bounded to a grid, and which a device can be oriented at random.) Specifically for the online data, 60 combinations of orientation/locations were randomly selected, and then sampled 110 times, resulting in 6600 measurement in total.

More details of the floor plan, and location of `online` and `offline` data can be seen in Figure 1.Circles serve as markers for the positions where `offline` measurements were conducted, while black squares indicate the locations of the six access points. The positions of the access points were provided in a separate file by the client.

For simplicity, this report will share the results found in the `offline` data set, but initial process of data cleaning can be directly applied to the `online` data as well because both sets share the same format.

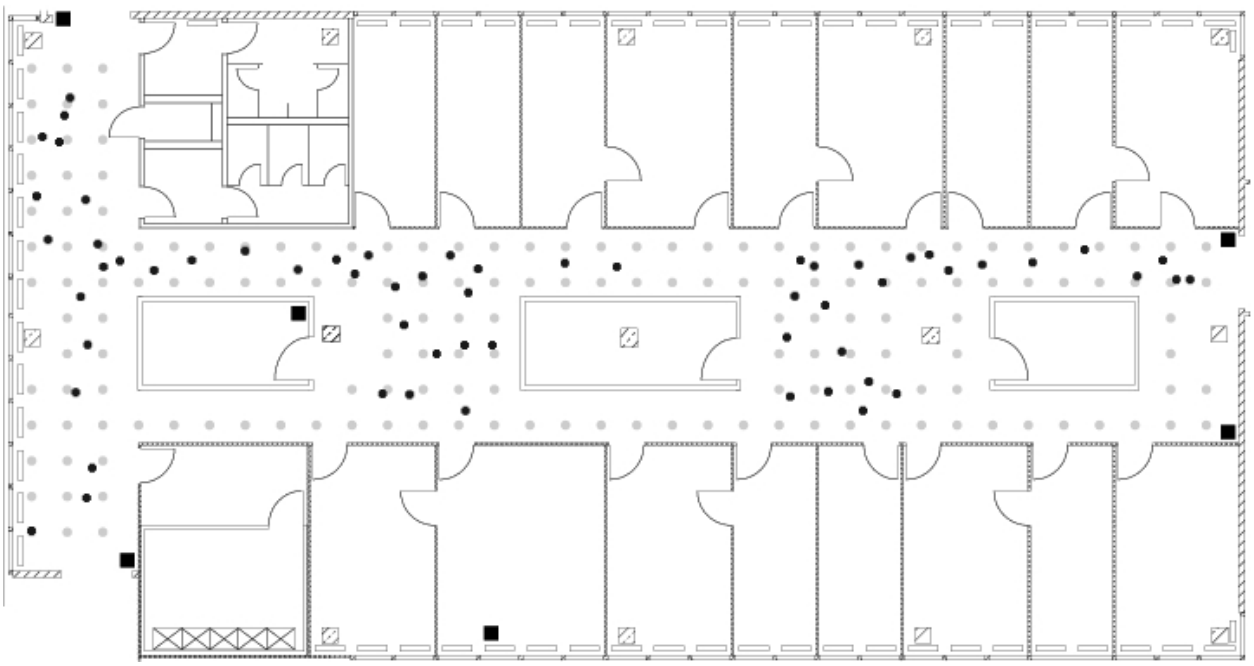

Figure 1: Flooplan location. Access points are squares. Grey dots are offline data locations and black dots are online data locations.

2. Data Processing - Describe the basic variable component of your data

This section provides the description of the data itself, and summarizes the steps of cleaning the data set, including basic statistic exploration.

2.1 Variable Description

According to documents provided by the client, the data contains the following variables:

- `time`: time in miliseconds since midnight 01/01/1970 UTC
- `scanMac`: IP address of the scanning device, in mm:mm:mm:ss:ss:ss.
- `pos`: the 3-D coordination of the scanning device (x,y,z)
- `orientation`: the scanning device's orientation.
- `mac`: the IP address of the access points.
- `signal`: signal strength in dBm.
- `channel`: the channel frequency.
- `type`: type of device (access point = 3, device in adhoc mode =1)

2.2 Data formatting

The raw data are stored in a .txt file. The first 3 lines of the data are characterized by the hash (#) symbol, followed by a row that contains all the variables in one line, separated by a semi-colon. A sample of the initial format of the data can be seen below. We start by eliminating the rows that start with the hash symbols from the data set using a `strsplit` function. The resulting data set contains a total of 146080 rows, and therefore 5312 rows were eliminated. This value, if divided by 3 (first 3 rows contain a hash) is 1770, which is close consistent with the expected number of locations (166) and angles (8). This means that each stack location/orientation combination that contains 110 samples (from herein `Location_orientation stack`) was separated by 3 hash symbols.

```
 [1] "t=1139643118358;"                    "id=00:02:2D:21:0F:33;"
 [3] "pos=0.0,0.0,0.0;"                    "degree=0.0;"
 [5] "00:14:bf:b1:97:8a=-38,2437000000,3;" "00:14:bf:b1:97:90=-56,2427000000,3;"
 [7] "00:0f:a3:39:e1:c0=-53,2462000000,3;" "00:14:bf:b1:97:8d=-65,2442000000,3;"
 [9] "00:14:bf:b1:97:81=-65,2422000000,3;" "00:14:bf:3b:c7:c6=-66,2432000000,3;"
[11] "00:0f:a3:39:dd:cd=-75,2412000000,3;" "00:0f:a3:39:e0:4b=-78,2462000000,3;"
[13] "00:0f:a3:39:e2:10=-87,2437000000,3;" "02:64:fb:68:52:e6=-88,2447000000,1;"
[15] "02:00:42:55:31:00=-84,2457000000,1"
```

Some variable names do not correspond to the ones given by the client, for example, `orientation` is `degree` in the data set, and the variables `type`, `channel`, `signal` do not have an explicit name. We also note that for `pos` the x,y,z variables are grouped together, and the `mac` variable includes `signal`, `channel` and `type` separated by comas. We can then distinguish between single variables (defined by a name and have one value), and secondly composite variables (defined by containing multiple values for one parameter).

The results of the `strplt` function on one row can be seen bellow.

```
 [1] "t"                 "1139643119002"     "id"
 [4] "00:02:2D:21:0F:33" "pos"               "0.0"
 [7] "0.0"               "0.0"               "degree"
[10] "0.0"               "00:14:bf:b1:97:8a" "-38"
[13] "2437000000"        "3"                 "00:0f:a3:39:e1:c0"
[16] "-54"               "2462000000"        "3"
[19] "00:14:bf:b1:97:90" "-57"               "2427000000"
[22] "3"                 "00:14:bf:b1:97:81" "-66"
[25] "2422000000"        "3"                 "00:14:bf:3b:c7:c6"
[28] "-69"               "2432000000"        "3"
[31] "00:14:bf:b1:97:8d" "-70"               "2442000000"
[34] "3"                 "00:0f:a3:39:e0:4b" "-78"
[37] "2462000000"        "3"                 "00:0f:a3:39:e2:10"
[40] "-83"               "2437000000"        "3"
[43] "00:0f:a3:39:dd:cd" "-65"               "2412000000"
[46] "3"                 "02:64:fb:68:52:e6" "-90"
[49] "2447000000"        "1"
```

We use these patterns to create a matrix with the variables. For this, we created a function that first, separated all the data separated by a semi-colon, a comma, or an equal symbol. Then, we selected the rows corresponding to `mac`, `signal`, `type` and created a matrix that has the information for the specific access point. Lastly, we bind all the information together in a large data frame that contains one row per location/orientation and access point.

We provide the structure of our data frame, along with the first 3 observations below.

```
          time           scanMac posX posY posZ orientation               mac
1 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:8a
2 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:90
3 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:0f:a3:39:e1:c0
  signal    channel type
1    -38 2437000000    3
2    -56 2427000000    3
3    -53 2462000000    3
```

2.3 Data Transformation

Before further exploration and analysis of the data, we conducted converted the variables into the correct types (as defined by the documents provided by the client).

The summary is as follows:

1) The variables `position`, `orientation`, `signal` and `channel` were converted to numerical values.
2) The variable `time` was converted into a time value using as origin midnight on January 1st, 1970. The original variable was kept in the data set as `rawTime` in case it becomes necessary for future analysis.
3) The variable `type` has binary values of 1 and 3. The documentation explains that `type = 3` corresponds to ad-hoc devices, that are not needed for the development and testing of the IPS and therefore, after removing the rows with a value of `type` equal to 3, we remove the variable from the data set.
4) For the exploration, we remove the `scanMac`, as information given by the client indicates that one devices was used.

The first rows of the formatted data can be seen in Table 1.

A quick analysis of the numerical data shows that posZ has only zero values (Table 2). This seemingly anomalous value is due to the fact that all of the readings were taken on one floor of the building. We, therefore, removed the posZ variable from the data set. Furthermore, we detect anomalous values for orientation that we report in the next section.

3. Data exploration

This section focus on the exploration of the data itself.

3.1 Orientation of hand-held devices

As shown in Table 2, the max value for `orientation` corresponded to 355.9 degrees. As provided by the client, the orientation of the hand-held device going in increments of 45 degrees. However, in practice, the measured orientations slightly deviate from these eight values (Fig. 2). We attribute this to human or measurements error, and decide to modify the data and approximate the orientations to those closer to their 45 degree angles.

Table 1: Data with transformed variables

time	posX	posY	orientation	mac	signal	channel	rawtime
2006-02-10 23:31:58	0	0	0	00:14:bf:b1:97:8a	-38	2.437e+09	1.139643e+12
2006-02-10 23:31:58	0	0	0	00:14:bf:b1:97:90	-56	2.427e+09	1.139643e+12
2006-02-10 23:31:58	0	0	0	00:0f:a3:39:e1:c0	-53	2.462e+09	1.139643e+12
2006-02-10 23:31:58	0	0	0	00:14:bf:b1:97:8d	-65	2.442e+09	1.139643e+12
2006-02-10 23:31:58	0	0	0	00:14:bf:b1:97:81	-65	2.422e+09	1.139643e+12
2006-02-10 23:31:58	0	0	0	00:14:bf:3b:c7:c6	-66	2.432e+09	1.139643e+12

Table 2: Numerical Data Summary

	posX	posY	posZ	orientation	signal	channel
	Min. : 0.00	Min. : 0.000	Min. :0	Min. : 0.0	Min. :-99.0	Min. :2.412e+09
	1st Qu.: 2.00	1st Qu.: 3.000	1st Qu.:0	1st Qu.: 90.0	1st Qu.:-69.0	1st Qu.:2.422e+09
	Median :12.00	Median : 6.000	Median :0	Median :180.0	Median :-60.0	Median :2.432e+09
	Mean :13.52	Mean : 5.897	Mean :0	Mean :167.2	Mean :-61.7	Mean :2.435e+09
	3rd Qu.:23.00	3rd Qu.: 8.000	3rd Qu.:0	3rd Qu.:270.0	3rd Qu.:-53.0	3rd Qu.:2.442e+09
	Max. :33.00	Max. :13.000	Max. :0	Max. :359.9	Max. :-25.0	Max. :2.472e+09

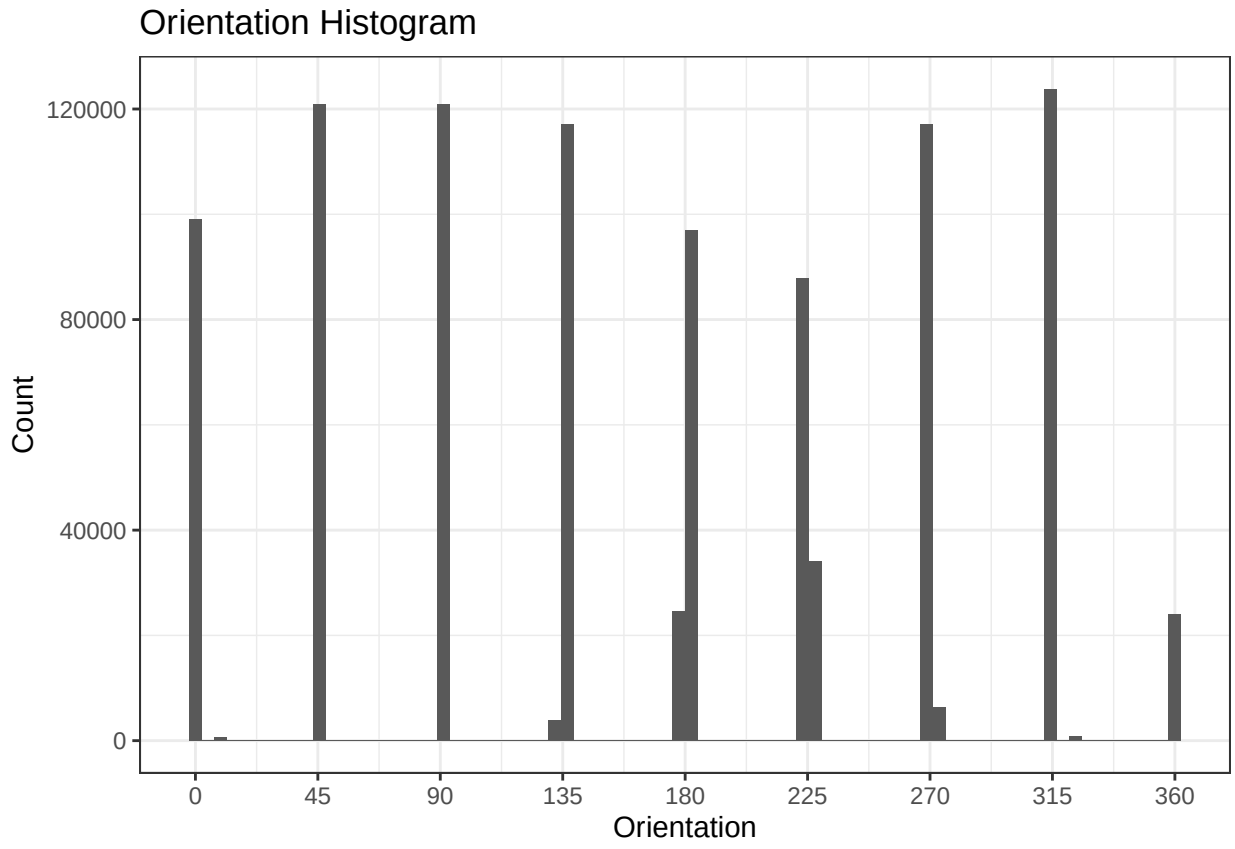

Figure 2: The figure corresponds to a histogram of the orientation. Here, it can be seen that some orientation fall outside the 45 degree increments, most likely because of errors in measurements or problems in having a consistent angle

Table 3: MAC addresses provided by client

MAC ID
Macs
00:0f:a3:39:e1:c0
00:14:bf:b1:97:8a
00:14:bf:3b:c7:c6
00:14:bf:b1:97:90
00:14:bf:b1:97:8d
00:14:bf:b1:97:81

Table 4: Number of observation by MAC address

MAC ID	Observations
00:0f:a3:39:e1:c0	145862
00:0f:a3:39:dd:cd	145619
00:14:bf:b1:97:8a	132962
00:14:bf:3b:c7:c6	126529
00:14:bf:b1:97:90	122315
00:14:bf:b1:97:8d	121325
00:14:bf:b1:97:81	120339
00:0f:a3:39:e0:4b	43508
00:0f:a3:39:e2:10	19162
00:04:0e:5c:23:fc	418
00:30:bd:f8:7f:c5	301
00:e0:63:82:8b:a9	103

3.2 Access points

The client provided a list of six access point locations (Table 3). However, the data collected includes 12 access points. This problem arises from the scanning device catching signals from wifi routers on other floors and nearby. We tally the number of observations (Table 4) for each access point and found seven access points with significantly higher number of observations than other access points.

We confirm the corresponding access points with the data provided by the client on the "Access Point Location" file and filter accordingly. Additionally, we add the data of the position values for the access point provided by the client to the data set. This allows us to track the location of the place where the sample was taken and the location of the access device emitting the signal (Fig. 3).

This will be useful for creating the model, as it will allow us to calculate the distance between the access point, the hand-held device, and signal strength.

To confirm the list of access points, we match them with the channel's frequency (Table 5). Since each access point coordinates with only one channel, we do not need the `channel` variable for our model.

4. Visualization

We first explore the relationship between number of measurements and position in the floor plan. Figure 4 depicts the relationship between position on the x, y axis and the number of observations. We observe that the observations are not equally distributed. It's possible that part of the information in the sample was tainted by significant noise that was subsequently removed during the exploration process.

Additionally, we plot the locations on the floor map, and confirm they overlap (Fig 5).

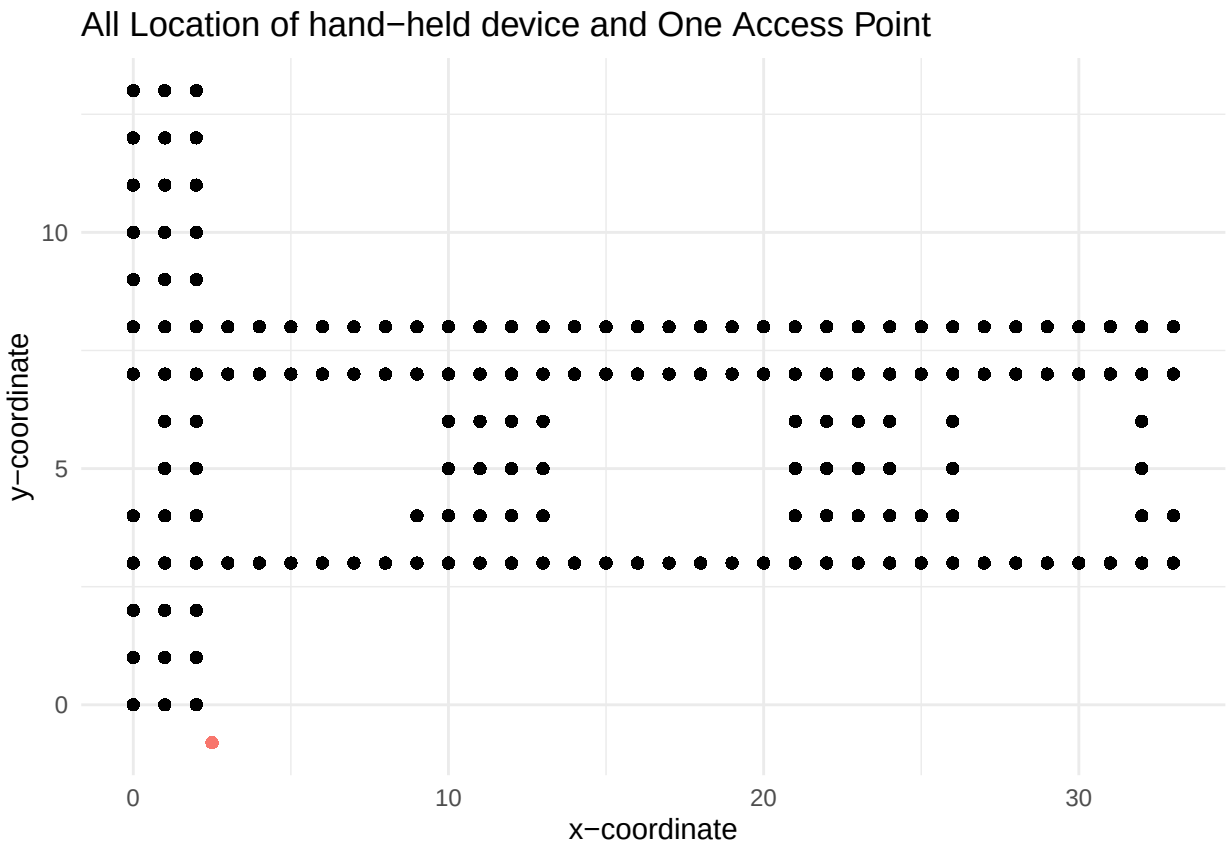

Figure 3: Location of Hand-hel devices (black) and one access point (red). Adding the access point location makes easy to determine the distance between device and access point and relate it to signal strength and orientation

Table 5: Grouped MAC addresses and corresponding channel

MAC ID	Channel
00:0f:a3:39:e1:c0	2.462e+09
00:14:bf:3b:c7:c6	2.432e+09
00:14:bf:b1:97:81	2.422e+09
00:14:bf:b1:97:8a	2.437e+09
00:14:bf:b1:97:8d	2.442e+09
00:14:bf:b1:97:90	2.427e+09

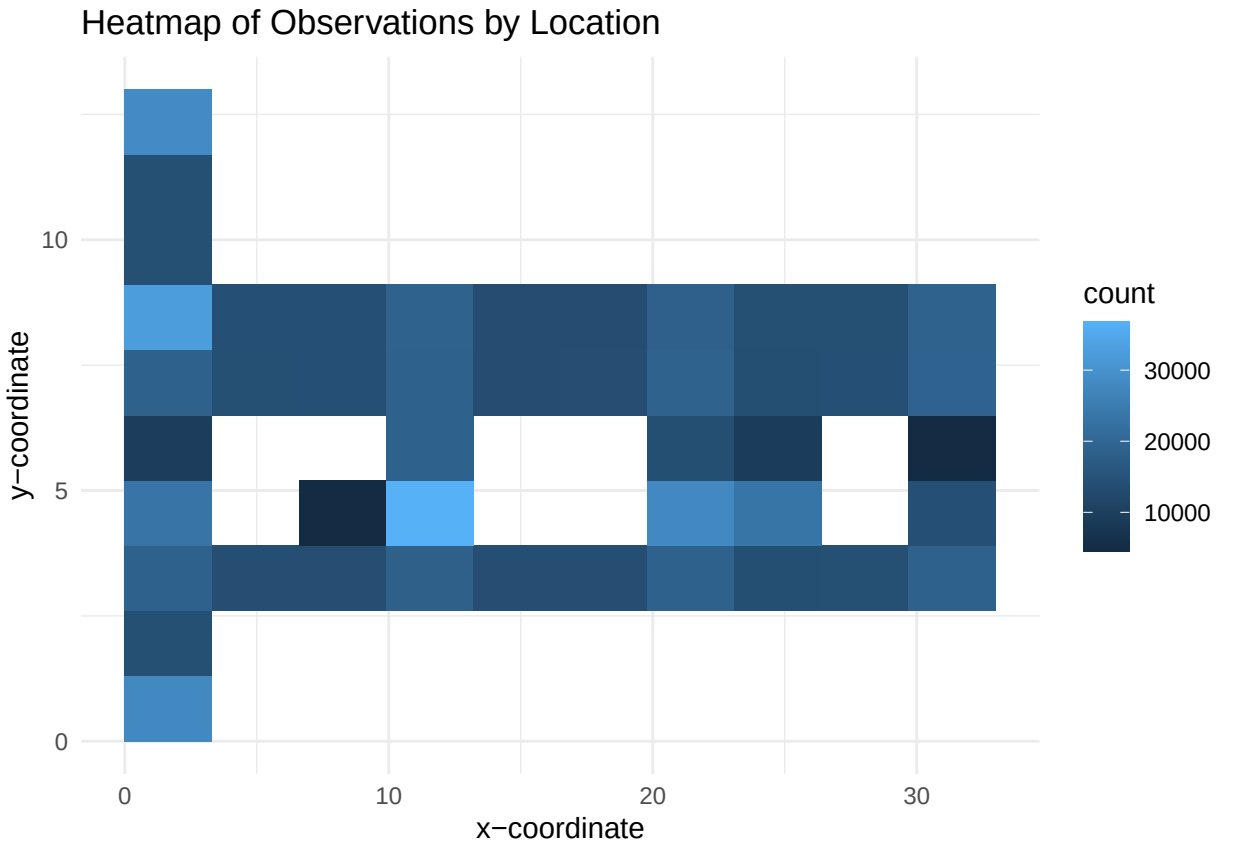

Figure 4: Heat map of strength signal samples by location

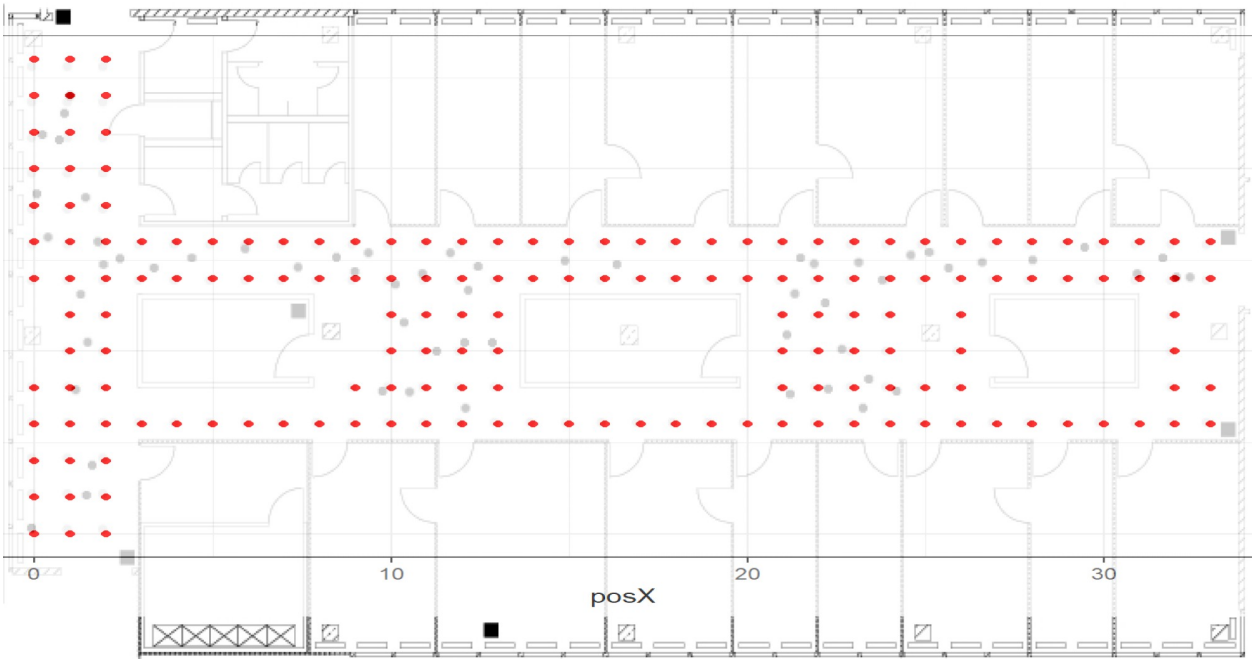

Figure 5: Flooplan location and offline data location. We see the data cleaning was sucessful in eliminating unwanted variables while preserving the geographical information

We analyzed the possibility of a stronger median signal for certain access points on the map depending on the orientation of the device (Fig 6). We note that some access points, particularly `00:14:bf:b1:97:90` has a stronger median signal, while the lowest corresponds to `00:14:bf:b1:97:8d`. This might be because of the location of access point; some will be closer to the sampling locations, while others on the extreme will be further away. Naturally, the strength signal will decrease as the device moves away from the router. What's interesting to notice is that orientation does seem to be influence to signal strength, at least in the 50 percentile behavior.

We have established that at least in the general behavior, orientation seems to influence signal strength. To further explore this, we create a box-plot lattice that:

1) Creates boxplots for each `mac` and `orientation` combination.
2) Groups the individual box-plots by strength signal and orientation `posX`, `posY`.

This plot (Fig. 7) allows to visualize strength signal values depending on orientation. For better visualization we subdivide the plots by access point. groups the different access points, and creates box-plots grouped by signal strength and orientation. The figure shows mainly two things. The first is that some access points have consistently stronger signals, while other have median values that are significantly lower. The second is that there's not a consistent behavior of orientation and signal strength (i.e., the signal strength for some is higher at some orientations, and lower for others). The third is that the median value for the signal varies according to the orientation of the device.

We also select one location `posX` and `posY` equal to zero, and explore the relation of signal strength and orientation, grouped by access point. The results are shown in figure 8. Here, we can see that signals are stronger for one access point, possibly because of proximity. Median values for angles greatly vary as well.

Lastly, we create a new column in the data frame that stores the distance between the access point and the hand-held device. A scatter plot of distance and signal strength is shown below (Fig. 9). The plot shows an inverse realtionship between signal strength and distance.

5. Results

The resulting exploration allows us to conclude several things:

1) Orientation seems to influence strength signal
2) Distance to the device is an important factor of signal strength and they are inversely related
3) Some access points have weaker signals in general; most likely, the ones located at the extremes on the area where the experiment was conducted.

6. Challenges

This section discusses challenges we encountered exploring the Indoor Positioning data.

6.1 Data formatting

The data the client provided is not in a format ready for analysis. The data file had multiple characteristics that need significant re-formatting:

1) It is a text file with minimal formatting.

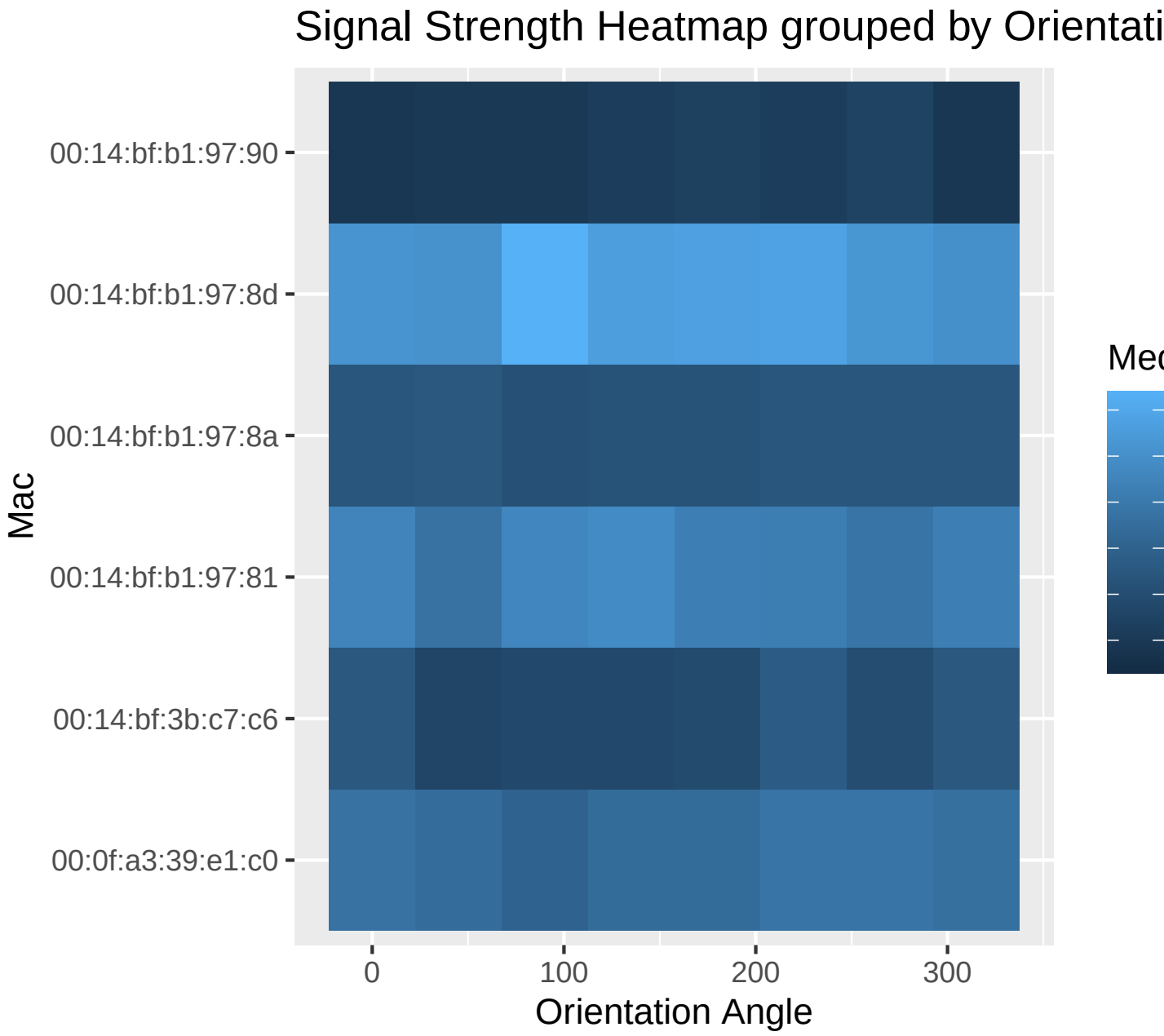

Figure 6: Heat map of strength signal grouped by orientation

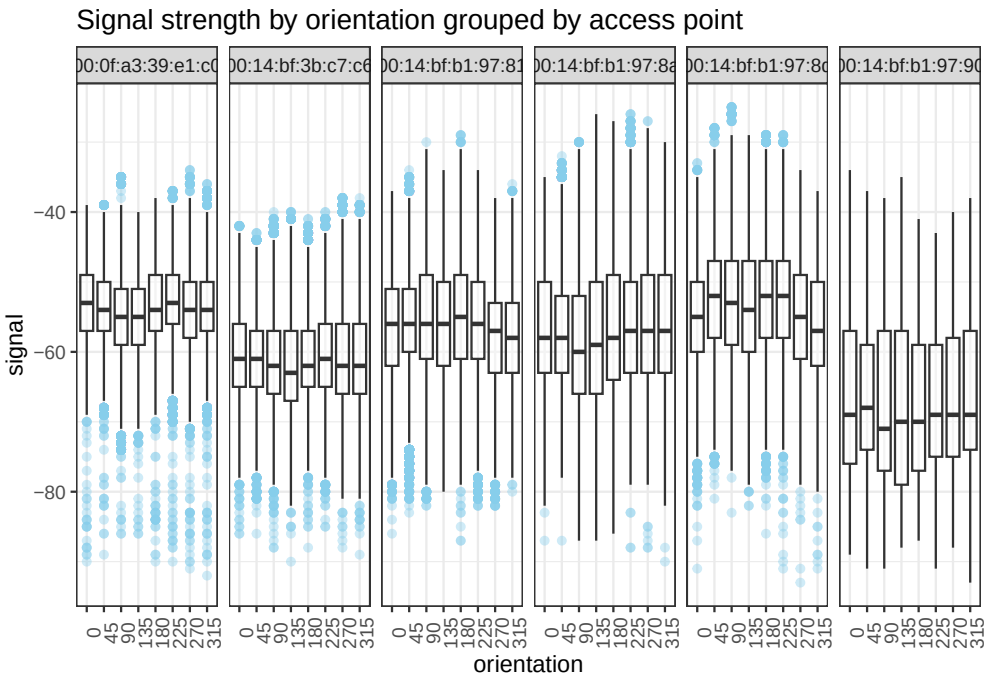

Figure 7: Box-plot of signal strength depending on orientation, grouped by access point

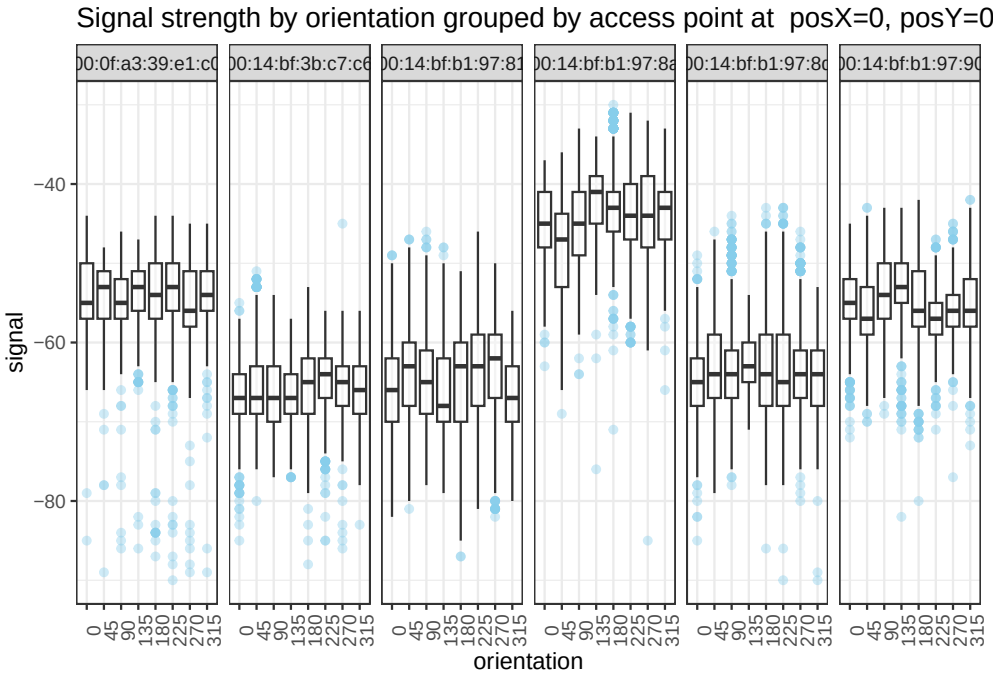

Figure 8: Box-plot of signal strength depending on orientation, grouped by access point in row 5

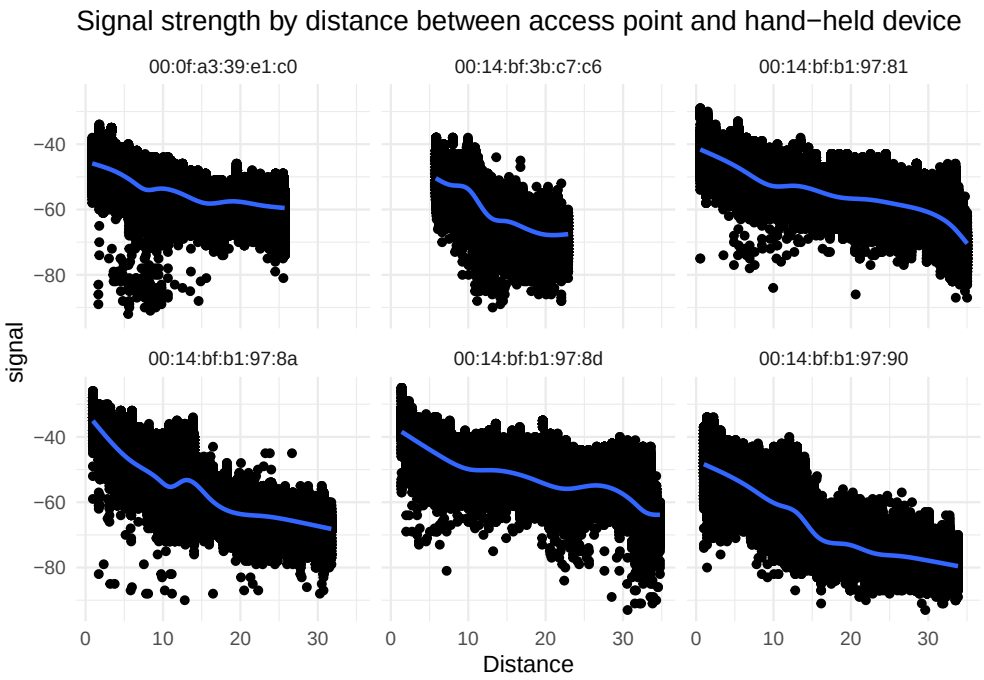

Figure 9: Scatter-plot of signal vs distance grouped by access point

2) The data was not presented in a table format and variables are separated by multiple different separators.

3) Each line corresponds to multiple observations and the number of observations on each line is different.

These challenges were solved by learning about the `string` function.

6.2 Data exploration

1) There are different counts between MAC addresses and channels. From the txt file, we realized that there are extra access points that are not included in the testing area.

2) We found that there is a one to one relationship between MAC address and channel for the seven devices. We had to delete channel from the 'offline' data set.

5.3 Lack of knowledge on the subject

None of the members of the team are experts or at least knowledgeable in IPS, so substantial learning had to be done to understand the data itself.