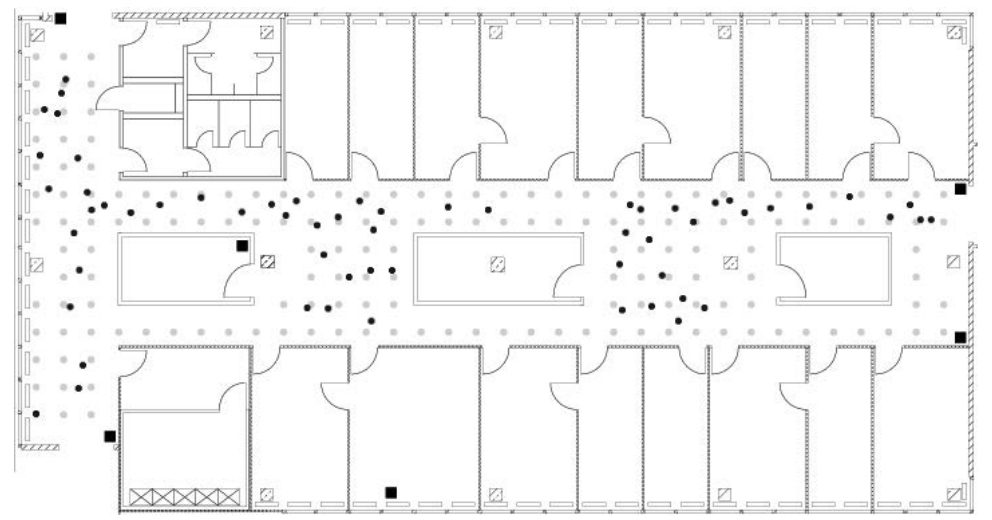
**Abstract**

**Introduction**

With the advent of Wi-Fi and local area networks, devices like Real-Time Location Systems (RTLS) can be leveraged for location positioning of an object in a specified area in real time. This is made possible by way of a continuous communication and feedback between a device held by the object being tracked and the beacons or receiver by the host. Examples of RTLS’s include Infrared, Bluetooth, Cellular, and Radio Frequency Identification (RFID). To operate RTLS, a scanning device is required to locate an object (such as a cell phone or laptop) based on the angle and coordinates of the object being tracked. You can use multiple scanning devices to triangulate the exact location of the object using the combination of angles in which the signal is triggering the respective receivers. Utilizing a series of wireless network signals in an office building, we will be able to detect the exact location of various objects in real time using different combinations of mac devices. Here, we can perform an unweighted and weighted k-nearest neighbors (k-NN) analysis to predict the location of the Online data using the offline data. To go further, we will be seeing if we can better predict the location of the online data using different combinations of the mac addresses available.

**Data**

For this project, we will be leveraging two separate datasets for our analysis. One of which is a reference set named “offline” which contains signal strength measurements from a hand-held device on a gridwork of 166 different points, all of which were spaced 1 meter apart. This gridwork is located in the hallways of a one floor building at the University of Mannheim. The other dataset is titled “online” which we will be using for testing our k-NN model to predict the location. This dataset includes 60 different locations chosen at random with 110 signals measured from them across each point. In figure 1.1 below, you can see a map of the “online” test locations (black dots) overlaid with the “offline” training locations (grey dots). Both datasets contain the same features and will require the same procedures for cleaning.



*Figure 1.1. The floor plan of our experimentation environment. This makes up a 1-floor building in the University of Mannheim with the offline (grey dots) points placed about a meter apart from one another throughout the building. In addition, you can see the placement of the online points scattered throughout for testing.*

|  |  |  |  |
| --- | --- | --- | --- |
| *Feature Name* | *Description* | *Original Data Type* | *Converted Data Type* |
| Time | Timestamp in milliseconds since midnight, Jan 1. 1970 | Character | Numeric |
| id | MAC address of the scanning device | Character | Character |
| pos | The physical coordinate of the scanning device which can be broken out in a latitudinal or longitudinal coordinate plane. | Character | Numeric |
| degree | Orientation of the user carrying the scanning device in degrees | Character | Numeric |
| MAC | MAC address of a responding peer (e.g., an access point or a device in adhoc mode) with the corresponding values for signal strength in dBm (decible milliwatts), the channel frequency and its mode (access point = 3, device in adhoc mode = 1 | Character | Character |

*Table 1.1. The raw data collected from* [*http://rdatasciencecases.org/Data/offline.final.trace.txt*](http://rdatasciencecases.org/Data/offline.final.trace.txt) *can be found here in it’s original form. It includes rudimentary details on the time in which the record was taken, the id number of the scanning device, the position of the reading, the angle of reading and the Mac Address of the responding peer device. This data will need to be cleaned in order for us to further analyze.*

The first cleaning method we will employ will be to break up the position variable into separate variables which we can use to triangulate the location. In our raw dataset, we have position values for latitude, longitude and elevation separated by commas, which we will convert into PosX, PosY and PosZ. Upon further cleaning, we were able to determine that there was only one unique value for PosZ at 0 (which made sense considering the experiment took place in a one story building), we had the liberty to drop the variable. Additionally, running a procedure to check on the number of unique variables in the ScanMac column yielded only a single unique value, so we can drop that one as well.

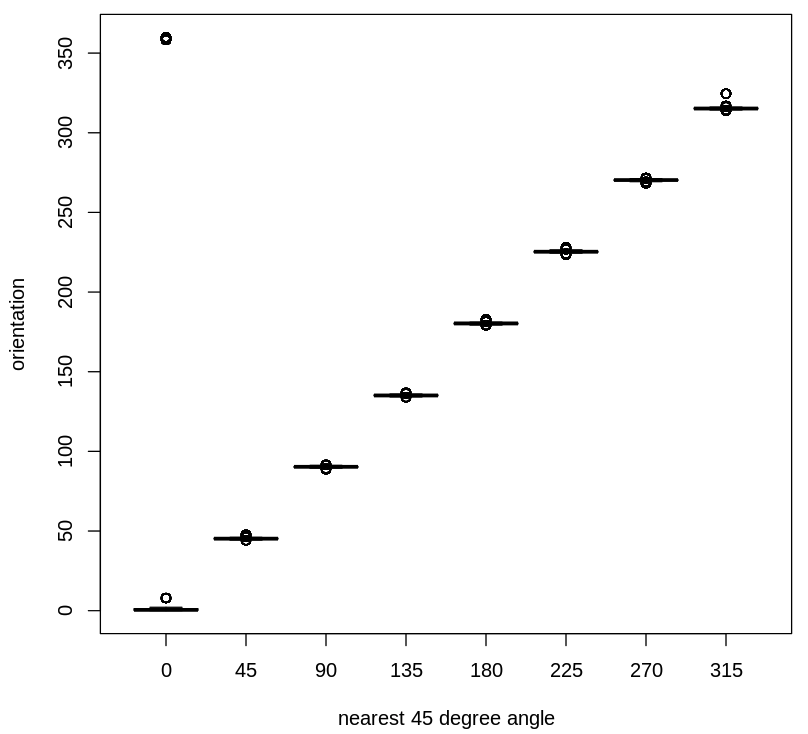
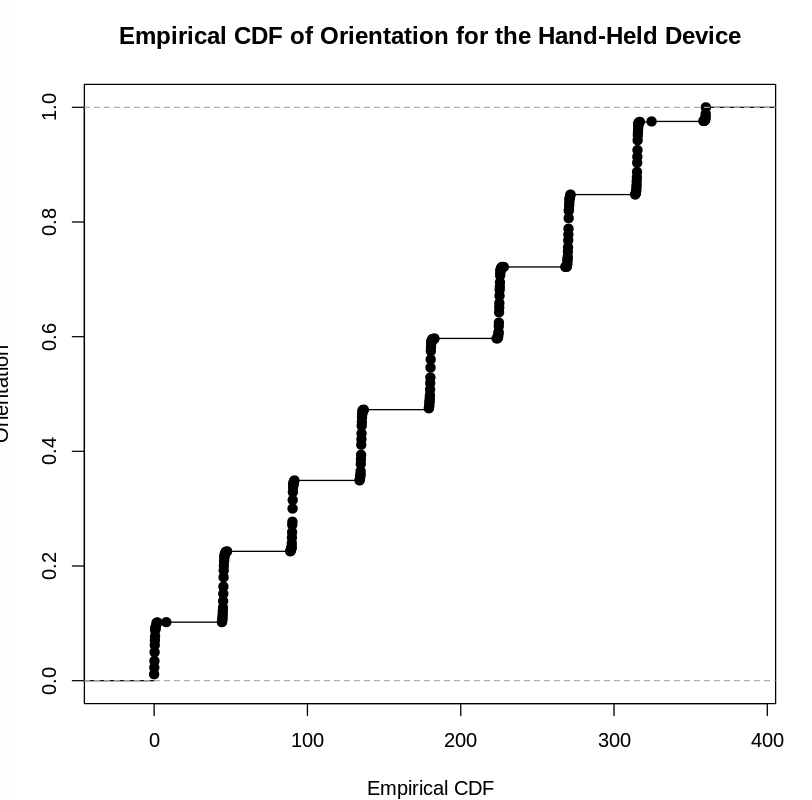
In our documentation, we found that our type of device was mixed between the values 1 and 3, which we may want to clarify more. Reviewing our documentation, we will only want to focus on fixed access points as that is more relevant to our study of predicting device locations using a fixed set of receivers. So, moving further, we will remove the adhoc instances in our dataset.

The Time measurement is something that we will want to make an adjustment to so that we can more easily analyze in the future. As mentioned prior, the time data is based on the number of milliseconds from a specific date (which could possibly be arbitrary), so we can change to a Year-Month-Day-Time format. But first, we can divide the number of milliseconds to seconds. This leaves us with the following features that we will use across our offline and online data. Additionally, we will remove the channel feature since it is strictly a character code that contains redundant identifiers of Mac Address, signal strength, frequency and mode that may play an unfair role in our predictive modeling.

|  |  |  |
| --- | --- | --- |
| Variable | Description | Datatype |
| Time | Year-Month-Day-Hours-Minute-Seconds | POSIXt |
| posX | X coordinate of the device | Numeric |
| posY | Y coordinate of the device | Numeric |
| Orientation | Orientation angle of the device (in degrees) | Numeric |
| Mac Address | The device address of the Mac reciever | Character |
| Signal | Signal Strength of the device (in dBM) | Numeric |
| Raw Time | The raw timestamp / 1000 | Numeric |

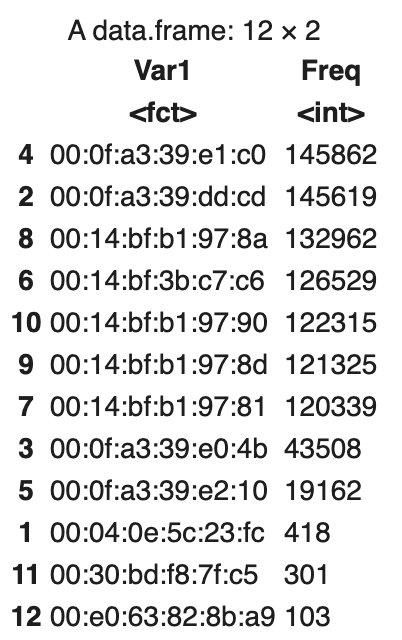
*Table 2.1 Our final data table following our cleaning procedure. Includes the dropping of our PosZ and scanMac value as well as our channel value.*

Next, we will look at the orientation column of our dataset, we can see that we have a wide variety of angles available in clusters around the expected angles (such as 179 or 181) as shown in figure 2.1. Since we are going to focus on measure signal strength at 8 orientations in 45-degree increments, we will round each of our orientations to the nearest 45 degree increment. Additionally, we will try to map values close to 360 so that they line up back to zero.



*Figure 2.1. The location of the orientation values as it relates to the empirical cdf. We can see that at each major orientation (such as 45, 90, 135, 180, ect) we are scattered around these values. So our cleaning procedure is going to round these to the nearest 45 degree angle. After making the adjustment, we can see that the new values look like they are more exact to the 8 angles we are using.*

The last cleaning procedure we will follow is to review the frequency of the mac addresses in place for our models. Since the documentation mentioned that the mac addresses ending in ‘c5’, ‘a9’, ‘fc’, ‘10’, and ‘4b’ were either not on the correct floor, or were not turned on the whole time, we will want to drop these addresses so that we are only working with the remaining 7 addresses that have a frequency of over 120,000, and dropping the ones with less, as per figure 3.1.

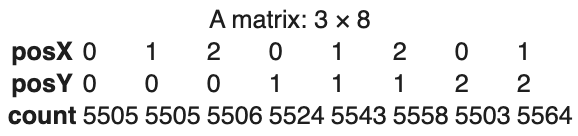


*Figure 2.1. Frequency count of Mac addresses in our study. We will be dropping the addresses that have a frequency count below 120,000.*

## **Exploratory Data Analysis**

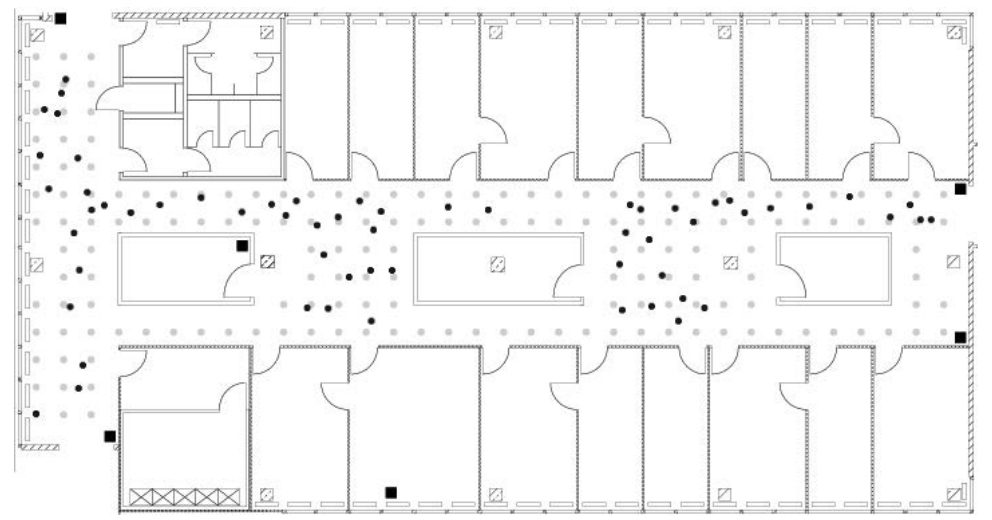
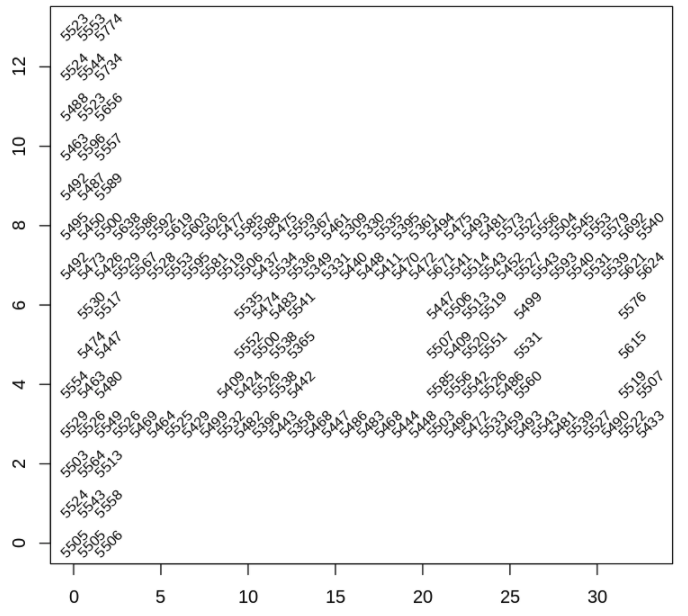
Now that we have clean online and offline data, we will analyze the remaining clean online and offline variables to detect patterns that may help to go into the story of our prediction model. This will include detail on the location and orientation of the Mac Addresses remaining in our dataset as well as the behavior of the signals in relation to the remaining variables.

Considering the position variables of position X and Y, we will endeavor to tally the rows in our data frame for the individual combinations of Xs and Ys. This will give us a total that is longer than the actual x, y locations at which the measurements were recorded, which we found in the documentation were empty. Removing these empty instances will leave us with a total of 166 observations recorded at each location.



*Table 3.1. Taking a count of the number of recordings at each position, which aligns to our 8 orientations. We can see that there are over 5,500 recordings at each position.*

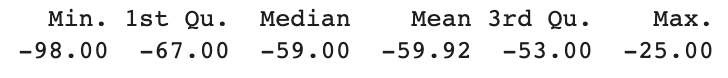
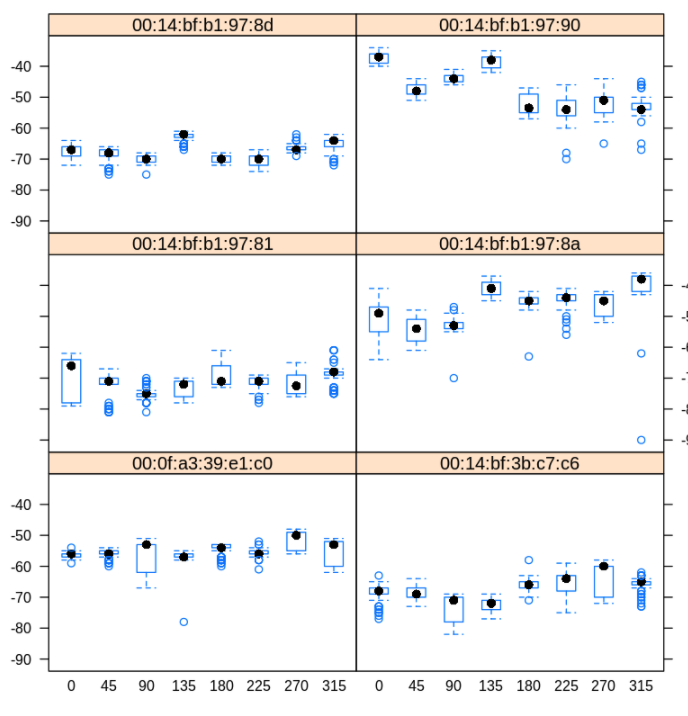
From here, we can visualize all 166 counts by adding the counts as texts to their respective locations, to do so, we can transpose the matrix so that the locations are columns of the matrix. In figure 3.1, we can see how the placements in our dataframe line up with the layout of our experimental room.



*Figure 3.1. The counts of the instances of each address based on the respective location within the experimental location. Considering that each are all relatively close in their count numbers, this suggests that each location was on for roughly the same period of time, which will give us an equal plane for comparison.*

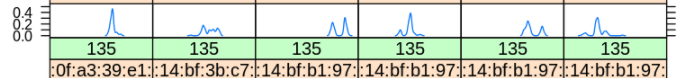
Next, we can work on investigating the properties and composition of the response variable we are interested in for our experiment, which is the signal strength of our devices. To analyze this, we will first review and explore the patterns in which our signal strength tends to behave. Does the signal strength behave the same at all locations? Do the Mac Addresses have different capabilities with regards to signal strength? Do does the signal strength change with repeated measurements over time?

In a laboratory setting, the documentation found that signal strength tends to decrease linearly over distance. However, our experimental environment has walls and humans interacting with it, so we may not see a similar behavior. So to account this, we will compare the signal strengths of different orientations at different access points by fixing the locations on the map, as indicated figure 4.1 for six of the seven mac addresses.



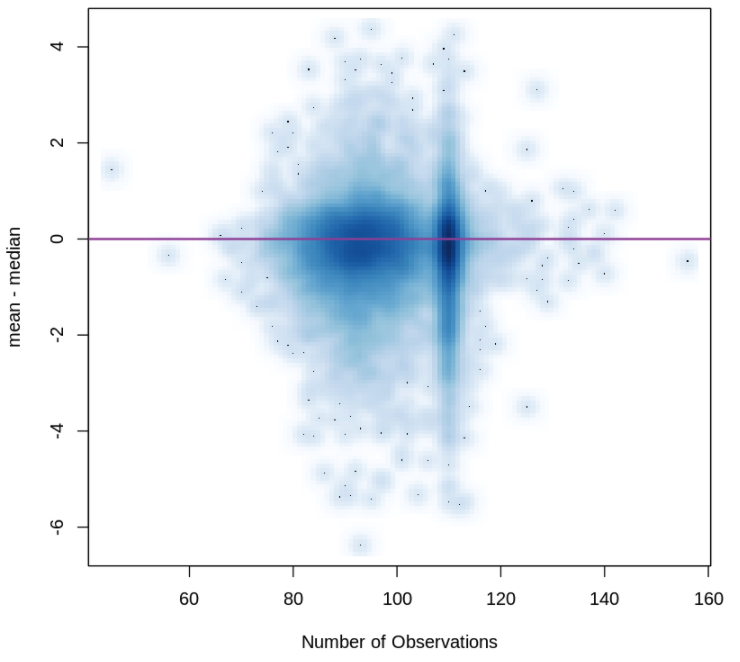
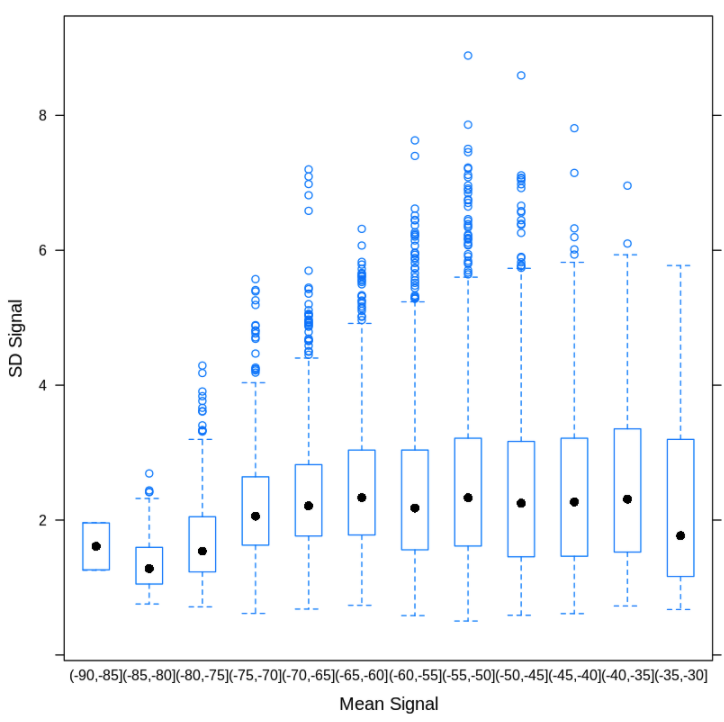
*Figure 4.1. Signal strength broken out by Mac Address and angle. Here, values such as -90 indicate a weak signal while values such as -25 indicate strong signals. With regards to the distributions on Mac Address, there isn’t much consistency outside of the pattern that the 135 degree angle sometimes has a higher signal strength.*

Looking further as one of our strongest signal strength values, 135 degrees (figure 5.1) and its distribution of signal across the different mac address, we can see there is not a normal distribution on any of the mac addresses on signal strength. This is not completely unexpected when we are working in a non-lab setting. As we have walls and other variables interfering with the signals, it will likely affect the distribution of mac addresses, especially since we don’t have documentation stating that our Mac addresses were placed in randomized locations.



*Figure 5.1.Distribution of signal strength for the angle 135 degrees. Many of the distributions across all Mac Addresses look relatively normal, with the center varying depending on the angle and the MacAddress*

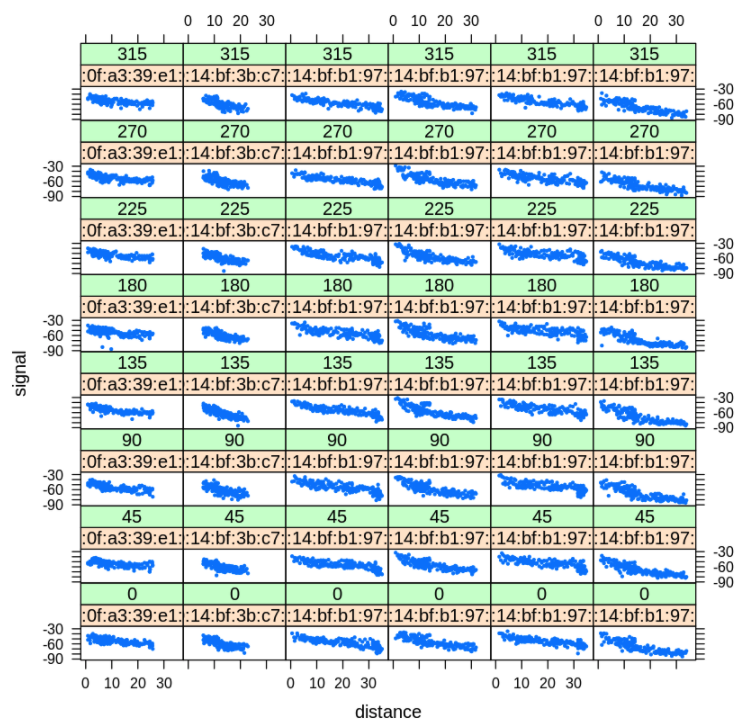
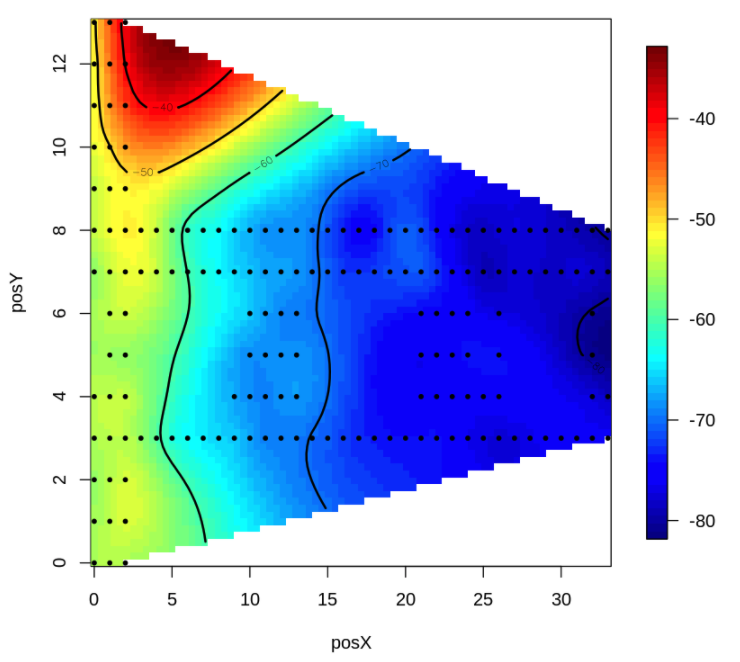
Next, we will look into the distribution of signal strength at each of the 166 locations, 8 angles and 6 access points. Instead of creating 1000s of different boxplots, we will review the summary statistics at each combination of access points. To start with the standard deviations of the average signal strength, we can use boxplots of sdSignal for subgroups of avgSignal through medianSignal against the different number of combinations of observations. Figure 6.1 shows the standard deviations based on the average signal strength between -90 and -30 (binned by 5). We can see that the weakest signals have the smallest standard deviations while the largest standard deviations belong to the stronger signals. As we model our signal strength behavior, we will need to account for this in our transformations to ensure our clustering algorithms run properly.



*Figure 6.1. Boxplots of the standard deviations of the various signal strengths as well as the delta of average and median skewness against the number of observations across all levels of signal strength.*

On the plot on Figure 6.1 that measures the distribution of skewness, we can see there is a bit of a lewt skew towards observations where the median is greater than the mean, which indicates a right skew nature to the data. To adjust for this, we can run a function that will help to smooth the differences between the mean and median within our final cluster model.

Looking further into signal strength and how distance plays a role in its distribution, we can smooth the signal strength over the region where it is measured to create a contour plot as demonstrated in Sigure 7.1. We added a color heat map using the fields package that can fit to a survey to the signal strength at the observed locations.



*Figure 7.1. Media Signal at who Access points at angle 0. The ark red region shows us the main access point of the specific address of 00:14:bf:b1:97:90. Additionally, we have a plot at every location angle showing the relationship between signal and distance.*

Figure 7.1 shows that we can find the location of the access point at the dark red region at the position X of 5 and the position Y of 12. Here, we can also confirm that we have a linear relationship of location and the strength of the signal. This is further verified in the scatter plots to the right indicating the linear relationship showing that signal diminishes over larger distances.