# [CSE590] MINI PROJECT – 1

**[ ADITI SINGH SBU ID (110285096) ADISINGH@CS.STONYBROOK.EDU NETWORK_DATA2 ]**

## 1. Introduction

### 1.1. Data Set:

- The Network Data 2 files comprised of the different types of signals that are transmitted in the MIL-STD-1553 Bus from/to traffic. The header of the columns defines the nature of the signals. For example, BC=>RT10 means that the Bus Controller (BC) is sending data to Remote Terminal (RT) #10.
- Along with the signal that is transmitted to and fro, some disturbances exist. The data files (CSVs) which were given consist of approximately 0.3 million rows and 55 columns, the first column being 'Time'.

### 1.2. Known:

- There are 2 CSV files with the same schema. Each file consists of 0.3 million rows of numerical data for 55 columns.
- The column headers are in (sender) => (receiver) format and there is duplication in the columns, indicating that different types of signals are sent from sender to receiver at a particular instance.
- There are 16 columns (RT2=>BC), 6 columns (RT4=>BC), 4 columns (RT6=>BC), 4 columns (RT10=>BC), 4 columns (BC=>RT10), 3 columns (RT14=>BC), 3 columns (BC=>RT15), and 14 columns (BC => RT16) and 1 column for Time.

## 2. Problem (or approach) statement

### 2.1. Problem:

To find out patterns in data using plots, also to find the Correlation between different signals i.e. correlation between data columns and look for some anomalies (if possible). Apart from the above also find out the differences between the two files using Root Mean Square Error (RMSE).

### 2.2. How to read the Signal transmission notations:

- The columns are labeled as **RT *X* -> BC.*Y*** or **BC -> RT *X.Y*.** This is due to the repetitive columns which defines different messages to be transmitted by the same device to another.
- **X :** defines the RT number. Example RT2 denotes Remote Terminal number 2
- **Y :** defines the **(column -1)** number for a particular device signal set. Example, RT 2-> BC.1 denotes 2nd column of RT2->BC transmission, whereas RT 2-> BC : denotes 1st column of RT 2-> BC transmission and so on.

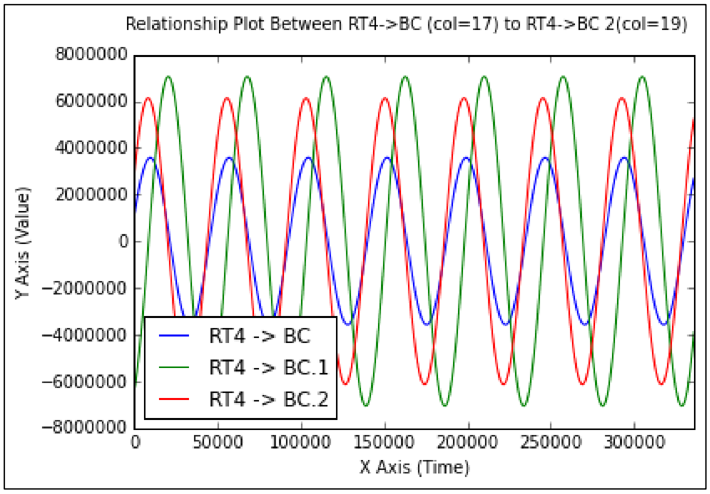
### 2.3. Approach:

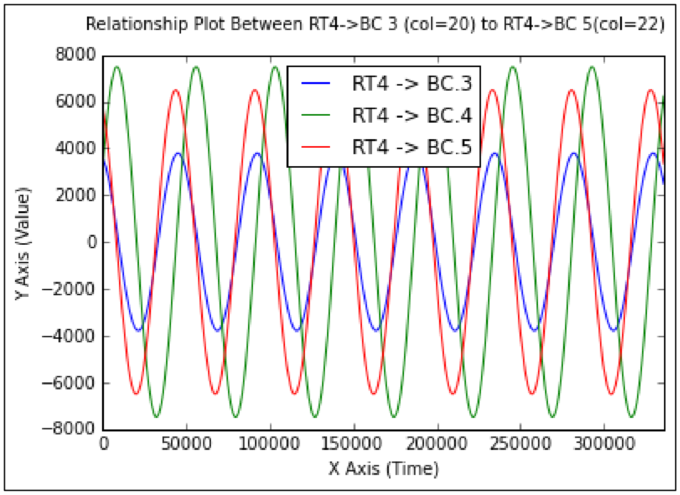
- First plot all columns against time to see definite patterns. Patterns in this case, also referred to periodicity in the signals
- Vary the sample sizes or take different time intervals to see if smaller patterns exist.
- Create Box Plots to see the variation in data columns (attributes). Median, SD, Quartiles can be evaluated. The 'whiskers' of the plot have few data points and can be taken as anomalies (noise), only if it is not symmetric.
- Find out the correlation between different attributes and create a heat map to display it. Get only columns that are highly correlated, either positively or negatively.
- To find the differences between the two datasets 'Root Mean Square Error' or 'Root Mean Square Deviation' is calculated.

# 3. Results and discussion

## 3.1. Some Example Plots:

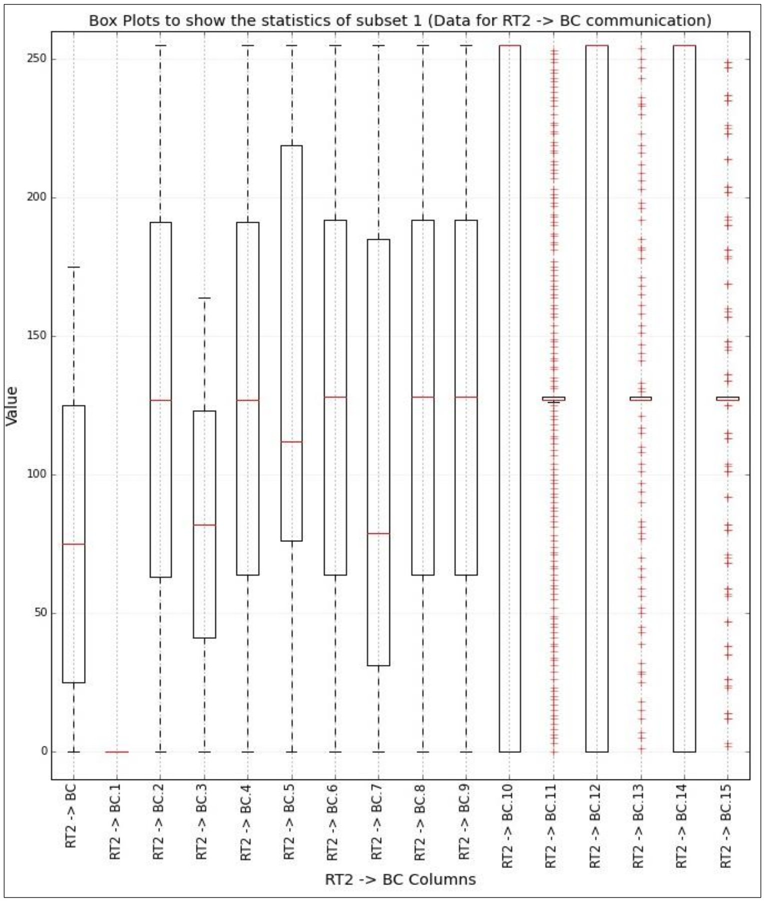
The plots described below show definite periodic waves (Sine & Cosine functions). Since sine waves are periodic, they are used as carrier signals. These sine waves are usually used in Global Positioning System [5.1] to describe position, velocity etc. On plotting other graphs we see periodicity and also observe some box graph & saw tooth graphs, such as those used as digital clocks.

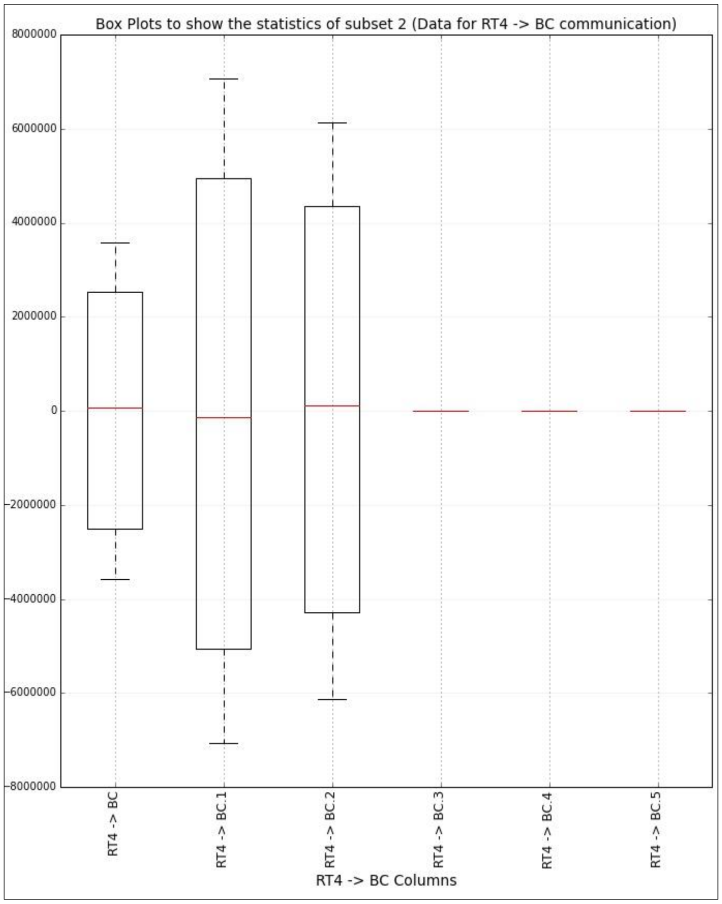




## 3.2 Box Plots used to describe the data – File 1:

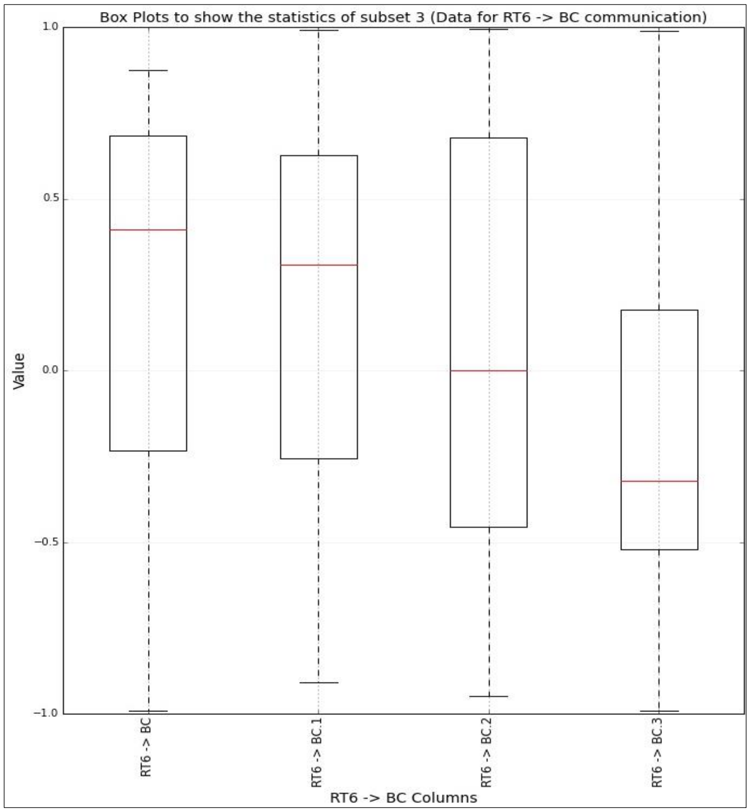
The box plot is used to describe distribution of the dataset. Eg. The mean, mode, quartiles etc. **The red line denotes the Median.** The dotted black lines are the 'Whiskers' and Red dotted lines on the whiskers are 'Fliers'.

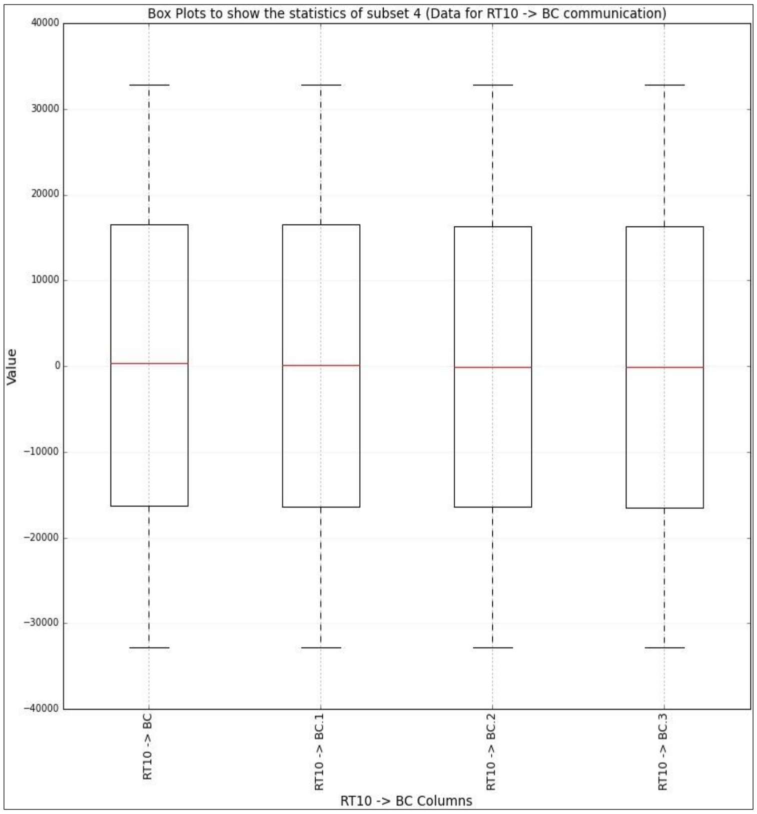


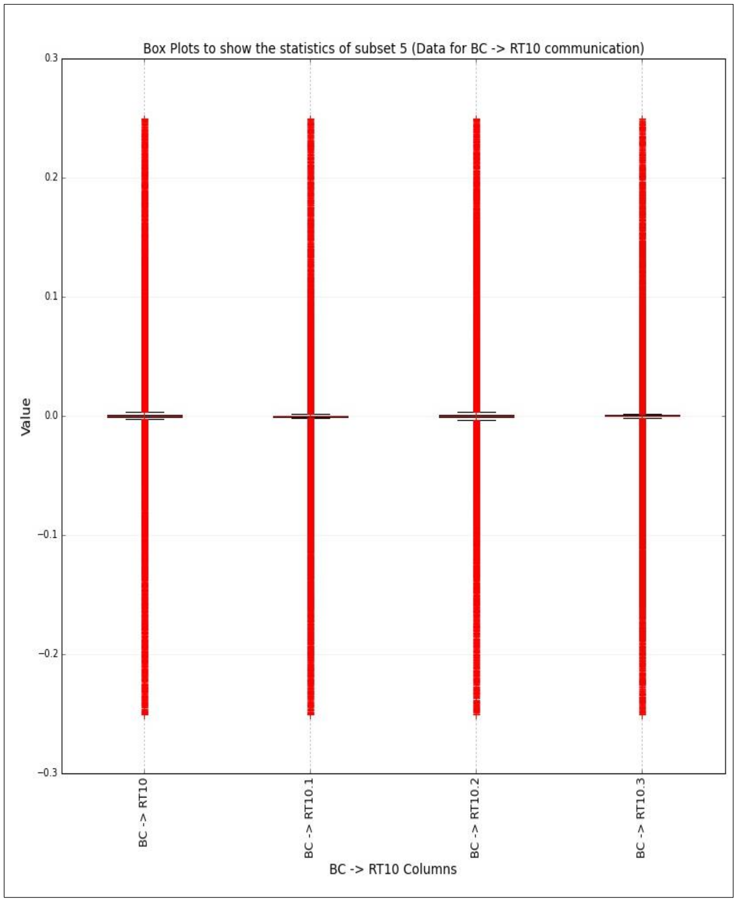


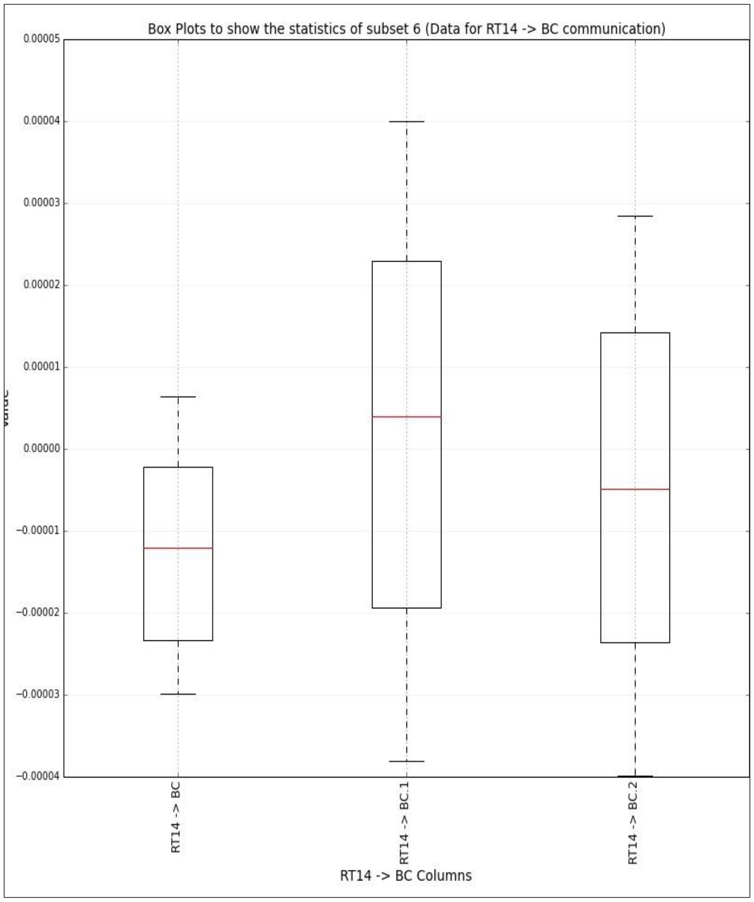
The highest black line is the Maximum value and the lowest black line shows the Minimum Value.

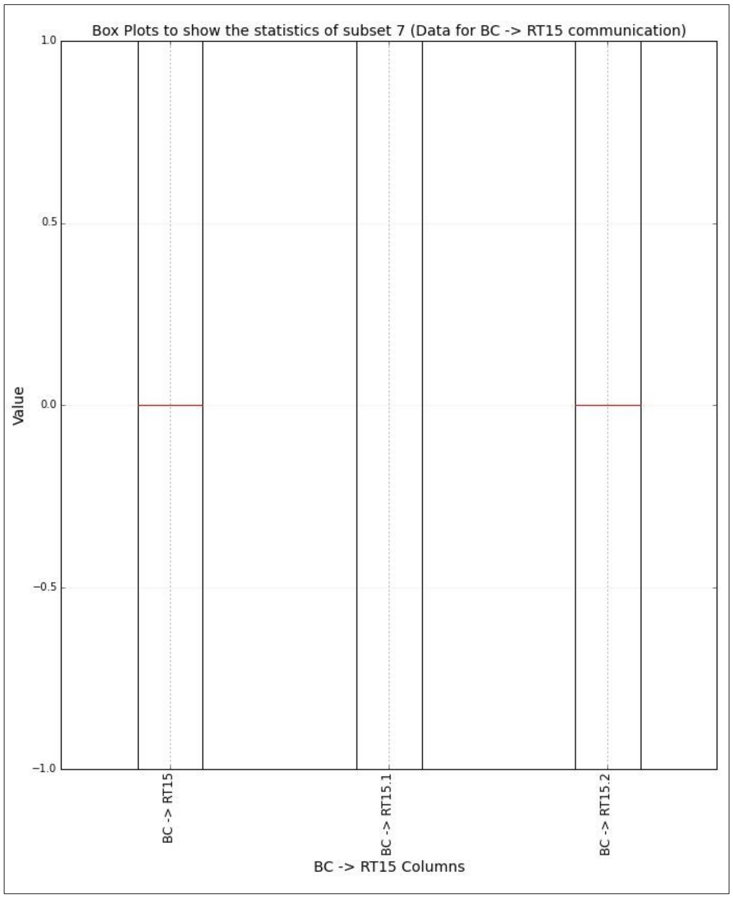
The box contains the major distribution between 1st quartiles. Beyond that the whiskers contain 2nd and 3rd quartiles.

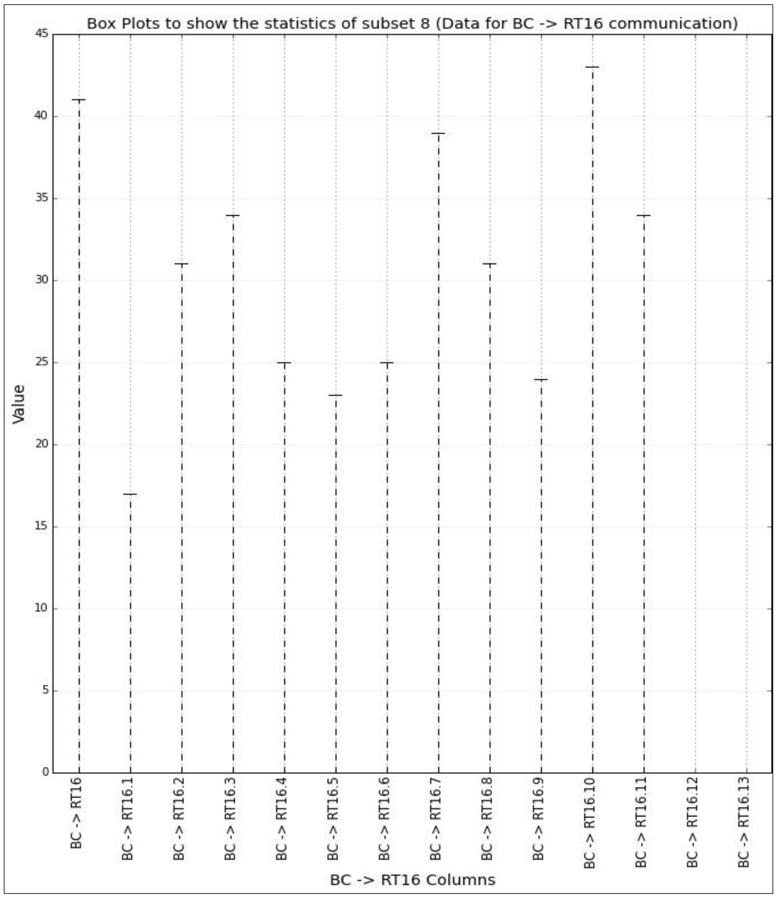












## 3.3. Root Mean Square Error:

| COLUMNS | RMSE |
|---|---:|
| RT2->BC (column #6) | 100.66 |
| RT2->BC (column #7) | 96.93 |
| RT2->BC (column #8) | 101.64 |
| RT2->BC (column #9) | 112.78 |
| RT2->BC (column #10) | 94.73 |
| RT2->BC (column #11) | 109.39 |
| RT4->BC (column #18) | 2612.00 |
| RT4->BC (column #19) | 2764.96 |
| RT4->BC (column #20) | 2723.55 |
| RT10->BC (column #28) | 25288.00 |
| RT10->BC (column #29) | 26143.80 |
| RT10->BC (column #30) | 25508.80 |
| RT10->BC (column #31) | 25316.00 |

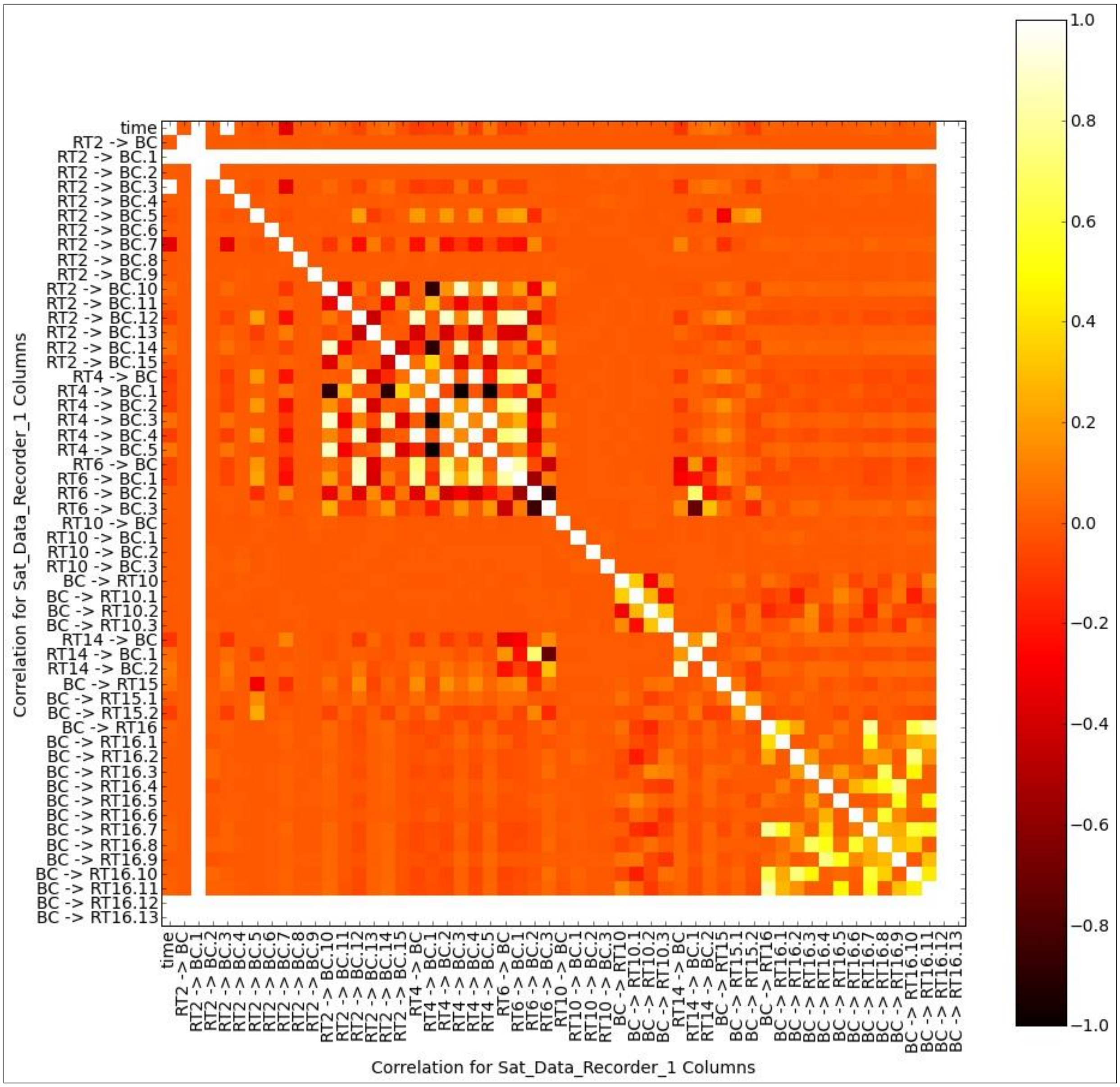
**Method:**

- Find the Difference between each coresponding values of both sets. Find the square of the result set.
- Find the average of the squared columns. Then find the square root of it.
- The resultant is the RMSE (Root Mean Squared Error)

**Observations:**

***Major differences in the two data sets occur for these Remote terminals (RT2, RT4 and RT10) while transmitting signals to Bus Controller.***

### 3.4 Correlation Matrix- Heat Map- File 1



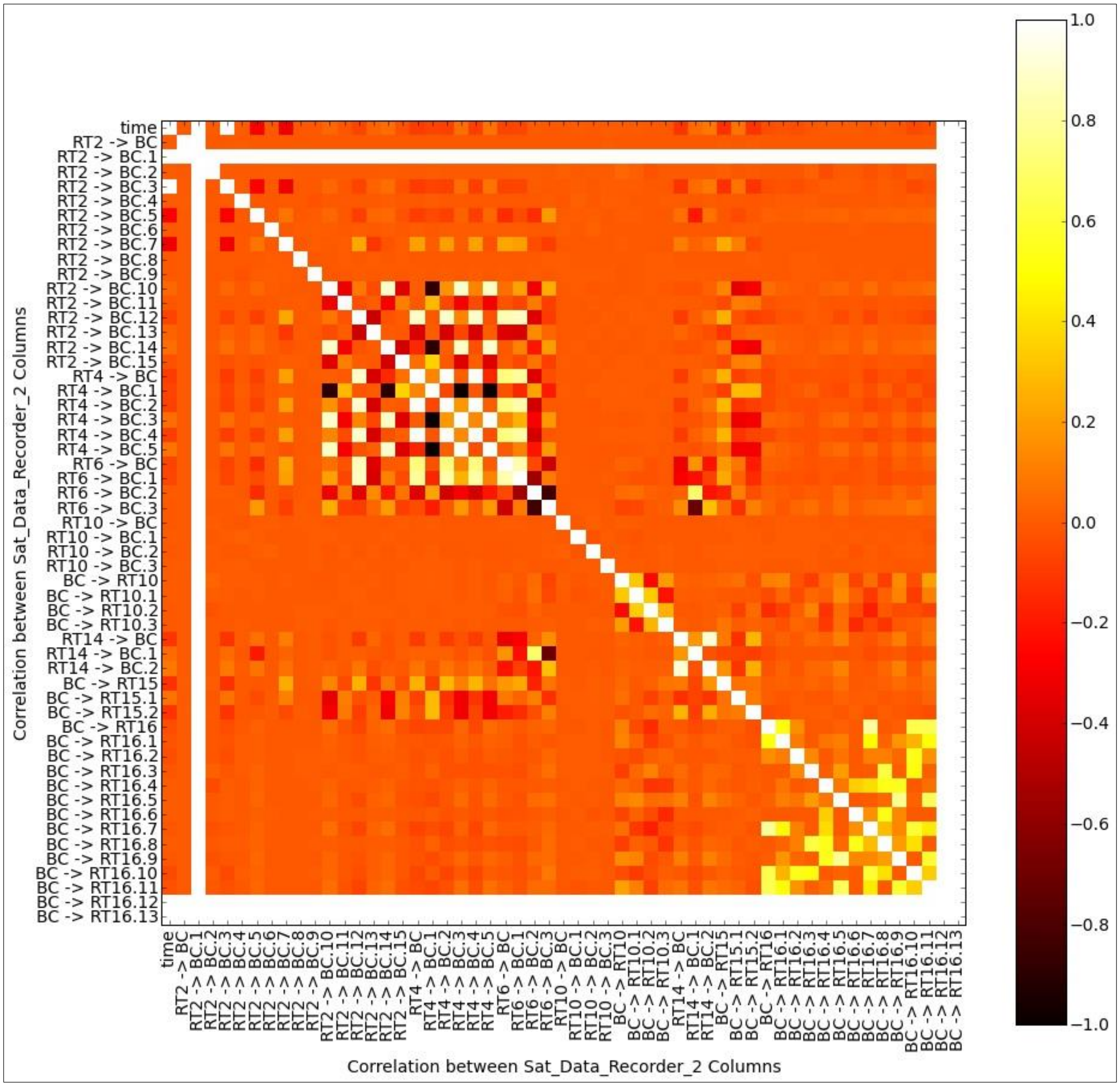
| Column Variable 1 | Column Variable 2 | Correlation Coefficient (CC) |
|---|---|---|
| RT4 -> BC.1 | RT4 -> BC.5 | -0.99905 |
| RT4 -> BC.1 | RT4 -> BC.3 | -0.99168 |
| RT2 -> BC.14 | RT4 -> BC.1 | -0.90069 |
| RT2 -> BC.10 | RT4 -> BC.1 | -0.89414 |
| RT6 -> BC.2 | RT6 -> BC.3 | -0.8487 |
| RT4 -> BC.4 | RT6 -> BC.1 | 0.823665 |
| RT6 -> BC | RT6 -> BC.1 | 0.827311 |
| RT4 -> BC.2 | RT6 -> BC.1 | 0.828857 |
| BC -> RT16 | BC -> RT16.11 | 0.832613 |
| RT2 -> BC.12 | RT6 -> BC | 0.84794 |
| RT2 -> BC.12 | RT6 -> BC.1 | 0.884951 |
| RT2 -> BC.10 | RT4 -> BC.5 | 0.888234 |
| RT2 -> BC.14 | RT4 -> BC.3 | 0.888286 |
| RT2 -> BC.10 | RT2 -> BC.14 | 0.891259 |
| RT2 -> BC.12 | RT4 -> BC | 0.893346 |

- The Negatively correlated columns are highlighted in green . *i.e.* ***CC < -0.8***
- The positively correlated graphs are highlighted in red. i.e. ***CC> 0.8***
- In the correlation map the left diagonal has a Correlation Coefficient (CC) of 1 always, since the variable is always correlated completely to itself.
- From the table enumerating the CC, we can form definite clusters of similar signal transmissions.
- The devices **RT4, RT6 , RT2 and RT14** are the devices which are much correlated (similar) to each other and thus have very high CC (almost 1).

|  |  |  |
|---|---|---|
| **RT2 -> BC.12** | **RT4 -> BC.2** | **0.900526** |
| **RT2 -> BC.12** | **RT4 -> BC.4** | **0.901278** |
| **RT2 -> BC.14** | **RT4 -> BC.5** | **0.901492** |
| **RT2 -> BC.10** | **RT4 -> BC.3** | **0.901615** |
| **RT14 -> BC** | **RT14 -> BC.2** | **0.917878** |
| **RT4 -> BC** | **RT4 -> BC.2** | **0.984955** |
| **RT4 -> BC.3** | **RT4 -> BC.5** | **0.985116** |
| **RT4 -> BC** | **RT4 -> BC.4** | **0.991573** |
| **RT4 -> BC.2** | **RT4 -> BC.4** | **0.999042** |
| **Time** | **RT2 -> BC.3** | **0.999981** |

- Only some sub-columns between a particular transmission set are highly correlated.
- This means that only some have similar signals.
- The signals also consist of Clocks eg. RT2->BC is a saw tooth graph denoting a clock.

### 3.5 Correlation Matrix- Heat Map- File 2.



| Column Variable 1 | Column Variable 2 | Correlation Coefficient |
|---|---|---|
| **RT4 -> BC.1** | **RT4 ->BC.5** | **-0.99905** |
| **RT4 -> BC.1** | **RT4 ->BC.3** | **-0.99168** |
| **RT2 -> BC.14** | **RT4 -> BC.1** | **-0.90069** |
| **RT2 -> BC.10** | **RT4 -> BC.1** | **-0.89414** |
| **RT6 -> BC.2** | **RT6 ->BC.3** | **-0.83574** |
| BC -> RT16.7 | BC -> RT16 | 0.809737 |
| RT4 -> BC.4 | RT6 ->BC.1 | 0.823833 |
| RT4 -> BC.2 | RT6 -> BC.1 | 0.829503 |
| RT6 -> BC | RT6 ->BC.1 | 0.837298 |
| RT2 -> BC.12 | RT6 ->BC | 0.856574 |
| RT2 -> BC.12 | RT6 ->BC.1 | 0.881211 |
| RT2 -> BC.14 | RT4 ->BC.5 | 0.888234 |
| RT2 -> BC.14 | RT4 -> BC.3 | 0.888286 |
| RT2 -> BC.10 | RT2 -> BC.14 | 0.891259 |
| RT2 -> BC.12 | RT4 ->BC | 0.89335 |
| RT2 -> BC.12 | RT4 ->BC.2 | 0.900527 |
| RT2 -> BC.12 | RT4 ->BC.4 | 0.901278 |
| RT2 -> BC.14 | RT4 -> BC.5 | 0.901492 |
| RT2 -> BC.14 | RT4 ->BC.3 | 0.901615 |
| RT14 -> BC | RT14 ->BC.2 | 0.917878 |
| RT4 -> BC | RT4 -> BC.2 | 0.984955 |
| RT4 -> BC.3 | RT4 -> BC.5 | 0.985116 |
| RT4 -> BC | RT4 -> BC.4 | 0.991573 |
| RT4 -> BC.2 | RT4 -> BC.4 | 0.999042 |
| Time | RT2 -> BC.3 | 0.999981 |

**Similar to the above correlation matrix description:**

- The Negatively correlated columns are highlighted in green . *i.e.* ***CC* < *-0.8***
- The positively correlated graphs are highlighted in red. i.e. ***CC> 0.8***
- In the correlation map the left diagonal has a Correlation Coefficient (CC) of 1 always, since the variable is always correlated completely to itself.
- From the table enumerating the CC, we can form definite clusters of similar signal transmissions.
- The devices **RT4, RT6 , RT2 and RT14** are the devices which are much correlated (similar) to each other and thus have very high CC (almost 1).
- Only some sub-columns between a particular transmission set are highly correlated.
- This means that only some have similar signals.
- The signals also consist of Clocks eg. RT2->BC is a saw tooth graph denoting a clock

## 4. Conclusion:

- From the Correlation Matrix we could find out certain columns which are highly correlated.
- The only transmission that seems noticeably similar to **Time** is **RT2 ->BC3.** *This means that this is a* **monotonically increasing graph**. *This can be to denote the* ***altitude*** *increase of a shuttle.*
- **A cluster** of similar devices can be created looking into the tables above which shows that the devices **RT4, RT6 , RT2 and RT14** are the devices which are much correlated (similar) to each other
- These High correlation coefficients exist for those transmissions which occur from similar systems. For example, similarity exists in transmission between two "**RT.x => BC and RT.y =>BC**" only and also one between a pair of **"BC=>RT.x and BC=>RT.y"** only. We can't see correlation between, let's say, a RT-> BC and BC->RT transmission.
- *This tells us that the BC has a different way to communicate with the RT , than a RT with BC.*
- When we observe the heat map, we can see that only the columns which are close to each other, i.e. points nearer to the left diagonal are more likely to be correlated than extreme edge points.
- A few differences exist between the two CSV files, but the order of the differences appear small. Thus on performing the root mean square error, we can evaluate the error existing in the individual columns.
- This shows very high differences between **RT10 ->BC** signals, **RT2->BC** (middle signals) and the beginning signals from **RT4->BC.**

## 5. References:

1. **Tutorial on MIL STD 1553 :** *http://aviftech.com/files/3313/5168/8298/1553_Tutorial.pdf*

2. **The Global Positioning System by James R. Clynch, 2003. (For understanding applications of sine waves.)**