Sailing Boat Performance Analysis

Internship Report

Chanon Jenakom
chanonjenakom@gmail.com

IMT Atlantique Bretagne-Pays de la Loire
Brest, France
1st June - 31st July, 2018

An internship report presented for the Cooperative Education Software & Knowledge Engineering (01219490), Bachelor's Degree of Software & Knowledge Engineering

Department of Computer Engineering
Kasetsart University
July 26, 2018

Acknowledgements

I would like to thank Professor Romain Billot for providing me with all the support I need during my internship here at IMT Atlantique. It has been a pleasure to work on this project with him. Not only his help with the project, but also his advice and information on the cultures and daily routines here in France has made my time here much more enjoyable.

I would also like to express my appreciation for Professor Philippe Lenca's warm welcome to me during the first days of my internship. Thank you for making sure everything is fine for me during my stay.

I want to thank all the staff who work here at IMT Atlantique. They have been very helpful in every situation and always help us without hesitation.

I would also like to thank all the students I met at Le Foyer des lves who has helped me with whatever I need whenever I visit the place. They are all very kind and friendly.

Moreover, I would also want to thank my friends from Kasetsart University who are also interns at IMT Atlantique with me for all the help and support they have provided for me, whether it be about the internship project or other daily life matters.

Most importantly, I would like to thank my parents and relatives who have always been supportive of me both before and during my internship at IMT Atlantique. Their supports means a lot to me in every way.

Abstract

Nowadays, boat sailing is one of the popular competitive sports around the world. In order to gain advantage over the opponents, the use of data analysis is required. With large amount of data collected from the sensors on the boat during each sailing session, it is very important to select and deploy suitable big data techniques and tools in order to effectively and efficiently understand the data.

As a means to understand which of the factors that were collected by the sensors have influence on the performance of the boat, each data point were categorized into different bins based on the wind speed and angle at the given moment. Once the bins were created, each of them was then analysed in details to find the correlations between the sensors and other related descriptive statistics. Finally, the principal components for each bin were computed along with the level of contributions from each of the original attributes in the dataset.

Contents

List of Figures

List of Tables

Objectives

Using the data retrieved from team Gitana 17, the main objectives of this project are:

1. Create the bins for each data points according to the wind statistics.
2. Construct descriptive statistics for each of the created bins.
3. Perform Principal Components Analysis on the created bins.

Chapter 1

Introduction

With a large amount of data collected from the sensors during Gitana 17's sailing sessions, it is crucial that these data are analyzed correctly and effectively. One of the first steps to understand these data is to understand the behaviors of each component on the boat and how winds can interact with them. It is also very important to take into account that different wind statistics such as its speed and its incident angle have different effects on the boat's performance, and must be analyzed separately. Furthermore, it is advisable that each separated data points, or bins should be small so that each member of the bin shares the same behaviors under the same wind conditions as much as possible.

Once the bins are created, it is then the time to analyze each of them to find the factors that affect the performance of the boat, namely its speed. However, it is not only the boat's speed that must be focused on, since it is possible, and very likely, that there are many other factors between each component of the boat. Furthermore, by discovering how each of the different sensors may affect each other, the number of sensors that have to be analyzed can be reduced for ease of computation and higher performance during real competitions. This is evident when sailing in critical wind conditions where different algorithms to determine the next course of actions are required on the fly.

This report will focuses on the methodology used in creating the bins, as well as the selection strategy used in order to get rid of bins with inadequate amount of data to be suitable for analysis. Moreover, the results of descriptive statistics and principal component analysis will be shown and discussed in details.

Chapter 2

Tools

- **NumPy** is used to perform matrix-related calculations and transformations, to help ease the task of data manipulation.
- **Pandas** is used as a tool to store and output data including the sensors data and the created descriptive statistics.
- **Matplotlib** is a library used for plotting data points in a variety of ways such as bar plots, line graphs, etc.
- **Seaborn** is another library used in conjunction with Matplotlib to perform functions that the latter does not provide, such as modifiable correlation heatmaps.
- **Chardet** is a library used to detect encodings of the file. It helps minimize the work required to read data in Pandas' dataframe.
- **Scikit-Learn** is used for standardizing the data as well as computing the principal component.
- **Jupyter Notebook** is used as a means to present the outcome in an easy-to-read manner to non programmers, with example codes and their outputs.

Chapter 3

Methodology

3.1 Preprocessing

In order to effectively use the dataset retrieve from Gitana 17, it is necessary to clean and perform data preprocessing techniques. A few of the techniques used were feature selection and feature creation.

Sensors Data					
Data	Hour	Max-1s-Foil-B-P-1-i-0-1	Max-1s-Foil-B-P-1-i-0-2	...	VarFilter-WTP-SelBoatSpd
08/06/2018	12:41:07	-28.70	-43.64	...	17.79
08/06/2018	12:41:08	-32.33	-55.92	...	17.68
08/06/2018	12:41:09	-27.51	-33.82	...	17.60
08/06/2018	12:41:10	-27.96	-34.98	...	17.62
08/06/2018	12:41:11	-34.06	-72.92	...	17.63

Table 3.1: Examples of Original Sensors Data

3.1.1 Feature Selection

In this step, a subset of important sensors is selected by hand during a meeting with a team member. However, there are still some sensors that while important, may need to be dropped for the analysis stages, such as some of the duplicate attributes from different stages of the data collection.

3.1.2 Feature Creation

The this next step, some of the selected sensors can be combined in order to reduce the computation power required to finally analyze the data. For example, the upper and lower sensors seen in Figure 3.1 can be combined for reduced number of features.

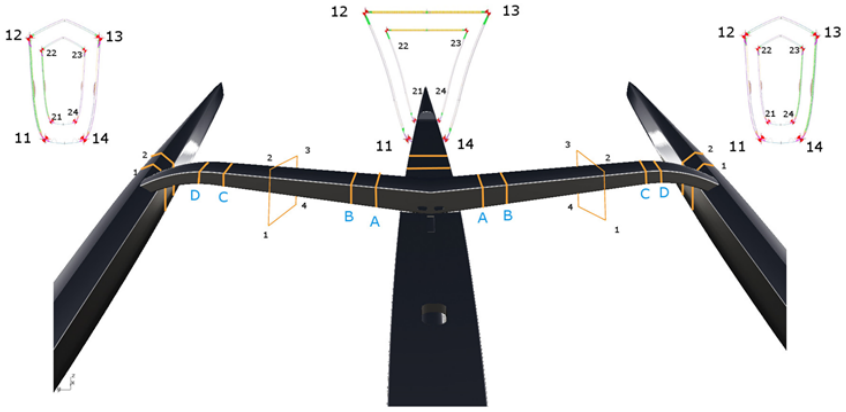

Figure 3.1: Sensors' Positions

3.1.3 Result

After feature selection creation, the result are as follows:

Sensors Data					
Date	Hour	Max-Board-Deformation-Inside-Port	Max-Board-Deformation-Outside-Port	...	Boat-Speed
08/06/2018	12:41:07	-57.51	94.68	...	17.79
08/06/2018	12:41:08	-100.27	263.56	...	17.68
08/06/2018	12:41:09	-45.66	246.61	...	17.60
08/06/2018	12:41:10	-39.15	151.98	...	17.62
08/06/2018	12:41:11	-169.08	163.32	...	17.63

Table 3.2: Examples of Preprocessed Sensors Data

3.2 Bins Creation

For the bins creation algorithm, the default size of the bins are 5 degrees by 2 knots, and its size is adjustable if the team sees fit. All the basic descriptive statistics are computed within the algorithm itself, including the correlations between features. The detailed implementation is shown in Listing 3.1.

```
1
2  def create_bins(df,
3                  wind_features=[x[1] for x in wind_features],
4                  target_feature=boat_speed_feature[1],
5                  statistics_cols=statistics_features,
6                  dx=5, dy=2, min_thresh=5,
7                  exclude=[x[1] for x in identifier_features + wind_features
     ]):
8      """Create bins"""
9      bins, corr = {}, {}
10     for max_x in range(-180, 180, dx):
11         for max_y in range(0, math.ceil(df[wind_features[1]].max()), dy):
12             query = '{0} >= {2} and {0} < {2}+{4} and {1} >= {3} and {1} <
     {3}+{5}' \
13                     .format(wind_features[0], wind_features[1], max_x,
     max_y, dx, dy)
14             binned_df = df.query(query)
15             bin_size = len(binned_df.index)
16             bin_corr, unsorted_corr = compute_sorted_corr(binned_df.drop(
     exclude, axis=1))
17             if bin_size >= min_thresh and bin_corr is not None:
18                 bin_name = 'bin_x{}to{}_y{}to{}'.format(max_x, max_x+dx,
     max_y, max_y+dy)
19                 for col, c_corr in unsorted_corr.items():
20                     if col in corr:
21                         corr[col].append(c_corr)
22                     else:
23                         corr[col] = [c_corr]
24                 bins[bin_name] = {'bin': binned_df, 'size': bin_size, 'corr
     ': bin_corr}
25     sorted_corr_df = pd.DataFrame(data=sort_corr({col: median(c_corr) for
     col, c_corr in corr.items()}), columns=statistics_cols)
26     return bins, dx, dy, max_x, max_y, sorted_corr_df
```

Listing 3.1: Bins Creation Algorithm

By using this algorithm, the resultant bins are shown in a grid layout, where each grid represents one bin, as shown in Figure C.2.

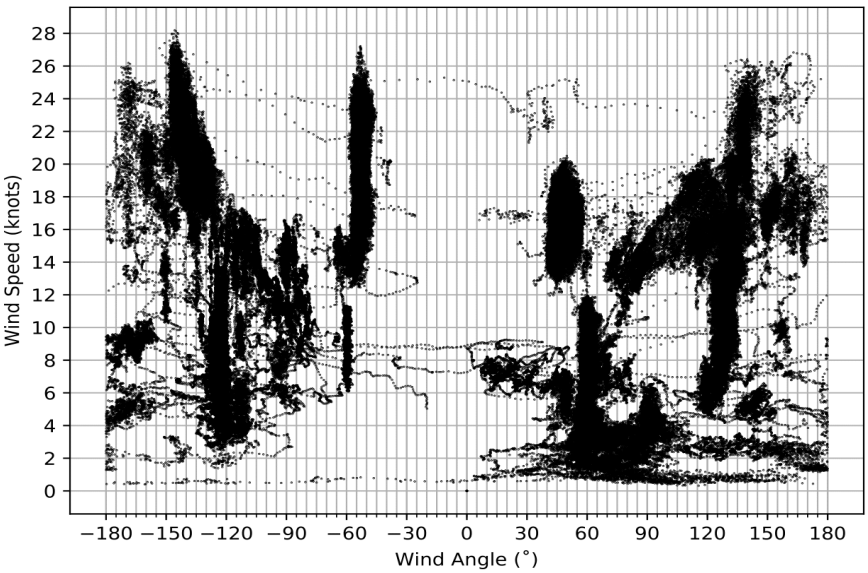

Figure 3.2: Bins Creation

Also, each of the bins contains descriptive statistics such as boxplots and heatmaps, where the examples can be seen in Appendix A.

3.3 Principle Component Analysis

In the next step, further analysis is done of the bins. For the purpose of this report, three of the bins have been selected to demonstrate the resultant analyses. The bins are:

- 90 degrees by 9 knots
- 140 degrees by 22 knots
- 50 degrees by 16 knots

Principal component analyses are performed on these bins in order to reduce the number of relevant features, along with the importance of each original sensors. The detailed implementation is shown in Listing 3.2.

```
def create_PCA(df,
               target_feature=boat_speed_feature[1],
               var_thresh=.75,
               column_name='PC_{}',
               exclude=[x[1] for x in identifier_features + wind_features]):
    """Create Principal Component Analysis"""
    df = df.drop(exclude, axis=1)
    df.dropna(axis=1, how='all', inplace=True)
    df.replace([np.inf, -np.inf], np.nan, inplace=True)
    df.fillna(df.mean(), inplace=True)
    X = df.drop(target_feature, axis=1).reset_index()
    X, y = StandardScaler().fit_transform(X), df[target_feature].
    reset_index()
    pca = PCA(var_thresh)
    principal_components = pca.fit_transform(X)
    columns = [column_name.format(i+1) for i in range(0, len(
    principal_components[0]))]
    pca_df = pd.DataFrame(data=principal_components, columns=columns)
    pca_df = pd.concat([pca_df, y], axis=1).drop(['?index'], axis=1)
    return pca, pca_df
```

Listing 3.2: PCA Computation

Chapter 4

Result

4.1 First Bin - -90 Degrees by 9 Knots

In the first bin, it can be seen from Figure 4.1 and Figure 4.2 that the data are quite sparse, with additional descriptive statistics available in Appendix A. Moreover, an example of the sensors contribution to each principal component in this bin is included in Figure 4.1 and Figure 4.2, along with an accompanying table (Table 4.1) labelling each of the numbers in the circle or correlations.

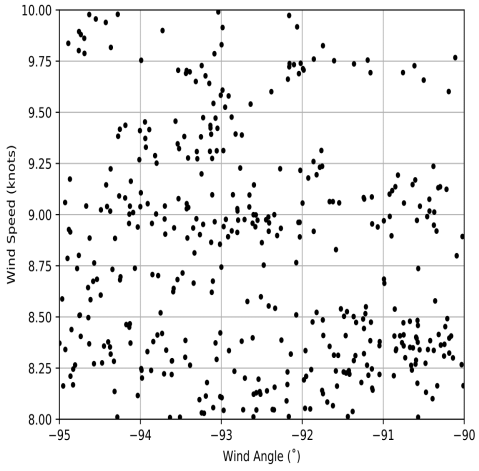

Figure 4.1: 90 degrees by 9 knots

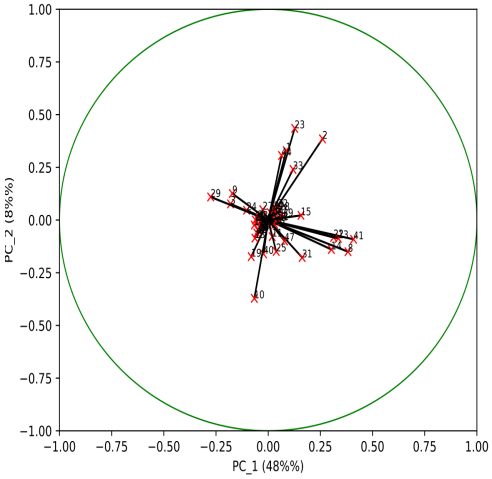

Figure 4.2: First Bin's Statistic

Labels Table	
Label	Feature
1	Max-Front-Beam-Lower-Forward-Rake-Starboard
2	Max-Front-Beam-Upper-Aft-Rake-Starboard
3	Max-Hull-To-Center
4	Max-Front-Beam-Upper-Aft-Rake-Port
5	Max-Front-Beam-Upper-Forward-Rake-Port
...	...

Table 4.1: Accompanying Table for Figure 4.1 and Figure 4.2

4.2 Second Bin - -140 Degrees by 22 Knots

In the second bin, it can be seen from Figure 4.3 and Figure 4.4 that the data are quite sparse, with additional descriptive statistics available in Appendix B. Moreover, an example of the sensors contribution to each principal component in this bin is included in Figure 4.3 and Figure 4.4, along with an accompanying table (Table 4.2) labelling each of the numbers in the circle or correlations.

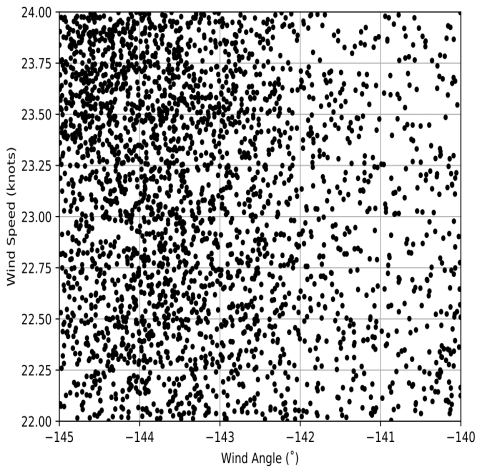

Figure 4.3: -140 Degrees by 22 Knots

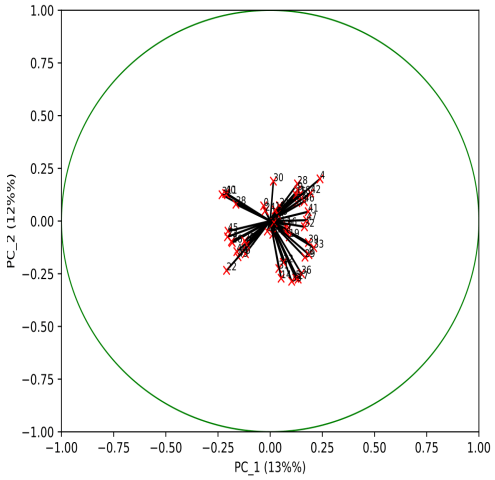

Figure 4.4: Second Bin's Statistic

Labels Table	
Label	Feature
1	Min-Front-Beam-Upper-Forward-Rake-Starboard
2	Max-Front-Beam-To-Starboard
3	Max-Hull-Temp-Starboard
4	Min-Front-Beam-Lower-Forward-Rake-Port
5	Max-Front-Beam-Lower-Aft-Rake-Port
...	...

Table 4.2: Accompanying Table for Figure 4.3 and Figure 4.4

4.3 Third Bin - 50 Degrees by 16 Knots

In the third bin, it can be seen from Figure 4.5 and Figure 4.6 that the data are quite sparse, with additional descriptive statistics available in Appendix C. Moreover, an example of the sensors contribution to each principal component in this bin is included in Figure 4.5 and Figure 4.6, along with an accompanying table (Table 4.3) labelling each of the numbers in the circle or correlations.

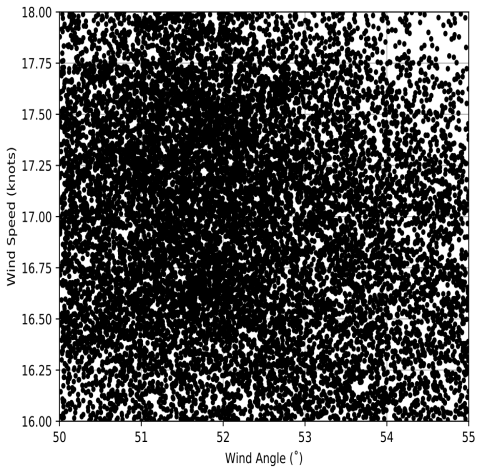

Figure 4.5: 50 Degrees by 16 Knots

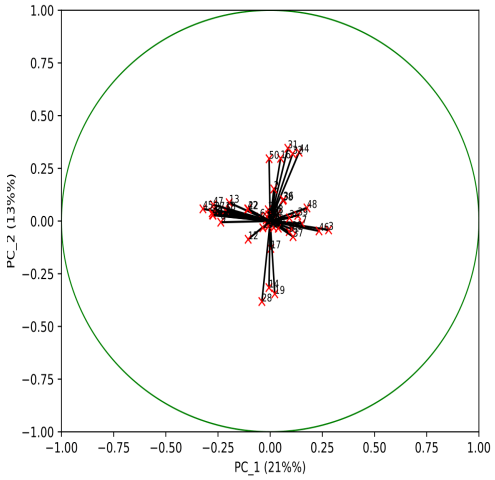

Figure 4.6: Third Bin's Statistic

Labels Table	
Label	Feature
1	Max-Front-Beam-Upper-Forward-Rake-Starboard
2	Min-Hull-Temp-Center
3	Min-Front-Beam-Upper-Forward-Rake-Port
4	Min-Front-Beam-Lower-Aft-Rake-Port
5	Max-Front-Beam-Upper-Aft-Rake-Port
...	...

Table 4.3: Accompanying Table for Figure 4.5 and Figure 4.6

Chapter 5

Analysis

5.1 Preprocessing

The preprocessing steps taken were efficient enough to come up with a result. However, there are much rooms for improvement in many areas.

5.1.1 Feature Selection

In this part, since some of the features contain duplicates within the data, as well as a lot of missing values at the time of this project. Moreover, despite the fact that the features were personally hand-picked in the earlier stage of this project, there were minor drop in qualities of the results of this project due to the aforementioned reason. By improving the feature selection techniques used, as well as means to dealing with missing data, better results can be expected.

5.1.2 Feature Creation

In this step of preprocessing, some of the sensors were combined together, such as the sensors which are in close proximity to each other. However, some of the sensors may be better off un-combined, since some of the information may be lost during the combination process. Further analysis into the difference in the amount of information these sensors may improve the understanding of these sensors, and thus improve the quality of this project.

5.2 Performance

For the most part of the algorithms, they were straight-forward implementations of the specified requirements, which means that there are rooms for further improvements to the designed algorithm, both runtime-wise and memory-wise. The main focus for major improvements in this area is to improve the way data are manipulated, since some parts of the algorithm may not be necessary for the desired outcome of this project.

Chapter 6

Conclusion

In conclusion, the data from Gitana 17 proves to be very useful for gaining better insights about how the wind conditions affect each sensors and how each of the sensors affects the boat's performance. Due to the high amount of data provided, many meaningful descriptive statistics were created.

Though some difficulties were posed by the data, such as duplicate and missing values, meaningful statistics were successfully constructed and analyzed through bins creation, where each bin contains a large enough number of data points while maintaining enough locality to be useful for the Gitana 17 team. Moreover, principal component analysis was performed to gain further insights into how each sensors of the boat is important to the boat's performance.

Bibliography

[1] NumPy: NumPy,
http://www.numpy.org

[2] Pandas: Python Data Analysis Library,
https://pandas.pydata.org

[3] Matplotlib: Matplotlib Version 2.2.2,
https://matplotlib.org

[4] Seaborn: Seaborn Version 0.9.0,
https://seaborn.pydata.org/index.html

[5] The Python Package Index: Chardet 3.0.4,
https://pypi.org/project/chardet/

[6] Sci-learn: sklearn.decomposition.PCA,
http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[7] Jupyter: Installing Jupyter,
http://jupyter.org/install

Appendix A

Descriptive Statistic 1

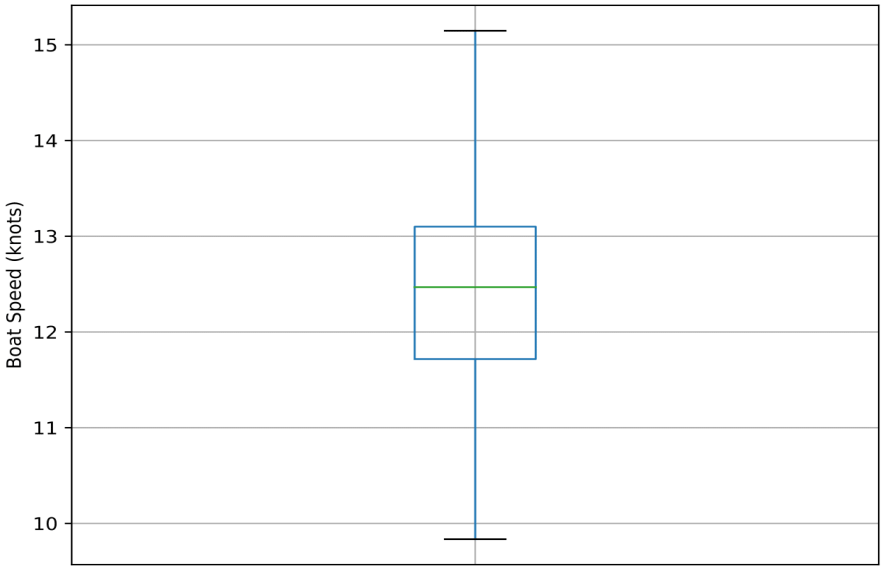

Figure A.1: First Bin's Boxplot

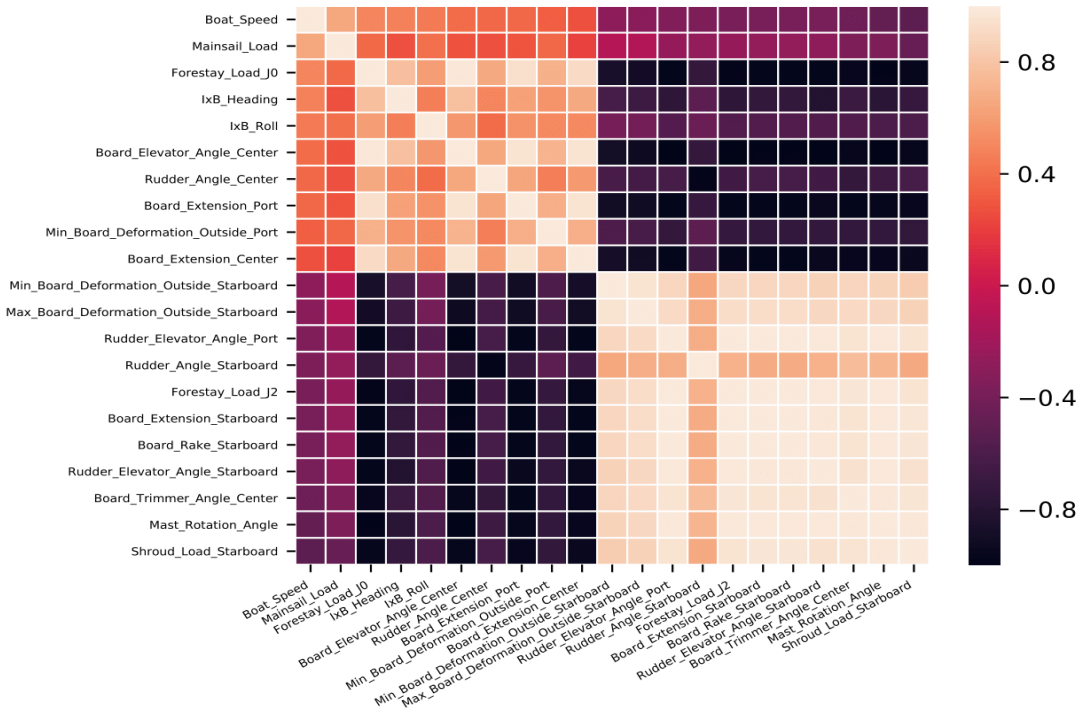

Figure A.2: First Bin's Heatmap

Appendix B

Descriptive Statistic 2

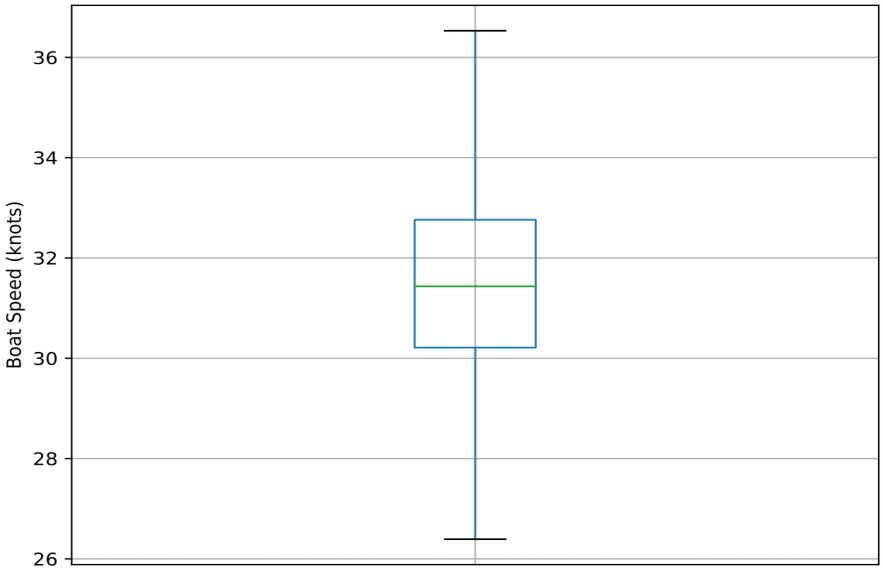

Figure B.1: Second Bin's Boxplot

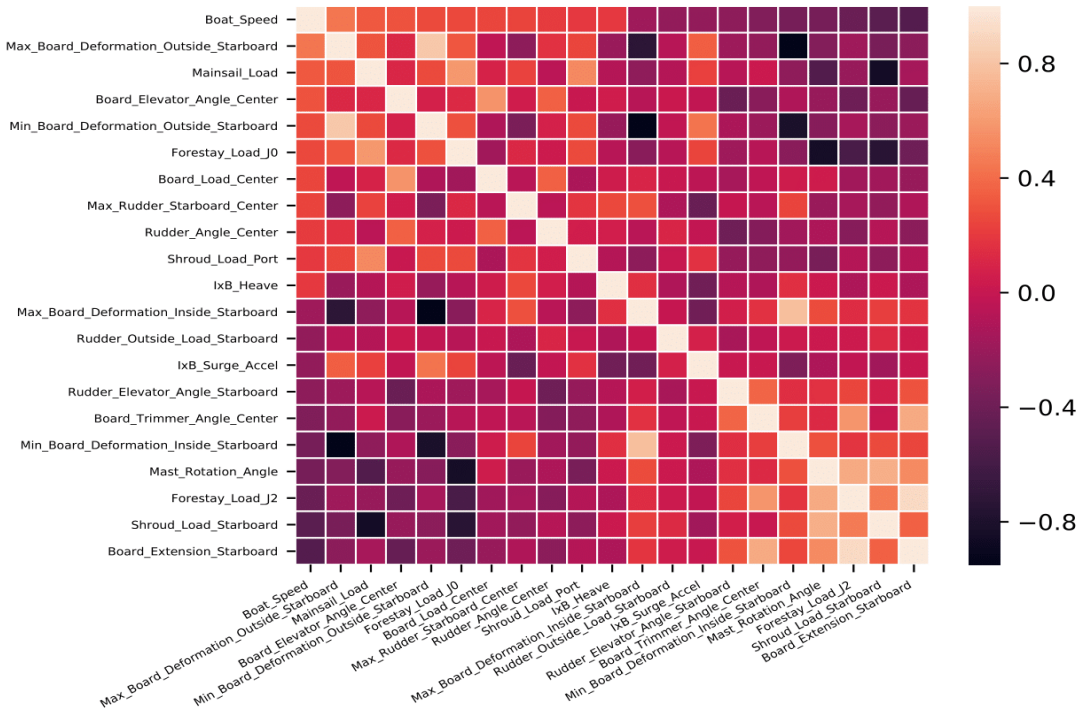

Figure B.2: Second Bin's Heatmap

Appendix C

Descriptive Statistic 3

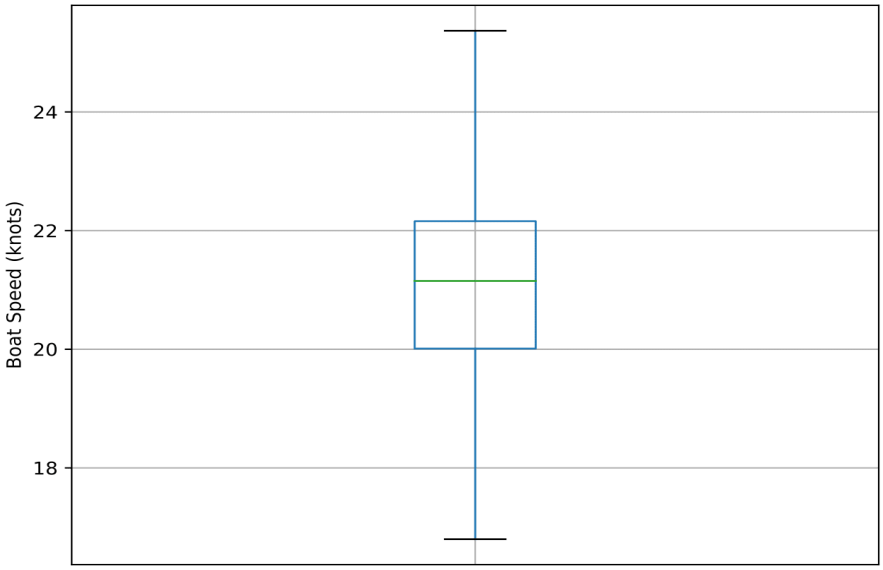

Figure C.1: Third Bin's Boxplot

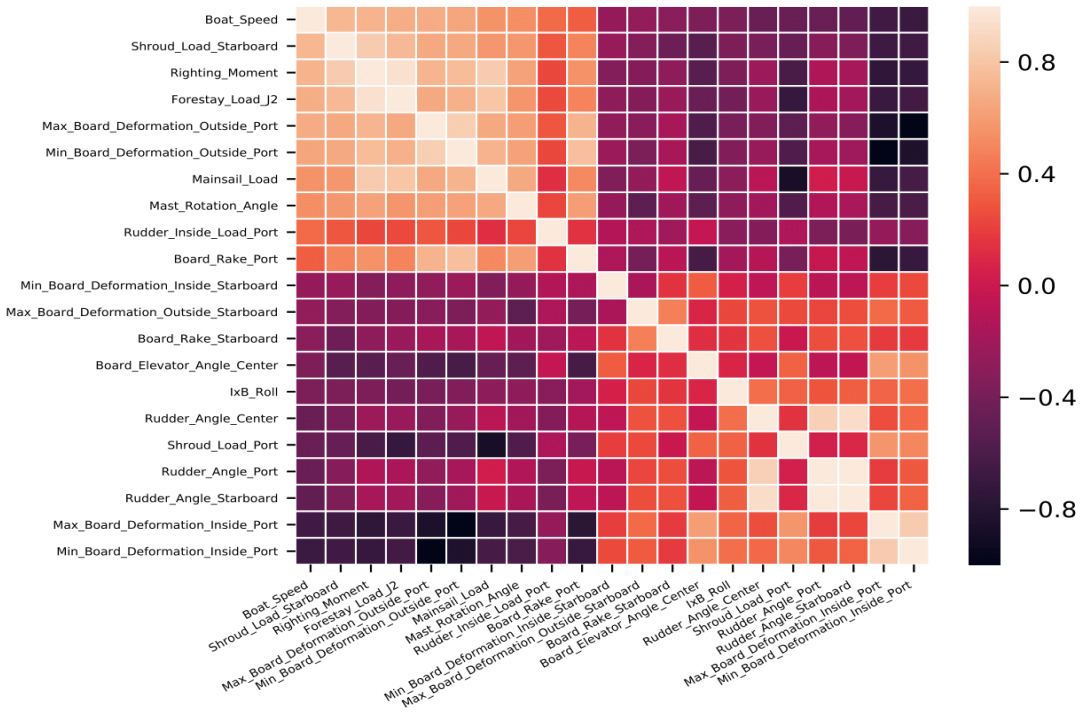

Figure C.2: Third Bin's Heatmap

Appendix D

Source Code

https://github.com/mickeycj/france-internship-project