Boston Buoy Data Analysis

MA 615 Team Project

Zixuan Liu (zliu203)

Chi Zhang (zc9714)

Jiaheng Li (jli305)

SEP 24th 2020

1. Introduction

1.1 Background Research and Project Aims

In seeking evidence of the forming of global warming, our group tried to reformat and analyze the data collected by a single weather buoy in the NOAA National Data Buoy Center. To make an approach to our research goal, we used Rstudio to modify the data and distracting our teamwork through Zoom and Github. During the process of handling the data, we made substitutions for NA data, transform the date-time data into POSIX numbers using lubridate, and determine an appropriate sampling frequency. After refining the data by using R coding, we finally have an objective view of our initial goal.

1.2.1 Variables

There are total 19 variables. The first 13 independent variables in this project is

“WDIR”- Wind Direction (WDIR): WNW (300 deg true)

“WSPD”-Wind Speed (WSPD): 9.7 kts

“GST”- Wind Gust (GST): 11.7 kts

“WVHT”- Wave Height (WVHT): 3.6 ft

“DPD”- Dominant Wave Period Dominant Wave Period (DPD): 13 sec

“APD”- Average Period Average Period (APD): 6.8 sec

“MWD”- Mean Wave Direction Mean Wave Direction (MWD): E (84 deg true)

“PRES”- Atmospheric Pressure Atmospheric Pressure (PRES): 29.71 in

“WTMP”- Water Temperature Water Temperature (WTMP): 61.3 °F

“DEWP”- Dew Point Dew Point (DEWP): 55.8 °F

"VIS- Visibility

“TIDE”- Tide

The dependent variable in this project is

“ATMP”- Air Temperature Air Temperature (ATMP): 67.5 °F

1.2.2 10 Observations

First, we download Historical data from NOAA National Data Buoy Center. Read data ofNDBC Station 44013, years from 1987 to 2019. We found out that this dataset contain a huge amount of data, total of 276411 rows. We first separate the raw data to 5 smaller one which contain same column within each set. then convert ‘YY’, ‘MM’, ‘DD’ to a single variable called ‘DATE’, and ‘HH’, ‘mm’ to ‘TIME’. The head six observations are listed below:

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WD <chr>	WSPD <chr>	GST <chr>	WVHT <chr>	DPD <chr>	APD <chr>
1	1987-01-01	1	00:00:00	290	08.0	10.0	02.70	11.10	08.60
2	1987-01-01	1	01:00:00	290	07.0	08.0	02.40	10.00	08.00
3	1987-01-01	1	02:00:00	290	06.0	08.0	02.50	11.10	08.30
4	1987-01-01	1	03:00:00	300	06.0	07.0	02.60	11.10	08.60
5	1987-01-01	1	04:00:00	290	05.0	06.0	02.70	12.50	08.70
6	1987-01-01	1	05:00:00	340	06.0	07.0	02.40	14.30	08.40

6 rows | 1-10 of 16 columns

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WD <dbl>	WSPD <dbl>	GST <dbl>	WVHT <dbl>	DPD <dbl>	APD <dbl>
1	1999-01-01	1	00:00:00	221	5.4	7.2	0.33	11.11	5.25
2	1999-01-01	1	01:00:00	218	5.6	7.3	0.31	11.11	5.51
3	1999-01-01	1	02:00:00	226	5.7	7.0	0.32	12.50	6.53
4	1999-01-01	1	03:00:00	228	5.7	7.3	0.31	11.11	6.17
5	1999-01-01	1	04:00:00	237	5.9	7.7	0.39	11.11	5.02
6	1999-01-01	1	05:00:00	235	5.7	7.5	0.42	11.11	4.81

6 rows | 1-10 of 16 columns

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WD <dbl>	WSPD <dbl>	GST <dbl>	WVHT <dbl>	DPD <dbl>	APD <dbl>
1	2000-01-01	1	00:00:00	315	0.8	1.5	0.54	10.00	4.55
2	2000-01-01	1	01:00:00	271	0.7	1.9	0.53	4.55	4.61
3	2000-01-01	1	02:00:00	232	2.3	3.1	0.52	4.76	4.78
4	2000-01-01	1	03:00:00	236	2.9	3.8	0.52	10.00	4.86
5	2000-01-01	1	04:00:00	232	4.1	4.9	0.50	10.00	5.00
6	2000-01-01	1	05:00:00	228	5.4	6.5	0.46	4.55	4.94

6 rows | 1-10 of 17 columns

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WD <dbl>	WSPD <dbl>	GST <dbl>	WVHT <dbl>	DPD <dbl>	APD <dbl>
1	2005-01-01	1	00:00:00	187	9.0	10.1	0.85	3.23	3.58
2	2005-01-01	1	01:00:00	193	8.3	10.5	0.78	3.85	3.57
3	2005-01-01	1	02:00:00	212	7.0	8.9	0.90	3.13	3.50
4	2005-01-01	1	03:00:00	210	8.8	11.2	1.00	4.17	3.70
5	2005-01-01	1	04:00:00	237	6.8	8.0	1.00	3.33	3.79

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WD <dbl>	WSPD <dbl>	GST <dbl>	WVHT <dbl>	DPD <dbl>	APD <dbl>
6	2005-01-01	1	05:00:00	210	5.7	6.7	0.84	3.57	3.82

6 rows | 1-10 of 17 columns

	DATE <date>	MONTH <dbl>	TIME <S3: hms>	WDIR <chr>	WSPD <chr>	GST <chr>	WVHT <chr>	DPD <chr>	APD <chr>
1	2007-01-01	1	00:00:00	209	2.2	3.1	0.39	5.00	4.32
2	2007-01-01	1	01:00:00	184	1.6	2.3	0.37	4.76	4.33
3	2007-01-01	1	02:00:00	173	2.8	4.0	0.33	4.17	4.24
4	2007-01-01	1	03:00:00	183	5.6	6.7	0.32	4.17	4.19
5	2007-01-01	1	04:00:00	201	6.0	7.1	0.32	2.50	3.38
6	2007-01-01	1	05:00:00	163	4.4	5.2	0.32	2.50	3.37

6 rows | 1-10 of 17 columns

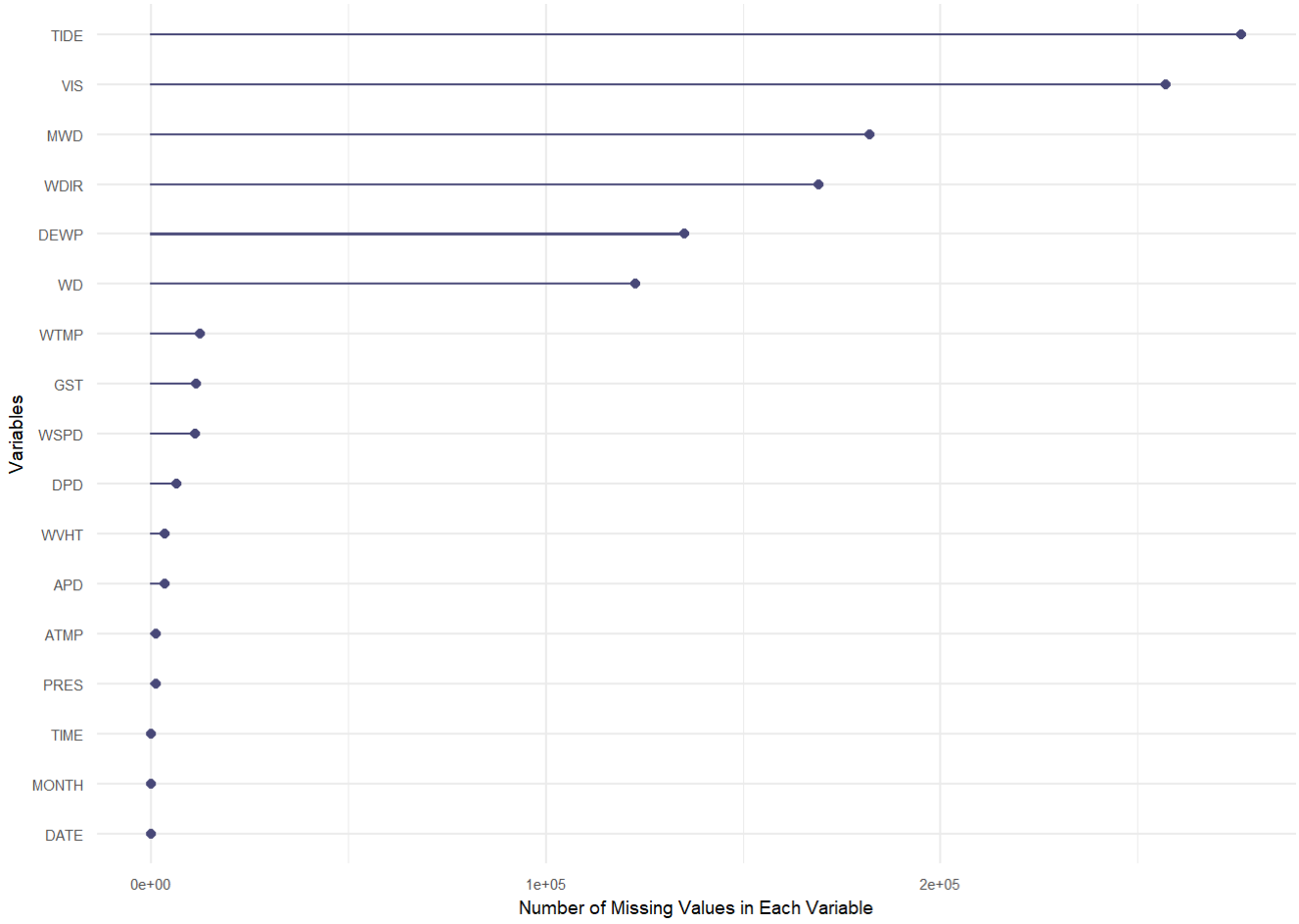
2. Summary Statistics and Data Visualization

2.1 Missing Values & Data Preprocessing

2.1.1 Missing Values

First We conduct basic data preprocessing. Missing values for dataset are shown in the histogram below.

```
##       DATE                MONTH             TIME            WDIR
##  Min.   :1987-01-01   Min.   : 1.000   Min.   :    0   Min.   :  1.0
##  1st Qu.:1995-03-23   1st Qu.: 4.000   1st Qu.:21000   1st Qu.:124.0
##  Median :2003-06-28   Median : 7.000   Median :42600   Median :203.0
##  Mean   :2003-07-08   Mean   : 6.535   Mean   :42451   Mean   :195.8
##  3rd Qu.:2011-07-04   3rd Qu.:10.000   3rd Qu.:64200   3rd Qu.:281.0
##  Max.   :2019-12-31   Max.   :12.000   Max.   :85800   Max.   :360.0
##                                                        NA's   :169114
##        WD              WSPD             GST              WVHT
##  Min.   :  0.0    Min.   : 0.000   Min.   : 0.0    Min.   :0.000
##  1st Qu.:127.0    1st Qu.: 3.600   1st Qu.: 4.3    1st Qu.:0.400
##  Median :211.0    Median : 5.500   Median : 6.6    Median :0.650
##  Mean   :197.8    Mean   : 6.079   Mean   : 7.4    Mean   :0.864
##  3rd Qu.:280.0    3rd Qu.: 8.100   3rd Qu.: 9.8    3rd Qu.:1.060
##  Max.   :360.0    Max.   :25.700   Max.   :32.4    Max.   :9.100
##  NA's   :122835   NA's   :11033    NA's   :11295   NA's   :3467
##       DPD              APD              MWD              PRES
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.0    Min.   : 964.6
##  1st Qu.: 4.550   1st Qu.: 3.900   1st Qu.: 78.0    1st Qu.:1010.2
##  Median : 7.690   Median : 4.780   Median : 94.0    Median :1015.7
##  Mean   : 7.378   Mean   : 5.007   Mean   :124.3    Mean   :1015.5
##  3rd Qu.:10.000   3rd Qu.: 5.900   3rd Qu.:129.0    3rd Qu.:1021.1
##  Max.   :25.000   Max.   :12.100   Max.   :360.0    Max.   :1045.8
##  NA's   :6319     NA's   :3467     NA's   :182225   NA's   :1117
##       ATMP             WTMP             DEWP             VIS
##  Min.   :-19.700   Min.   :-1.80   Min.   :-24.90   Min.   : 0.00
##  1st Qu.:  3.600   1st Qu.: 5.10   1st Qu.: -0.50   1st Qu.: 8.10
##  Median :  9.700   Median : 9.80   Median :  7.00   Median : 9.40
##  Mean   :  9.671   Mean   :10.49   Mean   :  6.28   Mean   :12.48
##  3rd Qu.: 16.700   3rd Qu.:15.70   3rd Qu.: 14.50   3rd Qu.:11.60
##  Max.   : 32.100   Max.   :27.80   Max.   : 26.10   Max.   :36.00
##  NA's   :1172      NA's   :12288   NA's   :135148   NA's   :257172
##    TIDE
##  Mode:logical
##  NA's:276411
##
##
##
##
##
```



The plot above shows that TIDE has the highest missing value.

2.1.2 Dealing With Missing Values

Due to the large number of missing values in this dataset, we decided to delete those variables which has over 10,000 missing values, for those variables which has less than 10,000 missing values, we use variable means to replace missing values. The new dataset called data_87_19 which still contain 107,611 rows.

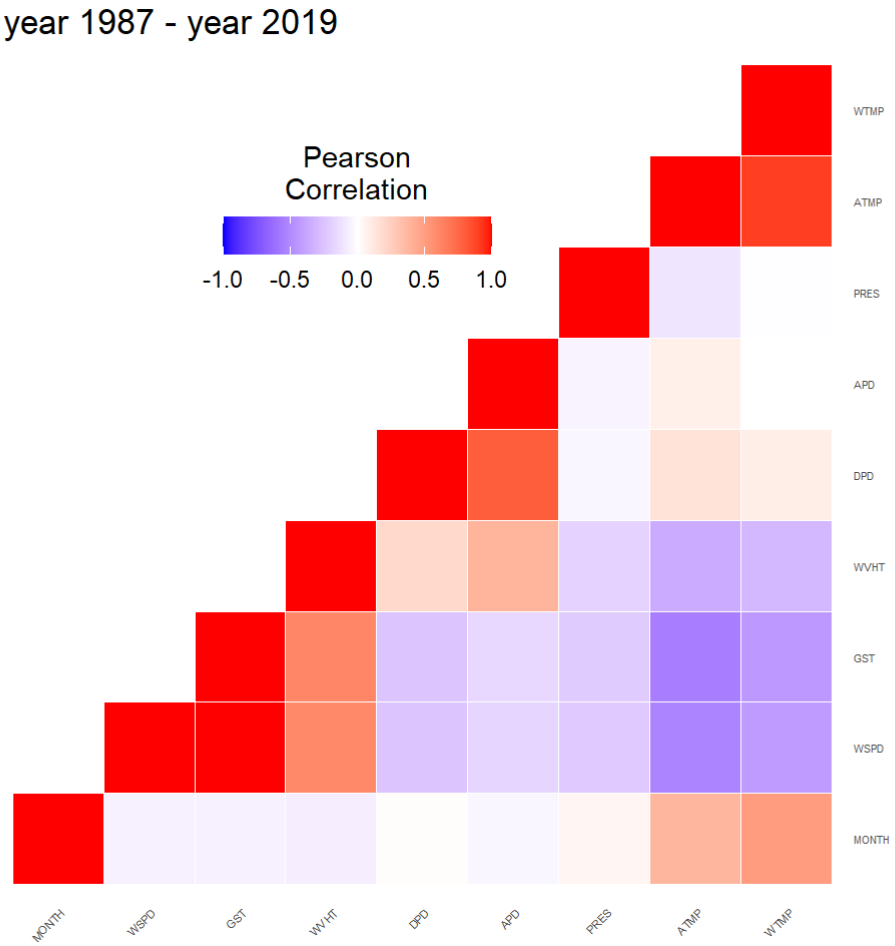
To simplify our model even more, We merge 24 row of hours into one day, that result to our final data 'final' and it has 11,576 rows.

DATE <date>	M... <dbl>	WSPD <dbl>	GST <dbl>	WVHT <dbl>	DPD <dbl>	APD <dbl>	PRES <dbl>	ATMP <dbl>
1987-01-01	1	4.000000	5.250000	2.0833333	11.900000	8.441667	1025.392	1.700000
1987-01-02	1	11.684211	14.578947	2.7631579	9.463158	6.873684	1004.784	3.057895
1987-01-03	1	10.041667	12.333333	1.9250000	10.387500	6.075000	999.575	-1.116667
1987-01-04	1	7.125000	8.875000	0.8250000	9.254167	4.770833	1017.875	-2.245833
1987-01-05	1	5.208333	6.333333	0.5833333	6.091667	4.270833	1021.879	0.287500
1987-01-06	1	4.291667	5.416667	0.5500000	5.270833	4.275000	1023.183	0.262500

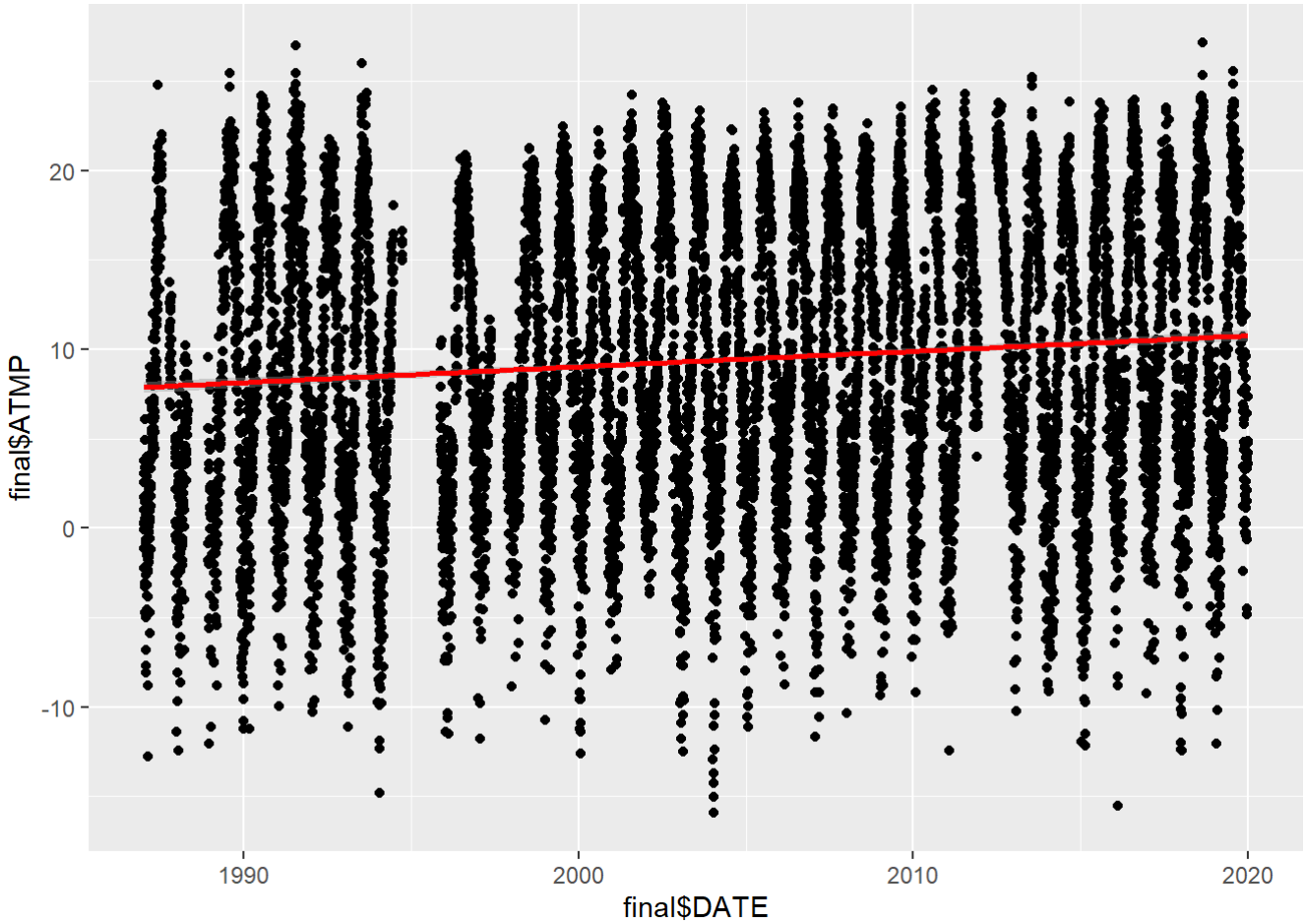
6 rows

2.1.2 Heatmap

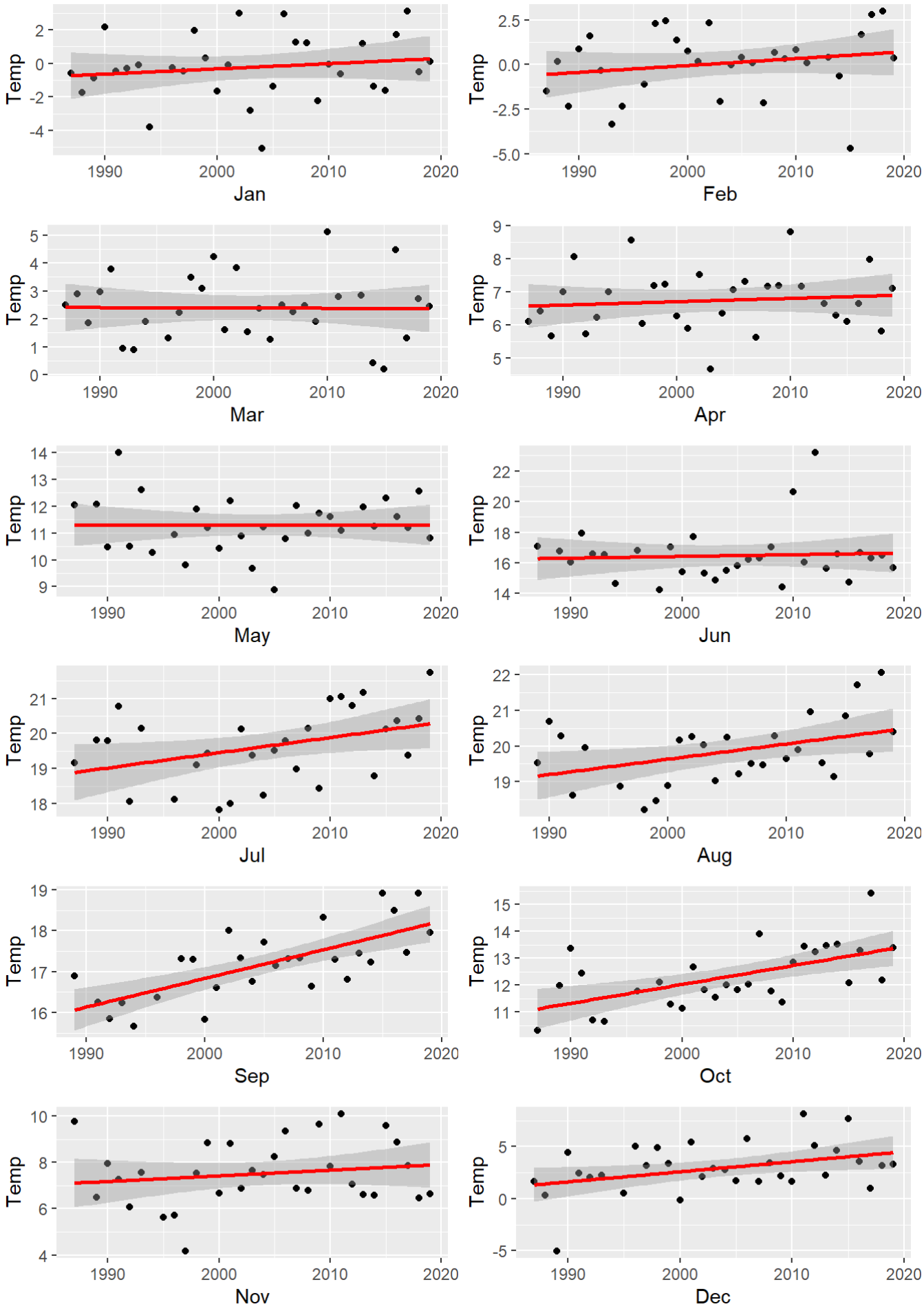
Shown in below is a correlation map for the year 1987 - year 2019 data that describes the relationship between the different features.



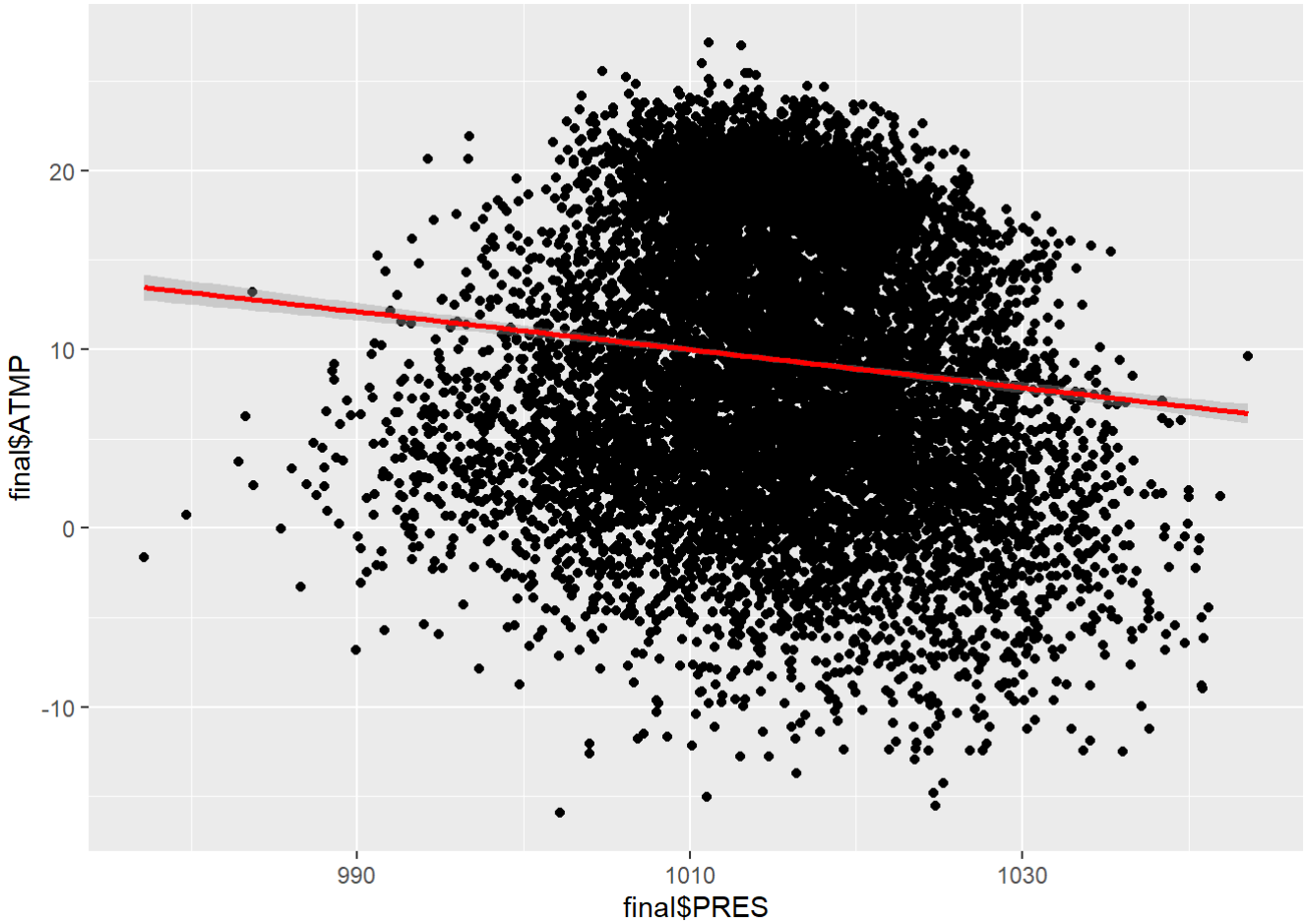
We can find that GST and ATMP have highest negative correlation, WSPD and GST have highest positive correlation.



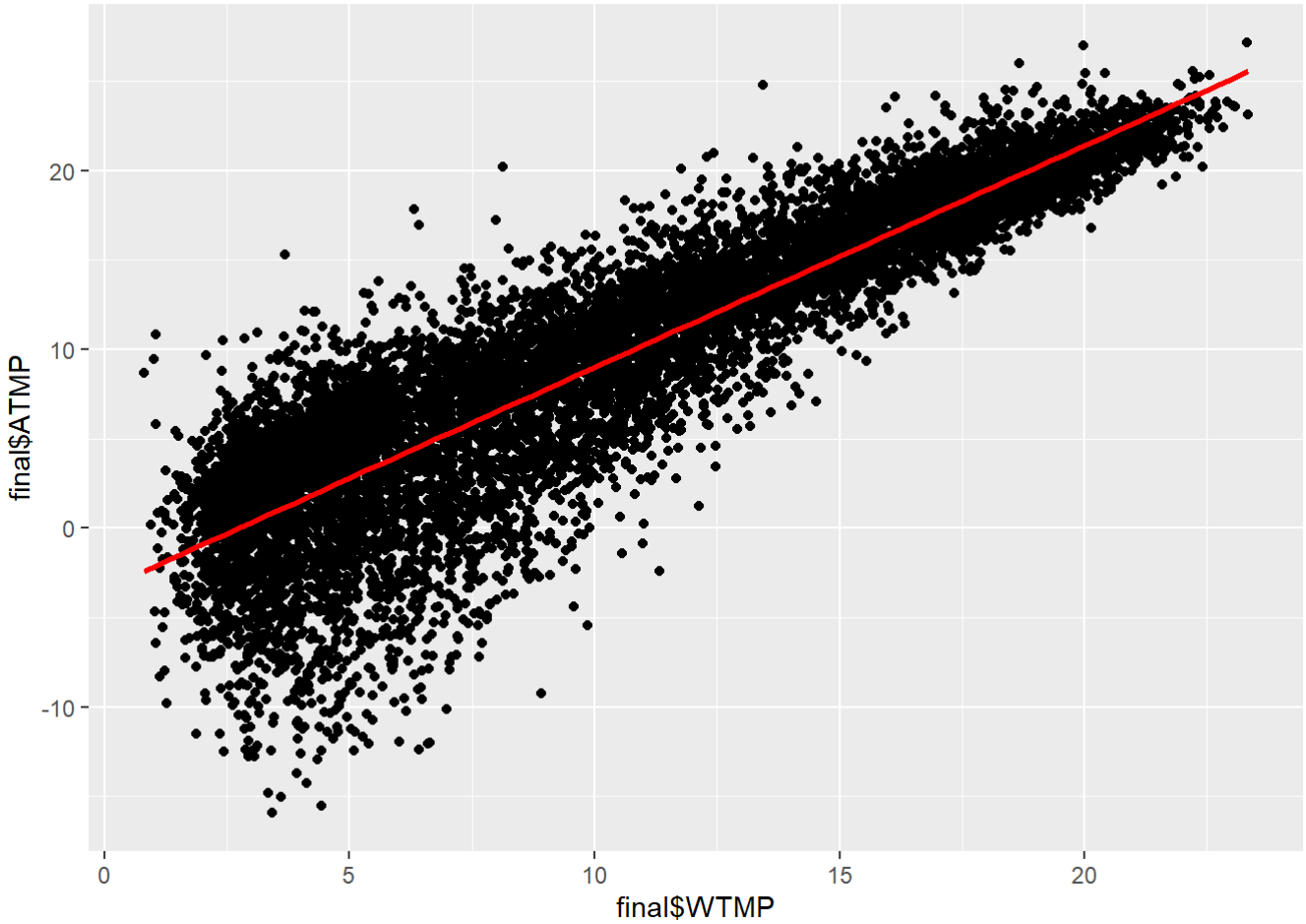
This chart shows the distribution of air temperature in different years. After using regression, we find that overall temperature by year has an increasing tendency. From 1987 to 2020, it has increased by around 2 degrees in Fahrenheit. It is considerable, and there might be more obvious temperature rise in some other spots.



To make the results more intuitive, we regressed the monthly average air temperature from 1987 to 2020. The results show that, especially in the second half of the year, the temperature increased significantly from 1987 to 2020. We may make a conclusion that global warming is relatively more serious in summer and autumn, and the rising tendency in spring and winter is not that obvious.



When the temperature increases, the air warms up and tends to rise - becoming thinner and lighter and therefore weighs less. As a result, atmospheric pressure decreases. From the previous chart, we know that the average temperature increases year by year, and according to physics, we know that atmospheric pressure will decrease with the increase of temperature in open environment. This graph confirms, on average, the inverse relationship between temperature and atmospheric pressure.



It is well known that water can conduct more heat than air. Every time the air temperature rises by one degree, the sea water temperature will rise more. As shown in the figure above, the air temperature is directly proportional to the sea water temperature. With global warming, the average temperature of sea water will also increase by year. As a result, marine pollution will be accelerated and various aquatic plants and animals will be affected.

4. Conclusion

4.1 Obstacles

The first-hand data we got from the weather buoy is kind of massive and rough. Some of the certain data was missing or cannot be observed from a part of the early stage, and we think that it might due to it the tech issue at that time. So our group used R code to transfer a part of those data as ‘NA’ and deleted the column with over 100000 (because if one column has too many missing values this column would not be useful and convincing enough to analyze the data). Also, the format of time in the original data set was divided into several columns as ‘MM’, ‘YY’, ‘DD’ , etc, which is too distracting during the process of organizing and analyzing the data set. So we first divided the data set into 5 groups which each one of these groups has exactly the same amount of column numbers within it. Then we used R code to modify these groups as one format and merged them into a whole data set. Last but not least, even after deleting and modifying our data set through the first three steps, the size of our database was still too huge to handle. So we used R code to take the mean of each data within a day to get a condensed and operable data set. After the above steps, we were finally able to use regressions and graphs to interpret the data set and find our project target.

4.2 Conclusion

After doing a certain amount of background research, we finally find out that there are 3 elements in our data sets are closely related to the existence of global warming which are ‘Air Temperature’(ATMP), ‘Water Temperature’ (WTMP), and ‘Atmospheric Pressure’ (PRES). To conducting our final conclusion, we are going to narrate our explanation by expanding the causing and relationship between these three elements. First, we noticed that the data of the air temperature observed by the weather buoy among the whole year is gradually growing up from 1987 to 2020. Visually, the increment of 1-2 Celsius degrees may not seem so significant, however, the thing we need to know is that for the climatic environment, sometimes even a slight change may cause a huge consequence of the butterfly effect. For instance, the heat energy brought by the rising temperature will provide huge kinetic energy to the air and ocean, resulting in disasters such as large or even super large typhoons, hurricanes, and tsunamis. Thus, this kind of increment can already be regarded as significant. With further analyzing the data using the regression model, we find out there is a conspicuous relation between the water temperature and the air temperature which are almost perfectly proportional to each other. This also verifies the example I gave in the previous paragraph. What’s more, the rising of the air temperature and water temperature also makes sense to another regression model we made for the atmospheric pressure. From the graphic, we can see there is a declining trend on the graphic of the atmospheric pressure. This change can be explained through a simple Physics formula: P=F/S. It can be seen from the formula that under the same area, the pressure is only related to F. As the temperature rises, the atmosphere becomes thinner and the density becomes smaller; therefore, in the specified area where the atmosphere is thinner, the F becomes smaller, so the air pressure would be relatively lower. As the three elements of ATMP, WTMP, and PRES are all involved in this simple formula, once one of them changes negatively, these three elements will produce a vicious circle which would be a harmful issue for our environment. In summary, we would say that base on the analysis of the data we received from the weather buoy in the NOAA National Data Buoy Center, we can infer that the existence of global warming is positive. Although the reasons behind the formation of global warming may still need more data and information to verify, all of us should pay more attention to this issue because if we want our living environment to become better, it will require the contribution of all the creatures living on this planet. # 5.Reference

National Data Buoy Center (https://www.ndbc.noaa.gov/station_page.php?station=44013)