HOMESTEAD NATIONAL MONUMENT

TEAM DATA WIZARDS

TEAM MEMBERS

MALIK ALZARAH
NISHI MAHATO
VINCENT HA
VISHEKHA TAMRAKAR

AGENDA

- INTRODUCTION
- RESEARCH QUESTIONS
- WHY THE QUESTIONS ARE IMPORTANT AND WHO SHOULD CARE
- DATA DESCRIPTION
- ANALYSIS APPROACH AND INTERPRETATION
- RECOMMENDATION
- REFERENCES

INTRODUCTION

- Homestead National Monument of America (HOME) is a unit of National Park Service
- Allows qualified people to claim up to 160 acres of land owned by the federal government in exchange for:
  - Five years of residence
  - Cultivation
  - Improving the property in every possible way.
- It allows a fair chance to any mankind to claim land for free in exchange for taking care of it keeping it free from contamination, persuading people to take the best care of it.
- The Homestead Act brought far-reaching effects on the landscape and in the lives of the people.

RESEARCH QUESTION

1. **Does precipitation affect the amount of E.coli for North Site and West Site?**

2. **How E.coli variable varied in North Site and West Site from the year 2013 till 2017?**

3. **Does Turbidity and Conductivity varies in North Site and West Site?**

4. **What factors(Dissolved Oxygen and Alkalinity) affect pH level of water in North Site and West Site?**

WHY THE QUESTIONS ARE IMPORTANT AND WHO SHOULD CARE

- Question No.1 is important because we want to check if there was contamination water (human or animal waste) flowing in the water stream.

- Question No.2 is the follow up question to the first question and it helps identify the changes in E.Coli colonies between the North and West sites.

- Question No.3 is to cover other types of water contamination (sediments from erosion, waste discharge, algae growth, sodium, chloride, oil,etc.) by looking at the Turbidity and Conductivity between the North and West sites.

- **Question No. 4** in a sense is important to figure out what Nutrient factors such as Alkalinity or Dissolved oxygen affect the quality in both North and West Site. Since, the dissolved oxygen. Dissolved oxygen in surface water is used by all forms of aquatic life; therefore, this constituent typically is measured to assess the "health" of lakes and streams. Also, the Alkalinity has the capacity of water to resist changes in pH that would make the water more acidic and therefore it would be interesting to know how these factors affect pH level of water

DATA DESCRIPTION

Where the data came from?

- Data was received from our client Ranger Jesse Bolli.
- Data set includes:
  - Water Temperature, pH level of the water, Conductivity, Turbidity, Nitrates, Total Phosphate, and Alkalinity Average

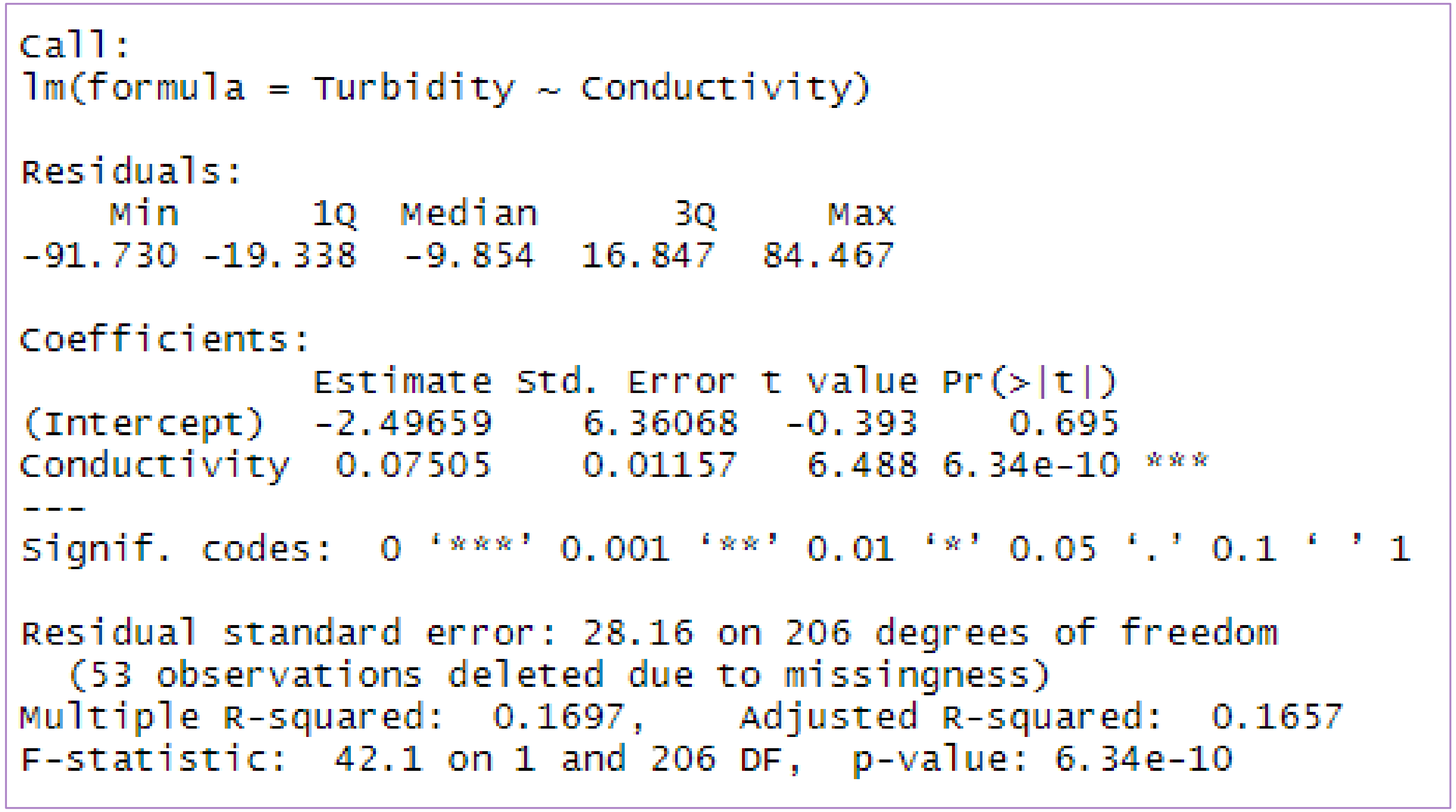
How we used the data?

- These data can be used for the data analysis but should not be used for commercializing and any means of business.
- Observations that were recorded from the year 2013 till 2017.
- The observations provided us with enough information to make an effective analysis.

Limitation of data.

- Unnecessary rows and columns.
- There are three separate columns for a date (Day, Month, Year).
- Columns that were supposed to hold numeric values just like other rows in the column but holds value such as NV instead of NULL or be an empty cell.
- Rows with missing values or no values.
- Same name has written in different patterns in the column Observation Site.

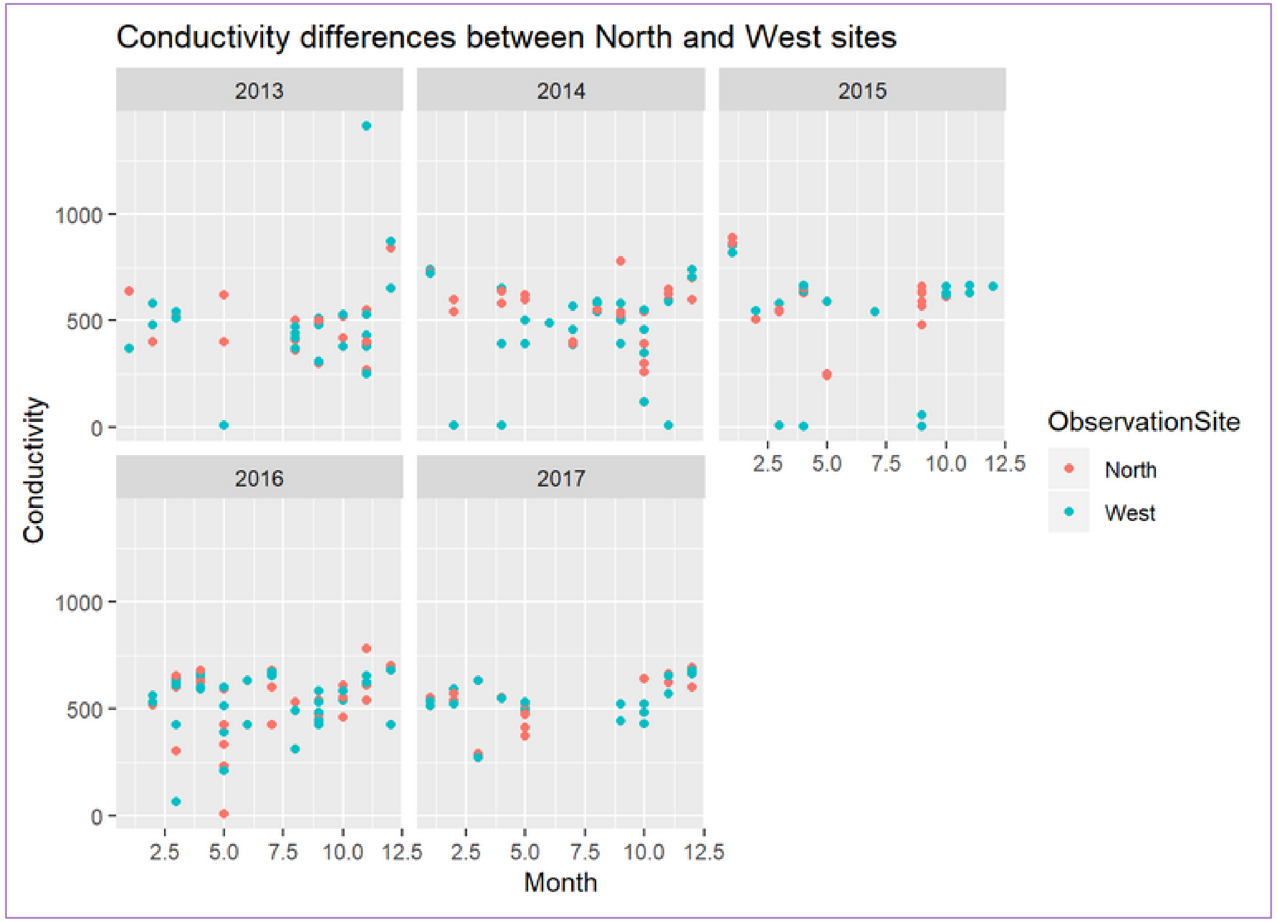
ANALYSIS APPROACH



```
Call:
lm(formula = Turbidity ~ Conductivity)

Residuals:
    Min      1Q  Median      3Q     Max 
-91.730 -19.338  -9.854  16.847  84.467 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.49659    6.36068  -0.393    0.695    
Conductivity  0.07505    0.01157   6.488 6.34e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.16 on 206 degrees of freedom
  (53 observations deleted due to missingness)
Multiple R-squared:  0.1697,	Adjusted R-squared:  0.1657 
F-statistic:  42.1 on 1 and 206 DF,  p-value: 6.34e-10
```

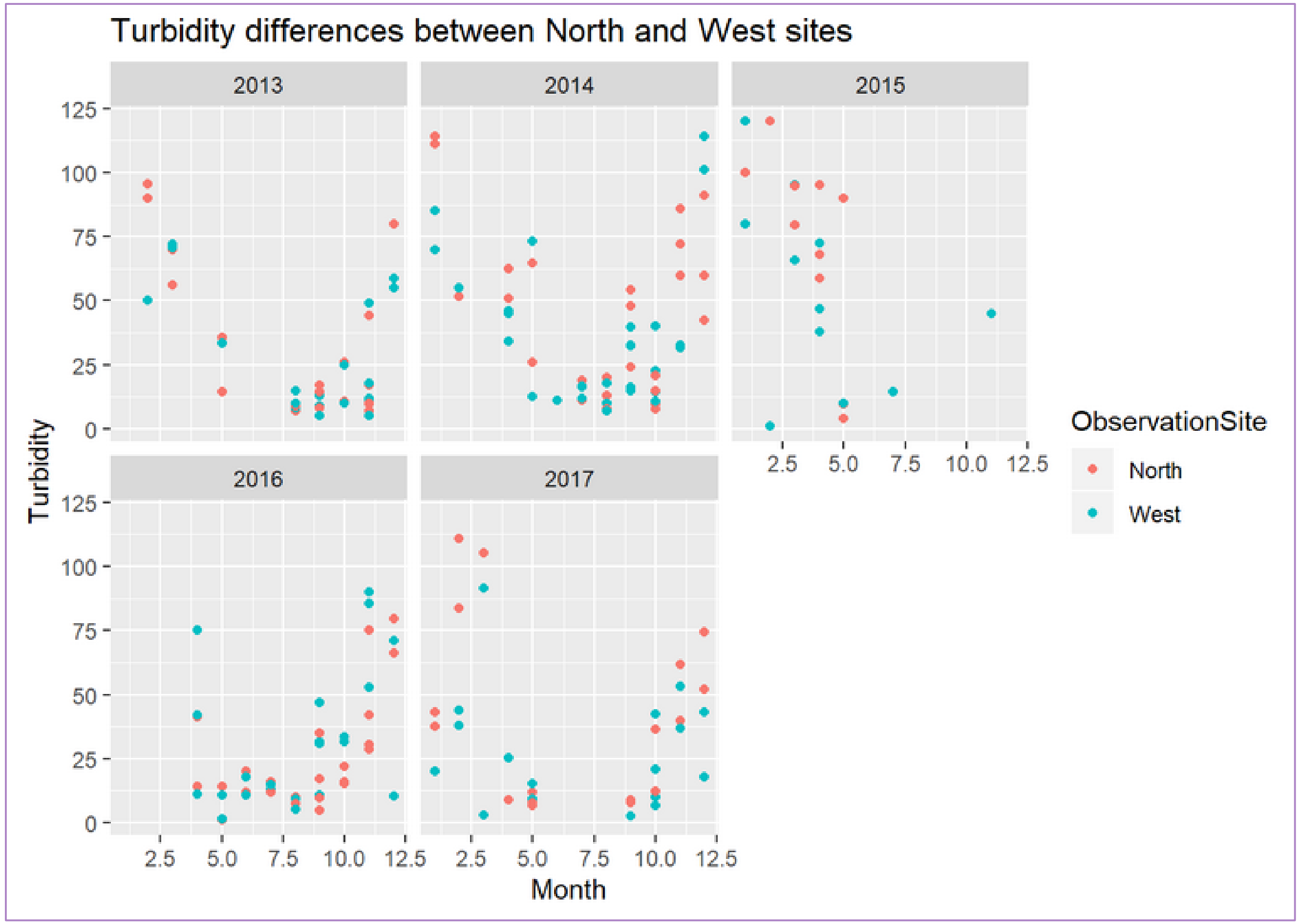
- Identify the connection between Turbidity and Conductivity
- Using basic linear regression
- Some observations got deleted during the process
- The result shows that there is no linear relation with the Adjusted R-squared of 0.1657

ANALYSIS INTERPRETATION



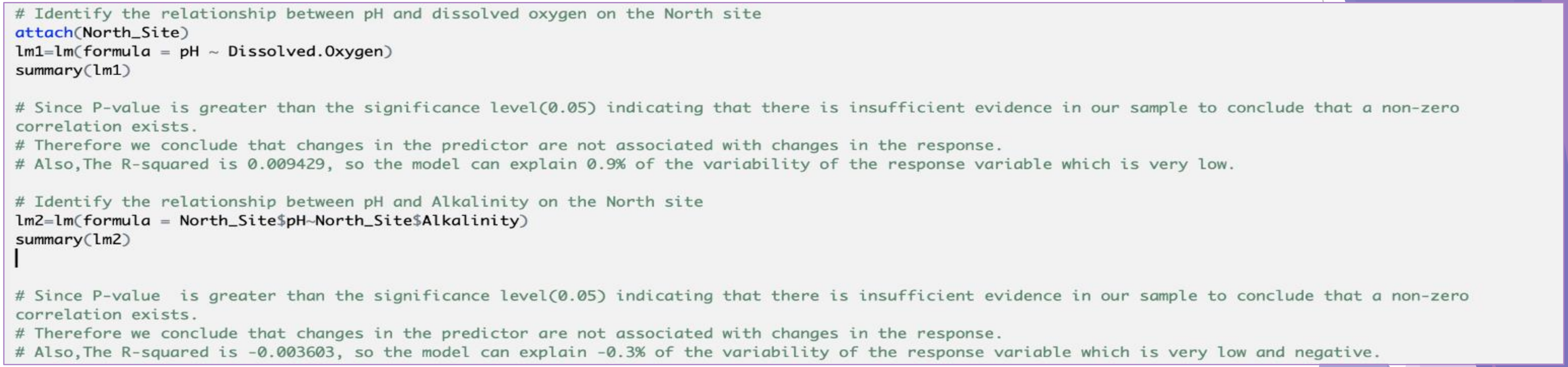
ggplot(data = dataset) + geom_point(mapping = aes(x = Month, y = Conductivity, color = ObservationSite)) + facet_wrap(~ Year, nrow = 2)+ggtitle("Conductivity differences between North and West sites")

ANALYSIS INTERPRETATION

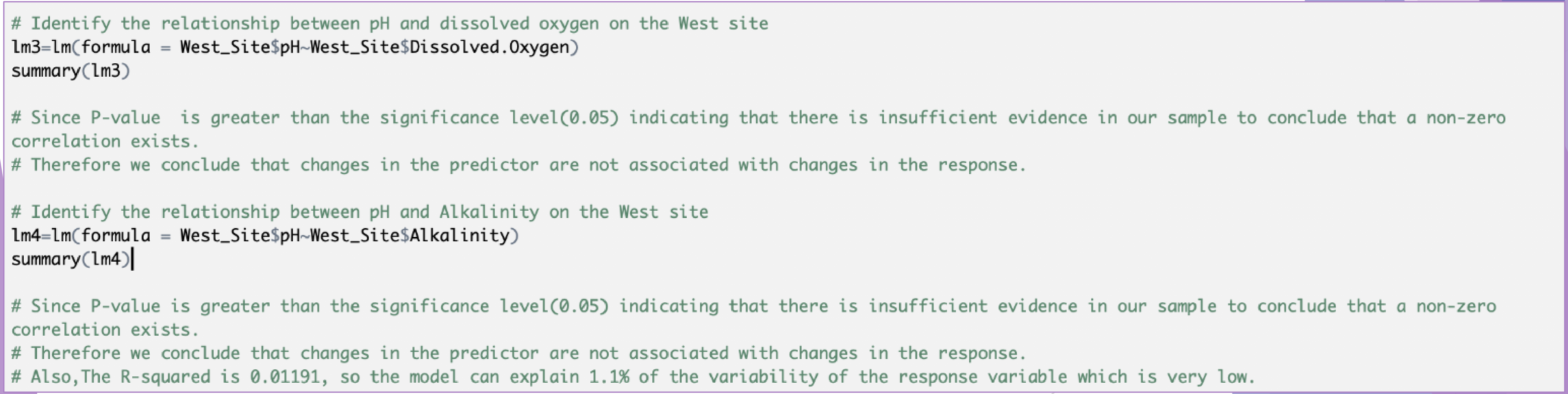


ggplot(data = dataset) + geom_point(mapping = aes(x = Month, y = Turbidity, color = ObservationSite)) + facet_wrap(~ Year, nrow = 2)+ggtitle("Turbidity differences between North and West sites")

ANALYSIS APPROACH

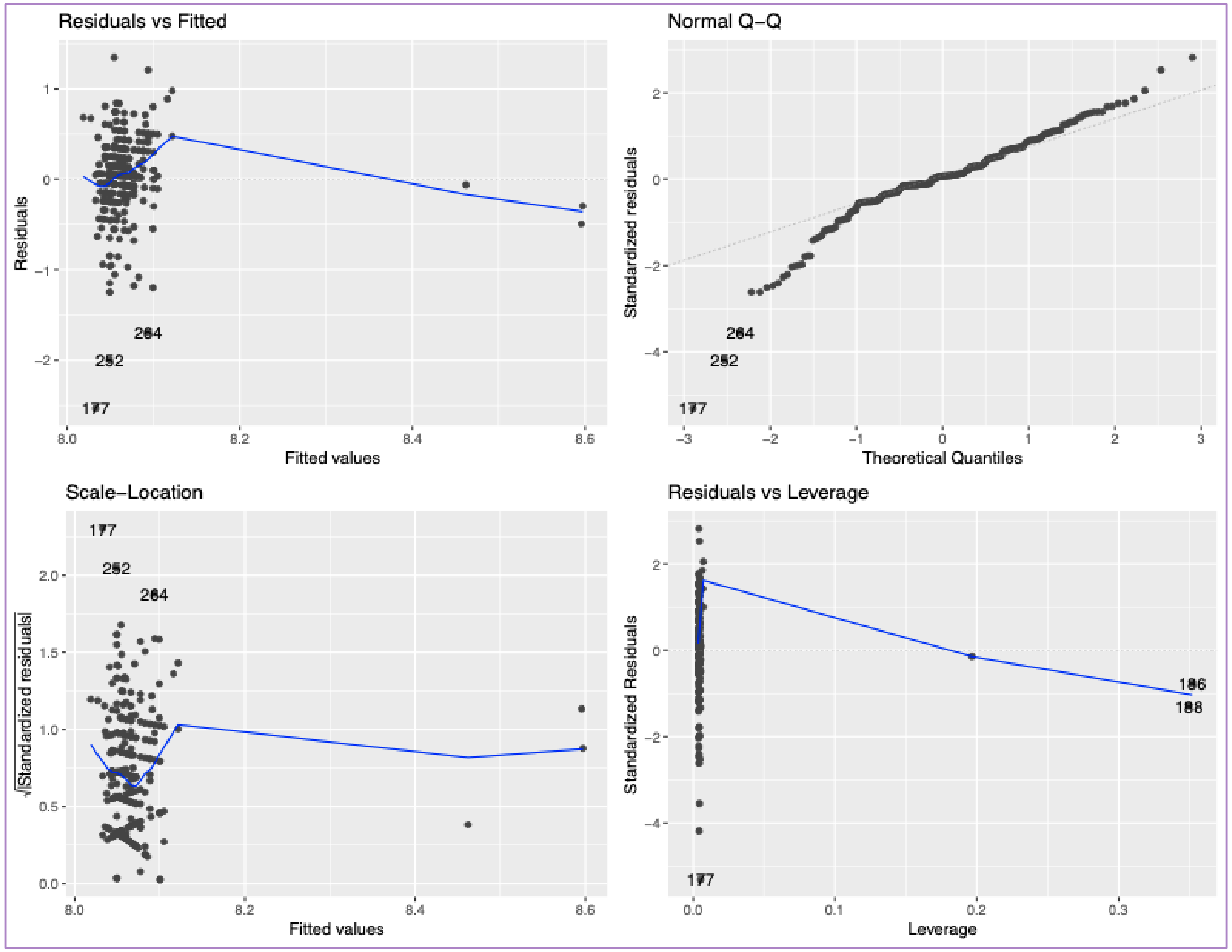


```
# Identify the relationship between pH and dissolved oxygen on the North site
attach(North_Site)
lm1=lm(formula = pH ~ Dissolved.Oxygen)
summary(lm1)

# Since P-value is greater than the significance level(0.05) indicating that there is insufficient evidence in our sample to conclude that a non-zero
correlation exists.
# Therefore we conclude that changes in the predictor are not associated with changes in the response.
# Also,The R-squared is 0.009429, so the model can explain 0.9% of the variability of the response variable which is very low.

# Identify the relationship between pH and Alkalinity on the North site
lm2=lm(formula = North_Site$pH~North_Site$Alkalinity)
summary(lm2)


# Since P-value  is greater than the significance level(0.05) indicating that there is insufficient evidence in our sample to conclude that a non-zero
correlation exists.
# Therefore we conclude that changes in the predictor are not associated with changes in the response.
# Also,The R-squared is -0.003603, so the model can explain -0.3% of the variability of the response variable which is very low and negative.
```



```
# Identify the relationship between pH and dissolved oxygen on the West site
lm3=lm(formula = West_Site$pH~West_Site$Dissolved.Oxygen)
summary(lm3)

# Since P-value  is greater than the significance level(0.05) indicating that there is insufficient evidence in our sample to conclude that a non-zero
correlation exists.
# Therefore we conclude that changes in the predictor are not associated with changes in the response.

# Identify the relationship between pH and Alkalinity on the West site
lm4=lm(formula = West_Site$pH~West_Site$Alkalinity)
summary(lm4)

# Since P-value is greater than the significance level(0.05) indicating that there is insufficient evidence in our sample to conclude that a non-zero
correlation exists.
# Therefore we conclude that changes in the predictor are not associated with changes in the response.
# Also,The R-squared is 0.01191, so the model can explain 1.1% of the variability of the response variable which is very low.
```

ANALYSIS INTERPRETATION



A simple linear regression was applied to predict a continuous outcome variable (y) which pH value of water in this case based on one single predictor variable (x) which is level of Dissolved Oxygen in water in this case for North Site.

The outcome of linear regression model was plotted using ggplot which provides us with the above mentioned plot.

Conclusion of ggplot

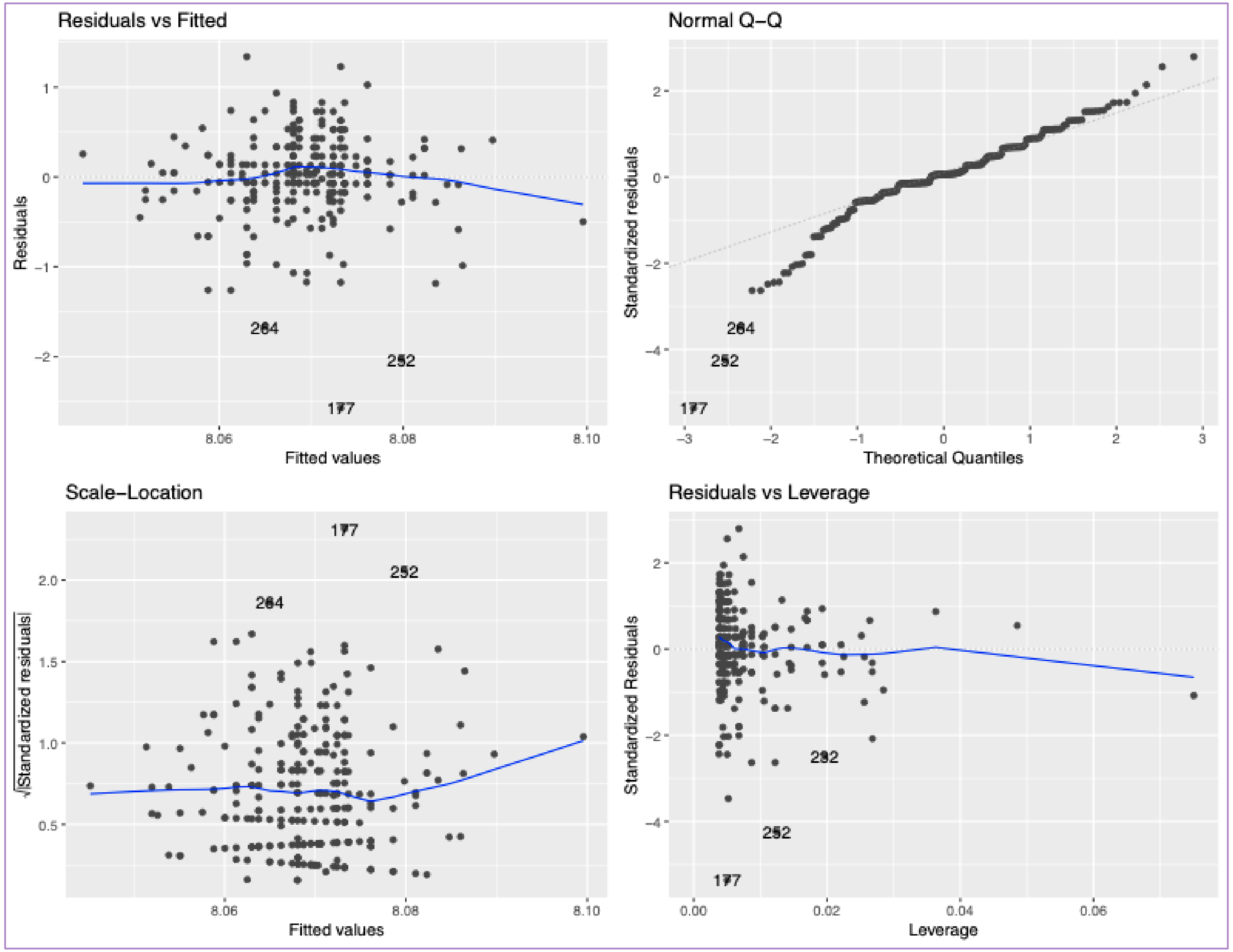
1.Residual Vs Fitted: Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, which is good. In this case there is almost no pattern in the residual plot. T**his suggests that we cannot assume linear relationship between the predictors and the outcome variables.**

2. Normal Q-Q: Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. **In this case many observations are far away from the expected line therefore not normally distributed**.

3. Standardized Residual Plot: The plot shows if residuals are spread equally along the ranges of predictors. It's good if we see a horizontal line with equally spread points. In this case, this is not the case. **It can be seen that the variability (variances) of the residual points is haphazard with the value of the fitted outcome variable, suggesting no non-constant variances in the residuals errors.**

4. Residual Vs Leverage: Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. I**n this case the two observations 186 and 188 show high leverage point and should be well investigated.**

ANALYSIS INTERPRETATION



A **simple linear regression** was applied to predict a continuous outcome variable (y) which pH value of water in this case based on one single predictor variable (x) which is level of Alkalinity in water in this case for North Site.

Conclusion of ggplot

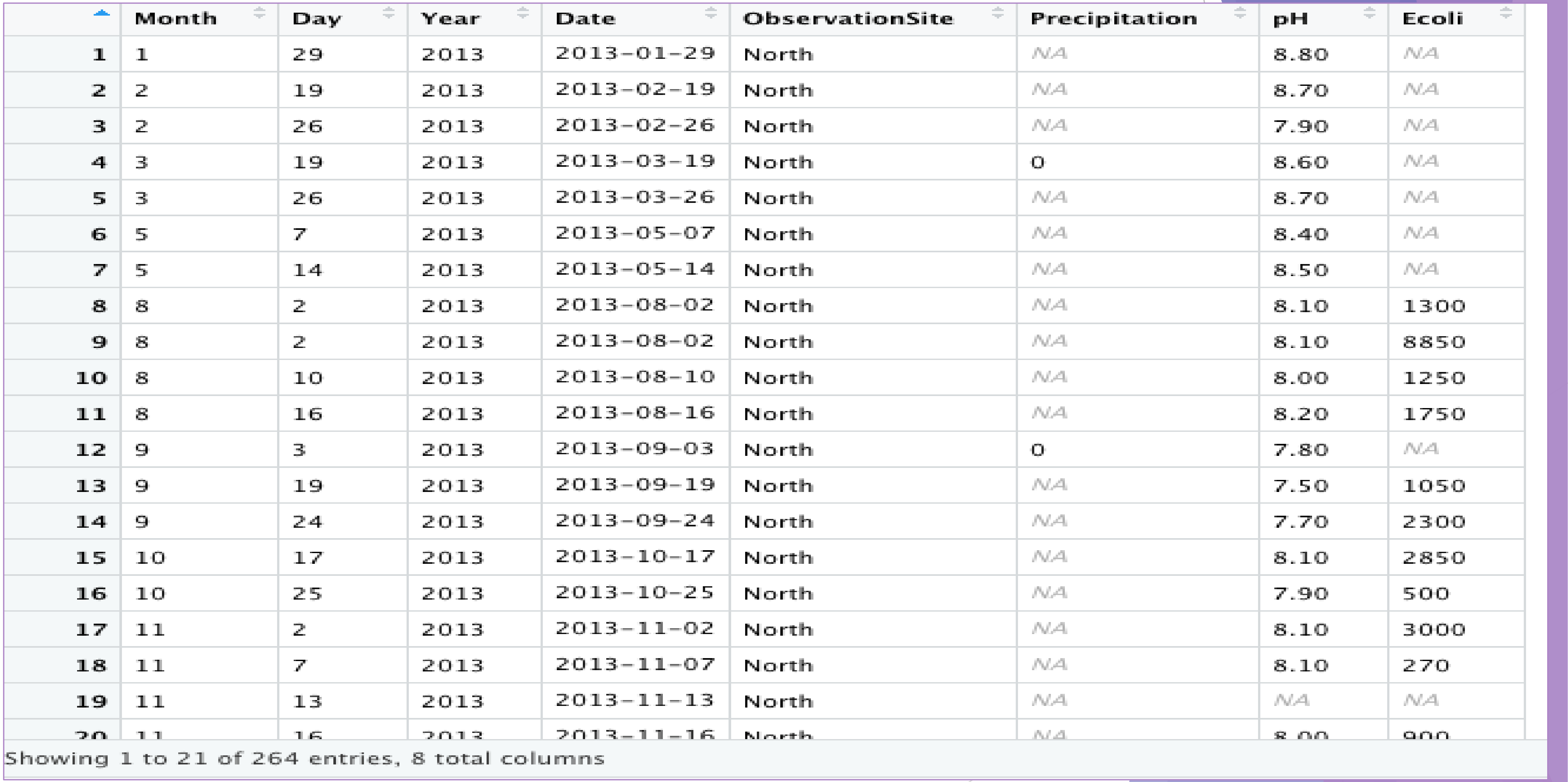
1.Residual Vs Fitted: Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, which is good. **In this case there is no fitted pattern in the residual plot. This suggests that we can assume a linear relationship between the predictors and the outcome variables.**

2. Normal Q-Q: Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. **In this case many observations are far away from the expected line therefore not normally distributed.**

3. Standardized Residual Plot: The plot shows if residuals are spread equally along the ranges of predictors. **It's good if we see a horizontal line with equally spread points. In this case, this is not the case. It can be seen that the variability (variances) of the residual points is random with the value of the fitted outcome variable, suggesting no non-constant variances in the residuals errors.**

4.Residuals Vs Leverage: Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. **In this case the three observations visible as high leverage points and should be well investigated.**

ANALYSIS APPROACH



	Month	Day	Year	Date	ObservationSite	Precipitation	pH	Ecoli
1	1	29	2013	2013-01-29	North	NA	8.80	NA
2	2	19	2013	2013-02-19	North	NA	8.70	NA
3	2	26	2013	2013-02-26	North	NA	7.90	NA
4	3	19	2013	2013-03-19	North	0	8.60	NA
5	3	26	2013	2013-03-26	North	NA	8.70	NA
6	5	7	2013	2013-05-07	North	NA	8.40	NA
7	5	14	2013	2013-05-14	North	NA	8.50	NA
8	8	2	2013	2013-08-02	North	NA	8.10	1300
9	8	2	2013	2013-08-02	North	NA	8.10	8850
10	8	10	2013	2013-08-10	North	NA	8.00	1250
11	8	16	2013	2013-08-16	North	NA	8.20	1750
12	9	3	2013	2013-09-03	North	0	7.80	NA
13	9	19	2013	2013-09-19	North	NA	7.50	1050
14	9	24	2013	2013-09-24	North	NA	7.70	2300
15	10	17	2013	2013-10-17	North	NA	8.10	2850
16	10	25	2013	2013-10-25	North	NA	7.90	500
17	11	2	2013	2013-11-02	North	NA	8.10	3000
18	11	7	2013	2013-11-07	North	NA	8.10	270
19	11	13	2013	2013-11-13	North	NA	NA	NA
20	11	16	2013	2013-11-16	North	NA	8.00	900

Showing 1 to 21 of 264 entries, 8 total columns

ANALYSIS APPROACH

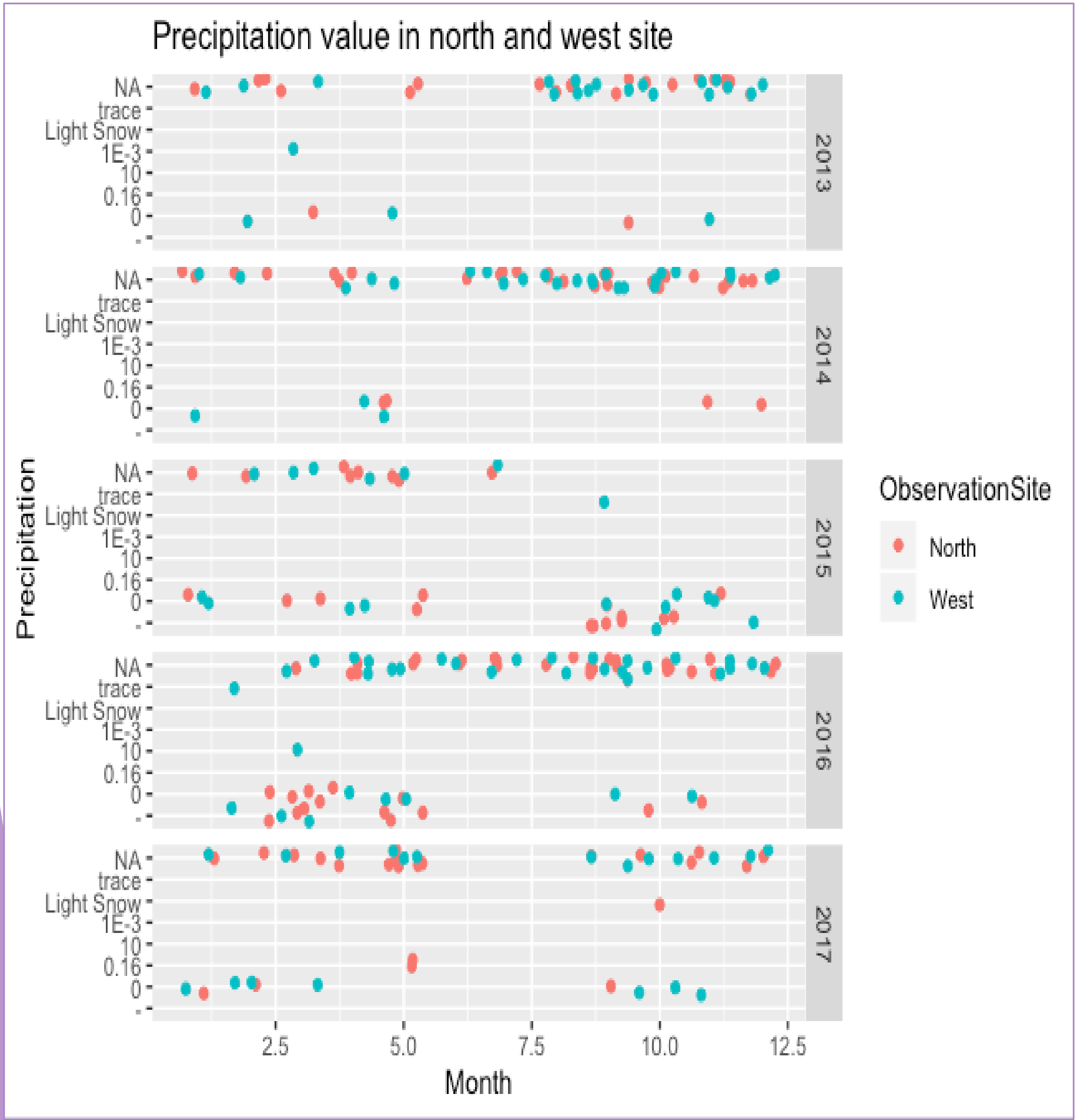
- **Ecoli in north and west site from 2013 to 2017**

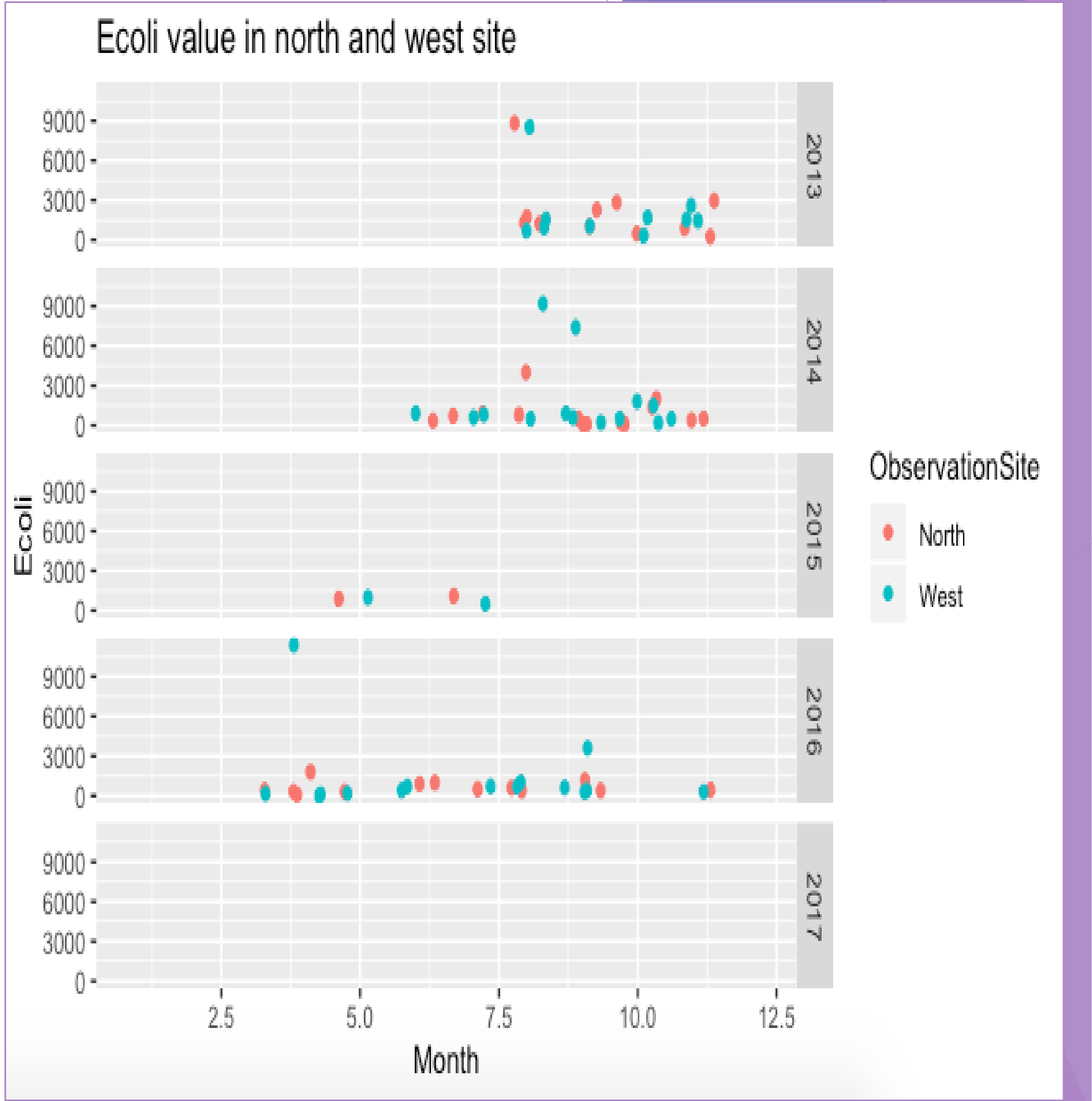
```
> ggplot (north_west_data, mapping = aes(x = Month, y = Ecoli, color = ObservationSite)) + geom_jitter()+
facet_grid(facets= vars(Year)) + ggtitle ("Ecoli value in north and west site")
```

- **Precipitation in north and west site from 2013 to 2017**

```
> ggplot (north_west_data, mapping = aes(x = Month, y = Precipitation, color = ObservationSite))
+geom_jitter()+ facet_grid(facets= vars(Year)) + ggtitle ("Precipitation value in north and west site")

>
```

ANALYSIS INTERPRETATION





RECOMMENDATION

- Recommendation Regarding Data Set
  - Remove calculation columns from the original dataset
  - Remove the use of color coded observations
  - Consistent in data standards
    - Column names
    - Observation site names
    - NA instead of empty or NV
    - Keep data type consistent
    - Decimal values
  - Document more precipitation data

- Analysis recommendation
  - Nutrients such as Alkalinity when found to be higher affects the quality of water and changes the pH level of water, however, from the analysis we prove that pH level of water can not be predicted by the level of Alkalinity in water.
  - Regarding Variable E.coli
  - → Inspect the nearby land.
  - → Educate people to learn more about the water quality and how should we take care of that.
  - → Precipitation data is also important to control the E.coli value. Precipitation data should be taken on a regular basis.

REFERENCES

1. Homestead National Monument of America Retrieved from https://en.wikipedia.org/wiki/Homestead_National_Monument_of_America

2. Aquatic invertebrate community trends and water quality at Homestead National Monument of America, Nebraska, 1996-2012 Retrieved from https://bioone.org/journals/transactions-of-the-kansas-academy-of-science/volume-116/issue-3-4/062.116.0301/Aquatic-Invertebrate-Community-Trends-and-Water-Quality-at-Homestead-National/10.1660/062.116.0301.short

3. National Park Service Retrieved From https://www.nps.gov/home/learn/historyculture/abouthomesteadactlaw.htm

4. How much water is there on Earth Retrieved From https://www.usgs.gov/special-topic/water-science-school/science/how-much-water-there-earth?qt-science_center_objects=0#qt-science_center_objects

5. Summary of the Clean Water Act Retrieved From https://www.epa.gov/laws-regulations/summary-clean-water-act

6. Water Quality Monitoring Retrieved From http://www.longwood.edu/cleanva/images/Sec5.WQMchapter.pdf

7. Chapter 5 Water Quality Conditions Retrieved From https://archive.epa.gov/water/archive/web/html/vms50.html

8. Cub creek water quality project Retrieved from https://www.nps.gov/home/learn/nature/cubcreekwaterquality.htm

9. pH and water. USGS science for a changing world Retrieved from https://www.usgs.gov/special-topic/water-science-school/science/ph-and-water?qt-science_center_objects=0#qt-scienc e_center_objects

10. Rock,C., & Rivera, B.(2014, march). Water Quality, E. coli and Your Health. The University of Arizona,College of Agriculture and Life Sciences. Retrieved from https://extension.arizona.edu/sites/extension.arizona.edu/files/pubs/az1624.pdf

ANY QUESTIONS?

THANK YOU

UNIVERSITY OF Nebraska Omaha