Jacob Dineen

**#HW 6**

**Unexecuted:**

#HW 6

##Jacob Dineen

#DUE 8/27 @ MIDNIGHT

#CLEAR ENVIRONMENT AND INSTALL INITIAL PACKAGES

rm(list = ls(all = TRUE))#Clear Enviroment

#Create function for installing/downloading packages

EnsurePackage<-function(x) {

x <- as.character(x)

if (!require(x,character.only=TRUE)) {

install.packages(pkgs=x, repos="http://cran.rproject.

org")

require(x,character.only=TRUE)

}

}

#Install Package needed for Data Visualization

EnsurePackage("ggplot2")

#STEP 1 LOAD THE DATA

airquality # Call Airquality dataset

airqual<- data.frame(airquality)# Save Airquality dataset as df

#STEP 2 DATA CLEANSING

str(airqual)

any(is.na(airqual)) #test if any NAs

summary(airqual) #37 NA's on OZONE. 7 NA's on SOLAR.R. NO OTHER NA's

mean(airqual$Ozone) #Run descriptive to see if Na's are impacting

mean(airqual$Ozone, na.rm = TRUE) #Could use na.rm as workaround, but will clean

str(na.omit(airqual)) #Using na.omit removed 42 rows of data from the df. will need to replace NA values instead.

airqual$Ozone[is.na(airqual$Ozone)] <- mean(airqual$Ozone,na.rm = TRUE) #Replace NA's with the mean of the column

airqual$Solar.R[is.na(airqual$Solar.R)] <- mean(airqual$Solar.R,na.rm = TRUE) #Replace NA's with the mean of the column

airqual #Call DF

any(is.na(airqual)) #Test if any NA's remain in data frame

#STEP 3 Understand the data distribution

###Histograms for each Variable.

#ozone

ggplot(airqual,aes(x=Ozone))+geom\_histogram(binwidth=5,color='white',fill="black")

#solar

ggplot(airqual,aes(x=Solar.R))+geom\_histogram(binwidth=5,color='white',fill="black")

#wind

ggplot(airqual,aes(x=Wind))+geom\_histogram(binwidth=5,color='white',fill="black")

#temp

ggplot(airqual,aes(x=Temp))+geom\_histogram(binwidth=5,color='white',fill="black")

#Month

ggplot(airqual,aes(x=Month))+geom\_histogram(bins=5,color='white',fill="black")

#Day

ggplot(airqual,aes(x=Day))+geom\_histogram(bins=31,color='white',fill="black")

###Boxplots

#Melt Data

EnsurePackage("reshape2")

airqual1 <- melt(airqual)

#Ozone Boxplot

Ozone <- airqual1[airqual1$variable== 'Ozone', ] #Save DF for filtered melt df

ggplot(Ozone, aes(variable, value )) + geom\_boxplot() #create boxplot on filtered melt df

#Wind Boxplot

Wind <- airqual1[airqual1$variable== 'Wind', ] #Save DF for filtered melt df

Wind$windround <- round(Wind$value) #New Col for Rounded Wind

Wind <- Wind [,-2] #Remove un-rounded value column

Wind

g1<- ggplot(Wind, aes(as.factor(variable), windround)) + geom\_boxplot() #create boxplot for rounded wind values

#Step 3 EXPLORE HOW DATA CHANGES OVER TIME

airqual$Date <- as.Date(paste(as.numeric("1973"), airqual$Month, airqual$Day, sep="/")) #Add concatenated Date column with 1973 as year

###Line Charts

#Ozone

ggplot(airqual,aes(x=Date,y=Ozone, group=1)) +

geom\_line(color="black")

#Temp

ggplot(airqual,aes(x=Date,y=Temp, group=1)) +

geom\_line(color="black")

#Solar.R

ggplot(airqual,aes(x=Date,y=Solar.R, group=1)) +

geom\_line(color="black")

#Wind

ggplot(airqual,aes(x=Date,y=Wind, group=1)) +

geom\_line(color="black")

# create one chart with 4 lines, each having a different color

# Rescale Wind values so that they won't be too close to the x axis, and store in a new column

airqual$scalewind <- (airqual$Wind\*10) #multiplied original values by 10 + Added new column

airclean <- airqual[ , c(1,2,4,7,8)] # create a new dataframe containing the four y variables and the x variable (Date).

airclean <- melt(airclean, id= ("Date")) # reframe the dataframe to stack all the y variables into a single column before visualization (sample on next slide)

# create one chart with four lines, using "Date" as x variable, and the four factors as y variables

ggplot(airclean , aes(x=Date,y=value, group=variable, color=variable)) + geom\_line() + geom\_line()

##STEP 4 LOOK AT ALL THE DATA VIA A HEATMAP

ggplot(airclean, aes(x=Date, y=variable, color=value)) + geom\_tile() + scale\_fill\_gradient(airclean, low="white", high="purple")

#Step 5 Look at all the data via a scatter chart

# Use data frame from step 2 to create a scatter chart, with "Wind" along x-axis and "Temp" along y-axis.

ggplot(airqual, aes(x=Wind, y=Temp)) + geom\_point(aes(size=Ozone,color=Solar.R)) # Additionally, set the points' size according to "Ozone" value and set the shade of color according to "Solar.R" value.

# Step 6: Final Analysis

# Do you see any patterns after exploring the data?

##I think just by looking at the line chart showing all four variables, we can see that there is a correlation between increases in the values as it relates to the 'Date'.

#When we look at the scatter plot, we can see that when temp is the highest, the wind is the lowest, suggesting an inverse,but correlated relationship.

#The individual line charts don't really help us determine any kind of relationship between the variables, although they work to show us more of an individual time series analysis for each.

#One thing that might be flawed about the dataset, is we're only seeing a sample of 'summer' months, really.

# What was the most useful visualization?

#I think the 4 piece line chart was the most useful in seeing correlation. The individual charts don't help to paint an aggregate picture.

#Looking at the scatter plot with all four variables plotted, we are able to see that solar and temp are positively correlated against an inverse relationship wind. When the temp is high, the solar value is generally higher, and the wind value is low, which, logically, makes sense. We can also see that when the wind is below the 10 threshold, we see a majority of our high output ozone values. Really when Wind is low, the other three variables are higher, and vice versa.

**EXECUTED**

> #HW 6

> ##Jacob Dineen

> #DUE 8/27 @ MIDNIGHT

>

> #CLEAR ENVIRONMENT AND INSTALL INITIAL PACKAGES

> rm(list = ls(all = TRUE))#Clear Enviroment

>

> #Create function for installing/downloading packages

> EnsurePackage<-function(x) {

+ x <- as.character(x)

+ if (!require(x,character.only=TRUE)) {

+ install.packages(pkgs=x, repos="http://cran.rproject.

+ org")

+ require(x,character.only=TRUE)

+ }

+ }

>

> #Install Package needed for Data Visualization

> EnsurePackage("ggplot2")

>

> #STEP 1 LOAD THE DATA

>

> airquality # Call Airquality dataset

Ozone Solar.R Wind Temp Month Day

1 41 190 7.4 67 5 1

2 36 118 8.0 72 5 2

3 12 149 12.6 74 5 3

4 18 313 11.5 62 5 4

5 NA NA 14.3 56 5 5

6 28 NA 14.9 66 5 6

7 23 299 8.6 65 5 7

8 19 99 13.8 59 5 8

9 8 19 20.1 61 5 9

10 NA 194 8.6 69 5 10

11 7 NA 6.9 74 5 11

12 16 256 9.7 69 5 12

13 11 290 9.2 66 5 13

14 14 274 10.9 68 5 14

15 18 65 13.2 58 5 15

16 14 334 11.5 64 5 16

17 34 307 12.0 66 5 17

18 6 78 18.4 57 5 18

19 30 322 11.5 68 5 19

20 11 44 9.7 62 5 20

21 1 8 9.7 59 5 21

22 11 320 16.6 73 5 22

23 4 25 9.7 61 5 23

24 32 92 12.0 61 5 24

25 NA 66 16.6 57 5 25

26 NA 266 14.9 58 5 26

27 NA NA 8.0 57 5 27

28 23 13 12.0 67 5 28

29 45 252 14.9 81 5 29

30 115 223 5.7 79 5 30

31 37 279 7.4 76 5 31

32 NA 286 8.6 78 6 1

33 NA 287 9.7 74 6 2

34 NA 242 16.1 67 6 3

35 NA 186 9.2 84 6 4

36 NA 220 8.6 85 6 5

37 NA 264 14.3 79 6 6

38 29 127 9.7 82 6 7

39 NA 273 6.9 87 6 8

40 71 291 13.8 90 6 9

41 39 323 11.5 87 6 10

42 NA 259 10.9 93 6 11

43 NA 250 9.2 92 6 12

44 23 148 8.0 82 6 13

45 NA 332 13.8 80 6 14

46 NA 322 11.5 79 6 15

47 21 191 14.9 77 6 16

48 37 284 20.7 72 6 17

49 20 37 9.2 65 6 18

50 12 120 11.5 73 6 19

51 13 137 10.3 76 6 20

52 NA 150 6.3 77 6 21

53 NA 59 1.7 76 6 22

54 NA 91 4.6 76 6 23

55 NA 250 6.3 76 6 24

56 NA 135 8.0 75 6 25

57 NA 127 8.0 78 6 26

58 NA 47 10.3 73 6 27

59 NA 98 11.5 80 6 28

60 NA 31 14.9 77 6 29

61 NA 138 8.0 83 6 30

62 135 269 4.1 84 7 1

63 49 248 9.2 85 7 2

64 32 236 9.2 81 7 3

65 NA 101 10.9 84 7 4

66 64 175 4.6 83 7 5

67 40 314 10.9 83 7 6

68 77 276 5.1 88 7 7

69 97 267 6.3 92 7 8

70 97 272 5.7 92 7 9

71 85 175 7.4 89 7 10

72 NA 139 8.6 82 7 11

73 10 264 14.3 73 7 12

74 27 175 14.9 81 7 13

75 NA 291 14.9 91 7 14

76 7 48 14.3 80 7 15

77 48 260 6.9 81 7 16

78 35 274 10.3 82 7 17

79 61 285 6.3 84 7 18

80 79 187 5.1 87 7 19

81 63 220 11.5 85 7 20

82 16 7 6.9 74 7 21

83 NA 258 9.7 81 7 22

84 NA 295 11.5 82 7 23

85 80 294 8.6 86 7 24

86 108 223 8.0 85 7 25

87 20 81 8.6 82 7 26

88 52 82 12.0 86 7 27

89 82 213 7.4 88 7 28

90 50 275 7.4 86 7 29

91 64 253 7.4 83 7 30

92 59 254 9.2 81 7 31

93 39 83 6.9 81 8 1

94 9 24 13.8 81 8 2

95 16 77 7.4 82 8 3

96 78 NA 6.9 86 8 4

97 35 NA 7.4 85 8 5

98 66 NA 4.6 87 8 6

99 122 255 4.0 89 8 7

100 89 229 10.3 90 8 8

101 110 207 8.0 90 8 9

102 NA 222 8.6 92 8 10

103 NA 137 11.5 86 8 11

104 44 192 11.5 86 8 12

105 28 273 11.5 82 8 13

106 65 157 9.7 80 8 14

107 NA 64 11.5 79 8 15

108 22 71 10.3 77 8 16

109 59 51 6.3 79 8 17

110 23 115 7.4 76 8 18

111 31 244 10.9 78 8 19

112 44 190 10.3 78 8 20

113 21 259 15.5 77 8 21

114 9 36 14.3 72 8 22

115 NA 255 12.6 75 8 23

116 45 212 9.7 79 8 24

117 168 238 3.4 81 8 25

118 73 215 8.0 86 8 26

119 NA 153 5.7 88 8 27

120 76 203 9.7 97 8 28

121 118 225 2.3 94 8 29

122 84 237 6.3 96 8 30

123 85 188 6.3 94 8 31

124 96 167 6.9 91 9 1

125 78 197 5.1 92 9 2

126 73 183 2.8 93 9 3

127 91 189 4.6 93 9 4

128 47 95 7.4 87 9 5

129 32 92 15.5 84 9 6

130 20 252 10.9 80 9 7

131 23 220 10.3 78 9 8

132 21 230 10.9 75 9 9

133 24 259 9.7 73 9 10

134 44 236 14.9 81 9 11

135 21 259 15.5 76 9 12

136 28 238 6.3 77 9 13

137 9 24 10.9 71 9 14

138 13 112 11.5 71 9 15

139 46 237 6.9 78 9 16

140 18 224 13.8 67 9 17

141 13 27 10.3 76 9 18

142 24 238 10.3 68 9 19

143 16 201 8.0 82 9 20

144 13 238 12.6 64 9 21

145 23 14 9.2 71 9 22

146 36 139 10.3 81 9 23

147 7 49 10.3 69 9 24

148 14 20 16.6 63 9 25

149 30 193 6.9 70 9 26

150 NA 145 13.2 77 9 27

151 14 191 14.3 75 9 28

152 18 131 8.0 76 9 29

153 20 223 11.5 68 9 30

> airqual<- data.frame(airquality)# Save Airquality dataset as df

>

> #STEP 2 DATA CLEANSING

>

> str(airqual)

'data.frame': 153 obs. of 6 variables:

$ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...

$ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...

$ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...

$ Temp : int 67 72 74 62 56 66 65 59 61 69 ...

$ Month : int 5 5 5 5 5 5 5 5 5 5 ...

$ Day : int 1 2 3 4 5 6 7 8 9 10 ...

> any(is.na(airqual)) #test if any NAs

[1] TRUE

> summary(airqual) #37 NA's on OZONE. 7 NA's on SOLAR.R. NO OTHER NA's

Ozone Solar.R Wind Temp Month Day

Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00 Min. :5.000 Min. : 1.0

1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00 1st Qu.:6.000 1st Qu.: 8.0

Median : 31.50 Median :205.0 Median : 9.700 Median :79.00 Median :7.000 Median :16.0

Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88 Mean :6.993 Mean :15.8

3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00 3rd Qu.:8.000 3rd Qu.:23.0

Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00 Max. :9.000 Max. :31.0

NA's :37 NA's :7

> mean(airqual$Ozone) #Run descriptive to see if Na's are impacting

[1] NA

> mean(airqual$Ozone, na.rm = TRUE) #Could use na.rm as workaround, but will clean

[1] 42.12931

> str(na.omit(airqual)) #Using na.omit removed 42 rows of data from the df. will need to replace NA values instead.

'data.frame': 111 obs. of 6 variables:

$ Ozone : int 41 36 12 18 23 19 8 16 11 14 ...

$ Solar.R: int 190 118 149 313 299 99 19 256 290 274 ...

$ Wind : num 7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...

$ Temp : int 67 72 74 62 65 59 61 69 66 68 ...

$ Month : int 5 5 5 5 5 5 5 5 5 5 ...

$ Day : int 1 2 3 4 7 8 9 12 13 14 ...

- attr(\*, "na.action")=Class 'omit' Named int [1:42] 5 6 10 11 25 26 27 32 33 34 ...

.. ..- attr(\*, "names")= chr [1:42] "5" "6" "10" "11" ...

>

> airqual$Ozone[is.na(airqual$Ozone)] <- mean(airqual$Ozone,na.rm = TRUE) #Replace NA's with the mean of the column

> airqual$Solar.R[is.na(airqual$Solar.R)] <- mean(airqual$Solar.R,na.rm = TRUE) #Replace NA's with the mean of the column

> airqual #Call DF

Ozone Solar.R Wind Temp Month Day

1 41.00000 190.0000 7.4 67 5 1

2 36.00000 118.0000 8.0 72 5 2

3 12.00000 149.0000 12.6 74 5 3

4 18.00000 313.0000 11.5 62 5 4

5 42.12931 185.9315 14.3 56 5 5

6 28.00000 185.9315 14.9 66 5 6

7 23.00000 299.0000 8.6 65 5 7

8 19.00000 99.0000 13.8 59 5 8

9 8.00000 19.0000 20.1 61 5 9

10 42.12931 194.0000 8.6 69 5 10

11 7.00000 185.9315 6.9 74 5 11

12 16.00000 256.0000 9.7 69 5 12

13 11.00000 290.0000 9.2 66 5 13

14 14.00000 274.0000 10.9 68 5 14

15 18.00000 65.0000 13.2 58 5 15

16 14.00000 334.0000 11.5 64 5 16

17 34.00000 307.0000 12.0 66 5 17

18 6.00000 78.0000 18.4 57 5 18

19 30.00000 322.0000 11.5 68 5 19

20 11.00000 44.0000 9.7 62 5 20

21 1.00000 8.0000 9.7 59 5 21

22 11.00000 320.0000 16.6 73 5 22

23 4.00000 25.0000 9.7 61 5 23

24 32.00000 92.0000 12.0 61 5 24

25 42.12931 66.0000 16.6 57 5 25

26 42.12931 266.0000 14.9 58 5 26

27 42.12931 185.9315 8.0 57 5 27

28 23.00000 13.0000 12.0 67 5 28

29 45.00000 252.0000 14.9 81 5 29

30 115.00000 223.0000 5.7 79 5 30

31 37.00000 279.0000 7.4 76 5 31

32 42.12931 286.0000 8.6 78 6 1

33 42.12931 287.0000 9.7 74 6 2

34 42.12931 242.0000 16.1 67 6 3

35 42.12931 186.0000 9.2 84 6 4

36 42.12931 220.0000 8.6 85 6 5

37 42.12931 264.0000 14.3 79 6 6

38 29.00000 127.0000 9.7 82 6 7

39 42.12931 273.0000 6.9 87 6 8

40 71.00000 291.0000 13.8 90 6 9

41 39.00000 323.0000 11.5 87 6 10

42 42.12931 259.0000 10.9 93 6 11

43 42.12931 250.0000 9.2 92 6 12

44 23.00000 148.0000 8.0 82 6 13

45 42.12931 332.0000 13.8 80 6 14

46 42.12931 322.0000 11.5 79 6 15

47 21.00000 191.0000 14.9 77 6 16

48 37.00000 284.0000 20.7 72 6 17

49 20.00000 37.0000 9.2 65 6 18

50 12.00000 120.0000 11.5 73 6 19

51 13.00000 137.0000 10.3 76 6 20

52 42.12931 150.0000 6.3 77 6 21

53 42.12931 59.0000 1.7 76 6 22

54 42.12931 91.0000 4.6 76 6 23

55 42.12931 250.0000 6.3 76 6 24

56 42.12931 135.0000 8.0 75 6 25

57 42.12931 127.0000 8.0 78 6 26

58 42.12931 47.0000 10.3 73 6 27

59 42.12931 98.0000 11.5 80 6 28

60 42.12931 31.0000 14.9 77 6 29

61 42.12931 138.0000 8.0 83 6 30

62 135.00000 269.0000 4.1 84 7 1

63 49.00000 248.0000 9.2 85 7 2

64 32.00000 236.0000 9.2 81 7 3

65 42.12931 101.0000 10.9 84 7 4

66 64.00000 175.0000 4.6 83 7 5

67 40.00000 314.0000 10.9 83 7 6

68 77.00000 276.0000 5.1 88 7 7

69 97.00000 267.0000 6.3 92 7 8

70 97.00000 272.0000 5.7 92 7 9

71 85.00000 175.0000 7.4 89 7 10

72 42.12931 139.0000 8.6 82 7 11

73 10.00000 264.0000 14.3 73 7 12

74 27.00000 175.0000 14.9 81 7 13

75 42.12931 291.0000 14.9 91 7 14

76 7.00000 48.0000 14.3 80 7 15

77 48.00000 260.0000 6.9 81 7 16

78 35.00000 274.0000 10.3 82 7 17

79 61.00000 285.0000 6.3 84 7 18

80 79.00000 187.0000 5.1 87 7 19

81 63.00000 220.0000 11.5 85 7 20

82 16.00000 7.0000 6.9 74 7 21

83 42.12931 258.0000 9.7 81 7 22

84 42.12931 295.0000 11.5 82 7 23

85 80.00000 294.0000 8.6 86 7 24

86 108.00000 223.0000 8.0 85 7 25

87 20.00000 81.0000 8.6 82 7 26

88 52.00000 82.0000 12.0 86 7 27

89 82.00000 213.0000 7.4 88 7 28

90 50.00000 275.0000 7.4 86 7 29

91 64.00000 253.0000 7.4 83 7 30

92 59.00000 254.0000 9.2 81 7 31

93 39.00000 83.0000 6.9 81 8 1

94 9.00000 24.0000 13.8 81 8 2

95 16.00000 77.0000 7.4 82 8 3

96 78.00000 185.9315 6.9 86 8 4

97 35.00000 185.9315 7.4 85 8 5

98 66.00000 185.9315 4.6 87 8 6

99 122.00000 255.0000 4.0 89 8 7

100 89.00000 229.0000 10.3 90 8 8

101 110.00000 207.0000 8.0 90 8 9

102 42.12931 222.0000 8.6 92 8 10

103 42.12931 137.0000 11.5 86 8 11

104 44.00000 192.0000 11.5 86 8 12

105 28.00000 273.0000 11.5 82 8 13

106 65.00000 157.0000 9.7 80 8 14

107 42.12931 64.0000 11.5 79 8 15

108 22.00000 71.0000 10.3 77 8 16

109 59.00000 51.0000 6.3 79 8 17

110 23.00000 115.0000 7.4 76 8 18

111 31.00000 244.0000 10.9 78 8 19

112 44.00000 190.0000 10.3 78 8 20

113 21.00000 259.0000 15.5 77 8 21

114 9.00000 36.0000 14.3 72 8 22

115 42.12931 255.0000 12.6 75 8 23

116 45.00000 212.0000 9.7 79 8 24

117 168.00000 238.0000 3.4 81 8 25

118 73.00000 215.0000 8.0 86 8 26

119 42.12931 153.0000 5.7 88 8 27

120 76.00000 203.0000 9.7 97 8 28

121 118.00000 225.0000 2.3 94 8 29

122 84.00000 237.0000 6.3 96 8 30

123 85.00000 188.0000 6.3 94 8 31

124 96.00000 167.0000 6.9 91 9 1

125 78.00000 197.0000 5.1 92 9 2

126 73.00000 183.0000 2.8 93 9 3

127 91.00000 189.0000 4.6 93 9 4

128 47.00000 95.0000 7.4 87 9 5

129 32.00000 92.0000 15.5 84 9 6

130 20.00000 252.0000 10.9 80 9 7

131 23.00000 220.0000 10.3 78 9 8

132 21.00000 230.0000 10.9 75 9 9

133 24.00000 259.0000 9.7 73 9 10

134 44.00000 236.0000 14.9 81 9 11

135 21.00000 259.0000 15.5 76 9 12

136 28.00000 238.0000 6.3 77 9 13

137 9.00000 24.0000 10.9 71 9 14

138 13.00000 112.0000 11.5 71 9 15

139 46.00000 237.0000 6.9 78 9 16

140 18.00000 224.0000 13.8 67 9 17

141 13.00000 27.0000 10.3 76 9 18

142 24.00000 238.0000 10.3 68 9 19

143 16.00000 201.0000 8.0 82 9 20

144 13.00000 238.0000 12.6 64 9 21

145 23.00000 14.0000 9.2 71 9 22

146 36.00000 139.0000 10.3 81 9 23

147 7.00000 49.0000 10.3 69 9 24

148 14.00000 20.0000 16.6 63 9 25

149 30.00000 193.0000 6.9 70 9 26

150 42.12931 145.0000 13.2 77 9 27

151 14.00000 191.0000 14.3 75 9 28

152 18.00000 131.0000 8.0 76 9 29

153 20.00000 223.0000 11.5 68 9 30

> any(is.na(airqual)) #Test if any NA's remain in data frame

[1] FALSE

>

> #STEP 3 Understand the data distribution

> ###Histograms for each Variable.

>

> #ozone

> ggplot(airqual,aes(x=Ozone))+geom\_histogram(binwidth=5,color='white',fill="black")

>

> #solar

> ggplot(airqual,aes(x=Solar.R))+geom\_histogram(binwidth=5,color='white',fill="black")

>

> #wind

> ggplot(airqual,aes(x=Wind))+geom\_histogram(binwidth=5,color='white',fill="black")

>

> #temp

> ggplot(airqual,aes(x=Temp))+geom\_histogram(binwidth=5,color='white',fill="black")

>

> #Month

> ggplot(airqual,aes(x=Month))+geom\_histogram(bins=5,color='white',fill="black")

>

> #Day

> ggplot(airqual,aes(x=Day))+geom\_histogram(bins=31,color='white',fill="black")

>

>

> ###Boxplots

>

> #Melt Data

> EnsurePackage("reshape2")

> airqual1 <- melt(airqual)

No id variables; using all as measure variables

>

> #Ozone Boxplot

> Ozone <- airqual1[airqual1$variable== 'Ozone', ] #Save DF for filtered melt df

> ggplot(Ozone, aes(variable, value )) + geom\_boxplot() #create boxplot on filtered melt df

>

> #Wind Boxplot

>

> Wind <- airqual1[airqual1$variable== 'Wind', ] #Save DF for filtered melt df

> Wind$windround <- round(Wind$value) #New Col for Rounded Wind

> Wind <- Wind [,-2] #Remove un-rounded value column

> Wind

variable windround

307 Wind 7

308 Wind 8

309 Wind 13

310 Wind 12

311 Wind 14

312 Wind 15

313 Wind 9

314 Wind 14

315 Wind 20

316 Wind 9

317 Wind 7

318 Wind 10

319 Wind 9

320 Wind 11

321 Wind 13

322 Wind 12

323 Wind 12

324 Wind 18

325 Wind 12

326 Wind 10

327 Wind 10

328 Wind 17

329 Wind 10

330 Wind 12

331 Wind 17

332 Wind 15

333 Wind 8

334 Wind 12

335 Wind 15

336 Wind 6

337 Wind 7

338 Wind 9

339 Wind 10

340 Wind 16

341 Wind 9

342 Wind 9

343 Wind 14

344 Wind 10

345 Wind 7

346 Wind 14

347 Wind 12

348 Wind 11

349 Wind 9

350 Wind 8

351 Wind 14

352 Wind 12

353 Wind 15

354 Wind 21

355 Wind 9

356 Wind 12

357 Wind 10

358 Wind 6

359 Wind 2

360 Wind 5

361 Wind 6

362 Wind 8

363 Wind 8

364 Wind 10

365 Wind 12

366 Wind 15

367 Wind 8

368 Wind 4

369 Wind 9

370 Wind 9

371 Wind 11

372 Wind 5

373 Wind 11

374 Wind 5

375 Wind 6

376 Wind 6

377 Wind 7

378 Wind 9

379 Wind 14

380 Wind 15

381 Wind 15

382 Wind 14

383 Wind 7

384 Wind 10

385 Wind 6

386 Wind 5

387 Wind 12

388 Wind 7

389 Wind 10

390 Wind 12

391 Wind 9

392 Wind 8

393 Wind 9

394 Wind 12

395 Wind 7

396 Wind 7

397 Wind 7

398 Wind 9

399 Wind 7

400 Wind 14

401 Wind 7

402 Wind 7

403 Wind 7

404 Wind 5

405 Wind 4

406 Wind 10

407 Wind 8

408 Wind 9

409 Wind 12

410 Wind 12

411 Wind 12

412 Wind 10

413 Wind 12

414 Wind 10

415 Wind 6

416 Wind 7

417 Wind 11

418 Wind 10

419 Wind 16

420 Wind 14

421 Wind 13

422 Wind 10

423 Wind 3

424 Wind 8

425 Wind 6

426 Wind 10

427 Wind 2

428 Wind 6

429 Wind 6

430 Wind 7

431 Wind 5

432 Wind 3

433 Wind 5

434 Wind 7

435 Wind 16

436 Wind 11

437 Wind 10

438 Wind 11

439 Wind 10

440 Wind 15

441 Wind 16

442 Wind 6

443 Wind 11

444 Wind 12

445 Wind 7

446 Wind 14

447 Wind 10

448 Wind 10

449 Wind 8

450 Wind 13

451 Wind 9

452 Wind 10

453 Wind 10

454 Wind 17

455 Wind 7

456 Wind 13

457 Wind 14

458 Wind 8

459 Wind 12

> g1<- ggplot(Wind, aes(as.factor(variable), windround)) + geom\_boxplot() #create boxplot for rounded wind values

>

> #Step 3 EXPLORE HOW DATA CHANGES OVER TIME

> airqual$Date <- as.Date(paste(as.numeric("1973"), airqual$Month, airqual$Day, sep="/")) #Add concatenated Date column with 1973 as year

>

> ###Line Charts

>

> #Ozone

> ggplot(airqual,aes(x=Date,y=Ozone, group=1)) +

+ geom\_line(color="black")

>

> #Temp

> ggplot(airqual,aes(x=Date,y=Temp, group=1)) +

+ geom\_line(color="black")

>

> #Solar.R

> ggplot(airqual,aes(x=Date,y=Solar.R, group=1)) +

+ geom\_line(color="black")

>

> #Wind

> ggplot(airqual,aes(x=Date,y=Wind, group=1)) +

+ geom\_line(color="black")

>

> # create one chart with 4 lines, each having a different color

>

> # Rescale Wind values so that they won't be too close to the x axis, and store in a new column

> airqual$scalewind <- (airqual$Wind\*10) #multiplied original values by 10 + Added new column

> airclean <- airqual[ , c(1,2,4,7,8)] # create a new dataframe containing the four y variables and the x variable (Date).

> airclean <- melt(airclean, id= ("Date")) # reframe the dataframe to stack all the y variables into a single column before visualization (sample on next slide)

>

> # create one chart with four lines, using "Date" as x variable, and the four factors as y variables

>

> ggplot(airclean , aes(x=Date,y=value, group=variable, color=variable)) + geom\_line() + geom\_line()

>

>

> ##STEP 4 LOOK AT ALL THE DATA VIA A HEATMAP

> ggplot(airclean, aes(x=Date, y=variable, color=value)) + geom\_tile() + scale\_fill\_gradient(airclean, low="white", high="purple")

>

>

> #Step 5 Look at all the data via a scatter chart

> # Use data frame from step 2 to create a scatter chart, with "Wind" along x-axis and "Temp" along y-axis.

> ggplot(airqual, aes(x=Wind, y=Temp)) + geom\_point(aes(size=Ozone,color=Solar.R)) # Additionally, set the points' size according to "Ozone" value and set the shade of color according to "Solar.R" value.

>

>

> # Step 6: Final Analysis

> # Do you see any patterns after exploring the data?

> ##I think just by looking at the line chart showing all four variables, we can see that there is a correlation between increases in the values as it relates to the 'Date'.

> #When we look at the scatter plot, we can see that when temp is the highest, the wind is the lowest, suggesting an inverse,but correlated relationship.

> #The individual line charts don't really help us determine any kind of relationship between the variables, although they work to show us more of an individual time series analysis for each.

> #One thing that might be flawed about the dataset, is we're only seeing a sample of 'summer' months, really.

>

> # What was the most useful visualization?

> #I think the 4 piece line chart was the most useful in seeing correlation. The individual charts don't help to paint an aggregate picture.

> #Looking at the scatter plot with all four variables plotted, we are able to see that solar and temp are positively correlated against an inverse relationship wind. When the temp is high, the solar value is generally higher, and the wind value is low, which, logically, makes sense. We can also see that when the wind is below the 10 threshold, we see a majority of our high output ozone values. Really when Wind is low, the other three variables are higher, and vice versa.