

airquality

About airquality

This data set contains the daily air quality measurements in New York, May to September 1973.

```
air_quality <- clean_names(airquality)
```

```
head(air_quality)
```

```
#>   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
```

```
dim(air_quality)
```

```
#> [1] 153  6
```

```
colnames(air_quality)
```

```
#> [1] "ozone" "solar_r" "wind" "temp" "month" "day"
```

```
glimpse(air_quality)
```

```
#> Rows: 153
```

```
#> Columns: 6
```

```
#> $ ozone   <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, 18, 14, ~
#> $ solar_r <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256, 290, 27~
#> $ wind    <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8.6, 6.9, 9~
#> $ temp    <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 68, 58, 64~
#> $ month   <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ day     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
```

Exploratory Data Analysis

Univariate non-graphical

```
summary(air_quality)
```

```
#>   ozone          solar_r          wind          temp
#> Min.   : 1.00   Min.   : 7.0   Min.   : 1.700   Min.   :56.00
#> 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
#> Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
#> Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
#> 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
#> Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
#> NA's   :37      NA's   :7
#>   month          day
#> Min.   :5.000   Min.   : 1.0
#> 1st Qu.:6.000   1st Qu.: 8.0
```

```
#> Median :7.000 Median :16.0
#> Mean   :6.993 Mean   :15.8
#> 3rd Qu.:8.000 3rd Qu.:23.0
#> Max.   :9.000 Max.   :31.0
#>
```

```
skimr::skim(air_quality)
```

Table 1: Data summary

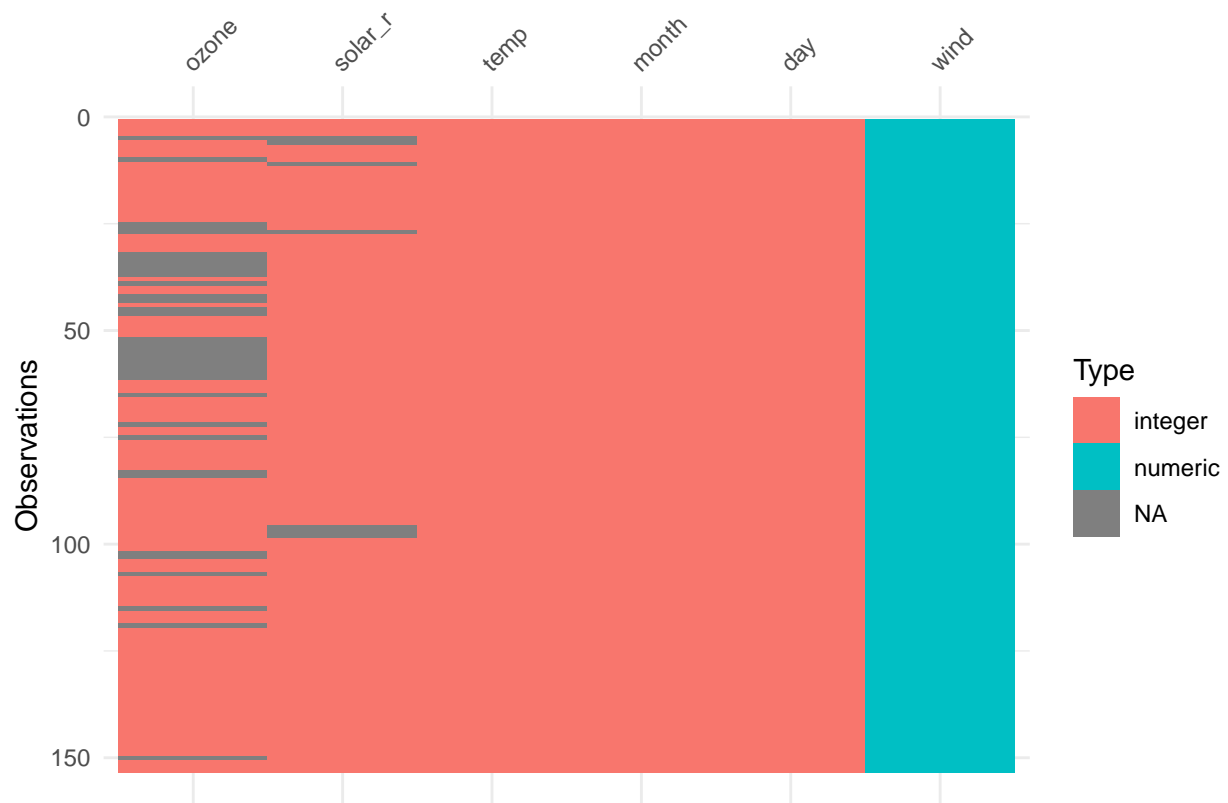
Name	air_quality
Number of rows	153
Number of columns	6
Column type frequency:	
numeric	6
Group variables	None

Variable type: numeric

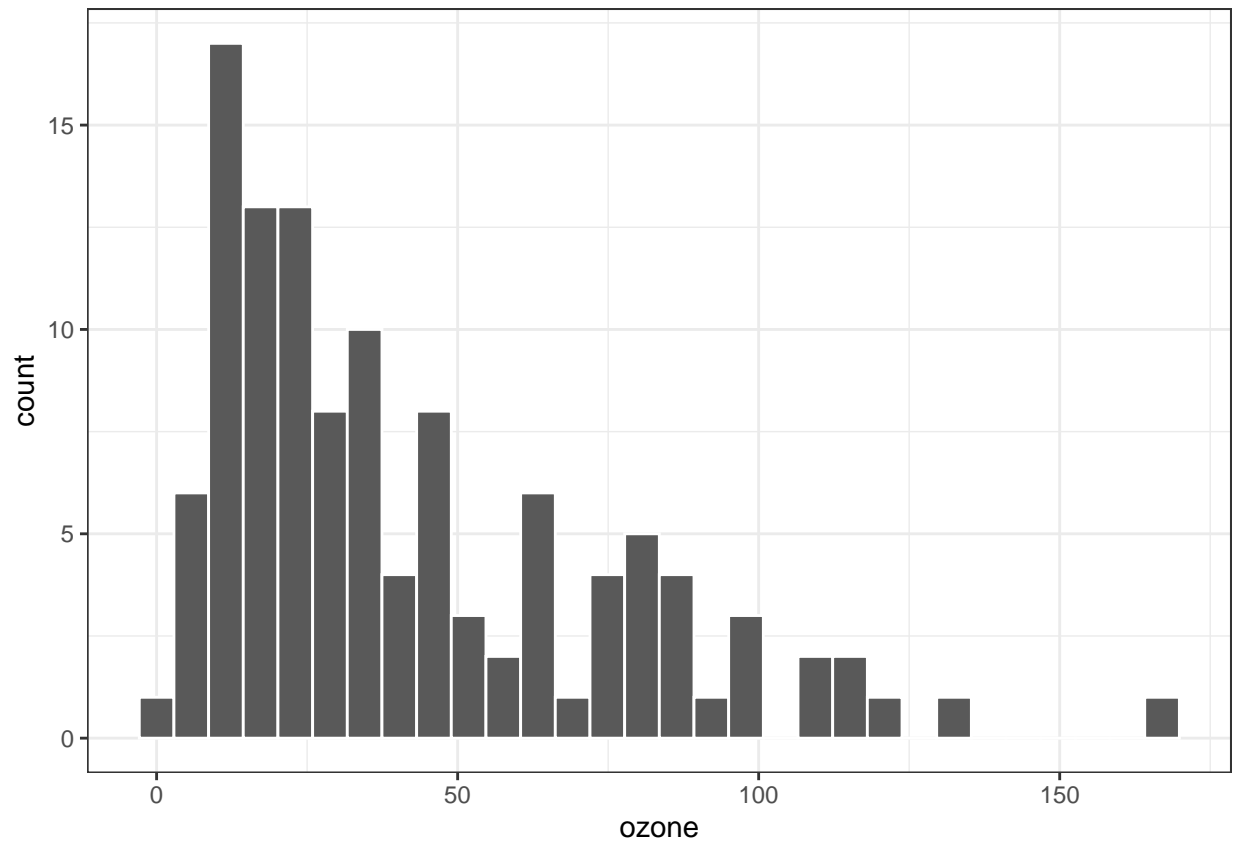
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ozone	37	0.76	42.13	32.99	1.0	18.00	31.5	63.25	168.0	
solar_r	7	0.95	185.93	90.06	7.0	115.75	205.0	258.75	334.0	
wind	0	1.00	9.96	3.52	1.7	7.40	9.7	11.50	20.7	
temp	0	1.00	77.88	9.47	56.0	72.00	79.0	85.00	97.0	
month	0	1.00	6.99	1.42	5.0	6.00	7.0	8.00	9.0	
day	0	1.00	15.80	8.86	1.0	8.00	16.0	23.00	31.0	

Univariate graphical

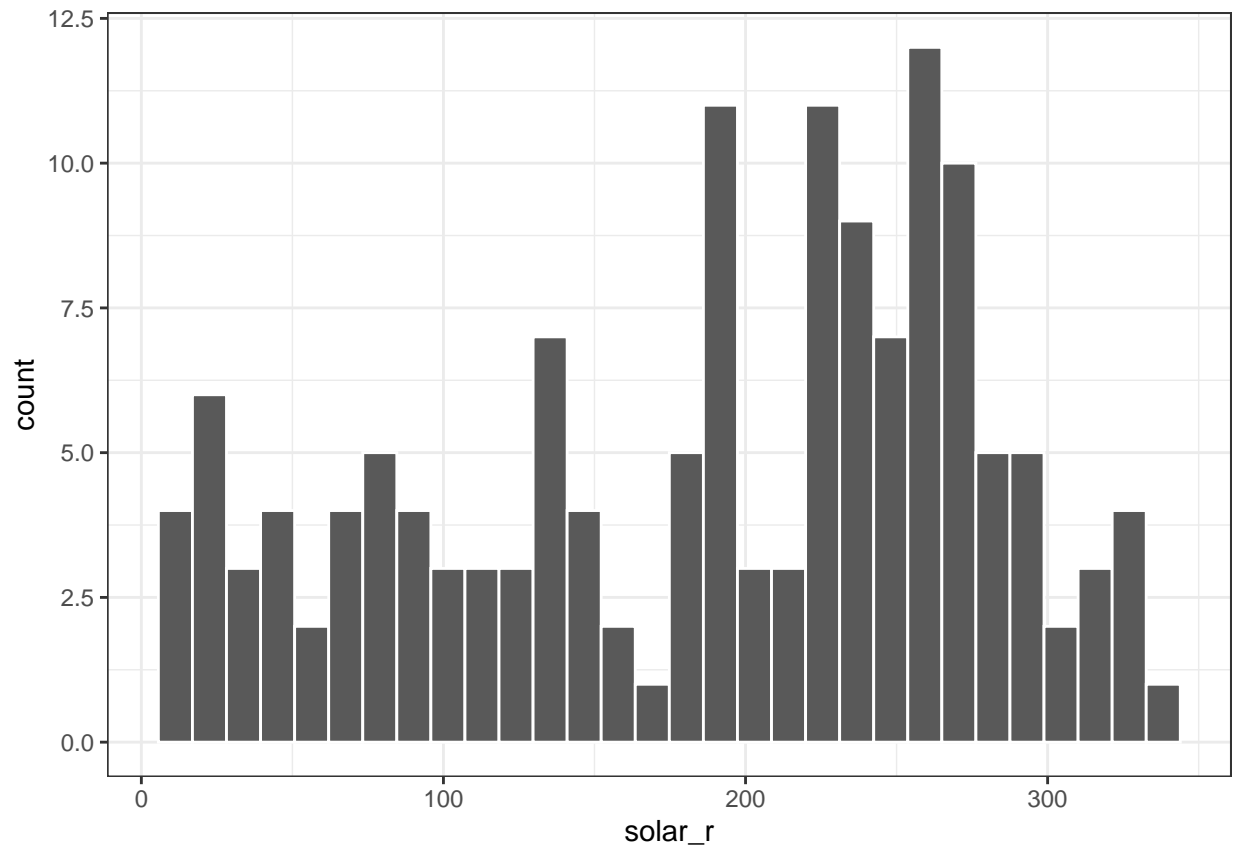
```
vis_dat(air_quality)
```



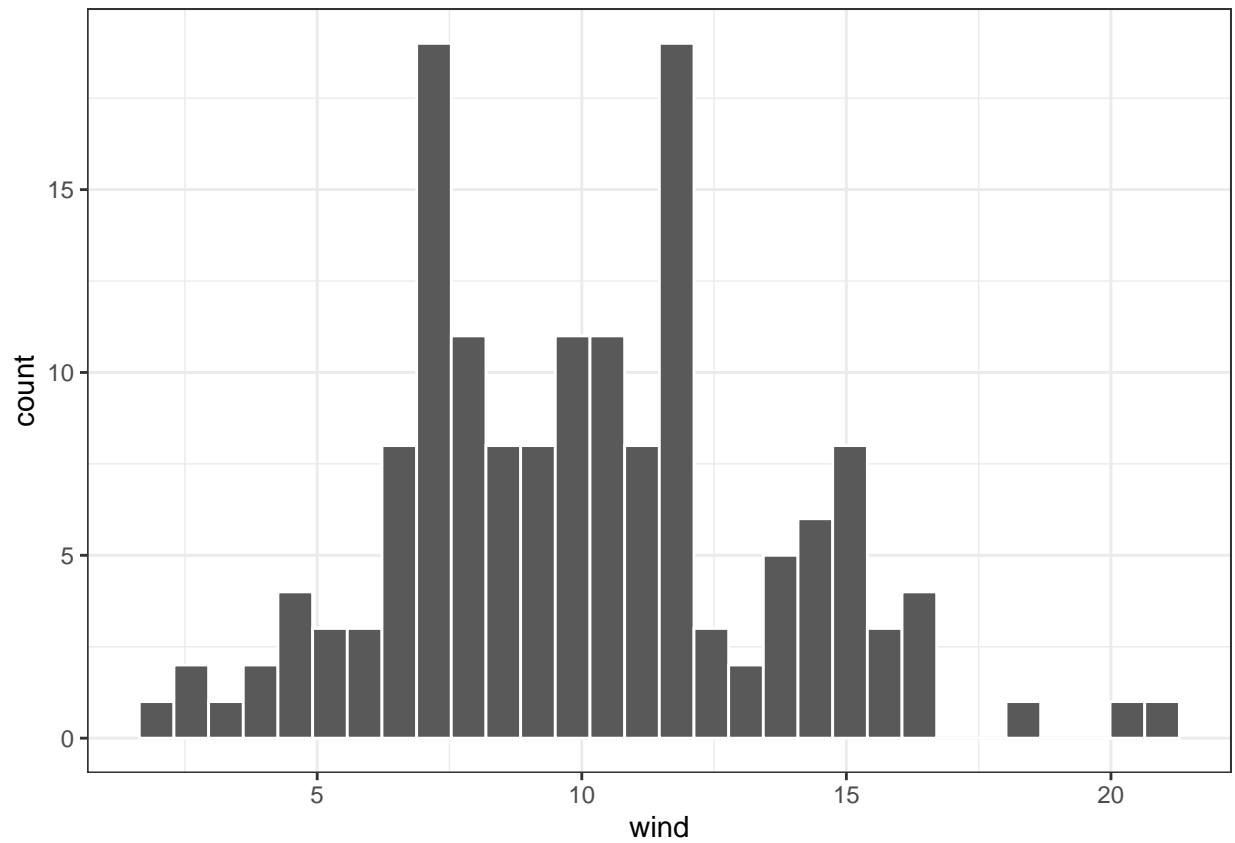
```
p1 <- ggplot(air_quality, aes(ozone)) +
  geom_histogram(color = "white")
p1
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 37 rows containing non-finite values (stat_bin).
```



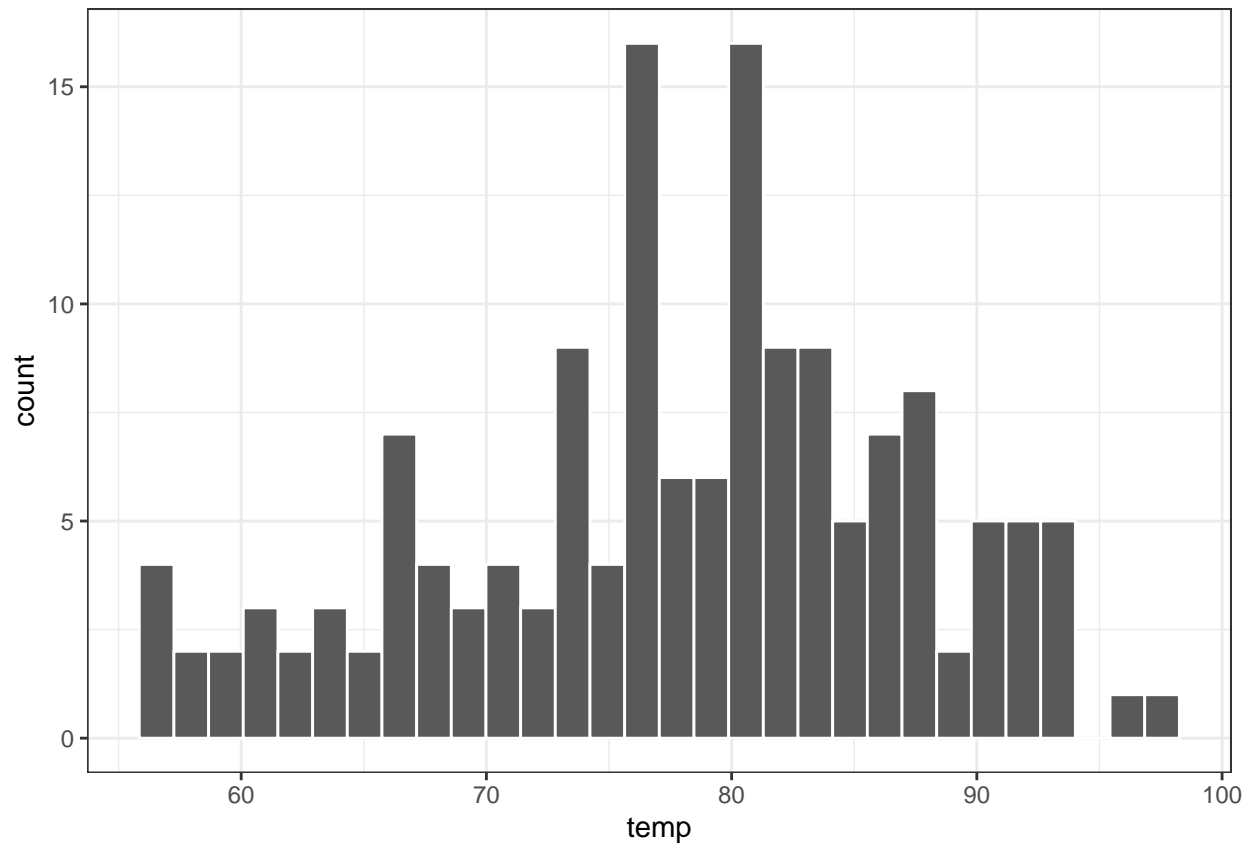
```
p2 <- ggplot(air_quality, aes(solar_r)) +  
  geom_histogram(color = "white")  
p2  
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.  
#> Warning: Removed 7 rows containing non-finite values (stat_bin).
```



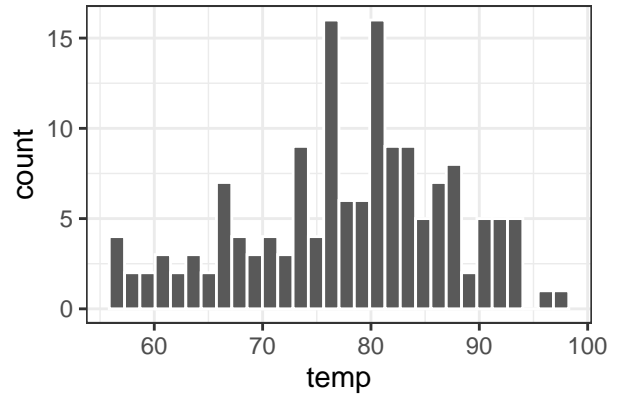
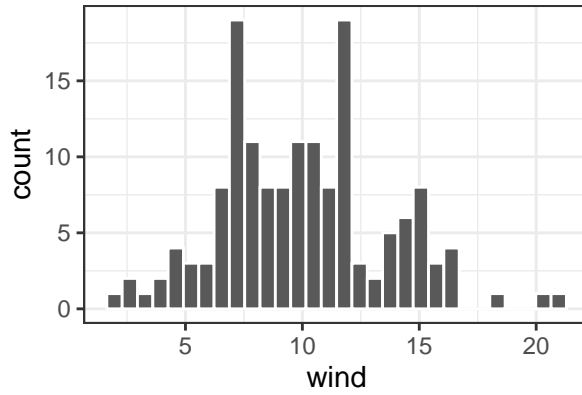
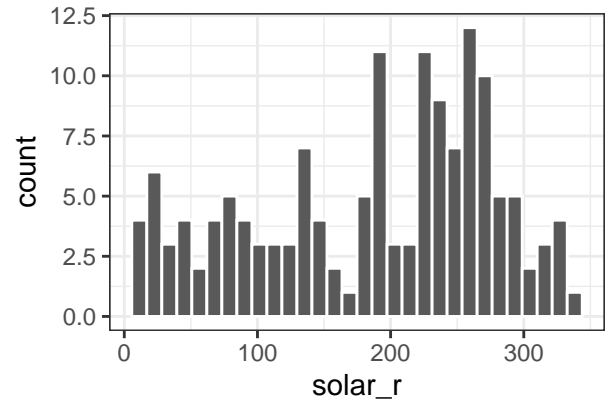
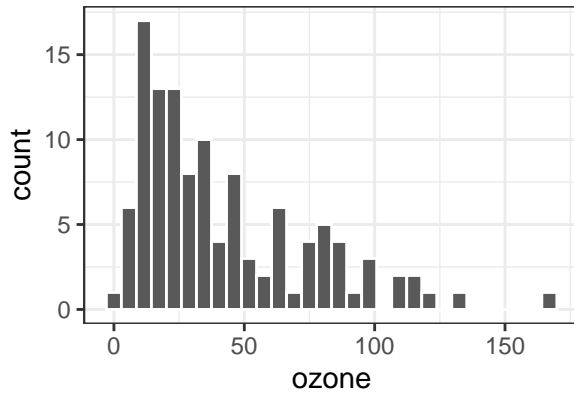
```
p3 <- ggplot(air_quality, aes(wind)) +  
  geom_histogram(color = "white")  
p3  
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



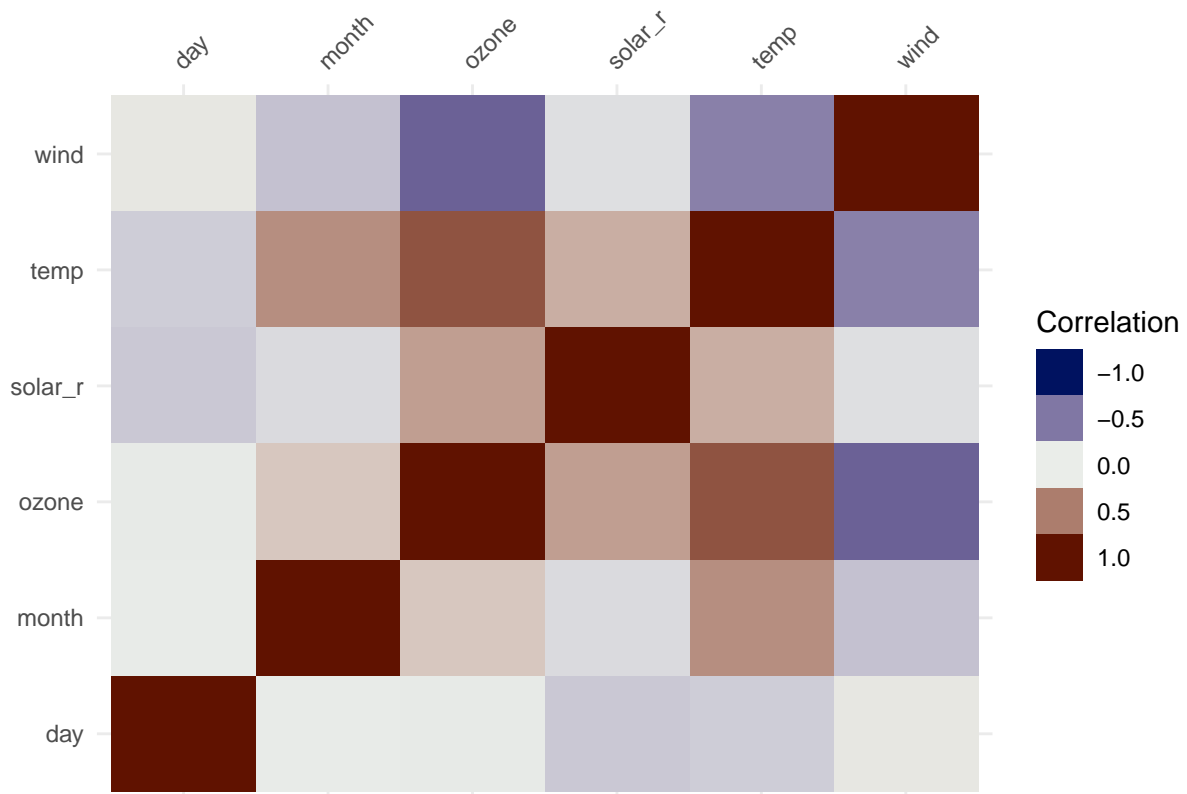
```
p4 <- ggplot(air_quality, aes(temp)) +  
  geom_histogram(color = "white")  
p4  
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
(p1 + p2) / (p3 + p4)
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 37 rows containing non-finite values (stat_bin).
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 7 rows containing non-finite values (stat_bin).
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
vis_cor(air_quality)
```

```
air_quality <- air_quality %>%
  mutate(
    date = lubridate::make_date(year = 1973, month = month, day = day)
  )
air_quality
#>   ozone solar_r wind temp month day      date
#> 1    41    190  7.4   67     5   1 1973-05-01
#> 2    36    118  8.0   72     5   2 1973-05-02
#> 3    12    149 12.6   74     5   3 1973-05-03
#> 4    18    313 11.5   62     5   4 1973-05-04
#> 5    NA     NA  14.3   56     5   5 1973-05-05
#> 6    28     NA  14.9   66     5   6 1973-05-06
#> 7    23    299  8.6   65     5   7 1973-05-07
#> 8    19     99 13.8   59     5   8 1973-05-08
#> 9     8     19 20.1   61     5   9 1973-05-09
#> 10   NA    194  8.6   69     5  10 1973-05-10
#> 11    7     NA  6.9   74     5  11 1973-05-11
#> 12   16    256  9.7   69     5  12 1973-05-12
#> 13   11    290  9.2   66     5  13 1973-05-13
#> 14   14    274 10.9   68     5  14 1973-05-14
#> 15   18     65 13.2   58     5  15 1973-05-15
#> 16   14    334 11.5   64     5  16 1973-05-16
#> 17   34    307 12.0   66     5  17 1973-05-17
#> 18    6     78 18.4   57     5  18 1973-05-18
#> 19   30    322 11.5   68     5  19 1973-05-19
#> 20   11     44  9.7   62     5  20 1973-05-20
```

#> 21	1	8	9.7	59	5	21	1973-05-21
#> 22	11	320	16.6	73	5	22	1973-05-22
#> 23	4	25	9.7	61	5	23	1973-05-23
#> 24	32	92	12.0	61	5	24	1973-05-24
#> 25	NA	66	16.6	57	5	25	1973-05-25
#> 26	NA	266	14.9	58	5	26	1973-05-26
#> 27	NA	NA	8.0	57	5	27	1973-05-27
#> 28	23	13	12.0	67	5	28	1973-05-28
#> 29	45	252	14.9	81	5	29	1973-05-29
#> 30	115	223	5.7	79	5	30	1973-05-30
#> 31	37	279	7.4	76	5	31	1973-05-31
#> 32	NA	286	8.6	78	6	1	1973-06-01
#> 33	NA	287	9.7	74	6	2	1973-06-02
#> 34	NA	242	16.1	67	6	3	1973-06-03
#> 35	NA	186	9.2	84	6	4	1973-06-04
#> 36	NA	220	8.6	85	6	5	1973-06-05
#> 37	NA	264	14.3	79	6	6	1973-06-06
#> 38	29	127	9.7	82	6	7	1973-06-07
#> 39	NA	273	6.9	87	6	8	1973-06-08
#> 40	71	291	13.8	90	6	9	1973-06-09
#> 41	39	323	11.5	87	6	10	1973-06-10
#> 42	NA	259	10.9	93	6	11	1973-06-11
#> 43	NA	250	9.2	92	6	12	1973-06-12
#> 44	23	148	8.0	82	6	13	1973-06-13
#> 45	NA	332	13.8	80	6	14	1973-06-14
#> 46	NA	322	11.5	79	6	15	1973-06-15
#> 47	21	191	14.9	77	6	16	1973-06-16
#> 48	37	284	20.7	72	6	17	1973-06-17
#> 49	20	37	9.2	65	6	18	1973-06-18
#> 50	12	120	11.5	73	6	19	1973-06-19
#> 51	13	137	10.3	76	6	20	1973-06-20
#> 52	NA	150	6.3	77	6	21	1973-06-21
#> 53	NA	59	1.7	76	6	22	1973-06-22
#> 54	NA	91	4.6	76	6	23	1973-06-23
#> 55	NA	250	6.3	76	6	24	1973-06-24
#> 56	NA	135	8.0	75	6	25	1973-06-25
#> 57	NA	127	8.0	78	6	26	1973-06-26
#> 58	NA	47	10.3	73	6	27	1973-06-27
#> 59	NA	98	11.5	80	6	28	1973-06-28
#> 60	NA	31	14.9	77	6	29	1973-06-29
#> 61	NA	138	8.0	83	6	30	1973-06-30
#> 62	135	269	4.1	84	7	1	1973-07-01
#> 63	49	248	9.2	85	7	2	1973-07-02
#> 64	32	236	9.2	81	7	3	1973-07-03
#> 65	NA	101	10.9	84	7	4	1973-07-04
#> 66	64	175	4.6	83	7	5	1973-07-05
#> 67	40	314	10.9	83	7	6	1973-07-06
#> 68	77	276	5.1	88	7	7	1973-07-07
#> 69	97	267	6.3	92	7	8	1973-07-08
#> 70	97	272	5.7	92	7	9	1973-07-09
#> 71	85	175	7.4	89	7	10	1973-07-10
#> 72	NA	139	8.6	82	7	11	1973-07-11
#> 73	10	264	14.3	73	7	12	1973-07-12

#> 74	27	175	14.9	81	7	13	1973-07-13
#> 75	NA	291	14.9	91	7	14	1973-07-14
#> 76	7	48	14.3	80	7	15	1973-07-15
#> 77	48	260	6.9	81	7	16	1973-07-16
#> 78	35	274	10.3	82	7	17	1973-07-17
#> 79	61	285	6.3	84	7	18	1973-07-18
#> 80	79	187	5.1	87	7	19	1973-07-19
#> 81	63	220	11.5	85	7	20	1973-07-20
#> 82	16	7	6.9	74	7	21	1973-07-21
#> 83	NA	258	9.7	81	7	22	1973-07-22
#> 84	NA	295	11.5	82	7	23	1973-07-23
#> 85	80	294	8.6	86	7	24	1973-07-24
#> 86	108	223	8.0	85	7	25	1973-07-25
#> 87	20	81	8.6	82	7	26	1973-07-26
#> 88	52	82	12.0	86	7	27	1973-07-27
#> 89	82	213	7.4	88	7	28	1973-07-28
#> 90	50	275	7.4	86	7	29	1973-07-29
#> 91	64	253	7.4	83	7	30	1973-07-30
#> 92	59	254	9.2	81	7	31	1973-07-31
#> 93	39	83	6.9	81	8	1	1973-08-01
#> 94	9	24	13.8	81	8	2	1973-08-02
#> 95	16	77	7.4	82	8	3	1973-08-03
#> 96	78	NA	6.9	86	8	4	1973-08-04
#> 97	35	NA	7.4	85	8	5	1973-08-05
#> 98	66	NA	4.6	87	8	6	1973-08-06
#> 99	122	255	4.0	89	8	7	1973-08-07
#> 100	89	229	10.3	90	8	8	1973-08-08
#> 101	110	207	8.0	90	8	9	1973-08-09
#> 102	NA	222	8.6	92	8	10	1973-08-10
#> 103	NA	137	11.5	86	8	11	1973-08-11
#> 104	44	192	11.5	86	8	12	1973-08-12
#> 105	28	273	11.5	82	8	13	1973-08-13
#> 106	65	157	9.7	80	8	14	1973-08-14
#> 107	NA	64	11.5	79	8	15	1973-08-15
#> 108	22	71	10.3	77	8	16	1973-08-16
#> 109	59	51	6.3	79	8	17	1973-08-17
#> 110	23	115	7.4	76	8	18	1973-08-18
#> 111	31	244	10.9	78	8	19	1973-08-19
#> 112	44	190	10.3	78	8	20	1973-08-20
#> 113	21	259	15.5	77	8	21	1973-08-21
#> 114	9	36	14.3	72	8	22	1973-08-22
#> 115	NA	255	12.6	75	8	23	1973-08-23
#> 116	45	212	9.7	79	8	24	1973-08-24
#> 117	168	238	3.4	81	8	25	1973-08-25
#> 118	73	215	8.0	86	8	26	1973-08-26
#> 119	NA	153	5.7	88	8	27	1973-08-27
#> 120	76	203	9.7	97	8	28	1973-08-28
#> 121	118	225	2.3	94	8	29	1973-08-29
#> 122	84	237	6.3	96	8	30	1973-08-30
#> 123	85	188	6.3	94	8	31	1973-08-31
#> 124	96	167	6.9	91	9	1	1973-09-01
#> 125	78	197	5.1	92	9	2	1973-09-02
#> 126	73	183	2.8	93	9	3	1973-09-03

```

#> 127    91    189  4.6   93    9    4 1973-09-04
#> 128    47     95  7.4   87    9    5 1973-09-05
#> 129    32     92 15.5   84    9    6 1973-09-06
#> 130    20    252 10.9   80    9    7 1973-09-07
#> 131    23    220 10.3   78    9    8 1973-09-08
#> 132    21    230 10.9   75    9    9 1973-09-09
#> 133    24    259  9.7   73    9   10 1973-09-10
#> 134    44    236 14.9   81    9   11 1973-09-11
#> 135    21    259 15.5   76    9   12 1973-09-12
#> 136    28    238  6.3   77    9   13 1973-09-13
#> 137     9     24 10.9   71    9   14 1973-09-14
#> 138    13    112 11.5   71    9   15 1973-09-15
#> 139    46    237  6.9   78    9   16 1973-09-16
#> 140    18    224 13.8   67    9   17 1973-09-17
#> 141    13     27 10.3   76    9   18 1973-09-18
#> 142    24    238 10.3   68    9   19 1973-09-19
#> 143    16    201  8.0   82    9   20 1973-09-20
#> 144    13    238 12.6   64    9   21 1973-09-21
#> 145    23     14  9.2   71    9   22 1973-09-22
#> 146    36    139 10.3   81    9   23 1973-09-23
#> 147     7     49 10.3   69    9   24 1973-09-24
#> 148    14     20 16.6   63    9   25 1973-09-25
#> 149    30    193  6.9   70    9   26 1973-09-26
#> 150    NA    145 13.2   77    9   27 1973-09-27
#> 151    14    191 14.3   75    9   28 1973-09-28
#> 152    18    131  8.0   76    9   29 1973-09-29
#> 153    20    223 11.5   68    9   30 1973-09-30

```

```

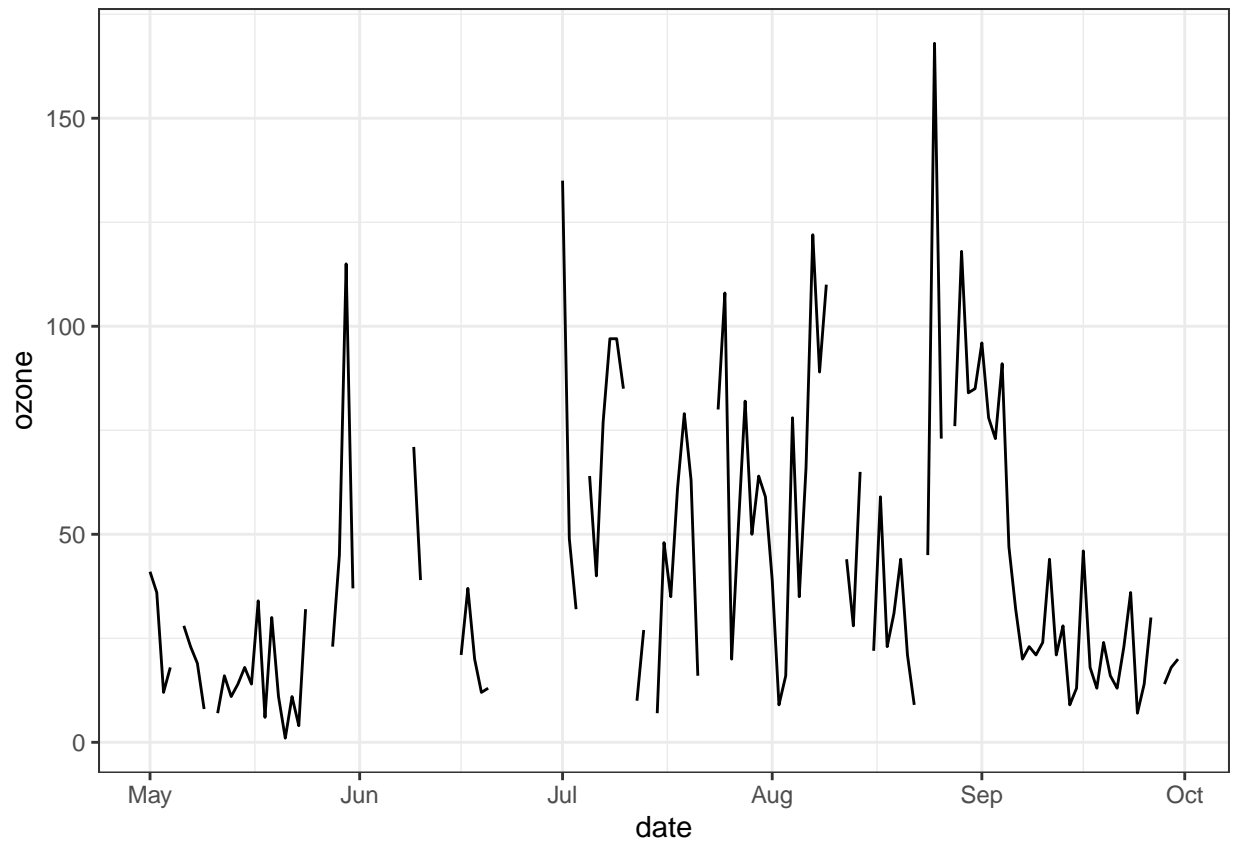
air_quality %>%
  summarize(sum(ozone < 0, na.rm = TRUE))
#>   sum(ozone < 0, na.rm = TRUE)
#> 1                             0

```

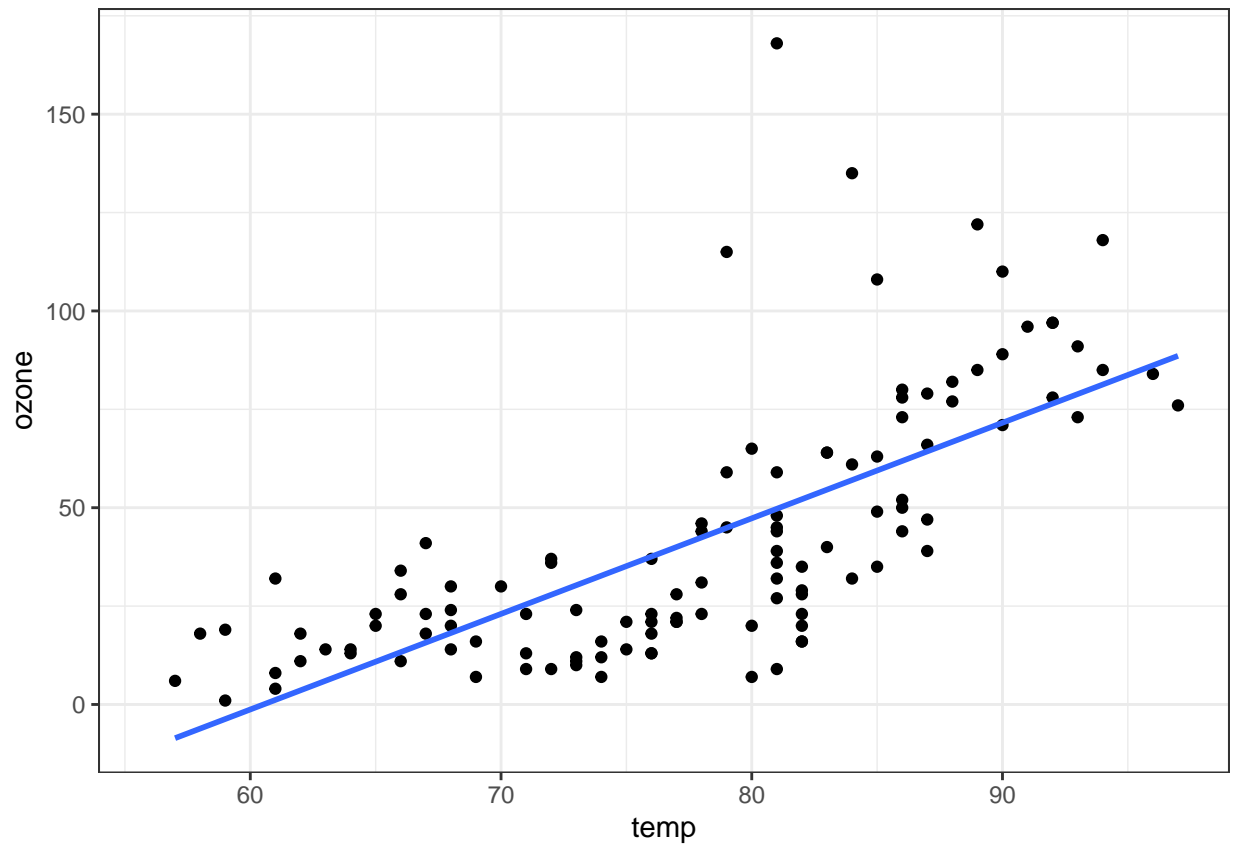
```

ggplot(air_quality, aes(date, ozone)) +
  geom_line()

```



```
ggplot(air_quality, aes(temp, ozone)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
#> `geom_smooth()` using formula 'y ~ x'
#> Warning: Removed 37 rows containing non-finite values (stat_smooth).
#> Warning: Removed 37 rows containing missing values (geom_point).
```



```
# Simple linear regression imputation
```

```
air_quality_imputed <- simputation::impute_lm(air_quality, ozone ~ temp)
```

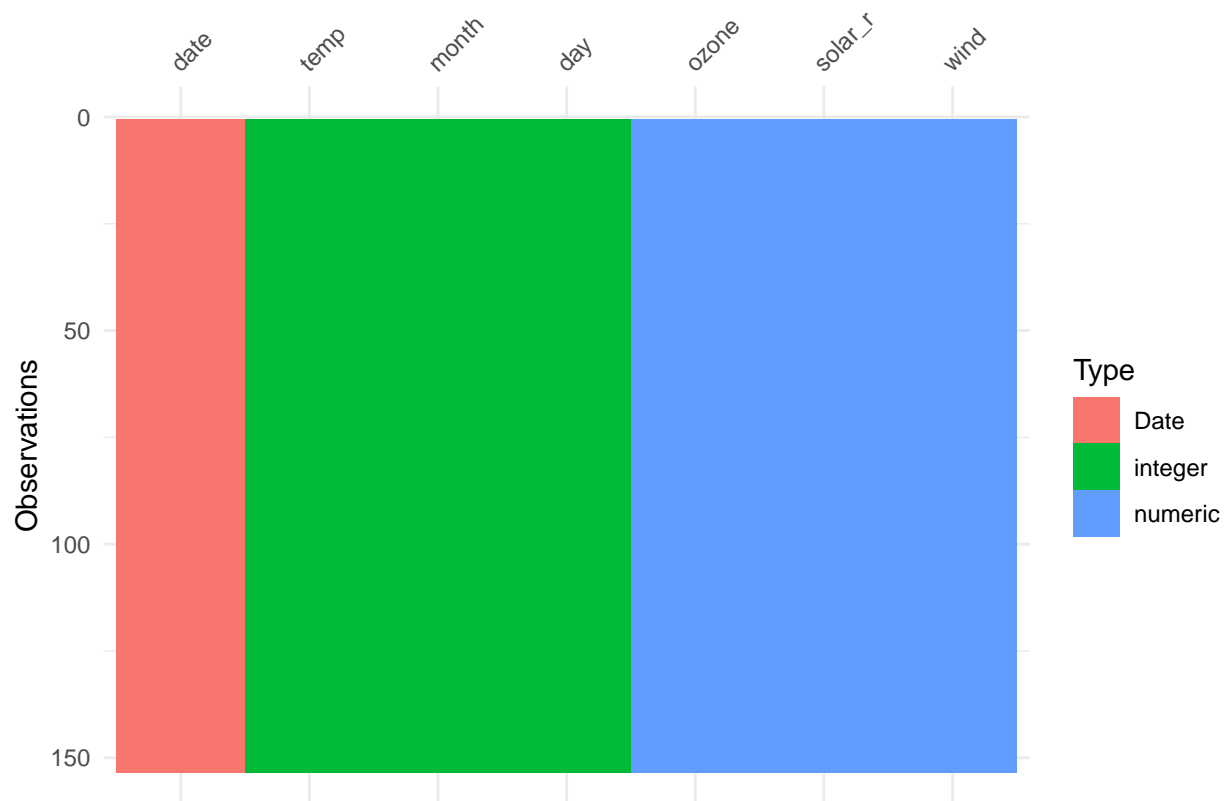
```
air_quality_imputed <- simputation::impute_lm(air_quality_imputed, solar_r ~ temp + ozone)
```

```
# Random forest imputation
```

```
air_quality_imputed2 <- simputation::impute_cart(air_quality, ozone ~ temp + wind + date)
```

```
air_quality_imputed2 <- simputation::impute_cart(air_quality_imputed2, solar_r ~ ozone + temp + wind + date)
```

```
vis_dat(air_quality_imputed)
```



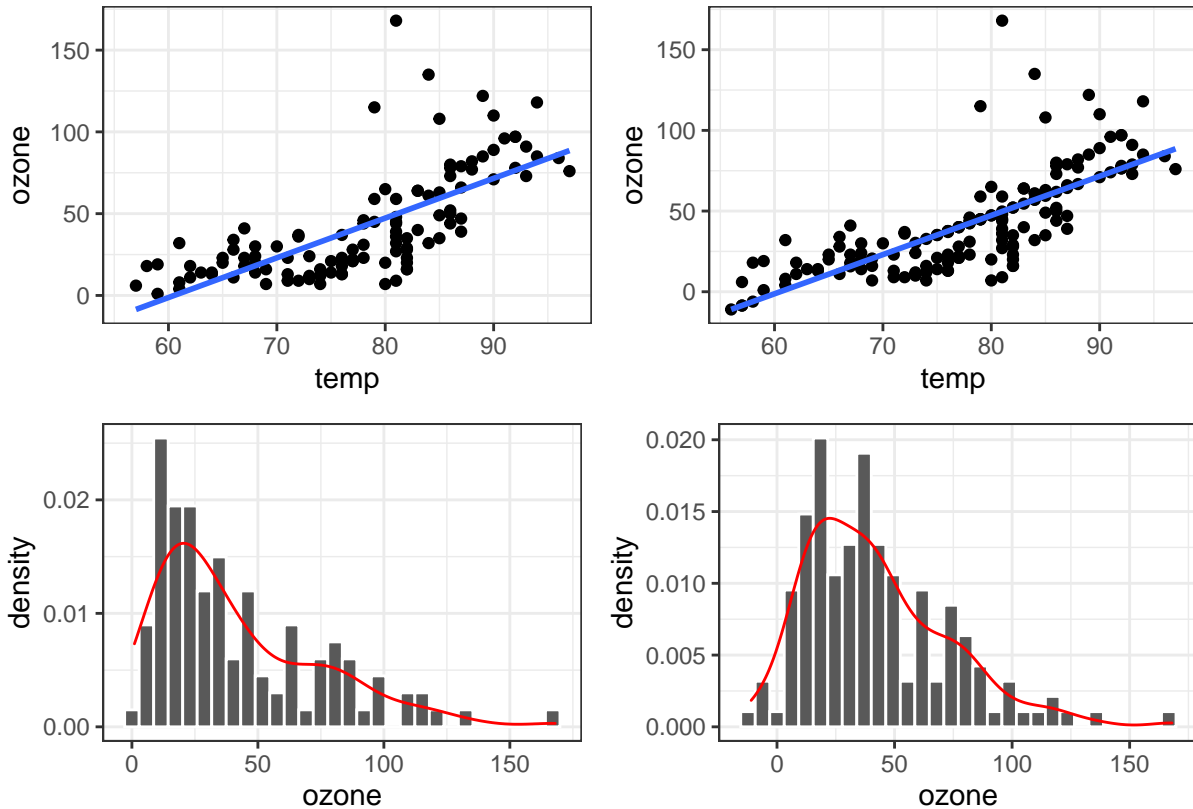
```
# ozone before and after imputation
before1 <- ggplot(air_quality, aes(temp, ozone)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

before2 <- ggplot(air_quality, aes(ozone)) +
  geom_histogram(aes(y = ..density..), color = "white") +
  geom_density(color = "red")

after1 <- ggplot(air_quality_imputed, aes(temp, ozone)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

after2 <- ggplot(air_quality_imputed, aes(ozone)) +
  geom_histogram(aes(y = ..density..), color = "white") +
  geom_density(color = "red")

(before1 + after1) / (before2 + after2)
#> `geom_smooth()` using formula 'y ~ x'
#> Warning: Removed 37 rows containing non-finite values (stat_smooth).
#> Warning: Removed 37 rows containing missing values (geom_point).
#> `geom_smooth()` using formula 'y ~ x'
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 37 rows containing non-finite values (stat_bin).
#> Warning: Removed 37 rows containing non-finite values (stat_density).
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

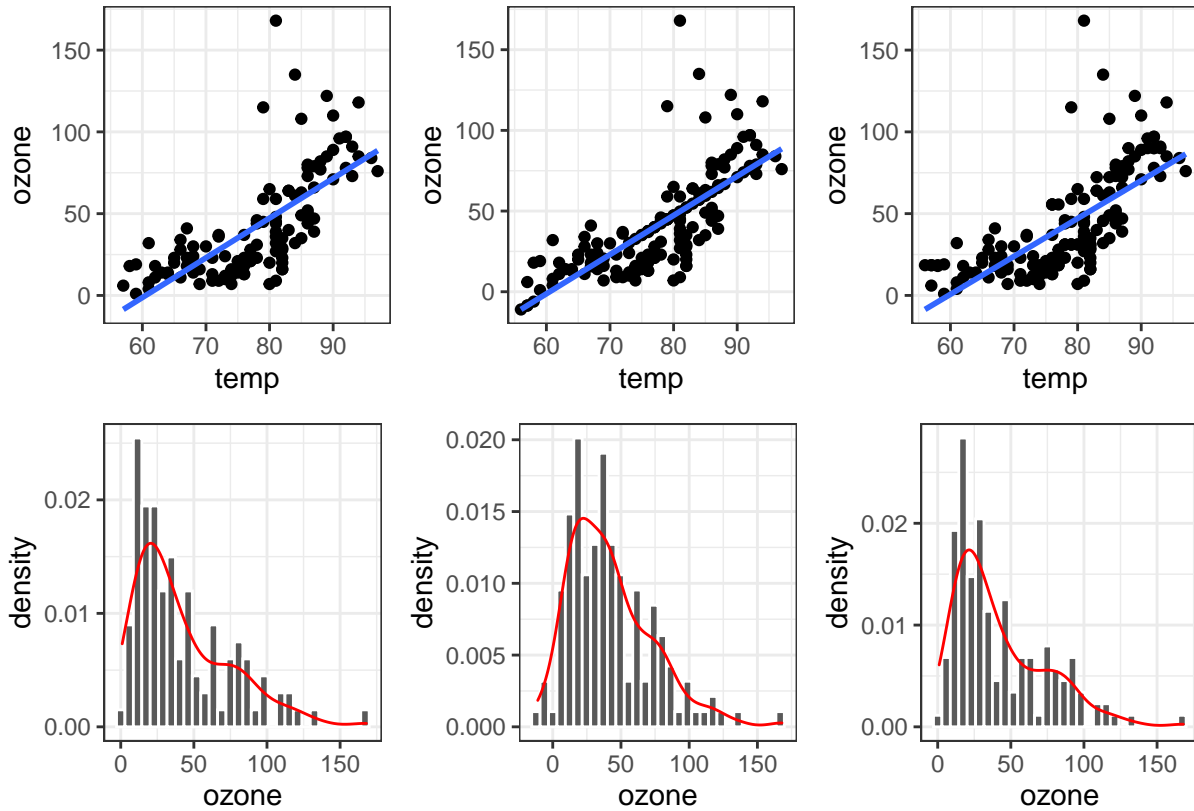


Negative ozone levels after imputation

```
after3 <- ggplot(air_quality_imputed2, aes(temp, ozone)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

after4 <- ggplot(air_quality_imputed2, aes(ozone)) +
  geom_histogram(aes(y = ..density..), color = "white") +
  geom_density(color = "red")

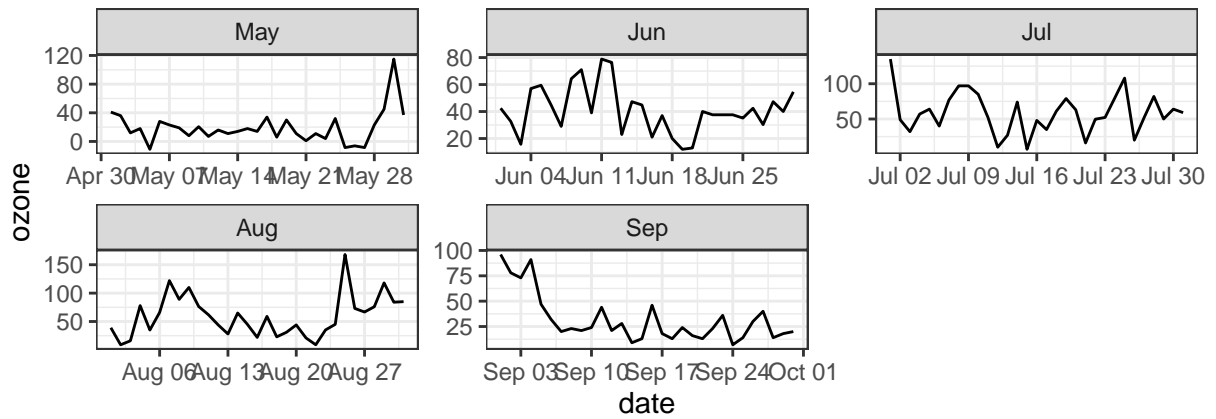
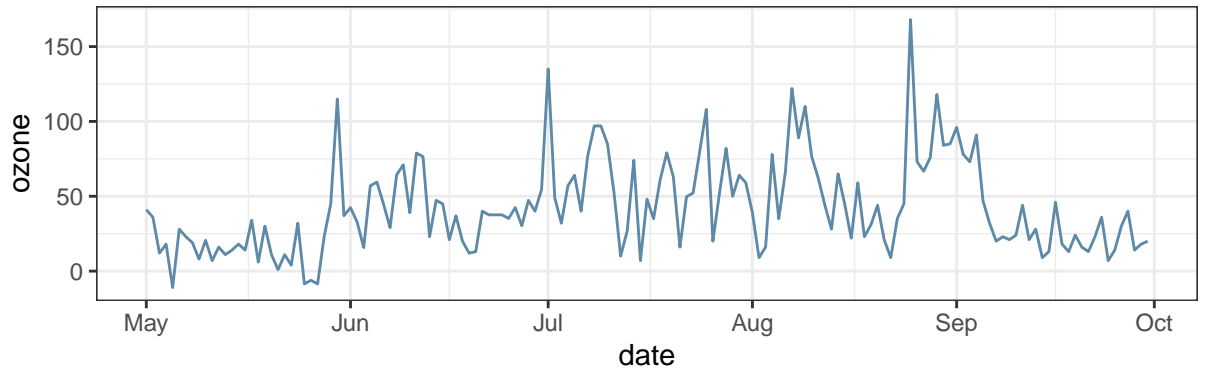
(before1 + after1 + after3) / (before2 + after2 + after4)
#> `geom_smooth()` using formula 'y ~ x'
#> Warning: Removed 37 rows containing non-finite values (stat_smooth).
#> Warning: Removed 37 rows containing missing values (geom_point).
#> `geom_smooth()` using formula 'y ~ x'
#> `geom_smooth()` using formula 'y ~ x'
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 37 rows containing non-finite values (stat_bin).
#> Warning: Removed 37 rows containing non-finite values (stat_density).
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
by_time <- ggplot(air_quality_imputed, aes(date, ozone)) +
  geom_line(color = "#5D8AA8")

by_month <- ggplot(air_quality_imputed, aes(date, ozone)) +
  geom_line() +
  facet_wrap(~lubridate::month(date, label = TRUE), scale = "free")

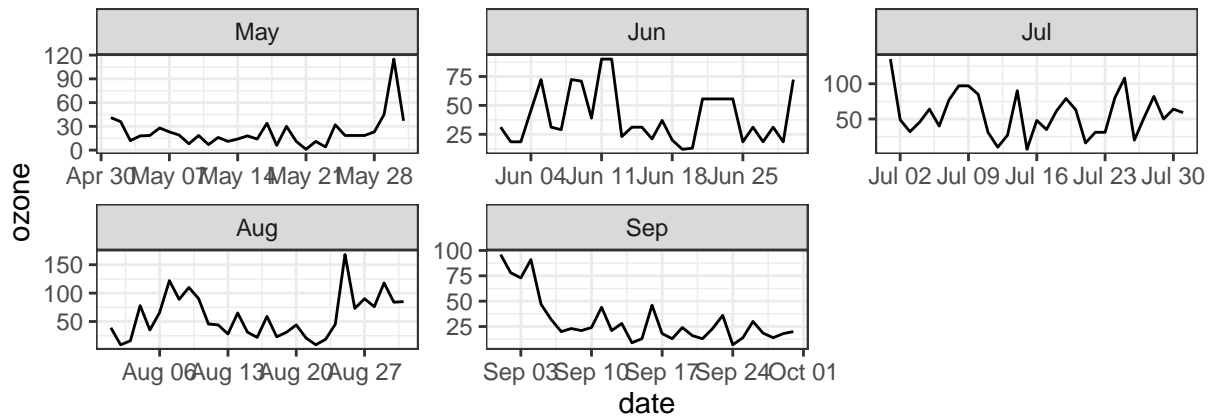
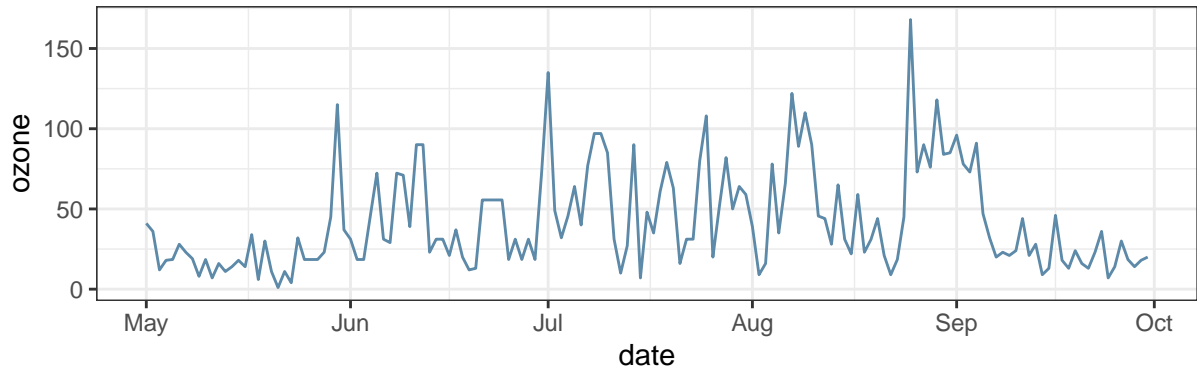
by_time / by_month
```



```
by_time2 <- ggplot(air_quality_imputed2, aes(date, ozone)) +
  geom_line(color = "#5D8AA8")

by_month2 <- ggplot(air_quality_imputed2, aes(date, ozone)) +
  geom_line() +
  facet_wrap(~lubridate::month(date, label = TRUE), scale = "free")

by_time2 / by_month2
```



```
ggplot(air_quality_imputed, aes(date)) +
  geom_line(aes(y = ozone), color = "red") +
  geom_line(data = air_quality, aes(y = ozone))
```

