Prof. Donna Ankerst, Stephan Haug

October 24, 2017

Problem H.1

Download the L.A. ozone data set from moodle course and read it into R using functions from the `readr` package (contained in the `tidyverse`). The data consists of nine predictor, one response (`ozone`) and one id variable.

a) Summarize the univariate distributions of the 9 predictor variables. Use the function `summary()` to produce a numerical summary of the data.

b) Change the format of the data set from wide

```
LAozone
```

```
## # A tibble: 330 × 11
##    ozone    vh  wind humidity  temp   ibh   dpg   ibt   vis   doy    id
##    <int> <int> <int>    <int> <int> <int> <int> <int> <int> <int> <int>
## 1      3  5710     4       28    40  2693   -25    87   250     3     1
## 2      5  5700     3       37    45   590   -24   128   100     4     2
## 3      5  5760     3       51    54  1450    25   139    60     5     3
## 4      6  5720     4       69    35  1568    15   121    60     6     4
## 5      4  5790     6       19    45  2631   -33   123   100     7     5
## 6      4  5790     3       25    55   554   -28   182   250     8     6
## 7      6  5700     3       73    41  2083    23   114   120     9     7
## 8      7  5700     3       59    44  2654    -2    91   120    10     8
## 9      4  5770     8       27    54  5000   -19    92   120    11     9
## 10     6  5720     3       44    51   111     9   173   150    12    10
## # ... with 320 more rows
```

to long

```
LAozone_long
```

```
## # A tibble: 2,970 × 3
##       id variable value
##    <int>    <chr> <int>
## 1      1       vh  5710
## 2      2       vh  5700
## 3      3       vh  5760
## 4      4       vh  5720
## 5      5       vh  5790
## 6      6       vh  5790
## 7      7       vh  5700
## 8      8       vh  5700
## 9      9       vh  5770
## 10    10       vh  5720
## # ... with 2,960 more rows
```

c) Now use the data in long format to create boxplots and histograms by using appropriate functions in the `ggplot2` package.

d) The boxplots in part c) are hard to compare due to the different scales of the predictor variables. Hence, before changing the format, the data should now be scaled (use `scale()`). Now create again the boxplots. Which variable is the most skewed one?

e) Draw a scatterplot of each of the predictor variables versus the response. Can you detect relationships between the predictors and response? Describe them shortly.

f) Convert the variable `doy` (day of the year) into a variable `season` with the two categories "April to September" and "October to March". Draw a scatterplot of `ozone` vs. `dpg`. Indicate the season for each observation with a different colour and a different character. Add a legend. Compare the figure to the scatterplot from e).