Technical University of Munich
Department of Mathematics

Prof. Donna Ankerst, Stephan Haug

October 24, 2017

Grading: For each ✓ you should give 2, for each ✓ 1 and for each ✓ 0.5 points. For each part there exist also other approaches to solve the problem. Hence you have to check if solutions, which are different than the proposed one, are also correct.

Remark: If in the end the total number of points is not an integer (moodle only knows integers), you have to round up. So e.g. a result of 8.5 will be evaluated as 9 in moodle.

Problem H.1 (10 points)

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).

Solution

a)
```
LAozone <- read_csv("LAozone.csv")
```

The command `summary(LAozone[, -c(1, 11)])` yields

```
summary(LAozone[,-c(1,11)])

##        vh             wind           humidity          temp
##  Min.   :5320   Min.   : 0.000   Min.   :19.00   Min.   :25.00
##  1st Qu.:5690   1st Qu.: 3.000   1st Qu.:47.00   1st Qu.:51.00
##  Median :5760   Median : 5.000   Median :64.00   Median :62.00
##  Mean   :5750   Mean   : 4.891   Mean   :58.13   Mean   :61.75
##  3rd Qu.:5830   3rd Qu.: 6.000   3rd Qu.:73.00   3rd Qu.:72.00
##  Max.   :5950   Max.   :21.000   Max.   :93.00   Max.   :93.00
##       ibh              dpg              ibt             vis
##  Min.   : 111.0   Min.   :-69.00   Min.   :-25.0   Min.   :  0.0
##  1st Qu.: 877.5   1st Qu.: -9.00   1st Qu.:107.0   1st Qu.: 70.0
##  Median :2112.5   Median : 24.00   Median :167.5   Median :120.0
##  Mean   :2572.9   Mean   : 17.37   Mean   :161.2   Mean   :124.5
##  3rd Qu.:5000.0   3rd Qu.: 44.75   3rd Qu.:214.0   3rd Qu.:150.0
##  Max.   :5000.0   Max.   :107.00   Max.   :332.0   Max.   :350.0
##       doy
##  Min.   :  3.00
##  1st Qu.: 90.25
##  Median :177.50
##  Mean   :181.73
##  3rd Qu.:275.75
##  Max.   :365.00
```

✓

b) Now we want to change the format of the data set. This is done by using the function `gather()` (see `?gather`).

```
LAozone_long <- gather(LAozone[,-1], variable, value, -id)
```

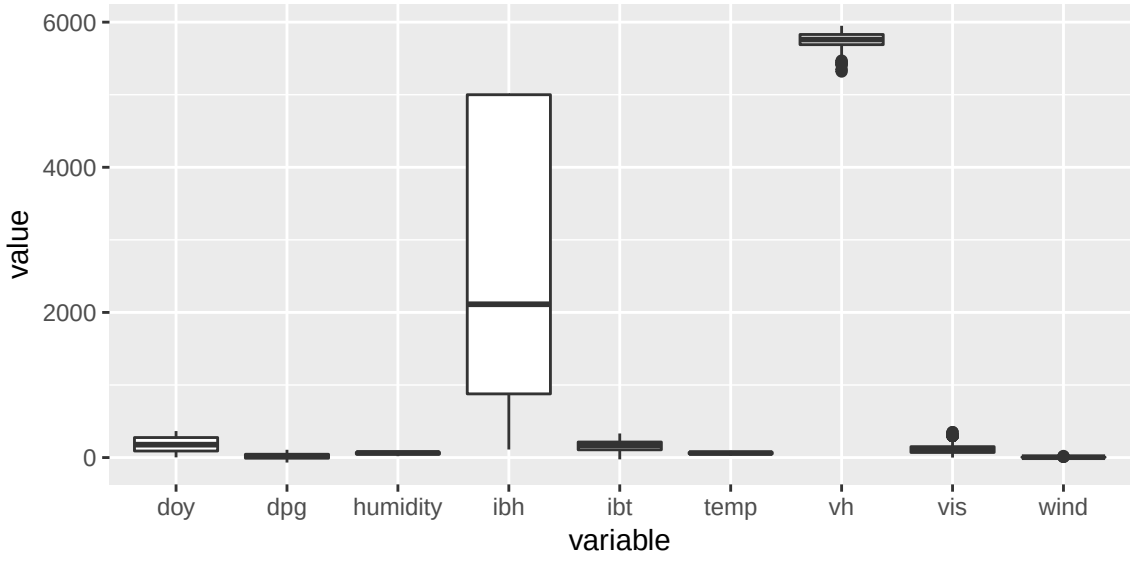
```
LAozone_long
```

```
## # 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
```

In the long format it will now be very easy to create the boxplots and histograms in part c) by using functions from the `ggplot2` package. Remember also that one main requirement for using `ggplot()` is, that the data is given as a data frame.

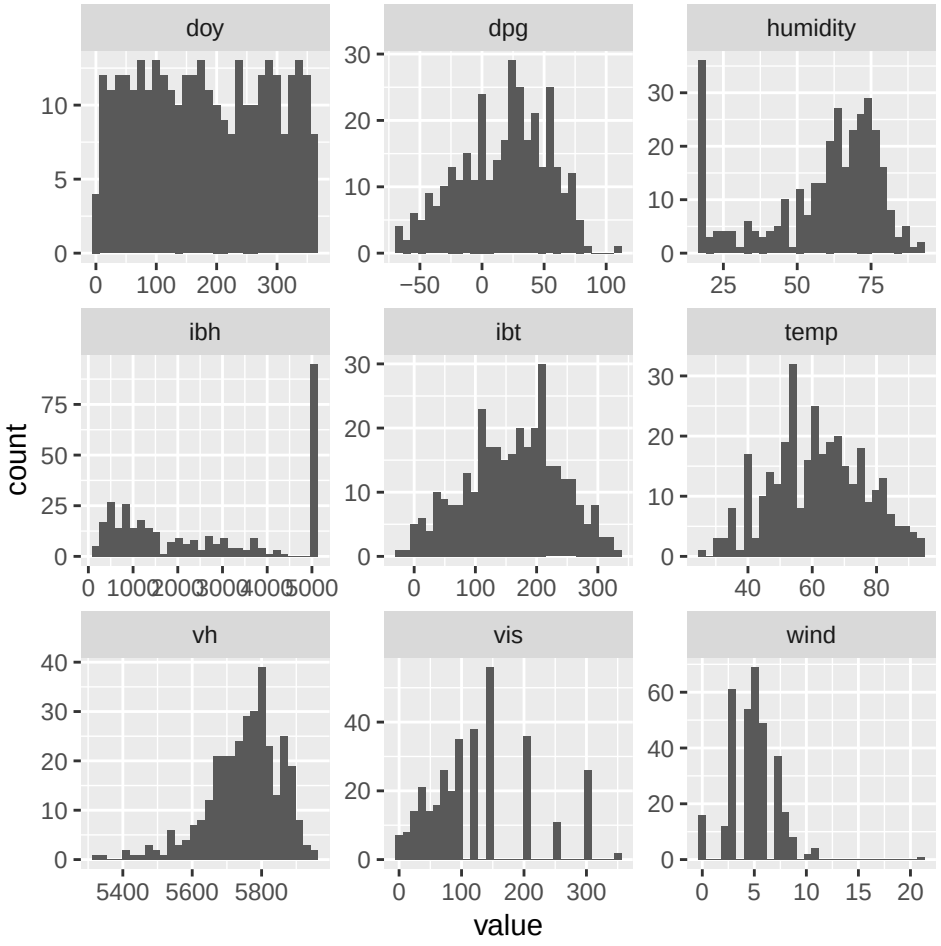
c) In the next step we summarize the distributions of the predictor variables by using boxplots and histograms. We use the function `ggplot()` together with the geoms `geom_boxplot()` and `geom_histogram()` to create the boxplots and histograms, respectively. See `http://ggplot2.tidyverse.org/reference/` for more details on those functions.

```
ggplot(LAozone_long, aes(x = variable, y = value)) + geom_boxplot()
```



To produce separate histograms one needs an additional function. We introduce a further facet by distinguishing between different histograms of `value` in terms of `variable`. `facet_wrap` then wraps does different panels in a 2d representation.

```
ggplot(LAozone_long, aes(value)) + facet_wrap(~ variable, scales = "free") +
  geom_histogram()
```



✓

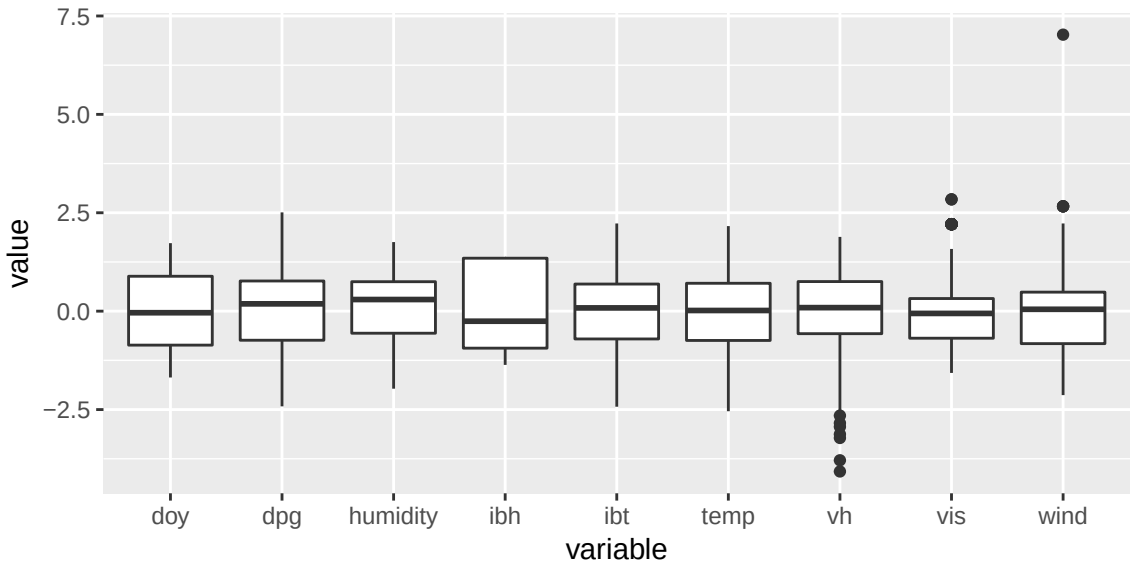
Remark: If the figure, containing the nine histograms, was created using the data in the wide format, then one just gets 0.5 points. Thus ✓ instead of ✓.

d) The predictor variables are measured on quite different scales. Therefore it is not easy to compare the variation of the predictor variables using the boxplots in part c). Hence we will now first scale the data and then create the boxplots.

```
LAozone_long_scaled <- gather(data.frame(scale(LAozone[,-c(1, 11)]),
                                         id = LAozone$id), variable, value, -id)
```

✓

```
ggplot(LAozone_long_scaled, aes(x = variable, y = value)) + geom_boxplot()
```



✓

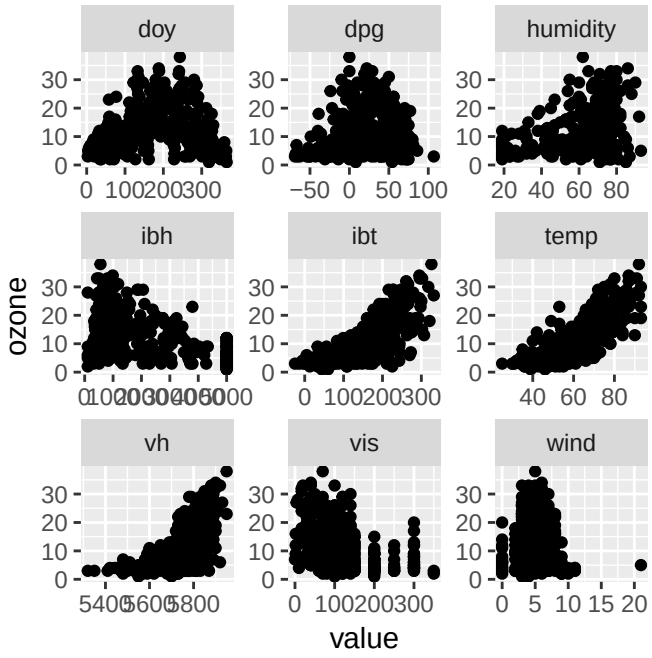
Still `ibh` shows the largest spread and is most skewed. ✓

e) Now we should create scatterplots for each of the predictor variables against the response. We still would like to use the long format, but of course now we have to also include the response `ozone`

```
LAozone_long <- gather(LAozone, variable, value, -id, -ozone)
```

Now we can again use `ggplot()` to create the scatterplots.

```
ggplot(LAozone_long, aes(x = value, y = ozone)) +
  facet_wrap(~ variable, scales = "free") + geom_point()
```



✓

For some of the predictors, we can indeed detect a relationship with the response. For exam-

ple, for `vh`, `humidity`, `temp` and `ibt`, the ozone level seems to increase with the value of the corresponding variable. For `temp` and `ibt` this relationship even appears to be quite linear. ✓ Further, `doy` and `dpg` seem to have quadratic relationship with the outcome. ✓

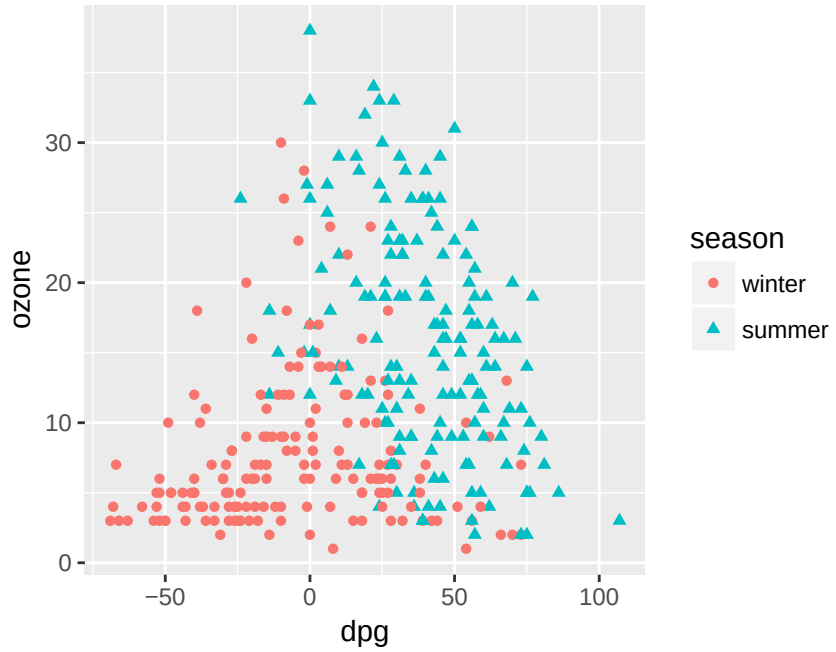
f) First we have to create the factor variable `season`. This is done by identifying (logical operation) the winter observation (the rest has to be summer) and converting this information to a factor variable.

```
LAozone$season <- factor(LAozone$doy <= 89 | LAozone$doy > 273,
                         levels = c(TRUE, FALSE), labels = c("winter", "summer"))
```

✓

Choosing different colours and including a legend is now very simple in `ggplot2`. We just use (inside `aes()`) a variable - in our case `season` - to specify different colours. A legend is then added by default. The style could of course be changed, but for the moment this is enough for us.

```
ggplot(LAozone, aes(x = dpg, y = ozone, col = season, shape = season)) +
  geom_point()
```



If colour and shape are different for the different seasons ✓. *If just one of them varies with season* ✓.

If we consider the two seasons separately, we detect a positive linear relationship for the winter season and a negative linear relationship for the summer season. ✓