Breast Cancer Prediction

Andrea Becsek

2 January 2020

```
library(tidyverse)
```

```
## -- Attaching packages ------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1       v purrr   0.3.2
## v tibble  2.1.3       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts --------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
library(ggcorrplot)
library(devtools)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
```

```
##     accumulate, when
```

```
## Loaded glmnet 2.0-18
```

Import data

```
data <- read.csv("../data/data.csv")
glimpse(data)
```

```
## Observations: 569
## Variables: 33
## $ id                      <int> 842302, 842517, 84300903, 84348301, 8435840...
## $ diagnosis               <fct> M, M, M, M, M, M, M, M, M, M, M, M, M, M, M...
## $ radius_mean             <dbl> 17.990, 20.570, 19.690, 11.420, 20.290, 12....
## $ texture_mean            <dbl> 10.38, 17.77, 21.25, 20.38, 14.34, 15.70, 1...
## $ perimeter_mean          <dbl> 122.80, 132.90, 130.00, 77.58, 135.10, 82.5...
## $ area_mean               <dbl> 1001.0, 1326.0, 1203.0, 386.1, 1297.0, 477....
## $ smoothness_mean         <dbl> 0.11840, 0.08474, 0.10960, 0.14250, 0.10030...
## $ compactness_mean        <dbl> 0.27760, 0.07864, 0.15990, 0.28390, 0.13280...
## $ concavity_mean          <dbl> 0.30010, 0.08690, 0.19740, 0.24140, 0.19800...
## $ concave.points_mean     <dbl> 0.14710, 0.07017, 0.12790, 0.10520, 0.10430...
## $ symmetry_mean           <dbl> 0.2419, 0.1812, 0.2069, 0.2597, 0.1809, 0.2...
## $ fractal_dimension_mean  <dbl> 0.07871, 0.05667, 0.05999, 0.09744, 0.05883...
## $ radius_se               <dbl> 1.0950, 0.5435, 0.7456, 0.4956, 0.7572, 0.3...
## $ texture_se              <dbl> 0.9053, 0.7339, 0.7869, 1.1560, 0.7813, 0.8...
## $ perimeter_se            <dbl> 8.589, 3.398, 4.585, 3.445, 5.438, 2.217, 3...
## $ area_se                 <dbl> 153.40, 74.08, 94.03, 27.23, 94.44, 27.19, ...
## $ smoothness_se           <dbl> 0.006399, 0.005225, 0.006150, 0.009110, 0.0...
## $ compactness_se          <dbl> 0.049040, 0.013080, 0.040060, 0.074580, 0.0...
## $ concavity_se            <dbl> 0.05373, 0.01860, 0.03832, 0.05661, 0.05688...
## $ concave.points_se       <dbl> 0.015870, 0.013400, 0.020580, 0.018670, 0.0...
## $ symmetry_se             <dbl> 0.03003, 0.01389, 0.02250, 0.05963, 0.01756...
## $ fractal_dimension_se    <dbl> 0.006193, 0.003532, 0.004571, 0.009208, 0.0...
## $ radius_worst            <dbl> 25.38, 24.99, 23.57, 14.91, 22.54, 15.47, 2...
## $ texture_worst           <dbl> 17.33, 23.41, 25.53, 26.50, 16.67, 23.75, 2...
## $ perimeter_worst         <dbl> 184.60, 158.80, 152.50, 98.87, 152.20, 103....
## $ area_worst              <dbl> 2019.0, 1956.0, 1709.0, 567.7, 1575.0, 741....
## $ smoothness_worst        <dbl> 0.1622, 0.1238, 0.1444, 0.2098, 0.1374, 0.1...
## $ compactness_worst       <dbl> 0.6656, 0.1866, 0.4245, 0.8663, 0.2050, 0.5...
## $ concavity_worst         <dbl> 0.71190, 0.24160, 0.45040, 0.68690, 0.40000...
## $ concave.points_worst    <dbl> 0.26540, 0.18600, 0.24300, 0.25750, 0.16250...
## $ symmetry_worst          <dbl> 0.4601, 0.2750, 0.3613, 0.6638, 0.2364, 0.3...
## $ fractal_dimension_worst <dbl> 0.11890, 0.08902, 0.08758, 0.17300, 0.07678...
## $ X                       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

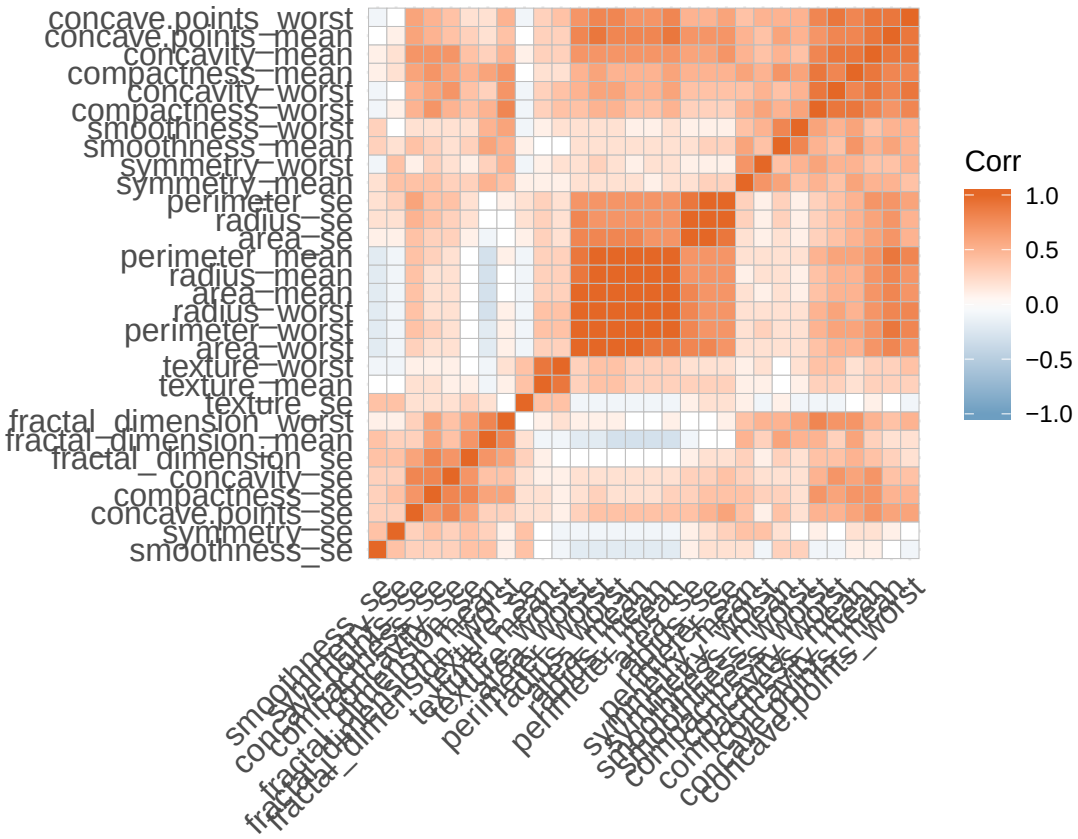
Check for NAs

Remove *id* and *X*

```
data %<>%
  dplyr::select(-c(id, X))
```

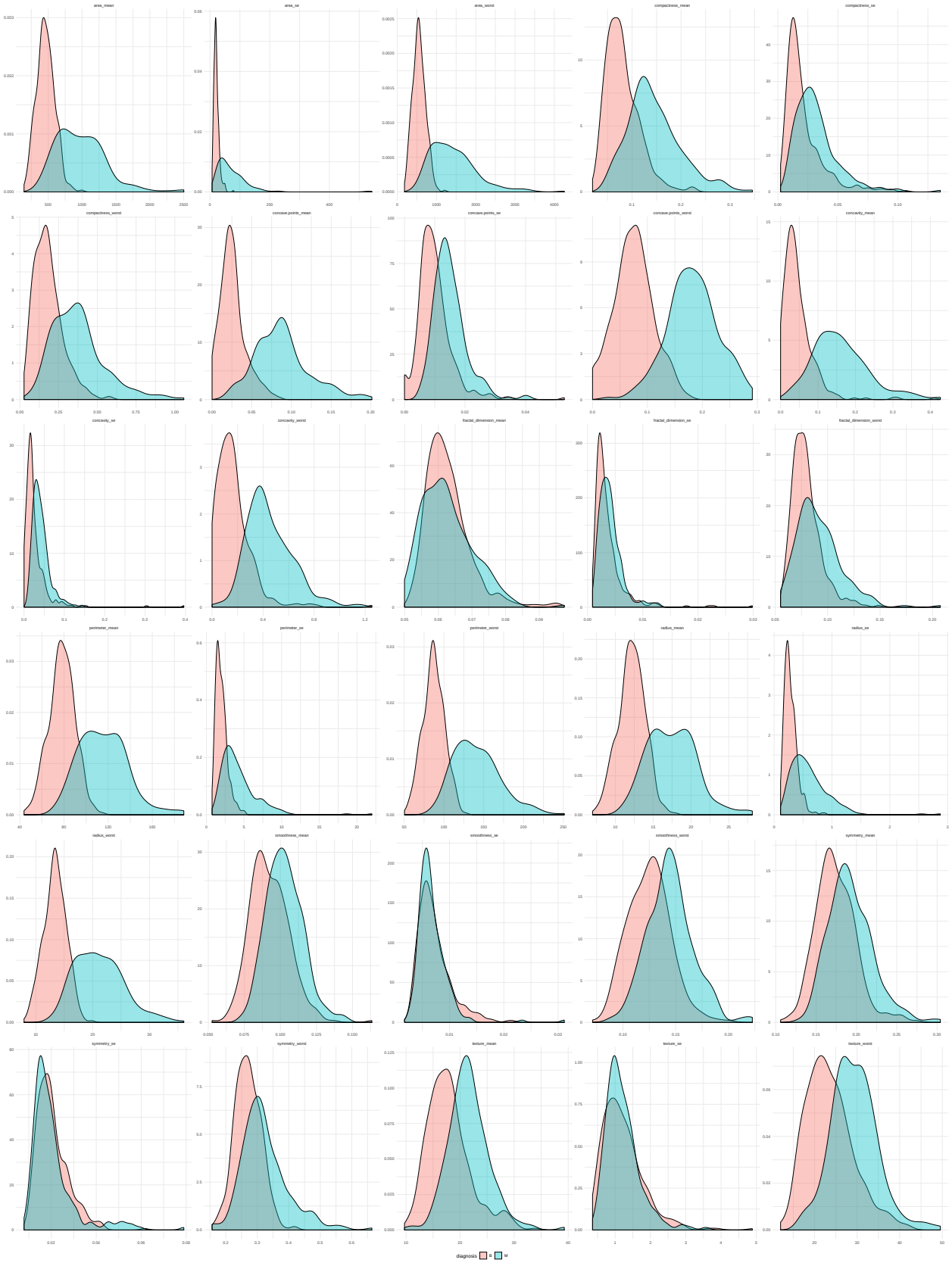
Correlation plot

```
corr <- data[,-1] %>%
          cor() %>%
           round(1)
ggcorrplot(corr, hc.order = TRUE,colors = c("#6D9EC1", "white", "#E46726"), ggtheme = ggplot2::theme_mi
```



Plot

```
data %>%
  group_by(diagnosis) %>%
  gather(key="var",value="value",-diagnosis) %>%
  ggplot(aes(x=value))+
  facet_wrap(~var,scales='free',nrow=6)+
  geom_density(aes(fill=diagnosis),alpha=0.4)+
  theme_minimal()+
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        legend.position="bottom")
```


area_mean
area_se
area_worst
compactness_mean
compactness_se
compactness_worst
concave.points_mean
concave.points_se
concave.points_worst
concavity_mean
concavity_se
concavity_worst
fractal_dimension_mean
fractal_dimension_se
fractal_dimension_worst
perimeter_mean
perimeter_se
perimeter_worst
radius_mean
radius_se
radius_worst
smoothness_mean
smoothness_se
smoothness_worst
symmetry_mean
symmetry_se
symmetry_worst
texture_mean
texture_se
texture_worst
diagnosis B M