Breast Cancer EDA

Jas Kainth

30/09/2020

What is the proportion of the dependent variable?

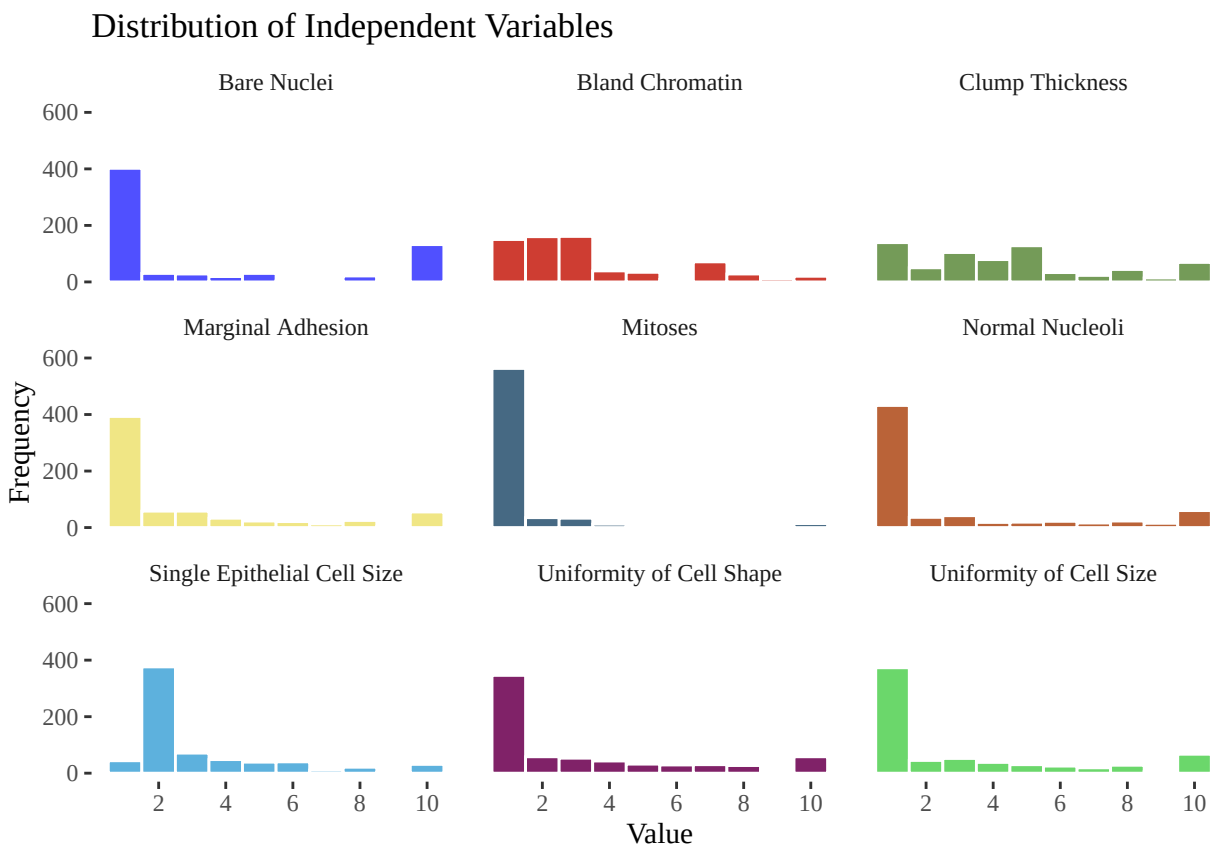
```
# What are the proportion of cases that we have?
data %>%
  mutate(for_table = case_when(class_logistic == 0 ~ "Benign",
            TRUE ~ "Malignant")) %>%
  pull(for_table) %>%
  table() %>%
  knitr::kable(col.names = c("Type", "Frequncy"))
```

Type	Frequncy
Benign	444
Malignant	239

There are a lot more Benign cases than there are Malignant. The proportion of benign is 0.6501 and 0.3499 for malignant.

How are the covariates (Independent Variables) distributed?

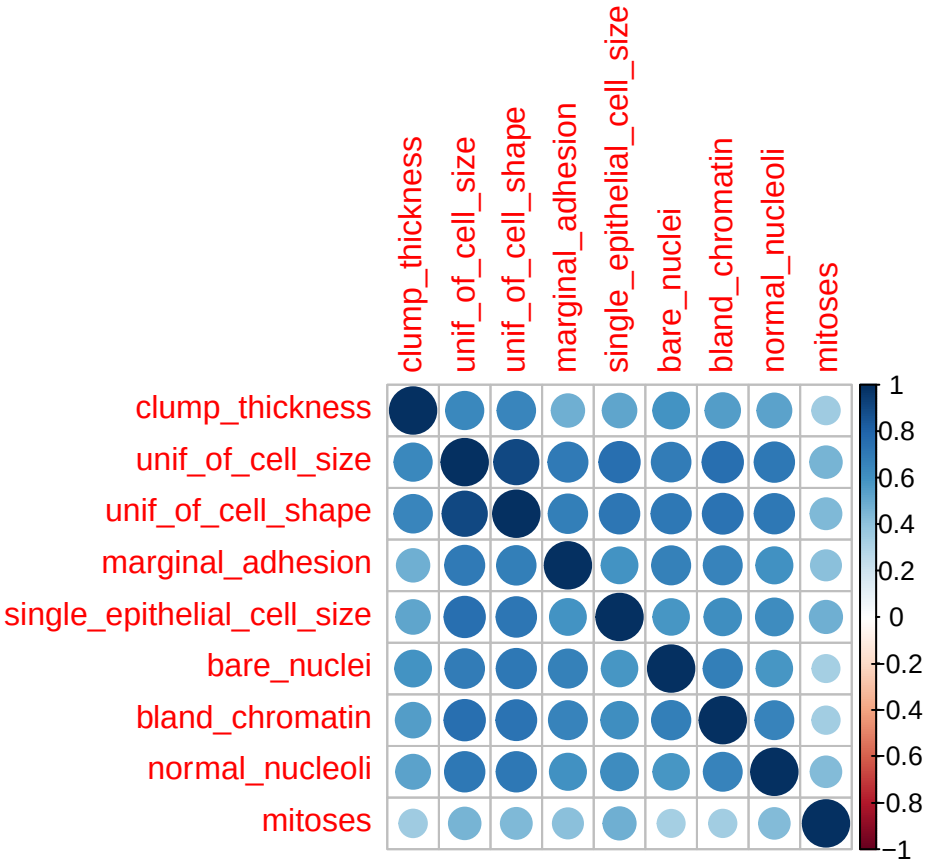
```
# How are most of the covariates distributed?
covariates <- data %>%
  select(-class, -class_logistic)
covariates %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(value)) +
  geom_histogram(aes(fill = variable), bins = 10, color = "white") +
  facet_wrap(~variable, labeller = as_labeller(labels)) +
  theme_tufte() +
  theme(legend.position = "none") +
  scale_fill_igv() +
  scale_x_continuous(breaks = c(2, 4, 6, 8, 10)) +
  scale_y_continuous(limits = c(0, 600)) +
  labs(title = "Distribution of Independent Variables",
       x = "Value",
       y = "Frequency")
```




```
# So most of these peak at 1, with some that also have mini peaks early on or near
# 10
# This is interesting, we might want to check the correlation among these variables
# to ensure there aren't some with high correlation
```

Most of them peak at 1, with some having mini peaks near the 2 extremes. These may be benign cases. It might be that as we get higher values for each of the independent variables that it is more likely the cancer type is malignant. However, to get more concrete results we will make some models during the statistical analysis. However, the fact that a lot of the distributions for these variables are similar might mean they are also highly correlated so let's check the correlation

```
corrplot::corrplot(cor(covariates))
```




```
# Other than mitosis, it seems like most other plots have a high correlation
# The good thing is that there is no high negative correlation which may cause the
# model to think they both are significant when in fact they are just "cancelling"
# each other out
# We may want to come back to this later
```

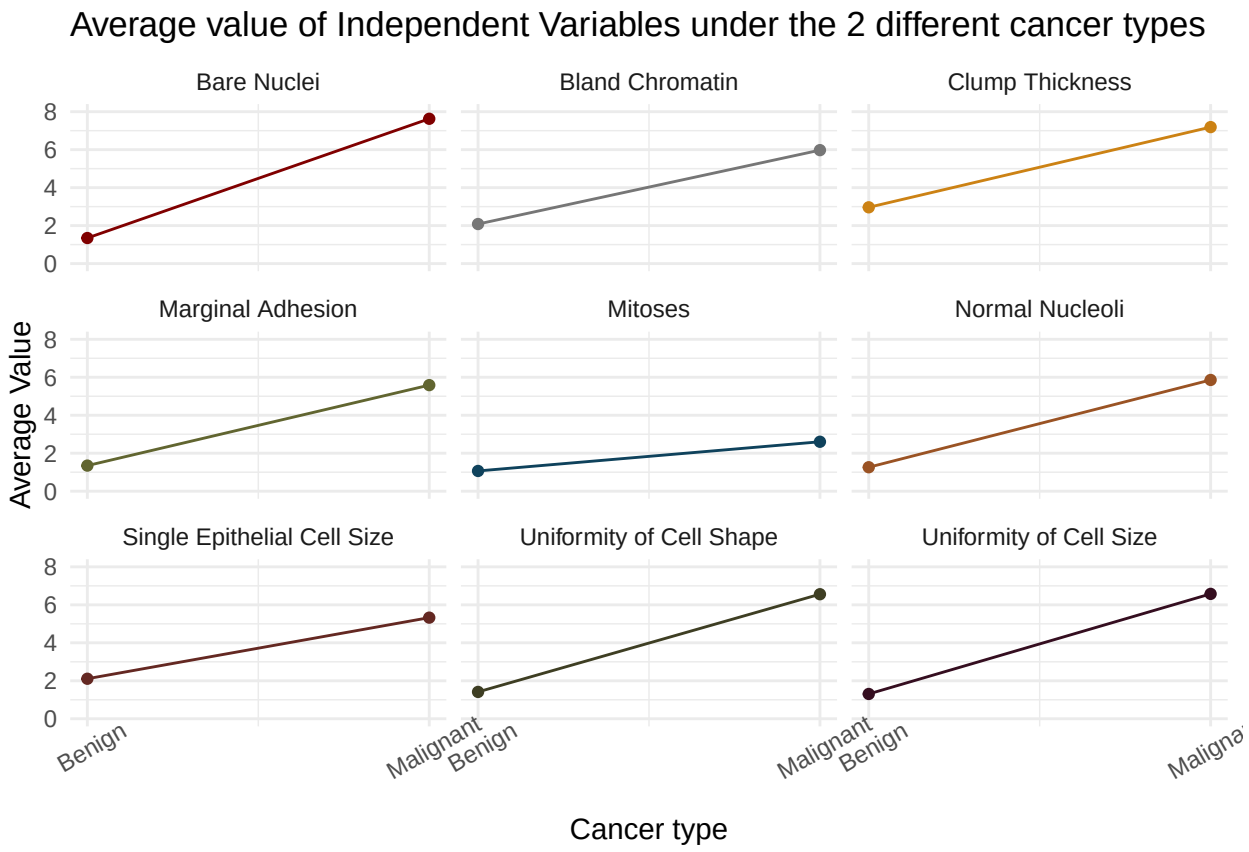
Other than mitosis, which still has a pretty high correlation with the other variables, the correlation is still high. Since we don't have any negative correlation, this means that our model won't think that the variables are both significant when in fact they are just "cancelling" each other our. However, because of this high correlation, this means we will get non-significant p-values for the covariates for variables that are, in fact, important. This means we will want to use other statistics which indicates the goodness of fit (like AIC or likelihood ratio tests) when we are selecting the optimal model.

How do the independent variables affect the dependent variable?

```
# Compare the predictor values to the dependent variables
# How do the averages vary over the two possible outcomes?
data %>%
  select(-class) %>%
  pivot_longer(cols = clump_thickness:mitoses, names_to = "variable",
               values_to = "value") %>%
  group_by(variable, class_logistic) %>%
  summarise(avg = mean(value),
            n = n()) %>%
  ggplot(aes(x = class_logistic, y = avg)) +
```

```
  geom_line(aes(color = variable)) +
  geom_point(aes(color = variable)) +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_color_uchicago(palette = "dark") +
  facet_wrap(~ variable, labeller = as_labeller(labels)) +
  scale_x_continuous(breaks = c(0, 1),
                     labels = c("Benign", "Malignant")) +
  scale_y_continuous(limits = c(0, 8)) +
  labs(title = "Average value of Independent Variables under the 2 different cancer types",
       x = "Cancer type",
       y = "Average Value") +
  theme(axis.text.x = element_text(angle = 30))
```

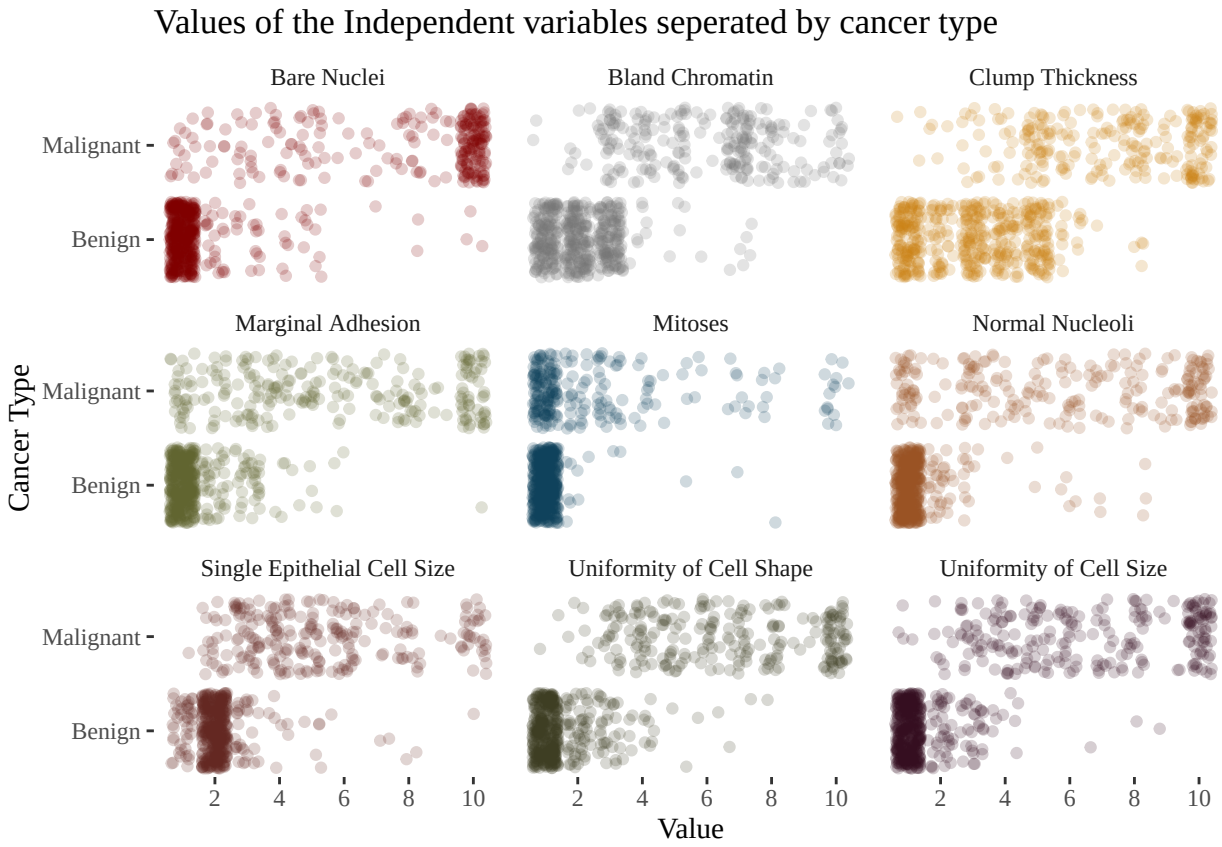
```
## `summarise()` regrouping output by 'variable' (override with `.groups` argument)
```



For every independent variable, going from benign to malignant, the average increases. This further supports our idea that since there are more benign cases and in the histogram, the majority of the points are near 1, the higher, more extreme values are probably the cases for malignant cancer.

```
# Plot all the points
# It's easier to see with a jitter plot rather than a point graph
data %>%
  select(-class) %>%
  pivot_longer(cols = clump_thickness:mitoses, names_to = "variable",
               values_to = "value") %>%
  ggplot(aes(x = value, y = class_logistic)) +
```

```
  geom_jitter(aes(color = variable), alpha = 0.2) +
  theme_tufte() +
  facet_wrap(~ variable, labeller = as_labeller(labels)) +
  scale_color_uchicago(palette = "dark") +
  scale_y_continuous(breaks = c(0, 1),
                     labels = c("Benign", "Malignant")) +
  scale_x_continuous(breaks = c(2, 4, 6, 8, 10)) +
  theme(legend.position = "none") +
  labs(title = "Values of the Independent variables seperated by cancer type",
       x = "Value",
       y = "Cancer Type")
```




```
# Most of these seem to have a trend where the higher the value, the more often
# y = 1 (malignant)
# Most of them have dark values where the predictor is 1 and y = 1 but that's
# because we have a lot of points at 1 for the predictors
# It does seem like all of them are important though
```

Most of these have a trend where the higher the value, the more likely you are to be in the malignant case. One interesting result which is visible is there are more malignant cases when the independent variable is close to 1 than there are benign cases when the independent variable is near 10.

So it does seem that all of these independent variables are important in predicting the results of cancer type. However, we will get insignificant p-values during the analysis phase due to the high correlation. There was not much to explore on the surface of this data since it is all quite simple and scaled.