CaseStudy1

Caterina Ponti, Daranaim Mahamad, Krrish Ghindani, Ursula Ontiveros

2024-09-16

```
#Loading the data frame
load("//Users/caterinaponti/Desktop/BSDS100/ramen.Rdata")

#summarize ramen data
summary(ramen)
```

```
##       Brand                    Variety         Style             Country
##  Nissin  : 381   Beef              :   7   Pack   :1531   Japan      : 352
##  Nongshim:  98   Chicken           :   7   Bowl   : 481   USA        : 323
##  Maruchan:  76   Artificial Chicken:   6   Cup    : 450   South Korea: 309
##  Mama    :  71   Vegetable         :   6   Tray   : 108   Taiwan     : 224
##  Paldo   :  66   Yakisoba          :   6   Box    :   6   Thailand   : 191
##  Myojo   :  63   Miso Ramen        :   5          :   2   China      : 169
##  (Other) :1825   (Other)           :2543   (Other):   2   (Other)    :1012
##      Stars           Top.Ten       perc_salt
##  Min.   :0.000           :2543   Min.   : 3.691
##  1st Qu.:3.250   2012 #1 :   1   1st Qu.:18.372
##  Median :3.750   2012 #10:   1   Median :19.340
##  Mean   :3.655   2012 #2 :   1   Mean   :18.951
##  3rd Qu.:4.250   2012 #3 :   1   3rd Qu.:20.198
##  Max.   :5.000   2012 #4 :   1   Max.   :22.870
##  NA's   :3       (Other) :  32
```

Number 1

```
#how many different brands in the data set
length(unique(ramen$Brand))
```

```
## [1] 355
```

There are 355 different brands in the data set.

Number 2

```
#turning Top.Ten data in a string column
ramen$Top.Ten <- as.character(ramen$Top.Ten, rm.na=TRUE)
#subsetting the year
years <- (substr(ramen$Top.Ten, 1, 4))

#printing unique values for year
print("Years with Top Ten data: ")
```

```
## [1] "Years with Top Ten data: "
```

```
unique(years)
```

```
## [1] ""     "2016" "2015" "2013" "2014" "2012"
```

Top ten data are from years: 2016, 2015, 2013, 2014, 2012.

Number 3

```
#ramen brands from the United States
USA.brands <- which(ramen$Country == 'USA', 'United States' )

print("Ramen Brands from the US:")
```

```
## [1] "Ramen Brands from the US:"
```

```
unique(ramen$Brand[USA.brands])
```

```
##  [1] Nissin                          Yamachan
##  [3] Jackpot Teriyaki                Lipton
##  [5] Pringles                        Myojo
##  [7] Daifuku                         Dream Kitchen
##  [9] Dr. McDougall's                 Shirakiku
## [11] Mama Pat's                      Goku-Uma
## [13] Gefen                           Farmer's Heart
## [15] Nongshim                        Maruchan
## [17] Roland                          Koyo
## [19] IbuRamen                        Fortune
## [21] Thai Smile                      Sapporo Ichiban
## [23] Crystal Noodle                  Authentically Asian
## [25] One Dish Asia                   Thai Pavilion
## [27] Osaka Ramen                     Annie Chun's
## [29] Snapdragon                      Miracle Noodle
## [31] Lotus Foods                     Sakura Noodle
## [33] Thai Kitchen                    Komforte Chockolates
## [35] Tasty Bite                      Star Anise Foods
## [37] Tradition                       Sun Noodle
## [39] S&S                             Right Foods
## [41] Hosoonyi                        Mexi-Ramen
## [43] Chikara                         US Canning
## [45] Tayho                           Fu Chang Chinese Noodle Company
## [47] Teriyaki Time                   Smack
## [49] Westbrae
## 355 Levels: 1 To 3 Noodles 7 Select 7 Select/Nissin A-One ... Zow Zow
```

Number 4

```
#subsetting to find place won by winning ramen
rating <- (substr(ramen$Top.Ten, 6, 7))
#selecting who won first place
top.1.indeces <- which(rating == '#1')

#getting the brands who won first place
top1.brands <- ramen$Brand[top.1.indeces]

#store top1.brands in a table
brand_counts <- table(top1.brands)

#look up for what elements in the list show up more than once
brands_more_than_once <- names(brand_counts[brand_counts > 1])

brands_more_than_once
```

```
## [1] "Mama"        "MyKuali"     "Prima Taste"
```

“Mama”, “MyKuali” and “Prima Taste” have won more than once first place.

Number 5

```
#Aggregating by brands and calculating the mean of each brand
brand.stars.average <- aggregate(ramen$Stars, by = list(ramen$Brand), FUN = "mean", na.r
m=FALSE)

#maximum average
max_average_stars <- max(brand.stars.average$x, na.rm = TRUE)
print(max_average_stars)
```

```
## [1] 5
```

```
#top brands with maximum average rating
top_brand <- brand.stars.average[brand.stars.average$x == max_average_stars, ]
print(top_brand$Group.1)
```

```
##  [1] ChoripDong            Daddy                 Daifuku
##  [4] Foodmon               Higashi               Jackpot Teriyaki
##  [7] Kiki Noodle           Kimura                Komforte Chockolates
## [10] <NA>                  MyOri                 Nyor Nyar
## [13] ORee Garden           <NA>                  Patanjali
## [16] Peyang                Plats Du Chef         Prima
## [19] Prima Taste           <NA>                  Seven & I
## [22] Song Hak              Takamori              Tao Kae Noi
## [25] The Bridge            The Ramen Rater Select Torishi
## 355 Levels: 1 To 3 Noodles 7 Select 7 Select/Nissin A-One ... Zow Zow
```

There are 27 brands whose average rating is the maximum rating.

Number 6

```
#loading libraries
library(ggplot2)
library(dplyr) #lead dplyr package to use group by
```

```
##
## Attaching package: 'dplyr'
```

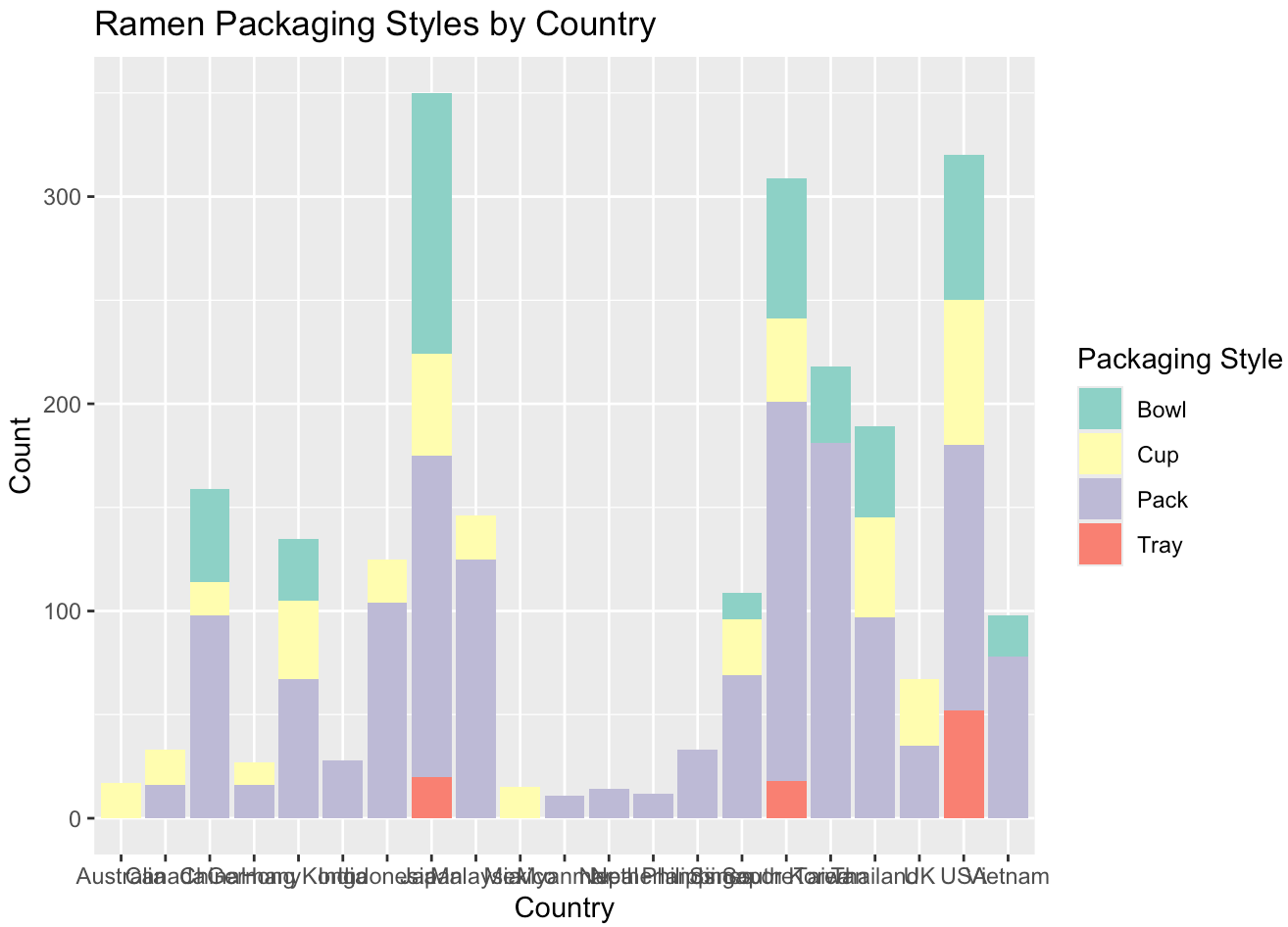
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#grouping by country and style and selecting only countries whose count is greater than
10
packaging_data <- ramen %>%
  group_by(Country, Style) %>%
  summarise(Count = n()) %>%
  filter(Count > 10)
```

```
## `summarise()` has grouped output by 'Country'. You can override using the
## `.groups` argument.
```

```
#plotting country against count filling bars by style
ggplot(packaging_data, aes(x = Country, y = Count, fill = Style)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Ramen Packaging Styles by Country",
       x = "Country",
       y = "Count",
       fill = "Packaging Style") +
  scale_fill_brewer(palette = "Set3")
```



From the graph, it looks like in countries like Myanmar, Netherlands, Philippines, India and Nepal, the only packaging style for ramen is Pack. In countries like Indonesia, Canada, South Korea, USA, Thailand, Taiwan Bowl is the most popular packaging style. Tray is used as packaging style only in few countries like South Korea, Japan and the USA. In Mexico and Australia the only packaging used is the cup. From these observations, we can conclude that packaging is related in some way to which country it is from.

Number 7

```
#number of ramen entries for country
country_ramen_count <- ramen %>%
  group_by(Country) %>%
  summarize(Ramen_Count = n()) %>%
  arrange(desc(Ramen_Count))

#selectign the country which produces the most ramen
most_ramen_country <- country_ramen_count %>%
  slice(1)

print(most_ramen_country) #Japan with 352 count
```

```
## # A tibble: 1 × 2
##   Country Ramen_Count
##   <fct>         <int>
## 1 Japan           352
```

The country that produces most ramen is Japan.

```
# Best ramen = best Stars average
# Group by Country and calculate average stars
country_average_stars <- ramen %>%
  group_by(Country) %>%
  summarize(Average_Stars = mean(Stars, na.rm = TRUE)) %>%
  arrange(desc(Average_Stars))

# Find the country with the highest average stars
best_country <- country_average_stars %>%
  slice(1)

print(best_country) #best country is Brazil
```

```
## # A tibble: 1 × 2
##   Country Average_Stars
##   <fct>           <dbl>
## 1 Brazil           4.35
```

The first way we thought about best ramen is by selecting which country had the highest average of stars from the ramen produced. It turn out Brazil with an average of 4.35 produces the “best” ramen. This answer differs from the country which produces most ramen: Japan.

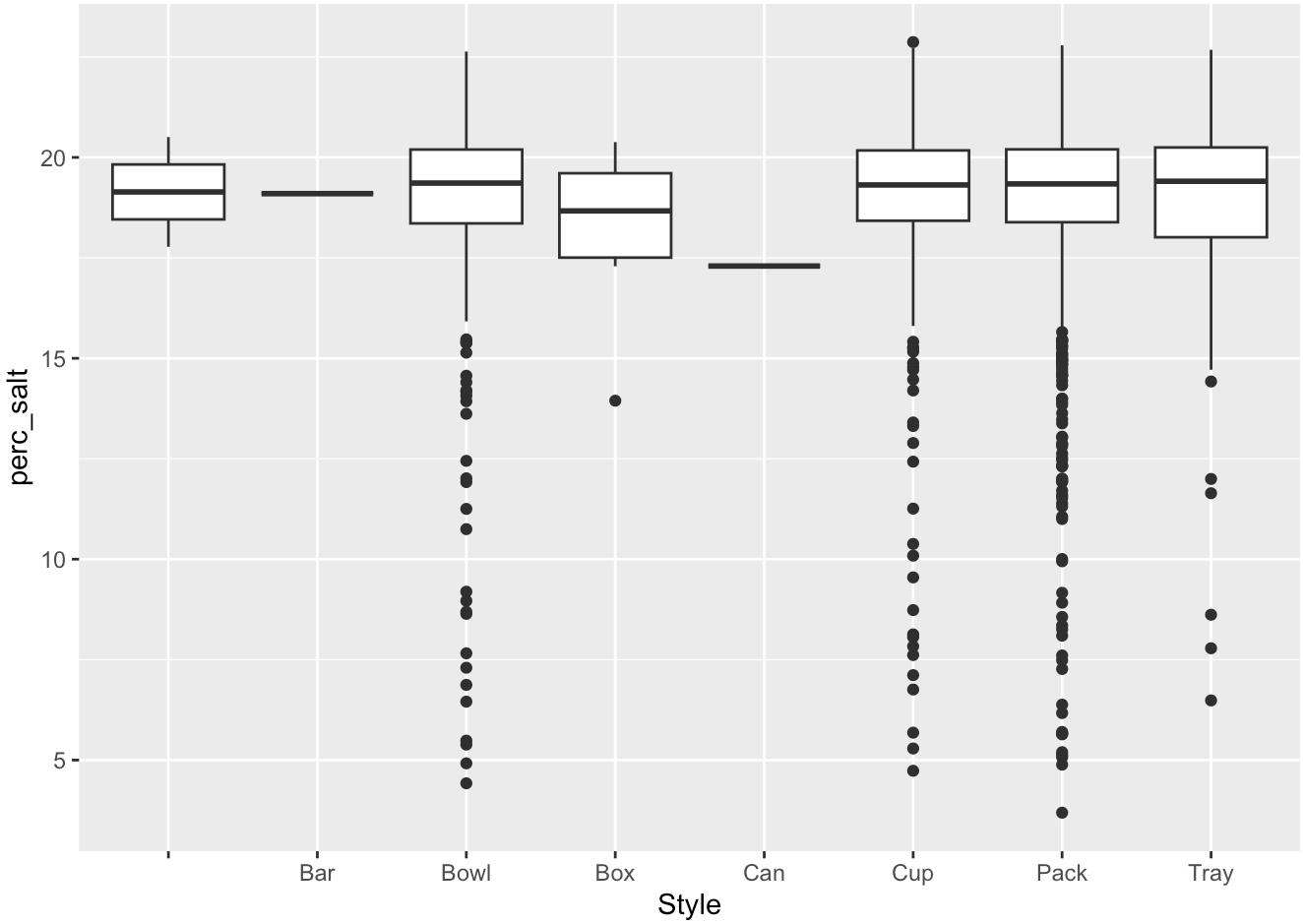
```
#Another way we thought about "best" ramen
#Best Ramen - country with most nominees in top 10
top.brand <- ramen$Brand[top.1.indeces]
top.countries <- ramen$Country[top.brand]

# countries in top 10 multiple times table with count
country_counts <- table(top.countries)
most_frequent_country <- names(which.max(country_counts))

most_frequent_country
```

```
## [1] "Japan"
```

Another way we thought about “best” ramen is by defining it as the the ramen with most nominee in the top 10. In this case, the country that makes best ramen is Japan. This answer is the same as the country which produces most ramen!

Question 8

```
library('ggplot2')
library('tidyr') #for drop_na() function

#more than 20 as count of style
saltiness <- ramen %>%
  drop_na() %>%
  group_by(Style, na.rm=TRUE)

#boxplot of Style and percent salt
ggplot(saltiness,  aes(x = Style, y = perc_salt)) + geom_boxplot()
```

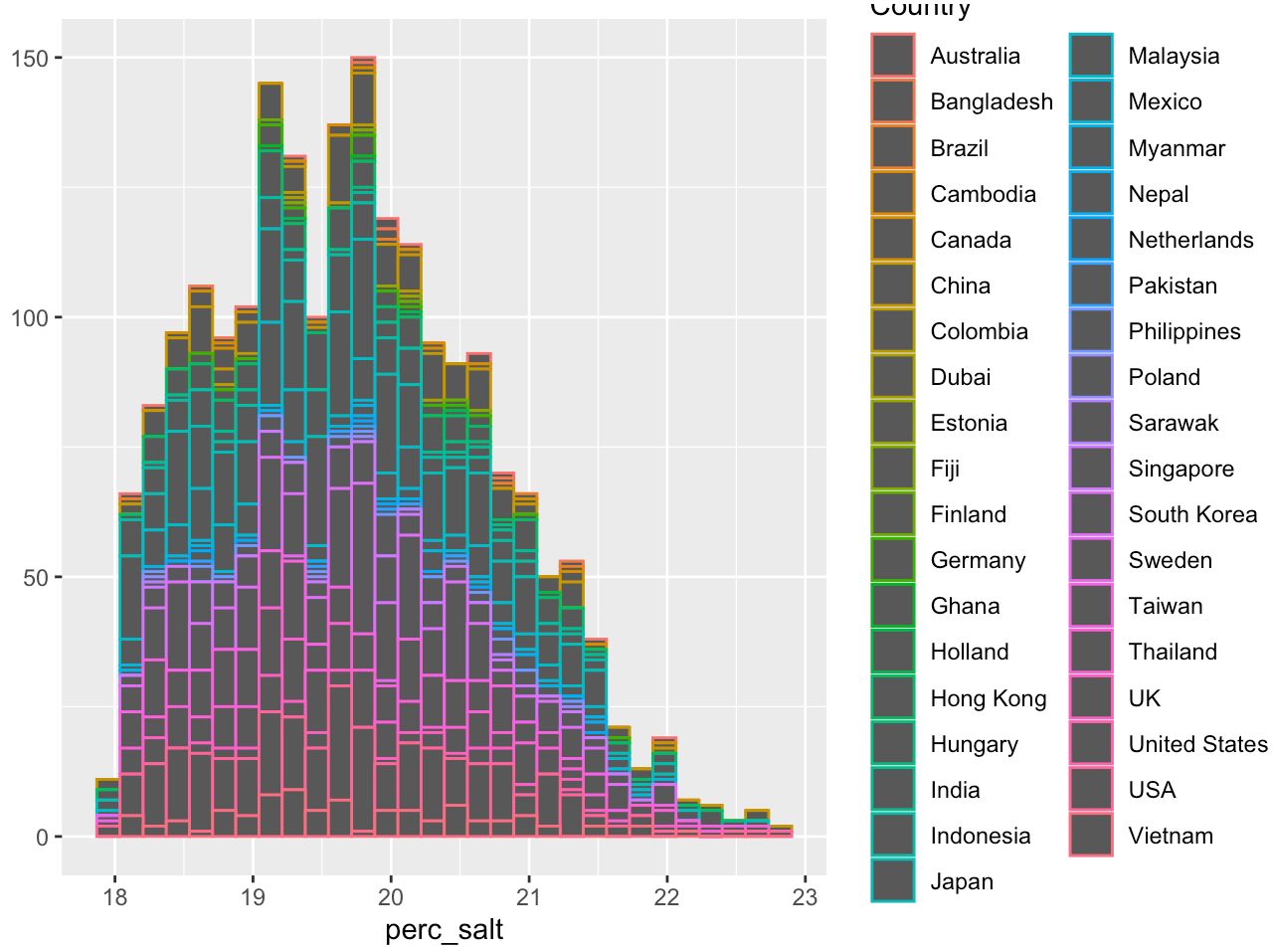


a. We can observe the distribution of percentage of salt for each packaging style. Cups, Packs, Trays and Bowls have the highest percentage of salt. Cans, Bars and Boxes have a settled percentage of salt.

```
# Filter ramen styles with more than 20 counts and perc_salt greater than 18
more_than_18 <- ramen %>%
  drop_na() %>%
  group_by(Style) %>%
  filter(n() > 20 & perc_salt > 18)

#plot percent of salt by country
qplot(x = perc_salt, data=more_than_18, colour = Country)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

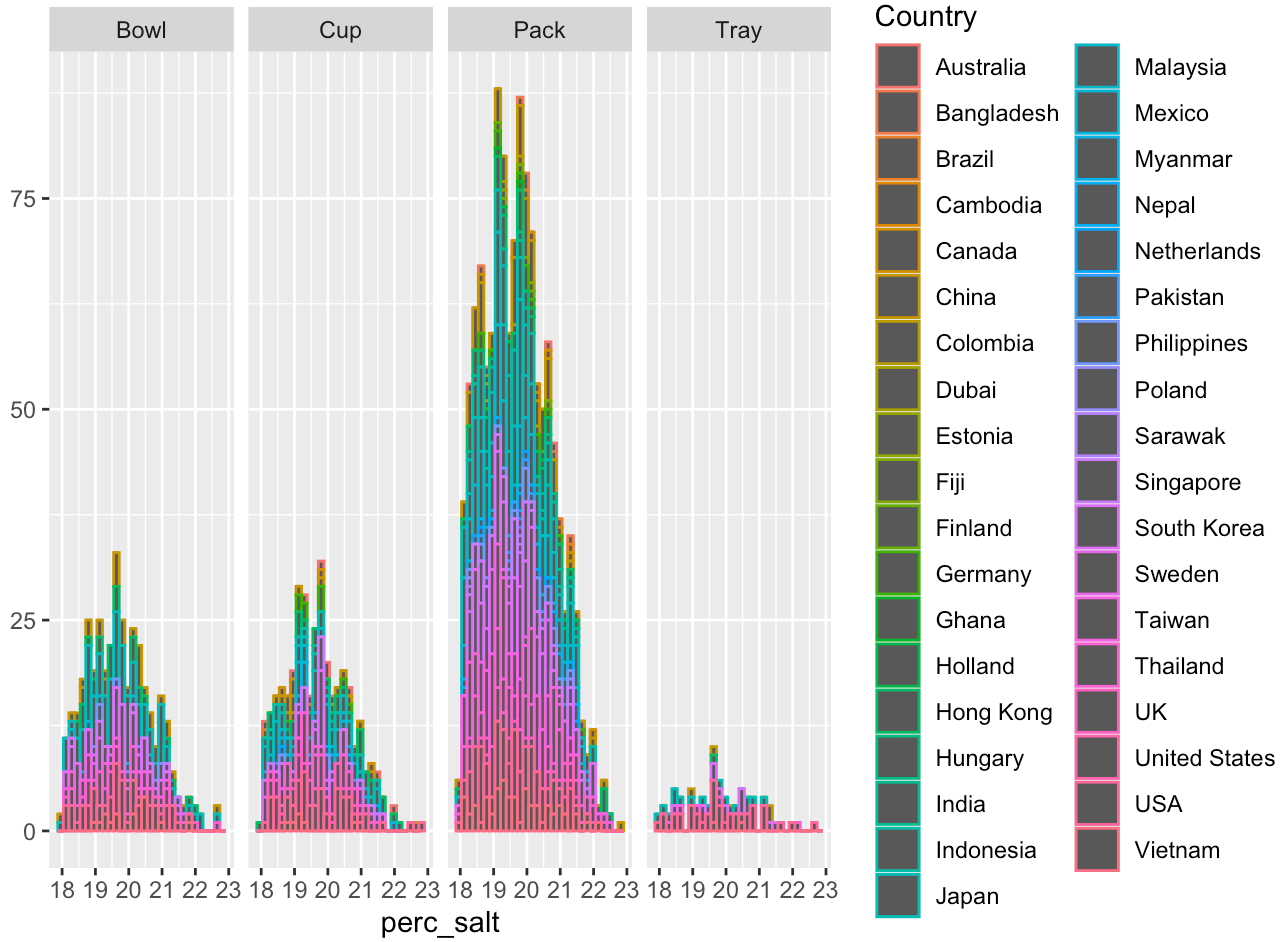
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



b. In this visualization, we can observe how the percentage of salt for each country is well distributed when looking at high percentage of salt. For the highest levels of salt, Australia, Japan, and the US seem to dominate.

```
#percent of salt by Style and country
qplot(x = perc_salt, data=more_than_18, facets = .~Style, colour = Country)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In this graph we can observe which country has the highest levels of salt based on packaging styles.

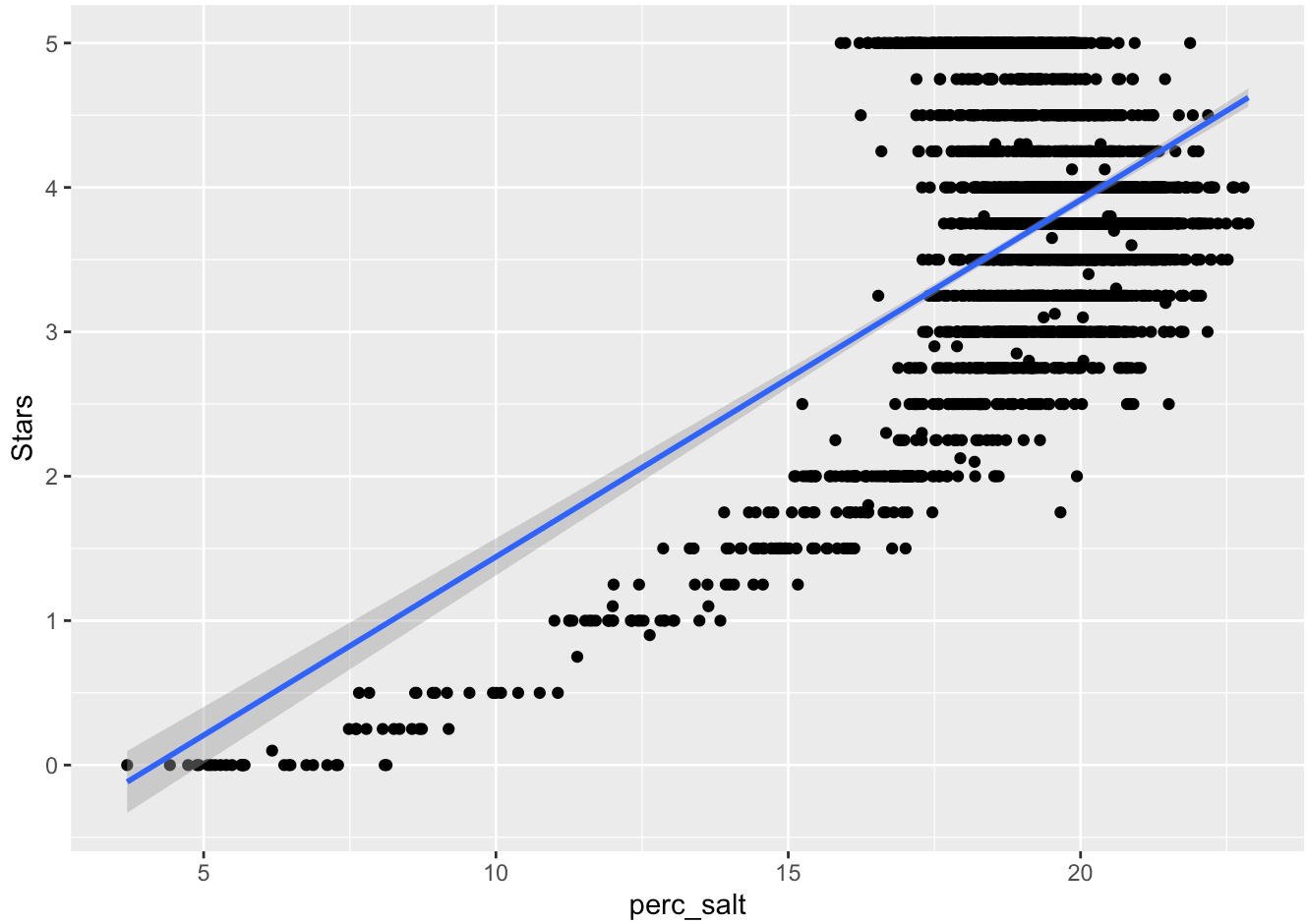
Question 9

```
#plot percent of salt agaist star rating
qplot(perc_salt, Stars, data = ramen) + geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

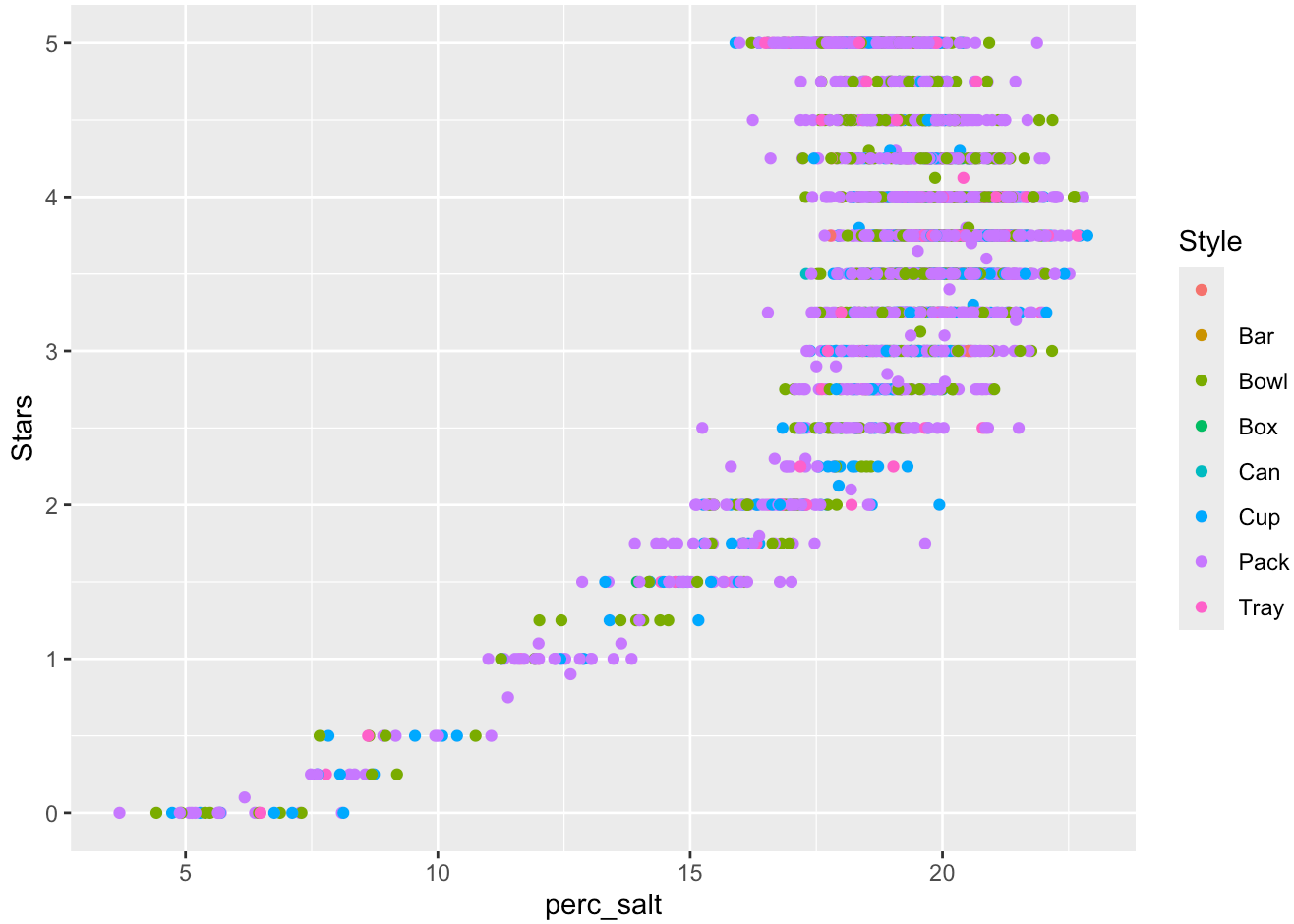


From this plot, we can notice that as the percentage of salt increases the star rating increases as well.

Question 10

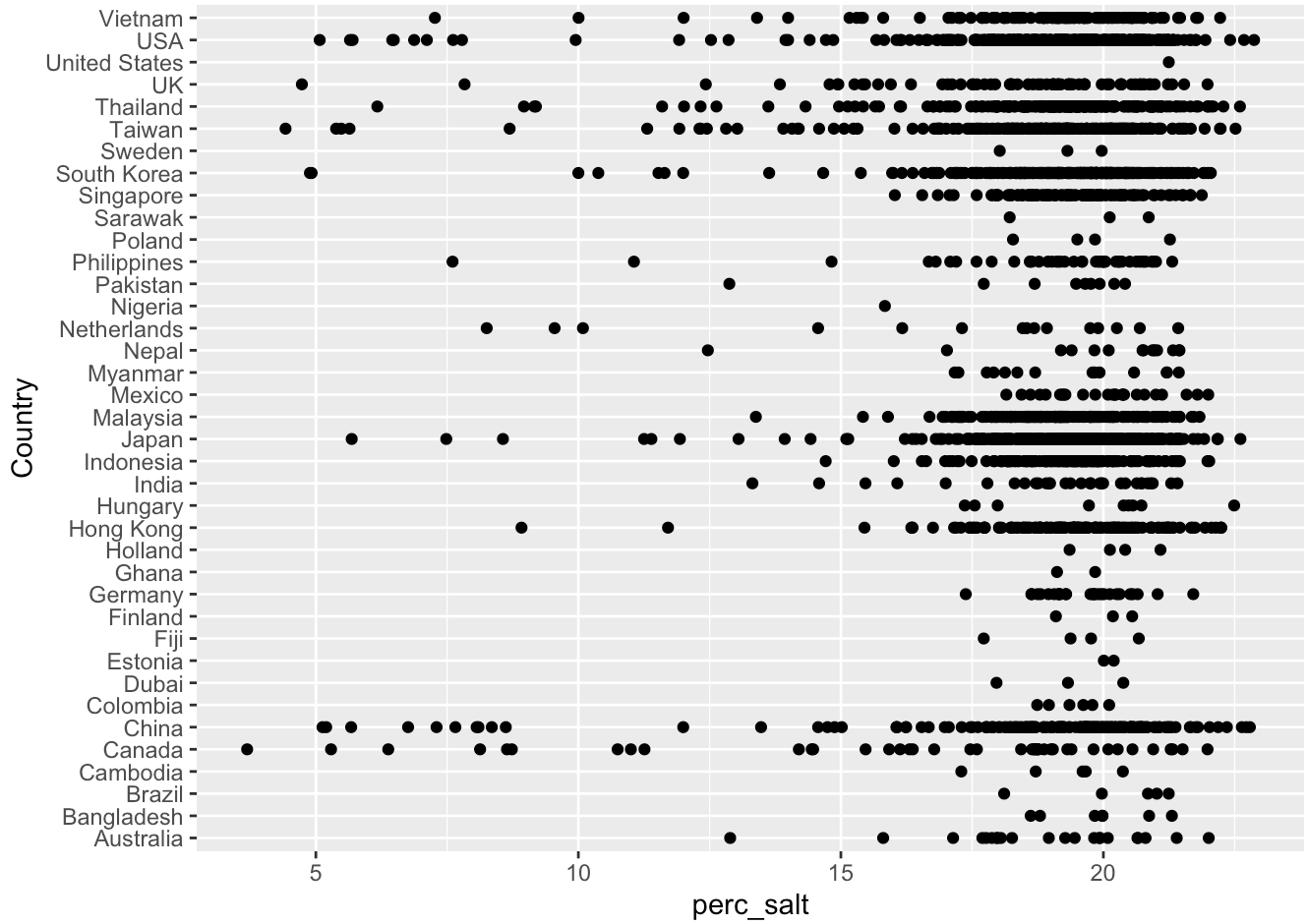
```
# Scatter plot of perc_salt vs Stars, colored by Style
qplot(perc_salt, Stars, data=ramen, color=Style)
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



From this visualization we can observe that Pack, Tray and Bowls have the highest percent of salt. Pack, Cup, and Bowl appear in the lowest percentage of salt. In general, all packaging styles seem to be equally distributed between all the percentages of saltiness.

```
#plot of percent of salt and country
qplot(perc_salt, Country, data=ramen)
```

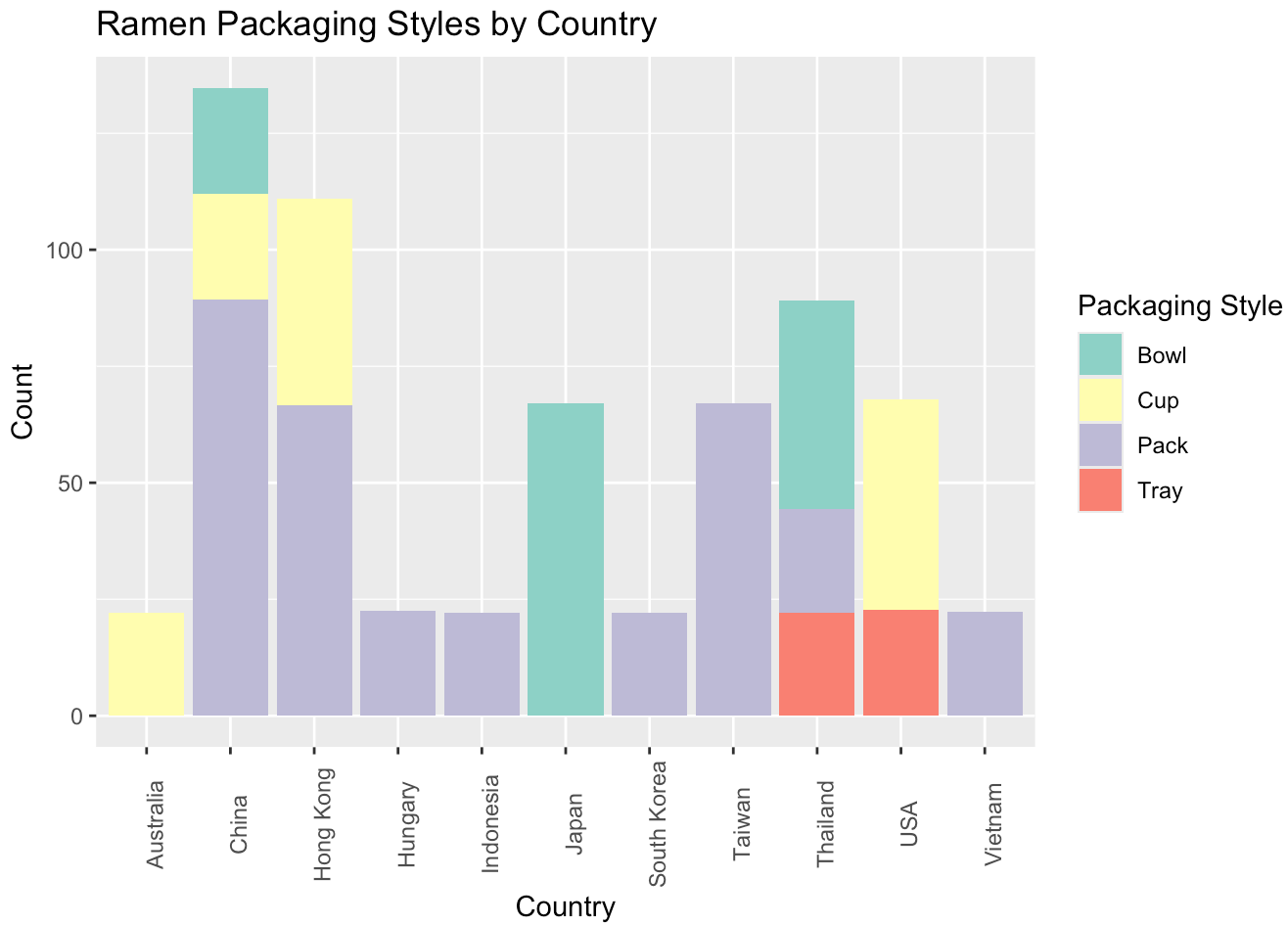


From this visualization we can observe how countries like China, Hong Kong, United States, Japan and Taiwan reach the highest of percentages of salt. UK, Canada, Colombia, Estonia, and Singapore tend to use smaller percentages of salt.

```
summary(ramen$perc_salt)
```

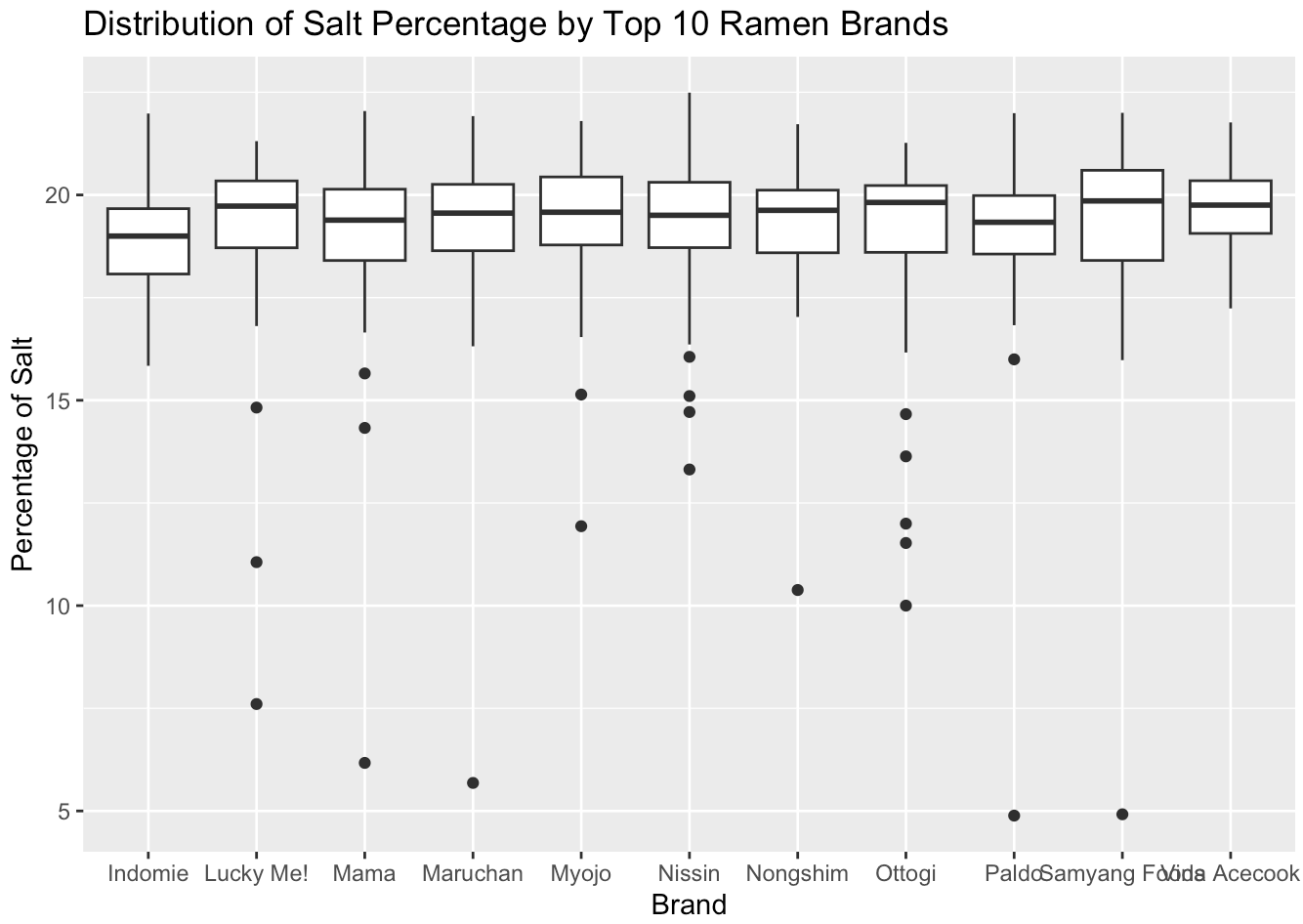
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.691  18.372  19.340  18.951  20.198  22.870
```

```
#creating a subset of more than 22 percent salt
more_than22_salt <- ramen %>%
  filter(perc_salt > 22)

#plotting countries with more than 22% of salt and adding Style
ggplot(more_than22_salt, aes(x = Country, y = perc_salt, fill = Style)) +
  geom_bar(stat = "identity", position = "stack") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Ramen Packaging Styles by Country",
       x = "Country",
       y = "Count",
       fill = "Packaging Style") +
  scale_fill_brewer(palette = "Set3")
```



We can observe which countries have the highest percentage of salt (top 10) and which packaging style are most popular in these countries. We can observe that packaging styles like Bowls, Cups and Packs have the highest percentage of salt (higher than 22 percent).

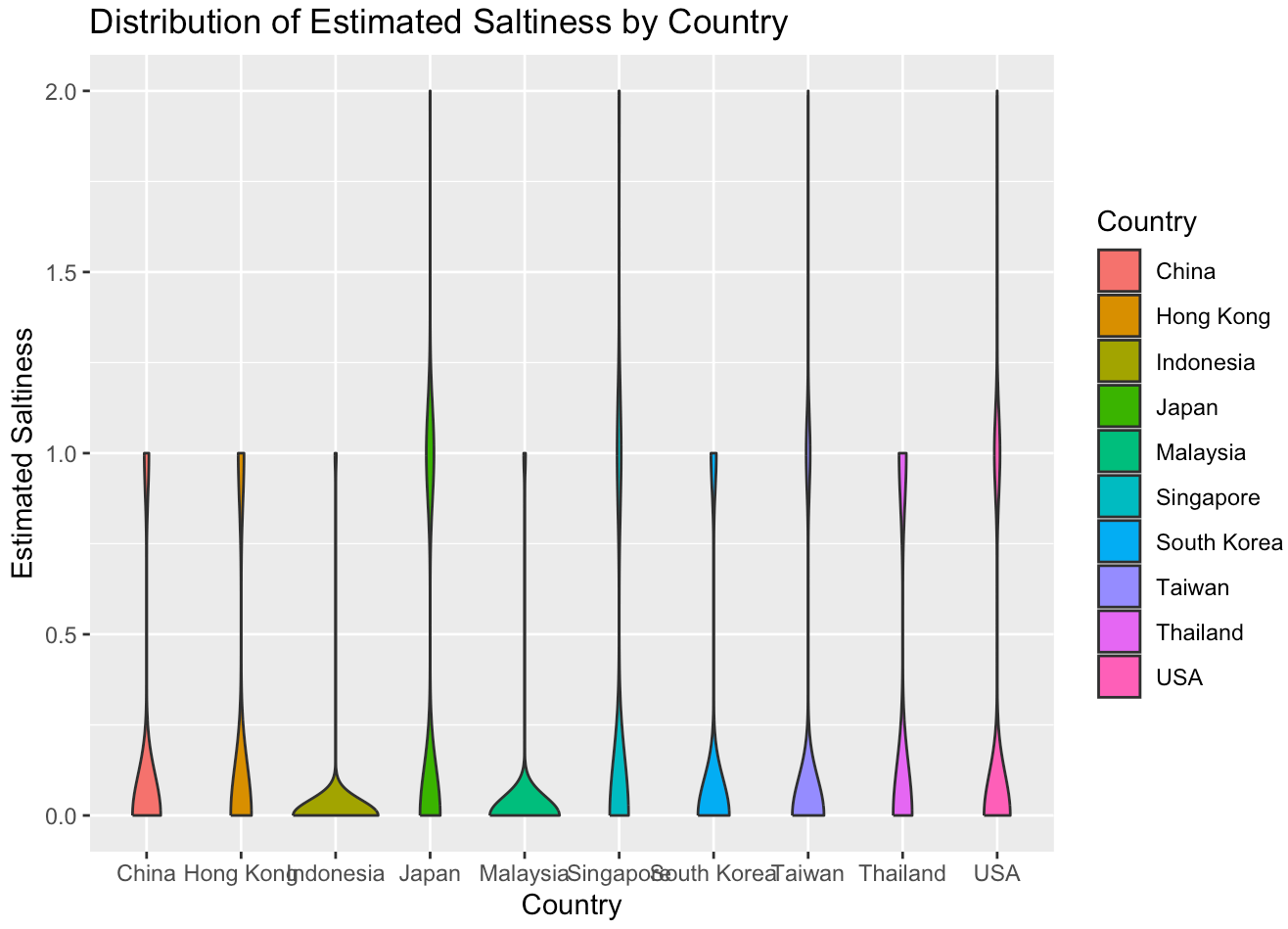
```
# Filter the top 10 brands
top_brands <- ramen %>%
  count(Brand) %>%
  top_n(10, n) %>%
  pull(Brand)

# plot percentage of salt by brand
ggplot(ramen %>% filter(Brand %in% top_brands), aes(x = Brand, y = perc_salt)) +
  geom_boxplot() +
  labs(title = "Distribution of Salt Percentage by Top 10 Ramen Brands",
       x = "Brand", y = "Percentage of Salt")
```



From this visualization we can observe that the top 10 brands (greater brand count) tend to use higher percentages of salt.

```
#loading libraries
library(tidyr)
library(stringr)
```

```
# Function to estimate saltiness based on keywords found in Variety variable
estimate_saltiness <- function(variety) {
   # Define keywords that indicate saltiness
  salt_words <- c("salt", "soy", "shoyu", "miso", "seafood", "fish", "shrimp", "crab")
  salt_score <- sum(str_count(tolower(variety), salt_words))
  return(salt_score)
}

# Add a new column for Saltiness based on the Variety
ramen <- ramen %>%
  mutate(Saltiness = sapply(Variety, estimate_saltiness))

# Get the top 10 countries with the most ramen varieties
top_countries <- ramen %>%
  count(Country) %>%
  top_n(10, n) %>%
  pull(Country)

# Create a violin plot for the distribution of saltiness in the top coumntries
ramen %>%
  filter(Country %in% top_countries) %>%
  ggplot(aes(x = Country, y = Saltiness, fill = Country)) +
  geom_violin() +
  labs(title = "Distribution of Estimated Saltiness by Country",
       x = "Country", y = "Estimated Saltiness")
```



Another way we evaluated saltiness is by looking for certain keywords present on Variety. We can observe that countries like USA, Taiwan, Singapore and Japan have the highest percentage of estimated saltiness based on Variety.

Question 11

One way to break down the ramen in 5 collection is by salt percentage levels.

```
#break down ramen into 5 collections of "similar" ramens
# Let's create 5 categories based on Stars rating and perc_salt

max(ramen$perc_salt)
```

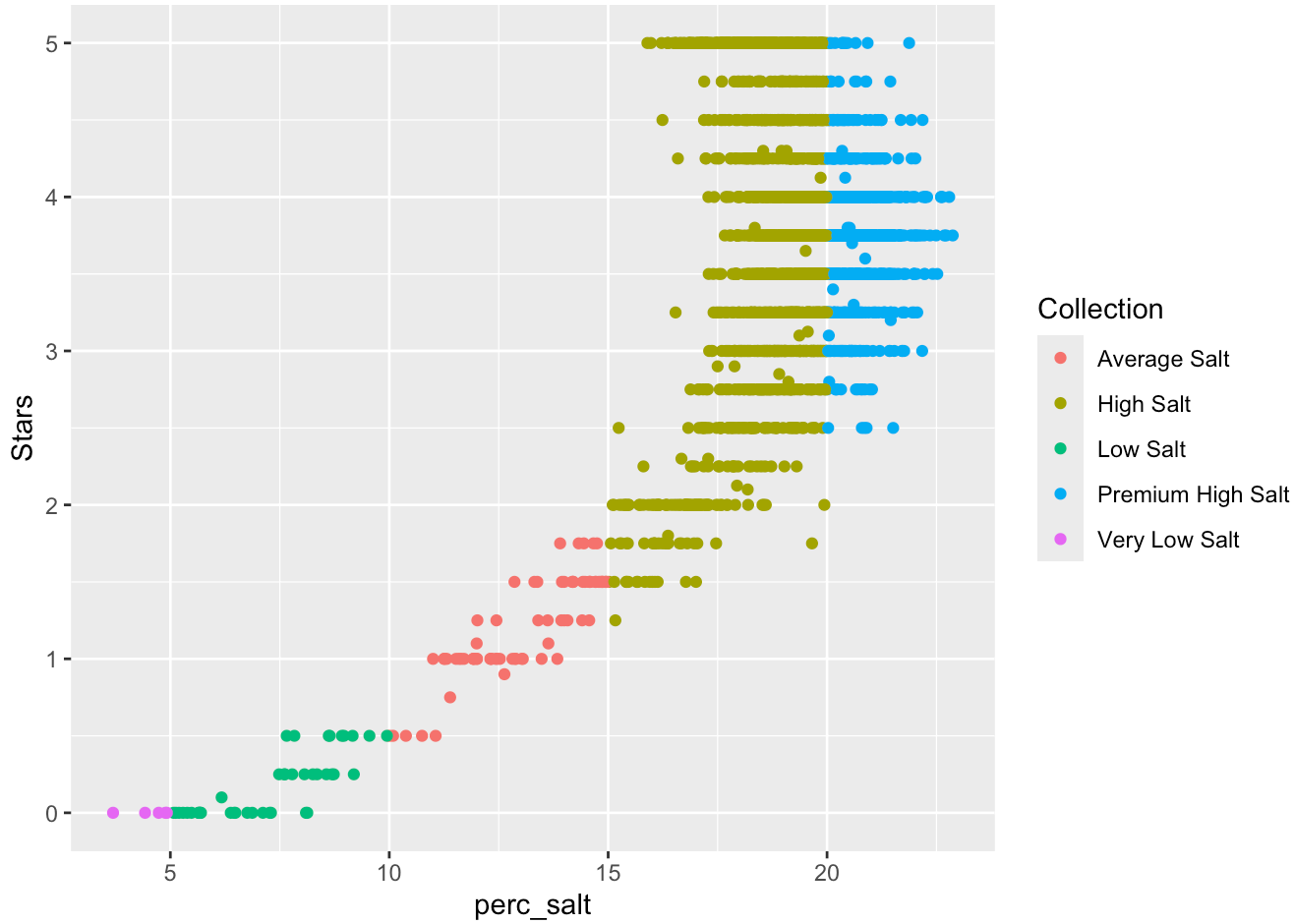
```
## [1] 22.87043
```

```
min(ramen$perc_salt)
```

```
## [1] 3.691179
```

```
ramen$Collection <- ifelse(ramen$perc_salt <= 5, "Very Low Salt",
                    ifelse(ramen$perc_salt >5 & ramen$perc_salt <=10, "Low Salt",
                    ifelse(ramen$perc_salt >10 & ramen$perc_salt <=15, "Average Salt",
                    ifelse(ramen$perc_salt >15 & ramen$perc_salt <=20, "High Salt",
                    ifelse(ramen$perc_salt > 20 & ramen$perc_salt <= 25, "Premium High S
alt",
                           "Ultra Premium High Salt")))))


category_counts <- table(ramen$Collection)

#Plotting percent of salt and stars and coloring by collection (level of saltiness)
qplot(x = perc_salt, y = Stars, data = ramen, color = Collection)
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

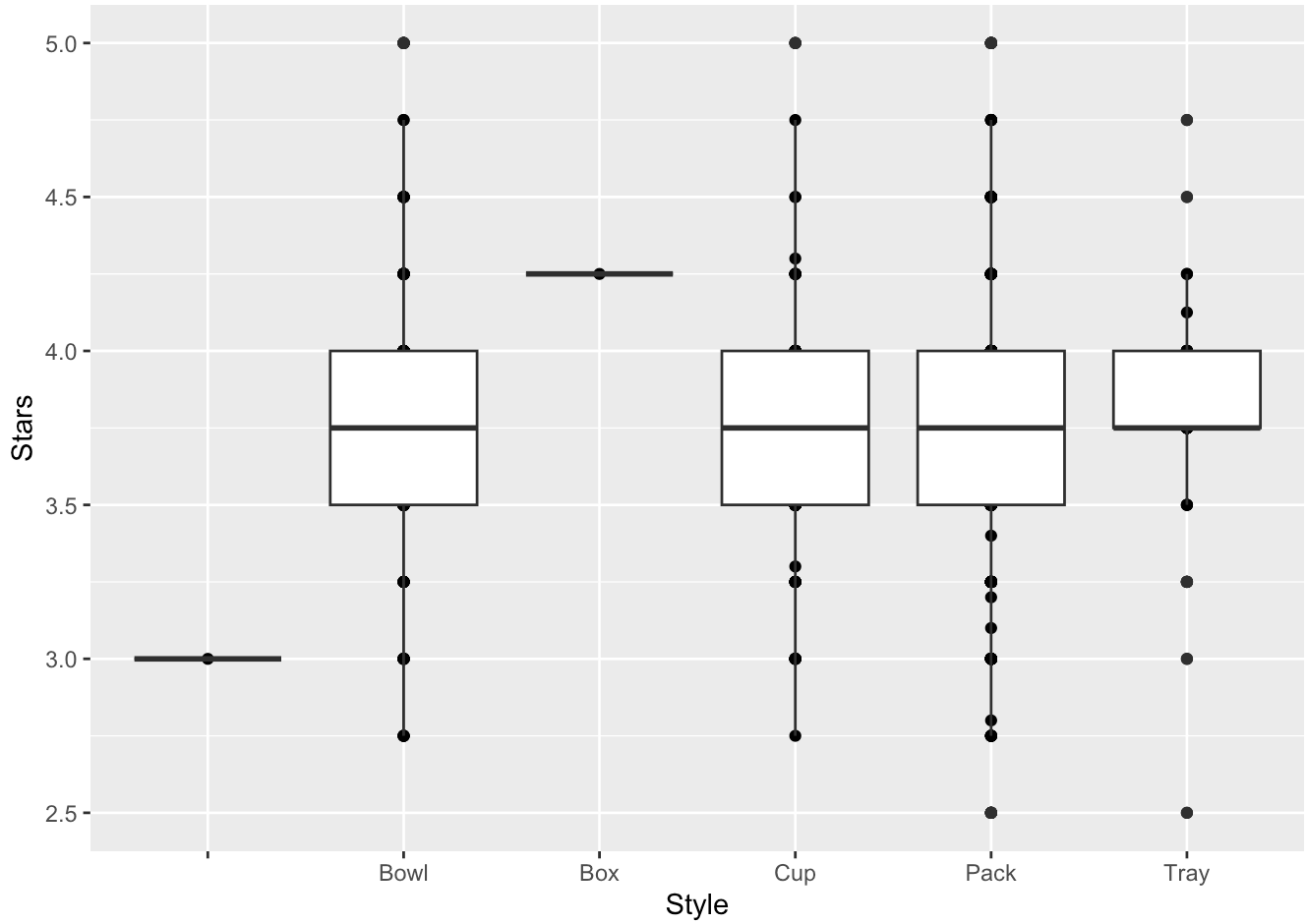


From this plot we can observe how highest levels of salt lead to higher star ratings.

```
#Analyzing only the Premium High Saltiness
premium_high_salt <- ramen %>%
  filter(Collection == "Premium High Salt")
premium_high_salt$Style <- as.factor(premium_high_salt$Style)

#boxplot of style and stars of only premium high salt
qplot(x = Style, y = Stars, data = premium_high_salt) + geom_boxplot()
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



This visualization of the premium level of salt shows us how distributed are the stars ratings based on packaging styles. As seen previously in other visualizations, bowls, cups and packs have higher percent of salt and higher star ratings.

```
#loading libraries
library('readr')
```

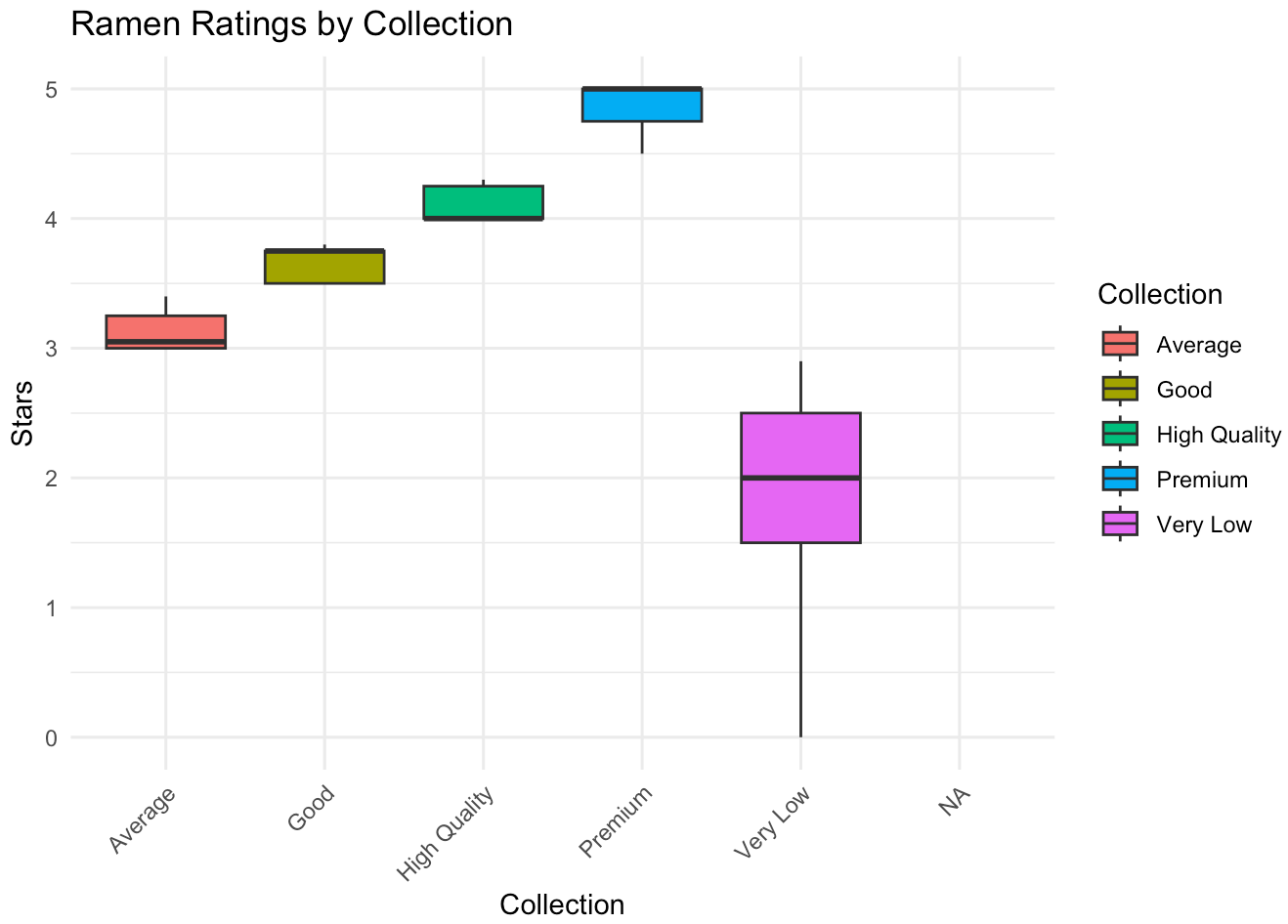
Another way to break down our data in 5 collection is based on the star rating.

```
#Another way to break down our data set in 5 categories is based on the star rating

ramen$Stars <- as.numeric(ramen$Stars)

ramen <- ramen %>%
  #break down n 5 categories based on ratings
  mutate(Collection = case_when(
    Stars >= 4.5 ~ "Premium",
    Stars >= 4 & Stars < 4.5 ~ "High Quality",
    Stars >= 3.5 & Stars < 4 ~ "Good",
    Stars >= 3 & Stars < 3.5 ~ "Average",
    Stars < 3 ~ "Very Low"
  ))

#grouping by collection (satr rating)
collection_summary <- ramen %>%
  group_by(Collection) %>%
  summarize(
    Count = n(),
    AvgStars = mean(Stars, na.rm = TRUE),
    TopCountries = paste(names(sort(table(Country), decreasing = TRUE)[1:3]), collapse =
", "),
    TopStyles = paste(names(sort(table(Style), decreasing = TRUE)[1:2]), collapse = ",
")
  )

#printing collections summary
print(collection_summary)
```

```
## # A tibble: 6 × 5
##   Collection   Count AvgStars TopCountries                    TopStyles
##   <chr>        <int>    <dbl> <chr>                           <chr>
## 1 Average        352     3.12 USA, Japan, Thailand            "Pack, Cup"
## 2 Good           691     3.63 South Korea, USA, Japan         "Pack, Cup"
## 3 High Quality   542     4.07 Japan, South Korea, USA         "Pack, Bowl"
## 4 Premium        585     4.86 Japan, Malaysia, South Korea    "Pack, Bowl"
## 5 Very Low       407     1.88 USA, Taiwan, Thailand           "Pack, Cup"
## 6 <NA>             3   NaN    South Korea, Malaysia, Australia "Pack, "
```

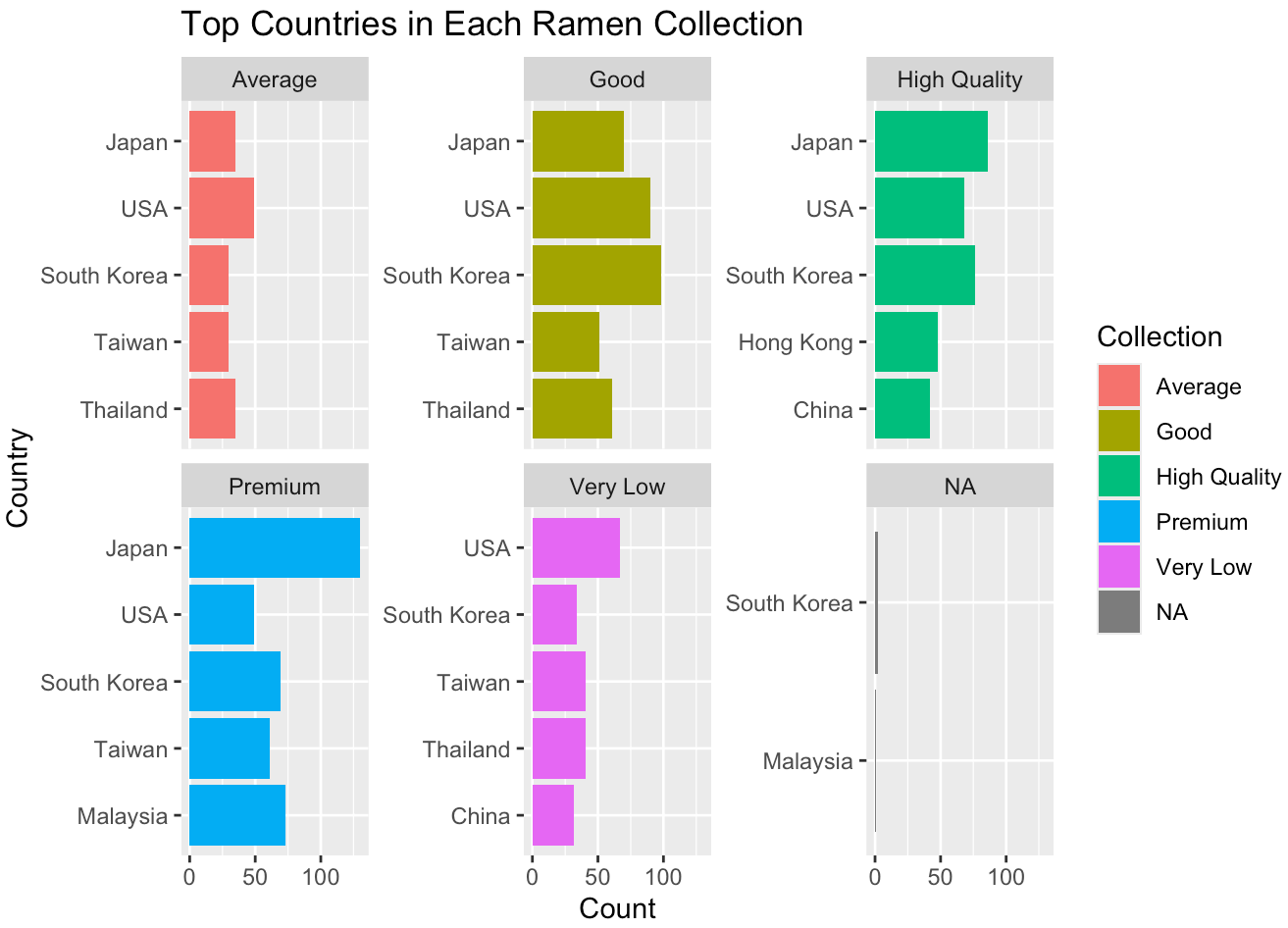
```
#plotting Collections against stars by collection
ggplot(ramen, aes(x = Collection, y = Stars, fill = Collection)) +
  geom_boxplot() +
  labs(title = "Ramen Ratings by Collection", x = "Collection", y = "Stars") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
ramen %>%
  group_by(Collection, Country) %>%
  summarize(Count = n()) %>%
  group_by(Collection) %>%
  top_n(5, Count) %>%
  ggplot(aes(x = reorder(Country, Count), y = Count, fill = Collection)) +
  geom_col() +
  facet_wrap(~Collection, scales = "free_y") +
  coord_flip() +
  labs(title = "Top Countries in Each Ramen Collection", x = "Country", y = "Count")
```

```
## `summarise()` has grouped output by 'Collection'. You can override using the
## `.groups` argument.
```



From the above visualization, we can observe that Japan has the highest best ratings. USA has the most average and very low ratings. Furthermore, USA ramen's ratings show up in all ratings categories Very Low, Average, Good, High Quality and Premium.

Question 12

```
#define a threshold
salt_threshold <- mean(ramen$perc_salt, na.rm = TRUE)
salt_threshold <- 20
#filter higher than salt_threshold and equal to 5 stars
top_stars_high_salt <- ramen %>%
  filter(Stars == 5 & perc_salt > salt_threshold)

#top Variety with higher concentration of salt and stars rating
unique_varieties <- unique(top_stars_high_salt$Variety)
print(unique_varieties)
```

```
##  [1] Viet Cuisine Bun Rieu Cua Sour Crab Soup Instant Rice Vermicelli
##  [2] Penang White Curry Instant Noodle
##  [3] Rice Noodle Seafood Flavour
##  [4] Tokyo Tokunou Gyokai Tonkotsu
##  [5] Chikin Ramen Donburi
##  [6] Curry Udon
##  [7] Cup Noodle Big Cheese Mexican Chilli
##  [8] Singapore Fish Soup La Mian
##  [9] Taste Of Malaysia Penang Hokkien Mee Ramen
## [10] Chow Mein Japanese Style Noodles Yakisoba
## [11] Singapore Chilli Crab La Mian
## [12] Dry Noodle Mandarin Noodle - Onion Oil Sauce
## 2413 Levels: "A" Series Artificial Chicken ... 三養라면 (Samyang Ramyun) (South Korean
Version)
```

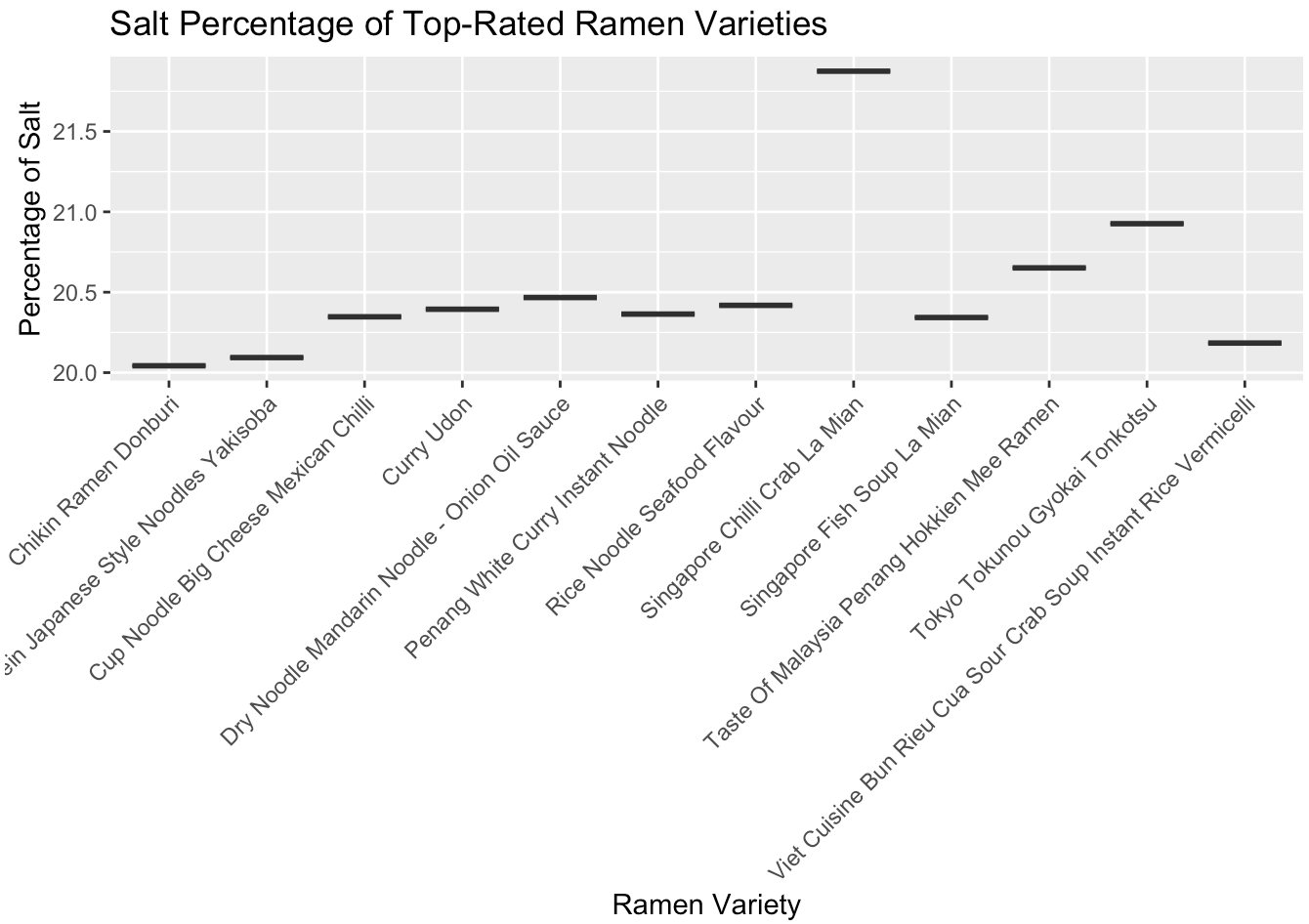
As observed in visualization #9, higher level of saltiness will lead to higher star ratings. Here we combined stars rating = 5 to greater than 20 percent level of saltiness and we got 12 best performing ramen Variety.

```
#summary statistics
summary_stats <- top_stars_high_salt %>%
  summarize(
    mean_salt = mean(perc_salt, na.rm = TRUE),
    min_salt = min(perc_salt, na.rm = TRUE),
    max_salt = max(perc_salt, na.rm = TRUE),
    count = n()
  )


print(summary_stats)
```

```
##   mean_salt min_salt max_salt count
## 1  20.50882 20.04261 21.87434    12
```

This is a summary statistics for top stars ramen with higher tnan 20 percentage of salt. Average percentage of salt is 20.50882.

```
# Create a plot of salt percentage for the top varieties
ggplot(top_stars_high_salt, aes(x = Variety, y = perc_salt)) +
  geom_boxplot() +
  labs(title = "Salt Percentage of Top-Rated Ramen Varieties",
       x = "Ramen Variety",
       y = "Percentage of Salt") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



These are the top ramen varieties and their respective percentage of salt. As we observed in previous visualizations, higher salt percentages tend to lead to higher ratings and overall better performance. These 12 varieties have high star ratings (5) and a high level of saltiness, which, as we observed earlier, correlates with better ratings. Therefore, the next best variety should likely fall within the range of these 12 varieties and have a salt percentage around 20.51%.