R Notebook

Code

iiData

cache me out side. how 'bout dat data

Hide

```
library(dplyr)
library(ggplot2)
library(randomForest)
data <- read.csv("ffclean6.csv")
data = data %>%
  mutate(mydate = as.Date(substr(last_modified_datetime, 1, 10)))
```

Question 1: France's score of it's own country

This is the only problem done in SQL

It seems that France gives french manufacters a higher than average score for all nutrition.

Conversly, it seems that the UK gives a lower nutrition score on average, and the countries that have France as a point of manufacturing score lower than average.

avg score of all countries = 7.98163664965167 31149 rows returned in 95ms from:

```
SELECT (CAST (nutrition_score_fr_100g  as integer)), countries_en
FROM ffclean6;
```

avg_score where country has france in it 8.06925675675676:

```
SELECT AVG( CAST (nutrition_score_fr_100g  as integer))
  FROM ffclean6
  WHERE countries_en LIKE '%France%' ;
```

avg_score where country does not contain france in it = 7.70384254920337

```
SELECT AVG(CAST (nutrition_score_fr_100g  as integer)), countries_en
FROM ffclean6
WHERE countries_en NOT LIKE '%France%' ;
```

avg of all countries based on UK scores7.71909210568558:

```
SELECT AVG(CAST (nutrition_score_uk_100g  as integer)), countries_en
FROM ffclean6;
```

avg of France based on UK scores7.75865709459459:

```
SELECT AVG( CAST (nutrition_score_uk_100g  as integer))
FROMG ffclean6
WHERE countries_en LIKE '%France%'
```

avg of NOT France based on UK stores 7.5936537689115:

```
SELECT AVG( CAST (nutrition_score_uk_100g  as integer))
FROM ffclean6
WHERE countries_en NOT LIKE '%France%'
```

Question 2: Predicting categorization of Sugary snack vs Dairies vs Fresh Food vs meats

Hide

```
q2data = data %>%
  filter(main_category_en == "Sugary snacks" |
         main_category_en == "Dairies" |
         main_category_en == "Fresh foods" |
         main_category_en == "Meats") %>%
  select(additives_n,
         ingredients_that_may_be_from_palm_oil_n,
         energy_100g,
         fat_100g,
         saturated_fat_100g,
         carbohydrates_100g,
         sugars_100g,
         fiber_100g,
         proteins_100g,
         salt_100g,
         sodium_100g,
         main_category_en,
         mydate)

q2data = q2data[complete.cases(q2data),]
q2data$main_category_en <- factor(q2data$main_category_en)
set.seed(420)

q2train = q2data %>%
  sample_frac(.9, replace = TRUE)
q2test = q2data %>%
  sample_frac(.1, replace = TRUE)
```

Hide

```
fit2 <- randomForest(as.factor(main_category_en) ~ additives_n +
                       ingredients_that_may_be_from_palm_oil_n +
                       energy_100g +
                       fat_100g +
                       saturated_fat_100g +
                       carbohydrates_100g +
                       sugars_100g +
                       fiber_100g +
                       proteins_100g +
                       salt_100g +
                       sodium_100g,
                     data = q2train,
                     importance=TRUE,
                     ntree= 2000)


predict2 <- predict(fit2, q2test)
conf2 = fit2$confusion
```

Answer to 2

Hide

```
conf2
```

Hide

```
q2data %>%
  ggplot(aes(x=mydate))+
    geom_histogram(aes(y = ..density..), bins = length(unique(q2data$mydate))) +
    geom_density()
```

Hide

```
q2data %>%
  ggplot(aes(mydate, main_category_en, color = main_category_en))+
   geom_point( alpha = .2)
```

Question 3: Predicting categorization of United States vs United Kingdom vs Germany vs Spain

Hide

```
q3data = data %>%
  filter(countries_en == "United States" |
         countries_en == "United Kingdom" |
         countries_en == "Germany" |
         countries_en == "Spain") %>%
  select(additives_n,
         ingredients_that_may_be_from_palm_oil_n,
         energy_100g,
         fat_100g,
         saturated_fat_100g,
         carbohydrates_100g,
         sugars_100g,
         fiber_100g,
         proteins_100g,
         salt_100g,
         sodium_100g,
         countries_en,
         mydate)

q3data = q3data[complete.cases(q3data),]
q3data$countries_eny_en <- factor(q3data$countries_en)
set.seed(420)

q3train = q3data %>%
  sample_frac(.9, replace = TRUE)
q3train$countries_en <- factor(q3train$countries_en)

q3test = q3data %>%
  sample_frac(.1, replace = TRUE)
q3test$countries_en <- factor(q3test$countries_en)
```

Hide

```
fit3 <- randomForest(as.factor(countries_en) ~ additives_n +
                       ingredients_that_may_be_from_palm_oil_n +
                       energy_100g +
                       fat_100g +
                       saturated_fat_100g +
                       carbohydrates_100g +
                       sugars_100g +
                       fiber_100g +
                       proteins_100g +
                       salt_100g +
                       sodium_100g,
                     data = q3train,
                     importance=TRUE,
                     ntree= 2000)


predict3 <- predict(fit3, q3test)
conf3 = fit3$confusion
```

Answer to 3

Hide

```
conf3
```

Hide

```
q3data %>%
  ggplot(aes(x=mydate))+
    geom_histogram(aes(y = ..density..), bins = length(unique(q3data$mydate))) +
    geom_density()
```

Hide

```
q3data %>%
  ggplot(aes(mydate, countries_en, color = countries_en))+
   geom_point( alpha = .2)
```

Question 5: Predicting categorization of United States vs United Kingdom vs Germany vs Spain

Hide

```
q5data = data %>%
  filter(pnns_groups_1 == "Fish Meat Eggs") %>%
  select(additives_n,
         ingredients_that_may_be_from_palm_oil_n,
         energy_100g,
         fat_100g,
         saturated_fat_100g,
         carbohydrates_100g,
         sugars_100g,
         fiber_100g,
         proteins_100g,
         salt_100g,
         sodium_100g,
         pnns_groups_2,
         mydate)

q5data = q5data[complete.cases(q5data),]
q5data$pnns_groups_2 <- factor(q5data$pnns_groups_2)
set.seed(420)

q5train = q5data %>%
  sample_frac(.9, replace = TRUE)
q5train$pnns_groups_2 <- factor(q5train$pnns_groups_2)

q5test = q5data %>%
  sample_frac(.1, replace = TRUE)
q5test$pnns_groups_2 <- factor(q5test$pnns_groups_2)
```

Hide

```
fit5 <- randomForest(as.factor(pnns_groups_2) ~ additives_n +
                       ingredients_that_may_be_from_palm_oil_n +
                       energy_100g +
                       fat_100g +
                       saturated_fat_100g +
                       carbohydrates_100g +
                       sugars_100g +
                       fiber_100g +
                       proteins_100g +
                       salt_100g +
                       sodium_100g,
                     data = q5train,
                     importance=TRUE,
                     ntree= 2000)


predict5 <- predict(fit5, q5test)
conf5 = fit5$confusion
```

Answer to #5

Hide

```
conf5
```

Hide

```
q5data %>%
  ggplot(aes(x=mydate))+
    geom_histogram(aes(y = ..density..), bins = length(unique(q5data$mydate))) +
    geom_density()
```

Hide

```
q5data %>%
  ggplot(aes(mydate, pnns_groups_2, color = pnns_groups_2))+
   geom_point( alpha = .2)
```