Machine Learning - Day 1

Exploratory analysis

Mariachiara Fortuna

Libraries

```
library(dplyr)
library(ggplot2)
```

DATA IMPORT

```
dc <-  file.path("data", "dc-wikia-data.csv") %>%
  read.csv(na.strings = "")
```

DATA EXPLORATION

Dataset structure

```
str(dc)
```

Summary information

```
summary(dc)
```

Simple stats

Range

```
min(dc$YEAR)

min(dc$YEAR, na.rm = T)

max(dc$YEAR, na.rm = T)

range(dc$YEAR, na.rm = T)
```

Mean and Sd

```r
mean(dc$APPEARANCES, na.rm = T)

sd(dc$APPEARANCES, na.rm = T)
```

Frequency tables

Univariate

```r
table(dc$EYE)
```

Bivariate

```r
ID_SEX_freq  <- table(dc$ID, dc$SEX)

ID_SEX_freq
```

```r
margin.table(ID_SEX_freq, margin = 1)
```

```r
prop.table(ID_SEX_freq)
```

```r
prop.table(ID_SEX_freq, margin = 2)
```

DATA MANIPULATION

Create a new column: Mutate

```r
max_year <- max(dc$YEAR, na.rm =T)

dc <-  dc %>%
  mutate(active_years = max_year - YEAR) %>%
  arrange(desc(active_years))
```

Recode factor

```r
dc <- dc %>%
  mutate(
    sex = case_when(SEX == "Female Characters" ~ "F",
                    SEX == "Male Characters" ~ "M",
                    TRUE ~ "other")
    )
```

Subsetting data (rows and columns)

```r
the_bad <- dc %>%
  filter(ALIGN == "Bad Characters") %>%
  select(name, ID, SEX, ALIVE, YEAR, APPEARANCES, FIRST.APPEARANCE, active_years)
```

More complex summary stats

```
the_bad %>%
  group_by(SEX) %>%
  summarize(n = n(),
            avg_appearance = mean(APPEARANCES, na.rm = T),
            avg_active_year = mean(active_years, na.rm = T))
```

Percentage

```
dc %>%
  group_by(SEX, ALIVE) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n))
```

DATA VISUALIZATION

Barplot

```
ggplot(data=dc, aes(x = SEX)) +
  geom_bar()
```

Barplot with grouping variable

```
ggplot(data=dc, aes(x = ALIVE, fill = SEX)) +
  geom_bar()
```

```
ggplot(data=dc, aes(x = YEAR, fill = SEX)) +
  geom_bar()
```

Percent values

```
YEAR_SEX_tab <- dc %>%
  group_by(YEAR, SEX) %>%
  summarize(n = n())

margin <- dc %>%
  group_by(YEAR) %>%
  summarize(tot = n())

perc_ys <- YEAR_SEX_tab %>%
  inner_join(margin) %>%
  mutate(perc = (n/tot)*100)


ggplot(data=perc_ys, aes(x = YEAR, y = perc, fill = SEX)) +
  geom_bar(stat = "identity")
```

Pie chart

```
eye_tab <- dc %>%
  filter(EYE != "") %>%
  group_by(EYE) %>%
  summarize(count = n())

 ggplot(data=eye_tab,
       aes(x=factor(1), y = count, fill = factor(EYE))) +
   geom_bar(width = 1, stat = "identity") +
   coord_polar(theta = "y")
```

Matrix plot

```
ggplot(data = dc, aes (x = EYE, y = HAIR)) +
  geom_count() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Histogram

```
ggplot(data = dc, aes(x = YEAR)) +
  geom_histogram()
```

```
ggplot(data = dc, aes(x = YEAR)) +
  geom_density()
```

Line plot

```
year_sex <- dc %>%
  group_by(YEAR, sex) %>%
  summarise(n = n())

ggplot(data = year_sex, aes(x = YEAR, y = n, color = sex)) +
  geom_line()
```

Scatterplot

```
ggplot(data = dc %>% filter(APPEARANCES<500),
       aes(x = active_years, y = APPEARANCES)) +
  geom_point()
```

```
ggplot(data = dc %>% filter(APPEARANCES<500),
       aes(x = YEAR, y = log(APPEARANCES), color = ALIVE)) +
  geom_point()
```

```
ggplot(data = dc %>% filter(APPEARANCES<500 & APPEARANCES>20),
       aes(x = YEAR, y = APPEARANCES, color = SEX)) +
  geom_point()
```

Carpet plot

```
sex_alive <- dc %>%
  group_by(SEX, ALIVE) %>%
  summarize(avg_app = mean(APPEARANCES, na.rm = T))

ggplot(data = sex_alive, aes(x = SEX, y = ALIVE)) +
  geom_tile(aes(fill = avg_app)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```