SQUAD Datathon Starter R Markdown

SQUAD

03/04/2024

This is a starter R Markdown to get you going! Note that you don't need to code in R-markdown (even though it is a good tool for reporting). You can use R scripts. The code shown here is basic and it is expected that you create visualizations much more beautiful than these boring ones :(
The starter notebook only contains some examples for Exploratory Data Analysis (EDA). To know more about modelling, please look into the R documentation.

Setup

```
# Load the required packages using pacman
# Tidyverse is used to play with tibbles in R
# ggplot2 is used for plotting which is available in tidyverse
pacman::p_load(tidyverse, dplyr)
```

1. Loading the data

"<-" is the assignment operator for a variable in R. You can still use "=" since R accepts both (weird if you ask me).

```
filename <- paste0("./data/train.csv") # assign filename
# make sure you use the absolute path

# Read in the data
df <- read_csv(filename) # read csv file
head(df, 10) # display the first 10 lines of the data
```

```
## # A tibble: 10 x 23
##    sex      age address family_size    parents_together mother_job father_job
##    <chr>  <dbl> <chr>   <chr>          <chr>            <chr>      <chr>
##  1 Female    18 Urban   Greater than 3 Apart            at_home    teacher
##  2 Male      16 Urban   Greater than 3 Together         health     other
##  3 Male      15 Urban   Greater than 3 Together         other      teacher
##  4 Male      16 Urban   Less than 3    Together         other      other
##  5 Female    17 Urban   Less than 3    Together         services   services
##  6 Female    17 Urban   Less than 3    Together         at_home    at_home
##  7 Female    18 Urban   Less than 3    Together         services   services
##  8 Female    15 Urban   Less than 3    Together         services   services
##  9 Female    15 Rural   Less than 3    Together         other      services
## 10 Female    19 Urban   Less than 3    Together         services   other
## # i 16 more variables: guardian <chr>, travel_time <dbl>, study_time <dbl>,
## #   failed_classes <dbl>, school_support <chr>, extra_curricular <chr>,
## #   want_higher <chr>, internet <chr>, romantic_rel <chr>, family_rel <dbl>,
```

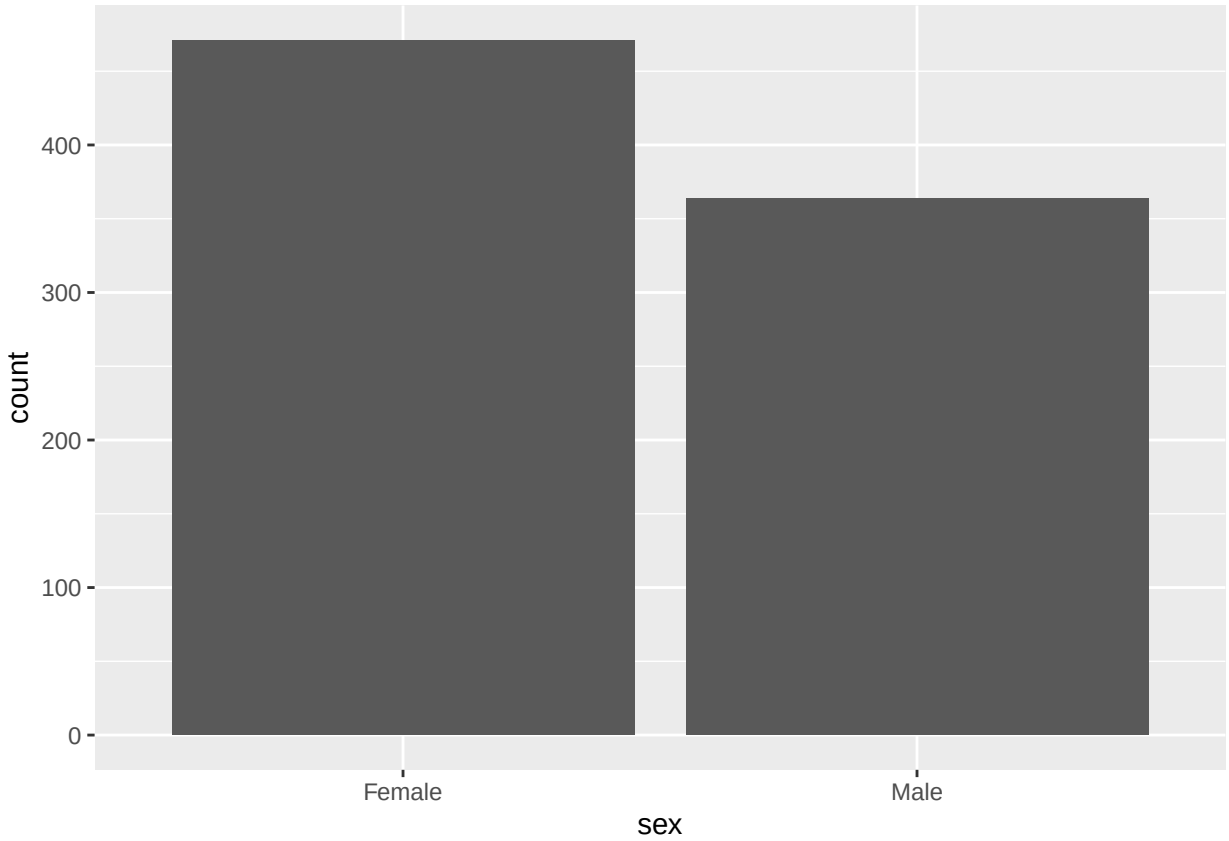
```
## #   free_time <dbl>, go_out <dbl>, workday_alcohol <dbl>,
## #   weekend_alcohol <dbl>, absences <dbl>, grade <dbl>
```

You can see data types for each feature when printing the head or the tibble. Tip: In R, you still need to factor them into relevant data types before modelling.

2. EDA

Countplot

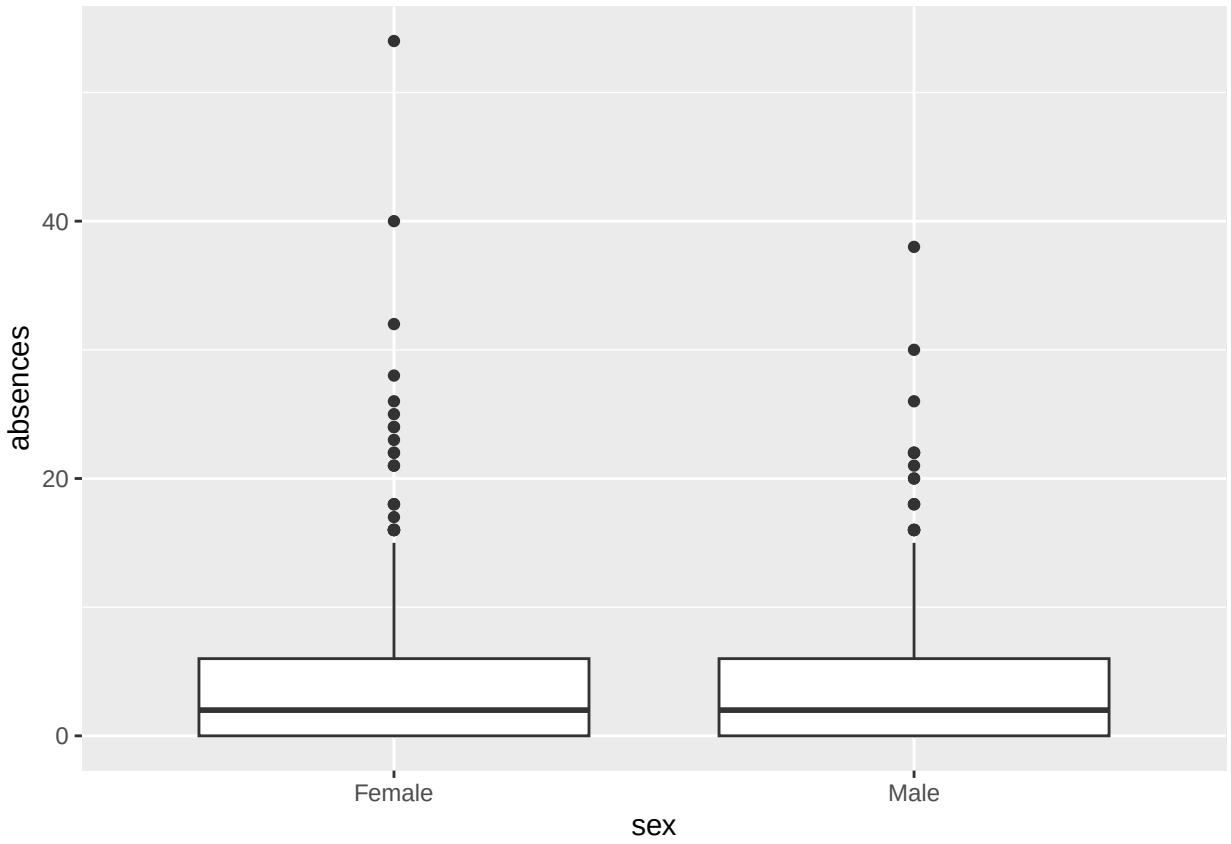
```
ggplot(df, aes(x = sex)) + # select data and variable
  geom_bar() # do a bar chart
```



Great! We have a basic count plot (or a bar chart) for our categorical variable. It is suggested that you go through the ggplot2 documentation since tweaking some parameters above can help you accommodate another categorical variable as additional information in this chart!

Boxplot

```
ggplot(df, aes(x=sex, y=absences)) + # select data and variables
  geom_boxplot() # do a boxplot
```



Boxplots contain important information that can be useful for data cleaning. These two plots should give you a lot of insights, but to win, it is recommended that you again go through the ggplot2 documentation to try different kinds of plots.

3. Filtering data

In R, you use a pipe (%>%) to perform operations on a tibble.

```
# Filter data using the filter() method provided by dplyr
df %>% filter(sex == "Male")
```

```
## # A tibble: 364 x 23
##    sex     age address family_size    parents_together mother_job father_job
##    <chr> <dbl> <chr>   <chr>          <chr>            <chr>      <chr>
##  1 Male     16 Urban   Greater than 3 Together         health     other
##  2 Male     15 Urban   Greater than 3 Together         other      teacher
##  3 Male     16 Urban   Less than 3    Together         other      other
##  4 Male     17 Urban   Less than 3    Apart            other      other
##  5 Male     17 Rural   Less than 3    Apart            teacher    other
##  6 Male     19 Rural   Greater than 3 Together         other      other
##  7 Male     17 Rural   Less than 3    Together         other      services
##  8 Male     15 Urban   Greater than 3 Together         teacher    other
##  9 Male     16 Urban   Greater than 3 Together         teacher    other
## 10 Male     19 Urban   Greater than 3 Together         services   at_home
## # i 354 more rows
## # i 16 more variables: guardian <chr>, travel_time <dbl>, study_time <dbl>,
## #   failed_classes <dbl>, school_support <chr>, extra_curricular <chr>,
## #   want_higher <chr>, internet <chr>, romantic_rel <chr>, family_rel <dbl>,
```

```
## #   free_time <dbl>, go_out <dbl>, workday_alcohol <dbl>,
## #   weekend_alcohol <dbl>, absences <dbl>, grade <dbl>
```

4. Additional tips

1. Identify categorical and numerical data correctly.

2. Creating plots can be easy, but study the plots carefully to derive presentable insights.
3. Identify the best method to show what information you want to convey and then apply it.
4. Preprocess data using any of the various encoding and scaling methods before applying a model to predict the grade.
5. When modelling, remember to use cross validation.

6. You can always go back to data preprocessing and EDA if the model doesn't perform well.

7. Models can sometimes require hyperparameter optimization to perform well.
8. Commonly used R packages (besides the ones used above): inspectdf, moments, carat

Great, you're all set to go!