CSC8631 Assignment Report

Marc Birkett

07/11/2021

```r
library(ProjectTemplate); load.project()
```

```
## Loading required package: digest

## Loading required package: tibble

## Project name: csc8631

## Loading project configuration

## Autoloading packages

##  Loading package: reshape2

## Loading required package: reshape2

##  Loading package: plyr

## Loading required package: plyr

##  Loading package: tidyverse

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
##  Loading package: stringr

##  Loading package: lubridate

## Loading required package: lubridate

##
## Attaching package: 'lubridate'

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

##  Loading package: DiagrammeR

## Loading required package: DiagrammeR

## Autoloading helper functions

##  Running helper script: globals.R

##  Running helper script: helpers.R

## Autoloading data

## Munging data

##  Running preprocessing script: 01-A.R

##  Running preprocessing script: 02-E.R

##  Running preprocessing script: 03-L.S.R

##  Running preprocessing script: 04-Q.R

##  Running preprocessing script: 05-S.A.R

##  Running preprocessing script: 06-T.M.R

##  Running preprocessing script: 07-V.S.R

##  Running preprocessing script: 08-S.S.R
```

CSC 8631 - Data Investigation with Student Data

Introduction

Report into investigation of Student Data using the CRISP-DM model. This report covers two iterations of the model and includes the processes of Business Understanding, Data Understanding, Data Preparation. The subprocesses I've chosen are to do the following steps:

- Import
- Tidy
- Visualise
- Understand
- Communicate

The project has been set up using ProjectTemplate to provide some structure and repeatability, which will be tested on a regular basis. Version control is provided by Git and this report created with R Markdown.

Libraries

- Readr - library to provide extra functionality to import the data from CSV. In this case it allows me to import that data and assign type.

- r - data management.

Relationships

The data was investigated to reveal no primary keys, however columns such as learner_id could be used to create local relationships.The initial table structure has been envisaged as per below.

Simple Entity Flow Diagram

Iteration 1

Iteration 1 will be used to investigate the data and generate a hypothesis for further investigation. It will go through the entire list of sub-processes outlined above. Once a hypothesis has been identified this will be further investigated in iteration 2.

1 - Import and Tidy

The data was imported into R using Readr into 8 data frames from the original csv files, for easier analysis, the initial data sets were split out by dataset number which resulted in too many separate data frames to investigate. This step was done using ProjectTemplate's munge functionality. Due to the fact that the detected data types were not consistent the following data manipulation was carried out.

- ID - to integer
- Date time - to the local POSIX date time format
- Logicals - converted to logical

Initial exploration of the data indicated that Learner_Id was a candidate for relating 6 of the 8 data frames as it was a common column, the data regarding video views was video centric and did not have learner id nor did the sentiment survey data.

2 Visualise

3 Understand

Iteration 2

Iteration 2 will further investigate the hypothesis identified in iteration 1 and will present the findings.

Findings

To answer the hypothesis XYZ the findings are that ABC

Conclusion

References

ReadR - https://readr.tidyverse.org/index.html DiagrammeR - https://rich-iannone.github.io/DiagrammeR/ Working with categorical data - https://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

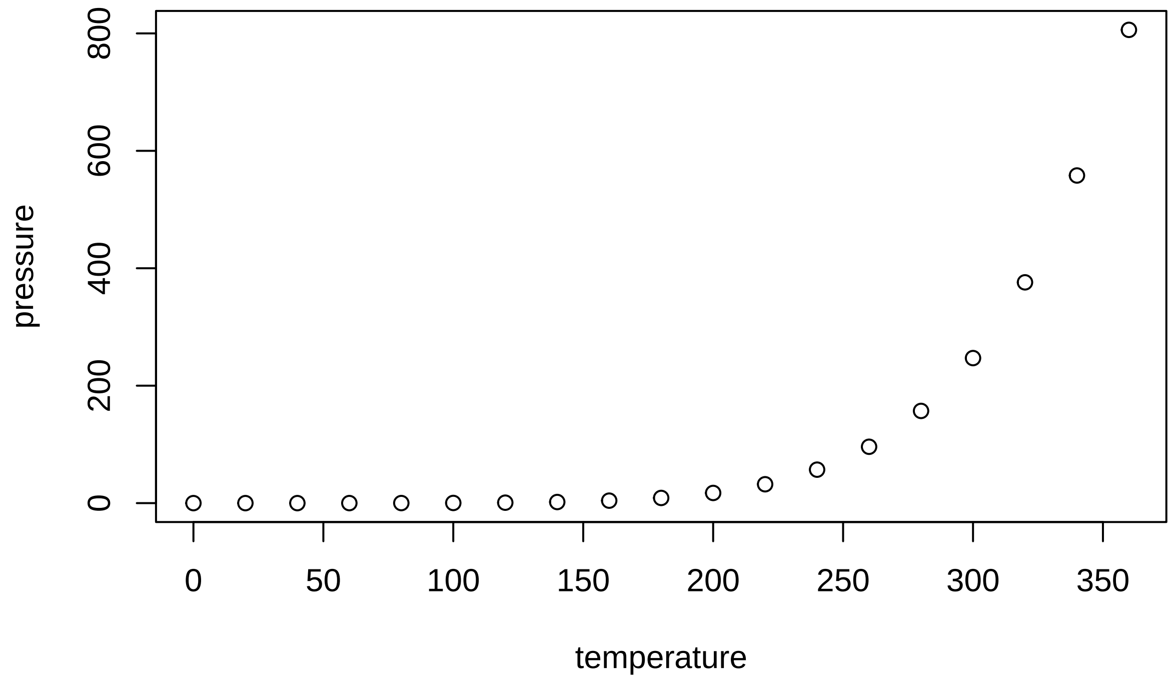
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.