CSC8631 Assignment Report

Marc Birkett

07/11/2021

CSC 8631 - Data Management and Exploratory Data Analysis

An investigation into an MDOC dataset using a CRISP-DM model. The model is the industry standard approach to data mining projects, it has six phases and is iterative. It is described as a "set of guardrails to help you plan organise and implement your data science (or machine learning) project."(Data Science Alliance, 2021).

The phases and tasks of a CRISP-DM project are outlined by the Data Science Alliance as per:

1. Business Understanding - what does the business need?

- Understand the business objective. What does the customer want to accomplish?
- Assess the situation. Determine resource availability, requirements, risks and contingencies
- Data Mining Goals - what does success look like?

2. Data Understanding - what data do we have / need? Identify, collect and analyse.

- Collect the initial data, aquire and load into data analysis tool.
- Describe the data, examine and document properties such as data format, number of records and identities
- Explore the data, query and visualise and idetntify the relationships.
- Verify the data quality and document any issues.

3. Data Preparation - prepare the final datasets for modelling.

- Determine which data set to be used and reasons for inclusion / exlusion.
- Clean the data
- Construct data, carry out any required feature engineering.
- Integrate data, create new datasets by combining data sources
- Re-format data as necessary, cast data types as required.

4. Modelling - what modelling techniques should we apply?

- Determine which models to try.
- Design tests.
- Build and assess the model.

5. Evaluation - which model best meets the business objectives?

- Evaluate the models against the business success criteria identified in step 1.

- Review the processes, was anything overlooked and everything properly executed?
- Determine next steps, what do we need to do to deploy.

6. Deployment - how do stakeholders access the results?

- Plan deployment, document.
- Plan monitoring and maintenance.
- Produce the final report.

This report will cover the first three steps of this model and a minimum of two iterations. I will attempt to implement all of the associated tasks. With the lack of formal requirements provided by the assignment it will be difficult to define a business need in the first iteration of the model therefore tasks such as defining the business objective and data mining goals will be defined for the second iteration as a result of the first. So, each therefore the two iterations of the model will be:

1. Iteration 1

- Data Understanding
- Data Preparation

2. Iteration 2

- Business Understanding - define and understand the hypothesis from Iteration 1.
- Data Understanding
- Data Preparation.

The investigative project will be completed in R using R Studio and has been set up using ProjectTemplate to provide structure and repeatability. Version control is provided by Git and this report created with R Markdown. Various libraries have been imported and used from the Tidyverse regarding data import and management (Dplyr, Readr) and visualisation (GGPlot2).

Iteration 1

A first iteration completing the understanding and preparation step will be carried out. The Future Learn MDOC dataset was downloaded as a zip file and reviewed. The data was supplied in csv files covering 8 different areas of the software over multiple stages, this was imported into R into 8 data frames from the original for easier analysis. The data was then reviewed.

Step 2: Data Understanding Upon initial import is was found that the detected data types were not consistent and so his was dynamically set on import. Particularly any ID property was set to be an integer and date time properties we converted to the local POSIX time format.

Upon investigation it was decided that some data would be disregarded due to lack of breath. Therefore, we have data covering the life cycle of a student including:

- **Enrollments**. Enrollment is a categorical dataset with n=37296 items and p=14 variables. It was found that the majority of data for all properties was "unknown" and therefore not suitable for further analysis.
  - **Primary Key** - "Learner_ID" of character data type.

- **Fields and format** - Properties include Country, Gender, Age Range, Highest Education Achieved, Employment Area and Employment Status, Role all of type character. Also date / time of enrolled, un-enrolled, fully participated and purchased statements.
- **Related data sets** - Team Member, Leaving Survey and Step Activity can be related on Learner_ID
- **Data Quality** - The data types across date times in multiple files were inconsistent and therefore they were cast on import. The vast majority of data for each categorical value is "Unknown". We don't have enough data to analyse.

- **Leaving Survey**. The leaving survey responses comprises of categorical, character and date based data regarding feedback from individuals which have left the course. The data frame has n=403 items and p=10 variables. The primary key of this table is ID which was cast to an integer in the import process.

  - **Primary Key** - "ID" which was cast to integer during import.
  - **Fields and format** - "ID" as integer, "Learner_ID" as character, "Left_at" as datetime, the date and time a student left the course, "leaving_reason" as characters, "last_completed_step" as character and "last_completed_step" as as date time, also the "last_completed_week_number" and "last_completed_step_number".
  - **Related data sets** - Enrollments on "Learner_ID". Investigation into the key fields of "last_completed_step", "last_completed_step_number" and "last_completed_week_number" revealed that "last_completed_step" was a concatenation of the other two fields and could be used to relate to the Step Activity data. This allows the identification of the stage that students leave the course. This can be seen in Figure 1.

```
## # A tibble: 5 x 3
##   last_completed_step last_completed_step_number last_completed_week_number
##   <chr>                                   <dbl>                       <dbl>
## 1 1.3                                         3                           1
## 2 1.19                                       19                           1
## 3 3.18                                       18                           3
## 4 1.3                                         3                           1
## 5 1.2                                         2                           1
```

  - **Data Quality** - Upon import it was necessary to standardise the data type across multiple files, id was cast to integer, last completed step to character and the last completed step and week numbers to integer. All date times were also cast to date time. Upon investigation of the data it was found that there were multiple leaving reasons which amounted to "lack of time", this field was merged to make it comparable to other values.

- **Step Activity**. This indicates the stage in the course that the related data occurred at. It is principally categorical data with n=423072 items and p=6 variables.

  - **Primary Key** - A compound key of "Learner_ID" of character data type and step (integer). This identifies data about the step the student was at.
  - **Fields and format** - "Learner_ID" of type character, "week_number" and "step_number" both of which are integer and "step" which is also an integer and is a concatenation of the two. First_Visited_At and "Last_Completed_At" are both date times and indicate when the student was at that stage.
  - **Related data sets** - Enrollments via "Learner ID", Question Response, Video Stats and Sentiment Survey via Step.
  - **Data Quality** - Data types were consistent across the files to be imported. Date times were cast on import.

- **Question Responses**. Each step in the step activity data frame has associated quiz questions and responses. This is stored in the questions data frame. The question responses data is categorical data with n=176463 items and p=10 variables.
  - **Primary Key** - "Learner_ID" of character data type.
  - **Fields and format** -
  - **Related data sets** -
  - **Data Quality** -
- **Video Views**. Each step in the step activity data frame also has associated video view data. The video stats data is multivariate data with n=65 items and p=29 variables. There are significantly more data variables in this relation then in those examined previously, and significantly less data items.
  - **Primary Key** -
  - **Fields and format** -
  - **Related data sets** -
  - **Data Quality** -
- **Weekly Sentiments**. This is completed by each student per week. The sentiment data has n=181 items and p=6 variables. This table has a categorical properly rating and character data to cover the sentiment. I will not be looking to cover sentiment analysis in this work so only the categorical value will be looked at.
  - **Primary Key** - "ID" has been cast to integer during the import process
  - **Fields and format** -
  - **Related data sets** -
  - **Data Quality** -

The relationships between data sets were documented as per Figure 1.

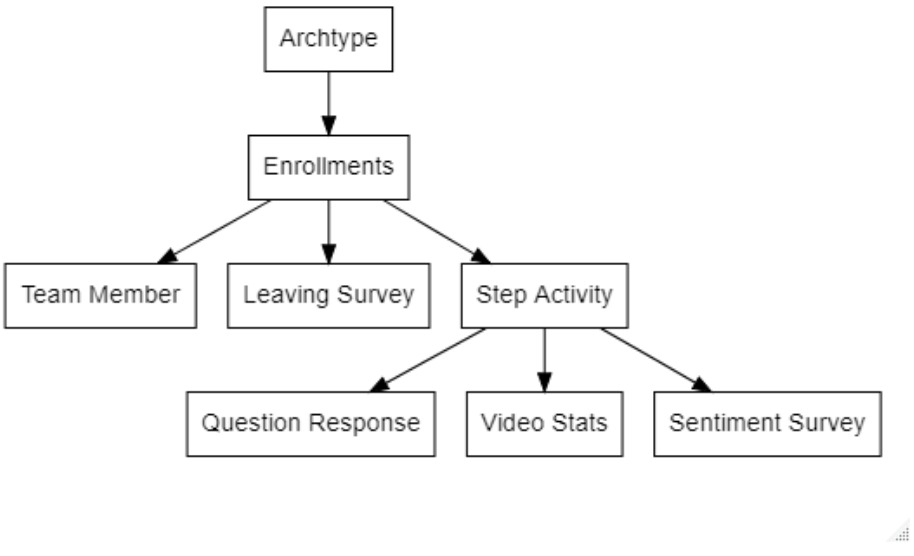


Figure 1: Relationships between data frames.

Data Preparation

3. Data Preparation - prepare the final datasets for modelling.

- Determine which data set to be used and reasons for inclusion / exclusion.
- Clean the data
- Construct data, carry out any required feature engineering.
- Integrate data, create new datasets by combining data sources
- Re-format data as necessary, cast data types as required.

Iteration 2

Business Understanding - Hypothesis

Data Understanding

Data Preparation

Conclusions

References

Data Science Alliance, 2021, What is CRISP-DM, https://www.datascience-pm.com/crisp-dm-2/, Accessed: 26/11/2021

The next related dataset to the enrollment data is the step activity data. This indicates the stage in the course that the related data occurred at. It is principally categorical data with n=423072 items and p=7 variables, although two more were subsequently added after feature engineeering.

The data includes the step number, the step start and end date with the learner_id. The end date has been used to indicate whether a step was complete for that student and how long the step took, this has resulted in 2 new columns - "isComplete" flag and "completedTime" in days, and allows us to calculate step completion stats and time scales per step as per below.

We can see that there are far more completed steps than incomplete, and that incomplete steps are generally earlier in the course. However, there are still students who complete early steps and subsequently fail to complete later ones. There also indicates a wide spread of time to complete each step, with a variance of 103.1396. An outlier of this is *Step 2.8* which takes everyone very few days to complete.

Due to adding the stage of the course we can also see that the time to complete each unit reduces as the course wears on, as does the number of students that complete each step.

The addition of the feature engineered values gives us the potential to use this dataset in further analysis. There is potential to look at incomplete steps in relation to course leavers, and investigate the short completion times of step 2.8.

Each step in the step activity data frame has associated quiz questions and responses. This is stored in the questions data frame. The question responses data is categorical data with n=176463 items and p=12 variables after the feature engineeering outlined below. The question responses table has the quiz question which includes the step number. Feature engineering was carried out to extract the correct step number and add it to a column, should there be a need to relate to the the activity steps.

An examination of the data was carried out to look at the volume of responses.

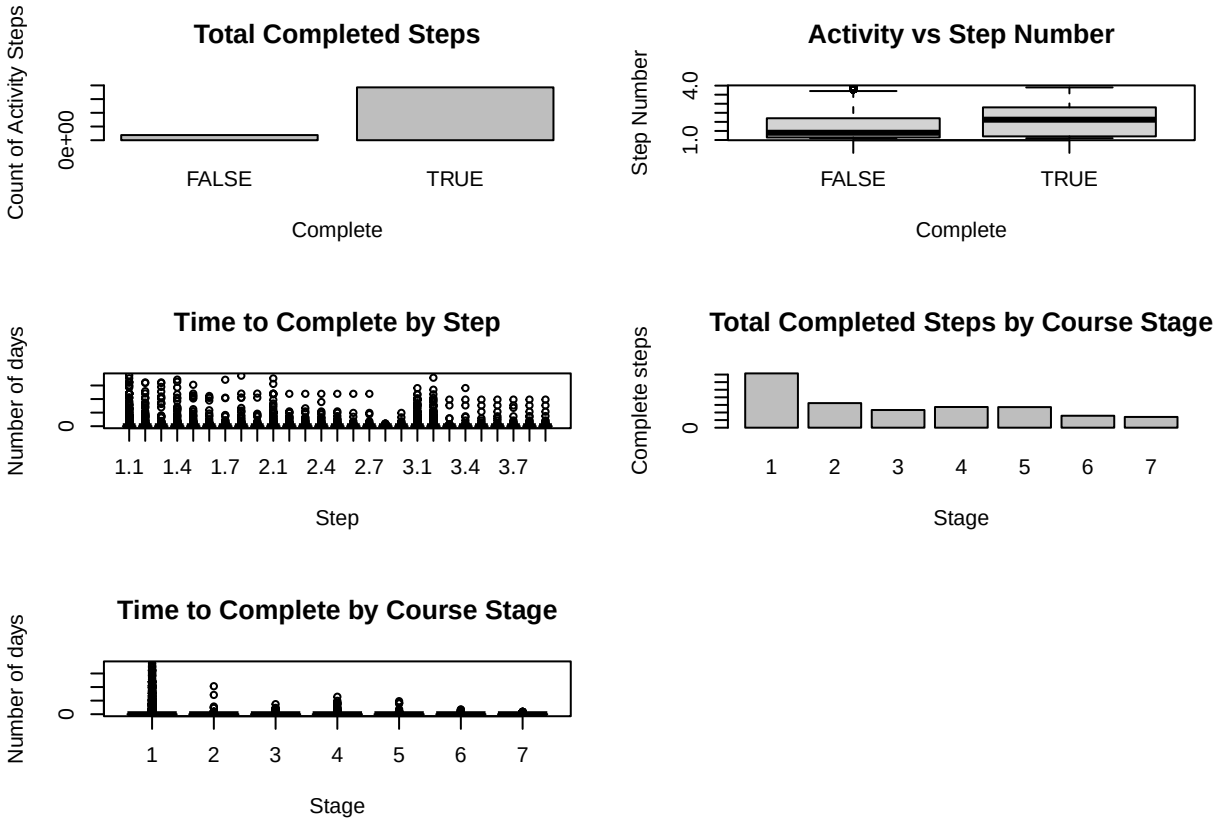


Figure 2: Step Activity Exploration

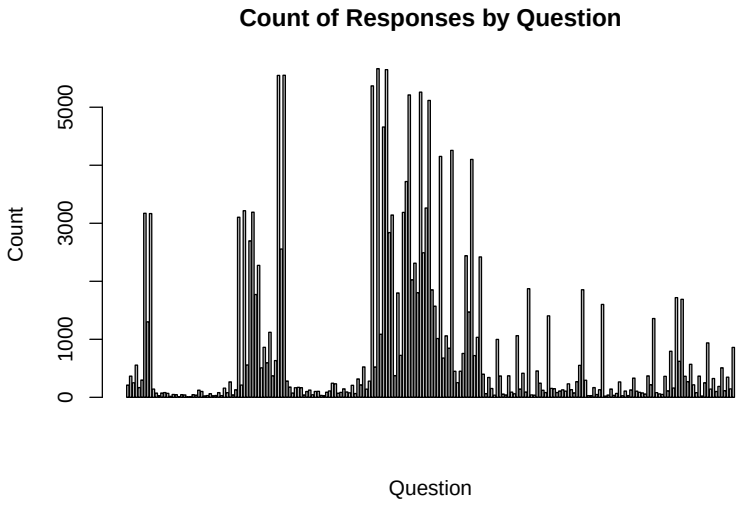


Figure 3: Quiz Question Responses

As we can see there are significant differences in number of times each question was answered over the length of the course and the as expected mix of correct and incorrect answers. There is the potential here to look into the mix of correct and incorrect answers, potentially investigating any relationship between low attainment on the quizzes throughout the course to the student subsequently leaving.

Each step in the step activity data frame also has associated video view data. The video stats data is multivariate data with n=65 items and p=29 variables. There are significantly more data variables in this relation then in those examined previously, and significantly less data items.

The data in this relation can be subsetted into data pertaining to percentage complete, type of device and location of the video view, so three data sets were created with just this data present. Some work was then done to convert the data from percentages to raw numbers, and then to pivot the result to allow some insights into the number of views, views by device and views by location in Figures 7, 8 and 9 below.

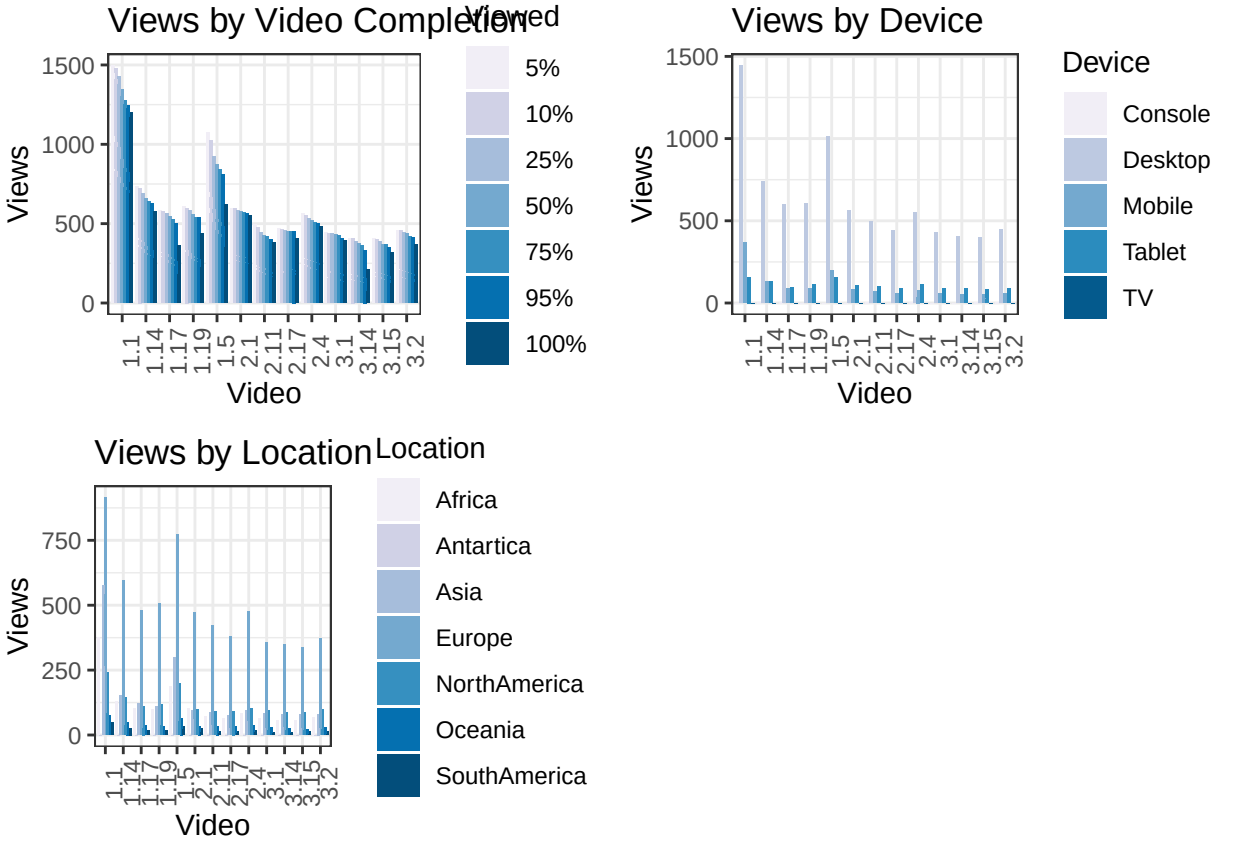


Figure 4: Video Views

It can be seen in the barplot for the Views by Video completion that the number of views which watched the whole video dropped off as they viewer got further through. For example we can see for the first video "1.1. Welcome to the course" that 1500 views completed 5% of the video, which drops to 1200 who completed 100%. This is then presented for each video throughout the course. As the course progresses the volume of students engaging with each video reduces reasonably significantly other than "1.5 Privacy online and offline" It can also be seen that some videos see a significant drop between 95% and 100% completion.

When viewing Views by Device it can be seen that no video was viewed on a tablet or TV, all views were by Console, Desktop or Mobile. It can be seen that by far the majority of views were on a Console throughout the length of the course, so it may be concluded the videos should be designed to play best on a console. The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

By viewing Views by Location it can be seen that the majority of views are from Europe with non at all from Antarctica and few from North America, Oceania and South America. The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

There is potential to link views with stages and we could potentially look at leavers by location and device by using this view data along with the leaver survey responses. We could also look at test performance related to video completion, location or devices.

The final data frame related to the course stage is the weekly sentiment data, this is completed by each student per week. The sentiment data has n=181 items and p=6 variables. This table has a categorical properly rating and character data to cover the sentiment. I will not be looking to cover sentiment analysis in this work so only the categorical value will be looked at.

As we can see from the weekly experience data below, across the three weeks of the course the total sentiment data provided by the student cohort went down from 73 to 45, although the experience ratings being provided remained consistent at approximately 90% providing a rating of 3.

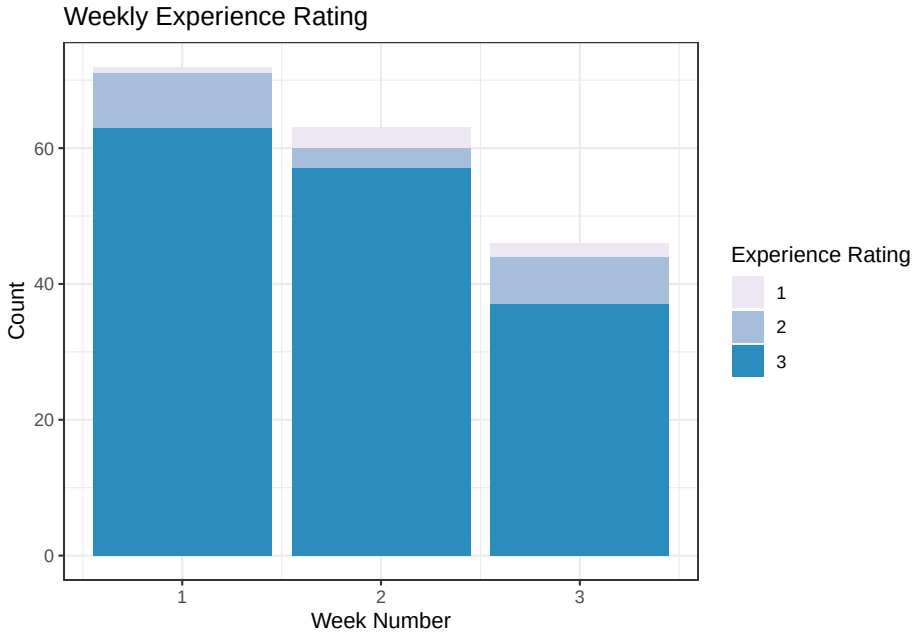


Figure 5: Sentiment Completions

It would be of value to understand this data as a percentage of the total patient cohort. Given the lack of a "learner_id" in this data it isn't immediately possible to relate the sentiments back to a specific learner, so this data could only be looked into by step. There may be value in understanding after which step students decided to stop providing feedback, this could be related on the "stage_id".

3 Conclusion and Hypothesis

There is a wide breath of data in the data set regarding activity throughout the course, and richness is provided by the video views data frame which includes data such as the device used, location of the student and so forth. The use of the step dataset allows us to investigate the stage of the course which events happened, and the leavers dataset allows us to look into leavers. Knowing that the location work has been completed previously by other students, I feel that there is enough data to investigate something around student performance, stage and the decision to leave.

I hyptothesise that the lower completion of videos and lower weekly sentiment scores increase the chances of a student choosing to leave the course. This will be investigated in iteration 2 of the CRISP-DM model.

Iteration 2 Hypothesis Test performance vs video views vs leavers

Iteration 2

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

Location Analysis Multivariate Analysis GGPlot2

Findings

To answer the hypothesis XYZ the findings are that ABC

Conclusion

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.

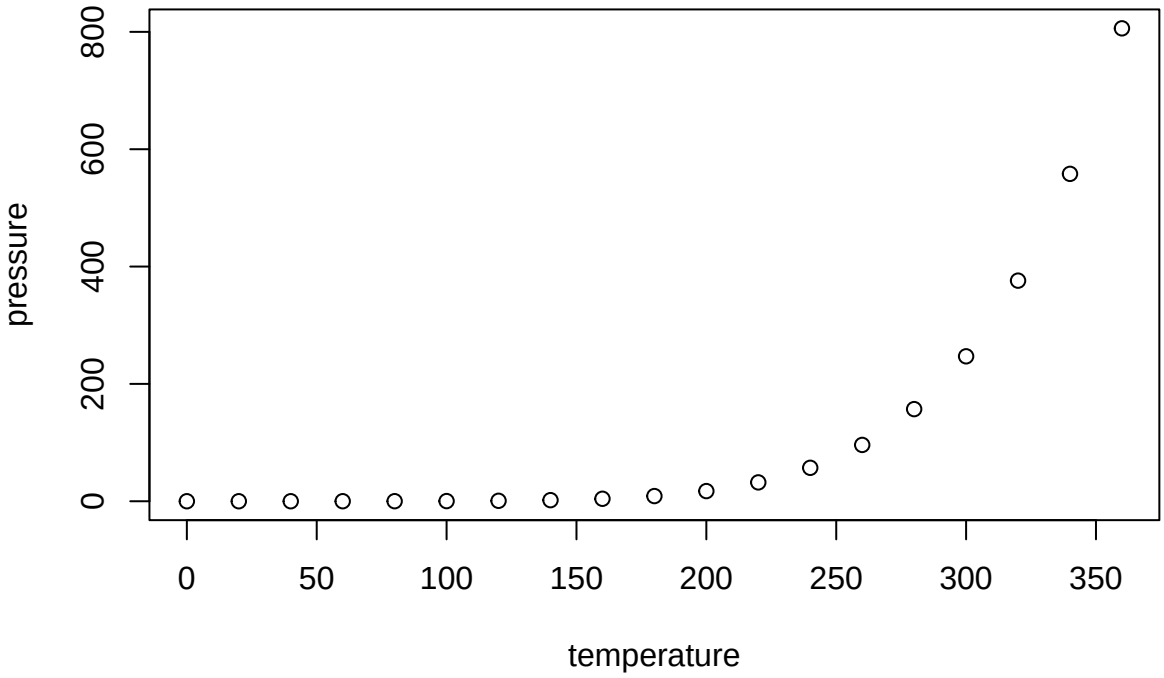
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.