

## A snappy project title

STDS AT2B Project Report

Team name

Student names and IDs

Statistical Thinking for Data Science TD School

University of Technology Sydney

UTS CRICOS 00099F



### Executive summary



Key result 1

Key result 2

STDS 36103 – AT2B2 project report structure example

### Table of Contents

[**EXECUTIVE SUMMARY 2**](#_heading=h.gjdgxs)

[**TABLE OF CONTENTS 3**](#_heading=h.30j0zll)

[**INTRODUCTION 5**](#_heading=h.1fob9te)

[PROBLEM STATEMENT 5](#_heading=h.3znysh7)

[RATIONALE 5](#_heading=h.2et92p0)

[PROJECT AIMS AND OBJECTIVES 5](#_heading=h.tyjcwt)

[**METHODOLOGY 8**](#_heading=h.3dy6vkm)

[METHODS OVERVIEW 8](#_heading=h.1t3h5sf)

[METHODS DETAILS 8](#_heading=h.4d34og8)

[**RESULTS 11**](#_heading=h.2s8eyo1)

[KEY FINDINGS 11](#_heading=h.17dp8vu)

[IN-DEPTH RESULTS 11](#_heading=h.3rdcrjn)

[*Does increased smoking per captia cause a decrease in fertility rates? 11*](#_heading=h.26in1rg)

[**CONCLUSION 14**](#_heading=h.lnxbz9)

[TAKE HOME MESSAGE 14](#_heading=h.35nkun2)

[DISCUSSION 14](#_heading=h.1ksv4uv)

[PROJECT LIMITATIONS AND CAVEATS 14](#_heading=h.44sinio)

[STAKEHOLDER ANALYSIS AND PROJECT OUTCOMES 14](#_heading=h.2jxsxqh)

[**ACKNOWLEDGEMENTS 16**](#_heading=h.z337ya)

[**REFERENCES 17**](#_heading=h.3j2qqm3)

[**APPENDICES 18**](#_heading=h.1y810tw)

**Keypoints:**

1. **Structure the dataset - while taking into consideration the correlation threshold > 0.2**
2. **Models - Compare two models (Comes under method details)**

**Links for the notebooks - https://www.kaggle.com/code/mohamedelsaadany/statistical-modeling-of-life-expectancy-data-r/notebook#4.-Feature-selection-methods**

1. **Prediction - Accuracy of the model**
2. **Results - Key findings**
3. **Conclusion - Project limitations and future improvements**
4. **Happiness score ()**

**Feedback:**

1. **There is a mismatch between the overall aim, predicting life expectancy, and the research questions.**
2. **SHAP values for feature importance may indicate which are better predictors**
3. **High correlating variables (GDP, traffic) - showcase these**
4. **Include the gif (showing all country data)**
5. **Highlight the other variables (not existing in the WHO data)**

Kyle Zawada. STDS 2022 3



# Introduction/ project background

### Introduction

#### Problem statement

State, in simple terms, what the problem this project tackles is. Use easy-to-understand language. Details can be provided in the main text. Keep it short and punchy.

#### Rationale

Outline the project background and reason the study is important. Start broad in the initial paragraphs, with high level background.

Become more specific with each paragraph, before dialling into information that is very specific to your research questions and aim.

Include references and key facts and numbers (x% of the population, Y million dollars spent).

You may want to include conceptual figures to support the project background. But these should not be based on data or results from the project. Anything generated from the project should be included in the relevant methods/results section.

#### Project aims and objectives

State the main aim of this research, ensuring that the aim matches to the research questions and objectives. If the specific questions are aimed at the impact of smoking on fertility rates, the aim of the project should not be too broad i.e. “the aim of this project was to explore how lifestyle choice impact health” but “the aim of this research was to explore the link between smoking and fertility rates”.

Followed the aim up with research questions that are answered by your results, remembering that at least one question should be answerable by a regression model.

To support our project aim, we asked the following questions:

1. Does increased smoking per captia cause a decrease in fertility rates?
2. Etc…

Notice that question 1 is a hypothesis that contains references to data. Question 1 could be stated as “Does smoking change fertility?”, but that is too vague. What do you mean by fertility or smoking? How would you say there has been a

change? Make sure your questions are crystal clear, and that you analyses and data are able to answer your research questions. These questions can then form the subheadings for your results section helping with report structure and navigation.

Then, you can break the project into objectives. These objectives should be high level descriptions of the key steps required to achieve your project aim. These objectives can inform the subheadings for your methods section.

To answer these questions we had the following objectives:

1. Identify relevant datasets
2. Something about getting data ready
3. Etc,

You can include a final, sentence or two to almost tease the project outcomes without too much detail. E.g. “We show that smoking cause a significant drop in fertility rates across countries, and is made worse when combined with other risk factors. We discuss these results and their implications in the context of...”



Methodology

### Methodology

#### Methods overview

Again, this overview serves to outline this section in easy-to-understand language for non-specialists. Use this to outline the main steps and approach without jargon or technical details.

#### Methods details

Here you go into detail about all relevant steps to execute the project. From data acquisition through to producing final outputs. There should be enough detail for the reader to fully recreate your project. Include sources for your datasets. Examples of things to include in your methods are:

* Data acquisition
* Data manipulation and summaries
* Exploratory analyses and data decisions (e.g. removing observations)
* Formal analyses, including at least one regression model

Reference relevant R libraries used for analyses and data manipulation. There’s no need to outline exactly step-by-step the cleaning/analysis process, e.g. “we loaded the data into R and then used the summary() function to…” as this will be clear in the code appendix. However, you should outline the main steps, e.g. “We explored data in R to identify missing values. We then did *x* to handle missing values before…”

Break your methods into relevant subheadings. Ensure your analyses are clear, making an explicit reference to each research question. E.g. “To answer whether increased smoking per capita caused decreased fertility rates, we ran a linear regression model with fertility as the response and smoking per captia as the explanatory variable. We then checked if there was a negative and significant slope coefficient for the model.” Notice how it is very clear what question the model is answering and how the model will be used to define our answer.

Methods can include figures and tables to summarise key data relevant to the project. Make sure you directly reference tables and figures in the text, e.g. “The total number of observations in the final dataset was 2,355 (table 1)”.

You should aim your language in the main text (outside of the summary statements) towards other data scientists. For the presentations you could simply say “we plan to build linear models to answer this question”, but here you need to be specific. What family of linear model? What is the model structure? What

metrics are you using to assess your models and why? You do not need to explain what logistic regression is, but you should explain why you needed to use it here.

**Table 1. An example table showing the number of observations by year and location for the final dataset.**

| **Year** | **Location** | **Number of**  **observations** |
| --- | --- | --- |
| 2019 | USA | 72 |
| UAE | 72 |
| Australia | 301 |
| Bahamas | 143 |
| Tattoine | 192 |
| Kings Landing | 147 |
| Ireland | 73 |
| The Shire | 71 |
| 2019 Total | 1,071 |
| 2020 | USA | 166 |
| UAE | 72 |
| Australia | 72 |
| Bahamas | 72 |
| Tattoine | 72 |
| Kings Landing | 60 |
| Ireland | 143 |
| The Shire | 144 |
| France | 72 |
| High Charity | 72 |
| Smallville | 72 |
| Valyria | 72 |
| Mordor | 71 |
| Lichtenstein | 64 |
| Australia 2 | 60 |
| 2020 Total | 1,284 |
| Grand Total | | 2,355 |



# Results

### Results

#### Key findings

Put your one or two biggest, most important, or most surprising results here. Again, high level and high impact.

#### In-depth results

Does increased smoking per captia cause a decrease in fertility rates?

Use your research questions as subheadings for your results to help guide the structure of the results. Start with the biggest results and then go into details. Do not interpret your results too much here. You could say something like “Linear regression showed a significant negative slope between per captia smoking and fertility rate (slope = -2.074, p.value < 0.01) suggesting that smoking negatively impacts fertility rate”. However, you shouldn’t add additional speculation or synthesis here, do not, for example, say “This supports research by so-and-so 2015, and is likely due to smoking impacting hormonal… etc, etc”.

You can also highlight potential outliers here for later discussion. And should include a statement of how model assumptions were checked and if they were met. Diagnostic plots and coefficient tables should mostly be included in a separate appendix instead of the main text. You can refer to these as such “linear regression assumptions were checked via diagnostic plots and were determined to be met (Appendix A, figure 1)”.

In some cases you may have tried one method (Poission regression), but found an assumption wasn’t met (e.g. overdispersion), and switched to another method (negative binomial regression). You can and should mention this here to show that initial ideas were tested but needed changing.

Take care with your figure design. Figures are typically some of the first things that people check when looking at a document for the first time; poor figures leave a poor first impression that could be hard to recover from. Take the time and effort to make your figures pop and look pleasing. Send the figure to someone outside of the project and ask for their honest opinion, if they find it hard to read or that it looks ugly, chances are it is! Make sure text is easily visible and replace any shortened variable names with descriptive names with units: e.g. ‘fert’ to ‘Average fertility rate (births per capita)’. Try not to stick with ggplots default colours and theme, even though it shouldn’t make a difference, everyone knows what a default ggplot looks like, and it makes people assume you haven’t thought about the figure design. If you want more control over your figures, export them to .pdf and open in a vector graphics editor such as InkScape.



**Figure 1. Example results figure to show hypothetical results for model accuracy for predicting fertility rates from per captia smoking. Make your figure legends self-contained and descriptive so the reader can understand them. For example, overall accuracy was high suggesting that per**

**capita smoking can be used to predict fertility rate. However, this figure doesn’t match with the research question. A better figure here would be a scatterplot showing the significant negative relationship between per captia smoking and fertility rate for all locations.**



Conclusion

### Conclusion

#### Take home message

What is the one thing--more than anything else in the world--that you want your reader to take away from your research? State that here with a follow up sentence about the implications or future work that this take home message inspires.

#### Discussion

Start the discussion by restating the aim of the project. If you achieved the aim, say so and provide evidence statements from your results. If you haven’t achieved your aim, provide a high level reason why you think this was the case before diving into more details further on.

The discussion should start specific and broaden back out. Think of it as the introduction in reverse; start with the project aims and objectives and then contextualise them in a broader context.

Try to thoughtfully connect the project outcomes with the wider themes underlying it. Smoking and fertility is one aspect of public health for example, what have you learned here that may have implications elsewhere.

Talk about any surprises and try to explain them. Things like outliers, or reversed hypotheses (what if smoking *increased* fertility rate in your study?). Whilst this is not a fully scientific article-style, try to look for and reference other relevant studies that support or maybe contradict your hypotheses or results.

#### Project limitations and caveats

Take time to be honest about the limitations of the study. This includes limitations in scope (amount of data, countries included, time periods included) and interpretation of outcomes (what your results do and so not tell you). When stating a limitation, it is important to appreciate and discuss it’s implications. For example, instead of just saying “One limitation was that all our data came from a single country, Australia.”, expand that, for example “Therefore we may expect our results to change if we included data from countries with lower GDP, because…, however, our results should hold across countries with similar socio- economic conditions”.

#### Stakeholder analysis and project outcomes

Here you can state what the key project outcomes are and the specific relevant stakeholders. Project outcomes can be an artefact (e.g. a regression model that can be used for prediction) or knowledge (e.g. a newly discovered relationship between your response and explanatory variables). Project outcomes are not

coefficient tables, data visualisations, or p.values. Think of an outcome as something someone else could use in some way.

Once your outcomes are defined, you should then state the key stakeholders that you envisage could benefit from it. If it is a predictive model, it could be a tool governments could use to make predictions for future planning. If it is knowledge, it could be something used by policy makers to inform the public or help draft legislation.

Outcomes and stakeholders do not need to be a huge list, identify the most important outcomes, and for those outcomes the main stakeholders.

### Acknowledgements

Typically here is where you can mention other people or organisations outside of the main project team that supported throughout the project. Probably not applicable here given that you are all working in teams with no engagement outside of that, but it is good to be aware of.

### References

Include any referenced materials here. Whilst the exact referencing style doesn’t matter, it should be consistent and roughly follow the APA, Harvard, or Nature styles. You can use a reference manager such as Endnote or Zotero to help manage references throughout.

### Appendices

You can include additional tables and figures as appendices to support the report. This should not be an infinite stream of every figure or analysis you can think of, but should include relevant additional information. Examples would be regression assumption plots (Q-Q, residuals) to support assumptions. It may include regression tables or statistical tests showing if normality has been met, or overdispersion tests for a Poission regression.

Everything here should serve some supporting purpose for the project. Make sure to check and justify as a team what goes into the appendices. Also, do not include additional text here outside of figure and table legends. All additional text will be ignored for marking the main report.