# CASA0007 Research Proposal

Student Number: 22186878

Date submitted: 2022-12-05

* Proposed research question
* Bibliography (single source inspiration paper)
* Description of source
* Explanation of value of research question (with reference to paper)
* Summary of methods that may be applied

## Planning Material

A good research question…

1. Cannot be answered with a single calculation
2. Is not vague (can be quite detailed, and split into subsections if required)
3. Asks about the real world (not just data or methods)
4. Is interesting, valuable, and ambitious

Workflow for hypothesis testing:

1. State the null and alternative hypotheses
2. State variables and data type
3. State assumptions
4. State test statistic used
5. State test results
6. State hypothesis conclusion

Assumptions for a *linear regression*:

1. Line relationship exists
2. Independent errors
3. Normally distributed errors
4. Equal variance of errors (uniform distribution of variance)

What to report in a *linear regression*:

1. Fitted equation
2. P-value
3. R2 value (and adjusted R2)
4. Scatter plot of data
5. Residuals vs fit plot

Workflow for *linear regression*:

1. Identify the regression problem
2. Explore the data (clean, visualise)
3. Select dependent and independent variables
4. Build and fit regression model
5. Conduct residual analysis
6. Refine the model
7. Interpret the model

Workflow for *cluster analysis*:

1. Standardise the data (Z-score, min-max rescaling)
2. Choose and apply a clustering method (K-means or hierarchical)
3. Assess clustering quality (SSE/Elbow diagram, Silhouette analysis)
4. Visualisation (cluster plot, elbow diagram, silhouette plot)
5. Follow up (Describe cluster characteristics, examine cluster centroids, compare against unconsidered variables/geography, consider analysing clusters separately)

## Resources

**Potential Data Sources**

|  |  |  |
| --- | --- | --- |
| **Theme** | **Description** | **Source** |
| Monkeypox | Monkeypox cases in the EU/EEA, by day and country | [European Centre for Disease Prevention and Control](https://www.ecdc.europa.eu/en/publications-data/data-monkeypox-cases-eueea) |
| COVID | National Flu and COVID-19 Surveillance reports UK | [UK Health Security Agency](https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2022-to-2023-season) |
| Indigenous health outcomes | Australian Prisoner Statistics by Indigenous status, Jurisdiction and Year | [ABS](https://www.abs.gov.au/statistics/people/crime-and-justice/prisoners-australia/2021#data-download) |
| HIV | HIV Annual Data Tables UK | [UK Health Security Agency](https://www.gov.uk/government/statistics/hiv-annual-data-tables) |
| HIV | Local Government Expenditure UK | [Ministry of Housing, Communities and Local Government](https://www.gov.uk/government/statistics/local-authority-revenue-expenditure-and-financing-england-2020-to-2021-budget-individual-local-authority-data) |

**Test Statistic Decision Tree**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Is the independent variable categorical or quantitative? | Cat. | Is the dependent categorical or quantitative? | C | Logistic Regression; Wald Test | | | | |
| Q | How many independent variables are there? | 1 | Simple regression; t-test | | |
| >1 | Multiple regression; f-test | | |
| Quant. | Is the dependent categorical or quantitative? | C | Chi-squared | | | | |
| Q | How many groups are being compared? | 2 | Are the standard deviations similar? | Y | Comparison t-test |
| N | Welch’s test |
| >2 | ANOVA or MANOVA | | |