**12 Marital Status and Pregnancy**

The ONS have released data regarding conceptions from 1998–2017 in the UK:

<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/datasets/conceptionstatisticsenglandandwalesreferencetables?fbclid=IwAR39nzs5Bnj4Xi_cs7Ismm830-yaIkUPKjZQRSE2GMRdi8jXcoIAaMeSnz4>

This data is particularly interesting as from September 2020 Sexual Education is made compulsory in all UK schools as part of the curriculum.

**Question:**

Does age and marital status impact a person’s decision to continue with their pregnancy? The analysis may include the following sub-questions:

• Which of the conceptions took place within marriage or civil partnerships and how many conceptions lead to an abortion?

• How did conceptions change over this time period both in and out of wedlock and within certain age categories?

**Module Title: Introduction to Statistical Data Analysis with R**

**Deadline for submission:27/01/2023 by 15:00**

**Assessment brief**

You need to perform an initial analysis of the data with the help of R and answer the Assessment research question on **Marital Status and Pregnancy indicated above**

You should describe your findings in a written report

The requirements and content are detailed below

You are expected to complete the following steps in your assessment:

**Data preparation and pre-processing.**

This may include downloading, re-shaping and merging data, selecting subsets, amending variables and setting correct variable types, creating new variables, dealing with missing values and outliers, and otherwise ensuring data integrity.

**Data analysis.**

The analysis must be aimed at answering your research question(s) and may include simple data summaries and statistical tests covered in the module as well as “advanced” methods outside the module (for example, linear regression or decision trees). If you decide to use a method outside the module, some background research into it is expected but not mastery of the method. The selected methods, both covered in the module and advanced, must be appropriate for your data and research question(s).

**Results presentation.**

You must make your data to tell a story and present your findings in a convincing and verifiable way, using appropriate tools including tables and visualizations

**Note: You must do all data preparation and analysis using R code. All the visualization (graph, histogram, bar charts…) should be done using R** (including downloading the data from the web, where possible). If you use any other software tools for certain steps, you should explain briefly why it is impossible or considerably more difficult to do these steps in R. You are expected to reasonably care about the efficiency of procedures and data structures you use.

**Assessment report**

Your written report should be aimed at a colleague (a junior data analyst) without specialist knowledge of the topic and must include the following elements:

• Brief introduction describing the context of the data and research.

• Streamlined account of your data preparation and pre-processing with motivation or justification for important steps.

• Brief general description of the resulting data set (at least the number of observations and the description, type and suitable summaries for each variable).

• Streamlined account of your data analysis, including justifications for the choice of methods and approaches. You should avoid general descriptions of methods without relating them to your data set and analysis; if you use “advanced” methods outside the module, you may give links to appropriate sources where necessary.

• Clear statement of your results and conclusions with a discussion of their reliability and possible alternative interpretations. You should also include a separate paragraph summarising your findings for a non-specialist audience (typically, for the general public, but you may choose to write a summary for a decision-maker).

• Short reflection section discussing the project at a meta-level, for example, what you have learned in the project, your motivation and challenges, your mistakes in the project, what additional analysis you could do if you had more time or knowledge, how the same research question(s) can be addressed with different data, what alternative research question(s) can be posed within the same aim, etc.

**The report will be typically 2500 to 4000 words.** The word count is notional and given as a guidance only (in other words, the markers will informally assess whether the length of your report is appropriate rather than apply a strict word counting procedure). Code snippets are not counted towards the report length. You are strongly advised to use informative visualizations whenever possible.

**Note: The R codes used would not be included in the word count**

You must include all R code necessary to reproduce your analysis in full. You can either interleave code and text (for example, using RMarkdown) or submit your code separately. You should not normally include dead ends, wrong variants, etc. (this approach is referred to as “streamlined” above), but you may mention them to provide motivation for your choices.

Your submission must include the following files:

• Your report as a single pdf file.

• Your R code if not included in the report.

• Your original data set(s) if your code does not download them from the web.

**Assessment marking criteria/rubric**

Your submission will be marked according to the standard University Grading Criteria within the following rubric:

• 10% for clarity and consistency of the submission (including the appropriate length of the report and meeting all requirements to the submission).

• 35% for correct and appropriate use of R. The precise allocation of marks within this item may depend on the details of your code but as a general guidance, the following four elements will be assessed (not necessarily with equal weight):

o the use of appropriate data types and structures;

o the range of appropriately used methods and packages;

o clarity and efficiency of your code, including the ease to re-use and modify it;

o appropriate formatting of outputs and visualisations.

• 15% for correct use of statistical procedures, including assigning appropriate types to variables and selecting suitable summaries, visualisations and analysis methods

• 10% for data preparation and pre-processing.

• 5% for the brief general description of the resulting data set.

• 5% for the statement of obtained results and conclusions.

• 20% for justifications, motivations, discussion and reflection.