**ST8002-Project**

The objective of this project is to provide you with the opportunity to apply the methods that we have been discussing to your own work.

**About the report**

Use the given template. This takes the form of three reports on three different datasets, one report for one dataset. (It has to be three different datasets. You can not use one dataset for two reports.)

Report format: Calibri, font size 12, single-line space, no spacing before or after. (The given template is formatted with such settings.)

**About the data**

The data should be related to your study/work background. (What is the point of taking 8001&8002, if you don’t have or can’t think of any data analysis problem in your field?)

You need to include a link to your data in your report, so that I can freely download the data by simply clicking the link. For example, upload your data into Google drive or OneDrive and allow anyone with the link to download.

The data have to be from open-data repositories (more information below) or collected by you and/or your colleague. If your data are collected by you and/or your colleague through, e.g., lab experiments. You don’t need to provide the original raw data, and you can do any kind of anonymization. (I won’t re-use or re-distribute your data.) What I need are the data that, by directly running your R code, I will get the same results as those in your report.

If you currently don’t have any data, you can download any relevant data (w.r.t. your study/work) from the following open data repositories:

* Smart Dublin (<https://data.smartdublin.ie/>), or any other country’s local government open data portal,
* Ireland - data.gov.ie (<https://data.gov.ie/>), or any other country’s central government open data portal,
* WHO’s open data repository (<https://www.who.int/data/gho/>), or any other international organisation’s open data portal,
* Amazon’s open data AWS (<https://registry.opendata.aws/>), or any other company’s open data portal,
* Kaggle (<https://www.kaggle.com/datasets>), or any other free-access website.

The above list is far from being complete. If you can’t find relevant data from the above open data repositories, you can look for other resources. Note that these resources must be open-access. This is to drive engagement and impact from open data repositories.

**About exploratory analysis**

For any data-analysis project, you always start with exploratory analysis, mainly through data visualization. The analysis differs for different types of data (continuous, discrete, ordinal, nominal). Here are some topics that you might include in your reports.

|  |  |
| --- | --- |
| **Topics** | **Marks** |
| (E1) contingency table | 3 |
| (E2) conditional probability and independent events | 3 |
| (E3) Pearson/Kendall/Spearman correlation | 3 |
| (E4) point-biserial correlation coefficient | 3 |
| (E5) correlation coefficient significance level | 3 |
| (E6) centre, spread, skewness & kurtosis measures | 3 |
| (E7) outlier detection | 5 |
| (E8) correspondence analysis | 7 |
| (E9) correlation matrix visualization | 3 |
| (E10) bar plots of contingency tables | 5 |
| (E11) Q-Q plot (not limited to Normal Q-Q plot) | 5 |
| (E12) scatterplot matrix | 5 |
| (E13) balloon plot | 5 |
| (E14) grouped bar charts  (i.e., bar charts of two or more features, appropriately grouped in one graph) | 7 |
| (E15) grouped boxplots  (i.e., boxplots of two or more features, appropriately grouped in one graph) | 7 |
| (E16) histogram overlayed with the density line | 7 |
| (E17) conditional histogram | 7 |
| (E18) ridgeline plots | 7 |
| (E19) biplots | 7 |
| (E20) heatmaps | 7 |
| (E21) word clouds | 7 |

**About statistical analysis**

Your exploratory analysis will help decide what type of statistical analysis you need to perform. Here are some topics that you might include in your reports.

|  |  |
| --- | --- |
| **Topics** | **Marks** |
| (S1) design of experiments (3), sample size determination (3), power of a test (3) |  |
| (S2) regression (5), prediction interval (3), coefficient of determination (3) |  |
| (S3) regression residual analysis | 5 |
| (S4) one-proportion z-test or two-proportion z-test (pooled or un-pooled) | 5 |
| (S5) one-sample t-test or two-sample t-test (pooled or un-pooled) | 5 |
| (S6) Tukey’s test | 5 |
| (S7) chi-square test for independence | 5 |
| (S8) chi-square test for goodness of fit | 5 |
| (S9) chi-square test for normality | 5 |
| (S10) Anderson–Darling test | 5 |
| (S11) Shapiro–Wilk test | 5 |
| (S12) Kolmogorov-Smirnov test | 5 |
| (S13) Cramer--von Mises test | 5 |
| (S14) D'Agostino normality test | 5 |
| (S15) Fisher’s exact test | 5 |
| (S16) binomial test | 5 |
| (S17) Mann–Whitney U test | 5 |
| (S18) Wilcoxon signed-rank test | 5 |
| (S19) Kruskal-Wallis test | 5 |
| (S20) Friedman test | 5 |
| (S21) McNemar test | 5 |
| (S22) ANOVA | 5 |
| (S23) ANCOVA | 5 |
| (S24) MANOVA | 5 |
| (S25) Factorial ANOVA | 5 |
| (S26) One-Way Repeated Measures ANOVA | 5 |

**Requirements**

* The three reports **together** (not each) should cover no fewer than 20 **different** topics with the total mark >= 100.
* You can perform hypothesis tests that are not listed above, each accounting for 5 marks.
* For statistical analysis:

(1) You have to state clearly the null hypothesis and the alternative hypothesis.

Marks will not be given, if rejecting/not rejecting the null hypothesis is obvious.

(2) You have to specify & justify the assumptions underlying every hypothesis test.

Marks will not be given for a hypothesis test, if the underlying assumptions are not valid w.r.t. the data.

(3) Explain the R output w.r.t. your problem.

You will lose marks, if you do not interpret (correctly) the hypothesis test results.

* For every graph:

(1) You will lose marks, if your graph does not give any useful information.

Do NOT plot multiple graphs (with different topics) for one set of data points, unless

they each gives additional and useful information.

(2) You should appropriately customize the axes, gridlines, colors, fonts, labels, and legend.

(3) You have to interpret your graphs.

**How to submit**

You need to submit two files:

1. your project report, saved as a “.pdf” file,
2. your R code, saved as a “.R” file.

To submit your files, click the title “8002 Project”, scroll down to the bottom, and you will find the submission area. You can resubmit as many times as you want before the deadline.

You can defer your 8002 assessment to the re-assessment session (in August next year), which will be counted as your first attempt. However, the assessment type of 8002 in the re-assessment session is an exam, not a project. To apply for deferral, the contact person is Simon Wilson <SWILSON@tcd.ie>.