



PSY403
Statistics and Data Analysis

Tutor-Marked Assignment 01
July 2022 Presentation

PSY403

TMA01

This tutor-marked assignment (TMA) is worth 20% of the final mark for PSY403 Statistics and Data Analysis.

You are to include the following particulars in your submission: Course Code, Title of the TMA, SUSS PI No., Your Name, and Submission Date.

Please upload this assignment to Turnitin by Friday, 26 August 2022, 11:55pm. Resubmissions are allowed before this cut-off time.

It is strongly recommended that you make an early submission to check the originality report and, if necessary, make amendments to your document for resubmission. Note that the Turnitin report is usually generated immediately after the first submission, however, subsequent reports may take up to one day to generate. Do note that Turnitin will not accept any further submissions **AFTER** the cut-off time. There is a 12-hour grace period after the cut-off time, which is not an extended deadline but solely meant for solving any technical problems that you may encounter while attempting to make a submission before the cut-off time. Please email Canvas Support immediately (with relevant screenshots and your TMA attached) and follow up with Canvas Support first thing in the morning to ensure that the problem is resolved before the grace period is over.

One late submission is allowed only if no prior submissions were made before the cut-off time. Do note that the Canvas system will automatically deduct penalty marks for every day that your assignment is late. With this automatic deduction, there will be no need to request for extensions from your tutor because your tutor does not have the mandate to over-ride the Canvas system settings. You will need to form your own judgement as to how many marks you are willing to forego for each extra day that you gain to work on your assignment.

Take care to ensure that you upload the **correct TMA document** to the **correct folder** of the **correct course**. Requests to transfer incorrectly uploaded documents to the correct folder will require an official appeal (and an administrative fee).

Backup your TMA at all times. Once you have uploaded your TMA (in Word document format only), retain the Turnitin digital receipt as evidence of a successful submission. View your submission to ensure that the entire document has been uploaded successfully.

Plagiarism and Collusion

The assignment is to be completed on your own. You may discuss the TMA with your course-mates; however, the assignment must be written independently. Do not share your notes, draft or final TMA with anyone before the marked TMAs are returned to you.

Avoid plagiarism by giving yourself sufficient time to research and understand the material so that you can write up your assignment in your own words. Quotations should be used sparingly. Simply citing the source of 'copied' chunks of text does not excuse it from plagiarism. Do ensure that any paraphrasing is done appropriately, even if you use text from your own work that you have submitted as part of another assignment for the same or another course.

The University takes a very serious view of plagiarism (passing off someone else's ideas as your own, or recycling of contents from your own earlier marked TMA from the same course or another course) and collusion (submitting an assignment which is the same or very similar to another student's). Both are very serious academic offences. Please refer to the Student Handbook on the penalties of plagiarism or collusion. You are strongly advised to submit your TMA early, check the plagiarism report yourself, and if needed revise and resubmit your TMA before the submission deadline.

Question 1 (Total 50 marks)

You are a data analytics student and recently joined 9Move, a ride hailing company based in New York City, as a data science intern. The ride hailing business is highly volatile and 9Move's business could be affected by several factors such as climate conditions and events such as holidays. You have been given 6 months' worth of data (RSS503_TMA_ride.csv) and your job is to analyse the dataset and extract insights on the demand for rides in order to help 9Move grow its business. The dataset contains the following variables:

- pickup_dt: Date and time of the pick-up.
- borough: NYC's borough.
- pickups: Number of pickups for the period.
- spd: Wind speed in miles/hour.
- vsb: Visibility in miles to the nearest tenth.
- temp: Temperature in Fahrenheit.
- dewp: Dew point in Fahrenheit.
- slp: Sea level pressure.
- pcp01: 1-hour liquid precipitation.
- pcp06: 6-hour liquid precipitation.
- pcp24: 24-hour liquid precipitation.
- sd: Snow depth in inches.
- hday: Being a holiday (Y) or not (N).
- start_year: The year of the pick up.
- start_month: The month of the pick up.
- start_hour: The hour of the pick up (24 hours)
- start_day: The day of the pick up (e.g. the 15th day of the month).
- week_day: The day of the week of the pick up.

At the end of your internship, you are to produce a report to the senior executives of 9Move, not exceeding 1,500 words, to discuss the company's business.

For your report, create data analytical outputs using an exploratory data analysis (EDA) to understand the data. The exploratory data analysis can contain summary statistics and data visualisations. Data visualisations should be constructed carefully to provide insights into the business. Data wrangling work, such as the creation of new variables that can generate new insights, is encouraged.

To solve the issue of analysing demand for rides, apply data visualisations and statistical analyses, and develop a business plan to help the company grow its business. The business plan should contain your recommended business strategies informed by your data work.

The general structure of the report is up to you to decide. However, the report should minimally have an introduction and conclusion section. The introduction section may offer a brief discussion on the ride hailing business, what you have done in the report, issues you have identified and data insights, and brief business recommendations. In other words, think about the introduction section as an extended abstract or executive summary. Your conclusion should briefly wrap up the work you have done.

Between the introduction and conclusion sections, you may craft out your sections depending on how you would like the report to flow. These sections should contain a description of the dataset, discussion of any statistical work for understanding the data (e.g. summary statistics), discussions of your data visualisations, and business recommendations.

Your report should be readable and accessible to the intended audience (i.e. senior executives with limited knowledge in data analytics and statistics).

The mark distribution is as follows:

1. Data understanding and data visualisations (30 marks)

You should provide a clear discussion on your data checks and data exploration work using appropriate statistical tools. You should construct the appropriate data visualisations to help the senior executives understand the business. Data wrangling and preprocessing work should be carefully motivated and discussed. Data insights should be clearly communicated.

2. Quality of the discussion and report (20 marks)

The report should be easy to read, have a good flow and structure. The target audience are the senior executives of 9Move. Therefore, the report should be as accessible to a lay person as possible.

You may employ one or more software of your choice to perform this assignment. You will be assessed on the quality of your data work and report, not the software employed.

Question 2 (Total 50 marks)

Many factors affect life expectancy. Previous studies focused mainly on the effects that socio-economic and demographic variables may have on life expectancy. Recent studies have considered the importance of immunisation for lifespan, such as immunisation against Hepatitis B, Polio, and Diphtheria.

You are an advisor to the government of Kuva Republic, a developing country. The Ministry of Interior Affairs believes that an indication of the quality of life is life expectancy. Your task is to understand based on data from countries other than Kuva Republic (RSS503_TMA_life.csv) what the most important factors for life expectancy are. At the end of your analysis, you are to apply regression analysis and write a short research brief detailing your findings to the Ministry.

The dataset contains the following variables:

- Country: Country
- Year: Year
- Status: Developed or Developing status
- Life expectancy: Life Expectancy in years
- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Infant deaths: Number of Infant Deaths per 1000 population
- Alcohol: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
- percentage expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)

- Hepatitis B: Hepatitis B (HepB) immunisation coverage among 1-year-olds (%)
- Measles: number of reported cases of Measles per 1000 population
- BMI: Average Body Mass Index of the entire population
- under-five deaths: Number of under-five deaths per 1000 population
- Polio: Polio (Pol3) immunisation coverage among 1-year-olds (%)
- Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunisation coverage among 1-year-olds (%)
- HIV/AIDS: Deaths per 1000 live births due to HIV/AIDS (0-4 years)
- GDP: Gross Domestic Product per capita (in USD)
- Population: Population of the country
- thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling: Number of years of schooling

Write a report, not exceeding 1,200 words, to the Ministry providing insights into the determinants of life expectancy based on data from other countries. Assume that the audience has some background in statistics.

The structure of the report is up to you to decide, but it should minimally contain an introduction and conclusion section. Your report should contain exploratory data analysis and you should create regression models to study the issue of life expectancy. You should carefully evaluate the results from your regression analysis. If there are caveats, do explain them carefully. For instance, you may appraise issues such as multicollinearity and implement the appropriate test. You may also consider if your estimates could be biased, if so, in which possible direction. You may discuss other issues that are relevant to your analysis and the interpretation of your results.

1. Data understanding and regression analysis (30 marks)

You should provide a clear discussion on your data checks and data exploration work using appropriate statistical tools. Data wrangling and preprocessing work should be carefully motivated and discussed. Data insights should be clearly communicated. Regressions should be carefully motivated and interpreted. For example, you should explain your choice of regressors in the model, your modelling strategies, or subsamples you have selected for the regression analysis.

2. Quality of the discussion and report (20 marks)

The report should be easy to read, have a good flow and structure. It should provide clear insights into the factors that are important for life expectancy. As this is a research brief to be read by technocrats, you may afford to be more technical in the presentation of your research findings in the report.

You may employ one or more software of your choice to perform this assignment. You will be assessed on the quality of your data work and report, not the software employed.

— End of TMA —