Statistic for AI and Data Science

Coursework 1 (Version 1)

1 Introduction

This document outlines the requirements for coursework 1, which is based on notebook 2 and uses the London Underground exit data. The coursework tests whether you can:

- Manipulate data in a data frame using Pandas
- Review data, plotting distributions and analysing what is shown.
- Present your results as a Jupyter notebook, preparing a document for a reader interested in travel patterns (not computer code) using appropriate writing and formatting

The requirements are in three parts.

1.1 Part 1: Calculating Morning and Evening Peak Proportions

The work on notebook 2 showed that the distribution of exit times is bi-modal, with a morning and evening peak. It is common to divide the operation of the underground into 6 periods.

Period	Hours Included
Early	H05, H06
AMPeak	**H07, H08, H09**
Interpeak	H10, H11, H12, H13, H14, H15
PMPeak	**H16, H17, H18**
Evening	H19, H20, H21
Late	H22, H23, H00, H01

We will focus on only the morning and evening peaks (i.e. periods AMPeak and PMPeak).

Complete the following steps:

1. Create a data frame (called something like 'peak_counts') with station as the index and two columns, one for the AM peak count and one for the PM peak count.
2. Use the total exits for each station to create further columns (in the same or a different Data Frame):
    a. The AM peak count as a proportion of the daily total for the station
    b. The PM peak count as a proportion of the daily total for the station
3. Clearly describe all the data you have created in the new data frame.

Notes

- The total exits were calculated in section 3 of notebook 2. You can copy this code, together with code to load the data frame, from there.
- You do not need to use any Python loops to complete this.
- You can create an empty Data Frame using `pd.DataFrame()` and then use `.assign` to add the new columns
- You may find it easier to use several steps rather creating the new data frame all at once.

1.2 Part 2: Plotting and Analysing Distributions

The aim of this section is to analyse the distributions of the data so that a reader understands more about the patterns of travel on the Underground. You should complete the following steps:

1. Use `.describe()` to generate the statistics of the distributions and also plot histograms of the two proportions.
2. Review and comment on what the results of step 1 show about travel patterns, considering:
   - The statistics of the distribution (e.g. the median, quartiles and range)
   - The shape of each distribution: is it symmetric or skewed.

Notes

- Your comments should be about patterns of travel, not just the shape of graphs. However, the shapes are quite complex and not easy to explain in full. You are not expected to be familiar with the geography of London
- Example: suppose that the distribution of the AMPeak proportion data showed that most stations have a high proportion of exits in the AM Peak, then you could explain that this appears to suggest that there are lots of stations that people go to in the morning.

1.3 Part 3: A Simple Classification of Stations

The aim of this section is to classify stations into mainly 'work', mainly 'residential' and 'other' and complete a simple evaluation. The classes are to be defined as follows:

1. Work: have (significantly) ‘more’ exits in the AM peak than in the PM peak
2. Residential: have (significantly) ‘more’ exits in the PM peak than in the AM peak
3. Other: the number exits in the AM and PM peaks are approximately equal

Decide how to implement ‘more’ (a ratio or a difference perhaps) and choose suitable thresholds, with a brief justification. Report your classification of the following station (all on the Northern Line).

Archway, Balham, Embankment, Goodge Street, Highgate, High Barnet, Leicester Square, Morden, South Wimbledon, Tottenham Court Road, Warren Street

Briefly explain whether these results are what you expect: see maps of the underground at https://tfl.gov.uk/maps/track/tube if you wish. Again, you are not expected to be familiar with the geography of London but the map shows which of the stations are at the ends of the line or the nearer the centre.

Note that the suggested rules are very simple and they are not expected to classify stations very accurately.

2 Submission Requirements

The following additional requirements are about how your work should be submitted.

1. You must submit a single .ipynb' file only.
2. The notebook must be executable without errors. It must read the original data file of exits (from the same directory as the notebook); the data file must not be changed. Rerun the complete notebook before submission, so that the cells are executed in order.
3. The notebook must be readable. Markdown cells should be used to organised the notebook with a title and section headings. The code cells should be short, alternating with text cells (using markdown). The markdown text should be written to a 'domain expertise' interested in how the data is being manipulated (rather than in how the code works).
4. You can use material from notebook 2 (e.g. code to load the CSV), but do include code that is not relevant to the requirement above.

3 Mark Scheme

The following table shows the mark scheme.

Section	Weighting	Criteria	Detailed Criteria
All	20%	Presentation of the document	The notebook has a clear structure, with a title and sections, all in the style of the notebooks provided on the module
			Document includes markdown text cells interleaved with code, suitably formatted. Writing addresses a 'domain expert' – a reader interested in transport patterns
All	20%	Correctness and clarity of the code	The notebook shows the code executed in order without errors. Code runs when all cells are executed in order
			Code organised in short segments, alternating with text explaining the operations on data. All the code presented in the notebook is needed. Appropriate use of library code (e.g. pandas), avoiding unnecessarily complex code
Part 1	20%	Correct calculation of the proportions	Correct construction of the data frame, including the proportions.
			Clear and complete description of the data.
Part 2	20%	Use and interpretation of histograms and statistics	Statistics and histograms correctly generated. Graphs are readable and clearly labelled
			Review and comments on travel patterns clear and reasonable
Part 3	20%	Classification of stations	The implementation of the classification is clear, with the classification clearly explained. The thresholds used have reasonable justifications.
			The classification scheme is applied to the specified Northern Line stations

4 Feedback

No feedback is provided until after the last hand-in date (i.e. no sooner than one week after the deadline, perhaps later). The following forms of feedback will be provided:

1. A sample answer.
2. General review comments on the main strengths and weaknesses for the class as a whole (written or in a review lecture)
3. The grades obtained by your answer on each of the detailed criteria. Clarifying comments may be added.

5 Available Resources

Notebooks 1 and 2 should provide examples of all the necessary forms of code. You are welcome to ask questions to clarify the requirements for the coursework or ask for guidance on the principles.