**Objective:** Analyse data from total spend data for 2010 – 2023

Provide an overall summary

Identify anomalies

**Extract from e-mail**:

*“In advance of the interview, you will need to analyse data on MPs expenses and prepare a presentation, assuming the audience is an internal audit team who have requested you analyse this data and share your findings.*

*The audit team has requested you analyse the “Total spend” data from the IPSA website, providing an overall summary as well as drawing their attention to anything of note. You may find this guide to MP’s expenses useful.*

*You will spend 10 minutes presenting your work at the start of the interview, followed by 5 minutes for questions from the interview panel. You will need to send your slides (or equivalent) and source files (e.g. workbook or code) to GIAARecruitment@giaa.gov.uk by 9am on Thursday 30 May.”*

**Understanding budget discrepancies (**[A guide to MPs’ staffing and business costs | IPSA (theipsa.org.uk))](https://www.theipsa.org.uk/a-guide-to-mps-business-costs)

There may be discrepancies between the total spent by an MP against their budget, as published in the “Total Spend” data for the financial year, and the total amount spent under the same budget, as calculated from the published individual claims for the MP.

These discrepancies occur for the following reasons:

While the total spend against each budget for each MP represents how much they spent under that budget in each financial year, not all individual claims that make up that total can be published. For example, some security or disability-assistance costs under the Office budget, Accommodation budget, or under “one-off costs”, will be counted in the published total spend, but will not be released on the individual claims data.

Some accruals and prepayments made around the end of a financial year may also have been published as individual entries in the former or previous financial year.

**Code steps (number denotes script)**

1. Import and construct a list for the years of data and continue until all are read
   1. Initialise variables for start of the year and create list
   2. Function to contruct url and to input start and end year placeholders
   3. While loop constructed to continue indefinitely until it encouters an error (no more data sets left)
      1. Inside the while loop, the url is constructed which takes the start and end year as inputs
      2. After data is imported, it creates a label for that year and adds a new column for each year
      3. A message is printed to indicate a successful import
      4. The start and end year of the loops are incremented by 1
2. Data Cleansing
   1. Replace symbols in financial variables (£ ,””) and if variable values have \n in it remove all values after and including the \n from the values
   2. Convert character variables to double if they contain predominantly numerical values and round numerical values to nearest pound
   3. Rename variables –
      1. Office maximum budget available = Office budget
      2. Staffing maximum budget available = Staffing budget
      3. Accommodation maximum budget available = Accommodation budget
      4. Travel and subsistence spend = Travel and subsistence (uncapped)
      5. Other costs spend = Other costs (uncapped)
      6. Winding-up spend = Wind-up spend
      7. Wind-up maximum budget available = Wind-up budget
      8. Winding-up budget = Wind-up budget
   4. Include start-up spend if it doesn’t exist on the datasets
   5. Replace missing values with 0’s for these variables
      1. Start-up spend
      2. Wind-up spend
      3. Accommodation budget
      4. Accommodation spend
   6. Derive variables:
      1. Derive ‘Subtotal of office running costs’ for 2022/23 (office spend + staff spend).
      2. Derive ‘Subtotal of other parliamentary costs’ for 2022/23 (Accommodation spend + Travel and subsistence spend)
      3. Derive ‘Overall total spend for this financial year’ for 2022/23 (Office spend + Staffing spend + Accommodation spend + Travel and subsistence (uncapped) + Other costs (uncapped))
   7. Derive ‘Office budget’ for year 2011/12 - all values are 0
   8. Replace negative values with positive values for variables with total or spend in their names
   9. Drop these variables:
      1. Drop all variables detailing ‘reason’ for budget
      2. Travel and subsistence maximum budget available
      3. Other costs maximum budget available
   10. Clean column names using Janitor package (issue with naming conventions)
   11. Bind all years to be in one dataset
   12. Replaces instances of missing values for numerical values with 0
   13. Final dataset (before data visualisations) named “processed\_data” and final list of variables (data set presents them in ‘cleaned’ format):
       1. MP’s name
       2. Constituency
       3. Office budget
       4. Office spend
       5. Remaining office budget
       6. Staffing budget
       7. Staffing spend
       8. Remaining staffing budget
       9. Subtotal of office running costs
       10. Accommodation budget
       11. Accommodation spend
       12. Remaining accommodation budget
       13. Travel and subsistence (uncapped)
       14. Other costs (uncapped)
       15. Subtotal of other parliamentary costs
       16. Overall total spend for this financial year
       17. pID
       18. Year
3. Test processed data against original imported data (in case processing caused unexpected changes)
   1. Set options to display values with non-scientific notation
   2. List variables to be selected
   3. Clean column names and remove symbols (£, etc.)
   4. Output unprocessed and processed data sets by year and select only specific variables
   5. List the years of interest
   6. Define the columns to select
   7. Function to compare totals and then subtracting totals to find absolute difference
      1. List the years
      2. Constructs the name of the two data sets (processed and unprocessed
      3. Checks both exist
      4. Selects the columns for both data sets and combines into a single table using p\_id as the unique identifier
      5. Rounds values
      6. Calculates absolute different between processed and unprocessed with the same names (except the .x and .y)
      7. Stores the data in a list using the year as the key
      8. Skips years if p\_id or a particular year is missing
4. Data visualisations for totals
   1. Rename, select, and reorder variables for data visuals
   2. Rename year
   3. Summarise total spend by year
   4. Prepare data for forecasting
   5. Fit ARIMA model
   6. Forecast for the next 3-years
   7. Prepare forecasted data for plotting
   8. Combine the past and projected data
   9. Calculate percentage change
   10. Create bar plot for total data, superimpose line, and place percentage change on top of the bars
   11. Similar steps for visuals 2 & 3 but without forecast
   12. Again, similar steps for remaining visuals but stacked bar chart for budget vs spend
5. Data visualisations for averages
   1. Select relevant columns
   2. Summary statistics calculation
   3. Convert to long format
   4. Create bar plots for each category
   5. Create box plots for each category
6. Data visualisations for outliers
   1. Select data
   2. Convert data to long format
   3. Identify outliers (standard IQR method)
   4. Create plots
   5. Select column
   6. Calculate years of service
   7. Merge years of service
   8. Summarise counts for different combinations
7. Benford’s law analysis
   1. Calculate Benford’s law probabilities
   2. Filter 0’s
   3. Check assumptions
   4. Plots for Benford’s law

**FINAL LIST OF VARIABLES** (after processing code – step 3):

1. MP’s name
2. Constituency
3. Office budget
4. Office spend
5. Remaining office budget
6. Staffing budget
7. Staffing spend
8. Remaining staffing budget
9. Wind-up Spend
10. Remaining wind-up budget
11. Subtotal of office running costs
12. Accommodation budget
13. Accommodation spend
14. Remaining accommodation budget
15. Travel and subsistence (uncapped)
16. Other costs (uncapped)
17. Subtotal of other parliamentary costs
18. Overall total spend for this financial year
19. pID
20. Year
21. Start-up maximum budget available
22. Start-up spend
23. Remaining start up budget
24. Winding up budget

**Additional Documents:**

[A guide to MPs’ staffing and business costs | IPSA (theipsa.org.uk)](https://www.theipsa.org.uk/a-guide-to-mps-business-costs) **-** Further Information related to task (minimum and maximum claim amount, transaction dates, negative values, staffing options, overspending and underspending, budget discrepancies, how claims are checked, thematic reviews, accommodation costs).

[Annual budgets, costs and claims | IPSA (theipsa.org.uk)](https://www.theipsa.org.uk/mp-staffing-business-costs/annual-publications) - Total Spend Data Source.

[IPSA – The Scheme of MPs’ Staffing and Business Costs 2024-25 (ctfassets.net)](https://assets.ctfassets.net/nc7h1cs4q6ic/5S1XNi2m1uVDWyEqpx3rdv/2afac427cd87da2aa504260951fa43b6/Seventeenth_edition_of_the_Scheme_2024-25.pdf) - What they can and can’t claim (page 12 – repayment of money to IPSA

**RAP:**

[Quality assurance of code for analysis and research — Quality Assurance of Code for Analysis and Research (best-practice-and-impact.github.io)](https://best-practice-and-impact.github.io/qa-of-code-guidance/intro.html) - Best Practice for RAP.

**Shiny**

[orgsurveyr/R at master · ukgovdatascience/orgsurveyr (github.com)](https://github.com/ukgovdatascience/orgsurveyr/tree/master/R) - RAP and R Shiny Example

**Previous work on MP’s expenses:**

[MP\_expenses\_main\_report.pdf (publishing.service.gov.uk)](https://assets.publishing.service.gov.uk/media/5a7dd58bed915d2acb6ee6a9/MP_expenses_main_report.pdf)

[PSPE-WP9-12.pdf (lse.ac.uk)](https://www.lse.ac.uk/government/Assets/Documents/pdf/research-groups/pspe/working-papers/PSPE-WP9-12.pdf)

**Auditing Analytics:**

[About the author | Audit Analytics with R (jonlin.ca)](https://auditanalytics.jonlin.ca/about-the-author.html) - Analytics for Auditing (standard cleaning data, data visualisation for auditing, anomalies, testing, machine learning) ([GitHub - jonlinca/auditanalytics: Source for book Audit Analytics with R)](https://github.com/jonlinca/auditanalytics)