

Analytics Project Goal: create something useful and helpful.		
<i>Your analysis should discover or showcase some useful findings to provide insightful data into specific questions.</i>		
Prep	Create Slack channel & add marker	SK 13th Nov
	Choose project management interface	GitHub projects (decided in meeting 13th Nov)
	Create Project, Breakdown tasks & add to project manager	TN - 18th Nov
	Create a GitHub Repo, add marker & CFG	TN - 18th Nov
Step 1: Frame the problem	The project objective(s) is/are....	Contained in HW2
Data availability will dictate topic!	The project performs the useful / helpful function of...	
	The project addresses the following questions / problems	
<i>Project creates something useful / helpful</i>	Analysis Q1:	
	Analysis Q2:	
<i>Objectives must be clearly defined</i>	Analysis Q3:	
	Analysis Q4:	
<i>Formulate exact questions that we are trying to answer/ Formulate exact problems that we are trying to solve</i>	The project addresses these questions / problems by providing the following insightful data	
	The project will discover / showcase the following useful findings ...	
	The chosen analysis topic is...	
	The analysis topic is relevant to the project objective because....	
Step 2: Collect the raw data	Data gathering: series of data captured for a specific time period, relevant records, values etc.	
	Where are you going to get data for your analysis?	Teleport for salary data, exchange-rate for currency, plus other adjacent sources for parallel analysis variables
	Use at least one API to fetch data	Teleport (x3 calls), exchange-rateapi.com
	Will you also download CSV or XLSX from somewhere?	Gender parity info (OECD), geonames country info, cost of living (World Data)
	How many different resources were used?	5 (or 4 if World Data cost of living stuff isn't used)
<i>Crucial: find suitable, relevant datasets (online sources, data libraries, APIs etc)</i>		All API data interwoven through the python code into one output csv. Then, this was joined/merged with other data sources using pandas in Jupyter Notebook part 2. Data was further normalised through SQL database creation.
<i>Data sources utilised in your research and analysis</i>	Combine resources to expand data series	

	The datasets are relevant because...	Salary information comprehensively provided for 52 jobs across 198 countries
	The dataset is suitable because...	As above (comprehensive and consistent variables) plus we can convert the salary information from the other Teleport data provied (i.e. currency code) using currency exchange rate data to translate the figures into those a UK analyst or audience can relativise.
Step 3: Prepare the data for analysis <i>Did students deal with missing or skewed values?</i>	Recommended but optional? Build a DB with data arranged in tables	Completed
	Clean the data	Records without salary data excluded (i.e. AQ Antarctica), three mismatched currency codes BYR, VEF, MRO corrected and one problematic country code (NA Namibia manually corrected). Corrected this manually in database but also implemented fixes in the python code and jupyter notebook directly to elimatate errors
	Enrich data with required additional values etc.	Salary data (including GBP conversion) enriched with paralell data about countries (population, area (i.e. population density metric)), currency code, gender pay parity metric and cost of living metric from World Data. All these were combined into a large dataframe that could then be split out according to the metrics which were desired for comparison and analysis.
	Correct handling anomalies: deal with missing / skewed values	See above
	<i>Correct handling anomalies: How did they solve these issues?</i>	Explained in analysis visualisation section, did specific visualisations (box plots and heat maps and distributions) to explore this
Code: Steps 3, 4, 5, 6 <i>All members must contribute some elements of code.</i>	Python code script in Jupyter Notebook <i>Has clear structure & defined sections</i> <i>Demonstrates clearly defined and accomplished stages of data sourcing, pre-processing, evaluation and visualisation.</i>	Section: loading data - Jupyter Notebook Part 1 and Part 2
		Section: cleaning data - Jupyter Notebook Part 2
		Section: transforming data - Jupyter Notebook Part 2
		Section: data analysis - Jupyter Notebook Part 2
		Section: data visualisation - Jupyter Notebook Part 2
		Section: data reporting - Jupyter Notebook Part 2
		Use at least one API to fetch data
		Effective use of Python - - Jupyter Notebook Part 1 for data retrieval, transforming, processing and combination.
		Jupyter Notebook Part 2 uses Python to combine data (pandas), clean data, refine into subsets and matplotlib and numpy to analyse and visualise the data
		Effective use of key scientific packages: Pandas , NumPy, MatplotLib

	Bonus code elements: Yep, all of them!	Verify: an instructor or assessor can run submitted scripts with the files provided and get expected results . 1. Machine Learning 2. SciKit Library 3. SQL database
Step 4: Explore the data	Review, understand the data, and perform summary statistics with descriptive analysis . <i>What evidence did the team present to support their findings?</i> The core questions have been answered as follows...	In Jupyter Notebook Part 2 and summarised in Canva documentation
Step 5: In-depth analysis <i>Credibility of research / analysis conclusion.</i>	Usually Machine Learning, AI analysis using regression, mathematical model building -- predictive analysis.	In Jupyter Notebook Part 2 and summarised in Canva documentation
Step 6: Communicate results	data visualisation findings summary interpretation of results recommendations conclusions	Effective use of Matplotlib to present findings All in Jupyter Notebook 2
Step 6: README File	Contains clear instructions how to execute the code. An instructor or assessor should be able to run submitted scripts with the files provided and get expected results.	README completed and uploaded to GitHub. All team members participated in testing both notebooks rigorously and contributed to collecting a list of dependencies.
Step 6: Project Documentation	Project Document: Overview	Concise yet detailed - explain every key point. Max 5-7 A4 pages long Include diagrams, and images with descriptive captions.
	Project Document: INTRODUCTION	<ul style="list-style-type: none"> ● Aims and objectives of the project ● Roadmap of the report
	Project Document: BACKGROUND	<ul style="list-style-type: none"> ● Any specific details about the project based on your chosen topic.
	Project Document: SPECIFICATION & DESIGN	<ul style="list-style-type: none"> ● Requirements technical and non-technical ● Design and architecture
	Project Document: IMPLEMENTATION AND EXECUTION	<ul style="list-style-type: none"> ● Development approach and team member roles ● Tools and libraries ● Implementation process (achievements, challenges, decision to change something)

Used to assess your project work and understand your approach to the project delivery. Insight into your architecture, testing and implementation strategy as a team.		<ul style="list-style-type: none"> ● Agile development (did your team use any agile elements like iterative approach, refactoring, code reviews, etc) ● Implementation challenges
	Project Document: DATA COLLECTION	<ul style="list-style-type: none"> ● What information do you need? ● What information is available? ● What is your data source? ● Describe how you collected the data (e.g., if you have used an API, briefly describe it).
	Project Document: CONCLUSION (Evaluation)	Evaluation of Project (confirmed by Stavros). Limitations of project / data, what went well, what didn't go well (what we would change if we did it again) what we would do if we could extend the project / Part 2 (further research options) comments about any lack of resources / knowledge which may have assisted, etc
Step 6: Project Group Activity Log	Individuals record their own work	
	Collate all tabs into a single pdf at the end	
Delivery	Deadline project: 11.59pm, Sunday the 17th December 2023.	
	Share presentations with your instructors shortly before Thurs 21st Dec.	
Step 6: Presentation (Thurs 21st Dec)	Produce Powerpoint Slide Deck	
	Project to be delivered in a group	
	Should last around 2-3 minutes , max 5 mins.	
	Use slide deck or similar (e.g. Google Slides, Prezi, Canva)	
Submission Elements	Include project demo where applicable.	
	1. A Project Document (PDF preferably) which reports on project work with clear project specification	
	2. Source code for the project (Jupyter notebook file)	
	NB: share via GitHub with instructor and https://github.com/CFGer	
	3. Separate data files : csv, xlsx, txt etc. Provide a file where your API fetched data is saved.	
	4. README file with clear instructions how to execute the code. An instructor or assessor should be able to run submitted scripts with the files provided and get expected results.	

	5. Comined individual Project Activity Log (xls) - 1 tab per person. Share with your assigned marker via Slack as a PDF!!	
	6. PPT slide deck (2-3 min length max) with key points for presentations. Share with your instructors shortly before Thurs 21st Dec.	
		Completed
		Needs to be done

[illegible]

Link	Title / Source	Area	Topic	Comments	Added
https://data.oecd.org/price/housing-prices.htm#indicator-chart	OECD	Housing Prices	Housing Prices per country		SK
https://data.oecd.org/earnwage/gender-wage-gap.htm the salary wage gap in % from 1970 to 2022 per country if we want to use it	OECD	Gender wages	Gender Wage Gap by country. csv contains data over time		AMG
https://data.oecd.org/earnwage/wage-levels.htm#indicator-chart	OECD	Wage levels	Wage Levels (high low) by country. csv contains data over time		AMG
https://data.oecd.org/earnwage/average-wages.htm#indicator-chart	OECD	Wage levels	Average wages by country. csv contains data over time		AMG
https://data.oecd.org/price/price-level-indices.htm#indicator-chart	OECD	Price levels relative to GDP	Price level indices: ratio of country purchasing power relative to market exchange rates.		SK
https://www.geonames.org/countries/	Geonames country codes	Country stats	Country codes (used by Teleport) and basic country stats		SK
https://www.worlddata.info/cost-of-living.php	World Data	These are different	Cost of Living 'Index'. Comparison chart of worldwide cost of living		SK
https://www.worlddata.info/quality-of-life.php?expats=0&stability=1&rights=1&health=1&safety=1&climate=1&costs=50&popularity=1#ranges	World Data	angles on the same data about cost of living per country	Adjustable weighting chart of worldwide cost of living. Can just use 'costs & income' as the only ranking factor. Could potentially use this to contrast with our per country/salary data		SK
https://teleport.org/cities/ https://developers.teleport.org/api/getting_started/ https://developers.teleport.org/api/reference/#/ https://developers.teleport.org/api/resources/	Main API Datsource Teleport	Country & City data, esp salaries for common jobs and 'quality of life'	<u>Country Level Info</u> https://api.teleport.org/api/countries/ - lists 252 countries, won't have information on all of them https://developers.teleport.org/api/resources/CountrySalaries/ - has salary percentile information for lists of different job		
Discarded Ideas / Sources					
NVD & CVE https://nvd.nist.gov/developers/vulnerabilities Sample CVE query via API https://services.nvd.nist.gov/rest/json/cves/2.0/?pubStartDate=2023-10-16T00:00:00.000Z&pubEndDate=2023-11-15T23:59:59.999Z&resultsPerPage=2000&noRejected CWE https://cwe.mitre.org/about/new_to_cwe.html	National Vulnerability Database (NVD) Common Vulnerability & Exploits (CVE) Common Weaknesses Enumeration (CWE)	Cybersec	See Slack post 1: https://autumncfgdegree2023.slack.com/archives/C06507J102K/p1700045181058639?thread_ts=1699956459.041179&cid=C06507J102K See Slack post 2: https://autumncfgdegree2023.slack.com/archives/C06507J102K/p1700127853337249?thread_ts=1699956459.041179&cid=C06507J102K		SK
https://ico.org.uk/action-weve-taken/data-security-incident-trends/	ICO DataSecurity Incident trends	Cybersec	Advantages Reputable source: UK Information Commissioners Office Great relevance: UK specific, cybersecurity Recent! 2019 - 2023. Recently updated 1st Nov 23! Analysis - scope: Raw dataset - not yet analysed in form of a report - scope for us to actually find some observations- Damn, they have produced an interactive display already using the data. We'd need to be able to say something different. Israel: 119000 rows of data!		SK
https://www.kaggle.com/code/benciaiello/cve-dataset https://www.kaggle.com/datasets/katehighnam/beth-dataset		Cybersec	CVE dataset. Wonder if we could cross reference this with MitreAtt&ck		
https://www.unb.ca/cic/datasets/index.html	Canadian Institute of Cybersec data resources	Cybersec, multiple datasets			SGH
https://www.worldbank.org/en/publication/globalindex#sec1	The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19	Global finances	Finance data by country, banked /unbanked, potential relevance to digital payments / crypto. Very large, detailed dataset, worldwide Reputable source		SK

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/datasets/healthstatelifeexpectancyallagesuk	UK Health Dept, life expectancy	Health			SHS
https://www.reddit.com/r/datasets/s/VZl3ZsZ64		Collection of datasets			SGH
https://vickel.notion.site/vickel/a360dea317234868a0f7cfb1ef249843?v=2923f79780214ec390b9f0fefbc1c002		Collection of datasets			SGH
https://opendata.nhsbsa.net/dataset/english-prescribing-data-epd	English Prescribing Dataset	Pharma			NP
https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/drugmisuseinenglandandwalesappendixatable	Drug Misuse England and Wales	Pharma	Data from 1995-2022 on drug misuse with demographics		NP
https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/natureofcrimefraudandcomputermisuse	ONS	Cybersec	UK reported cybercrime data (possibly could be merged with data from row 2)		NP
https://www.kaggle.com/datasets/thedevastator/chemicals-that-may-contribute-to-disease		Chemicals / Pharma	Chemicals That May Contribute to Disease		AM
https://www.kaggle.com/datasets/thedevastator/chemicals-in-cosmetics-what-s-really-in-your		Chemicals / Pharma	Chemicals in Cosmetics: What's Really in Your?		AM
https://www.cisa.gov/known-exploited-vulnerabilities-catalog	CISA	Cybersec	Known Exploited Vulnerabilities Catalog		AM
http://data.un.org/	UN	Various	Big variety of topics regarding population, gender or energy among others per country		AM
https://www.fao.org/faostat/en/#data	FAO	Various	Big variety of topics per country as Production, Food Security and Nutrition, SDG indicators, Food Balances, Trade, Prices, Cost and Affordability of a Healthy Diet, Food Value Chain or Climate Change: Agrifood systems emissions		AM
https://digital.nhs.uk/developer/api-catalogue	NHS Digital, API Catalogue	Health	117 API's related to NHS data		SHS
https://biobank.ndph.ox.ac.uk/showcase/	UK BioBank	Health	large-scale UK biomedical database and research resource		SHS
https://digital.nhs.uk/data	NHS Data	Health	NHS Data on health conditions, hospitals, patients, reasons for admissions etc		SHS
https://docs.google.com/document/d/1875QHePuWjo_4xl28B_wDs86FZgkcssG/mobilebasic	Google Sheet	Various	Websites with Data for Projects		TN
https://www.unb.ca/cic/datasets/index.html	UNB	Cybersecurity	Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers		TN
https://explore.data.parliament.uk/	Parliament UK	Political	Election data (API works)		NP
https://alpaca.markets/data	Market Data API > navigate to 'developers'	Various	Alpaca Data API is your new go-to stock data API for building trading apps & algorithmic		TN
https://developers.google.com/apis-explorer/#p/	Google API Explorer >	Various	There seems to be loads of options here > The Google APIs Explorer is a tool that helps you explore various Google APIs		TN
https://publicapis.dev/	Public API	Various	With quite a lot of API options and we can go through it > A collection of public APIs for developers, categorized and crowdsourced. Animals, books, cryptocurrencies, development, music, weather and much more. {		TN

9

Section	Section Weighting	Sub-section	Marks Available	Overall Weighting
Project Objectives & Final Result <i>how well did the project meet the original objectives?</i>	10%	Final results vs objective set	5	5.00%
		Performance of the final product	5	5.00%
Code Implementation <i>how well written and organised is the code?</i>	50%	Use of Python and key analytical libraries (NumPy, Pandas, Matplotlib)	10	9.10%
		Defined and accomplished data sourcing, pre-processing and evaluation	10	9.10%
		Code layout & readability	5	4.50%
		Relevance of chosen topic	5	4.50%
		Data sources utilised	5	4.50%
		Visualisation of data	10	9.10%
		Handling of anomalies	5	4.50%
		Credibility of research and analysis conclusions	5	4.50%
Project Documentation <i>is there a clear intro, background, discussion and conclusion?</i>	30%	Introduction	10	5.00%
		Background	5	2.50%
		Specifications and Design	15	7.50%
		Implementation and Execution	15	7.50%
		Result Reporting	10	5.00%
		Conclusion	5	2.50%
Group Project Presentation <i>how well did you present as a group? Was it clear and concise? Did you keep the audience engaged and answer questions effectively?</i>	10%	Evidence of Teamwork	5	3.30%
		Presentation Skills	5	3.30%
		Understanding of the Project	5	3.30%