Analytics Project Goal: create something useful and helpful.

Your analysis should discover or showcase some useful findings to provide **insightful data** into **specific questions.**

Your analysis should discover or showco	sse some useful findings to provide insightful data into specific questions.	
	Create Slack channel & add marker	SK 13th Nov
D	Choose project management interface	GitHub projects (decided in meeting 13th Nov)
Prep	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	
Step 1. Traine the problem	The project performs the useful / helpful function of	
Data availability will dictate topic!	The project addresses the following questions / problems	
	Analysis Q1:	
Project creates something useful /	Analysis Q2:	
helpful	Analysis Q3:	Contained in LIW2
Objectives must be clearly defined	Analysis Q4:	Contained in HW2
Objectives must be clearly defined	The project addresses these questions / problems by providing the	
Formulate exact questions that we are	following insightful data	
trying to answer/	The project will discover / showcase the following useful findings	
Formulate exact problems that we are	The chosen analysis topic is	
trying to solve	The analysis topic is relevant to the project objective becauses	
	Data gathering: series of data captured for a specific time period, relevant records, values etc.	
Step 2: Collect the raw data	Where are you going to get data for your analysis?	Teleport for salary data, exchange-rate for currency, plus other adjacent sources for paralell analysis variables
YOU NEED TO FIND GOOD DATA	Use at least one API to fetch data	Teleport (x3 calls), <u>exchange-rateapi.com</u>
Crucial: find suitable, relevant datasets	Will you also download CSV or XLSX from somewhere?	Gender parity info (OECD), geonames country info, cost of living (World Data)
(online sources, data libraries, APIs	How many different resources were used?	5 (or 4 if World Data cost of living stuff isn't used)
etc) Data sources utilised in your research		All API data interwoven through the python code into one output csv. Then, this was joined/merged with other data sources using pandas in Juptyer Notebook part 2. Data was further normalised through SQL
and analysis	Combine resources to expand data series	database creation.

	The datasets are relevant because	Salary information comprehensively provided for 52 jobs across 198 countries					
		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					
	The dataset is suitable because	figures into those a UK analyst or audience can relativise.					
	Recommended but optional? Build a DB with data arranged in tables	Completed					
	Clean the data	Records without salary data excluded (i.e. AQ Antartica), 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 elimate errors					
itep 3: Prepare the data for analysis Did students deal with missing or kewed values?		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					
	Enrich data with required additional values etc.	split out according to the metrics which were desired for comparison and analysis.					
	Correct handling anomalies: deal with missing / skewed values	See above					
	Correct handling anomalies: How did they solve these issues?	Explained in analysis visualisation section, did specific visualisations (box plots and heat maps and distributions) to explore this					
		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					
	Duth on so do soviet	Section: data visualisation - Jupyter Notebook Part 2					
	Python code script in Jupyter Notebook	Section: data reporting - Jupyter Notebook Part 2					
	in Jupyter Notebook	Use at least one API to fetch data					
Code: Steps 3, 4, 5, 6	Has clear structure & defined sections	Effective use of Python Jupyter Notebook Part 1 for data retrieval, transforming, processing and combination.					
All members must contribute some elements of code.	Demonstrates clearly defined and accomplished stages of data sourcing, pre-processing, evaluation and visualisation.	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					

		Verify: an instructor or assessor can run submitted scripts with the files
		provided and get expected results.
		1. Machine Learning
	Bonus code elements: Yep, all of them!	2. SciKit Library
		3. SQL database
	Review, understand the data,	In Jupyter Notebook Part 2 and summarised in Canva documentation
Step 4: Explore the data	and perform summary statistics with descriptive analysis.	
Step 4. Explore the data	What evidence did the team present to support their findings?	
	The core questions have been answered as follows	
Step 5: In-depth analysis	Usually Machine Learning, AI analysis using regression, mathematical	In Jupyter Notebook Part 2 and summarised in Canva documentation
Credibility of research / analysis conclusion.	model building predictive analysis.	
	data visualisation	Effective use of Matplotlib to present findings
Stan C. Communicate vesselts	findings summary	All in Jupyter Notebook 2
Step 6: Communicate results	interpretation of results	
	recommendations	
	conclusions	
	Contains clear instructions how to execute the code. An instructor or	README completed and uploaded to GitHub. All team members
Step 6: README File	assessor should be able to run submitted scripts with the files provided and	, , , , , , , , , , , , , , , , , , , ,
	get expected results.	collecting a list of dependencies.
	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	Aims and objectives of the project
	Project bocument. INTRoboerion	Roadmap of the report
	Project Document: BACKGROUND	Any specific details about the project based on your chosen topic.
	Project Document: SPECIFICATION & DESIGN	Requirements technical and non-technical
	Project bocument. Steel learnon & besidn	Design and architecture
		Development approach and team member roles
		• Tools and libraries
Step 6: Project Documentation	Project Document: IMPLEMENTATION AND EXECUTION	• Implementation process (achievements, challenges, decision to change something)

Used to assess your project work and understand your approach to the project		• Agile development (did your team use any agile elements like iterative approach, refactoring, code reviews, etc)
delivery. Insight into your architecture,		• Implementation challenges
testing and implementation strategy		What information do you need?
as a team.		What information is available?
	Project Document: DATA COLLECTION	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
	Individuals record their own work	
Step 6: Project Group Activity Log	Collate all tabs into a single pdf at the end	
Dalissam	Deadline project: 11.59pm, Sunday the 17th December 2023.	
Delivery	Share presentations with your instructors shortly before Thurs 21st Dec.	
	Produce Powerpoint Slide Deck	
	Project to be delivered in a group	
Step 6: Presentation (Thurs 21st Dec)	Should last around 2-3 min utes, max 5 mins.	
	Use slide deck or similar (e.g. Google Slides, Prezi, Canva)	
	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.	
Submission Elements	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

							v	veek1			week 2		2			we	ek 3				w	eek 4				v	eek 5			wee	k 6
luday.	Task	Start date	Fmd d-4-	Depend	0	124		11014-	, , , , ,	10	20/24/			1.20	27/22				1. 2.10	4/	c /	/12 0	1 0/4	1140	1115				417	0/40	20.4
Index 0	Team approach preparation		End date	ancies	Owner	13/	14/ 15	/ 16/ 1/	// 18/	19/	20/ 21/	22/ 23/	24/ 25/	. 26/	2// 28	7 29/	30/ 1,	1 2/	3/1	4/ 5/	6/1/	/12 8/	1 9/1	10/	11/ 14	2/ 13/	/ 14/ 1	.5/ 16,	/ 1 / / 1	8/ 19/	20/ 2
•	Create Github repository																									-					
0.1	(HW2 Q4.3)	13/11/2023	18/11/2023		Tatiana				~																						
0.2	Create SWOT analysis (HW2 Q4.2)	13/11/2023	30/11/2023		Whole team	✓																									
0.3	Create project log (HW2 Q4. 1)	13/11/2023	16/11/2023		Tatiana			~																							
0.4	Create timeline/gantt chart (HW2 Q4.4)	25/11/2023	28/11/2023		Alicia										<u>~</u>																
1	Look for an appropriate dataframe	13/11/2023			Whole team																								G		P r
1.1	Decide API to use	13/11/2023	26/11/2023		Whole team								V	DD															r		0
1.2	Decide Dataset to use	13/11/2023	26/11/2023		Whole team									~															0		j
	Determine potential issues																												u		е
1.3	from the dataframe and how to overcome it (HW2 Q3)	24/11/2023	28/11/2023	1.1, 1.2											~		0					Н							р		t
2	Determine Problem and Solution to cover																n e	-				m e							p r		p
2.1	Problem and questions to tackle (HW2 Q1)	25/11/2023	28/11/2023	1	Tatiana / Sam H									Y	DI)	v.					w o							o j		r e
2.2	Target audience (HW2 Q2)	25/11/2023	28/11/2023	1	Alicia / Tatiana									/	DE)	r					r							е		S
3	Data collection																k					k							C		е
3.1	API connection in Python	25/11/2023	30/11/2023	1	Sam K / Nicola										\checkmark		DD												t		n
3.2	Set tables into SQL	25/11/2023	30/11/2023	1.2	Sam H										\checkmark		DD 2					3									t
4	Data processing																												С		а
4.1	Connect dataframes to jupyter	27/11/2023	3/12/2023	3	Sam K / Nicola													~	DD										d		t i
4.2	Data cleaning	1/12/2023	7/12/2023	4.1	Whole team																D	D	/						е		0
5	Data Visualization																														n
5.1	Define plots to show	7/12/2023	10/12/2023	4	Whole team																			~							
5.2	Write plots code in jupyter Notebook	9/12/2023	16/12/2023	5.1	Nicola / Sam K / Alicia																							DD			
6	Project Presentation																														
6.1	Project documentation	11/12/2023	17/12/2023	1 to 5	Whole team																							DD			
6.2	Create presentation	16/12/2023	21/12/2023	5	Whole team																										DD
6.3	Present!	21/12/2023	21/12/2023	6.2	Whole team																										DD
																										-					

Title / Source	Area	Topic	Comments	Added
OECD	Housing Prices	Housing Prices per country		SK
OECD	Gender wages	Gender Wage Gap by country. csv contains data over time		AMG
OECD	Wage levels	Wage Levels (high low) by country. csv contains data over time		AMG
OECD	Wage levels	Average wages by country. csv contains data over time		AMG
OECD	Price levels relative to GDP	Price level indices: ratio of country purchasing power relative to market exchange rates.		SK
Geonames country codes	Country stats	Country codes (used by Teleport) and basic country stats		SK
World Data	These are different	Cost of Living 'Index'. Comparison chart of worldwide cost of living		SK
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
Main API Datsource Teleport	Country & City data, esp salaries for common jobs and 'quality of life'	Country Level Info 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		
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
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. Largel: 11900 rows of data!		SK
	Cybersec	CVE dataset. Wonder if we could cross reference this with MitreAtt&ck		
	Cybersec			
Canadian Institute of Cybersec data resources	Cybersec, multiple o	latasets		SGH
		Finance data by country, banked /unbanked, potential relevance to digital payments / crypto.		
	OECD OECD OECD OECD OECD OECD Geonames country codes World Data World Data World Data Main API Datsource Teleport National Vulnerability Database (NVD) Common Vulnerability & Exploits (CVE) Common Weaknesses Enumeration (CWE)	OECD Housing Prices OECD Gender wages OECD Wage levels OECD Price levels relative to GDP Geonames country codes Country stats World Data These are different angles on the same data about cost of living per country Main API Datsource Country & City data, esp salaries for common jobs and 'quality of life' National Vulnerability Database (NVD) Common Vulnerability & Exploits (CVE) Common Weaknesses Enumeration (CWE) ICO DataSecurity Incident trends Cybersec Cybersec Cybersec	DECD Housing Prices Housing Prices per country	Housing Prices Housing Prices Housing Prices Sender Wage Gaip by country, cox contains data over time

LIK Health Dent, life expectancy	Health		SHS
ok ricatin bept, inc expectancy		ote	SGH
	Collection of datase	ets	ЗСП
	Collection of datase	ats	SGH
	Concention of datas		3011
English Prescribing Dataset	Pharma		NP
Liigiisii Frescribiiig Dataset	FIIdIIIId		INF
Drug Misuse England and Wales	Pharma	Data from 1995-2022 on drug misuse with demographics	NP
ONS	Cybersec	UK reported cybercrime data (possibly could be merged with data from row 2)	NP
	,		
	Chemicals / Pharm	a Chemicals That May Contribute to Disease	AM
	Chemicals / Pharm	a Chemicals in Cosmetics: What's Really in Your?	AM
CISA	Cybersec	Known Exploited Vulnerabilities Catalog	AM
UN	Various	Big variety of topics regarding population, gender or energy among others per country	AM
		Big variety of topics per country as Production, Food Security and Nutrition, SDG indicators, Food Balances, Trade,	
FAO	Various	Prices, Cost and Affordability of a Healthy Diet, Food Value Chain or Climate Change: Agrifood systems emissions	AM
NHS Digital, API Catalogue	Health	117 API's related to NHS data	SHS
UK BioBank	Health	large-scale UK biomedical database and research resource	SHS
NHS Data	Health	NHS Data on health conditions, hospitals, patients, reasons for admissions etc	SHS
Google Sheet	Various	Websites with Data for Projects	TN
		Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and	
UNB	Cybersecurity	independent researchers	TN
Parliament UK	Political	Election data (API works)	NP
Market Data API > navigate to 'developers'	various	, , , , , , , , , , , , , , , , , , , ,	TN
Google API Explorer >	Various	There seems to be loads of options here > The Google APIs Explorer is is a tool that helps you explore various Google APIs	TN
		With quite a lot of API ontions and we can go through it > A collection of public APIs for developers, categorized and	
Public API	Various	crowdsourced. Animals, books, cryptocurrencies, development, music, weather and much more. (TN
	CISA UN FAO NHS Digital, API Catalogue UK BioBank NHS Data Google Sheet UNB Parliament UK Market Data API > navigate to 'developers' Google API Explorer >	Collection of datas Collectio	Collection of datasets Pharma Data from 1995-2022 on drug misuse with demographics Cybersec UK reported cybercrime data (possibly could be merged with data from row 2) Chemicals / Pharma Chemicals / Pharma Chemicals That May Contribute to Disease Chemicals / Pharma Chemicals in Cosmetics: What's Really in Your? CISA Cybersec Known Exploited Vulnerabilities Catalog UN 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 NHS Digital, API Catalogue Health 117 APIs related to NHS data UK Biodank Health NHS Data on health conditions, hospitals, patients, reasons for admissions etc NHS Data on health conditions, hospitals, patients, reasons for admissions etc Congle Sheet Various Websites with Data for Projects Conadian institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian Institute for Cybersecurity datasets are used around the world by universities, private industry, and independent researchers Canadian Institute for Cybersecurity datasets are used around the world by universities,

Data Question Ideas								
Gathered datasets	Source							
Salaries %iles of common jobs - by country	Teleport API							
Salaries %iles of common jobs - by city	Teleport API							
2022 gender wage gap (column) 2022 wage levels, high/low - by country	OECD csv							
(column)	OECD csv							
2022 average wages - by country (column)	OECD csv							
Geonames country codes & basic stats (table)	Geonames table							
Cost of living Index	World Data							
Cost of Housing Index	OECD csv							
	What dataset(s) will	be drawn on to answer this	augetien?					
Data Question (natural language)	Dataset 1	Dataset 2	Dataset 3	Analysis Steps	Visualisation	Why is this Q / it's output interesting?	What is the output useful for?	Who is the Q / output useful to?
Which countries offer the highest salaries for {each job}?		Dataset 2	Dataset 3	* Compare salary data for job roles across countries rather than by individual country.		It will be interesting to see which countries offer the best and worst pay for different jobs.	Teleport only currently offers the ability to	People wanting to assess, for a particular profession (i.e. their own!) where some of the best / worst places to live might be,
				* Rank these from lowest > highest * Add latitude and logitude of the country into the view.		Do certain countries offer the best pay for all jobs? Or does it vary by role? Are there any surprises in this?	levels across countries. In essence, our aggregation and visual comparison would extend the utility of Teleports' salary dataset	based solely on potential average salaries
Which countries offer the lowest salaries for {each job}?				view.		Materially, is it better as a woman to live in a country where average wages for that job are lower than other countries, but where the gender pay gap has less impact?	udidoet	
What are the average salaries for different jobs, when adjusted by the OECD gender wage gap for that country? Does the gender pay gap adjustment actually impact which is the best country for salary of {a particular job}?	Teleport salary data - country	OECD Gender Wage Gap (2022)		* Calculate the average salaries for each role across different countries (as above) * Use OECD gender pay gap percentages to calculate a new column of these average salaries by job across countries for women * Analyse the impact this has upon the rankings of where the top 5 / bottom 5 countries are for salary for that job. Does the gender pay gap influence the rankings	worst countries for pay has changed.	Seeing if a country level gender pay gap is significant enough to actually override the relative rankings of average salary levels. For example, there may be two countries which have identical average pay bands for a particular job - paywise, it will be equally lucrative to do that job in either country. However, if one country has a significant gender bias, does that knock it down in the ranking, for those people who a gender pay gap would affect?	Evaluating the impact of the gender pay gap across different countries. Does a large gender pay gap actually override average salary advantages in any country, for women?	Those interested in evaluating the gender pay gap particularly across different roles. For example, are there any professions which are particularly resilient to gender pay gaps, or to large gender pay gaps across the country as a whole? Also, could inform practical decisions for those who might be disadvantageously affected by a gender pay gap - women, or those for whom women are the main breadwinners - regarding where they want to live if informed by their earnings
How does Teleport's salary data (aggregated by country) compare to the OECD rankings of average wages by country?		OECD Average Wages - country (2022)		* Aggregate Teleports salary-by-job-by-country into overall average salary level * Optional: calculate and visualise the differences between values from both sources?	* A bar chart comparing Teleport with OECD average wages by country. * A visualisation of the scale of the differences? Are there any particular countries where the data is particularly divergent?	Informs about the consistency of Teleport's data relative to the OECD data on the same/very similar metric	Benchmarking and judging the accuracy of Teleports salary data (which comes from ???) relative to the OECD statistics for average wages	potential. Those interested in using Teleports' data and need to gauge the reliability (or at least the divergence) relative to the OECD gathered data.
Looking at the average salary by population density. To see a correlation using a scatter plot. Using regression analysis. step. further-using machine learning Do the countries with the highest overall salaries have the highest pay parity?								

Section	Section Weighting	Sub-section	Marks Available	Overall Weighting
Project Objectives & Final Result how well did the project meet the	10%	Final results vs objective set	5	5.00%
original objectives?		Performance of the final product	5	5.00%
		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 Implementation		Code layout & readability	5	4.50%
how well written and evaguised is the	50%	Relevance of chosen topic	5	4.50%
how well written and organised is the code?		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%
		Introduction	10	5.00%
Project Documentation		Background	5	2.50%
_	30%	Specifications and Design	15	7.50%
is there a clear intro, background,	30%	Implementation and Execution	15	7.50%
discussion and conclusion?		Result Reporting	10	5.00%
		Conclusion	5	2.50%
Group Project Presentation		Evidence of Teamwork	5	3.30%
how well did you present as a group? Was it clear and concise?	10%	Presentation Skills	5	3.30%
Did you keep the audience engaged and answer questions effectively?		Understanding of the Project	5	3.30%