Task:

1. Critically Evaluate the adoption of any one modern data analysis and

visualization technology for any specific practices (e.g.: Medical, Insurance,

business, etc.)

2. Data Analysis & Visualization using Oracle Analytics Cloud or another tool of your choice

**Task 1**

**Critical Evaluation of the adoption of modern data analysis and visualization technology for specific practices.**

The importance of Data Analysis and Visualization in today’s world can never be over-emphasized. The continuous evolution over the years has led to a systematic process of inspecting, transforming, and modeling data to extract useful information and insight. Due to the number and complexity of contemporary data sets, traditional approaches are no longer capable, which has considerably raised the need for advanced data analysis tools. In this report, I will focus on the Python programming language (as a modern data analysis and visualization technology) and the techniques adopted in the medical industry as it compares to finance, transportation, and other industries. Python is a general-purpose programming language that has emerged as a predominant tool for data analysis and visualization due to its open-source benefit, simplicity, adaptability, and extensive libraries. The libraries offered include Numpy, Pandas, Matplotlib, and Seaborn, among others. According to McKinney (2010), Python also offers data structures such as arrays, matrices, and data frames that are useful for storing and manipulating large datasets in multiple fields.

**The use of Python in the Medical Field**

Image processing, clinical data analysis, and predictive modeling are some of the reasons Python has been commonly adopted in the medical industry. The article by Neda R et al (2021) gave a systematic review of the importance of data analysis techniques and how they are used to analyze patient information from electronic health records. These techniques include clustering, classification, and time-series analysis. Clustering was used to group patients who have similar symptoms or characteristics of an illness. For instance, Wu et al. (2019) employed the use of clustering to categorize breast cancer patients according to the patterns of gene expression. This also helped to identify patients with a different illness and rate of progression. A classification model was used to predict the risk of re-admitting heart failure patients based on the patient’s demographics, clinical data, and laboratory results. In a study by Gotz et al (2019), time-series analysis was used with data collected over time to track changes in electroencephalogram (EEG) data with traumatic brain damage which led to better treatment decisions.

**Python in Other Industries**

A study by Wanan et al (2022) demonstrated the effectiveness of using Python and machine learning techniques in the financial industry. Scikit-learn and XGBoost are used for data processing, feature selection, and model development. A weighted boosted tree-based algorithm is used to predict the financial distress of firms based on financial ratios and other company-specific factors obtained from data of companies identified as financially distressed by the China Securities Regulatory Commission. To evaluate the model, they used metrics such as accuracy, precision, recall, and F1-score to identify the most important factors causing the distress. The author’s approach outperformed other traditional methods such as logistic regression and decision trees, indicating the superiority of the weighted boosted tree-based algorithm in handling imbalanced data.

A study of Analysis and visualization of accident severity (Kun et al 2022) depicts how it has been very effective in the transportation system. It was used to train and build machine learning algorithms to perform data analysis to predict accidents by adopting a LightGBM-TPE model. A hyperparameter tuning technique called tree-structured parzen estimator (TPE) was used to optimize the model’s performance from a dataset obtained from traffic accidents in the city of Rio de Janeiro, Brazil. Haocheng et al. (2022) stated that "LightGBM-TPE was used as the main algorithm for predicting the severity of traffic accidents". The authors used the Pandas library for data processing and Matplotlib for data visualization. Heat maps were used to visualize the correlation between different types of accidents (collision, number of cars, time, etc.) and the severity of the accidents. With this, a valuable tool was created to improve road safety and the occurrence of accidents.

The urban planning study (Yong Kang 2021) illustrated the use of Python libraries to visualize and analyze the positive externalities of urban underground spaces. They used spatial analysis and visualization techniques to identify the socio-environmental advantages of underground spaces and developed a framework to evaluate its benefits. For data analysis and visualization, they used pandas, geopandas, matplotlib, and seaborn. They also used the Arcpy module in conjunction with GIS software for geospatial analysis and data management. The results were then visualized on interactive maps using Folium and Leaflet libraries. This made it easier to understand the benefits of urban underground spaces due to positive externalities such as reduced energy consumption and improved air quality.

**Limitations of python**

A common issue in Python revolves around scalability when processing large datasets. The “Using Python libraries and k-Nearest neighbor’s algorithms” study noted that the memory requirement can be an issue when dealing with a large number of data points (Manuel M et al 2023). A possible solution presented by Ahmed O et al (2017) in a study on big data technologies is to use distributed computing frameworks such as Apache Spark which can handle larger data sets across multiple nodes.

Version management of Python libraries is another challenge observed. In “Python tools for stem cell transplantation” the authors expressed the need for more support and resources from the community of Python users (e.g. GitHub) to enhance and develop standardized tools and resources for Python-based analyses. In addition to tools like anaconda, conda, and virtual environment offering more efficient and streamlined usage (Joseph et al. 2019).

**Conclusion**

Python as a technology for analysis and visualization has demonstrated strong potential in various industries due to its versatility and availability of various libraries. The use of Python in the medical field and practice has led to the crucial development of automated tools for disease diagnosing, detection, and improved treatment. In other industries, it is useful for interpreting financial distress, urban development, and the severity of accidents. It is also used to aid in effective risk management and the development of prevention and management tactics. However, the adoption of Python practices requires continuous improvement and updates to cater to the ever-changing technological landscape in regards to standardization, documentation and compatibility.

**Task 2**

**Introduction**

A recruitment company is in need of expanding into specific countries with highly skilled populations in other to maximize its productivity. In other to make a data-driven decision, the company has decided to analyze skills, education, and earning datasets from across nine countries within Europe. The analysis is to help the company make informed decisions concerning the intended development towards quality selection and hiring process.

**Dataset**

The dataset contains nine countries which will be highlighted in my report.

**The skills needed by the industry**

Dataset source: <https://stats.oecd.org/Index.aspx?DataSetCode=S4J2022_NACE>

The skills needed by the industry data set aim to facilitate better adoption to changing skill needs by making available a database of skill imbalances indicators that is comparable across countries. This data set provides an overview of surplus and shortages of skills across countries. The idea of skills by industries considered by the company is to help them determine which industries will be worth more resources in consideration to changing demands or skill needs. This also will help the company understand the skills needed for the expansion of the recruitment company and identify the competitive skills across the industries.

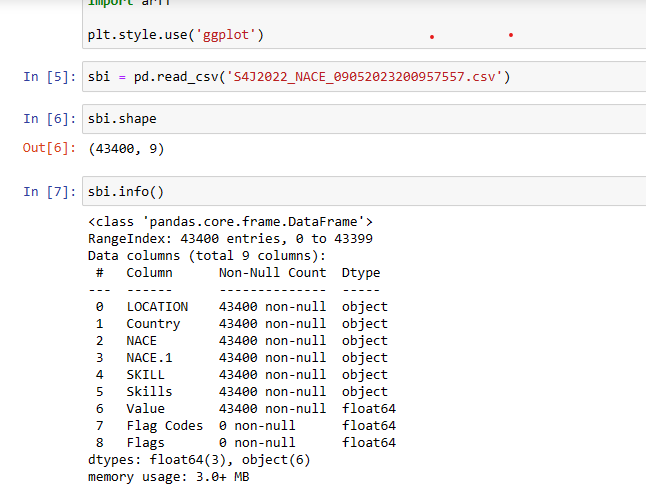
**Education and earning by field of study**

Dataset source:<https://stats.oecd.org/Index.aspx?DataSetCode=EAG_EARNINGS_FIELD>

The education and earnings data set compare demographics, education, and earnings based on the field of study. This dataset will help the company to determine what skill level in terms of age and gender to focus more resources on. This will also help the company to determine cost budgeting.

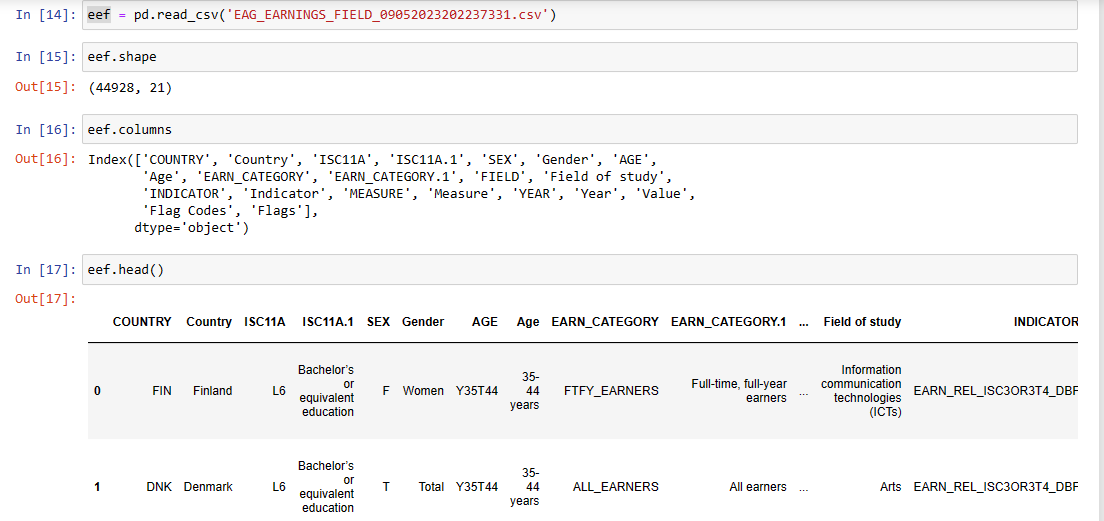
**Data preprocessing:**

**Fig 1**



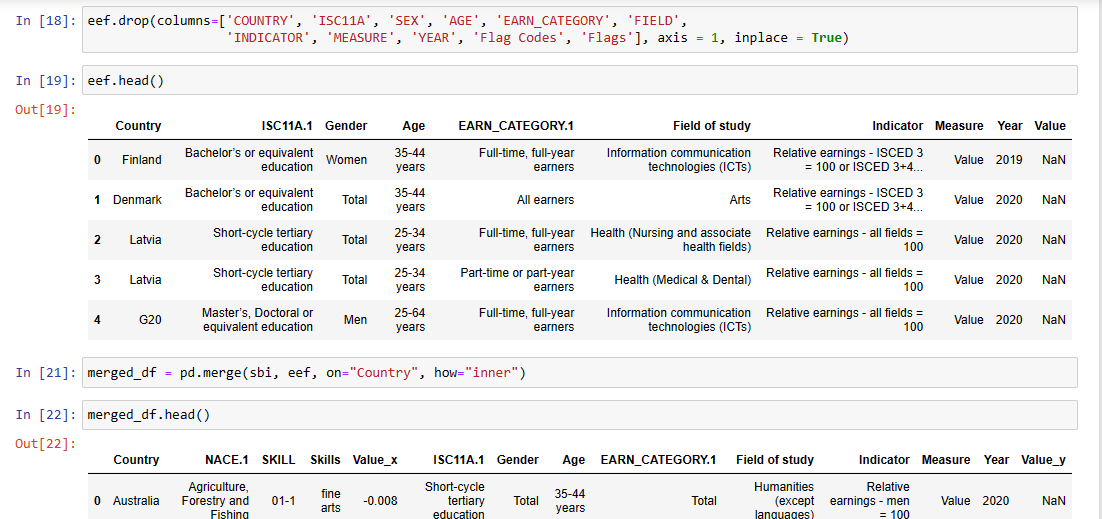
In Fig 1, I imported the first data set “Skills by industry”. The dataset contains 9 columns and 43400 rows. Using “pd.read\_csv” to upload the file that contained the dataset and used “isnull(). sum()” to inspect for null and missing values in the dataset. I discovered that flag codes and flags had over 43,400 missing values which I dropped from the dataset alongside other columns that are not relevant to the analysis using “. drop(columns=['LOCATION', 'NACE', 'Flag Codes’, ‘Flags'], axis = 1, inplace = True)”. I used .info() to get a comprehensive idea of the dataset and discovered it consists of object and float.

Fig 2



In Fig 2, I imported the second dataset “education earning by field”. I inspected and ensured the data was clean enough for the analysis. Using a similar process of importing and inspecting the first dataset, I discovered the dataset contains 21 columns and 44,928 rows. I inspected the columns using. head().

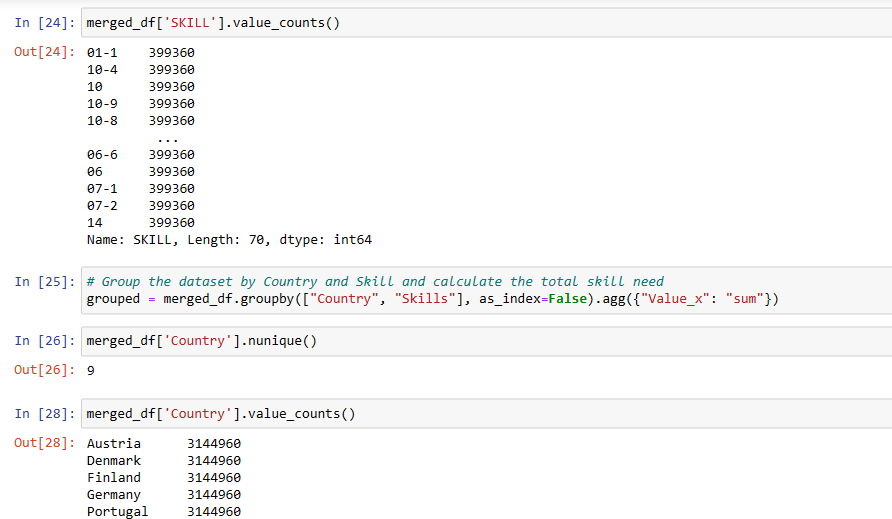
**Preparation and merging**

**Fig 3**

Having dropped the unwanted columns from the education and earnings field due to duplications using “.drop(columns=['COUNTRY', 'ISC11A', 'SEX', 'AGE', 'EARN\_CATEGORY', 'FIELD','INDICATOR', 'MEASURE', 'YEAR', 'Flag Codes', 'Flags'], axis = 1, inplace = True)”, I examined to be sure that I have the right set of columns for the analysis. I went ahead to merge the datasets using “merged\_df = pd.merge(sbi, eef, on="Country", how="inner")” and displayed the data frame to see the outcome of the merged dataset. I was satisfied with the output and convinced it was enough for analysis.

**Calculations**

Fig 4



By grouping the dataset according to country skillset and aggregating by the value of each skillset. I counted the skill types amounting to 70 different types of skills contained in the dataset. I grouped the dataset by country and skills while ignoring the index of the dataset in other to get a desirable result. A few examples of the skill types are active listening, adaptability, attitudes, values, writing, and so on.

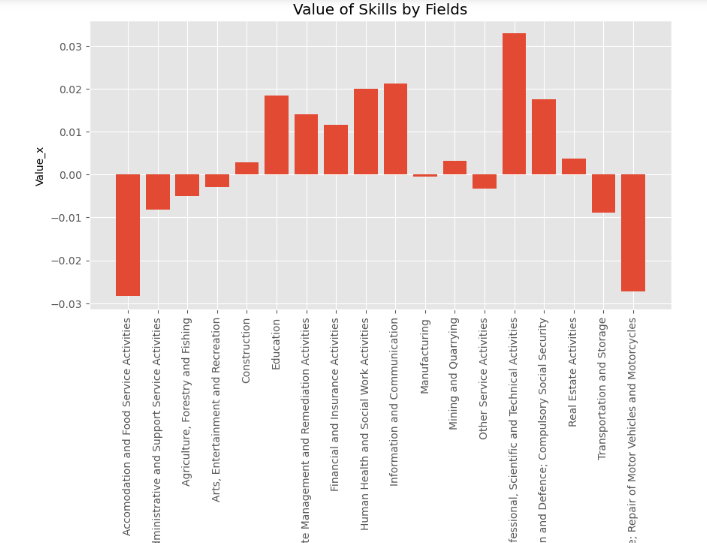
The dataset contains 9 countries which are:

* Austria
* Denmark
* Finland
* Germany
* Portugal
* Sweden
* Estonia
* Latvia
* Australia

**Visualization**

The importance of visualization is to help ease the comprehension of the findings. It is important to note that an untidy dataset will give a poor visualization. Hence the need to clean it before the graphical display of the data.

Fig 5

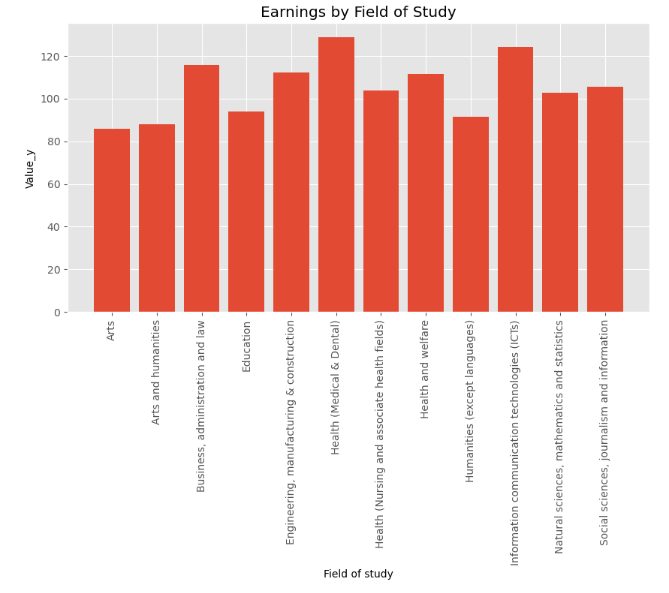


From Fig 5, it is obvious that the most valuable skills are those in the professional, scientific, and technical domains because they rank the highest across industries. The least ranked ones are in the food services, wholesale, and retail trade industries.

The financial implication of this for the recruitment company is that more budget allocation is required for skillsets in the professional, scientific, and technical domains as compared to the budget required for skillsets in the food services, wholesale, and retail trade industries.

The average ranked skillsets are those in the construction, mining, real estate, and entertainment industries. If the company wants to make relatively less expenses on skillsets or to conduct preliminary assessments, they may consider the skillset in these domains.

Fig 6



This graph represents the top earning fields of study. The graph revealed that professionals in the medical field earn the highest followed by ICT then business and administration. Art and humanities are the least earners.

The implication of hiring across this skillset is that the company should be ready to make higher budget allocations for professionals in the highest earning group and relatively less for professionals in the lower earning group.

The medical field is a more lucrative field of study but could demand high financial implications for the recruitment company.

**Conclusion:**

The purpose of the analysis was to help the company make a quality selection and hiring process for its expansion. The dataset helps to assess the values and the potential earnings of individuals who possess any of the categories of the skillsets. From the dataset analyzed, we observed the distribution of skillsets across nine countries in Europe to be highly skewed towards professional, scientific, and technical while education and earning are highly skewed towards the medical fields, ICT, and business administration.

This leads me to propose to the recruitment company to allocate a higher budget in order to cut through the competition cluster and hire the best hands. The earnings by field of study suggest that medical fields will be the most lucrative and attractive domain for investors.

**Reference**

Andrei Lobov & Tuan Anh Tran *Object-oriented approach to product design using extended NX Open API 2020;* <https://www.sciencedirect.com/science/article/pii/S2351978920319995>

Kun Li a, Haocheng Xu a & Xiao Liu *Analysis and visualization of accidents severity based on LightGBM-TPE 2020:* <https://www.sciencedirect.com/science/article/pii/S0960077922001977>

Yong-Kang Qiao a, Fang-Le Peng a, Xiao-Lei Wu b & Yong-Peng Luan *Visualization and spatial analysis of socio-environmental externalities of urban underground space use: Part 1 positive externalities 2021:* <https://www.sciencedirect.com/science/article/pii/S0886779821005162>

Wes McKinney *Data Structures for Statistical Computing in Python 2010:* [*https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf*](https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf)

Manuel Martín-Martín a, Manuel Bullejos b, David Cabezas c & Francisco Javier Alcalá *Using python libraries and k-Nearest neighbors algorithms to delineate syn-sedimentary faults in sedimentary porous media 2023:* [*https://www.sciencedirect.com/science/article/pii/S0264817223001897*](https://www.sciencedirect.com/science/article/pii/S0264817223001897)

Joseph Vadakara MD, Soumit Basu MD & PhD *Python Tools for Stem Cell Transplantation 2019:* [*https://www.sciencedirect.com/science/article/pii/S1083879118310310*](https://www.sciencedirect.com/science/article/pii/S1083879118310310)

Ahmed Oussous a, Fatima-Zahra Benjelloun a, Ayoub Ait Lahcen a b & Samir Belfkih *Big Data technologies: A survey 2018:* [*https://www.sciencedirect.com/science/article/pii/S1319157817300034*](https://www.sciencedirect.com/science/article/pii/S1319157817300034%20)

Neda Rostamzadeh, Sheikh S. Abdullah & Kamran Sedig *Visual Analytics for Electronic Health Records 2021:* [*https://www.mdpi.com/2227-9709/8/1/12*](https://www.mdpi.com/2227-9709/8/1/12)