## Index

1. [Project\_outline](#Project%20outline)
2. [environment](#environment)
3. [directory](#directory)
4. [Development\_Environment\_Setup](#Development%20Environment%20Setup)
5. [Troubleshooting](#Troubleshooting)

## Project Name

IMPACT OF REMOTE WORK ON PRODUCTIVITY AND PERFORMANCE

## Project outline

The data describes the questionnaire comprises approximately 180 questions in total. Analyzing pairs one by one becomes impractical given the number of question pairs is , totaling over 16,000 patterns. To efficiently investigate these question pairs, we have developed a statistical application as outlined below. The notebook titled “main” serves as the report file for this analysis, while “main.py” acts as an integration file, implementing all the functions. The process is divided into three main streams: “a\_data\_process.py”, “b\_stats\_approach.py”, and “c\_viz.py”, executed in sequential order.

## How to use this application

1. Please make sure to download all the files listed in the Directory below. All the files are organized in the directory structure.
2. Set your own environment variables. The application requires you to have Kaggle API and OpenAI API available. The former API is free to register, but the last is paid.
3. Please make sure to install all the packages listed in the requirements.txt
4. Run the main.ipynb.

## Environment

Please refer to the ‘requirement.txt’ in the same directory.

## Directory

A screenshot of a computer program

Description automatically generated

### File A

It is responsible for cleaning the survey’s raw data, ensuring its consistent usability. This file’s components are crucial; for example, it addresses missing value imputation. To maximize the information from the raw data, we opt for univariate imputation over Pair-wise or List-wise deletion. Although multivariate imputation could be more comprehensive, its systematic implementation to universally fill all types of missing values poses challenges. For categorical variables, missing values are imputed with the most frequent value or an ‘Empty’ text. To minimize data analyst bias, “Empty” text is used to clearly denote a cell as NULL if the missing value ratio exceeds 10%. Conversely, if the ratio is below this threshold, the most frequent value is used for imputation. For numerical variables, we impute missing values with either the mean or median value. Should the outlier ratio surpass 0.3%, we consider their influence significant and opt for median imputation, whereas mean imputation is applied otherwise.

### File B

It constitutes the core of the statistical analysis implementation, and file C facilitates visualization. Associated with file B, several helper functions act as individual statistical methods. For example, when a Bartlett test is required, “equal\_variance\_test.py” is called from file B.

### File C

It plays a role of visualization. The task is done with Matplotlib and Seaborn, which is very popular libaries to create graphs.

### File D

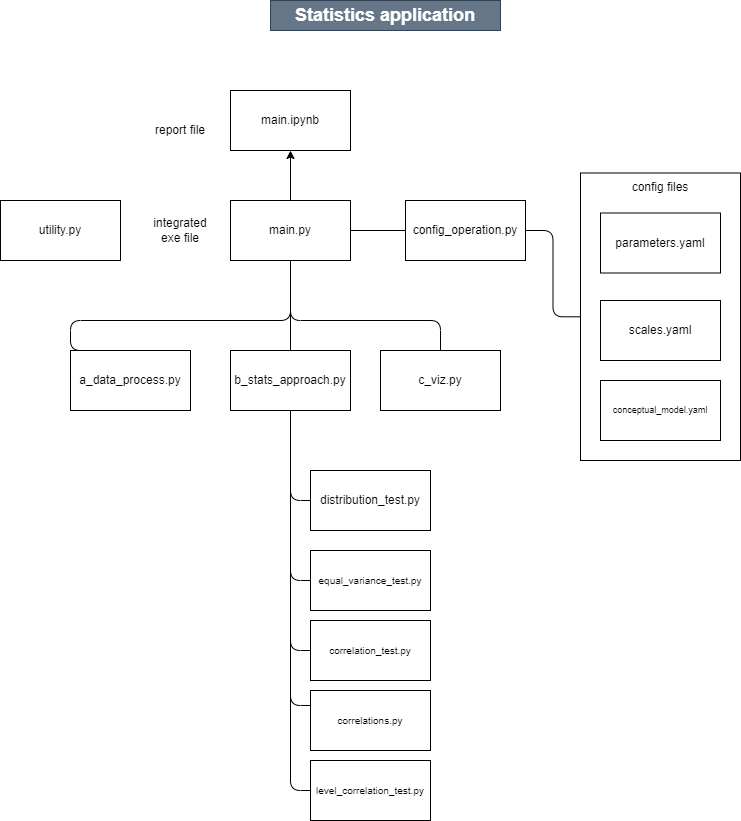
It is an independent file from File A to File D. This contains of code to execute ChatGPT expert, which is a customized ChatGPT to answer questions based on ohter rearches on the Internet.

### Main.py

It is for running all the functions defined in other files with specific argument parameters. It is directy called from Main.ipynb.

### Other files

we have predefined configurable parameters stored in YAML files. These files set hyperparameters, such as common question IDs. “Utility.py” includes useful functions for addressing common preprocessing tasks.



## Development Environment Setup

### environment variables

| variable names | roles |
| --- | --- |
| KAGGLE\_USERNAME | Username for Kaggle API |
| KAGGLE\_KEY | Sercret key for Kaggle API |
| OPENAI\_API\_KEY | Secret key for OpenAI API |

## Troubleshooting

### dotenv trouble

please make sure you set environment parameters in .env.

### I cannot find data files

Please make sure you follow the directory structure above.

If you cannot open the data files, you can download the data through Kaggle API.

Please refer to “how\_to\_use\_kaggle\_api.ipynb”

### I cannot run the d\_chatgpt.py

You need your OpenAI API key. It is pay-as-you-go.

### Module not found

please install all the packages listed in the requirements.txt