Coursework - Research Report

NOTE: This is a Group Assignment[1]

The Problem:

In this coursework, you are provided with a dataset which includes information on the various aspects of a building location and construction. The dataset includes lots of information such as location of the building, it's age, foundation, etc. Using the data, one can help predict the level of damage to the building that can be caused by an earthquake.

Such datasets can help concerned government agencies design better governance plans, including but not limited to improvement to existing infrastructure, better disaster management plans, etc.

Given the datasets, you are tasked to prepare a research report outlining the importance of dataset and the analysis that can be performed and how it will be useful for the concerned authorities.

See Canvas for submission deadlines.

[1] Refer to the Canvas page here on where to register your group and how your final individual mark will be calculated based on the Individual Contribution Report.

Coursework Report:

In this report, you are asked to use the dataset provided to perform a research analysis consisting of the following:

Introduction (20%)

This section should include an introduction to the field of data science, industry/application/domain that the dataset is representative of. You are required to formulate a problem statement and explain the proposed approach you will take to solve the identified problem. Identify different (research) questions (minimum 8, maximum of 10) that you plan to investigate as part of your research report. Define the objectives, intended research methodology and expected outcomes. Describe your workflow supported with figures.

Tools, Dataset and Preliminary Data Analysis (35%)

Programming Language and Tools (10%)**:** In this section you are required to discuss different tools (3-5) and methods available for data analysis. Discuss the advantages and disadvantages, features of them, the choice of programming language (Python) and the related libraries that you will be using. What are the advantages and disadvantages of using your selected tool and its associated libraries? Which functionality or features will be useful for your analysis and why?

Dataset(s) (5%)**:** In this part, discuss the dataset - its source, format, type of data, the source of the dataset and format used (structured/unstructured), etc. (Hint: Read the dataset and discuss the columns, type of data in the columns, etc.).

Initial Data Analysis (20%)**:** Discuss about the quality of data that you have. Are there many missing values, data is clean, etc.). Are different transformation methods required for your further analysis? Make sure that you discuss the different methods of data cleaning, different methods to account for missing data (Null values), pre-processing, and transform features. What are the different types of data wrangling skills (e.g., extraction, merging, and/or construction of analytical data sets) that can be used?

Hint: Refer to Pandas.ipynb file and Lecture slides.

Exploratory Data Analysis (40%)

Introduction to EDA (5%): In this section you need to discuss EDA. Also make sure to include discussion about if further data cleaning or transformation is required? What are the different approaches and which ones do you think will be suitable for the data in hand.

Descriptive Statistics (20%)**:** In this section, discuss the various statistical methods. Discuss which ones you think will be relevant to understand your data, help answer your research questions and why? You need to present a detailed discussion on various methods and their suitability for your use case.

Data Visualization (15%)**:** What is data visualization, how and why are they used. Discuss about the best practices. What are the different types of graphs (pie-chart/scatter plots/bar plots/histogram, etc.-you can use sample ones from the official website) and when are they used? Which visualization methods do you think will be appropriate to your research questions and why? (Note: You need to justify at least one visualization method for each of your research questions). You can reference to different plots from other sources (make sure to include appropriate references)

Hint: Refer to Pandas.ipynb and PandasAdvance.ipynb file and lecture slides.

Summary and Conclusion (3%)

Summarize your work and learnings.

References (2%)

Anything borrowed/copied from any external source needs to be referenced. Any referencing method can be used (IEEE preferred). Refer to Research Methods (Library talk), SASC or Library Canvas & tools like "Cite them Right" for referencing guidance.

##

Starting Guide:

1) There will be a discussion on the problem statement and how to approach it. Make sure to attend the lecture and take notes, as appropriate.

   Note: You are welcome to start with Jupyter Notebook and data analysis but not required to submit it as part of this report.

2) Refer to your notes and material from Research Methods canvas module

Jan 2021: https://canvas.kingston.ac.uk/courses/16422/pages/research-methods-jan-2021
Sept 2021: https://canvas.kingston.ac.uk/courses/19827/pages/research-methods

Instructions for submission:

You need to submit a **Microsoft Word document (*.doc,.docx*) or pdf (preferred) file only**. Any other file format is NOT acceptable and will be considered a non submission.

At the start of the report, add a section (including names and KU IDs of ALL the group members, and title of the report), a ToC (Table of Contents) and Table of Figures, etc.

Make sure that you have informed the module leader of your team and got approval before starting to work together on the report.

Assessment and Feedback:

You will be assessed primarily on (but not necessarily limited to) the following: completeness of the task (report content, see above), quality of the work (research questions, reasoning, etc.), presentation style (report structure, logical flow, grammatical and spelling errors, sentence structure etc.), appropriate references using an appropriate style. You will be provided feedback on your work by the deadline mentioned on the *Canvas Assessment and Feedback Schedule* page here.

Rubric:

The above description clarifies the marking structure. Below is just a summary copy of the marks for the research report. In the case of any discrepancies, the above description supersedes this one below.

Introduction (15%)		**15**
Tools, Dataset and IDA (35%)	**Programming Languages and tools**	**8**
	Dataset	**7**
	IDA	**20**
EDA (35%)	**Intro to EDA**	**5**
	Desc Stats	**10**
	Data Visualization	**20**
Summary and Conclusion (3%)		**10**
References (2%)		**2**

Please note that as part of this group report, ALL group members will initially receive the SAME marks. However, your final, individual marks will be decided based on your actual contributions which will be decided based on the Individual Contribution Report available here.

University Policy and Regulations

Any submission found/suspected of academic misconduct including but not limited to collusion, plagiarism and Purchasing or commissioning will NOT be marked and all the members of the group will be reported to the University for further action. Make sure you read the below guidelines available here:

Academic Misconduct (including Plagiarism)

To apply for extensions and mitigating circumstance, please refer to the guidelines here:

Extensions and Mitigating Circumstances Policy

Viva: We reserve the right to withhold your marks and invite you to a *viva* where you will be asked questions about your work to demonstrate ownership and understanding of your own submission. Whether you are asked to *viva* or not will depend on:

- Your engagement during scheduled classes
- Your engagement with the team
- Your formative submissions (i.e. the work you do for learning purposes rather than marks)

- The work you submit (if it's an interesting submission and we want to ask some questions – this is a GOOD thing, as you may end up with more marks!)

If you are asked to do a *viva*, then your final mark will be subjected to a multiplier factor as follows:

Multiplier	**Viva assessment**
1.2	Ownership and understanding of the artefact submitted was excellent, and gave an insight into not only their own work as-is but also possible alternative approaches, demonstrating more understanding of development techniques than the work alone provides.
1	Ownership and understanding of the artefact submitted is clearly established. The student responded to questions about their work with authority.
0.6	The student was able to answer most questions about their work, but there were some gaps in understanding of the artefact submitted.
0.3	The student was able to answer some questions and demonstrate only a limited understanding of their own submission. However, there was a clear disconnect between the quality of artefact and the students' understanding of the technology and principles encapsulated within.
0	The student was not able to provide any meaningful responses to questions about their work. Extremely limited, or no understanding of the artefact submitted was apparent. Alternatively, the student did not respond to the invitation to viva and/or did not attend the viva at the appointed time.

Please note that, as they contribute to an assessment event, *vivas* will be recorded to provide sight of such assessment events to moderators and external examiners. Where appropriate, the recordings will be used in other university regulatory processes.

Contact

Please contact your module leader, Dr Nabajeet Barman (n.barman@kingston.ac.uk) or your Course Leader Dr James Denholm-Price (J.Denholm-Price@kingston.ac.uk) if you have any questions or concerns.