**Team Project – Research Proposal**

1. **Objective:** The primary goal of the project is to clean and analyze the given datasets on PM2.5 concentrations in the city of Ho Chi Minh, Vietnam over the given period, and develop both descriptive statistics and inferential analysis required to identify monthly or yearly trends for PM2.5 in Ho Chi Minh, before and after the COVID-19 lockdown period (starts at end of March 2020). Based on the outputs, we aim to provide recommendations from a well-being perspective to the residents.

The changes in air pollution is compared with changes in stock market price of a company to analyse the effect of pollution on stock price.

1. **Team Members:**

Leader and team members (“Group 6”) are as follows:

* 1. Mohamed Ghayaas (Leader)
  2. Mary Ann Villamor
  3. Shalon
  4. Young Lee

1. **Research questions and Course of Project:**

Group 6 propose to determine the following matters using various statistical sets and for HCMC data before and after COVID-19:

1. Perform EDA – Exploratory data analysis on the data set to determine:
   1. Types of variables and explanations.
   2. Check various issues with the data quality, such as out of range scores, outliers and fix them
   3. Determine missing data, perform imputation using suitable techniques and explain the techniques employed.
   4. Use descriptive statistics and visualization techniques to support the EDA.
   5. Determine Pearson’s correlations among the variables. Which variables are highly correlated with each other and quantify the degree of correlation? Explain the interpretations
   6. Repeat the above correlation exercise with Spearman’s correlation. Are there any difference between the correlation values? If yes, explain the difference.
2. Use external dataset and determine the correlation of temperature, atmospheric pressure, humidity, visibility, and wind speed with the air pollution of HCMC city. Explain how air pollution impacts all other weather parameters.
3. List out inferences from the above analysis and give recommendations to the people of HCMC city from a wellbeing perspective
4. Feature Selection: Based on correlation analysis, determine the top features affecting the air pollution.
5. Build a time series model using the selected features to forecast/ Predict the air pollution level for the month Dec 2021 based on previous trends of air pollution. Perform feature engineering / scaling if required. Evaluate the model accuracy and explain the working of the model.
6. **Types of Analysis/Activities to-do:**
   1. **Tasks for cleaning:** Check for accuracy, Missing Data and Outliers, and handle them effectively to get a clean data set.
   2. **Data Exploration:** Identify the type of variables and explore the data by drawing out descriptive statistics and summarize the data in the given dataset.
   3. **Correlation and Linear Regression:** Identify the relationship between the data and demonstrate them by using Correlation and Bivariate Linear Regression techniques.
   4. **Comparison with External Dataset:** Compare the results from descriptive and inferential analysis to an external dataset to find patterns with the proposed case of study (Stock price variation).
   5. **Interpretation and Recommendations:** Interpret the outputs from different analysis and comparisons to give final recommendations for well-being of the Ho Chi Minh residents.
   6. **Visualization:** Use bars, graphs, charts and other techniques to visualize the analysis and outcomes, to support of the recommendations made.
   7. **Power-point presentation:** Conclude all the findings in a power-point presentation for presenting during the final submission.

(Note: This proposal is based on the preliminary initial analysis made by the Team and subject to changes depending on the availability of the data and outcome of analysis.)

1. **Task Assignments and timelines**

|  |  |
| --- | --- |
| **Activities** | **Indicative Timelines** |
| 1. **Data Cleaning**    * Each team member to take 1-2 files for analysis and compile. Team member will generate statistics to determine missing, null, outliers, etc.    * Team will use common filenames, variable name for consistency and easy harmonization of the data during merging or consolidated analysis. | 4th March 2021 |
| 1. **External Datasets** 2. Each team member will perform research of the available data for below aspects that can be used for analysis against HCMC air quality:    1. weather (temperature, Wind, Satellite view, etc.). For climate change (Global warming, water quality, etc.).    2. disease (impact of COVID19 and other diseases, etc.)    3. economic elements (Stock market index, Cryptocurrencies including Bitcoin), GDP/GNP, etc.)    4. demographic changes (Mortality, Birth rate, etc.)    5. industrial aspects (Factories, etc.)    6. transportation (number of cars, motorcycles, public transportations, ridership data, traffic info, etc.) 3. To discuss and shortlist which of the above factors will be used for further analysis. | 9th March 2021  12th March 2021 |
| 1. **Correlation and predictions**    1. Each team member will perform correlation and regression analysis to determine relationships and dependencies of the external factors against the air pollution data in HCMC.    2. To discuss and shortlist which of the above factors will be used for further analysis. | 16th March 2021  18th March 2021 |
| 1. **Report and Presentation slides** 2. Report Draft 1 3. Report Draft 2 4. FinalReport 5. **Draft Presentation Slides** 6. **Final Presentation slides** | 18th March 2021  23rd March 2021  27th March 2021  23rd March 2021  27th March 2021 |
| 1. **Weekly Progress Meetings with Monica** 2. Discussion 1 3. Discussion 2 4. Discussion 3 (with Draft Report) 5. Discussion 4 (Final Report and presentation slides) | 11th March  17th March  22nd March  25th March |
| 1. Actual Presentation | Week 13 |

**Appendix A – Preliminary external dataset sources**

1. <https://cleanairasia.org/>
2. <https://datasetsearch.research.google.com/>
3. <https://www.kaggle.com/datasets>
4. <https://www.data.gov/>
5. <https://datahub.io/collections>
6. <https://archive.ics.uci.edu/ml/datasets.php>
7. <https://earthdata.nasa.gov/>
8. <http://opendata.cern.ch/>
9. <https://apps.who.int/gho/data/node.home>
10. <https://www.bfi.org.uk/industry-data-insights>
11. <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
12. <https://crime-data-explorer.fr.cloud.gov/>
13. Source: <https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/>