CS 2704 – Final Project Instruction

Jong-Kyou Kim, PhD

Oct 31, 2023

1 Schedule

- Scheule
  - Nov 14 (Tue): Proposal presentation
  - Dec 5 (Tue): Final presentation
  - Dec 7 (Thu): Final report

2 Requirements of the proposal

- Choose the dataset
- Your github repository for the project
  - dataset
  - code
  - proposal
  - report
- Explain your hypothesis
- Plan for testing your hypothesis

3 Requirements of the final report

- Minimum requirement (70%)
  - Explain your hypothesis
    - (Compare it with your original hypothesis if changed)
  - Explain the data
    - Meta data (tables, columns, and the description of values)
    - Source of the data (preferrably publically available)
  - Descriptive analytics
    - Basic statistics
    - Visualize the description (such as correlation heatmap)
  - Predictive analytics

        - Explain the response variable
        - Choose the predictor variables
    - Submit the code

- Intermediate requirements (20%)
    - Analyze the dataset
        - Feature engineering of the predictor variables
        - Visualize the feature to explain the data
    - Submit the code

- Advanced requirement (10%)
    - Build a predictive model
    - Evaluate the predictive model
    - Submit the code and the collected data

4 An example of the hypothesis and the data sets

- Brief background
    - GDP per capita could explain how well a country is prepared for treating COVID-19 such that preventing the infection to be escalated to a serious condition.
    - We can find a significant correlation between the two variables

- Hypothesis: The fatality rate of covid-19 has a correlation with GDP per capita.

- Dataset:
    - `https://ourworldindata.org/coronavirus-source-data`
    - `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`

- Explain the dataset
    - Examples: Scatter plot of GDP vs. Fatality rate, heatmap of the correlation, etc.
    - The goal: Without having a prior knowledge, the readers or the audience can recognize the relationship between variales

- Descriptive analytics
    - The p-value is smaller than 0.05, therefore, the two variables are correlated
    - The references for the conclusion

- Predictive analytics
    - We used the linear regression to predict the fatality rate of COVID-19. We concluded that the prediction is not statistically significant.
    - The references for the conclusion

- Discussion and further research

  - Explain what have been useful or successful
  - Explain what were the theretical or practical challenged
  - Suggest future work for better understanding the dataset
    - This may include suggestions for more data collection

5 Recommendations for possible data source

- The following web pages are supposed to be free data sources. *Note: Some pages may contain broken links or possible phishing site. Bad guys seem to exploit the popularity of data analytics. Though I reviewed the following links, I might have missed something. Please inform me when you find something suspicious within the following.*

```
https://wikidata.org/
https://en.wikipedia.org/wiki/Wikipedia:Database_download
https://wiki.dbpedia.org/
https://www.data.gov/
https://www.usa.gov/developer
https://registry.opendata.aws/
https://www.nationalarchives.gov.uk/
https://archive.ics.uci.edu/ml/index.php
http://crawdad.org/
http://snap.stanford.edu/data/index.html
https://data.austintexas.gov/
https://registry.opendata.aws/
https://data.cityofchicago.org/
https://data.gov.uk/
https://www2.jpl.nasa.gov/srtm/
https://data.medicare.gov/
https://data.seattle.gov/
https://datasf.org/opendata/
https://www.dartmouthatlas.org/
https://www.bls.gov/
https://www.kiva.org/
https://www.faa.gov/data_research/
https://opendata.vancouver.ca/pages/home/
https://fred.stlouisfed.org/
https://stats.oecd.org/index.aspx
http://data.un.org/Explorer.aspx
https://www.ngdc.noaa.gov/ngdc.html
https://data.gov.uk/
https://data.worldbank.org/
https://pslcdatashop.web.cmu.edu/
https://data.gov.bc.ca/
https://www.archives.gov/research/alic/tools/online-databases.html
https://www.data.gv.at/veroeffentlichende-stellen/
https://daten.berlin.de/datensaetze
https://opendata.cityofnewyork.us/
https://dados.gov.pt/pt/
```

https://www.dati.gov.it/
https://dati.trentino.it/
https://www.google.com/publicdata/directory?hl=en_US&dl=en_US#!
https://www.google.com/publicdata/directory
https://developer.imdb.com/
http://usgovxml.com/
https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html
https://www.kaggle.com/
https://www.theguardian.com/data
https://github.com/awesomedata/awesome-public-datasets

6 Materials for your proposal

- Slides explaining your hypothesis
- URL or snippet of data
- Your guess to the expected output

7 Materials for your final presentation

- Slides
- Demonstration
  - Explain the code
  - Generate visualization

8 Sections for your final report

- Introduction and Background ($\leq$ 200 words)
- The hypothesis ($\leq$200 words)
- The analysis and the implication ($\leq$300 words)
- Conclusion ($\leq$200 words)
- References (as complete as possible)