Project Diary

Todd (Zhen) Zhang, Shubo Feng, Dan Wei, Zichen Zhou

2023-10-19

Overview of project diary:

- Week 9: Topic selection
  - Sep 25: Initial Ideas Generation
  - Sep 26-28: Feasibility Evaluation
  - Sep 29: Determination of Finial Topic and Individual Tasks Assignments
- Week 10: Data collection and wrangling
  - Oct 2-3: Data Collection
  - Oct 4-6: Data wrangling
- Week 11: Data analysis and project refinement
  - Oct 9-10: Data analysis
  - Oct 11-15: Add more risk factors data and refine data wrangling process
- Week 12: Reports and Presentation
  - Oct 16-17: Initial Project Reports
  - Oct 18: Presentation
  - Oct 19: Reports modification based on presentation feedbacks
  - Oct 20: Project submission

Details of project diary

Week 9: Topic selection

Sep 25 - Initial Ideas Generation:

Group meeting: Brain storms and idea generation. Potential topics proposed by group members at the initial stage:

- Topic 1: F1 formula races data analysis: driver and team performance comparisons. by Shubo
- Topic 2: New Zealand vehicle crashes reasons analysis: indetifying potential risk factors for vehicles accidents. by Dan Wei
- Topic 3: Treatment response and adverse events analysis for cancer patients who received immunotherapy. by Todd (Zhen) Zhang
- Topic 4: New Zealand cancer risk factors analysis: indetifying potential risk factors for cancer. by Zichen Zhou

Sep 26-28 - Feasibility Evaluation:

Topic background research and initial data collection for each potential topic. Topic 1 by Shubo, Topic 2 by Dan, Topic 3 by Todd, Topic 4 by ZiChen.

Sep 29 - Determination of Finial Topic and Individual Tasks Assignments:

Group meeting:

- **Determination of Finial Topic**: Demonstrating initial results of each potential topic by each members and determining the finial topic: **"New Zealand cancer risk factors analysis"**. Reason for topic selection: 1. Meaningful idea that benefit people's health; 2. Multiple data sources involved in this project, including cancer data, environmental factors data (e.g., earthquake, air quality, temperature…) and social factors data (e.g., income, work hours, smoking, drindking…).
- **Individual Tasks Assignments**:
  - Data collection and documentation: ZiChen
  - Project diary: Dan
  - Data wrangling: Todd, Shubo
  - Cancer Overview analysis: Dan, Todd
  - Cancer risk factors analysis: Shubo, Todd
  - Project Reports:
    - Background: Todd
    - Methods: Zichen, Dan, Shubo, Todd
    - Results: Dan, Shubo
    - Disscussion: Todd
    - Answers to assignments questions
  - Readme file: Shubo

Week 10: Data collection and wrangling:

Oct 2-3 - Data Collection

Zichen:

- Identified and collected cancer incidence data from Te Whatu Ora "Cancer Web Tool" database.
- Identified and collected environmental factors data:
  - Earthquake data from GeoNet website
  - Water quality and air quility data from LAWA database
  - Temperature data from Stats NZ
- Identified and collected social factors data:
  - New Zealand Health survey data from Ministry of Health Website

Oct 4-6 - Data wrangling

Shubo (with Julia):

- Cancer data:
  - format DHB region names
  - filter undefined regions
  - format cancer type names
  - identify characters in numeric variables and replace "S" with 0
  - Organize data into long data format
  - primary key: {DHB, year, sex, cancer}

Todd (with R):

- Health Survey data:
  - filter target population and type
  - modify DHB names to make it match to cancer data
  - extract year from date variable
  - Change data to wide data format, each column represent each risk factor
  - primary key: {DHB, year, sex}
- Environmental data (Earthquake, Air, Temperature, Water) :
  - filter data
  - turn coordinates into sf object
  - map to DHB geometry region
  - summarize data based on {DHB, year} to generate max, min, average, and frequency (only for earthquake data) value.
  - Change data to wide data format, each column represent each risk factor.
  - primary key: {DHB, year}
- Prepare clean data set for correlation analysis:
  - split cancer data into a list by "cancer". This list contains dataframes for each cancer types with primary key {DHB,year,sex}
  - Inner join each dataframe in the list with health survey data using {DHB, year, sex}
  - Inner join each dataframe in the list with environmental data using {DHB,year}

Week 10: Data analysis and project refinement

Oct 9-10: Data analysis

Cancer overview analysis:

- Dan (with R):
  - Age distribution

  - Gender variance of cancer types
  - Regional distribution
  - Most common region for each cancer
  - Temporal Trends of cancer incidence

- Todd (with R):
  - Heatmap plot visulization

Risk factors analysis:

- Shubo:
  - Volcano plot screening significant risk-factor cancer pairs
  - Generate tables for significant risk factors for each cancer types
  - correlation dot plot for better demonstration of the relationship between risk factors and cancer incidence.

Oct 11-15: Add more risk factors data and refine data wrangling process

Use github for team project management

Additional data:

- Zichen: Income, Education, Birth numbers, Work hours data

Refine data wrangling :

- Shubo (with Julia): Based on initial results, filtered cancer data with sex, each cancer only has one corresponding sex group, for example: only keep “female” data for breast cancer, only “male” data for prostate cancer, only “allsex” for lung cancer…’
- Todd (with R): create new final datasets based on sex filtered cancer data.

Refine Risk Factors analysis:

- Shubo (with R): regenerate volcano plot, tables and correlation dot plot based on new data.

Week 12: Reports and Presentation

Oct 16-17: Initial Project Reports

Project report:

- Report frame: Dan
- Background: Todd
- Methods: Zichen, Dan, Shubo, Todd
- Results: Dan, Shubo

- Disscussion: Todd
- Answers to assignment question: Zichen, Todd, Shubo

Data documentation:

- Zichen

PPT for presentation:

- Todd

Presentation Rehearsals within group:

- All members

Oct 18: Presentation

Presentation:

- Todd

Group Meeting for project reflection and lessons learned from feedbacks

- All members

Oct 19: Reports modification based on feedbacks

Project Modification based on presentation feedbacks:

- All members

Readme file:

- Shubo

Oct 20: Project submission

All members