https://edav.info/project.html

01-introduction – Xiaonan (how to render bookdown)

What

Why

How (workflow)

Team member

02-data (xiaonan, Kenny)

<https://edav.info/project.html>

Read data from website

Unzip to csv

Load csv to database mysql - Kenny

03-cleaning/transformation

1. Basic data analysis (outliers) –Gaoyi
2. Remove outliers
3. Data transformation (prepare for 05-results, any transformation put in this part) - Kenny

04-missing - Kenny

Pattern before removing outliers

Pattern after removing (outliers, missing)

05-results -by Monday night (12/7)

Read data from sql (filter: age >16, <100 | tripduration < 24h)

Consistent color for histogram, barchart, boxplot…, color=’black”, fill=’steelblue1’,

<https://www.nceas.ucsb.edu/sites/default/files/2020-04/colorPaletteCheatsheet.pdf>

Data exploration (tidyverse, other packages) + analysis

1. Time series (2018-2020 bike usage) - Jiayin
2. Age/gender vs bike usage (count, tripduration, speed, distance) - Gaoyi, cleverland
3. Weekday/weekend vs bike usage (hourly count) --Jiayin
4. travel between Manhattan and the outer boroughs, average hourly (choose one week, only weekdays) – zipcode dictionary (sql database) -- Xiaonan
5. weather vs bike usage (temperature, rain/snow) – python API -- Xiaonan
6. Top 5th/the most popular and the least popular trips (year, month) – Kenny

Most popular station

cleverland

Map (R package ggvis) <https://rpubs.com/jiayiliu/ggmap_examples>

Heatmap (count, dentity)

(A-> B) top 5th- paths - google map

1. Membership types - subscriber/customer vs bike usage (simple analysis is enough, mosaic) – Gaoyi

1: membership

2: Gender

3: Age

…..

Other plots: Likert

06-interactive

Map

Speed (age vs speed)

07-conclusion