initial creation of files + data

Version: 1.0

RestoreWorkspace: Default SaveWorkspace: Default AlwaysSaveHistory: Default

EnableCodeIndexing: Yes UseSpacesForTab: Yes NumSpacesForTab: 2 Encoding: UTF-8

RnwWeave: Sweave LaTeX: pdfLaTeX

title: "p8105_hw3_hs3478" author: "Huachen Shan" date: "2024-10-10" output: github_document always_allow_html: true

In [4]:
```
#```{r setup, include=FALSE}
# library(tidyverse)
# library(haven)
# library(kableExtra)
# library(leaflet)
# library(p8105.datasets)
# library(patchwork)
```

In [ ]:
```
knitr::opts_chunk$set(
        echo = TRUE,
        warning = FALSE,
  fig.width = 8,
  fig.asp = .6,
  out.width = "100%",
        dpi=300
)

theme_set(theme_minimal() + theme(legend.position = "bottom"))

options(
  ggplot2.continuous.colour = "viridis",
  ggplot2.continuous.fill = "viridis"
)

scale_colour_discrete = scale_colour_viridis_d
scale_fill_discrete = scale_fill_viridis_d
```

In [ ]:
```
## Problem 0:

# Create a public GitHub repo + local R Project: p8105_hw3_hs3478
# Create a single .Rmd file named p8105_hw3_hs3478.Rmd that renders to github_document
# Create a subdirectory (data) to store the local data files, and use relative paths to access these data files
# Submit a link to your repo via Courseworks: https://github.com/huachenshan/p8105_hw3_hs3478
```

In [ ]:

Problem 1:

```{r prob1_import, message=FALSE}
data("ny_noaa")

summary(ny_noaa)

ny_noaa =
  ny_noaa %>%
  mutate(
    tmin = as.numeric(tmin),
    tmax = as.numeric(tmax)
  ) %>%
  relocate(tmin, .before=tmax) %>%
  separate(date, c("year", "month", "date"), convert=TRUE)
```

Average Max Temp in Jan/July

```{r prob1_avgmax, message=FALSE}
ny_noaa %>%
  drop_na(tmax) %>%
  filter(
    month == 1 | month == 7
  ) %>%
  group_by(id, year, month) %>%
  summarise(mean_tmax = mean(tmax, na.rm=TRUE)) %>%
  ggplot(aes(x = year, y=mean_tmax, group = id)) +
  geom_point() + geom_line()+
  facet_grid(~month)
```

TMin vs TMax & Distribution of Snowfall

```
plot1 =
  ny_noaa %>%
```

```
  ggplot(aes(x=tmin, y=tmax)) +
  geom_hex()+
  labs(
    title = "Min Temp vs Max Temp ",
    x = "Tenths of Degrees in C",
    y = "Tenths of Degrees in C"
    ) +
  theme(plot.title = element_text(hjust = 0.5))

plot2 =
  ny_noaa %>%
  filter (
    snow > 0,
    snow < 100
    ) %>%
  group_by(year) %>%
  ggplot(aes(x=snow, fill = as.factor(year))) +
  geom_density(alpha = 0.4)+
  labs(
    title = "Distribution of Snowfall 0-100 mm",
    x = "Years",
    y = "Density"
    ) +
  theme(plot.title = element_text(hjust = 0.5))

plot1 + plot2
```

In [ ]:

```
## Problem 2:

```{r problem2_importdata, message=FALSE}

nhanes_demo = read_csv("data/nhanes/nhanes_covar.csv",
                       skip = 4) %>%
  janitor::clean_names() %>%
  drop_na() %>%
  filter(
    age >= 21
  ) %>%
  mutate(
    sex = ifelse(sex == 1, "Male", "Female"),
    education =
      case_match(
        education,
        1 ~ "Less than high school",
        2 ~ "High school equivalent",
        3 ~ "More than high school"
      ),
    education = factor(education, levels = c("Less than high school", "High school equivalent", "More than high school"))
  )

nhanes_acc_data = read_csv("data/nhanes/nhanes_accel.csv", show_col_types = FALSE) %>%
  janitor::clean_names() %>%
  distinct() %>%
  pivot_longer(
    cols = min1:min1440,
    names_to = "minute",
    values_to = "mims_data",
    names_prefix = "min"
  ) %>%
  mutate(minute = as.numeric(minute))

# join the demographic and accelerometer datasets
nhanes = inner_join(nhanes_demo, nhanes_acc_data, by="seqn")
# there are 22 patients who don't have any demographic data
```

Reader friendly table for the number of men and women in each education category, and create a visualization of the age distrib
utions for men and women in each education category.
```

```{r prob2_genderedu}

nhanes %>%
  distinct(seqn, .keep_all = TRUE) %>%
  group_by(education, sex) %>%
  count() %>%
  arrange(education) %>%
  pivot_wider(
    names_from = sex,
    values_from = n
  )%>%
  knitr::kable(col.names = c("Education Level", "Female (n)", "Male (n)"), caption = "NHANES Accelerometer Demographics Breakdown")

nhanes %>%
  distinct(seqn, .keep_all = TRUE) %>%
  group_by(education, sex) %>%
  ggplot(aes(x = age, fill = sex)) +
  geom_density(alpha = .3) +
  facet_grid(. ~ education) +
  labs(
    title = "Density Plots of Ages Across Genders & Education Levels",
    x = "Age (Years)",
    y = "Density"
    ) +
  theme(plot.title = element_text(hjust = 0.5))

```

Comments about the Total Activity Across Ages, comparing men and women and for each education level:


```{r prob2_total_activity, message=FALSE}

nhanes =
  nhanes %>%
  group_by(seqn) %>%
  mutate(
    total_activity = sum(mims_data)
  )

nhanes %>%
```

```
  distinct(seqn, .keep_all = TRUE) %>%
  ggplot(aes(x = age, y = total_activity, color = sex)) +
  geom_point(alpha = .8) +
  geom_smooth(se = FALSE) +
  facet_grid(. ~ education) +
  labs(
    title = "Scatterplots of Ages x Total Activity Across Education Levels",
    x = "Age (Years)",
    y = "Total Activity (MIMS-Unit)",
    colour="Sex"
    ) +
  theme(plot.title = element_text(hjust = 0.5))
```

Accelerometer Data Over 24 Hours for each education level and across sexes.

```{r prob2_acceldata_24hours, message=FALSE}

nhanes %>%
  ggplot(aes(x = minute, y = mims_data, color = sex)) +
  geom_point(size = 0.001, alpha = 0.2) +
  geom_smooth(se = FALSE, linewidth=0.5) +
  facet_grid(. ~ education) +
  labs(
    title = "24 Hour Accelerometer Data Across Education Levels by Sex",
    x = "Time (Minutes)",
    y = "Monitor-Independent Movement Summary (MIMS-unit)",
    colour="Sex"
    ) +
  theme(plot.title = element_text(hjust = 0.5))

```

In [ ]:

Problem 3: Citi Bikes

```{r prob3_import, message=FALSE}
citi_bike_jan2020 = read_csv("data/citibike/Jan 2020 Citi.csv") %>%
  janitor::clean_names() %>%
  mutate(
    year = 2020,
    month = "January")

citi_bike_jan2024 = read_csv("data/citibike/Jan 2024 Citi.csv") %>%
  janitor::clean_names()%>%
  mutate(
    year = 2024,
    month = "January")

citi_bike_july2020 = read_csv("data/citibike/July 2020 Citi.csv") %>%
  janitor::clean_names() %>%
  mutate(
    year = 2020,
    month = "July")

citi_bike_july2024 = read_csv("data/citibike/July 2024 Citi.csv") %>%
  janitor::clean_names() %>%
  mutate(
    year = 2024,
    month = "July")


citi_bike =
  bind_rows(citi_bike_jan2020, citi_bike_july2020, citi_bike_jan2024, citi_bike_july2024) %>%
  relocate(
    year, month
  ) %>%
  mutate(
    weekdays = factor(weekdays, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))
  )

#colSums(is.na(citi_bike))
```

```{r prob3_year_month_combos}

citi_bike %>%
  group_by(year, month, member_casual) %>%
  count() %>%
  arrange(year) %>%
  pivot_wider(
    names_from = member_casual,
    values_from = n
  ) %>%
  knitr::kable(col.names = c("Year", "Month", "Casual Rides (n)", "Member Rides (n)"), caption = "Citi Bike Rides by Year, Month, and Rider Type") %>%
  collapse_rows(columns = 1)
```

Here are the top 5 popular starting stations for July 2024.

* The most popular station is Pier 61 at Chelsea Piers, with 163 rides originating there.

```{r prob3_top5_start_stations}
citi_bike %>%
  filter(
    year == 2024,
    month == "July"
  ) %>%
  group_by(start_station_name) %>%
  count() %>%
  arrange(desc(n)) %>%
  head(5) %>%
  knitr::kable(col.names = c("Starting Station", "Number of Rides Originating Here"), caption = "The 5 Most Popular Starting Citi Bike Stations of July 2024")
```

(Saturday and Sunday). However, 2024 had Sundays with a smaller median than Saturdays and even dipped to the lowest of all days for January 2024.

```{r prob3_median, message=FALSE}

citi_bike %>%
  group_by(weekdays, month, year) %>%
  summarize(median = median(duration)) %>%
  ggplot(aes(x = weekdays, y = median, group=1)) +
```

```
  geom_line()+
  geom_point() +
  facet_grid(month ~ year) +
  labs(
    title = "Effects of Day of Week, Month, and Year on Median Ride Duration",
    x = "Weekday",
    y = "Median Ride Duration (Minutes)"
    ) +
  theme(
    panel.background = element_rect(fill = NA, color = "black"),
    axis.text.x=element_text(angle = 45, vjust = 1, hjust=1),
    plot.title = element_text(hjust = 0.5)
  )


```


```{r prob3_duration, message=FALSE}

citi_bike %>%
  filter(
    year == 2024
  ) %>%
  group_by(month, member_casual, rideable_type) %>%
  ggplot(aes(x = month, y = duration, fill = month)) +
  geom_violin(alpha=0.4)+
  facet_grid(member_casual ~ rideable_type) +
  labs(
    title = "Effects of Month, Membership Status, and Bike Type on the Distribution of Ride Duration",
    x = "Month",
    y = "Ride Duration (Minutes)",
    colour="Month"
    ) +
  theme(
    panel.background = element_rect(fill = NA, color = "black"),
    plot.title = element_text(hjust = 0.5)
  )


```