p8105_hw3_hs3478

Huachen Shan 2024-10-10

Problem 0:

In [ ]:
```
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 t
Submit a link to your repo via Courseworks: https://github.com/rysgpd/p8105_hw3_hs34
```

Problem 1:

The dataset consists of 7 columns, including variables for date, precipitation (in tenths of millimeters), snowfall (in millimeters), snow depth (in millimeters), minimum temperature (in tenths of degrees Celsius), and maximum temperature (in tenths of degrees Celsius), with a total of 2,595,176 rows of data. There are 747 unique IDs. Upon reviewing the summary, snowfall and snow depth contain a significant number of zeros and missing values, with 2,008,508 zeros and 381,221 missing values. The high number of zeros is reasonable, as snow is absent in most months, whereas rainfall has fewer zeros since it can occur throughout the year.

In [2]:
```
##       id                 date                 prcp               snow        
##  Length:2595176     Min.   :1981-01-01   Min.   :    0.00   Min.   :  -13  
##  Class :character   1st Qu.:1988-11-29   1st Qu.:    0.00   1st Qu.:    0  
##  Mode  :character   Median :1997-01-21   Median :    0.00   Median :    0  
##                     Mean   :1997-01-01   Mean   :   29.82   Mean   :    5  
##                     3rd Qu.:2005-09-01   3rd Qu.:   23.00   3rd Qu.:    0  
##                     Max.   :2010-12-31   Max.   :22860.00   Max.   :10160  
##                                          NA's   :145838     NA's   :381221 
##       snwd            tmax               tmin          
##  Min.   :   0.0   Length:2595176     Length:2595176    
##  1st Qu.:   0.0   Class :character   Class :character  
##  Median :   0.0   Mode  :character   Mode  :character  
##  Mean   :  37.3                                        
##  3rd Qu.:   0.0                                        
##  Max.   :9195.0                                        
##  NA's   :591786
```

Average Max Temp in Jan/July

The plot is quite cluttered due to the presence of over 700 IDs, making it difficult to interpret any clear structure beyond the general similarity in their fluctuations. January tends to be cooler than July, which is expected. There is a notable outlier in July during the late 1980s where temperatures are unusually lower compared to the rest. Similarly, another outlier appears in the early 1980s in January, where it's also colder than usual. Aside from these, there aren't many extreme outliers.

TMin vs TMax & Distribution of Snowfall

Based on the hex plot, it's clustered in the middle but there is variability on all aspects. Typically, it looks like as min goes up, max also goes up, but I'm confused at the outlier on the bottom (ex: -250 min and -400 max) which doesn't make sense. The density plot shows that the general pattern is very similar to each year, but the peaks are getting lower over time. There are many peaks (multimodal).

Problem 2:

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

The gender distribution seems fairly balanced across most education categories, except for the High School Equivalent group, where there are more males. There are also significantly more participants with education levels beyond high school—about twice as many compared to the other two categories, following a 1:1:2 ratio.

I opted for a density plot to visualize the age distributions for men and women in each education category, as the histogram was too rigid and hard to interpret. By overlapping the density plots, it became easier to compare the age distributions between genders.

For men, the distributions across all education levels appear to be bimodal, while for women, they are unimodal. In the "Less than High School" category, both men and women tend to be older, with the distribution skewed toward older populations. In the "High School Equivalent" category, women's distribution steadily rises and peaks around age 70, while for men, the distribution remains relatively flat until it declines after age 60.

In the "More than High School" category, the density for women peaks around age 30, with a value of approximately 0.025, indicating more younger women pursuing higher education. A similar trend is seen for men, though the peak is not as pronounced, with younger men showing higher density than older men in this group.

In [ ]:

Education Level	Female (n)	Male (n)
Less than high school	28	27
High school equivalent	23	35
More than high school	59	56

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

For all education levels and both sexes, there is a generally downward trend of total activity (minutes) measured. There are a few exceptions that have slight peaks around ages 60 (less than high school for both, more than high school for men) or 70 (high school equivalent). For high school equivalent and more than high school, women are almost always higher than men. For less than high school, younger women have higher activity until age 40, then the men have higher total activity. The trend lines start at the highest for less than high school, but ends at the lowest also for the less than high school level. More than high school seems to have less variability in general, but has the most extreme outliers.

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

From the plot below, it looks like both female and male populations for each education level showed very similar trends and patterns. Especially for the Less than high school level, those lines are practically the same. For the other two, women seem to have a slightly higher MIMS level for the entire day until the end when the lines converge again. The most separation is in the more than high school group. It's a general pattern where most people have low MIMS activity at the beginning/midnight (first 250 minutes), which makes sense because the activity begins to pick up quickly when people start to wake up. For less than high school, the peak is around noon and slowly goes down after that until the sharp drop off around 10 pm. The high school equivalent level has a more stable pattern and stays pretty stagnant until a little earlier than the less than high school group. Lastly, the more than high school level has a little higher plateau than the rest of the levels and similarly dips at the time that high school equivalent does. Lastly, the range for more than high school is the largest out of the three. Women seem to have more activity in the morning, while men have more activity in the evening. These go up to about 80 MIMS units.

Problem 3: Citi Bikes

Describing the resulting dataset:

This dataset has 99485 unique rides after combining all four data files. There are 33468 and 66017 rides in the months of January/July of 2020 and 2024 respectively. There are 79862 member rides total and 19623 casual rides total. The range of duration is from 1.00165 to 238.7798333 minutes. There are 43 rides without a start_station_name and 207 rides without an end_station_name. I have decided to keep these rides in the dataset because there are many questions that don't involve the station names, and all other variables are included.

A reader-friendly table for the total number of rides in each combination of year and month (January 2020, July 2020, January 2024, and July 2024).

This table shows that for each year/month combination, the number of member rides were overwhelmingly more than casual rides. By July 2024, there were about 3x as many member rides than casual. The number of rides from each year and month increased each time for both rider groups except for between July 2020 and January 2024 - casual rides decreased and member rides barely increased. This, however, does make sense because of the weather conditions, alongside more visitors, that make biking more enjoyable in the summer than winter months. The number of rides in July 2024 was almost 50,000 which shows how popular Citi Biking has become in New York City.

Looking at the effects of the day of the week, month, and year on median ride duration:

The impact of month, membership status, and bike type on the distribution of ride duration:

I chose to do violin plots because it shows the distribution and is clear to understand compared to box plots which were tried, but the IQR was so small that it was messy. All of the medians are very similar, so none of the factors really affect that. If anything, July is slightly higher than January for both years. The width of January's violin plots is larger than July's, showing a larger distribution of rides with smaller duration. July's, especially with electric bikes, had a less wide base and had more of the intermediate durations. For both Julys, there were many outliers that extended the range and length of the needle, which was very high for electric bikes. The range of electric bikes in January were the smallest range, even smaller for casual riders, which may make sense because electric bikes made the ride faster/shorter and they are more expensive by the minute. Classic bikes had longer ranges of durations in January than July, but it looks like it was only a few outliers instead of July which had a solid number of outliers at the extremes. There also seemed to be a smaller second peak (bimodal) for the classic/January/casual combo around 75 minutes.

In [ ]: