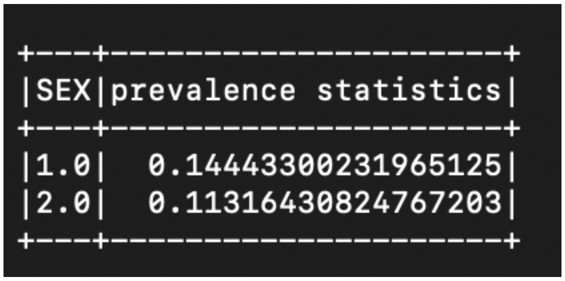
CS5304 Spring 2020
Assignment 1: Data Preparation
Queenie Liu (ql299)

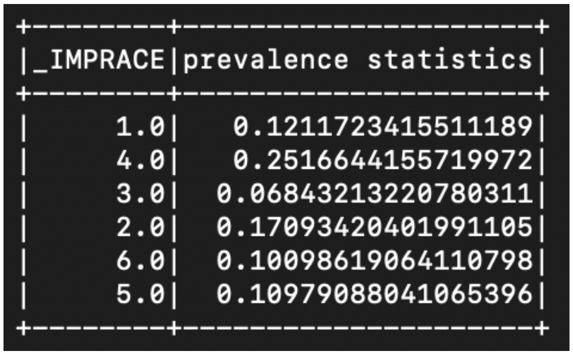
Question 1: Extract, Transform Load (ETL)

- Step 5:

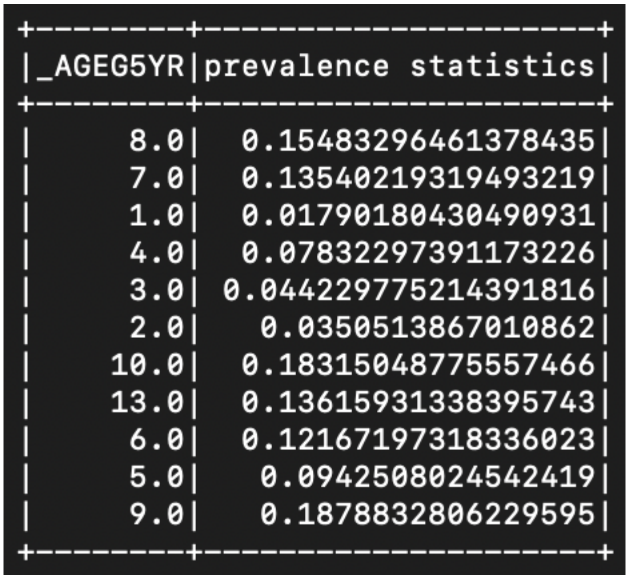
Result:



```
+---+---------------------+
|SEX|prevalence statistics|
+---+---------------------+
|1.0|  0.14443300231965125|
|2.0|  0.11316430824767203|
+---+---------------------+
```



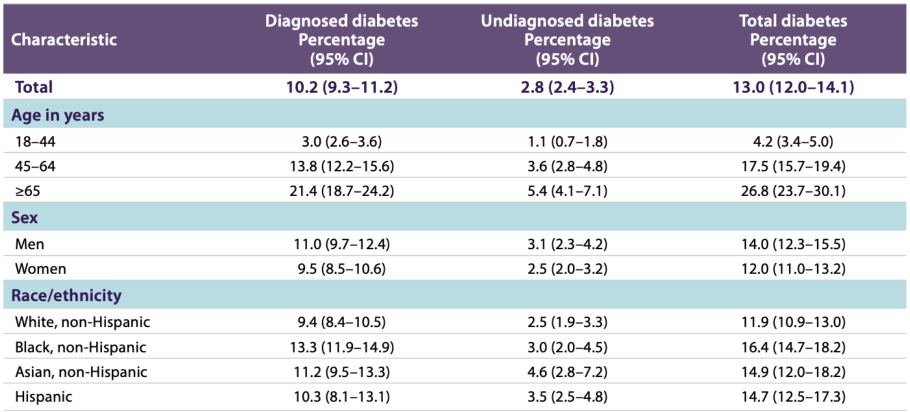
```
+--------+---------------------+
|_IMPRACE|prevalence statistics|
+--------+---------------------+
|     1.0|   0.1211723415511189|
|     4.0|   0.2516644155719972|
|     3.0|  0.06843213220780311|
|     2.0|  0.17093420401991105|
|     6.0|  0.10098619064110798|
|     5.0|  0.10979088041065396|
+--------+---------------------+
```



```
+--------+---------------------+
|_AGEG5YR|prevalence statistics|
+--------+---------------------+
|     8.0|  0.15483296461378435|
|     7.0|  0.13540219319493219|
|     1.0|  0.01790180430490931|
|     4.0|  0.07832297391173226|
|     3.0| 0.044229775214391816|
|     2.0|   0.0350513867010862|
|    10.0|  0.18315048775557466|
|    13.0|  0.13615931338395743|
|     6.0|  0.12167197318336023|
|     5.0|   0.0942508024542419|
|     9.0|   0.1878832806229595|
+--------+---------------------+
```

CDC:

Table 1a. Estimated crude prevalence of diagnosed diabetes, undiagnosed diabetes, and total diabetes among adults aged 18 years or older, United States, 2013–2016



Characteristic	Diagnosed diabetes Percentage (95% CI)	Undiagnosed diabetes Percentage (95% CI)	Total diabetes Percentage (95% CI)
Total	**10.2 (9.3–11.2)**	**2.8 (2.4–3.3)**	**13.0 (12.0–14.1)**
Age in years			
18–44	3.0 (2.6–3.6)	1.1 (0.7–1.8)	4.2 (3.4–5.0)
45–64	13.8 (12.2–15.6)	3.6 (2.8–4.8)	17.5 (15.7–19.4)
≥65	21.4 (18.7–24.2)	5.4 (4.1–7.1)	26.8 (23.7–30.1)
Sex			
Men	11.0 (9.7–12.4)	3.1 (2.3–4.2)	14.0 (12.3–15.5)
Women	9.5 (8.5–10.6)	2.5 (2.0–3.2)	12.0 (11.0–13.2)
Race/ethnicity			
White, non-Hispanic	9.4 (8.4–10.5)	2.5 (1.9–3.3)	11.9 (10.9–13.0)
Black, non-Hispanic	13.3 (11.9–14.9)	3.0 (2.0–4.5)	16.4 (14.7–18.2)
Asian, non-Hispanic	11.2 (9.5–13.3)	4.6 (2.8–7.2)	14.9 (12.0–18.2)
Hispanic	10.3 (8.1–13.1)	3.5 (2.5–4.8)	14.7 (12.5–17.3)

Sex: '1' = male
'2' = female
Race: '1' = 'White, non-Hispanic'
'2' = 'Black, non-Hispanic'

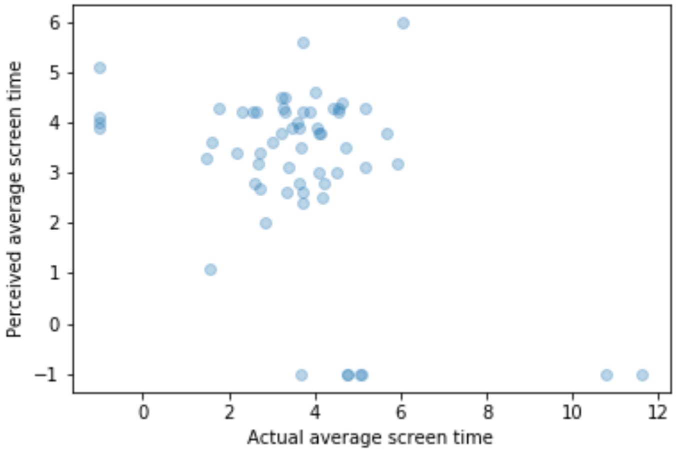
‘3’ = 'Asian, non-Hispanic'
‘4’ = 'American Indian, non-Hispanic'
‘5’ = 'Hispanic'
‘6’ = 'Other, non-Hispanic’

According to my mapping, sex and age are quite similar to the CDC Statistics. There are some variance regarding to ‘Asian’ and ‘Hispanic’. These might be resulted from I have one more category towards race which is ‘other’. To improve the accuracy, I might implement 95% confidence interval.

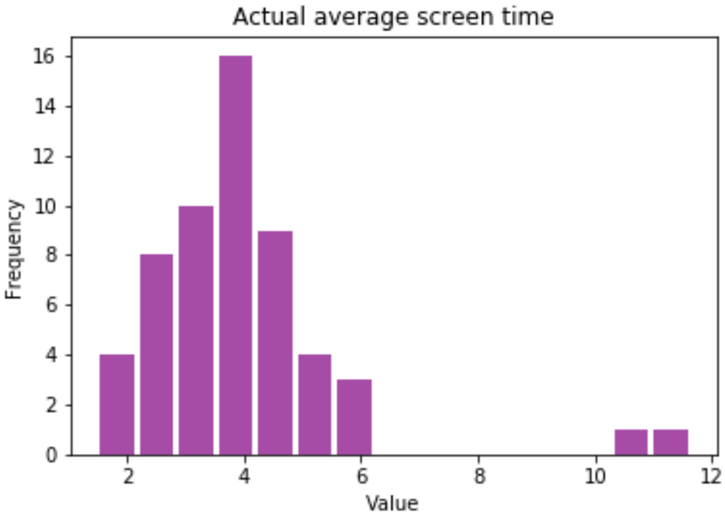
Question 2: Dealing with messy and missing data

- Case 1: Actual screen time.

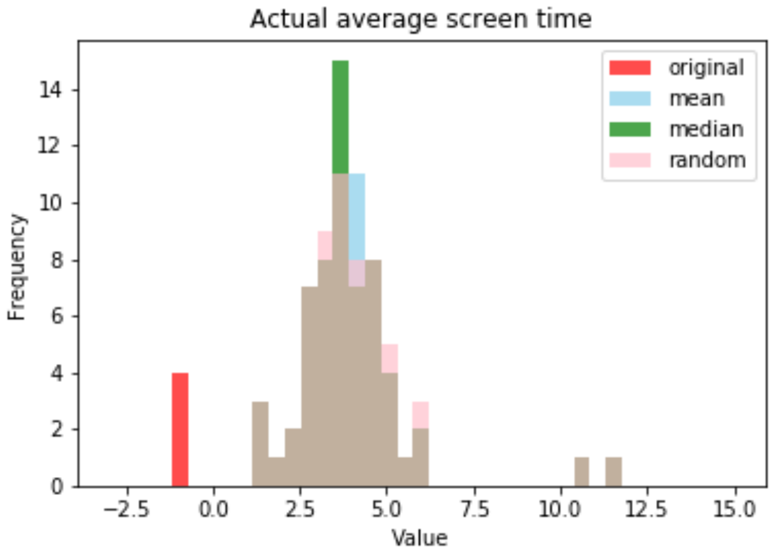
A. I plot a scatterplot of both the actual screen time and the perceived screen time as the following. From the plot, I can discover that the missing values for this feature are represented as negative values.



B. By removing the missing values, I plot a histogram for the actual average screen time. It does have outliers and is skewed which is a right-skewed distribution that has positive skewness.

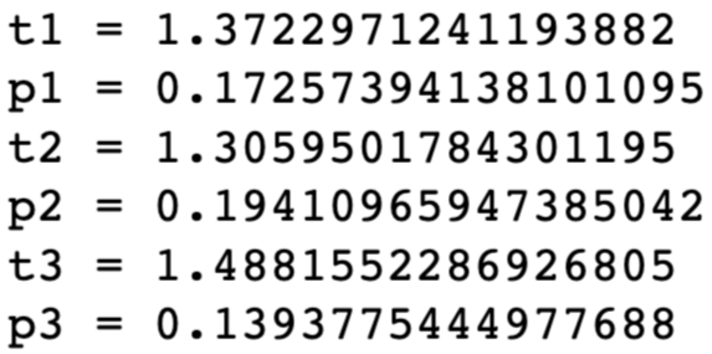


C. I overlay a figure with the original distribution and the 3 new distributions. For method 3, I choose the random value from a range from 2 to 6 since from the original distribution we can find that most participants have an actual average screen time between 2 and 6. The distributions are slightly similar to each other.



D. ‘p1’ represents the p-value between mean and this population distribution.
‘p2’ represents the p-value between median and this population distribution.
‘p3’ represents the p-value between random value and this population distribution.

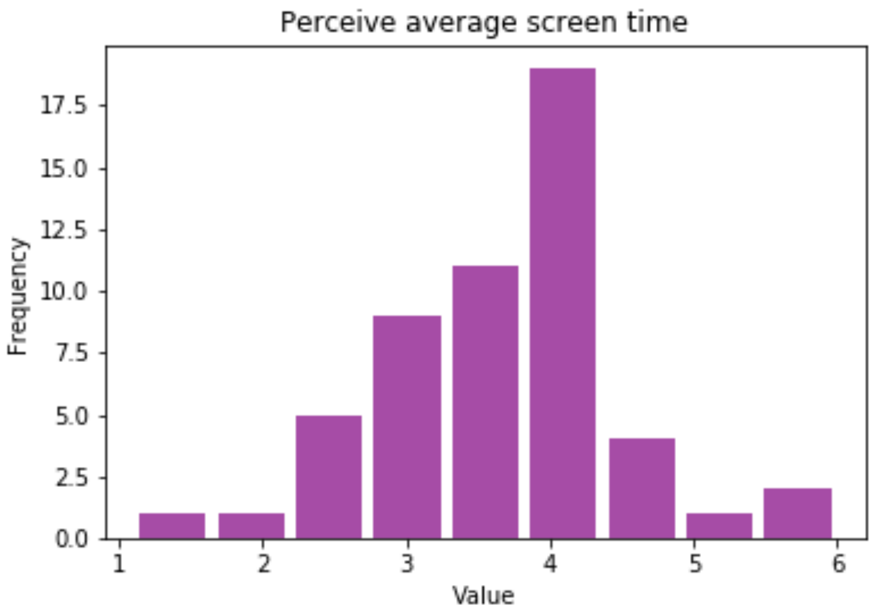
As we can see, although the value might change since the population distribution is generated by np.random, the relationship between these 3 p-values is always [p3<p1<p2]. This means that the distribution filling with median is closest to the original distribution.



```
t1 = 1.3722971241193882
p1 = 0.17257394138101095
t2 = 1.3059501784301195
p2 = 0.19410965947385042
t3 = 1.4881552286926805
p3 = 0.1393775444977688
```

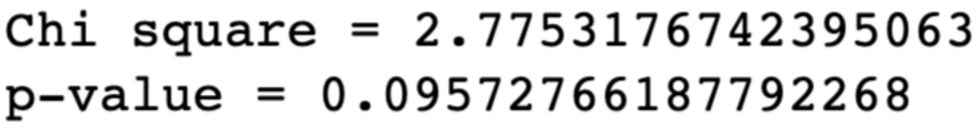
- Case 2:

  A. After removing the missing values, there are outlierS in this data. This distribution is left-skewed with negative skewness.



  B. According to the calculations through code, 4 of them are intense phone users.

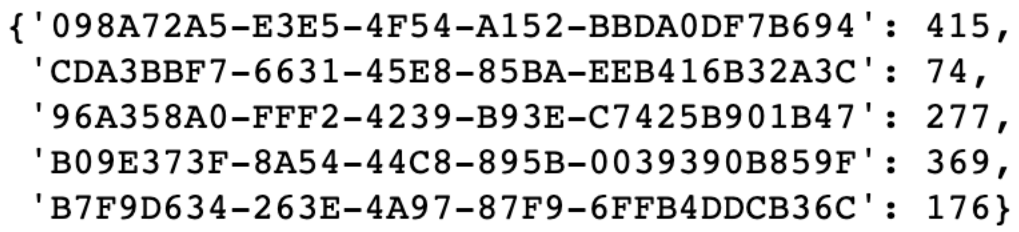
  C. The p-value is 0.09572766187792268. Since the p-value is larger than 0.05, there is no confident evidence showing that they are independent. Thus, it is possible to have correlation. This feature might be MNAR.



```
Chi square = 2.775317674239506З
p-value = 0.09572766187792268
```

- Case 3:

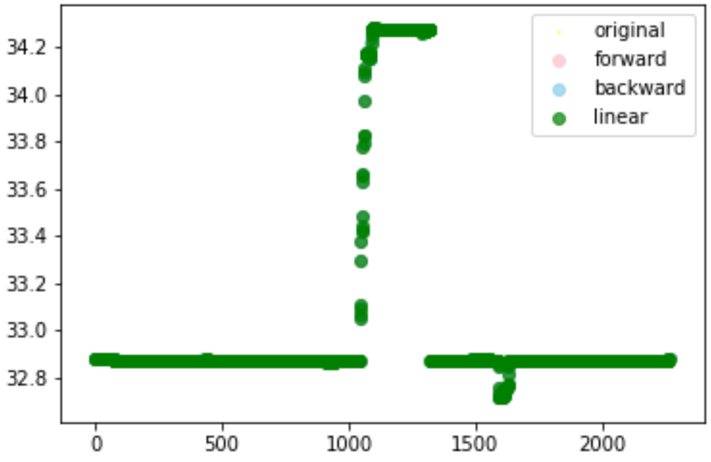
  A. I calculate by extracting all minutes the location is missing and the battery level is below 20% and then check if they matched. If so, I report this participant as the type of

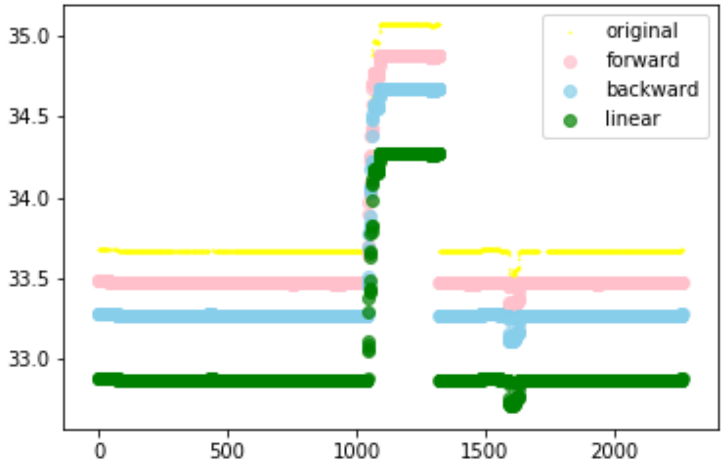


```
{'098A72A5-E3E5-4F54-A152-BBDA0DF7B694': 415,
 'CDA3BBF7-6631-45E8-85BA-EEB416B32A3C': 74,
 '96A358A0-FFF2-4239-B93E-C7425B901B47': 277,
 'B09E373F-8A54-44C8-895B-0039390B859F': 369,
 'B7F9D634-263E-4A97-87F9-6FFB4DDCB36C': 176}
```

participant we want to find. Their uuid and the length in minutes is represented in the above screenshot.

B. The four traces are completely overlapping from the left plot. To better analyze, I add {0.8, 0.6, 0.4} to make it obvious to see the pattern for each trace. If I was to use this dataset for further analysis, I will choose linear interpolation.

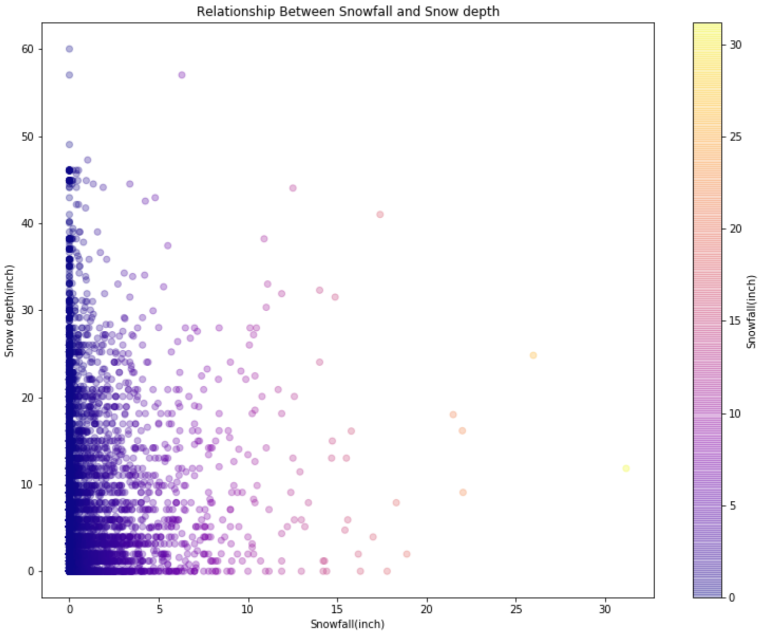




Question 3: Data visualization

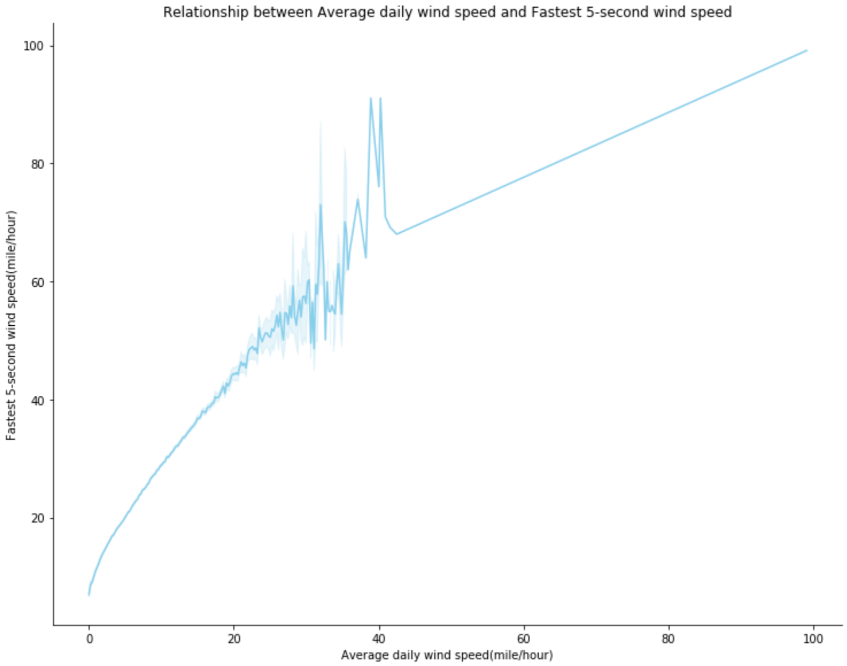
- Plot 1: Relationship between snowfall and snow depth

I use matplotlib to plot a scatterplot with x denotes snowfall in inches and y denotes snow depth in inches. From the plot, we can observe that most of the stations has ranges of snowfall between 0-10 inches and snow depth between 0-30 inches.



- Plot 2: Relationship between average daily wind speed and fastest 5-second wind speed.

I use seaborn to draw a relplot with x denotes average daily wind speed (mile/hour) and y denotes fastest 5-second wind speed. From the plot, we can obviously find that it seems like a logarithmic function which is strictly increasing but the fastest 5-second wind speed tends to be closer to the average daily wind speed as it continues to increase.


Relationship between Average daily wind speed and Fastest 5-second wind speed
Fastest 5-second wind speed(mile/hour)
Average daily wind speed(mile/hour)