p2_write_up

Name: Yue Liu

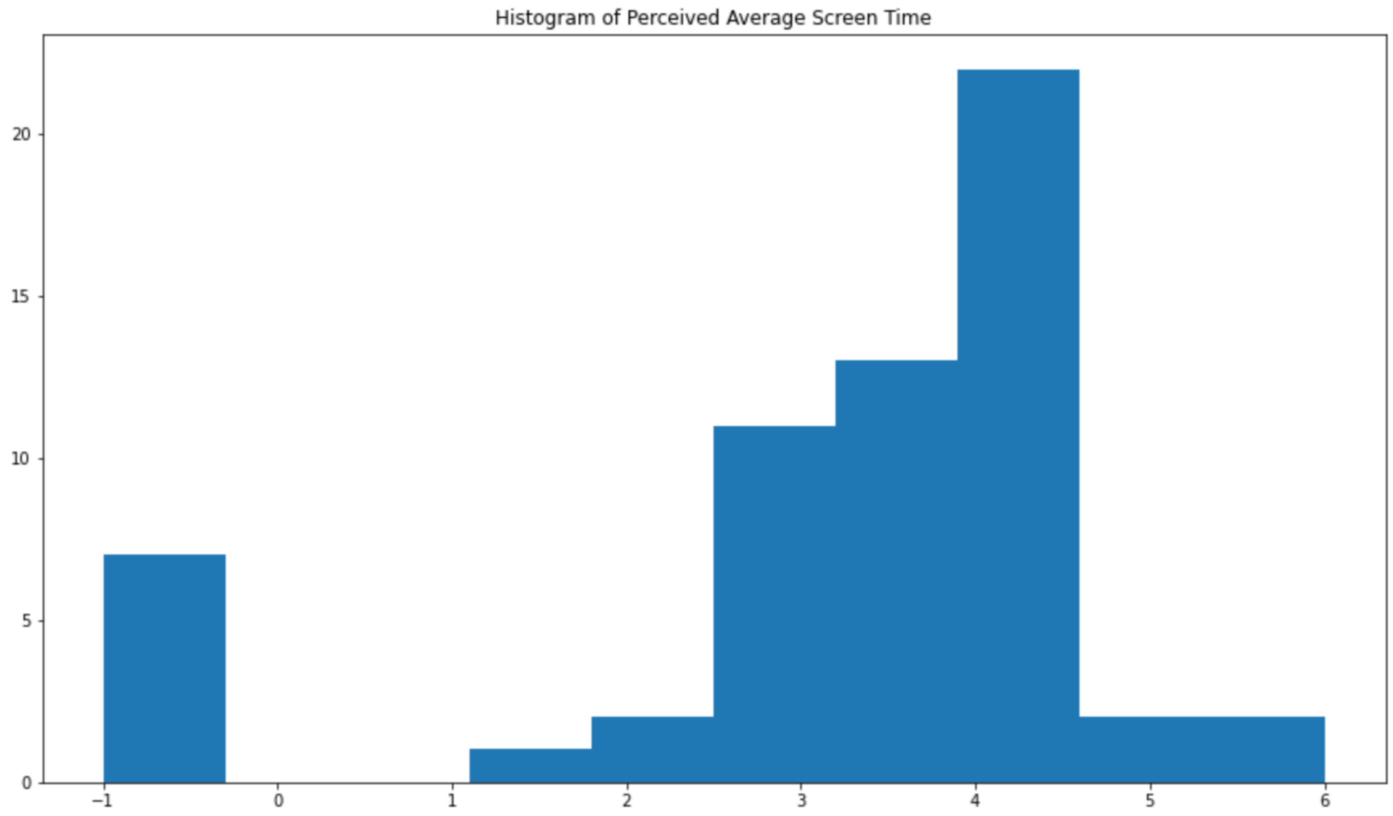
NetID: yl992

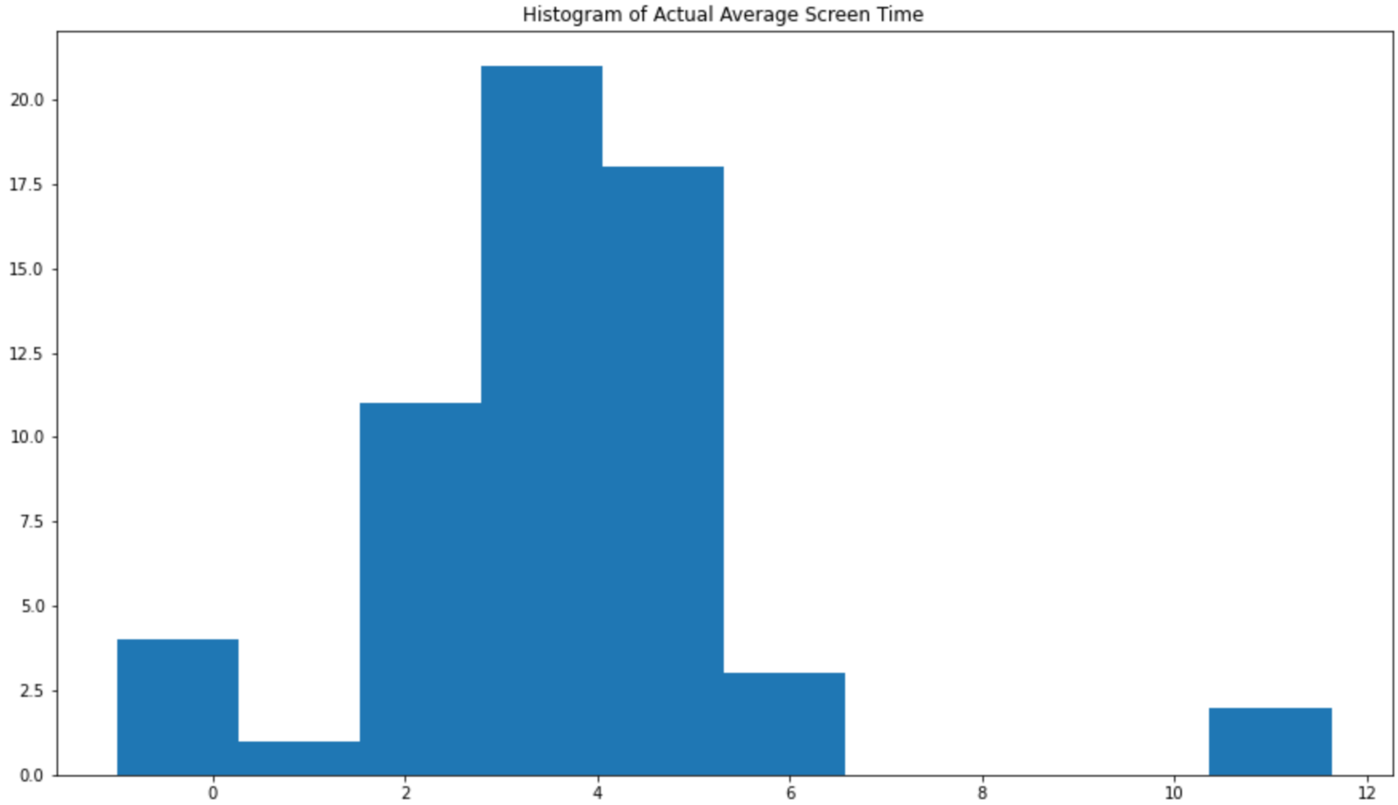
Question 2

Case 1

Writeup Answer to Problem A: How are missing values represented for this feature?

From the histogram, the missing value for perceived average screen time and actual average screen time is set to -1.

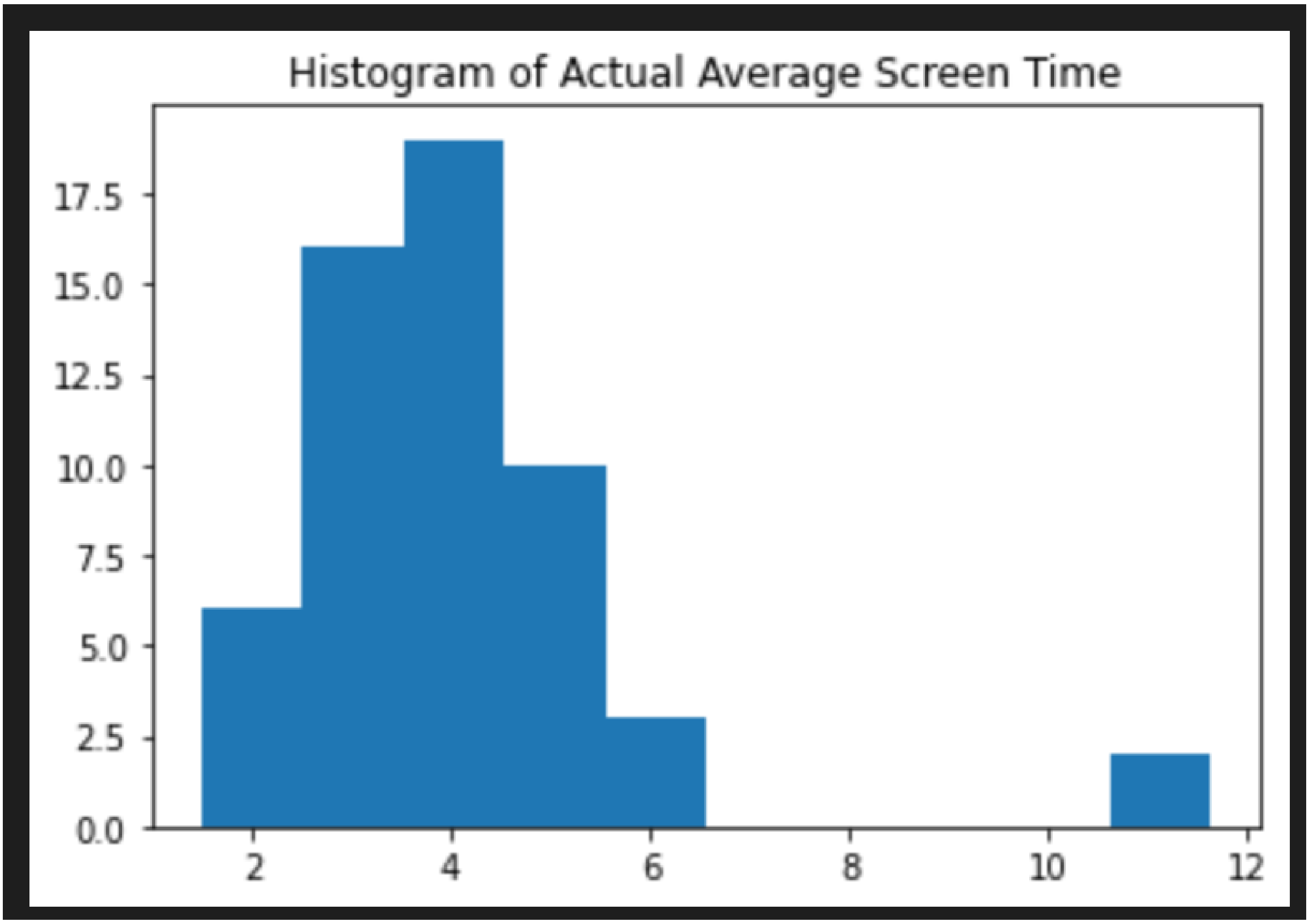




Writeup Answer to Problem B: Does it have outliers? If so, how many? Is it skewed? If so, is it left skewed or right skewed? What’ s the skewness?

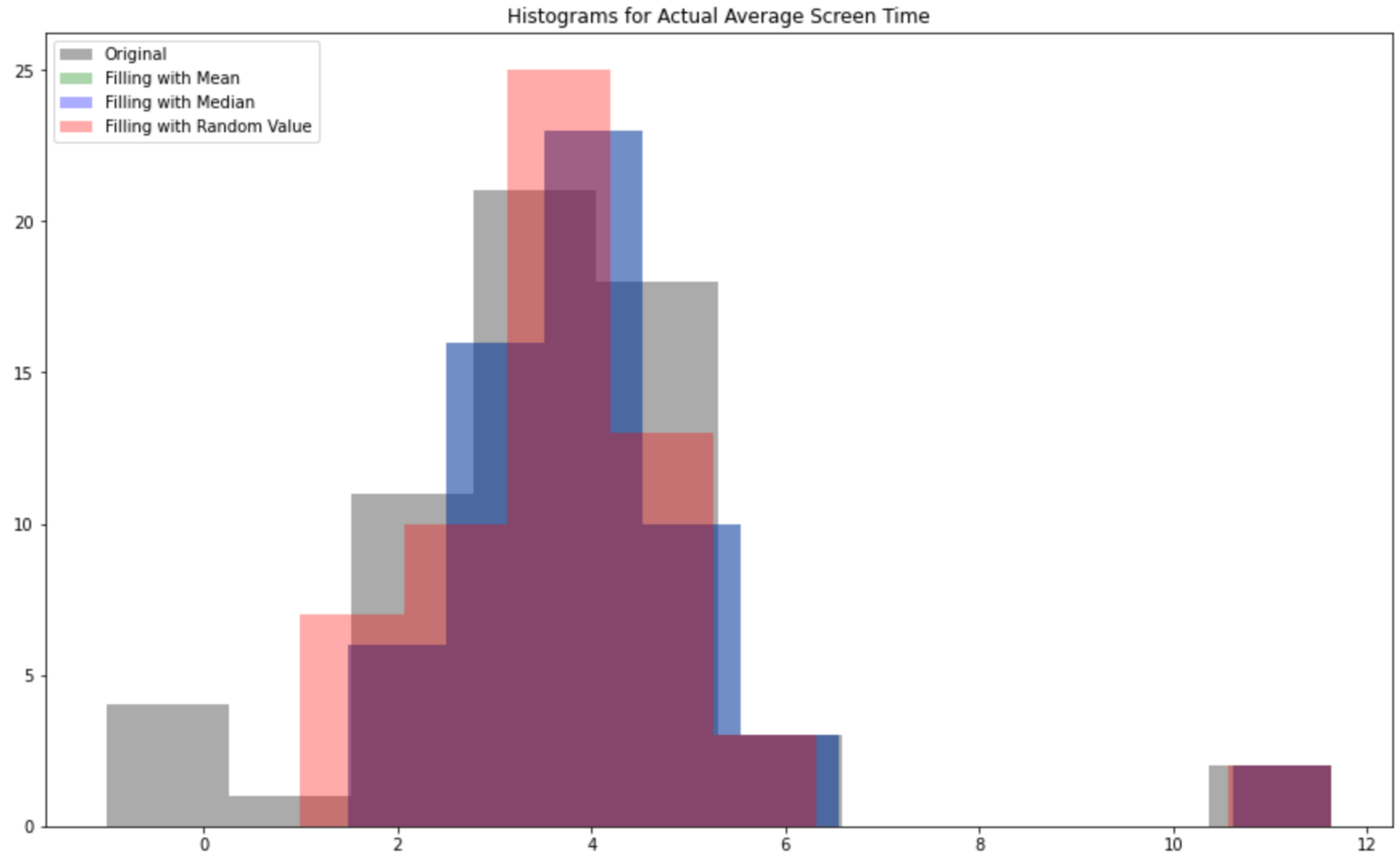
Yes it has 2 outliers, 10.78 and 11.63.

The data is highly skewed to the right with a skegness of 2.46. It is quite apparent simply looking at the histogram blow.



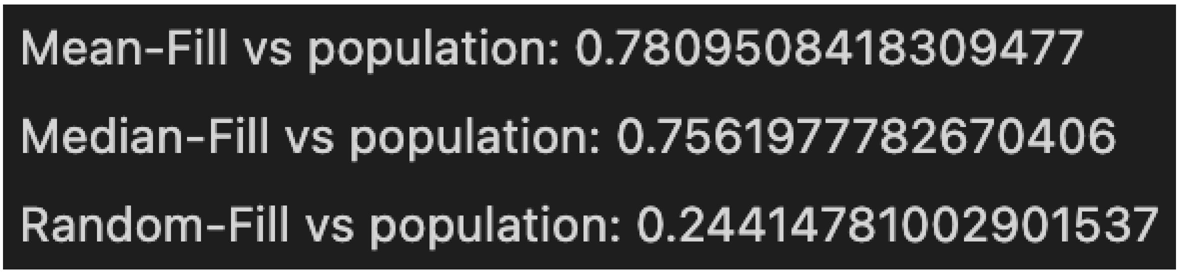
Writeup Answer to Problem C: How did you choose the random value from method 3)? How do the distributions look like after you implement the three filling methods? (Compare them)

The random value is a random number from 0 to the maximum number of filtered dataset without outliers. The data looks to have less variance, range, and noise after using the filling methods.



Answer to Problem D: Report the three p-values. Which one of the filling methods reconstruct this feature to be closest to the research distribution? Why do you think this is the case?

From the calculated p-values, filling with mean reconstructing this feature to be closest to the research distribution because it has the highest P value. A higher P value indicates that the distribution is closer to the true population.



```
Mean-Fill vs population: 0.7809508418309477
Median-Fill vs population: 0.7561977782670406
Random-Fill vs population: 0.24414781002901537
```

Case 2

Writeup Answer to Problem A: Does it have outliers?If so, how many? Is it skewed? If so, is it left skewed or right skewed? What’ s the skewness?

Yes, there are two outliers, 1.1 and 6.0. The data is slightly skewed to the left with skewness value of -0.214.

Writeup Answer to Problem B: How many of them are intense phone users?

There are 5 intense phone users.

Writeup Answer to Problem C: What is the p-value? Do you think they are correlated? What does this mean? Do you think this feature is MAR or MNAR?

The P value is 0.99, which is greater than 0.05, I fail to reject the null hypothesis, therefore, they are correlated and the feature is MNAR.

Case 3

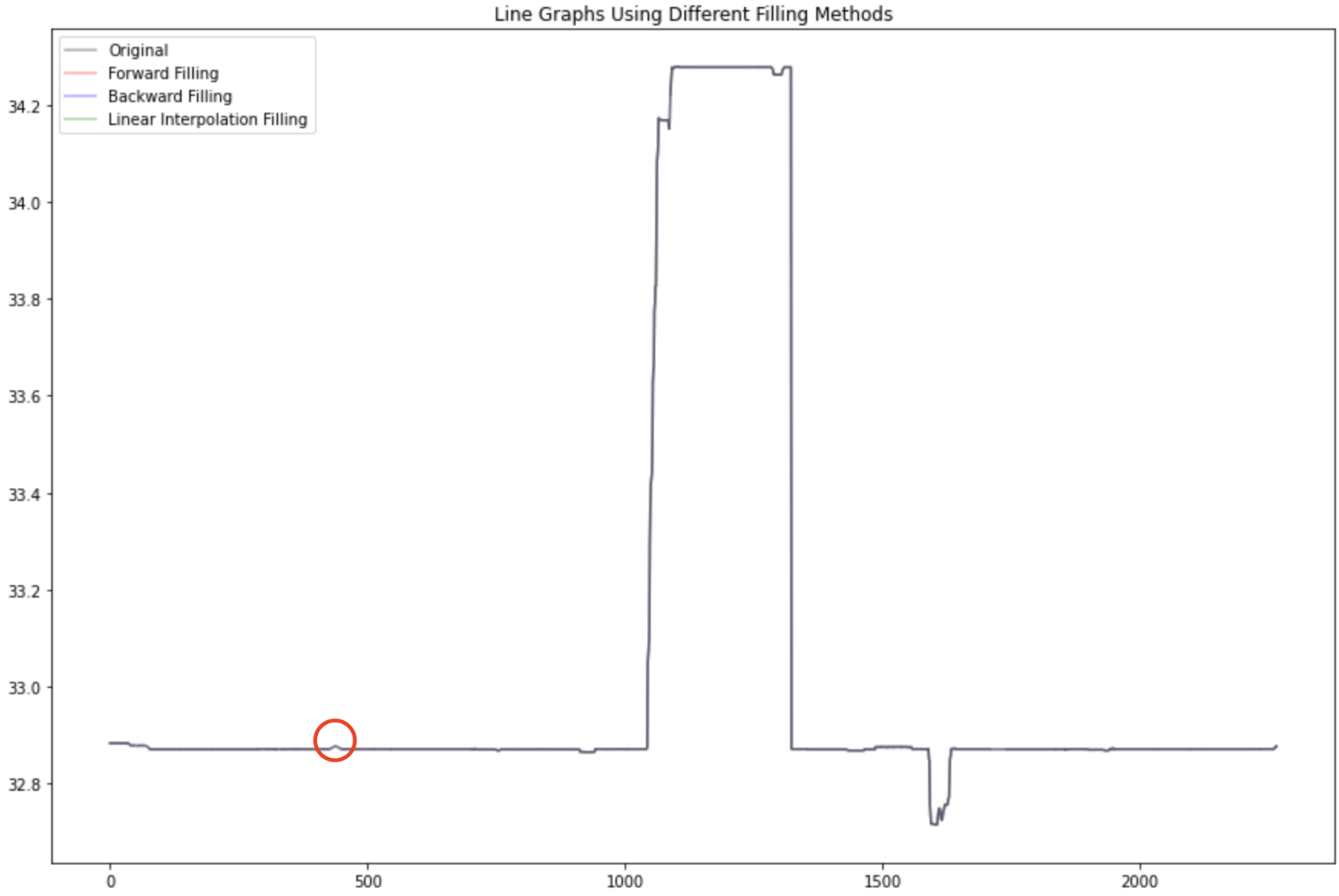
Writeup Answer to Problem A: Explanation of implementation

"people_2" is a dictionary that stores the people who behaves as 2) and number of minutes of location data have lost. "count" keep tracks of how many minutes of data is lost for each people. "b_count" keep tracks of how many minutes of data has battery level lower than 15%.

For each entry in the sensorData, I will look for all data points that have battery < 15% and raw_latitude is missing. For each data points, I will add one to the "count" and "b_count", if the data is not missing but battery < 15%, I will only add it to "b_count". Eventually, if more than 95% of the time that a user have < 15% battery level and data is missing at the same time, it means that it is a consistent behavior of the user and thus the data will be appended to people_2.

Writeup Answer to Problem B: Compare the 4 traces. What do you see? If you were to use this dataset for further analysis, which filling method will you choose?

The 4 lines generally overlaps each other. But if I were to choose a filling method, I would choose the linear interpolation. If you zoom in very closely, there is some difference around 500 where there is a small bump, and linear interpolation is the smoothest.



Thanks for grading. Have a nice day :)