Project 3 - Statistics and Visualization

Ji-Eun Park

Background

Urban Ministries of Durham (UMD) is a non-profit organization offering food, shelter and other services necessary for people in need of those aids. This project is designed to bring meaningful insights from the supply data gathered since around 1990's which is almost since the foundation of this organization (1983).

Project Goal

The aim of this project is to provide UMD with useful and invaluable information about the organization and the clients and provide analytic information and suggest insightful ideas for UMD.

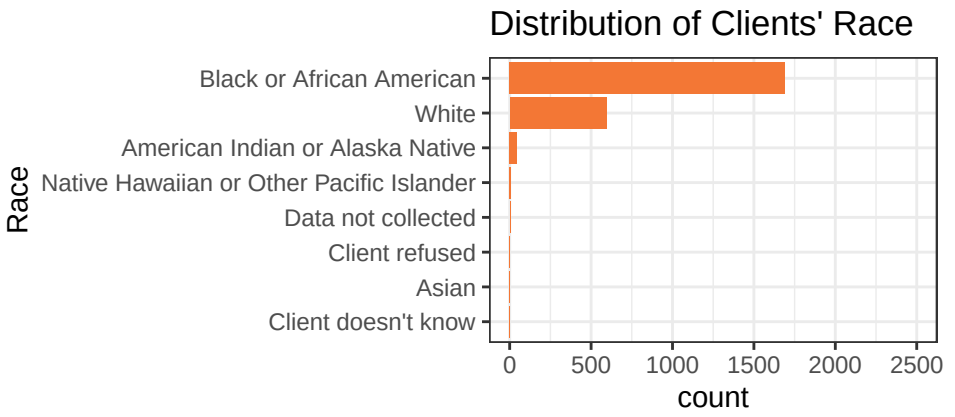
Data

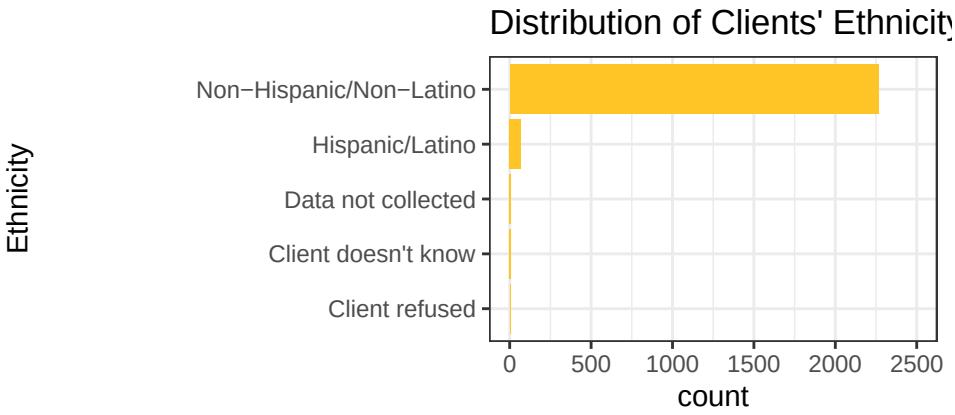
UMD has provided 16 datasets including such as client/provider information, disability at entry/exit, health insurance at entry/exit, income at entry/exit, noncash support received entry/exit, etc.

Analysis Results

1. Demographics of UMD clients

i) Race and Ethnicity

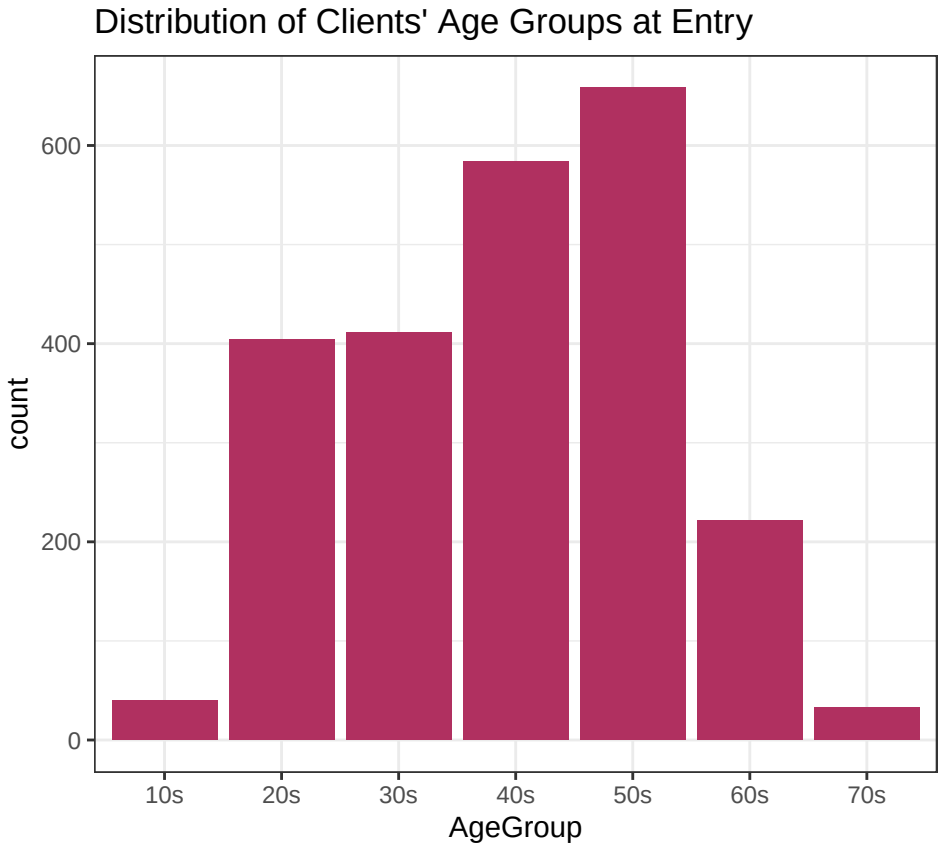




As you can see in the plots, majority of UMD's clients are Black or African American, or white. From ACS (American Community Survey) data collected in 2019, the racial composition of Durham is: White(47.96%), Black or African American(39.65%), Asian(5.18%), Native American (0.25%), Native Hawaiian or Pacific Islander (0.05%). Considering this fact, we can see that there is a very high ratio of Black/African American clients.

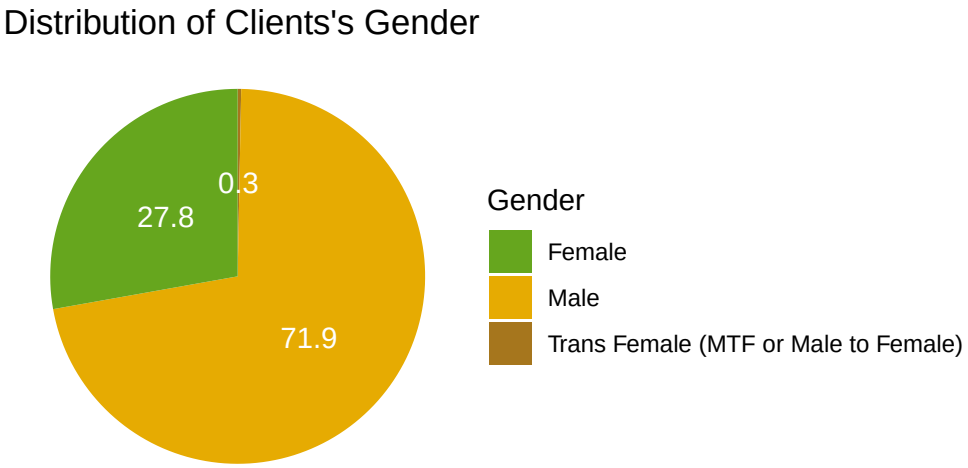
(reference: http://worldpopulationreview.com/us-cities/durham-population/)

ii) Age Groups



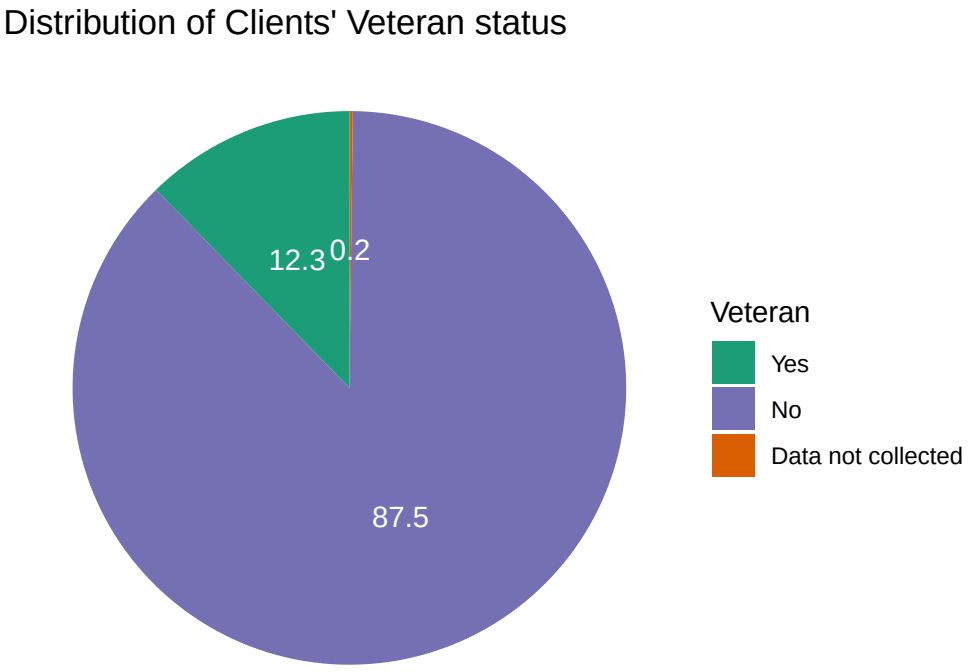
Results show that majority of the clients are in their 20s to 50s, where 40-50s have the highest proportion.

iii) Gender



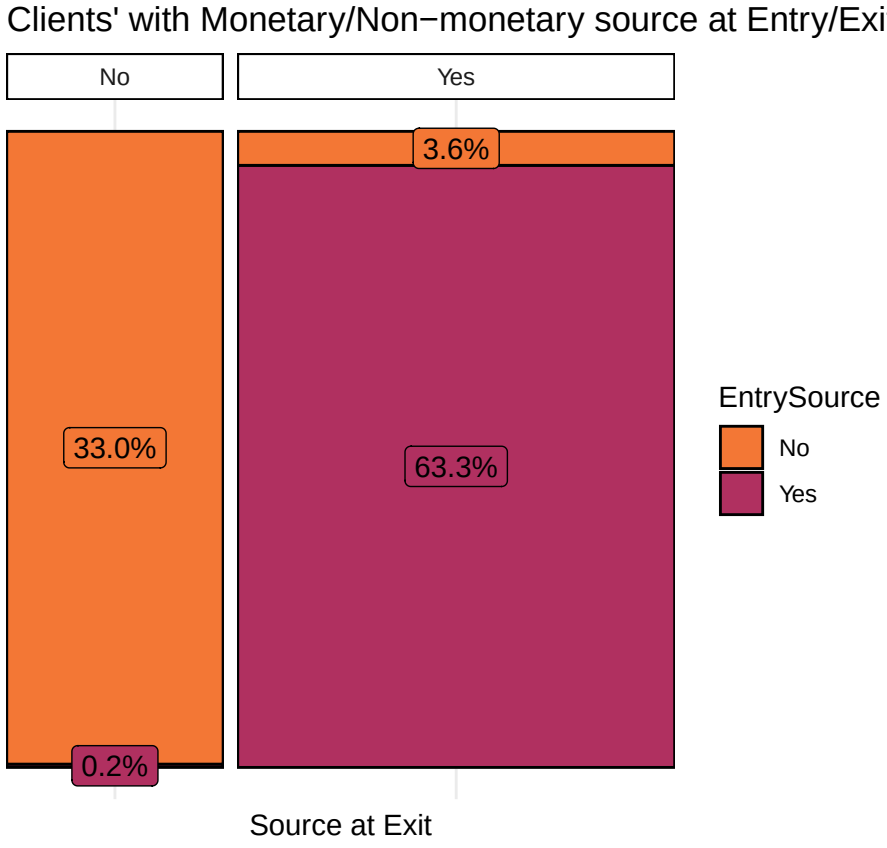
You can see that the clients at UMD have 72% male clients and 28% female clients. Very minor, but there also exists some MTF clients.

iv) Veteran Status



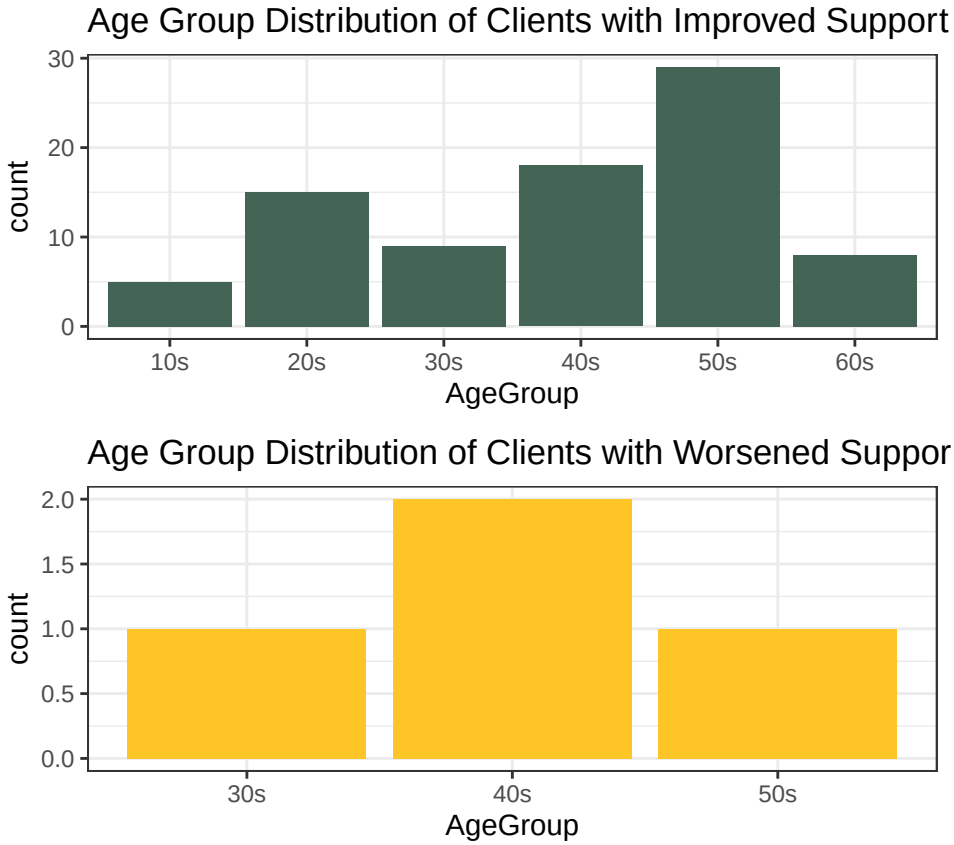
12.3% of the clients are veterans.

2. How did the income and non-cash support change from entry to exit? Were there any improvements?



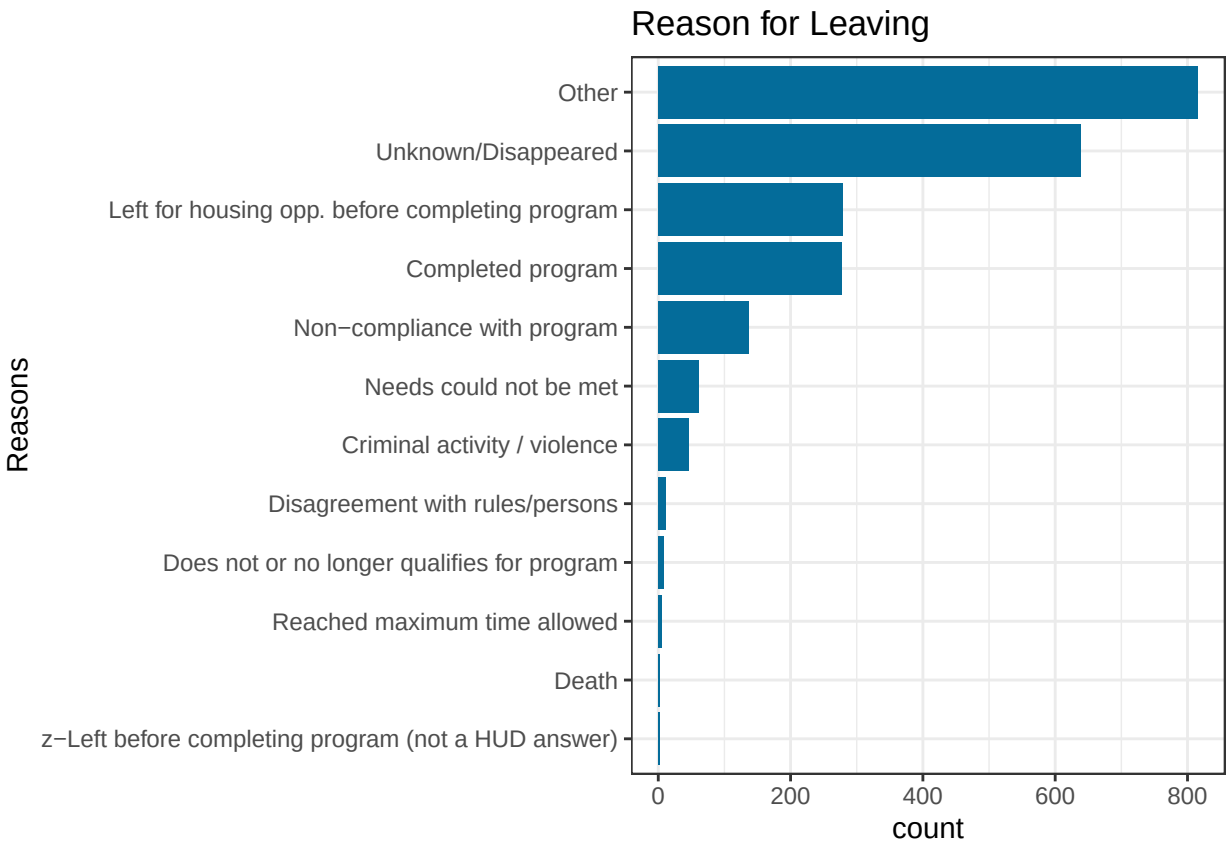
The above plot summarizes the ratio of clients who had monetary or non-monetary source of support at entry versus exit of the program. As we can observe, around 96% of the clients did not have change in the status, either did not have any source at entry and exit, or had a source at entry and exit.

Only 3.6% of the clients who did not have source at entry found a source at exit. And although it is minor, 0.2% of the clients lost their existing source at exit.



We can see that considering the age distribution of the clients, there is a higher chance of clients in their 20s to have better chance in improving their income status. We can guess that there are more demands for young workers.

3. What are the major reasons why the clients left?



We can see that the major reason for leaving is unknown or clients disappeared. This may imply that there are some underlying problems that clients are facing and had to leave without telling UMD. There are many possibilities but one possibility is that the status of the clients did not improve while staying at UMD. It is important to look deeper into client care.

Also, many clients left due to completion of program or housing opportunities, around 200 clients (out of 2352) had to leave due to non-compliance or due to needs not met. This is also something UMD might want to look into, in order to provide more service and goods to variety of people.

Let's look deeper on who left for each reason.

Table 1: Age distribution in clients reason for leaving is unknown/disappeared

AgeGroup	count	totalcount	percentage
10s	11	40	0.275
20s	123	404	0.304
30s	129	411	0.314
40s	151	584	0.259
50s	176	658	0.267
60s	43	222	0.194
70s	5	33	0.152

Table 2: Age distribution in unsatisfied clients

AgeGroup	count	totalcount	percentage
10s	1	40	0.025
20s	14	404	0.035
30s	13	411	0.032
40s	11	584	0.019
50s	12	658	0.018
60s	8	222	0.036
70s	2	33	0.061

Table 3: Age distribution in clients complete program

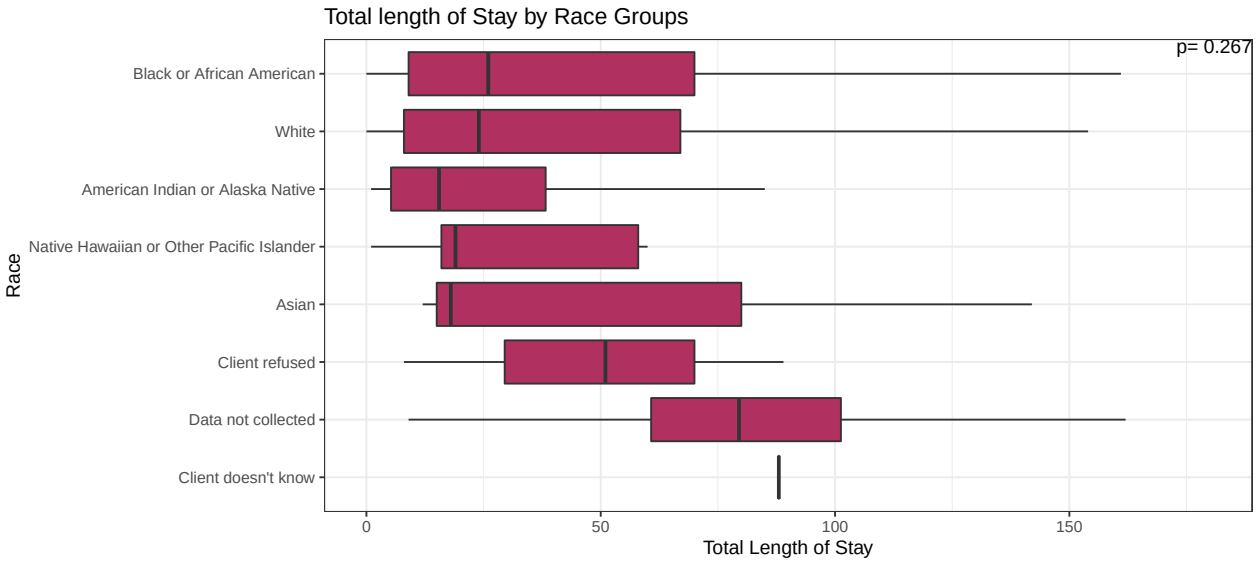
AgeGroup	count	totalcount	percentage
10s	3	40	0.075
20s	30	404	0.074
30s	28	411	0.068
40s	72	584	0.123
50s	93	658	0.141
60s	38	222	0.171
70s	13	33	0.394

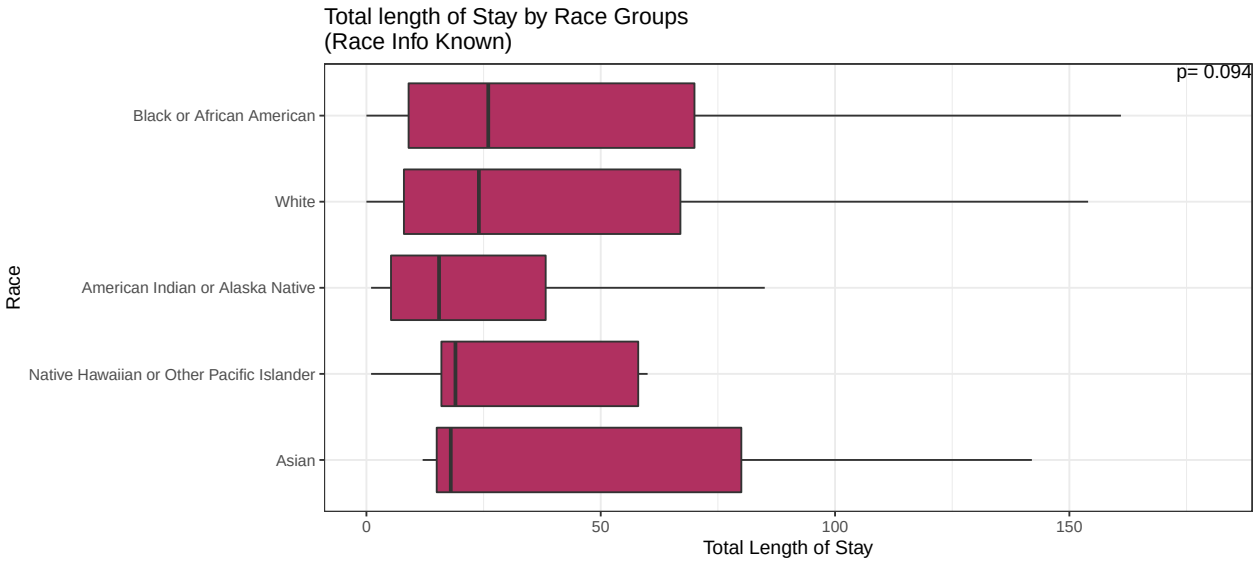
One noticeable characteristics of patients for the major reasons for leaving is that clients from 20s and 30s were more likely to disappear or leave without giving reason compared to 40s to 60s who were more likely to stay until completion of the program.

This may indicate that the UMD service is focused on elders but not youngs.

4. Relationship between length of stay and client demographics

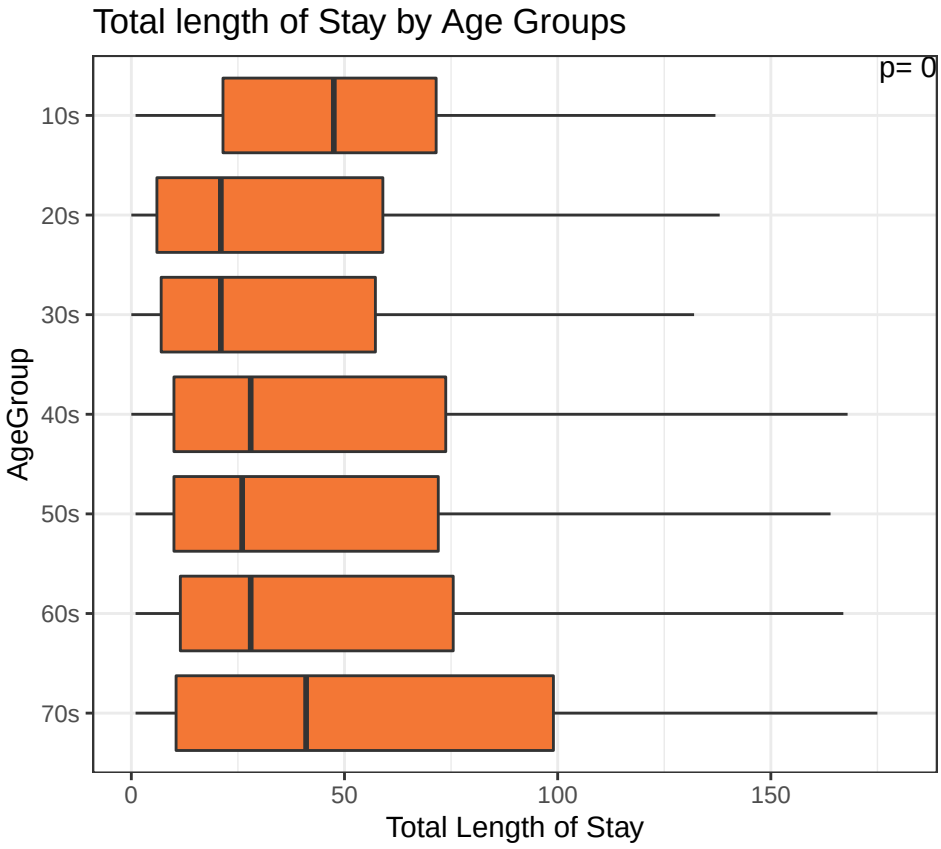
i) Race





Above are the boxplots which shows the distribution of the total length of stay clients in each Race group. The first one is based on the entire dataset, while the second one is only based on data with Race group known. I used the Kruskal Wallis test, which is a non-parametric multigroup comparison test to see if there's significant difference in the total length of stay among race groups. However, both results does not show significant difference (p=0.267, 0.094) under the significance level 0.05.

ii) Age Group

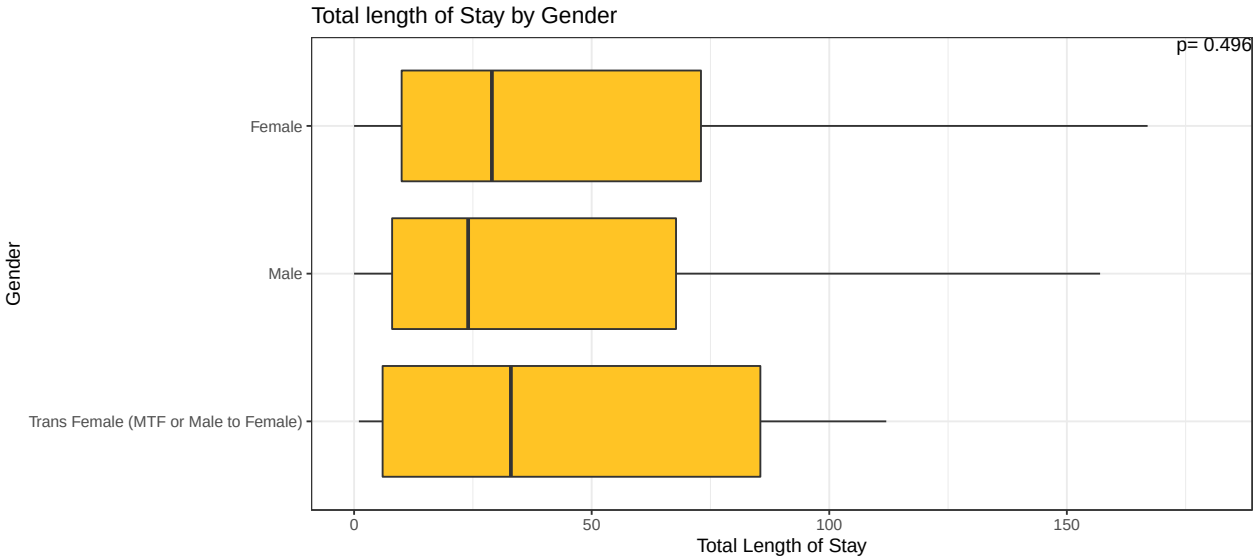


Above are the boxplots which shows the distribution of the total length of stay clients in each Age group.

Our Kruskal-Wallis test showed significant difference in total length of stay between some Age groups. Based on pairwise wilcoxon test we know that the following groups have significant difference: 40s vs 20s,50s vs 20s,60s vs 20s,70s vs 20s,40s vs 30s,50s vs 30s,60s vs 30s,70s vs 30s.

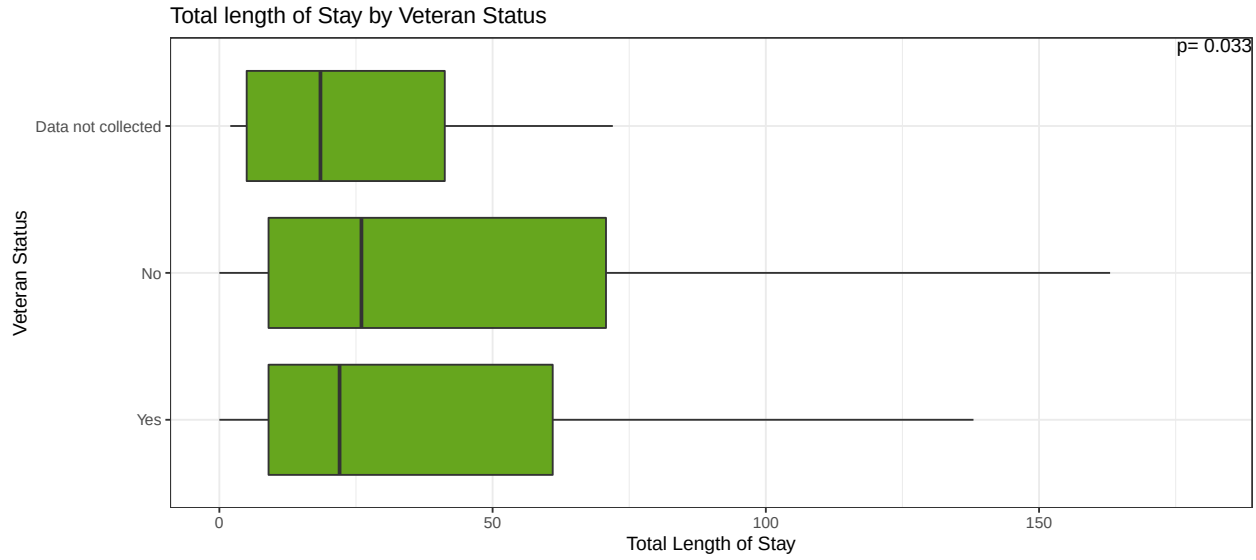
Hence, we can see that clients in their 20s, 30s tend to have a significantly lower total length of stay.

iii) Gender



Same process was done on Gender groups, but there were no significant difference between Gender groups.

iv) Veteran status

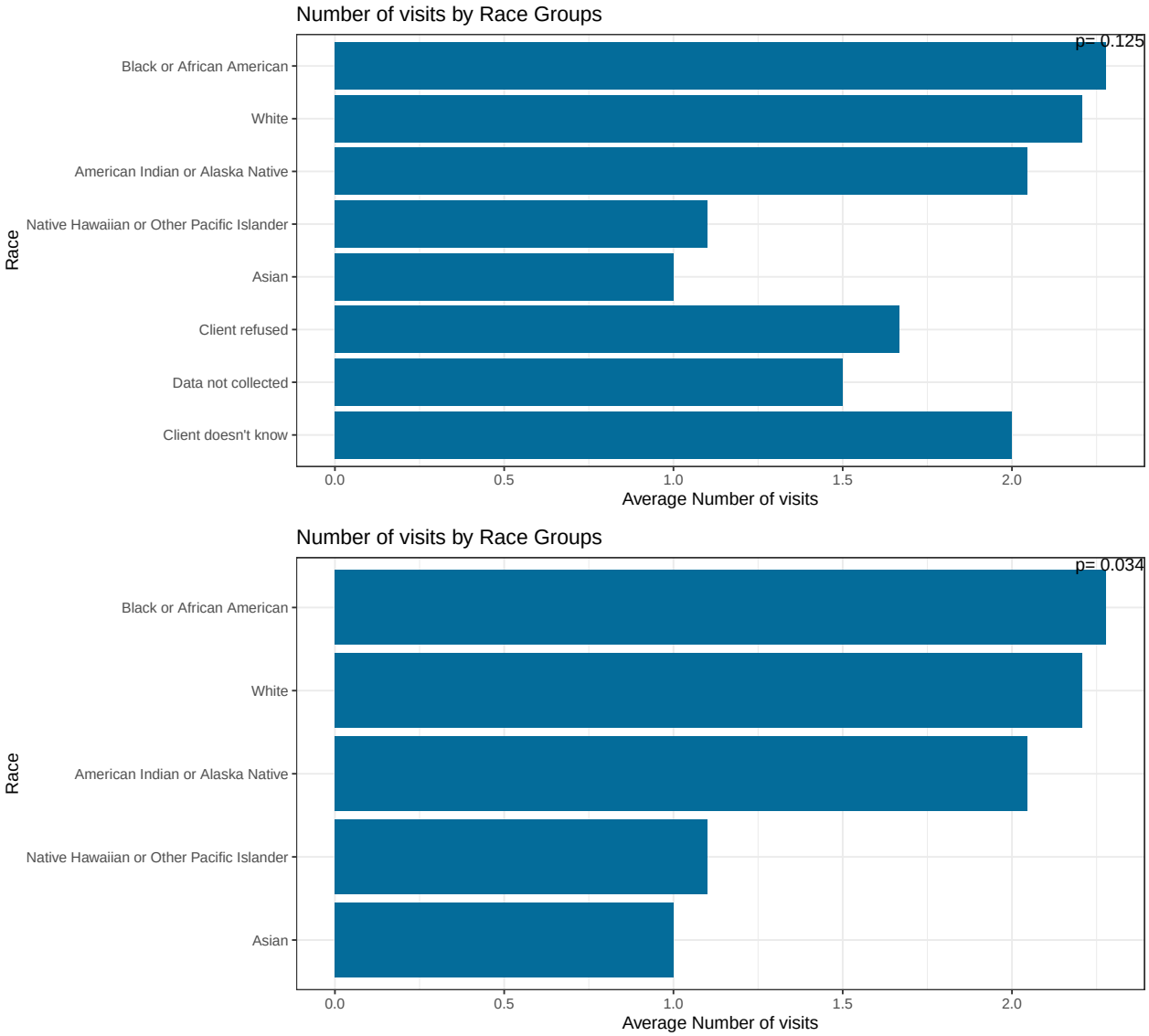


For Veteran status, the difference was checked for both 1)whole data, and 2)data without clients with unkown veteran status. Interestingly both Kruskal Wallis test results show there is significant difference in total length of stay between veteran status. Since we only have two veteran status (veteran vs not veteran), it is implied from the boxplot that veterans have lower total length of stay thn non veterans.

5. Relationship between client recurrence and demographics

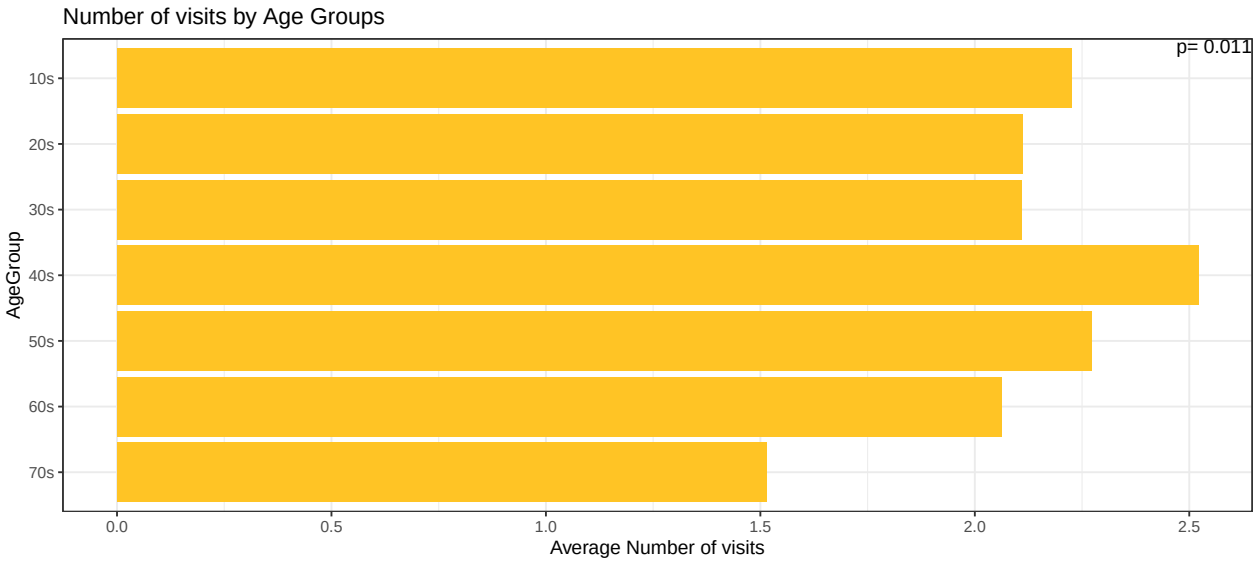
Same process as question 4. Although we are using non-parametric methods to test significance different groups, since the values are very small, barplots will be used instead of boxplots.

i) Race



Same process was done with number of visits instead of total length of stay. Results show that when data with unkown race were removed, there are significant differences between some Race groups. Based on wilcoxon rank sum test, Native Hawaiian/Other Pacific Islander vs Black/African American,Native Hawaiian/Other Pacific Islander vs White have significant difference. Hence we can see Native Hawaiian or Other Pacific Islanders have significantly less recurrence compared to White or Black/African American clients.

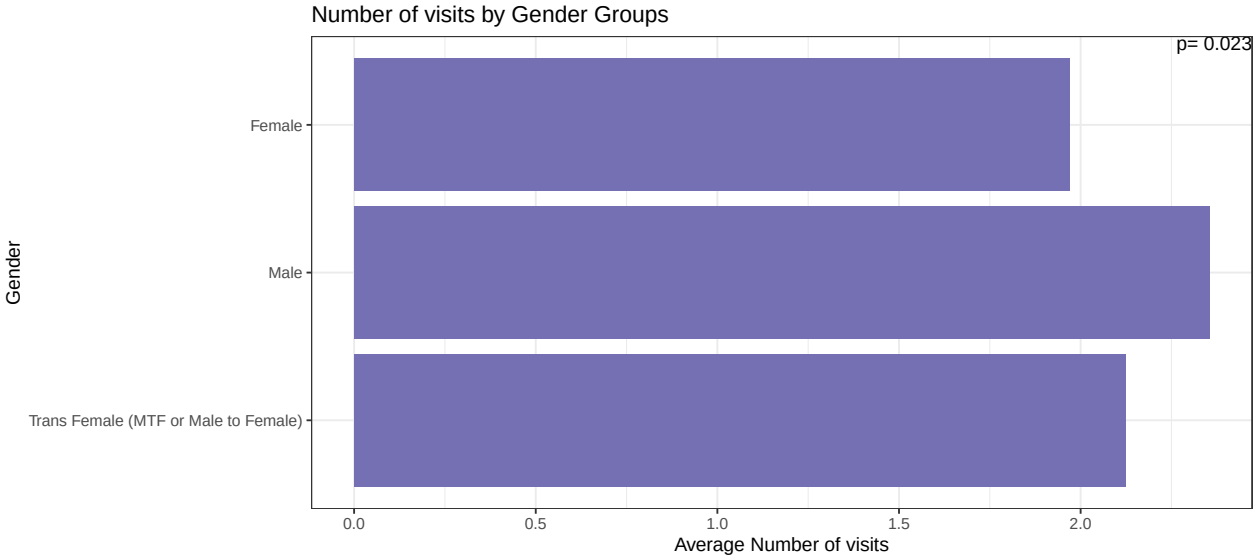
ii) Age Group



Kruskal-Wallis test showed significant difference in number of visits between some Age groups. Based on pairwise wilcoxon test we know that the following groups have significant difference: 70s vs 10s,60s vs 40s,70s vs 40s,60s vs 50s,70s vs 50s.

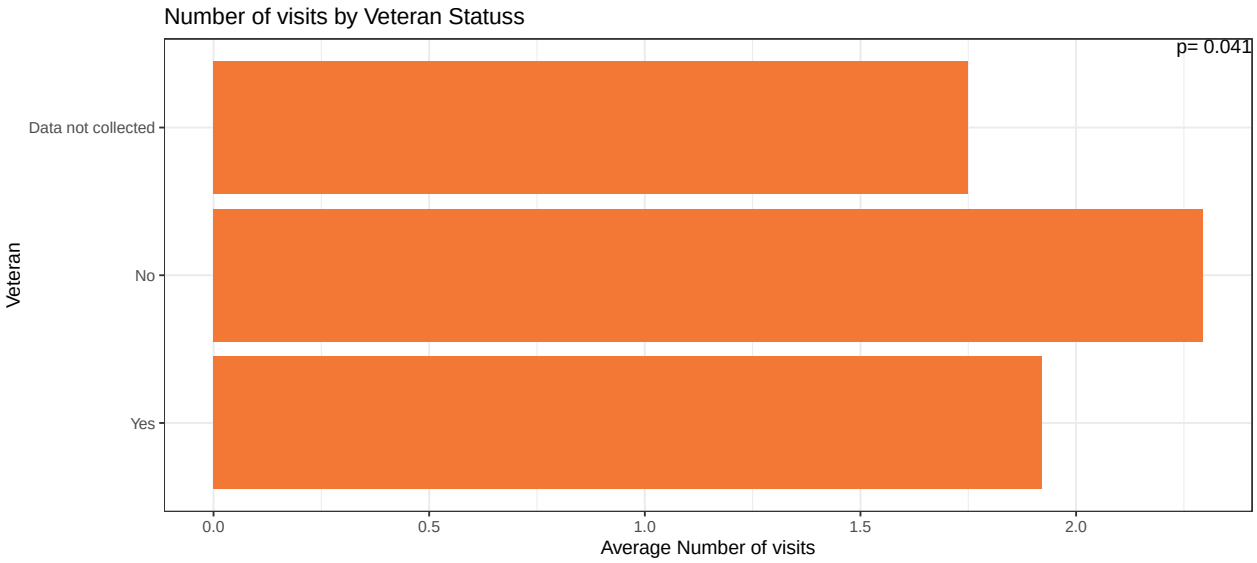
Hence, we can see that clients in their 60s and 70s tend to have a significantly lower recurrence.

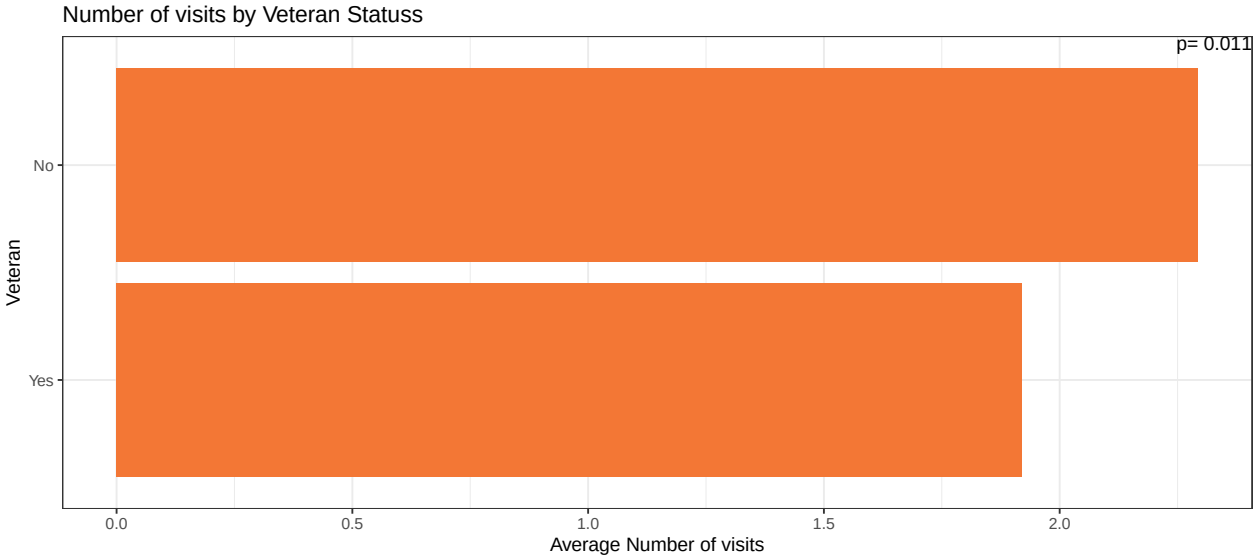
iii) Gender



Same process was done on Gender groups, and results show that there are significant differences between some Gender groups. Based on wilcoxon rank sum test, Male vs Female have significant difference. To be specific, male clients have longer stay than female clients.

iv) Veteran status





For Veteran status, the difference was checked for both 1)whole data, and 2)data without clients with unkown veteran status. Since we only have two veteran status (veteran vs not veteran), it is implied from the boxplot that veterans have lower recurrence than non veterans.

Conclusion