UMD Data Analysis: Final Project

Chuwen Liu

Introduction

The dataset shows some variables that represent support for the people in need by Urban Ministries of Durham, whose aim is to help ending the homeless in the neighborhood.

In this project, we are provided different kind of datasets which contain the different features' statistics summary and demographic data at two different status: entry and exit. we could explore some potential phenomenon based on the summary of entry data and exit data.

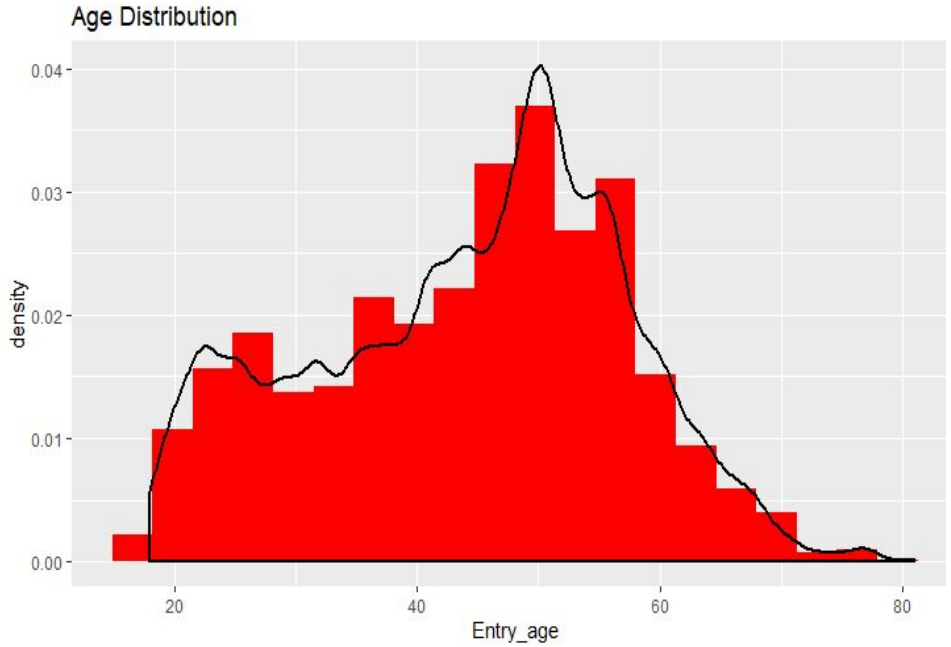
Goals

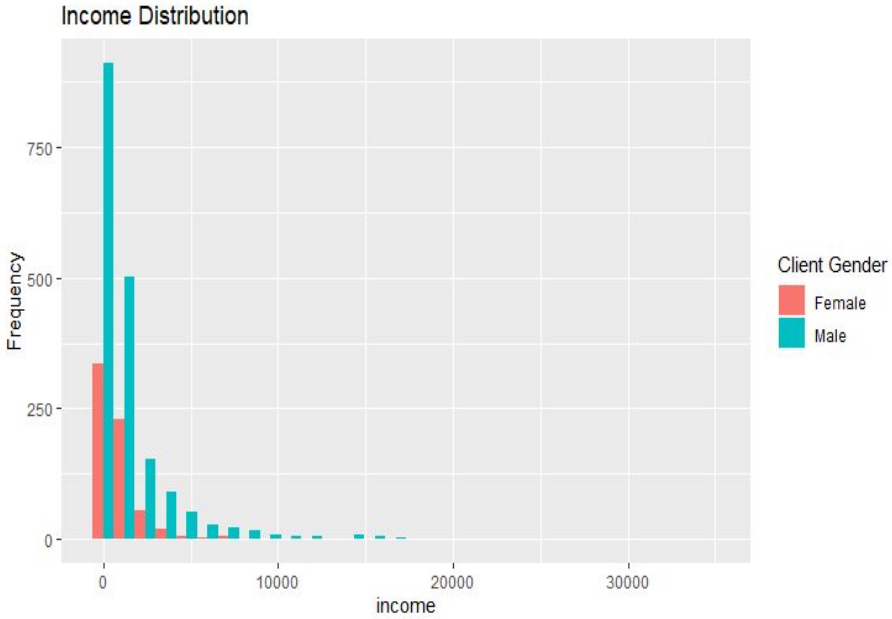
Will the Duration of Staying be affected by some features of clients, such as gender, race, ….? Try to build a linear model to see if it works well.

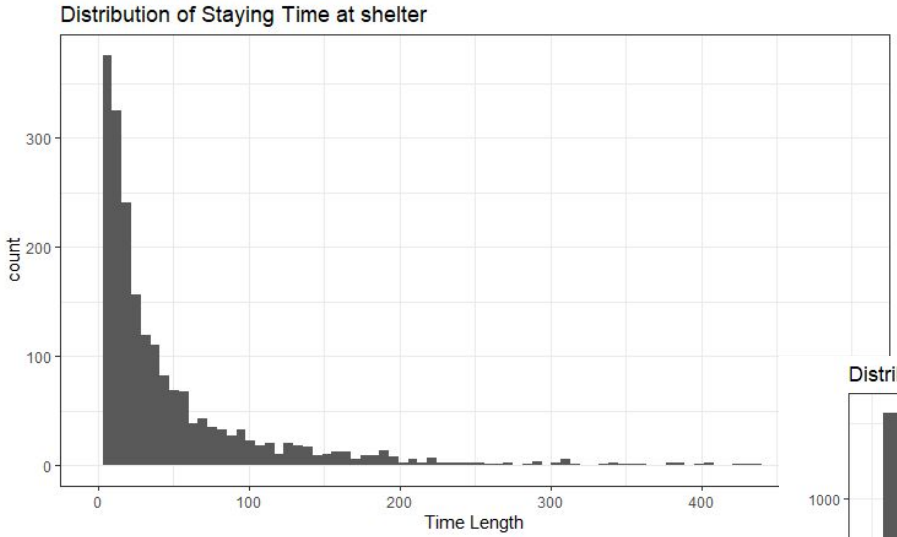
Explore the description and correlation among different data.

Will season affect the entry? Will some variables have influence on the clients' entry month?

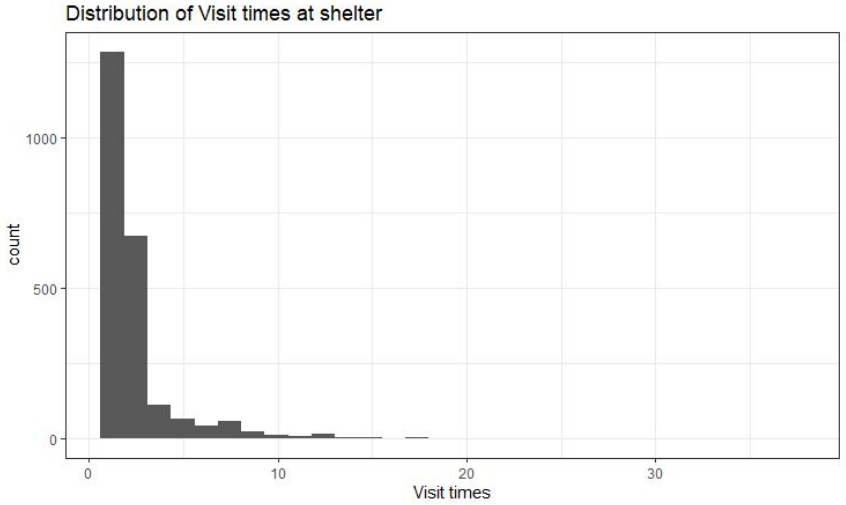
Distribution of some variables



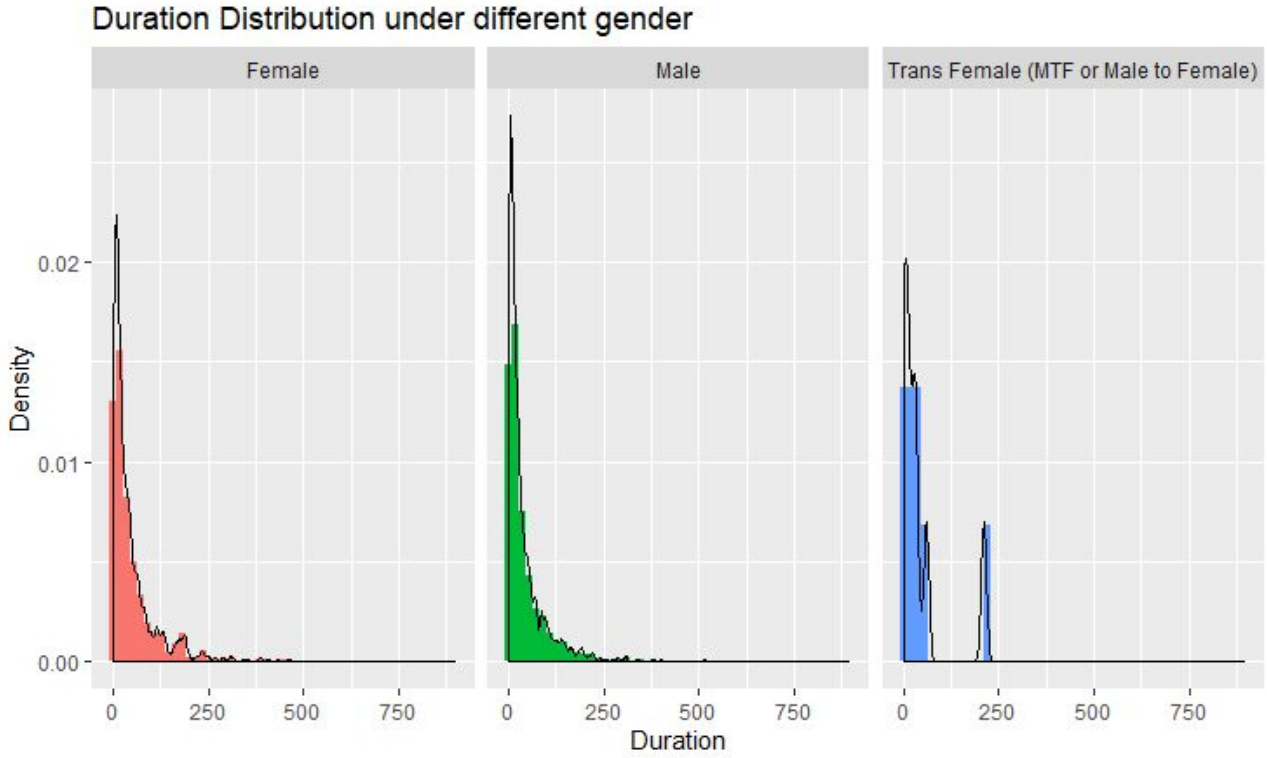




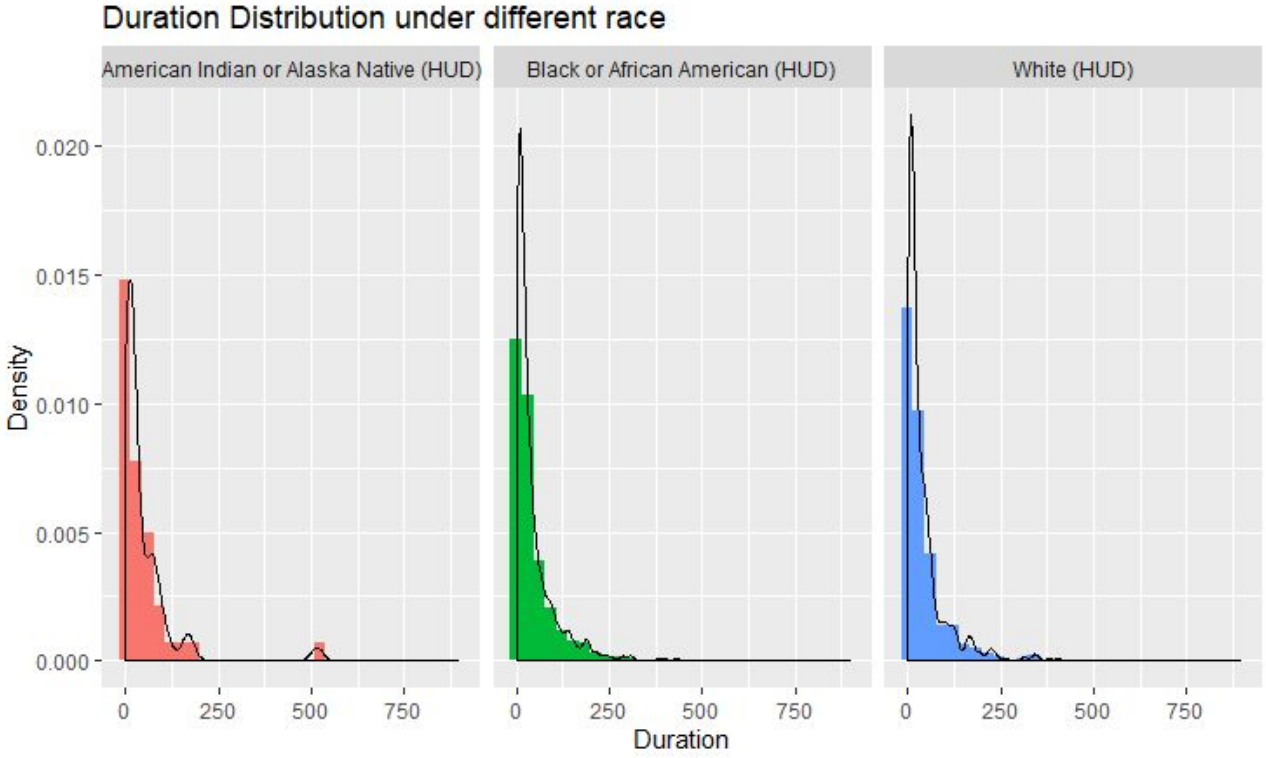
The two distribution are both highly distributed around 0, and the most times who visit the shelter is 37. The longest staying time is 894 days.



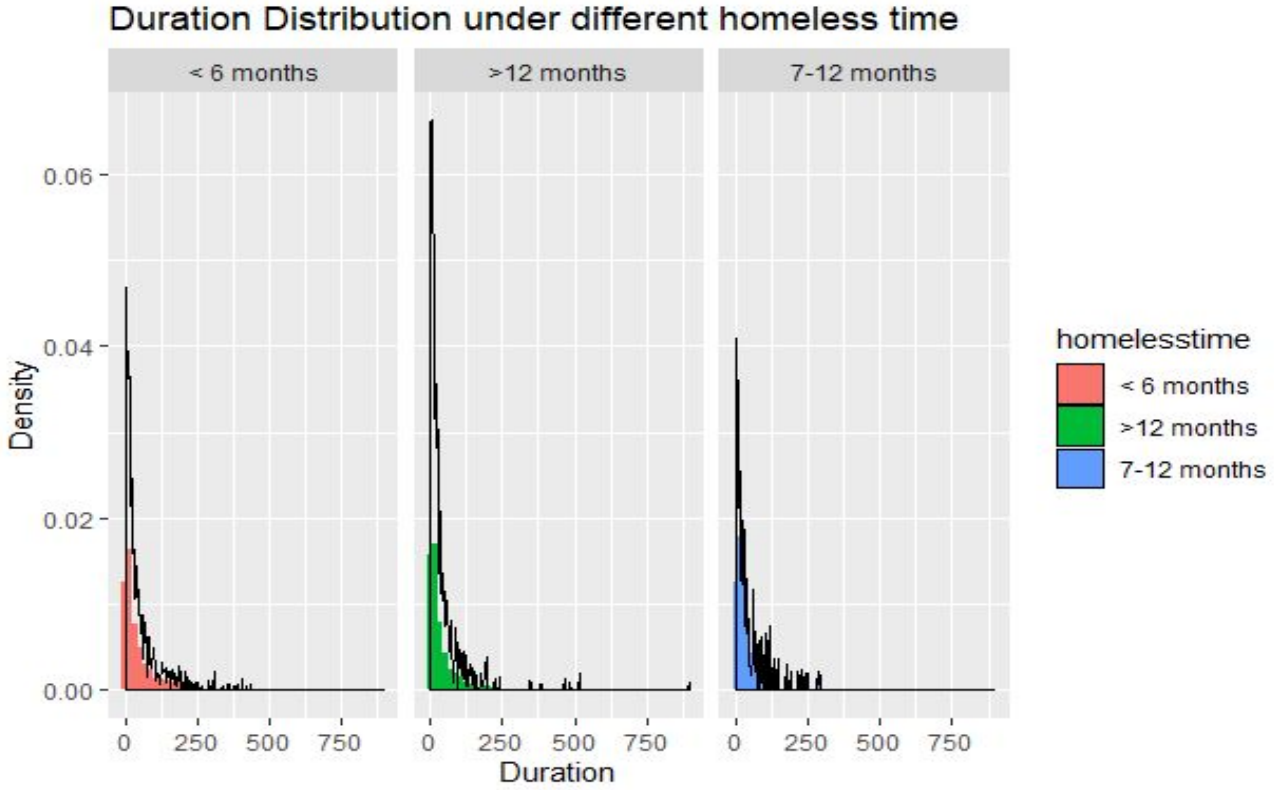
Correlation with Duration



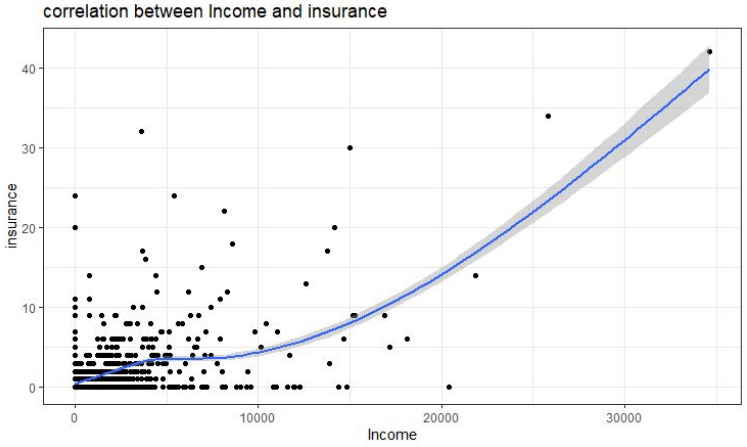
Correlation with Duration



Correlation with Duration

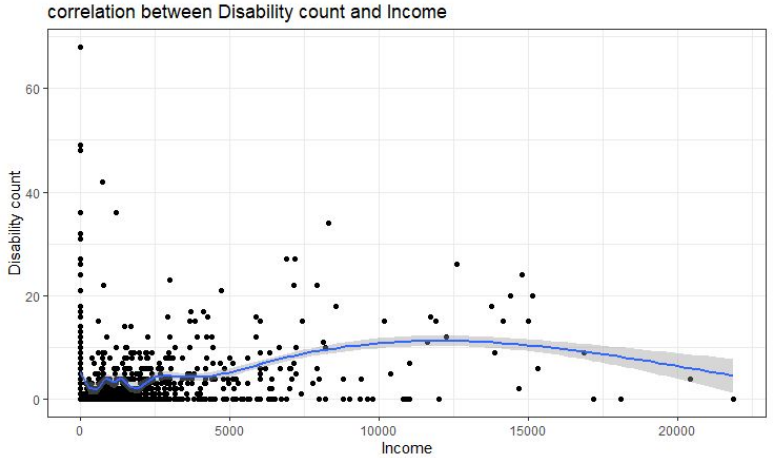


Some interesting Correlation

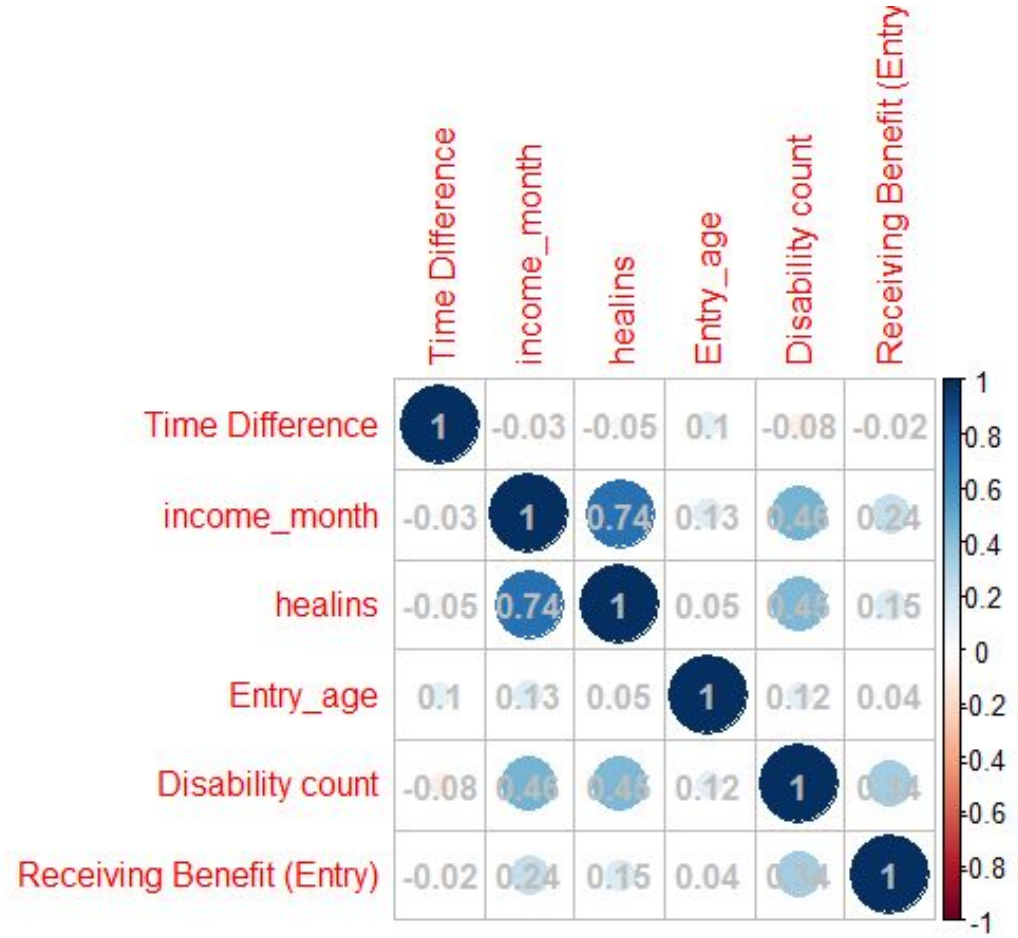


For who got higher monthly income, usually they will have more kinds of health insurance.

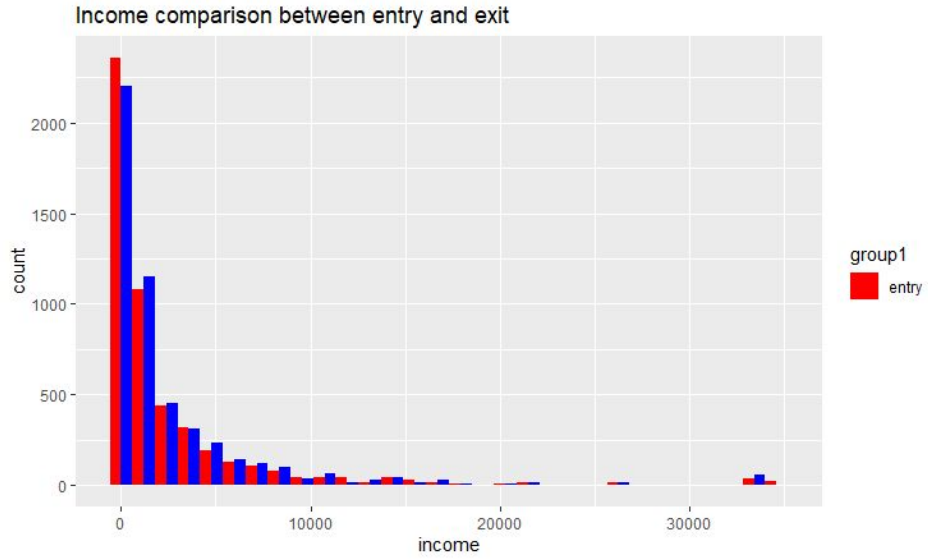
For who got higher monthly income, usually they will have less disability; but for whose monthly income is small, since the cardinality is large, we found more points gathering near 0

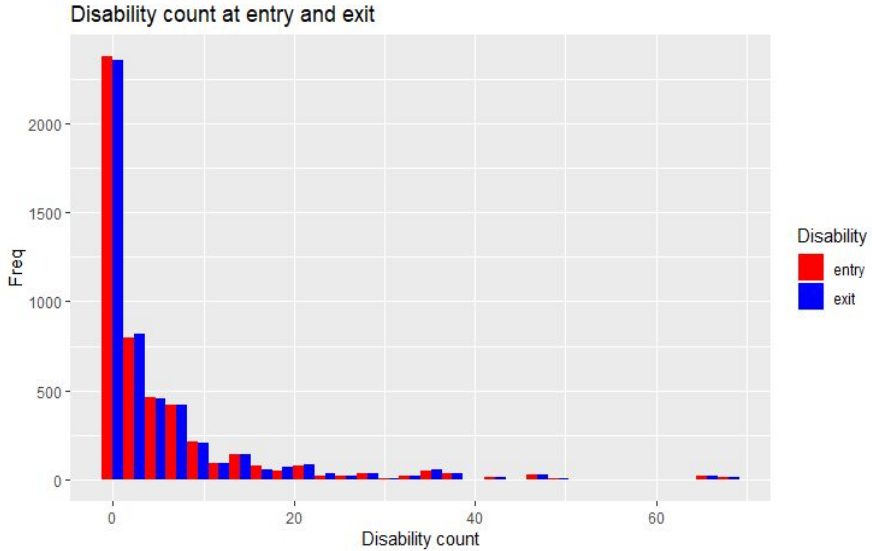


From the Correlation plot, we know that the income, disability and insurance count have moderate level linear correlation.

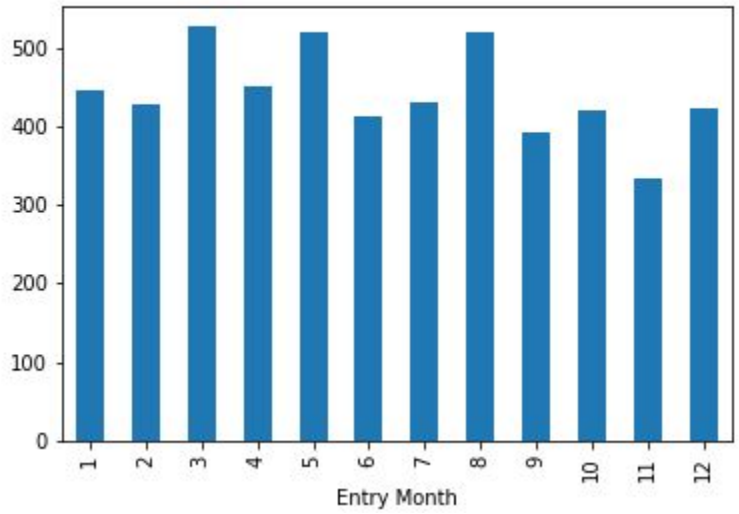


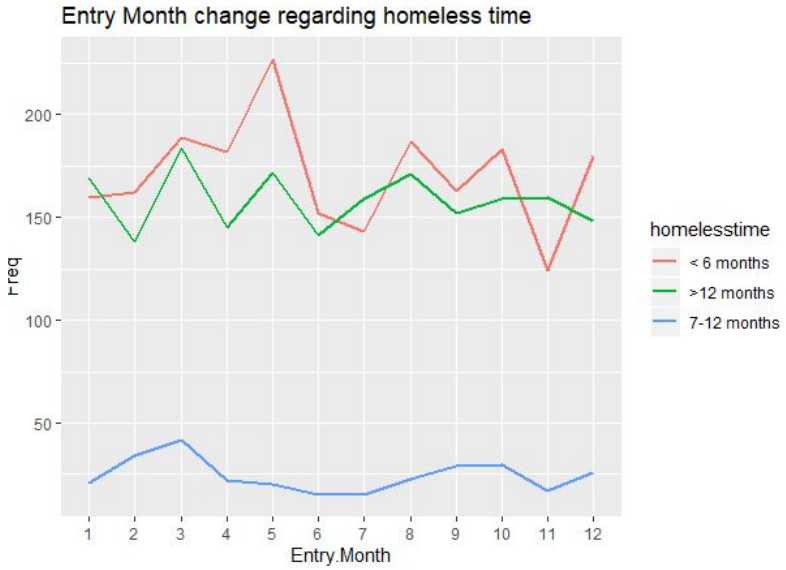
Comparisons between entry and exit





Any monthly trend?





We find around November is almost a time that less people enter the shelter.

Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	18.0497254	10.1765406	1.774	0.076196	.
Entry_age	0.5896283	0.0804013	7.334	2.7e-13	***
`Client Gender`Male	-5.7619078	2.3535638	-2.448	0.014402	*
raceBlack or African American (HUD)	-2.7672987	7.8457708	-0.353	0.724323	
raceWhite (HUD)	-4.9020090	8.0219895	-0.611	0.541187	
`Client Veteran Status`Yes (HUD)	-9.4855934	3.3628351	-2.821	0.004815	**
`Client Ethnicity`Non-Hispanic/Non-Latino (HUD)	12.0942522	5.6656433	2.135	0.032849	*
income_month	0.0002789	0.0003197	0.872	0.383093	
healins	-0.2051081	0.2882905	-0.711	0.476839	
`Disability count`	-0.5101565	0.1158894	-4.402	1.1e-05	***
`Receiving Benefit (Entry)`	0.1521957	0.3649462	0.417	0.676675	
homelesstime>12 months	-8.3048505	2.1536475	-3.856	0.000117	***
homelesstime7-12 months	-7.3223762	4.0636846	-1.802	0.071636	.

Duration of staying is our response variable, and we include all the continuous and categorical variables at first.

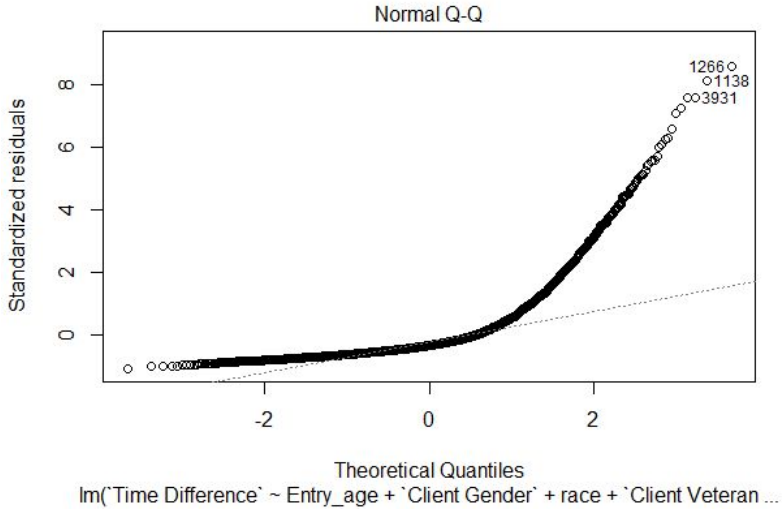
Linear Model

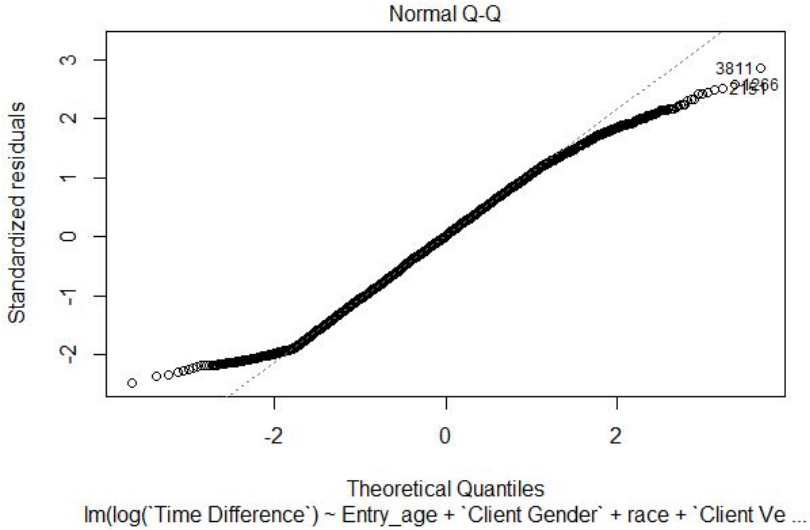
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.34965	6.70248	1.693	0.090466	.
Entry_age	0.62717	0.08076	7.766	1.02e-14	***
`Client Gender`Male	-5.04849	2.35181	-2.147	0.031882	*
`Client Veteran Status`Yes (HUD)	-10.76816	3.37672	-3.189	0.001439	**
`Client Ethnicity`Non-Hispanic/Non-Latino (HUD)	13.94352	5.55068	2.512	0.012042	*
`Disability count`	-0.52480	0.10360	-5.066	4.25e-07	***
homelesstime>12 months	-7.90170	2.17177	-3.638	0.000278	***
homelesstime7-12 months	-6.06640	4.09788	-1.480	0.138851	

When we drop some non-significant variables, we found gender, veteran status, ethnicity, homeless time and disability count are very important variables to be included in the model.

Try log normal Distribution





Conclusions

Most people are around 50-60 years old, black man. For the duration time, we found that in fact age, gender, race, ethnicity have little effect on it. But seems who are homeless for more year is more possible to stay for shorter time.

In linear regression model, We found that those who are veteran, young people usually stay less time at shelter; those who are homeless more than one year will be less possible to stay longer than those whose homeless time is less than 6 months. That is interesting since it is against our usual idea.

50% people didn't complete the program when they select to leave. I think our next step analysis is to figure out what leads to their departure, is it correlated to any other feature?