Main Report

Team M

September 26, 2017

Introduction

Marriage is such a big topic in human life and always a very strong predictor of people's life satisfaction and happiness level. However, not all marriages are eternal happy endings and a lot of them nowadays end in divorce. If we can understand the factors that influence the quality of marriage, we may be able to discover more preventive measures to help more people maintain a lifelong happy marriage.

To answer this question, our group examined the 2011 UK Census data and limited our scope to the marital status of a sample of British people.

through descriptive and inferential statistics

To clarify, we define divorce rate here as (total number of divorce or separated)/(total number of marriage). Thus, we exclude all people who have not yet married, and the people who end their marriages out of other reasons such as the other partner's death.

Description

The dataset we examined is the 5% sample data of the 2011 UK Census. It contains 2,848,155 observations and 121 variables. Each observation represents a UK citizen Since the dataset is very large and there are too many variables, we decided to exclusively focus on the relationship between marital status (marstat) and other variables. In brief, we would like to examine the factors that may influence people's marriage and lead to higher divorce rate. Therefore, we selected a total of 17 variables that may be highly correlated with marital status.

We can attach a table of all variables we examined here:

Methods

a really brief description of the R packages we used. We can load all our libraries in this page:

```
library(dplyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(stargazer)
library(readr)
library(RCurl)
library(foreign)


# Load data
url <- "https://raw.githubusercontent.com/dlouhasha/TheGreatWork/master/data/filtered_dataset.csv"
filtered_data <- getURL(url)
filtered_data <- read.csv(textConnection(filtered_data))
```

```
fd <- filtered_data
fd <- as.data.frame(fd)
```

Descriptive Statistics and Correlation Tests

Simple descriptive statistis including bar charts, pie charts, density graph... etc. We can show a few correlation tests, chi-squre tests and graphs

Social and Economic Status & Marriage

Religion and Culture & Marriage

Age & Marriage

Region/Geography & Marriage

An article on divorce rate by region piqued our interests (http://www.dailymail.co.uk/news/article-3201497/Wish-weren-t-ten-divorce-hot-spots-Britain-sea-Blackpool-worst-place-live-want-happy-marriage.html). The article lists the top 10 divorce hot spots and they are all coastal cities. In this paper we are keen to explore whether our data attests to the findings of that article.

The data shows a significant peak in inner London which shows that the percentage of people in bad marriage (either divorced or separated) is much higher than elsewhere. We would like to conduct a hypothesis testing to see if region and marital status are independent. The test we adopt here is chi-squared test at 5% significance level because we are exploring the correlations between two categorical variables.

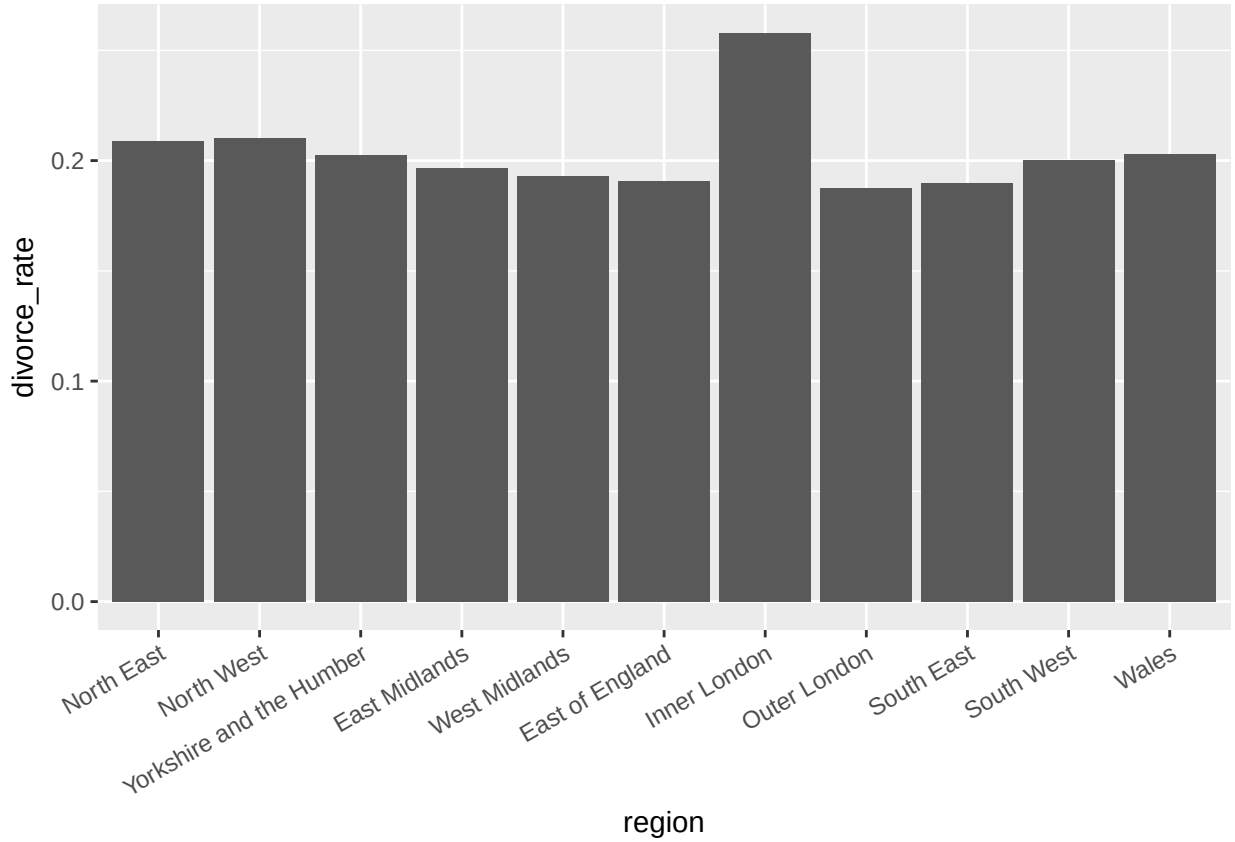
The p-value is less than 5% and it means that the we can reject the null hypothesis that region and marital status are independent. From the residuals of chi-squared test, we see that married couples are greatly underrepresented in Inner London, but divorced and separated couples are over represented. Interestingly Inner London has a very high proportion of same-sex couples.

We use the data to plot a mosaic plot which essentially represents the residual table graphically.

```
# select the data relevant for region
region1 <- fd %>% select(binary_marstat, age, aggdtwpew11g, region, marstat, transport, wpzhome, Highes

#explore marital status with region
region_count<- region1 %>% select(marstat, region) %>% group_by(region, marstat)%>% summarise(count = n

region_count%>% mutate(divorce_rate = (Divorced + Separated)/(Divorced + Married + Separated))%>% ggplo
```




```
# chi-squared test
region_test <- region_count[,-1]
rownames(region_test) <- as.data.frame(region_count)[,1]
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
chi_region <- chisq.test(as.data.frame(region_test))
chi_region
```

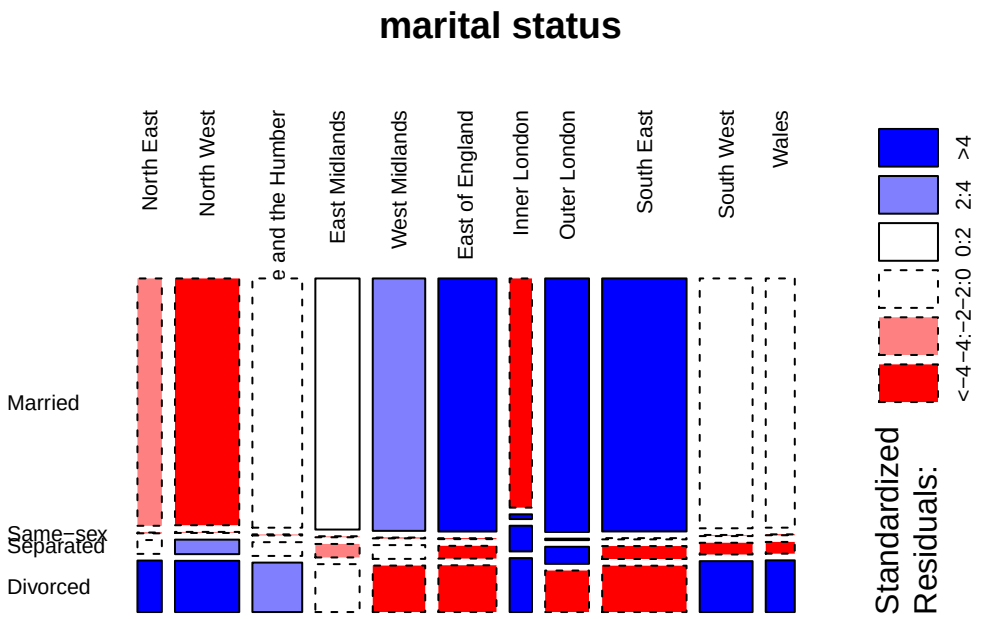
```
##
##  Pearson's Chi-squared test
##
## data:  as.data.frame(region_test)
## X-squared = 5517.9, df = 30, p-value < 2.2e-16
```

```
round(chi_region$residuals, 3)
```

```
##                          Married Same-sex Separated Divorced
## North East                -2.267   -4.707    -0.801    6.343
## North West                -4.514   -3.927     2.637    9.472
## Yorkshire and the Humber  -0.845   -2.241    -1.604    3.148
## East Midlands              1.439   -3.507    -2.833   -1.179
## West Midlands              3.242   -6.573    -1.494   -5.508
## East of England            4.322   -5.741    -7.515   -4.849
## Inner London             -17.610   42.124    41.564   10.845
## Outer London               4.289    3.883    14.929  -18.429
## South East                 5.093   -0.129    -7.530   -7.498
## South West                -0.118   -2.089   -12.891    7.557
```

```
## Wales                    -0.689    -4.321    -9.541    7.408
```

```
#Mosaicplot
mosaicplot(region_test, shade = TRUE, las=2,
           main = "marital status")
```



We further perform t-test to compare the divorce rate of Inner London and all the other regions. Inner London has a divorce rate of approximately 25%, higher than the average of 20% across all regions. The difference is statistically significant.

```
#mean_London <- fdata$binary_marstat[fdata$region == "Inner London"]
outside_london <- fd$binary_marstat[fd$region != "Inner London"]
t.test(fd$binary_marstat, fd$binary_marstat[fd$region == "Inner London"])
```

```
##
##  Welch Two Sample t-test
##
## data:  fd$binary_marstat and fd$binary_marstat[fd$region == "Inner London"]
## t = -29.485, df = 60679, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05857028 -0.05126873
## sample estimates:
## mean of x mean of y
## 0.1990274 0.2539469
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "North East"])
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "North East"]
## t = -7.0448, df = 67505, p-value = 1.875e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.015065598 -0.008507167
## sample estimates:
## mean of x mean of y
## 0.1965882 0.2083746
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "North West"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "North West"]
## t = -12.204, df = 204130, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01514510 -0.01095363
## sample estimates:
## mean of x mean of y
## 0.1965882 0.2096375
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "Yorkshire and the Humber"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "Yorkshire and the Humber"]
## t = -4.1936, df = 151500, p-value = 2.747e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007284591 -0.002644151
## sample estimates:
## mean of x mean of y
## 0.1965882 0.2015525
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "East Midlands"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "East Midlands"]
## t = 0.59654, df = 132400, p-value = 0.5508
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001687283  0.003163742
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1958499
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "West Midlands"])
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  outside_london and fd$binary_marstat[fd$region == "West Midlands"]
## t = 3.8376, df = 162210, p-value = 0.0001243
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.002135541 0.006594027
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1922234
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "East of England"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "East of England"]
## t = 6.1722, df = 183610, p-value = 6.749e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.004557332 0.008798462
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1899103
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "Outer London"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "Outer London"]
## t = 8.0127, df = 131890, p-value = 1.131e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.007395683 0.012185378
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1867976
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "South East"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "South East"]
## t = 8.0449, df = 291970, p-value = 8.667e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.005586210 0.009184904
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1892026
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "South West"])
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  outside_london and fd$binary_marstat[fd$region == "South West"]
## t = -2.6886, df = 162800, p-value = 0.007177
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0053381059 -0.0008366652
## sample estimates:
## mean of x mean of y
## 0.1965882 0.1996756
```

```
t.test(outside_london, fd$binary_marstat[fd$region == "Wales"])
```

```
##
##  Welch Two Sample t-test
##
## data:  outside_london and fd$binary_marstat[fd$region == "Wales"]
## t = -3.7516, df = 82014, p-value = 0.0001758
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.008699317 -0.002728801
## sample estimates:
## mean of x mean of y
## 0.1965882 0.2023022
```

The article aforementioned states that the factors contributing to the high divorce rate in coastal regions are due to the high deprivation and possibly sheer boredoem in winter. Our data shows otherwise. London is anything but a deprived and boring city. What could have contributed to the different in findings?

Secondly, we suspect there are other more salient influencing factors and this secton of our paper will explore that.

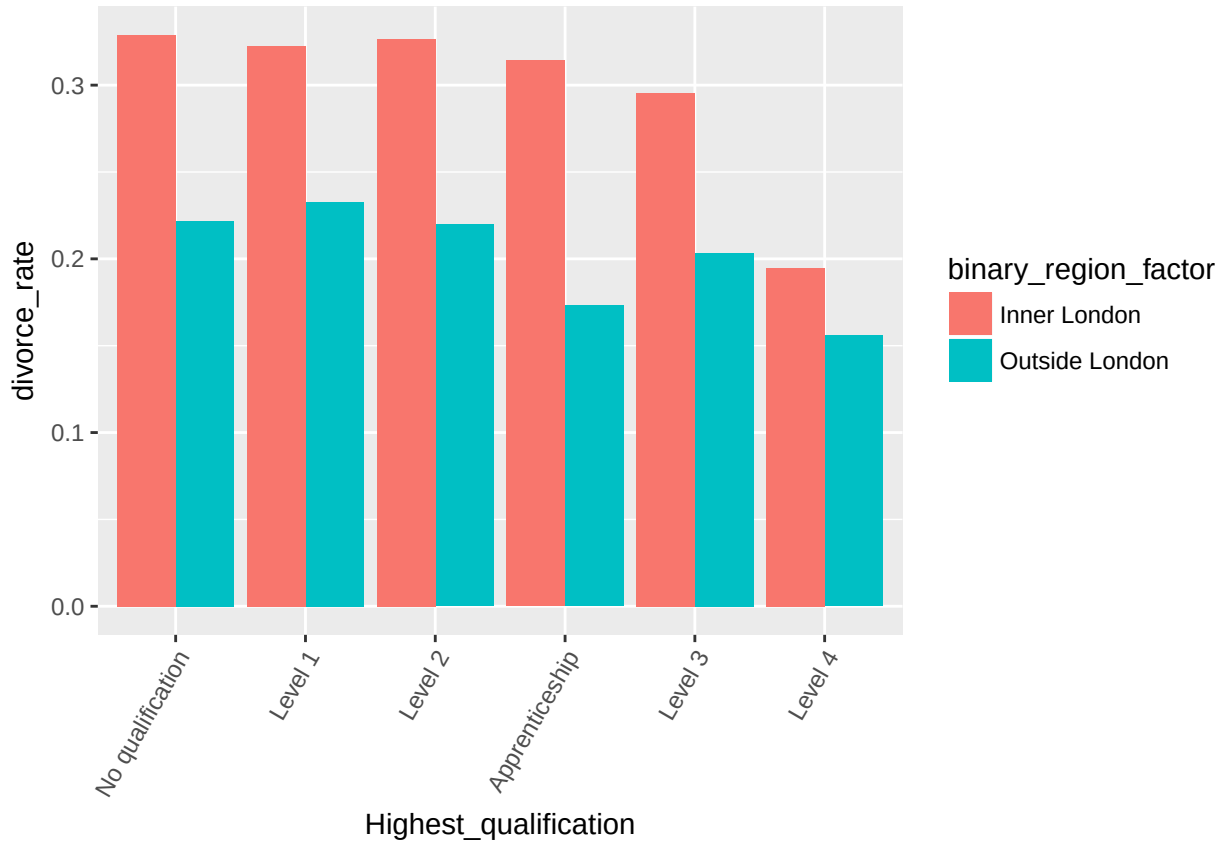
qualification and region

We start by looking at the highest qualifications achieved by residents in each region. We do so by plotting the proportion of people with each qualifications upon the total population of that region. London has the highest proportion of people with a level 4 qualifications. Does that contribute to the higher divorce rate in London? It is contradictory to what we observe when looking at the correlations between qualifications and divorce rate. We found that the divorce rate among highly educated people tend to be lower. Let's now look at among people with the same qualifications, what's the proportion of bad marriage.

When qualification is the same, does region affect marital status?

We look at the divorce rate of each qualification across regions. Overall the divorce rate is slightly lower among people with high qualifications, which is consistent with our findings. But we still observe a high divorce rate in inner London which can't be explained by qualifications.

```
# london and non-london
region1 %>% select(binary_marstat, region, Highest_qualification, binary_region_factor) %>% filter(!is.
```

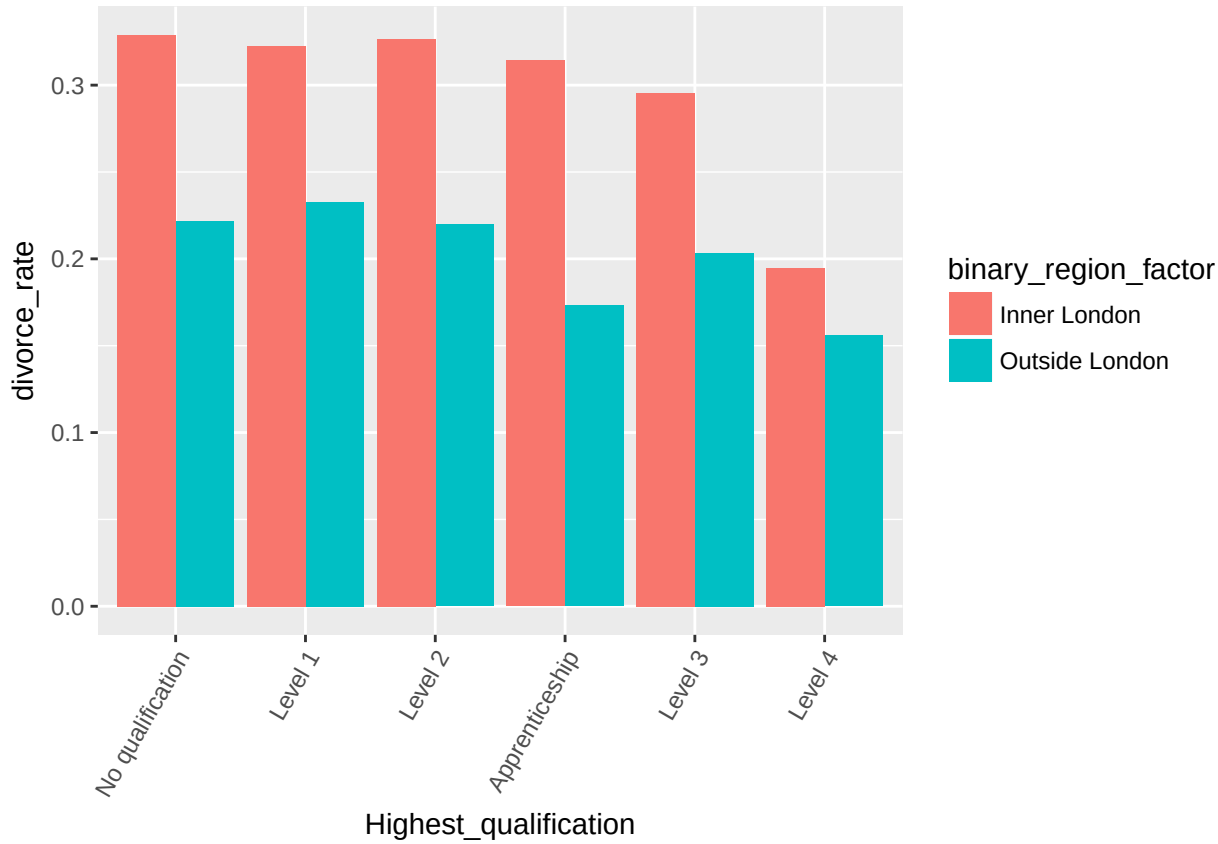


We moved on to explore hours worked and region, but do not find any apparent relationships as the number of hours worked across all regions seem to be even.

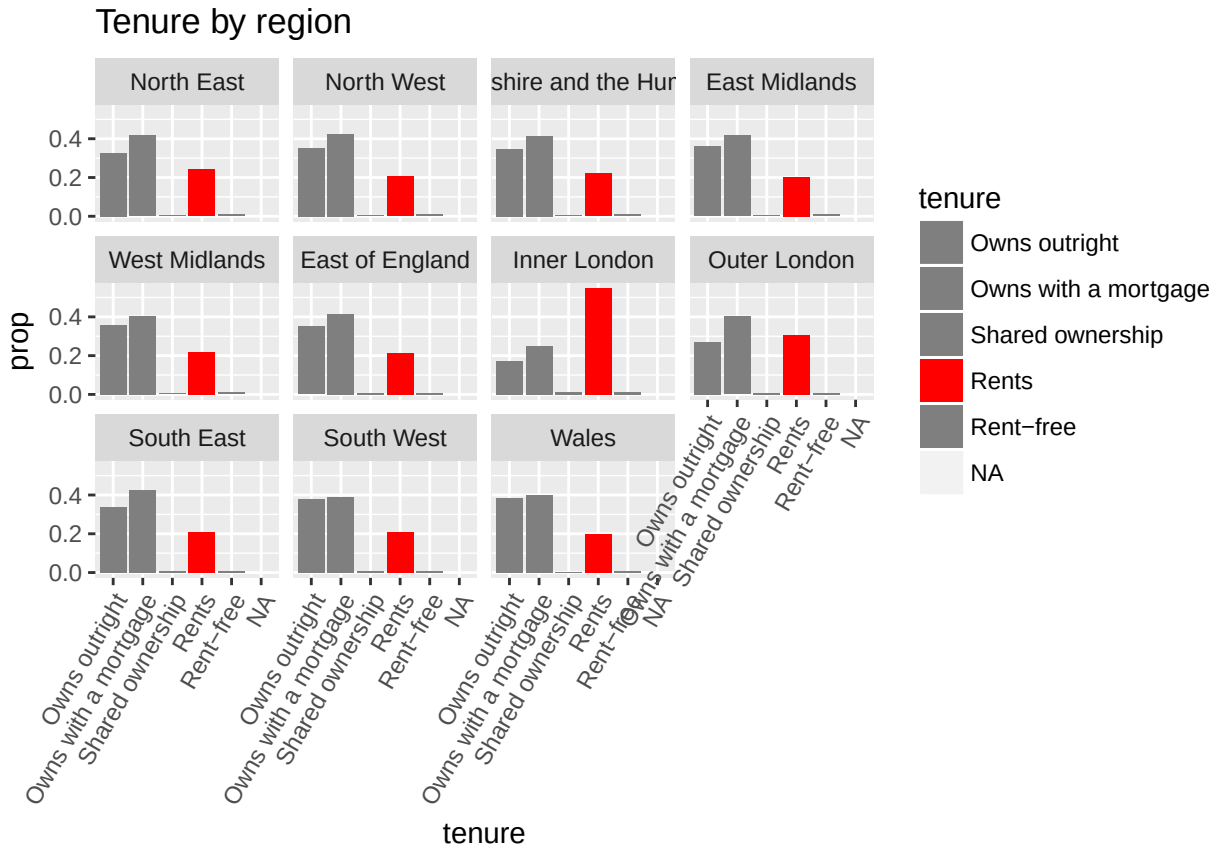
Type of dweling and region and marital status

We continue to explore other variables in our list and see if we could find any pattern. When we compare the types of dwelling across regions, inner London has the highest percentage of people staying in rented property.

```
region1 %>% select(binary_marstat, region, Highest_qualification, binary_region_factor) %>% filter(!is.
```




```
fd %>% dplyr::select(region, marstat, tenure) %>% filter(marstat != 'Single' & marstat != 'Widowed')%>%
  ungroup() %>% group_by(region)%>%mutate(prop = count /sum(count)) %>% ggplot(aes(x = tenure, y = prop
```



We next look at the marital status of people staying in different types of dwelling. Not surprisingly separated and divorced people stay in rented property, whereas people marriage often opt to own a house either by outright ownership or mortgage.

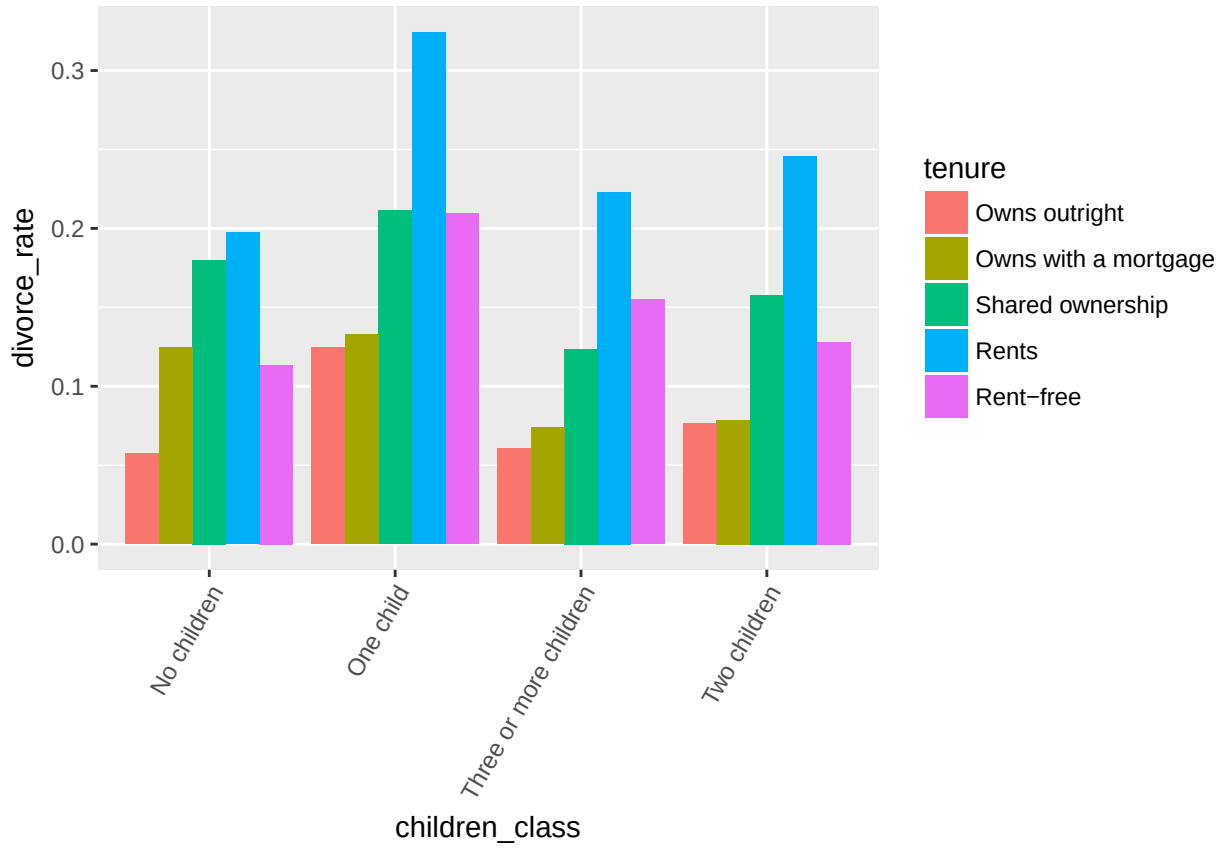
Next we want to look at the marital status, types of dwelling and region. We find the proportion of people in each types of dwelling upon total population in each marital status within each region. We observe that in London a majority of people staying in rented properties regardless of their marital status, but even more so among separated and divorced people. On the other hand, most people in marriage opt to stay in owned property in other regions, even among the divorced and separated couples.

However we can't imply that there is a causal relationship.

Dependent childre and tenure

```
family_compo <- fd%>% select(dpcfamuk11, tenure, binary_marstat, binary_region_factor) %>% mutate(no_of_

family_compo %>% select(binary_marstat, tenure, binary_region_factor, children_class) %>% filter(!is.na
```



Transport and marital status

Other than types of dwelling, we want to further analyse how people commute to work. As reports have shown that the longer it takes for people to get to work, the lower the happiness level.

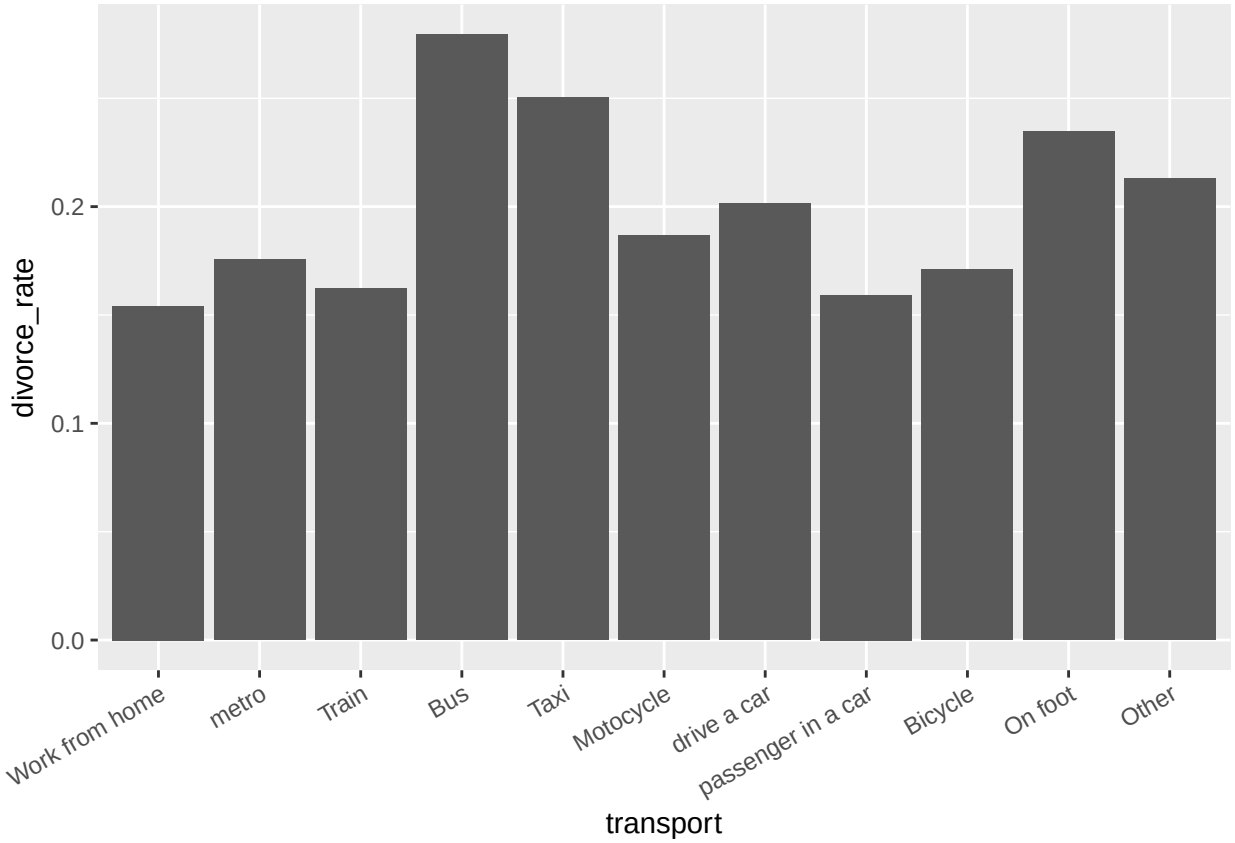
As we do not have data on the duration of daily commute from home to work, but we have means of transport and distance of home from work to simulate that.

We first look at the percentage of bad marriage among total married and previously married people across different means of commute. We find that bus shows a higher divorce rate.

Divorce rate by transport

```
transport_count<- region1 %>% select(marstat, transport) %>% filter(!is.na(transport))%>% group_by(trans

transport_count%>% mutate(divorce_rate = (Divorced + Separated)/(Divorced + Married + Separated))%>% gg
```



To further test our hypothesis that means of commute influence marital status, we run a chi-squared test. We get a p-value below 5% so we can conclude that it does have impact to some extent.

Chi-squared test

```
transport_test <- transport_count[,-1]
rownames <- as.data.frame(transport_count)[,1]
rownames(transport_test) <- as.data.frame(transport_count)[,1]
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
chi_transport <- chisq.test(as.data.frame(transport_test))
chi_transport
```

```
##
##  Pearson's Chi-squared test
##
## data:  as.data.frame(transport_test)
## X-squared = 5400.6, df = 30, p-value < 2.2e-16
```

```
#round(chi_region$residuals, 3)
```

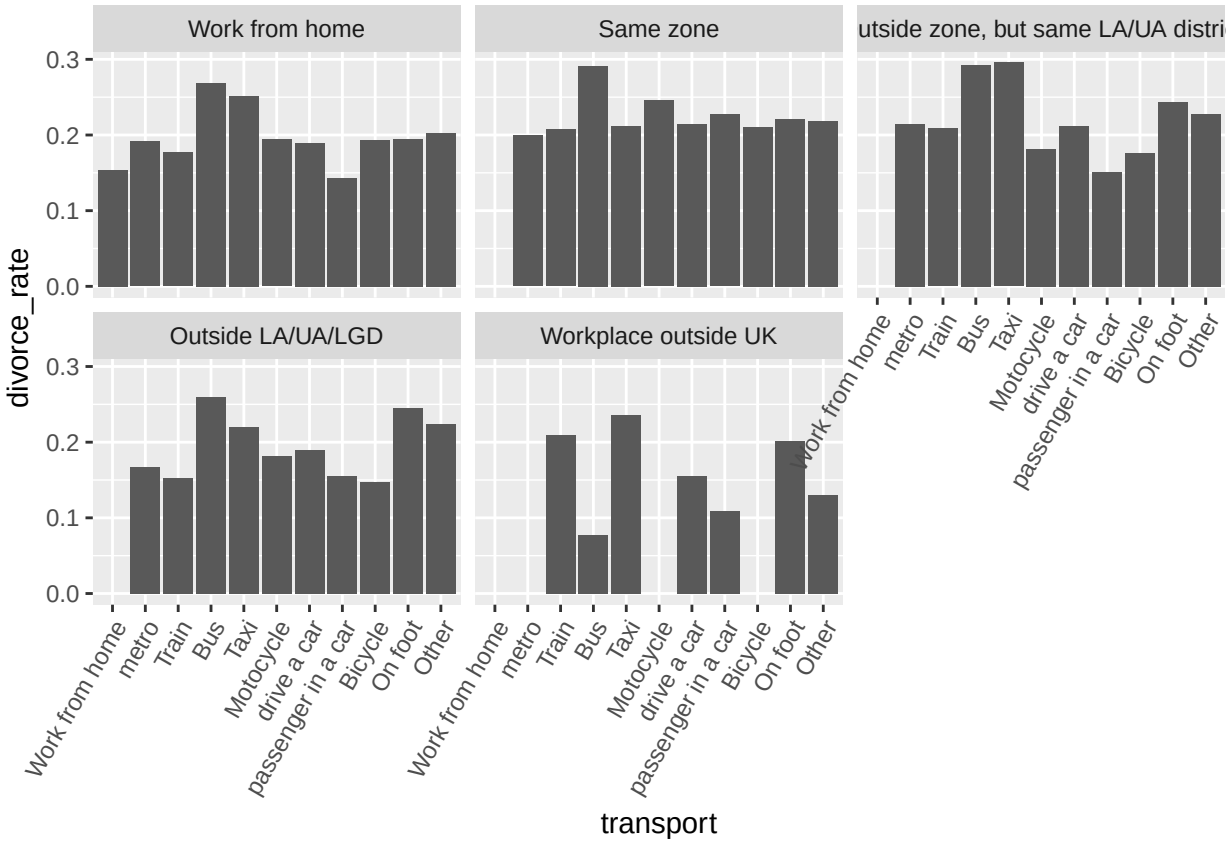
Divorce rate by transport per distance to work

Means of transport alone does not tell us much, but combined with distance from home to work it could provide some insights. We find the divorce_rate for each commute methods by distance to work. We plot two graphs one is with transport on the x-axis, the other with distance to work on x-axis. There is a slight increase in divorce rate as the distance to work increases across all means of transport. When we compare the

different means of transport for the same distance to work, there are more divorced and separated couples who are travelling by bus.

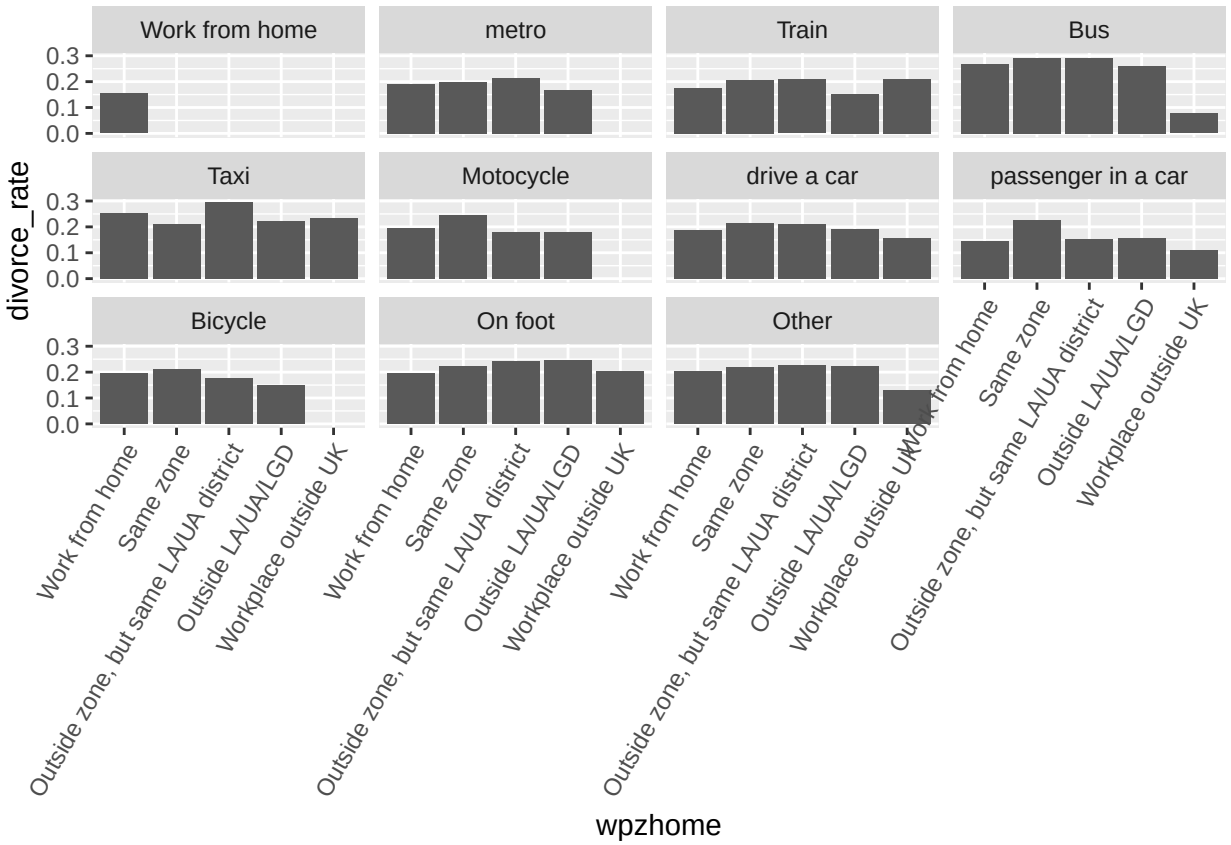
```
# x: transport, wrap by distance to work
region1 %>% select(marstat, transport, wpzhome)%>% filter(!is.na(wpzhome)) %>% group_by(wpzhome, transp
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```




```
# x: distrance to work, wrap by transport
region1 %>% select(marstat, transport, wpzhome)%>% filter(!is.na(wpzhome)) %>% group_by(wpzhome, transp
```

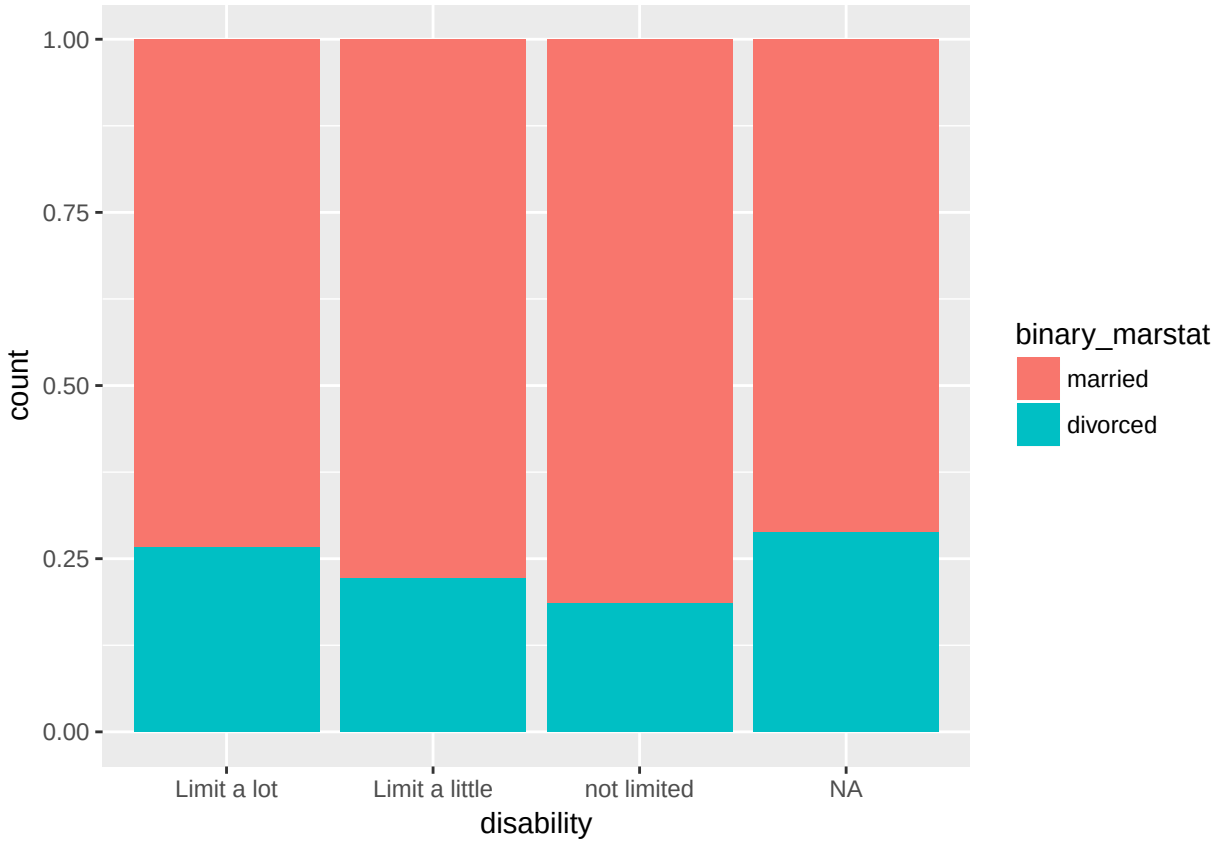
```
## Warning: Removed 3 rows containing missing values (position_stack).
```

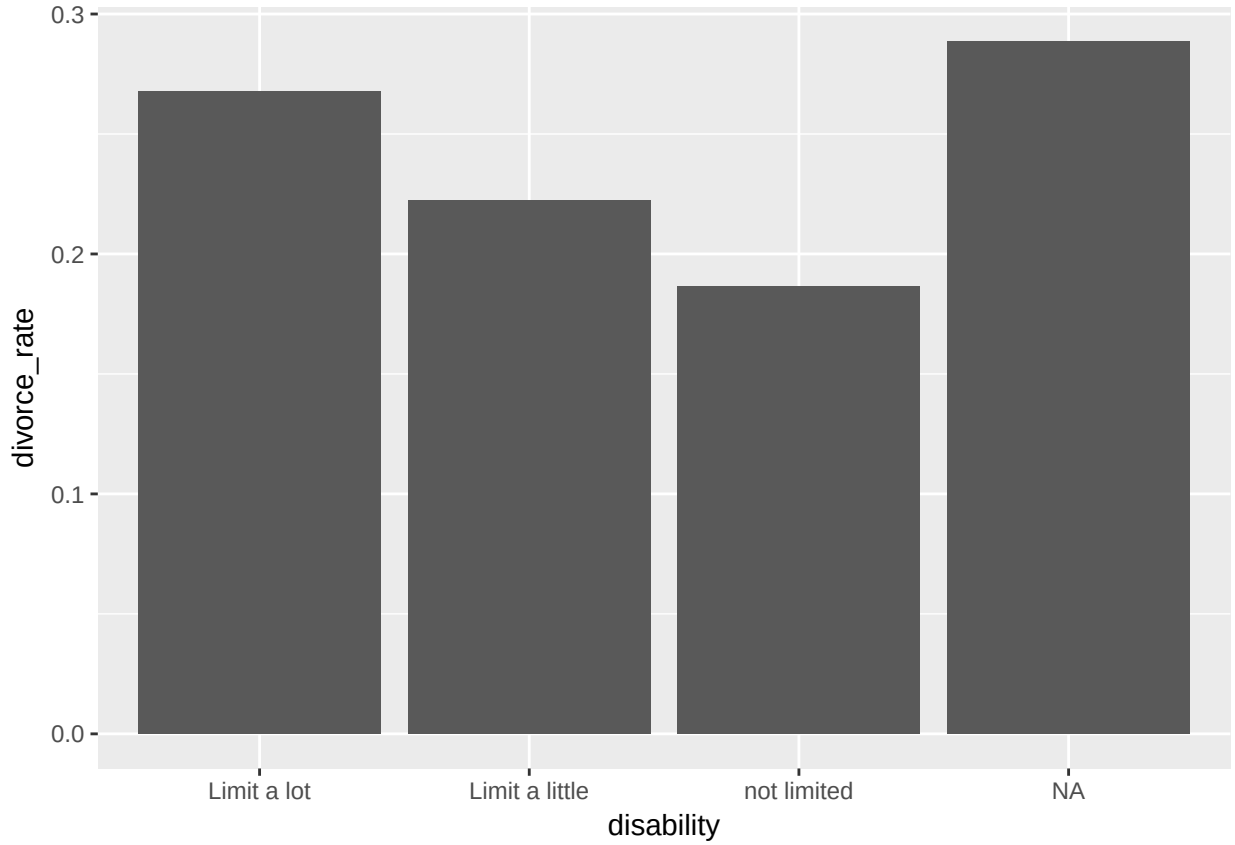


Health & Marriage

```
##
##  Pearson's Chi-squared test
##
## data:  mine$binary_marstat and mine$disability
## X-squared = 5428.6, df = 2, p-value < 2.2e-16
```

A chi-square test of the two variables indicate that the correlation between marital status and long-term health problem is significant. This finding suggests that the two variables are not independent of each other. Long-term health problem will influence marital status.


1.00
0.75
0.50
0.25
0.00
count
Limit a lot
Limit a little
not limited
NA
disability
binary_marstat
married
divorced



The graph shows that the less severe the long-term health problem, the less the divorced rate.

Logistic Regression

A comprehensive regression analysis of all correlated variables

Hypothesis Testing and Prediction Model (Helene's)

Conclusion

Our findings will go here

References

Main Github Repository: https://github.com/dlouhasha/TheGreatWork

Appendix

A list of all variables and descriptions