Gallery of Missing value visualization with in R

در این پروژه تلاش شد با استفاده از کتابخانه‌های tidyverse, dplyr, ggplot2, gridExtra, tidyr و naniar به مصورسازی داده‌های گمشده پرداخته شود.

داده‌های به کار گرفته شده در پروژه، داده‌های custdata می‌باشد که در درس مبانی علم داده‌ها در مبحث داده‌های گمشده و دور افتاده مورد بررسی قرار گرفته شد.

این داده‌ها مربوط به مشتریان بیمه می‌باشند که شمای کلی و خلاصه ای از داده‌ها به صورت زیر است.

```
> names(custdata)
 [1] "custid"       "sex"          "is.employed"  "income"       "marital.stat" "health.ins"   "housing.type" "recent.move"  "num.vehicles"
[10] "age"          "state.of.res"
>
```

```
> summary(custdata)
     custid          sex          is.employed         income                   marital.stat health.ins                              housing.type
 Min.   :   2068   F:440   Mode :logical   Min.   :     0   Divorced/Separated:155   Mode :logical   Homeowner free and clear    :157
 1st Qu.: 345667   M:560   FALSE:73        1st Qu.: 14700   Married           :516   FALSE:159       Homeowner with mortgage/loan:412
 Median : 693403           TRUE :599       Median : 35000   Never Married     :233   TRUE :841       Occupied with no rent       : 11
 Mean   : 698500           NA's :328       Mean   : 53567   Widowed           : 96                   Rented                      :364
 3rd Qu.:1044606                           3rd Qu.: 67000                                            NA's                        : 56
 Max.   :1414286                           Max.   :615000
                                           NA's   :1
 recent.move      num.vehicles        age              state.of.res
 Mode :logical   Min.   :0.000   Min.   :  0.0   California  :100
 FALSE:820       1st Qu.:1.000   1st Qu.: 38.0   New York    : 71
 TRUE :124       Median :2.000   Median : 50.0   Pennsylvania: 70
 NA's :56        Mean   :1.916   Mean   : 51.7   Texas       : 56
                 3rd Qu.:2.000   3rd Qu.: 64.0   Michigan    : 52
                 Max.   :6.000   Max.   :146.7   Ohio        : 51
                 NA's   :56                      (Other)     :600
```

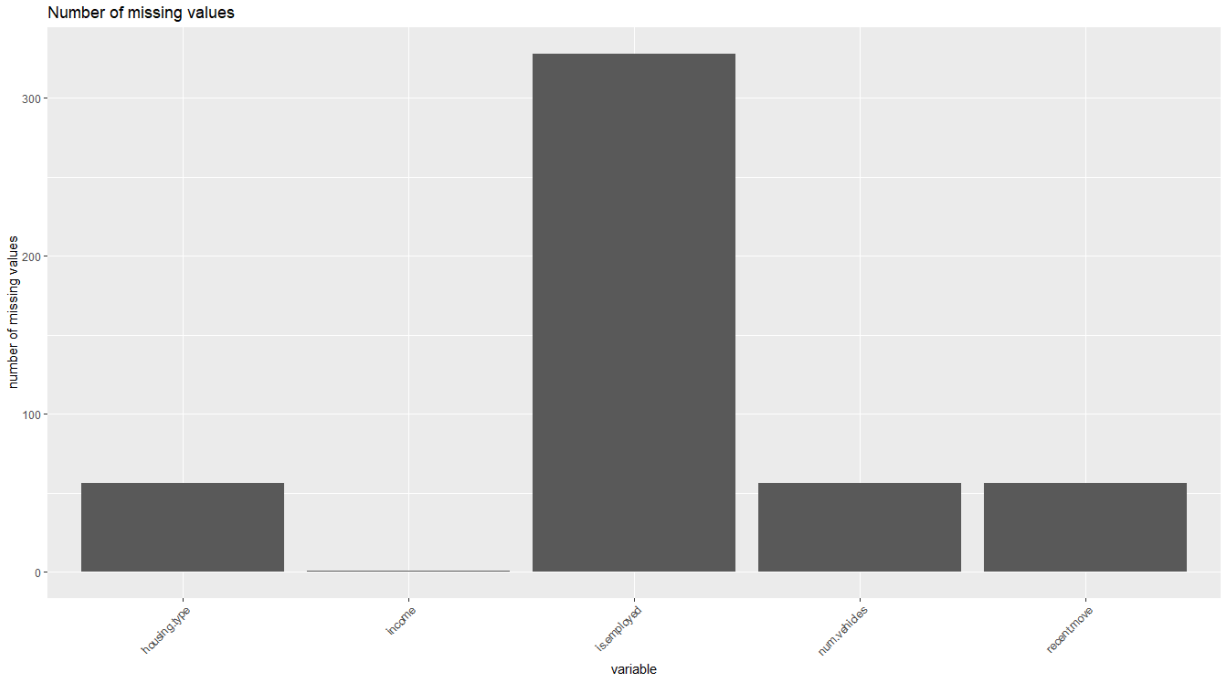
مقادیر گمشده در متغیرهای وضعیت اشتغال، وضعیت مسکن، وضعیت داشتن نقل مکان و تعداد وسیله نقلیه هر خانوار وجود دارد. (همچنین یک داده‌ی گمشده در متغیر درامد وجود دارد که در واقع مربوط به مقدار منفی بوده و به عنوان داده‌ی گمشده آن را تعریف کرده ام).

همچین کد زیر تعداد متغیرهایی که دارای مقادیر گمشده می‌باشند را نمایش می‌دهد.

```
> n_var_miss(df)
[1] 5
```

با استفاده از کتابخانه‌های نام برده شده نمودارهای جالبِ زیر داده‌های گمشده را به تصویر کشیده اند.

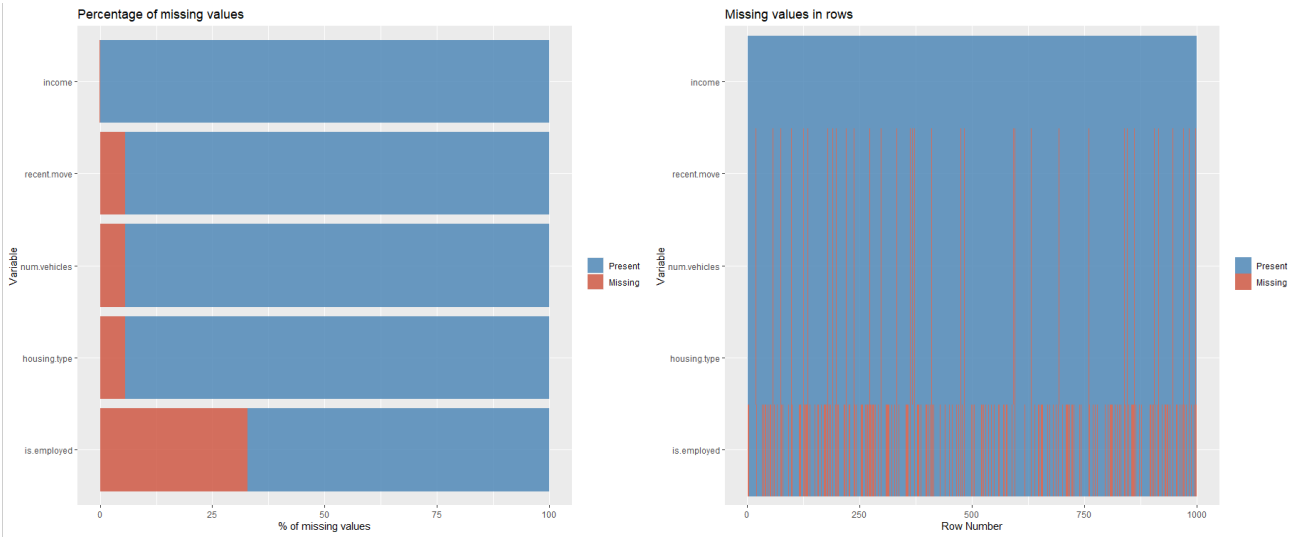
در شکل زیر نمودار میله‌ای ۵ متغیری که شامل داده گمشده می‌باشد، نمایش داده شده و همانطور که مشاهده می‌شود متغیر وضعیت اشتغال دارای بیشترین فراوانی (تقریبا ۳۳۰ داده گمشده) و متغیر درامد دارای کمترین فراوانی (یک داده گمشده) و سه متغیر دیگر هم با تعدادی نزدیک به ۶۰ داده گمشده در رتبه دوم قرار دارند. (که این نتایج در خلاصه آماری نیز مشاهده شد)



در گام بعد نمودارهای زیر را برای مصورسازی داده‌های گمشده استفاده کرده و این نتایج را در آن‌ها هم می‌توان مشاهده کرد.

با توجه به نمودار سمت چپ حدودا ۳۰٪ داده‌های متغیر وضعیت اشتغال در قسمت قرمز رنگ قرار دارد که آن مربوط به داده‌های گمشده می‌باشد و برای سه متغیر دیگر، این مقدار حدودا برابر ۶٪ و برای متغیر درامد نیز مقدار خیلی ناچیزی می‌باشد.

همچنین خطوط قرمز رنگ در نمودار سمت راست محل مقادیر گمشده را نشان می‌دهد که در مکان‌هایی که تراکم این خطوط بیشتر است، داده‌های گمشده نیز بیشتر می‌باشند.

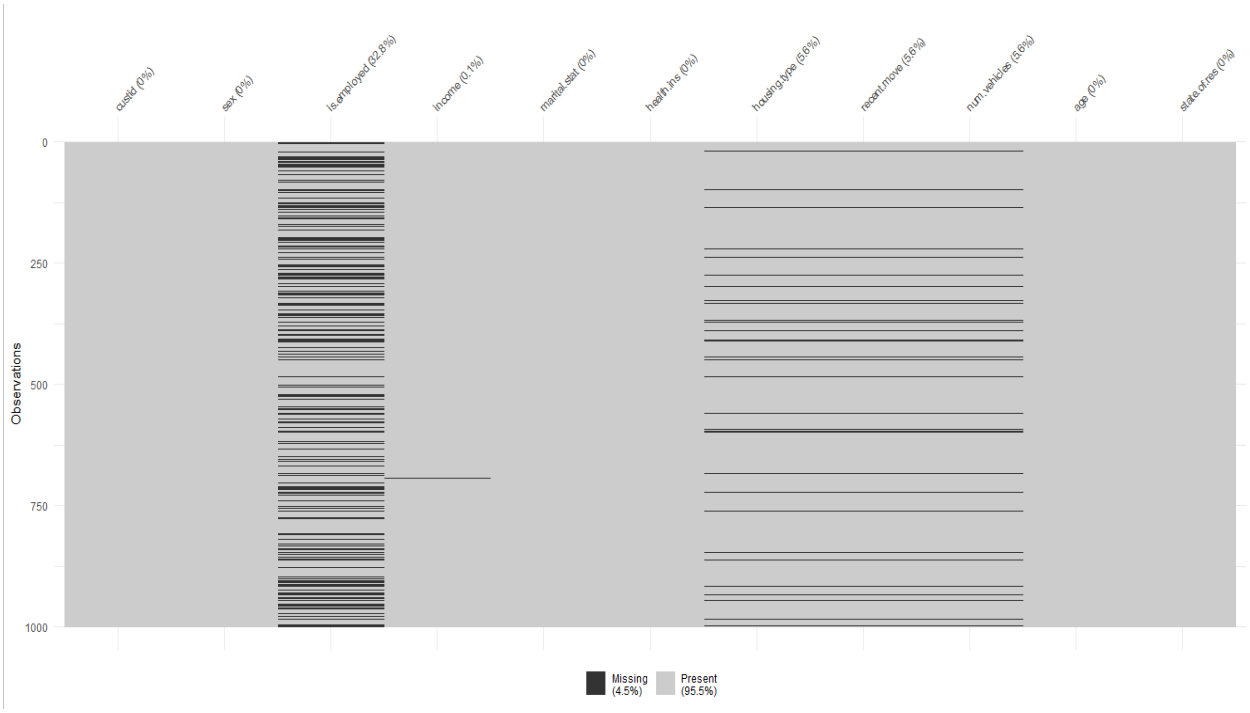


همچنین نمودار زیر مانند آخرین نمودار بررسی شده در بالا می‌باشد که علاوه بر نمایش محل داده‌های گمشده در هر متغیر، درصد وجود آن‌ها را هم مشخص کرده است.

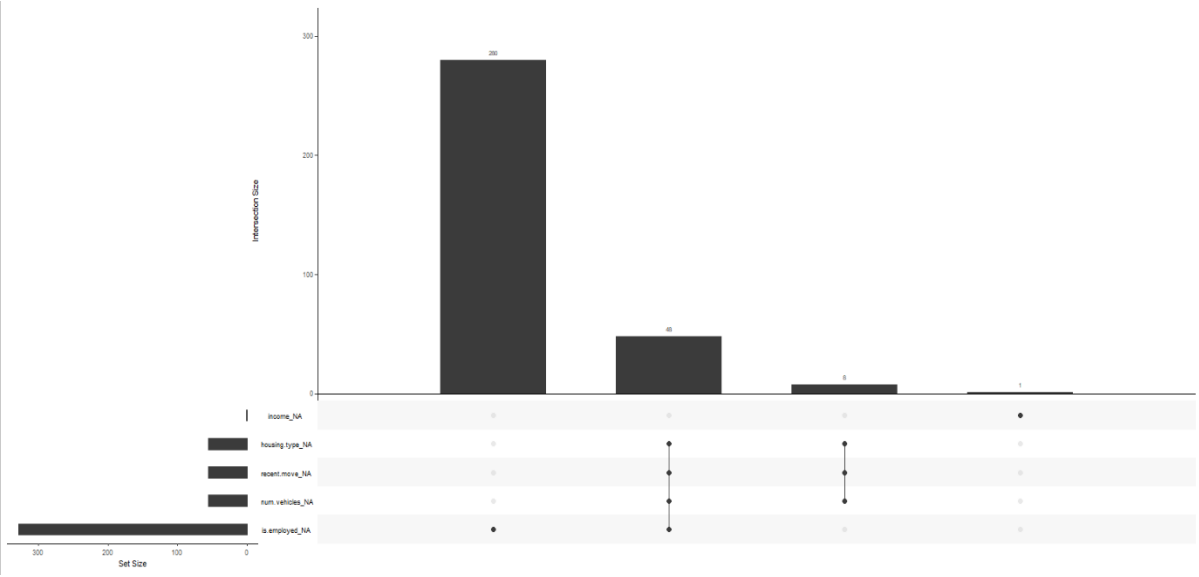
نکته‌ای که در این نمودارها وجود دارد، همان میزان تراکم خطوط و نزدیکی آن‌ها به هم می‌باشد که نشان دهنده میزان داده‌های گمشده در آن محل است.

برای مثال حدودا در مشاهدات ۵۰ ام با توجه به خطوط مشکی پر رنگ (که از کنار هم قرار گرفتنِ چند خط متوالی به وجود آمده اند) میزان گمشدگی زیاد و بیشتر از بقیه‌ی نمونه‌ها می‌باشد و برای متغیر درامد مقدار گمشده در بین مشاهدات ۷۰۰ ام و برای سه متغیر دیگر هم، مقادیر گمشده حول و حوش مشاهده ۶۰۰ ام بیشتر از بقیه می‌باشد.

که به طور کلی ۹۵,۵٪ داده‌ها مشاهده شده و ۴,۵٪ درصد آن‌ها گمشده می‌باشند.

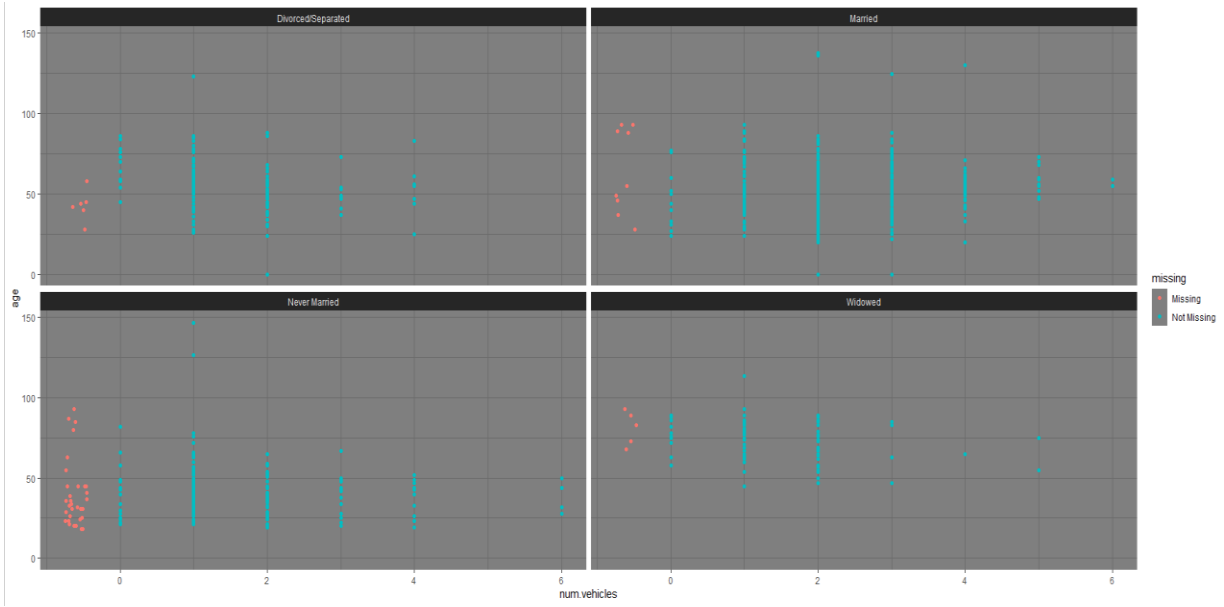


این نتایج در نموار زیر نیز قابل مشاهده می‌باشد.

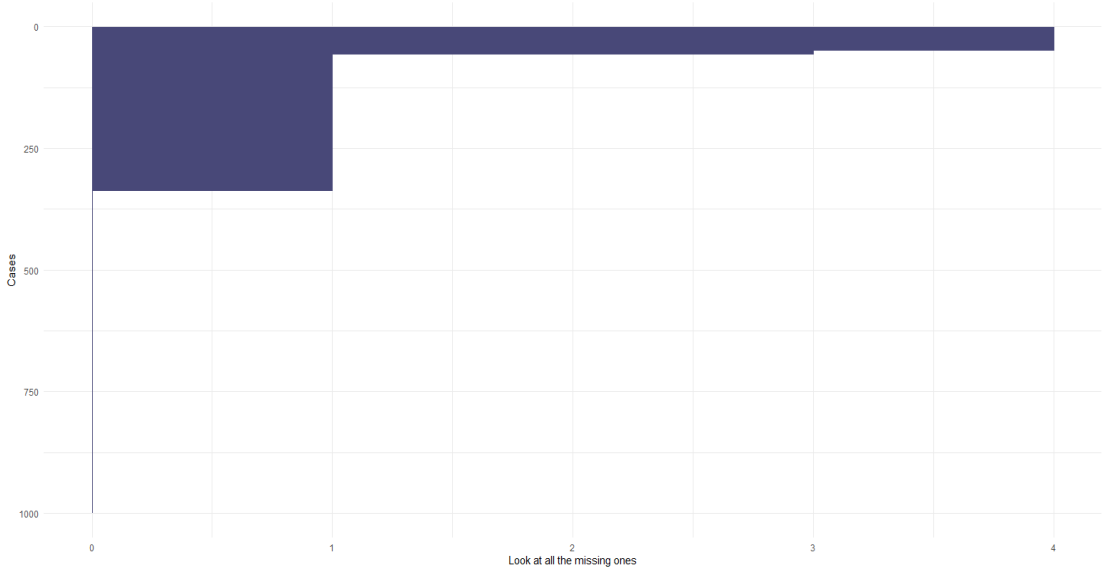


در نمودار بعد به تفکیک متغیرگسسته‌ی وضعیت تاهل، میزان داده‌های گمشده برای دو متغیر تعداد وسیله نقلیه و سن مشاهده می‌شود که با توجه به نمودار در می‌یابیم که ۱) رابطه بین این دو متغیر خطی نیست، ۲) مقادیر گمشده نسبت به مقادیر مشاهده شده به ترتیب در بین افراد بیوه (که دارای سنین ۶۰ تا ۹۰ سال هستند) کمتر از همه، بعد از آن افراد طلاق گرفته (در سنین ۳۰ تا ۶۰ سال) و بعد افراد متاهل (در سنین ۲۵ تا ۹۰ سالگی) و در آخر هم افراد مجرد (افراد بین ۲۰ تا ۹۰ سال) می‌باشد. ۳) این مقادیر گمشده بیشتر بین افرادی است که وسیله نقلیه ندارند.

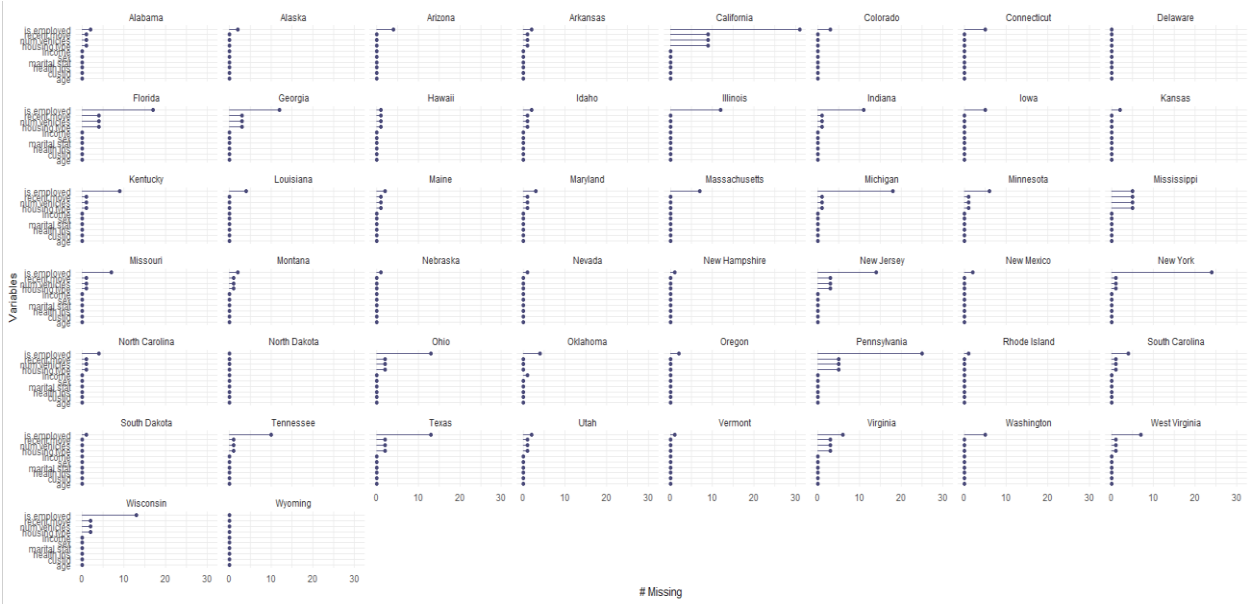
همچنین با توجه به اطلاعات حاصل شده شاید بتوان گفت در اینجا افراد مجرد محافظه کارانه تر از بقیه افراد به خصوص افراد بیوه عمل می‌کنند و اطلاعات کاملی از خودشان به جای نمی‌گذارند و یکی از دلایل علت وجود داده‌های گمشده آن هم با تعداد بیشتر، می‌تواند این موضوع باشد.



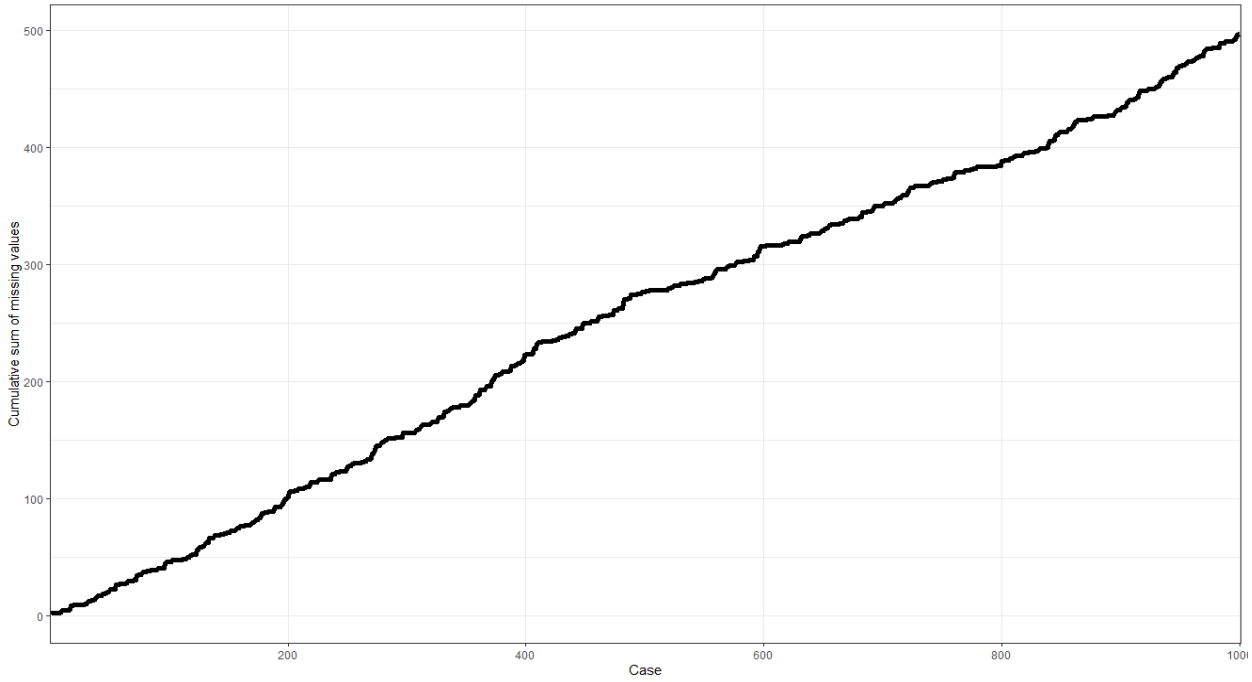
نمودار زیر نیز فراوانی تجمعی مقادیر گمشده را نشان می‌دهد که با توجه به آن بیشتر داده‌های گمشده برای متغیر وضعیت اشتغال در بین مشاهدات ۰ تا ۵۰ می‌باشد و برای متغیرهای دیگر نیز تا مشاهدات ۲۰۰ ام را در بر می‌گیرد.

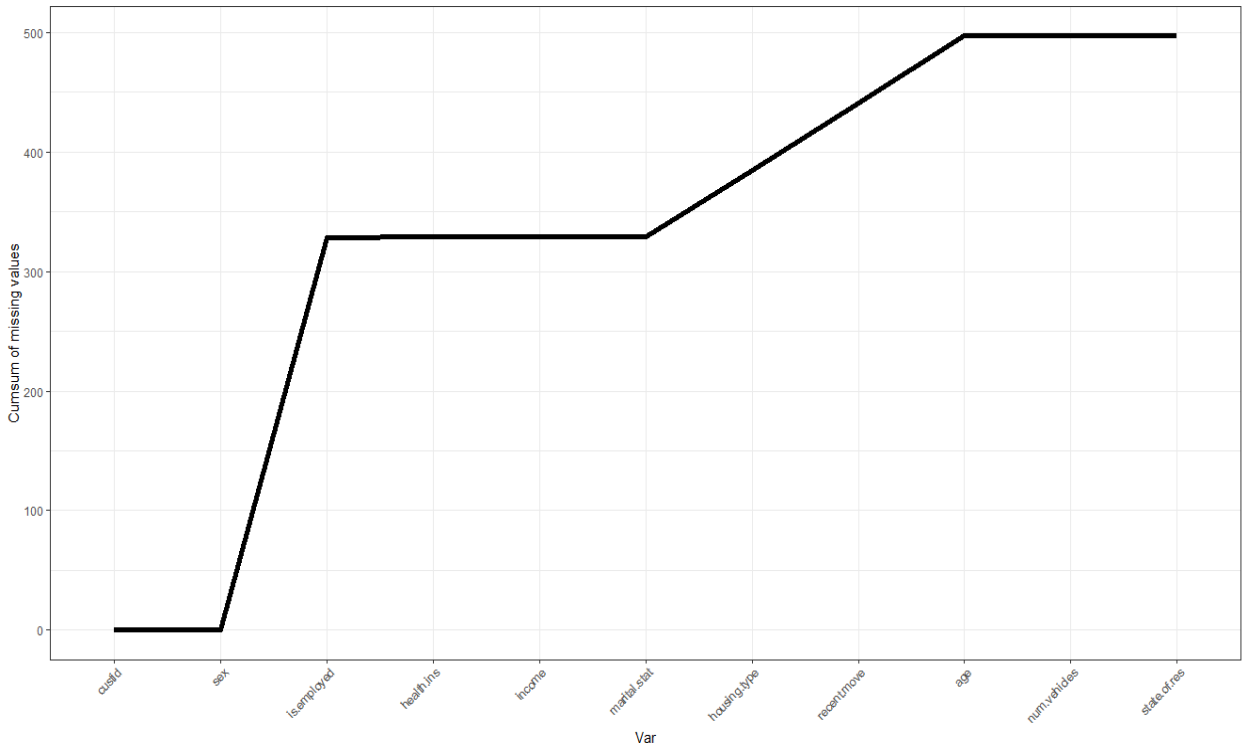


نمودار بعدی میزان مقادیر گمشده در ایالت‌های مختلف را در ۵ متغیر نشان می‌دهد که بیشترین میزان گمشدگی در ایالت‌های کالیفرنیا، نیویورک، پنسیلوانیا و فلوریدا و متغیر وضعیت اشتغال قرار دارد. (احتمالا جمع آوری اطلاعات و داده‌ها در این ایالت‌ها به خوبی صورت نگرفته، باید روند جمع آوری اطلاعات را در این بخش‌ها بیشتر بررسی کنند!!! البته یک عامل موثر که اصلا در اینجا بررسی نشده است، می‌تواند جمعیت و شلوغی باشد که روی تعداد مقادیر گمشده نیز تاثیر می‌گذارد).

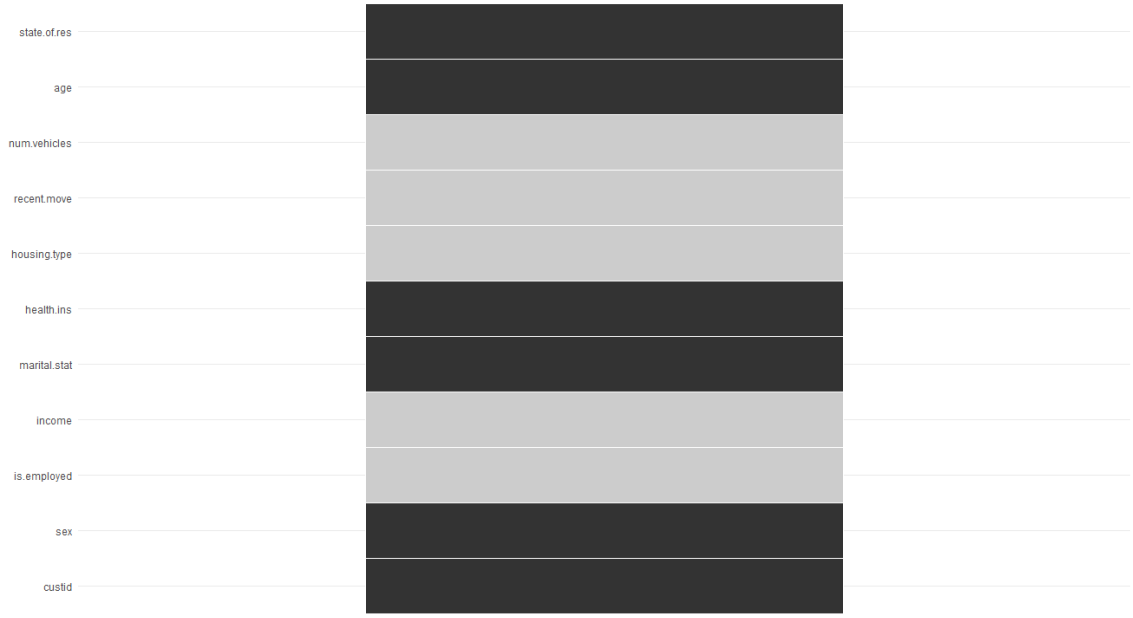


همچنین دو شکل زیر به ترتیب از بالا به پایین نمودار فراوانی تجمعی داده‌های گمشده نسبت به نمونه‌ها و نسبت به متغیرها را نمایش داده اند.



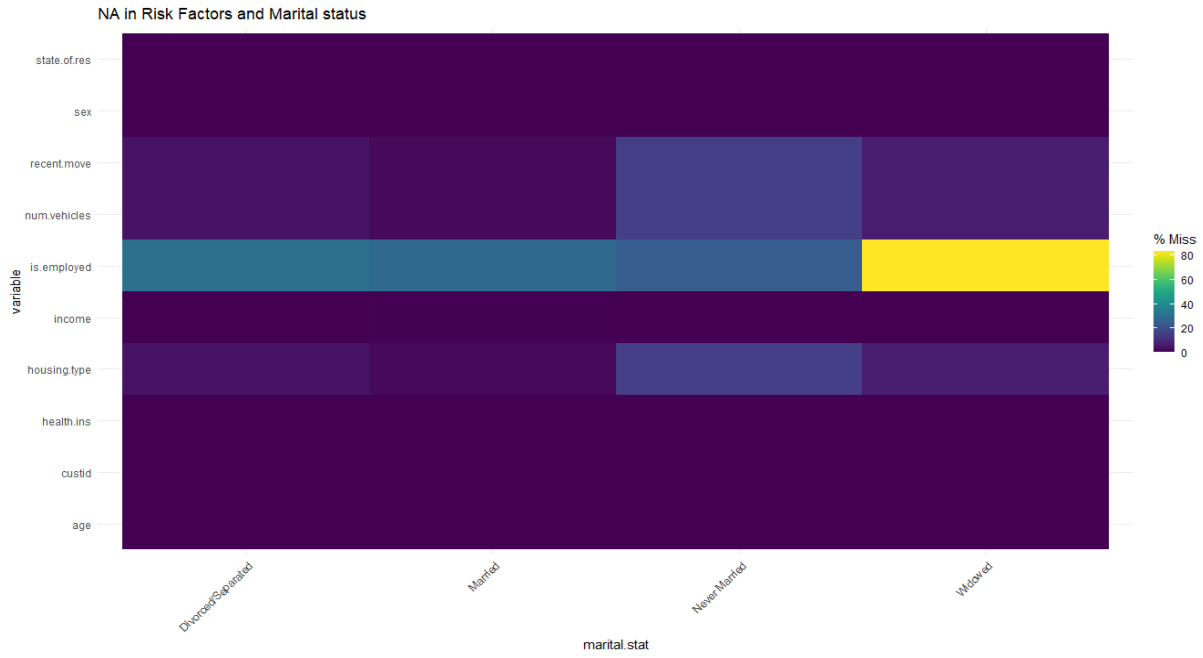


بخش‌های طوسی در نمودار زیر نمایانگر متغیرهایی است که دارای داده‌های گمشده و بخش‌های مشکی نمایانگر داده‌ها مشاهده شده (observed) می‌باشد.



نمودار آخر نیز مقادیر گمشده را با توجه به وضعیت تاهل نشان می‌دهد که همانطور که مشاهده می‌شود، مقادیر گمشده با توجه به رنگ‌های در نظر گرفته شده به ترتیب در بین افراد بیوه در متغیر وضعیت اشتغال بیشتر می‌باشد، بعد از آن در بین افراد مطلقه در همین متغیر و سپس افراد متاهل و مجرد است.

بعد از آن هم بیشترین مقدار در متغیر وضعیت مسکن برای افراد مجرد و بعد از آن در متغیر وضعیت نقل مکان بین همین افراد، سپس بین افراد مطلقه در متغیر وضعیت مسکن بعد هم متغیر وضعیت نقل مکان می‌باشد.



در اینجا نتایج متفاوتی با نمودار ص۴ حاصل شد به دلیل اینکه آن شکل، مربوط به نمودار پراکنش است و دو متغیر تعریف شده، مسلما تاثیر گذار هستند اما در اینجا تنها فراوانی مورد نظر بوده است.

همچنین کدهای مورد استفاده در فایلی به نام missingcode پیوست شده است.

Visualization of missing data