Project Title

A STAT 139 Final Project

Yuyue Wang, Xiangru Shu, Chengye Liu, Chia Chi (Michelle) Ho

Due December 13, 2017

Abstract

Introduction

- Obesity is an exerbating problem in the US.
- Explore the association of 21 different factors with bmi, 13 of which are behavior-related factors such as the typical number of hours sleep per night

Methods

- Data description
  - data source is NHANES 2013-2014
  - Variables of interest
  - Only consider adults of age 20 or above
- Data preprocessing & assumptions
  - Merge data by participant sequence number
  - Exclude don't know/refused/missing values — discuss implications in limitations
- Perform EDA
- Fit regression models
- Check assumptions

Results

Exploratory Data Analysis

Limitations

Conclusions

Appendix

Appendix I: Data preprocessing

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##         gender            age           race            edu       marriage
##              0              0              0              0              0
##        famsize      famincome    alcohol12yr     alcoholfrq        grocery
##              0              0              0              0              0
##         eatout       delivery           milk  meals_nothome meals_fastfood
##              0              0              0              0              0
##      depressed  sleep_trouble       activity         tv_hrs       sleep_hr
##              0              0              0              0              0
##          smoke            bmi      bmi_class
##              0              0              0
```

```
##         gender            age           race            edu       marriage
##       "factor"      "integer"       "factor"       "factor"       "factor"
##        famsize      famincome    alcohol12yr     alcoholfrq        grocery
##      "integer"       "factor"       "factor"      "integer"      "integer"
##         eatout       delivery           milk  meals_nothome meals_fastfood
##      "integer"      "integer"       "factor"      "integer"      "integer"
##      depressed  sleep_trouble       activity         tv_hrs       sleep_hr
##       "factor"       "factor"       "factor"       "factor"      "integer"
##          smoke            bmi      bmi_class
##       "factor"      "numeric"    "character"
```

Appendix II: Exploratory Data Analysis

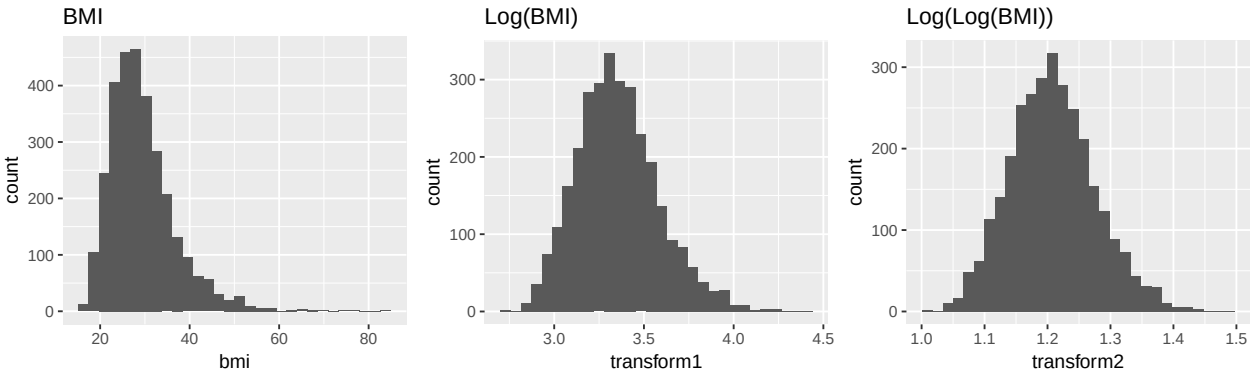
Response Variable (bmi)

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

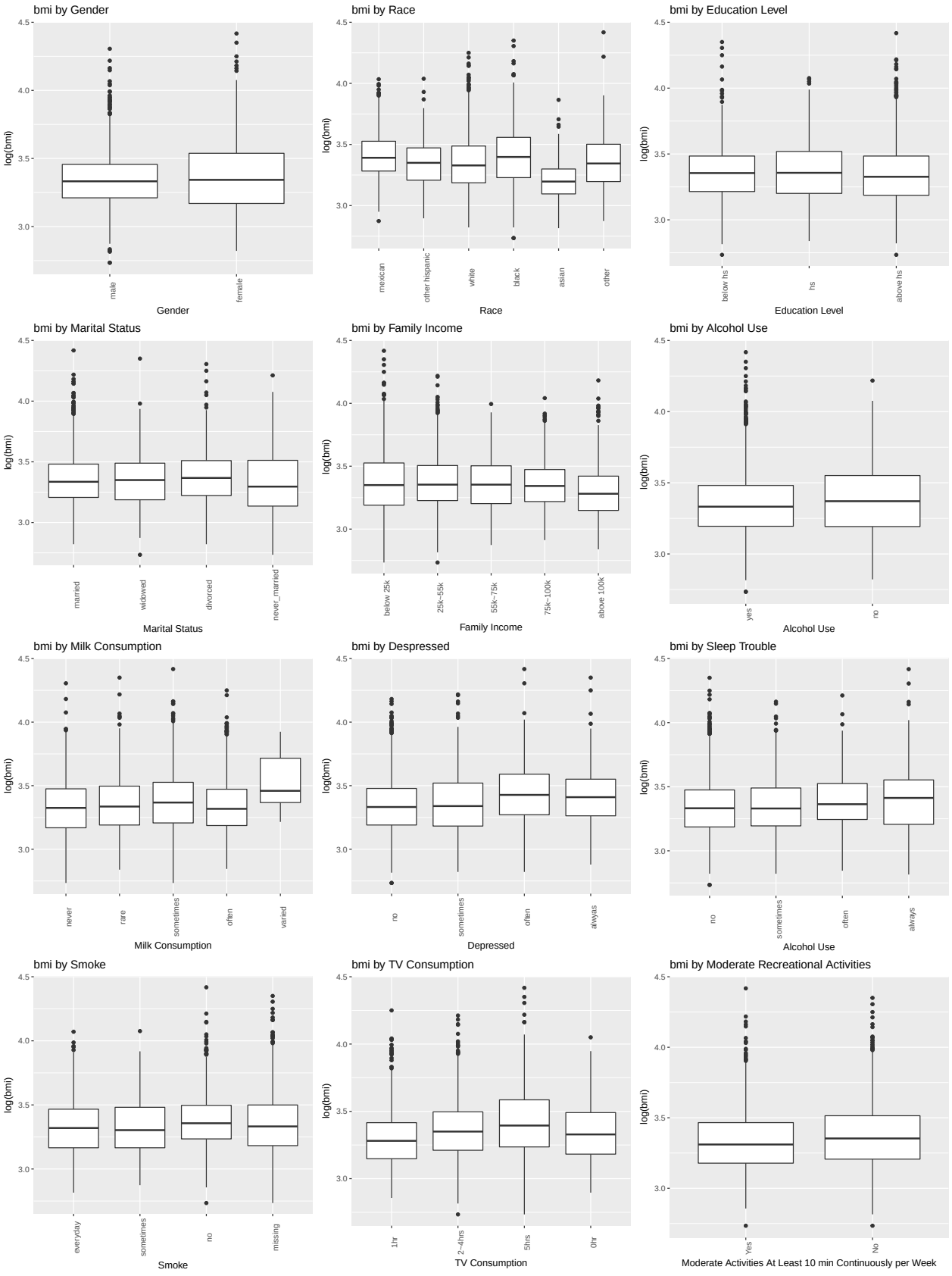
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Predictor Variables

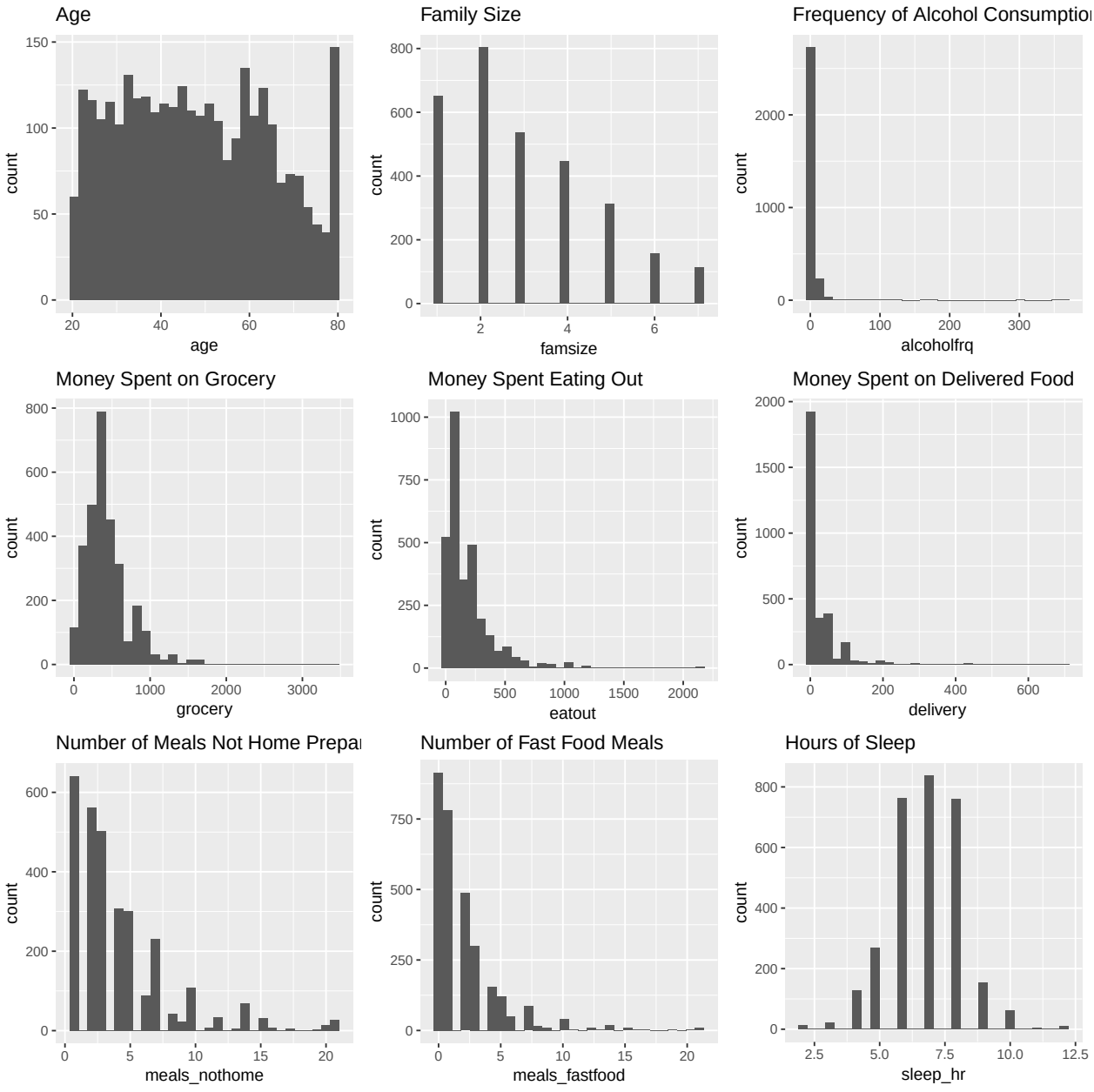
Categorical Variables


bmi by Gender
log(bmi)
male
female
Gender
bmi by Race
mexican
other hispanic
white
black
asian
other
Race
bmi by Education Level
below hs
hs
above hs
Education Level
bmi by Marital Status
married
widowed
divorced
never_married
Marital Status
bmi by Family Income
below 25k
25k~55k
55k~75k
75k~100k
above 100k
Family Income
bmi by Alcohol Use
yes
no
Alcohol Use
bmi by Milk Consumption
never
rare
sometimes
often
varied
Milk Consumption
bmi by Despressed
no
sometimes
often
always
Depressed
bmi by Sleep Trouble
no
sometimes
often
always
Alcohol Use
bmi by Smoke
everyday
sometimes
no
missing
Smoke
bmi by TV Consumption
1hr
2~4hrs
5hrs
0hr
TV Consumption
bmi by Moderate Recreational Activities
Yes
No
Moderate Activities At Least 10 min Continuously per Week

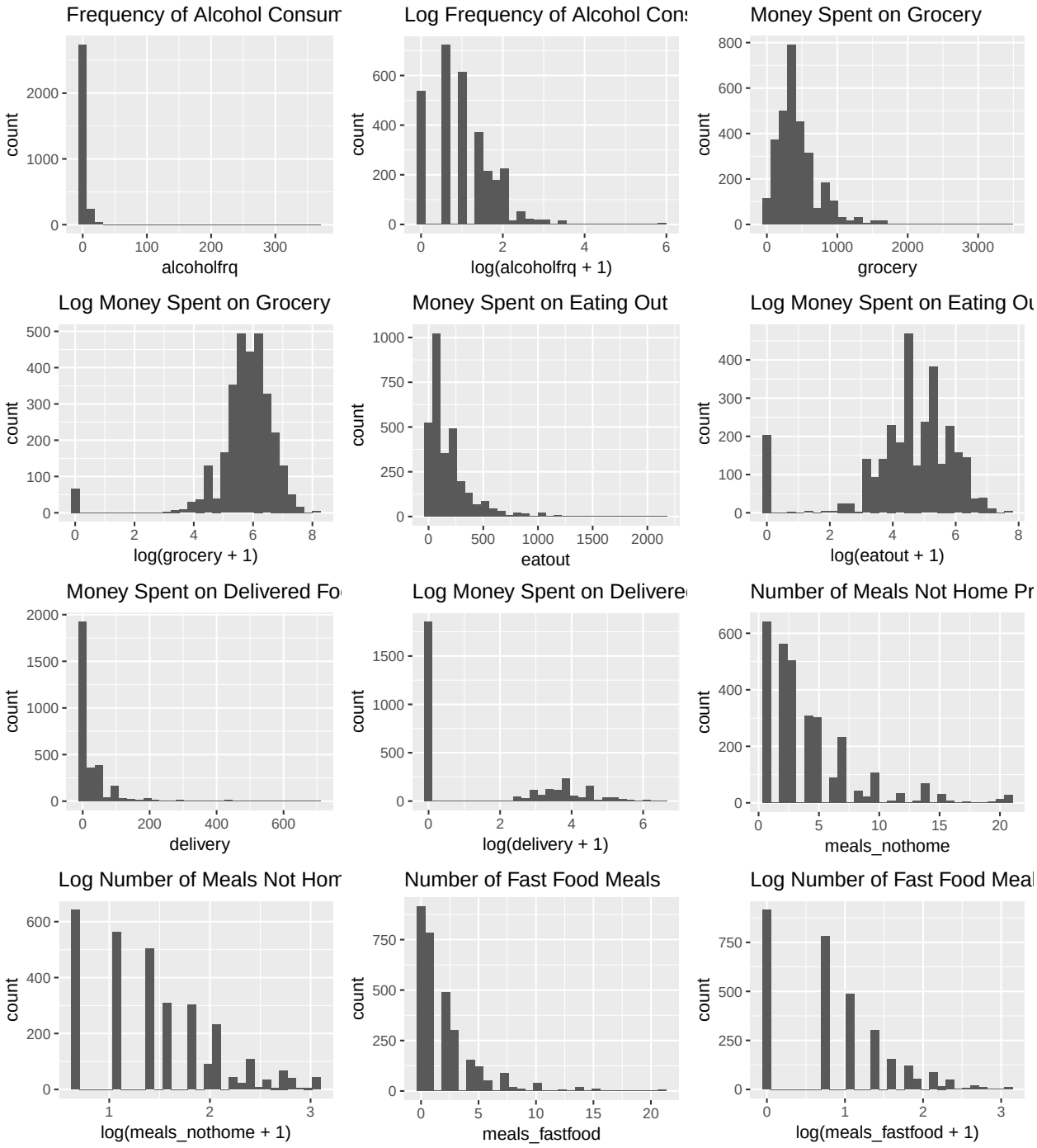
Numeric Variables

Distribtuion of numeric variables

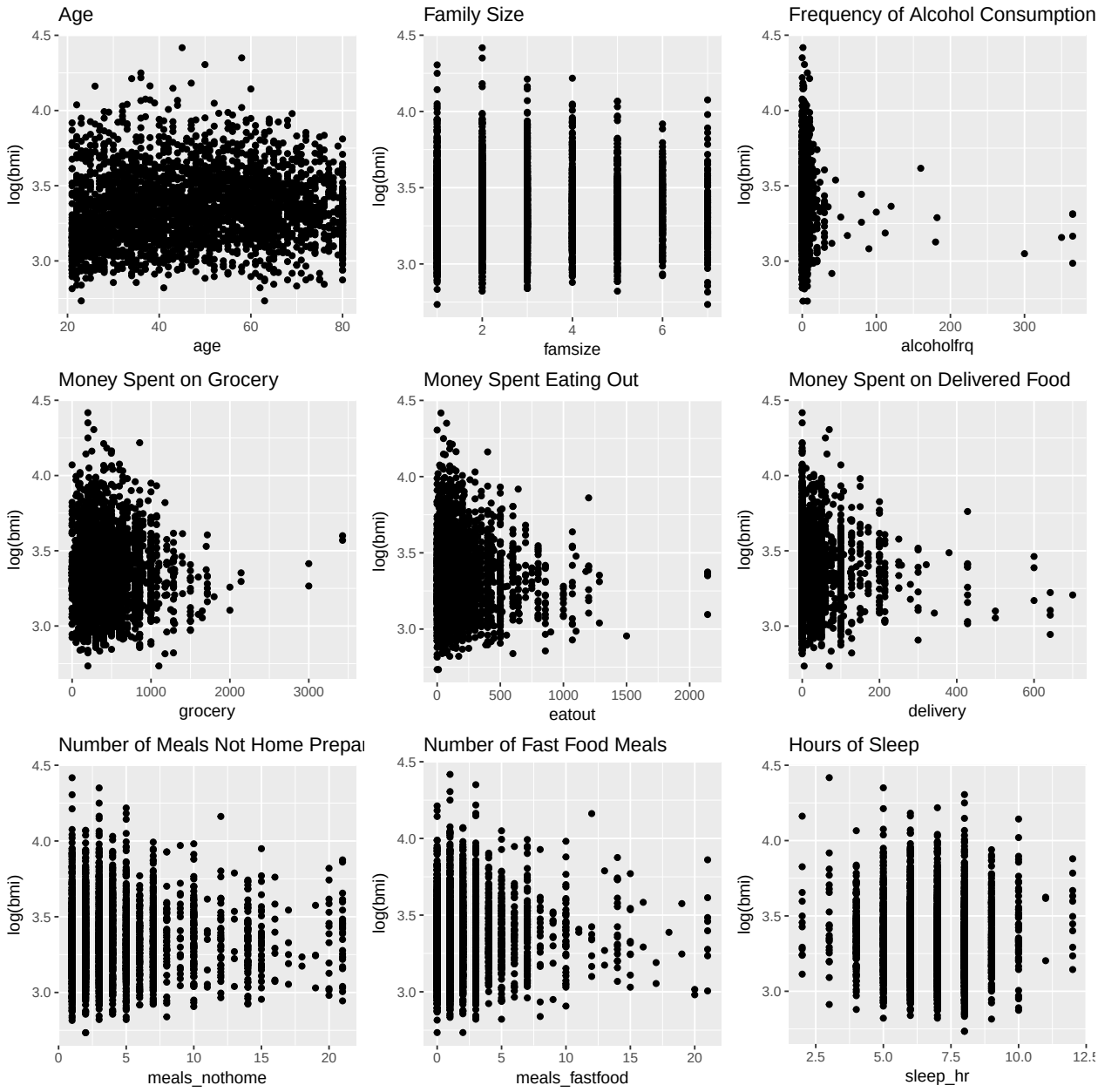
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

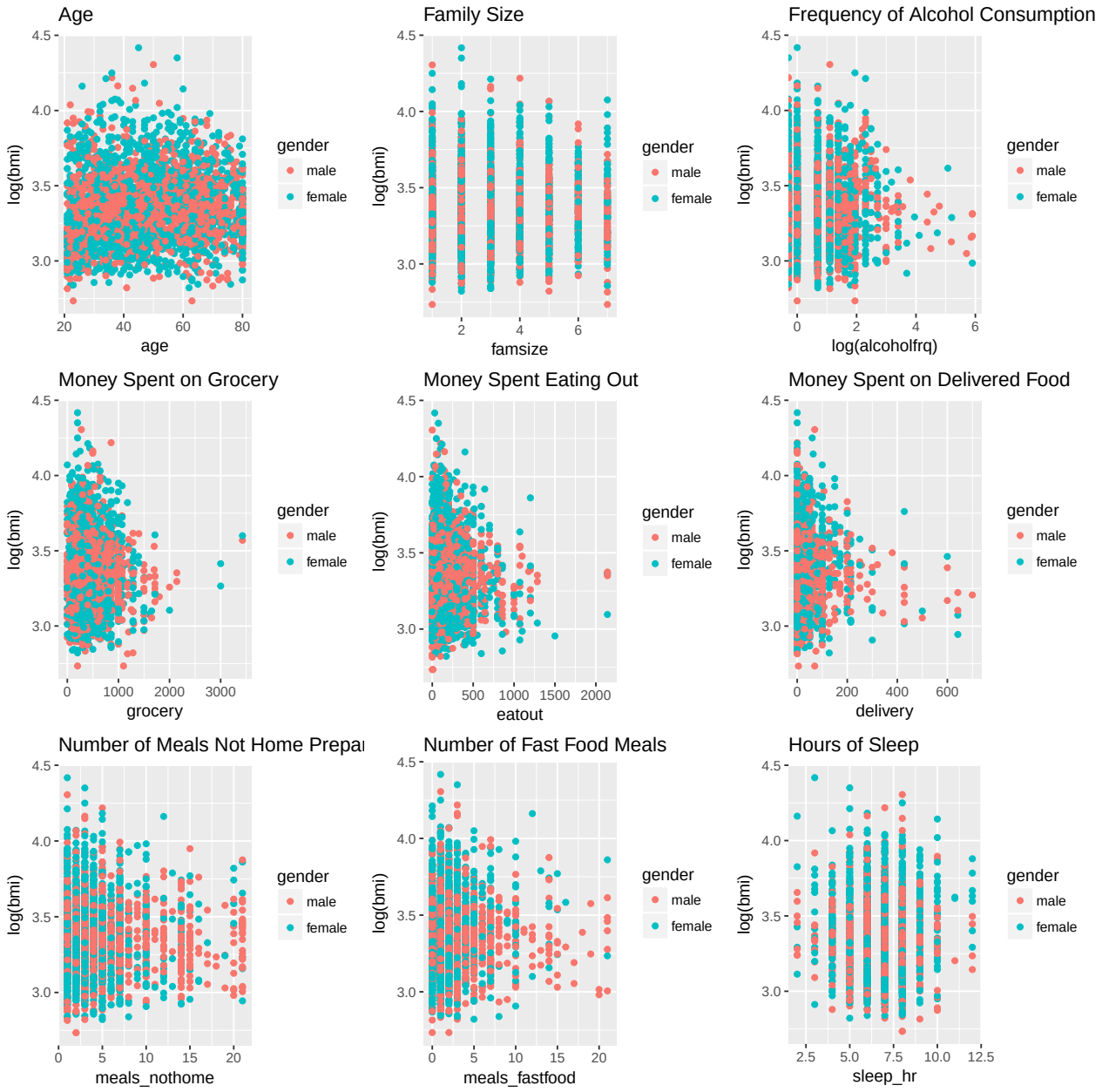


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Response vs. numeric distribution


Age
log(bmi)
age
Family Size
famsize
Frequency of Alcohol Consumption
alcoholfrq
Money Spent on Grocery
grocery
Money Spent Eating Out
eatout
Money Spent on Delivered Food
delivery
Number of Meals Not Home Prepar
meals_nothome
Number of Fast Food Meals
meals_fastfood
Hours of Sleep
sleep_hr



parallel corrd plot to figure out if interaction may be helpful