Data Overview

Matthew Coleman

2/12/2020

Data Overview Write R code to complete the following and include as an appendix in your Rmarkdown file. – Read the dataset(s) in. – Report the dimensions of each. – Summarize the missingness in the data. – Split the data into training, validation (optional), and test sets. Report dimensions. – Using only your training set: * Summarize the response variable using numerical or graphical summaries. * Fit a very basic model to the training set. e.g. a logistic regression with 1 or 2 explanatory v

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.2     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bnb <- read.csv('AB_NYC_2019.csv')
dims <- dim(bnb)

sprintf('Our dataset has %d observations and %d attributes', dims[1],dims[2])
```

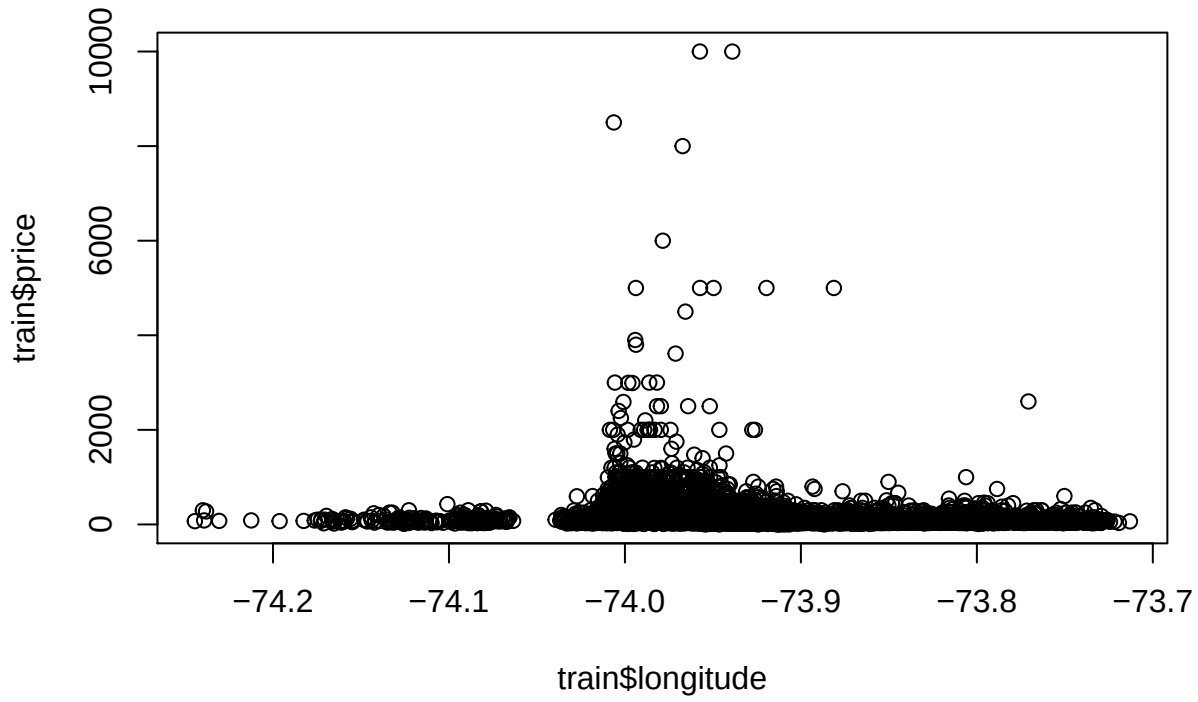
```
## [1] "Our dataset has 48895 observations and 16 attributes"
```

```
num_com <- sum(complete.cases(bnb))

per_com <- sum(complete.cases(bnb))/nrow(bnb) * 100

cat(sprintf('there are %d number of complete cases, which amounts to %2f percent of the dataset',
            num_com, per_com ))
```

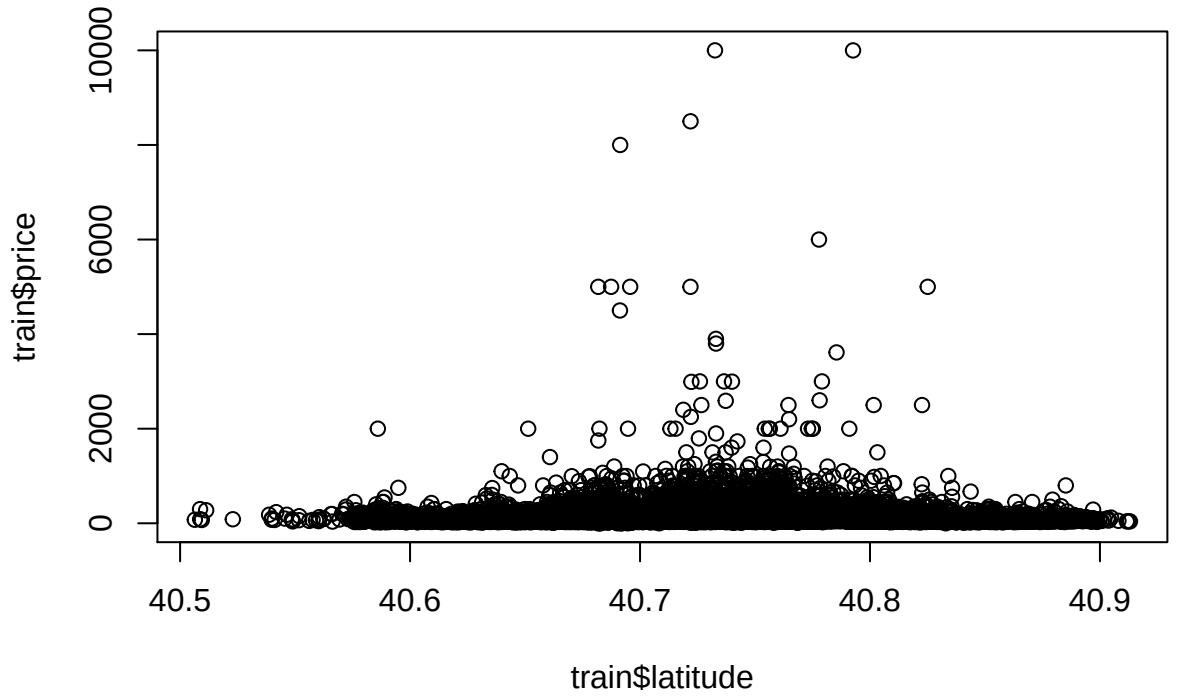
```
## there are 38843 number of complete cases, which amounts to 79.441661 percent of the dataset
```

```
bnb <- bnb[complete.cases(bnb),]
set.seed(123)

train.ind <- sample(1:nrow(bnb),nrow(bnb)*.80 )

train <- bnb[train.ind,]
test <- bnb[-train.ind,]
```

```r
plot(train$longitude, train$price)
```




```r
plot(train$latitude, train$price)
```




```r
hist(train$price)
```

As we can see by the pair plot, there is a serious issue of overplotting, which may require that we reduce the number of observations in the dataset we analyze. We may also need log transformations on some of the variables such as the price. We will dive more in depth with the data cleaning when we begin to work more with the dataset.

```
train <- train %>% mutate(price_greater = ifelse(train$price>median(train$price),1,0))
```

```
gfit <- glm(price_greater~longitude + latitude, data = train, family = 'binomial')
summary(gfit)
```

```
##
## Call:
## glm(formula = price_greater ~ longitude + latitude, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8750  -1.1100  -0.2026   1.0506   2.8491
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1371.6715    28.2620  -48.53   <2e-16 ***
## longitude     -16.6905     0.3437  -48.56   <2e-16 ***
## latitude        3.3724     0.2219   15.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43078  on 31073  degrees of freedom
```

```
## Residual deviance: 39835  on 31071  degrees of freedom
## AIC: 39841
##
## Number of Fisher Scoring iterations: 4
```