Course 2 Section 1.17 - Your turn

Jiaying Wu

04/10/2020

```r
library(tidyverse)
library(naniar)
library(visdat)
```

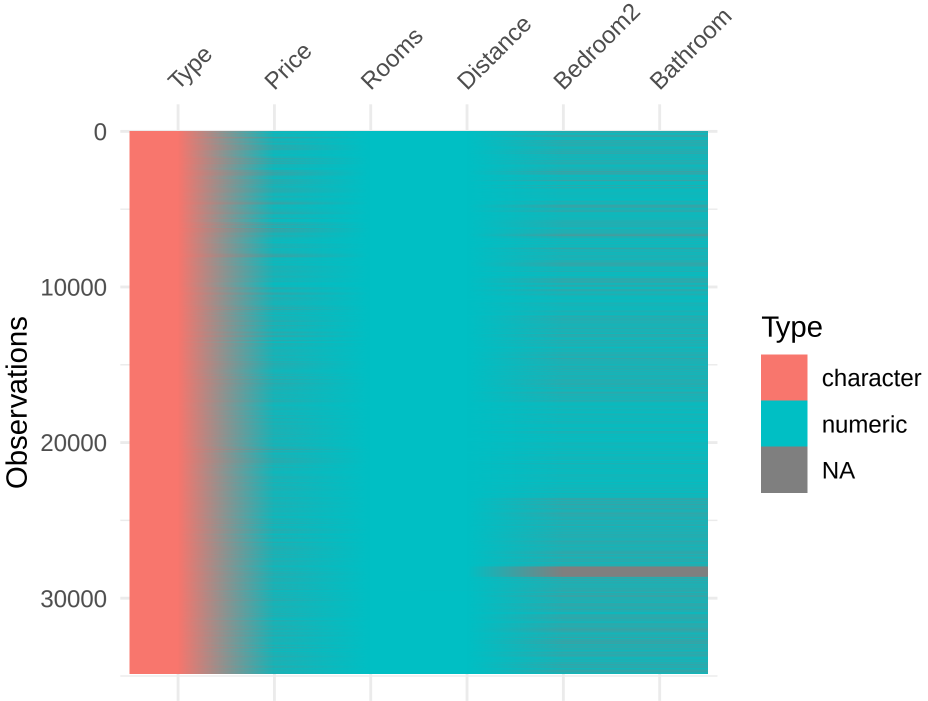
```r
# Read the data
houses <- read_csv(here::here("data", "Melbourne_housing_FULL.csv"))
```

```
## Warning: 5 parsing failures.
##   row           col expected actual
## 18524 Propertycount a double   #N/A '/Users/jiayingwu/Desktop/R4-Material/data/Melbourne_housing_FUL
## 26889 Propertycount a double   #N/A '/Users/jiayingwu/Desktop/R4-Material/data/Melbourne_housing_FUL
## 29484 Distance      a double   #N/A '/Users/jiayingwu/Desktop/R4-Material/data/Melbourne_housing_FUL
## 29484 Postcode      a double   #N/A '/Users/jiayingwu/Desktop/R4-Material/data/Melbourne_housing_FUL
## 29484 Propertycount a double   #N/A '/Users/jiayingwu/Desktop/R4-Material/data/Melbourne_housing_FUL
```

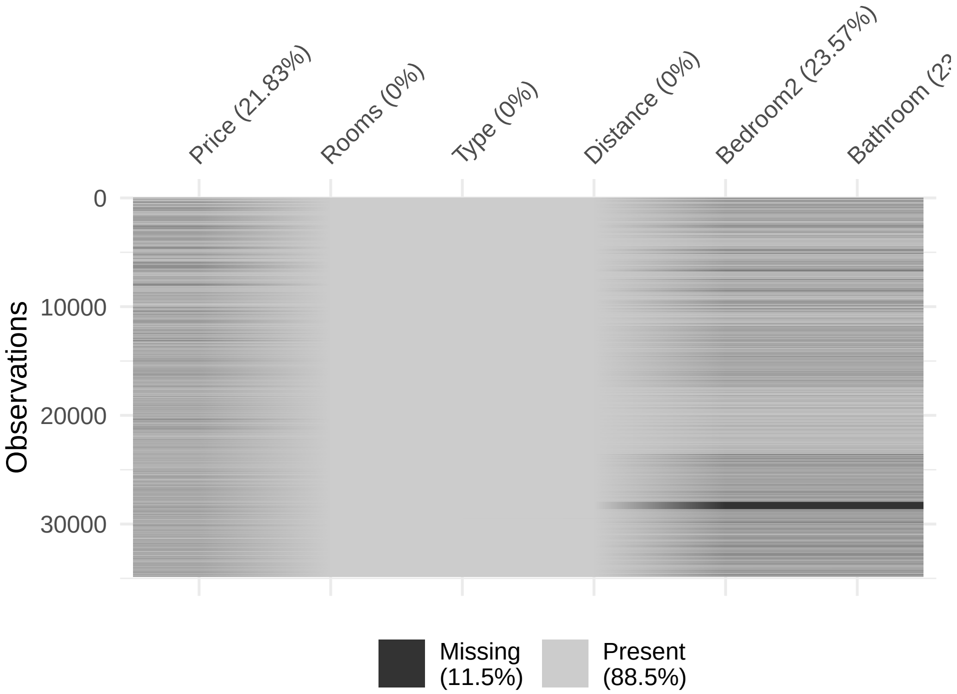
Examine house prices in Melbourne, Australia

Q1. Make an overview plot of the full data. Which variables have missings? Focus only on the variables Price, Rooms, Type, Distance, Bedroom2, Bathroom.

```r
# 1. Keep only variables Price, Rooms, Type, Distance, Bedroom2, Bathroom
houses_sub <- houses %>%
  select(Price, Rooms, Type, Distance, Bedroom2, Bathroom)

# Map of data coloured by variable type and NA
houses_sub %>%
  vis_dat()
```



```r
# Missingness map
houses_sub %>%
  vis_miss()
```



Q2. Make a missing values summary of all the data. What proportion of observations are missing on Price?

```
# 2. Missing variables summary table
miss_var_summary(houses_sub)
```

```
## # A tibble: 6 x 3
##   variable n_miss pct_miss
##   <chr>     <int>    <dbl>
## 1 Bathroom   8226 23.6
## 2 Bedroom2   8217 23.6
## 3 Price      7610 21.8
## 4 Distance      1  0.00287
## 5 Rooms         0  0
## 6 Type          0  0
```

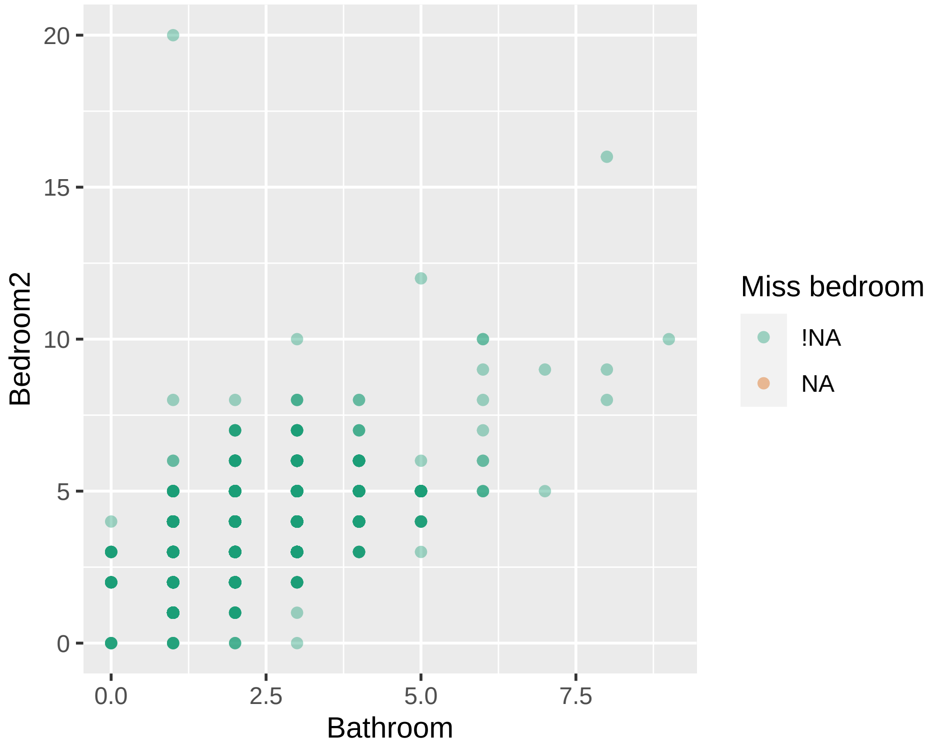
Q3. Remove the observations that have missing values on Price because this is the response variable that we want to ultimately predict. You can't build a stable model of house price if you don't know the price.

```
# 3. Remove missing house price values
houses_sub <- houses_sub %>%
  filter(!is.na(Price))
```

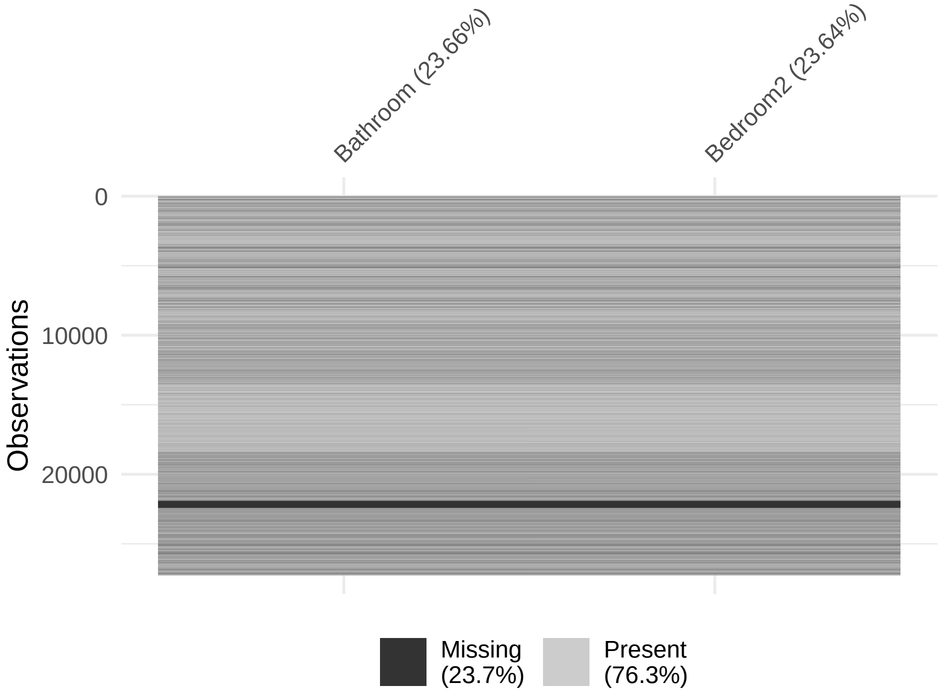
Q4. Make the shadow matrix, and plot Bathroom vs Bedroom2 coloured by missingness on Bedroom2. Why don't any missing values show up?

```
# 4. Scatter plot of bath vs. bed coloured by missingess in bed
houses_sub_shadow <- houses_sub %>%
  bind_shadow()

houses_sub_shadow %>%
  ggplot(aes(x = Bathroom, y = Bedroom2, colour = Bedroom2_NA)) +
  geom_point(alpha = 0.4) +
  # Dark2 palette to accommodate colour blindness
  scale_colour_brewer("Miss bedroom", palette = "Dark2")
```

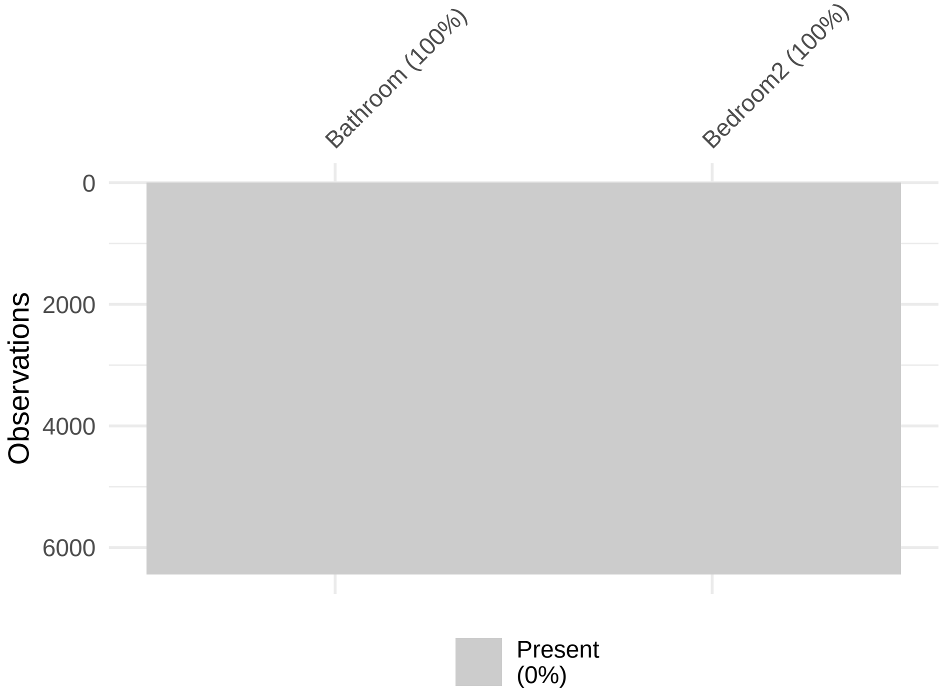
```
## Warning: Removed 6447 rows containing missing values (geom_point).
```



```r
# Missingness map with just bathroom and bedroom
houses_sub_shadow %>%
  select(Bathroom, Bedroom2) %>%
  vis_miss()
```



```r
# Missingness map based on missings in bedroom
houses_sub_shadow %>%
  select(Bathroom, Bedroom2) %>%
  filter(is.na(Bedroom2)) %>%
  vis_miss()
```



Missing values don't show because all missing values in bedroom are also missing in bathroom

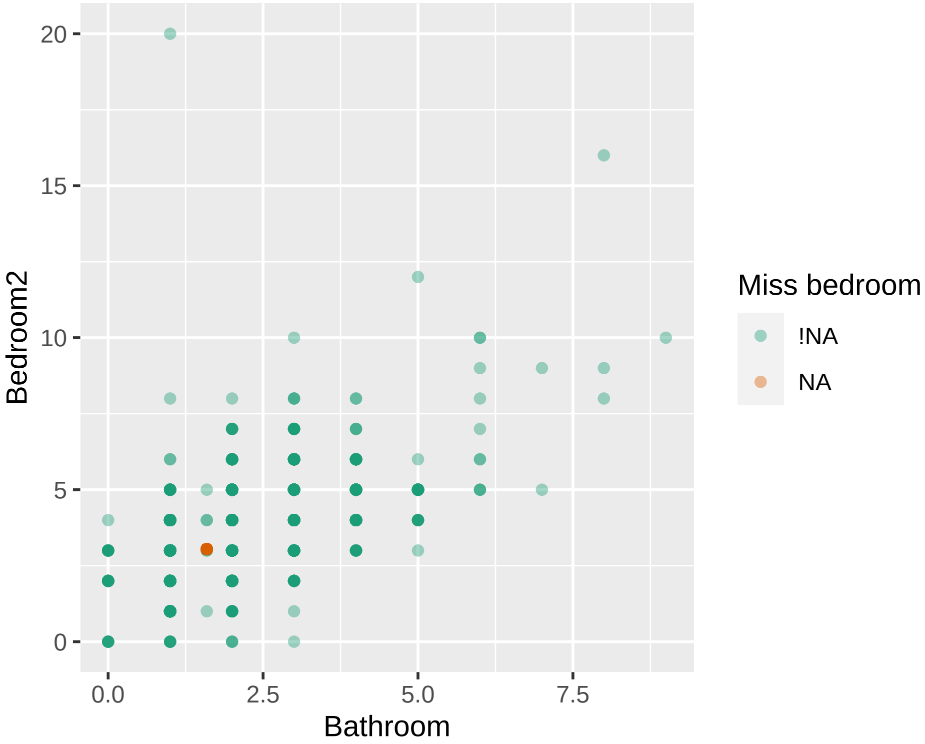
Q5. Impute the missing values for Bedroom2 and Bathroom, by using mean imputation.

```r
# 5. Impute missing values in bed and bath with their mean
houses_sub_shadow_mean <- houses_sub_shadow %>%
  mutate(Bedroom2 = ifelse(is.na(Bedroom2), mean(Bedroom2, na.rm = TRUE), Bedroom2),
         Bathroom = ifelse(is.na(Bathroom), mean(Bathroom, na.rm = TRUE), Bathroom))

# Mean bed and bath
houses_sub_shadow %>%
  summarise(mean_bed = mean(Bedroom2, na.rm = TRUE),
            mean_bath = mean(Bathroom, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   mean_bed mean_bath
##      <dbl>     <dbl>
## 1     3.05      1.59
```

Q6. Make a plot of the two variables, with the imputed values coloured. Describe the pattern that you see.

```r
# 6. Scatter plot of bath vs. bed coloured by imputed values
houses_sub_shadow_mean %>%
  ggplot(aes(x = Bathroom, y = Bedroom2, colour = Bedroom2_NA))  +
  geom_point(alpha = 0.4) +
  scale_colour_brewer("Miss bedroom", palette = "Dark2")
```

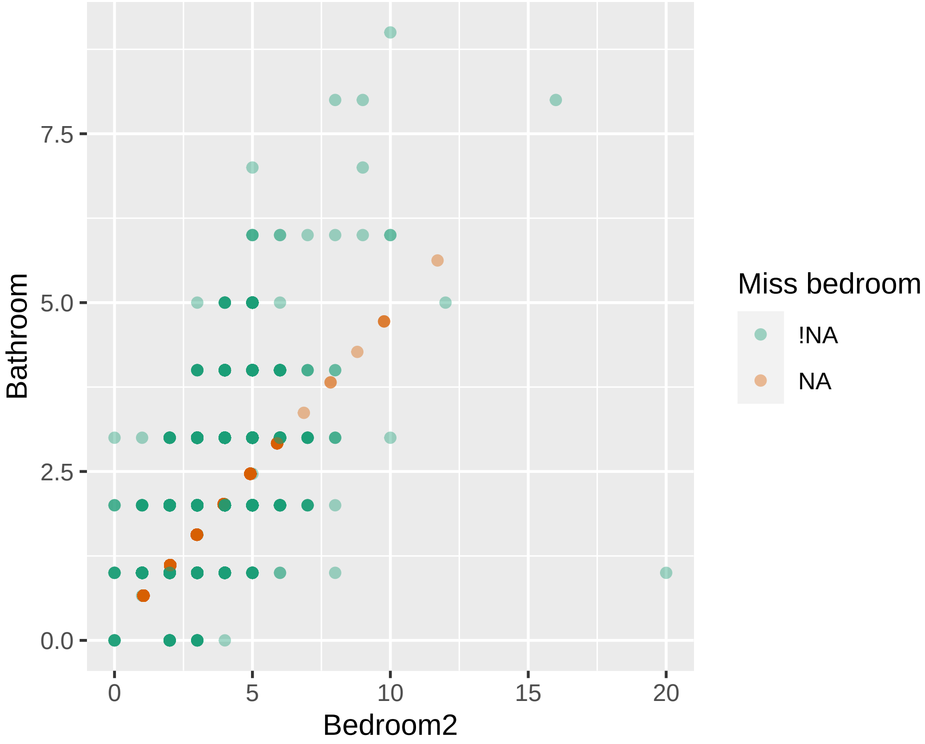


Q7. Impute the missing values for Bedroom2 and Bathroom, using a linear model on the variable Rooms.

```r
# 7. Linear regression

# Use houses_sub_shadow and not houses_sub_shadow_mean because houses_sub_shadow_mean
# has already imputed missing values in bath and bed (used the mean)

# Run a linear regression of bedroom on room
br2 <- lm(Bedroom2 ~ Rooms, data = houses_sub_shadow)

# Run a linear regression of bathroom on room
ba <- lm(Bathroom ~ Rooms, data = houses_sub_shadow)
```

Q8. Make a plot of the two variables, with the imputed values coloured. Is this better or worse than the mean value imputed values? Explain your thinking.

```r
# 8. Scatter plot after inputation with linear regression

# Impute missing values for  bedroom and bathroom based on above regression
houses_sub_shadow_linreg <- houses_sub_shadow %>%
  mutate(Bedroom2 = ifelse(is.na(Bedroom2), predict(br2, new = houses_sub_shadow), Bedroom2),
         Bathroom = ifelse(is.na(Bathroom), predict(ba, new = houses_sub_shadow), Bathroom))

# Scatter plot
ggplot(houses_sub_shadow_linreg,
       aes(x = Bedroom2, y = Bathroom, colour = Bedroom2_NA)) +
  geom_point(alpha = 0.4) +
  scale_colour_brewer("Miss bedroom", palette = "Dark2")
```


Bathroom
Bedroom2
Miss bedroom
!NA
NA