EDA2

Emily Goldfarb

2023-12-07

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

EDA using vehicles.csv

```
data <- read_csv("vehicles.csv")
```

```
## Rows: 426880 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (18): url, region, region_url, manufacturer, model, condition, cylinder...
## dbl   (6): id, price, year, odometer, lat, long
## lgl   (1): county
## dttm  (1): posting_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

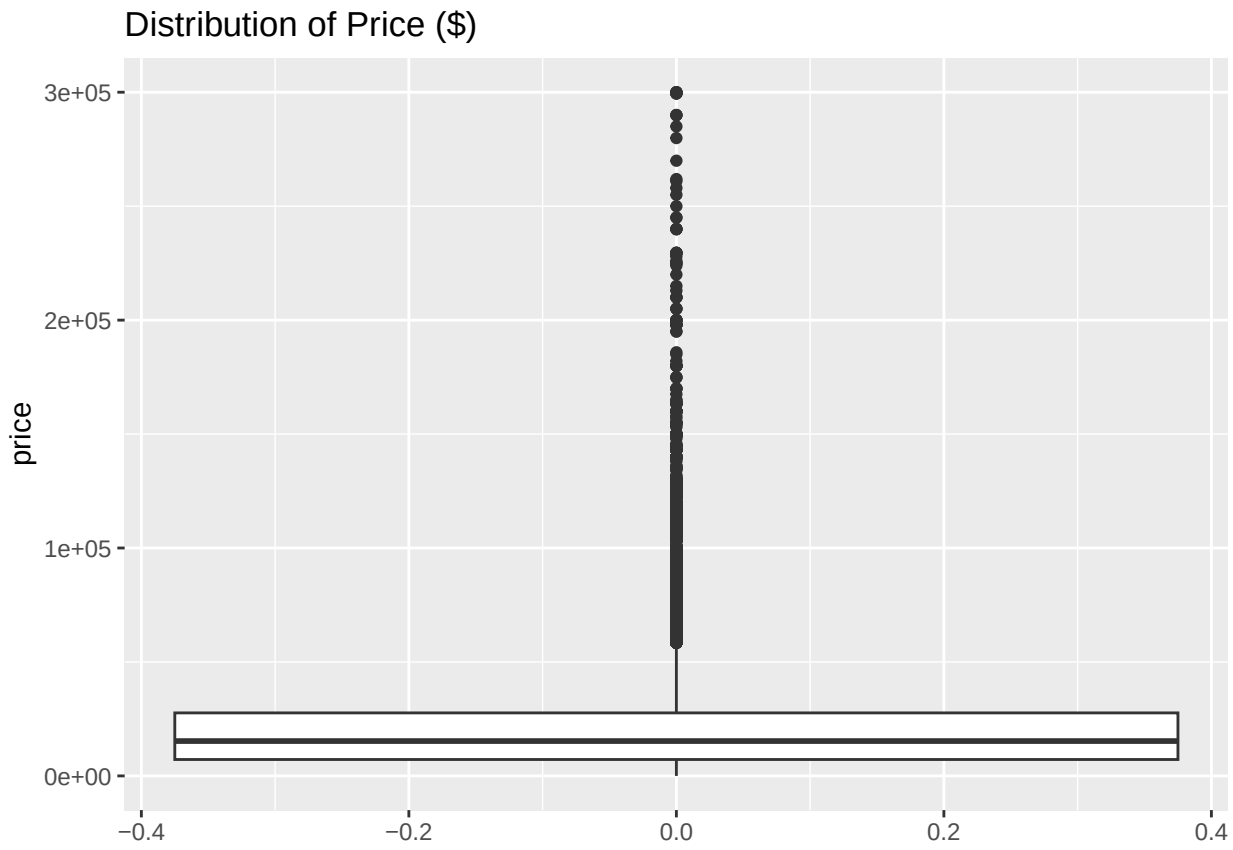
Removing cols based on python notebook preprocessing Data 2

```
data_rc <- data %>% select(-c(county, description, image_url, region_url,url))


# setting max expected price to be 300 K, price > 0, odometer < 999999 based on
# python notebook preprocessing Data 2

data1 <- data_rc %>% filter(price < 300000, price > 0, odometer < 999999)
```

Examining breakdown of price and odometer by vehicle type and transmission type

Distribution of price, boxplots with and without outliers

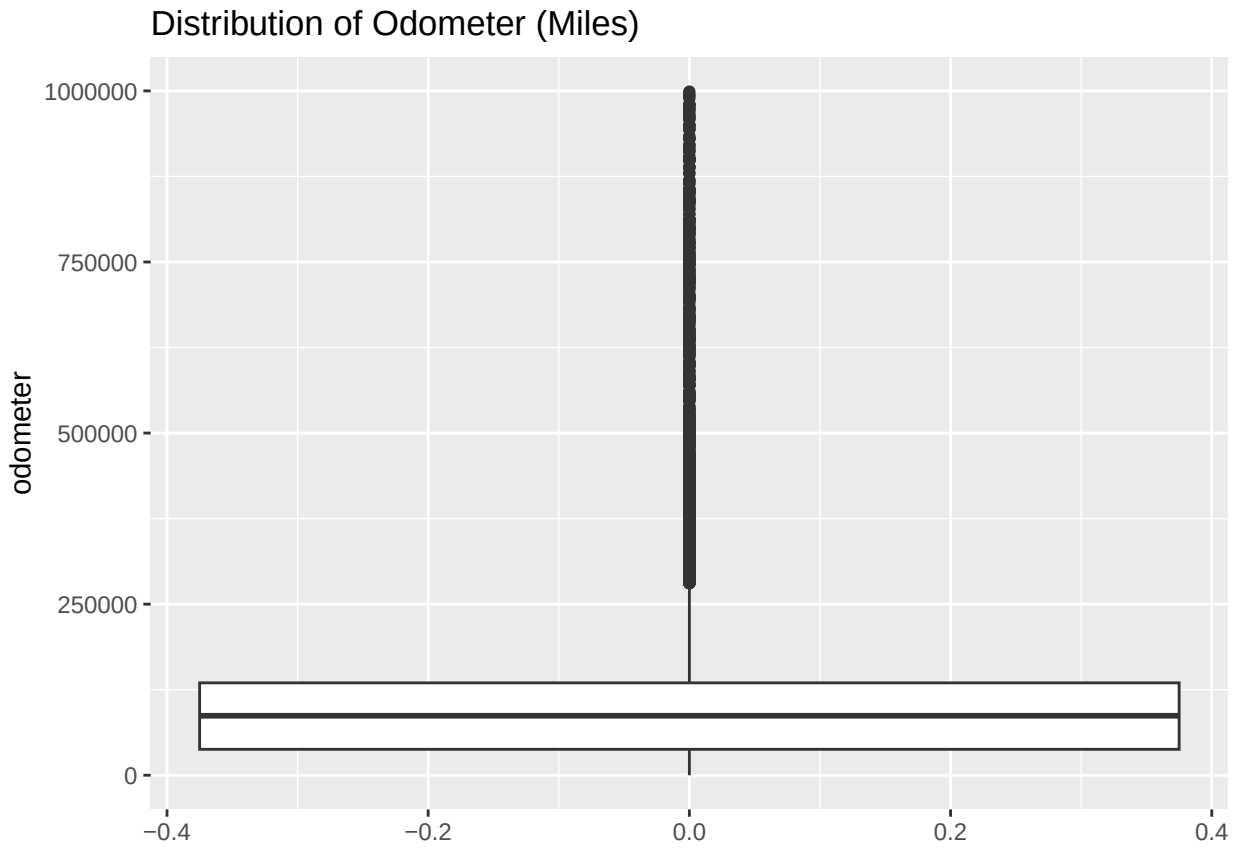
```
ggplot(data = data1, aes(y = price)) + geom_boxplot() +
  ggtitle("Distribution of Price ($)")
```



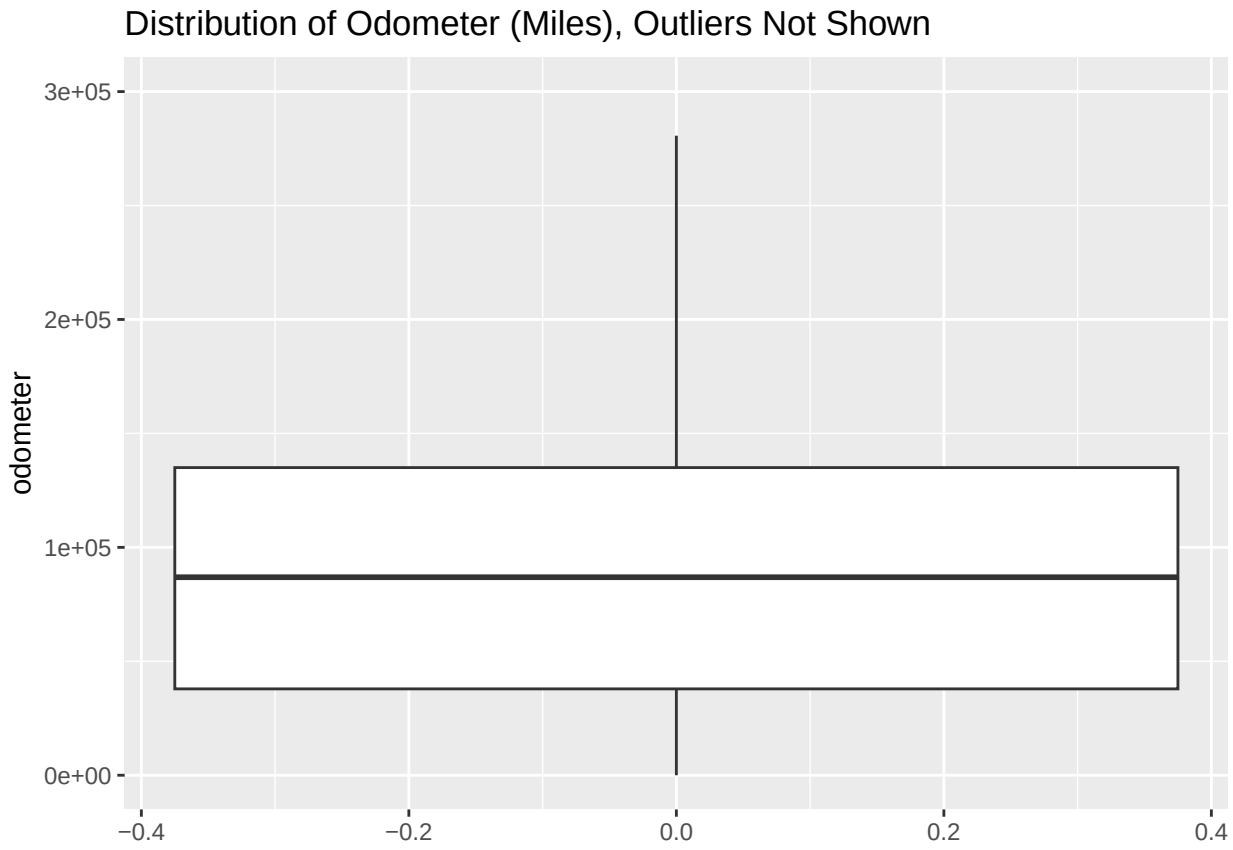

```
ggplot(data = data1, aes(y = price)) + geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(0, 100000))+
  ggtitle("Distribution of Price ($), Outliers Not Shown")
```

Distribution of odometer, boxplots with and without outliers

```
ggplot(data = data1, aes(y = odometer)) + geom_boxplot() +
  ggtitle("Distribution of Odometer (Miles)")
```

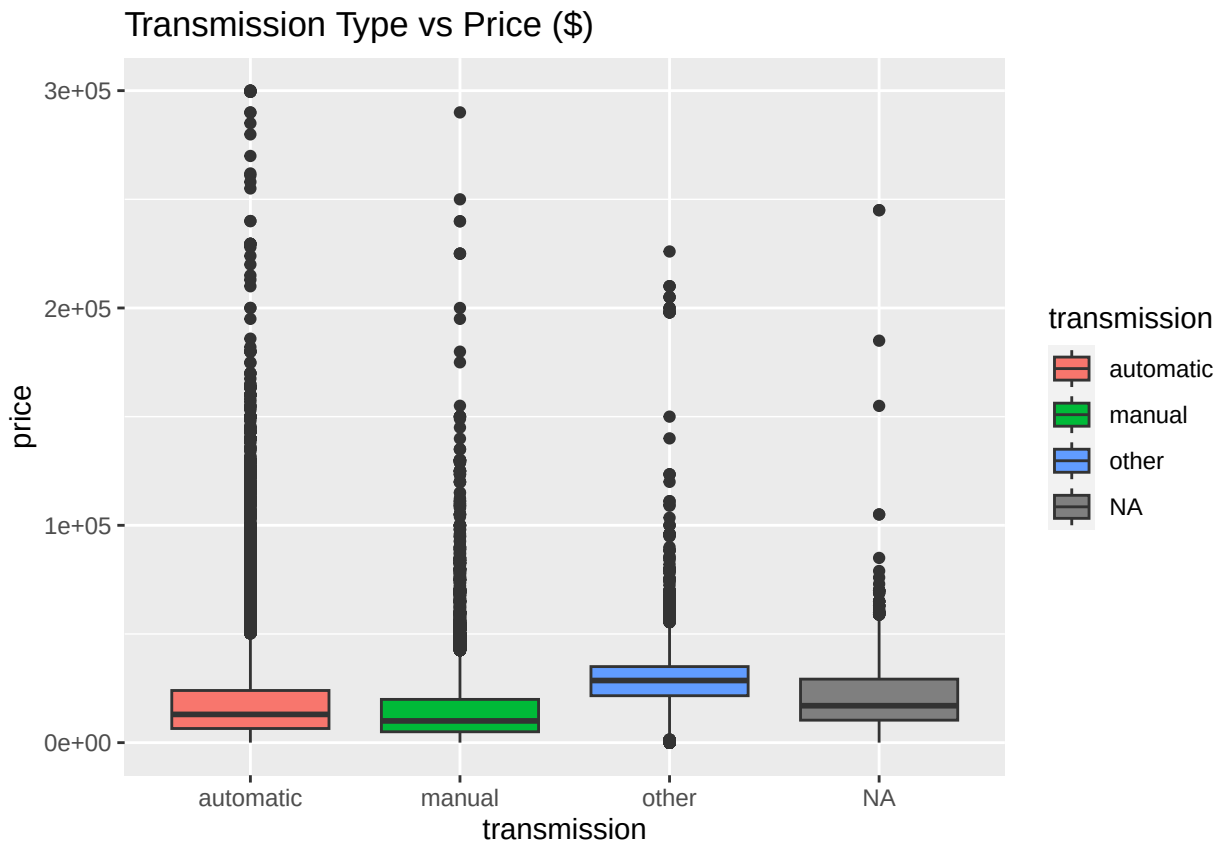



```
ggplot(data = data1, aes(y = odometer)) + geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(0, 300000))+
  ggtitle("Distribution of Odometer (Miles), Outliers Not Shown")
```



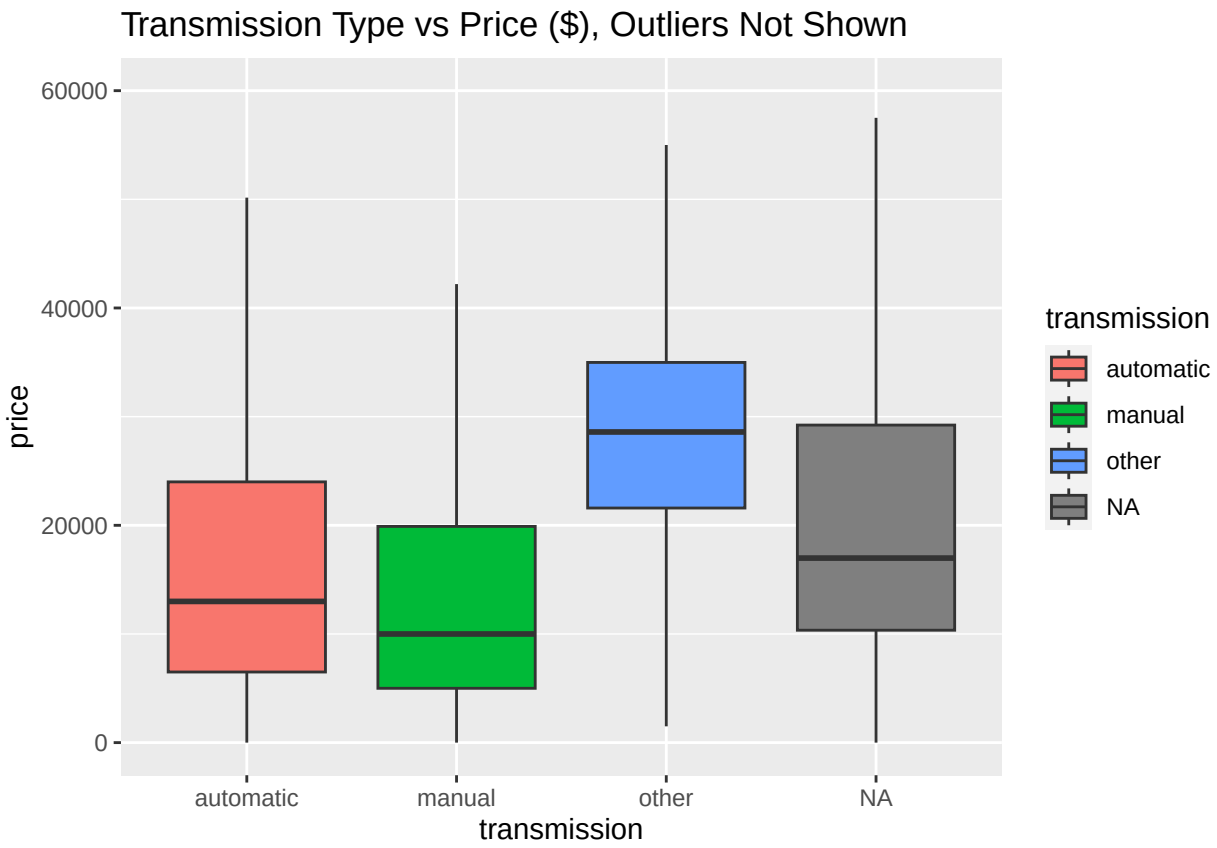
Price vs transmission type with outliers boxplot

```
ggplot(data1, aes(x = transmission, y = price, fill = transmission))+
  geom_boxplot() + ggtitle("Transmission Type vs Price ($)")
```



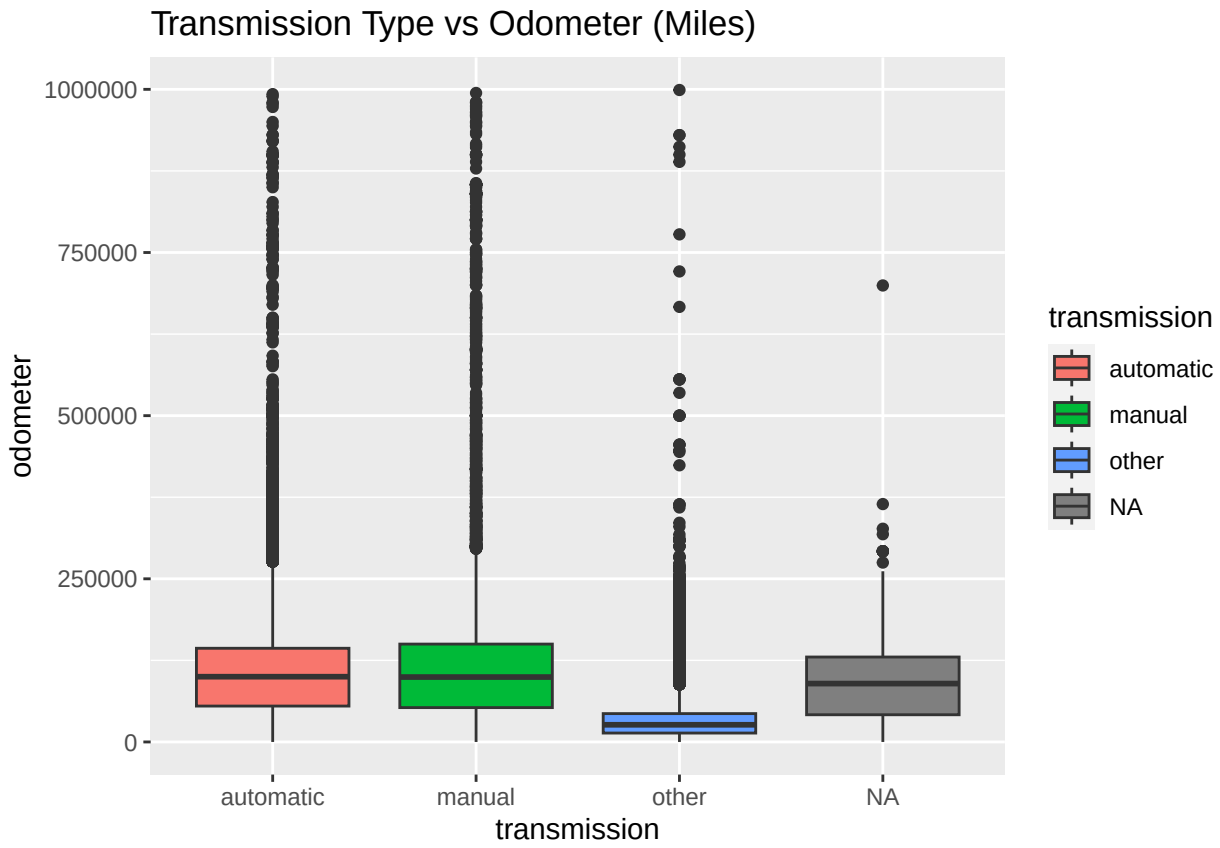
Price vs transmission type with no outliers

```
ggplot(data1, aes(x = transmission, y = price, fill = transmission))+
  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(0, 60000))+
  ggtitle("Transmission Type vs Price ($), Outliers Not Shown")
```



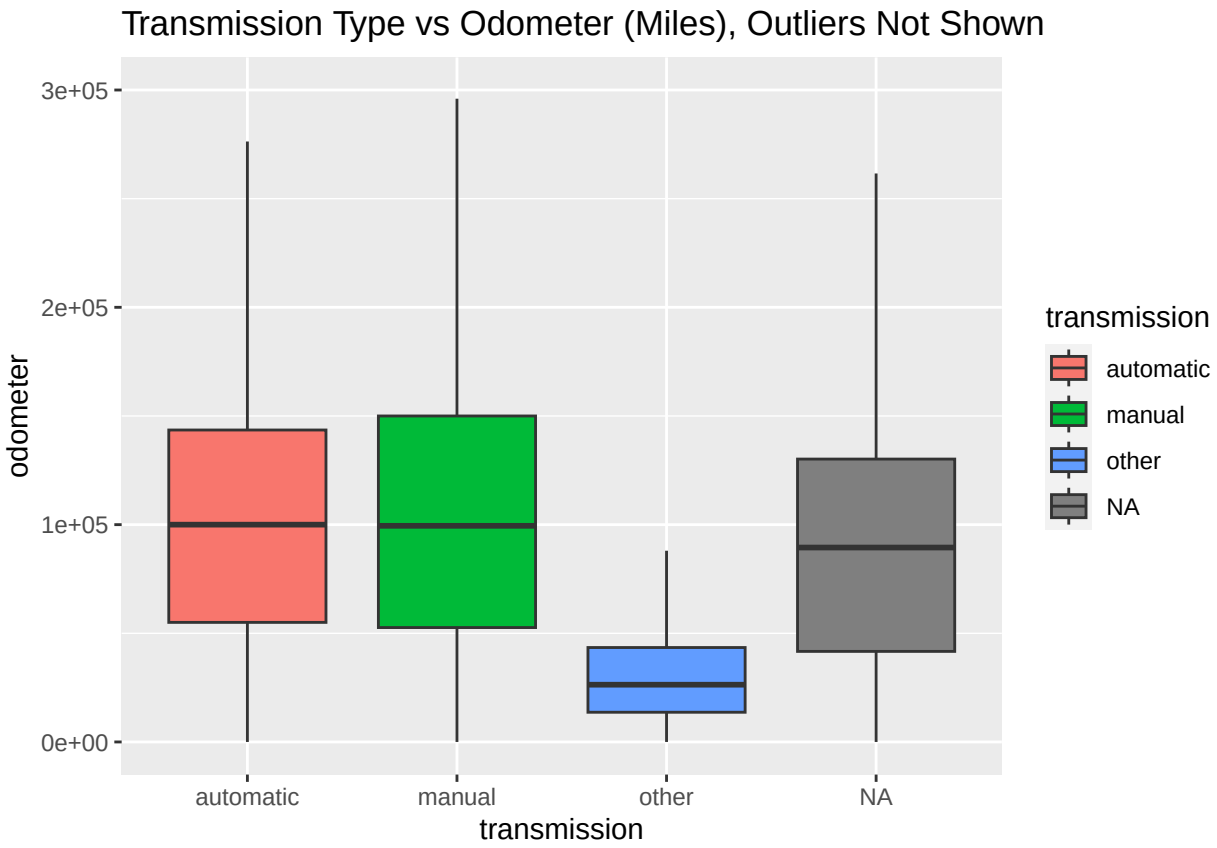
Odometer vs transmission type with outliers boxplot

```
ggplot(data1, aes(x = transmission, y = odometer, fill = transmission))+
  geom_boxplot() + ggtitle("Transmission Type vs Odometer (Miles)")
```



Odometer vs transmission type with no outliers

```
ggplot(data1, aes(x = transmission, y = odometer, fill = transmission))+
  geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = c(0, 300000))+
  ggtitle("Transmission Type vs Odometer (Miles), Outliers Not Shown")
```



Price vs odometer Plot is exclusively in the saved images folder to mitigate strain on R studio

```
scatter <- ggplot(data1, aes(x = price, y = odometer)) + geom_point() +
  ggtitle("Price ($) vs Odometer (Miles)")
```