Cleaning Messy Data Lab

Ben Goldstone

10/2/2023

Part 1

Importing Dataset

```
# Read in Dataset
CarPrices = read_csv("~/CSVs/CarPrices.csv")
```

```
## Rows: 50 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Car, Price
## dbl (1): Mileage
```

```
##
## 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.
```

```
head(CarPrices)
```

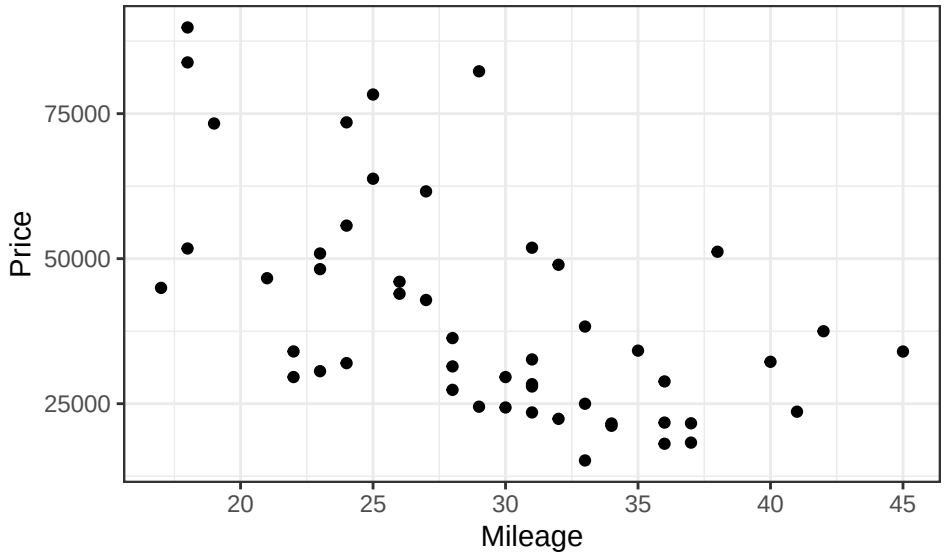
```
## # A tibble: 6 x 3
##   Car              Mileage Price
##   <chr>              <dbl> <chr>
## 1 Ford Fusion           41 $23,625.00
## 2 Ford Focus            36 $18,100.00
## 3 Ford Edge             30 $29,595.00
## 4 Ford Flex             23 $30,605.00
## 5 Ford Explorer         24 $31,995.00
## 6 Ford Expedition       21 $46,630.00
```

Format Price as Numeric

```
CarPrices$Price = parse_number(CarPrices$Price)
head(CarPrices)
```

```
## # A tibble: 6 x 3
##   Car              Mileage Price
##   <chr>              <dbl> <dbl>
## 1 Ford Fusion           41 23625
## 2 Ford Focus            36 18100
## 3 Ford Edge             30 29595
## 4 Ford Flex             23 30605
## 5 Ford Explorer         24 31995
## 6 Ford Expedition       21 46630
```

```
gf_point(Price~Mileage,data=CarPrices)
```

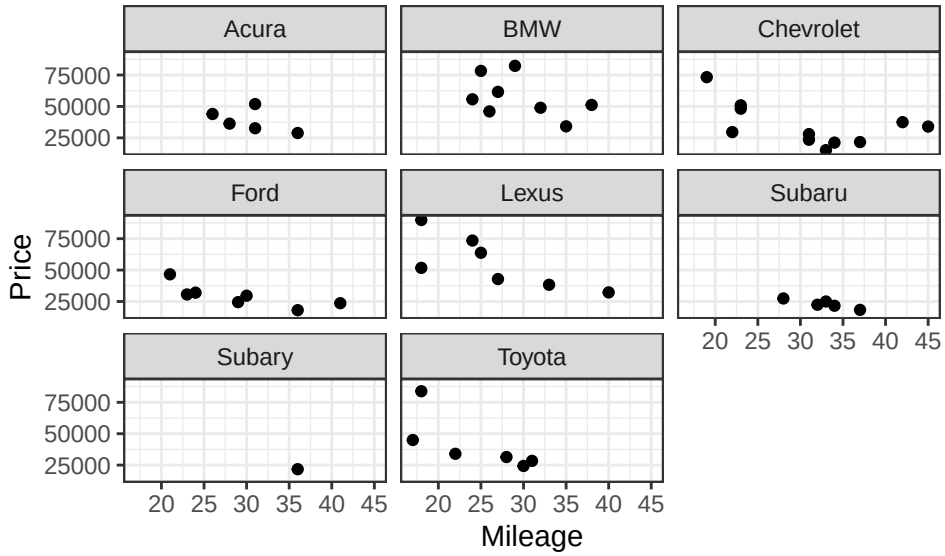


Separate Car into Make and Model

```
CarPrices = CarPrices %>% separate(Car,c("Make","Model"),extra="merge",fill="left")
head(CarPrices)
```

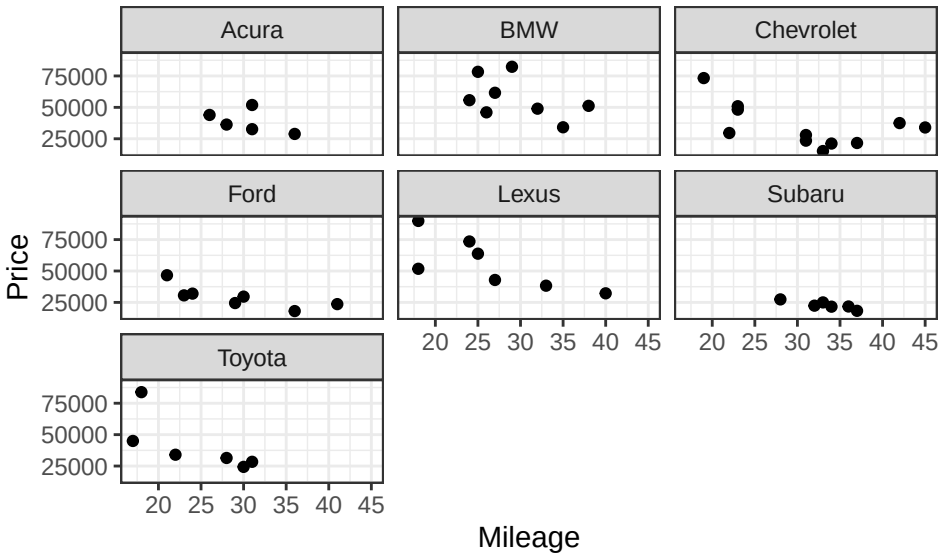
```
## # A tibble: 6 x 4
##   Make  Model      Mileage Price
##   <chr> <chr>        <dbl> <dbl>
## 1 Ford  Fusion          41 23625
## 2 Ford  Focus           36 18100
## 3 Ford  Edge            30 29595
## 4 Ford  Flex            23 30605
## 5 Ford  Explorer        24 31995
## 6 Ford  Expedition      21 46630
```

```
gf_point(Price~Mileage|Make,data=CarPrices)
```



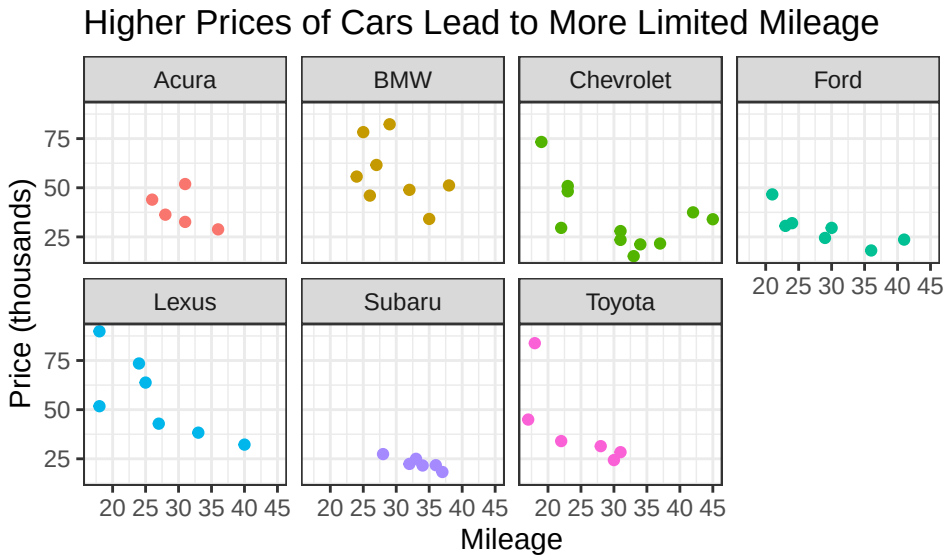
Replace Subary with Subaru

```
CarPrices$Make = str_replace(CarPrices$Make, "Subary","Subaru")
gf_point(Price~Mileage|Make,data=CarPrices)
```



Refining the Scatter Plot

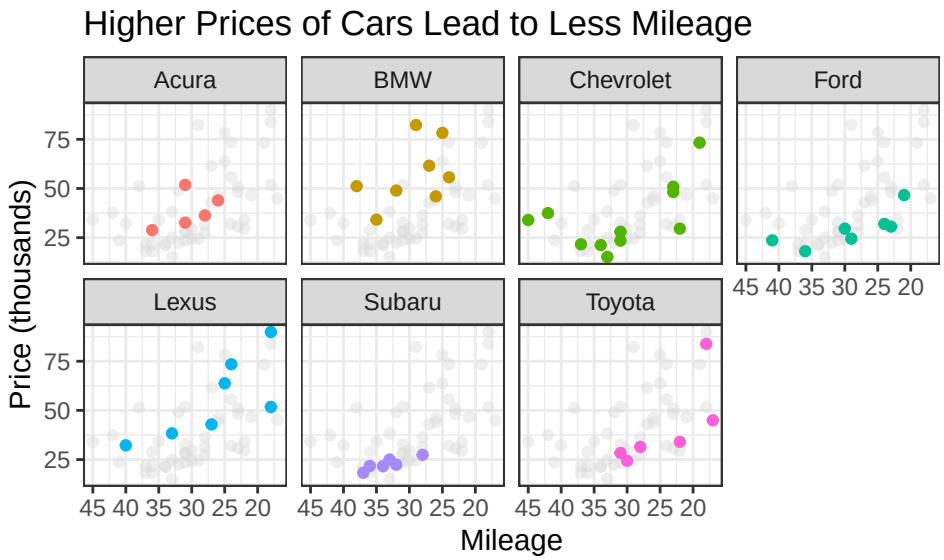
```
gf_point(Price/1000~Mileage,data=CarPrices,color=~Make) %>%
  gf_facet_wrap(~Make,nrow=2) %>%
  gf_theme(legend.position="none") %>%
  gf_labs(title="Higher Prices of Cars Lead to More Limited Mileage", y="Price (thousands)")
```



Add Other Data Points Refinement

```
gf_point(Price/1000~Mileage,data=transform(CarPrices,Make=NULL),color="grey85",alpha=0.4)%>%
  gf_point(Price/1000~Mileage,data=CarPrices,color=~Make) %>% gf_facet_wrap(~Make,nrow=2) %>%
  gf_theme(legend.position="none") %>%
```

```
  gf_labs(title="Higher Prices of Cars Lead to Less Mileage", y="Price (thousands)") %>% gf_refine(scal
```



Part 2

Importing and Exploring Dataset

```
Music = read_csv("~/CSVs/Music.csv")
```

```
## Rows: 52 Columns: 7
```

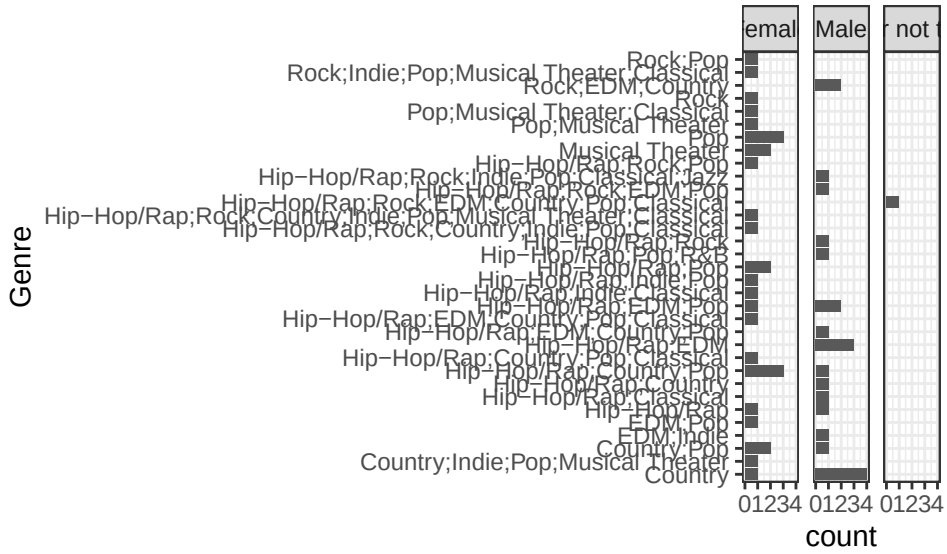
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Year, Gender, Days, Genre, Platform, Where
## dbl (1): Hours
```

```
##
## 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.
```

```
head(Music)
```

```
## # A tibble: 6 x 7
##   Year      Gender Days   Hours Genre                   Platform          Where
##   <chr>     <chr>  <chr>  <dbl> <chr>                   <chr>             <chr>
## 1 Junior    Female 7 days     5 Hip-Hop/Rap;Country;Pop Spotify;Youtube   Car
## 2 Junior    Male   7 days     4 Hip-Hop/Rap;EDM         Soundcloud;Youtu~ Room
## 3 Junior    Female 7 days     5 Hip-Hop/Rap;EDM;Pop     Spotify           Gym
## 4 Sophomore Male   5 days     2 Hip-Hop/Rap;Rock        Spotify;Soundclo~ Car
## 5 Freshman  Male   4 days     3 EDM;Indie               Soundcloud        Walki~
## 6 Senior    Male   7 days     2 Rock;EDM;Country        Apple Music;Yout~ Room
```

```
gf_barh(~Genre|Gender,data=Music)
```



Format Genre Column

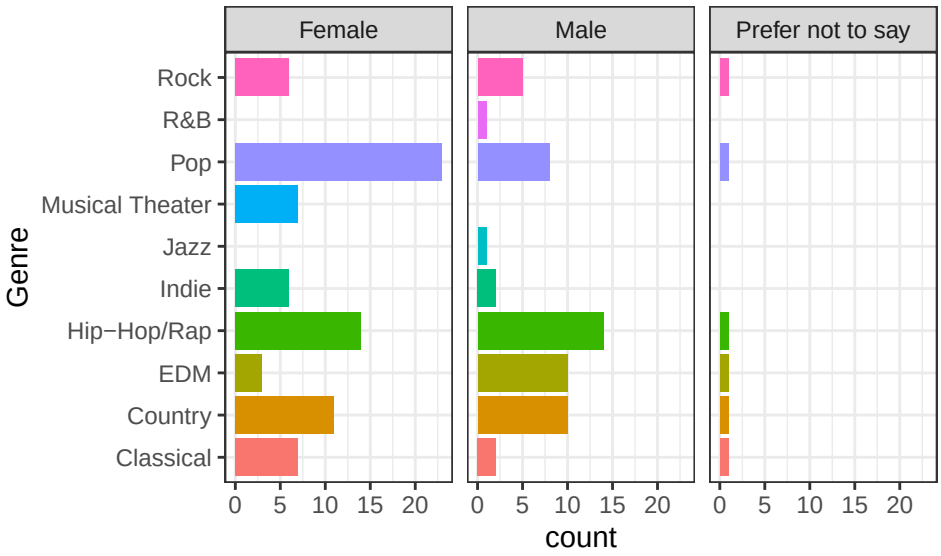
```
NewMusic = Music %>%
  # Separates Genre separated by a semicolon.
  separate(Genre, c("A1","A2","A3","A4","A5","A6","A7"),sep=";") %>%
  # Collapses A1:A7 into their own rows.
  gather("DummyGenre","Genre",5:11) %>%
  # Removes NA values.
  na.omit() %>%
  # Removes Platform and Dummy Genre.
  select(-c("Platform","DummyGenre"))
```

```
## Warning: Expected 7 pieces. Missing pieces filled with `NA` in 51 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
head(NewMusic)
```

```
## # A tibble: 6 x 6
##   Year      Gender Days   Hours Where   Genre
##   <chr>     <chr>  <chr>  <dbl> <chr>   <chr>
## 1 Junior    Female 7 days     5 Car     Hip-Hop/Rap
## 2 Junior    Male   7 days     4 Room    Hip-Hop/Rap
## 3 Junior    Female 7 days     5 Gym     Hip-Hop/Rap
## 4 Sophomore Male   5 days     2 Car     Hip-Hop/Rap
## 5 Freshman  Male   4 days     3 Walking EDM
## 6 Senior    Male   7 days     2 Room    Rock
```

```
gf_barh(~Genre|Gender,data=NewMusic,fill=~Genre) %>% gf_theme(legend.position="none")
```


Female
Male
Prefer not to say
Genre
Rock
R&B
Pop
Musical Theater
Jazz
Indie
Hip−Hop/Rap
EDM
Country
Classical
0
5
10
15
20
count