Assignment 9 - Answer Key

For this assignment, you'll be using the lemons dataset, which is a subset of the dataset used for a Kaggle competition described here: https://www.kaggle.com/c/DontGetKicked/data. Complete the following steps:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(knitr)
library(modelr)
```

```
df<-read_csv("training.csv",n_max=1e6)
```

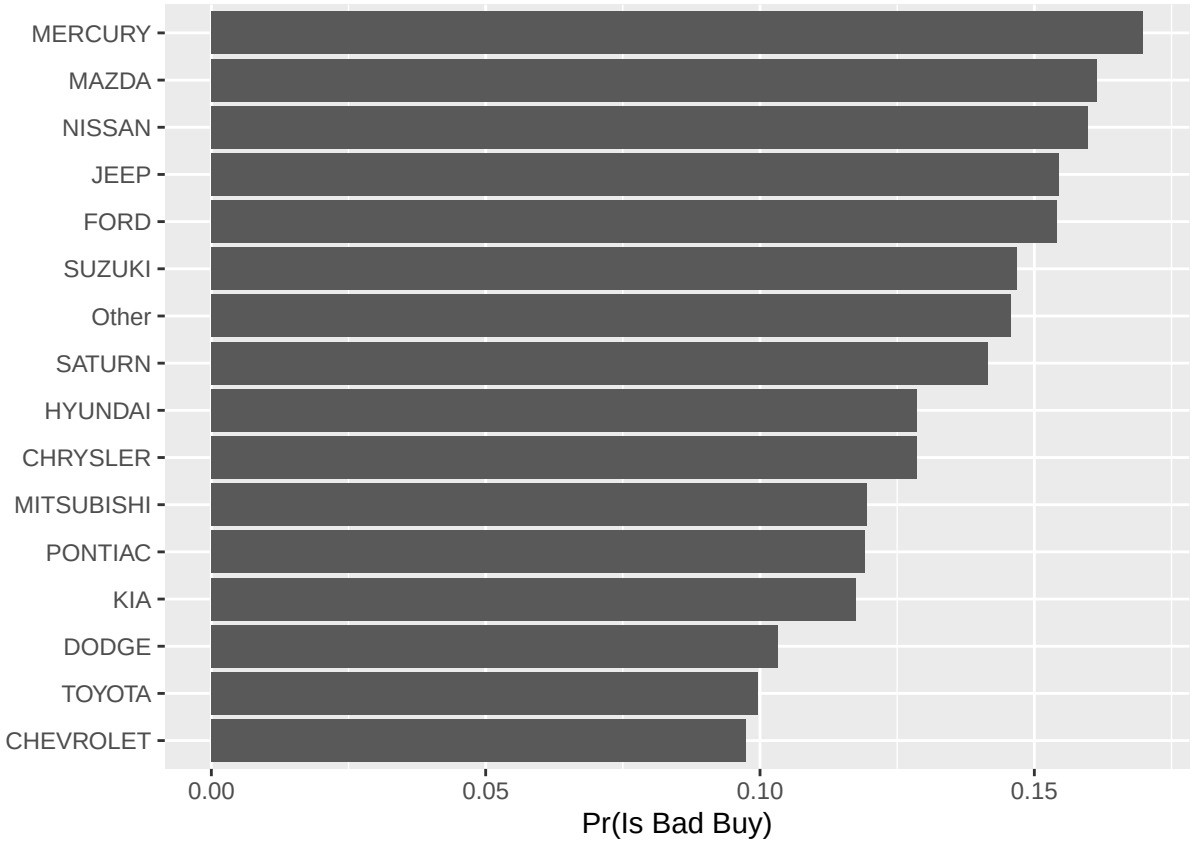
```
## Rows: 72983 Columns: 34

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (24): PurchDate, Auction, Make, Model, Trim, SubModel, Color, Transmissi...
## dbl (10): RefId, IsBadBuy, VehYear, VehicleAge, VehOdo, BYRNO, VNZIP1, VehBC...

##
## 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.
```

1. Using the lemons dataset, plot the probability of a car being a bad buy by make.

```
df%>%
  mutate(Make=fct_lump_prop(Make,prop=.01,other_level = "Other"))%>%
  group_by(Make)%>%
  summarize(mean_isbadbuy=mean(IsBadBuy,na.rm=TRUE))%>%
  mutate(Make=fct_reorder(Make,mean_isbadbuy))%>%
  ggplot(aes(y=mean_isbadbuy ,x=Make) )+
  geom_col()+
  coord_flip()+
  ylab("Pr(Is Bad Buy)")+
  xlab("")
```



2. Create a table that shows the probability of a car being a bad buy by make.

```
df%>%
  mutate(Make=fct_lump_prop(Make,prop=.01,other_level = "Other"))%>%
  group_by(Make)%>%
  summarize(mean_isbadbuy=mean(IsBadBuy,na.rm=TRUE))%>%
  arrange(-mean_isbadbuy)%>%
  mutate(mean_isbadbuy=round(mean_isbadbuy,2))%>%
  rename(`Pr(Bad Buy)`=mean_isbadbuy)%>%
  kable()
```

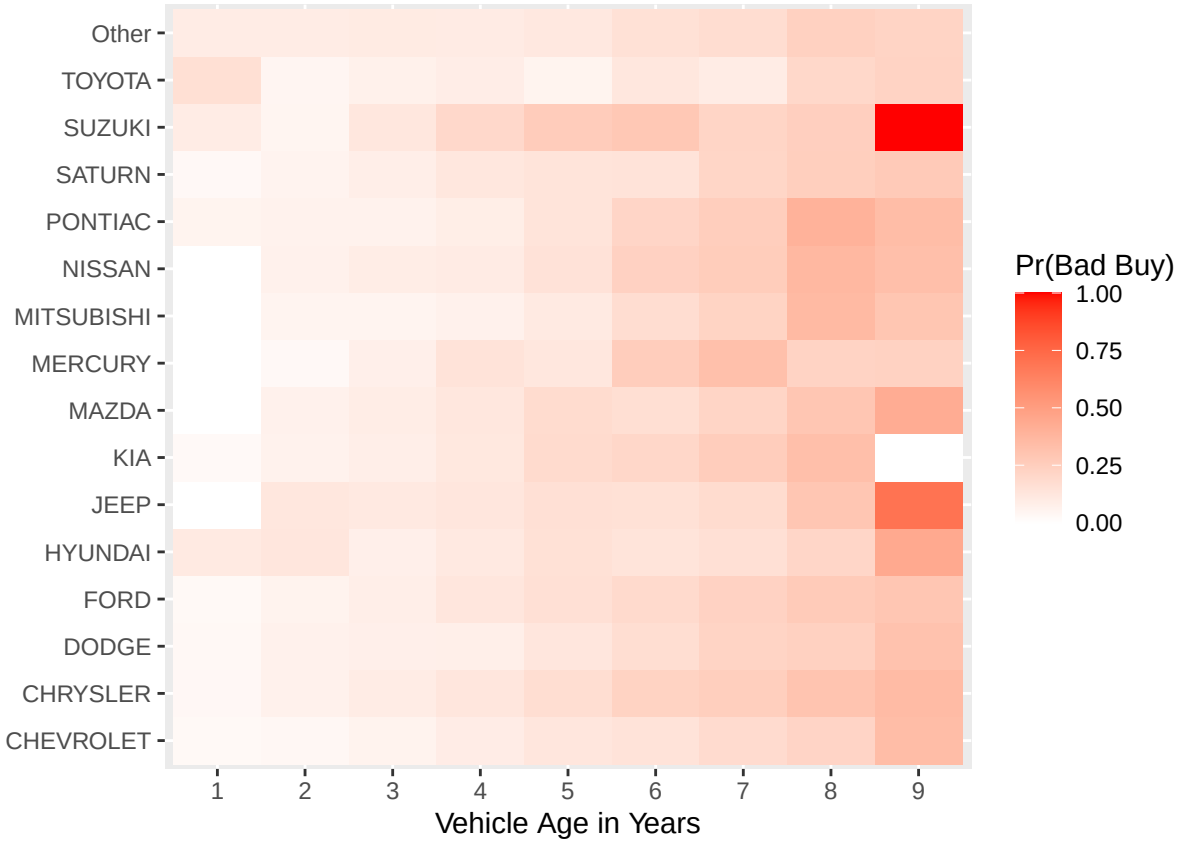
Make	Pr(Bad Buy)
MERCURY	0.17
MAZDA	0.16
NISSAN	0.16
JEEP	0.15
FORD	0.15
SUZUKI	0.15
Other	0.15
SATURN	0.14
HYUNDAI	0.13
CHRYSLER	0.13
MITSUBISHI	0.12
PONTIAC	0.12

Make	Pr(Bad Buy)
KIA	0.12
DODGE	0.10
TOYOTA	0.10
CHEVROLET	0.10

3. Create a heatmap of the probability of a car being a bad buy by make and vehicle age.

```
df %>%
  mutate(Make = fct_lump_prop(Make, prop = .01, other_level = "Other"))%>%
  group_by(Make, VehicleAge) %>%
  summarize(mean_isbadbuy = mean(IsBadBuy, na.rm = TRUE)) %>%
  rename(`Pr(Bad Buy)` = mean_isbadbuy) %>%
  drop_na() %>%
  filter(VehicleAge > 0) %>%
  ggplot(aes(
    y = as.factor(Make),
    x = as.factor(VehicleAge),
    fill = `Pr(Bad Buy)`
  )) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  xlab("Vehicle Age in Years") + ylab("")
```

```
## `summarise()` has grouped output by 'Make'. You can override using the `.groups` argument.
```



4. Create a plot of your choosing that shows the probability of a car being a bad buy by year and make.

```
df%>%
  mutate(Make=fct_lump_prop(Make,prop=.01,other_level = "Other"))%>%
  group_by(Make,VehYear)%>%
  summarize(mean_isbadbuy=mean(IsBadBuy,na.rm=TRUE))%>%
  mutate(Make=fct_reorder(Make,mean_isbadbuy))%>%
  rename(`Pr(Bad Buy)`=mean_isbadbuy)%>%
  drop_na()%>%
  filter(VehYear!=2010)%>%
  ggplot(aes(x=VehYear,y=`Pr(Bad Buy)`,color=Make))+
  geom_point()+
  facet_wrap(~Make)+
  theme(legend.position = "none")+xlab("")
```

```
## `summarise()` has grouped output by 'Make'. You can override using the `.groups` argument.
```