BAN 502 Project

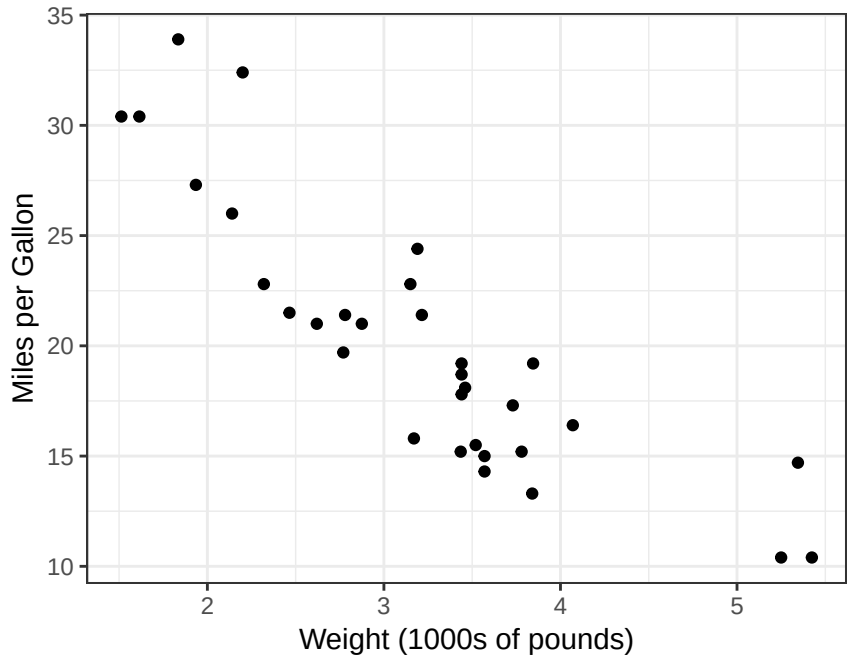
Part 1 - Data Exploration, Preparation, and Visualization

The objective of this project is to predict rain based upon historical weather data. Weather prediction can be critical to business success particular in industries like event planning, transportation, etc. This a binary classification problem with a response variable, "RainTomorrow" with levels "No" and "Yes". You should seek to develop models that maximize predictive accuracy. Naive accuracy (acheived by predicting that all rows will be in the majority class) for this dataset is 0.78. Note that the naive accuracy is likely to be different after your training/testing split.

Part 1 Deliverables There are two deliverables for Part 1 of the project:
1. A short (less than 2 pages) write-up describing your work on this deliverable.
2. Your R Markdown file with your work.

Part 1 Objectives
The objectives for Part 1 are as follows:
1. Gain understanding/intuition regarding the data and the variables.
2. Identify missingness and appropriately deal with missingness in the data.
3. Visualize the relationships between the variables and the response variable (RainTomorrow). Identify variables that may be strong predictors of the response. You may wish to include one or two charts in you written (non-R Markdown) deliverable. See the R code below for an example of how to save ggplot output to an image file:

```
options(tidyverse.quiet = TRUE)
library(tidyverse)
#Create a basic ggplot graph
ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point() +
  theme_bw() + labs(x = "Weight (1000s of pounds)", y = "Miles per Gallon")
```




```
ggsave("demo.png") #save the ggplot graph as a PNG format image file (saves in working directory)
```

```
## Saving 4.5 x 3.5 in image
```

The Data Download the "rain.csv" file from Canvas. The variables in the dataset are described below:

- Date: The date of observation.
- Location: The common name of the location of the weather station.
- MinTemp: The minimum temperature in degrees celsius.
- MaxTemp: The maximum temperature in degrees celsius.
- Rainfall: The amount of rainfall recorded for the day in mm.
- WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight.
- WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight.
- WindDir9am: Direction of the wind at 9am.
- WindDir3pm: Direction of the wind at 3pm.
- WindSpeed9am: Wind speed (km/hr) averaged over 10 minutes prior to 9am.
- WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 3pm.
- Humidity9am: Humidity (percent) at 9am.
- Humidity3pm: Humidity (percent) at 3pm.
- Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am.
- Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm.
- Cloud9am: Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky while an 8 indicates that it is completely overcast.
- Cloud3pm: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values.
- Temp9am: Temperature (degrees C) at 9am.
- Temp3pm: Temperature (degrees C) at 3pm.
- RainToday: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.
- RainTomorrow: The response variable. Did it rain tomorrow?