Homework #1

Dustin Leatherman

9/13/2020

```
knitr::opts_chunk$set(echo = TRUE, fig.width = 10, warning = FALSE, message = FALSE)
library(tidyverse)
library(knitr)
library(kableExtra)
library(grid)
library(gridExtra)
library(broom)
library(ggfortify)

ozone <- read.csv("~/Downloads/ozone.csv")
```

1

Create a table with the overall (across sites and days) mean, standard deviation, and percent missing.

```
ozone_pivot <-
  ozone %>%
  # Pivot Day columns to Rows
  pivot_longer(starts_with("Day"), names_to = "day", values_to = "value") %>%
  mutate(
    # convert day to integer for easier processing
    day = as.integer(str_replace(day, "Day.", ""))
  )

ozone_pivot %>%
  summarise(
    mean = mean(value, na.rm = TRUE),
    sd = sd(value, na.rm = TRUE),
    perc_na = sum(is.na(value)) / n()
  ) %>% kable(
    caption = "Overall Summary Statistics"
  ) %>% kable_styling(bootstrap_options = "striped", latex_options = "hold_position")
```

Table 1: Overall Summary Statistics

mean	sd	perc_na
51.27333	17.26207	0.0432246

2

a. Compute the mean, variance, and percent missing for each of the n sites;
b. Make a histogram of each variable (all three histograms should have n observations);

c. create scatter plots of each pair of these variables (each of the three plots should have n points)

```
ozone_summary <-
  ozone_pivot %>%
  group_by(Station.ID) %>%
  summarise(
    mean = mean(value, na.rm = TRUE),
    sd = sd(value, na.rm = TRUE),
    perc_na = sum(is.na(value)) / n(),
    var = var(value, na.rm = TRUE)
  )

ozone_summary %>%
  select(-var) %>%
  # There are 1000+ stations. Limit for viewing purposes
  head(10) %>%
  kable(
    caption = "A sample of Station Observation Summaries"
  ) %>% kable_styling(bootstrap_options = "striped", latex_options = "hold_position")
```

Table 2: A sample of Station Observation Summaries

Station.ID	mean	sd	perc_na
1	39.55608	16.93143	0.1612903
2	40.67765	11.44378	0.0000000
3	44.82777	14.17176	0.0000000
4	38.78773	11.18569	0.0322581
5	41.24200	12.40991	0.0645161
6	42.80556	13.98266	0.1290323
7	41.88306	16.61801	0.0000000
8	44.81452	19.68811	0.0000000
9	41.15229	12.93453	0.0000000
10	43.51293	17.36322	0.0645161

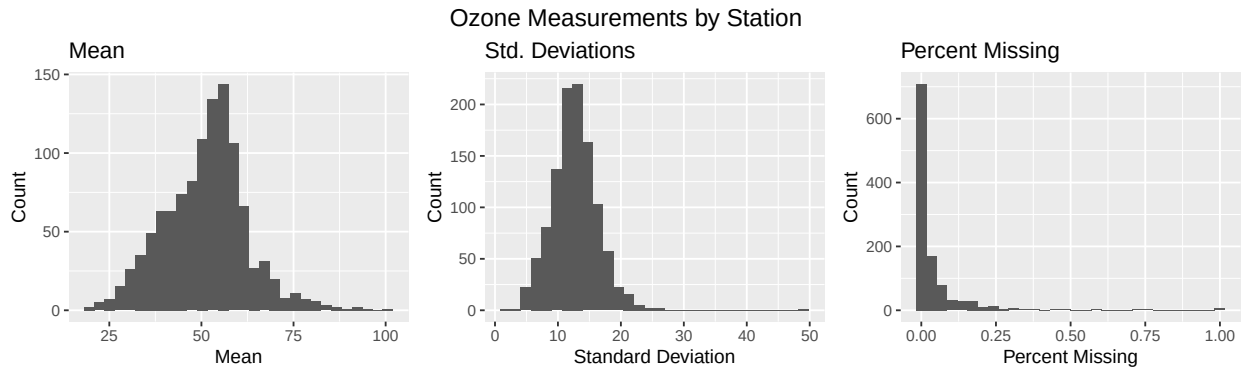
```
plot_mean <-
  ozone_summary %>%
  ggplot(aes(x = mean)) +
    geom_histogram() +
    labs(x = "Mean", y = "Count", title = "Mean")

plot_sd <-
  ozone_summary %>%
  ggplot(aes(x = sd)) +
    geom_histogram() +
    labs(x = "Standard Deviation", y = "Count", title = "Std. Deviations")

plot_percmis <-
  ozone_summary %>%
  ggplot(aes(x = perc_na)) +
    geom_histogram() +
    labs(x = "Percent Missing", y = "Count", title = "Percent Missing")

grid.arrange(plot_mean, plot_sd, plot_percmis, ncol = 3,  top = textGrob("Ozone Measurements by Station
```

```
                          gp=gpar(fontsize=14,font=1),just=c("center")))
```

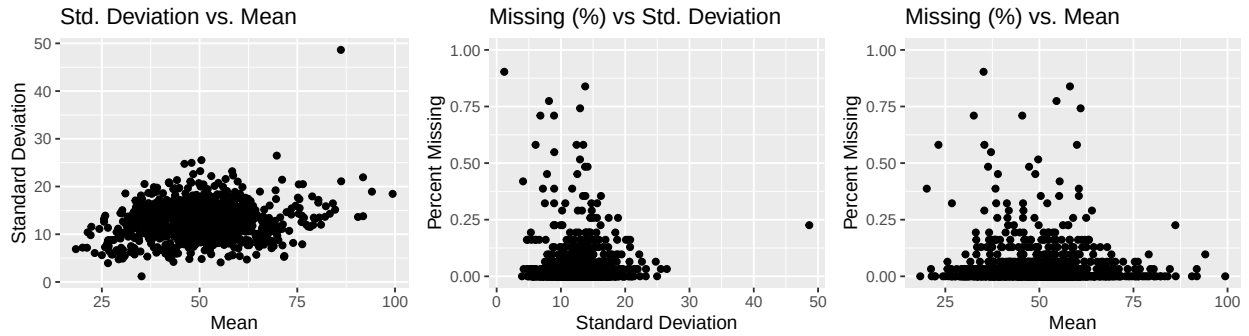



```
plot_mean <-
  ozone_summary %>%
  ggplot(aes(x = mean, y = sd)) +
    geom_point() +
    labs(x = "Mean", y = "Standard Deviation", title = "Std. Deviation vs. Mean")

plot_sd <-
  ozone_summary %>%
  ggplot(aes(x = sd, y = perc_na)) +
    geom_point() +
    labs(x = "Standard Deviation", y = "Percent Missing", title = "Missing (%) vs Std. Deviation")

plot_percmis <-
  ozone_summary %>%
  ggplot(aes(x = mean, y = perc_na)) +
    geom_point() +
    labs(y = "Percent Missing", x = "Mean", title = "Missing (%) vs. Mean")

grid.arrange(
  plot_mean,
  plot_sd,
  plot_percmis,
  ncol = 3,
  top = textGrob("Ozone Measurement Summaries by Station",
                          gp=gpar(fontsize=14,font=1),just=c("center"))
)
```



3

Conduct a linear regression with response equal to the site's mean and the site's variance and percent missing as covariates.

```
model1 <- lm(mean ~ var + perc_na, data = ozone_summary)

tidy(model1) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped", latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	48.5895500	0.6328189	76.782713	0.0000000
var	0.0170994	0.0029618	5.773334	0.0000000
perc_na	-10.5407498	3.6119795	-2.918275	0.0035916

```
autoplot(model1)
```