Assignment 1

Shale

1/19/2022

1 Load data

```
pm_df <- readxl::read_excel(here("HW1/CES4.xlsx"), sheet = "CES4.0FINAL_results") %>%
  clean_names() %>% select(census_tract, total_population, california_county,
                          low_birth_weight, pm2_5, poverty) %>%
  drop_na() %>%
  mutate(low_birth_weight = round(as.numeric(low_birth_weight), 2))

pm_df$california_county <- as.factor(pm_df$california_county)
```

2 Answers

2.0.1 (a) What is the average concentration of PM2.5 across all census tracts in California?

10.1523683 is the mean PM 2.5 concentration across all California census tracts.

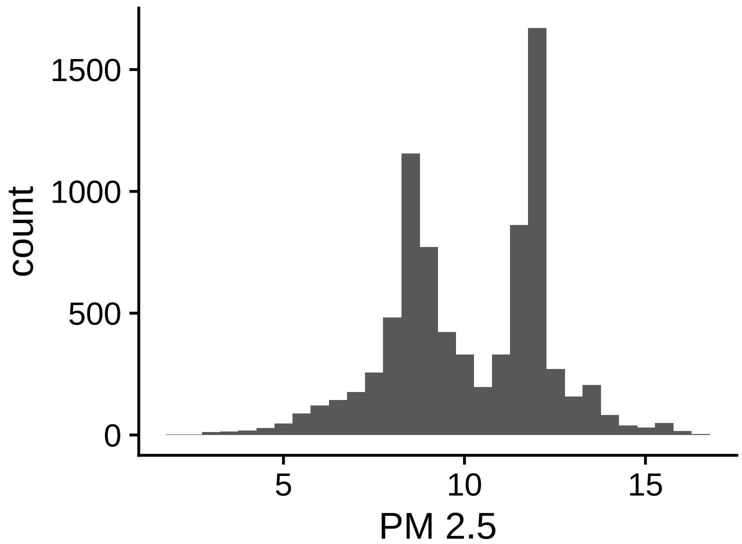
2.0.2 (b) What county has the highest level of poverty in California?

```
county_data <- pm_df %>%  mutate(pov_num = (poverty/100) * total_population) %>%
  select(california_county, total_population, pov_num) %>%
  group_by(california_county) %>% summarise(across(where(is.numeric), sum))  %>%
  mutate(acc_pov = pov_num / total_population)
```

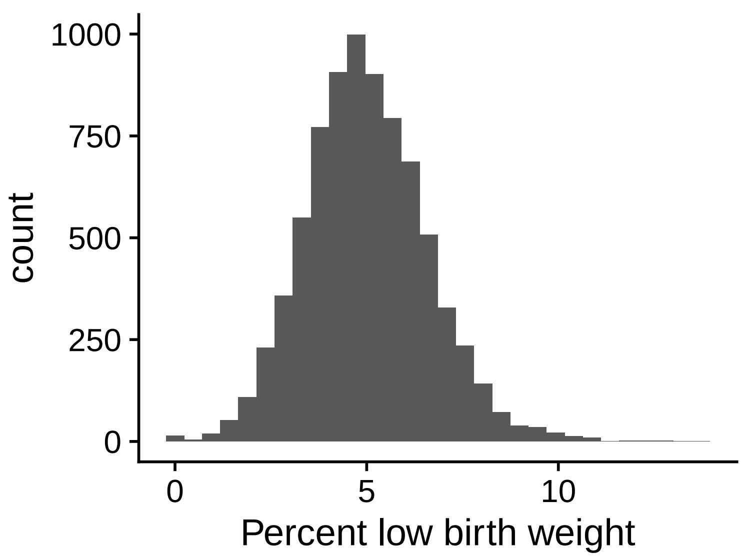
Ventura county has the highest level of poverty in a single census tract (tract 6111003012). However, Tulare county has the highest overall mean poverty level at 50.147635 percent.

2.0.3 (c) Make a histogram depicting the distribution of percent low birth weight and PM2.5.

```
ggplot(pm_df, aes(x=pm2_5)) +
  geom_histogram()+
  theme_cowplot(14)+
  labs(x = "PM 2.5")
```




```
ggplot(pm_df, aes(x=low_birth_weight)) +
  geom_histogram() +
  theme_cowplot(14) +
  labs(x = "Percent low birth weight")
```



Note: PM 2.5 is distinctly bimodal.

2.0.4 (d) Estimate a OLS regression of LowBirthWeight on PM25. Report the estimated slope coefficient and its heteroskedasticity-robust standard error. Interpret the estimated slope coefficient. Is the effect of PM25 on LowBirthWeight statistically significant at the 5%?

```
lbw_model = estimatr::lm_robust(data = pm_df, low_birth_weight ~ pm2_5)
huxtable::huxreg(lbw_model, error_pos = "right")
```

As the table shows, an OLS model of the effect of pm2__5 on LBW gives a pm2__5 coefficient of 0.118, which means that for a 1 microgram per cubic meter (standard units inferred, not provided in metadata) increase in annual pm2__5 a 0.118 percent increase in LBW would be expected. The heteroskedasticity-robust standard error is 0.008. This effect is statistically

	(1)	
(Intercept)	3.800 ***	(0.089)
pm2_5	0.118 ***	(0.008)
N	7805	
R2	0.025	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

significant at the 5% level based on a p<0.001.

Part (e) removed from assignment 1

2.0.5 (f) Add the variable Poverty as an explanatory variable to the regression in (d). Interpret the estimated coefficient on Poverty. What happens to the estimated coefficient on PM25, compared to the regression in (d). Explain.

```
bwp_model = estimatr::lm_robust(data = pm_df, low_birth_weight ~ pm2_5 + poverty)
huxtable::huxreg(lbw_model, bwp_model, error_pos = "right")
```

	(1)		(2)	
(Intercept)	3.800 ***	(0.089)	3.544 ***	(0.085)
pm2_5	0.118 ***	(0.008)	0.059 ***	(0.008)
poverty			0.027 ***	(0.001)
N	7805		7805	
R2	0.025		0.117	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

The table above shows that in the second model (with poverty included) the coefficient for pm2_5 decreases from about .12 to .06; this is because some of the variability in LBW that the first model was attributing to pm2_5 is actually caused by poverty (the higher R2 value in the second model also indicates that including poverty in the model does help explain more of the variability in LBW than pm2_5 alone). Specifically, the poverty coefficient in the second model suggests that for every one percent increase in poverty, an increase of 0.027 percent LBW would be expected.

2.0.6 (g) From the regression in (f), test the null hypothesis that the effect of PM2.5 is equal to the effect of Poverty.

Looking at the confidence intervals (using `summary(bwp_model)`), it is clear that PM 2.5 has a bigger impact on birth weights than poverty does - we can reject the null hypothesis that their effects are the same at the 95% CI. The lower bound of the PM 2.5 CI is 0.04285, while the upper bound of the poverty CI is 0.02940. This means that even if our point estimates for the model coefficients for PM 2.5 and poverty are not exactly right (specifically, if we overestimated the effect of poverty and underestimated the effect of PM 2.5, which would

bring the two estimates closer together), the impact of PM 2.5 will still be greater than the impact of poverty in determining birth weight.

```
hyp = car::linearHypothesis(bwp_model, c("poverty=pm2_5"), white.adjust = "hc2")
```

Another way to consider this is with the `linearHypothesis()` function above, which shows that we can reject the null hypothesis that the effect is the same for poverty and PM 2.5 at the 5% CI with a p value of 0.0002426.