Dani DiTomasso
Professor Eleish
Data Analytics
22 October 2024

Assignment 3 Written Responses

1a. Comparing Summaries of COVID cases and deaths from 2020 to 2021

```
> summary(covid2020.ny.cases.no_outliers$cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    84.0   250.0   575.9   693.0  3919.0
> summary(covid2021.ny.cases.no_outliers$cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    112    3189    5118    7583    9014   35370
> summary(covid2020.ny.deaths.no_outliers$deaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    5.00   18.53   27.00  139.00
> summary(covid2021.ny.deaths.no_outliers$deaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    53.0    85.0   115.5   131.0   551.0
```

When considering the summaries of data on Covid-19 cases and deaths within New York counties in 2020 and 2021, there are some striking and surprising trends that seem to emerge. Firstly, in both the variables of “cases” and “deaths”, all numbers seem to drastically increase from 2020 to 2021, despite the fact that the initial outbreak occurred in the beginning of 2020. Additionally, when comparing reported Covid-19 cases from 2020 to 2021, the mean of the data increased by a total of approximately 1216 percent. Whereas for Covid-19 deaths, the mean increased by approximately 523 percent from 2020 to 2021, which is about half of the percentage increase for cases in New York.

1b. Histograms for the “Cases” and “Deaths” in NY from 2020 and 2021

Upon plotting the data for both “cases” and “deaths” reported in the state of New York in both 2020 and 2021, there are some general similarities but also some important differences. When comparing 2020 to 2021, the data for both Covid cases and deaths in the histogram starts by being extremely skewed to the left, meaning that there was a much higher presence of zero Covid-19 cases or deaths within the year of 2020. Upon shifting to 2021, a more “regular” or “bell-curve” shape starts to take place, with the data points shifting to the right, showing a significant increase in both Covid cases and deaths in 2021.

1c. ECDF’s and QQ-Plots for Covid “Cases” and “Deaths” in NY in 2020 vs 2021

When observing the ECDF diagrams for both “cases” and “deaths” reported in the state of New York in both 2020 and 2021, the relationship that is observed can clearly be seen as an exponential increase. This indicates that from 2020 to 2021, Covid cases and deaths in the state of New York grew at a rapid rate, and then towards the end of 2021, plateaued by staying at a consistent, yet high number. The QQ-Plots for each of the variables show a similar trend, with a lower number in the beginning, only to increase drastically as rnorm increases.

2. Summary of plots for 2a, 2b, and 2c, with data subset being Nassau and Suffolk counties (2 counties found in the state of New York that make up Long Island).

Similarly to the entire state of New York, Long Island saw a significant increase in both Covid cases and deaths from 2020 to 2021. For example, histograms show an extremely high density of case numbers being around 40,000 to 50,000 in 2020, whereas in 2021 the case numbers range increases to 150,000 to 250,000. Furthermore, ECDF plots also show an exponential increase in cases and deaths from 2020 to 2021, reflecting the trends of the entire state of New York, just on a smaller population scale.

3a. Summary of the linear model run for NY housing prices

```
Call:
lm(formula = PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny.house.price.compl)

Residuals:
     Min       1Q   Median       3Q      Max
-4626408  -317281  -128791   181455  2131231

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  264540.961  16786.888   15.76   <2e-16 ***
BEDS          -5029.322   5783.781   -0.87    0.385
BATH         209515.665   9551.450   21.93   <2e-16 ***
PROPERTYSQFT    113.762      9.585   11.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 500700 on 4238 degrees of freedom
Multiple R-squared:  0.3249,   Adjusted R-squared:  0.3245
F-statistic:   680 on 3 and 4238 DF,  p-value: < 2.2e-16
```

Based on the summary run on the linear model, the BATH variable is the one that has the most effect on PRICE. One reason is due to the P-value being 2e-16, which shows a highly significant relationship with price. In contrast however, BEDS has a p-value of 0.385, which is not statistically significant. Furthermore, the t-value for BATH is 21.93, which is much higher than the t-values for BEDS (-0.87) and PROPERTYSQFT (11.87), which shows more significance than both BATH and PROPERTYSQFT. While PROPERTYSQFT also shows a significance on price (due to the p-value) BATH is the most relevant in terms of significance and effect size.

3b. Summary of the linear model run for NY housing data subset where the PRICE < 500,000

```
Call:
lm(formula = PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny.house.subset)

Residuals:
    Min      1Q  Median      3Q     Max
-383218  -70614   -3829   79948  200447

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.397e+05  9.658e+03  24.815  < 2e-16 ***
BEDS         3.574e+03  3.377e+03   1.058   0.2902
BATH         5.025e+04  6.769e+03   7.424  2.1e-13 ***
PROPERTYSQFT 8.849e+00  4.252e+00   2.081   0.0376 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 102100 on 1243 degrees of freedom
Multiple R-squared:  0.0679,   Adjusted R-squared:  0.06565
F-statistic: 30.18 on 3 and 1243 DF,  p-value: < 2.2e-16
```

Based on the summary run on the linear model subset, the BATH variable is still the one that has the most effect on PRICE. One reason is due to the P-value being 2.1e-13, which shows a highly significant relationship with price. Furthermore, the t-value for BATH is 7.424, which is much higher than the t-values for BEDS (1.058) and PROPERTYSQFT (2.081), which shows more significance than both BATH and PROPERTYSQFT. However it is important to note that while PROPERTYSQFT is still also significant, it is much less significant compared to the entire dataset and to the BATH variable.

References

Zach. “How to Remove Outliers in R.” *Statology*, 6 Aug. 2020, www.statology.org/remove-outliers-r/.

---. “The Difference between T-Values and P-Values in Statistics.” *Statology*, 30 Aug. 2021, www.statology.org/t-value-vs-p-value/.