hw5

David Jeong

Packages

```r
# Load packages
library(tidyverse)
library(tidymodels)
library(knitr)
```

Data

```r
# Load Data
money <- read_csv("data.csv")
```

Regression Analysis

```r
# Filter data so that it represents the true amount of money that should have been collect
money_train <- money |>
  filter(BRINK == 0)

# Fit linear model
money_fit <- linear_reg() |>
  set_engine("lm") |>
  fit(CON ~ CITY, data = money_train)

# Neatly display model estimates to 3 digits
tidy(money_fit) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	268913.791	359194.396	0.749	0.462
CITY	204.413	52.627	3.884	0.001

```r
# Predict the amount of money that should have been collected for when time period = 21 (C
new_obs <- tibble(
  CITY = 6613
)

predict(money_fit, new_obs)
```

```
# A tibble: 1 x 1
     .pred
     <dbl>
1 1620699.
```

```r
# Filter data so that it represents money Brink's Inc. should have collected
money_train_2 <- money |>
  filter(BRINK == 1)

# Fit linear model
money_fit_2 <- linear_reg() |>
  set_engine("lm") |>
  fit(CON ~ CITY, data = money_train_2)

# Neatly display model estimates to 3 digits
tidy(money_fit_2) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	500459.417	345740.309	1.448	0.162
CITY	156.892	49.676	3.158	0.005

```r
# Augment data
money_Brink <- augment(money_fit_2$fit)

# Estimate total amount of money that Brink's Inc. should have collected
money_Brink |>
```

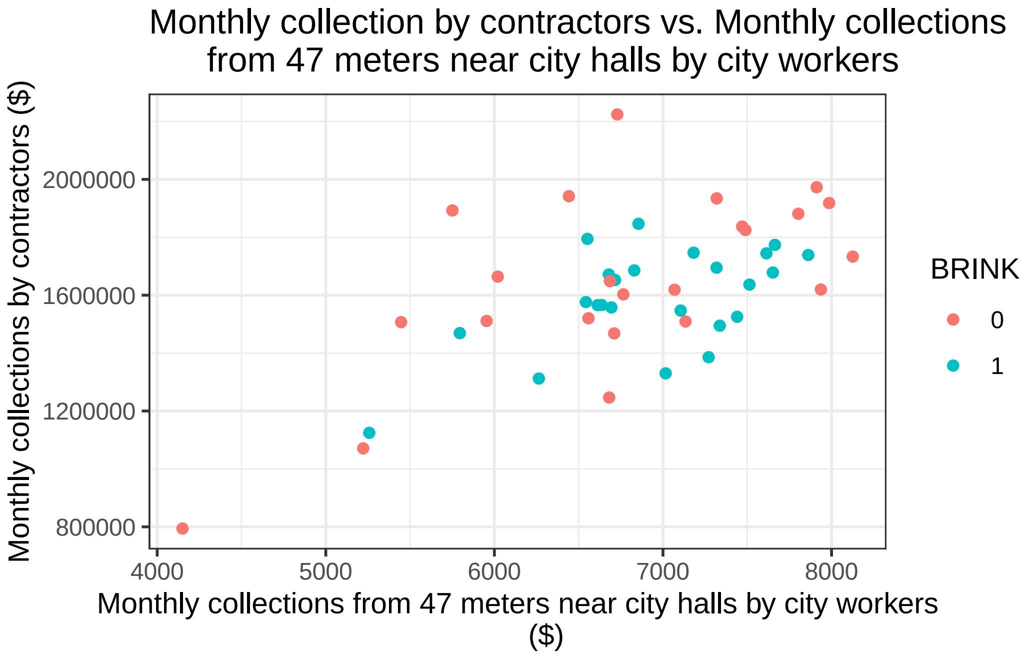
```
  summarize(total_collect = sum(.fitted))
```

```
# A tibble: 1 x 1
  total_collect
          <dbl>
1     38119433.
```

Data Visualization

```
# Make BRINK categorical
money <- money |>
  mutate(BRINK = factor(BRINK))

# Visualize data between CITY and CON
ggplot(money, aes(x = CITY, y = CON, color = BRINK)) +
  geom_point() +
  labs(title = "Monthly collection by contractors vs. Monthly collections
      from 47 meters near city halls by city workers",
       x = "Monthly collections from 47 meters near city halls by city workers
       ($)",
       y = "Monthly collections by contractors ($)") +
  theme_bw()
```



We see a clear outlier at the very lower-left corner of the scatter plot shown, in which monthly collections by city workers is near \$4000 (to be exact, \$4150) and the monthly collections by contractors is \$794191.