Correlation Check - Sharing Results

Abraham Cheung

2023-03-31

Summary

This file assesses the correlation between the three variables in our data: penalties, trips, and deployment. We find that there is a significant and relatively strong correlation in between penalties and trips; and penalties and deployment (p < 0.001). We also find a significant and very strong correlation between trips and deployment (p < 0.001). Because of the strong correlation, we can state with more certainty in our APP report that penalties happen in areas with more trips; penalties occur in areas with more deployment; more trips happen in places that have greater deployment. We also conclude that due to the strength of the correlation, we might not want to conduct a linear regression because the possible issue of collinearity.

Set up

```
pacman::p_load("tidyverse", "readxl", "sf")

# file path objects
data_files_dir <- file.path("..", "output", "files")
```

Load Data

The data used for the correlation check has already been cleaned and transformed. In summary, we limited observations to March 2019 to September 2022 since that time frame intersects for all three data sets (penalty, trip, and average deployment data). The data frame shows the number of penalties, number of trips (by origin and by destination), and average deployment for each month for each Neighborhood Council.

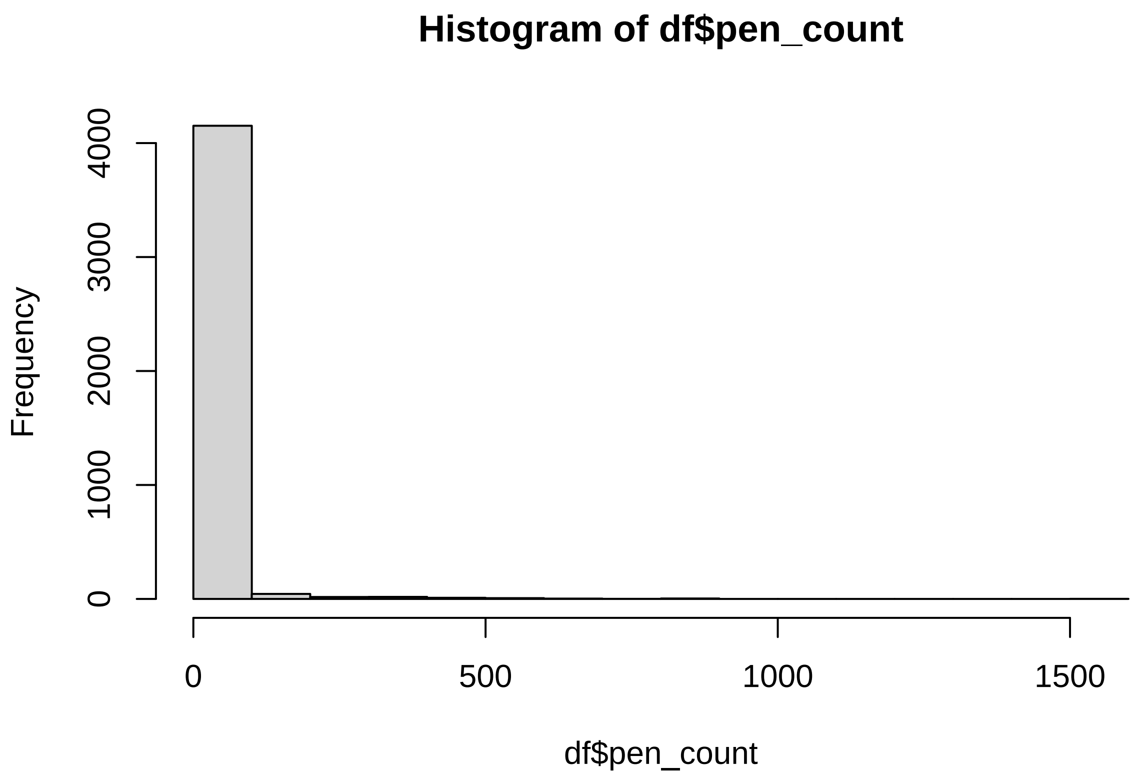
```
getwd()
## [1] "C:/Project/APP/APP/scripts"
df <- read.csv(file.path(data_files_dir, "correlation_data.csv"))

glimpse(df)
## Rows: 4,257
## Columns: 7
## $ NC_ID     <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ~
## $ cert_name <chr> "ARLETA NC", "ARLETA NC", "ARLETA NC", "ARLETA NC", "ARLETA ~
## $ month     <chr> "2019-03-01", "2019-04-01", "2019-05-01", "2019-06-01", "201~
## $ pen_count <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ avg_depl  <int> 0, 0, 2, 3, 2, 2, 2, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ dest_ct   <int> 0, 24, 35, 45, 59, 94, 52, 71, 10, 6, 1, 0, 2, 0, 0, 0, 6, 9~
## $ origin_ct <int> 0, 24, 37, 44, 57, 77, 35, 57, 7, 5, 1, 0, 2, 0, 0, 0, 2, 9,~
```

```
head(df)
##   NC_ID cert_name      month pen_count avg_depl dest_ct origin_ct
## 1     6 ARLETA NC 2019-03-01         0        0       0         0
## 2     6 ARLETA NC 2019-04-01         0        0      24        24
## 3     6 ARLETA NC 2019-05-01         0        2      35        37
## 4     6 ARLETA NC 2019-06-01         0        3      45        44
## 5     6 ARLETA NC 2019-07-01         0        2      59        57
## 6     6 ARLETA NC 2019-08-01         0        2      94        77
```

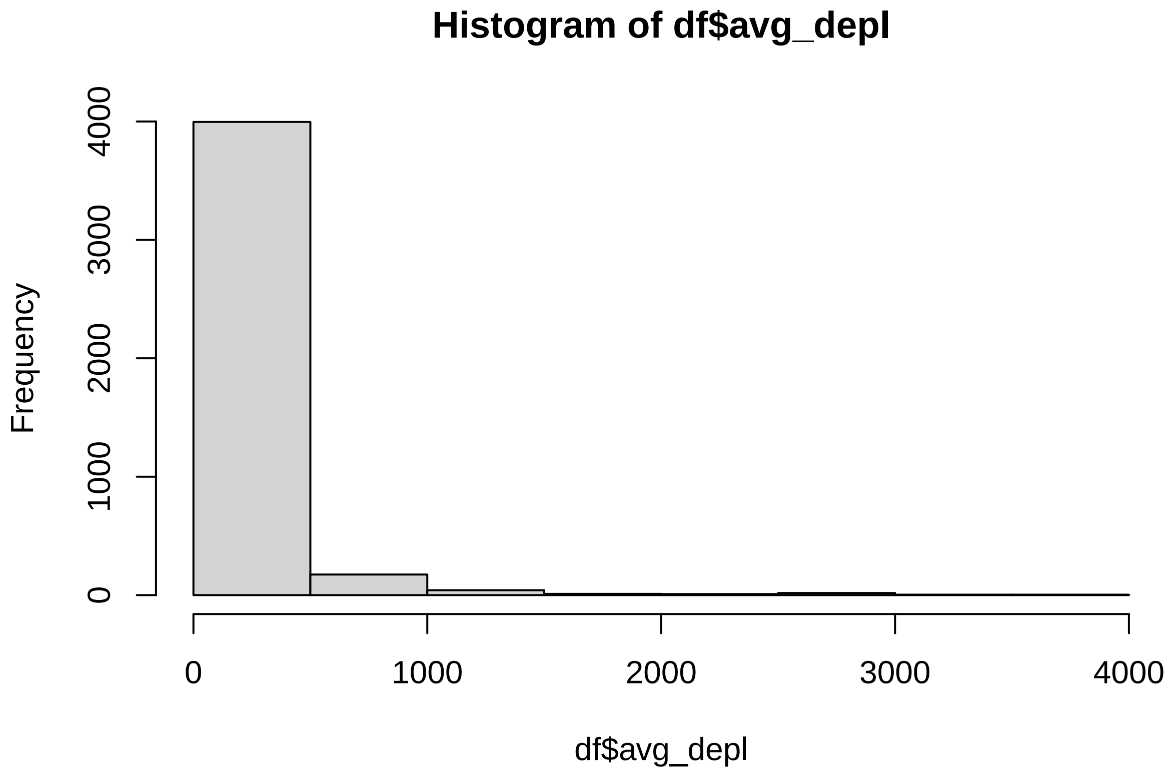
Normality Check

Our data is not normally distributed, so we used Spearman's rho instead of Pearson's r.

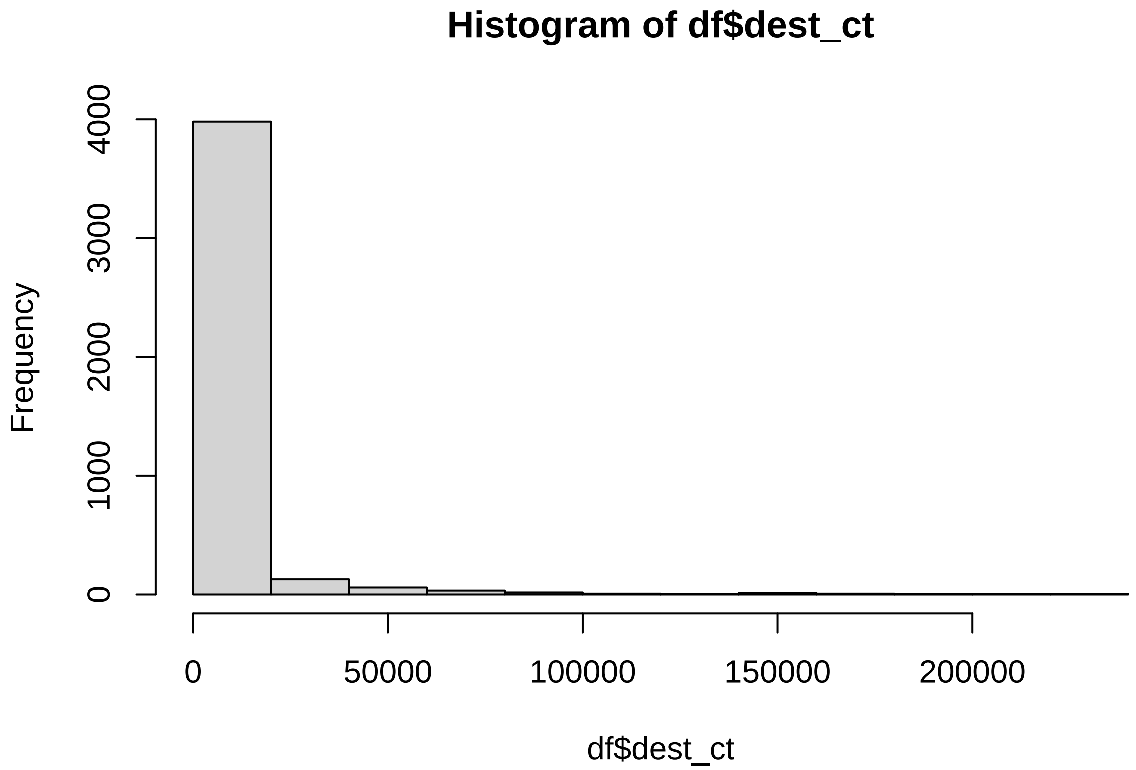
```
hist(df$pen_count)
```



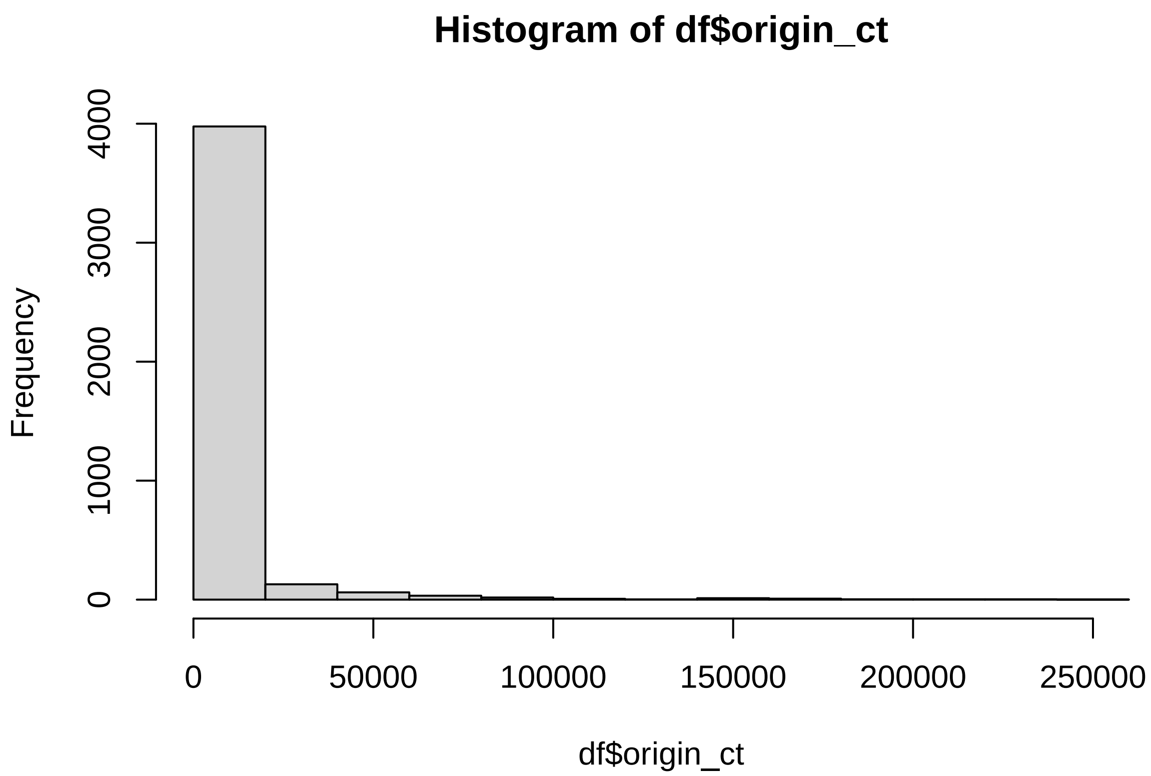
```
hist(df$avg_depl)
```



```
hist(df$dest_ct)
```



```
hist(df$origin_ct)
```



Spearman's Rho

The correlation between penalties and deployment/trips is relatively strong (rho > 0.60). The correlation between deployment and trips is very strong (rho > 0.96).

```
df %>%
    select(pen_count, avg_depl, dest_ct, origin_ct) %>%
    cor(method = "spearman")
##           pen_count  avg_depl   dest_ct origin_ct
## pen_count 1.0000000 0.6468522 0.6271429 0.6266363
## avg_depl  0.6468522 1.0000000 0.9685564 0.9686789
## dest_ct   0.6271429 0.9685564 1.0000000 0.9978118
## origin_ct 0.6266363 0.9686789 0.9978118 1.0000000
```

Kendall's Tau

We also calculated correlation with Kendall's Tau to verify our results. The correlation between penalties and deployment/trips is also relatively strong (rho > 0.50). The correlation between deployment and trips is also very strong (rho > 0.88).

```
df %>%
    select(pen_count, avg_depl, dest_ct, origin_ct) %>%
    cor(method = "kendall")
##           pen_count  avg_depl   dest_ct origin_ct
## pen_count 1.0000000 0.5278084 0.5024109 0.5017615
## avg_depl  0.5278084 1.0000000 0.8844991 0.8816748
## dest_ct   0.5024109 0.8844991 1.0000000 0.9690034
## origin_ct 0.5017615 0.8816748 0.9690034 1.0000000
```

Testing Spearman's Rho

All combinations for the correlation are statistically significant ($p < .001$)

```
cor.test(df$pen_count, df$avg_depl, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
##
## data:  df$pen_count and df$avg_depl
## S = 4540630239, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.6468522
cor.test(df$pen_count, df$dest_ct, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
##
## data:  df$pen_count and df$dest_ct
## S = 4.794e+09, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.6271429
cor.test(df$pen_count, df$origin_ct, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
##
## data:  df$pen_count and df$origin_ct
## S = 4800559084, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.6266363
cor.test(df$avg_depl, df$dest_ct, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
##
## data:  df$avg_depl and df$dest_ct
## S = 404289584, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.9685564
cor.test(df$avg_depl, df$origin_ct, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
```

```
##
## data:  df$avg_depl and df$origin_ct
## S = 402714426, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.9686789
cor.test(df$origin_ct, df$dest_ct, method = "spearman", alternative = "greater", exact = FALSE)
##
##  Spearman's rank correlation rho
##
## data:  df$origin_ct and df$dest_ct
## S = 28135146, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.9978118
```