EDA_ksarshy

Arshdeep Sahota

2025-06-27

Introduction

This exploratory data analysis investigates the relationship between infrastructure damage, specifically pothole frequency, and urban population density. The data used for this project is from a mock dataset of 10 U.S. cities with data on `potholes_reported`, `population_density`, and `annual_infrastructure_spending`.

Research Question:
Do cities with higher population density experience more potholes?
Is infrastructure spending related to fewer potholes reported?

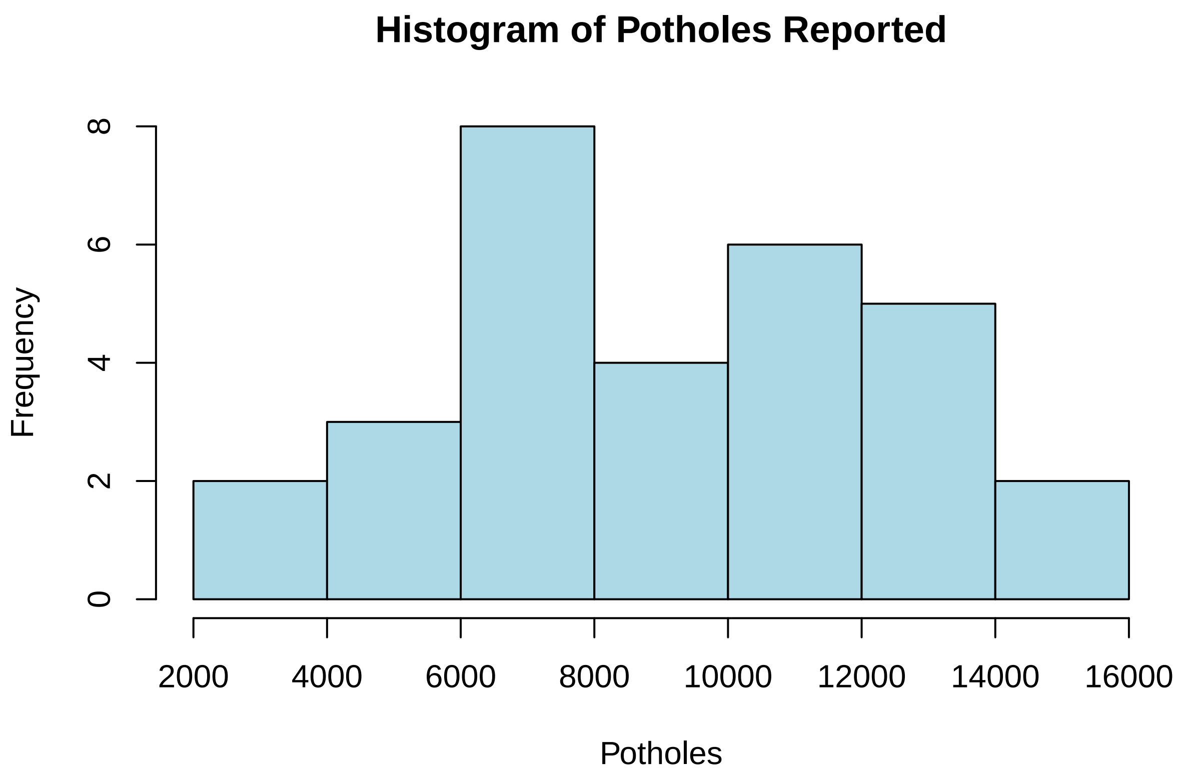
Univariate Exploration

Potholes Reported

```
summary(pothole_data$potholes_reported)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2264    7071    9488    9254   11534   14906
```

```
hist(pothole_data$potholes_reported,
     main = "Histogram of Potholes Reported",
     xlab = "Potholes", col = "lightblue")
```



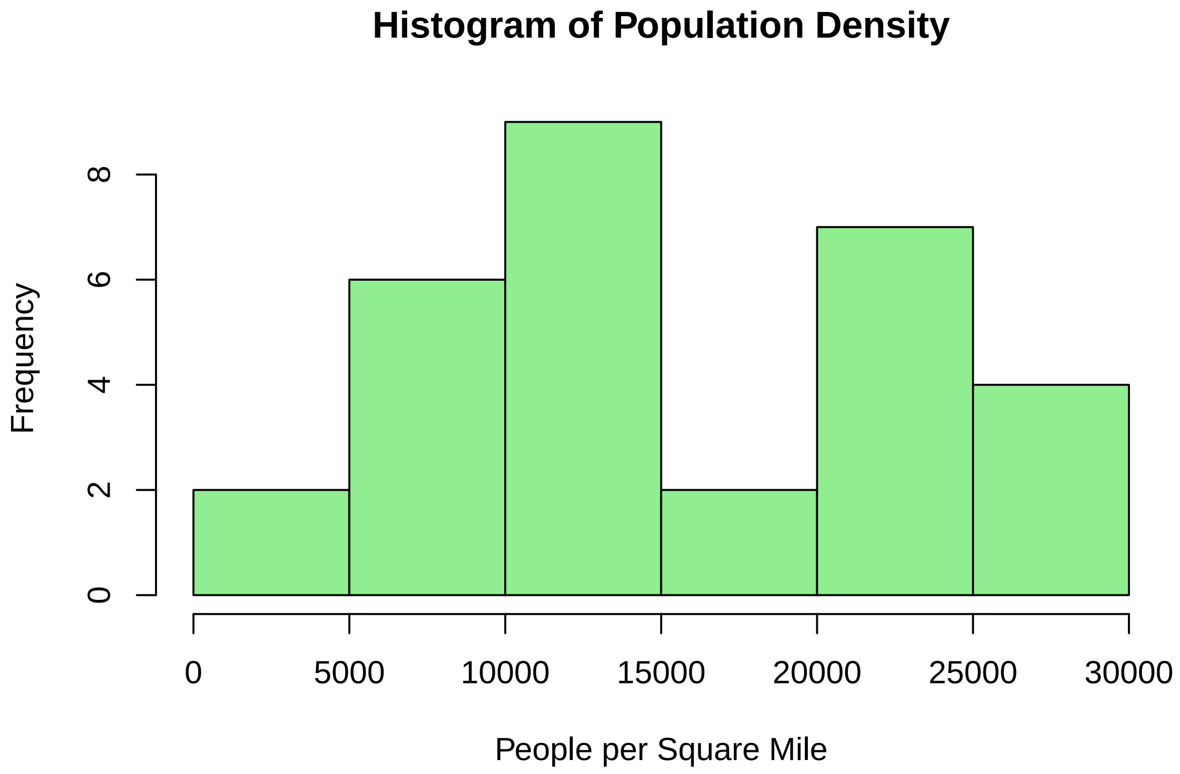
Summary: Potholes reported range from a few thousand to over 11,000, with variation across cities.

Population Density

```
summary(pothole_data$population_density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4202    9941   12420   15679   23600   28409
```

```
hist(pothole_data$population_density,
     main = "Histogram of Population Density",
     xlab = "People per Square Mile", col = "lightgreen")
```



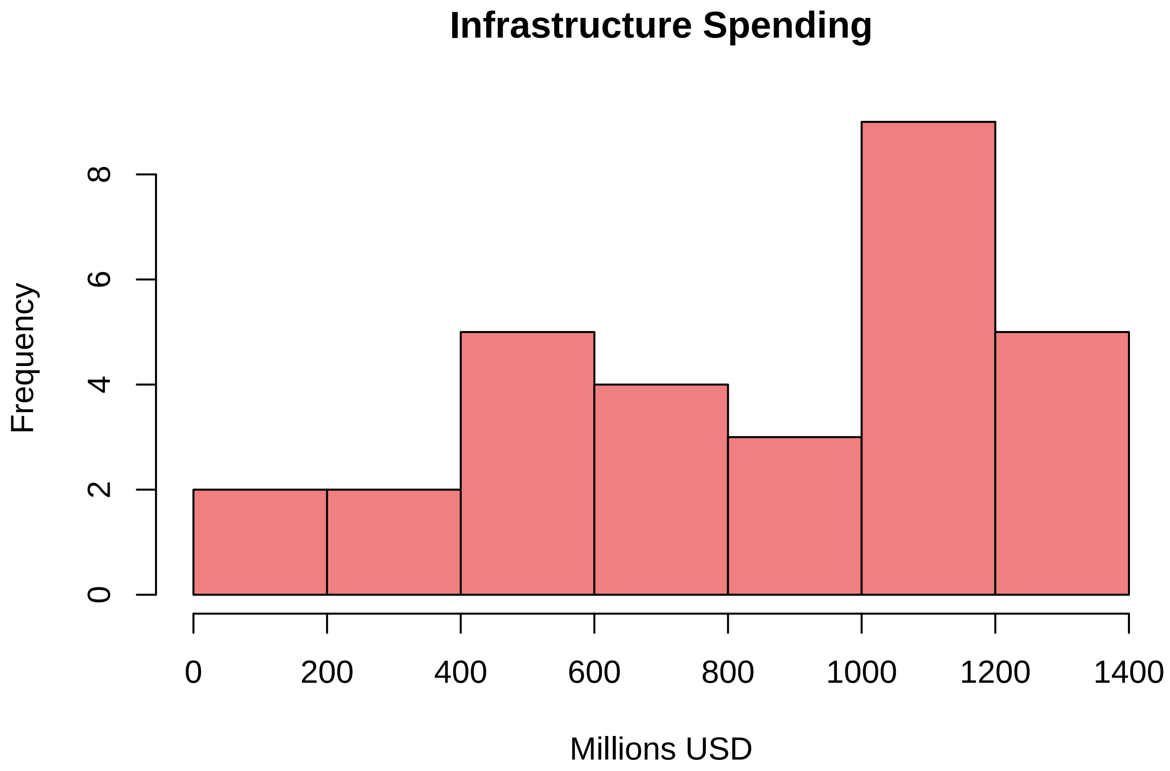
Summary: Cities show large variation in density, from ~3,000 to ~25,000 people per square mile.

Infrastructure Spending

```
summary(pothole_data$annual_infrastructure_spending)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   117.0   497.8   895.5   837.0  1114.5  1400.0
```

```
hist(pothole_data$annual_infrastructure_spending,
     main = "Infrastructure Spending",
     xlab = "Millions USD", col = "lightcoral")
```



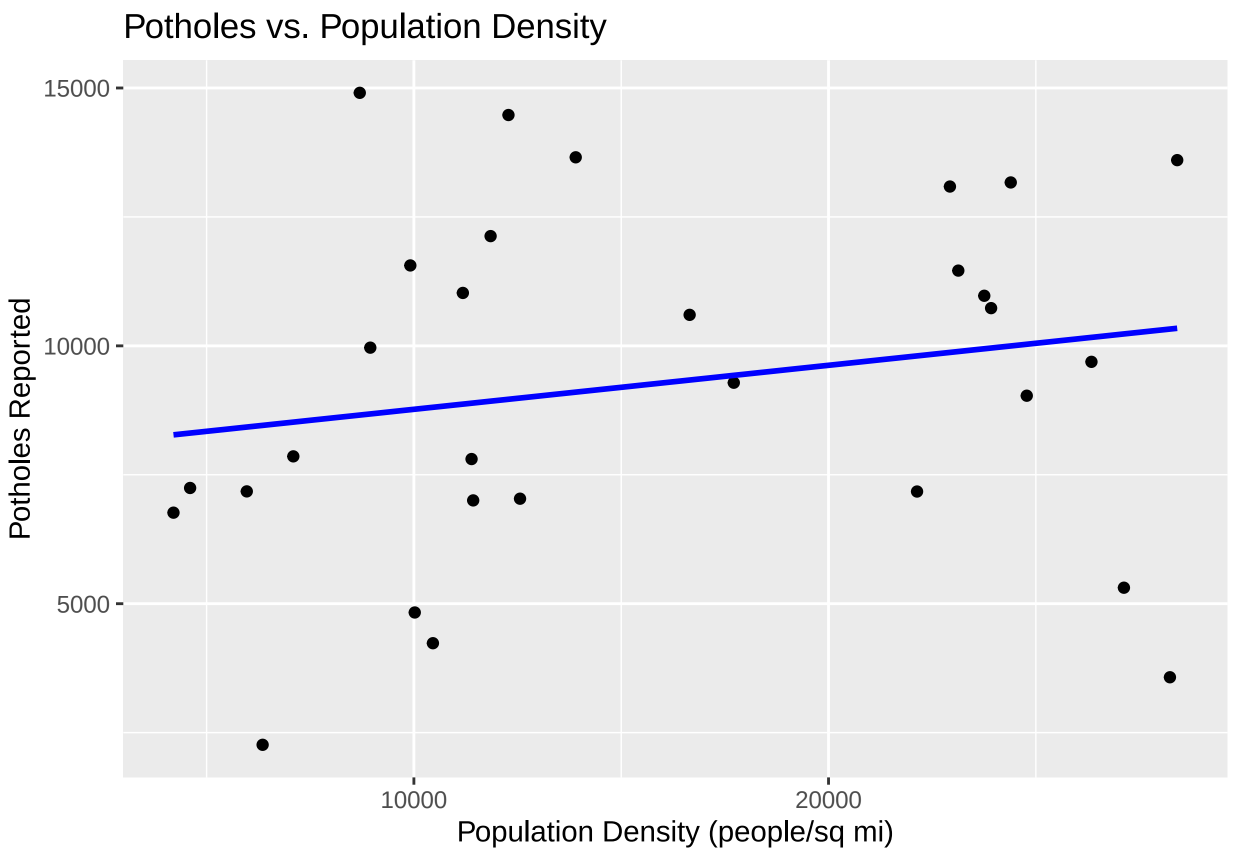
Summary: Spending ranges from ~$100M to over $1B. Higher spending might reflect higher need.

Bivariate Exploration

Potholes vs. Population Density

```
ggplot(pothole_data, aes(x = population_density, y = potholes_reported)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Potholes vs. Population Density",
       x = "Population Density (people/sq mi)",
       y = "Potholes Reported")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

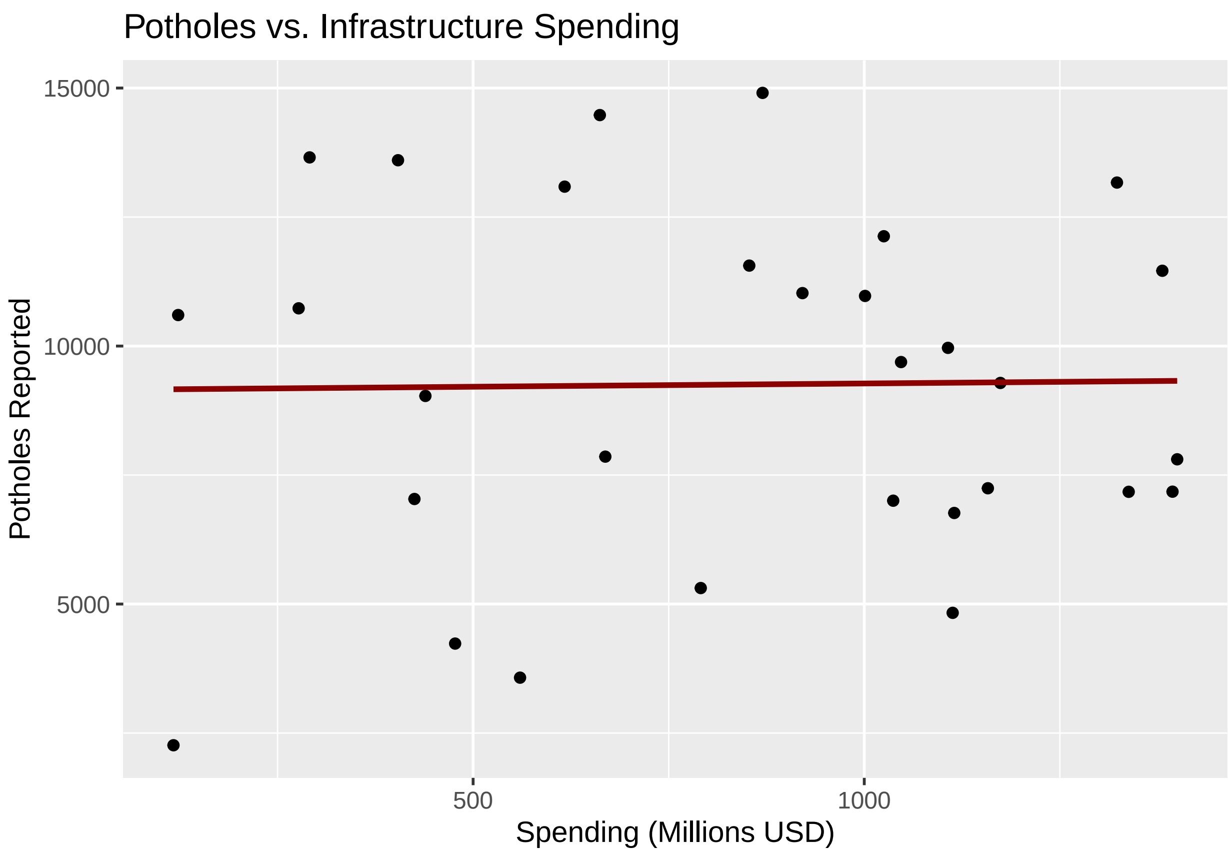


Summary: Positive trend — denser cities tend to have more potholes reported.

Potholes vs. Infrastructure Spending

```r
ggplot(pothole_data, aes(x = annual_infrastructure_spending, y = potholes_reported)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  labs(title = "Potholes vs. Infrastructure Spending",
       x = "Spending (Millions USD)",
       y = "Potholes Reported")
```

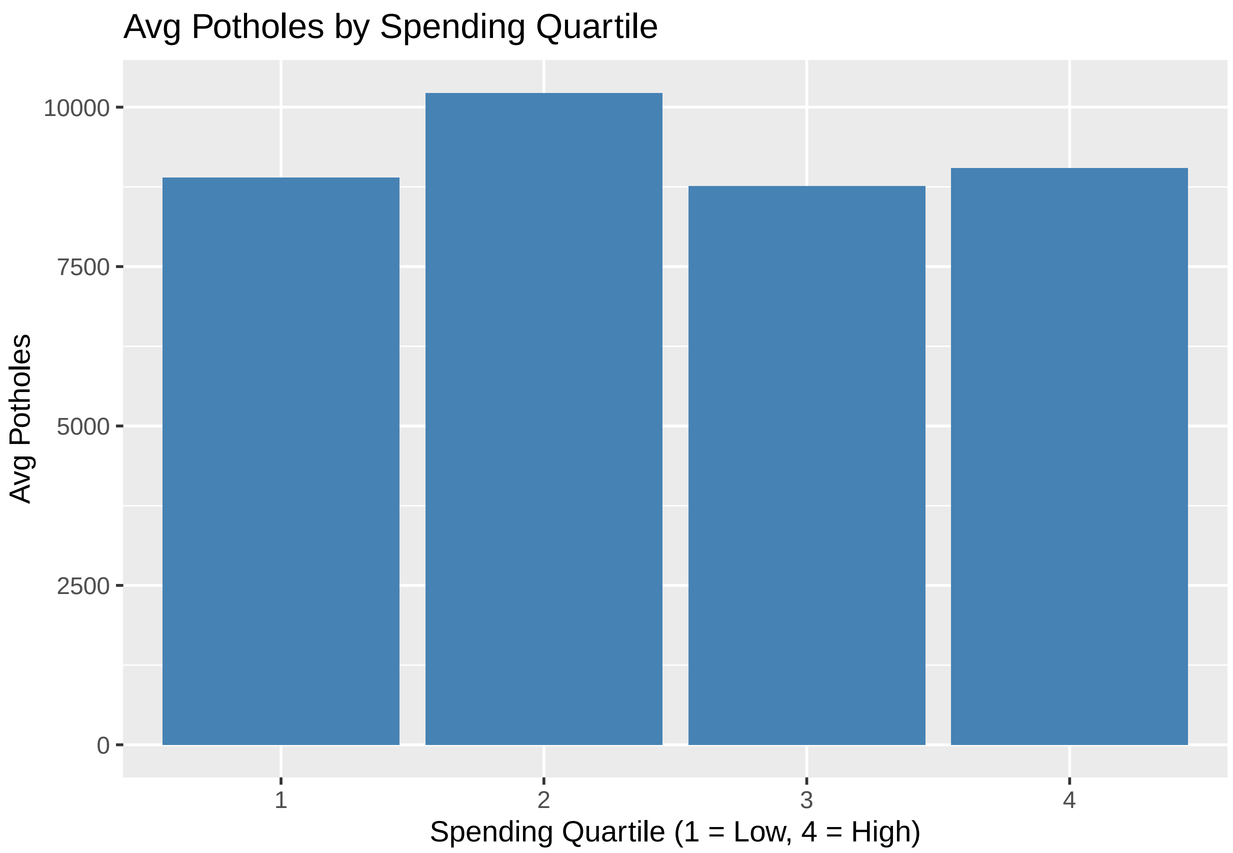
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Summary: Slight negative trend — higher spending may reduce potholes, but not strongly.

Potholes by Spending Quartile

```
pothole_data %>%
  mutate(spending_quartile = ntile(annual_infrastructure_spending, 4)) %>%
  group_by(spending_quartile) %>%
  summarise(mean_potholes = mean(potholes_reported)) %>%
  ggplot(aes(x = factor(spending_quartile), y = mean_potholes)) +
  geom_col(fill = "steelblue") +
  labs(title = "Avg Potholes by Spending Quartile",
       x = "Spending Quartile (1 = Low, 4 = High)",
       y = "Avg Potholes")
```



Summary: Higher spending doesn't always lead to fewer potholes. Context matters.

Conclusion

This EDA shows:

- Potholes increase with population density.
- Infrastructure spending may reduce potholes, but the effect is unclear.
- Future studies should include weather and road age.
- No formal statistical tests were run. This is purely descriptive.