Marine Traffic Data Analysis: Technical Summary

Statistical Data Analytics using R - Project Report
Date: October 15, 2025

---

1. Data Acquisition - 4-Quadrant Scraping

```
coordinates = ["00", "01", "10", "11"]

    for coord in coordinates:
        x = coord[0]
        y = coord[1]
        url = f"https://www.marinetraffic.com/getData/get_data_json_4/z:2/X:{x}/Y:{y}/station:0"

        print(f"Fetching data for X:{x}, Y:{y}...")
        driver.get(url)

        # Wait for page to load and get the pre tag content
        time.sleep(2)  # Give it time to load
        page_source = driver.page_source

        # Extract JSON from page
        try:
            # Find the body text which contains the JSON
            body = driver.find_element(By.TAG_NAME, "body")
            json text = body.text
            data = json.loads(json_text)

            if "data" in data and "rows" in data["data"]:
                rows = data["data"]["rows"]
                all_rows.extend(rows)
                print(f"Loaded {len(rows)} records from X:{x}, Y:{y}")
            else:
                print(f"Warning: No rows found for X:{x}, Y:{y}")
        except Exception as e:
            print(f"Error parsing JSON for X:{x}, Y:{y}: {e}")
return all_rows
```

Result: 10,379 vessels from 4 quadrants

---

2. Categorical Enrichment

```
type_mapping <- df %>%
  filter(!is.na(SHIPTYPE) & !is.na(TYPE_NAME)) %>%
  group_by(SHIPTYPE) %>%
  summarize(TYPE_NAME = first(TYPE_NAME))

df <- df %>%
  left_join(type_mapping, by = "SHIPTYPE") %>%
  mutate(TYPE_NAME = coalesce(TYPE_NAME.x, TYPE_NAME.y))
```

Coverage: 37% → 100%

---

3. ROT Imputation

```
df %>%
  filter(STATUS_NAME %in% c('At Anchor', 'Moored')) %>%
  pull(SPEED) %>%
  summary()

df %>%
  filter(SPEED > 10, HC_DIFF_ABS < 5)
```

Decision: Impute ROT = 0

---

4. Speed Distribution

```
hist(df$SPEED, breaks=50, main="Speed Distribution")
hist(df$SPEED[df$SPEED > 0], breaks=50)
```

```
df %>%
  filter(SPEED > quantile(SPEED, 0.99, na.rm=TRUE)) %>%
  select(SHIPNAME, SPEED, SHIPTYPE, LENGTH)
```

Finding: Max 405 knots (GPS errors)

---

5. Heading-Course Misalignment

```
both <- both %>%
  mutate(
    HC_DIFF = ifelse(
      abs(HEADING - COURSE) <= 180,
      HEADING - COURSE,
      ifelse(HEADING - COURSE > 180, HEADING - COURSE - 360, HEADING - COURSE + 360)
    ),
    HC_DIFF_ABS = abs(HC_DIFF)
  )

hist(both$HC_DIFF_ABS, breaks=50)
summary(both$HC_DIFF_ABS)
```

Result: 50% at 0°, 75% < 1° - Not a distress indicator

---

6. Length-Width Ratio

```
df <- df %>%
  filter(!is.na(LENGTH), !is.na(WIDTH), WIDTH > 0) %>%
  mutate(LW_RATIO = LENGTH / WIDTH)

df %>%
  group_by(SHIPTYPE) %>%
  summarize(
    count = n(),
    median_ratio = median(LW_RATIO, na.rm=TRUE),
    mean_ratio = mean(LW_RATIO, na.rm=TRUE),
    sd_ratio = sd(LW_RATIO, na.rm=TRUE)
  ) %>%
  arrange(desc(median_ratio))
```

Vessel Type	Median LW_RATIO
Fishing Vessels	6.8
Tankers	5.8
Cargo Vessels	5.2

---

7. Correlation Analysis

```
numeric_cols <- df %>% select_if(is.numeric)
cor_matrix <- cor(numeric_cols, use = "pairwise.complete.obs")

library(corrplot)
corrplot(cor_matrix, method = "color", type = "upper",
         tl.cex = 0.8, addCoef.col = "black", number.cex = 0.7)
```

Feature Pair	Correlation
LENGTH ↔ WIDTH	r = 0.72
WIDTH ↔ DWT	r = 0.89
HEADING ↔ COURSE	r = 0.93

---

8. Speed by Type

```
df %>%
  group_by(SHIPTYPE) %>%
  summarize(
    count = n(),
    mean_speed = mean(SPEED, na.rm=TRUE),
    median_speed = median(SPEED, na.rm=TRUE),
    max_speed = max(SPEED, na.rm=TRUE)
  )
```

---

9. DWT vs Length

```
df %>%
  filter(!is.na(LENGTH), !is.na(DWT), DWT > 0) %>%
  ggplot(aes(x = LENGTH, y = DWT, color = SHIPTYPE)) +
  geom_point(alpha = 0.6) +
  scale_y_log10() +
  labs(title = "DWT vs Length (Log Scale)")
```

---

10. Linear Regression

```
model_data <- df %>% filter(!is.na(LENGTH), !is.na(WIDTH), WIDTH > 0, LENGTH > 0)

lm_model <- lm(WIDTH ~ LENGTH, data = model_data)
summary(lm_model)
```

Model: WIDTH = 2.47 + 0.1234 × LENGTH
Performance: $R^2 = 0.518$, $p < 2.2e\text{-}16$

---

11. Diagnostics

```
par(mfrow = c(2, 2))
plot(lm_model)
```

Issues: Heteroscedasticity, non-normality, outliers

---

12. Stratified Visualization

```r
ggplot(model_data, aes(x = LENGTH, y = WIDTH, color = SHIPTYPE)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "black") +
  labs(title = "Length-Width Relationship by Vessel Type")
```

Summary

Metric	Value
Total Vessels	10,379
API Quadrants	4 (00, 01, 10, 11)
Categorical Improvement	37% → 100%
LENGTH-WIDTH Corr	r = 0.72
Model R^2	0.518

Course: Statistical Data Analytics using R (SDAUR)
Framework: R (tidyverse, ggplot2, corrplot)
Source: MarineTraffic.com API