Reddit Dataset Appendix

1.1: Dataset Overview

This dataset contains Reddit comments referencing four major U.S. shipping companies:

- USPS
- UPS
- DHL
- FedEx

The data were collected using the Reddit API. Raw posts and comments were cleaned to remove formatting artifacts, boilerplate content, and non-informative text. Sentiment scores were computed using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model.

Total observations analyzed: **2,006 comments**

Company Breakdown:

Company	Number of Comments
USPS	793
UPS	597
DHL	411
FedEx	205

1.2: Unit of Observation

Each row in the Reddit dataset represents one individual Reddit comment referencing a shipping company.

- Observational unit = single comment
- Each observation contains metadata and computed sentiment measures

1.3: Variable Documentation

1. Variable: company

- Type: Categorical (string)
- Description: Identifies which shipping company the Reddit comment references.
- Possible Values: USPS, UPS, FedEx, DHL

2. Variable: subreddit

- Type: String
- Name of the subreddit where the comment was posted.
- Used for contextual metadata.

3. Variable: post_id

- Type: String
- Unique identifier of the Reddit submission (thread) containing the comment.

4. Variable: comment_id

- Type: String
- Unique identifier for the specific Reddit comment.

5. Variable: comment_created_utc

- Type: Integer (Unix timestamp)
- Time the comment was created (UTC format).

6. Variable: comment_created_dt

- Type: Datetime
- Human-readable version of the creation timestamp.

7. Variable: comment_body

- Type: String
- Original raw comment text as retrieved from Reddit.

8. Variable: comment_score

- Type: Integer
- Net upvote score of the comment at time of scraping.

9. Variable: parent_id

- Type: String
- Identifier of the parent comment or submission.
- Allows reconstruction of conversation threads.

10. Variable: text_raw

- Type: String
- Initial preprocessed version of the comment text.

11. Variable: text_clean

- Type: String (text)
- Description: Cleaned version of the original Reddit comment.
- Preprocessing steps included:
    - Removal of boilerplate text
    - Lowercasing
    - Removal of formatting artifacts
- This variable is the input for sentiment analysis.

12. Variable: keep_basic

- Type: Boolean
- Indicator for whether the comment passed basic cleaning filters.

13. Variable: is_boilerplate

- Type: Boolean
- Indicates whether the comment was identified as boilerplate or auto-generated content.

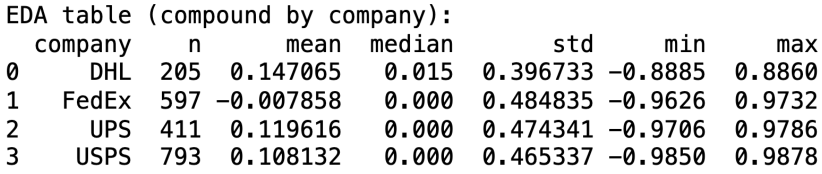
14. Variable: keep_final

- Type: Boolean
- Indicates whether the comment was retained in the final analysis dataset.

15. Variable: Vader_compound

- Type: Continuous numeric

- Range: -1 to +1
- Description: VADER compound sentiment score computed from text_clean.
- Interpretation:
    - Values close to +1 → Highly positive sentiment
    - Values close to -1 → HIghly negative sentiment
    - Values near 0 → Neutral sentiment



```
EDA table (compound by company):
  company    n      mean  median       std     min     max
0     DHL  205  0.147065   0.015  0.396733 -0.8885  0.8860
1   FedEx  597 -0.007858   0.000  0.484835 -0.9626  0.9732
2     UPS  411  0.119616   0.000  0.474341 -0.9706  0.9786
3    USPS  793  0.108132   0.000  0.465337 -0.9850  0.9878
```

Observation:

- FedEx has a slightly negative mean sentiment.
- DHL has the highest sentiment score.

16. Variable: vader_pos

- Type: Numeric (0 to 1)
- Proportion of text classified as positive sentiment.

17. Variable: vader_neu

- Type: Numeric (0 to 1)
- Proportion of text classified as neutral sentiment.

18. Variable: vader_neg

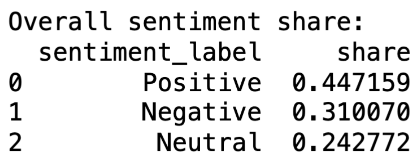
- Type: Numeric (0 to 1)
- Proportion of text classified as negative sentiment.

19. Variable: Sentimental_label

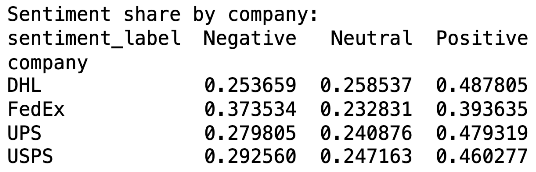
- Type: Categorical
- Description: Three-class sentiment label derived from VADER compound score.

Classification rule:

- compound ≥ 0.05 → Positive
- compound ≤ -0.05 → Negative
- Otherwise → Neutral



```
Overall sentiment share:
  sentiment_label     share
0        Positive  0.447159
1        Negative  0.310070
2         Neutral  0.242772
```



```
Sentiment share by company:
sentiment_label  Negative   Neutral  Positive
company
DHL              0.253659  0.258537  0.487805
FedEx            0.373534  0.232831  0.393635
UPS              0.279805  0.240876  0.479319
USPS             0.292560  0.247163  0.460277
```

20. Variable: neg_bin

- Type: Binary (0/1)
- Indicator variable for binomial test.
- 1 = Negative sentiment
- 0 = Not negative

1.4: Figures and Visualization

1. Number of Comments per Company (Bar Chart)

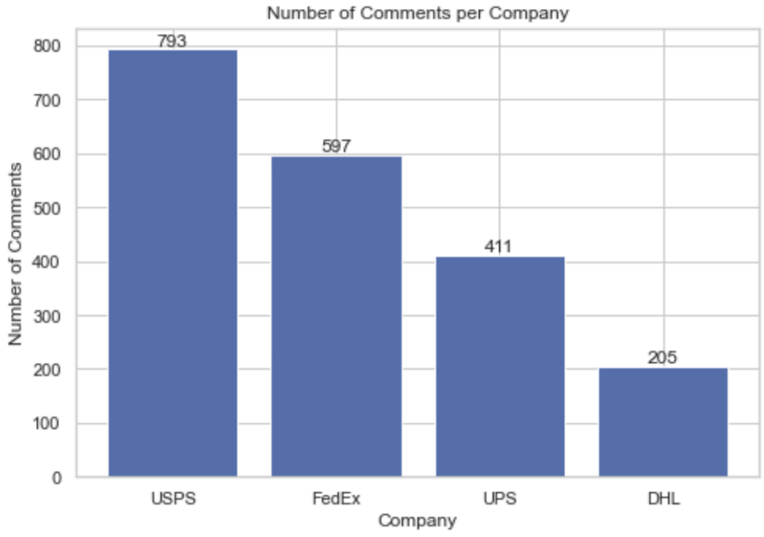


Figure Description:
This bar chart displays the number of Reddit comments collected for each shipping company (USPS, FedEx, UPS, DHL).

Variables Used:

- Company
- Comment count per company

What the Figure Shows:

- USPS has the largest sample size (793 comments).
- FedEx follows with 597 comments.
- UPS has 411 comments.
- DHL has the smallest sample size (205 comments).

Interpretation:
The sample sizes are unequal across companies. USPS is discussed more frequently on Reddit, while DHL has the smallest presence. This difference in sample size is important when interpreting statistical results, especially in ANOVA and binomial tests.

2. Overall Distribution of VADER Compound Sentiment (Histogram)

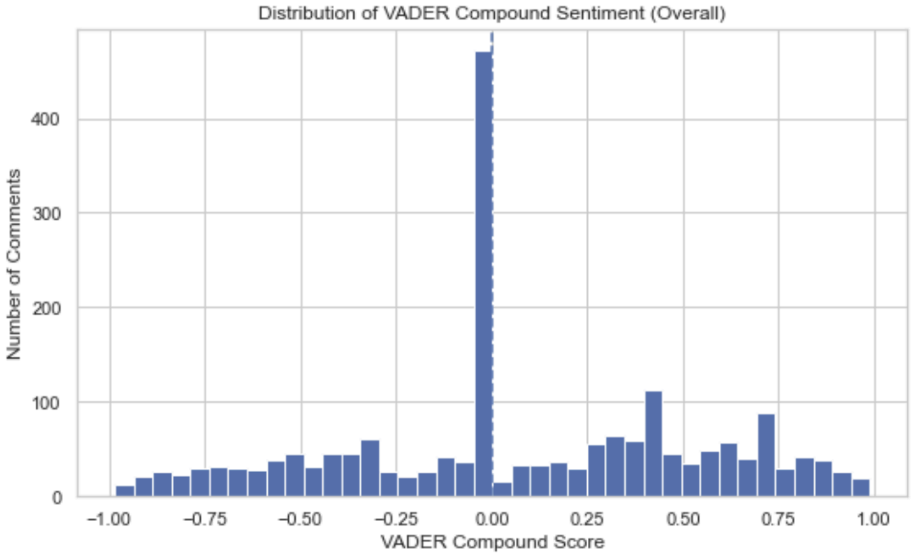


Figure Description:
This histogram shows the distribution of the vader_compound sentiment score across all 2,006 Reddit comments.

Variable Used:

- vader_compound (range: -1 to +1)

What the Figure Shows:

- A large concentration of scores around 0 (neutral).
- A visible positive skew in the distribution.
- Fewer extremely negative scores compared to moderate positive scores.

Interpretation:
Most Reddit comments are either neutral or moderately positive. Strongly negative sentiment is present but not dominant overall.

3. Distribution of VADER Compound by Company (Faceted Histograms)

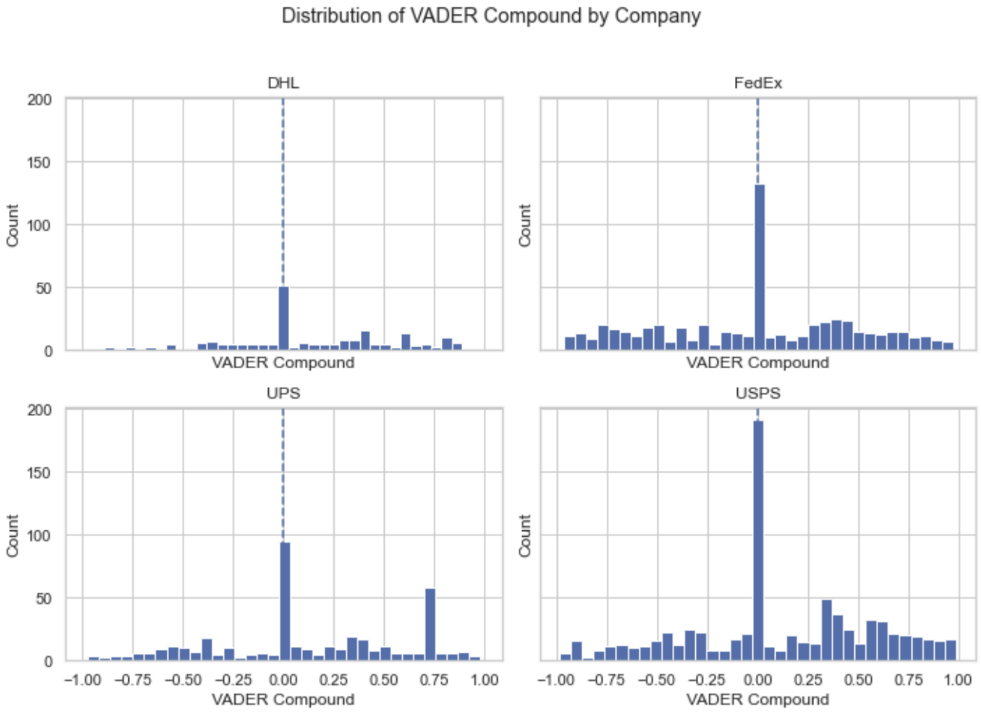


Figure Description:
This set of four histograms shows the sentiment distribution separately for DHL, FedEx, UPS, and USPS.

Variable Used:

- Vader_compound
- company

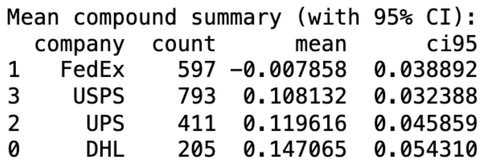
What the Figure Shows:

- FedEx has a noticeably larger mass of negative values compared to others.
- DHL and UPS show more positive clustering.
- USPS has a fairly balanced distribution but slightly positive leaning.

Interpretation:
The shape of the sentiment distribution differs by company. FedEx appears to receive relatively more negative comments, while DHL appears more positively skewed.

This visual evidence supports the need for formal hypothesis testing (ANOVA and Chi-square).

4. Mean VADER Compound by Company (95% Confidence Intervals)



```
Mean compound summary (with 95% CI):
  company  count      mean      ci95
1   FedEx    597 -0.007858  0.038892
3    USPS    793  0.108132  0.032388
2     UPS    411  0.119616  0.045859
0     DHL    205  0.147065  0.054310
```

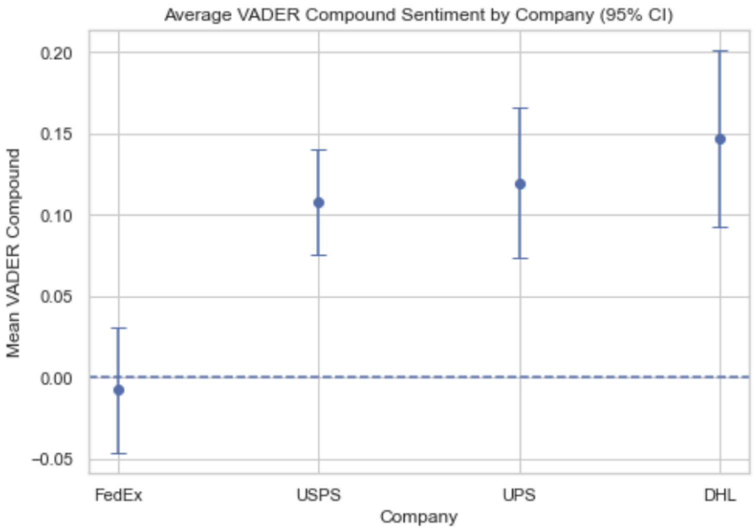


Figure Description:
This point plot displays the mean compound sentiment for each company with 95% confidence intervals.

Variables Used:

- Vader_compound
- company

What the Figure Shows:

- FedEx mean ≈ -0.008 (slightly negative / near neutral)
- USPS mean ≈ 0.108
- UPS mean ≈ 0.120
- DHL mean ≈ 0.147

Confidence intervals for FedEx do not overlap strongly with DHL.

Interpretation:
FedEx is the only company with a near-zero or slightly negative average sentiment.
DHL has the highest average sentiment.
UPS and USPS fall in between.

The visual separation between FedEx and other companies aligns with significant ANOVA and Tukey HSD results

5. VADER Compound Distribution by Company (Boxplot)

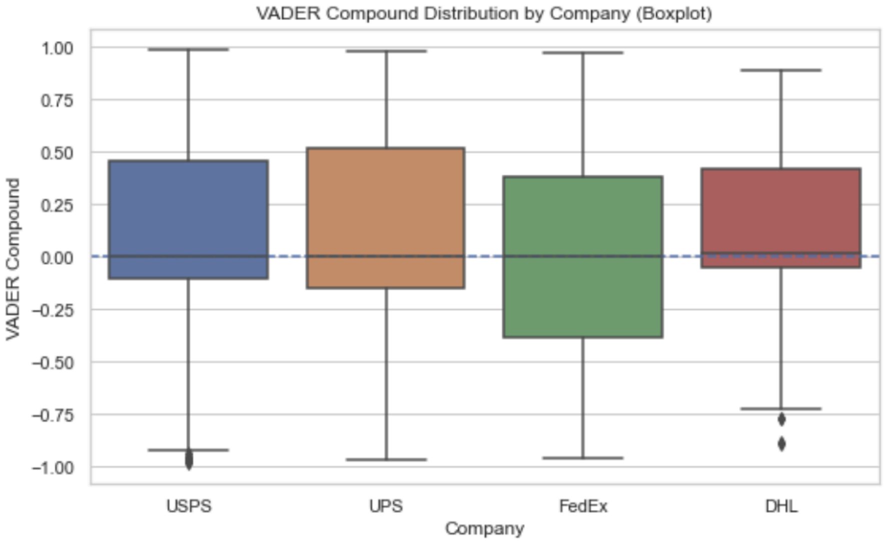


Figure Description:
This boxplot compares the distribution of compound sentiment scores across companies.

Variables Used:

- Vader_compound
- company

What the Figure Shows:

- Median sentiment is highest for DHL.
- FedEx has a lower median and greater spread toward negative values.
- All companies show wide variability.
- Outliers are present in both positive and negative extremes.

Interpretation:
There is clear variation in sentiment central tendency across companies.
FedEx shows more negative tail behavior.
DHL shows a stronger positive central tendency.

This visually reinforces the ANOVA results indicating mean differences.

6. Overall Sentiment Share (Pie / Table Visualization)

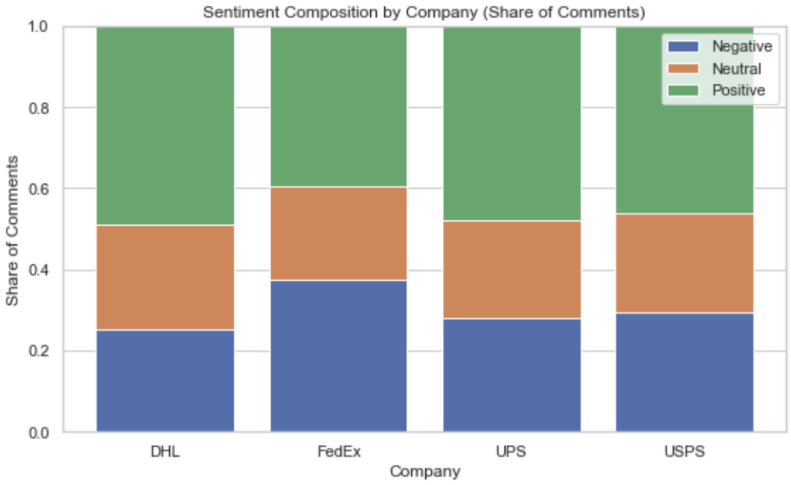


Figure Description:
This visualization summarizes the overall proportion of sentiment categories across all companies.

Variables Used: sentiment_label

Overall Sentiment Breakdown:

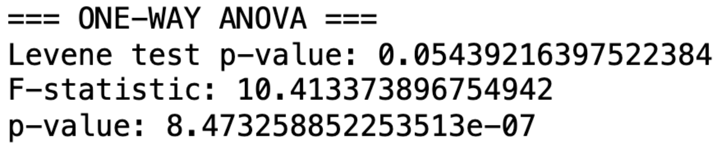
- Positive: 44.7%
- Negative: 31.0%
- Neutral: 24.3%

Interpretation:
Reddit discussions about shipping companies are predominantly positive overall.
Negative sentiment represents less than one-third of comments.

This explains why the binomial test (>50% negative) was not significant.

1.5: Statistical Procedures

1. One-Way ANOVA



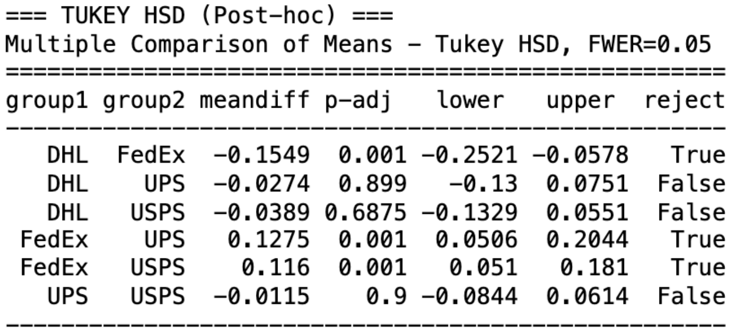
```
=== ONE-WAY ANOVA ===
Levene test p-value: 0.05439216397522384
F-statistic: 10.413373896754942
p-value: 8.473258852253513e-07
```

Purpose: Test whether mean VADER sentiment differs across companies.

Results: F-statistic = 10.41, p-value < 0.001

Conclusion:
Mean sentiment differs significantly across at least two companies.

2.Turkey HSD Post-Hoc Test



```
=== TUKEY HSD (Post-hoc) ===
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
   DHL  FedEx  -0.1549  0.001 -0.2521 -0.0578   True
   DHL    UPS  -0.0274  0.899   -0.13  0.0751  False
   DHL   USPS  -0.0389 0.6875 -0.1329  0.0551  False
 FedEx    UPS   0.1275  0.001  0.0506  0.2044   True
 FedEx   USPS    0.116  0.001   0.051   0.181   True
   UPS   USPS  -0.0115    0.9 -0.0844  0.0614  False
----------------------------------------------------
```

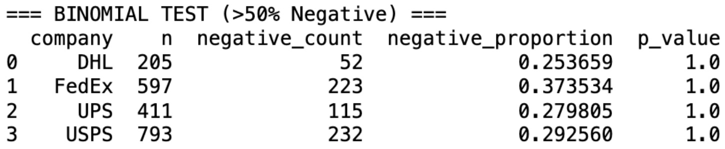
Purpose: In greater detail, test the significance differences in sentiment found between each company.

Significant differences found between:

- DHL vs FedEx
- FedEx vs UPS
- FedEx vs USPS

FedEx exhibits significantly lower sentiment than multiple competitors.

3. Binomial Test (>50% Negative)

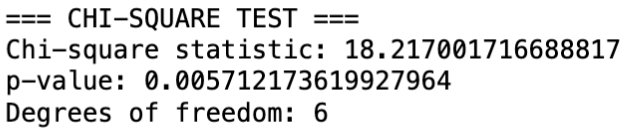


```
=== BINOMIAL TEST (>50% Negative) ===
  company    n  negative_count  negative_proportion  p_value
0     DHL  205              52             0.253659      1.0
1   FedEx  597             223             0.373534      1.0
2     UPS  411             115             0.279805      1.0
3    USPS  793             232             0.292560      1.0
```

Purpose:
Test whether any company has a majority (>50%) of negative comments.

Result:
No company has more than 50% negative sentiment (all p-values = 1.0).

4. Chi-Square Test of Independence



```
=== CHI-SQUARE TEST ===
Chi-square statistic: 18.217001716688817
p-value: 0.005712173619927964
Degrees of freedom: 6
```

Purpose:
Test whether the sentiment category (Positive/Neutral/Negative) depends on the company.

Results:

- Chi-square = 18.22
- p-value = 0.0057
- Degrees of freedom = 6

Conclusion: Sentiment composition differs significantly across companies.