# Problem 2

### 1. Summarization

First, I aggregated the trip-level data into a comprehensive route-level summary. For each of the 77 routes, I computed key statistics to understand their individual performance and characteristics. This summary includes the total number of trips, average and standard deviation of distance, duration, and load, as well as cost metrics.

#### (i) Sample Route Summary (Top 5 rows)

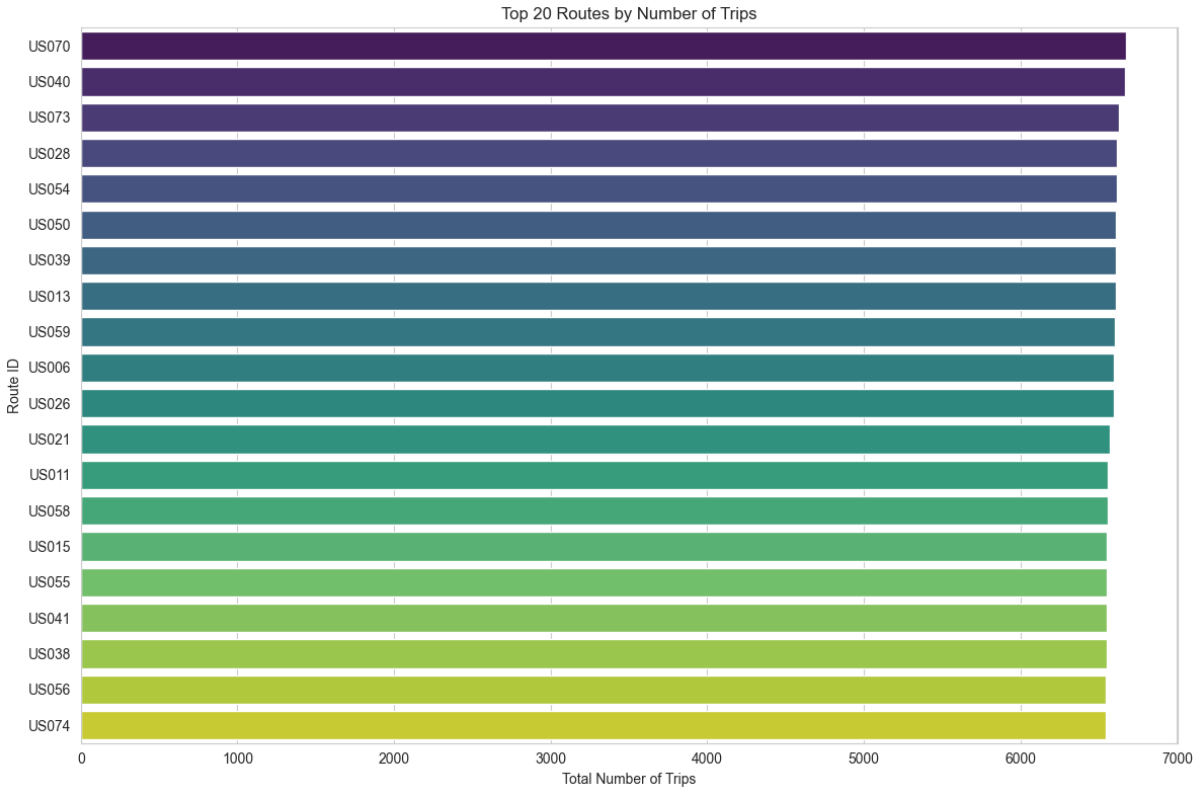
Click here to view the aggregated data

| Route ID | Total Trips | Avg Distance (km) | Std Distance (km) | Avg Duration (hrs) | Std Duration (hrs) | Avg Load (tons) | Std Load (tons) | Avg Cost per km (USD) | Std Cost per km (USD) | Total Distance (km) | Total Cost (USD) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| US001 | 6,444 | 46.01 | 6.91 | 5.91 | 0.45 | 15.99 | 2.58 | 0.845 | 0.030 | 2,229,677.13 | 1,884,548.41 |
| US002 | 6,393 | 153.03 | 3.05 | 2.63 | 0.25 | 16.04 | 2.60 | 0.729 | 0.026 | 978,316.29 | 712,711.61 |
| US003 | 6,538 | 361.98 | 7.28 | 5.32 | 0.41 | 16.02 | 2.59 | 0.728 | 0.026 | 2,366,644.54 | 1,724,046.76 |
| US004 | 6,449 | 708.10 | 14.15 | 11.41 | 0.85 | 16.10 | 2.61 | 0.846 | 0.030 | 4,566,530.91 | 3,861,581.21 |
| US005 | 6,448 | 499.18 | 10.15 | 8.24 | 0.62 | 15.99 | 2.59 | 0.846 | 0.030 | 3,218,729.30 | 2,721,942.11 |

#### (ii) Visualizations of Route Characteristics

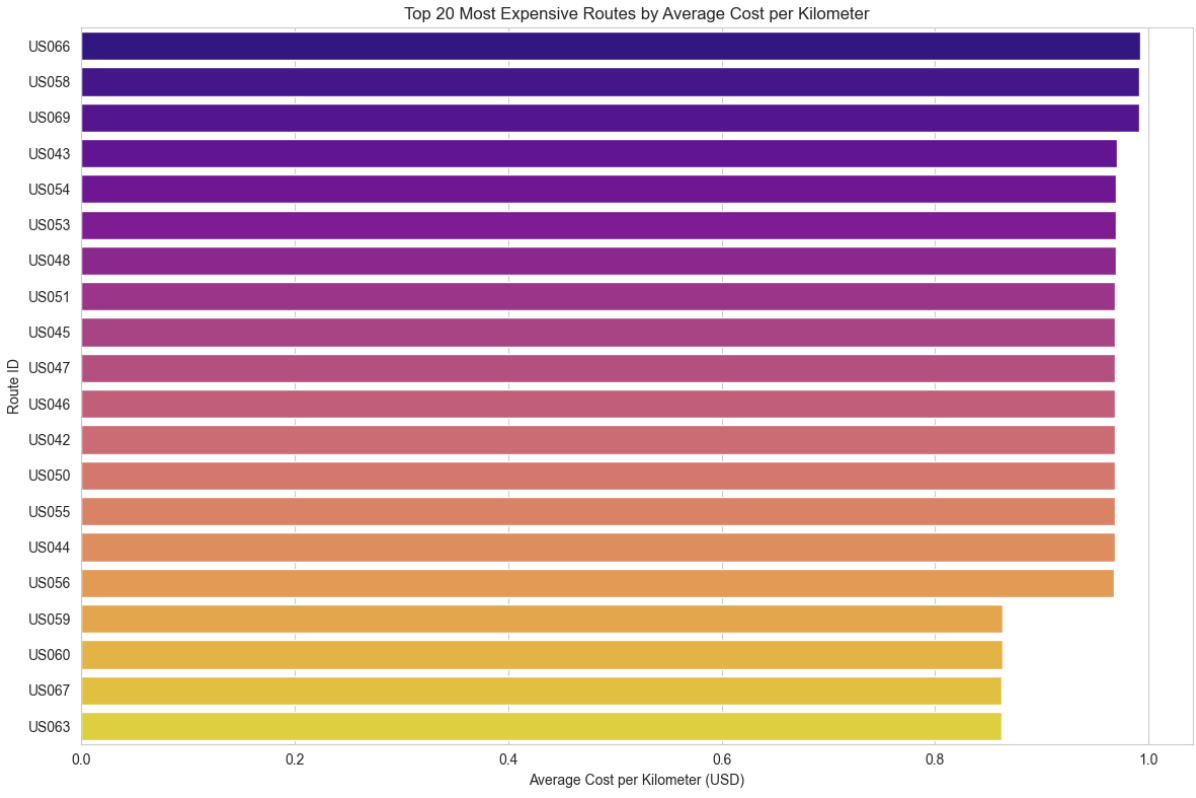
To better understand the differences between the routes, I created a series of visualizations:

- **Top 20 Routes by Trip Volume:**



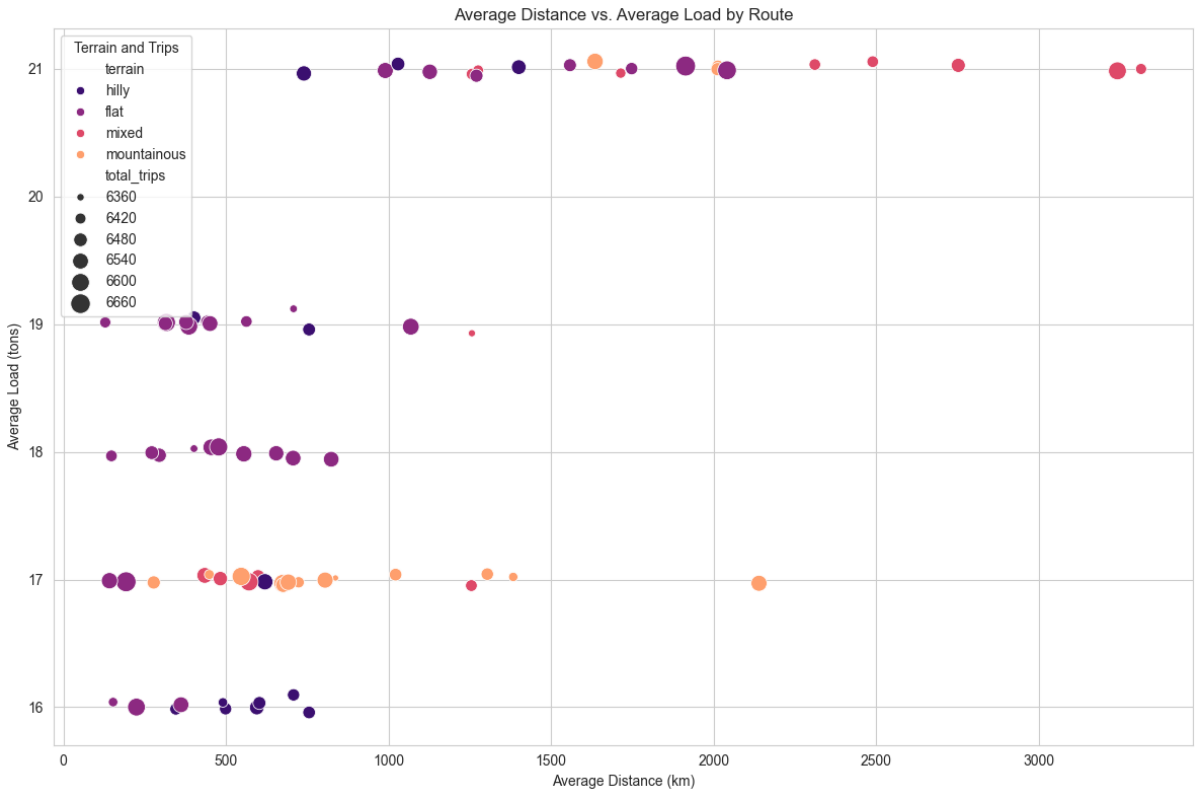

The number of trips is fairly evenly distributed across the top routes, with most of the busiest routes handling a similar volume of traffic. This suggests a well-distributed logistics network.

- **Top 20 Most Expensive Routes:**



There is a clear distinction in cost per kilometer between routes. The most expensive routes are significantly more costly than the average, which often correlates with challenging terrain such as mountains.

- **Average Distance vs. Load:**



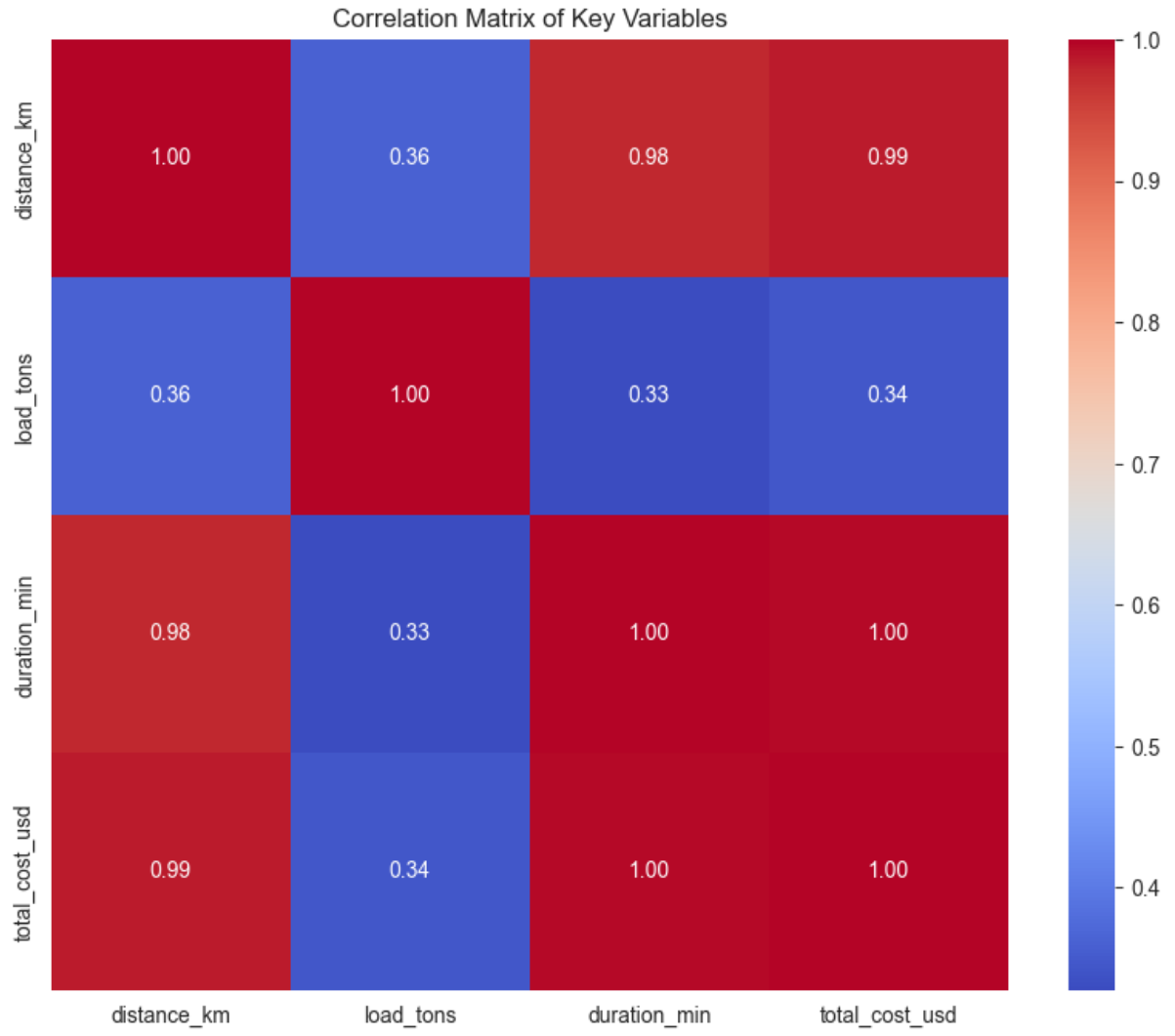

This scatter plot reveals that longer-distance routes tend to carry heavier loads. It also shows that routes with mountainous terrain are present across all distance ranges, while flat terrain routes are more common for shorter to medium distances.

**2. Exploratory Data Analysis (EDA)**

**I then performed an exploratory data analysis to uncover deeper insights into the factors influencing trip duration and costs.**

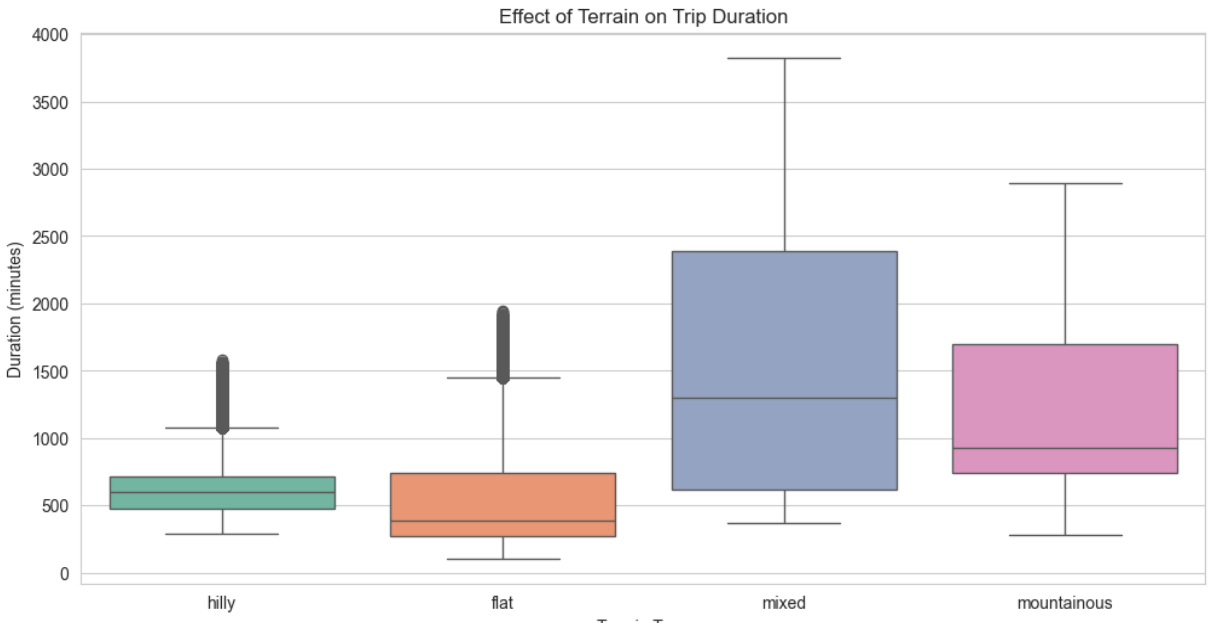
**Key Findings:**

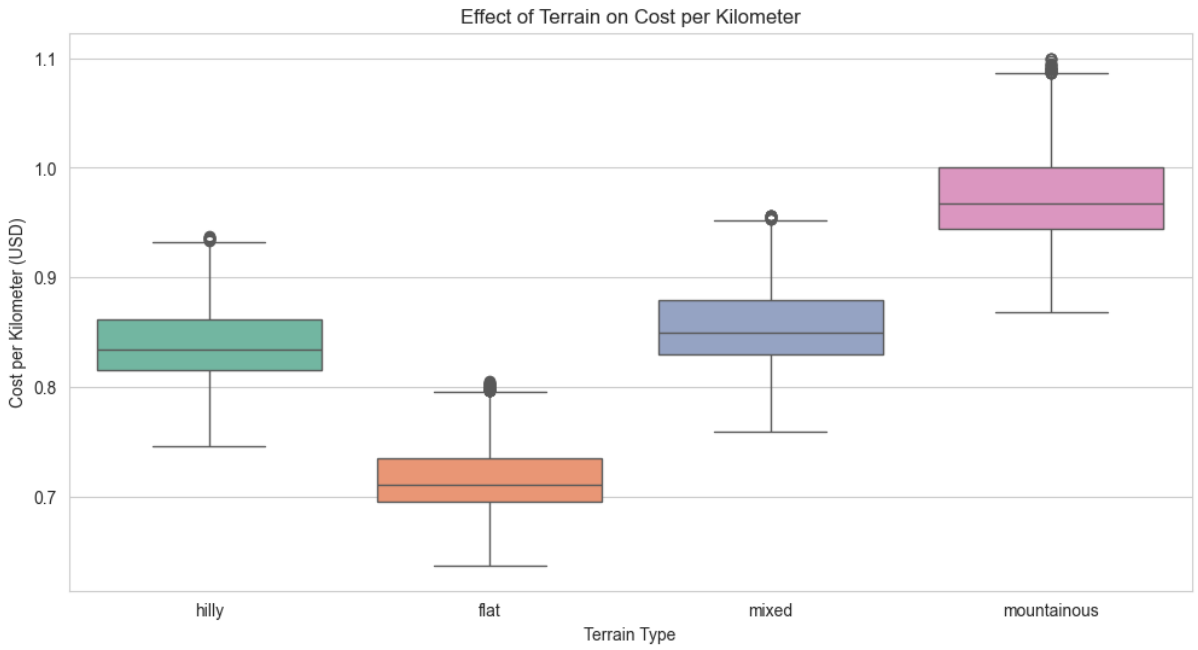
- **Impact of Distance, Load, and Terrain:**



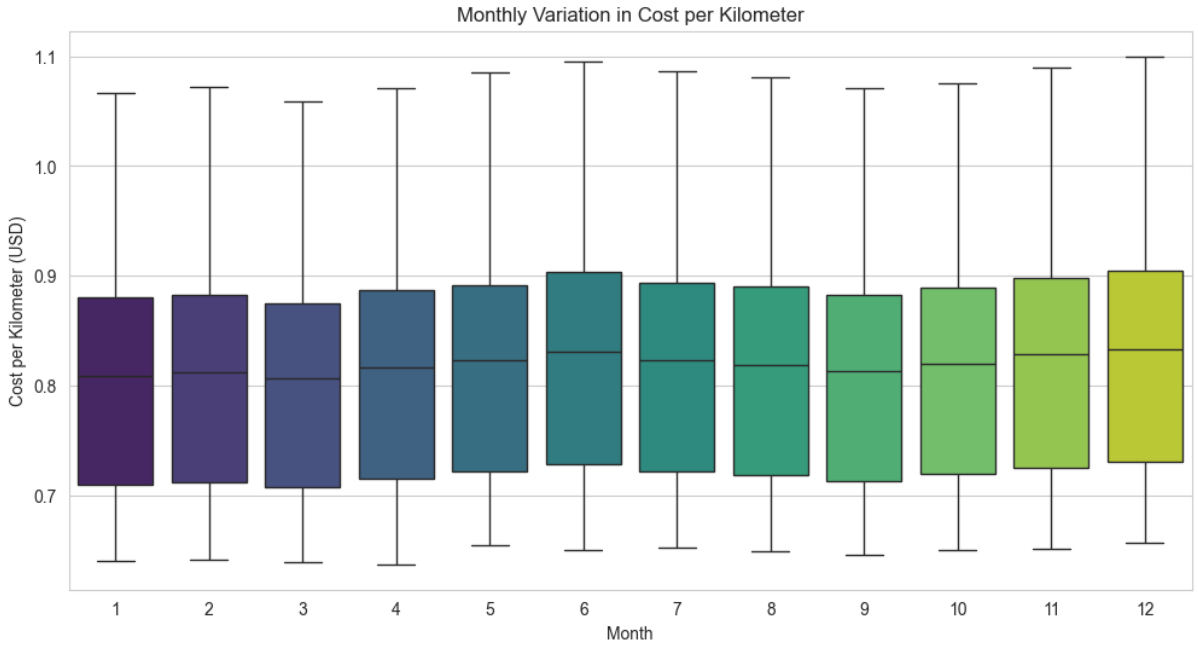

- There is a very strong positive correlation between distance and both duration and total cost, which is expected.

- **Terrain has a significant impact on efficiency.**
  - As you can see in the boxplots below, trips on mountainous terrain take longer and are more expensive per kilometer compared to trips on flat or hilly terrain.



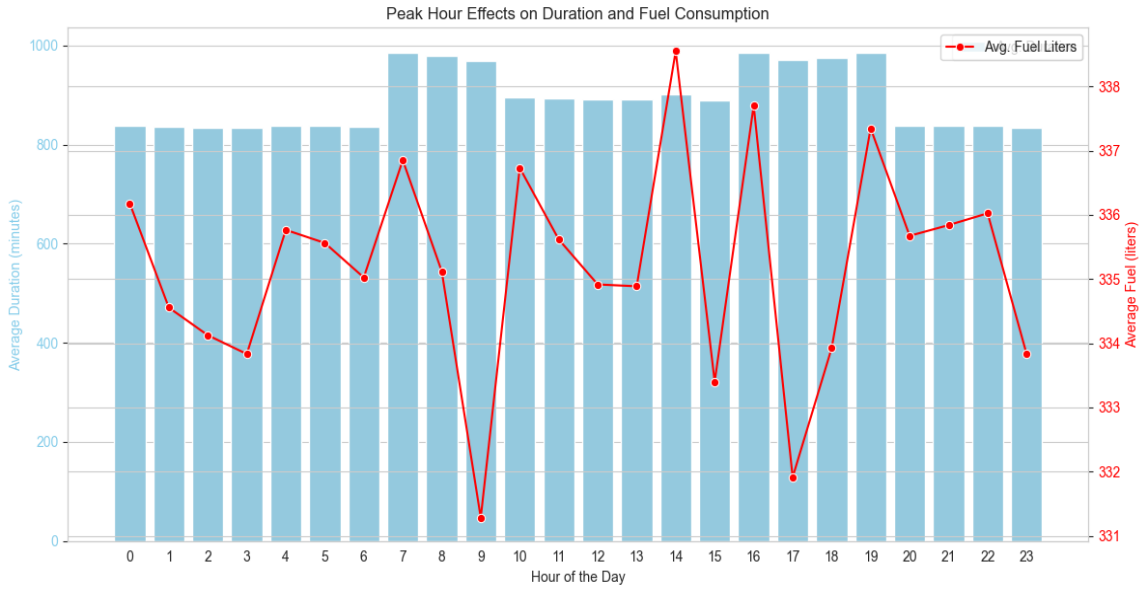



- **Seasonal and Temporal Patterns:**
    - **Monthly Variations:**




The cost per kilometer shows slight variations throughout the year, with a minor increase in the latter half of the year. This could be due to seasonal demand, weather conditions, or other factors.

- **Peak Hour Effects:**




The analysis of hourly data shows that trips starting in the early morning (around 5-7 AM) have a slightly longer average duration, which might be related to morning rush hour traffic in urban areas. Fuel consumption, however, remains relatively stable throughout the day.

- **Most and Least Efficient Routes:**
  - The most efficient routes are predominantly shorter-distance corridors on flat terrain, such as "San Diego, CA" to "Los Angeles, CA" (US040).
  - The least efficient routes are almost all located in mountainous terrain, such as "San Francisco, CA" to "Denver, CO" (US066), which are the most expensive on a per-kilometer basis.

| Category | Route ID | Origin | Destination | Avg. Cost/km (USD) | Terrain |
|---|---|---|---|---|---|
| **Most Efficient** | **US040** | **San Diego, CA** | **Los Angeles, CA** | **0.70** | **flat** |
| | **US041** | **San Francisco, CA** | **Sacramento, CA** | **0.70** | **flat** |
| | **US016** | **Chicago, IL** | **Kansas City, MO** | **0.71** | **flat** |
| **Least Efficient** | **US066** | **San Francisco, CA** | **Denver, CO** | **0.99** | **mountainous** |
| | **US058** | **Los Angeles, CA** | **Denver, CO** | **0.99** | **mountainous** |
| | **US069** | **Portland, OR** | **Denver, CO** | **0.99** | **mountainous** |

**3. Clustering Analysis**

To identify natural groupings of routes with similar characteristics, I performed a K-Means clustering analysis on the aggregated route-level data. Based on the "Elbow Method," I chose to group the routes into 4 distinct clusters.

**Interpretation of Clusters**

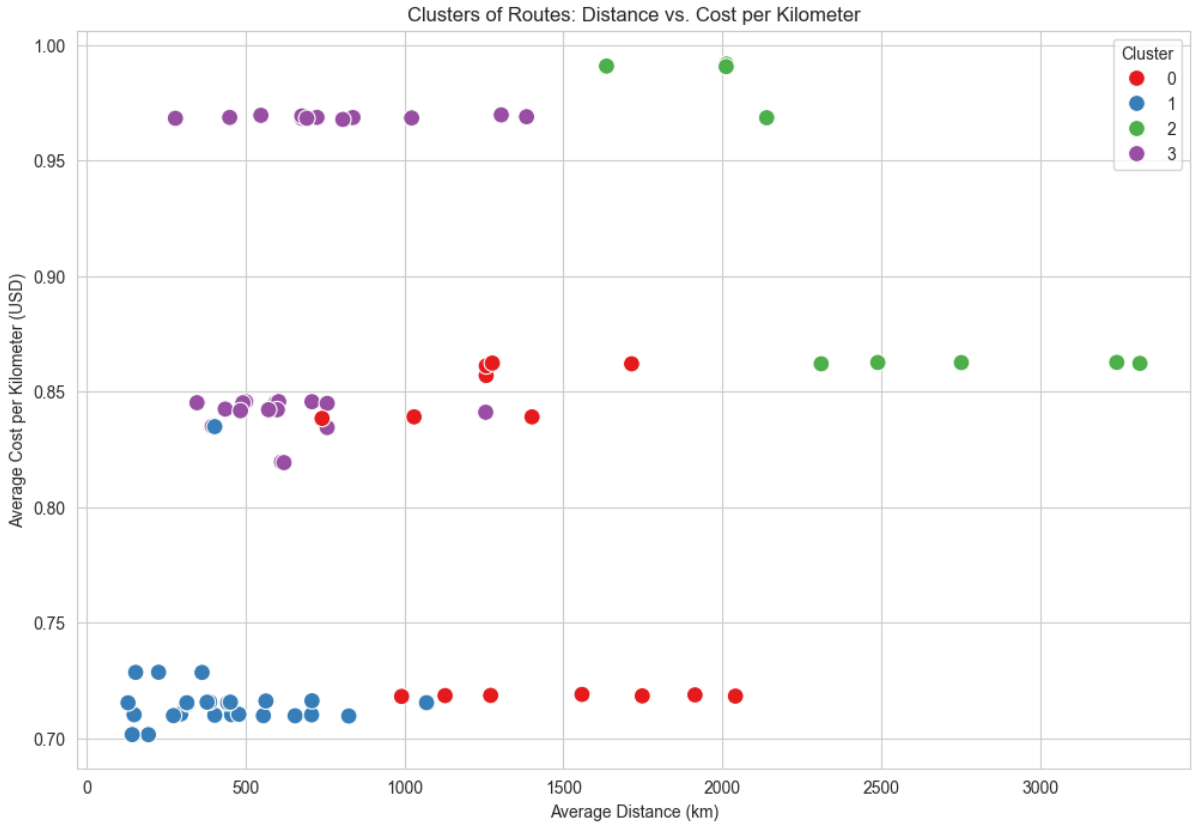
The clustering revealed a clear and meaningful segmentation of the routes, which can be interpreted in business terms as follows:

- **Cluster 0: Long-Haul, Mixed-Terrain Routes (14 routes)**
    - Characteristics: These are long-distance routes (average of 1380 km) with high loads. They have moderate costs and are mostly on mixed or hilly terrain.
    - Business Interpretation: These are major cross-country corridors that form the backbone of the logistics network.
- **Cluster 1: Short-Haul, Low-Cost Routes (26 routes)**
    - Characteristics: These routes are short (average of 424 km) and have the lowest average cost per kilometer. They are predominantly on flat terrain.
    - Business Interpretation: These are highly efficient, regional routes, likely used for local distribution and "last-mile" logistics.
- **Cluster 2: Ultra Long-Haul, High-Cost Routes (9 routes)**
    - Characteristics: This cluster contains the longest routes (average of 2435 km) and also the most expensive ones. They often traverse mountainous terrain.
    - Business Interpretation: These are challenging, transcontinental routes that are costly to operate but essential for connecting distant economic centers.
- **Cluster 3: Mid-Haul, High-Cost/Hilly Routes (28 routes)**
    - Characteristics: These are medium-distance routes (average of 683 km) with high costs, almost always on hilly or mountainous terrain.
    - Business Interpretation: These are regional connector routes that are more expensive to run due to challenging terrain, requiring more fuel and maintenance.

**Visualizing the Clusters**

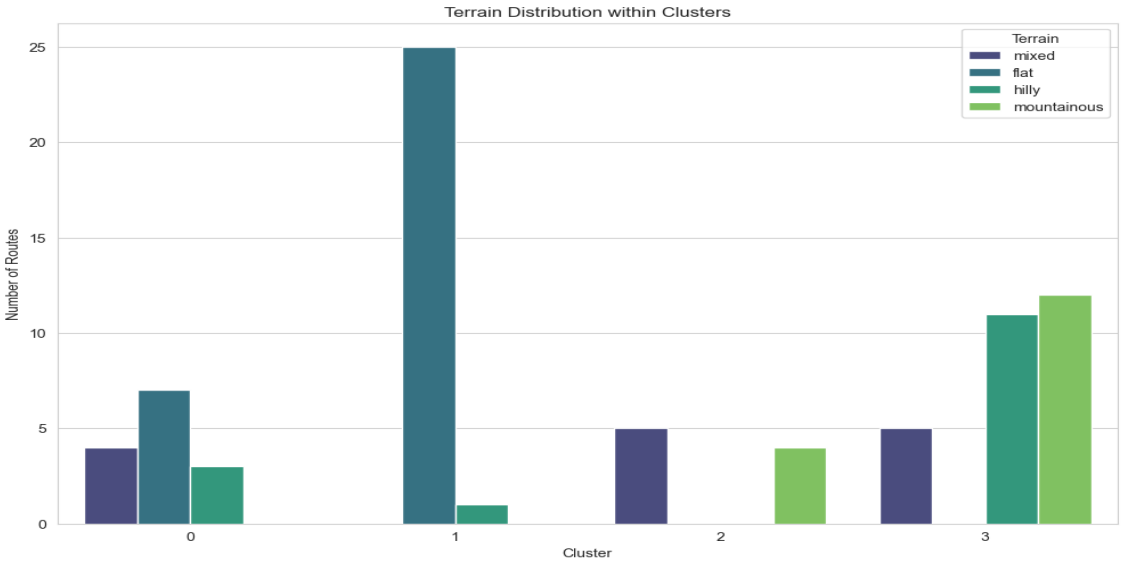
The following visualizations confirm the interpretation of the clusters:

- **Clusters by Distance and Cost:**




The scatter plot clearly separates the clusters based on their average distance and cost profiles.

- **Terrain Distribution within Clusters:**




The count plot shows a strong relationship between the clusters and terrain type. For example, Cluster 1 is almost entirely flat routes, while Clusters 2 and 3 are dominated by mountainous and hilly terrains.

**Conclusion**

**This analysis provides a comprehensive overview of the trucking operations described in the dataset. The key takeaways are:**

- **Efficiency is Driven by Terrain and Distance:** Flat terrains and shorter distances are strongly associated with lower costs and higher efficiency.
- **Clear Route Segmentation:** The clustering analysis successfully grouped the routes into four distinct operational categories (short-haul, long-haul, etc.), which can be used for strategic planning, resource allocation, and pricing.
- **Actionable Insights:** The findings can help in optimizing logistics, such as by adjusting pricing for more challenging routes, planning fuel stops more effectively based on hourly consumption patterns, and identifying opportunities for cost savings on the most and least efficient corridors.